Queueing Theory

Queueing system: a buffer (waiting room), service facility (one or more servers) and a scheduling policy (first come first serve, etc.)

We are interested in what happens when a stream of customers (jobs) arrive to such a system (throughput, sojourn time, number in system, server utilization, etc.)
Terminology

A/B/c/K queue (due to Kendall)
- $A$ - arrival process, interarrival time distr.
- $B$ - service time distribution
- $c$ - no. of servers
- $K$ - capacity of buffer

Does not specify scheduling policy
Standard values for A and B

$M$ - exponential distribution ($M$ is for Markovian)

$D$ - deterministic (constant)

$GI; G$ - general distribution
Some basic relations

some notation

- $C_n$ - customer $n$, $n = 1, 2, ...$
- $a_n$ - arrival time of $C_n$
- $d_n$ - departure time of $C_n$
- $\alpha(t)$ - number of arrivals by time $t$
- $\delta(t)$ - number of departures by time $t$
- $N(t)$ - no. in system at time $t$

$$N(t) = \alpha(t) - \delta(t)$$
Elementary Queueing Theory
**Little's Law**

additional notation

\( \gamma(t) \) - total time spent by all customers in system during interval \((0,t)\)

\[
\gamma(t) = \sum_{n=1}^{\alpha(t)} \min\{d_n, t\} - \alpha_n = \int_0^t N(s)ds
\]

\( T_t \) - average time spent in system during \((0,t)\) by customers arriving in \((0,t)\)

\[
T_t \equiv \frac{\gamma(t)}{\alpha(t)}
\]

\( N_t \) - average number of customers in system during \((0,t)\)

\[
N_t \equiv \frac{\gamma(t)}{t}
\]

It follows that \( N_t = \lambda_t T_t \)

Assume \( N_t, \lambda_t, \) and \( T_t \) have limits \( N, \lambda, \) and \( T. \)

**Little's Law:** \( N = \lambda T \)

Can be applied to a large variety of systems.
Utilization Law for Single Server Queue

- $X$ - stationary service time with mean $E[X]$
- $Y$ - state of server, $Y=0$ if idle, $Y=1$ if busy
- $\rho$ - server util., $\rho = P(Y=1)$
Utilization Law for Single Server Queue

\[ T = E[X] \]

\[ N = P(Y=0) \times 0 + P(Y=1) \times 1, \]
\[ = P(Y=1), \]
\[ = \rho \]

- Little's law yields

Utilization Law: \[ \rho = \lambda E[X] \]
**M/M/1 Queueing System**

- arrivals - Poisson process with rate $\lambda$
- service times - exponentially distributed with rate $\mu$ (mean $1/\mu$)
- $\rho \equiv \lambda / \mu$
- model as BD process; state $N(t)$ equals number of customers in system at time $t$
M/M/1 Queueing System

- $\lambda < \mu$ for steady state solution; let $N = \lim_{t \to \infty} N(t)$

- $\pi_i = P(N = i)$ satisfies
  \[ \mu \pi_i = \lambda \pi_{i-1}, \quad i = 1, \ldots \]
  or
  \[ \pi_i = \rho \pi_{i-1}, \quad i = 1, \ldots \]
M/M/1 Queueing System

Solution is

$$\pi_i = \rho^i \pi_0, i = 0, ...$$

Coupled with relation $$\sum_{i=0}^{\infty} \pi_i = 1$$ yields

$$\pi_i = \rho^i (1 - \rho), i = 0, ...$$

with mean $$E[N]$$

$$E[N] = \rho / (1 - \rho)$$
**M/M/1: Waiting Times**

- $X_n$ service time of $n$-th customer, $X_n \overset{d}{=} X$ where $X$ is exponential rv
- $W_n$ waiting time of $n$-th customer in queue (excl. service)
- $T_n$ sojourn (response) time of $n$-th customer; $T_n = W_n + X_n$
- When $\rho < 1$, steady state solution exists and $X_n, W_n, T_n \rightarrow X, W, T$

Q: what is $E[W]$?
Let $M$ denote number of customers in system at time of arrival of tagged customer.

Wait time is $X_1 + X_2 + \ldots + X_{M-1} + R$ where \{\(X_i\}\} are s.t.s of customers in queue and $R$ is remaining service time of customer in service.

- $X_i$ - exponential rv with mean $1/\mu$
- $R$ - exponential rv with mean $1/\mu$ because of memoryless property of expo. distr.

$$E[W \mid M=m] = m/\mu, \quad m = 0,1, \ldots$$
Waiting Times

\[ E[W] = \sum_{m=0}^{\infty} P(M = m)E[W \mid M = m] \]
\[ = \frac{1}{\mu} \sum_{m=0}^{\infty} mP(M = m) \]

Now, \( P(M = m) = P(N = m) \) (see next page).

Therefore,
\[ E[W] = E[N]/\mu = \frac{1}{\mu} \times \frac{\rho}{1 - \rho} \]

and
\[ E[T] = E[W] + 1/\mu = \frac{1/\mu}{1 - \rho} = \frac{1}{\mu - \lambda} \]
PASTA Property

PASTA: Poisson Arrivals See Time Averages

Consider a general system

- Poisson arrival process with rate $\lambda$
- $M(t)$ - system at time $t$ given that an arrival occurs in $(t, t + \Delta t)$.
- $N(t)$ system state at time $t$

$$P(M(t) = n) = \frac{P(N(t) = n \mid \text{arrival in } (t, t + \Delta t))}{P(N(t) = n, \text{arrival in } (t, t + \Delta t))}$$
$$= \frac{P(\text{arrival in } (t, t + \Delta t))}{P(N(t) = n)P(\text{arrival in } (t, t + \Delta t))}$$
$$= \frac{P(N(t) = n)}{P(\text{arrival in } (t, t + \Delta t))}$$
$$= P(N(t) = n)$$
Sojourn Time Distribution

Focus now on $f_T(t)$, $t \geq 0$

Condition on $M = n$,

$$f_{T|M}(t|n) = \frac{\mu(\mu t)^n e^{-\mu t}}{n!}, \quad t \geq 0; n = 0, 1, \ldots$$

Removal of conditioning

$$f_T(t) = \sum_{n=0}^{\infty} (1 - \rho)\rho^n \frac{\mu(\mu t)^n e^{-\mu t}}{n!}$$

$$= (1 - \rho)\mu e^{-\mu t} \sum_{n=0}^{\infty} (\rho\mu t)^n / n!,$$

$$= (\mu - \lambda)e^{-\mu t} e^{\lambda t}$$

$$= (\mu - \lambda)e^{-(\mu-\lambda)t}$$

$T$ is an exponential rv with mean $1/(\mu - \lambda)$

The following expression can be derived for $F_W(t)$:

$$F_W(t) = 1 - \rho e^{-(\mu-\lambda)t}, \quad t \geq 0$$
Sojourn Times

The $r$-percentile of the distribution of $T$ is easy to calculate:

$$\alpha_T(r) = E[T] \ln \left( \frac{100}{100 - r} \right)$$

where

$$P(T \geq \alpha_T(r)) = \frac{100 - r}{100}$$

Examples

$$\alpha_T(90) = 2.3 \ E[T],$$
$$\alpha_T(95) = 3 \ E[T],$$
Sojourn Times

**stretch factor** is ratio $E[T]/E[X]$. It is important because user's expectation is usually set by performance under light load where $E[T] \approx E[X]$. When stretch factor gets too long, user gets dissatisfied.

Factor of 5 often considered limit.

Example for $M/M/1$: $E[T]/E[X] = 1/(1-\rho)$. $\rho < .8$ ensures stretch factor less than 5.
Example

Given one GIP machine (10^9 instr./sec.)
- each job executes no. of instr. that is exponentially distr. with mean 5 \times 10^7 instructions
- Poisson arrival process, \( \lambda = 15 \) jobs/sec.

Questions that can be asked of this system
1. what is expected response time?
2. what is expected waiting time in queue?
3. what is expected no. of jobs waiting in queue?
4. what is max. supportable throughput yielding avg. delay \( \leq 5 \) sec.?
5. what is maximum response time for any job?
6. what is \( P(N \geq 10) \)?
7. what is 90-percentile?
M/M/1/K System

- arrivals - Poisson process with rate $\lambda$
- service times - exponentially distributed with rate $\mu$ (mean $1/\mu$)
- finite capacity of $K$ customers. Customers that arrive and find queue full are rejected (lost).
- model as BD process; state $N(t)$ equals number of customers in system at time $t$

\[ \pi_i = \rho \pi_{i-1} = \rho^i \pi_0 \quad i = 1, \ldots, K \]
Calculation of $\pi_0$

- $\lambda \neq \mu$

$$\sum_{i=0}^{K} \pi_i = \pi_0 \sum_{i=0}^{K} \rho^i = \pi_0 \frac{1 - \rho^{K+1}}{1 - \rho}$$

$$\sum_{i=0}^{K} \pi_i = 1 \Rightarrow \pi_0 = \frac{1 - \rho}{1 - \rho^{K+1}}$$

and

$$\pi_i = \frac{1 - \rho}{1 - \rho^{K+1}} \rho^i, \quad i = 0, 1, \ldots, K$$

- $\lambda = \mu$

$$\sum_{i=0}^{K} \pi_i = \pi_0 \sum_{i=0}^{K} \rho^i = \pi_0 (K + 1)$$

or

$$\pi_i = \frac{1}{K + 1}, \quad i = 0, 1, \ldots, K$$
**Expected Number in System**

- $\lambda \neq \mu$

\[
E[N] = \sum_{i=0}^{K} i\pi_i = \frac{1 - \rho}{1 - \rho K + 1} \sum_{i=0}^{K} i\rho^i = \frac{\rho}{1 - \rho} - \frac{(K + 1)\rho K + 1}{1 - \rho K + 1}
\]

- $\lambda = \mu$

\[
E[N] = \sum_{i=0}^{K} i\pi_i = \frac{1}{K + 1} \sum_{i=0}^{K} i = \frac{1}{K + 1} \frac{K(K + 1)}{2} = \frac{K}{2}
\]
Other Metrics

- throughput = \((1 - \pi_0) \mu\) or \(\lambda(1 - \pi_K)\)
- Prob. of buffer overflow = \(\pi_K\)
- \(E[T] = E[N]/\text{throughput} = E[N]/((1 - \pi_0) \mu)\)

be careful to use correct value of throughput which is not \(\lambda\)
One GIP machine revisited

Given one GIP machine (10^9 instr./sec.)
- each job executes no. of instr. that is exponentially distr. with mean $5 \times 10^7$ instructions
- Poisson arrival process, $\lambda = 15$ jobs/sec.
One GIP Machine Example Revisited

Suppose that the OS can only support a finite number of jobs, say $K$, and that a job arriving to find $K$ jobs already in the system is rejected.

- what value should $K$ be to ensure that the rejection probability is less than 0.005

More generally, assume that we want to choose minimum $K$ such that

$$\pi_K \leq \varepsilon \quad 0 < \varepsilon < 1$$
Suppose that the OS can only support a finite number of jobs, say $K$, and that a job arriving to find $K$ jobs already in the system is rejected.

- what value should $K$ be to ensure that the rejection probability is less than .005

More generally, assume that we want to choose minimum $K$ such that

$$\pi_K \leq \varepsilon \quad 0 < \varepsilon < 1$$

Consider the case, $\rho < 1$. Substituting in for $\pi_K$

$$\frac{1 - \rho}{1 - \rho^{K+1}} \rho^K \leq \varepsilon,$$

or

$$\rho^K (1 - \rho) \leq \varepsilon (1 - \rho^{K+1})$$

- Collecting terms,

$$\rho^K (1 - \rho (1 - \varepsilon)) \leq \varepsilon$$
One GIP Machine Example Revisited

Taking the logarithm of both sides,

\[ K \ln \rho \geq \ln \frac{\epsilon}{1 - \rho(1 - \epsilon)}, \]

or

\[ K \geq \frac{\ln \frac{\epsilon}{1 - \rho(1 - \epsilon)}}{\ln \rho} \]

Returning to our problem, \( K \geq 13.65 \)

Therefore \( K = 14 \) suffices.
M/M/c System

- arrivals - Poisson process with rate $\lambda$
- service times - exponentially distributed with rate $\mu$ (mean $1/\mu$)
- $c$ identical servers
- model as BD process; state $N(t)$ equals number of customers in system at time $t$
Balance equations:

\[
\pi_{i-1}\lambda = \pi_i \mu, \quad i \leq c, \\
\pi_{i-1}\lambda = \pi_i c \mu, \quad i > c
\]

Solution to balance equations

\[
\pi_i = \begin{cases} 
\frac{\lambda^i}{i! \mu^i} \pi_0, & 0 \leq i \leq c, \\
\frac{\lambda^i}{c! c^i - c \mu^i} \pi_0, & c < i
\end{cases}
\]
M/M/c Metrics

- prob. a customer has to wait

\[ P(\text{Wait}) = \sum_{n=c}^{\infty} \pi_n, \]
\[ = \frac{\rho^c}{\pi_0 c! (1 - \rho/c)} \]
\[ = \frac{\rho^c}{\pi_c (1 - \rho/c)} , \]
\[ = \frac{\rho^c/(c-1)!}{(c-\rho)[\sum_{n=0}^{c-1} \rho^n/n! + \rho^c/((c-1)!(c-\rho))]}. \]

- this last expression is called Erlang's C formula, \( C(c, \rho) \)

- waiting time statistics

\[ P(W \leq t) = 1 - C(c, \rho) e^{-(c\mu-\lambda)t}, \quad t \geq 0 \]
\[ E[W] = \frac{C(c, \rho)/\mu}{c - \rho}. \]
M/M/∞ system

Infinite server system (aka delay server)

- arrivals - Poisson process with rate $\lambda$
- service times - exponentially distributed with rate $\mu$ (mean $1/\mu$)
- each user gets its own server
- model as BD process; state $N(t)$ equals number of customers in system at time $t$
**M/M/∞ system**

Balance equations:

\[ \pi_{i-1}\lambda = \pi_i i\mu, \quad i = 0, 1, \ldots \]

Solution to balance equations

\[ \pi_i = \frac{\lambda^i}{i!\mu^i}\pi_0, \quad i = 0, 1, \ldots \]

\[ = e^{-\rho} \frac{\lambda^i}{i!\mu^i} \]
Machine Repairman Model

- K machines
- each machine fails at rate $\lambda$
- repairman repairs machines in FCFS order at rate $\mu$

Note similarity to a timesharing system. If we assume that user think times are exponentially distributed and service times are exponentially distributed we get the following BD process.
Balance equations:

\[ \pi_{i-1} \lambda (K - i + 1) = \pi_i \mu, \quad i = 1, \ldots, K \]
Machine Repairman Model

Solution to balance equations

$$\pi_i = \frac{K!}{(K-i)!} \left(\frac{\lambda}{\mu}\right)^i \pi_0, \quad i = 0, 1, \ldots, K$$

where

$$\pi_0 = B(K, \mu/\lambda) = \frac{(\mu/\lambda)^K/K!}{\sum_{i=0}^{K} (\mu/\lambda)^i/i!}$$

- server utilization is $1 - B(K, \mu/\lambda)$
- throughput at server, $\gamma$, is $\gamma = \mu (1 - B(K, \mu/\lambda))$
Q: what is $E[W]$, expected waiting time in queue?

$$E[W] = \frac{K}{\mu(1 - B(K, \mu/\lambda))} - \frac{1}{\lambda} - \frac{1}{\mu}$$
Machine Repairman Model

Q: what is \( E[W] \), expected waiting time in queue?

- Each customer has a cycle time \( C = Y + W + X \) where \( Y \) is think time (operating time) with \( E[Y] = 1/\lambda \).

Avg. cycle time is

\[
E[C] = E[Y] + E[W] + E[X],
\]

\[
= E[W] + 1/\lambda + 1/\mu
\]

- Thruput, \( \gamma \), is

\[
\gamma = K \left( \frac{1}{E[W] + 1/\lambda + 1/\mu} \right)
\]

- Therefore,

\[
E[W] = \frac{K}{\gamma} - 1/\lambda - 1/\mu,
\]

\[
= \frac{K}{\mu(1 - B(K, \mu/\lambda))} - 1/\lambda - 1/\mu
\]
Machine Repairman Model

Let \( N_q \) denote number waiting in queue. Little's law yields

\[
E[N_q] = \gamma E[W],
\]

\[
= K - \gamma \times \frac{1}{\lambda} - \gamma \times \frac{1}{\mu}
\]

Other metrics:

\[
E[N] = K - \gamma \frac{1}{\lambda}
\]

\[
E[T] = K \frac{1}{\gamma} - \frac{1}{\lambda}
\]
Interlude: Transforms

- why transforms? - makes life easier
- non-negative integer rvs - z-transform, probability generating function (pgf)
- nonnegative, real valued rvs - Laplace transform (LT)
**z-transform**

**Defn.** Have rv $X$ that takes nonnegative integer values $p_k = P(X = k)$, $k=0,1,...$

$$G_X(z) \equiv E[z^X] = \sum_{k=0}^{\infty} p_k z^k$$
**z-transforms: examples**

- X is geometric rv, $p_k = (1 - p)p^{k-1}$

  $$G_X(z) = \sum_{k=1}^{\infty} (1 - p)p^{k-1}z^k = \frac{(1 - p)z}{1 - pz},$$

- Poisson distr., $p_k = \lambda^k e^{-\lambda}/k!$

  $$G_X(z) = \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!}z^k$$

  $$= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda z)^k}{k!},$$

  $$= e^{-\lambda(1-z)}$$
Benefits

- **moments**: e.g., average

\[
\frac{dG_X(z)}{dz} = \sum_{k=1}^{\infty} kp_k z^{k-1}
\]

which translates to

\[
E[X] = \sum_{k=1}^{\infty} kp_k z^{k-1} \bigg|_{z=1} = \frac{dG_X(z)}{dz} \bigg|_{z=1}
\]

similarly

\[
\frac{d^2G_X(z)}{dz^2} \bigg|_{z=1} = E[X^2] - E[X]
\]
Benefits

- solution of difference equations: will look at soon
- convolution: let $X, Y$ be independent rvs with pgfs $G_X(z)$ and $G_Y(z)$. Let $U = X + Y$,

$$G_U(z) = G_X(z)G_Y(z)$$
Solving M/M/1 Queue with Transforms

$$(\lambda + \mu)\pi_i = \begin{cases} i \neq 0 & \lambda \pi_{i-1} + \mu \pi_{i+1} + \{i = 0\} \mu \pi_0 \quad i = 0, 1, \ldots \end{cases}$$

Multiplying by $z^i$, using $\rho = \lambda / \mu$, summing over $i$

$$(1 + \rho) \sum_{i=0}^{\infty} \pi_i z^i = \rho z \sum_{i=0}^{\infty} \pi_i z^i + z^{-1} \sum_{i=1}^{\infty} \pi_i z^i + \pi_0$$

or

$$(1 + \rho) G_N(z) = \rho z G_N(z) + z^{-1}(G_N(z) - \pi_0) + \pi_0$$
Solving M/M/1 Queue with Transforms

\[(1 + \rho)G_N(z) = \rho z G_N(z) + z^{-1}(G_N(z) - \pi_0) + \pi_0\]

Multiplying by \(z\) and rearranging yields

\[(\rho z^2 - (1 + \rho)z + 1)G_N(z) = (1 - z)\pi_0\]

Now

\[(\rho z^2 - (1 + \rho)z + 1)G_N(z) = (1 - z)\pi_0\]

which substituted into the above expression yields

\[G_N(z) = \frac{\pi_0}{1 - \rho z},\]

\[= \frac{1 - \rho}{1 - \rho z}\]
Inversion of z-transforms

We have looked at

\[ \{p_0, p_1, \ldots \} \longrightarrow G_X(z) \]

How about

\[ G_X(z) \overset{\text{invert}}{\longrightarrow} \{p_0, p_1, \ldots \} \]

The definition is:

\[ p_n = \frac{1}{n!} \left. \frac{d^n G_X(z)}{dz^n} \right|_{z=0} \]

Not always easy to compute and not focus of lecture

Note that \( p_0 = G_X(0) \) and \( G_X(1) = 1 \)
Laplace transforms

**Defn.** Given nonnegative, real valued rv $X$ with pdf $f_X(x)$ ($f_X(x) = 0, x < 0$)

**Example.** $X$ exponential rv, $f_X(x) = \lambda e^{-\lambda x}$
Laplace transforms

Moments:

\[ E[X] = - \frac{d}{ds} F_X^*(s) \bigg|_{s=0} \]

\[ E[X^i] = (-1)^i \frac{d^i}{ds^i} F_X^*(s) \bigg|_{s=0} \]
Laplace transforms

Defn. Given nonnegative, real valued rv $X$ with pdf $f_X(x)$ ($f_X(x) = 0, x > 0$)

$$F_X^*(s) \equiv E[e^{-sX}] = \int_0^\infty f_X(x)e^{-sx} \, dx$$

Example. $X$ exponential rv, $f_X(x) = \lambda e^{-\lambda x}$

$$F_X^*(s) = \int_0^\infty \lambda e^{-x(\lambda+s)} \, dx = \frac{\lambda}{\lambda + s}$$

Moments:

$$E[X] = -\frac{d}{ds} F_X^*(s) \bigg|_{s=0}$$

$$E[X^i] = (-1)^i \frac{d}{ds} F_X^*(s) \bigg|_{s=0}$$
Properties of Laplace Transform

**Convolution:** if $X_1, X_2, \ldots, X_n$ are independent, nonnegative rvs with transforms $F^{*}_{X_1}(s), \ldots, F^{*}_{X_n}(s)$, $Y = \sum_i X_i$, then

$$F^{*}_{Y}(s) = F^{*}_{X_1}(s)F^{*}_{X_2}(s) \cdots F^{*}_{X_n}(s)$$

Example, if $Y$ is n-th order Erlang
Properties of Laplace Transform

**Convolution:** if $X_1, X_2, \ldots, X_n$ are independent, nonnegative rvs with transforms $F_{X_1}(s), \ldots, F_{X_n}(s)$, $Y = \sum_i X_i$, then

$$F_Y(s) = F_{X_1}(s)F_{X_2}(s)\cdots F_{X_n}(s)$$

- Example, if $Y$ is $n$-th order Erlang

$$F^*(s) = \left(\frac{\lambda}{\lambda + s}\right)^n$$
Properties of Laplace Transform

Let $X_1, X_2, \ldots$ be iid nonnegative rvs with LT $F_X^*(s)$, let $N$ be a discrete rv with pgf $G_N(z)$.

Let $Y = X_1 + \ldots + X_N$

What is $F_Y^*(s)$?

$$F_Y^*(s) = G_N(F_X^*(s))$$