TITLE BANDWIDTH AND BUFFER ALLOCATION IN A MULTISERVICE ENVIRONMENT, A PRIORITY CASE

AUTHOR(S) A.S. Drigas, E.N.Protonotarios

INSTITUTION Department of Electrical Engineering
National Technical University of Athens
42 Patission Street
106 82 Athens, Greece

SUMMARY

This article is consentrated on resource alloction aspects, in an ISDN or in a computer network where we distinguish more than one packet classes, in cases of random contention for M indentical resources from 2 or more statistically different packet types. In particular attention is focused on sharing bandwidth among voice and data virtual circuits, and sharing buffer before a multiserver system. These two problems can be commonly formulated as will be showed in this article.
The problem generally speaking is to determine the optimal policy for accepting or rejecting a call when the type of the requesting packet is known as well as the state vector with components the numbers of customers of each type that are in service. The optimal choice of buffer size and bandwidth is involved in the design of service facility as well as the rules of sharing resources among users.

The objective of this study is to develop analytical models and computational algorithms for the determination of the state subset with better performance for slotted time systems with call traffic modeled as stationary independent arrival processes and with service time modeled as a general discrete time distribution. The parameters optimized are these that are generally accepted as throughput, utilization and blocking of the system.

The term policy is usually reffering to the determination of the acceptable states of the system or in other words the operation of accepting or rejecting a call when the type of the requesting customer is known as well as the system state characterized by the allocation policy. The state is a vector with components the numbers of customers of each type that are in service.

1. INTRODUCTION

A priority model for buffer allocation for two or more packet classes that are going to be served by a multiserver and sharing bandwidth among several packet types, especially but not only withing the ISDN framework, are under consideration in this study. We examine a priority model because is crusial for some implementations, a class of packets to have priority in service, for example voice packets within an ISDN. The article is consentrated on these points although the method can be inplemented in a variety of problems that appeared in the literature. Some of these problems including the previous are numbered following: a) A number of user types try to get access to a host. There is an upper limit to the number of virtual circuits that can terminate to the host, b) k types of jobs are looking forward to be served by a limited number of processors in a multiprocessor machine or in a computer network [2], c) several types of customers contending to set up virtual circuits through a limited bandwidth channel in an ISDN [8], d) in some cases a memory of limited sized is chared among some packet types within a computer communication network [1],[4],[5],[6]. Especially in a ISDN framework communities of data and voice packets share a common memory before they get served by any server. Several articles presented in the literature considering similar problems of buffer and bandwidth allocation, [1],[7],[8],[9]. The usual goal of such studies is the determination of the optimal policy for a specific allocation scheme or the choise of the better policy among several of them, and the development of computational methods for that determination or choise.

2. THE MATHEMATICAL MODEL

Consider a multiserver system with a common waiting area including a total accomodation of N storage places. This buffer is shared between two or more packet types.

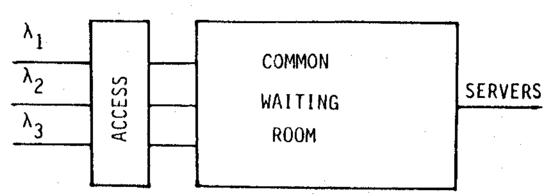

Fig. 1 Multiserver system with common waiting room.

Each packet from one type is going to be served with a first come first served policy. The service time for any customer type is not constant and is modeled as a general discrete time distribution. So the number of customers that are going to be served in a time slot ranges from 0 to Z, where Z is the number of serves. $\mu_1^{(j)}$ express the departure distribution probabilities for 1st class of packets, $\mu_1^{(1)}$ is the probability one packet of the 1st class to depart, $\mu_1^{(Z)}$ is the probability Z packets of 1st class to depart. Also $\lambda_1^{(i)}$ express the arrival distribution probabilities for 1st class packets, $\lambda_1^{(0)}$ is the probability no packet of 1st type to arrive, $\lambda_1^{(s)}$ is the probability s packets of 1st type to arrive, and is also modeled as a general discrete time distribution. The same holds for the other packet types. $\lambda_k^{(i)}$, is the arrival distribution probability for kth packet type as well as $\mu_k^{(i)}$, is the departure distribution probability.

Accordinly, for the bandwidth allocation problem, N is the bandwidth capacity, $\mu_j^{(i)}$ express the probability under which we receive i call clear packets from j packet class or better 1 call clear packet from j packet class which terminate a virtual circuit that used i units of bandwidth capacity, and finally $\lambda_j^{(i)}$ express the probability to receive i call set up packets from j packet type.

When a packet, or in terms of bandwidth allocation problem, a call set up packet, requests for service at its arrival, may be accepted or rejected according to the policy and the number of each packet type that there is already in the waiting area, or the virtual circuits that are already in action. We denote $Q_1(t_n)=n_1$, $Q_2(t_n)=n_2$, etc, the number of each packet type that are waiting for service or virtual circuits in action, at the beginning of the t_n slot. And that is what we call a policy now. A policy is the decision to accept or reject a call, when $Q_1(t_n)$, $Q_2(t_n)$, etc is know as well as the identity of the requesting packet. Also with no loss of generality we can assume that the arrivals of packets or call set up packets for buffer or bandwidth allocation problems respectively, at each time slot, take place exactly at the end of the slot. At the begining of each slot the multiserver system chooses by chance one of the waiting packet types and serves a number of it's population according to the probability distribution during the slot. In the bandwidth allocation problem, that means that we make the assumption that in each time slot we could receive call clear packet only from one type. That assumption is a close approximation to the reality if we choose short enough time slots.

We observe that any state of the system, (exept when the storage place or bandwidth is full) comes from another minus one or some packets or virtual circuits, of any type. As a result we understand that any policy must be a coordinate convex set of admissible states, as originally stated by Aein [3].

3. CALCULATION OF THE TRANSITION PROBABILITIES, GENERAL EQUATION

3.1 Possible transition for the two packet types case. One packet of a class could arrive at max in a slot.

Considering the state (n_1, n_2) lying in the interior of the state space in R^2, we have seven possible transitions into it. These are depicted in Fig.2 and are:

(i) Arrival of two packets, one of each type with service completion of one customer from either type (transitions from (n_1,n_2-1)).

(ii) Arrival of one packet of any type, with service completion of a customer of same type (transitions for (n_1,n_2)).

(iii) Arrival of one packet of either type, with service completion of a packet of the other type (transitions from (n_1-1,n_2+1) or (n_1+1,n_2-1)).

(iv) No new arrivals, with service completion of a customer from either type (transitions from (n_1,n_2+1) or (n_1+1,n_2)).

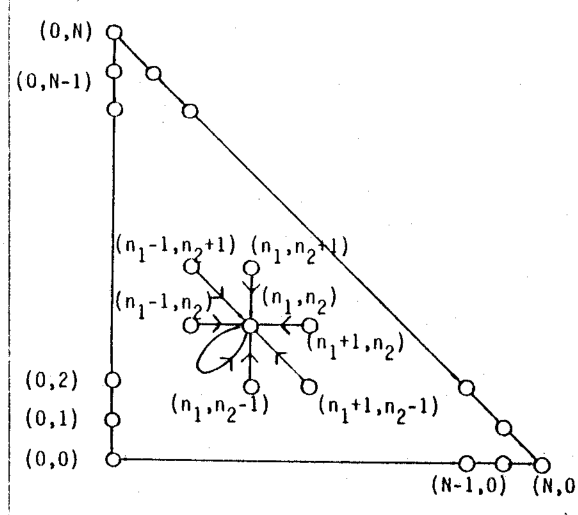

Fig. 2. State transition diagram for the two packet types case.

3.2. General equation

The general equation for the two packet types and for the arrival of one packet of a type at max in a slot case is:

$$Pr\{n_1,n_2)\} = \sum_{I_1=0}^{1} \sum_{I_1=0}^{1} \sum_{J=0}^{1}$$

$$[\prod_{i=1}^{2} \lambda_i^{(Ii)})A_J Pr\{(N_1-I_1),(N_2-I_2)\}]Y \qquad (1)$$

Where $N_i=n_i$ if $i<>k$ and $N_i=n_i+1$ if $i=J$ and Y is a coefficient of policy and $Y(n_1,n_2)=0$ if (n_1,n_2) doesn't belongs to Ω and $Y(n_1, n_2)=1$ if (n_1,n_2) belongs to Ω, where Ω is the set of admissible states. Also A_J is the priority coefficient and express the percentage priority to the J class of packets.

Now we can generalise this equation to hold to the case that s packets from one type could arrive in a slot, so the previous equation becomes:

$$\Pr\{n_1,n_2\} = \sum_{I_1=0}^{s} \sum_{I_2=0}^{s} \sum_{J=1}^{1}$$

$$[(\prod_{i=1}^{2} \lambda_i^{(Ii)}) A_J \Pr\{(N_1-I_1),(N_2-I_2)\}]Y \quad (2)$$

This shape of general equation, and with the use of policy's coefficient, makes possible for that equation to be able to express any state general or boundary, for any policy, exept for the state (0,0) and for the four states around it, which are common in any policy but we need some special equations there, which must include the phenomenon that in the (0,0) state, there isn't any packet to be served, so zero packets will be served under probability 1. So that, there are some possible transitions according to the packets that could arrive, that there aren't in other state anywhere in Ω.

So we have for Pr{0,0} = second part of the general equation + $\lambda_1^{(0)}\lambda_2^{(0)}$ Pr{0,0}. Similar equations hold for the four states around (0,0) state.

Now we can futher more generalize the previous equation for two packet types to hold under the situation that three or more packets classes, (w packet classes), could be distinguished in the network.

So we have for three packet types:

$$\Pr\{n_1,n_2,n_3\} = \sum_{I_1=0}^{s} \sum_{I_2=0}^{s} \sum_{I_3=0}^{s} \sum_{\zeta=1}^{z}$$

$$[(\prod_{i=1}^{3} \lambda_i^{(I_i)})/(\sum_{i=1}^{3} n_i - \sum_{i=1}^{3} I_i + \zeta)] \, .$$

$$. \sum_{k=1}^{3} \mu_k^{(\zeta)} . (n_k+\zeta-I_k) . \Pr\{(N_1-I_1),(N_2-I_2),$$

$$(N_3-I_3)\} . Y \quad (3)$$

And for w packe types we have:

$$\Pr\{n_1,n_2,\ldots,n_w\} = \sum_{I_1=0}^{s_1} \sum_{I_2=0}^{s_2} \ldots \sum_{I_w=0}^{s_w} \sum_{\zeta=1}^{z}$$

$$[(\prod_{i=1}^{w} \lambda_i^{(I_i)}) / (\sum_{i=1}^{w} n_i - \sum_{i=1}^{w} I_i + \zeta)] \, .$$

$$. \sum_{k=1}^{w} \mu_k^{(\zeta)} (n_k+\zeta-I_k) . \Pr\{N_1-I_1),\ldots,(N_i-I_i),\ldots,$$

$$(N_w-I_w)\} . Y \quad (4)$$

where $N_i = \begin{cases} n_i & \text{if } i \neq k \\ n_i+\zeta & \text{if } i = k \end{cases}$

4. THE METHOD

Let $P_n=[P_{n0},\ldots,P_{nm},\ldots,P_{nm}]^T$, where T means transposed and m depends on the policy. P_n for the three packet types case, contais all the probabilities Pr{n,i,j} for every addmissible i,j, as well as, for the two packet types case, contains all the probabilities Pr{n,i} for every addmissible i. From the construction of general equation (two or three customer types) and indepentently of the policy we are following, we can perceive that the system mooves to P_0, from $P_0,P_1,\ldots,P_n,\ldots,P_z$, where P_n is an one dimentional array that mentioned before. That observation leads us to perceive that we are able to express P_z as a function of $P_0,P_1,\ldots,P_z$, so we can write

$$P_z = A_z^{(0)}P_z + A_z^{(1)}P_{z-1} + \ldots + A_z^{(z)}P_0 \quad (5)$$

Generally speaking the system is mooved to P_{k-z} only for $P_{k-z-s}, P_{k-z-s+1},\ldots,P_{k-z},\ldots,P_k$.

So is possible to express P_k as

$$P_k=A_k^{(0)}P_k+A_k^{(1)}P_{k-1}+\ldots+A_k^{(z)}P_{k-z}+\ldots$$

$$+A_k^{(z+s)}P_{k-z-s} = \sum_{I=0}^{z+s} A_k^{(I)}P_{k-I} . \quad (6)$$

That means that we can express every P_k in terms of $P_0,P_1,\ldots,P_{z-1}$ as follows:

$$P_z = A_z^{(0)}P_z + A_z^{(1)}P_{z-1} + \ldots + A_z^{(z)}P_0$$

$$P_z = (I-A_z^{(0)})^{-1})A_z^{(1)}P_{z-1}+\ldots+A_z^{(z)}P_0)$$

$$P_z = B_z^{(z-1)}P_{z-1} + \ldots + B_z^{(0)}P_0 \quad (7)$$

Where $A_z^{(0)},\ldots,A_z^{(1)},\ldots,A_z^{(z)}$ are arrays of appropriate dimensions and

$$B_z^{(I)} = (I-A_z^{(0)})^{-1} \, . \, A_z^{(z-1)} \quad (8)$$

Now generally have

$$P_k = (I-A_k^{(0)})^{-1}(A_k^{(1)}P_{k-1} + \ldots +$$

$$+ A_k^{(z)}P_{k-z} + \ldots + A_k^{(z+s)}P_{k-z-s}) =$$

$$= (I-A_k^{(0)})^{-1}[A_k^{(1)}(B_{k-1}^{(0)}P_0+\ldots+B_{k-1}^{(z-1)}P_{z-1})+$$

$$+ A_k^{(2)}(B_{k-2}^{(0)}P_0 + \ldots + B_{k-2}^{(z-1)}P_{z-1}) +$$

$$\vdots$$

$$+ A_k^{(z+s)})B_{k-z-s}^{(0)}P_0 + \ldots + B_{k-z-s}^{(z-1)}P_{z-1})] =$$

$$= B_k^{(0)}P_0 + \ldots + B_k^{(z-1)}P_{z-1} \quad (9)$$

where

$$B_k^{(0)} = (I-A_k^{(0)})^{-1} \sum_{I=1}^{z+s} A_k^{(I)}B_{k-I}^{(0)}$$

$$\vdots$$

$$B_k^{(z-1)} = (I-A_k^{(0)})^{-1} \sum_{I=1}^{z+s} A_k^{(I)}B_{k-I}^{(z-1)}$$

So it is possible to express every $P_r\{i,j,k_I\}$ state that belongs to the k_I plane vertical to the k_I point of k axis as a function of P_0, $P_1,\ldots,P_{z-1}$ (array P_0 contains Pr{i,j,0} for every admissible i,j. Same holds for $P_1,P_2\ldots$). As a result we are able that way to express every P_k as a function of P_0, $P_1,\ldots,P_{z-1}$. That is result of using transition probabilities for every (i,j,k) state, exept for a set of states Ω_1. Ω_1 set contains all these states that have a distance (in the k orientation) that is smaller than z from the policy's surface. If we write now transition equations for all the states that belong to Ω_1, we take a linear homogeneous equation system. The number of these equations are the same as the number of states (i,j,k) for k=0. Substituting one of these equations with the normalization equation

$$\sum \sum \sum Pr\{i,j,k\} = 1 \quad \text{with } (i,j,k) \in \Omega \qquad (10)$$

we take a linear non homogeneous system of equations that can be solved. (the description of the method is made for three customer types. For more customer types an identical procedure to this can be implementend).

Once we have calculated probabilities for every admissible state of the system, we calculate performance characteristics which are: idlenes = Pr{0,0,0}, Utilization = = 1-Pr{0,0,0}, Rejection = $\sum_{\Omega_3} Pr\{i,j,k\} \cdot \lambda_1^{(\zeta)}$, where Ω_3 is all the states which have a distance from policy's surface smaller than s. And finally throughput = 1-Rejection.

5. CONCLUSION

The method that developed enables us to evaluate the performance of all the bandwidth and buffer in a multiservice environment allocation schemes, for two or more customer types, which give a coordinate convex state diagram and to choose the better policy for any special allocation scheme we need.

REFERENCES

[1] Gerard J. Foschini, B.Gopinath, "Sharing Memory Optimally", IEEE Trans. Commun., Vol. COM-31, pp.352-360, March 1963.

[2] G.I.Foschini, B.Gopinath and J.F.Hayes, "Optimum Allocation of servers to two types of Competing Customers", IEEE trans. Commun. Vol. COM-29, pp.1051-1055, July 1981.

[3] J.M.Aein and O.S.Kosovych, "Satellite capacity allocation", Proc. IEEE, Special Issue on Satellite Communications, Vol.65, pp. 332-341, March 1977.

[4] F.Kamoun and L.Kleinrock, "Analysis of chared finite storage in a computer network node environment under general traffic conditions", IEEE Trans., Commun., Vol.COM-28, pp. 922-1003, July 1980.

[5] J.S.Kaufman, "Blocking in a shared resource environment", IEE Trans. Commun., Vol.COM-29, pp.1474-1481, Oct.1981.

[6] M.I.Irland, "Buffer management in a packet switch", IEEE Trans. Commun., Vol. COM-26, Mar. 1978.

[7] G.Latouche, "Exponential server sharing a finite storage: Comparison of space allocation policies", IEEE Trans., Com., Vol. COM-28, pp. 910-915, June 1980.

[8] B.Kraimech and M.Schawarz, "Analysis of Traffic Access Control Strategies in Integrated Service Networks", IEEE Trans. Com., Vol.33, pp.1085-1093, October 1985.

[9] A.S.Drigas, E.N.Protonotarios, "BUFFER ALLOCATION FOR THREE OR MORE CUSTOMER TYPES", EURINFO 88.

This work has supported by general secretariat of research and technology of Greece (Project Ariadne).