QUEUEING SYSTEMS

Nicole Bäuerle
University of Ulm, Germany

Keywords: queueing discipline, service facilities, waiting time, virtual waiting time, Little's formula, Pollaczek-Khintchine formula, Erlang's loss system, multiclass queueing network, product form distribution.

Contents

1. Introduction
2. Design of Queueing Systems
2.1 Arrivals
2.2 Service Facilities
2.3 Queueing Discipline
3. Performance Measures and Special Queues
3.1 M/M/1 Queue
3.2 M/M/∞ Queue
3.3 M/M/m Queue
3.4 M/M/1/K Queue
3.5 Erlang's Loss System
3.6 M/GI/1 Queue
3.7 GI/M/1 Queue
4. Little's Formula
5. Queueing Networks and Examples
5.1 Jackson Networks
5.2 Kelly Networks
5.3 Re-entrant Lines
Glossary
Bibliography
Biographical Sketch

Summary

This paper summarizes the main results of classical queueing theory and some recent developments in multiclass queueing networks. In the first part, the different characteristics of a queueing system are discussed, such as arrival process, service facilities, queueing discipline, and service times.

The main focus of this paper lies on the long run behavior of queueing systems. In particular we investigate the $M/M/1$ queue, the $M/M/\infty$ queue, the $M/M/m$ queue, the $M/M/1/K$ queue, Erlang's loss system, the $M/GI/1$ queue and the $GI/M/1$ queue. We derive a stability condition for these queues and give the limiting distribution of the number of waiting jobs in the system or the limiting distribution of the waiting time. The important Little's Formula is explained and at the end we shortly introduce some special queueing networks, so-called open multiclass networks.

1. Introduction

Queueing systems are one of the oldest and most widely investigated topics of Markov models. The main evolution of this field has taken place in three waves. The first wave appeared around 1918 with the emerging telephony technology and the investigations of the Danish mathematician A.K. Erlang. In the seventies, fresh impetus has been given by problems that arise in computer systems. Last but not least, the investigation of re-entrant networks for semiconductor manufacturing plants brought some new aspects in the nineties. Therefore, typical applications of queueing systems are in the area of computer systems, data transmission, telecommunication and manufacturing systems. More recent applications include ISDN and the Internet. Queueing situations arise, when a limited capacity has to be shared by incoming jobs. For example in time-sharing computer systems jobs are created by users, which then compete for processing capacity, storage capacity and input/output facilities. Interesting performance measures of these systems are e.g. response or waiting times, server utilization and system throughput. In the following, we will look at some simple queueing systems and open multiclass queueing networks.

2. Design of Queueing Systems

Queueing systems can be distinguished according to the following criteria, which constitute the basic design of the queue.

2.1 Arrivals

Jobs can arrive one at a time or in *batches*, where the batch size can be random. We denote by τ_n the random *interarrival time* between job n and $n + 1$ (or batch n and $n + 1$). The jobs that arrive are taken from a pool of jobs, which can be finite or infinite.

2.2 Service Facilities

The *waiting room* can be finite or infinite. In a system with finite waiting room, arriving customers are blocked when all places of the waiting room are occupied. Blocked customers can get lost or will retry to enter the queue after a random time. In this case, we speak of a *retrial* queue. Queueing systems also differ in the number of servers available. An infinite number of servers are also allowed. Every arriving job will then get into service immediately.

2.3 Queueing Discipline

Jobs can be served one at a time or in batches. We denote by σ_n the random service time of job n (batch n). If jobs are served one at a time, there has to be an algorithm that determines the order in which the waiting jobs are served. The most common queueing disciplines are

FCFS First Come, First Served (or FIFO: First In, First Out). Jobs are served in the order of arrivals. This is the usual service discipline if nothing else is mentioned.

LCFS Last Come, First Served (or LIFO: Last In, First Out). Whenever the server has finished a job, he/she will continue with the latest arrived job.

SIRO Serve In Random Order. The next job to be served is picked randomly from the waiting ones.

PS Processor Sharing. The server devotes his/her capacity equally to the waiting jobs. i.e. if n jobs are waiting, each job receives $\frac{1}{n}$ of the server's capacity.

RR Round Robin or time-sharing. The server spends a fixed time Δ on one job (or less, when service completion occurs) and switches afterwards to the next waiting job.

If different job classes are served according to priorities given to these classes, we call a service *preemptive*, if the server switches instantaneously upon arrival to a new job of higher priority. Vice versa, a service is called *non-preemptive* if the server always finishes the service.

In what follows, we will restrict our exposition to *simple queueing systems*. Simple queueing systems are those, where arrival and service is one by a time and the sequences of interarrival times and service times are stationary. For simple queueing systems a shorthand notation of the form $\alpha/ß/m/n/K$ has been established, where

α gives the type of interarrival distribution.

$ß$ gives the type of service distribution.

m gives the number of servers.

n gives the size of the waiting room.

K gives the number of jobs in the system.

α and $ß$ can take for example the following values:

M exponential distribution (Markovian); the corresponding sequence of interarrival or service times consists of independent, exponentially distributed random variables.

D deterministic, fixed time.

G general distribution.

GI general distribution; however, the corresponding sequence consists of independent random variables.

E_k Erlang distribution with k phases.

PH phase-type distribution.

M^X batch arrivals with exponential interarrival times and batch size distribution as the random variable X.

Often the notation is truncated to $\alpha/ß/m$ in which case the waiting room and the number of jobs are supposed to be infinite.

3. Performance Measures and Special Queues

Basic processes that describe the queueing system are e.g.

X_t	the number of jobs in the system at time t.
W_n	the waiting time of job n.
V_t	the virtual waiting time at time t. This is the sum of the (remaining) waiting times of all jobs which are in the system at time t.

In general, one is interested in the long-run behavior of queueing systems, i.e. we are interested in X_t, V_t and W_n, when t or n is large. Mathematically this means we have to determine the limit distributions of the stochastic processes. When the interarrival and service distributions are independent exponentially distributed, the preceding processes are Markov processes and we can apply the methods described in the article on *Markov Models* to identify limit distributions. We will outline this procedure in the next sections.

Other interesting performance measures are the *response time*, *throughput* and *server utilization*. The response time is defined as the time between arrival of a job and its departure and the throughput is the average number of jobs leaving the system per unit time. The utilization of servers U in the case of only one server is the probability that the server is busy. If there are m identical servers and r are busy on average, then $U = \frac{r}{m}$.

An important characteristic of queueing systems is the *traffic intensity*. In a $G/G/m$ queue the traffic intensity ρ is defined by

$$\rho = \frac{\mathbb{E}[\sigma_k]}{m\mathbb{E}[\tau_k]}. \qquad (1)$$

Hence, the traffic intensity is the ratio of expected service time and expected interarrival time. Loynes (1962) has shown for $G/G/1$ queues (with only stationary interarrival and service times) that the process (W_n) is stable if $\rho < 1$, which means that the sequence of distributions of W_n converges to a proper distribution as n tends to infinity. In what follows we will show for several Markovian queueing systems that if $\rho < 1$, i.e. the number of arrivals is less than the number of potential departures, the system is stable (positive recurrent). If $\rho > 1$ the system is unstable (transient) and if $\rho = 1$ it is null-recurrent.

3.1 M/M/1 Queue

Here we have one server and an unlimited waiting room. Arrivals occur according to a Poisson process with intensity λ and the service times are independent and exponentially

distributed with parameter μ. The traffic intensity in this model is $\rho = \frac{\lambda}{\mu}$ and the process (X_t) which gives the number of waiting jobs is a birth-and-death process with birth rates $\lambda_i \equiv \lambda$ and death rates $\mu_i = \mu \min\{1, i\}$, $i \in \mathbb{N}_0$. According to the birth-and-death process example described in *Markov Models*) we obtain that the $M/M/1$ queue is positive recurrent if and only if $\rho < 1$ and the limit distribution π of the number of waiting jobs is given by the following geometric distribution

$$\pi_i = (1-\rho)\rho^i, \quad i = 0, 1, 2, \ldots . \tag{2}$$

π_i is the probability that there are i jobs waiting in the long run. In particular, $\pi_0 = 1 - \rho$ is the probability that the server is idle. In other words, $U = \rho$ is here the server utilization.

The throughout in this system is equal to λ.

3.2 M/M/∞ Queue

In this case, we have infinitely many servers and each arriving job will immediately receive service. Arrivals occur according to a Poisson process with intensity λ and the service times at each server are independent and exponentially distributed with parameter μ. Let us denote $\eta = \frac{\lambda}{\mu}$ (note that the traffic intensity is here formally 0). The process (X_t) which gives the number of waiting jobs is a birth-and-death process with birth rates $\lambda_i \equiv \lambda$, and death rates $\mu_i = i\mu$, $i = 0, 1, \ldots,$. According to the birth-and-death process example described in Markov Models we obtain that the $M/M/\infty$ queue is always positive recurrent and the limit distribution π of the number of waiting jobs is given by the following Poisson distribution

$$\pi_i = e^{-\eta}\frac{\eta^i}{i!}, \quad i = 0, 1, 2, \ldots . \tag{3}$$

-
-
-

Bibliography

Asmussen S. (1987). *Applied Probability and Queues*, 318 pp. Chichester: John Wiley & Sons. [Presents a comprehensive treatment of queueing systems.]

Buzacott J.A. and Shanthikumar J.G. (1993). *Stochastic Models of Manufacturing Systems*, 553 pp. Englewood Cliffs, New Jersey: Prentice Hall. [Contains many applications of manufacturing systems.]

El-Taha M. and Stidham S. (1999) *Sample-Path Analysis of Queueing Systems*, 295 pp. Boston: Kluwer Academic Publisher. [This book provides a state-of-the-art treatment of sample path arguments.]

Gross D. and Harris C.M. (1985). *Fundamentals of Queueing Theory*, 587 pp. New York: Wiley. [Classical textbook on queueing systems.]

Kelly F.P. and Williams R.J. (1995) Stochastic Networks. *IMA Volumes in Mathematics and its Applications*, **71**. New York: Springer-Verlag. [Collection of recent papers on stochastic networks.]

Kleinrock L. (1975). *Queueing Systems*. Vol.1*: Theory, Vol.*2*: Computer Applications*. New York: John Wiley & Sons. [Contains many applications in computer science.]

Loynes R.M. (1962). The Stability of a Queue With Non-Independent Inter-Arrival and Service Times. *Proc. Camb. Philos. Soc.* **58**, 497-520. [Stability of stationary input queues.]

Sennott L.I. (1999). *Stochastic Dynamic Programming and the Control of Queueing Systems*, 328 pp. New York: John Wiley & Sons. [Investigates control problems which arise in queueing systems.]

Serfozo R. (1999). *Introduction to Stochastic Networks*, 300 pp. New York: Springer-Verlag. [Presents an advanced treatment of queueing networks]

Biographical Sketch

Nicole Bäuerle was born in 1968. She received the M.S. degree in mathematics & economics in 1992 , the Ph.D. degree in mathematics in 1994 and the Habilitation in mathematics in 1999, all from the University of Ulm. Since 2001 she has been Associate Professor at the University of Ulm. Her research interests include the analysis and control of stochastic processes with applications in telecommunication, computer science, finance and insurance. Dr. Bäuerle received two awards in 1996 for her Ph.D. thesis: one from the German special interest group in stochastics and the other one from the Ulmer Universitätsgesellschaft. She is currently Editor of the journal *Stochastic Models*.