Fluid Model for a Traffic Congestion Prediction

Masataka OHTA
Systems Development Laboratory, Hitachi, Ltd.
1099, Ohzenji, Asao-Ku
Kawasaki-Shi, 215 JAPAN

ABSTRACT

Congestion prediction makes traffic management more efficient, thereby improving the grade of service. To predict congestion, it is necessary to calculate quickly the transient behavior that is about to occur in a network. Many works on transient analysis have been done using complicated calculation procedures. Those methods cannot be applied to congestion prediction.

In this paper, we present a fluid model for congestion prediction. In a congested situation, traffic flow in a network can be regarded as a deterministic process rather than a stochastic process. Thus, we can apply a fluid model to calculate the transient behavior of a network. Our fluid model makes calculation procedures simpler and faster. Furthermore, advanced intelligent networks make it possible to implement our proposed prediction method.

1. INTRODUCTION

With the advance of the information-oriented society, communication networks play an important role for human beings. To provide more advanced services, communication networks have been evolving to the Intelligent Network(IN)\(^1\). Because of this evolution, the scale of communication networks has become large. As the result of this growth, information sources in a network, such as a data base, are shared by many users. Thus, there will be increased occurrences of concentrated traffic accessing common information sources. Namely, the traffic is directed to the same node by the demand to access common information sources. It is known that such a state of concentrated traffic leads to an unstable state and congestion at one node sometimes is spread over the neighbor nodes\(^2\). To ensure network quality of services and efficient use of available network facilities, advanced traffic control is essential in network management.

Dynamic routing\(^3,4\) is an advanced traffic control technique which improves network probabilities by changing routing patterns according to the variation of the traffic offered to the network. However, changing routing patterns is not sufficient to cope with the congestion; traffic restriction is also necessary. Once a node becomes congested, the throughput is not improved immediately, even if the traffic restriction is executed at once. The throughput of the node shows the hysteresis curve\(^5,6\). If we can predict congestion of a node, we can avoid the congestion.

Congestion prediction makes traffic management more efficient, which improves the grade of service.

To predict congestion, it is necessary to calculate quickly the transient behavior that is about to occur in a network. Many works on transient analysis have been done. However, even in the most basic model (M/M/1\(^7\), M/M/S/K\(^8\) and M(t)/M/S/S\(^9\)), complicated and sophisticated calculation procedures are required. Diffusion approximation method\(^10\) also is complicated because it requires solution of a two-dimensional partial differential equation. Thus, these methods cannot be applied to congestion prediction.

To analyze a dynamic control of a network, Filipiak\(^11\) has proposed the method for the transient analysis using a fluid approximation. His basic idea is to describe the transient behavior of a system, such as M/M/S/K, as in the following equation:

\[
\frac{dx(t)}{dt} = -\mu G(x(t)) + \lambda(t)
\]

where \(x(t)\) is the number of customers in the system, \(\mu\) is the service rate and \(\lambda\) is the arrival rate. \(G(x(t))\) is approximated using the steady-state characteristics. However, \(G(x(t))\) is expressed by a high-order polynomial, and it is difficult to determine the polynomial which is the most suitable for the model. In an overload condition, the transient analysis can be done by the more simplified method\(^12\). In this paper, we adopt the method given in Ref.\(^{12}\) to predict congestion.

The outline of this paper is as follows. The basic approach is shown in Sec.2. In Sec.3, we expand the method in Ref.\(^{12}\) to analyze the M(t)/PHr/K system, and we show that our approach is quite effective to analyze the transient behavior in an overload condition. In Sec.4, we employ our approach to a circuit switched network and show the fluid model for a congestion prediction. Finally, in Sec.5, we show the numerical results in the fully connected network.

2. BASIC APPROACH TO CALCULATE THE BLOCKING PROBABILITIES

In this section, we show our basic approach to calculate the transient blocking probability in a finite capacity system which is required for the congestion prediction. Consider a system with finite capacity, which is shown in Fig.1. In Fig. 1, we let \(\lambda(t)\) and \(\alpha(t)\) be the mean arrival rate and the mean departure rate at time \(t\), and we regard \(\lambda(t)\) and \(\alpha(t)\) as a deterministic continuous process; namely, a fluid flow process is applied. Then we consider the transient blocking probability \(B(t)\) as

\[
B(t) = \frac{\lambda(t) - \alpha(t)}{\lambda(t)} \text{ (the usage rate of resource is 100\%)}
\]

\[
= \frac{\lambda(t) - \alpha(t)}{\lambda(t)} \cdot \frac{dS(t)}{S}
\]

(2)

where \(dS(t)\) denotes the probability that the usage rate of resource is 100\% at time \(t\), and \(S\) in \(dS(t)\) means the number of resources. The meaning of eq.\(^2\) is as follows. Under the
condition that the usage rate of resource is 100% (namely, all resources are occupied), this system can accept the offered load up to the number of departure customers at time \( t \), and the remaining offered load which exceeds the number of departure customers at time \( t \) is blocked.  

Next we consider \( o(t) \). Using the mean number of busy resources \( x(t) \), \( o(t) \) can be expressed as  
\[
o(t) = \mu x(t)
\]  
where \( \mu^{-1} \) is the mean holding time. Letting the state probability \( P_i(t) \) for the system shown in Fig.1 be  
\[
P_i(t) = \text{the probability that the number of occupied resources is } i \text{ at time } t.
\]  
\( x \) is defined as  
\[
x(t) = \sum_{i=0}^{S} i \cdot P_i(t)
\]  
When the system in Fig.1 is the M/M/S/S system, \( x \) satisfies the following differential equation.  
\[
dx(t) = -\mu(x(t) + \lambda(t) (1 - P_0(t)) - P_S(t))
\]  
Under the overload condition \( \lambda(t) > \mu - S \), we can consider  
\[
\begin{align*}
   \lim_{t \to \infty} x(t) &= S \\
   \lim_{t \to \infty} P_S(t) &= 1
\end{align*}
\]  
Then, eq.(6) is  
\[
\begin{align*}
   \frac{dx}{dt} &= -\mu x + \lambda & \text{for } 0 \leq x < S \\
   \frac{dx}{dt} &= -\mu x & \text{for } x = S
\end{align*}
\]  
Eq.(10) can be interpreted as the fluid model. In this paper we call eq.(10) the equation of motion. Next we consider \( dS(t) \) in eq.(2). Please note the following relation  
\[
dS(t) = P_S(t)
\]  
Namely, \( dS(t) \) means the probability that the usage rate of resource is 100%, and the other hand, \( P_S(t) \) has the other meaning of "the time congestion probability." For example, this difference can be found in the steady state probability. These are  
\[
\begin{align*}
   \lim_{t \to \infty} dS(t) &= 1 \\
   \lim_{t \to \infty} P_S(t) &= \text{Erlang-B}
\end{align*}
\]  
Eq.(12) means that in the overload condition, the probability that all resources are occupied will be \( \lambda \) and the time congestion probability \( P_S(t) \) will be expressed by the Erlang-B formula. When the input process is Poissonian process, \( P_S(t) \) can be regarded as the blocking probability. Then, the relation between \( dS(t) \) and \( P_S(t) \) is  
\[
P_S(t) = \frac{\lambda(t) - o(t)}{\lambda(t)} \cdot dS(t)
\]  
\( dS(t) \) can be obtained by the following manner. In the overload condition, once a resource is occupied, it never returns to idle status. Namely, as soon as a resource is released by a customer, it will be occupied by another customer. Assuming that the input process is Poissonian, the state transition process for the state that the occupancy rate of resources is \( \mu - S \) becomes the pure-birth process as shown in Fig.2. Solving the state equation for this pure-birth process, we have  
\[
dS(t) = 1 - e^{-\lambda(t)} \sum_{i=0}^{S-1} \frac{A(t)(S-1-i)}{(S-1-i)!}
\]  
where  
\[
A(t) = \int_0^t \lambda(s) ds
\]  
Using the results presented in this section, we can calculate the transient blocking probabilities easily. Namely, the transient blocking probability can be obtained by calculating eq.(2) with eq.(9,13) and \( \lambda(t) \).  

### 3. FLUID MODEL FOR A SINGLE NODE

The solution of eq.(10) is  
\[
x = e^{-\mu t} \left[ e^{\lambda t} x_0 + \int_0^t \lambda(s) e^{-\mu(s-t)} ds \right] \text{ for } 0 \leq x < S
\]  
When \( \lambda(t) \) is equal to 0, \( x \) decreases exponentially. Thus we can consider that the life time of \( x \) has an exponential distribution and the system which can be described by eq.(10) has an exponential service time. Furthermore, we regard this system as one stage in a phase type distribution. Fig.3 shows the M/COX/SS system. We express the behavior of the number of customers in a phase in Fig.3 using eq.(10). Namely, letting \( x_i \) be the number of customers in the phase \( i \), we have  
\[
\begin{align*}
   \frac{dx_i}{dt} &= -\mu_i x_i + g(t) \\
   \frac{dx_r}{dt} &= -\mu x_r + \beta_{r-1} \mu_{r-1} x_{r-1}
\end{align*}
\]  
where  
\[
g(t) = \lambda(t) \sum_{i=1}^{r} \mu_i x_i \\
g(t) = \sum_{i=1}^{r-1} \mu_i x_i + \mu x_r \text{ for } \sum_{i=1}^{r} x_i = S
\]  
Eq.(19-a) means that when the number of the occupied resources is smaller than \( S \), the system can accept all of the offered load. Eq.(19-b) means that the system can accept only the offered load as much as the number of departing customers. Moreover, in a phase-type distribution, \( o(t) \) can leave the system. Then, \( o(t) \) in eq.(2) is  
\[
o(t) = \sum_{i=1}^{r} \beta_{i} \mu_i x_i + \mu x_r
\]
Using eqs. (2), (13) and (16), we can calculate the transient blocking probabilities in the M/COX_r/S/S system.

To obtain the explicit expression, we consider the M/E_r/S/S system. We let
\[ a_i = 1, \quad \beta_i = 0, \quad \mu_i = r \mu \quad \text{for } i = 1, \ldots, r \]  
and to simplify the analysis we let
\[ \alpha = 1, \quad \beta = 0, \quad P_i = r \quad \text{for } i = 1, \ldots, r \]  
(21)
and to simplify the analysis we let
\[ M_1 = 1. \]  
the solution of eq. (18) with the initial condition
\[ x_j(0) = 0 \quad (i = 1, \ldots, r) \]  
for \( EX_i \leq \beta \) is
\[ x_i(t) = \frac{1}{r} - \frac{1}{(i-1)!} (\mu r t)^i \quad \text{for } \sum_{j=1}^r x_j < \beta \]  
(22)
and the solution for \( EX_i = \beta \) is
\[ x_i(t) = \sum_{j=1}^r y_j e^{-\mu r t} \quad \text{for } \sum_{j=1}^r x_j = \beta \]  
(23)
where
\[ y_r = \sum_{l=0}^{r+1} y_l \quad \text{for } r = 2m \]  
(24-a)
\[ y_r = \sum_{l=0}^{r} y_l \quad \text{for } r = 2m + 1 \]  
(24-b)
and
\[ y_l = \exp(\mu r \cos(2\pi l r)) \quad \{ A_l \cos(\mu r \sin(2\pi l r) t) + B_l \sin(\mu r \sin(2\pi l r) t) \} \]  
(25)
In eq. (24), we let \( m \) stand as a positive integer. We can obtain \( x_i \) \((i = 1, \ldots, r)\) for \( \sum_{j=1}^r x_j = \beta \) easily using eq. (18). \( A_l \) and \( B_l \) in eq. (25) are integral constants which can be determined by the initial conditions. Using the results presented here, \( o(t) \) for the M/E_r/S/S system is given by
\[ o(t) = \lambda \left( 1 - \sum_{i=1}^r \frac{(\mu r t)^i}{(i-1)!} e^{-\mu r t} \right) \quad \text{for } \sum_{j=1}^r x_j < \beta \]  
(26)
\[ o(t) = \mu r \cdot y_r e^{-\mu r t} \quad \text{for } \sum_{j=1}^r x_j = \beta \]  
(27)
Fig. 4 shows the numerical results for the M/E_3/10/10 system and the exact transient solution which is obtained by solving the state equations numerically. This figure shows our approach is accurate. Fig. 5 shows the numerical results with several cases \((r = 1, 3, 5, 10, 20, 50)\). We can see that the curve oscillates for large \( r \) values. We can understand this phenomenon by considering the M/D/S/S system because the \( r \)-stage Erlang distribution approach to the constant distribution when \( r \to \infty \). When an overload is applied to the system which has a constant service time, all resources will be occupied in a moment. This situation will continue in a constant time. During this situation, all offered loads are rejected and the blocking probability is 1. After this period, all resources are released in a moment, and then all offered loads are accepted by the system until all resources are occupied. During this period, the blocking probability is 0. Thus, in the M/D/S/S system, the blocking probability oscillates between 0 and 1. The M/E_r/S/S system approaches the M/D/S/S system as the number of stage \( r \) becomes large. Therefore, the transient blocking probability shows an oscillation.

Figs. 6 and 7 show the numerical results in another phase-type distribution with the exact solutions obtained by solving the state equations numerically. These figures also show our approach is accurate and acceptable for calculating the transient blocking probabilities in an overload condition.
4. FLUID MODEL FOR A COMMUNICATION NETWORK

In this section, we employ the basic approach mentioned in Sec.3 to a communication network. Consider a circuit switched network as a communication network. In a circuit switched network, when a new call encounters a congestion of one route, it selects another route using a routing algorithm as shown in Fig.8, where \( R(i,j,k) \) is defined as

\[
R(i,j,k) = k - \text{th route for the destination } j \text{ at office } i. \tag{28}
\]

Fig.8 A routing strategy

To apply the fluid model shown in Sec.2 and 3, we introduce the state transition diagram shown in Fig.9. Fig.9 represents a routing strategy. A new call tries to seize a circuit in the first route. If the call finds an idle circuit, the service of the call will be completed with the mean holding time \( 1/\mu \) sec. If the call finds that all circuits are occupied, the call tries to seize a circuit in the 2nd route. If the call finds an idle circuit, the service of the call will be completed with the mean holding time \( 1/\mu \) sec. If the call finds

\[
\begin{align*}
\frac{dx(i,j,i,R(i,j,1))}{dt} &= -\mu x(i,j,i,R(i,j,1)) + B(i,j) \lambda(i,j) \tag{29-a} \\
\frac{dx(i,j,i,R(i,j,2))}{dt} &= -\gamma x(i,j,i,R(i,j,2)) + B(i,R(i,j,1)) \lambda(i,j) \tag{29-b} \\
\frac{dx(i,j,i,R(i,j,2),j)}{dt} &= -\mu x(i,j,i,R(i,j,2),j) + B(i,R(i,j,2),j) \gamma x(i,j,i,R(i,j,2)) \tag{30-a} \\
\frac{dx(i,j,2)}{dt} &= -\mu x(i,j,2) + B(i,R(i,j,2),j) \gamma x(i,j,i,R(i,j,2)) \tag{30-b}
\end{align*}
\]

where

\[
\begin{align*}
\lambda(i,j) &= \text{mean call originating rate for the destination } j \text{ at office } i \\
\mu &= \text{mean holding rate} \\
\gamma &= \text{mean call completion rate}
\end{align*}
\]

The meanings of \( \lambda(t) \), \( \mu(t) \), and \( S(t) \) in eq.(2) for \( B(i,j) \) are

\[
\begin{align*}
\lambda(t) &= \text{call arrival rate to the circuits between office } i \text{ and } j \\
\mu(t) &= \text{call completion rate for the circuits between office } i \text{ and } j \\
S(t) &= \text{probability of the usage rate}
\end{align*}
\]

Moreover, to reduce the amount of calculation time, we adopt other approximations. These are

\[
\begin{align*}
\lambda(t) &= \sum \frac{1}{k} \text{ for } X(i,j) \leq S(i,j) \\
\mu(t) &= \mu \text{ for } X(i,j) = S(i,j) \\
S(t) &= 100\% \text{ for } X(i,j) = S(i,j)
\end{align*}
\]

The equation of motion eq.(29),(30) and (31) by numerical calculation. We show the calculation procedure in Fig. 10.

By adopting the present states as initial conditions, which are the present number of occupied circuits and the present call originating rates, we can predict a congestion in a communication network.
5. CONGESTION PREDICTION

Let us apply the congestion prediction method in Sec.4 to the fully connected network shown in Fig.11. To solve the equations of motion, we adapt the following assumptions.

1) The circuits are in the both-way operation.
2) The routing is performed only by the originating office.
3) The 1st route is the direct route to the terminating office and other routes have only one transient office.
4) After the call encounters congestion at the tandem office, the originating office performs the routing only one time.

![Fig.11 The fully connected network](image)

The blocking probabilities in Fig.11 can be obtained by the following manner. To calculate the blocking probabilities, we must know the call arrival rate to the circuits between office $i$ and $j$. Letting $A(i,j)$ be this arrival rate, $A(i,j)$ is given by

$$A(i,j) = A_0(i,j) + A_1(i,j) + A_2(i,j) + A_3(i,j)$$

where

$$A_0(i,j) = \lambda(i,j)$$

$$A_1(i,j) = \sum_{l,m \in R(i,l,m)} B(i,R(i,l,k))$$

$$A_2(i,j) = \sum_{l,m \in R(i,l,m)} B(R(i,l,m-1), 0, \gamma \cdot x(i,l, i, R(i,l,m)))$$

$$A_3(i,j) = \sum_{l,m \in R(i,l,m)} \gamma \cdot x(k, i, k, j)$$

$A_0$ means the offered traffic to the first route. $A_1$ means the offered traffic after the $m$ times routing, as shown Fig. 12(a). $A_2$ means the offered traffic encountered congestion at the tandem office $j$ after $m-1$ times routing as shown in Fig.12(b). $A_3$ means the offered traffic at the tandem office $j$ after $m$ times $m \equiv 2$ routing as shown in Fig. 12(c). Using eq(35),(31),(32) and (33), we can predict the congestion with the procedures shown in Fig.10. Moreover, the throughput of the calls originating from office $i$ destined for office $j$ can be calculated

**Throughput for**

$$\lambda(i,j) = \text{call completion rate / offered rate}$$

$$= \sum_{k=1}^{N} x(i,j, R(i,j,k), j) / \lambda(i,j)$$

By evaluating eq.(39), we can predict whether throughput degradation will occur in the near future. By evaluating eq.(33), we can know the cause of the throughput degradation.

The calculation result is shown in Fig.13 where we assume $N=4$ and the offered load from office 1 to office 2 changes 2.5 Erlang to 125 Erlang at time 0. Fig.13 shows the throughput of the whole network and the throughput of the individual origin-destination pairs. Using these calculation results, we can predict that the overload of the one origin-destination pair(1-2) will decrease the throughput of the whole network in the near future.

![Fig.12 The traffic patterns offered to the circuits i-j](image)

![Fig.13 An example of prediction](image)

Applying this method for congestion prediction, we present an advanced network traffic management system, which is shown in Fig.14. The advanced intelligent networks will make it possible to implement our proposed prediction method.

![Fig.14 The network traffic management system](image)
6. CONCLUSIONS

We have presented a fluid model which enables us to predict traffic congestion. The accuracy of our fluid model has been confirmed by comparison with exact solutions in simple models. We have applied our prediction algorithm to the N-fully connected network.

Our algorithm for the prediction is quite simple, and the calculation procedure also is simple. Our algorithm can be applied to any type of network. However, the existing public networks have more than 100 nodes. For this case, calculation time will be long. Thus we need further study to improve the calculation speed, such as application of an analog computer which is suitable for solving differential equations. In small scale networks, such as private networks, our method can be applied. Therefore, our present target is the private network.

ACKNOWLEDGMENT

The author would like to thank Professor Tadao Saito of University of Tokyo and Dr. Yoshitaka Takahashi of NTT Communication Switching Laboratory for participating in helpful discussions and their useful comments. The author is also grateful to Dr. Joe Brand of Hitachi Telecom USA and Dr. Zenya Koono, Dr. Koichi Yamamoto and Mr. Kazuhiko Ohmachi of Hitachi Ltd. for their helpful comments.

REFERENCES