Queueing Theory

Jeff Gill

2003

1 Introduction

Social scientists often study situations in which there is excess demand for some set of limited resources. Queueing theory is a subfield of stochastic modeling where the process by which demanders of a specific service or resource are assumed to wait for accommodation. The methodology is widely used in many literatures to describe assembly-line type processes and services in which the completion time is indeterminant. Developed initially to analyze telephone traffic and industrial throughput (Erlang 1935; Brockmeyer, Halstrom and Jensen 1960), it interprets the results from imposing a set of rules that describe three aspects of provision: the manner in which demanders arrive in the system, the capabilities and resources of the supplier, and how individual cases are treated relative to others.

2 Classic Queueing Theory

The first component of the model stipulates that *transactions arrive* stochastically according to an assumed parametric form. Queueing theory models are much easier to describe if we can claim that these arrivals are independent and serial (not arriving simultaneously). It is not required that these transactions stay in the assigned or selected queue until served; they can leave immediately due to queue length (balking), leave after waiting a specified time without being served (reneging), or possibly change to another available queue in the system (jockeying). The parameterized arrivals rate can change with the time index (non-stationarity), or remain fixed over time (stationarity). In addition, different characteristics can be attached to transactions denoting things like priority, complexity, or flags indicating behavior once in the system.

The second major component of the model is a description of the system's *servicing resources.* Attributes include the number of service queues (or channels), the existence of priority queues, layers of service requirements, and buffering or storage capacity. These features and the way they are arranged determine how efficiently the system can process incoming transactions.

The third and final part of the basic structure is the *set of rules governing queue discipline.* These include the order in which queued transactions are served, priority arrangements, and possibly allowances for "cheating". The two most common arrangements for transaction service are first-in/first-out, and first-in/last-out. In addition, priority arrangements can be defined focus on the handling of important transactions as they enter the system; preemptive-priority cases get served directly usually by displacing a transaction being served, non-preemptive-priority cases are placed at the front of the queue and therefore are accommodated by the next available server.

3 Terminology for Queueing Theory

There exists specialized notation that describes the assumptions of any given queueing theory model called Kendall notation (named after Sir David Kendall). This has the form: A/B/σ/k/M/Z, where A is the distribution of arrivals, B is the service time distribution, σ is the number of servers, k is the total number of transactions that can be either served or stored in the queue, M is the population size from which arrivals are drawn, and Z is a description of the queue discipline.

So for example, M/D/3/8/40,000/FIFO might describe a barber shop in a small town with: exponential arrivals, deterministic service (identical cuts), 3 barbers, 5 seats for waiting customers, in a medium sized town, and serving in order of arrival. It is typical, however, to drop the last three terms in the list when they are obvious from the context of the problem, especially the population size which cannot be controlled and often does not affect the model specification.

The most straightforward and common model specification is M/M/σ: exponentially distributed inter-arrival time (therefore Poisson distributed arrivals for time period t), exponentially distributed service times, and σ servers. Most texts (cf. Gross and Harris 1998) start with these systems because, with mild assumptions, many common empirical situations can be effectively described.

4 References

Brockmeyer, E., Halstrom, H.L., and Jensen, Arne . 1960. *The life and works of A.K. Erlang.* Kobenhavn: Akademiet for de Tekniske Videnskaber.

Erlang, A. K. 1935. *Fircifrede logaritmetavler og andre regnetavler til brug ved undervisning og i praksis.* Kobenhavn: G.E.C. Gads.

Gross, Donald, and Harris, Carl M. 1998. *Fundamentals of Queueing Theory.* Third Edition. New York: Wiley & Sons.