PERFORMANCE ANALYSIS AND APPLICATION OF A TWO-PRIORITY PACKET QUEUE

AT&T TECHNICAL JOURNAL

Chung-Yin Lo

Chung-Yin Lo *is a member of technical staff in the ISN (Information Systems Network) Software Development Department at AT&T Information Systems in Lincroft, New Jersey. He works on TIM (trunk interface module) performance analysis and ISN network management. Mr. Lo joined AT&T Bell Laboratories in 1982 and has a Ph.D. in mathematics from Queen's University (Kingston, Ontario, Canada) and a D.Sc. in physics from Massachusetts Institute of Technology.*

This paper analyzes the performance of a packet nonpreemptive (PNP), two-priority queue that is modeled as a message M/G/1 system. We derive the average in-queue waiting time for a packet of given priority and the average message delays. To overcome mathematical complexity, this derivation employs the well-known but seldom used G/G/1 conservation law. Because of the PNP discipline, the moments of low-priority traffic do not influence the average delay of the high-priority traffic. This model is compatible with the queueing discipline of the ISN trunk interface module (TIM). Among all the known practical queueing disciplines, the TIM discipline results in the shortest delay for short messages at reasonable expense to the long messages. Furthermore, the TIM discipline favors middle-sized time-sharing messages over long file-transfer messages, although not to the same extent as the round-robin algorithm.

Perspective

Information Systems Network (ISN) is AT&T's entry into the local-area network (LAN) market. With enhancements, AT&T ISN can be operated as a wide-area network (WAN), integrating geographically dispersed LANs—a distinct feature. Toward this end, the trunk interface module (TIM) has been developed[1] to transport data at 9.6 kb/s to 1.544 Mb/s over digital facilities. Alternate routing capability to ensure network reliability is under development.[2]

The queueing delays introduced when messages of widely different lengths share a trunk through the TIM are critical to the performance of long-haul computer-communication networks that are based on packet switches. In particular, one does not want long file transfers to delay single keystrokes from asynchronous terminals or short time-sharing messages. This is the basic drawback of the first-come,first-served

(FCFS) queueing discipline.

Short Messages and Queueing Disciplines. The usual way to favor customers who require short service is to force anyone who needs more than a given amount of service to relinquish the server periodically and wait for another turn.

For instance, with the round-robin algorithm, arriving customers join the end of a single queue; work their way to the front; receive a quantum of service; and, if their total service requirement has not been met, return to the end of the queue. Although round robin favors customers who require short service, it forces any arriving customer to wait, not only for earlier customers who have not yet received any service, but also for those who are awaiting their second, third, etc., quantum of service.

AT&T's Datakit® virtual circuit switch (VCS) is designed even more to favor customers who require only one quantum of service. For Datakit VCS, Riddle[3] proposed a two-queue discipline in which arriving customers join an FCFS queue and then join a round-robin queue, if they require more than one quantum of service. The FCFS queue has nonpreemptive priority over the round robin. When the server completes each quantum of service, it checks if any customers are waiting in the FCFS queue and attends to them first. This Datakit queueing discipline was analyzed extensively for 56-kb/s trunks by Fraser and Morgan[3] who used a message M/G/1 model.

In ISN, the TIM implements a new, two-priority trunk queueing discipline that Hluchyj[4] proposed. With this queueing discipline, short interactive messages (e.g., single-character traffic-requiring echoes) are given priority over larger messages that fill several ISN packets. Moreover, instead of maintaining separate queues in the trunk module for each channel (Riddle's approach), the TIM's queueing discipline uses only two queues. Thus, much of the packet handling can be done in hardware rather than software, allowing trunk speeds as high as 2 Mb/s.

Because it gives short interactive messages priority over packets of larger messages, the TIM queueing discipline favors short messages even further. Consequently, from the law of conservation of unfinished work, the larger message packets necessarily have longer queueing delays than with the other disciplines mentioned.

Overview of TIM and Analysis. The TIM's queueing discipline is a packet nonpreemptive (PNP) two-priority discipline in which the length of a packet determines the priority. Another algorithm ensures that packets from the same message retain their order. The service quantum of the TIM discipline is a packet, not a message as in the usual two-priority queueing discipline.

This design has the effect of limiting the queueing delays that long messages cause for short interactive messages. Therefore, the TIM discipline gives less delay for short messages than a two-priority queueing discipline that has messages as its service units. However, the TIM's unique design also requires that we derive new delay formulas for analyzing its performance. The derivations would be difficult because the packet arrivals, in general, are not Poisson processes.

While message arrivals can be modeled by Poisson arrivals,[3,5] packet arrivals generally cannot be modeled as such, so the average packet delays are difficult to calculate directly. The essence of a Poisson process is that the arrivals of large numbers of elements are independent.[6,7] But packets from the same message are not independent, because bits within a message were generated *continuously*.

Mathematically, a one-server system with Poisson arrivals of messages is M/G/1,[3,5] but the corresponding system—in terms of packets—is G/G/1. If packet arrivals could be modeled as Poisson processes, the problem of TIM performance analysis would have been easier because existing average-delay formulas could be used. Unfortunately, it is impossible to justify such a model because it requires independence of packets, even from the same message.

Approach of Analysis and Results. The standard assumptions for a message M/G/1 priority system are:

- The incoming data traffic is a superposition of independent Poisson processes.

- All incoming messages from the same Poisson process belong to the same priority class.[5,8]

For the TIM discipline, however, the second assumption's validity is obscure, because TIM message priorities are classified according to a message's length: Is the length less than a certain number of bytes?

To use the standard assumptions, we must be able to show that an arbitrary Poisson process can be decomposed into Poisson processes, each belonging to only one priority class. However, the processes decomposed from a general Poisson process may not be Poisson. Fortunately, as Appendix A shows, we are able to do the Poisson decompositions for this case.

Before we derive the delay formulas, we first analyze the problem. We compare it with related phenomena to grasp the physical meaning of the related mathematical expressions and, then, infer what the characteristics of the delay formulas should be. Further, we study related queueing theory to find an appropriate approach to the derivation.

For an M/G/1 queue, the average-packet-delay formulas are in closed form. As we will show, the effect of different message-length distributions is clear. Furthermore, the moments of the low-priority traffic do not influence the average delay of the high-priority traffic. In our derivations, we use the G/G/1 queue conservation law[9] and physical understanding to overcome the difficulties associated with the G/G/1 nature of the packet arrivals. This shows that the G/G/1 conservation law deserves more attention.

The average-message-delay formulas are exactly in closed form and provide a better physical picture of the queueing process. This picture shows that the TIM discipline favors middle-sized time-sharing messages over long file-transfer messages. We compare the average-message-delay formulas with delay formulas for a similar, but *synchronized* and *preemptive* packet-priority queueing discipline that Hluchyj and others analyzed.[5] We conclude that the TIM discipline gives shorter delay not only for low-priority messages but also for high-priority messages. Interestingly, the corresponding terms from these two sets of formulas have exactly the same form, while the noncorresponding terms reflect the different nature of these queueing disciplines. Moreover, as the packet size goes to zero, both sets of delay formulas—as expected—give the same delays at the limit.

Based on our analysis and resulting formulas, the TIM discipline results in the shortest average delay for short messages (at reasonable expense to the longer messages) among all known practical queueing disciplines. The limited influence of low-priority traffic on the average delay of high-priority traffic can be controlled by design.

Finally, to show the application of these delay formulas, we calculate the parameters for Fraser and Morgan's data traffic model.[3]

Remarks. This derivation differs substantially from the other traditional derivations of queueing delay formulas. Analysis is not separate from the formula derivation; instead, it is an indispensable part. While such a method of derivation is common in theoretical physics, it may not be viewed as strictly rigorous in terms of pure mathematics. For this complicated problem, the author believes that this derivation method is better because, at each stage, it also makes the physical meanings of the formulas apparent.

Queueing Model and Conservation Laws

Here, we summarize the important points of the M/G/1 data traffic model and the queueing disciplines, and identify steady-state statistics that are of interest. The conservation laws, which are crucial to the derivation of delay formulas for our queueing model, are discussed for average packet delays. The notation established here is used later in the performance analysis.

Data Traffic Model M/G/1 Queue. Assume that all messages consist of a finite number of packets, each, at most, of length L. Moreover, only the last packet of a message can possibly have a length smaller than L. Also assume that there are n independent Poisson arrival processes, each consisting of messages from only one priority class.

Let the arrival rates and the message-length (service time) distributions be, respectively,

$$\lambda_i \text{ and } b_i(x); \quad i = 1, 2, \ldots, n \tag{1a}$$

for n independent message Poisson arrival processes.

Let the first moment and second moment of $b_i(x)$ be, respectively,

$$\bar{x}_i \text{ and } \overline{x_i^2}; \quad i = 1, 2, \ldots, n \tag{1b}$$

Also, let

$$\rho_i = \lambda_i \bar{x}_i; \quad i = 1, 2, \ldots, n \tag{1c}$$

Thus, ρ_i is the traffic load from the ith Poisson arrival process. Let us define $\overline{W}_{oi}$ as the average remaining service time for a message from the ith process found by a new arrival from the ith process. According to reference 8 (Vol. II, page 109),

$$\overline{W}_{oi} = \frac{1}{2}\rho_i \frac{\overline{x_i^2}}{\bar{x}_i}; \quad i = 1, 2, \ldots, n \tag{1d}$$

Let the averages of $\bar{x}_i$ and $\overline{x_i^2}$ be, respectively

$$\bar{x} \text{ and } \overline{x^2} \tag{1e}$$

Let the sum of all λ_i and ρ_i be, respectively, the total arrival rate λ and the total load ρ. Thus,

$$\rho = \lambda \bar{x} \tag{1f}$$

Let $\overline{W}_o$ be the sum of all $\overline{W}_{oi}$. Thus,

$$\overline{W}_o = \frac{1}{2}\rho \frac{\overline{x^2}}{\bar{x}} \tag{1g}$$

Also, $\overline{W}_o$ is the average remaining service time for a message found by a new arrival.

From equations (1f) and (1g), both parameters ρ and $\overline{W}_o$ are additive.

Let the messages be classified into two distinct classes, say the high-priority class h, and the low-priority class l. For instance, let messages from the first m ($< n$) processes belong to class h and the rest belong to class l. To denote the averages within each class and the sums from each class, we use the notation

$$\lambda_h, \bar{x}_h, \overline{x_h^2}, \text{ and } \overline{W}_{oh}$$

$$\lambda_l, \bar{x}_l, \overline{x_l^2}\ \rho_l, \text{ and } \overline{W}_{ol} \tag{1h}$$

Then, for the usual message nonpreemptive two-priority (MNP) queueing discipline, the average in-queue waiting time for the high- and low-priority messages are, respectively,[8]

$$\overline{W}_h = \frac{\overline{W}_o}{1 - \rho_h} \tag{2a}$$

$$\overline{W}_l = \frac{\overline{W}_h}{1 - \rho} \tag{2b}$$

As we shall see next, the average delay formulas for the TIM discipline are closely related to formula (2), although additional parameters are needed.

PNP Discipline and the TIM. In the TIM, we implement a new two-priority queueing discipline that is nonpreemptive in terms of packets. Data that leaves for the trunk is placed either in the high-priority queue (HPQ)—if the incoming packet contains less than 16 data bytes—or in the low-priority queue (LPQ). Because the last packet of a long message can contain less than 16 bytes, an algorithm maintains the order of packets from the same message[4] (Figure 1). In practice, this means that packets from the same message have the same priority.

The trunk empties the HPQ before serving the LPQ, but both queues are scheduled according to the first-

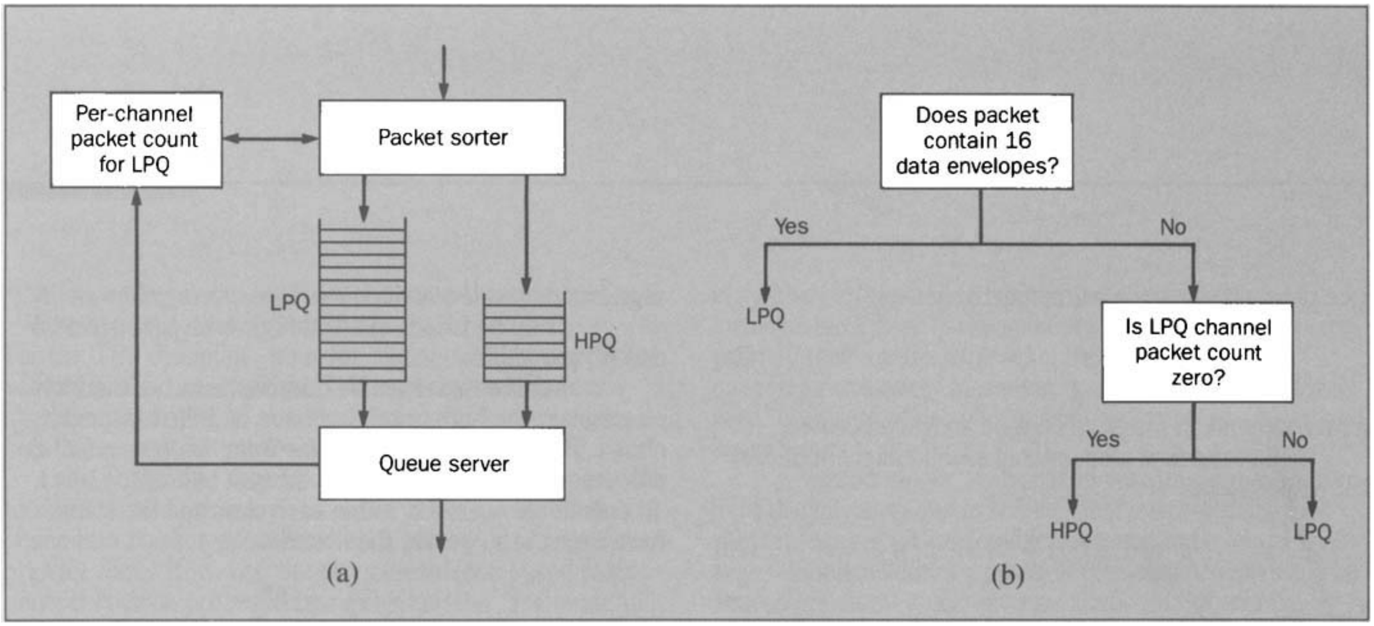


Figure 1. An algorithm maintains the order of packets from the same message. (a) Trunk interface module (TIM) discipline; (b) packet sorting process. HPQ = high-priority queue; LPQ = low-priority queue.

in, first-out discipline. The service units are packets that may have variable lengths, from three to 18 bytes (two address bytes overhead). In other words, an incoming high-priority message can interrupt the service to a low-priority message after any units of packets from the low-priority message have been serviced. Thus, in terms of messages, the TIM queueing discipline is neither non-preemptive nor preemptive, because a low-priority message may not be served as a whole without interruption.

In short, the TIM discipline is a PNP queueing discipline whose service unit is a packet and the length of a message determines its priority. Both the PNP discipline and the usual message two-priority discipline satisfy the packet G/G/1 conservation law.[9] See Appendix B.

However, the TIM discipline might not be compatible with the M/G/1 traffic model, which requires that the message-length distributions be zero either before or after a fixed length L (a design specification). In other words, we must be able to express an arbitrary Poisson arrival process in terms of Poisson arrival processes that have a special property; otherwise, such a requirement might be too restrictive to serve as a general assumption. However, it was not clear that such decompositions are possible.

We would like to emphasize that the compatibility is verified. As we show in Appendix A, a Poisson arrival process with an arbitrary message-length distribution can be decomposed into two Poisson arrivals with message-length distributions that satisfy the above requirements because of the TIM discipline.

Packet Delay and G/G/1 Law. It is difficult to calculate the average packet delays directly, because packet arrivals are generally not a Poisson process. Fortunately, we have found an indirect method.

It turns out that the G/G/1 queueing conservation law[9] is a powerful tool for calculating the average delays of packets. The advantage of using a conservation law is that we can avoid the complicated direct calculations to obtain the average delay formulas. But first, we should understand the conservation law.

M/G/1 and G/G/1 laws. Schrage's G/G/1 queueing conservation law (Appendix B) is an extension of Kleinrock's M/G/1 queueing conservation law.[10] Both laws can be written in exactly the same form:

$$\sum_i \rho_i \overline{W}_i = \overline{U} - \overline{W}_o \quad (3)$$

where ρ_i and $\overline{W}_i$ are, respectively, the load and the average waiting time of the ith class of customers; $\overline{U}$ is the average of unfinished work; and, as before, $\overline{W}_o$ is the average remaining service time for the customer found by a new arrival.

Both conservation laws require nonpreemptive discipline (i.e., no disruptions of service and conservation of unfinished work). While the M/G/1 law requires the Poisson arrival assumption, the G/G/1 law requires neither Poisson arrival nor independence. It only requires that equilibrium distributions exist! This is about as general as anyone can expect. Therefore, the G/G/1 law should be able to replace the M/G/1 law in any usage.

Although well-known, the G/G/1 law is not used much, while the M/G/1 law is cited almost everywhere that a conservation law is used. Besides the M/G/1 law's having been discovered much earlier, there are few known cases where only the G/G/1 law can be used. Apparently, it is very difficult to do mathematical modeling analysis without the famous Poisson-arrival assumption.

But for calculating the average packet delays of message M/G/1, the G/G/1 conservation law is almost a perfect tool, as we shall show for nonpreemptive disciplines. Furthermore, the G/G/1 law is far more useful and deserves due attention in our derivation of the average message delay formulas for the preemptive TIM discipline.

The conservation law does not prevent reducing the overall average delay. Instead, it provides insight on how to reduce the overall average waiting time. If a queueing discipline favors short messages (packets), then the resulting overall-average-waiting time has to be shorter than for the FCFS discipline.

Thus, an implicit necessary assumption of the conservation law is: Within each class, messages (packets) are served independent of their service time. Also, the assumption that the queueing discipline is nonpreemptive is only a *sufficient* condition for the M/M/1 queue. The *necessary* assumption is that it be a preemptive resume, work-conserving discipline, as Coffman and Mitrani's work shows.[11]

Averages of packet and message delays. To understand the relationship between the averages of packet delays and message delays, we first establish such a relationship for cases where the queueing discipline is message nonpreemptive. Then, the average message delays satisfy the M/G/1 conservation laws and, if we consider some average packet delays instead of the related average message delays, also satisfy G/G/1 conservation laws.

To establish the relationship between an average message delay and its related average packet delay, consider the jth average packet delay and the other average message delays. It follows that the G/G/1 conservation law implies

$$\begin{aligned}\sum_i \rho_i \overline{W}_i - \rho_j \overline{W}_j + \rho_j[P]\,\overline{W}_j[P] \\ = \overline{U} - \overline{W}_o + \overline{W}_{oj} - \overline{W}_{oj}[P]\end{aligned} \tag{4}$$

where $\overline{W}_j[P]$ is the average packet delay of the jth class; $\overline{W}_{oj}[P]$ is the average remaining service time for the jth-class packet, found by a newly arrived jth-class packet; and $\rho_j[P]$ is the traffic load for the jth-class packet.

In this form, the first term on the left-hand side of formula (4) and the first two terms on the right-hand side are readily identified as terms in the M/G/1 conservation law for average message delays. As Schrage pointed out,[9] the expectation value of unfinished work is determined strictly by the arrival process and is independent of discipline.

The quantity $\rho_j[P]$ should be equal to ρ_j because a physical quantity—such as traffic load—is the same, whether defined from the packet or message viewpoint. For convenience, we write $\overline{W}_{oj}$ and $\overline{W}_{oj}[P]$ in these forms:

$$\overline{W}_{oj} = \frac{1}{2}\rho_j B_j \tag{5a}$$

$$\overline{W}_{oj}[P] = \frac{1}{2}\rho_j Z_j \tag{5b}$$

Here is the motivation for formula (5b). While $\overline{W}_{oj}[P]$ might not have exactly the same form as in (1d), Z_j is independent of the packet traffic loads and is a function of the jth packet-length distribution. Some simple algebraic calculation leads to the relation

$$\overline{W}_j[P] - \overline{W}_j = \frac{1}{2}(B_j - Z_j) > 0 \qquad (6a)$$

The left-hand side is the difference between the average delays of packets and of messages. Because messages are not interrupted, this difference should be independent of traffic loads. Therefore, Z_j is independent of traffic loads and can only be a function of the jth packet-length distribution; see Appendix B. If all packets from the jth class have the same length L, then we have as expected (and verified in Appendix B)

$$Z_j = L \qquad (6b)$$

The meaning of relation (5b) is now clear. The load ρ_j is the probability that a newly arrived packet will see a packet is already being served. Z_j is the length of the new packet, if all the packets have the same length. The factor 1/2 accounts for the randomness of packet arrivals. If most packets have a length L, then Z is very close to L. Notice that the presence of the parameter Z_j is characteristic of the average-packet-delay formulas because of the G/G/1 conservation law.

Performance Analysis

Now, we analyze the PNP discipline and consider the TIM discipline as a special case.

To simplify our notations, we use the relations

$$\overline{W}_{oh} = \frac{1}{2}\rho_h A, \; A = \overline{x_h^2}/\overline{x}_h \qquad (7a)$$

$$\overline{W}_{oh} = \frac{1}{2}\rho_l B, \; B = \overline{x_l^2}/\overline{x}_l \qquad (7b)$$

$$\overline{W}_o[P] = \frac{1}{2}\rho_l Z \qquad (7c)$$

$$\overline{W}_o = \overline{W}_{oh} + \overline{W}_{ol} \qquad (7d)$$

Z is very close to L because the low-priority messages, on the average, are long and only a message's last packet can possibly be less than full.

Let $\overline{W}_h$ and $\overline{W}_l$ be, respectively, the average waiting time for high- and low-priority messages; and let $\overline{W}[P]$ be the average waiting time for low-priority packets. For the TIM, $\overline{W}_h$ is also the average waiting time for high-priority packets, because the lengths of high-priority messages are equivalent to the packet lengths.

With the above notation, for example, the average message delay for the FCFS discipline[8] is

$$\overline{W}_F = \frac{1}{2}\frac{A\rho_h + B\rho_l}{1 - \rho} \qquad (8)$$

For the usual message two-priority discipline, we have [see formulas (2a) and (2b)]

$$\overline{W}_h = \frac{1}{2}\frac{A\rho_h + B\rho_l}{1 - \rho_h} \qquad (9a)$$

$$\overline{W}_l = \frac{\overline{W}_h}{1 - \rho} \qquad (9b)$$

Derivations of Average-Packet-Delay Formulas. The PNP discipline has the unique characteristic of limiting how low-priority messages affect the queueing delays of high-priority messages. Therefore, this queueing discipline results in less delay for high-priority messages than a message nonpreemptive two-priority (MNP) discipline. When compared with the MNP discipline, this characteristic gives longer delays for low-priority messages whose transmission may be disrupted.

But if the low-priority messages are, at most, one full packet in length, the delays for the PNP and MNP disciplines are practically the same. This gives us some physical insight into the PNP discipline for special cases, because the MNP discipline's delay formulas are readily available in any textbook on queueing theory.

Delay for high-priority messages. Through mathematical analysis and physical insights, we shall establish the average high-priority message waiting time. The average high-priority packet and message delays are related by formula (6a); see Appendix B. Therefore, it is enough to derive the average high-priority message delay formula. Because of the two-priority queueing discipline's nature, the high-priority message queueing delay is simple.

Suppose that there are only high-priority messages. Then, the PNP discipline is equivalent to MNP without low-priority traffic. From formula (9a), because $\rho_l = 0$, the average high-priority message (packet) delay for this special case is

$$\overline{W}(h) = \frac{1}{2}\frac{A\rho_h}{1 - \rho_h} \qquad (10a)$$

When low-priority message traffic is also present, we generally have

$$\overline{W}_h < \overline{W}(h) + \frac{L}{2} \qquad (10b)$$

because the PNP discipline is nonpreemptive in terms of packets. (That is, a low-priority packet must continue to go out on the trunk, even if a high-priority packet arrives during its transmission.) The number L is the full packet length, and the factor 1/2 accounts for the randomness of the arrival of high-priority messages.

Relation (10b) gives a *bound* to the influence of low-priority messages. If the queueing discipline were nonpreemptive in terms of messages, such a bound would not be possible according to formula (9a). In formula (9a), B increases as the second moment of the low-priority messages increases. In short, $\overline{W}_h$ is largely *independent* of the low-priority message traffic, if the high-priority message buffer is large enough to prevent overflow. (The overflowed high-priority messages will go into the low-priority buffer in the TIM.)

To obtain the exact expression for $\overline{W}_h$, first we consider the special case where low-priority messages are one full packet long. (Here, $B = L$, and the PNP queueing discipline is equivalent to MNP.) According to formula (9a),

$$\overline{W}_h = \overline{W}(h) + \frac{L}{2}\frac{\rho_l}{1 - \rho_h} \qquad (10c)$$

We interpret the factor $\rho_l / (1 - \rho_h)$ as the probability that a low-priority packet will be injected into a gap between high-priority message trains. Because the sources of the low- and high-priority messages are independent, this probability should not depend on the length distributions of low-priority messages. Therefore, formula (10c) is valid as long as the message lengths are integer multiples of full packets.

From formula (10c), notice that the parameter L—a design specification—controls how low-priority messages influence the queueing delay of high-priority messages. Also, if the packet length L is smaller, then the delay would be smaller. In general, the delay should be smaller because not all the packets are of full length L. However, the value of the average delay should be close, because only the last packet of a message can possibly be partially full and low-priority messages are generally long, consisting of many packets.

Formula (10c) is not only a good approximation for the general case, but also is probably the only *practical* formula because, in practice, we don't really know the packet-length distributions. However, for completeness in theory, we shall write the formula for the general case as follows:

$$\overline{W}_h = \frac{1}{2}\frac{A\rho_h + Z\rho_l}{1 - \rho_h} \qquad (11)$$

In formula (11), Z has replaced L in formula (10c). This replacement has little practical meaning because we don't know Z's value, except that it is close to L. Theoretically, this is because of formulas (7c) and (6b).

Note that the parameter Z in formula (11) has replaced the parameter B in formula (9a). This means that the moments of the low-priority traffic have no influence on the average delay of the high-priority traffic. In practice, this will result in a small average delay for high-

Table I. Data Traffic Model for Analysis of Datakit VCS

Traffic type	Distribution	Average length (characters)	Arrival rate
1	Deterministic	1	λ
2	Exponential	40	$\lambda/10$
3	Exponential	512	$\lambda/100$

Table II. Decomposition into Poisson Processes

	Arrival rate	
Traffic type	HP	LP
1	λ	0
2	$0.0313\ \lambda$	$0.0687\ \lambda$
3	$0.00029\ \lambda$	$0.00971\ \lambda$

NOTE: HP = high priority; LP = low priority.

priority messages in spite of a heavy load of low-priority long messages as formula (10b) suggests. The limited influence of the low-priority traffic can be controlled by the design specification L—as indicated in formula (10c)—which further supports the ISN design of using a 16-byte length for short packets.

Thus, formula (11) is valid for any PNP discipline, of which the TIM discipline is a special case.

Delay for low-priority messages. Now, we shall use the G/G/1 conservation law to derive the average-waiting-time formula for a low-priority packet. (The PNP discipline is nonpreemptive only for packets, and the packet arrivals are not a Poisson process.)

For this case, the G/G/1 conservation law takes the form

$$\rho_h \overline{W}_h + \rho_l \overline{W}[P] = \overline{U} - \{\overline{W}_o[P] + \overline{W}_{oh}\} \qquad (12)$$

where $\overline{U}$ is the average of unfinished work, $\overline{W}[P]$ is the average low-priority packet waiting time, and the last term is the average-remaining-service time for all high-priority messages and low-priority packets. Because the PNP queueing discipline conserves work, we can express $\overline{U}$ as the average delay of the FCFS discipline.[8] In other words,

$$\overline{U} = \overline{W}_F \qquad (13)$$

Using formulas (7), (8), and (11) and some algebraic calculations, equation (12) gives the expression

$$\overline{W}[P] = \frac{\overline{W}_h}{1-\rho} + \frac{1}{2}\frac{B-Z}{1-\rho} \qquad (14a)$$

Thus, we complete our derivations for packet delay formulas of PNP.

Packet delay formula parameters. In formula (11), the value L provides the upper bound for delay, which can only be slightly larger for any real values of Z. The delay $\overline{W}[P]$ is not sensitive to a slight change in the value of Z, as formula (14b) will show. In the practical usage of formula (14), B's accuracy is the dominating factor.

The second term in (14a) makes up the difference between the assumptions of packet Poisson arrivals and message Poisson arrivals. For a fixed load, the first term can be bounded without traffic characteristic studies. We, therefore, conclude that a packet Poisson arrival assumption (which requires that packets be independent even from the same message) is unrealistic because longer messages are always present.

Contrarily, the second term cannot be bounded without data traffic studies, because the parameter B in formula (14a) depends on the second moment of the message-length distributions of low-priority messages. Therefore, message Poisson arrivals include more realistic modeling.

Average-Message-Delay Formulas. While average packet delays are useful for data network design, the average-message-delay formula should be more useful for

Table III. Moments of Effective Message Length

Parameter	Value	Parameter	Value
$\bar{x}_h$	3.1847	$\overline{[x]}_l$	118.5597
A	3.7029	B	701.2319
$\bar{x}_l$	128.2696	ρ_h/ρ	0.24614

NOTE: The parameter Z takes the larger approximate value 18.
The time length unit is *byte/S*, where S is the speed of the trunk.

understanding the real meaning of the average packet delays. What our customers see are message delays. (We define a message delay as the difference between the time that any part of a message remains in the queueing system and the message's processing time.)

Delay for low-priority messages. To derive the average-message-delay formula for low-priority messages, we write (14a) in an alternative form,

$$\overline{W}[P] = \frac{1}{2}\frac{A\rho_h + B\rho_l}{(1-\rho_h)(1-\rho)} + \frac{1}{2}(B - Z) + \frac{1}{2}(B - Z)\frac{\rho_h}{1-\rho_h} \qquad (14b)$$

In formula (14b), the sum of the first two terms is the average low-priority packet delay of the MNP discipline. The second term is the difference between the average packet delay and the average message delay for a nonpreemptive discipline. While both the second and last terms have a common factor $(B - Z)$, they have very different physical meanings. The last term accounts for disruptions in the service of low-priority messages to serve high-priority messages. This term is insignificant for light high-priority traffic—a small price to bound the influence of ρ_l on $\overline{W}_h$.

Because the PNP discipline is preemptive in terms of messages, the derivation of the average message delay is not as straightforward as for a nonpreemptive discipline. We must first understand the delay mechanism.

Physically, the average delay for the first packet of a low-priority message is the same as the average message delay of the MNP discipline. Analysis shows that the average message delay is given by

$$\overline{W}_l = \frac{1}{2}\frac{A\rho_h + B\rho_l}{(1-\rho_h)(1-\rho)} + \overline{[x]}\frac{\rho_h}{1-\rho_h} \qquad (15)$$

where x is the length of a low-priority message; $[x]$ equals x minus the length of the message's last packet; and $\overline{[x]}$ is the average of $[x]$ over the message-length distribution of the low-priority messages. The function $[x]$ would probably explain, in part, the difficulty of a direct, strictly mathematical derivation of the average-message-delay formulas. It is interesting to note that the second term in formula (14b) does not enter formula (15). Furthermore, for a low-priority message of length x, the average waiting time is

$$\overline{W}_l(x) = \frac{1}{2}\frac{A\rho_h + B\rho_l}{(1-\rho_h)(1-\rho)} + [x]\frac{\rho_h}{1-\rho_h} \qquad (16)$$

Formula (16) shows that the PNP discipline does favor middle-size time-sharing messages over long file-transfer messages. When the low-priority messages are M/D/1 processes and all the packets are full, the disruption term in the average-packet-delay formula (14b) is, as expected, one-half the corresponding term in the average-message-delay formula (16). The lengthy analysis that leads to formulas (15) and (16) is given in Appendix B.

By comparing formulas (15) and (14b), we can see

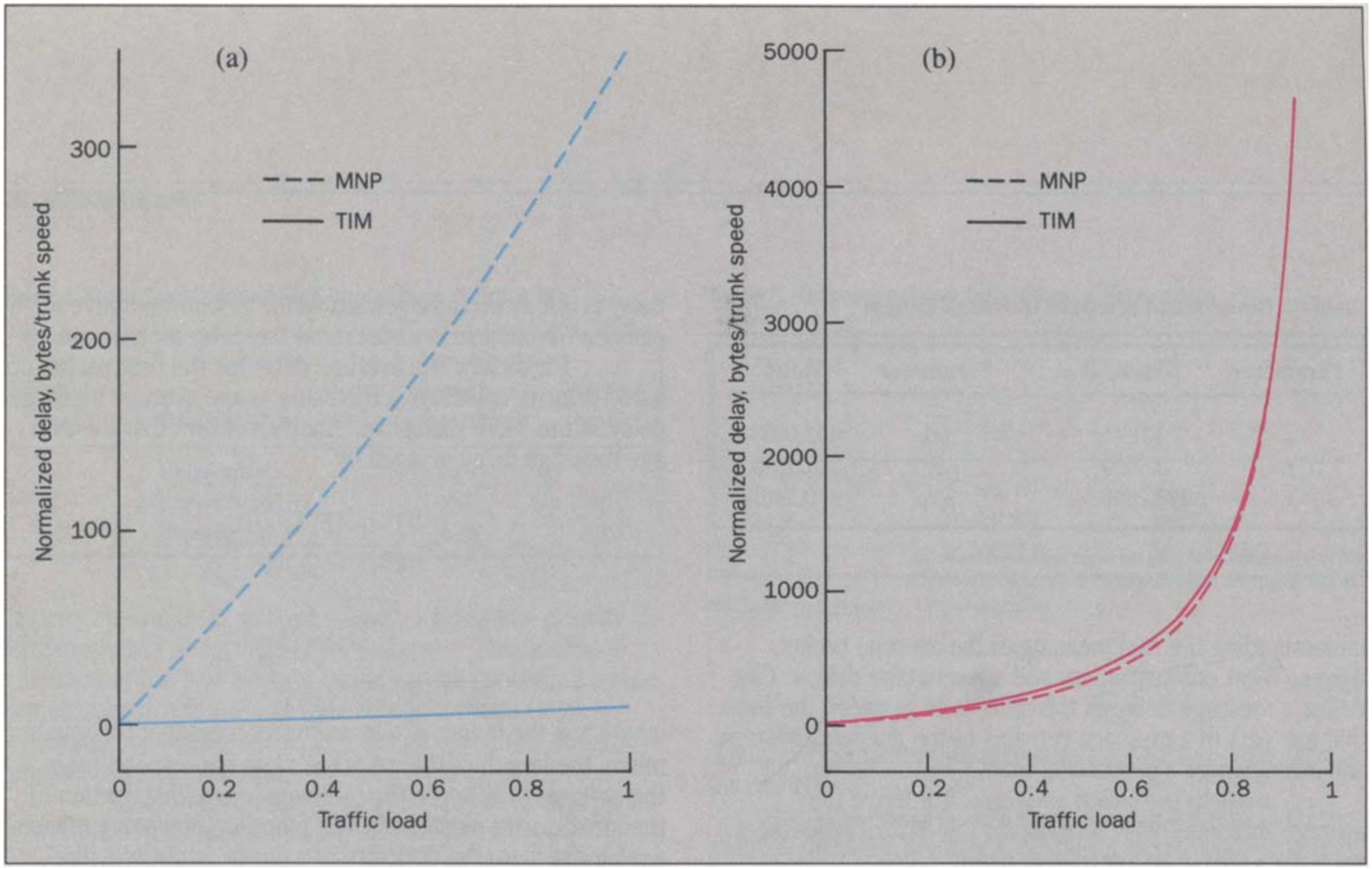


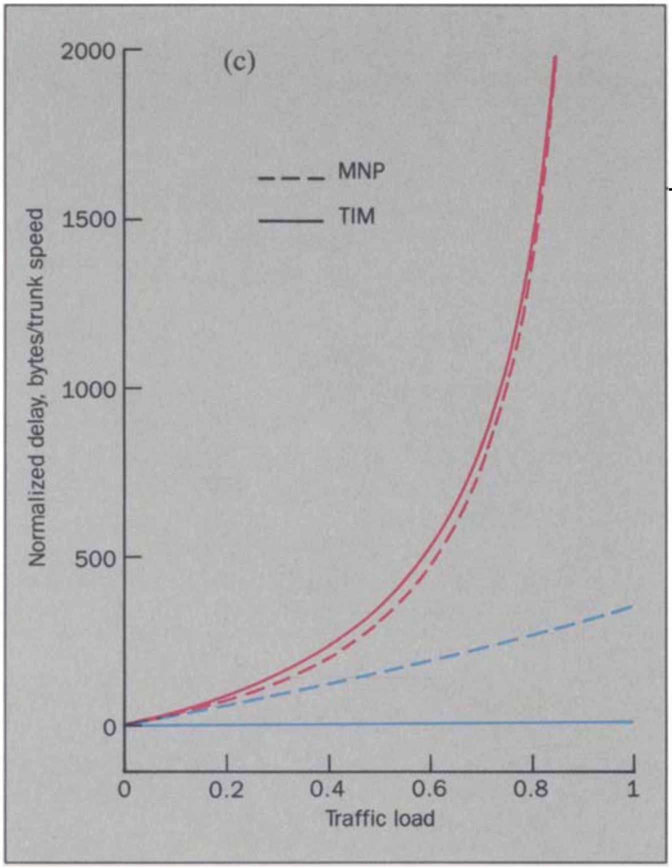
Figure 2. Normalized packet delay for the message nonpreemptive (MNP) two-priority queueing discipline, and the TIM, a packet nonpreemptive (PNP) two-priority discipline. (a) High-priority messages (blue); (b) low-priority messages (red); and (c) both high- and low-priority messages (blue and red, respectively). The average in-queue waiting time for high-priority messages with PNP is less than 2.8 percent.

that the average packet delay and average message delay are close if the high-priority traffic load is not heavy. From formula (16), we can see that longer messages have, on the average, longer delay. But such delay differences—caused by the length of the messages—could be considered, except for very long messages, as second-order effects under light, high-priority traffic conditions.

Average message delay and M/G/1 law. If the low-priority messages are M/M/1 processes, then the corresponding disruption terms have the same limit as L goes to zero. Then, the relationship between average packet delay and average message delay is the same as for a message nonpreemptive discipline [see formula (6)]. As for the round-robin discipline, the M/M/1 processes have a way to get around the message preemptiveness of work-conserving disciplines after the average is taken.[8]

We have shown that formula (6a) is valid because average message delays satisfy the M/G/1 conservation law and average packet delays satisfy the G/G/1 conservation law. Conversely, the validity of formula (6a) implies the validity of the M/G/1 conservation law. In other words, the above analysis shows that, if the low-priority messages are M/M/1 processes, then the M/G/1 conservation law is satisfied, at least for the limit as the full-packet length L goes to zero.

As Coffman and Mitrani[11] show, this relationship is exact for any value of L. For preemptive resume, work-conserving disciplines, they prove that the M/G/1 conservation law is satisfied if the job streams are exponential. One might argue that the conditions of their work are not exactly satisfied in our analysis because we do not assume that the high-priority message streams are exponential.



But in the M/G/1 conservation law, the exponential condition neutralizes the effects of message transmission interruptions and, in the PNP discipline, high-priority messages are not interrupted.

If the low-priority message stream is exponential, then the parameter Z can be calculated accurately because of the M/G/1 conservation law. Because Z is not greater than L, the M/G/1 conservation law provides a check for the correctness of the average delay formulas. However, in practice, we cannot use the M/G/1 conservation law to calculate the value of the parameter Z because the TIM discipline results in a lower-end cutoff at L, for each message-length distribution of low-priority messages.

Comparison with an SPP discipline. Hluchyj, Tsao, and Boorstyn analyzed the synchronized preemptive priority (SPP) queueing discipline,[5] where the service unit is also a packet. This discipline's requirements are: *synchronization*, and all packets are the *same* length.

Intuitively, one can expect that a nonpreemptive discipline would give less delay for low-priority messages than a similar preemptive one. However, one could guess that the PNP discipline would probably give longer delay for high-priority messages. But because of synchronization, this is not necessarily true. Our analysis shows that the PNP discipline gives shorter delays even for high-priority messages.

For easier comparison, we first convert the formulas in reference 5 for the two-priority case into our notation. The relevant formulas for comparison are

$$\overline{W}_{11} = \frac{L}{2} + \frac{1}{2}\frac{A\rho_h}{1-\rho_h} \tag{17a}$$

and

$$\overline{W}_{2j} = \frac{1}{2}\frac{L}{1-\rho_h} + \frac{1}{2}\frac{A\rho_h + B\rho_l}{(1-\rho_h)(1-\rho)} + (j-1)\frac{L}{1-\rho_h} \tag{17b}$$

where $\overline{W}_{11}$ is the average waiting time for the first packet of the high-priority messages, and $\overline{W}_{2j}$ is the average waiting time for the jth packet of a low-priority message. Because the transmission of a high-priority message is not interrupted, $\overline{W}_{11}$ is the average waiting time of a high-priority message. Notice that the first term of formula (17a) results from synchronization.

The average waiting time of a low-priority message of length x ($= jL$) needs some calculation. From the average waiting time of the jth packet, we should subtract the service time of the ($j-1$) packets. In other words, the average waiting time of this low-priority message is $\overline{W}_{2j} - (j-1)L$ and, therefore,

$$\overline{W'}_h = \overline{W}_{11} \tag{18a}$$

and

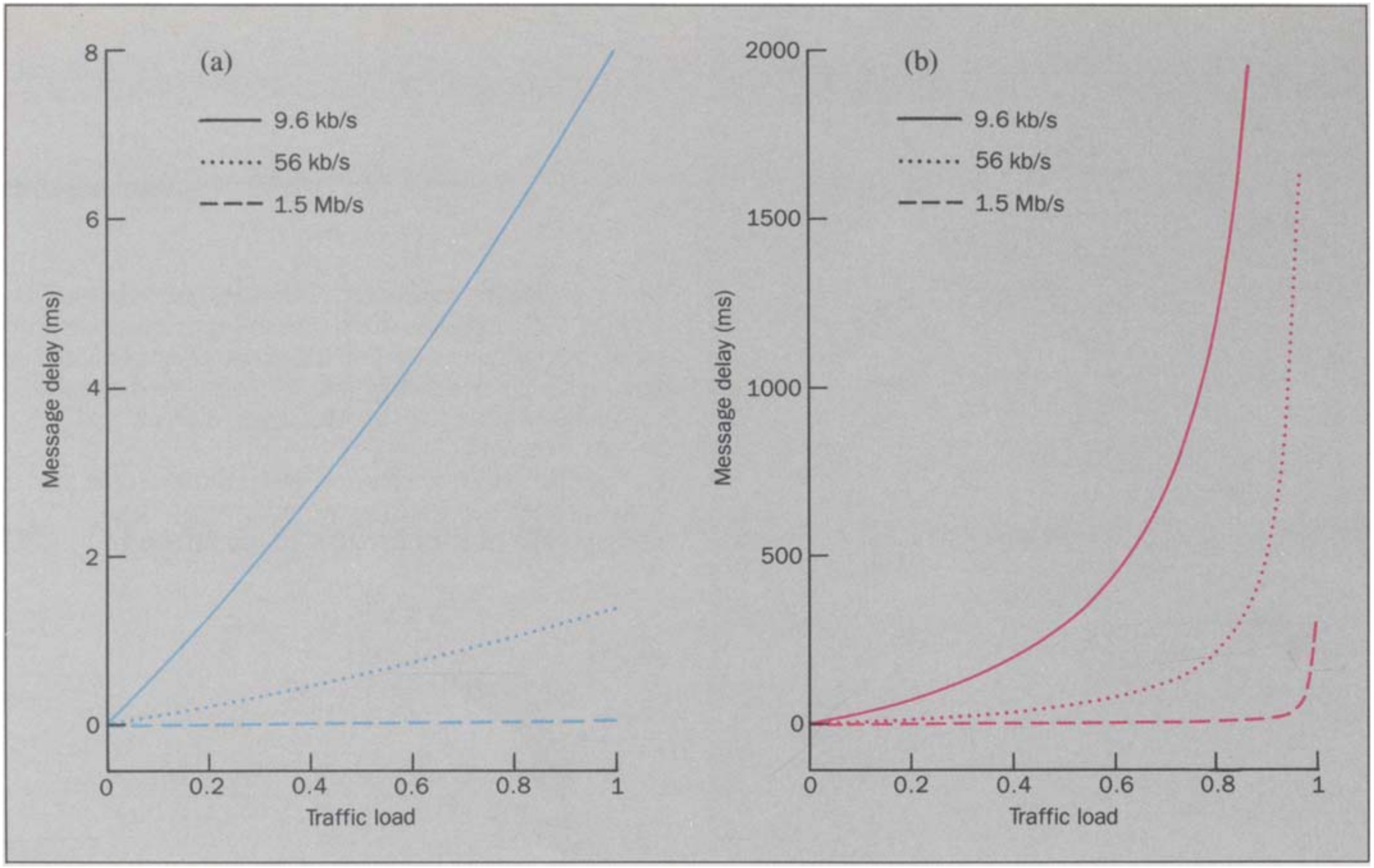


Figure 3. Average message in-queue waiting time at trunk speeds of 9.4 kb/s, 56 kb/s, and 1.544 Mb/s for the TIM discipline. (a) High-priority messages (blue); (b) low-priority messages (red); and (c) both high- and low-priority messages (blue and red, respectively).

$$\overline{W}'_l(x) = \frac{1}{2}\frac{L}{1-\rho_h} + \frac{1}{2}\frac{A\rho_h + B\rho_l}{(1-\rho_h)(1-\rho)} + [x]\frac{\rho_h}{1-\rho_h} \qquad (18b)$$

After comparing formula (18) with (11) and (16), we conclude that the PNP discipline gives smaller delays even for high-priority messages.

Moreover, in these two sets of formulas, the corresponding terms are exactly the same. While this fact is physically meaningful, it is far from obvious in formulas (14b) and (17b), because there the terms that accounted for the disruption effects are so different. These terms, which have no corresponding terms in the other set of formulas, reflect the different nature of the queueing disciplines. Also, as expected, these two sets of delay formulas give the same delays at the limit as L goes to zero.

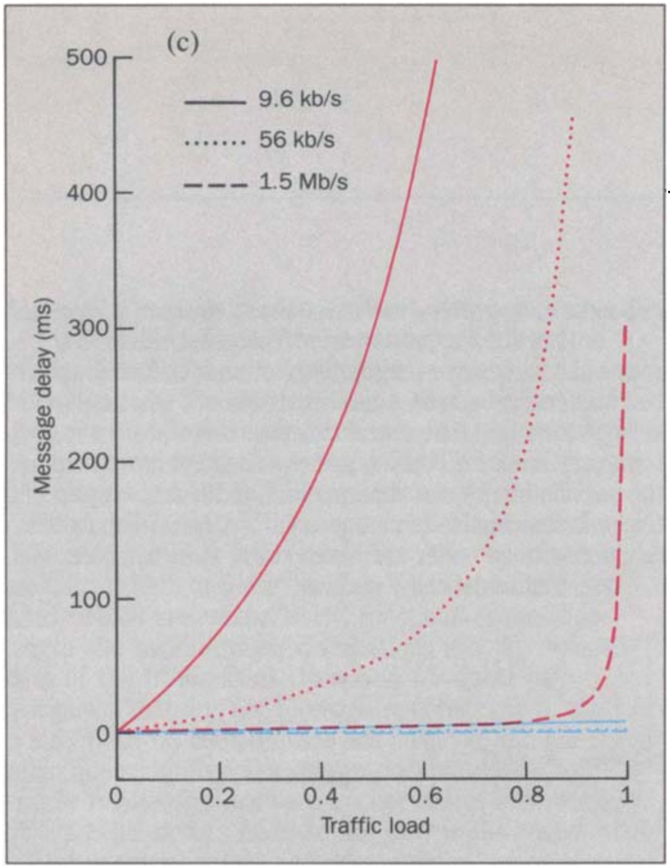
An Example

This example of delay formulas for the TIM uses the established data-traffic model, Table I, that Fraser and Morgan[3] used to analyze Datakit VCS. The traffic types in the table represent

- Type 1—single-character, terminal-to-host transmissions from asynchronous terminals
- Type 2—time-sharing, host-to-terminal responses
- Type 3—host-to-host file transfers.

First, we decompose each Poisson process into a high- and a low-priority Poisson process using the method described in Appendix A. Table II presents the results. Then for each decomposed Poisson process, we calculate the first and second moments of the effective message length, which includes the packet overhead. Table III presents these parameters.

While packet overhead for low-priority messages



increases the low-priority traffic load less than 13 percent, it almost triples the high-priority traffic load. The real load capacity is reduced from one to 0.757. However, we cannot avoid such an overhead for any statistical multiplexer because address bytes must be added for correct delivery. With greater full-packet length, the capacity will increase slowly because of short messages. For example, if the full-packet length were 64, the capacity would be 0.807.[3] The type-3 traffic has a strong influence on the parameter B, which is the dominant factor for average low-priority delays. If this were the MNP discipline, the high-priority messages would have had 68 times longer average delays. But because of the TIM discipline, the average queueing delay for high-priority messages is generally small.

To compare the PNP with MNP, as illustrated in Figure 2, the average in-queue waiting time for high-priority messages is less than 2.8 percent. The average in-queue waiting time for low-priority messages adds another 4 percent for 60-percent trunk utilization, but only 2 percent for 80-percent utilization.

For the same traffic characteristics and traffic load, the average delays are inversely proportional to the trunk speeds. Thus, in terms of utilization, higher speed trunks have decided advantages over slower trunks (i.e., for the same data-traffic characteristics, the higher speed trunk should have higher utilization). In formula (15), we implicitly assumed that the access speeds of messages are at least as fast as the trunk speed. Owing to the effect of the slow access-line speeds,[3,12] the advantage of a higher speed trunk is even larger. To illustrate this, Figure 3 shows the average message in-queue waiting time for trunk speeds of 9.6 kb/s, 56 kb/s, and 1.544 Mb/s.

Conclusion and Discussion

Based on the above analysis we conclude that, among all the known practical queueing disciplines, the TIM discipline results in the shortest average delay for short messages (at reasonable expense to long messages). This is critical to the performance of long-haul computer-communications networks because short messages are more delay sensitive. Because of the TIM discipline, the moments of the low-priority traffic have no influence on the average delay of the high-priority traffic. The TIM discipline, to some extent, favors middle-size time-sharing messages over long file-transfer messages. Thus, it would allow higher utilization of a trunk.

Faster transmission of a long message can be achieved if it masquerades as short messages (i.e., by being broken up into smaller 15-byte packets). This feature is useful, for instance, for network management. Also, the average message delay would be reduced when the TIM discipline is combined with the module round-robin algorithm that is used in the ISN backplane.

Our delay formulas are valid for any PNP disciplines of which the TIM discipline is a special case. As a result of this analysis, we recognize that some of the TIM discipline's advantages result from the general PNP discipline, not just the TIM's priority logic. This observation would be useful for other applications. In our analysis, the usefulness of the G/G/1 conservation law has been further demonstrated. Finally, our method can be used for the multipriority discipline.

Acknowledgments

The author appreciates the fruitful discussions with members of the Systems Engineering and Network Center, in particular P. K. Verma, R. P. Kelly, J. A. Newell, and J. J. Sikora. The author would like to acknowledge S. P. Morgan and G. Sencer for valuable suggestions. Finally, the author would like to thank S. Boone, M. Braff, and M. G. Hluchyj for helpful comments.

References

1. AT&T, "An Introduction to ISN Information Systems Network," AT&T Customer Information Center, 2855 North Franklin Rd., P.O. Box 1991, Indianapolis, Indiana 46219, Document No. 555-300-020, Iss. 1, February 1986.
2. AT&T product announcement, International Communication Association Conference, Dallas, Texas, May 7, 1985.
3. A. G. Fraser and S. P. Morgan, "Queueing and Framing Disciplines for a Mixture of Data Traffic Types," *The Bell System Technical Journal*, Vol. 63, No. 6, July-August 1984, pp. 1061-1087.
4. M. G. Hluchyj, "Packet Switch Trunk Circuit Queueing Arrangement," U.S. Patent No. 464,533, issued February 17, 1987.
5. M. G. Hluchyj, C. D. Tsao, and R. R. Boorstyn, "Performance Analysis of a Pre-emptive Priority Queue with Applications to Packet Communication Systems," *The Bell System Technical Journal*, Vol. 62, No. 10, December 1983, pp. 3225-3245.
6. C. Palm, *Ericsson Techics*, Vol. 44, 1943, pp. 1-189.
7. A. Y. Khintchine, *Mathematical Methods in the Theory of Queueing*, 2nd ed., Griffin, London, 1969.
8. L. Kleinrock, *Queueing Systems*, Vol. I & II, J. S. Wiley & Sons—Interscience, New York, 1976.
9. L. Schrage, "An Alternative Proof of a Conservation Law for the Queue G/G/1," *Operations Research*, Vol. 18, 1970, pp. 185-187.
10. L. Kleinrock, *Communication Nets: Stochastic Message Flow and Delay*, McGraw-Hill, New York, 1964; reprinted by Dover Publications, Inc., New York, 1972.
11. "A Conservation Law for a Wide Class of Queueing Disciplines," *Naval Research Logistics Quarterly*, Vol. 12, 1965, pp. 181-192.
12. E. G. Coffman, Jr., and I. Mitrani, *Operations Research* Vol. 28, No. 3, Part II, May-June 1980, pp. 810-821.
13. D. Anick, D. Mitra, and M. M. Sondhi, "Stochastic Theory of a Data-handling System with Multiple Sources," *The Bell System Technical Journal*, Vol. 61, No. 8, October 1982, pp. 1871-1894.

Appendix A. Decomposition of a Poisson Process

Let P be a Poisson arrival process with arrival rate λ, and a service-length distribution $b(x)$. Let A and B be two servers, where A serves customers who require service time less than L and B serves the other customers. Because a Poisson process does not have memory, the service-length distributions in A and B are, respectively, the normalized functions of the original distribution function with an upper and lower cutoff at L.

Mathematically, we have

$$b(x) = b(x)\,\{\chi_{[O,L)}(x) + \chi_{[L,\infty)}(x)\} \tag{A1}$$

where $$\chi_{[a,b)}(x) = \begin{cases} 1 & \text{if } x \in [a,b) \\ 0 & \text{otherwise} \end{cases}$$

Let

$$b_h(x) = b(x)\,\chi_{[O,L)}(x) \tag{A2}$$

$$b_l(x) = b(x)\,\chi_{[L,\infty)}(x) \tag{A3}$$

Then, we have

$$b(x) = b_h(x) + b_l(x) \tag{A4}$$

Let

$$c_h = \int b_h(x)\;dx \tag{A5}$$

$$c_l = \int b_l(x)\;dx \tag{A6}$$

Then, c_h and c_l are the probability that a customer goes to A and B, respectively, for service. Because a Poisson process does not have memory, these probabilities are independent of the other arrivals. Therefore, the two arrival processes to A and B are, respectively, Poisson arrivals with arrival rates $c_h\lambda$ and $c_l\lambda$. Their service length distributions are, respectively,

$$b_h(x)/c_h \text{ and } b_l(x)/c_l \tag{A7}$$

Appendix B. Average Packet and Message Delays

We shall discuss the relationship between the average packet delays and average message delays for an M/G/1 system. These relationships, depending on the queueing disciplines used, can be simple or complicated.

For message nonpreemptive disciplines, the difference between the average packet delay and the corresponding average message delay is one-half the difference between two parameters, the "average" message length and the "average" packet length (see text), which are, respectively, functions of message-length and packet-length distributions and are independent of the traffic loads. However, for queueing disciplines that are not message nonpreemptive, the relationship can be complicated. For instance, for the round-robin queueing discipline, there does not seem to be a simple relationship between packet delays and message delays. This occurs because not only is the round-robin queueing discipline not nonpreemptive in terms of messages, but also the packets from different messages are well mixed.

The PNP queueing discipline, while not nonpreemptive in terms of messages, maintains the order of the messages from the same class (i.e., messages of the same class go out in the same order that they arrived, and their packets are not intermixed). This characteristic makes it easier to calculate the average message delays from the average packet delays and vice versa. We shall show this relationship here.

Before doing so, we shall do some simple calculations that support the conclusions in the main text.

Nonpreemptive Disciplines. Let us consider the special case that all the packets have length L. Let y be the waiting time of a message M of length NL. Then, the waiting time for the first packet of the message M is y; the waiting time for the second packet is $y + L$; and the waiting time for the ith packet is $y + (i - 1)L$. Consequently, the average waiting time for the packets from the message M of length x is

$$y + \frac{1}{2}(N - 1)L = y + \frac{1}{2}(x - L) \qquad \text{(B1)}$$

It thus follows that, for this case, the difference between the average packet delay and the message delay is

$$\frac{1}{2}(x - L) \qquad \text{(B2)}$$

Then, we must take the overall average in terms of packets. Let $b(x)$ be the message-length distribution. The packet distribution in terms of x is

$$\frac{(x/L)\,b(x)}{\bar{x}/\mathrm{L}} = \frac{xb(x)}{\bar{x}} \qquad \text{(B3)}$$

It follows that such an average is

$$\frac{1}{2}\left[\frac{\overline{x^2}}{\bar{x}} - L\right] \qquad \text{(B4)}$$

Comparing formula (B4) with (6a), we get relation (6b): $Z_j = L$. At the same time, we have also shown through our calculations that Z_j is a function of the jth packet-length distribution.

The PNP Discipline. Because the PNP discipline is not nonpreemptive in terms of messages, relationship (6b) between the average packet delay and the corresponding average message delay is generally not valid. More detailed analysis is required to obtain the average-message-delay formulas. For high-priority messages, the relationship between the average packet delay and the average message delay is the same as formula (6b), because there is no interruption between packets of the same high-priority message. (For the TIM discipline, the average high-priority packet delay and message delay are the same.)

Thus, it remains to derive the average-message-delay formula for low-priority messages. To obtain this formula, we compare the average-packet-delay formulas for the PNP and MNP disciplines. Observe that, if the arrivals of high-priority messages do not interrupt the transmission of a low-priority message, then the waiting time for this message is the *same* whether the queueing discipline is MNP or PNP. However, if an interruption does occur, then this low-priority message will have larger delay than the delay due to the MNP discipline.

To avoid unnecessary mathematical complications, we consider the case when all packets have length L. Consider a message M that consists of N packets. The waiting time of its first packet is y. However, message M is interrupted once by the arrival of high-priority messages, and the duration of this interruption is z. Then, y would be also the waiting time of message M if we used the MNP discipline.

Suppose that the interruption occurs after the ith packet is transmitted. Then, among the packets of message M, the average packet delay would be

$$y + \frac{[(N-i)z + N(N-1)L/2]}{N} \quad \text{(B5)}$$

where L is the packet length. However, we must also consider that the interruption can occur with equal probability for $i = 1, \ldots, (N-1)$. Then, for these packets, the average delay should be

$$y + \frac{1}{N(N-1)} \sum_{i=1}^{N-1} \left\{ (N-i)z + \frac{N(N-1)L}{2} \right\}$$

$$= y + \frac{(N-1)L}{2} + \frac{z}{2} \quad \text{(B6)}$$

Formula (B6) remains the same even if more than one interruption occurred, because the interruptions are independent of each other due to the memoryless nature of Poisson processes. However, z in formula (B6) then would be the *total* duration of the interruptions. Notice that, in formula (B6), the sum of the first and second terms would be the average waiting time of these packets if we used the MNP discipline and that the interruption effect for the average delay of these packets is only one-half the interruption effect for message M. More important, the factor 1/2 is *independent* of the message size NL.

To obtain the correct expression for the interruption term of the average message delay, we must take the averages of z in terms of packets and messages. For convenience, we shall change our notation slightly.

Let x be the length of a low-priority message and $z(x)$ be the total duration of the interruptions for a low-priority message of length x. Because high-priority message arrivals are Poisson processes, on the average, the disruption effect for a message of length x should be proportional to $(x - L)$. In other words, we have

$$z(x) = C(x - L) \quad \text{(B7)}$$

where C is a function that is independent of x. Then, the average of $z(x)$ in terms of messages is

$$C(\bar{x} - L) \quad \text{(B8)}$$

To take the average in terms of packets, as before, we use the distribution function (B3). Then, the average of $(1/2)z(x)$ in terms of packets is

$$\frac{C}{2}\left[\frac{\overline{x^2}}{\bar{x}} - L \right] \quad \text{(B9)}$$

Comparing formula (B9) with the interruption term in formula (14b), we have

$$C = \frac{\rho_h}{1 - \rho_h} \quad \text{(B10)}$$

because $Z = L$ for this case. Thus, we complete the derivation of the average low-priority message delay formula from the average low-priority packet delay formula. The complete formula has the following form:

$$\overline{W}_l = \frac{1}{2}\frac{A\rho_h + B\rho l}{(1-\rho_h)(1-\rho)} + \overline{[x]}\frac{\rho_h}{1-\rho_h} \quad \text{(B11)}$$

where $[x] = x$ minus the length of the last packet.

From formula (B7), we may further state that, for a message of length x, the average waiting time is

$$\overline{W}_l(x) = \frac{1}{2}\frac{A\rho_h + B\rho_l}{(1-\rho_h)(1-\rho)} + [x]\frac{\rho_h}{1-\rho_h} \quad \text{(B12)}$$

Let $\overline{W}_l(x, P_l)$ be the average waiting time of the last packet. Then, we have

$$\begin{aligned}\overline{W}_l(x, P_l) &= \overline{W}_l(x) + [x] \\ &= \frac{1}{2}\frac{A\rho_h + B\rho_l}{(1-\rho_h)(1-\rho)} + \frac{[x]}{1-\rho_h} \quad \text{(B13)}\end{aligned}$$

The average waiting time of the jth packet can be obtained similarly, simply by substituting $(j-1)L$ for $[x]$ in formula (B13).

Remarks on G/G/1 Conservation Law. Here, we would like to discuss the G/G/1 conservation law in connection with the arrivals of packets. In the original proof of the G/G/1 conservation law, there is an implicit assumption that the service-time-distribution function is independent of the arrival process. When the last packet of a low-priority message is not full, this condition might not be satisfied because the length of the related message determines the length of the last packet. This is an interesting academic problem.

However, because the G/G/1 conservation law is valid when the length of the last packet of a long message is full, the average-packet-delay formula (14) is at least a highly accurate approximation for the general case. Notice that the length of long messages is, on the average, much longer than the packet length. Moreover, we can consider virtual subpackets that have a length of $1/n$ of the full packet length. As the limit of the integer n gets very large, the length of a virtual subpacket approaches zero. Then, we can apply the G/G/1 conseration law to these virtual subpackets. This will justify formula (14) as a highly accurate approximation. If we consider that the instantaneous message arrival assumption is only a mathematical idealization of the realistic situation, then further discussion has little practical meaning and is only of academic interest. This is beyond the scope of this application-oriented paper.

Finally, the validity of formulas (15) and (16) does not depend on the G/G/1 conservation law being valid for the case when the last packet of a long message is not full length. The reason is that the derivation of these formulas are generalization of the full-length case based on physical considerations.

(Manuscript received April 29, 1986)