A mean-value approach for M/G/1 priority queues

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics and Computing Science

Memorandum COSOR 83 - 09

A mean-value approach for M/G/1 priority queues

by

Jan van Doremalen

Eindhoven, the Netherlands

March 1983

A MEAN-VALUE APPROACH FOR M/G/1 PRIORITY QUEUES

by

Jan van Doremalen

Abstract.

This note deals with a mean-value approach for M/G/1 priority queues. Using the residual life-time formula, Little's formula and the fact that Poisson arrivals see time averages, we derive schemes to evaluate mean response times, mean queue lengths and mean waiting times for the respective priority classes.

1. Introduction

This note deals with a queueing system, where R independent Poisson arrival streams with rates λ_r, $r = 1,2,\ldots,R$, are to be served by a single server infinite capacity queue. The service times are independent and distributed according to distribution functions G_r for stream r, $r = 1,2,\ldots,R$. The mean w_r and the second moment m_r of G_r are assumed to be finite. The service discipline is first-come first-served but for priorities. Two priority rules, the preemptive resume and the head-of-the-line priorities, will be discussed.

It is our purpose to show an elegant derivation of schemes to compute mean system times, mean queue lengths and mean waiting times for customers of the successive streams. These schemes will be based on the following three important results:

(i) the PASTA-property, i.e. Poisson arrivals see time averages
(ii) Little's formula
(iii) the expected residual life-time formula.

The schemes are not new and can be found for instance in Takacs [1964] and Wolff [1970]. It is the elegant and exact way of deriving them which is of interest.

2. Some notations

The following notations will be used

S_r : mean response time of a stream r customer (an r-customer).

D_r : mean waiting time for an r-customer, i.e. the mean time between arrival moment and the moment service starts for the first time.

C_r : mean completion time for an r-customer, i.e. the mean time until service completion from the moment service starts for the first time.

Q_r : mean number of waiting r-customers.

L_r : mean number of r-customers in the service completion phase.

ρ_r : $\lambda_r w_r$, the utilization factor of the server for r-customers.

We note that the concept of service completion time can be found in Gaver [1961]. Furthermore, it should be noted that L_r in fact gives the probability that there is some r-customer in the service completion phase, since L_r is the mean of a random variable which can only take on the values 0 and 1.

3. A single server queue with preemptive resume priorities

A queueing system with R independent Poisson arrival streams will be analyzed. An r_1-customer has a higher priority than an r_2-customer if $r_1 < r_2$. A customer interrupts at his arrival the service of a lower priority customer. The service of this lower priority customer is resumed at the moment there are no higher priority customers left in the system.

First step in the analysis is the evaluation of the mean service completion time C_r of an r-customer, $r = 1,2,\dots,R$. His "effective" service time is influenced by interrupts of higher priority customers. If we match the mean service completion time with the total amount of work to be done during this completion time, we find

$$C_r = w_r + \sum_{i=1}^{r-1} \lambda_i C_r w_i , \qquad r = 1,2,\dots,R ,$$

where we use the Poisson character of the arrival streams, so

$$(1) \qquad C_r = \frac{w_r}{1 - \sum_{i=1}^{r-1} \rho_i} .$$

Wolff [1982] showed under rather general assumptions that customers arriving according to a Poisson process, see the system as if in time-equilibrium, the so-called PASTA-property.

A first consequence is that if a customer finds upon arrival an r-customer in his service completion phase, the remaining "effective" work to be done for the r-customer is given by the expected residual life-time formula, namely

$$\frac{m_r}{2w_r} .$$

Note that the independence of service time and interrupts is essential in this reasoning.

Another consequence of the PASTA-property is that an arriving customer sees in the average Q_r r-customers having received no service yet. Furthermore, with probability L_r there will be an r-customer in the service completion phase.

We now are able to give a mean value relation for the mean waiting time D_r of an r-customer, matching the waiting time and the total amount of work to be done during the waiting time.

(2) $$D_r = \sum_{i=1}^{r} Q_i w_i + \sum_{i=1}^{r} L_i \frac{m_i}{2w_i} + \sum_{i=1}^{r-1} \lambda_i D_r w_i .$$

The last term on the RHS denotes the amount of higher priority work entering during the waiting time of the customer.

Little's formula (confer Little [1961]) gives the relation

(3) $$Q_r = \lambda_r D_r .$$

Furthermore, the fraction of time there is some r-customer in the service completion phase is given by $\lambda_r C_r$. As we have noted in Section 2 this fraction equals L_r also, and thus

(4) $$L_r = \lambda_r C_r .$$

It should be observed that this is a consequence of Little's formula also.

The mean response time S_r of r-customers, of course, is given by

(5) $$S_r = C_r + D_r .$$

Now, the equations (1) through (5) give a scheme to evaluate mean waiting times, mean queue lengths and mean response times for the R streams.

We note that D_r is determined by

$$(6) \qquad D_r = \frac{\sum_{i=1}^{r} \rho_i \frac{m_i}{2w_i}}{\left(1 - \sum_{i=1}^{r-1} \rho_i\right)\left(1 - \sum_{i=1}^{r} \rho_i\right)}$$

a result corresponding with results of for instance Takacs [1964: Formula 66] and Wolff [1970: Formula 31] .

4. A single server queue with head-of-the-line priorities

The system is as in Section 3 but for the fact that a customer does not interrupt the service of a lower priority customer. Now service completion time and "effective" service time coĭncide and we have

$$(7) \qquad C_r = w_r .$$

Using the same arguments as in Section 3 we find as a relation for the mean waiting time D_r of an r-customer,

$$(8) \qquad D_r = \sum_{i=1}^{r} Q_i w_i + \sum_{i=1}^{R} L_i \frac{m_i}{2w_i} + \sum_{i=1}^{r-1} \lambda_i D_r w_i$$

and again we have, with the relations $Q_r = \lambda_r D_r$, $S_r = C_r + D_r$ and $L_r = \rho_r$, $r = 1,2,\ldots,R$, a scheme to compute the relevant mean values.

We note that D_r is determined by

(9) $$D_r = \frac{\sum_{i=1}^{R} \rho_i \frac{m_i}{2w_i}}{\left(1 - \sum_{i=1}^{r-1} \rho_i\right)\left(1 - \sum_{i=1}^{r} \rho_i\right)}$$

a result which corresponds with the results of for instance Takacs [1964: Formula 68] and Wolff [1970: Formula 33].

5. Conclusions

We have derived in an elegant way relations between mean values in queueing systems with priorities. It should be observed that the way of reasoning can be extended to more complex situations, for example time-sharing systems.

The interest of the schemes also lies in the field of finding approximations in queueing networks where stations with priority rules are considered.

6. References

- Gaver, D.P. [1961]: "A Waiting Line with Interrupted Service including Priorities", J.R. Stat. Soc. B25 : 73 - 90.
- Little, J.D.C. [1961]: "A proof of the queueing formula: L = λW", OR 9 : 383 - 387.
- Takacs, L. [1964]: "Priority Queues", OR 12 : 63 - 74.
- Wolff, R.D. [1970]: "Work Conserving Queues", JAP 7 : 327 - 337.
- Wolff, R.D. [1982]: "Poisson Arrivals See Time Averages", OR 30 : 223 - 231.