Outline

• Finite source queue $M/M/c//K$
• Queues with impatience (balking, reneging, jockeying, retrial)
• Transient behavior
• Advanced Queue
  – Batch queue
    • Bulk input queue $M[X]/M/1$
    • Bulk service queue $M/M[Y]/1$
  – Erlangian queue
    • Erlangian queue $M/E_k/1$
    • Erlangian queue $E_k/M/1$
Queues with finite source, M/M/c//K queue

- Infinite calling population
  - Arrival rate is constant, $\lambda$
  - Example, the outpatient of hospital, passengers of bus station, etc.

- Finite calling population
  - Arrival rate is variant
  - Example, $K$ machines, breakdown rate $\lambda$, service rate of maintenance is $\mu$, # in maintenance is $n$, arrival rate is $(K-n)\lambda$
M/M/c///K

- Population size: K
- Arrival rate of outside customers: (K-n)λ
- Service rate of each server: μ
- Special case of birth-death process
  - Birth rate: \( \lambda_n = (K-n)\lambda \)
  - Death rate: \( \mu_n = (n \land c)\mu \)
State transition rate diagram

- State transition rate diagram

\[ K\lambda \quad (K-1)\lambda \quad (K-2)\lambda \quad (K-c+1)\lambda \quad (K-c)\lambda \quad (K-c-1)\lambda \quad \lambda \]

\[ 0 \quad 1 \quad 2 \quad \ldots \quad c \quad c+1 \quad \ldots \quad K \]

\[ \mu \quad 2\mu \quad 3\mu \quad c\mu \quad c\mu \quad c\mu \quad c\mu \]

- With local balance equation, we have

\[
\pi_n = \begin{cases} 
\frac{K!}{(K-n)!} \frac{r^n \pi_0}{n!}, & n = 0, 1, \ldots, c-1 \\
\frac{K!}{(K-n)!} \frac{r^n \pi_0}{c^{n-c} c!}, & n = c, c+1, \ldots, K
\end{cases}
\]
Steady state distribution

• Steady state distribution of M/M/c//K

\[ \pi_n = \begin{cases} \binom{K}{n} r^n \pi_0, & n = 0, 1, \ldots, c - 1 \\ \frac{n!}{c^{n-c} c!} r^n \pi_0, & n = c, c + 1, \ldots, K \end{cases} \]

• \( \pi_0 \) can be calculated with the Law of total probability, but the form is complicated
Special case: M/M/1//K

• When c=1, we have

\[ \pi_n = \binom{K}{n} n! \rho^n \pi_0 = \frac{K! \rho^n \pi_0}{(K-n)!} \]

and \[ \pi_0 = \left[ \sum_{n=0}^{K} \frac{K! \rho^n}{(K-n)!} \right]^{-1} \]

• Analysis of other metrics is similar
Average queue length and mean waiting time

- Average number of customers in the system
  \[ L = \sum_{n=0}^{K} n\pi_n \]

- Effective arrival rate
  \[ \lambda_{eff} = \sum_{n=0}^{K} (K - n)\lambda\pi_n = (K - L)\lambda \]

- Average queue length
  \[ L_q = L - \frac{\lambda_{eff}}{\mu} = L - (K - L)r \]

- Mean waiting time and sojourn time
  \[ W = \frac{L_q}{\lambda_{eff}} = \frac{L_q}{(K - L)\lambda} \quad T = \frac{L}{\lambda_{eff}} = \frac{L}{(K - L)\lambda} \]

Li Xia, Tsinghua Univ.
Example: models with spares

- M/M/c//K queue can model the production system with spares (Gross’ book p.88)
  - $n$: number of failed machines
  - $c$: number of technicians to repair machine
  - At most $K$ machines in operation, $Y$ spares
  - Every machine has a failure rate $\lambda$; when fails, if has spares, immediately replace the failed one; when repaired, become spare

- Analyze the queue similar to M/M/c//K
- When $Y \to \infty$, equivalent to M/M/c queue
Queues with impatience

• Customer is not patient during its waiting period
• model for practical situation
  – Call center, crowded service facility, etc.
• types of queues with impatience
  – Queues with balking (not joining when crowded)
  – Queues with reneging (leave the line with probability)
  – Queues with jockeying (moving among multiple lines)
  – Retrial queues (balk when crowded and rejoin the line after a random time)
M/M/1 with Balking

• One example is finite storage queue M/M/c/K
• Balking function is to decrease the arrival rate when system becomes crowded (discouraged)
  – \{b_n\}, b_{n+1} < b_n for all n=0,1,2,...
  – arrival rate: \lambda_n = \lambda b_n
  – for example, b_n = 1/n, or 1/n^2, or e^{-\alpha n}
• Waiting time is more reasonable than queue length, so
  – Estimate the mean service time and set b_n = e^{-\alpha n}/\mu
M/M/1 with reneging

• All arrivals join the queue, but estimate the waiting time and leave the line if with long waiting time

• Reneging function \( r(n) \), total leaving rate at \( n \)
  – \( r(0)=r(1)=0 \)
  – Equivalent to a variant service rates, \( \mu + r(n) \)
  – Similar analysis to birth-death process
  – Example: \( r(n) = e^{\alpha n/\mu}, \ n>1 \)

Li Xia, Tsinghua Univ.
Other queues with impatience

• Queues with jockeying
  – Difficult to analyze, no meaningful results

• Retrial queues
  – Intensively studied, good model for many practical systems, such as call center
  – Particular results with insights to queuing theory
  – Discuss it later
Transient behavior

• Balance equation for equilibrium behavior or steady-state distribution
• Analysis for transient behavior is much more complicated
  – M/M/1/1 and M/M/1, as example
Transient behavior of M/M/1/1

- Differential-difference equation (optional)

\[
\frac{d\pi_0(t)}{dt} = \mu \pi_1(t) - \lambda \pi_0(t)
\]

\[
\frac{d\pi_1(t)}{dt} = -\mu \pi_1(t) + \lambda \pi_0(t)
\]

\[\pi_0(t) + \pi_1(t) = 1\]

- So,

\[\pi_1'(t) + (\lambda + \mu)\pi_1(t) = \lambda\]

- We have

\[\pi_0(t) = \frac{\mu}{\lambda + \mu} + [\pi_0(0) - \frac{\mu}{\lambda + \mu}]e^{-(\lambda + \mu)t}\]

\[\pi_1(t) = \frac{\lambda}{\lambda + \mu} + [\pi_1(0) - \frac{\lambda}{\lambda + \mu}]e^{-(\lambda + \mu)t}\]

- Chalk drawing, curve of probability
Transient behavior of M/M/1

- Differential-difference equation (optional)

$$
\pi'_n(t) = - (\lambda + \mu) \pi_n(t) + \lambda \pi_{n-1}(t) + \mu \pi_{n+1}(t), \quad n > 0
$$

$$
\pi'_0(t) = - \lambda \pi_0(t) + \mu \pi_1(t)
$$

- Z-transform and Laplace-transform

- Sum all of the equations and have

$$
P(z, s) = \frac{z^{i+1} - \mu(1-z)P_0(s)}{(\lambda + \mu + s)z - \mu - \lambda z^2}
$$

- With initial condition $n(0) = i$
Transient behavior of M/M/1

• Reverse transform and obtain (optional)

\[
p_n(t) = e^{-(\lambda+\mu)t} \left[ \rho^{(n-i)/2} I_{n-i}(2t\sqrt{\lambda\mu}) + \rho^{(n-i-1)/2} I_{n+i+1}(2t\sqrt{\lambda\mu}) \right. \\
\left. + (1 - \rho) \rho^n \sum_{j=n+i+2}^{\infty} \rho^{-j/2} I_j(2t\sqrt{\lambda\mu}) \right]
\]

where modified Bessel function of the first kind

\[
I_n(x) := \sum_{k=0}^{\infty} \frac{(y/2)^{n+2k}}{k!(n+k)!}, \quad n > -1
\]
Advanced Markovian queues

• Batch queue
  – Bulk input queue $M^{[X]}/M/1$
  – Bulk service queue $M/M^{[Y]}/1$

• Erlangian queue
  – Erlangian queue $M/E_k/1$
  – Erlangian queue $E_k/M/1$
Bulk input queue $M[X]/M/1$

- Batch arrival, Poisson process, arrival rate $\lambda$
  - number of customers in each arrival is an integer random number $X$
  - Probability of $X$ is $c_n = \text{Pr}[X=n]$
- State transition rate diagram
  - $X=\{1,2\}$ as example, pp.118 of Gross’ book
Balance equation of M^{[X]}/M/1

• Global balance equation

\[ 0 = - (\lambda + \mu) p_n + \mu p_{n+1} + \lambda \sum_{k=1}^{n} p_{n-k} c_k , \quad n \geq 1 \]

\[ 0 = -\lambda p_0 + \mu p_1 \]

• To solve these equations, use Z-transforms of c and p
Balance equation of $M^{[X]}/M/1$

- Multiply $z^n$ on both sides and sum all

$$0 = -\lambda \sum_{n=0}^{\infty} p_n z^n - \mu \sum_{n=1}^{\infty} p_n z^n + \frac{\mu}{z} \sum_{n=1}^{\infty} p_n z^n + \lambda \sum_{n=1}^{\infty} \sum_{k=1}^{n} p_{n-k} c_k z^n$$

- We have

$$0 = -\lambda P(z) - \mu [P(z) - p_0] + \frac{\mu}{z} [P(z) - p_0] + \lambda C(z) P(z)$$

- Thus,

$$P(z) = \frac{\mu p_0 (1-z)}{\mu (1-z) - \lambda z [1-C(z)]}, \quad |z| \leq 1$$
Average metrics of $M^{[X]}/M/1$

- Let $z=1$, obtain the average metrics

$$
p_0 = 1 - \rho, \quad r = \lambda / \mu, \quad \rho = rE[X]
$$

$$
L = \frac{r(E[X] + E[X^2])}{2(1 - \rho)} = \frac{\rho + rE[X^2]}{2(1 - \rho)}
$$

$$
L_q = L - (1 - p_0) = L - \rho
$$

$$
T = \frac{L}{\lambda E[X]}, \quad W_q = \frac{L_q}{\lambda E[X]}
$$
Example of $M^{[X]}/M/1$

- $X$ is a constant $K$, apply the results we have

$$L = \frac{\rho + K\rho}{2(1-\rho)} = \frac{K+1}{2} \frac{\rho}{1-\rho}, \quad \rho = \frac{K\lambda}{\mu}$$

- What’s the difference from $M/M/1$ with $K\lambda$?
- Equivalent to $M/M/1$ with $K\lambda$ and a scale factor $(K+1)/2$
  - Worse performance metrics
  - Why? more bursty
Bulk service queue $M/M^K/1$

• Similar to $M/M/1$, except the server serves $K$ customers at a time, called $M/M^K/1$. $K$ is a constant.

• Two model for different service mode
  – Partial-batch
    • Start service no matter $n<K$
  – Full-batch
    • Start service until $n\geq K$
M/M\(^{[K]}\)/1 v.s. M/M/1

• formulas of metrics are similar
  – Formulas for distribution, L, T, L\(_q\), W\(_q\) are similar

• similar to M/M/1 with service rate K\(\mu\)
  – \(r_0\) is the root of characteristic equation
  – \(\rho = \lambda / K\mu\) is the utilization factor of M/M/1
  – which one is bigger?
    • Guess: \(r_0 > \rho\) (because M/M\(^{[K]}\)/1 is more bursty \(\rightarrow\) more crowded)

  – Metrics of M/M\(^{[K]}\)/1 is worse than those of M/M/1
Erlangian queues and $M/E_k/1$

• Exponential distribution
  – Memoryless property, simplify the analysis
  – Limitation, more general cases

• Erlang distribution for service time and interarrival time
  – Sum of K-stage exponential distributions
  – More general model

• Other distributions
  – PH, Coxian, hyperexponential, etc.
Erlang distribution

- k-stage Erlang distribution, each stage is exponentially distributed with rate $k\mu$
- By convolution, pdf is

$$f(x) = \frac{k\mu(k\mu x)^{k-1}}{(k-1)!} e^{-k\mu x}$$

$$E[X] = \frac{1}{\mu}, \quad Var[X] = \frac{1}{k\mu^2}, \quad c_x = \frac{1}{\sqrt{k}}$$
Erlang distribution

- Distribution with 2-parameters, $c_x < 1$, to model smooth stochastic process
  - Approximate empirical distribution, use the mean and variance to match $\mu$ and $k$
- If $c_x > 1$, we can use other distribution model, such as hyperexponential distribution
M/E\textsubscript{k}/1 queue

• Service time is Erlang-k distribution
  – k stage, each stage is exponential with k\mu

• State definition
  – (n,i): n customers and customer in service is in phase i, i=1,2,\ldots,k. (k is the first phase and 1 is the last phase)

• Global balance equation
Average metrics of M/E\(_k\)/1

- Consider equivalent queue with state (n-1)k+i, # of phase requests

- Equivalent to M\([k]\)/M/1, average # of customer is average # of phase requests

\[
L = \frac{k+1}{2} \frac{\rho}{1-\rho}, \quad \rho = \frac{\lambda}{\mu}
\]

- So,

\[
W_q = \frac{k+1}{2k} \frac{\rho}{\mu(1-\rho)}, \quad L_q = \lambda W_q = \frac{k+1}{2k} \frac{\rho^2}{(1-\rho)}
\]

\[
T = W_q + \frac{1}{\mu}, \quad L = L_q + \rho
\]

Can we use PASTA to do mean value analysis?
Example of M/E\(_k\)/1

• Suppose a queuing system
  – \(\lambda=16/\text{h}\), Poisson arrival
  – Mean service time 2.5min, std. deviation is 1.25min
    • not exponentially distribution
    • Use Erlang-k to approximate, \(C_x=0.5=(1/k)^{0.5}\), \(k=4\)
  – M/E\(_4\)/1 queue, with \(\lambda=4/15/\text{min}\), \(\mu=0.4/\text{min}\), \(\rho=2/3\)
  – So,

\[
L_q=5/6, \ W_q=25/8\text{min}
\]
$E_k/M/1$ queue

- The arrival process is not Poisson arrival
  - Erlang-$k$ distribution with mean $1/\lambda$, each stage is exponential distribution with $k\lambda$
- State: $(n,i)$, or change to $nk+i$
- The balance equation is identical to that of full-batch bulk-service queue $M/M^k/[k]/1$ with $k\lambda$ arrival and constant batch size $k$
Average metrics of $E_k/M/1$

- The distribution is similarly as
  \[ p_j^{(P)} = \rho(1-r_0)r_0^{j-k}, \quad j \geq k-1, \rho = \lambda / \mu \]

- The distribution of $E_k/M/1$ is
  \[ p_n = \sum_{j=nk}^{nk+k-1} p_j^{(P)} = \rho(1-r_0)r_0^{nk-k} (1 + r_0 + \ldots r_0^{k-1}) \]
  \[ = \rho(1-r_0^k)r_0^{nk-k} \]

- We have
  \[ L = \frac{\rho}{1-r_0^k}, \quad L_q = L - \rho, \quad T = L / \lambda, \quad W_q = T - \frac{1}{\mu} \]