Queuing Theory

Basic properties, Markovian models, Networks of queues, General service time distributions, Finite source models, Multiserver queues
Kendall’s Notation for Queuing Systems

A/B/X/Y/Z:

A = interarrival time distribution
B = service time distribution
   G = general (i.e., not specified); M = Markovian (exponential); D = deterministic
X = number of parallel service channels
Y = limit on system pop. (in queue + in service); default is $\infty$
Z = queue discipline; default is FCFS (first come first served)
   Others are LCFS, random, priority
Other Notation

Random variables:

- $X(t)$ = Number of customers in the system at time $t$
- $S$ = Service time of an arbitrary customer
- $W^*$ = Amount of time an arbitrary customer spends in the system
- $W_Q^*$ = Amount of time an arbitrary customer spends waiting for service

Often we are most interested in the averages or expectations:

- $L = E[X(t)]$
- $W = E[W^*]$
- $L_Q = \text{Average number in the queue}$
- $W_Q = E[W_Q^*]$
- $L_S = \text{Average number in service}$
Little’s Formula

\( w_n^* \) is the time spent in the system by the \( n^{\text{th}} \) customer.

Assume these system times are uniformly finite and let
\( w = \lim_{k \to \infty} \frac{1}{k} \sum_{n=1}^{k} w_n^* \)
be the customer average time spent in the system.

Also, let
\[ \overline{X} = \lim_{s \to \infty} \frac{1}{s} \int_{0}^{s} X(t) dt \]
be the time average number of jobs in the system.

Then under very general conditions,
\[ \overline{X} = \lambda_a w \]
where \( \lambda_a \) is the arrival rate. This is usually written \( L = \lambda_a W \).

*The mean number of customers in the system is proportional to the mean time in the system!*
Heuristic Proof of Little’s Formula

Two ways to compute \( B(T) = \sum_{n=1}^{N} w_n^* = \int_0^T X(t)dt \)
Mean time in system in \((0,T)\): \( W(T) = B(T)/N = B(T)/A(T) \)
Mean number in system in \((0,T)\): \( \bar{X}(T) = B(T)/T \)

Then \( \lim_{T \to \infty} \bar{X}(T) = \lim_{T \to \infty} \frac{B(T)}{A(T)} \frac{A(T)}{T} = \lim_{T \to \infty} \lambda(T)W(T) \), or \( L = \lambda^a W \)
Other Little’s Formulae

In queue: \( L_Q = \lambda_a W_Q \)
In service: \( L_s = \lambda_a E[S] \)

and their implications…

Expected number of busy servers \( L_s = \lambda_a E[S] = \rho \)
Expected number of idle servers \# servers \( - \rho \)
Single server utilization \( L_s = \lambda_a E[S] = \rho \)
Single server prob. of empty system \( 1 - \rho \)
Observation Times

\[ P_n = \lim_{t \to \infty} P\{X(t) = n\}, n = 0, 1, \ldots \]

\( a_n = \) Proportion of arriving customers that find \( n \) in the system

\( d_n = \) Proportion of departing customers that leave behind \( n \) in the system

If customers arrive one at a time and are served one at a time then

\[ a_n = d_n \]

But these proportions may not match the limiting probability of having \( n \) in the system (long run proportion of time that \( n \) are in the system)

However, if arrivals follow a Poisson process then \( a_n = P_n \)

This is known as \textit{PASTA (Poisson Arrivals See Time Averages)}
M/M/1 Model

Single server, Poisson arrivals (rate $\lambda$), exponential service times (rate $\mu$)
CTMC (birth-death) model:
M/M/1 Steady-State Probabilities

\[ P_n = \lim_{t \to \infty} P\{X(t) = n\}, n = 0,1,\ldots \]

Level-crossing equations:
\[ \lambda P_n = \mu P_{n+1}, n = 0,1,\ldots \]

Solve for \( P_n \) in terms of \( P_0 \)
\[ P_n = \left(\frac{\lambda}{\mu}\right)^n P_0, n = 1,2,\ldots \]

Then use the facts that \( \sum_{n=0}^{\infty} P_n = 1 \), \( \sum_{n=0}^{\infty} r^n = (1-r)^{-1} \) if \( r < 1 \)
and substitute \( \rho = \frac{\lambda}{\mu} \)
to get
\[ P_n = \rho^n (1-\rho), n = 0,1,\ldots \text{ if } \rho < 1 \]
M/M/1 Performance Measures

Steady-state expected number of customers in the system

\[ L = \sum_{n=0}^{\infty} n P_n = \frac{\rho}{1 - \rho}, \quad \text{if } \rho < 1 \]

Mean time in system

\[ W = \frac{L}{\lambda_a} = \frac{1}{\mu (1 - \rho)} = \frac{1}{\mu - \lambda} \quad \text{by Little's Formula} \]

Mean time in queue

\[ W_Q = W - \frac{1}{\mu} = \frac{\lambda}{\mu (\mu - \lambda)} \]

Mean number in queue

\[ L_Q = \lambda W_Q = \frac{\lambda^2}{\mu (\mu - \lambda)} \]
M/M/1 Performance Measures

Distribution of time in system (FCFS)

\[ P[W^* \leq x] = \sum_{n=0}^{\infty} a_n P[W^* \leq x \mid \text{arrival sees } n \text{ customers}] \]

\[ = \sum_{n=0}^{\infty} (1 - \rho) \rho^n P[W^* \leq x \mid W^* \sim \text{gamma}(n+1, \mu)] \]

\[ = \sum_{n=0}^{\infty} (1 - \rho) \rho^n \sum_{l=n+1}^{\infty} e^{-\mu x} (\mu x)^l / l! = 1 - e^{-\mu (1 - \rho) x}, x \geq 0 \]

Exponential with parameter \( \mu(1 - \rho) \)

Reversibility: Departure process is Poisson with rate \( \lambda \)
Finite Capacity: M/M/1/N Model

Single server, exponential service times (rate $\mu$)
Poisson arrivals (rate $\lambda$) as long as there are $\leq N$ in the system
CTMC (birth-death) model:
M/M/1/N Steady-State Probabilities

\[ P_n = \lim_{t \to \infty} P\{X(t) = n\}, \quad n = 0, 1, \ldots \]

Level-crossing equations:
\[ \lambda P_n = \mu P_{n+1}, \quad n = 0, 1, \ldots, N-1 \]

Solve for \( P_n \) in terms of \( P_0 \)

Then use the facts that
\[ \sum_{n=0}^{\infty} P_n = 1, \quad \sum_{n=0}^{N} r^n = \frac{1-r^{N+1}}{1-r} \quad (\text{note: } r \text{ need not be } < 1) \]

to get
\[ P_n = \frac{(\lambda/\mu)^n}{1-(\lambda/\mu)^{N+1}} \left(1 - \frac{\lambda}{\mu}\right), \quad n = 0, 1, \ldots, N \]
M/M/1/N Performance Measures

(In the unlikely event that $\lambda = \mu$, for $n = 1, \ldots, N$, $P_n = P_0 = 1/(N + 1)$)

Steady-state expected number of customers in the system, $L$, has a “messy” closed form
Mean time in system $W = \frac{L}{\lambda_a}$ by Little's Formula

but here, $\lambda_a$ is the rate of arrival into the system $\lambda_a = \lambda (1 - P_N)$

$$W_Q = W - \frac{1}{\mu} \quad L_Q = \lambda (1 - P_N) W_Q$$
Tandem Queue

If arrivals to the first server follow a Poisson process and service times are exponential, then arrivals to the second server also follow a Poisson process and the two queues behave as independent M/M/1 systems:

$$P\{n \text{ customers at server 1 and } m \text{ customers at server 2}\} = \left(\frac{\lambda}{\mu_1}\right)^n \left(1 - \frac{\lambda}{\mu_1}\right) \left(\frac{\lambda}{\mu_2}\right)^m \left(1 - \frac{\lambda}{\mu_2}\right)$$
Open Network of Queues

- $k$ servers, customers arrive at server $k$ from outside the system according to a Poisson process with rate $r_k$, independent of the other servers.
- Upon completing service at server $i$, customer goes to server $j$ with probability $P_{ij}$, where $\sum_j P_{ij} \leq 1$.

For $j = 1, \ldots, k$, the total arrival rate to server $j$ is $\lambda_j = r_j + \sum_{i=1}^{k} \lambda_i P_{ij}$.

The number of customers at each server is independent and

If $\lambda_j < \mu_j$ for all $j$, then $P\{n \text{ customers at server } j\} = \left(\frac{\lambda_j}{\mu_j}\right)^n \left(1 - \frac{\lambda_j}{\mu_j}\right)$

that is, each acts like an independent M/M/1 queue!
Closed Queuing Network

- $m$ customers move among $k$ servers
- Upon completing service at server $i$, customer goes to server $j$ with probability $P_{ij}$, where $\sum_j P_{ij} = 1$
- Let $\pi$ be the stationary probabilities for the Markov chain describing the sequence of servers visited by a customer:

$$\pi_j = \sum_{i=1}^{k} \pi_i P_{ij}, \quad \sum_j \pi_j = 1$$

Then the probability distribution of the number at each server is

$$P_m(n_1, n_2, \ldots, n_k) = C_m \prod_{j=1}^{k} \left( \frac{\pi_j}{\mu_j} \right)^{n_j} \quad \text{if } \sum_j n_j = m$$
CQN Performance

Computation of the normalizing constant $C_m$ to get the stationary distribution can be lengthy; but may be mostly interested in the throughput $\lambda_m = \sum_{j=1}^{j} \lambda_m(j)$ where $\lambda_m(j)$ is the arrival rate to (and departure rate from) $j$.

**Arrival Theorem**: In the CQN with $m$ customers, the system as seen by arrivals to server $j$ has the same distribution as the whole system when it contains only $m-1$ customers.

This leads to mean value analysis to find $\lambda_m(j)$ along with $W_m(j) = \text{the average time a customer spends at server } j$, and $L_m(j) = \text{the average number of customers at server } j$. 
Mean Value Analysis

Solve iteratively:

\[ W_m(j) = \frac{1 + L_{m-1}(j)}{\mu_j} \]

\[ L_m(j) = \lambda_m(j)W_m(j), \text{ where } \lambda_m(j) = \pi_j \lambda_m \]

\[ \lambda_m = \frac{m}{\sum_{i=1}^{k} \pi_i W_m(i)} \quad \text{throughput} \]

Begin with \( W_1(j) = \frac{1}{\mu_j} \)
M/G/1
Best combination of tractability & usefulness

• Assumption of Poisson arrivals may be reasonable based on Poisson approximation to binomial distribution
  – many potential customers decide independently about arriving (arrival = “success”),
  - each has small probability of arriving in any particular time interval
• Probability of arrival in a small interval is approximately proportional to the length of the interval – no bulk arrivals
• Amount of time since last arrival gives no indication of amount of time until the next arrival (exponential – memoryless)
M/G/1
Best combination of tractability & usefulness

• Exponential distribution is frequently a bad model for service times
  – memorylessness
  – large probability of very short service times with occasional very long service times

• May not want to use one of the “standard” distributions for service times, either
  – in a real situation, collect data on service times and fit an empirical distribution

• Distributions of number of customers in the system and waiting time depend on service time distribution to be specified
M/G/1
Best combination of tractability & usefulness

- Assumption of Poisson arrivals may be reasonable based on Poisson approximation to binomial distribution
  - many potential customers decide independently about arriving (arrival = “success”),
  - each has small probability of arriving in any particular time interval
- Distributions of number of customers in the system and waiting time depend on service time distribution
M/G/1 Performance

How many customers? How much time?

$S$ is the length of an arbitrary service time (random variable)
$\lambda$ is the arrival rate of customers; define $\rho = \lambda E[S]$ and assume it is $< 1$.

Expected values can be found from generalizing Little’s formula from # customers in the system to amount of work in the system:

An arriving customer brings $S$ time units of work:

The *time average amount of work* in the system ($V$)

$= \lambda \ast$ the *customer average amount of work* remaining in the system
Work Content

$W_Q^*$ is the (random variable) waiting time in queue

Expected amount of work per customer is

$$E \left[ SW_Q^* + \int_0^S (S - x) \, dx \right]$$

Work remaining

$$= E \left[ SW_Q^* \right] + \frac{E \left[ S^2 \right]}{2}$$

If a customer’s service time

is independent of own wait in queue, get average work in system

$$V = \lambda E \left[ S \right] E \left[ W_Q^* \right] + \frac{\lambda E \left[ S^2 \right]}{2}$$
Mean waiting time

\( W_Q = \) Customer mean waiting time = average work in the system when a customer arrives

From PASTA, \( W_Q = V \). Therefore, (Pollaczek-Khintchine formula)

\[
W_Q = \lambda E[S]W_Q + \frac{\lambda E[S^2]}{2} \Rightarrow W_Q = \frac{\lambda E[S^2]}{2(1 - \lambda E[S])}
\]

And the other measures of performance are:

\[
L_Q = \lambda W_Q = \frac{\lambda^2 E[S^2]}{2(1 - \lambda E[S])}, W = W_Q + E[S], L = \lambda W
\]
Priority Queues

Different types of customers may differ in importance.

- Type $i$ customers arrive according to a Poisson process with rate $\lambda_i$ and require service times with distribution $G_i$, $i = 1, 2$.
- Type 1 customers have (nonpreemptive) priority:
  - service does not begin on a type 2 customer if there is a type 1 customer waiting.
  - If a type 1 customer arrives during a type 2 service, the service is continued to completion.

What is the average wait in queue of a type $i$ customer, $W_Q^i$
Two customer types w/o priority

M/G/1 model with \( \lambda = \lambda_1 + \lambda_2 \) \( G(x) = \frac{\lambda_1}{\lambda} G_1(x) + \frac{\lambda_2}{\lambda} G_2(x) \)

Average work in system is

\[
V = \frac{\lambda E[S^2]}{2(1 - \lambda E[S])} = \frac{\lambda \left( (\lambda_1 / \lambda) E[S_1^2] + (\lambda_2 / \lambda) E[S_2^2] \right)}{2 \left( 1 - \lambda \left( (\lambda_1 / \lambda) E[S_1] + (\lambda_2 / \lambda) E[S_2] \right) \right)}
\]

\[
= \frac{\lambda_1 E[S_1^2] + \lambda_2 E[S_2^2]}{2 \left( 1 - \lambda_1 E[S_1] - \lambda_2 E[S_2] \right)}
\]

If the server is not allowed to be idle when the system is not empty, this quantity is the same for the system with priority.
Two customer types with priority

Let $V^i$ be the average amount of type $i$ work in the system

$$V^i = \lambda_i E[S_i] W_Q^i + \frac{\lambda_i E[S_i^2]}{2}$$

in queue  in service

$$V_Q^i, V_S^i$$

Now focus on a type 1 customer. *Waiting time* = *amt. of type 1 work in system* + *amt. of type 2 work in service* when this customer arrives, so

$$W_Q^1 = V^1 + V_S^2 = \lambda_1 E[S_1] W_Q^1 + \frac{\lambda_1 E[S_1^2]}{2} + \frac{\lambda_2 E[S_2^2]}{2}$$
Two customer types with priority

\[ W_Q^1 = \frac{\lambda_1 E[S_1^2] + \lambda_2 E[S_2^2]}{2(1 - \lambda_1 E[S_1])} \quad \text{if } \lambda_1 E[S_1] < 1 \]

But a type 2 customer has to wait for everyone ahead, plus any type 1 customers who arrive during the type 2 wait, so

\[ W_Q^2 = V + \lambda_1 E[S_1]W_Q^2 \Rightarrow W_Q^2 = \frac{\lambda_1 E[S_1^2] + \lambda_2 E[S_2^2]}{2(1 - \lambda_1 E[S_1] - \lambda_2 E[S_2])(1 - \lambda_1 E[S_1])} \]

\[ \text{if } \lambda_1 E[S_1] + \lambda_2 E[S_2] < 1 \]
M/M/k Model

$k$ identical machines in parallel, Poisson arrivals (rate $\lambda$), exponential service times (rate $\mu$)
CTMC (birth-death) model:
M/M/k Steady-State Probabilities

Level-crossing equations:

\[ \lambda P_n = (n + 1) \mu P_{n+1}, n = 0, 1, \ldots, k - 1 \]
\[ \lambda P_n = k \mu P_{n+1}, n = k, k + 1, \ldots \]

Define \( \rho = \frac{\lambda}{k \mu} \), solve for \( P_n \) in terms of \( P_0 \)

\[
P_n = \begin{cases} 
\frac{(k \rho)^n}{n!} P_0, & n = 0, 1, \ldots, k \\
\frac{k^n \rho^n}{k!} P_0, & n = k + 1, k + 2, \ldots 
\end{cases}
\]

Then use the facts that \( \sum_{n=0}^{\infty} P_n = 1 \), \( \sum_{n=0}^{\infty} r^n = (1 - r)^{-1} \) if \( r < 1 \)

to get

\[
P_0 = \left\{ \sum_{n=0}^{k-1} \frac{(k \rho)^n}{n!} + \frac{(k \rho)^k}{(1 - \rho) k!} \right\}^{-1} \quad \text{if } \rho < 1 \quad \rho \text{ is the utilization of each server}
\]
M/M/k Performance Measures

Steady-state expected number of customers in the system

\[ L = \sum_{n=0}^{\infty} nP_n = k \rho + \left\{ \frac{(k \rho)^k}{k!} \left( \frac{\rho}{(1-\rho)^2} \right) \right\} P_0, \text{ if } \rho < 1 \]

Mean flow time

\[ W = \frac{L}{\lambda} = \frac{1}{\mu} + \left( \frac{1}{k \mu - \lambda} \right) \left( \frac{(k \rho)^k}{k!(1-\rho)} \right) P_0 \text{ by Little's Formula} \]

Expected waiting time

\[ W_Q = W - \frac{1}{\mu} \]

Expected number in the queue (\( k \rho \) is expected number of busy servers)

\[ L_Q = \lambda W_Q = L - k \rho \]
Erlang Loss System

M/M/k/k system: k servers and a capacity of k: an arrival who finds all servers busy does not enter the system (is lost)

\[ P_n = \frac{(k \rho)^n}{n!} P_0, n = 0, 1, \ldots, k \]

\[ P_i = \frac{(k \rho)^i}{i!} \left/ \sum_{n=0}^{k} \frac{(k \rho)^n}{n!} \right., i = 0, \ldots, k \]

Above is called Erlang’s loss formula, and it holds for M/G/k/k as well, if \( k \rho = \lambda E[S] \)