This Week’s Agenda

Last Time

- CTMC

Today

- A Primer on Queueing Theory
Denote the system by $A/B/k/c - S$, where

- $A$ is the inter-arrival time distribution
- $B$ is the service time distribution at each server
- $k$ is the number of servers
- $c$ is the queue capacity (omit if $\infty$)
- $S$ is the service discipline (omit if FIFO)

Some common distributions of inter-arrival and service times:

- $M$: Markov, i.e. exponential; e.g. $M/M/1$, $M/M/k$
- $D$: Deterministic; e.g. $M/D/k$
- $E_m$: Erlang-$m$ (common with phone call arrivals at toll offices)
- $H_m$: Hyper-$m$
- $G$: a general distribution
Basic Parameters and Performance Measures

Input Parameters

- Packet arrival processes (from the outside)
- Distributions of service times at servers
- Service disciplines at servers: FIFO, LIFO, Random, Priorities (rush order first, shortest processing time first, ...)
- Service capacity (nodes have single or multiple servers)
- Queue capacity (finite or infinite)

Output Performance Measures

- Distributions of waiting times at nodes
- Distributions of sojourn times (wait + service times)
- Distributions of number of packets at nodes
- Distributions of busy times at servers
- Packet loss rates
- Throughput
- ...
Common Parameters and Measures for $A/B/k/c$ Systems

The averages

- $\lambda$: mean arrival rate
- $\mu$: mean service time at each server
- $\rho = \lambda/\mu$: traffic intensity
- $\bar{N}$: mean number of packets in the system
- $\bar{N}_q$: mean number of packets in the queue
- $\bar{W}$: mean waiting time overall, also called response time
- $\bar{W}_q$: mean waiting time in the queue
- $r$: throughput (departure rate)
- $u$: utilization

The random variables

- $N(t)$: number of packets in the system at time $t$
- $\bar{N}$: number of packets in the system at steady state
- $\tau$: invariant distribution (if any) of $\{N(t)\}_{t \geq 0}$
Little’s Formulas

Think: each packet pays money to wait and/or to be serviced.

\[
\text{Mean rate system earns} = \lambda \times \text{Mean amount a packet pays}
\]

The following are called Little’s Formulas

- Each packet pays 1$ per unit time in the system, then
  \[
  \bar{N} = \lambda \cdot \bar{W}
  \]

- Each packet pays 1$ per unit time in the queue, then
  \[
  \bar{N}_q = \lambda \cdot \bar{W}_q
  \]
The $M/M/1$ Queue

- A BDP with constant birth rate $\lambda$, constant death rate $\mu$

- Stability condition

\[ C = \sum_{n=0}^{\infty} \rho^n < \infty \text{ iff } \rho < 1 \text{ iff } \rho \lambda < \mu \]

- Steady state probabilities

\[ \tau_n = \frac{1}{C} \rho^n = (1 - \rho) \rho^n \quad n \geq 0 \]

- Fraction of time system has $n$ packets is $\tau_n (\lambda + \mu)$

- It then follows that

\[ \Pr[N \geq m] = 1 - \sum_{n=0}^{m-1} \tau_n = \rho^m \]

\[ \bar{N} = \sum n \tau_n = \frac{\rho}{1 - \rho} \]

\[ \bar{W} = \frac{\bar{N}}{\lambda} = \frac{1}{\mu - \lambda} \]

\[ \bar{W}_q = \bar{W} - \frac{1}{\mu} = \frac{\rho}{\mu - \lambda} \]

\[ r = (1 - \tau_0) \mu = \lambda \]

\[ u = 1 - \tau_0 = \rho \]
$M/M/1$: Some Performance Graphing

![Graph](image)

- Mean number of packets in the system vs. Utilization
- Mean Response Time vs. Utilization

$x/(1-x)$

$1/(1-x)$
The $M/M/k$ Queue

- This is a birth and death process with constant birth rate $\lambda$ and death rates

$$\mu_i = \begin{cases} 
i \mu & 1 \leq i \leq k \\
\kappa \mu & i > k \end{cases}.$$ 

- The stability condition is

$$C = \sum_{n=0}^{k-1} \frac{\rho^n}{n!} + \frac{\rho^k}{k!} \sum_{i=0}^{\infty} \left( \frac{\rho}{k} \right)^i < \infty,$$

which holds iff $\rho < k$.

- At equilibrium,

$$\tau_0 = \left[ \sum_{n=0}^{k-1} \frac{\rho^n}{n!} + \frac{\rho^k}{k!} \frac{1}{1 - \rho/k} \right]^{-1} \tau_0 \frac{\rho^n}{n!} \frac{1}{1 - \rho/k}$$

$$\tau_n = \begin{cases} \tau_0 \frac{\rho^n}{n!} & 1 \leq n \leq k \\
\tau_0 \frac{\rho^n}{k! k^{n-k}} & n > k \end{cases}$$
The $M/M/1/c$ Queue

- This is a birth and death process with constant birth rate $\lambda$ and constant death rate $\mu$, state space $\{0, 1, \ldots, c\}$

- If $\rho \neq 1$,
  \[
  \tau_n = \begin{cases} 
  \rho^n \frac{(1-\rho)}{1-\rho^{c+1}} & n \leq c \\
  0 & n > c 
  \end{cases}
  \]

- If $\rho = 1$,
  \[
  \tau_n = \begin{cases} 
  \frac{1}{c+1} & n \leq c \\
  0 & n > c 
  \end{cases}
  \]

- Thus,
  \[
  \tilde{N} = \begin{cases} 
  \frac{\rho}{1-\rho} - \frac{c+1}{1-\rho^{c+1}} \rho^{c+1} & \rho \neq 1 \\
  c/2 & \rho = 1 
  \end{cases}
  \]

- Loss probability
  \[
  p_{\text{loss}} = \tau_c = \begin{cases} 
  \rho^n \frac{(1-\rho)}{1-\rho^{c+1}} & \rho \neq 1 \\
  \frac{1}{c+1} & \rho = 1 
  \end{cases}
  \]
$M/M/1/30$: Some Performance Graphing

\[ \frac{x}{1-x} - x^{31} \cdot \frac{31}{1-x^{31}} \]

\[ x^{30} \cdot \frac{1-x}{1-x^{31}} \]
Simple Comparisons of Queueing Systems

Question

Intuitively, is it better (in terms of response time) to have

an $M/M/1$ with service rate $10\mu$ or

an $M/M/10$ with service rate $\mu$, 
Simple Comparisons of Queueing Systems

Figure 1: $W_a(x)$: waiting time of the $M/M/10$ queue
$W_b(x)$: waiting time of the $M/M/1$ queue