Queueing Theory

Queueing Theory

- **Specification of a Queue**
  - Source
    - Finite
    - Infinite
  - Arrival Process
  - Service Time Distribution
  - Maximum Queueing System Capacity
  - Number of Servers
  - Queue Discipline

Queueing Theory(cont.)

- Specification of a Queue(cont.)
  - ❍ Traffic Intensity (λ/μ)

    Note: $E[s] / E[\tau] = \lambda E[s] = \lambda/\mu$
  - ❍ Server Utilization
  - ❍ Probability that N customers are in the system at time t.

Queueing Theory(cont.)

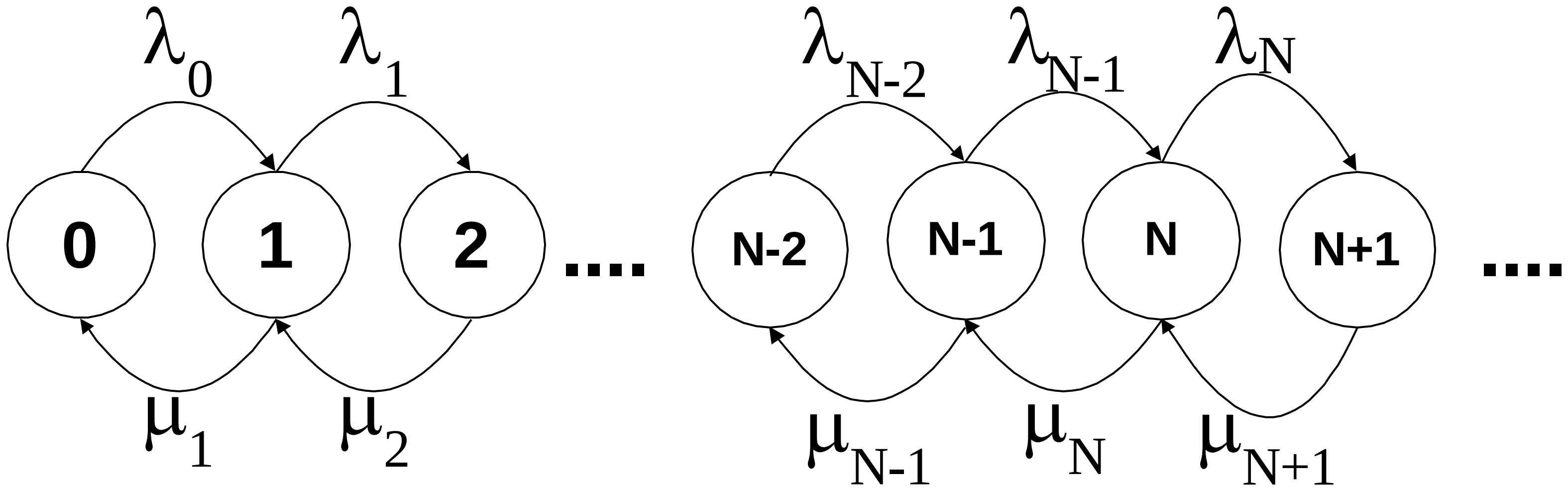
- Relationships:
  - $L = \lambda W$ (L: avg # in the system)
  - $L_q = \lambda W_q$ (L_q : avg # in queue)
  - $W = W_q + 1/\mu$ (W: avg waiting time in sys.)

    (W_q: avg waiting time in queue)

⌖Note: All four(L, L_q, W , W_q) can be determined after ONE is found

Birth-And-Death Process

State:



In the long run, we have:
Rate IN = Rate Out Principle

Birth-And-Death Process(cont.)

- Equation Expressing This:

State	Rate In = Rate Out
0	$\mu_1 P_1 = \lambda_0 P_0$
1	$\lambda_0 P_0 + \mu_2 P_2 = (\lambda_1 + \mu_1)\ P_1$
2	$\lambda_1 P_1 + \mu_3 P_3 = (\lambda_2 + \mu_2)\ P_2$
....	....
N-1	$\lambda_{N-2} P_{N-2} + \mu_N P_N = (\lambda_{N-1} + \mu_{N-1})\ P_{N-1}$
N	$\lambda_{N-1} P_{N-1} + \mu_{N+1} P_{N+1} = (\lambda_N + \mu_N)\ P_N$
....	....

Birth-And-Death Process(cont.)

- Finding Steady State Process:

State

0: $P_1 = (\lambda_0 / \mu_1)\, P_0$

1: $P_2 = (\lambda_1 / \mu_2)\, P_1 + (\mu_1 P_1 - \lambda_0 P_0) / \mu_2$

$= (\lambda_1 / \mu_2)\, P_1 + (\mu_1 P_1 - \mu_1 P_1) / \mu_2$

$= (\lambda_1 / \mu_2)\, P_1$

$$= \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} P_0$$

Birth-And-Death Process(cont.)

- *Finding Steady State Process(cont.):*

State

n-1: $P_n = (\lambda_{n-1} / \mu_n) P_{n-1} + (\mu_{n-1}P_{n-1} - \lambda_{n-2}P_{n-2}) / \mu_n$

$= (\lambda_{n-1} / \mu_n) P_{n-1} + (\mu_{n-1}P_{n-1} - \mu_{n-1}P_{n-1}) / \mu_n$

$= (\lambda_{n-1} / \mu_n) P_{n-1}$

$$= \frac{\lambda_{n-1}\lambda_{n-2}\cdots\lambda_0}{\mu_n\mu_{n-1}\cdots\mu_1} P_0$$

Birth-And-Death Process(cont.)

- Finding Steady State Process(cont.):

N: $P_{n+1} = (\lambda_n / \mu_{n+1}) P_n + (\mu_n P_n - \lambda_{n-1} P_{n-1}) / \mu_{n+1}$

$= (\lambda_n / \mu_{n+1}) P_n$

$$= \frac{\lambda_n \lambda_{n-1} \cdots \lambda_0}{\mu_{n+1} \mu_n \cdots \mu_1} P_0$$

To Simplify:

Let $C = (\lambda_{n-1}\ \lambda_{n-2} \ldots.\ \lambda_0) / (\mu_n\ \mu_{n-1} \ldots\ldots\ldots\ \mu_1)$

Then $\mathbf{P_n = C_n P_0}$, N = 1, 2, ...

M/M/1

Recall:

$\rho \quad = \lambda / \mu \quad < 1$ (for steady-state)

$C_n \quad = (\lambda / \mu)^n = \rho^n$, for n = 1, 2, ...

$P_n \quad = C_n \cdot P_0$

The requirement that $\sum_{n=0}^{\infty} P_n = 1$

$$\Rightarrow [1 + \sum_{n=1}^{\infty} C_n] P_0 = 1$$

$$\Rightarrow P_0 = 1/(1 + \sum_{n=1}^{\infty} C_n)$$

$$= 1/(1 + \sum_{n=1}^{\infty} \rho^n)$$

$$= 1/(\rho^0 + \sum_{n=1}^{\infty} \rho^n) \qquad (\rho^0 = 1)$$

M/M/1(cont.)

$$P_0 = 1 \Big/ \left(\sum_{n=0}^{\infty} \rho^n\right)$$

$$= \left(\sum_{n=0}^{\infty} \rho^n\right)^{-1}$$

$$= [1/(1-\rho)]^{-1}$$

$$= 1 - \rho$$

Thus, $P_n = (1 - \rho)\, \rho^n$, for n = 0, 1, 2,...

Note:

1) $\sum_{i=0}^{n} X^i = (1 - X^{n+1})/(1 - X)$, for any x,

2) $\sum_{i=0}^{\infty} X^n = 1/(1 - X)$, if $|x| < 1$.

M/M/1(cont.)

Consequently,

$$L = \sum_{n=0}^{\infty} n(1-\rho)\rho^n$$

$$= (1-\rho)\rho \sum_{n=0}^{\infty} \frac{d}{d\rho}\rho^n = (1-\rho)\rho \frac{d}{d\rho}\left(\sum_{n=0}^{\infty}\rho^n\right)$$

$$= (1-\rho)\rho \frac{d}{d\rho}\frac{1}{(1-\rho)} = (1-\rho)\rho\frac{1}{(1-\rho)^2}$$

$$= \rho/(1-\rho)$$

or

$$= \lambda/(\mu-\lambda)$$

M/M/1(cont.)

Similarly,

$$L_q = \sum_{n=1}^{\infty}(n-1)P_n$$

$$= \sum_{n=1}^{\infty} nP_n - \sum_{n=1}^{\infty} P_n$$

$$= \sum_{n=0}^{\infty} nP_n - (\sum_{n=0}^{\infty} P_n - P_0)$$

$$= L - 1(1 - P_0)$$

$$= \rho/(1-\rho) - 1 + (1-\rho)$$

$$= \rho^2/(1-\rho), \text{or}$$

$$= \lambda^2/\mu(\mu-\lambda)$$

M/M/1 Example I

Traffic to a message switching center for one of the outgoing communication lines arrive in a random pattern at an average rate of 240 messages per minute. The line has a transmission rate of 800 characters per second. The message length distribution (including control characters) is approximately exponential with an average length of 176 characters. Calculate the following principal statistical measures of system performance, assuming that a very large number of message buffers are provided:

M/M/1 Example I (cont.)

- (a) Average number of messages in the system
- (b) Average number of messages in the queue waiting to be transmitted.
- (c) Average time a message spends in the system.
- (d) Average time a message waits for transmission
- (e) Probability that 10 or more messages are waiting to be transmitted.
- (f) 90th percentile waiting time in queue.

M/M/1 Example I (cont.)

- E[s] = Average Message Length / Line Speed = {176 char/message} / {800 char/sec} = 0.22 sec/message or
- μ = 1 / 0.22 {message / sec} = 4.55 message / sec
- λ = 240 message /min = 4 message / sec
- ρ = λ E[s] = λ / μ = 0.88

M/M/1 Example I (cont.)

- (a) $L = \rho / (1 - \rho) = 7.33$ (messages)
- (b) $L_q = \rho^2 / (1 - \rho) = 6.45$ (messages)
- (c) $W = E[s] / (1 - \rho) = 1.83$ (sec)
- (d) $W_q = \rho \times E[s] / (1 - \rho) = 1.61$ (sec)
- (e) P [11 or more messages in the system] $= \rho^{11} = 0.245$
- (f) $\pi_q(90) = W \ln\{(100-90)\,\rho\}$ $= W \ln(10\rho)$ $=$ 3.98 (sec)

M/M/1 Example II

A branch office of a large engineering firm has one on-line terminal that is connected to a central computer system during the normal eight-hour working day. Engineers, who work throughout the city, drive to the branch office to use the terminal to make routine calculations. Statistics collected over a period of time indicate that the arrival pattern of people at the branch office to use the terminal has a Poisson (random) distribution, with a mean of 10 people coming to use the terminal each day. The distribution of time spent by an engineer at a terminal is exponential, with a

M/M/1 Example II (cont.)

mean of 30 minutes. The branch office receives complains from the staff about the terminal service. It is reported that individuals often wait over an hour to use the terminal and it rarely takes less than an hour and a half in the office to complete a few calculations. The manager is puzzled because the statistics show that the terminal is in use only 5 hours out of 8, on the average. This level of utilization would not seem to justify the acquisition of another terminal. What insight can queueing theory provide?

M/M/1 Example II (cont.)

- $\{10 \text{ person} / \text{day}\} \times \{1 \text{ day} / 8\text{hr}\} \times \{1\text{hr} / 60 \text{ min}\}$
  $= 10 \text{ person} / 480 \text{ min}$

  $= 1 \text{ person} / 48 \text{ min}$

  $\Longrightarrow \lambda = 1 / 48 \ (\text{person} / \text{min})$
- 30 minutes : 1 person = 1 (min) : 1/30 (person) $\Longrightarrow \mu = 1 / 30 \ (\text{person} / \text{min})$
- $\rho = \lambda / \mu = \{1/48\} / \{1/30\} = 30 / 48 = 5 / 8$

M/M/1 Example II (cont.)

- Arrival Rate $\lambda = 1 / 48$ (customer / min)
- Server Utilization $\rho = \lambda / \mu = 5 / 8 = 0.625$
- Probability of 2 or more customers in system $P[N \geq 2] = \rho^2 = 0.391$
- Mean steady-state number in the system $L = E[N] = \rho / (1 - \rho) = 1.667$
- S.D. of number of customers in the system $\sigma_N = \text{sqrt}(\rho) / (1 - \rho) = 2.108$

M/M/1 Example II (cont.)

- Mean time a customer spends in the system $W = E[w] = E[s] / (1 - \rho) = 80$ (min)
- S.D. of time a customer spends in the system $\sigma_w = E[w] = 80$ (min)
- Mean steady-state number of customers in queue $L_q = \rho^2 / (1 - \rho) = 1.04$
- Mean steady-state queue length of nonempty Qs $E[N_q \mid N_q > 0] = 1 / (1 - \rho) = 2.67$
- Mean time in queue $W_q = E[q] = \rho \times E[s] / (1 - \rho) = 50$ (min)

M/M/1 Example II (cont.)

- Mean time in queue for those who must wait

  $E[q \mid q > 0] = E[w] = 80$ (min)

- 90th percentile of the time in queue

  $\pi_q(90) = E[w] \ln (10\ \rho)$

  $= 80 * 1.8326$ $= 146.6$ (min)

- 90th percentile of the time in system

  $\pi_w(90) = 2.3 * E[w] = 184$ (min)

M/M/1 Example II (cont.)

- Defined by equation $P[w \leq \pi_w(90)] = 0.9$ response time of system $\pi_w(90)$ - amount of time in the system such that 90% of all arriving customers spend less than this amount of time in the system

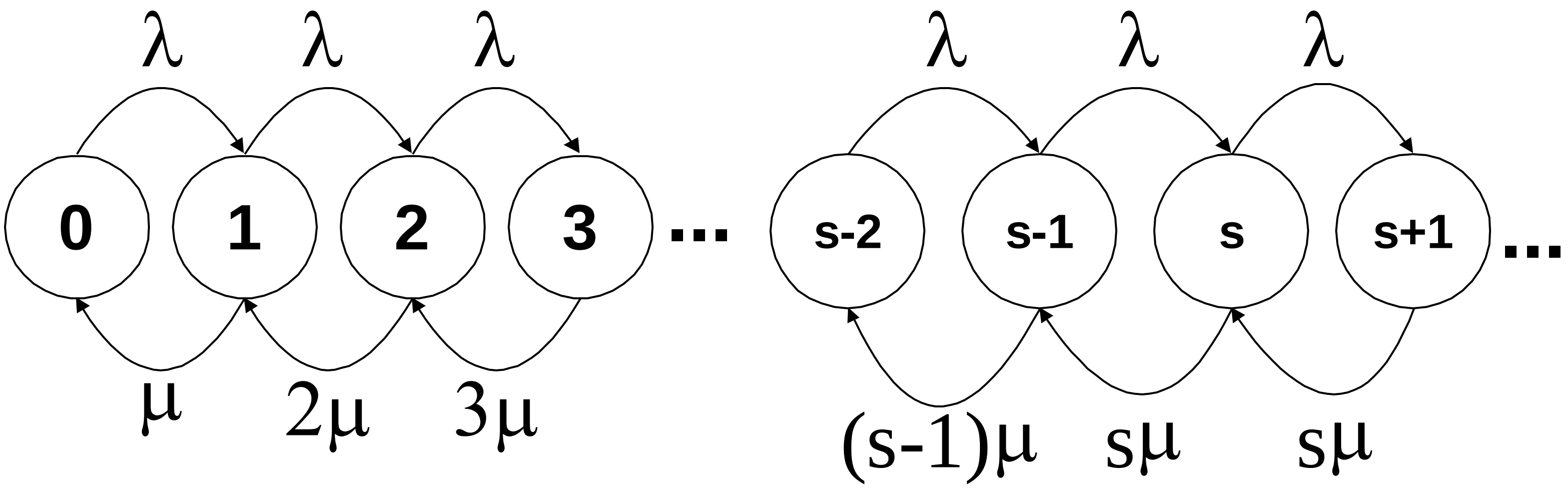
M/M/s (s > 1)

Recall: $\lambda_n = \lambda$, for n = 0, 1, 2,.....

$\mu_n = n\,\mu$, for n = 1, 2,..., s

$= s\,\mu$, for n = s, s+1,...

Rate Diagram



M/M/s (cont.)

State	Rate In = Rate Out
0	$\mu P_1 = \lambda P_0$
1	$2\mu P_2 + \lambda P_0 = (\lambda + \mu)\ P_1$
2	$3\mu P_3 + \lambda P_1 = (\lambda + 2\mu)\ P_2$
….	……………….
s-1	$s\mu P_s + \lambda P_{s-2} = \{\lambda + (s-1)\mu\}\ P_{s-1}$
s	$s\mu P_{s+1} + \lambda P_{s-1} = (\lambda + s\mu)\ P_s$
s+1	$s\mu P_{s+2} + \lambda P_s = (\lambda + s\mu)\ P_{s+1}$
….	……………….

M/M/s (cont.)

- Now, solve for P_1 , P_2, P_3... in terms of P_0

$P_1 = (\lambda / \mu) P_0$

$P_2 = (\lambda / 2\mu) P_1 = (1/2!) \times (\lambda / \mu)^2 P_0$

$P_3 = (\lambda / 3\mu) P_2 = (1/3!) \times (\lambda / \mu)^3 P_0$

....

$P_s = (1/s!) \times (\lambda / \mu)^s P_0$

$$P_{s+1} = (1/s) \times (\lambda / \mu) P_s = \frac{(\lambda/\mu)^s}{s!} \times \frac{\lambda}{s\mu} P_0$$

M/M/s (cont.)

$$P_{s+2} = (1/s)\cdot(\lambda/\mu)P_{s+1} = \frac{(\lambda/\mu)^s}{s!}\left(\frac{\lambda}{s\mu}\right)^2 P_0$$

$\ldots$

$$P_{s+j} = (1/s)\cdot(\lambda/\mu)P_{s+j-1} = \frac{(\lambda/\mu)^s}{s!}\left(\frac{\lambda}{s\mu}\right)^j P_0$$

$\ldots$

M/M/s (cont.)

Therefore, if we denote $P_n = C_n \times P_0$,

then $C_n = \dfrac{(\lambda / \mu)^n}{n!}$, for n = 1, 2, ..., s.

and $C_n = \dfrac{(\lambda / \mu)^s}{s!}\left(\dfrac{\lambda}{s\mu}\right)^{n-s}$, for n = s+1, s+2,...

$$= \frac{(\lambda/\mu)^n}{s!s^{n-s}}$$

M/M/s (cont.)

So, if $\lambda < s\mu$ =>

$$P_0 = 1 \Big/ \left\{ \sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \sum_{n=s}^{\infty} (\lambda/s\mu)^{n-s} \right\}$$

$$= 1 \Big/ \left\{ \sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \frac{1}{1-(\lambda/s\mu)} \right\}$$

$$P_n = \frac{(\lambda/\mu)^n}{n!} P_0, \qquad \text{if } 0 \le n \le s$$

$$= \frac{(\lambda/\mu)^n}{s!\,s^{n-s}} P_0, \qquad \text{if } s \le n$$

M/M/s (cont.)

Now solve for L_q: Note, $\rho = \lambda / s\mu$

$$L_q = \sum_{n=s}^{\infty} (n - s) P_n$$

$$= \sum_{j=0}^{\infty} j P_{s+j} \; ; \;\; \text{Note, } n = s + j$$

$$= \sum_{j=0}^{\infty} j \frac{(\lambda / \mu)^s}{s!} \rho^j P_0$$

$$= P_0 \frac{(\lambda / \mu)^s}{s!} \rho \sum_{j=0}^{\infty} \frac{d}{d\rho} \rho^j$$

M/M/s (cont.)

$$L_q = P_0 \frac{(\lambda/\mu)^s}{s!} \rho \frac{d}{d\rho} \sum_{j=0}^{\infty} \rho^j$$

$$= P_0 \frac{(\lambda/\mu)^s}{s!} \rho \frac{d}{d\rho} \frac{1}{(1-\rho)}$$

$$= P_0 \frac{(\lambda/\mu)^s}{s!} \frac{\rho}{(1-\rho)^2}$$

M/M/s (cont.)

$$L_q = P_0 \frac{(\lambda / \mu)^s}{s!} \frac{\rho}{(1-\rho)^2}$$, $\rho = \lambda / s\mu$

(L_q : avg # in queue)

$W_q = L_q / \lambda$ (W_q: avg waiting time in Q)

$W = W_q + 1 / \mu$ (W: avg waiting time in sys.)

$L = \lambda (W_q + 1/\mu)$ (L: avg # in the system)

$= L_q + \lambda / \mu$

Steady-State Parameters of M/M/s Queue

$$\rho \qquad = \lambda \,/\, s\mu$$

$$P_0 = \left\{ \left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} \right] + \left[\left(\frac{\lambda}{\mu} \right)^s \frac{1}{s!} \frac{s\mu}{s\mu - \lambda} \right] \right\}^{-1}$$

$$= \left\{ \left[\sum_{n=0}^{s-1} \frac{(s\rho)^n}{n!} \right] + \left[(s\rho)^s \frac{1}{s!} \frac{1}{1-\rho} \right] \right\}^{-1}$$

$$P(L(\infty) \geq s) = \{(\lambda/\mu)^s P_0\} \,/\, \{s!(1- \lambda/s\mu)\}$$

$$= \{(s\rho)^s P_0\} \,/\, \{s!\ (1 - \rho)\}$$

Steady-State Parameters of M/M/s Queue (cont.)

$L \quad = s\rho + \{(s\rho)^{s+1} P_0\} / \{s\,(s!)\,(1 - \rho)^2\}$

$\quad\;\; = s\rho + \{\rho\, P\,(L(\infty) \geq s)\,\} / \{1 - \rho\}$

$W \;= L / \lambda$

$W_q = W - \;1/\mu$

$L_q = \lambda\, W_q \qquad\qquad = \{(s\rho)^{s+1}$
$P_0\} / \{s\,(s!)\,(1 - \rho)^2\} \qquad\qquad = \{\rho\, P\,(L(\infty) \geq$
$s)\,\} / \{1 - \rho\}$

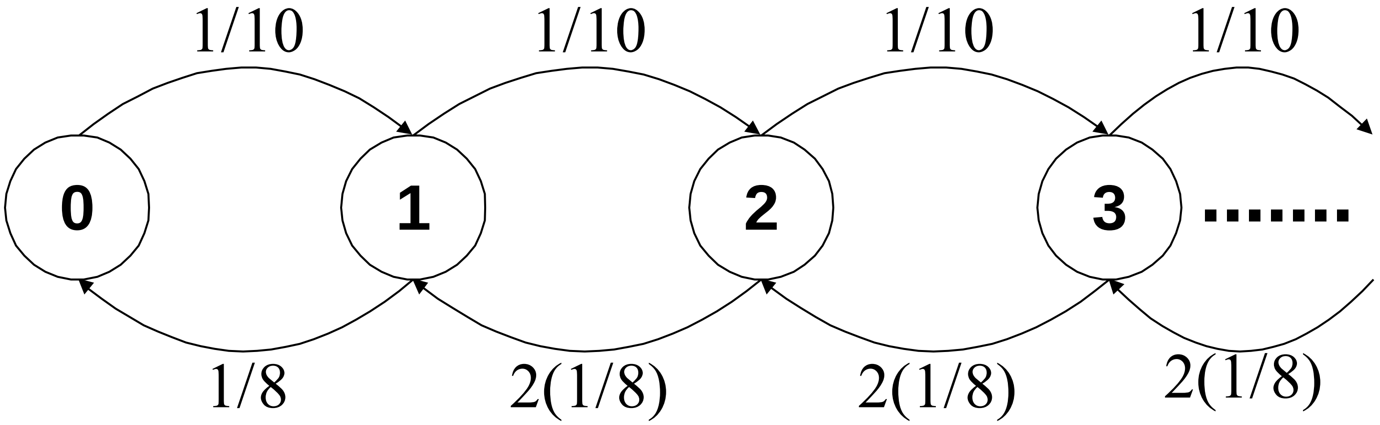
$L - L_q = \lambda / \mu = s\rho$

M/M/s Case Example I

Example:

M/M/2 ; s = 2

$\lambda = 1/10$, $\mu = 1/8$ (=service rate/server)



$$\rho = \lambda / s\mu = \{1/10\} / \{2(1/8)\} = 0.4$$

M/M/s Case Example I (cont.)

$$P_0 = 1 \Big/ \left\{ \frac{(0.8)^0}{0!} + \frac{(0.8)^1}{1!} + \frac{(0.8)^2}{2!} \times \frac{1}{1-0.4} \right\}$$

$$= 0.429 \text{ } (\cong 43\% \text{ of time, system is empty})$$

$$\text{as compared to s = 1:} \quad P_0 = 0.20$$

$$L_q = P_0 \frac{(\lambda / \mu)^s}{s!} \frac{\rho}{(1-\rho)^2}$$

$$= 0.429 \times \{0.8^2 \times 0.4\} / \{2! \times (1-0.4)^2\}$$

$$= 0.152$$

M/M/s Case Example I (cont.)

$W_q = L_q / \lambda = 0.152 / (1/10) = 1.52$ (min)

$W = W_q + 1 / \mu = 1.52 + 1 / (1/8) = 9.52$ (min)

What proportion of time is both repairman busy? (long run)

$P(N \geq 2) = 1 - P_0 - P_1$
$= 1 - 0.429 - 0.343$ =
0.228 (Good or Bad?)

M/M/s Example II

Many early examples of queueing theory applied to practical problems concerning tool cribs. Attendants manage the tool cribs while mechanics, assumed to be from an infinite calling population, arrive for service. Assume Poisson arrivals at rate 2 mechanics per minute and exponentially distributed service times with mean 40 seconds.

M/M/s Example II (cont.)

λ = 2 per minute, and μ = 60/40 = 3/2 per minute.

Since, the offered load is greater than 1, that is, since, $\lambda / \mu = 2 / (3/2) = 4/3 > 1$, more than one server is needed if the system is to have a statistical equilibrium. The requirement for steady state is that $s > \lambda / \mu = 4/3$. Thus, at least s = 2 attendants are needed. The quantity 4/3 is the expected number of busy server, and for $s \geq 2$, $\rho = 4 / (3s)$ is the long-run proportion of time each server is busy. (What would happen if there were only s = 1 server?)

M/M/s Example II (cont.)

Let there be s = 2 attendants. First, P_0 is calculated as

$$P_0 = \left\{ \left[\sum_{n=0}^{1} \frac{(4/3)^n}{n!} \right] + \left[\left(\frac{4}{3} \right)^2 \frac{1}{2!} \frac{2(3/2)}{2(3/2) - 2} \right] \right\}^{-1}$$

$$= \{1 + 4/3 + (16/9)(1/2)(3)\}^{-1}$$

$$= \{15 / 3\}^{-1} = 1/5 = 0.2$$

The probability that all servers are busy is given by

$$P(L(\infty) \geq 2) = \{(4/3)^2 (1/5)\} / \{2!(1- 2/3)\}$$

$$= (8/3)(1/5) = 0.533$$

M/M/s Example II (cont.)

Thus, the time-average length of the waiting line of mechanics is $L_q = \{(2/3)(8/15)\} / (1 - 2/3) = 1.07$ mechanics

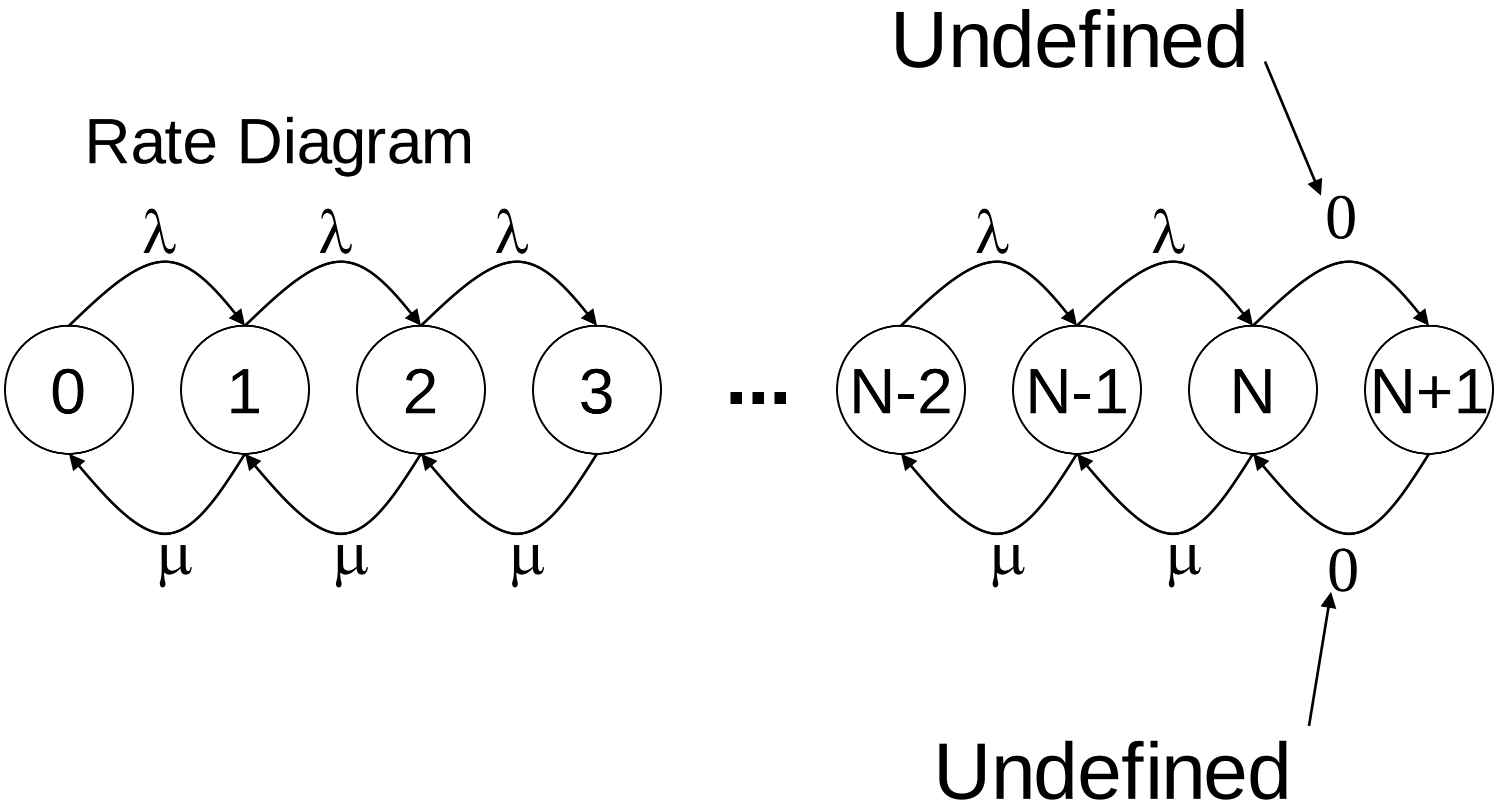
and the time-average number in system is given by $L = L_q + \lambda/\mu = 16/15 + 4/3 = 12/5 = 2.4$ mechanics

Using Little's relationships, the average time a mechanic spends at the tool crib is $W = L / \lambda = 2.4 / 2 = 1.2$ minutes

while the avg time spent waiting for an attendant is $W_q = W - 1/\mu = 1.2 - 2/3 = 0.533$ minute

M/M/1/N (single server)

Rate Diagram



M/M/1/N (cont.)

1. Form Balance Equations:

2. Solve for P_0:

$$\sum_{n=0}^{N} P_n = 1 \quad \text{or}$$

$$P_0 + (\lambda/\mu)^1 P_0 + \cdot\cdot\cdot + (\lambda/\mu)^N P_0 = 1$$

$$P_0 \{1 + (\lambda/\mu)^1 + \cdot\cdot\cdot + (\lambda/\mu)^N \} = 1$$

$$P_0 = 1 / \{ \sum_{n=0}^{N} (\lambda/\mu)^n \}$$

$$= 1 \Big/ \left\{ \frac{1-(\lambda/\mu)^{N+1}}{1-(\lambda/\mu)} \right\}$$

$$= (1 - \rho) / (1 - \rho^{N+1})$$

M/M/1/N (cont.)

So, $$P_n = \left\{ \frac{1-\rho}{1-\rho^{N+1}} \right\} \rho^n$$, for n = 0, 1, 2, ..., N

Hence,

$$L = \sum_{n=0}^{N} nP_n$$

$$= \frac{1-\rho}{1-\rho^{N+1}} \rho \sum_{n=0}^{N} \frac{d}{d\rho} \rho^n$$

$$= \frac{1-\rho}{1-\rho^{N+1}} \rho \frac{d}{d\rho} \sum_{n=0}^{N} \rho^n$$

M/M/1/N (cont.)

$$L = \frac{1-\rho}{1-\rho^{N+1}}\rho\frac{d}{d\rho}\left\{\frac{1-\rho^{N+1}}{1-\rho}\right\}$$

$$= \rho\frac{-(N+1)\rho^{N}+N\rho^{N+1}+1}{(1-\rho^{N+1})(1-\rho)}$$

$$= \frac{\rho}{(1-\rho)} - \frac{(N+1)\rho^{N+1}}{(1-\rho^{N+1})}$$

M/M/1/N (cont.)

As usual (when s = 1)

$L_q = L - (1 - P_0)$

$W = L / \lambda_e$, where $\lambda_e = \lambda (1 - P_N)$

$W_q = L_q / \lambda_e$

M/M/1/N Example

The unisex barbershop can hold only three customers, one in service and two waiting. Additional customers are turned away when the system is full. Determine the measures of effectiveness for this system. The traffic intensity is $\lambda / \mu = 2 / 3$.

The probability that there are three customers in the system is computed by $P_n = P_3 = \{(1-2/3)\ (2/3)^3\}\ \mathbf{/}\ \{ 1 - (2/3)^4\} = 8 / 65 = 0.123$

M/M/1/N Example (cont.)

The expected # of customers in the shop is given by

$$L = \frac{2/3\{1-4(2/3)^3+3(2/3)^4\}}{\{1-(2/3)^4\}(1-2/3)} = \frac{66}{65}$$

$$= 1.015 (customers)$$

Now, the effective arrival rate, λ_e , is given by

$\lambda_e = \lambda\ (1 - P_n) = 2(1 - 8/65) = 2 \times 57 / 65 = 114/65$

$= 1.754$ (customers/hour)

Then W can be calculated as

$W = L / \lambda_e = 1.015 / 1.754 = 0.579$ (hour)

M/M/1/N Example (cont.)

In order to calculate L_q, first determine P_0 as

$P_0 = (1 - \rho) / (1 - \rho^{N+1}) = (1 - 2/3) / \{1 - (2/3)^4\}$

$= \{1/3\} / \{65/81\} = 27 / 65$

$= 0.415$

Then the average length of the queue is given by

$L_q = L - (1- P_0) = 1.015 - (1 - 0.415)$

$= 0.43$ (customer)

M/M/1/N Example (cont.)

Note that 1- P_0 = 0.585 is the average number of customers being served, or equivalently, the probability that the single server is busy. Thus the server utilization, or proportion of time the server is busy in the long run, is given by ρ = 1- P_0 = λ_e / μ = 0.585

Finally, the waiting time in the queue is determined by Little's equation as W_q = L_q / λ_e = 0.43 / 1.754 = 0.245 (hour)

M/M/1/N Example (cont.)

The reader should compare these results to those of the unisex barbershop before the capacity constraint was placed on the system. Specifically, in systems with limited capacity, the traffic intensity λ / μ can assume any positive value and no longer equals the server utilization $\rho = \lambda_e / \mu$.

Note that server utilization decreases from 67% to 58.5% when the system imposes a capacity constraint.

M/M/1/N Example (cont.)

Since P_0 and P_3 have been computed, it is easy to check the value of L using equation $L = \sum_{n=0}^{N} nP_n$

To make the check requires computation of P_1 & P_2: P_1

$= \{(1 - 2/3)(2/3)\} / \{1- (2/3)^4\} = 18/65 = 0.277$

Since $P_0 + P_1 + P_2 + P_3 = 1$, P_2

$= 1 - P_0 - P_1 - P_3 = 1 - 27/65 - 18/65 - 8/65$

$= 12 / 65$

$= 0.185$

M/M/1/N Example (cont.)

$$\Rightarrow L = \sum_{n=0}^{N} nP_n$$

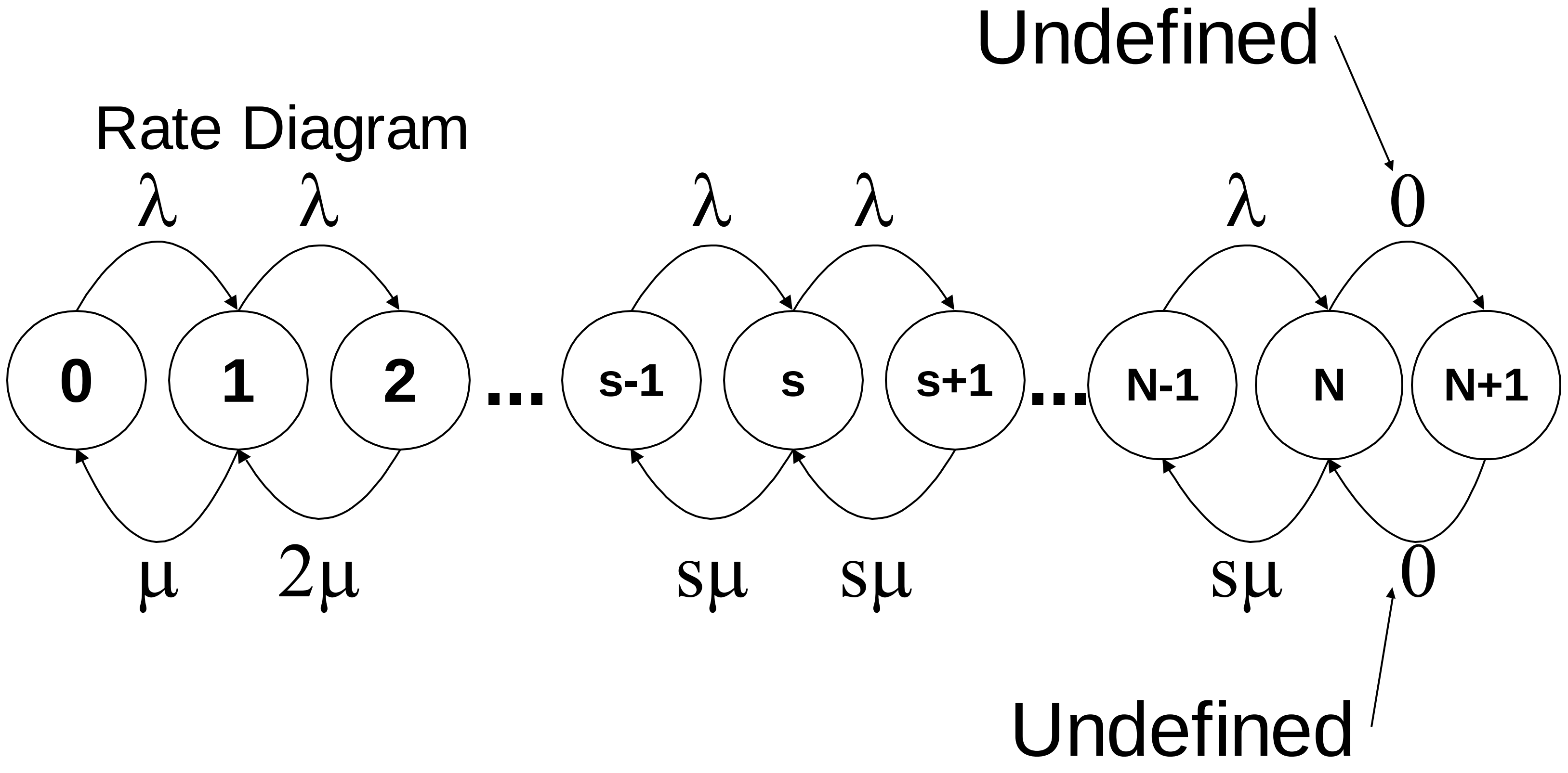
$$= 0\times(27/65) + 1\times(18/65) + 2\times(12/65) + 3\times(8/65)$$

$$= 66 / 65$$

$$= 1.015 \text{ (customer)}$$

which is the same value as the expected number computed.

M/M/s/N


Undefined
Rate Diagram
λ
λ
λ
λ
λ
0
0
1
2
...
s-1
s
s+1
...
N-1
N
N+1
μ
2μ
sμ
sμ
sμ
0
Undefined

Steady-State Parameters of M/M/s/N

$$P_0 = 1 \Big/ \left[1 + \sum_{n=1}^{s} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!} \sum_{n=s+1}^{N} (\lambda/s\mu)^{n-s} \right]$$

$$P_n = \frac{(\lambda/\mu)^n}{n!} P_0, \quad \text{for } n = 1, 2, \ldots s$$

$$= \frac{(\lambda/\mu)^n}{s!s^{n-s}} P_0, \quad \text{for } n = s, s+1, \ldots N$$

$$= \; 0, \text{ for } n > N$$

Steady-State Parameters of M/M/s/N (cont.)

$$L = \sum_{n=0}^{s-1} nP_n + L_q + s(1 - \sum_{n=0}^{s-1} P_n)$$

$$L_q = \frac{P_0(\lambda/\mu)^s \rho}{s!(1-\rho)^2}\left[1 - \rho^{N-s} - (N-s)\rho^{N-s}(1-\rho)\right]$$

Note: W and W_q are obtained from these quantities just as shown for the single server case.

Steady-State Parameters of M/G/1 Queue

- $\rho = \lambda / \mu$
- $L = \rho + \{\lambda^2 (\mu^{-2} + \sigma^2)\} / \{2 (1 - \rho)\}$
  $= \rho + \{\rho^2 (1 + \sigma^2 \mu^2)\} / \{2 (1 - \rho)\}$
- $W = \mu^{-1} + \{\lambda (\mu^{-2} + \sigma^2)\} / \{2 (1 - \rho)\}$
- $W_q = \{\lambda (\mu^{-2} + \sigma^2)\} / \{2 (1 - \rho)\}$
- $L_q = \{\lambda^2 (\mu^{-2} + \sigma^2)\} / \{2 (1 - \rho)\}$
  $= \{\rho^2 (1 + \sigma^2 \mu^2)\} / \{2 (1 - \rho)\}$
- $P_0 = 1 - \rho$

M/G/1 Example

There are two workers competing for a job. Able claims an average service time which is faster than Baker's, but Baker claims to be more consistent, if not as fast. The arrivals occur according to a Poisson process at a rate of $\lambda = 2$ per hour. (1/30 per minute). Able's statistics are an average service time of 24 minutes with a standard deviation of 20 minutes. Baker's service statistics are an average service time of 25 minutes, but a standard deviation of only 2 minutes. If the average length of the queue is the criterion for hiring, which worker should be hired?

M/G/1 Example (cont.)

- For Able, $\lambda = 1/30$ (per min), $\mu^{-1} = 24$ (min), $\rho = \lambda / \mu = 24/30 = 4/5$ $\sigma^2 = 20^2 = 400(\text{min}^2)$ $L_q = \{\lambda^2 (\mu^{-2} + \sigma^2)\} / \{2 (1 - \rho)\}$ $= \{(1/30)^2 (24^2 + 400)\} / \{2 (1-4/5)\}$ $= 2.711$ (customers)
- For Baker, $\lambda = 1/30$ (per min), $\mu^{-1} = 25$ (min), $\rho = \lambda / \mu = 25/30 = 5/6$ $\sigma^2 = 2^2 = 4(\text{min}^2)$ $L_q = \{(1/30)^2 (25^2 + 4)\} / \{2 (1-5/6)\}$ $= 2.097$ (customers)

M/G/1 Example (cont.)

Although working faster on the average, Able's greater service variability results in an average queue length about 30% greater than Baker's. On the other hand, the proportion of arrivals who would find Able idle and thus experience no delay is $P_0 = 1 - \rho = 1 / 5 = 20\%$, while the proportion who would find Baker idle and thus experience no delay is $P_0 = 1 - \rho = 1 / 6 = 16.7\%$. On the basis of average queue length, L_q, Baker wins.

Steady-State Parameters of M/E$_k$/1 Queue

$$L = \frac{\lambda}{\mu} + \frac{1+k}{2k}\frac{\lambda^2}{\mu(\mu-\lambda)} = \rho + \frac{1+k}{2k}\frac{\rho^2}{1-\rho}$$

$$W = \frac{1}{\mu} + \frac{1+k}{2k}\frac{\lambda}{\mu(\mu-\lambda)} = \mu^{-1} + \frac{1+k}{2k}\frac{\rho\mu^{-1}}{1-\rho}$$

$$W_q = \frac{1+k}{2k}\frac{\lambda}{\mu(\mu-\lambda)} = \frac{1+k}{2k}\frac{\rho\mu^{-1}}{1-\rho}$$

$$L_q = \frac{1+k}{2k}\frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{1+k}{2k}\frac{\rho^2}{1-\rho}$$

M/E$_k$/1 Example

Patient arrive for a physical examination according to a Poisson process at the rate of one per hour. The physical examination requires three stages, each one independently and exponentially distributed with a service time of 15 minutes. A patient must go through all three stages before the next patient is admitted to the treatment facility. Determine the average number of delayed patients ,L_q , for this system.

M/E$_k$/1 Example (cont.)

If patients follow this treatment pattern, the service-time distribution will be Erlang of order k=3. The necessary treatment parameters are λ = 1/60 per minute and μ = 1/45 per minute; thus

$$L_q = \frac{1+k}{2k}\frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{1+3}{2\times 3}\frac{(1/60)^2}{(1/45)(1/45-1/60)}$$

$$= \frac{2}{3}\frac{135}{60} = \frac{3}{2}(patients)$$

Steady-State Parameters of M/D/1 Queue

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

$$W = \frac{1}{\mu} + \frac{1}{2}\frac{\lambda}{\mu(\mu-\lambda)} = \mu^{-1} + \frac{1}{2}\frac{\rho\mu^{-1}}{1-\rho}$$

$$W_q = \frac{1}{2}\frac{\lambda}{\mu(\mu-\lambda)} = \frac{1}{2}\frac{\rho\mu^{-1}}{1-\rho}$$

$$L_q = \frac{1}{2}\frac{\lambda^2}{\mu(\mu-\lambda)} = \frac{1}{2}\frac{\rho^2}{1-\rho}$$

M/D/1 Example

Arrivals to an airport are all directed to the same runway. At a certain time of the day, these arrivals are Poisson distributed at a rate of 30 per hour. The time to land an aircraft is a constant 90 seconds. Determine L_q, W_q, L and W for this airport. In this case λ= 0.5 per minute, and $1/\mu = 1.5$ minutes, or $\mu = 2/3$ per minute.

M/D/1 Example (cont.)

The runway utilization is

$\rho \quad = \lambda / \mu = (1/2) / (2/3) = 3/4$

The steady-state parameters are given by

$L_q \quad = \{(3/4)^2\} / \{2 (1 - 3/4)\}$

$= 9 / 8 = 1.125$ aircraft

$W_q = L_q / \lambda = (9/8) / (1/2) = 2.25$ minutes

$W \quad = W_q + 1 / \mu = 2.25 + 1.5 = 3.75$ minutes

$L \quad = L_q + \lambda / \mu = 1.125 + 0.75 = 1.875$ aircraft

Steady-State Parameters of M/G/∞ Queue

$P_0 = e^{-\lambda/\mu}$

$P_n = \{e^{-\lambda/\mu}(\lambda/\mu)^n\} / n!$, $n = 0, 1, ...$

$W = 1 / \mu$

$W_q = 0$

$L = \lambda / \mu$

$L_q = 0$

M/G/∞ Example

Prior to introducing their new on-line computer information service, The Connection must plan their system capacity in terms of the number of users that can be logged on simultaneously. If the service is successful, customers are expected to log on at a rate of λ = 500 per hour, according to a Poisson process, and stay connected for an average of $1/\mu$ = 20 minutes (or 1/3 hour). In the real system there will be an upper limit on simultaneous users, but for planning purpose The

M/G/∞ Example (cont.)

Connection can pretend that the number of simultaneous users is infinite. An M/G/∞ model of the system implies that the expected number of simultaneous users is L = λ/μ = 500(3) = 1500, so a capacity greater than 1500 is certainly required. To ensure that they have adequate capacity 95% of the time, The Connection could allow the number of simultaneous users to be the smallest value s such that

M/G/∞ Example (cont.)

$$P(L(\infty) \le s) = \sum_{n=0}^{s} P_n$$

$$= \sum_{n=0}^{s} \{e^{-1500}(1500)^n\} / n! \ge 0.95$$

A capacity of s=1564 simultaneous users satisfies this requirement.

Steady-State Parameters of M/M/s/K/K Queue

$$P_0 = \left[\sum_{n=0}^{s-1} \binom{K}{n} \left(\frac{\lambda}{\mu}\right)^n + \sum_{n=s}^{K} \frac{K!}{(K-n)!s!s^{n-s}} \left(\frac{\lambda}{\mu}\right)^n \right]^{-1}$$

$$P_n = \binom{K}{n} \left(\frac{\lambda}{\mu}\right)^n P_0, \qquad n = 0, 1, ..., s\text{-}1$$

$$= \frac{K!}{(K-n)!s!s^{n-s}} \left(\frac{\lambda}{\mu}\right)^n P_0, \quad n = s, s+1, ... K$$

Steady-State Parameters of M/M/s/K/K Queue (cont.)

$$L = \sum_{n=0}^{K} nP_n$$

$$L_q = \sum_{n=s+1}^{K} (n-s)P_n$$

$$\lambda_e = \sum_{n=0}^{K} (K-n)\lambda P_n$$

$$\begin{aligned} W &= L / \lambda_e \\ W_q &= L_q / \lambda_e \\ \rho &= (L - L_q) / s \\ &= \lambda_e / s\mu \end{aligned}$$

M/M/s/K/K Example

There are two workers that are responsible for 10 milling machines. The machines run on the average of 20 minutes, then require an average 5-minute service period both times exponentially distributed. Therefore, $\lambda = 1/20$ and $\mu = 1/5$. Determine the various measures of performance for this system.

M/M/s/K/K Example (cont.)

All of the performance measures depend on P_0

$$P_0 = \left[\sum_{n=0}^{2-1} \binom{10}{n} \left(\frac{5}{20}\right)^n + \sum_{n=2}^{10} \frac{10!}{(10-n)!2!2^{n-2}} \left(\frac{5}{20}\right)^n \right]^{-1}$$

$$= 0.065$$

Using P_0 we can obtain the other P_n, from which we can compute the average number of machines waiting for service

$$L_q = \sum_{n=2+1}^{10} (n-2)P_n$$

$$= 1.46 (\text{machines})$$

M/M/s/K/K Example (cont.)

The effective arrival rate

$$\lambda_e = \sum_{n=0}^{K}(K-n)\lambda P_n$$

$$= \sum_{n=0}^{10}(10-n)(1/20)P_n$$

$$= 0.342(\text{machines}/\min)$$

and the average waiting time in the queue

$$W_q \quad = L_q / \lambda_e = 4.27 \text{ (minutes)}$$

Similarly, we can compute the expected number of machines being serviced or waiting to be served

$$L = \sum_{n=0}^{K} nP_n = \sum_{n=0}^{10} nP_n = 3.17(\text{machines})$$

M/M/s/K/K Example (cont.)

The average number of machines being serviced is given by $L - L_q = 3.17 - 1.46 = 1.71$ (machines)

since the machines must be running, waiting to be served, or in service, the average number of running machines is given by $K - L = 10 - 3.17 = 6.83$ (machines)

A frequently asked question is: What will happen if the number of servers is increased or decreased?

M/M/s/K/K Example (cont.)

If the number of workers in this example increases to three(s=3), then the time-average number of running machines increases to K - L = 7.74 (machines) an increase of 0.91 machine, on the average.

Conversely, what happens if the number of servers decreases to one? Then the time-average number of running machines decreases to K - L = 3.98 (machines)

M/M/s/K/K Example (cont.)

The decrease from two to one server has resulted in a drop of nearly three machines running, on the average.

This example illustrates several general relationships that have been found to hold for almost all queues. If the number of servers is decreased, delays, server utilization, and the probability of an arrival having to wait to begin service all increase.