THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 57 October 1978 Number 8

A Statistical Analysis of Bell System Building Fires, 1971–1977

By A. R. ECKLER

(Manuscript received December 23, 1977)

About 200 fires in Bell System buildings and adjacent grounds (excluding Western Electric) are reported to AT&T each year; the actual number of fires that occur may be somewhat higher. The dollar damage of reported fires (excluding only the $60 million fire in New York City on February 27, 1975) is reasonably modeled above the median by a log normal probability density function. This paper introduces a detailed taxonomy of fires, showing substantial differences in their frequency and costliness. The paper concludes with various special topics: (i) an analysis of employee injuries and service interruptions caused by fire; (ii) the correlation of business hours with fire frequency and building occupancy with fire severity; (iii) the methods used to fight fires; (iv) an analysis of multiple fires in buildings and of a cluster of fires in the Greater New York area in March 1975.

I. INTRODUCTION

Fires occurring in Bell System operating company or Long Lines telephone buildings or adjacent grounds, but not fires in Western Electric plants, are reported to AT&T on a standard form entitled "American Telephone and Telegraph Company Fire Report—Buildings" (Form E-5000), issued in 1962 and revised in 1969 and 1976. This paper analyzes approximately 1500 of these reports, covering fires that occurred in the years 1971 through 1977. Although Bell System summary statistics on fires go back a decade or more, few reports on fires prior to 1971 are, apparently, now available. The issuance of a fire report is governed by the following definition of a fire:

> Any occurrence that produces heat or flame and smoke in telephone company property or leased space, that affects service, causes property or equipment damage, and/or endangers inhabitants.

For the purpose of this paper, this has been interpreted to mean that a fire is characterized by an open flame, arcing, or sparks, visible smoke, or a combination of these; if the fire is out before it is detected, the site is marked by ashes, charred areas, or discoloration. Furthermore, an explosion is counted as a fire. However, a burning odor unaccompanied by smoke that cannot be traced to evidence of the above nature is not counted as a fire. All fires on company-occupied premises, either owned or leased, are supposed to be reported. Specifically, in this paper a fire is included if (*i*) it begins on non-Bell property and spreads to Bell property, damaging it (including water damage by fire fighters) or (*ii*) it occurs in a vehicle parked on Bell premises. However, a fire is not included if (*i*) it occurs in a PBX, telephone closet, or similar place in a building owned or leased by the telephone user, (*ii*) it occurs in a Bell-owned car or truck off Bell premises, (*iii*) it occurs on adjacent property but does not spread to Bell property, (*iv*) it occurs in Bell System outside plant such as manholes and cables, or (*v*) it occurs in a Bell-owned building not being used for telephone purposes and slated for eventual demolition and replacement (often, these are vacant but sometimes they are rented to tenants for a few months). Car and truck fires off Bell premises are recorded on a standard form entitled "American Telephone and Telegraph Company Fire Report—Motor Equipment" (Form E-5000 ME).

It is quite clear that reported fires do not represent all the building fires in the Bell System under the above definitions. Very small fires, such as a lighted match dropped on a carpet and immediately stamped out, are rather unlikely to be reported. Also, since fires are relatively rare events, employees may not be aware of the reporting procedure to be followed.

The "Report of Abnormal Service Conditions" (Form E-3877), telephoned to AT&T by operating companies when telephone service is threatened or interrupted, includes a few building fires (10 to 15 per year). Table I shows that over two-thirds of these fires are also included in the fire reports. If service reports and fire reports are filed independently, this yields an approximate estimate of under-reporting of fires.

Table I — Reports of abnormal service conditions involving building fires

	1972	1973	1974	1975	1976	1977	Total
Fire report	8	13	5	11	11	7	55
No fire report	3	3	3	4	7	1	21

This paper is restricted to fires reported to AT&T, and the reader should keep this potential under-reporting of Bell System fires in mind. Specifically, always ask the question: Are reported fires typical of *all* fires with respect to the characteristic under discussion?

There are several reasons why damage information in the fire report should be regarded with caution. These estimates are highly rounded (to quantities such as \$100, \$200, \$500); furthermore, they are usually made a day or so after the fire, long before the actual bills are in. The dollar values presumably reflect replacement costs, and do not allow for depreciation; furthermore, there is no indication whether labor costs associated with clean-up and repair have been included. If the fire does less than \$32 or so in damage, there is a strong possibility that it will be rounded down to zero; over 30 percent of all fires are so reported. There is no reporting mechanism for providing AT&T with more accurate follow-up reports of fire costs. However, trends in costs and comparisons of different cost distributions should be relatively immune to these problems.

II. BELL SYSTEM BUILDING FIRE EXPERIENCE

This section, the core of the paper, summarizes Bell System building fire experience. The first two parts analyze year-by-year changes in the number of fires per year and the probability density function of fire damage for the Bell System as a whole. The third part presents a detailed taxonomy of fires, showing which kinds are most frequent or most costly.

2.1 Number of fires per year

Table II and Fig. 1 summarize the number of Bell System building fires (excluding 73 Bell Canada fires, since Bell Canada left the System in 1975) which occurred from 1971 through 1977. The upper set of points in Fig. 1 includes all fires, no matter how small the damage, but the lower set includes only those fires with reported damages of \$32 or more. A statistical chi-squared test on the Poisson counts[1] rejects (at the 0.02 level) the null hypothesis that the average number of fires per year is constant; in fact, the figure suggests that there has been a downward drift. However, this inhomogeneity can be explained by differential diligence from year to year in reporting very small fires, for the same test on the homogeneity of Poisson counts confirms that more expensive fires occur at a constant average rate of a little more than 100 per year.

Table II — Bell System fire frequency, 1971–1977

	1971	1972	1973	1974	1975	1976	1977
All fires	244	209	186	191	194	174	212
Fires over \$32	120	102	97	100	95	108	104

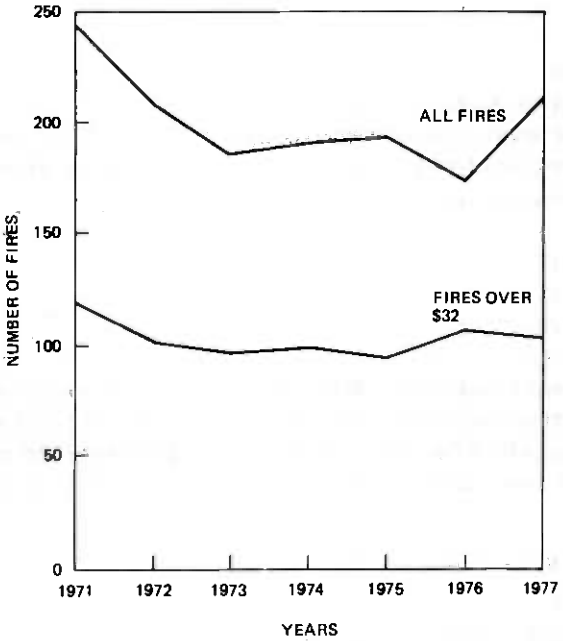


Fig. 1—Bell System building fire frequency, 1971–1977.

Even if inhomogeneities in the fire frequency had been revealed by the data, this would not necessarily have been reason for complacency or alarm. The Bell System continuously evolves in many ways; for example, the number of operating company employees and construction dollars have had year-to-year declines in the 1971–77 period. Furthermore, AT&T updated its fire protection policies and introduced fire-retardant materials in telephone equipment in the late 1960s; the benefits of these and similar actions have spread through the Bell System during the 1971–77 period. Changes in Bell System fire frequency, if they occur in the future, are likely to be complicated functions of many Bell System characteristics.

The important role played by small fires can be illustrated in another way. Counting C&P as one company, there are 19 operating companies plus Long Lines in the Bell System; it is a straightforward matter to calculate the expected number of fires for each of these in 1971–76 under the assumption that fires per million square feet per year are constant over the Bell System. One can then look at the likelihood of the actual counts, based on Poisson distributions having the expected counts as their means. It turns out that eight operating companies are in the upper 10-percent tail of the Poisson distribution (one at the 0.99999971 level), and six more are in the lower 10-percent tail of the Poisson distribution (one at the 0.0000002 level). However, if only fires greater than $32 are

considered, most of this inhomogeneity vanishes; three companies lie in the upper 10-percent tail and four in the lower 10-percent tail (two companies should be in each tail because of normal statistical fluctuations). There is strong evidence of differential reporting of small fires among operating companies as well as for different years; fires which do significant damage (more than $32) are more reliably reported than small ones.

It is somewhat more meaningful to normalize Bell System fire frequency by relating it to floor area. The overall Bell System fire frequency is approximately 0.7 fires per million square feet of floor space per year. If the 15 percent of external (and roof) fires are removed from the total, the rate is reduced to 0.6. (However, if allowance is made for unreported fires as discussed above, the rate is increased to 0.85.) Although quantitative comparisons of this with other industries are hard to ascertain, one study by Factory Mutual Research[2] for the Naval Facilities Engineering Command divides properties on naval bases into three risk categories: less than 1 fire per million square feet per year (communications facilities, clinics, electronic data processing facilities, hospitals, outside storage, offices, child care centers, schools, vacant buildings, mobile equipment, warehouses), between 1 and 3 fires (aerospace manufacturing facilities, churches, cold storage plants, laundries, cafeterias, stores, theaters), and more than 3 fires (gasoline stations, barracks, clubs, laboratories, utilities and power plants, homes, recreational areas). To get a breakdown of Bell System fire frequency by type of space, it is useful to subdivide Bell System floor space into analogous categories. However, this task is hampered by the lack of centrally compiled statistics on Bell System occupancy. Based on rough estimates of Bell System floor areas (obtained, in the first three cases, by scaling up New Jersey Bell floor areas), the number of reported fires per million square feet per year is about 0.3 in switchrooms, 1.0 in power rooms, 0.2 in cable vaults, 0.5 in Community Dial Offices, and 1.0 in repeater huts and microwave stations.

2.2 Probability density function of fire damage

It is useful to summarize the Bell System data on fire damage by means of a probability density function characterized by a small number of parameters. Because of the extremely wide range of fire damage (most are a few hundred dollars or less, but fires exceeding $100,000 have occurred each year, and the Second Avenue fire in New York in 1975 was valued at $60 million), it is necessary to restrict oneself to probability density functions in which the independent variable is expressed in logarithms. Two common ones exist—the Weibull and the log normal; the latter turned out to fit the data quite well and is the one that was eventually selected.

Figure 2 shows the empirical cumulative distribution functions of fire damage for each of the seven years. (The totals do not agree with those in Section 2.1 because some fire reports omit damage estimates.) Because of the already-mentioned tendency of fire reports to round small-damage fires down to zero, only the upper half of the distribution (above the 50th percentile) is shown. The ith largest fire damage for a year having n fires is plotted at the point (x,y) corresponding to (log dollars, $1 - (i - 1)/n$); to avoid clutter, only the fires corresponding to i = 1(1) 10(2) 20(5) 50(10) 100 have actually been plotted.

In view of the noticeably greater variability of the data in the upper tail of the distribution, it was decided to fit a straight line to each cumulative distribution by a least-squares line fitted to the 50th, 60th, 70th, 80th, and 90th percentiles listed in Table III. (In this regression, the independent variables are the Gaussian deviates of the percentiles, and the dependent variables are the fire damages in log dollars.) After it was found that the slopes of the six lines did not significantly differ from each other, the model was reformulated: seven lines parallel to each other were fitted to the data instead. This was accomplished by reducing the 50th through 90th percentiles of the 1972 through 1977 data to the 1971 level by subtracting the average difference of the damage, as shown in the final column of Table III. The common value of the standard deviation (the slope) is 1.386 (in log dollars), and the median fire damage is given for each year in Table IV.

It is frequently useful to know the average fire damage as well. Because of the skewed nature of the log normal distribution, this is ordinarily a much larger value than the median. If m and s represent the mean and standard deviation of the log normal distribution in log dollars, the mean of the distribution in dollars is given by the formula:[3]

$$M = \exp(m \log_e 10 + (s \log_e 10)^2/2).$$

The values of M for the various years are also given in Table IV.

Why go through this involved procedure to calculate the average fire damage when an unbiased estimate of this quantity can be easily obtained by taking an average of the recorded fire costs? Unfortunately, the variance of such an estimate is quite large, for it depends almost entirely upon the values of the half-dozen largest fires in the set. The estimate M given above is based on m and s, which are far more representative of all the fires in the sample, not just those in the extreme tail.

It is likely that the observed differences in the median fire damages from one year to another are, at least in part, due to the inflation in repair costs over these years. To obtain an estimate of the percentage inflation rate, one can fit a straight line to the estimated median fire values (in log dollars), using the year as the independent variable. The slope of this

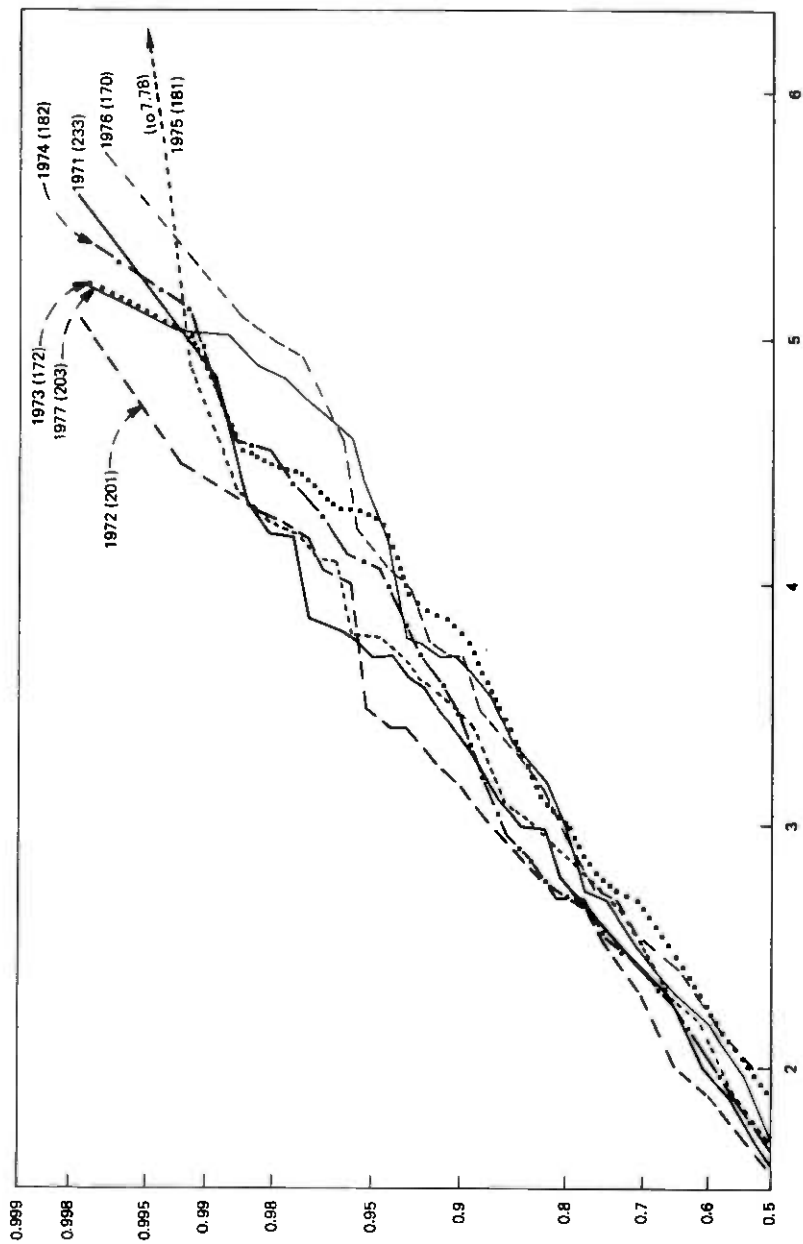


Fig. 2—Cumulative distribution function of fire damage, 1971–1977 (number of fires in parentheses).

Table III — Adjustment of fire damage (in log dollars) to 1971 level

Year	Percentile (Gaussian deviate) 50 (0.000)	60 (0.253)	70 (0.524)	80 (0.842)	90 (1.282)	
1971	1.54	2.00	2.40	2.72	3.37	
1972 original	1.59	1.88	2.30	2.70	3.18	
corrected	1.67	1.96	2.38	2.78	3.26	0.08
1973 original	1.86	2.18	2.69	3.00	3.79	
corrected	1.56	1.88	2.39	2.70	3.49	−0.30
1974 original	1.70	2.00	2.40	2.72	3.48	
corrected	1.65	1.95	2.35	2.67	3.43	−0.05
1975 original	1.65	2.10	2.42	2.88	3.43	
corrected	1.56	2.01	2.33	2.79	3.34	−0.09
1976 original	2.00	2.30	2.49	2.95	3.69	
corrected	1.72	2.02	2.21	2.67	3.41	−0.28
1977 original	1.70	2.18	2.48	3.01	3.70	
corrected	1.49	1.97	2.27	2.80	3.49	−0.21
Predict value	1.60	1.95	2.33	2.77	3.38	

Table IV — Median and average damage per fire for years 1971–1977

	Median	Average
1971	$40	$ 6481
1972	$33	$ 5390
1973	$79	$12931
1974	$45	$ 7272
1975	$50	$ 7973
1976	$76	$12349
1977	$65	$10511

fitted line is 0.04 in log dollars, or a rate of 10 percent per year, somewhat larger than the inflation rate corresponding to the well-known consumer price index from 1971 through 1977.

The 1975 New York fire, the most costly one in Bell System history, is not consistent with the log normal distribution. Its $60 million damage corresponds to a logarithmic cost of 7.778, which (after subtracting the average of the 1971–1977 median fire costs and dividing by the standard deviation) yields a standard normal variable equal to 4.37. This translates into a probability of only 0.0000062 that a fire randomly drawn from the log normal distribution will be this costly; if fires occur at the average rate of 200 per year, there is only a 50 percent chance that a fire this damaging will occur in 550 years corresponding to 1971–1977 experience: $(1-0.0000062)^{550(200)} = 0.506$.

Are there any other fires besides the New York one which are not consistent with the log normal distribution? Figure 3 depicts the 10 most costly fires in the Bell System during 1971–1977, compared with the 1971 log normal predicted line. To allow for inflation, all fires have been translated in damage values to hypothetical 1971 levels using the correction factors given in the final column of Table III. Clearly, the log-normal model fits all fires but the New York one reasonably well.

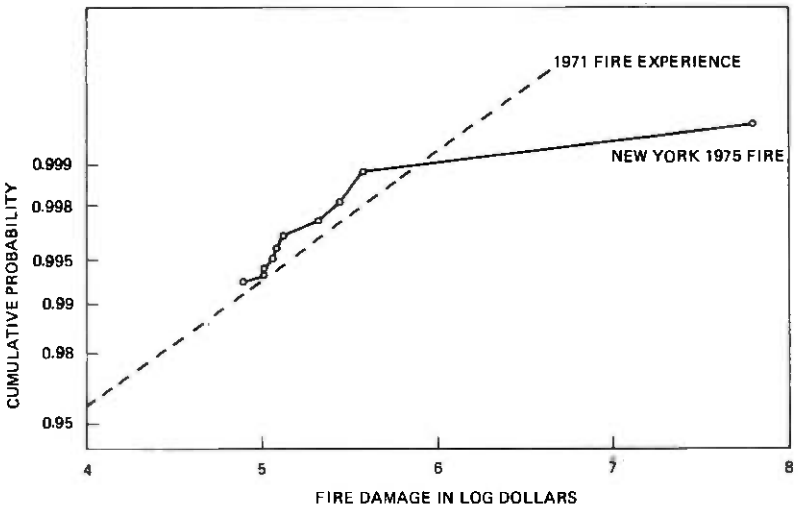


Fig. 3—Upper tail of fire damage distribution (10 most costly fires, 1971–1977).

Instead of believing that an event this unusual has occurred, the statistician prefers to conclude that the New York fire is a sample of one drawn from a probability density function of fire damage *different* from the ones shown in Fig. 2. With such a small sample, it is impossible to derive the distribution function of fire damage in the underlying population, nor is it easy to say what fraction of Bell System fires come from such a population. Although detailed statistics of the damage distribution of 3013 building fires during 1960–1970 are not available, the pattern of the most costly fires in that 11-year period ($1.67 million, $0.70 million, $0.50 million, $0.40 million, $0.19 million, $0.18 million, $0.15 million, $0.14 million, $0.12 million, . . .) is consistent with Fig. 2. Therefore, the fraction of building fires in the Bell System that do *not* follow this damage pattern is very small—one out of 3013 + 1483 = 4496 fires in 18 years! The best estimate of the fraction is 0.0002, and a 95-percent confidence interval enclosing the true but unknown fraction is (0.0000056, 0.00125).

It is hazardous to characterize the variability of the cumulative distribution function of fire damage based on only seven fitted lines, one for each year. The standard deviation of the difference between an observed percentile and its fitted value is approximately 0.07 in log dollars, for percentiles between 50 and 90. However, this is a misleadingly small number if one is interested in predicting the fire damage corresponding to a specified percentile of the cumulative distribution in a future year. As already noted, there is an increase of about 10 percent per year in fire costs, but the year-to-year fluctuation of the fitted lines around this 10-percent rate is considerable. For example, if one attempted to predict the fitted median fire damage of 1972 on the basis of the fitted median

fire damage in 1971 with 10-percent inflation added, one would have overestimated by 0.12 in log dollars—inflation suggested an increase of 0.04, but in reality the 1972 fitted line was 0.08 below the 1971 one. Other year-to-year errors are even larger, and a rough estimate of the standard deviation of the error in estimating the following-year fit from the preceding-year one is 0.2 in log dollars. The standard deviation of the error in predicting an actual percentile in a future year is the root-mean-square sum of these two standard deviations, or again about 0.2 (since the smaller error is swamped by the larger). In dollars, this corresponds to a multiplicative factor of 1.6; thus, if one estimates a future fire damage to be (say) $1000, there is only a two-thirds chance that the actual fire damage will lie between $600 and $1600. If one is interested in predicting extreme percentiles, such as the damage of the largest fire to be expected in a future year, the errors are likely to be far larger.

2.3 Taxonomy of fires

Bell System building fires can be classified in a number of ways. Table V and Fig. 4 present a hierarchical classification in which fires are first sorted out by place of origin (under control of Bell System employees, or not under their control), then by fire type, and finally by the equipment in which it originated. Fire type is related to, but not identical with, the well-known classification of fires by fire extinguisher type: paper fires, electrical fires, oil and grease fires. A fire needs three things to ignite—oxygen, a fuel, and a source of heat; if the fuel is especially volatile, almost any source of heat will do the job, but if it is less volatile, the particular source is of greater concern. The following two-level description of fire type has been adopted:

- Fires with volatile fuels (oil, gas, gasoline, grease, etc.) regardless of heat source.
- Fires with less volatile fuels (paper, wood, insulation, etc.).
  - Electrical sparks or short circuits.
  - Overheating (placing a flammable substance too near a properly functioning heat-producing source, as a chimney or space heater).
  - Heat-generating tools (used too near flammable substances).
  - Smoking and matches (whether deliberately set or accidental).

Classification is not always as simple as this would suggest; for instance, electrical malfunction sometimes results in an overheated resistor which actually starts the fire (these fires have been classed as electrical). Note that fires in certain equipment can appear in several different places in Table V or Fig. 4. For example, a fire in a furnace will appear under Volatile (Fuel Oil) if this is involved; otherwise, it will appear under Electrical (Building Equipment, Furnace). Similarly, a fire in a stove can appear under Volatile (Grease), Overheating (Stoves), or Electrical

Table V — Explanation of fire taxonomy

External to Bell System: earthquakes, lightning strokes, power wire crosses and surges, fires beginning on non-Bell property, water main breaks—but not fires caused by interruption of commercial power

Internal to Bell System

Volatile Fuels

Fuel Oil: boiler explosions, fuel oil line leaks, oil in cans or on floor

Gas: explosions of gas furnaces, propane heaters, gas pipeline breaks

Gasoline: gas pumps, Bell System or employee vehicles located on Bell property—but not on assignment away from Bell property

Tar Kettles: contractor fires

Hot Grease: grease or fat associated with stoves and grills

Other: floor sealers, adhesives, calcium hypochlorite, windshield washing solvent, butane lighter, anti-static spray, oxygen tank, lacquer thinner, etc.

Nonvolatile Fuels

Overheating

Furnaces, Heaters: gloves on furnaces, cartons stored nearby—not including fires of volatiles

Stoves: papers in vicinity, coffee pot overheating, food bag in microwave oven—not including fires of volatiles

Light Bulbs

Engine Exhaust: ignition of building structure adjacent to emergency engine exhaust pipe

Heat-Generating Tools: acetylene and propane torches, soldering irons, grinding wheels, Cadwelders (including hot solder deposited in waste containers)—not including fires involving volatiles

Smoking, Matches

Trash Containers: ashtrays, wastebaskets, janitor carts and bags, scrap wire bags, rubbish rooms, trash compactors

Loose Paper: fires in paper or wood scraps not in trash containers (often regarded as due to arson)

Mops, Cloths: fires in janitorial closets caused by mops picking up smoldering cigarettes

Chairs, Beds, Drapes: fires in upholstered furniture in lounges or quiet rooms (often attributed to smoking)

Cable Well Bags

Paper Records, Cartons: a heterogeneous category including fires in paper supplies, books, bulletin boards, etc.

Nonpaper Supplies: fires in stored telephone supplies containing no obvious paper or cardboard (often regarded as arson, as ignition by cigarette is not easy)

Outside Fires

Trash Containers: truck-away containers in parking lots; piles of loose lumber or trash associated with construction activity

Grass, Shrubs

Vehicles, Telephone Equipment: night deposit boxes, cartons stored outside, employee cars, cable reels, plastic conduit (often regarded as arson)

Construction Activity: miscellaneous fires in construction areas not obviously associated with volatiles (tar kettles), trash piles or heat-producing tools (usually attributed to smoking)

Construction Supplies, Roofing

Electrical

Building Power

Vaults, Transformers: commercial power entrance facilities, including transformer vaults and entrance ducts

Panels, Electrical Closets: main commercial power switchboard, and branches terminating in panel boxes (wall-mounted, sometimes in separate closets)

Local Wiring: fires in distributive wiring of commercial power, including plugs in sockets—but not fires in appliances or known to be in fluorescent lights

Building Appliances

Fluorescent Lights: defective ballasts

Local Air Conditioning: window air conditioners or free-standing room units, humidifiers

Heaters: either portable or wall-mounted types

Fans: ceiling-mounted exhaust fans, pedestal fans, portable fans

Building Equipment

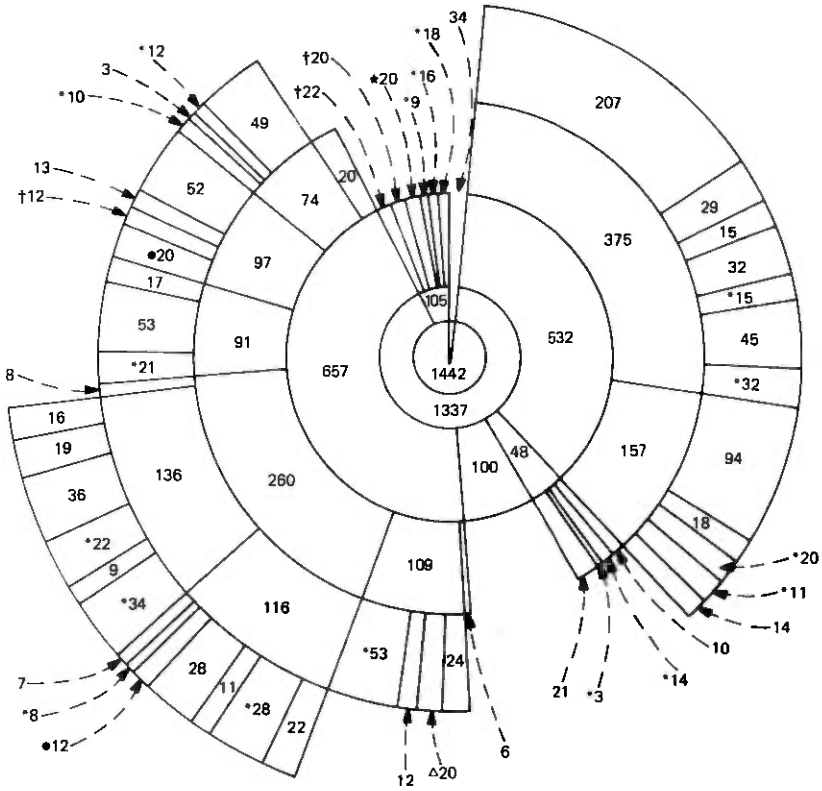
Table V (cont)

Central Air Conditioning: fan motors, compressors, condensers, chilled water pumps
Elevators: motors, control circuits (including dumbwaiters)
Furnace: fires not obviously associated with volatiles, including mechanical failure also (belt leaving sheave)
Other: air dryers, vacuum pumps, motor controllers and control centers, ultrasonic cleaning machine, electric toilets, fire pump control cabinets, garage air compressors, portable battery chargers, sump pumps
Food Appliances: coffee pots, stoves, portable defrosters, refrigerators and freezers, water coolers, sandwich and vending machines—not including fires attributed to volatiles or overheating
Office Appliances
Copiers: fires in office copiers, including ones which are attributable to paper jams as well as electrical malfunction (source often difficult to determine from report)
Computers, Calculators: desk calculators, computer processing equipment
Teletype: those not directly associated with switching equipment
Other: offset press, envelope inserter, CRT service order machine, enclosing machine, assignment wheel, typewriter, conveyor belt motor
Automobiles: fires not associated with gasoline
Telephone Equipment
Cable Vault: fires in cabling (usually in open splices)
Power Room Equipment
Emergency Engine: load boxes, alternators, start motors, switches and related controls—not fires caused by overheating of exhaust duct
Battery: electrolytic leakage or cell overheating, and fires in associated circuitry
Generator: in motors or associated control circuitry
Rectifier: includes converter and inverter fires
Power Plant: fires in 130-volt power panels, or in general power controls such as the 412B power plant or Uninterrupted Power Source equipment
Cabling: principally in DC power cables, and often due to craftsman error
Switchroom Equipment
Main Distributing Frame: usually fires in open splices
Test and Operator Boards
ESS Switchers: includes TSPS, and closely associated equipment such as teletypewriters
Carrier: primarily N and T carrier, and often in unattended remote locations
Radio: includes mobile radio, and often in unattended remote locations
EM Switchers: includes closely associated equipment such as teletypewriters; fires usually in relays, markers, fuses, step-by-step switches, line finders, etc.
Other: fires in auxiliary equipment, and incompletely identified switchroom fires (most of these probably associated with EM switching)

(Food Appliances), depending upon the nature of the fire. Note also that this taxonomy cannot be used to determine the number of fires that occur in a given type of Bell System space (switchroom, power room, utility room, cafeteria, hall, etc.); in general, a wide variety of different fires can occur in a given location.

In Table V and Fig. 4, 67 building fires occurring in Bell Canada from 1971 through 1974 are also included to provide as large a statistical base as possible.

Figure 4 depicts the relative frequency of different types of Bell System building fires. The relative seriousness of these different types is presented in Figs. 5 and 6, in which the damage of each fire type is plotted (on triangular graph paper) according to a trinomial probability density



3	TELETYPE
*3	LIGHT BULBS
6	AUTOMOBILES
7	M. D. FRAME
8	CABLE VAULT
*8	TEST, OPERATOR
9	POWER PLANT
*9	TAR
10	FURNACES
*10	OTHER
11	RADIO
*11	CONSTRUCT. ACT.
12	FURNACES
*12	COMPUTERS
†12	HEATERS
•12	ESS
13	LOCAL AIR COND.
14	CONSTRUCT. SUPP.
*14	STOVES
15	MOPS, CLOTHS
*15	CABLE WELL BAGS
16	EMERGENCY ENGINE
*16	GREASE
17	VAULTS, TRANS-FORMERS
18	GRASS, SHRUBS
*18	OTHER
19	BATTERIES
20	FOOD APPLIANCES
*20	TELCO EQUIP.
†20	GAS
•20	PORTABLE FANS
★20	GASOLINE
△20	ELEVATORS
21	ENG. EXHAUST
*21	LOCAL WIRING
22	OTHER
*22	RECTIFIERS
†22	FUEL OIL
24	OTHER
28	CARRIER
*28	ELECTRO-MECHANICAL
29	LOOSE PAPER
32	CHAIRS, BEDS, DRAPES
*32	NON-PAPER SUPPLIES
34	EXTERNAL
*34	CABLING
36	GENERATORS
45	PAPER RECORDS, CARTONS
48	OVERHEATING
49	COPIERS
52	FLUORESCENT LIGHTS
53	PANELS, ELECTRICAL CLOSETS
*53	CENTRAL AIR CONDITIONING
74	OFFICE APPLIANCES
91	BUILDING POWER
94	TRASH
97	BUILDING APPLIANCES
100	HEAT GENERATING TOOLS
105	VOLATILE
109	BUILDING EQUIP-MENT
116	SWITCHROOM
136	POWER ROOM
157	OUTSIDE
207	TRASH IN CON-TAINERS
260	TELEPHONE EQUIP-MENT
375	INSIDE
532	SMOKING, MATCHES
657	ELECTRICAL
1337	HEAT SOURCES
1442	INTERNAL

Fig. 4—A taxonomy of Bell System building fires, 1971–1977 (1476 total, including Bell Canada, 1971–1974).

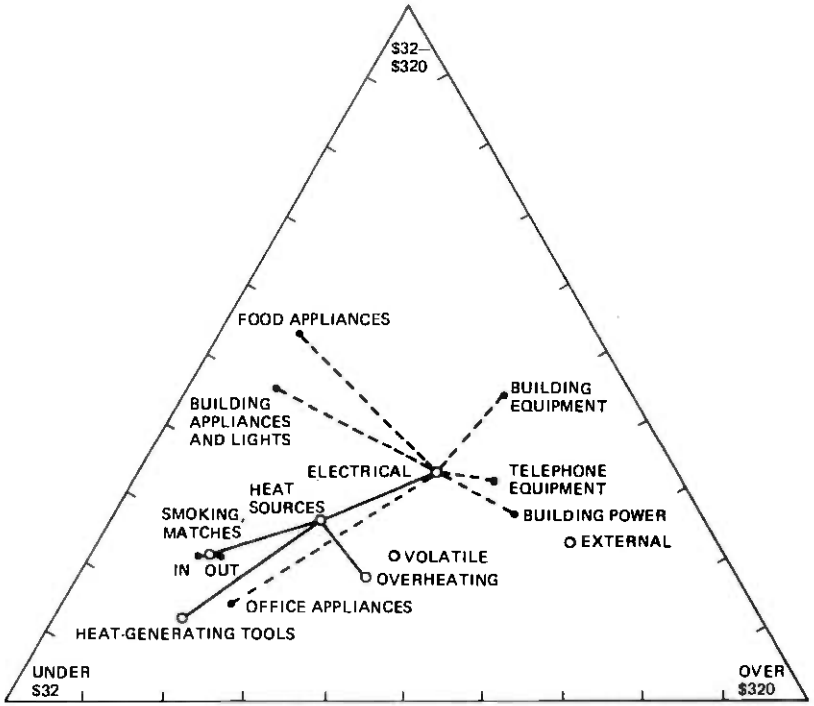


Fig. 5—A trinomial distribution of fire damage.

function: the fraction of fires less than $32, the fraction of fires between $32 and $320, and the fraction of fires above $320. (These limits were selected because they divide all Bell System building fires into approximately equal parts.) Thus, points plotted near the lower left corner of Fig. 5 or 6 correspond to fires with typical damage under $32, and points plotted near the lower right corner, to fires with typical damage over $320.

By using Figs. 4, 5, and 6 in concert, one can learn a great deal about the impact of different fires upon the Bell System. For example, inside trash container fires are relatively common (207 fires in seven years), but rarely cause much damage (73 percent under $32); on the other hand, battery fires are considerably rarer (19 fires in 7 years) but are far more costly when they do occur (53 percent over $320).

If fire frequencies are examined year by year for each of the categories in the taxonomy, few patterns of interest emerge. However, the decline in fires related to construction and installation activity is noteworthy. Although overall construction dollars discounted for inflation have remained in a narrow range from 1971 through 1977 ($5.6 to $6.8 billion, in 1967 terms), construction-related fires have declined in every year, from 44 in 1971 to 7 in 1977. The four most relevant categories are listed in Table VI.

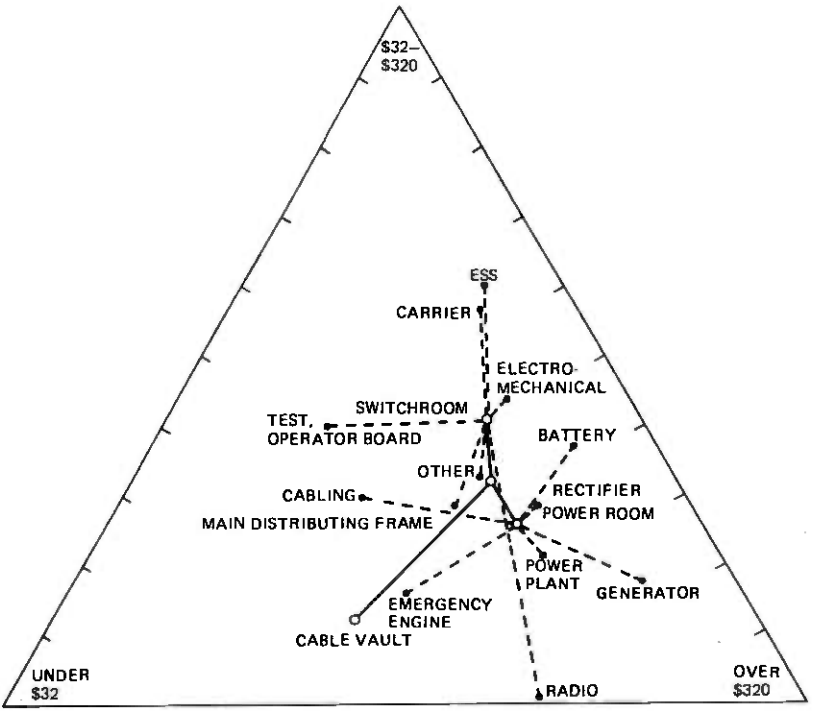


Fig. 6—A trinomial distribution of telephone equipment fire damage.

Table VI — Decline in fires related to construction activity

	1971	1972	1973	1974	1975	1976	1977
Outside fires (construction activity)	6	1	3	0	0	1	0
Outside fires (construction material)	6	3	2	2	0	1	0
Heat-generating tools	24	20	18	15	10	8	5
Power room equipment (cabling)	8	12	2	2	5	3	2
Total	44	36	25	19	15	13	7

III. SPECIAL STUDIES OF BUILDING FIRES

This section of the paper shows how the information on the fire report form can be used to carry out various studies involving specific aspects of Bell System building fires. In particular, four topics are introduced: (*i*) an analysis of employee injuries and service interruptions caused by fires, (*ii*) the correlation of business hours with fire frequency and the correlation of building occupancy with fire severity, (*iii*) the methods employed in fighting Bell System building fires, and (*iv*) an analysis of correlated fire events in buildings, and the cluster of fires in the New York City area in March 1975. These topics are meant to be illustrative rather than exhaustive; others could easily be developed.

3.1 Definitions of fire damage

Dollar loss is, perhaps, the most well-known definition of fire damage, but it is not the whole story. In the Bell System, two other measures may also be appropriate:

(*i*) With the strong Bell System emphasis on safety on the job, it is of interest to study injuries to employees occurring as the result of fires.

(*ii*) With the strong Bell System emphasis on reliable service to the customer, it is of interest to study service interruptions occurring as the result of fires.

Of the 1483 fires between 1971 and 1977 (including Bell Canada) for which reports were filed, one fire resulted in the death of a contract (non-Bell) worker servicing an air-conditioning system and the injury of three Bell System employees, and 19 other fires resulted in the injury of a total of 5 contract workers and 16 Bell System employees. There are several possible reasons for the relatively high injury rate of contract employees: They may typically work with more volatile substances (eight of the fires, including four of the five fires involving contract personnel, were characterized by explosive ignition of highly volatile liquids or gases), and they are less likely to be well-instructed in safe working procedures and well-motivated to follow them than Bell employees.

These injury statistics can be put in broader perspective by comparing them with the 1976 estimates of fire injuries and deaths throughout the United States prepared by the National Fire Protection Association.[4] There were 2.94 million fires in the United States resulting in 108,000 injuries and 8,800 deaths; if Bell System fire experience was comparable, the 1483 reported fires would have resulted in 55 injuries (instead of 24) and 4.5 deaths (instead of 1). Looking at the data from a different perspective, 600,000 operating company and contract employees working 8 hours per day, 5 days per week, have 1/1470 of the potential exposure to fire of 210,000,000 United States residents living 24 hours per day, and therefore should incur 73 injuries and 6 deaths per year (instead of 3.4 and 0.1, respectively). Even though the environment and characteristics of United States residents and Bell employees are markedly different, it is clear that the Bell System has an excellent safety record with respect to fires when they do occur.

Service interruptions due to building fires, although rare, are slightly more frequent than injuries. Interruptions can broadly be divided into two classes:

(*i*) Fires that destroy interoffice trunk circuits resulting in possible delays caused by increased congestion on alternate routes.

(*ii*) Fires that deny service to individual telephones.

It can be argued that the latter loss is of much greater importance to the Bell System; as long as the delays are not large, subscribers may not even be aware of the former impediment.

Clearly, some buildings are more likely than others to be the site of fires resulting in service impairment; garages and office buildings have little or no likelihood of this. More specifically, some areas of a central office are more vulnerable than others to service-impairing fires—the areas of greatest risk are those containing individual subscriber lines (cable vault, main distributing frame, first stage of switcher) or those which contain unduplicated equipment.

If a fire denies service to individual telephones, a rough measure of its impact is the number of days of lost service multiplied by the number of exchanges affected, called exchange-days for brevity. (An exchange can contain up to 10,000 lines, not including extension telephones or PBXs, but the number of assigned lines in a typical exchange will be considerably less.) Fire reports do not call for information on service interruption, so that more precise measures than exchange-days are hard to calculate; often, this is only a rough estimate. Table VII lists the most serious service-impairing fires (as measured by exchange-days) encountered between 1971 and 1977. In addition to the 22 fires in this table, 20 more fires affected trunks, principally carrier and radio circuits, for various lengths of time.

3.2 Some relationships between fires and people

Two truisms associated with fires are: (*i*) fires are at least in part caused by human activity, and consequently are more frequent during those hours that a building is occupied, and (*ii*) fires are less costly if they can be detected and fought quickly. Thus, one expects a few costly fires at night or on weekends (or at unattended buildings, such as Community Dial Offices or repeaters), and numerous but inexpensive fires at attended buildings during business hours. To what extent do the data support these truisms?

Table VIII shows there is a mild (but statistically significant at the 0.002 probability level, using a chi-squared test of goodness of fit)

Table VII—Building fires which impaired service to individual subscribers 1971–1977

Feb 27, 1975	New York, N.Y.	270 exchange-days
Nov 10, 1971	New York, N.Y.	1.5 exchange-days
Feb 11, 1971	Long Island City, N.Y.	1.0 exchange-days
May 19, 1973	Peekskill, N.Y.	0.5 exchange-days

In addition, there were 18 fires in operating companies which resulted in less than 0.1 exchange-days of service impairment.

Table VIII — Frequency of building fires by month

January	116	April	139	July	133	October	132
February	134	May	116	August	147	November	99
March	144	June	105	September	112	December	114

Table IX — Frequency of building fires by day of week

Saturday	122	Monday	236	Thursday	282
Sunday	96	Tuesday	240	Friday	253
		Wednesday	252		

seasonality to fires, with maxima in March and August and minima in June and November. Table IX demonstrates that there is a strong difference in fire incidence between business days and weekends; weekend fires occur less than half as frequently. However, there are no statistically significant differences among the different weekdays. Finally, Table X and Fig. 7 exhibit a strong relationship between fire frequency and the time of day, with a minimum around 5 a.m. and a broad maximum around noon. Note that the fire incidence rises steeply in the morning, but falls off much more gradually at night. This function follows fairly closely the number of on-premise employees (including contract labor), with a delay factor to allow for the fact that a certain number of fires smolder awhile before being discovered.

Table X — Time of day of discovery of building fires

Night and Morning				Afternoon and Evening			
12–1	30	6–7	30	12–1	97	6–7	74
1–2	24	7–8	47	1–2	98	7–8	56
2–3	26	8–9	77	2–3	102	8–9	55
3–4	24	9–10	100	3–4	84	9–10	52
4–5	16	10–11	96	4–5	108	10–11	43
5–6	23	11–12	105	5–6	70	11–12	37

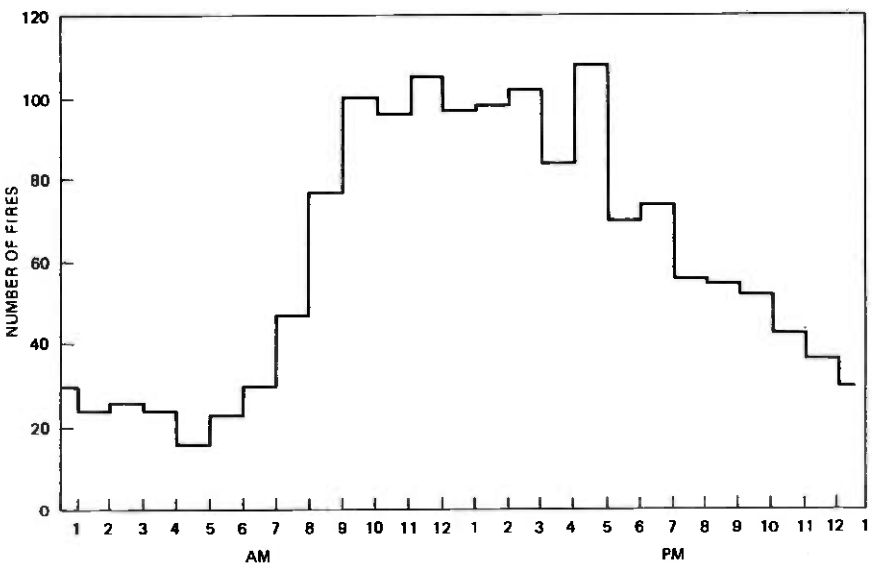


Fig. 7—Frequency of fires as function of time of day.

From 1971 through 1977 (Bell Canada fires 1971–1974 only), a total of 159 fires occurred in buildings unoccupied at the time the fire broke out. Of these, 44 fires self-extinguished, 4 were put out by sprinker systems, and 1 by a Halon system. (In one of the remaining fires, a sprinkler system inside the building was activated by the heat of an external fire, but it played no role in putting the fire out.) Fitting a log probability density function by a least-squares line as described in Section 2.2 (to the 50, 60, 70, 80, 90, and 95 percentiles) the estimated median fire damage turns out to be 2.88 ($630), and the slope is 1.316 (in log dollars). In other words, the median damage of a fire in an unoccupied building is approximately 10 times as large as the median damage of a fire in an occupied building. Put another way, out of the 26 Bell System fires of at least $50,000 damage, 15 were in unoccupied buildings; this supports the statement that about half the Bell System building fire damage occurs in unoccupied buildings.

Although Tables IX and X strongly suggest that fires occur less often in unoccupied buildings than occupied ones, it is difficult to establish a causal relationship. Telephone demand (and hence electrical activity in telephone central offices) also reduces at night and on weekends; could this result in fewer telephone equipment fires and building power fires as well, regardless of the number of people present? Furthermore, it is impossible to estimate the average number of fires per million square feet of unoccupied space per year unless one knows the areas of Bell System buildings as a function of their occupancy: $f(x)$ square feet occupied x or more hours per week, for $0 \le x \le 168$. Unfortunately, these data are not available from the operating companies, and it would take a substantial effort to generate $f(x)$ for the 20,000 or more buildings in the Bell System. If $f(x)$ were known, it might be possible to predict the effect (with respect to fire) of such actions as dispersing Bell System switching equipment into a large number of small unmanned central offices close to the subscriber instead of a few larger central offices at a greater distance.

It would be desirable to extend this study to see if there is any difference in fire severity as a function of the distance to the nearest person in occupied buildings. Unfortunately, the fire report form does not give such information; however, the crude analysis given in Table XI may be suggestive. These figures should be interpreted with considerable caution, because the mixture of fires may not be the same; for example, fires detected by electrical means are likely to be expensive equipment fires, whereas fires detected by smoke or odor are likely to include a large number of inexpensive trash fires in addition to equipment fires. Thus, the fact that the median damage for fires detected by equipment or smoke alarms is greater than the median damage of fires detected by heat or odor should not be regarded as a demonstration of ineffectiveness of the former.

Table XI — Median fire damage in occupied buildings 1971–1975

	Number of Fires	Median Damage (Dollars)
People in other room, fire detected by equipment or smoke alarm	92	200
People in other room, fire detected by odor, noise, or light	205	30
People enter room in which fire is located (for other reasons)	154	20
People already in same room as fire	320	10

3.3 Methods used to fight building fires

During the 1971–76 period, Bell System regular or contract employees took action with respect to 1088 fires, either by fighting it themselves, calling the fire department, or both; this represents 86 percent of all building fires that occurred in that period. Fires in which employees were not involved are of two types: (*i*) those detected by outsiders who called the fire or police department or who (in one instance) extinguished the fire themselves, (*ii*) those detected by telephone people which were already out, or which self-extinguished before any action was taken (and the fire department was not notified).

Table XII shows that certain occupational groups—inside craft and, to a lesser extent, office worker and building mechanic—are the ones most likely to deal with Bell System building fires. Inside craft includes occupational titles such as switchmen, powermen, splicers, combinationmen, test deskmen, framemen, and central office maintenance; office workers include clerks, stockmen, cafeteria workers, service representatives, engineers, and other white-collar occupations, including management above supervision; building mechanics include titles such as building engineers, watch engineers, building maintenancemen, elevator mechanics, building electricians, and building technicians. Note that three occupations—construction, janitor, and guard—are likely to include substantial numbers of contract employees. The occupation was not specified in 12 percent of building fires.

Fires can be fought in many different ways, and these are summarized in Table XIII. Informal methods ordinarily involve blowing out or

Table XII — Distribution of Bell System occupations fighting fires

Occupation	Fraction of Fires
Auto mechanic	0.02
Janitor	0.06
Office worker	0.14
Inside craft	0.41
Operator	0.03
Guard	0.04
Western Electric	0.04
Construction	0.08
Install/repair	0.04
Building mechanic	0.14

Table XIII — Fraction of fires in which various fire-fighting methods were used

	Fire Department		
	Not Call	Call, not Need	Call and Need
Informal methods	0.16	0.05	0.01
Extinguisher or hose	0.42	0.14	0.07
Call fire department only	–	0.03	0.12

smothering the fire, throwing a glass (or a pail) of water on it, or turning off the electricity. The fire department was considered to be called and needed if they played a significant role in putting out the fire; if the fire was out (or almost out) when they arrived, and they assisted only in clean-up or smoke evacuation, they were considered to be called but not needed. A substantial fraction of fires are put out with the aid of an extinguisher and no notification of the fire department; in only 20 percent of all fires was the fire department really necessary. These fractions remain much the same for paper fires or electrical fires, but volatile fires are less likely to be fought using informal methods without notifying the fire department (0.04 instead of 0.16), and more likely to involve extinguishers combined with an unneeded fire department (0.21 instead of 0.14) or a call to the fire department with no attempt to fight the fire (0.24 instead of 0.15).

Bell fire-fighters using extinguishers or hoses almost always select the proper tools for the job; in only 13 cases was water apparently used (in whole or in part) on an electrical or a volatile fire. On the other hand, it is worth noting that women are infrequent users of extinguishers or hoses; out of 682 fires in which these tools were used, only 18 (about 2.6 percent) involved women.

Figs. 8, 9, and 10 (all plotted on triangular graph paper) give a more detailed look at the different fire-fighting exposures and techniques encountered by various occupations. Not surprisingly, auto mechanics and (to a lesser extent) installation and repair personnel (based at garages) encounter far higher percentages of volatile fires than the other groups; inside craft and Western Electric encounter more electrical fires than others, whereas janitors and (to a lesser extent) office workers, construction workers, and guards encounter paper fires. Operators and guards are far more likely to call the fire department than fight the fire, but inside craft, Western Electric and (to a lesser extent) construction workers, janitors, and building mechanics are unlikely to do so. Fires encountered by automobile mechanics, guards, or installation and repair personnel are the most likely to require professional assistance; fires associated with Western Electric or inside craft, the least.

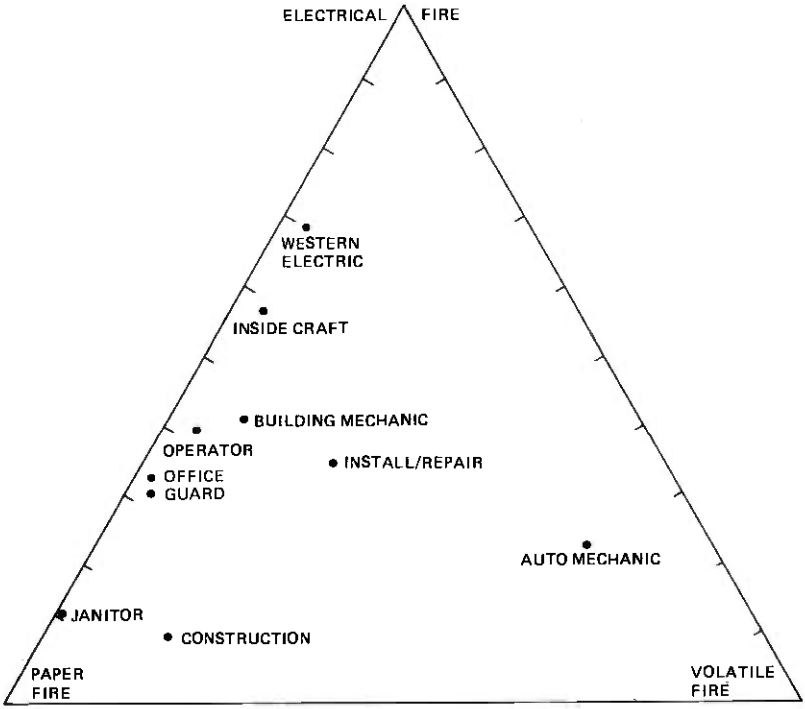


Fig. 8—Occupational exposure to fighting paper-electrical-volatile fires.

3.4 Correlated fire events

Many Bell System buildings have had more than one fire in the 1971–76 time period; in fact, one building has had 10. Statistical methods can be used to assess multiple fire events, to determine whether they are attributable to statistical fluctuations of fires occurring independently and at random, or to correlated events between fires. Correlated fires can arise for various reasons; the most common one is arson, but an undiagnosed electrical fault can lead to repeated occurrences of fire as well.

There are two distinct ways in which the possible correlation between fires at a given building can be examined. First, if x fires have been observed, one can ask if they cluster in a small period of time, rather than spreading out over the entire period. Second, one can ask whether x fires is excessive for that building, given Bell System fire experience. This is a somewhat more difficult assessment, for one must decide how to normalize the building with respect to the Bell System. Floor area is the most plausible candidate, but in view of the relationship between fires and people exhibited in Section 3.2, the number of people in the building may be a better normalization. In any event, one must be quite cautious in deciding whether or not a given Bell System building is more fire-prone

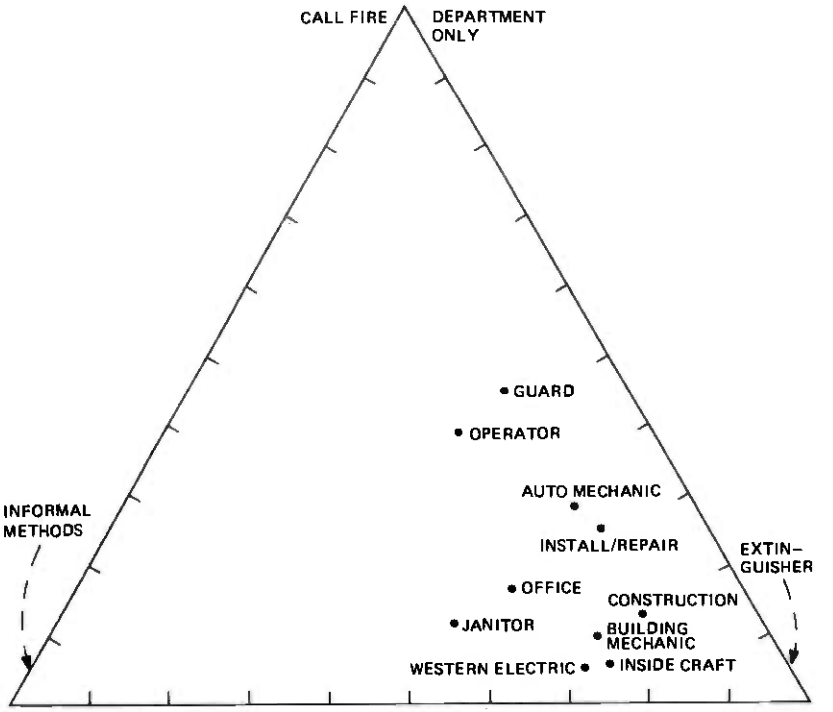


Fig. 9—Fire-fighting techniques used by various occupations.

than others, simply on the basis of a larger-than-average number of fires.

If x fires have occurred in a building, the cumulative probability density function of the smallest time-spacing between any two consecutive fires can be written[5]

$$\text{Prob(minimum spacing} \leq t) = 1 - (1 - (n-1)t)^n,$$

where n is the total number of fires and t is normalized with respect to the total time-interval (for example, if two fires occur 10 days apart in a 6-year period, t is equal to 10/2192, or 0.00456). If fires occur independently, and at random at a building throughout the time interval, this probability is distributed uniformly between zero and one; on the other hand, if there is correlation between fires (the occurrence of a fire raises the chance of another fire occurring in the near future), then there will be an excessive number of small values of the probability. Table XIV summarizes the probabilities associated with all multiple-fire buildings. The right-hand column clearly indicates that there are more buildings with small probabilities than with large ones; a chi-squared test of goodness-of-fit confirms that this result is not explainable by random fluctuation (at the 0.0000001 level).

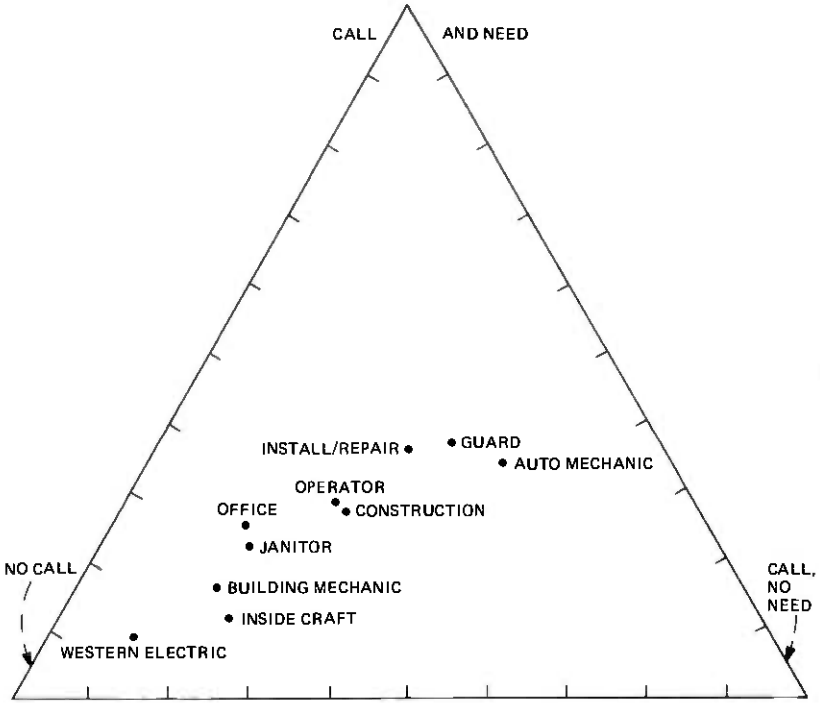


Fig. 10—Propensity of various occupations to call fire department.

Table XIV — Distribution of minimum time-interval probabilities for Bell System buildings with two or more fires, 1971–1976

Probability (minimum spacing ≤t)	Number of Buildings
0–0.1	44 (17)
0.1–0.2	26 (21)
0.2–0.3	17 (15)
0.3–0.4	17
0.4–0.5	10
0.5–0.6	15
0.6–0.7	8
0.7–0.8	8
0.8–0.9	10
0.9–1.0	9
	164 (130)

However, the inflation in probabilities does not extend beyond 0.3, for a similar chi-squared test on the last seven values yields a value of only 6.91, significant at the 0.3 level.

If one examines the fire reports for those 87 buildings for which the probabilities are less than 0.3, it is not hard to identify pairs of fires that appear to have some common factor. If these 34 buildings (20 percent of all multiple-fire buildings) are subtracted from the total, the parenthesized values in Table XIV result, and the corresponding chi-squared

test of goodness-of-fit is far more plausible under the hypothesis of randomness and independence: it is 14.46, significant at the 0.11 level. In other words, it is possible to detect most, if not all, of the correlation in multi-fire buildings by a reading of the fire reports; there does not appear to be much additional correlation present.

Most of the correlated events in the 34 buildings removed from the analysis are arsonous in nature; only six appeared to have other causes:

(*i*) A fire in the power panel of the turbine room of one building occurred within 10 minutes of a fire in the turbine sensing unit in a neighboring building.

(*ii*) Two fires 14 days apart occurred in the AC busway serving an ESS office.

(*iii*) An electrical fire in aisle 31 on the third floor of a mobile radio center was followed 98 days later by an electrical fire in aisle 32 on the same floor.

(*iv*) Two fires 26 days apart were caused by careless use of a cutting torch during modification of a building by a contractor.

(*v*) A fire in a coffee urn in a ladies' lounge was followed by a fire in a stove in the same lounge 10 days later.

(*vi*) A fire in a 48-volt generator in a power room was followed 21 days later by a fire in a 24-volt generator in the same room.

Unfortunately, this statistical technique cannot be used to identify arson in Bell System buildings if the arsonist acts only once; many such fires can be effectively made to look like accidents (for example, a cigarette carelessly thrown into a wastebasket). It is only when the arsonist strikes twice within a reasonably short period of time (say, six months) that his presence is almost always suspected.

In principle, a similar statistical analysis could be performed on the number of fires in each building in the Bell System normalized with respect to floor area (or other indicator of size or activity); however, floor area data on the 20,000 or more buildings in the Bell System is widely dispersed and not readily available for analysis. To give some flavor of the possible calculations, Table XV presents statistics on all buildings in the Bell System having eight fires or more during 1971–1976. The

Table XV — Buildings in the Bell System with eight or more fires, 1971–1976

	Floor Area (thous. sq. ft.)	Obs. Fires	Exp. Fires	Pr(fires ≥ observed)
Fresno, Cal.	179	10	0.76	0.00000001
Detroit, Mich.	769	9	3.25	0.0063
Manhattan, N.Y.	516	8	2.18	0.0019
Bronx, N.Y.	328	8	1.39	0.00010
Washington, D.C.	167	8	0.71	0.0000008
Bell System: 1194 fires, floor area 282.1 million sq. ft.				

expected number of fires in Table XV is calculated by multiplying the total fires in the Bell System by the ratio of the building floor area to the Bell System floor area; the probability of observing this number of fires or more, given the expected number, is calculated by means of the Poisson probability density function

$$\Pr(x \text{ or more fires}) = \sum_{i=x}^{\infty} \exp(-\lambda)\lambda^i/i!.$$

In four of these buildings, a reading of the fire reports clearly points to an arsonist at work. The fifth building, in Detroit, has no obvious pattern of fires, but it is likely that an event of probability 0.0063 could have occurred by chance, given the large number of Bell System buildings. (In a list of 20,000 buildings, there is a 50-50 chance that the probability in the final column of Table XV will be less than 0.000025 in at least one case, even assuming all the buildings have the same underlying fire propensity per square foot. To assess the correctness of this assumption, one would have to look at the probabilities associated with *all* of the buildings.) Furthermore, the minimum spacing between any pair of fires in the Detroit building (12 days) is not unusual; a random sample of 9 fires will produce a shorter minimum spacing 33 percent of the time.

The typical fire in the Bell System receives very little publicity; usually, only a few of the workers in the building know about it. (Since 58 percent of all fires do not involve the fire department, newspaper coverage is likely to be sparse.) It is of interest, therefore, to assess the impact of a Bell System fire which generated enormous publicity in a metropolitan area—the February 27, 1975 fire at 204 Second Avenue in Manhattan. Among other things, newspapers reported a rash of fires in other telephone buildings in the area during the month that repairs were being made; it was suggested that the fire publicity might have encouraged latent arsonists elsewhere in New York Telephone Company.

One can examine the fire data to see whether or not such an allegation is true, or whether the number of fires that occurred in the next month can be explained as a not-untypical fluctuation in the pattern of fires over the entire seven years. Table XVI gives the number of months (out of 84) in which 0, 1, 2, $\cdots$ fires were observed in Manhattan, in all five boroughs of New York, and in the Greater New York metropolitan area (specifically, all fires in the five boroughs, in the Nassau and Westchester operating areas of New York, and in the Essex, Raritan, and Hudson operating areas of New Jersey). The "Obs" column gives the actual number of months that the indicated number of fires were observed, and the "Poi" column gives the expected number of fires if a Poisson probability density function is fitted to the data

The one-parameter Poisson distribution does not fit the data particularly well; in fact, a chi-squared test of goodness-of-fit rejects the model

Table XVI—Distribution of number of fires by months, 1971–1977, in the Greater New York area

	Number of Months with Indicated Fires in								
	Manhattan			5 Boroughs			Greater N.Y.		
Fires	Obs	Poi	NB	Obs	Poi	NB	Obs	Poi	NB
0	44	37.8	41.0	26	19.7	22.8	15	9.2	12.9
1	20	30.2	26.5	23	28.5	26.8	17	20.3	20.6
2	13	12.0	11.1	13	20.7	18.3	23	22.5	19.4
3	7	3.2	3.8	17	10.0	9.5	8	16.6	13.9
4				3	3.6	4.1	12	9.2	8.5
5				2	1.1	1.6	4	4.1	4.6
6							4	1.5	2.3
7							1	0.5	1.1
m		0.80			1.45			2.21	
s^2		0.98			1.79			3.06	
r		3.44			6.19			5.78	
c		4.31			4.27			2.61	

at probability level of 0.03, 0.04, and 0.06, respectively. There is some evidence that fires tend to cluster in months more than a Poisson model would predict; note that the number of months with zero fires or with a large number of fires generally exceeds expectations. In such a situation, the two-parameter negative binomial distribution (also known as the Polya distribution, and often used in studies of accident-proneness) provides a better fit to the data. In the negative binomial, it is assumed that m, the mean of the Poisson distribution, is itself distributed according to the gamma distribution $c^r m^{r-1} \exp(-cm)/\Gamma(r)$. The probability of 0, 1, 2, 3, · · · observations in a cell is given by the successive terms of the series

$$\left(\frac{c}{c+1}\right)^r \left[1, \frac{r}{c+1}, \frac{r(r+1)}{2!(c+1)^2}, \frac{r(r+1)(r+2)}{3!(c+1)^3}, \cdots\right],$$

where c and r are estimated from the mean and variance of the data by the formulas

$$c = m/(s^2 - m), r = cm.$$

The negative binomial fit to the data is given in the "NB" column of Table XVI; the fit is considerably improved.

As far as fire reports are concerned, March 1975 (the month following the Second Avenue fire) witnessed three fires in Manhattan, one in Queens, and one in the suburbs. (These numbers do not tally exactly with newspaper-reported fires for several reasons: One Manhattan building fire inexplicably failed to generate a fire report, and a couple of fires occurred on customer premises or in outside plant, which are not covered by the fire report; on the other hand, one fire included here was not reported to the fire department and did not appear in the papers.) Using the Poisson model, the estimated probability of three or more fires in

Manhattan in one month is 0.048; of four or more in all five boroughs, 0.060; of five or more in the Greater New York area, 0.074. Using the negative binomial model, these probabilities increase to 0.064, 0.079, and 0.102, respectively. There is some evidence that March 1975 was an unusually busy month for telephone building fires in Manhattan; however, there is less evidence that it was an unusually busy month for fires in the five boroughs or the Greater New York area. In other words, the influence of the February 27 fire upon building fire statistics during March decreases as ever-larger geographical areas are considered—a hardly surprising result.

IV. CONCLUSIONS

The Bell System has a very good record with respect to building fires. About 200 fires per year were reported between 1971 and 1977 in Bell System operating company buildings, or on roofs or grounds; of these, inside fires occurred at a rate of approximately 0.6 per million square feet per year. (However, unreported fires may increase this figure by 30 percent or more.) All fires but the New York fire on February 27, 1975 appear to be well modeled by a log-normal probability density function of damage with a median value of $30 to $80 (or a mean value of $5,000 to $13,000); about one percent of all fires exceeds $100,000 in damage. The New York fire demonstrates that there is a small, but finite, chance of far more damaging fires; the best estimate of the probability of fires not following the log-normal damage distribution is 0.0002, based on 1960–1977 experience of 4496 fires. A tenth of all fires and half of all fire damage occurs in building unoccupied at the time of the fire. There is considerable evidence that fires occur in clusters; about 20 percent of all multiple-fire buildings had two or more fires that occurred near in time under similar circumstances. Furthermore, the enormous newspaper publicity of the New York fire may have been responsible for a modest but statistically significant increase in telephone building fires in the Greater New York area during the following month.

REFERENCES

1. C. A. Bennett and N. L. Franklin, *Statistical Analysis in Chemistry and the Chemical Industry*, New York: Wiley, 1954, p. 612.
2. L. M. Krasner and S. A. Wiener, "The Feasibility of Quantitatively Analyzing Investments in Loss Prevention Activities," Factory Mutual Research report FMRC 19257 (April 1973), pp. 14–15.
3. K. A. Brownlee, *Statistical Theory and Methodology in Science and Engineering*, New York: Wiley, 1965.
4. L. Derry, "A Study of United States Fire Experience, 1976," Fire Journal, *71*, No. 6 (November 1977).
5. D. F. Votaw, "The Probability Distribution of a Random Linear Set," Annals of Mathematical Statistics, *17* (1946), pp. 240–244.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 8, October 1978
Printed in U.S.A.

Boundary Integral Solutions of Laplace's Equation

By J. L. BLUE

(Manuscript received January 26, 1978)

Although Laplace's equation is simple, the region over which it is to be solved is often complicated. Both the shape of the region and the boundary conditions can induce solutions Φ which are singular at isolated points on the boundary of the region.

Boundary integral equation methods are well-suited to the problem, reducing a two-dimensional partial differential equation to a one-dimensional integral equation. Unfortunately, the standard boundary integral equation methods lead to an ill-conditioned set of linear equations, restricting the achievable accuracy in the approximate solution.

This paper describes an improved boundary integral method. A new integral equation is derived. Laplace's equation is reduced to solving two coupled, one-dimensional integral equations. The resulting linear equations are well-conditioned.

A program package for solving Laplace's equation has been developed. The package solves Laplace's equation in two dimensions or in three dimensions with axial symmetry. The region may extend to infinity, and may be multiply-connected. In addition to smooth basis functions, the program automatically includes appropriate singular basis functions, greatly improving the achievable accuracy for regions with corners.

I. INTRODUCTION

Laplace's equation frequently arises in modeling physical problems, especially in electromagnetism, in thermal flow, and in fluid flow. In two dimensions, Laplace's equation is

$$\frac{\partial^2 \Phi}{\partial x^2} + \frac{\partial^2 \Phi}{\partial y^2} = 0,$$

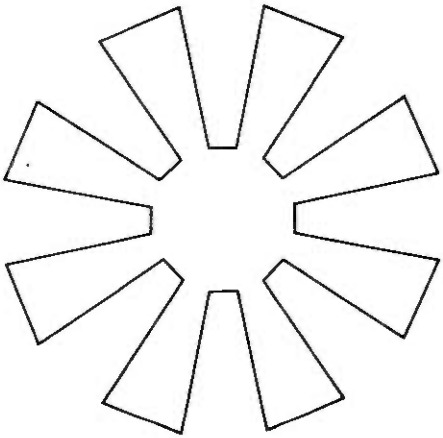


Fig. 1—Region used in an analysis of an electrostatic lens.

and in three dimensions,

$$\frac{\partial^2\Phi}{\partial x^2} + \frac{\partial^2\Phi}{\partial y^2} + \frac{\partial^2\Phi}{\partial z^2} = 0.$$

To complete the specification of a particular problem, a region on which to solve Laplace's equation must be specified, plus boundary conditions on the boundary of the region.

As compensation for the simplicity of the partial differential equation, the region over which Laplace's equation is to be solved is often complicated. Figure 1 shows the region used by the author in an unpublished analysis of an electrostatic lens. The solution is singular at the re-entrant corners. (By singular, we mean that Φ has a finite limit as the corner is approached, but that some derivatives of Φ do not have a finite limit.) The singularity is a consequence of the region itself, not of any particular boundary conditions. In fact, the solution is singular unless very special boundary conditions are prescribed.

Even with a rectangular region, the solution can be singular at isolated points. Figure 2 is an example, a thin-film capacitor with metal top and bottom contacts. To obtain its capacitance, Laplace's equation must be solved inside the rectangle. The boundary conditions are $\Phi = 1$ on the top contact, $\Phi = 0$ on the bottom contact, and zero normal derivative, $\partial\Phi/\partial n = 0$, on the remainder of the boundary. (The definition of the normal derivative is given in the next section.) At the center edge of the top contact, the solution is singular.

Standard methods for elliptic partial differential equations include finite difference and finite element methods. Both methods require a grid, usually rectangular or triangular, everywhere inside the region. Thus the region must be bounded. Both methods are difficult to apply

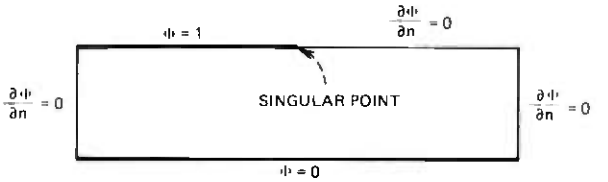


Fig. 2—Region used in an analysis of a thin-film capacitor. The potential is singular at the center of the top side.

to complicated regions; if the true solution has singularities, accuracy is usually poor unless heroic measures are taken. Neither method is suitable for a package for general regions and boundary conditions.

Laplace's equation is the simplest elliptic partial differential equation, and has been the subject of a great deal of analysis. Special methods for Laplace's equation are available, methods that do not work for general elliptic partial differential equations.

Special methods for Laplace's equation include the so-called "fast Poisson solvers."[1] They can quickly solve $\nabla^2\Phi(x,y) = f(x,y)$ if the region and boundary conditions are sufficiently simple. However, even Fig. 2 is not simple enough because of the mixed boundary conditions on the top boundary. The fast Poisson solvers have great utility for special problems, but are not appropriate for a general Laplace package. Recent research (Ref. 2, for example) indicates how these methods may be extended in the future.

1.1 The boundary integral equation method

The most useful special method for Laplace's equation is the boundary integral equation method. The basic method has been known for many years,[3,4] but has enjoyed a renewed popularity since the advent of large digital computers. A few representative references are Refs. 5 to 9. A two-dimensional partial differential equation is reduced to a one-dimensional integral equation. Similarly, a three-dimensional partial differential equation can be reduced to a two-dimensional integral equation. The integral equation involves only the geometry and the values of Φ and $\partial\Phi/\partial n$ on the boundary. Multiply-connected regions pose no added difficulty. After the integral equation has been solved approximately, another integral can be done to evaluate $\nabla\Phi$ and Φ at any point inside the region.

The boundary integral equation method has been quite successful, providing fast and inexpensive solutions for Laplace's equation in two dimensions. The usual implementation does have several difficulties. First, the integral equation is a Fredholm integral equation of the first kind for $\partial\Phi/\partial n$, and consequently is ill-conditioned. (For either a Dirichlet or a Neumann problem, a well-conditioned integral equation is

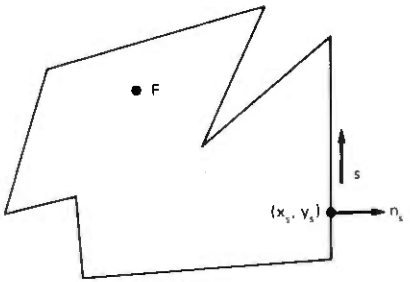


Fig. 3—Illustration of the definitions of s, x_s, y_s, $\mathbf{n}_s$, and F.

available, but not for mixed boundary conditions.) Matrices generated for approximate solutions to the integral equation are ill-conditioned, and $\partial\Phi/\partial n$ cannot be found accurately. Second, three-dimensional problems with axial symmetry are essentially two-dimensional problems, but no provision is made for their solution. (The next two objections do not apply to Ref. 9.) Third, the unknown Φ and $\partial\Phi/\partial n$ are approximated by low-order polynomials on sections of the boundary. Convergence requires many coefficients if accuracy of more than a few percent is required. Finally, no provision is provided for dealing with singularities.

The present paper describes an improved boundary integral method which counters all the above difficulties. Two coupled integral equations are used; the combination leads to a well-conditioned matrix. Both Φ and $\partial\Phi/\partial n$ can be obtained accurately. Higher-order approximations for the unknown Φ and $\partial\Phi/\partial n$ are used. Corner singularities are recognized automatically, and special approximating functions are used. Axisymmetric problems are solved by the same program. Typical problems cost only a few dollars to run. A reliable estimate of the accuracy of the approximate Φ is available.

1.2 The Laplace package

The method described in this paper has been implemented in the Laplace program package. A user's guide for the Laplace package, with several examples, is available separately.[10] The program package is written in EFL,[11] an extended Fortran language. The output of the EFL compiler is portable Fortran.

The package solves Laplace's equation in two dimensions or in three dimensions with axial symmetry. Two-dimensional regions must be bounded by straight-line segments. Three-dimensional regions must be figures of revolution whose cross section in the (r,z) plane is bounded by straight-line segments. The region may extend to infinity, but the boundary must not extend to infinity. The region may be multiply-

connected. On each line segment, either Φ or $\partial\Phi/\partial n$ may be specified as a boundary condition. In addition to smooth basis functions, the program automatically includes the appropriate singular basis functions greatly improving the achievable accuracy for regions with corners.

Section II discusses the mathematical basis for the boundary integral equation method. Section III describes the implementation of the method. Possible extensions to the program package are discussed in Section IV. Section V has results for a sample problem. The appendix derives the new integral equation used.

II. INTEGRAL EQUATION FORMULATION

In this section, Laplace's equation is formulated as a pair of coupled integral equations. The two-dimensional partial differential equation is reduced to a pair of one-dimensional integral equations.

We wish to solve

$$\nabla^2\Phi(x,y) = \frac{\partial^2\Phi}{\partial x^2} + \frac{\partial^2\Phi}{\partial y^2} = 0 \tag{P1a}$$

for (x,y) in a region D with boundary Γ. D may be multiply-connected, in which case Γ has several distinct parts. For now, we discuss only the two-dimensional "interior" problem, with D a finite region. At the conclusion of this section, we discuss the two-dimensional "exterior" problem, with D an infinite region, and the three-dimensional axisymmetric problem, both interior and exterior.

As in Fig. 3, let (x_s,y_s) be the coordinates of the point at arc length s, and denote $\phi(s) = \Phi(x_s,y_s)$. We will use Φ for the potential of a general point, and ϕ for a point on Γ. Let $\mathbf{n}_s$ be the outward-pointing unit normal vector at s. For a point s not at a vertex of Γ, define

$$\psi(s) \equiv \lim_{(x,y)\to(x_s,y_s)} \mathbf{n}_s \cdot \nabla\Phi(x,y).$$

The notation $\partial\phi/\partial n_s$ is also used for the right side of the above definition.

For the problem to be well-posed, a boundary condition must be given at each point of Γ.[12] The Laplace package allows the specification

$$\phi(s) = b_1(s) \text{ on part of } \Gamma, \text{ say } \Gamma_1 \tag{P1b}$$

$$\psi(s) = b_2(s) \text{ on the remainder of } \Gamma, \text{ say } \Gamma_2. \tag{P1c}$$

For any fixed point $F = (x_F,y_F)$, the Green's function, or fundamental solution to Laplace's equation, is

$$G(x,y;x_F,y_F) = -\tfrac{1}{2}\ln\left[(x-x_F)^2 + (y-y_F)^2\right].$$

Except at point F, $\nabla^2 G(x,y;x_F,y_F) = 0$.

Now let (x,y) be any point strictly inside D. Green's boundary identity is[3]

$$2\pi\Phi(x,y) = \int_\Gamma [\psi(s)G(x_s,y_s;x,y) - \phi(s)\mathbf{n}_s \cdot \nabla_s G(x_s,y_s;x,y)]ds.$$

The gradient operator, ∇_s, operates on the x_s and y_s. The above equation is usually abbreviated as

$$2\pi\Phi(x,y) = \int_\Gamma \left[\psi(s)G - \phi(s)\frac{\partial G}{\partial n_s}\right] ds, \tag{1}$$

with the arguments of G left implicit.

If (x,y) is a point at arc length t on a smooth part of Γ, it may be shown[3,13] that

$$\pi\phi(t) = \oint_\Gamma \left[\psi(s)G - \phi(s)\frac{\partial G}{\partial n_s}\right] ds. \tag{2}$$

The integral is now a Cauchy principal-value integral at $s = t$.

Suppose that the correct $\phi(s)$ and $\psi(s)$ are not known, but only approximate values $\phi^*(s)$ and $\psi^*(s)$ are known. Then the function $\Phi^*(x,y)$ defined by

$$2\pi\Phi^*(x,y) = \int_\Gamma \left[\psi^*(s)G - \phi^*(s)\frac{\partial G}{\partial n_s}\right] ds$$

exactly obeys Laplace's equation for (x,y) strictly inside D. Φ^* will not obey the correct boundary conditions as (x,y) approaches the boundary unless $\phi^*(s)$ and $\psi^*(s)$ are chosen correctly.

Thus the boundary integral equation method is one of the class of "particular solution" methods.[14,15] Any approximate solution obeys the partial differential equation exactly, but only obeys the boundary conditions approximately. The advantage over the usual particular solution methods for Laplace's equation, as seen in Ref. 16, for example, is that the boundary integral particular solutions incorporate the exact boundary of the region and do not require a restricted region. They are more complicated to calculate, but are appropriate for the region.

Equation (2) may be used to obtain an integral equation for ϕ^* and ψ^*.

$$\pi\phi^*(t) = \oint_\Gamma \left[\psi^*(s)G - \phi^*(s)\frac{\partial G}{\partial n_s}\right] ds. \tag{3a}$$

Letting $\mathbf{R}$ be the vector from point t to point s, and R the length of $\mathbf{R}$, (3a) may be written as

$$\pi\phi^*(t) = \oint_\Gamma \left[\frac{\mathbf{n}_s \cdot \mathbf{R}}{R^2}\phi^*(s) - \ln(R)\psi^*(s)\right] ds \tag{3b}$$

This boundary integral equation has been used for many years for solving Laplace's equation.[5–9]

For example, if $\phi(s) = b_1(s)$ is given on all of Γ (Dirichlet problem), set $\phi^*(s) = \phi(s)$, and the above is then an integral equation for the unknown $\psi^*(s)$. Thus a two-dimensional partial differential equation has been reduced to a one-dimensional integral equation. An approximate solution may be obtained by expanding $\psi^*(s)$ in an appropriate set of basis functions, and taking a finite number of these.

$$\psi^*(s) = \sum_{j=1}^{N} a_j f_j(s).$$

The f's are piecewise constant functions in Ref. 6 and piecewise quadratic in Ref. 7. We discuss an appropriate set of f's later. The integral equation (3a) then becomes

$$\sum_{j=1}^{N} a_j \oint_\Gamma f_j(s) G \, ds = \oint_\Gamma b_1(s) \frac{\partial G}{\partial n_s} ds + \pi b_1(t).$$

If $M = N$ points t_i are chosen at which to make this equation hold exactly (collocation), a set of N linear equations for the N unknowns, a_j, is obtained. If M points t_i, $M > N$, are chosen, an over-determined set of linear equations is obtained for the a_j's. This reduces the sensitivity of the approximate solution to the exact choice of the t_i. The equations are

$$\sum_{j=1}^{N} A_{ij} a_j = r_i, \; j = 1,2,\ldots,M,$$

where

$$A_{ij} = \oint_\Gamma f_j(s) G(s,t_i) ds$$

$$r_i = \oint_\Gamma b_1(s) \frac{\partial G(s,t_i)}{\partial n_s} ds + \pi b_1(t_i).$$

These may be solved in a least-squares sense, say, by a standard subroutine.[17]

In addition to the obvious advantages of this formulation, there is a well-known disadvantage. The integral equation for $\psi^*(s)$ is a Fredholm integral equation of the first kind,[18] of the type

$$u(x) = \int_0^1 H(x,y) v(y) dy,$$

where u and H are known and v is to be determined. This kind of integral equation is *ill-conditioned* (sometimes called *ill-posed*); it is difficult to obtain accurate solutions for v.[19,20] The reason for the difficulty is easy

to see. Since ν appears only inside the integral, its high-frequency components are not well-determined; by the Riemann-Lebesgue lemma,

$$\lim_{n\to\infty} \int_0^1 H(x,y) \sin (ny) dy = 0,$$

if H is not too badly behaved. The difficulty numerically is that the matrix $\{A_{jk}\}$ is ill-conditioned. Small errors in calculating elements A_{jk} or r_j lead to much-magnified errors in the coefficients a_k. If a sequence of approximate solutions with increasing N is done, the larger matrices are increasingly ill-conditioned. The typical failure mode is that $\psi^*(s)$ does not converge as N increases, after a certain point; rather, spurious and unphysical oscillations in $\psi^*(s)$ are seen.

Various methods of ameliorating the difficulty have been suggested, such as regularization[21] and matrix singular-value decomposition.[19,20,22] Better yet is to derive a Fredholm integral equation of the second kind.

In the appendix, we derive the following identity for t any point at a smooth part of Γ.

$$\pi\psi(t) = \fint_\Gamma \left\{ \psi(s) \frac{\partial G}{\partial n_t} - [\phi(s) - \phi(t)] \frac{\partial^2 G}{\partial n_s \partial n_t} \right\} ds. \tag{4a}$$

For the Dirichlet problem, this identity leads to a Fredholm integral equation of the second kind for $\psi^*(s)$ and is not ill-conditioned. Apparently, but surprisingly, (4a) is new. The integral equation derived from (4a) may be written as

$$\pi\psi^*(t) = \fint_\Gamma \left\{ [\phi^*(s) - \phi^*(t)] \times \frac{2(\mathbf{n}_s \cdot \mathbf{R})(\mathbf{n}_t \cdot \mathbf{R}) - R^2 \mathbf{n}_s \cdot \mathbf{n}_t}{R^4} + \frac{\mathbf{n}_t \cdot \mathbf{R}}{R^2} \psi^*(s) \right\} ds. \tag{4b}$$

With two integral equations, one well-conditioned for ϕ^* and the other for ψ^*, problem (P1) can be reduced to a set of linear equations with a well-conditioned matrix. If fitting point t_j is on Γ_1, where $\phi(s)$ is specified, use (4). If t_j is on Γ_2, where $\psi(s)$ is specified, use (3). A coupled pair of linear integral equations results. Analogously to the Dirichlet problem discussed earlier, appropriate basis functions and fitting points can be chosen, and the problem reduced to a set of linear equations. Some of the complications will be covered in later sections of the paper.

For the Dirichlet problem, ϕ given everywhere on Γ, (4) cannot be used everywhere. Since (4) is independent of the zero of potential, (4) alone will lead to a singular matrix, of rank $N - 1$. Special methods may be used for dealing with rank-deficient matrices, or the other equation, (3), may be used at some fitting points.

2.1 Exterior two-dimensional problems

We now consider solving Laplace's equation in an infinite region, D, exterior to a finite boundary, Γ. To have a unique solution, it is insufficient to specify either $\phi(s)$ or $\psi(s)$ at each point of Γ. In addition, the behavior of $\Phi(x,y)$ far from Γ must be specified. Use standard polar coordinates, (r,θ), with $r = \sqrt{x^2 + y^2}$ and suppose

$$\lim_{r\to\infty} \Phi(x,y) = \frac{\Psi_\infty}{2\pi} \ln \frac{1}{r} + \Phi_\infty + O(1/r),$$

where Φ_∞ and Ψ_∞ are constants. If Ψ_∞ is specified, then a unique solution can be found.[12] Ψ_∞ is the negative of total flux extending to infinity.

The earlier equations apply with small changes. For example, (3a) must be replaced by

$$\pi\phi^*(t) = 2\pi\Phi_\infty^* + \int_\Gamma \left[\phi^*(s) \frac{\partial G}{\partial n_s} - \psi^*(s) G \right] ds,$$

and the unknown Φ_∞^* must also be found.

The user specifies Ψ_∞ as well as boundary conditions. The Laplace package calculates an approximate value for Φ_∞^* as well as for ϕ^* and ψ^* on Γ.

2.2 Three-dimensional axisymmetric problems

Most of the preceding two-dimensional analysis needs only minor changes for the three-dimensional axisymmetric problem. Unlike the two-dimensional problem, the same formulation is adequate for interior and exterior three-dimensional problems. We use standard cylindrical coordinates (r,θ,z). We wish to solve

$$\nabla^2\Phi(r,\theta,z) = \frac{1}{r}\frac{\partial}{\partial r}\left(r \frac{\partial \Phi}{\partial r} \right) + \frac{1}{r^2}\frac{\partial^2 \Phi}{\partial \theta^2} + \frac{\partial^2 \Phi}{\partial z^2} = 0$$

in an axisymmetric region $\overline{D}$; $\overline{D}$ is formed as a figure of revolution by rotating a region D, with boundary Γ, about the z-axis. The boundary conditions must also be independent of θ.

For any fixed point $F = (x_F, y_F, z_F)$, the Green's function is

$$G(x,y,z;x_F,y_F,z_F) = \frac{1}{[(x-x_F)^2 + (y-y_F)^2 + (z-z_F)^2]^{1/2}} \equiv \frac{1}{R_3}$$

The integral equation corresponding to (3) is[13]

$$2\pi\phi^*(t) = \int_\Omega \left[\psi^*(s) G - \phi^*(s) \frac{\partial G}{\partial n_s} \right] d\Omega.$$

The integral is an area integral on the boundary, the surface of revolution.

Without loss of generality, let t be at $\theta = 0$. Let Γ be the intersection of Ω with any plane $\theta =$ constant. Express the area integral as an iterated integral over θ and s, arc length along Γ. Since ϕ^* and ψ^* are independent of θ, all the θ dependence is in G and $\partial G/\partial n_s$. R_3 may be expressed as

$$R_3^2 = R_m^2 - 2r_s r_t(1 + \cos\theta),$$

where

$$R_m^2 = (r_s + r_t)^2 + (z_s - z_t)^2.$$

With much manipulation, the θ integration may be performed, giving

$$2\pi\phi^*(t) = \oint_\Gamma \left\{ 4r_s K(m)\psi^*(s) + \left[\frac{4r_s \mathbf{n}_s \cdot \mathbf{R}_2}{R_m^2(1-m)} E(m) + 2\mathbf{n}_s \cdot \mathbf{e}_r [K(m) - E(m)] \right] \phi^*(s) \right\} \frac{ds}{R_m}, \quad (5)$$

where $\mathbf{e}_r$ and $\mathbf{e}_z$ are unit vectors in the $\theta = 0$ plane, and $\mathbf{R}_2$ is the vector in the $\theta = 0$ plane from t to s.

$$\mathbf{R}_2 = (r_s - r_1)\mathbf{e}_r + (z_s - z_t)\mathbf{e}_z$$

$$m = \frac{4r_s r_t}{R_m^2}.$$

K and E are the usual complete elliptic integrals.[23]

$$E(m) = \int_0^{\pi/2} (1 - m\sin^2 u)^{1/2} du$$

$$K(m) = \int_0^{\pi/2} (1 - m\sin^2 u)^{-1/2} du.$$

The integral equation corresponding to (4) is considerably more complicated.

$$2\pi\psi^*(t) = \oint_\Gamma \frac{4r_s}{R_m^3} \left\{ \left[\mathbf{n}_t \cdot \mathbf{R}_2 \frac{E}{1-m} - 2r_s \mathbf{n}_t \cdot \mathbf{e}_r \frac{K-E}{m} \right] \psi^*(s) + \left[(\mathbf{n}_s \cdot \mathbf{e}_r)(\mathbf{n}_t \cdot \mathbf{e}_r)(2E - K) - \mathbf{n}_s \cdot \mathbf{n}_t \frac{E}{1-m} + \frac{(\mathbf{n}_s \cdot \mathbf{R}_2)(\mathbf{n}_t \cdot \mathbf{R}_2)}{R_m^2(1-m)} \left(\frac{2(2-m)}{(1-m)} E - K \right) + \frac{2}{R_m^2} [(\mathbf{n}_s \cdot \mathbf{e}_r)(\mathbf{n}_t \cdot \mathbf{R}_2) r_t - (\mathbf{n}_t \cdot \mathbf{e}_r)(\mathbf{n}_s \cdot \mathbf{R}_2) r_s] \times \left(\frac{E}{1-m} + \frac{K-E}{m} \right) \right] [\phi^*(s) - \phi^*(t)] \right\} ds. \quad (6)$$

K and E have been used as abbreviations for $K(m)$ and $E(m)$.

2.3 Error estimates

If (x,y) is strictly inside D, the approximate potential (in two dimensions) obeys

$$2\pi\Phi^*(x,y) = \int_\Gamma \left[\psi^*(s)G - \phi^*(s) \frac{\partial G}{\partial n_s} \right] ds.$$

If (x,y) is on a smooth part of Γ,

$$\pi\Phi^*(x,y) = \fint_\Gamma \left[\psi^*(s)G - \phi^*(s) \frac{\partial G}{\partial n_s} \right] ds.$$

If (x,y) is at a vertex of Γ, the π is replaced by the interior angle of the vertex. Similarly, $\nabla\Phi^*$ inside D and $\partial\Phi^*/\partial n$ on Γ may be obtained by the analog of (4).

The function Φ^* as defined above exactly obeys Laplace's equation inside D. Therefore, by the maximum principle,[24] the maximum error in Φ^* occurs somewhere on Γ.

$$\left| \Phi^*(x,y) - \Phi(x,y) \right| \leq \max_s \left| \Phi^*(x_s,y_s) - \phi(s) \right|.$$

For the Dirichlet problem, a rigorous error bound is in principle possible by finding the largest discrepancy between Φ^* and the boundary condition ϕ. However, finding the error bound can be more expensive than solving the integral equation.

For mixed boundary conditions, a rigorous bound is in general impossible. The above bound is still correct, but is not useful, since the true Φ is not known on all of the boundary. For certain restricted regions, another rigorous error bound can be obtained.[25] For these restricted regions,

$$\left| \Phi^*(x,y) - \Phi(x,y) \right| \leq \max_{\Gamma_1} \left| \Phi^*(x_s,y_s) - \phi(s) \right| + R_D \max_{\Gamma_2} \left| \frac{\partial \Phi^*(x_s,y_s)}{\partial n_s} - \psi(s) \right|,$$

where R_D is the maximum perpendicular distance from any point of Γ_2 to any other point of Γ. This bound is in principle possible to compute, but is expensive in practice.

An error *estimate* is available at no extra cost in the Laplace package, because of the method of solution. The over-determined system of linear equations is solved in a least-squares sense, minimizing the total fitting error (TFE),

$$\left[\sum_{i=1}^{M} \sum_{j=1}^{N} (A_{ij}a_j - r_j)^2 \right]^{1/2},$$

and returning this error. For the implementation discussed in the next section, with $M \approx 3N/2$, numerical experiments indicate that the TFE is a reliable upper bound on the error in the potential on the boundary, and a substantial overestimate of the error away from the boundary.

III. IMPLEMENTATION

Section II was general and discussed mathematics; we now become more specific and discuss numerical analysis. We also discuss some of the myriad details necessary to make a computer program feasible.

3.1 Geometry

Section II considered regions of arbitrary shape. The current implementation of the Laplace package requires that Γ be composed of finite straight-line segments. This is in contrast to the usual practice in analysis, of requiring that Γ be a smooth curve everywhere. Many practical problems have corners in their geometries, so it is essential to be able to handle such boundaries. It is easier to analyze exact corners than "smooth" geometries with a very small radius of curvature rather than a corner. In the following, each of the straight-line segments is called a *side*.

3.2 Basis functions

In the previous section, the choice of the basis functions $f_k(s)$ was left arbitrary. However, the particular choice made strongly affects the accuracy and efficiency of the program. At least four factors should be considered.

(*i*) The basis functions should be able to model the behavior of $\phi(s)$ and $\psi(s)$ with only a few functions.

(*ii*) If enough basis functions are used, they should be able to approximate $\phi(s)$ and $\psi(s)$ arbitrarily well.

(*iii*) The basis functions should be a well-conditioned set, so that small errors in doing the integrals do not lead to large errors in the approximate solution.

(*iv*) The integrals of the basis functions times G, $\partial G/\partial n_s$, and $\partial^2 G/\partial n_s \partial n_t$ must be tractable, either analytically or numerically.

Historically, (*iv*) has been dominant. Symm[6] approximated curved boundaries by straight-line segments and used piecewise constant basis functions. Hayes[7] allowed boundaries to be straight-line segments or arcs of circles, and used piecewise quadratic basis functions. Blue[9] allowed straight-line boundaries and allowed piecewise polynomial basis functions. For all these choices, the integrals in (3) and (4) can be done

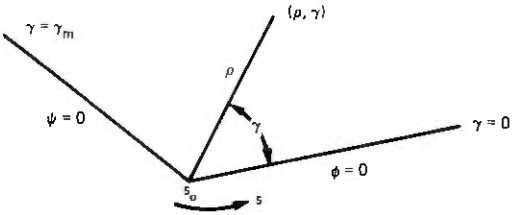


Fig. 4—Local polar coordinates used for eq. (7).

analytically. The above authors did not treat the axisymmetric problem, but the integrals in (5) and (6) are intractable analytically.

If the boundary has corners, $\psi(s)$ can be infinite at the corners, and piecewise polynomial basis functions will not be able to approximate $\psi(s)$ well. For example, in Fig. 4, suppose the boundary conditions are as shown, and the interior angle is γ_m. Choose polar coordinates (ρ,γ) centered at the vertex, with angle $\gamma = 0$ on the $\phi = 0$ side; let $s = s_0$ at the corner. Then for small ρ, Φ has an expansion (in two dimensions)[26]

$$\Phi(\rho,\gamma) = \sum_{n=1}^{\infty} C_n \rho^{\alpha_n} \sin(\alpha_n \gamma), \tag{7}$$

where

$$\alpha_n = \frac{(n - 1/2)\pi}{\gamma_m}.$$

Thus Φ is identically zero on the line $\gamma = 0$. On the line $\gamma = \gamma_m$, the normal derivative is identically zero.

On the line $\gamma = 0$, where $s_0 \geq s$, the normal derivative is

$$\psi(s) = \sum_{n=1}^{\infty} \alpha_n C_n (s - s_0)^{\alpha_n - 1}.$$

For the case $\gamma_m > \pi/2$, we have $\alpha_1 < 1$, and we expect $\psi(s)$ to be infinite at the corner, unless C_1 happens to be zero. For $\psi(s)$ to be approximated accurately with only a few basis functions, one of them should be $\alpha_1(s - s_0)^{\alpha_1 - 1}$ with the correct α_1.

Similarly, on the line $\gamma = \gamma_m$, where $s \leq s_0$, the potential is

$$\phi(s) = \sum_{n=1}^{\infty} C_n (s_0 - s)^{\alpha_n} (-1)^{n+1}.$$

Therefore a basis function $(s_0 - s)^{\alpha_1}$ is needed on the $\gamma = \gamma_m$ side. In fact, only a single unknown coefficient, C_1, need be introduced to deal with the worst part of $\phi(s)$ and $\psi(s)$ at s_0. It may also be desirable to include a few of the less singular basis functions. The Laplace package includes singular basis functions for which $\alpha < \alpha_{max}$, with a default value α_{max}

= 1, and does not include any singular function whose α is within 0.1 of an integer.

Similar singular basis functions are used at corners where ϕ is specified on both sides and where ψ is specified on both sides. Since the expansion (7) is not necessarily convergent far from the vertex, the singular basis functions centered at a corner are used only on the two sides meeting at that corner.

For axisymmetric problems, the expansion (7) does not hold, but the exponent of the singularity is the same for off-axis points, since the singularity depends only on the highest order derivatives in the differential equation.[26] For singularities on the axis, the exponents are different,[27] and singular functions have not yet been implemented.

Additional "smooth" basis functions are also needed. A well-conditioned family of basis functions is B-splines. A brief description of some of their properties follows.[28,29] B-splines are defined on a line divided into intervals by *knots*. B-splines of order k are piecewise polynomials of degree $k - 1$. Each B-spline is nonnegative, has exactly one maximum, and has local support. The sum of B-splines at any point is identically one. In any interval, exactly k B-splines are nonzero; each B-spline is nonzero in at most k intervals. The B-splines on a line are uniquely determined by the knots, which may be multiple. At a knot with multiplicity m, a kth-order B-spline has $k - m - 1$ continuous derivatives. If $m = k - 1$, the B-spline is only continuous; if $m = 1$, the B-spline has $k - 2$ continuous derivatives.

The user chooses a k, the same for all sides, and the number of interior knots on each side. Mesh spacing proceeds according to the following rules. A vertex is called *singular* if its expansion, as in (7), has $\alpha_1 < 0.9$. If a singular basis function is used, the vertex is called *compensated*. If the two vertices delimiting a side are each either compensated or nonsingular, the interior knots on the side are spaced uniformly. Otherwise, the interior knots are spaced closer together near uncompensated singular vertices.

For a given mesh, higher-order B-splines are potentially more accurate, since the approximation error can be $O(h^k)$ [30], where h is the maximum mesh. However, the integrals for higher order splines are more difficult, and there are more unknown spline coefficients for higher k. Currently, the Laplace package restricts k to be 2, 3, or 4.

3.3 Boundary conditions

If $\phi(s)$ or $\psi(s)$ is specified as an arbitrary function, the integrals involving the boundary conditions require special methods. Instead, the Laplace package does a least-squares fit of the boundary condition to a B-spline. A separate fit is done on each side; the same order and mesh

are used as specified by the user for that side for the unknown $\psi^*(s)$ or $\phi^*(s)$.

For the Neumann problem—ψ specified on all of Γ—the problem is undetermined up to an additive constant in Φ. A solution exists for the interior problem only if $\int_\Gamma \psi(s)\,ds$ is exactly zero. The Laplace package currently will not solve the Neumann problem; ϕ must be specified on at least one side.

3.4 Integrals

With polynomial basis functions and boundaries composed of straight-line segments, the integrals in (3a) and (4a) can be done analytically. This can cause conditioning problems; B-splines are a well-conditioned basis only if calculated properly.[29] The integrals in (5) and (6) cannot be done analytically, even with polynomial basis functions. With singular basis functions like $(s - s_0)^\alpha$, none of the integrals can be done analytically unless α is special.

All the necessary integrals can be done accurately and efficiently by the numerical methods described in this section. We first consider (3a) and (4a), for any fixed t. For straight-line boundaries, the integral over Γ is divided up into a sum of integrals over the line segments. We consider only a single segment, and eliminate any subscript referring to the segment. On the segment, $\mathbf{n}_s$ is constant, and $\mathbf{R}$ may be written as

$$\mathbf{R} = R_\perp \mathbf{n}_s + (s - s_\perp)\mathbf{e}_s,$$

where $\mathbf{e}_s$ is a unit vector along the side. Also expand $\mathbf{n}_t$ as

$$\mathbf{n}_t = n_\perp \mathbf{n}_s + n_\parallel \mathbf{e}_s.$$

The portions of (3a) and (4a) from the segments are

$$\pi\phi^*(t) = \int_{s_1}^{s_2} \left\{ \frac{R_\perp}{R_\perp^2 + (s - s_\perp)^2}\phi^*(s) - \tfrac{1}{2}\ln\left[R_\perp^2 + (s - s_\perp)^2\right]\psi^*(s) \right\} ds \quad (3c)$$

$$\pi\psi^*(t) = \int_{s_1}^{s_2} \left\{ [\phi^*(s) - \phi^*(t)] \times \frac{2R_\perp[n_\perp R_\perp + n_\parallel(s - s_\perp)] - [R_\perp^2 + (s - s_\perp)^2]n_\perp}{[R_\perp^2 + (s - s_\perp)^2]^2} + \frac{n_\perp R_\perp + n_\parallel(s - s_\perp)}{R_\perp^2 + (s - s_\perp)^2}\psi^*(s) \right\} ds. \quad (4c)$$

We first consider the case where the point t is not on the segment in question. Then the Green's function parts of the integrals are not sin-

gular; either $R_\perp \neq 0$ or $s_\perp$ is not in the interval $[s_1,s_2]$. As an example, look at the integral with the logarithm in it, and look at one term of the expansion of $\psi^*(s)$, with the basis function $B_i(s)$. Further divide the segment into subintervals, between knots of the spline, so that over each subinterval $B_i(s)$ is a polynomial. Over each subinterval, the integrand is a polynomial times a nonsingular function. Gauss-Legendre quadrature is ideal for such integrands, if the order of the quadrature rule can be determined *a priori*. (An automatic quadrature method could be used, such as Refs. 31 or 32, but these are usually less efficient.) Computing the order of the quadrature rule necessary can be done using the results of numerical experimentation. If s_{j_1} and s_{j_2} are the ends of the subinterval, let $s_c = (s_{j_1} + s_{j_2})/2$ and $h = s_{j_2} - s_{j_1}$. The change of variable $u = 2(s - s_c)/h$ changes the term in question to

$$-\frac{h}{4}\int_{-1}^{1} [\ln (h^2/4) + \ln [a^2 + (u - b)^2]]B_i(s_c + hu/2)du,$$

where $a = 2R_\perp/h$ and $b = 2(s_\perp - s_c)/h$. Since the integral is over a single mesh interval of B_i, we expect the error to be no worse than the worst error in any of the kth-order B-splines with k-fold knots at -1 and 1, and no interior knots, since the latter B-splines vary more rapidly over the interval. Thus we look only at the errors in these k B-splines; call them $\overline{B}(u)$ to distinguish them.

Now consider the family of integrals

$$I_j(a,b,k) = \int_{-1}^{1} \ln [a^2 + (u - b)^2]\overline{B}_j(u)du.$$

Let $E_j(a,b,k,n)$ be the error in evaluating $I_j(a,b,k)$ by an n point Gauss-Legendre quadrature rule, and let

$$E(a,b,k,n) = \left[\sum_{j=1}^{k} E_j(a,b,k,n)^2\right]^{1/2}.$$

Numerical experiments show that in the (a,b) plane, the locus of constant $E(a,b,k,n)$ is approximately an ellipse. For given n and desired accuracy, ϵ, there is an ellipse with semi-axes $A(k,n,\epsilon)$ and $B(k,n,\epsilon)$ so that the error is satisfactory if a and b are outside the ellipse, or

$$\left[\frac{a}{A(k,n,\epsilon)}\right]^2 + \left[\frac{b}{B(k,n,\epsilon)}\right]^2 \geq 1.$$

For doing the integrals $I_j(a,b,k)$ to accuracy ϵ, the functions $A(k,n,\epsilon)$ and $B(k,n,\epsilon)$ are determined experimentally for a series of values of n. (The default values are $n = 4, 6, 8, 10, 12$, and 16, and $\epsilon = 10^{-6}$.) For any particular a and b, the smallest satisfactory n is used. If the largest n available is insufficient, then the interval is divided; this is seldom

necessary. In the Laplace package, $A(4,n,\epsilon)$ and $B(4,n,\epsilon)$ are used for $k \leq 4$.

The other integrals in (3a) and (4a) are done similarly, using numerically derived ellipses for B-spline basis functions. For singular basis functions, Gauss-Jacobi quadrature formulas are used; these are Gauss quadrature formulas on (0,1) with weight function $x^{\alpha-1}$. Different quadrature formulas are used for each of the unique α's used in singular basis functions. The same ellipses as calculated for B-splines are used; slightly smaller ellipses could be used, but the gain in efficiency is small.

The Gauss quadrature formulas are calculated portably by the method of Sack and Donovan,[33] using programs in the PORT library.[34]

The integrals of (5) and (6) are somewhat more complicated than those of (3a) and (4a), but are no harder numerically. The same ellipses are used.

If point t is on the line segment, then the Green's functions in the integrals have singularities. The integrals (3c) and (4c) simplify somewhat, since $\mathbf{n}_s = \mathbf{n}_t$, $\mathbf{n}_\parallel = 0$, $\mathbf{n}_\perp = 1$, and $R_\perp = 0$. The ϕ^* term in (3c) is identically zero; the ψ^* term is

$$-\tfrac{1}{2} \int_{s_1}^{s_2} \ln [(s-t)^2] \psi^*(s)\, ds.$$

The ψ^* term in (4c) is identically zero; the ϕ^* term is

$$-\int\!\!\!\!\!\!-_{s_1}^{s_2} [\phi^*(s) - \phi^*(t)] \frac{ds}{(s-t)^2}.$$

Special care must be taken to get accurate approximations to these singular integrals.

First consider a B-spline basis function, $B_i(x)$, again dividing (s_1,s_2) into subintervals. If t is not in the subinterval in question, then the previous methods are adequate. (The Laplace package never takes t to be exactly at a knot.) For the logarithmic integral, the subinterval including t is, for some positive δ_1 and δ_2,

$$\begin{aligned}
&-\tfrac{1}{2} \int_{t-\delta_1}^{t+\delta_2} \ln [(s-t)^2] B_i(s)\, ds \\
&\quad = -\int_0^{\delta_1} \ln (u) B_i(t-u) du - \int_0^{\delta_2} \ln (u) B_i(t+u) du \\
&\quad = -\delta_1 \int_0^1 \ln (\nu) B_i(t - \nu\delta_1) d\nu - \ln (\delta_1) \int_0^{\delta_1} B_i(t-u) du \\
&\qquad - \delta_2 \int_0^1 \ln (\nu) B_i(t + \nu\delta_2) d\nu - \ln (\delta_2) \int_0^{\delta_2} B_i(t+u) du.
\end{aligned}$$

The integrals with ln (ν) are done by Gauss quadrature with weight function ln (ν); the others are done by Gauss-Legendre quadrature. The Gauss quadrature formulas with logarithmic weight function are also calculated portably by the method of Ref. 33.

The Cauchy principal-value integral, for the subinterval including t, is

$$-\int_{t-\delta_1}^{t+\delta_1} \frac{B_i(s) - B_i(t)}{(s-t)^2}\,ds$$

$$= -B'_i(t)\int_{t-\delta_1}^{t+\delta_1} \frac{ds}{s-t} - \int_{t-\delta_1}^{t+\delta_2} \frac{B_i(s) - B_i(t) - (s-t)B'_i(t)}{(s-t)^2}\,ds.$$

The first integral is done analytically. The second has no singularity at $s = t$ and is done analytically as

$$-\int_{t-\delta_1}^{t+\delta_2} \left[\frac{1}{2}B_i''(t) + \frac{1}{6}(s-t)B_i'''(t) + \ldots\right]ds.$$

This is adequate for low-order B-splines. For high-order B-splines, more care would be necessary.

Now consider integrals with singular basis functions and with t on the same segment as s. If $\psi(s)$ is given on the segment, $\phi^*(s)$ may have $(s - s_1)^\alpha$ or $(s_2 - s)^\alpha$ terms; however, if $\psi(s)$ is given, (3) is always used for t on the side, and the ϕ^* terms vanish because $R_\perp = 0$. If $\phi(s)$ is given on the segment, $\psi^*(s)$ may have $(s - s_1)^{\alpha-1}$ or $(s_2 - s)^{\alpha-1}$ terms; however, if $\phi(s)$ is given, (4) is almost always used, and the ψ^* terms vanish because $R_\perp = 0$. The exception, when $\phi(s)$ is given on a segment and (3) is used, occurs only for the Dirichlet problem, ϕ given on all of Γ. Then (3) is used at the central fitting point of each side, and we need integrals of the form

$$\pm\tfrac{1}{2}\alpha\int_{s_1}^{s_2} \ln\,[(s-t)^2](s-s_1)^{\alpha-1}ds.$$

As much as possible of the integral

$$\int_{s_1}^{t} \ln\,(t-s)(s-s_1)^{\alpha-1}ds,$$

starting from s_1, is done using Gauss-Jacobi quadratures. The remainder has only a logarithmic singularity. It and the integral form t to s_2 are done by Gauss quadrature with a logarithmic weight function, as described earlier in this section.

3.5 Complete elliptic integrals

The complete elliptic integrals $K(m)$ and $E(m)$ are necessary for the axisymmetric problem. Suitable expansions are[35]

$$K(m) = P_K(1 - m) - Q_k(1 - m) \ln (1 - m)$$

$$E(m) = P_E(1 - m) - Q_E(1 - m) \ln (1 - m).$$

Polynomial approximations for the Ps and Qs are given in Ref. 35. The argument $(1 - m)$ is used instead of m to avoid excessive error as $m \to 1$, i.e., as $s \to t$ in (5) and (6).

The combination $D(m) = [K(m) - E(m)]/m$ is also needed. As $m \to 0$, $D(m) \to \pi/4$. For $D(m)$, another approximation of the above type was generated.

3.6 Fitting points

At least N fitting points are needed to determine the N unknown coefficients of the basis functions. The work to calculate the matrix is proportional to the number of points used. If more than N points are used, the sensitivity of the solution to the placement of the points is diminished, as is the amplification of any small errors in calculating matrix elements. In the Laplace package, approximately f times N fitting points are used; the default value of f is 1.5. In the subinterval between each pair of knots, the number of fitting points is f times the number of unknowns associated with the subinterval, rounded up. The fitting points are uniformly spaced within each subinterval.

3.7 Scaling, constraints, and matrix solution

Each row of the matrix corresponds to applying either (3) or (4) at one fitting point, t_i. To keep the solution approximately independent of the scaling of the region, each row corresponding to (4) is multiplied by the length of the side containing t_i.

For an interior problem, $\int \psi^*(s)\, ds = 0$. For an exterior two-dimensional problem, $\int \psi^*(s)\, ds = \Psi_\infty$. Either restriction may be written as a linear equality constraint on the unknown coefficients. When the matrix equations are solved by QR factorization, such linear constraints can easily be enforced using a method described by Lawson and Hanson.[36]

3.8 Portability

A portable stack allocation mechanism[34] is used for all temporary storage. The program is written in EFL.[11] The output of the EFL compiler is portable Fortran.

Two parts of the program are not portable. For $\epsilon \neq 10^{-6}$, new ellipses are necessary. The approximations to the complete elliptic integrals are accurate to about 10^{-8}.

IV. POSSIBLE EXTENSIONS

In this section, we discuss several extensions to the Laplace package which could be implemented if there were sufficient incentive, and the difficulties involved with each. Combinations of the individual extensions pose further difficulties, but will not be discussed.

4.1 Higher-order B-splines and higher-accuracy integrals

B-splines of order higher than 4 are useful if very high accuracy solutions are desired. The only change necessary to allow higher-order B-splines or higher-accuracy integrals is to change the ellipse semi-axes. This feature was not included in the Laplace package, since calculating the ellipses portably for any specified accuracy and B-spline order requires too much code. Alternate methods for doing integrals to any specified accuracy are under consideration.

4.2 Singular basis functions for axisymmetric problems

At a vertex away from the axis of revolution, an expansion similar to (7) will hold; the exponents $\{\alpha_n\}$ are the same as for the two-dimensional problem with the same shape as the cross section of the figure of revolution. At a vertex on the axis of revolution, the exponents are different; on-axis singular functions have not been implemented.

4.3 General linear boundary conditions

In some applications, it is desirable to solve Laplace's equation with the general linear boundary conditions on Γ

$$a(s)\phi(s) + b(s)\psi(s) = c(s),$$

with a and b simultaneously nonzero. Then (3a), say, would become

$$\pi\phi^*(t) = \oint_\Gamma \left\{ \left[\frac{\mathbf{n}_s \cdot \mathbf{R}}{R^2} - \frac{a(s)}{b(s)} \ln\left(\frac{1}{R}\right) \right] \phi^*(s) + \ln\left(\frac{1}{R}\right) \frac{c(s)}{b(s)} \right\} ds.$$

The difficulty here is in choosing a method for accurately evaluating the integrals involving a/b and c/b, unless a/b and c/b are restricted drastically, say, to being constants on each of the boundary line segments.

4.4 Curved boundaries

Many applications have part or all of the boundary as a smooth curve, which the user might not wish to approximate by straight-line segments. In principle, all that is needed to allow Γ to be any smooth curve is a parameterization of x_s, y_s, and $\mathbf{n}_s$ as a function of s. Again, the difficulty is in doing the integrals accurately and efficiently. The ellipse method would not be directly applicable. In addition, some of the integrals which

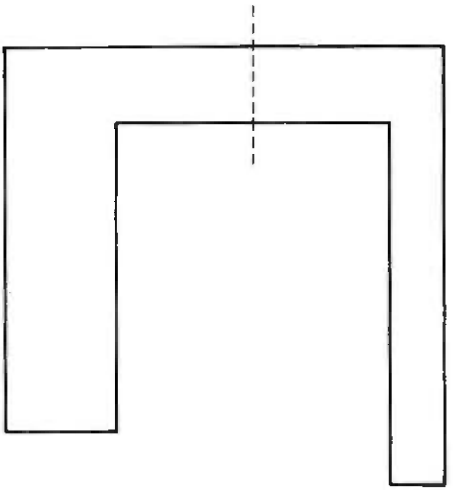


Fig. 5—A region which could be handled more easily by breaking it into two regions with an interface.

vanished identically for straight-line boundaries would not vanish. For example, the ϕ^* term of (3b) for s and t on the same segment vanishes if the segment is a straight line, because $\mathbf{n}_s \cdot \mathbf{R}$ is identically zero. If the line segment is curved, $\mathbf{n}_s \cdot \mathbf{R}/R^2$ in general has a finite limit as $s \rightarrow t$, and this term needs to be kept.

4.5 Interfaces

In other applications, there may be interfaces. A common problem is $\nabla^2 \Phi_1 = 0$ in region D_1 with boundary Γ_1, $\nabla^2 \Phi_2 = 0$ in region D_2 with boundary Γ_2, and interface conditions on the common portions of Γ_1 and Γ_2. Typical interface conditions are $\phi_1 = \phi_2$ and $\kappa_1 \psi_1 = \kappa_2 \psi_2$, where κ_1 and κ_2 are given constants.

Implementing this extension would require a significant change in data structure, but otherwise would be easy. No new types of integrals would arise. The singular basis functions at corners which are also points on the common boundary depend on κ_1 and κ_2 as well as the angles.[37]

This extension would also be useful for some single-region problems. A typical example is Laplace's equation inside a U-shaped region, Fig. 5. This could be broken artificially into two regions as shown, with interface conditions $\phi_1 = \phi_2$ and $\psi_1 = \psi_2$. The full region requires approximately $3\,N^2$ integrals, if N is the number of unknowns. The two half-size regions would each have $N/2$ unknowns, plus a few extra for the boundary values on the dotted line. Each region would require somewhat more than $3(N/2)^2$ integrals, so that the total work would be somewhat more than half.

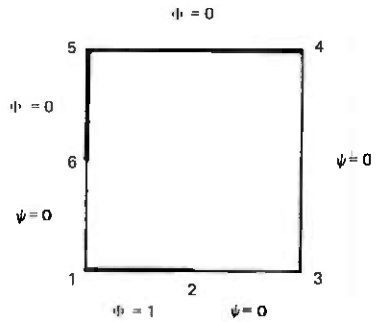


Fig. 6—Geometry for a sample problem.

V. AN EXAMPLE

Figure 6 gives the geometry and the boundary conditions for a sample problem. The boundary conditions are $\psi = 0$ on the light sides, $\phi = 1$ on the bottom dark side, and $\phi = 0$ on the top (L-shaped) dark side. This problem was solved approximately with third-order and fourth-order B-splines, and with various numbers of interior knots. Each side had the same number of interior knots. Some of the information is summarized in Table I. N is the number of basis functions. The running time is given in seconds for a Honeywell 6070 computer. TFE is the total fitting error, as defined in Section II. $\int \psi ds$ is over the side from 1 to 2. The next column gives ϕ at vertex 3. The approximate Φ at $(1/2,1/2)$ is the final column; vertex 1 is at (0,0) and vertex 3 is at (2,2). The time goes approximately as N^2; most of the work is in calculating elements of the matrix. Solving the matrix takes time proportional to N^3, but the proportionality constant is smaller than that of the N^2 term. Figure 7 is a log-log plot of TFE against N; TFE seems to be converging as N^{-3} for third-order splines and as N^{-4} for fourth-order splines. These rates of convergence are the optimum rates for approximating smooth functions by B-splines,[30] it is of interest to see them apparently applying for nonsmooth functions. The

Table I

kord	nknots	N	time	TFE	$\int \psi\, ds$	ϕ at 3	$\Phi(1/2,1/2)$
3	0	15	1.19	0.1067	0.997242	0.5482	0.6189
3	1	21	1.98	0.0255	1.000287	0.5069	0.6196
3	2	27	3.21	0.0231	0.999932	0.5011	0.6193
3	3	33	4.91	0.0123	0.999951	0.4999	0.6192
3	4	39	7.09	0.0080	0.999965	0.4997	0.6192
4	0	21	1.85	0.0235	0.999838	0.5041	0.6192
4	1	27	2.89	0.0135	0.999958	0.4997	0.6194
4	2	33	4.58	0.0066	0.999956	0.4987	0.6193
4	3	39	6.52	0.0034	0.999961	0.4995	0.6192
4	4	45	9.29	0.0024	0.999968	0.4995	0.6193

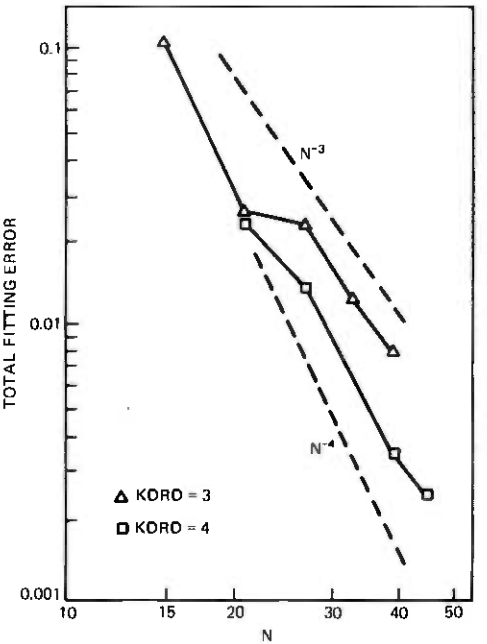


Fig. 7—Total fitting error (TFE) vs the number of basis functions, for third-order and fourth-order splines. Lines proportional to N^{-3} and to N^{-4} are shown for comparison.

last three columns appear to have converged to the accuracy allowed by the finite precision of the calculations. Matrix elements are calculated to a relative precision of about 10^{-6}, and the matrix has a condition number on the order of a few hundred, so accuracy of a few parts in 10^4 is all that can be expected for boundary values. $\int \psi(s)\, ds$ can be more accurate, since the integration can average out the boundary errors.

For these examples, the same number of knots was used on each side. Other examples may require differing numbers of knots on different sides. The intuition of the user is valuable in deciding on the number of knots per side.

APPENDIX

Derivation of Integral Equation (4)

Equation (4) can be derived in various ways. We use a derivation modeled on the derivation of (3) as sketched in Ref. 13. Start with Green's identity in two dimensions.

$$\int \int_D (u\nabla^2 v - v\nabla^2 u)\, dA = \int_\Gamma \left(u \frac{\partial v}{\partial n_s} - v \frac{\partial u}{\partial n_s} \right) ds.$$

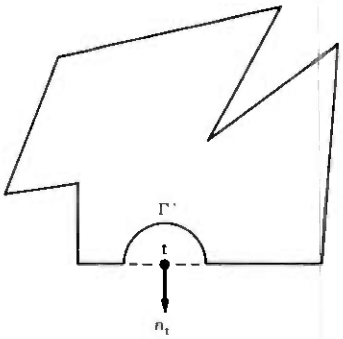


Fig. 8—Integral equation (4).

This identity is usually stated to hold for u and v which are C^2 inside D and C^1 in $D + \Gamma$, but is more generally true. For example, the condition on u can be weakened to include u's which have corner singularities as discussed in the body of the paper. The region D need not be simply-connected.

Pick any fixed point (x_t,y_t) at arc length t on Γ, at a smooth part of Γ. Let $\mathbf{n}_t$ be the outward-pointing normal at t. Choose

$$u(x,y) = \Phi(x,y) - \Phi(x_t,y_t)$$

$$v(x,y) = \mathbf{n}_t \cdot \nabla G(x,y;x_t,y_t),$$

and apply Green's identity to the region D', which is D minus a sector of a circle, with radius ϵ, centered at t (Fig. 8). Let Γ' be the circle sector. In D', $\nabla^2 u = 0$ and $\nabla^2 v = 0$, so the area integral is zero.

Consider the Γ' integral, and use polar coordinates (r,θ) centered at t. Let $\mathbf{e}_r$ be the unit vector at (r,θ) pointing away from the point t. On Γ', $\mathbf{n}_s = -\mathbf{e}_r$,

$$v = -\mathbf{n}_t \cdot \mathbf{e}_r/\epsilon,$$

$$\partial v/\partial n_s = -\mathbf{n}_t \cdot \mathbf{e}_r/\epsilon^2.$$

Expand $\Phi(x,y)$ about (x_t,y_t). For (x,y) on Γ',

$$\Phi(x,y) = \Phi(x,y_t) + \epsilon \mathbf{e}_r \cdot \nabla\Phi(x_t,y_t) + O(\epsilon^2),$$

$$\frac{\partial \Phi}{\partial n}(x,y) = -\frac{\partial \Phi}{\partial r} = -\mathbf{e}_r \cdot \nabla\Phi(x_t,y_t) + O(\epsilon),$$

where $\nabla\Phi(x_t,y_t)$ is an abbreviation for $\nabla\Phi(x,y)|_{x_t,y_t}$. The integrals over Γ' may be evaluated explicitly in the limit as $\epsilon \to 0$.

$$\lim_{\epsilon\to 0}\int_\Gamma u\frac{\partial v}{\partial n_s}\,ds = \lim_{\epsilon\to 0}\int_0^\pi [\epsilon \mathbf{e}_r\cdot\nabla\Phi(x_t,y_t)+O(\epsilon^2)][-\mathbf{n}_t\cdot\mathbf{e}_r/\epsilon^2]\epsilon\,d\theta$$

$$= -\lim_{\epsilon\to 0}\int_0^\pi [\mathbf{e}_r\cdot\nabla\Phi(x_t,y_t)+O(\epsilon)][\mathbf{n}_t\cdot\mathbf{e}_r]\,d\theta$$

$$= -\frac{\pi}{2}\,\mathbf{n}_t\cdot\nabla\Phi(x_t,y_t) = -\frac{\pi}{2}\,\psi(t).$$

In performing the integral, we used the identity

$$\int_0^\pi (\mathbf{a}\cdot\mathbf{e}_r)(\mathbf{b}\cdot\mathbf{e}_r)\,d\theta = \frac{\pi}{2}\,\mathbf{a}\cdot\mathbf{b},$$

true for constant vectors **a** and **b**. The other integral is evaluated similarly.

$$\lim_{\epsilon\to 0}\int_\Gamma\left(-\nu\frac{\partial u}{\partial n_s}\right)ds$$

$$= \lim_{\epsilon\to 0}\int_0^\pi\left[\frac{\mathbf{n}_t\cdot\mathbf{e}_r}{\epsilon}\right][-\mathbf{e}_r\cdot\nabla\Phi(x_t,y_t)+O(\epsilon)]\epsilon\,d\theta$$

$$= -\frac{\pi}{2}\,\psi(t).$$

Thus the integral over Γ' gives $-\pi\psi(t)$, in the limit $\epsilon\to 0$. Again in the limit $\epsilon\to 0$, the integral over the remainder of Γ becomes a Cauchy principal-value integral, and (4) is obtained. The argument depends only on the most singular terms in the Green's function, and so is easily generalized to the axisymmetric case.

REFERENCES

1. F. Dorr, "The Direct Solution of the Discrete Poisson Equation on a Rectangle," SIAM Rev., *12* (1970), pp. 248–263.
2. B. L. Buzbee, G. H. Golub, and C. W. Neilson, "On Direct Methods for Solving Poisson's Equation," SIAM J. Numer. Anal., *7* (1970), pp. 627–656.
3. O. D. Kellogg, *Foundations of Potential Theory*, New York: Dover, 1953 (reprint of original 1929 edition).
4. N. I. Muskhelishvili, *Singular Integral Equations*, P. Noordhoff, Gröningen, Holland, 1953.
5. M. A. Jaswon, "Integral Equation Methods in Potential Theory. I," Proc. Roy. Soc. London, *A275* (1963), pp. 23–32.
6. G. T. Symm, "Integral Equation Methods in Potential Theory. II," Proc. Roy. Soc. London, *A275* (1963), pp. 33–46.
7. J. K. Hayes, "The LAPLACE FORTRAN Code," Los Alamos Scientific Laboratory Report LASL-4004, October, 1968.
8. R. F. Harrington, *Field Computation by Moment Methods*, New York: MacMillan, 1968.
9. J. L. Blue, "On Finding the Admittance Matrix of a Thin-Film Network by Solving the Reduced Wave Equation in Two Dimensions," J. Comp. Phys., *7* (1971), pp. 327–345.
10. J. L. Blue, "Boundary Integral Solutions of Laplace's Equation," Bell Laboratories Computing Science Technical Report No. 60, April, 1977.
11. S. I. Feldman, private communication.

12. G. Hsiao and R. C. MacCamy, "Solution of Boundary Value Problems by Integral Equations of the First Kind," SIAM Rev., *15* (1973), pp. 687–705.
13. R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. 2, New York: Interscience, 1962.
14. S. Bergman, *Integral Operators and Partial Differential Equations*, Ergebnisse der Mathematic und Ihrer Grenzgebiete, Neue Folge, Heft 23, 1961.
15. I. N. Vekua, *New Methods for Solving Elliptic Equations*, Amsterdam: North-Holland Publishing Company, 1967.
16. J. R. Cannon, "The Numerical Solution of the Dirichlet Problem for Laplace's Equation by Linear Programming," SIAM J. Appl. Math., *12* (1964), pp. 233–237.
17. P. Businger and G. H. Golub, "Linear Least Squares Solutions by Householder Transformations," Num. Math., *7* (1965), pp. 269–276.
18. J. A. Cochran, *Analysis of Linear Integral Equations*, New York: McGraw-Hill, 1972.
19. R. J. Hanson, "A Numerical Method for Solving Fredholm Integral Equations of the First Kind Using Singular Values," SIAM J. Numer. Anal., *8* (1971), pp. 616–622.
20. J. M. Varah, "On the Numerical Solution of Ill-Conditioned Linear Systems with Applications to Ill-Posed Problems," SIAM J. Numer. Anal., *10* (1973), pp. 257–267.
21. A. N. Tikhonov, "Solution of Nonlinear Integral Equations of the First Kind," Soviet Math. Dokl., *5* (1964), pp. 835–838.
22. C. L. Lawson, "Applications of Singular Value Analysis," *Mathematical Software*, ed. J. Rice, New York: Academic Press, 1971.
23. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, New York: Dover, 1965.
24. M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, Englewood Cliffs: Prentice-Hall, 1967.
25. J. D. P. Donnelly, "Eigenvalues of Membranes with Reentrant Corners," SIAM J. Numer. Anal., *6* (1969), pp. 47–61.
26. R. S. Lehman, "Developments at an Analytic Corner of Solutions of Elliptic Partial Differential Equations," J. Math. and Mech. *8* (1959), pp. 727–760.
27. J. A. Morrison, "Charge Singularity at the Vertex of a Slender Cone of General Cross-Section," SIAM J. Appl. Math. *33* (1977), pp. 127–132.
28. H. B. Curry and I. J. Schoenberg, "On Polya Frequency Functions IV: The Fundamental Spline Functions and their Limits," J. of Anal. and Math., *17* (1966), pp. 71–107.
29. C. deBoor, "On Calculating with B-Splines," J. Approx. Th., *6* (1972), pp. 50–62.
30. C. deBoor, "On Uniform Approximation by Splines," J. Approx. Th., *1* (1968), pp. 219–235.
31. J. L. Blue, "Automatic Numerical Quadrature," B.S.T.J., *56*, No. 9 (November 1977), pp. 1651–1678.
32. T. N. L. Patterson, Algorithm 68, "Algorithm for Automatic Numerical Integration over a Finite Interval," Comm. ACM, *16* (1973), pp. 696–699.
33. R. A. Sack and A. F. Donovan, "An Algorithm for Gaussian Quadrature Given Modified Moments," Num. Math., *18* (1972), pp. 465–478.
34. P. A. Fox, A. D. Hall, and N. L. Schryer, "The PORT Mathematical Subroutine Library," ACM Trans. Math. Software, *4* (1978), pp. 104–126.
35. J. F. Hart, et al., *Computer Approximations*, New York: John Wiley, 1968.
36. C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, Englewood Cliffs: Prentice-Hall, 1974, Chapter 20.
37. J. A. Lewis and J. McKenna, "The Field of a Line Charge Near the Tip of a Dielectric Wedge," B.S.T.J., *56*, No. 3 (March 1976), pp. 335–342.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 8, October 1978
Printed in U.S.A.

Failure Mechanisms and Reliability of Low-Noise GaAs FETs

By J. C. IRVIN and A. LOYA

(Manuscript received February 24, 1978)

The degradation and failure of low-noise GaAs FETs have been accelerated by various stress-aging techniques including storage at elevated temperatures with and without bias, exposure to humid atmospheres with and without bias, and temperature cycling. Several time-temperature-bias-induced catastrophic failure mechanisms have been observed, all involving the Al gate metallization. These mechanisms are Au-Al phase formation, Al electromigration, and electrolytic corrosion. Each of these processes results ultimately in an open gate. Accelerated aging also produces gradual, long-term degradation in both dc and RF characteristics, though the two are not always correlated. In fact, contrary to some expectations, contact resistance may increase almost two orders of magnitude without significant degradation in the noise figure or gain of a low-noise transistor. Besides contact resistance, other mechanisms such as traps in the channel are thought to play a role in the degradation of RF properties. It was found that all the important degradation mechanisms are bias-sensitive and that aging without bias gives erroneously long lifetime projections.

The cumulative failure distributions for the mechanisms observed approximate a log-normal relation with standard deviations between 0.6 and 1.4. The relevant degradation or failure processes have activation energies near 1.0 eV, which give rise to projected median lifetimes at 60°C (channel temperature) over 10^7 hours and corresponding failure rates (excepting infant mortality) under 40 FITs (40 per 10^9 device-hours) at 20 years of service.

I. INTRODUCTION

This paper describes the goals, experimental methods, and results of a study of the reliability of low-noise gallium arsenide field-effect transistors[1] involving about 1500 devices and 1.5 million device-hours of aging. The ultimate purpose of this work is twofold: (*i*) to calculate

the probable failure rate as a function of time; (*ii*) to identify the failure and degradation mechanisms and propose corrective action where possible. To estimate the failure rate of the device for any given operational conditions, each of these mechanisms should be characterized in terms of the nature of its cumulative failure distribution, the median life and standard deviation of the distribution, and the activation energy of the mechanism.

Since the reliability of a GaAs FET depends intimately on its structural details, especially the choice of metallization, the structure of the devices studied is described in Section II of this report. The various acceleration methods, measurement techniques, and other aspects of the experimental program will be discussed in Section III. The failure modes observed may be categorized as either sudden or gradual. The former are marked by a complete collapse of dc and RF properties and are almost always associated with a failure of the gate metallization. They are the subject of Section IV. (Burn-out due to undesirable voltage pulses is considered a matter of handling technique or circuit design and is not investigated in the present work.) The gradual failures involve degradation of the important RF properties, especially the noise figure and the gain. There is an associated, though not well correlated, change in the observable dc characteristics. The gradual degradation of low-noise GaAs FETs is discussed in Section V. In Section VI, the pertinent failure statistics are summarized and some cumulative failure distributions are shown. Finally, estimated failure rates under typical operational conditions are presented in Section VII, together with some prognoses with regard to other operating environments.

II. THE STRUCTURE

Two slightly different versions of low-noise GaAs FETs were studied, differing primarily in the details of the gate bonding pad. In the earlier form, shown in Fig. 1, the Al gate metallization extends under the entire bonding area which is covered by a titanium-platinum-gold final metallization.[1] In the later version, the bonding area is separated laterally from the Al to which it is connected by the Ti-Pt-Au final metal. This is shown in Fig. 2. In both cases, the gate bonding pads, as well as the source and drain bonding pads, lie on the semi-insulating substrate. A schematic cross-sectional view of the source and drain contacts is given in Fig. 3. The ohmic contact consists of a layer of 88-percent Au/12-percent Ge, topped successively by a layer each of silver and gold and then alloyed. A final metallization of Ti-Pt-Au, as described above, is applied on top of the alloyed ohmic contact.

The active n-type layer is 3000 to 6000 Å thick with a donor density of approximately 1×10^{17} cm^{-3}. An n+ layer, about 3000 Å thick and with a donor density around 2×10^{18} cm^{-3} underlies the source and drain

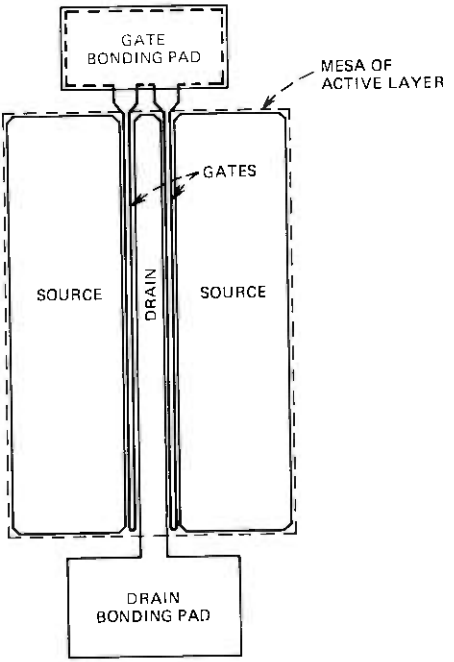


Fig. 1—Plan view of early low-noise GaAs FET.

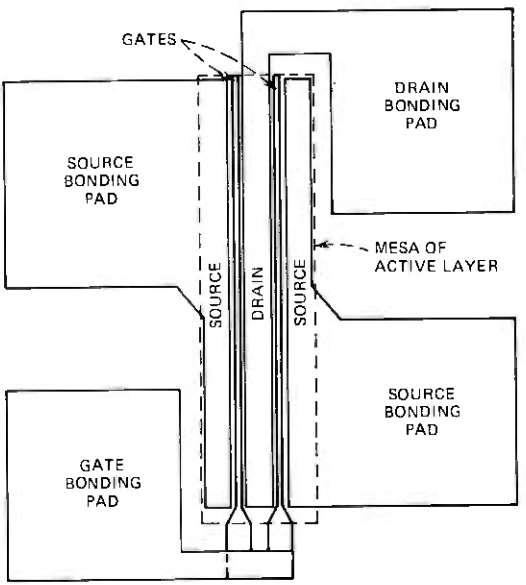


Fig. 2—Plan view of later model low-noise GaAs FET.

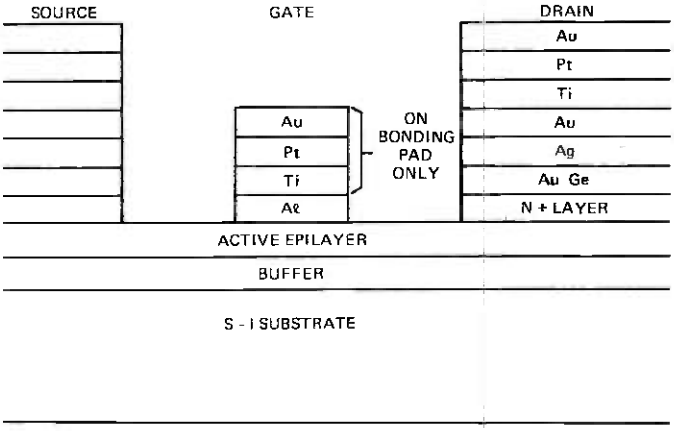


Fig. 3—Cross section of GaAs FET contact metallurgy (relative thicknesses not to scale).

contacts. A buffer layer with a donor density $< 10^{13}$ cm^{-3} and 2 to 5 μm thick separates the active layer from the semi-insulating (Cr-doped) substrate. Except in a few cases, no passivation layers were present on the finished chip. The chip size is 0.5 mm square and 50 μm thick. Each chip is bonded in a 2.5-mm square package which is hermetically sealable.

III. EXPERIMENTAL METHOD

To accelerate the degradation or failure of the GaAs FETs, a number of methods were used. The primary of these (about 1,000,000 device-hours) was aging at elevated temperatures, both with and without bias. The ambient was air at temperatures of 88°, 180°, 220°, 250°, and 275°C. The bias duplicated normal operating values consisting of 5V on the drain and a gate bias of −0.1 to −2.V, as necessary to produce 15 mA of drain current.[2] At this bias, the elevation of the channel temperature above ambient is estimated to be about 8°C. Due to the wide-band instability of GaAs FETs, RF oscillations will readily occur even at high temperatures. Such oscillations are in themselves sometimes destructive and they may also produce instantaneous, or by rectification, dc bias values of unknown and uncontrolled magnitudes. Thus considerable effort was devoted to the suppression of oscillations by various means. Dissipative media (Eccosorb), RC networks, and ferrite beads were employed, with various degrees of effectiveness. One principal difficulty was the incompatibility of some of the stabilizing components with the high temperatures involved and the fact that it is desirable to place such stabilization as near the FET as possible. Ferrite beads were the most effective and usually succeeded in quelling oscillation. Zener diodes were also employed in both the drain and gate supplies to protect the FET from destructive voltage transients.

While aging units under bias, the dc bias values could be monitored and were recorded daily. Catastrophic failures, that is, short or open circuits in the drain or gate, were thereby readily observed. RF properties could only be determined by periodic removal of the units and testing in either a tunable or a fixed-tuned amplifier. Thus, at intervals which ranged from 100 to 1000 hours, groups of FETs were temporarily removed from the aging environment and dc and RF measurements were performed. The dc characterization consisted of photographing the output characteristics from which the saturated drain current, I_{DSS}, and the low-field source-drain resistance, R_S, i.e., (dV_{DS}/dI_{DS}) at $V_G = V_{DS} = 0$, were determined. The RF parameters measured were the noise figure, NF, and the associated gain, G, at 4 GHz and 15 mA, 5V drain bias. In the case of the tunable amplifier used in the earlier stages of this study, the minimum NF was obtained; the NF obtained in the fixed-tuned amplifier (a modified Western Electric 652A[2]) was near-minimum, but not actually optimized for each device.

Another failure-acceleration technique used was storage under bias in air of 85-percent relative humidity at 85°C (referred to hereafter as 85/85). In these experiments (about 150,000 device-hours), only the gate was biased at $-4\frac{1}{2}$ or -6V with respect to the grounded drain; the source floated. Some devices were aged without bias, of course, as was also the case at the higher temperatures. The reverse leakage and the continuity of the gate were checked hourly, then daily, and finally weekly in these experiments, which varied in duration from a few hours to a year. Periodic RF measurements were generally not performed on these devices since catastrophic failure due to electrolytic corrosion of the gate was the mechanism studied. The purpose of the 85/85 experiments was to determine the integrity of "hermetically sealed" packages, the presence of corrosive contaminants therein, and the effectiveness of various waterproofing or passivation coatings.

To test the security of the thermocompression bonds of the 25-μm diameter gold leads to the source, drain, and gate bonding pads as well as to test the hermetic seal, devices were cycled, under bias and without bias, between $-40°$ and $+125°$C. The continuity of the bonds was tested before and after cycling as well as during cycling, in a few cases. Some devices were also thermal-shocked by alternate immersions in freezing and boiling water. These tests will not be discussed further, since in no case (out of 28,500 bond-cycles) was an open bond observed. In fact, no open bonds have been encountered among any of the over 1500 devices tested, before or after the various aging regimes described above. No centrifugal or vibration tests were employed in this program.

Lastly, a few lots of FETs have been aged without acceleration—that is, under normal operating dc bias (no RF) at room temperature (27°C), totaling 250,000 device-hours.

IV. CATASTROPHIC FAILURES

4.1 Au-Al phase formation

The most common cause of complete dc and RF failures encountered in this study was related to the formation of Au-Al compounds (a version of this on Si devices is known as purple plague). A dramatic example is shown in Fig. 4. In this case, the force of thermocompression bonding to the top Au layer has ruptured the integrity of the intervening Ti-Pt layer and caused contact between the Au top layer and wire and the bottom Al layer at the end of the gate structure. However, the loss of gate continuity is not due to embrittlement and subsequent parting of the bond nor to the high resistance of the Au-Al compound. Fed by the surplus of available Au, the Au-Al system (of which Au_5Al_2 is the favored end product) acts like a sink for the surrounding Al and has produced voids in the Al gate structure (the Kirkendall effect). The voids in the gate are visible in Fig. 4. The presence of Ga may catalyze this reaction

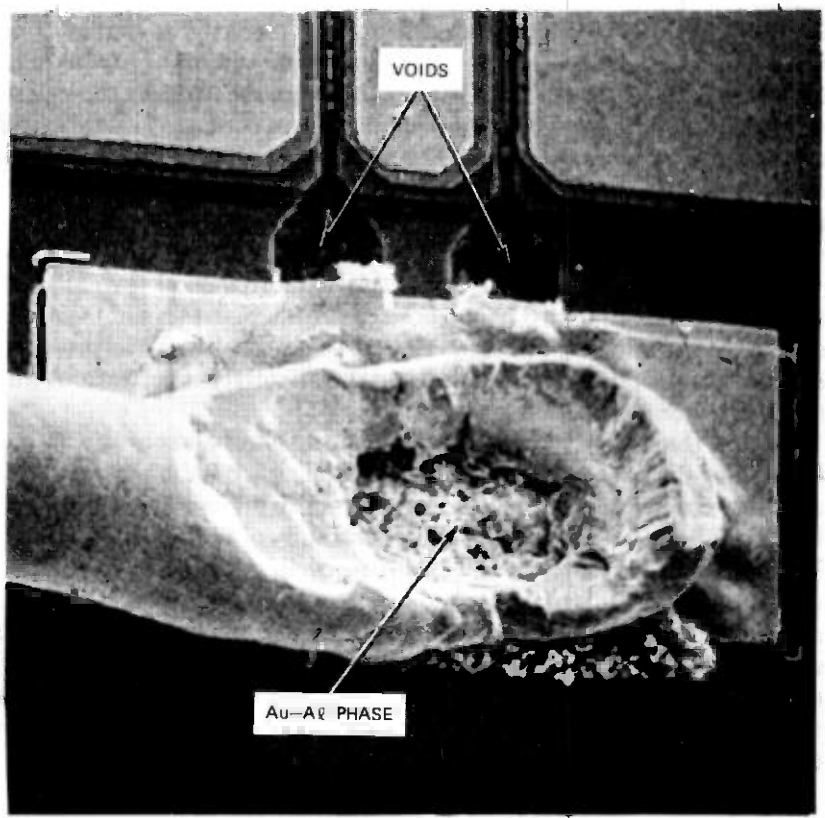


Fig. 4—SEM photo of Au-Al phase at gate-bonding pad of early model FET after aging at 250°C with bias.

as does Si in the case of Si devices[3] (or other devices with SiO_2 layers). Figures 5 and 6 show other examples of Au-Al interaction leading to open gates. Note that the FETs of Figs. 5 and 6 employ the layout shown in Fig. 2, in which the gate bonding pad is separated laterally from the Al structure. However, Au and Al still were able to interdiffuse due to a slight mask misalignment. Both the devices shown in Figs. 4 and 5 were

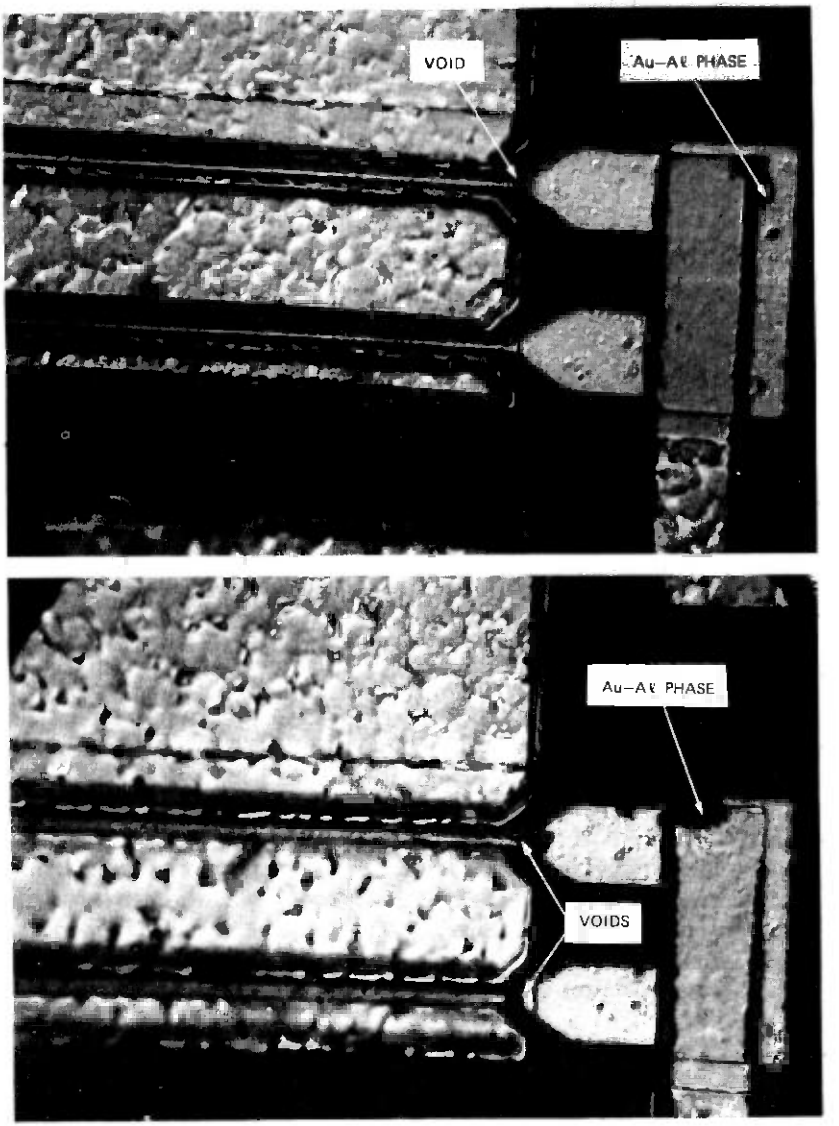


Fig. 5—Optical photos of Au-Al phase and consequent open gates in later model FETs after aging 144 hours at 250°C with bias.

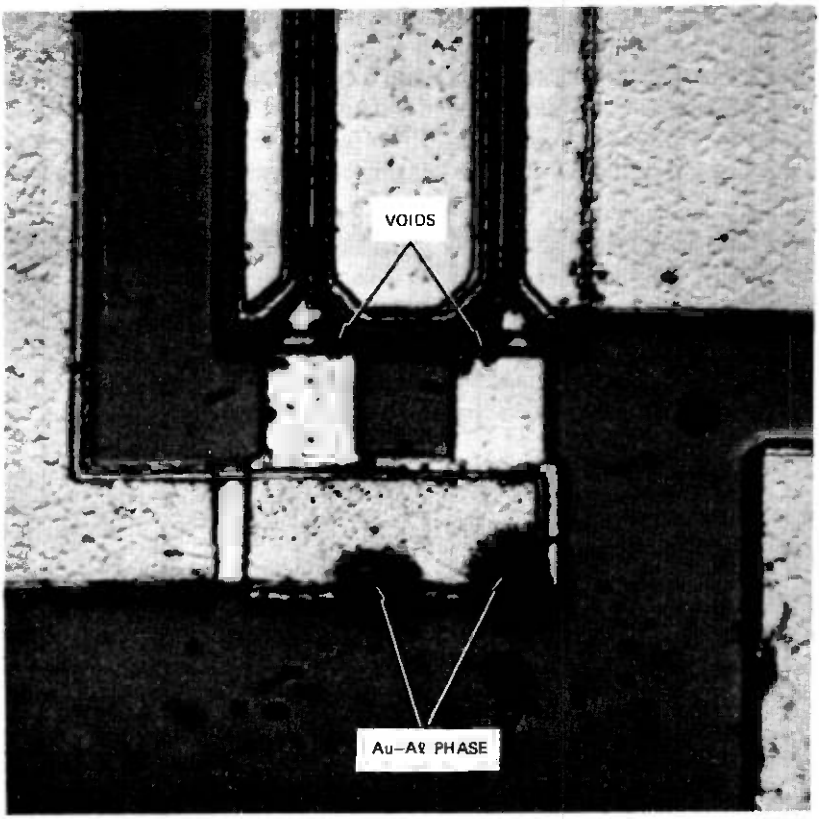


Fig. 6—Optical photo of Au-Al phase and consequent open gates in later model FET after aging at 144 hours at 250°C without bias. (This device is from a slice much more prone to Au-Al phase formation than the device of Fig. 5.)

aged under bias. The FET of Fig. 6 was aged at the same temperature (250°C) without bias.*

The detailed processes of this failure mechanism have not been fully unravelled, but a number of pertinent observations are summarized below.†

(*i*) Every FET that has failed due to an open gate (more than 100 have been examined) exhibits a Au-Al interaction site somewhere in the region where Au and Al overlap. This site may appear to be insignificantly small.

(*ii*) The median time to failure due to open gates is 3 to 10 times longer

* Even with perfect registration, Au-Al contact is expected to occur eventually due to diffusion through the Ti-Pt barrier, though no such cases have been observed so far in these experiments. A mask modification which eliminates the Au layer from the bridge between bonding pad and gate virtually prevents any Au-Al reaction.

† The authors are indebted to A. T. English for his assistance in analyzing the inter-diffused gate structures.

among units aged without bias than among identical devices aged at the same temperature with bias.

(*iii*) In all bias-aged failures, a void appears in the gate stripe just where it broadens. This is the point of maximum current density in the gate metallization. Voids may also (but do not always) appear in the broad Al area or at the mesa step.

(*iv*) Failures among units aged without bias have voids scattered generally about the broad Al regions and frequently over the mesa step, but usually not at the aforementioned point of maximum current density.

It is apparent from these observations that there is a decided dependence on the presence of bias. It is important to note that, at 250°C, gate leakage currents at operating bias are one to two orders of magnitude larger than at room temperature, i.e., 20 to 200 μA instead of 1 to 5 μA. Even so, the maximum gate current densities during bias aging are calculated to be only 1×10^4 A/cm^2. This value is generally considered "safe" with regard to electromigration at 250°C. Furthermore, no correlation is found between failure and the gate currents of individual units during aging. Thermal dissipation in biased units in 250°C ambient raises the channel temperature to approximately 258°C. However, unbiased units in 275°C ambient have a much lower incidence of open gates than the 258°C bias-aged units. Thus, this aspect of self-heating cannot explain the bias dependence. Also, joule self-heating within the gate stripe itself is calculated to cause less than 1°C temperature rise, which, of course, also fails to justify much electromigration at these apparently modest current densities. However, current densities in the gate structure are not accurately calculable, since actual Al cross sections vary with the topography of the surface, especially at the mesa edge. It is known that electromigration is influenced by grain size and very little of the voluminous electromigration literature treats stripe widths as small as the 1-μm gates involved here. Thus, electromigration in conjunction with Au-Al phase formation is tentatively thought to be responsible for gate failures in bias aging. Electromigration would transport Al down the stripe away from the bonding pad while Au-Al phase formation causes diffusion in the opposite sense. Perhaps electromigration inhibits Al atoms near the gate throat from replacing the atoms just downstream in the wider portion of the structure (where the current density is less) which are being drawn by diffusion toward the Au-Al compound. The formation of voids may be accelerated by this tug-of-war situation.

If the above hypotheses regarding this gate failure mechanism are correct, the temperature dependence would be quite complex. The activation energy of Au-Al phase formation has been variously reported with values between 0.6 and 1.0 eV.[4] (In any case, the interaction with Ga may alter these values.) Al electromigration (at constant current

density) is reported to have an activation energy of 0.5 to 0.7 eV.[5] However, as mentioned earlier, the gate leakage current itself is temperature-sensitive. The latter two effects together, it is calculated, should give the electromigration an effective activation energy of 1.2 eV.

The experimental situation, unfortunately, is not much clearer. On slices (of the type in Fig. 2) where Au-Al phase formation occurs, its occurrence is quite erratic, depending as it does on slight vagaries in alignment, lift-off, etching, and other details of pattern formation. Thus, among devices aged *without bias*, the median life (ML) varies greatly from slice to slice, though the activation energy observed is fairly consistent and near 1 eV. Electromigration varies as the second or third power of current density which, in turn, differs widely from one unit to another. However, no failures have been observed which appeared to be due to electromigration alone. When units are aged *under bias* at elevated temperature, both mechanisms are thought to be operative, though the degree of dominance by the one mechanism or the other probably varies both among devices and as a function of time during the course of void formation. Initially, phase formation is probably dominant, but when the cross-sectional area has been diminished enough and the local current density increases, electromigration becomes more important. In principle, it is inappropriate to use an activation energy to characterize this joint process consisting of two mechanisms. However, since both mechanisms are expected in this case to have an activation energy near 1 eV, as described above, it is a useful approximation to apply an "activation energy" to the combined effect. As expected, the experimental data are not entirely consistent, but are grouped about a value of 1.0 eV.

No typical ML can be cited for bias-aged devices, since many slices are entirely free of this mechanism. However, in the worst case, an ML of 94 hours at 250°C with bias has been observed. It is important to note that this mechanism has been observed in the present study in devices bias-aged at temperatures as low as 180°C in times as short as 240 hours. Weaver and Brown detected Au-Al interdiffusion at 84°C in 3 hours.[6] Thus, Au-Al phase formation and subsequent destruction of GaAs FETs in which the choice of metallurgy and layout permits this combination cannot be dismissed as an exclusively high temperature phenomenon. However, an appropriate layout can completely eliminate the possibility of Au-Al phase formation.

4.2 Electrolytic corrosion

Figure 7 is an example of electrolytic gate corrosion. The corrosion shown was produced by a 2-hour exposure to an atmosphere of 85°C/85% RH with 6 V negative bias on the gate. The device was uncapped. This corrosion is clearly electrolytic, since in the absence of gate bias no sig-

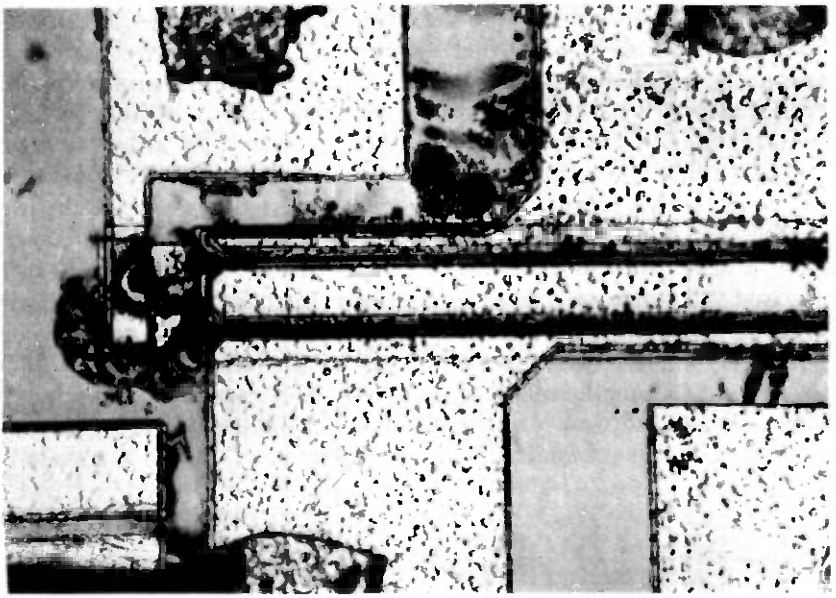


Fig. 7—Optical photograph of FET after 48 hours of aging, uncapped, in 85°C/85% RH humidity chamber, showing electrolytic corrosion of gate.

nificant corrosion is observed. Electrolytic corrosion of unprotected Al structures is well known, of course, from reliability studies of silicon devices.[7] The unusually close electrode spacing and consequent high fields in GaAs FETs as well as the minuteness of the Al gates make them prime candidates for this failure mechanism, in humid conditions. The acceleration factors relative to both varying humidity and temperature have already been reported in the Si device literature[8,9] and is summarized in Section VI.

Various passivation or protective coatings have been proposed for the prevention of electrolytic corrosion in GaAs FETs. Schemes that only coat the GaAs, such as grown oxides, would not be expected to be effective. The highly irregular topography of GaAs FETs complicates the task of achieving a continuous, impervious, pinhole-free, protective film. Equally important is the requirement that the film have small dielectric constant and low microwave loss; otherwise, the sacrifice in microwave performance is unacceptable. None of the films explored in this study fulfills all these specifications perfectly.

A hermetically sealed package can provide permanent protection against electrolytic corrosion from external humidity and without any sacrifice in RF performance, at least at frequencies where a package can be tolerated. It is suspected, however, as observed already among Si devices,[10] that residual impurities entrapped inside the package can produce destructive electrolytic corrosion, although the seal remains

intact. Water and chlorine are the chief offenders, and the trace amount of both which may be adsorbed on the package interior surface are apparently sufficient to produce corrosion. Patches of unremoved photoresist may also harbor enough impurities to cause corrosion. Figure 8 shows the corroded gate of a sealed device that failed after 240 hours under bias in 85/85 though no leak was detectable after removal from the chamber. A Krypton 85 radio-tracer technique was used for leak detection, which has a sensitivity in this case of 10^{-8} std cm^3/s. Though a leak below the detectable limit cannot be excluded and might have caused the corrosion, the Si experience[10] must be borne in mind and residual contamination suspected. This would be confirmed by the discovery of electrolytic corrosion in sealed devices aged at 80° to 90°C in dry air. Among the relatively few devices (30) aged in this manner in the present study, no corrosion has been observed. However, among a group of devices which had failed optical inspection due to unusually

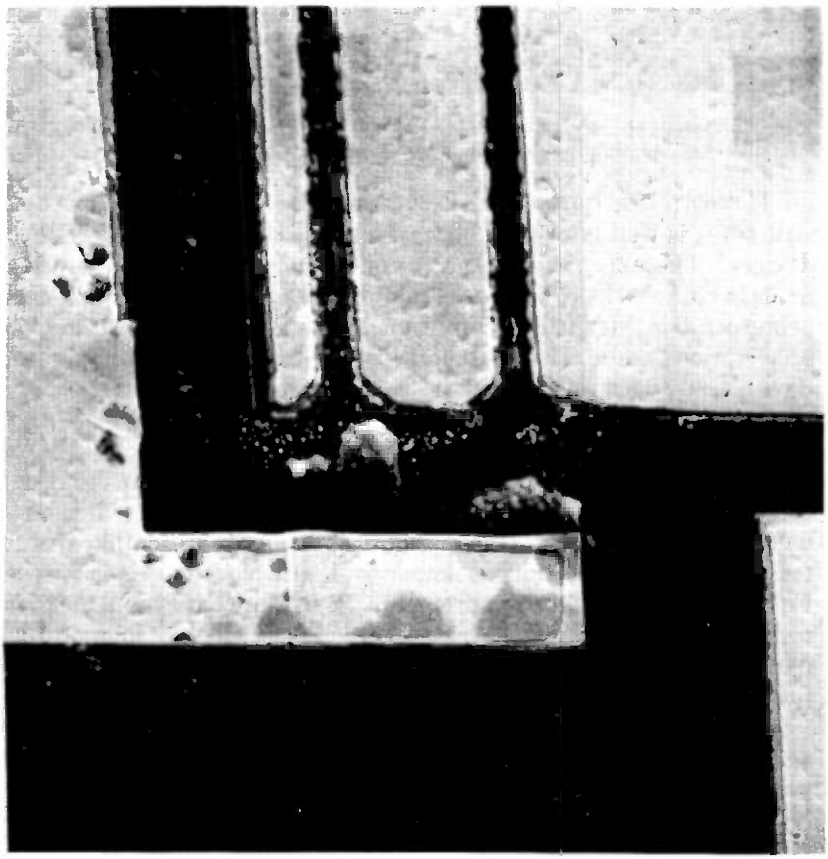


Fig. 8—SEM photograph of corroded gate of FET after 1816 hours of aging, sealed and leak-tight, in humidity chamber. Analysis reveals traces of Cl at corrosion site.

large amounts of photoresist remaining on the chip and which were sealed and aged in 85/85, the incidence of corroded gates was 50 percent in 2000 hours. Altogether, the incidence of electrolytic gate corrosion among "clean" devices which passed optical inspection and which also passed the leak test before and after aging has been about 1 percent in 2000 hours.

V. GRADUAL DEGRADATION MECHANISMS

5.1 High-temperature effects

One of the earliest GaAs FET degradation mechanisms to be discussed was an increase in contact resistance.[11] According to well supported models,[12,13] Ga diffuses out of the crystal into the contact metallization at elevated temperatures. The resulting Ga vacancies probably act as acceptors, compensating the donors in the n-type lamina immediately adjacent to the contact, and thereby increase the contact resistance. The capacity of the metallization for absorbing Ga, or the effectiveness of an interposed barrier to the transport of the Ga, speed or inhibit the degradation process, respectively. In the ohmic contact structure described in Section II and illustrated in Fig. 3, the heavy final gold layer is the largest potential sink for migrating Ga, while the Ag and Ti-Pt layers act as an impeding barrier. As will be seen, however, other factors (such as the alloying cycle) must also play a role in the degradation of ohmic contacts.

By means of special test patterns and an appropriate computer program, the contact resistivity, ρ_c, and the channel resistance of a gateless device, $R(ch)$, were measured on certain FET slices before and after aging at 250°C, both with and without bias (0.3 A/cm, the same current per unit source width as in an operating device). A number of actual FETs from the same slice were also aged at the same temperature, with and without bias, and the usual parameters measured (R_S, I_{DSS}, NF, and G). Some slices (which will be designated Class I) showed virtually no change in any parameters after 500 hours, with or without bias, in either test patterns or actual FETs. Other slices (designated Class II), though nominally identical to Class I in design and fabrication, showed startling changes in certain dc parameters, as summarized for one slice in Table I. Typical values of R_S for unaged FETs were 15 to 30 ohms, of which the contact resistance contribution is only 0.2 to 0.5 ohms. The remainder

Table I — Changes in various parameters after 500 hours of aging at 250°C for a Class II slice

	Test Patterns		Actual FETs			
	ρ_c	$R(ch)$	R_S	I_{DSS}	NF	G
With bias	+5000%	0	+50%	−16%	0	0
Without bias	+5000%	0	+20%	−5%	0	0

of R_S is the resistance of the channel and of the semiconductor portions of the source and drain regions. Thus, it would appear plausible that a 5000-percent increase in contact resistance, as shown in Table I, would cause an approximate doubling of R_S. However, it is noted that the change in ρ_c was not affected by bias, whereas the change in R_S and I_{DSS} was very much bias dependent. (The latter bias dependence is also seen in Fig. 9.) The change in R_S (and corresponding change in I_{DSS}) is therefore not wholly attributable to contact deterioration.

The origin of the bias-dependent component of R_S has not been determined. One possibility is recombination enhanced defect formation,[14] though the hole production at the drain and under the gate seems too small for this effect. It is also suggested that the bias-dependent degradation may be related to the gate, since the test patterns, which were unaffected by bias, have no gates. It is also not understood what the essential difference is between Class I and Class II slices—and the continuous spectrum of behavior between these two extremes. It is thought that the least-controlled processing step may be the contact alloying, which is therefore tentatively blamed for at least a part of the slice-to-slice variation in aging behavior. Fortunately, perhaps because the degraded dc qualities of contacts are capacitively bypassed by RF signals,[15] these wide fluctuations in the degradation of dc parameters are not reflected in the RF performance.

Figure 9 shows the average values of R_S, I_{DSS}, NF, and G for two groups of FETs from the same slice (a Class II slice) aged at 250°C, one group with and the other without bias. Though both the dc and RF characteristics degrade faster with bias than without, it is seen that, while R_S doubles, the noise figure and gain only deteriorate by 0.2 to 0.5 dB. Another example is shown in Fig. 10, where NF and G degrade only 0.2 and 0.3 dB, respectively, while again R_S doubles. (The actual contact resistance increased 50-fold.) Two other interesting cases, both representative of many, are shown in Figs. 11 and 12. The devices of Fig. 11 suffered only negligible changes in R_S, I_{DSS}, and NF after 1300 hours of bias-aging at 250°C, though the gain declined about 0.6 dB. Figure 12 shows the data from a group of devices in which none of the measured dc or RF parameters changed significantly in 1500 hours of bias-aging. The results presented in Figs. 9 through 12 may be summarized as follows:

(*i*) There are significant differences among slices in the way the dc and RF characteristics change upon aging.

(*ii*) Radical deterioration of contact resistance (5000 percent) or of source-drain resistance (100 percent) are accompanied by only minor degradation of RF performance; conversely, NF and G may degrade slightly, even though R_S remains constant.

(*iii*) The median life at 250°C under bias, where failure is defined as

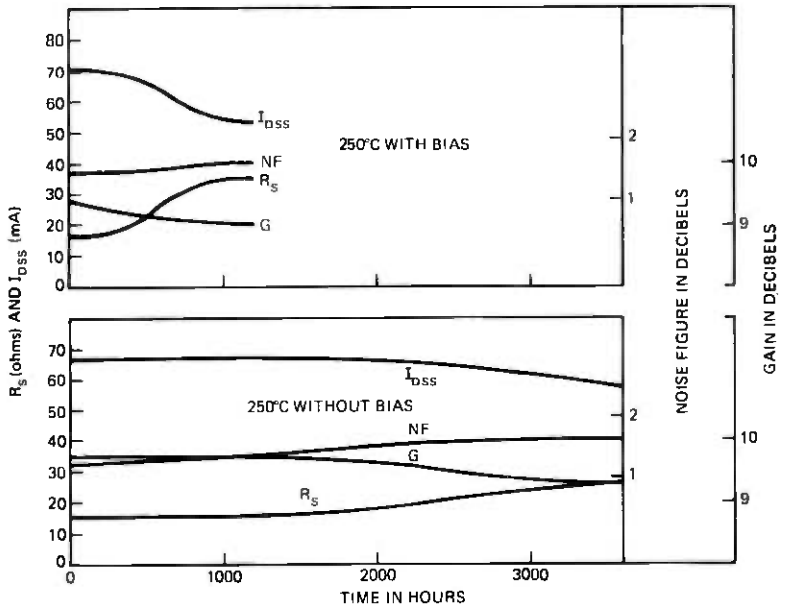


Fig. 9—Plot of median R_S, I_{DSS}, NF, and associated gain as a function of aging time at 250°C for two groups of GaAs FETs from slice (1); Group A aged with bias and Group B aged without bias. (Note acceleration due to bias.)

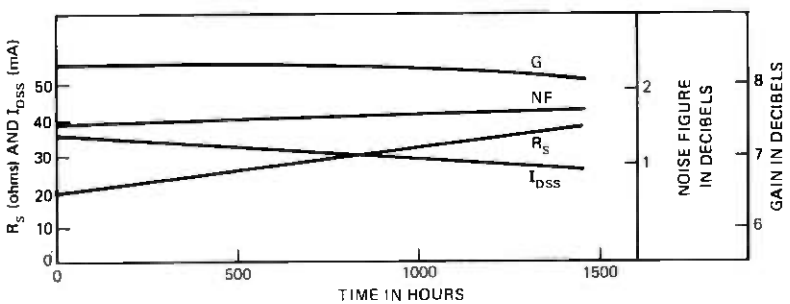


Fig. 10—Plot of median R_S, I_{DSS}, NF, and associated gain of a group of GaAs FETs from slice (2) as a function of aging time at 250°C with bias. (Note NF and G degrade only 0.2 dB, though R_S increases 90 percent.)

an NF or G degradation of equal to or more than 0.2 or 0.8 dB, respectively, is at least 1500 hours.

The observed activation energy of RF degradation is 0.8 to 1.0 eV. It may be noted that many diffusion phenomena within or on the surface of semiconductors, such as might produce traps or scattering centers, have activation energies near 1.0 eV.

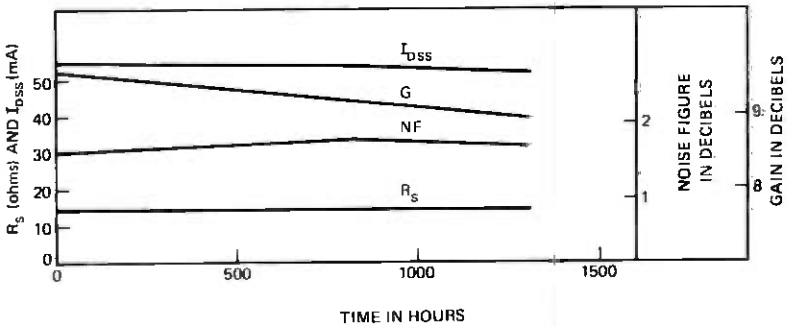


Fig. 11—Plot of median R_S, I_{DSS}, NF, and associated gain of a group of GaAs FETs from slice (3), as a function of aging time at 250°C with bias. (Note NF and G degrade slightly, though R_S in this case remains unchanged.)

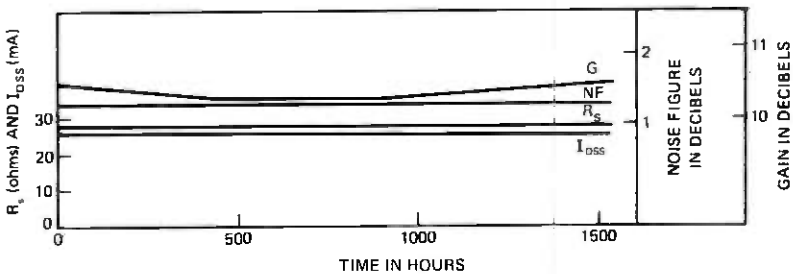


Fig. 12—Plot of median R_S, I_{DSS}, NF, and associated gain of a group of GaAs FETs from slice (4) as a function of aging time at 250°C with bias. (Note that all measured properties remain essentially unchanged.)

VI. FAILURE STATISTICS AND UNOBSERVABLES

6.1 Cumulative failure distributions

The statistics obtained from an aging study depend to some extent upon the definition of failure. A definition of failure can be tailored to a specific failure mechanism and thus be used to sort out data that are relevant exclusively to that mode. Two definitions have been used in various stages of the present investigation.

A. Catastrophic—collapse or radical change in dc output characteristics, usually due to a short or open circuit in one or more of the three electrodes. This definition is especially appropriate for study of the catastrophic mechanisms discussed in Section IV, but ignores any degradation of RF performance not associated with a large change in dc behavior.

B. RF degradation, exclusively—requires that any units that suffer catastrophic failure be subtracted from the population and not counted in the statistics. Figures 9, 10, 11, and 12 were based on such a population. The degree of permitted RF deterioration should be set with system requirements in mind. In this study, unless otherwise specified, a device

was considered to have failed, RF-wise, if the noise figure increased 0.2 dB or more, or the associated gain changed (up or down) by 0.8 dB or more as measured at 4 GHz in a fixed-tuned amplifier.

Figure 13 is a log-normal plot of the cumulative percent failures of type B as a function of aging time for 31 devices representing four separate slices. The aging was performed with bias at 250°C air ambient. The channel temperature is estimated to be 8°C warmer. A few of the devices were sealed, but the majority were not. No difference has been observed in the aging behavior of sealed versus unsealed FETs at this temperature. The data are seen to fit reasonably a straight line, making allowance for the statistical vagaries of small samples, which means they approximate a log-normal distribution. The standard deviation estimate, s, of the line is about 1.3, obtained from the operational calculation

$$s = \ln[t(50)/t(16)],$$

where it is noted the natural logarithm is used and $t(50)$ and $t(16)$ are the times corresponding to 50 and 16 percent cumulative failure, respectively. The median life of this group is about 1700 hours. The results shown in Fig. 13 are typical of the RF degradation observed in this study.

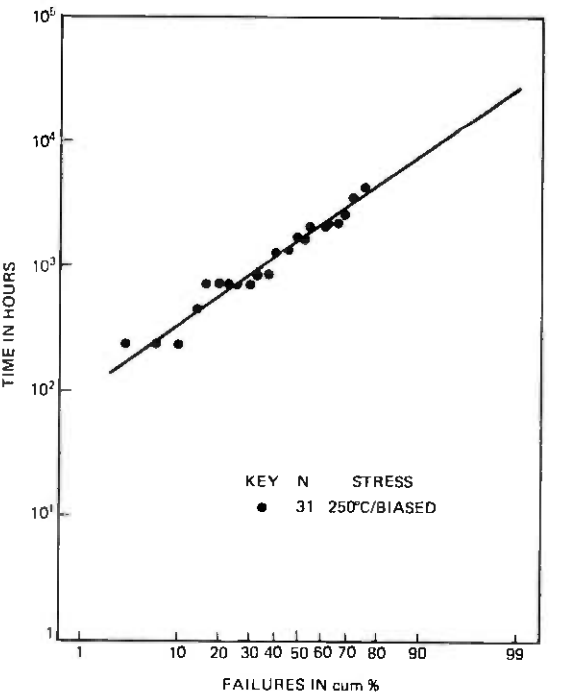


Fig. 13—Log-normal plot of cumulative failure distribution (RF degradation) of a group of GaAs FETs aged with bias at 250°C.

Figure 14 shows the cumulative failure distributions at 250°C of two groups of units from one slice, one aged with bias and the other without bias. The MLs are about 100 hours and 350 hours, respectively, with nearly the same standard deviation of $s = 1$. The failures in this case are all type A and due to Au-Al phase formation, plus an apparent assist from electromigration in the biased group, as discussed in Section 4.1. This slice was unusually susceptible to the Au-Al phase problem and is chosen here to illustrate the failure statistics of that mechanism.

6.2 Humidity acceleration factors

Electrolytic corrosion is accelerated by increased humidity and temperature mainly as a result of and in proportion to the increased electrical conductivity of the surface. The problem has been most recently studied by Sbar and Kozakiewicz,[9] who give acceleration factors with respect to 85°C/85% RH for various encapsulations and temperature/humidity conditions. Though the absolute value of conductance on a GaAs surface may differ from that on a Si, Si_3N_4, or alumina surface, the temperature and humidity dependence are expected to be similar. For 60°C/5% RH (a choice which will be justified later), the Sbar-Kozakiewicz results indicate an acceleration factor of 2×10^5 with respect to 85/85. For a condition of 60°C/25% RH, the factor is about 10^4. Both values apply to an unencapsulated device. For a perfectly sealed device, of course, the external humidity has no effect. The only acceleration of electrolytic

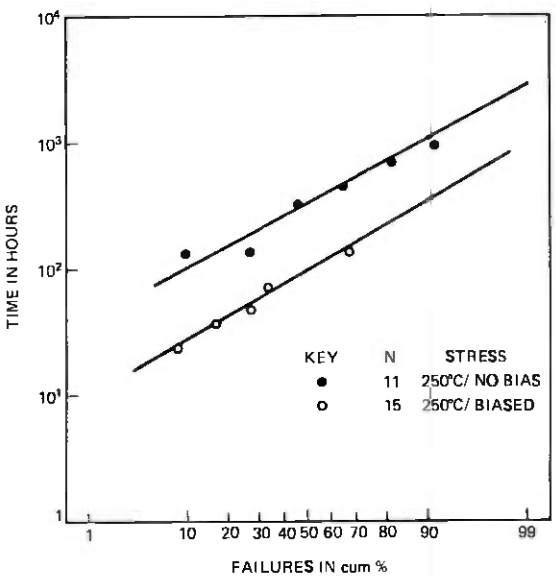


Fig. 14—Log-normal plot of cumulative failure distributions of two groups of GaAs FETs aged at 250°C with and without bias. (All failures were due to gate destruction by Au-Al phase formation plus, in one case, a bias-dependent factor.)

corrosion would be that due to the temperature elevation, assuming there are sufficient contaminants inside the package to produce an electrolyte, but that their release is not temperature-sensitive. The acceleration factor at 85°C relative to 60°C is approximately 3.

6.3 Statistical error

During the course of this study, numerous small changes have been made in the design or fabricational methods of the device under investigation. Since it was desirable to appraise the reliability aspects of each of these variations and both facilities and device are limited in supply, the strategy has been to age many small lots of FETs rather than fewer and larger lots. The relatively small sample size (10 to 20) raises questions about the validity of the statistics obtained. It should be pointed out, therefore, that with a sample size of 10, one can be 90 percent confident that the true ML (i.e., the ML of an infinitely large sample) would be somewhere between 45 and 225 percent of the observed value, assuming a log-normal distribution with a standard deviation of 1. Thus, in the next section where failure rates are calculated, error tolerances may be attached to the values given by noting that in the area where most of the data fall, a factor of 2 in ML produces a factor of approximately 2 to 3 in failure rate.

6.4 Possibility of unobserved failure mechanisms

A relatively small number of FETs have been aged under bias for periods of 0.5 to 1.5 years at moderate or room temperatures (more precisely, 30 units at 180°C for 6000 hours, 10 units at 88°C for 4600 hours, and 30 units at 27°C for 8000 to 14,000 hours, all in air ambient). Judging from the statistics obtained at higher temperatures, provided that only the same mechanisms prevail, no failures or degradation should be apparent in these modest times at lower temperature, except possibly for the Au-Al phase-migration problem. Indeed, this turns out to be the case: five units of one lot of 10 at 180°C have failed due to the Au-Al syndrome, but otherwise no degradation appears in any of the units.

The lack of any failures among the other units provides some lower bound to the activation energy of any failure mechanism which may be important at lower temperature but is obscured at higher temperatures by phenomena with higher activation energies. (This sort of insidious situation has been encountered in Pt-GaAs IMPATTs.[16]) It will be assumed that such a failure mechanism would have a log-normal failure distribution with a standard deviation of 1. It is noted furthermore that the absence of a single failure in a lot of 30 indicates with 90 percent confidence that the true percent failure cannot exceed 10 percent. Finally, it is noted that the ML of any such so-far-unobserved failure mechanism must be at least 1700 hours at 250°C, since that is the longest

observed ML at that temperature due to recognized mechanisms. Combining these arguments leads to the conclusion that the minimum activation energy of a failure mechanism active at 180°C, but obscured at 250°C is 0.7 eV. However, the minimum activation energy for a failure mechanism dominant at 27° to 180°C but not yet observed in this program is only 0.03 eV. The latter figure is rather alarming. It means, in conjunction with the high temperature data, that if such a hypothetical mechanism exists, the projected ML at room temperature would be 30,000 hours and the ML would be only negligibly accelerated by elevated temperatures. The duration of the present reliability program is insufficient to rule out such a possibility. However, it is reassuring that other investigators have reported room temperature tests of low-noise GaAs FETs of similar metallurgy in excess of six years without any failures.[17]

6.5 Infant mortality

Few instances of infant mortality have been found in this study. There are two reasons: (*i*) rigorous optical inspection of all chips before mounting and discarding of any units which appear mechanically or electrically defective after mounting eliminate most devices that might otherwise be candidates for early failure; (*ii*) the small size of the samples used further diminishes the probability of encountering anomalous devices representing only a small proportion of the population. Thus, this work sheds no light on the nature of such early failures except that, with the present fabrication, inspection, and testing procedures, their occurrence is less than 1 percent. The failure and degradation mechanisms discussed here and the failure rates projected pertain to the main body of the population. It must be anticipated that some cases of infant mortality will accompany large-scale production and deployment of this device.

VII. ESTIMATION OF FAILURE RATES

Given the nature of the failure distributions at an elevated temperature and their respective activation energies, and making the all-important assumption that the mechanisms studied at elevated temperature are also the dominant ones at room temperature, and with the further assumption that the nature of each failure distribution is not temperature-dependent (i.e., it stays log-normal with the same s), it remains only to specify the operating conditions in order to calculate the probable failure rates in the field. The maximum ambient temperature in a Bell System radio relay application is 52°C (125°F). The corresponding maximum channel temperature would be 60°C. Though the annual average temperature would certainly be considerably lower, 60°C will be taken as the channel temperature for calculation of failure rates. Three separate cases will be considered.

7.1 Case I: No catastrophic mechanisms

In this case, it is assumed all fabricational steps have been faultless, assuring the absence of any contact between Au and Al, a contamination-free chip and package interior, and a leak-tight seal. The catastrophic failure mechanisms are therefore precluded, and only long-term degradation of RF properties is of concern. The median lifetime (type B) at 258°C was found to be about 1700 hours, and the associated activation energy will be taken as 0.8 eV. Thus, the projected ML at 60°C would be 4×10^7 hours. Taking a standard deviation of 1.0 and using Goldthwaite's curves,[18] the failure rate after 20 years of service is found to be less than 10^{-2} FIT (1 FIT = 1 failure in 10^9 device-hours). It should be noted that, with a standard deviation of 1.5, the projected failure rate would be 2 FITs and with $s = 2$, the failure rate is 30 FITs, i.e., the difference between $s = 1$ and $s = 2$ is more than 3 orders of magnitude in failure rate. Thus, an accurate knowledge of the standard deviation is vital to the accurate forecasting of failure rates. However, the low confidence levels of the statistics do not justify quibbling over the real value of s, and the predicted failure rates are small in any case (but do not include infant mortality).

7.2 Case II: Au-Al phase formation dominant

As mentioned in Section 4.1, the ML due to Au-Al phase formation at 250°C has been observed to be as short as 94 hours, though it exceeds observation times in many cases. Based on this shortest observed ML and the smallest observed activation energy of 0.5 eV, a worst-case prediction is obtained, indicating that for an unscreened product the failure rate could go over 10,000 FITs, i.e., 1 percent per 1000 device-hours. However, a reliability qualification test of each slice can be used to assure that the ML due to the Au-Al/electromigration syndrome is no less than 500 hours at 250°C. Assuming a relatively conservative value of 0.8 eV for the activation energy (from the wide range observed of 0.5 to 1.6 eV), an ML at 60°C of 1×10^7 hours is projected. Taking $s = 1$, as found in Fig. 14, gives an estimated failure rate of 0.6 FIT in a 20-year service period. Taking $s = 1.5$, as found in occasional slices also dominated by the Au-Al failure mechanism, gives a failure prediction of 40 FITs. The latter value is considered a realistic upper limit for devices of the type shown in Fig. 2 subjected to a reliability screening procedure (and is therefore the value quoted in the abstract).

7.3 Case III: Electrolytic corrosion dominant

If unsealed, unprotected low-noise GaAs FET chips were employed in an amplifier in which the housing was not hermetically sealed, electrolytic corrosion as described in Section 4.2 would be expected. Since in some radio relay applications, the waveguide is pressurized with 5-

percent RH air, this value for the ambient will be considered for the first example. (It should be noted that 30°C/25% RH air becomes 5% RH air when heated, at constant water vapor content, to 60°C). For unprotected units in 85/85, an ML of about 3.5 hours due to electrolytic corrosion has been observed. Using the appropriate acceleration factor of 2×10^5, as described in Section 6.2, an ML of 7×10^5 hours at 60°C/5% RH is predicted. The corresponding failure rate is about 1000 FITs. For a second example, an atmosphere of 60°C/25% RH is chosen, which may be obtained by heating 32°C (90°F)/100% RH air up to 60°C. In this case, an ML of 3.5×10^4 hours and a failure rate of about 20,000 FITs after 5 years are predicted.

As a third example, it might be assumed that the GaAs FET chip is unprotected and the amplifier is hermetically sealed, but not adequately free of contaminants. Indeed, in view of the large amount of surface within an amplifier and the difficulty of giving it a high-temperature vacuum bakeout, it very likely would contain dangerous amounts of residual impurities. An ML of about 2000 hours has been observed in this study with contaminated packages at 85°C. The acceleration factor relative to 60°C in this case is only 3, as discussed in Section 6.2. Thus, an ML of 6000 hours might be anticipated for this amplifier with unsealed, unprotected FETs and a first-year failure rate of over 50,000 FITs.

It is emphasized that the above three examples of electrolytic corrosion assume unsealed, unprotected (unpassivated) devices. In the case of a clean, hermetically sealed device, electrolytic corrosion is effectively prevented, and no failures due to that mechanism are expected.

VIII. CONCLUSIONS

Two catastrophic failure mechanisms were found in this study of low-noise GaAs FETs, not including voltage transients which are considered primarily a problem of handling technique and circuit design. One of these mechanisms is Au-Al phase formation occurring at the junction of the Al gate and its Au bonding pad. This mechanism is enhanced by bias through what appears to be electromigration, though positive evidence of the latter is lacking. In a worst case, this mechanism could give rise to failure rates as high as 10,000 FITs, though with appropriate slice screening, values in the neighborhood of 1 to 50 FITs appear more likely. Proper design and fabrication methods can eliminate this mechanism entirely. The other catastrophic failure mechanism is electrolytic corrosion of the Al. In a humid environment or in a contaminated package, failure rates again in the order of 10,000 FITs might be anticipated. However, hermetic sealing in a contaminant-free package eliminates this problem. It is noted that both these failure mechanisms are related to the choice of an Al gate. They are not peculiar to GaAs FETs, but are well known as causes of failure in Si devices.

In the absence of catastrophic failure, a long-term, gradual degradation of noise figure and gain is observed. This effect is only weakly correlated with increase of contact resistance and is apparently more strongly influenced by other factors such as the formation of traps and scattering centers. The median lifetime due to this gradual RF degradation is estimated to be over 10^7 hours at a channel temperature of 60°C. The corresponding failure rate after 20 years of service is less than 2 FITs.

All the important failure modes were accelerated by the presence of drain and gate bias. Aging without bias would give erroneously optimistic predictions.

IX. ACKNOWLEDGMENTS

The authors are indebted to many colleagues whose contributions significantly aided this work. Special thanks are due to J. P. Beccone, W. L. Boughton, J. V. DiLorenzo, A. T. English, D. E. Iglesias, L. C. Luther, F. M. Magalhaes, W. C. Niehaus, Mrs. Y. C. Nielsen, R. H. Saul, and W. O. Schlosser.

REFERENCES

1. B. S. Hewitt, H. M. Cox, H. Fukui, J. V. DiLorenzo, W. O. Schlosser, and D. E. Iglesias, "Low Noise GaAs MESFETs: Fabrication and Performance," 1977 GaAs and Related Compounds (Edinburgh), 1976 (Inst. Phys. Conf. Ser. 33a), p. 246.
2. R. H. Knerr and C. B. Swan, "A Low-Noise GaAs FET Amplifier for 4 GHz Radio," B.S.T.J., *57*, No. 3 (March 1978), p. 479.
3. B. Selikson, "Failure Mechanisms in Integrated Circuit Interconnect Systems," 6th Annual Proc. Rel. Phys. Symp. (IEEE) (1968), p. 201.
4. E. Philofsky, "Purple Plague Revisited," Solid-State Electronics, *13* (October 1970), p. 1391.
5. I. A. Blech and E. S. Meieran, "Electromigration in Thin Al Films," J. Appl. Phys., *40* (February 1969), p. 485.
6. C. Weaver and L. C. Brown, "Diffusion in Evaporated Films of Au-Al," The Phil. Mag., *7* (1961), p. 1.
7. B. Reich and E. B. Hamkim, "Environmental Factors Governing Field Reliability of Plastic Transistors and Integrated Circuits," 10th Annual Proc. Rel. Phys. Symp. (IEEE) (1972), p. 82.
8. D. S. Peck and C. H. Zierdt, Jr., "Temperature-Humidity Acceleration of Metal-Electrolysis Failure in Semiconductor Devices," 11th Annual Proc. Rel. Phys. Symp. (IEEE) (1973), p. 149.
9. N. L. Sbar and R. P. Kozakiewicz, "New Accleration Factors for Temperature, Humidity, Bias Testing," to appear in 16th Annual Proc. Rel. Phys. Symp. (IEEE) (1978).
10. A. Shumka and R. R. Piety, "Migrated-Gold Resistive Shorts in Microcircuits," 13th Annual Proc. Rel. Phys. Symp. (IEEE) (1975), p. 93.
11. T. Irie, I. Nagasako, H. Kohza, and K. Sekido, "Reliability Study of GaAs MESFETs," IEEE Trans. on Microwave Th. and Tech., *MTT-24* (June 1976), p. 321.
12. K. Ohata and M. Ogawa, "Degradation of Au-Ge Ohmic Contact to n-GaAs," 12th Annual Proc. Rel. Phys. Symp. (IEEE) (1974), p. 278.
13. A. Christou and K. Sleger, "Precipitation and Solid Phase Formation in Au(Ag)/Ge Based Ohmic Contacts for GaAs FETS," 6th Biennial Conf. on Active Microwave Semiconductor Devices and Circuits, Cornell, 1977.
14. L. C. Kimerling, "New Developments in Defect Studies in Semiconductors," IEEE Trans. on Nuclear Sci., *NS-23* (1976), p. 1497.
15. J. C. Irvin and R. L. Pritchett, "Nonohmic Contacts for Microwave Devices," Proc. IEEE (Corres.), *58* (November 1970), p. 1845.
16. W. C. Ballamy and L. C. Kimerling, "Premature Failure in Pt-GaAs IMPATTS-Recombination Assisted Diffusion as a Failure Mechanism," Tech. Digest IEDM (IEEE) (1977), pp. 90–92.

17. D. A. Abbott and J. A. Turner, "Some Aspects of GaAs MESFET Reliability," IEEE Trans. on Microwave Th. and Tech. *M-24* (June 1976), p. 317.
18. L. R. Goldthwaite, "Failure Rate Study for the Log-Normal Lifetime Model," Proc. 7th Nat'l. Symp. on Reliability and Quality Control, *208* (January 1961). [This curve was reprinted in the 9th Annual Proc. Rel. Phys. Symp. (IEEE) (1971), p. 78].

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 8, October 1978
Printed in U.S.A.

Estimation of Point-to-Point Telephone Traffic

By J. P. MORELAND

(Manuscript received February 17, 1978)

Estimates of point-to-point telephone traffic are required for the current and the long-range planning of the Bell System's Public Switched Network. Because of the potentially immense volume of data which must be processed, these estimates are typically based upon small samples of total traffic and, therefore, can have large statistical errors. In this paper, we develop a model for quantifying the accuracy of point-to-point traffic measurements as a function of sample size and traffic parameters. Together with a worth-of-data model, not described here, our results can be used to establish a cost-optimal sampling rate for point-to-point traffic measurement systems. However, our results have been used to establish 20 percent as an upper bound on a cost-optimal sampling rate for a usage measurement system and 10 percent for an attempt-only measurement system. We show, however, that the attempt-based estimate is, for sampling rates greater than about 2 percent, less accurate than the usage-based estimate. We also show how the accuracy of point-to-point load estimates can be improved by employing a ratio-estimate which combines point-to-point and trunk-group measurements; however, in practical applications, we find that the improvement is not significant.

I. INTRODUCTION

Trunk-group and point-to-point traffic data systems provide the measurements of telephone traffic which are used for the current and the long-range planning of the Bell System's Public Switched Network. Trunk-group data systems provide estimates of the traffic offered to existing trunk groups. Normally, an estimate of trunk-group offered load is based upon a direct measurement of the average number of busy trunks, the average attempt count, and the average overflow count.[1]

Point-to-point traffic data systems provide estimates of the telephone traffic which originates at one and terminates at the other of a specific pair of network points not necessarily joined by a single trunk group; for

example, the end-office pair (A_1,B_1) of Fig. 1. In the trunk-provisioning process, estimates of point-to-point offered loads are required to plan for the introduction of new trunk groups and the rehoming of end-offices or tandems. In general, they are also used, as a supplement to trunk-group measurements, in the network disassembly process (the process that converts measured loads on trunk groups which receive overflow traffic to first-route loads) and in the network assembly process (the process which converts projected first-route loads to total offered loads). Moreover, with the possible introduction of dynamic traffic routing, our studies have shown that the trunk-provisioning process will require more extensive use of point-to-point data than is required in the present hierarchical fixed-routing network.

Estimates of point-to-point offered loads cannot, in general, be derived from trunk group measurements since trunk groups typically carry more than one point-to-point load. Instead, estimates of point-to-point offered loads are derived from detailed records of the origin, destination, and, when available, holding times of individual calls. (When holding times are not available, a load estimate can be based upon an attempt count measurement together with an exogenous estimate of mean holding time; see Section 3.2.)

To reduce the costs for recording and processing point-to-point data, most existing measurement systems have been designed to record only a small sample of total traffic. For example, the Centralized Message Data System (CMDS, see Section II) provides estimates of point-to-point loads derived from a 5-percent sample of all toll calls. But while sampling reduces the cost of providing point-to-point data, it also introduces statistical measurement errors that reduce the accuracy and, hence, the worth of the data.

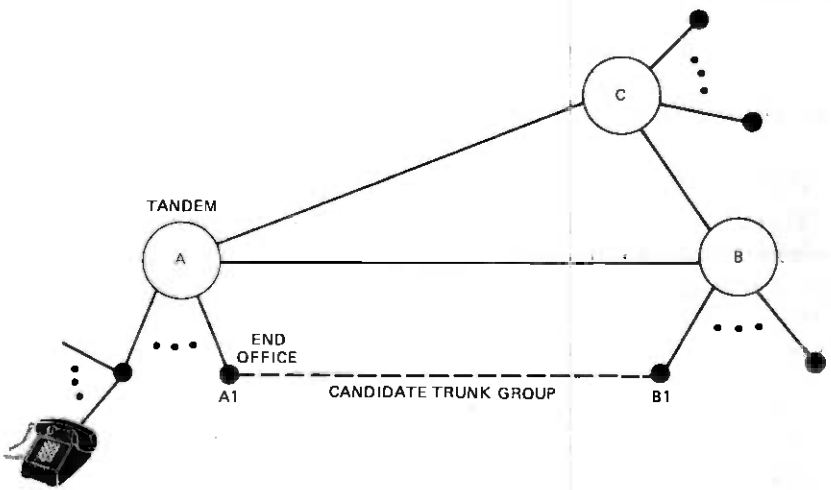


Fig. 1—An application of point-to-point data: planning new trunk groups.

In this paper, we develop a model for quantifying the accuracy of point-to-point traffic measurements as a function of sample size and traffic parameters. Together with a worth-of-data model[2] for quantifying the cost impact of data errors on the network provisioning process, this data accuracy model can be used to establish the trunk-engineering requirements for point-to-point data systems.

In Section II, we describe how point-to-point loads are measured by CMDS and we develop a model for quantifying sampling error. While this is a specific example, the methods and results are directly applicable to other (existing and proposed) point-to-point traffic measurement systems. In Section III we use our model to analyze three methods for estimating point-to-point loads: one based upon a usage measurement, one upon an attempt count together with an exogenous estimate of holding time, and one upon a combination of point-to-point and trunk-group measurements. A summary is given in the last section, and the required statistical results are developed in Appendices A and B.

II. POINT-TO-POINT MEASUREMENTS

For toll traffic, the major source of point-to-point data is provided by the Centralized Message Data System. In this section, we describe the CMDS data base and model the various sources of error.

2.1 The CMDS data base

For every point-to-point traffic item (defined by originating and terminating end-office prefix codes), the CMDS data base provides an *estimate* of both the total number of calls and the associated usage (i.e., sum of holding times) for calls that originate during a time-consistent hour over 20 consecutive business days.

These estimates are based upon a 5-percent sample of the total number of calls processed by the toll billing equipment in each Regional Accounting Office (RAO). Figure 2 illustrates the process. Automatic Message Accounting (AMA) tapes are periodically shipped to a Regional Accounting Office where they are processed to produce sequential records of the origin, destination, and conversation time of individual calls. As these records are processed for customer billing, the record for every 20th call is transmitted (in a batch mode) to the CMDS computer in Kansas City, where they are sorted and summarized to provide estimates of individual point-to-point loads.

2.2 Sources of error

Since estimates of point-to-point offered loads are based upon measurements made over several time-consistent hours, and since source loads are known to vary from day to day, our model will account for statistical errors due to both the finite measurement interval and day-

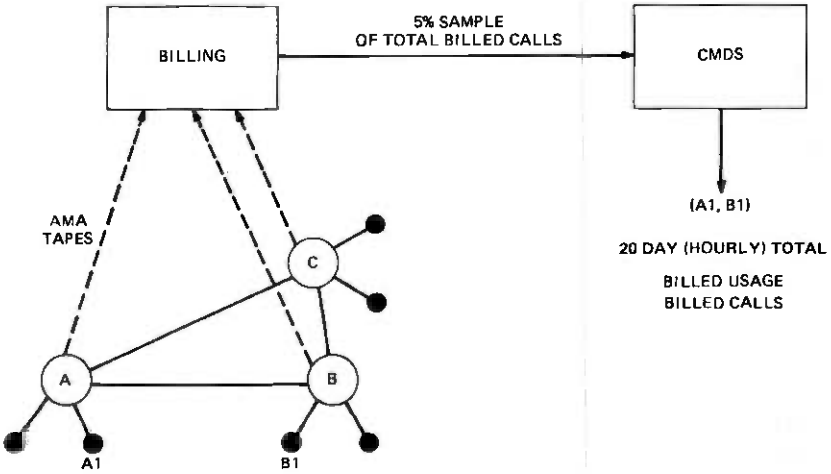


Fig. 2—Centralized message data system.

to-day load variation.[3] Furthermore, since the measurements are obtained from a 5-percent sample of total traffic, our model will also account for variations in the sample size for individual point-pairs. That is, depending upon the position of calls in the sequence of message records (from which the 5-percent sample is obtained) the actual sample size for an individual point-pair can be more, or less, than 5 percent. In Section 2.3, we develop a model for quantifying these sources of error.

The CMDS data base excludes toll traffic which is not billed. In addition, of course, to blocked calls, CMDS also excludes call set-up and ringing time (for both completed and noncompleted calls), directory assistance calls, and official calls which are not detailed billed. Estimates of this nonbilled usage are, therefore, an additional source of error for CMDS-based load estimates. However, our studies have shown that this error is negligible in comparison with sampling error and, therefore, it will not be accounted for by our model. (Section 3.2 describes a method for estimating nonbilled usage.)

2.3 Mathematical model

Estimates of point-to-point offered loads are normally based upon measurements made over K disjoint time-consistent intervals $I_1, \ldots, I_K$, each of length t (typically, $K = 20$ and $t = 1$ hour). We assume that the distribution of realized loads can be described by the model used by Hill and Neal[3] to explain the observed variation of trunk-group offered loads. Thus, during I_j, we assume that call arrivals are Poisson-distributed* with rate λ_j and that call-holding times are independent and exponen-

* Point-to-point offered loads correspond to trunk-group first-offered (Poisson) loads; hence, it is appropriate to set the peakedness factor, z, of Ref. 3 to unity.

tially distributed with mean h. Furthermore, in accordance with the model for day-to-day load variation developed in Ref. 3, the loads $\alpha_i = \lambda_i h$, $i = 1,\ldots,K$, are assumed to be independent and identically distributed with mean $a = \lambda h$ and variance

$$v_d = \max\left\{0, 0.13a^{\phi} - \frac{2a}{t/h}\right\}, \tag{1}$$

where ϕ is a parameter that describes the level of day-to-day variation. For engineering applications, we use $\phi = 1.5$, 1.7, or 1.84, which are referred to, respectively, as low, medium, or high day-to-day variation. For first-routed and point-to-point traffic, $\phi = 1.5$ is usually appropriate.

To model the sampling process, we assume that in the sequence of message records, each call associated with a given point-pair is included in the sample with the same probability p (for CMDS, $p = 0.05$); i.e., we assume a multinomial distribution for the numbers of sampled calls belonging to given point-pairs. (For CMDS, the actual distribution is more closely approximated by a hypergeometric distribution; however, since the number of calls belonging to a given point-pair is a small fraction of the total number of calls processed by an RAO, our simplifying assumption introduces no significant loss of accuracy.)

Let N_j denote the number of arrivals during I_j and let h_{ij} be the holding time of the ith arrival in I_j. Then, with $\delta_{ij} = 1$ if the ith call is included in the sample and zero otherwise,

$$c = \sum_{j=1}^{K} \sum_{i=1}^{N_j} \delta_{ij} \tag{2}$$

is the total number of sampled calls during $I = \Sigma_{j=1}^{K} I_j$, and

$$u = \sum_{j=1}^{K} \sum_{i=1}^{N_j} h_{ij}\delta_{ij} \tag{3}$$

is the corresponding usage.

III. LOAD ESTIMATES

In this section, we analyze three procedures for estimating point-to-point loads. The first estimate, $\hat{a}^{(1)}$, is based upon the usage measurement, u; the second estimate, $\hat{a}^{(2)}$, upon the attempt count, c; and the third estimate, $\hat{a}^{(3)}$, upon a combination of point-to-point and trunk-group measurements. (Although these do not exhaust the possible estimates, they do form the basis for analyzing more complex estimates; for example, an estimate of the offered load at 10 a.m. could be based upon a combination of the measured loads at 9, 10, and 11 a.m.) In each case, we use mean square error (MSE) to measure the accuracy of the load estimate, i.e., if $\hat{a}$ denotes an estimate of the mean offered load, then

$$\begin{aligned} \text{MSE}\{\hat{a}\} &= E\{\hat{a} - a\}^2 \\ &= \text{Var}\{\hat{a}\} + E^2\{\hat{a} - a\}. \end{aligned} \tag{4}$$

3.1 Estimate 1

Since u [eq. (3)] is a p-sample of usage over K intervals each of length t,

$$\hat{a}^{(1)} = \frac{1}{Kpt} u \tag{5}$$

is an estimate of the corresponding average offered load a.

In Appendix A, we show that $\hat{a}^{(1)}$ is unbiased, i.e.,

$$E\{\hat{a}^{(1)}\} = a, \tag{6}$$

and has variance

$$\text{Var}\{\hat{a}^{(1)}\} = \frac{1}{K}\left\{\frac{2a}{pt/h} + v_d\right\}, \tag{7}$$

hence, from eq. (4),

$$\text{MSE}\{\hat{a}^{(1)}\} = \frac{1}{K}\left\{\frac{2a}{pt/h} + v_d\right\}. \tag{8}$$

In (8), the first term $\{2a/pt/h\}$ represents the combined effects of the finite measurement interval and deviations from the average sample size. The second term $\{v_d\}$ is due to (day-to-day) variations in the source load. Of course, the factor K is due to averaging measurements over K independent intervals.

Figure 3 displays the root-mean-square (RMS) error of $\hat{a}^{(1)}$ (in percent of mean load) as a function of average offered load for sampling rates of 5 and 100 percent. The results for a 5-percent sample apply when the offered load is estimated using CMDS data, while those for a 100-percent

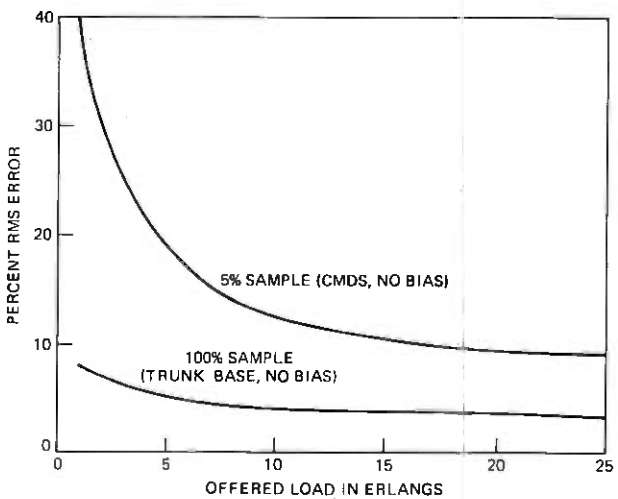


Fig. 3—Sampling error vs offered load.

sample apply when the load estimate is based directly upon trunk-group measurements. As noted in Section 2.2, errors in estimates of nonbilled usage, associated with CMDS estimates, are negligible in comparison with sampling error. Also, in applying our results to trunk-group measurements, we assume that u [eq. (3)] adequately approximates the actual usage *during* the measurement interval I; i.e., we assume that the edge effects are negligible (see Ref. 3). Furthermore, our studies have shown that the additional variance caused by discretely sampling the usage with a 100-second-scan Traffic Usage Recorder is negligible when compared with the variance caused by day-to-day load variation. The results shown in Fig. 3 assume the standard measurement interval (t = 1 hour, K = 20), low day-to-day variation (ϕ = 1.5), and h = 250 seconds.

Note that estimates based upon a 5-percent sample can have errors that are large relative to those based upon a 100-percent sample. For example, for an offered load of 5 erlangs (typical of base year prove-in loads for new high-usage trunk groups), the RMS error for a 5-percent sample is about 20 percent, compared with an RMS error of about 5 percent for a 100-percent sample. Similarly, for an offered load of about 15 erlangs (typical of loads offered to existing Long Lines high-usage trunk groups), an estimate based upon a 5-percent sample has an RMS error of about 12 percent, while for a 100-percent sample, the RMS error is about 4 percent.

Figure 4 displays the percent RMS error of $\hat{a}^{(1)}$ as a function of the sampling rate p for offered loads of 5 and 15 erlangs. The important result to note is that the statistical variability of $\hat{a}^{(1)}$ does not decrease appreciably as the sampling rate is increased beyond about 20 percent. This occurs since the contribution of day-to-day load variation is independent of the sampling rate, and above a sampling rate of about 20

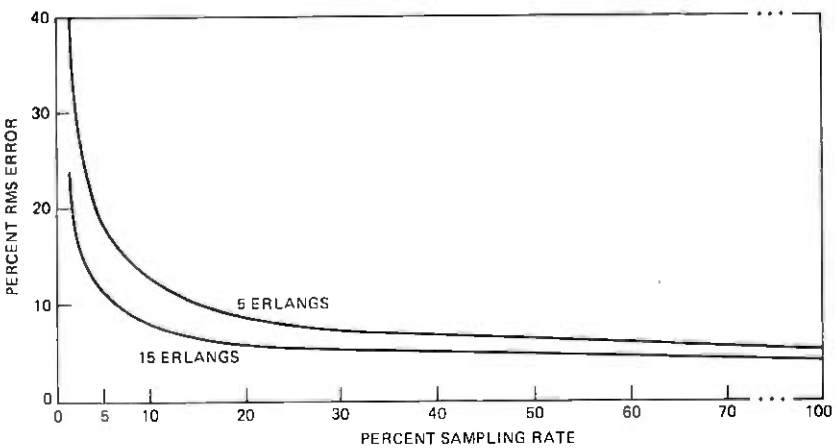


Fig. 4—Sampling error vs sampling rate.

percent it becomes the dominant source of error. Of course, any improvement in accuracy is significant if the associated benefits justify the increased cost for data collection and processing. However, using the worth-of-data model of Ref. 2, which quantifies the cost impact of data errors on the provisioning of direct final (i.e., nonalternate route) trunk groups, we have established 20 percent as an *upper bound* on a cost-optimal sampling rate. To establish an actual cost-optimal sampling rate, however, we require a worth-of-data model which applies to more general network configurations (i.e., alternate routing networks); such a model is currently being formulated.

3.2 Estimate 2

The usage-based estimate of offered load (i.e., Estimate 1) is derived from attempt count and holding-time measurements. In this section, we analyze an alternative estimate based upon an attempt count together with an exogenous estimate of the corresponding mean holding time. The data collection and processing costs for an attempt-based point-to-point data system are less than for a usage-based system; however, we will show that the load estimates are substantially less accurate.

Let $\hat{h}$ (a constant) denote an estimate of the mean holding time h. Then, since c [eq. (2)] is the total number of sampled calls during I, c/Kpt is an estimate of the mean attempt rate λ and, therefore,

$$\hat{a}^{(2)} = \frac{1}{Kpt} c\hat{h} \tag{9}$$

is an estimate of the average offered load a.

In Appendix A we show that

$$E\{\hat{a}^{(2)}\} = \lambda\hat{h} \tag{10}$$

so that $\hat{a}^{(2)}$ is biased whenever $\hat{h} \neq h$, and

$$\text{Var}\{\hat{a}^{(2)}\} = \left(\frac{\hat{h}}{h}\right)^2 \frac{1}{K}\left\{\frac{a}{pt/h} + v_d\right\}; \tag{11}$$

hence, from (4),

$$\text{MSE}\{\hat{a}^{(2)}\} = \left(\frac{\hat{h}}{h}\right)^2 \frac{1}{K}\left\{\frac{a}{pt/h} + v_d\right\} + \lambda^2(\hat{h} - h)^2. \tag{12}$$

Clearly, $\text{MSE}\{\hat{a}^{(2)}\}$ depends upon the error $(\hat{h} - h)$. In practice, the same estimate $\hat{h}$ would be applied to a collection of point-pairs (e.g., all point-pairs within an operating company, or all point-pairs served by a common trunk group), and our studies have found that the corresponding distribution of errors $(\hat{h} - h)$ has a coefficient of variation of at least 20 percent. Accordingly, our numerical results will assume that $\hat{h}$ is in-error by 20 percent.

Figure 5 displays the percent RMS error of $\hat{a}^{(2)}$ as a function of sampling rate for an offered load of 5 erlangs. We assume the same numerical values for K, t, ϕ, and h as in Fig. 3, and we assume $\hat{h}/h = 1.2$. For purposes of comparison, Fig. 5 also displays the percent RMS error of $\hat{a}^{(1)}$, as given previously in Fig. 4.

We draw two conclusions from the results shown in Fig. 5. First, whereas 20 percent is a reasonable upper bound on sampling rate for a usage-based measurement system, a sampling rate of about 10 percent is sufficient for an attempt-based system. Of course, if $\hat{h}$ were known to be in error by more (less) than 20 percent, a sampling rate of less (more) than 10 percent would be appropriate. But if we know only the coefficient of variation of the distribution of h (which we assume to be 20 percent), then the average value of MSE$\{\hat{a}^{(2)}\}$, with respect to this distribution, cannot be significantly reduced by increasing the sampling rate beyond 10 percent. Second, we note that an estimate based upon measured usage is, for sampling rates greater than about 2 percent, more accurate than an estimate based upon an attempt count. (For sampling rates less than 2 percent, the standard deviation of the measured holding time exceeds that of the estimate $\hat{h}$; hence, $\hat{a}^{(2)}$ is relatively more accurate in this range.)

In view of the above results, we conclude that usage measurements (when available) are preferable to attempt counts for estimating point-to-point loads. However, our studies have shown that the attempt count provides a more accurate basis for estimating CMDS nonbilled usage than does the measured (billed) usage. That is, with CMDS data, an estimate of the form $\hat{a} = (u + \hat{\beta}c)/Kpt$ is employed, where the first term $\{u/Kpt\}$ is an estimate of billed load and the second term $\{\hat{\beta}c/Kpt\}$ is an estimate of nonbilled load. Thus, $\hat{\beta}$ can be interpreted as an estimate

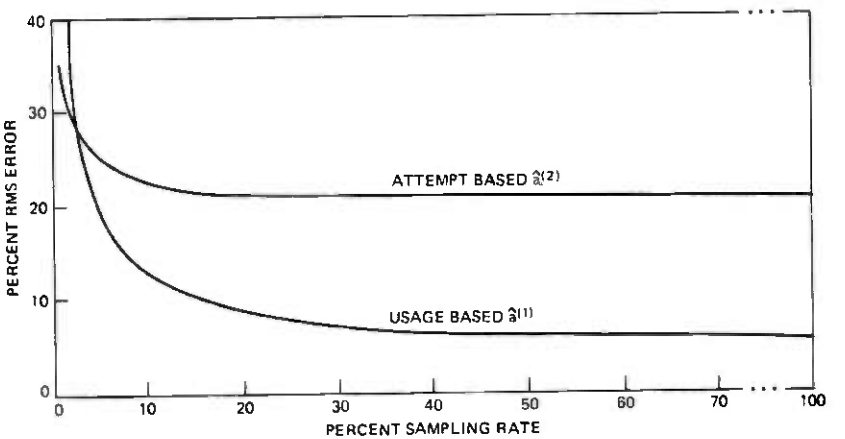


Fig. 5—Comparison of attempt-based and usage-based offered-load estimates (offered load = 5 erlangs).

of an average nonbilled holding time per billed attempt. Furthermore, we have shown that a small additional improvement can be obtained by employing a load-dependent combination of u and c to estimate billed load. However, this additional improvement is not significant.

3.3 Estimate 3

In this section, we show how the statistical variability of point-to-point load estimates can be reduced by combining point-to-point and trunk-group measurements. The procedure we describe has been proposed as a means for improving the accuracy of CMDS-based load estimates; however, we show that the improvement is not significant.

Consider a trunk group whose total offered load is the sum of N point-to-point first-offered loads. For the ith load, $i = 1,\ldots,N$, let a_i denote the mean load and let $\hat{a}_i^{(1)}$ be the (point-to-point) usage-based estimate of a_i. Furthermore, let $\hat{T}$ denote the estimate of trunk-group offered load based upon trunk-group usage measurements and let $\hat{A} = \sum_{i=1}^{N} \hat{a}_i^{(1)}$ denote the corresponding estimate (for the same measurement interval) based upon point-to-point usage data. Since $\hat{T}$ is based upon a 100-percent sample, the difference $(\hat{T} - \hat{A})$ measures the sum of the errors relative to the realized loads in the individual estimates $\hat{a}_i^{(1)}$. By assigning a fraction (w_i) of this difference to the individual estimates $\hat{a}_i^{(1)}$, we obtain a new estimate of a_i; i.e.,

$$\hat{a}_i^{(3)} = \hat{a}_i^{(1)} + w_i(\hat{T} - \hat{A}). \tag{13}$$

In Appendix B, we show that an approximation to a minimum-variance linear estimate of a_i is obtained when $w_i = \hat{a}_i^{(1)}/\hat{A}$. Thus, we have the ratio-estimate

$$\hat{a}_i^{(3)} = \frac{\hat{a}_i^{(1)}}{\hat{A}}\hat{T}. \tag{14}$$

Since $\hat{a}_i^{(1)}$ appears as a summand in $\hat{A}$, the ratio $\hat{T}/\hat{A}$ is negatively correlated (or tends to vary inversely) with $\hat{a}_i^{(1)}$. Physically, it is this negative correlation which makes $\hat{a}_i^{(3)}$ statistically less variable than $\hat{a}_i^{(1)}$.

By employing a first-order Taylor series approximation to $\hat{a}_i^{(3)}$, we obtain in Appendix A the following approximations for the mean and variance of $\hat{a}_i^{(3)}$:

$$E\{\hat{a}_i^{(3)}\} \approx a_i \tag{15}$$

and

$$\mathrm{Var}\{\hat{a}_i^{(3)}\} \approx \frac{2a_i}{Kpt/h}\{1 - f_i(1-p)\} + \frac{1}{K}v_{di}, \tag{16}$$

where

$$f_i = \frac{a_i}{\sum_{j=1}^{N} a_j} \tag{17}$$

is the fraction of the total load contributed by the ith point-pair. Also, since our results are not significantly affected by differences in mean holding times, we have assumed that each point-to-point load has the same mean holding-time, h. From (4), (15), and (16) we have

$$\text{MSE}\{\hat{a}_i^{(3)}\} \approx \frac{2a_i}{Kpt/h}\{1 - f_i(1-p)\} + \frac{1}{K}v_{di}. \tag{18}$$

Figure 6 displays the percent RMS error of $\hat{a}_i^{(3)}$ as a function of offered load for several values of the parameter f_i. We assume a sampling rate of 5 percent (the CMDS sampling rate) and the same numerical values for K, t, h, and ϕ as in Fig. 3. Note that $\hat{a}_i^{(3)}$ is more accurate than $\hat{a}_i^{(1)}$ and that the relative difference in accuracy is a maximum when f_i equals one (since $\hat{a}_i^{(3)} = \hat{T}$ when $f_i = 1$) and approaches zero as f_i approaches zero (since the variance of $\hat{T}/\hat{A}$ approaches zero and, hence, $\hat{a}_i^{(3)}$ approaches $\hat{a}_i^{(1)}$ as f_i approaches zero).

The results of Fig. 6 are perhaps more striking when viewed in terms of the reciprocal of f_i, which can be interpreted as the number (N') of equal-sized point-to-point loads corresponding to f_i. That is, the relative difference in accuracy of $\hat{a}_i^{(1)}$ and $\hat{a}_i^{(3)}$ is a rapidly decreasing function of N'; for N' greater than 4, the relative difference is less than about 4 percentage points.

Typically, trunk groups carry a large number of point-to-point loads, each of which represents a small fraction ($f_i \ll 1$) of total offered load. In this region, Fig. 6 shows that the use of trunk-group measurements provides only a small improvement in the quality of CMDS-based load estimates. Again, any improvement is significant if the associated ben-

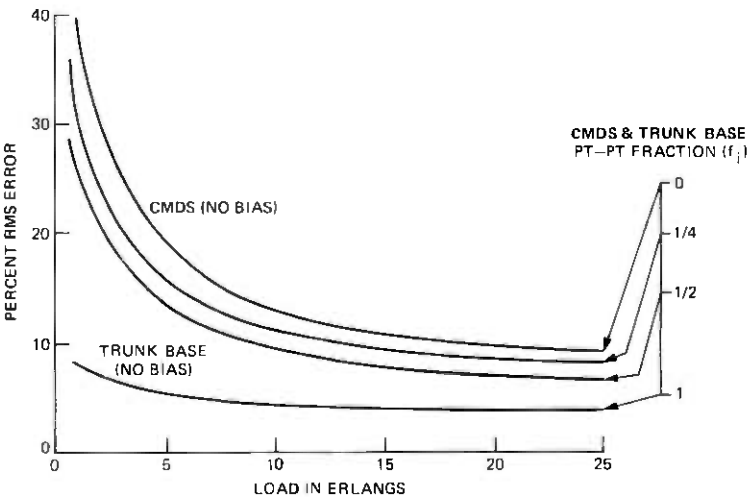


Fig. 6—Reduction in rms error afforded by combining point-to-point and trunk-base measurements.

efits justify the increased development and data processing costs. However, based upon the worth-of-data model of Ref. 2, we have concluded that employing trunk-group measurements will not significantly reduce the statistical errors associated with CMDS-based estimates of point-to-point loads.

IV. SUMMARY AND CONCLUSIONS

We have developed a model for quantifying the accuracy of point-to-point traffic measurements as a function of sampling rate and traffic parameters. Using this model, we have established 20 percent as an upper bound on a cost-optimal sampling rate for a usage-based measurement system and 10 percent for an attempt-based system. Furthermore, for sampling rates greater than a few percent and loads in the range of engineering interest, our results show that a usage-based load estimate is more accurate than an attempt-based load estimate. We also showed that the accuracy of (CMDS) load estimates could be improved by employing a ratio estimate that combines point-to-point and trunk-group measurements; however, in practical applications, the improvement is not significant. Our results, together with a worth-of-data model,[2] can be used to establish requirements for point-to-point traffic measurement systems.

APPENDIX A

Mean and Variance of Load Estimates

A.1 Estimate 1

From eqs. (3) and (5) and Ref. 4,

$$E\{\hat{a}^{(1)}\} = \frac{1}{Kpt} \sum_{j=1}^{K} E\left\{E\left\{\sum_{i=1}^{N_j} h_{ij}\delta_{ij} | N_j\right\}\right\}. \tag{19}$$

Since the h_{ij} and δ_{ij} are independent, and since arrivals during I_j are Poisson-distributed with rate λ_j, we have

$$\begin{aligned} E\left\{E\left\{\sum_{i=1}^{N_j} h_{ij}\delta_{ij} | N_j\right\}\right\} &= E\{N_j ph\} \\ &= phE\{E\{N_j | \lambda_j\}\} \\ &= phE\{\lambda_j t\} \\ &= ph\lambda t. \end{aligned} \tag{20}$$

Substituting (20) into (19) gives

$$\begin{aligned} E\{\hat{a}^{(1)}\} &= \lambda h \\ &= a. \end{aligned} \tag{21}$$

Furthermore, since the measurements during each interval I_j are uncorrelated, (3) and (5) give

$$\mathrm{Var}\{\hat{a}^{(1)}\} = \frac{1}{(Kpt)^2} \sum_{j=1}^{K} \mathrm{Var}\left\{ \sum_{i=1}^{N_j} h_{ij}\delta_{ij} \right\}. \tag{22}$$

From Ref. 4,

$$\begin{aligned}\mathrm{Var}\left\{ \sum_{i=1}^{N_j} h_{ij}\delta_{ij} \right\} &= E\left\{ \mathrm{Var}\left\{ \sum_{i=1}^{N_j} h_{ij}\delta_{ij} | N_j \right\} \right\} \\ &\quad + \mathrm{Var}\left\{ E\left\{ \sum_{i=1}^{N_j} h_{ij}\delta_{ij} | N_j \right\} \right\} \\ &= E\{N_j h^2 p(2-p)\} + \mathrm{Var}\{N_j ph\} \\ &= \lambda t h^2 p(2-p) + p^2 h^2 \mathrm{Var}\{N_j\}. \end{aligned} \tag{23}$$

Again, given λ_j, N_j is Poisson-distributed; hence, $\mathrm{Var}\{N_j | \lambda_j\} = E\{N_j | \lambda_j\} = \lambda_j t$.

Thus it follows that

$$\begin{aligned}\mathrm{Var}\{N_j\} &= E\{\mathrm{Var}\{N_j | \lambda_j\}\} \\ &\quad + \mathrm{Var}\{E\{N_j | \lambda_j\}\} \\ &= E\{\lambda_j t\} + \mathrm{Var}\{\lambda_j t\} \\ &= \lambda t + t^2 \mathrm{Var}\{\lambda_j\}. \end{aligned} \tag{24}$$

Substituting (24) into (23) gives

$$\mathrm{Var}\left\{ \sum_{i=1}^{N_j} h_{ij}\delta_{ij} \right\} = 2\lambda t h^2 p + p^2 t^2 h^2 \mathrm{Var}\{\lambda_j\}. \tag{25}$$

The offered load during I_j is $\alpha_j = \lambda_j h$; hence, $v_d = \mathrm{Var}\{\alpha_j\} = h^2 \mathrm{Var}\{\lambda_j\}$.

Thus, from (22) and (25), we have

$$\mathrm{Var}\{\hat{a}^{(1)}\} = \frac{1}{K}\left\{ \frac{2a}{pt/h} + v_d \right\}. \tag{26}$$

We now develop an expression, which we require in Section A.3, for

$$\mathrm{Cov}\{\hat{a}^{(1)}, \hat{a}^{(1)}|_{p=1}\},$$

where

$$\hat{a}^{(1)}|_{p=1} = \frac{1}{Kt} \sum_{j=1}^{K} \sum_{i=1}^{N_j} h_{ij} \tag{27}$$

corresponds to a sampling rate of 100 percent. Thus, from (5) and (27), we have

$$\mathrm{Cov}\{\hat{a}^{(1)}, \hat{a}^{(1)}|_{p=1}\} = \frac{1}{K^2 t^2 p} \sum_{j=1}^{K} \mathrm{Cov}\left\{ \sum_{i=1}^{N_j} h_{ij}\delta_{ij}, \sum_{i=1}^{N_j} h_{ij} \right\}. \tag{28}$$

From Ref. 4,

$$\mathrm{Cov}\left\{ \sum_{i=1}^{N_j} h_{ij}\delta_{ij}, \sum_{i=1}^{N_j} h_{ij} \right\} = E\left\{ \mathrm{Cov}\left\{ \sum_{i=1}^{N_j} h_{ij}\delta_{ij} | N_j, \sum_{i=1}^{N_j} h_{ij} | N_j \right\} \right\}$$

$$+ \operatorname{Cov}\left\{E\left\{\sum_{i=1}^{N_j} h_{ij}\delta_{ij}|N_j\right\}, E\left\{\sum_{i=1}^{N_j} h_{ij}|N_j\right\}\right\}. \quad (29)$$

We first expand

$$\begin{aligned}
&\operatorname{Cov}\left\{\sum_{i=1}^{N_j} h_{ij}\delta_{ij}|N_j, \sum_{i=1}^{N_j} h_{ij}|N_j\right\} \\
&\triangleq E\left\{\left(\sum_{i=1}^{N_j} h_{ij}\delta_{ij}|N_j\right)\left(\sum_{i=1}^{N_j} h_{ij}|N_j\right)\right\} - E\left\{\sum_{i=1}^{N_j} h_{ij}\delta_{ij}|N_j\right\} E\left\{\sum_{i=1}^{N_j} h_{ij}|N_j\right\} \\
&\quad = N_j\{p2h^2 + (N_j - 1)ph^2\} - N_j^2 ph^2 \\
&\quad = N_j ph^2. \quad (30)
\end{aligned}$$

Substituting (30) into (29) and using (24) gives

$$\begin{aligned}
\operatorname{Cov}\left\{\sum_{i=1}^{N_j} h_{ij}\delta_{ij}, \sum_{i=1}^{N_j} h_{ij}\right\} &= ph^2 E\{N_j\} + \operatorname{Cov}\{N_j ph, N_j h\} \\
&= ph^2 E\{N_j\} + ph^2 \operatorname{Var}\{N_j\} \\
&= 2phat + pt^2 v_d. \quad (31)
\end{aligned}$$

Thus, from (28) and (31), we have

$$\operatorname{Cov}\{\hat{a}^{(1)}, \hat{a}^{(1)}|_{p=1}\} = \frac{1}{K}\left\{\frac{2a}{t/h} + v_d\right\}. \quad (32)$$

A.2 Estimate 2

Using expansions similar to those of A.1 it follows, from (2) and (9), that

$$\begin{aligned}
E\{\hat{a}^{(2)}\} &= \frac{\hat{h}}{Kpt}\sum_{j=1}^{K} E\left\{E\left\{\sum_{i=1}^{N_j}\delta_{ij}|N_j\right\}\right\} \\
&= \hat{h}\lambda \quad (33)
\end{aligned}$$

and

$$\begin{aligned}
\operatorname{Var}\{\hat{a}^{(2)}\} &= \left(\frac{\hat{h}}{Kpt}\right)^2 \sum_{j=1}^{K} \operatorname{Var}\left\{\sum_{i=1}^{N_j}\delta_{ij}\right\} \\
&= \left(\frac{\hat{h}}{h}\right)^2 \frac{1}{K}\left\{\frac{a}{pt/h} + v_d\right\}. \quad (34)
\end{aligned}$$

A.3 Estimate 3

An approximation for the mean and variance of $\hat{a}_i^{\{3\}}$ is obtained by expanding the right-hand side of eq. (14) in a three-dimensional Taylor series about the point $\{E\{\hat{T}\}, E\{\hat{A}\}, E\{\hat{a}_i^{(1)}\}\}$. To first order, this gives the approximations

$$E\{\hat{a}_i^{\{3\}}\} \approx \frac{E\{\hat{T}\}}{E\{\hat{A}\}} E\{\hat{a}_i^{(1)}\} \quad (35)$$

and

$$\mathrm{Var}\{\hat{a}_i^{(3)}\} \approx \mathrm{Var}\{\hat{a}_i^{(1)}\} + f_i^2\mathrm{Var}\{\hat{A}\} + f_i^2\mathrm{Var}\{\hat{T}\} - 2f_i\mathrm{Cov}\{\hat{a}_i^{(1)}, \hat{A}\} + 2f_i\mathrm{Cov}\{\hat{a}_i^{(1)}, \hat{T}\} - 2f_i^2\mathrm{Cov}\{\hat{A}, \hat{T}\}, \quad (36)$$

where

$$f_i = \frac{a_i}{\sum_{j=1}^{N} a_j}. \quad (37)$$

Since the trunk-group measurement $\hat{T}$ corresponds to a sampling rate of 100 percent (i.e., $p = 1$), we have

$$\hat{T} = \hat{A}|_{p=1} = \sum_{i=1}^{N} \hat{a}_i^{(1)}|_{p=1}. \quad (38)$$

Hence, from (21), (35), and (38), it follows that

$$E\{\hat{a}_i^{(3)}\} \approx a_i. \quad (39)$$

We assume that the daily source loads (for different point-pairs) are uncorrelated;* hence, the estimates $\hat{a}_i^{(1)}$ are uncorrelated. Furthermore, since our results are not significantly affected by differences in the mean holding times, we assume that each point-to-point load has the same mean holding time, h. Thus, we have

$$\mathrm{Var}\{\hat{A}\} = \sum_{i=1}^{N} \mathrm{Var}\{\hat{a}_i^{(1)}\} = \frac{1}{K}\sum_{i=1}^{N}\left\{\frac{2a_i}{pt/h} + v_{di}\right\}, \quad (41)$$

$$\mathrm{Var}\{\hat{T}\} = \mathrm{Var}\{\hat{A}|_{p=1}\} = \frac{1}{K}\sum_{i=1}^{N}\left\{\frac{2a_i}{t/h} + v_{di}\right\}, \quad (42)$$

and

$$\mathrm{Cov}\{\hat{a}_i^{(1)}, \hat{A}\} = \mathrm{Var}\{\hat{a}_i^{(1)}\} = \frac{1}{K}\left\{\frac{2a_i}{pt/h} + v_{di}\right\}. \quad (43)$$

Also, from (32) and (38),

$$\mathrm{Cov}\{\hat{a}_i^{(1)}, \hat{T}\} = \mathrm{Cov}\{\hat{a}_i^{(1)}, \hat{a}_i^{(1)}|_{p=1}\} = \frac{1}{K}\left\{\frac{2a_i}{t/h} + v_{di}\right\} \quad (44)$$

* We have shown that our results are independent of the covariance structure of the daily source loads; for simplicity, we assume that they are uncorrelated.

and

$$\text{Cov}\{\hat{A},\hat{T}\} = \sum_{i=1}^{N} \text{Cov}\{\hat{a}_i^{(1)}, \hat{T}\}$$

$$= \frac{1}{K}\sum_{i=1}^{N}\left\{\frac{2a_i}{t/h} + v_{di}\right\}. \tag{45}$$

Combining (26), (36), and (41) through (45) gives

$$\text{Var}\{\hat{a}_i^{(3)}\} \approx \frac{1}{K}\left\{\frac{2a_i}{pt/h} + v_{di}\right\} + \frac{2(1-p)}{Kpt/h} f_i^2 \sum_{j=1}^{N} a_j - \frac{4(1-p)}{Kpt/h} f_i a_i$$

or, since $f_i \sum_{j=1}^{N} a_j = a_i$,

$$\text{Var}\{\hat{a}_i^{(3)}\} \approx \frac{2a_i}{Kpt/h}\{1 - f_i(1-p)\} + \frac{1}{K}v_{di}. \tag{46}$$

APPENDIX B

Minimum Variance Estimate

In this appendix, we show that Estimate 3 can be obtained as an approximation to a minimum variance linear estimate. Thus, from eq. (13)

$$\hat{a}_i^{(3)} = \hat{a}_i^{(1)} + w_i\{\hat{T} - \hat{A}\}. \tag{47}$$

This estimate is unbiased, i.e., $E\{\hat{a}_i^{(3)}\} = E\{\hat{a}_i^{(1)}\} = a_i$, and has variance

$$\text{Var}\{\hat{a}_i^{(3)}\} = \text{Var}\{\hat{a}_i^{(1)}\} + 2w_i\text{Cov}\{\hat{a}_i^{(1)}, \hat{T} - \hat{A}\} + w_i^2\text{Var}\{\hat{T} - \hat{A}\}. \tag{48}$$

The value of w_i which minimizes the variance satisfies the equation

$$\frac{\partial\, \text{Var}\{\hat{a}_i^{(3)}\}}{\partial w_i} = 0, \tag{49}$$

which implies that

$$w_i = \frac{\text{Cov}\{\hat{a}_i^{(1)}, \hat{A} - \hat{T}\}}{\text{Var}\{\hat{T} - \hat{A}\}}. \tag{50}$$

From Appendix A, it follows that

$$w_i = \frac{a_i}{\sum_{j=1}^{N} a_j}. \tag{51}$$

Now if a_i is estimated by $\hat{a}_i^{(1)}$ so that w_i is estimated by $\hat{a}_i^{(1)}/\hat{A}$, eq. (47) becomes

$$\hat{a}_i^{(3)} = \frac{\hat{a}_i^{(1)}}{\hat{A}}\hat{T}. \tag{52}$$

Q.E.D.

REFERENCES

1. S. R. Neal and A. Kuczura, "A Theory of Traffic-Measurement Errors for Loss Systems with Renewal Input," B.S.T.J., *52*, No. 6 (July-August 1973), pp. 967–990.
2. R. L. Franks, H. Heffes, J. M. Holtzman, and S. Horing, "A Model Relating Measurement and Forecast Errors to the Provisioning of Direct Final Trunk Groups," Proceedings of the 8th International Teletraffic Congress, Melbourne, Australia, November 1976.
3. D. W. Hill and S. R. Neal, "Traffic Capacity of a Probability-Engineered Trunk Group," B.S.T.J., *55*, No. 7 (September 1976), pp. 831–842.
4. H. H. Panjer, "On the Decomposition of Moments by Conditional Moments," Amer. Statist., *27* (October 1973), pp. 170–171.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 8, October 1978
Printed in U.S.A.

Buffering of Slow Terminals

By A. G. FRASER, B. GOPINATH, and J. A. MORRISON

(Manuscript received January 23, 1978)

In a previous paper, a model for queuing processes with correlated inputs was analyzed. To illustrate the application of those results, we model a concentrator of slow terminals. The sources of data, the terminals, generate data at a slower rate than the output speed of the buffer. In special cases, we obtain closed-form expressions for the generating function of the equilibrium queue size distribution. For the general case, we describe a computational procedure to obtain the distribution and the average of queue size in steady state. The numerical results obtained using this procedure are presented for a family of problems in which each message consists of two packets separated by a fixed time interval.

I. INTRODUCTION

A data communications network may be constructed by connecting together terminals and switching nodes so that each terminal is connected to just one node and the nodes are connected together in a more-or-less redundant fashion. All connections are by means of transmission lines. Those between terminals and nodes are called access lines, while those between one node and another are called trunk lines (Fig. 1). For economy, the access lines commonly have smaller bandwidth than the trunk lines.

The character of the traffic carried by a data network depends in part upon the type of terminal connected to it. Keyboard and display terminals transmit and receive data messages that are typically less than a few hundred characters in length. These terminals operate at speeds up to 1.2 Kb/s. Batch stations transmit and receive data in larger quantities and typically operate at speeds up to 9.6 Kb/s. Trunk lines, and the connections to computers, typically operate at about 50 Kb/s. Thus, we find that traffic in a data network is not uniformly distributed either among the terminals or among the various transmission lines.

Analysis of delay and the probability of queue overflow is most simply

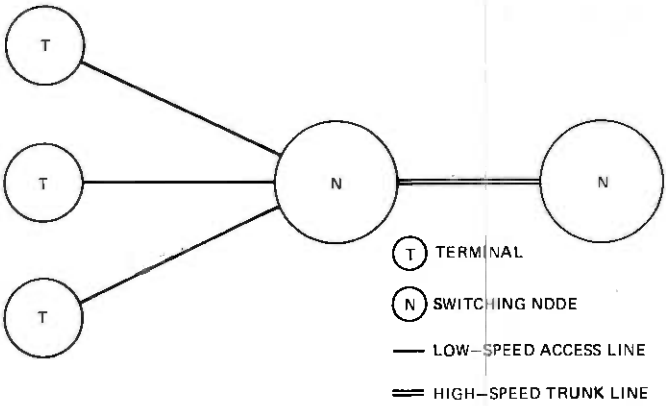


Fig. 1—Network topology.

obtained by assuming that all traffic is uniformly distributed between the terminals and that all transmission lines operate at the same speed. Using these assumptions, Chu[1,2] has studied two cases. In one case, equal size packets (individual characters perhaps) are generated randomly by the terminals. In another case, messages are generated at random, but each message consists of a random number of packets which all enter the network in one instant. The case of mixed input traffic was studied by Chu and Liang.[3]

Consider the situation when terminals randomly generate messages consisting of several fixed-size packets. The data are fed into a switching node over access lines that are substantially slower than the trunk lines used to carry data out of the node. Several packets may be transmitted on a trunk line in the time that it takes to transmit one packet on an access line (Fig. 2). Thus, packets of one message which are transmitted consecutively by a terminal will arrive periodically at the node, and the period will be greater than the period of packet transmissions on the

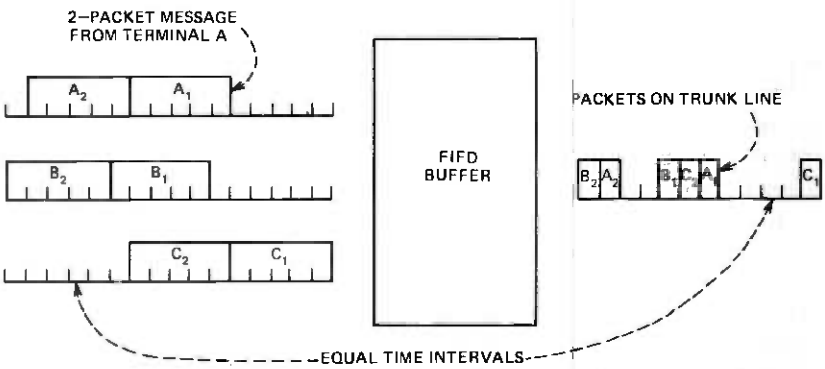


Fig. 2—Example of packet flow at a node.

trunk line. Given that packet arrivals are not entirely random, one would expect the buffer storage requirements in the switch to be less than is indicated by Chu and Liang's analyses.

In this paper, we study the behavior of a switching node that receives data from a (large) number of low-speed access lines. The data are received in the form of packets of a fixed size. As the packets arrive, they are placed in a buffer, which is a first-in-first-out queue. In an actual communications network, the buffer has a finite size, and a packet is lost if the buffer is full when it attempts to enter it. The buffer transmits packets at a uniform rate onto a high-speed trunk, provided the buffer is not empty. A crucial question is how large the buffer should be in order that the probability of packet loss should be less than 10^{-4}, say. In the mathematical analysis of the single queue corresponding to this buffer, we consider a buffer of unlimited size so that no overflow is possible, and we calculate the steady-state probability that the buffer content (i.e., the number of packets in the buffer, or queue size) exceeds the proposed size of the finite buffer. We refer to this quantity as the probability of overflow, since it is usually used to estimate the actual overflow probability, when the probability of packet loss is very small.

Some of the symbols used are listed in Section II. The mathematical model, which includes assumptions concerning the packet arrivals, is discussed in Section III. This model was analyzed by two of the authors,[4] and formulas for calculating the equilibrium queue size distribution are summarized in Appendix A. These formulas involve some marginal distributions, and some polynomials which have to be determined. Explicit analytical expressions for the coefficients in these polynomials are given for some particular examples in Section IV. The computation of the coefficients for more general examples is discussed in Section V. The computation of the marginal distributions is discussed in Section VI. Some numerical results are presented in Section VII.

II. NOTATION

b_n	buffer content at time n
z_n	number of packets entering buffer in time interval $(n,n+1]$
$k+1$	number of time intervals required for message arrival
α_j^i	nonnegative integers with $\alpha_0^i > 0$
$(x_n^1, \ldots, x_n^l)$	independent identically distributed vector of nonnegative integer valued random variables.

III. MATHEMATICAL MODEL

The behavior of a switching node is modeled by considering the state of the buffer at discrete times. Suppose that the buffer transmits one packet in a unit time interval. If b_n denotes the buffer content at time n, the buffer content at time $n+1$ is

$$b_{n+1} = b_n - 1 + z_n \quad \text{if } b_n \geq 1$$
$$= z_n \quad \text{if } b_n = 0$$

or equivalently

$$b_{n+1} = (b_n - 1)^+ + z_n, \tag{1}$$

where z_n is the number of packets entering the buffer in the time interval $(n,n+1]$. The peculiar feature of our queuing problem arises from the fact that the access lines are slower than the trunk line on which the output of the buffer is transmitted. For example, the 9.6 Kb/s access lines operate approximately at only one-fifth of the speed of a 50 Kb/s trunk. Then, in our discrete model, messages arriving from the access lines will consist of a random number of packets separated by five units of time.

Let us first focus our attention on the case when there are two packets to each message. If the packets are separated by d units of time, then the number of packets entering the buffer in the interval $(n,n+1]$ will be equal to the number of first packets in messages arriving in that interval, plus the number of second packets in messages whose first packets arrived d intervals earlier. Then $z_n = x_n + x_{n-d}$, where x_n is the number of first packets arriving in the interval $(n,n+1]$. Now consider the case in which the messages are transmitted by N independent sources. The probability is zero that the first packet of a message from a source enters the buffer in a unit time interval if the first packet of a message from that source entered the buffer in one of the previous $(2d-1)$ time intervals, and otherwise it is p. Then

$$\Pr\{x_n = i_0 | x_{n-j} = i_j, \quad j = 1,2,\ldots\} = \binom{N-I}{i_0} p^{i_0}(1-p)^{N-I-i_0},$$

where $I = \sum_{j=1}^{2d-1} i_j$. Thus, x_n and x_{n-d} are not independent random variables. However, if $N \to \infty$ with $Np = \lambda$ fixed, then $\Pr\{x_n = i_0\} \to (e^{-\lambda}\lambda^{i_0})/i_0!$, independently of x_{n-j}, $j = 1,2,\ldots$. Hence, if the number of sources is large, it is reasonable to assume that the random variables x_n, $n = 0,1,\ldots$, are independently and identically distributed, and we will not restrict ourselves to the Poisson distribution.

Next we consider the case where each message consists of either two packets, separated by d units of time, or of only one packet. If we let x_n^1 denote the number of first packets of two-packet messages arriving in the interval $(n,n+1]$, and x_n^2 denote the number of single packet messages arriving in the same interval, then $z_n = x_n^1 + x_{n-d}^1 + x_n^2$. In order to allow for randomness in the number of packets to a message (either one or two), x_n^1 and x_n^2 may be dependent on each other. For example, if the number of packets in a message is two with fixed probability $1-\rho$, and one with probability ρ, where $0 \leq \rho \leq 1$, then $E(t_1^{x_n^1} t_2^{x_n^2}) =$

$\Theta[(1-\rho)t_1 + \rho t_2]$, where $\Theta(t)$ is the distribution for the number of messages arriving in a unit time interval. As before, when there is a large number of independent sources, it is reasonable to assume that (x_n^1, x_n^2), $n = 0,1,\ldots$, are independent identically distributed vector random variables.

There are some obvious generalizations of the above special arrival processes. For instance, the random number of packets to a message may be as large as $m \geq 2$. Also, there may be slow access lines with several different speeds, so that the packets in a message may be separated by d_i units of time, $i = 1,\ldots,r$. This led two of the authors[4] to consider arrival processes of the form

$$z_n = \sum_{i=1}^{l} \sum_{j=0}^{k} \alpha_j^i x_{n-j}^i, \tag{2}$$

where the vector nonnegative integer valued random variables $\{(x_n^1,\ldots,x_n^l)\}$ are independent and identically distributed, and α_j^i are nonnegative integers with $\alpha_0^i > 0$. In Appendix A we summarize the relevant results pertaining to the steady-state distribution of the queue size corresponding to (1) subject to (2).

IV. EXPLICIT EXAMPLES

To calculate the generating function of the steady-state queue size from (37) and (38), it is necessary to know the polynomials $c_r(s)$, $r = 1,\ldots,k$, as well as the quantities $\phi_{rv}(s)$, $r = 0,\ldots,k$. Hence, from (39), we need to know the constants c_j. We now give explicit analytical results for some particular examples.

We first consider the arrival process

$$z_n = x_n^1 + x_{n-k}^1 + x_n^2 + x_n^3, \tag{3}$$

where $k \geq 2$, and

$$E(t_1^{x_n^1} t_2^{x_n^2} t_3^{x_n^3}) = \Theta[(1-\rho)t_1 + \rho t_2]\Psi(t_3), \tag{4}$$

with $0 \leq \rho \leq 1$ fixed. This corresponds to arrivals from two different classes of sources. One class of sources sends messages which consist either of two packets separated by k units of time with probability $1-\rho$, or of one packet with probability ρ. The other class of sources sends messages which consist of just one packet. From (2), (3), and (4), and the definition of v_{rn} in (31), it follows that

$$\begin{aligned} \phi_{rv}(s) &= \Theta(s)\Psi(s), \quad r = 0,\ldots,k-1, \\ \phi_{kv}(s) &= \Theta[(1-\rho)s^2 + \rho s]\Psi(s). \end{aligned} \tag{5}$$

It remains to give the values of c_j, which we do for $k = 2$ and 3, with $0 \leq \rho \leq 1$, and for $k = 4$ with $\rho = 0$ and $\Psi(s) \equiv 1$. It is shown in Appendix C how to determine which constants c_j occur in (46). It is also shown how the values of c_j were calculated in the case $k = 3$.

Let

$$\Theta(s) = \sum_{i=0}^{\infty} p_i s^i, \quad \Psi(s) = \sum_{i=0}^{\infty} q_i s^i. \tag{6}$$

For example, the case $p_i = e^{-\lambda}\lambda^i/i!$ corresponds to a Poisson distribution for the number of messages arriving from the first class of sources in a unit time interval. For $k = 2$ the nonzero coefficients c_j are

$$c_{00}, \qquad c_{11} = (1 - \rho)p_1 q_0 c_{00}, \tag{7}$$

with $c_{00} + c_{11} = 1$. For $k = 3$ the nonzero coefficients c_j are

$$c_{000}, \qquad c_{011} = (1 - \rho)p_1 q_0 c_{000},$$

$$c_{122} = (1 - \rho)^2 p_1^2 q_0^2 c_{000}, \qquad c_{222} = (1 - \rho)^2 p_0 p_2 q_0^2 c_{000}, \tag{8}$$

and

$$c_{111} = \frac{(1 - \rho)q_0[p_1(1 + p_0 q_1) + 2\rho p_0 p_2 q_0]}{[1 - (1 - \rho)p_0 p_1 q_0^2]} c_{000}, \tag{9}$$

with $\Sigma c_j = 1$.

For $k = 4$ we take $\rho = 0$ and $\Psi(s) \equiv 1$, corresponding to $z_n = x_n + x_{n-4}$, with $E(s^{x_n}) = \Theta(s)$. In this case, there are 14 nonzero coefficients c_j, namely

$$c_{0000} = c_0, \qquad c_{0011} = p_1 c_0, \qquad c_{0122} = p_1^2 c_0,$$

$$c_{1233} = p_1^3 c_0, \qquad c_{0222} = p_0 p_2 c_0, \qquad c_{3333} = p_0^2 p_3 c_0,$$

$$c_{1333} = c_{2333} = c_{2233} = p_0 p_1 p_2 c_0, \tag{10}$$

and

$$c_{0111} = p_1 \Delta c_0, \qquad c_{1122} = p_1^2 \Delta c_0,$$

$$c_{1222} = p_1^2(1 + p_0 p_1)\Delta c_0, \qquad c_{2222} = p_0 p_2(1 + p_0 p_1)\Delta c_0,$$

$$c_{1111} = p_1[1 + p_0^2(p_1^2 + p_0 p_2)(1 + p_0 p_1)]\Delta c_0, \tag{11}$$

where

$$\Delta = \{1 - p_0 p_1[1 + p_0^2(p_1^2 + p_0 p_2)(1 + p_0 p_1)]\}^{-1}. \tag{12}$$

We now consider a class of arrival processes z_n for which the polynomials $c_r(s)$, $r = 1, \ldots, k$, are, in fact, constants. Then the first two moments of the equilibrium queue size distribution, Ey_0 and $E(y_0^2)$, may be expressed, with the help of (59) to (61) and (63), in terms of the first three moments of v_{rn}, since $c'_r(1) \equiv 0$ and $c''_r(1) \equiv 0$. The class of arrival processes we consider corresponds to

$$\alpha_j^i > 0, \quad j = 0, \ldots, j_i, \quad i = 1, \ldots, l,$$

$$\alpha_j^i = 0, \textit{ otherwise}, \tag{13}$$

in (2). We will show that $y_{0n} = 0$ implies $y_{rn} = 0, r = 1, \ldots, k$, and hence that

$$\phi(0, s_1, \ldots, s_k) = \mu = 1 - E v_{kn}. \tag{14}$$

This implies, from (39), that $c_r(s) = \mu, r = 1, \ldots, k$.

Now, since $b_n = y_{0n}$, it follows from (29) and (13) that $y_{0n} = 0$ implies that

$$x^i_{n-j-1} = 0, \quad j = 0, \ldots, j_i, \quad i = 1, \ldots, l. \tag{15}$$

If $r + 1 \geq j_i$, then $\alpha^i_j = 0$ for $j \geq r + 1$. If, on the other hand, $r + 1 \leq j_i$, then $x^i_{n-j+r} = 0$ for $r + 1 \leq j \leq j_i$, and $\alpha^i_j = 0$ for $j > j_i$. Hence, $\alpha^i_j x^i_{n-j+r} = 0$ for $j = r + 1, \ldots, k, i = 1, \ldots, l, r = 0, \ldots, k - 1$. Thus, from (30), $y_{0n} = 0$ implies $y_{rn} = 0, r = 1, \ldots, k$, as was asserted.

As a particular example, we consider messages which consist either of one packet or of two or more packets (up to a maximum number) which arrive in consecutive time intervals. Such is the case when the access lines have the same speed as the output line. For this example, the arrival process is

$$z_n = \sum_{i=1}^{k+1} \sum_{j=0}^{i-1} x^i_{n-j}. \tag{16}$$

Hence, from (2), $l = k + 1$ and

$$\alpha^i_j = \begin{matrix} 1, & j = 0, \ldots, i - 1, i = 1, \ldots, k + 1 \\ 0, & j = i, \ldots, k, i = 1, \ldots, k \end{matrix} \tag{17}$$

so that (13) is satisfied, and $c_r(s) = \mu, r = 1, \ldots, k$. From (31) it follows that

$$\mu^i_r = \min(i, r + 1), \quad r = 0, \ldots, k, \quad i = 1, \ldots, k + 1 \tag{18}$$

and

$$v_{0n} = \sum_{i=1}^{k+1} x^i_n,$$

$$v_{rn} = \sum_{i=1}^{r} i x^i_n + (r + 1) \sum_{i=r+1}^{k+1} x^i_n, \quad r = 1, \ldots, k. \tag{19}$$

Let $\rho_i \geq 0, i = 1, \ldots, k + 1$, be the probability that there are i packets in a message, where $\sum_{i=1}^{k+1} \rho_i = 1$. Then,

$$E(t_1^{x_n^1} \cdots t_{k+1}^{x_n^{k+1}}) = \theta\left(\sum_{i=1}^{k+1} \rho_i t_i\right). \tag{20}$$

Hence, from (19),

$$\phi_{0v}(s) = \theta(s),$$

$$\phi_{rv}(s) = \theta\left(\sum_{i=1}^{r} \rho_i s^i + \sum_{i=r+1}^{k+1} \rho_i s^{r+1}\right), \quad r = 1, \ldots, k. \tag{21}$$

V. COMPUTATION OF $c_r(s)$

As described in Appendix A [see (34)], the constants $\{c_j\}$ determine the polynomials $c_r(s)$ from (39). In principle, using a program like ALTRAN,[5] which performs symbolic manipulations, it is possible to substitute (34) into (33) and equate coefficients of like powers on both sides of (33) to get the equations for the $\{c_j\}$. However, we will describe an alternate method that was used for obtaining the numerical results presented in Section VII. From (36) it is seen that

$$\mu c_{j_1 j_2 \ldots j_k} = P_{0 j_1 j_2 \ldots j_k}. \tag{22}$$

The method presented here arrives at equations for $\{P_{0j}\}$ directly from the equations for $\{P_i\}$: from (32)

$$P_{\mathbf{i}} = \sum_{j_0=0, \mathbf{j} \in A} p_{i_0 - j_1, i_1 - j_2, \ldots i_k - j_k} P_{\mathbf{j}} + \sum_{j_0 > 0, \mathbf{j} \in A} p_{i_0 - j_1 + 1, i_1 - j_2 + 1, \ldots, i_k - j_k + 1} P_{\mathbf{j}}. \tag{23}$$

The sequence of programs that were used to determine $P_{0 j_1 \ldots j_k}$ are described briefly below.

(*i*) The first step is to generate a sufficient number of equations from (23). Starting with $\mathbf{i} = (0,0, \ldots, 0)$, the right-hand side of (23) is found in symbolic form. Using a test for determining admissible states, every index $\mathbf{j}$ appearing on the right-hand side of (23) that corresponds to states not communicating with $(0,0, \ldots, 0)$ is omitted. Corresponding to each new index that arises on the right-hand side of (23), a new equation is generated from (23) by setting $\mathbf{i}$ equal to the new index. This process is terminated when every new index generated has $i_0 \geq k$. This whole process is repeated starting with each index with $i_0 = 0$, i.e., for each $P_{0 j_1 \ldots j_k}$. The total number of equations and the total number of unknowns are counted. The output of this program is this set of equations in symbolic form. The number of equations turns out to be always less than the number of unknowns. However, we know[4] a set of linear homogeneous equations must exist for $P_{0 j_1 \ldots j_k}$. In the examples considered, visual inspection revealed a few substitutions which made the number of equations one less than the number of unknowns. The details of a specific procedure for accomplishing this, in the case $l = 1$ in (2), have recently been given by Massey and Morrison.[6]

(*ii*) The normalizing constant which determines $P_{0 j_1 \ldots j_k}$ uniquely is easily seen to be $(\mu / \Sigma P_{0 j_1 \ldots j_k})$. This program starts with one of the unknowns found in step (*i*) above and sets its value equal to 1. Then it determines the maximum number of other unknowns that can be determined from this recursively, i.e., without having to invert any matrix. The best unknown to fix, the one that minimizes the number of unknowns to be solved by inverting a matrix, is selected and set equal

to 1. The matrix corresponding to the equations for the other unknowns, and the right-hand sides for these equations, are then generated in symbolic form. The output of this program is then a subroutine which generates the coefficients of $c_r(s)$ in symbolic form, as functions of the given probabilities. Some excerpts are shown in Tables I and II. In Table I is the "triangular" part of the equations for $k = 5$. $R_{i_0 \ldots i_5}$ here is $P_{i_0 \ldots i_5}$ normalized such that $P_{000000} = 1$ and $P(i) = \Pr\{x_n = i - 1\}$. In Table II are equations for generating the coefficients of $c_r(s)$ from $R_{i_0 \ldots i_5}$.

VI. COMPUTATION OF MARGINAL DISTRIBUTIONS

Equations (37) and (38) give expressions for the generating functions of the equilibrium distributions of y_{jn}, $j = 0,1, \ldots, k$. As in (30), y_{0n} is b_n the queue length. Hence, the equilibrium queue size distribution has as its generating function $\phi_0(s)$. If $\pi_{0j} = \lim_{n\uparrow\infty} \Pr\{b_n = j\}$, $j = 0,1, \ldots$, then $\phi_0(s) = \sum_{j=0}^{\infty} \pi_{0j} s^j$. One way to find $\{\pi_{0j}\}$ is to start with $\phi_k(s)$ and iterate using (38), thus obtaining an expression for $\phi_0(s)$, then to inverse transform $\phi_0(s)$. For example, using $s = e^{-j\omega}$, we can treat $\phi_0(s)$ as a function of ω, and then finding $\{\pi_{0j}\}$ corresponds to finding the Fourier series for $\phi_0(e^{-j\omega})$.

We will present a different method here. Generally, the quantities of interest are

$$\Pi_{0j} = \sum_{i=0}^{j} \pi_{0i} = \lim_{n\uparrow\infty} \Pr\{b_n \leqslant j\} \quad \text{for } j \leqslant N,$$

Table I

$R_{111111} = ((1-P(1))/P(1))*R_{000000}$
$R_{111112} = (P(2)/P(1))*R_{000000}$
$R_{222223} = (P(2)/P(1))*R_{111111}$
$R_{222224} = (P(3)/P(2))*R_{111112}$
$R_{333335} = (P(3)/P(2))*R_{222223}$
$R_{333336} = (P(4)/P(3))*R_{222224}$
$R_{444447} = (P(4)/P(3))*R_{333335}$
$R_{444448} = (P(5)/P(4))*R_{333336}$
$R_{000011} = P(1)*R_{111112}$
$R_{111123} = (P(2)/P(1))*R_{000011}$
$R_{111133} = P(1)*R_{222224}$
$R_{222235} = (P(3)/P(2))*R_{111123}$
$R_{222245} = (P(2)/P(1))*R_{111133}$
$R_{222255} = P(1)*R_{333336}$
$R_{333347} = (P(4)/P(3))*R_{222235}$
$R_{333357} = (P(3)/P(2))*R_{222245}$
$R_{333367} = (P(2)/P(1))*R_{222255}$
$R_{333377} = P(1)*R_{444448}$
$R_{000122} = P(1)*R_{111123}$
$R_{000222} = P(1)*R_{111133}$
$R_{111234} = (P(2)/P(1))*R_{000122}$
$R_{111334} = (P(2)/P(1))*R_{000222}$
$R_{111244} = P(1)*R_{222235}$
$R_{111344} = P(1)*R_{222245}$
$R_{111444} = P(1)*R_{222255}$
$R_{222346} = (P(3)/P(2))*R_{111234}$
$R_{222356} = (P(2)/P(1))*R_{111244}$
$R_{222446} = (P(3)/P(2))*R_{111334}$
$R_{222456} = (P(2)/P(1))*R_{111344}$
$R_{222556} = (P(2)/P(1))*R_{111444}$
$R_{222366} = P(1)*R_{333347}$
$R_{222466} = P(1)*R_{333357}$
$R_{222566} = P(1)*R_{333367}$
$R_{222666} = P(1)*R_{333377}$
$R_{001233} = P(1)*R_{111234}$
$R_{001333} = P(1)*R_{111244}$
$R_{002233} = P(1)*R_{111334}$
$R_{002333} = P(1)*R_{111344}$
$R_{003333} = P(1)*R_{111444}$
$R_{112345} = (P(2)/P(1))*R_{001233}$
$R_{112445} = (P(2)/P(1))*R_{001333}$
$R_{113345} = (P(2)/P(1))*R_{002233}$
$R_{113445} = (P(2)/P(1))*R_{002333}$
$R_{114445} = (P(2)/P(1))*R_{003333}$
$R_{112355} = P(1)*R_{222346}$
$R_{112455} = P(1)*R_{222356}$
$R_{112555} = P(1)*R_{222366}$
$R_{113355} = P(1)*R_{222446}$
$R_{113455} = P(1)*R_{222456}$
$R_{113555} = P(1)*R_{222466}$
$R_{114455} = P(1)*R_{222556}$
$R_{114555} = P(1)*R_{222566}$

Table II

$C(1,1) = R_{000000} + R_{000011} + R_{000111} + R_{000122} + R_{000222} + R_{001111} + R_{001122}$
$C(1,1) = C(1,1) + R_{001222} + R_{001233} + R_{001333} + R_{002222} + R_{002233} + R_{002333} + R_{003333}$
$C(1,2) = R_{011111} + R_{011122} + R_{011222} + R_{011233} + R_{011333} + R_{012222} + R_{012233}$
$C(1,2) = C(1,2) + R_{012333} + R_{012344} + R_{012444} + R_{013333} + R_{013344} + R_{013444} + R_{014444}$
$C(1,3) = R_{022222} + R_{022233} + R_{022333} + R_{022344} + R_{022444} + R_{023333} + R_{023344}$
$C(1,3) = C(1,3) + R_{023444} + R_{024444}$
$C(1,4) = R_{033333} + R_{033344} + R_{033444} + R_{034444}$
$C(1,5) = R_{044444}$
$C(2,1) = R_{000000} + R_{000011} + R_{000111} + R_{000122} + R_{000222}$
$C(2,2) = R_{001111} + R_{001122} + R_{001222} + R_{001233} + R_{001333} + R_{011111} + R_{011122}$
$C(2,2) = C(2,2) + R_{011222} + R_{011233} + R_{011333}$
$C(2,3) = R_{002222} + R_{002233} + R_{002333} + R_{012222} + R_{012233} + R_{012333} + R_{012344}$
$C(2,3) = C(2,3) + R_{012444} + R_{022222} + R_{022233} + R_{022333} + R_{022344} + R_{022444}$
$C(2,4) = R_{003333} + R_{013333} + R_{013344} + R_{013444} + R_{023333} + R_{023344} + R_{023444}$
$C(2,4) = C(2,4) + R_{033333} + R_{033344} + R_{033444}$
$C(2,5) = R_{014444} + R_{024444} + R_{034444} + R_{044444}$
$C(3,1) = R_{000000} + R_{000011}$
$C(3,2) = R_{000111} + R_{000122} + R_{0001111} + R_{001122} + R_{011111} + R_{011122}$
$C(3,3) = R_{000222} + R_{001222} + R_{001233} + R_{002222} + R_{002233} + R_{011222} + R_{011233}$
$C(3,3) = C(3,3) + R_{012222} + R_{012233} + R_{022222} + R_{022233}$
$C(3,4) = R_{001333} + R_{002333} + R_{003333} + R_{011333} + R_{012333} + R_{012344} + R_{013333}$
$C(3,4) = C(3,4) + R_{013344} + R_{022333} + R_{022344} + R_{023333} + R_{023344} + R_{033333} + R_{033344}$
$C(3,5) = R_{012444} + R_{013444} + R_{014444} + R_{022444} + R_{023444} + R_{024444} + R_{033444}$
$C(3,5) = C(3,5) + R_{034444} + R_{044444}$
$C(4,1) = R_{000000}$
$C(4,2) = R_{000011} + R_{000111} + R_{001111} + R_{011111}$
$C(4,3) = R_{000122} + R_{000222} + R_{001122} + R_{001222} + R_{002222} + R_{011122} + R_{011222}$
$C(4,3) = C(4,3) + R_{012222} + R_{022222}$
$C(4,4) = R_{001233} + R_{001333} + R_{002233} + R_{002333} + R_{003333} + R_{011233} + R_{011333}$
$C(4,4) = C(4,4) + R_{012233} + R_{012333} + R_{013333} + R_{022233} + R_{022333} + R_{023333} + R_{033333}$
$C(4,5) = R_{012344} + R_{012444} + R_{013344} + R_{013444} + R_{014444} + R_{022344} + R_{022444}$
$C(4,5) = C(4,5) + R_{023344} + R_{023444} + R_{024444} + R_{033344} + R_{033444} + R_{034444} + R_{044444}$

where N is some initially selected constant. The method presented here determines Π_{0j}, $j \leqslant N$, explicitly in a finite number of additions and multiplications. Let $\phi_r(s) = \Sigma_{j=0}^{\infty} \pi_{rj} s^j$ and $\Pi_{rj} = \Sigma_{i=0}^{j} \pi_{ri}$. We can write (37) as

$$\phi_k(s) = \frac{(1-s)\phi_{kv}(s) \cdot c_k(s)}{\phi_{kv}(s) - s}. \tag{24}$$

Denote $\phi_k(s)/(1-s)$ by $\psi_k(s)$, so that $\psi_k(s) = \Sigma_{j=0}^{\infty} \Pi_{kj} s^j$ for $|s| < 1$. Therefore,

$$\psi_k(s) = \frac{c_k(s) \cdot \phi_{kv}(s)}{\phi_{kv}(s) - s}, \quad |s| < 1. \tag{25}$$

Similarly defining $\psi_r(s) = \Sigma_{j=0}^{\infty} \Pi_{rj} s^j$ for $r = k-1, k-2, \ldots, 0$, we have

$$\psi_r(s) = s^{-1}[\psi_{r+1}(s) - c_{r+1}(s)]\phi_{rv}(s). \tag{26}$$

The functions $\phi_{rv}(s)$ are, of course, determined from the distributions of x_n^i and the constants α_j^i.

For each r, the process of determining Π_{rj} from $\Pi_{r+1,j}$ can be described as follows. Subtract the known constants $c_{r+1,j}$ from $\Pi_{r+1,j}$ for $j \leqslant$ degree

of $c_{r+1}(s)$. Since $\Pi_{r+1,0} = c_{r+1,0}$, the sequence δ_{rj} corresponding to $s^{-1}[\psi_{r+1}(s) - c_{r+1}(s)]$ is such that $\delta_{rj} = 0$ for $j < 0$ and

$$\delta_{rj} = \begin{cases} \Pi_{r+1,j+1} - c_{r+1,j+1} \text{ for } 1 \leqslant j+1 \leqslant \textit{degree of } c_{r+1}(s) \\ \Pi_{r+1,j+1} \text{ for } j+1 > \textit{degree of } c_{r+1}(s) \end{cases}$$

In order to get $\{\Pi_{rj}\}$, we now convolve this sequence $\{\delta_{rj}\}$ with the sequence corresponding to $\phi_{rv}(s)$, say $\{p_{rj}\}$. Therefore, $\Pi_{rj} = \sum_{i=0}^{j} \delta_{r,j-i} p_{ri}$. This process involves only a finite number of multiplications and additions as long as only a finite number of Π_{rj}'s are sought. Notice that in order to determine Π_{0j}, $j \leqslant N$, we have to start with values of Π_{kj} for $j \leqslant N + k$.

We will now show that there is a recursion which allows us to compute Π_{kj}, $j \leqslant N + k$, in a finite number of arithmetic operations. Let Π'_{kj} correspond to $\phi_{kv}(s)/(\phi_{kv}(s) - s)$ and let $\phi_{kv}(s) = \sum_{j=0}^{\infty} p_{kj} s^j$. Then, equating coefficients of like powers of s on both sides of

$$\sum_{j=0}^{\infty} \Pi'_{kj} s^j = \frac{\phi_{kv}(s)}{\phi_{kv}(s) - s}, \tag{27}$$

we can derive the following.

$$\Pi'_{k0} = 1,$$

$$\Pi'_{kj} = (p_{kj} + \Pi'_{k,j-1} - \sum_{i=1}^{j} p_{ki} \Pi'_{k,j-i})/p_{k0}, \quad j = 1,2,\ldots. \tag{28}$$

Once the Π'_{kj} have been determined, we convolve the sequence $\{\Pi'_{kj}\}$ with $\{c_{kj}\}$ to get $\{\Pi_{kj}\}$, as seen from (25). Summarizing, we have shown that Π_{0j} for $j \leqslant N$ can be determined from $c_r(s)$, $r = 1, \ldots, k$ and $\phi_{rv}(s)$, $r = 0,1, \ldots, k$, by performing only a finite number of multiplications and additions. The method described was used in calculating the probabilities presented in Section VII.

VII. AN EXAMPLE OF THE CALCULATIONS

We calculated the queue size distributions for various traffic intensities Ez_n for the following queuing models:

(*i*) $b_{n+1} = (b_n - 1)^+ + 2x_n, \quad k = 0$
(*ii*) $b_{n+1} = (b_n - 1)^+ + x_n + x_{n-3}, \quad k = 3$
(*iii*) $b_{n+1} = (b_n - 1)^+ + x_n + x_{n-4}, \quad k = 4$
(*iv*) $b_{n+1} = (b_n - 1)^+ + x_n + x_{n-5}, \quad k = 5,$

where the i.i.d. random variables x_n are assumed to be distributed according to the Poisson law. These cases are referred to equivalently by referring to the value of k. The objective was to determine what buffer size would suffice for a concentrator used to buffer terminals that generate two packets per message but are slower than the trunk line. We

present here some of the results. In Fig. 3 the abscissa corresponds to traffic intensity: average number of packets transmitted on the trunk line per unit of time. In the steady state, the probability that the queue size exceeds that shown on the ordinate is less than 10^{-4} for each value of traffic intensity. The cases when all packets arrive at once ($k = 0$) or "infinitely apart" ($k \uparrow \infty$) are shown, as well as the case $k = 5$. When $k = 0$, the queue size corresponding to each value of traffic intensity is either twice that of the case $k = \infty$, or one less than twice.[4] These two cases are further compared in Fig. 4, this time for buffer sizes 20 and 40, and the logarithm to base 10 of the probability of the queue size exceeding 20 and 40 is plotted.

For fixed traffic intensities, the change in the probability of overflow as a function of buffer size can be seen from Figs. 5, 6, and 7. From these figures we can see that for low traffic intensities and low buffer sizes there is a difference between the batch case $k = 0$, and $k = 5$ (see Fig. 5), but for larger traffic intensities the difference decreases substantially. From the formulas (37) and (38), it can be shown that the tail of queue size distribution is geometric for each value of k. Furthermore, for all finite values of k, the common ratio is the same as that of the case when $k = 0$, so that the similarity in the behavior of queue size distributions for large values of queue size is as expected. The slopes of the curves marked $k = 5$ in Figs. 5, 6, and 7 approach those of the curves marked $k = 0$ for large values of the abscissa. Hence, in applications where probabilities

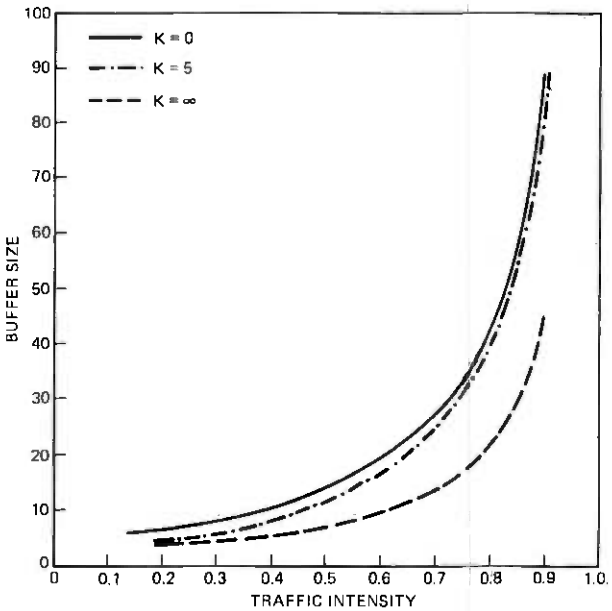


Fig. 3—Buffer size vs traffic intensity for $k = 0$, 5 and ∞, and probability of overflow $< 10^{-4}$.

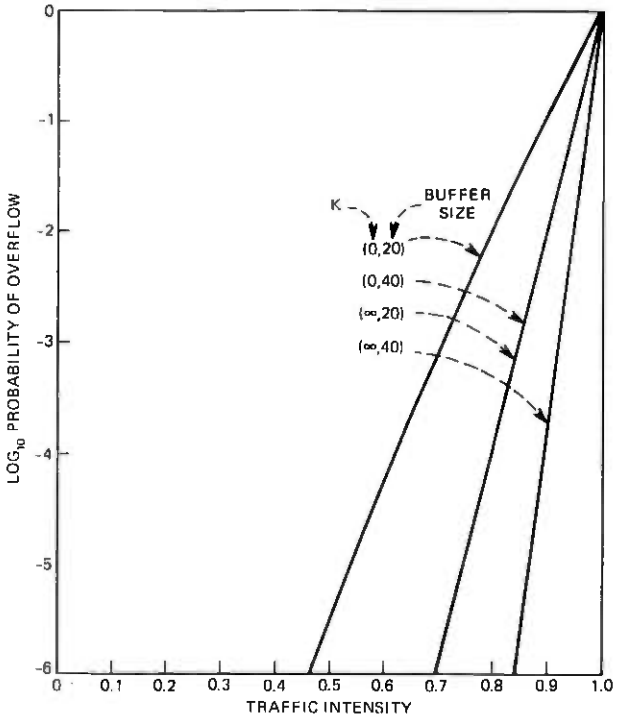


Fig. 4—Probability of overflow vs traffic intensity for $k = 0$ and ∞, and buffer sizes 20 and 40.

of overflow of 10^{-7} or smaller are required, the relative slowness of the sources of packets does not seem to reduce the buffer size required. Packets may be assumed to arrive simultaneously for the purpose of estimating the buffer size.

APPENDIX A

Summary of Formulas

We here summarize formulas for calculating the equilibrium queue size distribution. With a model for the input process z_n as in (2), the queue size b_n is described by

$$b_{n+1} = (b_n - 1)^+ + \sum_{i=1}^{l} \sum_{j=0}^{k} \alpha_j^i x_{n-j}^i. \tag{29}$$

Various formulas pertaining to (29) were derived.[4] The reader is referred to Ref. 4 for proofs of the formulas presented here. Define $y_{0n} = b_n$ and, for $r = 0,1,\ldots,k-1$,

$$y_{r+1,n} = y_{rn} + \sum_{i=1}^{l} \sum_{j=r+1}^{k} \alpha_j^i x_{n-j+r}^i. \tag{30}$$

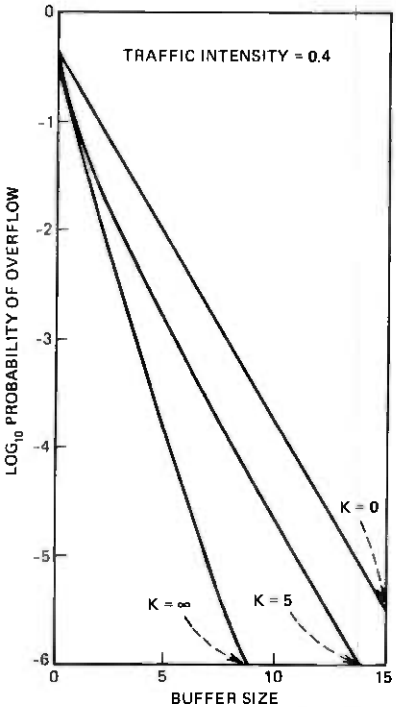


Fig. 5—Probability of overflow vs buffer size for $k = 0$, 5 and ∞, and traffic intensity 0.4.

Note that $y_{0n} \leqslant y_{1n} \leqslant \ldots \leqslant y_{kn}$. The vector process $(y_{0n}, y_{1n}, \ldots, y_{kn})$, $n = 0,1,2, \ldots$, is Markov under the assumption that $(x_n^1, x_n^2, \ldots, x_n^l)$, $n = 0,1,2, \ldots$, is a vector sequence of independent identically distributed random variables. The Markov chain, denoted by S, which corresponds to $(y_{0n}, y_{1n}, \ldots, y_{kn})$, was shown to be positive recurrent when $Ez_n < 1$. The states of S are those which communicate with $(0,0, \ldots, 0)$, since it is assumed that the buffer is empty and no one is transmitting initially, at $n = 0$. Let

$$\sum_{j=0}^{r} \alpha_j^i = \mu_r^i, \quad \sum_{i=1}^{l} \mu_r^i x_n^i = v_{rn}, \quad r = 0,1, \ldots, k. \tag{31}$$

The probabilities that enter the calculations turn out to be only those corresponding to the random variables v_{rn}, $r = 0,1, \ldots, k$. Let $p_{i_0,i_1,\ldots,i_k} = \Pr\{v_{0n} = i_0, v_{1n} = i_1, \ldots, v_{kn} = i_k\}$. Then the transition probabilities for S are given by:

$$P_{\mathbf{i}}^{n+1} = \sum_{j_0=0, \mathbf{j} \in A} p_{i_0-j_1, i_1-j_2, \ldots, i_k-j_k} P_{\mathbf{j}}^n + \sum_{j_0>0, \mathbf{j} \in A} p_{i_0-j_1+1, i_1-j_2+1, \ldots, i_k-j_k+1} P_{\mathbf{j}}^n. \tag{32}$$

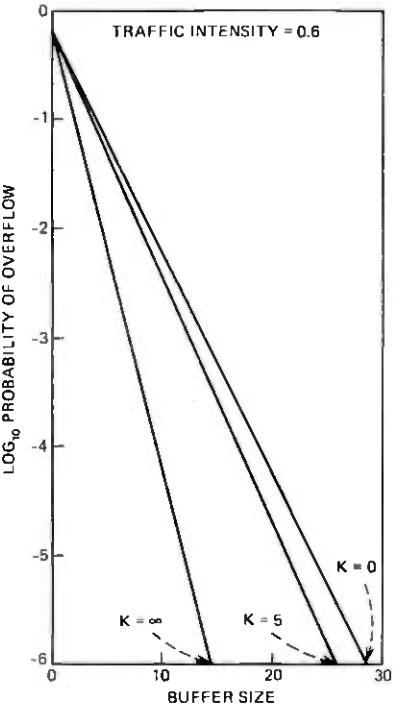


Fig. 6—Probability of overflow vs buffer size for $k = 0$, 5 and ∞, and traffic intensity 0.6.

The sums in (32) are over those indices $j_0, j_1, \ldots, j_k$ which correspond to states communicating with $(0,0,\ldots,0)$ denoted by A. As mentioned above, when $Ez_n < 1$, then $\lim_{n\uparrow\infty} P_i^n = P_i$ exists and P_i satisfies (32) with P_i^n and P_i^{n+1} both replaced by P_i. The generating function corresponding to P_i,

$$\phi(s_0,s_1,\ldots,s_k) = \sum_{i\epsilon A} P_i s_0^{i_0} s_1^{i_1} \ldots s_k^{i_k}$$

satisfies

$$\phi(s_0,s_1,\ldots,s_k) = [\phi(1,s_0,s_1,\ldots,s_{k-2},s_{k-1}s_k) \prod_{i=0}^{k} s_i^{-1} + \left(1 - \prod_{i=0}^{k} s_i^{-1}\right) \phi(0,s_0,s_1,\ldots,s_{k-2},s_{k-1}s_k)]\phi_v(s_0,s_1,\ldots,s_k), \quad (33)$$

where

$$\phi_v(s_0,s_1,\ldots,s_k) = E \prod_{i=0}^{k} s_i^{v_{in}}.$$

It can be shown[4] that ϕ has the representation

$$\phi(s_0,s_1,\ldots,s_k) = \sum_j c_j \theta_j(s_0,\ldots,s_k), \quad (34)$$

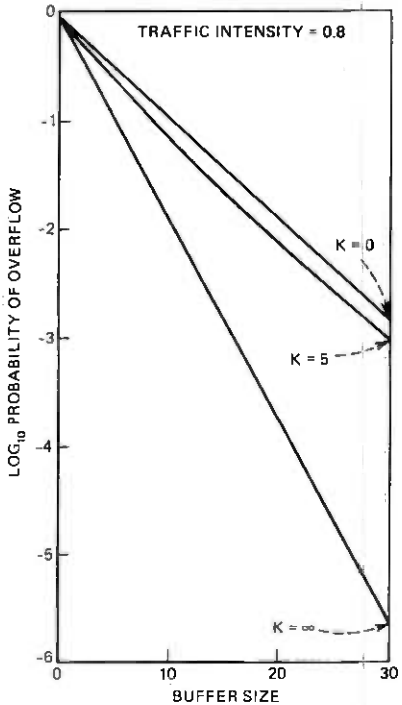


Fig. 7—Probability of overflow vs buffer size for $k = 0$, 5 and ∞, and traffic intensity 0.8.

where $\theta_j(s_0, \ldots, s_k)$ are solutions of

$$\theta_j(s_0,s_1, \ldots, s_k) = [\theta_j(1,s_0,s_1, \ldots, s_{k-2},s_{k-1}s_k) \prod_{i=0}^{k} s_i^{-1} + \mu(1 - \prod_{i=0}^{k} s_i^{-1}) s_0^{j^1} s_1^{j^2} \ldots s_{k-2}^{j^{k-1}} (s_{k-1}s_k)^{j^k}] \phi_v(s_0,s_1, \ldots, s_k), \quad (35)$$

with $\mu = 1 - Ez_n = 1 - Ev_{kn}$. Substitution of (34) into (33) yields a system of equations for c_j which together with $\Sigma c_j = 1$ uniquely determine

$$\phi(0,s_1, \ldots, s_k) = \mu \sum_j c_j s_1^{j^1} \ldots s_k^{j^k}. \quad (36)$$

Once $\phi(0,s_1, \ldots, s_k)$ is found, the generating functions corresponding to the marginal distributions, namely $\lim_{n \uparrow \infty} Es^{y_{in}} = \phi_i(s)$, are solutions of

$$\phi_k(s) = \frac{(1 - s^{-1})c_k(s)\phi_{kv}(s)}{1 - s^{-1}\phi_{kv}(s)}, \quad (37)$$

where $\phi_{jv}(s) = Es^{v_{jn}}$, $j = 0,1, \ldots, k$, and, for $r = k - 1, \ldots, 1, 0$,

$$\phi_r(s) = [s^{-1}\phi_{r+1}(s) + (1 - s^{-1})c_{r+1}(s)]\phi_{rv}(s). \quad (38)$$

Here the polynomials $c_r(s)$, $r = 1, \ldots, k$ are of finite degree and are given by

$$c_r(s) = \phi(0,1, \ldots, \overset{r}{s}, 1, \ldots, 1) = \mu \sum_{\mathbf{j}} c_{j_1 j_2 \ldots j_r \ldots j_k} s^{j_r}. \tag{39}$$

Then (37) uniquely determines $\phi_k(s)$. Using (38) k times yields $\phi_0(s)$, which by definition of y_{0n} [see (30)] is the generating function of steady-state queue size.

An alternate representation of the generating function corresponding to P_i is obtained by setting $u_j = \Pi_{i=j}^{k} s_i$, $j = 0, \ldots, k$ and defining

$$\phi(s_0, s_1, \ldots, s_k) = \Phi(u_0, u_1, \ldots, u_k). \tag{40}$$

Then, corresponding to (33),

$$\begin{aligned}\Phi(u_0, u_1, \ldots, u_k) = {} & [u_0^{-1}\Phi(u_0, u_0, u_1, \ldots, u_{k-1}) \\ & + (1 - u_0^{-1})\Phi(0, u_0, u_1, \ldots, u_{k-1})]\Phi_v(u_0, u_1, \ldots, u_k),\end{aligned} \tag{41}$$

where

$$\Phi_v(u_0, u_1, \ldots, u_k) = E\left(\prod_{r=0}^{k} u_r^{w_{rn}}\right), \quad w_{rn} = \sum_{i=1}^{l} \alpha_r^i x_n^i. \tag{42}$$

It follows from (41) that, for $j = 0,1, \ldots, k-2$, $(k \geq 2)$,

$$\begin{aligned}\Phi(s, \ldots, s, u_1, \ldots, u_{k-j}) = {} & [s^{-1}\Phi(s, \ldots, s, u_1, \ldots, u_{k-j-1}) \\ & + (1 - s^{-1})\Phi(0, s, \ldots, s, u_1, \ldots, u_{k-j-1})] \\ & \times \Phi_v(s, \ldots, s, u_1, \ldots, u_{k-j}),\end{aligned} \tag{43}$$

and

$$\begin{aligned}\Phi(s, \ldots, s, u_1) = {} & [s^{-1}\Phi(s, \ldots, s) \\ & + (1 - s^{-1})\Phi(0, s, \ldots, s)]\Phi_v(s, \ldots, s, u_1).\end{aligned} \tag{44}$$

If we set $u_1 = s$ in (44), and solve for $\Phi(s, \ldots, s)$, we obtain

$$\Phi(s, \ldots, s) = \frac{(1 - s)\Phi(0, s, \ldots, s)\Phi_v(s, \ldots, s)}{[\Phi_v(s, \ldots, s) - s]}. \tag{45}$$

It was shown[4] that $\Phi(0, u_1, \ldots, u_k)$ is a multinomial independent of u_k. From (43) to (45), $\Phi(s, u_1, \ldots, u_k)$ may be expressed in terms of $\Phi(0, s, \ldots, s, u_1, \ldots, u_{k-j-1})$, $j = 0, \ldots, k-2$, and $\Phi(0, s, \ldots, s)$. If we let $s \to 0$ in this expression, and equate $\Phi(0, u_1, \ldots, u_k)$ with the finite part, we obtain a system of homogeneous linear equations for the coefficients in the multinomial. In general, we also obtain a (consistent) set of homogeneous linear equations from finiteness conditions. The normalization condition is $\Phi(0,1, \ldots, 1) = \mu$. From (36) and (40), it follows that

$$\Phi(0, u_1, \ldots, u_k) = \mu \sum_{\mathbf{j}} c_{j_1 \ldots j_k} u_1^{j_1} \prod_{i=2}^{k} u_i^{j_i - j_{i-1}}. \tag{46}$$

Also, from (39),

$$c_r(s) = \Phi(0, s, \ldots, s, 1, \ldots, 1). \quad (47)$$

Formulas for calculating the first two moments of the equilibrium queue size distribution are derived in Appendix B.

APPENDIX B

First and Second Moments

We here derive expressions for the first two moments of $y_r = \lim_{n\to\infty} y_{rn}$. We note that y_0 represents the equilibrium queue size. By definition,

$$\phi_r(s) = E(s^{y_r}), \quad \phi_{rv}(s) = E(s^{v_{rn}}), \quad (48)$$

for $r = 0, \ldots, k$. Hence,

$$\phi_r(1) = 1, \quad \phi'_r(1) = Ey_r, \quad \phi''_r(1) = E(y_r^2) - Ey_r \quad (49)$$

and

$$\phi_{rv}(1) = 1, \quad \phi'_{rv}(1) = Ev_{rn}, \quad \phi''_{rv}(1) = E(v_{rn}^2) - Ev_{rn}. \quad (50)$$

We also note that, from (39), since $\Sigma c_j = 1$,

$$c_r(1) = \mu = 1 - Ev_{kn}, \quad r = 1, \ldots, k. \quad (51)$$

From (37),

$$\phi_k(s) = c_k(s)\phi_{vk}(s)f(s), \quad f(s) = \frac{(1-s)}{[\phi_{kv}(s) - s]}, \quad (52)$$

and, from (38),

$$\phi_r(s) = [s^{-1}\phi_{r+1}(s) + (1 - s^{-1})c_{r+1}(s)]\phi_{rv}(s), \quad r = 0, \ldots, k-1. \quad (53)$$

If we differentiate (53), we obtain

$$\phi'_r(s) = [-s^{-2}\phi_{r+1}(s) + s^{-1}\phi'_{r+1}(s) + s^{-2}c_{r+1}(s) + (1 - s^{-1})c'_{r+1}(s)] \times \phi_{rv}(s) + [s^{-1}\phi_{r+1}(s) + (1 - s^{-1})c_{r+1}(s)]\phi'_{rv}(s). \quad (54)$$

If we set $s = 1$ in (54), and use (49) to (51), we obtain

$$Ey_r = Ey_{r+1} + Ev_{rn} - Ev_{kn}, \quad r = 0, \ldots, k-1. \quad (55)$$

Next, if we differentiate (52), we find that

$$\phi'_k(s) = [c'_k(s)\phi_{kv}(s) + c_k(s)\phi'_{kv}(s)]f(s) + c_k(s)\phi_{kv}(s)f'(s). \quad (56)$$

But

$$\phi_{kv}(s) = 1 + (s-1)\phi'_{kv}(1) + \frac{1}{2}(s-1)^2\phi''_{kv}(s) + \frac{1}{6}(s-1)^3\phi'''_{kv}(1) + \ldots. \quad (57)$$

Hence, from (52),

$$f(s) = \frac{1}{[1-\phi'_{kv}(1)]} + \frac{(s-1)\phi''_{kv}(1)}{2[1-\phi'_{kv}(1)]^2} + \frac{(s-1)^2}{2}\left\{\frac{\phi'''_{kv}(1)}{3[1-\phi'_{kv}(1)]^2} + \frac{[\phi''_{kv}(1)]^2}{2[1-\phi'_{kv}(1)]^3}\right\} + \ldots. \quad (58)$$

If we let $s \to 1$ in (56), and use (49) to (51) and (58), we obtain

$$Ey_k = \frac{c'_k(1)}{(1-Ev_{kn})} + Ev_{kn} + \frac{[E(v_{kn}^2) - Ev_{kn}]}{2(1-Ev_{kn})}. \quad (59)$$

From (55), it follows that

$$Ey_j = Ey_k + \sum_{r=j}^{k-1} Ev_{rn} - (k-j)Ev_{kn}, \quad j = 0, \ldots, k-1. \quad (60)$$

This determines Ey_j, and in particular Ey_0, in view of (59).

For the second-order moments, if we differentiate (54) and set $s = 1$, we obtain

$$E(y_r^2) = E(y_{r+1}^2) - Ey_{r+1} + Ey_r + E(v_{rn}^2) - Ev_{rn} + 2(1-Ev_{rn})(Ev_{kn} - Ey_{r+1}) + 2c'_{r+1}(1), \quad (61)$$

for $r = 0, \ldots, k-1$. Finally, if we differentiate (56) and let $s \to 1$, and use the relationship

$$\phi'''_{kv}(1) = E(v_{kn}^3) - 3E(v_{kn}^2) + 2E(v_{kn}), \quad (62)$$

which follows from (48), we obtain, after some simplification,

$$E(y_k^2) = Ey_k + \frac{[c''_k(1) + 2(Ev_{kn})c'_k(1)]}{(1-Ev_{kn})} + \frac{[E(v_{kn}^2) - Ev_{kn}]c'_k(1)}{(1-Ev_{kn})^2} + \frac{[E(v_{kn}^3) - Ev_{kn}]}{3(1-Ev_{kn})} + \frac{[E(v_{kn}^2) - Ev_{kn}]^2}{2(1-Ev_{kn})^2}. \quad (63)$$

An expression for $E(y_0^2)$ may be obtained from k applications of (61), with the help of (59), (60), and (63).

APPENDIX C

Determination of Constants

We first show how to determine which constants c_j occur in (46) for the example of (3). From (2) and (30),

$$y_{0n} = b_n, \; y_{r+1,n} - y_{rn} = x^1_{n-k+r}, \; r = 0, \ldots, k-1. \quad (64)$$

Hence,

$$\Phi(u_0, u_1, \ldots, u_k) = \lim_{n\to\infty} E\left(u_0^{b_n} \prod_{r=1}^{k} u_r^{x^1_{n-k+r-1}}\right). \quad (65)$$

But from iteration of (29), it is evident that

$$b_n = 0 \Rightarrow \sum_{j=1}^{l} z_{n-j} \leqslant l - 1, \quad l = 1, \ldots, k, \tag{66}$$

where z_n is given by (2). Hence, for the example of (3),

$$b_n = 0 \Rightarrow \sum_{j=1}^{l} x_{n-j}^1 \leqslant l - 1, \quad l = 1, \ldots, k. \tag{67}$$

The inequalities in (67) determine the admissible vectors $(x_{n-k}^1, \ldots, x_{n-1}^1)$ corresponding to $b_n = 0$, and thus, from (65), which constants c_j occur in (46). Note, in particular, that $x_{n-1}^1 = 0$ implies that $\Phi(0,u_1, \ldots, u_k)$ is independent of u_k.

For purposes of illustration, we show how to calculate the values of c_j for the example of (3), subject to (4), in the case $k = 3$. From the above procedure it is found that

$$\Phi(0,u_1,u_2,u_3) = \chi(u_1,u_2) = \mu(c_{000} + c_{111}u_1 + c_{011}u_2 + c_{222}u_1^2 + c_{122}u_1u_2). \tag{68}$$

But from (2) and (42),

$$\Phi_v(u_0,u_1,u_2,u_3) = \Theta[(1-\rho)u_0u_3 + \rho u_0]\Psi(u_0). \tag{69}$$

Hence, from (45),

$$\Phi(s,s,s,s) = \frac{(1-s)\chi(s,s)\Theta[(1-\rho)s^2 + \rho s]\Psi(s)}{|\Theta[(1-\rho)s^2 + \rho s]\Psi(s) - s|}. \tag{70}$$

Then, from (44), we obtain

$$\Phi(s,s,s,u_1) = \frac{(1-s)\chi(s,s)\Theta[(1-\rho)su_1 + \rho s]\Psi(s)}{|\Theta[(1-\rho)s^2 + \rho s]\Psi(s) - s|}. \tag{71}$$

Also, from (43), we have

$$\Phi(s,s,u_1,u_2) = [s^{-1}\Phi(s,s,s,u_1) + (1-s^{-1})\chi(s,s)] \times \Theta[(1-\rho)su_2 + \rho s]\,\Psi(s), \tag{72}$$

and

$$\Phi(s,u_1,u_2,u_3) = [s^{-1}\Phi(s,s,u_1,u_2) + (1-s^{-1})\chi(s,u_1)] \times \Theta[(1-\rho)su_3 + \rho s]\Psi(s). \tag{73}$$

From (71) to (73), it follows that

$$\Phi(s,u_1,u_2,u_3) = \frac{(1-s)}{s}\Theta[(1-\rho)su_3 + \rho s]\Psi(s) \times \left\{\Theta[(1-\rho)su_2 + \rho s]\frac{\Psi(s)}{s} \times \left[\frac{\Theta[(1-\rho)su_1 + \rho s]\Psi(s)}{|\Theta[(1-\rho)s^2 + \rho s]\Psi(s) - s|} - 1\right]\chi(s,s) - \chi(s,u_1)\right\}. \tag{74}$$

From finiteness at $s = 0$ we deduce, with the help of (6), that

$$\chi(0,\mu_1) = \mu[1 + (1 - \rho)p_1q_0\mu_1]c_{000}. \tag{75}$$

Hence, from (68),

$$c_{011} = (1 - \rho)p_1q_0c_{000}. \tag{76}$$

If we now let $s \to 0$ in (74), and equate coefficients, we obtain, in addition to (76), the relations

$$c_{122} = (1 - \rho)^2p_1^2q_0^2c_{000},\ c_{222} = (1 - \rho)^2p_0p_2q_0^2c_{000}, \tag{77}$$

and

$$\begin{aligned} c_{111} = {} & (1 - \rho)q_0(p_1 + p_0p_1q_1 + 2\rho p_0p_2q_0)c_{000} \\ & + (1 - \rho)p_0p_1q_0^2(c_{111} + c_{011}) - p_0q_0c_{122}. \end{aligned} \tag{78}$$

If we substitute the values of c_{011} and c_{122}, from (76) and (77), into (78), we obtain (9).

REFERENCES

1. W. W. Chu, "Buffer Behavior for Poisson Arrivals and Multiple Synchronous Constant Outputs," IEEE Transactions on Computers, *C-19*, No. 6 (June 1970), pp. 530–534.
2. W. W. Chu, "Buffer Behavior for Batch Poisson Arrivals and Single Constant Output," IEEE Transactions on Communication Technology, *COM-18*, No. 5 (October 1970), pp. 613–618.
3. W. W. Chu and L. C. Liang, "Buffer Behavior for Mixed Input Traffic and Single Constant Output Rate," IEEE Transactions on Communications, *COM-20*, No. 2 (April 1972), pp. 230–235.
4. B. Gopinath and J. A. Morrison, "Discrete-Time Single Server Queues with Correlated Inputs," B.S.T.J. *56*, No. 9 (November 1977), pp. 1743–1768.
5. W. S. Brown, "ALTRAN User's Manual," Fourth Edition, Bell Laboratories, 1977.
6. W. A. Massey and J. A. Morrison, "Calculation of Steady-State Probabilities for Content of Buffer with Correlated Inputs," B.S.T.J., *57*, No. 9 (November 1978).

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 8, October 1978
Printed in U.S.A.

Speaker Verification by Human Listeners over Several Speech Transmission Systems

By C. A. McGONEGAL, L. R. RABINER, and B. J. McDERMOTT

(Manuscript received January 6, 1978)

Although a great deal has been learned about how speakers are verified, both by humans and by machines, several factors have not yet been studied. One of these factors is the effect of the transmission system (over which the message is communicated) on the accuracy with which verification is achieved. This factor is potentially an important one for digital communications problems over telephone lines where the transmission system could vary from one which gives a high-quality coded representation of the signal (e.g., log PCM) to a low-bit-rate vocoder. The purpose of this paper is to demonstrate the effects of three speech transmission systems on verification accuracy by human listeners. It is shown that the false alarm rate (i.e., a customer is rejected) is significantly higher when the test and reference utterances are transmitted by different systems than when transmitted by the same system. The miss rate (i.e., an imposter is accepted) is not significantly different for similar comparisons except for one of the conditions. The overall conclusion of this experiment is that speaker verification by human listeners cannot be performed as accurately over mixed speech transmission systems as over the same transmission system.

I. INTRODUCTION

Speaker verification, both automatically by machine and by human listeners, is an important problem in the area of man-machine communication by voice.[1–8] The verification problem has applications in the business community for such things as voice banking by telephone, credit card transactions (including charging of telephone calls), and access of privileged or confidential information.

As shown in Fig. 1, the speaker verification problem, either by human listeners or by machine, has two aspects—the creation of a reference pattern (i.e., the training phase) and the determination of similarity between a test and a reference pattern (i.e., the testing phase). When

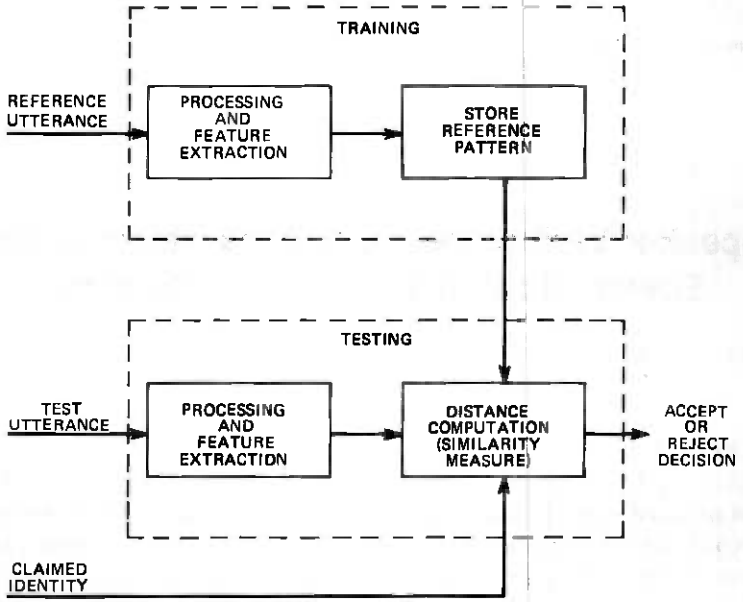


Fig. 1—Block diagram of the speaker verification problem.

verification is to be performed over telephone lines, an additional factor complicates the problem—namely, the transmission system used in the telephone plant. With the increased reliance on digital speech-processing techniques, such as waveform coders and linear prediction methods, the interesting possibility arises that the test pattern for speaker verification may have been coded or vocoded, whereas the reference pattern may not have been subjected to the same processing. Past experiments by Rosenberg[2] have studied the problem of verification both by human listeners and with an automatic system using natural speech for both the test and reference patterns. The purpose of this experiment is to evaluate how several speech transmission systems affect the process of speaker verification by human listeners. In future work, we will investigate the parallel problem—how these same factors affect automatic methods of speaker verification.

The organization of this paper is as follows. Section II describes the way in which the evaluation was carried out. This section includes a description of the speech transmission systems, as well as the experimental procedure used to measure system performance. In Section III, the experimental results are presented in terms of a signal detectability model, and in Section IV the results are discussed.

II. EXPERIMENTAL EVALUATION

For the experiment to be described below, both the reference and test utterances were preprocessed by one of the following three transmission systems:

(*i*) Bandpass filtering from 100 to 2600 Hz.

(*ii*) Adaptive differential pulse code modulation (ADPCM) coding, followed by bandpass filtering from 100 to 2600 Hz.

(*iii*) Linear predictive vocoding (LPC), followed by bandpass filtering from 100 to 2600 Hz.

The bandwidth of all three systems was set to 2500 Hz, in accordance with the requirements of the ADPCM coder, to ensure that the speech bandwidth was not a factor in determining relative verification accuracy.

The ADPCM coder used in this experiment was a simulation of the coder built by Bates,[9] based on the work of Cummiskey et al.[10] Figure 2 is a block diagram of the ADPCM system. The input signal is band-pass-filtered from 100 to 2600 Hz and sampled at a 6000-Hz rate. A 4-bit adaptive quantizer was used to code the difference signal, giving an overall bit rate of 24 kb/s for the coder. The step-size multiplier of the quantizer ranged over a 41-dB range (i.e., the ratio between the largest and smallest step size was 114 to 1). A first-order predictor was used with a multiplier of $\alpha = 0.9375$. Signal levels were chosen so that the coder was operating at approximately the optimum point.

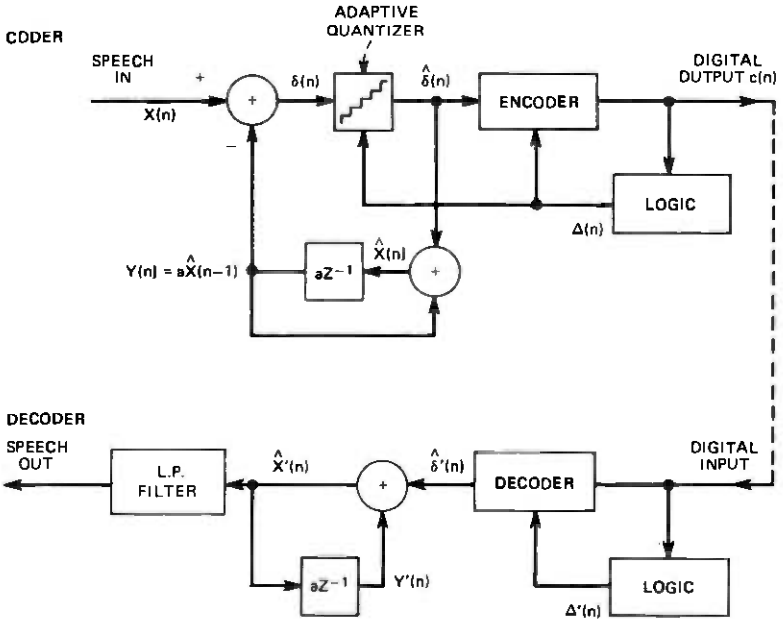


Fig. 2—Block diagram of an ADPCM coder.

A block diagram of the LPC vocoder is given in Fig. 3. The implementation was based on the autocorrelation method of linear prediction.[11–13] Pitch detection and voiced-unvoiced decision were performed using the modified autocorrelation pitch detector of Dubnowski et al.[14] The input signal was sampled at a 10-kHz rate, and a 12-pole LPC analysis was done with a pitch adaptive, variable frame size, at a rate of 100 frames per second.[15] No quantization of the LPC parameters was used in this experiment.

2.1 Data base for the evaluation

To evaluate the three transmission systems, a data base was designed which included

(*i*) 16 speakers designated "customers":
8 male.
8 female.
(*ii*) 62 speakers designated "imposters":
31 male.
31 female.
(*iii*) 2 sentences:
"We were away a year ago"—male utterance.
"I know when my lawyer is due."—female utterance.
(*iv*) 3 versions of each utterance:
bandpass filtered speech—SP.
ADPCM coded and filtered speech—ADPCM.
LPC vocoded and filtered speech—LPC.

The set of male utterances used in this study were those used by Rosenberg in his earlier work.[4] New recordings were made for the set of female utterances. Both male and female speakers recorded 10 utterances over a period of several weeks. The imposters provided just one recording each.

2.2 Experimental procedures

To test the effects on verification of combinations of different speech systems for the reference and test utterances, a paired-comparison test was used. A block diagram of the experimental arrangement used is shown in Fig. 4. Each test pair consisted of a comparison utterance and a challenge utterance. The comparison utterance was always a customer utterance processed by one of the three transmission systems. The challenge utterance was either an imposter utterance (customer-imposter pair) or one of the remaining nine utterances of the same customer (customer-customer pair) processed by one of the three systems.

Ten analog tapes were prepared. Each tape consisted of only male or female utterances with 48 customer-customer and 48 customer-imposter pairs randomly presented. The eight customers were presented in each

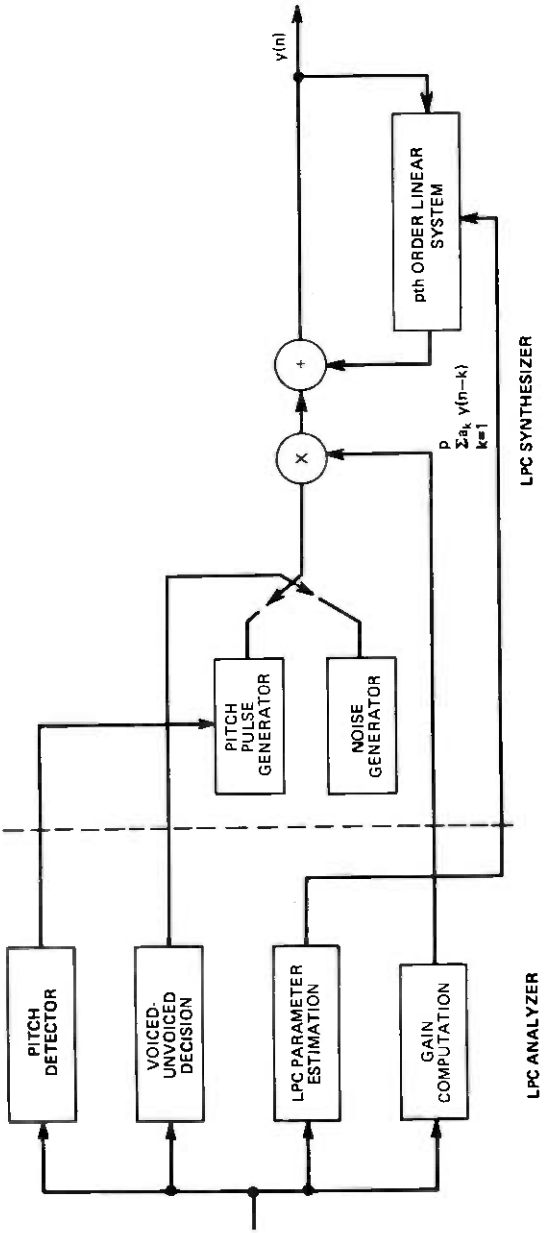


Fig. 3—Block diagram of an LPC vocoder.

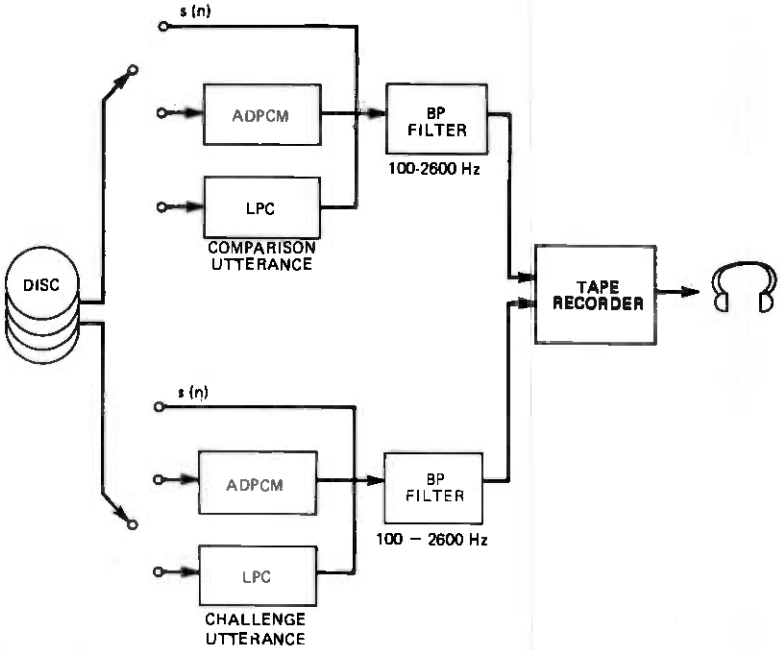


Fig. 4—Block diagram of the experimental arrangement.

of the transmission system combinations in both types of paired comparisons. When the transmission system combinations were heterogeneous, four of the eight customers were presented in each of the two orders.

Two tapes, one containing the male utterances and one containing the female utterances, were presented over headphones to five different groups of six naive subjects who were seated in a soundproof booth. The subjects were asked to indicate whether the comparison and challenge utterances were spoken by the same or by different speakers. They received no training for the experiment and were given no instructions as to the costs of any type of error.

III. EXPERIMENTAL RESULTS

The subject responses can be interpreted according to signal detection theory[16]—i.e., as a hypothesis test. For any trial in the test, there were two input hypotheses (same speakers or different speakers) and two possible subject responses (SAME and DIFFERENT). Therefore, each trial can be represented by the intersection of one of the input alternatives and one of the response alternatives as indicated in Fig. 5. There are two types of errors (*false alarm* and *miss*) and two types of correct responses (*hit* and *correct rejection*) associated with each trial. A *false alarm* (the rejection of a customer) is defined as a subject response of DIFFERENT

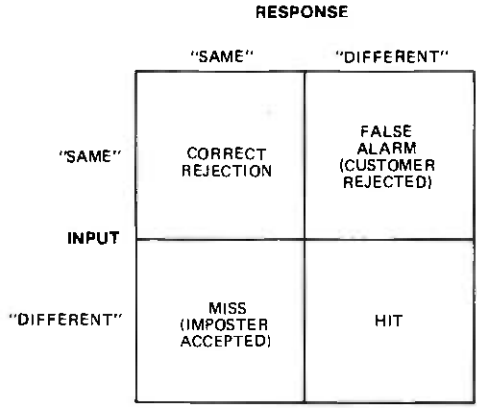


Fig. 5—The various response classifications for detecting an imposter.

when both utterances are spoken by the same speaker. A *miss* (the acceptance of an imposter) is defined as a subject response of SAME when the challenge utterance was spoken by a different speaker. A *hit* (acceptance of a customer) is defined as a subject response of DIFFERENT when the challenge utterance was spoken by a different speaker. A *correct rejection* (rejection of an imposter) is defined as SAME when both utterances are spoken by the same speaker.

The false alarm rates for male and female customers are shown in Fig. 6. The customer false alarm rates are represented by vertical bars—one bar per customer for each pair of transmission systems. The percentage of time a customer was rejected varied for each customer in a group and also between groups. Although customers were asked to record their sentence the same way at each session, several had dramatic pitch changes. Since the subjects were not familiar with the customer voices, they tended to reject those customers. In general, the false alarm rates were fairly low and in many cases less than 10 percent.

The miss rates for the male and female customers are shown in Fig. 7. The percentage of time an imposter was accepted also varied greatly among customers. As seen in this figure, the miss rates were generally higher than 15 percent for all transmission pairs except for the LPC-ADPCM pair.

An alternative way of displaying the information in the subject data is in terms of the likelihood ratio. The likelihood ratio, l, is a good measure of signal detectability and is defined as

$$l = \frac{P\,(\text{hit})}{P\,(\text{false alarm})},$$

where

$$P\,(\text{hit}) = 1 - P\,(\text{miss}).$$

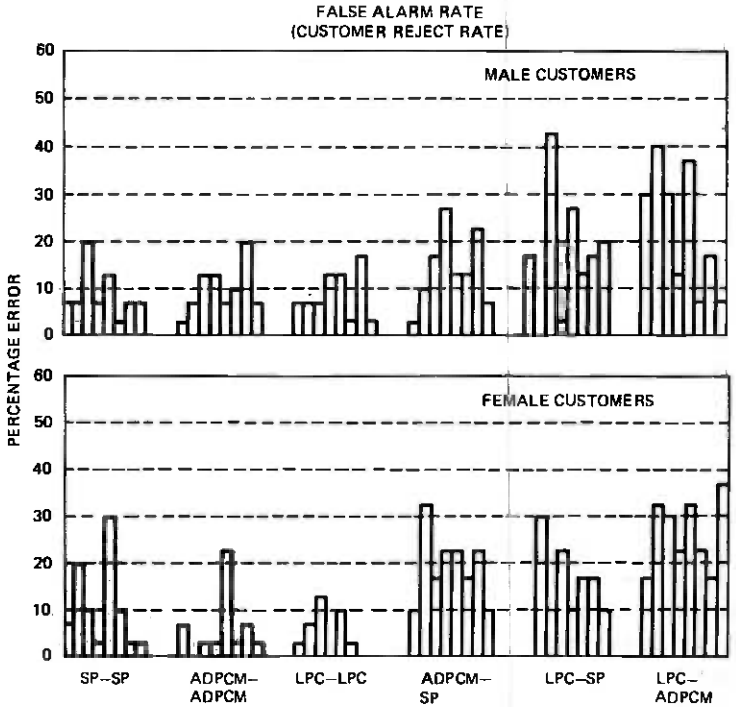


Fig. 6—False alarm rates for male and female customers for all transmission system pairs.

Figure 8 is a plot of the likelihood ratios for the male and female customers for transmission system pairs. The height of a vertical bar is the ratio of the number of times a subject correctly detected an imposter to the number of times a subject rejected a customer. Therefore, the larger the ratio (the higher the vertical bar), the better the subject performance for that transmission pair. As seen in this figure, the ratios are highest among the homogeneous systems for both male and female customers and lowest among the mixed systems. For several of the female customers, there were no false alarms so the likelihood ratios are infinite. Again, the large amount of variation among customers can be seen.

Because of the high variation among the customers (as seen in the preceding figures), the data for the false alarm and miss rates were pooled on the basis of median error scores rather than mean error scores. The median errors of the eight customers for each pair of transmission systems are shown in Fig. 9. Both a chi-square and a Fisher test[17] were applied to the median data to determine when significant differences existed between (*i*) the male and female customer medians for each pair of transmission systems (no significant differences were indicated) and (*ii*) the combined male and female customer medians of each pair of

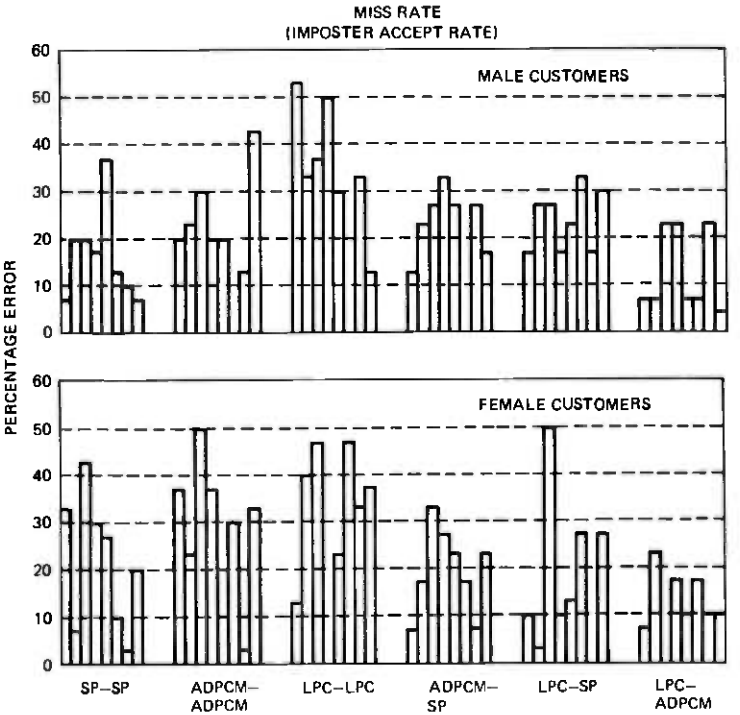


Fig. 7—Miss rates for male and female customers for all transmission pairs.

transmission systems. The following significant differences were found.

(*i*) The false alarm rate for mixed systems, that is, when the test and reference utterances were processed by different transmission systems, is significantly different from that of homogeneous systems.

(*ii*) The false alarm rate for the SP-SP pair was significantly different from all other transmission pairs.

(*iii*) The miss rates for mixed and homogeneous systems were not significantly different except for two transmission pairs. The LPC-ADPCM system had a significantly lower miss rate than the other system pairs, and the LPC-LPC pair had a significantly higher miss rate than any other transmission system pair.

Finally, Fig. 10 shows the overall error rates, that is, the average of the false alarm rates and miss rates, for each speech transmission pair. The overall error rate is between 10 and 20 percent for all speech transmission pairs. The lowest overall error rate is observed for SP-SP and ADPCM-ADPCM transmission pairs. There is no significant difference in the overall error rate for any of the system combinations.

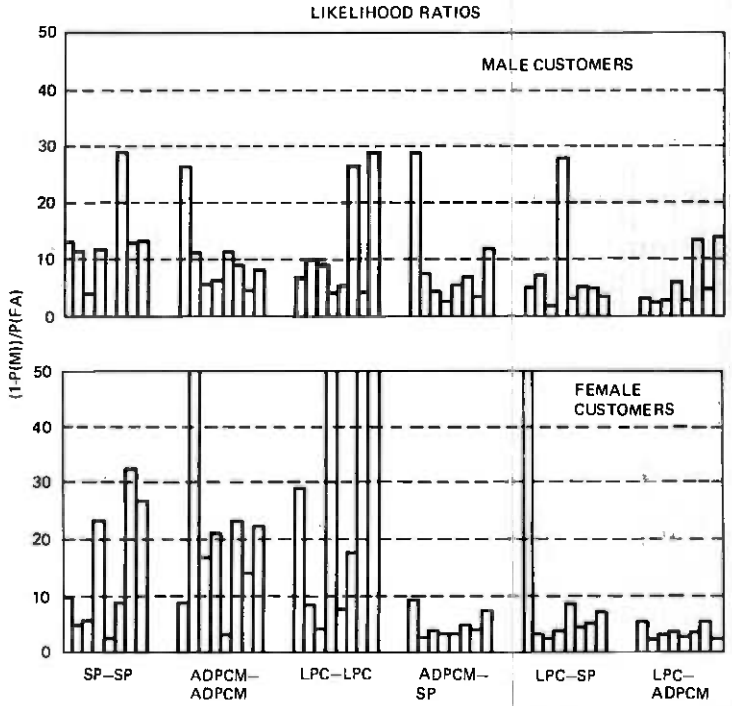


Fig. 8—Likelihood ratios for male and female customers for all transmission pairs.

IV. DISCUSSION OF THE RESULTS

The results (i.e., the false alarm and miss rates for the speech transmission pairs) can be interpreted in terms of the dimensions of difference between transmission systems. There are three main differences between system pairs, as shown in Fig. 11. These are:

(*i*) No difference—the same transmission system is used for both the test and reference utterances. The transmission system pairs in this category are SP-SP, ADPCM-ADPCM, and LPC-LPC.

(*ii*) One dimension of difference—the transmission system pair consists of one utterance transmitted over a clear channel and the other utterance processed over an ADPCM or LPC system. The transmission system pairs in this category are ADPCM-SP and LPC-SP.

(*iii*) Two dimensions of difference—two very different transmission systems are used for the test and reference utterances. Only the LPC-ADPCM pair is in this category.

In category (*i*), the median customer rejection (false alarm) rates were very low, and the median imposter acceptance (miss) rates were very high. This result reflects a subject bias toward responding SAME when the test and reference utterances are processed over the same transmission system. The low customer rejection rates in this category also indicate that subjects can easily verify customer-customer pairs.

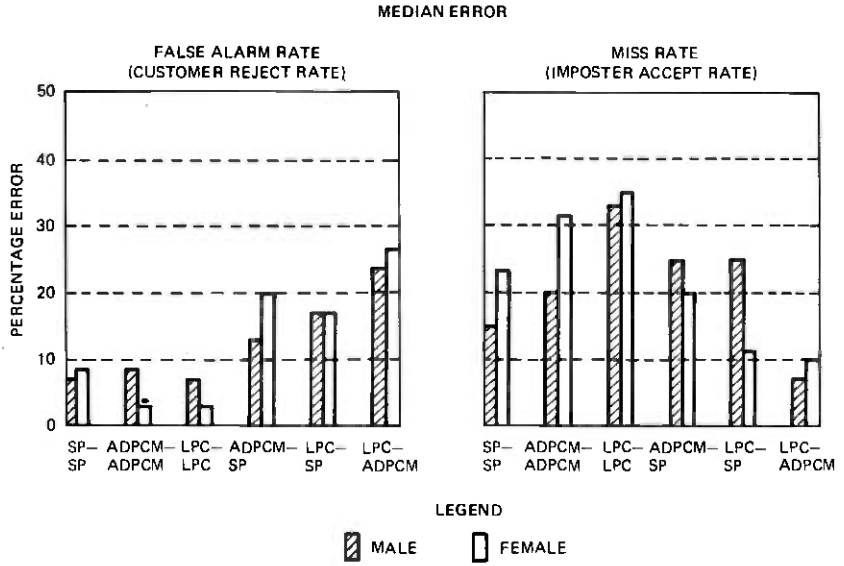


Fig. 9—Median false alarm rates and median miss rates for all customers and for all transmission pairs.

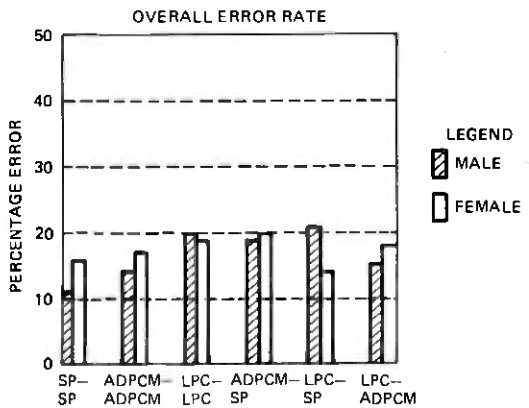


Fig. 10—Overall error rates for all customers and for all transmission pairs.

In category (*ii*), the median error rates were approximately the same for both customer-customer and customer-imposter pairs. For these cases, about 15 to 20 percent of the time a customer would be rejected and an imposter would be accepted. Whether a customer or an imposter was processed over either one of the transmission systems seemed to make very little difference. This result indicates that the subjects were confused by the pairing of an ADPCM or LPC system with a natural speech utterance.

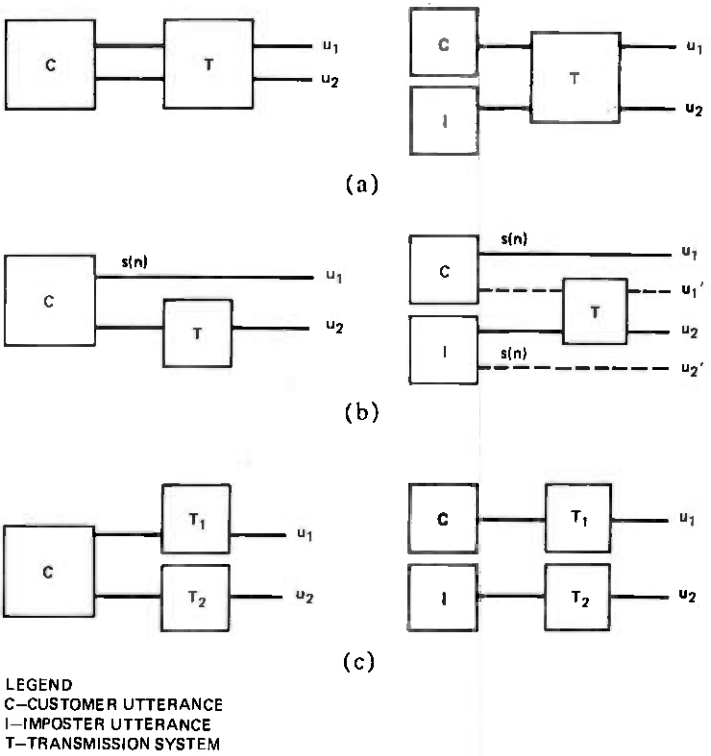


Fig. 11—Dimensions of difference between transmission system pairs. (a) No difference—SP-SP, ADPCM-ADPCM, LPC-LPC. (b) One dimension of difference—ADPCM-SP, LPC-SP. (c) Two dimensions of difference—LPC-ADPCM.

In category (*iii*) (i.e., LPC-ADPCM pair), the median customer rejection rate was very high and the median imposter acceptance rate was very low. The speech quality produced by the ADPCM and LPC systems is extremely different, and the results seem to reflect a subject bias toward responding DIFFERENT in this situation.

The overall conclusion is that the speaker verification task by human listeners is easiest when homogeneous systems are used and is significantly more difficult when mixed systems are used for the test and reference patterns.

V. COMPARISONS WITH PREVIOUS WORK

Since the male utterances used in Rosenberg's experiment[4] were also used in this experiment, the male SP-SP error rates can be compared. The median false alarm and miss rates for the two experiments are shown in Table I. The median false alarm rate observed by Rosenberg was 3.3 percent, which is about two times smaller than the error rate seen in this study. The median miss rate observed by Rosenberg was only 2.8 percent,

Table I — Comparison of male SP-SP median error rates

Experiments	False Alarm Rate	Miss Rate
Current experiment	7%	15%
Rosenberg[4]	3.3%	2.8%*

* The average miss rate after completion of 25 percent of the listening sessions was six percent.

which is considerably less than that observed here. Ideally, the error rates for the two experiments should be the same. However, several differences between the two experiments may have influenced the results. These are:

(*i*) The bandwidth used in Rosenberg's experiment was 4 kHz, whereas the bandwidth used here was 2.5 kHz. The 2.5 kHz bandwidth was required by the ADPCM system, which had a sampling frequency of 6 kHz.

(*ii*) The written instructions given to the subjects differed in both experiments. In Rosenberg's experiment, the subjects were divided into two groups. One group was provided with instructions intended to lower the false alarm rate, while the other group was provided with instructions intended to lower the miss rate. In this experiment, all subjects received the same instructions with no intent to lower either type of error rate.

(*iii*) In this experiment, no repeat judgments were obtained from any one subject, but in Rosenberg's experiment, 32 repeated judgments were obtained from each subject for customer-customer pairs and 4 judgments were obtained for each customer-imposter pair. Even though there was no prior training in either experiment, Rosenberg noted a training effect imbedded in his data. He found an average miss rate of 6 percent after the completion of 25 percent of the listening sessions. This rate is approximately two times less than the miss rate we observed. No drop was noticed with regard to the false alarm rate.

(*iv*) The last factor that may have influenced the difference in the results is the fact that Rosenberg's experiment consisted entirely of SP-SP test presentations. In this experiment, the SP-SP presentations were randomly combined with all other transmission pair presentations.

VI. SUMMARY

The purpose of this experiment was to show the effect of different transmission systems on speaker verification accuracy by human listeners. It was shown that when the reference and test utterances were recorded from different transmission systems, the false alarm rate was significantly larger than when they were recorded from the same transmission system. However, with one exception, the miss rates were essentially equivalent, independent of the transmission system. As such, it is concluded that speaker verification by humans cannot be performed as accurately when different transmission systems are used.

VII. ACKNOWLEDGMENT

The authors wish to acknowledge the assistance of A. E. Rosenberg in providing the male data base from his previous experiments in speaker verification tasks.

REFERENCES

1. J. L. Flanagan, "Computers that Talk and Listen: Man-Machine Communication by Voice," Proc. IEEE, *64*, No. 4 (April 1976), pp. 405–433.
2. A. E. Rosenberg, "Automatic Speaker Verification: A Review," Proc. IEEE, *64*, No. 4 (April 1976), pp. 475–487.
3. R. C. Lummis, "Speaker Verification by Computer Using Speech Intensity for Temporal Registration," IEEE Trans. Audio Electroacoust., *AU-21* (1973), pp. 80–89.
4. A. E. Rosenberg, "Listener Performance in Speaker Verification Tasks," IEEE Trans. Audio Electroacoust., *AU-21* (1973), pp. 221–225.
5. A. E. Rosenberg, "Evaluation of an Automatic Speaker Verification System over Telephone Lines," B.S.T.J., *55*, No. 6 (JULY–August 1976), pp. 723–744.
6. S. K. Das and W. S. Mohn, "A Scheme for Speech Processing in Automatic Speaker Verification," IEEE Trans. Audio Electroacoust., *AU-19* (1971), pp. 32–43.
7. G. R. Doddington, "A Method of Speaker Verification," Ph.D. dissertation, Univ. Wisconsin, Madison, 1970.
8. E. Bunge, "Automatic Speaker Recognition by Computers," in Proc. Carnahan Conf. Crime Countermeasures, 1975.
9. S. L. Bates, "A Hardware Realization of a PCM-ADPCM Code Converter," M.I.T. M.S. Thesis, Dept. of Elec. Eng. and Comp. Sc., January 1976.
10. P. Cummiskey, N. S. Jayant, and J. L. Flanagan, "Adaptive Quantization in Differential PCM Coding of Speech," B.S.T.J., *52*, No. 7 (September 1973), pp. 1105–1118.
11. J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*, Springer-Verlag, 1976.
12. J. D. Markel and A. H. Gray, Jr., "A Linear Prediction Vocoder Simulation Based Upon an Autocorrelation Method," IEEE Trans. on Acoustics, Speech, and Signal Proc., *ASSP-22*, No. 2 (April 1974), pp. 124–134.
13. B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," J. Acoust. Soc. Amer., *50*, (1971), pp. 637–655.
14. J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner, "Real-Time Digital Hardware Pitch Detector," IEEE Trans. Acoust., Speech, Signal Proc., *ASSP-24* (February 1976), pp. 2–8.
15. L. R. Rabiner, "On the Use of Autocorrelation Analysis for Pitch Detection," IEEE Trans. Acoust., Speech, and Signal Proc., *ASSP-25*, No. 1 (February 1977), pp. 24–33.
16. D. M. Green and J. A. Swets, *Signal Detection and Psychophysics*, New York: Wiley, 1966.
17. S. Siegel, *Nonparametric Statistics for Behavioral Sciences*, New York: McGraw-Hill, 1956.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 8, October 1978
Printed in U.S.A.

Efficient Utilization of Satellite Transponders via Time-Division Multibeam Scanning

By A. S. ACAMPORA and B. R. DAVIS

(Manuscript received November 28, 1977)

The space segment of a satellite system is proposed wherein a fixed number of identical transponders are shared among a larger number of spot beam regions which completely span a large total service area. Time-division multiple-access techniques are employed, and each transponder is rapidly scanned over appropriately defined group pairs of spot beam regions, thereby establishing full coverage and full interconnectivity. The service is matched to the nonuniform traffic requirements exhibited among the various spot beam regions, reliability can be optimized since all transponders are identical, and each transponder is utilized with an efficiency of 100 percent. A mathematical proof is presented which shows that the traffic can always be assigned on a nonconflicting basis, and an efficient assignment technique is described.

I. INTRODUCTION

The trend toward higher frequency communication satellites employing multiple spot beams affords significant capacity advantages relative to lower frequency, wide-coverage area systems, since the allocated spectral band can be reused in the various spot beams.[1,2] When used in conjunction with digital modulation techniques and time-division multiple-access, the various coverage regions are readily interconnected via an onboard satellite switch operating in the time-division multiplex mode. In addition to the frequency reuse capability, the down-link transmitter power requirements are generally reduced because the antenna gain is higher than for a wide coverage area system.

Despite these advantages, however, multiple spot beam satellites have several distinct drawbacks. These are generally associated with conflicting requirements concerning reliability,[3] coverage and blackout areas,[4] efficient transponder utilization, and nonuniform traffic density demands.

In the following sections we explore these conflicting requirements and review some partial solutions proposed to date. Then we present a space segment configuration for a satellite communication system which provides high reliability with a minimum of redundancy, access from any location within a wide service area, and up to 100-percent efficiency in transponder capacity utilization matched to an arbitrary nonuniform density of traffic demand over the entire service region. This system employs N identical transponders which are shared on a time-division basis among $M \geq N$ antenna ports spanning the entire service region. Starting with the traffic demand matrix, we give a mathematical proof that the desired arrangement is always possible, and present an assignment algorithm.

Such a system might find applicability to a geosynchronous satellite operating in the 12/14-GHz band. From synchronous orbit, the 3-dB contour of a beam radiated from a $2\frac{1}{4}$-m antenna would cover about 1 percent of the continental United States. Total United States coverage, then, would require about 100 such beams. Not only is the offered traffic nonuniformly distributed over the subregions, but within most such subregions the traffic is far too small to justify deployment of a dedicated wideband transponder to each. Moreover, from a practical viewpoint, the number of onboard satellite transponders is limited to the range of 10 to 20 by weight, power, and cost constraints. Through proper time-division assignment, these 10 to 20 transponders can efficiently provide service to the entire United States.

II. PROBLEMS OF MULTIPLE SPOT BEAM SATELLITES

A major problem in multibeam satellite design is one of transponder reliability. Unlike area coverage systems wherein the allocated band is divided among several transponders and service is provided via frequency division multiple access, it is desirable to serve each spot beam of a multibeam satellite system with a single transponder. With this approach, the required number of transponders is kept from becoming prohibitive, and the weight of the communications subsystem is minimized. However, sufficient redundancy must be provided to ensure high reliability for each transponder since single failures would preclude continuing service to the area serviced by that transponder. By contrast, for area coverage systems using frequency division multiple access, isolated failures merely cause a slight increase in the demand presented to the surviving transponders.

A second problem in multibeam satellite systems concerns efficient utilization of the satellite transponders. In general, the traffic demands from the various coverage areas (or footprints) are nonuniform. Thus, to utilize each transponder fully, the capacity of each must be tailored to the traffic demand of the area covered by that transponder. A tech-

nique for achieving such a custom fit has been reported,[5] wherein the bit rate of each beam is selected as a fixed multiple of some basic rate. At the satellite, each uplink beam is demultiplexed into several basic rate bit streams, switched, and then remultiplexed into downlink beams. One disadvantage of this scheme is that onboard demodulation and remodulation is required. However, a more serious disadvantage in such a system is the need for nonidentical transponders, which precludes sharing a common pool of spare transponders among all beams, and the reliability of the system suffers.[3]

A third problem of multibeam satellites involves means of accessing traffic from areas not within the footprint of some spot beam. Several solutions have been proposed,[4] involving sharing the spectrum between spot beams and an area coverage beam. These have the disadvantage that the area coverage transponders are different from the spot beam transponders and have higher power requirements to compensate for the loss of antenna gain. Also, the fixed spot beam transponders (assumed identical) are not matched to traffic requirements of the area served.

Another solution to the access problem involves the use of a steerable spot beam which can be rapidly scanned across the entire service region via a phased array antenna, thereby providing universal coverage.[6] When used in conjunction with a multitude of fixed spot beams, the resulting hybrid system has the advantages of frequency reuse, high antenna gain, and identical transponders, and hence is the most attractive proposal among those reported to date. A similar system which provides for beam scanning by appropriate excitation of feedhorn clusters has also been proposed.[7] However, such systems do not utilize the transponders efficiently, because of nonuniform traffic demands from the various areas covered.

III. TIME DIVISIDN MULTIBEAM SCANNING SATELLITE

To enable frequency reuse via a multibeam satellite system employing identical transponders, so that all transponders are used at maximum efficiency and a uniform grade of service is provided over the service area, we propose to generalize upon the scanning-beam approach.

Consider a satellite employing N identical wideband transponders, each with a capacity or throughput of C units. The diameter of the satellite antenna and the resulting beamwidth determine the number M of distinct footprints needed to provide service anywhere throughout the required service area. In general, M may be much greater than N, but in what follows we only require that $M \geq N$.

The system traffic can be represented by an $M \times M$ matrix $[t_{ij}]$ as shown:

$$[t_{ij}] = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1M} \\ t_{21} & t_{22} & \cdots & t_{2M} \\ & & \vdots & \\ t_{M1} & t_{M2} & \cdots & t_{MM} \end{bmatrix}. \quad (1)$$

The element t_{ij} represents the traffic originating in beam i and destined for somewhere in beam j. Each footprint might contain several ground stations, so t_{ij} represents the sum of the traffic from all stations within beam i which is directed to stations within beam j.

It is not necessary that the traffic matrix be symmetric, and a loopback feature is possible, i.e., we do not require $t_{ij} = t_{ji}$, nor do we require $t_{ii} = 0$. Of course, $t_{ij} \geq 0$.

Two requirements must be imposed on the traffic matrix $[t_{ij}]$. First, since the total capacity of the satellite is equal to NC (N transponders each of capacity C), we require that

$$T = \sum_{i=1}^{M} \sum_{j=1}^{M} t_{ij} \leq NC. \quad (2)$$

The second requirement is that the traffic originating from or destined for a particular beam should not exceed the capacity of one transponder, i.e.,

$$\text{Row sum } R_i = \sum_{j=1}^{M} t_{ij} \leq C \quad i = 1,2,\ldots M \quad (3)$$

$$\text{Column sum } S_j = \sum_{i=1}^{M} t_{ij} \leq C \quad j = 1,2,\ldots M. \quad (4)$$

The transponders are utilized with 100-percent efficiency when (2) is satisfied as an equality. This equation may be interpreted as establishing the minimum number N of transponders required. Conditions (3) and (4) are necessary because no two transponders can be connected to a common spot beam (either uplink or downlink) on a noninterfering basis.

If the total offered traffic equals the sum of the transponder capacities, we have the potential for 100-percent utilization. We will show that it is possible to interconnect the various uplink beams, transponders, and downlink beams such that this is achieved. We do this on a time-division basis by enabling each of the N transponders to access any of the M receive (uplink) antenna ports and any of the M transmit (downlink) antenna ports. Figure 1 shows the use of two $M \times N$ crossbar type switches which enable any required interconnection. Alternatively, the appropriate interconnections could be achieved by using N phased array antennas as shown in Fig. 2.

It remains to be shown that all the offered traffic can be allocated among the N transponders on a noninterfering basis, i.e., at any instant of time, the N transponder inputs are each connected to a different receive port, and the N transponder outputs are each connected to a dif-

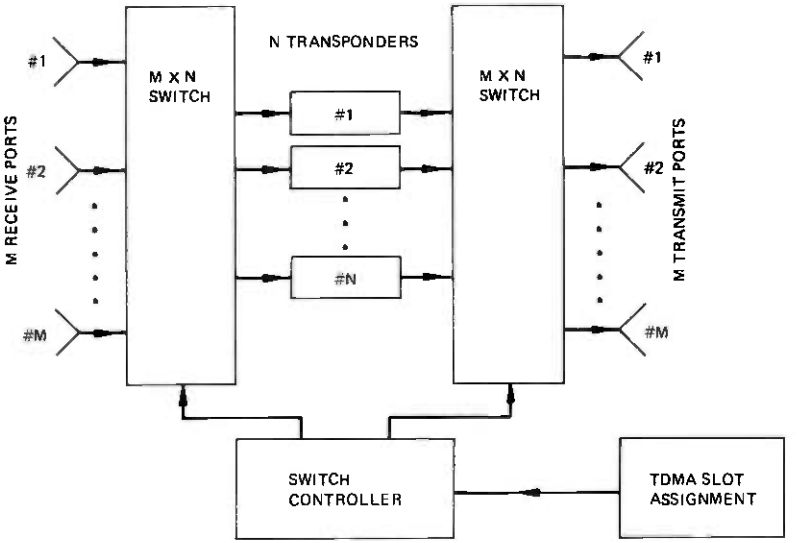


Fig. 1—Satellite communication subsystem for rapid TDMA scanning of multiple transponders using two $M \times N$ crossbar switches.

ferent transmit port. The theorem below guarantees that such an assignment is always possible.

Definition: A diagonal of a matrix $[t_{ij}]$ is a K-tuple $D = \{d_1, d_2, \ldots d_K\}$, where each member is a nonzero element of the matrix and no two elements appear in the same row or same column of the matrix. The length of the diagonal is K (the number of elements) and the diagonal is said to cover the K rows and K columns from which the elements are taken.

Theorem 1: In a traffic matrix $[t_{ij}]$ *for which* $T = \Sigma_{i=1}^{M}\Sigma_{j=1}^{M} t_{ij}$ *equals NC and for which no row or column sum exceeds C, a diagonal of length N exists which covers all rows and columns which sum to C exactly (if any).*

The proof of this theorem is somewhat lengthy and is presented in the appendix.

For convenience, we will assume that the elements t_{ij} of the traffic matrix are integers, representing the traffic as multiples of some basic unit such as, for example, one voice channel.

We shall assign traffic to the various transponders as follows: Let the TDMA frame consist of C time slots, each representing one unit of traffic. There are N such frames, one belonging to each transponder. In the traffic matrix $[t_{ij}]$, find any diagonal of length N which covers all rows and columns summing to C (if any). Theorem 1 guarantees this is always possible. From these N diagonal elements, extract one unit of traffic from each and assign one unit to each of the N transponders. Since the traffic assigned to the transponders (for this time slot) originates from different

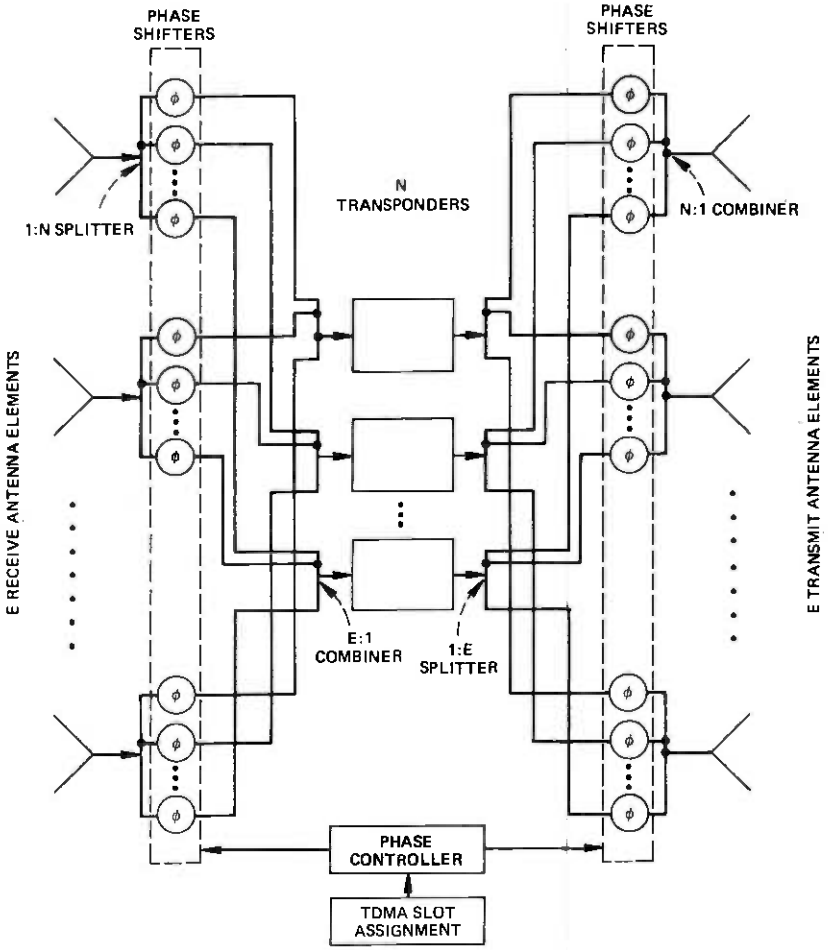


Fig. 2—Satellite communication subsystem for rapid TDMA scanning of multiple transponders using receive and transmit phased-array antennas. Each transponder can be steered independently to M transmit and M receive spot beam regions.

uplink beams and are directed to different downlink beams, then the traffic has been assigned on a noninterfering basis.

Now since N units of traffic have been removed from the matrix, the reduced matrix has a total traffic of $NC - N = N(C - 1)$ units. Furthermore, each transponder has $C - 1$ units of traffic carrying capacity left, and no row or column of the reduced matrix sums to more than $C - 1$. The latter is true because every row and column which summed to C in the original matrix has had one unit of traffic removed (because of the way the diagonal was constructed).

At this stage, we have the same situation as we started with, except $C - 1$ replaces C. By the same technique, we can assign another N units

of traffic to the next time slot in each transponder, and end up with a matrix with remaining traffic $N(C-2)$ in which no row or column sums to more than $C-2$. Each transponder has then $C-2$ time slots unallocated. Hence, we can repeat this procedure until all transponder time slots are used and no traffic remains unallocated.

Thus, the nonuniform demands of a traffic matrix can be met by N identical transponders each operating at 100-percent utilization efficiency. We also note that, although the method described was for a matrix for which eq. (2) was satisfied as an equality (i.e., $T = NC$), it also applies to a matrix for which $T < NC$, because we can always pad such a matrix with dummy traffic[8] until $T = NC$. The assignments corresponding to the dummy traffic can be ignored, and simply reflect the fact that the available transponder capacity exceeds the demand.

The assignments are not unique, and it may be possible to extract more than one unit of capacity per diagonal element at a time. This is desirable from a practical point of view as it minimizes the number of times the $M \times N$ switches have to be reconfigured during one frame period. To achieve this, it seems desirable to choose the N diagonal elements from large elements in the rows and columns with the largest sums, if possible. The maximum traffic extractable is $t = \min(t_1, t_2)$, where t_1 = smallest element on the diagonal and $C - t_2$ is the largest row or column sum among the rows and columns not covered by the diagonal.

As an example, consider the matrix below with $N = 3$ and $C = 13$.

		Downlink beam j				
	t_{ij}	1	2	3	4	R_i
Uplink	1	3	6	2	1	12
beam	2	6	4	0	0	10
i	3	0	1	6	2	9
	4	2	0	2	4	8
	S_j	11	11	10	7	$39 = T$

In Fig. 3 we show the successive reductions of the matrix as the traffic is assigned to transponders. The diagonal elements chosen are circled, and rows and columns which sum to (the reduced value of) C are marked with an asterisk.

The corresponding traffic assignments to the transponders are shown in Fig. 4. The switch must be reconfigured six times per frame for this solution.

IV. CONCLUSIONS

Although the system described has been presented in terms of subdividing the transponder capacity by time division, it is applicable to any other method of subdividing the transponder, e.g., by frequency division or by a combination of time and frequency division. In a frequency-division system, the smallest subdivision unit of capacity would

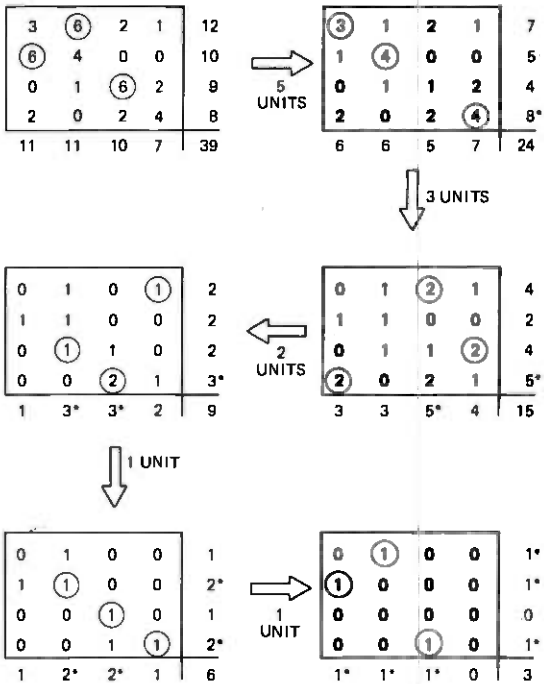


Fig. 3—Illustrative reduction of a 4 × 4 traffic matrix. The matrix contains 39 units of traffic, and there are three transponders, each of capacity 13.

usually be larger than for a time-division system, and transponder linearity would be an important consideration as far as crosstalk is concerned. In a time-division system, transponder nonlinearities are more tolerable.

For the system proposed, reliability of the transponders could be provided by the usual method of having a standby transponder for every transponder in use or perhaps sharing a standby with two operational transponders. However, an interesting alternative is to provide $N' > N$ transponders and use $M \times N'$ switches at input and output. In this way, failed transponders can be excluded by simply modifying the switching sequence, and we have a pool of spare transponders which can be used to supply replacements for any that fail.

The output switch at the satellite will generally operate at a high power level, and switching time may be a significant factor. In the phased-array realization, the equivalent problem is the time taken to steer the beam from one area to another. In either case, reducing the number of switch reconfigurations per frame will minimize any overhead time due to switching delays. Since the switching sequence, once decided upon, is not changed from frame to frame, a search for an efficient switching sequence is worthwhile.

In practice, other considerations besides switching delays would also

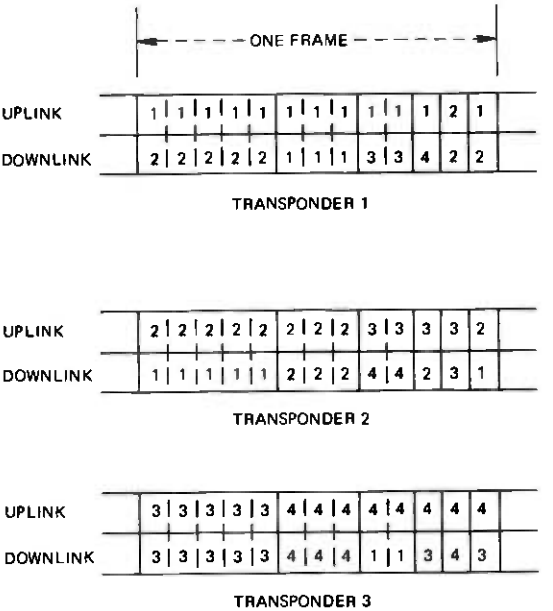


Fig. 4—TDMA frame assignment for the example of Fig. 3. The numbers 1 through 4 appearing within each frame correspond to the spot beam coverage areas 1 through 4 of the 4 × 4 traffic matrix.

be of importance. One additional consideration would be the interference between stations in adjacent beam-coverage areas. This interference would be reduced by ensuring that the stations did not transmit or receive during the same time slot, or by using different polarizations. Alternatively, adaptive interference cancellation can be performed for the phased-array implementation. Constraints imposed by considerations such as these, however, may result in more than the minimum number of transponders being required to transmit the traffic.

We have described a system which enables efficient utilization of transponder capacity, while at the same time providing service over a wide area with a uniform grade of service, identical transponders in the satellite which can be coupled with a very efficient standby method for maintaining transponder reliability, and a system which can be adapted to changing traffic demands by simply altering the switching sequence at the satellite.

APPENDIX

Proof of Theorem 1

The definition of a diagonal and its length are found in the main text.

The proof is approached by first establishing a number of lemmas which can then be used in the proof of the theorem.

The theorem as proved here is slightly more general than that presented in the main text, since the existence of a diagonal of the required type is demonstrated for $(N-1)C < T \leq NC$ rather than just for $T = NC$.

Lemma 1: If in a matrix $[t_{ij}]$ *there exist diagonals* $D_1 = \{a_1, a_2, \ldots a_N\}$ *of length* N *and* $D_2 = \{b_1, b_2, \ldots b_L\}$ *of length* $L \geq N$, *then it is possible to construct:*

(*i*) *A diagonal* D_3 *of length* L *which covers all the rows and columns covered by* D_1.

(*ii*) *A diagonal* D_4 *of length* $\geq L$ *which covers the rows* D_1 *and the columns of* D_2.

(*iii*) *A diagonal* D_5 *of length* $\geq L$ *which covers the columns of* D_1 *and the rows of* D_2.

Proof:

(*i*) Form disjoint sets S_m, $m = 1,2, \ldots$ from the elements of D_1 and D_2 as follows:

(*a*) Once an element has been allocated to a set, it is not considered further.

(*b*) To form set S_m, choose an initial element from among those not yet allocated to a set.

(*c*) If b_j is on the same row or column as an $a_i \in S_m$, then $b_j \in S_m$.

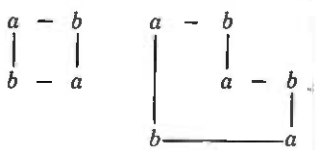
(*d*) If a_r is on the same row or column as a $b_t \in S_m$ then $a_r \in S_m$.

(*e*) Continue adding elements to S_m using (*c*) and (*d*) until no more can be added. If unallocated elements still remain, form a set S_{m+1} starting at step (*b*).

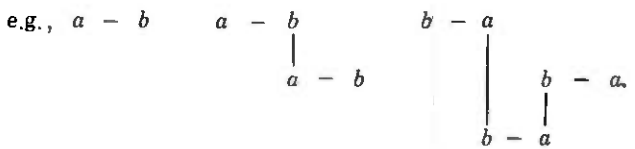
(*ii*) The sets have the property that an element from one set cannot share a row or column with an element from another set.

(*iii*) The sets are of the following varieties:

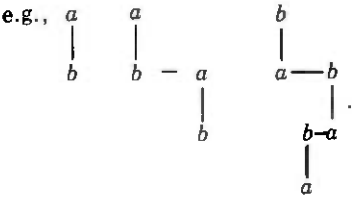
V1: Sets with equal number of elements from D_1 and D_2 which cover the same rows and columns e.g., a, b coincident,



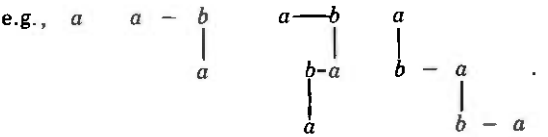
V2: Sets with equal numbers of elements from D_1 and D_2 which cover the same rows, but not the same columns,



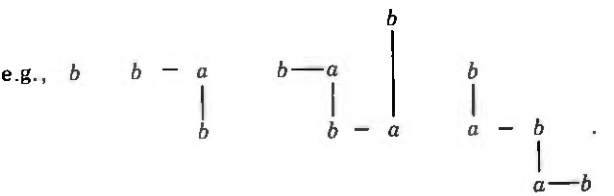
V3: Sets with equal numbers of elements from D_1 and D_2 which cover the same columns, but not the same rows,



V4: Sets with one more element from D_1 than from D_2, in which the elements from D_1 cover all the rows and columns of the elements from D_2,



V5: Sets with one more element from D_2 than from D_1, in which the elements from D_2 cover all the rows and columns of the elements from D_1,



Not all varieties need be present, but it is easily seen that there must be at least $L - N$ sets in V5.

Let $V5' = \{\text{any } L - N \text{ sets from } V5\}$.

Then

$$D_3 = \{a_i \notin V_5', b_j \in V_5'\}$$

$$D_4 = \{a_i \in V_1, b_j \in V_2, a_i \in V_3, a_i \in V_4, b_j \in V_5\}$$

$$D_5 = \{a_i \in V_1, a_i \in V_2, b_j \in V_3, a_i \in V_4, b_j \in V_5\}$$

are diagonals of the type required.

Lemma 2: If the maximum length of a diagonal in a matrix $[t_{ij}]$ *is* N, *and the row and column sums do not exceed* C, *then*

$$T_N \triangleq \sum_i \sum_j t_{ij} \leq NC.$$

Proof: If all the nonzero elements are in at most N rows (or columns), then by summing over these rows (or columns) we obtain $T_N \leq NC$.

We therefore need only consider the case where there are more than N rows and more than N columns which contain nonzero elements.

Let $D_1 = \{a_1, a_2, \ldots a_N\}$ be a diagonal of length N covering columns $j_1, j_2, \ldots j_N$. There must be another column j_{N+1} with a nonzero element x. One of the elements of D_1, say a_r, must be on the same row as x, otherwise $\{a_1, a_2, \ldots a_N, x\}$ would form a diagonal of length $N + 1$.

Remove the row containing x and a_r from the matrix. We will show that the reduced matrix has a maximum diagonal length of $N - 1$.

In the reduced matrix we have $D_1' = \{a_1, \ldots a_{r-1}, a_{r+1}, \ldots a_N\}$ of length $N - 1$ and suppose there is also a diagonal $D_2 = \{b_1, b_2, \ldots b_N\}$ of length N (in the reduced matrix). By Lemma 1 we can find from D_1' and D_2 a diagonal D_3 of length N which covers all the rows and columns covered by D_1'. Since $N - 1$ of the elements of D_3 are in columns $j_1 \ldots j_{r-1}, j_{r+1} \ldots j_N$, then both columns j_r, j_{N+1} cannot be covered by the Nth element. Hence, D_3 augmented by either a_r or x would form a diagonal of length $N + 1$ in the original matrix. Hence, no such diagonal D_2 exists.

The reduced matrix satisfies the same conditions as the original matrix except $N - 1$ replaces N.

$$T_N = T_{N-1} + R \leq T_{N-1} + C,$$

where

$$T_N = \text{sum of elements in original matrix}$$

$$T_{N-1} = \text{sum of elements in reduced matrix}$$

$$R = \text{sum of elements in row removed} \leq C.$$

Hence, $T_N \leq NC$ if $T_{N-1} \leq (N - 1)C$. Since $T_0 = 0$, an inductive argument establishes the result.

Lemma 3: In a matrix $[t_{ij}]$ for which the row and column sums do not exceed C, and for which $T \triangleq \Sigma_i \Sigma_j t_{ij}$ satisfies $(N - 1)C < T \leq NC$ for some integer N, there exists a diagonal of length $L \geq N$.

Proof: Let L be the maximum diagonal length.

By Lemma 2, $T \leq LC$

But $T > (N - 1)C$, so $L > N - 1$,

Hence, a diagonal of length $L \geq N$ exists.

Lemma 4: In a matrix $[t_{ij}]$ for which the row and column sums do not exceed C, and for which $T = \Sigma_i \Sigma_j t_{ij}$ satisfies $(N - 1)C < T \leq NC$ for some integer N, there exists a diagonal of length $N'' \geq N$ which covers all rows and columns which sum to C exactly.

Proof: The submatrix consisting only of the P rows which sum to C has, by Lemma 3, a diagonal $D_1 = \{a_1, a_2, \ldots, a_P\}$ of length P, because its el-

ements sum to PC. This diagonal covers all the P rows summing to C. Note that $P \leq N$.

By Lemma 3, the original matrix has a diagonal $D_2 = \{b_1, b_2, \ldots, b_N\}$ of length N. By Lemma 1, we can construct from D_1 and D_2 a diagonal D_2' of length $N' \geq N$ which covers all the columns of D_2 and all the rows of D_1.

Let $D_1' = \{a_1', \ldots, a_Q'\}$ be a diagonal of length Q of the submatrix consisting only of the Q columns which sum to C exactly. Note that $Q \leq N \leq N'$.

Then by Lemma 1 we can construct from D_1' and D_2' a diagonal D_2'' of length $N'' \geq N'$ which covers all the columns of D_1' and all the rows of D_2' (and hence all the rows of D_1).

Hence, D_2'' covers all the rows and columns which sum to C exactly.

Theorem: In a matrix $[t_{ij}]$ for which the row and column sums do not exceed C, and for which $T \triangleq \Sigma_i \Sigma_j t_{ij}$ satisfies $(N-1)C < T \leq NC$ for some integer N, there exists a diagonal of length N which covers all rows and columns which sum to C exactly.

Proof: By Lemma 4, a diagonal $D_1 = \{a_1, a_2, \ldots a_L\}$ of length $L \geq N$ exists which covers the P rows and Q columns which sum to C exactly.

Divide D_1 into disjoint subdiagonals S_1, S_2, and S_3 with L_1, L_2, and L_3 elements, respectively, with $L_1 + L_2 + L_3 = L$.

S_1 = {elements of D_1 in both a row and a column summing to C}

S_2 = {elements of D_1 in either a row or a column summing to C, but not both}

S_3 = {elements of D_1 in neither a row nor a column summing to C}

If $L_1 + L_2 \leq N$, then a diagonal consisting of the $L_1 + L_2$ elements from S_1 and S_2 plus any $N - L_1 - L_2$ elements of S_3 is a diagonal of length N covering all rows and columns summing to C.

Hence, we need only consider the case $L_1 + L_2 > N$. Note that $P + Q = 2L_1 + L_2 > N$ also.

Consider the submatrix consisting of P rows and Q columns containing only those elements (including zero elements) which lie in both a row and a column summing to C. We know S_1 is a diagonal of length L_1 of this submatrix. The sum of the elements in the submatrix = T' = {sum of elements of original matrix in those P rows} − {sum of elements of original matrix in the same P rows, but which do not lie in the columns summing to C}. Hence, $T' \geq (P + Q - N)C$, since the first sum is PC and the second cannot exceed $(N - Q)C$.

By Lemma 3, the submatrix has a diagonal S_4 of length $P + Q - N = 2L_1 + L_2 - N > L_1$ (since $L_1 + L_2 > N$). By Lemma 1 we can construct from S_1 and S_4 a diagonal S_1' of length $2L_1 + L_2 - N$, which covers all the rows and columns covered by S_1.

Now S_1' covers $L_1 + L_2 - N$ rows summing to C and $L_1 + L_2 - N$ columns summing to C not covered by S_1 and which must have been cov-

ered by S_2. Hence, form a new subdiagonal S_2' by deleting from S_2 the elements in these rows and columns. Then S_1' and S_2' cover different rows and columns.

Now S_1' is a diagonal of length $L_1' = 2L_1 + L_2 - N$ and S_2' is a diagonal of length $L_2' = L_2 - 2(L_1 + L_2 - N)$. Thus, the elements of S_1' and S_2' form a diagonal which covers all the rows and columns which sum to C and its length is $L_1' + L_2' = N$.

REFERENCES

1. L. C. Tillotson, "A Model of a Domestic Satellite Communication System," B.S.T.J., *47*, No. 10 (December 1968), pp. 2111–2137.
2. R. Cooperman and W. G. Schmidt, "Satellite Switched SDMA and TDMA Systems for Wideband Multibeam Satellite," ICC Conference Record, 1973.
3. A. S. Acampora, "Reliability Considerations for Multiple Spot Beam Communication Satellites," B.S.T.J., *56*, No. 4 (April 1977), pp. 575–596.
4. D. O. Reudink, A. S. Acampora, and Y. S. Yeh, "Spectral Reuse in 12 GHz Satellite Communication Systems," ICC Conference Record, 1977.
5. H. W. Arnold, "An Efficient Digital Satellite Technique for Serving Users of Differing Capacities," ICC Conference Record, 1977.
6. D. O. Reudink and Y. S. Yeh, "A Scanning Spot Beam Satellite System," B.S.T.J., *56*, No. 8, (October 1977), pp 1549–1560.
7. E. A. Ohm, in preparation.
8. Y. Ito et al., "Analysis of a Switch Matrix for an SS/TDMA System," Proc. IEEE, *65*, No. 3 (March 1977), pp. 411–419.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 8, October 1978
Printed in U.S.A.

On Blocking Probabilities for a Class of Linear Graphs

By F. R. K. CHUNG and F. K. HWANG

(Manuscript received October 20, 1977)

In a t-stage linear graph, all vertices are arranged in a sequence of stages such that any edge goes between a vertex in stage i and a vertex in stage i + 1 for some i. Lee first proposed the use of t-stage linear graphs for studying the blocking performance of t-stage switching networks. In his model, each edge is in either of two states, busy or idle, and the states of the edges are independent. Furthermore, an edge connecting a vertex in stage i to a vertex in stage i + 1 has the constant probability p_i of being busy. In the current paper, we use Lee's model to compare the blocking probabilities of different linear graphs. In particular, a t-stage linear graph is said to be superior to another t-stage linear graph if the blocking probability of the former never exceeds that of the latter for any choice of the p_i. For a class of linear graphs known as SP-canopies, we give simple necessary and sufficient conditions that one t-stage linear graph is superior to another.

I. INTRODUCTION

We consider a t-stage linear graph with a source (the vertex of the first stage) and a sink (the vertex of the last stage). All vertices are lined up in a sequence of stages such that any edge goes from a vertex in stage i to a vertex in stage $i + 1$ for some i, while each edge can be in either of the two states, *busy or idle*. The linear graph is said to be *blocked* if every path joining the source and the sink contains a busy edge. Lee[1] first proposed the concept of linear graphs in his study of the blocking performance of switching networks. We use Lee's model and follow his independence assumptions, namely, that the probabilities of occupancy for edges being busy in successive stages are independent. Thus, we may assume that any edge connecting a vertex in stage i with a vertex in stage $i + 1$ has some probability p_i of being busy for $1 \leq i \leq t - 1$. The sequence $(p_1,p_2,\cdots,p_{t-1})$ will be called *link occupancies* of the t-stage linear graph. A linear graph is said to be *superior* to another if, for any given

link occupancies, the blocking probability of the first graph does not exceed that of the second graph.

In this paper, we restrict ourselves to a special kind of linear graph, called an *SP-canopy*. By definition, the smallest SP-canopy is a series combination of two edges. Any other *SP-canopy* is either a *parallel* combination of two smaller SP-canopies or a *series* combination of a smaller SP-canopy and an edge. For readers familiar with graph-theoretic terminology, an SP-canopy can be viewed as the union of two rooted trees with identified sets of terminal nodes such that the union is a planar graph (see Fig. 1a for an example).

Let e be an edge from a vertex **a** in stage i to a vertex **b** in stage $i + 1$. We define $\lambda(e)$ to be the ratio of the outdegree of **a** to the indegree of **b**. If all $\lambda(e)$, where e ranges over all edges between stage i and stage $i + 1$ (for a fixed i), have the same value λ_i for $1 \leq i \leq t - 1$, then this linear graph is said to be a *regular* linear graph (see Fig. 1b). Thus, a regular linear graph is associated with a unique degree sequence $(\lambda_1,\lambda_2,\cdots,\lambda_{t-1})$. In the case in which the regular linear graph is an SP-canopy, it can be uniquely represented by the degree sequence. Define $\lambda^* = \max_{1\leq i\leq t-1} \{\lambda_1\lambda_2\cdots\lambda_i\}$. It is easy to verify that λ^* is just the number of distinct paths from the source to the sink. A regular SP-canopy is said to be a *symmetric* SP-canopy if $\lambda_i\lambda_{t-i} = 1$ for $1 \leq i \leq t - 1$. Thus, the degree sequence of a symmetric SP-canopy can be written, abbreviated as $(\lambda_1,\cdots,\lambda_{\lfloor (t-1)/2 \rfloor})$, when $\lfloor x \rfloor$ denotes the greatest integer not exceeding x. The linear graph

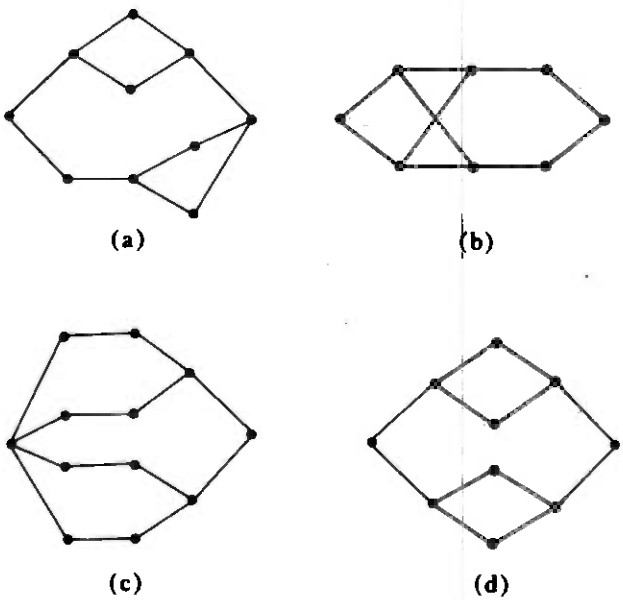


Fig. 1—(a) An SP-canopy. (b) A regular linear graph. (c) A regular SP-canopy. (d) A symmetric SP-canopy.

in Fig. 1c is a regular SP-canopy, and the linear graph in Fig. 1d is a symmetric SP-canopy.

We say a degree sequence $(\lambda_1,\lambda_2,\cdots,\lambda_{t-1})$ *majorizes* another degree sequence $(\lambda_1', \lambda_2',\cdots,\lambda_{t-1}')$ if and only if $\lambda_1\lambda_2\cdots\lambda_i \geq \lambda_1'\lambda_2'\cdots\lambda_i'$ for every i, $1 \leq i \leq t-1$. If we consider the set S_{t,λ^*} of all regular canopies with a fixed number t of stages and a fixed number λ^* of distinct paths, we see that $(\lambda^*,1,\cdots,1,(\lambda^*)^{-1})$ majorizes the degree sequence of any other regular SP-canopy in S_{t,λ^*}. In Ref. 2, comparisons are made involving all symmetric SP-canopies with fixed t and λ^* for t odd, with the conclusion that the symmetric SP-canopy with the degree sequence $\lambda^*,1,\cdots,1((t-1)/2$ 1s) is superior to all the others and the symmetric SP-canopy with degree sequence $1,\cdots,1,\lambda^*$ $((t-1)/2$ 1s) is inferior to all the others. In this paper, we prove a stronger and more general result which says that one regular SP-canopy is superior to another one if and only if the degree sequence of the first one majorizes the degree sequence of the second one.

II. SYMMETRIC SP-CANOPIES

In this section, we study symmetric SP-canopies with degree sequences of the form $(\lambda_1,\lambda_2,\cdots,\lambda_{\lfloor (t-1)/2 \rfloor})$.

First, we prove a few auxiliary lemmas needed in the proof of the main result.

Lemma 1: Define

$$F(x) = (1 - a(1 - b^x))^{k/x}, \text{ where } 0 \leq a, b \leq 1.$$

Then $F(x)$ is monotone nondecreasing for $1 \leq x \leq k$.

Proof: We consider several cases:

Case 1: $0 < a, b < 1$.

$$\frac{dF}{dx}(x) = (1 - a(1 - b^x))^{k/x} \cdot \left(-\frac{k}{x^2} \ln(1 - a(1 - b^x)) + \frac{kab^x \ln b}{x(1 - a(1 - b^x))} \right).$$

We define

$$G(a) = -\frac{1}{x} \ln (1 - a(1 - b^x)) + \frac{ab^x \ln b}{1 - a(1 - b^x)}.$$

It is easy to verify that $G(0) = G(1) = 0$. Furthermore, by setting $dG/da = 0$, we obtain the unique solution a_0 which satisfies

$$a_0 = \frac{1}{1 - b^x} + \frac{xb^x \ln b}{(1 - b^x)^2}.$$

Since $d^2G/da^2\,(a_0) < 0$, a_0 is indeed a maximum. Therefore, $G(a) > 0$ for all $0 < a < 1$. Thus dF/dx is positive for $0 < a, b < 1$.

Case 2: $a = 0$ or $b = 1$.

We have $F(x) = 1$ and $dF/dx = 0$.

Case 3: $a = 1$, then

$$F(x) = b^k, \; dF/dx = 0.$$

Case 4: $a \neq 0, 1$ and $b = 0$

$$F(x) = (1-a)^{k/x}.$$

$$\frac{dF}{dx}(x) = -\frac{k}{x^2}(1-a)^{k/x}\ln(1-a) > 0.$$

In any case, dF/dx is nonnegative. Therefore, Lemma 1 is proved.

For a given vector $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_{t-1})$, $1 \geq \alpha_i \geq 0$, and a sequence of positive real numbers $(\beta_1, \cdots, \beta_{\lfloor (t-1)/2 \rfloor})$, we define the following function:

$$f(\beta_{\lfloor (t-1)/2 \rfloor}) = \begin{cases} (1-\alpha_{(t-1)/2} \cdot \alpha_{(t+1)/2})^{\beta_{\lfloor (t-1)/2 \rfloor}} & \text{if } t \text{ is odd} \\ (1-\alpha_{t/2})^{\beta_{\lfloor (t-1)/2 \rfloor}} & \text{if } t \text{ is even.} \end{cases}$$

and

$$f(\beta_i, \beta_{i+1}, \cdots, \beta_{\lfloor (t-1)/2 \rfloor}) = (1 - \alpha_i \alpha_{t-i}(1 - f(\beta_{i+1}, \cdots, \beta_{\lfloor (t-1)/2 \rfloor})))^{\beta_i}$$

for $i = 1, 2, \cdots, \lfloor (t-1)/2 \rfloor - 1$.

Lemma 2: If the sequence $(\beta_1, \beta_2, \cdots, \beta_{\lfloor (t-1)/2 \rfloor})$ *majorizes the sequence* $(\beta_1', \beta_2', \cdots, \beta_{\lfloor (t-1)/2 \rfloor}')$, *then we have*

$$f(\beta_1, \beta_2, \cdots, \beta_{\lfloor (t-1)/2 \rfloor}) \leq f(\beta_1', \beta_2', \cdots, \beta_{\lfloor t-1/2 \rfloor}').$$

for any vector $(\alpha_1, \cdots, \alpha_{t-1})$ *satisfying* $1 \geq \alpha_i \geq 0$ *for all* i.

Proof: It is easily checked that Lemma 2 is true for $t = 2$ and 3. By the induction assumption, we have

$$f(\beta_1\beta_2/\beta_1', \beta_3, \cdots, \beta_{\lfloor (t-1)/2 \rfloor}) \leq f(\beta_2', \beta_3', \cdots, \beta_{\lfloor (t-1)/2 \rfloor}'),$$

since $(\beta_1\beta_2/\beta_1', \beta_3, \cdots, \beta_{\lfloor (t-1)/2 \rfloor})$ majorizes the sequence

$$(\beta_2', \beta_3', \cdots, \beta_{\lfloor (t-1)/2 \rfloor}').$$

It is also clear that the following holds:

$$f(\beta_1', \beta_1\beta_2/\beta_1', \beta_3, \cdots, \beta_{\lfloor (t-1)/2 \rfloor}) \leq f(\beta_1', \beta_2', \cdots, \beta_{\lfloor (t-1)/2 \rfloor}').$$

Therefore, it suffices to show

$$f(\beta_1, \beta_2, \cdots, \beta_{\lfloor (t-1)/2 \rfloor}) \leq f(\beta_1', \beta_1\beta_2/\beta_1', \beta_3, \cdots, \beta_{\lfloor (t-1)/2 \rfloor}).$$

Now we have

$$\begin{aligned} f(\beta_1, \beta_2, \cdots, \beta_{\lfloor (t-1) \rfloor}) &= (1 - \alpha_1\alpha_{t-1}(1 - f(\beta_2, \cdots, \beta_{\lfloor (t-1)/2 \rfloor})))^{\beta_1} \\ &= (1 - \alpha_1\alpha_{t-1}(1 - (1 - \alpha_2\alpha_{t-2}(1 - f(\beta_3, \cdots, \beta_{\lfloor (t-1)/2 \rfloor})))^{\beta_2}))^{\beta_1}. \end{aligned}$$

We set

$$a = \alpha_1\alpha_{t-1},$$

$$b = 1 - \alpha_2\alpha_{t-2}(1 - f(\beta_3, \cdots, \beta_{\lfloor (t-1)/2 \rfloor})),$$

$$k = \beta_1\beta_2.$$

Then we have

$$f(\beta_1,\beta_2,\cdots\beta_{\lfloor (t-1)/2\rfloor}) = (1 - a(1 - b^{\beta_2}))^{k/\beta_2}.$$

Similarly,

$$f(\beta_1',\beta_1\beta_2/\beta_1',\beta_3,\cdots,\beta_{\lfloor (t-1)/2\rfloor}) = (1 - a(1 - b^{\beta_1\beta_2/\beta_1'}))^{\beta_1' k/\beta_1\beta_2}.$$

Since $\beta_2 \leq \beta_1\beta_2/\beta_1'$, by Lemma 1 we have the following:

$$f(\beta_1,\beta_2,\cdots,\beta_{\lfloor (t-1)/2\rfloor}) \leq f(\beta_1',\beta_1\beta_2/\beta_1',\beta_3,\cdots,\beta_{\lfloor (t-1)/2\rfloor}) \leq f(\beta_1',\beta_2',\cdots,\beta_{\lfloor (t-1)/2\rfloor}'),$$

and Lemma 2 is proved.

Theorem 1: Consider two t-stage symmetric SP-canopies with degree sequences $(\lambda_1,\cdots,\lambda_{\lfloor (t-1)/2\rfloor})$ *and* $(\lambda_1',\cdots,\lambda_{\lfloor (t-1)/2\rfloor}')$ *respectively. Then the first linear graph is superior to the second if and only if the sequence* $(\lambda_1,\cdots,\lambda_{\lfloor (t-1)/2\rfloor})$ *majorizes the sequence* $(\lambda_1',\cdots,\lambda_{\lfloor (t-1)/2\rfloor}')$.

Proof: If we let $\alpha_i = 1 - p_i$, for $1 \leq i \leq t-1$, in Lemma 2, it is easy to see that the blocking probability $P(\lambda_1,\cdots,\lambda_{\lfloor (t-1)/2\rfloor})$ for the symmetric SP-canopy with degree sequence $(\lambda_1,\cdots,\lambda_{\lfloor (t-1)/2\rfloor})$ has the same value as $f(\lambda_1,\cdots,\lambda_{\lfloor (t-1)/2\rfloor})$. Thus, the fact that $(\lambda_1,\cdots,\lambda_{\lfloor (t-1)/2\rfloor})$ majorizes $(\lambda_1',\cdots,\lambda_{\lfloor (t-1)/2\rfloor}')$ implies the symmetric SP-canopy with degree sequence $(\lambda_1,\cdots,\lambda_{\lfloor (t-1)/2\rfloor})$ is superior to the symmetric SP-canopy with degree sequence $(\lambda_1',\cdots,\lambda_{\lfloor (t-1)/2\rfloor}')$.

We also want to show that, if the symmetric SP-canopy with degree sequence $(\lambda_1,\cdots,\lambda_{\lfloor (t-1)/2\rfloor})$ is superior to the symmetric SP-canopy with degree sequence $(\lambda_1',\cdots,\lambda_{\lfloor (t-1)/2\rfloor}')$, then it is necessary to have $(\lambda_1,\cdots,\lambda_{\lfloor (t-1)/2\rfloor})$ majorizing $(\lambda_1',\cdots,\lambda_{\lfloor (t-1)/2\rfloor}')$. Suppose, on the contrary, that there exists an integer k, $1 \leq k \leq \lfloor (t-1)/2\rfloor$ such that $\Pi_{i=1}^{k} \lambda_i < \Pi_{i=1}^{k} \lambda_i'$. We consider the link occupancies $(p_1,\cdots,p_{t-1})$, where $p_i = p_{t-i} = 1 - \epsilon$ if $1 \leq i \leq k$ and $p_i = p_{t-i} = 0$ if $k \leq i \leq \lfloor (t-1)/2\rfloor$. Then for ϵ sufficiently small, we have

$$\begin{aligned} P(\lambda_1,\cdots,\lambda_{\lfloor (t-1)/2\rfloor}) &= P(\lambda_1,\cdots,\lambda_k) \\ &= (1 - \epsilon^2(1 - P(\lambda_2,\cdots,\lambda_k))^{\lambda_1} \\ &= 1 - (\lambda_1\epsilon^2 + 0(\epsilon^4))(1 - P(\lambda_2,\cdots,\lambda_k)) \\ &= 1 - \epsilon^{2k} \prod_{i=1}^{k} \lambda_i + 0(\epsilon^{2k+2}). \end{aligned}$$

Similarly, $P(\lambda_1',\cdots,\lambda_{\lfloor (t-1)/2\rfloor}')$ is approximately $1 - \epsilon^{2k} \Pi_{i=1}^{k} \lambda_i'$. Thus we have $P(\lambda_1',\cdots,\lambda_{\lfloor (t-1)/2\rfloor}') < P(\lambda_1,\cdots,\lambda_{\lfloor (t-1)/2\rfloor})$. This contradicts the fact that the symmetric SP-canopy with degree sequence $(\lambda_1,\cdots,\lambda_{\lfloor (t-1)/2\rfloor})$ is superior to the symmetric SP-canopy with degree sequence $(\lambda_1',\cdots,\lambda_{\lfloor (t-1)/2\rfloor}')$, and Theorem 1 is proved.

III. REGULAR SP-CANOPIES

From definitions in Section I, it can be easily seen that a regular SP-canopy is either a parallel combination of copies of a smaller regular SP-canopy or series combination of a smaller regular SP-canopy and an edge. A regular SP-canopy has many special properties, described in the following lemma.

Lemma 3. Let $(\lambda_1,\cdots,\lambda_{t-1})$ *be the degree sequence of a regular SP-canopy* G.

(i) $\lambda_1\lambda_{t-1}$ *is either an integer or the reciprocal of an integer. If* $\lambda_1\lambda_{t-1}$ *is an integer,* G *is a parallel combination of copies of a smaller regular SP-canopy as shown in Fig. 2a, where* G' *has degree sequence* $(\lambda_1\lambda_{t-1},\lambda_2,\cdots,\lambda_{t-2})$. *If* $\lambda_1\lambda_{t-1}$ *is the reciprocal of an integer,* G *is a parallel combination of copies of a small regular SP-canopy as shown in Fig. 2b where* G'' *has degree sequence* $(\lambda_2,\cdots,\lambda_{t-2},\lambda_{t-1}\lambda_1)$.

(ii) $\prod_{i=1}^{t-1} \lambda_i = 1$.

(iii) *If* $\lambda_k > 1$ *for some* k, $1 \le k < t-1$, *then* $\lambda_i \ge 1$ *for all* $i < k$. *If* $\lambda_{k'} < 1$ *for some* k', $1 < k' \le t-1$, *then* $\lambda_1 \le 1$ *for all* $i > k'$.

Proof: Since G is a regular SP-canopy, the configuration of G can be easily shown to be either as in Fig. 2a or as in Fig. 2b. If G is as in Fig. 2a, G is a parallel combination of k copies of a regular SP-canopy (for some k) which is a series combination of G' and an edge. Let G' have degree sequence $(\lambda_1',\lambda_2',\cdots,\lambda_{t-2}')$. It is clear that the degree sequence of G is $(k\lambda_1',\lambda_2',\cdots,\lambda_{t-2}',k^{-1})$ and $\lambda_1\lambda_{t-1} = (k\lambda_1')k^{-1} = \lambda_1'$ is an integer. If G is as in Fig. 2b, G is a parallel combination of k' copies of a regular SP-canopy (for some k') which is a series combination of an edge and G''. Let G'' have degree sequence $(\lambda_1'',\lambda_2'',\cdots,\lambda_{t-2}'')$. It is easy to see that the degree sequence of G is $(k',\lambda_1'',\lambda_2'',\cdots\lambda_{t-2}''(k')^{-1})$ and $\lambda_1\lambda_{t-1} = k'(\lambda_{t-2}''(k')^{-1}) = \lambda_{t-2}''$ is the reciprocal of an integer. Since one of the two cases must occur, (i) is proved.

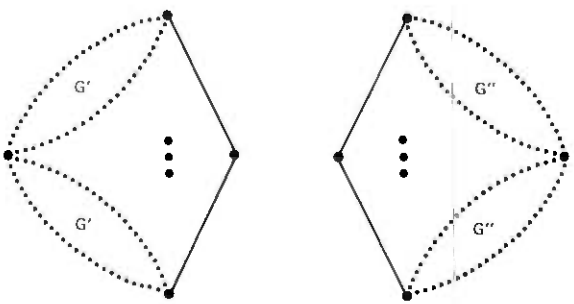


Fig. 2—Regular SP-canopy G.

We may assume without loss of generality that G is as in Fig. 2a. By the induction assumption, we have
$\Pi_{i=1}^{t-2} \lambda_1' = 1$. Thus, we have

$$\prod_{i=1}^{t-1} \lambda_i = k\lambda_1' \left(\prod_{i=2}^{t-2} \lambda_i' \right) k^{-1} = \prod_{i=1}^{t-2} \lambda_i' = 1,$$

and (ii) is proved.

Finally, λ_k being an integer implies $\lambda_k' = \lambda_k$ is an integer if $k > 1$. By the induction assumption, $\lambda'_j, j = 1,\cdots,k-1$, is an integer. Therefore, $\lambda_1 = k\lambda_1'$ is an integer and $\lambda_j = \lambda'_j, j = 2,\cdots,k$ integer. The other half of (iii) can be similarly verified. This proves Lemma 3.

We note that (ii) holds for the degree sequence of any regular linear graphs. However, (i) and (iii) are not true for series-parallel, regular linear graphs.

For a given vector, $\alpha = (\alpha_1,\cdots,\alpha_{t-1})$, $0 \le \alpha_i \le 1$ and positive real numbers $(\beta_1,\cdots,\beta_{t-1})$, we define the following function g by

$$g(\beta_1,\cdots,\beta_{t-1};\alpha) = \begin{cases} (1-\alpha_1(1-g(\beta_2,\cdots,\beta_{t-1}\beta_1;\alpha_{t-2,2})))^{\beta_1} \text{ if } \beta_1\beta_{t-1} \le 1, \\ (1-\alpha_{t-1}(1-g(\beta_1\beta_{t-1},\beta_2,\cdots,\beta_{t-2};\alpha_{t-2,1})))^{1/\beta_{t-1}} \text{ otherwise,} \end{cases}$$

where $\alpha_{k,i}$ is the vector $(\alpha_i,\cdots,\alpha_{i+k-1})$ and $g(\beta_i;\alpha_{1,i}) = 1 - \alpha_i$. Note that

$$g(\beta_1,\cdots,\beta_{t-1};\alpha) = g(\beta_{t-1}^{-1},\beta_{t-2}^{-1},\cdots,\beta_2^{-1},\beta_1^{-1};\bar{\alpha})$$

where

$$\bar{\alpha} = (\alpha_{t-1},\alpha_{t-2},\cdots,\alpha_2,\alpha_1).$$

Lemma 4: For any vector $\alpha = (\alpha_1,\cdots,\alpha_{t-1})$, if the sequence $(\beta_1,\beta_2,\cdots,\beta_{t-1})$ majorizes the sequence $(\beta_1',\beta_2',\cdots\beta_{t-1}')$, then we have

$$g(\beta_1,\beta_2,\cdots,\beta_{t-1};\alpha) \le g(\beta_1',\beta_2',\cdots,\beta_{t-1}';\alpha).$$

Proof: It is easy to see that Lemma 4 is true for $t = 2$. When $t = 3$, we have

$$g(\beta_1,\beta_2,;\alpha) = (1-\alpha_1\alpha_2)^{\beta_1} \le (1-\alpha_1\alpha_2)^{\beta_1'} = g(\beta_1',\beta_2';\alpha).$$

It suffices to consider $t \ge 4$. We note that
$(\beta_1,\beta_2,\cdots,\beta_{t-1})$ majorizes $(\beta_1',\cdots,\beta_{t-1}')$ if and only if $(\beta_{t-1}^{-1},\cdots,\lambda_2^{-1},\beta_1^{-1})$ majorizes $((\beta_{t-1}')^{-1},\cdots,(\beta_2')^{-1},(\beta_1')^{-1})$.
Therefore we may assume, without loss of generality, that $\beta_1\beta_{t-1} \le 1$ (we may consider the inverse sequence otherwise). Then we have

$$g(\beta_1,\cdots,\beta_{t-1};\alpha) = (1-\alpha_1(1-g(\beta_2,\cdots,\beta_{t-2},\beta_{t-1}\beta_1;\alpha_{t-2,2})))^{\beta_1}.$$

Let us consider several possibilities for $\beta_1'\beta_{t-1}'$.
Case (i): $\beta_1'\beta_{t-1}' \le 1$.
Since $\beta_1' \le \beta_1$, and $\beta_1'\beta_{t-1} \le 1$, we have

$$g(\beta_1',\beta_1\beta_2/\beta_1',\beta_3,\cdots,\beta_{t-1};\alpha) = (1-\alpha_1(1-g(\beta_1\beta_2/\beta_1',\beta_3,\cdots,\beta_{t-1}\beta_1';\alpha_{t-2,2})))^{\beta_1'}.$$

We want to show that

$$g(\beta_1,\cdots,\beta_{t-1};\alpha) \le g(\beta_1',\beta_1\beta_2/\beta_1',\beta_3,\cdots,\beta_{t-1};\alpha).$$

We note that $\beta_2(\beta_{t-1}\beta_1) = (\beta_1\beta_2/\beta_1')(\beta_{t-1}\beta_1')$. It suffices to consider the following two cases.

(a) $\beta_1\beta_2\beta_{t-1} \le 1$.

$$g(\beta_1,\cdots,\beta_{t-1};\alpha) = (1-\alpha_1(1-(1-\alpha_2(1-g(\beta_3,\cdots,\beta_{t-2},\beta_{t-1}\beta_1\beta_2;\alpha_{t-3,3})))^{\beta_2}))^{\beta_1}$$

and

$$g(\beta_1',\beta_1\beta_2/\beta_1',\beta_3,\cdots,\beta_{t-1};\alpha) = (1-\alpha_1(1-(1-\alpha_2(1-g(\beta_3,\cdots,\beta_{t-2},\beta_{t-1}\beta_1\beta_2;\alpha_{t-3,3})))^{\beta_1\beta_2/\beta_1'}))^{\beta_1'}$$

By using Lemma 1 and the fact that $\beta_2 \le \beta_1\beta_2/\beta_1'$, it can be easily seen that

$$g(\beta_1,\cdots,\beta_{t-1};\alpha) \le g(\beta_1',\beta_1\beta_2/\beta_1',\beta_3,\cdots,\beta_{t-1};\alpha).$$

(b) $\beta_1\beta_2\beta_{t-1} > 1$.

The proof is similar to that of Case (a), and the proof is omitted.

The next step is to show that

$$g(\beta_1',\beta_1\beta_2/\beta_1',\beta_3,\cdots,\beta_{t-1};\alpha) \le g(\beta_1',\beta_2',\cdots,\beta_{t-1}';\alpha).$$

We note that $\beta_1'\beta_{t-1}' \le 1$ and

$$g(\beta_1',\cdots,\beta_{t-1}';\alpha) = (1-\alpha_1(1-g(\beta_2',\cdots,\beta_{t-1}'\beta_1';\alpha_{t-2,2})))^{\beta_1'}.$$

Because $(\beta_1\beta_2/\beta_1',\beta_3,\cdots,\beta_{t-1}\beta_1')$ majorizes $(\beta_2',\beta_3',\cdots,\beta_{t-2}',\beta_{t-1}'\beta_1')$, we have

$$g(\beta_1,\cdots,\beta_{t-1};\alpha) \le g(\beta_1',\beta_1\beta_2/\beta_1',\beta_3,\cdots,\beta_{t-1};\alpha) \le g(\beta_1',\beta_2',\cdots,\beta_{t-1}';\alpha).$$

Case (ii): $\beta_1'\beta_{t-1}' > 1$.

Since $(\beta_1,\beta_2,\cdots,\beta_{t-1})$ majorizes $(\beta_1,\beta_2,\cdots,\beta_{t-2}\beta_{t-1}\beta_1,\beta_1^{-1})$ and $\beta_1\beta_1^{-1} = 1$, we have, from Case (i), that

$$g(\beta_1,\beta_2,\cdots,\beta_{t-1};\alpha) \le g(\beta_1,\beta_2,\cdots,\beta_{t-2}\beta_{t-1}\beta_1,\beta_1^{-1};\alpha)).$$

It suffices to show that

$$g(\beta_1,\beta_2,\cdots,\beta_{t-2}\beta_{t-1}\beta_1,\beta_1^{-1};\alpha) \le g(\beta_1',\beta_2',\cdots,\beta_{t-1}';\alpha).$$

It is easy to see that

$$(\beta_1,(\beta_{t-2}\beta_{t-1}\beta_1)^{-1},\beta_{t-3}^{-1},\cdots,\beta_2^{-1},\beta_1^{-1}) \text{ majorizes } ((\beta_{t-1}')^{-1},\cdots,(\beta_2')^{-1},(\beta_1')^{-1})$$

and $\beta_1\beta_1^{-1} \leq 1$, $(\beta'_{t-1})^{-1}(\beta'_1)^{-1} \leq 1$.

From Case (i) we have

$$g(\beta_1,(\beta_{t-1}\beta_{t-1}\beta_1)^{-1},\beta_{t-3}^{-1},\cdots,\beta_2^{-1},\beta_1^{-1};\overline{\alpha}) \leq g((\beta'_{t-1})^{-1},\cdots,(\beta'_2)^{-1},(\beta'_1)^{-1};\overline{\alpha})$$

where $\overline{\alpha} = (\alpha_{t-1}, \alpha_{t-2},\cdots,\alpha_2,\alpha_1)$.

Therefore

$$g(\beta_1,\beta_2,\cdots,\beta_{t-2}\beta_{t-1}\beta_1,\beta_1^{-1};\alpha) \leq g(\beta'_1,\beta'_2,\cdots,\beta'_{t-1};\alpha)$$

and Lemma 4 is proved.

It is easy to see that the blocking probability $P(\lambda_1,\cdots,\lambda_{t-1})$ of a regular SP-canopy with degree sequence $(\lambda_1,\cdots,\lambda_{t-1})$ for any link occupancy $\alpha = (\alpha_1,\cdots,\alpha_{t-1})$ has the same value as $g(\lambda_1,\cdots,\lambda_{t-1};\alpha)$. By using Lemma 4, the following theorem can be proved by techniques similar to those used in the proof of Theorem 1.

Theorem 2: Consider two t-stage regular SP-canopies with degree sequences $(\lambda_1,\cdots,\lambda_{t-1})$ *and* $(\lambda'_1,\cdots,\lambda'_{t-1})$, *respectively. Then the first linear graph is superior to the second if and only if the sequence* $(\lambda_1,\cdots,\lambda_{t-1})$ *majorizes* $(\lambda'_1,\cdots,\lambda'_{t-1})$.

IV. REMARKS AND EXAMPLES

In Fig. 3 we list a few examples. The degree sequence of the symmetric SP-canopy G_1 in Fig. 3a is $(3,2,1,2^{-1},3^{-1})$. The degree sequence of the symmetric SP-canopy G_2 in Fig. 3b is $(2,3,1,3^{-1},2^{-1})$. The degree sequence of the regular SP-canopy G_3 in Fig. 3c is $(2,3,3^{-1},1,2^{-1})$, and the

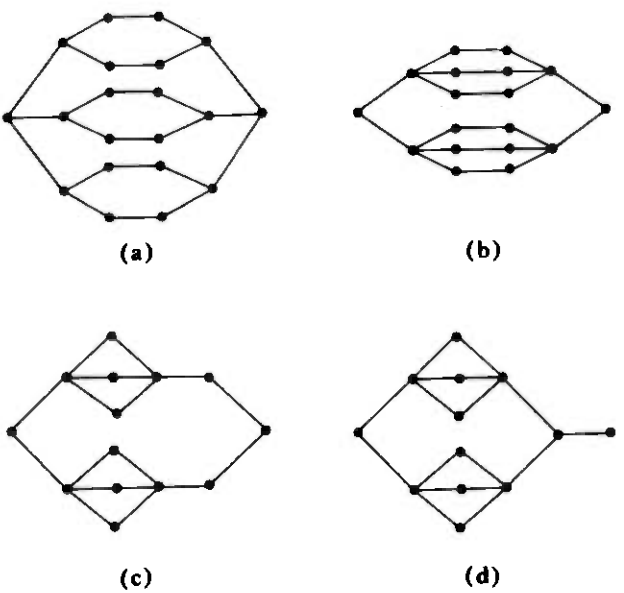


Fig. 3—Examples.

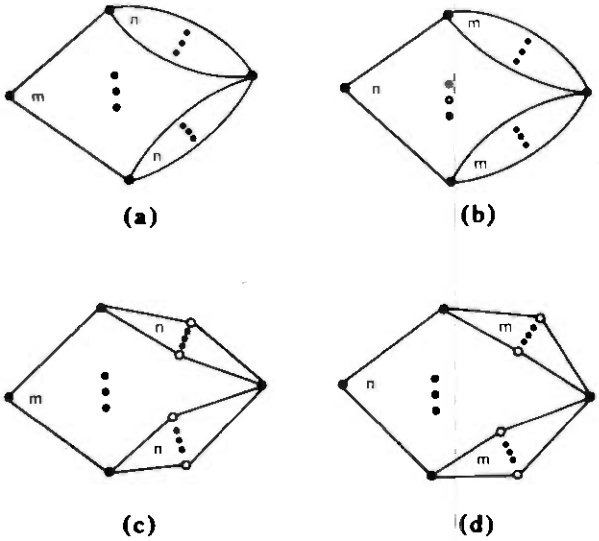


Fig. 4—Linear graphs.

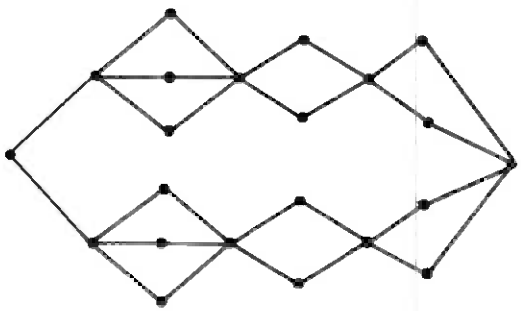


Fig. 5—Series-parallel, regular linear graph.

degree sequence of the regular SP-canopy in Fig. 3d is $(2,3,3^{-1},2^{-1},1)$. By Theorems 1 and 2, we note that, since $(3,2,1,2^{-1},3^{-1})$ majorizes $(2,3,1,3^{-1},2^{-1})$ and so forth, then G_1 is superior to G_2, which is superior to G_3, which is superior to G_4.

Although the result in this paper only involves SP-canopies, they can easily be generalized in the following ways:

(*i*) Every edge between stage i and state $i+1$ can be interpreted as a linear graph G_i.

(*ii*) Some linear graphs with multiple edges can be viewed as SP-canopies by adding imaginary stages or vertices so that we could then apply our results. For example, in order to compare linear graphs in Figs. 4a and 4b, we put an imaginary stage between the second and third stages. The resulting linear graphs are shown in Figs. 4c and 4d, respectively.

Suppose $m \geq n$ for the linear graphs in Fig. 4. It is easily seen that the degree sequence $(m,n,(mn)^{-1})$ of the linear graph in Fig. 4c majorizes the degree sequence $(n,m,(mn)^{-1})$ of the linear graph in Fig. 4d. Thus, we know that the linear graph in Fig. 4c is superior to that in Fig. 4d, and we may therefore conclude that the linear graph in Fig. 4a is superior to that in Fig. 4b.

More generally, we may consider the class of all series-parallel, regular linear graphs. For example, the linear graph in Fig. 5 is a series-parallel, regular linear graph but not a regular SP-canopy.

Is it true that a series-parallel regular linear graph is superior to another if its degree sequence majorizes the degree sequence of the other? We conjecture this is true. However, it seems that it cannot be proved by the methods we used in this paper.

REFERENCES

1. C. Y. Lee, "Analysis of Switching Networks," B.S.T.J., *34*, No. 61 (November 1955), pp. 1287–1315.
2. F. R. K. Chung and F. K. Hwang, "A Problem on Blocking Probabilities in Connecting Networks," Networks, 7 (1977), pp. 37–58.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 8, October 1978.
Printed in U. S. A.

An Analysis of 16 kb/s Sub-Band Coder Performance: Dynamic Range, Tandem Connections, and Channel Errors

By R. E. CROCHIERE

(Manuscript received December 26, 1977)

In this paper, we examine the performance of sub-band encoding under a number of constraints which can exist in practical digital communications systems. In particular, we investigate the effects of varying input signal levels, tandem connections, and channel errors on the performance of sub-band coders. A coder bit rate of 16 kb/s is used in all the simulations. The dynamic range performance is evaluated for a 50-dB range of input signal levels. Tandem connections of up to four sub-band coders in tandem are examined. Finally, the effects of random channel errors on the performance of sub-band coders is examined for bit error probabilities of up to 10 percent. A robust coder design with partial bit error protection is also proposed for use in very high channel error environments.

Three different performance measures were used in these simulations, the conventional signal-to-noise ratio, a segmental signal-to-noise ratio, and an LPC distance measure. By comparing the results of these various performance measures and from informal assessments of subjective quality, we gain some new insights into the advantages and disadvantages of these measures in terms of their usefulness in predicting coder quality.

I. INTRODUCTION

Sub-band coding has recently been proposed as a technique for obtaining relatively good quality digital speech at a bit rate of 16 kb/s.[1-3,16] This quality is subjectively comparable to that of 24 kb/s ADPCM (adaptive differential PCM),[1,2] and it is generally acceptable for some types of digital communications applications where relatively low transmission rates are required.

In a practical communications system, the quality of digital speech can be affected and degraded by a number of factors. The input speech

levels to the coders may vary over a relatively broad range, and the coders may not necessarily be driven at their optimum input levels. In a communications network, digital coders may be linked with other types of digital or analog systems, and it is possible that several tandem connections of the same type of coder may occur in a given transmission path. Finally, channel errors may occur in a digital system, and it is important to understand how the performance of digital coders are affected by these errors.

In this paper, we present the results of a series of experiments designed to assess the performance and robustness of 16-kb/s sub-band coding in practical communications environments. The dynamic range of the coder is evaluated over a 50-dB range of input signal levels. Tandem connections of sub-band coders of up to four links are examined. The effect of channel errors is examined for error probabilities as high as 0.1. Finally, several methods for improving the robustness of sub-band coding in the presence of channel errors are examined.

II. THE SUB-BAND CODER

Sub-band coding is a waveform coding technique in which the speech band is partitioned into typically four or five sub-bands by bandpass filters. Each sub-band is then lowpass-translated to dc, sampled at its Nyquist rate, and then digitally encoded using adaptive PCM (APCM) encoding. By this process of dividing the speech band into sub-bands, each sub-band can be preferentially encoded according to perceptual criteria for that band. On reconstruction, sub-band signals are decoded and bandpass-translated back to their original bands. They are then summed to give a replica of the original speech signal.

A particularly attractive implementation of the sub-band coder, in terms of hardware, is based on an integer band sampling approach.[1,2] This approach uses the samplers both for discretizing the sub-band signals as well as for doing the lowpass and bandpass translations, i.e., the modulation is achieved with an impulse train instead of with sine and cosine signals. This implementation is illustrated in Fig. 1. Bandpass filters BP_1 to BP_N in the transmitter and receiver serve to partition the input speech into sub-bands. The coders and decoders encode the sub-band signals and the multiplexer combines these digital signals into a single bit stream for transmission over the digital channel. In addition, the multiplexer inserts synchronizing bits into the bit stream for the purpose of synchronizing the operation of the transmitter and receiver.

Table I shows the choice of bands and bit allocations used in the 16-kb/s coder. The coder is a 5-band design which was proposed in Ref. 2. Column 2 shows the frequency range covered by each sub-band. The bit allocation refers to the number of bits/sample used by the coders in each

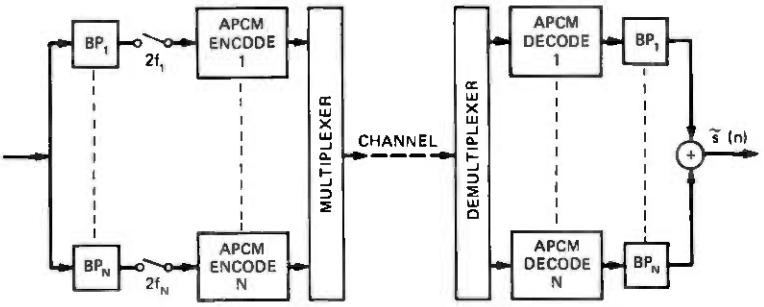


Fig. 1—An integer-band sampling implementation of the sub-band coder.

Table I — 16 kb/s 5-band sub-band coder

Band	Band Edges (Hz)	Sampling Freq (Hz)	Min Step-Size (dB)	Bit Allocation	Kb/s
1	178–356	356	(Ref)	4	1.42
2	296–593	593	0	4	2.37
3	533–1067	1067	0	3	3.20
4	1067–2133	2133	−3	2	4.27
5	2133–3200	2133	−8	2	4.27
SYNC	—			—	0.47
					16.00

sub-band. As seen from the table, more accuracy is allowed for encoding the lower bands for reasons explained in Ref. 2.

The frequency range of the coder extends from 200 to 3200 Hz. A plot of this frequency response is shown in Fig. 2. As seen in this figure, two small notches appear in the frequency response at 1067 and 2133 Hz.

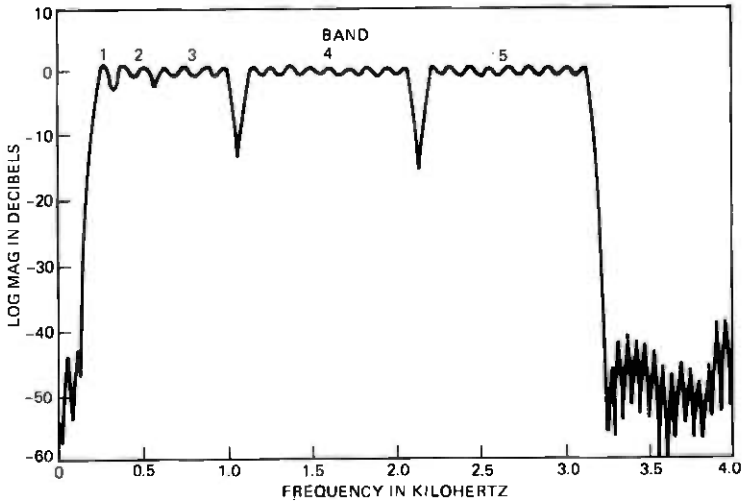


Fig. 2—Frequency response of the 5-band coder in Table I.

These notches are due to the transition bands of the filters in bands 4 and 5. Subjectively, these notches are not very perceptible. Bands 1 to 3 are overlapped to avoid such notches at lower frequencies. The filters are sharp cutoff, 200 tap, FIR filters.

Column 4 in Table I refers to the ratio of minimum allowed step sizes of the APCM coders (expressed in decibels), with the minimum step size of band 1 being the reference. This choice of minimum step sizes is different than that suggested in Ref. 2 and was found to give a better matching of the dynamic ranges of the sub-bands.

In Section VI, we propose an alternate design for a 4-band coder which can be used in a high channel error environment.

III. PERFORMANCE MEASURES

3.1 Conventional s/n

Several objective measures were used to evaluate the performance of the sub-band coder. In this section we briefly define each of these measures.

The most commonly used measure of performance of digital coders has been the conventional signal-to-noise ratio (s/n) evaluated over an utterance of speech. The speech power is defined as

$$\hat{s} = \sum_m x^2(m) \tag{1}$$

and the noise power is defined as

$$\hat{n} = \sum_m (x(m) - y(m))^2, \tag{2}$$

where $x(m)$ and $y(m)$ are the input and output signals of the coder, respectively, and the summations in (1) and (2) are taken over the entire speech utterance. The conventional s/n is then defined as

$$s/n = 10 \log(\hat{s}/\hat{n}). \tag{3}$$

In measuring the input and output signals to the sub-band coder, it is generally desirable to compensate for the effects of filtering in the coder, particularly effects of group delay. This is done by the circuit arrangement shown in Fig. 3. The input speech signal $s(m)$ is sub-band-coded to form the output speech signal $y(m)$. It is also filtered with the same filters used in the sub-band coder to generate a compensated reference signal $x(m)$ which is used as the input signal in (1) and (2). Thus, the s/n ratio defined here is strictly a measure of coder distortions and is not affected by bandlimiting or group delay in the coder.

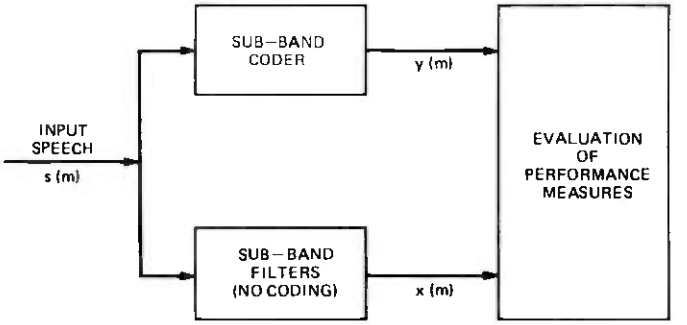


Fig. 3—Circuit for evaluating signal-to-noise ratios of the sub-band coder.

3.2 Segmental s/n

While the s/n measure is perhaps the most widely used criterion in measuring coder distortion, it has also long been known that it does not correlate well with subjective performance.[2,4,5] Another definition of signal-to-noise ratio, however, recently proposed by Noll,[6,7] does appear to correlate better with subjective performance. This measure is based on s/n measurements made over short segments of speech which are typically about 20 ms in duration. An average over all of the segments in the speech utterance is then taken to obtain a composite measure of performance for the entire utterance. If $(s/n)_i$ corresponds to the signal-to-noise ratio in decibels for a segment, i (computed in the same manner as in (3)), the segmental s/n, (SEG), is then defined as

$$\text{SEG} = \frac{1}{N}\sum_{i=1}^{N}(s/n)_i, \tag{4}$$

where it is assumed that there are N 20 ms segments in the speech utterance.

Several problems occur in this definition of segmental s/n when regions of silence exist in the speech utterance. In segments where the input signal $x(n)$ is essentially zero, any slight noise will give rise to large negative $(s/n)_i$ ratios, and these segments may unduly dominate the average in (4). To prevent this anomaly, we first identify those segments which correspond to silence and exclude them from the average in (4). This is achieved by means of a simple threshold. Let $\hat{s}_i$ represent the speech energy in a segment, i, so that

$$\hat{s}_i = \frac{1}{K}\sum_{m=1}^{K} x^2(m), \tag{5}$$

where K corresponds to the number of speech samples in the segment. Then the segment will be included in the computation of SEG in (4) if its energy exceeds a threshold σ_t^2; that is, if $\hat{s}_i > \sigma_t^2$. If it does not exceed this threshold, it is not included in the average in (4). Furthermore, to

prevent any one segment from dominating the average, we also limit the value of $(s/n)_i$ to a range of -10 to $+80$ dB. That is, $-10 \leq (s/n)_i \leq 80$ dB.

It remains to determine the threshold σ_t^2 for the speech/silence decision. To establish this threshold, we coded a speech utterance composed of three concatonated sentences in which there was about 30 percent silence in the entire utterance. Figure 4 shows a plot of SEG as a function of σ_t. The threshold σ_t was varied from 0 to 32767 corresponding to the range of signal values representable in the 16-bit integer word length of the computer. The dashed line in Fig. 4 shows the number of segments included in the SEG measure. As seen in the figure when the threshold σ_t was below 3, virtually all the silence intervals were included in the SEG measure. The low $(s/n)_i$ in these regions essentially dominated the sum, resulting in values of SEG of about 1 dB. When the threshold, σ_t, was raised to a value of 10, nearly all the silence regions were eliminated from the measure, and the value of SEG rose to about 9 dB. At a threshold of $\sigma_t = 30$, the value of SEG reached a plateau of about 10.3 dB. Therefore, the threshold was chosen to be $\sigma_t = 30$ for all SEG measurements. The conventional s/n for this same utterance was 10.8 dB.

3.3 LPC distance measure

A third performance measure that was used is the LPC distance measure proposed by Itakura.[8,9] This measure is based on an all-pole model of speech of the form

$$s(n) = \sum_{m=1}^{P} a(m)s(n-m) + Gu(n), \tag{6}$$

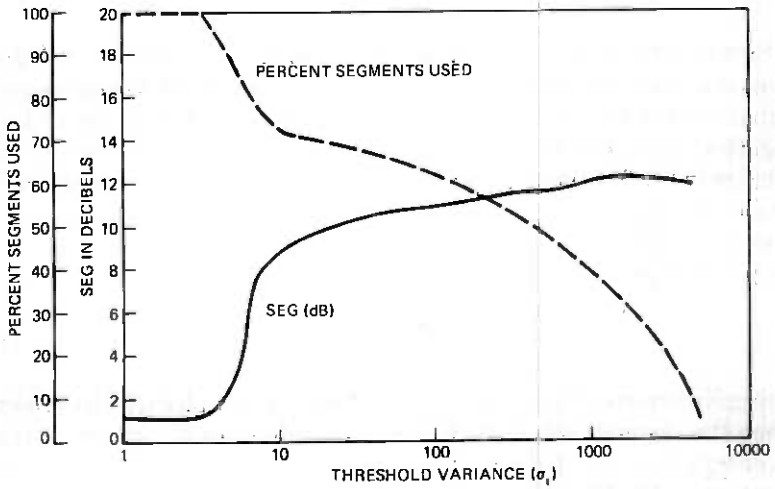


Fig. 4—Segmental s/n as a function of the speech/silence threshold σ_t.

where $s(n)$ is the sampled speech signal, $a(m)(m = 1, \ldots p)$ are the coefficients of an all-pole filter which models the resonances of the speech production mechanism, p is the number of modeled poles, G is the gain of the filter, and $u(n)$ is the excitation source for the all-pole filter.

The LPC distance measure for a segment, K, of speech (typically 20 ms in duration) is then defined as

$$d_{1k} = \log\left[\frac{\mathbf{a}_k V_b \mathbf{a}_k^t}{\mathbf{b}_k V_b \mathbf{b}_k^t}\right], \tag{7}$$

where

$\mathbf{a}_k$ = LPC coefficient vector $(1, a_1, \ldots a_n)$ measured for the kth frame of the original (reference) speech signal $s(n)$,

$\mathbf{b}_k$ = LPC coefficient vector measured for the kth frame of the coded (or processed) speech is $s'(n)$,

and V_b is the speech correlation matrix of $s'(n)$ whose elements V_{ij} are defined as

$$V_{bij} = v(|i-j|) = \sum_{n=1}^{N-|i-j|} s'(n)s'(n+|i-j|), \tag{8}$$

where $s'(n)$ is the processed speech signal. The overall distance measure for the speech utterance is then determined as the average over the N segments in the utterance,

$$\bar{d}_1 = \frac{1}{N}\sum_{k=1}^{N} d_{1k} \tag{9}$$

By interchanging the roles of the reference and processed speech, a second distance measure can similarly be defined in the form[10]

$$d_{2k} = \log\left[\frac{\mathbf{b}_k V_a \mathbf{b}_k^t}{\mathbf{a}_k V_a \mathbf{a}_k^t}\right] \tag{10}$$

and

$$\bar{d}_2 = \frac{1}{N}\sum_{k=1}^{N} d_{2k}. \tag{11}$$

An average distance measure can now be defined as

$$\bar{d} = \frac{1}{2}(\bar{d}_1 + \bar{d}_2) \tag{12}$$

which is the measure used in this paper. The LPC distance measure $\bar{d}$ is basically a measure of dissimilarity between the spectra of the processed and unprocessed speech. It is therefore useful in measuring the spectral distortion introduced by the coder. If the processed and unprocessed utterances are identical, then the distance $\bar{d}$ is zero. For small differences

between the processed and unprocessed signals, $\bar{d}$ will typically have a small positive value less than 1. Large values of $\bar{d}$, greater than 1, generally indicate significant spectral differences between the processed and unprocessed signals.

IV. DYNAMIC RANGE

The dynamic range of the sub-band coder is determined by the ratio of maximum to minimum allowed step-sizes in the APCM quantizers. In this work, we used a ratio of $\Delta_{max}/\Delta_{min} = 128$, which leads to an effective dynamic range of about 30 to 35 dB over which the quality remains relatively constant. Typically, if the $\Delta_{max}/\Delta_{min}$ ratio is increased, the dynamic range of the coder increases (within limits) by about 6 dB per doubling of the ratio.

To improve the performance of the sub-band coder at the low end of its dynamic range we also used a mid-rise/mid-tread switch in the APCM coders.[11] This extended the useful range of the coders by about 6 dB and eliminated the low-level tones and idle channel noise generated by the APCM coders.

Figure 5 shows the results of the s/n and the SEG measures for the coder for input signal levels over a range of about 50 dB. The measurements were made for a speech segment composed of two sentences, "High altitude jets whiz past screaming" and "A lathe is a big tool," spoken by two different male speakers. As seen in the figure, the conventional s/n is high in the granular noise region of the coder (input levels less than −10 dB) and drops rapidly in the over-load region (input levels greater than 0 dB). It is controlled primarily by the high-energy region in the speech utterance. At low input levels, the s/n measure is typically too large. It fails to account for the low-level granular noise of the coders,

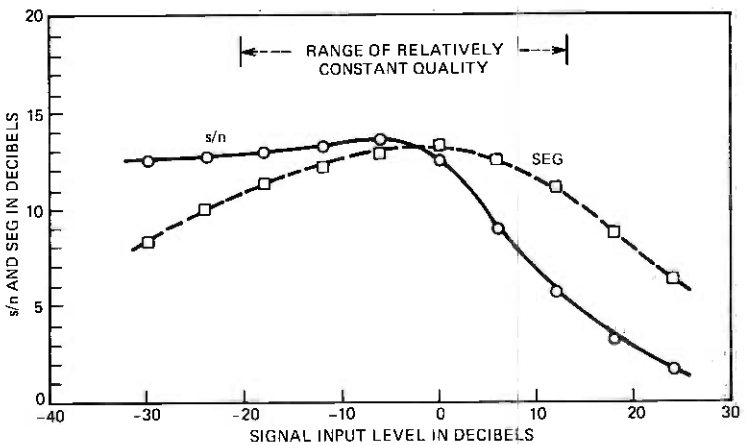


Fig. 5—Dynamic range of the sub-band coder: s/n and SEG measurements.

which can be subjectively disturbing. In the overload region, the s/n measure overemphasizes the clipping in the high-energy parts of the coded speech.

The SEG measure agrees much better with our informal observations of quality. It is more sensitive to the granular noise at low levels and less sensitive to the overload in very loud parts of the speech. It essentially treats all time segments on an equal basis and does not favor high or low parts of the utterance.

In examining the performance of the sub-band coder, it is also instructive to observe the performance of the individual APCM coders used in the sub-bands. Results for s/n and SEG measurements for the 4-, 3-, and 2-bit coders are presented in Fig. 6, where the results of the 4-bit coder are obtained from measurements of sub-bands 1 and 2, the results of the 3-bit coder are obtained from sub-band 3, and results for the 2-bit coder are obtained from sub-bands 4 and 5. The solid lines refer to s/n measurements, and the dashed lines refer to SEG measurements. An important consideration in the design of the sub-band coder is that the dynamic range in each of the sub-bands be aligned so that, at the optimum input level, each sub-band is operating at its peak performance. This alignment is determined by the choice of maximum and minimum step sizes in the coders in each sub-band. The relative values of minimum step sizes (expressed in decibels) that we used are given in column 4 of Table I, which resulted in the alignment of the dynamic ranges shown in Fig. 6.

Figure 7 shows the results of the LPC distance measurements on the sub-band coder. The measure was made between $x(m)$ and $y(m)$ according to the arrangement in Fig. 3 and, therefore, does not take into account the spectral distortions due to the filters or notches between the bands. At the optimum input level, the value of the LPC distance is 0.12.

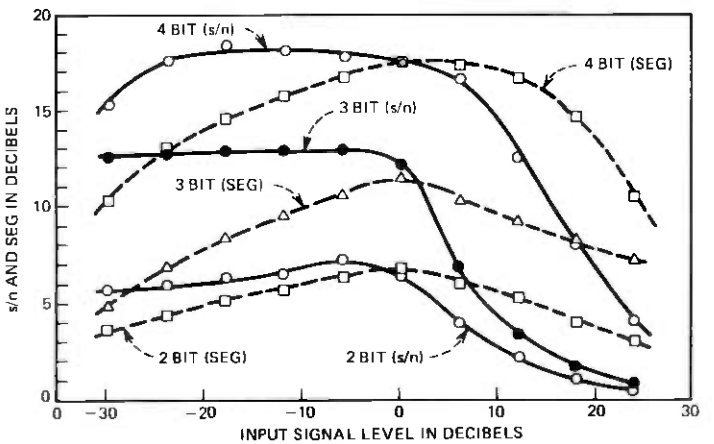


Fig. 6—Dynamic range of the individual APCM coders in the sub-bands.

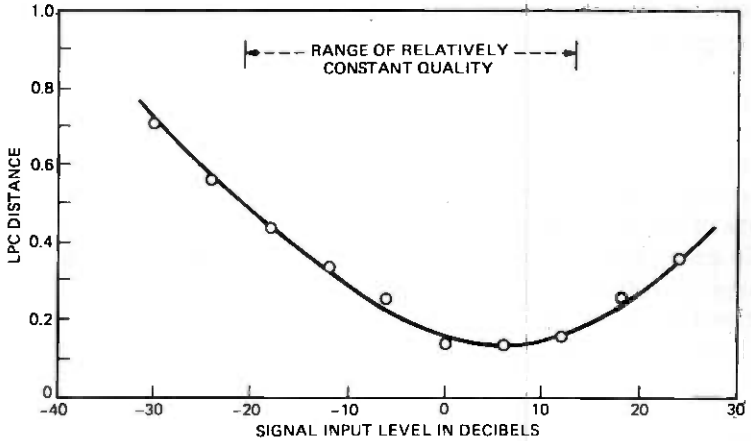


Fig. 7—LPC distance as a function of the input signal level.

For low input levels, i.e., in the granular noise region, it goes up to 0.56 at a −24-dB input level. At high input levels (+24 dB) in the overload or clipping region of the coder, the LPC distance goes up to 0.36. Thus, the spectral distortion is typically greater in the granular noise region than in the overload region of the coder.

To determine the effect of the bandpass filters and the notches in frequency response of the filter, a second LPC distance measurement was made across the sub-band coder according to the arrangement shown in Fig. 8. The input speech was delayed by a flat delay equal to the delay of the filters. This reference signal and the output of the coder were then both filtered with a 200- to 3200-Hz bandpass filter giving the signals $\hat{x}(m)$ and $\hat{y}(m)$. The purpose of the bandpass filters on the outputs is so that the spectral differences outside of the 200- to 3200-Hz band of interest do not affect the LPC distance measure.

When we measured the LPC distance between $\hat{x}(m)$ and $\hat{y}(m)$ by this method, we obtained a distance of 0.58 for the sub-band coder (operating at the optimum input level of 0 dB). We then removed the quantizers

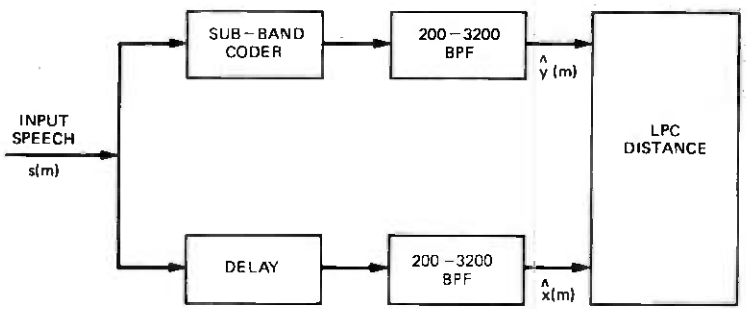


Fig. 8—Circuit arrangement for measuring the total LPC distance of the sub-band coder (including the effect of the filters) in a 200- to 3200-Hz bandwidth.

from the coder and measured only the effects of the sub-band filtering. This resulted in a distance of 0.53 between $\hat{x}(m)$ and $\hat{y}(m)$. This distance is strictly due to the passband ripples, sharp transition bands, and notches in the frequency response of the coder as seen in Fig. 2. Although the contribution to the LPC distance due to the filters was greater than that due to quantization noise, their subjective effects cannot necessarily be weighted in the same way. Subjectively, the effects of the sharp cutoff filters and the notches do not strongly affect the quality or intelligibility of the coder.

V. TANDEM CONNECTIONS

Computer simulations of tandem connections of sub-band coders were made for up to four coders in tandem. Two types of tandem connections were considered in this experiment. The first type of tandem link consists of a sub-band coder followed by 16-bit linear PCM as shown in Fig. 9a. A parallel link of sub-band filters, shown in dotted lines, was also simulated in order to generate reference signals to facilitate s/n and SEG measurements.

In the second type of link shown in Fig. 9b, we simulated the effects of a digital-to-analog conversion and a resampling of the signals between each coder. This simulation was achieved by means of an all-pass filter which was inserted between the tandem links. Again, a reference link of sub-band filters was also simulated to facilitate signal-to-noise ratio measurements. The effect of the all-pass filter is to disperse the phase of the coder output so that the succeeding coders cannot synchronize their levels from link to link. Figure 10 is a plot of the group delay of the all-pass filter that was used.

Figure 11 shows the results of s/n and SEG measurements for the tandem connections as a function of the number of tandem links. The solid lines refer to s/n measurements and the dashed lines refer to SEG measurements. The upper two curves refer to measurements made on the sub-band-to-PCM links in Fig. 9a, and the lower curves refer to measurements made on the sub-band-to-analog connections of Fig. 9b.

As seen in the figure, for the sub-band-to-analog connections, the s/n and SEG measures drop by roughly 3 dB per doubling of the number of tandem links, indicating that the quantization noise contributed by each link adds independently of other links.

In the sub-band-to-PCM connection, however, it is seen that the quantizer distortions do not add independently. After the first encoding, the succeeding coders tend to synchronize their quantizer levels to those of the first coder, and in this way they do not add any further distortion to the signal. This result is somewhat surprising in view of the fact that the quantizers in the sub-bands are separated by interpolating and de-

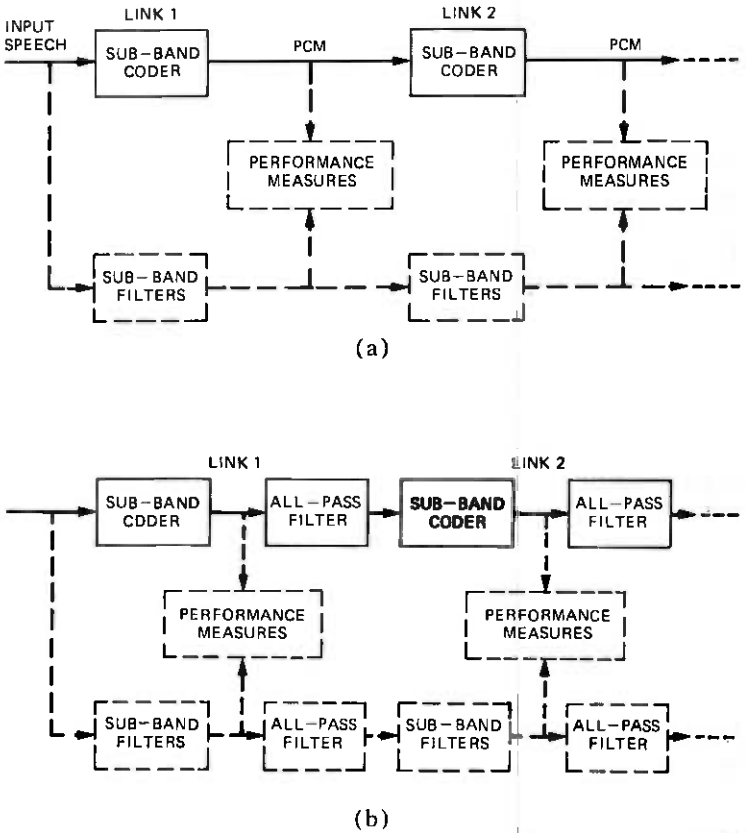


Fig. 9—Circuits for measuring performance of tandem connections of sub-band coders. (a) Sub-band/PCM links. (b) Sub-band/analog links.

cimating filters and all the sub-bands are summed at the outputs of the coders between links. In one example, we observed an s/n of 6.8 dB in the fourth sub-band of the first link. In the succeeding links, the s/n of this same coder went up to 18 dB in the fourth sub-band due to this synchronization effect.

Figure 12 shows the corresponding results for the LPC distance measurements on the tandem connections. Here again we see that the sub-band-to-PCM link performs better than the sub-band-to-analog link. A maximum LPC distance of 0.29 was observed for four sub-band-to-analog tandem connections, indicating that successive tandem connections do not excessively distort the spectrum of the coded speech over that of the initial coding.

Based on informal listening, the quality of two tandem connections does not appear to be much different than that of one encoding. For three sub-band/analog encodings, the differences become apparent, and with four links the differences are clearly noticeable.

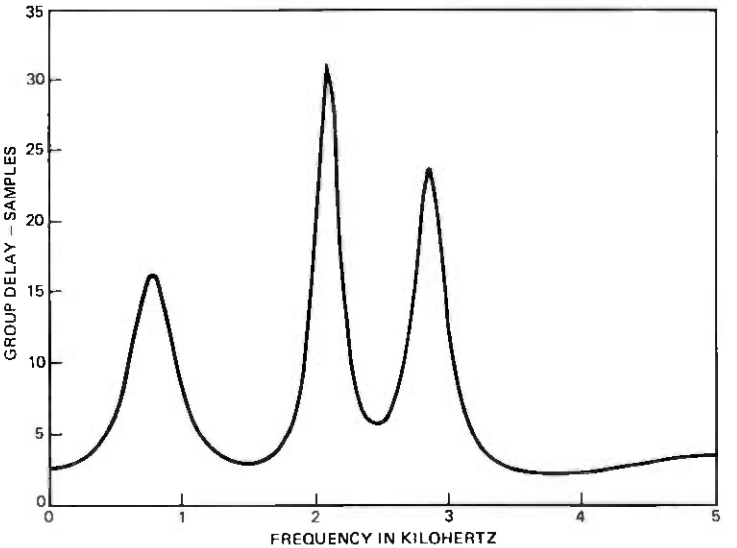


Fig. 10—Group delay as a function of frequency for the all-pass filters used to simulate analog links.

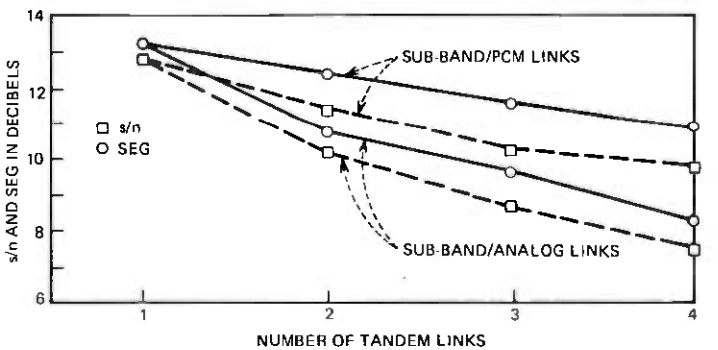


Fig. 11—*s/n* and SEG measurements for the sub-band/PCM and sub-band/analog tandem connections.

VI. CHANNEL ERRORS

The analysis of the sub-band coder performance under channel errors constituted the largest part of our experimental investigations. The coder performance was analyzed for bit error probabilities of up to 10 percent. We first analyzed the individual 4-, 3-, and 2-bit APCM coders in each of the sub-bands in order to assess their performance separately under channel errors. We then examined the use of a robust step-size adaption algorithm[12] in order to enhance the performance of these individual coders. For the 4- and 3-bit coders, we also investigated the use of partial bit error protection of the sign and most significant bits in the coders.

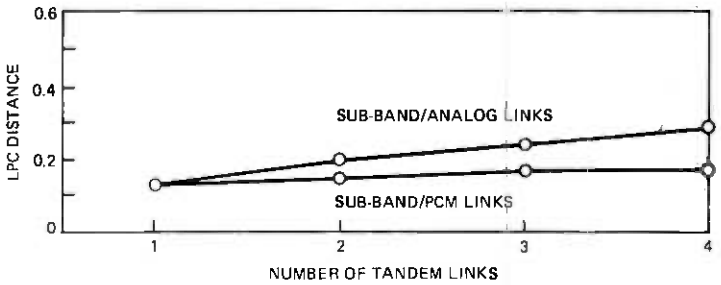


Fig. 12—LPC distance measurements for the tandem connections.

Based on these results, we then considered three overall sub-band coder designs. The first design was the 5-band coder described in Section II. The second design was the same coder with the robust step-size adaption algorithm for its APCM coders. In the third design, we considered a 4-band coder with a reduced bandwidth and a slightly lower bit rate. The remainding bits were applied to a partial bit error protection scheme to enhance its robustness under conditions of very high channel errors. We then analyzed and compared the performance of these three coders under channel errors.

6.1 The robust quantizer

The step-size adaption algorithm used in the sub-band coder is based on the one-word step-size memory scheme proposed by Jayant, Flanagan, and Cummiskey.[4,13] The coder input signal is quantized to one of 2^B levels, where B is the number of bits in the coder. The step-size adaption circuit examines the quantizer output bits for the $(r-1)$th sample and computes the quantizer step-size, Δ_r, for the rth sample according to the relation

$$\Delta_r = \Delta_{r-1} M(L_{r-1}), \tag{13a}$$

where

$$\Delta_{\min} \leqslant \Delta_r \leqslant \Delta_{\max} \tag{13b}$$

and where Δ_{r-1} is the step-size used for the $(r-1)$th sample. $M(L_{r-1})$ is a multiplication factor whose value depends on the quantizer magnitude level L_{r-1} at time $r-1$. It can take on one of 2^{B-1} values $M_1, M_2, \ldots M_{2^{(B-1)}}$. If the lower-magnitude quantizer levels are used at time $r-1$, a value of $M(L_{r-1}) = M_i$ less than one is used to reduce the next step-size. If upper magnitude levels are encountered, a value of M_i greater than 1 is chosen. In this way, the coder continuously adapts its step-size in an attempt to track the short-time variance of the input signal.

A disadvantage of the above adaption scheme is that, once a step-size

error occurs, it remains in error until the maximum or mininum step-size is reached. A modification of this algorithm, proposed by Goodman and Wilkinson[12] allows for the step-size computation to be less sensitive to past errors. This "robust" algorithm is based on the relation

$$\Delta_r = (\Delta_{r-1})^{\beta} \cdot M(L_{r-1}), \tag{14a}$$

where

$$\Delta_{\min} \leq \Delta_r \leq \Delta_{\max}. \tag{14b}$$

The parameter β is chosen to be slightly less than 1, and it determines how rapidly the effects of past errors are dissipated. In the limit when β goes to 1, the algorithm reduces to that of (13a).

As the value of β is reduced, the M values must be adjusted to compensate for its effect on the step-size adaptation. As shown in Ref. 12, this compensation can be obtained by a simple scaling of the M values. If $\hat{M}_i$ represent the ideal M values for step-size adaption when $\beta = 1$, then the new M values, denoted as M_i, are approximately

$$M_i = G\hat{M}_i \qquad i = 1, 2, \ldots 2^{B-1}, \tag{15}$$

where G is a scaling factor that is dependent on β *and* on the expected value of Δ_r. In computer simulations, we determined G by optimizing the performance of the coders as a function of this scaling factor. Figure 13 is a plot of G as a function β that was used in our simulations. It is based on an expected value of Δ_r in the range of 500 to 5000, typically encountered in our computer simulations. As seen in the figure, when β varies from 1 down to 15/16 optimum scaling factor, G, increases from a value of 1 to about 1.5.

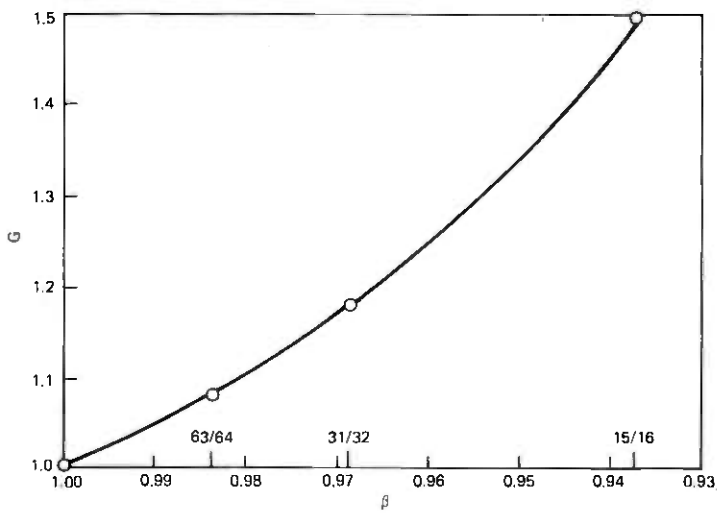


Fig. 13—Multiplier scaling factor, G, as a function of β, used for computer simulations.

6.2 Performance of individual coders under channel errors

The performance of the individual APCM coders was examined, in terms of s/n and SEG measurements, as a function of the bit-error rate and the robust quantizer parameter, β. Figures 14a to 14c show the results for the s/n measurements for the 4-, 3-, and 2-bit coders, respectively, as a function of bit-error rate where the bit-error rate corresponds to random channel errors. Figures 15a to 15c show similar results for the SEG measurements. Four values of β were used: 1, 63/64, 31/32, and

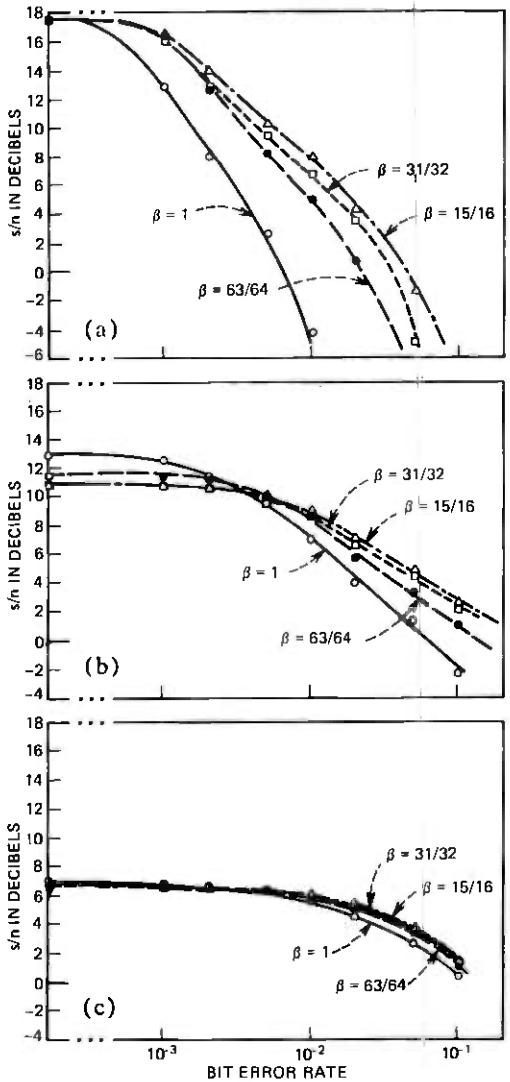


Fig. 14—s/n performance of the APCM coders as a function of the bit error rate. (a) 4-bit coder. (b) 3-bit coder. (c) 2-bit coder.

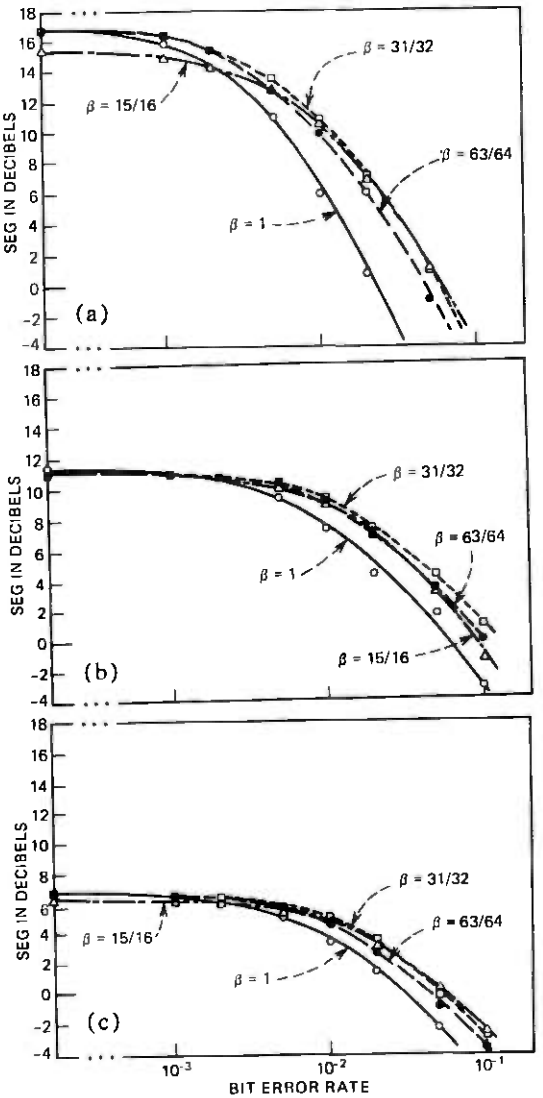


Fig. 15—SEG performance of the APCM coders as a function of the bit error rate. (a) 4-bit coder. (b) 3-bit coder. (c) 2-bit coder.

15/16. Each value of β corresponds to one curve in the plots. The bit error rate is plotted on a log scale and covers a range of 10^{-4} to 10^{-1}.

Several conclusions can be drawn from the results in Figs. 14 and 15. It is seen that the 4-bit coder is the most vulnerable to channel errors and that the 2-bit coder is the least vulnerable. Fortunately, the 4-bit coder receives the most improvement from the use of a robust step-size algorithm. The 2-bit coder, however, receives the smallest improvement from the robust algorithm.

The robust quantizer does not seem to affect the s/n and SEG measurements when no channel errors are present until the value of β is reduced below a value of about 31/32. This is assuming that the M values in the coder are appropriately scaled as discussed in the preceding section. If the M values are not properly scaled, then the performance of the coders will be significantly reduced as β decreases. For example, the performance of the 3-bit coder drops by about 6 dB in s/n and 3 dB in the SEG measure when β is reduced from 1 to 31/32 and the M values are not scaled according to (15).

The optimum choice for the robust quantizer parameter, β, for protection against channel errors appears to be about 31/32.

6.3 Partial bit error correction

Since the 4- and 3-bit coders are the most vulnerable to channel errors, the lower sub-bands which use these coders are affected the most by channel errors. Subjectively, these are also the most important bands since distortions in these bands quickly deteriorate the quality of the sub-band order.

One way to maintain the quality in these lower sub-bands is to provide for some partial bit-error correction in the transmission of these coder bits. In this section, we investigate the effect of sign and/or most significant magnitude bit protection on the performance of the 3- and 4-bit APCM coders.

To provide for error correction of transmitted bits, extra parity bits must be transmitted by the coder.[14,15] The degree of error protection that is achieved is strongly dependent on the design of the error protection block codes, the bit error rate of the channel, and the percentage of additional redundant bits that are transmitted for error protection. Fortunately, since the lower sub-bands typically have low sampling rates and therefore low transmission rates, the additional transmission rate required to provide partial bit-error correction of some of the bits in these lower sub-bands should be relatively small compared to the overall transmission rate of the coder. In this work, we have avoided issues of specific designs of block codes for bit error correction. We have instead assumed that ideal or nearly ideal error protection can be achieved. The results that we present should therefore be interpreted as upper bounds on what can be achieved, given a sufficient amount of extra transmission rate for error protection.

Figure 16a and 16b show results of s/n and SEG measurements on the 4-bit APCM coder, as a function of the bit error rate, for several bit error protection schemes. In all the results, a robust quantizer with $\beta = 31/32$ is used. The solid line shows the performance when no bit-error correction is used. The long dashed curve shows the results when the sign bit is ideally protected, and the short dashed line shows the results when

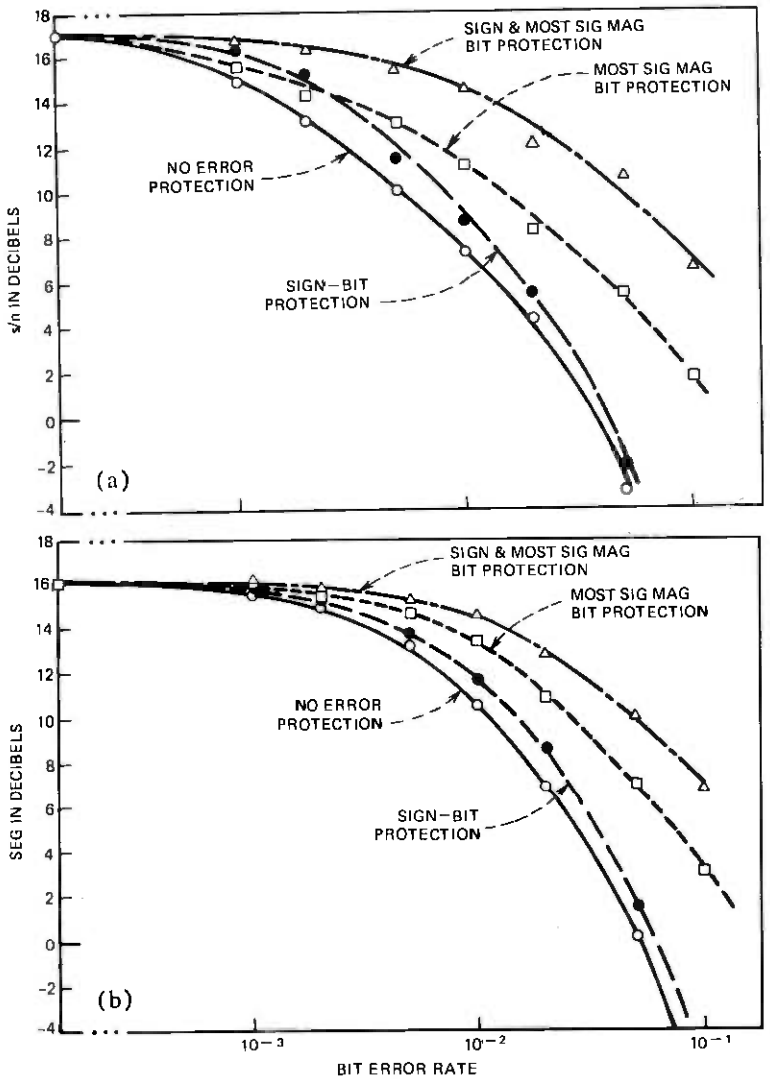


Fig. 16—Performance of the 4-bit APCM coder with partial error protection. (a) s/n measurements. (b) SEG measurements.

the most significant magnitude bit is ideally protected. Finally, the long and short dashed curve shows the performance when both the sign bit and the most significant magnitude bit are protected. As seen in the figure, protection of the most significant magnitude bit alone gives a better performance than when the sign bit is protected alone. This occurs because an error in the sign bit results in a single isolated error, whereas an error in the most significant magnitude bit causes a step-size error which propagates for many samples. At high bit-error rates, significant

improvements in coder performance are possible with bit-error protection.

Figure 17 shows similar results for the 3-bit APCM coder. In this case, the protection of only one bit was considered, either the sign bit (long dashed line) or the most significant magnitude bit (short dashed line). It is seen that protecting the sign bit leads to about the same improvement as the most significant magnitude bit. In comparing Figs. 16 and 17, it can be seen that the improvement of the 3-bit coder performance with error protection in high channel errors is not as large as the improvement obtained for the 4-bit coder.

6.4 A sub-band coder design for high channel errors

As noted in the previous section, when high channel errors are encountered, it is possible to divert a part of the transmission rate to the protection of bits in the lower sub-band(s). In this way, some of the coder quality at low channel error rates can be traded for more robustness of the coder at high channel error rates. In this section, we consider an example of such a design.

Table II shows the choice of bands and bit allocations for a 4-band

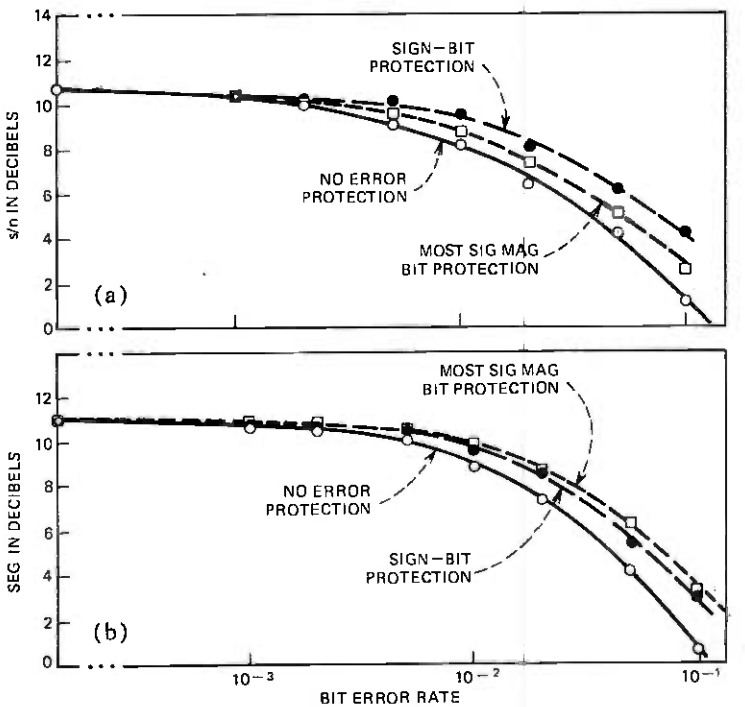


Fig. 17—Performance of the 3-bit APCM coder with partial error protection. (a) s/n measurements. (b) SEG measurements.

Table II — 16 kb/s 4-band coder with partial bit error correction

Band	Band Edges (Hz)	Sampling Freq (Hz)	Min Step-Size (dB)	Bit Allocation	Kb/s
1	250–500	500	(Ref)	4	2.0
2	500–1000	1000	−1.9	3	3.0
3	1000–2000	2000	−6	2	4.0
4	2000–3000	2000	−10	2	4.0
SYNC AND ERROR CORRECTION					3.0
					16.0

16-kb/s coder with partial bit-error correction in the lowest band. The frequency response of this coder is shown in Fig. 18. In comparison to the 5-band coder, it is seen that this coder has a narrower overall bandwidth and an additional notch in its frequency response. Thus, the quality of this coder tends to be more reverberant than that of the 5-band design. The LPC distance measure for this coder, measured according to Fig. 8, was 0.82 compared to 0.58 for the 5-band coder. When the sub-band filters alone were measured, a distance of 0.69 was observed compared to 0.53 for the 5-band coder.

In trade for this reduced quality, the 4-band coder has 3 kb/s of remaining transmission rate or 18.75 percent of its total transmission rate which can be used for bit error protection. This is applied to the protection of the sign and most-significant magnitude bits of the 4-bit coder in the first sub-band.

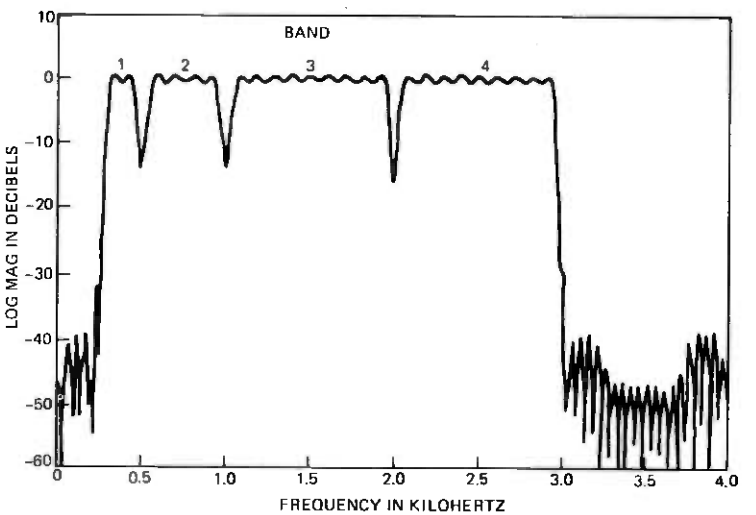


Fig. 18—Frequency response of the 4-band coder in Table II.

6.5 Overall performance of the sub-band coders with channel errors

In this section, we present the results of computer simulations of three different sub-band coder designs under conditions of channel errors. The simulations were made with random channel errors with error rates of up to 10^{-1}. The first coder, coder A, is the 5-band coder in Table I with no robust quantizer (i.e., $\beta = 1$). Coder B is the same 5-band design with a robust quantizer with $\beta = 31/32$. Coder C is the 4-band design, in Table II, for high channel errors. It has a robust quantizer, $\beta = 31/32$, and assumes ideal error protection of the sign and most-significant bit in its first sub-band (the 4-bit APCM coder).

Figure 19 shows the results of the s/n and SEG measurements for the three coders as a function of the bit error rate. Coders A and B, the 5-band designs, have the best performance and quality at very low error rates. The use of the robust quantizer does not significantly reduce the performance of Coder B (assuming the M values are scaled properly) at low error rates. The 4-band design has a somewhat lower quality at low error rates due to its reduced bandwidth and lower effective transmission rate.

As the bit error rate increases, the performance of the unprotected coder, coder A, drops rapidly. Channel error distortions are noticeable at error rates of 2×10^{-3}. At error rates of 5×10^{-3} and 10^{-2}, the quality drops rapidly and at error rates of 2×10^{-2} the coder is essentially unintelligible.

The use of the robust quantizer significantly improves the performance of the 5-band coder for moderate error rates. Coder B has noticeable degradations in quality at bit-error rates of about 10^{-2}. At error rates of $2 \times 10^{-}2$, this quality degrades rapidly and at error rates of 5×10^{-2} the coder starts to become unintelligible.

The 4-band coder, coder C, holds up well for error rates up to about 2×10^{-2} before the effect of channel errors becomes noticeable. Its quality, however, is slightly lower to begin with. At error rates of 10^{-1}, the quality degrades sharply although the coder still appears to be quite intelligible.

Figure 20 shows the results of the LPC distance measure on the three coders. These results do not appear to agree well with s/n and SEG measures nor do they agree well with our informal subjective observations. For example, at bit error rates of 10^{-2} the LPC distance of coder A is 0.35, indicating that the coder should have reasonably good quality. In fact, the subjective quality of the coder at this point was significantly degraded. Also, the LPC distance failed to sufficiently distinguish the differences in quality between coders B and C at high error rates.

To investigate this problem in more detail, we plotted the individual segmental LPC distances d_{1k} and d_{2k} defined in (7) and (10) as a function of time (measured in segments). Figure 21a shows these results for coder

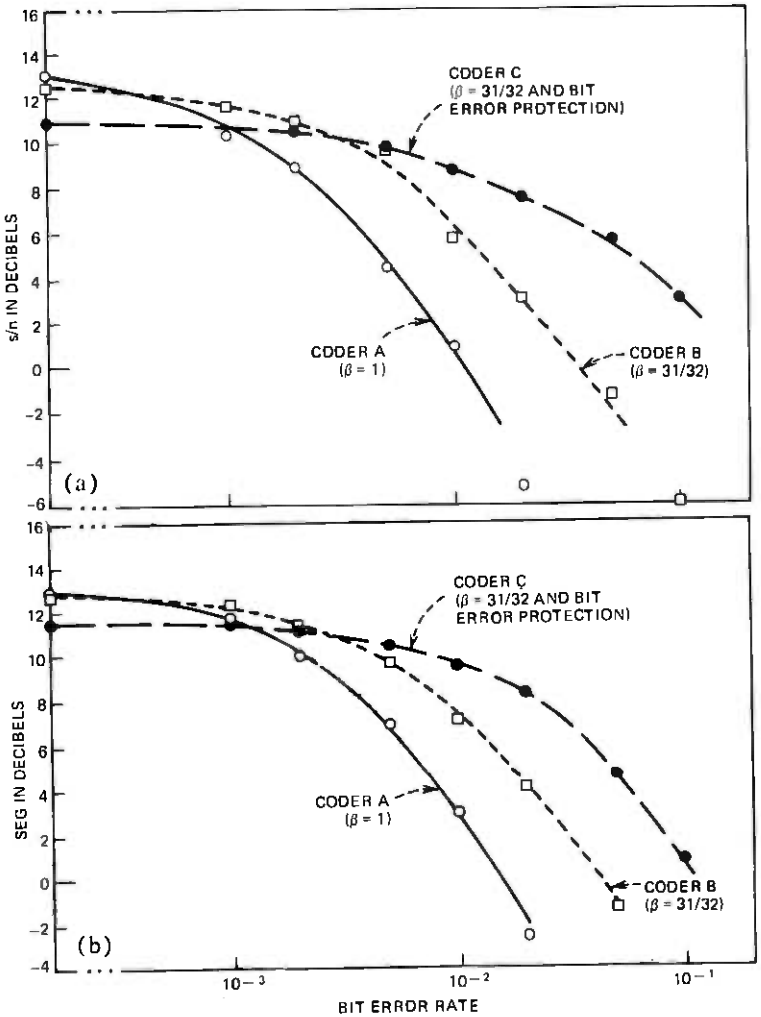


Fig. 19—Performance of the three sub-band coder designs under channel errors. (a) s/n measurements. (b) SEG measurements.

A at a bit error rate of 10^{-2} for two concatenated sentences. From this plot, it becomes clear as to what is happening. On the average, the coder performance is quite good. However, in about 10 or 12 isolated segments, severe distortions were observed where channel errors occurred in lower sub-bands. Because of these isolated errors, the entire sentence sounds poor in quality. Figure 21b shows similar results for coder A at error rates of 5×10^{-2}. Again, it is seen that there are numerous segments in which the distortions are intolerable; however, on the average, the distortion was not that bad; i.e., it was below 1. Subjectively, the presence of these large errors made the sentence virtually unintelligible.

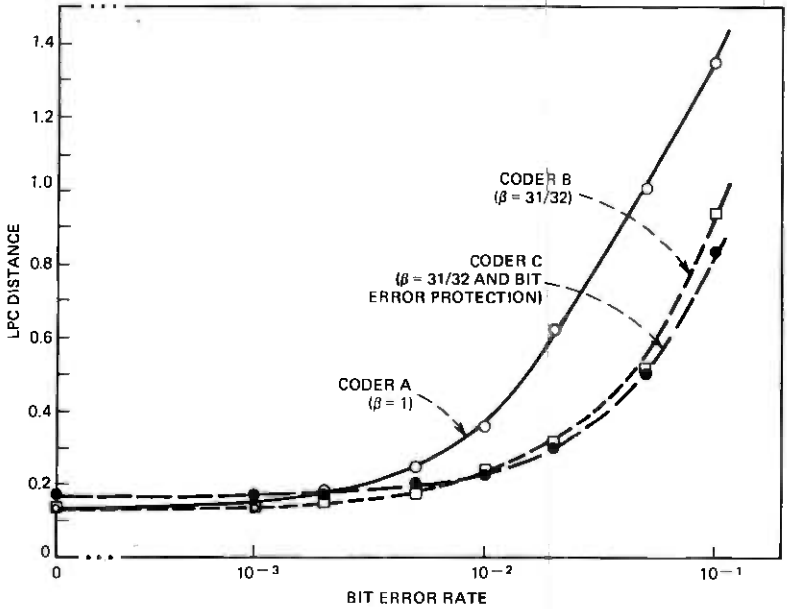


Fig. 20—LPC distance as a function of bit-error rate for the three sub-band coder designs.

From these observations, we conclude that the LPC distance, used properly, is in fact a good indicator of quality. However, when an overall measure of quality for an utterance is required, something more sophisticated than a simple mean of the segmental distances must be used. This is particularly important in the case of channel errors.

VII. CONCLUSIONS

In summary, a number of general conclusions can be drawn from the results of this work.

(*i*) When the maximum-to-minimum step-size ratios of the APCM coders is 128 and the dynamic range of sub-bands are properly aligned, the quality of the coder remains relatively constant over a range of input levels of about 30 dB. This range increases by about 6 dB per doubling of this step-size ratio. The idle channel noise performance of the coder can be improved by the use of a mid-rise/mid-tread switch on the quantizers in the APCM coders.

(*ii*) For tandem connections of sub-band coders with conversion to analog format between links, the signal-to-noise ratio drops by roughly 3 dB per doubling of the number of tandem coders. When linear phase FIR filters are used in the coders and they are connected by PCM links, the step sizes of the coders tend to synchronize, and the performance of the tandem connection improves.

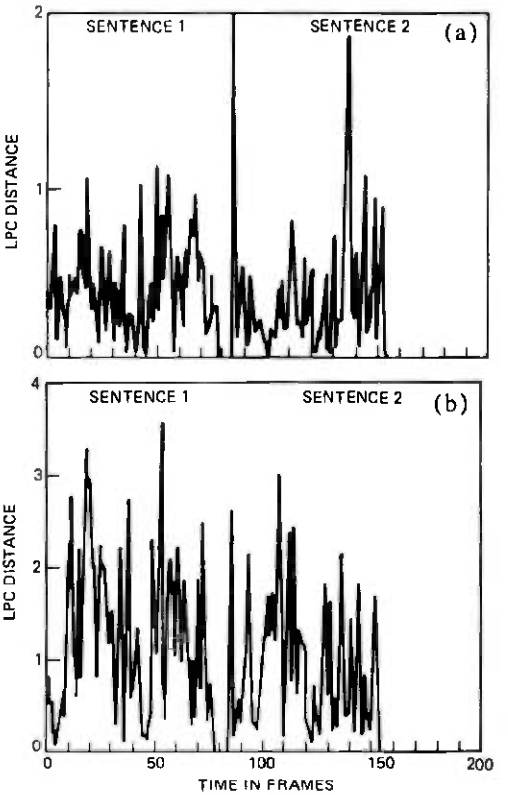


Fig. 21—LPC distance as a function of time for (a) coder A at a bit error rate of 10^{-2} and (b) coder A at a bit error rate of 5×10^{-2} (note a difference in scale).

(*iii*) The effects of channel errors in an unprotected sub-band coder are first observed at bit error rates of about 2×10^{-3}. At error rates of 2×10^{-2}, the quality of the coder is essentially unintelligible. When a robust quantizer algorithm is used, errors are first noticeable at bit error rates of about 10^{-2}, and at error rates above 5×10^{-2} the coder becomes unintelligible. When both a robust quantizer and partial bit-error protection is used in the lower sub-band(s), the effect of channel errors is not significant until error rates of about 2×10^{-2} are reached and the coder appears to be intelligible at error rates as high as 10^{-1}. The above results are based on the assumption that sufficient protection is provided for the synchronization and parity bits so that no loss of synchronization occurs between high channel errors and that nearly ideal error protection is possible for the coder bits which are protected in the partial bit-error protection scheme.

REFERENCES

1. R. E. Crochiere, S. A. Webber, and J. L. Flanagan, "Digital Coding of Speech in Sub-Bands," B.S.T.J., *55*, No. 8 (October 1976), pp 1069–1085.
2. R. E. Crochiere, "On the Design of Sub-band Coders for Low Bit Rate Speech Communication," B.S.T.J., *56*, No. 5 (May-June 1977), pp. 747–770.
3. D. Esteban and C. Galand, "Application of Quadrature Mirror Filters to Split Band Voice Coding Schemes," 1977 IEEE Int'l. Conf. on Acoust., Speech and Sig. Proc. (May 1977), pp. 191–195.
4. N. S. Jayant, "Digital Coding of Speech Waveforms: PCM, DPCM, and DM Quantizers," Proc. IEEE, *62* (May 1974), pp. 611–632.
5. N. S. Jayant, (ed.) *Waveform Quantization and Coding*, New York: IEEE Press, 1976.
6. P. Noll, "Adaptive Quantization in Speech Coding Systems," Int. Zurich Seminar on Digital Communication (IEEE) (October 1976), pp. B3.1 to B3.6.
7. R. Zelinski and P. Noll, "Adaptive Transform Coding of Speech Signals," IEEE Trans. Acoust., Speech and Sig. Proc., *ASSP-25* (August 1977), pp. 299–309.
8. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoust., Speech and Signal Processing, *ASSP-23*, (Feburary 1975), pp. 67–72.
9. M. R. Sambur and N. S. Jayant, "LPC Analysis/Synthesis from Speech Inputs Containing Quantizing Noise or Additive White Noise," IEEE Trans. Acoust., Speech and Sig. Proc., *ASSP-24* (December 1976), pp. 488–494.
10. L. R. Rabiner, private communication.
11. R. E. Crochiere, "A Mid-Rise/Mid-Tread Quantizer Switch for Improved Idle-Channel Performance in Adaptive Coders," B.S.T.J., this issue, pp. 2953–2955.
12. D. J. Goodman and R. M. Wilkinson, "A Robust Adaptive Quantizer," IEEE Trans. Commun. (November 1975), pp. 1362–1365.
13. P. Cummiskey, N. S. Jayant, and J. L. Flanagan, "Adaptive Quantization in Differential PCM Coding of Speech," B.S.T.J., *52*, No. 7 (September 1973), pp. 1105–1118.
14. R. W. Hamming, "Error Detecting and Error Correcting Codes," B.S.T.J., *29*, No. 2 (April 1950), pp. 147–160.
15. E. R. Berlekamp (ed.), *Key Papers in the Development of Coding Theory*, New York: IEEE Press, 1974.
16. R. E. Crochiere, J. L. Flanagan, and S. A. Webber, "Digital Speech Communication System for Minimizing Quantizing Noise," U.S. Patent 4,048,443, September 13, 1977.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 8, October 1978
Printed in U.S.A.

A Mid-Rise/Mid-Tread Quantizer Switch for Improved Idle-Channel Performance in Adaptive Coders

By R. E. CROCHIERE

(Manuscript received November 17, 1977)

A mid-rise/mid-tread switch is proposed for improving the idle channel performance of an adaptive quantizer. The method incorporates the advantages of both mid-rise and mid-tread quantizer characteristics.

In adaptive waveform coding such as ADPCM (adaptive differential PCM), ADM (adaptive delta modulation),[1,2] and sub-band coding,[3,4] the quantizer step-size in the coder varies in accordance with the short-time energy of the signal being coded in order to take advantage of its non-stationary properties. In practice, these types of coders generally have minimum and maximum limits on their step-size. Furthermore, they often use a mid-rise quantizer characteristic as depicted in Fig. 1 where x denotes the input signal level, $\hat{x}$ denotes the discrete output signal levels, and Δ denotes the quantizer step-size. This mid-rise characteristic is desirable because of its symmetry and because it uses the 2^B possible levels of a B-bit coder efficiently.

A disadvantage of this mid-rise characteristic is that it cannot represent a zero output level. During very low, or zero, input signal intervals (such as silent regions in speech), the output of the coder must be $\pm\Delta_{\min}$, where $\Delta_{\min}$ is the minimum step-size in the coder. Generally, $\Delta_{\min}$ is chosen to be small enough so that this signal is very low. Unfortunately, in many coder designs the sign of the output signal varies in a systematic pattern which can be perceived even for very low values of $\Delta_{\min}$.

For example, in ADPCM with a zero input level, the coder output level can oscillate between $\pm\Delta_{\min}$, creating (in a speech coder) a low level tone at half the sampling rate. Because of the sensitivity of our hearing mechanisms to tones, this tone can be perceived even at very low levels (i.e., very small $\Delta_{\min}$s). In sub-band coders, this problem is compounded further by the fact that sub-bands of speech are lowpass-translated to dc, encoded and decoded, and then bandpass-translated back to their

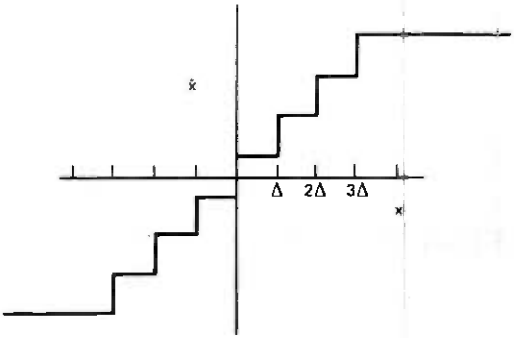


Fig. 1—Mid-rise quantizer characteristic.

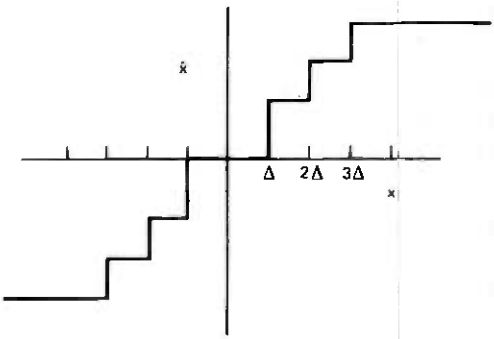


Fig. 2—Mid-tread quantizer characteristic.

respective bands. This modulation process can translate such small tones directly into the middle of the speech band where they are particularly noticeable. Even a dc level of $+\Delta_{\min}$ or $-\Delta_{\min}$ will appear as a low-level tone when modulated to the middle of the speech band. Furthermore, decaying exponentials, in which the quantizer systematically uses its lowest level as its step-size decays to $\Delta_{\min}$, will appear as decaying tones in this situation.

One way to alleviate the above situation is to use a mid-tread quantizer characteristic as shown in Fig. 2. Unfortunately, this characteristic has an odd number of levels (if it is symmetric) or it must be nonsymmetric about zero. Therefore, it does not use the 2^B possible levels of a B-bit quantizer efficiently.

For adaptive quantizers, fortunately, there is a solution to the above problem. Since the quantizer step-size Δ varies with the short-time energy of the signal being coded, it also tells us when the input signal level is near zero. We propose a mid-rise to mid-tread switch on the decoder quantizer output which occurs when the step-size Δ falls below some threshold Δ_{th}. That is, when $\Delta \leq \Delta_{th}$, the two lowest levels of the quantizer in the decoder are switched to zero to give a mid-tread char-

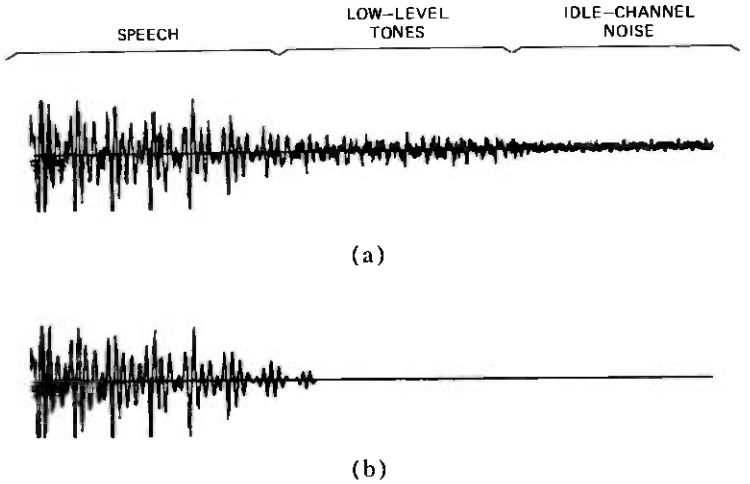


Fig. 3—(a) Output of a sub-band coder with mid-rise quantizers as the step-size is decaying. (b) The same output using the mid-rise/mid-tread switch.

acteristic as in Fig. 2. When $\Delta > \Delta_{th}$, the characteristic in Fig. 1 is used because of its more efficient use of levels. Typically, a practical choice of Δ_{th} is about 1.5 $\Delta_{\min}$ to 3 $\Delta_{\min}$.

The above mid-rise/mid-tread switch has been used successfully in a sub-band coder. It greatly improved the performance of the sub-band coder when it was driven at the low end of its dynamic range. The slight amount of center-clipping introduced by the mid-tread characteristic (at very low input levels) was found to be greatly preferred to the low-level tones and idle channel noise of the mid-rise quantizer characteristic. Figs. 3a and 3b show the results of a sub-band coder output, as the quantizer step-sizes are decaying, for coders with and without the mid-rise/mid-tread switch. As seen in the figure, the low level tones and the coder noise are completely eliminated by the mid-rise/mid-tread switch.

REFERENCES

1. N. S. Jayant, "Digital Coding of Speech Waveforms: PCM, DPCM, and DM Quantizers," Proc. IEEE, *62* (May 1974), pp. 611–632.
2. N. S. Jayant (ed.) *Waveform Quantization and Coding*, New York IEEE Press, 1976.
3. R. E. Crochiere, S. A. Webber, and J. L. Flanagan, "Digital Coding of Speech in Sub-bands," BSTJ, *55*, No. 8 (October 1976), pp. 1069–1085.
4. R. E. Crochiere, "On the Design of Sub-band Coders for Low Bit Rate Speech Communications," B.S.T.J., *56*, No. 5 (May–June 1977), pp. 747–770.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 8, October 1978
Printed in U.S.A.

Zone-Balanced Networks and Block Designs

By F. R. K. CHUNG

(Manuscript received November 30, 1977)

A switching network can be viewed as a collection of interconnected crosspoints that provides connection between input terminals and output terminals. Many networks have the property that the set of input terminals and output terminals can be partitioned into a number of zones such that the requests for connection between an input terminal and an output terminal in the same zone are more likely than those connecting terminals in different zones. In this paper, we study the structure of switching networks of this type by the use of block designs.

I. INTRODUCTION

We shall consider multistage switching networks composed of rectangular switches. For an input terminal u and an output terminal u', the *channel graph for u and u'*, denoted by $G(u,u')$, is defined to be the union of all paths that can be used to connect u and u'. (A channel graph is also called a linear graph.) A network is said to be *balanced* if all channel graphs $G(u,u')$, where u is in the set I of input terminals and u' is in the set Ω of output terminals, are isomorphic.[1,2] A network is said to be *zone-balanced* if it has two nonisomorphic channel graphs, say G_1 and G_2, so that the channel graph $G(u,u')$ is isomorphic to G_1 if u and u' are in the same zone, and $G(u,u')$ is isomorphic to G_2 if u and u' are in different zones. (See Fig. 1. Note that the switches in the network can be viewed as nodes of the corresponding graph.) G_1 is called the *internal graph* and G_2 is called the *external graph* of the switching network.

A zone-balanced network usually consists of three parts. The primary part consists of a few stages where traffic distribution takes place within each zone so that each input terminal has sufficient access to the central stage. The secondary part is the central stage which provides interconnections between different zones. The tertiary part plays the same role for the output terminals as the primary part does for the input terminals.

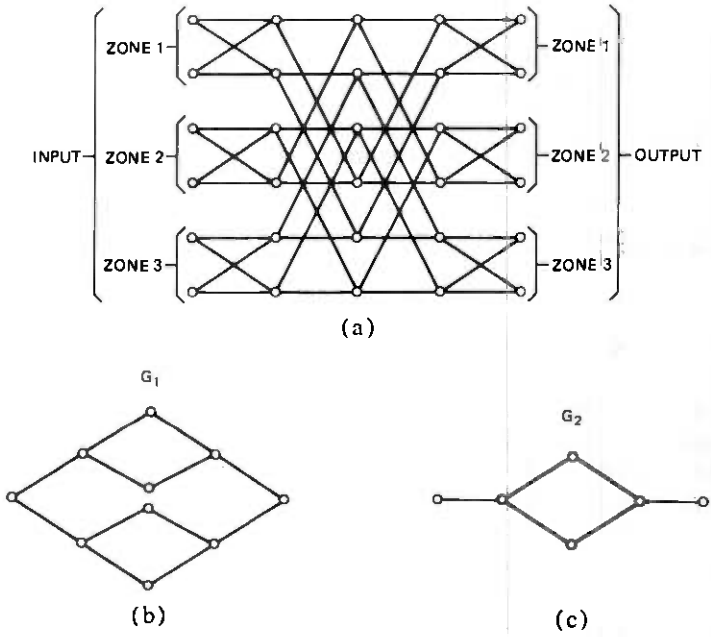


Fig. 1—(a) A zone-balanced network. (b) The internal graph. (c) The external graph.

The primary part is, in fact, composed of a number of *distribution networks*, each of which provides traffic distribution within each zone. We first investigate distribution networks in Section II. In Section III, we introduce the concept of block designs, which will then be used for connecting the distribution networks in the primary part and the switches in the central stage (Sections IV and V). In Section VI, we study zone-balanced networks of more general types.

II. DISTRIBUTION NETWORKS

Figure 2a illustrates an example of a distribution network. The *distribution graph* for an input terminal u is defined to be the union of all paths containing u. In a distribution network, the distribution graphs for any two input terminals are isomorphic. The distribution graph of the distribution network in 2a is shown in 2b. The labeling function f is explained later in this paper.

We consider a distribution network M_s, which is an s-stage network with switches in stage i having size $n_i \times m_i$ for $1 \leq i \leq s$. The distribution graph is then as shown in Fig. 3.

The switch sizes ($n_i \times m_i$, $1 \leq i \leq s$) and the number s of stages are generally dependent upon the traffic loads and the number of input terminals in each zone in order to reduce the "cost" (the number of

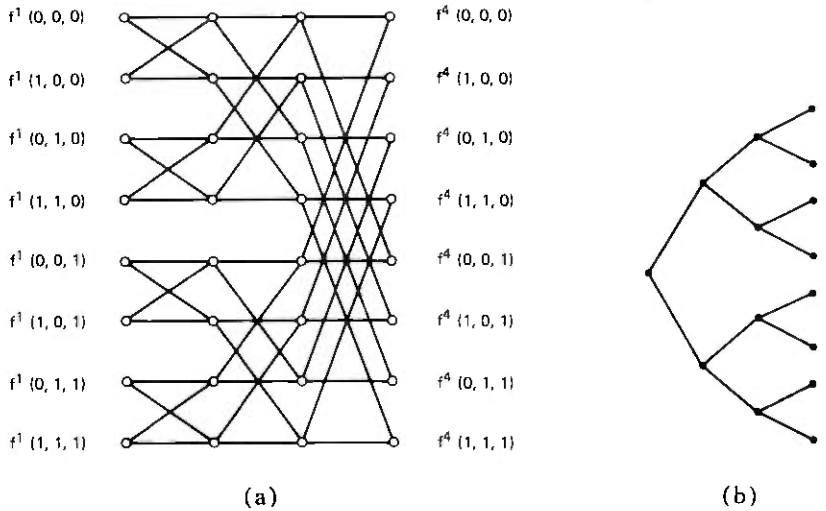


Fig. 2—(a) A distribution network. (b) The distribution graph.

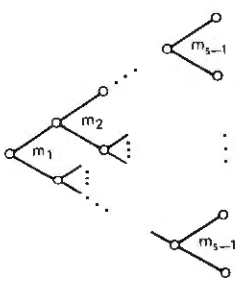


Fig. 3—A general distribution graph.

crosspoints) of the network.[3,4] Here we will only study the structure of distribution networks and will not be concerned with the complicated problem of determining s and n_i, m_i, $1 \leq i \leq s$.

A distribution network can usually be constructed recursively. Figure 4 illustrates a *complete* distribution network in which the number of inlet lines is the product of n_i, $1 \leq i \leq s$, and the number of outlet lines is the product of m_i. We note that, in a complete distribution network, we have $p = p'$ in Fig. 4 and exactly one link exists between a copy of M_{i-1} and a switch in the last stage of M_i for $1 \leq i \leq s$.

Here is an explicit method for the interconnection in the complete distribution network M_s. While the notation may appear at first to be somewhat complicated, it will turn out to be very useful and precise in specifying the link connections of the network.

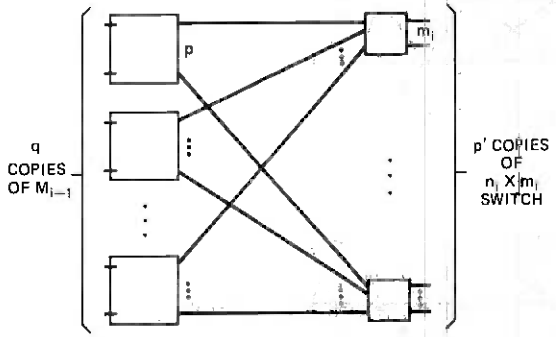


Fig. 4—A complete distribution network.

Let $f^j(i_1,i_2,\ldots,i_{s-1})$ denote the $(i_1 + i_2m_1 + \ldots + i_{s-1}m_1 \ldots m_{s-2} + 1)$th switch in stage j where $0 \le i_q \le m_q$, $1 \le q \le s-1$.

Let $f^j_q(i_1,i_2,\ldots,i_{s-1})$ denote the $(q+1)$th outlet line of the switch $f^j(i_1,\ldots,i_{s-1})$ where $0 \le q < m_j$. Then we have:

$$f^j_q(i_1,\ldots,i_{s-1}) \text{ is connected to}$$

$$f^{j+1}(i_1,\ldots,i_{j-1},q,i_{j+1},\ldots,i_{s-1})$$

for all j, $1 \le j < s$.

Sometimes the values of n_i, m_i, $1 \le i \le s$, do not allow complete connections between consecutive stages, i.e., $p < p'$ and the number of inlet lines of the zone is not equal to the product of n_i, $1 \le i \le s$ (see Fig. 5). These distribution networks are said to be incomplete distribution networks. In this case, an appropriate connection, according to some rules to assign links cyclically, usually can result in a distribution network of this type (called cyclic distribution network). Here we give a scheme of constructing a cyclic distribution network in which copies of M_{i-1} and

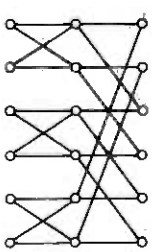


Fig. 5—An incomplete distribution network.

the switches of the last stage of M_i are connected according to the connection pattern of a regular bipartite graph as follows.

Consider a bipartite graph with sets of nodes A and B. A link connects a node in A to a node in B. A bipartite graph is said to be *regular* if the degree of each node in A is equal to some integer x and the degree of each node in B is equal to some integer y. It then follows that $ax = by$ where $a = |A|$, $b = |B|$. We restrict ourselves to the case that $a \geq y$, $b \geq x$. We will show that the condition $ax = by$ is sufficient for the existence of such a regular bipartite graph. We consider the following two possibilities.

Case 1: a and y are relatively prime. In this case, y divides x. Let $\alpha_1, \ldots, \alpha_a$ denote nodes in A and $\beta_1, \ldots, \beta_b$ denote nodes in B. We will then connect α_i to β_j where $j \equiv i\,x/y + k \pmod{b}$ where $1 \leq k \leq x$, and $1 \leq j \leq b$. It is easy to see that the resulting graph is a regular bipartite graph (see Fig. 6a).

Case 2: Let d be the greatest common divisor of a and y. Then we can construct, by Case 1, a regular bipartite graph G' on sets of nodes A' and B' where $|A'| = a' = a/d$ and $|B'| = b$. The degrees of nodes in A' are all equal to $x' = x$ and the degrees of nodes in B' are all equal to $y' = y/d$. Now, we construct the regular bipartite graph on A and B as follows: The set A can be viewed as d copies of A'. The connection between each copy of A' and B is the same as G'. It is easily verified that the resulting graph is regular and a node in A or B has degree x or y, respectively (see Fig. 6b).

Now we can construct the incomplete distribution network M_i by connecting a copies of M_{i-1} and b copies of $n_i \times m_i$ switches if $ap = bn_i$ (p is the number of outlet lines of M_{i-1}), according to the regular bipartite graph we described above by taking each copy of M_{i-1} as a node in A and each switch $n_i \times m_i$ as a node in B.

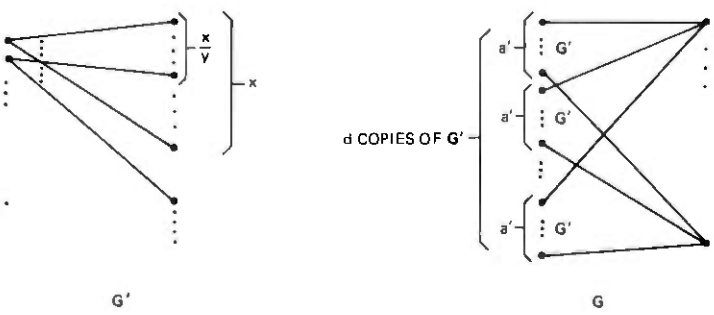


Fig. 6—Regular bipartite graphs.

We note that there are many nonisomorphic regular bipartite graphs of the given sizes. Therefore, there are many different incomplete distribution networks of the same "cost" (number of crosspoints). In Section IV and V we see some incomplete distribution networks obtained by some methods other than the cyclic connections described here. Those methods are based on a basic concept in combinatorics called a *block design*, which we next describe.

III. BLOCK DESIGNS

A (v,b,r,k,λ) block design is a family of subsets $X_1,X_2,\ldots,X_b$ of a v-element set X, satisfying the following conditions:

(*i*) Each X_i has k elements, $1 \le i \le b$.

(*ii*) Each 2-element subset of X is a subset of exactly $\lambda > 0$ of the sets $X_1,\ldots,X_b$.

Properties (*iii*) and (*iv*) follow immediately from (*i*) and (*ii*).

(*iii*) Each element of X is in exactly r of the sets $X_1,\ldots,X_b$.

(*iv*) $r(k-1) = \lambda(v-1)$ and $bk = vr$.

For example, the following is a (7,7,3,3,1) block design.

$$X_i = \{i, i+1, i+3\} \pmod 7 \text{ for } 1 \le i \le 7.$$

The reader is referred to Refs. 5 and 6 for the existence and construction of various classes of block designs.

IV. THREE-STAGE ZONE-BALANCED NETWORKS

Let $X_1,\ldots,X_b$, be a (v,b,r,k,λ) block design. We will construct a three-stage zone-balanced network having internal graphs containing r paths and external graphs containing λ paths (see Fig. 7a and b), and having v switches in the first or third stage. The input terminals of the same zone go to the same switch. Let $y_1,\ldots,y_v$ be switches in the first

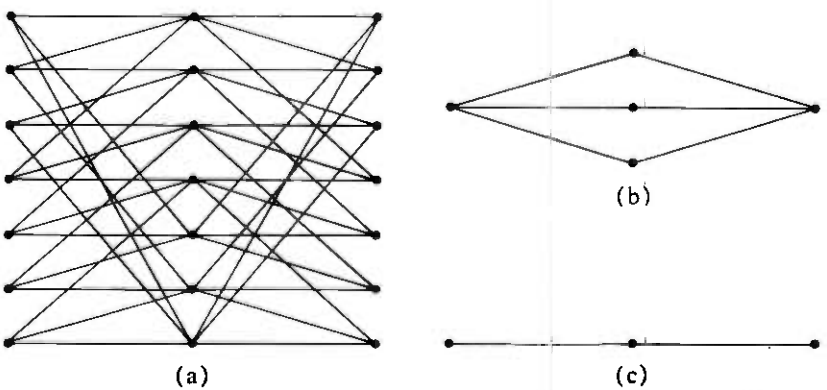


Fig. 7—(a) A zone-balanced network constructed by using (7,7,3,3,1) block design. (b) The internal graph. (c) The external graph.

stage (the primary part), each of which has r output lines. Let $z_1, \ldots, z_b$ be switches of size $k \times k$ in the second stage (the secondary part). Then we connect y_i to z_j if and only if i is an element of X_j (see Fig. 7 for the (7,7,3,3,1) block design). The connection between the second and the third stage is a mirror image of that between the first and the second stage. It is easily verified that the resulting network is zone-balanced and has internal and external graphs as shown in Fig. 7b and c, respectively. This is the basic model of zone-balanced networks.

V. MULTISTAGE ZONE-BALANCED NETWORKS

In this section, we give explicit constructions for various types of multistage zone-balanced networks. Roughly speaking, we construct these multistage zone-balanced networks by replacing each switch in the first or the last stage of the three-stage zone-balanced network described in Section III by a distribution network. The internal graphs in these zone-balanced networks depend on the switch sizes in the distribution networks. Suppose a distribution network has s stages and has switch of size $n_i \times m_i$ in stage i, $1 \leq i \leq s$. Then the internal graph in the zone-balanced network has $m_1 m_2 \ldots m_s$ paths as shown in Fig. 8a. The external graph depends heavily upon the linking pattern between the distribution networks in the primary part (or tertiary part) and the switches in the central stage. In general, it would be desirable to have the external graph as "spread-out" as possible. We will not define "spread-out" rigorously here. For example, the graph in Fig. 10c is more spread-out than the graph in Fig. 1c, although they contain the same number of paths. It can be shown[7,8] that the more spread-out the channel graph is, the less the traffic congestion will be. The external graph in the zone-balanced network we construct will be as shown in Fig. 8b.

In this section, we restrict ourselves to the case in which the number of input terminals is the same as the number of output terminals and the zone sizes are equal, i.e., $|I| = |\Omega|$, $I = I_1 \cup \ldots \cup I_v$, $\Omega = \Omega_1 \cup \ldots \cup \Omega_v$

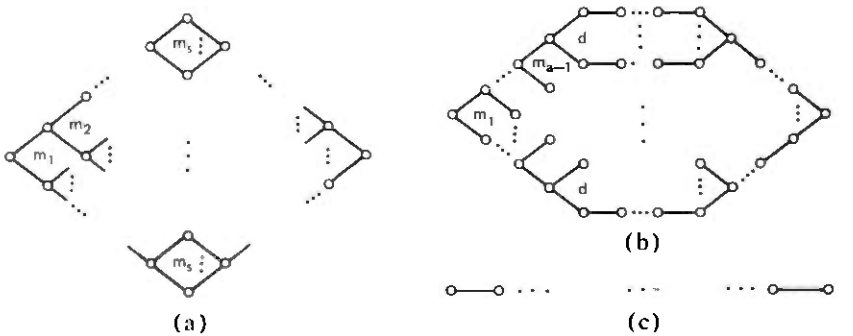


Fig. 8—(a) A general multistage internal graph. (b) A general multistage external graph. (c) A path.

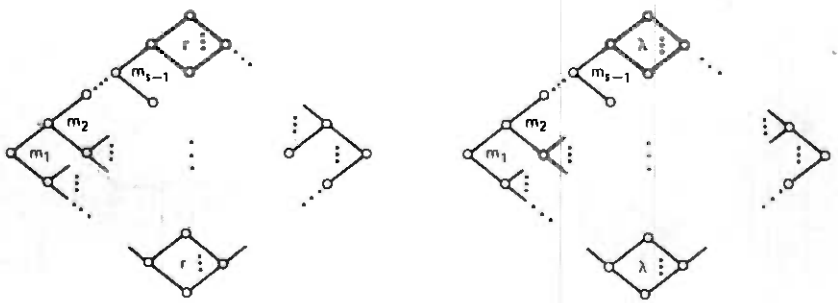
and $|I_i| = |\Omega_j|$, $1 \le i,j \le v$. An input terminal u, $u \epsilon I_i$, and an output terminal u', $u' \epsilon \Omega_j$, are said to be in the same zone if and only if $i = j$.

We first present a simple technique, called a Type 1 construction, for construcing zone-balanced networks. The distribution network M_s in the primary part has the property that every switch in the last stage has r output lines, i.e., $m_s = r$. Let us assume M_s has q switches in the last stage. Thus, M_s has qr output lines. We will construct a zone-balanced network with internal graphs containing qr paths and external graphs containing $q\lambda$ paths (see Fig. 9a and b) as follows.

Let the central stage consist of bq switches of size $k \times k$. Then the q'th switch of the v'th copy of M_s in the primary part is connected to the $[(b'-1)q + q']$th switch in the central stage for any b' with $v' \epsilon X_{b'}$. We connect the central part and the tertiary part in the same way (symmetrically) that the primary and central parts are connected.

For example, using the (3,3,2,2,1) block design $X_1 = \{1,3\}$, $X_2 = \{1,2\}$, $X_3 = \{2,3\}$, we obtain the network shown in Fig. 10. We note that the networks in Fig. 1a and Fig. 10a have the same number of crosspoints, but the channel graph in Fig. 10c is more spread-out than the channel graphs containing r paths, provided a (v,b,r,k,λ) block design exists in Fig. 1a, though their "cost" (number of crosspoints) are the same.

Now we give a construction (Type 2) of another class of zone-balanced networks which has external graphs containing a path and internal graphs containing r paths, provided a (v,b,r,k,λ) block design exists where $\lambda = 1$. The primary part consists of v copies of a complete distribution network which has r output lines. Thus, r is the product $m_1 m_2 \ldots m_s$. The central stage consists of b switches of size $k \times k$. We will connect the r output lines of the distribution network to the switches in the central stage as follows: One of the output lines of the v'th copy of the distribution network will be connected to the b'th switch in the central stage for any b' such that v' is contained in the block $X_{b'}$. In Fig. 11, we have an example which is constructed by using a (9,12,4,3,1) block design. It can be easily verified that the internal graph is as shown in Fig. 8a and the external graph is as shown in Fig. 8c.



(a) (b)

Fig. 9—(a) An internal graph. (b) An external graph.

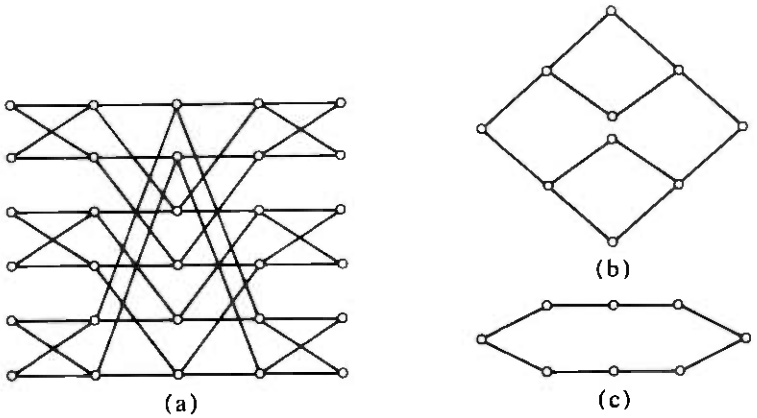


Fig. 10—(a) A type-1 zone-balanced network constructed by using (3,3,2,2,1) block design. (b) The internal graph. (c) The external graph.

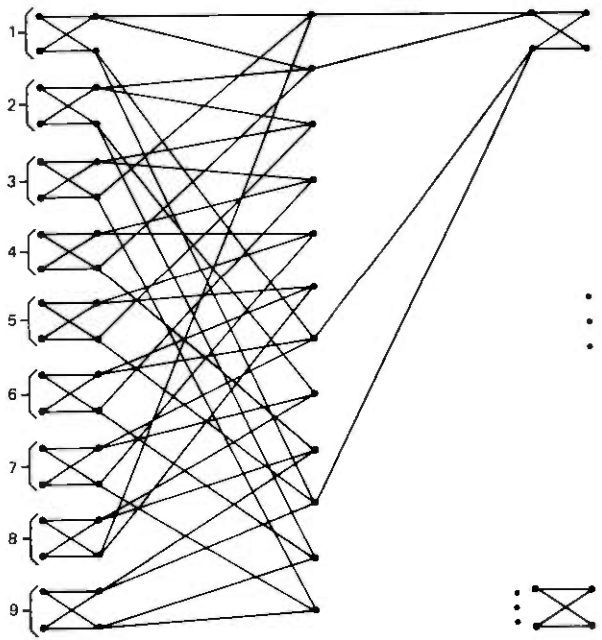


Fig. 11—A type-2 zone-balanced network constructed by using (9,12,4,3,1) block design.

Type 3 zone-balanced networks are variations of Type 2 zone-balanced networks. The primary part consists of copies of a distribution network which is not necessarily a complete distribution network. In Figure 12a, we have an example which is constructed by using (3,3,2,2,1) design. Its internal and external graphs are shown in Figs. 12b and c, respectively. The distribution network usually has *rw* output lines for some integer

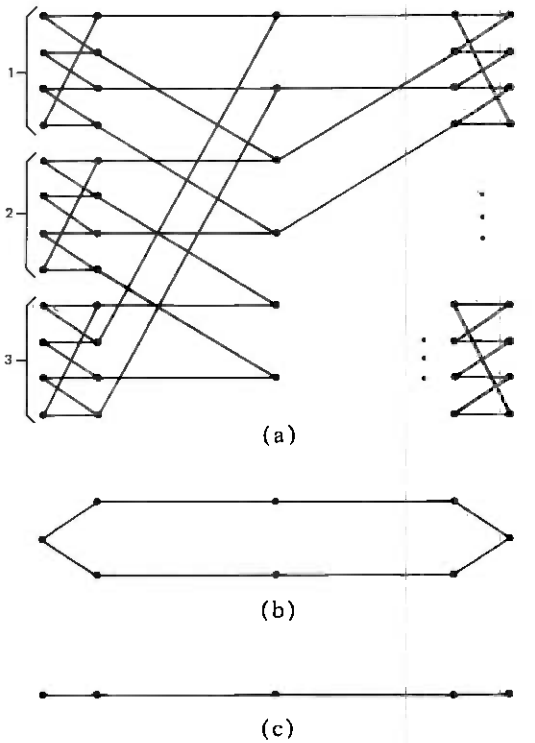


Fig. 12—(a) A type-3 zone-balanced network constructed by using (3,3,2,2,1) block design. (b) The internal graph. (c) The external graph.

w. The central stage consists of bw switches which can be partitioned into b subsets, each of which has w switches. Among the rw output lines of the v'th copy of the distribution network, there are w output lines which are connected to the w switches, respectively, in the b'th subset if $X_{b'}$ contains v' and any input line of the distribution network does not have access to more than one output line of these w output lines. The example in Fig. 12 more or less reveals the scheme for constructing zone-balanced networks of this type. The detail of the construction is omitted.

Type 4 zone-balanced networks are generalizations of Type 2 zone-balanced networks. A Type 4 zone-balanced networks has external graph containing λw paths and internal graphs containing rw paths, provided a (v,b,r,k,λ) block design exists. The primary part consists of v copies of a complete distribution network which has s stage and has switches of size $n_i \times m_i$ in ith stage. Thus, rw is the product $m_1 m_2 \ldots m_s$. Since we want the external graph as spread-out as possible, we will restrict ourselves to use block designs with $\lambda = 1$. (By using block design with $\lambda > 1$, the subsequent scheme for constructing Type 4 zone-balanced

network can be applied with some modification. However, the resultant networks will usually have internal graphs which are not as spread-out as possible.) The internal graph and the external graph are as shown in Fig. 8b. Thus, we will choose w in the form $m_1 m_2 \ldots m_{a-1} d$ where d divides m_a. The connection between the primary part and the central part is more complicated than that in the Type 2 construction. However, it can be explicitly specified by the following simple method.

Let $f_q^s(i_1, \ldots, i_{s-1}; v')$ denote the $(q+1)$th outlet lines of the $(i_1 + i_2 m_1 + \ldots + i_{s-1} m_1 \ldots m_{s-2} + 1)$th switch in stage s of the v'th copy of M_s.

The central stage of the zone-balanced network consists of bw switches of size $k \times k$. We define

$$f^c(i_1, \ldots, i_a; b') \text{ to be the}$$

$$(i_1 + i_2 m_1 + \ldots + i_a m_1 \ldots m_{a-1} + (b'-1)\, m_1 \ldots m_{a-1} d + 1)\text{th switch,}$$

where

$$0 \le i_q < m_q,\ 1 \le q < a, 0 \le i_a < d \text{ and } 0 < b' \le b.$$

Let $X_1, X_2, \ldots, X_b$ denote the sets of a (v,b,r,k,λ) block design with $\lambda = 1$. For any element $y \epsilon X = \cup_i X_i$, we say the ith y-set is X_j if X_i is the ith set containing y, i.e., $|\{X_q : y \epsilon X_q, 1 \le q \le j\}| = i$.

First, we consider the special case when $d = m_a$.

$$f_q^s(i_1, \ldots, i_{s-1}; v') \text{ is connected to}$$
$$f^c(i_1, \ldots, i_a; b')$$

if $v' \epsilon X_{b'}$ and the

$$(i_{a+1} + i_{a+2} m_{a+1} + \ldots + i_{s-1} m_{a+1} \ldots m_{s-2} + q m_{a+1} \ldots m_{s-1} + 1)\text{th } v'\text{-set}$$

is $X_{b'}$.

Now, if d is a proper divisor of m_a, the above scheme has to be modified slightly. Note that i_a can be written as $i_a' + i_a'' d$ where $0 \le i_a' < d$. Then we have:

$$f_q^s(i_1, \ldots, i_{s-1}; v') \text{ is connected to}$$
$$f^c(i_1, \ldots, i_{a-1}, i_a'; b')$$

if $v' \epsilon X_{b'}$ and the

$$(i_a' + i_{a+1} d + \ldots + i_{s-1} d m_{a+1} \ldots m_{s-2} + q d m_{a+1} \ldots m_{s-1} + 1)\text{th } v'\text{-set}$$

is $X_{b'}$.

We note that the f_q^s (or f^c) is just a digital expression for address as-

signments of the output lines in stage s. The last $s - a$ digits (i.e., $i_a, \ldots, i_{s-1}$ of $f_q^s(i_1, \ldots, i_s)$) are used to find the location of the "block" of switches in the central stage and the first a digits are used to specify the location of the switch in that block to which the output lines of $f^s(i_1, \ldots, i_{s-1})$ should be connected. Figure 13a gives an example using the (13,13,4,4,1) design where $X_1 = \{i, i+1, i+3, i+9\} \pmod{13}$. Its internal and external graphs are also shown in Fig. 13b and c.

We note that, by modifying the construction mentioned above, we could easily obtain zone-balanced networks having external graphs not necessarily as spread-out as possible. For example, let us modify the definition of f_q^s, the digital expression for address assignments of output lines of the switches in stage s. Instead of using the last a digits, we use the first a digits to assign the location of the "block" of switches in the central stage and use the last $s - a$ digits to specify the location of the switch in the "block" to which the output line should be connected. The resultant zone-balanced network then has the external graph which is the least spread-out channel graph among all possible graphs having the same number of paths. In general, we can use arbitrary x digits to specify the location of the block (as long as the necessary condition on divisibility is satisfied) to obtain zone-balanced networks having various external graphs.

We also note that it is possible to derive a generalized version of Type 3 zone-balanced networks using incomplete distribution networks. However, it is more complicated and has more constraints than Type 4 construction. We shall not discuss it here.

VI. ZONE-BALANCED NETWORKS OF MORE GENERAL TYPES

We note that the right half of a zone-balanced network is itself an incomplete distribution network. These incomplete distribution networks, called BD-distribution networks, seem to distribute traffic more evenly and have richer combinatorial properties than the cyclic incomplete distribution networks described in Section II. One of the obvious reasons is that a BD-distribution network together with its mirror image gives a zone-balanced network, whereas a cyclic incomplete distribution network does not.

Suppose the set of input and output terminals can be partitioned into a number of zones which can themselves then be partitioned into several areas such that requests for connecting terminals in the same area (zone) are more likely than those for connecting terminals in different areas (zones). In such a network, there usually are four channel graphs, say G_1, G_2, G_3, G_4, such that $G(u,u')$ is isomorphic to G_1, if u and u' are in the same zone and area; $G(u,u')$ is isomorphic to G_4 if u and u' are in different zones and areas; $G(u,u')$ is isomorphic to G_2 if u and u' are in the same zone but different areas; $G(u,u')$ is isomorphic to G_3 if u and

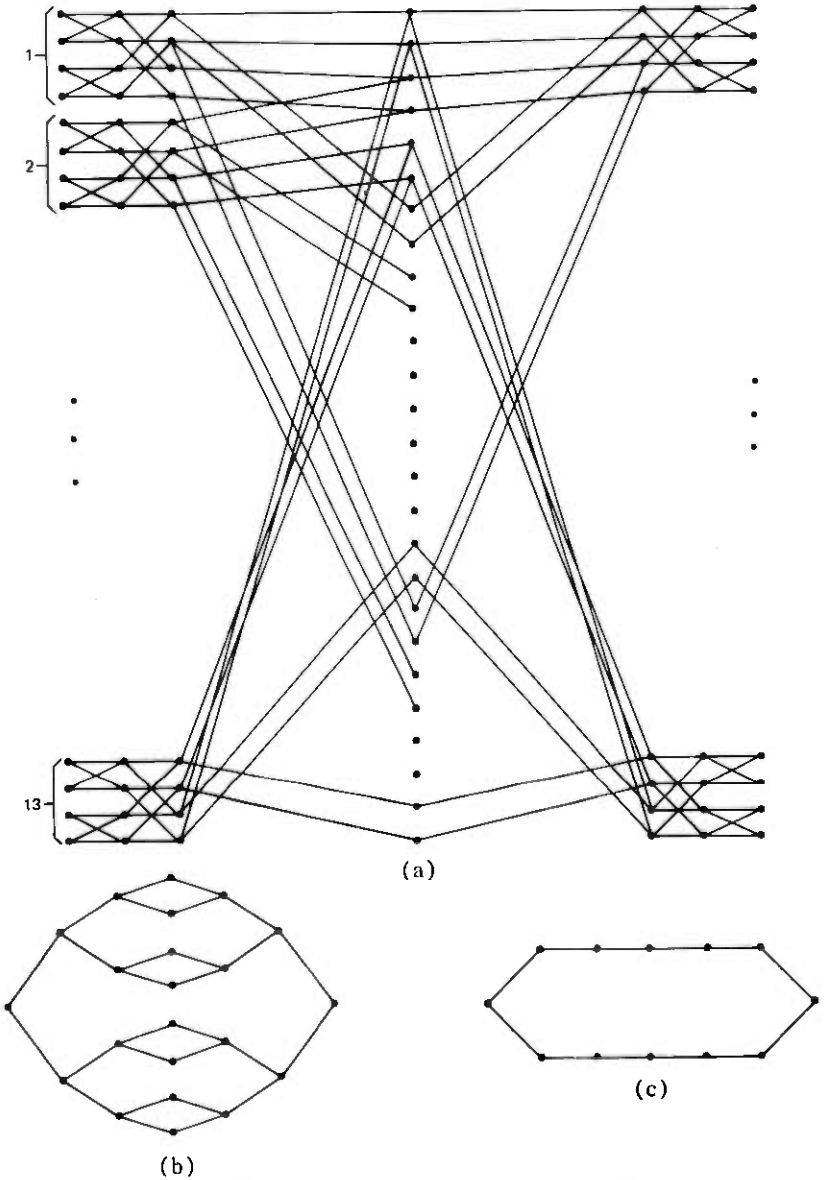


Fig. 13—(a) A type-4 zone-balanced network constructed by using (13,13,4,4,1) block design. (b) The internal graph. (c) The external graph.

u' are in the same area but different zones. These networks will be called multizone-balanced networks, and can be built by schemes similar to those described in Section V. A multizone-balanced network can also be viewed as a combination of three parts. However, the primary part usually consists of copies of a BD-distribution network.

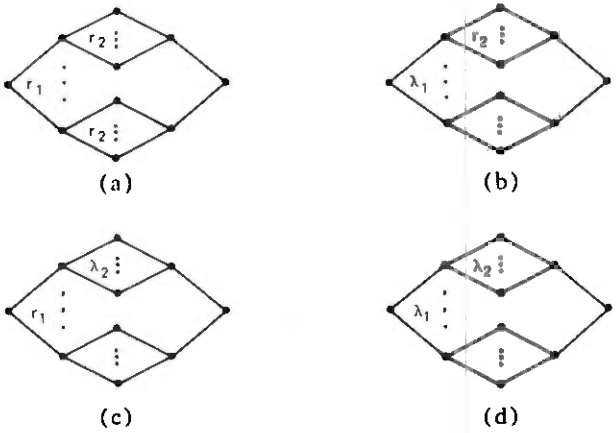


Fig. 14—Some channel graphs for multizone-balanced networks.

A simple model of a five-stage multizone-balanced network can be described as follows

Let $X_1, \ldots, X_{b_1}$ be a $(v_1, b_1, r_1, k_1, \lambda_1)$ block design and $Y_1, \ldots, Y_{b_2}$ be a $(v_2, b_2, r_2, k_2, \lambda_2)$ block design. The distribution network M_s for each zone is exactly the right half of the three-stage zone-balanced network described in Section IV, and the number of output lines of a switch in the last stage of M_s is r_2. The primary part consists of v_2 copies of M_s, and the central part consists of switches of size $k_2 \times k_2$. The ith switch of the jth copy of M_s is connected to the $((b' - 1)b_1 + i)$th switch if j is an element of $Y_{b'}$. Again, the tertiary part is connected to the central part in the same way (symmetrically) that the primary and central parts are connected.

It is easy to verify that the channel graphs G_1, G_2, G_3, G_4 are as shown in Fig. 14.

We could modify the above multizone-balanced network by replacing the switches in the primary part or the tertiary part by distribution networks using the same method used in Section V. However, we will not discuss this here.

Suppose the number of input terminals and the number of output terminals are not equal, say $|\Omega|$ is a multiple of $|I|$. We can still construct zone-balanced networks by combining copies of zone-balanced networks which have the same number of input and output terminals.

Hagelbarger[9] first proposed the use of block designs for constructing switching networks in 1973. Some other techniques for constructing various classes of networks have been investigated in Refs. 10 and 11. Hopefully, more structures in combinatorics and graph theory can be employed as useful techniques to construct interesting classes of switching networks.

REFERENCES

1. F. R. K. Chung and F. K. Hwang, "A Problem on Blocking Probabilities in Connecting Networks," Networks, *7* (1977), pp. 185–192.
2. F. K. Hwang, "Balanced Networks," Conference Record of 1976 International Conference on Communications, pp. 7-13–7-16.
3. F. R. K. Chung, "Optimal Multistage Networks," IEEE Transactions on Communications, to appear.
4. A. Lotze, "Optimum Link Systems," 5th ITC, New York, 1967, pp. 242–251.
5. R. J. Collens, private communication.
6. M. Hall, Jr., *Combinatorial Theory*, Waltham, Mass: Blaisdell, 1967.
7. F. R. K. Chung and F. K. Hwang, "On Blocking Probabilities for Switching Networks," B.S.T.J., *56*, No. 8 (October 1977), pp. 1431–1446.
9. D. W. Hagelbarger, unpublished work.
10. F. R. K. Chung, "On Switching Networks and Block Designs," Conference Records of the Tenth Asilomar Conference on Circuits, Systems, and Computers, 1976, pp. 212–218.
11. F. K. Hwang and S. Lin, "Construction of Balanced Switching Networks," to appear in Networks.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 8, October 1978
Printed in U.S.A.

The Construction for Symmetrical Zone-Balanced Networks

By F. K. HWANG and T. C. LIANG

(Manuscript received January 27, 1978)

For many real networks, the input and output switches can often be partitioned into subsets called zones such that switches in the same zone have a certain community of interest and are more likely to request a connection. Such a network will be called a zone-balanced network if, among other regularity conditions, the channel graph between an input switch u and an output switch v is either isomorphic to a graph G_1 if u and v are in the same zone, or isomorphic to a graph G_2 if otherwise. In this paper, we continue the study of using balanced incomplete block design for the construction of zone-balanced networks. We introduce some new methods to construct a wide class of such networks, which include some previous constructions as special cases.

I. INTROOUCTION

The topology of a switching network can often be represented by a graph by taking switches as vertices and links as edges. By this representation, a *multistage (switching) network* is a graph the vertex-set of which can be naturally partitioned into s subsets $V_1,\ldots,V_s$ and the edge-set into $s-1$ subsets $E_1,\ldots,E_{s-1}$, for some number s, so that E_i connects V_i to V_{i+1}. (We do not allow multiple edges between two vertices.) Vertices in V_1 correspond to the input switches of the network and vertices in V_s correspond to the output switches. Let $u\epsilon V_1$ and $v\epsilon V_s$. Then the *channel graph* $G(u,v)$ is the union of all paths connecting u to v in the network. A multistage network is said to be *regular* if every vertex in V_i has the same number of edges in E_{i-1} and the same number of edges in E_i. A regular multistage network is *balanced* if the channel graphs $G(u,v)$ over all $u\epsilon V_1$ and all $v\epsilon V_s$ are isomorphic.

For many real networks, the input and output switches can often be partitioned into subsets called *zones* such that switches in the same zone have a certain community of interest and are more likely to request a connection. (In this paper, we are concerned only with connection be-

tween an input switch and an output switch.) Such a network will be called a *zone-balanced network* if it is regular and there exists two graphs G_1 and G_2 so that $G(u,v)$ is isomorphic to G_1 if u and v are in the same zone and $G(u,v)$ is isomorphic to G_2 if not. G_1 and G_2 will be referred to as the *intrazone* and the *interzone* channel graphs, respectively. A zone-balanced network is said to be *symmetrical* if it is symmetrical with respect to the center stage or the two stages in the middle.

A *balanced incomplete block design* (abbreviated as BIBD) with parameters (v,b,r,k,λ) is a family of *blocks*, with each block being a k-subset of the set $\{1,2,\ldots,v\}$, satisfying the following properties:

(*i*) Every element in the set $\{1,2,\ldots,v\}$ appears in exactly r blocks.
(*ii*) Every pair of elements in the set $\{1,2,\ldots,v\}$ appears together in exactly λ blocks.

BIBDs have long been a favorite subject for mathematicians and statisticians. The reader is referred to Ref. 1 for the existence and construction for many BIBDs. The use of BIBDs for constructing zone-balanced networks was first studied in Ref. 2. Some further constructions were given in Ref. 3. In this paper, we give some methods for such constructions. The zone-balanced networks constructed previously, as well as in this paper, are all symmetrical.

II. SOME PRELIMINARY RESULTS

A zone-balanced network is called *canonical* if each zone consists of a single input switch and a single output switch. Therefore, a CZBN (canonical zone-balanced network) can be viewed as a prototype for a full-fledged network with the same interzone and intrazone channel graphs. The mechanism for expanding a CZBN into a full-fledged network is the operation of "parallel expansion," which was first introduced by Takagi[4] and by Timperi and Grillo.[5] For an s-stage network N, a (k,j) left (right) parallel expansion means taking k copies of N and identifying their subgraphs from stage j to stage s (from stage 1 to stage j). Figure 1 gives some examples of parallel expansion. It is clear that parallel expansion preserves the isomorphisms of the interzone and intrazone channel graphs.

Next we introduce a method which we will use later to describe the connection between switches in two adjacent stages. To use this method, every switch in the two adjacent stages should be labeled by a subset of a given set. Then two switches in the adjacent stages should be connected if the label of one is contained in the label of the other. This type of connection will be called a *labeled-subset connection.*

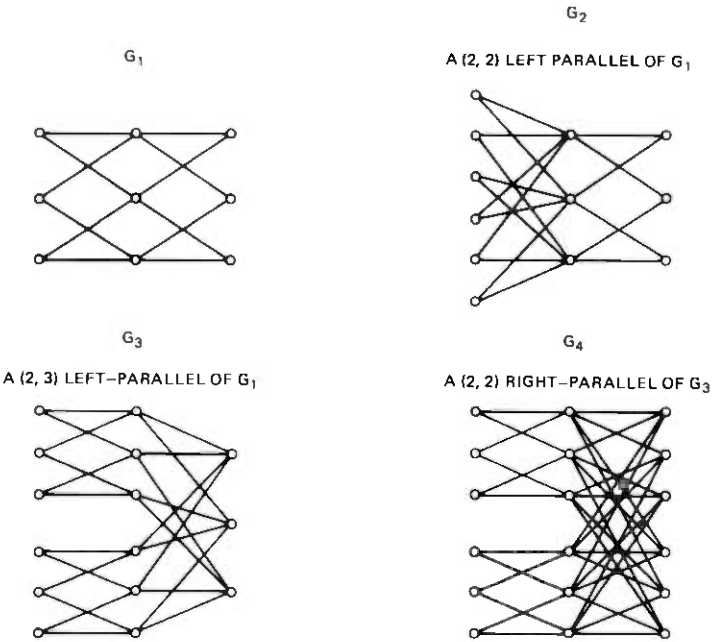


Fig. 1—Examples of parallel expansion.

III. A RECURSIVE CONSTRUCTION FOR CZBN

A 2-stage CZBN with v zones is necessarily a $v \times v$ complete bipartite graph; hence, its construction is trivial. We now give a construction for a 3-stage CZBN (noting that a 3-stage channel graph is uniquely determined by its number of paths).

Theorem 1: Suppose that a (v,b,r,k,λ)*-BIBD exists. Then we can construct a 3-stage CZBN with* v *zones which has* r *paths in its intrazone channel graph and* λ *paths in its interzone channel graph.*

Proof: Take b switches of V_2 and label each of them by a distinct block of the given BIBD. Take v switches of $V_1(V_3)$ and label each of them by a distinct element of $Z = \{1,2,\ldots,v\}$. Apply a labeled-subset connection between V_2 and $V_1(V_3)$. It is easy to verify that the resulting network is the one specified in Theorem 1.

Example 1: Let the given BIBD have parameters (7,7,3,3,1) and have blocks (1,2,4), (2,3,5), (3,4,6), (4,5,7), (5,6,1), (6,7,2), and (7,1,3). Figure 2 gives a 3-stage CZBN with 7 zones.

We now give a recursive construction for a symmetrical s-stage CZBN for $s \geq 4$.

Theorem 2: Suppose that an s*-stage CZBN with* k *zones exists which has* G_1 *and* G_2 *as its intrazone and interzone channel graphs. Fur-*

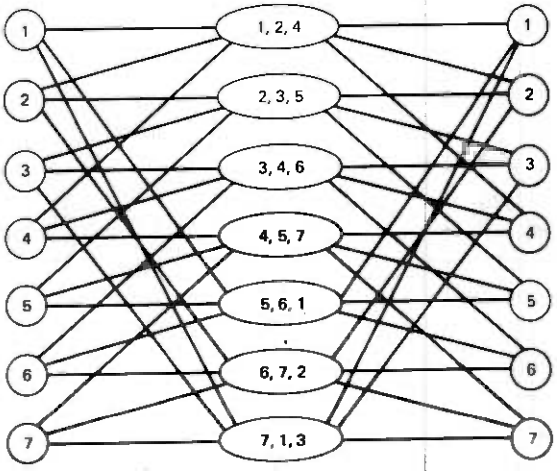


Fig. 2—A 3-stage CZBN.

thermore, suppose that a (v,b,r,k,λ) *BIBD exists. Then there exists an* $(s+2)$*-stage CZBN with* v *zones which has the channel graphs as shown in Fig. 3.*

Proof: Let N be the given s-stage CZBN and assume that every input (output) switch of N is labeled by the zone it belongs to. Take b copies of N and let N_i denote the ith copy. Replace the k zones in N_i by the k elements in the ith block of the given BIBD. Take v switches of $V_1(V_3)$ and label each switch by a distinct element of the set $Z = \{1,2,\ldots,v\}$. Apply a labeled-subset connection between $V_1(V_3)$ and the input (output) switches of the b copies of N. It is easy to verify that the resulting network is indeed the one specified in Theorem 2.

Corollary: *Suppose that a* (v,b,r,k,λ) *BIBD exists. Then we can construct a 4-stage CZBN with* v *zones such that its intrazone channel graph consists of* r *disjoint paths and its interzone channel graph consists of* λ *disjoint paths.*

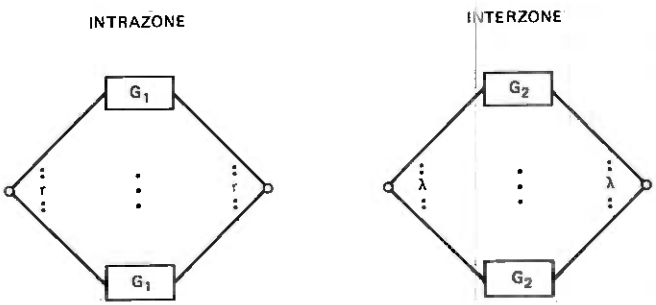


Fig. 3—Channel graphs for Theorem 2.

Example 2: Using the same BIBD as in Example 1, we obtain the 4-stage CZBN as shown in Fig. 4.

IV. A CONSTRUCTION FOR CZBNS USING A BALANCED PARTITION OF BLOCKS

Consider a (v,b,r,k,λ) design, and let F_i denote the subfamily of blocks containing element i. A partition of F_i is said to be *balanced* with parameters (p,d) if the following conditions are satisfied:

(*i*) F_i is divided into p disjoint parts such that each part consists of r/p blocks.

(*ii*) Exactly $d + 1$ distinct elements appear in each part.

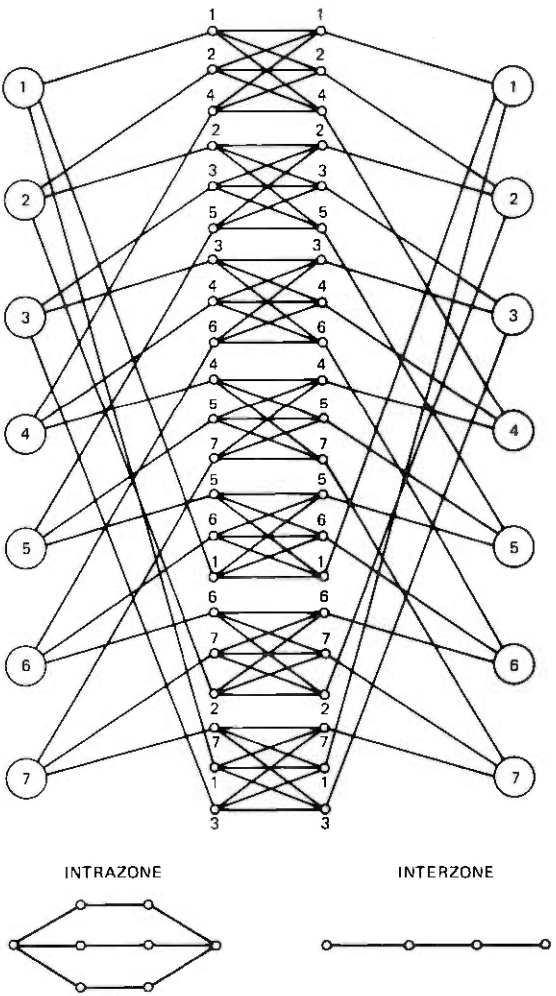


Fig. 4—A 4-stage CZBN.

(*iii*) For any element $i' \neq i$ contained in part j, the number of blocks in part j containing i' is a constant. (From (*i*) and (*ii*), this constant must also be independent of j.)

We say F_i has a 2-step balanced partition with parameters (p_1,d_1,p_2,d_2) if F_i has a balanced partition $P_1,\ldots,P_{p_1}$, with parameters (p_1,d_1) and each P_j has a balanced partition with parameters (p_2,d_2). Similarly, we can define a t-step balanced partition of F_i with parameters $(p_1,d_1,p_2,d_2,\ldots,p_t,d_t)$.

Note that any t-step nested partition of a set N induces a partial ordering which can be represented by a $(t+2)$-level rooted tree. Suppose that N has n elements. Then the first level of the tree corresponds to the crudest partition, namely, a single node representing the set N itself, and the $(t+2)$-level of the tree corresponds to the finest partition, namely, n nodes each representing a single element of N. The t intermediate levels of the tree correspond to the t partitions sequentially. By taking two copies of this tree and identifying their nodes at the $(t+2)$-level, we obtain a $(2t+3)$-stage symmetrical network. This mapping from a nested partition to a multistage network is critically used in the following theorem.

Theorem 3: *Consider a* (v,b,r,k,λ) *BIBD and let* F_i *be the subfamily of blocks containing the element* i. *Suppose that for each* F_i, $i = 1,2,\ldots,v$, *there exists a* t*-step balanced partition with the parameters* $(p_1,d_1,p_2,d_2,\ldots,p_t,d_t)$. *Then there exists a* $(2t+3)$*-stage* ($(2t+4)$*-stage*) CZBN *which has channel graphs as shown in Fig. 5:* $(q = r/\Pi_{l=1}^{t} p_l)$. (To obtain the channel graphs for the $(2t+4)$-stage CZBN, replace each vertex in the center stage by the graph O–O.)

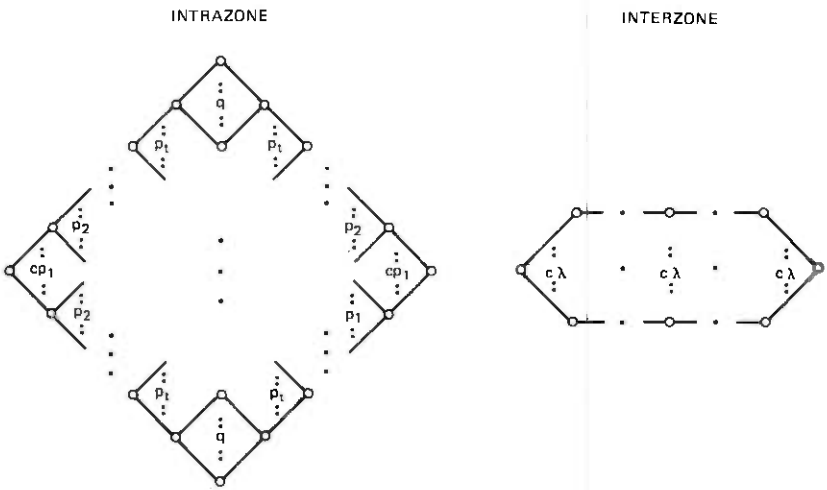


Fig. 5—Channel graphs for Theorem 3.

Proof (only for the case that n is odd): For each F_i, $i = 1,\ldots,v$, construct a $(2t + 3)$-stage symmetrical network by using the given t-step balanced partition. The union of these v networks (they overlap at the center stage, since the F_i's overlap) yields a CZBN with channel graphs as specified in Fig. 5 with $c = 1$. Taking c copies of these networks and identifying their first stages and last stages, we obtain the desired CZBN. That the constructed network is "balanced" is a consequence of the partition being balanced.

Corollary: *Suppose a* $(v,b,r,k,1)$ *BIBD exists and* $\Pi_{l=1}^{t} p_l$ *divides* r. *Then a CZBN with channel graphs as specified in Fig. 5 exists.*

Proof: With $\lambda = 1$, any partition which satisfies condition (i) of a balanced partition is a balanced partition. The same is true for a t-step partition. Therefore, when $\Pi_{l=1}^{t} p_l$ divides r, then a t-step balanced partition with parameters $(p_1,\ldots,p_t)$ always exists (the parameters d_ls are determined by p_ls).

Note that by applying Theorem 2 several times to the network constructed in Theorem 3, we can obtain CZBNs with various types of channel graphs. In particular, we obtain the following:

Theorem 4: *Suppose that a sequence of BIBDs with parameters* $(v_j,b_j,r_j,k_j,\lambda_j)$, $j = 1,2,\ldots,m$ *exists. Furthermore, suppose* $k_j = v_{j+1}$ *for* $j = 1,2,\ldots,m-1$, $\lambda_m = 1$, *and* $\Pi_{l=1}^{t} p_l$ *divides* r_m. *Then there exists a* $(2t + 2m + 1)$*-stage* $((2t + 2m + 2)$*-stage) CZBN which has channel graphs as shown in Fig. 6:* $(q = r_m/\Pi_{l=1}^{t} p_l)$.

Proof: Use the $(v_m,b_m,r_m,k_m,\lambda_m)$ BIBD to construct a $(2t + 3)$-stage CZBN from Theorem 3. Then apply Theorem 2 $m - 1$ times.

Note that, if we take c copies of each k out of v combination, we obtain a (v,b,r,k,λ) BIBD with $b = c\binom{v}{k}$, $r = c\binom{v-1}{k-1}$ and $\lambda = c\binom{v-2}{k-2}$. By setting $k = v$, it is clear that a (v,r,r,v,r) BIBD always exists. The zone-balanced networks constructed in Ref. 3 are thus seen to be special cases of the networks specified in Theorem 4 by setting $\lambda_j = r_j$ for $j = 1,2,\ldots,m - 1$. (The conditions that $\lambda = 1$ and $\Pi_{l=1}^{t} p_l$ divides r_m are not explicitly stated in Ref. 3, but a check with the author of Ref. 3 has verified their necessity.)

V. A GENERALIZATION

We can generalize the definition of zone-balanced network to *partially zone-balanced network* in which every pair of zones is classified into one of the k *associate classes*. The channel graphs of all intrazone pairs of the ith associate are isomorphic to a graph G_i regardless of which pair is chosen. The number of the ith associates of a given zone should be independent of which zone is chosen. Just as balanced incomplete block designs are a natural tool for the construction of zone-balanced networks,

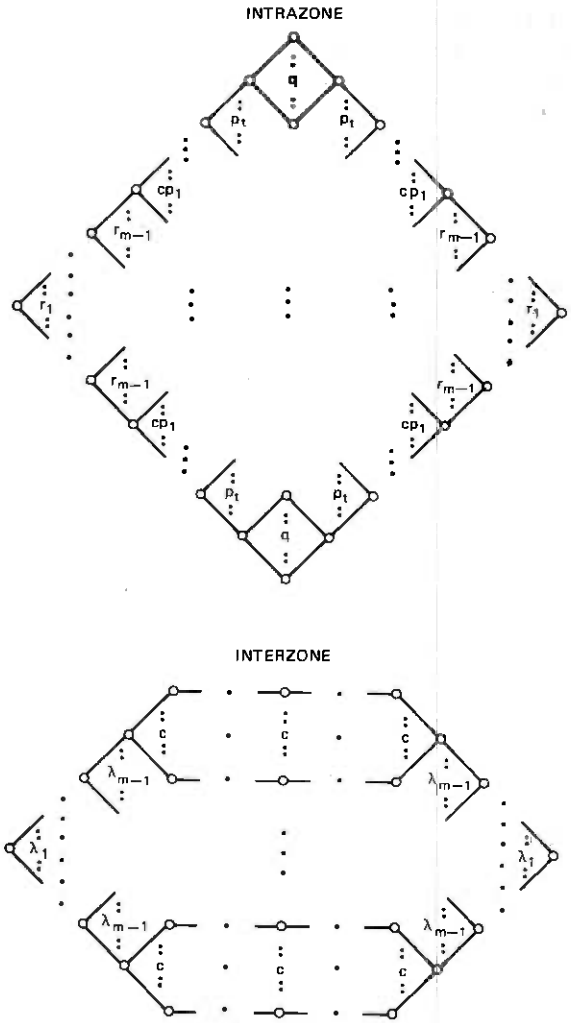


Fig. 6—Channel graphs for Theorem 4.

we can use partially balanced incomplete block designs to construct partially zone-balanced networks, and results similar to those given in this paper can be obtained. However, the partially balanced incomplete block design is really too strong for our construction, since we do not require that for every pair of zones X and Y of the ith associate, the number of zones which are the jth associate of X and the kth associate of Y should be independent of X and Y. This suggests that some design weaker than the partially balanced incomplete block design should be studied for this purpose.

After the completion of this paper, we learned that the author of

Ref. 3 had just revised her paper into a more complete and general account.[6] However, the main difference between our construction and her construction remain as follows: (*i*) Her construction uses only one BIBD, while ours uses many BIBDs sequentially. (*ii*) The method of using balanced partition of blocks is unique in our construction.

REFERENCES

1. M. Hall, Jr., *Combinatorial Theory*, Blaisdell, 1967.
2. F. K. Hwang, "Link Designs and Probability Analysis for a Class of Connecting Networks," to appear in Networks.
3. F. R. K. Chung, "On Switching Networks and Block Designs," Conference Records of the Twentieth Annual Asilomar Conference on Circuits, Systems, and Computers, Pacific Grove, California, 1976, pp. 212–218.
4. K. Takagi, "Optimal Channel Graph of Link System," Electronics and Comm. in Japan, *54-A* (1971), pp. 1–10.
5. G. Timperi and D. Grillo, "Structural Properties of a Class of Link Systems," Alta Frequenza, *XLI* (1972), pp. 279–289.
6. F. R. K. Chung, "Zone-Balanced Networks and Block Designs," B.S.T.J., this issue, pp. 2957–2971.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 8, October 1978
Printed in U.S.A.

The Reliability of 302A Numerics

By A. S. JORDAN, R. H. PEAKER, R. H. SAUL, H. J. BRAUN, and H. H. WADE

(Manuscript received December 9, 1977)

To assess the long-term reliability of Western Electric 7-segment 302A numerics, accelerated forward bias-aging (10 mA, 60° and 125°C) and thermal cycling (−40° to 125°C) experiments have been performed. In treating the bias-aging data, we strictly differentiate between LED chip and digit failure and show that the times to chip failure—defined here as the times required to reach a normalized efficiency of $r = \eta/\eta_0 = 0.5$—are lognormally distributed. The analysis of the data was facilitated by a novel computer-graphics routine which provides for each digit on test a bar-by-bar description of the time-evolution of r. The median life and standard deviation at 125°C and 10 mA for chips are 2400 hours and ~0.4, respectively. Furthermore, we find that the failure distribution for digits can be obtained from the chip distribution by a simple probabilistic consideration. The good accord demonstrated between the experimental data and the theoretical curve derived from the diffusion theory of red GaP LED degradation indicates that the predominant mode of degradation in bias-aging of 302A devices is that of the LED.

The thermal cycling response of 302A numerics encapsulated in Hysol 1700 epoxy is excellent. Similar to bias-aging, the chip failure distribution is lognormal, and chip and digit failures are interrelated by the probabilistic law. For a temperature excursion between −40° and 125°C, the median number of cycles to failure is 23,350 for chips and 300 for digits. The cause of failure is identified by electrical testing as open wire bonds.

Finally, the acquired data permit the estimation of the mean times to failure (MTTF) and failure rates beyond infant mortality of 302A numerics in a specific application such as Transaction telephone sets under realistic operating conditions. The overall reliability of these devices is excellent, characterized by an MTTF of 10^6 hours and a maximum failure rate of less than 1 FIT for bias-aging over a 20-year service life. Failures due to broken wires are estimated to yield an MTTF of 10^9 hours and a failure rate of $\lesssim 4$ FITs at 20 years of service.

I. INTRODUCTION

The 302A and 302C red numeric displays designed* by Bell Laboratories and manufactured by Western Electric—Reading consist of 8 red GaP chips arranged as 7 segments into a single digit and a right-hand decimal point. Each chip is mounted in a separate reflector and connected with either a common cathode (302A) or common anode (302C). In both cases, the device is encapsulated in Hysol 1700 epoxy.†

In general, the effects of long-term thermomechanical, environmental, and electrical operating conditions on device performance are determined concurrently with device development. The major objective of this work was to obtain reliability information on 302A numeric display devices by means of forward bias-aging and thermal cycling at accelerated rates, which enables us to predict their long-term behavior, beyond infant mortality, when used in typical Bell System applications.

First, we provide an outline of the experimental procedures employed in the acquisition of failure data by high temperature forward bias-aging and wide temperature-range thermal cycling. Then, the handling, treatment, and graphic display of the copious amount of information generated by the bias-aging experiments are discussed. Second, the time evolution of the relative luminescent efficiency ($r = \eta_t/\eta_0$) of 302A numerics during bias-aging is compared to the predictions of the diffusion theory of red GaP degradation.[1] Moreover, the failure distribution is established for the entire sample of chips. Next, we determine the failure distribution for thermal cycling and attempt to locate the thermomechanical weak points of a bonded chip. Finally, we discuss the evaluation of the mean time to failure (MTTF) by a variety of criteria (i.e., alternative definitions of chip and digit failure). It is shown on the basis of probabilistic arguments that the failure distribution for digits can be calculated from the chip distribution. Likewise, in the case of thermal cycling, chip and digit distributions are convertible. The parametric values under normal operating conditions can be extrapolated from accelerated failure data by means of semi-empirical correlations. When they are combined with a device utilization model for a specific application, the MTTFs and failure rates induced by the forward bias and temperature cycling can be readily estimated.

II. EXPERIMENTAL

2.1 Bias-aging

Seventeen 302A numeric devices were selected for long-term elevated temperature forward bias-aging. Prior to aging (at $t = 0$), the light output of every segment was measured at 25°C by a standard technique.[2]

* The device was designed by C. R. Paola.
† A product of the Hysol Division, Dexter Corporation.

Briefly, the output was determined segment by segment over a wide range of pulsed test current inputs between 1 and 100 mA (including 2, 5, 10, 20, and 50 mA). At each pulsed current, the duty cycle was adjusted to assure a constant 1 mA dc average current to minimize junction heating. The segment electroluminescence was detected by a PIN 10 diode and its output after amplification was handled by an appropriate software program on an HP 9830A minicomputer to yield a table of information stored on a computer file for all devices under testing.* The table contains the light output of each segment of a given device measured over the entire range of measuring currents employed. The lowest and average initial light outputs were 0.021 and 0.035 millicandella/mA, respectively.

Following initial testing, the 17 devices were split into two groups (9 and 8 units in each group). One of the groups was aged at 60°, the other one at 125°C, in ovens continuously purged with filtered N_2. The devices were placed in trays, each holding three devices, connected to power supplies providing 10 mA dc forward bias. Periodically, all the devices were removed from the ovens to determine the effect of bias-aging on light output. Before performing the measurements by the above-described procedure, the devices were allowed to cool for two hours to 25°C.

2.2 Thermal cycling

Ten 302A numerics were subjected to continuous temperature cycling without bias in a controlled-environment chamber. Each cycle consisted of cooling the devices in the chamber from room temperature (~25°C) to a cold dwelling point at −40°C, then reversing to a warm dwelling point at 125°C, and finally returning to ~25°C. At each temperature extreme, the dwelling time was about 20 minutes. The cooling and heating rates never exceeded 5°C/min.

The devices were mounted in ceramic sockets attached to a combination aluminum and phenolic test fixture through which electrical connections could be made for periodic checks. As a result of progressive thermal cycling, dark segments could be visually observed at the standard forward current of 10 mA dc. All the defective chips were faulty at both temperature extremes as well as at room temperature. For the duration of the first 100 cycles, checks were made at about every 5 cycles; thereafter, the test interval was increased to approximately 25 cycles.

* The automated test facility was developed by J. W. Mann.

III. DATA HANDLING PROCEDURE

Whenever the degraded light output is measured, each device in the sample is characterized by 8 segments × 8 electrical values (including the decimal point and the device voltage at 10 mA). For a total sample of 17 devices, as many as 1088 numbers are acquired per test at all current levels. Obviously, computer storage, retrieval, and treatment of the data together with a suitable graphic display are required for detailed analysis. Therefore, a time-sharing program has been developed which plots the relative electroluminescent efficiency in a unique manner. First, the program LED/EFFCAL reads the permanently stored file of raw data generated by each light output measurement of any t. A short terminal dialog permits the user to name the type of device involved because, in addition to the 302A, the program is also applicable to other classes of numerics as well as to a group of 10 discrete LED chips. Moreover, one can specify the aging temperature and room temperature test current and one of the two plotting scales. Then, a file is written which includes the device identification number, aging temperature, total accumulated aging time, test current, and the number of devices in test under listed conditions. Moreover, an array is created from the raw data which, for every digit of each device,* contains, segment by segment, the relative efficiency as a function of the elapsed aging time.

As shown in Fig. 1, in accord with the accepted convention, the segments are designated by the alphabetic codes $A, B \ldots G$. For the A segment of the ith device, the relative efficiency at t is given by

$$r_A^i = \frac{\eta_A^i(t)}{\eta_A^i(0)}, \tag{1}$$

where the conversion factor relating light output to electroluminescent efficiency η was cancelled. The array includes r_A^i for each time the light output has been measured. In addition, it may be interesting to know the relative degradation of any segment in comparison with the mean value. For this reason, we devised the following statistical indicators:

(i) Device or digit mean, $\bar{r}_i$

$$\bar{r}_i = \frac{\sum_{A=1}^{7} r_A^i}{7}. \tag{2}$$

(ii) Grand mean, $\bar{r}$

$$\bar{r} = \frac{\sum_{i=1}^{n}\sum_{A=1}^{7} r_A^i}{7n} = \frac{\sum_{i=1}^{n} \bar{r}_i}{n}, \tag{3}$$

where n is the number of devices tested under identical conditions.

* Although in the case of a 302A numeric each device corresponds to a digit, allowance is made in the program for numerics which consist of as many as 4 digits per device.

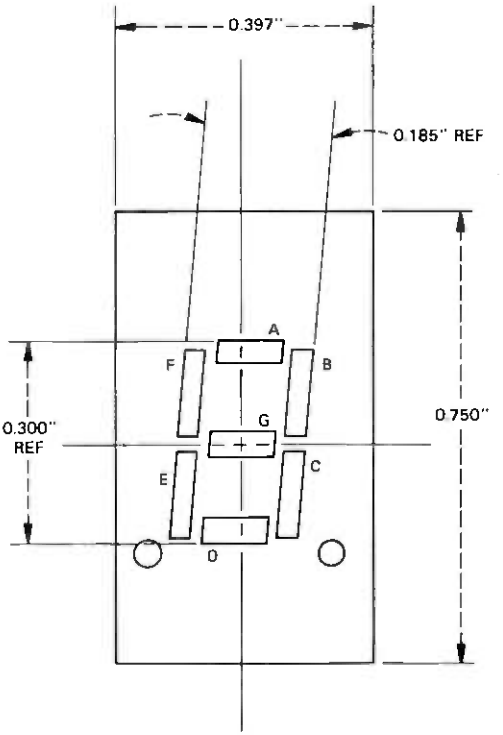


Fig. 1—Top view of a 302A device showing letter designation of segments.

(*iii*) Standard deviation—upper and lower confidence limits of the grand mean at ~68-percent confidence level, $\bar{r}_u$ and $\bar{r}_\ell$

$$\bar{r}_u = \bar{r} + \frac{\left(\sum_{i=1}^{n}\sum_{A=1}^{7}(r_A^i - \bar{r})^2\right)^{1/2}}{7n} \tag{4a}$$

and

$$\bar{r}_\ell = \bar{r} - \frac{\left(\sum_{i=1}^{n}\sum_{A=1}^{7}(r_A^i - \bar{r})^2\right)^{1/2}}{7n}. \tag{4b}$$

At the beginning of the file written for a given set of test conditions (i.e., aging temperature and current) $\bar{r}$, $\bar{r}_u$ and $\bar{r}_\ell$ are listed in a time sequence. At the end of the file, we find $\bar{r}_i$ as a function of time for each digit.

A batch program named LED/EFFPLT accepts these arrays to produce hard-copy plots of r versus time either using the rapid output STARE 3 system or the FR80 microfilm plotter, especially suitable for white prints or viewgraphs. The two options for scales are $\log_{10} r$ versus square-root of time and r versus $\log_{10}$time. In Fig. 2, a computer-generated graphic

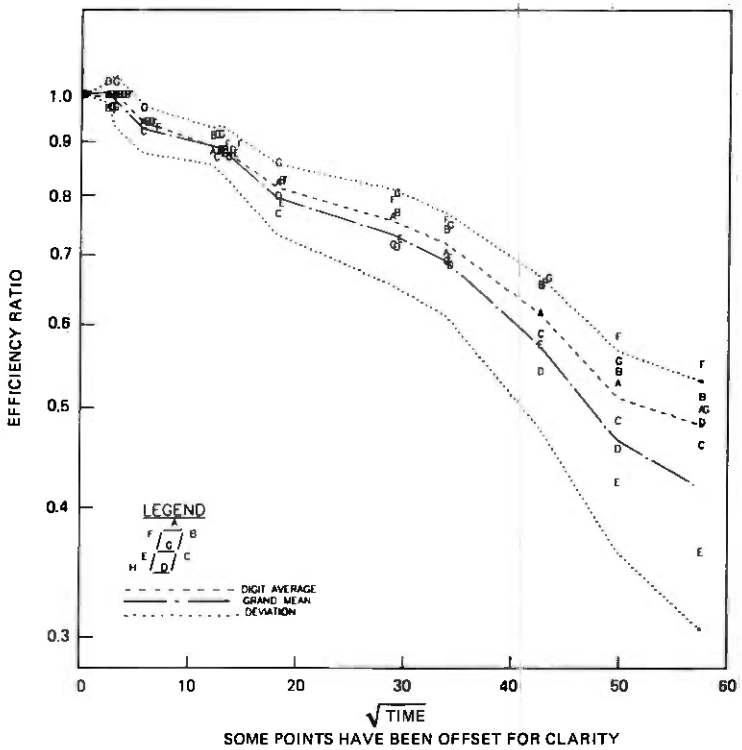


Fig. 2—Computer-generated plot of $\log_{10}$ normalized efficiency ($r = \eta/\eta_0$) versus time $^{1/2}$ [hours$^{1/2}$] for a 302A device aged at 125 °C and 10 mA. The dashed, dash-dot, and dot-dot lines are the digit mean [eq. (2)], grand mean [eq. (3)], and standard deviation of grand mean [eq. (4)], respectively.

output is presented for a 302A device aged at 125°C and 10 mA dc and measured at 10 mA in the $\log_{10} r$ versus $\sqrt{t}$ projection. Note that, for each segment of the digit, the symbol is its alphabetic designation in the order shown in Fig. 1. When overlap of the letters interferes with clarity, a slight horizontal displacement of the symbol along the time axis has been introduced into the plotting routine. In addition to the discrete r values of the individual bars, we also display the device mean, grand mean, and its lower and upper confidence limits by continuous dashed, dash-dot, and dot-dot lines, respectively. In Fig. 3, the data for the same device are presented using the r versus $\log t$ scale option. It can be readily seen that, at the chosen testing intervals, the former scale provides an evenly spread distribution of points at long aging times, while the latter (r vs. $\log t$) achieves well-separated spacings at short times. Up to 3200 hours, the normalized efficiency of the 302A is apparently independent of time when bias-aged at 60°C. This is shown in Fig. 4 on a $\log_{10} r$ versus $\sqrt{t}$ plot. Although the r scale is magnified here compared to Fig. 2, it does not appear to illustrate more than measurement fluctuations. Of course, this

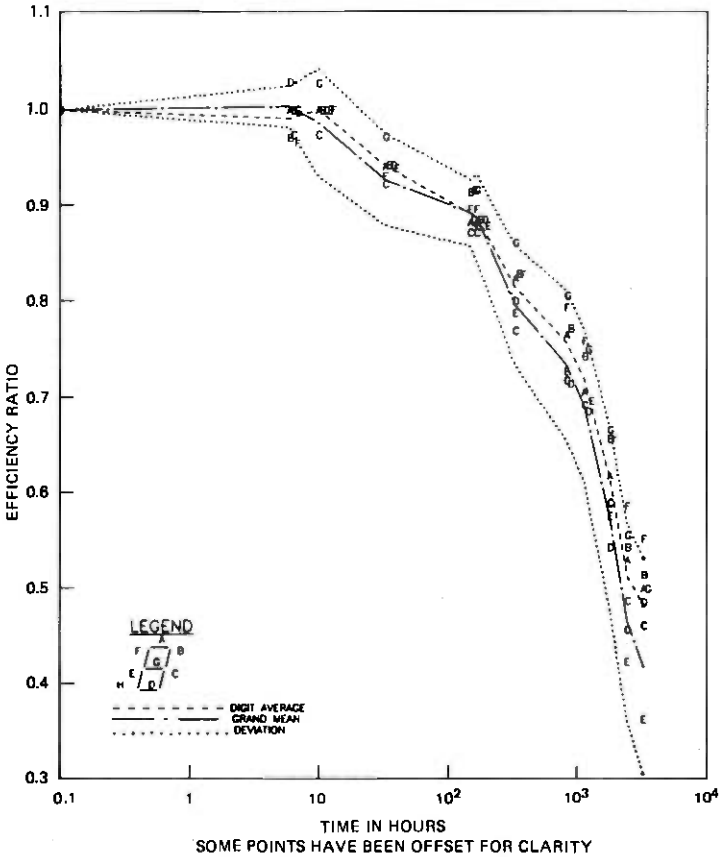


Fig. 3—Computer-generated plot of normalized efficiency ($r = \eta/\eta_0$) versus $\log_{10}$ time for a 302A device aged at 125°C and 10 mA. The dashed, dash-dot, and dot-dot lines are the digit mean [eq. (2)], grand mean [eq. (3)], and standard deviation of grand mean [eq. (4)], respectively.

is consistent with the activation energy of red GaP LED degradation (0.7 eV)[5] which leads to very little change in r at 60°C in the first few thousand hours.

These detailed computer-generated figures are very handy in a rapid assessment of numeric degradation and also as an intermediate step for further analysis. In particular, one can immediately see to what extent an individual chip departs from the digit and grand means. In addition, as shown in the next section, the MTTF of the failure distribution can be readily evaluated from the plots under a variety of failure definitions.

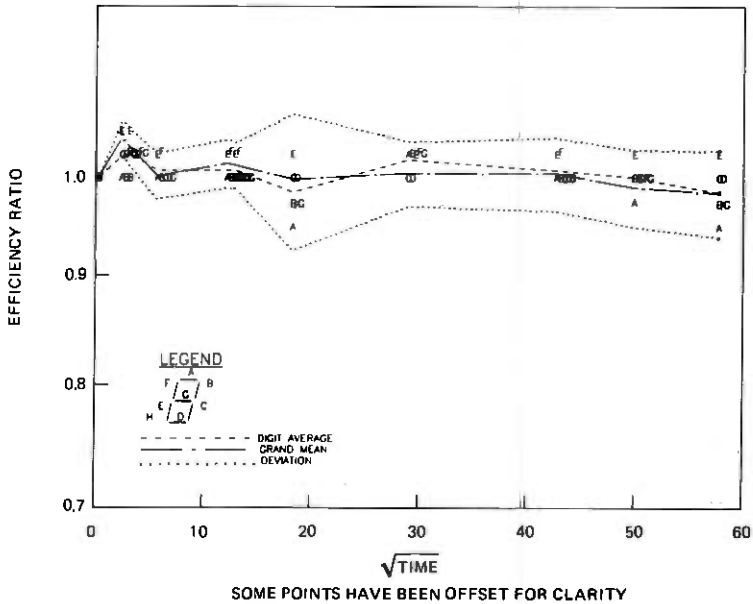


Fig. 4—Computer-generated plot of $\log_{10}$ normalized efficiency ($r = \eta/\eta_0$) versus time$^{1/2}$ [hours]$^{1/2}$ for a 302A device aged at 60°C and 10 mA. The dashed, dash-dot, and dot-dot lines are the digit mean [eq. (2)], grand mean [eq. (3)], and standard deviation of grand mean [eq. (4)], respectively.

IV. RESULTS AND DISCUSSION

4.1 Bias aging

To compare the bias-aging results on 302A numerics with existing information involving discrete red GaP LED chips, we have replotted from Fig. 3 the grand mean of the normalized efficiency and its standard deviation as a function of $\log_{10}t$ in Fig. 5. Superimposed on the same figure is the calculated course of degradation at 125°C ambient temperature and 10 mA stress current (132°C junction temperature). The theoretical curve is based on the diffusion theory of red GaP LED degradation.[1]

Recently, the time evolution of nonradiative centers, thought to be responsible for the long-term degradation of red GaP LEDs, has been modeled. It has been postulated that degradation is due to the diffusion and accumulation of an undesirable impurity or point defect through the depletion layer as the p-n junction potential, which retards defect motion, is reduced by the forward voltage. An explicit analytical expression between r and t was derived which provided a good fit to the degradation data for discrete diodes obtained at various junction temperatures and stress currents. The equation for r is of the form[1]

$$r(t) = \frac{1}{\sqrt{1 + r_0 + \gamma\varphi(t/\theta)}}, \tag{5a}$$

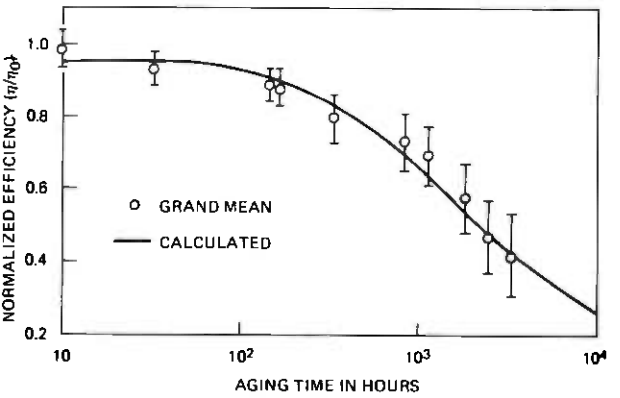


Fig. 5—Normalized efficiency versus $\log_{10}$ time for 302A numerics aged at 125°C and 10 mA. The full line is calculated from the diffusion theory of red LED degradation [eq. (5)]. The data points are the grand mean and its standard deviation for all segments of the sample (see Fig. 3).

where φ is an infinite series, γ is a constant which is proportional (among other quantities) to the initial undesirable impurity concentration and electron lifetime of a diode lot, and r_0 reflects the rapid drop in η at relatively short aging times. The quantity θ is related to the diffusivity (D_0), activation energy (ΔH_a), and stress current (I_{stress}) according to

$$\frac{1}{\theta} = \alpha I_{\text{stress}} e^{-\Delta H_a/kT}, \tag{5b}$$

where α is a known constant of proportionality.

To calculate the degradation curve in Fig. 5 from eqs. (5), we made use of the parameters listed in Ref. 1. However, on account of the lot-dependent properties of γ, a small upward adjustment in its value was required to achieve optimum description of the data. Without this change, the computed r values would be approximately 10 percent above the plotted ones at times in excess of 1000 hours.

The good agreement seen between the experimental normalized efficiencies for 302A numerics and the theoretical curve indicates that these red devices do not exhibit failure modes in addition to LED degradation. This finding is corroborated by the 125°C storage aging of numerics which has not discolored the Hysol 1700 encapsulant. Thus, it appears that the optical coupling efficiency of 302A numerics is invariant during bias aging.

Since the normalized efficiency of individual segments visibly deviates from the grand mean (Fig. 2), the failure of 302A numerics must be distributed by some statistical law. Detailed digit-by-digit failure plots similar to Fig. 2 permit the determination of the failure distribution as a consequence of bias aging. The following possible failure criteria are worthy of exposition in some detail:

(*i*) Chip failure: Whenever any one of the $7n$ chips in the samples reaches $r_A^i = 0.5$, that time is denoted as the time to chip failure (TTCF).

(*ii*) Digit failure: The time for the first chip in each device to attain $r_A^i = 0.5$ is designated as the time to digit failure (TTDF).

(*iii*) Digit mean failure: The time for any digit mean $\bar{r}_i$ [eq. (2)] to equal 0.5 is named the time to digit mean failure (TTDMF);

(*iv*) Grand mean failure: When the grand mean normalized efficiency [eq. (3)] $\bar{r} = 0.5$, we speak of time to grand mean failure (TTGMF).

Each one of these quantities possesses a mean except TTGMF.

The total number of chips in our sample of 8 aged at 125°C is $8 \times 7 = 56$. In Fig. 6, we present on lognormal graph paper the time to chip failure as a function of cumulative failure percent.[3,4] The TTCF values were taken from computer-generated plots for each 302A digit, identical in form with Fig. 2, and then rank-ordered to provide the cumulative failure. It should be noted that TTCFs above ~3400 hours were obtained by linear extrapolation on the $\sqrt{t}$ plots.[5] Hence, the longer the time to chip failure is, the less accurate the TTCF value becomes. Fortunately, this is probably not important except in the case of the last two points.

The linearity of the TTCF plot in Fig. 6 indicates that the failure distribution for 302A numeric chips is lognormal, as is the case for 1A opto-isolators,[6] which utilize GaP LEDs and also for numerous semiconductor devices.[7] Least-square analysis provides the following lognormal parameters for the chip failure distribution.[4]

$$\mu_a = \ln t_{ma} = 7.783$$
$$\sigma_a = 0.37,$$

where the median life, t_{ma}, is 2400 hours. It can be seen that the distribution is very tight, as σ_a is quite small.

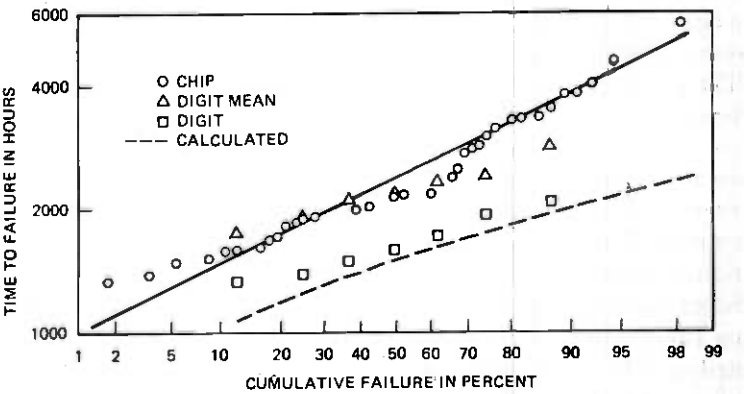


Fig. 6—Time to failure versus cumulative failure percent for 302A numerics aged at 125°C and 10 mA. The symbols O, □, and Δ correspond to failure criteria (*i*), (*ii*), and (*iii*), respectively (see text). The solid line is a least-square line for chip failure, while the dashed line is calculated from eq. (9) for digit failures.

The time to digit mean failure as a function of cumulative failure percent is also shown in Fig. 6. Each one of the computer-generated plots for each device, similar to Fig. 2, yields one value of TTDMF. Because each time value is the result of averaging seven normalized efficiencies, the standard variation is smaller than for TTCF. However, the median life t_{mc} = 2200 hours, which is almost the same as for chip failure.

Perhaps the most important distribution for numerics is the time to digit failure, since it is reasonable to assume that if any one chip in a 7-bar numeric loses half its efficiency, the displayed numerals may not be correctly discriminated by eye. The TTDF distribution is also shown in Fig. 6. Its construction from the computer-generated digit-by-digit graphic outputs is self-evident. We can see that in comparison with a t_{ma} of 2400 hours for chips, the median life for digits, t_{mb}, is reduced to 1600 hours with an accompanying decrease in σ. In Table I, we summarize the median lives obtained by a variety of methods. It can be seen that all the values but the median of TTDF are closely spaced.

However, the chip and digit failure distributions are related by the laws of probability. The reliability function for chip failure R_a is the complement of the plotted cumulative failure Q_a, hence at any time t

$$R_a(t) = 1 - Q_a(t). \tag{6}$$

R_a expresses the probability that the chip will survive to t. If the chip failures in the device are independent, then the reliability of the digit, R_b, assuming device failure if any one of the segments fails ($r_A^i \leq 0.5$), is given by[8]

$$R_b(t) = R_a^7(t). \tag{7}$$

Obviously, the TTDF definition is consistent with eq. (7). The cumulative failure function for device failure, Q_b, is of the form

$$Q_b(t) = 1 - R_b(t) = 1 - R_a^7(t). \tag{8}$$

Finally, a combination of eqs. (6) and (8) yields

$$Q_b(t) = 1 - (1 - Q_a(t))^7. \tag{9}$$

The cumulative device failure function Q_b for 302A numerics can be

Table I — Median lives for a 302A red numeric at 125°C and 10 mA

Method	t_m (hours)
Chip failure	2400
Digit mean failure	2200
Digit failure	1600
Grand mean failure*	2300
Diffusion theory (r = 0.5)*	2200

* Not a median but a single value.

readily obtained from the full line (lognormal distribution) in Fig. 6 and eq. (9). The dashed line in Fig. 6 is the calculated Q_b. Considering the small sample size in terms of digits, the calculated line is a surprisingly good representation of the cumulative failure data for digits. The median life appropriate for Q_b is 1500 hours, which should be compared with the empirical result of 1600 hours. Therefore, it matters very little how the median lives are computed (chip versus device), as long as the results are correctly interpreted. Furthermore, although the median lives obtained by various criteria nearly coalesce, this may not be true of the MTTFs and failure rates as those quantities also involve the standard deviations which, according to the slopes over the data in Fig. 6, are quite variable.

4.2 Thermal cycling

In thermal cycling, chip failure is sudden and manifests itself as a dark segment on testing. In analogy with the definition invoked in the section on bias-aging, digit failure occurs at the number of thermal cycles at which the very first segment fails to light up. In Fig. 7, we present a lognormal projection of the number of cycles to failure versus cumulative failure percent, both in terms of chip as well as device failure, for 302A red numerics encapsulated in Hysol 1700 epoxy.

Again, as in the case of bias-aging, the failure distribution follows the lognormal pattern, as it plots as a straight line on the lognormal graph-paper. Least-square analysis yields the following parameters for chip failure:

$$\mu_{tc} = \ln t_{mtc} = 10 \text{ and } \sigma_{tc} = 3.27,$$

where $t_{mtc} = 23{,}350$ is the median number of cycles to failure (MCTF). The large standard deviation corresponds to a widely spread failure distribution. This is also clear from the steepness of the data and the least-square (solid) line in Fig. 7.

The probabilistic equation derived in the previous section on bias-aging to calculate the digit failure distribution from information on chips is also valid for thermal cycling. Applying eq. (9) to the lognormal line for chips, we can calculate the number of thermal cycles to failure as a function of cumulative failure function for digits. The dashed line in Fig. 7 represents the digit cumulative failure function. The digit MCTF is 300. It is apparent that there is excellent agreement between the theoretical line and the data points for digit failure.

The effect of increasing the number of digits in the same package is also shown in Fig. 7. The dash-dot line is the cumulative failure function for a hypothetical 302A-like device consisting of four digits (28 bars). It is obvious that, with increasing complexity, there is a drastic drop in the MCTF with a somewhat compensating drop in standard deviation.

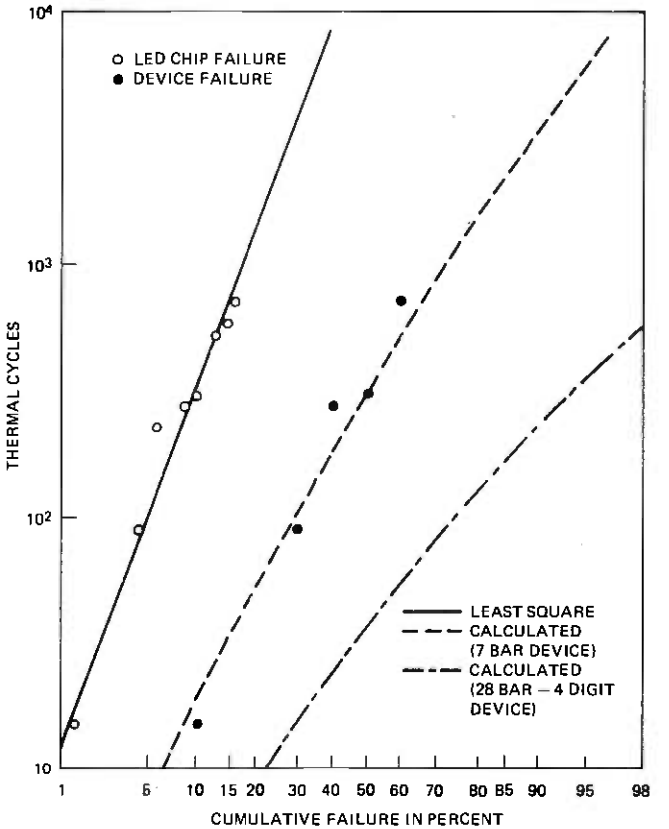


Fig. 7—Number of thermal cycles to failure versus cumulative failure percent for 302A numerics, encapsulated in Hysol 1700 epoxy, thermally cycled between −40° and 125°C. The solid line is a least-square line for chip failure, while the dashed line is calculated from eq. (9) for digit failure.

To locate the source of thermal cycle failure, a number of devices with numerous chip failures were lapped and polished until the gold leads to the diodes were exposed. A detailed view of the lapped area of a 302A numeric is illustrated in Fig. 8. The polishing of the lens continued to the point at which the anode connecting Au wire was broken between the lead frame and the die-bonding pad. Due to the ambiguity of viewing the location of failure, simple electrical checks were performed. Using cathode pins as the common terminal, a fine needle point probe with +2V bias was successively applied to the ends of the severed Au wire and to the bonding pads by piercing the plastic. As a result of such tests, we determined that, after thermal cycling, open circuits develop which are evenly distributed between breaks in the neck of the ball bond and the heel of the wedge bond.

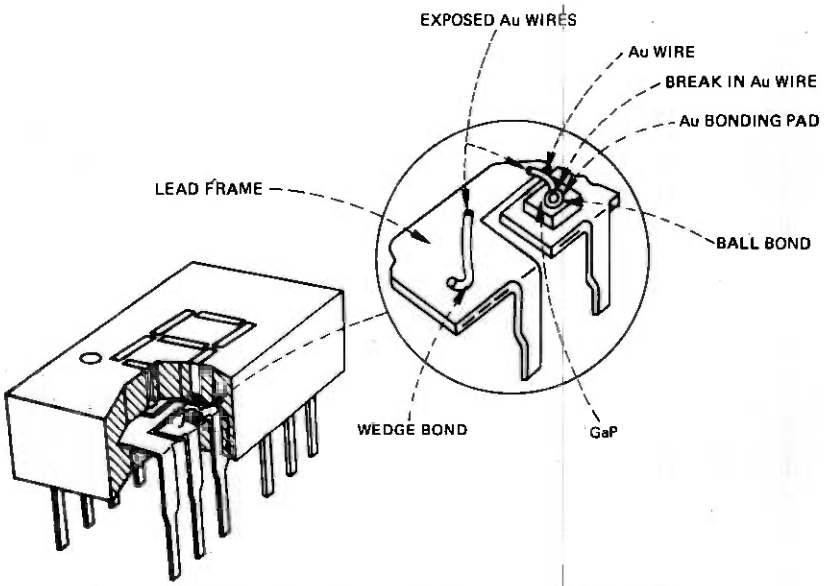


Fig. 8—Cutaway view of a 302A numeric showing the exposed bonds.

4.3 Reliability calculations—device utilization model

It should be very strongly emphasized that the adequacy of a given median life and σ combination based on exhaustive tests can only be judged if the device designer possesses information from systems engineering on its anticipated application. Hence, a device utilization model for the 302A numeric is essential.

One use of the 302A display is in Transaction telephone sets. If we envision very frequent operation, such as in airline terminals, fresh information may appear on the displays as often as every minute for a 30-s duration. The devices operate at 10 mA dc, which leads to a 7°C junction heating. Thus, if the ambient temperature is 27°C, the devices are exposed to 34°C. During the off-state, the time (30 s) is too short to reach thermal equilibrium with the ambient. Therefore, we assume a temperature fluctuation or cycling excursion of no more than ΔT = 2°C.

The model in combination with the reliability data derived herein permit the estimation of the failure rates and MTTFs during device operation under forward bias, and also on account of temperature variations. We shall give the reliability-associated properties for both LED chips and digits at 20 years of service.

(*i*) Bias Aging: The accepted activation energy for red LED degradation is 0.7 eV.[5] With this value, the Arrhenius-law multiplier between the aging temperature (132°C) and use temperature (34°C) becomes 604. Applying the multiplier to the chip and digit median lives, t_{ma} = 2400

and t_{mb} = 1600 hours, respectively, we obtain at the use temperature*

$$t_{ma}(34°\text{C}) = 1.45 \times 10^6 \text{ hours}, \; t_{mb} = 9.66 \times 10^5 \text{ hours}.$$

The standard deviations are usually taken as temperature-independent constants, and their values are

$$\sigma_a = 0.37 \text{ and } \sigma_b = 0.27.$$

It should be noted that the device failure distribution (Q_b) is not strictly lognormal and σ_b results from the linearization of the dashed line in Fig. 6.

The above parameters yield the following MTTFs† and maximum failure rates for a 20-year‡ service life beyond infant mortality:[3,4]

$$\text{MTTF}_a = 1.57 \times 10^6 \text{ hours and } \text{MTTF}_b = 1.01 \times 10^6 \text{ hours}$$

and

$$\lambda_a \ll 1 \text{ FIT and } \lambda_b \ll 1 \text{ FIT}.$$

These failure rates for bias aging are outstandingly low. This is a consequence of the fact that both the chip and device distributions, as shown in Fig. 6, are very tight, corresponding to a small σ. If it is assumed that the investigated lot was atypical and occasionally σ_a and σ_b may become as large as 1 and 0.9, respectively, then we obtain for the λs

$$\lambda_a = 90 \text{ FITs and } \lambda_b = 150 \text{ FITs},$$

indicating, as expected, that the failure rate for chips is less than for digits.

(*ii*) Thermal Cycling: To relate the median number of cycles to failure, MCTF, obtained by long-term wide ΔT excursion experiments to the small ΔT excursions encountered in use, an acceleration factor is required. We have estimated this factor on the basis of previous work on gold beam fatigue. Dais and Howland[9] have shown, for rubber encapsulated devices tested in the plastic deformation domain, that the magnitude of the temperature excursion between ΔT = 400° and 45°C is a monotonically decreasing function of the median number of cycles to failure. We can characterize the dependence of ΔT on cycles by a power law with an exponent increasing from ~ -0.5 (Coffin's Law[10]) to a limiting value of ~ -0.1. In addition, theoretical analysis of typical device structures indicates that the magnitude of the maximum elastic ΔT span for rubber encapsulant is about twice that for epoxy.[11] Thus,

* The subscripts a and b denote chip and device reliability properties, in accordance with the definitions in Section 4.1.

† $\text{MTTF} = t_m e^{\sigma^2/2}$.

‡ The effective service life is only 10 years due to the duty factor of 0.5.

it seems reasonable to take $\Delta T = 20°C$ as the transition temperature between elastic and plastic deformation for the epoxy encapsulant. Consequently, by extrapolating the results of Dais and Howland[9] between $\Delta T = 165°$ and 20°C, we obtain $\sim 3 \times 10^7$ as the approximate acceleration factor appropriate for 302A numerics. Since a temperature cycling range of only 2°C is anticipated in use, the deformation always remains elastic. Hence, the acceleration factor and failure rates given here should be considered as lower and upper bounds, respectively.

A combination of the acceleration factor and the experimental values MCTF_{expa} (23350) and MCTF_{expb} (300 cycles) for chip (a) and digit (b) failure, respectively, yields

$$\text{MCTF}_{usea} = 7 \times 10^{11} \text{ cycles and } \text{MCTF}_{useb} = 9 \times 10^9 \text{ cycles}$$

$$\text{and } \sigma_{tca} = 3.27 \text{ and } \sigma_{tcb} = 2,$$

where σ_{tcb} is from the hypothetical linearization of the dashed line in Fig. 7. Since in operation there are 60 cycles/hour, the median lives become

$$t_{mtca} = 1.2 \times 10^{10} \text{ hours and } t_{mtcb} = 1.5 \times 10^8 \text{ hours.}$$

The above parameters yield the following MTTFs and failure rates, λ_{tc}, at 20 years of service life:[3,4]

$$\text{MTTF}_{tca} = 2.5 \times 10^{12} \text{ hours and } \text{MTTF}_{tcb} = 1.1 \times 10^9 \text{ hours}$$

$$\lambda_{tca} = 2 \text{ FITS and } \lambda_{tcb} = 4 \text{ FITs.}$$

In conclusion, we find that the 302A red numeric performs very reliably in specific applications such as the Transaction telephone. We have shown that the long-term failure rates associated with LED degradation and junction heating induced thermal cycling are very low, namely, no more than 4 FITs over 20 years of continuous service.

V. ACKNOWLEDGMENTS

We are grateful to G. A. Dodson for a critical review of the manuscript. We appreciate the useful practical advice provided by C. R. Paola, B. Johnson, and J. W. Mann.

REFERENCES

1. A. S. Jordan and J. M. Ralston, J. Appl. Phys., *47* (1976), p. 4518.
2. J. M. Ralston, Rev. Scientific Instruments, *43* (1972), p. 876.
3. J. Aitchison and J. A. C. Brown, "The Lognormal Distribution," Cambridge: Cambridge University Press, 1957.
4. A. S. Jordan, Microelectronics and Reliability, *18*, No. 3 (1978), in press.
5. J. M. Ralston, unpublished work.
6. R. H. Saul, E. H. Nicollian, and D. A. Harrison, unpublished work.

7. D. S. Peck and C. H. Zierdt, Jr., Proc. IEEE, *62* (1974), p. 185.
8. "Reliability Engineering," ARINC Research Corporation, W. H. VonAlven, editor, Englewood Cliffs, N.J.: Prentice-Hall, 1965.
9. J. L. Dais and F. L. Howland, unpublished work.
10. S. S. Manson, "Thermal Stress and Low Cycle Fatigue," New York: McGraw-Hill, 1966.
11. J. L. Dais, unpublished work.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 8, October 1978
Printed in U.S.A.

Intelligible Crosstalk Performance of Voice-Frequency Customer Loops

By K. I. PARK

(Manuscript received March 9, 1978)

A methodology for evaluating the intelligible crosstalk performance of voice-frequency customer loops is developed in this paper. Using this methodology, intelligible crosstalk probabilities are calculated for a representative sample of loops in the plant. The effect of gain on loop crosstalk performance is then evaluated for a particular example of gain application where length-dependent gain ranging approximately from −1 to 9 dB is added. Two possible locations of gain application are evaluated: the central office and the telephone set. Presently, no crosstalk performance objectives exist for loops. For planning purposes, however, an intelligible crosstalk probability of 0.1 percent has been used in the past as a limit for satisfactory performance. In comparison with this limit, the crosstalk performance of the present loop plant (loops without gain) is satisfactory. For the particular example of gain application considered in this paper, gain applied at the central office has only a small effect on loop crosstalk performance. However, gain applied at the telephone set degrades loop crosstalk performance significantly, increasing the crosstalk probability above the 0.1 percent level on about 15 percent of the sample loops evaluated.

I. INTRODUCTION

A telephone user occasionally receives an extraneous speech signal as a result of interference between communications circuits, which is referred to as crosstalk. Crosstalk not only produces annoyance to the affected customer but also constitutes loss of another customer's privacy when it is intelligible, and is an important concern in transmission systems design and planning. For example, if, with the advancement of loop electronics, gain devices are applied on loops to enhance the speech signal level, the maximum allowable amount of gain and the location of its application may be restricted by the resulting crosstalk performance degradation. In this paper, a methodology is developed for evaluating

the intelligible crosstalk performance of voice-frequency customer loops that can be used in loop transmission systems design and planning. In particular, the methodology can be used in (*i*) establishing loop crosstalk performance objectives, (*ii*) allocating the objectives to components of the loop plant, such as cable facilities, central office switches, and customer-premises wiring, and (*iii*) evaluating effects of new technology and new loop design rules on crosstalk performance.

Intelligible crosstalk performance is measured by the crosstalk probability, which is defined as the probability that a customer will hear one or more intelligible crosstalk words during a call. The crosstalk probability on customer loops depends on the probability distributions of such random variables as call holding time, quiet interval between calls, disturbing talker volume, crosstalk path loss, circuit noise, and disturbed-listener hearing acuity. These underlying probability distributions in turn depend on telephone connection configurations and crosstalk coupling loss characteristics of the multipair cables used for loops.

The loop crosstalk evaluation methodology developed in this paper can be divided into three basic parts as shown by the block diagram of Fig. 1: a cable crosstalk coupling model, a telephone connection model, and a crosstalk probability model. The cable crosstalk coupling model provides equations for calculating near-end and far-end crosstalk coupling losses between customer loop wire-pairs in multipair cables. The model contains adjustable parameters, which are estimated by fitting the model to measured crosstalk coupling loss data. The telephone connection model describes typical intraoffice (loop-to-loop) telephone

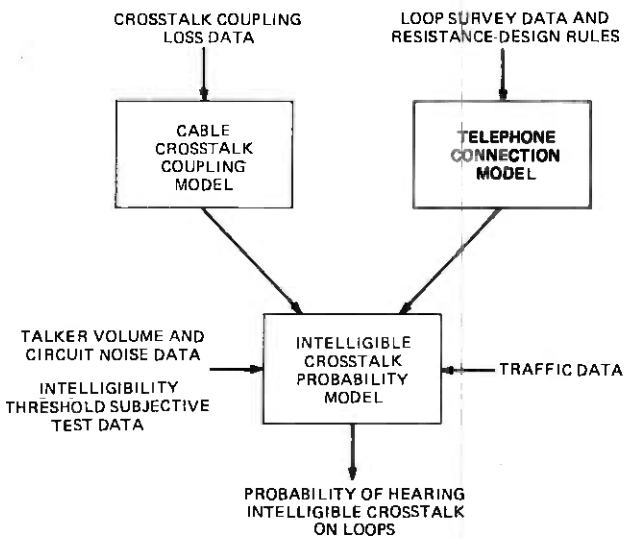


Fig. 1—Modeling of loop crosstalk.

connections as the disturbed connections and identifies potential crosstalk exposures to other intraoffice or toll connections. For the purposes of this study, loop characteristics, such as length, loading, and loss, are described, based on either theoretical loop design rules[1] or the information obtained from the loops sampled in the 1964 Loop Survey.[2] The crosstalk probability model, the last of the three parts shown in Fig. 1, combines the information provided by the preceding two models with data on traffic activity on loops, talker volume, circuit noise, and listener hearing acuity, and determines, by a Monte Carlo simulation, the crosstalk probabilities for loops.

The methodology developed in this paper provides the following features:

(*i*) By virtue of the analytical cable crosstalk coupling model introduced here, the loop crosstalk performance can be evaluated as a function of loop length, rather than only for the fixed length for which measurements are available.

(*ii*) The distribution of crosstalk probability is obtained for all the loops sharing a cable of arbitrary length, or for loops of different lengths sampled from the loop plant.

(*iii*) The telephone connection model developed here is general enough to include a number of different crosstalk exposures, such as near-end and far-end crosstalk occurring in the disturbed customer's loop and near-end and far-end crosstalk occurring in the loop of the customer at the other end of the disturbed connection.

(*iv*) The effect of gain on crosstalk is evaluated for gain applied at the telephone set as well as for gain applied at the central office.

(*v*) For disturbing talkers' speech volumes, the latest speech volume data obtained in 1976[3] is used.

A number of studies on the subject of crosstalk in general were made previously at Bell Laboratories, including those by T. C. Spang, B. E. Davis, M. G. Mugglin, D. H. Morgen,[4] and P. M. Lapsa.[5] Lapsa, in particular, considered a loop crosstalk problem similar to one specific case of the present study—the case of the effect of gain applied at the central office. Focusing primarily on long rural loops with gain applied at the central office and considering near-end crosstalk (NEXT) at the central office as the major crosstalk exposure, he assumed an "electrically long" loop—sufficiently long to render the NEXT coupling loss independent of length—and used measured NEXT coupling loss data. For disturbing talkers' speech volumes, Lapsa used McAdoo's speech volume data obtained in 1960.[6] He concluded that gain of 6 dB or less applied at the central office would be acceptable. In comparison with this, the results of the present paper on the effect of the central office gain are more optimistic because of, among other things, the use of more recent coupling loss and speech volume data in the present study, as discussed in

Section 3.1. In the case of the effect of gain applied at the telephone set, no similar study was made previously that can be compared with the present study.

Section II describes the three basic models constituting the methodology shown in Fig. 1 and determines the probability distributions of the underlying random variables. Section III evaluates the loop crosstalk probability in detail. Section IV is the summary of the loop crosstalk probability evaluation results.

II. METHODOLOGY

2.1 Twisted multipair cable crosstalk coupling model

Crosstalk performance of a customer loop depends on, among other things, the electromagnetic coupling characteristics between the loop and the other loops sharing the same twisted multipair cable. An analytical model was developed to provide equations for the near-end and far-end crosstalk coupling losses between wire-pairs in a cable as a function of frequency, cable length, and terminating impedances. Such a model is necessary because coupling loss measurements are available only for certain frequencies, cable lengths, and terminating conditions. A detailed derivation of the model is described in an unpublished work by the author.[7] In this section, this cable crosstalk coupling model is described in general terms.

A twisted multipair cable consists of a number of twisted wire-pairs stranded together. Each wire-pair is used as a loop, which is permanently assigned to a customer as the transmission path between his telephone set and the serving central office. Although the wire-pairs in a cable are isolated from one another, a certain amount of electromagnetic coupling between simultaneously active pairs is unavoidable.

As illustrated in Fig. 2, crosstalk is referred to as near-end crosstalk (NEXT) when the signal source on the disturbing pair and the point of crosstalk reception on the disturbed pair are at the same end of the cable, and far-end crosstalk (FEXT) when they are at the opposite ends of the cable. The difference in decibels between the disturbing power and the received crosstalk power is referred to as coupling loss. Referring to Fig. 2, NEXT and FEXT coupling losses from pair j into pair i, denoted by NEXT_{ij} and FEXT_{ij}, are defined by the following equations:

$$\text{NEXT}_{ij} = V_{j(\text{disturbing,near-end})} - V_{i(\text{disturbed,near-end})} \quad (1)$$

$$\text{FEXT}_{ij} = V_{j(\text{disturbing,far-end})} - V_{i(\text{disturbed,far-end})}, \quad (2)$$

where $V_{j(\text{disturbing,near-end})}$ and $V_{j(\text{disturbing,far-end})}$ are the disturbing signal powers at the source and the far end on the disturbing pair, pair j, expressed in decibels relative to a reference power; and $V_{i(\text{disturbed,near-end})}$ and $V_{i(\text{disturbed,far-end})}$ are the crosstalk signal powers at the near end and

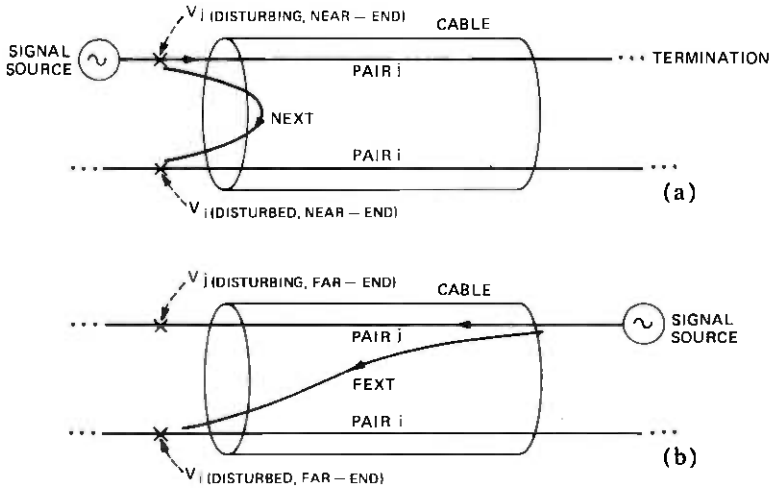


Fig. 2—Definition of NEXT and FEXT: NEXT or FEXT coupling loss from pair j into pair i is the decibel difference between the disturbing volume V_j and the crosstalk volume V_i measured at the points shown by X. (a) Near-end crosstalk (NEXT). (b) Far-end crosstalk (FEXT).

the far end on the disturbed pair, pair i, expressed in decibels relative to a reference power.

Crosstalk performance of a multipair cable can be characterized by determining NEXT and FEXT coupling losses defined by eqs. (1) and (2) for all possible combinations among its wire-pairs. In this paper, the coupling losses are determined analytically by the cable crosstalk coupling model mentioned earlier.[7] The model provides equations for NEXT and FEXT coupling losses as a function of frequency, cable length, and the terminating impedances of the disturbing and disturbed pairs. It contains certain adjustable parameters which are dependent on the proximity between pairs in a cable and which can be determined by fitting the model to measured crosstalk coupling loss data.

The model was fitted to recent crosstalk data measured at Bell Laboratories, Atlanta, on a typical cable used in the loop plant. The data consisted of the NEXT and FEXT coupling losses of 300 pair-to-pair combinations (all possible combinations) in a 25-pair, 26-gauge, nonloaded polyethylene insulated cable (PIC), measured at eight different frequencies (2, 3, 5, 10, 28, 56, 76, and 150 kHz). The length of the measured cable was 3 kft, and all pairs were terminated in pure resistive, 600 ohms at both ends. For each of the 300 pair-to-pair combinations, the model parameters were determined by the least-squares method. Two examples of the results of fitting the model to the data are shown in Fig. 3, where the abscissa is frequency and the ordinate NEXT coupling loss

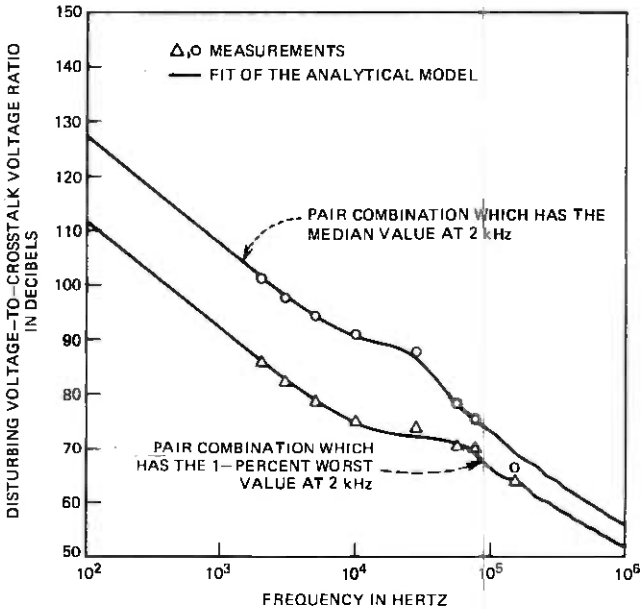


Fig. 3—Two examples of the results of fitting the analytical cable crosstalk coupling model to NEXT coupling loss data measured on a 3-kft, 25-pair, 26-gauge, nonloaded PIC cable with pure resistive 600-Ω terminations.

in decibels. The Δs and Os show the measurements* and the solid curves, the theoretical coupling losses fitted by the model. The rms errors between the measurements and the fitted values for these two particular pair combinations are 0.8 and 1.2 dB, respectively.

Figure 4 presents the cumulative distribution functions (CDFs) of the voice frequency (1 kHz) NEXT and FEXT coupling losses of all the 300 pair-to-pair combinations of the 25-pair cable, calculated by the model for an arbitrarily chosen reference cable length of 1 kft. The far-end and near-end terminating impedances were fixed at (900–j300) ohms and (600 + j200) ohms, respectively, the average terminating impedances of the loops at the central office and at the telephone set, estimated from the 1964 Loop Survey.[2]

Coupling losses vary with cable length. Figures 5 and 6 show length translation factors normalized to 1 kft, as calculated by the model for the voice-frequency NEXT and FEXT coupling losses. Figure 5 shows that, beyond a certain length, in this case about 30 kft, the translation factor for NEXT no longer changes with length. A cable longer than this is referred to as electrically long. From Fig. 5, the NEXT loss at such an electrically long length is about 7 dB smaller than the NEXT loss at 1 kft. Figure 6 shows that FEXT loss keeps decreasing with cable length without saturation.

* Δs and Os in Fig. 3 identify the pair combinations with the 1-percent worst and the median NEXT coupling loss among the 300 measurements at 2 kHz, respectively.

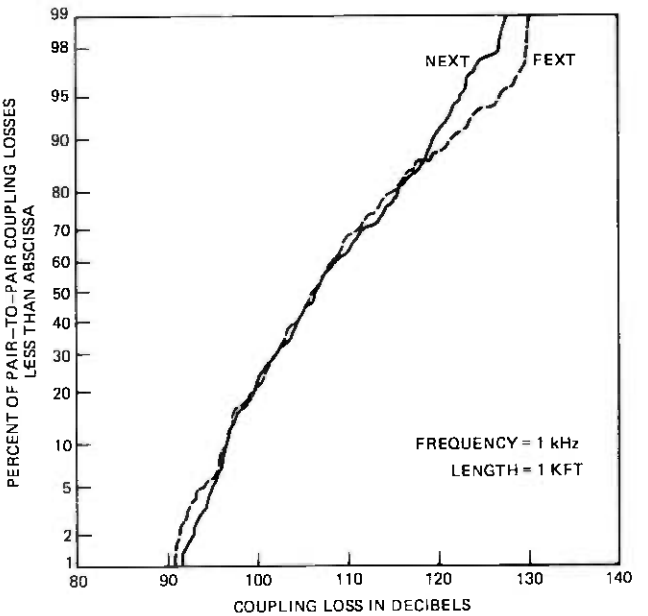


Fig. 4—The cumulative distribution functions (CDFs) of the voice-frequency (1 kHz) NEXT and FEXT coupling losses calculated by the theoretical model with cable length fixed at 1 kft for the 25-pair, 26-gauge, nonloaded PIC cable. The coupling losses at other cable lengths are obtained by using the length translation factors calculated by the model, shown in Figs. 5 and 6.

The NEXT and FEXT coupling losses at lengths other than 1 kft can be obtained by subtracting the corresponding length translation factors determined from Figs. 5 and 6 from the 1-kft coupling losses shown in Fig. 4. For example, the 1-percent worst NEXT coupling loss at 1 kft is, from Fig. 4, about 91 dB and the 1-percent worst NEXT coupling loss at an electrically long length, say 50 kft, is obtained to be 84 dB by subtracting the length translation factor of about 7 dB, determined from Fig. 5, from the 1-kft loss, 91 dB.

The data used to determine the model parameters were measured on an unspliced, laboratory cable. In the plant, several reels of cable may be spliced to form a single long cable. PIC cables are straight spliced; that is, pair identifications on the first reel are maintained over the subsequent reels. This type of splicing has theoretically no effect on the model prediction. For randomly spliced cables, such as pulp cables, the splicing may have some effect because pair locations change over the subsequent reels. At present, there are no appropriate field measurements that can be used to examine the effect of random splicing on crosstalk. However, other things being equal, random splicing should render the crosstalk prediction by the model somewhat conservative (pessimistic) because, with such splicing, the worst crosstalk pair combination of the first reel

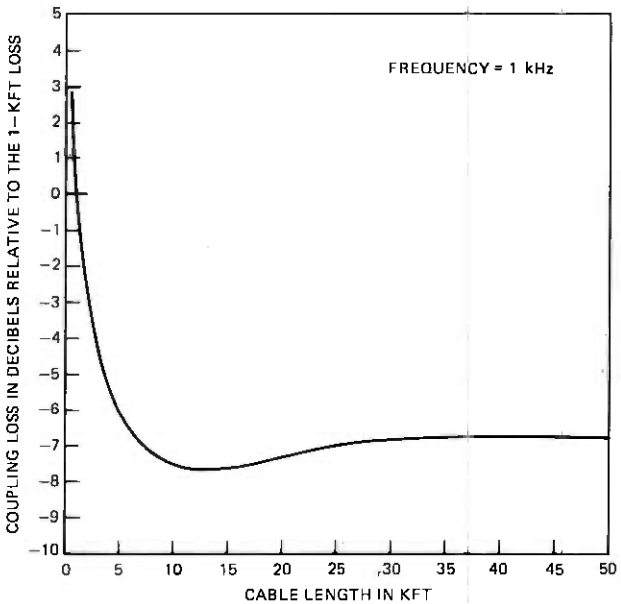


Fig. 5—Length translation factor normalized to 1 kft, calculated by the theoretical model for the voice-frequency NEXT.

would not necessarily be the worst combination in the subsequent reels.

The coupling losses discussed above represent the coupling losses of nonloaded cables, which make up the majority* of the loops in the plant. At the present time, there is no theoretical means of predicting the effect of loading on crosstalk coupling losses. Based on Bell Laboratories coupling loss data measured on loaded cables, it is assumed that, other conditions being equal, loaded cables, which make up a relatively small fraction of the loop plant, have approximately 3 dB smaller NEXT losses than nonloaded cables at 1 kHz. For FEXT, the same FEXT coupling losses are used for both nonloaded and loaded cables.

2.2 Telephone connection model

A model of telephone connections is described in this section to identify potential crosstalk exposures and determine the distributions of received crosstalk volume and other random variables affecting crosstalk performance. On connections involving trunks as well as loops, the crosstalk on trunks is dominant. To evaluate the loop crosstalk taken alone, intraoffice connections, consisting of two loops connected at the central office, are considered the disturbed connections. Intraoffice

* The 1964 Loop Survey (Ref. 2) shows that 84 percent of the loops sampled in the survey are nonloaded loops.

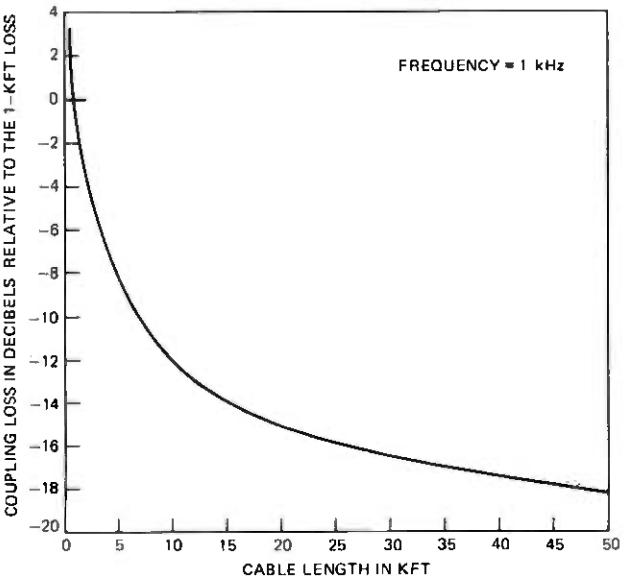


Fig. 6—Length translation factor normalized to 1 kft, calculated by the theoretical model for the voice-frequency FEXT.

connections have relatively low circuit noise, providing low masking on crosstalk intelligibility, and thus are in general most susceptible to intelligible crosstalk. As the disturbing connections, both toll and intraoffice connections are considered.

As shown in Fig. 7, a consumer at one end of an intraoffice connection is subject to the following four potential crosstalk exposures: NEXT and FEXT occurring in his own loop, and NEXT and FEXT occurring in the loop of the customer at the other end of the disturbed connection. Comparing Fig. 2 with Fig. 7, the cable end where the coupling losses are defined corresponds to the telephone set line-terminals for the first two crosstalk exposures and the central office loop terminations for the latter two exposures. For convenience, therefore, the first two exposures will be referred to in this paper as "line terminal NEXT" (LTNEXT) and "line terminal FEXT" (LTFEXT) and the latter two as "central office NEXT" (CONEXT) and "central office FEXT" (COFEXT). Of these four crosstalk exposures, LTNEXT is, in general, most important because, with this exposure, the disturbing talker's volume is attenuated only by the coupling loss between the two loops involved, and there are no additional losses in the crosstalk path. In the other three exposures (LTFEXT, CONEXT, and COFEXT), the disturbing talker's volume is attenuated by loop losses in addition to coupling losses, and thus the crosstalk from such an exposure is less likely to be intelligible than LTNEXT. On the other hand, if gain is applied on loops in the future, the relative importance of the four crosstalk exposures may change depending on the lo-

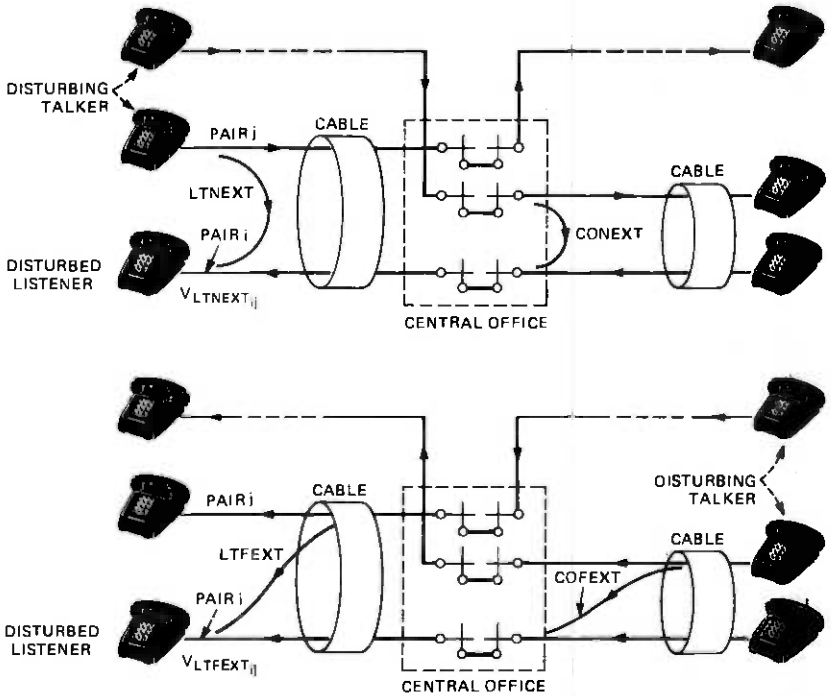


Fig. 7—Four types of potential crosstalk exposures of the intraoffice (loop-to-loop) connection: line-terminal NEXT (LTNEXT), line-terminal FEXT (LTFEXT), central office NEXT (CONEXT), and central office FEXT (COFEXT).

cation of the gain application. The effect of gain on crosstalk will be discussed in Section 2.4.

The crosstalk level in VU (volume units) received at the line-terminals of a disturbed customer's telephone set is the speech level (in VU) at the disturbing talker's telephone set minus the loss (in decibels) of the crosstalk path from the disturbing talker to the disturbed listener. The loss of the crosstalk path includes the loss from the disturbing talker to the point of crosstalk coupling, the coupling loss and the loss from the point of crosstalk coupling to the disturbed customer's telephone set. The crosstalk level on pair i received from pair j for the four crosstalk exposures, denoted by $V_{\text{LTNEXT}_{ij}}$, $V_{\text{LTFEXT}_{ij}}$, $V_{\text{CONEXT}_{ij}}$ and $V_{\text{COFEXT}_{ij}}$, are given by the following equations:

$$V_{\text{LTNEXT}_{ij}} = V_{\text{LT}_j} - \text{LTNEXT}_{ij} \tag{3}$$

$$V_{\text{LTFEXT}_{ij}} = V_{\text{CO}_j} - L_2 - \text{LTFEXT}_{ij} \tag{4}$$

$$V_{\text{CONEXT}_{ij}} = V_{\text{CO}_j} - \text{CONEXT}_{ij} - L_2 \tag{5}$$

$$V_{\text{COFEXT}_{ij}} = V_{\text{LT}_j} - L_1 - \text{COFEXT}_{ij} - L_2, \tag{6}$$

where V_{LT_j} and V_{CO_j} denote the disturbing talker volume at the line

terminals and the central office, LTNEXT_{ij}, LTFEXT_{ij}, CONEXT_{ij} and COFEXT_{ij} denote NEXT and FEXT coupling losses at the line terminals and the central office [as defined by eqs. (1) and (2)], and L_1 and L_2 denote the losses of the loops in the two cables involved in the loop-to-loop disturbed connection. Since talker volume may be assumed to have a same distribution on all pairs in a given cable, the subscript j may be dropped from the disturbing talker volume in the above equations.

The electrical talker volume as measured at the serving central office was determined by a recent survey undertaken by Bell Laboratories to be nearly normally distributed with a mean of −22.2 VU (volume unit) and a standard deviation of 4.6 dB for intraoffice calls and a mean of −21.6 VU and a standard deviation of 4.5 dB for toll calls.[3] These latest speech volume data are used in this paper. These data show that there is very little difference in talker volume statistics between intraoffice and toll calls in contrast to the 1960 McAdoo speech volume data,[6] which showed a mean of −24.8 VU with a standard deviation of 7.3 dB for intraoffice calls and a mean of −16.8 VU with a standard deviation of 6.4 dB for toll calls. The standard deviation of the new speech data is considerably smaller than that of the McAdoo data.

The crosstalk volume equations for LTNEXT and COFEXT, eqs. (3) and (6), involve the electrical volume at the telephone set line terminals of the disturbing talker, V_{LT}. Presently, talker volume statistics at the telephone set line terminals are not available. To obtain the line-terminal talker volume statistics, as a function of loop length, from the central office statistics, the following expressions apply:

$$m_{V_{\text{LT}}}(x) = \{m_{V_{\text{CO}}} + m_{E_2}\} - E_1(x) \tag{7}$$

$$s_{V_{\text{LT}}} = (s^2_{V_{\text{CO}}} - s^2_{E_2})^{1/2}, \tag{8}$$

where $m_{V_{\text{LT}}}(x)$ and $s_{V_{\text{LT}}}$ denote the mean and standard deviation of the talker electrical volume at the telephone set line-terminals, $m_{V_{\text{CO}}}$ and $s_{V_{\text{CO}}}$ the mean and standard deviation of the talker volume measured at the central office, $E_1(x)$ the acoustic-to-electric transducer power loss,* as a function of loop length x, between the input acoustic pressure applied at the telephone set transmitter and the output voltage produced at the telephone set line terminals, and m_{E_2} and s_{E_2} the mean and standard deviation of the acoustic-to-electric transducer power loss between the acoustic pressure at the telephone set transmitter and the output voltage at the loop termination at the central office.

In (7), the term in the braces translates the mean electrical talker volume at the central office, $m_{V_{\text{CO}}}$, into the mean acoustic pressure at the transmitter by adding the mean acoustic-to-electric power loss, m_{E_2},

* These transducer power losses are similar to, but different from, the EARS (Electro-Acoustic Rating System) losses discussed in Section 2.4.1: these power losses are frequency-weighted in a different manner than the EARS losses.

averaged over a representative population of loops of various lengths. This translation assumes that the talker acoustic pressure at the transmitter is not correlated with loop length. The subtraction of $E_1(x)$, the acoustic-to-electric power loss at a given loop length x, from the term in the braces translates the mean acoustic pressure into the mean electrical speech volume at the line terminals for that specific loop length x. Figure 8 shows the mean electrical speech volume at the telephone set line terminals as a function of loop length, obtained by eq. (7) from the mean central office talker volume of −22.2 VU of intraoffice calls presented in Ref. 3. From (8), the standard deviation of the line-terminal talker volume is determined to be 3.9 dB.

Circuit noise received at the end of the intraoffice (loop-to-loop) connection is the power sum of three independent noises: (*i*) the far-end talker's carbon transmitter noise (N_1), attenuated by the losses of the two loops of the connection, (*ii*) the noise of the far-end talker's loop including the noise contributed by the central office (N_2), attenuated by the loss of the near-end loop, the disturbed listener's loop, and (*iii*) the noise of the near-end loop (N_3):

$$N = (N_1 - L_1 - L_2) \oplus (N_2 - L_2) \oplus N_3, \tag{9}$$

where L_1 and L_2 denote the losses of the two loops in the connection and

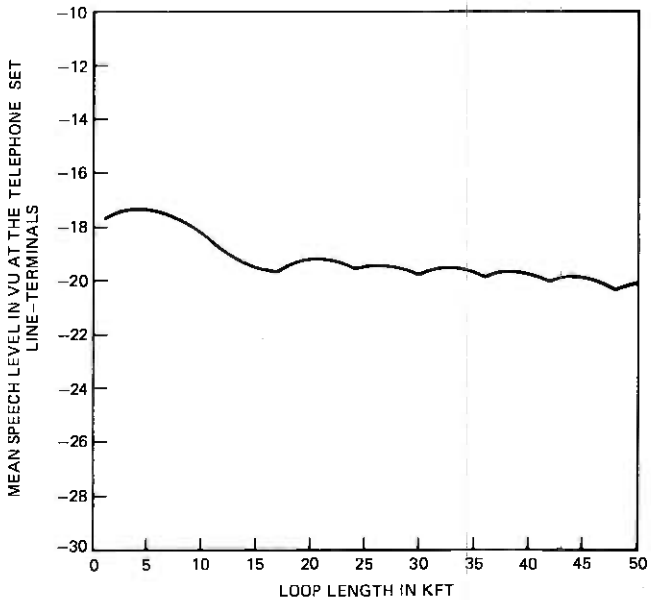


Fig. 8—Mean electrical talker volume at the telephone set line terminals as a function of loop length, obtained from eq. (7) with the central office mean talker volume, −22.2 VU, of intraoffice calls.

⊕ represents the power sum operator.* The far-end talker's carbon transmitter noise is assumed to have a constant value of 10.2 dBrnC.† The 1964 Loop Survey[2] shows that loop noise has little correlation with loop length. Based on the 1964 Loop Survey data, loop noise is assumed to be normally distributed with a mean of −1.1 dBrnC without the central office noise and a mean of 5.6 dBrnC with the central office noise. The standard deviation of loop noise is assumed to be 12.5 dB, both with and without the central office noise. By a Monte Carlo evaluation of eq. (9) with the above component noise statistics, the mean and the standard deviations of the total received noise of the intraoffice connection were determined as a function of the disturbed listener's loop length. For example, for a loop-to-loop connection, with the length of both loops fixed at 7 kft, the mean and standard deviation of the received noise are determined to be 10.5 dBrnC and 8.5 dB, respectively.

2.3 Crosstalk probability model

The discussions hitherto have been concerned with the determination of crosstalk coupling losses, received crosstalk levels for potential crosstalk exposures, and received circuit noise. Whether or not a customer will actually receive intelligible crosstalk, however, is a random event. A mathematical model is developed in this section to evaluate the probability of hearing intelligible crosstalk on loops.

For a customer to receive intelligible crosstalk, the following two conditions must be met simultaneously. First, a potential disturbing circuit must become active during the period when the customer under consideration is engaged in a telephone conversation. Given that the first condition has been met, exposing the customer to crosstalk, the second condition is that the received crosstalk level must exceed the disturbed customer's intelligibility threshold in the presence of circuit noise. The probability that a customer on loop pair i will receive intelligible crosstalk from another loop pair, pair j, in the same cable, denoted by P_{ij}, is expressed by the following equation:

$$P_{ij} = \Pr\{\text{pair } j \text{ active/pair } i \text{ active}\} \times \Pr\{V_{ij} > T(N)\}, \qquad (10)$$

where V_{ij} denotes the crosstalk volume on pair i received from pair j and $T(N)$ denotes intelligibility threshold in the presence of circuit noise N.

The probability of activity coincidence between loops, the first probability in the right-hand side of (10), depends on the distributions of call holding time and quiet interval between calls on loops. This probability was determined in Ref. 5 for average busy-hour loop traffic to be

$$\Pr\{\text{pair } j \text{ active/pair } i \text{ active}\} = 0.17. \qquad (11)$$

---

* $A \oplus B = 10 \log_{10}(10^{A/10} + 10^{B/10})$.

† L. M. Padula, Bell Laboratories, private communication.

The probability of crosstalk intelligibility, the second probability in the right-hand side of (10), depends on the distributions of crosstalk volume, circuit noise, and listener intelligibility threshold. The received crosstalk volume and circuit noise are determined by (3) through (6) and (9), with the distributions discussed in Section 2.2. Listener intelligibility threshold, which is determined by subjective tests, is defined quantitatively as the speech level at which a subject is just able to understand one or more words of the crosstalk content presented to him in the presence of masking noise.[8]

Intelligibility threshold increases as a function of noise. When noise is relatively high, the increase in intelligibility threshold with noise is linear, that is, decibel for decibel. At low noise levels, the relationship between intelligibility threshold and noise is nonlinear: in this region of noise, as noise is decreased toward an infinitely small value, intelligibility threshold approaches a constant rather than continuously decreasing, indicating a human ear's absolute threshold independent of noise. This functional relationship between intelligibility threshold and noise can be expressed by the following equation in terms of a random variable independent of noise, T_0, and a term varying nonlinearly with noise:

$$T(N) = T_0 + (N \oplus 12.3) \text{ VU}, \tag{12}$$

where $\oplus$ represents the power sum operator defined previously. The above equation is a mathematical expression of the intelligibility threshold data presented by T. K. Sen.[8] Sen's data show that T_0 is normally distributed with a mean of -95 VU and a standard deviation of 2.5 dB for a crosstalk coupling mechanism with a flat frequency spectrum. Sen also observed that the mean of T_0 should be lowered by 2 dB to -97 VU for a crosstalk coupling mechanism with coupling losses that roll off with frequency by 6 dB per octave. Since, as can be seen in Fig. 3, crosstalk coupling losses over the voice band have a 6-dB per octave roll-off, T_0 is assumed to have a mean of -97 VU and a standard deviation of 2.5 dB.

Substituting (12) into (10), we have the following expression for the probability of crosstalk intelligibility:

$$\Pr\{V_{ij} > T(N)\} = \Pr\{V_{ij} - T_0 - (N \oplus 12.3) > 0\}. \tag{13}$$

Because of the power sum, $(N \oplus 12.3)$, analytical evaluation of the above equation is not possible even for normally distributed random variables. A simple but crude way of treating the term $(N \oplus 12.3)$ would be to approximate it with a normal variate. However, such an approximation will result in pessimistic results because the normality assumption allows an infinitely low value for the term when, in fact, the random term $(N \oplus 12.3)$ can never be smaller than 12.3. In this paper, therefore, the above probability is evaluated by a Monte Carlo method.

To apply a Monte Carlo method, the above equation is manipulated in the following manner. For a fixed value of noise, say $N = n_k$, and assuming normal distributions for other random variables, it can be shown that the crosstalk intelligibility probability is given in terms of the standardized normal cumulative distribution function Φ as follows:

$$\Pr\{V_{ij} > T(N)/N = n_k\} = \Pr\{V_{ij} - T_0 - (n_k \oplus 12.3) > 0\} = \Phi\left[\frac{\{m_{V_{ij}} - m_{T_0} - (n_k \oplus 12.3)\}}{(s^2_{V_{ij}} + s^2_{T_0})^{1/2}}\right], \quad (14)$$

where

$$\Phi(a) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2}\, dx.$$

The Monte Carlo evaluation procedure of (13) then consists of generating a sequence of random numbers according to the distribution of noise N and evaluating the following average:*

$$\Pr\{V_{ij} > T(N)\} = \frac{1}{M} \sum_{k=1}^{k=M} \Phi\left[\frac{\{m_{V_{ij}} - m_{T_0} - (n_k \oplus 12.3)\}}{(s^2_{V_{ij}} + s^2_{T_0})^{1/2}}\right], \quad (15)$$

where M is the number of random samples drawn for noise N, $m_{V_{ij}}$ and $s_{V_{ij}}$ are the mean and standard deviation of the received crosstalk volume determined for a given crosstalk path using (3) through (6), and m_{T_0} and s_{T_0} are the mean and standard deviation of the random variable T_0 of (12).

Using the last equation and the activity coincidence probability of (11) in (10), the probability that a customer on pair i will receive intelligible crosstalk from pair j, P_{ij}, is given by

$$P_{ij} = 0.17 \times \text{eq. (15)}. \quad (16)$$

Finally, the crosstalk probability that the customer on pair i will receive intelligible crosstalk from any of the remaining N-1 pairs in the N-pair cable, denoted by P_i, is given by, assuming small P_{ij}:

$$P_i = \sum_{j=1}^{j=N} P_{ij},\ i = 1,2,\ldots,N,\ j \neq i. \quad (17)$$

2.4 *Effect of gain*

The received crosstalk volume and circuit noise equations, (3) through (6) and (9), assume no gain devices on loops, as is the case in the current loop plant. With the advancement of loop electronics, the present loop

* Lapsa (Ref. 5) evaluated a similar probability by numerical evaluation of convolution integrals.

design rules may change in the future and require application of gain on loops. The effect of gain on crosstalk performance is discussed in this section for a particular example of gain application where the required gain is determined as a function of loop length to meet a certain constant loop loss.

2.4.1 Loudness loss of telephone connections

Voice communications over a telephone connection are accomplished by conversion of a talker's acoustic pressure at the transmitter into an electrical signal, transmission of the electrical signal over a transmission medium to the receiving telephone set at the far end and reconversion of the received electrical signal into an acoustic pressure at the listener's receiver. The loudness of the speech perceived by the listener depends on the magnitude of the talker's acoustic pressure and the loss and frequency characteristics of the transmitter, the receiver and the transmission medium. The loudness loss between the input and output acoustic pressure of a connection is quantified by means of the Electro-Acoustic Rating System (EARS), and is referred to as the EARS loss of the connection. For a complete and extensive discussion on the subject of EARS, the reader may refer to Ref. 9.

For interoffice or toll connections, the transmission path consists of one or more trunks in tandem between the two end offices, which are in general derived on carrier facilities, plus a loop at each end. The EARS loss of such a connection is given by the sum of the transmit loop rating (TLR) of the talker's loop (transmit loop) and the receive loop rating (RLR) of the listener's loop (receive loop), plus the electrical loss of the intervening trunks. For intraoffice connections, the transmission path consists of two loops connected together at the central office. The EARS loss of such a loop-to-loop connection is approximately the sum of the TLR and the RLR of the two loops.

The TLR is defined in terms of an acoustic pressure spectrum specified by the EARS methods at the transmitter of a telephone set and the resulting EARS frequency-weighted, electrical voltage (EARS voltage) produced at the transmit loop termination at its central office. The RLR is defined in terms of an EARS voltage applied at the central office termination of a receive loop and the resulting acoustic pressure produced at the telephone set receiver at the other end of the receive loop. The TLR and RLR have a unit analogous to decibels and are loss-like quantities in the sense that an algebraically larger TLR and RLR respectively correspond to a lower output EARS voltage at the central office and a lower output acoustic pressure at the receive telephone set.

Under the present loop design rules, both TLR and RLR vary with loop length, and consequently the EARS loss over the local portion* of a connection varies with the lengths of the two loops. A recent study[10] examined the possibility of providing a constant EARS loss for the local portions of all connections, regardless of loop length. Such a loss plan would permit EARS loss equalization of intraoffice (loop-to-loop) connections for all loop lengths, but would require changing loop design rules to allow for incorporation of gain. Since application of gain would raise the crosstalk level, the maximum amount of allowable gain may be limited by the consequent crosstalk performance degradation.

The amount of gain required for loop EARS loss equalization depends primarily on three factors: (*i*) the constant EARS loss objective for local portions, (*ii*) allocation of the EARS loss objective to the TLR and RLR, and (*iii*) the present values of TLR and RLR, which are determined largely by the length of the loop. In this paper, the required gain is determined as a function of loop length to meet a constant TLR of −21 dB and RLR of 27 dB, regardless of loop length, which amount to a constant EARS loss of 6 dB for intraoffice (loop-to-loop) connections for all loop lengths. This constant EARS loss of 6 dB, allocated as −21 dB to TLR and 27 dB to RLR, was examined as a possible alternative in the recent study mentioned previously[10] to evaluate long-term loss plans for the loop plant.

Presently, loops are designed according to the resistance-design rules[1] that control the electrical losses of loops by limiting loop resistance and requiring load coils when the length exceeds 18 kft. The resistance-design rules are applied with respect to the longest loop among the loops sharing a same cable, and thus the rest of the loops in the same cable would exhibit less loss. The longest loop, or the maximum-loss loop, in a cable assumed to conform to the resistance-design rules will be referred to as a theoretical resistance-design loop.

The TLR and RLR are shown in Figs. 9 and 10 as a function of loop length for a theoretical resistance-design loop.† The constant TLR and RLR are indicated by dashed horizontal lines. The amount of gain required to meet the constant TLR or RLR is then given by the difference between the horizontal line and the length-dependent curve. The required gain is shown in Fig. 11 as a function of loop length. The required transmit and receive loop gains range approximately from −3 to 9 dB and from −1 to 4 dB, respectively. At short loop lengths, the required gain is negative, indicating that a loss, rather than a gain, is required for the loop loss equalization.

* In this study, the local portion of a connection refers to that part of the connection which comprises the loop plus the telephone set at each end of the connection.

† The TLR and RLR shown in these figures were calculated with a computer program developed by F. B. Stallman, Bell Laboratories.

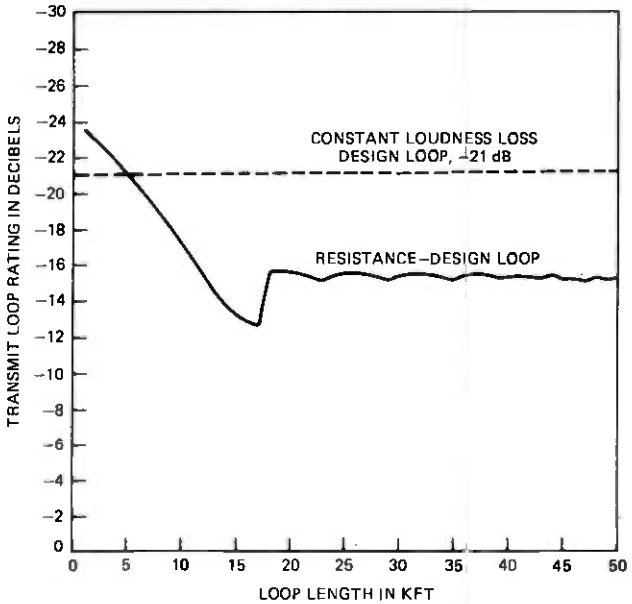


Fig. 9—Transmit loop rating (TLR) of the theoretical resistance-design loop and a constant loudness loss design loop.

2.4.2 Effect of gain on crosstalk volume and noise

For a given amount of gain, the effect on crosstalk performance depends on the location of its application. In this paper, we consider two possible locations: the telephone set and the central office.

Referring to Fig. 12, which is the same as Fig. 7 except for the gain, the crosstalk volume equations, (3) through (6), are modified for gain applied at the telephone set as shown below:

$$V_{\mathrm{LTNEXT}_{ij}} = V_{\mathrm{LT}} + G_{T_2} - \mathrm{LTNEXT}_{ij} + G_{R_2} \tag{18}$$

$$V_{\mathrm{LTFEXT}_{ij}} = V_{\mathrm{CO}} - L_2 - \mathrm{LTFEXT}_{ij} + G_{R_2} \tag{19}$$

$$V_{\mathrm{CONEXT}_{ij}} = V_{\mathrm{CO}} - L_2 - \mathrm{CONEXT}_{ij} + G_{R_2} \tag{20}$$

$$V_{\mathrm{COFEXT}_{ij}} = V_{\mathrm{LT}} + G_{T_1} - L_1 - \mathrm{COFEXT}_{ij} - L_2 + G_{R_2}. \tag{21}$$

The received noise equation, (9), is modified as follows:

$$N = (N_1 + G_{T_1} - L_1 - L_2 + G_{R_2}) \oplus (N_2 - L_2 + G_{R_2}) \oplus (N_3 + G_{R_2}). \tag{22}$$

Referring to Fig. 13, gain applied at the central office will not affect LTNEXT but will affect LTFEXT, CONEXT, and COFEXT. Since the latter three types of crosstalk exposures are in general less significant than the first, the effect of the gain is less pronounced when applied at the central office than at the telephone set. However, depending on loop length, the

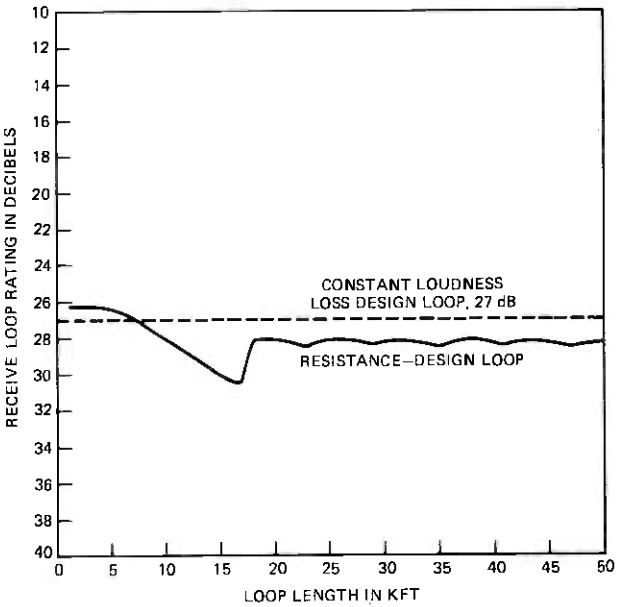


Fig. 10—Receive loop rating (RLR) of the theoretical resistance-design loop and a constant loudness loss design loop.

amount of required gain might be sufficiently large to make these crosstalk exposures significant. The following equations give crosstalk volumes and circuit noise when gain is applied at the central office:

$$V_{\mathrm{LTNEXT}_{ij}} = \text{same as eq. (3)} \tag{23}$$

$$V_{\mathrm{LTFEXT}_{ij}} = V_{\mathrm{CO}} + G_{R_2} - L_2 - \mathrm{LTFEXT}_{ij} \tag{24}$$

$$V_{\mathrm{CONEXT}_{ij}} = V_{\mathrm{CO}} + G_{R_1} - \mathrm{CONEXT}_{ij} + G_{T_1} + G_{R_2} - L_2 \tag{25}$$

$$V_{\mathrm{COFEXT}_{ij}} = V_{\mathrm{LT}} - L_1 - \mathrm{COFEXT}_{ij} + G_{T_1} + G_{R_2} - L_2 \tag{26}$$

$$N = (N_1 - L_1 + G_{T_1} + G_{R_2} - L_2) \oplus (N_2 + G_{T_1} + G_{R_2} - L_2) \oplus N_3. \tag{27}$$

III. RESULTS

The loop crosstalk probabilities were determined first for theoretical resistance-design loops and then for the 1100 loops sampled in the 1964 Loop Survey.[2] In each case, the crosstalk probabilities were determined both without gain and with gain. In the case of loops with gain, two possible locations of gain application were evaluated: the telephone set and the central office. Sections 3.1 and 3.2 present the crosstalk probabilities determined for the theoretical resistance-design loops and the actual loops, respectively.

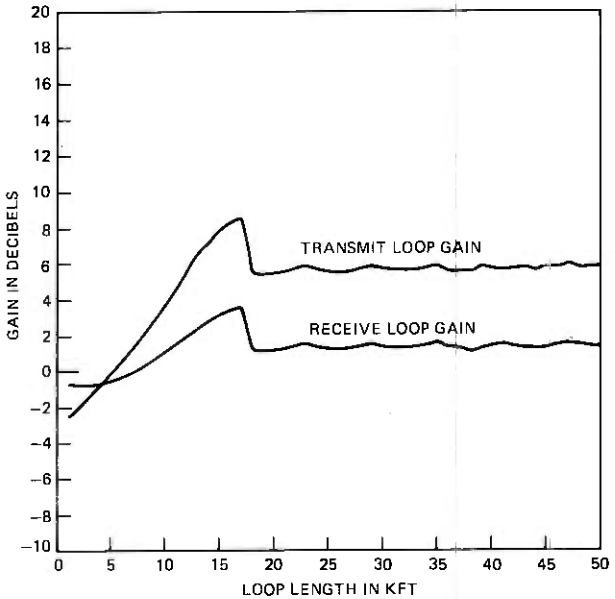


Fig. 11—The required gain on the transmit and receive loop derived from the curves of Figs. 9 and 10.

3.1 Theoretical resistance-design loop

As discussed in Section 2.3, the crosstalk probability for a loop is obtained by summing the crosstalk probabilities between that loop and the rest of the loops in the same cable, considering the four potential crosstalk exposures shown in Fig. 7: LTNEXT, LTFEXT, CONEXT, and COFEXT. The crosstalk probability for loop pair i, P_i, for example, is obtained first by determining the probability P_{ij} for all j, $j \neq i$, by eq. (16) in connection with eq. (3) through (6) for the four crosstalk exposures and then summing P_{ij} over j, as expressed by eq. (17). The crosstalk probability P_i so determined for loop pair i will be referred to as the total crosstalk probability of the pair, and represents the probability of receiving intelligible crosstalk on that loop from any of the remaining loops in the cable through any of the four possible crosstalk exposures.

Table I presents the total crosstalk probabilities calculated for each of the 25 loops of the 25-pair cable used in the cable crosstalk coupling model, with intraoffice type disturbing and disturbed connections. The crosstalk probability with the toll type disturbing connection was almost the same as that with the intraoffice type disturbing connection because, as discussed in Section 2.2, there was very little difference between the intraoffice and toll talker volume statistics.[3] All probabilities discussed hereafter are the probabilities with the intraoffice type disturbing connection.

For the particular results shown in Table I, the two loops of the dis-

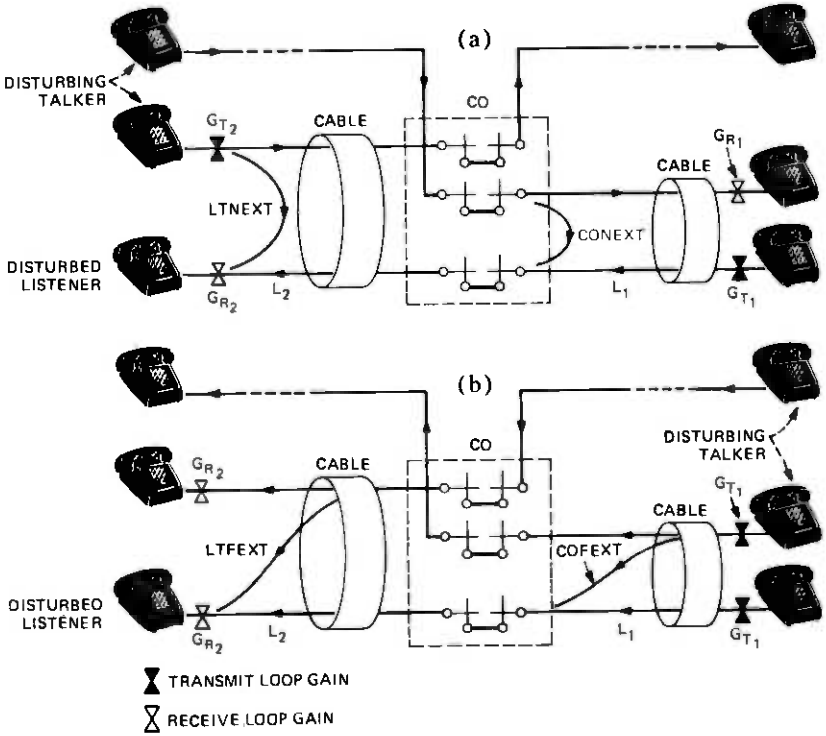


Fig. 12—Application of gain at the telephone set. (a) NEXT. (b) FEXT.

turbed connections were both assumed to be 7 kft, which was estimated to be a typical length of the Bell System loops, based on the 1964 Loop Survey.[2] The loop length dependence of the crosstalk probability is discussed later. As can be seen in this table, there is a wide difference in crosstalk probability between pairs in a cable: the highest crosstalk probability is 3.19×10^{-4} percent (pair 18), the median probability, 1.45×10^{-5} percent (pair 14), and the smallest probability, 1.15×10^{-6} percent (pair 25).

The crosstalk probability of the worst loop, pair 18, was evaluated as a function of the disturbed customer's loop length as presented in Fig. 14. Unlike the disturbed customer's loop, which is permanently assigned to the customer, the other loop of the disturbed connection occurs randomly, depending on the called party. The length of this latter loop was fixed at 7 kft, the representative length mentioned previously. The dashed curves show the crosstalk probabilities for the four exposures, LTNEXT, LTFEXT, CONEXT, and COFEXT, and the solid curve shows the total crosstalk probability, the sum of the four probabilities. As can be seen, the probability of LTNEXT is dominant at all loop lengths except at lengths less than about 2 kft at which the probability of COFEXT is dominant.

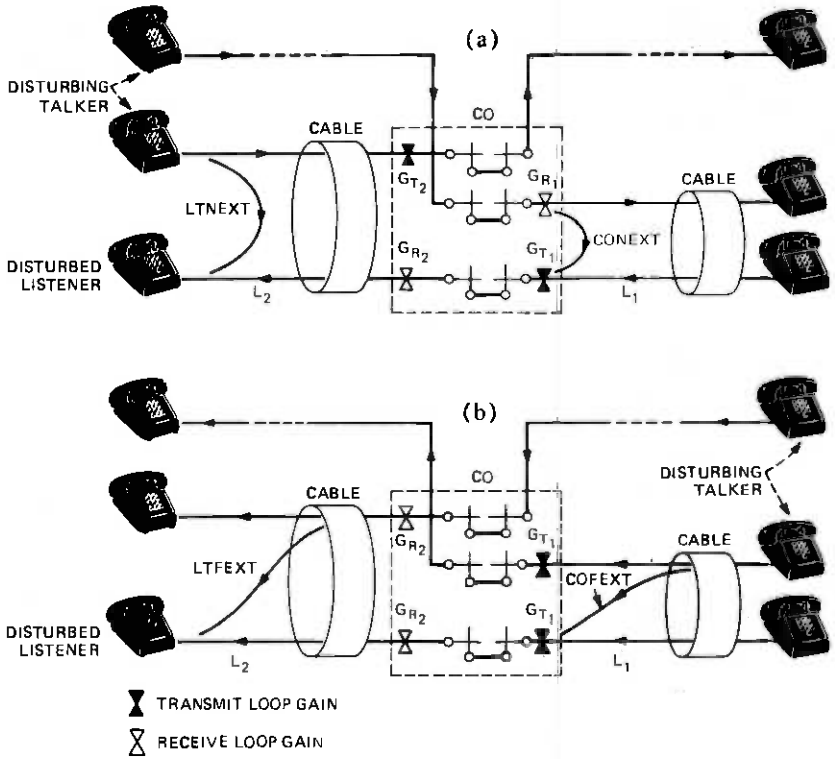


Fig. 13—Application of gain at the central office. (a) NEXT. (b) FEXT.

Since LTNEXT is the dominant crosstalk, the pattern of variation with loop length of the total crosstalk probability in Fig. 14 is determined by the pattern of the LTNEXT probability variation. The behavior of the LTNEXT probability with loop length can be explained by considering the corresponding crosstalk volume equation (3). As can be seen from Figs. 5 and 8, both NEXT coupling loss and line-terminal electrical speech level decrease with increasing loop length. At short loop lengths, since NEXT coupling loss decreases with loop length much faster than disturbing speech volume, the received crosstalk volume of LTNEXT, and consequently the LTNEXT probability, increases with loop length. As loop length is increased further, however, NEXT coupling loss approaches a saturation, that is, the length translation factor given in Fig. 5 does not change, whereas disturbing speech volume still decreases steadily with loop length. Therefore, the received crosstalk volume for LTNEXT, and consequently the LTNEXT probability, decreases as loop length is increased beyond a certain point; in this case, about 9 kft.

According to the resistance-design rules,[1] a cable is loaded when its length exceeds 18 kft. As discussed in Section 2.1, a loaded cable is assumed to have a NEXT coupling loss 3 dB less than a nonloaded cable.

Table I — The total crosstalk probabilities of the 25 loops of the 25-pair, 26-gauge, nonloaded PIC cable, obtained by treating each loop as a 7 kft, theoretical resistance-design loop engaged in an intraoffice (loop-to-loop) connection

Rank	Pair No.	Crosstalk Probability (%)
1	18	3.19×10^{-4}
2	8	2.82×10^{-4}
3	10	2.21×10^{-4}
4	4	1.93×10^{-4}
5	7	1.43×10^{-4}
6	19	1.23×10^{-4}
7	5	1.22×10^{-4}
8	20	1.20×10^{-4}
9	24	6.29×10^{-5}
10	22	4.54×10^{-5}
11	11	4.02×10^{-5}
12	2	3.69×10^{-5}
13	14	1.45×10^{-5}
14	15	1.29×10^{-5}
15	13	1.02×10^{-5}
16	12	1.00×10^{-5}
17	9	4.41×10^{-6}
18	23	3.64×10^{-6}
19	6	3.06×10^{-6}
20	17	2.60×10^{-6}
21	16	2.45×10^{-6}
22	1	2.01×10^{-6}
23	21	1.96×10^{-6}
24	3	1.33×10^{-6}
25	25	1.15×10^{-6}

The sudden increase in the LTNEXT probability at 18 kft is due to the 3-dB drop in NEXT coupling loss with loading. At loop lengths greater than 18 kft, both disturbing talker's electrical signal level and NEXT coupling loss are fairly constant with loop length, and the LTNEXT probability does not change much with loop length.

The effect of gain on the crosstalk probability of the theoretical resistance-design loop is shown in Fig. 15. The solid curve is the total crosstalk probability without gain, the same curve as that shown in Fig. 14, and the two dashed curves are the total crosstalk probability with gain at the telephone set and at the central office, respectively. Without gain, the total crosstalk probability of the theoretical resistance-design loop does not exceed 0.002 percent at all loop lengths. Gain applied at the central office shows very little effect on the crosstalk probability. This is because gain applied at the central office does not affect LTNEXT, the dominant crosstalk, as shown by eq. (23). However, with gain applied at the telephone set line terminals, the total crosstalk probability of the theoretical resistance-design loop can increase up to as much as 0.5 percent, depending on loop length.

Currently, no crosstalk objectives exist for loops. However, for planning purposes, a crosstalk probability of 0.1 percent has generally been used in the past as a limit for satisfactory loop crosstalk performance.

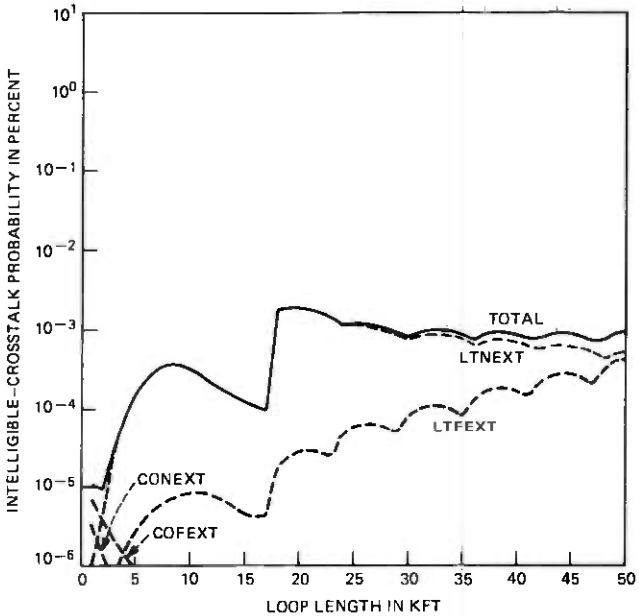


Fig. 14—Crosstalk probabilities of the theoretical resistance-design loop, without gain, evaluated for the worst pair (pair 18) of the 25-pair, 26-gauge PIC cable.

In comparison with this limit, the crosstalk performance of the present resistance-design loops is, from Fig. 15, more than satisfactory.

For the particular example of gain application considered in this paper, gain applied at the telephone set can cause a significant degradation in loop crosstalk performance, depending on loop length. To relate the increase in crosstalk probability to the amount of gain applied, one may compare Figs. 11 and 15. Figure 15 shows that, with gain applied at the telephone set, the crosstalk probability exceeds the 0.1-percent level, the limit mentioned previously, at about 12 kft of loop length. From Fig. 11, one may find that the required gain assumed at this length is 6 dB for the transmit loop and 2 dB for the receive loop, which amounts to a total gain of 8 dB on a crosstalk path. The maximum allowable telephone set gains at other loop lengths and for other values of permitted crosstalk probability can be determined similarly.

With gain applied at the central office, the crosstalk performance of the theoretical resistance-design loop still remains well below the level of 0.1-percent crosstalk probability, for the entire range of gain considered, where the maximum transmit and receive loop gains were about 9 and 4 dB. In a similar evaluation made previously, Lapsa[5] concluded that 9 dB of gain applied at the central office would be excessive. Because of the differences in the methodology as well as in the coupling loss and speech volume data used in the evaluation, a direct comparison between

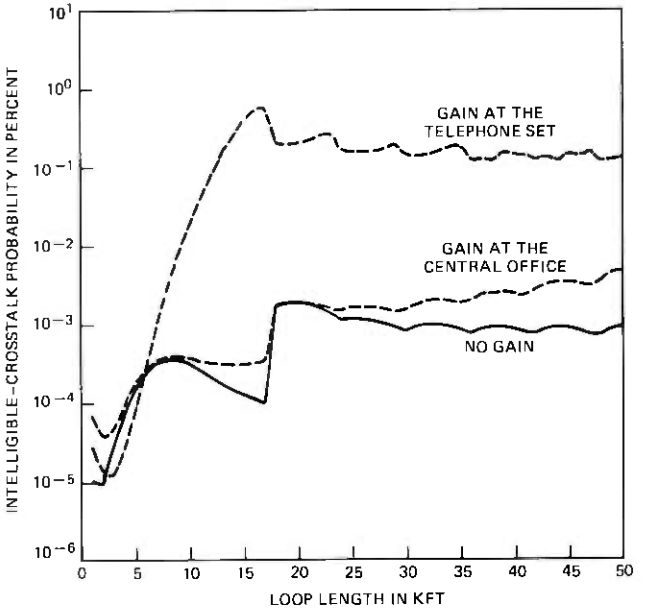


Fig. 15—The effect of gain on the crosstalk probability of the theoretical resistance-design loop, evaluated for the worst pair (pair 18) of the 25-pair, 26-gauge PIC cable.

the previous results and the present results is not possible. Nevertheless, the present results on the effect of the central office gain are, in general, somewhat optimistic in comparison with the previous results because of, among other things, the use of more recent coupling loss and speech volume data in the present study.*

3.2 The 1964 survey loops

The 1964 Loop Survey results[2] provide such information as length and loading conditions on 1100 loops sampled in the plant. Using this information, the crosstalk probabilities were calculated for the 1100 sample loops, first without gain and then with gain assumed either at the telephone set or at the central office. Each loop was treated as though it was the worst pair in a cable, such as pair 18 of Table I. This worst-case evaluation was made because, due to the permanent assignment of a loop to a customer, poor crosstalk performance would focus on a single customer rather than being distributed among many customers.

The total crosstalk probabilities calculated for the 1100 sample loops are shown in Fig. 16 as a scatter plot, where the abscissa is the length and

* The more recent coupling loss data used in the present study show better crosstalk performance than the coupling loss data used in the previous study. As discussed in Section 2.2, the speech volume data used in the present study show a much smaller standard deviation than the McAdoo data used in the previous study, the smaller speech volume variability yielding a smaller crosstalk probability.

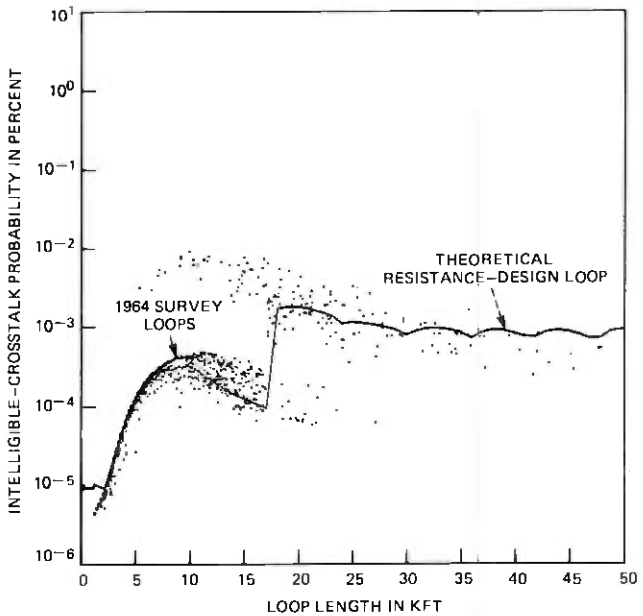


Fig. 16—Scatter plot of the total crosstalk probabilities of the 1964 survey loops, without gain, obtained by assuming that each loop was the worst pair in a cable.

the ordinate the crosstalk probability. For comparison, the total crosstalk probability of the theoretical resistance-design loop is superimposed as a solid curve, which is the same curve as that shown in Fig. 14. The cumulative distribution functions (CDFs) of the crosstalk probabilities of the 1100 sample loops without gain are presented in Fig. 17, where the solid curve shows the CDF of the total crosstalk probability and the dashed curves show the CDFs of the LTNEXT, LTFEXT, CONEXT, and COFEXT probability.

The effect of gain on the crosstalk performance of the sample loops was evaluated with the required gain determined by the difference between the constant TLR of −21 dB and RLR of 27 dB mentioned previously and the actual TLR and RLR, which were calculated from the information provided by the 1964 Loop Survey. The results are compared with the crosstalk probability determined for the present plant (loops without gain) in Fig. 18. The solid curve is the CDF of the crosstalk probability of the sample loops without gain (the same curve as that shown in Fig. 17) and the two dashed curves show the CDFs of the crosstalk probabilities with gain applied at the telephone set and at the central office, respectively.

Without gain, the total crosstalk probability is less than 0.01 percent for all the sample loops; the median is 3×10^{-4} percent. This indicates that the crosstalk performance of the present loop plant is more than

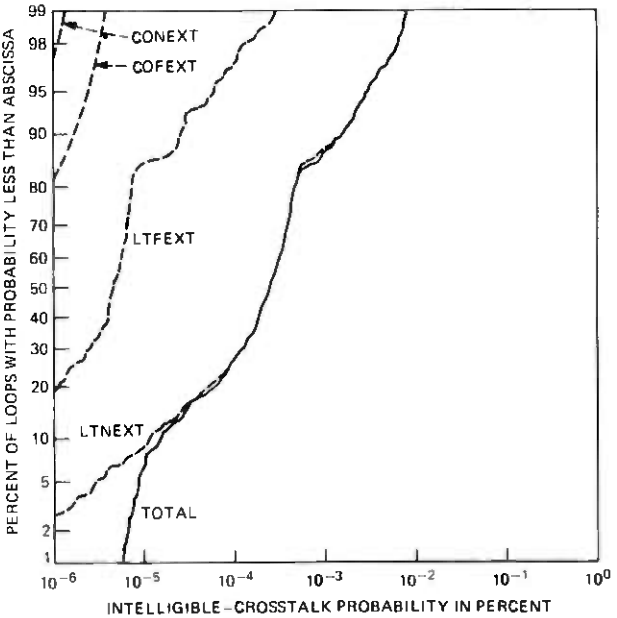


Fig. 17—Cumulative distribution functions of the crosstalk probabilities of the 1964 survey loops, without gain, obtained by assuming that each loop was the worst pair in a cable.

satisfactory in comparison with the 0.1 percent crosstalk probability limit. Gain applied at the central office shows only a small effect on the distribution of the loop crosstalk probabilities in the plant. However, gain applied at the telephone set changes the distribution of the loop crosstalk probabilities significantly, increasing the crosstalk probability above the 0.1-percent level on about 15 percent of the sample loops evaluated.

IV. SUMMARY OF THE RESULTS

The intelligible crosstalk probability is defined as the probability that a customer will hear one or more intelligible crosstalk words during a call. The intelligible crosstalk probability for a loop is obtained by summing the probabilities of intelligible crosstalk between that loop and the rest of the loops in the same cable, considering the four potential crosstalk exposures shown in Fig. 7. Using the methodology developed in Section II, the crosstalk probabilities have been calculated first for theoretical maximum-loss resistance-design loops[1] as a function of loop length and then for the 1100 loops of various lengths sampled from the loop plant in the 1964 Loop Survey.[2]

The crosstalk probabilities were obtained first for loops as they exist in the present plant, that is, loops without gain. The effect of gain devices on the loop crosstalk probabilities was then evaluated for a particular

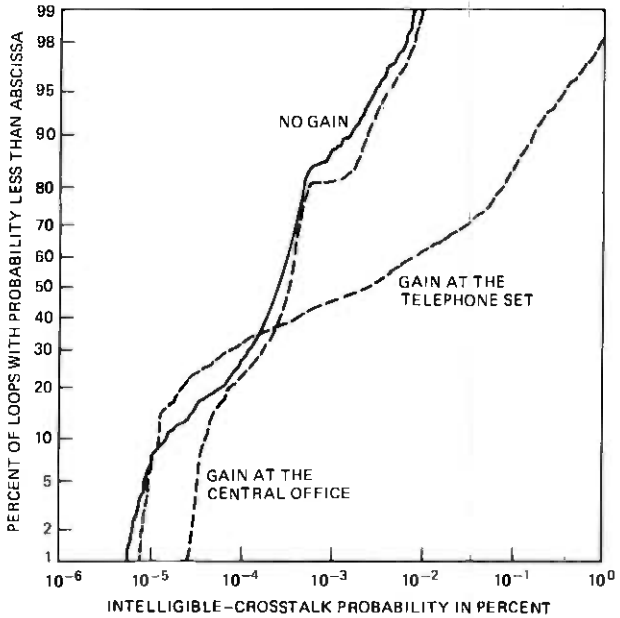


Fig. 18—The effect of gain on the distribution of the total crosstalk probabilities of the 1964 survey loops, obtained by assuming that each loop was the worst pair in a cable.

example of gain application. In this example, the assumed gain was determined as a function of loop length to meet a constant TLR (Transmit Loop Rating) of −21 dB and RLR (Receive Loop Rating) of 27 dB, regardless of loop length, which would equalize the EARS (Electro-Acoustic Rating System)* loss of intraoffice (loop-to-loop) connections at a constant value of 6 dB. For this particular example, gain required for a loop in its transmit direction and receive direction ranged roughly from −3 to 9 dB and from −1 to 4 dB, respectively. Two possible locations of gain application were evaluated: the central office and the telephone set.

Table I shows rank-ordered crosstalk probabilities of the 25 theoretical resistance-design loops without gain in a 25-pair cable, determined with loop length fixed at 7 kft, a representative length of Bell System loops. Figure 15 presents the crosstalk probability of the worst of the 25 loops (pair 18 in Table I) as a function of loop length for the three different cases: loops without gain (the present plant), loops with gain at the central office, and loops with gain at the telephone set. Figure 18 presents the cumulative distribution functions (CDFs) of the crosstalk probabilities of the 1100 sample loops obtained by treating each sample loop as the worst loop in a cable (such as pair 18 of Table I).

Presently, no crosstalk objectives exist for loops. For planning pur-

* See Section 2.4.1 of this paper and Ref. 9 for the discussion of EARS, TLR, and RLR.

poses, however, a crosstalk probability of 0.1 percent has generally been used as a limit for satisfactory loop crosstalk performance. In comparison with this limit, the crosstalk performance of the present loop plant (loops without gain) is more than satisfactory, as can be seen in Fig. 18. With gain at the central office, the crosstalk probability still remains well below the 0.1-percent level for all the sample loops, and thus gain applied at the central office does not appear to have any significant effect on loop crosstalk performance for the entire range of gain considered. However, with gain applied at the telephone set, the crosstalk probability exceeds the 0.1-percent level on about 15 percent of the loops evaluated. These results indicate that, for the particular example of gain application considered in this paper, gain applied at the telephone set may cause a significant crosstalk performance degradation.

V. ACKNOWLEDGMENTS

The author wishes to thank P. C. Lopiparo for devoting considerable time to many helpful discussions during this study, J. Kreutzberg and his associates for providing the crosstalk coupling loss data, F. P. Duffy and his associates for providing the speech volume data obtained from the 1975–1976 Loop Signal Power Survey, F. B. Stallman for providing the computer program used in calculating some of the loop parameters, and A. M. Lessman, T. C. Spang, and J. L. Sullivan for their useful comments.

REFERENCES

1. Members of the Technical Staff, Bell Laboratories, *Transmission Systems for Communications*. Revised Fourth Edition, February 1970.
2. P. A. Gresh, "Physical and Transmission Characteristics of Customer Loop Plant," B.S.T.J., *48*, No. 10 (December 1969), pp. 3337–3385.
3. F. P. Duffy, W. C. Ahern, and J. A. Maher, "Speech Signal Power in the Switched Message Network," B.S.T.J., *57*, No. 7 (September 1978), pp. 2695–2726.
4. D. H. Morgen, "Expected Crosstalk Performance of Analog Multichannel Subscriber Carrier Systems," IEEE Trans. Commun., *COM-23*, No. 2 (February 1975), pp. 240–245.
5. P. M. Lapsa, "Calculation of Multidisturber Crosstalk Probabilities—Application to Subscriber-Loop Gain," B.S.T.J., *55*, No. 7 (September 1976), pp. 875–903.
6. K. L. McAdoo, "Speech Volumes on Bell System Message Circuits—1960 Survey," B.S.T.J., *42*, No. 5 (September 1963), pp. 1999–2012.
7. K. I. Park, unpublished work.
8. T. K. Sen, "Masking of Crosstalk by Speech and Noise," B.S.T.J., *49*, No. 4 (April 1970), pp. 561–584.
9. J. L. Sullivan, "A Laboratory System for Measuring Loudness Loss on Telephone Connections," B.S.T.J., *50*, No. 8 (October 1971), pp. 2663–2739.
10. A. M. Lessman, unpublished work.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 8, October 1978
Printed in U.S.A.

On the Stability of Interconnected Systems*

By I. W. SANDBERG

(Manuscript received March 3, 1978)

Theorems are presented concerning conditions for the input-output stability of interconnected dynamical systems. Results in the area of input-output stability are often partitioned into two categories: small-gain type-results and passivity-type results. The main theorem given here does not fall into either of these categories, but is most closely related to the passivity-type results. The theorem involves a new class of interconnection operators that is a substantial generalization of the familiar set of nonnegative operators defined on a space of vector-valued functions.

I. INTRODUCTION

In this paper, theorems are presented concerning the input-output stability of interconnected systems.† Results in the area of input-output stability are often partitioned into two categories: small-gain type results and passivity-type results. The main theorem given here does not fall into either of these categories, but is most closely related to the passivity-type results. The theorem involves a new class of interconnection operators that is a substantial generalization of the familiar set of nonnegative operators defined on a space of vector-valued functions.

The mathematical model considered throughout the paper is described in Section II, and results of a general nature concerning the model are given in Section III. The case in which the interconnection operator has a certain matrix representation is treated in considerable detail in Section IV. In Section 4.5, a specific example is given of a stable interconnected system for which the interconnection matrix does not meet the nonnegative-definiteness requirement of the criterion given in Ref. 7, which contains the most pertinent earlier stability result.

* This paper was presented at the 1978 IEEE Symposium on Circuits and Systems (New York, May 17–19, 1978).

† For background material in book form concerning intput-output stability, see, for example, Refs. 1–4. Interconnected systems (which are systems whose natural or artificial decomposition into subsystems plays a prominent role in their mathematical analysis) have been considered by many researchers. See, for example, Refs. 3, 5, 6, and 7. Although some interesting and significant results have been obtained concerning the stability of interconnected systems, the theory is very much in its initial stages of development.

The main purpose of the paper is to introduce a new concept that is believed to be useful. No attempt is made to present the sharpest possible stability results that the concept can be used to obtain.

II. THE MODEL

2.1 Preliminaries

Let K denote a real linear space that contains a normed linear inner-product space L with inner product $(\cdot,\cdot)$ and norm $|\cdot|$ related by $|f| = (f,f)^{1/2}$ for $f \in L$. (Of particular interest to us is the case in which L is the set L_2 of all real Lebesgue square-integrable functions defined on the half line $[0,\infty)$ with the usual inner product, and K is the "extended" set E_2 of real functions defined on $[0,\infty)$ such that each function is square integrable on $[0,\tau]$ for any nonnegative number τ.)

For each $\tau \geq 0$, let P_τ denote a linear mapping of K into L (e.g., if $K = E_2$, let P_τ be defined by $(P_\tau f)(t) = f(t)$ for $t \in [0,\tau]$ and $f(t) = 0$ for $t > \tau$, where f is an arbitrary element of E_2).

Let K, L, and P_τ be such that (i) $g \in L$ if and only if $g \in K$ and $\sup_\tau |P_\tau g| < \infty$, (ii) $|g| = \sup_\tau |P_\tau g|$ for $g \in L$, and (iii) $(P_\tau f,g) = (P_\tau f,P_\tau g)$ and $|P_\tau f| \leq |f|$ for f and g in L and $\tau \geq 0$.

We let L^n and K^n, in which n is any positive integer, denote the n-fold Cartesian product of L and K, respectively. The norm of an element $h = (h_1,h_2,\ldots,h_n)$ of L^n is denoted by $|h|$ and is defined by $|h| = (\Sigma_i |h_i|^2)^{1/2}$.

It is assumed that L contains n elements $e_1,e_2,\ldots,e_n$ such that $|e_i| = 1$ for each i and $(e_i,e_j) = 0$ for $i \neq j$.*

We say that an operator T that maps K into itself (i.e., an operator T in K) is *causal* if and only if $P_\tau T = P_\tau T P_\tau$ on K for all $\tau \geq 0$.

2.2 The basic equations

Throughout the paper, attention is focused on an interconnected system governed by

$$x_i + \sum_{j=1}^{n} A_{ij}B_j x_j = y_i, \quad i = 1,2,\ldots,n, \tag{1}$$

in which $(A.1)$: x_i and y_i belong to K for all i, and A_{ij} and B_j are causal operators in K for all i and j.

In (1), each B_j is associated (sometimes somewhat indirectly) with a subsystem, and the A_{ij} ordinarily take into account the way in which the subsystems interact. Typically, it is not difficult to show the existence of a solution $x_1, x_2, \ldots, x_n$ of (1) for any given $y_1, y_2, \ldots, y_n$ under some weak additional hypotheses. (Successive-approximation type arguments of the kind commonly used in connection with nonlinear Volterra integral equations often suffice.)

* This assumption is used only in Section IV.

We assume that (*A.2*): each B_j is nonnegative in L, in the sense that each B_j maps L into L and there exists a nonnegative constant α such that

$$(B_j w, w) \geq \alpha |w|^2 \tag{2}$$

for $w \in L$ and all j. It is assumed also that (*A.3*): each A_{ij} maps L into itself, and there is a positive constant γ such that

$$|A_{ij} w| \leq \gamma |w| \tag{3}$$

for $w \in L$ and all i and j.*

It is often convenient to write (1) in the form

$$x + ABx = y, \tag{4}$$

in which $x = (x_1, x_2, \ldots, x_n)$, $y = (y_1, y_2, \ldots, y_n)$, and A and B are the mappings of K^n into K^n defined by $(Af)_i = \Sigma_j A_{ij} f_j$ and $(Bf)_i = B_i f_i$ for all $f \in K^n$ and each i.

2.3 Definition of stability

We say that (4) is *L-stable* if and only if $y \in L^n$ implies that $x \in L^n$ with $|x| \leq \rho(|y|)$ for some nonnegative continuous function ρ that depends only on A and B and is defined on the nonnegative reals such that $\rho(0) = 0$.

III. S_β AND THE MAIN THEOREM

In the following definition and hypothesis, β is a nonnegative number.

Definition of S_β: S_β is the set of operators H in L^n with the following property: For each $v \in L^n$ such that $|v| \neq 0$, there is an index k such that $|v_k| \neq 0$ and $(v_k, (Hv)_k) \geq \beta |v_k|^2$.

The definition of S_β is related to one of the equivalent definitions of a "P-matrix." [9]

H.1: If $\beta = 0$, there is a positive constant δ such that

$$|B_j w| \leq \delta |w| \tag{5}$$

for $w \in L$ and all j.

Our main result is the following.

Theorem 1: Let H.1 (and A.1 through A.3) hold. Then (4) is L-stable if $A \in S_\beta$ *with* $\alpha + \beta > 0$.

---

* In order to focus attention only on essentials, we are proceeding with some assumptions that are stronger than necessary. It will become clear that (2) and (3) (and also (5) of Section III) could have been replaced with somewhat weaker inequalities (similar, for example, to some of those used in Section 5.3 of Ref. 8). Similarly, if for example there are positive constants α_j such that $(B_j w, w) \geq \alpha_j |w|^2$ for $w \in L$ and all j, and if A in (4) is represented by an $n \times n$ matrix in the sense of Section IV with $[I + \alpha \text{ diag } (\alpha_j)]$ invertible, then it is clear that x satisfies an equation similar to (4) in which each $B_j x_j$ is replaced with $(B_j x_j - \alpha_j x_j)$ and A and y are modified accordingly. Consideration of such a modified equation often enables a useful trade-off to be made between requirements on A and the degree of positiveness of the B_j.

Proof of Theorem 1: Let *H.1* and *A.1* through *A.3* be satisfied, let $y \in L^n$, let $x \in K^n$ be a solution of (4), and assume that $A \in S_\beta$ with $\alpha + \beta > 0$.

Suppose first that $\beta = 0$, in which case $\alpha > 0$. Let $\tau \geq 0$ be arbitrary. Using (1) and the casuality of the A_{ij}, we have

$$P_\tau x_i + P_\tau \sum_j A_{ij} P_\tau B_j x_j = P_\tau y_i \tag{6}$$

and

$$(P_\tau B_i x_i, P_\tau x_i) + \left(P_\tau B_i x_i, P_\tau \sum_j A_{ij} P_\tau B_j x_j\right) = (P_\tau B_i x_i, P_\tau y_i) \tag{7}$$

for $i = 1,2,\ldots,n$. Since $A \in S_\beta$, and $(P_\tau B_i x_i, P_\tau \Sigma_j A_{ij} P_\tau B_j x_j) = (P_\tau B_i x_i, \Sigma_j A_{ij} P_\tau B_j x_j)$ for each i, there is an index k such that $(P_\tau B_k x_k, P_\tau \Sigma_j A_{kj} P_\tau B_j x_j) \geq 0$ and hence such that

$$(P_\tau B_k x_k, P_\tau x_k) \leq (P_\tau B_k x_k, P_\tau y_k). \tag{8}$$

By the Schwarz inequality and the fact that $(P_\tau B_k x_k, P_\tau y_k) = (P_\tau B_k P_\tau x_k, P_\tau y_k) = (B_k P_\tau x_k, P_\tau y_k)$, $(P_\tau B_k x_k, P_\tau y_k) \leq |B_k P_\tau x_k| \cdot |P_\tau y_k|$. Therefore, using $(P_\tau B_k x_k, P_\tau x_k) = (B_k P_\tau x_k, P_\tau x_k)$ as well as (2), (5) and (8), we have

$$\alpha |P_\tau x_k|^2 \leq \delta |P_\tau x_k| \cdot |P_\tau y_k| \tag{9}$$

and consequently, with $c = \delta \alpha^{-1}$,

$$|P_\tau x_k| \leq c |P_\tau y_k|.$$

The argument given above shows that $|P_\tau x_k| \leq c|y|$ for some k (which might depend on x and τ). Let J denote any nonempty proper subset of $\{1,2,\ldots,n\}$ with the following property. For $j \in J$, there is a constant c_j that depends only on c, δ, and γ such that $|P_\tau x_j| \leq c_j |y|$. Using (1),

$$x_i + \sum_{j \notin J} A_{ij} B_j x_j = y_i - \sum_{j \in J} A_{ij} B_j x_j, \quad i \notin J. \tag{10}$$

The left side of (10) is basically the same in form as the left side of (1). With r the number of elements contained in J, let the elements of $(\{1,2,\ldots,n\} - J)$ be $j_1, j_2, \ldots, j_{(n-r)}$ ordered so that $j_1 < j_2 < \cdots < j_{(n-r)}$. With respect to that ordering, let the mapping of $K^{(n-r)}$ into itself associated with (10) that corresponds to A be denoted by A_J. Since *A.3* holds, each A_{ij} maps the zero element of L into itself, and it is a simple matter to verify that A_J belongs to, so to speak, S_β with n replaced with $(n - r)$. Thus, by the argument given above, we find that there is an index $l \notin J$ such that

$$|P_\tau x_l| \leq c \left| P_\tau \left(y_l - \sum_{j \in J} A_{lj} B_j x_j \right) \right|. \tag{11}$$

Using $|P_\tau x_j| \le c_j|y|$ for $j \in J$, as well as (11) and the causality of the A_{ij} and the B_j, we have

$$|P_\tau x_l| \le c\left(|y| + \left|\sum_{j\in J} P_\tau A_{lj}B_j x_j\right|\right)$$

$$\le c\left(|y| + \sum_{j\in J} |P_\tau A_{lj}P_\tau B_j P_\tau x_j|\right)$$

$$\le c\left(|y| + \gamma\delta \sum_{j\in J} |P_\tau x_j|\right)$$

$$\le c\left(1 + \gamma\delta \sum_{j\in J} c_j\right)|y|.$$

Let $\omega_1,\omega_2,\ldots,\omega_n$ be defined by $\omega_1 = c$ and

$$\omega_i = c\left(1 + \gamma\delta \sum_{j=1}^{(i-1)} \omega_j\right), \quad i = 2,3,\ldots,n.$$

We have shown that given x, τ, and i, $|P_\tau x_i| \le d_i|y|$ for some $d_i \in \{\omega_1,\omega_2,\ldots,\omega_n\}$. Since $\omega_n = \max_i \omega_i$, we have

$$\sum_{i=1}^{n} |P_\tau x_i|^2 \le n\omega_n{}^2|y|^2 \quad \text{for all } \tau \ge 0,$$

which shows that $x \in L^n$ and that $|x|$ is suitably bounded in terms of $|y|$. This completes the proof for the case in which $\beta = 0$.

The proof for the $\beta > 0$ case is similar. Using primarily (7) and the hypothesis that $A \in S_\beta$, we find that

$$\beta|P_\tau B_k x_k|^2 \le |P_\tau B_k x_k| \cdot |P_\tau y_k|$$

for some k. Therefore,

$$|P_\tau B_k x_k| \le \beta^{-1}|y| \tag{12}$$

for some k. By proceeding essentially as indicated above, we can show that

$$|P_\tau B_i x_i| \le \Omega_n|y| \quad \text{for all } i \text{ and } \tau \ge 0, \tag{13}$$

in which Ω_n is the number defined by $\Omega_1 = \beta^{-1}$ and

$$\Omega_i = \beta^{-1}\left(1 + \gamma \sum_{j=1}^{(i-1)} \Omega_j\right)$$

for $i = 2,3,\ldots,n$.

From (6) and (13),

$$|P_\tau x_i| \le |P_\tau y_i| + \sum_j |A_{ij}P_\tau B_j x_i|$$

$$\le |y| + \sum_j \gamma\Omega_n|y|$$

$$\le (1 + n\gamma\Omega_n)|y|$$

for each i and all $\tau \geq 0$. Since this shows that $x \in L$ and that $|x|$ is bounded as required, our proof is complete.

3.1 Comments

To see that $\bigcup_{\eta \geq 0} S_\eta$ contains the familiar set of nonnegative operators defined on a space of vector-valued functions, let H be any mapping of L^n into itself such that

$$\sum_{i=1}^{n} (w_i, (Hw)_i) \geq \sigma |w|^2, \quad w \in L^n, \tag{13}$$

in which σ is a nonnegative constant. From (13) it is clear that for each $w \in L^n$ with $|w| \neq 0$ there is an index k such that $|w_k| \neq 0$ and $(w_k, (Hw)_k) \geq \beta |w|^2$ in which $\beta = \sigma n^{-1}$. Since $|w|^2 \geq |w_k|^2$, we observe that $H \in S_\beta$.

Theorem 1 is of course a result concerning the L-boundedness of solutions of (4).* By modifying the hypotheses and proof of Theorem 1 in a direct manner, an analogous result can be obtained concerning the L-continuity of solutions (i.e., concerning the L-boundedness of the difference $(x_a - x_b)$ of a solution x_a of (4) that corresponds to $y = y_a$ and a solution x_b that corresponds to $y = y_b$ with $(y_a - y_b) \in L$). With regard to the necessary modifications of the hypotheses concerning A, the following definition, in which $\beta \geq 0$, plays a central role.

Definition of T_β: T_β is the set of operators H in L^n with the following property: For each u and v in L^n such that $|u - v| \neq 0$, there is an index k such that $|u_k - v_k| \neq 0$ and $(u_k - v_k, (Hu)_k - (Hv)_k) \geq \beta |u_k - v_k|^2$.

In order to be more explicit, let (*A.1′*) denote the assumption that $x_a + ABx_a = y_a$ and $x_b + ABx_b = y_b$ in which each A_{ij} and B_j are causal operators in K and x_a, x_b, y_a, and y_b belong to K. Let *A.2′* be the hypothesis obtained from *A.2* by replacing "$(B_j w, w) \geq \alpha |w|^2$ for $w \in L$ and all j" with "$(B_j u - B_j v, u - v) \geq \alpha |u - v|^2$ for u and v in L and all j," and let *A.3′* and *H.1′* be the hypotheses obtained in a similar manner from *A.3* and *H.1*, respectively.

Our L-continuity result (whose proof is omitted) is the following.

Theorem 2: Let H.1′ and A.1′ through A.3′ be satisfied, let $(y_a - y_b) \in L$, and let $A \in T_\beta$ with $\alpha + \beta > 0$. Then $(x_a - x_b) \in L$, and there is a nonnegative continuous function ρ defined on $[0, \infty)$ that depends only on A and B such that $\rho(0) = 0$ and $|x_a - x_b| \leq \rho(|y_a - y_b|)$.

* Results along the lines of Theorem 1 for cases in which B is more general than assumed here but both A and B are nonnegative operators are given in Ref. 8, where the stability of interconnected systems in the sense of Section 2.2 is not explicitly discussed. A nonnegative-operator approach to the stability of interconnected systems, as well as its relation to other approaches, is discussed in Ref. 7.

We shall refer to the following two hypotheses.

H.2: Each B_i is a continuous mapping in L that maps the zero element into itself, and there are positive constants c_1 and c_2 such that for all i

$$(B_i u, u) \geq c_1 |B_i u|^2 \tag{14}$$

$$|B_i u - B_i v| \leq c_2 |u - v| \tag{15}$$

for u and v in L.

H.3: $K = E_2$, $L = L_2$, and for each $\tau \geq 0$ P_τ is the operator associated with E_2 in the example given in Section 2.1.

There are many cases in which (14) holds* for all u for some $c_1 > 0$, but there is no positive α such that (2) is satisfied for all w.† On the other hand, it is clear that there is a positive c_1 with the property that (14) is met when (2) holds with $\alpha > 0$ for all w and there is a positive constant δ such that $|B_i w| \leq \delta |w|$ for all w.

Corollary 1: Let H.2 and H.3 (as well as A.1 and A.3) be satisfied. If A $\in S_0$, then (4) is L-stable.

Proof: Assume that the hypotheses of the corollary are satisfied and let I and I_n, respectively, denote the identity operators in K and K^n. With regard to the following lemma, two elements u and v of E_2 are taken to be the same if and only if $|P_\tau(u - v)| = 0$ for all $\tau \geq 0$.

Lemma 1: Let H.3 hold, let F be a continuous mapping of L_2 into itself such that for some positive constant $c < 1$ we have

$$|Fu - Fv| \leq c|u - v| \quad \textit{for } u \textit{ and } v \textit{ in } L_2, \tag{16}$$

and let F also be a causal mapping of E_2 into E_2. Then $(I - F)^{-1}$ exists and is causal on both L_2 and E_2.

Proof of Lemma 1: Let the hypotheses of the lemma be met. In view of (16) and the continuity of F, the equation $x - Fx = h$ with $h \in L_2$ has in L_2 a unique solution x which is given by $x = \lim_{n \to \infty} x^{(n)}$ in which $x^{(n)} = h + Fx^{n-1}$ for $n \geq 1$ and $x^{(0)} = h$. Thus, $(I - F)^{-1}$ exists on L_2, and since $h + F(\cdot)$ is causal on L_2 so is $(I - F)^{-1}$.

Now let $h \in E_2$, and for each $\tau \geq 0$ let z_τ be the unique element of L_2 that satisfies $z_\tau - Fz_\tau = P_\tau h$. Since $(I - F)^{-1}$ is causal on L_2, it is clear that $P_{\tau_1} z_{\tau_2} = P_{\tau_1} z_{\tau_1}$ for $\tau_2 \geq \tau_1$. Let x be the element of E_2 defined by the condition that $P_\tau x = P_\tau z_\tau$ for all $\tau \geq 0$. For any $\tau \geq 0$, $P_\tau x - P_\tau Fx = P_\tau z_\tau - P_\tau FP_\tau x = P_\tau z_\tau - P_\tau FP_\tau z_\tau = P_\tau z_\tau - P_\tau Fz_\tau = P_\tau h$. Therefore x satisfies $x - Fx = h$. Suppose that x_1 and x_2 in E_2 satisfy $h = x_1 - Fx_1 = x_2 - Fx_2$

* This type of inequality is among those used in Ref. 8.

† We mention two simple examples: Let $L = L_2$ and let B_i be defined by the condition that for each $t \geq 0$, $(B_i w)(t) = w(t)$ for $|w(t)| \leq 1$ and $(B_i w)(t) = \text{sgn}(w(t))$ for $|w(t)| > 1$. Then (14) with $c_1 = 1$ holds for all $u \in L_2$, but there is no $\alpha > 0$ for which (2) is satisfied for all $w \in L_2$. It is not difficult to show that a similar conclusion is reached when B_i is the convolution operator in L_2 with impulse response e^{-t}.

with $|P_\tau(x_1 - x_2)| \neq 0$ for some $\tau \geq 0$. Since F is causal, we have $P_\tau h_1 = P_\tau h_2$ in which $h_1 = P_\tau x_1 - FP_\tau x_1$ and $h_2 = P_\tau x_2 - FP_\tau x_2$. This contradicts the fact that $(I - F)^{-1}$ is causal on L_2. Thus, x is the unique solution in E_2 of $x - Fx = h$, which means that $(I - F)^{-1}$ exists on E_2, because h is arbitrary. In view of the fact that the solution x of $x - Fx = h$ satisfies $P_\tau x = P_\tau z_\tau$ where $z_\tau - Fz_\tau = P_\tau h$ for every $\tau \geq 0$, it is evident that the operator $(I - F)^{-1}$ on E_2 is causal. This proves the lemma.

Let c be a positive constant such that $c < \min(c_1, c_2^{-1})$. By Lemma 1, $(I_n - cB)^{-1}$ exists on K^n and $B(I_n - cB)^{-1}$ is causal on K^n and maps L^n into itself. In particular, the equation $x + ABx = y$ can be written as

$$h + (A + cI_n)B(I_n - cB)^{-1}h = y$$

in which $h = x - cBx$. From $A \in S_0$, it follows at once that $(A + cI_n) \in S_c$.

Also, from $h = x - cBx$ and the fact that B is causal on K^n and satisfies $|Bu| \leq c_2|u|$ for $u \in L^n$, with $cc_2 < 1$, we have $|P_\tau x| \leq (1 - cc_2)^{-1}|P_\tau h|$ for $\tau \geq 0$.

Therefore, by Theorem 1, to complete the proof it suffices to observe that for any $w \in L^n$,

$$\begin{aligned}(B(I_n - cB)^{-1}w, w) &= (Bu, u - cBu) \\ &= (Bu, u) - c|Bu|^2 \\ &\geq 0\end{aligned}$$

in which of course $u = (I_n - cB)^{-1}w$.

IV. RESULTS CONCERNING THE MATRIX CASE

Of importance in the theory of interconnected systems is the special case in which A is represented by a real $n \times n$ matrix a with elements a_{ij}, in the sense that for each i,

$$(Aw)_i = \sum_{j=1}^{n} a_{ij}w_j, \quad w \in L^n.$$

Throughout this section, "$A \in M$" means that A has such a representation with representation matrix a, and, assuming that $A \in M$, $U_0(U)$ denotes the set of representation matrices such that $A \in S_0$ ($A \in S_\beta$ with $\beta > 0$). In addition, $P_0(P)$ denotes the set of real square matrices with nonnegative (positive) principal minors.

Proposition 1: If (A.1 through A.3 are satisfied and) $A \in M$ with $a \notin P_0$, then (4) is not L-stable for some B.

Proof: Let the hypotheses be met, and let 1_n denote the identity matrix of order n. From $a \notin P_0$, it follows that there is a diagonal matrix $d =$

diag $(d_1,d_2,\ldots,d_n)$ with $d_i > 0$ for all i such that $(d + a)$ and hence $(1_n + ad^{-1})$ are singular.*

For each i, let B_i be defined by the condition that $B_i w = d_i{}^{-1} w$ for $w \in K$. Let v be any real nonzero n-vector that is annihilated by $(1_n + ad^{-1})$, and let e be any element of K different from the zero element θ of L. With $x = ve$, we have $x + ABx = \theta$. This shows that (4) is not L-stable for the particular B constructed.

In order to proceed to the first theorem in this section, it is necessary to introduce the following definitions. For each $w \in L^n$, a_w denotes the matrix obtained from a by replacing a_{ij} with $a_{ij}(w_i,w_j)$ for all i and j. For any $w \in L^n$ with $|w| \neq 0$, $\Gamma(a_w)$ denotes the matrix obtained from a_w by deleting both the ith row and ith column for all $i \in \{i: |w_i| = 0\}$.

Theorem 3: Let $A \in M$. We have $a \in U (a \in U_0)$ if and only if $\Gamma(a_w) \in P (a_w \in P_0)$ for each $w \in L^n$ with $|w| \neq 0$.

Proof of Theorem 3: We shall use two lemmas. With regard to the first of the lemmas, M_n denotes the normed linear space of real $n \times n$ matrices, with the usual Euclidean norm, and C denotes $\{u \in M_n$: there is a $w \in L^n$ with $|w| = 1$ such that $u_{ij} = (w_i,w_j)$ for all i and $j\}$.

Lemma 2: C is compact.

Proof of Lemma 2: The set C is obviously bounded. To show that C is closed, let $u^{(1)},u^{(2)},\ldots$ be a sequence of elements of C that converges to some element $\tilde{u}$ of M_n.

Given a real n-vector v, for each $w \in L^n$ and its corresponding element u of C, we have $v^{tr}uv = (\Sigma_i v_i w_i, \Sigma_i v_i w_i) \geq 0$.† Thus, each $u^{(j)}$ is nonnegative definite, and therefore it follows that $\tilde{u}$ is nonnegative definite. In view of the fact that $Tr(u^{(j)}) = 1$ for all j, it also follows that $Tr(\tilde{u}) = 1$.

Since $\tilde{u}$ is nonnegative definite and has unit trace, there is an orthogonal matrix T and a diagonal matrix $D = \text{diag}\,(d_1,d_2,\ldots,d_n)$ with $d_j \geq 0$ and $\Sigma_j d_j = 1$ such that

$$\tilde{u} = TDT^{tr} = \sum_l d_l T_l (T_l)^{tr},$$

in which T_l is the lth column of T. Referring to the pairwise mutually orthogonal elements $e_1,e_2,\ldots,e_n$ mentioned in Section 2.1, let z in L^n be defined by

$$z = \sum_i d_i{}^{1/2} T_i e_i.$$

Using the orthonormality of the e_i, and the fact that for each l the sum of the squares of the components of T_l is unity, it is not difficult to verify that $|z| = 1$, and that (z_i,z_j) is equal to the i,jth element of $\tilde{u}$ for all i and

* A proof is given in Ref. 10. Another proof can be obtained from the fact that, since $a \notin P_0$, there is (see Ref. 9) a real nonzero n-vector q such that $q_i(aq)_i < 0$ for every i such that $q_i \neq 0$.

† The superscript "tr" denotes transpose.

j. This shows that C is closed, and completes the proof of the lemma.*

The following lemma is proved in Ref. 9.

Lemma 3: A real square matrix m belongs to $P(P_0)$ if and only if $v_k(mv)_k > 0$ ($v_k \neq 0$ and $v_k(mv)_k \geq 0$) for some k for each real nonzero vector v of dimension equal to the order of m.

In order to prove the theorem, suppose initially that $\Gamma(a_w) \in P$ for every $w \in L^n$ with $|w| \neq 0$. By Lemma 3, for each $w \in L^n$ with $|w| = 1$ we have $(\Gamma(a_w)v)_k > 0$ for some index k, when all of the components of the vector v of compatible dimension are unity. Thus, $\max_i \sum_{j=1}^n a_{ij}(w_i,w_j)$, which we view as a function of the matrix u whose elements are the (w_i,w_j), is positive for each w in L^n with unit norm. Since $\max_i \sum_{j=1}^n a_{ij}(w_i,w_j)$ is obviously a continuous function of u, and, by Lemma 2, C is compact, there is a $\sigma > 0$ such that

$$\min_{u \in C} \max_i \sum_{j=1}^n a_{ij}(w_i,w_j) = \sigma. \tag{17}$$

Therefore, for each $w \in L^n$ with $|w| = 1$ there is an index k such that

$$\sum_{j=1}^n a_{kj}(w_k,w_j) \geq \sigma |w|^2 \geq \sigma |w_k|^2,$$

from which we see that for each $w \in L^n$ with $|w| \neq 0$, there is a k such that $|w_k| \neq 0$ and

$$\sum_{j=1}^n a_{kj}(w_k,w_j) \geq \sigma |w_k|^2.$$

Thus, $a \in U$.

To show that $a \in U_0$ when $a_w \in P_0$ (and hence $\Gamma(a_w) \in P_0$) for each $w \in L^n$ with $|w| > 0$, we observe that then, by Lemma 3, for each $w \in L^n$ with $|w| \neq 0$ we have $(\Gamma(a_w)v)_k \geq 0$ for some k when the components of v are all unity. Therefore, for each $w \in L^n$ with $|w| \neq 0$, there is a k such that $|w_k| \neq 0$ and

$$\sum_{j=1}^n a_{kj}(w_k,w_j) \geq 0,$$

which means that $a \in U_0$.

Suppose now that for some $w \in L^n$ with $|w| > 0$ we have $\Gamma(a_w) \notin P(a_w \notin P_0)$. Then, by Lemma 3, there is a nonzero vector v such that $v_k(\Gamma(a_w)v)_k \leq 0$ ($v_k(\Gamma(a_w)v)_k < 0$) for every k such that $v_k \neq 0$. Thus, by multiplying each w_i for which $|w_i| \neq 0$ by the appropriate component of v, it is a simple matter to construct a $z \in L^n$ for which $|z| \neq 0$ and $\Sigma_j a_{ij}(z_i,z_j) \leq 0$ for all i ($\Sigma_j a_{ij}(z_i,z_j) < 0$ for all i such that $|z_i| \neq 0$). This completes the proof of the theorem.

Corollary 2: If $n \geq 3$, $U(U_0)$ is a proper subset of the matrices of order n in $P(P_0)$.

* Of some peripheral interest is the fact that it is not necessary to assume that L is complete.

Proof: To see that $U(U_0)$ is a subset of $P(P_0)$, let $a \in U(U_0)$, let e be any element of L such that $|e| = 1$, and let w be the element of L^n defined by $w_i = e$ for all i. Thus $a_w = a$, and, by Theorem 3, $a \in P(P_0)$.

In order to show that for $n \geq 3$ there is a matrix of order n in $P(P_0)$ that is not contained in $U(U_0)$, observe that it is sufficient to consider the $n = 3$ case, and let $a^{(+)}$ and $a^{(0)}$ be defined by

$$a^{(+)} = \begin{bmatrix} 1.1 & 1 & -10 \\ 1 & 1.1 & 1 \\ 1 & 1 & 1.1 \end{bmatrix},$$

$$a^{(0)} = \begin{bmatrix} 1 & 1 & -10 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

We have $a^{(+)} \in P$ and $a^{(0)} \in P_0$. Let $w \in L^3$ be given by

$$w = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} e_1 + \begin{pmatrix} 0 \\ 1 \\ 10 \end{pmatrix} e_2,$$

in which e_1 and e_2 are orthogonal elements of L with unit norm. It is a simple matter to verify that

$$a_w^{(+)} = \begin{bmatrix} 1.1 & 1 & -10 \\ 1 & 2.2 & 11 \\ 1 & 11 & 111.1 \end{bmatrix},$$

$$a_w^{(0)} = \begin{bmatrix} 1 & 1 & -10 \\ 1 & 2 & 11 \\ 1 & 11 & 101 \end{bmatrix}$$

and that we have $\det|a_w^{(+)}| < 0$ and $\det|a_w^{(0)}| < 0$. By Theorem 3, this shows that $a^{(+)}(a^{(0)}) \notin U(U_0)$, which completes the proof.

Corollary 3: If $n = 2$, $U(U_0) = P(P_0)$ restricted to 2×2 matrices.

This follows directly from Theorem 3.*

4.1 Definitions

Let N denote the set of real symmetric nonnegative definite matrices m of order n with $m_{ii} > 0$ for some i. For each $m \in N$, let a_m denote the $n \times n$ matrix whose i,jth element is $a_{ij}m_{ij}$ for all i and j,† and let $\Lambda(a_m)$ denote the matrix obtained from a_m by deleting row i and column i for each $i \in \{j: m_{jj} = 0\}$.

Corollary 4: We have $a \in U(U_0)$ if and only if $\Lambda(a_m) \in P(a_m \in P_0)$ for each $m \in N$.

Proof: The proof of Lemma 2‡ shows that a real matrix m of order n belongs to N if and only if there is a w in L^n such that $|w| \neq 0$ and $m_{ij} = (w_i, w_j)$ for all i and j. Thus, Corollary 4 follows from Theorem 3.

4.2 Introduction to Corollary 5

In order to present our next corollary, we need the following additional definitions: Let $S(m)$ denote the set of all matrices obtainable from a given real $n \times n$ matrix m by replacing each *off-diagonal* element m_{ij} of m with $r_{ij}m_{ij}$, where the r_{ij} are real numbers that satisfy $r_{ij} = r_{ji}$ and $|r_{ij}| \leq 1$. Let R denote $\{m \in P: S(m) \subset P\}$, and, similarly, let $R_0 = \{m \in P_0: S(m) \subset P_0\}$.

When $n = 2$ and P and P_0 are restricted to 2×2 matrices, we have $R = P$ and $R_0 = P_0$. On the other hand, if we let $a^{(+)}(\lambda)$ and $a^{(0)}(\lambda)$, respectively, denote the matrices obtained from $a^{(+)}$ and $a^{(0)}$ of the proof of Corollary 2 by multiplying the (1,3) and (3,1) elements by a scalar variable λ, then $a^{(+)}(1) \in P$ and $a^{(0)}(1) \in P_0$, but $a^{(+)}(0) \notin P$, and, similarly, $a^{(0)}(0) \notin P_0$. This shows that $R(R_0)$ is a proper subset of the $n \times n$ matrices in $P(P_0)$ when $n \geq 3$.§ Two familiar classes of matrices contained, for example, in R_0 are the set of row-sum dominant matrices and the set of column-sum dominant matrices.

Corollary 5: We have $a \in U(U_0)$ if either

(i) $a \in R(R_0)$.

(ii) There are diagonal matrices d_1 and d_2 of order n with positive diagonal elements such that d_1ad_2 is positive definite (nonnegative definite).¶

Proof: Suppose first that $a \in R(R_0)$. Let w be any element of L^n such that $|w| \neq 0$, let $c = \text{diag}(c_1, c_2, \ldots, c_n)$ in which for all i, $c_i = 0$ if $|w_i|$

* The proof of Corollary 2 shows that $U \subset P$ and $U_0 \subset P_0$ for $n \geq 2$.

† In other words, let a_m denote the "Schur product" of a and m.

‡ Lemma 2 is used in the proof of Theorem 3.

§ It will become clear that this proposition also follows from Corollary 2 and Corollary 5.

¶ As usual, we say that a real square matrix m is positive definite (nonnegative definite) if and only if the symmetric part of m is the matrix of a positive definite (nonnegative definite) quadratic form.

$= 0$ and $c_i = |w_i|^{-1}$ if $|w_i| > 0$, and let $\Gamma(ca_wc)$ denote the matrix obtained from (ca_wc) by deleting the rows and columns corresponding to the indices i for which $c_i = 0$. In view of the fact that $|(w_i,w_j)| \leq |w_i| \cdot |w_j|$ for each i and j, we see that $\Gamma(ca_wc) \in P(P_0)$ and hence* that $\Gamma(a_w) \in P(P_0)$. By Theorem 3, $a \in U(U_0)$.

At this point, we need the following lemma.

Lemma 4†: *If $p = \{p_{ij}\}$ and $q = \{q_{ij}\}$ are real square matrices of the same order, with p positive definite and q symmetric, nonnegative definite, and such that $q_{ii} > 0$ for all i, then $r = \{p_{ij}q_{ij}\}$ is positive definite.*

Proof of Lemma 4: Let p and q be as indicated, and let k denote the order of p. The proof of Lemma 2 shows that L_2 contains k functions $f_1, f_2, \ldots, f_k$ such that

$$q_{ij} = \int_0^\infty f_i(t)f_j(t)dt \quad \text{for all } i \text{ and } j.$$

With v any real nonzero k-vector and with λ the smallest eigenvalue of the symmetric part of p, we have

$$\begin{aligned} v^{tr}rv &= \sum_{i,j} v_iv_jp_{ij} \int_0^\infty f_i(t)f_j(t)dt \\ &= \int_0^\infty \sum_{i,j} p_{ij}v_if_i(t)v_jf_j(t)dt \\ &\geq \int_0^\infty \lambda \sum_i (v_if_i(t))^2dt \\ &> 0, \end{aligned}$$

which shows that r is positive definite.

To complete the proof of the corollary, suppose that d_1ad_2 is positive definite, with d_1 and d_2 as described, and let $m \in N$. By Lemma 4, $\Lambda\{d_1a_md_2\}$ (i.e., $\Lambda\{a_m\}$ with a replaced with d_1ad_2) is positive definite and hence it belongs to P. Therefore, $\Lambda\{a_m\} \in P$, and, by Corollary 4, $a \in U$.

The proof for the case in which d_1ad_2 is nonnegative definite is essentially the same, and is omitted.

4.3 Comments Regarding Corollary 5

In light of the fact that $R_0 = P_0$ restricted to 2×2 matrices when $n = 2$, the following special result is a direct consequence of Corollaries 5 and 1, and the content of the proof of Proposition 1.

Proposition 2: Let $n = 2$ and $A \in M$. Let H.2 and H.3 (as well as A.1 and

---

* Here and in another part of the proof, we use the easily proved result that a real square matrix m belongs to $P(P_0)$ if and only if $d_1md_2 \in P(P_0)$ for every pair of compatible diagonal matrices d_1 and d_2 with positive diagonal elements.

† A proof that the conclusion of Lemma 4 holds when q is positive definite is given in Ref. 11.

A.3) be satisfied. Then (4) is L-stable for every B if and only if $a \in P_0$.

An example of a matrix a that is nonnegative definite and such that $a \notin R_0$ is given by

$$a = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2^{1/2} & 2^{1/2} \\ 1 & 2^{1/2} & 2^{1/2} \end{bmatrix},$$

since

$$\det \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2^{1/2} & 2^{1/2} \\ 0 & 2^{1/2} & 2^{1/2} \end{bmatrix} < 0.$$

Similarly, a very simple example of an $a \in R_0$ such that $d_1 a d_2$ is nonnegative definite for no suitable d_1 and d_2 is

$$a = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Of some interest is the fact that $a \in U$ if a is an M-matrix (i.e., if a has positive principal minors and nonpositive off-diagonal elements); in that case there is a diagonal matrix d with positive diagonal elements such that ad is strongly row-sum dominant* and therefore ad and consequently a belong to R.

Theorem 4: Let $A \in M$. *If* $a_{ii} = 0$ *for all* i, *then* $a \in U_0$ *if and only if* $a \in R_0$.

Proof: The "if part" is a special case of Corollary 5.

Suppose that $a \in U_0$ with $a_{ii} = 0$ for all i, and suppose also that $a \notin R_0$ in which case there is an element b of $S(a)$ such that $b \notin P_0$. Let b be given by $b_{ii} = 0$ for all i, and $b_{ij} = r_{ij} a_{ij}$ with $r_{ij} = r_{ji}$ for $i \neq j$. Choose n real numbers $r_{11}, r_{22}, \ldots, r_{nn}$ so that the $n \times n$ matrix m given by $m_{ij} = r_{ij}$ for all i and j is nonnegative definite. Observe that $m \in N$. Since $a_m = b$, by Corollary 4, we have a contradiction to the supposition that $a \in U_0$. Therefore, $a \in R_0$ when $a \in U_0$ and $a_{ii} = 0$ for all i, which completes the proof of the theorem.

4.4 Comment regarding Theorem 4

We can have $a \in P_0$ with $a_{ii} = 0$ for all i, and $a \notin R_0$. For example, let

$$a = \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

* See the theorem given on page 387 of Ref. 12.

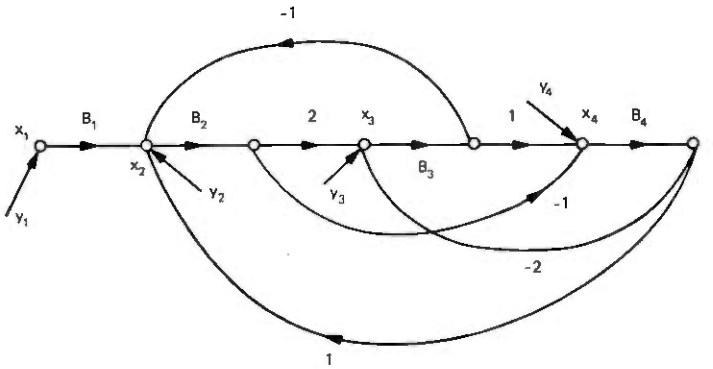


Fig. 1—Flow graph of an interconnected system.

Then $a \in P_0$, and

$$\det \begin{bmatrix} 0 & -r_{12} & r_{13} \\ r_{12} & 0 & 0 \\ 0 & r_{23} & 0 \end{bmatrix} = r_{23}r_{12}r_{13} < 0$$

for, say, $r_{23} = r_{12} = -r_{13} = 1$.

4.5 A Specific example of an L-stable interconnected system

Assume that *H.2* and *H.3* (as well as *A.1* and *A.3*) are satisfied. For the system described in flow-graph form in Fig. 1, we have

$$y_1 = x_1$$

$$y_2 = x_2 - B_1x_1 + B_3x_3 - B_4x_4$$

$$y_3 = x_3 - 2B_2x_2 + 2B_4x_4$$

$$y_4 = x_4 + B_2x_2 - B_3x_3.$$

Here $A \in M$, with

$$a = \begin{bmatrix} 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & -1 \\ 0 & -2 & 0 & 2 \\ 0 & 1 & -1 & 0 \end{bmatrix}.$$

To see that $a \in R_0$, consider the matrix

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ -r_{21} & 0 & r_{23} & -r_{24} \\ 0 & -2r_{23} & 0 & 2r_{34} \\ 0 & r_{24} & -r_{34} & 0 \end{bmatrix}.$$

We observe that its determinant vanishes and every 1×1 and 2×2 principal minor is nonnegative for all real values of the r_{ij}. It is a simple matter to verify that its principal minor of order three obtained by deleting the first row and first column vanishes for all values of the r_{ij}, and it is clear that every other principal minor of order three also vanishes for all values of the r_{ij}.

Since $a \in R_0$, by Corollary 1 and either Corollary 5 or Theorem 4, the system described in Fig. 1 is L-stable.

Another way to prove that the system in Fig. 1 is L-stable is as follows. Since $H.2$ holds, $|B_1 u| \leq c_2 |u|$ for $u \in L$. It therefore suffices to show the L-stability of the system obtained from the flow graph in Fig. 1 by deleting B_1, x_1, and y_1. That can be done with the aid of Corollary 5 by verifying that the interconnection matrix a of the modified system has the property that there is a 3×3 diagonal matrix d with positive diagonal elements such that da is nonnegative definite.

REFERENCES

1. C. A. Desoer and M. Vidyasagar, *Feedback Systems: Input–Output Properties*, New York: Academic Press, 1975.
2. J. M. Holzman, *Nonlinear System Theory*, Englewood Cliffs, N.J.: Prentice-Hall, 1970.
3. A. N. Michel and R. K. Miller, *Qualitative Analysis of Large Scale Dynamical Systems*, New York: Academic Press, 1977.
4. J. C. Willems, *The Analysis of Feedback Systems*, Cambridge, Mass.: M.I.T. Press, 1971.
5. F. N. Bailey, "The Application of Lyapunov's Second Method to Interconnected Systems," SIAM J. Contr., *3* (1966), pp. 443–462.
6. F. M. Callier, W. S. Chan, and C. A. Desoer, "Input–Output Stability Theory of Interconnected Systems Using Decomposition Techniques," IEEE Trans. Circuits and Systems, *CAS*-23 (1976), pp. 714–728.
7. M. K. Sundareshan and M. Vidyasagar, "L_2-Stability of Large-Scale Dynamical Systems: Criteria via Positive Operator Theory," IEEE Trans. on Automatic Control, *AC*-22 (1977), pp. 396–398.
8. I. W. Sandberg, "Some Results on the Theory of Physical Systems Governed by Nonlinear Functional Equations," B.S.T.J., *44*, No. 7 (September 1965), pp. 871–898.
9. M. Fiedler and V. Pták, "Some Generalizations of Positive Definiteness and Monotonicity," Numer. Math., *9*, No. 2 (1966), pp. 163–172.
10. I. W. Sandberg and A. N. Willson, Jr., "Some Theorems on Properties of DC Equations of Nonlinear Networks," B.S.T.J., *48*, No. 1 (January 1969), pp. 1–34.
11. R. Bellman, *Introduction to Matrix Analysis*, New York: McGraw Hill, 1970, p. 95.
12. M. Fiedler and V. Pták, "On Matrices with Non-Positive Off-Diagonal Elements and Positive Principal Minors," Czechoslovac Math. Journal, *12*, (1962), pp. 382–400.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 8, October 1978
Printed in U.S.A.

A Note Concerning Optical-Waveguide Modulation Transfer Functions

By. I. W. SANDBERG

(Manuscript received March 9, 1978)

A necessary and sufficient condition is given for the modulation-transfer-function of certain multimode optical fiber guides to be zero free in the closed right-half of the complex plane, and to be structurally stable with respect to that property. The condition is of interest, for example, in connection with the possibility of determining the phase of a modulation-transfer-function from its amplitude.

I. INTRODUCTION AND PRELIMINARIES

Reference 1 considers the range of validity of a Hilbert-transform approach in which the measured magnitude of the modulation-transfer-function of an optical fiber guide is used to compute the guide's impulse response.* It is argued there that a key "minimum-phase assumption" can fail to be satisfied in important cases, and a few closely related experimental and analytical results are presented.

The purpose of this note is to report on a result along the same lines as a proposition given in Ref. 1 to the effect that, for a fiber guide that can propagate a finite number of discrete modes without mode mixing, the modulation-transfer-function (more precisely, the Laplace transform version of the modulation-transfer-function) is zero-free in the closed right half of the complex plane, and that property is structurally stable in a certain sense, if and only if a certain condition is met. The theorem described in Section II is concerned with a more realistic and far more interesting case in which mode mixing is not ruled out. In particular, the result provides further detailed support for the conclusion reached in

* By "the guide's impulse response" is meant the output power of the guide excited by a unit impulse of optical power. The modulation-transfer-function $G(\omega)$ is the envelope response of the fiber guide to an incoherent optical signal sinusoidally modulated at angular frequency ω. To the extent that certain approximations hold, (Ref. 2), the impulse response is the Fourier transform of the modulation-transfer-function. The reason for considering the Hilbert transform approach is that it is often desirable to determine the impulse response of fiber guides by methods other than direct time-domain measurement, and, while $|G(\omega)|$ can be measured easily, it is at the present time difficult to accurately measure the phase of $G(\omega)$. (See Refs. 3 and 4.)

Ref. 1. (i.e., that "nonminimum-phase behavior" is likely to arise, and can arise, in important actual cases). As in Ref. 1, it is the introduction and use of the concept of structural stability of a mathematical property which, for the case considered, enables a result to be obtained that is complete and easy to interpret.*

We consider, as in Ref. 1, a class of optical fiber guides for which appropriate approximations can be made so that the modulation-transfer-function $G(\omega)$ of a guide can be written in the form

$$G(\omega) = \int_T e^{-i\omega\tau} da(\tau), \tag{1}$$

in which T denotes a closed, finite, real interval whose end points depend on the refractive indices of the core and cladding, and $a(\tau)$ is a real-valued nondecreasing function.† It is assumed throughout that $a(\tau)$ satisfies the normalization condition

$$\int_T da(\tau) = 1.$$

As mentioned previously, in Ref. 1 attention is focused on the class of fiber guides that can propagate n discrete modes without mode mixing.‡ In that case, $G(\omega)$ can be written as

$$\sum_{j=1}^{n} d_j e^{-i\omega\tau_j}, \tag{2}$$

in which each d_j is a positive number that represents the initial excitation of the jth mode, and $\tau_1 < \tau_2 < \ldots < \tau_n$.

In this note we suppose that the right side of (1) can be expressed as

$$\sum_{j=1}^{k_F} d_j e^{-i\omega\tau_j} + \int_T e^{-i\omega\tau} b(\tau) d\tau, \tag{3}$$

in which: k_F is a positive integer (the motivation for using the subscript F will become clear shortly), $d_j > 0$ for $1 \leq j \leq k_F$, $\tau_1 < \tau_2 < \ldots < \tau_{k_F}$, and $b(\tau)$ is a bounded piecewise-continuous§ nonnegative function which takes into account mode mixing. It is also assumed that τ_1 has the following property: $b(\tau) = 0$ for $\tau \in T$ with $\tau < \tau_1$ (which, of course, allows the possibility, but does not require, that τ_1 is equal to the lower endpoint

* The general problem of determining when the Hilbert-transform approach (i.e., when the so-called Kramers-Kronig transformation) is valid is of interest in many fields (see, for example, Ref. 5).

† Thus, roughly speaking, $da(\tau)$ in (1) can be replaced with $f(\tau)d\tau$ in which the function $f(\tau)$ is nonnegative and may contain impulses corresponding to discrete modes. See Refs. 2 and 3 for the relevant background material. We are assuming that material dispersion can be neglected.

‡ Typically, $n > 100$.

§ The piecewise-continuity assumption appears to be adequate for applications. For basically a somewhat more general version of the theorem stated in Section II, see Section 2.2.

of T). Since (3) is a specialization of the right side of (1), we also have $\tau_j \varepsilon T$ for each j, as well as

$$\sum_{j=1}^{k_F} d_j + \int_T b(\tau)d\tau = 1.$$

In physical terms, b is the relative-power density function associated with the nondiscrete modes, and the idealized impulse response of the guide is the inverse Fourier transform of (3). The observable impulse response of the guide (i.e., the impulse response of the guide-detector combination) is a somewhat smoothed version of the idealized response, with the smoothing provided by the detector (see Ref. 3).

The important assumption that $d_1 > 0$ and that $b(\tau) = 0$ for $\tau \varepsilon T$ with $\tau < \tau_1$ means that a discrete mode corresponding to the smallest modal delay is propagated. This assumption appears to be reasonable for at least some interesting classes of guides. For example, if a guide with a step-index profile is short enough to neglect mode conversion phenomena, then it is not unreasonable to assume that $G(\omega)$ has the form given in (2) with τ_1 the modal delay corresponding to the fundamental mode. In a real fiber, geometrical perturbations couple energy among the modes so that the distribution of modal delays changes continuously from a discrete set to a continuum as the fiber length L increases. Experimental evidence indicates that the assumption is reasonable at least if the guide is not too long.* (For a particular fiber, there is a characteristic coupling length L_c such that for $L > L_c$, it is difficult in the time domain to isolate discrete modes with appreciable energy.)

II. THE RESULT

In this section, z denotes a complex variable and $F(z)$ is defined by

$$F(z) = \sum_{j=1}^{k_F} d_j e^{-z\tau_j} + \int_T e^{-z\tau} b(\tau)d\tau \tag{4}$$

for each z. (Of course, if $G(\omega)$ denotes (3), then $G(\omega) = F(i\omega)$, and $F(z)$ is simply the Laplace transform of the generalized function whose Fourier transform is $G(\omega)$.)

In order to state our result, consider an arbitrary function $H(z)$ of the same type as $F(z)$. More explicitly, let $H(z)$ be given for all z by

$$H(z) = \sum_{j=1}^{k_H} \delta_j e^{-zt_j} + \int_T e^{-z\tau} \beta(\tau)d\tau, \tag{5}$$

in which k_H is a positive integer (not necessarily equal to k_F), and the δ_j, the t_j, and $\beta(\tau)$ satisfy the restrictions imposed on the corresponding terms in (4). Let S denote the set of all such functions $H(z)$.

* The writer is indebted to his colleague I. P. Kaminow for a helpful discussion concerning the significance of the assumption described above.

For the purpose of defining a "distance" between F and an arbitrary element H of S, let J_F, J_H, and J, respectively, denote the sets of numbers $\{1,2,\ldots,k_F\}$, $\{1,2,\ldots,k_H\}$, and $\{1,2,\ldots,\min(k_F,k_H)\}$, and, with y a real variable, let $q(y)$ denote any continuous nondecreasing function of y such that $q(0) = 0$.

Let the "distance" $\rho(F,H)$ between F and any H in S be defined by*

$$\rho(F,H) = \sum_{j \in J} |d_j - \delta_j| + \sum_{j \in (J_F - J)} d_j + \sum_{j \in (J_H - J)} \delta_j + q\left(\max_{j \in J} |\tau_j - t_j|\right) + \max_{u,v \in T} \left| \int_u^v [b(\tau) - \beta(\tau)] d\tau \right|. \quad (6)$$

Each term on the right side of (6) has a direct interpretation. In particular,

$$\sum_{j \in (J_F - J)} d_j + \sum_{j \in (J_H - J)} \delta_j,$$

in which at most one sum is nonzero, reflects the extent to which terms in one of the two finite sums in (4) and (5) do not have counterparts in the other. Also,

$$\int_u^v [b(\tau) - \beta(\tau)] d\tau$$

is an integral of the difference of two power-density functions, and, roughly speaking, if

$$\max_{u,v \in T} \left| \int_u^v [b(\tau) - \beta(\tau)] d\tau \right|$$

is sufficiently small, then, for practical purposes, the functions b and β are indistinguishable in the sense that the observable impulse response of the guide is essentially unchanged if b is replaced with β. (The portion of the idealized impulse response that does not contain impulses is g defined by $g(t) = b(t)$ for $t \in T$ and $g(t) = 0$ for $t \notin T$. If, for example, the smoothing introduced by the detector is modeled by a filter with impulse response r given by $r(t) = \rho^{-1}$ for $t \in [0, \rho]$ and $r(t) = 0$ otherwise, in which ρ is a small positive constant, then the observable version of $g(t)$ is $\rho^{-1} \int_{t-\rho}^{t} g(\tau) d\tau$ for each t. Similarly, if instead $r(t) = 0$ for $t < 0$, $r(0)$ is finite, and the derivative of r is absolutely integrable on $[0,\infty)$, then an integration by parts shows that the observable version of g is essentially unchanged when b is replaced with a sufficiently nearby β in the sense indicated above.)

Our result, the theorem given below, provides an answer to the following question: When is it true that $F(z) \neq 0$ for $\mathrm{Re}(z) \geq 0$ *and* that

* We adopt the convention that a sum over the empty set is zero.

property of F is *structurally stable* in the sense that there is a positive constant ϵ (which can be thought of as a "tolerance") such that $H(z) \neq 0$ for $\mathrm{Re}(z) \geq 0$ for every H in S such that $\rho(F,H) \leq \epsilon$.

Theorem: We have $F(z) \neq 0$ for $Re(z) \geq 0$ with that property of F structurally stable, if and only if

$$d_1 > \sum_{\substack{j \in J_F \\ j \neq 1}} d_j + \int_T b(\tau) d\tau.$$

Note that the theorem* does not rule out the possibility that the condition given in the theorem is violated and $F(z)$ is zero-free in the closed right-half plane. (In fact, an example given in Ref. 1 shows that the possibility can occur. Essentially the same example can be used to show also that if the condition is violated, then it need not be the case that *all* functions in S "sufficiently close" to but different from F have a zero in $\mathrm{Re}(z) \geq 0$.) On the other hand, at present it appears that there are complex, and for practical purposes impossible-to-specify, additional relationships among the τ_j, the d_j, and b that, in particular take into account geometrical perturbations along the length of a real guide. The theorem shows that, when additional information is unavailable, it is not possible to prove that $F(z)$ is zero-free in the closed right-half plane whenever

$$d_1 < \sum_{\substack{j \in J_F \\ j \neq 1}} d_j + \int_T b(\tau) d\tau$$

(which, in view of the normalization condition concerning a, is equivalent to the statement that the discrete mode with the smallest delay has relative power less than $\frac{1}{2}$) and, in the sense indicated above, the τ_j, the d_j, and b are known only to within some tolerance, no matter how small the tolerance is.

A proof of the theorem is given in the next section.

* As indicated earlier, one application of the theorem is that it provides further detailed support for the material reported on in Ref. 1. For the benefit of the reader who has not read Ref. 1, we mention that a much simplified version of essentially the proof given in Section 2.1 can be used to show that if $k_F \geq 2$, if $b(\tau)$ and $\beta(\tau)$ in (4) and (5), respectively, are each replaced by the zero function, if S is further restricted so that $k_H = k_F$ and $\delta_j = d_j$ for $j = 1,2,\ldots, k_F$ for all $H \in S$, and if $\rho(F,H)$ is instead $q(\max_{j \in J} |\tau_j - t_j|)$ [i.e., just the fourth term in the sum on the right side of (6)], then: $F(z) \neq 0$ for $\mathrm{Re}(z) \geq 0$ and there is an $\epsilon > 0$ such that $H(z) \neq 0$ for $\mathrm{Re}(z) \geq 0$ for every H in the corresponding S with $\rho(F,H) \leq \epsilon$, if and only if $d_1 > \sum_{j=2}^{k_F} d_j$. This result is basically a slight generalization of the comparable proposition in Ref. 1.

2.1 Proof of the theorem

For the reader's convenience, we first repeat some of the material described above. We have $T = [T_1, T_2]$ in which T_1 and T_2 are real numbers such that $T_1 < T_2$, and S denotes the set of all functions of the complex variable z of the form (5) where k_H is a positive integer, $\delta_j > 0$ for $1 \leq j \leq k_H$, $T_1 \leq t_1 < t_2 < \ldots < t_{k_H} \leq T_2$, $\beta(\tau)$ is a nonnegative bounded piecewise-continuous* function defined on T such that $\beta(\tau) = 0$ for $\tau \in T$ with $\tau < \tau_1$, and

$$\sum_{j=1}^{k_H} \delta_j + \int_T \beta(\tau) d\tau = 1.$$

The "distance" $\rho(F,H)$ between any $H \in S$ and the particular element F of S given by (4), is defined by (6).

Proof of the "If" Part: Suppose that

$$d_1 > \sum_{\substack{j \in J_F \\ j \neq 1}} d_j + \int_T b(\tau) d\tau.$$

With

$$r = d_1 - \sum_{\substack{j \in J_F \\ j \neq 1}} d_j - \int_T b(\tau) d\tau,$$

let ϵ satisfy $0 < \epsilon < (1/4)r$. For each $H \in S$ with $\rho(F,H) \leq \epsilon$, we see that $|d_1 - \delta_1| \leq \epsilon$,

$$\sum_{\substack{j \in J \\ j \neq 1}} |d_j - \delta_j| \leq \epsilon, \quad \sum_{j \in (J_F - J)} d_j \leq \epsilon, \quad \sum_{j \in (J_H - J)} \delta_j \leq \epsilon,$$

as well as

$$\left| \int_T b(\tau) d\tau - \int_T \beta(\tau) d\tau \right| \leq \epsilon,$$

and therefore

$$\delta_1 > \sum_{\substack{j \in J_H \\ j \neq 1}} \delta_j + \int_T \beta(\tau) d\tau.$$

Thus, for $\mathrm{Re}(z) \geq 0$ and $H \in S$ with $\rho(F,H) \leq \epsilon$, we have

$$|e^{zt_1} H(z)| = \left| \delta_1 + \sum_{\substack{j \in J_H \\ j \neq 1}} \delta_j e^{-z(t_j - t_1)} + \int_T e^{-z(\tau - t_1)} \beta(\tau) d\tau \right|$$

* It will become evident that the theorem holds also if "piecewise continuous" is replaced with either "Riemann integrable" or "Lebesgue integrable." In this connection, see Section 2.2.

$$\geq \delta_1 - \left| \sum_{\substack{j \in J_H \\ j \neq 1}} \delta_j e^{-z(t_j - t_1)} + \int_T e^{-z(\tau - t_1)} \beta(\tau) d\tau \right|$$

$$\geq \delta_1 - \sum_{\substack{j \in J_H \\ j \neq 1}} \delta_j - \int_T \beta(\tau) d\tau. \tag{7}$$

Since the right side of (7) is positive, it is clear that $H(z) \neq 0$, which completes the proof of the "if" part.

Proof of the "Only If" Part: Suppose now that

$$d_1 \leq \sum_{\substack{j \in J_F \\ j \neq 1}} d_j + \int_T b(\tau) d\tau,$$

and let $\epsilon > 0$ be given.

Let $k_G = \max(k_F, 2)$, let $\eta = \min((1/6)\,\epsilon, (1/2)\,d_1)$, and let $\delta_1 = d_1 - \eta$. If $k_F > 1$, let $\delta_2 = d_2 + \eta$ and $\delta_j = d_j$ for $j \in \{j : j \in J_F; j \neq 1,2\}$, and if $k_F = 1$, let $\delta_2 = \eta$ and $\tau_2 = T_2$. Then the function G defined (for all z) by

$$G(z) = \sum_{j=1}^{K_G} \delta_j e^{-z\tau_j} + \int_T e^{-z\tau} b(\tau) d\tau$$

belongs to S, and we have (if we set $H = G$):

$$\sum_{j \in J} |d_j - \delta_j| + \sum_{j \in (J_F - J)} d_j + \sum_{j \in (J_G - J)} \delta_j \leq \frac{1}{3}\epsilon \tag{8}$$

and the strict inequality

$$\delta_1 < \sum_{\substack{j \in F_G \\ j \neq 1}} \delta_j + \int_T b(\tau) d\tau. \tag{9}$$

Let δ_τ denote $\min\{(\tau_{j+1} - \tau_j) : 1 \leq j, j+1 \leq k_G\}$, and let $\Delta = \sup_{t \in T} b(\tau)$. With $B = \{\tau \in T : b(\tau) > 0\}$, let s_1 and s_2 denote $\inf B$ and $\sup B$, respectively, when B has nonzero measure, and T_1 and T_2, respectively, otherwise.

Choose any $\delta_\epsilon > 0$ such that $q(\delta_\epsilon) \leq (1/3)\,\epsilon$, and let δ be any positive number such that

$$\delta < \min\left(\frac{1}{2}\delta_\epsilon, \frac{1}{4}\delta_\tau, \frac{1}{3}\epsilon, \epsilon(18\Delta)^{-1}, \frac{1}{4}(s_2 - s_1)\right). \tag{10}$$

Let $\omega = \pi\delta^{-1}$, and let K_δ denote the set of numbers of the form $\tau_1 + k\delta$, with k an odd positive integer. Clearly, $\exp[-i\omega(t - \tau_1)] = -1$ for $t \in K_\delta$.

Choose $t_1 = \tau_1$, and (using $k_G \geq 2$ and $2\delta < (1/2)\,\delta_\tau$) for each $j = 2,3,\ldots,k_G$ choose a t_j in $K_\delta \cap [\tau_1,T_2]$ such that $|\tau_j - t_j| \leq 2\delta$. Since $2\delta < (1/2)\,\delta_\tau$ and $2\delta < \delta_\epsilon$, we have $t_1 < t_2 < \ldots < t_{k_G}$ and $q(\max_{j \in J}|\tau_j - t_j|) \leq (1/3)\epsilon$.

We see that the "distance" between F and the element E of S given by

$$E(z) = \sum_{j=1}^{k_G} \delta_j e^{-zt_j} + \int_T e^{-z\tau} b(\tau)\,d\tau$$

is at most $(2/3)\,\epsilon$. It therefore suffices to show that there is an H in S defined by

$$H(z) = \sum_{j=1}^{k_G} \delta_j e^{-zt_j} + \int_T e^{-z\tau} \beta(\tau)\,d\tau$$

with

$$\max_{u,v \in T} \left| \int_u^v [b(\tau) - \beta(\tau)]\,d\tau \right| \leq \frac{1}{3}\epsilon$$

such that $H(z) = 0$ for some z with $\mathrm{Re}(z) > 0$.

Let $L = K_\delta \cap (s_1,s_2)$. Since $\delta < 1/4\,(s_2 - s_1)$,† L contains at least two points. Let the points in L be $p_1,p_2,\ldots,p_n$, ordered so that $p_1 < p_2 < \ldots < p_n$. Let σ be a positive number such that $\sigma < \delta$, $p_1 - s_1 > \sigma$, and $s_2 - p_n > \sigma$. With $I(u,v)$ denoting $\int_u^v b(\tau)\,d\tau$ for $u,v \in [s_1,s_2]$, let $\beta_\sigma(\tau)$ be defined for $\tau \in T$ by $\beta_\sigma(\tau) = f(t - p_1)I(s_1,p_1 + \delta) + f(t - p_2)I(p_2 - \delta, p_2 + \delta) + \ldots + f(t - p_n)I(p_n - \delta,s_2)$, in which $f(t) = (2\sigma)^{-1}$ for $|t| \leq \sigma$ and $f(t) = 0$ otherwise. Since

$$\int_T \beta_\sigma(\tau)\,d\tau = I(s_1,p_1 + \delta) + I(p_2 - \delta,p_2 + \delta) + \ldots + I(p_n - \delta,s_2) = \int_T b(\tau)\,d\tau, \quad (11)$$

we see that the function H_σ defined by

$$H_\sigma(z) = \sum_{j=1}^{k_G} \delta_j e^{-zt_j} + \int_T \beta_\sigma(\tau) e^{-z\tau}\,d\tau$$

belongs to S.

Using $I(s_1,p_1 + \delta) \leq 3\delta\Delta$,‡ $I(p_j - \delta,p_j + \delta) \leq 2\delta\Delta$ for $j = 2,\ldots,(n - 1)$, $I(p_n - \delta,s_2) \leq 3\delta\Delta$,

$$\int_{T_1}^t b(\tau)\,d\tau = \int_{T_1}^t \beta_\sigma(\tau)\,d\tau$$

for $t = s_1,p_1 + \delta,p_2 + \delta,\ldots,p_{n-1} + \delta,s_2$, and the fact that $b(\tau)$ and $\beta_\sigma(\tau)$ are nonnegative, we have

$$\left| \int_{T_1}^t b(\tau)\,d\tau - \int_{T_1}^t \beta_\sigma(\tau)\,d\tau \right| \leq 3\delta\Delta$$

† See (10).
‡ Notice that $p_1 - s_1 \leq 2\delta$.

for all $t \in T$. It follows at once that for $u,v \in T$,

$$\left| \int_u^v [b(\tau) - \beta_\sigma(\tau)]d\tau \right| \leq 6\delta\Delta$$
$$< \frac{1}{3}\epsilon.^\dagger$$

Therefore $\rho(F,H_\sigma) \leq \epsilon$, uniformly for σ as described.

Let $P(z)$ be defined for all z by

$$P(z) = \sum_{j=1}^{k_G} \delta_j e^{-zt_j} + e^{-zp_1}I(s_1,p_1 + \delta)$$
$$+ e^{-zp_2}I(p_2 - \delta,p_2 + \delta) + \ldots + e^{-zp_n}I(p_n - \delta,s_2).$$

Let α be a real variable. Since $t_j \in K_\delta$ for $j = 2,3,\ldots,k_G$, and $p_j \in K_\delta$ for $j = 1,2,\ldots,n$, and (9) and (11) hold, $P(\alpha + i\omega)\exp[(\alpha + i\omega)t_1]$[‡] is real and negative when $\alpha = 0$. On the other hand, $P(\alpha + i\omega)\exp[(\alpha + i\omega)t_1]$ is positive for all sufficiently large α. Thus, $P(z_1) = 0$ for some z_1 with $\mathrm{Re}(z_1) > 0$.

The function $P(z)$ is analytic in z throughout the complex plane. Since it is not identically zero and is analytic at $z = z_1$, its zero at $z = z_1$ is isolated. Therefore, there exists a constant $r > 0$ such that $r < \mathrm{Re}(z_1)$ and, with $\Gamma = \{z:|z - z_1| = r\}$, $P(z) \neq 0$ for $z \in \Gamma$. It follows that $\min\{|P(z)|:z \in \Gamma\}$ is positive.

Using the fact that

$$(2\sigma)^{-1}\int_{t-\sigma}^{t+\sigma} e^{-z\tau}d\tau = e^{-zt}w(\sigma z),$$

in which $w(\sigma z) = (2\sigma z)^{-1}(e^{z\sigma} - e^{-z\sigma})$, we have

$$H_\sigma(z) = P(z) + [W(\sigma z) - 1]y(z), \tag{12}$$

where

$$y(z) = e^{-zp_1}I(s_1,p_1 + \delta) + e^{-zp_2}I(p_2 - \delta,p_2 + \delta)$$
$$+ \ldots + e^{-zp_n}I(p_n - \delta,s_2).$$

The function $|y(z)|$ is bounded on Γ, and $\max\{|w(\sigma z) - 1|:z \in \Gamma\}$ (and hence $\max\{|[w(\sigma z) - 1]y(z)|:z \in \Gamma\}$) can be made arbitrarily small by choosing σ to be sufficiently small. By Rouché's theorem, for sufficiently small σ, $H_\sigma(z)$ has a zero inside Γ (and hence in $\mathrm{Re}(z) > 0$). This completes the proof.

---

† See (10).

‡ Recall that $\omega = \pi\delta^{-1}$.

2.2 Comment

It is apparent that we have proved a somewhat stronger result than the theorem stated. Suppose that the definition of S is changed to the extent that $\beta(\tau)$ need not be piecewise continuous, but merely Lebesgue integrable. Then it it is clear that the "if" part of the theorem remains true. More importantly, the proof shows that if

$$d_1 < \sum_{\substack{j \in J_F \\ j \neq 1}} d_j + \int_T b(\tau) d\tau,$$

then, for any $\epsilon > 0$, there is an H in S of the form (5) with the-following properties: $\rho(F,H) \leq \epsilon$, $H(z)$ has a zero in the open right-half of the complex plane, $t_1 = \tau_1, k_H = k_F, \delta_j = d_j$ for $j = 1,2,\ldots,k_F$, $\beta(\tau)$ is piecewise constant, and the smallest closed real interval containing the support of $b(\tau)$ (which might possibly be the "empty interval") also contains the support of $\beta(\tau)$.

REFERENCES

1. I. W. Sandberg, I. P. Kaminow, L. G. Cohen, and W. L. Mammel, "On the Phase of the Modulation Transfer Function of a Multimode Optical-Fiber Guide," B.S.T.J., *57*, No. 1 (January 1978), pp. 99–110.
2. S. D. Personick, "Baseband Linearity and Equalization in Fiber Optic Digital Communication Systems," B.S.T.J., *52*, No. 7 (September 1973), pp. 1175–1194.
3. L. G. Cohen, H. W. Astle, and I. P. Kaminow, "Wavelength Dependence of Frequency-Response Measurements in Multimode Optical Fibers," B.S.T.J., *55*, No. 10 (December 1976), pp. 1509–1524.
4. L. G. Cohen, H. W. Astle, and I. P. Kaminow, "Frequency Domain Measurements of Dispersion in Multimode Optical Fibers," Appl. Phys. Lett., *30* (1977), pp. 17–19.
5. R. H. Young, "Validity of the Kramers-Kronig Transformation used in Reflection Spectroscopy," J. Opt. Soc. Am., *67* (April 1977), pp. 520–523.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 8, October 1978
Printed in U.S.A.

Some Extensions of the Ordering Techniques for Compression of Two-Level Facsimile Pictures

By F. W. MOUNTS, A. N. NETRAVALI, and K. A. WALSH

(Manuscript received September 2, 1977)

We present extensions of our earlier published ordering techniques for efficient coding of two-level (black and white) facsimile pictures. Ordering techniques use the two-dimensional correlation present in spatially close picture elements to change the relative order of transmission of elements in a scan line so as to increase the average length of the runs of consecutive black or white elements in the ordered line, making the data more amenable to one-dimensional run-length coding. The extensions that we consider allow us to use different run-length codes to match the statistics of different parts of the ordered data, and to drop certain runs from transmission. Computer simulations using the eight standard CCITT pictures, which have a resolution of approximately 200 dots/inch, indicate that these extensions can result in transmission bit rates which are about 11 to 21 percent lower than the ordering schemes described in our earlier work. The entropies vary between 0.021 and 0.125 bits/pel for the eight pictures.

I. INTRODUCTION

Coding of two-tone (black and white) facsimile pictures has gained considerable importance in the past few years, as is evidenced by a large number of papers as well as by a variety of facsimile communication systems. More and more sophisticated coding algorithms are being used which depend upon the two-dimensional spatial correlation present in picture data. This trend is understandable when one realizes that the cost of digital circuits and memories is decreasing faster than the cost of transmission.

This paper presents some extensions of our ordering schemes[1,2] for efficient coding of facsimile pictures. In the basic ordering scheme we make a prediction of the present element using the surrounding previously transmitted picture elements and classify it as "good" or "bad," depending upon the probability of the prediction being in error, condi-

tioned on the specific values of the surrounding elements. We then change the relative order of the prediction errors corresponding to picture elements along a scan line using the "goodness" of the prediction in such a way as to increase the average run-length of the black and/or white elements and then transmit the run-lengths.

This paper has several objectives. First, we give the entropy results using our earlier ordering schemes on the CCITT (International Telegraph and Telephone Consultative Committee) images. This will allow a comparison with the many coding algorithms proposed by other workers since the CCITT images are widely available. This was not possible from the results presented in our earlier paper where we had used locally generated picture material. The second objective is to present certain extensions of the ordering schemes and give results of computer simulations. The following extensions are presented: (*i*) Since good and bad regions of the ordered prediction errors have different statistics, two sets of run-length codes can be used. It is not necessary to specify the location of the boundary between the good and bad regions to the receiver. (*ii*) Runs across the good-bad region boundary can be bridged wherever advantageous, even if the color of the element changes across the boundary. (*iii*) A specified run in each line of data can be omitted from transmission since the number of elements in a line is fixed. The length of the omitted run can be derived at the receiver if a line sync code is transmitted at the end of each line.

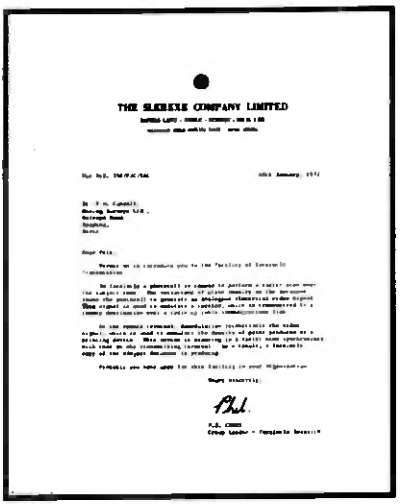
Computer simulations indicate that entropies ranging between 0.021 and 0.125 bits/pel for the eight CCITT pictures are possible using these extensions. This represents a 11- to 21-percent decrease over the ordering techniques of our earlier paper.[1]

II. CODING ALGORITHMS

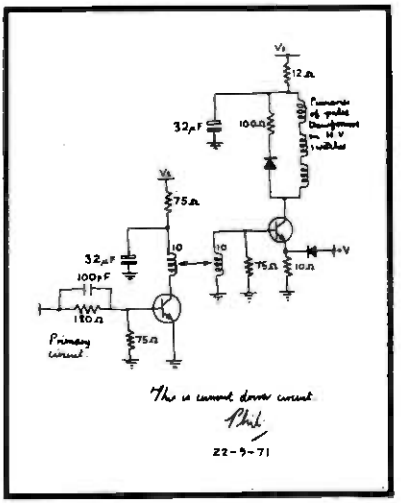
In this section, we describe our coding algorithms in detail and present results of the computer simulations. The pictures used for simulations are the eight CCITT pictures which have a resolution of approximately 200 dots/inch. Each picture consists of 2128 lines with 1728 picture elements (pels) in each line. Copies of these pictures are shown in Figs. 1a through 1h. As a measure of performance, we used the sample first-order entropy of run-length statistics. We computed the average black and white run-lengths and the entropy of black and white runs using, for example, the formula

$$E_w = -\sum_i \frac{n_i}{N} \cdot \log_2 \frac{n_i}{N}, \tag{1}$$

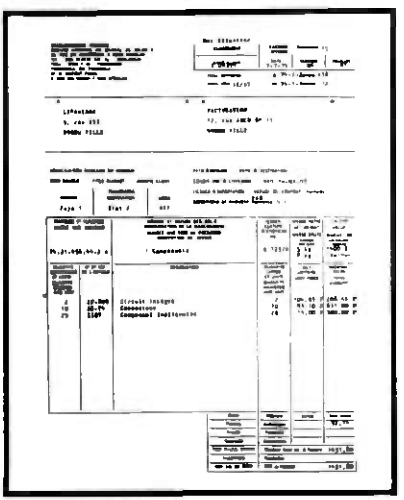
where E_w is the entropy of the white run-lengths, n_i is the number of white runs of length i, and N is the total number of white runs. Using these and eq. (2), we computed the entropy, E, in bits/pel by:



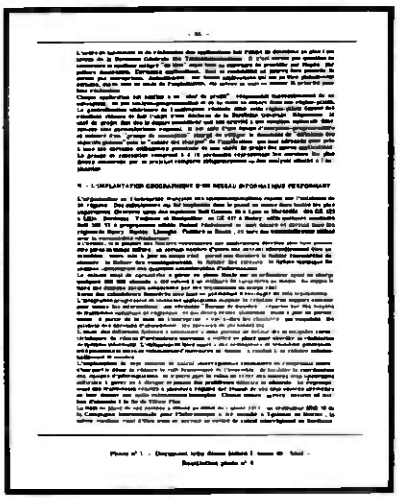
(a)



(b)



(c)

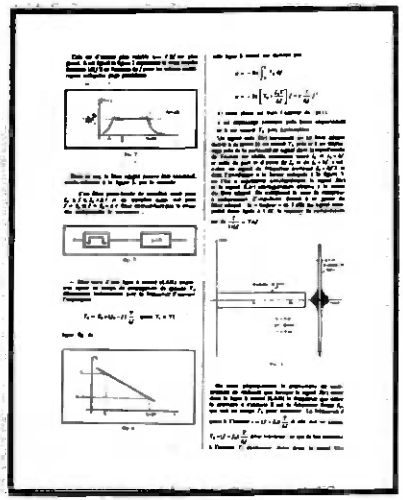


(d)

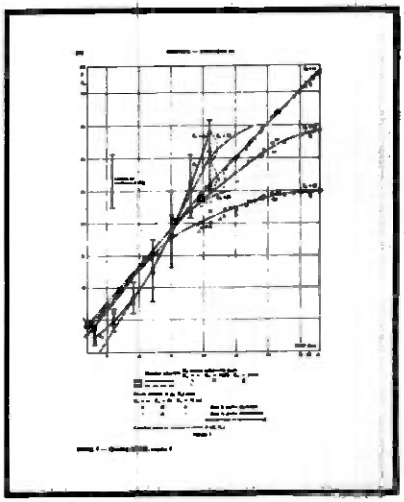
Fig. 1—The eight (a to h) CCITT pictures used for computer simulation. Each picture consists of 2128 lines with 1728 pictures elements in each line and has an approximate resolution of 200 dots/inch. (*Figs. 1e through 1h on next page*)

$$E = \frac{E_w \cdot N_w + E_b \cdot N_b}{r_w N_w + r_b N_b}, \tag{2}$$

where E_b is the entropy of the black run statistics, r_w, r_b are the average white and black run-lengths, respectively, N_w, N_b are the number of white and black runs, respectively, and E is the entropy in bits/pel. The above numbers are computed for the entire picture (1728 × 2128 pels) using, on the sides and top of the picture, a border of white elements surrounding the actual picture.

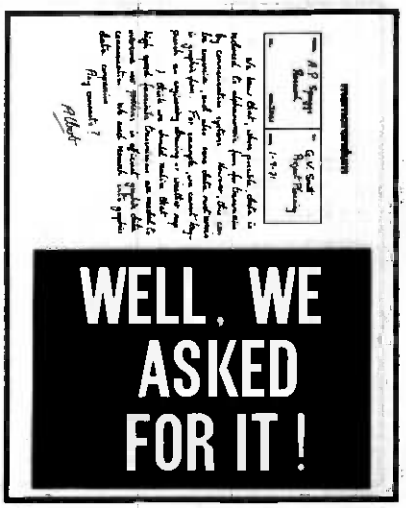


(e)



(f)

(g)



(h)

Fig. 1 (*Continued from previous page*).

2.1 Prediction algorithm

The first step in the ordering algorithm consists of making a prediction of the present picture element using the already transmitted surrounding picture elements. We define a state S_i using the four surrounding picture elements $\{X_j\}_{j=1,\ldots,4}$ as shown in Fig. 2. There are 16 states. The predictor is developed in a standard way[3–5] as the one which minimizes the probability of making an error, given that a particular state has occurred. Thus the predictor $C(S_i)$, for a given state S_i, is given by:

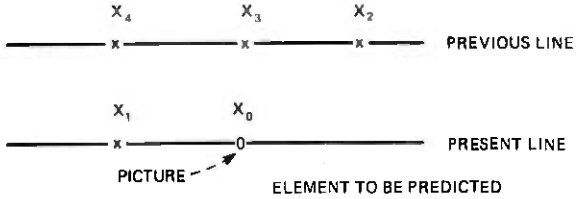


Fig. 2—Configuration for state definition.

$$C(S_i) = \text{"black," if } P(X_0 = \text{"black"} \mid S = S_i) > 0.5$$
$$= \text{"white," otherwise,}$$

where $P(\cdot|\cdot)$ is the conditional probability measured for the picture. For convenience, we represent the color of the picture elements by "1" and "0," "1" for black and "0" for white. The predictor varies from picture to picture; however, the variation is not great, as shown in our earlier paper.[1] The predictor for a typical picture [CCITT picture 2 (Fig. 1b)] is shown in Table I.

2.2 Ordering algorithms with one set of run-length codes

In this section, we give the simulation results using our earlier ordering algorithms. First, in Table II, for the purposes of comparison, we give the entropies of the run-length statistics from the raw picture data as well as from the prediction error data. As expected, the entropies of the run-lengths of the prediction errors show about 0.7 to 24 percent decrease over the entropies of the run-lengths of raw data. The decrease is smaller for the busier pictures such as the CCITT pictures 4 and 7.

Next, we simulated the ordering algorithm of Ref. 1. As explained there, this algorithm can be illustrated by considering a memory containing 1728 cells (equal to the number of elements per line). Let the cells

Table I—State-dependent prediction for CCITT picture 2 (Fig. 1b)

	State S_i				$P(X_0/S_i)$		Predicted
	X_1	X_2	X_3	X_4	$X_0 = 0$	$X_0 = 1$	Value X_0
S_0	0	0	0	0	1.000	0.000	0
S_1	1	0	0	0	0.300	0.700	1
S_2	0	1	0	0	0.777	0.223	0
S_3	1	1	0	0	0.006	0.994	1
S_4	0	0	1	0	0.822	0.178	0
S_5	1	0	1	0	0.055	0.945	1
S_6	0	1	1	0	0.323	0.677	1
S_7	1	1	1	0	0.001	0.999	1
S_8	0	0	0	1	1.000	0.000	0
S_9	1	0	0	1	0.690	0.310	0
S_{10}	0	1	0	1	0.971	0.029	0
S_{11}	1	1	0	1	0.154	0.846	1
S_{12}	0	0	1	1	0.996	0.004	0
S_{13}	1	0	1	1	0.200	0.800	1
S_{14}	0	1	1	1	0.708	0.292	0
S_{15}	1	1	1	1	0.012	0.988	1

Table II—Entropy comparisons for different coding algorithms. The entropy numbers do not include certain housekeeping bits (e.g., line sync, color of the beginning run in a line)

Coding Algorithm		Entropy (bits/pel) CCITT Image Number							
No.		1	2	3	4	5	6	7	8
1	One-dimensional run-length coding	0.0505	0.0447	0.0914	0.1652	0.0988	0.0679	0.1791	0.0870
2	Run-length coding of prediction errors	0.0466	0.0373	0.0693	0.1640	0.0795	0.0482	0.1678	0.0678
3	Run-length coding of ordered prediction errors; one set of codes, goodness threshold = 0.1.	0.0390	0.0267	0.0571	0.1396	0.0652	0.0366	0.1400	0.0463
	Run-length coding of ordered prediction errors; all "white" state = good state	0.0398	0.0305	0.0592	0.1424	0.0673	0.0392	0.1442	0.0569
4	Run-length coding of ordered prediction errors; two sets of codes, goodness threshold = 0.1, good—bad boundary run broken	0.0356	0.0247	0.0547	0.1287	0.0613	0.0351	0.1284	0.0430
5	Run-length coding of ordered prediction errors; two sets of codes, goodness threshold = 0.1, good—bad boundary bridged	0.0351	0.0233	0.0527	0.1282	0.0596	0.0335	0.1274	0.0419
6	Run-length coding of ordered prediction errors; one set of codes, goodness threshold = 0.1, first run dropped	0.0338	0.0201	0.0501	0.1320	0.0579	0.0298	0.1326	0.0390
7	Run-length coding of ordered prediction errors; two sets of codes, goodness threshold = 0.1, good—bad boundary bridged; last decodable run dropped	0.0324	0.0210	0.0506	0.1239	0.0569	0.0312	0.1250	0.0398

of this memory be numbered from 1 to 1728. We classify the states used for predictors into two categories, good or bad. Good states are those for which the probability of the prediction being in error, conditioned on that state, is less than a given threshold (defined as the goodness threshold). All the other states are bad. In the process of ordering, if the first element of the present line has a state which is classified as good, we put the prediction error corresponding to it in memory cell 1; if, on the other hand, the state is classified as bad, we put the prediction error in memory cell 1728. We continue in this manner: the prediction error for the ith element of the present line is put in the unfilled memory cell of the smallest or the largest index, depending on whether the state corresponding to the ith element is good or bad. When the memory is filled, its cells are read in numerical order and the contents are run-length encoded. It is easy to see that the present line can be uniquely reconstructed from the knowledge of the run-lengths of the ordered line, since the ordering information is known to the receiver. The efficiency of such ordering depends upon the threshold used for classifying the states into good or bad. Table II shows two examples, one in which the goodness threshold was 0.1 and the other in which only one state (corresponding to all four surrounding elements being zero) is classified as good. A goodness threshold of 0.1 appears to be acceptable among the many thresholds that we used in our simulations. Comparing entropies corresponding to the ordered and unordered prediction errors, we see that ordering reduces the entropy by about 15 to 32 percent, depending on the picture used. Also, ordering of the prediction errors brings entropies down by 15 to 47 percent of the run-length coding of raw data. It should be noted that in each of the above cases the predictor was optimized for the particular picture.

2.3 Ordering algorithms with two sets of codes

Statistics of the run-lengths in the good and bad regions of the ordered prediction errors are quite different. As an example, for CCITT picture 2 (Fig. 1b), 98.5 percent of the pels fall in the good region of which 99.9 percent are correctly predictable, whereas the bad region contains only 1.5 percent of the total elements of which 73 percent are correctly predictable. Thus, the average run-lengths in the good region are much larger than in the bad region. Such a variation in the statistics can be exploited by using two different sets of run-length codes for the good and bad regions, respectively. The algorithm* would then operate as follows: First, we put the ordered prediction errors in the memory as before; then, the contents of the memory are run-length coded with one set of codes in the good region and a different set of codes in the bad region.

* This algorithm is related to the one proposed by Preuß (Ref. 5). It is discussed here mainly for completeness and was motivated by the communication we received from him (Ref. 6).

Switching from one set of codes to the other is done at the boundary of the good-bad region even though the ordered line may not have a new run at the boundary. This process will break the run at the boundary between the good and bad region of the ordered line, whereas the ordering technique discussed in Section 2.2 continues the run (whenever possible) across the boundary of the good-bad region. This procedure is continued until all the runs from the memory are exhausted.

At the receiver, the coded run-lengths for a complete line are held in a memory. Good or bad runs are decoded from the memory as needed.

The results of computer simulations for the ordering scheme with two sets of codes are shown in Table II. These results use a goodness threshold of 0.1. Comparing the entropies from algorithms with one and two sets of codes, it is seen that with two sets of codes about 4 to 8 percent improvement is possible. This is the opposite conclusion* from that given in our earlier paper, which used a different source material.[1] For the pictures used in Ref. 1, we had found that ordering schemes with two sets of codes resulted in 10 to 18 percent higher entropies than the entropies obtainable with one set of codes. This may have been a result of the small size of the pictures used for the simulation (an array of 256× 256 picture elements).

2.4 Ordering algorithms with two sets of codes and bridging of good-bad boundary

Use of two sets of run-length codes described in the previous subsection resulted in the breaking of a run at the boundary of the good-bad region since part of the run may be in the good region and the other part may be in the bad region. To avoid breaking the run, which extends across the boundary, we code the boundary run using the run-length code of the good region or the bad region as follows: If the boundary run is first required as a bad run in the process of decoding the run-lengths at the receiver, it is coded as a bad-run; otherwise, it is coded as a good run. The method in which the receiver decodes the bridged run is similar to the one given in the next subsection. Results of such a scheme are shown in Table II. Bridging of the run across the boundary results in an improvement of about 0.39 to 6 percent over nonbridging. As would be expected, the percent improvement is smaller for busier pictures.

2.5 Ordering algorithms with dropped runs

In most facsimile communication systems a code for the line sync is sent at the end of each line of coded data. Since the number of elements in a line is fixed, this is redundant. A run can be dropped from each line

* We thank D. Preuβ for showing us data from his simulations which first demonstrated this fact.

as long as the receiver knows the position of the dropped run. In the ordered line a large benefit can be derived by dropping the first good run, since it is generally the longest. This also avoids transmission of run-length codes for lines with no prediction errors. Table II shows results of the simulation of a scheme in which the first good run from the ordered prediction errors is dropped from transmission, and the rest of the runs are transmitted by using one set of run-length codes. Dropping the run reduces the entropy to between 0.020 and 0.133 bits/pel which is a 5 to 25 percent reduction compared to the case where all the runs are sent.

It is also possible to drop a run from transmission when two sets of codes are used for the run-lengths in the good or bad regions. In this case, the first run cannot be dropped since the receiver switches between the two sets of codes depending on the past decoded data. However, the last run that the receiver needs to decode may be dropped. We have simulated a scheme in which the good-bad region boundary is bridged and the last decodable run is dropped. To explain the scheme, consider a line made up of run-lengths of ordered prediction errors as shown in Fig. 3. We use two sets of codes and start transmitting codewords corresponding to run-lengths $G_1, G_2, \cdots, B_2, B_1$ of the ordered line, appropriately switching the code in the good and bad regions. The receiver decodes these run-lengths as needed. To bridge the boundary run and drop the last decodable run, we use the following rules:

(*i*) If there are no runs in the good region, drop the last run in the bad region, i.e., B_m.

(*ii*) If there are no runs in the bad region, drop the last run in the good region, i.e., G_n.

(*iii*) If the last two runs required by the receiver in the decoding process are G_n and B_m (in either order), drop the runs G_n and B_m. This is done independently of the color of prediction errors in G_n and B_n.

(*iv*) If the last two runs required by the receiver are from the bad region and at least one good region run has occurred, then if
- (*a*) color of B_m is a "1," bridge G_n and B_m, code it using the good region code, and drop B_{m-1}.
- (*b*) color of B_m is a "0," drop B_m.

(*v*) If the last two runs required by the receiver are from the good region and at least one bad region run has occurred, then if
- (*a*) color of G_n is a "1," bridge G_n and B_m, code it using a bad region code, and drop G_{n-1}.
- (*b*) color of G_n is a "0," drop G_n.

Rules (*iv*) and (*v*) allow us to drop a run of 0s rather than a run of 1s, since runs of 0s usually have longer lengths than runs of 1s. Also, it is possible to bridge the runs at the boundary independent of the color change across the boundary of the good and bad region. Thus, the above

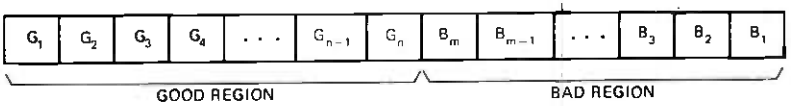


Fig. 3—Ordered run-lengths.

strategy allows dropping a run from transmission, bridging runs across the boundary (whenever it is advantageous, even if colors change), and the use of two separate sets of codes for the good and bad regions.

At the receiver, the coded run-lengths are held in memory and decoded as needed. A running total of the number of elements from decoded run-lengths is kept. If all the run-lengths have been decoded from the receiver memory and an additional run is required, this running total is subtracted from the total number of elements in a line, and the result is taken as the length of the next run. If the result is zero, then the next run is taken to be of opposite color, as usual, and decoding proceeds until the end of the line. The simulations using the above scheme decreased the entropy to between 0.021 and 0.125 bits/pel as shown in Table II. For busy images this scheme does better than the scheme which uses only one set of codes and drops the first run. However, for quieter pictures the performance is reversed.

III. DISCUSSION AND SUMMARY

We have described in this paper schemes for efficient coding of two-level (black and white) facsimile pictures. These were extensions of our earlier schemes which ordered the prediction errors before run-length coding. The most sophisticated extension presented here results in an entropy of between 0.021 and 0.125 bits/pel. Our computer simulations indicate that use of two sets of codes for good and bad regions of the ordered pictures results in about 4 to 8 percent decrease in entropy compared to using only one set of codes; whereas using two sets of codes, bridging the good-bad boundary run, and dropping the last decodable run decreases the entropy by 11 to 21 percent.

It should be mentioned that this is not a definitive coding system study. We have not considered many important factors crucial to the success of any coding system such as the run-length codes and their picture dependence and the effect of transmission errors.

IV. ACKNOWLEDGMENTS

We would like to thank Dr. Ronald Arps of IBM—Los Gatos, California, and Professor Hans G. Musmann of Technical University of Hanover, West Germany, for supplying us with the digitized CCITT images. Special thanks are due to Dr. Dieter Preuß, also of Technical

University of Hanover, West Germany, who sent us several communications describing results of his simulations which helped clear the differences between our earlier results[1] and those given in Section 2.3.

REFERENCES

1. A. N. Netravali, F. W. Mounts, and E. G. Bowen, "Ordering Techniques for Coding of Two-Tone Facsimile Pictures," B.S.T.J., *55*, No. 10 (December 1976), pp. 1539–1552.
2. A. N. Netravali, F. W. Mounts, and J. D. Beyer, "Techniques for Coding Dithered Two-Level Pictures," B.S.T.J., *56*, No. 5 (May–June 1977), pp. 809–819.
3. J. S. Wholey, "The Coding of Pictorial Data," IRE Trans. Information Theory, *IT-7* (April 1961), pp. 99–104.
4. H. Kobayashi and L. R. Bahl, "Image Data Compression by Predictive Coding," IBM J. Res. Devel., 1974, pp. 164–179.
5. D. Preuβ, "Comparison of Two-Dimensional Facsimile Coding Schemes," Int. Conf. Commun. (June 1975), pp. 7-12–7-16.
6. D. Preuβ, private communication.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 57, No. 8, October 1978
Printed in U.S.A.

Free Electron Laser

by A. HASEGAWA

(Manuscript received March 6, 1978)

An introductory guide to the basic mechanisms of the free electron laser is presented. The laser gain originates from the stimulated Raman or Compton backscattering of a pump electromagnetic field by a relativistic electron beam. The condition of optimization of the gain, the maximum operation frequency, and the optimum output power are obtained in terms of the beam parameters and the magnitude of the pump magnetic field.

I. INTRODUCTION

Recent observations of amplification of submillimeter[1] and infrared[2] electromagnetic waves using a relativistic electron beam (REB) have created interest in applying the mechanism to produce a high-power, tunable laser in the infrared to visible range as well as in speculating the possibility of constructing an X-ray laser.

This paper introduces the basic mechanism of the amplification processes and discusses the limitations in the power and frequency referring to the presently available REBs. A nonspecialist should be able to follow the contents without referring to special references.

Section II introduces Lorentz transformation of various variables between the beam and the laboratory frames, which are used in succeeding sections.

One of the important discussions presented here is the distinction between the stimulated Compton and stimulated Raman scattering. When the scattering occurs by an excitation of a single particle state, uncorrelated free-streaming motion of electrons, it is called the stimulated Compton scattering; if it occurs by an excitation of plasmon, the collective plasma oscillation of the electrons, it is called the stimulated Raman scattering. In most cases, the stimulated Compton scattering has a gain which is too small to be useful for practical purposes. Hence, the limitation in the output frequency is decided by whether or not the relativistic electron beam can be operated in the stimulated Raman regime. The beam current density and the energy spread is the decisive factor for this, as shown in Section III.

The gain calculations based on classic mechanics are presented for both processes in Sections IV and V. The classic calculation is justified when the scattered photon density is large so that the photons can be regarded as consisting of a continuous fluid. This occurs when the number of photons in a box of its wavelength (λ) cubed is much larger than unity; that is, when $\lambda^3 P/(\hbar\omega c) \gg 1$, where P is the electromagnetic power and c is the speed of light.

Some design examples using presently available REBs are shown in Section IV. MKS units are used throughout this paper. Definitions of the notations and subscripts used are listed below.

z: coordinate taken in the direction of the beam velocity.
x,y: coordinates perpendicular to the beam velocity
m: electron rest mass
p: momentum
P: power
v_0: beam velocity
v_g: group velocity
E: electric field intensity
B: magnetic flux density
c: speed of light, 3×10^8 m/s
γ: $(1 - v_0^2/c^2)^{-1/2}$ [eq. (5)]
H_0: beam energy
γ_0: H_0/mc^2 [eq. (21)]
ω_p: plasma angular frequency, frame invariant
k_0: $2\pi/\lambda_0$ (λ_0 is the periodicity of the helical winding of the pump magnetic field, Fig. 1)
ω_0: $k_0 c$
ϵ_0: space dielectric constant, 8.854×10^{-12} F/m
v_T: thermal speed in the beam frame [eq (17) and (35)]
$\Delta\gamma/\gamma$: relative energy spread of the beam in the laboratory frame
Γ: temporal gain
ω_i: incident electromagnetic wave angular frequency, which corresponds to the pump frequency in the beam frame
k_i: incident wavenumber, beam frame
ω_s: scattered electromagnetic wave angular frequency, beam frame
k_s: scattered wavenumber, beam frame
ω_l: longitudinal electrostatic wave angular frequency, beam frame
k_l: longitudinal wavenumber, beam frame
$B_\perp$: transverse pump magnetic field, laboratory frame
k_D: Debye wavenumber ω_p/v_T in the beam frame
ω_{cr}: angular frequency of transition from stimulated Raman to

stimulated Compton scattering in the laboratory frame [eq. (37)]

J_0: beam current density

v_i: amplitude of oscillating velocity of electrons due to the incident (pump) wave [eq. (24)]

Subscript L: quantities in the laboratory frame

Subscript B: quantities in the beam frame

Subscript l: longitudinal wave, beam frame

Subscript s: scattered wave, beam frame

Subscript i: incident wave, beam frame

Subscript $\perp$: component perpendicular to z.

II. LORENTZ TRANSFORMATIONS

To understand the dynamics of the REB, we must first refresh our memory of the Lorentz transformations which are relevant to our problem. If we take z axis in the direction of the beam velocity as in Fig. 1 and use subscripts L and B to represent the laboratory and the beam frame, the Lorentz transformations of the coordinate z and time t for a REB with the velocity v_0 are given by (for example, see Ref. 3):

$$z_B = \gamma(z_L - v_0 t_L) \tag{1}$$

or

$$z_L = \gamma(z_B + v_0 t_B), \tag{2}$$

and

$$t_B = \gamma\left(t_L - \frac{v_0}{c^2} z_L\right), \tag{3}$$

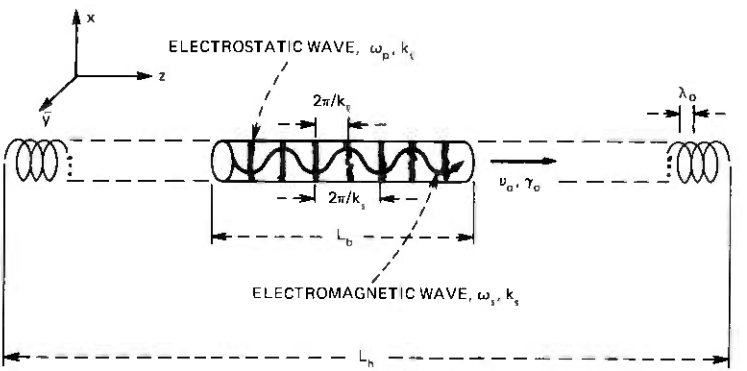


Fig. 1—Schematic diagram of a free electron laser which utilizes the helical magnetic pump field. The helical current produces a periodic magnetic field which induces longitudinal electrostatic oscillations in the beam. A nonlinear interaction between the induced longitudinal oscillation and the periodic pump field produces an electromagnetic wave which propagates in the direction of the beam. This process can be viewed, in the beam frame, as a stimulated backscattering of the pump field by the electrons in the beam. Since the scattered wave propagates at the same speed as the beam itself, the beam length, L_b, can be a size of several wavelengths in the beam frame. However, the length of the helical field, L_h, should be such that enough e-folding gain can be obtained. The minimum e-folding distance is obtained in eq. (91). L_h should therefore be much larger than L_m in this equation. Typically, L_m is on the order of 1 m.

or

$$t_L = \gamma \left(t_B + \frac{v_0}{c^2} z_B \right), \tag{4}$$

where

$$\gamma = \left(1 - \frac{v_0^2}{c_2} \right)^{-1/2} \tag{5}$$

and c is the speed of light. Similarly, the electric field intensity $\mathbf{E}$ and the magnetic flux density $\mathbf{B}$ are transformed to

$$E_{Bz} = E_{Lz}, \tag{6}$$

$$\mathbf{E}_{B\perp} = \gamma(\mathbf{E}_{L\perp} + \mathbf{v}_0 \times \mathbf{B}_L), \tag{7}$$

and

$$B_{Bz} = B_{Lz}, \tag{8}$$

$$\mathbf{B}_{B\perp} = \gamma \left(\mathbf{B}_{L\perp} - \frac{1}{c^2} \mathbf{v}_0 \times \mathbf{E}_L \right), \tag{9}$$

where subscript $\perp$ shows the component perpendicular to the beam velocity. Equation (7) indicates that a transverse magnetic field which is static but spacially periodic in the z direction with the periodicity $2\pi/k_0$ creates an oscillating electric field in the beam frame with the frequency given by $\gamma k_0 v_0$. Transformations of velocities are obtained by taking the derivatives of (2) and (4),

$$v_{Lz} = \frac{v_{Bz} + v_0}{1 + v_0 v_{Bz}/c^2}. \tag{10}$$

The beam has transverse velocity modulation due to the $\mathbf{v}_0 \times \mathbf{B}_{L\perp}$ Lorentz force. The Lorentz transformation becomes

$$\begin{aligned} v_{L\perp} &= \frac{v_{B\perp}}{\gamma(1 + v_0 v_{Bz}/c^2)} \\ &\simeq \frac{1}{\gamma} v_{B\perp}. \end{aligned} \tag{11}$$

The Lorentz transformations for frequency and the wave number are obtained by considering the phase factor $k_L z_L + \omega_L t_L$ of a wave in the laboratory frame, $\exp i(kz + \omega t)$; we take a wave propagating against the beam direction to consider the back scattering.

$$k_L z_L + \omega_L t_L = \gamma \left(k_L + \frac{v_0}{c^2} \omega_L \right) z_B + \gamma(\omega_L + k_L v_0) t_B; \tag{12}$$

hence

$$k_B = \gamma \left(k_L + \frac{v_0}{c^2} \omega_L \right), \tag{13}$$

$$\omega_B = \gamma(\omega_L + k_L v_0). \tag{14}$$

One important aspect of this result is that the frequency seen by the beam is γ times the laboratory frequency ω_L *plus* γ times the Doppler shifted laboratory frequency $k_L v_0$. An electromagnetic pump wave propagating against the beam direction (whose dispersion relation is given by $\omega = kc$) has a frequency given by $\gamma(\omega + kv_0) \simeq 2\gamma\omega$ when observed in the beam frame. Similarly, the frequency ω_s of the back-scattered light which faces little frequency shift from the incident light in the beam frame becomes $2\gamma\omega_s$ when observed in the laboratory frame. Hence, the frequency of the back-scattered light in the laboratory frame is given approximately by $4\gamma^2$ times the incident (pump) frequency in the laboratory frame.

The pump frequency can be dc when a periodic magnetic field is used. In this case, the frequency of the scattered wave is given by $2\gamma^2 k_0 v_0$, where k_0 is the wave number of the periodicity λ_0, $k_0 = 2\pi/\lambda_0$, of the magnetic field (see Fig. 1).

In addition to these quantities, we need the transformation of the plasma frequency, ω_p, the beam thermal speed v_T, the beam oscillating velocity in the transverse direction due to the pump field $v_\perp$, and the growth rate Γ.

Since the Lorentz contraction increases the density by γ and the mass also by a factor γ, the plasma frequency, ω_p $(= e^2 n/\epsilon_0 m)^{1/2}$ (where e is the electron change, n the beam density, and ϵ_0 the space dielectric constant), is frame invariant.

The thermal speed in the beam frame v_T can be expressed in terms of the energy spread of the beam in the laboratory frame as follows. From the definition of γ in (5),

$$v_0^2 = c^2\left(1 - \frac{1}{\gamma^2}\right). \tag{5'}$$

Hence the velocity spread δv_0 in the laboratory frame is expressed in terms of the spread in γ,

$$\delta v_0 = c\,\frac{\Delta\gamma}{\gamma^3}. \tag{15}$$

Now if we use the Lorentz transformation of v_z, (10),

$$\delta v_0 = \delta v_{Lz} = \frac{\Delta v_{Bz}}{\gamma^2(1 + v_0 v_{Bz}/c^2)} \simeq \frac{1}{\gamma^2}\Delta v_{Bz} = \frac{v_T}{\gamma^2}, \tag{16}$$

because $v_{Bz} = 0$. Hence from (16) the thermal speed in the beam frame is obtained:

$$v_T = c\,\frac{\Delta\gamma}{\gamma}. \tag{17}$$

Next, we obtain the oscillating transverse velocity in the beam frame. We consider the example of periodic magnetic pump. In this case, the beam kinetic energy H_0 does not change due to the presence of the pump. If we introduce γ_0 to represent the total kinetic energy of the beam,

$$H_0 = c(p_L^2 + m^2c^2)^{1/2}$$
$$\equiv mc^2\gamma_0, \tag{18}$$

where p_L is the momentum in the laboratory frame (H_0 is not frame invariant, but we delete subscript L for this quantity). The velocity components in the transverse and z directions are obtained in terms of p_L as

$$v_{L\perp} = \frac{\partial H_0}{\partial p_{L\perp}} = \frac{1}{m\gamma_0} p_{L\perp} \tag{19}$$

$$v_{Lz} = v_0 = \frac{1}{m\gamma_0} P_{Lz}. \tag{20}$$

If we substitute (19) and (20) into (17), we can obtain the relation between γ_0 and γ as defined in (5),

$$\gamma_0^2 = \gamma^2 \left(1 + \gamma_0^2 \frac{v_{L\perp}^2}{c^2}\right). \tag{21}$$

This expression shows that γ can be significantly different from γ_0 even if $v_{L\perp}^2/c^2 \ll 1$. With these preparations, we can now obtain $v_{B\perp}$ in terms of the pump magnetic field. The equation of motion of an electron in the presence of a transverse helical pump magnetic field $\mathbf{B}_\perp$ ($B_\perp \cos k_0 z$, $B_\perp \sin k_0 z$, 0) is given by

$$\frac{d\mathbf{p}_{L\perp}}{dt} = m\gamma_0 \frac{d\mathbf{v}_{L\perp}}{dt} = -e(\mathbf{v}_0 \times \mathbf{B}_\perp), \tag{22}$$

since γ_0 is constant. If we assume $v_0 \gg v_{L\perp}$, $z = v_0 t$, (22) can be immediately integrated to give

$$v_{L\perp} = \left(\frac{eB_\perp}{m\gamma_0 k_0} \cos(k_0 v_0 t), \frac{eB_\perp}{m\gamma_0 k_0} \sin(k_0 v_0 t), 0\right). \tag{23}$$

As will be seen, we need only the magnitude of the oscillating velocity in the beam frame, $|v_{B\perp}|$, which may be obtained from (23) and (11),

$$|v_{B\perp}| = \frac{\gamma eB_\perp}{\gamma_0 m k_0} (= v_i). \tag{24}$$

This gives the relation between the oscillation amplitude of the electrons in the beam frame and the pump magnetic field in the laboratory frame.

We now consider the transformation of the growth rate Γ. If a wave with slowly varying amplitude $A_B(z_B, t_B)$ grows in time and space at a

temporal growth rate Γ_B in the beam frame, A_B satisfies the following equation

$$\frac{\partial A_B}{\partial t} + v_{Bg}\frac{\partial A_B}{\partial z_B} = \Gamma_B A_B, \tag{25}$$

where v_{Bg} is the group velocity in the beam frame. If we use (2) and (4), $\partial/\partial t_B$ and $\partial/\partial z_B$ can be expressed in terms of derivatives in the laboratory frame.

$$\frac{\partial}{\partial t_B} + v_{Bg}\frac{\partial}{\partial z_B} = \gamma\left(1 + \frac{v_{Bg}v_0}{c^2}\right)\frac{\partial}{\partial t_L} + \gamma(v_0 + v_{Bg})\frac{\partial}{\partial z_L}. \tag{26}$$

If we substitute (26) into (25), we see

$$\frac{\partial A_B}{\partial t_L} + \frac{v_0 + v_{Bg}}{1 + v_{Bg}v_0/c^2}\frac{\partial A_B}{\partial z_L} = \frac{\Gamma_B}{\gamma(1 + v_{Bg}v_0/c^2)}A_B. \tag{27}$$

The amplitude in the laboratory frame is linearly proportional to A_B. Hence (27) gives the Lorentz transformation of the group velocity as well as the growth rate, i.e.,

$$v_{Lg} = \frac{v_0 + v_{Bg}}{1 + v_{Bg}v_0/c^2} \simeq \frac{1}{2}(v_0 + v_{Bg}), \tag{28}$$

$$\Gamma_L = \frac{\Gamma_B}{\gamma(1 + v_{Bg}v_0/c^2)} \simeq \frac{\Gamma_B}{2\gamma}. \tag{29}$$

III. STIMULATED COMPTON OR STIMULATED RAMAN SCATTERING?

We consider here the basic processes of the stimulated scattering *in the beam frame.* If we designate the frequency and wave number of the incident (pump) wave by ω_i and k_i and those of the scattered (amplified) wave by ω_s and k_s, the frequency and wave number of the longitudinal oscillation excited in the beam (which is a stationary electron plasma in the beam frame) are given by

$$\omega_l = \omega_i - \omega_s, \tag{30}$$

$$\mathbf{k}_l = \mathbf{k}_i - \mathbf{k}_s. \tag{31}$$

We note here that the incident and scattered waves are electromagnetic waves, hence $\omega_i/k_i \approx \omega_s/k_s \approx c$, while the longitudinal wave in the electron plasma has a phase velocity, ω_l/k_l, much smaller than the speed of light.

To consider the backscattering, which is needed to utilize the frequency up conversion as discussed in Section II, as well as to maximize the gain, we must take $\mathbf{k}_s\cdot\mathbf{k}_i = -|k_s||k_i|$. The incident wave propagates against the beam direction, hence $\mathbf{k}_i = -|k_i|\hat{z}$. Thus $|k_l| = |k_s| + |k_i|$.

Now the longitudinal mode in the electron beam has the plasma dispersion relation given by

$$1 - \frac{\omega_p^2}{k_1^2} \int_{-\infty}^{\infty} \frac{\partial f_0/\partial v}{v - (\omega + i0)/k_1} dv = 0, \tag{32}$$

where $f_0(v)$ is the velocity distribution function of the beam electrons in the beam frame and is assumed to be nonrelativistic. If we solve (32) for ω, we have

$$\omega \simeq \omega_p \qquad \text{if } k_1 \ll k_D, \tag{33}$$

$$\omega \simeq k_1 v_T[1 - i0(1)] \qquad \text{if } k_1 \gg k_D, \tag{34}$$

where

$$v_T = \left[\int_{-\infty}^{\infty} v^2 f_0 dv \right]^{1/2} \tag{35}$$

is the thermal speed of the electrons and $k_D = \omega_p/v_T$ is the Debye wave number, both in the beam frame. Equations (33) and (34) indicate that if the wave number is larger than the Debye wave number, the collective property of the plasms oscillation is lost. The large imaginary part in (34) is the consequence of the Landau damping.

Now the dispersion relation of the electromagnetic wave is given by

$$\omega^2 = c^2 k^2 + \omega_p^2. \tag{36}$$

If we use the dispersion relations for ω_i and ω_s [which satisfies (36)] and ω_1 [which satisfies (32)], the resonant conditions, Eqs. (31) and (32), can be plotted in (ω,k) diagram. For the case of backscattering, the plots are shown in Fig. 2 (for the case of $k_1 \ll k_D$) and Fig. 3 (for the case of $k_1 \gg k_D$). In these figures, the arrows show the direction in which the state with energy $\hbar\omega_i$ and momentum $\hbar k_i$ decays into two other states with energy $\hbar\omega_s$, and $\hbar\omega_1$ and momentum $\hbar k_s$ and $\hbar k_1$. The decay process shown in Fig. 2 describes the stimulated Raman scattering and that in Fig. 3 the stimulated Compton scattering.

Both figures show backscattering because k_i and k_s have opposite signs. We see from these figures that if $\omega_s \gg \omega_p$, $|k_1| \simeq 2|k_s|$. Hence for a given quality of a beam if $\omega_s (= k_s c)$ is increased, k_1 which may be initially smaller than k_D becomes larger than k_D at some value of ω_s. Hence, there exists a critical frequency of the scattered wave (which corresponds to the lasing frequency in the beam frame) above (below) which scattering process becomes Compton (Raman). If we write this critical angular frequency in the laboratory frame as ω_{cr}, that is, the actual lasing frequency, ω_{cr} can be expressed in terms of the beam quality. Using

$$\omega_{cr} = 2\gamma\omega_s$$
$$\omega_s = ck_s$$
$$k_1 = 2k_s = k_D,$$

we have, with eq. (17),

$$\omega_{cr} = \gamma k_D c$$
$$= \gamma\omega_p(\gamma/\Delta\gamma). \tag{37}$$

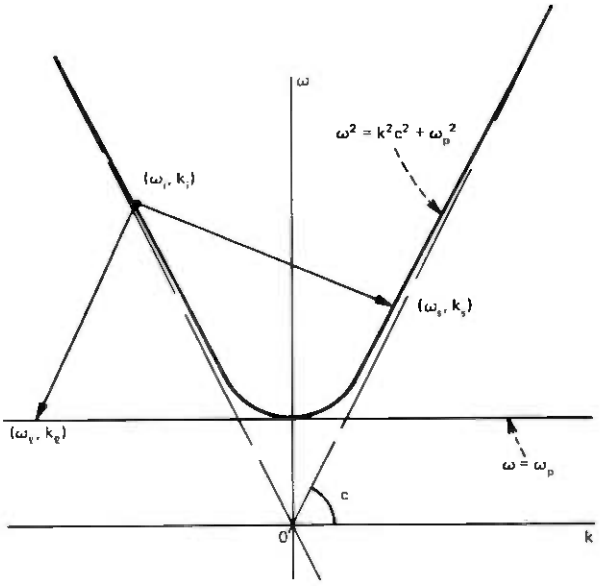


Fig. 2—Dispersion diagram of the electromagnetic wave and plasma wave in the beam frame. This diagram shows the stimulated Raman scattering process. The arrow indicates the direction of decay of the incident wave with frequency and wave number given by ω_i, k_i into a longitudinal oscillation with frequency ω_p and wavenumber k_l and a backscattered electromagnetic wave with frequency ω_s and wavenumber k_s.

Thus the critical frequency depends on the relative spread of the beam energy observed in the laboratory frame, $\Delta\gamma/\gamma$, as well as the beam density and γ. Since the plasma frequency is frame-invariant, it may be expressed in terms of the current density J_0 of the beam. Equation (37) then becomes

$$\omega_{cr} = 8.14 \times 10^6\, \gamma(\gamma/\Delta\gamma)J_0^{1/2}. \tag{37'}$$

Since MKS units are used, J_0 is in the unit of A/m^2. This expression is an important criterion in designing the laser, because at $\omega > \omega_{cr}$ it should operate in the stimulated Compton regime and the growth rate becomes pessimistically small. For a practical purpose, $\omega = \omega_{cr}$ is the high-frequency limitation of a free electron laser.

IV. THE STIMULATED RAMAN SCATTERING

In this section, we derive the growth rate in the stimulated Raman regime. A number of authors have derived the growth rate using different methods. The classic mechanical calculation is much simpler than the quantum mechanical one and is well justified for a stimulated process because a large number of photons are produced at a very early stage of the process. Tytovich's book[4] and a review paper by Kaw et al.[4] are some of the appropriate references on this subject.

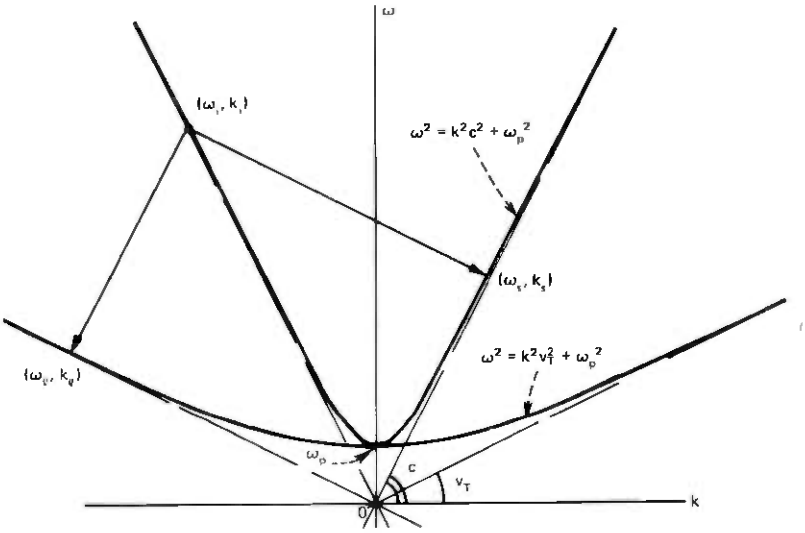


Fig. 3—The dispersion diagram that shows the stimulated Compton scattering process. When the wave number of the induced longitudinal oscillation k_1 is larger than the Debye wave number k_D, the induced longitudinal oscillation in the beam electrons becomes uncorrelated. In this case, the scattering occurs by the sum of Compton scattering by individual electrons. Since the induced wave number k_1 is proportional to the lasing frequency, when the lasing frequency is increased, the scattering process changes from the stimulated Raman to the stimulated Compton.

Attempts have been made to obtain the gain in the laboratory frame using a rather complicated nonlinear relativistic dynamics.[6,7] As has been shown, the gain and all the other parameters can be Lorentz-transformed into the laboratory frame, it is much simpler to do the nonrelativistic calculation in the beam frame. Thus we do the analysis in the beam frame. Referring to Fig. 2, we consider a large amplitude incident wave propagating in the negative z-direction with transverse electric field given by

$$\mathrm{Re}E_i \exp i(k_i z + \omega_i t), \tag{38}$$

where k_i and ω_i are positive. E_i is related to the pump field in the laboratory frame through the Lorentz transformation shown in eq. (7). In particular, if the static periodic magnetic field is used, E_i is given by

$$|E_i| = \gamma v_0 B_L \simeq \gamma c B_L, \tag{39}$$

where B_L is the amplitude of the rippled or helical magnetic field in the direction perpendicular to the beam.

To simplify the analysis, we assume the variation of E_i and all the other field quantities in the transverse direction is negligible. This assumption may be justified if the beam diameter is much larger than all the wavelengths involved.

To obtain the growth rate, we consider a test electromagnetic wave

(the scattered wave) which propagates in the direction of the beam and which is excited by a nonlinear current density produced by the product of the incident field and the induced longitudinal density perturbation in the beam.

From the Maxwell equation, the electric field of the scattered wave $\mathbf{E}_s$ satisfies the wave equation

$$\nabla^2 \mathbf{E}_s - \frac{1}{c^2} \frac{\partial^2 \mathbf{E}_s}{\partial t^2} = \mu_0 \frac{\partial \mathbf{J}_s}{\partial t}, \tag{40}$$

where the current density consists of the linear (self-consistent) portion, $\mathbf{J}_s^L$, and the nonlinear portion $\mathbf{J}_s^{NL}$, which is produced by the incident field,

$$\mathbf{J}_s = \mathbf{J}_s^L + \mathbf{J}_s^{NL}, \tag{41}$$

where

$$\mathbf{J}_s^L = -en_0 \mathbf{v}_s \tag{42}$$

and

$$\mathbf{J}_s^{NL} = -en_l \mathbf{v}_i. \tag{43}$$

$\mathbf{v}_s$ is the electron velocity modulation due to the scattered field

$$\frac{d\mathbf{v}_s}{dt} = -\frac{e}{m} \mathbf{E}_s, \tag{44}$$

while $\mathbf{v}_i$ is the modulation due to the pump field. In the case of a helical field pump, $\mathbf{v}_i$ is given by eq. (24),

$$\mathbf{v}_i = \frac{\gamma e \mathbf{B}_\perp}{\gamma_0 m k_0}, \tag{45}$$

and n_l is the density modulation due to the induced longitudinal oscillation in the beam, which satisfies the continuity equation,

$$\frac{\partial n_l}{\partial t} + \nabla \cdot (n_0 \mathbf{v}_l) = 0, \tag{46}$$

with

$$\frac{d\mathbf{v}_l}{dt} = -\frac{e}{m} \mathbf{E}_l. \tag{47}$$

$\mathbf{E}_l$ is the electric field of the longitudinal oscillation.

If we Fourier-transform (43), $\mathbf{J}_s^{NL}$ contains two frequency components, one the Stokes mode, $\omega_i - \omega$ and the other the anti-Stokes mode, $\omega_i + \omega$, where ω is the frequency of the induced longitudinal oscillation. To obtain the growth rate due to the stimulated Raman scattering, we need to retain only the Stokes mode. (We discuss the effect of anti-Stokes mode later.) If we Fourier-transform eqs. (40) to (44), retaining only the Stokes mode, we have

$$\left\{k_s^2 - \frac{1}{c^2}[(\omega_i - \omega)^2 - \omega_p^2]\right\} \mathbf{E}_s = -i(\omega_i - \omega)\mu_0 e n_1^* \mathbf{v}_i, \tag{48}$$

where * shows the complex conjugate.

If we express n_1 in terms of $\mathbf{E}_1$, using Eqs. (46) and (47),

$$n_1 = n_0 \frac{e}{m} \frac{\mathbf{k}_1 \cdot \mathbf{E}_1}{i\omega^2}, \tag{49}$$

eq. (48) becomes

$$D_s(k_s, \omega_i - \omega)\mathbf{E}_s = \omega_s(\mathbf{k}_1 \cdot \mathbf{E}_1^*)\mathbf{v}_i, \tag{50}$$

where

$$D_s(k, \omega) = k^2c^2 + \omega_p^2 - \omega^2, \tag{51}$$

and $\omega_1 \simeq \omega_p$ is used in evaluating the right-hand side of (50). $D_s = 0$ gives the linear dispersion relation of the scattered electromagnetic wave. Equation (50) shows that the dispersion relation is modified by the incident electromagnetic wave and the induced longitudinal wave.

To close the equations, we now must express $\mathbf{E}_1$ in terms of $\mathbf{E}_s$ and $\mathbf{v}_i$. The set of equations that describe the longitudinal mode are Poisson's equation,

$$\nabla \cdot \mathbf{E}_1 = -\frac{en_1}{\epsilon_0}, \tag{52}$$

and the continuity equation (46), both of which are linear, and the equation of motion,

$$\frac{d\mathbf{v}_1}{dt} = -\frac{e}{m}(\mathbf{E}_1 + \mathbf{v}_i \times \mathbf{B}_s + \mathbf{v}_s \times \mathbf{B}_i). \tag{47'}$$

The continuity equation is linear because the electromagnetic wave is incompressible, $n_s = n_i = 0$. This means that the current density for the longitudinal mode is given by $-en_0\mathbf{v}_1$. Hence, the only nonlinearity comes from the Lorentz force, $\mathbf{v} \times \mathbf{B}$, in eq. (47'). Note that we dropped the corresponding nonlinear term in the calculation of $\mathbf{J}_s^{NL}$ because it is smaller than the term retained by the factor of v_i/c. Also note that we used the linear relation, eq. (47), to express n_1 to evaluate the coupling term $n_1\mathbf{v}_i$ of (50) because it was a higher order correction there. If we use the Maxwell equation,

$$\omega\mathbf{B} = \mathbf{k} \times \mathbf{E}, \tag{53}$$

the nonlinear terms in (47') become

$$\begin{aligned}
&(\mathbf{v}_i \times \mathbf{B}_s^* + \mathbf{v}_s^* \times \mathbf{B}_i) \\
&= \left(\mathbf{v}_i \times \frac{\mathbf{k}_s \times \mathbf{E}_s^*}{\omega_s} + \mathbf{v}_s^* \times \frac{\mathbf{k}_i \times \mathbf{E}_i}{\omega_i}\right) \\
&\simeq \frac{1}{\omega_s}(\mathbf{v}_i \cdot \mathbf{E}_s^*)(\mathbf{k}_s - \mathbf{k}_i) \\
&= -\frac{1}{\omega_s}(\mathbf{v}_i \cdot \mathbf{E}_s^*)\mathbf{k}_1,
\end{aligned} \tag{54}$$

where we used $\mathbf{k}_s \cdot \mathbf{v}_i = 0$. Hence the total longitudinal velocity modulation is given by

$$\mathbf{v}_l = \frac{1}{i\omega}\frac{e}{m}\left(\mathbf{E}_l - \frac{\mathbf{v}_i \cdot \mathbf{E}_s^*}{\omega_s}\mathbf{k}_l\right). \tag{55}$$

If we use this expression in (46) and (52), we have

$$D_l(k_l,\omega)\mathbf{k}_l \cdot \mathbf{E}_l = -k_l^2 \frac{\mathbf{v}_i \cdot \mathbf{E}_s^*}{\omega_s}, \tag{56}$$

where

$$D_l(k,\omega) = 1 - \frac{\omega_p^2}{\omega^2}, \tag{57}$$

and $D_l = 0$ gives the linear dispersion relation for the longitudinal mode. Noting that $\mathbf{E}_s$ is parallel to $\mathbf{v}_i$ in eq. (50), eqs. (50) and (56) present the set of coupled equations between the scattered wave and the induced longitudinal wave,

$$D_s \mathbf{E}_s = \omega_s(\mathbf{k}_l \cdot \mathbf{E}_l^*)\mathbf{v}_i, \tag{50}$$

$$D_l\mathbf{k}_l \cdot \mathbf{E}_l = -k_l^2 \frac{\mathbf{v}_i \cdot \mathbf{E}_s^*}{\omega_s}\frac{\omega_p^2}{\omega^2}, \tag{56}$$

through the velocity modulation by the incident wave $\mathbf{v}_i$. The dispersion relation of the coupled system is given by eliminating $\mathbf{k}\cdot\mathbf{E}_l$ and $\mathbf{E}_s$ from these equations,

$$D_s(k_s,\omega_i - \omega)D_L(k_l,\omega) + \frac{\omega_p^2}{\omega^2}k_l^2 v_i^2 = 0. \tag{58}$$

If $k_l v_i$ is much smaller than ω_l, eq. (58) may be solved for a small frequency deviation $\Delta\omega$ from the frequency given by the linear dispersion relation by expanding D_s and D_L as

$$D_s(k_s,\omega_i - \omega) = D_s(k_s,\omega_s) + \left.\frac{\partial D_s}{\partial \omega}\right|_{k_s,\omega_s}\Delta\omega = 0 + 2\omega_s\Delta\omega, \tag{59}$$

while

$$D_l(k_l,\omega) = D_l(k_l,\omega_l) + \left.\frac{\partial D_l}{\partial \omega}\right|_{k_l,\omega_l}\Delta\omega = 2\Delta\omega/\omega_p. \tag{60}$$

Substituting (59) and (60) into (58), we have

$$\Delta\omega = \pm\frac{i}{2}|k_l v_i|\left(\frac{\omega_p}{\omega_s}\right)^{1/2}. \tag{61}$$

The imaginary part in $\Delta\omega$ gives the Raman growth rate in the beam frame Γ_B^R, hence

$$\Gamma_B^R = \frac{1}{2}|k_1 v_i| \cdot \left(\frac{\omega_p}{\omega_s}\right)^{1/2}. \tag{62}$$

In the case of the periodic magnetic pump, v_i is related to $B_\perp$ through eq. (45). The growth rate in this case is then given by

$$\Gamma_B^R = \frac{\gamma^2 eB_\perp}{\gamma_0 m}\left(\frac{\omega_p}{\omega_s}\right)^{1/2}. \tag{63}$$

The gain in the laboratory frame is simply given by $\Gamma_B/2\gamma$ as shown in eq. (29).

We note here that the ratio γ^2/γ_0 can be expressed in term of γ through (21),

$$\frac{\gamma^2}{\gamma_0} = \frac{\gamma_0}{1 + \gamma_0^2 v_{L\perp}^2/c^2}. \tag{64}$$

This expression indicates that a level exists in the velocity modulation $v_{L\perp}$, or the pump strength $B_\perp$, that produces a maximum growth rate. This is because an excessively large modulation deflects the beam too much in the transverse direction, which results in reducing the value of γ. There are different ways by which the growth rate can be optimized depending on the choice of fixed quantities. In any case, the maximum growth is achieved by selecting

$$v_{L\perp}^2\, \gamma_0^2/c^2 \sim 1,$$

or in terms of the modulation magnetic field,

$$\frac{eB_\perp}{m}\frac{1}{k_0 c} = \frac{eB_\perp}{m}\frac{1}{\omega_0} \sim 1. \tag{65}$$

When the pump intensity is large such that the growth rate Γ_B becomes larger than the plasma frequency, that is, if

$$|k_1 v_i| > (\omega_p \omega_s)^{1/2}, \tag{66}$$

the longitudinal mode loses its linear property. In this regime, the growth rate should be obtained from (58) without expanding $D_l(k_l,\omega)$ around k_l,ω_p.[8] The growth rate is then modified to

$$\Gamma_B^0 = \left[\frac{\omega_p^2 k_l^2 v_i^2}{2\omega_i}\right]^{1/3}. \tag{67}$$

This regime is often called the oscillating two-stream instability (OTSI).[9]

If the pump amplitude is further increased, we should include the effect of the anti-Stokes mode which is simultaneously coupled in. The dispersion relation including the anti-Stokes mode becomes

$$1 + \frac{k_l^2 v_i^2 \omega_p^2}{D_l(k_l,\omega)\omega^2}\left[\frac{1}{D_s(k_s,\omega_i - \omega)} + \frac{1}{D_s(k_s^+,\omega_i + \omega)}\right] = 0, \tag{68}$$

where k_s^+ is the wave number of the scattered anti-Stokes mode. The growth rate in this regime is shown to be proportional to v_i^2, and it corresponds to the modulation instability (for example, see Ref. 10) of the pump wave.

V. STIMULATED COMPTON SCATTERING

Here we obtain the gain in the stimulated Compton regime. As was discussed in Section III, if the wave number of the longitudinal oscillation induced in the beam electrons is larger than the Debye wave number, $k_D (= \omega_p/v_T)$, the collective nature of the longitudinal mode is lost. The scattering then occurs by the individual electrons.

Because distribution of velocities exists in the beam electrons, to obtain the total scattering gain we must average over the velocity distribution. If we look at Fig. 3, we see that the resonant condition of the stimulated Compton scattering in the beam frame is given by

$$\omega_i - \omega_s = |k_1| v_T, \tag{69}$$

$$|k_i| + |k_s| = |k_1|. \tag{70}$$

As we have seen in the case of the stimulated Raman scattering, we must obtain J_s^{NL} to calculate the effect of the pump on the scattered mode in (40). In the present case, the Fourier amplitude of J_s^{NL} is again given by

$$J_s^{NL} = -en_1^* v_i; \tag{71}$$

however, the calculation of n_1^* is more complicated because of the averaging over the velocity distribution.

To obtain n_1^*, we use the Vlasov equation, which includes the nonlinear force term produced by the $\mathbf{v} \times \mathbf{B}$ force as seen previously.

$$\frac{\partial f_1}{\partial t} + v_z \frac{\partial f_1}{\partial z} + \frac{F_1^{NL}}{m} \frac{\partial f_0}{\partial v_z} = 0, \tag{72}$$

where f_1 and f_0 are the perturbed (which represents the induced density modulation) and unperturbed velocity distribution function of electrons in the beam frame, v_z is the z component of velocity, and F_1^{NL} is the nonlinear force acting upon electrons at the frequency $\omega = \omega_1$,

$$F_1^{NL} = -e(\mathbf{v}_i \times \mathbf{B}_s + \mathbf{v}_s \times \mathbf{B}_i). \tag{73}$$

In (72), the linear force produced by the self-field, eE_1/m, is ignored because the induced longitudinal field is nonresonant; that is, $D_l(k_1,\omega_1) \neq 0$, due to the heavy Landau damping, and hence its amplitude is small. If we Fourier-transform eqs. (72) and (73) and take only the Stokes term, we have

$$f_1 = \frac{\partial f_0/\partial v_z}{i(k_1 v_z - \omega)} \frac{e}{m} \frac{k_1}{\omega_s} \mathbf{v}_i \cdot E_s^*. \tag{74}$$

The induced charge density n_1 is then obtained by integrating this expression over v_z,

$$en_1 = i\frac{k_1^2\,\epsilon_0 \mathbf{v}_i \cdot \mathbf{E}_s^*}{\omega_s}\chi_1, \tag{75}$$

where χ_1 is the susceptibility of an electron gas,

$$\chi_1 = -\frac{\omega_p^2}{k_1^2}\int \frac{\partial f_0/\partial v_z}{v - (\omega + i0)/k_1}dv_z. \tag{76}$$

The dispersion relation for the scattered wave is now obtained by substituting (75) for the expression for the nonlinear current density, (71), and using it in the wave equation for the scattered electric field, (48).

$$[(\omega_i - \omega)^2 - \omega_p^2 - c^2k_s^2]E_s = |v_i|^2\chi_1^* k_1^2 E_s. \tag{77}$$

If we solve for $\omega \simeq \omega_i - \omega_s + \Delta\omega$, we have

$$\Delta\omega = -\frac{\chi_1^*}{2\omega_s}k_1^2|v_i|^2. \tag{78}$$

The temporal growth rate is obtained from the imaginary part of χ_1^*. From Eq. (76), we see

$$\text{Im}\,\chi_1 = -\frac{\chi_p^2}{k_1^2}\pi\int \delta(v - \omega/|k_1|)\frac{\partial f_0}{\partial v_z}dv_z. \tag{79}$$

If we take the Maxwellian velocity distribution for f_0 in the beam frame,

$$f_0 = \frac{1}{\sqrt{2\pi}\,v_T}e^{-v^2/2v_T^2}, \tag{80}$$

$$\text{Im}\,\chi_1 = \frac{\omega_p^2}{k_1^2v_T^2}\sqrt{\frac{\pi}{2e}} \simeq 0.76\frac{\omega_p^2}{k_1^2v_T^2}. \tag{79'}$$

The Compton growth rate Γ_B^c is now obtained from (78) and (79'),

$$\Gamma_B^c \simeq 0.4\frac{\omega_p^2}{\omega_s}\frac{|v_i|^2}{v_T^2}. \tag{81}$$

If we compare the Compton growth rate Γ_B^c with the Raman growth rate, (62), we see a qualitative difference. The Compton growth rate is proportional to the pump amplitude squared, while the Raman growth rate is proportional to the pump amplitude itself.

If the pump amplitude is increased such that $v_i > v_T$, it has been shown by Hasegawa et al.[11] that the pump field effectively increases the velocity spread by $\mathbf{v}_i \times \mathbf{B}_i$ force and thus decreased the gain. The proof was made for an electromagnetic wave pump, but it is believed that even when the helical magnetic pump is used, the similar effect appears when the beam enters into the magnetic field and suddenly see the magnetic field pressure, $B_1^2/2\mu_0$. The Compton gain for such a case becomes[11]

$$\Gamma_B^c \simeq 0.3\frac{\omega_p^2}{\omega_s}\frac{v_i}{c}\left(\frac{c}{v_T}\right)^{3/2}. \tag{82}$$

One important remark should be made here. We obtained Raman and Compton gains by taking the asymptotic limits of $k_l \ll k_D$ and $k_l \gg k_D$, respectively, to have simple analytic expressions. However, this does not mean that the gain at the transition regime cannot be obtained, nor that an abrupt transition exists between the two regimes. In fact, the unified dispersion relation which covers the entire regime can be obtained by using the Vlasov equation and by simply including the self-consistent electric field $\mathbf{E}_l$ in (72). If we further allow a situation that the scattered wave may not propagate in the beam direction, the unified dispersion relation which is expressed in the form of eq. (68) becomes

$$1 - \frac{k_l^2 \chi_l^* (k_l,\omega)}{1 + \chi_l^* (k_l,\omega)} \left[\frac{|\mathbf{k}_s \times \mathbf{v}_i|^2}{k_s^2 D_s(k_s,\omega_i - \omega)} + \frac{|\mathbf{k}_s^+ \times \mathbf{v}_i|^2}{k_s^{+2} D_s(k_s^+,\omega_i + \omega)} \right] = 0. \quad (83)$$

The gain for the entire regime is obtained by numerically solving this equation for ω.

VI. LIMITING GAIN AND OUTPUT POWER

In the previous two sections, temporal growth rates for stimulated Raman and stimulated Raman scatterings were obtained. We summarize the result in the following, by using $k_l \simeq 2|k_i| \simeq 2\omega_s/c$, and $\omega_s \sim \omega_i$.

Raman gain (beam frame)

$$\Gamma_B^R = \frac{|v_i|}{c} (\omega_p \omega_i)^{1/2}, \qquad \text{if } \frac{|v_i|}{c} \ll \left(\frac{\omega_p}{\omega_i}\right)^{1/2}, \quad (84)$$

$$\Gamma_B^0 = \left(2 \frac{|v_i|^2}{c^2} \omega_p^2 \omega_i\right)^{1/3}, \qquad \text{if } \frac{|v_i|}{c} \gg \left(\frac{\omega_p}{\omega_i}\right)^{1/2}. \quad (85)$$

Compton gain (beam frame)

$$\Gamma_B^c = 0.4 \frac{|v_i|^2}{v_T^2} \frac{\omega_p^2}{\omega_i}, \qquad \text{if } \frac{v_i}{v_T} < 1, \quad (86)$$

$$\Gamma_B^c = 0.3 \frac{|v_i|}{c} \frac{\omega_p^2}{\omega_i} \left(\frac{c}{v_T}\right)^{3/2}, \qquad \text{if } \frac{v_i}{v_T} > 1. \quad (87)$$

The gain in all cases depends on the pump intensity v_i. If one uses the helical magnetic pump, as we have shown in Section VI, an optimum value exists in the pump magnetic field $B_\perp$, which is given by eq. (64). The corresponding velocity v_i becomes $|v_i|/c \simeq 1/\sqrt{2}$. If we use this value, the Raman (OTSI) and Compton gains become

$$\Gamma_{B\max}^R \simeq (\omega_p^2 \omega_i)^{1/3}, \quad \text{applicable for } \omega_i \ll \frac{\gamma}{2\Delta\gamma} \omega_p, \quad (88)$$

$$\Gamma_{B\max}^c \simeq 0.2 \left(\frac{\gamma}{\Delta\gamma}\right)^{3/2} \frac{\omega_p^2}{\omega_i}, \quad \text{applicable for } \omega_i \gg \frac{\gamma}{2\Delta\gamma} \omega_p. \quad (89)$$

Here $\omega_i = 2\Delta\gamma/\gamma\ \omega_p$ corresponds to the critical frequency, eq. (37) between the two regimes, that is the incident frequency for $k_l = k_D$.

We see that the growth rate increases gradually as ω_i is increased and then decreases in proportion to ω_i^{-1}. If we take an example of a best quality beam with $\Delta\gamma \sim 10^{-3}\,\gamma$, $\Gamma^c_{B\,\max}$ at the critical frequency is given approximately by

$$\Gamma^c_{B\,\max} \simeq 0.4\,\omega_p \left(\frac{\gamma}{\Delta\gamma}\right)^{1/2} \simeq 12\,\omega_p.$$

On the other hand, at the same frequency,

$$\Gamma^R_{B\,\max} \sim \omega_p \left(\frac{\gamma}{2\Delta\gamma}\right)^{1/3} \simeq 7.8\,\omega_p.$$

This indicates that, at the critical frequency, the Raman and Compton gains are approximately the same. If we now express the plasma frequency in terms of the beam current density J_0, $\omega_p = 8.14 \times 10^6\sqrt{J}$, hence the maximum growth rate in the beam frame is approximately given by $\Gamma_{B\,\max} \sim 10\,\omega_p \sim 10^8\sqrt{J}$. As an example, if we take a nominal parameter of "microtron" [12] beam with a current of 1 A with the cross section of 1 mm^2, $J = 10^6$ A/m^2. Thus, $\Gamma_{B\,\max} \simeq 10^{11}$ sec^{-1}. We also note that the gain in the laboratory frame Γ_L is given by $\Gamma_B/2\gamma$. For a nominal value of $\gamma = 10^3$, the laboratory frame gain is 5×10^8 sec^{-1}. Hence the e-folding distance $L = c/\Gamma_L \simeq 1$ m. The e-folding distance at a lower frequency becomes shorter in proportion to $\omega_i^{-1/3}$, while at a higher frequency becomes longer in proportion to ω_i.

These arguments may be summarized as follows. If we define the critical frequency given by (37) as the limiting frequency that the free electron laser can operate, the minimum e-folding distance in the laboratory frame L_m and ω_{cr} can be expressed in terms of J_0, γ and $\gamma/\Delta\gamma$.

The maximum lasing frequency, f_{cr}:

$$f_{cr} = \frac{\omega_{cr}}{2\pi} = 1.3 \times 10^6\,\gamma \left(\frac{\gamma}{\Delta\gamma}\right) [J_0(\mathrm{A/m^2})]^{1/2}\ \mathrm{Hz}. \tag{90}$$

The minimum e-folding distance, L_m:

$$L_m = \frac{c}{\Gamma_{L\,\max}} = 93\gamma \left(\frac{\Delta\gamma}{\gamma}\right)^{1/3} [J_0(\mathrm{A/m^2})]^{1/2}\ \mathrm{m}. \tag{91}$$

Condition to achieve L_m:

$$\frac{eB_\perp}{m} = k_0 c = \frac{\omega_{cr}}{2\gamma^2}$$

or

$$B_\perp(W/m^2) = 1.8 \times 10^{-11}\frac{f_{cr}}{\gamma^2}. \tag{92}$$

Note that the beam pulse length (Fig. 1) is not a crucial parameter so long as it is longer than, say, $10\,k_1^{-1}$ because it runs at the same speed as the

scattered light. If we take again the previous examples of microtron,[12] $J_0 = 10^6$ A/m^2, $\gamma = 10^2$, and $\gamma/\Delta\gamma = 10^3$, we have

$$f_{cr} = 1.3 \times 10^{14} \text{ Hz}$$

$$L_m = 0.93 \text{ m}$$

$$B_\perp = 2.3 \times 10^{-1} \text{ W/m}^2$$

$$\lambda_0 = 2\pi/k_0 = \frac{2\gamma^2 c}{f_{cr}} = 4.6 \times 10^{-2} \text{ m}.$$

Let us now discuss the maximum output power of the laser. Because L_m is on the order of 1 m, it takes a relatively long system to achieve the saturation in gain. But let us assume that the system is infinitely long and ask ourselves what causes the saturation of the gain.

As we have found, when the energy spread of the beam becomes large so that $k_1 < k_D$, the gain drops in proportion to ω_i^{-1}. When the scattered power is increased, it produces a larger $\mathbf{v} \times \mathbf{B}(= \mathbf{v}_i \times \mathbf{B}_s)$ force which traps the beam electrons and increases its energy spread. The trapping potential ϕ_t due to the Lorentz force $\mathbf{v}_i \times \mathbf{B}_s$ in the beam frame is obtained from

$$\left|\frac{\partial \phi_t}{\partial z}\right| = |k_1 \phi| \simeq |v_i B_s^*|$$

or

$$\phi_t = \frac{1}{k_1} |v_i| |B_s|. \tag{93}$$

The effective thermal speed v_{Teff} produced by the trapping potential ϕ_t is

$$v_{\text{Teff}} = \left(\frac{2e\phi_t}{m}\right)^{1/2}. \tag{94}$$

We can consider that the saturation occurs when $k_1 \simeq \omega_p/v_{\text{Teff}}$ because if v_{Teff} is made larger than this critical value, the gain changes from Raman to Compton. Hence, the maximum amplitude of the magnetic field of the scattered wave is given by

$$k_1 = \frac{\omega_p}{(2e\phi_t/m)^{1/2}} = \frac{\omega_p}{(2e|v_i||B_s|/k_1 m)^{1/2}}, \tag{95}$$

or by solving B_s using $|v_i| \simeq c$, we have

$$B_s = \frac{m}{e}\frac{\omega_p^2}{ck_i} = \frac{m}{e}\frac{\omega_p^2}{\omega_i}. \tag{96}$$

If we operate at the maximum gain, $\omega_i = \omega_{cr}/2\gamma = \omega_p(\gamma/\Delta\gamma)/2$. Hence, we must use as the maximum scattered field

$$B_s = 2\frac{m}{e}\omega_p \frac{\Delta\gamma}{\gamma}, \tag{97}$$

and the corresponding electric field is

$$E_s = cB_s. \tag{98}$$

If we Lorentz-transform these fields to the laboratory frame according to eqs. (5) and (7), we have

$$B_{Ls} = 2\gamma B_s$$

and

$$E_{Ls} = 2\gamma E_s. \tag{99}$$

Hence, the maximum output power P_m is given by

$$\begin{aligned} P_m &= E_{Ls}B_{Ls}/\mu_0 \\ &= 16\gamma^2 \left(\frac{m}{e}\right)^2 \omega_p^2 \left(\frac{\Delta\gamma}{\gamma}\right)^2 \frac{c}{\mu_0} \\ &= 16 \left(\frac{\Delta\gamma}{\gamma}\right)^2 P_{\text{Beam}}, \end{aligned} \tag{100}$$

where P_{Beam} is the beam kinetic power density,

$$P_{\text{Beam}} = mc^3\gamma n. \tag{101}$$

Equation (100) shows that the conversion efficiency is roughly given by $16(\Delta\gamma/\gamma)^2$. This may be misleading, because it shows that the poorer quality beam gives better efficiency. This comes from the dependency of B_s on ω_i^{-1} so that the lower the frequency the longer the saturation field. When a poor quality beam is used, the efficiency may become better but with a sacrifice of lowering the laser frequency.

If we use the same example of parameters, $\gamma = 10^2$, $\Delta\gamma/\gamma = 10^{-3}$ and 1 A beam, the maximum output power of the laser becomes 800 W.

VII. CONCLUSION

Use of stimulated backscattering of a pump field by a relativistic electron beam for a tunable laser was discussed. The temporal gain and the e-folding distance in the laboratory frame are obtained for both stimulated Raman and stimulated Compton scattering regimes. It is shown that in the stimulated Compton regime, the gain drops in proportion to the lasing frequency hence is not a practical regime to deploy. If we consider that the transition frequency from the Raman to the Compton regime is the maximum lasing frequency, the lasing frequency can be obtained as a function of the beam energy γ, the relative energy spread of the beam $\Delta\gamma/\gamma$, and the current density J_0 as shown in (90). The e-folding distance corresponding to this frequency is shown in eq. (91). For a nominal value of the available relativistic electron beam, these quantities become approximately 10^{14} Hz and 1 m. The maximum power output corresponding to this operation condition is also obtained and

shown to be given by (100). Again for the nominal value of the beam parameter, the output laser power becomes about one kilowatt. These results indicate that the use of a relativistic beam with γ of 100 and $\Delta\gamma/\gamma$ of 10^{-3} can produce a tunable laser with an optimum operating frequency approaching to the visible. However, extending this process into X-ray regime seems extremely difficult.

VIII. ACKNOWLEDGMENTS

The author would like to thank B. M. Kincaid for his interest in this problem and many valuable discussions on available experimental data, and to H. Ikezi and K. Mima for valuable technical discussions.

REFERENCES

1. V. L. Granatstein, S. P. Schlesinger, M. Herndon, R. K. Parker, and J. A. Pasour, "Production of Megawatt submillimeter pulses by Stimulated Magneto-Raman Scattering," App. Phys. Lett., *30* (1977), pp. 384–386.
2. D. A. G. Deacon, L. R. Elias, J. M. J. Madey, G. R. Ramian, H. A. Schwettman, and T. I. Smith, "First Operation of Free Electron Laser," Phys. Rev. Lett., *38*, pp. 892–894.
3. L. D. Landau and E. M. Lifshitz, *The Classical Theory of Field*, New York: Pergamon Press, 1975, pp. 9–65.
4. V. N. Tsytovich, "*Nonlinear Effects in Plasma*," New York: Plenum Press, 1970.
5. P. K. Kaw, W. L. Kruer, C. S. Lin, and K. Nishikawa, "Parametric Instabilities in Plasma," *Advances in Plasma Physics*, Vol. 6, Ed. by A. Simon and W. B. Thompson, New York: Academic Press, 1976, pp. 1–270.
6. P. Sprangle and A. T. Drobot, "Stimulated Backscattering from Relativistic Unmagnetized Electron Beams," Proc. Free Electron Generators of Coherent Radiation Workshop, Telluride, Colorado, August 17–21, 1977.
7. F. A. Hopf, P. Meystre, M. O. Scully, and W. H. Louisell, "Strong-Signal Theory of a Free Electron Laser," Phys. Rev. Lett., *37* (1977) pp. 1342–1944.
8. V. P. Silin, "Parametric Resonance in a Plasma," Sov. Phys. JETP, *21* (1965) pp. 1127–1134, and K. Nishikawa, "Parametric Excitation of Coupled Waves, I, General Formulation," J. Phys. Soc. Japan, *24* (1968) pp. 916–922.
9. J. Drake, P. Kaw, Y. C. Lee, G. Schmidt, C. S. Liu, and M. N. Rosenbluth, "Parametric Instabilities of Electromagnetic Waves in Plasmas," Phys. Fluids, *17* (1974), pp. 778–785.
10. A. Hasegawa, *Plasma Instabilities and Nonlinear Effects*, Heidelberg: Springer-Verlag 1975, pp. 201–204.
11. A. Hasegawa, K. Mima, P. Sprangle, H. H. Szu, and V. L. Granatstein, "Limitation in Growth Time of Stimulated Compton Scattering in X-ray Regime," Appl. Phys. Lett., *29*, (1976), pp. 542–544.
12. P. M. Lapstolle and A. L. Septier, eds. *Linear Accelerators*, North-Holland, 1970, pp. 553–567.

Contributors to This Issue

Anthony S. Acampora, B.S.E.E., 1968, M.S.E.E., 1970, Ph.D., 1973, Polytechnic Institute of Brooklyn; Bell Laboratories, 1968—. Mr. Acampora initially worked in the fields of high power microwave transmitters and radar system studies and signal processing. Since 1974, he has been studying high capacity digital satellite systems. His current research interests are modulation and coding theory, time division multiple access methods, and efficient frequency re-use techniques. Member Eta Kappa Nu, Sigma Xi, IEEE.

James L. Blue, A.B., 1961, Occidental College; Ph.D., 1966, California Institute of Technology; Bell Laboratories, 1966—. Mr. Blue has done research in noise theory for avalanche diodes and in modeling of semiconductor devices, and was involved in the development of computer aids for testing of integrated circuits. He is now a member of the Computing Mathematics Research Department, where he is involved in mathematical modeling, research in numerical methods, and the development of numerical software.

H. J. Braun, Bell Laboratories, 1941—. Mr. Braun has worked on the development of the solderless wrapped connection and automatic machine wiring, and on the design of equipment for semiconductor fabrication. He is currently engaged in the mechanical design and packaging of opto-isolators.

Fan R. K. **Chung,** B.S., 1970, National Taiwan University; Ph.D., 1974, University of Pennsylvania; Bell Laboratories, 1974—. Mrs. Chung's current interests include combinatorics, graph theory, and the analysis of algorithms. She is presently investigating various problems in the theory of switching networks.

Ronald E. Crochiere, B.S., (E.E.) 1967, Milwaukee School of Engineering; M.S. (E.E.), 1968, Ph.D. (E.E.), 1974, Massachusetts Institute of Technology; Raytheon Co., 1968–1970; Bell Laboratories, 1974—. Mr. Crochiere is presently engaged in research activities in speech communications, speech coding, and digital signal processing. Member, IEEE, Sigma Xi, ASSP-DSP Subcommittee.

Bruce R. Davis B.E., 1960, B.Sc., 1963, Ph.D., 1969, University of Adelaide, Australia. Mr. Davis has been with the University of Adelaide since 1964 and at present is a Senior Lecturer in Electrical Engineering. His research interests are in the field of communication systems. During 1970 he was with Bell Laboratories, Holmdel, New Jersey, studying various aspects of mobile radio communications, and again in 1977 when he was involved in satellite systems research. Member, IEEE.

A. Ross Eckler, B.A., 1950, Swarthmore College; Ph.D., 1954, Princeton University; Bell Laboratories, 1954—. For much of his Bell Laboratories career, Mr. Eckler worked on a variety of probabilistic models in the military area, particularly missile guidance codes, target coverage,and anti-ballistic missile allocation; since 1972, he has consulted on a number of statistical problems in Bell System operations, including subject-matter fields such as building fires, investment tax credit on new construction, equal employment opportunity, and outside plant cable installation. He is currently Head of the Common Systems Analysis Department, and a member of Sigma Xi, Phi Beta Kappa, and the American Statistical Association.

A. G. Fraser, B.Sc. (aero. engin.), 1958, Bristol University; Ph.D. (computing science), 1969, Cambridge University; Bell Laboratories, 1969—. Mr. Fraser has been engaged in computer and data communications research. His work includes the Spider network of computers and a network-oriented file store. Prior to joining Bell Laboratories, he was at Cambridge University where he wrote the file system for the Atlas 2 computer. In 1977 he was appointed Head, Computer Systems Research Department. Member, IEEE, ACM, British Computer Society.

B. Gopinath, M.Sc. (Math.), 1964, University of Bombay; Ph.D. (E.E.) 1968, Stanford University; research associate, Stanford University, 1967–1968; Alexander von Humbolt research fellow, University of Göttingen, 1971–1972; Bell Laboratories, 1968—. Mr. Gopinath is engaged in applied mathematics research in the Mathematics and Statistics Research Center.

Akira Hasegawa, B.S., M.S., Osaka University; Ph.D., 1964, University of California, Berkeley; Associated Professor, Osaka University, 1964–1968; Bell Laboratories, 1968—. Mr. Hasegawa's primary fields

of interest are plasma physics, space physics, nonlinear optics, and fluid dynamics. Since 1971, he has been serving as Adjunct Professor at Columbia University. He is a Fellow of American Physical Society, Senior Member, IEEE and Member, American Geophysical Union, Physical Society of Japan, and Sigma Xi.

Frank K. Hwang, B.A., 1960, National Taiwan University; M.B.A., City University of New York; Ph.D. (Statistics), 1968, North Carolina State University; Bell Laboratories, 1967—. Mr. Hwang spent the fall of 1970 visiting the Department of Mathematics of National Tsing-Hua University. He has been engaged in research in statistics, computing science, discrete mathematics, and switching networks.

John C. Irvin, B.A., 1949, Miami University (Ohio); M.A., 1953, Ph.D. (Physics), 1957, University of Colorado; Bell Laboratories, 1957—. Mr. Irvin has engaged in research on the properties of bulk Si and diffused layers in Si and on Si interfaces. He has been involved in the development of microwave semiconductor devices including GaAs varactors, mixer diodes, Gunn diodes, IMPATT diodes, and most recently the reliability aspects of the GaAs FET. Senior member, IEEE; member, American Physical Society, Sigma Xi, and Phi Beta Kappa.

Andrew S. Jordan, B.S. (Metallurgy), 1959, Pennsylvania State University; Ph.D. (Metallurgy), 1965, University of Pennsylvania; Bell Laboratories, 1965—. Mr. Jordan has worked mainly in the area of compound semiconductors. He had been involved in the growth, phase equilibria, and impurity incorporation of ZnTe, CdTe, GaP, and GaAs. More recently, he has studied the degradation and reliability of GaP LEDs. Currently, he is engaged in modeling GaAs crystal growth. Member, Electrochemical Society.

T. C. Liang, B.S. (math.), 1972, and M.S. (applied math.), 1976, National Tsing-Hua University; Telecommunication Laboratories, 1976—. Mr. Liang has been engaged in research in statistics and switching networks.

A. Loya, Assoc. E.E., 1955, Pennsylvania State College, U.S.N. 1955–1959, Bell Laboratories, 1960—. Mr. Loya has worked on developing transistor fabrication, Si interfaces involving dry oxides (MOS),

IMPATT diodes, and BARITT diodes, and is presently working on the reliability and stress aging of the GaAs FET.

Barbara J. McDermott, B.A. (Psychology), 1949, University of Michigan; M.A. (Psychology), 1963, Columbia University; Haskins Laboratories, 1950–1959; Bell Laboratories, 1959—. Ms. McDermott has worked on speech quality evaluation and multidimensional scaling analysis. Member, Acoustical Society of America.

Carol A. McGonegal, B.S. (cum laude) (mathematics), 1974, Fairleigh Dickinson University; M.S. (computer science), 1977, Stevens Institute of Technology; Bell Laboratories, 1967—. Ms. McGonegal is a member of the Acoustics Research Department, where she has worked on problems in digital filter design, digital speech processing, computer voice response, and speaker verification.

James P. Moreland, B.S.E.E., 1964; M.S.E.E., 1964; Ph.D. (E.E.), 1967, Ohio State University; Research Associate, Electroscience Laboratory, 1964–1968, Instructor, Electrical Engineering, 1967–1968, both Ohio State University; Bell Laboratories, 1968—. At Ohio State, Mr. Moreland worked on studies of scattering theory and optical heterodyne detection. At Bell Laboratories, he has been concerned with clock-synchronization schemes for digital communications networks, optical-fiber transmission studies, and traffic and facility network planning. Member, IEEE, Eta Kappa Nu, Tau Beta Pi, and Sigma Xi.

John A. Morrison, B.Sc., 1952, King's College, University of London; Sc.M., 1954, and Ph.D., 1956, Brown University; Bell Laboratories, 1956—. Mr. Morrison has done research in various areas of applied mathematics and mathematical physics. He has recently been interested in queuing problems associated with data communications networks. He was a Visiting Professor of Mechanics at Lehigh University during the fall semester, 1968. Member, American Mathematical Society, SIAM, IEEE, Sigma Xi.

F. W. Mounts, E.E., 1953, M.S., 1956, University of Cincinnati; Bell Laboratories, 1956—. Mr. Mounts has been concerned with research in efficient methods of encoding pictorial information for digital television

and graphics systems. Member, Eta Kappa Nu; Senior Member, IEEE.

Arun N. Netravali, B. Tech. (Honors), 1967, Indian Institute of Technology, Bombay, India; M.S., 1969, and Ph.D. (E.E.), 1970, Rice University; Optimal Data Corporation, 1970–1972; Bell Laboratories, 1972—. Mr. Netravali has worked on problems related to filtering, guidance, and control for the space shuttle. At Bell Laboratories, he has worked on various aspects of signal processing. He is presently Head of the Visual Communication Research Department. He has been on the adjunct faculty at Rutgers University since 1976. Member, Tau Beta Pi, Sigma Xi; Senior Member, IEEE.

Kun I. Park, B.S., 1966, Seoul National University; M.S., 1968, Ph.D., 1972, University of Pennsylvania; Bell Laboratories, 1973—. Mr. Park has worked primarily on transmission performance objectives and requirements for public and private telephone networks. Member, IEEE, Sigma Xi.

Robert H. Peaker, Associate in Applied Science, 1964, Lowell Institute; B.S. (E.E.), 1976, New Jersey Institute of Technology; Bell Laboratories, 1967—. Mr. Peaker has been engaged in experimental work on electroluminescent materials and devices. He is currently working on electroluminescent device reliability.

Lawrence R. Rabiner, S.B. and S.M., 1964, Ph.D. (electrical engineering), Massachusetts Institute of Technology; Bell Laboratories, 1962—. From 1962 through 1964, Mr. Rabiner participated in the cooperative plan in electrical engineering at Bell Laboratories. He worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently he is engaged in research on speech communications and digital signal processing techniques. He is coauthor of *Theory and Application of Digital Signal Processing* (Prentice-Hall, 1975). Former President, IEEE G-ASSP Ad Com; former Associate Editor, G-ASSP Transactions; former member, Technical Committee on Speech Communication of the Acoustical Society. Member, G-ASSP Technical Committee on Speech Communication, IEEE Proceedings Editorial Board, Eta Kappa Nu, Sigma Xi, Tau Beta Pi. Fellow, Acoustical Society of America and IEEE.

Robert H. Saul, Ph.D. (Metallurgy and Materials Science), 1967, Carnegie-Mellon University; Bell Laboratories, 1967—. Mr. Saul worked on the growth, fabrication, and characterization of III–V semiconductor LEDs for a variety of opto-electronic devices. More recently, he has been involved with reliability of opto-electronic devices. He is supervisor of the Device Reliability and Electroluminescent Materials Group. Member, Electrochemical Society, Sigma Xi, and Tau Beta Pi.

Irwin W. Sandberg, B.E.E., 1955, M.E.E., 1956, and D.E.E., 1958, Polytechnic Institute of Brooklyn; Bell Laboratories, 1958—. Mr. Sandberg has been concerned with analysis of radar systems for military defense, synthesis and analysis of active and time-varying networks, several fundamental studies of properties of nonlinear systems, and with some problems in communication theory and numerical analysis. His more recent interests include macroeconomics and the economic theory of large corporations. Fellow and member, IEEE; member, American Association for the Advancement of Science, Eta Kappa Nu, Sigma Xi, and Tau Beta Pi.

Harry H. Wade, Associate (Electronic Technology), 1964, Philco Technical Institute; Bell Laboratories, 1965—. Mr. Wade has worked in the development of solid-state microwave power sources (IMPATT, TRAPATT). He is currently engaged in development of the linear opto-isolator.

Kenneth A. Walsh, A.A.S. (E.E.), Kent State University, Salem, Ohio, 1969; Bell Laboratories, 1969—. Mr. Walsh's work has been mainly concerned with the investigation of efficient digital coding techniques, using both hardware and software methods, with application to video and facsimile transmissions.