Proceedings of 5th Open Workshop on High Speed Networks, March 1996, Paris, pp. 7.11-7.18.

Performance Evaluation of Saturated Priority Systems with Feedback

Stefan Bodamer

University of Stuttgart
Institute of Communications Switching and Data Technics
Seidenstraße 36
D-70174 Stuttgart
email: bodamer@ind.uni-stuttgart.de

Abstract

An important performance aspect of modern switching systems and signalling networks is the response time to user requests. The overall system consists of nodes communicating by exchanging messages. In each node incoming messages pass a series of service phases with possibly different priorities. Therefore a node can approximately be modelled as an M/G/1 priority system with feedback. The mean transfer times of different messages are a useful measure to describe the system behaviour.

In this paper, a generic model is presented including batch arrivals, random branches, and forks. Furthermore different preemption and queueing disciplines can be defined. The model is evaluated with regard to phase utilization and mean transfer time through a single phase or a chain of phases. The analysis uses an efficient and easy to implement algorithm based on the method of moments. Interesting effects occur if the system is saturated. In this case the offered load will become greater than 1 and only higher priority phases are passed in finite transfer times.

1. Introduction

Modern telecommunication networks can be characterized as large and complex distributed systems. They consist of various nodes exchanging signalling information via a signalling network. Within a node the information generally passes a protocol stack which might be implemented as a couple of software processes running on a single processing unit. These processes again interact by exchanging messages. In order to minimize the time to user requests, e.g. the call setup delay, it is necessary to evaluate the performance of the network.

One approach to get the network performance is to investigate the network as a whole. However, the model then becomes rather complex. Exact analysis of this complex model is not feasible at all while on the other hand simulation often causes unacceptable run times.

If the network consists of loosely coupled nodes decomposition and aggregation techniques have evolved to be a promising way for performance modelling ([1], [19]). In this approach the information stream between nodes is supposed to approximately have the Markov property. Each node is then represented by an M/G/1 queueing system which can be solved separately. The arrival rates to the subsystem can be obtained by a message flow analysis for the network. After the subsystems have been evaluated aggregation of the subsystem delays yields the total response time to a user request introduced by the network. The processing of messages in the subsystem is expressed by so-called message chains composed of several phases with possibly different priorities and service times. So we have an M/G/1 priority system with feedback as the model for a node.

An introduction to the analysis of M/G/1 systems without priorities can be found in [12]. Takagi shows in [18] how response time distributions for M/G/1 priority systems without feedback can be obtained. Response time distributions for priority systems with feedback have been investigated among others by Enns ([5], [6]) and Fontana ([7], [8], [9], [10]). However, they had to make some restrictions which are not valid for the model needed here.

While analysis of an M/G/1 priority system with regard to transfer time distribution functions is still a rather complex task the method of moments introduced by Cobham in [3] is an alternative approach. The goal here is to calculate only the first moment (i.e. the mean value) of response times. This method has been applied to feedback priority systems by Simon ([17]) and Paterok ([15], [16]). In this paper, the method of moments will be extended to models with different queueing and preemption strategies. Moreover the influence of saturation, i.e. an offered load greater than 1, on system behaviour will be considered (Section 4). The exact model will first be described in Section 2. The analysis of the generic model by the method of moments will be shown in Section 3. In Section 5 some results of an example case study will be presented.

2. Model Description

In the following a generic M/G/1 model consisting of a single server with infinite queueing capacity will be presented. This abstract model covers the most important cases necessary for modelling communication subsystems.

Jobs arriving to the system are served in several phases forming a message chain. The stochastic arrival processes are Poissonian. Batch arrivals are also possible. Arrival rates and batch sizes can be individually defined for each message chain. Within a chain message flow can be influenced by random branch and fork elements (Fig. 1).

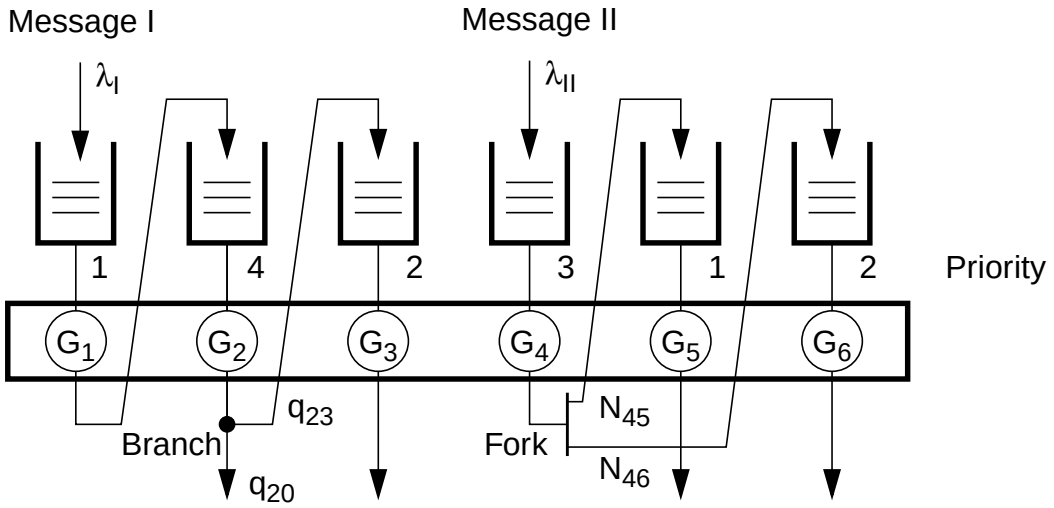

Figure 1: Example of a system with two message chains

If the server is busy on the arrival of a job the job is put into the queue. The sequence of the jobs in the queue is determined by the priority values of the jobs. On arrival or feedback a job changes its priority according to the priority value associated with the subsequent phase. Among jobs with the same priority the queueing discipline controls which job is to be serviced next. Considered queueing disciplines are first-come-first-serve (FCFS), last-come-first-serve (LCFS), and random order of service.

The service times of the various phases are distributed according to GI distribution functions. For non-preemptive systems the mean service times and the coefficients of variation are sufficient to describe the stochastic service processes.

The preemption behaviour of the system is defined by the global preemption distance (see [11]). A job arriving in phase A will interrupt a job currently served in phase B if the difference of the priority values of A and B is greater than or equal to the preemption distance. The preemption strategy associated with the corresponding phase determines how an interrupted job behaves. If the preemption strategy is “preemptive resume” service continues at the point where it has been interrupted when the job re-enters service. The strategy “preemptive repeat” defines that the work that has already been done on an interrupted job is lost. In this case the service time on re-entry to the server may be resampled according to the service time distribution (“preemptive repeat with resampling”) or it may be the same as in the first service attempt (“preemptive repeat without resampling”).

3. Analysis

In order to get the mean values for transfer and waiting times in a model derived from the generic model described in section 2 the method of moments is applied. This approach uses three fundamental laws of traffic theory:

- Little’ law [14] expressing that the mean value of the number of customers N in an arbitrary system depends in a linear way on the arrival rate λ and the expectation value of the sojourn time T_S in the system:

$$E[N] = \lambda \cdot E[T_S] \tag{1}$$

- the PASTA theorem [20] which proclaims that a customer arriving in a Poisson stream sees the system in its average state
- the result from renewal theory [13] for the expectation value of the residual life time T_R of a customer whose service time T has known mean value and variance:

$$E[T_R] = \frac{1}{2} \cdot E[T] + \frac{Var[T]}{2 \cdot E[T]} \tag{2}$$

The basic principle of the method of moments is to tag a random customer and to follow his way through the system. For this customer the mean delay is composed of some partial delays he suffers from different classes of jobs. In the case of a non-preemptive system without feedback the delay of a tagged job until his service is finished arises from three classes of customers:

- the customer currently served when the tagged job arrives
- customers with higher or equal priority already being in the queue upon the tagged job’s arrival
- customers arriving while the tagged job is in the queue and overtaking him because of higher priority

The partial delays can be calculated by usage of the fundamental laws described above. Summarizing the partial delays for the tagged customer leads to a linear equation system for the mean transfer times through the phases which can easily been solved.

For the simple M/G/1 queueing system with only one phase we have to solve a single equation:

$$w = \rho \cdot t_R + \Omega \cdot h = \rho \cdot \frac{h}{2} \cdot (1 + c^2) + \lambda \cdot w \cdot h \tag{3}$$

where w is the mean waiting time, h is the mean service time, Ω is the mean queue length, t_R is the mean residual service time, c is the coefficient of variation of the service time, and $\rho = \lambda \cdot h$ is the mean utilization of the system. Solving this yields Pollaczek-Khintchine's well-known formula for the mean waiting time:

$$w = \frac{1 + c^2}{2} \cdot \frac{\rho}{1 - \rho} \cdot h \tag{4}$$

In a system with feedback things become more difficult. If the tagged job re-enters the system in a feedback the PASTA theorem can not be applied because the feedback stream does not necessarily have the Markov property. In this case the re-entering customer does not see the system in its global mean state. Therefore the mean state at feedback must be sequentially derived from the mean state at arrival which is identical to the global mean state. In [16] and [17] this is done by complex mathematical formulas while we used an efficient and easy to implement algorithm that directly imitates priority queueing within the system (see [2]). This algorithm is iteratively applied to all phases within a chain. First it provides the calculation of the delay equation for the current phase. Moreover the algorithm maintains the mean system state at departure of the phase. This state at departure can then be interpreted as the state at arrival when the algorithm is applied to the following phase.

For preemptive systems the decomposition into job categories has to be refined. Furthermore some auxiliary quantities have to be computed for each phase like the probability for not being interrupted or the expected residual service time. Those quantities have already been derived in [4] and [16] for preemptive resume and preemptive repeat with resampling while for preemptive repeat without resampling they can be found in [2].

4. Saturated Systems

The system becomes saturated if the total offered load is greater than 1. In this case the number of customers entering the system within a time period Δt is greater than the number of customers served within Δt. Therefore the number of jobs in the system increases so that stationarity is no longer given. The system gets instable.

Regarding a saturated priority system one can establish that in a certain range of load the number of jobs waiting for service in higher priority phases still shows a stationary behaviour while it increases to infinity in lower priority phases. In other words, lower priority phases get that share of utilization that has been left by higher priority phases. Therefore the output rate $\lambda_{out,\,i}$ of a lower priority phase i is reduced by a factor x_i compared to the input rate $\lambda_{in,\,i}$:

$$x_i = \frac{\lambda_{out,i}}{\lambda_{in,i}} \tag{5}$$

This "reduction factor" can be used to introduce the so-called "critical" priority and to distinguish three phase classes according to their behaviour at a certain load:

- Phases with a priority greater than the critical priority have a reduction factor of 1. This means that there is no reduction of data flow in these phases. The mean values of partial queue length and transfer time are finite for these phases.

- The reduction factor for phases with a priority value equal to the critical priority takes on values between 0 and 1. Customers are still served in these phases but the service rate is lower than the offered rate. The mean values of partial queue length and transfer time go to infinity.
- Phases with a priority lower than the critical priority have a reduction factor of 0, i.e. no jobs are served in those phases. Here the mean value of the partial queue length also goes to infinity while transfer time is undefined. All phases following that phase within a chain have input rates of 0.

For systems with FCFS or random order queueing discipline it can be shown that all phases whose priorities are equal to the critical priority have the same reduction factor. Within a message chain the equation

$$\lambda_{in,i} = q \cdot \lambda_{out,j} \tag{6}$$

holds where j is the predecessor phase of i. The factor q is equal to unity unless there is a branch or a fork at the transition from phase i to phase j.

Now let n be the total number of phases in the system and k the number of phases with critical priority. Applying the previous results and considering the possibility of more than one phase with critical priority being part of the same chain one can state that

$$\lambda_{out,i} \sim x^{b_i} \qquad \text{with } 0 \le b_i \le k,\ 1 \le i \le n. \tag{7}$$

In this formula b_i denotes the number of preceding phases with critical priority in the chain containing phase i including phase i if its priority value is critical. With h_i being the mean service time of phase i we then have the phase utilization

$$\rho_i = \lambda_{out,i} \cdot h_i = a_i \cdot x^{b_i}. \tag{8}$$

Finally the sum of the utilization factors of all n phases in the system must equal unity in the saturated case. So the definition equation for the reduction factor x is a non-linear equation of the form

$$\sum_{i=1}^{n} a_i \cdot x^{b_i} = 1 \qquad \text{with } 0 \le a_i \le 1,\ 0 \le b_i \le k \le n \tag{9}$$

which can be solved by numerical approximation for higher degrees. After x has been calculated all flow rates and utilization factors can be computed. The transfer time analysis is then performed using the method of moments as described in Section 3. The only influence of saturation is now that merely phases with a finite transfer time according to the previous classification have to be considered.

5. Example

As an example the results for the system depicted in Fig. 2 will now be presented. The system consists of three chains which are composed of phases with priority values 1, 2, and 3. If we choose the global preemption distance to have a value greater than 2 we get a non-preemptive system. All phases in the system have deterministic service times with mean values $h_1 = h_4 = h_5 = h_6 = h_7 = 1$ and $h_2 = h_3 = 5$. The arrival rates are variable maintaining a constant ratio $\lambda_1 \div \lambda_2 \div \lambda_3 = 3 \div 2 \div 4$. Batch arrivals are not considered in this example.

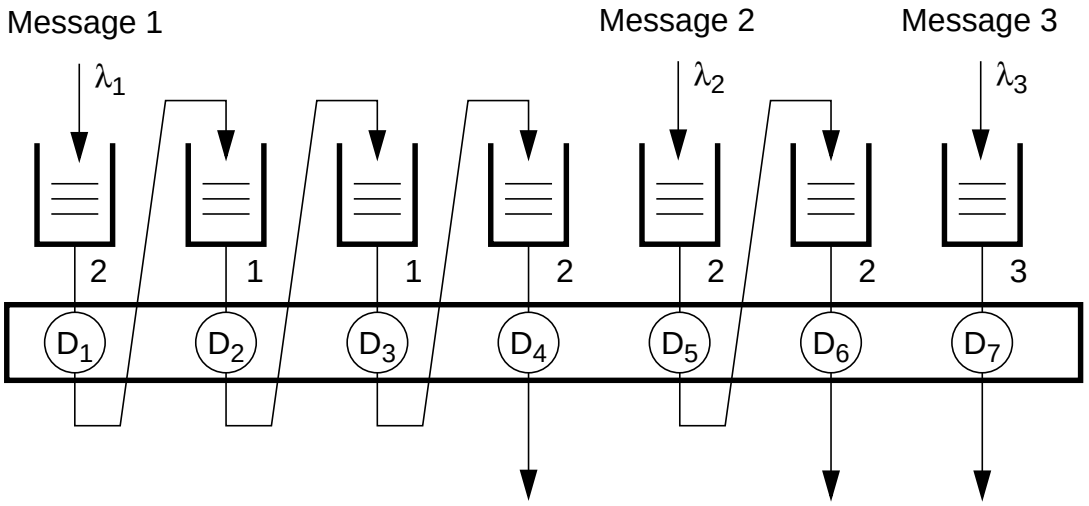

Figure 2: System with three message chains

The diagrams for the chain utilizations are depicted in Fig. 3. One can recognize that the utilization of chain 1 containing phases with the least priority 1 is reduced first. Although phase 4 has a priority value greater than the critical priority at this point it suffers from the throttling in the preceding phases. On the other hand, utilization of chain 2 is not affected by the collapse of chain 1. Utilization of this chain increases up to the load where 2 becomes the critical priority. In the following load range we have a non-linear degradation of chain utilization in chain 1 and 2. This is caused by the fact that both chains contain phases with the critical priority related to the corresponding load range.

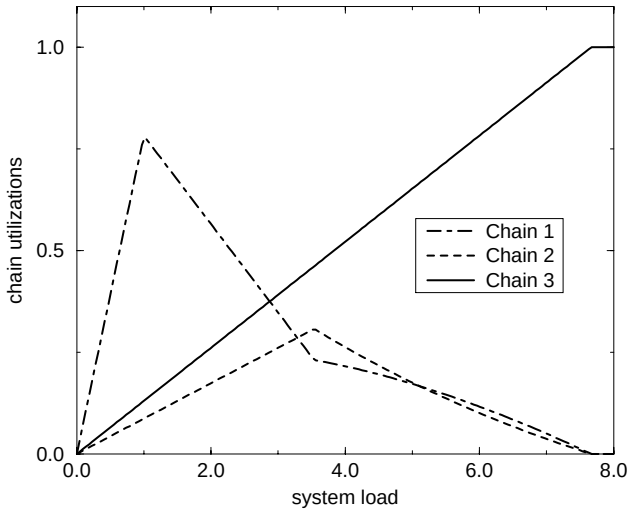

Figure 3: Chain utilizations

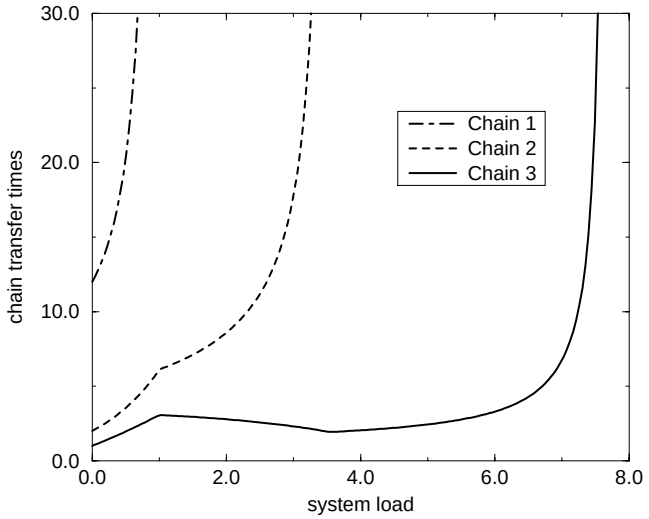

Figure 4: Chain transfer times

When we look at the transfer times (Fig. 4) we find that especially the curve for the transfer time through chain 3 is interesting. This chain consists of a single phase with highest priority. The diagram shows that there are sharp bends at the points where lower priority phases reach saturation. In the load range where 1 is the critical priority we even have a decreasing transfer time in chain 3. This is because the partial delay of a chain 3 job, which is caused by the residual service times of phases 2 and 3, decreases when lower priority utilization goes down.

6. Conclusions

When using decomposition and aggregation techniques for the performance evaluation of communication networks problems are often reduced to the evaluation of M/G/1 priority systems with feedback. A generic model for that kind of systems has been presented. The model considers that messages are processed sequentially in phases with different properties.

The M/G/1 priority model with feedback can be evaluated by the method of moments. Based on a few fundamental laws this technique divides the total mean delay of a tagged job into partial delays which can be calculated separately. Using an efficient algorithm the mean value analysis of transfer times can easily be implemented in a tool.

When load is increased priority systems show an interesting behaviour. Higher priority phases still can have a finite transfer time while stationarity is no longer given for lower priority phases. It has been shown in an example case study that this may even lead to a decreasing transfer time for higher priority phases.

Acknowledgement

The results presented in this paper are the output of the author’s diploma thesis. This thesis was guided by M. Schopp and M. Bafutto whom the author would like to thank for their advice and some basic ideas.

References

[1] M. Bafutto, P. J. Kühn, G. Willmann, “Capacity and Performance Analysis of Signaling Networks in Multivendor Environments”, *IEEE Journal on Selected Areas in Communication*, Vol. 12, No. 3, pp. 490-500, Apr. 1994

[2] S. Bodamer, *Mittlere Durchlaufzeiten in gesättigten M/G/1-Prioritätensystemen mit Rückkopplungen*, Diploma Thesis, Institute of Communications Switching and Data Technics, University of Stuttgart, 1994

[3] A. Cobham, “Priority Assignments in Waiting Line Problems”, *Operations Research*, Vol. 2, pp. 70-76, 1954

[4] R. W. Conway, W. L. Mexell, L. W. Miller, *Theory of Scheduling*, Addison-Wesley, Reading, Mass., 1967

[5] E. G. Enns, “Some Waiting Time Distributions for Queues with Multiple Feedback and Priorities”, *Operations Research*, Vol. 17, pp. 519-525, 1969

[6] E. G. Enns, “A Simple Method of Calculating Queue Size Distributions of Priority Queueing Systems with Feedback”, *Proc. of the 6th International Teletraffic Congress (ITC)*, Munich, Sep. 1970, paper 314

[7] B. Fontana, “Queue with Two Priorities and Feedback: Joint Queue-Length Distribution and Response Time Distributions for Specific Sequences”, *Proc. of the 10th International Teletraffic Congress (ITC)*, Montreal, June 1983, paper 4.1-8

[8] B. Fontana, C. D. Bersoza, “Stationary Queue-Length Distributions in an M/G/1 Queue with Two Non-Preemptive Priorities and General Feedback”, *Proc. of the Second International Symposium on Performance of Computer Communication Systems*, Zürich, Mar. 1984

[9] B. Fontana, C. D. Bersoza, “M/G/1 Queue with Two Non-Preemptive Priorities and Feedback: Response Time Distributions for any Particular Sequence”, *Proc. of the 3rd International Seminar on Teletraffic Theory*, Moscow, June 1984, paper 16, pp. 113-116

[10] B. Fontana, C. D. Bersoza, “M/G/1 Queue with N-Priorities and Feedback: Joint Queue-Length Distributions and Response Time Distributions for any Particular Sequence”, *Proc. of the 11th International Teletraffic Congress (ITC)*, Kyoto, 1985, paper 3.3A-4.1

[11] U. Herzog, *Verkehrsfluß in Datennetzen*, Habilitationsschrift, Institute of Communications Switching and Data Technics, University of Stuttgart, 1973

[12] L. Kleinrock, *Queueing Systems Volume I: Theory*, John Wiley & Sons, New York et al., 1975

[13] L. Kleinrock, *Queueing Systems Volume II: Computer Applications*, John Wiley & Sons, New York et al., 1976

[14] J. D. C. Little, “A Proof of the Queueing Formula L=λW”, *Operations Research*, Vol. 9, pp. 383-387, 1961

[15] M. Paterok, O. Fischer, “Feedback Queues with Preemption-Distance Priorities”, *ACM SIGMETRICS Performance Evaluation Review*, Vol. 17, No. 1, pp. 136-145, May 1989

[16] M. Paterok, *Warteschlangensysteme mit Rückkopplungen und Prioritäten*, Arbeitsberichte des Instituts für Mathematische Maschinen und Datenverarbeitung, vol. 23, no. 12, Friedrich Alexander Universität Erlangen Nürnberg, Germany, Oct. 1990

[17] B. Simon, “Priority Queues with Feedback”, *Journal of the ACM*, Vol. 31, pp. 134-149, 1984

[18] H. Takagi, *Queueing Analysis Volume I: Vacation and Priority Systems Part I*, North-Holland, Amsterdam et al., 1991

[19] G. Willmann, P. J. Kühn, “Performance Modeling of Signaling System No. 7”, *IEEE Communications Magazine*, Vol. 28, No.7, pp. 44-56, 1990

[20] R. W. Wolff, “Poisson Arrivals see Time Averages”, *Operations Research*, Vol. 20, pp. 223-231, 1982