VOLUME XXV OCTOBER, 1946 NO. 4

THE BELL SYSTEM TECHNICAL JOURNAL

DEVOTED TO THE SCIENTIFIC AND ENGINEERING ASPECTS OF ELECTRICAL COMMUNICATION

AMERICAN TELEPHONE AND TELEGRAPH COMPANY
NEW YORK

50¢ per copy *$1.50 per Year*

THE BELL SYSTEM TECHNICAL JOURNAL

Published quarterly by the
American Telephone and Telegraph Company
195 Broadway, New York, N. Y.

SUBSCRIPTIONS

Subscriptions are accepted at $1.50 per year. Single copies are 50 cents each. The foreign postage is 35 cents per year or 9 cents per copy.

PRINTED IN U. S. A.

The Bell System Technical Journal

Vol. XXV *October, 1946* *No. 4*

A Study of the Delays Encountered by Toll Operators in Obtaining an Idle Trunk

By S. C. RAPPLEYE

THE aim of the Bell System is to give the fastest possible toll service consistent with costs. The aim of the Intertoll Trunk Engineer is to provide the proper number of trunks in each group to obtain that objective. His problem is to gauge the effect of his work on the overall speed of service.

Overall speed of toll service is the elapsed interval from the filing of a call until conversation starts or until there is a definite report about the called party. This overall speed includes the operating time or interval required for the operators to establish the connection; the subscriber time or interval required for the calling party to give the details of the call, for the called party to answer his telephone, etc.; and the circuit delay time or interval of waiting for a trunk to become idle. This last factor may be termed the *trunk speed* interval.

The proportion of this trunk speed to the overall speed is an important factor in determining the number of trunks to be provided. If it is a large proportion of the total, a marked improvement may be expected as a result of providing more trunks. Conversely, if the trunk speed is a small proportion of the total, the improvement to be expected as a result of providing more trunks will be small also, with a diminishing rate of improvement until the trunk speed ceases to be a factor. The trunk speed in turn depends upon three factors:

Group size—number of trunks to the called city or in the direction of the called city.

The per cent. usage—the degree to which the trunks are kept busy in carrying the load offered.

Holding time—the length of time that a trunk is in use each time it is used.

Since the trunk speed depends in part on the per cent. usage, it follows that this interval will be longer in the busiest hour when the usage is greatest and will be shorter in the hours which are less busy. Consequently, the trunk speed interval over the total day will be much less than in the busiest hour.

Purpose of Study

Earlier information, based on data assembled in Cleveland in 1929 and 1930, was formulated as a series of relationships between varying degrees of loading (in terms of busy hour per cent. use) on trunk groups of different sizes and the *overall* speed of service. These relationships were set forth in a table which was to be used as a guide to the trunk provision needed to accomplish a desired overall service result.

The table also furnished the *percent* calls encountering an NC (no circuit) condition but made no specific reference to the average *duration* of NC although from the data shown it could be inferred and demonstrated that other factors, such as operating method, operating and party delays, normally have a more pronounced influence on the total day overall speed of service than the busy hour trunk provision. That being so, as changing conditions since 1930 have affected these other factors, either in the direction of faster or slower service, the relationships in terms of overall speed of service shown in that table have become less valuable as engineering guides.

The purpose of the current study, therefore, was to improve the engineering and management tools used in determining the number and arrangement of trunks required to attain faster toll service so that the investment in facilities may be used as effectively as possible.

Study Procedure

The study was based on the premise that if the size of group, per cent. usage and holding time are known, the trunk speed can be determined and will remain constant under that particular set of conditions. With this constant known, it would then become possible to construct from analyses of overall speed of service data for groups, offices, areas or networks the going relationship between the trunk speed of service and the overall speed of service and to predict with reasonable assurance the effect on the overall speed which would be brought about by changes in the group sizes or traffic characteristics. The effect of foreseen changes in operating method, force conditions or the character of the toll traffic on the overall speed can be estimated separately and taken into consideration in determining the basis of trunk provision. With such information available, trunks can be provided where they will be most effective. This is especially important during periods of major change such as the transition from war to peacetime conditions or from the ringdown to the dial method of toll operation.

The problem was therefore to determine the average delay in securing a trunk with various sizes of groups at various levels of usage with a view to:

Stating that portion of the overall speed of service which results from inability to secure a circuit, and

Constructing engineering tables based on a preselected constant circuit delay or trunk speed of service.

Arrangements were made with several Associated Companies to furnish data for this purpose which would show:

1. The average overall speed of service on different sizes of groups under various conditions of loading.
2. The minimum average overall speed interval on these same groups at times when circuit provision was not a factor, i.e., when NC conditions were not encountered.

The speeds obtained in Item 2 were subtracted from those obtained in Item 1, the difference representing that portion of the overall speed which can be attributed to circuit delay, or the trunk speed.

In order to determine these trunk speeds it was necessary to obtain from several sources as much data as possible of the following nature:

Per cent. circuit usage, by hours, as derived from group busy timing registers on selected groups of various sizes. Hours during which the traffic over a group was handled subject to posted delay were disregarded.

The number of originating terminal calls handled over the groups during the hours corresponding to the usage data and the average speed of service on these calls. The call and speed of service data were summarized first to include all calls and then separately for calls not encountering NC. Correction was made for transfer of tickets to point-to-point positions by subtracting from the speed shown on each such ticket an interval representing the average length of time required to send a ticket to point-to-point positions in the office in which the data were obtained, provided the transfer time was included in the overall speed interval. This interval of transfer time is not properly chargeable as part of the trunk speed.

These data were obtained for trunk groups of various sizes ranging from one up to eighteen trunks. To secure a comparable amount of data for the smaller groups which handle fewer calls, it was necessary to include more of the smaller groups or to continue the record for a longer period of time on such groups.

The data for all hours of the day or evening were useful because as the volume of traffic recedes from the busy hour the data are typical of the busy hour condition of other groups engineered on a more liberal basis. The very light hours also show the minimum speed interval which can be obtained when lack of an available circuit is not a factor.

Five Associated Companies obtained data at eleven toll offices on 112 intertoll groups having 561 trunks. Approximately 17,000 calls (occurring during hours when the groups were at least 40% busy) were included.

The data were summarized by size of group and by circuit usage. Separate counts were maintained for groups with and without alternate routes and for person and station traffic. The following tabulation shows the type of data available for each point, i.e., each size of group at each level of usage.

ONE TRUNK—WITH ALTERNATE ROUTE—61–70% USE

% Use	Type	All Calls			Calls Not Encountering NC		
		Minutes	Calls	Speed	Minutes	Calls	Speed
61–65	Station	111	23		39	17	
	Person	109	38		68	32	
66–70	Station	103	25		33	21	
	Person	117	28		54	25	
Total		440	114	3.86	194	95	2.04

Average Speed—All Calls ..... 3.86 Mins.
Average Speed—NC not Encountered ..... 2.04 Mins.

Average Delay Due to NC (Trunk Speed) ..... 1.82 Mins.

The results were plotted for each level of usage by steps of 10% as shown in Fig. 1, using a 3-point moving average to smooth out the deviations and to establish a more definite trend. Each of these curves was then redrawn in relation to the others and combined results are shown on Fig. 2.

The delay intervals indicated in Fig. 2 represent the total delay which resulted from the fact that there was no circuit available when the operator was first ready to make use of one. It includes not only the time spent in waiting for a circuit to become idle but also the time required for the operator herself to return to that call if she had engaged in some other work in the meantime. If the operator is not free to utilize the circuit as soon as it becomes available some other operator may use it for another, later call. The subsequent call is then delayed less than the average, or not at all, but the original call is delayed longer than the average. While the delays experienced by individual calls may vary considerably from the average, the data have been treated in terms of averages for engineering purposes.

MATHEMATICAL FORMULAE

The summarized data were referred to the different mathematical expressions frequently applied to trunking problems, such as the Poisson, Erlang "B" and Erlang "C" formulae. It was found that the observed average NC delays were considerably shorter than the theoretical average delays in those formulae which make allowance for variable holding times, such as would be encountered in local trunking where the average trunk use is short and the deviations from average on a percentage basis are apt to be appreciable.

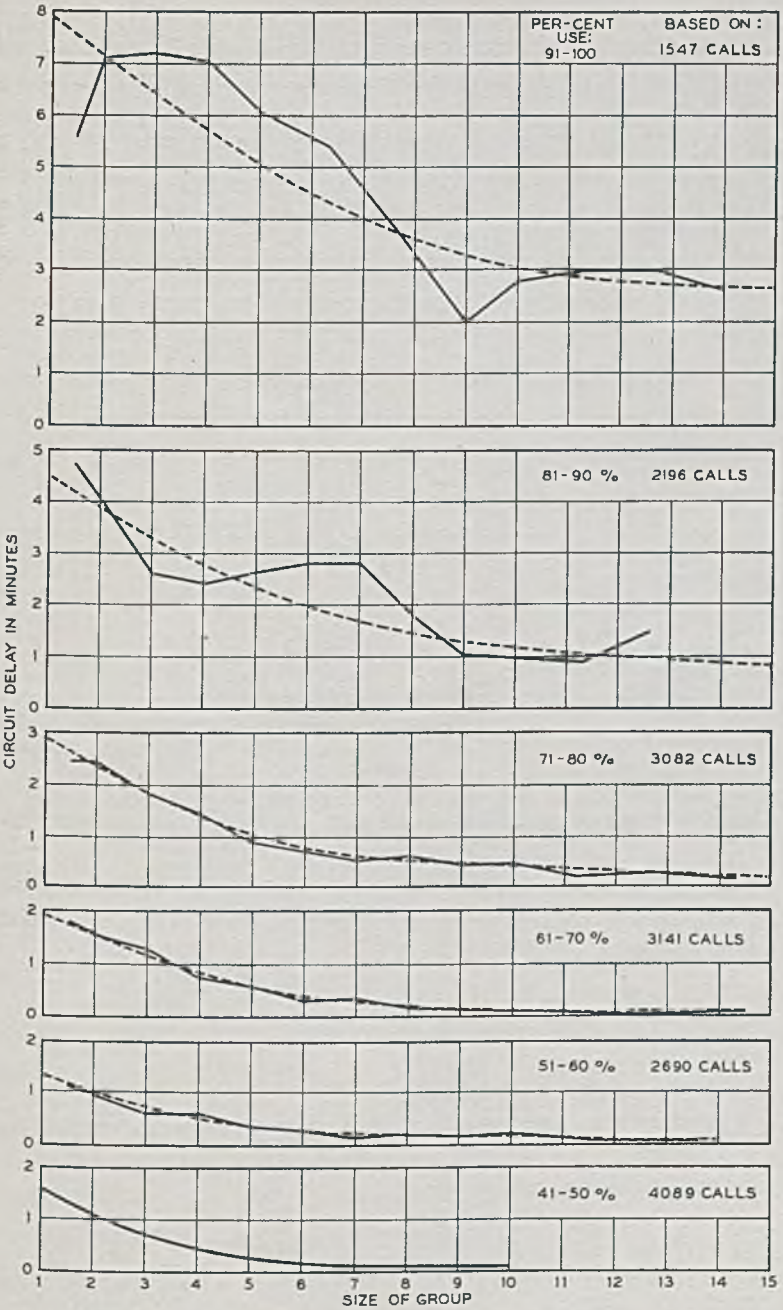


Fig. 1—Average circuit delay on all calls (with alternate routes where authorized). Circuit delay = average speed on all calls minus average speed on calls which did not encounter NC. Based on 3-point moving average.

However, further reference to mathematical studies of telephone traffic indicated that a delay theory based on constant holding times, first developed

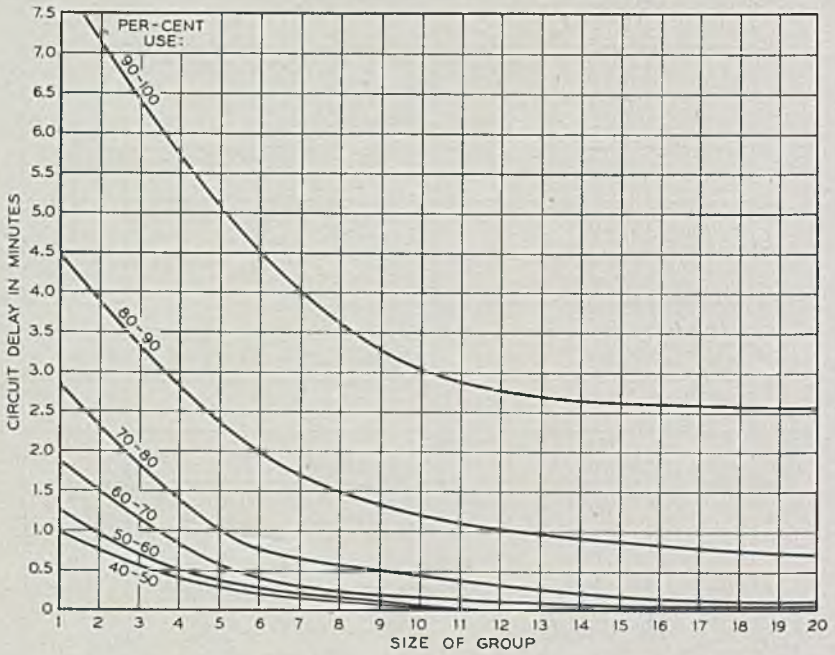


Fig. 2—Average circuit delay on all calls (with alternate routes where authorized). Circuit delay = average speed on all calls minus average speed on calls which did not encounter NC. Combined curves based on 16,745 calls.

by Felix Pollaczek in Germany and amplified by C. D. Crommelin in England, closely approximated the empirical data. This formula[1] is:

$$d = \sum_{w=1}^{\infty} e^{-aw} \left[\sum_{u=wc}^{\infty} \frac{(aw)^u}{\lfloor u} - \frac{c}{a} \sum_{u=wc+1}^{\infty} \frac{(aw)^u}{\lfloor u} \right]$$

In which

d = average delay on all calls

a = average simultaneous calls submitted to a group of c trunks (trunk hours)

c = number of trunks in group

It may be quite reasonable that the Pollaczek constant holding time formula should better represent toll delays than an exponential holding time formula since the toll charge and perhaps other factors ordinarily cause these calls

[1] C. D. Crommelin, "Delay Probability Formulae," *P.O.E.E. Journal*, Jan. 1934, p. 266

to exhibit considerably less percentage deviation from their average than is found in the exponential distribution.

In order to compare the empirical data with the Pollaczek formula it was necessary to assume a holding time per attempt since the formula expresses the delays in terms of the average interval of use whenever the circuit is in use. The average holding time as reported by the companies for the groups included in the study was 8.3 minutes per message. Recent data show 1.42 attempts per call disposed of. Relating this figure to the 8.3 minutes results in an average holding time per circuit use of 5.85 minutes. Six

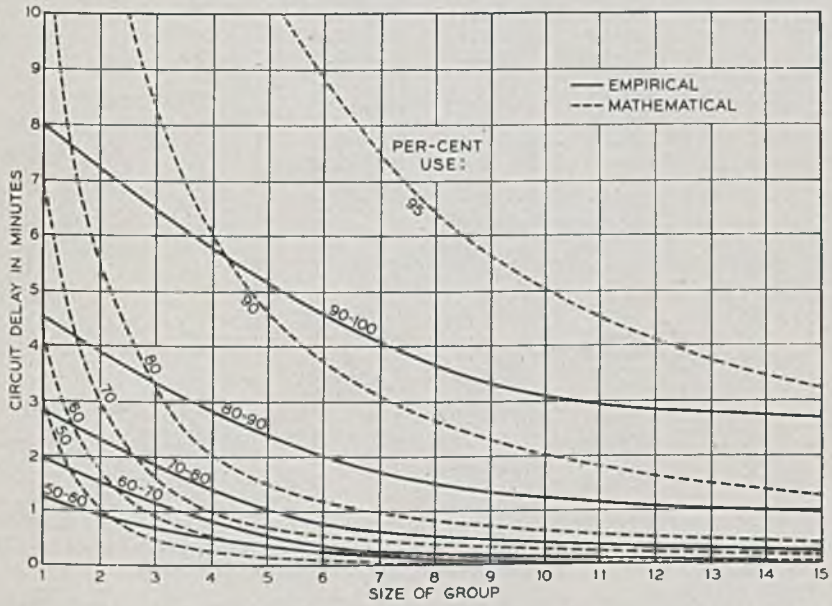


Fig. 3—Average circuit delay on all calls.

Comparison of empirical data (with alternate routes) and Pollaczek formula (no alternate routes) using a 6 minute holding time.

minutes is therefore well within the limits of accuracy required for this purpose.

Curves were prepared from the Pollaczek formula for various levels of usage at a 6-minute holding time per attempt. The corresponding curves derived from the empirical data were then superimposed for comparative purposes as shown in Fig. 3. It will be seen that the shape and levels of the curves are very similar *except* for the smaller groups on which the effect of alternate routes tends to reduce the average length of delay.

As a further check on the validity of the Pollaczek formula, the delay data from the Cleveland (1929–1930) study were expressed in terms of hold-

ing times for different group sizes at different levels of usage and compared with delay intervals developed from the formula. This comparison is

COMPARISON OF 1945 STUDY WITH CLEVELAND STUDY OF 1929–30
Based on 3.5 Min. HT per Circuit Attempt
or 5.25 Min. HT per Message

No. of Trunks	% Use	Minutes Delay		% of H.T.	
		Cleveland	1945	Cleveland	1945
1	55–60	.8	.8	21	22
1	65–70	1.1	1.2	30	35
1	75–80	1.5	1.9	43	53
1	85–90	2.7	3.0	76	86
3	55–60	.3	.5	09	13
3	65–70	.5	.8	14	22
3	75–80	.9	1.3	24	36
3	85–90	1.8	2.4	51	68
6	55–60	.1	.1	03	04
6	65–70	.3	.3	09	08
6	75–80	.6	.6	16	16
6	85–90	1.3	1.7	36	48
10	55–60	.1	—	02	01
10	65–70	.3	.1	07	03
10	75–80	.4	.3	12	08
10	85–90	1.0	1.0	28	28
14	55–60	.1		01	
14	65–70	.2	.1	05	02
14	75–80	.3	.2	09	05
14	85–90	.8	.7	23	19

The minutes of circuit delay shown above for the Cleveland study are derived by subtracting the minimum speed of 1.65 minutes from the actual overall speed for the various sizes of groups and levels of usage. The comparable 1945 figures are taken from Fig. 5.

It will be noted that the principal differences occur on the smaller groups at the higher levels of usage. This is undoubtedly due to the fact that the alternate routes are more heavily loaded today than they were in 1929–30 and therefore are less helpful in absorbing the overflow from the first route.

The holding time used in this comparison as probably typical of 1929–30 is derived as follows:

Conversation time	3.00 minutes
Operating time	2.25 minutes
	5.25 minutes
No. of Circuit Uses per Message	1.50
HT per Circuit Use (5.25 ÷ 1.50)	3.50

Fig. 4

shown in Fig. 4. It will be seen that there is substantial agreement between the two sets of delay factors, such differences as there are being explained in the notes on that figure.

The delay in securing a circuit varies directly with the holding time, i.e., a call waiting for a circuit will be delayed twice as long if the group is handling 10-minute calls as would be the case with 5-minute calls. This is best illustrated by one call awaiting access to a single circuit group. The new call may appear at any time during the progress of the existing call, but the average delay for many such delayed calls will be one-half the holding time of the existing call. If the existing call uses the circuit for five minutes, the new call will wait $5 \div 2 = 2.5$ minutes. If the existing call uses the circuit for ten minutes, the new call will wait five minutes. It should be noted that the average delay depends upon the average length of time that the circuit is in use each time that it is used, in other words, the holding time per circuit attempt.

The Cleveland study did not go into this phase in detail, the statement being made that the effect of holding time on speed of service "is slight." This is so when considering the *overall* speed of service, with which that study was primarily concerned, because of the weight of operating and subscriber time intervals. Reference to that study shows that the circuit delay increased about in proportion to the holding time when a minimum operating and subscriber time interval is subtracted from the overall speed, as follows:

	Holding Time		
	5′	7.5′	10′
Total Overall Speed	2.2	2.6	3.0
Minimum Operating and Subscriber Time Interval	1.6	1.6	1.6
Average NC Delay-Trunk Speed	.6	1.0	1.4

Combination of Mathematical and Empirical Methods

Because of the apparent close agreement between the Pollaczek delay formula and two representative large samples of actual NC delay data taken at different periods and under widely different conditions, 1930 and 1945, the Pollaczek formula can be used for deriving expressions of the average duration of NC with intertoll trunk operation without alternate routes. The effect of alternate routes in reducing the duration of NC can be shown with sufficient accuracy for practical needs from the empirical data. The curves shown in Fig. 5 were constructed on this basis. For large groups the formula has been extended in Fig. 6.

It is advantageous to have available an acceptable mathematical formula for expressing the relationship between the loads carried by the trunk groups and the length of time that the average call will be delayed because of NC conditions, i.e., the part played by trunk provision in the overall speed of service. With such a formula results can be predicted for any given set of

assumptions without going through the burdensome process of accumulating actual data in reliable volume and with useful frequency. The formula

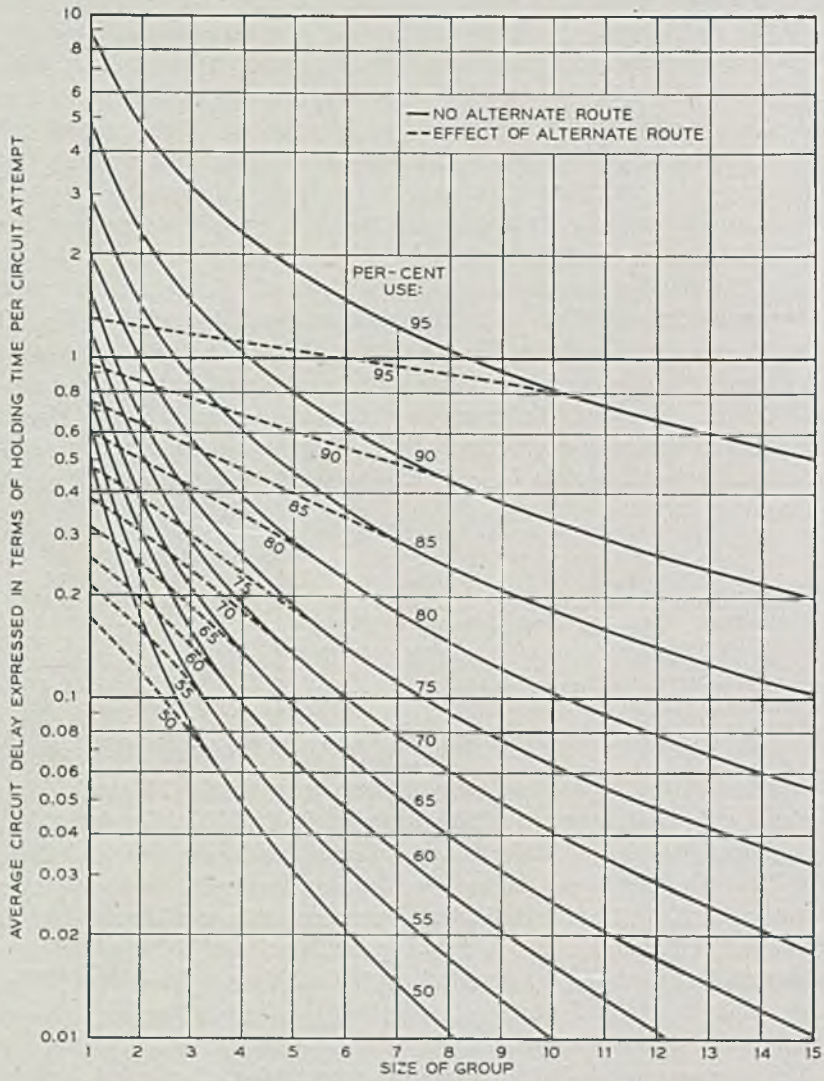


Fig. 5—Average circuit delay on all calls expressed in terms of the holding time per circuit attempt.

also provides a convenient means of checking the adequacy of the trunk facilities in any unusual situations which may be observed from time to time.

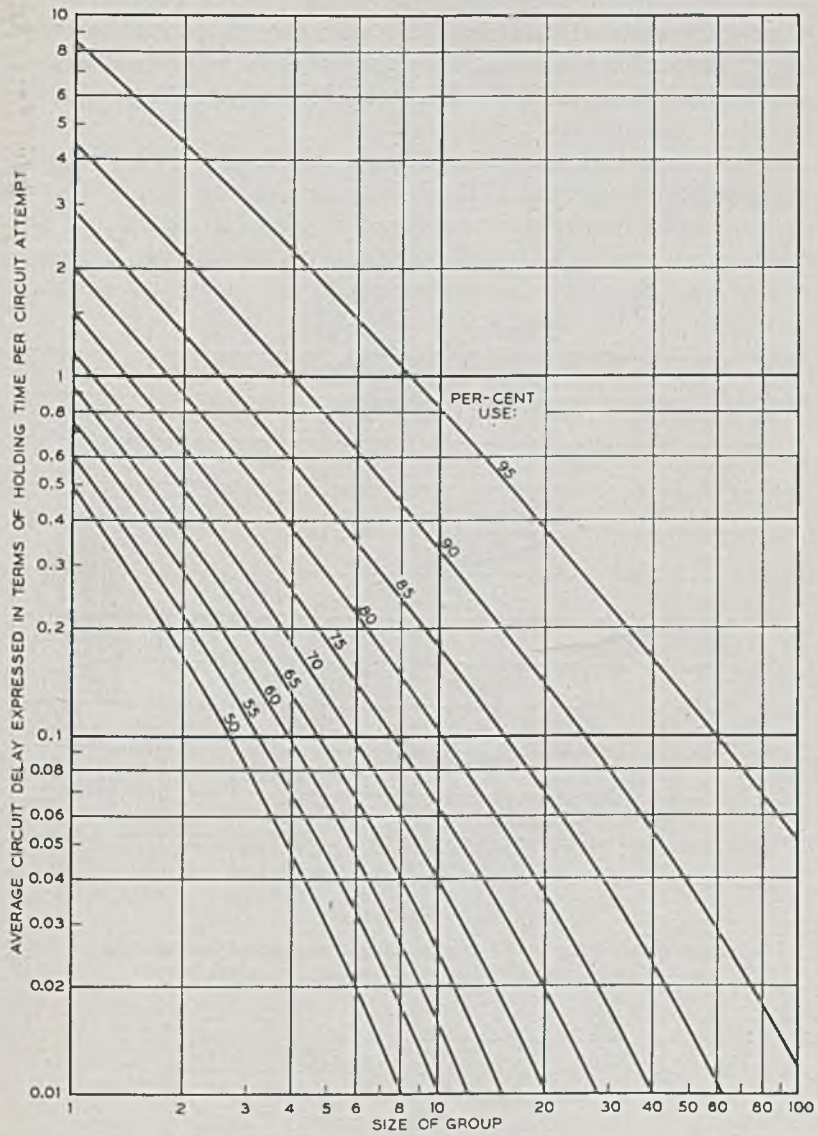


Fig. 6—Average circuit delay on all calls expressed in terms of the holding time per circuit attempt.
Based on Pollaczek formula.

In the case of the alternate route effect, where a variable is introduced which the formula does not encompass, it may be necessary later to recheck

this factor by observed data should the conditions governing the use of alternate routes change substantially or should the actual results at some future time on groups provided with alternates be found to differ from those currently predicted.

Delays are expressed in Figs. 5 and 6 as a percentage of the holding time per *average use* of the trunk. The holding time factors used in intertoll trunk engineering generally are expressed in terms of the holding time *per message.* Therefore, in order to use the curve conveniently it is necessary to reduce the holding time per message to the holding time per trunk

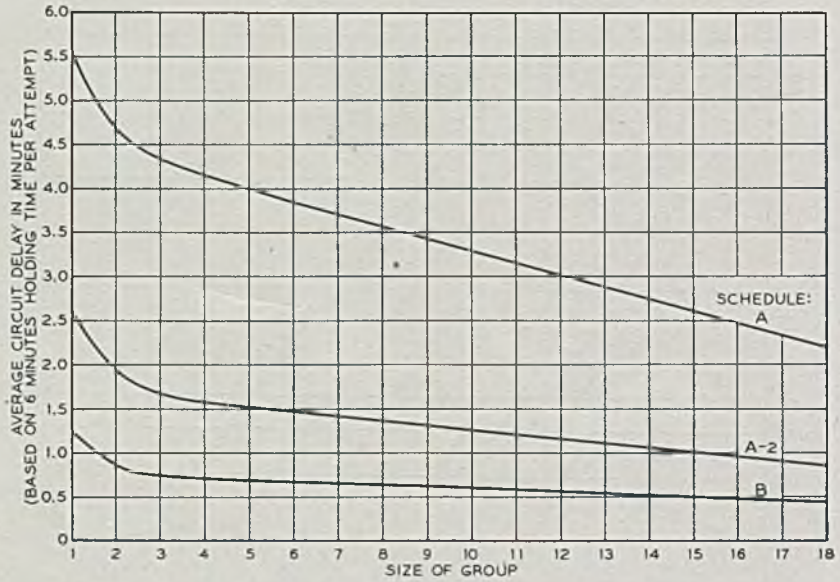


Fig. 7—Average circuit delay on all calls resulting from present intertoll trunk capacity schedules (without alternate routes)—Using Pollaczek formula.

use or attempt. This can be done with sufficient accuracy for this purpose by a ratio of 1.5 outward attempts per call disposed of.

Description of Engineering Tables

Having determined the average circuit delay, as previously described, the delays which result from the present Intertoll Trunk Capacity Tables A, A2, and B can be determined by referring the percentages of use on these tables to the curves in Figs. 5 and 6. Figure 7 indicates that each of these existing tables results in variable delays depending upon the size of the group.

The curves in Figs. 5 and 6 can also be used to construct new capacity tables which should produce a relatively uniform delay for any size of group. If we select an average delay of three-tenths of a holding time as an example

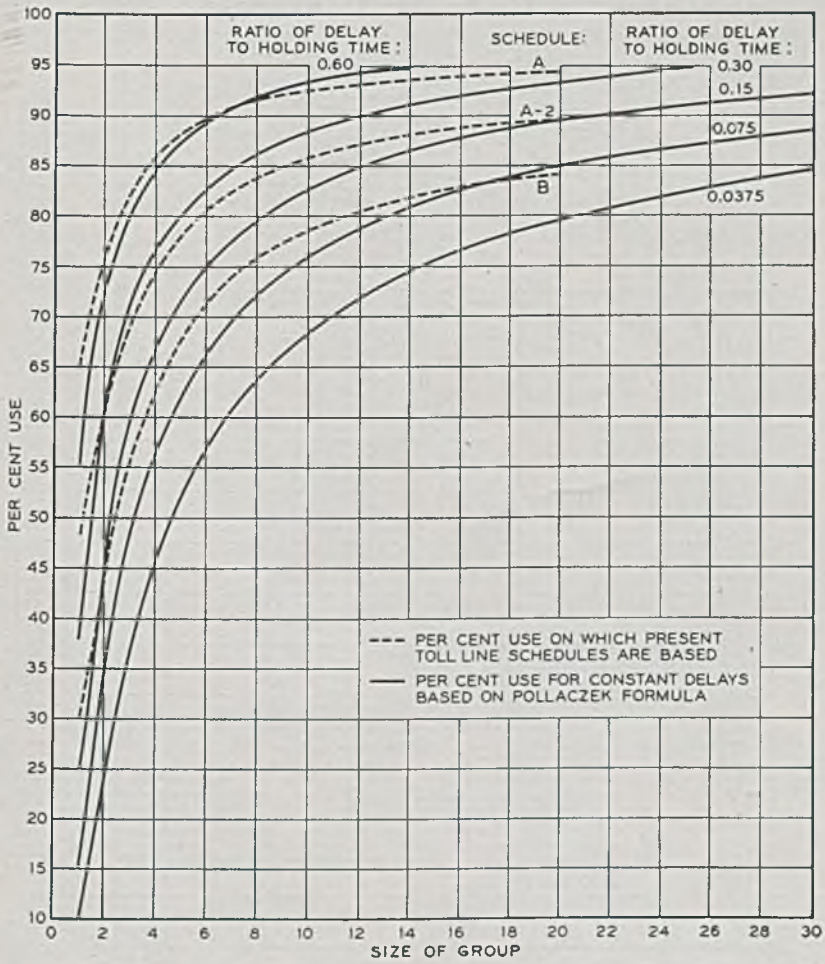


Fig. 8—Per cent. use with present intertoll trunk schedules and for constant delays based on Pollaczek formula.

and follow the .3 line across these curves, we see that a two-trunk group can be kept busy 60.8 per cent. of the time; a three-trunk group can be in use about 71.1 per cent. of the time; about 76.7 per cent. for four trunks, etc.

Figure 8 shows the per cent. usage for groups of from one to thirty trunks

which result in average delays of .0375, .0750, .15, .30, and .60 of a holding time. The usage obtained from present capacity tables is also shown for comparison. From this figure it will be seen that five new tables based on these average delays will adequately cover the field encompassed by those now in use. The usage curves for these five selected delay intervals were

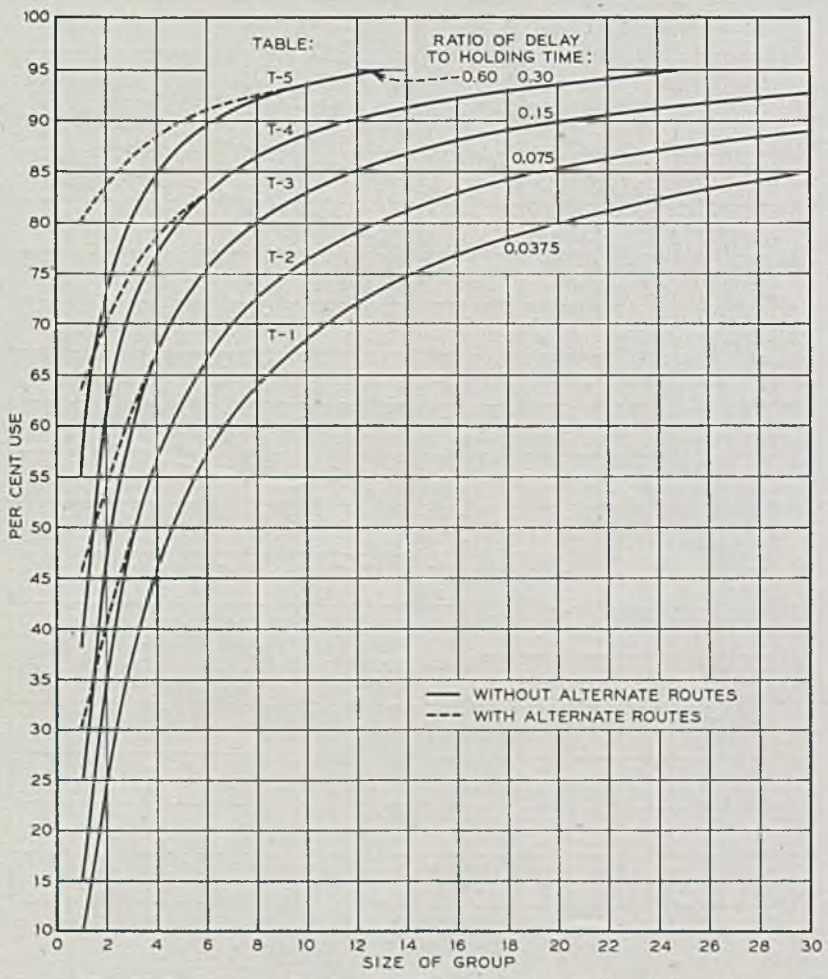


Fig. 9—Per cent. use for constant delays based on Pollaczek formula.

redrawn in Fig. 9, on which the effect of alternate routes is also shown. Capacity Tables T-1 to T-5 (Fig. 10) were constructed from this curve, Table T-1 representing the shortest (.0375) delay and Table T-5 the longest (.60) delay.

For situations where it may be necessary for service reasons to provide trunks on a basis more liberal than Table T-1, the interlocal trunk capacity

tables should be used. These latter tables, constructed from the Poisson formula, express a service relationship in terms of the percentage of calls *encountering NC* rather than the average *duration of NC*. When trunks are provided on a basis as liberal as that implied by the interlocal tables the frequency of NC is deemed to be the more important service consideration because the duration of NC is so short as to be of lesser consequence.

Since the Pollaczek formula expresses the delay as a percentage of the holding time per *circuit attempt* but the intertoll trunk engineer is accustomed to dealing with holding times per *message* the delays indicated for each of Tables T-1 to T-5 have been expressed in the latter terms, using a 1.5 ratio of attempts per message. This is consistent with previous computations. With Table T-4 and a five-minute message holding time, for example, we have:

5-minute HT per message ÷ 1.5 = 3.33 minutes HT per attempt
3.33 minutes x .30 delay = 1.0 minute delay

This one minute delay shown in the heading of Table T-4 is, therefore, actually the delay per attempt when the holding time per message is five minutes. The delay per attempt is the important criterion because the "speed of toll service" as quoted from service observations is generally the speed of the first attempt and the two are, therefore, comparable.[2]

As the usage approaches 100 per cent. there may be an indeterminate backed-up potential demand and the normal relationship between service and loading no longer holds true. From other data assembled for this purpose, it appears that about 96–97 per cent. represent the practical upper limit of usage, beyond which trunk speeds can not be accurately predicted. For practical reasons, therefore, group capacities have not been computed for percentages of use higher than 97 even though, from a theoretical viewpoint, the curves derived from the Pollaczek formula would permit extending the usage virtually to 100 per cent. This would also apply to Tables T-1 and T-2 if they were extended above 75 trunks.

Relation of Circuit Delays in Busy Hour to Total Day

Up to this point the trunk speed of service has been discussed in terms of the busiest hour. However, the overall speed of service is generally quoted in terms of the total day. Therefore, one additional step is necessary, namely to determine the relationship of the trunk speed in the busy hour to that of the total day.

[2] There is one exception to the statement that the "speed of toll service" is the speed on the first attempt. That is the case of a built-up connection where an NC condition is encountered at an intermediate office which persists so long that the first circuit is released. When the connection is established it is at least the second attempt. The full time interval during which the ticket is held awaiting completion is included in the speed quoted on that call. Similar intervals were also included in the empirical data for this study and the results are, therefore, comparable in this case also.

Fig. 10—Intertoll trunk capacity tables

Trunk Speed when average H.T. per Msg. is: 4–6 Mins. 7–9 Mins 10–12 Mins.	Table T-1 .13 Minutes .20 Minutes .28 Minutes				Table T-2 .25 Minutes .40 Minutes .55 Minutes				Table T-3 .5 Minutes .8 Minutes 1.1 Minutes				Table T-4 1.0 Minutes 1.6 Minutes 2.2 Minutes				Table T-5 2.0 Minutes 3.2 Minutes 4.4 Minutes			
	No Alt. Route		Alt. Route		No Alt. Route		Alt. Route		No Alt. Route		Alt. Route		No Alt. Route		Alt. Route		No Alt. Route		Alt. Route	
No. of Trunks	% Use	Minutes Capacity	% Use	Minutes Capacity	% Use	Minutes Capacity	% Use	Minutes Capacity	% Use	Minutes Capacity	% Use	Minutes Capacity	% Use	Minutes Capacity	% Use	Minutes Capacity	% Use	Minutes Capacity	% Use	Minutes Capacity
1	10.0	6	—	—	15.0	9	30.0	18	25.0	15	50.0	30	38.3	23	63.4	38	55.0	33	80.0	48
2	25.0	30	—	—	35.0	42	40.0	48	48.3	58	53.3	64	60.8	73	70.0	84	73.3	88	83.4	100
3	37.2	67			48.9	88			60.0	108	61.1	110	71.1	128	75.0	135	80.6	145	86.6	156
4	46.3	111			56.7	136			67.5	162			76.7	184	78.3	188	85.0	204	88.4	212
5	52.7	158			62.3	187			72.3	217			80.3	241	80.7	242	87.6	263	90.0	270
6	57.5	207			66.7	240			75.4	272			83.1	299			89.5	322	91.1	328
7	61.2	257			70.0	294			78.1	328			85.0	357			91.0	382	91.9	386
8	64.2	308			72.7	349			80.2	385			86.5	415			92.1	442	92.5	444
9	66.7	360			74.8	404			81.9	442			87.6	473			93.0	502		
10	68.8	413			76.5	459			83.1	499			88.7	532			93.6	561		
11	70.6	466			78.0	515			84.2	556			89.6	591			94.1	621		
12	72.2	520			79.3	571			85.3	614			90.3	650			94.5	680		
13	73.6	574			80.4	627			86.2	672			90.9	709			94.9	740		
14	74.8	628			81.4	684			86.9	730			91.4	768			95.2	800		
15	75.9	683			82.3	741			87.6	788			91.9	827			95.5	860		
16	76.9	738			83.1	798			88.1	846			92.3	886			95.8	926		
17	77.8	793			83.9	855			88.6	904			92.6	945			96.1	980		
18	78.6	848			84.5	912			89.0	962			93.0	1,005			96.3	1,040		
19	79.3	904			85.0	969			89.5	1,021			93.4	1,065			96.5	1,100		
20	80.0	960			85.5	1,026			90.0	1,080			93.7	1,125			96.6	1,160		
21	80.7	1,017			86.0	1,084			90.4	1,139			94.0	1,185			96.8	1,220		
22	81.3	1,074			86.5	1,142			90.8	1,198			94.3	1,245			97.0	1,280		
23	82.0	1,131			87.0	1,200			91.1	1,257			94.5	1,305						
24	82.5	1,188			87.4	1,258			91.4	1,316			94.8	1,365						
25	83.0	1,245			87.7	1,316			91.7	1,375			95.0	1,425						
30	85.0	1,530			89.7	1,606			92.8	1,670			95.9	1,725						
35	86.5	1,816			90.5	1,901			93.7	1,968			96.4	2,025						
40	87.8	2,106			91.5	2,196			94.6	2,268			96.8	2,325						
45	88.8	2,398			92.3	2,491			95.1	2,568										
50	89.7	2,692			92.9	2,786			95.6	2,868										
55	90.5	2,987			93.4	3,081			96.0	3,168										
60	91.2	3,282			93.8	3,376			96.3	3,468										
65	91.7	3,577			94.3	3,675			96.6	3,768										
70	92.2	3,872			94.7	3,975			96.8	4,068										
75	92.7	4,170			95.0	4,275			97.0	4,368										

Note:—The Trunk Speeds of Service indicated at the heading of each Table represent the average busy hour delay in securing a trunk on each attempt, when the holding times per message are as shown.

To develop this relationship it was necessary first to determine a typical distribution of traffic throughout the day, i.e., the ratio of the busy hour to each of the other hours. Actual delays experienced on a particular group may deviate somewhat from those developed herein to the extent that the actual distribution varies from the typical distribution.

The probable total day circuit delays are derived from Figs. 5 and 6 and from a typical distribution of traffic based on a five-day record on each of 20 groups in Ohio and 24 groups in Illinois.

Hours	No. of Calls	% of Total Traffic in 14-hour day (Used as weighting factor)
1 (Busy Hour)	100	12.80
1	90	11.53
1	90	11.53
1	80	10.25
1	80	10.25
1	75	9.62
1	70	8.98
1	65	8.35
1	55	7.07
5	75	9.62
14	780	100.0
10	20	
24	800	

It will be noted that the above distribution shows a busy hour which is 12.5% of the total 24-hour day but that the weighting factors are based on a total day of 14 hours. This corresponds to the normal service observing period so that the results will be comparable with the overall total day speeds obtained from service observations.

The total day delays for Tables T-1 to T-5 for each size of group were computed as illustrated in the following sample calculation:

TABLE T-5—FIVE TRUNKS

Hours	% of BH	% Use in BH (Table T-5)	% Use Each Hour	Circuit Delay (Fig. 6)	Weight Factor	Weighted Delay
1	at 100	× 87.6 =	87.6	.60	× 12.80% =	.0768
1	90		78.8	.26	11.53	.0300
1	90		78.8	.26	11.53	.0300
1	80		70.2	.13	10.25	.0133
1	80		70.2	.13	10.25	.0133
1	75		65.7	.10	9.62	.0096
1	70		61.3	.07	8.98	.0063
1	65		56.9	.06	8.35	.0050
1	55		48.2	.03	7.07	.0021
5				None	9.62	.0000
14					100.00	.1864

The figure .1864 derived above represents the average delay expressed as a per cent. of the holding time per circuit use (attempt). The last step, therefore, is to relate this figure to the message holding times contemplated in Table T-5, as follows:

Holding Time Per Message		Ratio of Attempts Per Message		Holding Time Per Attempt		% of H. T. Delayed		Average Delay
5 min.	÷	1.50	=	3.33	×	.1864	=	.62 min.
8 min.	÷	1.50	=	5.33	×	.1864	=	1.00 min.
11 min.	÷	1.50	=	7.33	×	.1864	=	1.37 min.

The results of similar calculations are summarized in Fig. 11.

As pointed out previously, actual delays experienced will deviate from those shown in Fig. 11 to the extent that the actual hourly distribution varies from that which has been used. If a particular group has a higher per cent busy hour the total day delays should be less than indicated. Conversely, if the group has a lower per cent busy hour the delays should be greater. However, the variations in distribution which are most likely to be encountered in practice will not have any marked effect on the total day delays except possibly for groups of about five trunks or less which are loaded as heavily as indicated in Tables T-4 and T-5.

Per Cent. NC Encountered

The per cent. of calls delayed by NC as noted by the operators on the tickets analyzed for this study was plotted for each level of usage in steps of 10% as shown in Fig. 12, using a 3 point moving average. Each of these curves was then redrawn in relation to the others and the combined results are in Fig. 13. The results are very similar to those obtained in the Cleveland study of 1929–30, as shown in the same figure.

It should be noted that there is a difference between the per cent. calls encountering NC and the per cent. NC existing. In the present study no data were obtained to indicate NC existing. However, the Cleveland study included such data which showed that the NC existing follows the Erlang "B" formula (Fig. 14) in this respect. The individual points in Fig. 14 were derived by selecting from the Erlang "B" table of overflows the point at which the call-seconds carried (offered minus overflow) gave the desired level of usage.

The difference between NC existing and NC encountered may be due to several factors, some of which are suggested below:

1. Effect of alternate routes.

No. of Trunks	Table T-1			Table T-2			Table T-3			Table T-4			Table T-5		
	Holding Time—In Minutes														
	5	8	11	5	8	11	5	8	11	5	8	11	5	8	11
	Trunk Speed—Busy Hour—In Minutes														
	.13	.20	.28	.25	.40	.55	.5	.8	1.1	1.0	1.6	2.2	2.0	3.2	4.4
	Trunk Speed—Total Day—In Minutes														
1	.09	.14	.20	.19	.31	.43	.38	.60	.83	.67	1.07	1.47	1.20	1.92	2.64
2	.08	.13	.18	.16	.25	.34	.29	.47	.65	.52	.83	1.14	.92	1.48	2.03
3	.07	.11	.16	.14	.22	.30	.25	.39	.54	.43	.69	.95	.77	1.24	1.70
4	.06	.10	.14	.12	.19	.26	.22	.35	.48	.39	.62	.85	.68	1.09	1.50
5	.06	.09	.13	.11	.18	.24	.20	.32	.44	.35	.56	.77	.62	1.00	1.37
6	.05	.09	.12	.10	.17	.23	.18	.29	.40	.33	.52	.72	.57	.91	1.25
7	.05	.08	.11	.10	.15	.21	.17	.28	.38	.31	.49	.67	.53	.85	1.17
8	.05	.08	.11	.09	.14	.20	.16	.26	.36	.29	.46	.63	.50	.80	1.10
9	.05	.07	.10	.09	.14	.19	.16	.25	.34	.27	.44	.60	.48	.76	1.05
10	.04	.07	.10	.08	.13	.18	.15	.24	.33	.26	.42	.58	.46	.73	1.00
11	.04	.07	.09	.08	.13	.18	.14	.23	.32	.25	.41	.56	.44	.70	.96
12	.04	.07	.09	.08	.12	.17	.14	.22	.31	.24	.39	.54	.43	.68	.94
13	.04	.06	.09	.08	.12	.17	.13	.21	.29	.23	.37	.52	.41	.66	.91
14	.04	.06	.09	.07	.12	.16	.13	.21	.29	.23	.36	.50	.40	.64	.88
15	.04	.06	.08	.07	.11	.15	.13	.20	.28	.22	.35	.48	.39	.63	.86
16	.04	.06	.08	.07	.11	.15	.12	.20	.27	.22	.35	.48	.38	.61	.84
17	.04	.06	.08	.07	.11	.15	.12	.19	.26	.21	.34	.47	.38	.60	.83
18	.04	.06	.08	.07	.11	.15	.12	.19	.26	.21	.33	.45	.37	.59	.81
19	.03	.05	.07	.06	.10	.14	.11	.18	.25	.20	.33	.45	.37	.59	.81
20	.03	.05	.07	.06	.10	.14	.11	.18	.24	.20	.32	.44	.36	.58	.79
21	.03	.05	.07	.06	.10	.14	.11	.18	.24	.20	.32	.43	.35	.57	.78
22	.03	.05	.07	.06	.10	.13	.11	.17	.24	.19	.31	.43	.35	.56	.77
23	.03	.05	.07	.06	.09	.13	.11	.17	.23	.19	.31	.42	.35	.55	.76
24	.03	.05	.07	.06	.09	.13	.10	.17	.23	.19	.31	.42	.34	.55	.76
25	.03	.05	.07	.06	.09	.13	.10	.17	.23	.19	.30	.41	.34	.54	.75
30	.03	.05	.06	.05	.09	.12	.10	.15	.21	.18	.28	.39	.33	.52	.72
35	.03	.04	.06	.05	.08	.11	.09	.15	.21	.17	.27	.37	.32	.51	.70
40	.03	.04	.06	.05	.08	.10	.09	.14	.20	.16	.26	.36	.31	.50	.68
45	.03	.04	.06	.05	.07	.10	.09	.14	.19	.16	.26	.35	.30	.49	.67
50	.02	.04	.05	.05	.07	.10	.09	.14	.19	.16	.25	.34	.30	.48	.66
55	.02	.04	.05	.04	.07	.10	.08	.14	.19						
60	.02	.04	.05	.04	.07	.10	.08	.13	.18						
65	.02	.04	.05	.04	.07	.09	.08	.13	.18						
70	.02	.04	.05	.04	.07	.09	.08	.13	.18						
75	.02	.04	.05	.04	.06	.09	.08	.13	.18						

Fig. 11—Relation of trunks speeds in busy hour to total day

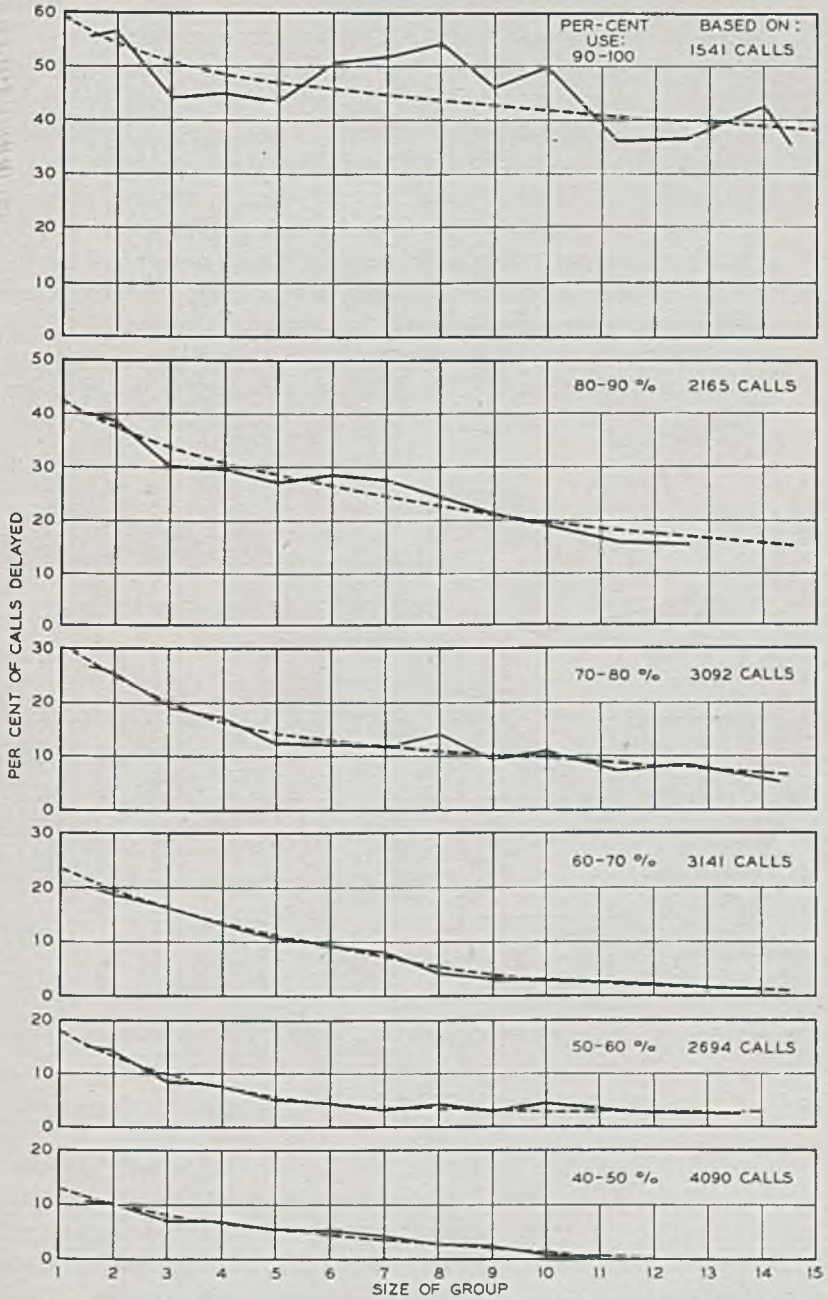


Fig. 12—Per cent. of calls delayed by NC as noted by the operators (with alternate routes where authorized).

2. The operators did not make NC notations on the tickets in a certain proportion of the cases where NC was actually encountered.
3. Because the operator does not test for an idle trunk with machine-like finality there are probably many cases where she does not consider that an NC condition existed if it was of such short duration that it did not materially affect her ability to secure a trunk.
4. The possibility of some "limited sources" effect in the case of small groups. The number of people in Newark who have occasion to call York, Pa. must be relatively small since only one trunk is provided. Therefore, while the trunk is in use on one call there is less likelihood that a second call will be originated than would be the case if there were a greater community of interest between the two places. Although the NC condition exists during the period of one call, no NC is encountered unless there is a second and overlapping call.

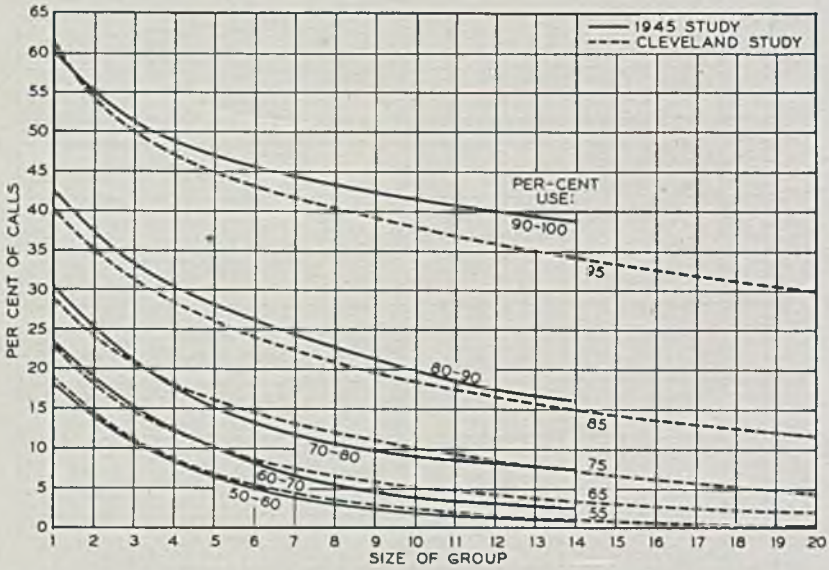


Fig. 13—Per cent. of calls delayed by NC (with alternate routes where authorized).

Because of the importance of Items 2 and 3 above, both of which involve the testing of trunks by operators, the empirical data should be used as representative of per cent. NC encountered with ringdown operation (operator testing) and the Erlang "B" per cent. NC existing should be used as representative of intertoll dialing conditions (mechanical testing).

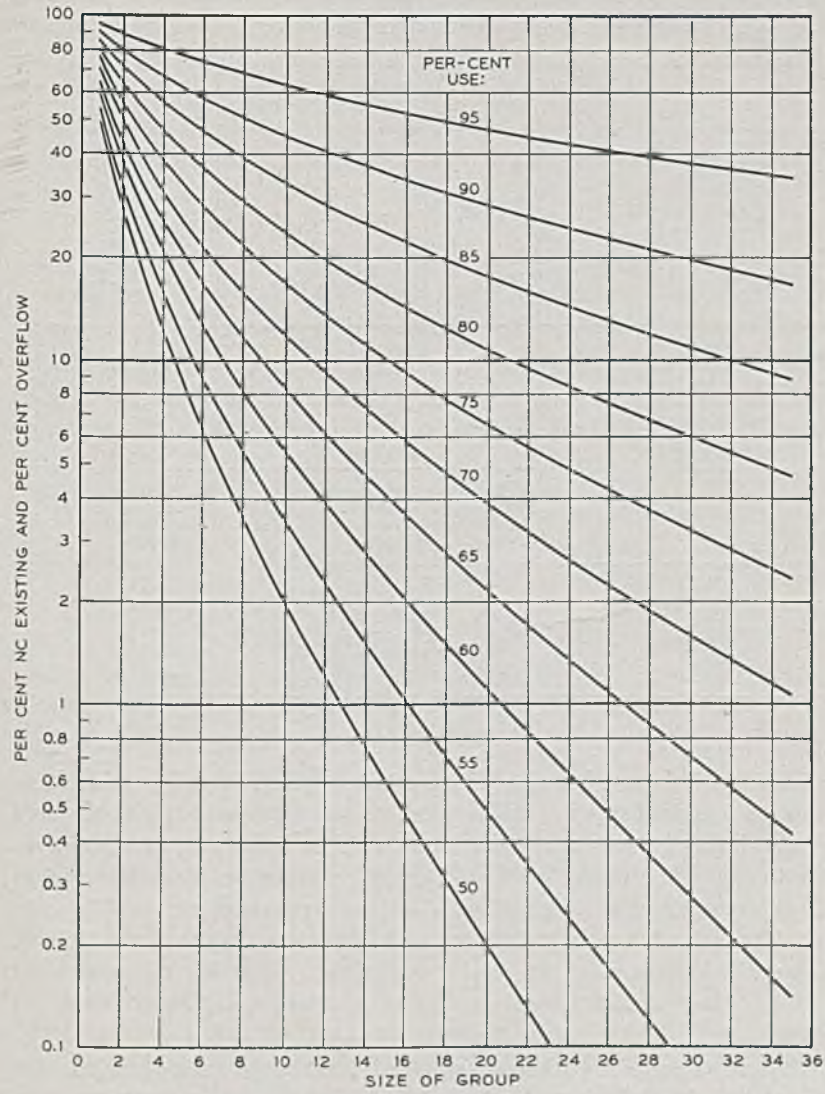


Fig. 14—Per cent. NC existing and per cent. overflow at various levels of usage. Based on Erlang "B" formula.

It is interesting to note that each of the proposed Capacity Tables T-1 to T-5 results in a fairly uniform percentage of NC encountered by the operators as determined from the empirical data (Fig. 13) and also a fairly uniform

percentage of NC existing as determined from the Erlang "B" formula (Fig. 14). Using Table T-1 as an example, the NC conditions are as follows:

% Use	No. of Trunks Table T-1 from Fig. 9	%NC Encountered from Fig. 13	% NC Existing from Fig. 14
50	4.6	6.5	10.0
55	5.5	6.0	10.0
60	6.7	6.0	10.7
65	8.4	6.5	10.9
70	10.7	7.0	11.0
75	14.1	7.5	11.0
80	20.0	7.0	10.7
85	30.0		10.8

Similar comparisons made with the other capacity tables indicate similar uniformity in the most frequently used portion of the tables, i.e., up to 20 or 30 trunks. The results are as follows:

Capacity Table	% NC Encountered by Operators (from empirical data) With Alternate Routes for the Small Groups	% NC Existing (from Erlang "B) Without Alternate Routes
T-1	6–7	10–11
T-2	9–11	18–19
T-3	14–18	27–29
T-4	21–26	39–42
T-5	33–35	55–57

Conclusion

Since the primary function of an intertoll trunk capacity table is to translate a desired speed of toll service into the number of trunks required for that level of service, the table used should be indicative, within reasonable limits, of the probable effect of trunk provision on the overall speed. For this reason, tables which reflect a uniform service situation will be more useful in intertoll trunk engineering and administration than the present tables which have inherent service variations. Capacity tables such as Tables T-1 to T-5 will therefore be substituted for present Schedules A, A2 and B.

The author gratefully acknowledges the helpful cooperation of those in the several Associated Companies who participated in collection of the empirical data. Thanks are also extended to A. S. Mayo for his guiding hand; to R. I. Wilkinson and F. F. Shipley for their helpful comments; to K. W. Halbert and Miss C. A. Lennon who computed the necessary extensions of the Pollaczek formula; and to Miss E. B. Schaller for her skill in preparing the numerous curves.

Spark Gap Switches for Radar

By F. S. GOUCHER, J. R. HAYNES, W. A. DEPP and E. J. RYDER

Introduction

AN ESSENTIAL feature of radar is the generation, by means of an oscillator, of high-energy pulses of short duration, repeated many times a second. The energy for these pulses is furnished to the oscillator from a power supply in a variety of ways. One of the most widely used of these is the "line type modulator" in which a pulse-forming network made up of a series of condensers and inductances is charged from the power supply through a choke and is then discharged by a switch so that a substantially constant current will flow for a predetermined short time through the primary of a pulse transformer coupled to the oscillator. This switch is, therefore, an essential component of this type of modulator.

To meet the pulsing requirements of radar as it developed during the war, this line modulator switch was required to withstand thousands of volts between pulses and to carry hundreds of amperes for the pulse duration which was of the order of microseconds. Also, the switching operation had to be repeated from a few hundred to a few thousand times a second for a total operating time of hundreds of hours. Furthermore, the dissipation of energy within the switch had to be very small in comparison with the energy delivered to the oscillator for efficient operation.

The switch which had the widest application in this type of modulator was that employing an electric spark. Of over 50,000 radars of various types manufactured by the Western Electric Co. during the war, over half employed the electric spark in switching. One form of this switch was a rotary spark gap, operating in air, in which the timing of breakdown was controlled mechanically. These gaps were successfully adapted to a variety of radar types including airborne radar. However the demands for a more compact and lighter weight switch capable of operating at lower voltages for airborne radar led to the development of fixed sealed unit type gaps which, when connected in series, can be broken down electrically in a simple circuit.

Many problems had to be solved in the development of these switches. They required a considerable amount of study, and with the aid of new techniques developed during the war, a number of significant measurements have been made which have extended our knowledge of sparks generally. It is the object of this paper to describe the results of some of these studies,

as well as to describe the essential characteristics of a variety of spark gap switches which were used in such numbers that they may be considered as an important contribution to the war effort.

I. Rotary Spark Gap Switches For Low Voltage Circuits

Rotary gaps were used successfully as switches in some of the earlier radar systems developed by Bell Telephone Laboratories. The switching voltages in the modulator circuits were relatively high, being in excess of 20 kilovolts. No trouble was encountered in switching at the required pulsing rates nor in obtaining satisfactorily long life. Fortunately the sparks tend to move about the electrode surfaces uniformly and the rate of erosion is such that with tungsten or molybdenum electrodes a uniformly small change in electrode dimensions is achieved which in no way interferes with satisfactory operation over long periods of time.

A difficulty was encountered, however, when the switching voltage was reduced to lower values, as required for applications in which the power supply voltages were limited. The gaps failed to break down regularly.

A particular application in which this difficulty was encountered was one in which the power supply was limited to 4 kilovolts, and in which 80 ampere pulses of one microsecond duration were required every 600 microseconds. The modulator circuit used was that shown schematically in Fig. 1 (a). The pulse-forming network includes the condenser elements which are charged through the choke and discharged by the spark gap designed to break down at the required pulsing rate of 1600 per second. The load is the primary of a pulse transformer coupled to a magnetron and is closely equivalent to a 50-ohm resistance. The constants of the circuit are such that following the discharge of the network it is recharged sinusoidally along the solid line of Fig. 1 (b) to a peak value of approximately 8000 volts in 600×10^{-6} seconds, at which point breakdown must again occur and the operation be repeated. The dashed line is the approximate path of the charging voltage wave when breakdown at the peak fails to occur.

A rotary spark gap was designed to meet these pulsing conditions. In this gap there are four fixed and four moving electrodes as indicated in Fig. 1 (a). These electrodes are tungsten rods 3 mm in diameter and about 15 mm in length mounted with their axes parallel and so spaced that the moving electrodes pass very close to the fixed electrodes with an overlap of about one-half their length. The speed of the moving electrodes is such that in the region of near approach the maximum gradients are those indicated in Fig. 1 (c). The solid curve shows the gradients when breakdown takes place at the required time and the dashed curve the gradients when breakdown fails to occur. Although the latter greatly exceed the normal

dielectric strength of air, sparking failed to take place a large fraction of the time.

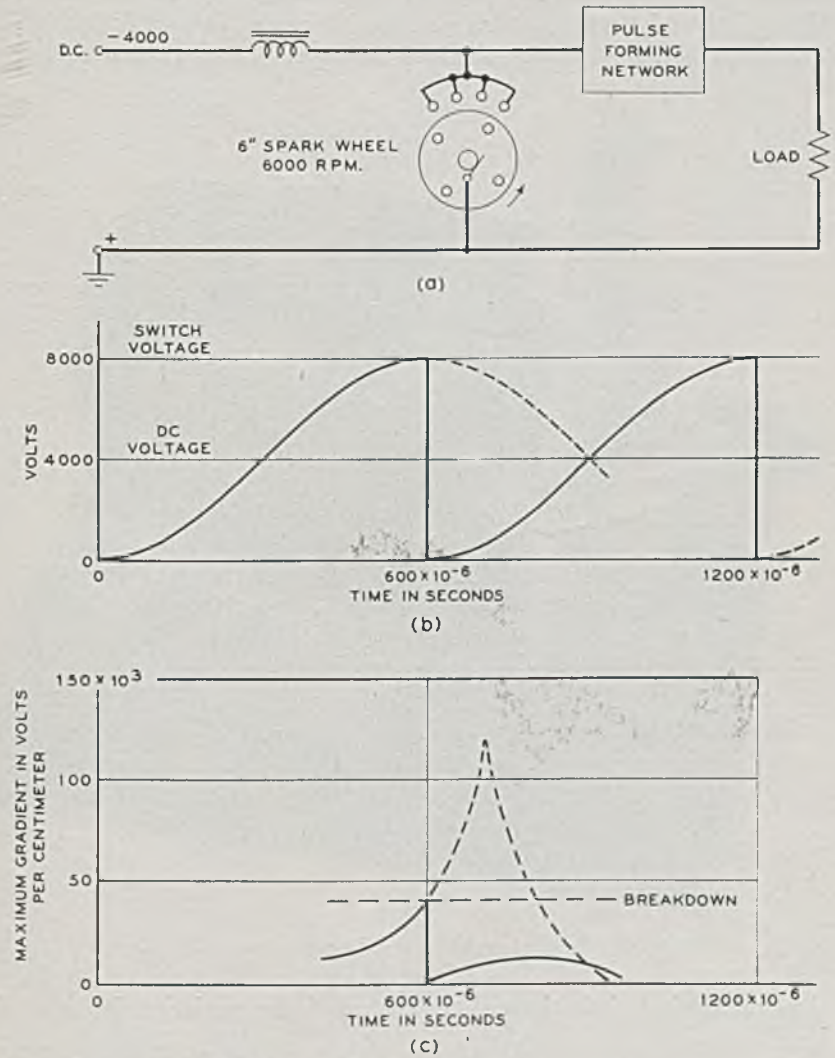


Fig. 1—(a) Line modulator circuit with rotary spark gap switch, (b) switch voltage vs. time, (c) maximum voltage gradient between electrodes vs. time.

Experiment indicated that this was caused by spark delay time, as irradiation of the cathodes by means of an ultra-violet lamp produced 100%

breakdown. This method of reducing spark delay time was not practical, however, and other means were sought. A solution was found through a rediscovery of the efficacy of corona prior to breakdown which came about through the introduction of a properly placed sharp edge on the cathode. Although this edge was apart from the sparking area of the cathode, 100% breakdown of the gaps was obtained and spark delay time so reduced

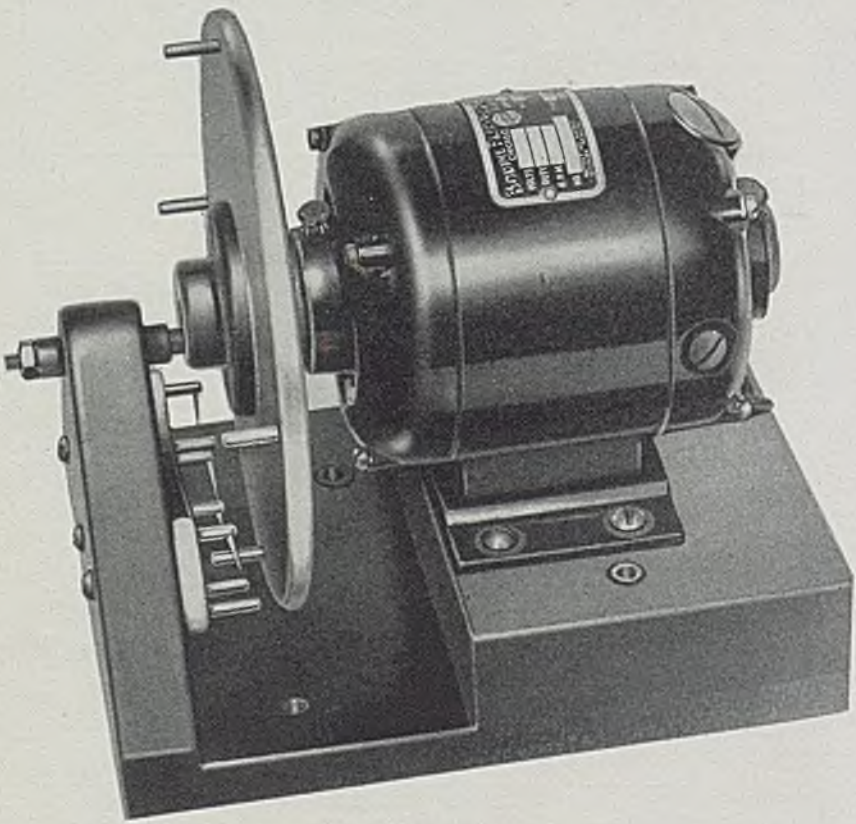


FIG. 2—Experimental model of rotary gap showing corona points.

that mechanical limitations alone controlled the variation in time of breakdown.

The essential features of the gap as finally developed for this project are shown in the photograph of an experimental model, Fig. 2, and in the perspective drawing and accompanying diagram, Fig. 3. The electrodes are of tungsten as in the earlier design, and corona points are introduced by the addition of the rods holding sharp metal points mounted on the same metal

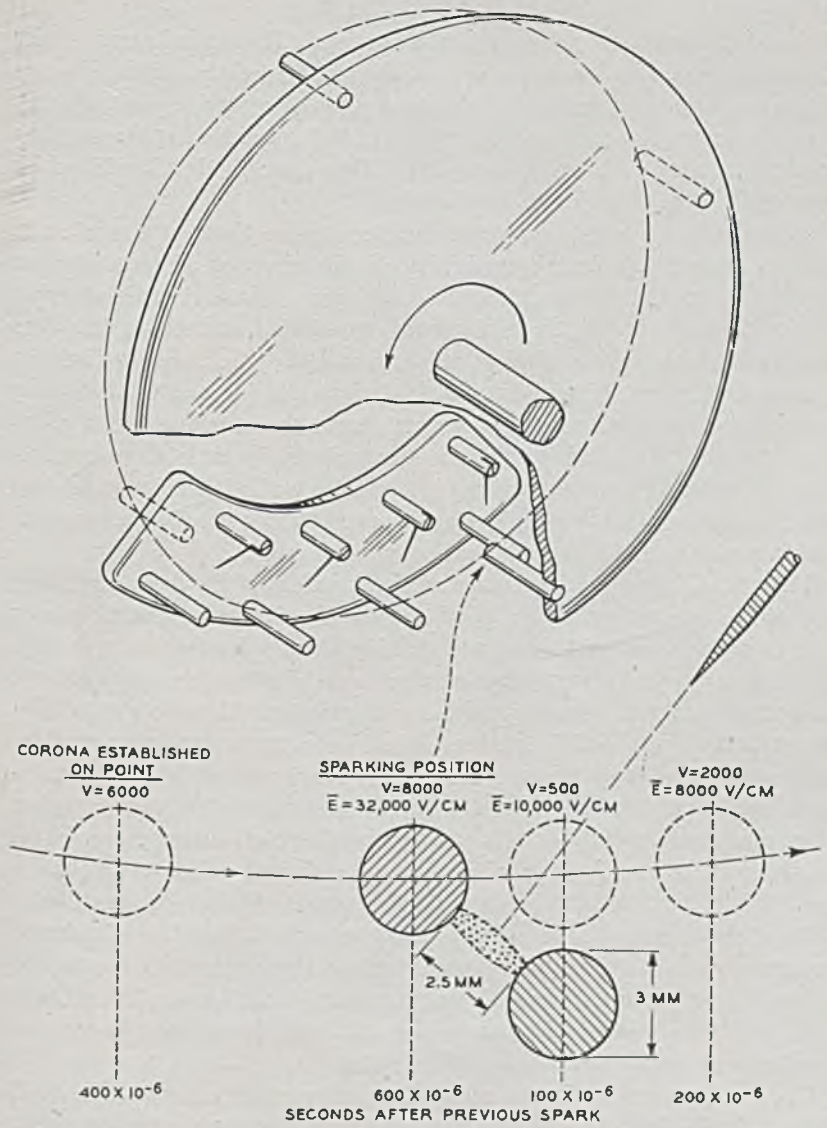


Fig. 3—Above, perspective drawing of rotary gap showing arrangement of corona points. Below, diagram showing voltage (V), and mean voltage gradient ($\bar{E}$) at various times in the spark-over region.

base as the fixed electrodes. The moving electrodes pass between the fixed electrodes and their associated corona points. This arrangement is clarified in the diagram which is a section through a plane normal to the electrode

axes and passing through the region of overlap. The shaded areas are for the sparking position as indicated and the location of the corona point is shown to scale. Experiment shows that when the moving electrode has reached the position corresponding to 400×10^{-6} seconds after the previous spark, corona is established on the point. Thus the cathode is irradiated for 200×10^{-6} seconds prior to breakdown.

No serious erosion problem was encountered when these gaps were operated for many hundreds of hours in air. No deterioration of the points was observed when their locations were properly adjusted so as to avoid sparking over to them. The cathode erosion rate is so low that appreciable flats were produced only after a hundred hours of operation. The anode erosion was estimated to be less than one-tenth of that of the cathode, and was doubtless associated with a small amount of reverse current shown to be present. The magnitude of the cathode erosion rate for tungsten in air is about twenty-five fold less than that for tungsten in hydrogen under the same conditions which indicates that oxygen plays an important and somewhat unexpected role in making practical the operation of these gaps.

There was, however, a serious corrosion problem when these gaps were adapted to airborne radar because of the necessity for sealing the modulator unit in a container capable of maintaining atmospheric pressure at high altitudes. Spark discharges in air are attended by the formation of both ozone and oxides of nitrogen, the latter combining with moisture to form nitrous and nitric acids. These reached such concentrations under continuous operation in the container that they were damaging to all enclosed equipment because of their corrosive action. A solution for this was arrived at after considerable study on the part of the Chemical Department. This consisted of the use of a copper impregnated activated carbon as an absorbent. With this absorbent a life of 500 hours was shown to be possible.

Over 10,000 rotary gap switches of this type were manufactured and used successfully in both ship and airborne radars. However, under the urge to reduce the weight of all possible components used in airborne radar and even to eliminate the necessity for pressurizing, the development of glass-enclosed fixed gaps as switches was diligently pursued.

The authors would like to acknowledge the cooperation of Mr. N. I. Hall of the Whippany Laboratories whose responsibility it was to engineer and develop these rotary gap switches for manufacture.

II. Fixed Gaps

Preliminary experiment indicated that a series of fixed gaps could be made to operate satisfactorily as a modulator switch. A study was therefore made to determine the most suitable gas atmosphere, electrode material and gap design for use in sealed gaps. This led to the development of a unit

type gap, two or more of which could be operated in series. The first unit type gap had an aluminum cathode and a hydrogen-argon gas atmosphere. Later, under the urge for higher peak powers, mercury cathode gaps were developed. Details of this study and development will be discussed in this section.

(a) *Triggering Gaps in Series*

An alternative to a rotary gap in which the timing of spark breakdown is controlled mechanically was the use of a fixed gap, the breakdown of which is controlled electrically. One method of accomplishing this was to use a third electrode to which an impulse voltage was applied periodically at double the frequency of the resonant charging circuit. This voltage breaks down one gap with a discharge of energy furnished by the trigger circuit, which in turn causes a breakdown of the main gap, either through a modification of the field in this gap or through the addition of ions which reduce its breakdown voltage. This type of gap, however, required a strong air blast to de-ionize the gaps and, because of this, its use obviously presented no great improvement over the rotary gap. It was well known that the rate of de-ionization is greater the smaller the gap, so an attempt was made to trigger without air blast a number of smaller gaps which when connected in series would withstand the full switch voltage as employed in the rotary gap.

The arrangement used was that shown in Fig. 4. Six tungsten pins, 3 mm in diameter, were mounted with their axes parallel and spaced to give five 0.5 mm gaps. The switch voltage was divided by means of equal high resistances connected across the gaps, and a highly damped bi-directional trigger pulse was applied to the four middle pins through capacity coupling as shown. Corona points were also connected in such a way that the cathode of each of the gaps is irradiated in order to reduce the spark delay time.

By an appropriate adjustment of the circuit elements it was demonstrated that this series of gaps could be broken down by the trigger pulse and de-ionized with sufficient rapidity so that no air blast was required.

Although no attempt will be made here to elucidate the detailed steps in the triggering of the five gaps just described, we can get a qualitative idea of the process by considering a simple two-gap and three-gap circuit which, it turned out, was all that was required for the various applications of fixed gaps as they were eventually developed.

In the two-gap circuit, Fig. 5 (a), if the first half cycle of the trigger pulse and the switch voltage are both positive, gap 1 will break down when the potential at the mid point, due to the sum of the switch voltage and that of the trigger, is equal to the gap breakdown voltage, which for the moment we shall consider as singly valued. This effectively shorts gap 1 and throws the full switch voltage across gap 2 which in turn will break down provided this

switch voltage is equal to or greater than the breakdown voltage of one gap. The gaps will operate for all switch voltages up to a value equal to twice the breakdown voltage of one gap when both gaps will break down without the addition of trigger. This, then, is the maximum operating voltage and the ratio of maximum to minimum operating voltage is two to one on the basis of this simple picture.

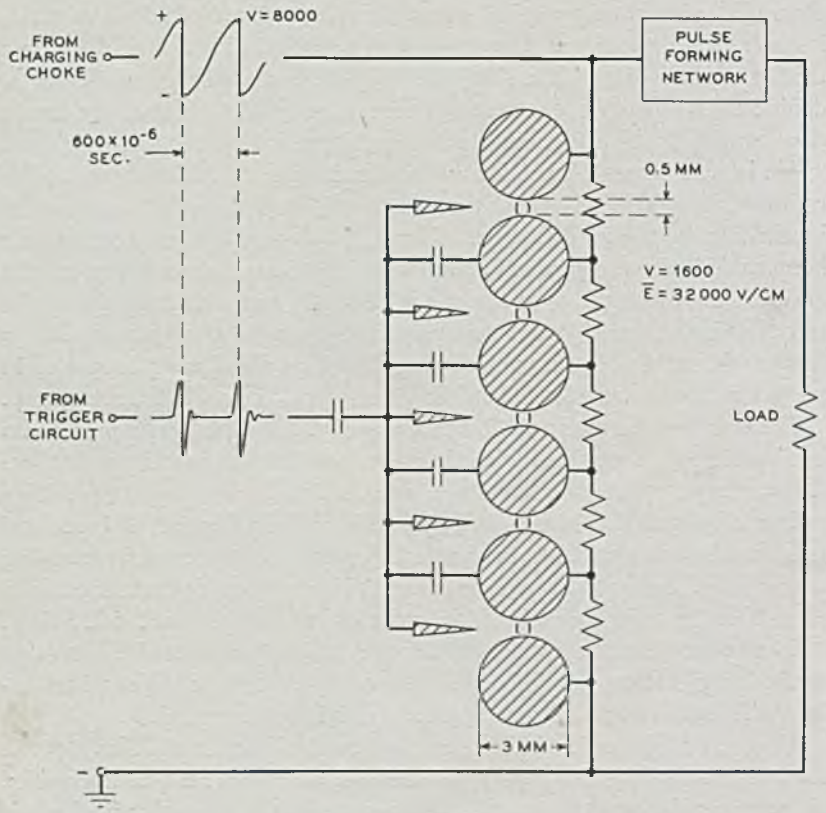


Fig. 4—Line modulator circuit with fixed gap switch composed of five 0.5 mm. air gaps triggered electrically.

In the case of the three-gap circuit, Fig. 5 (b), gaps 1 and 2 may be broken down by the simultaneous application of a trigger pulse through capacity coupling. The circuit elements can be so chosen that gap 1 first breaks down leaving enough trigger on gap 2, over and above that furnished by the switch voltage, to break it down. The full switch voltage is then applied across gap 3 and it will break down for values of switch voltage in excess of the

single gap breakdown voltage. In this case the switch voltage may be increased to a value three times that of the breakdown of one gap before the three gaps can break down without addition of trigger. Thus the ratio of maximum to minimum operating voltage is three to one. Ideally this ratio may be increased by the addition of more gaps.

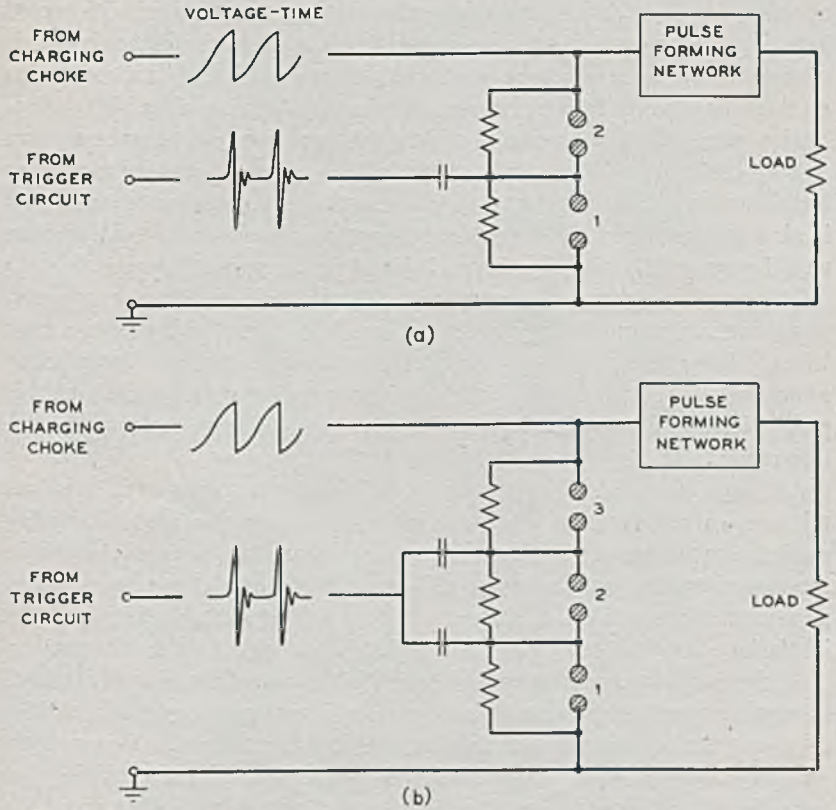


Fig. 5—Line modulator circuit (a) using two fixed gaps as switch, (b) using three fixed gaps as switch.

The operating characteristics of actual gaps do not conform exactly to this simple picture as we shall see later. This is because the breakdown voltage of a gap is not singly valued but depends on a variety of conditions such as rate of rise of applied voltage, pulsing rate, and the energy of the pulse, as well as the type of gap employed. However, we may regard it as a qualitatively correct picture of the operating characteristics of series gaps.

A more complete description of operating characteristics will be given in a

later section, but, in view of the fact that the gap itself plays an important part in these characteristics, it seems desirable to describe first the gap types with which we have to deal.

(b) *The Hydrogen-Argon Aluminum Cathode Gap*

Following the successful triggering of fixed gaps in air without the use of air blast for their de-ionization, experiments were undertaken with sealed gaps in various gas atmospheres using simple rod electrodes having their axes parallel. A large number of gases were tested and the conclusion reached that hydrogen was the most satisfactory because of its high de-ionization rate. With it fewer and wider gaps were required to meet a given pulsing condition. Three 4 mm. gaps in hydrogen at pressures somewhat less than atmospheric were approximately equivalent to the five 0.5 mm. gaps in air already referred to. Thus, from this point of view, the use of hydrogen would very greatly simplify the problem of making practical gaps.

The spark in hydrogen, particularly with relatively small peak currents, was, however, unsatisfactory in that it terminated in a high-pressure glow with a high cathode drop rather than the low drop required for efficient switching. The addition of about 25% argon corrected this and about this proportion was used successfully in the gaps with which we are concerned in this report.

Although the required operating conditions were met with this gas mixture, cathode erosion or sputtering was so excessive with all readily available cathode materials that this factor appeared as the chief obstacle in the way of making practical gaps. The sputtered material was deposited on all surfaces in the form of a fine powder which eventually destroyed the insulation, thereby limiting the useful life of the gaps to a few hours.[1]

A promising lead was, however, obtained in the case of aluminum cathodes. It was observed that some of the sputtered material deposited on the anodes opposite the cathodes from which it was removed. This deposit was reasonably compact and smooth, which suggested the possibility of reducing by gap design the extent of harmful scattering. This might be achieved by increasing the amount of sputtered cathode material which is deposited on the anode or returned to the cathode within the sparking area.

The tube, Fig. 6, was an early attempt in this direction. This tube had three 4 mm. gaps between flat electrodes, the cathode surfaces having raised portions to confine the sparking within their areas. The gaps were

[1] At about this time we learned that the British had developed sealed gaps triggered by means of an auxiliary electrode and known as "Trigatrons." These were high pressure gaps containing argon with a small amount of oxygen to reduce sputtering of the electrodes. The life of these gaps was determined by the time required to clean up this oxygen. Though these were tried it was decided to follow an independent development avoiding if possible all clean up effects.

operated successfully for somewhat over 100 hours before enough scattered material accumulated to interfere with gap insulation. A uniform spark distribution was maintained throughout this time and measurement showed that aluminum was removed quite uniformly from the raised portion of the cathode to a depth of only a fraction of a millimeter. An equally thick

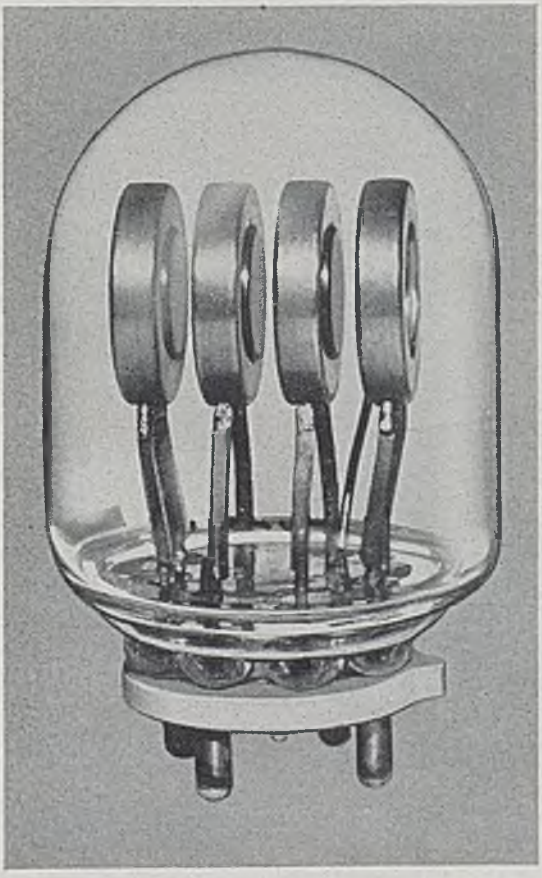


Fig. 6—Three gap tube having aluminum electrodes and a hydrogen-argon atmosphere ——actual size.

though somewhat rougher deposit was formed on the opposing anode surface, thereby retaining the gap spacing very satisfactorily. About 30 milligrams of loose material were scattered throughout the tube.

A more drastic but also more successful design change was introduced by making three separately enclosed gaps, one of which is shown in the photo-

graph and radiograph, Fig. 7. In these gaps the sparking area of the cathode was hemispherical in shape, partly surrounding a spherical anode. These gaps were operating successfully at the end of 1000 hours. The scattered material was deposited on only a portion of the glass envelope of

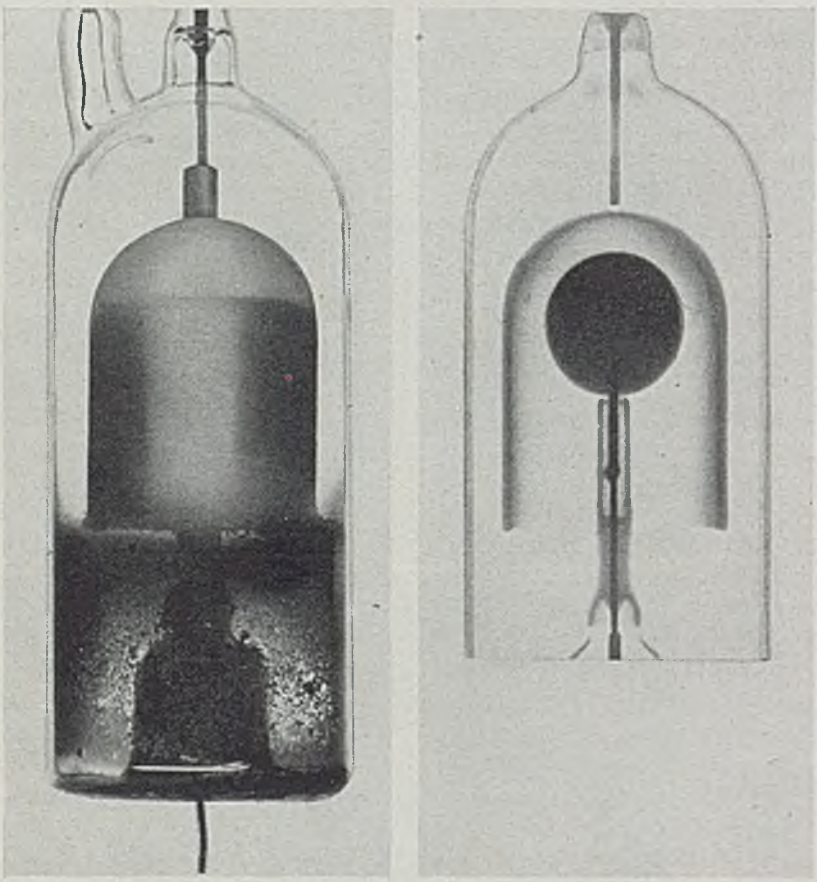


Fig. 7—Photograph and radiograph—actual size—of the first unit type gap having a re-entrant aluminum cathode, a spherical aluminum anode and a hydrogen-argon gas atmosphere, after operating 1000 hours with a 40 ampere pulse of one microsecond duration repeated 1660 times a second.

each gap, as shown in the photograph. The extent of the material removed from the cathode and deposited on the anode, as shown in the radiograph, was such as to cause no marked change in gap spacing. Furthermore, the operating range remained substantially constant throughout the 1000 hours of operation as shown in Fig. 8. This was an important observation since it

indicated that there is no gas clean-up effect associated with gap operation, a fact that was later proved by careful measurement of gas pressure before and after operating gaps of this type. A section through the anode of this gap, Fig. 9 (a), shows that the anode deposit is not compact but assumes the form of a coral-like structure. This low-density deposit must, however, be electrically equivalent to a compact surface as shown by the constancy of the operating characteristics with time.

In view of the success of this design it was decided to develop gaps of the unit type having anodes well enclosed by the cathode surfaces. An attempt to make a more practical gap is that shown in the photograph and radio-

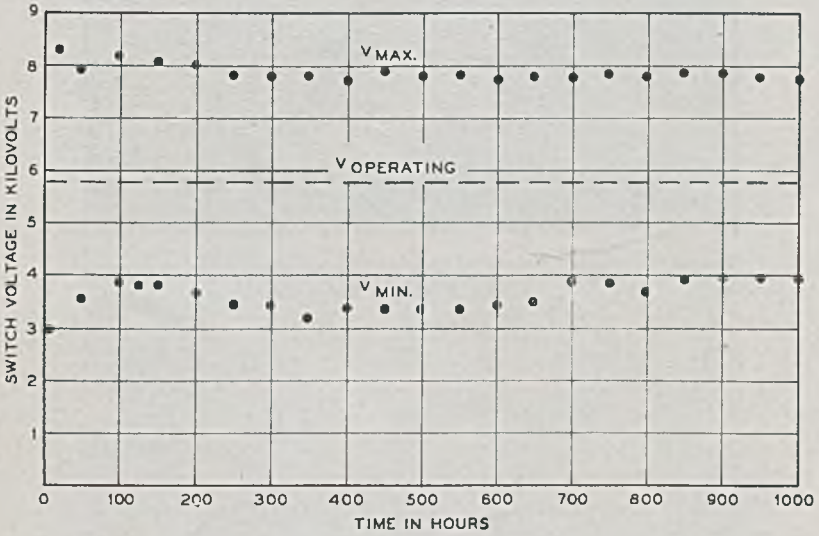


Fig. 8—Maximum and minimum operating voltages as a function of time, for three unit gaps of the type shown in Fig. 7, when operated in series.

graph, Fig. 10, both of which were taken after 750 hours operating time. In this gap the anode is an aluminum rod rounded at the end mounted concentrically with the enclosing cathode which has a hemispherical closed end. The corona point was added to facilitate starting. Because of the higher anode gradient the sparking was confined to the end region of the tube as indicated in the radiograph, and for this reason we have designated this design an "end sparking tube". A section through the anode, Fig. 9 (b), shows a deposit which in this case is compact due to the fact that the moving spark is confined to a smaller area than in the previous tube, Fig. 7 (a). It is to be noted also that the scattered material is less in extent than that

obtained with the first design, pointing to a more effective trapping of the sputtered material.

Weight loss measurements made under a variety of pulsing conditions show that though the rate of cathode erosion is somewhat dependent on the

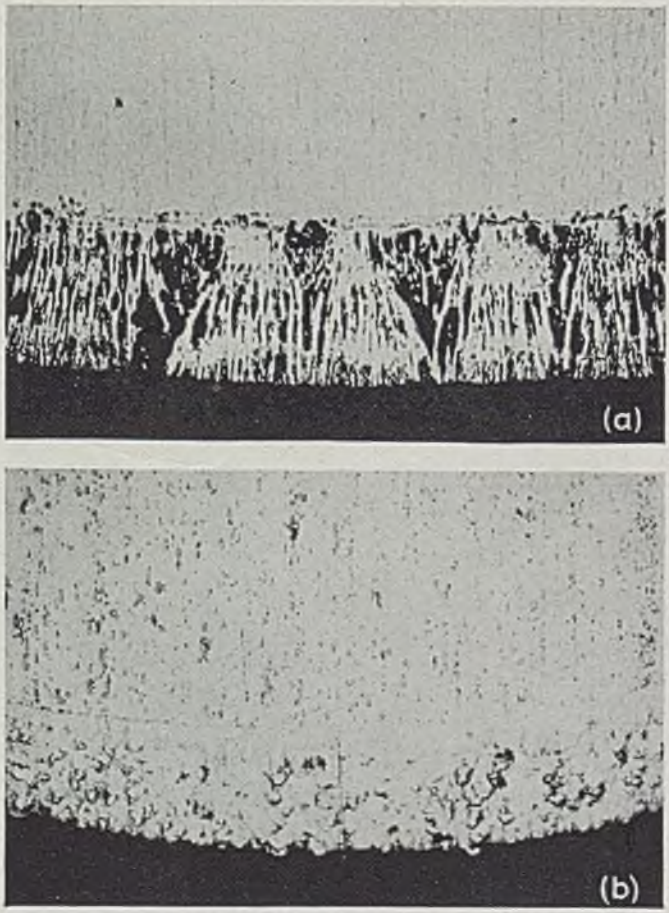


Fig. 9—Photomicrographs (×50) of sections showing anode deposits for (a) unit gap shown in Fig. 7, (b) unit gap shown in Fig. 10.

pulse duration, it depends to a much greater extent on gap design. Erosion rate measurements in terms of grams per coulomb are shown in Fig. 11 for the two gap designs there indicated and for pulse durations varying from one to five microseconds. It is clear that the open gap type of design in which the cathode is small leads to a loss which is at least five fold greater than

that of the "end sparking" type of gap. The smaller loss in the case of the latter shows that much more material returns to the cathode for resputtering

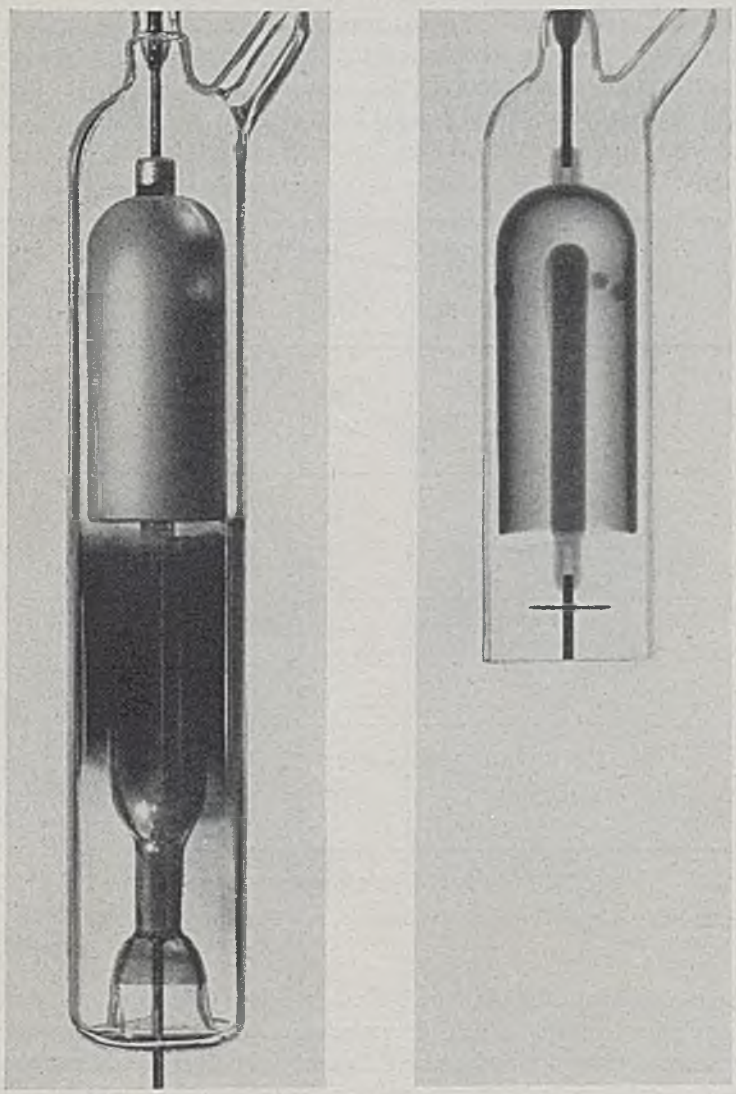


Fig. 10—Photograph and radiograph—actual size—of end sparking unit gap, after operating 700 hours with a 65-ampere pulse of one microsecond duration and repeated 1660 times a second.

than in the case of the former and supports the use of the unit type gap in which this process can be utilized. A cylindrical cathode enclosing a rod

anode also behaves in this way and its erosion rate differs but little from the "end sparking" type of tube; in fact, the practical gaps to be described in II-(f) are essentially of this type.

With these facts in mind it would appear that gaps could be designed to meet a variety of pulsing conditions if the total number of ampere hours for a pre-assigned life were known, for the electrode areas could be so adjusted that the changes in gap spacing would be as small as required. Analysis of the gradients associated with the end sparking type of gap shows that there

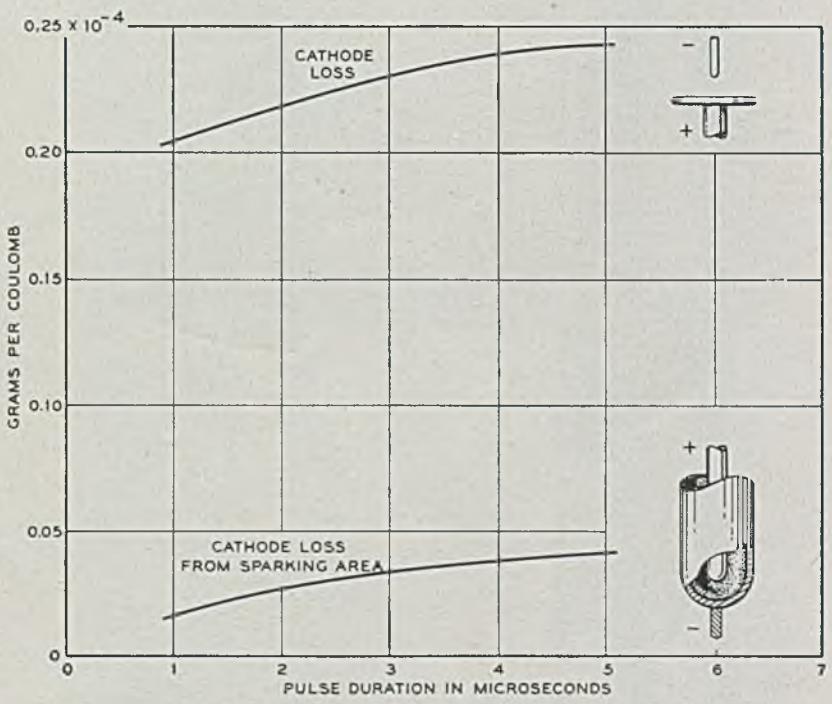


Fig. 11—Cathode loss, in grams per coulomb, as a function of pulse duration showing effect of gap design.

can be a considerable build-up on the anode before there is much change in the maximum gradient which determines the spark-over voltage.

Experience with gaps designed for a variety of pulsing conditions showed that substantial anode build-ups could be tolerated without interfering with operating conditions, but not as much as theory would predict for an unexpected factor had a controlling influence on gap life. This factor was the failure of the spark to keep moving under certain conditions with the result that spikes were grown on the anode which introduced a rapid deterioration

of the operating range due to an increase in anode gradient and also in part to a decrease in gap spacing.

Both the relatively large anode build-up, which may be tolerated without interference with gap operation, and the nature of spike growth, which limited useful life, are illustrated in the radiographs, Fig. 12. It is to be noted that the spike is almost of uniform cross section along its length and radiographs made at various stages of its formation show that growth takes place

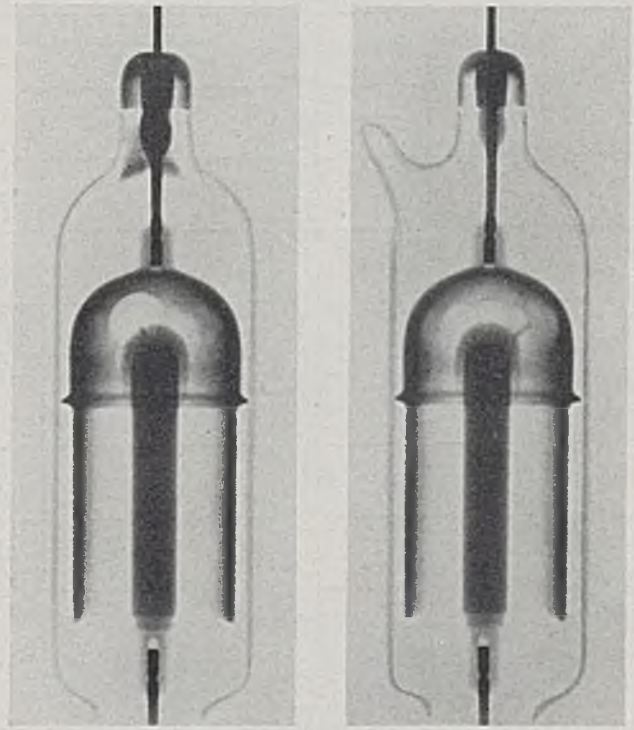


Fig. 12—Radiographs showing two views of the uniform deposit and subsequent spike growth on the anode of an end sparking unit gap.

at its end. This indicates a high concentration of negative ions in the vapor prior to deposit on the anode.

Life test data in which the pulse repetition rate was kept constant at 200 per second are shown in Fig. 13 for gaps having a fixed spacing but in which the peak current is varied (a), and for gaps having a variety of spacings but in which the peak current is kept constant (b). The life is measured in terms of hours to the beginning of spike growth. Both "end sparking" and "side sparking" tubes were employed in the tests. These data clearly show that

if lives longer than 500 are to be obtained, there is a limiting peak current of about 70 amperes with gap spacings of 250 mils or with a peak current of 70 amperes there is a minimum spacing of 250 mils. Similar data were obtained indicating a different critical spacing for other pulsing conditions.

This factor of a critical gap spacing imposed an important restriction on gap design for it was desirable to make gap spacing as small as possible for any given project. This follows because of gap size and weight, also—as

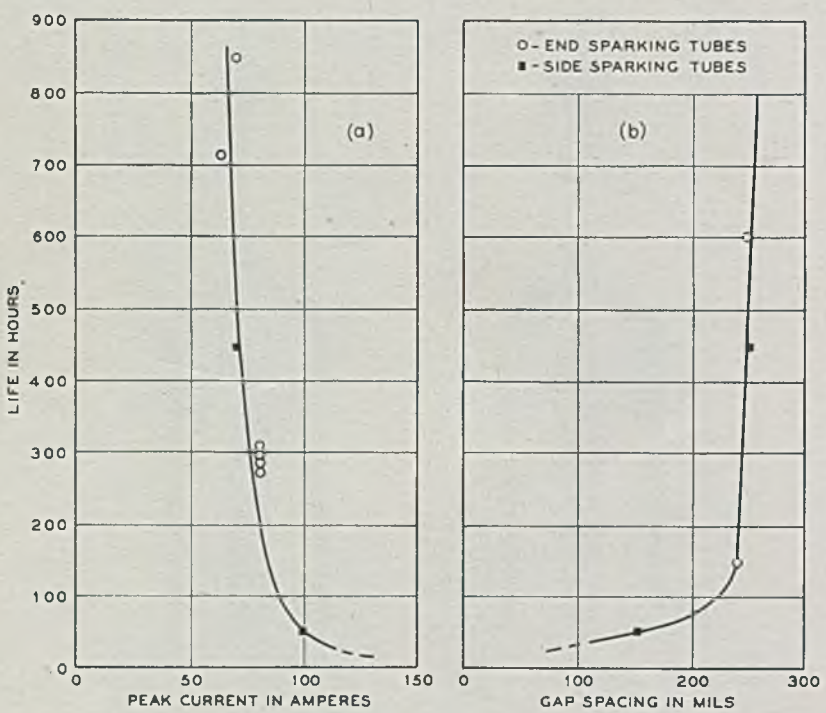


Fig. 13—Life in hours measured to the beginning of spike growth obtained with 5-microsecond pulses repeated 200 times a second (a) for a 60-mil gap and various peak currents, (b) for a fixed peak current of 70 amperes and various gap spacings.

we shall see in II-(e)—because of switching efficiency. This led to the development of a variety of unit gaps as described in II-(f).

(c) *The Mercury Cathode Gap*

Early in the study of the aluminum cathode gap it was realized that the sputtering difficulty might be largely if not entirely eliminated through the use of mercury as a cathode and the suppression of reverse current to avoid sputtering of the anode. It was shown that simple mercury pool cathode

gaps could switch peak powers in the megawatt range for long periods of time with stable operating characteristics. Under the urge for still higher powers than those which were handled by the aluminum cathode gaps, experiments were undertaken to develop a mercury cathode type of gap.

The main difficulty in the way of using mercury as a cathode is a mechanical one, as the conditions of operation of spark gap switches, particularly for

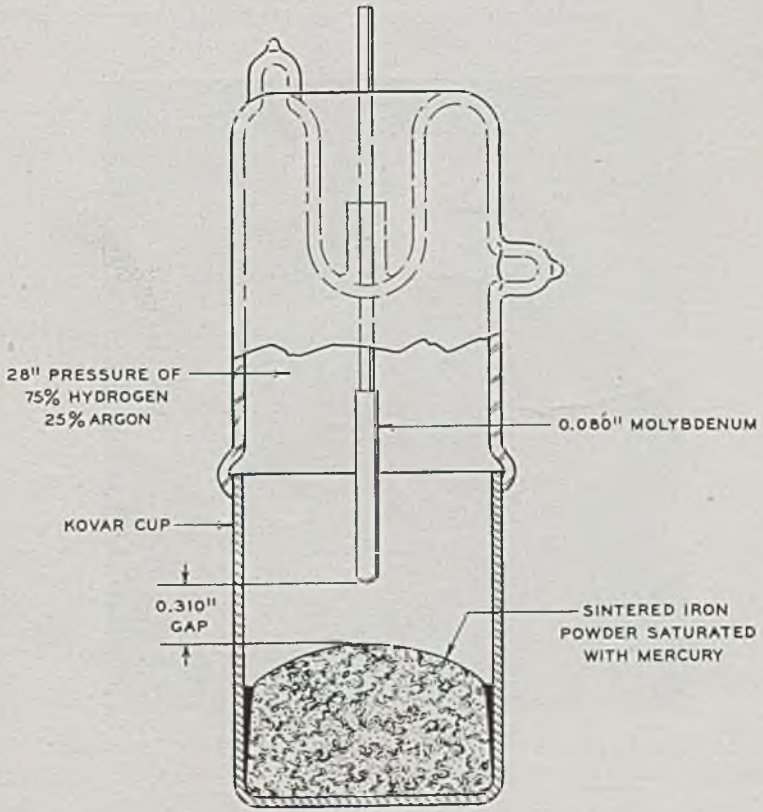


Fig. 14—Schematic drawing of an iron sponge mercury cathode unit gap—actual size.

airborne radar, demand that the sparking surface be rendered substantially quiescent. Preliminary experiments were made with metal baffles as damping agents and with metal wicks to furnish a mercury sparking surface. The latter led to the development of a sintered iron sponge saturated with mercury as the best means of obtaining a satisfactory cathode.

The constructional details of one of the earliest tubes of this type are given in the sketch, Fig. 14. The sintered iron sponge, a cross section of which is

shown in the photomicrograph Fig. 15, is about 60% porous. It was prepared by pressing iron powder in the Kovar cup and sintering in a hydrogen atmosphere. A special heat treatment to remove oxide made it possible to fill all pores of the sponge with mercury and to supply a mercury film on its surface. Under sparking conditions mercury from this film is evaporated and is condensed on the tube walls, eventually returning to the cathode. Due to capillary action the film is continuously replenished. This film protects the iron sponge from sputtering provided that there is sufficient

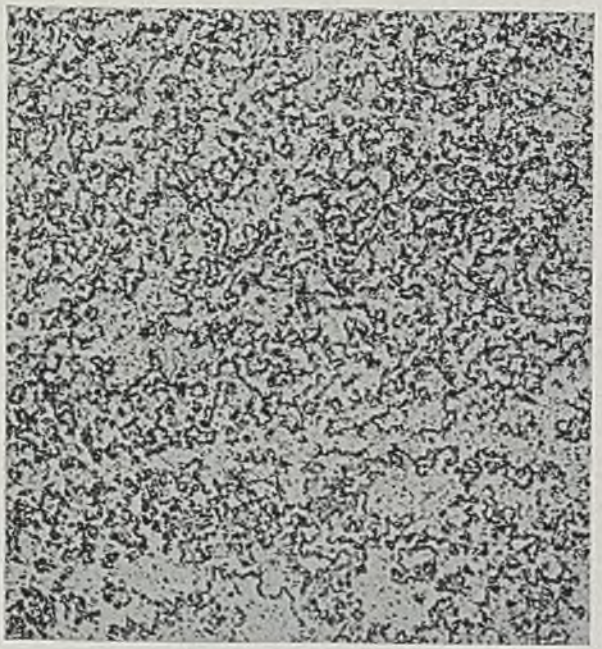


Fig. 15—Photomicrograph (×15) of a section through a sintered iron sponge showing porosity.

cooling of the cathode to maintain the mercury film at a temperature below its boiling point.

These gaps are not temperature sensitive as are most electronic devices containing mercury. This is because the mercury vapor plays no essential role in the spark discharge, as indicated by the fact that dissipation measurements—discussed in II-(e)—show its dependence on the hydrogen-argon rather than on the nature of the cathode material. With adequate cathode cooling gaps of this type operate satisfactorily over a range of ambient temperature at least from −50°C to over 100°C. Practical gaps constructed

with iron-sponge mercury cathodes were developed to the manufacturing stage, as discussed in II-(f).

In addition to being capable of switching higher peak powers than aluminum cathode gaps, the mercury cathode gaps can be designed to have superior operating characteristics. Through the use of small radius anodes not possible with the aluminum cathode gaps, a wider operating range and much less time "jitter" can be attained. The small anodes build up corona at voltages less than those of breakdown, thus furnishing radiation prior to breakdown. For special applications, gaps have been developed having a range approaching 3 to 1 in a two-gap circuit, capable of switching 10 megawatts peak power, for many hundreds of hours, and having a time "jitter" of less than 0.02 microseconds at the operating voltage.[2, 3]

(d) *Starting and Operating Characteristics*

It has already been stated in II-(a) that starting and operating characteristics of series gaps cannot be interpreted simply because, under the circuit conditions of rapidly varying voltage, the breakdown voltage of a spark gap is not singly valued. Because of spark formation time the minimum voltage at which a spark gap will break down increases as the rate of rise of the voltage across it increases. Further, due to spark delay time, the voltage across the gap at breakdown is usually still higher than this minimum value. It is therefore impossible to designate a unique breakdown voltage of a spark gap when the voltage across it is increasing with time. It is, however, possible to find a practical minimum and maximum breakdown voltage for a particular rate of rise of voltage. The difference between this maximum and minimum value is a measure of the maximum spark delay time. It is for the purpose of reducing this spark delay time that corona points (or radium) are introduced, and it will be shown that the value of both spark delay time and spark formation time have an important bearing on the operational characteristics of fixed gaps.

In addition to rate of voltage rise, the breakdown voltage of a spark gap depends on the amount of ionization in the gap due to a previous spark. When a spark discharge stops, a column of highly ionized gas is left in the gap. Although this column is rapidly de-ionized by recombination and diffusion of ions, a lower breakdown voltage is found for many microseconds in consequence of this residual ionization. The minimum value of the breakdown voltage of the gap is therefore a function of the time

[2] F. S. Goucher, J. R. Haynes and E. J. Ryder, High Power Series Gaps Having Sintered Iron Sponge-Mercury Cathode, P.B. 19640, U. S. Department of Commerce, Office of the Publication Board.

[3] J. R. Dillinger, Operation of Sintered Iron Sponge-Mercury Cathode Type Series Gaps at S.C.I., A.E.W. and 5 Microsecond Conditions, P.B. 13270, U. S. Department of Commerce, Office of the Publication Board.

after the spark ceases and is called the re-ignition voltage of the gap. It will be shown that this re-ignition voltage determines to a large extent the starting voltage of the fixed gaps.

Before describing the sequence of events required for starting and operating, it is desirable to define our terms more precisely than we have defined them up to this point. The minimum operating voltage is the lowest switch voltage at which the tubes will continue to break down 100% of the time under the action of the trigger pulse, and the maximum operating voltage is that higher switch voltage at which spontaneous breakdown of the series of gaps never occurs. Thus the operating range of voltage is that which includes those voltages existing across the series of gaps, at the time of application of trigger pulse, for which the tubes always break down under the action of the trigger pulse but never before. Starting voltage is defined as the minimum value of d-c voltage at which a series of gaps can be made to break down under the action of the trigger pulse. Starting thus differs fundamentally from operating in that while operating demands that the series of gaps always breaks down under the application of the trigger pulse, starting requires only that the gaps break down once in many trigger pulses occurring in a fraction of a minute. Thus, a starting voltage is always lower than the minimum operating voltage. However, due to the doubling of the switch voltage when starting occurs, the d-c power supply voltage required to start may be higher than the d-c power supply voltage at the minimum.

The results of a quantitative oscillographic analysis of starting and operating characteristics of a pair of preproduction W. E. 1B22 tubes[4] are now presented in detail, for they are qualitatively representative of all spark gaps. These tubes operate in a two-gap circuit, a schematic of which is shown in Fig. 5 (a). The analysis is carried out by an examination of the voltage-time wave which occurs at the point of application of the trigger pulse, the midpoint of the two gaps. It will help in understanding the oscillograms[5] which follow if it is borne in mind that the voltage across gap 1 is the voltage shown on the oscillogram with respect to ground or "O" voltage, while the voltage across gap 2 is the voltage shown on the oscillogram with respect to the switch voltage.

The sequence required for starting is shown in Fig. 16 (a). Just before the application of the trigger pulse the voltage at the midpoint of the two gaps is half that of the applied d-c by virtue of the resistance divider. When the trigger pulse is applied, the voltage rises to A (3.8 kv) which is the minimum breakdown voltage for these tubes with voltage rates of rise encountered in the trigger pulse. Gap 1, therefore, may break down at A

[4] These tubes contain both corona points and radium to reduce spark delay time (see II-(e)).

[5] The time scales of these oscillograms are expanded in regions of very rapidly reversing voltage in order to make clear the sequence of events.

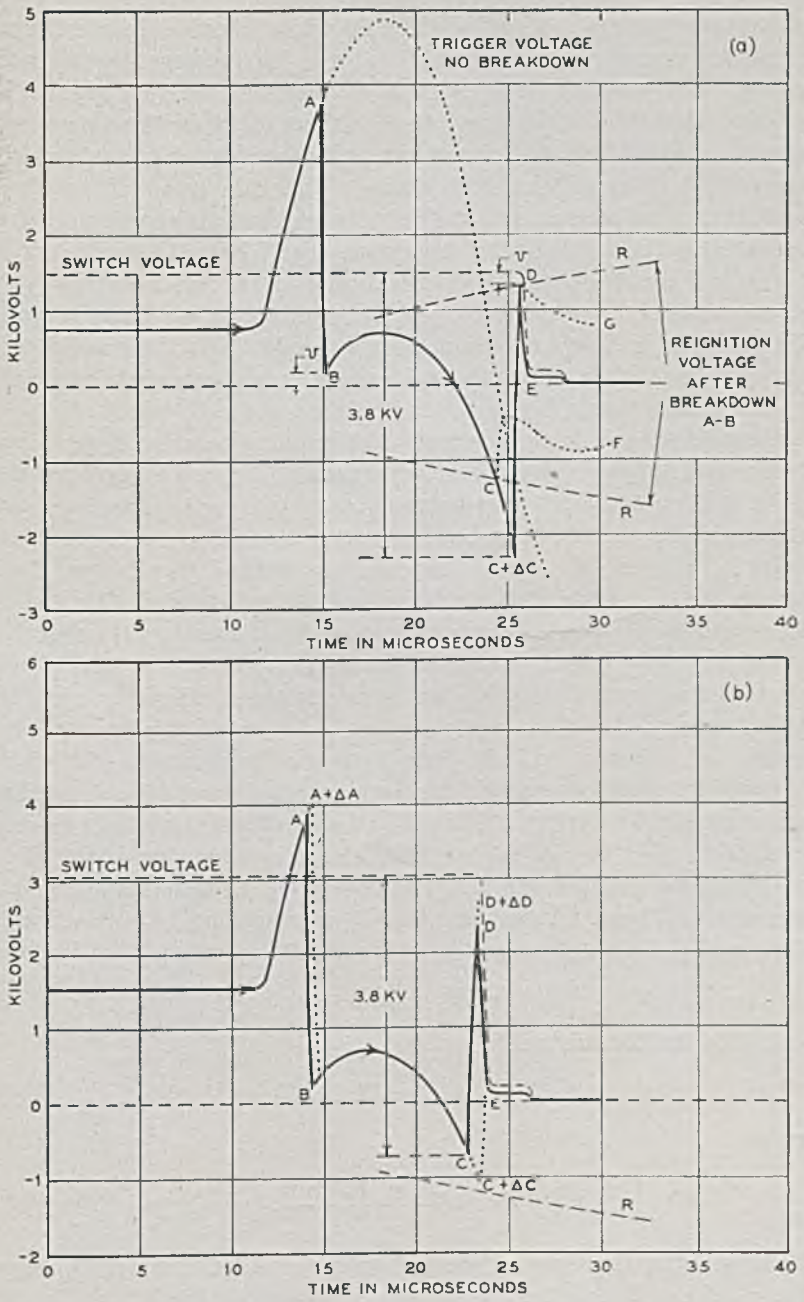


Fig. 16—Oscillographic traces of voltage vs. time as measured at the mid-point of a two-gap circuit during breakdown of 1B22 tubes, (a) for starting, (b) for operation at minimum switch voltage.

passing a low-energy spark supplied by the trigger circuit. In consequence of this, the voltage drops sharply to B and then the discharge stops since the voltage (v) remaining is insufficient to maintain the discharge. This voltage, called the extinguishing voltage, is about 0.2 kv for these low energy sparks. Gap 1 is now ionized and has the independently measured re-ignition voltage characteristics, R, as shown. Under the action of the trigger pulse the voltage then proceeds to $C + \Delta C$ when gap 2 may break down since it has the minimum required voltage across it (3.8 kv). When this occurs, the voltage rises sharply to D, which falls short of the switch voltage by the amount of the extinguishing voltage (v). At this point gap 1 may re-ignite. If this occurs both gaps are simultaneously conducting and the switch voltage drops to L while passing the high-current pulse of energy from the network. This sequence occurs relatively infrequently.

Because of spark delay time, instead of breaking down at A, gap 1 may break down at some higher voltage, or not at all. Instead of gap 2 breaking down at $C + \Delta C$, gap 1 may break down in the reverse direction at any voltage higher than C, its re-ignition voltage, and is only prevented from doing so by spark delay time. Also, because of this delay time, gap 1 will usually fail to re-ignite at D, its re-ignition voltage, and since D is also the extinguishing voltage (v) for gap 2, the potential will drop to G under control of the trigger pulse. If any one of these things occurs the gaps will not start on that particular application of trigger pulse. However, since the pulses are applied at the rate of many hundred a second, it is usually only a fraction of a second until the desired sequence is obtained.

From the conditions essential for the consummation of each of the three steps necessary for starting, it follows that the starting switch voltage V_{dc} must be equal to $A - (R + \Delta C)$ or $v + R$, whichever is the greater. Since R, the re-ignition voltage, increases with time, $A - (R + \Delta C)$ decreases while $v + R$ increases with time. A minimum for V_{dc} will, therefore, be obtained when the period of the trigger voltage wave is such that when gap 2 breaks down,

$$A - (R + \Delta C) = v + R, \tag{1}$$

and since also for this minimum

$$V_{dc} = A - (R + \Delta C) \tag{2}$$

we get

$$V_{dc} = \frac{A - \Delta C + v}{2}. \tag{3}$$

By substituting the observed constant values of A, ΔC and v in (3) we get $V_{dc} = 1.5$ kv, which is the value of switch voltage depicted in the diagram. This diagram is, therefore, that for optimum period of the trigger voltage

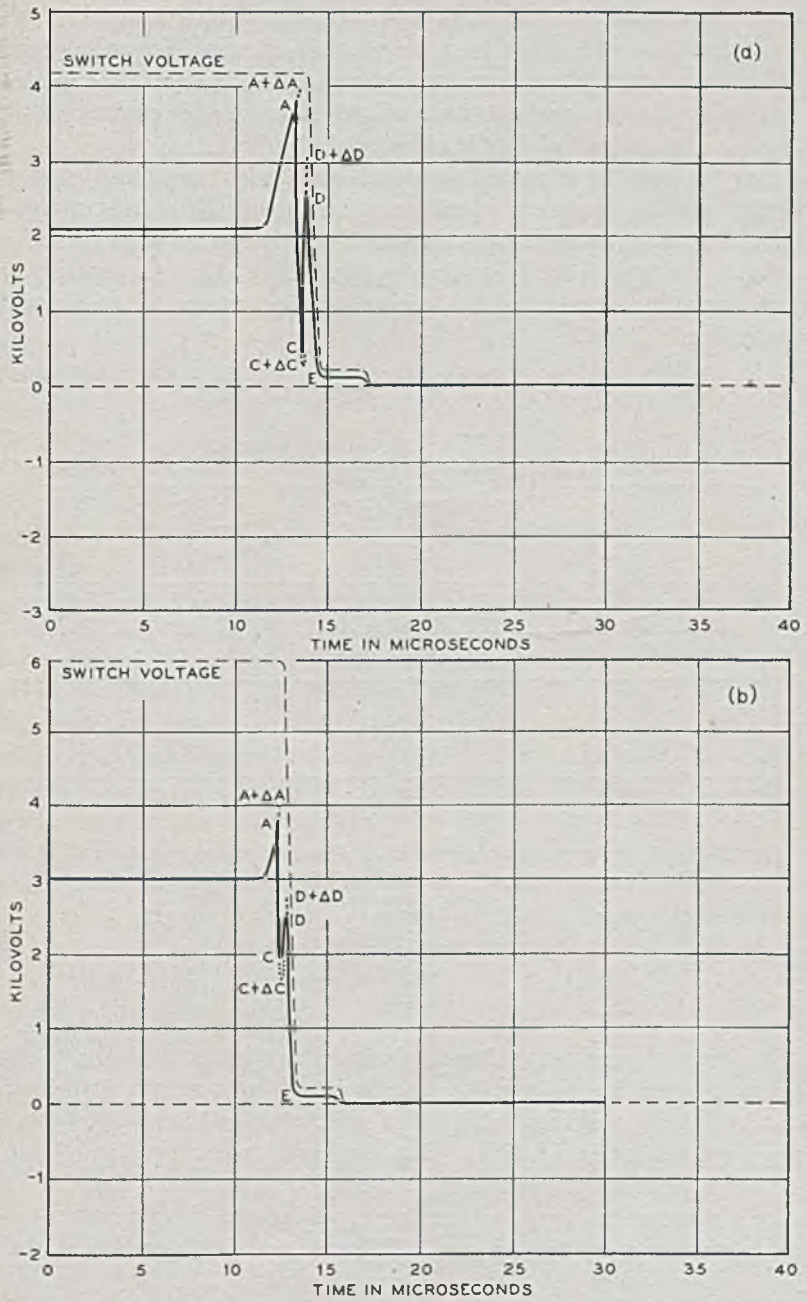


Fig. 17—Oscillographic traces of voltage vs. time as measured at the mid-point of a two-gap circuit during breakdown of 1B22 tubes (a) for normal operating switch voltage (b) for operation at maximum switch voltage.

wave. That this is actually a minimum was demonstrated experimentally by varying the period of the trigger pulse. V_{dc} increased for pulse periods both greater than and less than that shown in the diagram. The increase was small and so is of no great practical interest, but it does confirm the prediction made on the basis of the above analysis.

After the tubes have started the switch voltage is nearly double the d-c voltage, and the tubes will operate continuously if the switch voltage is above the minimum operating voltage. The sequence of events near the minimum operating voltage is shown in Fig. 16 (b). During operation the spark delay time is much less than during starting, as indicated by a smaller voltage is established.

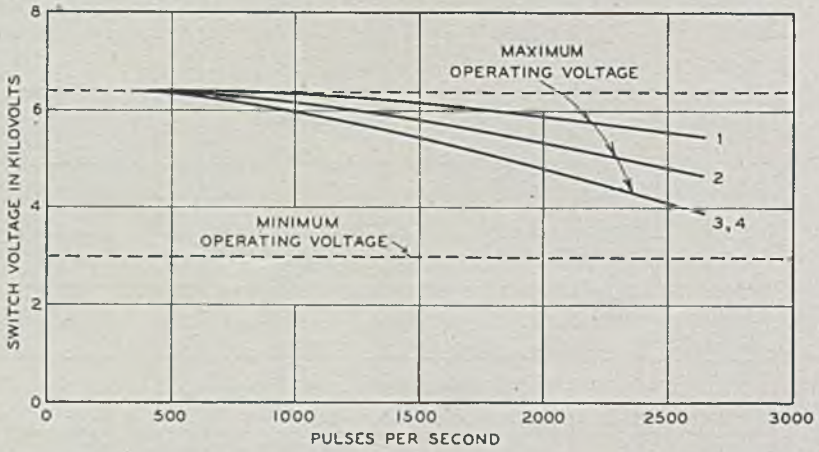


Fig. 18—Maximum operating voltage of 1B22 tubes in a two-gap circuit as affected by pulse repetition rate for a variety of pulsing conditions as follows:

Curve	Pulse Duration in Microseconds	Load in Ohms
1	0.75	55
2	0.75	30
3	0.75	15
4	1.50	30

Of course, no spark gap tubes are designed to operate very close to their minimum operating voltage. A margin of safety is always maintained. The characteristics of these tubes with a switch voltage at a practical operating voltage is shown in Fig. 17 (a). Gap 1 breaks down between A and $A + \Delta A$ and before gap 1 is extinguished gap 2 breaks down between C and $C + \Delta C$. Since in this case both gaps are conducting simultaneously, the main pulse passes without re-ignition of gap 1. The voltage at the midpoint of the two discharges rises to a value between D and $D + \Delta D$, due to the rapid change of spark impedance. This sequence always takes place since ample margin is provided.

If the switch voltage has been increased to a value near the maximum operating voltage, the voltage-time characteristic shown in Fig. 17 (b) results. Exactly the same sequence occurs as before. However, if the voltage be slightly increased above the value shown, the gaps can break down spontaneously during the network charging cycle and before the application of the trigger pulse, even though the value of A is some 20% greater than the charging voltage applied to the gap. This is the expected effect of spark formation time on minimum breakdown voltage since the rate of rise of trigger voltage is far higher than that of the network charging voltage. When spontaneous breakdown occurs, because of circuit conditions, both the rate of rise of the voltage of the network charging cycle and its peak value are increased. Since the switch voltage arrives at a higher value in a shorter time, spontaneous breakdown is most likely to occur again. The effect is cumulative so that, after a few increasingly frequent cycles, an arc is established. It is clear that this arcing must never be allowed to occur in the operating range.

These characteristics were taken while using a current pulse of 0.75 μs duration at a repetition rate of 1000 per second and a 30 ohm resistance load. This produced a peak current at the maximum operating voltage closely closely equal to the switch voltage divided by twice the resistance load, or about 100 amperes. Under these conditions, due to the relatively low pulse repetition rate, there is little residual ionization in the gaps at the time of the next pulse, so that the gaps have closely recovered their maximum breakdown voltage. However, as the pulse rate is increased, thus decreasing the time between pulses, the value of switch voltage at which the gaps break down spontaneously is found to decrease due to residual ionization. Thus the maximum operating voltage is a function of the pulse repetition rate.

The decrease of the maximum operating voltage as a function of pulse rate, for these tubes, is shown in Fig. 18 for a variety of pulsing conditions.

Curve 2 was obtained with the 0.75 μs pulse and a 30-ohm load. It will be observed that the maximum operating voltage decreases with pulse rate in the expected manner.

If the peak current of the pulse be decreased, fewer ions are produced in the spark and so at any given time after the pulse one would expect less residual ionization in the gaps. Curve 1 was obtained by keeping the pulse duration the same as before but increasing the load resistance to 55 ohms. Thus the current at a given switch voltage was reduced to 30/55 of its former value. It will be seen that, as predicted, the drop of maximum operating voltage with increased pulse repetition rate is less.

Conversely, if the current is increased the opposite effect is produced. Curve 3 was obtained by decreasing the load resistance to 15 ohms while keeping the pulse duration constant. This gives twice the peak current at

the same switch voltage as that of Curve 2 with a resultant increased residual ionization and a decrease of maximum operating voltage at the higher pulse repetition rates.

If, instead of doubling the current, the pulse duration be doubled, a similar increase in residual ionization is produced. Curve 4 was obtained by doubling the pulse duration (1.5 μs) and using a 30 ohm load. Thus, while the current is the same as Curve 2, the current pulse has twice the duration. It will be observed that for these pulses, doubling the time of pulse is the equivalent of doubling the current.

One might expect that the minimum operating voltage would also decrease as the pulse repetition rate is increased. However, experimentally it is found that, for these tubes, the minimum operating voltage is nearly constant and, therefore, independent of residual ionization. This result is produced largely because the maximum breakdown voltage of the gaps at the extremely high rate of voltage-time change encountered in triggering at the minimum is little affected by this amount of residual ionization.

Since the minimum is nearly constant the operating range of voltage of these tubes is a decreasing function of the pulse repetition rate, current, and pulse duration. This is in general true of all fixed spark gaps; however, the amount of decrease of operating range depends on the spark gap spacing, gas atmosphere and geometry of the electrodes.

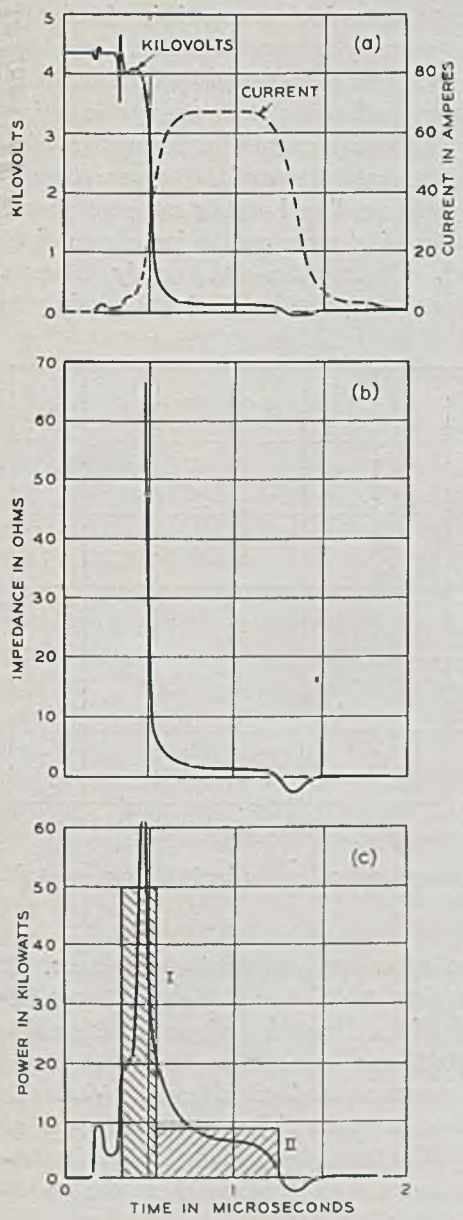
(e) *Dissipation and Switching Efficiency*

In II-(d) we considered the voltage-time relationships leading to the simultaneous breakdown of series gaps. In this subsection we will consider the voltage and current relationships with time during this breakdown, and their bearing on spark dissipation and switching efficiency.

In Fig. 19 (a) are shown a voltage-time and current-time trace obtained oscillographically with a pair of 1B22 gaps. The voltage is measured across both gaps and corresponds to the dotted traces shown for switch voltage in Fig. 17 (a). The current pulse is shown in proper time relationship with the voltage trace. Similar traces are obtained for any pulse duration and peak current. These, then, may be considered as typical of all pulses produced by spark switching with these gaps.

In Fig. 19 (b) is plotted the impedance of both gaps with time, from which we see that the impedance of this switch falls rapidly in a small fraction of a microsecond to an average value of only a few ohms while the main current pulse is passing. The tail of the trace showing a negative impedance is due not to the gaps but to inductance inherent in their leads.

The solid trace, Fig. 19 (c), shows the product of voltage and current in kilowatts plotted against time. The integrated area of this plot corresponds to the dissipation per pulse of both gaps. This area is independent of the



(a) Voltages vs. time and current vs. time for 0.75-microsecond pulse.

(b) Impedance vs. time.

(c) Instantaneous power dissipated in gaps vs. time—solid trace from oscillographic, shaded areas from calorimetric measurements.

Fig. 19—Pulse characteristics of two 1B22 tubes operated in series.

pulse repetition rate, enabling one to determine the gap dissipation for any project by multiplying the loss per pulse by the repetition rate.

This area can be divided into two parts as suggested by the two shaded blocks I and II. The first part corresponds to the energy dissipated initially by the trigger and then by the pulse forming network in the brief transient period when the voltage across and the current through the gaps are changing rapidly. The former is comparatively small and usually can be neglected. The latter attains a maximum value of power when the impedance of the gaps approximates that of the load. The second so-called steady state part, corresponding to block II, represents the energy lost during the main pulse

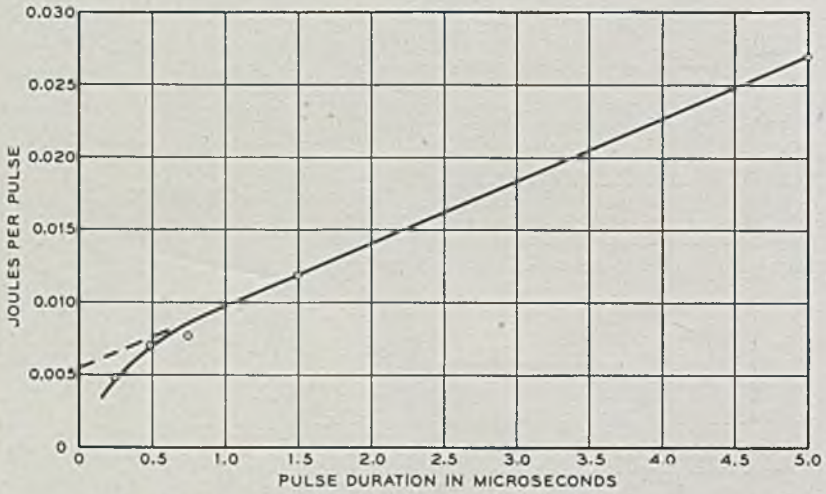


Fig. 20—Dissipation per gap per pulse vs. pulse duration for 1B22 gaps operated in series with a peak current of 70 amperes.

when the impedance of the gaps is low and comparatively constant. Its value will depend on both the pulsing conditions and the gaps themselves.

A calorimetric study was made of the dissipation of gaps as affected by various parameters. This method was superior to the oscillographic approach in that it afforded greater accuracy and ease of measurement. The curve, Fig. 20, shows observations in terms of joules per pulse per gap obtained calorimetrically with the 1B22 type tube as a function of pulse duration in microseconds. The peak current in all cases was 70 amperes and the trigger energy was included. It is clear that for pulse durations greater than 0.5 microseconds the dissipation D in joules per pulse per gap is given by

$$D = A + Bt \tag{1}$$

where A is the intercept of the extrapolated solid straight line through the value of ΔA and ΔC. There are reasons for believing that this is due primarily to the higher electrode temperature, but is doubtless aided by residual ionization left over by the high-energy sparks now passing. As a consequence the first gap always breaks down at voltages intermediate between A and $A + \Delta A$. Gap 2 always breaks down at voltages between C and $C + \Delta C$, below the re-ignition voltage of gap 1, and gap 1 always re-ignites at voltages between D and $D + \Delta D$ allowing the main pulse of current to current to pass at a voltage E. However, if the switch voltage is decreased, $C + \Delta C$ will occasionally cross the re-ignition voltage characteristic R of gap 1. Gap 1 can then re-ignite and thus the gaps will not fire on the application of that trigger pulse. A second way in which the gaps can miss is by failure of gap 1 to re-ignite at D. Even though either of one of these events occurs only once in many thousands of pulses, a minimum operating observed points and where B is the slope of this line. The shaded blocks I and II of Fig. 19 (c) were obtained from values of the two terms A and Bt, respectively, showing graphically the agreement between the calorimetric and the oscillographic methods.

As a result of calorimetric measurements on a wide variety of gaps having either aluminum or mercury cathodes and operated under a wide variety of pulsing conditions, we have been able to establish an empirical formula for the dissipation D in joules per pulse per gap in terms of these gap parameters and pulsing conditions as follows:

$$D = 5.7(10)^{-7} I_p S + [40 + 3.9(10)^{-2} p^{0.4} S] I_p t. \qquad (2)$$

Here I_p is the peak current in amperes, S the gap spacing in mils, p the gas pressure of hydrogen-argon in inches of mercury, and t is the duration in seconds of an idealized square-top wave equivalent in ampere-seconds to the actual current wave. This formula holds for either aluminum or mercury cathodes and is independent of gap design. It is modified only slightly when pure hydrogen is substituted for the hydrogen-argon mixture, the constant $3.9(10)^{-2}$ becoming $3.1(10)^{-2}$. It is based on many measurements in which the parameters covered the following ranges:

Parameter	Range
S	40–350 mils
p	28–50″ Hg.
t	1–6 × 10^{-6} seconds
I_p	45–1070 amperes

After calculating the value of D from Equation (2) the dissipation in watts per gap for any project is obtained by multiplying by the pulse repetition rate. This equation does not include the trigger energy dissipated which usually

can be neglected but which can be measured independently and added if so desired.

It is to be noted that the first or transient term of the formula is unaffected by pulse duration and argon content and depends at least to a first approximation on only the peak current and length of spark. The numerical constant includes the time of this transient, the average gradient during this period, and a factor to reduce the peak current to an average value. The portion of the second or steady state term within the brackets represents the average voltage across a gap when it is highly conducting and is approaching the characteristics of a steady arc. This average voltage is separated into two parts. The first part, 40 volts, is the sum of the cathode and anode drops arising from space charges at the electrodes. The second part is the voltage drop along the positive column which has a pressure dependent uniform gradient and which is of the order of 100 volts per cm. It is only this gradient which is perceptibly altered when argon is added to the hydrogen.

From this formula it is possible to calculate the switching efficiency for any design of gap and set of pulsing conditions within the specified range of parameters covered by the formula. Calculation shows that with three gaps in series the switching efficiency in all projects was at least 90%, whereas with two gaps in series it was in most cases as high as 96%.

(f) *Development of Fixed Gaps for Manufacture*

The designs of the fixed gaps for manufacture were dictated by the requirements of particular modulators. Under the code number of each of the gaps a brief description is given of the electrical and mechanical requirements which had to be met.

W.E. 1B22

The 1B22 fixed gap tube is an aluminum cathode type with a hydrogen-argon filling. An exterior and a cross-sectional view are shown in Fig. 21. This fixed gap tube was developed for the modulator of an airborne radar known initially as ASH and later an AN/APS-4. In this modulator two tubes are used in series to switch a peak power of about 105 kilowatts into a W.E. 725A magnetron. It was desirable that the peak voltage in the modulator section be kept fairly low so that the circuit would perform satisfactorily at high altitude even when the pressurizing container was damaged. Furthermore, the equipment was to be very compact and light in weight.

In order to meet the requirements of this radar, two tubes were used in series with a peak switching voltage of 4 kilovolts. They were required to pass a current pulse of 67 amperes for 0.75 microseconds at two repetition

rates, one of 600, the other of 1000 pulses per second. They were also to operate for short periods at 2.25 microseconds and 330 pulses per second. The main problems in the design of such a tube were those of obtaining an adequate service life and a sufficiently low starting voltage.

As pointed out in II-(b), the life of an aluminum cathode gap of this type is critically dependent upon the anode-cathode spacing. For this tube a

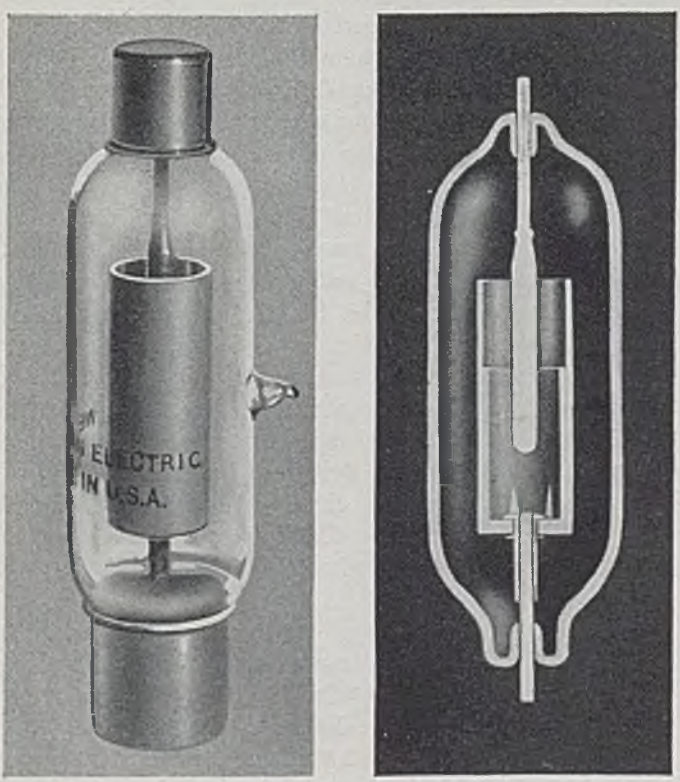


Fig. 21—Western Electric 1B22 spark gap tube.

spacing of approximately 150 mils was selected, the gas pressure being 20 inches of 75% hydrogen and 25% argon. This gave a life of about 500 hours for the 0.75 microsecond pulse, and a much shorter life for the 2.25 microsecond pulse. However, since the latter pulse duration is used only a small part of the time, the service life proved to be adequate. In order to obtain the maximum life from each tube, it was necessary that the anode and the cathode depart no more than a few mils from concentricity. Otherwise the sparking would not be uniformly distributed radially, leading to a

non-uniform anode build-up and a shortened life. Furthermore, in order to prevent failure of the tube, due to sputtered material destroying the insulation of the interior glass walls, the inside diameter of the cathode was enlarged near the open end, thus confining the sparking to the deeper portion of the cathode cylinder.

As discussed in II-(d) the starting voltage of a pair of fixed gap tubes is particularly important. The operating voltage of the tubes in this case is approximately 4 kilovolts which is derived from the resonant charging of the pulse shaping network condensers from a high voltage supply of about 2.2 kilovolts. The open circuit voltage of this supply is about 2.7 kilovolts. This, then, is the voltage available for starting the gaps. In order to make the gaps start at a voltage well below this value, corona points were introduced at the end of the cathode opposite the end of the anode, a small quantity of radium was also introduced in this region, and the anode diameter was reduced to the lowest value consistent with long life. The effectiveness of the corona points and the radium was reduced by the sputtered material during the life of the tube, but the irregular deposition of this sputtered material favored the production of corona and actually reduced the starting voltage to a lower value than that for a new tube.

The tube was designed for fuse clip mounting but it was found that the acceleration imparted to the tube when it was snapped into heavy clips was greater than that encountered in flying service. Accordingly, a special mounting was devised so that the tube would not be broken when being installed in the radar set. By the end of the war these tubes had been installed in approximately 15,000 radar equipments.

W.E. 1B29

The 1B29 fixed gap tube is similar in constructional details to the 1B22 except that it is smaller, the gap spacing being only 90 mils. An exterior and a cross-sectional view of the tube are shown in Fig. 22.

The gaps were designed to switch 2.8 kilovolts and to pass a peak current of about 27 amperes for 0.75 microseconds at a repetition rate of 2000 pulses per second. The main design problems were those of adequate life and stability of tube drop during conduction.

The small size of these gaps resulted in a life of only 300 hours which was, however, quite adequate for this application. As pointed out in II-(b) the argon was added to the hydrogen to ensure a uniform low impedance on sparking. The extremely small peak current required an increase in the amount of argon to 50% instead of the usual 25%.

In mechanical construction, the 1B29 is essentially a scaled-down 1B22. Because of the smaller size of the tube, no new problems existed in making it rugged.

Sufficient tubes were manufactured to supply approximately 3000 radar equipments.

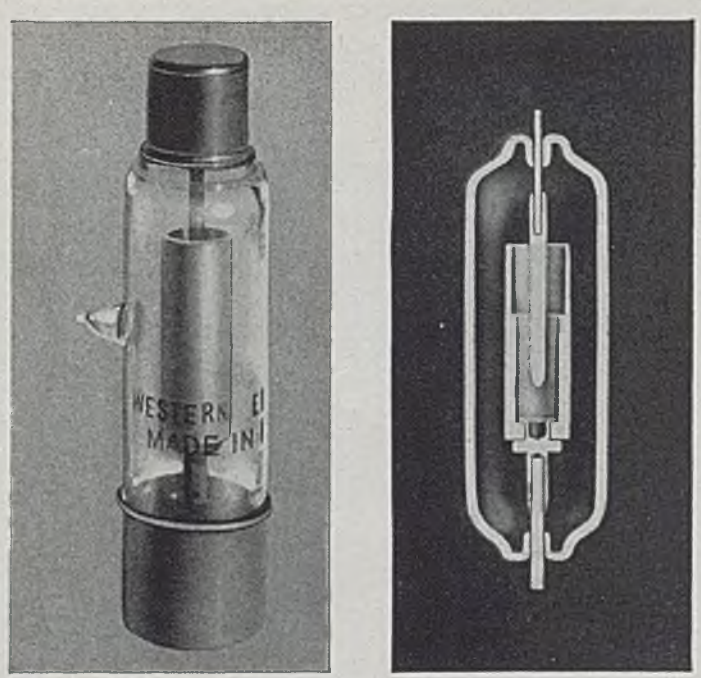


Fig. 22—Western Electric 1B29 spark gap tube.

W.E. 1B31

The 1B31 fixed gap was also an aluminum cathode gap, with a gap spacing of 300 mils and 24 inches of 75% hydrogen and 25% argon. An exterior and a cross-sectional view are shown in Figure 23. This gap was developed for an airborne radar. This modulator was required to furnish a peak power of 230 kilowatts to a W.E. 2J53 magnetron. The modulator was also to provide a range of pulse durations and repetition rates extending from 0.25 microseconds at 1600 pulses per second to 5.0 microseconds at 200 pulses per second.

In order to meet these requirements, two 1B31 tubes were used with a peak switch voltage of 8 kilovolts and a peak current of 75 amperes. By using a 300 mil spacing, a life greater than 500 hours was obtained at 200 pulses per second, 5 microseconds and 75 amperes. The other operating conditions were less severe from the life standpoint.

The wide spacing used meant a considerable increase in the size of the cath-

ode over the previous designs. To make this tube rugged, both electrodes were supported from large diameter kovar-to-glass seals. During assembly the cathode end of the tube was open so that a tool could be inserted to hold the anode concentric with respect to the cathode, while its supporting member was sealed to the glass. A cup was then brazed in to cover the cathode opening.

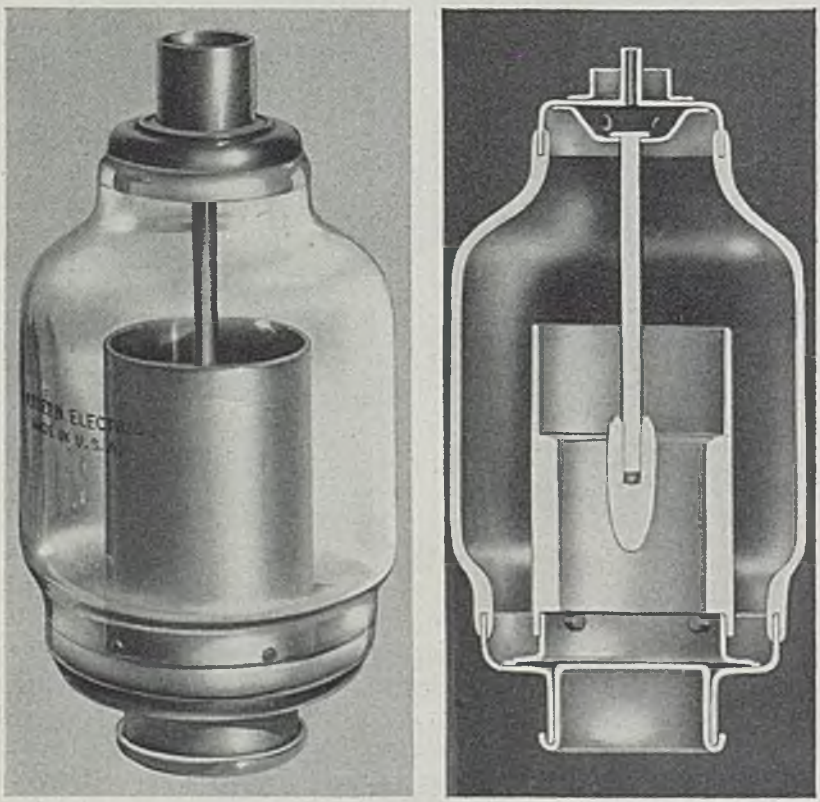


Fig. 23—Western Electric 1B31 spark gap tube.

Several hundred models of this tube were made in the laboratory and performed satisfactorily in the circuit. Due to circuit design changes, however, these tubes did not go into large scale manufacture.

W.E. 1B42

The 1B42 fixed gap tube departs considerably in design from the 1B22 and 1B29 in that mercury instead of aluminum was used as cathode. Its

construction is illustrated in Fig. 24. This tube was developed for radars which were for long range search on shipboard. In these modulators three tubes were used in series to switch a peak power of 0.8 megawatts and 1.4

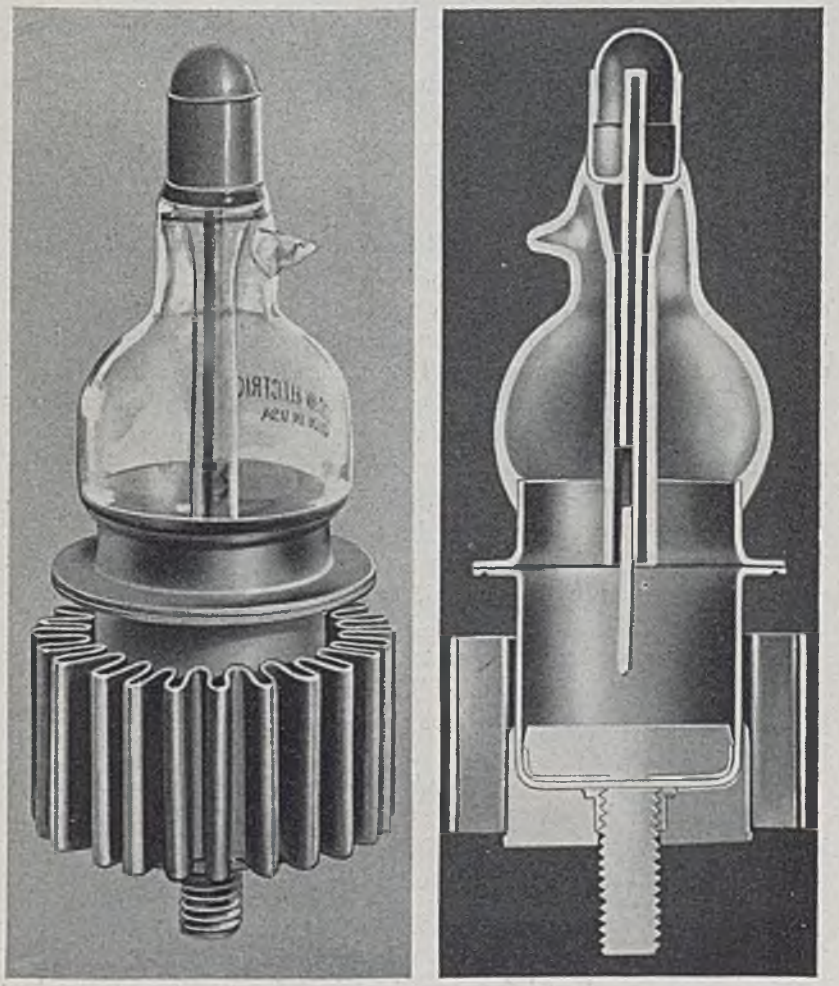


Fig. 24—Western Electric 1B42 spark gap tube.

megawatts, respectively, into high power triode oscillators. These modulators were to be capable of operation at either half or full power. The equipment was to be capable of withstanding the shock and vibration normally encountered on shipboard.

The series of gaps was required to operate with peak switch voltages

varying from 10.5 to 17.1 kilovolts, and to pass a maximum current of 200 amperes for 6 microseconds at a repetition rate of 180 pulses per second. They were also required to operate with 1.5 microsecond pulses at 600 pulses per second. The main electrical design problems were those of obtaining a wide voltage operating range and an adequate life with large peak currents and long pulses.

As discussed in II-(c), the use of an iron sponge mercury cathode with a molybdenum rod anode provided a wide voltage operating range as well as a long life with 200 ampere, 6 microsecond pulses. The mercury sponge cathode also met the vibration and shock requirements of shipboard operation.

In order to secure good wetting of the sintered sponge, which was essential to a long life, a special construction, as shown in Fig. 24, was used. The sponge was sintered directly into the bottom of a Kovar cup which had six radial vanes welded into it. These served to anchor the sintered material as well as to conduct the heat away from the center of the cathode. After the anode assembly and glass envelope were attached to the upper Kovar flange, the two sub-assemblies were welded together by means of a single ring weld. This allowed a minimum of handling of the sintered material and eliminated all glass work after the sintered material was inside the tube.

The processing of the tube consisted of first evacuating and then of heating the lower portion to 800°C while passing purified hydrogen through the tube. After the sponge had partially cooled, the mercury was introduced and wetting took place instantly.

Since the temperature of the center of the sponge must be kept below the boiling point of mercury, in addition to the internal vanes described, the Kovar cup was soldered into a block of copper to which was attached a folded copper radiator.

Several hundred models of the tube were made in the laboratory and delivered to the Navy and to equipment manufacturers. Full manufacturing information was turned over to the Navy which in turn issued a contract for the procurement of several thousand tubes.

Ratings

The ratings of the four different models of spark gap tubes developed by the Laboratories are summarized in Table 1. In order to permit the use of these gaps under a wide variety of operating conditions, yet prevent the simultaneous application of the maximum values of peak current, pulse duration, and repetition rate, a special system of rating was evolved. In addition to placing a maximum value on each of these three quantities a maximum value was also placed on the product of any two of these quan-

tities. For instance, one of these products would prevent the use of very high peak currents along with very long pulses, a combination which would give a very short life, especially with aluminum cathode gaps. Or another product would prevent the use of the tube at both high peak currents and high repetition rates, a condition which would not allow adequate deionization between pulses. The later types of gaps were rated in this manner.

(g) *Evaluation of the Fixed Gap as a Modulator Switch*

In order to compare the performance of fixed gaps in radar modulators with that of other switching devices, as well as to assess their future possibilities, we may consider them with respect to the following points.

TABLE I
RATINGS OF W. E. FIXED SPARK GAP TUBES

Tube Type	Repetition Rate—pps		Peak Current a Max.	Pulse Duration μs Max.	Micro-Coulombs per Pulse Max.	Operating Voltage Range—2 gap Ckt. kv		Operating Voltage Range—3 gap Ckt. kv		Peak Trigger Voltage 3 gap Ckt. kv Nominal	Peak Trigger Voltage 3 gap Ckt. kv Nominal	Voltage Required for Starting kV	
	Min.	Max.				Min.	Max.	Min.	Max.			2 gap Ckt. Min.	3 gap Ckt. Min.
1B29	500	2100	30	0.75	—	2.6	3.0	—	—	3.0	—	1.9	—
1B22	300	1100	75	0.75	—	3.8	5.4	—	—	5.0	—	2.5	—
1B31	200	1600	300	5.0	375	7.3	9.2	—	—	8.0	—	6.1	—
1B42	160	1500	300	6.1	1280	9.0	11.4	10.5	17.1	10.0	15.0	6.5	8.5

(More complete information on the above tubes is contained in the JAN Specifications for individual tubes.)

1) *Peak current*—The present coded tubes cover a range of currents from 20 amperes to 300 amperes. Experimental tubes have been tested up to 1000 amperes, and indications are that even larger currents are possible.
2) *Switch voltage*—The present tubes cover a range from 2.6 to 17.1 kilovolts. Experimental tubes have been tested up to 30 kilovolts.
3) *Peak power output*—With the limits of peak currents and voltages on the present tubes, power outputs of 25 kilowatts to 2.2 megawatts are possible. Experimental tubes were made which were capable of furnishing 15 megawatts. Much larger power outputs seem possible.
4) *Pulse duration*—The maximum range of pulse durations covered by any of the present tubes is from 0.25 to 6 microseconds. For pulses shorter than 0.25 microseconds the efficiency of these tubes would decrease rapidly. Pulse durations much greater than 6 microseconds, however, could probably be used if proper attention is given to cooling.

5) *Pulse repetition rate*—A range of 160 to 2100 pulses per second has been covered by coded tubes. Experimental tubes with very short gaps have been tested up to 10,000 pulses per second. However, the design of tubes for practical operation in this region would entail considerable effort.
6) *Operating voltage range*—Although a given set of tubes may exhibit a wide range of operating voltage on a laboratory test, the rated range must be considerably less because of manufacturing variations and changes during life. However, since most radar modulators operate at a fixed power level, this limitation is not a serious one.
7) *Trigger requirements*—The spark gap tubes require a high-voltage low-current trigger supply. While this is more difficult to obtain than the low-voltage supplies required by some other modulators, it caused no real difficulty in practice.
8) *Time jitter*—Although the time jitter of coded tubes is of the order of one microsecond, experimental tubes have been made which have, at the operating voltage, a jitter of the order of one hundredth of a microsecond.
9) *Efficiency*—The switching efficiency for all of the past applications of fixed gaps has been in the range of 90 to 96 percent, which makes the fixed gap one of the most efficient switching devices for radar.
10) *Simplicity of manufacture*—Since the unit type of fixed gap has only two elements of simple geometry, its manufacture is relatively easy.
11) *Dependability*—The dependability of the fixed gap has been demonstrated by its satisfactory performance in its extensive application.

Acknowledgments

The authors wish to acknowledge the valuable help given by many who cooperated in this project. In particular, they would like to thank Dr. S. B. Ingram for his criticisms and suggestions in regard to the development of coded sealed gaps. Also, we are especially indebted to Messrs. H. W. Weinhart and C. Depew for their invaluable help in the design and construction of the large variety of gaps required for this development.

Coil Pulsers for Radar

By E. PETERSON

RADAR systems in current use radiate short bursts of energy developed by pulsing a high-frequency generator, usually a magnetron. One means of developing the requisite impulses employs a non-linear coil and is termed a coil pulser. Such pulsers are found in substantial numbers among the Navy's complement of precision radars. Most fire control radars on surface vessels are equipped with them, and all modern radar installations on submarines are so equipped for search and for torpedo control.

History of Development

Coil pulsers had their origin in the magnetic harmonic generators first built for the telephone plant. Multi-channel carrier telephone systems in general use throughout the Bell System require numbers of carriers, harmonically related in frequency. These are derived from non-linear coil circuits[1] which convert energy supplied by a sine wave input into regularly spaced, sharply peaked pulses.

When development was started on precision radars, one of these circuits generating a power peak of a few hundred watts, several microseconds in duration, was adapted to the purpose.[2] Its output was shaped and amplified by vacuum tubes of sufficient power to key or modulate the ultra-high-frequency generator of the radar transmitter. All early fire-control radars were made up in this way; hundreds are still in use.

The next development of pulsers for fire-control radars was directed toward higher-powered pulses, shorter in duration for good range resolution. These had to be provided by a small package pulser, small enough and rugged enough to mount integrally with the magnetron and the antenna. The power rectifier was to be located at any convenient distance, and the rectified voltage had to be low enough to permit the use of standard low-voltage cables. These requirements put high vacuum tubes at a disadvantage in handling the finally developed pulses. Pulse transformers had not attained their present state of perfection in dealing with short pulses at this early stage and the pulser therefore had to work the magnetron directly.

[1] Magnetic Generation of a Group of Harmonics, by Peterson, Manley and Wrathall, *B.S.T.J.*, vol. XVI, p. 437, 1937.

[2] Fire-Control Radars, by Tinus and Higgins, *B.S.T.J.*, January, 1946.

One arrangement developed by W. Shockley to meet these requirements used a thyratron as a switch to generate pulses. High vacuum tubes were used at low voltages for comparatively long-time intervals in the driving circuit. Deficiencies of the thyratrons available at that time prevented the generation of pulse powers as high as required. With the earlier experience on low-level coil pulsers in mind, it was natural to think of using a non-linear coil for switching pulses at high level, in place of the thyratron. Much development was required to arrive at suitable circuits embodying the basic ideas, to build non-linear coils capable of withstanding high voltages, to proportion the circuit elements for efficient operation at specified powers and pulse durations, and to shape the output pulse to the desired flat-topped form.

This development resulted in a power pulser mounted in an oil-filled steel box, with associated high vacuum tubes of the sturdiest sort mounted externally, operated from a 1200 volt d-c. supply. It was suitable for installation integral with the antenna, and rugged enough to withstand gun blast and shock. Life of the pulser box components is long, and performance stable with time and temperature. The time of pulse emission is linked precisely to the input wave, practically independent of voltage and frequency variations over a suitable range. Such precision timing, or freedom from jitter, permits starting the indicator equipment in advance of the pulse emission time so that target ranges may be accurately measured. The power rectifier voltage is much lower than that of the pulse applied to the magnetron, and the pulser works directly into the magnetron without requiring an intermediary pulse transformer.

Subsequent developments left unchanged the general form of the circuit and its mounting, but were devoted to achieving various pulse widths, powers, and pulsing rates to suit different applications. Pulse widths covered a range from two-tenths to over one microsecond, peak powers ranged from 100 to 1000 kw, and pulsing rates ranged from 400 to 3600 pulses per second.

Non-Linear Coil Structures

An idea of the general form and makeup of non-linear coils used in various radar developments can be had from the photograph of Fig. 1. All cores shown there are made of molybdenum permalloy tape, one mil thick. Insulation is electroplated on the tape in a silicic acid bath, and the tape is wound in ring form. After the standard magnetic anneal of 1000°C in hydrogen, the coating of insulation a fraction of a mil thick adheres firmly to the tape.

The smallest coil shown in Fig. 1 seen just in front of the oil filled container in which it is mounted is used for low-power pulse generation. Its core weighing 7 grams is wound on an isolantite form.

The two larger coils shown are used in power pulsers. Their cores are made up of self-supporting rings. The smaller coil has a core weight of one kilogram and is used at voltages up to 25 kv. for the generation of power peaks of the order of 100–250 kw. Phenol fibre is used to support and position the core and winding. The larger core has a weight of 13 kg. and is used at a voltage of 40 kv. in a pulser generating power peaks of one megawatt. Glass-bonded mica and built-up mica are used for support and

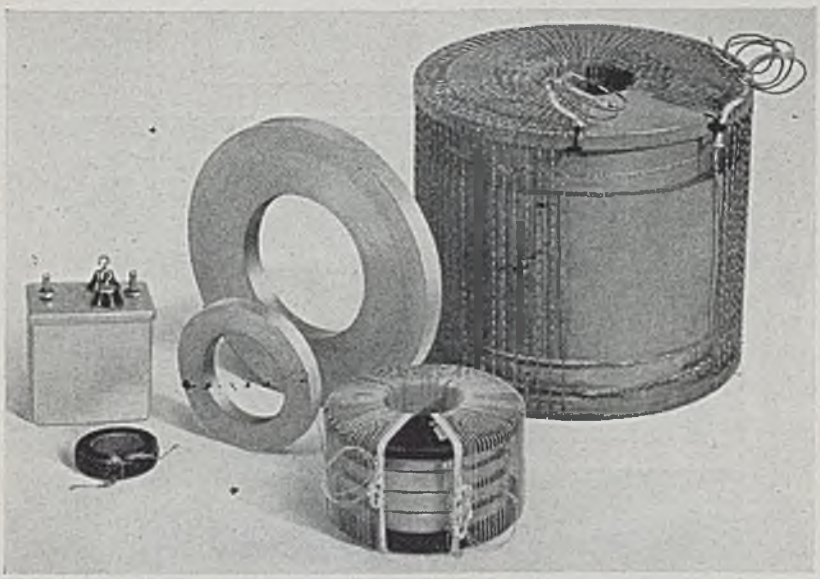


Fig. 1—Non-linear coils used in various radar transmitters. The smallest coil at the left, seen in front of its container, is used for low power pulse generation. The two larger coils are used in power pulsers developing 200 KW and 1000 KW peaks, respectively. The core rings of molybdenum permalloy tape are assembled into the coils shown.

positioning of the core rings and windings. The coils are assembled with other passive elements of the pulser network and the whole immersed in oil.

Operating principles of the two types of pulser circuits in which these coils are used are now to be discussed.

Low-Level Pulser

A schematic of the circuit used for developing low-power pulses is shown in Fig. 2a. Sinusoidal driving current (i_1) is introduced from the left, and a sharply peaked wave (i_2) is developed in the right-hand mesh. A resonant circuit (L_1C_1) serves to prevent dissipation of the generated pulse in the input mesh, and to tune out the input reactance at the driving frequency.

Capacitance and resistance elements (C_2R_2) in conjunction with the non-linear coil (L_2) make up the output mesh.

A complete cycle of the input wave is depicted in Fig. 2c, placed to correspond with the *B-H* loop of the non-linear core shown above it in Fig. 2b.

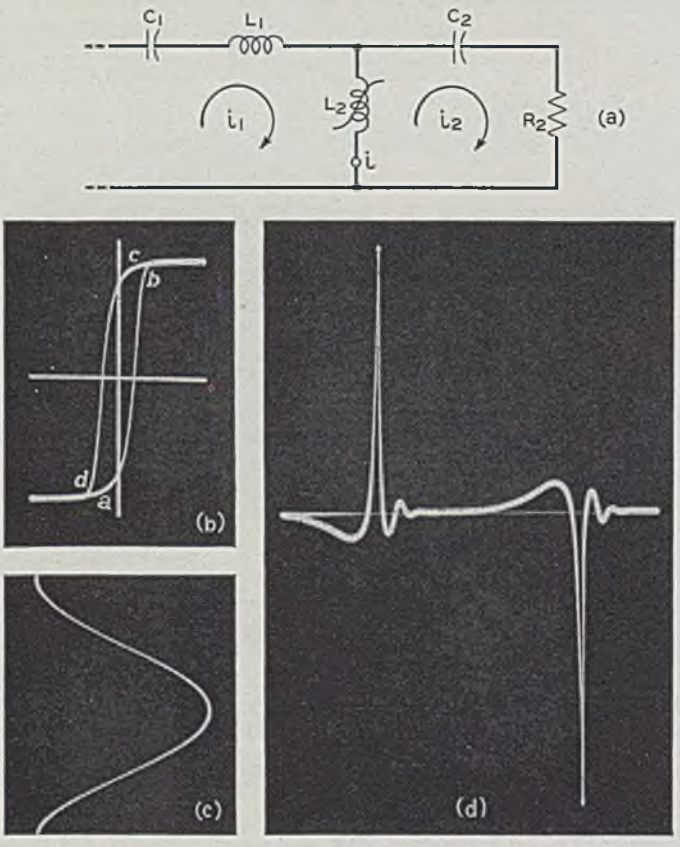


Fig. 2—Low-level coil pulser.

(a) Circuit diagram showing input tuning C_1L_1, non-linear coil L_2, output condenser C_2, and load resistor R_2.

(b) *B-H* loop of non-linear coil, with letters marking transitions between permeable and saturable regions.

(c) Sinusoidal input current wave scaled and placed to correspond with the horizontal scale of Fig. 2b.

(d) Pulsed output wave, i_2 as ordinate; i_1 as abscissa.

Action of the circuit is now to be followed throughout a cycle, starting with the input wave at its maximum negative excursion, condenser C_2 uncharged, and the core in its lower saturation region. Here the slope of the *B-H* loop and the corresponding differential permeability and inductance are

small. Hence the voltage drop across the coil is small. Little current flows in the output mesh, and practically all the input current flows through the coil. Matters are much different during the next interval in which the increase of current in L_2 brings the core into the permeable region *a-b*. Here the differential permeability is large so that part of the input current is diverted to the output mesh, charging the output condenser until upper saturation is reached at *b*. There the coil inductance falls to a low value, switching most of the condenser voltage across the load resistance. A current pulse accordingly develops in the output mesh lasting until the condenser charge is exhausted. The form of the current pulse shown in Fig. 2d approaches that of a highly damped sinusoid, and the pulse duration and magnitude are functions of the three elements of the discharge mesh. During the next half-period of the input wave, the same situation develops as in the first half-period, except that the corresponding currents and voltages throughout are reversed in sign.

According to this description the non-linear coil acts like a switch which automatically shifts the inductance from relatively high to relatively low values at specific coil currents. When the core is driven well into saturation, as is the case here, the ratio of these two inductances can be made large, usually in the neighborhood of several thousand. One feature of its action important from the efficiency standpoint is that the pulse occurs for the most part in the saturation region, where the contribution to eddy loss is small. The principal core loss occurs in the permeable region while the output condenser is charging, when variation of current through the coil occurs at a relatively slow rate.

In low-level radar applications the pulser output feeds a vacuum tube amplifier biased so that pulses of just one polarity are passed, the other oppositely poled pulse being cut off.

Since the range sweep of the radar receiver is initiated prior to pulse emission, the pulse should occur at a time linked precisely to the input wave. Otherwise the received pulse would be blurred introducing an uncertainty in measuring target range. No blurring (jitter) is visible with normal coil pulser operation. To get a measure of any variations which might be associated with core magnetization, tests were performed on a communication circuit in which jitter occurring at an audio rate would show up as noise. Measurements with a sensitive noise meter indicate the corresponding variation of pulse emission time to be smaller than 10^{-9} second.

Power Pulser

Operating Principles

The power pulser has the same type of discharge circuit as the low-level pulser just discussed. It differs in using a d-c. rather than an a-c. power

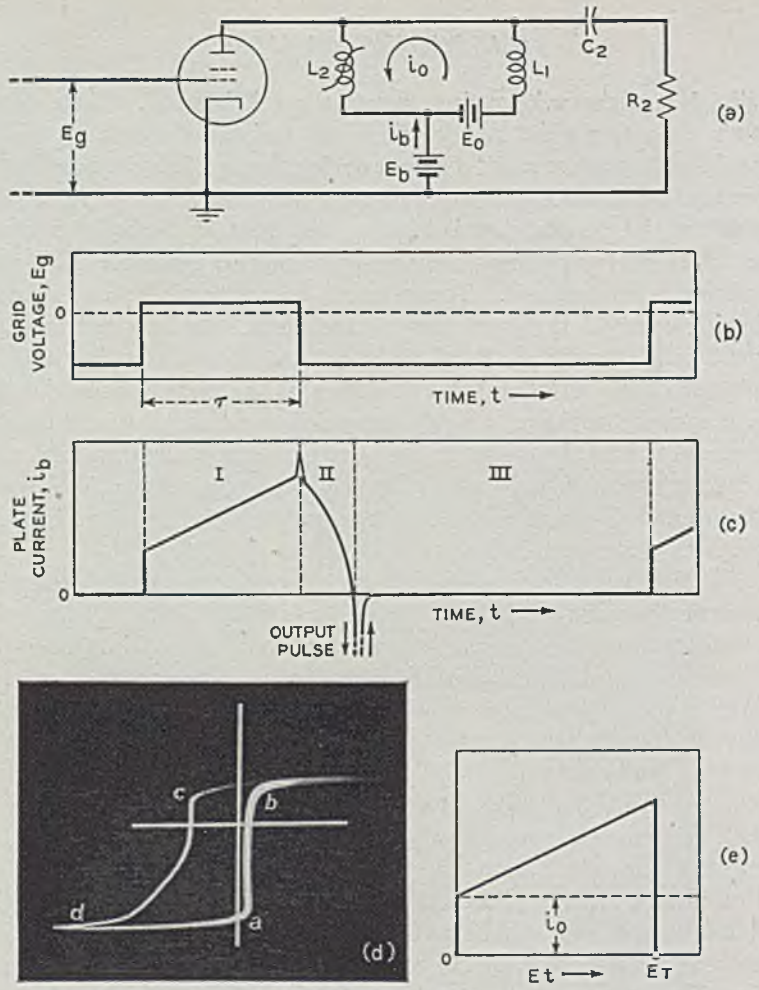


Fig. 3—Power pulser.

(a) Simplified circuit diagram showing charging tube at left, bias supply E_0, plate power supply E_b, linear coil L_1, non-linear coil L_2, output condenser C_2, and load resistor R_2.

(b) Rectangular wave of grid voltage impressed upon the tetrode of Fig. 3a. The tube conducts during the time τ in each cycle, and is cut off outside that interval.

(c) Plate current wave (i_b) corresponding to time scale of (b). During interval I, current is drawn through the paralleled inductors and the charging tube. At the end of this interval the tube is cut off and remains so until the start of the next cycle. During II, the magnetically stored energy is transferred to the condenser through R_2. At the same time the non-linear coil is brought toward saturation. During III saturation is reached; energy stored in C_2 is transferred to the load resistor through L_2 in a short pulse.

(d) *B-H* loop of non-linear coil used in the circuit of (a). Letters mark the most important transitions. During interval I magnetization proceeds from the lower left through *a* up to *b*; during II magnetization decreases past *c* down to *d*, and during III it extends far beyond the limits of the Figure to the left, returning to the neighborhood of *d* upon completion of the output pulse.

(e) Plot of current in linear coil during charging interval I against the product of coil voltage and time. Enclosed area represents energy stored in the linear coil. The rectangular area under the dashed line drawn through i_0 represents that part of the stored energy which varies with bias current.

source, and in charging the load condenser by a free, rather than by a forced oscillation. Energy for the free oscillation is taken from the d-c. source in a preliminary operation, in which energy is stored in a linear inductor. This preliminary operation consists in closing a d-c. path from the plate power supply through the linear inductor by means of a high vacuum tube, permitting current to build up with time. After a predetermined time has elapsed, the tube circuit is opened, the d-c. path is thereby interrupted, and energy stored in the inductor transfers to the load condenser. In this way the voltage to which the load condenser is charged can be made many times greater than the voltage of the plate power supply. The simplified circuit of Fig. 3(a) will serve to bring out salient operating features. Conduction of the tetrode at the left is controlled by a rectangular wave of grid voltage (Fig. 3b) developed by a multivibrator (not shown) which swings the grid from a potential below cutoff to one just above cathode potential. The plate power source E_b feeds two inductors in parallel, L_1 being linear, and L_2 non-linear. A small biasing voltage E_0 drives polarizing current i_0 through the two inductors in series.

The preliminary operation which serves to transfer energy from the main power source to the inductors is initiated when the tetrode grid is driven positive. Current from the main source builds up through the paralleled inductors and the tetrode as indicated on Fig. 3c, interval *I*. The region in which the non-linear coil works may be seen from the hysteresis loop of Fig. 3d. Its operating point is displaced to the left of the origin near *d* by the bias current. When the tetrode conducts, current in the non-linear coil rises rapidly at first in the lower saturation region until *a* is reached. The rise thereafter is comparatively small and slow in traversing the permeable region *a-b*, while at the same time current builds up in the linear coil at a much greater and practically uniform rate. When the core of L_2 reaches saturation near *b* its inductance again drops, preventing further rise of current in L_1. At this time the tetrode is driven below cutoff and remains out of the picture until the start of the next cycle.

The second interval, in which energy is transferred from the linear inductor to the load condenser, starts with the cutoff of tetrode current. This transfer is effected in an oscillation with frequency determined mainly by the paralleled inductors and the load condenser. In this interval II of Fig. 3c, current through the non-linear coil falls suddenly at first from *b* to *c* and then more slowly from *c* to *d*. The rate of change in region *c-d* is much greater than that in *a-b* as indicated by the fainter trace in Fig. 3d, so that eddy currents in the core are increased and the slope of the descending branch of the loop reduced correspondingly. Thus some of the energy previously stored in the linear inductor is used up in completing the magnetization cycle and this part, consequently, is not available for transfer to the load

condenser. The maximum voltage to which C_2 is charged in this interval is made much greater than that of the d-c. power source (E_b). The ratio of these two voltages depends upon the ratio of the inductance charging time in the preceding interval to the oscillation period. Both factors can be varied over wide limits, and step-up ratios of roughly ten to twenty are generally used.

The third interval starts with magnetization of the non-linear core near point d on the loop, where the inductance again drops. This situation is precisely the same as that previously described for the low-power pulser. As a result the condenser discharges through the load resistance at the time indicated in Fig. 3c, driving the core far into saturation with a field of the order of a hundred oersteds. This field extends too far to the left of point d to be shown in Fig. 3(d). Here the differential permeability approaches unity, and the correspondingly low inductance permits a rapid build-up of pulse current. Evidently but one pulse is produced each time the tetrode conducts, and the number of pulses produced per second is changed simply by varying the input frequency without requiring any circuit change, power dissipations permitting.

Energy storage in the linear coil depends upon its inductance, upon the bias current, and upon the peak current reached during the tetrode conduction interval. A plot of the current in L_1 against the product of time and of voltage across the coil permits this energy to be represented as an area (Fig. 3e). Evidently a given area can be made up by varying the relative sizes of its component triangle and rectangle, only the latter varying with bias current. If for example the bias is reduced to zero, the rectangle would vanish and the peak current would have to be increased to attain the original amount of stored energy. The higher maximum current requires more cathode emission of the tetrode and leads to greater plate power dissipation. Thus in addition to determining the energy stored, the amount of bias is one of the factors determining power dissipation capacity and emission which must be provided in the driving tube or tubes. Additional factors enter to make a bias corresponding to d (Fig. 3d) the most favorable from an efficiency standpoint.

The operating principles developed above in terms of a simplified circuit have been applied to a number of practical circuit forms which are described in the sections following.

Load Circuit

In radar applications the useful load is a magnetron which takes the place of the linear resistance previously considered. Since the magnetron viewed at its input terminals acts essentially like a negatively biased rectifier, additional means must provide for the flow of condenser charging current in a

direction opposite to that of the discharge pulse. This takes the form of a suitably poled diode shunted around the magnetron input terminals. After the main discharge pulse is completed, reactive elements are left with some little energy which tends to redistribute throughout the network. In course of redistribution, additional pulses of lower energy may occur shortly after the main pulse is completed. This tendency is a harmful one if the after-pulses are large, since echoes from short-range targets are obscured. Suppression of after-pulses is assisted by shunting around the diode-magnetron a linear inductance known as a clipping choke. This added inductance slows down the rate at which energy is redistributed, and permits the diode to fulfill its second function of dissipating the greater part of the residual energy. The shunting inductor, too, is made to fill a second function. Through provision of a bifilar winding, it passes heating current to the filament of the magnetron, thereby eliminating the need for high-voltage insulation otherwise required in the filament transformer.

Magnetic Bias

Several arrangements have been worked out for supplying various amounts of bias, some of them using a separate source, others being self-biased.[3] In general the use of external bias leads to a lower demand on the driving tetrode and is associated with pulse production at best efficiency. Circuits dispensing with an external bias source are that much more convenient in use, where the added tube demand and the lower efficiency corresponding can be handled without undue increase of the tube complement. In general the energy delivered to the magnetron is roughly 25 to 55 per cent of the plate energy input, with the higher figure applying to the higher outputs and external bias.

Transformer Coupling

In some cases it is convenient to equip the non-linear coil with primary and secondary windings providing voltage transformation or isolation to avoid adding a transformer for that purpose. The first case arises in the higher-powered pulsers, where the load condenser has to be charged to a voltage greater than the driving tetrode can withstand. For the Western Electric 5D21 tubes customarily used, voltage breakdown occurs near 20 kv, while condenser voltages in certain of the pulsers reach 30 and 40 kv. This situation calls for a step-up ratio from primary to secondary to fit the required potentials. The need for isolation may be illustrated by reference to Fig. 3a where the bias battery E_0 is shown maintained at the plate supply potential above ground. To avoid the resulting insulation problems in a

[3] One widely used circuit using a small amount of self-bias was developed by L. G. Kersta and E. E. Crump.

rectifier built to supply bias, a secondary winding is readily provided on the non-linear coil for connection to the linear coil and to the bias rectifier, which can then be maintained with one side at ground potential.

In either case whenever coupled windings are employed, the inside winding is invariably made to carry the discharge pulse. This provision results in minimum saturation inductance, since the inner winding is brought as close to the magnetic core as the voltage breakdown strength of the intervening dielectric permits. This winding is disposed as uniformly as possible around the core to avoid leakage which would add to the saturation inductance, and so limit the rate of current build-up in the pulse. The other winding can then be disposed with generous spacings, and with partial core coverage if desired.

Pulse Shaping

The oscillation frequency of the magnetron is determined primarily by its internal structure, although it is to some extent a function of the impressed potential. Departure of the driving wave from perfectly rectangular form permits the oscillation frequency to vary during the pulse, to an extent depending upon the size and duration of the departure and upon the characteristics of the magnetron.[4] Frequency modulation thus produced disperses energy over the spectrum. With the receiver band width limited to reduce noise and interference, one effect of this spreading of energy over the spectrum is to cut down the strength of the observed echo. For this reason, other things being equal, rectangularity must be approximated well enough to make the wasted energy a small fraction of that usefully employed.

It is convenient to regard the rectangular wave as synthesized by a series of harmonically related sine waves of appropriate magnitudes. The fundamental component according to this concept has a half period equal to the duration of the pulse, and the other components, progressively smaller in amplitude, have frequencies which are odd harmonics of the fundamental. In the low-power pulser with its rounded discharge wave the harmonic waves are quite small in amplitude. To approach the flat-topped discharge wave necessary in the power pulsers, harmonic components must be built up. This can be done by providing additional resonances in the discharge circuit at the wanted harmonic frequencies.

With the close spacing between circuit elements and their proximity to the pulser box walls, parasitic capacitances of appreciable magnitude add to those normally present. These involve dielectrics of low loss and, since the circuit elements and connecting wires are firmly fixed in position, they are fairly well reproduced. They can be used, therefore, in conjunction

[4] The Magnetron as a Generator of Centimeter Waves, by Fisk, Hagstrum, and Hartman, *B.S.T.J.*, April, 1946.

with added reactors of small size to provide harmonic resonances needed to shape the discharge pulse. These help to bring up the third and fifth, and in some cases higher harmonics.

Results after shaping are shown in Fig. 4 for two extremes of pulse width. The shorter pulse, roughly a quarter microsecond in average duration,

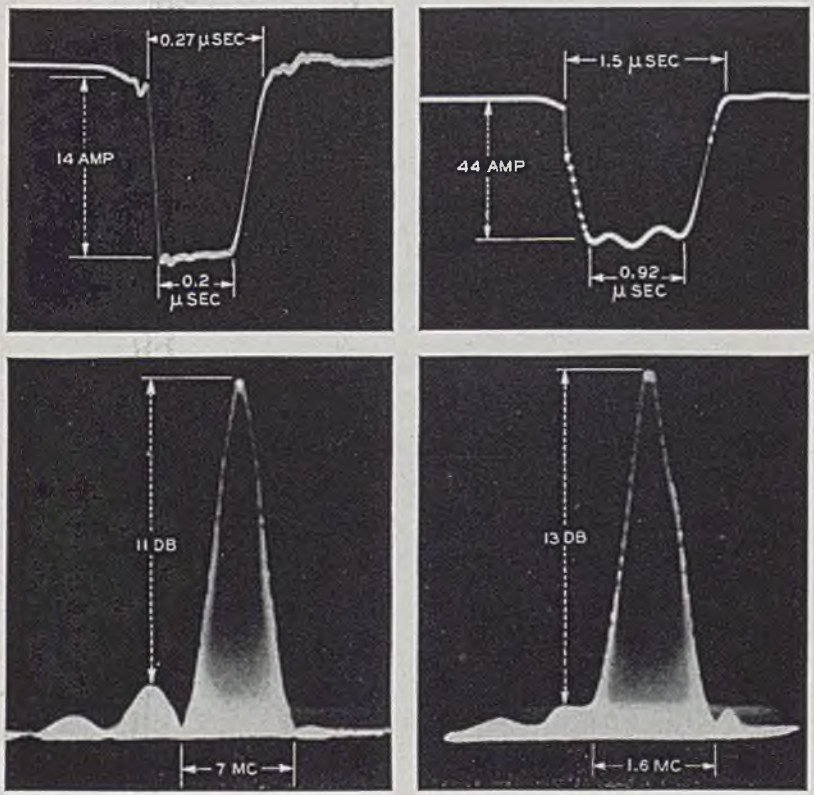


Fig. 4—Shaped magnetron current pulses, together with the radio frequency spectrograms corresponding. Pulse at upper left indicates presence of high harmonics; pulse at upper right shows strong fifth harmonic and little at higher harmonics. The band width of the main energy lobe, and the dispersion of energy outside that band in both cases indicate negligibly small effect attributable to frequency modulation.

evidences the presence of fairly high harmonics. The wider pulse, roughly one and a quarter microseconds in average duration, has a strong fifth harmonic and some even harmonics as well.[5] Below each pulse is shown a spectrogram of the corresponding magnetron high-frequency output, which

[5] Magnetron currents are shown rather than voltages, since current is a far more sensitive indicator of performance.

represents energy as a function of frequency. Different magnetrons are used with the two pulses; their operating frequencies and power capacities differ widely. Apparently frequency modulation exists in both cases to a small extent indicated by the departure of each spectrogram from symmetry about a vertical axis. Detailed study, however, shows that the band

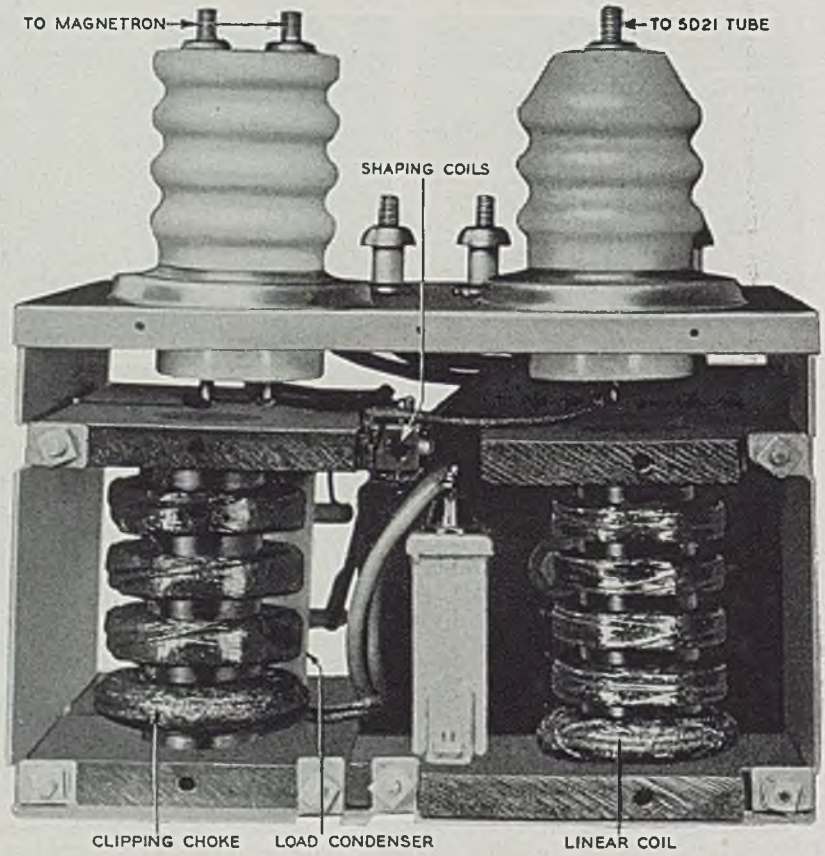


Fig. 5—Typical power pulser network.

width of the main energy lobe differs inappreciably from that with zero frequency modulation, and that the dispersion of energy attributable to frequency modulation is negligibly small. The pulses shown therefore provide satisfactory performance with their respective magnetrons.

Pulser Box

The form in which the typical power pulser network appears is shown in Fig. 5. Power peaks generated by the particular network shown are of the

order of 150 kw, with pulse durations of the order of a half microsecond. The non-linear coil here is similar to the one kilogram model pictured in Fig. 1; it is mounted on a panel back of the linear inductor indicated on the Figure. The two larger insulators are used to support high-voltage terminals, the double terminal at the left connecting to the cathode and heater of the magnetron and the single terminal at the right connecting to the tetrode plate. The smaller terminals provide lower-voltage connections including those to the plate power supply of 1000–1500 volts, the bias source where required, and the heating power supply for the magnetron. In use the network is sealed into a closely fitting oil-filled container.

ACKNOWLEDGMENT

The development of coil pulsers was a cooperative enterprise involving a number of different groups in the Laboratories. Design and engineering of the research models were the work of J. M. Manley, P. A. Reiling, L. R. Wrathall, W. R. Bennett, L. W. Hussey and E. M. Roschke. Magnetic cores were developed under the direction of R. M. Bozorth and E. E. Schumacher. Production models were engineered under the direction of F. J. Given. The achievement of successful coil pulsers, moreover, owes much to the efforts of W. H. Doherty and his radar development group.

Linear Servo Theory

By ROBERT E. GRAHAM

The servo system is a special type of feedback amplifier, usually including a mixture of electrical, mechanical, thermal, or hydraulic circuits. With suitable design, the behavior of these various circuits can be described in the universal language of linear systems. Further, if the servo system is treated in terms of circuit response to sinusoidally varying signals, it then becomes possible to draw upon the wealth of linear feedback amplifier design based on frequency analysis.

This paper discusses a typical analogy between electrical and mechanical systems and describes, in frequency response language, the behavior of such common servo components as motors, synchro circuits, potentiometers, and tachometers. The elementary concepts of frequency analysis are reviewed briefly, and the familiar Nyquist stability criterion is applied to a typical motor-drive servo system. The factors to be considered in choosing stability margins are listed—system variability, noise enhancement, and transient response. The basic gain-phase interrelations shown by Bode are summarized, and some of their design implications discussed. In addition to the classical methods, simple approximate methods for calculating dynamic response of servo systems are presented and illustrated.

Noise in the input signal is discussed as a compelling factor in the choice of servo loop characteristics. The need for tailoring the servo loop to match the input signal is pointed out, and a performance comparison given for two simple servos designed to track an airplane over a straight line course. The use of subsidiary or local feedback to linearize motor-drive systems, and predistortion of the input signal to reduce overall dynamic error are described.

1. Introduction

A SIMPLE servo system is one which controls an output quantity according to some required function of an input quantity. This control is of the "report back" type. That is, some property of the output is monitored and compared against the input quantity, producing a net input or "error" signal which is then amplified to form the output. The first statement defines the servo as a transmission system; the second, as a feedback loop. The problem of servo design is then to fashion the desired transmission properties while meeting the stability requirements of the feedback loop. This is the familiar design problem of the feedback amplifier.

2. The Servo Circuit

The design of linear feedback amplifiers has been developed to a high degree in terms of frequency response; that is, in terms of circuit response to sinusoidal signals.[1] The servo system is a special type of feedback amplifier, and usually can be made fairly linear. Thus, it is logical to analyze and design the servo circuit on a frequency response basis. Also,

[1] See "Network Analysis and Feedback Amplifier Design," by H. W. Bode, D. Van Nostrand Co., 1945.

servo systems usually are combinations of electrical, mechanical, thermal, or hydraulic circuits. In order to describe the behavior of these various circuits in homogeneous terms, it is desirable to recognize the analogous relationships established by similarity of the underlying differential equations. Before proceeding to a discussion of frequency analysis, a typical analogy between electrical and mechanical systems will be described.

2.1 *Electrical-Mechanical Analogy*

Confining the discussion to rotating mechanical systems, the analogy which will be chosen here puts voltage equivalent to torque, and current to rotational speed. This choice leads to the array of equivalents shown in Fig. 1; inductance, capacity, and resistance corresponding to inertia, compliance, and mechanical resistance, respectively. Charge is equivalent to angular displacement, and both kinetic and potential energy are self-analogous. The ratio of voltage to current, or torque to speed has the dimensions of resistance. In an interconnected electro-mechanical system, the ratio of voltage to speed or torque to current may be called a transfer resistance. Similarly, the ratio of voltage to angular displacement, or of torque to charge, is a transfer stiffness (reciprocal of capacity or compliance).

Some commonly used devices for coupling between electrical and mechanical circuits are shown in Figs. 2 and 3. The motor, Fig. 2a, is used to convert an electrical current i into a mechanical speed or "current" $\dot{\theta}$ $(= d\theta/dt)$. The electrical control current i is produced by the voltage difference between an applied emf e and a counter-rotational emf (not indicated), acting upon the total electrical mesh resistance R_e.* In the mechanical circuit, a torque proportional to i forces a "current" $\dot{\theta}$ through the mechanical load R_m, J.

An equivalent mechanical mesh directly relating shaft speed to the applied emf is shown in Fig. 2b. A fictitious generated torque $\mu_t e$ acts upon the mechanical load through an apparent mechanical resistance R_m'. μ_t is a transfer constant determined both by the motor properties and the electrical mesh resistance R_e. R_m' is similarly governed and is inversely proportional to R_e.

The motor may be compared to a vacuum tube having an amplification factor μ_t and a plate resistance R_m'. However, the motor is usually much more a bilateral coupling element than the vacuum tube, due to the effect of the counter emf upon the electrical mesh.

The potentiometer, tachometer, and synchro circuit shown in Fig. 3 are all means for converting a mechanical quantity to an electrical one. All three are substantially unilateral coupling elements. The potentiometer

* R_e includes both the source resistance and the motor winding resistance.

delivers an output voltage e proportional to its shaft angle θ. Thus, the ratio of e to θ is a transfer stiffness constant S_t. The synchro circuit consists of a synchro generator connected to a control transformer, and delivers an output voltage e proportional to $\theta_1 - \theta_2$, the angular difference between

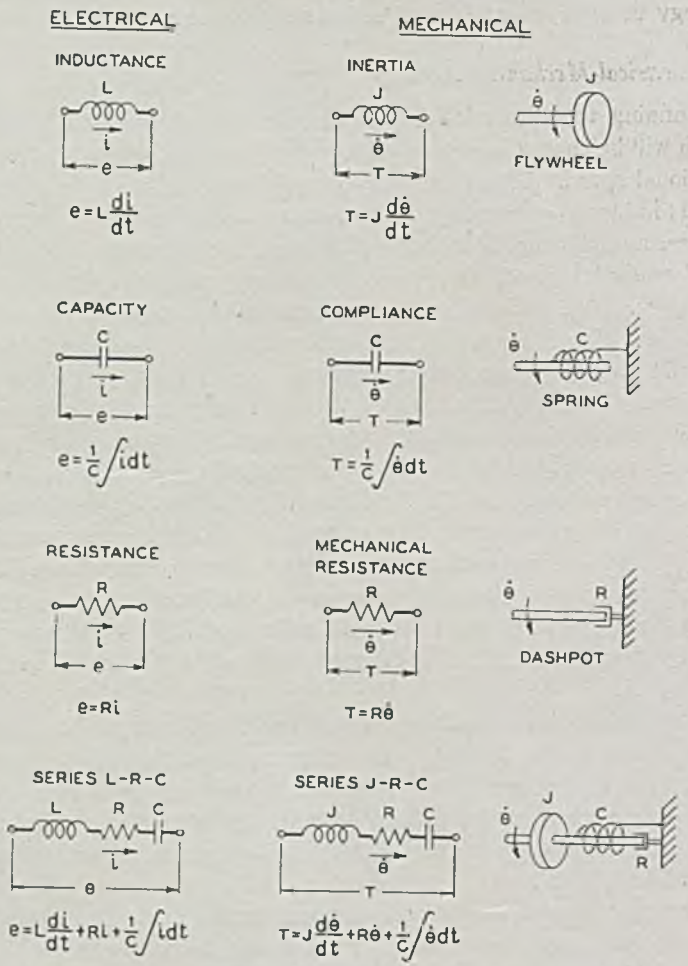


Fig. 1—Electrical-mechanical analogy.

the two shafts. Thus the action of the synchro pair is that of a combined transfer stiffness and differential. The tachometer is a generator which produces an output voltage e proportional to $\dot{\theta}$, the angular speed of its shaft. The ratio of e to $\dot{\theta}$ is a transfer resistance constant R_t.

There are many other specific devices used to convert from mechanical to electrical quantities. Most are equivalent to the potentiometer or synchro circuit, one such being a lobing radar antenna, which delivers a voltage proportional to an angular difference. A different, less widely used device is the accelerometer, a generator which delivers an output voltage proportional to the angular acceleration of its shaft. Its characteristic is that of a transfer inductance or inertia.

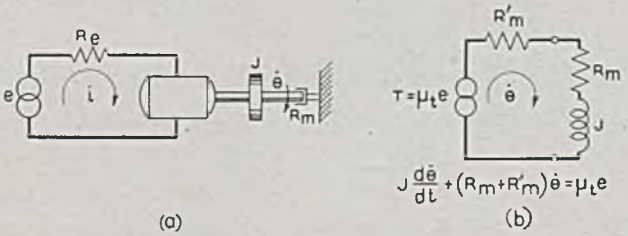


Fig. 2—Motor as a transfer device. (a) Motor and load. (b) Equivalent mechanical mesh.

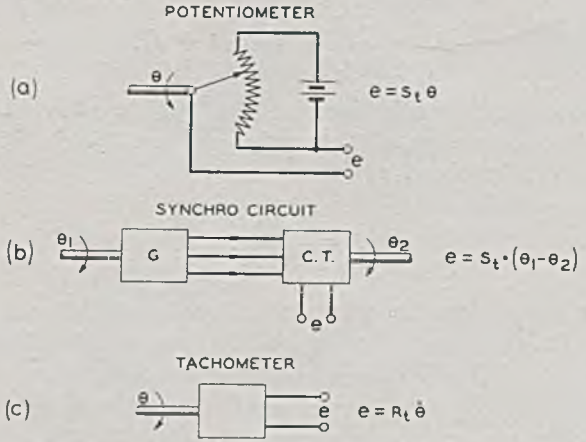


Fig. 3—Mechanical-electrical transfer devices.

2.2 *Frequency Analysis*

A brief review of the basic concepts of periodic analysis will be presented. It is assumed that the driving force applied to a network may be analyzed in terms of a series of sinusoidal components of various amplitudes, frequencies, and phases. The network response to each sinusoidal component is then evaluated, and the over-all result obtained by a summation of all such elementary responses. This is the formal procedure. Actually

it is often unnecessary to perform these precise operations in order to obtain a broad picture of the network behavior.

The method for determining the network response to a sinusoidal input is developed as follows. It is assumed that the circuit parameters are constant, independent of signal amplitude. Then, as indicated in Fig. 1, a single R-L-C or R-J-C mesh may be represented by a constant-coefficient linear differential equation. Choosing the electrical mesh for illustration,

$$(Lp + R + 1/Cp)i(t) = e(t),$$

where $p^n = d^n/dt^n$, $1/p = \int dt$, and $e(t)$, $i(t)$ are the mesh voltage and current respectively. This may be further abbreviated as

$$Z(p)i(t) = e(t), \tag{1}$$

where $Z(p) = Lp + R + 1/Cp$. For purposes of frequency analysis we are interested only in the forced or steady-state solution of (1), where $e(t)$ is a sinusoidal voltage $E \sin \omega t$. This steady-state solution is

$$i(t) = \frac{E}{|Z(j\omega)|} \sin (\omega t + \phi),$$

where $j\omega$ has replaced p in the function $Z(p)$, and the phase shift ϕ is the negative of the phase angle of the complex number $Z(j\omega)$.[2] This result is conventionally abbreviated as

$$I = \frac{E}{Z(j\omega)}, \tag{2}$$

where I is a complex number whose magnitude equals the peak amplitude of the current, and whose phase angle gives the associated phase shift. The function $Z(j\omega)$ is called the impedance of the mesh.

The relationship between torque, angular speed, and mechanical impedance is of course the same as expressed by (2). That is,

$$\theta = \frac{T}{Z(j\omega)}, \tag{2.1}$$

where θ is the complex peak amplitude of the sinusoidally varying speed, T is the peak amplitude of the applied sinusoidal torque, and $Z(j\omega)$ is the mechanical impedance obtained by substituting $j\omega$ for p in the operator $Z(p) = Jp + R + 1/Cp$. Since the angular displacement is the time integral

[2] ω is used to represent frequency in radians/sec, or 2π times frequency in cycles/sec.

of the speed, the expression for θ may be obtained by dividing both sides of equation (2.1) by p or, for the periodic case, by $j\omega$. Thus

$$\theta = \frac{T}{j\omega Z(j\omega)}. \tag{2.2}$$

The function $j\omega Z(j\omega)$ is the complex stiffness of the mechanical mesh. The phase shift of θ relative to θ is -90 degrees, as seen from a comparison of (2.1) and (2.2).

For an electro-mechanical network consisting of a number of interconnected meshes, a set of simultaneous differential equations of the type of (1) may be written. If $j\omega$ is substituted for p in these equations, there results a set of simultaneous algebraic equations which lead directly to the steady-state periodic solution. If a sinusoidal voltage or torque is applied at some mesh of the network, the resulting sinusoidal current or speed response in some other mesh is given by, using the notation of (2),

$$(\text{Response}) = \frac{(\text{Cause})}{Z_t(j\omega)},$$

where $Z_t(j\omega)$ for the chosen pair of meshes is obtained from the solution of the algebraic equations. $Z_t(j\omega)$ is called a transfer impedance, and may express the ratio of a voltage to current or speed, or of a torque to current or speed. The above relation also may be written as

$$(\text{Response}) = Y_t(j\omega) \cdot (\text{Cause}),$$

where $Y_t(j\omega) = 1/Z_t(j\omega)$ is called the transfer admittance between the two chosen parts of the network. In this form the response amplitude is obtained by multiplying the forcing sinusoid by $|\, Y_t(j\omega)\,|$, while the phase shift is given directly by the angle of $Y_t(j\omega)$. The transfer ratio between like or analogous quantities at two parts of the network is similarly a complex function of frequency, having the dimensions of a pure numeric.

Servo systems usually consist largely of elementary networks isolated by unilateral coupling devices (vacuum tubes, potentiometers, etc.). Thus, over-all transfer ratios often may be evaluated by taking the product of a number of simple "stage" transfer ratios, rather than by solving a large array of simultaneous equations. If the absolute magnitudes of the transfer ratios or "transmissions" are expressed in decibels[3] of logarithmic gain, both the over-all gain and phase shift of a number of tandem stages may be obtained by simple addition of the individual stage gain and phase values.

The transfer ratios of the conversion devices shown in Figs. 2 and 3

[3] The gain in decibels for a given transfer ratio is taken to be 20 $\log_{10}$ of the absolute value of the ratio.

may be written by inspection. Referring to Fig. 2b, the transfer admittance of a motor with resistance and inertia load may be written as

$$\frac{\dot{\theta}}{E} = \frac{\mu_t}{j\omega J + R_m + R'_m},$$

$$= \frac{\mu_t}{J} \cdot \frac{1}{j\omega + \omega_m}, \tag{3}$$

where

$$\omega_m = \frac{R_m + R'_m}{J}.$$

ω_m is the reciprocal of the time-constant of the motor and control mesh, and is 2π times the "corner" frequency at which the inertial impedance just equals the apparent mechanical resistance. Writing the transfer characteristic in terms of shaft position, rather than speed,

$$\frac{\theta}{E} = \frac{\mu_t}{J} \cdot \frac{1}{j\omega(j\omega + \omega_m)}. \tag{3.1}$$

For values of ω small compared with ω_m, θ/E is proportional to $1/j\omega$. This factor has a phase shift of -90 degrees and approaches infinity as ω approaches zero. This is merely a statement in frequency analysis language that the motor shaft angle is the time integral of the applied voltage, for slowly changing voltage. For more rapidly varying voltage, such that ω is large compared to ω_m, θ/E is proportional to $1/(j\omega)^2$ or $-1/\omega^2$, the angular variation being shifted -180 degrees with respect to the voltage variation.

The transfer ratio of the potentiometer, Fig. 3a, may be written as

$$\frac{E}{\theta} = S_t, \tag{4.1}$$

while for the tachometer, Fig. 3c,

$$\frac{E}{\theta} = R_t, \tag{4.2}$$

or

$$\frac{E}{\theta} = j\omega R_t. \tag{4.3}$$

Usually at some point in the system a compensating or "equalizing" network will be included to modify the transfer ratio of the basic components to the desired over-all transmission characteristic. Frequently this equalizer is incorporated in the electrical section of the servo because

of the comparative ease with which electrical circuit components may be assembled in desired combinations. The transfer characteristic of the equalizer may be simple or complicated, but in general may be written in the form,

$$\mu_e \sim \frac{(j\omega + \omega_1)(j\omega + \omega_3) \cdots}{(j\omega + \omega_2)(j\omega + \omega_4) \cdots}, \tag{5}$$

where the constants ω_1, ω_2, etc. may be real or complex. The synthesis of equalizing networks is a well known art and will not be discussed here, particularly since most of the equalization characteristics used in present servo systems can be realized with simple networks.

2.3 *Simple Servo System (Single Loop)*

The simple servo system may be divided into two basic parts, an amplifying circuit and a monitoring or comparison circuit. Such a division is

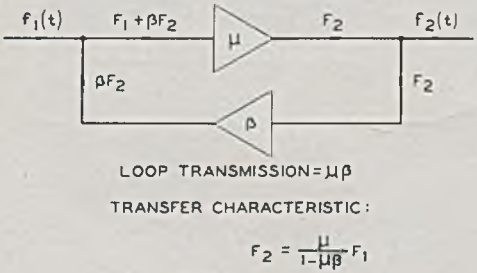


Fig. 4—Simple servo system.

indicated in Fig. 4, where μ and β are the transfer characteristics of the amplifying and monitoring parts, respectively. F_1 and F_2 represent typical sinusoidal components of the total input and output quantities $f_1(t)$ and $f_2(t)$,[4] while μ and β are complex-valued functions of $j\omega$ as described in the previous section.

The return signal βF_2 from the monitoring circuit is added to the servo input F_1 to form a net μ circuit input $F_1 + \beta F_2$. The servo transfer characteristic is found by setting

$$F_2 = \mu(F_1 + \beta F_2),$$

from which

$$F_2 = \frac{\mu}{1 - \mu\beta} F_1. \tag{6}$$

[4] That is, F_1 and F_2 are complex quantities employed in the same fashion as I in equation (2).

The closed system formed by the two basic circuits in tandem is of course a feedback loop, the loop transmission characteristic being given by $\mu\beta$.

Any desired form of servo transfer ratio may be obtained by an unlimited number of μ and β circuit combinations.[5] However, the β characteristic, which is usually determined by a passive network or an inherent property of a monitoring device, tends to be more stable with time and varying signal amplitude than that of the μ circuit, which may include vacuum tubes, motors, and other variable components. Consequently, it is desirable to have the servo transfer characteristic largely dependent upon the β circuit properties alone. This may be accomplished by making the loop trans-

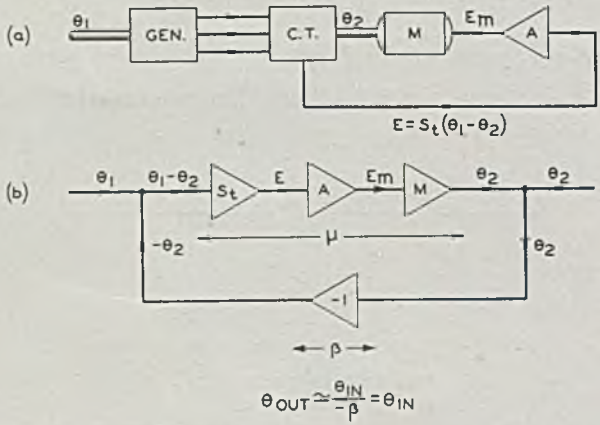


Fig. 5—Synchro follow-up system.

mission $\mu\beta$ very large compared to unity over the essential frequency range of the servo input signal $f_1(t)$. Under this condition, equation (6) becomes,

$$F_2 \simeq \frac{F_1}{-\beta}, \qquad |\mu\beta| \gg 1. \tag{6.1}$$

Thus the external transfer characteristic is set by β.* If, for instance, F_1 and F_2 are similar or analogous quantities and it is desired to have the servo output a replica of the input, β may be chosen as -1, yielding $F_2 \simeq F_1$.

It is not always easy to determine the basic parts μ and β of a servo by inspection of a schematic diagram of the system. An example is furnished by the synchro follow-up system shown in Fig. 5a. As previously discussed, the characteristic of the synchro comparison circuit is that of a differential

[5] Feedback stability requirements place certain restrictions on the permissible forms of $\mu\beta$. This will be discussed in the next section.

* The error arising from the approximate nature of (6.1) will be discussed in the next section as one type of "servo error."

transfer stiffness S_t, the voltage output of the control transformer being given by $S_t(\theta_1 - \theta_2)$. However, as seen from the modified diagram of Fig. 5b, the β characteristic is simply -1, the transfer constant S_t appearing in the μ circuit. Thus if the loop transmission is kept large, the essential relation demanded between θ_1 and θ_2 does not depend upon the value of S_t

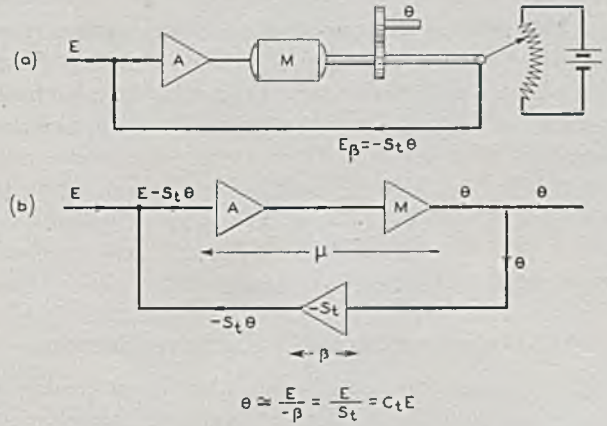


Fig. 6—Potentiometer loop.

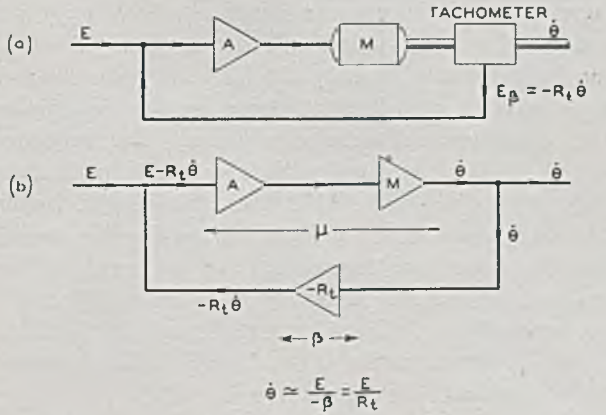


Fig. 7—Tachometer loop.

(as is obvious from physical considerations). This result also applies to a radar angle-tracking loop, where the received deviation or error signal is proportional to the difference between the angular coordinate of the target and that of the antenna system.

Figures 6 and 7 represent two servo systems where the input is electrical and the output mechanical. In Fig. 6 a potentiometer is used as a monitor-

ing device, the transfer stiffness in this case appearing in the β circuit. If θ is regarded as the output, then $\beta = -S_t$,* and the transfer characteristic is, for high loop gain,

$$\theta \simeq \frac{E}{-\beta} = \frac{E}{S_t} = C_t E,$$

where $C_t = 1/S_t$. Thus the over-all characteristic between input voltage and angular displacement is simply a transfer compliance constant. In Fig. 7 a tachometer monitor is used. Regarding angular speed $\dot{\theta}$ as the output, then $\beta = -R_t$, the transfer resistance of the tachometer. The transfer equation is thus

$$\dot{\theta} \simeq \frac{E}{-\beta} = \frac{E}{R_t} = g_t E,$$

where $g_t (= 1/R_t)$ is a transfer conductance.

3. Design of Simple Linear Servo Systems

The majority of servo systems in use, while often greatly extended in space and frequently including highly diversified transmission elements, may be represented by one essential feedback loop. However, a well designed servo often will incorporate numerous subsidiary or local feedback loops around stages of the system, in order to obtain a desired degree of linearity or performance with easily obtainable circuit components. Common examples of such local feedback loops are electrical feedback around vacuum tube amplifiers, and tachometer (velocity) feedback around motor-drive systems. These subsidiary feedback loops are almost always designed so that they are individually stable when the over-all feedback loop is opened (assuming the method employed for opening the over-all loop does not disturb the impedance terminations of the local feedback stages). If it is thus assumed that any subsidiary loops are individually stable, then the primary servo loop design may be treated simply as that of a single loop, whose over-all loop transmission is obtained by taking the product of the external transfer ratios of the various stages.

The design of a single loop servo may be divided into the design of the loop transmission $\mu\beta$, and one of the remaining parts, μ or β. As previously described, it is usually desirable to fix β according to the required basic input-output relationship of the servo, thus leaving $\mu\beta$ as a single characteristic to be chosen.

* It is assumed here that the transmission of the μ circuit is basically positive. The negative signs associated with S_t of Fig. 6 and R_t of Fig. 7 are then introduced (by poling) to make the loop transmission $\mu\beta$ essentially negative. This stipulation ensures what is commonly called "negative feedback," when the loop delay is zero.

As usual, specification of the form of $\mu\beta(j\omega)$ is beset by a series of performance objectives on the one hand and a set of physical limitations and restrictions on the other. Assuming the relationship expressed by (6.1) to be the required one, it would seem desirable to make $\mu\beta(j\omega)$ very large compared to unity at all frequencies. However, there are reasons why this is neither possible nor actually desirable. As the value of ω is increased, the loop transmission is eventually controlled by parasitic circuit elements such as distributed capacity and inductance in the electrical circuits, and distributed inertia, compliance, and backlash in the mechanical circuits. The effect of these parasitic elements at the higher signal frequencies is to cause $| \mu\beta |$ to decrease as a very high order of $1/\omega$ with increasing frequency. It will be shown, however, that feedback stability considerations require the loop transmission to be decreasing comparatively slowly through the frequency region where $| \mu\beta |$ is of the order of unity. Thus $\mu\beta$ must be reduced below unity at a frequency sufficiently low to avoid excessive contribution from the parasitics.

The presence of "noise" or undesired disturbances in the servo input signal is another compelling factor in the design of the loop characteristic. Input noise is harmful both in causing spurious output fluctuations and in overloading the power stages of the servo system. Both of these effects are reduced by narrowing the frequency band of the servo transfer characteristic. Referring to the expression for the transfer characteristic given by (6), it may be seen that a restricted transfer bandwidth may be obtained by reducing μ and $\mu\beta$ well below unity at a small value of signal frequency.[6]

On the other side of the picture is the requirement of fidelity in maintaining the desired input-output relationship. Undue narrowing of the transfer bandwidth of the servo results in large dynamic error, the magnitude of which depends both upon the character of the input signal and upon the chosen transfer characteristic.

The optimum design of a servo system, for a specified input signal and noise, thus is a compromise between dynamic error and output noise fluctuations, with stability considerations and parasitic circuit elements restricting the possible choice of loop transmission characteristics.

3.1 *Stability of Single Loop Systems*

The word *stable* as applied to a servo system is used here to imply a system whose transient response decreases with increasing time. It is possible

[6] When the β characteristic is under suitable design control, another method is available. Thus if β is made to rise in the frequency region of the desired transfer cut-off, and if $\mu\beta$ is maintained large beyond this region, (6.1) shows that the desired restriction is effected. For a given transfer characteristic, this cut-off method requires a wider frequency range for $\mu\beta$ and is thus more vulnerable to the effects of parasitic circuit elements. However, shaping of both the μ and β circuits permits a more rapid cut-off of the servo transfer characteristic than is possible with μ circuit shaping alone.

to determine the stability of a completed servo design by obtaining the transient solution of its differential equation. Though often very tedious, this is fairly straightforward. However, this method of procedure often is of little help either in guiding the initial design or in predicting the necessary changes, should the trial design be found unstable. The addition of even one circuit element to a design will generally create an entirely new differential equation whose solution must be found.

An alternative method for determining servo system stability, based on frequency analysis, furnishes the necessary information in a form which greatly facilitates the design procedure. This method is relatively simple to apply, even when the system has a large number of meshes and a high order differential equation, and the additive effects of minor circuit modifications are easily evaluated.

The stability of a single loop servo system—or of a primary loop, when the subsidiary loops are individually stable—may be investigated by plotting the negative of the loop transmission, $-\mu\beta(j\omega)$, on a complex plane for real values of ω ranging from minus infinity to plus infinity. (The negative sign is introduced because the loop transmission $\mu\beta$ is generally arranged to have an implicit negative sign, apart from network phase shifts. Thus $-\mu\beta$ is a positive real number when the network phase shift is zero.) *Then the necessary and sufficient criterion for system stability is that the resulting closed curve must not encircle or intersect* the point $-1,0$.* This type of plot is commonly called a Nyquist diagram, and is widely used in the design of electrical feedback amplifiers. An added stipulation is necessary if $\mu\beta(j\omega)$ becomes infinite at a real value of ω, say ω'. In this case an infinitesimal positive real quantity ϵ must be added to $j\omega$; that is, the function to be plotted is $\mu\beta(j\omega + \epsilon)$. This has no effect upon the plot except in the neighborhood of the singularity, where $\mu\beta(j\omega + \epsilon)$ is caused to traverse an arc of infinite radius as ω is varied through the value ω'.

As seen from (3.1), inclusion of a motor in a servo loop of the type shown in Figs. 5 and 6 will cause an infinite loop transmission at $\omega = 0$, assuming there is transmission around the remainder of the loop at zero frequency. The motor is the only commonly encountered circuit element capable of producing an infinite loop transmission at real frequencies.

In order to illustrate the use of the Nyquist diagram, a motor servo system of the type shown in Fig. 6 will be chosen. Again referring to equation (3.1), it may be seen that the transfer ratio of the motor and potentiometer is,

* This criterion is due to H. Nyquist, "Regeneration Theory," *B. S. T. J.*, January 1932. Detailed descriptions of stability criteria for single and multiple loop systems are given by Bode, loc. cit., and by L. A. MacColl, "Fundamental Theory of Servomechanisms," D. Van Nostrand Co., 1945.

$$\frac{E_\beta}{E_m} = \frac{-S_t \mu_t}{J} \cdot \frac{1}{j\omega(j\omega + \omega_m)}.$$

It is assumed that the amplifier includes an equalizing network such that the over-all amplifier characteristic is

$$A\left(\frac{j\omega + \omega_1}{\omega_1}\right)\left(\frac{\omega_2}{j\omega + \omega_2}\right)^3,$$

where A, ω_1, and ω_2 are positive real constants. The loop transmission is given by the product of these two transfer factors and thus may be written as

$$\mu\beta(j\omega) = -\frac{\omega_0}{j\omega}\left(\frac{\omega_m}{j\omega + \omega_m}\right)\left(\frac{j\omega + \omega_1}{\omega_1}\right)\left(\frac{\omega_2}{j\omega + \omega_2}\right)^3, \tag{7}$$

where ω_0 is a positive real constant given by

$$\omega_0 = \frac{AS_t \mu_t}{\omega_m J} = \frac{AS_t \mu_t}{R_m + R_m'}.$$

The three factors in parenthesis have been so grouped that they all approach unity for small values of ω. Thus the low-frequency behavior of $-\mu\beta$ is given by $\omega_0/j\omega$. This quantity has a pole at $\omega = 0$, so that it is necessary to plot the function

$$-\mu\beta(j\omega + \epsilon) \simeq \frac{\omega_0}{j\omega + \epsilon}, \tag{7.1}$$

in the neighborhood of $\omega = 0$. As ω is, in succession, a small negative quantity, zero, and a small positive quantity; the expression of (7.1) is correspondingly a large positive imaginary, a large positive real, and a large negative imaginary. Thus (7.1) traverses an infinite arc from the positive imaginary axis to the negative imaginary axis as ω increases through the value zero.

Assuming the numerical values

$$\begin{aligned} \omega_0 &= 200 \text{ sec}^{-1} \\ \omega_m &= 1 \quad \text{“} \\ \omega_1 &= 10 \quad \text{“} \\ \omega_2 &= 200 \quad \text{“} \end{aligned},$$

equation (7) may be rewritten as

$$-\mu\beta(j\omega) = \frac{200}{j\omega}\left(\frac{1}{j\omega + 1}\right)\left(\frac{j\omega + 10}{10}\right)\left(\frac{200}{j\omega + 200}\right)^3. \tag{7.2}$$

The phase angle of $-\mu\beta$ in degrees is, from (7.2),

$$B = -90 - \tan^{-1}\omega + \tan^{-1}\frac{\omega}{10} - 3\tan^{-1}\frac{\omega}{200}, \tag{7.3}$$

while the absolute magnitude is given by

$$|\mu\beta| = \frac{200}{\omega}\sqrt{\left(\frac{1}{1+\omega^2}\right)\left(\frac{10^2+\omega^2}{10^2}\right)\left(\frac{200^2}{200^2+\omega^2}\right)^3}. \qquad (7.4)$$

The Nyquist diagram of (7.2) is shown in Fig. 8. To emphasize the important features, radial magnitudes have been plotted on a logarithmic

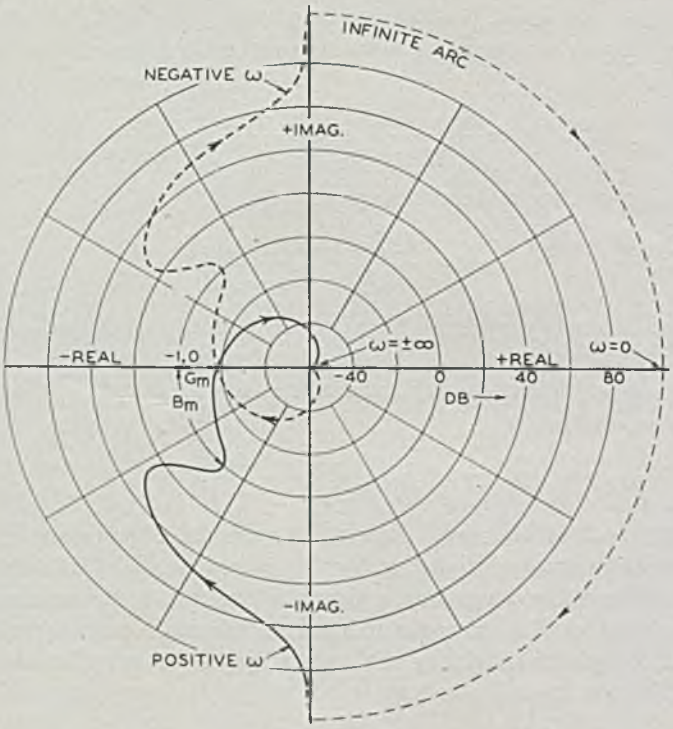


Fig. 8—Nyquist diagram of $-\mu\beta$.

scale.[7] The arrows indicate the direction of traversal as ω is varied from $-\infty$ to $+\infty$. The infinite arc traversed as ω varies through zero is indicated symbolically by the dotted semicircle in the right half plane.[8] As is the case for any physical system, the plot for negative values of ω is simply the mirror image of the positive frequency plot.

Since the polar plot does not encircle or intersect the "critical" point $-1,0$,

[7] Except in the immediate neighborhood of the origin, where a linear scale must be employed to plot the value $\mu\beta = 0$.

[8] The exact shape of this arc is of no consequence.

the system is seen to be stable.[9] From a practical standpoint it is necessary to know not only that a design is stable, but that it has sufficient margin against instability. The need for proper stability margin arises from two general considerations. First, the loop transmission of the physical system will vary with time due to aging, temperature changes, line voltage fluctuations, etc. Also the physical embodiment will depart from the paper design due to errors of adjustment and measurement, and to the effects of unallowed-for parasitic elements. Second, a design which is too near instability will have an undesirable transient response—large overshoots and persistent oscillations—and will unduly enhance noise in the input signal.

Stability margin is measured in a sense by the minimum displacement between the polar plot and the point $-1,0$. In feedback amplifier design, two numbers often are taken as a measure of margin against instability. These are called the *gain margin* and the *phase margin*. The gain margin, G_m, measures the amount, in decibels, by which the magnitude of $\mu\beta$ falls short of unity, at a phase angle of ± 180 degrees. The numerical value of gain margin for the system of Fig. 8 is about 18 db, which is the required increase in amplifier gain to make the servo unstable.[10] That is, this increase in amplifier gain would multiply the curve of Fig. 8 by a constant factor such that it would intersect the point $-1,0$. The phase margin, B_m, is equal to the absolute magnitude of the angle between $-\mu\beta$ and the negative real axis, at $|\mu\beta| = 1$. Figure 8 illustrates a phase margin of about 50 degrees. That is, if the points on the curve where $|\mu\beta| = 1.0$ were swung toward the negative real axis by about 50 degrees they would coincide with the point $-1,0$, and the servo would be unstable.

The type of transient response obtained with reasonable gain and phase margins is indicated in Fig. 9, which shows the response of the illustrative servo system to an input step. The initial overshoot is about 25%, and the oscillation damps out very quickly. For the general case, (6) may be rewritten in the form

$$F_2 = \left[\frac{-\mu\beta}{1 - \mu\beta}\right] \cdot \frac{F_1}{-\beta}. \tag{8}$$

The relation $F_2 = F_1/-\beta$ may be regarded as the desired one, with the bracketed factor acting as the inevitable modifier. Then if the quantity

[9] With more complicated systems it may not be obvious whether or not the plot encircles $-1, 0$. A simple test employs a vector with its origin at the $-1, 0$ point and its tip on the curve. If the vector undergoes zero net rotation as it traces along the curve from $\omega = 0$ to $\omega = \infty$, the curve does not encircle the critical point.

[10] In some servo systems a *decrease* in amplifier gain also may cause instability. Such systems are still covered by the polar plot criterion of stability, and are commonly called "Nyquist stable," or "conditionally stable."

$-\mu\beta$ exhibits gain and phase margins of the order of 10 db and 50 degrees respectively, the transient response of the modifying factor to a step function will be well-damped and generally not overshoot more than about 25%. If the gain margin is sufficient, the phase margin usually will be the dominant factor in determining the size of the initial overshoot. The required phase margin for critical damping depends upon the exact shape of $\mu\beta(j\omega)$, but in general is about 60 degrees. The gain margin needed in a particular design will depend upon the expected variability of the loop transmission. Radar tracking loops should usually have gain margins of the order of 15

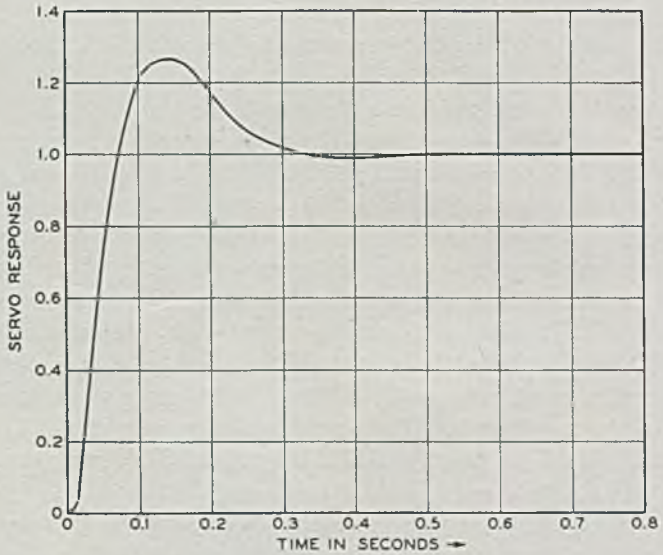


Fig. 9—Transient response of illustrative servo system.

db or more because of the large number of factors which may cause the loop gain to vary.

While the polar diagram gives a clear picture of stability considerations, it is usually more convenient for design purposes to plot the gain and phase of $-\mu\beta$ as separate curves on a logarithmic frequency scale, for positive values of ω. This is illustrated in Fig. 10, for the sample servo system. Under two commonly met conditions, the requirement for single loop[11] stability on this type of plot is simply that the absolute value of phase shift be less than 180 degrees at zero db gain ($|\mu\beta| = 1$). The conditions are that the connective polarity be such as to make $-\mu\beta$ positive when the

[11] Again, multiple loop systems may be included if all subsidiary loops are individually stable.

network phase shifts vanish, and that the gain curve cross zero db at only one frequency.[12]

An advantage of this logarithmic diagram is that commonly encountered forms of $|\mu\beta|$ vary as $\omega^{\pm n}$ for intermediate or asymptotic frequency regions, and thus plot as corresponding straight line segments. From (7.4) it may be seen that the illustrative form of $|\mu\beta|$ behaves, in turn, as $200\omega^{-1}$, $200\omega^{-2}$, $20\omega^{-1}$, and $1.6 \times 10^{8}\omega^{-4}$, as ω is increased. These asymptotic lines are drawn in lightly in Fig. 10, the actual gain describing smooth transi-

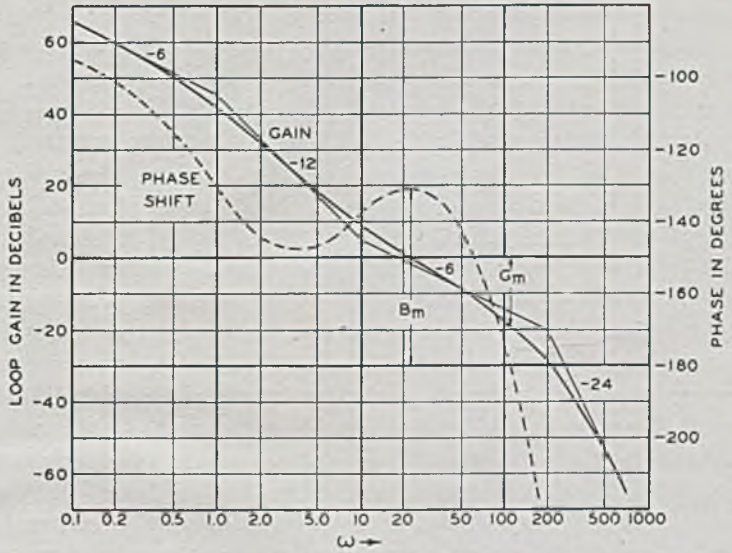


Fig. 10—Loop characteristic of illustrative servo system.

tions between adjacent asymptotes. Since the logarithmic slope of $\omega^{\pm k}$ is $\pm 6k$ db/octave,[13] the successive asymptotic slopes in Fig. 10 are -6, -12, -6, and -24 db/octave. The junctures of adjacent asymptotes occur at values of ω of 1, 10, and 200. These juncture frequencies are called "corner" frequencies, and may be seen from (7.2) to coincide with the real constants or "roots" added to $j\omega$ in each of the three factors in parentheses. The corner associated with the factor $200/j\omega$ is of course at $\omega = 0$. The corner of the last, or cubed factor is a multiple one, joining two asymptotes differing in slope by 18 db/octave. From a knowledge of such corner

[12] This discussion assumes that $\mu\beta$ has a low-pass configuration; that is, only the high frequency cut-off is considered. If $\mu\beta$ has a ow-frequency cut-off also, then a corresponding requirement identical with the above must be added.

[13] An octave is taken to be a 2:1 span in frequency, and 6 db is of course very closely equivalent to a 2:1 increase in $|\mu\beta|$.

frequencies, and the fact that the actual gain curve lies 3 db from an isolated simple asymptotic corner, the gain curve can usually be drawn in without further computation.[14] The phase curve also is easily constructed by adding up the elementary phase curves associated with the various corners. As may be seen from (7.3), these component phase curves all will have the same shape on a logarithmic frequency plot, merely being shifted along the frequency scale. The phase contributed by each simple corner will be $\pm$ 45 degrees at the corner frequency, the sign depending upon whether the associated root appears in the numerator or in the denominator.

It is an extremely important fact that the very requirement of stability imposes an unambiguous interrelationship between the gain and the phase shift of most types of transfer characteristic! By the general mathematical methods leading to the previously discussed stability criteria, Bode[15] has shown that this is true for the broad class of network structures commonly used in feedback loop design. That is, if either the transfer gain or phase shift is specified *at all frequencies*, the attendant phase or gain can be computed without further information. This class of networks is called *minimum phase*. Any stable structure composed of lumped circuit elements will have a transfer characteristic of the minimum phase type, provided it does not include an all-pass section.[16] All-pass characteristics are seldom used in the design of feedback loops, since their inclusion in the loop always reduces the stability margins achievable with a given high-frequency cut-off. Thus the unique interrelationship between phase and gain may be assumed for the loop characteristic $-\mu\beta$ in single-loop feedback systems. The nature of this relationship is discussed in detail by Bode. Briefly, the phase shift at any frequency ω_c is proportional to a weighted average of the gain slope in db/octave, over the entire logarithmic frequency scale. The weighting factor sharply emphasizes gain slopes in the immediate vicinity of ω_c, while the contributions of gain slopes at remote frequencies are reduced in proportion to the logarithmic frequency span from the particular frequency ω_c.[17] For transfer characteristics of the form $\omega^{\pm k}$, having a constant gain slope of $\pm 6k$ db/octave,[18] the associated phase shift is also constant and equal to $\pm 90k$ degrees. For transfer functions which behave approximately as $\omega^{\pm k}$ over a finite frequency span, the phase shift

[14] The corner frequency concept is less useful if the roots are complex. However a great many servo systems are so constructed that $\mu\beta$ has only real roots.

[15] Loc. cit. Also see "Relations between attenuation and phase in feedback amplifier design," by H. W. Bode, *B. S. T. J.*, July 1940, p. 421.

[16] An all-pass section is one which has constant gain but varying phase shift versus frequency, and is usually composed of a lattice, bridged T, or other bridge type circuit.

[17] About 60% of the area under this weighting function lies between $\omega = 0.5\ \omega_c$ and $\omega = 2\ \omega_c$, 80% between 0.25 ω_c and 4 ω_c.

[18] That is, for transfer characteristics whose *absolute magnitude* is given by $\omega^{\pm k} \cdots$.

of $\pm 90k$ degrees is approached more and more closely as the length of span is increased.

This may be observed qualitatively from the transfer characteristic of Fig. 10. For $\omega \ll 1$, the gain slope is -6 db/octave, and the phase shift approaches -90 degrees. For $1 < \omega < 10$, the average gain slope is about -10 db/octave, and the phase shift near $\omega = 3$ is -148 degrees (instead of $-90 \times 10/6 = -150$ degrees). As ω increases toward 200, the phase shift increases rapidly due to the asymptotic slope of -24 db/octave, finally approaching -360 degrees ($-90 \times 24/6$) for $\omega \gg 200$.

Foreknowledge of the inevitable gain-phase relationship is of great value to the servo designer, in making clear the comparatively small class of realizable gain-phase combinations and thus averting attempts at non-physical designs. For example the design use of too-rapidly falling loop gain characteristics in the region of the high-frequency gain cross-over (that is, near zero db loop gain) is not permissible because of the large negative phase shifts which must accompany the steep gain slopes. Another way of stating the advantage of an early realization of the gain-phase laws is to say that the designer is assured in advance that any paired gain and phase characteristics which he chooses within the basic restrictions will be achievable with stable physical networks.[19]

3.2 *Dynamic Error*

A servo system is usually designed to transmit some class of input functions with a required degree of fidelity. This class of functions may reduce substantially to one specific input signal whose time variation or whose frequency spectrum is known, or it may include a great variety of signals which have certain properties in common. In the latter case it is conceivable that definite limits may be placed upon the allowable amplitude ranges of the input signal and its various time derivatives, or certain limiting frequency spectrum characteristics may be specified for the input function.

Servomechanisms are subject to several types of transmission error. The systematic error, or dynamic error, is predictable from knowledge of the *noise-free* input signal and of the transfer frequency characteristic of the servo system. For simplicity, the discussion of error will be limited to the case where the output signal is desired to be a replica of the input, and where $\beta = -1$. Thus the loop transmission $\mu\beta$ becomes simply $-\mu$. The input-output relationship as given by (6) is therefore

$$F_2 = \frac{\mu}{1 + \mu} F_1, \tag{9}$$

[19] With some necessary reservations as to practicable dissipation constants and parasitic circuit constants.

where F_1 and F_2 are again typical sinusoidal components of the input and output respectively. Thus the corresponding *sinusoidal error component* may be written as

$$\Delta = F_1 - F_2 = \frac{F_1}{1 + \mu}. \tag{9.1}$$

The methods which may be used to determine the actual dynamic error $\Delta(t)$ from (9.1) depend both upon the nature of $f_1(t)$ and the type of information available about $f_1(t)$. If the input signal is a known periodic function, $\Delta(t)$ may be found by applying (9.1) for each sinusoidal component of the input and summing the resulting terms. If the input is non-periodic in character, then the error may be calculated from the Fourier integral expression

$$\Delta(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{F_1(\omega)}{1 + \mu} e^{j\omega t} \, d\omega, \tag{10}$$

where $F_1(\omega)$ represents the continuous frequency spectrum of $f_1(t)$, as obtained from

$$F_1(\omega) = \int_{-\infty}^{\infty} f_1(t) e^{-j\omega t} \, dt. \tag{10.1}$$

The problems of calculating $F_1(\omega)$ from $f_1(t)$ and $\Delta(t)$ from $F_1(\omega)$ often may be avoided by consulting well-known tabular lists of paired time and frequency functions.[20]

Equation (9.1) may be used as a broad guide in selecting the type of μ characteristic best suited to a particular input signal. It has been mentioned previously that because of input noise and parasitic circuit elements, the servo transfer bandwidth usually should be kept as narrow as possible, consistent with dynamic error requirements. The transfer characteristic $\mu/(1 + \mu)$ will be closely equal to unity while $|\mu| \gg 1$, will rise slightly[21] in the region where $|\mu| \backsimeq 1$, and fall off as μ when μ is small compared with unity. The "cross-over" frequency, for which $|\mu| = 1$, may be taken as a rough measure of the transfer bandwidth. Thus, the requirement of minimizing the bandwidth may be restated as that of minimizing the cross-over frequency, while holding the dynamic error within specified limits. Reasoning in a general way, this requirement may be met by designing μ so that the amplitudes of the sinusoidal error components, as given by

[20] An excellent list is given by G. A. Campbell and R. M. Foster in a Bell System monograph "Fourier Integrals for Practical Application," September, 1931. A table of Laplace Transforms, which also may be used, is given by M. F. Gardner and J. L. Barnes in "Transients in Linear Systems," John Wiley and Sons Inc., 1942.

[21] Assuming a phase margin of the order of 60 degrees.

(9.1), are roughly constant with frequency over the servo band. *This demands that μ have somewhat the same frequency distribution as the input signal spectrum* (for $|\mu| \gg 1$). Because of stability requirements and complexity of the necessary apparatus, this rule can usually be followed over only a part of the servo frequency band, especially when the input signal spectrum falls off very rapidly with increasing frequency. However, even a rough adherence to this desired relation is usually of real worth in reducing the noise errors of the servo. An illustration of this will be given in a later section.

3.21 *Approximate Calculation of Dynamic Error*

Frequently the servo requirement is to transmit, with great accuracy, a type of signal whose frequency spectrum falls off very rapidly with increasing frequency. As may be seen from (9.1) this demands very large values of loop transmission μ at the lower frequencies where the input signal energy is concentrated, but permits a rapidly dropping loop transmission versus frequency commensurate with the falling amplitude spectrum of the input signal. Such a rapid reduction in loop gain is practicable while $|\mu| \gg 1$. However stability considerations force a more gradual gain reduction as the region of gain cross-over is approached. As a result, contributions to the servo error from this frequency region may be neglected compared with those from the lower frequencies. This suggests a series expansion of (9.1) in the form,

$$\Delta = [a_0 + a_1(j\omega) + a_2(j\omega)^2 + a_3(j\omega)^3 + \cdots] F_1 , \tag{11}$$

where a_0 , a_1 , etc. are real constants.

Because of the assumed rapid drop in component amplitude F_1 with increasing frequency it is often unnecessary to take account of more than a few terms of the expansion.[22]

It is easy to show that (11) may be rewritten on a time basis to give the total dynamic error as

$$\Delta(t) = a_0 f_1(t) + a_1 \dot{f}_1(t) + a_2 \ddot{f}_1(t) + \cdots , \tag{12}$$

where $(\dot{}) = d(\ )/dt$. Thus the coefficient a_0 gives the error component proportional to input displacement. Similarly, a_1 and a_2 are the coefficients of the error components due to input velocity and input acceleration, respectively. For a great many motor-drive servo systems the loop transmission μ approaches infinity as $1/j\omega$ when ω approaches zero. This en-

[22] The series may be said to converge rapidly in a practical sense, for the following reason: For small values of ω the higher order terms are negligible. For values of ω sufficiently large that the high order terms may no longer be neglected the coefficient F_1 has become so small as to make the contribution of the entire series negligible.

Again for the rapidly converging case, this system will have principally an *acceleration* error.

Type 3. $-6, -12$ *db/octave,* $\mu = \omega_0\omega_1/j\omega(j\omega + \omega_1)$

This is perhaps the most commonly encountered characteristic in simple servos. The corresponding error expansion is

$$\Delta(t) = \frac{1}{\omega_0}\dot{f}_1(t) + \frac{1}{\omega_0\,\omega_1}\ddot{f}_1(t) - \frac{2}{\omega_0^2\,\omega_1}\dddot{f}_1(t) - \cdots, \qquad (\omega_0 \gg \omega_1). \quad (12.3)$$

and the principal error for this type system thus is a combination of *velocity* and *acceleration* components. Either the velocity or the acceleration error component may be predominant, depending upon the various parameters.

3.3 *Noise Errors*

The typical sinusoidal component of servo error due to noise (unwanted signals or irregularities) in the input signal may be written as*

$$\Delta_n = \frac{\mu}{1+\mu} N, \qquad (13)$$

where N represents the corresponding sinusoidal component of the input noise. If the noise signal $n(t)$ is known, the total noise error $\Delta_n(t)$ may be calculated from (13) in the ways described for the dynamic error. However, the noise input is seldom known in this sense, although certain outstanding components sometimes may be estimated and their effects evaluated. On the other hand the average disturbance due to random input noise, of the kind described as "thermal noise" in electrical circuits, may easily be calculated. This type of noise has constant amplitude versus frequency, and the total power in the output noise error may be found from

$$P_n = K \int_0^\infty \left|\frac{\mu}{1+\mu}\right|^2 d\omega, \qquad (14)$$

where K is a constant dependent upon the input noise power.

Input noise also causes overloading of the power amplifier and overheating of the motor. These effects are aggravated by the falling transfer characteristic versus frequency of the motor, as seen from the following discussion. The servo transfer characteristic is maintained approximately at unity out to the cross-over frequency. However the transfer ratio of the motor, equation (3.1), will be falling at least at 6 *db*/octave, usually at 12 *db*/octave, at frequencies below this point.[23] Thus the transfer

* Again assuming $\beta = -1$.

[23] Assuming that the mechanical load impedance is a series combination of resistance and inertia.

sures that a_0 and thus the displacement error will be zero, leaving principally the velocity and acceleration errors to be considered.

The coefficients a_0, a_1, a_2, etc. may be calculated easily for any particular case. For illustration, the three common forms of μ characteristic shown in Fig. 11 will be examined. (As previously discussed, the designated forms of μ need hold only for $|\mu| \gg 1$.)

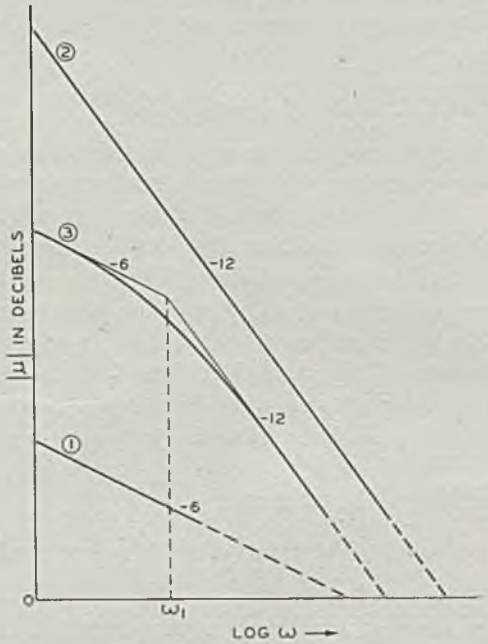


Fig. 11—Elementary μ forms.

Type 1. -6 *db/octave,* $\mu = \omega_0/j\omega$

The error expansion becomes,

$$\Delta(t) = \underline{\frac{1}{\omega_0}\dot{f}_1(t)} - \frac{1}{\omega_0^2}\ddot{f}_1(t) + \cdots . \tag{12.1}$$

For the combination of high accuracy and rapidly converging input spectrum, the first term is the only one of importance. Thus this type of system has essentially a *velocity* error.

Type 2. -12 *db/octave,* $\mu = (\omega_0/j\omega)^2$

Here the error is

$$\Delta(t) = \underline{\frac{1}{\omega_0^2}\ddot{f}_1(t)} - \frac{1}{\omega_0^4}\dddot{f}_1(t) + \cdots . \tag{12.2}$$

characteristic (loop closed) from the servo input up to the motor and power amplifier must rise correspondingly with frequency, out to the cross-over point. Again assuming input noise of the uniform amplitude versus frequency type, the total noise power at the motor input is therefore,

$$P_{nm} = K_1 \int_0^\infty \left| \frac{\mu}{1+\mu} \right|^2 (\omega^2 + \omega_m^2)\omega^2 \, d\omega. \tag{15}$$

Again, ω_m is the reciprocal time-constant of the motor and K_1 is a proportionality constant. If ω_m is less than about half the cross-over frequency, then the noise power at the motor input increases as the fifth power of the bandwidth of the servo transfer characteristic.[24] Thus, if the input signal/noise ratio is small, this effect may be an important design consideration.

Still other servo errors may result from local extraneous signals or from coulomb and static frictional effects. These error sources are in a somewhat different class from those discussed previously, in that they are more nearly under the designer's control. That is, such extraneous signals and friction may be kept small by proper design and the residual friction effects further reduced by the use of local feedback. In the absence of local feedback, the servo error resulting from frictional or other torque disturbances at the output shaft readily is found to be

$$\Delta_T = \frac{T}{S(j\omega)} \cdot \frac{1}{1+\mu}. \tag{16}$$

$S(j\omega)$ is the actual stiffness (loop opened) of the output mesh, and T is the disturbing torque. T conceivably may represent static or coulomb friction, load-torque irregularities due to fluctuating running-friction, or wind torque. Again assuming the mechanical impedance to be resistance and inertia in series, the mechanical stiffness is, from (2.2), $S(j\omega) = j\omega(R + j\omega J)$. Thus the error is

$$\Delta_T = \frac{T}{j\omega(R + j\omega J)} \cdot \frac{1}{1+\mu}, \tag{16.1}$$

and the apparent output stiffness (loop closed) is

$$S' = j\omega(R + j\omega J)\,(1+\mu). \tag{16.2}$$

If T is taken as the static load torque, the resulting static error is found by setting $\omega = 0$ in (16.1). Assuming that μ behaves as $\omega_0/j\omega$ when ω approaches zero, the static error is

$$\Delta_T = \frac{T}{\omega_0 R}, \tag{16.3}$$

[24] This assumes a constant functional form for the transfer characteristic. However, the statement holds approximately, even with considerable variation in this form.

and the apparent low-frequency stiffness is $\omega_0 R$, being the ratio of the mechanical resistance to the velocity error coefficient. It may be noted that the static error will vanish if the loop transmission approaches infinity more rapidly than $1/\omega$ as ω approaches zero.

3.4 *Comparison of μ Characteristics for a Particular Input Signal*

In order to illustrate the advantages of shaping the loop characteristic for a particular input signal, a brief discussion will be given of the design of an automatic radar loop to track an airplane in azimuth over a constant linear-velocity course. The servo configuration is that given by Fig. 5b, θ_1 being the azimuth angle of the target and θ_2 the corresponding antenna angle. The lobing radar antenna has been assumed to take the place of the

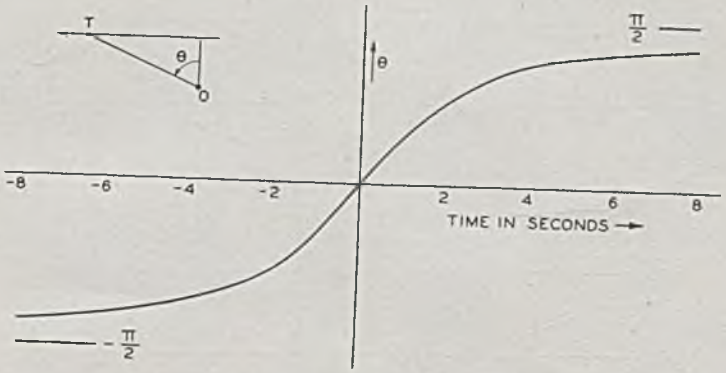


Fig. 12—Azimuth angle for constant linear velocity course.

synchro pair. Thus $\beta = -1$, and an error signal proportional to $\theta_1 - \theta_2$ is developed.[25]

Assuming a constant linear-velocity course having a maximum azimuth rate of 30 degrees/sec, the target azimuth angle is given by [26]

$$\theta_1(t) = \tan^{-1} .524t, \tag{17}$$

which is plotted in Fig. 12.

This course will develop a maximum azimuth acceleration $\ddot{\theta}_1$ of ± 10.3 degrees/sec^2 and a maximum $\dddot{\theta}_1$ of -16.4 degrees/sec^3. The continuous frequency spectrum of $\theta_1(t)$ may be found from (10.1) to be

$$F_1(\omega) = \pi \frac{e^{-1.9|\omega|}}{j\omega}. \tag{18}$$

[25] Assuming a low elevation course.

[26] The azimuth angle has been so taken that zero azimuth is obtained at the point of nearest approach.

A logarithmic plot of $|F_1(\omega)|$ is shown in Fig. 13.* It may be seen that the input signal spectrum falls at 6 *db*/octave for $\omega \ll 0.5$, at 12 *db*/octave for $\omega = 0.524$, and 30 *db*/octave at $\omega = 2.1$.

Assuming that the permissible dynamic error is 0.3 degree, a comparison will be made between the type 1 and type 3 loop characteristics of the previous section. For the type 1 system, which will have essentially a pure velocity error, (12.1) shows the required value of ω_0 to be 30/0.3 or 100.

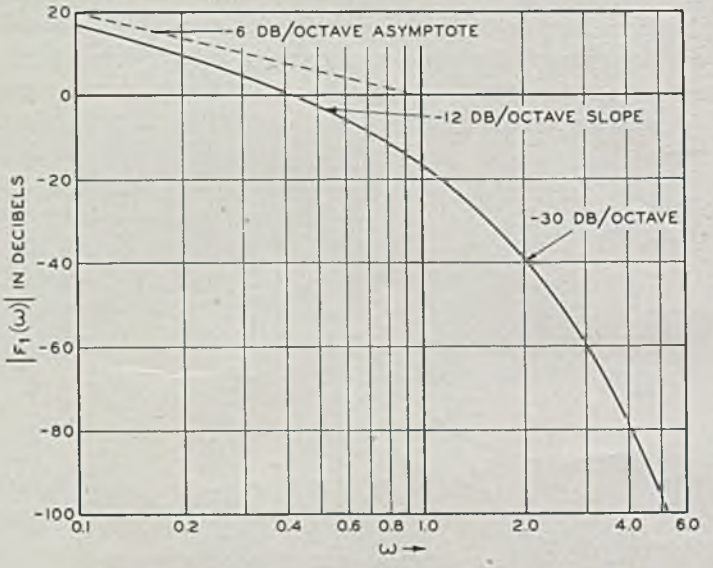


Fig. 13—Target frequency spectrum for constant velocity course.

Thus $\mu = 100/j\omega$. Figure 14 shows a logarithmic plot of the corresponding $|\mu|$. This characteristic departs rapidly from the shape of input signal spectrum given by Fig. 13, as ω is increased above 0.1.

The type 3 characteristic permits a considerably better match. Choosing a compromise value for ω_1 of 0.1, (12.3) may be used to calculate the necessary value of ω_0 as 415. Thus the loop transmission becomes $\mu = 41.5/j\omega(j\omega + 0.1)$. Figure 14 shows a plot of the corresponding $|\mu|$, modified near the gain cross-over to satisfy the stability requirements. This curve is a considerably better average match for the target frequency-spectrum up to $\omega = 1$. The resulting type 3 system has a predominant acceleration error as judged from the maximum velocity and acceleration errors of .072 degree and 0.25 degree respectively.

The total dynamic error curves for the constant-velocity course are given

* $|F_1(\omega)| = \pi$ has been taken as the zero *db* level.

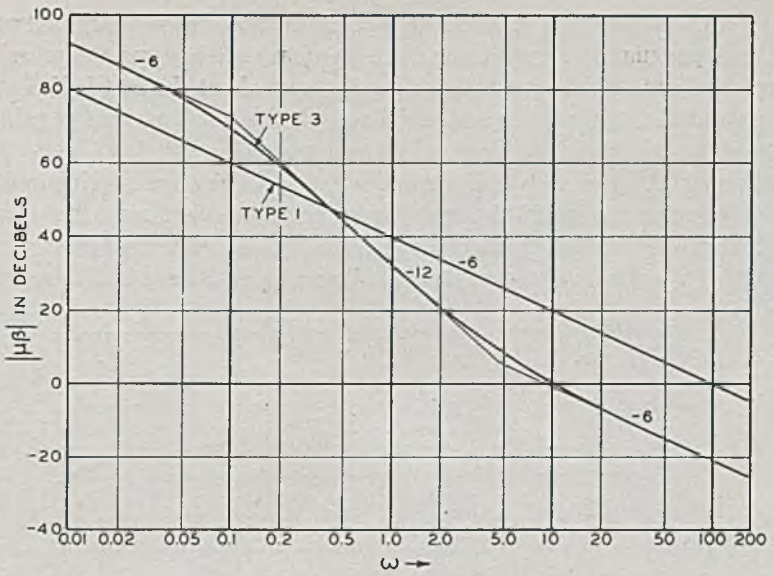


Fig. 14—Tracking loop characteristics.

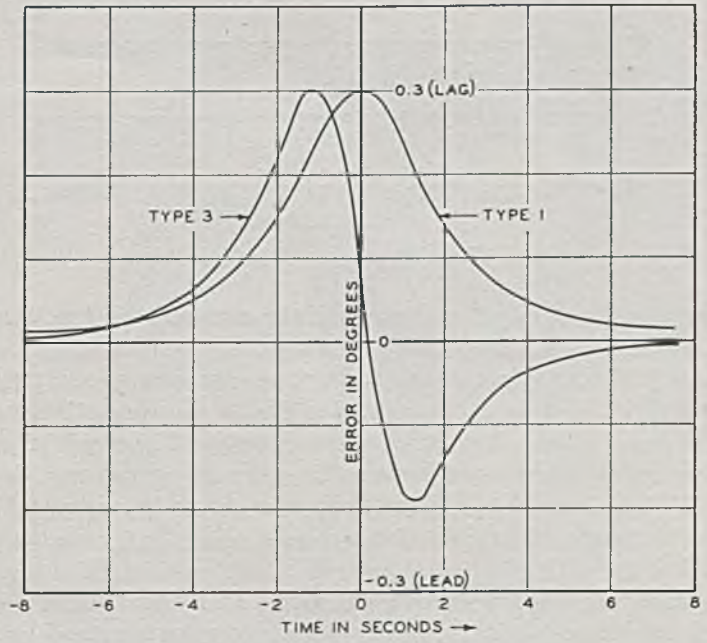


Fig. 15—Tracking errors for constant velocity course.

in Fig. 15. The velocity error of the type 1 system is always a lagging error and is maximum at the point of nearest approach. The type 3 composite of velocity and acceleration errors is lagging over about the first half of the course and leading for the second half, having lead and lag maxima at points closely grouped about the point of nearest approach.

Although the two loop characteristics develop the same maximum dynamic error on the specified target course, their transient responses to an input step differ widely, as may be seen from Fig. 16. The rise time for the type 1 loop is about .03 second compared with an initial rise in 0.17 second

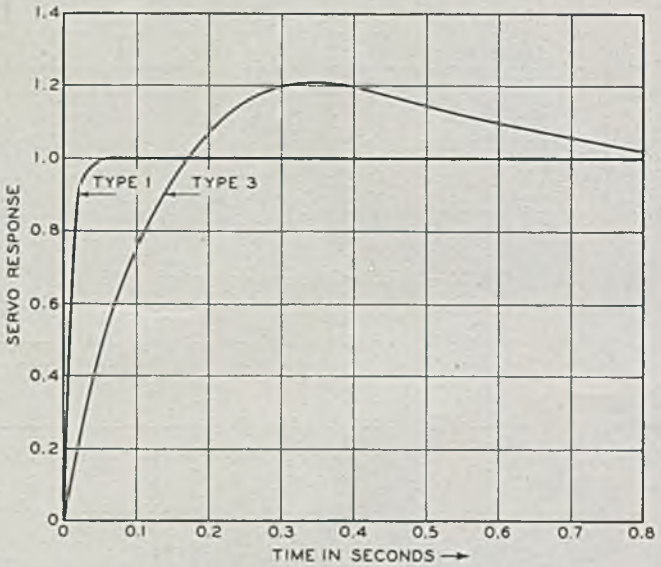


Fig. 16—Transient response of tracking servos.

for the type 3 system. Also, because of the overshoot the type 3 system requires about 0.7 second to settle within 5% of the equilibrium value.

For a final comparison of the two systems the corresponding transfer characteristics, $\mu/(1 + \mu)$, are plotted in Fig. 17 on arithmetic amplitude and frequency scales. It may be seen that the type 1 system is vulnerable to noise and interfering signals over a far wider frequency range than the type 3. Again assuming uniform input noise versus frequency, (14) may be used to show that the ratio of output noise power for the two systems is about 7.5:1.

Thus the luxury of crisp transient response as obtained with the type 1 system may demand a heavy penalty in terms of output fluctuations due to noise and other unwanted signal variations. This is a clear illustration of

the necessity for designing the servo loop to match the type of input signal to be transmitted, particularly for radar tracking systems where the "unwanted variations" are ever present.

3.5 *Use of Local Feedback*

There are many examples of the use of local or subsidiary feedback in servo systems. The more common of these include feedback around vacuum tube power amplifiers to obtain improved linearity and impedance properties, and over-all feedback around amplifier and motor-drive systems to

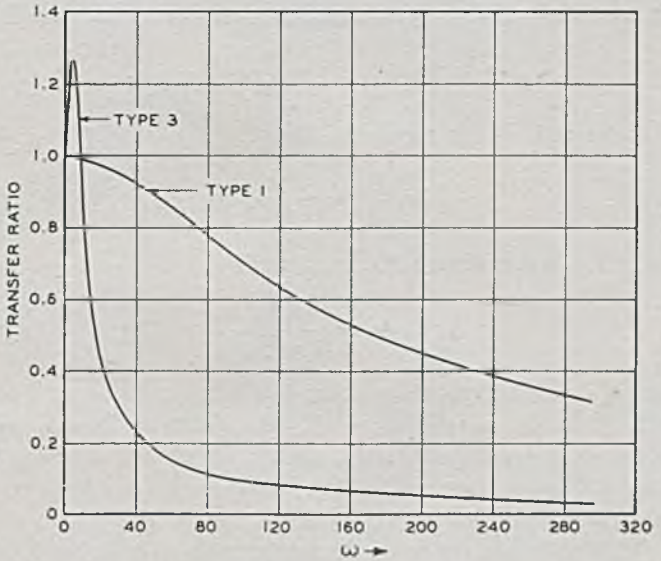


Fig. 17—Frequency response of tracking servos.

suppress frictional effects, increase output stiffness, and modify the inherent frequency characteristics of the basic components.[27] The tendency toward "β circuit dependency" as previously discussed also produces greater constancy of the stage transfer characteristics with time, temperature, etc.

Perhaps the simplest and most useful kind of local feedback is negative tachometer (velocity) feedback around motor-drive systems. This type of feedback widens the transfer frequency band of the drive system by reducing its time-constant, and increases the linear speed range of the motor. This may be illustrated by referring back to Fig. 7, which shows a typical tachometer loop. Assuming the transfer ratio of the amplifier to be a con-

[27] In a slightly different class are the servo systems used to provide automatic frequency and gain control in radio systems.

stant A, the transfer ratio of the motor and amplifier without feedback is, from (3.1),

$$\mu_T = \frac{\theta}{E}\,(\text{loop open}) = \frac{\mu_0}{J}\frac{1}{j\omega(j\omega + \omega_m)}, \tag{19}$$

where the constant $\mu_0 = A\mu_t$. (To avoid confusion with primary loop quantities, the tachometer loop will be represented by the symbols μ_T and β_T, rather than μ and β.) The quantity ω_m was defined as the ratio $(R_m + R_m')/J$ (see Fig. 2b), and is the reciprocal of the motor time constant. Replacing $(R_m + R_m')$ by R for convenience, (19) may be rewritten as

$$\mu_T = \frac{\theta}{E}\,(\text{loop open}) = \frac{\mu_0}{j\omega(R + j\omega J)}. \tag{19.1}$$

The transfer ratio of the tachometer is

$$\beta_T = \frac{E_\beta}{\theta} = -j\omega R_t,$$

and thus the loop transmission characteristic is

$$\mu_T\beta_T = -\frac{\mu_0 R_t}{R + j\omega J}. \tag{20}$$

For values of ω small compared with ω_m this loop transmission is constant and closely given by $\mu_T\beta_T(0) = -\mu_0 R_t/R$. When $\omega \gg \omega_m$, $\mu_T\beta_T$ approaches the form $-\mu_0 R_t/j\omega J$, and thus falls off at 6 *db*/octave. Consequently the maximum phase shift of the factor $-\mu_T\beta_T$ is -90 degrees, and no stability problem arises for the local tachometer loop.[28]

From (19.1) and (20), the over-all transfer ratio with feedback is

$$\begin{aligned}\frac{\theta}{E}\,(\text{loop closed}) &= \frac{\mu_T}{1 - \mu_T\beta_T},\\ &= \frac{\mu_0}{j\omega(R + \mu_0 R_t + j\omega J)}.\end{aligned} \tag{21}$$

Comparing (21) with (19.1), it may be seen that the sole effect of the tachometer feedback upon the over-all transfer ratio has been to add an apparent "ohmic" friction or mechanical resistance $\mu_0 R_t$ to the original value R. (It will be shown that this increase in apparent mechanical resistance also is effective in increasing the mechanical output impedance, although no power is dissipated in the added component $\mu_0 R_t$.)

[28] Actually, the effects of parasitic elements always modify this situation somewhat, especially if unusually high loop transmission is sought. However tachometer loops often require little or no stabilizing equalization.

Equation (21) also may be written as

$$\frac{\theta}{E}\,(\text{loop closed}) = \frac{\mu_0}{J}\,\frac{1}{j\omega(j\omega + \omega_m')}, \tag{21.1}$$

where $\omega_m' = (R + \mu_0 R_t)/J$ is the new corner frequency.

The change in over-all transfer ratio due to the tachometer feedback is shown in Fig. 18. The solid line diagrams A and B are the transfer gains without feedback and with feedback, respectively.[29] At low frequencies such that $\omega \ll \omega_m$, the feedback reduces the transfer ratio by the factor ω_m'/ω_m, the ratio of the two corner frequencies.[30] In order to restore this low-frequency loss in transmission, it is necessary to provide an added

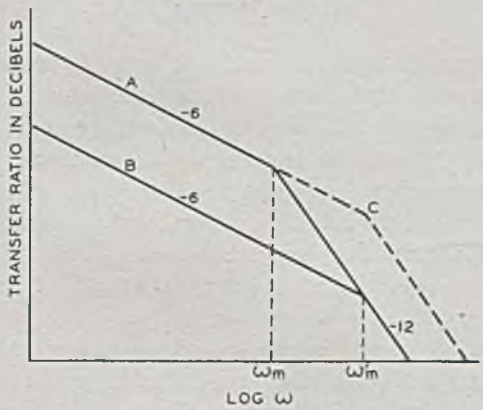


Fig. 18—Effect of tachometer feedback on motor characteristic.

amplification ω_m'/ω_m. If this is accomplished by increasing μ_0 and decreasing R_t so that the product $\mu_0 R_t$ remains constant,[31] the resulting transfer ratio will be that shown by the dotted lines C in Fig. 18. Comparing A and C, it may be seen that the net result of applying tachometer feedback and increasing the amplifier gain is to widen the transfer bandwidth by the factor ω_m'/ω_m. The required increase in amplification is the cost of widening the transfer bandwidth either by tachometer feedback or by non-feedback means, such as the use of a "forward-acting" equalizer in the amplifier. (However, such forward acting equalization fails to provide the increased over-all linearity and mechanical impedance obtained by the feedback method.) At frequencies sufficiently high that $\omega \gg \omega_m'$, the change in transfer ratio due to the feedback disappears, the mechanical inertia becoming the controlling element.

[29] The straight line asymptotes have been drawn instead of the actual gain curves.

[30] This is also the factor by which the feedback reduces the output speed obtained for a steady input voltage, neglecting circuit non-linearites and coulomb friction.

[31] This ensures a fixed loop transmission, and thus an unchanging value for ω_m'.

For ω small compared with ω_m', (21) becomes

$$\frac{\theta}{E}\,(\text{loop closed}) \simeq \frac{\mu_0}{j\omega(R+\mu_0 R_t)}, \qquad (\omega \ll \omega_m').$$

If the tachometer feedback is substantial ($\omega_m' \gg \omega_m$), this may be further approximated as

$$\frac{\theta}{E}\,(\text{loop closed}) \simeq \frac{1}{j\omega R_t}, \qquad \begin{pmatrix} \omega \ll \omega_m' \\ \omega_m' \gg \omega_m \end{pmatrix}. \qquad (21.2)$$

and the corner frequency becomes

$$\omega_m' \simeq \frac{\mu_0 R_t}{J}, \qquad (\omega_m' \gg \omega_m).$$

Thus for reasonably high feedback, the over-all transfer ratio (21.2) depends only upon the tachometer characteristic, being substantially independent of changes in the original mechanical resistance R or the amplifier-motor factor μ_0. The corner frequency ω_m' is similarly independent of changes in R, although still a direct function of μ_0. Thus the principal non-linearity of two-phase induction motors, namely variation in electrical damping with speed, is effectively suppressed by this type of local feedback, and systems employing such motors up to 80% of their synchronous speed may be designed on a linear basis.

The increase in mechanical impedance due to the feedback may be shown by assuming a torque disturbance T applied at the output shaft. Without feedback, the resulting speed disturbance is

$$\theta\,(\text{loop open}) = \frac{T}{Z_m} = \frac{T}{R + j\omega J}.$$

With feedback, the corresponding shaft speed disturbance becomes

$$\theta\,(\text{loop closed}) = \frac{T}{Z_m} \cdot \frac{1}{1 - \mu_T \beta_T},$$

$$= \frac{T}{R + \mu_0 R_t + j\omega J}.$$

Thus the apparent mechanical resistance, and therefore the protection against frictional torques, has been multiplied by a factor $(1 + \mu_0 R_t/R) = \omega_m'/\omega_m$. If the motor-drive system with tachometer feedback is employed in a simple follow-up system of the type of Fig. 5, equation (16.3) shows that the resulting low-frequency output-shaft stiffness will be $\omega_0(R + \mu_0 R_t)$ or $(\omega_m'/\omega_m)\omega_0 R$.[32] Therefore the output stiffness has

[32] The low-frequency loop transmission of the follow-up loop is again taken to be $\omega_0/j\omega$.

been increased by the factor ω_m'/ω_m over that obtained without the use of local feedback, assuming identical follow-up loop characteristics ($\mu\beta$) for the two cases. The ratio ω_m'/ω_m thus directly measures the feedback eduction of static and low-speed errors of the follow-up system due to torque disturbances. In practice the resulting increase in static accuracy may be of the order of 10 to 100 times.

3.6 *Error Reduction by Non-Feedback Means*

In situations where the noise associated with the input signal is small, it may be desirable to reduce the dynamic errors obtained with a given servo system by the use of forward-acting equalization external to the loop.

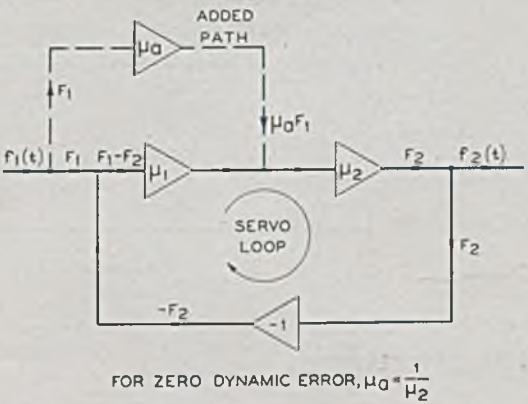


Fig. 19—Forward-acting error compensation.

That is, the dynamic error characteristic may be computed, and the servo input or output modified by supplementary networks in such a fashion as to reduce the over-all error.

An illustrative arrangement, which is suitable when the input member is accessible,[33] is shown in Fig. 19. For convenience the servo is taken to be a simple follow-up system having $\beta = -1$. The μ circuit is shown divided into two parts, μ_1 and μ_2. Typically, μ_1 may be the transfer stiffness of a synchro pair (Fig. 3b), and μ_2 the transfer characteristic of a motor-drive system. The normal dynamic error component for such a loop, omitting the dotted line, has been shown to be $F_1/(1 + \mu)$. If an additional signal $\mu_a F_1$ is obtained from the input member and injected into the system as shown by the dotted line, then

$$F_2 = \frac{\mu}{1+\mu} F_1 + \frac{\mu_a \mu_2}{1+\mu} F_1,$$

$$= \frac{\mu + \mu_a \mu_2}{1+\mu} F_1.$$

[33] This is not the case for a radar tracking loop, for instance.

Thus the over-all error becomes

$$F_1 - F_2 = \left(1 - \frac{\mu + \mu_a \mu_2}{1 + \mu}\right) F_1 ,$$

or

$$F_1 - F_2 = \frac{1 - \mu_a \mu_2}{1 + \mu} F_1 . \tag{22}$$

If the added transmission path is so designed that

$$\mu_a = \frac{1}{\mu_2} , \tag{23}$$

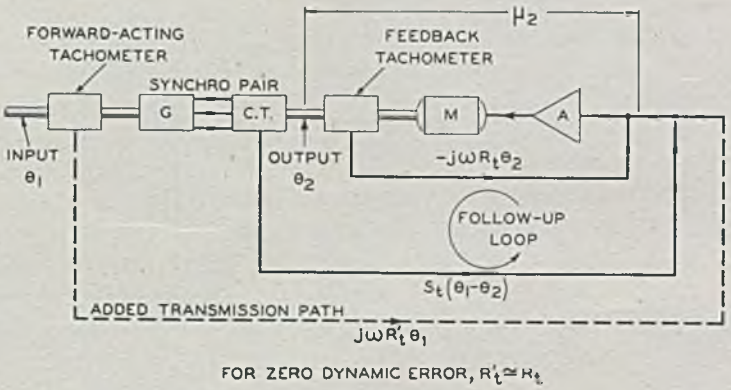


Fig. 20—Forward-acting tachometer system.

then $F_1 = F_2$, and the dynamic error vanishes. Thus the desired form of the added transmission depends only upon the μ_2 portion of the loop characteristic. It will not be possible to satisfy the condition given by (23) exactly, especially at the higher frequencies where noise enhancement and parasitic effects will become increasingly important. However, it is often possible to obtain the proper form for μ_a over the range of frequencies responsible for the bulk of the dynamic error. If μ_a has the proper frequency characteristic but is too large by 10%, for instance, it may be seen from (22) that there still remains a 10/1 increase in dynamic accuracy.

The foregoing method is especially applicable when μ_2 represents the transfer characteristic of a motor-drive system employing tachometer feedback, as shown in Fig. 20. Here the basic input-output comparison is obtained by means of the synchro pair, while a tachometer coupled to the input shaft provides the error-reducing signal. Thus the transmission μ_a is equal to $j\omega R_t'$, where R_t' is the tachometer transfer resistance. The expression for

μ_2 is given approximately by (21.2) as $1/j\omega R_t$. Thus, by (23), $R_t' \simeq R_t$ for substantial cancellation of the dynamic error (at frequencies small compared with ω_m'). That is, the output voltages of the two tachometers must closely annul each other when the input and output shafts are travelling at the same speed. Since the tachometers may be closely alike and excited from the same supply line, it is comparatively easy to keep their transfer ratios closely matched. In practice an error reduction of 20/1 is readily maintained by this method.

The error compensation scheme described above does not change the loop characteristic $\mu\beta$ of the basic servo loop, and thus does not create new stability problems. Its use to obtain high servo accuracy is desirable when the input noise is small and when a high loop gain is difficult to obtain because of parasitic elements or equipment complexities.

Abstracts of Technical Articles by Bell System Authors

Computation of Interfacial Angles, Interzonal Angles, and Clinographic Projection by Matrix Methods.[1] W. L. Bond. A way of setting up the general crystallographic axes a, b, c on unit orthogonal axes x, y, z is used to afford a matrix method of computing interfacial angles and zonal angles. It also affords a method of making clinographic projections.

A Current Distribution for Broadside Arrays which Optimizes the Relationship between Beam Width and Side-Lobe Level.[2] C. L. Dolph. A one-parameter family of current distributions is derived for symmetric broadside arrays of equally spaced point sources energized in phase. For each value of the parameter, the corresponding current distribution gives rise to a pattern in which (1) all the side lobes are at the same level; and (2) the beam width to the first null is a minimum for all patterns arising from symmetric distributions of in-phase currents none of whose side lobes exceeds that level.

Design curves relating the value of the parameter to side-lobe level as well as the relative current values expressed as a function of side-lobe level are given for the cases of 8-, 12-, 16-, 20-, and 24-element linear arrays.

Paper Capacitors Containing Chlorinated Impregnants—Mechanism of Stabilization.[3] L. Egerton and D. A. McLean. The stabilization of paper capacitors containing chlorinated aromatic impregnants with small quantities of organic additives is well established commercially. Although for practical reasons anthraquinone was chosen for initial commercial application, other quinones are also effective, as are the nitroaromatics, maleic anhydride, and sulfur. Evidence is given that the mechanism of stabilization consists of the formation of barrier films on the electrodes. These barrier films, which in certain cases may cover only the active points on the electrode surface, reduce the catalytic decomposition of the chlorinated impregnant by the electrode metal, prevent attack of the electrodes by liberated hydrogen chloride, and hinder electrolytic action. It appears likely that the film-forming properties of the stabilizers are dependent upon their oxidizing power. A secondary effect of stabilizers may be the formation of complexes with aluminum chloride to diminish the activity of the latter or change the nature of the reactions which it induces. Conductivity measurements in

[1] *American Mineralogist*, Vol. 31, pp. 31–42 (1946).
[2] *Proc. I.R.E. and Waves and Electrons*, June 1946.
[3] *Indus. and Engg. Chemistry*, May 1946.

HCl-saturated chlorinated diphenyl containing soluble additives are in line with known hydrogen-bonding tendencies of the additives. Compounds which are strong organic bases do not stabilize capacitors.

Quartz Crystals for Electrical Circuits.[4] R. A. HEISING. This book is a compendium of information, both theoretical and practical, on quartz crystal plates, their design and manufacture. It embodies the vast experience of the Bell Telephone Laboratories in research and in manufacture of quartz crystals. It originated from a series of lectures given by the members of the Laboratories technical staff who had carried out the early studies and developments in this field. By this means, engineers were trained for the immense expansion in crystal manufacture required to meet the demand of our military forces during the War. These lectures have been reorganized and rewritten, and are published together in this comprehensive book. Articles covering some of the various chapters have appeared in the Bell System Technical Journal.

The treatment covers in full the theory and practice of the preparation of quartz crystals, the instruments used, including new types developed for special purposes, the problems encountered in the various uses of quartz crystals, and the full details of the methods devised for their solution. The various processing chapters, dealing with cutting and grinding, plating and other topics of equal importance, include much information that appears for the first time in any book. The account of the manufacturing process is most complete. There are discussions of new practical methods of adjustment to frequency, of the new performance indicator, of a new type of crystal cut that operates at very low frequencies, and many new developments that represent notable advances in crystal technology.

Geometrical Characterizations of Some Families of Dynamical Trajectories.[5] L. A. MACCOLL. The chief problem considered in this paper is that of obtaining a set of geometrical properties which shall completely characterize the five-parameter family of trajectories of an electrified particle moving in an arbitrary static magnetic field. A solution of the problem is found in the form of a set of four principal and four subsidiary properties. A geometrical characterization, in the form of a set of two properties, is also given of the four-parameter family of trajectories of an electrified particle moving in an arbitrary static magnetic field with an arbitrarily prescribed value of the total energy. Various other properties of the families of curves are discussed, and the paper closes with a brief consideration of some analogous problems in which the particle moves in a fixed plane.

[4] Published by D. Van Nostrand Company, Inc., New York, N. Y., 1946.
[5] *Amer. Math. Soc. Trans.*, July 1946.

Comparison of Natural and Synthetic Hard Rubbers.[6] G. G. WINSPEAR, D. B. HERRMANN, F. S. MALM, and A. R. KEMP. GR-S, nitrile, and natural hard rubbers are compared as regards compounding, processing, vulcanization, and physical and dielectric properties. Natural rubber and GR-S compounds intermediate in sulfur content between hard and soft rubber also are compared. GR-S and nitrile rubber compositions suitable for commercial ebonite fabrication are described. Extensive breakdown of the basic copolymers has little effect on the physical properties of synthetic ebonites. The time required for the beginning of exothermic reaction in vulcanization is longer for GR-S than for natural rubber ebonites. Rockwell hardness is greater for GR-S. Some GR-S ebonites are penetrated to the same depth as natural ebonites, with a greater tendency toward instantaneous recovery. The two are similar in impact strength, but the ability to withstand a sharp bend is characteristic of natural ebonites alone. The latter are superior to GR-S ebonites in heat deformation below 60° C., but above this temperature the reverse is true and nitrile ebonites are superior to both. GR-S ebonites are more stable and nitrile ebonites less stable chemically than natural ebonites. GR-S ebonite dust as a filler increases brittleness. A diatomaceous earth improves the processing properties of GR-S hard rubbers. The adverse effect of ultraviolet light on surface resistivity is reduced when a GR-S hard rubber is filled with whiting. Natural and GR-S hard rubbers are alike in dielectric behavior.

Signal and Noise Levels in Magnetic Tape Recording.[7] D. E. WOOLDRIDGE. The primary object of the work described here was to determine what properties of the tape and associated magnetic elements are responsible for the noise and signal output levels of magnetic recordings and, if possible, to display in specific equations the pertinent relationships connecting noise and signal levels with the physical properties of the tape and polepieces. In the course of the study, methods appeared for decreasing the noise and increasing the useful signal reproduced from magnetic tape. These methods and some of the use that Bell Telephone Laboratories and Western Electric have made of them are mentioned in the discussion. While some of the work described in this paper has implications for more than one type of magnetic recording process, perpendicular recording on tape is the actual subject matter dealt with. In every case discussed, the record medium was 0.050 inch wide and 0.0022 inch thick. Except where otherwise noted, a chrome-steel tape was used at a speed of 16 inches per second.

[6] *Indus. and Engg. Chemistry*, July 1946.
[7] *Elec. Engg., Trans. Sec.*, June 1946.

Contributors to This Issue

WALLACE A. DEPP, B.S. in Electrical Engineering, University of Illinois, 1936; M.S. 1937. Bell Telephone Laboratories, 1937–. Mr. Depp has been primarily engaged in the development of gas filled tubes.

FREDERICK S. GOUCHER, A.B., Acadia University, 1909; A.B., Yale University, 1911; M.A. Yale, 1912; Ph.D. in physics, Columbia University, 1917; D.Sc. (Hon.), Acadia University, 1934. Engineering Department, Western Electric Co., 1917–18; Research, University College, London, 1919; Research Laboratories, General Electric Co., Ltd., England, 1919–26; Bell Telephone Laboratories, 1926–. In the Physical Research Department engaged in a fundamental study of contacts with reference to carbon microphones and switching apparatus, and during the war in the development of switching devices used in radar.

ROBERT E. GRAHAM, B.S. in E.E., Purdue University, 1937. Mr. Graham joined the technical staff of the Bell Telephone Laboratories in 1937, and has since been engaged in television research. During the war period he worked on radar and automatic tracking problems.

J. R. HAYNES, B.S. in physics, University of Kentucky, 1930. Bell Telephone Laboratories, 1930–. In the Physical Research Department engaged in contact studies and during the war the development of switching devices used in radar.

E. PETERSON, Cornell University, 1911–14; Brooklyn Polytechnic, E.E., 1917; Columbia University, M.A., 1923, Ph.D., 1926; Lecturer in Electrical Engineering, 1934–40. Electrical Testing Laboratories, 1915–17; Signal Corps, U. S. Army, 1917–19; Western Electric Company, Engineering Department, 1919–25; Bell Telephone Laboratories, 1925–. Dr. Peterson's work has dealt with non-linear circuits and circuit elements.

SNYDER C. RAPPLEYE, War Alumnus, Cornell University, 1919; New York Telephone Company, 1921–1923; Traffic Division of Operation and Engineering Department, American Telephone and Telegraph Company since 1923. Recently Mr. Rappleye has been working on intertoll trunk engineering methods.

E. J. Ryder, E.E., Polytechnic Institute of Brooklyn, 1935. Engineering Department, Western Electric Co., 1922–25; Bell Telephone Laboratories, 1925–. In the Physical Research Department engaged in contact studies and during the war the development of switching devices used in radar.