

Fundamentals of Queuing Models

Michela Meo

Maurizio M. Munafò

Michela.Meo@polito.it - Maurizio.Munafò@polito.it

Copyright

Quest'opera è protetta dalla licenza *Creative Commons NoDerivs-NonCommercial*. Per vedere una copia di questa licenza, consultare:
<http://creativecommons.org/licenses/nd-nc/1.0/>
oppure inviare una lettera a:
Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

This work is licensed under the *Creative Commons NoDerivs-NonCommercial* License. To view a copy of this license, visit:
<http://creativecommons.org/licenses/nd-nc/1.0/>
or send a letter to
Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.

Modeling a TLC network

- Modeling and simulation of a network often involves common components
 - Network elements (memory buffers, transmission lines,...)
 - Protocols and system interactions through message exchanges
- Analytical tools for modeling these components
 - Network elements -> Queues and queuing networks
 - Protocols -> State machines
- Simulation is often used when the analytical tools are not powerful enough

Queuing Models

- The basic element for the modeling of several real world systems is a *queue*
 - Post office
 - Highway tollbooth
 - Shop counter
 - ... almost anything else ...
- Since queues are so important in modeling, it's useful to reassess a few key concepts

Queue Characteristics

- We can give a very general and high level description of a queuing system:
 - "Customers" require a "service" from one or more "servers", waiting their turn in a "waiting line"
- What are the "customers"? How do they behave? What is a "service"? For each system there is a specific answer to these and all the other questions we can ask...

Queue Characteristics

- The Calling Population
 - The number of possible customers who can arrive at the system and request a service
 - Usually we suppose infinite population (a new customer can arrive in any moment)
 - Finite population is used only in very specific cases
- System Capacity
 - The maximum number of customers in the waiting line or system
 - Limited space in the waiting line -> when line is full new customers cannot enter the system
 - For some systems we can suppose unlimited capacity

Queue Characteristics

- The Arrival Process
 - Infinite-population models characterized by the *aggregate interarrival time* (time between the arrival of two customers)
 - Scheduled times or random times (with a probability distribution)
 - Single or group (*batches*) arrivals
 - Batches with fixed or random size
 - Finite-population models usually characterized by the *individual interarrival time*

Queue Characteristics

- The Poisson Arrival Process
 - Very important model for random arrivals
 - If A_n is the interarrival time between customer $n-1$ and customer n , then for a Poisson arrival process A_n is exponentially distributed with mean $1/\lambda$ time units

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{elsewhere} \end{cases}$$
 - The arrival rate is λ customers per time unit
 - Several nice statistical properties

Queue Characteristics

- Queue behavior, or the actions of the customers in the waiting line
 - Wait until service
 - Leave if the queue is too long
 - The customer does not enter the system or is ejected from the system by a specific queue policy
 - Leave if queue is moving too slow
 - Impatient customers
 - Move among queues if chosen one is too slow

Queue Characteristics

- Queue Discipline
 - For single services: the order the customers will be chosen for service when a server becomes free
 - First-In-First-Out (FIFO)
 - Last-In-First-Out (LIFO)
 - Random order
 - Shortest Processing Time First (SPT)
 - Priority service
 - For multiple services: policy for serving the customers
 - Round-robin
 - Processor sharing
 - Time-slicing, with or without priority and preemption

Queue Characteristics

- Service times or duration
 - Independent and identically distributed random variables
 - Depending on the customers, by type, class or priority
 - Depending on the state of the system
- Service mechanisms
 - Single server
 - Multiple parallel servers
 - Unlimited servers

Queue Characteristics

- Kendall Notation: A/B/c/N/K
 - A → Interarrival time distribution
 - B → Service time distribution
 - c → Number of parallel servers
 - N → system capacity
 - K → size of the calling population
- Possible A and B: M (exponential), D (constant), E_k (Erlang k), G (arbitrary), GI (general independent), ...
- If N and K are infinite, they are dropped from the notation
 - M/M/1/∞/∞ → M/M/1

Performance Measures

- In studying queuing systems we are almost always interested in long-run, steady-state performance indexes
- System parameters
 - Arrival rate of the customers λ [customers/time unit]
 - Service rate of one server μ [customers/time unit]
- Indexes
 - Average number of customers in the system L or in the queue L_Q
 - Average time spent by customers in the system W or in the queue W_Q and its distribution
 - Probability of having n customers in the system P_n
 - Server utilization ρ
 - Loss probability

Performance Measures

- System stability
 - For a single server queue, it must be $\lambda/\mu < 1$
 - When $\lambda/\mu > 1$, the rate at which customers arrive at the system is larger than the rate at which the server can provide service
 - customers are delayed more and more
 - the size of the waiting line increases in time at rate $\lambda - \mu$
 - For a queue with c servers, it must be $\lambda/\mu < c$

Performance Measures

- Conservation equation: Little's law
 - $L = \lambda W$
 - Valid, in its general form, for any queuing system with any queue and service discipline, if the system is stable
- Server utilization ρ
 - Defined as the fraction of time a server is busy

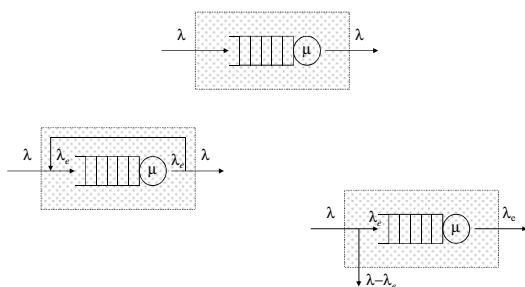
$$\rho = \frac{\lambda}{c\mu}$$

Performance Measures

- Loss probability
 - When the waiting line is limited, even if $\rho < 1$, due to the random nature of the system, it is possible for an arriving customer not to find room in queue
 - Depending on the context, we talk about a loss, dropping, or blocking event and we are interested in its probability

Performance Measures

- Overall system behavior



Queueing Networks

- Some queue systems can be combined in complex networks of queues
 - Arrival rate λ_i at queue i is a combination of the arrival rate from outside the system and the arrival rate from the other queues
 - For queue i , utilization is $\rho_i = \lambda_i / c\mu_i$ and $\rho_i < 1$ is required for the queue to be stable
 - If no customer is destroyed in the system, the departure rate is equal to the arrival rate



Protocols

- Not all the systems can be modeled by queues and queuing networks
- Complex systems are often characterized by the interaction of several components and the evolution of this interaction in time
- Protocols are typical examples:
 - Interaction among the elements inside each communicating system
 - Interaction among remote communicating systems



Protocols

- Modeling using state machines (or finite state automata)
 - Identification of the protocol states and their relations
 - Description of the system evolution as passage through the states
- Performance measures
 - Time spent in each state
 - Recurrence time
 - No. of visits to each state in the time unit