Queueing Theory

Carey Williamson
Department of Computer Science
University of Calgary

UNIVERSITY OF CALGARY

Outline

- Plan:
  - Introduce basics of Queueing Theory
  - Define notation and terminology used
  - Discuss properties of queueing models
  - Show examples of queueing analysis:
    - M/M/1 queue
    - Variations on the M/G/1 queue
    - Open queueing network models
    - Closed queueing network models

UNIVERSITY OF CALGARY

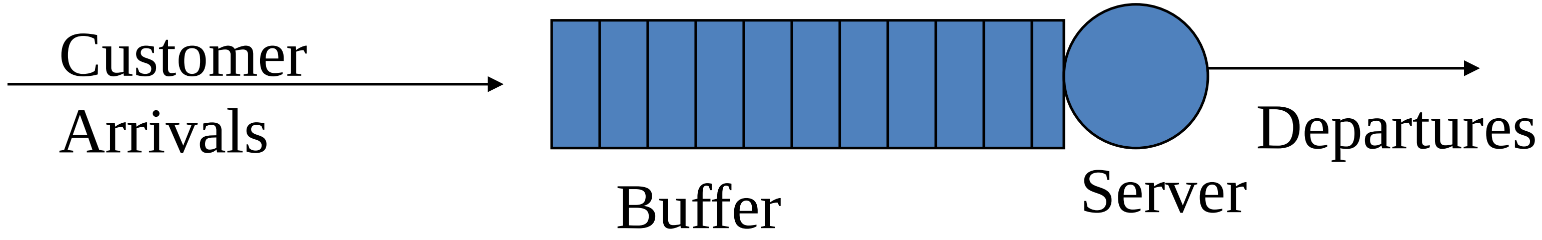
Queueing Theory Basics

- Queueing theory provides a very general framework for modeling systems in which customers must line up (queue) for service (use of resource)
  - Banks (tellers)
  - Restaurants (tables and seats)
  - Computer systems (CPU, disk I/O)
  - Networks (Web server, router, WLAN)

UNIVERSITY OF CALGARY

Queue-based Models

- Queueing model represents:
  - Arrival of jobs (customers) into system
  - Service time requirements of jobs
  - Waiting of jobs for service
  - Departures of jobs from the system
- Typical diagram:



Why Queue-based Models?

- In many cases, the use of a queueing model provides a quantitative way to assess system performance
  - Expected waiting time for service
  - Number of buffers required to limit loss of customers
  - Response time (e.g., Web page download time)
  - Throughput (e.g., job completions per second)
- Reveals key system insights (properties)
- Often with efficient, closed-form calculation

UNIVERSITY OF CALGARY

Caveats and Assumptions

- In many cases, using a queueing model has the following implicit underlying assumptions:
  - Poisson arrival process
  - Exponential service time distribution
  - Single server
  - Infinite capacity queue
  - First-Come-First-Serve (FCFS) discipline (a.k.a. FIFO)

Note: important role of memoryless property!

Advanced Queueing Models

- There is a tonne of published work on variations of the basic model:
  - Correlated arrival processes
  - General (G) service time distributions
  - Multiple servers
  - Vacationing servers
  - Finite capacity systems
  - Other scheduling disciplines (non-FIFO)
- We will start with the basics!

UNIVERSITY OF CALGARY

Queue Notation

- Queues are concisely described using the Kendall notation, which specifies:
  - Arrival process for jobs {M, D, G, …}
  - Service time distribution {M, D, G, …}
  - Number of servers {1, n}
  - Storage capacity (buffers) {B, infinite}
  - Service discipline {FIFO, PS, SRPT, …}
- Examples: M/M/1, M/G/1, M/M/c/c

UNIVERSITY OF CALGARY

The M/M/1 Queue

- Assumes Poisson arrival process, exponential service times, single server, FCFS service discipline, infinite capacity for storage, with no loss
- Notation: M/M/1
  - Markovian arrival process (Poisson)
  - Markovian service times (exponential)
  - Single server (FCFS, infinite capacity)

UNIVERSITY OF CALGARY

The M/M/1 Queue (cont'd)

- Arrival rate: λ (e.g., customers/sec)
  - Inter-arrival times are exponentially distributed and independent with mean $1 / \lambda$
- Service rate: μ (e.g., customers/sec)
  - Service times are exponentially distributed and independent with mean $1 / \mu$
- System load: $\rho = \lambda / \mu$

  $0 \leq \rho \leq 1$ (also known as utilization factor)
- Stability criterion: $\rho < 1$ (single server systems)

UNIVERSITY OF CALGARY

Queue Performance Metrics

- $\overline{N}$: Avg number of customers in system as a whole, including any in service
- Q: Avg number of customers in the queue (only), excluding any in service
- W: Average waiting time in queue (only)
- T: Avg time spent in system as a whole, including waiting time plus service time
- Note: Little's Law: $\overline{N} = \lambda T$

UNIVERSITY OF CALGARY

M/M/1 Queue Results

- Average number of customers in the system: $N = \rho / (1 - \rho)$
- Variance: $Var(N) = \rho / (1 - \rho)^2$
- Waiting time: $W = \rho / (\mu (1 - \rho))$
- Time in system: $T = 1 / (\mu (1 - \rho))$

UNIVERSITY OF CALGARY

The M/D/1 Queue

- Assumes Poisson arrival process, deterministic (constant) service times, single server, FCFS service discipline, infinite capacity for storage, no loss
- Notation: M/D/1
  - Markovian arrival process (Poisson)
  - Deterministic service times (constant)
  - Single server (FCFS, infinite capacity)

UNIVERSITY OF CALGARY

M/D/1 Queue Results

- Average number of customers:
  $Q = \rho/(1 - \rho) - \rho^2 / (2 (1 - \rho))$
- Waiting time: $W = x \rho / (2 (1 - \rho))$ where x is the mean service time
- Note that lower variance in service time means less queueing occurs ☺

UNIVERSITY OF CALGARY

The M/G/1 Queue

- Assumes Poisson arrival process, general service times, single server, FCFS service discipline, infinite capacity for storage, with no loss
- Notation: M/G/1
  - Markovian arrival process (Poisson)
  - General service times (must specify F(x))
  - Single server (FCFS, infinite capacity)

UNIVERSITY OF CALGARY

M/G/1 Queue Results

- Average number of customers:

  $Q = \rho + \rho^2 (1 + C^2) / (2 (1 - \rho))$ where C is the Coefficient of Variation (CoV) for the service-time distribution F(x)
- Waiting time:

  $W = x \rho (1 + C^2) / (2 (1 - \rho))$ where x is the mean service time from distribution F(x)
- Note that variance of the service time distn could be higher or lower than for exponential distn!

UNIVERSITY OF CALGARY

The G/G/1 Queue

- Assumes general arrival process, general service times, single server, FCFS service discipline, infinite capacity for storage, with no loss
- Notation: G/G/1
  - General arrival process (specify G(x))
  - General service times (must specify F(x))
  - Single server (FCFS, infinite capacity)

UNIVERSITY OF CALGARY

Queueing Network Models

- So far we have been talking about a queue in isolation
- In a queueing network model, there can be multiple queues, connected in series or in parallel (e.g., CPU, disk, teller)
- Two versions:
  - Open queueing network models
  - Closed queueing network models

UNIVERSITY OF CALGARY

Open Queueing Network Models

- Assumes that arrivals occur externally from outside the system
- Infinite population, with a fixed arrival rate, regardless of how many are in the system
- Unbounded number of customers are permitted within the system
- Departures leave the system (forever)

UNIVERSITY OF CALGARY

Closed Queueing Network Models

- Assumes that there is a finite number of customers, in a self-contained world
- Finite population; arrival rate varies depending on how many and where
- Fixed number of customers (N) that recirculate in the system (forever)
- Can be analyzed using Mean Value Analysis (MVA), recurrence relations, and balance equations