King Fahd University of Petroleum & Minerals Computer Engineering Dept

COE 402 – Computer Systems Performance Evaluation

Term 043

Dr. Ashraf S. Hasan Mahmoud

Rm 22-148-3

Ext. 1724

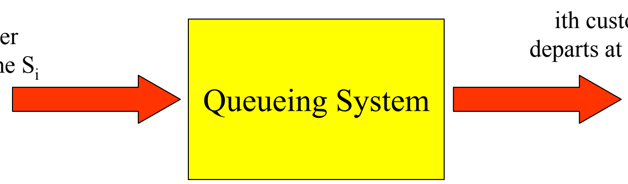
Email: ashraf@ccse.kfupm.edu.sa

Queuing Model

- **Consider the following system:**

A(t) $n(t) = A(t) – D(t)$ D(t)

ith customer arrives at time S_i



ith customer departs at time D_i

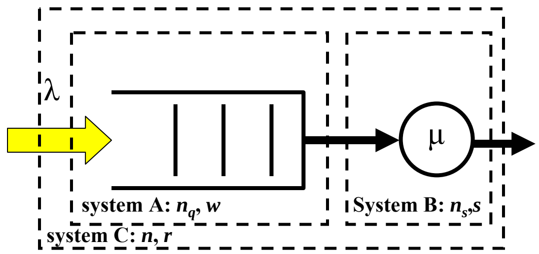
$r_i = D_i – A_i$

$w_i = r_i – S_i$
$= D_i – A_i – S_i$

A(t) – number of arrivals in (0, t]
D(t) – number of departures in (0, t]
n(t) – number of customers in system in (0,t]
r_i – duration of time spent in system for i^{th} customer – **textbook call this response time (*r*)**
w_i – duration of time spent waiting for service for i^{th} customer

Queuing Model

- The model is characterized using the following quantities:
  - λ – mean arrival rate of customers = 1/E[τ] (remember τ is the interarrival time)
  - s = service time per job
  - μ = mean service rate = 1/E[s]
  - n = number of job in system → E[n]
  - n_q = number of jobs in buffer → E[n_q]
  - ns = number of jobs in server → E[ns]
  - r = response time or total time in system → E[r]
  - w = waiting time → E[w]



Example: Queueing System

Problem: A data communication line delivers a block of information every 10 microseconds. A decoder check each block for errors and corrects the errors if necessary. It takes 1 microsecond to determine whether the block has any errors. If the block has one error it takes 5 microseconds to correct it and it has more than 1 error it takes 20 microseconds to correct the error. Blocks wait in the queue when the decoder falls behind. Suppose that the decoder is initially empty and that the number of errors in the first 10 blocks are: 0, 1, 3, 1, 0, 4, 0, 1, 0, 0.

a) Plot the number of blocks in the decoder as a function of time.

b) Find the mean number of blocks in the decoder

c) What percent of the time is the decoder empty?

Example: Queueing System – cont'd

Solution:

Interarrival time = 10 μsec

Service time = 1 if no errors
1+5 if 1 error
1+20 if more than 1 error

The queue parameters (A, D, S, and W) are shown below:

Block #:	1	2	3	4	5	6	7	8	9	10
Arrivals:	10	20	30	40	50	60	70	80	90	100
Errors:	0	1	3	1	0	4	0	1	0	0
Service:	1	6	21	6	1	21	1	6	1	1
Departs:	11	26	51	57	58	81	82	88	91	101
Waiting:	0	0	0	11	7	0	11	2	0	0

Example: Queueing System – cont'd

Solution:

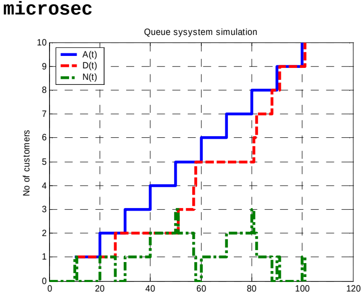
Using the previous results and knowing that

n(t) = A(t) – D(t)

One can produce the following results

```
Average no of customers in system =    0.950
Average customer waiting time     =    3.100 microsec
Maximum simulation time           =  101.000 microsec
Duration server busy              =   65.000
Server utilization                =   0.6436
Server idle                       =   0.3564
```

The following Matlab code can be used to solve this queue system (Note the code is general – it solves any system provided The Arrivals vector A, and the service vector S)



Example: Queueing System – cont'd

```
%
% Problem 9.3 - Leon Garcia's book
clear all
A = [10:10:100];
Errors = [0 1 3 1 0 4 0 1 0 0];
S = zeros(size(A));
D = zeros(size(A));
%
% this loop to computes service times
for i=1:length(A);
    if (Errors(i)==0)  S(i) = 1;
    else
        if (Errors(i)==1) S(i) = 6;
        else
            S(i) = 21;
        end
    end
    %
    % this section computes the departure time for the ith user
    if (i>1) % this is not the first user
        if (D(i-1) < A(i)) D(i) = A(i) + S(i);
        else
            D(i) = D(i-1) + S(i);
        end
    else
        D(i) = A(i)+S(i);
    end
    %
    % compute waiting time
    W(i) = D(i) - A(i) - S(i);
end
%
% Compute N(t)
T    = [];  % time axis
T(1) = 0;   % time origin
N    = [];  % number of cutomers
N(1) = 0;   % initial condition
k    = 2;   % place for next insert
A_max = A(length(A)); % last arrival instant
i = 1;      % index for arrivals
j = 1;      % index for departures
t = 0;      % system time

while (t < A_max)
    t = min(A(i), D(j));
    if (t == A(i))
        N(k) = N(k-1) + 1;
        T(k) = t;
        k    = k + 1;
        i    = i + 1;  % get next arrival
    else % departure occurs
        N(k) = N(k-1) - 1;
        T(k) = t;
        k    = k + 1;
        j    = j + 1; % get next departure
    end
end
%
% record remaining departure instants
for i=j:1:length(D)
    t    = D(i);
    N(k) = N(k-1) - 1;
    T(k) = t;
    k    = k + 1;
end

k = k - 1; % decrement k to get real size of N and T
%
% compute means
MeanW                = mean(W);
T_Intervales         = T(2:k)-T(1:k-1);
MeanN                = sum(N(1:k-1).*T_Intervales) / T(k);
IdleDurationsIndex = find(N(1:k-1) ~= 0);
Utilization          = sum(T_Intervales(IdleDurationsIndex))/T(k);
%
```


Example: Queueing System – cont'd

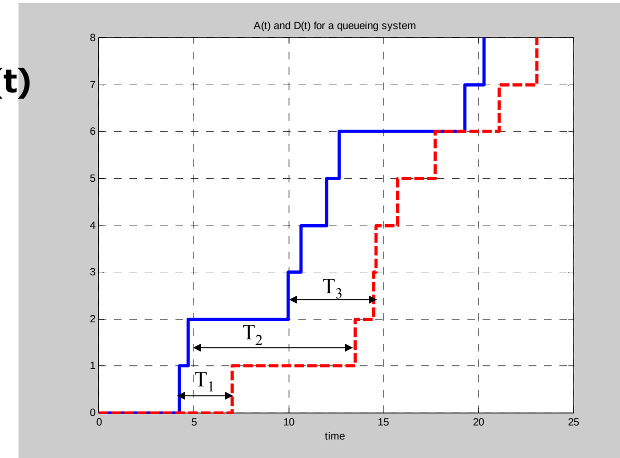
```
% Display results
fprintf('Block #:  '); fprintf('%3d  ', [1:1:length(A)]); fprintf('\n');
fprintf('Arrivals: '); fprintf('%3d  ', A); fprintf('\n');
fprintf('Errors:   '); fprintf('%3d  ', Errors); fprintf('\n');
fprintf('Service:  '); fprintf('%3d  ', S); fprintf('\n');
fprintf('Departs:  '); fprintf('%3d  ', D); fprintf('\n');
fprintf('Waiting:  '); fprintf('%3d  ', W); fprintf('\n');
fprintf('\n\n');
fprintf('Average no of customers in system = %7.3f\n', MeanN);
fprintf('Average customer waiting time     = %7.3f microsec\n', MeanW);
fprintf('Maximum simulation time           = %7.3f microsec\n', T(k));
fprintf('Duration server busy              = %7.3f microsec\n', ...
         sum(T_Intervales(IdleDurationsIndex)));
fprintf('Server utilization                = %7.4f\n', Utilization);
fprintf('Server idle                       = %7.4f\n',1.0-Utilization);
%
% Plot results
figure(1)
h = stairs(T, N); grid
set(h, 'LineWidth', 3);
xlabel('Time');
ylabel('No of customers in system, N(t)');

figure(2);
[AT, AA] = stairs(A, cumsum(ones(size(A))));
[DT, DD] = stairs(D, cumsum(ones(size(D))));
[NT, NN] = stairs(T, N);
h = plot(AT, AA, '-', DT, DD,'--r', NT, NN,'-.'); grid
set(h, 'LineWidth', 3);
title('Queue sysystem simulation');
ylabel('No of customers');
xlabel('Time');
legend('A(t)', 'D(t)', 'N(t)', 0);

figure(3);
h = stem(W); grid
set(h, 'LineWidth', 3);
ylabel('Waiting time');
xlabel('Customer index');
LegendStr = ['MeanW = ' num2str(MeanW)];
legend(LegendStr, 0);
```

Number of Customers in System

- **Blue curve: A(t)**
- **Red curve: D(t)**
- **Total time spent in the system for all customers = area in between two curves**



Little's Formula

- **Little's formula:**

$$E[n] = \lambda E[r]$$

Holds for many service disciplines and for systems with arbitrary number of servers. It holds for many interpretations of the system as well

Example 1:

- **<u>Problem</u>: Let $n_s(t)$ be the number of customers being served at time t, and let τ denote the service time. If we designate the set of servers to be the "system" then Little's formula becomes:**

$$E[n_s] = \lambda E[s]$$

Where $E[n_s]$ is the average number of busy servers for a system in the steady state.

Example 1: cont'd

Note: for a single server $n_s(t)$ can be either 0 or 1 → $E[n_s]$ represents the portion of time the server is busy. If p_0 = Prob[$n_s(t)$ = 0], then we have

$$1 - p_0 = E[n_s] = \lambda E[s], \text{ Or}$$
$$p_0 = 1 - \lambda E[s]$$

The quantity $\lambda E[s]$ is defined as the utilization, U, for a single server. Usually, it is given the symbol ρ

$$\rho = \lambda E[s]$$

For a c-server system, we define the utilization (the fraction of busy servers) to be

$$\rho = \lambda E[s] / c$$

The M/M/1 Queue

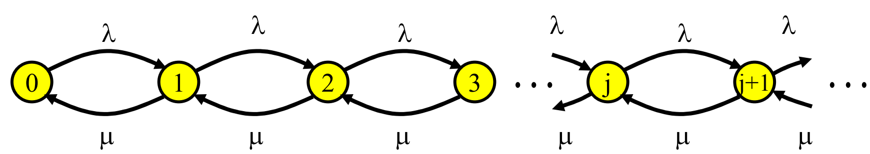
- **Consider a single server system where customers arrive according to a Poisson process of rate λ**
  - **➔ inter-arrival times are iid exponential r.v. with mean $E[\tau] = 1/\lambda$**
- **Assume the service times are iid exponential r.v. with mean $E[s] = 1/\mu$**
- **Assume the inter-arrival times and service times are independent**
- **Assume the system can accommodate unlimited number of customers**

The M/M/1 Queue – cont'd

- **What is the steady state pmf of n(t), the number of customers in the system?**
- **What is the PDF of *r*, the total customer delay in the system or the response time (as used in the textbook)?**

The M/M/1 Queue – cont'd

- **Consider the transition rate diagram for M/M/1 system**



- **Note:**
  - **System state – number of customers in systems**
  - **λ is rate of customer arrivals**
  - **μ is rate of customer departure**

The M/M/1 Queue – Distribution of Number of Customers

- **Writing the global balance equations for this Markov chain and solving for Prob[n(t) = j], yields (refer to previous example)**

$$p_j = \text{Prob}[n(t) = j]$$
$$= (1-\rho)\rho^j \quad \text{for } j=0,1,2,\ldots,$$

for $\rho = \lambda/\mu < 1$

Note that for $\rho = 1$ → arrival rate λ = service rate μ

The M/M/1 Queue – Expected Number of Customers

- **The mean number of customers (in the whole system = buffer + server) is given by**

$$E[n] = \sum_j j \, Prob[N(t) = j]$$

$$= \rho / (1-\rho) \qquad \text{for } \rho < 1$$

- **You can also show that the variance is equal to**

$$Var[n] = \sum_j (j-E[n])^2 \, Prob[n(t) = j]$$

$$= \rho / (1-\rho)^2 \qquad \text{for } \rho < 1$$

The M/M/1 Queue – Mean Customer Delay

- **The mean total response time for the system, E[r], is found using Little's formula**

$$E[r] = E[n] / \lambda$$
$$= (1/\mu) / (1-\rho)$$
$$= 1/(\mu - \lambda)$$

- **Actually, the response time CDF can be shown to be**

$$F(r) = 1 - e^{-r\mu(1-\rho)} \qquad r \geq 0$$

- **Can you fine $F^{-1}(r)$ for this CDF?**

The M/M/1 Queue – Mean Queueing Time

- **The mean waiting time in queue, E[w], is given by**

$$\mathbf{E[w] = E[r] - E[s]}$$
$$\mathbf{= \rho / (1-\rho)\ \ E[s]}$$

- **The CDF for the waiting time can be shown to be**

$$F(w) = 1 - \rho e^{-w\mu(1-\rho)} \qquad w \geq 0$$

- **Can you fine $F^{-1}(w)$ for this CDF?**

The M/M/1 Queue – Mean Number in Queue

- **Again we employ Little's formula:**

$$\mathbf{E[n_q] = \lambda E[w]}$$

Another way of finding E[n_q] is using

$$E[n_q] = \sum_{n=1}^{\infty}(n-1)p_n = \sum_{n=1}^{\infty}(n-1)(1-\rho)\rho^n$$

$$\mathbf{= \rho^2/(1-\rho)}$$

Remember:

server utilization $\rho = \lambda/\mu = 1-p_0$

All previous quantities E[n], E[r], E[w], and E[n_q] $\rightarrow \infty$ as $\rho \rightarrow 1$

Busy Period

- **Def: The time interval between two successive idle intervals**

- **Mean busy period =** $$\frac{1}{\mu(1-\rho)}$$

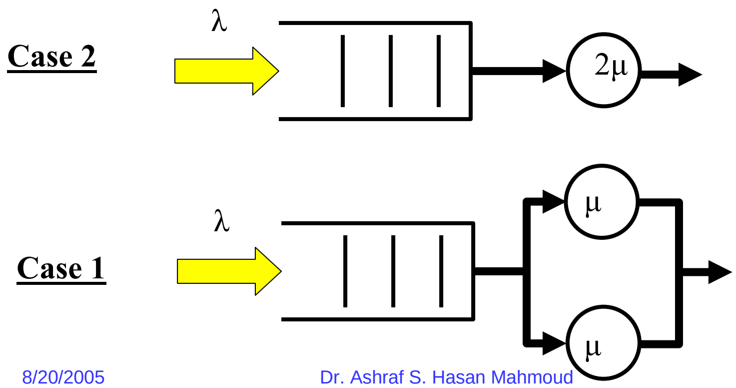
- **The textbook provides lots of other formulas in regard to the busy period.**

Scaling Effect for M/M/1 Queues

- **Consider a queue of arrival rate λ whose service rate is μ**
  - $\rho = \lambda/\mu$,
  - **The expected delay E[r] is given by**
    $$E[r] = (1/\mu) / (1-\rho)$$
- **If the arrival rate increases by a factor of K, then we either**
  1. **Have K queueing systems, each with a server of rate μ**
  2. **Have one queueing system with a server of rate K μ**
- **Which of the two options will perform better?**

Scaling Effect for M/M/1 Queues – cont'd

- **Example: K = 2: M/M/1 and M/M/2 systems with the same arrival rate and the same maximum processing rate**



Scaling Effect for M/M/1 Queues – cont'd

- **Case 1: K queueing systems**
  - **Identical systems**
  - **E[r] is the same for all – E[r] = (1/ μ) / (1- ρ)**
- **Case 2: 1 queueing system with server of rate Kμ**
  - **ρ for this system = (Kλ) /(K μ) = λ/ μ – same as the original system**
  - **E[r′] = (1/(K μ)) / (1- ρ) = (1/K) E[r]**
- **Therefore, the second option will provide a less total delay figure – significant delay performance improvement!**

Example 31.1

- **<u>Problem:</u> On a network gateway, measurements show that the packets arrive ata mean rate of 125 packets per second (pps) and the gateway take about 2 milliseconds to forward them. Using an M/M/1 model:**

(A) Analyze the gateway.

(B) What is the probability of buffer overflow had only 13 buffers?

(C) How many buffers do we need to keep packet loss below one packet per million?

Example 31.1 – cont'd

- **Solution:**

A) Analyzing the gateway:
Arrival rate, λ = 125 pps
Service rate, $\mu = 1/0.002 = 500$ pps
➔ Gateway utilization, $\rho = \lambda / \mu = 0.25$
Prob[n packets in gateway] = $p_n = (1-\rho)\rho^n$
$= (0.75)(0.25)^n$
Mean # of packets in gateway, $E[n] = \rho/(1-\rho)$
$= 0.25/0.75 = 0.33$
Mean time in gateway, $E[r] = 1/(\mu-\lambda) = 1/(500-125) = 2.66$ msec

B) Prob[buffer overflow] = Prob[more than 13 packets in gateway]
$= p_{14}+p_{15}+p_{16}+ \ldots$
$= (1-\rho)\{\rho^{14} + \rho^{15} + \rho^{16} + \ldots\}$
$= (1-\rho)\,\rho^{14}\{1 + \rho + \rho^2 + \ldots\}$
$= \rho^{14} = 3.7\times10^{-9}$

C) To limit the probability of loss to less than 10^{-6} ➔ Using (B)
Prob[buffer overflow] $\leq 10^{-6}$
➔ Prob[more than n packets in gateway] $\leq 10^{-6}$
➔ $\rho^{n+1} \leq 10^{-6}$ ➔ $n \leq \log(10^{-6})/\log(\rho) - 1 = 9.97 - 1$ ➔ choose n = 9 buffers

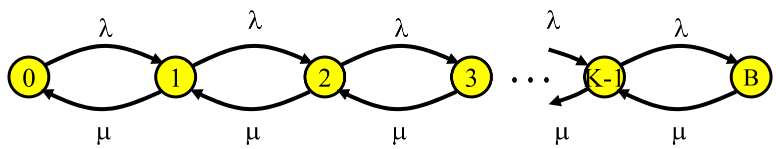
Note –
(1) the solution in the textbook has typos in parts (B) and (C)
(2) The solution for parts (B) and (C) are approximate – the more accurate model should be M/M/1/B – see the solution for this example after the M/M/1/B material

M/M/1/B – Finite Capacity Queue

- Consider an M/M/1 with finite capacity B $< \infty$
- For this queue – there can be at most B customers in the system
  - 1 being served
  - B-1 waiting
- A customer arriving while the system has B customers is BLOCKED (does not wait)!

M/M/1/B – Finite Capacity Queue – cont'd

- Transition rate diagram for this queueing system is given by:
  - n(t) - A continuous-time Markov chain which takes on the values from the set {0, 1, …, B}



M/M/1/B – Finite Capacity Queue – cont'd

- **The global balance equations:**

  $\lambda \quad p_0 = \mu p_1$

  $(\lambda + \mu)p_j = \lambda p_{j-1} + \mu p_{j+1}$ **for j=1, 2, …, B-1**

  $\mu \quad p_B = \lambda p_{B-1}$

➔ Prob[n(t) = j] = p_j **j=0,1, …, B;** $\rho < 1$

$= (1- \rho)\, \rho^{j}/(1- \rho^{B+1})$

When $\rho = 1$, $p_j = 1/(B+1)$ **(all states are equiprobable)**

Will this system become unstable for ρ = 1? Why?

- **Two important numbers:** p_0 **and** p_B
  - p_0 **is the probability of the server being idle –** $p_0 = (1- \rho)/(1- \rho^{B+1})$
  - p_B **is the probability of an arrival being blocked (or system overflow) –** $p_B = (1- \rho)\, \rho^{B}/(1- \rho^{B+1})$

M/M/1/B – Mean Number of Customers

- **Mean number of customers, E[n] is given by:**

$$E[n] = \sum_{j=0}^{B} j \Pr[n(t) = j]$$

$$= \begin{cases} \dfrac{\rho}{1-\rho} - \dfrac{(B+1)\rho^{B+1}}{1-\rho^{B+1}} & \rho < 1 \\ B/2 & \rho = 1 \end{cases}$$

M/M/1/B – Blocking Rate

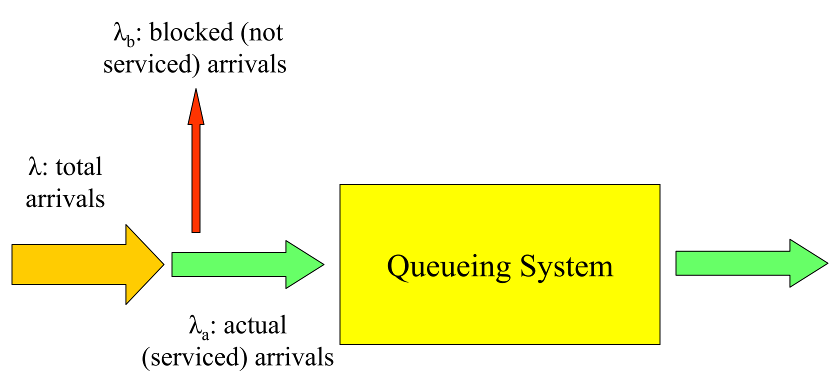
- **A customer arriving while the system is in state K is BLOCKED (does not wait)!**
- **Therefore, rate of blocking, λ_b is given by**

$$\lambda_b = \lambda p_B$$

- **The actual arrival rate into the system is λ_a given**

$$\lambda_a = \lambda - \lambda_b$$
$$= \lambda(1 - p_B)$$

M/M/1/B – Blocking Rate – cont'd



M/M/1/B – Mean Response Time

- **The mean total response time, E[r] is given by**

$$E[r] = E[n] / \lambda_a$$

Example 31.1 – Redoing Part (B) and (C)

- **<u>Problem:</u> On a network gateway, measurements show that the packets arrive ata mean rate of 125 packets per second (pps) and the gateway take about 2 milliseconds to forward them.**

(A) Analyze the gateway.

(B) What is the probability of buffer overflow had only 13 buffers?

(C) How many buffers do we need to keep packet loss below one packet per million?

Example 31.1 – cont'd

- **Solution:**

B) Prob[buffer overflow] = p_{13}

$= (1-\rho)\,\rho^{13}/(1-\rho^{13+1})$

$= 1.1\times10^{-8}$

C) To limit the probability of loss to less than 10^{-6} ➔ Using (B)
Prob[buffer overflow] $\le 10^{-6}$

➔ $p_B \le 10^{-6}$

➔ $(1-\rho)\,\rho^{B}/(1-\rho^{B+1}) \le 10^{-6}$

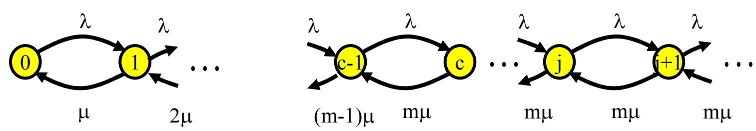
➔ $\rho^{B}/(1-\rho^{B+1}) \le 10^{-6} / (1-\rho) = 1.33\times10^{-6}$

The above can be solved numerically for the value of B

➔ $B \ge 9.6$ = 10 buffers

Multi-Server Systems: M/M/m

- **The transition rate diagram for a multi-server M/M/m queue is as follows:**
  - **Departure rate = $k\mu$ when k servers are busy**



Multi-Server Systems: M/M/m – cont'd

- **Writing the global balance equations:**

  $\lambda \quad p_0 = \mu p_1$

  $j\mu \quad p_j = \lambda p_{j-1} \quad$ for $j=1, 2, \ldots, m$

  $m\mu \quad p_j = \lambda p_{j-1} \quad$ for $j= m, m+1, \ldots$

➔

$p_j = a^j/j!\ p_0 \qquad$ (for $j=1, 2, \ldots, m-1$) and

$p_j = \rho^{j-m}/m!\ a^m\ p_0 \qquad$ (for $j=m, m+1, \ldots$)

where $a = \lambda/\mu$ and $\rho = a/m$

Multi-Server Systems: M/M/m – cont'd

- **To find p_0, we resort to the fact that $\Sigma\, p_j = 1$**

➔

$$p_0 = \left\{ \sum_{j=0}^{m-1} \frac{a^j}{j!} + \frac{a^m}{m!}\frac{1}{1-\rho} \right\}^{-1}$$

Two important quantities:
- p_0, and
- Prob[W >0]

The probability that an arriving customer has to wait

Prob[W > 0] = Prob[N ≥ m]

$= p_m + p_{m+1} + p_{m+2} + \ldots$

$= p_m/(1- \rho)$

$= p_0\, a^m/\{m!\,(1-\rho)\}$

In the textbook, this quantity is denoted by

Erlang-C formula

Multi-Server Systems: M/M/m – cont'd

- **The mean number of customers in queue (waiting):**

$$E[n_q] = \sum_{j=m}^{\infty}(j-m)\Pr[n(t)=j]$$

$$= \sum_{j=m}^{\infty}(j-m)\rho^{j-m} p_m$$

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

$$= \frac{\rho}{1-\rho}\Pr[W>0]$$

Multi-Server Systems: M/M/m – cont'd

- **The mean waiting time in queue:**

$$E[w] = E[n_q]/\lambda$$

- **The mean total response time in system:**

$$E[r] = E[w] + E[s]$$

$$= E[w] + 1/\mu$$

- **The mean number of customers in system:**

$$E[n] = \lambda E[r]$$

$$= E[n_q] + a$$

Why?

- **In addition to the formulas above the textbook gives the corresponding formulas for the variances too.**

Example 2:

- **A company has a system with four private telephone lines connecting two of its sites. Suppose that requests for these lines arrive according to a Poisson process at rate of one call every 2 minutes, and suppose that call durations are exponentially distributed with mean 4 minutes. When all lines are busy, the system delays (i.e. queues) call requests until a line becomes available. Find the probability of having to wait for a line.**

Example 2: cont'd

- **Solution:**
  $\lambda = ½, 1/\mu = 4, m = 4 \rightarrow a = \lambda/\mu = 2$
  $\rightarrow \rho = a/m = ½$

$p_0 = \{1+2+2^2/2!+2^3/3!+2^4/4!\ (1/(1-\rho))\}^{-1}$
$= 3/23$

$p_m = a^m/m!\ p_0$
$= 2^4/4! \times 3/23$

$\text{Prob}[W > 0] = p_m/(1-\rho)$
$= 2^4/4! \times 3/23 \times 1/(1-1/2)$
$= 4/23$
≈ 0.17

Waiting/Response Time Distribution for M/M/m

- **See textbook**

$$F(w) = 1 - \frac{p_c}{1-\rho} e^{-m\mu(1-\rho)w} \qquad w > 0$$

- **The q-percentile can be computed as follows:**

$$\begin{aligned} w_q &= \max\left\{0, \frac{1}{m\mu(1-\rho)} \ln\left(\frac{100 * \Pr[W>0]}{100-q}\right)\right\} \\ &= \max\left\{0, \frac{E[w]}{\Pr[W>0]} \ln\left(\frac{100 * \Pr[W>0]}{100-q}\right)\right\} \end{aligned}$$

Example 13.2:

- **<u>Problem</u>: Students arrive at the university computer center in a Poisson manner at an average rate of 10 per hour. Each student spends an average of 20 minutes at the terminal, and the time can be assumed to be exponentially distributed. The center currently has five terminals. Some students have been complaining that waiting times are too long.**

- **Analyze the center using a queueing model.**

$$p_0 = \left\{ \sum_{j=0}^{m-1} \frac{a^j}{j!} + \frac{a^m}{m!} \frac{1}{1-\rho} \right\}^{-1}$$

Example 13.2: cont'd

- **Solution**:

The center model: M/M/5
Arrival rate, λ = 1/6 student/min
Service rate, μ = 1/20 student/min
➔ Center utilization: a = λ / μ = 3.3333
ρ = a/m = 3.3333/5 = 0.6667 ← avg server utilization

p_0 = {1 +10/3+(10/3)^2/2!+(10/3)^3/3!+ (10/3)^4/4!+(10/3)^5/5!/(1-2/3)}$^{-1}$
= {31.4938}$^{-1}$ = 0.0318

p_m = a^m/m! p_0
= (10/3)^5 / 5! * 0.0318
= 0.1091

Prob[W>0] = p_m/(1- ρ)
= 0.1091/(1-2/3)
= 0.3271 ← This is the probability that a student has to wait!

$$E[n_q] = \frac{\rho}{1-\rho} \Pr[W > 0]$$

Example 13.2: cont'd

- **Solution**:

Avg # of students waiting in the center, E[n_q]
E[n_q] = ρ/(1- ρ)Prob[W>0]
= (2/3)/(1-2/3)* 0.3271
= 0.6543 students

Avg waiting time for a student, E[w]
E[w] = E[n_q]/ λ
= 0.6543 / (1/6)
= 3.9259 ≈ 4 minutes

Avg time spent in the centre, E[r]
E[r] = E[w] + E[s]
= 4 + 1/(1/20)
= 24 minutes

Avg/variance # of students in the center, E[n]
E[n] = λ E[r]
= (1/6) * 24 = 4 students

Var[n] = … = 479 students^2

$$w_q = \max\left\{0, \frac{1}{m\mu(1-\rho)}\ln\left(\frac{100*\Pr[W>0]}{100-q}\right)\right\}$$

Example 13.2: cont'd

- **Solution:**

Avg # of students using the terminals, E[n_s]

E[n_s] = E[n] – E[nq]
= 4 - 0.6543
= 3.35 students

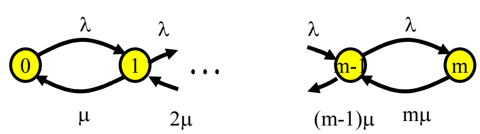
The 90-percentile of the waiting time, $w_{0.9}$

w0.9 = max{0, 4/0.3271 ln(10*0.3271)}
= 14 minutes

THUS, only 10% of the students have to wait more than 14 minutes!!

Multi-Server Systems: M/M/m/m

- **The transition rate diagram for a multi-server with no waiting room (M/M/m/m) queue is as follows:**
  - **Departure rate = kμ when k servers are busy**



PMF for Number of Customers for M/M/m/m

- **Writing the global balance equations, one can show:**

$$p_j = a^j/j!\ p_0 \qquad \text{(for } j=0, 1, \ldots, m\text{)}$$

where a = λ/μ (the offered load)

- **To find p_0, we resort to the fact that $\sum p_j = 1$**

$$p_0 = \left\{ \sum_{j=0}^{m} \frac{a^j}{j!} \right\}^{-1}$$

Erlang-B Formula

- **Erlang-B formula is defined as the probability that all servers are busy:**

$$\Pr[n = m] = p_m$$
$$= \frac{a_m / m!}{1 + a + a^2/2! + \ldots + a^m/m!}$$

Expected Number of customers in M/M/m/m

- **The actual arrival rate *into* the system:**

$$\lambda_a = \lambda(1 - p_m)$$

- **Average total delay figure:**

$$E[r] = E[s] = 1/\mu$$

Why?

- **Average number of customers:**

$$E[n] = \lambda_a E[s] = \lambda_a / \mu$$