34-70 IX

FINITE SOURCE PRIORITY QUEUING SYSTEMS

by

G. Joachim B r a n d t

Institute for Switching and Data Technics

University of Stuttgart

Federal Republic of Germany

Proc. Internat. Seminar on Telecom.
Helsinki 1970

INTRODUCTION

The investigation of waiting lines is a basic problem in computerized data-transmitting networks, in computer-systems and in many other technical, but also economical situations. In my lecture some aspects of finite source queuing systems shall be treated.

The general queuing system has n servers and s waiting places. If at least one server is free, arriving calls are served immediately. If all servers are busy, arriving calls can wait in one of the s waiting places. Calls arriving during a time interval, when all servers as well as all waiting places are occupied, cannot be handled and are lost. This general loss-delay-system includes two special cases: For s=0 we have the loss-system, where each call is lost which cannot be served immediately, and for $s \to \infty$ we have the delay-system, where no call is lost.

The rules according to which the calls are served constitute the service discipline. There are several different possible disciplines, each of which has practical applications.

For FIRST COME - FIRST SERVED discipline the loss-delay-system can be calculated accordingly to the well-known formulae of ERLANG, if POISSON-traffic is offered. That means if the interarrival-time and the service-time is distributed negative exponentially. The solution of the waiting system with other traffic assumptions are

./..

connected with the names MOLINA, FRY, POLLACZEK, KHINTCHINE, CROMMELIN, PALM, COHEN and others. LOTZE has derived a method for the calculation of the traffic characteristics of delay-systems by means of the traffic characteristics of the loss-system.

For a finite number of POISSON-sources, the loss-delay-system has been solved by BAUER and STÜRMER. Before discussing the finite model, let us introduce some notations. When j sources are busy, then j calls are in the system. We say that this system is in state j. The mean arrival-time in state j is the inverse of the arrival rate λ_j. For an infinite number of sources the arrival rate is independent of j:

$$\lambda_j = \lambda \qquad (1)$$

In other words, when the number of subscribers is so large that the arrival process is not remarkably influenced by the number j of busy sources, then we are allowed to apply the infinite model. On the other hand, when the number of subscibers is of the same magnitude as the number of devices n+s, then we have to apply the finite source model.

A typical finite source situation exists, when terminals are connected to a time-shared computer processor. Another example is the group of subsribers, which is connected with its preselection stage. These subsribers form finite sources, which are allowed to wait preceding the preselectors. The number of sources is denoted by q and in general, we assume that all traffic sources have the same intensity. In this case, the arrival rate is a linear function of the number (q-j) of free sources:

$$\lambda_j = \frac{q-j}{q} \cdot \lambda_0 \qquad (2)$$

λ_0 is the arrival rate, when all sources are free. The greater the number of busy sources, the less is the arrival rate λ_j.

We have introduced the mean interarrival-time. Now we come to the service times which are assumed to be negative-exponentially distributed with the mean service-time h. Thus, we can define the termination rate μ_j:

$$\mu_j = \begin{cases} \frac{j}{h} & \text{for } j \leq n \\ \frac{n}{h} & \text{for } j > n \end{cases} \qquad (3)$$

At most n calls can be served at the same time, so that the terminating rate for $j>n$ equals the termination rate for $j=n$. The variable of state X (t) gives the number j of calls in the system at time t. The probability of state j, that means the probability that j calls are in the system at time t is denoted by $P_j(t)$:

$$P_j(t) = P\{X(t)=j\} \qquad (4)$$

We are interested in the stationary probabilities of state, which means in the time-independent solution:

$$P_j(t) = P_j \qquad (5)$$

The solution of the stationary probabilities for the loss-delay-system with a finite number of traffic sources are well known.

In order to obtain the mean arrival rate λ, we average the λ_j with the probability P_j that j calls are in the system:

$$\lambda = \sum_{j=0}^{n+s} \lambda_j \cdot P_j \qquad (6)$$

The arrival rate λ is the mean number of arriving calls per time unit. In the finite case the mean arrival rate λ is a traffic parameter, which is dependent on the probabilities of state and thereby on the number n of servers and on the number s of waiting-places.

Instead of λ, the parameter

$$A = h \cdot \lambda \qquad (7)$$

known as traffic offered, is often used.

In many waiting-systems, not all calls are of equal importance. In this case, we may associate a priority index i with each class where i=1 denotes the class with highest priority and i=r the class with lowest priority. Combining the calls of class 1,2,... until i to one group we refer to this group as class $\leq i$. A call of class $\leq r$ is an arbitrary call without regard of its priority class.

If calls of higher priority are not allowed to interrupt calls of lower priority, being in service, then we speak of non-preemptive priorities. Under this discipline, a call, which is in service, cannot be interrupted bythe arrival of a higher priority call.Thus, the priority call has to wait at the head of the waiting line until a server becomes free. This discipline was introduced by COBHAM, KESTEN and RUNNEBERG and has been studied as a multi-service system by W. WAGNER.

In the preemptive case a call of higher priority has the right to interrupt the service of a lower priority call. This discipline was introduced by WHITE and CHRISTIE, HEATHCOTE and TAKACS. The preemptive discipline can be divided in two cases, depending on the manner in which an interrupted call is served upon its reentry. In the preemptive resume case the interrupted call resumes service from the point where it was interrupted. In the preemptive repeat case the interrupted call starts its service anew.

Next, let us consider a finite number of sources for a system with r different priority classes.

We have to distinguish two different priority source models:

(1) The multiple finite-source model, in which the q sources are composed of r groupes, containing ${}_1q$, ${}_2q$,... until ${}_rq$ sources. The total number of sources is the sum of all ${}_iq$:

$$q = \sum_{i=1}^{r} {}_iq \qquad (8)$$

This model seems to be suitable for military networks. Another example for the application of this model are real-time computer systems, whose traffic-sources can be classified as follows: Programs of the operating system have highest priority, real-time jobs second priority and batch-jobs the lowest priority.

(2) The single finite-source model, in which each of the q traffic-sources can produce calls of each priority class. In this model, the priority of a call can be chosen by the subscribers themselves. The higher a priority is chosen the more expensive are the fees. This model is suitable for example for time-sharing systems with several terminals or for teletype networks. A call of class i shall be produced with probability ${}_ip$, so that the sum of all ${}_ip$ equals unity:

$$\sum_{i=1}^{r} {}_ip = 1 \qquad (9)$$

In this finite source priority model, the arrival of a call of class i affects the arrivals of other classes, while in the multiple finite-source priority model the arrivals of the different classes are independent. Consequently, the multiple case is easier to handle than the single case.

For $q \to \infty$ and ${}_iq \to \infty$ for all i, the two finite-source priority models approach the infinite source model.

Chapter II

THE PREEMPTIVE LOSS-DELAY-SYSTEM WITH A FINITE NUMBER OF SINGLE SOURCES

No investigations are known dealing with the complicated single source model and preemptive service discipline. In this chapter, we investigate this case for the preemptive resume discipline, where an interrupted call waits until its service can be continued. It follows from the negative exponential distribution of the service time that the remaining service time is distributed negative exponentially with the same mean as the total service time. This has an important consequence for the preemptive system, Since the remaining service time of an interrupted call and the total service time of an interrupting call have the same distribution, the interruption has no influence upon the termination process: With respect to the service time, two equivalent calls are exchanged. Consequently, we need not distinguish non-interrupted and interrupted calls.

In order to describe the state of the preemptive loss-delay-system we consider the class $\leq i$, which is the union of the classes 1,2,... until i. The state variable ${}_{\leq i}X(t)$ gives the number of calls of class $\leq i$ staying in the system at time t. Thus, the total number of calls in the system is given by ${}_{\leq r}X(t)$.

In the infinite case, the following basical sentence can be proved: The state variable ${}_{\leq i}X(t)$ behaves as if the calls of the classes i+1, i+2,... until r were non-existent . Consequently, ${}_{\leq i}X(t)$ behaves exactly like the state variable X(t) of the priorityless system, to which only the calls of class $\leq i$ are offered. Specificly, the total state process ${}_{\leq r}X(t)$ of the preemptive system behaves in the infinite case as if the classification into priority classes did not exist.

However, in the finite case the arrival process depends on the total number of calls in the system that means on the total state variable ${}_{\leq r}X(t)$. We combine the two state variables ${}_{\leq i}X(t)$

and $_{\leq r}X(t)$ and investigate the probabilities of state $_{\leq i}P_{k,j}(t)$:

$$_{\leq i}P_{k,j}(t) = P\{ {}_{\leq i}X(t)=k \ , \ {}_{\leq r}X(t)=j\} \qquad (10)$$

In the above, $_{\leq i}P_{k,j}(t)$ is the probability that at time t k calls of class $\leq i$ are in the system at the same time as a total number of j calls. Considering the different events which can arise within the short time Δt, the following system of differential equations can be derived for the time-dependent probabilities of state $_{\leq i}P_{k,j}(t)$:

$j=0,1,\ldots,n+s-1 \qquad k=0,1,\ldots,j$

$$\begin{aligned} {}_{\leq i}P'_{k,j}(t) = {} & {}_{\leq i}P_{k,j-1}(t)\cdot {}_{>i}\lambda_{j-1} + {}_{\leq i}P_{k-1,j-1}(t)\cdot {}_{\leq i}\lambda_{j-1} + \\ & + {}_{\leq i}P_{k,j+1}(t)\cdot(\mu_{j+1}-\mu_k) + {}_{\leq i}P_{k+1,j+1}(t)\cdot\mu_{k+1} + \\ & - {}_{\leq i}P_{k,j}(t)\cdot(\lambda_j+\mu_j) \end{aligned}$$

$j=n+s \qquad k=0,1,\ldots,j$

$$\begin{aligned} {}_{\leq i}P'_{k,j}(t) = {} & {}_{\leq i}P_{k,j-1}(t)\cdot {}_{>i}\lambda_{j-1} + {}_{\leq i}P_{k-1,j-1}(t)\cdot {}_{\leq i}\lambda_{j-1} + \\ & + {}_{\leq i}P_{k-1,j}(t)\cdot {}_{\leq i}\lambda_j - {}_{\leq i}P_{k,j}(t)\cdot({}_{\leq i}\lambda_j+\mu_j) \end{aligned}$$

$j=k=n+s$

$$\begin{aligned} {}_{\leq i}P'_{k,j}(t) = {} & {}_{\leq i}P_{k-1,j-1}(t)\cdot {}_{\leq i}\lambda_{j-1} + {}_{\leq i}P_{k-1,j}(t)\cdot {}_{\leq i}\lambda_j + \\ & - {}_{\leq i}P_{j,j}(t)\cdot\mu_j \end{aligned}$$

The sum of all probabilities of state equals unity:

$$\sum_{j=0}^{n+s}\sum_{k=0}^{j} {}_{\leq i}P_{k,j}(t) = 1$$

When we sum up the above equations for fixed j over k=1,2,... until j, then we get a system of differential equations for the probabilities that a total of j calls are in the system at time t. The resulting differential system is identical with the differential system for the probabilities of state in the case of the priorityless system. The fact of the identical systems leads to an important result: In case of finite single priority sources, the state process $_{\leq i}X(t)$ of class $\leq i$ does not behave independently of the classes i+1, i+2,... until r as it does in the infinite case, but the

total state process $_{\leq r}X(t)$ behaves as if the classification into priority classes did not exist. This sentence is in agreement with the general sentence in case of the infinite model.

As a consequence, the mean total arrival rate $_{\leq r}\lambda$ of the preemptive system is identical with the mean arrival rate λ of the priorityless system.

Substituting the stationary condition

$$_{\leq i}P_{k,j}(t) = {}_{\leq i}P_{k,j} \qquad (13)$$

into the differential system, this changes into a linear system of equations with rank $\frac{1}{2}\cdot(n+s+1)\cdot(n+s+2)-1$. For this linear system I derived two different algorithms which allow an efficient numerical evaluation.

In the following final portion of my paper several numerical results will be given. In these examples, the traffic characteristics are shown as a function of the total mean arrival rate $_{\leq r}\lambda$. The greater this mean arrival rate, the more calls are offered to the system.

For all examples I have chosen a loss-delay-system with n=2 servers, s=3 waiting-places and r=2 priority classes. The probability of class 1 is equal to $\frac{1}{3}$, and consequently the probability of class 2 is equal to $\frac{2}{3}$. The mean service time is h=1.

Figure 1 shows the infinite case. Obviously the total loss increases with increasing arrival rate. The loss $_1B$ of the higher priority class is smaller than the total loss B and the loss $_2B$ of the calls of lower priority is greater than B. The probability $_2V$ that a call of class 2 is displaced increases at first with increasing arrival rate. Then, after having reached a maximum, $_2V$ decreases for high values of $_{\leq r}\lambda$. This can be explained as follows:

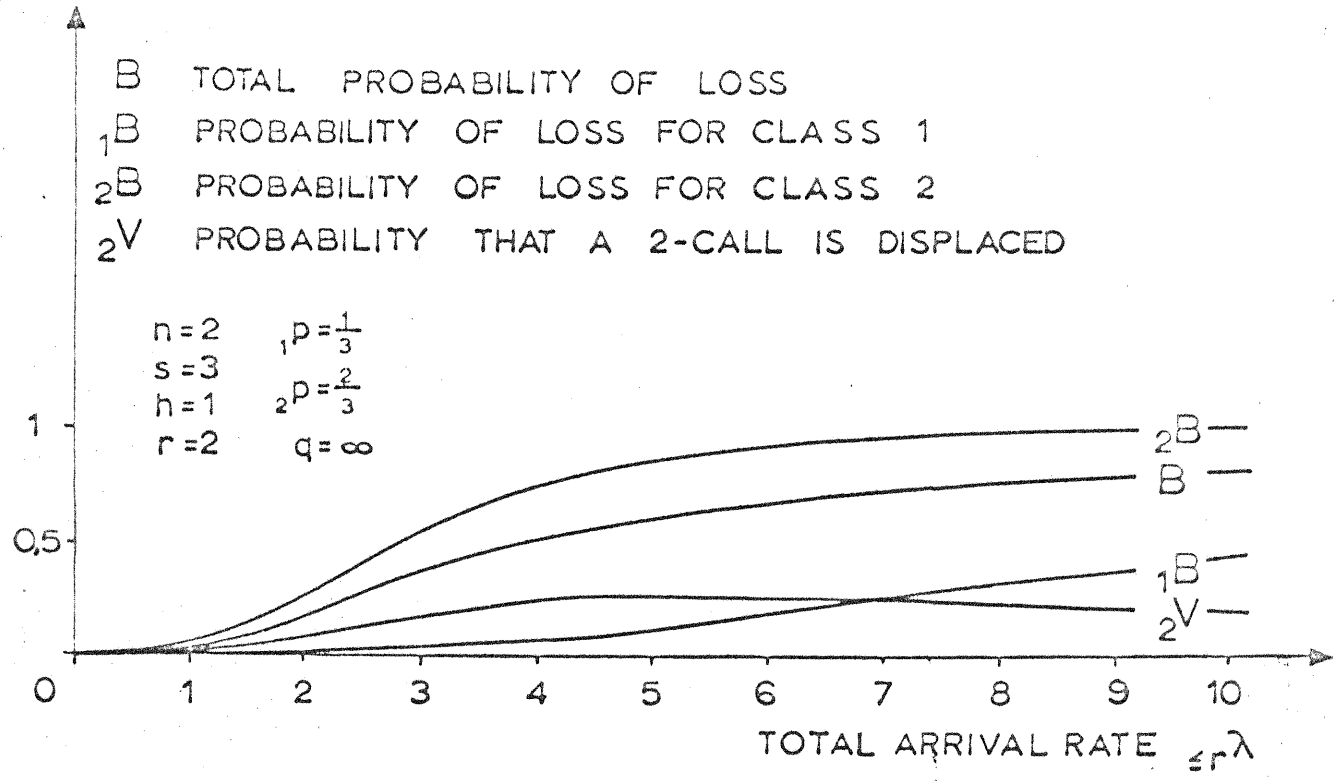

Figure 1: Some traffic characteristics as a function of the total arrival rate $_{\leq r}\lambda$ in the infinite case.

In case of a high arrival rate the system is occupied by calls of class 1. Therefore, only few calls enter the system. Since only few calls of class 2 are present in the system, only few calls of class 2 can be displaced.

The critical situations where losses and displacements are possible, are those with a great number of busy sources. For the finite model we pointed out in chapter I that the arrival rates of these states are less than the mean arrival rate, which in the infinite case is valid for all states j. Therefore, the probability of loss and of displacement must in the finite case be less than in the infinite case. This consideration proves to be true in figure 2 and 3. In these figures the ordinates are subdivided logarithmically.
Figure 2 shows the probability of loss for class 1. The lowest curve corresponds to q=7 sources. With an increasing number of sources the probability of loss for class 1 increases until it

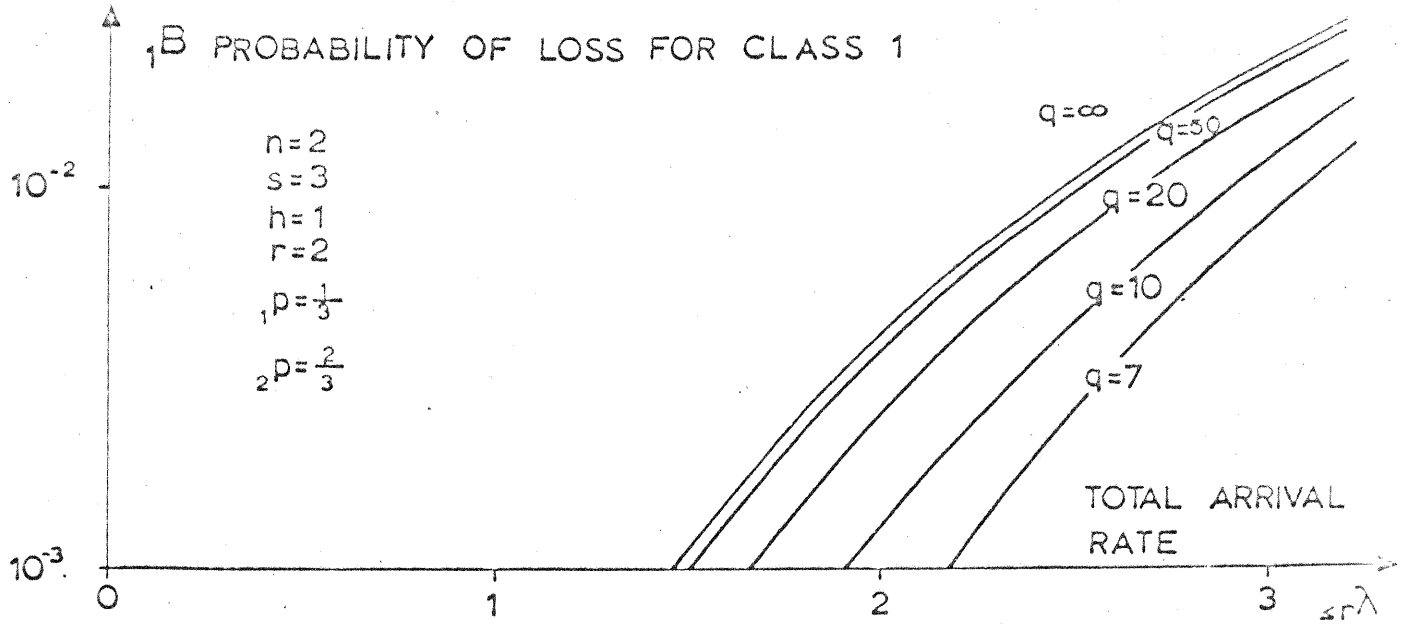

Figure 2: The probability of loss for class 1 as a function of the total arrival rate $_{\leq r}\lambda$

approaches the values for q=∞, namely the infinite source model.

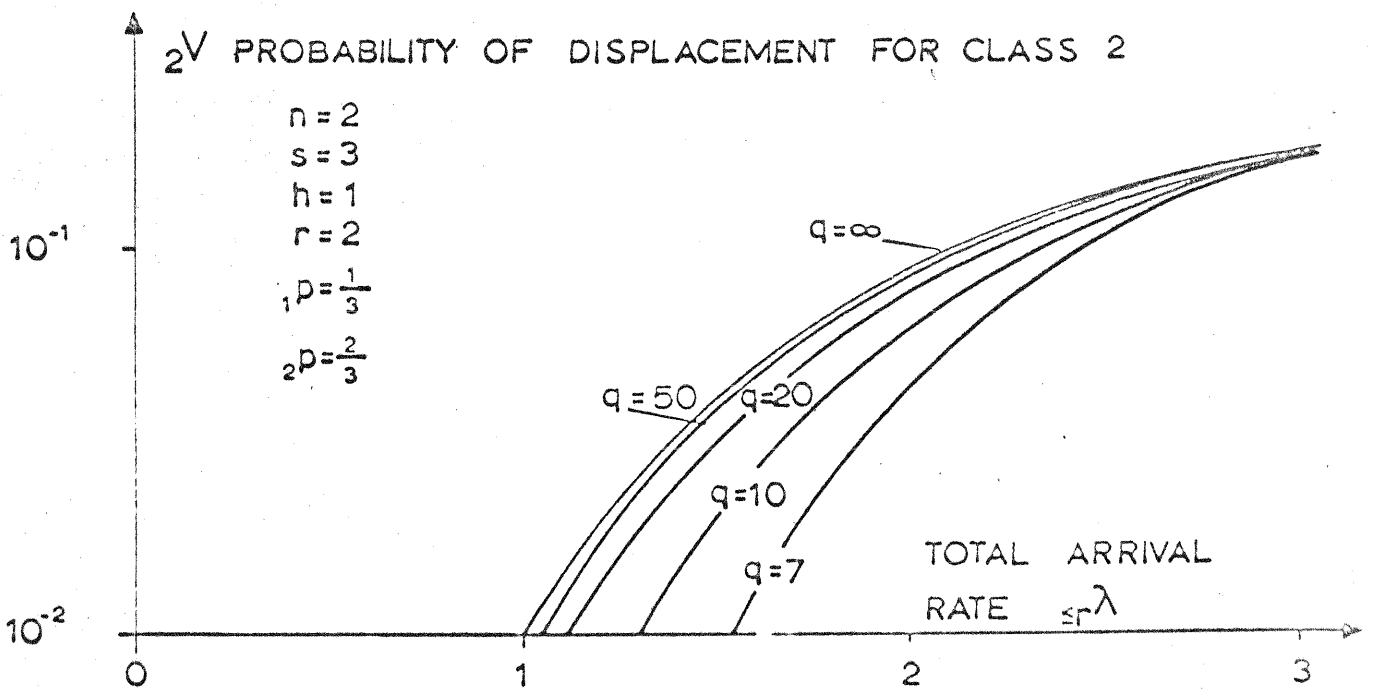

Figure 3: The probability of displacement for class 2 as a function of the total arrival rate $_{\leq r}\lambda$

The same behaviour is valid for the probability $_2V$ that a call of class 2 is displaced, as it is shown in figure 3.

Finally, for the same loss-delay-system, the behaviour of the traffic carried by the servers shall be discussed as a function of the mean total arrival rate. The traffic carried by the servers of class i is the mean number of servers occupied by class i.

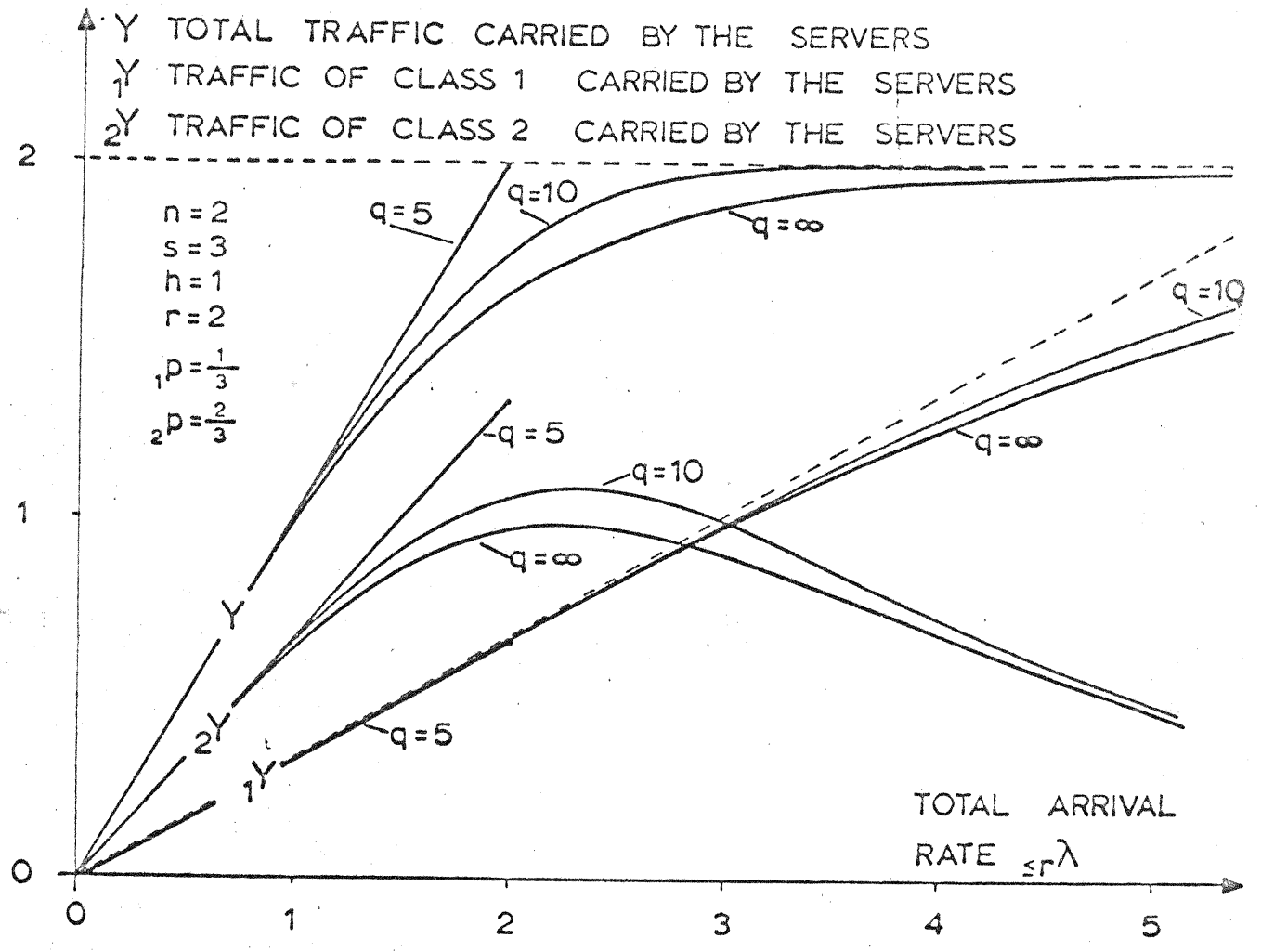

Figure 4: The traffic carried by the servers as a function of the total arrival rate $_{\leq r}\lambda$

Figure 4 shows that in the infinite case the total traffic Y increases with increasing arrival rate until it approaches the limiting value 2. The traffic $_1Y$ of class 1 does not only increase with increasing mean arrival rate, but its part of the total traffic Y increases also so that for higher arrival rates class 1 displaces class 2 from the system. Consequently, the traffic $_2Y$ of class 2 decreases after having reached a maximum.

On the other hand we consider the system, when the number of sources is equal to the number of devices, in this case q=n+s=5. When all q=n+s calls are busy the system is occupied but no further calls can arrive. Thus, we have no losses. In this case of a pure delay system the dependence of the carried traffic is represented by the straight lines corresponding to the total traffic Y and the traffics of class 1 or 2, respectively.

For the loss-delay-system with finite q>n+s the functions are between the extremes q=∞ and q=n+s as shown for a number of q=10 sources in the figure 4.

With these examples I conclude my introduction into the broad field of finite source priority queuing systems.

REFERENCES

G.J.BRANDT Das preemptive Warteverlustsystem
Ph.D.Thesis, University of Stuttgart

N.K.JAISWAL Priority Queues
Academic Press

W.WAGNER Ober ein kombiniertes Warte-Verlust-System mit Prioritäten
Ph.D.Thesis, 1968, University of Stuttgart