CONGESTION MECHANISMS AND OVERLOAD CONTROL IN TELEPHONE NETWORKS

O. HASHIDA, S. NAKAJIMA, K. OKADA, M. SHINOHARA

Musashino Electrical Communication Laboratory, NTT, Tokyo, Japan

ABSTRACT

In telephone networks, exchanges and trunks are dimensioned to meet the required grade of service under estimated engineering loads. In practice, loads offered to a network vary largely over the estimated load levels. In case of network overload, call attempts to a part of a network run several to tens of times normal, as people call to check on the status of relatives and acquaintances. When these unusual overloads are offered to all or a part of a network, congestion phenomena occur without network controls.

In this paper, as congestion phenomena, throughput deterioration, lockup, hysterisis and so on are investigated, and their mechanisms are clarified, mainly by computer simulations. From consideration of congestion mechanisms, algorithms are proposed for area-focused overloads, which occur most frequently in the Japanese telephone network. It is confirmed that these algorithms are able to prevent throughput deterioration and that the throughput remains at normal levels, regardless of overload levels.

1. INTRODUCTION

The telephone network in Japan is a huge automatic system which consists of more than 6,000 exchange offices, more than 1.5 million toll transmission circuits and more than 40 million telephone terminals. Generally, switching systems and trunks are dimensioned to meet the required grades of service under the estimated engineering loads (normal and high loads). In practice, loads offered to the network vary over the estimated load levels. It is desirable to absorb annual and recurrent load variations in the dimensioned network provision. However, it is very uneconomical to absorb large load variations, which greatly exceed the estimated load levels, in the network provision. For example, in case of disastrous accidents, such as typhoons and earthquakes, call attempts to a part of the network run several to tens of times normal, as people call to check on the status of relatives and acquaintances.

Traffic congestions caused by these unusual overloads occur several tens of times a year in the Japanese telephone network. More than 95 percent of them are due to terminating calls focused to specific areas or subscribers (1). Causes of area-focused overloads are typhoons, earthquakes, floods and fires. Those of subscriber-focused overloads are telephone seat reservations and telephone requests for television or radio programs.

When unusual overloads are offered to a network without adequate control methods, the call completion ratio degrades because of network congestion (2)(3)(4). In this case, although network facilities are fully utilized, those calls which are holding or have used these facilities encounter blocking or time-out because of traffic congestions in other parts of the network, and the overall completion ratio of the network degrades. Therefore, network control methods are needed for preventing these congestion states and for making efficient use of network facilities. Network controls must retain functions to observe traffic in real time, to detect overload levels, and to control the network for preventing it from performance deterioration.

For network performance criteria, grade of service and network facility efficiencies are significant from the subscriber's point of view, and from the network administration's point of view, respectively. Since grade of service degrades badly under congested states, network efficiency is achieved as a criterion for measuring network performance. This paper uses throughput, namely the number of completed calls per time unit, as network efficiency, which corresponds to revenue production rate of a network.

When unusual overloads are offered to the whole or a part of a network without adequate network control, congestion phenomena, such as throughput deterioration, lockup and hysterisis can occur. Throughput deterioration is a phenomenon where throughput remains with increasing offered loads and is due to congestion propagation, network resource contention and so on. Lockup can occur when the resources are each occupied by calls requesting connection to the other and throughput reduces to zero. This is "direct store and forward lockup", and can be regarded as a kind of congestion propagation which is a cause of throughput deterioration. Hysteresis is shown by a hysterisis curve of throughput versus load characteristics: once throughput degrades under overload conditions, it only recovers from deteriorated states when the load is considerably reduced. This can occur due to various causes. Investigation of mechanisms for these congestion phenomena enables developing network control techniques to prevent throughput deterioration and to recover the network from deteriorated states.

Under overload conditions, in general, loss system is prone to become saturated in throughput characteristics and waiting systems, especially in tandem queues, is prone to fall into throughput deterioration because of inefficient usage of resources. In a telephone network, speech paths are established in a loss system and inter-office signals are dealt with in a waiting system. Therefore, it is necessary to pay attention to inter-office signaling schemes in investigation of congestion mechanisms.

2. THROUGHPUT DETERIORATION

Throughput of a toll telephone network increases in proportion to the offered load under light load conditions, and it reaches saturation as

ITC-9

HASHIDA / NAKAJIMA / OKADA / SHINOHARA-1
the load becomes heavy. When the load greatly exceeds the engineered load level, it is probable that the throughput will decrease inversely. This throughput deterioration characteristic depends greatly upon switching systems composing the network, especially upon signal sending and receiving procedures, which concern congestion propagation phenomenon.

Two kinds of networks were studied from the signal dealing procedure's point of view. One is separate register-sender network (S-network), whose switching systems (S-switching systems) are separately equipped with incoming-registers (IRs) and outgoing-senders (OSs). The other is combined register-sender network (C-network), whose switching systems (C-switching systems) are equipped with register-senders (RSs) which work as both IRs and OSs. Appendix 1 and 2 show the call processing flows concerning seizures of IRs and OSs in an S-switching system, and RSs in a C-switching system. Throughput deterioration is due to the mechanisms: incoming call increase, congestion propagation, and resource contention. These can be explained as follows, by using a network model of exchanges in tandem shown in Fig. 1.

(a) Incoming call increase
Exchange A is to be assumed under focused overload. As soon as the number of calls to exchange A increases, the trunk group from exchange B to A becomes congested and incompletely called due to outgoing-trunk (OGT) block increase in exchange B. Because of shorter trunk holding times for incompletely called calls, in comparison with those of completed calls, the average incoming-trunk (ICT) holding times in exchange B decreases, which keeps ICT group un congested. Instead of increasing calls to exchange A, in consequence, the number of calls arriving at IRs (RSs in case of a C-switching system) in exchange B increase markedly and IRs (RSs) reach congestion.

(b) Congestion propagation
When IRs (RSs) in exchange B is congested by the mechanism (a), OSs (RSs in case of a C-switching system) in exchange C are also becoming congested, owing to prolong of those holding times for OSs, which include waiting times for IRs (RSs) in exchange B. When exchange C is a C-switching system, RS congestion in exchange C provokes OS (RS) congestion in exchange D. Thus, in a C-network, congestion evolves through the medium of RSs more readily than in an S-network. Especially, lockup phenomenon can occur when RSs in two exchanges are each full with calls waiting for RSs in the other exchange. This phenomenon will be analyzed in the following Section.

(c) Resource contention
Even if OS (RS) efficiency remains very high in exchange C, most of OSs (RSs) are occupied by calls for exchange A (focused area) whose OS (RS) holding times are very long. Therefore, the traffic intensity carried by OSs (RSs) in exchange C decreases. Especially, the number of completed calls to non-focused areas decreases so markedly that total throughput for a network degrades.

Resource contention can be illustrated by a simple model shown in Fig. 2. Two kinds of calls share N resources corresponding to OSs (RSs) in exchange C (see Fig. 1): call 1 for a focused area and call 2 for a non-focused area. Let $\lambda_i$ and $\mu_i$ be Poisson arrival rates and exponential service rates for call i (i=1,2). The equilibrium state probabilities $P_{ij}$ that the numbers of calls 1 and 2 in the system are i and j, are given by the follows.

$$P_{ij} = c\left(\frac{\lambda_1}{\mu_1}, \frac{\lambda_2}{\mu_2}\right)$$

(C: normalizing constant)
Throughputs are obtained easily from this equation. Figure 3 shows throughput characteristics where the offered traffic intensity of call 1 ($=\lambda_1/\mu_1$) increases, while the offered traffic intensity of call 2 is fixed. When $\lambda_1/\mu_1$ exceeds 1.1, the total throughput deteriorates by the decrease in completed calls.

The above mentioned overload characteristics were confirmed by large-scale simulations. Network models with four hierarchical levels closely follow the actual telephone network in Japan. One example is shown in Fig. 4. The total throughput characteristics for this network, under focused overloads offered to office TC48,
are shown in Fig. 5. The C-network overload characteristics are worse than those of S-network, owing to congestion propagation and lockup phenomenon. Figures 6 and 7 show the numbers of incompleted calls classified by causes for S- and C-networks respectively. When the overload is not heavy, the principal cause for incompletion is "trunk block". Under a heavy overload condition, a certain switching system (for example, exchange B in Fig. 1) becomes congested and most of incompleted calls are due to "OS block" or to "time-out" rejection in seizing an IR (RS) of the next exchange. In this stage, throughputs deteriorate. Figure 8 shows throughput characteristics classified by destinations. Incompleted calls to non-focused areas are caused mainly by resource contention. This corresponds to the characteristics shown in Fig. 3.

3. LOCKUP

Lockup phenomenon can occur when RSs in two adjacent exchanges are both full with calls waiting for RSs in the other exchange and call connection process is halted, owing to lack of RSs in the exchange for which calls are bound. Similar phenomena are observed in a packet switched network, in a computer operating system, and so on (5)(6).

Consider a queuing network with two RS groups, RS1 and RS2, and two classes of calls, class 1 and class 2, as shown in Fig. 9. Each RS, has
time-out interval $T_i$: a call becomes lost when its waiting time exceeds $T_i$. The queuing discipline is first-come, first-served. A call of class 1 receives service first at $RS_1$, where the service time is $h_1$ seconds. When service at $RS_1$ is over, the call seeks $RS_2$. When $RS_2$ is available, the call proceeds to $RS_2$, where the service time is $h_2$ seconds. However when $RS_2$ is not available, the call waits for $RS_2$, while holding one of $RS_1$.

The mean flow rates $\lambda_i$ in the model can be calculated by flow conservation law, if the mean flow rates $\lambda_i$ for class $i$ and the probabilities $R_{ij}$ of call rejection rate at $RS_i$ are given:

$$\lambda_i = (1-R_i)\lambda_i (1+1,2)$$

$$\lambda_i = (1-R_i)\lambda_i (1+1,2,1)$$

The problem of finding probabilities $R_{ij}$, under the assumptions that the input is a Poisson arrival with rate $\lambda$, and the holding time is an exponentially distributed random variable with mean $\mu_i$, has been solved (7) (see Appendix 3). A set of equations can be solved by interactive method. Figure 10 shows throughput characteristics for completed calls $\lambda$, as a function of offered calls $\lambda_i$, where

$$\lambda = \lambda_1 = \lambda_2, \lambda_0 = \lambda_1 + \lambda_2$$

For a $\lambda_i$ in some range, two values of throughput are found. One value indicates the state where completion rate is high, waiting time is small and usage is normal. The other indicates the state where completion rate is low, waiting time is prolonged, nearly $T_i$, and usage is high, almost 100 percent.

To confirm this two-solution property, simulation was run on the same model. Figure 11 shows time-varying data for completed calls of both classes 1 and 2. Two kinds of seemingly stable states can be observed in it. One state, which is "non-lockup state", has high call completion rate, short queue and normal usage. The other, which is "lockup state", has low call completion rate, long queue and high usage.

4. HYSTERESIS

Once the throughput degrades due to overload, it does not recover, even if the offered load decreases. This is hysteresis phenomenon. One of the causes for this phenomenon is repeated attempt. Once repeated calls have increased, the trial number per subscriber is far more increased and the total number of offered calls, including repeated attempts, does not decrease as much as the amount of restricted subscribers.

In this section, this hysteresis phenomenon caused by repeated attempts will be analyzed by a simplified probabilistic flow model. Similar results have been obtained in the analyses of a multi-stage alternate routing network (8), a waiting system whose throughput depends on the number of calls staying in the system (9), and so on.

Consider a service system $S$ with repeated attempts shown in Fig. 12 and assume that $C$, calls carried by service system $S$ per second, is a function of $Z$, calls offered to service system $S$ per second.

$$C = C(Z)$$

The stationary flow rate equations in Fig. 12 by flow rate conservation law are

$$Z_1 + \lambda Z_2 + \delta Z_1 = f(Z_1)$$

where

$$\lambda : \text{fresh call flow rate (calls/sec)}$$

$$Z_2 : \text{blocked call flow rate (calls/sec)}$$

$$\delta : \text{reattempt probability}$$

From Eq. (3) and Eq. (2), we obtain

$$C(Z) = \frac{1-\delta}{\delta} Z_1 (1-\lambda)$$

Consider the case where service system $S$ has throughput deteriorating characteristics $C(Z)$, as shown in Fig. 13. Common control switching systems and their networks without effective control are examples (10). Figure 13 shows the solutions of Eq. (4) as the intersections of a curve and straight lines, which correspond to the left-hand side and the right-hand side in Eq. (4), respectively. For some values of $\lambda$ and $\delta$, there can be more than one intersection (a, b, and c in case of $\lambda = \Lambda$). The state is stable at the points a and c, but unstable at the point b. That is, at the point a or c, the state returns to where it was even if $Z$, deviates slightly, since the feedback
gain \(|\gamma|<1\), where \(\gamma\) is defined as follows.

\[
\gamma = \frac{\Delta f_f}{\Delta z} = \frac{f(\Delta z)}{\Delta z} \left(1 - \frac{sc(t)}{\Delta z} \right) (5)
\]

At the point b, if \(z\) deviates slightly due to stochastic fluctuation, the state greatly separates from where it was and reaches the point a or c, since the feedback gain \(|\gamma|\) is greater than one.

In case \(\lambda = \lambda_2\) and the state is at d, a slight increase in \(\lambda\) makes the state shift from d to e. This transition is due to positive feedback caused by the increases in blocked calls \(\Delta z_f\), in repeated calls \(\Delta z_r\), and in offered load \(z\). This transition from d to e is not reversible. Similar transition occurs when \(\lambda = \lambda_1\). Owing to these two avalanche like phenomena, the state changes in different manners, depending on increase or decrease of \(\lambda\). This is called "hysteresis phenomenon" and is illustrated in Fig. 14 as the relation between fresh calls \(\lambda\) and throughput \(c\).

\[
\text{Throughput} \ (\text{calls/sec})
\]

\[
\text{Offered calls} \ z_t \ (\text{calls/sec})
\]

\[
\lambda_1 \quad \lambda_2 \quad \lambda_2 \quad \lambda_1
\]

\[
\Delta z_f \quad \Delta z_r \quad \Delta z_f \quad \Delta z_r
\]

\[
\text{Transition}
\]

\[
\text{Fig. 13} \quad \text{Throughput as the solutions of Eq. (4)}
\]

Occurrence of this phenomenon depends on the values of \(f\) and the throughput characteristics of \(S\). A necessary and sufficient condition that the system shows hysteresis characteristics due to repeated attempts is that the following inequality holds for some \(z\):

\[
\Delta \frac{sc(t)}{\Delta z} < \frac{c(t)}{f} - 1 \quad \text{(6)}
\]

In other words, for a given throughput versus offered load characteristic, the hysteresis cannot occur if reattempt probability is small enough.

5. OVERLOAD CONTROL

Overload controls are required in order to prevent throughput deterioration and to maintain the maximum throughput of a telephone network. The basic principle of overload controls is to prevent the throughput deterioration mentioned in Section 2. From considerations of congestion mechanisms, cancellation of alternate routing might be adequate for incoming call increase and limitation of IR waiting time for congestion propagation and lockup. However, the main throughput deterioration mechanism is based on resource contention, in which calls having a low probability of completion to a focused area waste the network resources and disturb those calls bound for other areas, which would otherwise be successful otherwise. Therefore, the most effective overload control under focused overload conditions is to regulate originating calls to the focused area under the maximum throughput capacity of the network. The control becomes more effective as the regulation points situate nearer to the network entrances, i.e. originating local offices. This overload control is called "code blocking".

As quick responses are required for network congestion, automatic code blocking is desirable. Algorithms for code blocking are composed of detection of focused area codes and decision of regulation rates according to congestion levels. Two algorithms are proposed as automatic code blocking. One is based on trunk and IR or OS congestion states. When focused overloads are offered to a network, the incoming trunks and IRs of exchanges in the focused area, or IRs and OSs of preceding exchanges are congested, as stated in Section 2. The area codes which belong to the congested exchanges can be regarded as focused areas.

The trunk and IR states of toll incoming exchanges are supervised periodically. When congestion states exceed the predetermined threshold level, the area code for the exchange and regulation rate are transferred to toll outgoing exchanges, where calls for the focused area are restricted according to regulation rate. The regulation is strengthened or lightened according to the congestion states. Regulation rate is a function of the congestion states and the regulation rate at previous control period. When the congestion state and regulation rate are below the reset threshold level, the regulation is stopped.

In this algorithm, congested area is regarded as focused area. Hence, if congestion propagates to exchanges in the neighborhood of the focused area, the calls to non-focused areas, i.e. the neighborhood of the focused area, might be restricted.

The second algorithm is based on the incompletion rate classified by destination codes. When focused overloads are offered to the network, the incompletion rate for the focused area code becomes high as lost calls increase because of trunk block, OS block and time-out rejection in seizing IR. Therefore, area codes with high incompletion rates can be regarded as focused area codes. As the incompletion rate positively correlates with the focused overload level, required regulation rate can be estimated by the incompletion rate.

The problem in this algorithm is where to measure the incompletion rate. One method is to measure outgoing call incompletion rate at originating stages. In this case, each exchange can control originating calls autonomously by using the incompletion rate data measured by itself. The other method is to measure the incompletion rate of incoming calls close to terminating stages. The measuring spots are at the comparatively high level in the downward hierarchical chain. This measurement requires the function of transferring regulation area codes and regulation rates to outgoing stages as well as the first algorithm. The focused area codes are estimated more accurately, by the incompletion rates measured close to terminating stages, than by those at originating stages.
The effectiveness of overload controls for focused overloads was evaluated by computer simulation studies. The model is the same as Fig. 4 in Section 2. In addition to the normal engineered load, the traffic offered to TC#8 from all EOs except EO#3 was increased by several times. It was confirmed that the controls, such as cancellation of alternate routing and limitation of IR waiting time, were not sufficient under heavy focused overload and the code blocking was effective under the overload.

Figure 8 shows the effectiveness of the automatic code blocking algorithm mentioned above in this simulation. The throughput is shown as a function of a factor of engineered load to TC#8. Without control, the throughput begins to decrease at two times the engineered load. With control, the focused area is detected and calls to the area are restricted at the outgoing stages. From Fig. 8, it is confirmed that the algorithms are able to prevent throughput deterioration and to maintain the throughput at about the maximum network capacity, regardless of overload levels.

Figure 15 shows the characteristics of the algorithm under focused overload, which is based on incompletion rate classified destination codes measured at terminating stages. The incompletion rate and regulation rate are shown as a function of time from the onset of the focused overload. The incompletion rate for a focused area code begins to increase almost at the beginning of the overload and the regulation is started at the same time.

So far, it has been made clear that receiving and sending procedures for inter-office signals have large influence on occurrence of congestion phenomena in a telephone network with channel associated signaling system. However, a network with common channel signaling system would behave differently under overload conditions. The overload characteristics for it are also under study.

ACKNOWLEDGEMENT

The authors wish to acknowledge the continued guidance and the kind assistance of Dr. K.Habara, the director, and Dr. K.Gotoh, the deputy director of Switching Development Division in Musashino ECL.

REFERENCES

(4) L.A.Gimpelson, "Network Management: Design and Control of Communications Networks," Electrical Communication, Volume 49, Number 1, 1974.

6. CONCLUSION

Mechanisms of congestion phenomena, such as throughput deterioration, lockup and hysteresis, are investigated by simulations and approximate analyses for partial models. On the basis of the investigation results, two algorithms for overload controls are proposed for area-focused overloads, which occur most frequently in the Japanese telephone network, and their effectiveness is confirmed. At present, NTT is planning to introduce an area congestion control system, which automatically restricts calls destined for congested areas at the originating exchanges, all over the network (1).

Investigations are now in progress for subscriber-focused overloads, which occur most frequently next to area-focused overloads. A control algorithm, which detects focused subscribers by counting the number of blocked calls for each subscriber and regulates calls headed to focused subscribers at the originating local exchanges, is confirmed to be effective.
Appendix 1

Call processing flow for S-switching system

Appendix 2

Call processing flow for C-switching system

Appendix 3

By setting $ \theta_i = A_i, H_i$ and $P_i = \theta_i / N_i$, we get

$$R_i = \frac{\alpha_i}{N_i} \cdot e^{-\alpha_i / H_i} \cdot P(O_i)$$

$$W_i = \frac{\alpha_i}{N_i} \cdot \left( \frac{N_i / H_i + A_i + A_i \cdot T_i \cdot (N_i / H_i - A_i)}{N_i / H_i - A_i} \right)$$

$$P(O_i) = \left( \frac{\alpha_i}{N_i} \right)^{N_i / H_i}$$

Where

- $N_i$: The number of servers at RS_i
- $A_i$: The input flow rate at RS_i
- $H_i$: The effective holding time at RS_i

They are given as follows

$$\lambda_i = \lambda_{i1} + \lambda_{i2}$$

$$A_i = \lambda_{i1} + \lambda_{i2}$$

$$H_i = \frac{\lambda_{i1} + \lambda_{i2}}{\lambda_{i1} + \lambda_{i2}} + \frac{\lambda_{i1}}{\lambda_{i1} + \lambda_{i2}} \cdot (h_i + W_i)$$

$$H_2 = \frac{\lambda_{i2}}{\lambda_{i1} + \lambda_{i2}} \cdot (h_i + W_i) + \frac{\lambda_{i1}^2}{\lambda_{i1} + \lambda_{i2}^2} \cdot h_2$$