Performance Modeling and Design of Computer Systems- Ch 2 Queueing Theory Terminology

Debobroto Das Robin

Kent State University

Spring 2020

drobin@kent.edu

Overview

Queueing theory Terminology-1

- **Service Order**: order in which jobs will be served by the server.Assume First-Come-First-Served (FCFS) if not explicitly mentioned
- **Average Arrival rate** λ, at which jobs arrive to the server. Ex. $\lambda = 3$ jobs/sec).
- **Mean Interarrival Time** Avg time between successive job arrivals. (e.g., $1/\lambda = 1/3$ sec).
- **Service Requirement** Time it would take the job to run on this server if there were no other jobs around (no queueing). Random Variable S
- **Mean Service Time($E(S)$)** This is the expected value of S , namely the average time required to service a job on this CPU, where "service" does not include queueing time.
- **Average Service Rate**μ, at which jobs are served. $\mu = 1/E[S]$

Queueing theory Terminology-2

- **Response Time, Turnaround Time, Time in System, or Sojourn Time** (T) : a job's response time by $T = t_{depart} - t_{arrive}$
- **Waiting Time or Delay** (T_Q): time that the job spends in the queue, not being served. It is also called the "time in queue" or the **"wasted time."** $E[T] = E[T_Q] + E[S]$.
- Under FCFS service order, waiting time is the time from when a job arrives to the system until it first receives service.
- **Number of Jobs in the System (N)**: This includes those jobs in the queue, plus the one being served (if any).
- **Number of Jobs in Queue** (N_Q): This denotes only the number of jobs waiting (in queue).

Queueing theory Terminology-3

- **Device Utilization** (ρ_i) is the fraction of time device i is busy.
- Suppose we watch a device i for a long period of time. Let τ denote the length of the observation period. Let B denote the total time during the observation period that the device is non-idle (busy). Then $\rho_i = \frac{B}{\tau}$
- **Device Throughput** (X_i): the rate of job completions at device/system i (e.g., jobs/sec).
- Let C denote the total number of jobs completed at device i during time τ . Then $X_i = \frac{C}{\tau}$
- **Relation:** $X_i = \frac{C}{\tau} = \frac{C}{B} * \frac{B}{\tau} = \frac{1}{\frac{B}{C}} * \frac{B}{\tau} = \frac{1}{E[S]} * \rho_i = \mu_i * \rho_i$

Classfication of Queueing Networks

Open Networks

- open queueing network has external arrivals and departures
- Example
  - CPU uses a time-sharing scheduler to serve a queue of jobs waiting for CPU time
  - Router in a network serves a queue of packets waiting to be routed.

Open Networks: Example

Network of Queues with Probabilistic Routing

- Server i receives external arrivals ("outside arrivals") with rate r_i .
- Server i also receives internal arrivals from some of the other servers.
- A packet that finishes service at server i is next routed to server j with probability p_{ij} .
- Multiple **"class"** of the packet, may have different probability according to routing scheme

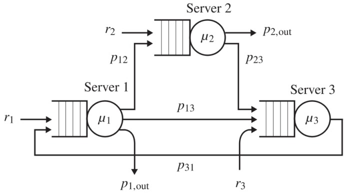


Figure 2.3. Network of queues with probabilistic routing.

Open Networks: Example

Network of Queues with Probabilistic Routing

- Real application in internet
  - Wire delay can be replaced by a server with some rate matching with wire delay
  - **Goal**: is to predict RTT
  - **Deterministic Varation**: instead of P_{ij}, specific path to next server

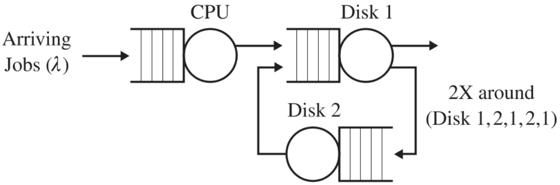
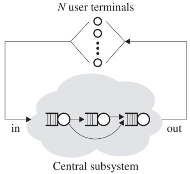


Figure 2.4. Network of queues with non-probabilistic routing.

Classfication of Queueing Networks

Closed Networks-1

- Closed queueing networks have no external arrivals or departures. They can be classified into two categories
- Example
  - **Interactive (Terminal-Driven) Systems**

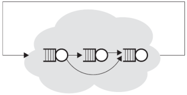


- Terminals represent users who each send a job to the "central subsystem" and then wait for a response.
- The central subsystem is a network of queues.
- A user cannot submit next job before previous job returns.
- Thus, the number of jobs in the system is fixed (equal to the number of terminals). Called the load or **MPL (multiprogramming level)**
- $E[T] = E[R] + E[Z]$: $E[R]$ = Avg time to get from in to out, $E[Z]$ = user thinking time

Classfication of Queueing Networks

Closed Networks-2

- **Batch Systems**



  - an interactive system with a think time of zero with dfrnt goal
  - Goal: For a batch system, the goal is to obtain high throughput, so that as many jobs as possible are completed overnight.
  - **in = out** for the batch system.
  - X is the number of jobs crossing "out" per second

Classfication of Queueing Networks

Open vs Closed Networks-2

- Open Systems
  - Throughput, X , is independent of the μ_i
  - X is not affected by doubling the μ_i
  - Throughput and response time are not related.
- Closed Systems
  - Throughput, X , is dependent of the μ_i
  - If we double all the μ_i holding N constant, then X changes
  - Higher throughput $\Longleftrightarrow$ Lower avg. response time