Advances of simple queues

• Queues with priority
• Retrial Queue
Queues with priority

• For single server queue
  – Non-preemptive queue with many classes
  – Preemptive queue with many classes
  – General service time M/G/1
  – FCFS M/M/1 with 2 classes
Non-preemptive queue with many classes

- Non-preemptive M/M/1 with 2 classes in Lecture5
- A general case, with many classes, k=1,...,r, the priority 1>2>...>r.
- Mean value analysis to find the average metrics
Non-preemptive M/M/1 with r classes

• Parameters of system
  – Poisson arrival rate $\lambda_k$, $k=1,...,r$
  – Exponential service time $\mu_k$, $k=1,...,r$
  – Priority $1>2>...>r$

• Define
  $$\rho_k = \frac{\lambda_k}{\mu_k}, \quad \sigma_k = \sum_{i=1}^{k} \rho_i \quad (\sigma_0 = 0, \sigma_r = \rho)$$
  – $\sigma_r = \rho < 1$ for stable system
Non-preemptive M/M/1 with r classes

• Consider a customer of priority i arrives, with PASTA theorem, we have

\[
W_q^{(i)} = \sum_{k=1}^{i-1} E[S'_k] + \sum_{k=1}^{i} E[S_k] + E[S_0^{(i)}]
\]

Note: You can also apply PASTA to compute \(T^{(i)}\), obtain the same recursive equation

• With Little’s law,

\[
E[S_k] = \frac{\lambda_k W_q^{(k)}}{\mu_k} = \rho_k W_q^{(k)}
\]

\[
E[S'_k] = \frac{\lambda_k W_q^{(i)}}{\mu_k} = \rho_k W_q^{(i)} \quad (k < i)
\]

• So,

\[
W_q^{(i)} = W_q^{(i)} \sum_{k=1}^{i-1} \rho_k + \sum_{k=1}^{i} \rho_k W_q^{(k)} + E[S_0^{(i)}]
\]
Non-preemptive M/M/1 with $r$ classes

- We have \[ (1 - \sigma_{i-1})W_q^{(i)} = \sum_{k=1}^{i} \rho_k W_q^{(k)} + E[S_0^{(i)}] \]
  also, \[ (1 - \sigma_i)W_q^{(i)} = \sum_{k=1}^{i-1} \rho_k W_q^{(k)} + E[S_0^{(i)}] \]

- Replace $i$ with $i-1$ in the first equation, we have \[ (1 - \sigma_{i-2})W_q^{(i-1)} = \sum_{k=1}^{i-1} \rho_k W_q^{(k)} + E[S_0^{(i-1)}] \]

- Combining two equations, we obtain \[ (1 - \sigma_i)W_q^{(i)} = (1 - \sigma_{i-2})W_q^{(i-1)} - E[S_0^{(i-1)}] + E[S_0^{(i)}] \]
Non-preemptive M/M/1 with r classes

• Since queue is non-preemptive, all the same
  \[ E[S_0^{(i)}] = E[S_0^{(j)}] = E[S_0] \]

• So, \( (1 - \sigma_i)W_q^{(i)} = (1 - \sigma_{i-2})W_q^{(i-1)} \), for \( i > 1 \)

• Initial condition, \( i=1 \), we have
  \[ (1 - \sigma_1)W_q^{(1)} = E[S_0^{(1)}] \Rightarrow W_q^{(1)} = \frac{E[S_0]}{1 - \sigma_1} \]

• By induction, we have
  \[ W_q^{(i)} = \frac{E[S_0]}{(1 - \sigma_{i-1})(1 - \sigma_i)} \]
Residual service time $E[S_0]$

- In general, we have

$$E[S_0] = \Pr\{\text{busy}\}E[S_0|\text{busy}] = \rho E[S_0|\text{busy}]$$

- further

$$E[S_0|\text{busy}] = \sum_{k=1}^{r} \frac{1}{\mu_k} \frac{\rho_k}{\rho} \implies E[S_0] = \sum_{k=1}^{r} \frac{\rho_k}{\mu_k}$$

- So,

$$W_q^{(i)} = \frac{1}{(1-\sigma_{i-1})(1-\sigma_i)} \sum_{k=1}^{r} \frac{\rho_k}{\mu_k} \quad L_q^{(i)} = \lambda_i W_q^{(i)}$$

- $$L_q = \sum_{i=1}^{r} L_q^{(i)}, \quad W_q = \sum_{i=1}^{r} \frac{\lambda_i}{\lambda} W_q^{(i)}$$

- Question: what’s the relation with FCFS M/M/1?

- how to assign the priority to improve system performance?

  High priority to high service rate class, to make $\sigma_1$ small
Extension, non-preemptive M/G/1 with r classes

• All the analysis remains the same, except
  – The residual service time $E[S_0]$

• area of triangles: $E[X_k^2]/2 = E[S_0|busy]E[X_k]$

$$E[S_0|busy \text{ type } k] = \frac{E[X_k^2]}{2E[X_k]}$$

- $E[S_0|busy]$: average height of triangle
- $E[X]$: the average length of triangle
- $E[X^2]/2$: the average area of triangle
Extension, non-preemptive M/G/1 with r classes

- Mean residual service time is

\[
E[S_0] = \rho \sum_{k=1}^{r} \frac{\rho_k}{\rho} \frac{E[X_k^2]}{2E[X_k]} = \frac{\lambda}{2} \sum_{k=1}^{r} \frac{\lambda_k}{\lambda} E[X_k^2] = \frac{\lambda E[S^2]}{2}
\]

- Mean waiting time of type i customers is

\[
W_q^{(i)} = \frac{\lambda E[S^2] / 2}{(1-\sigma_{i-1})(1-\sigma_i)}
\]

where \(E[S^2] = \sum_{k=1}^{r} \frac{\lambda_k}{\lambda} E[X_k^2]\)

2\textsuperscript{nd} moment of service time of customers on average
2-class M/M/1 FCFS without priority

• Two classes customers
  – Arrival rate: $\lambda_1, \lambda_2$
  – Service rate: $\mu_1, \mu_2$
  – No priority, FCFS

• M/H$_2$/1 queue
  – Arrival rate: $\lambda = \lambda_1 + \lambda_2$
  – Service rate: hyperexponential distribution
    • Choose $\mu_1$ with probability $\lambda_1/\lambda$
    • Choose $\mu_2$ with probability $\lambda_2/\lambda$
2-class M/M/1 FCFS without priority

- Utilization factor $\rho_1 = \frac{\lambda_1}{\mu_1}, \quad \rho_2 = \frac{\lambda_2}{\mu_2}, \quad \rho = \rho_1 + \rho_2$
  - $\rho_1$: ratio that server is busy in serving customer 1
  - $\rho_2$: ratio that server is busy in serving customer 2
  - $\rho$: ratio that server is busy

- Question, why?
  - Probability of customer type in the queue?
    * $\lambda_1/\lambda$
  - Probability of customer type in the busy server?
    * $\rho_1/\rho$
2-class M/M/1 FCFS without priority

- Equivalent service rate
  \[ 1/\mu = \lambda_1/\lambda^* 1/\mu_1 + \lambda_2/\lambda^* 1/\mu_2 \]

- PASTA theorem
  \[ W_q = L_q \frac{1}{\mu} + \rho_1 \frac{1}{\mu_1} + \rho_2 \frac{1}{\mu_2} = \lambda W_q \frac{1}{\mu} + \rho_1 \frac{1}{\mu_1} + \rho_2 \frac{1}{\mu_2} \]
  - So,
    \[ W_q = \frac{\rho_1/\mu_1 + \rho_2/\mu_2}{1 - \rho} \]
  - \( W_q \) is identical for different type of customers, while \( W \) is different

\[ L^{(1)}_q = \lambda_1 W_q, \quad L^{(2)}_q = \lambda_2 W_q, \quad L_q = \lambda W_q = L^{(1)}_q + L^{(2)}_q \]

\[ W^{(1)} = W_q + \frac{1}{\mu_1}, \quad W^{(2)} = W_q + \frac{1}{\mu_2}, \quad W = W_q + \frac{\lambda_1}{\lambda} \frac{1}{\mu_1} + \frac{\lambda_2}{\lambda} \frac{1}{\mu_2} = \frac{\lambda_1}{\lambda} W^{(1)} + \frac{\lambda_2}{\lambda} W^{(2)} \]

\[ L^{(i)} = \lambda_i W^{(i)} = L^{(i)}_q + \rho_i, \quad (i = 1, 2) \]

\[ L = \lambda W = L^{(1)} + L^{(2)} \]
Comparison with M/M/1 with equivalent service rate

• 1-class M/M/1
  – Same arrival rate $\lambda$
  – Equivalent service rate
    • $1/\mu = \lambda_1/\lambda * 1/\mu_1 + \lambda_2/\lambda * 1/\mu_2$

• Comparison
  – 2-class M/M/1 has worse performance metrics, why?
    • Higher variability of service time than exponential
    • Hyperexponential ($c_x > 1$) v.s. exponential ($c_x = 1$)
2-class M/M/1 with non-preemptive priority and identical service rates

• We discussed before in Lecture 5
  – Identical service rates:

\[ W^{(1)} = \frac{(1 + \rho_2) / \mu}{1 - \rho_1}, \quad L^{(1)} = \frac{(1 + \rho_2) / \rho_1}{1 - \rho_1} \]

\[ W^{(2)} = \frac{[1 - \rho_1(1 - \rho_1 - \rho_2)] / \mu}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}, \quad L^{(2)} = \frac{[1 - \rho_1(1 - \rho_1 - \rho_2)] \rho_2}{(1 - \rho_1)(1 - \rho_1 - \rho_2)} \]
2-class M/M/1 with non-preemptive priority and different service rates

- Different service rates
  \[ \rho_1 = \frac{\lambda_1}{\mu_1}, \quad \rho_2 = \frac{\lambda_2}{\mu_2}, \quad \rho = \rho_1 + \rho_2 \]

- For type 1 customer, PASTA theorem
  \[ W_q^{(1)} = L_q^{(1)} \frac{1}{\mu_1} + \rho_1 \frac{1}{\mu_1} + \rho_2 \frac{1}{\mu_2} = \lambda_1 W_q^{(1)} \frac{1}{\mu_1} + \rho_1 \frac{1}{\mu_1} + \rho_2 \frac{1}{\mu_2} \]
  So,
  \[ W_q^{(1)} = \frac{\rho_1 / \mu_1 + \rho_2 / \mu_2}{1 - \rho_1} \]

  \[ L_q^{(1)} = \lambda_1 W_q^{(1)}, \quad W^{(1)} = W_q^{(1)} + \frac{1}{\mu_1}, \quad L^{(1)} = \lambda_1 W^{(1)} \]
2-class M/M/1 with non-preemptive priority and different service rates

- For type2 customers, PASTA theorem

\[ W_{q}^{(2)} = L_{q}^{(2)} \frac{1}{\mu_2} + L_{q}^{(1)} \frac{1}{\mu_1} + \lambda_1 W_{q}^{(2)} \frac{1}{\mu_1} + \rho_1 \frac{1}{\mu_1} + \rho_2 \frac{1}{\mu_2} \]

\[ = \rho_2 W_{q}^{(2)} + \rho_1 W_{q}^{(2)} + L_{q}^{(1)} \frac{1}{\mu_1} + \rho_1 \frac{1}{\mu_1} + \rho_2 \frac{1}{\mu_2} \]

- Substitute \( L_{q}^{(1)} \), we have

\[ W_{q}^{(2)} = \frac{\rho_1 / \mu_1 + \rho_2 / \mu_2}{(1 - \rho_1)(1 - \rho)} \]

\[ L_{q}^{(2)} = \lambda_2 W_{q}^{(2)}, \quad W^{(2)} = W_{q}^{(2)} + \frac{1}{\mu_2}, \quad L^{(2)} = \lambda_2 W^{(2)} \]

\[ L_{q} = L_{q}^{(1)} + L_{q}^{(2)}, \quad L = L^{(1)} + L^{(2)}, \quad W_{q} = \frac{\lambda_1}{\lambda} W_{q}^{(1)} + \frac{\lambda_2}{\lambda} W_{q}^{(2)}, \quad W = \frac{\lambda_1}{\lambda} W^{(1)} + \frac{\lambda_2}{\lambda} W^{(2)} \]

Li Xia, Tsinghua Univ.
Comparison of 4 non-preemptive models

• Refer to pp.149 of Gross’ book
  – (a). M/M/1, equivalent service rate
  – (b). Two priorities, equal service rates
  – (c). Two priorities, two service rates
  – (d). No priority, two service rates

• Parameter setting
  • \( \lambda = \lambda_1 + \lambda_2 \)
  • \( \frac{1}{\mu} = \frac{\lambda_1}{\lambda} \frac{1}{\mu_1} + \frac{\lambda_2}{\lambda} \frac{1}{\mu_2} \)
Comparison of 4 non-preemptive models

<table>
<thead>
<tr>
<th>Versus</th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(b)</td>
<td>$L_q: (a)=(b)$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| (c)    | $L_q: (c) < (a)$  
  $\text{iff } \mu_1 > \mu_2$ | $L_q: (c) < (b)$  
  $\text{iff } \mu_1 > \mu_2$ |   |
| (d)    | $L_q: (d) \geq (a)$ | $L_q: (d) \geq (b)$ | $L_q: (c) < (d)$  
  $\text{iff } \mu_1 > \mu_2$ |

- (a). M/M/1, equivalent service rate $\mu$
- (b). Two priorities, equal service rate $\mu$
- (c). Two priorities, two service rates $\mu_1$ and $\mu_2$
- (d). No priority, two service rates $\mu_1$ and $\mu_2$
Retrial queue

• Important queueing model in practice
  – Call center, e-ticket booking, etc.

Primary arrival → Service facility → Customer leaving

Rearrival from orbit → Orbit

Balking customer

Li Xia, Tsinghua Univ.
Retrial queue

• Complicated queueing model
  – If orbit waiting time is 0, equal to M/M/1 with random service discipline
  – Markovian model
    • Waiting time in orbit is exponential
    • No impatient customer
M/M/1 retrial queue

• Parameters
  – Arrival rate $\lambda$, service rate $\mu$, orbit retrial rate $\gamma$
  – No impatient customer
  – System state $(i,n)$: $i \in \{0,1\}$, # of customer in service; $n \in \{0,1,2,\ldots\}$, # of customers in orbit;
M/M/1 retrial queue

- State transition rate diagram

1: server busy

0: server idle

- Global balance eq.

\[(\lambda + n\gamma) p_{0,n} = \mu p_{1,n}, \quad (n \geq 0)\]

\[(\lambda + \mu) p_{1,n} = \lambda p_{0,n} + (n+1)\gamma p_{0,n+1} + \lambda p_{1,n-1}, \quad (n \geq 1)\]

\[(\lambda + \mu) p_{1,0} = \lambda p_{0,0} + \gamma p_{0,1}\]
Average metrics of M/M/1 retrial queue

• Use Z-transform to calculate the distribution

• Average performance metrics

\[ L_o = \frac{\rho^2}{1-\rho} \frac{\mu + \gamma}{\gamma} \quad W_o = \frac{\rho}{\mu - \lambda} \frac{\mu + \gamma}{\gamma} \]

\[ L = \frac{\rho}{1-\rho} \frac{\lambda + \gamma}{\gamma} \quad W = \frac{1}{\mu - \lambda} \frac{\lambda + \gamma}{\gamma} \]

– As \( \gamma \to \infty \), approach to M/M/1 queue!
M/M/1 retrial queue with impatience

- When the customer is denied service, it will
  - Enter orbit to retry with prob. $q$
  - Abandon the system with prob. $1-q$

- Similar analysis
M/M/c retrial queue

• More complicated, use approximation

• Assumption
  – When $\gamma$ is small, retrials see time average
  – Do simplications
Phase-Type distribution

• A generalization of concept of phases

• PH distribution: the time to enter an absorbing state in Markov process
  – PH distribution has two terms, \((\alpha, \mathbf{T})\)
    • \(\alpha\) is the initial distribution
    • \(\mathbf{T}\) is part of the transition probability/rate matrix \(Q\)

• For example, Erlang-2 distribution:

\[
\begin{bmatrix}
  -\mu & \mu & 0 \\
  0 & -\mu & \mu \\
  0 & 0 & 0
\end{bmatrix}
\]

\(\alpha = (1,0)\)
**PH distribution**

- **Coxian distribution**

- **Hyberexponential distribution**

\[
Q = \begin{bmatrix}
-\mu & \mu p_1 & \mu (1 - p_1) \\
0 & -\mu & \mu \\
0 & 0 & 0
\end{bmatrix}
\]

\[\alpha = (1,0)\]

with an initial distribution \(\alpha = (q,1-q)\)
The calculation of pdf of PH distribution

• Use the C-K equation and matrix calculation of Q

• Calculate the transient behavior of absorbing state, \( p_n(x) \).

\[
F(x) = 1 - \alpha e^{Tx} e
\]

– \( p_n(x) = P\{X<x\} \), so \( p_n(x) \) is cdf of PH and \( p'_n(x) \) is pdf.

• Chalk writing, use hyperexponential distribution as example

– For any PH distribution, we can similarly analyze its pdf based on its Q matrix of Markov process