Performance Specification, Call-processing Capacity, and Overload Control for
SPC Systems with Extremely High Peak Traffic

Shuichi SUMITA and Shin-ichi NAKAGAWA
NTT Laboratories
3-9-11, Midori-cho, Musashino-shi, Tokyo 180, Japan
Phone: 81-422-59-4440, Fax: 81-422-59-3290, e-mail: sumita@hashi.ntt.jp

Abstract
This paper analyzes real traffic data collected from our ISDN systems, and determines the
effects on call-processing capacity and overload control. This analysis leads us to conclude that
the BHCA definition alone is not sufficient to represent call-processing capacity and therefore it
should be extended in order to adapt to a new traffic condition that we are facing, taking account
of the non-stationarity of traffic. We propose a set of switching system performance
specifications, which is based on a traffic model abstracted from the collected data. Our traffic
model is a mixture of stationary Poisson traffic and momentarily focused traffic. This
momentarily focused traffic, which we call impulse, is approximated by batch traffic. Based on
this model, we propose the acceptable height of an impulse and the acceptable interarrival time
of impulses as additional performance specifications, together with BHCA. The additional
performance specifications are determined so as to meet GOS requirements that are in effect
during some short interval after the impulse arrives at the system. Finally, we also propose an
overload control scheme in accordance with a new set of performance specifications.

1. Introduction

Switching systems that support narrow-band ISDN services have been in
operation in Japan since 1987. Sufficient traffic data for ISDN circuit switched
services has been collected to identify traffic characteristics. It is now possible to
compare these with the characteristics of telephone traffic; the traffic stream at a local
telephone switching node is Poissonian, because a large number of users are
accommodated at a node and each user generates only a few calls per unit time.

Traffic characteristics depend on the users' behavior. Interestingly, this makes
ISDN traffic characteristics slightly different from telephone traffic characteristics.
Traffic data collected at an ISDN switching system reveals that traffic is sometimes
focused at a switching node during an extremely short interval (for example, 1 sec.), and this momentarily focused traffic
happens only a few times during the busy
hour, however the number of calls offered during this short interval exceeds the call-handling capacity of that switching node. This focused traffic is mixed with stationary Poisson traffic, so that a resulting traffic type that we observe over one hour is a mixture of stationary and non-stationary traffic types.

What we should note in handling this mixture of traffic is that the focused traffic causes the system to change from a stationary situation to a transient one every time it comes about at the system. This means that the assumption of stationarity for traffic, and consequently the stationarity of system behavior no longer holds and thus the engineering of the switch should be carried out under non-stationary conditions. Non-stationarity, in turn, raises new questions concerning traffic engineering, in particular GOS requirements for the switch, switch capacity specifications and representation, and overload control schemes. The following questions are specifically important:

- Is the definition of BHCA sufficient to represent call-processing capacity even for non-stationary traffic?
- If not, how can we define a call-processing capacity that is appropriate for non-stationary traffic?

The aim of this research is to answer these questions by focusing on a particular traffic type that was found in our ISDN systems.

This paper first presents real traffic data collected in our ISDN systems, and analyzes its effects on both call-processing capacity and overload control. These analyzes lead us to conclude that the BHCA definition alone is not sufficient to represent the call-processing capacity. Therefore, to adapt to new traffic demands, the definition should be extended to take the non-stationarity of traffic into account. We propose a set of switching systems performance specifications, which is based on a traffic model abstracted from the collected data. One notable feature of our traffic model is that the momentarily focused traffic is approximated by batch traffic, which is called impulse in this paper. Together with BHCA, the acceptable height of an impulse and the acceptable interarrival time of impulses should also be specified. Numerical examples are provided based on the new performance definition. Finally, an overload control scheme is proposed.

2. ISDN Traffic Characteristics

Figure 1 illustrates an ISDN system configuration [1]. The D70 is a digital switching system for local applications, and accommodates analog subscriber lines. The ISM is an interface module, which provides circuit termination, concentration, and intermodule distribution for 64Kbit/s, 384Kb/s, and 1.536Mb/2 channels and packet multiplexing. We collected traffic data for ISDN circuit-switched services to identify traffic characteristics and to compare them to telephone traffic characteristics.

Figure 2(a) shows an example of traffic characteristics at one ISM. The
number of originating calls, or equivalently, the number of set-up signals per two seconds in one hour is plotted. We see that the traffic is not peaky. In fact, we confirmed that the traffic is almost Poissonian by computing the squared coefficient of variation of the number of originating calls. There are no particular problems with handling this type of traffic because the methods established for telephone traffic can be used to design the switching system.

Figure 2(b) shows an example of traffic data collected at another ISM. We see in this figure that the traffic is sometimes focused during an extremely short period and the focused traffic happens a few times in one hour. There are two explanations about why this focused traffic takes place at the ISM. One is due to the following users' circumstances; one PBX installed at the head office in a company automatically generates many calls simultaneously, each of which connects between the head office and one branch of the company; after connections are established, the same documents are sent from the head office to branches. The set-up signals for these connections arrive continuously at the ISM through the D-channel. The other case occurs when different users may try to establish connections at virtually the same time.

A significant feature of Fig. 2(b) is that impulses come to the switch only a few times per one hour, but the height of one impulse, i.e., the number of simultaneous arrivals, is much higher than the number of arrivals at other times. This means that an impulse itself may cause a momentary overload if the number of simultaneous arrivals is far greater than the call-processing capacity represented by BHCA. By the term "momentary" we mean that if the interval between two consecutive arrivals of impulses is sufficiently large, the overload caused by the impulse might disappear before the next impulse comes. However, the impulse itself momentarily causes the grades of services for calls in progress worse.

Overload control is therefore necessary even for the momentary overload, just as it is necessary for an overload of long duration [2]. A question of how many simultaneous arrivals are acceptable is crucial to designing a control scheme, because the control scheme needs this number in the following way; the impulse can be permitted to enter the system if its height is lower than the threshold, otherwise it is regulated.

One possible way to determine the threshold is to take a small interval $t$, which is on the order of a second, and to make it equal to $Vt/3600$, where $V$ is the call-processing capacity of the ISM represented by BHCA. This linear transformation of BHCA into the capacity for a short interval has a weak point because BHCA explicitly indicates nothing about how many call attempts the system can handle during an extremely short interval. Furthermore, the BHCA definition does not assure a particular user of a particular GOS, though it does
assure a whole set of users accommodated to
the switch that GOS requirements are met if
the switch operates at below its capacity.
This could be a serious problem if all
impulses came from the same user, because
that user would think that the GOS is always
bad. In addition to BHCA, we should
therefore specify the GOS that is defined for
only a short interval occurring just after an
impulse arrives. Accordingly, we should
also introduce a set of new call-processing
capacity definitions in order to design and
engineer a control system that can handle the
impulse traffic shown in Fig. 2(b).

3. Performance Specifications

First of all, we model the traffic
offered to the ISM as shown in Fig. 3. This
model is abstracted from Fig. 2(b),
representing a mixture of two types of
traffic: Poisson traffic, which is
characterized by mean rate \( \lambda \), and an impulse
traffic, which is characterized by two
parameters \((m, b)\). The parameter \( m \) is the
maximum number of simultaneous arrivals,
and the parameter \( b \) is the interval between
two consecutive arrivals of impulses.

We should add both the maximum
acceptable number of simultaneous arrivals
(the impulse height), \( m_{\text{max}} \), and acceptable
interarrival of impulses, \( b_{\text{min}} \), to the
ordinary performance measure BHCA, \( V \).
From this we propose a triple
\((V, m_{\text{max}}, b_{\text{min}})\) as a performance measure
for switching systems. The ordinary BHCA
represents the capacity for a long period
(typical time-scale is one hour), i.e., the call-
processing capacity when the system is
subject to stationary traffic [3]. A pair of
\((m_{\text{max}}, b_{\text{min}})\) represents the performance
over a short period (typical time-scale is on
the order of tens of seconds), especially the
performance just after the system is subject
to a sudden surge in traffic such as an
impulse.

After defining the performance in this
way, we go on to define the GOS
specification during a short period, which is
closely related to transient characteristics of
the system. We define the following three
items as performance measures during a
short interval [4], but use only one in our
calculations. First, suppose that the impulse
arrives at time \( t = 0 \).

[Queue length]
Let \( L(t) \) be queue length in the system at time
\( t \). Let \( R_1(T, k) \) be the ratio of \( T \) to the total
time during which the queue length \( L(t) \) is
greater than the threshold \( k \) over an interval
\((0, T)\). \( R_1(T, k) \) is given by

\[
R_1(T, k) = \frac{1}{T} \int_0^T 1(L(t) > k) \, dt,
\]

where \( 1(\cdot) \) is an indicator function. The
requirement for \( R_1(T, k) \) is specified.

[Waiting time]
Let \( t_n \) be the time when the \( n \)th call
originates. Let \( W_n \) be the response time for
the \( n \)th call. The performance measure is
the ratio of the number of originating calls
during \((0, T)\) to the number of calls during
\((0, T)\) whose response time is greater than
the threshold \( w, R_2(w, T) \). This is given by

\[
R_2(w,T) = \frac{\sum_{n=1}^{\infty} I(t_n \in [0,T]) \cdot I(W_n > w)}{\sum_{n=1}^{\infty} I(t_n \in [0,T])}
\]

The requirement for \( R_2(w, T) \) is specified.

[Recovery time]

In this case, the performance measure is the time from \( t = 0 \) to the first time thereafter when the system falls to the same state as \( t = 0 \). The requirement for this time should be specified.

The approach that we take is to use the second item (waiting time) and third item (recovery time). First, we determine \( b_{\text{min}} \) as the time interval from the impulse arrival to the first time thereafter when the system is restored to the same state as just before the impulse arrival. Queue length or processor utilization can be chosen as the state of the system. Next, we specify the GOS requirements as the GOS defined during \( b_{\text{min}} \).

In our ISDN system, three principal measures are defined: call set-up delay, alert signaling delay, and connection release delay, as shown in Fig. 4. For each delay, we have standard specification values that should be satisfied during the busy hour. Thus, the call-processing capacity represented by BHCA, \( V \), is determined to satisfy these requirements. For example, 95% of the delay distribution of A+B is less than 180 ms and 95% of the delay distribution of C is less than 100 ms. In the same way, a standard value for the GOS during some short interval that starts just after the impulse arrives should be provided. We provide less strict specifications for both A+B and C during \( b_{\text{min}} \). Ninety-five percent of the distribution of A+B during \( b_{\text{min}} \) is less than 2.9 second; the maximum of C during \( b_{\text{min}} \) is less than 500 ms. The value \( m_{\text{max}} \) is determined to satisfy these GOS requirements during \( b_{\text{min}} \).

Figure 5 shows an example of the maximum acceptable simultaneous arrivals \( m_{\text{max}} \) and the acceptable impulse interval \( b_{\text{min}} \). This figure shows how these parameters vary with the increase of switch capacity represented by BHCA. The horizontal axis shows the ordinary switch capacity when only stationary Poisson calls are offered to the system during one hour. This example shows the number of Poisson call attempts that a processor can handle at processor utilization of 95%. The vertical axis shows both \( m_{\text{max}} \) and \( b_{\text{min}} \) when stationary Poisson calls with a load equal to processor utilization of 70% are offered to the system as background calls, and the remaining real time, which is equal to a processor utilization of 25%, can be allocated to the processing of impulse traffic. Thus, \( b_{\text{min}} \) is the time duration from the impulse arrival to the first time thereafter when processor utilization falls to 70%.

The figure can be interpreted in the
following way. If a processor call-processing capacity is $X$ BHCA, as indicated in the figure, and it is operating at a processor utilization of 70\%, then the acceptable number of simultaneous arrivals should be less than $M$ and the interarrival of two impulses should be greater than $D$, to guarantee the GOS for short intervals.

The acceptable number of simultaneous arrivals increases as the BHCA increases. The acceptable interarrival of impulses decreases slightly as the BHCA increases. Although the requirements for $A+B$ and $C$ should both be satisfied, in most cases $m_{\text{max}}$ is determined according to the requirement for $C$.

### 4. Overload Control

In the previous section we discussed the necessity for specifying GOS during a short interval, and provided a method for defining it. We also provided the acceptable impulse traffic type according to this specification of short-term GOS requirements. Our next step is to design an overload control scheme that is in accordance with this short-term GOS specification.

First, we will identify two types of overload. One type is caused by impulse traffic, which happens if the height of the impulse is greater than the acceptable number of simultaneous arrivals $m_{\text{max}}$, or if the interarrival time between two impulses is less than the acceptable interval $b_{\text{min}}$. The second overload type is caused by Poisson traffic, which occurs if the arrival rate is greater than the call-processing capacity represented by BHCA. This type of overload continues for a long time, as opposed to momentary overload.

Momentarily focused traffic could be smoothed by delaying it. This would be acceptable if the delay caused by smoothing could meet the GOS specified during a short period after the impulse arrives. Thus, a regulation mechanism for momentary overload should be considered along with the smoothing mechanism. We have implemented a smoothing mechanism by changing the queueing structure, as shown in Figure 6. The smoothing works in the following way; in normal situations, all messages arriving at the system go through Q1 and Q2. If the system detects an impulse, it routes traffic to Q3. The messages routed to Q3 are periodically transferred to Q1. Once they are transferred to Q3, they are processed as in normal situations. Queue Q3 is used for smoothing momentarily focused traffic.

The features of our proposed control algorithm are a detection mechanism for long and short overloads, and a smoothing mechanism as mentioned above. The long-term overload can be detected by processor utilization that is measured during a fixed time interval. The length of this interval is denoted by $Y$. The momentary focused traffic should be detected by the number of originating calls. This number should be measured at an interval of less than $Y$. If the
long-term overload is detected, a regulation mechanism is activated. If the number of originating calls during a fixed short interval is greater than some threshold, a smoothing mechanism is activated. Further, if this number exceeds the acceptable number of simultaneous arrivals \( m_{\text{max}} \), a regulation mechanism is activated. The main difference between this control scheme and previous ones is that there are two observation points: one is for long-term overload and the other for short-term overload.

Figures 8-(a) and 8-(b) show examples of transient characteristics of the system during 40 seconds after the impulse arrives, which were obtained by computer simulation. In both figures, the vertical thick lines show the number of originating calls per 2 seconds; the solid line shows transient behavior of processor utilization which is measured every 2 seconds. In our simulations, normal processor utilization is 60%, i.e., Poisson calls are offered to the system up to processor utilization of 60%. In case (a) the impulse with 40 simultaneous arrivals happens at time 5 seconds; in case (b) 60 simultaneous arrivals at the same time. In both cases a smoothing mechanism is activated immediately after the impulse arrives, and continues to operate until the number of originating calls per 2 seconds falls below some threshold. The number of 40 simultaneous arrivals is acceptable, while the number of 60 simultaneous arrivals is unacceptable. Therefore, we see from Fig. 8-(b) that regulation is activated from 8 seconds to 16 seconds, where the regulation is released when the processor utilization falls below 70%. Here note that we simulated only overload control that can be activated if processor utilization exceeds the threshold (95%), because we would like to understand processor overload characteristics if we do not implement overload control for short-term overloads. We see that even if we do not implement overload control that uses the maximum acceptable number of simultaneous arrivals \( m_{\text{max}} \), a regulation mechanism for long-term overload certainly works well, and it can regulate even short-term overloads, however, it is activated too late. Therefore, it is useful to implement an overload control mechanism for short-term overloads, though it complicates the call-processing mechanism.

5. Conclusion

We addressed the necessity of GOS specification for short intervals in order to do traffic engineering under non-stationary traffic, and provided a method for defining it. We also provided the acceptable traffic type according to this specification of the short-term GOS requirements. In addition, we presented overload control in conjunction with a smoothing mechanism to cope with extremely high peak traffic and showed that this overload control scheme exhibits fairly good characteristics.

One remaining problem is to extend what we propose in this paper to more general non-stationary traffic.
References

Figure 1 ISDN switching system configuration

Figure 2 ISDN traffic characteristics example
Figure 3 Traffic model for performance specification

Figure 4 GOS example

Figure 5 Acceptable batch size and interval

Figure 6 Queueing structure for smoothing
Basic Principle:

- Coping with short-term (or batch calls) and long-term overloads
- Two detection points for both of these overloads
- Short-term overload (or batch calls) is accepted as much as possible, by smoothing it.

1. **Measure some items**
   - Processor utilization $\eta$
   - Number of set up signals during an extremely short period, $m$

2. **Judge whether or not overload happens**
   - $\eta > \eta_{\text{max}} \rightarrow$ Activate regulation
   - $m > m_{\text{max}} \rightarrow$ Activate regulation
   - $\eta_{\text{max}}$ : Maximum acceptable processor utilization (95%)
   - $m_{\text{max}}$ : Maximum acceptable batch size
   - $\eta \leq \eta_1 \rightarrow$ Release regulation

**Figure 7 Overload control scheme**

**Figure 8 Overload characteristics examples.**