ABSTRACT

Network management deals with network operation as contrasted with dimensioning, and its objective is to ensure optimum use of available facilities in the face of unpredicted load or facility failure. Control methods aim primarily to prevent regenerative delay among common control switching centers. Recent work considers the effect of controls on customer and traffic subsets within the network, improved methods to cope with overloads caused by "focused calling", and better means for collecting and presenting network data to the managers.

Two examples are discussed. One considers the effect on originating exchange dial tone performance of an overload control designed to protect against tandem exchange congestion in a metropolitan network. The second describes new control methods which rely on the automatic detection of traffic parcels encountering low completion. This information makes possible selective controls which are shown by simulation to improve network performance significantly under conditions of focused calling.

Mr. Range has observed that the term "network management" enjoys a wide variety of meanings. The work I will discuss today falls comfortably within the more specific definition of CCITT — the use of traffic overload controls to ensure optimum use of available network facilities. In the United States network management action in the field takes place typically on a time scale ranging from a few minutes to a day or two. Thus we are talking about a problem of network operation as contrasted with one of design or dimensioning. The purpose of my talk today will be to provide a necessarily brief overview of our current areas of interest in network management, and to illustrate by example some recent results.

It is convenient to characterize the telephone network in the USA as a collection of many local, or metropolitan, networks interconnected by the national toll, or long distance, network as illustrated in Figure 1. The triangular network to the left represents such a metropolitan network in which the local and tandem exchanges are arranged in a two-level hierarchy. A typical metropolitan network would contain several tandems, and the metropolitan New York network, our largest, uses 58 tandems for a variety of purposes. In practice the configurations of metropolitan networks vary widely from city to city depending particularly on the predominant type of switching equipment. The long distance, or toll, network, shown to the right, contains over 1000 transit exchanges usually called "toll offices". Together with the local exchanges these exchanges are arranged in a 5-level hierarchical network. At the top of the hierarchy are 12 regional switching centers fully interconnected by final trunk groups. Two of these regional exchanges are located in Canada at Montreal and Regina.

USA TELEPHONE NETWORK 1973

Figure 1
Following the introduction of Direct Distance Dialing 20 years ago, the proportion of customer dialed toll calls grew rapidly. By 1960 severe congestion was being observed in the toll network on peak days such as Christmas and Mother's Day and the phenomenon of regenerative switching delay was identified. Thus early work in network management was aimed at solving toll network problems. Experience and analysis led to the present controls which have proven to be fully adequate for protecting the network against the types of overloads which can occur on these peak days.

Interest in metropolitan overload problems increased in the mid 1960s as the effects of switching congestion became evident in multi-tandem networks when they were stressed, for example, during exceptionally long and heavy snowstorms. The phenomenon of regenerative switching delay was seen to be similar to that observed in the toll network, but in the metropolitan case the traffic backed into the originating local exchanges, interacting with dial tone performance and introducing an added dimension to the network management problem. Current work on network management is directed toward both the metropolitan and toll networks.

As practiced in the field, network management is a cooperative effort. Its success in fact depends upon the high degree of cooperation among several hundred people responsible for network management in the 21 Bell System companies and in the connecting independent telephone companies. Training of managers has also been a cooperative effort. The Bell System provides an intensive one-week course at Morristown, New Jersey called "the network management school" and taught by experienced network managers. As of May 1973 the school had graduated 3500 students of whom over 400 were employed outside of the Bell System.

Coordination in the field is maintained over a collection of dedicated voice lines, operated as party lines and arranged to conform to the switching network hierarchy. The National Control Center in New York is shown in Figure 2 as it appeared in 1967 at the time of the 5th ITC. The figure displays critical congestion indicators from each regional switching center and trunk group busy indicators for the major interregional trunk groups.

It is customary to describe specific network management controls as being either "expansive" or "protective". Expansive controls involve clever manipulation of routing to gain the use of idle capacity not available under the design routing. Protective controls, on the other hand, usually involve the blocking, or restriction, of certain categories of traffic to prevent the degradation of network carried load due to switching delay. I will discuss the status of expansive controls briefly before turning to the protection problems which have claimed more of our recent attention.

On a typical peak day, e.g., this past Mother's Day, North American network managers introduce and remove over 1000 temporary routings called "reroutes". Although total holiday network load is less than that of an average business day, it is distributed differently and the reroutes attempt in part to utilize business day capacity to which the holiday load would not normally have access. Interregional reroutes also take advantage of time-zone differences. Figure 3 shows two of the Mother's Day reroutes, east coast routing via California early in the day, and an international reroute later on between the Canadian cities of Vancouver and Winnipeg via New York.

Figure 3

At present nearly all rerouting is done manually. Automatic expansive routing requires more interexchange signaling capacity than now exists or can be provided economically with existing techniques. While we have some plans, they depend on the deployment of common channel interoffice signaling. Enhanced signaling is important to maintain control of the routing and, equally important, to retain control of the traffic measurements taken for network design purposes.

A few cases of automatic expansive routing do exist for special purposes. Figure 4 illustrates a technique used to provide full access on international circuits while permitting both switching center and circuit diversity. For example, circuits to Tel Aviv are divided between the Pittsburgh and New York gateway switching centers. Calls to Tel Aviv can overflow from one center to the other, but only if one or more international circuits are idle leaving the other machine. The intergateway group of 42 trunks is dedicated to international use and is dimensioned to be nonblocking. Telegraph channels transmit the group idle information needed to prevent "ring around the rosy" — i.e., the circular routing which would otherwise result when all circuits to Tel Aviv are busy. Routing to London is similar, as shown in the figure, and involves all three gateway centers.
Turning to the problems of network protection, the phenomenon which must be guarded against is, of course, regenerative switching delay caused by senders in one machine queuing for senders in another and so on. Paul Burke described this effect at the Congress in 1967, and his simulation results illustrate the problem very nicely. Figure 5 is Burke’s curve showing the now familiar phenomenon of carried network load falling with increasing offered load. In Figure 5 the network is operating in the steady state. The transient case shown in Figure 6 is particularly instructive. The key point is how the nature of the customer blockage changes, under continuing reattempt pressure, from trunk shortage in the early stages to sender timeouts in the later stages. Once into intense switching congestion, the trunk occupancy actually declines as short ineffective attempts replace normal holding time messages. In the days before adequate control methods were available, toll machine managers observed this phenomenon with understandable frustration.

Most switching overloads are controllable by the existing methods of cancellation of alternate routing, directionalizing of two-way trunks against the overloading traffic, and, where stored program machines permit, the restricting of low completion rate traffic on the basis of called number. This last technique is known as code blocking. While the existing methods are generally adequate, operating questions continue to arise in the field, and opportunities for improvement come with the development of new stored program control switching systems.

PROBLEMS OF CURRENT INTEREST

EFFECT OF CONTROLS ON CUSTOMER

BETTER CONTROL OF FOCUSED OVERLOADS

FASTER CONTROL RESPONSE

IMPROVED MANAGEMENT INFORMATION

Figure 7 lists some problems of current and continuing interest. The question of the effect of controls on the customer is a question of equity of service under conditions of overload in the presence of controls. I will describe a specific example below. Looking at the second problem area, a focused overload is characterized by heavy calling directed toward one telephone number, local exchange or geographic area. Existing controls, devised to combat peak day problems, fall well short of
optimal for controlling focused overloads. I will describe an improved, more general control method which looks very promising.

Manual protection methods tend to be slow both in the application and removal of controls. Reliability is also troublesome in metropolitan networks where the availability and the network management training of exchange personnel is necessarily limited; thus the need for faster, more automatic controls. Lastly, we are devoting substantial effort to the planning and design of a new computer-based system for collecting traffic data for engineering and operational purposes. The data will be gathered centrally in real time so that it can be made available for network management purposes, both as a means of surveillance and as a measurement base for automatic control. A major challenge is the definition of methods or processing the large volume of input data to extract and display only what the network manager needs or wants to know.

Turning now to the first of my two examples, a useful automatic control in the toll network has been DOC, Dynamic Overload Control. Under typical operations, DOC measures the number of incoming trunks waiting for senders at the input of a toll switching exchange (actually the number of start relays, a quantity which approximates the number of waiting trunks). When that number exceeds a preset threshold, a signal is sent to connecting exchanges to reduce the load offered to the requesting toll exchange, generally by the cancellation of calls alternate routing to the interconnecting trunk group. Thus under DOC the exchange will accept only as many call attempts as it can safely handle without excessive delay for common control.

DOC has also been found to be useful in multitenant metropolitan networks to prevent the spread of sender delay among the tandems and ultimately into the local exchanges. The experimental Local DOC, or LDOC, system was well tested in Washington, D. C., during a week-long snowstorm and results were reported by J. L. Laude. As applied to a metropolitan tandem, LDOC controls the traffic, if necessary, in the originating exchanges homing on that tandem. Calls cancelled by LDOC go directly to reorder (a distinctive tone or announcement) in the originating office and the customer is free to reattempt sooner than with no control, though certainly not encouraged to do so. Under overload conditions, originating register equipment is also stressed and dial tone can be delayed. The question raised by local exchange administrators was "does the imposition of controls to protect the tandem network affect local dial tone service?" In particular does controlled-call blockage cause increased dial tone delay through the (potential) mechanism of increased reattempts?

To answer this question, and others relating to control deployment, a metropolitan network model was constructed for computer simulation. As shown in Figure 8, the metropolitan area is divided geographically into three sectors, each served by one tandem. All local exchanges within a sector home on the sector tandem, and the tandems are interconnected by final trunk groups. The model contains 8000 trunk lines distributed in full direct (nonalternate-route), primary high usage, intermediate high usage and final groups in proportions representative of the New York metropolitan network. The sender and marker functions of the tandem exchanges are represented by queuing models appropriate to the crossbar tandem system. Similar models represent the No. 1 crossbar and No. 5 crossbar systems including the dial tone mechanisms. Under LDOC, a tandem can only control traffic in the local exchanges within its own sector.

A form of overload typical of large metropolitan networks consists of heavy intersector traffic added to a background load which is distributed in proportion to the traffic for which the network was engineered. Such an overload, called a "skewed" overload, arises from increased urban-suburban calling during exceptionally bad weather and during some forms of transportation stoppages. In this example, the model network is subjected to a skewed overload, the background being the average busy season load for which the trunks are engineered. Blocked calls reattempt with a probability of .8 and the time to reattempt is exponentially distributed. Results of the simulation are shown in Figure 9 which plots dial tone delay in percent of calls delayed beyond 3 seconds versus the offered load in percent of background load, where 100% means engineered load. The curves show the network average dial tone delay for the no control and control cases. The absolute levels of the curves depend on the reatempt parameters, but two general observations can be made. LDOC does not degrade local dial tone service, it helps it. Secondly, it doesn't help it very much. Under conditions of serious overload, service is still poor (the 10-high-day objective for dial tone delay is not more than 8%).

**Figure 8**

**Figure 9**
The simulation showed that the modest improvement in dial tone delay under DCC control resulted from reduced reattempts due partly to more effective tandem switching but primarily to higher completion rates on the less congested local routes. Local exchange common control, freed by DCC from waiting for tandem senders, was available to handle more calls.

While DCC is useful in controlling general overloads, in its present form it is largely ineffective for controlling focused overloads where the traffic needing to be controlled is heading down the hierarchy instead of up. Focused overloads come in all shapes and sizes. A water main break, flooding the cable vault of a large business workhorse, the ESS, found themselves 25 to 35 times normal as people called to check on the status of friends and relatives. Toll exchanges normally switching 3 or 4 percent of their calls to California found themselves with a 100% call attempt overload on common control. Failure rates to California at times ran as high as 95%, and due to the presence of the California traffic, blockage ran high between points far removed from the disaster area.

The network was kept under control primarily by direction-alizing trunks outward from Southern California and against the incoming pressure. In some groups 90% of the trunks were made busy to inbound traffic. Figure 10 shows the major points in the network where controls were taken (the larger circles in the figure) and illustrates the wide extent of the overload. Not shown on the figure are all the operator locations where attempts were limited as well.

The determination of HTR will enable 4ESS to detect a serious overload problem within five minutes of its onset. The existence of the HTR information will enable it to respond much more selectively than is possible with existing network management controls. Two controls are planned: selective trunk reservation (STR) and selective dynamic overload control (SDOC). The former responds to trunk congestion, the latter to machine congestion. If neither form of congestion exists, no controls will be taken even though one or more codes is identified as HTR.

The simulation showed that the modest improvement in dial tone delay under DCC control resulted from reduced reattempts due partly to more effective tandem switching but primarily to higher completion rates on the less congested local routes. Local exchange common control, freed by DCC from waiting for tandem senders, was available to handle more calls.

While DCC is useful in controlling general overloads, in its present form it is largely ineffective for controlling focused overloads where the traffic needing to be controlled is heading down the hierarchy instead of up. Focused overloads come in all shapes and sizes. A water main break, flooding the cable vault of a large business workhorse, the ESS, found themselves 25 to 35 times normal as people called to check on the status of friends and relatives. Toll exchanges normally switching 3 or 4 percent of their calls to California found themselves with a 100% call attempt overload on common control. Failure rates to California at times ran as high as 95%, and due to the presence of the California traffic, blockage ran high between points far removed from the disaster area.

The network was kept under control primarily by direction-alizing trunks outward from Southern California and against the incoming pressure. In some groups 90% of the trunks were made busy to inbound traffic. Figure 10 shows the major points in the network where controls were taken (the larger circles in the figure) and illustrates the wide extent of the overload. Not shown on the figure are all the operator locations where attempts were limited as well.

The determination of HTR will enable 4ESS to detect a serious overload problem within five minutes of its onset. The existence of the HTR information will enable it to respond much more selectively than is possible with existing network management controls. Two controls are planned: selective trunk reservation (STR) and selective dynamic overload control (SDOC). The former responds to trunk congestion, the latter to machine congestion. If neither form of congestion exists, no controls will be taken even though one or more codes is identified as HTR.

The determination of HTR will enable 4ESS to detect a serious overload problem within five minutes of its onset. The existence of the HTR information will enable it to respond much more selectively than is possible with existing network management controls. Two controls are planned: selective trunk reservation (STR) and selective dynamic overload control (SDOC). The former responds to trunk congestion, the latter to machine congestion. If neither form of congestion exists, no controls will be taken even though one or more codes is identified as HTR.
selective controls

4 ess

<table>
<thead>
<tr>
<th>traffic reserved</th>
<th>congestion level of</th>
<th>requesting machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>none</td>
<td></td>
</tr>
<tr>
<td>htr only</td>
<td>low</td>
<td></td>
</tr>
<tr>
<td>htr other alternate route</td>
<td>high</td>
<td></td>
</tr>
</tbody>
</table>

Figure 12

Selective DOC operates in a similar manner, but now the stimulus comes from a connecting switching machine which is sensing congestion. As shown in the slide, the responses are similar to those for STR. The lowest level of machine congestion will cause 4 ESS to restrict only HTR traffic if such exists and only if it is headed toward the requesting machine. At the higher congestion threshold all alternate route traffic is restricted as in nonselective DOC.

Together the new controls are both more powerful and more general than existing controls. They are more powerful in that they can cope with a wider range of overloads, and more general because they subsume the functions of the existing more specialized controls (DOC, DRE, PRE) and can be applied without hierarchical restriction. STR and SDOC can also be operated beneficially and compatibly with machines operating under the older controls, a requirement for orderly introduction.

To evaluate the effectiveness of the new 4 ESS controls, the computer simulator was modified to include STR and SDOC and was applied to the model toll network shown in Figure 13. This network represents three Regional centers and their subtending machines in a three level hierarchy. It has been used extensively at the Network Management School to study a wide variety of problems. The following results come from a focused overload headed toward toll office 66. In addition to the engineered load, the offered traffic to 66 (without reattempts) is increased eight times and the originating load from 66 by two times. Reattempts occur with the probability .8 and are distributed exponentially in time from the preceding call failure. The resulting attempt pressure is comparable to that experienced following the earthquake.

Figure 14 shows the number of messages in progress in the network versus time from the onset of overload. Engineered capacity is 1260 calls. With no control the network degrades to 19 percent of capacity. With the optimum use of existing controls the network can be held only to 37 percent. Under STR alone, the network holds above 1260 for some time but eventually succumbs to machine congestion. Once the trunks begin to carry ineffective attempts their occupancy drops and STR is no longer effective. Adding SDOC prevents the machine congestion and holds the network steady at peak load.
More detailed results confirm the effectiveness of the selective controls. Completion, both to and from the focus, is substantially increased over the nonselective control cases. Significant benefits also accrue when not all of the machines are assumed to have the 4 ESS selective controls, as will be the case for some time. While existing machines cannot themselves detect the presence of HTR traffic they can take selective controls once informed of HTR codes by connecting 4 ESS machines. Simulations of other types of overload yield similar results, though the amount of improvement is greatest for focused overloads. The simulation results suggest the possibility of a low cost deployment strategy. The STR control is accomplished entirely within the switching system. SDOC, on the other hand, requires telemetry and its associated cost. Given the initial effectiveness of STR as shown in Figure 14, would it be possible to obtain adequate protection with only a limited deployment of SDOC? The answer appears to be yes.

That concludes my brief tour through our current network management activities. Having grappled with the problems of traffic overloads for most of my time, let me close firmly and clearly on the note that network management methods are in no way a substitute for proper and adequate traffic engineering and equipment provision. Yet experience in the U. S. has shown that no matter how well we do the engineering job, in large common control networks good management methods are still needed. Some time ago Walt Hayward summarized that view very nicely, I think, by relating it to ship construction. He observed that sensible, conservative design requires lifeboats and water tight bulkheads, while their provision in no way lessens the need to construct the hull to stand up to the routine rigors of ocean voyaging.

REFERENCES

