THE BELL SYSTEM
TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING
ASPECTS OF ELECTRICAL COMMUNICATION

Volume 55 January 1976 Number 1

Loss-Noise-Echo Study of the Direct Distance Dialing Network

By T. C. SPANG

(Manuscript received June 23, 1975)

A review of the current VIA Net Loss plan was made using a simulation model of the transmission characteristics of telephone connections. This model allowed the determination of the optimal loss which balanced the degradation due to talker echo and the degradation caused by its control on primary speech. The review reaffirmed current methods with some changes to allow for the emerging digital network.

I. INTRODUCTION

In any telephone connection, energy can be reflected back to the talking customer at any impedance discontinuity or at any junction between four-wire and two-wire circuits. This energy manifests itself to the talking customer as an echo of his own voice. If the reflected energy has sufficient amplitude and delay, it can be annoying and can interfere with the talker's normal speech process. Through the years, considerable effort has been made to reduce the amplitude of the talker echo by using impedance-matching procedures. However, it is not feasible to completely eliminate echo. Loss is therefore introduced into the connection to further attenuate the echo energy. The amount of loss needed increases with the delay of the echo signal relative to the talker's voice signal.[1] This loss also attenuates the talker's voice signal received at the far end of the connection. When the amount of loss needed for talker echo control causes undue degradation of the received voice signal, echo suppressors[2] are inserted into the connection. Echo

suppressors function by selectively and dynamically changing loss in each direction of the connection in response to talker activity.

Currently, the amount of loss and the distance at which echo suppressors are applied are determined by the VIA Net Loss (VNL) plan. This plan was developed about 20 years ago.[3] Since that time more extensive use has been made of carrier systems. These systems have considerably less propagation delay than voice-frequency cable systems. Thus, there was a question whether the amount of loss specified by the VNL plan could not be reduced and the distance increased at which echo suppressors are introduced. Also, there was some question as to appropriateness of the through- and terminal-balance requirements used to specify the maximum allowable amount of impedance mismatch, and it was felt that some simplification of the loss plan could possibly result in some operational and economic benefits.

In addition, it will be possible with the introduction of the No. 4 Toll ESS switching system[4] for a digital trunk to be switched in digital form to another trunk without requiring digital channel banks to decode the signal to voice frequencies. However, present VNL design rules require loss to be introduced after the receiving channel bank equipment of each trunk. In a digital office, this would require either that a signal be decoded, loss inserted, and re-encoded; or that the encoded digital representation be changed by digital processing to a lower signal level. Either of these techniques would introduce additional cost and transmission impairments.

Because of these questions, a thorough review of the methods to control talker echo was initiated. This review included examining the appropriateness of using loss, balance requirements, and echo suppressors for talker echo control. To achieve this review, a comprehensive program was initiated to determine the transmission characteristics of the network, and to develop by subjective tests the relationship between measurable parameters and customer opinion of the quality of service. (The results of these studies are being published in other articles.[1,5]) These studies were then used in the development of a computer-simulation model which allowed the investigation of the optimal trade-off between the degradation due to talker echo and the degradation caused by its control on the primary speech. This paper reports on the results and recommendations arrived at by this investigation.

II. SUMMARY OF RECOMMENDATIONS

This review examined the appropriateness of various methods for talker echo control on three types of connections which, although not

all occurring today, may all occur in the near future. These types are (*i*) connections using analog or digital facilities switched via analog switching systems—referred to as the analog network, (*ii*) connections using only digital facilities switched digitally—referred to as the digital network, and (*iii*) connections using portions of both types of networks —referred to as the mixed analog-digital network. This review led to the following recommendations.

(*a*) The VNL plan provides nearly optimal loss for connections over the analog network, and need not be changed.

(*b*) Some simplification of the VNL plan is possible. However, at this time, there appears to be no significant improvement in quality or economic benefit in making a change to such a plan for the analog network.

(*c*) There are significant quality, technological, and economic reasons for having a different loss plan for connections using the emerging digital network. The recommended plan is (1) a fixed-loss plan from Class 5 to Class 5 office of 6 dB for all lengths of digital connections and (2) the establishment of the nominal transmission level for digital toll offices of -3.TLP (transmission level point); a change from -2 TLP of analog toll offices.

(*d*) The two loss plans for the analog and digital networks are compatible. Connections using portions of both networks have transmission quality intermediate between that of the analog and digital networks.

(*e*) Terminal balance requirement objective should be that the echo return loss distribution for all types of toll-connecting trunks (TCT) should achieve or exceed a distribution with 50 percent $\geqq 22$ dB, and none less than 16 dB. All trunks connected to digital offices must meet this objective. This should be viewed as a long-term goal for other offices. Currently, two-wire trunks on analog switching systems will be allowed to meet their present lower requirement of 50 percent $\geqq 18$ dB, and none less than 13 dB.

(*f*) The present through-balance requirement at a two-wire switching point appears satisfactory. The through-balance requirement states that the echo-return loss distribution should achieve or exceed a distribution with 50 percent $\geqq 27$ dB, and none less than 21 dB.

(*g*) In general, echo suppressors should be applied at a trunk length where the improvement in transmission quality outweighs their inherent risks. Based on these considerations, echo suppressors in both the analog and digital network should be applied on

Table I — Regional center-regional center echo suppressor application rules

Echo suppressors should be applied on all RC-RC trunks except:

White Plains—Wayne
White Plains—Pittsburgh
White Plains—Rockdale
White Plains—St. Louis
Wayne—Pittsburgh
Wayne—Rockdale
Wayne—St. Louis
Pittsburgh—Norway
Pittsburgh—Rockdale
Pittsburgh—St. Louis
St. Louis—Dallas
St. Louis—Norway
St. Louis—Rockdale
Dallas—Pittsburgh
Dallas—Rockdale
Dallas—San Bernardino
Dallas—Wayne
Dallas—White Plains
Sacramento—San Bernardino

high-usage trunks greater than 1850 route miles in length rather than the present 1565-mile length. Regional-center-to-regional-center trunks should be equipped with echo suppressors, except for those trunk groups between the offices listed in Table I.

III. METHODOLOGY

The review process that led to the above recommendations was achieved through a computer-simulation model which allowed the investigation of the optimal trade-off between the degradation due to talker echo and the degradation caused by its control on the received speech. The approach used in the computer model was to duplicate within the computer the routing and transmission characteristics that could have occurred on a sample of actual calls in the network. From this information, estimates of the distributions of loss, noise, echo-path loss, and delay were obtained for various approaches to the control of talker echo. The merits of each approach were evaluated by using the transmission parameter distributions obtained from the model to predict customer opinion of quality of service. Estimates of customer opinion were obtained by subjective tests[1] in which customers rated a call having a given set of transmission parameters on a scale of "excellent," "good," "fair," "poor," and "unsatisfactory." By combining this information with the distributions of occurrences of the parameters, estimates were obtained of the expected percentage of customers who would rate a call "good or better" (good and excellent) or "poor or

worse" (poor and unsatisfactory) if a large number of calls are made. These estimates are referred to as "grade-of-service."

A call using the DDD network involves the telephone set on the customer's premises, the communication path (the loop) to his local Class 5 office, a number of interconnected trunks to the far end Class 5 office, the far end loop and telephone set. The path between Class 5 offices chosen through the network is dictated by the Network Switching Plan. In this plan, the continental United States is divided currently into ten regional center areas. Within each region, the offices are interconnected in a hierarchical manner by final trunk groups. The most general hierarchical structure is illustrated in Fig. 1. In many cases, calls from a Class 5 office may be routed directly to a Class 1, 2, or 3 office and the chain of offices may not contain all five classes of offices. Regions are connected to another region by high-usage trunk groups or RC-RC final trunk groups. A high-usage trunk group may be established between any two offices whenever sufficient traffic exists.

The actual route chosen through the network depends on the availability of trunks at each of the offices in the hierarchical chain. At a

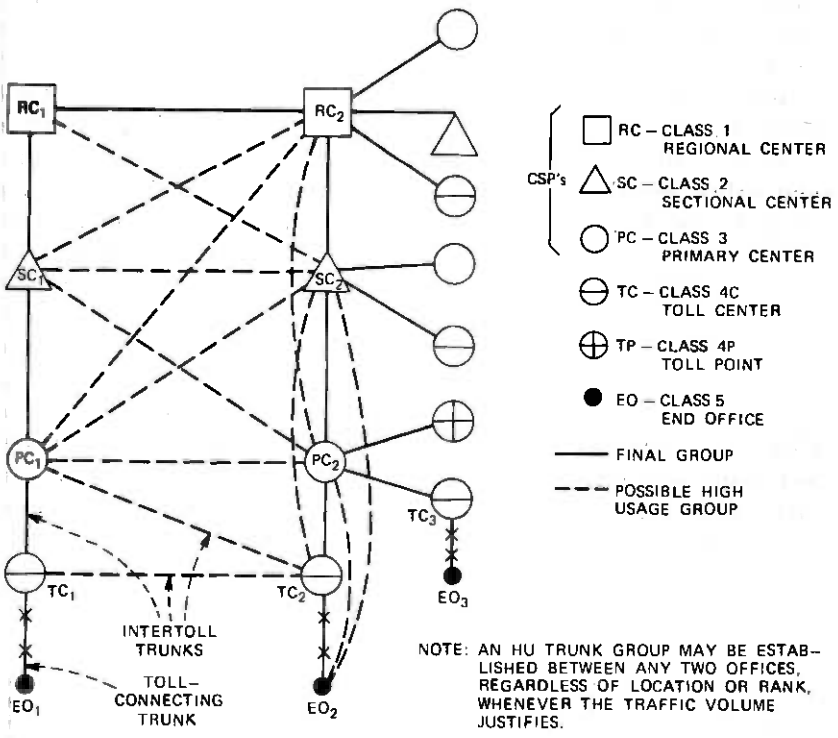


Fig. 1—Hierarchical switching plan for DDD network.

given office, preference is given to the high-usage trunk group to the lowest-ranking office in the distant region. If all of the trunks in that group are busy, a high-usage trunk group to the next highest office in the distant region is chosen if it exists. If the high-usage trunk groups to all the offices in the distant region are busy, the final trunk group to the next office in its own hierarchical chain is chosen.

Figure 2 shows the hierarchical chains and the high-usage trunk groups that actually could be used for a call from the Brooksville, Florida-to-Cincinnati, Ohio, Class 5 offices. There are four possible routes, in order of their probability of occurrence:

(*i*) Brooksville–Jacksonville–Cincinnati,
(*ii*) Brooksville–Jacksonville–Pittsburgh–Cincinnati,
(*iii*) Brooksville–Jacksonville–Rockdale–Cincinnati,
(*iv*) Brooksville–Jacksonville–Rockdale–Pittsburgh–Cincinnati.

The simulation model duplicates with the computer the routing and the transmission properties of the Class 5 to Class 5 portion of a sample of 1500 calls made by customers throughout the United States.[6] Because each call has the possibility of several routes, it is impractical to determine the precise route used for a particular call. The approach used was to incorporate in the model all of the routes that each call could have taken and weigh the results obtained from each of these routes by an estimate of the probability of occurrence of a given route.

These routing data provide information as to the number of trunks and airline* length of each trunk in a connection from the originating to the terminating Class 5 office. It does not, however, contain information as to the loss, noise, and delay that occur on these trunks. Fortunately, trunk surveys have indicated that the transmission properties of a trunk are determined more by trunk length than by its geographical route. Thus, considerable simplification was obtained without much decrease in accuracy by assigning to each trunk the transmission characteristics derived from system-wide trunk statistics.[5,7,8] Mathematically, the loss, noise, and delay characteristics were expressed as normally distributed random variables with means and standard deviations that are functions of the length of the trunk. These functions are shown in Fig. 3.

In developing the estimated performance of the connections, the model individually evaluates each of the routes. Given the basic in-

* The airline length of a trunk is the straight line distance between two offices. The route length of a trunk is the actual length of the trunk and is always longer than the airline length. The model estimate of the ratio of the route length to airline length is given as part of Fig. 3. The airline length of an entire connection is the straight-line distance between local Class 5 offices.

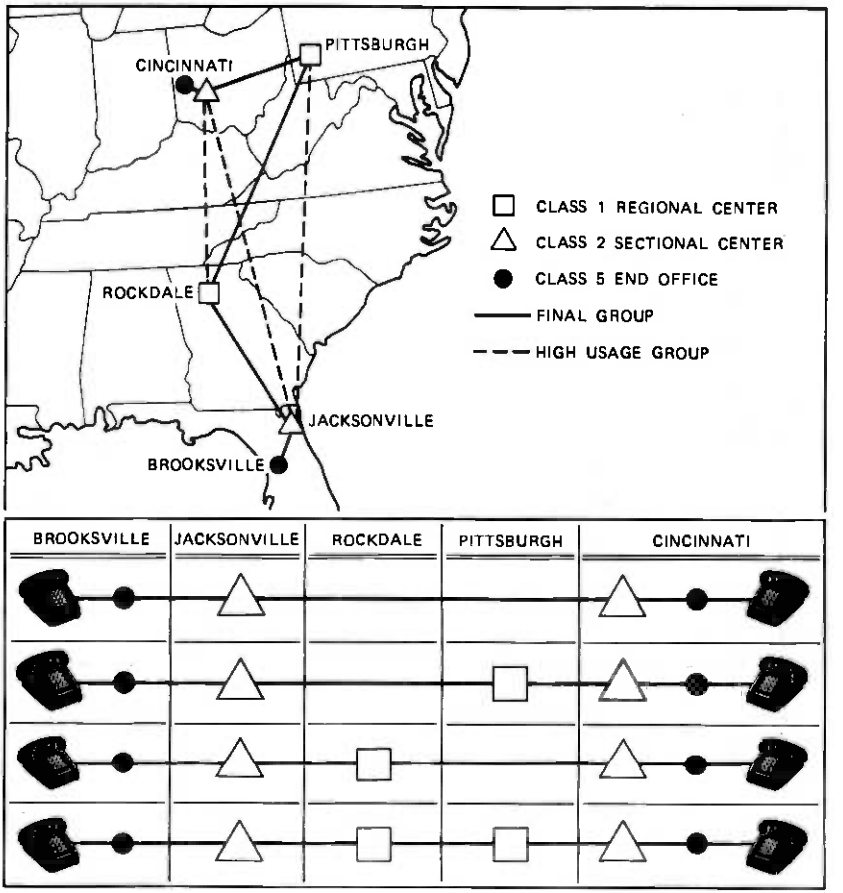


Fig. 2—Actual routing of call between Brooksville, Florida and Cincinnati, Ohio.

formation on the number of trunks and distances between offices, the model selects a random value from the appropriate distributions of delay, noise, and loss for each trunk. The delay value is added to the sum of delays from the previous trunks, the noise value in power is added to the previous noise value and attenuated by the loss value, and the loss is summed in dB to the sum of previous loss values. The process is then repeated for all trunks of a connection to obtain one estimate of the loss, noise, and delay which could occur on this connection. To obtain estimates of the range in values that could occur, the process of selecting values and summing is repeated many times, the exact number depending on the probability of the connection.

Once having an estimate of the entire distribution, estimates of the echo-path loss, the noise-loss grade-of-service, talker-echo grade-of-

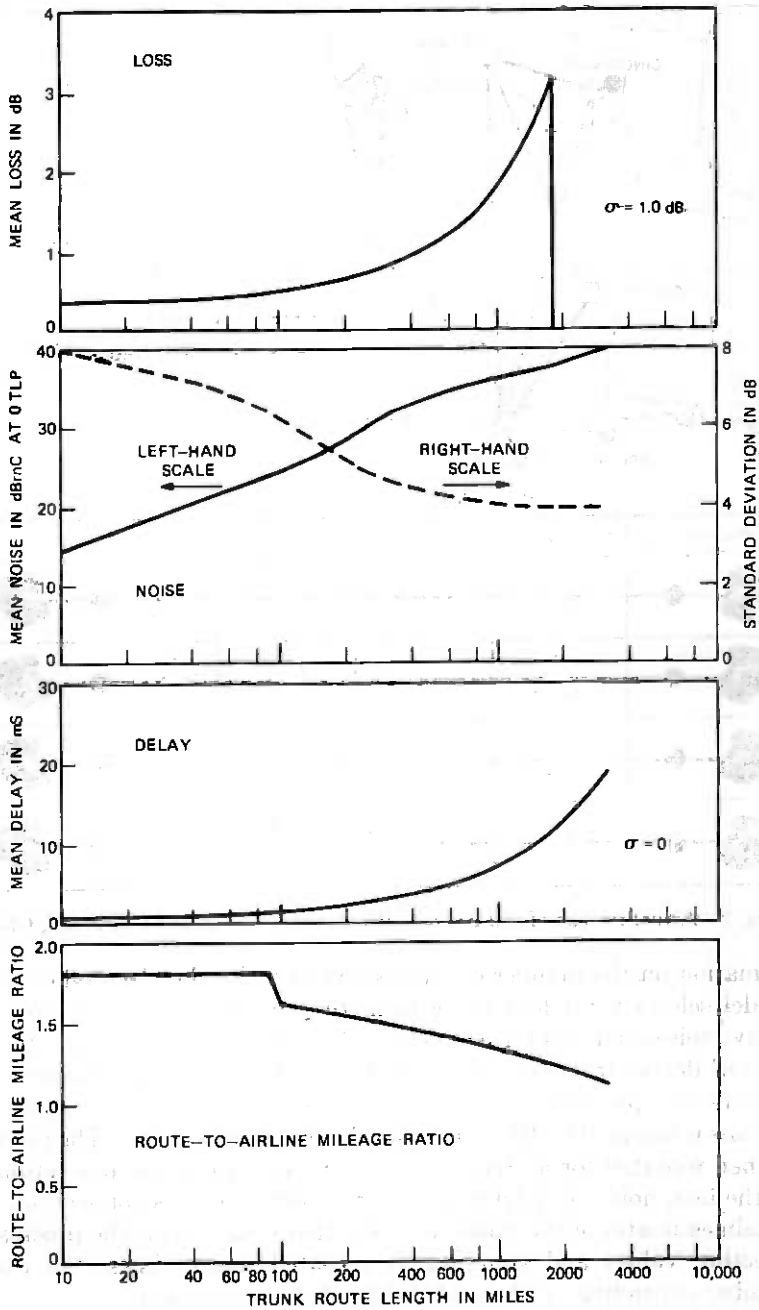


Fig. 3—Assumed trunk loss, noise, delay, and route-to-airline ratio for simulation model.

service, and noise-loss-echo grade-of-service are calculated using the subjective test results reported in Ref. 1. The parameters of the subjective test models were obtained from the loss, noise, and delay results by the formulae given in Fig. 4. In these computations, the average characteristics of loops and telephone sets were used since the emphasis in this study was on the change in the transmission characteristics of the Class 5 to Class 5 portion of the network. The echo-path loss measures the amount of attenuation of the echo energy reflected back to the talking customer. It includes the effects of the echo reflected at the impedance mismatch between the toll-connecting trunk and the customer loop at the Class 5 office and the echo reflections occurring

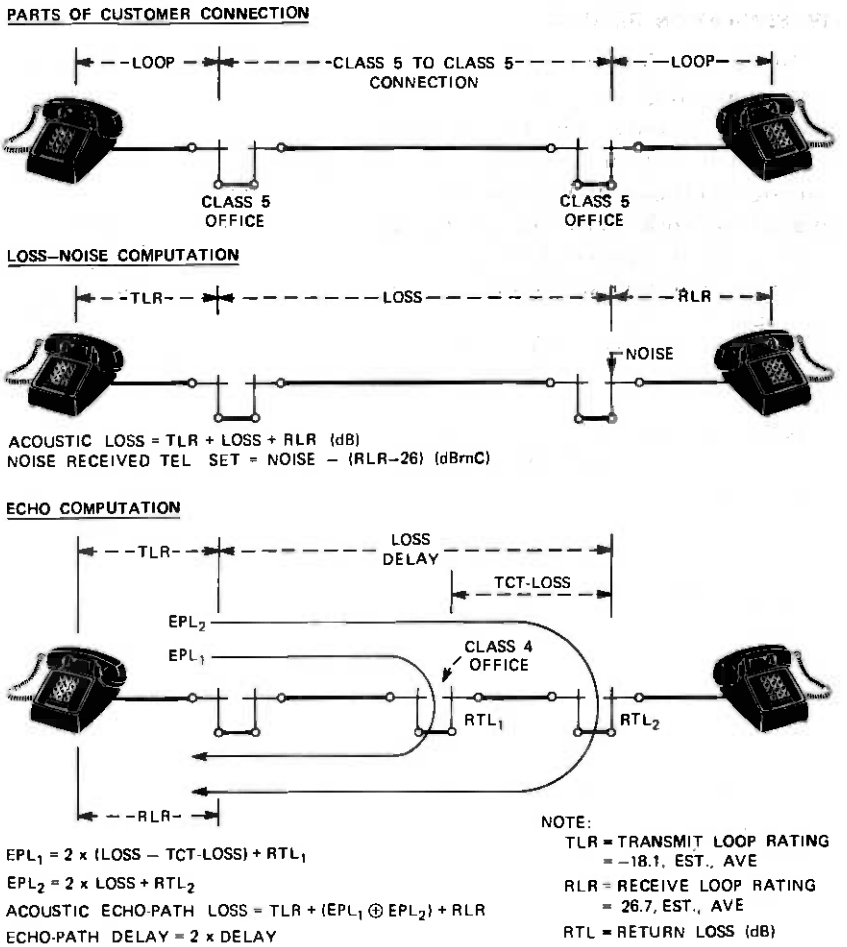


Fig. 4—Formulae used for deriving the parameters of the subjective test models.

in the trunk between the Class 4 toll office and the Class 5 office as measured by terminal balance procedures.

A number of assumptions were made in developing the model. Indeed, an important concept in modeling is to provide just sufficient detail to provide valid results. Too much information complicates the model and increases the computation time and cost. An indication of the validity of the model is obtained by comparing the loss, noise, and delay results with those of the 1969/1970 Connection Suvery[9] and the 1972 Echo Survey.[5] As shown in Tables II and III, the mean and standard deviation predicted by the model agree in all length categories within the statistical accuracy of the model and the surveys.

IV. SIMULATION RESULTS

In the near future, it will be possible to form a connection by using analog and digital facilities switched together via analog and digital switching systems. The resulting connection will be analogous to a connection formed over (*i*) a purely analog network consisting of analog facilities switched via analog switching machines, (*ii*) a purely digital network consisting of digital facilities switched digitally, or (*iii*) a mixed analog-digital network with both types of facilities and switches. This review examined the appropriateness of various methods of talker echo control for each of the analogous three types of networks. For convenience, the results given in this section are grouped according

Table II — Comparison of estimated loss and noise from model and 1969/1970 Connection Survey

A. Loss

Connection Length (Airline Miles)	Model		1969/1970 Survey	
	Mean (dB)	SD (dB)	Mean (dB)	SD (dB)
0–180	6.5	1.9	6.5 ± 0.7	2.0
180–725	7.6	2.1	7.3 ± 0.4	2.3
725–2900	7.8	2.5	7.7 ± 0.5	2.5

B. Noise

Connection Length (Airline Miles)	Model		1969/1970 Survey	
	Mean (dB)	SD (dB)	Mean (dB)	SD (dB)
0–180	21.9	6.0	18.7 ± 3.7	8.3
180–725	28.8	4.2	29.3 ± 0.6	3.1
725–2900	32.4	4.2	32.7 ± 0.6	3.5

Table III — Comparison of estimated round-trip delay and echo path loss from model and the 1972 Echo Survey

A. Round-Trip Absolute Delay

Connection Length (Airline Miles)	Model		1972 Echo Survey (1000-Hz Delay)	
	Mean (dB)	SD (dB)	Mean (dB)	SD (dB)
180–360	11.4	3.53	11.7 ± 0.8	3.4
360–725	16.4	5.51	16.4 ± 0.5	3.8
725–1450	23.0	5.77	24.8 ± 2.1	5.0
1450–2900	36.2	5.30	37.3 ± 1.3	6.1

B. Echo Path Loss

Connection Length (Airline Miles)	Model		1972 Echo Survey	
	Mean (dB)	SD (dB)	Mean (dB)	SD (dB)
180–360	22.7	4.4	23.1 ± 1.6	5.7
360–725	23.9	4.5	24.3 ± 2.2	6.8
725–1450	25.2	4.9	24.6 ± 2.2	6.3
1450–2900	21.0	4.6	23.3 ± 2.1	6.4

to the type of network. Within each of these three groups, control of talker echo by the use of loss, balance, and echo suppressors is discussed.

4.1 Analog network

4.1.1 Optimal loss

Loss in a connection reduces the amount of talker echo returned to the talker; however, it also attenuates the speech received from the far-end customer. Thus, there is a trade-off between the two types of degradation. Actually, a customer's opinion of his call is based on a joint assessment of these effects, since he experiences both effects during portions of his conversation. This joint assessment is measured through the loss-noise-echo grade-of-service estimates.

Quantitatively, the good or better value of this grade-of-service estimate will be slightly lower than the value of either the good or better loss-noise grade-of-service or the talker-echo grade-of-service which is individually lowest. This effect can be seen by examining Fig. 5. This figure contains an estimate of the percentage of customers who would rate a call good or better in terms of the acoustic loss of an end-to-end connection with average loops and a Class 5 to Class 5 connection of about 1300 airline miles. Also plotted are the individual loss-noise grade-of-service and echo grade-of-service. At low values

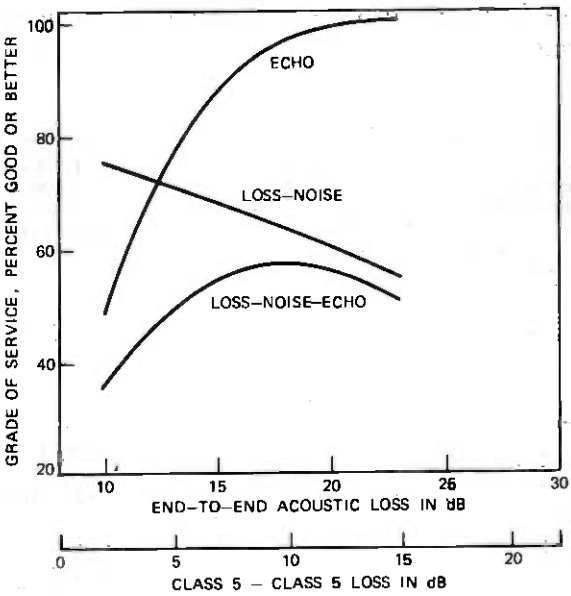


Fig. 5—Optimal loss for a 1270-airline-mile connection.

of loss, the value of the loss-noise-echo grade-of-service is determined by the echo grade-of-service. As the loss is increased, it follows the improvement in the value of the echo grade-of-service until the echo value exceeds the loss-noise grade-of-service. For loss values higher than this value, the loss-noise-echo grade-of-service follows the loss-noise grade-of-service.

Thus, the good or better loss-noise-echo grade-of-service as a function of loss increases to some maximum value and then decreases. The value of loss at which the good or better loss-noise-echo grade-of-service is a maximum is defined as the optimal loss. Because of the functional relationship between the good or better and poor or worse grade-of-service estimates, the optimal loss is also the value of loss which minimizes the poor or worse loss-noise-echo grade-of-service. For this particular example, the optimal loss is 18-dB acoustic loss, or about 10-dB connection loss between Class 5 offices.

The value of optimal loss for a given connection will depend upon the noise and delay associated with that connection. The value of loss will therefore vary between connections. A feeling for the range in optimal loss is obtained by plotting, as a function of airline length of the connection, the value at which 10, 50, or 90 percent of the connections would have an optimal loss value less than that value. The optimal loss for connections using analog facilities is shown in Fig. 6.

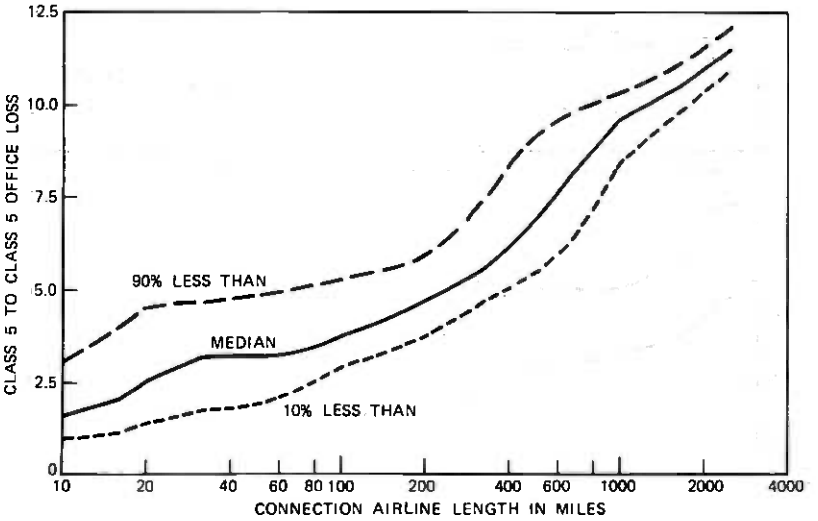


Fig. 6—Optimal loss for connections routed over analog network.

This figure shows that, for short distances, very little added Class 5 to Class 5 loss is desirable. As the length of the connection is increased, more loss is required to counteract the effect of increased delay on the talker-echo portion of the joint grade-of-service.

The grade-of-service obtained using this optimal loss is shown in Fig. 7 with the first graph showing, as a function of the connection airline distance, the talker-echo grade-of-service without echo suppressors being applied, and the next two graphs showing the loss-noise and loss-noise-echo grades-of-service. The set of curves plotted on each graph indicates, as a function of airline length, the value at which 10, 50, or 90 percent of the connections would have a grade-of-service greater than that value. For long connections, the loss-noise grade-of-service value is primarily due to noise, not loss. Some improvement in loss-noise grade-of-service would be possible with less loss. However, it would be small in comparison to the degradation in echo grade-of-service; thus, the optimal loss value tends to favor the control of echo.

These curves also indicate that satisfactory talker-echo grade-of-service can be obtained with loss control for all length connections. However, as is discussed later, somewhat better grade-of-service can be achieved on long connections by means of echo suppressors.

4.1.2 Loss allocation

The optimal loss value increases with increased end-to-end connection length. This variation is achieved in the network by allocating

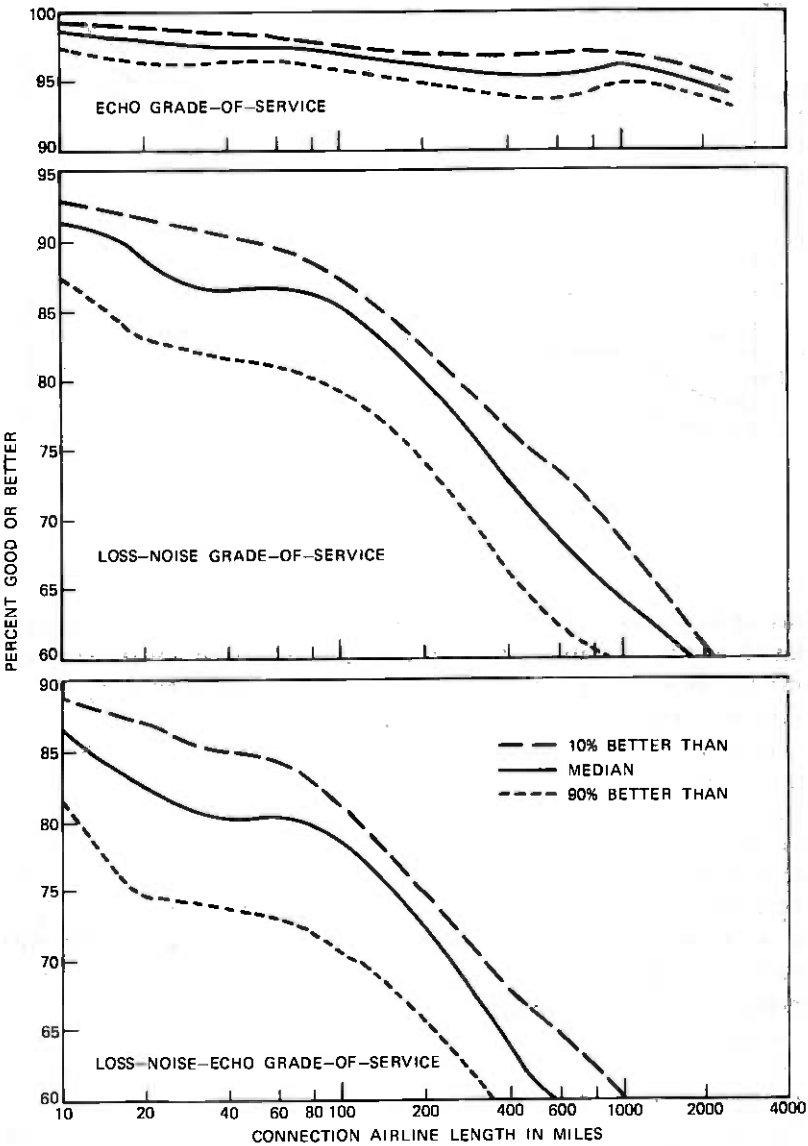


Fig. 7—Grades-of-service obtained for connections routed over the analog network if it were designed with optimal loss.

loss to the individual trunks of the connection such that the proper amount of loss is obtained when they are connected together. However, the trunks are usually used for both short connections and long connections. Any allocation plan must, therefore, be devised such that the

proper amount of loss is obtained when these trunks are used in all length connections. At the same time, the allocation plan must also be easily implemented and administered. In general, these constraints mean that the actual connection loss can only approximate the optimal loss.

Practically, there appear to be three types of approximate allocation approaches:

(*i*) Fixed Loss—All connections are designed to have the same connection loss. An example of this plan is the switched digital network plan discussed in Section 4.2.

(*ii*) Compensated Loss—Trunks are assigned just enough added loss to compensate for their increased delay. An example of this plan is the current VNL plan.

(*iii*) Overcompensated Loss—The longer trunks are assigned more loss than needed to compensate for the delay of short trunks which normally occur as part of a longer connection.

In the remaining parts of this section, we examine the advantages and disadvantages of these three types of plans for use in the analog network.

4.1.2.1 Fixed loss plan. In a fixed loss plan, all connections have the same loss, irrespective of their length. This requires somewhat of a compromise with the optimal loss policy, which indicates that long connections should have more loss than short connections. A fixed loss plan is a satisfactory plan only if the increase in optimal loss with distance is not too large. Figure 8 shows the optimal loss for the analog network as a function of length of the connection. Two lines are drawn on this curve to represent two possible values for fixed loss. The higher value near 9 dB would approximate the optimal loss for longer connections out to some mileage where an echo suppressor would be used. However, it would provide 3 to 6 dB too much loss for shorter connections. The lower value of loss of 6 dB would provide a much better match to the required loss on short connections, but would provide insufficient loss for the longer connections. Thus, a fixed value of connection loss does not appear suitable for the analog network. Some variation in loss is needed with the length of the connection. This type of plan, however, is being adopted for the switched digital network because its optimum loss has less variability than that of the analog network. This is discussed in more detail in Section 4.2.1.

4.1.2.2 Compensated loss plan. In a length-compensated loss plan, loss is inserted in each trunk in proportion to the added amount of delay introduced by that particular trunk. Since the variation in optimal

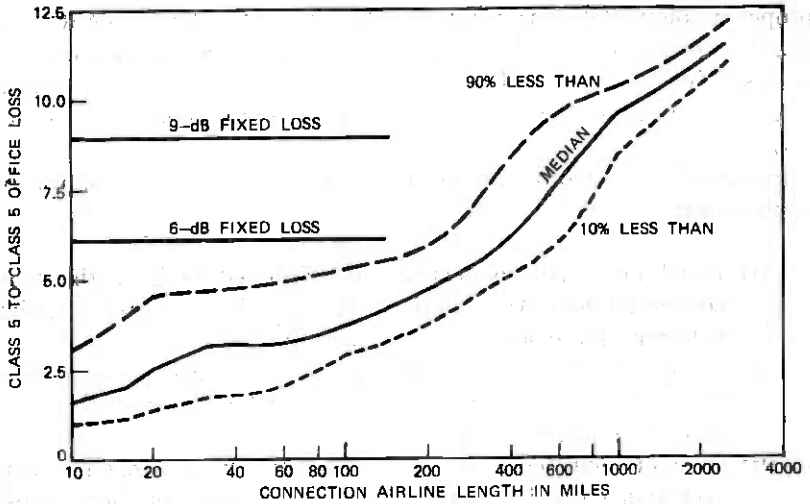


Fig. 8—Analog network connection loss of two fixed loss plans compared with optimal loss.

loss is primarily due to this effect, this type of plan provides a very good approximation to the desired loss.

The current VNL design plan is a prime example of this type of plan. The VNL plan has two classes of trunks, intertoll (ITT) and toll-connecting (TCT). The loss of intertoll trunks is determined for carrier trunks by the VNL formula (VNL = 0.4 + 0.0015 × trunk-length), while the loss of toll-connecting trunks is 2.5 dB plus that given by the VNL formula. The 0.0015-dB loss added for each mile in the carrier trunk VNL formula approximately compensates for the added propagation delay. The 0.4-dB factor which was originally introduced to compensate for loss variability in effect compensates for the delay of the carrier terminals.

Figure 9 is a comparison of the optimum loss with the loss obtained on connections using the VNL loss plan.* This figure indicates that the VNL plan provides slightly too much loss for short connections but is about optimal for long connections. Comparison with the median optimal loss for connections greater than 1000 miles is shown in Fig. 10. This graph shows the optimal loss for mean values of 18 and 22 dB and infinite return loss as measured by terminal balance procedures at the Class 4 office. The graph indicates that, with a mean value of 18-dB terminal balance return, the current network performance, the optimal value of loss should be about 0.5 dB greater than that supplied

* In these comparisons, it is assumed that no echo suppressors are used and the loss given by the VNL formula is used for all length intertoll trunks.

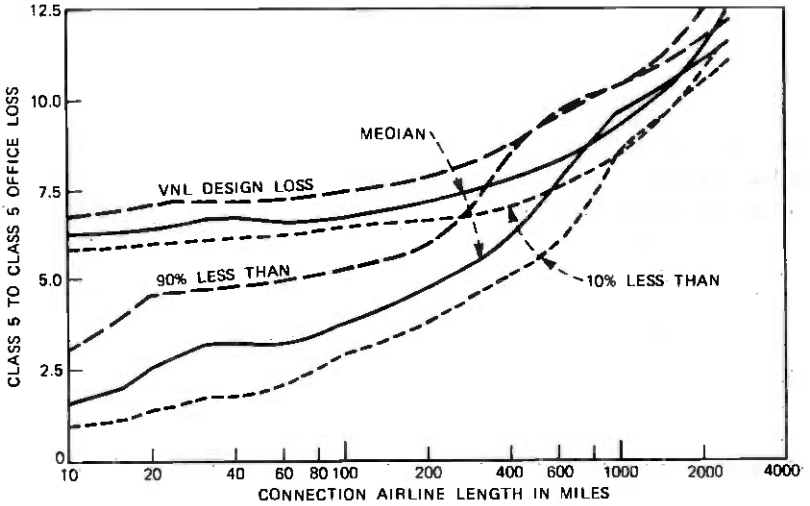


Fig. 9—Analog network connection loss obtained using VNL loss plan compared with optimal connection loss.

by VNL design. In general, the loss for VNL design corresponds most closely to the optimal loss with a mean return loss of 22 dB. This supports the long-term objective of a mean return loss at the Class 4 office of 22 dB as discussed in Section 4.1.3.

Although the VNL plan provides more than optimum loss for short connections, the difference is not sufficiently great to have any appreciable effect on grade-of-service. This can be seen by comparing

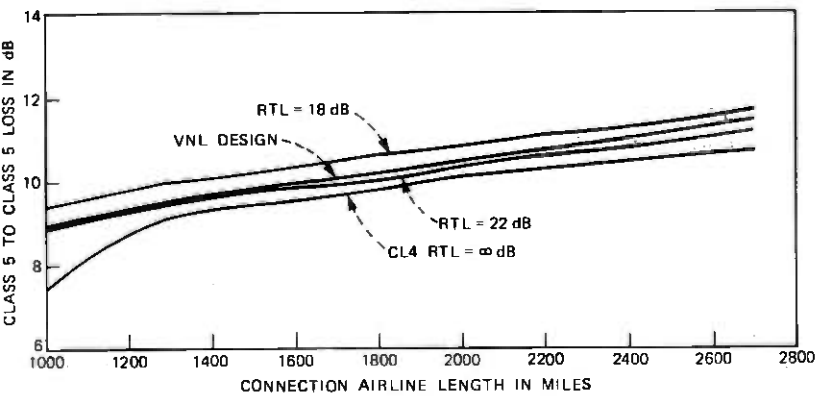


Fig. 10—Median analog network connection loss for connections with lengths greater than 1000 miles compared with median optimal connection loss obtained with average Class 4 terminal balance of ∞, 22, and 18 dB.

the talker-echo grade-of-service and the loss-noise grade-of-service curves using the VNL plan as given in Fig. 11 with those using optimal loss given in Fig. 7. Thus, the VNL plan is a very satisfactory plan for loss allocation in terms of providing nearly optimal performance for the analog network.

4.1.2.3 Overcompensated loss plan. There is one basic problem with the VNL loss plan. The realization of the increased loss with distance tends to make this plan difficult to administer. Currently, the 0.0015-dB-per-mile increase in loss is approximated by a 0.3-dB step about every 200 miles up to 1850 miles, the echo suppressor application length. This means that there are nine loss steps. Trunks having all these steps could exist in any given toll office. To initially install and maintain these trunks, the design loss of each individual trunk must be determined from record information. An overcompensated loss plan would simplify this process by reducing the reliance on recorded information by decreasing the number of loss steps. In such a plan, the longer trunks have more loss than is actually needed to compensate for the delay introduced by short trunks. The short trunks can then be operated at near-zero loss with improved performance when they are used in short connections.

Examination of the optimal loss curves and the facility types used for toll-connecting trunks indicated that about 6 dB of end-to-end loss is about optimum for short connections and approximately 9 dB is optimum for long connections. In addition, it is desirable for ease of conversion not to change the loss of toll-connecting trunks. With VNL design, toll-connecting trunks have an average loss of 2.9 dB which, for the new plan, will be assumed to be rounded to 3.0 dB.

Numerous possible plans could satisfy these broad guidelines. The differences between the various plans is the manner in which the variation in loss between short and long connections is achieved, i.e., the number and size of the loss steps and their application distances. Of the more than half-dozen plans evaluated with the simulation model, the most promising alternative to VNL was the simplest plan with only one step. Other plans with more steps did not appear to have any appreciable grade-of-service advantage to compensate for their added administrative complexity.

The most promising plan was to operate intertoll trunks less than 600 route miles at 0 dB loss, and trunks greater than 600 miles at 3 dB loss. Figure 12 shows the change in loss-noise grade-of-service, echo grade-of-service, and loss-noise-echo grade-of-service compared to that obtained for the VNL plan. This indicates that the loss-noise grade-of-service would be improved by about 3 percent, while the echo grade-of-

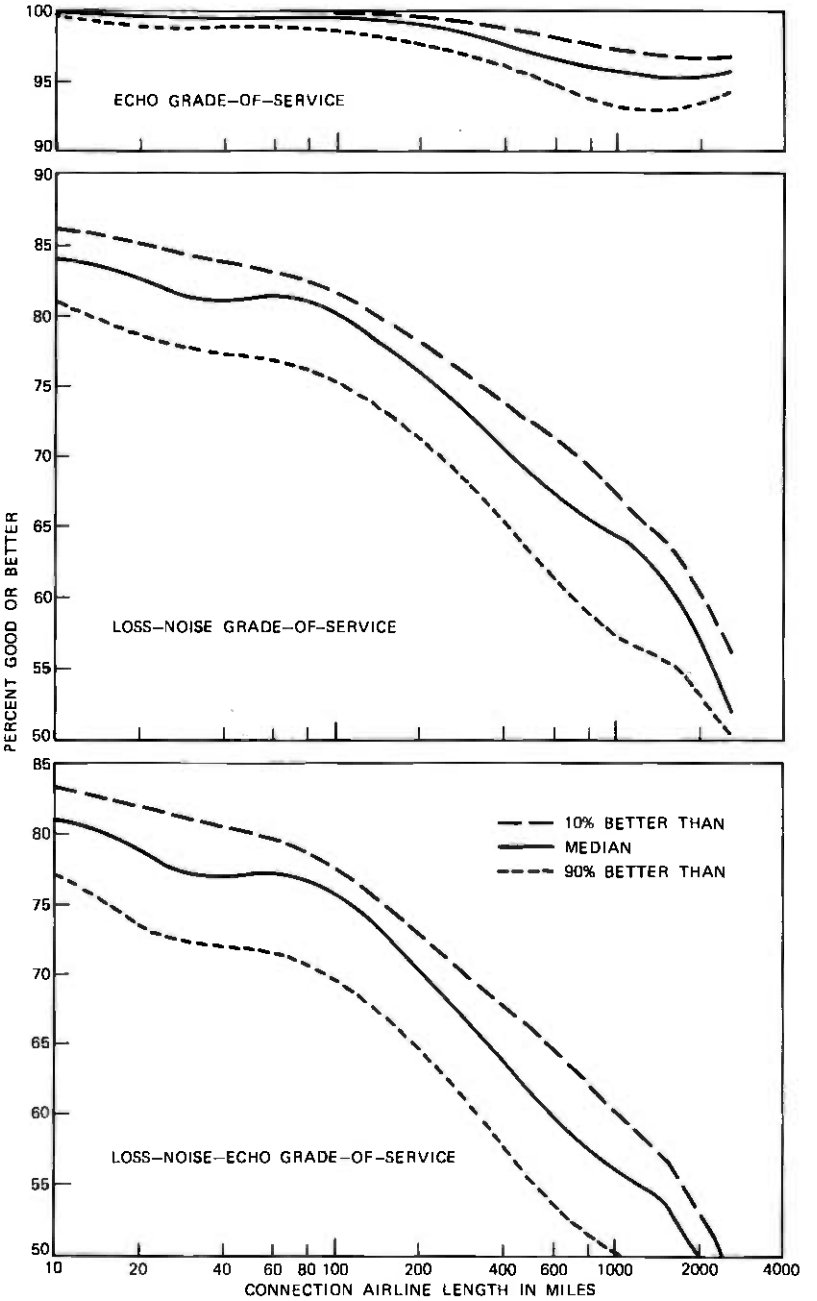


Fig. 11—Grades-of-service for analog network using VNL loss design. Loss inserted in all length trunks—no echo suppressors.

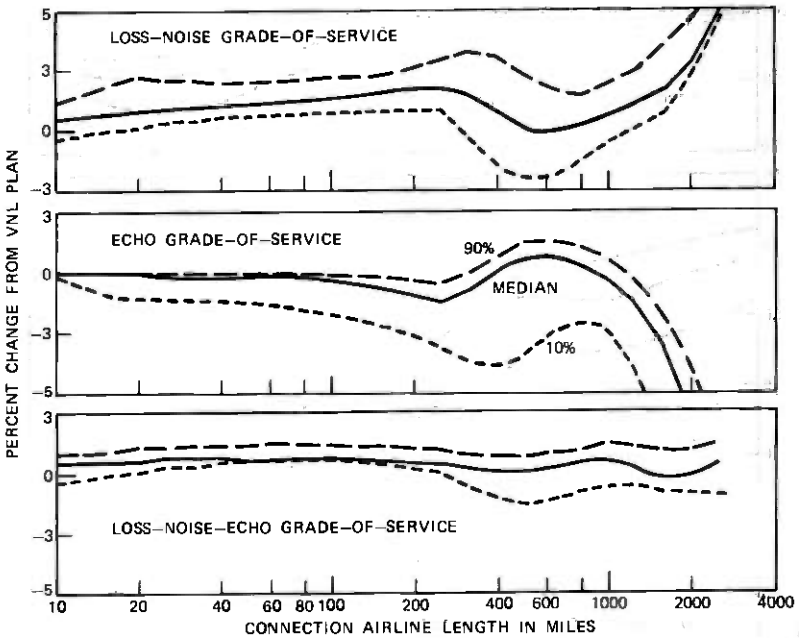


Fig. 12—Change in grades-of-service for possible new loss plan having intertoll trunks designed to 0 dB up to 600 miles and 3 dB for trunks greater than 600 miles.

service is degraded by about the same amount for connections greater than about 300 miles. Loss-noise-echo grade-of-service shows a modest improvement over that obtained by the VNL loss plan.

The simulation study, therefore, indicated that this plan was satisfactory. However, the simulation model predicts most accurately the average performance, not the performance that might occur on alternate routes with very low probability of occurrence, such as calls routed through sectional and regional centers. Although some performance degradation on these types of calls can be allowed because of their low probability of occurrence, extremely poor loss-noise due to multiple trunks with 3-dB loss, or poor echo performance due to the delay of a large number of carrier terminals, must be avoided.

An investigation indicated that multiple 3-dB trunks (trunks greater than 600 miles) could occur on calls routed through sectional centers toward several of the regional centers. This investigation also indicated that these types of calls would often have, with the new plan, an echo grade-of-service of around 70 percent good or better and a very high reliance on properly operating echo suppressors when they were used.

Because of these factors, the plan was altered to assign:

Intertoll Trunks	Loss	Mileage
Intraregional	0.5 dB	All lengths
Interregional	0.5 dB	<1000 mi.
	3.0 dB	>1000 mi.

Figure 13 shows the change in loss-noise grade-of-service, echo grade-of-service, and loss-noise-echo grade-of-service compared to that obtained using the VNL plan. These curves indicate that it provides essentially the same loss-noise-echo grade-of-service as the VNL plan at all connection lengths with some improvement in loss-noise grade-of-service at the expense of the echo grade-of-service. Since the joint loss-noise-echo grade-of-service is essentially the same, this plan is a suitable alternative to the VNL plan. Its main attraction is its administrative simplicity. It also has the advantage that only those trunks greater than 165 miles, about 44 percent, would have to be changed from current VNL design. However, at the present time there appears to be no appreciable long-term administrative economic benefit to

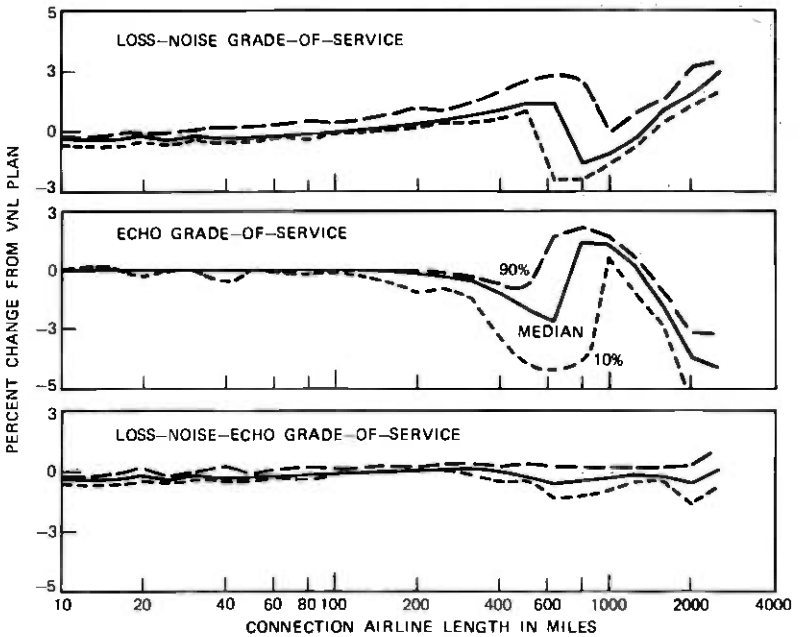


Fig. 13—Change in grades-of-service for the modified new loss plan having intertoll trunks designed to 0.5 dB up to 1000 miles and 3 dB for trunks greater than 1000 miles.

this plan. Since there would be some expense and some administrative difficulties during a changeover to this plan, continued use of the Via Net Loss plan was recommended.

4.1.3 Balance

The loss design plans assume that the predominant echo reflection occurs at the impedance mismatch of the toll-connecting trunk and the customer loop with any additional sources of echo being controlled by through and terminal balance requirements. The appropriateness of the present echo through and terminal balance requirements was reviewed as part of this study.

Intertoll trunk transmission facilities are designed on a four-wire basis that prevents intermediate echoes. However, many switching machines used for interconnection of intertoll trunks are two-wire and, therefore, require hybrids to effect the necessary four-wire to two-wire conversion. The amount of echo returned at these conversions depends on the impedance mismatch of the hybrids and the office cabling. This impedance mismatch is controlled by a through balance requirement which states that the distribution of echo return loss for the office shall achieve or exceed a statistical distribution having a mean of 27 dB and a minimum of 21 dB.

If two-wire offices are used for Class 3 or lower toll offices, the echo returned from an office just meeting this requirement will be 10 dB lower than that from other sources. This amount will have a negligible effect on the overall echo performance. Also, the requirement can usually be met or exceeded in current offices. Thus, the requirement appears to be satisfactory.

The effect on talker-echo performance of impedance irregularities at the distant toll office, and up to and within its Class 5 office, is controlled by terminal balance requirements. The current requirements allow a distribution of the terminal balance return loss values within an office with the median and minimum of the distribution being equal to or exceeding

	Median	Minimum
Two-wire facilities	18 dB	13 dB
Four-wire facilities	22 dB	16 dB.

The estimated mean value of the current network derived from the echo survey[6] is 18 dB.

In establishing a connection, a toll-connecting trunk having a given value of terminal balance return loss is assigned to the connection. Thus, on any given connection a customer experiences the effect of some return loss value from those which are allowed. The effect of a

specific value was evaluated by assuming that all connections in the model had that specific value. The grade-of-service obtained was then compared with the ideal situation with no additional echo returned from Class 4 offices. This comparison indicated that the good or better echo grade-of-service with VNL loss design was only minimally affected by a Class 4 return loss of 22 dB and about 2-percent decrease for 18 dB. Of more importance, however, is the effect of the allowed lower values. In this case, a 3-percent decrease occurred for a return loss of 16 dB and 8-percent decrease occurred for 13 dB.

Based on this analysis, it appears that the four-wire facility requirement is satisfactory. However, it is undesirable to have 50 percent of all calls using two-wire facilities experiencing a 3- to 8-percent decrease in echo grade-of-service. Some improvement appears warranted with all toll-connecting trunks ultimately meeting or exceeding the four-wire facility requirement of a distribution with a median terminal-balance echo-return loss of 22 dB, minimum 16 dB. Thus, this requirement has been adopted as a long-term objective. However, it is recognized that, in many situations, this objective cannot be met currently. The current requirement should still be used for existing facilities, with emphasis placed on reducing the number of trunks having an echo return loss less than 16 dB.

4.1.4 Echo suppressor application rules

As indicated in the discussion of the optimal loss, loss can be used to control talker echo for all length connections using terrestrial facilities within the continental United States. However, the loss causes some degradation of the loss-noise grade-of-service. One can reduce the amount of this degradation by introducing echo suppressors. However, they cause some degradation, even under ideal conditions.

In actuality, there is, in addition, a potential risk that an echo suppressor might not be installed properly or maintained properly. In such a case, the degradation from the echo suppressor could be much greater than that caused by added amounts of loss. In general, echo suppressors should be applied at a trunk length which minimizes their inherent risks. To avoid more than one echo suppressor in a connection, echo suppressors should only be applied on trunks between regions, either on high-usage trunks or on regional-center-to-regional-center trunks.

4.1.4.1 High-usage interregional intertoll trunks. The application of an echo suppressor with zero trunk loss causes an improvement of the loss-noise-echo grade-of-service. Figure 14 shows the effect on the connection loss-noise-echo grade-of-service of applying echo suppressors on a high-usage trunk of a given mileage. When an echo suppressor is

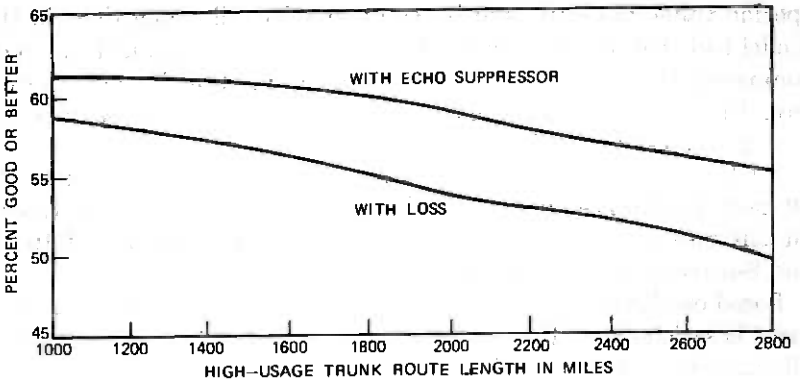


Fig. 14—Median loss-noise-echo grade-of-service for connection greater than 1000 miles when the high-usage trunk is designed according to VNL loss-design plan, or 0 dB loss and an echo suppressor.

applied, the grade-of-service jumps from the curve "with loss" to that value given by the curve marked "with echo suppressors." The latter curve takes into account the grade-of-service effect of nonperfect properly working echo suppressors. As can be seen, the amount of improvement obtained increases as the application distances increase. Figure 15a shows the percent improvement in the loss-noise-echo grade-of-service with the use of an echo suppressor. Figure 15b is a curve of the percentage of trunks existing at that mileage. It indicates that there is about 3-percent grade-of-service improvement if echo suppressors are applied to trunks with length of 1565 route miles. This improvement increases to about 5 percent at 1800 miles. Below 1800 miles, the improvement is less and the number of trunks that require echo suppressors increases markedly, as shown in the bottom part of Fig. 15b. Indeed, there are 50 percent more trunks in the 1600- to 1800-mile range than the 1800- to 2000-mile range. Thus, if echo suppressors were supplied below 1800 miles, there would be a substantial increase in the number of echo suppressors for a smaller, more questionable grade-of-service improvement. Furthermore, there would be a greater potential risk from improperly installed and maintained echo suppressors. For trunks greater than 1800 miles, a 5-percent improvement in average grade-of-service which can be achieved by applying echo suppressors above 1800 miles is considered significant. Thus, for high-usage trunks, it was recommended that a value of approximately 1800 miles be used rather than the previously used 1565 miles. For administrative reasons, a value of 1850 miles was chosen.

When echo suppressors are applied, the trunk loss for trunks greater than 1565 miles was reduced to 0 dB. With the shift of the echo sup-

pressor application mileage to 1850 miles, some nonzero value of loss is needed for trunks between 1565 and 1850 miles. Optimum loss and administrative considerations indicated that a value of loss of either 2.6 or 2.9 dB is required. In general, these values of loss produce the same loss-noise grade-of-service, since at this distance the grade-of-service is mainly determined by the signal-to-noise ratio, while the echo grade-of-service improves with the higher values of loss. Thus, on an overall basis, a value of 2.9 dB was recommended.

Figure 16 shows the resulting loss-noise, echo, and loss-noise-echo grade-of-service for VNL loss design with echo suppressors applied at 1850 miles. The application of echo suppressors improves the loss-noise and loss-noise-echo grade-of-service for calls of lengths greater than about 1000 airline miles.

4.1.4.2 RC-RC intertoll trunks. The majority of traffic switched through a regional center (Class 1) is calls from local Class 5 offices directly homed on that office or alternate route traffic from local offices homed on the next lower office (Class 2). In these cases, the regional office is acting as a Class 4 or Class 3 office. However, because of the hierarchical

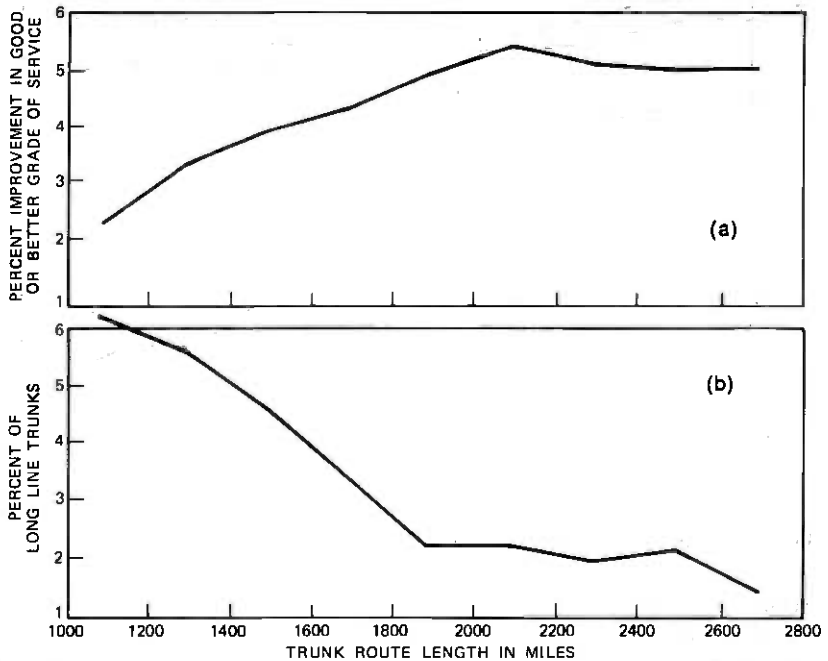


Fig. 15—(a) Percent improvement in loss-noise-echo grade-of-service with echo suppressors. (b) Percentage of trunks requiring echo suppressors if the echo suppressor is applied on high-usage trunks greater than the given length.

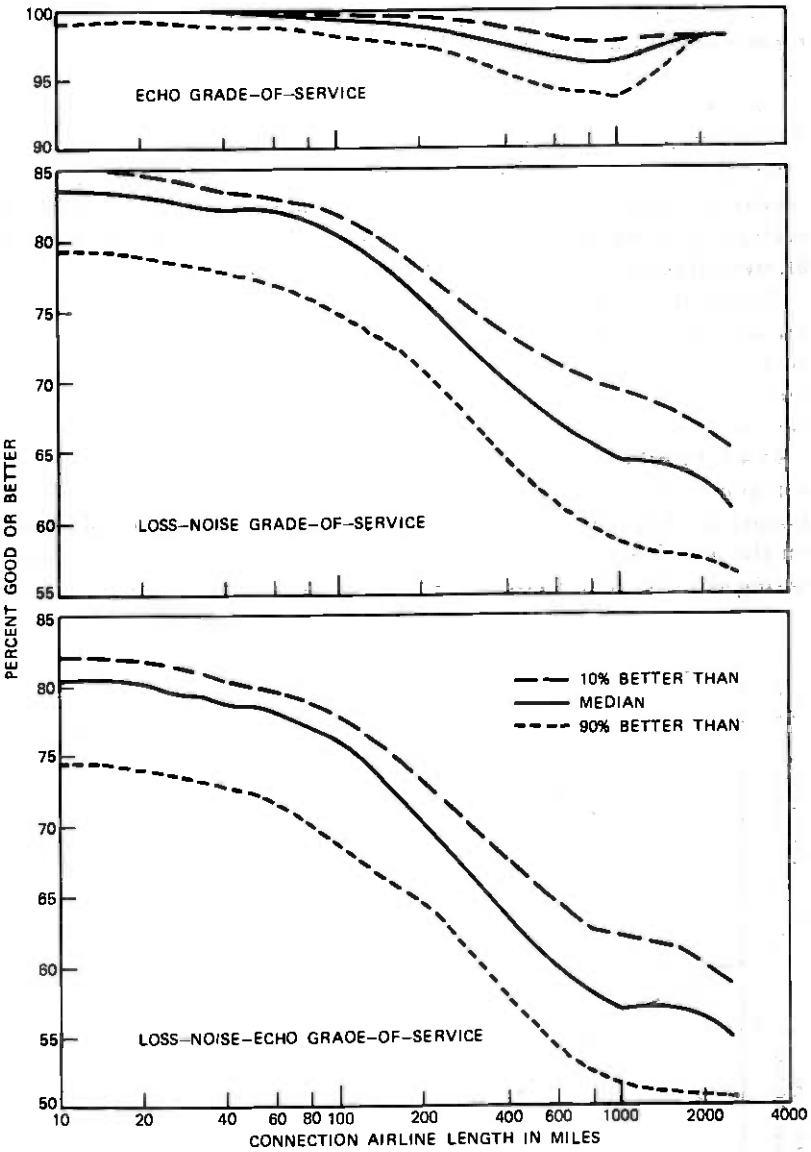


Fig. 16—Grades-of-service for analog network using VNL loss design with echo suppressors applied on high-usage trunks greater than 1850 route miles.

structure of the switching network, there is some traffic where the regional center is part of the final routing of a call originating lower in the hierarchy. These calls will have a larger number of trunks in tandem and greater trunk mileage than calls routed lower in the hierarchy or more directly routed through a regional center.

In general, the more directly routed calls through a regional center should have the same loss-noise-echo grade-of-service as calls routed through lower class offices. Final routed traffic could be allowed somewhat lower loss-noise-echo grade-of-service, but their grade-of-service should not be severely degraded. These considerations tend to require the use of echo suppressors at shorter distances than the 1850 route miles for high-usage intertoll trunks.

As part of this study, the need for echo suppressors was reviewed using the loss-noise-echo grade-of-service approach used for high-usage trunks. This study indicated that the trunk groups between the regional centers shown in Table I do not need echo suppressors and should be designed according to Via Net Loss design rules. All other trunk groups should be designed to 0 dB and have echo suppressors.

4.2 Digital network

With the advent of No. 4 Toll ESS,[3] it is possible to perform direct digital switching of bits coming from digital transmission systems. Economic studies of No. 4 Toll ESS indicates considerable economic advantages in using this approach. However, the VNL loss plan requires loss to be inserted in each trunk. This would require either that the digital signal be decoded to an analog signal, loss inserted, and the signal recoded, or that the encoded signal level be changed by some digital processing technique so that, when it is decoded, a lower signal level would be obtained. Either of these techniques would introduce additional cost and transmission impairments.

An alternative approach is to operate digital intertoll trunks at zero loss so that no conversion is required. This would mean that end-to-end connections over purely digital facilities would have a fixed loss. This section examines the appropriateness of using this approach in terms of the optimal loss concept. Section 4.3 examines the case where a call is routed over portions of both types of networks.

4.2.1 Optimal loss

For the all-digital network, the voice signal is digitally encoded and decoded only at the Class 5 office and only bits are switched at higher-class offices. This means that the noise on a call is that due to one pair of encoders and decoders. The delay has three components; the propagation delay, the delay of one pair of terminals, and the delay of digital buffering within each switching machine. The propagation delay is about the same as for the analog network, but the terminal and switching machine delays, which are about equal, are less than the delay of analog terminals. Thus, the connection delay is less than delay experienced on the same length call in the analog network. Also, the terminal balance return loss at a digital Class 4 office will have an

average of more nearly 22 dB since economic studies indicate cost advantages in using digital carrier at very short distances.

The reduced noise and delay and increased Class 4 average return loss changes the optimal loss. Figure 17 plots, as a function of connection airline distances, the value at which 10, 50, or 90 percent of the connections would have an optimal loss value less than that value. A comparison of this curve with that of the analog network (Fig. 6) indicates that the amount of required loss is reduced. At 1000 airline miles, the 90-percentile optimal loss is reduced from 10.5 to 7.5 dB and the median from 9.5 to 6.5 dB.

The loss-noise, echo, and loss-noise-echo grade-of-service using this optimal loss is shown in Fig. 18. The roll-off in loss-noise grade-of-service is due to the increased loss for control of echo, since the noise is constant with distance. The shape of loss-noise-echo grade-of-service curves is essentially determined by the loss-noise grade-of-service except at longer lengths where a slight additional effect occurs due to the echo grade-of-service.

4.2.2 Loss allocation and level plan

As indicated in Section IV, it is highly desirable to have any required loss inserted before the digital encoder or after the decoder, so that no bit conversion is needed at intermediate offices. This would mean that all connections are designed to a fixed loss value. Since the optimal loss increases with the length of a connection, any fixed loss value will be a compromise between the need for higher loss for long connections and lower loss for short connections. This compromise is more appropriate to the digital network than the analog network, since the range in optimal loss is considerably less. In making this compromise, one would

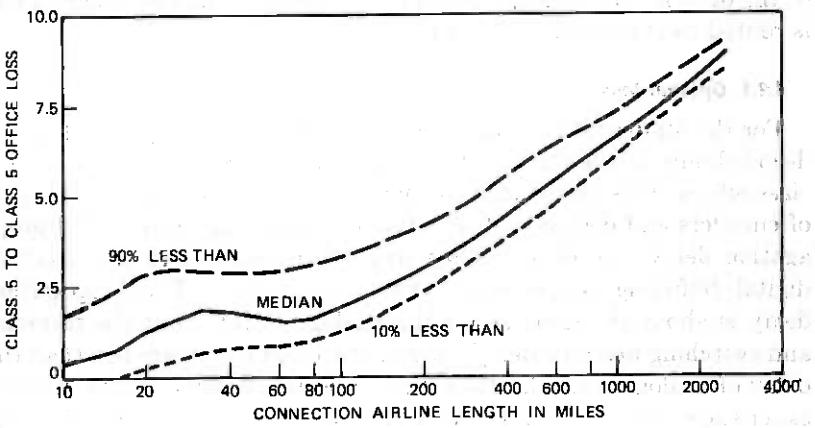


Fig. 17—Optimal loss for connections routed over digital network.

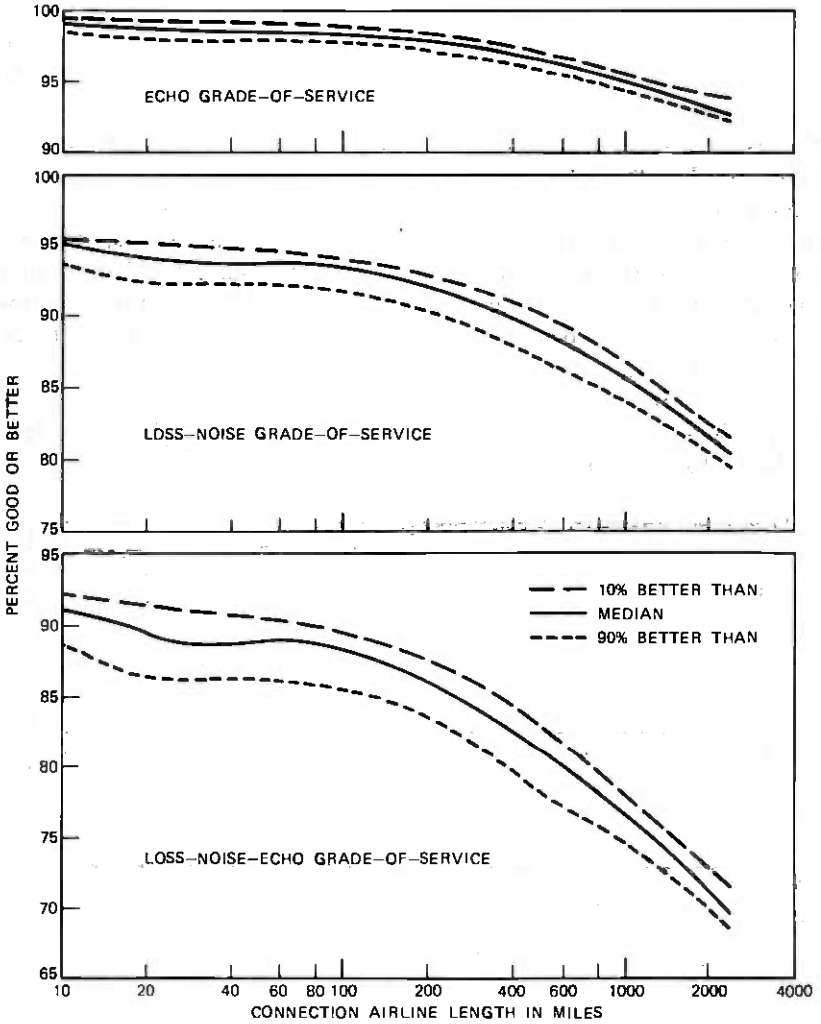


Fig. 18—Grades-of-service obtained for connections routed over digital network if it were designed with optimal loss.

like to have the same loss-noise-echo grade-of-service at about 1000 airline miles as that obtained with optimal loss, and at least as good a loss-noise-echo grade-of-service for short calls as currently obtained with VNL design. In general, since echo grade-of-service for longer calls is quite sensitive to changes in loss, the best compromise value tends to be one near the optimal loss value for longer connections. In the study, various values of loss were investigated with 6 dB of end-to-end loss chosen as best fulfilling the above philosophy.

Figure 19 shows the loss-noise, echo, and loss-noise-echo grade-of-service for connections switched on an all-digital network with 6 dB of end-office-to-end-office loss. This figure includes the effect of echo suppressors which will be discussed in the next section. Comparison of the loss-noise-echo grade-of-service values with those with optimal loss indicates that 6-dB loss provides, for longer connections, about the same grade-of-service as optimal loss and, for short connections, about the same grade-of-service as that obtained for the 10-percent worst optimal connections. Comparison of Figs. 16 and 19 indicates that the loss-noise grade-of-service for the short digital connection is better than the 10-percent best values obtained with the VNL plan. Thus, this plan provides satisfactory performance and was adopted for calls derived on digital facilities and switched digitally. Although this case is not possible today, growth studies indicate that this might be possible in the 1980s or 1990s.

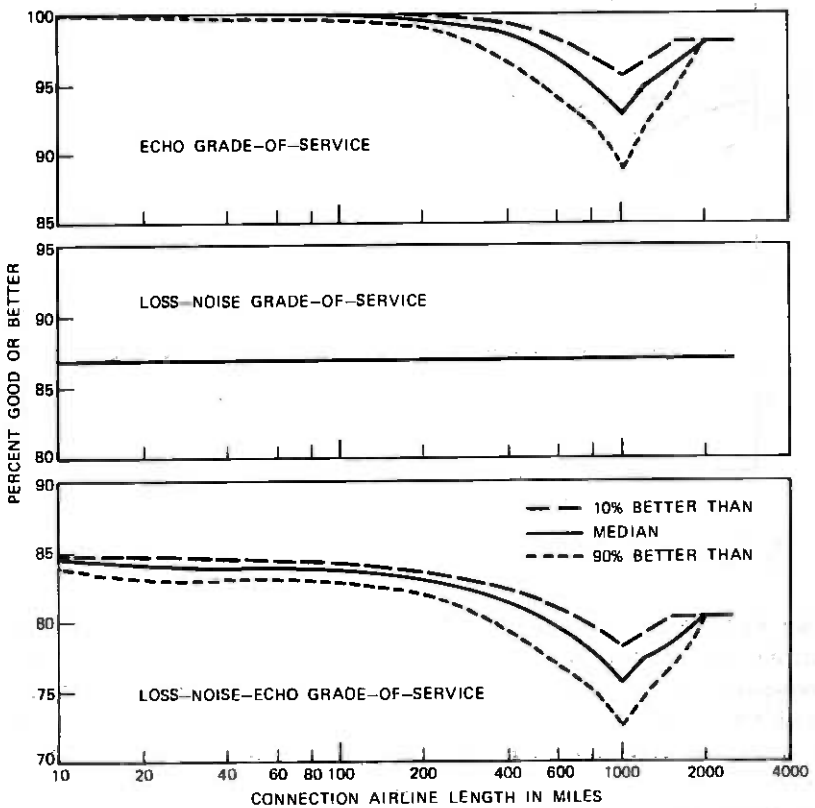


Fig. 19—Grades-of-service obtained on connections routed over digital network when it is designed to a fixed 6-dB loss and a median Class 4 terminal balance of 22 dB.

To avoid decoding at intermediate toll offices, the 6-dB end-to-end connection loss must be achieved by inserting loss in the toll-connecting trunks either in the transmit direction before the digital encoder or in the receive direction after the digital decoder. This loss is allocated equally (3 dB) based on the need for approximately the same connection loss when digital trunks are connected in tandem with analog trunks and maintenance considerations. The maintenance considerations are:

(*i*) The measured loss including test pads should be the same in each direction of transmission.
(*ii*) Test-tone levels should be at standard levels at the input and output of carrier systems.
(*iii*) Standard loop-around tests for digital carrier systems should be preserved.
(*iv*) The transmission level of existing analog offices should remain at 0 TLP and −2 TLP for local and toll offices.

Although the toll-connecting trunk is required to have 3 dB loss, there is no loss adjustment for a trunk on a digital facility connected to the digital office by a digital interface. Since the outgoing switch of a local office is defined as a 0 TLP, the necessary loss can only be achieved within the above constraints by having the end of the trunk at an effective −3 TLP. Actually, this transmission-level point does not exist in the normal sense of a definable analog signal, since a signal at the end of a trunk exists only as a bit stream within the office. The level only exists when the bits are decoded into an analog signal. Since the bit representation is not changed within the digital portion of the office, the output of the digital interface used for trunk testing is defined as a −3 TLP rather than the standard −2 TLP used with analog offices. This definition leads to the level plan for toll-connecting trunks connected to a digital office as shown in Fig. 20.

4.2.3 Balance

Digital switching offices will be four-wire; therefore, through balance considerations do not apply. However, toll-connecting trunks will originate or terminate at two-wire Class 5 offices and could use two-wire facilities. Thus, terminal balance measurements are required at digital toll offices.

The echo performance of the digital network is inherently more sensitive than the analog network to the terminal balance return loss requirement, because of the lack of loss in the intertoll trunks. The discussion in the previous section was predicated on a distribution with a median value of 22 dB for the terminal balance of digital offices.

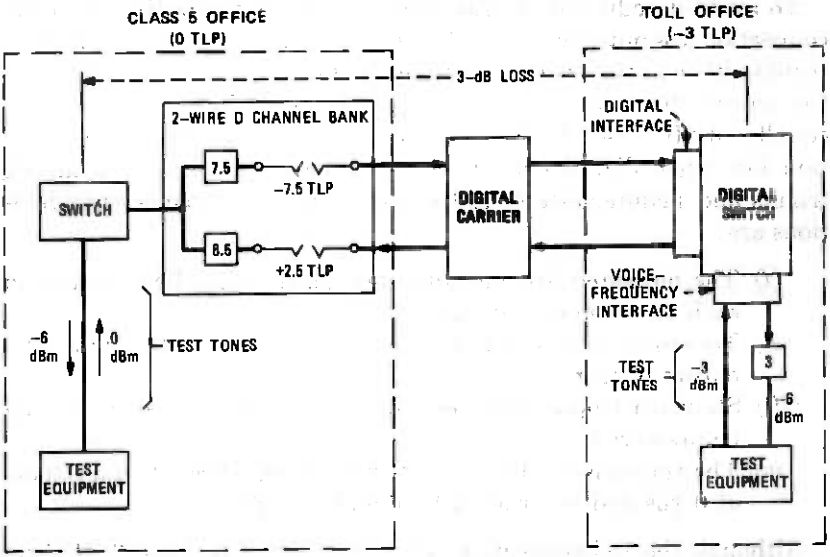


Fig. 20—Levels of a 3-dB toll-connecting trunk when connected through a digital interface to a digital office designated as a −3 TLP (transmission level point).

Figure 21 shows the echo grade-of-service for the digital network with the Class 4 median terminal balance assumed to be 18 dB. This indicates a drop of around 5 percentage points from that obtained with a median terminal balance of 22 dB (Fig. 19). This amount of drop is judged significant since it is not accompanied by an increase in loss-noise grade-of-service. Thus, the distribution of return losses for toll-

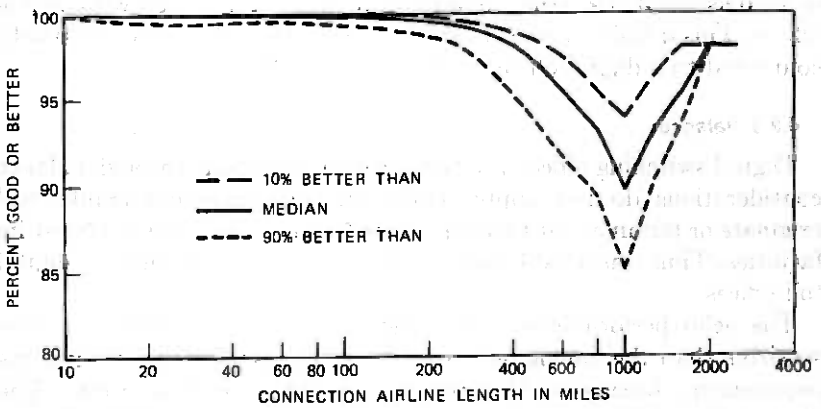


Fig. 21—Echo grade-of-service obtained on connection routed over digital network having a median Class 4 terminal balance of 18 dB.

connecting trunks in a digital office must achieve or exceed a statistical distribution having a median value of 22 dB, minimum of 16 dB. This agrees with the long-term objective for the analog network but is higher than that currently required for analog two-wire toll-connecting trunks. It is expected that the requirement is achievable when applied to all toll-connecting trunks as one group because of the economic advantage of trunks on digital carrier facilities.

4.2.4 Echo suppressor application rules

The loss-noise-echo grade-of-service for the digital network continuously decreases for longer distances because of the decrease in the echo grade-of-service. The amount of decrease is controlled by the echo suppressor application distance. Figure 19 shows the loss-noise, echo, and loss-noise-echo grade-of-service for the digital network when the echo suppressors are applied on high-usage trunks at the analog network mileage of 1850 route miles. The application of echo suppressors at this trunk distance affects the grade-of-service of calls of airline length greater than about 1000 miles. The minimum loss-noise-echo and echo grade-of-service appears satisfactory even though the echo grade-of-service is somewhat lower than that obtained for the analog network (Fig. 11). The joint loss-noise-echo grade-of-service is, however, significantly better. Since the joint assessment is the primary consideration, it appears that the echo suppressor rules for the analog network can be used for the digital network. A similar conclusion is obtained for application of echo suppressors on RC-RC trunks.

4.3 Mixed digital network

For sound transmission quality and technological and economic reasons, the analog and digital networks are being designed to different loss and level plans. Yet in many instances, a connection will be established over portions of both networks. One area of concern was the loss of the interconnecting trunks between an analog and a digital office. The definition of the digital interface as a −3 TLP and the maintenance constraints listed in Section 4.2.2 causes a trunk using digital facilities from an analog toll office to a digital toll office to have an effective loss of 1 dB, as shown in Fig. 22.

The loss of 1 dB on trunks using a digital facility is about the same loss as that required by the VNL design plan for a 500-mile trunk. Thus, this loss is higher than that normally used for trunks of less than this distance and lower for trunks longer than 500 miles. In general, this could mean that there might be either a loss-noise or an echo grade-of-service degradation for connections using these trunks.

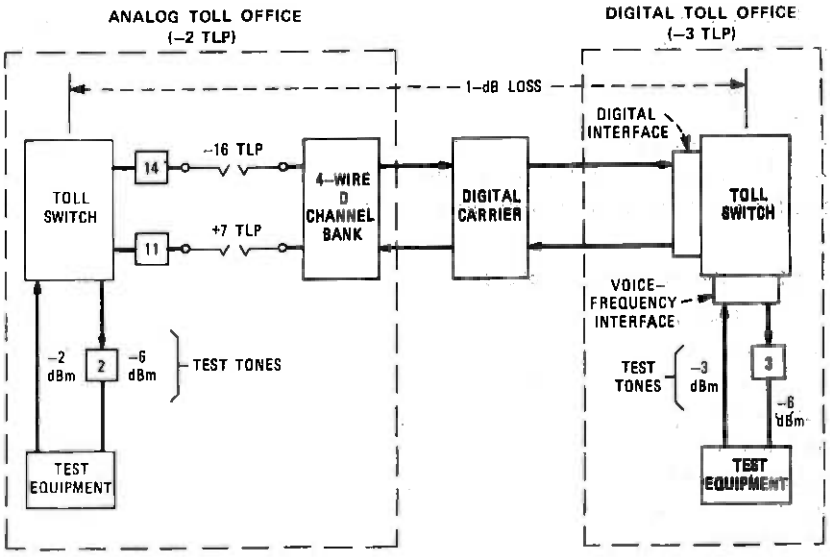


Fig. 22—Levels and loss of a digital intertoll trunk between an analog toll office (−2 TLP) and a digital office (−3 TLP).

The loss-noise and echo grade-of-service of connections containing one or two 1-dB interconnecting trunks was determined. The simulation results for both the cases indicate, as shown in Fig. 23, that the loss-noise performance of connections is improved because of the decreased noise on digital facilities. The echo performance is the same as that for connections using VNL-designed trunks. The improvement in loss-noise grade-of-service could be slightly greater if these trunks were designed for zero loss. Actually, the 1-dB loss provides 2-dB more path loss for talker-echo control. This is considered advantageous, since these trunks interconnect the digital network which is designed with minimum amount of loss and the analog network which has a large variability in return losses. Thus, the transmission performance of connections using both networks with digital interconnecting trunks appears to be quite satisfactory.

However, the simulation results for a trunk using an analog facility designed to a fixed loss of 1 dB indicated a loss-noise degradation for short connections and an echo degradation for medium length connections. Because of the grade-of-service degradation, it was recommended that all interconnecting trunks using analog facilities be designed to the same loss plan as normal analog trunks.

With these loss recommendations for interconnecting trunks, the transmission performance of connections using portions of both net-

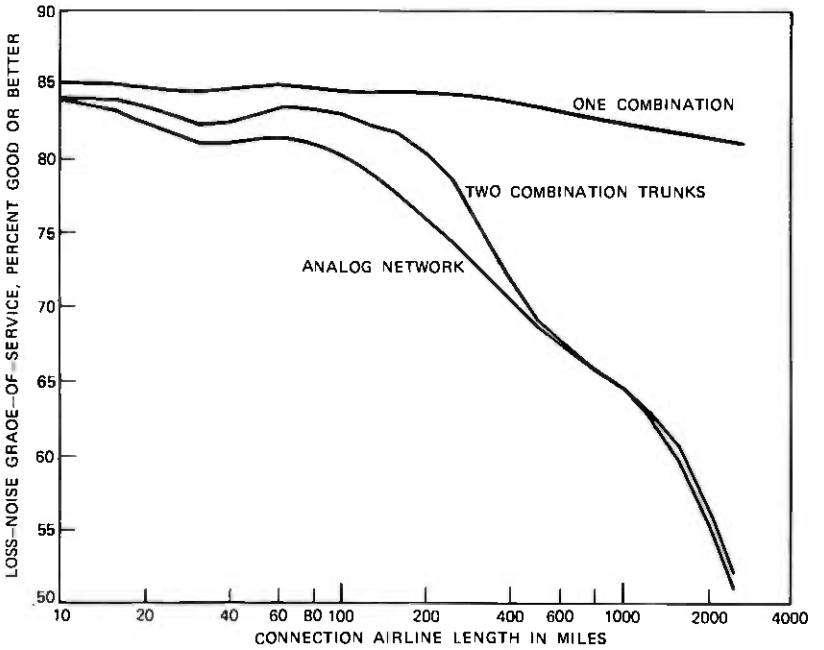


Fig. 23—Median loss-noise grade-of-service for a combined digital-analog connection and a pure analog connection.

works will be satisfactory. This allows introduction of isolated portions of a digital network such as digital switching machines and digital intertoll trunks. As continued growth occurs, these portions will be interconnected digitally, enabling connections to be routed on purely digital facilities.

V. ACKNOWLEDGMENTS

The author would like to thank his co-workers for their advice and help with this work, particularly P. Green, P. Mullins, and A. Scheavitz for implementing the computer simulation model, and E. J. Anderson and H. C. Franke for the transmission level plan for the digital network.

REFERENCES

1. J. L. Sullivan, "Is Transmission Satisfactory? Telephone Customers Help Us Decide," Bell Laboratories Record, *52*, No. 3 (March 1974), pp. 90–98.
2. P. T. Brady and G. K. Helder, "Echo Suppressor Design in Telephone Communications," B.S.T.J., *42* (November 1963), pp. 2893–2917.
3. H. R. Huntley, "Transmission Design of Intertoll Telephone Trunks," B.S.T.J., *32* (September 1953), pp. 1019–1036.
4. H. E. Vaughan, "An Introduction to No. 4 ESS," *IEEE 1972 International Switching Symposium Record*, IEEE, June 1972, pp. 19–25.

5. F. P. Duffy, G. K. McNees, I. Nasell, and T. W. Thatcher, "Echo Performance of Toll Telephone Connections in the United States," B.S.T.J., *54* (February 1975), pp. 207–241.
6. I. Nasell, "Some Transmission Characteristics of Bell System Toll Connections," B.S.T.J., *47* (July-August 1968), pp. 1001–1018.
7. I. Nasell, C. R. Ellison, and R. Holmstrom, "The Transmission Performance of Bell System Intertoll Trunks," B.S.T.J., *47* (October 1968), pp. 1561–1613.
8. J. E. Kessler, "The Transmission Performance of Bell System Toll Connecting Trunks," B.S.T.J., *50* (October 1971), pp. 2741–2777.
9. F. P. Duffy and T. W. Thatcher, Jr., "Analog Transmission Performance on the Switched Telecommunications Network," B.S.T.J., *50* (April 1971), pp. 1311–1347.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 1, January 1976
Printed in U.S.A.

The Jitter Performance of Phase-Locked Loops Extracting Timing From Baseband Data Waveforms

By D. L. DUTTWEILER

(Manuscript received June 3, 1975)

Phase comparators used in phase-locked loops extracting symbol timing from baseband data waveforms typically only produce a useful error signal when a data transition occurs. This gating of the error signal by data transitions makes the natural model for studying jitter in such phase-locked loops time-varying and difficult to analyze. In this paper we show how, under good approximation, it can be simplified to a time-invariant model that is easily analyzed. Using this model, we study jitter accumulation along chains of digital repeaters with phase-locked-loop timing extractors. A numerical example is given.

I. INTRODUCTION

To decode a baseband data waveform, a clock signal giving the proper sampling time must be available. Pilot tones are sometimes transmitted along with the data waveform for this purpose but, alternately, timing can be derived directly from the data waveform itself. One approach to self-timing is to let the data waveform passed through a memoryless nonlinearity ring a tuned circuit with a resonant frequency close to the nominal signaling rate.[1–8] Another technique, which in general involves more circuitry but gives superior performance, is to use a phase-locked loop (PLL).

The element of a PLL extracting symbol timing that is most interesting functionally is its phase comparator. Numerous realizations are possible. Almost all these realizations are similar, however, in that they only produce an error signal when a data transition occurs. The gating of the error signal by data transitions in a PLL extracting symbol timing makes the analysis of such a PLL potentially quite different from the analysis of phase-locked loops used in other applications.

For purposes such as studying timing acquisition, the effect of transition gating is adequately modeled as a multiplication of the

phase comparator gain in the presence of a data transition by the probability p of a data transition. After this approximation is made, the loop has the conventional structure. For studying jitter, however, this model is not satisfactory because transition gating is itself often an important source of jitter that of necessity is neglected when the average-gain approximation is made. As will be discussed more fully later, transition gating produces jitter by chopping the static phase error needed to pull the voltage-controlled oscillator (VCO) from its natural frequency to the line frequency and thereby injecting wideband noise into the loop.

The purpose of this paper is to develop a time-invariant linear model for a PLL with transition gating that adequately mirrors, even for jitter-studying purposes, the effects of transition gating. The model we obtain differs from the linear model obtained by simply making the average-gain approximation in the addition of an additive wideband noise representative of the chopping of the static phase error by transition gating.

With the time-invariant linear model available, it is straightforward to calculate the spectrum and variance of the jitter on the output of a PLL extracting symbol timing from a data waveform. We conclude with a discussion of jitter accumulation along chains of digital repeaters with PLL timing extractors.

Saltzberg[9] and Roza[10] have both previously done excellent work analyzing phase-locked loops extracting symbol timing. Our work is closer to Saltzberg's, and we use many of his results. Saltzberg was able to obtain equations for the jitter spectrum and jitter variance directly from the natural gated model. The equations we obtain from the approximate model are similar, but they are obtained with much less difficulty and without needing to assume one-sided intersymbol interference.

Roza's analysis is at a more abstract level. His equations for the spectrum and variance of the jitter do not require an exact specification of the phase comparator being used as do Saltzberg's and ours but, consequently, they relate the jitter spectrum and variance to more abstract quantities with less physical meaning.

II. GENERAL MODEL

The general PLL model we assume is shown in Fig. 1. The phase $\phi(t)$ of the data signal and the phase $\theta(t)$ of the timing output from the PLL are measured in slots (fractions of a symbol interval) and assumed to be slowly varying in comparison to a symbol interval. The loop is assumed to be in lock.

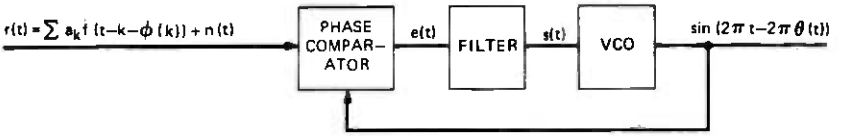


Fig. 1—General model for a PLL timing extractor.

The input data waveform will be taken to have the form*

$$r(t) = \sum a_k f[t - k - \phi(k)] + n(t), \tag{1}$$

where $n(t)$ is additive noise, $f(t)$ is the standard pulse, and the $\{a_k\}$ are the data taking on the values ± 1. The normalization of the symbol interval to unity is without loss of generality and is equivalent to agreeing to measure time (as well as phase) in slots. The assumption of polar (± 1) rather than, say, unipolar (0, 1) signaling is also without loss of generality. The assumption of binary data is restrictive. A generalization of our results for multilevel data would be nontrivial, but should be possible.

Note that, in writing the output of the VCO as $\sin [2\pi t - 2\pi\theta(t)]$, we are not precluding the possibility of the VCO center frequency being offset from the line frequency. With the loop open and a center frequency offset of Δf, the VCO output is expressible as $\sin [2\pi t - 2\pi\theta(t)]$ with

$$\theta(t) = \Delta f t + \theta_0. \tag{2}$$

III. PHASE DETECTORS

To proceed, we need an equation for the error signal $e(t)$ produced by the phase comparator and, thus, we must specify precisely the phase comparator to be considered. In this section, we develop equations for the error signals produced by two different phase comparators: the zero-crossing phase comparator described by Saltzberg[9] and the dead-zone quantizer (DZQ) phase comparator used in the T4M repeater.[11] Characterizing the DZQ phase comparator will allow presenting numerical results.

It happens that the error signal $e(t)$ produced by the DZQ phase comparator is described by equations identical to those describing the error signal produced by the zero-crossing phase comparator if certain constants relating to the pulse shape $f(t)$ and certain random variables relating to the channel noise $n(t)$ are redefined. We expect that the error signals from other phase comparators with transition gating probably also have this same functional form or one very similar to it.

* Summations and integrals written without limits are to be assumed to be from $-\infty$ to $+\infty$.

For this reason, it is felt that, although the analysis of Sections IV through VII is only directly applicable for a zero-crossing phase comparator or a DZQ phase comparator, it can probably be readily extended for any other phase comparator exhibiting transition gating.

3.1 Zero-crossing detector

The zero-crossing phase comparator produces an error signal $e_{ZC}(t)$ that is a train of weighted pulses with separation $T = 1$. The pulse weights are proportional to the time difference between zero crossings of the data waveform and the VCO output. For a PLL with a closed-loop response cutting off well below the symbol rate, as should always be the case, the exact phase of this pulse train and shape of its pulses are unimportant, and we can to good approximation take it to be given by

$$e_{ZC}(t) = \sum e_k \delta(t - k), \tag{3}$$

where $\delta(t)$ is the Dirac delta function and the $\{e_k\}$ are weights.

A typical received data waveform,

$$r(t) = \sum a_k f(t - k - \phi_k) + n(t), \tag{4}$$

where

$$\phi_k \triangleq \phi(k) \tag{5}$$

is drawn in Fig. 2a. It is assumed that $n(t)$ and $\{\phi_k\}$ are identically zero, and that the standard equalized pulse $f(t)$ has a raised cosine shape (see Fig. 2b). Notice that, whenever there is no data transition, there is no zero crossing. In this case, the zero-crossing phase detector sets e_k equal to zero. When the noise $n(t)$ or the input phase samples $\{\phi_k\}$ are not identically zero, or when $f(t)$ is not of its nominal raised-cosine shape, the received waveform is no longer as clean as that shown in Fig. 2a. Nonetheless, it will still maintain the same basic character with zero crossings occurring only with associated data transitions.

Saltzberg cleverly shows in Ref. 9 that, to within excellent approximation, e_k is given by

$$e_k = \alpha_1 d_k(\theta_k - \phi_k - w_k). \tag{6}$$

The proportionality of e_k to the phase error $\theta_k - \phi_k$ through the proportionality constant α_1 is the desired phase comparator response. This desired response is degraded by the on-off gating of the variable

$$d_k = \begin{cases} 1, & a_k \neq a_{k+1} \\ 0, & a_k = a_{k+1} \end{cases} = (1 - a_k a_{k+1})/2 \tag{7}$$

associated with data transitions and the additive disturbance w_k given

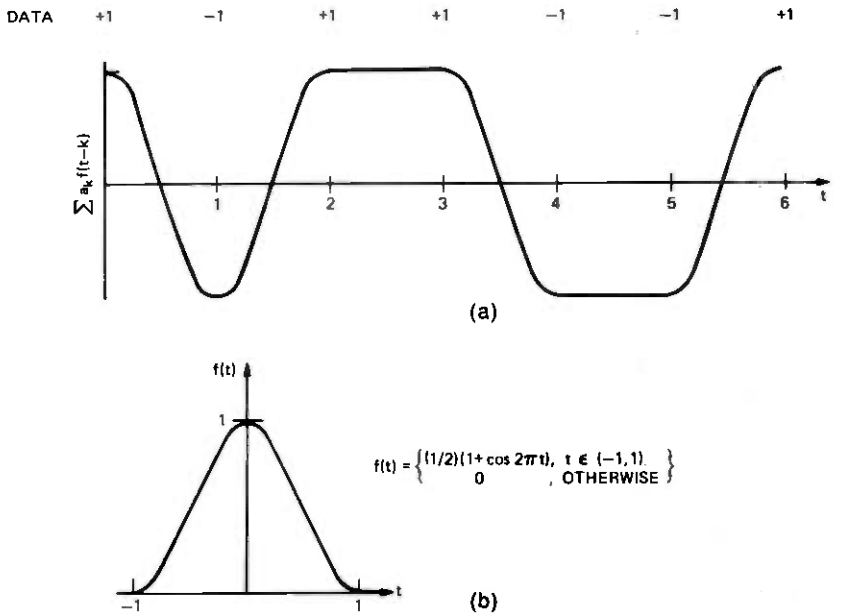


Fig. 2—(a) Nominal received waveform. (b) Raised-cosine pulse.

by

$$w_k = a_k v_k + a_k \sum_m a_{k-m}\epsilon_m, \tag{8}$$

where

$$v_k = \frac{n(k + \frac{1}{2} + \phi_k)}{[\dot{f}(\frac{1}{2}) - \dot{f}(-\frac{1}{2})]} \tag{9}$$

and

$$\epsilon_k = \begin{cases} 0, & k = 0, 1 \\ f(k + \frac{1}{2}), & k \neq 0, 1 \end{cases}. \tag{10}$$

[The pulse $f(t)$ has, without loss of generating, been centered so that $f(\frac{1}{2}) = f(-\frac{1}{2})$]. The two components $a_k v_k$ and $a_k \sum_m a_{k-m}\epsilon_m$ of w_k account for the shift in zero crossings of the received waveform by additive channel noise and by intersymbol interference, respectively. We shall not derive (6) here, since Saltzberg's derivation is quite clear and the derivation of $e_{DZQ}(t)$ in the next section and the appendix, being similar, presents the key ideas.

3.2 DZQ phase detector

The operation of the zero-crossing phase comparator is conceptually simple, but its implementation appears difficult. The dead-zone

quantizer (DZQ) phase comparator, which was first proposed by W. G. Hammett, is more readily implemented and is being used in the T4M repeater. By describing and mathematically characterizing it here, we establish a base for numerical results to be given later.

A block diagram of the DZQ phase comparator is shown in Fig. 3a. Its inputs are the received waveform

$$r(t) = \sum a_k f(t - k - \phi_k) + n(t) \tag{11}$$

and the VCO output $\sin [2\pi t - 2\pi\theta(t)]$. Its only block requiring explanation is the DZQ, which is a memoryless nonlinearity with the transfer function shown in Fig. 3b.

The operation of this phase comparator is illustrated in Fig. 4, where waveforms at the inputs and outputs of all the blocks in Fig. 3a are shown. It is assumed that the data sequence is as indicated, $n(t) = 0$, $\phi(t) = 0$, $f(t)$ has raised cosine shape (nominal for the T4M repeater), and $\theta(t) = \frac{1}{8}$ (that is, the VCO phase is lagging 45 degrees). The true error signal shown on the next-to-last line can be taken, for mathematical purposes, as the train of weighted impulses shown on the last line with the impulse weights equal to the integral of their associated pulses. Notice in Fig. 4 that the phase-comparator output is nonzero only when there is a data transition.

As in the zero-crossing phase comparator, the effect of additive channel noise and intersymbol interference is to make the weights of the impulses in the phase-comparator output differ from strict pro-

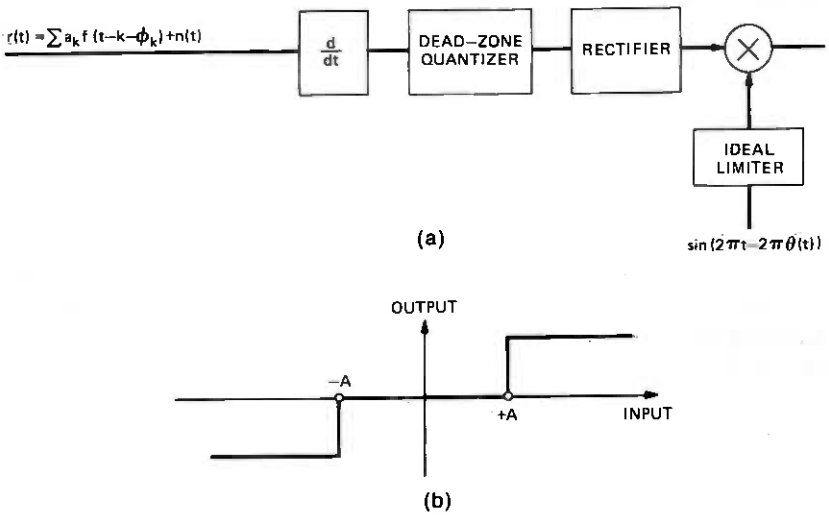


Fig. 3—(a) DZQ phase comparator. (b) Instantaneous transfer function of a dead-zone quantizer.

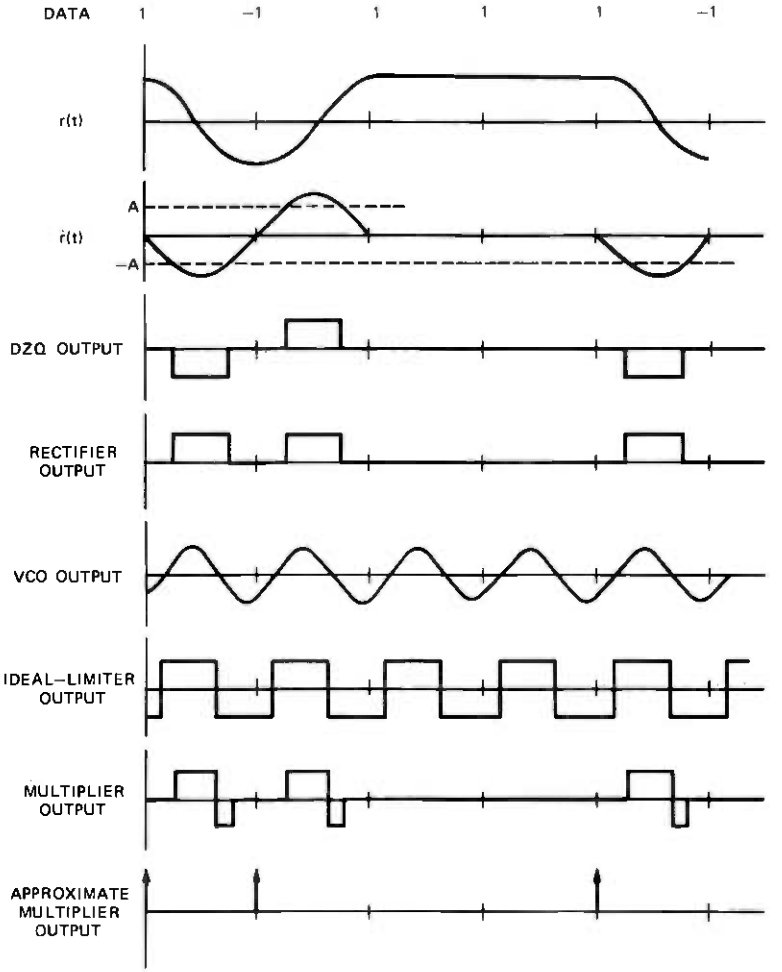


Fig. 4—Waveforms in DZQ phase comparators.

portionality to the difference between $\phi(t)$ and $\theta(t)$ by a small random amount [through $n(t)$ and the data $\{a_k\}$]. Defining the weights of the impulses in the last line of Fig. 4 as $\{e_k\}$, that is, defining

$$e_{DZQ}(t) = \sum e_k \delta(t - k), \tag{12}$$

we show in the appendix that

$$e_k = \alpha_1 d_k(\theta_k - \phi_k - w_k), \tag{13}$$

where all symbols except the channel noise variables $\{v_k\}$ and the intersymbol interference constants $\{\epsilon_k\}$ are as in the discussion of the

zero-crossing phase comparator. The changed definitions are

$$v_k = \frac{\dot{n}(k + \phi_k + \tau^-)}{(2\beta^-)} + \frac{\dot{n}(k + \phi_k + \tau^+)}{(2\beta^+)} \tag{14}$$

and

$$\epsilon_k = \begin{cases} 0, & k = 0, 1 \\ \dot{f}(k + \tau^-)/(2\beta^-) + \dot{f}(k + \tau^+)/(2\beta^+), & k \neq 0, 1 \end{cases}, \tag{15}$$

where

$$\beta^{\pm} = \dot{f}(\tau^{\pm} - 1) - \dot{f}(\tau^{\pm}) \tag{16}$$

and τ^+ and τ^- are defined as in Fig. 5. The pulse $f(t)$ has without loss of generality been centered so that

$$\tau^- + \tau^+ = 1. \tag{17}$$

Thus, the only difference in the two phase comparators is in the precise way in which the random variables $\{v_k\}$ are related to the additive channel noise and the constants $\{\epsilon_k\}$ to the pulse tails.

IV. SWITCHED LINEAR MODEL

Using eqs. (3) and (6) as a characterization of the phase comparator and realizing that the phase of the vco output is the integral of the signal on its input, we can model the PLL of Fig. 1 as shown in Fig. 6. In Fig. 6, Δf is the difference between the center frequency of the vco and the line frequency, which we have normalized to unity. The function $H(s)$ is the transfer function of the low-pass filter normalized so that $H(0) = 1$, and the open loop dc gain α is given by

$$\alpha = \alpha_1\alpha_2\alpha_3,$$

where α_1 is the phase-comparator gain [previously defined by eq. (6)], α_2 is the dc gain of the low-pass filter, and α_3 is the vco gain.

It will be convenient to mix discrete and continuous notation as in Fig. 6. For a precise interpretation of figures such as these, discrete

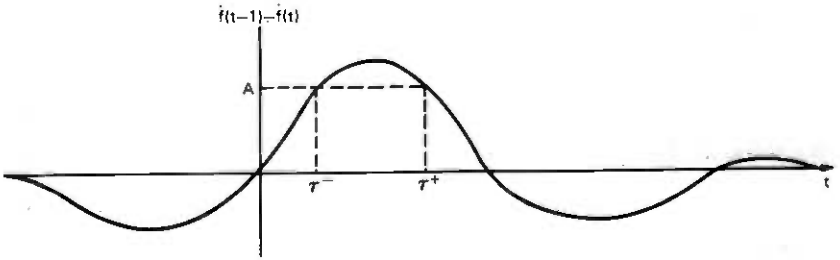


Fig. 5—Defining figure for τ^- and τ^+.

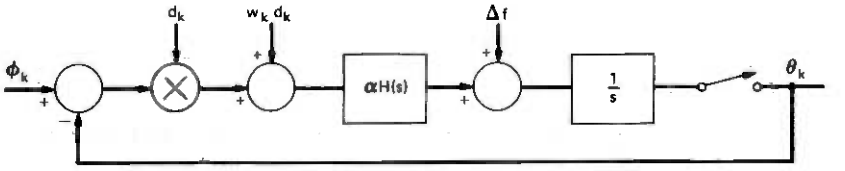


Fig. 6—Time-varying model for PLL.

labels such as θ_k and $w_k + \phi_k$ are to be taken as shorthand for $\sum \theta_k \delta(t - k)$ and $\sum (w_k + \phi_k)\delta(t - k)$.

The model of Fig. 6 is linear, but it varies in time as the gating variable d_k opens and closes the loop. A direct analysis of jitter using this model is made difficult by the gating.

V. TIME-INVARIANT LINEAR MODEL

The key to obtaining a time-invariant approximation is to consider the nature of the signal $(\phi_k - \theta_k)d_k$ at the gate output in Fig. 6. Letting p denote the probability of a data transition, we have

$$(\phi_k - \theta_k)d_k = (\phi_k - \theta_k)p + (\phi_k - \theta_k)(d_k - p). \tag{18}$$

The term $(\phi_k - \theta_k)p$ in this expansion represents the low frequency (in comparison to the symbol rate) response of the gate to the low frequency signal $\phi_k - \theta_k$ on its input. This signal completes the feedback path. Since d_k has mean p and to good approximation is not correlated with $(\phi_k - \theta_k)$, the other component is a wideband signal with no dc. It represents jitter-producing wideband noise injected into the loop by the gating.

Conceptually, the phase error $\phi_k - \theta_k$ has three components: a static offset, which we denote by μ, necessary to pull the vco from its quiescent frequency to the line frequency; a low-frequency component present when $\phi(t)$ changes faster than the loop can track; and jitter. Our key approximation is

$$(\phi_k - \theta_k)(d_k - p) \doteq \mu(d_k - p). \tag{19}$$

In the PLL timing extractor of the T4M repeater, μ is about 100 times the standard deviation of the jitter when the vco offset is a worst case 60 parts per million. Also $\phi(t)$, which is typically jitter from previous repeaters, is within the closed-loop bandwidth so that it can be tracked. Thus, for this system approximation (19) is excellent unless the center frequency of the vco happens to be such that μ is unusually small. In this case, however, approximating the wideband power injected by $(\phi_k - \theta_k)(d_k - p)$ by the wideband power injected by

$$\mu(d_k - p) \doteq 0$$

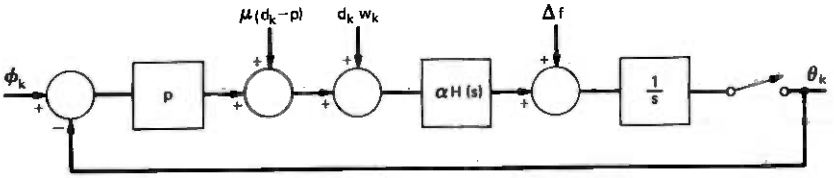


Fig. 7—Approximation to time-varying model.

is still good since the jitter-producing noise injected into the loop by additive channel noise and intersymbol interference is much stronger than that injected by transition gating. We expect that, for most other well-designed phase-locked loops, approximation (19) will be good.

Approximation (19) substituted in (18) gives

$$(\phi_k - \theta_k)d_k \doteq (\phi_k - \theta_k)p + \mu(d_k - p). \tag{20}$$

With this approximation, the model of Fig. 6 becomes the time-invariant model of Fig. 7, which can be further simplified to the model of Fig. 8.

It remains to calculate the static offset μ. Refer to Fig. 6. Since there can be no steady-state dc signal at the integrator input, the dc on the output of the low-pass filter must equal $-\Delta f$, and the dc on its input must equal $-\Delta f/\alpha$. Therefore, with overscores denoting mean values,

$$\overline{(\phi_k - \theta_k)d_k} = -\frac{\Delta f}{\alpha} - \overline{w_k d_k}. \tag{21}$$

We noted earlier that $\phi_k - \theta_k$ and d_k are approximately uncorrelated. Thus,

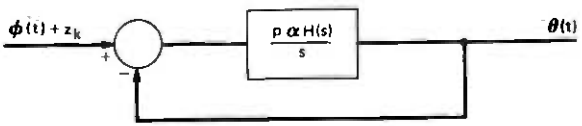
$$\overline{(\phi_k - \theta_k)d_k} = \overline{(\phi_k - \theta_k)}\,\overline{d_k} = \mu p \tag{22}$$

and

$$\mu = -\frac{\Delta f}{p\alpha} - \frac{\overline{w_k d_k}}{p}. \tag{23}$$

The mean $\overline{w_k d_k}$ is a function of the noise statistics and data statistics. It can be shown that if the data are independent of the noise and have stationary statistics,

$$\overline{w_k d_k} = 0, \tag{24}$$



$z_k = (1/p)\,(\omega_k d_k + \mu(d_k - p)) + \Delta f/\alpha$

Fig. 8—Time-invariant linear model for PLL.

in which case

$$\mu = -\frac{\Delta f}{p\alpha}. \tag{25}$$

The model of Fig. 8 gives the output phase $\theta(t)$ as the sum of the input phase $\phi(t)$ passed through a filter with transfer function

$$G(s) = \frac{p\alpha H(s)}{s + p\alpha H(s)} \tag{26}$$

and added noise

$$z_k = \left(\frac{1}{p}\right)\left[w_k d_k + \mu(d_k - p) + \frac{\Delta f}{\alpha}\right] \tag{27}$$

passed through the same filter. Since the model is fixed and linear, it is easily analyzed. Although in the remainder of the paper we concentrate on finding the output-phase spectrum $S_\theta(f)$ for random $\phi(t)$, $n(t)$, and $\{a_k\}$, exact calculations of $\theta(t)$ can be made for given $\phi(t)$, $n(t)$, and $\{a_k\}$.

VI. SPECTRUM AND VARIANCE

The spectrum $S_\theta(f)$ of $\theta(t)$ can be calculated under various assumptions on the statistics of $\phi(t)$, $n(t)$, and the data sequence $\{a_k\}$. We make the calculation here assuming

(*i*) $n(t)$, $\phi(t)$, and the data $\{a_k\}$ are statistically independent and stationary random processes.

(*ii*) The random variables $\{a_k\}$ are independent.

Assumption (*i*) is reasonable if $\phi(t)$ is not to represent accumulated jitter from previous timing extractors.

Let p' denote the probability a data bit equals 1 and let $q' = 1 - p'$. Then

$$p = 2p'q'. \tag{28}$$

Notice that the transition probability is at most $\frac{1}{2}$, which is achieved with equiprobable data.

Assumption (*i*) implies that the driving processes $\phi(t)$ and $\{z_k\}$ in Fig. 8 are independent.* Thus, the spectrum $S_\theta(f)$ of $\theta(t)$ is given by

$$S_\theta(f) = |G(f)|^2 S_\phi(f) + |G(f)|^2 S_z(f), \tag{29}$$

* The $\{z_k\}$ process is a function of the $\{a_k\}$ and $\{v_k\}$ processes. Equation (9) (zero-crossing phase comparator) and eq. (14) (DZQ phase comparator) both give the $\{v_k\}$ process as a function of both the $\phi(t)$ and $n(t)$ processes. Thus, the independence of the $\phi(t)$ and $n(t)$ processes does not at first seem sufficient for the independence of the $\phi(t)$ and $\{v_k\}$ processes. However, the functional forms of both (11) and (18) are such that, in both cases, the univariate statistics of the $\{v_k\}$ process are independent of the statistics of the $\phi(t)$ process [the stationarity of the $n(t)$ process is used here] and moreover to within the slowly-varying $\phi(t)$ assumption already made, the multivariate statistics are also. Thus, the $\phi(t)$ and $\{v_k\}$ processes are independent to excellent approximation.

where $S_\phi(f)$ is the spectrum of $\phi(t)$ and $S_z(f)$ is the spectrum of the process

$$z(t) = \sum_{k=-\infty}^{\infty} z_k \delta(t - k - \lambda), \tag{30}$$

with λ being an epoch-randomizing uniform $[0, 1]$ random variable, the addition of which is necessary to make $z(t)$ stationary. Assuming the loop to be narrowband so that the cutoff of the low-pass function $G(f)$ is well within the band of frequencies over which the spectrum of the wide-band process $z(t)$ is flat, we have, to excellent approximation,

$$S_\theta(f) = |G(f)|^2 S_\phi(f) + |G(f)|^2 S_z(0). \tag{31}$$

Since the variance σ_θ^2 of $\theta(t)$ is the integral of $S_\theta(f)$, we have to this same approximation

$$\sigma_\theta^2 = \int |G(f)|^2 S_\phi(f) df + S_z(0) \int |G(f)|^2 df. \tag{32}$$

The first term of (31) and the first term of (32) are associated with the input phase $\phi(t)$ and will not be considered further. The more interesting second terms are associated with the jitter added by the timing extractor. Notice that $S_z(0)$ provides a measure of the timing extractor's performance.

It can be shown through straightforward but tedious manipulations that

$$S_z(0) = \frac{[C_v(0) - C_v(1)]}{p} + \frac{\mu^2(2 - 3p)}{p} - 2\mu(\epsilon_1 - \epsilon_{-2}) + \sum_{k=1}^{\infty} (\epsilon_k + \epsilon_{-k} - \epsilon_{k-1} - \epsilon_{-k-1})^2, \tag{33}$$

where $C_v(k)$ is the covariance function of the sequence $\{v_k\}$. The first term in (33) is associated with additive channel noise and will usually be negligible in comparison to the remaining terms, which are associated with the pulse shape $f(t)$ through the intersymbol-interference constants $\{\epsilon_k\}$ and the vco offset Δf through μ.

It is straightforward to show for either phase comparator that, if $f(t)$ is a symmetric pulse, then

$$\epsilon_k = \epsilon_{-k-1}, \qquad k = 1, 2, \cdots \tag{34}$$

and therefore

$$S_z(0) = [C_v(0) - C_v(1)]/p + \mu^2(2 - 3p)/p. \tag{35}$$

For unknown μ, this is certainly the best that can be done. The design

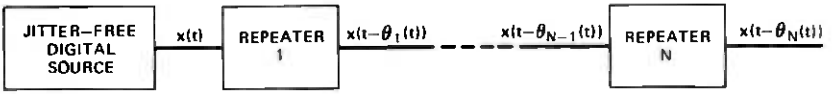


Fig. 9—Jitter accumulation along a chain of repeaters.

conclusion to be drawn from (35) and (31), therefore, is that symmetric pulse shapes are best for timing purposes.

VII. CHAINS

The results of the previous section are easily extended for a cascade of PLL timing extractors. Cascades are of engineering interest for the study of chains of digital repeaters with PLL timing extractors. In a cascade like that of Fig. 9, we call $\theta_N(t)$ *accumulated jitter* and

$$\Delta_N(t) = \theta_N(t) - \theta_{N-1}(t) \tag{36}$$

alignment jitter. The name alignment jitter comes from the fact that $\Delta_N(t)$ represents the amount by which the actual sampling time in the Nth repeater is misaligned from the ideal sampling time.

In Fig. 9, the phase of the output of the $(n-1)$th repeater is the phase of the input to the nth. This fact and the fixed linear model of Fig. 8 imply the model of Fig. 10, where

$$G_n(s) = \frac{p\alpha_n H_n(s)}{s + p\alpha_n H_n(s)}. \tag{37}$$

This general model is fixed and linear, and thus not impossible to analyze by conventional methods for any parameters of interest. However, to derive insight and to obtain reasonably concise expressions for $\theta_N(t)$ and $\Delta_N(t)$, it is necessary to make further assumptions that reduce the number of parameters. We assume

(*i*) The additive channel noise is negligible and therefore the random variables $\{v_k^{(n)}\}$ may be taken as zero.
(*ii*) The equalized pulse shapes $f_n(t)$ at the inputs to all the repeaters are identical.
(*iii*) The vco offsets Δf_n are all identical.
(*iv*) The filters $H_n(s)$ and open-loop dc gains α_n are all identical.

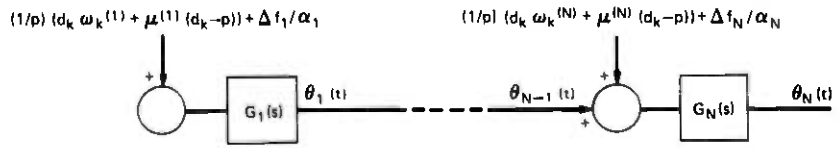


Fig. 10—Model for jitter accumulation along a chain of PLL timing extractors.

Assumption (*i*) is in general reasonable. Assumptions (*ii*) through (*iv*) are less reasonable, but are still interesting in that they are worst-case assumptions. The first two align the input signals in Fig. 10. The last gives the most chain gain at the peak frequency for a specified amount of peaking in each repeater.

With assumptions (*i*) through (*iv*), the model of Fig. 10 simplifies to that of Fig. 11, from which it is apparent that (we are making the wideband noise approximation again)

$$S_{\theta_N}(f) = \left| \sum_{n=1}^{N} G^n(f) \right|^2 S_z(0) \tag{38}$$

and

$$S_{\Delta_N}(f) = |G^n(f)|^2 S_z(0). \tag{39}$$

Interestingly, the model of Fig. 11 is identical to that used by Byrne, Karafin, and Robinson[6] to analyze a chain of digital repeaters with tuned-circuit timing extractors. One important difference, however, between the two analyses is that we have available a formula for $z(t)$ in terms of physically meaningful parameters. Byrne, Karafin, and Robinson did not have such a formula and had to experimentally determine $S_z(0)$.

The variances $\sigma^2_{\Delta_N}$ and $\sigma^2_{\theta_N}$ of $\Delta_N(t)$ and $\theta_N(t)$ are, from (38) and (39), given by

$$\sigma^2_{\Delta_N} = J(N) S_z(0) \tag{40}$$

and

$$\sigma^2_{\theta_N} = I(N) S_z(0), \tag{41}$$

where

$$J(N) = \int |G^N(f)|^2 df \tag{42}$$

and

$$I(N) = \int \left| \sum_{n=1}^{N} G^n(f) \right|^2 df. \tag{43}$$

The large N behavior of $J(N)$ and $I(N)$ depends critically on whether or not $G(f)$ exhibits peaking (i.e., whether or not there exists a frequency f such that $|G(f)| > 1$). When $G(f)$ exhibits peaking, it can be shown by Laplace's method (see, for example, Ref. 12 or

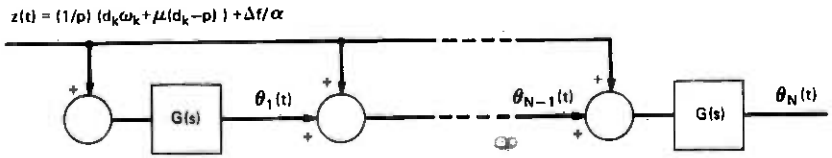


Fig. 11—Simplified model for jitter accumulation along a chain.

Ref. 13) that for large N

$$J(N) \doteq \hat{J}(N) = 2|G(f_p)|^{2N}\sqrt{\pi/[-N\ddot{L}(f_p)]} \tag{44}$$

and

$$I(N) \doteq \hat{I}(N) = \hat{J}(N)\left|\frac{G(f_p)}{1 - G(f_p)}\right|^2, \tag{45}$$

where f_p is the frequency at which $|G(f)|$ is maximum,

$$L(f) = \ln |G(f)|, \tag{46}$$

and $\ddot{L}(f_p)$ is the second derivative of $L(f)$ evaluated at f_p.

When there is no peaking, $G(f)$ is everywhere less than one except at the origin where it must equal one. Thus, as N increases, $|G^N(f)|^2$ and $\sum_{n=1}^{N} |G^n(f)|^2$ become increasingly narrower in bandwidth with amplitudes 1 and N^2 respectively at zero frequency. Let

$$G(f) = \exp\{L(f) + j\phi(f)\} \tag{47}$$

with $L(f)$ and $\phi(f)$ real, and approximate $L(f)$ and $\phi(f)$ near the origin by

$$\begin{aligned} L(f) &\doteq L(0) + \dot{L}(0)f + (\tfrac{1}{2})\ddot{L}(0)f^2 \\ &\doteq (\tfrac{1}{2})\ddot{L}(0)f^2 \end{aligned} \tag{48}$$

and

$$\begin{aligned} \phi(f) &= \phi(0) + \dot{\phi}(0)f \\ &= \dot{\phi}(0)f. \end{aligned} \tag{49}$$

The quantities $L(0)$ and $\dot{L}(0)$ equal zero because $G(f)$ equals one at the origin and $|G(f)|$ is even. The phase $\phi(0)$ equals zero since $G(0)$ is real. Equations (48) and (49) lead after much manipulation to the large N approximations

$$J(N) \doteq \sqrt{\pi/N[-\ddot{L}(0)]} \tag{50}$$

and

$$I(N) \doteq 2\pi N/\dot{\phi}(0). \tag{51}$$

If $G(f)$ is a single-pole characteristic, that is, if

$$G(f) = \frac{1}{1 + jf/f_0}, \tag{52}$$

then (68) reduces to

$$I(N) \doteq 2\pi N f_0, \tag{53}$$

which is consistent with a result in Ref. 6.

Notice the great difference in the asymptotic behavior of $J(N)$ and $I(N)$ in the two situations. With peaking, $J(N)$ and $I(N)$ are asymp-

totically exponential [eqs. (44) and (45)], whereas without peaking $J(N)$ actually decreases with N for large N and $I(N)$ only grows linearly. The rapid growth of (44) and (45) with N indicates that, to control jitter accumulation along a long chain, there must be very little if any peaking.

VIII. NUMERICAL RESULTS

We have used the theory developed here to numerically characterize the jitter performance of the timing extractor in the T4M digital repeater. The exact numbers obtained are, of course, only of direct interest to designers of this repeater, but their relative sizes should be indicative of other timing extractors, and thus of more general interest.

Data on the T4M line are scrambled so that an assumption of independent data is justified, and, furthermore, the probability p of a data transition equals $\frac{1}{2}$. Additive channel noise is a negligible jitter source. The timing loop has a worst-case vco offset of 60 parts per million, a dc gain

$$\alpha = 8 \times 10^{-3}, \tag{54}$$

and a filter

$$H(f) = \frac{(1 + jf/f_1)(1 + jf/f_2)}{(1 + jf/f_3)(1 + jf/f_4)^2(1 + jf/f_5)(1 + jf/f_6)}, \tag{55}$$

where f_1, f_2, f_3, f_4, f_5, and f_6 are, respectively, 3.65×10^{-6}, 3.65×10^{-5}, 1.82×10^{-7}, 7.30×10^{-5}, 1.45×10^{-4}, and 1.09×10^{-3} cycles per slot.*

The reasons for the rather involved form of this filter are of no consequence here. However, it should be pointed out that some of the poles and zeros are fixed by circuit parasitics and are not introduced intentionally.

The shape $f(t)$ of the received and equalized standard pulse depends on repeater spacing and other parameters. A typical waveform $f(t)$ appears in Fig. 12. With a slicing level A fixed at $\frac{3}{4}$ the maximum of $|f(t) - f(t-1)|$, the associated intersymbol interference constants $\{\epsilon_k\}$ are as in Table I. Substituting these values and $p = \frac{1}{2}$ in (33) gives

$$S_z(0) = \mu^2 - 2\mu(0.0090) + 1.61 \times 10^{-4}. \tag{56}$$

For $|\Delta f| \leqq 60 \times 10^{-6}$, we have

$$|\mu| = \left|\frac{\Delta f}{p\alpha}\right| \leqq 0.015. \tag{57}$$

* Assuming $T = 1$ is equivalent to measuring time in slots and thus frequency in cycles per slot.

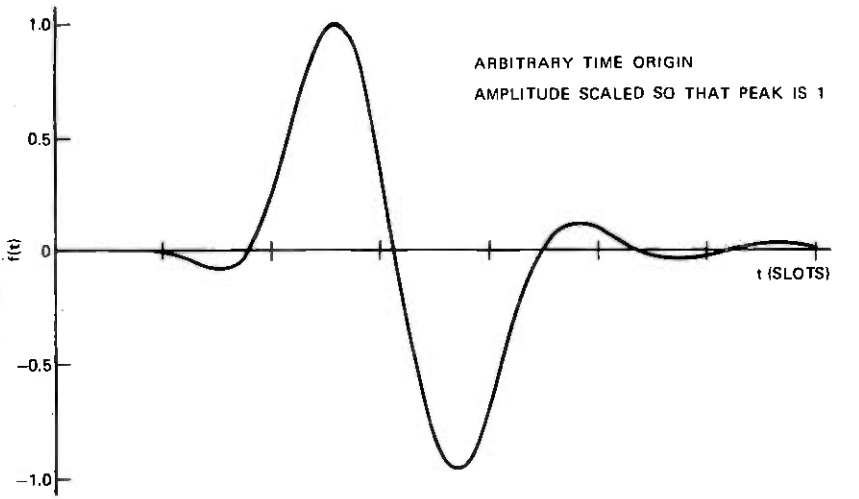


Fig. 12—Output of differentiator.

With $\mu \in [-0.015, 0.015]$,

$$S_z(0) \in [80 \times 10^{-6}, 656 \times 10^{-6}]. \tag{58}$$

The minimum is achieved by

$$\mu = +0.009 \tag{59}$$

and the maximum by

$$\mu = -0.015. \tag{60}$$

All three terms of (33) are of comparable magnitude for these parameters, indicating that for this particular repeater the jitter produced by intersymbol interference is comparable to that produced by transition gating.

Table I — Numerical values for the constants $\{\epsilon_k\}$ having magnitudes greater than 0.00005. The constants ϵ_0 and ϵ_{-1} equal zero by definition

ϵ_{-3}	−0.0009
ϵ_{-2}	0.0194
ϵ_{-1}	0.0000
ϵ_0	0.0000
ϵ_1	0.0284
ϵ_2	−0.0002
ϵ_3	0.0003
ϵ_4	−0.0005
ϵ_5	0.0001

Table II — Integral values and their approximations in dB

N	$\hat{J}(N)$	$J(N)$	$\hat{I}(N)$	$I(N)$
1	−40.0	−38.2	−27.2	−38.2
2	−40.7	−38.9	−28.0	−34.0
3	−40.9	−39.0	−28.1	−31.6
5	−40.6	−38.8	−27.8	−28.6
7	−39.8	−38.3	−27.1	−26.4
10	−38.4	−37.2	−25.7	−23.8
20	−32.7	−32.1	−19.9	−17.2
30	−26.3	−26.0	−13.6	−11.4
50	−12.9	−12.7	−0.2	+0.9
70	0.8	1.0	13.6	14.2
100	21.8	21.9	34.6	35.0
200	92.8	92.9	105.6	105.8
300	164.4	164.5	177.2	177.3
500	308.4	308.4	321.1	321.2

With $p = \frac{1}{2}$, $\alpha = 0.008$, and $H(f)$ as given by (55), $G(f)$ has 0.75 dB of peaking. In Table II we give $\hat{J}(N)$, $J(N)$, $\hat{I}(N)$, $I(N)$, $\sigma^2_{\Delta N}$, and $\sigma^2_{\theta N}$ [as defined by eqs. (44), (42), (45), (43), (40), and (41)] for various values of N. The entries for $\sigma^2_{\Delta N}$ and $\sigma^2_{\theta N}$ assume a worst-case $S_x(0)$ of 6.56×10^{-4}. The convergence of $\hat{J}(N)$ to $J(N)$ and $\hat{I}(N)$ to $I(N)$ for large N is readily apparent. It is also apparent that jitter accumulation limits line lengths on this particular system to about 70 repeaters where the rms alignment jitter is 0.03 slot (10 degrees). The entries in the table for larger N need qualification. One could not expect to measure them since, in that length line, lock would be lost. Our analysis assumes lock and makes many small signal approximations.

IX. CONCLUSIONS

By chopping the phase error $\phi_k - \theta_k$, transition gating injects jitter-producing wideband noise into a PLL extracting symbol timing from a baseband data waveform. If the wideband noise injected by the signal $(\phi_k - \theta_k)(d_k - p)$ is approximated as that injected by $\mu(d_k - p)$, the model for analyzing jitter in a PLL with transition gating becomes time invariant and can be easily analyzed.

X. ACKNOWLEDGMENTS

Ta-Mu Chien first introduced the author to the general problem of characterizing the jitter performance of digital repeaters with PLL timing extractors and, in particular, the theoretical problems associated with transition gating. Much of the timing-extractor circuitry in the T4M repeater was designed by J. A. Bellisio, who provided its characterization data. Throughout our work, M. R. Aaron shared

freely of his insight into this particular problem and his general knowledge of digital transmission.

APPENDIX

Calculation of Error Signal

In this appendix, we find the weights $\{e_k\}$ for the DZQ phase comparator as a function of the input phase, the VCO phase, the data sequence $\{a_k\}$, the standard pulse shape $f(t)$, and the additive channel noise $n(t)$. The procedure followed is quite similar to that used by Saltzberg[9] to find the weights $\{e_k\}$ for the zero-crossing phase comparator.

Assume first, for argument's sake, that $a_k = -1$ and $a_{k+1} = +1$. Then

$$r(t) = -f(t - k - \phi_k) + f(t - k - 1 - \phi_{k+1}) + \sum_{m \neq k, k+1} a_m f(t - m - \phi_m) + n(t). \quad (61)$$

The input jitter ϕ_k contains only low frequencies in comparison to the symbol rate. Thus, the near neighbors of ϕ_k approximately equal it and, assuming $f(t)$ is insignificantly small far out on its tails (as it must be for the eye to be open and regeneration possible), we can reasonably take $r(t)$ for $t \in [k, k+1]$ as

$$r(t) = -f(t - k - \phi_k) + f(t - k - 1 - \phi_k) + \sum_{m \neq k, k+1} a_m f(t - m - \phi_k) + n(t). \quad (62)$$

Define

$$q(t) = \sum_{m \neq k, k+1} a_m \dot{f}(t - m - \phi_k) + \dot{n}(t). \quad (63)$$

Then

$$\dot{r}(t) = -\dot{f}(t - k - \phi_k) + \dot{f}(t - k - 1 - \phi_k) + q(t). \quad (64)$$

Let τ^- denote the time in $[0, 1]$ when $\dot{f}(t-1) - \dot{f}(t)$ makes an upward crossing of the DZQ slicing level A and let τ^+ denote the time in $[0, 1]$ when it makes a downward crossing (see Fig. 5). During the time interval $[k + \phi_k, k + 1 + \phi_k]$, $\dot{r}(t)$ will appear as in Fig. 13. Define τ_k^- so that the input to the dead-zone quantizer makes an upward crossing of the slicing level A (that is, a pulse starts at the DZQ output) at time $k + \phi_k + \tau_k^-$. Denote the slope of $\dot{f}(t-1) - \dot{f}(t)$ at time τ^- by β^-. That is, let

$$\beta^- = \dot{f}(\tau^- - 1) - \dot{f}(\tau^-). \quad (65)$$

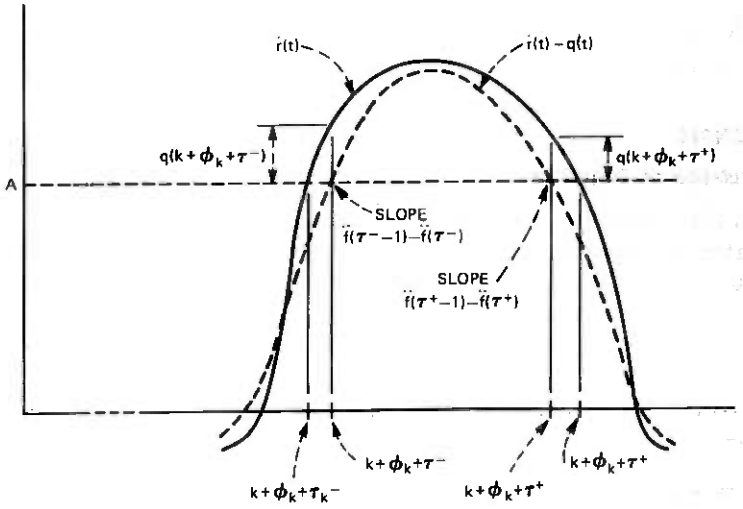


Fig. 13—Linear extrapolation for τ^- and τ^+.

To good approximation, we have (see Fig. 13)

$$\beta^- = \frac{q(k + \phi_k + \tau^-)}{\tau^- - \tau_k^-} \tag{66}$$

or, after rearrangement,

$$\tau_k^- = \tau^- - q(k + \phi_k + \tau^-)/\beta^-. \tag{67}$$

Defining τ_k^+ as the time when the pulse at the output of the dead-zone quantizer ends and defining

$$\beta^+ = \dot{f}(\tau^+ - 1) - \dot{f}(\tau^+), \tag{68}$$

we have, by a similar argument,

$$\tau_k^+ = \tau^+ - q(k + \phi_k + \tau^+)/\beta^+. \tag{69}$$

The output of the vco is $\sin [2\pi t - 2\pi\theta(t)]$. Assuming the phase-locked loop is in lock, that is, that $\theta(t) - \phi(t)$ is, at most, a small fraction of a slot, a downward crossing of zero by $\sin [2\pi t - 2\pi\theta(t)]$ must occur near the middle of the interval $[k + \phi_k, k + 1 + \phi_k]$. Define s_k so that

$$k + \phi_k + \tfrac{1}{2} + s_k$$

denotes the time of this occurrence. Then

$$[2\pi(k + \phi_k + \tfrac{1}{2} + s_k) - 2\pi\theta(k + \phi_k + \tfrac{1}{2} + s_k)] \bmod 2\pi = \pi. \tag{70}$$

The phase $\theta(t)$ is slowly varying with respect to the symbol interval.

Therefore,

$$\theta(k + \phi_k + \tfrac{1}{2} + s_k) \doteq \theta(k) = \theta_k. \tag{71}$$

With this approximation, (70) becomes

$$[2\pi(k + \phi_k + \tfrac{1}{2} + s_k) - 2\pi\theta(k)] \bmod 2\pi = \pi, \tag{72}$$

which implies ($\phi_k - \theta_k$ and s_k are small)

$$s_k = \theta_k - \phi_k. \tag{73}$$

From Fig. (14), it is clear that e_k must be proportional to

$$\tfrac{1}{2} + s_k - \tau_k^- - [\tau_k^+ - (\tfrac{1}{2} + s_k)], \tag{74}$$

which simplifies to

$$1 + 2s_k - \tau_k^- - \tau_k^+ \tag{75}$$

or, after substituting (74) for s_k,

$$2(\theta_k - \theta_k) + 1 - \tau_k^- - \tau_k^+. \tag{76}$$

Using the approximations (67) and (69) for τ_k^- and τ_k^+ and denoting the constant of proportionality relating e_k and (76) by $\alpha_1/2$, we have

$$e_k = \alpha_1(\theta_k - \phi_k) + \alpha_1(1 - \tau^- - \tau^+)/2 + \alpha_1\left\{\frac{q(k + \phi_k + \tau^-)}{2\beta^-} + \frac{q(k + \phi_k + \tau^+)}{2\beta^+}\right\}. \tag{77}$$

Let

$$v_k = \frac{\dot{v}(k + \phi_k + \tau^-)}{2\beta^-} + \frac{\dot{v}(k + \phi_k + \tau^+)}{2\beta^+} \tag{78}$$

and

$$\epsilon_k = \frac{\dot{f}(k + \tau^-)}{2\beta^-} + \frac{f(k + \tau^+)}{2\beta^+} \tag{79}$$

and assume that the pulse $f(t)$ has without loss of generality been centered so that

$$\tau^- + \tau^+ = 1. \tag{80}$$

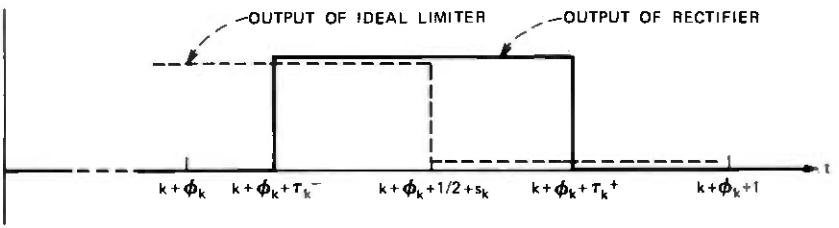


Fig. 14—Inputs to multiplier.

Then, using (63), we obtain

$$e_k = \alpha_1(\theta_k - \phi_k + v_k + \sum_{m \neq k, k+1} a_m \epsilon_{k-m}). \tag{81}$$

To derive (81), we assumed a_k equal to -1 and a_{k+1} equal to $+1$. If the signs are reversed, a similar analysis shows

$$e_k = \alpha_1(\theta_k - \phi_k - v_k - \sum_{m \neq k, k+1} a_m \epsilon_{k-m}). \tag{82}$$

If $a_k = a_{k+1}$,

$$e_k = 0, \tag{83}$$

since no transition occurs. By defining

$$d_k = \begin{cases} 1, & a_k \neq a_{k+1} \\ 0, & a_k = a_{k+1} \end{cases} = \frac{(1 - a_k a_{k+1})}{2}, \tag{84}$$

we can combine all three of these equations, and write for all a_k, a_{k+1}

$$e_k = \alpha_1 d_k(\theta_k - \phi_k - a_k v_k - a_k \sum_{m \neq k, k+1} a_m \epsilon_{k-m}). \tag{85}$$

REFERENCES

1. M. R. Aaron, "PCM Transmission in the Exchange Plant," B.S.T.J., *41*, No. 1 (January 1962), pp. 99–141.
2. E. D. Sunde, "Self-Timing Regenerative Repeaters," B.S.T.J., *36*, No. 4 (July 1957), pp. 891–938.
3. W. R. Bennett, "Statistics of Regenerative Digital Transmission," B.S.T.J., *37*, No. 6 (November 1958), pp. 1501–1542.
4. H. E. Rowe, "Timing in a Long Chain of Regenerative Binary Repeaters," B.S.T.J., *37*, No. 6 (November 1958), pp. 1543–1598.
5. M. R. Aaron and J. R. Gray, "Probability Distribution for the Phase Jitter in Self-Timed Reconstructive Repeaters for PCM," B.S.T.J., *41*, No. 2 (March 1962), pp. 503–558.
6. C. J. Byrne, B. J. Karafin, and D. B. Robinson, "Systematic Jitter in a Chain of Digital Regenerators," B.S.T.J., *42*, No. 6 (November 1963), pp. 2679–2714.
7. J. M. Manley, "The Generation and Accumulation of Timing Noise in PCM Systems—An Experimental and Theoretical Study," B.S.T.J., *48*, No. 3 (March 1969), pp. 541–613.
8. T. Takaci, T. Saito, and K. Mano, "Evaluation of Phase Jitter in a Limiter-Tuned Circuit PCM Self-Timing System in Band-Limited Baseband Gaussian Channel," Report of the Research Institute of Electrical Communication, Tohoku University, *22*, 1971, pp. 41–55.
9. B. R. Saltzberg, "Timing Recovery for Synchronous Binary Data Transmission," B.S.T.J., *46*, No. 3 (March 1967), pp. 593–622.
10. E. Roza, "Analysis of Phase-Locked Timing Extraction Circuits for Pulse Code Transmission," IEEE Trans. on Communications, *COM-22*, September 1974, pp. 1236–1249.
11. F. D. Waldhauer, "A Two-Level 274 Mb/s Regenerative Repeater for T4M," Proceedings of the 1975 International Conference on Communications, June 1975.
12. A. Papoulis, *The Fourier Integral and Its Applications*, New York: McGraw-Hill, 1962.
13. J. T. Harvey and J. W. Rice, "Random Timing Noise Growth in a Cascaded Digital Regenerator Chain," IEEE Trans. on Communications, *COM-21*, August 1973, pp. 969–971.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 1, January 1976
Printed in U.S.A.

Some Properties of the Variance of the Switch-Count Load

By A. DESCLOUX

(Manuscript received August 6, 1975)

Under equilibrium conditions, the load carried by a service system is defined as the average amount of traffic handled per unit of time. An unbiased estimate of this parameter is provided by the switch-count load, which is obtained by recording the number of busy servers at regular time intervals and then taking the arithmetic mean of these observations.

Formulas for the variance of this measurement (which are applicable to delay-and-loss systems with either finite- or infinite-source inputs and arbitrary defection rates) were derived in a previous paper[1]; a program for their computation is now available and has been used to explore effects of parameter changes on the switch-count load variance. The purpose of the present paper is to describe results of this investigation and, in particular, to draw attention to two properties which may be unexpected: (i) the variance of the switch-count load does not always decrease when waiting positions are added, and (ii) the variance of the carried-load estimate obtained from continuous observation over a given time interval is not a lower bound for the variance of load estimates calculated from a finite sequence of recording of the number of busy servers.

I. INTRODUCTION

When statistical equilibrium prevails, the load carried by a group of servers is defined as the average amount of traffic handled per unit of time. In telephony, this parameter is often evaluated by "switch-counting."[2,3] According to this method, the number of servers in use is recorded at regular time intervals; these numbers are then added together and their sum, divided by the number of observations (scans), is an unbiased estimate of the carried load. This measurement is called hereafter the switch-count load to distinguish it from estimates based on continuous observation. The latter are obtained by dividing the aggregated usage of the servers by the length of the measurement interval and can be viewed as limits of switch-count load measurements in which the number of scans tends to infinity while the length of the observation period is kept unchanged.

The problem of finding the variance of the switch-count load in loss systems with exponential service times has attracted a good deal of attention. Description of some earlier contributions to this subject can be found in Ref. 1.

In a recent paper,[1] formulas for the variance of the switch-count load were derived for delay-and-loss systems with state-dependent input rates and exponential service times. (As is customary, the state of the system at some instant, t, is defined as the number of customers who are either being served or are waiting at that time.) More precisely, the assumptions made here are as follows:

(*i*) Calls originate at rate λ_n (>0) whenever the system is in state n. We consider here only the two cases where the λ_n's are either independent of n (Poisson input) or are proportional to the number of idle sources (finite-source input).

(*ii*) Requests which originate when no free server is available are either delayed or lost; they are delayed if a waiting position is available and lost otherwise. While waiting, the requests are allowed to defect from the system at the same constant individual rate, j.

(*iii*) The service times are exponentially distributed; they are also independent of each other and of the state of the system. The average service time is taken throughout as the unit of time.

As in Ref. 1, we make the assumption that the number of servers and the number of waiting positions are finite. In the computations, however, we have to be more restrictive since numerical as well as storage problems may limit their ranges. But this, as it turns out, does not preclude investigations of rather large systems, and even of queues with Poisson input and infinite waiting room, so long as they can be approximated by systems that are computationally manageable. From a practical point of view, this treatment of the delay systems with unboundable queues has been found to be satisfactory even when they are nearly saturated.

A program for the computation of the formulas derived in Ref. 1 has been written and has been used to investigate the effects of parameter changes on the variance of the switch-count load—special attention being paid to the influence of delays on the variability of this measurement. The purpose of this expository paper, which serves as a complement to Ref. 1, is to describe the numerical results obtained thus far; these, in turn, suggest some general qualitative characteristics of the switch-count load variance that may be unexpected at times. These properties are "read off" the graphs and tables; they are

stated and explained informally in the following discussions. Their proofs will be presented in a subsequent paper.

The original goal of the computations was to determine the scope of the approach developed in Ref. 1 and, principally, to determine its limits of accuracy as the number of servers and/or waiting positions become large. Some conclusions in this regard appear in Ref. 1. We mention here only that a thorough examination of the numerical stability of the computations was carried out for systems with as many as 400 devices and that, from a practical point of view, no significant loss in accuracy could be detected over this range. (By device, we mean here either a server or a waiting position.)

II. NOTATION AND DEFINITIONS

Following is a list of symbols and definitions used throughout:

- c = number of servers,
- d = number of devices (=number of servers + number of waiting positions),
- s = number of sources (s is considered to be infinite whenever the input is Poissonian),
- a = offered load in erlangs (this symbol is used only when the input is Poissonian),
- λ = demand rate of an idle source—i.e., of a source that is neither being served nor waiting for service,
- Λ = $s \cdot \lambda$ (this symbol as well as λ are used only in the case of finite-source inputs),
- $h.t.$ = average service-time (used throughout as the unit of time),
- j = individual defection rate of the waiting requests (j is Palm's j-factor; it is equal to zero whenever waiting requests do not defect and to infinity when waiting is not allowed),
- T = length of the observation period (in multiples of $h.t.$),
- n = number of recordings (scans) made during the observation period,
- τ = interval between consecutive recordings of the number of busy servers in multiples of $h.t.$,
- $N_c(t)$ = number of busy servers at time t ($0 \leqq N_c(t) \leqq c$),
- $L_n(T)$ = switch-count load based on n scans spaced τ ($= T/n$) apart over an interval of length T.

Clearly, the n observations which enter in the computation of $L_n(T)$ can be made in many ways. If we take the beginning of the observation period as the time origin, then for any θ such that $0 \leqq \theta \leqq \tau$, the

instants θ, $\theta + \tau$, $\theta + 2\tau$, $\cdots$, $\theta + (n-1)\tau$, constitute a possible scanning sequence. Under equilibrium conditions, the statistical properties of $L_n(T)$ are independent of θ and, for the sake of definiteness, we shall set it equal to τ so that

$$L_n(T) = \frac{1}{n}[N_c(\tau) + N_c(2\tau) + \cdots + N_c(n\tau)], \qquad n\tau = T.$$

When the switch-count load is measured as described above, we shall say, whenever emphasis is needed (and only then), that the measurement is of type I. Sections III through VI as well as VIII pertain only to these measurements. Type II measurements are introduced and dealt with in Section VII, while pertinent numerical examples are presented in Tables I through IV.

III. QUEUING EFFECT

Superficially, it would seem that, as time elapses, the number of busy servers is less volatile when waiting is allowed than when it is not. The reason sometimes advanced to support this view is that a comparison of a loss system with a delay system having the same number of servers would reveal that, for a given offered load, the mean number of busy servers tends to be smaller for the loss system than for the delay system, and that the "holes" in the carried-load process (see Fig. 1) of a loss system would be shortened and partially filled if the blocked calls were allowed to wait. If this were the case, the traffic fluctuations would be dampened and the conclusion could then be drawn that the variance of the switch-count load must decrease as the number of waiting positions increases. (Throughout this and the next three sections, the scanning rate is assumed to be fixed.) Ac-

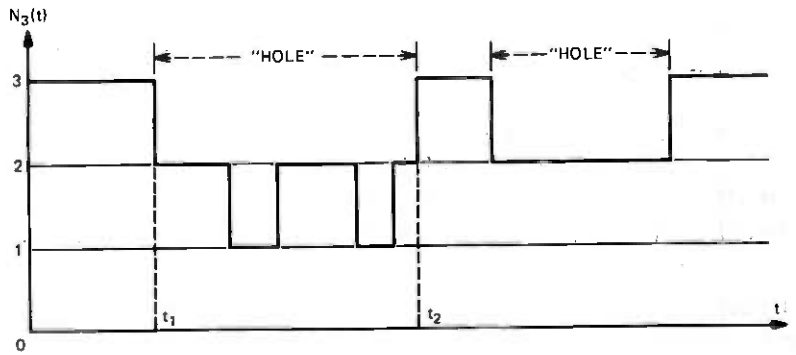


Fig. 1—Carried-load process in a loss system with three servers.

cordingly, for a given offered load, the variance of the switch-count load would be largest for loss systems and could therefore serve as an upper bound for the switch-count load variance in delay-and-loss systems.

Under the present conditions, the preceding argument is readily seen to be fallacious. Indeed, let t_1 be an instant at which an interval of full server occupancy terminates and let t_2 be the instant when, for the first time after t_1, all the servers are again occupied (see Fig. 1). Because of the assumptions made here regarding the input and the disposition of the requests, the behavior of the carried-load process over (t_1, t_2) is unaffected by what took place prior to t_1; hence, there cannot be any tendency for the "holes" to be filled since the distribution of their lengths remains the same and the evolution of the process over such "holes" is unchanged by the occurrence of delays. The only thing that happens is that the intervals of full uninterrupted occupancy tend to be of longer duration when queuing is permitted; the "holes" themselves are merely shifted.

The question of whether or not the variance of the switch-count load always decreases as the number of waiting positions increases has a clear-cut answer: it is "no," since the behavior of the switch-count load regarded as a function of the number of devices depends essentially on the offered load. We state next a few general properties.

So long as the offered load is light, the variance of the switch-count load increases monotonically as the number of waiting positions increases (see Fig. 2a). At higher loads, which, however, still fall below c, the variance first increases and then decreases monotonically toward a positive value as the number of waiting positions increases (see Fig. 2b). A similar behavior can be observed when $a = c$ and at "moderate" loads in excess of c, but with one difference, namely that the variance now tends to zero as the number of waiting positions tends to infinity (see Fig. 2c). Finally, when $a > c$ and is sufficiently high, the variance decreases monotonically and tends to 0 as the number of waiting positions increases (see Fig. 2d). It can be shown that the variance of the switch-count load always behaves in this manner and that, regardless of the number of servers, the lengths of the observation period, and the number of scans, each of the four patterns sketched in Fig. 2 does occur for some values of the offered load. Figure 3 depicts this rather intricate behavior of the switch-count load variance in the case of a four-server system with Poisson input and no defection from the queue.

For all values of T, the variance of $L_1(T)$ is equal to the variance, $\sigma^2_{c,a}$, of the equilibrium distribution of the number of busy servers.

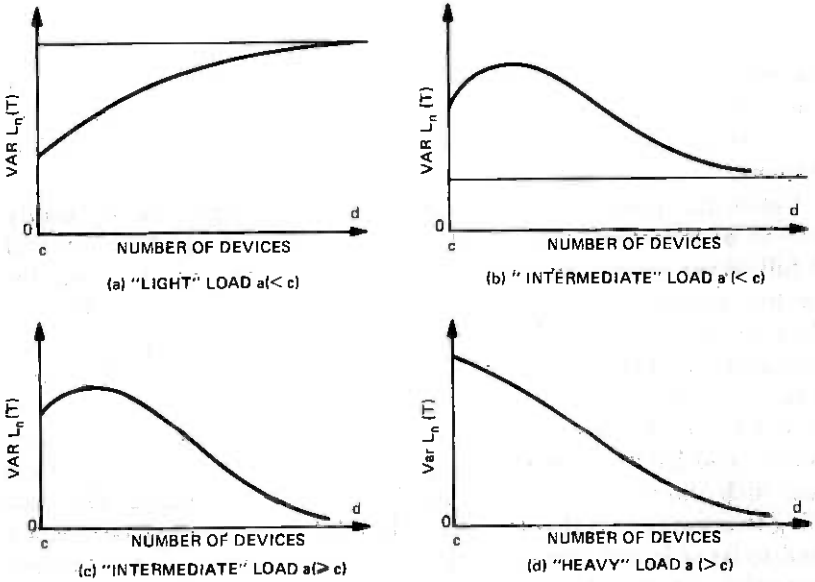


Fig. 2—Variance of switch-count load vs number of devices.

The latter, therefore, also display the characteristic behavior of the switch-count load variance. We can of course proceed in the opposite direction: we can first determine the behavior of $\sigma^2_{c,d}$ and, hence, of Var $L_1(T)$, and then anticipate some of the properties of the variance of $L_n(T)$. This is done next.

Let $P_{c,d}(n)$ be the equilibrium probability of n servers busy in a system with c servers and d devices. Let $\mathcal{P}_{c,d} \equiv \{P_{c,d}(0), P_{c,d}(1), \cdots, P_{c,d}(c)\}$, $\mathcal{P}_c \equiv \mathcal{P}_{c,c}$ and $\mathcal{U}_c$ be the distribution whose total probability mass is concentrated at c. Under the present conditions, the state probabilities are governed by the familiar birth-and-death equations and these imply that the ratios between the probabilities $P_{c,d}(n)$, $n = 0, 1, \cdots, c-1$, are independent of d. Hence,

$$\mathcal{P}_{c,d} = q_d\mathcal{P}_c + (1 - q_d)\mathcal{U}_c, \qquad 0 \leqq q_d \leqq 1,$$

where $(1 - q_d)$ is the probability that at least one waiting position is occupied ($q_c = 1$).

Let m_c be the mean of $\mathcal{P}_c$ and $\sigma_c^2 \equiv \sigma^2_{c,c}$ its variance. Simple calculations show that

$$\sigma^2_{c,d} = q_d\sigma_c^2 + q_d(1 - q_d)(c - m_c)^2.$$

Now, regarded as a function of q only,

$$V_c(q) \equiv q\sigma_c^2 + q(1 - q)(c - m_c)^2$$

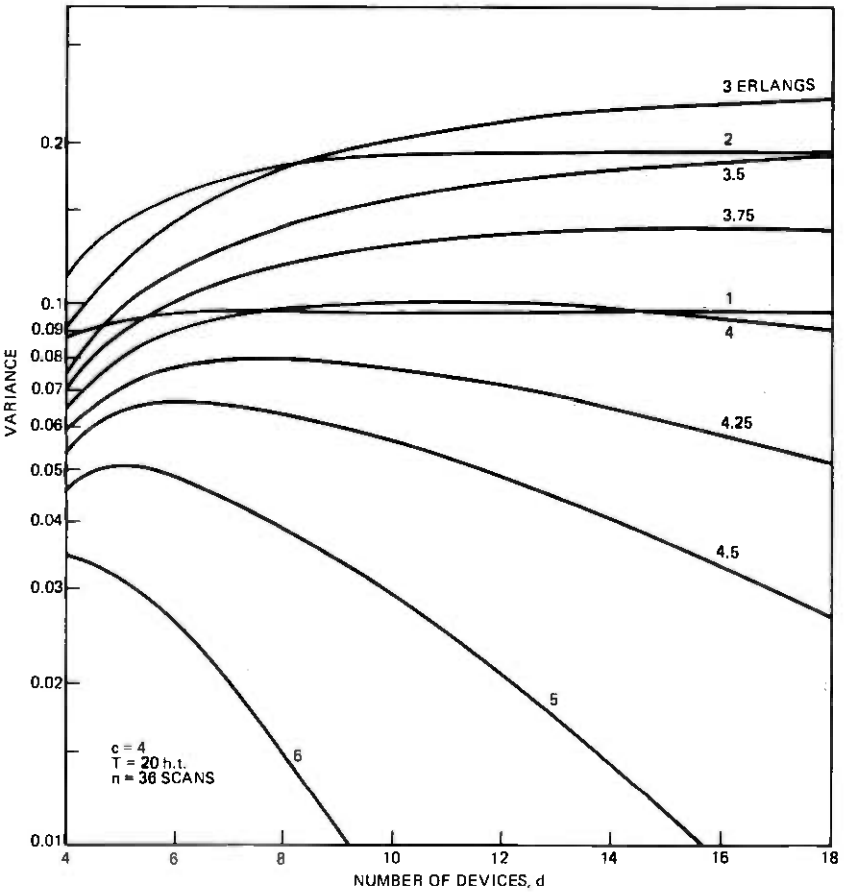


Fig. 3—Variance of the switch-count load (Poisson input).

is increasing whenever

$$q^* \equiv \frac{\sigma_c^2 + (c - m_c)^2}{2(c - m_c)^2} > q \qquad (\geqq 0),$$

and decreasing when $q^* < q$.

As the load tends to ∞ (c being kept fixed), q^* remains bounded away from 0 while q_d and q_{d+1} tend to 0. Hence, for sufficiently large values of d the probabilities q_d and q_{d+1} are both smaller than q^* and, since $q_{d+1} < q_d$ for all d's, we have

$$\sigma^2_{c,d+k+1} - \sigma^2_{c,d+k} = V_c(q_{d+k+1}) - V_c(q_{d+k}) < 0, \qquad k = 0, 1, 2, \cdots.$$

Conversely, as the load tends to 0, q^* tends to $\frac{1}{2}$ while q_d and q_{d+1} tend to 1. Hence, for sufficiently small loads, $q^* < q_{d+1} < q_d$ and the preceding inequality is reversed.

These considerations explain how the behavior of $\sigma^2_{c,d} = \text{Var } L_1(T)$, as d varies, is governed by two simple facts, namely that (*i*) transfers of "sufficiently small" probability masses to the sample value that is farthest from the mean lead to distributions with greater variances, and that (*ii*) an increased concentration of the probability mass in the vicinity of the mean is accompanied by a decrease of the variance. As we have just seen, such changes of the probability distribution of the number of busy servers can be induced by changing the number of waiting positions. Thus, so long as the offered load, a, does not exceed a certain bound (a_1 in the example of Fig. 4), Var $L_1(T)$ increases monotonically as d increases. Conversely, whenever a is larger than a specific value (a_2 in Fig. 4), Var $L_1(T)$ decreases monotonically as d increases. However, because of the discreteness of d, this monotonic decrease may occur over a wider range (for $a \geqq a_3$ in Fig. 4; a_3, in this instance, falls just short of 4). Finally, there is an intermediate range ($a_1 \leqq a < a_3$ in Fig. 4) where Var $L_1(T)$ first increases and then decreases monotonically as d increases from c to ∞.

We now turn our attention to Var $L_n(T)$, $n > 1$. The behavior of this variance is more difficult to elucidate because $L_n(T)$ is now the arithmetic mean of n correlated random variables. The informal argument used in the preceding paragraphs may nevertheless be modified so as to cover this new situation. For given a, c, and d ($>c$),

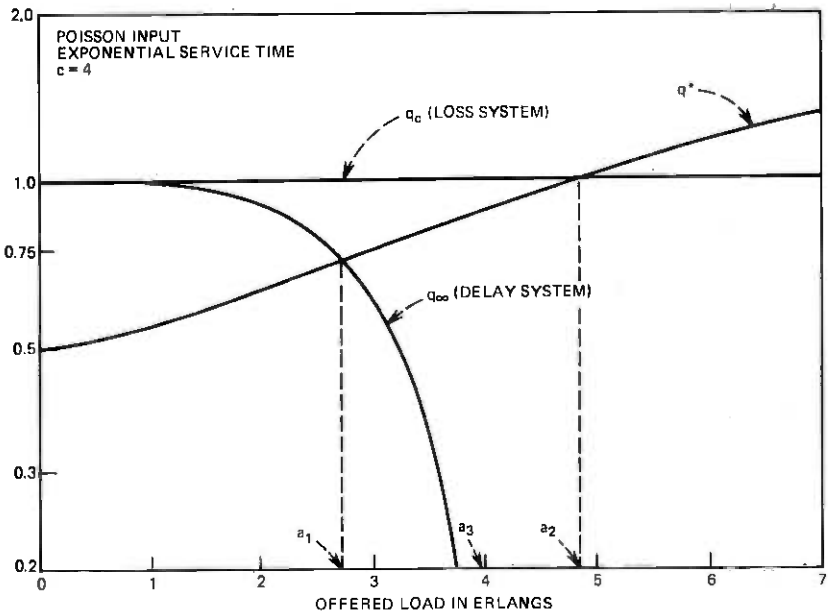


Fig. 4—Parameters q_c, q_∞, and q^* as functions of the offered load.

consider the aggregate (ensemble) of all possible switch-count load measurements over a given time interval of length T. This finite aggregate may be split into two disjoint classes: Class 1 includes all the measurements for which the number of busy devices does not exceed c at any of the scanning instants $\tau, 2\tau, 3\tau, \cdots, n\tau$; Class 2 comprises all the other measurements. [The possible values of Class 1 measurements are $0, 1/n, 2/n, \cdots, c(=cn/n)$, and those of Class 2 are $c/n, (c+1)/n, \cdots, c$.] Given that a measurement is of Class 1, let $\mathcal{P}_c^{(1)}$ be its conditional distribution. Similarly, let $\mathcal{P}_{c,d}^{(2)}$ be the conditional distribution of the Class 2 measurements.

Under the present assumptions, $\mathcal{P}_c^{(1)}$ is identical to the distribution of $L_n(T)$ for $d = c$ and is thus independent of d. The roles played by $\mathcal{P}_c$ and $\mathcal{U}_c$ above are now taken over by $\mathcal{P}_c^{(1)}$ and $\mathcal{P}_{c,d}^{(2)}$, respectively, and the distribution, $\mathcal{Q}_{c,d}$, of $L_n(T)$ is therefore given by

$$\mathcal{Q}_{c,d} = \mathfrak{q}_d \mathcal{P}_c^{(1)} + (1 - \mathfrak{q}_d)\mathcal{P}_{c,d}^{(2)},$$

where $\mathfrak{q}_d$ is the probability that a measurement is of Class 1.

By assumption, the system is in equilibrium at time 0 and the probability that at least one waiting position is occupied at any given scanning instant increases with d. Thus, as d becomes larger, the proportion of Class 2 measurements increases. It also stands to reason, however, that the values of these measurements tend to be larger than those of Class 1 and that their magnitudes also increase with d. Thus, we may expect a greater proportion of relatively high load measurements as d increases. As before, and so long as the offered load is sufficiently small, the appearance of these "more extreme" values produce a scattering of the probability masses and this will tend to magnify the variance of $L_n(T)$. But when the offered load exceeds a certain level, the Class 1 measurements are, on the average, just about as large as those of Class 2 and increases of d make "relatively small" carried-load measurements less likely. This leads to a concentration of the probability masses about the mean of $L_n(T)$ and brings down its variance. These effects are clearly visible in Figs. 3, 5, and 6.

IV. EFFECT OF DEFECTIONS

In the preceding section we have attempted to explain how changes in the number of waiting positions affect the variance of the switch-count load. The arguments advanced depend on the fact that "more queuing" tends to increase the probability of full server-occupancy (the all-server-busy state) at the expense of the probabilities of the states $0, 1, 2, \cdots, c-1$, while leaving the ratios between the latter unchanged. For fixed values of c and d, similar transfers of probability masses—with analogous consequences for Var $L_n(T)$—can be induced

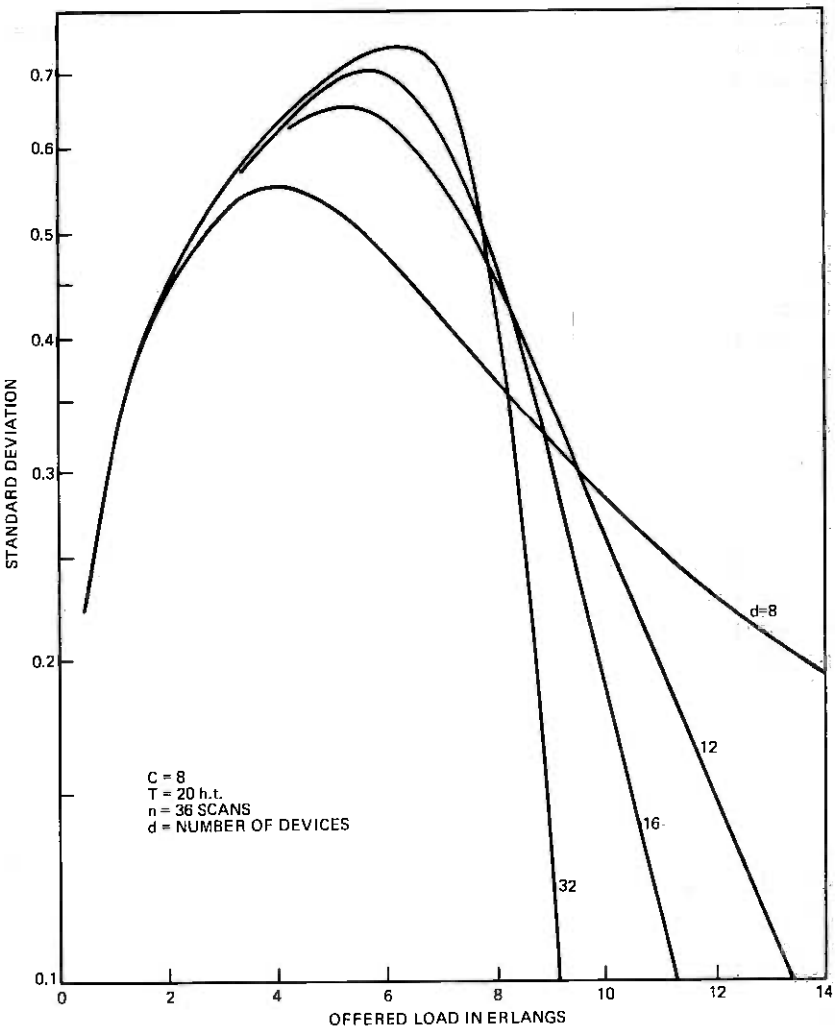


Fig. 5—Standard deviation of the switch-count load (Poisson input).

by varying the rates at which requests may defect from the waiting line.

If we assume, for example, that the requests defect at a constant individual probability rate, j, then the probability of full occupancy increases monotonically as j decreases. (Note that when $j = 0$, all delayed calls wait until served, and that whenever $j = \infty$, no waiting ever occurs and all the blocked calls are "lost." The familiar "blocked-calls-held" assumption corresponds to a j value of 1.) We may therefore expect that, for a given offered load, the variance of $L_n(T)$ will,

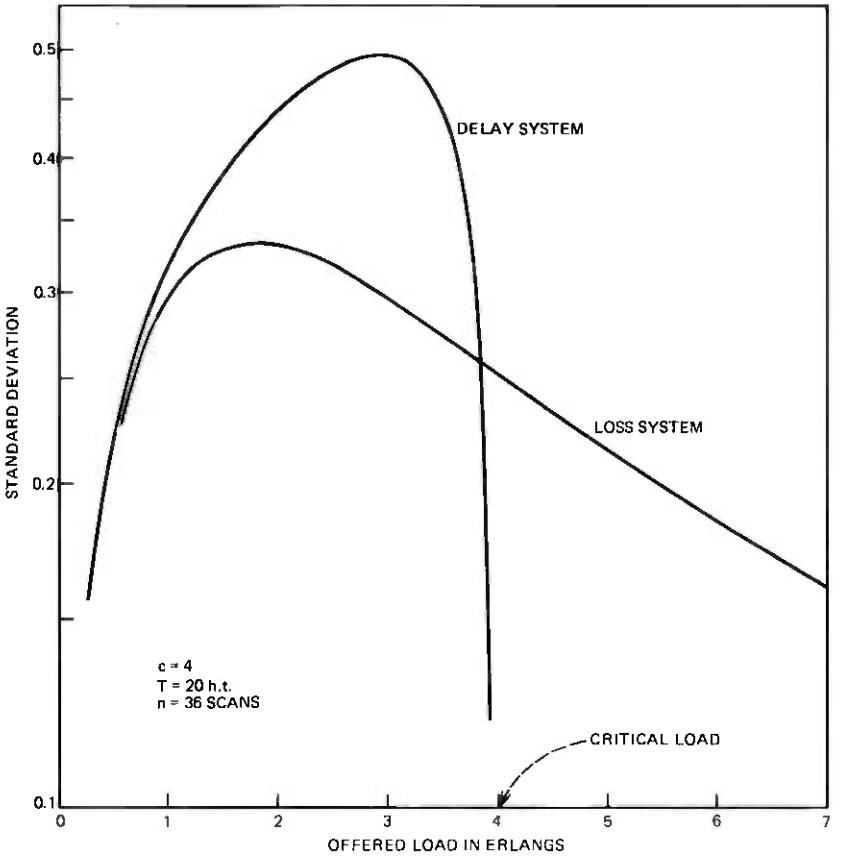


Fig. 6—Standard deviation of the switch-count load, delay vs loss system (Poisson input).

as j varies, display a similar behavior to that observed when the number of waiting positions changes. That this is indeed the case can be seen in Fig. 7. Hence, we may state the following: For given a, c, and d, and decreasing values of j, the variance of the switch-count load increases for sufficiently small values of the offered load and decreases whenever a is large enough. Also, there are values of the offered load for which intermediate defection rates do not imply intermediate values of Var $L_n(T)$.

V. FINITE-SOURCE EFFECT

A similar situation obtains—and admits of a similar explanation—when the input is generated by a finite number of sources (see Fig. 8). In this case, as the number of sources increases, so does the probability

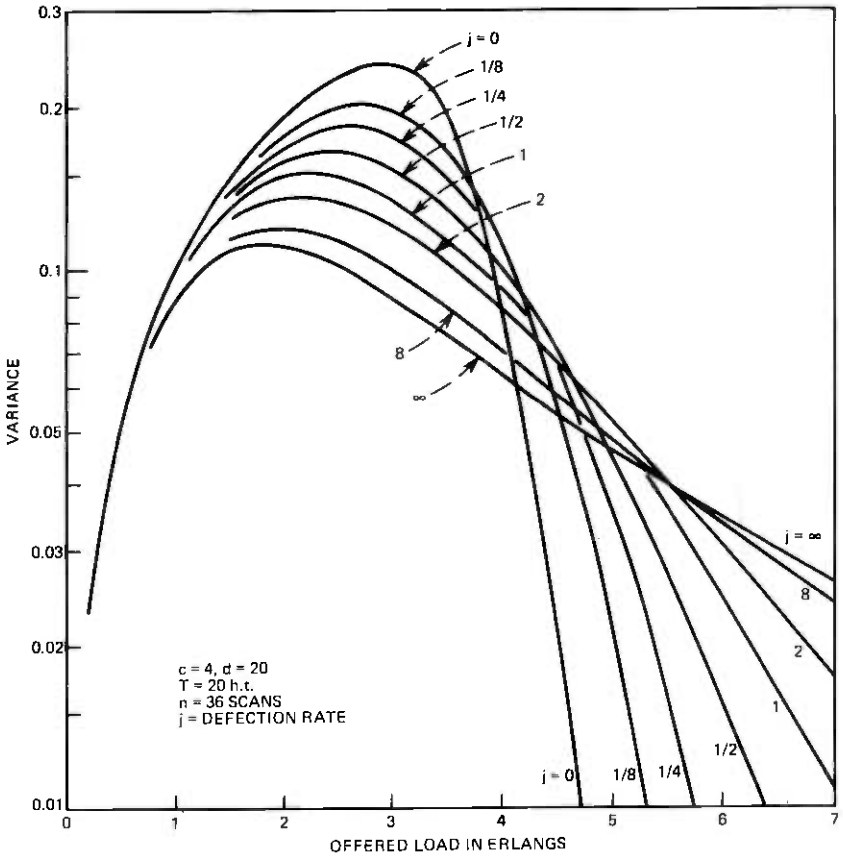


Fig. 7—Effect of defection rate on variance of the switch-count load, delay-and-loss system (Poisson input).

of waiting and Var $L_n(T)$, as a function of a, behaves as above. However, in these systems, the (overall) offered load is somewhat elusive as it depends not only on λ and the number of sources, s, but also on the structural parameters c and d and on the defection rate. Thus, the dependence of Var $L_n(T)$ on the number of sources is shown in Fig. 8 for prescribed values of $\Lambda = s \cdot \lambda$.

Since

$$\lambda_n = (s - n)\lambda = \Lambda\left(1 - \frac{n}{s}\right), \qquad n = 0, 1, \cdots, d - 1,$$

the rate λ_n, for any given n, therefore increases with s. This in turn implies that, for fixed Λ, the offered load increases with s. Hence, if the abscissa in Fig. 8 had been the offered load (instead of Λ), the

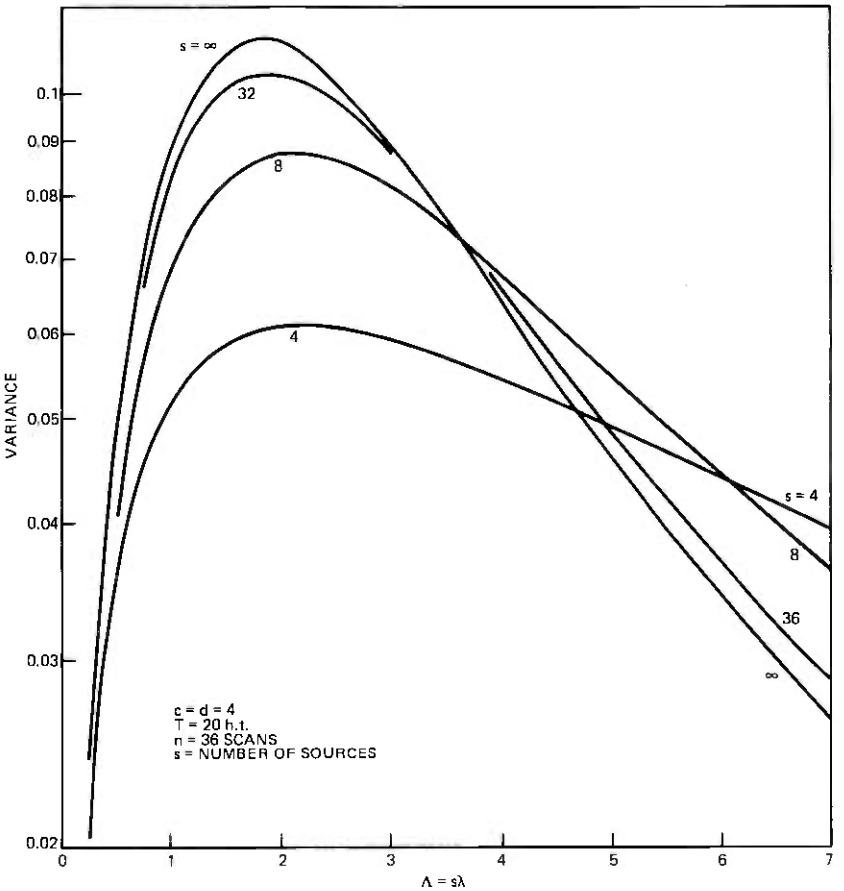


Fig. 8—Variance of the switch-count load, loss system (Poisson and finite-source inputs).

plotted points, with the exception of the origin and of those on the $s = \infty$ curve, would have been moved to the right. It is apparent, however, that such a change of abscissa would not have destroyed the overall incidence pattern depicted in Fig. 8.

VI. ANOTHER VIEW

We have studied thus far the effects of various parameter changes on Var $L_n(T)$ for known values of the (individual or overall) demand rates. This approach is particularly convenient since it gives us a direct handle on the state probabilities and, as we have seen, the behavior of Var $L_n(T)$ was then readily predictable in these terms. However, no general monotonicity property emerged within this frame-

work. But if, instead of the offered load, we use the carried load as primary variable, we obtain a very different and far less intricate picture which, in turn, leads to a simple general rule.

From Figs. 9 to 12, we may infer that:

(*i*) For a given number of servers and a given value of the carried load, Var $L_n(t)$ decreases monotonically as the number of waiting positions decreases (see Fig. 9).

(*ii*) For given c and d, and a prescribed value of the carried load,

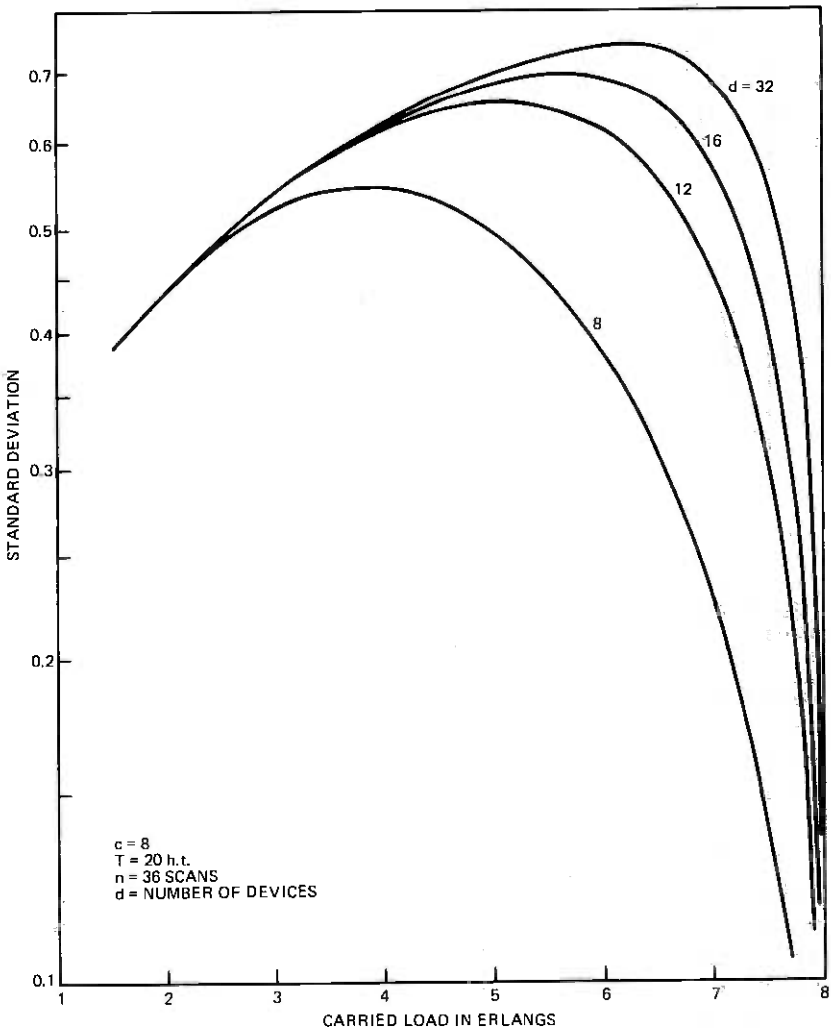


Fig. 9—Standard deviation of the switch-count load (Poisson input).

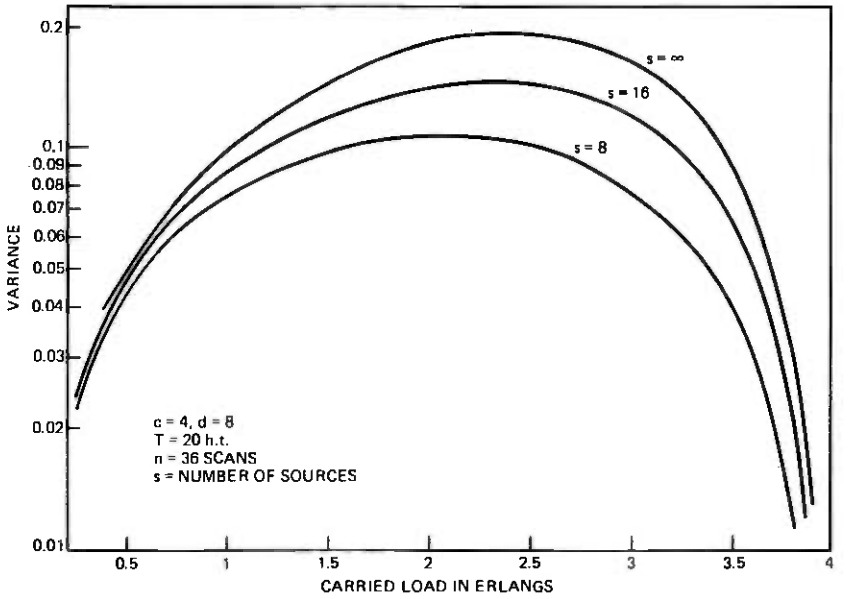


Fig. 10—Variance of the switch-count load, delay-and-loss system (Poisson and finite-source inputs).

Var $L_n(T)$ decreases monotonically as the number of traffic sources decreases (see Figs. 10 and 11).

(*iii*) For given c and d, and a fixed value of the carried load, Var $L_n(T)$ decreases monotonically as the defection rate, j, increases (see Fig. 12).

We note that these decreases of Var $L_n(T)$ are accompanied, in all cases, by increases in the number of calls that must be offered to maintain the carried load at a prescribed level. This brings us back to a "hole-filling" argument (cf. Section III) with a new twist, the prescription of the carried load that makes it operative. And now we see that, at a given occupancy, the "holes" are filled by allowing less rather than more queuing! Indeed, with less queuing, the average length of the busy periods is shortened and it is, therefore, necessary to fill the "holes" partially so as to maintain the carried load at its designated level. And with such compensatory fillings, the variance of the switch-count load can be expected to—and actually does—go down. These arguments provide us with an intuitive justification of inferences (*i*) through (*iii*) since, for a constant carried load, either decreasing the number of waiting positions or the number of sources or, alternatively, increasing the defection rate tends to reduce queuing.

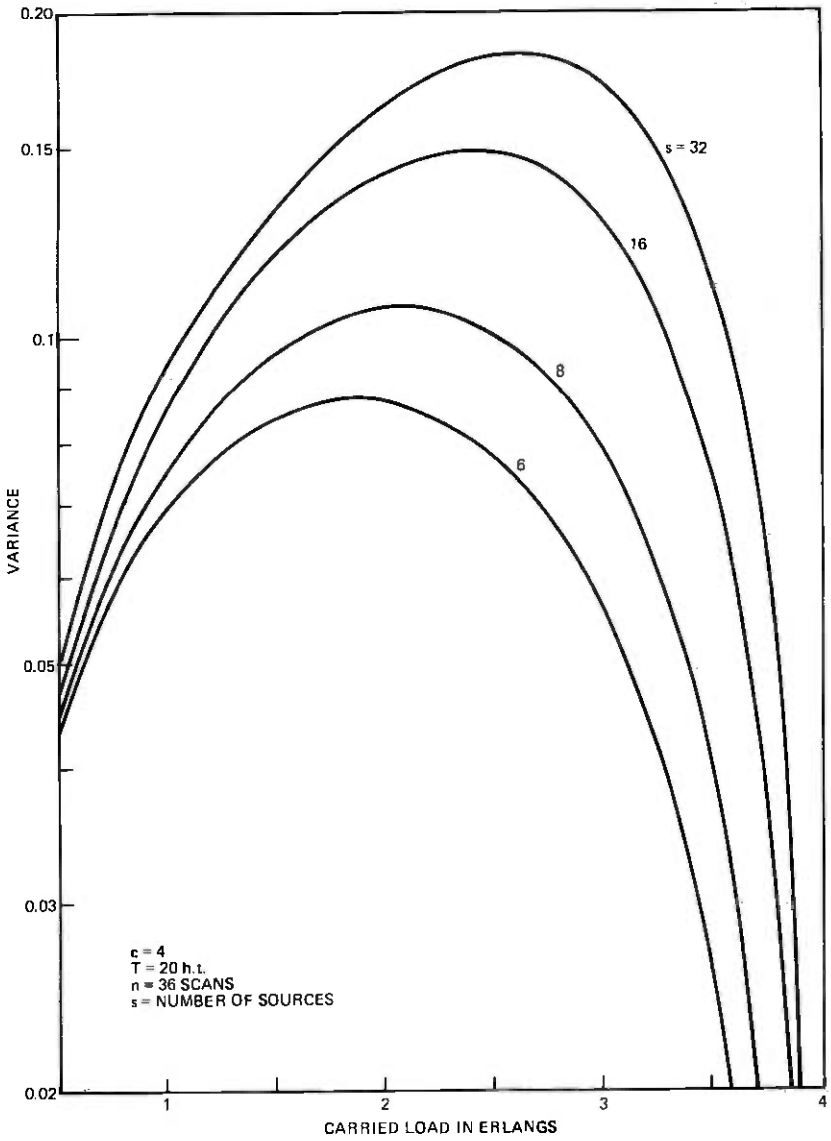


Fig. 11—Variance of the switch-count load, delay system (finite-source input).

The preceding considerations remain valid if one prescribes the blocking probability instead of the carried load. This is borne out by the data presented in Fig. 13.

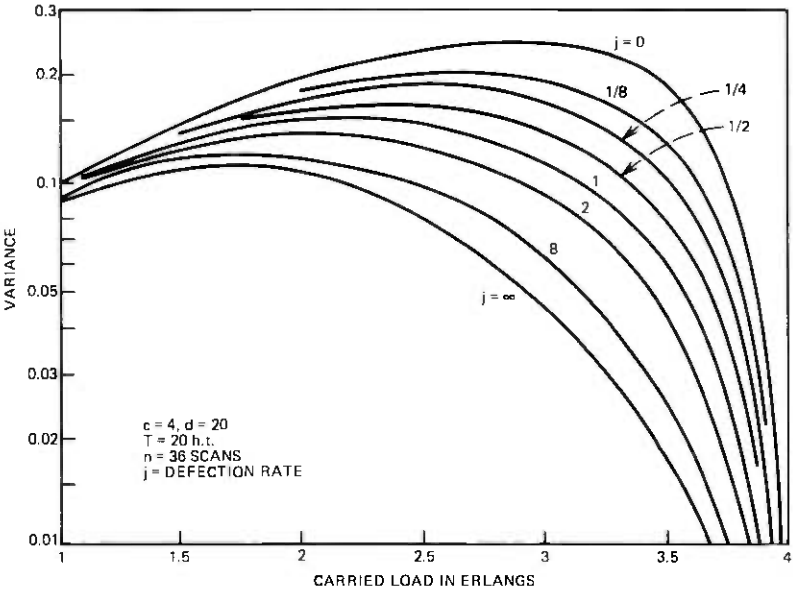


Fig. 12—Effect of the defection rate on the variance of the switch-count load, delay-and-loss system (Poisson input).

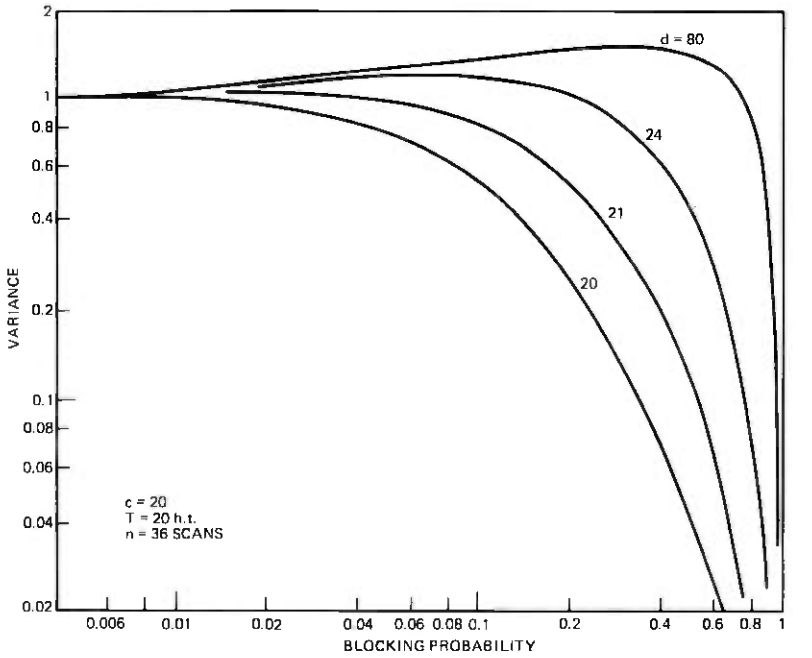


Fig. 13—Variance of the switch-count load vs blocking probability (Poisson input).

VII. EFFECT OF THE SCANNING RATE

We start with the statements of two properties. (In the sequel, c, d, s, and j are assumed fixed.)

(i) For a given value of the offered load and a given length of the observation period, the variance of the switch-count load decreases monotonically as the number of scans increases.

(ii) For a given offered load and a given number of scans, the variance of the switch-count load decreases monotonically as the length of the observation period increases.

All that is needed to prove the last assertion is a straightforward application of the familiar formula for the variance of a sum of correlated random variables (Ref. 4, pp. 229 ff) and use of the fact that the covariance, $R(t)$, between two observations of the number of busy servers made t apart decreases as t increases.

For given T and n, we have always assumed thus far that the n scans were made $\tau = T/n$ apart. Under this circumstance, it can be proved that (i) above is satisfied. But, as we shall see, (i) may fail to hold if the n scanning instants are chosen in a different way. We shall make use of this unexpected fact to show that the carried-load measurement obtained by continuous observation is never a minimum-variance estimate of the carried load.

For given T and n, the switch-count load was defined by the relation

$$L_n(T) \equiv \frac{1}{n}[N_c(\tau) + N_c(2\tau) + \cdots + N_c(n\tau)], \qquad n\tau = T.$$

In the following discussion, this measurement is said to be of type I and, accordingly, we shall designate it by $L_n^{\mathrm{I}}(T)$ instead of $L_n(T)$. The switch-count load measurements of type II are defined as follows:

$$L_n^{\mathrm{II}}(T) \equiv \frac{1}{n}\{N_c(0) + N_c(\tau') + \cdots + N_c[(n-1)\tau']\},$$

where $\tau' = T/(n-1)$. As defined here, measurements of type II differ from those of type I in that a recording is made at each end of the observation period (see Fig. 14). To avoid minor qualifications, we also define $L_1^{\mathrm{II}}(T)$ by setting it equal to $L_1^{\mathrm{I}}(T)$.

Figure 14 shows that each measurement of type I may also be regarded as a measurement of type II, and conversely. Thus, for $\tau = T/n$, the two random variables $L_n^{\mathrm{I}}(T)$ and $L_n^{\mathrm{II}}(T-\tau)$ are equidistributed for all n. In the present context, however, it is useful to make a distinction between type I and type II measurements because, as shown next, their respective variances do not behave in the same manner as the number of scans increases.

The load measurements $L_n^{\mathrm{I}}(T)$ and $L_n^{\mathrm{II}}(T)$ are sums of n correlated, but identically distributed, random variables, and their respective variances are, therefore, given by (see Ref. 4, p. 229):

$$\operatorname{Var} L_n^{\mathrm{I}}(T) = \frac{1}{n^2}\{nR(0) + 2(n-1)R(\tau) + 2(n-2)R(2\tau) + \cdots + 2R[(n-1)\tau]\}, \qquad \tau = T/n,$$

and

$$\operatorname{Var} L_n^{\mathrm{II}}(T) = \frac{1}{n^2}\{nR(0) + 2(n-1)R(\tau') + 2(n-2)R(2\tau') + \cdots + 2R[(n-1)\tau']\}, \qquad \tau' = T/(n-1),$$

where $R(0) = \sigma_{c,d}^2$.

For given T and n, the spacing between successive scans is greater for type II than for type I measurements ($\tau' > \tau$). Hence, since R decreases monotonically as its argument increases, we have $R(k\tau') < R(k\tau)$ for all k. It is then readily seen from the two preceding variances formulas that

$$\operatorname{Var} L_n^{\mathrm{I}}(T) > \operatorname{Var} L_n^{\mathrm{II}}(T), \qquad n = 2, 3, \cdots.$$

Furthermore, it is easy to prove that

$$\operatorname{Var} L_\infty(T) = \lim_{n\to\infty} \operatorname{Var} L_n^{\mathrm{I}}(T) = \lim_{n\to\infty} \operatorname{Var} L_n^{\mathrm{II}}(T),$$

where $L_\infty(T)$ is the observed carried load obtained by continuous measurement over $(0, T)$. But whereas, according to statement (i), above, $\operatorname{Var} L_n^{\mathrm{I}}(T)$ decreases monotonically towards $\operatorname{Var} L_\infty(T)$ as n tends to infinity, $\operatorname{Var} L_n^{\mathrm{II}}(T)$ first decreases to a value that lies below $\operatorname{Var} L_\infty(T)$ and then increases monotonically towards $\operatorname{Var} L_\infty(T)$ (see Fig. 15 and Tables I through IV, where all entries for which $\operatorname{Var} L_n^{\mathrm{II}}(T) < \operatorname{Var} L_\infty(T)$ are italicized).

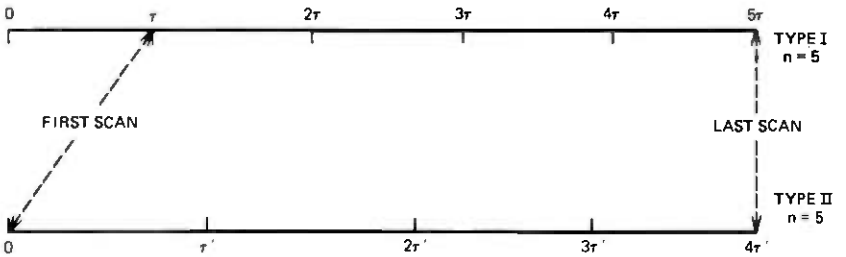


Fig. 14—Type I vs type II measurements.

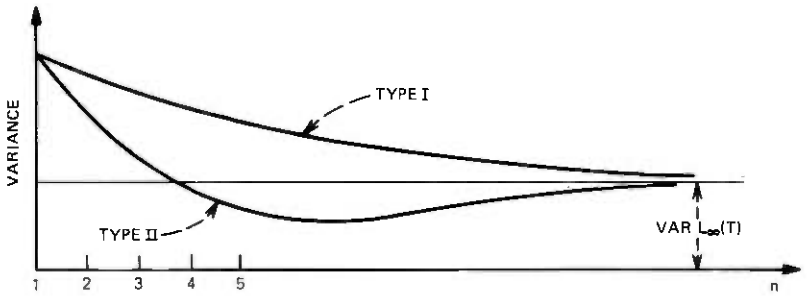


Fig. 15—Variance of type I and type II measurements.

Table I — Variance of the switch-count load for various offered loads

Poisson input; loss system, $c = d = 4$; $T = 15.36$ *h.t.*

No. of Scans	a = 0.25 erlang		a = 0.5 erlang		a = 1 erlang	
	Type I	Type II	Type I	Type II	Type I	Type II
13	0.034332	0.032435	0.067979	0.064228	0.124918	0.118142
17	0.032724	0.031185	0.064770	0.061727	0.118634	0.113110
21	0.031931	0.030650	0.063187	0.060653	0.115531	0.110921
25	0.031484	*0.030391**	0.062294	*0.060132*	0.113778	0.109840
33	0.031025	*0.030184*	0.061377	*0.059713*	0.111976	*0.108942*
49	0.030685	*0.030112*	0.060698	*0.059565*	0.110642	*0.108574*
65	0.030562	*0.030129*	0.060454	*0.059597*	0.110162	*0.108597*
∞	0.030401	0.030401	0.060133	0.060133	0.109529	0.109529
No. of Scans	a = 1.5 erlangs		a = 2 erlangs		a = 2.5 erlangs	
	Type I	Type II	Type I	Type II	Type I	Type II
13	0.156473	0.148318	0.163109	0.155147	0.154579	0.147698
17	0.147517	0.140800	0.152063	0.145398	0.142050	0.136159
21	0.143076	0.137442	0.146553	0.140916	0.135747	0.130709
25	0.140563	0.135736	0.143421	0.138570	0.132147	0.127783
33	0.137974	*0.134244*	0.140186	0.136419	0.128411	0.125000
49	0.136053	*0.133606*	0.137778	*0.135196*	0.125619	*0.123268*
65	0.135362	*0.133431*	0.136909	*0.134950*	0.124609	*0.122822*
∞	0.134449	0.134449	0.135761	0.135761	0.123270	0.123270
No. of Scans	a = 3 erlangs		a = 3.5 erlangs		a = 4 erlangs	
	Type I	Type II	Type I	Type II	Type I	Type II
21	0.119526	0.115301	0.102950	0.099532	0.088157	0.085446
25	0.115590	0.111901	0.098788	0.095772	0.083855	0.081434
33	0.111485	0.108576	0.094420	0.092017	0.079310	0.077355
49	0.108399	0.106382	0.091116	0.089435	0.075847	0.074466
65	0.107279	*0.105741*	0.089910	0.088626	0.074577	0.073517
129	0.106169	*0.105386*	0.088713	*0.088057*	0.073311	*0.072768*
257	0.105885	*0.105491*	0.088406	*0.088076*	0.072986	*0.072712*
∞	0.105790	0.105790	0.088303	0.088303	0.072876	0.072876

*All entries for which Var $L_n^{II}(T)$ < Var ∞ (T) appear in italics.

Table II — Variance of the switch-count load for various offered loads

Poisson input; delay-and-loss system, c = 4, $d = 80$; $T = 15.36$ *h.t.*

No. of Scans	$a = 0.25$ erlang		$a = 0.5$ erlang		$a = 1$ erlang	
	Type I	Type II	Type I	Type II	Type I	Type II
13	0.034364	0.032464	0.068701	0.064901	0.136799	0.129203
17	0.032756	0.031215	0.065492	0.062411	0.130537	0.124388
21	0.031963	0.030680	0.063910	0.061345	0.127445	0.122331
25	0.031515	*0.030421*	0.063017	*0.060829*	0.125698	*0.121338*
33	0.031056	*0.030214*	0.062100	*0.060417*	0.123902	*0.120549*
49	0.030716	*0.030143*	0.061421	*0.060275*	0.122572	*0.120290*
65	0.030594	*0.030160*	0.061177	*0.060310*	0.122093	*0.120367*
∞	0.030432	0.030432	0.060855	0.060855	0.121463	0.121463
No. of Scans	$a = 1.5$ erlangs		$a = 2$ erlangs		$a = 2.5$ erlangs	
	Type I	Type II	Type I	Type II	Type I	Type II
13	0.202358	0.191037	0.261618	0.246937	0.306471	0.289678
17	0.193599	0.184477	0.251394	0.239647	0.296204	0.282872
21	0.189262	0.181690	0.246311	0.236590	0.291074	*0.280080*
25	0.186807	*0.180358*	0.243426	*0.235159*	0.288151	*0.278822*
33	0.184278	*0.179324*	0.240445	*0.234105*	0.285122	*0.277981*
49	0.182401	*0.179031*	0.238225	*0.233917*	0.282856	*0.278012*
65	0.181725	*0.179175*	0.237423	*0.234165*	0.282035	*0.278374*
∞	0.180831	0.180831	0.236361	0.236361	0.280944	0.280944
No. of Scans	$a = 3$ erlangs		$a = 3.5$ erlangs		$a = 3.75$ erlangs	
	Type I	Type II	Type I	Type II	Type I	Type II
13	0.316095	0.300266	0.242082	0.232099	0.149011	0.143657
17	0.307459	0.294979	0.236873	0.229033	0.146173	0.141970
21	0.303119	*0.292863*	0.234234	*0.227809*	0.144733	*0.141286*
25	0.300638	*0.291949*	0.232727	*0.227285*	0.143905	*0.140988*
33	0.298056	*0.291418*	0.231147	*0.226994*	0.143037	*0.140812*
49	0.296116	*0.291620*	0.229954	*0.227144*	0.142380	*0.140874*
65	0.295410	*0.292014*	0.229518	*0.227396*	0.142140	*0.141003*
∞	0.294469	0.294469	0.228934	0.228934	0.141817	0.141817

An immediate consequence of this phenomenon is that $L_\infty(T)$ is *not* a minimum variance estimate of the load carried over intervals of length T.

To shed some light on how this behavior of Var $L_n^{II}(T)$ comes about, we consider the following simple examples.

Let X_1, X_2, and X_3 be identically distributed random variables with variances σ^2, and assume that Cov (X_1, X_2) = Cov $(X_2, X_3) = \sigma^2\theta$ and that Cov $(X_1, X_3) = \sigma^2\theta^2$, where $0 \leqq \theta \leqq 1$. (It is readily shown that this particular choice of the covariances is legitimate.)

Table III — Variance of the switch-count load for various lengths of observation period

Poisson input, $a = 2$ erlangs; loss system, $c = d = 4$

No. of Scans	$T = 0.48$		$T = 0.96$		$T = 1.92$	
	Type I	Type II	Type I	Type II	Type I	Type II
1	1.392290	1.392290	1.392290	1.392290	1.392290	1.392290
2	1.207470	*1.073203*	1.073203	*0.902186*	0.902186	0.758072
3	1.172790	*1.086189*	1.011540	*0.890833*	0.799280	*0.674767*
5	1.154976	*1.106250*	0.979610	*0.908806*	0.744683	*0.664710*
9	1.148037	*1.122470*	0.967124	*0.929260*	0.723056	*0.678623*
17	1.145806	*1.132783*	0.963103	*0.943647*	0.716056	*0.692850*
33	1.145168	*1.138609*	0.961952	*0.952112*	0.714048	*0.702227*
49	1.145042	*1.140660*	0.961725	*0.955143*	0.713652	*0.705726*
65	1.144997	*1.141707*	0.961644	*0.956700*	0.713510	*0.707550*
∞	1.144938	1.144938	0.961537	0.961537	0.713323	0.713323
No. of Scans	$T = 3.84$		$T = 7.68$		$T = 15.36$	
	Type I	Type II	Type I	Type II	Type I	Type II
1	1.392290	1.392290	1.392290	1.392290	1.392290	1.392290
2	0.758072	0.701766	0.701766	0.696191	0.696191	0.696145
3	0.599079	0.521641	0.489334	0.469113	0.465107	0.464138
5	0.508964	*0.446917*	0.390314	0.320960	0.288003	0.282078
9	0.471908	*0.434391*	0.287193	0.264847	0.190330	0.181270
17	0.459714	*0.439503*	0.265409	*0.252440*	0.152063	0.145398
33	0.456197	*0.445779*	0.259008	*0.252168*	0.140186	0.136419
49	0.455501	*0.448493*	0.257734	*0.253109*	0.137778	*0.135195*
65	0.455252	*0.449974*	0.257277	*0.253786*	0.136909	*0.134950*
∞	0.454923	0.454923	0.256674	0.256674	0.135761	0.135761

Under these conditions, we have

$$\operatorname{Var} \frac{X_1 + X_3}{2} = \frac{\sigma^2}{4} (2 + 2\theta^2)$$

and

$$\operatorname{Var} \frac{X_1 + X_2 + X_3}{3} = \frac{\sigma^2}{9} (3 + 4\theta + 2\theta^2).$$

Hence,

$$\operatorname{Var} \frac{X_1 + X_2 + X_3}{3} > \operatorname{Var} \frac{X_1 + X_3}{2},$$

whenever $3 - 8\theta + 5\theta^2 < 0$ or, equivalently, when $3/5 < \theta < 1$. Thus, the preceding inequality holds so long as the correlation, θ, between X_2 and either X_1 or X_3 is sufficiently large to wipe out the accuracy gains that usually accrue by increasing the sample sizes.

The preceding model applies without alteration to single-server loss systems with Poisson input. If, in this instance, we take $X_1 = N_1(0)$,

Table IV — Variance of the switch-count load for various lengths of observation period

Poisson input, $a = 2$ erlangs; delay-and-loss system, $c = 4$, $d = 80$

No. of Scans	$T = 0.48$		$T = 0.96$		$T = 1.92$	
	Type I	Type II	Type I	Type II	Type I	Type II
1	1.652174	1.652174	1.652174	1.652174	1.652174	1.652174
2	1.481714	*1.352457*	1.352457	*1.173156*	1.173156	*0.987875*
3	1.449704	*1.367446*	1.294822	*1.172862*	1.075209	*0.931136*
5	1.433250	*1.387236*	1.264970	*1.194538*	1.023479	*0.984834*
9	1.426839	*1.402759*	1.253287	*1.215878*	1.002992	*0.954575*
17	1.424778	*1.412527*	1.249524	*1.230361*	0.996354	*0.971246*
33	1.424188	*1.418022*	1.248446	*1.238769*	0.994449	*0.981700*
49	1.424071	*1.419963*	1.248234	*1.241764*	0.994073	*0.985534*
65	1.424030	*1.420938*	1.248158	*1.243298*	0.993938	*0.987619*
∞	1.423975	1.423975	1.248057	1.248057	0.993760	0.993760
No. of Scans	$T = 3.84$		$T = 7.68$		$T = 15.36$	
	Type I	Type II	Type I	Type II	Type I	Type II
1	1.652174	1.652174	1.652174	1.652174	1.652174	1.652174
2	0.987875	0.869191	0.869191	0.831006	0.831006	0.826225
3	0.832771	0.713694	0.649697	0.591226	0.568974	0.555158
5	0.746706	*0.663085*	0.514116	0.459861	0.382168	0.360649
9	0.711598	*0.663341*	0.454556	*0.419120*	0.287082	0.268209
17	0.700048	*0.674474*	0.434212	*0.414679*	0.251394	0.239647
33	0.696712	*0.683611*	0.428232	*0.418021*	0.240445	*0.234105*
49	0.696051	*0.687254*	0.427040	*0.420157*	0.238225	*0.233917*
65	0.695814	*0.689194*	0.426612	*0.424232*	0.237423	*0.234165*
∞	0.695502	0.695502	0.426047	0.426047	0.236361	0.236361

$X_2 = N_1(T/2)$, $X_3 = N_1(T)$, then $\theta = \exp[-(1 + a)(T/2)]$ and $\operatorname{Var} L_3^{II}(T) > \operatorname{Var} L_2^{II}(T)$, provided T is small enough. Hence, $\operatorname{Var} L_n^{II}(T)$ does not necessarily decrease as n increases. This simple system is used next to construct examples in which $\operatorname{Var} L_2^{II}(T)$ is smaller than $\operatorname{Var} L_\infty(T)$.

For $c = d = 1$ and Poisson input, it can be shown that

$$R(t) = \sigma^2 e^{-(1+a)t}.$$

Since, however,

$$\begin{aligned}\operatorname{Var} L_\infty(T) &= \lim_{n\to\infty} \operatorname{Var} L_n^{I}(T) = \lim_{n\to\infty} \operatorname{Var} L_n^{II}(T) \\ &= \lim_{n\to\infty} \frac{1}{n^2}\{nR(0) + 2(n-1)R(\tau) + 2(n-2)R(2\tau) \\ &\qquad + \cdots + 2R[(n-1)\tau]\} \\ &= \frac{\sigma^2}{T^2}\int_0^T (T-t)\cdot R(t)dt,\end{aligned}$$

a simple calculation shows that

$$\text{Var } L_\infty(T) = \frac{2\sigma^2}{(1+a)T^2}\left[T - \frac{1}{1+a} + \frac{e^{-(1+a)T}}{1+a}\right].$$

Consequently,

$$\text{Var } L_2^{\text{II}}(T) < \text{Var } L_\infty(T)$$

whenever

$$4(\zeta - 1 + e^{-\zeta}) > \zeta^2(1 + e^{-\zeta}), \qquad \zeta \equiv (1+a)T.$$

The preceding inequality is satisfied so long as

$$0 < \zeta \equiv (1+a)T < 2,$$

and, for any given a (>0), Var $L_\infty(T)$ therefore exceeds Var $L_2^{\text{II}}(T)$ provided $0 < T < 2/(1+a)$. (It is easy to find examples of multi-server systems for which Var $L_2^{\text{II}}(T) <$ Var $L_\infty(T)$. Illustrations of this type can be found in Tables I to IV).

The preceding results admit of the following generalization: For any given n $(\geqq 2)$ and a, the inequality Var $L_n^{\text{II}}(T) <$ Var $L_\infty(T)$ holds provided T is small enough (see Tables III and IV).

We note next that the behavior of Var $L_n^{\text{II}}(T)$ as a function of n is an immediate consequence of the following three properties: For any given a and T,

(*i*) There is an n such that Var $L_n^{\text{II}}(T) <$ Var $L_{n+1}^{\text{II}}(T)$.
(*ii*) If Var $L_n^{\text{II}}(T) <$ Var $L_{n+1}^{\text{II}}(T)$ for some n, then Var $L_{n+k}^{\text{II}}(T)$ $<$ Var $L_{n+k+1}^{\text{II}}(T)$ for all k's.
(*iii*) $\lim_{n\to\infty}$ Var $L_n^{\text{II}}(T) =$ Var $L_\infty(T)$.

Only the last of these three properties, which, of course, also holds for type I measurements, can be regarded as evident. The other two do not admit of a simple explanation since they reflect numerical attributes of the covariance function whose impact is hard to anticipate. Hence, the fact that (*i*) is valid for type II but not for type I measurements appears to be essentially fortuitous. That Var $L_n^{\text{I}}(T)$ displays, as n varies, a simpler and probably more frequently observed behavior than Var $L_n^{\text{II}}(T)$ does not invalidate the preceding remarks since the monotonicity of Var $L_n^{\text{I}}(T)$, as n varies, does not follow in an obvious way from general principles.

When all the parameters but n are prescribed, the ratios and, hence, the inequalities between the variances of either type I and/or type II measurements are independent of σ^2. These inequalities can then be expressed in terms of the correlation function $\rho(\cdot) \equiv R(\cdot)/\sigma^2$, and it

turns out that, for any n and all $k \geqq 0$,

$$\text{Var } L^{II}_{n+k}(T) < \text{Var } L_\infty(T),$$

provided $\rho(\tau)$ is large enough. Some properties of Var $L^{II}_n(T)$, regarded as a function of n, are expressed next in terms of ρ.

Except for the offered load, a, which is allowed to vary, let all parameters be prescribed and let $\rho(t)$ be the correlation between two observations made t apart. As sketched in Fig. 16, $\rho(t)$ has the following properties valid for all t's (see also Figs. 17 and 18 for closely related results):

(*i*) In loss systems, $\rho(t)$ decreases monotonically as a increases and tends to 0 as a tends to infinity.

(*ii*) In delay systems, $\rho(t)$ increases monotonically as a increases and tends to 1 as a tends to the critical load c.

Let $\mathbf{n}$ be the smallest n for which $L^{II}_n(T) < L_\infty(T)$. As stated above, this inequality is satisfied whenever $\rho(\tau)$ is large enough. Hence, by means of (*i*) and (*ii*), we may conclude that

(1) In loss systems, $\mathbf{n}$ cannot increase as a increases (see Table I).

(2) In delay systems, $\mathbf{n}$ cannot decrease as a increases (see Table II).

Furthermore, since for all systems considered here, $\rho(t)$ decreases monotonically as t increases we have

(3) In delay-and-loss systems with arbitrary defection rates, $\mathbf{n}$ cannot decrease as the length of the observation period increases (all other parameters being kept fixed). This is illustrated in Tables III and IV.

As the number of waiting positions increases, so does $\rho(t)$ and we therefore also have

(4) In delay-and-loss systems with fixed a, c, and j, $\mathbf{n}$ cannot increase as d increases. (This assertion can be partially checked by comparing the data of Table I with those of Table II.)

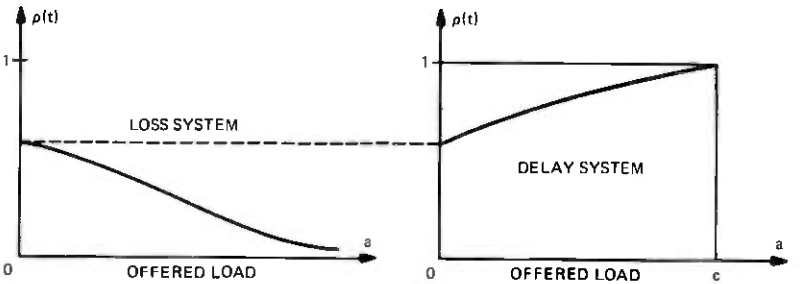


Fig. 16—Correlation between two observations made t apart vs the offered load.

[Note that for the values assigned to the parameters, the delay-and-loss system of Table II actually behaves like a (pure) delay system and, practically speaking, this table pertains to such a system. In the title of this table, however, the term "delay-and-loss" and the statement that $d = 80$ are retained so as not to obscure the conditions under which the computations were made.]

As can be seen from Tables I to IV, the difference between Var $L_n^{II}(T)$ and Var $L_\infty(T)$ is quite small and is certainly negligible in practical situations. The phenomenon studied in this section is of interest, however, because it contradicts a well-rooted feeling that a greater amount of information cannot entail a loss of accuracy; more important, however, is its implication that the common notion that the variance of the switch-count load can be regarded as the sum of the variances of the "source load" and of the switch-count error[2] is not unconditionally valid.

VIII. THE AUTOCOVARIANCE FUNCTION FOR SEQUENCES OF LOAD MEASUREMENTS

For the purpose of forecasting and/or controlling traffic volumes on trunk groups and switching devices, carried-load measurements are frequently performed over successive (nonoverlapping) intervals. The statistical analysis of such sequences of observations depends essentially on a knowledge of the autocovariance function (defined below). We shall therefore show how it can be computed by means of the variance formula derived in Ref. 1 and then describe some properties of the corresponding autocorrelation function.

From here on, we assume that all measurements are of type I and designate by $L_n(t, T)$ the switch-count load over $(t, t + T]$ [$L_n(0, T) \equiv L_n(T)$]. Then the autocovariance function, $\mathcal{R}_{nk}(T)$, for a sequence of observations performed over the nonoverlapping intervals $(0, T]$, $(T, 2T]$, $(2T, 3T]$, $\cdots$, is defined by

$$\mathcal{R}_{nk}(T) \equiv \text{Cov}\,\{L_n(0, T), L_n[kT, (k + 1)T]\}, \qquad k = 0, 1, \cdots.$$

For given n and T, this covariance is easily calculated for any value of k as soon as Var $L_{mn}(mT)$ is known for $m = 1, 2, \ldots, k + 1$. Indeed, for $k = 1$ we have (by the formula for the variance of sums of correlated variables)

$$4 \text{ Var } L_{2n}(2T) = 2 \text{ Var } L_n(T) + 2 \text{ Cov } [L_n(0, T), L_n(T, 2T)],$$

so that

$$\mathcal{R}_{n1}(T) \equiv \text{Cov } [L_n(0, T), L_n(T, 2T)] = 2 \text{ Var } L_{2n}(2T) - \text{Var } L_n(T).$$

Similarly, for $k = 2$, we have

$$9 \text{ Var } L_{3n}(3T) = 3 \text{ Var } L_n(T) + 4 \text{ Cov } [L_n(0, T), L_n(T, 2T)] + 2 \text{ Cov } [L_n(0, T), L_n(2T, 3T)],$$

so that

$$\mathcal{R}_{n2}(T) = \tfrac{9}{2} \operatorname{Var} L_{3n}(3T) - \tfrac{3}{2} \operatorname{Var} L_n(T) - 2\mathcal{R}_{n1}(T).$$

Hence, by a simple inductive process, we obtain the following expression:

$$\mathcal{R}_{nk}(T) = \frac{k^2}{2} \operatorname{Var} L_{kn}(kT) - \frac{k}{2} L_n(T) - (k-2)\mathcal{R}_{n2}(T) - (k-3)\mathcal{R}_{n3}(T) - \cdots - 2\mathcal{R}_{n,k-1}(T).$$

By means of these formulas the (auto)correlation function, $\Gamma(k) \equiv \mathcal{R}_{nk}(T)/\sigma^2$, was computed for some loss, delay, and delay-and-loss systems. These results, which are presented in Figs. 17 and 18, suggest the following properties. In loss systems, the coefficient of correlation $\Gamma(k)$ for fixed $k(\geq 1)$ increases monotonically as the load decreases and satisfies the following inequalities:

$$0 \leq \Gamma(k) \leq \Gamma_0(k) < 1,$$

where

$$\Gamma_0(k) \equiv \lim_{a \to 0} \Gamma(k).$$

Furthermore,

$$\Gamma_\infty(k) \equiv \lim_{a \to \infty} \Gamma(k) = 0.$$

In (pure) delay systems, the behavior of $\Gamma(k)$ is quite different. In this case, $\Gamma(k)$, for fixed $k \geq 1$, increases monotonically as a increases and

$$\Gamma_0(k) \leq \Gamma(k) \leq \Gamma_\infty(k) = 1,$$

where $\Gamma_0(k)$ and $\Gamma_\infty(k)$ are defined as above. (Note that $\Gamma_0(k)$ is independent of the number of waiting positions.)

The dependence of $\Gamma(k)$ on a is somewhat more complicated in delay-and-loss systems. In this last instance, $\Gamma(k)$, for fixed $k(\geq 1)$, first increases as a increases, reaches a maximum $\mathbf{\Gamma}(k)$, and then decreases monotonically as a further increases. $\Gamma(k)$ now satisfies the following inequalities:

$$\Gamma_\infty(k) = 0 \leq \Gamma(k) \leq \mathbf{\Gamma}(k),$$

where

$$\mathbf{\Gamma}(k) \equiv \max_a \Gamma(x) > \Gamma_0(k).$$

It is easy to show that $\mathcal{R}_{nk} = \sigma^2\Gamma(k)$ is asymptotically exponential for large values of k. But, as can be inferred from the behavior of $\Gamma(k)$ (see Figs. 17 and 18), deviation from exponentiality is rather pronounced for small k's. Hence, the assumption sometimes made in

practice that $\mathcal{R}_{nk}$, regarded as a function of k, is exponential requires further investigation. In this connection, it seems that, at the very least, any fitting covariance function, $\mathcal{R}^*_{nk}$, should not be subjected to the requirement that $\mathcal{R}^*_{n0}(T) = \text{Var } L_n(T)$.

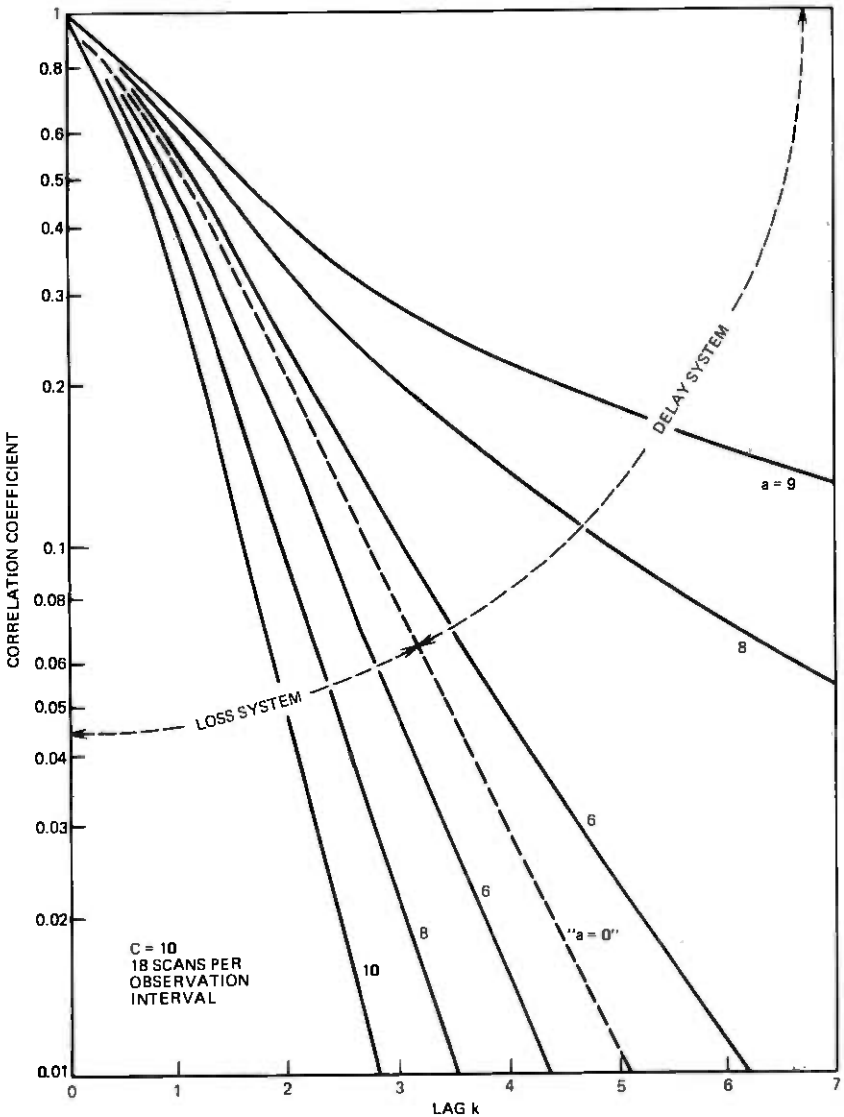


Fig. 17—Correlation between load measurements made over nonoverlapping intervals of unit length (Poisson input).

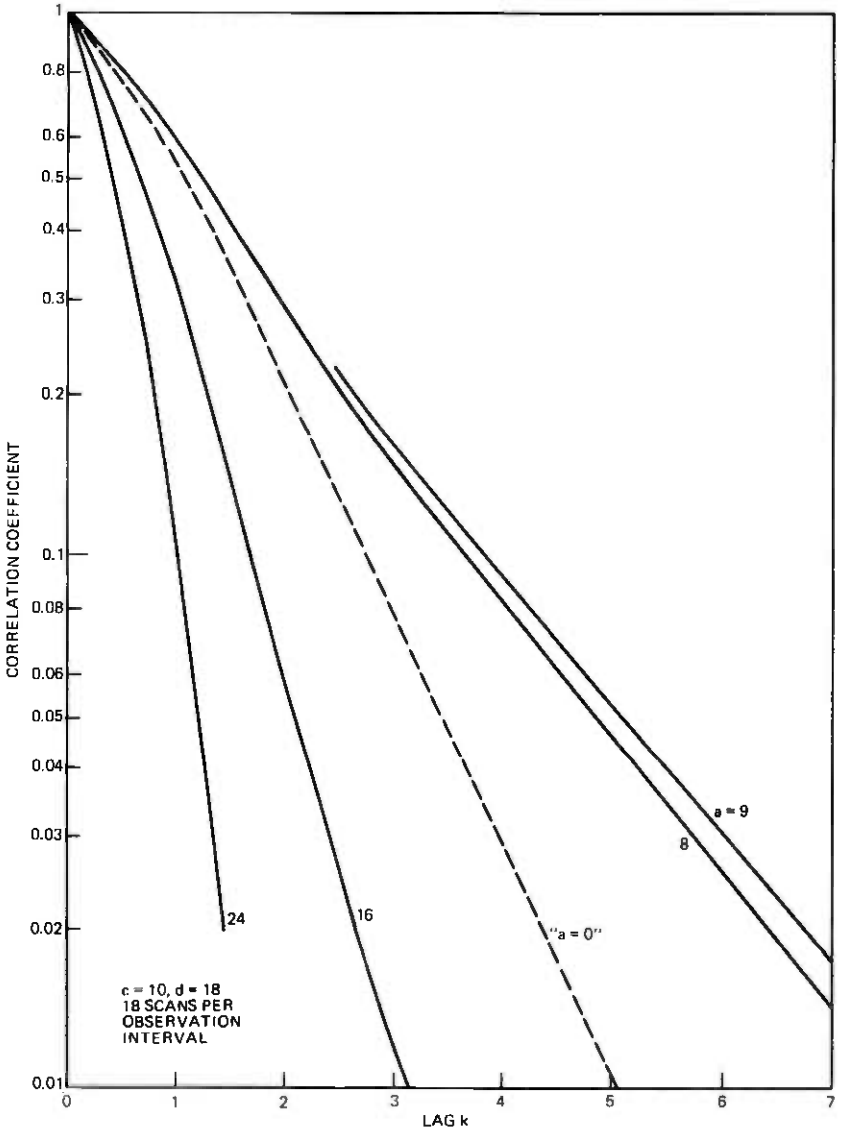


Fig. 18—Correlation between load measurements made over nonoverlapping intervals of unit length (Poisson input, delay-and-loss system).

IX. CONCLUSIONS

In this paper, we have presented numerical examples that shed considerable light on the behavior of the variance of the switch-count load. In particular, they show that, for relatively low offered loads, the

variance in question increases as more waiting is allowed by the system while the converse holds at sufficiently high offered loads. But when the same variance is studied in terms of the carried load, a much simpler picture emerges: for a fixed value of this parameter the variance of the switch-count load always increases when either the number of waiting positions and/or the number of sources increase. And a decrease in the defection rates has a similar effect on the variance of the switch-count load as an increase in the number of waiting positions. As we have shown, these properties can be explained by a combination of simple probability and traffic considerations.

But the results of this paper are not exclusively qualitative. On the contrary, the charts illustrate that waiting, in general, affects the magnitude of the switch-count load variance to a degree that cannot be ignored in practice.

The reasonings by which we have explained the qualitative behavior of the switch-count load variance can be expected to hold for more general inputs and service-time distributions. The basis for this statement is that the "hole-filling" argument of Section VI remains meaningful since, to maintain the carried load at a given level, more calls must be offered when fewer of them are allowed to wait, and this fact, of course, is not affected by the shapes of the interarrival and holding-time distributions.

An unexpected result of our investigation is that the continuous load measurements are not minimum variance estimates of the carried load and that (discrete) scanning does not necessarily entail a loss of accuracy.

Finally, as we have shown, the variance formulas derived in Ref. 1 make it possible to compute exactly the autocovariance function for sequences of switch-count load measurements which are thus brought within the purview of time-series analysis. This, in turn, should help evaluate the performance of traffic-control methods based on load measurements.

REFERENCES

1. A. Descloux, "Variance of Load Measurements in Markovian Service Systems," B.S.T.J., *54*, No. 7 (September 1975), pp. 1277–1300.
2. W. S. Hayward, Jr., "The Reliability of Telephone Traffic Load Measurements by Switch Counts," B.S.T.J., *31*, No. 2 (March 1952), pp. 357–377.
3. V. E. Beneš, "The Covariance Function of a Simple Trunk Group, with Applications to Traffic Measurements," B.S.T.J., *40*, No. 1 (January 1961), pp. 117–148.
4. W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1, 3rd ed. New York: John Wiley, 1967.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 1, January 1976
Printed in U.S.A.

An Algorithmic Procedure for Designing Hybrid FIR/IIR Digital Filters

By M. R. CAMPBELL, R. E. CROCHIERE, and L. R. RABINER

(Manuscript received July 3, 1975)

An algorithmic procedure for designing hybrid FIR/IIR digital filters is proposed and evaluated in this paper. The design is implemented as a two-stage optimization in which a Hooke and Jeeves optimization procedure is used to optimize the IIR component of the filter and the McClellan et al. optimization procedure is used to optimize the FIR component of the filter. To evaluate this method, a set of eight low-pass filters were designed in which a single complex conjugate pole pair was used as the IIR component. The resulting designs were compared and contrasted with standard IIR low-pass filters and the optimal FIR, linear-phase, low-pass filter in terms of multiplications, storage, and group delay properties.

I. INTRODUCTION

A wide variety of digital filter-design methods have been proposed and studied in the past several years.[1–9] Generally, these design methods can be classified as analytical or algorithmic, depending on the form of solution of the approximation problem which is used. Additionally, the resulting designs are classified as finite impulse response (FIR) or infinite impulse response (IIR), depending on the filter properties. FIR filters have the property that they can be easily designed to have exactly linear phase. Furthermore, linear-phase FIR filters can be designed to approximate an arbitrary magnitude response to within given tolerances by using a sufficiently high order. IIR filters cannot achieve a linear-phase response exactly, but are capable of approximating sharp cutoff filters with considerably lower order filters than are required for the FIR designs which meet identical magnitude specifications.[10] Thus, for many practical filter applications, there is a trade-off between the exact linear-phase response obtainable using an FIR filter and the reduced filter order obtainable using an IIR filter.

In this paper, a hybrid design is proposed which bridges the gap, somewhat, between the FIR and IIR filters. The hybrid filter is a particular class of IIR filter where the degree of the numerator of its system

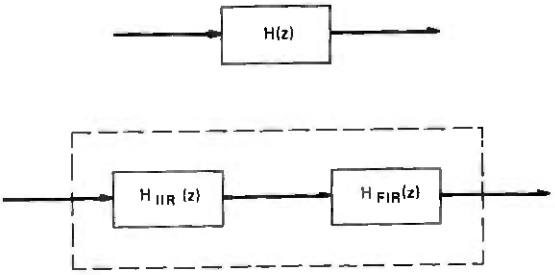


Fig. 1—Hybrid FIR/IIR filter.

function is significantly higher than the degree of the denominator. Thus, it is reasonable to consider the hybrid design as a cascade of an Mth-order IIR filter with Nth-order FIR filter, as shown in Fig. 1, where $N \gg M$. In the designs discussed here, M is set to 2* and N is in the range of 27 to 51. The idea behind this hybrid design is to incorporate the good features of both IIR and FIR filters, yielding a resulting filter which can meet arbitrary design specifications with a significantly smaller filter order than is required by the FIR filter alone, and with a smaller group delay variation (i.e., more phase linearity) than is normally obtained by the IIR filter alone.

Based on the above discussion, the hybrid filter of Fig. 1 can be put in the form

$$H(z) = H_{\text{IIR}}(z)H_{\text{FIR}}(z), \tag{1}$$

where, by assuming a second-order denominator as mentioned above,

$$H_{\text{IIR}}(z) = \frac{1}{(1 - \rho e^{j\theta}z^{-1})(1 - \rho e^{-j\theta}z^{-1})} \tag{2}$$

and

$$H_{\text{FIR}}(z) = \sum_{n=0}^{N-1} h_F(n)z^{-n}, \tag{3}$$

where ρ is the radius of the pole in the z plane and θ is the pole angle. (Although we have used a second-order IIR filter in eq. (2) and in the examples of this paper, the design approach is general and can be applied to other orders of IIR sections as well.) The overall filter is decomposed into a cascade of sections for two reasons. The first is to emphasize the fact that the hybrid filter is most readily realized as a cascade of an IIR filter with an FIR filter. In this way, the direct form structure can be used to realize the FIR section so as to fully utilize the symmetry in the impulse response due to the linear phase condition.

* Cases where $M = 1$ were also studied but have not led to any useful results.

The second reason that this cascade of sections is used is because it separates the design problem into two distinct parts; one which is fairly simple and one which can be solved using a well-known FIR approximation method.[11]

To obtain the best second-order denominator hybrid filter approximation to the desired specifications, the parameters ρ and θ of the IIR filter are systematically varied, and an optimal FIR filter is obtained for each set of parameters using the design program of McClellan et al.[11] An alternative method of obtaining such a hybrid design would be to use a different type of algorithmic procedure (e.g., Ref. 8 or 9) and simultaneously vary both the numerator coefficients and the denominator coefficients. Unfortunately, since the numerator order is so much higher than the denominator order, these optimization methods are not always successful.

In the remainder of this paper we discuss the hybrid filter-design algorithm and present some typical results on low-pass filter designs obtained with this method. Then we compare and contrast the resulting designs with some standard IIR low-pass filters including Butterworth, Chebyshev, and elliptic filters, and with the optimal FIR filter. The bases of comparison are the number of multiplications per sample, the group delay variation, and the storage requirements. We conclude with a discussion of the properties of the hybrid filters.

II. DESIGN ALGORITHM

Since the algorithm is being applied to the design of low-pass filters (although it is equally useful for any arbitrary magnitude function), the desired frequency response of the system $H_I(e^{j\omega})$ is of the form

$$|H_I(e^{j\omega})| = \begin{cases} 1 & 0 \leqq \omega \leqq 2\pi F_p \\ 0 & 2\pi F_s \leqq \omega \leqq \pi, \end{cases} \tag{4}$$

where F_p and F_s are the passband and stopband edges, respectively. The tolerances are δ_p in the passband and δ_s in the stopband. Thus, $|H_I(e^{j\omega})|$, the magnitude response of the composite system of Fig. 1, satisfies the inequalities

$$\begin{aligned} 1 - \delta_p \leqq |H_I(e^{j\omega})| &\leqq 1 + \delta_p \qquad 0 \leqq \omega \leqq 2\pi F_p \\ 0 \leqq |H_I(e^{j\omega})| &\leqq \delta_s \qquad 2\pi F_s \leqq \omega \leqq \pi. \end{aligned} \tag{5}$$

The method in which the individual magnitude responses of the IIR filters and the FIR filters are chosen is shown in Fig. 2. Initial values are chosen for the IIR coefficients ρ and θ, based on the location of the highest Q-pole pair of an elliptic filter which meets the tolerance scheme of eq. (5). The reason for this choice will be discussed later.

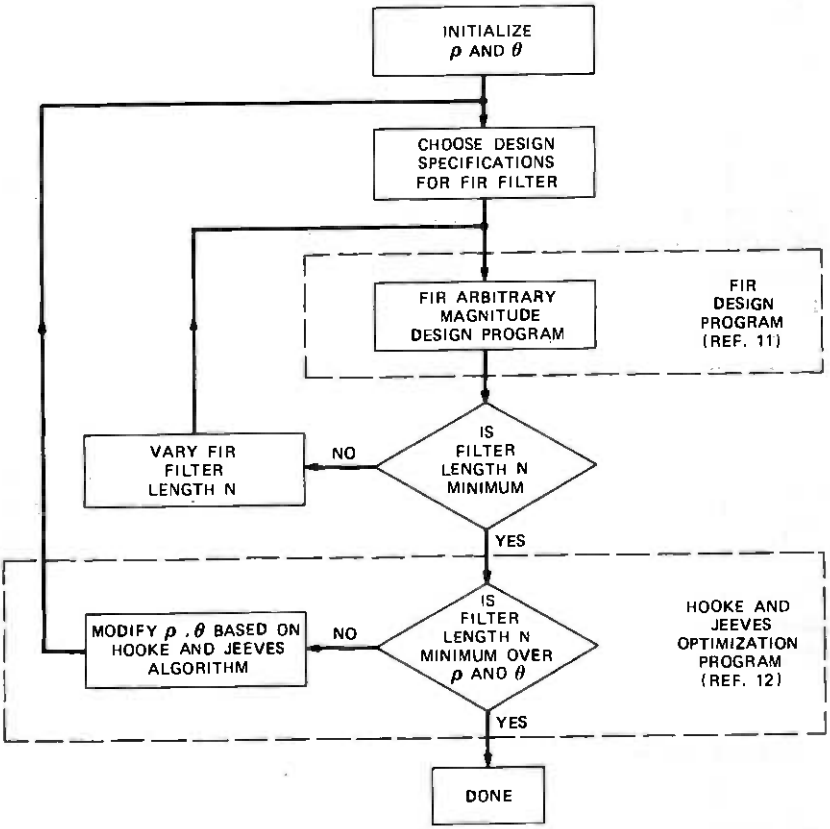


Fig. 2—Block diagram of the design algorithm.

Based on the initial values of ρ and θ, a set of design specifications for the FIR filter is obtained from eqs. (1) and (5) of the form

$$\begin{aligned} \frac{1-\delta_p}{|H_{\text{IIR}}(e^{j\omega})|} \leqq |H_{\text{FIR}}(e^{j\omega})| \leqq \frac{1+\delta_p}{|H_{\text{IIR}}(e^{j\omega})|} \qquad & 0 \leqq \omega \leqq 2\pi F_p \\ 0 \leqq |H_{\text{FIR}}(e^{j\omega})| \leqq \frac{\delta_s}{|H_{\text{IIR}}(e^{j\omega})|} \qquad & 2\pi F_s \leqq \omega \leqq \pi. \end{aligned} \tag{6}$$

Equation (6) specifies the appropriate parameters (the band edges, desired values, and weighting) for the McClellan et al. arbitrary magnitude FIR design program.[11] The final parameter required is N, the filter length. An initial guess of the value of N is made, and the design program yields the optimal FIR filter for the given specifications. A control loop is used to find the smallest value of N which can be used to meet the given input specifications.

Once the minimum N (for the given values of ρ and θ) is obtained, an outer loop optimization program is used to vary ρ and θ to obtain the minimum value of N as a function of ρ and θ. The algorithm used to find the optimum values of ρ and θ is the well-known Hooke and Jeeves optimization.[12] Generally, the Hooke and Jeeves algorithm is capable of optimizing a continuous function of (several) continuous variables. However, for this problem the function $[N(\rho, \theta)]$ is not continuous, but instead is discrete (integer) valued. To handle the problems created by this discrete-valued function, fairly careful control over the variation of ρ and θ had to be maintained. To illustrate this point, Fig. 3 shows a typical contour of the variation of N as a function of ρ.

The minimum of this function occurs at the point labeled A. However, because $N(\rho, \theta)$ is discrete, it is possible for the optimization algorithm to prematurely terminate on flat regions such as those

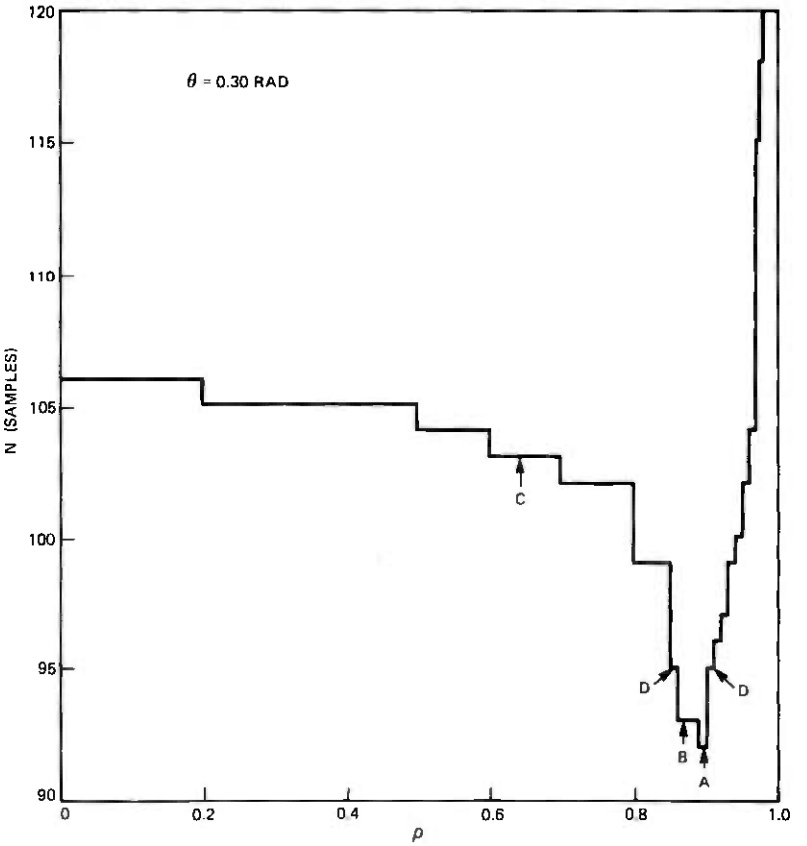


Fig. 3 —Typical contour of N as a function of ρ.

labeled B and C. To minimize these difficulties, a careful choice is required of the initial step sizes, $\Delta\rho$ and $\Delta\theta$, used by the Hooke and Jeeves algorithm to vary ρ and θ. For example, if $\Delta\rho$ is too large, the algorithm may not find the "valley" of the function. Similarly, if $\Delta\rho$ is too small, the algorithm may "hang up" on a flat region as it reduces its step size. Successful choices of $\Delta\rho$ and $\Delta\theta$ appear to be approximately one-half the distance across the "valley" of the function (e.g., one-half the distance between the two points labeled D in Fig. 3). For the examples in the next section, the choices $\Delta\rho = 0.005$, $\Delta\theta = 0.02$ (radians) yielded good results. Figure 4 shows another example of the variation of N as a function of ρ and θ. Here we can observe that there are two reciprocal minima of N as a function of ρ. If the optimum ρ_0 outside the unit circle is found, the corresponding optimum value of ρ inside the unit circle can be found by taking the reciprocal of ρ_0.

A factor that strongly affects the efficiency of the algorithm is the initial choice of ρ and θ. If accurate initial estimates of ρ and θ are used, they can result in a significant speed-up of the design method. As stated earlier, good estimates were found to be the location of the highest Q-pole pair of an elliptic filter which meets the same tolerance scheme. Other good estimates appear to be the highest Q-pole pair locations of the Chebyshev 2 and Chebyshev 1 designs which meet the same tolerance scheme. For greater assurance that an optimum has been found, several of these starting points may be tried to see if the algorithm converges to the same value of ρ and θ.

Another factor that strongly affects the efficiency of the algorithm is the strategy of the control loop for varying N (see Fig. 2). Generally, it was found that a good initial guess of N after changing ρ and θ is the previous value of N. An efficient strategy for increasing or decreasing N to find its new minimum then appeared to be a "tree" search (or log search). That is, N is incremented or decremented by an amount ΔN, depending on whether the previous choice of N yields a design which meets the specifications in eq. (6). On the next trial, ΔN is reduced by one-half, and the process is repeated until $\Delta N = 1$. The search can be terminated sooner if at any stage the FIR design is sufficiently close to the tolerance requirements (e.g., within 1 percent) given in (6). Other variations on this strategy are also possible.

III. EXPERIMENTAL RESULTS

Using the hybrid filter-design algorithm of Fig. 2, several low-pass filters were designed, ranging from narrow-band to wide-band designs. Table I gives the filter specifications (band edges and ripple tolerances) for eight low-pass filters, along with the resulting IIR pole position, the length of the FIR section, and the length of an optimal linear-phase

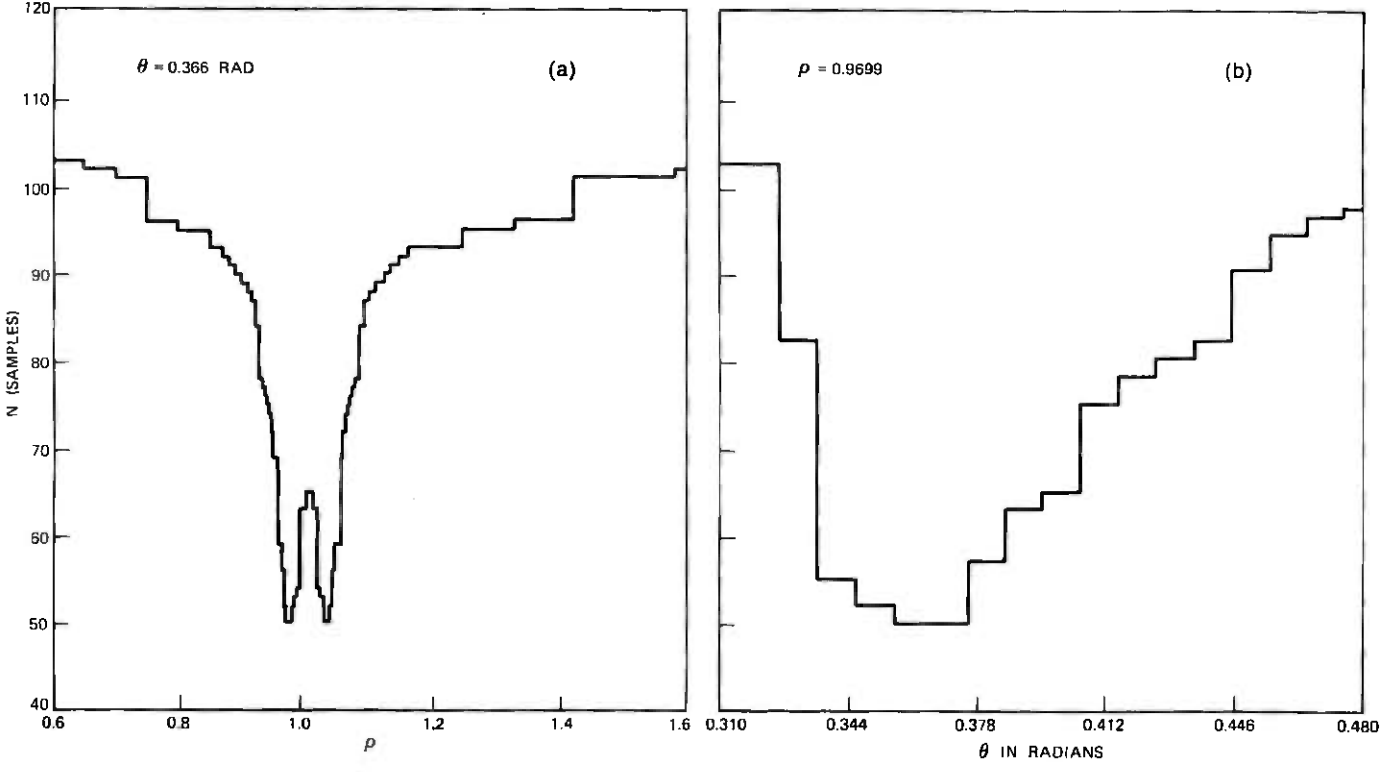


Fig. 4—Example of the contour of (a) N as a function of ρ, and (b) N as a function of θ.

Table I — Low-pass filter design

Filter Example	Filter Specifications				Optimum Hybrid FIR/IIR Design			Equivalent Optimum FIR Design	Highest Q-Pole of Equivalent Elliptic Design	
	F_p	F_s	δ_p	δ_s	ρ	θ(rad)	N	N_{FIR}	ρ_e	θ_e(rad)
1	0.05	0.075	0.01	0.001	0.9636	0.3363	39	109	0.9776	0.3238
2	0.1	0.125	0.01	0.001	0.9761	0.6817	50	107	0.9807	0.6367
3	0.15	0.175	0.01	0.001	0.9787	0.9979	51	106	0.9768	0.9530
4	0.2	0.225	0.01	0.001	0.9859	1.3133	51	105	0.9746	1.2684
5	0.05	0.125	0.005	0.0005	0.9634	0.3601	31	44	0.9534	0.3401
6	0.1	0.175	0.005	0.0005	0.9390	0.6805	27	43	0.9453	0.6605
7	0.15	0.225	0.005	0.0005	0.9582	0.9935	31	43	0.9482	0.9736
8	0.2	0.275	0.005	0.0005	0.9516	1.3127	29	42	0.9416	1.2927

FIR filter which also meets the design specifications. The first four examples are filters with a fairly narrow transition band (0.025), whereas the last four examples are filters with a wider transition band (0.075). As seen in Table I, the reduction in length of the FIR filter in the hybrid, from the optimal FIR linear-phase filter, is on the order of 1.5:1 to 2:1—i.e., there is a fairly significant reduction in FIR filter length.

Figures 5 to 7 show examples of the magnitude responses of filters nos. 1, 5, and 2, respectively. Figure 5a shows the log magnitude response of the IIR filter, Fig. 5b shows the log magnitude response of the FIR section, and Figs. 5c and 5d show the linear and log magnitude responses respectively, of the composite filter. It can be seen from these figures that the response of the IIR filter makes the requirements for the FIR section significantly easier to obtain. For example, the required tolerance near $\omega = \pi$ is on the order of -15 dB in order that the composite response be more than 60 dB down at this frequency.

Figures 6 and 7 show examples of some undesirable characteristics of the frequency response which can occur in the composite filter. In these cases, there is a ripple of the magnitude response in the transition band of the filter.* Since there is no real constraint on the composite filter response in the transition band, this behavior is not unreasonable. The question becomes one of whether or not a filter with a ripple in the transition band is acceptable. In general, the answer to this question is that it depends on the intended application. In some cases, this behavior is acceptable, in others it is not.

Table II gives the approximate height of the transition band ripple for each example of Table I. In example 5 (see Fig. 6d), the transition band ripple is down by 17 dB and may be perfectly acceptable. In example 2 (see Fig. 7d), the ripple peak amplitude is about 1.06—i.e.,

* Note that the pure FIR filter does not have such a peak in the don't care region.

Table II — Transition band ripple for Table I examples

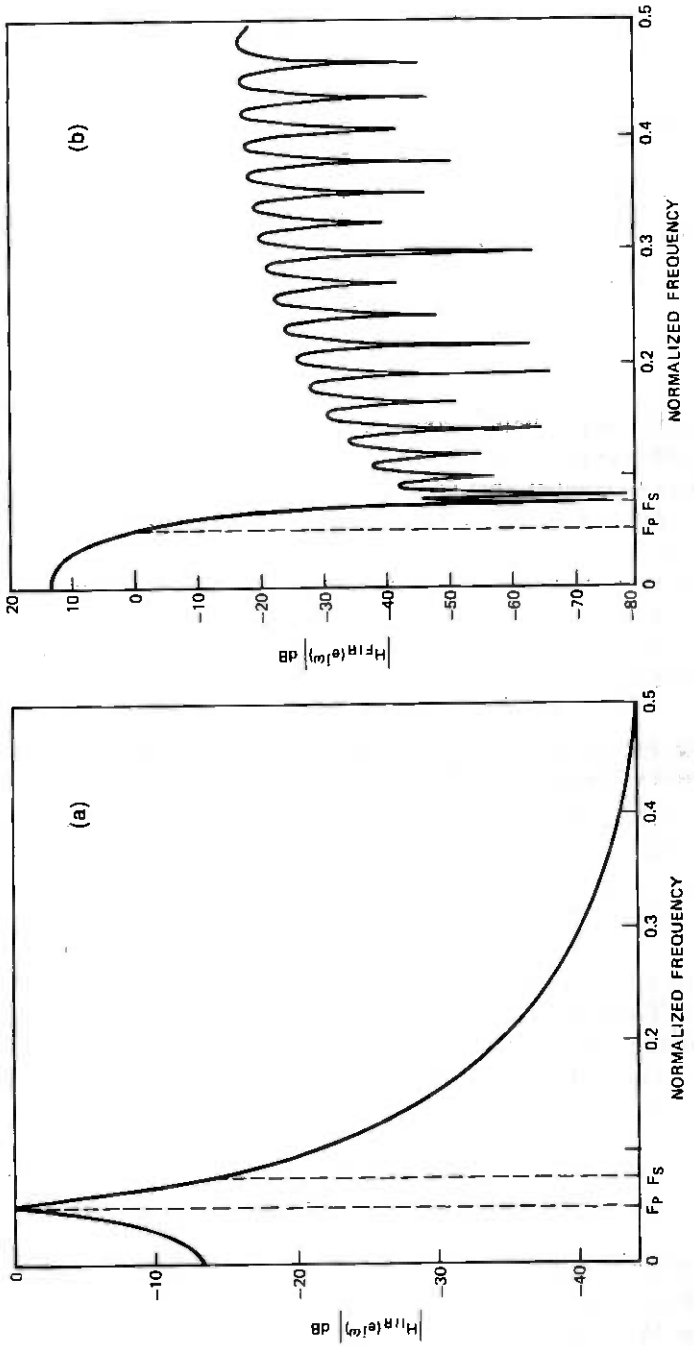
Example	Height of Transition Band Ripple	
	Magnitude	dB
1	—	
2	1.06	0.5
3	1.16	1.29
4	1.52	3.64
5	0.14	−17
6	0.022	−33
7	0.11	−19
8	0.063	−24

it exceeds the passband maximum response by about 6 percent. Generally, such large ripples render the filter useless in many applications. One way to combat this nonmonotonicity of the magnitude response in the transition band would be to constrain the pole of the IIR section to lie within the passband of the filter. Another possibility would be to constrain the response of the FIR section beyond the passband edge to guarantee monotonic behavior of the magnitude response. Either of these alternatives would lead to higher values of the FIR filter duration, thereby somewhat negating the gains of using the IIR filter.

IV. COMPARISON OF HYBRID FILTER DESIGNS TO OTHER CONVENTIONAL DESIGNS

The hybrid FIR/IIR design represents a trade-off between an FIR design and an IIR design. The usefulness of such an approach depends strongly on how it compares with other conventional designs. To illustrate where in this scope of design alternatives a hybrid approach might be competitive, we have compared the hybrid examples designed in Section III to examples of five other conventional low-pass filter designs. These designs included the optimum finite impulse response (FIR) design, the Butterworth (BUT) design, the Chebyshev type 1 and 2 (CHEB1, CHEB2) designs, and the Cauer elliptic (ELLIPT) design.

In Fig. 8, the various designs are compared on the basis of the number of multiplications required for their implementation. In the FIR designs and the FIR parts of the hybrid designs, the symmetry of the impulse response was exploited. In the IIR designs, it was assumed that a conventional implementation of cascaded second-order sections was used and that coefficients of value 1 and 2 are not implemented with multiplies. Figure 8a shows the results for examples 1 to 4 in Table I and Fig. 8b shows the results for examples 5 to 8.


(a)
(b)
$|H_{IIR}(e^{j\omega})|$ dB
$|H_{FIR}(e^{j\omega})|$ dB
NORMALIZED FREQUENCY
F_P F_S
0
0.2
0.3
0.4
0.5
20
10
0
−10
−20
−30
−40
−50
−60
−70
−80

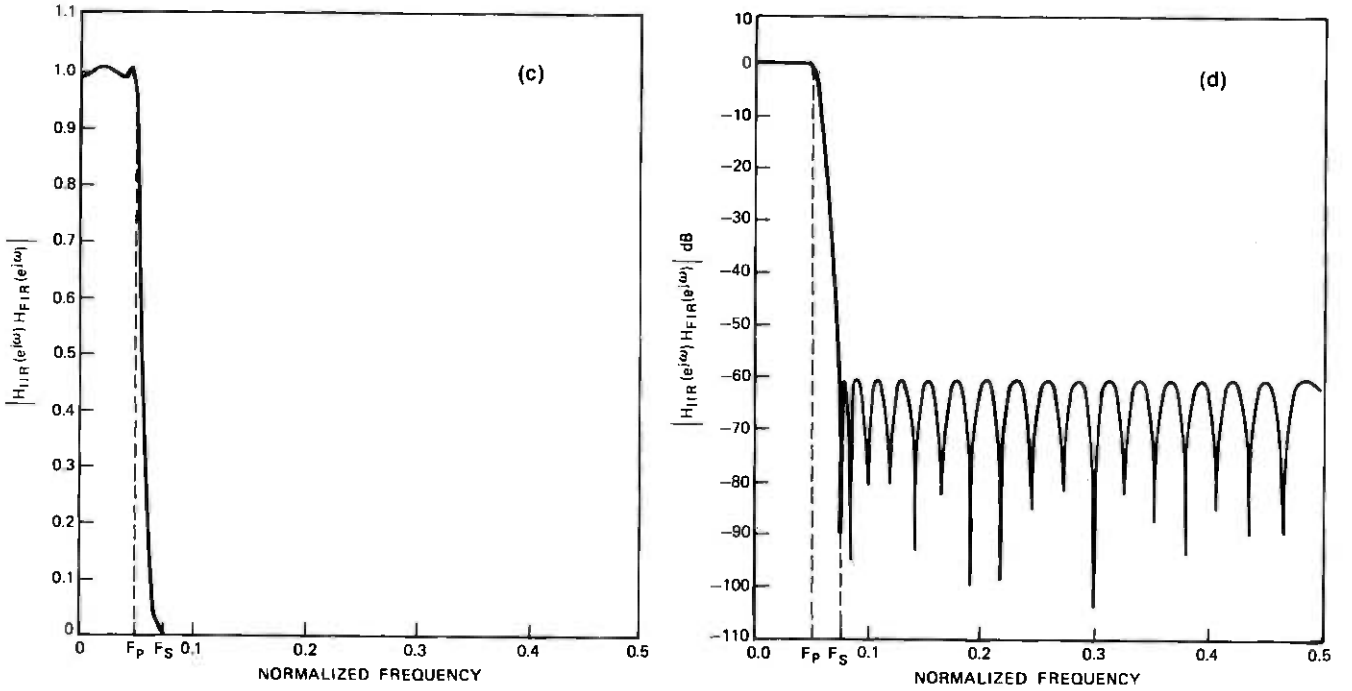
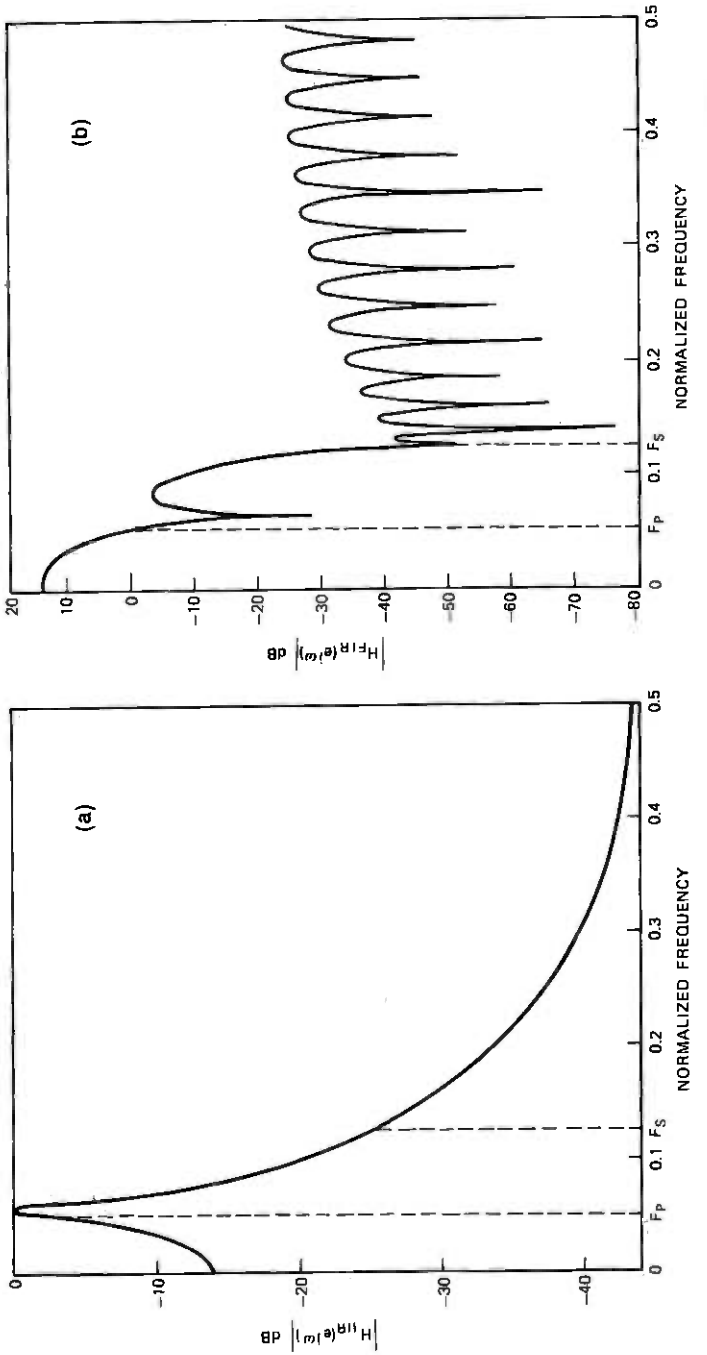


Fig. 5—Frequency responses of example 1. (a) Log magnitude plot of IIR section. (b) Log magnitude plot of FIR section. (c) Linear plot of composite magnitude response of FIR/IIR design. (d) Log magnitude plot of FIR/IIR design.


(a)
(b)
$|H_{IIR}(e^{j\omega})|$ dB
$|H_{FIR}(e^{j\omega})|$ dB
NORMALIZED FREQUENCY
NORMALIZED FREQUENCY
0
0.1 F_S
F_P
0.2
0.3
0.4
0.5
20
10
0
−10
−20
−30
−40
−50
−60
−70
−80

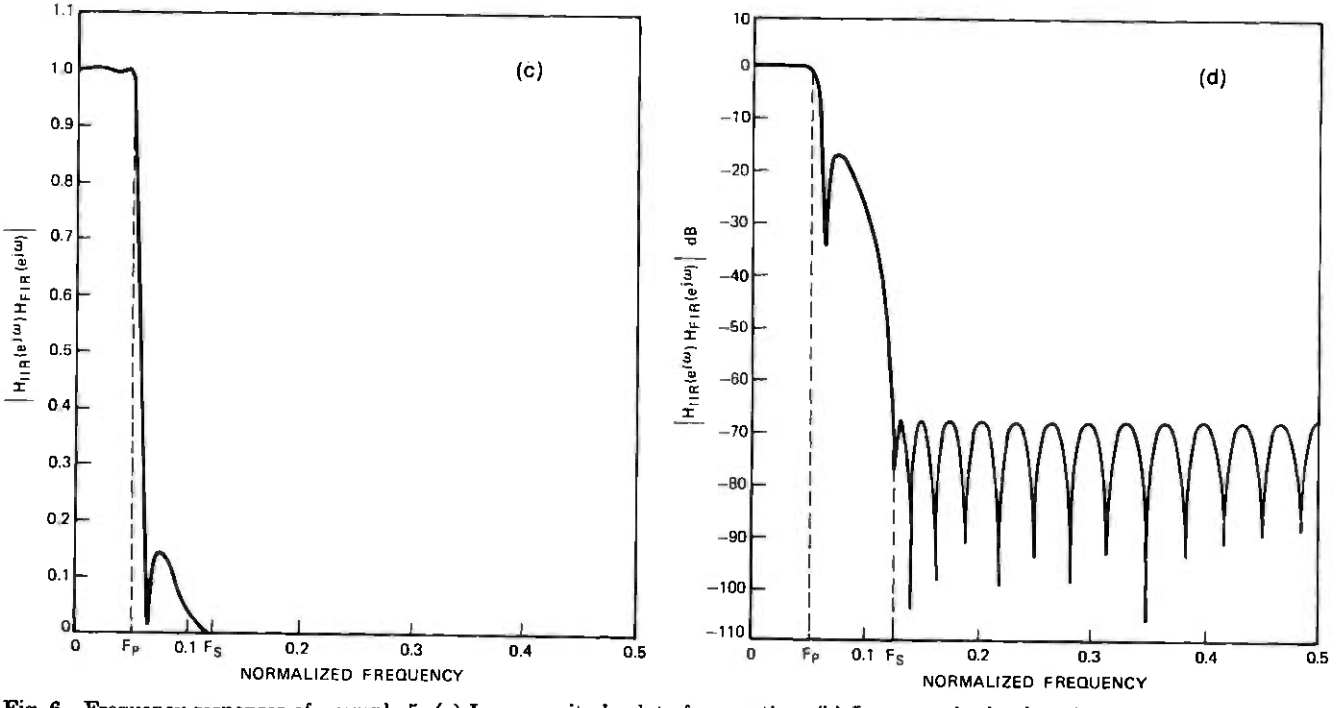
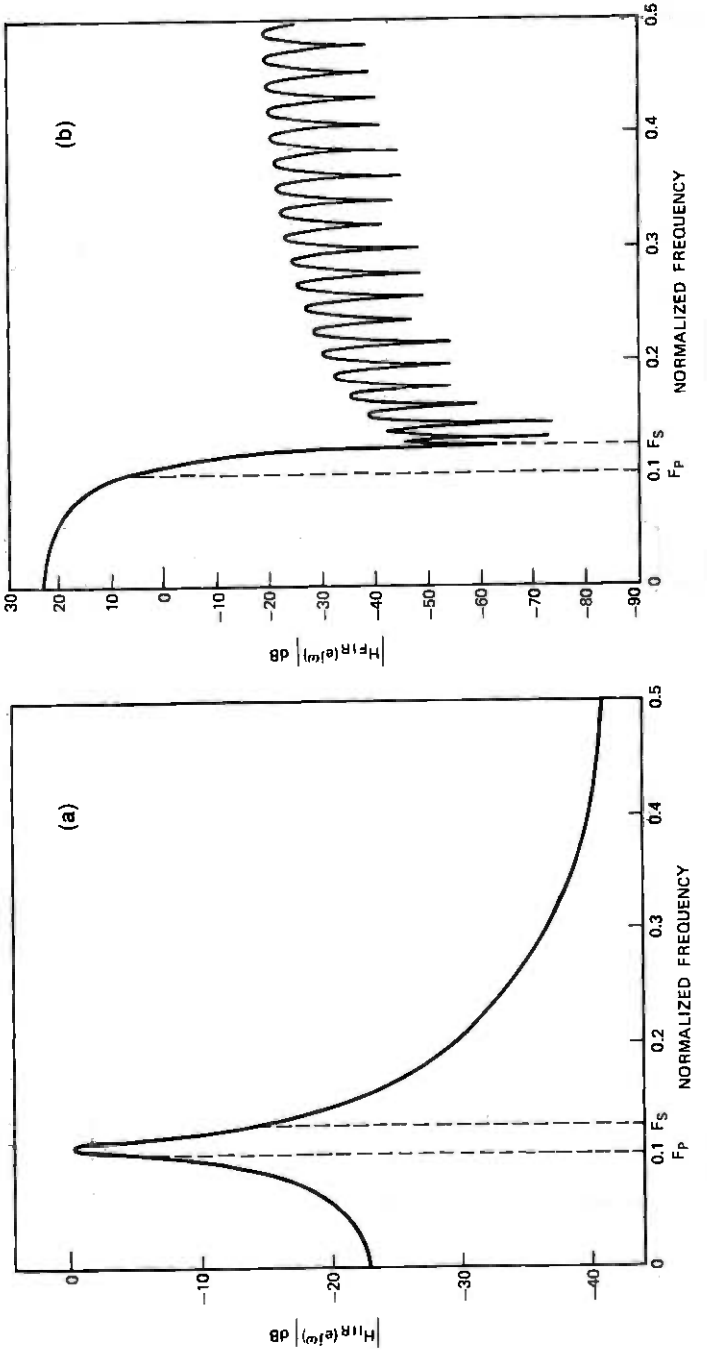


Fig. 6—Frequency responses of example 5. (a) Log magnitude plot of IIR section. (b) Log magnitude plot of FIR section. (c) Linear plot of FIR/IIR cascade. (d) Log magnitude plot of FIR/IIR cascade.


(a)
(b)
$|H_{IIR}(e^{j\omega})|$ dB
$|H_{FIR}(e^{j\omega})|$ dB
NORMALIZED FREQUENCY
F_P
F_S
0
0.1
0.2
0.3
0.4
0.5
-10
-20
-30
-40
-50
-60
-70
-80
-90
10
20
30

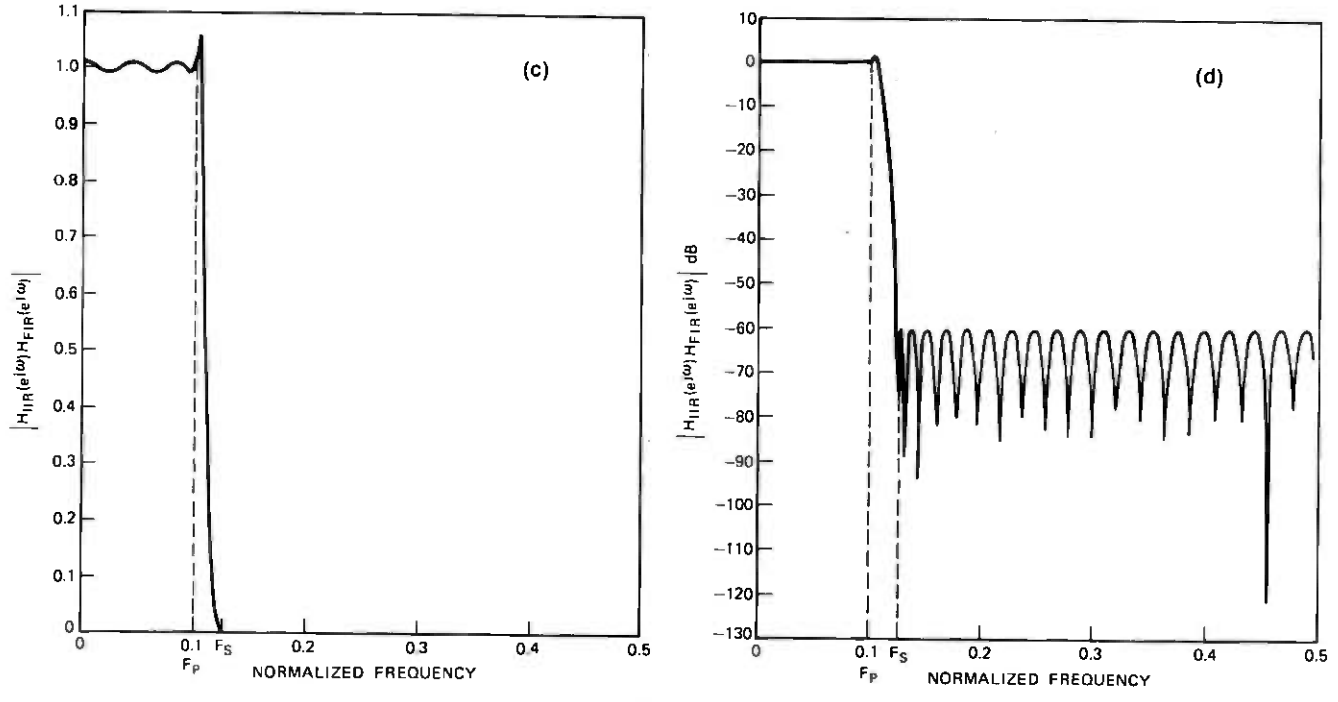


Fig. 7—Frequency responses for example 2. (a) Log magnitude plot of IIR section. (b) Log magnitude plot of FIR section. (c) Linear plot of FIR/IIR cascade. (d) Log plot of FIR/IIR cascade.

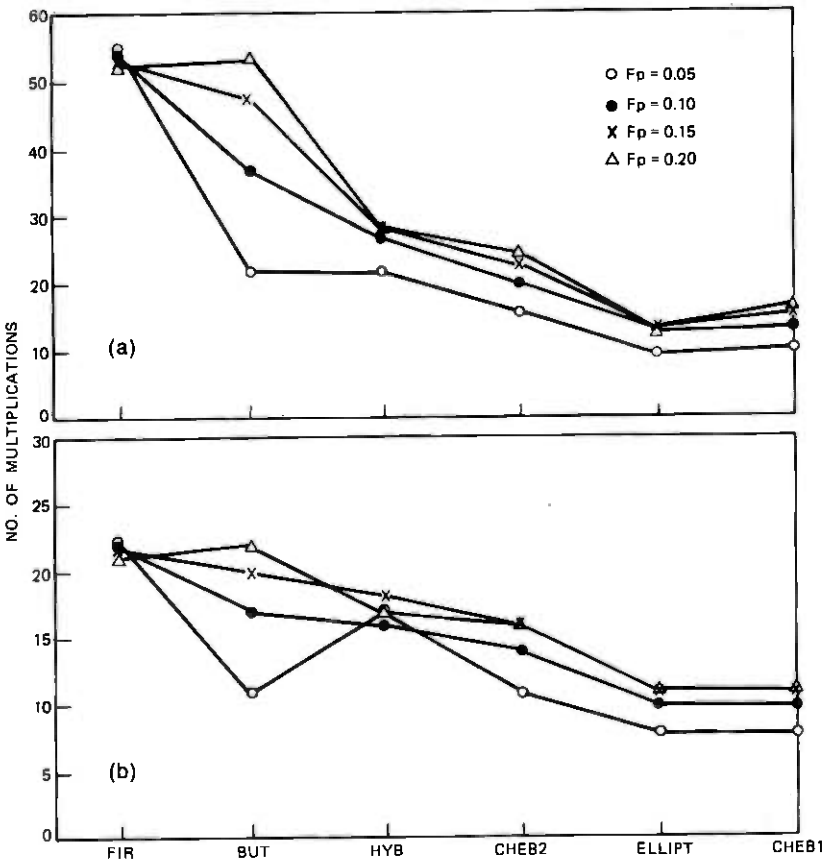


Fig. 8—Comparison of filter designs on the basis of the number of required multiplications.

The results of the comparison in Fig. 8 indicate that the hybrid design is often more efficient than the Butterworth design and is slightly less efficient than a Chebyshev 2 design for the same magnitude specifications.

A second criterion used for comparison was the group delay. Figure 9 shows plots of the group delay for each filter design for example 8. The FIR filter had the largest fixed delay at zero frequency, but its group delay was exactly flat (i.e., zero dispersion) across all frequencies. The hybrid design had a relatively flat component across most of the passband with most of its dispersion near the edge of the passband and within the transition band. The next least dispersive design appeared to be the Chebyshev 2 design, followed by the Butterworth, elliptic, and Chebyshev 1 designs.

Another measure of dispersion used to compare the designs was the maximum minus the minimum group delay across the passband. These results are plotted in Fig. 10. Obviously, the FIR design had exactly zero dispersion in these examples. The next best contenders appeared to be the hybrid and the Chebyshev 2 designs, followed by the elliptic, Butterworth, and Chebyshev 1 designs. In three examples, the order of the Butterworth filter was larger than what could be accommodated by the available design program, so these results were not included. It was essentially the large required order of the Butterworth filters that prevented them from having favorable group delay characteristics.

A final comparison was made on the number of data storage locations (i.e., state variables) necessary for the implementation of the

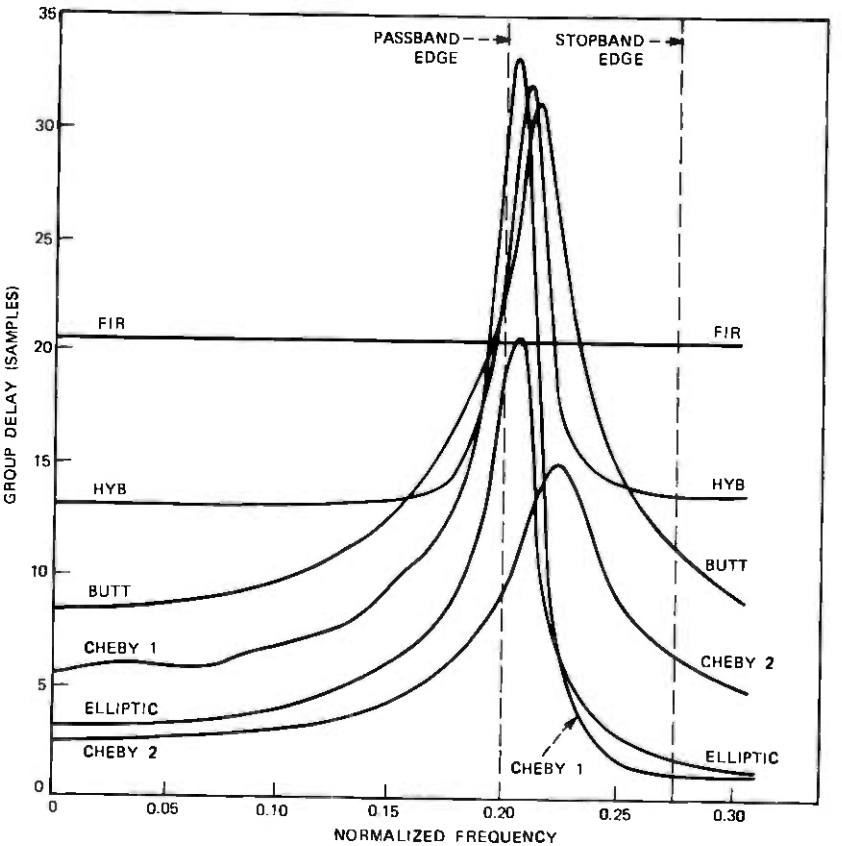


Fig. 9—Comparison of group delays of various filter designs for example 8.

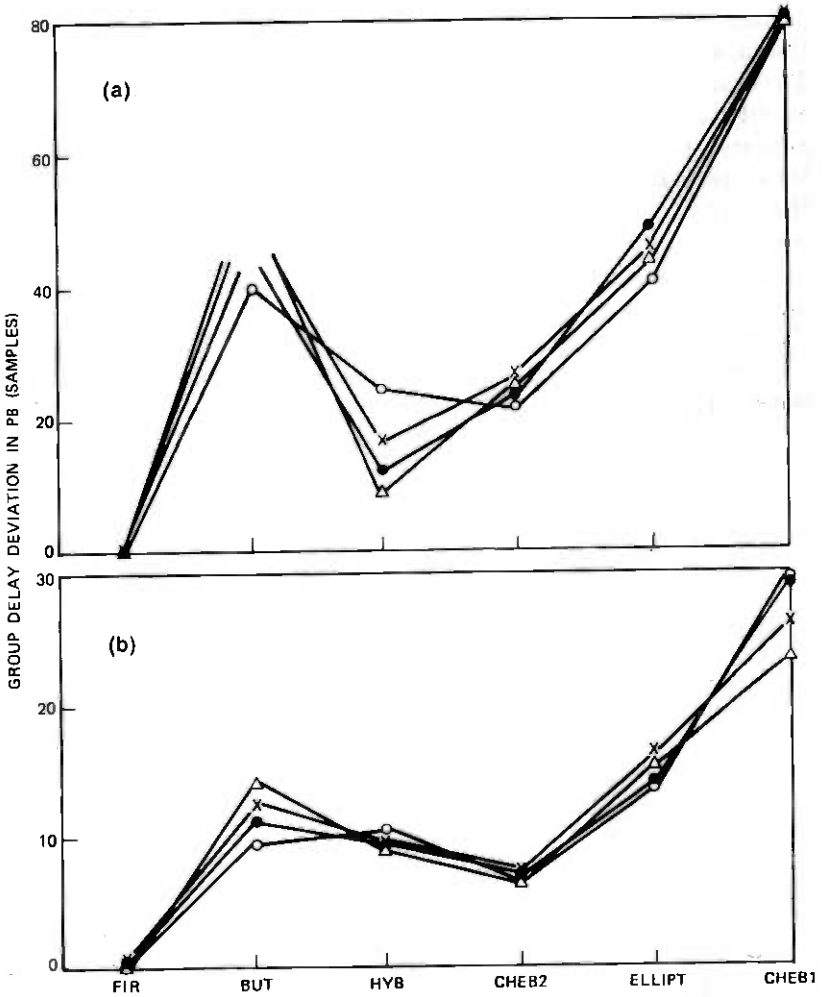


Fig. 10—Comparison of filter designs on the basis of group delay deviation (max.–min.) in the passband.

filters. These results are given in Fig. 11. In this respect, the FIR designs and the hybrid designs did not compare as favorably to the recursive designs.

By cross-comparing the above characteristics, it is possible to obtain a good insight into the various trade-offs involved in each filter design. Obviously, no single approach stands out as being superior over all other designs in all respects. The choice of a given design must be weighted according to the needs of a specific application.

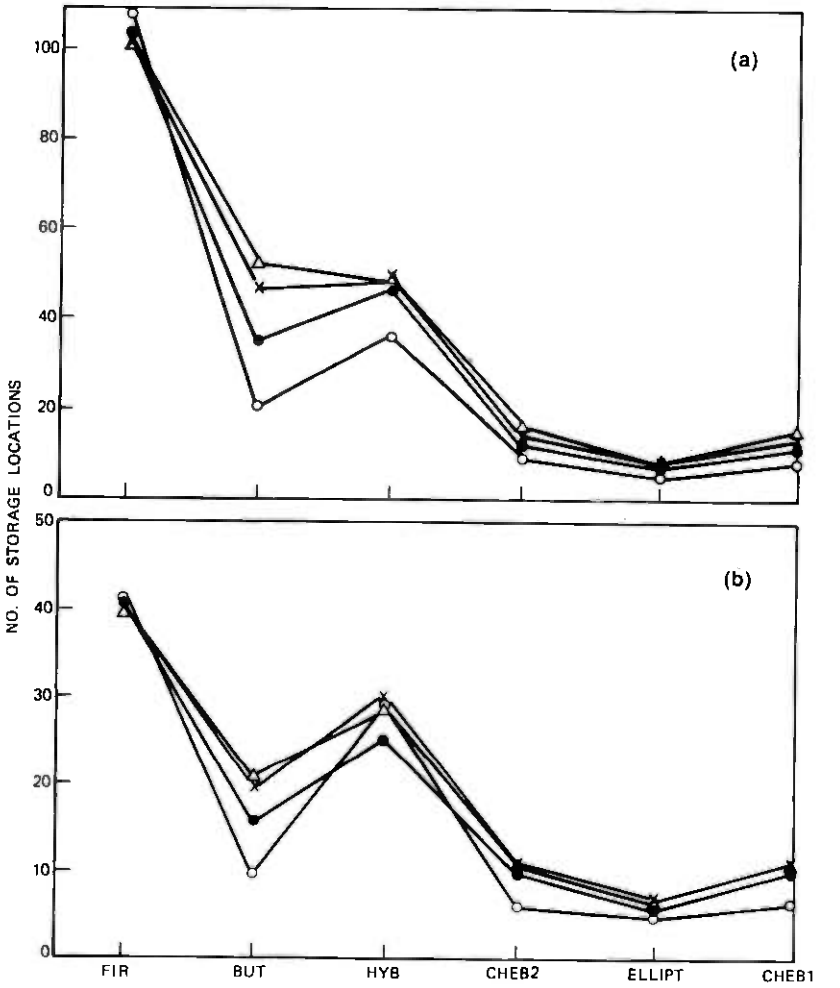


Fig. 11—Comparison of filter designs on the basis of required number of storage locations for state variables.

V. CONCLUSIONS

An algorithmic procedure has been proposed for designing hybrid FIR/IIR digital filters. The procedure is based on the use of a well-known FIR design algorithm for designing the FIR part of the filter, and it is coupled with a well-known optimization algorithm which is used to design the IIR part of the filter. A set of low-pass filters designed in this way were found to have characteristics in between those of optimal FIR designs and conventional IIR designs.

Several properties of the design algorithm have been discussed in detail as well as methods of choosing initial starting points and techniques for speeding up the algorithm. Because of the nature of the FIR/IIR design, it was found that ripples could occur in the transition band. In some examples, these ripples were found to be objectionable, i.e., their amplitude exceeded that of the passband gain although the constraints imposed by the algorithm for the passband and stopband regions were completely met. Thus, an issue which needs further investigation is that of incorporating additional constraints in the algorithm to control the amplitude of the transition band ripples which can occur in such a hybrid filter design.

REFERENCES

1. J. F. Kaiser and F. F. Kuo, *Systems Analysis by Digital Computer*, New York: John Wiley, 1966.
2. L. R. Rabiner, B. Gold, and C. A. McGonegal, "An Approach to the Approximation Problem for Nonrecursive Digital Filters," IEEE Trans. Audio and Electroacoust., *AU-18*, No. 2 (June 1970), pp. 83–106.
3. L. R. Rabiner, J. H. McClellan, and T. W. Parks, "FIR Digital Filter Design Techniques Using Weighted Chebyshev Approximation," Proc. IEEE, *63*, No. 4 (April 1975), pp. 595–610.
4. C. M. Rader and B Gold, "Digital Filter Design Techniques in the Frequency Domain," Proc. IEEE, *55*, No. 2 (February 1967), pp. 149–171.
5. C. S. Burrus and T. W. Parks, "Time Domain Design of Recursive Digital Filters," IEEE Trans. Audio and Electroacoust., *AU-18*, No. 2 (June 1970), pp. 137–141.
6. K. Steiglitz, "Computer-Aided Design of Recursive Digital Filters," IEEE Trans. Audio and Electroacoust., *AU-18*, No. 2 (June 1970), pp. 123–129.
7. A. G. Deczky, "Synthesis of Recursive Digital Filters Using the Minimum P-Error Criterion," IEEE Trans. Audio and Electroacoust., *AU-20*, No. 4 (October 1972), pp. 257–263.
8. L. R. Rabiner, N. Y. Graham, and H. D. Helms, "Linear Programming Design of IIR Digital Filters With Arbitrary Magnitude Function," IEEE Trans. Acoust., Speech, and Signal Processing, *ASSP-22*, No. 2 (April 1974), pp. 117–123.
9. D. E. Dudgeon, "Recursive Filter Design Using Differential Corrections," IEEE Trans. Acoust., Speech, and Signal Processing, *ASSP-22*, No. 6 (December 1974), pp. 443–448.
10. L. R. Rabiner, J. F. Kaiser, O. Herrmann, and M. T. Dolan, "Some Comparisons Between FIR and IIR Digital Filters," B.S.T.J., *53*, No. 2 (February 1974), pp. 305–331.
11. J. H. McClellan, T. W. Parks, and L. R. Rabiner, "A Computer Program for Designing Optimum FIR Linear Phase Digital Filters," IEEE Trans. Audio and Electroacoust., *AU-21*, No. 6 (December 1973), pp. 506–526.
12. J. L. Kuester and J. H. Mize, *Optimization Techniques With Fortran*, New York: McGraw-Hill, 1973, pp. 309–319.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 1, January 1976
Printed in U.S.A.

Filter Response of Nonuniform Almost-Periodic Structures

By H. KOGELNIK

(Manuscript received September 8, 1975)

The filter response of nonuniform, almost-periodic structures, such as corrugated optical waveguides, is investigated theoretically. The filter process, leading to reflection of a band of frequencies near the Bragg frequency, is treated as a contradirectional coupled-wave interaction and shown to obey a Riccati differential equation. The nonuniformity of the structure is represented by a tapering in the coupling strength (e.g., the depth of the corrugation) and by a chirp in the period of the structure. For small reflectivities, the filter response is a Fourier transform of the taper function. For large reflectivities, the Riccati equation was evaluated numerically and plots are given for the response of filters with linear and quadratic tapers and with linear and quadratic chirps.

I. INTRODUCTION

Recent papers[1,2] have reported the fabrication of wavelength-selective filters in the form of corrugated, thin-film, optical waveguides and have described the evaluation of the filter characteristics by means of tunable dye lasers. These papers include a numerical comparison of experimental and theoretical results on the nature of the filter response caused by a gradual tapering in the corrugation or grating strength (which we call "tapers"), or by a gradual variation of the effective grating period (called "chirping"). The present paper describes in some detail the theory by which these numerical results were obtained; it offers a discussion of the general characteristics of nonuniform grating filters, such as their scaling properties and the symmetry of the filter response; and it provides a more complete collection of plots representing numerical results for linear and quadratic tapers in the coupling strengths as well as linear and quadratic chirps in the effective grating period. Periodic waveguides serve as band-rejection filters with a center frequency determined by the Bragg condition

$$K = 2\beta(\omega), \tag{1}$$

where $K = 2\pi/\Lambda$ is the grating constant or spatial frequency, Λ the grating period, and $\beta(\omega)$ the propagation constant of the waveguide mode of interest. The filter mechanism results from the backward scattering of the light by the periodic structure which we describe as a contradirectional coupled-wave interaction. For its analysis, we use the coupled-wave formalism[3,4] which allows adequately for the depletion of the incident light.

Problems similar to ours have been encountered in other coupled-wave processes. Codirectional coupling in nonuniform structures has been investigated in connection with microwave directional couplers,[5–7] holography,[8] and tapered optical directional couplers.[9] Our present problem of contradirectional coupling in nonuniform periodic structures has been considered by Uchida[10] who analyzed structures with exponential tapers and by Hill and coworkers[11–13] who use a numerical method based on the iteration of a pair of coupled-mode integral equations. Another possible method is to approximate the nonuniform structure by a set of short uniform elements, each represented by a known matrix, and use matrix multiplication to calculate the properties of the compound overall structure. Kermisch[8] has applied such a technique to codirectional coupling in hologram gratings. As described below, we have used yet another method where we reduce the pair of coupled-wave equations to a single Riccati differential equation, which can then be solved by tested numerical techniques such as the Runge-Kutta method.

II. RICCATI EQUATION FOR THE REFLECTION COEFFICIENT

Following coupled-wave theory, we assume a periodic, single-mode waveguide with an electromagnetic field which can be represented by two contradirectional coupled waves in the form

$$E(z) = R(z)e^{-j\beta_0 z} + S(z)e^{j\beta_0 z}, \tag{2}$$

where R and S are the complex amplitudes of the forward- and backward-running mode. These amplitudes are linked by the standard coupled-wave equations[3,4]

$$R' + j\delta R = -j\kappa S e^{-j\phi} \tag{3}$$

and

$$S' - j\delta S = j\kappa R e^{j\phi}, \tag{4}$$

where we have allowed for a phase-shift ϕ of the periodicity relative to the origin ($z = 0$). The prime indicates differentiation with respect to z. The coupling coefficient κ is related to the amplitude of the wave-

guide perturbation (e.g., the height of the film corrugation), and formulas for specific guides with specific perturbations are given in the literature.[3,4,14] The measure $\delta(\omega)$ indicates the frequency deviation from the Bragg condition and is defined as[4]

$$\delta = \beta - K/2 = \beta - \beta_0 \approx \Delta\omega/v_g, \tag{5}$$

where $\beta_0 = K/2$ is the propagation constant at the center (Bragg) frequency, $\Delta\omega$ is the radian frequency deviation from that frequency, and v_g is the group velocity of the guide.

We consider, now, structures in which both the coupling coefficient $\kappa(z)$ and the grating phase $\phi(z)$ are slowly varying functions of z, indicating the nonuniformity in the grating parameters. We assume that the structure has a length L and extends from $z = -L/2$ to $z = L/2$. The boundary conditions for our scattering problem are then

$$R(-L/2) = 1, \qquad S(L/2) = 0. \tag{6}$$

The key to the reduction of the coupled-wave equations to a single differential equation is the definition of a local reflection coefficient $\rho(z)$,

$$\rho = \frac{S}{R} e^{-j\phi}. \tag{7}$$

The z-derivative of this is

$$\rho' = \left(\frac{S'}{R} - \frac{SR'}{R^2} - j\phi' \frac{S}{R} \right) e^{-j\phi}. \tag{8}$$

Combining the above expressions with (3) and (4), we obtain a Riccati differential equation for ρ which is of the form

$$\rho' = j(2\delta - \phi')\rho + j\kappa(1 + \rho^2). \tag{9}$$

The boundary condition for this equation follows from (6) as

$$\rho(L/2) = 0. \tag{10}$$

Our quantity of interest is the reflection coefficient $\rho(-L/2)$ of the entire structure, or the corresponding reflectivity $\rho\rho^*$. What we call filter response is the dependence of this reflectivity on the frequency (or wavelength) of the light. Results for the response curves of specific taper functions are provided in later sections, but first we study some of their general properties.

III. FOURIER-TRANSFORM RESPONSE AT LOW REFLECTIVITIES

When the reflectivities are low, we expect a simple relationship between the taper function $\kappa(z)e^{j\phi(z)}$ and the response function $\rho(\delta)$.

This relationship follows from the assumption of an undepleted incident wave or the use of a first Born approximation.[15,16] It is easily derived from the Riccati equation (9) by substituting a new variable σ defined by

$$\rho = \sigma e^{j(2\delta z - \phi)}, \tag{11}$$

which obeys the differential equation

$$\sigma' = j\kappa[e^{-j(2\delta z - \phi)} + \sigma^2 e^{j(2\delta z - \phi)}]. \tag{12}$$

When reflectivities are low, the term proportional to σ^2 can be neglected, and we can integrate this equation with the result

$$\sigma(-L/2) = -j \int_{-L/2}^{L/2} dz \cdot \kappa(z) \cdot e^{-j(2\delta z - \phi)}. \tag{13}$$

This shows that, apart from phase factors, the response function $\rho(\delta)$ is the Fourier transform of the taper function $\kappa(z) \cdot e^{j\phi(z)}$.

IV. SYMMETRY OF FILTER RESPONSE

Let us now consider the conditions for which the filter response is symmetric, i.e., for which

$$\rho\rho^*(\delta) = \rho\rho^*(-\delta). \tag{14}$$

To simplify our discussion, we assume that the filter structure can be described by a real and positive function $\kappa(z)$ and an arbitrary phase function $\phi(z)$. Then the Fourier transform relation (13) predicts that the filter response is symmetric (a) if $\phi(z) = 0$, or (b) if both κ and ϕ are symmetric, i.e., if $\kappa(z) = \kappa(-z)$ and $\phi(z) = \phi(-z)$. Relation (13) is valid for small reflectivities only, but we get the same answer from the Riccati equation (9), which is valid for all reflectivities. To demonstrate this, we take the complex conjugate of (9) and write the result in the form

$$(-\rho^*)' = j(-2\delta + \phi')(-\rho^*) + j\kappa[1 + (-\rho^*)^2]. \tag{15}$$

If we replace δ with $-\delta$ and ϕ' with $-\phi'$, this differential equation for $(-\rho^*)$ is the same as the differential equation (9) for ρ, and (as the boundary conditions are also the same) it predicts the same reflectivity $\rho\rho^*$. While replacing δ with $-\delta$ simply means that we are looking at the other side of the center frequency, replacing ϕ' with $-\phi'$ means that we have replaced the original filter structure with another one. For a structure with symmetrical $\kappa(z)$ and $\phi(z)$, we have $\phi' = -\phi'(-z)$, and the above replacement means that we have physically reversed the filter. Since the structure is lossless and reciprocal, the reflectivity is

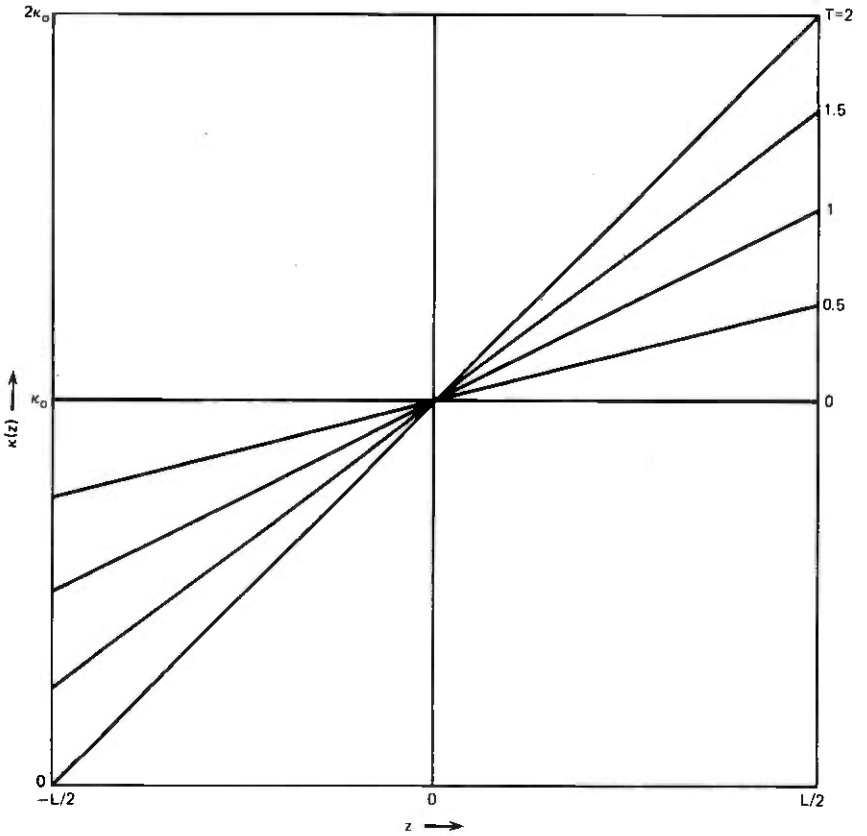


Fig. 1—Linear taper functions $\kappa(z)$ for the taper constants $T = 0$, 0.5, 1, 1.5, and 2.

the same at the input and output port of the filter, and condition (b) is proven. Condition (a) for $\phi' = 0$ is easily proven by inspection of (15).

V. SCALING OF FILTER RESPONSE

In the general case, the Riccati equation (9) has to be evaluated numerically. To make the numerical results more broadly applicable, it is convenient to introduce the normalized quantities z/L, δL, κL, and $\phi' L$. It follows from (9) that two filters of different length (labeled 1 and 2) have the same reflectivity at the frequencies $\delta_1 L_1 = \delta_2 L_2$ if their taper functions obey the scaling rule

$$\begin{aligned} \kappa_1\left(\frac{z}{L_1}\right) L_1 &= \kappa_2\left(\frac{z}{L_2}\right) L_2 \\ \phi_1'\left(\frac{z}{L_1}\right) L_1 &= \phi_2'\left(\frac{z}{L_2}\right) L_2. \end{aligned} \tag{16}$$

VI. TAPERED FILTERS

Tapered filters are structures where the coupling coefficient $\kappa(z)$ varies along the length of the device. For simplicity, we consider here tapered filters with no chirp ($\phi = 0$). At the center frequency ($\delta = 0$), we can write the Riccati equation (9) of such a filter in the form

$$\frac{d\rho}{1 + \rho^2} = j\kappa(z)dz, \tag{17}$$

which is easily integrated to yield for the reflection coefficient

$$\rho(-L/2) = -j \tanh \int_{-L/2}^{L/2} dz.\kappa(z). \tag{18}$$

This is similar to a result found by Kermisch[8] for codirectional interactions in holograms.

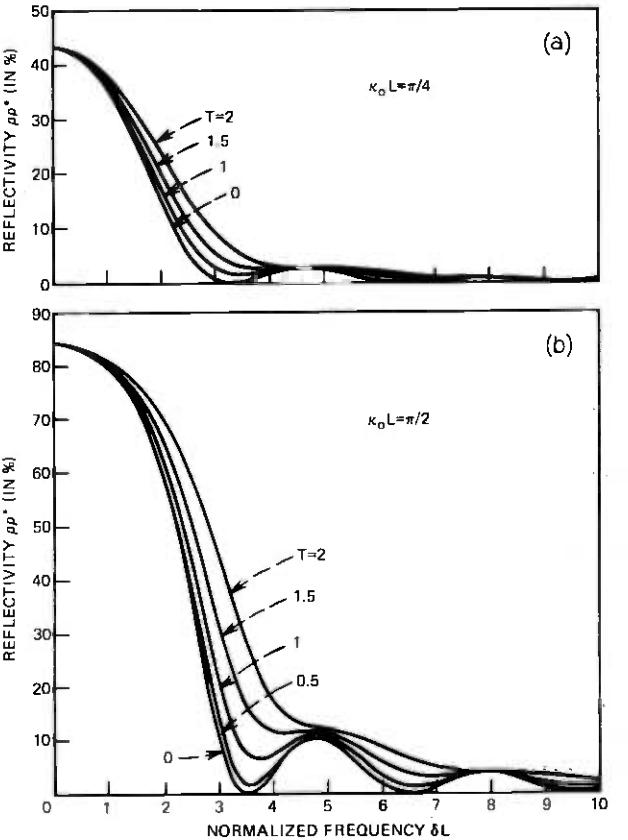


Fig. 2—Frequency response of filters with a linear taper for the taper constants $T = 0, 0.5, 1, 1.5$, and 2 and values of (a) $\kappa_0 L = \pi/4$, (b) $\kappa_0 L = \pi/2$, (c) $\kappa_0 L = 3\pi/4$, and (d) $\kappa_0 L = \pi$.

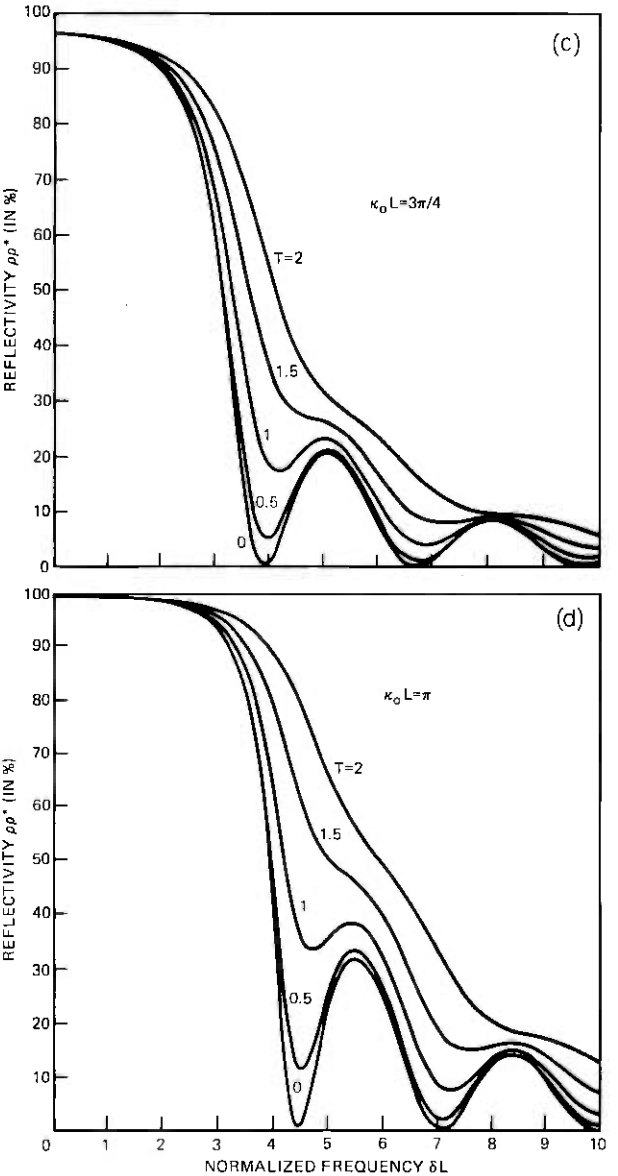


Fig. 2 (continued).

6.1 Linear tapers

In a linearly-tapered filter structure, the coupling coefficient $\kappa(z)$ varies as

$$\kappa = \kappa_0(1 + Tz/L), \tag{19}$$

where the constant T indicates the degree of the taper. Figure 1 shows a sketch of this taper function for five values of T including the cases of no tapering ($T = 0$) and full tapering ($T = 2$). From (18) and (19), we calculate the reflectivity of linearly-tapered filters at the center frequency ($\delta = 0$) as

$$\rho\rho^* = \tanh^2(\kappa_0 L). \tag{20}$$

According to rule (a) of the previous section, the filter response of tapered filters is symmetric. Figures 2a through 2b show the filter response of linearly tapered filters for positive frequency deviations and five values of the taper constant T. The curves are the result of a numerical evaluation of (9) using the fourth-order Runge-Kutta method. Response curves are shown for values of $\kappa_0 L = \pi/4$, $\pi/2$, $3\pi/4$, and π, and we notice a washing-out of the zeros in the response curve with increasing amounts of tapering.

6.2 Quadratic tapers

A filter with a quadratic taper is characterized by a coupling coefficient $\kappa(z)$ that varies as

$$\kappa = \kappa_0\left(1 - \frac{T}{12} + Tz^2/L^2\right). \tag{21}$$

The particular form of this expression has been chosen to make the reflectivity at center frequency ($\delta = 0$) independent of the constant T, and equal to

$$\rho\rho^* = \tanh^2(\kappa_0 L) \tag{22}$$

as calculated from (18). Figure 3 shows quadratic taper functions for five values of T. For positive T the curves are concave upward, and for negative valued T they are convex upward. Figures 4a through 4d show the numerically evaluated response curves for the same values of T and four typical values of $\kappa_0 L$. We note that the concave upward tapers produce very high side-lobe levels in their response characteristics, while the side-lobe levels can be very low for the convex upward tapers. The side-lobe levels of the tapers with $T = -6$ (emphasized by the thicker lines) are the lowest with about 2 percent in reflectivity.

VII. CHIRPED FILTERS

A variation of the grating period along the length of the filter is called a "chirp." Chirped filter characteristics may also be due to a variation of a waveguide parameter such as the refractive index or

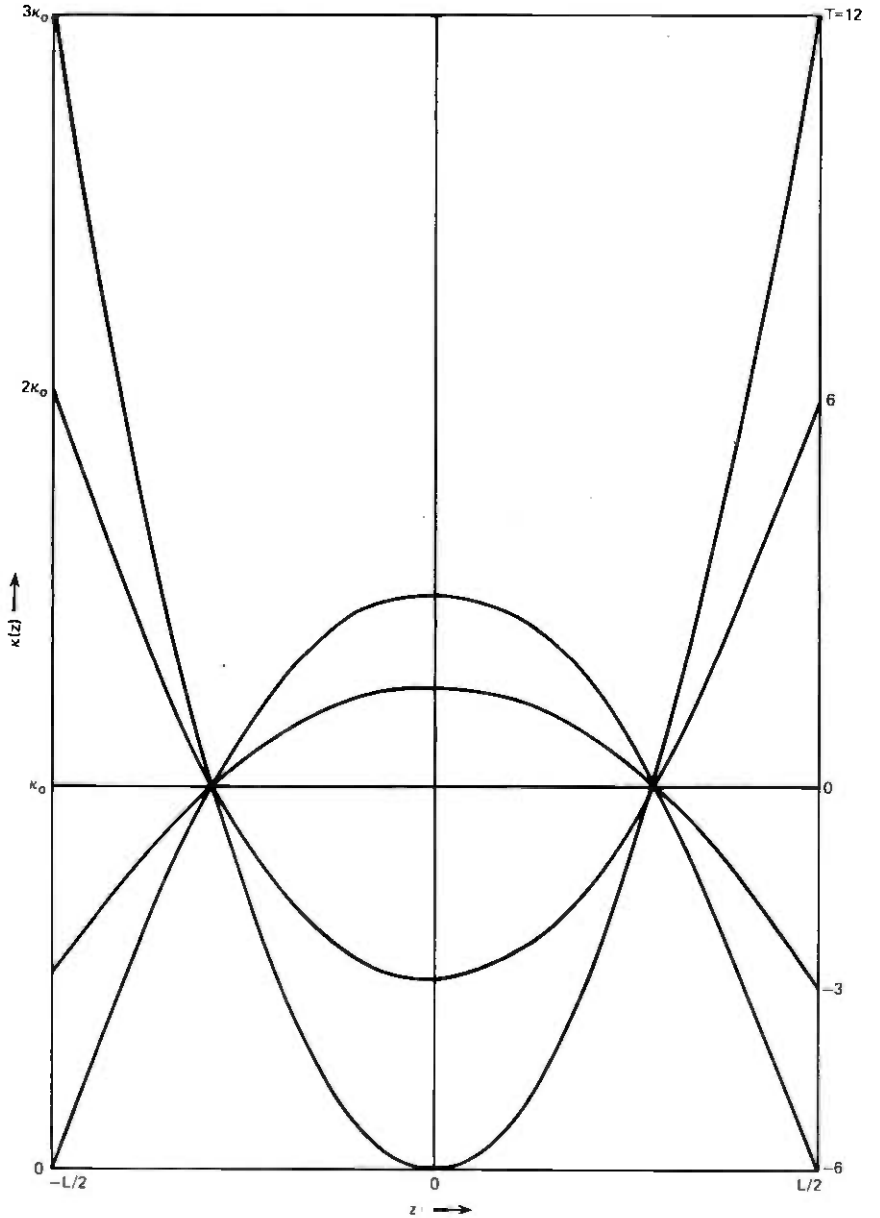


Fig. 3—Quadratic taper functions $\kappa(z)$ for the taper constants $T = -6, -3, 0, 6,$ and 12. The curves intersect at the points $z = \pm L/2 \cdot \sqrt{3}$ and $\kappa = \kappa_0$.

the guide height or width. We have chosen the phase function $\phi(z)$ to represent all variations of this kind. We are, therefore, dealing with a perturbation proportional to cos $(Kz + \phi)$ with constant spatial frequency K and a variable phase shift $\phi(z)$, which can also be viewed as a perturbation with a variable, or chirped, spatial frequency $K + \Delta K(z)$, where

$$\Delta K(z) = \phi'(z), \tag{23}$$

or as a perturbation of variable period $\Lambda + \Delta\Lambda(z)$, where

$$\frac{\Delta\Lambda(z)}{\Lambda} = -\frac{\Delta K(z)}{K}. \tag{24}$$

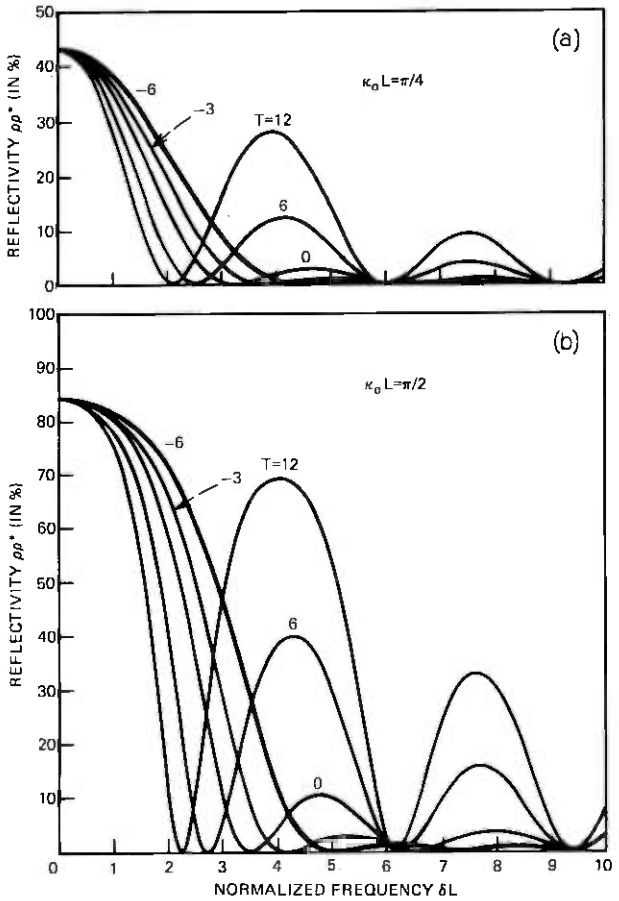


Fig. 4—Frequency response of filters with a quadratic taper for the taper constants $T = -6$, -3, 0, 6, and 12, and values of (a) $\kappa_0 L = \pi/4$, (b) $\kappa_0 L = \pi/2$, (c) $\kappa_0 L = 3\pi/4$, and (d) $\kappa_0 L = \pi$.

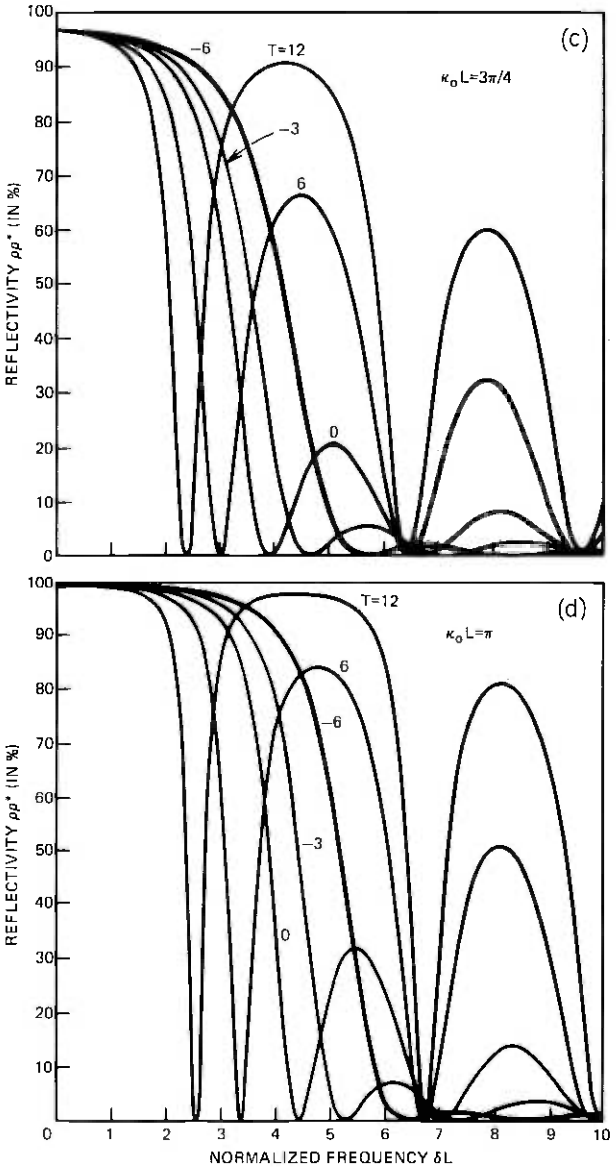


Fig. 4 (continued).

A chirped grating period implies that the (local) Bragg frequency changes along z, which results in a general broadening of the filter response.

7.1 Linear chirp

A periodic structure with a linear chirp is described by a linearly varying spatial frequency which we write in the form

$$\Delta K(z) = \phi'(z) = 2Fz/L^2, \tag{25}$$

where the constant F is a measure for the degree of the chirp. The

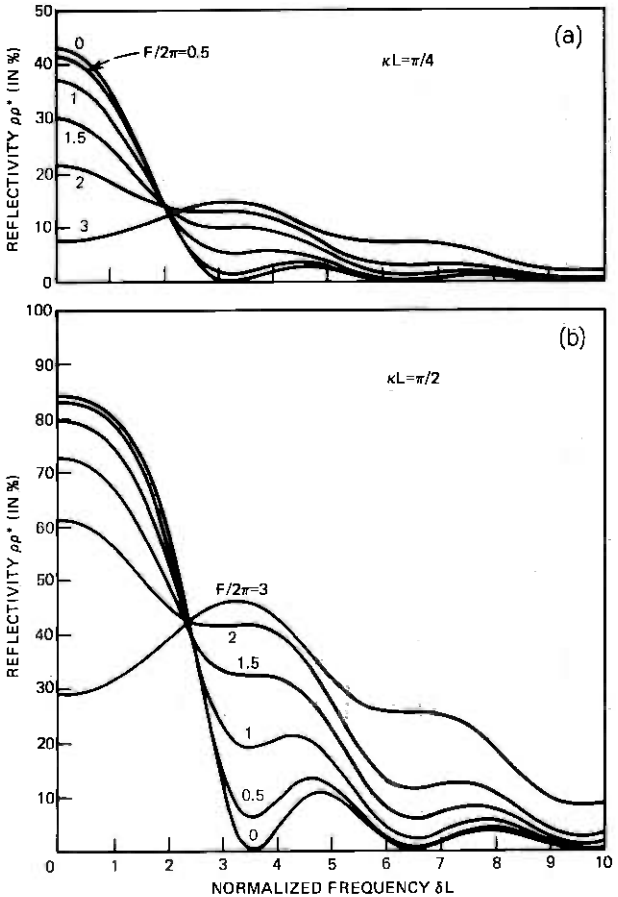


Fig. 5—Frequency response of filters with a linear chirp for the chirp constants $F/2\pi = 0$, 0.5, 1, 1.5, 2, and 3, and for κL-values of (a) $\pi/4$, (b) $\pi/2$, (c) $3\pi/4$, and (d) π.

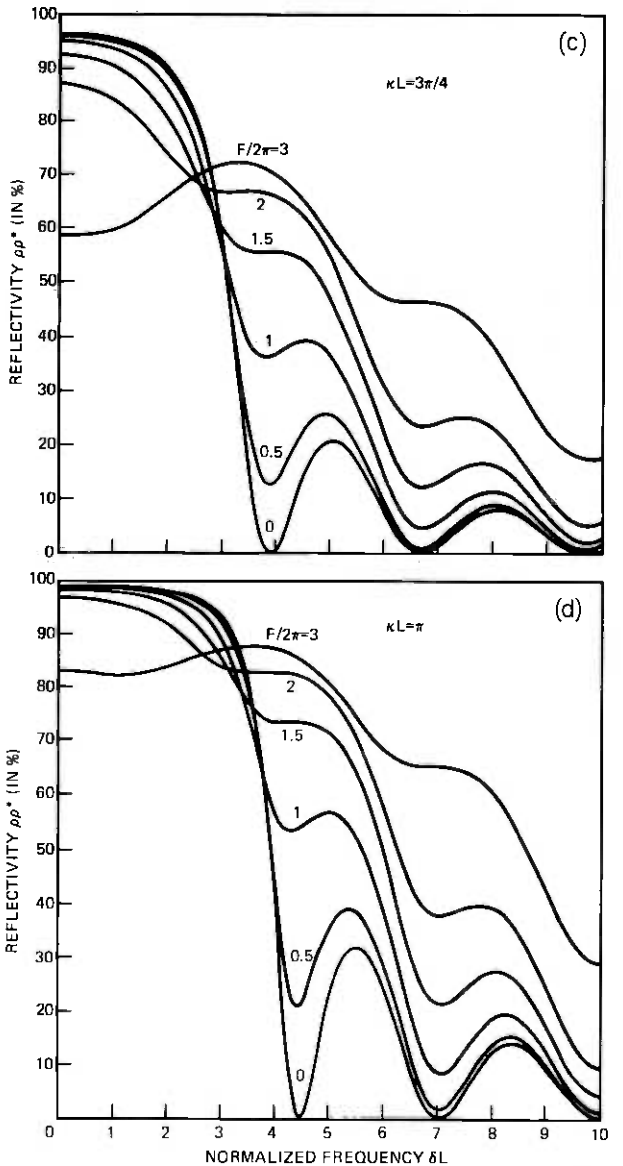


Fig. 5 (continued).

corresponding phase function is

$$\phi(z) = Fz^2/L^2. \tag{26}$$

For the grating period we have

$$\frac{\Delta\Lambda(z)}{\Lambda} = -\frac{\Delta K(z)}{K} = -\frac{F}{2\pi}\cdot\frac{\Lambda}{L}\cdot\frac{2z}{L}. \tag{27}$$

For the difference $\Delta\Lambda$ of the grating periods at the center ($z = 0$) and end ($z = L/2$) of the device, we obtain

$$\frac{\Delta\Lambda}{\Lambda} = \frac{\Lambda(0) - \Lambda(L/2)}{\Lambda} = \frac{F}{2\pi}\cdot\frac{\Lambda}{L}, \tag{28}$$

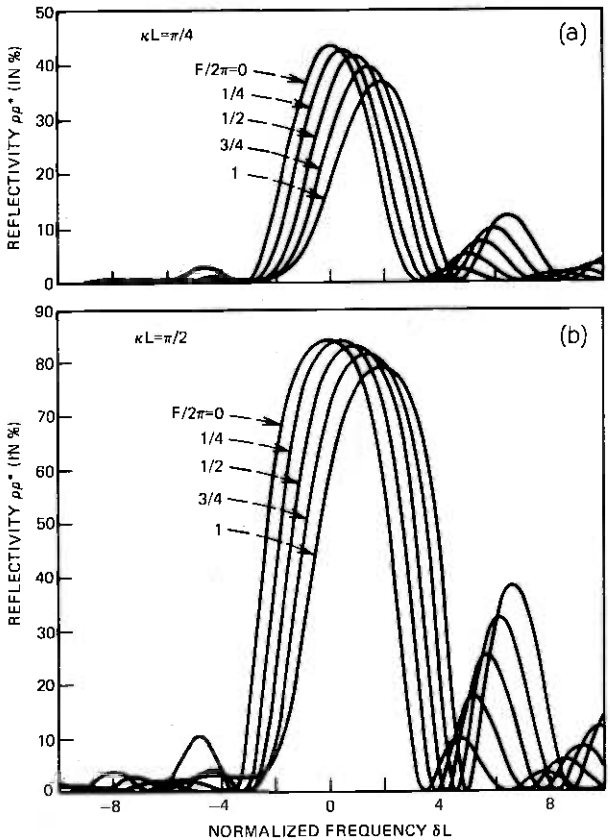


Fig. 6—Frequency response of filters with a quadratic chirp for the chirp constants $F/2\pi = 0$, 0.25, 0.5, 0.75, and 1, and for κL-values of (a) $\pi/4$, (b) $\pi/2$, (c) $3\pi/4$, and (d) π.

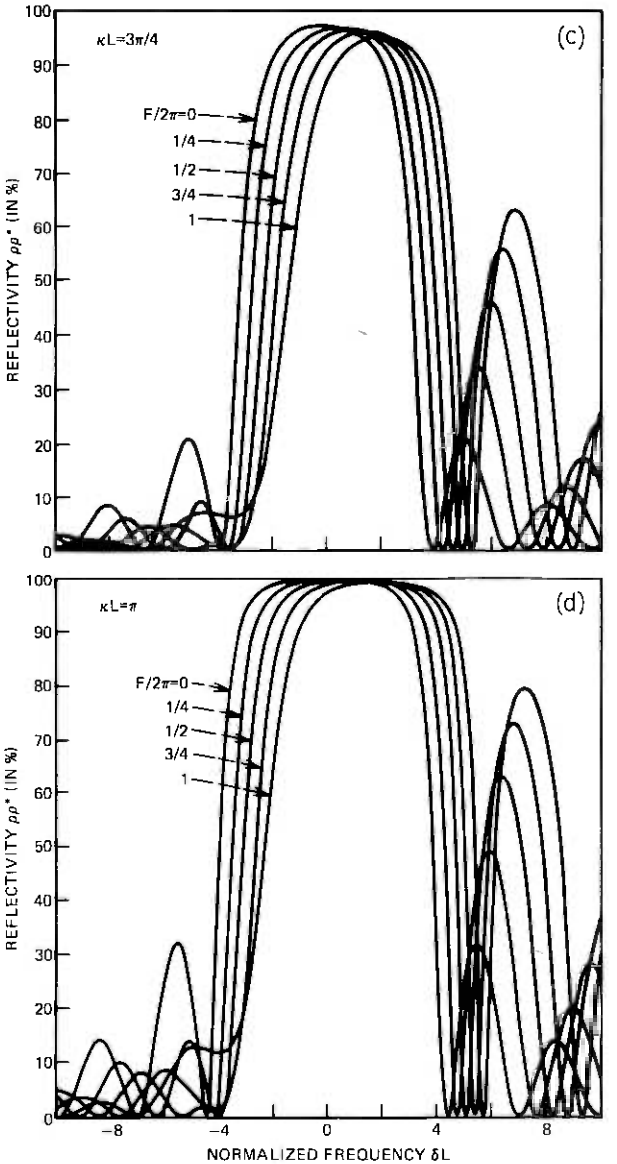


Fig. 6 (continued).

which allows us to express the chirp parameter F in terms of the total number of periods $N = L/\Lambda$ in the form

$$F = 2\pi \cdot N \cdot \Delta\Lambda/\Lambda. \tag{29}$$

According to symmetry rule (b), the response of a filter with a linear chirp is symmetric. The discussion in Section IV implies that a reversal of the sign of F will not change the filter response since this operation simply means that we have turned the two-port structure around to measure the reflectivity at the other port. The filter response as obtained from a numerical evaluation of (9) using the fourth-order Runge-Kutta method is shown in Figs. 5a through 5d. Again, we have chosen for κL the values $\pi/4$, $\pi/2$, $3\pi/4$, and π. We have included the curves for the unchirped filters ($F = 0$) and shown the responses of five chirped filters up to a chirp constant $F/2\pi = 3$. We note that the chirp washes out the zeros and broadens the response.

7.2 Quadratic chirp

For a periodic structure with a quadratic chirp, we write the spatial frequency variation in the form

$$\Delta K(z) = \phi' = 12Fz^2/L^3 \tag{30}$$

using F as a chirp parameter. With this we get for the phase function

$$\phi = 4F(z/L)^3. \tag{31}$$

The variation of the grating period is then given by

$$\frac{\Delta\Lambda(z)}{\Lambda} = -3\,\frac{F}{2\pi}\cdot\frac{\Lambda}{L}\cdot\left(\frac{2z}{L}\right)^2, \tag{32}$$

and the difference between the periods in the middle of the device ($z = 0$) and the device ends ($z = \pm L/2$) becomes

$$\frac{\Delta\Lambda}{\Lambda} = \frac{\Delta\Lambda(0) - \Delta\Lambda(L/2)}{\Lambda} = 3\cdot\frac{F}{2\pi}\cdot\frac{\Lambda}{L}. \tag{33}$$

Again, we can express the chirp parameter F in terms of the total number of periods by

$$3F = 2\pi \cdot N \cdot \Delta\Lambda/\Lambda. \tag{34}$$

The filter response of a periodic structure with a quadratic chirp is asymmetric. If we reverse the sign of F in this case, we are dealing with a different structure which has a different filter response. According to the discussion of Section IV, this new response curve is the mirror image of the response of the original filter relative to the center frequency. The response, as obtained by a numerical evaluation of (9),

is shown in Figs. 6a through 6d for four values of κL. Curves are given for chirp parameters up to $F/2\pi = 1$. We note that with increasing chirps the asymmetry of the response curve increases with the side-lobe levels increasing on one side and decreasing on the other side of the center frequency. Most of the zeros in the response are preserved and shifted somewhat, and there is a small shift in the frequency of peak reflectivity. There appears to be no significant broadening of the main lobe of the response for the parameters selected.

VIII. CONCLUSIONS

We have described the filter characteristics of nonuniform periodic waveguides on the basis of coupled-wave theory. We have also shown that the filter response is described by a Riccati differential equation and have presented solutions for linear and quadratic tapers of the coupling coefficient and for linear and quadratic chirps in the grating period. The results can be used as an aid in filter design and to predict the effect of imperfections introduced during device fabrication. Quadratic tapers appear to be a promising choice when filter characteristics with low side-lobe levels are desired.

IX. ACKNOWLEDGMENT

The author wishes to thank Mrs. Diane Vitello for performing the numerical evaluations.

REFERENCES

1. D. C. Flanders, H. Kogelnik, R. V. Schmidt, and C. V. Shank, "Grating Filters for Thin-Film Optical Waveguide," Appl. Phys. Lett., *24* (February 1974), pp. 194–196.
2. R. V. Schmidt, D. C. Flanders, C. V. Shank, and R. D. Standley, "Narrow-Band Grating Filters for Thin-Film Optical Waveguides," Appl. Phys. Lett., *25* (December 1974), pp. 651–652.
3. A. Yariv, "Coupled-Mode Theory for Guided-Wave Optics," IEEE J. Quantum Electron., *QE-9* (September 1973), pp. 919–933.
4. H. Kogelnik, "Theory of Dielectric Waveguides," in *Integrated Optics*, ed. T. Tamir, Heidelberg: Springer, 1975.
5. J. S. Cook, "Tapered Velocity Couplers," B.S.T.J., *34*, No. 4 (July 1955), pp. 807–822.
6. A. G. Fox, "Wave Coupling by Warped Normal Modes," B.S.T.J., *34*, No. 4 (July 1955), pp. 823–852.
7. W. H. Louisell, "Analysis of the Single Tapered Mode Coupler," B.S.T.J., *34*, No. 4 (July 1955), pp. 853–870.
8. D. Kermisch, "Nonuniform Sinusoidally Modulated Dielectric Gratings," J. Opt. Soc. Amer., *59* (November 1969), pp. 1409–1414.
9. M. G. F. Wilson and G. A. Teh, "Tapered Optical Directionally Coupler," IEEE Trans. Microw. Theory Tech., *MTT-23* (January 1975), pp. 85–92.
10. N. Uchida, "Calculation of Diffraction Efficiency in Hologram Gratings Attenuated Along the Direction Perpendicular to the Grating Vector," J. Opt. Soc. Amer., *63* (March 1973), pp. 280–287.
11. K. O. Hill, "A Periodic Distributed Parameter Waveguide for Integrated Optics," Appl. Opt., *13*, No. 8 (August 1974), pp. 1853–1856.

12. M. Matsuhara and K. O. Hill, "Optical-Waveguide Band-Rejection Filters: Design," Appl. Opt., *13*, No. 12 (December 1974), pp. 2886–2888.
13. M. Matsuhara, K. O. Hill, and A. Watanabe, "Optical-Waveguide Filters: Synthesis," J. Opt. Soc. Amer., *65* (July 1975), pp. 804–809.
14. D. Marcuse, *Theory of Dielectric Optical Waveguides*," New York: Academic Press, 1974.
15. S. E. Miller, "Coupled Wave Theory and Waveguide Applications," B.S.T.J., *33*, No. 3 (May 1954), pp. 661–719.
16. M. G. Cohen and E. I. Gordon, "Acoustic Beam Probing Using Optical Techniques," B.S.T.J., *44*, No. 4 (April 1965), pp. 693–721.

Contributors to This Issue

Michael R. Campbell, B.S.E.E., Syracuse University; M.S.E., 1973, Columbia University; Bell Laboratories, 1971—. Mr. Campbell has been concerned with the switched digital data system and hybrid FIR/IIR filter design. His current assignment is design of service circuits for a new ESS-based SDDS switch.

Ronald E. Crochiere, B.S., (E.E.) 1967, Milwaukee School of Engineering; M.S. (E.E.) and Ph.D. (E.E.), 1968 and 1974, Massachusetts Institute of Technology; Bell Laboratories, 1974—. Mr. Crochiere is presently engaged in research activities in speech communications and digital signal processing. Member, IEEE, Sigma Xi, IEEE Acoustics Speech, and Signal Processing Group Technical Committee on Digital Signal Processing.

Alfred Descloux, Math. Dipl., 1948, Swiss Federal Institute of Technology, Zürich; Ph.D. (Mathematical Statistics), 1961, University of North Carolina; Bell Laboratories, 1956—. Mr. Descloux spent 1955–1956 on the staff of the University of Washington where he taught mathematics and statistics. At Bell Laboratories, he has been concerned chiefly with the application of probability theory to traffic problems. Member, Institute of Mathematical Statistics, American Mathematical Society.

Donald L. Duttweiler, B.E.E., 1966, Rensselaer Polytechnic Institute; M.S. (E.E.), 1967, and Ph.D. (E.E.), 1970, Stanford University; Bell Laboratories, 1970—. At Bell Laboratories, Mr. Duttweiler has studied various digital communications problems. Member, Eta Kappa Nu, Tau Beta Pi, Sigma Xi.

Herwig Kogelnik, Dipl. Ing., 1955; Dr. techn., 1958, Technische Hochschule Wien, Austria; D. Phil., 1960, Oxford University, England; Bell Laboratories, 1961—. Mr. Kogelnik is head of the Coherent Optics Research Department. He is engaged in research in optical communications, lasers, and optical devices. Member, American Physical Society, American Association for the Advancement of Science; Fellow, IEEE, American Optical Society.

Lawrence R. Rabiner, S.B., S.M., 1964, Ph.D., 1967, Massachusetts Institute of Technology; Bell Laboratories 1962—. Mr. Rabiner has 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. Coauthor, *Theory and Application of Digital Signal Processing*. Member, Eta Kappa Nu, Sigma Xi, Tau Beta Pi; Fellow, Acoustical Society of America; President, IEEE G-ASSP Ad Com; member, G-ASSP Technical Committee on Digital Signal Processing, G-ASSP Technical Committee on Speech Communication, IEEE Proceedings Editorial Board, Technical Committee on Speech Communication of the Acoustical Society; former Associate Editor of the G-ASSP Transactions.

Thomas C. Spang, B.E., 1956, M.E., 1958, and Ph.D., 1961, Yale University; Bell Laboratories, 1960—. Mr. Spang has been responsible for system engineering studies to establish recommendations for transmission performance of speech signals. Currently, he is a supervisor in the Private Network Planning Department.

THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 55, No. 1, January 1976
Printed in U.S.A.

B.S.T.J. BRIEFS

A Simple Method for Estimating Five-Minute Point Rain-Rate Distributions Based on Available Climatological Data

By W. Y. S. CHEN

(Manuscript received August 27, 1975)

This brief presents a method for estimating 5-minute rain-rate distributions based on climatological data for over 250 locations as published annually by the National Oceanic and Atmospheric Administration (NOAA). A *maximized* 5-minute rain-rate distribution, which is believed to be more appropriate for use in predicting rain attenuation, is also introduced.

The sample form in Table I, showing the NOAA method of computation of maximum short-time rates of rainfall for the years 1896–1935, is taken from Ref. 1. On this form are shown the storms of July 14, 1912, and September 2, 1922, at Washington, D. C. The times of beginning and ending of the storms and other data are entered on the "Obs. Precip." line, which shows the precipitation at consecutive 5-minute intervals up to 50 minutes from the beginning of the period of excessive precipitation, as copied from Weather Bureau records. The "Increment" entries are computed directly from the figures on the line above. The maximum precipitation shown for any storm is the maximum for the time period indicated by the figure in the column heading, and is determined by selection from the 5-minute increments. For example, in the 1912 storm, the maximum precipitation indicated for 5 minutes occurs in the sixth increment, that for 10 minutes is found by combining the fifth and sixth increments, and that for 20 minutes is found by combining the fourth, fifth, sixth, and seventh increments. In the 1922 storm, the stated maximum for 5 minutes is half of the eleventh increment, which is for 10 minutes. The 10-minute maximum is the eleventh increment, but the 15-minute maximum is the second, third, and fourth together.

Table I — Sample form of computation of maximum short-time rates of rainfall (from Ref. 1)

U. S. Department of Agriculture
Compiler.....

Intense Precipitation Records

	Station	Duration of Storm		Total Precip.	Time Excess Prec. Began	Prec. Prior Excess Rate	Time in Minutes													
		From	To				5	10	15	20	25	30	35	40	45	50	60	80	100	120
1912	Date 1912 July 14																			
	Obs. Precip.	p 5.40	p 10.20	2.93	p 6.01	0.02	0.05	0.25	0.59	0.93	1.38	1.78	2.14	2.38	2.50	2.54				
	Increment						0.05	0.20	0.34	0.34	0.45	0.40	0.36	0.24	0.12	0.04				
	Max. Precip.						0.45	0.85	1.21	1.55	1.89	2.13	2.33	2.45	2.50	2.54				
1922	Date 1922 Sept. 2																			
	Obs. Precip.	a 1.55	a 10.25	5.16	a 5.32	0.66	0.21	0.41	0.67	0.93	1.10	1.25	1.37	1.51	1.65	1.81	2.35	2.96	3.45	
	Increment						0.21	0.20	0.26	0.26	0.17	0.15	0.12	0.14	0.14	0.16	0.54	0.61	0.49	
	Max. Precip.						0.27	0.54	0.72	0.93	1.10	1.25	1.37	1.51	1.71	1.94	2.35	2.96	3.45	

The present NOAA method, adopted with data for the calendar year 1936, gives the maximum fall of precipitation for periods of 5 to 180 minutes, the amounts being taken for the periods in which the fall is greatest for the given length of time without regard to a prescribed starting or sampling time and is tabulated for 5, 10, 15, 20, 30, 45, 60, 80, 100, 150, and 180 minutes. A sample[2] of the NOAA climatological data for maximum precipitations in 1972 at Miami, Florida is shown in Table II. Yarnell reported that the maximum 5-minute precipitation, as determined by the newer method (post-1935), exceeds the maximum of the earlier method (pre-1935) by 8 to 10 percent. Similarly, comparisons for the 1-hour maxima showed a 4- to 5-percent increase.[1]

The procedure for obtaining 5-minute rain-rate distributions, based on available climatological data since 1896, is as follows.

(*i*) Multiply the maximum precipitations for the time periods 5 to 180 minutes by the following factors:

Time Periods in Minutes	F
5	0.917
10	0.922
15	0.926
20	0.931
30	0.938
45	0.948
60	0.957
80	0.966
100	0.973
120	0.979
150	0.986
180	0.990

This table is made up from the formula

$$F = 1/(1 + 0.096e^{-0.0126t}),^*$$

where t is the time in minutes. (Omit this step for 1935 and earlier data.)

(*ii*) Convert the maximum precipitations into increment rates by taking differences between all pairs of adjacent elements in succession.

* This formula was obtained by fitting an exponential function to the two observed values reported by Yarnell: $1/F = 1.09$ at $t = 5$ minutes and $1/F = 1.045$ at $t = 60$ minutes.

Table II — Excessive short-duration rainfall (from Ref. 2)

Year 1972

Station and Date	Maximum Precipitation in Inches (5 to 180 Minutes)											
	5	10	15	20	30	45	60	80	100	120	150	180
Miami												
Feb 2	0.25	0.45	0.49	0.60	0.75	0.79	0.81	0.85	0.86	0.87	0.87	0.88
Mar 5	0.26	0.44	0.49	0.55	0.55	0.55	0.55	0.55	0.58	0.78	0.78	0.78
Mar 17	0.28	0.31	0.36	0.38	0.39	0.40	0.40	0.40	0.40	0.40	0.40	0.40
Apr 1	0.25	0.36	0.40	0.53	0.55	0.62	0.63	0.63	0.67	0.75	0.77	0.82
May 6	0.14	0.27	0.37	0.40	0.58	0.70	0.82	1.05	1.25	1.37	1.40	1.50
May 10	0.19	0.32	0.39	0.48	0.55	0.70	0.76	0.80	0.81	0.81	0.81	0.81
May 18	0.45	0.65	0.85	1.05	1.55	1.72	1.75	1.76	1.76	1.76	1.76	1.76
May 18	0.35	0.48	0.63	0.78	0.87	1.07	1.15	1.15	1.15	1.15	1.15	1.15
May 19	0.25	0.44	0.55	0.71	0.93	0.94	1.02	1.05	1.07	1.07	1.07	1.07
May 20	0.29	0.53	0.60	0.66	0.77	0.82	0.87	0.92	0.92	0.92	0.92	0.92
May 22	0.25	0.35	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40	0.40
May 26	0.60	1.16	1.40	1.42	1.62	1.90	2.07	2.07	2.07	2.09	2.11	2.11
May 31	0.28	0.31	0.32	0.32	0.34	0.34	0.34	0.34	0.34	0.44	0.65	0.65
Jun 1	0.40	0.70	0.80	0.90	0.91	0.95	1.07	1.09	1.11	1.11	1.11	1.11
Jun 11	0.60	0.90	1.00	1.20	1.80	1.95	2.00	2.01	2.05	2.07	2.17	2.24
Jun 11	0.30	0.52	0.60	0.65	0.83	1.20	1.38	1.45	1.50	1.53	1.56	1.65
Jun 12	0.40	0.62	0.70	0.73	0.90	1.15	1.32	1.52	1.74	1.74	1.79	1.82
Jun 26	0.25	0.33	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35
Jul 11	0.55	0.85	1.00	1.05	1.07	1.08	1.08	1.08	1.08	1.08	1.08	1.08
Jul 19	0.22	0.30	0.50	0.60	0.80	0.89	0.91	0.91	0.92	1.16	1.16	1.18
Jul 31	0.25	0.50	0.52	0.54	0.57	0.59	0.62	0.70	0.74	0.77	0.82	0.84
Aug 6	0.35	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Aug 15	0.41	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
Aug 23	0.22	0.43	0.47	0.50	0.55	0.57	0.58	0.85	0.90	0.93	0.95	0.95
Aug 28	0.45	0.65	0.86	1.20	1.70	1.85	1.97	1.99	2.00	2.01	2.05	2.05
Sep 4	0.27	0.32	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35
Sep 4	0.20	0.33	0.40	0.45	0.62	0.65	0.65	0.65	0.65	0.65	0.65	0.65
Sep 24	0.30	0.55	0.60	0.65	0.92	0.93	0.95	1.05	1.20	1.22	1.25	1.27
Oct 3	0.22	0.34	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35
Oct 28	0.33	0.60	0.68	0.75	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78
Nov 15	0.27	0.36	0.43	0.47	0.50	0.53	0.58	0.62	0.68	0.72	0.75	0.75
Dec 4	0.20	0.35	0.40	0.45	0.55	0.75	0.90	1.00	1.07	1.15	1.18	1.20
Dec 15	0.30	0.36	0.39	0.43	0.55	0.65	0.68	0.68	0.68	0.68	0.82	1.08

(*iii*) Obtain the uniform 5-minute rain rates for those elements with time intervals greater than 5 minutes. For example, between 20 and 30 minutes, assume two 5-minute intervals with equal rain rates that are equal to half of the 10-minute increment rate; similarly, between 30 and 45 minutes, 5-minute increment rates are one-third of the 15-minute increment rate, etc.

(*iv*) Find the distribution based on the 5-minute rain rates as obtained above.

Twenty randomly selected rainstorms taken from Ref. 1, Table I, were used to test the accuracy of the method. The result in Fig. 1

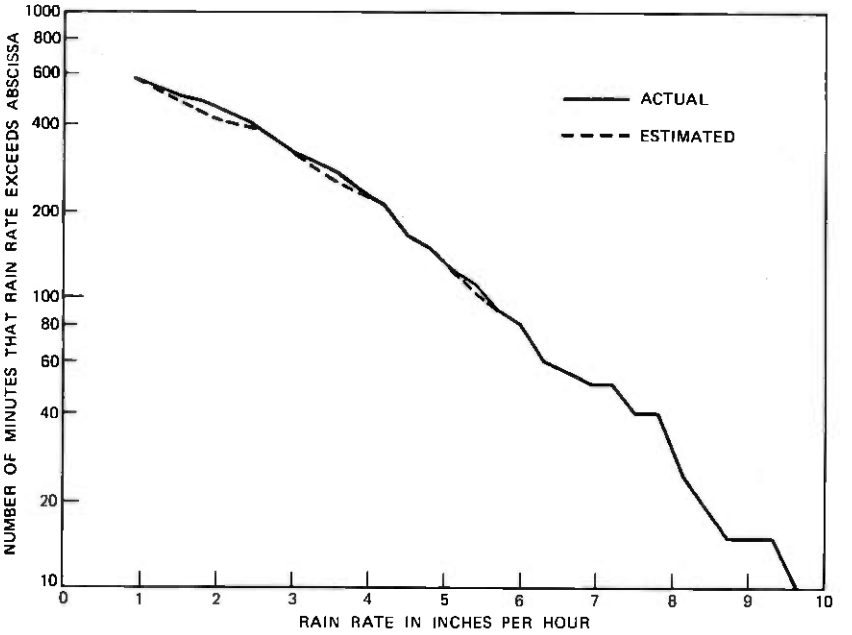


Fig. 1—Five-minute rain-rate distributions for 20 randomly selected, heavy rainstorms.

shows that the distribution computed by this method is in good agreement with the actual 5-minute rain-rate distribution. The actual distribution was obtained from the "Increment" entries in Table I and the estimated distribution (Fig. 1) was obtained from the "Maximum for Period" entries with those for periods 25, 35, 40, and 50 minutes deleted. The deletion was made to enable the original data to conform to the published data. It is seen from Fig. 1 that the error of the estimated distribution is small compared with that of the actual distribution. The trend of the error indicates that the higher the rain rate, the smaller the error. The maximum error in the range of interest is in the order of 5 percent.

The 1972 Miami data given in Table II were also processed using this method. Thirty-three heavy rainstorms occurred at Miami, Florida, in 30 days during 1972. The resulting 5-minute rain-rate distribution is shown by the dashed curve in Fig. 2.

A similar distribution can be obtained from the post-1935 NOAA data using the method by skipping the first step. The distribution curve thus obtained will lie above the 5-minute randomly sampled distribution and will be called the maximized 5-minute distribution.

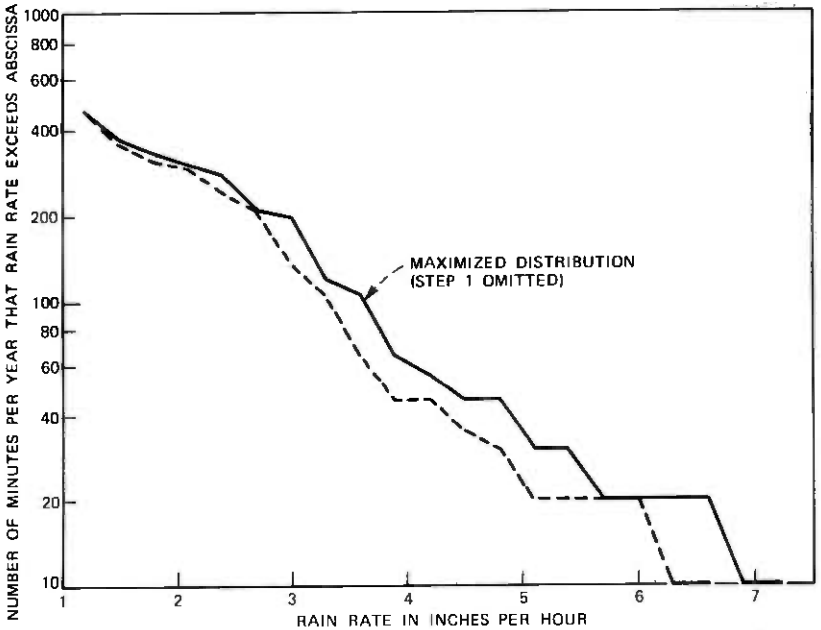


Fig. 2—Five-minute rain-rate distributions for Miami, Florida (1972).

Since this new distribution is expected to be closer to the 1-minute rain-rate distribution, it is believed to be more useful than its corresponding 5-minute rain-rate distribution for the purposes of attenuation estimation.[3] The maximized 5-minute rain-rate distribution for Miami, Florida, in 1972 was obtained and the results compared with the 5-minute rain-rate distribution obtained above, as shown by the solid curve in Fig. 2. It should be noted that, in a low rain-rate region of the distribution (<3 in./hr.), there exists a problem of "thresholds." This is the subject of the companion paper.[4]

REFERENCES

1. D. L. Yarnell, *Rainfall Intensity-Frequency Data*, U. S. Department of Agriculture, Miscellaneous Publication No. 204, August 1935.
2. "Climatological Data," U.S. Dept. of Commerce, NOAA Annual 1972, Asheville, North Carolina, *23*, No. 13.
3. W. F. Bodtmann and C. L. Ruthroff, "Rain Attenuation on Short Radio Paths: Theory, Experiment and Design," B.S.T.J., *53*, No. 7 (September 1974), pp. 1329–1349.
4. S. H. Lin, "Dependence of Rain-Rate Distribution on Rain-Gauge Integration Time," B.S.T.J. this issue, pp. 135–141.

Dependence of Rain-Rate Distribution on Rain-Gauge Integration Time

By S. H. LIN

(Manuscript received September 17, 1975)

An important problem in designing terrestrial and earth-satellite radio systems at frequencies above 10 GHz is the radio outage caused by rain attenuation. Since rain-rate statistics vary significantly from year to year, long-term rain-rate distributions are needed in radio path engineering to meet a given reliability objective. This paper and Ref. 2 together describe a method and the necessary data sources to obtain 20-year distributions of 5-min rain rates at about 250 locations in the U. S. A.

In this paper, a "5-min rain rate" corresponds to the average value of the randomly varying rain rate in a 5-min interval and is calculated as $\Delta H/T$, where ΔH is the 5-min accumulated depth of rainfall and $T = 5$ min $= 1/12$ h is the rain-gauge integration time. Similarly, a "T-min rain rate" is the average rain rate in a T-min interval.

The "excessive short duration rainfall data" for about 250 locations in the U. S. A., published by the National Climatic Center,[1] records details of those rainfalls which exceed a threshold which is a function of rain-gauge integration time T as shown in Table I. For example, the thresholds are 76 and 20 mm/h for $T = 5$ and 60 min, respectively. A method for obtaining 5-min rain-rate distributions from these long-term ($\geq$ 20 years) data is described in a companion paper.[2]

A straightforward extension of this method allows us to obtain rain-rate distributions appropriate to other integration times, such as 10, 15, 30, and 60 min. Figure 1 shows 20-year distributions with various integration time in the New York metropolitan area. The distributions for the New York metropolitan area represent an average of those obtained from LaGuardia Airport, Central Park, and Newark Airport.

The thresholds noted in Table I, however, indicate that the rain-rate distributions processed from these data are accurate only in the range above the thresholds. The 5-min rain-rate distribution above 76 mm/h is generally sufficient for engineering terrestrial radio paths at frequencies above 10 GHz in the eastern or midwestern U. S. A. For engineering long terrestrial paths in the western U. S. A. or for millimeter-wave satellite radio links, distributions for rain rates less than 76 mm/h are needed. Fortunately, the long-term hourly precipitation data published by The National Climatic Center[3,4] contain all rainfalls. As a result, the low-rain-rate portion of the 5-min rain-rate distribution

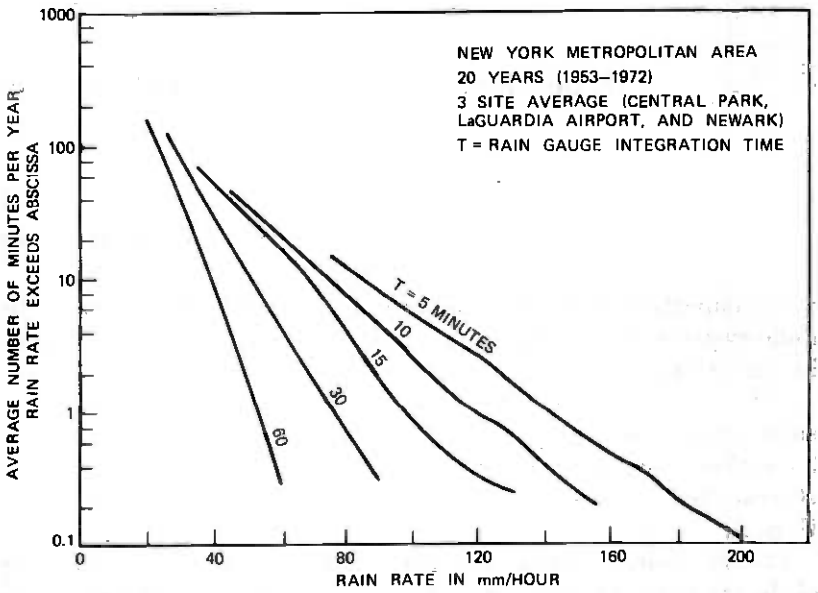


Fig. 1—20-year distributions of rain rates in the New York metropolitan area.

can be estimated by applying an empirical conversion factor, which we now derive, to the hourly rain-rate distribution.

From Fig. 1, we obtain the relationship between 5-min rain rate and 60-min rain rate at several probability levels. This is displayed in Fig. 2.

Table I — Thresholds of excessive short duration rainfalls

Duration T (min)	Minimum Depth of Recorded Rainfall (in.)	Threshold (T-Minute Average-Rain Rate, mm/hr)
5	0.25	76.2
10	0.30	45.7
15	0.35	35.6
20	0.40	30.5
30	0.50	25.4
45	0.65	22.0
60	0.80	20.3
80	1.00	19.1
100	1.20	18.3
120	1.40	17.8
150	1.70	17.3
180	2.00	16.9

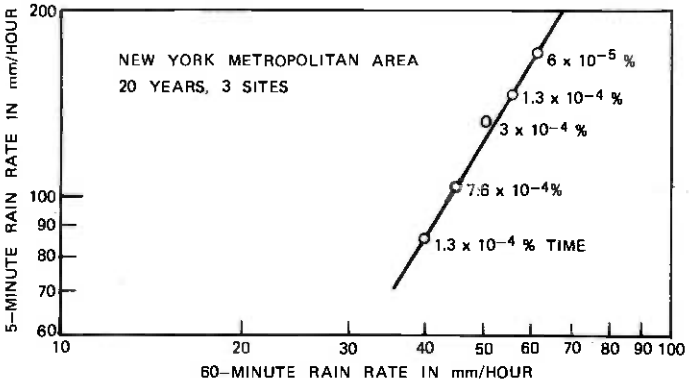


Fig. 2—Relationship between 5-min rain rate and 60-min rain rate at the same probability level in the New York metropolitan area.

Additional results from Miami, Florida and McGill Observatory, Canada,[5] shown in Fig. 3, indicate that this relationship varies significantly with geographic location. However, such geographic variations are removed if the rain rates of Fig. 3 are normalized with respect to

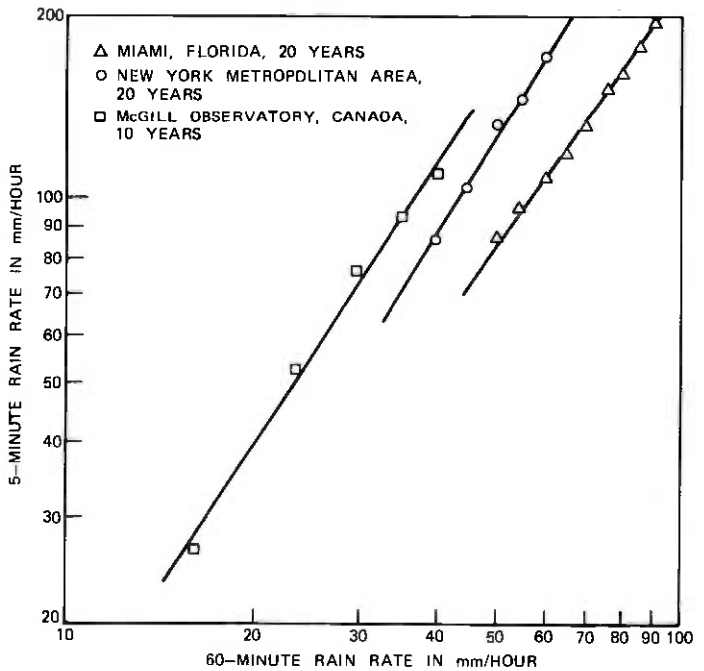


Fig. 3—Relationships between 5-min rain rate and 60-min rain rate at the same probability level in New York, Miami, and McGill Observatory.

a geographic-dependent, characteristic rain rate R_c. Let us define a dimensionless variable,

$$X(T) = \frac{R(T)}{R_c}, \tag{1}$$

as a measure of the normalized rain rate $R(T)$. Since the purpose of introducing R_c is to remove the geographic variable factor in the relationships in Fig. 3, R_c must be geographically dependent. A reasonable choice for R_c is the average 60-min rain rate $\bar{R}(60)$. For Miami, the New York metropolitan area, and McGill Observatory, $\bar{R}(60)$ is 3.6, 1.7, and 1.1 mm/h, respectively. Figure 4, which incorporates this normalization, indicates that the long-term relationship between $X(5)$ and $X(60)$ is almost the same for Miami, the New York metropolitan area, and McGill Observatory. Figure 5 shows similar results for $T = 10$ and 30 min. These relationships can be approximately described by

$$X(T) \simeq a(T) \cdot [X(60)]^{\eta(T)}, \tag{2}$$

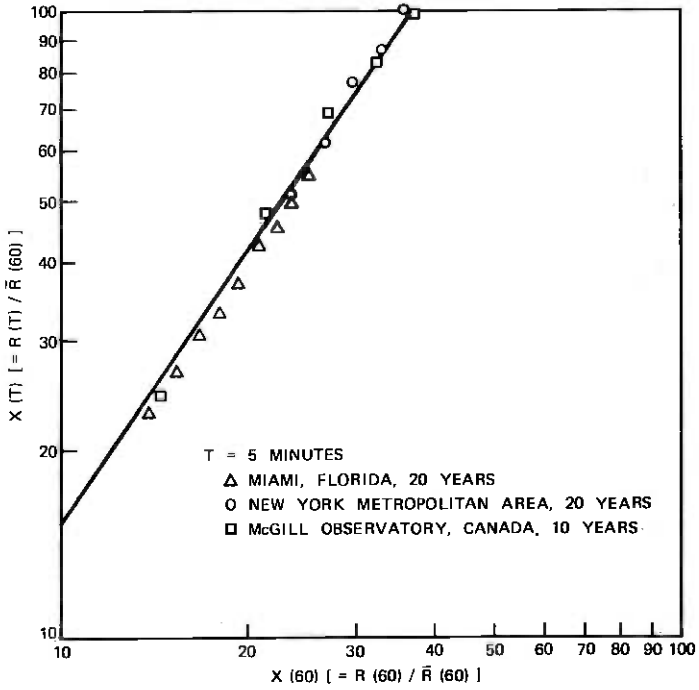


Fig. 4—Relationship between normalized 5-min rain rate and normalized 60-min rain rate at the same probability level.

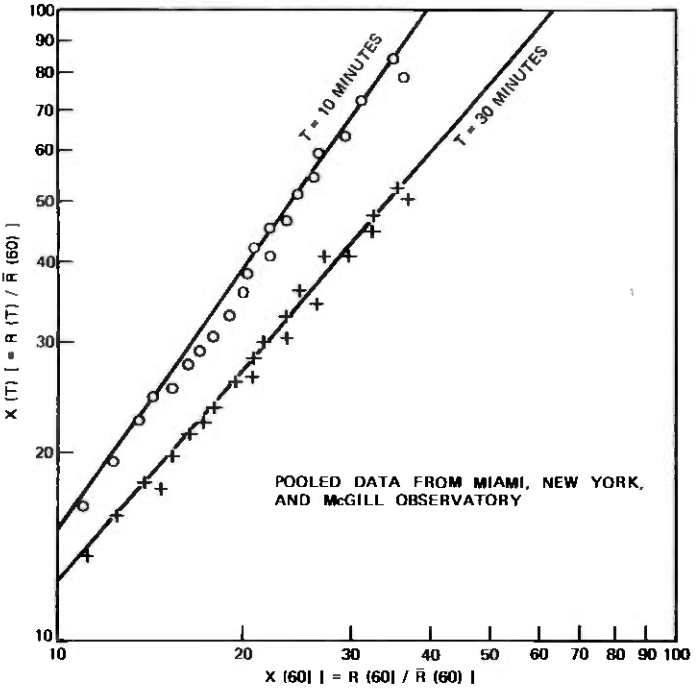


Fig. 5—Relationships among normalized 10-min, 30-min, and 60-min rain rates at the same probability level.

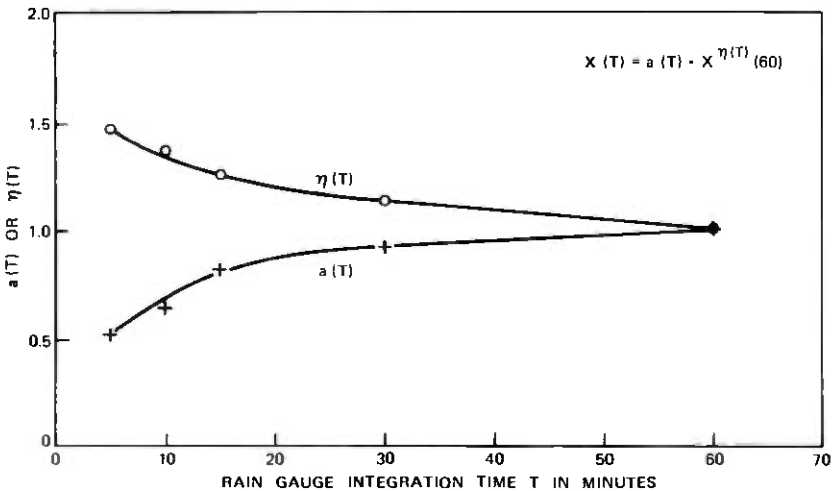


Fig. 6—Empirical dependence of normalized rain rate on rain-gauge integration time T at a given probability level.

where the coefficient $a(T)$ and the exponent $\eta(T)$, as functions of integration time T, are plotted in Fig. 6. We note that the relationships in Figs. 4 through 6 hold only in the long term (e.g., 20 years). The short-term (e.g., <5 years) results can deviate significantly from the long-term average relationship.

The moderate region (30 mm/h $\leqq R \leqq$ 76 mm/h) of the 5-min rain-rate distribution can be estimated by applying the empirical conversion factor to the 60-min rain-rate distribution whose threshold is 20 mm/h in the excessive short-duration rainfall data. Furthermore, long-term hourly precipitation data are available for about 3000 locations[3,4] in the U. S. A. The National Climatic Center has processed some of these raw data to obtain 10-yr (1951–1960) distributions of hourly precipitations at 105 locations.[4] Therefore, we have a procedure and the necessary data sources to obtain long-term distributions of 5-min rain rate covering the entire range of interest to the design of both terrestrial and earth-satellite radio systems. For example, Fig. 7 shows

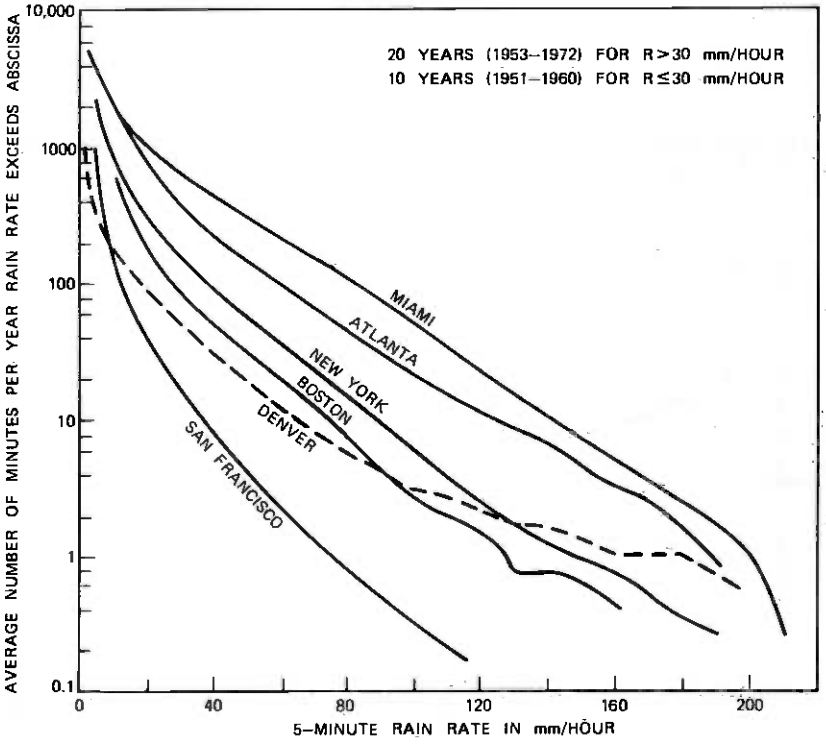


Fig. 7—Long-term distributions of 5-min rain rates in Miami, Atlanta, New York, Boston, Denver, and San Francisco.

the 5-min rain-rate distributions obtained by this method for Miami, Atlanta, New York, Denver, Boston, and San Francisco.

REFERENCES

1. "Climatological Data, National Summary," annual issues since 1950, U. S. Department of Commerce, National Oceanic and Atmospheric Administration, National Climatic Center, Federal Building, Asheville, North Carolina 28801. Also available from the Superintendent of Documents, Government Printing Office, Washington, D. C. 20402. The excessive short-duration rainfall data prior to 1950 are published in the Monthly Weather Review, the U. S. Meteorological Yearbook (last published for the period 1943–1949), and the Report of the Chief of the Weather Bureau (last published for 1931).
2. W. Y. S. Chen, "A Simple Method for Estimating Five-Minute Rain-Rate Distributions Based on Available Climatological Data," B.S.T.J., this issue, pp. 129–134.
3. "Hourly Precipitation Data," published for each state in the U. S. A., Superintendent of Documents, Government Printing Office, Washington, D. C. 20402. The hourly precipitation data for approximately 3000 locations are also stored on magnetic tapes in the Computer Center of The National Climatic Center, Federal Building, Asheville, North Carolina 28801.
4. "Climatography of the United States No. 82, Decennial Census of United States Climate—Summary of Hourly Observations (1951–1960)." National Climatic Center, Federal Building, Asheville, North Carolina 28801.
5. G. Drufuca and I. I. Zawadzki, "Statistics of Rain Gauge Records," Inter-Union Commission on Radio Meteorology (IUCRM) colloquium on "The Fine Scale Structure of Precipitation and Electromagnetic Propagation," Nice, France, October 23–31, 1973, Conference Record, Vol. 2, available from Dr. I. Revah, co-chairman of the conference, Department RSR, CNET, 38 Rue du General Leclerc, 92131 Issy les Moulineaux, France.