**M/M/c Queues**

- c servers rather than 1
- Exponential inter-arrival and service times

**Examples:**
- multiprocessors
- multiple servers

**Basic Assumptions**
- c servers, identical service rate \( \mu \)
- single queue for waiting customers
- arrival rate \( \lambda \)

**Birth-Death process applies**

\[ \lambda_n = \lambda \text{ for all } n \]

\[ M_n = \begin{cases} n \mu & n = 0, 1, \ldots, c \\ m \mu & n > c \end{cases} \]

0 \[\rightarrow\] 1 \[\xrightarrow{\mu} 2 \]

\[ j \xrightarrow{\mu} (j-1) \]

\[ j \xrightarrow{\mu} (j+1) \]

\[ j \gg c \]
$m/m/c$ Queues

From Birth-Death process solution

$$\Pi_n = \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} \quad n = 1, 2, \ldots, \infty$$

with $\rho = \frac{\lambda}{\mu}$

(Note: usage of $\rho$ varies between $\rho = \frac{\lambda}{\mu}$ and $\rho = \frac{\lambda}{c \mu}$)

Given

$$\Pi_n = \begin{cases} 
\frac{\rho^n}{n!} & \text{if } n < c \\
\frac{\rho^n}{c^n c^{n-c}} \cdot \Pi_0 & \text{if } c \leq n
\end{cases}$$

where

$$\Pi_0 = \left[ \frac{\rho^c}{c! \left(1 - \frac{\rho}{c}\right)} + \sum_{n=0}^{c-1} \frac{\rho^n}{n!} \right]^{-1}$$
Erlang C formula

Phone system modeled as M/M/c queue gives probability an arriving call waits

\[ E(c, \rho) = \sum_{n=0}^{\infty} \frac{\rho^n}{c^n n!} = \prod_{n=1}^{\infty} \frac{\rho^n}{c^n n!} = \prod_{n=0}^{\infty} \frac{\rho^n}{c^n n!} \]

\[ = \prod_{n=0}^{\infty} \frac{\rho^n}{c^n n!} \cdot \frac{1}{1 - \frac{\rho}{c}} = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]

\[ = \frac{\rho^c}{c!} \cdot \frac{1}{1 - \frac{\rho}{c}} \]
M/M/1c Queues

Mean number $\bar{N}$ in system

$$\bar{N} = \rho + \frac{\rho}{c} \cdot \frac{E(c_e)}{1 - \frac{\rho}{c}}$$

Expected number of jobs in service

$$\bar{n}_s = \sum_{n=1}^{c} n \pi_n + \sum_{n=c}^{\infty} c \cdot \pi_n = c \frac{\lambda}{\mu}$$

Mean response time, use Little's Law

$$\bar{R} = \frac{1}{\lambda} + \frac{1}{\mu c} \cdot \frac{E(c_e)}{1 - \frac{\rho}{c}}$$
Comparison between $M/M/1$ & $M/M/c$

Consider 2 ways of organizing servers

- $\lambda$ split
- $\frac{1}{2} \lambda$ $\frac{1}{2} \lambda$ $\frac{1}{2} \lambda$
- $0 0 ... 0$
- $c$
- $\lambda$
- $0 0 ... 0$
- $c$

- Is there a difference, e.g. in time in system?

YES!

Note it is possible on left to have free server and job in queue implies ....?
Comparison between $m/m/1$ & $m/m/c$

mean response time

c $m/m/1$ queues

$$\bar{t} = \frac{1}{\mu - \lambda/c}$$

1 $m/m/m$ queue

$$\bar{t} = \frac{1}{\lambda} + \frac{1}{\mu c} \cdot \frac{E(s|e)}{(1 - \frac{p}{q})}$$
Use of formulae to assess system behavior

Question:
- What performance for a given set of parameters?
- What parameters needed to achieve particular performance goals?

Examples:
- What happens to mean response time if \( \lambda \) doubles?
- What happens to mean response time if we double # servers
- How many servers do we need for given \( \lambda, \mu \) to achieve target mean response time
Example

Observation of jobs arriving to server with $c = 10$ CPUs is that

$\lambda = 5$

$\frac{1}{\mu} = 1$

Use m/m/c formula to estimate mean number in $Q$
and mean response time

$N_q = 0.036 \quad \bar{R} = 1.00$

virtually instantaneous response

Suppose arrival rate increases to $\lambda = 8$ after a
promotion campaign

$N_q = 1.63 \quad \bar{R} = 1.20$

marked relative increase in $N_q$, only 20%
increase in $\bar{R}$
Example \( \mu = 1, \quad \lambda = 3.9 \)

Find smallest number of servers \( c \) s.t. time in queue is less than 20\% of time in service

<table>
<thead>
<tr>
<th>( c )</th>
<th>( R_q )</th>
<th>( \frac{1}{\mu} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>9.45</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0.47</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0.12</td>
<td>1</td>
</tr>
</tbody>
</table>

Just evaluate at different \( c \)'s.
Finite buffers, at most $B$ jobs in system

Using birth-death process framework

$$
\lambda_n = \lambda \quad \text{for } n=0, 1, \ldots, B
$$

$$
\mu_n = \begin{cases} 
    n \mu & \text{for } n < c \\
    c \mu & \text{for } n = c, c+1, \ldots, B 
\end{cases}
$$

**NOTE:** only makes sense for $B \geq c$ else servers idled
State occupancy probs, $\rho = \lambda / \mu$

$$\Pi_n = \begin{cases} 
\frac{\rho^n}{n!} \cdot \Pi_0 & n = 1, 2, \ldots, c-1 \\
\frac{\rho^n}{c!c^{n-c}} \cdot \Pi_0 & n = c, c+1, \ldots, B
\end{cases}$$

With

$$\Pi_0 = \left\{ \left[ 1 - \left( \frac{\rho}{c} \right)^{B-c-1} \right] + \frac{(1 - (\rho / c)) e^c}{m! (1 - e/\rho)} + \sum_{n=1}^{c-1} \frac{\rho^n}{n!} \right\}^{-1}$$

Use these to compute:
- Mean response time
- Mean queue lengths
- "Effective" arrival rate
Given λ, "effective" arrival rate is rate of arrivals when they are accepted — i.e., not in state B
\[ \tilde{\lambda} = \lambda \cdot (1 - \Pi_B) \]

Use effective arrival rate with Little's Law to get mean response time
\[ \bar{\tau} = \frac{\bar{n}}{\tilde{\lambda}} = \frac{\bar{n}}{\lambda (1 - \Pi_B)} \]

Utilization of a server is
\[ \frac{\tilde{\lambda}}{\mu} = \rho (1 - \Pi_B) \]
Other Queues of Interest

Exact solutions for

\( M/Er/1 \) \( Er \) is sum of \( r \) exponentials

\( M/m/1 \) with bulk arrivals

\( M/D/1 \)

\( G/m/1 \)

\( M/G/1 \)

\( G/G/1 \)
Problem — at instant when job arrives, a job in service has state
- unlike exponential, distribution of remaining service depends on length of time in service already

FINESSE
- define state transitions at embedded points where there is no memory
  DEPARTURE INSTANTS

If job has been in service for time , how many arrivals from Markovian arrival process?

Poisson ( )

Consider embedded DTMC
Consider \# jobs left in queue at the point of a job’s departure. This will be state of DTMC:

From state $i > 0$

- transition to state $j$ if $j-i+1$ arrivals during the service time of the customer entering service.

Example: At departure of $n^{th}$ customer,

\[
\begin{array}{c|c|c|c}
\text{J}_{n+2} & \text{J}_{n+1} & \text{J}_n \\
\hline
3 & \rightarrow & 5
\end{array}
\]

$\text{J}_{n+3}$, $\text{J}_{n+4}$, $\text{J}_{n+5}$ arrived.

From state $i = 0$

- transition to $j$ if $j$ arrivals during service time of next arrival.
$M/G/1$

$\alpha_i = \text{Prob } i \text{ arrivals in a service time}$

Given service time $s$, probability of $i$ arrivals is

$\text{Prob a Poisson (}\lambda s)\text{ has value } i$

$\alpha_i = \int_0^\infty g(s) \frac{(\lambda s)^i}{i!} e^{-\lambda s} \, ds$

where $g(s)$ is pdf of service time distribution

e.tc.
\( M/\infty/1 \)

Solution of \( \kappa_i \) gives distribution of queue length at job departure

Mean queue length

\[ \bar{q} = \sum_{i=0}^{\infty} i \cdot \kappa_i \]

Magic bit. Limiting mean queue length at departures is limiting mean queue length!
Pollaczek-Khinchin formula for mean queue length

\[ M/G/1 \]

Define

- \( C_n \): \( n^{th} \) customer to enter system
- \( T_n \): Arrival time of \( n^{th} \) customer
- \( t_n \): \( T_n - T_{n-1} \)
- \( X_n \): Service time for \( n^{th} \) customer
- \( q_n \): Number left in queue by departure of \( C_n \)
- \( V_n \): Number of customers that arrive while \( C_n \) is receiving service
Formula for $q_{n+1}$

Case $q_n > 0$

$$q_{n+1} = q_n - 1 + v_{n+1}, \text{ when } q_n > 0$$
Case $q_n = 0$

$M/G/1$

$\text{SERVER}$

$\text{QUEUE}$

$q_n = 0$

$9_{n+1} \text{ left behind}$

$9_{n+1} = V_{n+1}$

$q_n = 0$
Combine expressions

\[ q_{n+1} = \begin{cases} 
q_n - 1 + V_{n+1} & \text{if } q_n > 0 \\
V_{n+1} & \text{otherwise}
\end{cases} \]

Define SHIFTED DISCRETE STEP FUNCTION

\[ \Delta_k = \begin{cases} 
1 & \text{for } k = 1, 2, \ldots \\
0 & \text{for } k = 0
\end{cases} \]

and re-write

\[ q_{n+1} = q_n - \Delta_q + V_{n+1} \]

Note: In limit as \( n \to \infty \), \( E[q_{n+1}] = E[q_n] \)
\[ n_{t+1} = n_t - \Delta n_t + V_{n+1} \]

\[ \Rightarrow E[n_{n+1}] = E[n_t] - E[\Delta n_t] + E[V_{n+1}] \]

\[ \Rightarrow \lim_{n \to \infty} E[n_{n+1}] = \lim_{n \to \infty} E[n_t] - E[\Delta n_t] + E[V_{n+1}] \]

\[ \Rightarrow \lim_{n \to \infty} E[\Delta n_t] = \lim_{n \to \infty} E[V_{n+1}] \]

We can compute

\[ \lim_{n \to \infty} E[V_{n+1}] = \int_0^\infty g(s) \sum_{k=0}^\infty \frac{(\lambda s)^k}{k!} e^{-\lambda s} ds \]

Mean of Poisson \( \lambda \)

\[ = \int_0^\infty g(s) \lambda s \]

\[ = \lambda \cdot E[\text{service time}] \]
\[
\lim_{n \to \infty} \mathbb{E}[\Delta_{q_n}] = \mathbb{E}[\Delta_{\tilde{q}}]
\]

\[
= \sum_{k=0}^{\infty} \Delta_k \mathbb{P}[\tilde{q} = k]
\]

\[
= \Delta_0 \mathbb{P}[\tilde{q} = 0] + \Delta_1 \mathbb{P}[\tilde{q} = 1] + \ldots.
\]

Recall definition of \( \Delta_i \):

\[
= 0 \cdot \mathbb{P}(\tilde{q} = 0) + 1 \cdot \mathbb{P}(\tilde{q} > 0)
\]

\[
= \mathbb{P}(\tilde{q} > 0)
\]

\[
= \mathbb{P}(\text{busy system})
\]

\[
= \lambda
\]

\[
= \mathbb{E}[\text{service time}]
\]

\[
= \rho
\]
 BUT we didn't get to \( E[\tilde{q}] \)!

From \( q_{n+1} = q_n - \Delta q_n + V_{n+1} \)

desire

\[
q_{n+1}^2 = (q_n - \Delta q_n + V_{n+1})^2
\]

\[
= q_n^2 + \Delta q_n^2 + V_{n+1}^2 - 2q_n \Delta q_n + 2q_n V_{n+1} - 2\Delta q_n V_{n+1}
\]

Note that \( \Delta_i^2 = \Delta_i \) and \( q_n \Delta_n = q_n \)

which brings us to

\[
E[q_{n+1}^2] = E[q_n^2] + E[\Delta q_n] + E[V_{n+1}^2] - 2E[q_n] + 2E[q_n V_{n+1}] - 2E[\Delta q_n V_{n+1}]
\]

Note that \( q_n \) and \( V_{n+1} \) are independent

\( \Delta q_n \) and \( V_{n+1} \) are independent

\[
E[q_{n+1}^2] = E[q_n^2] + E[\Delta q_n] + E[V_{n+1}^2] - 2E[q_n] + 2E[q_n]E[V_{n+1}] - 2E[\Delta q_n]E[V_{n+1}]
\]
leads to

$$0 = E[\Delta g] + E[v^2] - 2E[g] + 2E[g]E[v] - 2E[\Delta g]E[v]$$

but we've seen already that $E[\Delta g] = E[v] = \rho$

which brings us to

$$E[g] = \rho + \frac{E[v^2]}{2(1-\rho)}$$

the only unknown is $E[v^2]$

Aside: 2nd moment of Poisson $\lambda S$ is $(\lambda S)^2 + \lambda S$

(from variance = second moment - mean$^2$)

Then

$$E[v^2] = \int_0^\infty g(s) \cdot E[A(\lambda, s)^2] \, ds$$

$$= \int_0^\infty g(s)[(\lambda s)^2 + \lambda s] \, ds$$

$$= \lambda^2 E[S^2] + \lambda E[S]$$

$S$ the service time
\[ \bar{q} = \rho + \rho^2 \frac{1 + C_b^2}{Z(1-\rho)} \]

\[ \text{Pollaczek-Khinchin mean value formula} \]

\[ \bar{q} = \rho + \rho^2 \frac{\lambda \text{var}(S) + \lambda \text{E}[S] - \rho}{Z(1-\rho)} \]

\[ \text{let } C_b \text{ be coefficient of variation of service time dist.} \]

\[ C_b = \frac{\text{var}(S)^{\frac{1}{2}}}{\text{E}[S]} \quad \text{var}(S) = \text{E}[S^2] - (\text{E}[S])^2 \]

BUNCH OF ALGEBRA LATER:
$M/G/1$

\[ \hat{q} = p + p^2 \frac{(1 + C_b^2)}{2(1-p)} \]

**Example**: Exponential service time: $C_b = 1$

\[ \hat{q} = p + p^2 \frac{2}{2(1-p)} = p + \frac{p^2}{2(1-p)} = \frac{2p(1-p) + p^2}{2(1-p)} = \frac{2p}{2(1-p)} = \frac{p}{1-p} \]

Deterministic $\hat{q}$: $C_b^2 = 0$

\[ \hat{q} = p + \frac{p^2}{2(1-p)} = \frac{p}{1-p} - \frac{p^2}{2(1-p)} \]

So $M/D/1$ has fewer customers on average than $M/M/1$. 