Fig. 6. Distribution of the number of needed reswitchings to unblock the system. Random hunting strategy. (a) Extensivity. The distribution is independent of N . (b) $n = r$; the distribution is independent of any structural parameter.

Fig. 7. Global comparison between random and packing hunting strategies. Average number R of reswitched calls per new arrival versus the traffic demand t_D , computer simulations. Packing hunting strategy is the best.

Bernoulli's simultaneous connection distribution for inlet and outlet blocks is assumed:

$$P_R(n, r, t_D) = \prod_{j=0}^{r-n} \frac{n-1-j}{r-j} t_D^j (2-t_D)^{2n-r-2}. \quad (4)$$

Expressions (2) and (4) solve exactly in the assumption of our

model the rearrangeability problem for the random hunting strategy.

IV. CONCLUSION

In three-stage Clos networks a sizeable saving of hardware can be obtained by the reconfiguration of calls. To study this point we have considered the application of the maximum entropy principle with constraints in its simpler form, combined with the concept of thermodynamic limit.

In this limit we have shown that the probability of unblocking any blocked state by rerouting j calls in progress is independent of b and decreases exponentially when the label j tends to infinity.

Expressions (2) and (4) give the exact solution to the problem for the random hunting strategy.

The results have become apparent through the matrix representation we used and, consequently, can be applied to all systems having the same logical representation (time-division switching systems [7], [8], time-slot assignment in communication satellite systems).

REFERENCES

- [1] V. E. Beneš, *Mathematical Theory of Connecting Networks*. New York: Academic, 1965.
- [2] M. Kubale, "Worst-case analysis of some algorithm for rearranging three-stage connection networks," presented at the 9th Int. Teletraffic Cong., Torremolinos, Spain, 1979.
- [3] E. Bonomi, M. R. Feix, and G. Hebuterne, "Théorie de l'information: Structure et modélisation d'un réseau de connexion," *Ann. Télécommun.*, vol. 35, Jan.-Feb. 1980.
- [4] A. Jajszczyk and J. Rajska, "The effect of choosing the switches for rearrangements in switching networks," *IEEE Trans. Commun.*, vol. COM-28, Oct. 1980.
- [5] T. Jaynes, "Where do we stand on maximum entropy?," in *The Maximum Entropy Formalism*, R. D. Levine and M. Tribus, Eds., M.I.T., Cambridge, MA, 1979.
- [6] E. Bonomi, J. L. Lutton, and M. R. Feix, "Telephone networks: Statistical mechanics and nonrandom connecting procedures," in *Numerical Methods in the Studies of Critical Phenomena*, J. Della Dora, J. Demongeot, and B. Lacolle, Eds. New York: Springer-Verlag, 1981.
- [7] M. Karnaugh, "Loss of point-to-point traffic in three-stage circuit switches," *IBM J. Res. Develop.*, May 1974.
- [8] M. Akyama, "Rearrangement in time-division system," *Electron. Commun. Japan*, vol. 51-A, no. 4, 1968.

A Communications System Which Prioritizes the Queues, Service Channels, and Traffic

DAVID T. LOKERSON, MEMBER, IEEE

Abstract—A new priority communications protocol has been defined and analyzed. This M/M/C/K system is unique because priorities are assigned not only to the incoming traffic, but also to the queues and the service channels (which may process calls only of an equal or higher priority). This nonpreemptive priority assignment technique

Paper approved by the Editor for Communication Switching of the IEEE Communications Society for publication without oral presentation. Manuscript received August 31, 1981; revised June 3, 1982. This work is based on a dissertation submitted in partial fulfillment of the D.Sc. degree from the George Washington University, Washington, DC.

The author is with the U.S. Army Electronics Research and Development Command, Harry Diamond Laboratories, Adelphi, MD 20783.

significantly improves the grade of service and reduces the average waiting time for high priority calls while reducing only slightly the system's utilization factor.

I. INTRODUCTION

Most communications systems prioritize the traffic, but not the queues or service channels. For nonpreemptive disciplines, the probability of receiving immediate service can be increased for the higher priority calls if the service channels are assigned a priority and are permitted to service traffic of only an equal or higher priority. In general, this system may have P priority levels while each priority has a queue capacity of $Q[p]$ and a service channel capacity of $C[p]$. The traffic demand for each priority is assumed to be an independent, infinite population source with separately defined random call generation rates of $\lambda[p]$, and average call completion rates of $\mu[p]$.

Each priority system is similar to an $M/M/C[p]/Q[p]$ system until some of the calls having priorities greater than the lowest priority would be delayed. During this busy period, this nonpreemptive priority discipline shifts the traffic demand so that it effectively expands the channel capacity for the higher priority calls and reduces the effective capacity for the lower priority traffic.

Various strategies are possible to determine which of many idle channels should be assigned to service a newly arriving call. An assign-to-highest idle channel strategy will be used in the analysis which follows, but the same analytical technique may be used for assign-to-lowest or other strategies. This priority technique has the advantage over other methods in that it reserves some channel capacity for the exclusive use of the highest priority traffic. When the highest priority channel capacity has been allocated to match the highest priority traffic, the probability of delay for that traffic can be kept low in spite of heavy lower priority traffic demand. Under this matched condition, the decrease in system throughput may be an acceptable penalty for the high probability of immediate service for the highest priority traffic.

The analytical methods to provide closed form solutions to queueing systems such as a nonprioritized, multiple server system with a single queue has been widely published [1]–[3]. These papers also demonstrate how to evaluate the performance of the system for prioritized traffic that share the same queues and service channels. They show that prioritization does not alter the probability of delay and blockage. In contrast, this paper evaluates the behavior of a unique priority system which changes these probabilities.

Simple Queues

Simple queueing systems are frequently assumed to have a single source of calls which have Markovian (i.e., exponential time) distributions for both the call generation and the service times. These are also called Poisson distributions and are used because they model a memoryless random process. The average call generation and call duration (i.e., service) rates are λ and μ , respectively. The number of service channels is C , and the maximum queue capacity is Q . The maximum number of calls within the system, L , equals the number of lines required to carry both calls being held in the queue and calls being serviced. Therefore, $L = Q + C$. Closed form solutions are derived by defining the state of the system as determined by the number of calls in the system. There are $(L + 1)$ of these states which range from no calls in the system to L calls. Equations

of balance between the rate of leaving and entering each of these $(L + 1)$ states can be written in terms of the call generation rate, the call service rate, and the probability of being in the previous and present states. For a steady state solution, the average rate of leaving each state must equal the average rate of entering that state; so these $(L + 1)$ balance equations define the required equality of these rates. An additional constraint equation forces the sum of the state probabilities to be unity; so these $(L + 2)$ equations permit the $(L + 1)$ state probabilities to be uniquely determined. The traffic demand D is also called the call intensity and equals λ/μ , which is equivalent to the average number of channels that would be required to continuously service all calls that would like to be serviced. (Not all of these calls are serviced because newly arriving calls are blocked and cleared from the system when the queue is full. Also, a large queue would be required to store the calls so that the service channels would be continuously busy since the traffic is arriving randomly.) The probability that each service channel is busy if all calls were serviced, ρ , is equal to D/C and $\lambda/\mu C$, and is usually called the utilization factor. The performance of systems can be measured in many ways. The probability of a call receiving immediate service, I , and the probability of being blocked from service, B , are meaningful measures which will be used later. (Many authors in communications use the grade of service, GOS, which is equivalent to the probability of being delayed and is equal to $1 - I - B$.)

Prioritized Queues and Service Channels

Fig. 1 illustrates this unique priority system. The solution of this nonpreemptive priority system requires the use of a method analogous to that described for simple queues, but this method becomes much more complex because there are so many more states. In addition, the balance equations have many more terms to define the effects of the call generation and completion for each priority. These balance equations also define the channel assignment strategy through the definition of the change in state caused by the generation or completion of a call. The complexity is increased because the priority of the generated call, and the priority of both the call being completed and the channel servicing that call, will determine the change in state. This cross coupling to many states prevents a simple solution of the balance equations. Even the simple three priority case shown in Fig. 1 is defined by 90 states (which are listed in Table I).

The first column in Table I is the sequential value assigned to the states. The next six columns define each state in terms of the number of calls in each queue and the priority of the call in each service channel. The highest priority queue and service channel are identified as $Q3$ and $C3$, respectively. The last three columns define how the system responds to a newly arriving call of each priority (while in that state). An "I" indicates that the call is serviced immediately with no delay because an acceptable channel is idle. A "D" indicates that the call will be delayed by being held in the queue until an acceptable channel becomes available. A "B" indicates that the call is blocked and cleared without being serviced because the queue is full. The solution of state probabilities for the system shown in Fig. 1 requires the solution of a 90 by 90 matrix which would have 8100 elements. Of these, 430 elements have nonzero values; but most of these elements are the result of summing up to six dual factors. Six terms are required to include the influence of the call generation and call completion

Fig. 1. Three priority system, traffic flow diagram.

TABLE I
DEFINITION OF ALLOWED STATES AND THE SERVICE
PROVIDED TO NEW CALLS

STATE VALUE	STATE						SERVICE FOR PRIORITY			STATE VALUE	STATE						SERVICE FOR PRIORITY		
	Q3	Q2	Q1	C3	C2	C1	3	2	1		Q3	Q2	Q1	C3	C2	C1	3	2	1
1	0	0	0	0	0	0	I	I	I	46	0	1	0	0	3	1	I	B	D
2	0	0	0	0	0	0	I	I	I	47	0	1	0	0	3	1	I	B	D
3	0	0	0	0	0	0	I	I	I	48	0	1	0	0	3	1	I	B	D
4	0	0	0	0	0	0	I	I	I	49	0	1	0	0	3	1	I	B	D
5	0	0	0	0	0	0	I	I	I	50	0	1	0	0	3	1	I	B	D
6	0	0	0	0	0	0	I	I	I	51	0	1	0	0	3	1	I	B	D
7	0	0	0	0	0	0	I	I	I	52	0	1	0	0	3	1	I	B	D
8	0	0	0	0	0	0	I	I	I	53	0	1	0	0	3	1	I	B	D
9	0	0	0	0	0	0	I	I	I	54	0	1	0	0	3	1	I	B	D
10	0	0	0	0	0	0	I	I	I	55	0	1	0	0	3	1	I	B	D
11	0	0	0	0	0	0	I	I	I	56	0	1	0	0	3	1	I	B	D
12	0	0	0	0	0	0	I	I	I	57	0	1	0	0	3	1	I	B	D
13	0	0	0	0	0	0	I	I	I	58	0	1	0	0	3	1	I	B	D
14	0	0	0	0	0	0	I	I	I	59	0	1	0	0	3	1	I	B	D
15	0	0	0	0	0	0	I	I	I	60	0	1	0	0	3	1	I	B	D
16	0	0	0	0	0	0	I	I	I	61	0	1	0	0	3	1	I	B	D
17	0	0	0	0	0	0	I	I	I	62	0	1	0	0	3	1	I	B	D
18	0	0	0	0	0	0	I	I	I	63	0	1	0	0	3	1	I	B	D
19	0	0	0	0	0	0	I	I	I	64	0	1	0	0	3	1	I	B	D
20	0	0	0	0	0	0	I	I	I	65	0	1	0	0	3	1	I	B	D
21	0	0	0	0	0	0	I	I	I	66	0	1	0	0	3	1	I	B	D
22	0	0	0	0	0	0	I	I	I	67	1	0	0	0	3	2	B	D	D
23	0	0	0	0	0	0	I	I	I	68	1	0	0	0	3	2	B	D	D
24	0	0	0	0	0	0	I	I	I	69	1	0	0	0	3	2	B	D	D
25	0	0	1	0	0	0	I	I	I	70	1	0	0	0	3	2	B	D	D
26	0	0	1	0	0	0	I	I	I	71	1	0	0	0	3	2	B	D	D
27	0	0	1	0	0	0	I	I	I	72	1	0	0	0	3	2	B	D	D
28	0	0	1	0	0	0	I	I	I	73	1	0	0	0	3	2	B	D	D
29	0	0	1	0	0	0	I	I	I	74	1	0	0	0	3	2	B	D	D
30	0	0	1	0	0	0	I	I	I	75	1	0	0	0	3	2	B	D	D
31	0	0	1	0	0	0	I	I	I	76	1	0	0	0	3	2	B	D	D
32	0	0	1	0	0	0	I	I	I	77	1	0	0	0	3	2	B	D	D
33	0	0	1	0	0	0	I	I	I	78	1	0	0	0	3	2	B	D	D
34	0	0	1	0	0	0	I	I	I	79	1	0	0	0	3	2	B	D	D
35	0	0	1	0	0	0	I	I	I	80	1	0	0	0	3	2	B	D	D
36	0	0	1	0	0	0	I	I	I	81	1	0	0	0	3	2	B	D	D
37	0	0	1	0	0	0	I	I	I	82	1	0	0	0	3	2	B	D	D
38	0	0	1	0	0	0	I	I	I	83	1	0	0	0	3	2	B	D	D
39	0	0	1	0	0	0	I	I	I	84	1	0	0	0	3	2	B	D	D
40	0	0	1	0	0	0	I	I	I	85	1	0	0	0	3	2	B	D	D
41	0	0	1	0	0	0	I	I	I	86	1	0	0	0	3	2	B	D	D
42	0	0	1	0	0	0	I	I	I	87	1	0	0	0	3	2	B	D	D
43	0	0	1	0	0	0	I	I	I	88	1	0	0	0	3	2	B	D	D
44	0	0	1	0	0	0	I	I	I	89	1	0	0	0	3	2	B	D	D
45	0	0	1	0	0	0	I	I	I	90	1	0	0	0	3	2	B	D	D

NOTES: 1. Priority 3 is the highest priority.
2. Service types are:
I = immediate (because an idle channel of an acceptable priority is available).
D = delayed (because no channel having a priority at or below that priority is idle).
B = blocked and cleared from being served (because the queue for that priority is full).

rates for each of the three priorities. Each term has an additional factor to weight these rates by the probability of being in the previous state. Since there would be a high probability of making an error in defining these elements, a Fortran computer program was written which determines which states could influence the balance equation for a given state. The 90 balance equations can be written as a product of a matrix M and a vector P , where each element of P is the steady state probability of the system being in that state. The Fortran program calculated the values of each element of the matrix, $M[90, 90]$ (which is determined by the call generation rates, the call completion rates, and the permitted state transitions). The balance equations are then written as

$$0 = M[S, S] * P[S]. \quad (1)$$

Since each element of $P[S]$ can be multiplied by a constant and (1) will still hold, the first element of $P[S]$ was chosen to be unity. Then, M was partitioned:

$$0 = \begin{bmatrix} f & \vdots & V'[W] \\ T[r] & \vdots & R[r, W] \end{bmatrix} * \begin{bmatrix} 1 \\ \vdots \\ X[W] \end{bmatrix} \quad (2)$$

where

$$r = P - 1$$

$$W = S - 1.$$

Now, f is a scalar; V , T , and X are vectors; and R is a matrix. Then

$$-f = V'[W] * X[W] \quad (3a)$$

and

$$-T[r] = R[r, W] * X[W]. \quad (3b)$$

Now, (3b) is solved for the unknown elements of X by inverting the matrix R :

$$X[W] = -R^{-1}[W, r] * T[r]. \quad (4)$$

Since the state probabilities $P[i]$ must add to unity, they are found as follows:

$$P[1] = \frac{1}{1 + \sum_{j=1}^{S-1} X[j]} \quad (5a)$$

and

$$P[i] = \frac{X[i-1]}{1 + \sum_{j=1}^{S-1} X[j]}. \quad (5b)$$

Once the state probabilities have been found, Table I defines which state probabilities must be summed to determine a given performance index. For example, the probability of

immediate service for a priority one call is the sum of the state probabilities for the states 1, 5, 9, 13, 17, and 21.

II. DISCUSSION OF RESULTS

As shown in Fig. 1 when $C[p] = 1$ and $Q[p] = 1$ for all p , there are six variables, $\lambda[p]$ and $\mu[p]$ for each of the three probabilities. The average call service rates were set at $\mu[1] = 0.5$, $\mu[2] = 1$, $\mu[3] = 2$ for this example because the highest priority (i.e., priority three traffic) has the shortest duration calls in most systems. The call generation rates were nominally set to $\lambda[1] = 0.75$, $\lambda[2] = 1.5$, and $\lambda[3] = 3$. This choice of parameters permits the nominal traffic demand to be 1.5 for each priority. To show how a changing demand affects the service, the traffic demand for one priority was varied from zero to four while the other priorities were held constant at 1.5.

The performance of this prioritized system was compared with two types of systems which will be called nonprioritized. One of these nonprioritized systems was called "dedicated" because it is a simple queue system where no traffic flow was permitted between priority systems. Therefore, the system performance was evaluated for $Q = 1$ and $C = 1$ while the traffic demand was varied from zero to four. The other nonprioritized system assumed that the lines and channels were "fully shared." Therefore, $Q = 3$ and $C = 3$ with the traffic demand varying from three (i.e., $1.5 + 1.5 + 0$) to seven (i.e., $1.5 + 1.5 + 4$).

Even though these are called nonprioritized systems because the queues and channels are not prioritized, it has been shown [1] that intermixed prioritized traffic would have the same probability of being delayed or blocked as traffic which was not prioritized because assigning priorities to traffic only determines the order of service and not the number of calls in the system. Therefore, prioritizing the traffic modifies only the waiting time and not the probability of blockage or delay.

Figs. 2-7 each show how the system performance changes with a variable traffic demand for five types of traffic: 1) nonprioritized, dedicated system (solid line), 2) nonprioritized, fully shared system (long dashed line), 3) prioritized system's highest priority (3) traffic (dash-dot-dashed line), 4) prioritized system's medium priority (2) traffic (short dashed line), and 5) prioritized system's lowest priority (1) traffic (dotted line). Figs. 2, 4, and 6 show the probability of immediate service (i.e., the probability of no delay), while Figs. 3, 5, and 7 show the probability of the call being blocked and cleared. Figs. 2 and 3 vary only the highest priority (3) traffic demand, Figs. 4 and 5 vary only the medium priority (2) traffic demand, and Figs. 6 and 7 vary only the lowest priority (1) traffic demand. It is important to remember that the solid line on each figure is for the dedicated system performance, so it should be compared only with the curve for the priority which is stated as being the variable. This assures a proper comparison for the same varying traffic demand.

Figs. 2, 4, and 6 show that the probability of immediate service is, as expected, highest for the highest priority, lower for the medium priority, and lowest for the lowest priority. In all three figures, the probability of receiving immediate service for both the medium and highest priority equals or exceeds that of the fully shared system for all of the illustrated traffic demand conditions. Fig. 2 shows that the highest priority traffic has more than three times the probability of receiving immediate service when compared with the fully

KEY TO CURVES

-
	Priority 1 Traffic (Lowest Priority)
	Call completion rate = 0.5, Demand = 1.5 or 0 to 4
-----	Priority 2 Traffic (Medium Priority)
	Call completion rate = 1.0, Demand = 1.5 or 0 to 4
-----	Priority 3 Traffic (Highest Priority)
	Call completion rate = 2.0, Demand = 1.5 or 0 to 4
————	Nonpriority, Dedicated
	L = 2, C = 1, Q = 1, Demand = 0 to 4
———	Nonpriority, Fully Shared
	L = 6, C = 3, Q = 3, Demand = 3 to 7
-

Fig. 2. Probability of no delay versus priority 3 traffic demand.

Fig. 3. Probability of blockage versus priority 3 traffic demand.

Fig. 4. Probability of no delay versus priority 2 traffic demand.

Fig. 5. Probability of blockage versus priority 2 traffic demand.

Fig. 6. Probability of no delay versus priority 1 traffic demand.

Fig. 7. Probability of blockage versus priority 1 traffic demand.

shared system. The medium priority traffic is also served better than the fully shared case. For this case of constant 1.5 medium priority traffic demand, the probability of immediate service when the highest priority traffic demand is near zero is a little greater than 30 percent because medium priority traffic demand cannot be serviced solely by the single medium priority channel. The medium priority traffic must therefore frequently (but not continuously) use the lowest priority channel. Since this is a nonpreemptive protocol, some of these calls must wait for the lowest priority channel to finish servicing the lowest priority call.

Fig. 4 shows that the variable medium priority traffic demand does not perform as well as the dedicated channel because the highest priority traffic demand of 1.5 decreases the effective channel capacity which is available to the medium priority traffic. The lowest priority traffic experiences probabilities of immediate service of less than 15 percent.

Fig. 6 shows that the probability of immediate service for the lowest priority traffic drops very rapidly from the fully shared case as the lowest priority traffic demand is increased. The highest priority traffic performs nearly twice as well as the medium priority traffic. The coupling of the lowest traffic demand on the higher priority traffic is clearly visible, but it is not severe.

Figs. 3, 5, and 7 show that the probability of the highest priority traffic being blocked and cleared is almost always less than for the fully shared case. Fig. 3 illustrates that the highest priority typically has less than half the blocking probability of the dedicated system for low traffic demands, and less than two-thirds at the high traffic demand. The probability of being blocked for the lower priority traffic is higher than for the highest priority traffic, but it is interesting to note that even the lowest priority traffic is not blocked to even the 80 percent level under heavy traffic loading. This is because the lowest priority calls in the queue enter the service channel for those brief instants when the other random traffic is momentarily light. Once that call enters the system, that call is not preempted. Since the lowest priority call service rate averages one-quarter of the highest priority's service rate, this lowest

priority service channel is not immediately available to the highest priority traffic during the peak busy period until that call has been fully serviced. The effect of unequal service rates is important in determining the shape of these curves.

Fig. 5 shows that the probability of the highest priority traffic being blocked is significantly decoupled from the medium priority traffic demand. Its probability of being blocked varies from a little less than 20 percent to only about 30 percent. It is coincidental that the medium priority traffic demand follows the dedicated traffic demand closely.

Fig. 7 confirms that the blocking of the higher priority traffic is very insensitive to the lowest priority traffic demand. A goal of this priority system design was to provide this decoupling; but it should be noted that when the lowest priority traffic demand is less than 0.4, the effect of some lowest priority, long service time calls causes even the highest priority traffic to experience a higher probability of blockage than for the fully shared case.

III. CONCLUSIONS

The unique priority system shown in Fig. 1 provides an improved performance for the higher priority traffic. The analytical solution was solved using the classical method with balance equations; but even with the simple system shown in Fig. 1, the closed form equations are so complex that the manual computation of matrix elements was abandoned. Instead, a computer program was written to compute the values of the elements of the balance equation matrix. If the number of service channels or the number of priorities is increased only a small amount, the number of allowed states and the computer memory and run times increase very rapidly, so this analytical approach is limited to only simple problems. It should be noted that this is a typical problem in many communications and priority system analyses [4].

The assign-to-highest prioritized service protocol has been shown to provide a unique alternate approach to a priority system design which can provide high performance at a small cost in overall system utilization. This system was analyzed using a method of state identification which is new and which can be expanded to larger systems and to other system protocols (e.g., assign-to-lowest). Interesting extensions of this analysis could determine the blocking effect caused by a limited input line capacity (as opposed to only a queue size limit). The development of an approximation method to permit a rapid analysis of large systems would be useful. The time domain simulation which was used to confirm the accuracy of the matrix solution ran significantly faster than the matrix solution when a random simulation error of about 4 percent was permitted.

REFERENCES

- [1] P. M. Morse, *Queues, Inventories, and Maintenance*. New York: Wiley, 1958.
- [2] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*. New York: Wiley, 1974.
- [3] M. Reiser, "Performance evaluation of data communications systems," *Proc. IEEE*, vol. 70, pp. 171-196, Feb. 1982.
- [4] J. M. Aein and O. S. Kosovych, "On capacity allocation strategies for DoD communications satellites," Sci. Technol. Div., Inst. Defense Analyses, Arlington, VA, paper P-1155, 1976.