INTERNATIONAL CONFERENCE ON DATA COMMUNICATIONS TECHNOLOGY DEPT. OF ELECTRONIC AND COMPUTER ENGINEERING 21-23 SEPT., 1988 NIHE, LIMERICK, IRELAND

PROVISIONAL PROGRAMME

DCT CONFERENCE

April 1988

NIHE
Limerick's Technological University
Plassey Technological Park, Limerick, Ireland. Tel. 061 333644 Telex 26959

Bandwidth and Buffer allocation in a Multiservice environment for three or more customer types

by

A.S. Drigas, E.N. Protonotarios

Department of Electrical Engineering
National Technical University of Athens
42 Patission Street
106 82 Athens, Greece
Tel: +301 3612600, Tlx: 221682 NTUA GR, Fax: +301 7757501

Abstract : Resource allocation aspects are considered in cases of random contention for M identical resources from k statistically different customer types. In particular attention is focuced on sharing bandwidth among customer classes within the ISDN framework or allocating packets in a buffer before a multiserver system. These two problems can be commonly formulated as will be showed in this article.

The general problem is to determine the optimal policy for accepting or rejecting a call when the type of the requesting customer 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 the ones generally accepted as throughput, utilization and blocking of the system.

The term policy is usually reffered 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 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

Buffer allocation for three or more packet classes that are going to be served by a multiprocessor machine and sharing bandwidth among several packet types, especially but not only within the ISDN framework, are under consideration in this study. The aspect of a limited number of resources that are shared among several customer or packet classes, is a well known problem in the area of computer communications and ISDN, as well as in concurrent and decentralized computer systems. Examples of such situations 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 size is chared among some packet types within a computer communication network [1],[4],[5],[6]. Especially in an ISDN framework communities of data, voice and video 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 among three packet classes.

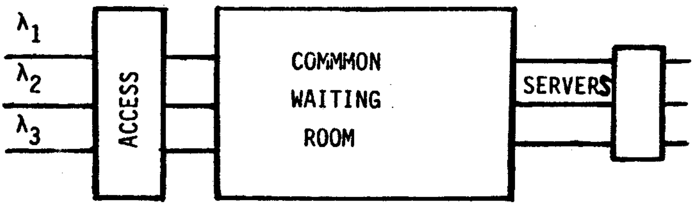

Fig. 1. Multiserver system with common waiting room.

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 servers. $\mu_1^{(i)}$ express the departure distribution probabilities for 1st class of packets, $\mu_1^{(1)}$ is the probability one packet of 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, $\lambda_1^{(o)}$ 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. $\lambda_2^{(i)}$, $\lambda_3^{(i)}$ are the arrival distribution probabilities for 2nd and 3rd packet types as well as $\mu_2^{(i)}$, $\mu_3^{(i)}$ are the departure distribution probabilities respectively.

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 requests for service at its arrival, or in terms of bandwidth allocation problem, a call accepted or rejected according to the policy and the number of each packet type that there is already in the waiting room, or the virtual circuits that are already in action. We denote $Q_1(tn) = n_1$, $Q_2(tn) = n_2$, $Q_3(tn) = n_3$ the number of each packet type that are waiting for service or V.C. in action, at the beginning of the tn slot. And that is what we call a policy now. A policy is the decision to accept or reject a call, when $Q_1(tn)$, $Q_2(tn)$, $Q_3(tn)$ is known 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 time 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 can observe that any state of the system, (exept when the storage place or bandwidth is full) comes from another minus one or some packets 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 STATE TRANSITION PROBABILITIES

3.1. Possible transitions

Let's begin from the case that one packet of a type could arrive at max in a slot, in a two server system. Considering the state (n_1,n_2,n_3) lying in the interior of the state space in R^3, there are 32 states from which could happen transitions in to it, that are figured in fig. 2 and are :

I) No new arrivals, with service completion of packets of a type from 0 to 2, [transitions from (n_1,n_2,n_3), (n_1,n_2,n_3+1), (n_1,n_2,n_3+2), (n_1,n_2+1,n_3), (n_1,n_2+2,n_3), (n_1+1,n_2,n_3), (n_1+2,n_2,n_3)].

II) Arrival of one packet of any type with service completion of a or more packets of any of other types, [transitions from (n_1+1,n_2-1,n_3), (n_1+1,n_2,n_3-1), (n_1+2,n_2-1,n_3), (n_1+2,n_2,n_3-1), (n_1-1,n_2+1,n_3), (n_1,n_2+1,n_3-1), (n_1-1,n_2+2,n_3), (n_1,n_2+2,n_3-1), (n_1-1,n_2,n_3+1),

(n_1,n_2-1,n_3+1), (n_1-1,n_2,n_3+2), (n_1,n_2-1,n_3+2)].

III) Arrival of one packet of any type with service completion of a or more packets of the same type, [transitions from (n_1,n_2,n_3), (n_1-1,n_2,n_3), (n_1,n_2-1,n_3), (n_1,n_2,n_3-1)].

IV) Arrival of one packet of any type with no service completion of any packet, [transitions from (n_1-1,n_2,n_3), (n_1,n_2-1,n_3), (n_1,n_2,n_3-1)].

V) Arrival of two packets of any two types with no service completion of any packet, [transitions from (n_1-1,n_2-1,n_3), (n_1-1,n_2,n_3-1), (n_1,n_2-1,n_3-1)].

VI) Arrival of two packets of any two types, with service completion of a or two packests of the third type, [transitions from (n_1-1,n_2-1,n_3+1), (n_1-1,n_2-1,n_3+2), (n_1-1,n_2+1,n_3-1), (n_1-1,n_2+2,n_3-1), (n_1+1,n_2-1,n_3-1), (n_1+2,n_2-1,n_3-1)].

VII) Arrival of two packets of any two types with service completion of one or two packets of one of these two types, [transitions from (n_1-1,n_2,n_3), (n_1-1,n_2+1,n_3), (n_1,n_2-1,n_3), (n_1+1,n_2-1,n_3), (n_1,n_2,n_3-1), (n_1+1,n_2,n_3-1), (n_1-1,n_2,n_3+1), (n_1,n_2+1,n_3-1), (n_1,n_2-1,n_3+1)].

VIII) Arrival of one packet of all types with service completion of 0 to 2 packets from any type, [transitions from (n_1-1,n_2-1,n_3-1), (n_1,n_2-1,n_3-1), (n_1+1,n_2-1,n_3-1), (n_1-1,n_2,n_3-1), (n_1-1,n_2+1,n_3-1), (n_1-1,n_2-1,n_3), (n_1-1,n_2-1,n_3+1)].

Extending the above to the w-types case, single at max arrivals and z at max departures, in R^W, we find $2^W(z.w+1)$ transitions into $(n_1,..,n_W)$. More specifically we have w.w transitions if a packet from any type arrives with service completion of a packet from any type, or w.w!/[m!(w-m)!] transitions if m packets form any type arrive. So we have a total of $w.\sum_{m=0}^{W} w!/[m!(w-m)!] = w.2^W$ transitions. Considering now that can happen from 0 to z departures, we have $2^W + z.w.2^W = 2^W(z.w+1)$ total transitions.

3.2. Equations for three packet types

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

$$\Pr\{n_1,n_2,n_3\} = \sum_{I_1=0}^{1}\sum_{I_2=0}^{1}\sum_{I_3=0}^{1}\sum_{\zeta=1}^{z}\Big[\Big(\prod_{i=1}^{3}\lambda_i^{(I_i)}\Big)\Big/\Big(\sum_{i=1}^{3}n_i-\sum_{i=1}^{3}I_i+\zeta\Big)\Big].$$

$$\cdot\sum_{k=1}^{3}\mu_k^{(\zeta)}.(n_k+\zeta-I_k).\Pr\{(N_1-I_1),(N_2-I_2),(N_3-I_3)\}.Y$$

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

and Y is a coefficient of policy and $Y(n_1,n_2,n_3) = 0$ if $(n_1,n_2,n_3) \notin 0$ and $Y(n_1,n_2,n_3) = 1$ if $(n_1,n_2,n_3) \in \Omega$, where Ω is the set of admissible states.

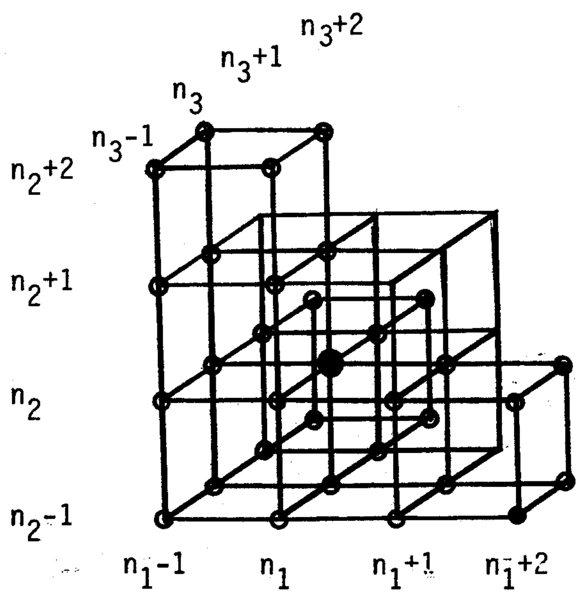

Fig. 2. 32 states around (n_1,n_2,n_3) from which can happen transitions into it.

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 (0,0,0) and for the seven states around it, that are common in any policy but we need some special equations there, which must include the phenomenon that in the (0,0,0) 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 states anywhere in Ω.

So we have for Pr{0,0,0} = second part of general equation + $\lambda_1^{(o)}\lambda_2^{(o)}\lambda_3^{(o)}$.Pr{0,0,0}. Similar equations hold for the seven states around (0,0,0) state.

3.3. General equation

The previous general equation for three packet types can be extended to the case that s_k packets of k type could arrive at max in a slot, simply extending the limits of three first sums of one to s_k. This general equation for three packet classes can also be extended to the case of w packet classes as follows:

$$\Pr\{n_1,n_2,\dots,n_w\} = \sum_{I_1=0}^{s_1}\sum_{I_2=0}^{s_2}\cdots\sum_{I_w=0}^{s_w}\sum_{\zeta=1}^{z}\Big[\Big(\prod_{i=1}^{w}\lambda_i^{(I_i)}\Big)\Big/\Big(\sum_{i=1}^{w}n_i-\sum_{i=1}^{w}I_i+\zeta\Big)\Big].$$
$$.\sum_{k=1}^{w}\mu_k^{(\zeta)}(n_k+\zeta-I_k).\Pr\{(N_1-I_1),\dots,(N_i-I_i),\dots,(N_w-I_w)\}.Y$$

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

4. THE METHOD

The construction of general equations, (three customer types, s maximum arrivals and z maximum departures) permits us to observe that the system mooves to P_o, from $P_o,P_1,\dots,P_n,\dots,P_z$, where $P_n = [P_{no},\dots,P_{ni},\dots,P_{nm}]^T$ is an one dimension array with the probabilities Pr{n,i,j} for every admissible i,j. That observation leads us to perceive that we are able to express P_z as a function of $P_o,P_1,\dots,P_z$, so we can write

$$P_z = A_z^{(o)}P_z + A_z^{(1)}P_{z-1} + \dots + A_z^{(z)}P_o$$

Generally speaking system is mooved to R_{k-z} only from $P_{k-z-s},P_{k-z-s+1},\dots,P_{k-z},\dots,P_k$.

So is possible to express P_k as

$$P_k = A_k^{(o)}P_k + A_k^{(1)}P_{k-1} + \dots + A_k^{(z)}P_{k-z} + \dots + A_k^{(z+s)}P_{k-z-s} = \sum_{I=0}^{z+s} A_k^{(I)}P_{k-I}$$

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

$$P_z = A_z^{(o)} P_z + A_z^{(1)} P_{z-1} + \ldots + A_z^{(z)} P_o \iff$$

$$P_z = (I - A_z^{(o)})^{-1} (A_z^{(1)} P_{z-1} + \ldots + A_z^{(z)} P_o) \iff$$

$$P_z = B_z^{(z-1)} P_{z-1} + \ldots + B_z^{(o)} P_o$$

where $A_z^{(o)}, .., A_z^{(I)}, \ldots, A_z^{(z)}$ are arrays of appropriate dimensions, and

$$B_z^{(I)} = (I - A_z^{(o)})^{-1} . A_z^{(z-I)}$$

Now generally have

$$\begin{aligned}
P_k &= (I - A_k^{(o)})^{-1} (A_k^{(1)} P_{k-1} + .. + A_k^{(z)} P_{k-z} + .. + A_k^{(z+s)} P_{k-z-s}) = \\
&= (I - A_k^{(o)})^{-1} [A_k^{(1)} (B_{k-1}^{(o)} P_o + .. + B_{k-1}^{(z-1)} P_{z-1}) + \\
&\qquad + A_k^{(2)} (B_{k-2}^{(o)} P_o + .. + B_{k-2}^{(z-1)} P_{z-1}) + \\
&\qquad \vdots \\
&\qquad + A_k^{(z+s)}) B_{k-z-s}^{(o)} P_o + .. + B_{k-z-s}^{(z-1)} P_{z-1})] = \\
&= B_k^{(o)} P_o + \ldots + B_k^{(z-1)} P_{z-1}
\end{aligned}$$

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

$$\vdots$$

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

So it is possible to express every $Pr\{i,j,k_T\}$ for the state (i,j,k_I) that belongs to the k_I plane, (plane vertical to the k_I point of k axis) as a function of $P_o, P_1, .., P_{z-1}$ (array P_o 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 $P_o, P_1, .., P_{z-1}$. That is a 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 korientation) 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

$$\sum_i \sum_j \sum_k \Pr\{i,j,k\} = 1 \quad \text{with } (i,j,k) \in \Omega,$$

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 implemented).

Once we have calculated probabilities for every admissible state of the system, we calculate performance characteristics which are : idleness = Pr{0,0,0}, Utilization = 1-Pr{0,0,0}, Rejection = $\sum_{\Omega_3} \Pr\{i,j,k\} \cdot \lambda_I^{(\zeta)}$, 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 three 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 1983.

[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 shared finite storage in a computer network node environment under general traffic conditions", IEEE Trans, Commun., Vol.COM-28, pp. 992-1003, July 1980.

[5] J.S. Kaufman, "Blocking in a shared resource environment", IEEE 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. Commun., Vol., COM-28, pp. 910-915, June 1980.

[8] B. Kraimeche and M. Schwartz, "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).