King Fahd University of Petroleum & Minerals
Computer Engineering Dept

COE 541 – Design and Analysis of Local Area Networks

Term 041

Dr. Ashraf S. Hasan Mahmoud

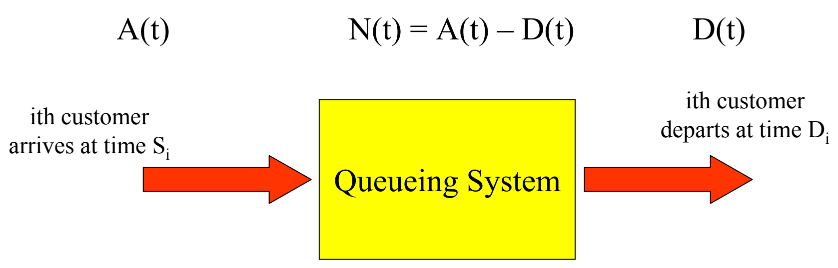
Rm 22-148-3

Ext. 1724

Email: ashraf@ccse.kfupm.edu.sa

Queuing Model

- **Consider the following system:**



$T_i = D_i - A_i$

$W_i = T_i - S_i$
$= D_i - A_i - S_i$

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]
T_i – duration of time spent in system for ith customer
W_i – duration of time spent waiting for service for ith customer

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

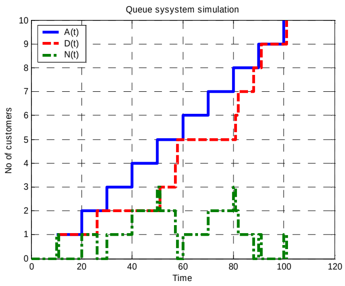
Solution:

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

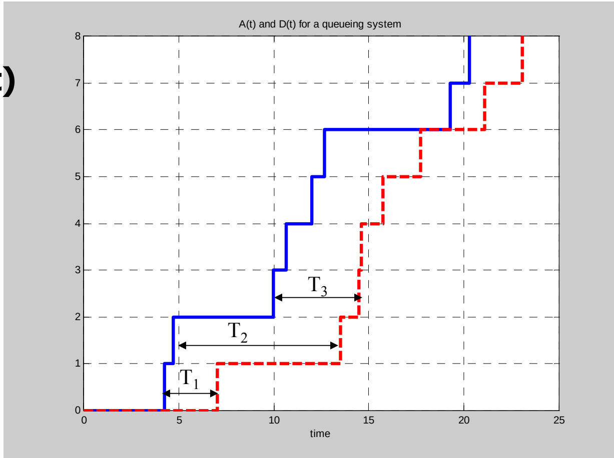
```
% Display results
fprintf('Block #:  '); fprintf('%3d  ', [1:1:length(A)]); fprintf('\n');
fprintf('Arrivals: '); fprintf('%3d  ', A); fprintf('\n');
fprintf('Errors:   '); fprintf('%3d  ', Errors); fprintf('\n');
fprintf('Service:  '); fprintf('%3d  ', S); fprintf('\n');
fprintf('Departs:  '); fprintf('%3d  ', 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:**

$$E[N] = \lambda E[T]$$

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 Ns(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"m then Little's formula becomes:**

$$E[Ns] = \lambda E[\tau]$$

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

Example 1: cont'd

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

$$1 - p_0 = E[Ns] = \lambda E[\tau], \text{ Or}$$
$$p_0 = 1 - \lambda E[\tau]$$

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

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

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

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

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 $1/\lambda$**
- **Assume the service times are iid exponential r.v. with mean $1/\mu$**
- **Assume the inter-arrival times and service times are independent**
- **Assume the system can accommodate unlimited number of customers**

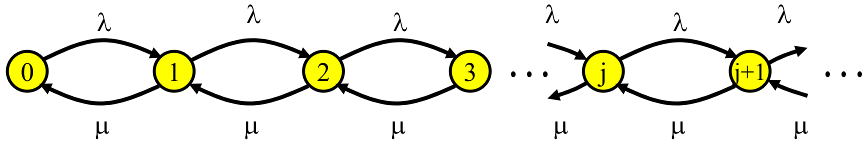
The M/M/1 Queue – cont'd

- **What is the steady state pmf of N(t), the number of customers in the system?**

- **What is the PDF of T, the total customer delay in the system?**

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)**

$$p_j = \text{Prob}[N(t) = j]$$
$$= (1-\rho)\rho^j$$

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 customer is given by**

$$E[N] = \sum_j j \, \text{Prob}[N(t) = j]$$

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

The M/M/1 Queue – Mean Customer Delay

- **The mean total customer delay in the system is found using Little's formula**

$$\begin{aligned} E[T] &= E[N]/\lambda \\ &= (\rho/\lambda) / (1-\rho) \\ &= 1/(\mu - \lambda) \end{aligned}$$

The M/M/1 Queue – Mean Queueing Time

- **The mean waiting time in queue is given by**

$$\begin{aligned} E[W] &= E[T] - E[\tau] \\ &= \rho / (1-\rho) \quad E[\tau] \end{aligned}$$

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

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

$$E[Nq] = \lambda E[W]$$

$$= \rho^2 / (1-\rho)$$

Remember:

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

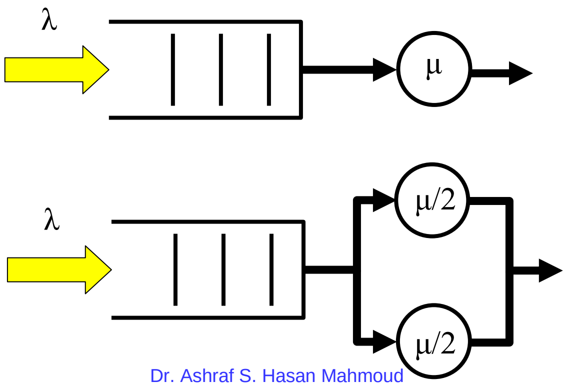
All previous quantities E[N], E[T], E[W], and E[Nq] → ∞ as $\rho \rightarrow 1$

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[T] is given by**
    $E[T] = (1/\mu) / (1-\rho)$
- **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 $K\mu$**
- **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[T] is the same for all – E[T] = $(1/\mu)$ / $(1-\rho)$**
- **Case 2: 1 queueing system with server of rate $K\mu$**
  - **ρ for this system = $(K\lambda)$ /$(K\mu)$ = λ/μ – same as the original system**
  - **E[T′] = $(1/(K\mu))$ / $(1-\rho)$ = $(1/K)$ E[T]**
- **Therefore, the second option will provide a less total delay figure – significant delay performance improvement!**

Arriving Customer's Distribution

- **Let Na be the number of customers found in the system by a customer arrival**
- **Prob[Na = k] ← is the arriving customer distribution**
- **(Refer to handout for proof) –**

$$\text{Prob}[Na = k] = \text{Prob}[N(t) = k] = (1-\rho)\rho^k$$

where Prob[N(t) = k] is the customer distribution at any time!! –

- **This is valid only for a POISSON ARRIVAL!**

Delay Distribution for M/M/1

- **We have shown before the mean delay, E[T] = $(1/\mu) / (1-\rho)$**
  - **But what is the distribution for T?**
- **An arriving customer see's k customers ahead**
  - **Has to wait for k iid exp r.v. service times, each with mean $1/\mu$**
  - **Then, our arriving customer will go to service for an exp r.v. service time of mean $1/\mu$**

Delay Distribution for M/M/1 – cont'd

- **Therefore, total delay, T, is the sum of k+1 iid exponential r.v. each with mean 1/μ**

- **The conditional (Na = k) distribution of T is given by the Gamma PDF (refer to Probability Theory slides)**

$$f_T(x/N_a = k) = \frac{(\mu x)^k}{k!}\mu e^{-\mu x} \qquad x > 0$$

Delay Distribution for M/M/1 – cont'd

- **The PDF of T can be found be de-conditioning on Na -**

$$\begin{aligned} f_T(x) &= \sum_{k=0}^{\infty} f_T(x/N_a = k)\Pr[N_a = k] \\ &= \sum_{k=0}^{\infty} \frac{(\mu x)^k}{k!}\mu e^{-\mu x}(1-\rho)\rho^k \\ &= (\mu - \lambda)e^{-(\mu-\lambda)x} \qquad x > 0 \end{aligned}$$

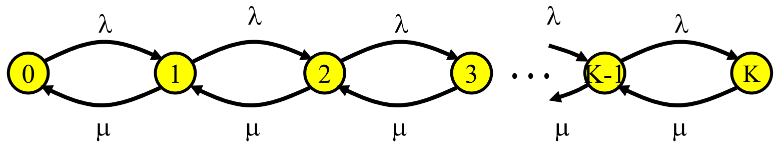
Therefore, the total delay, T, is a random variable *exponentially distributed* with mean 1/(μ-λ)

M/M/1/K – Finite Capacity Queue

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

M/M/1/K – 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, …, K}**



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

- **The global balance equations:**

  $\lambda p_0 = \mu p_1$

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

  $\mu p_K = \lambda p_{K-1}$

➔ Prob[N(t) = j] = p_j = $(1-\rho)\rho^j/(1-\rho^{K+1})$ j=0,1, …, K; $\rho<1$

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

M/M/1/K – Mean Number of Customers

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

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

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

M/M/1/K – Blocking Rate

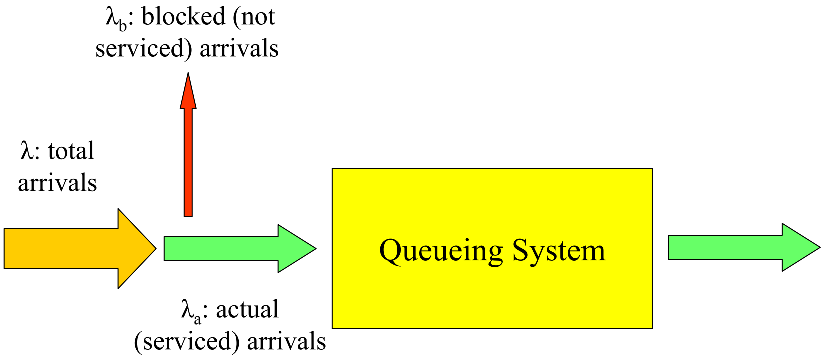
- **A customer arriving while the system is in state K is BLOCKED (does not wait)!**
- **Therefore, rate of blocking, λ_b is given by**

$$\lambda_b = \lambda\, p_K$$

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

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

M/M/1/K – Blocking Rate – cont'd



test

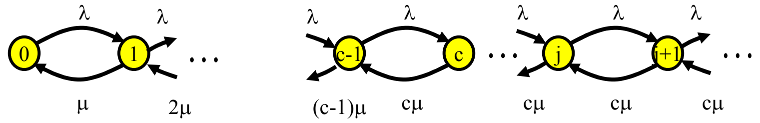
M/M/1/K – Mean Delay

- **The mean total delay E[T] is given by**

$$E[T] = E[N] / \lambda_a$$

Multi-Server Systems: M/M/c

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



Multi-Server Systems: M/M/c – cont'd

- **When k servers are busy, the time until the next departure is given by:**
  $X = \min(\tau_1, \tau_2, \ldots, \tau_k)$

where τ_i are iid exponential r.v. with mean $1/\mu$

The CDF for X is given by (refer to definition)

$Prob[X > t] = Prob[\min(\tau_1, \tau_2, \ldots, \tau_k) > t]$
$= Prob[\tau_1>t, \tau_2>t, \ldots, \tau_k>t]$
$= Prob[\tau_1>t]\ Prob[\tau_2>t] \ldots Prob[\tau_k>t]$
$= e^{-\mu t}\ e^{-\mu t} \ldots e^{-\mu t}$
$= e^{-k\mu t}$

Therefore, the time till the next departure (X) is an exponentially distributed r.v. with mean $1/(k\mu)$

Multi-Server Systems: M/M/c – cont'd

- **Writing the global balance equations:**

$\lambda \quad p_0 = \mu p_1$
$j\mu \quad p_j = \lambda p_{j-1}$ for $j=1, 2, \ldots, c$
$c\mu \quad p_j = \lambda p_{j-1}$ for $j= c, c+1, \ldots$

➔

$p_j= a^j/j!\ p_0$ (for $j=1, 2, \ldots, c$) and
$p_j= \rho^{j-c}/c!\ a^c\ p_0$ (for $j=c, c+1, \ldots$)

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

Multi-Server Systems: M/M/c – cont'd

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

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

The probability that an arriving customer has to wait

Prob[W > 0] = Prob[N ≥ c]
= $p_c + p_{c+1} + p_{c+2} + \ldots$
= $p_c/(1-\rho)$

Erlang-C formula

Multi-Server Systems: M/M/c – cont'd

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

$$E[N_q] = \sum_{j=c}^{\infty} (j-c) \Pr[N(t)=j]$$
$$= \sum_{j=c}^{\infty} (j-c)\rho^{j-c} p_c$$
$$= \frac{\rho}{(1-\rho)^2} p_c$$
$$= \frac{\rho}{1-\rho} \Pr[W>0]$$

Multi-Server Systems: M/M/c – cont'd

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

$$E[W] = E[N_q]/\lambda$$

- **The mean total delay in system:**

$$E[T] = E[W] + E[\tau]$$
$$= E[W] + 1/\mu$$

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

$$E[N] = \lambda E[T]$$
$$= E[N_q] + a$$

Why?

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:**
  $\lambda = ½$, $1/\mu = 4$, $c = 4$ ➔ $a = \lambda/\mu = 2$
  ➔ $\rho = a/c = ½$

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

$p_c = a^c/c!\ p0$
$= 2^4/4!\ X\ 3/23$

Prob[W > 0] = $p_c/(1-{}^{r})$
$= 2^4/4!\ X\ 3/23\ X\ 1/(1-1/2)$
$= 4/23$
≈ 0.17

Waiting Time Distribution for M/M/c

- **An arriving customer to the system, either**
  - **Does not wait, if number of busy servers is less than c**
  - **Does wait if number of busy servers is c**

- **If there are k > 0 customers waiting (as observed by an arriving customer), the total waiting time for the arriving customer = the sum of: remaining service time of the earliest job to finish + service time for these k customers**
  - **i.e. $W = \tau + \tau_1 + \tau_2 + \ldots + \tau_k$, where τ's ~ iid exponentially distributed r.v. with mean $E[\tau] = 1/\mu$**

Waiting Time Distribution for M/M/c – cont'd

Observation of an arriving customers

No of busy servers?

Less than c	Greater or equal to c
Prob = $p_0 + p_1 + \ldots + p_{c-1}$	Prob = $p_c + p_{c+1} + \ldots$
Customer does not wait, OR W = 0	Customer waits: If there are k waiting, then T = sum of k+1 iid exp r.v with parameter μ Remember k=0, 1, 2, …

Waiting Time Distribution for M/M/c – cont'd

- **We have seen before that (given there are k ahead), the distribution of W follows the gamma distribution with parameter cμ. I.e.**

$$f_W\left(x / N = c+k\right) = \frac{(c\mu x)^k}{k!} c\mu e^{-c\mu x} \qquad x > 0, k = 0,1,2,\ldots$$

Waiting Time Distribution for M/M/c – cont'd

- **We can find the overall pdf of W given N >= c (i.e. summing over all ks) as follows:**

$$f_W(x/W>0) = \sum_{k=0}^{\infty} f_W(x/N=c+k)\Pr[N=c+k] \qquad x>0$$

- **Equivalently, we can write:**

$$F_W(x/W>0) = \sum_{k=0}^{\infty} F_W(x/N=c+k)\Pr[N=c+k] \qquad x>0$$

Waiting Time Distribution for M/M/c – cont'd

- **But (refer to handout for proof)**

Pro[N = c + k/N ≥c] = (1-ρ)ρ^k k=0,1,2 …

- **Substituting in previous formula for F_W(x/W>0) and simplifying, yields**

$$F_W(x/W>0) = 1 - e^{-c(1-\rho)x} \qquad x>0$$

This is all assuming the customer will have to wait!!

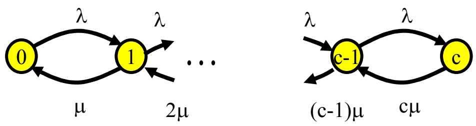
Waiting Time Distribution for M/M/c – cont'd

- **The general expression for the CDF (waiting and not waiting):**

$$F_W(x) = \Pr[W=0]\times 1 + F_W(x/W>0)\Pr[W>0]$$
$$= 1-\Pr[W>0]e^{-c\mu(1-\rho)x} \qquad x>0$$
$$= 1-\frac{p_c}{1-\rho}e^{-c\mu(1-\rho)x} \qquad x>0$$

Multi-Server Systems: M/M/c/c

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



PMF for Number of Customers for M/M/c/c

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

  $p_j = a^j/j!\ p_0$ (for j=0, 1, …, c)

where a = λ/μ (the offered load)

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

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

Erlang-B Formula

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

$$\Pr[N = c] = p_c$$
$$= \frac{a_c / j!}{1 + a + a^2/2! + \ldots + a^c/c!}$$

Expected Number of customers in M/M/c/c

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

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

- **Average total delay figure:**

$$E[T] = E[\tau]$$

Why?

- **Average number of customers:**

$$E[N] = \lambda_a E[\tau]$$

M/G/1 Queues

- **Poisson arrival process (i.e. exponential r.v. interarrival times)**
- **Service time: general distribution $f_\tau(x)$**
  - **For M/M/1, $f_\tau(x) = \mu e^{-\mu x}$ for $x > 0$**
- **The state of the M/G/1 system at time t is specified by**
  1. **N(t)**
  2. **The remaining (residual) service time of the customer being served**

The Residual Service Time

- **Mean residual time (see example and derivation in handout) is given by**

$$E[R] = \frac{E[\tau^2]}{2E[\tau]}$$

Mean Waiting Time in M/G/1

- **The waiting time of a customer is the sum of the residual service time R′ of the customer (if any) found in service and the Nq(t)= k-1 service time of the customers (if any) found in queue**

$$\begin{aligned} E[W] &= E[R'] + E[Nq]\,E[\tau] \\ &= E[R'] + \lambda E[W]E[\tau] \\ &= E[R'] + \rho\, E[W] \end{aligned}$$

Mean Waiting Time in M/G/1 – cont'd

- **But residual service time R' (as observed by an arriving customers) is either**
  - **0 is the server is free**
  - **R if the server is busy**

- **Therefore, mean of R' is given by**

$$\begin{aligned} E[R'] &= 0 \times Pro[N(t)=0] + E[R](1-Pro[N(t)=0]) \\ &= E[\tau^2]/(2E[\tau]) \times \rho \\ &= \lambda E[\tau^2]/2 \end{aligned}$$

Mean Waiting Time in M/G/1 – cont'd

- **Substituting back, yields**

$$E[W] = \frac{\lambda E[\tau^2]}{2(1-\rho)}$$

$$= \frac{\lambda(\delta^2_\tau + E[\tau]^2)}{2(1-\rho)}$$

$$= \frac{\rho\,(1 + C_\tau^2)}{2(1-\rho)} E[\tau]$$

Remember:
- $E[\tau^2] = \delta^2_\tau + E[\tau]^2$
- $C^2_\tau = \delta^2\tau/E[\tau]^2$

Pollaczek-Khinchin (P-K) Mean Value Formula

Mean Delay in M/G/1 – cont'd

- **The mean waiting time, E[T] is found by adding mean service time to E[W]:**

$$E[T] = E[\tau] + E[W]$$

$$= E[\tau] + \frac{\rho\,(1 + C_\tau^2)}{2(1-\rho)} E[\tau]$$

Example 3:

- **Problem: Compare E[W] for M/M/1 and M/D/1 systems.**

- **Answer:**

M/M/1: service time, τ, is exponential r.v. with parameter μ

$\rightarrow E[\tau] = 1/\mu$, $E[\tau^2] = 2/\mu^2$, $\delta^2_\tau = 1/\mu^2$, $C^2_\tau = 1$

M/D/1: service time, τ, is constant with value $\tau = 1/\mu$

$\rightarrow E[t] = 1/\mu$, $E[\tau^2] = 1/\mu^2$, $\delta^2_\tau = 0$, $C^2_\tau = 0$

Example 3: cont'd

- **Answer: cont'd**

Substitute in P-K mean value formula

M/M/1:

$$E[W_{M/M/1}] = \frac{\lambda E[\tau^2]}{2(1-\rho)} = \frac{\rho}{(1-\rho)} E[\tau]$$

M/D/1:

$$E[W_{M/D/1}] = \frac{\lambda E[\tau^2]}{2(1-\rho)} = \frac{\rho}{2\,(1-\rho)} E[\tau]$$

$$= \frac{1}{2} E[W_{M/M/1}]$$

The waiting time in an M/D/1 queue is half of that of an M/M/1 system

M/G/1 with Priority Service Discipline

- **Handles K priority classes of customers**
- **Head-of-line priority service discipline**
- **Type k ={1, 2, …, K} arrive according to Poisson arrival process**
- **A separate queue is kept for each priority class**
- **Server utilization from type k customers:**

$$\rho_k = \lambda_k E[\tau_k]$$

- **Total server utilization**

$$\rho = \rho_1 + \rho_2 + \ldots + \rho_K < 1$$

for a stable system

- **Assume class 1 is the highest priority while class K is the lowest**

Mean Waiting Time in M/G/1 with Priority Service Discipline

- **An arriving customer of type 1 finds $N_{q1}(t)$ = k1 type 1 customers in queue**
- **Assuming FCFS for each queue**
- **The mean waiting time for type one customer:**

$$E[W_1] = E[R''] + E[N_{q1}]\, E[\tau_1]$$

where E[R''] is the residual time of the customer (if any) found in service

Mean Waiting Time in M/G/1 with Priority Service Discipline – cont'd

- **We also know (Little's formula) that:**

$$E[N_{q1}] = \lambda_1\, E[W_1]$$

Substituting and solving for $E[W_1]$, yields,

$$E[W_1] = E[R''] \,/\, (1-\rho_1)$$

Mean Waiting Time in M/G/1 with Priority Service Discipline – cont'd

- **Consider a type 2 customer – Because of the priority scheme one can write**

$$E[W_2] = E[R''] + E[N_{q1}]\,E[\tau_1] + E[N_{q2}]\,E[\tau_2] + E[M_1]\,E[\tau_1]$$

Where

- $E[R'']$ is the residual time of the customer (if any) found in service
- $E[N_{q1}]\,E[\tau_1]$ time to service already existing class 1 customers (remember $E[N_{q1}] = \lambda_1 E[W_1]$)
- $E[N_{q2}]\,E[\tau_2]$ time to service already existing class 2 customers (remember $E[N_{q2}] = \lambda_2 E[W_2]$)
- $E[M_1]\,E[\tau_1]$ time to service class 1 customers arriving during our customer waiting time - $E[M_1] = \lambda_1 E[W_2]$

Mean Waiting Time in M/G/1 with Priority Service Discipline – cont'd

- **$E[M_1]$ is given by**

$$E[M_1] = \lambda_1 E[W_2]$$

- **Substituting and solving for $E[W_2]$, yields,**

$$E[W_2] = \frac{E[R'']}{(1 - \rho_1)(1 - \rho_1 - \rho_2)}$$

Mean Waiting Time in M/G/1 with Priority Service Discipline – cont'd

- **In general we can show the mean waiting time for a customer of type k, $E[W_k]$ is given by**

$$E[W_k] = \frac{E[R'']}{(1 - \rho_1 - \ldots - \rho_{k-1})(1 - \rho_1 \ldots - \rho_k)}$$

Mean Waiting Time in M/G/1 with Priority Service Discipline – cont'd

- **What is E[R'']?**
- **Remember R'' is the residual service time of a customer (if any) found in service – of any type**
- **Recall that mean residual time E[R''] is computed by**

$$E[R''] = \lambda E[\tau^2]/2 \quad \text{(refer to slide 52)}$$

But $E[\tau^2]$ for which type of customers?

Mean Waiting Time in M/G/1 with Priority Service Discipline – cont'd

- **$E[\tau^2]$ – is the mean service-time squared for ANY type:**

$$E[\tau^2] = (\lambda_1/\lambda)E[\tau_1^2] + (\lambda_2/\lambda)E[\tau_2^2] + \ldots + (\lambda_K/\lambda)E[\tau_K^2]$$

where $\lambda = \lambda_1 + \lambda_2 + \ldots \lambda_K$

Mean Waiting Time in M/G/1 with Priority Service Discipline – cont'd

- **Therefore, the mean waiting time of type k customers:**

$$E[W_k] = \frac{\sum_{j=1}^{K} \lambda_j E[\tau_j^2]}{2(1-\rho_1-\cdots-\rho_{k-1})(1-\rho_1-\cdots-\rho_k)}$$

- **The mean delay for type k customer is then equal to**

$$E[T_k] = E[W_k] + E[\tau_k]$$

M/G/1 Analysis Using Embedded Markov Chain

- **Pollaczek-Khinchin (P-K) Transform Equation**

$$G_N(z) = \frac{(1-\rho)(z-1)\hat{\tau}(\lambda(1-z))}{z-\hat{\tau}(\lambda(1-z))}$$

See derivation in handout

where:

- **$G_N(z)$: moment generating function of the r.v. N(t)**
- **$\hat{\tau}(s)$ is the Laplace transform of r.v. τ**

Example 4:

- **<u>Problem</u>: Use the P-K transform equation to find the steady state pmf of an M/M/1**
- **<u>Answer</u>:**

For an M/M/1 the steady state pmf for N(t) is given by (refer to slide 13)

$$p_j = \text{Prob}[N(t) = j] = (1-\rho)\rho^j$$

Example 4: cont'd

- **Answer: cont'd**

The moment generating function, $G_N(z)$, is then given by

$$G_N(z) = \sum_{j=0}^{\infty} p_j z^j$$
$$= \sum_{j=0}^{\infty} (1-\rho)\rho^j z^j$$
$$= \frac{(1-\rho)}{(1-\rho z)}$$

Example 4: cont'd

- **Answer: cont'd**

Now let's use the P-K transform and see if we get the same answer!

For M/M/1, τ is exp r.v → the pdf for τ is

$$f_\tau(t) = \mu e^{-\mu t} \qquad t > 0$$

The Laplace transform of τ is given by

$$\hat{\tau}(s) = \int_0^{\infty} f_\tau(t) e^{-st} dt$$
$$= \frac{\mu}{s+\mu}$$

Example 4: cont'd

- **Answer: cont'd**

Therefore, $\hat{\tau}(\lambda(1-z))$ is given by

$$\hat{\tau}(\lambda(1-z)) = \frac{\mu}{\lambda(1-z)+\mu}$$

We are now in a position to substitute in the P-K transform equation

Example 4: cont'd

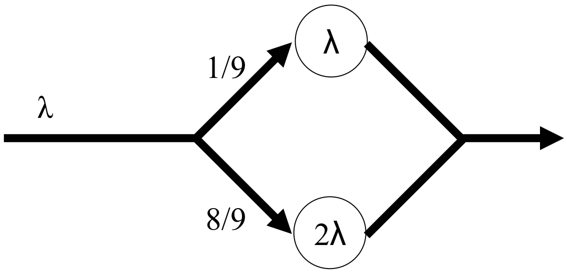
- **Answer: cont'd**

$$\begin{aligned}
G_N(z) &= \frac{(1-\rho)(z-1)\hat{\tau}(\lambda(1-z))}{z-\hat{\tau}(\lambda(1-z))} \\
&= \frac{(1-\rho)(z-1)(\mu/\lambda(1-z)+\mu)}{z-(\mu/\lambda(1-z)+\mu)} \\
&= \frac{(1-\rho)(z-1)\mu}{(\lambda-\lambda z+\mu)z-\mu} \\
&= \frac{(1-\rho)}{(1-\rho z)}
\end{aligned}$$

Which the same M.G.F for N(t) derived previously!

Example 5:

- **<u>Problem</u>: M/H$_2$/1**



What is Prob[N(t) = k] =?

Example 5: cont'd

- **<u>Answer</u>:**

The pdf of the service time, τ, is

$$f_\tau(t) = \frac{1}{9}\lambda e^{-\lambda t} + \frac{8}{9}2\lambda e^{-2\lambda t} \qquad t > 0$$

The mean service time, E[τ] is given by

E[τ] = (1/9)X 1/λ + (8/9)X 1/(2λ)
= 5/(9λ)

➔ ρ = λ E[τ] = 5/9

The Laplace transform is given by

$$\hat{\tau}(s) = \frac{1}{9}\frac{\lambda}{s+\lambda} + \frac{8}{9}\frac{2\lambda}{s+2\lambda}$$

and

$$= \frac{18\lambda^2 + 17\lambda s}{9(s+\lambda)(s+2\lambda)}$$

Example 5: cont'd

- **Answer**:

Substituting λ(1-z) for every s in the previous expression, and writing $G_N(z)$, yields,

$$G_N(z) = \frac{(1-\rho)(z-1)\hat{\tau}(\lambda(1-z))}{z-\hat{\tau}(\lambda(1-z))}$$

$$= \frac{(1-\rho)(35-17z)(z-1)}{9(2-z)(z-7/3)(z-5/3)}$$

Partial Fraction Expansion – How?

$$= (1-\rho)\left\{\frac{1/3}{1-3z/7} + \frac{2/3}{1-3z/5}\right\}$$

Example 5: cont'd

- **Answer**:

Therefore, $G_N(z)$ is given by

$$G_N(z) = (1-\rho)\left\{\frac{1}{3}\sum_{k=0}^{\infty}\left(\frac{3}{7}\right)^k z^k + \frac{2}{3}\sum_{k=0}^{\infty}\left(\frac{3}{5}\right)^k z^k\right\}$$

Since the coefficient of z^k is Prob[N(t) = k], then we finally have:

$$\Pr[N(t)=k] = \frac{4}{27}\left(\frac{3}{7}\right)^k + \frac{8}{27}\left(\frac{3}{5}\right)^k \qquad k = 0,1,\cdots$$

Total Delay Distribution for M/G/1 System

- **If, T is the total delay variable, then the Laplace transform of T is given by (see handout for derivation)**

$$\widehat{T}(s) = \frac{(1-\rho)s\widehat{\tau}(s)}{s-\lambda+\lambda\widehat{\tau}(s)}$$

P-K transform equation

- **The pdf for T, $f_T(t)$, is obtained by inverting the above expression analytically or numerically**

Waiting Time Distribution for M/G/1 System

- **Since T = W + τ ➔ Therefore,**

$$\widehat{T}(s) = \widehat{W}(s)\widehat{\tau}(s)$$

- **Hence, the Laplace transform of the waiting time is given by**

$$\widehat{W}(s) = \frac{(1-\rho)s}{s-\lambda+\lambda\widehat{\tau}(s)}$$

P-K transform equation

Example 6:

- **Problem: Verify the result obtained previously for the total delay time distribution of an M/M/1 queue using P-K transform equations for M/G/1 systems**

- **Answer: for M/M/1 the service time, τ, is exp r.v. ➔** $f_{\tau}(t) = \mu e^{-\mu t} \qquad t > 0$

or $$\hat{\tau}(s) = \frac{\mu}{s + \mu}$$

Example 6: cont'd

- **Substituting in the P-K transform equations**

$$\hat{T}(s) = \frac{(1-\rho)s\mu}{(s+\mu)(s-\lambda)+\lambda\mu}$$

$$= \frac{(1-\rho)\mu}{s-(\lambda-\mu)}$$

Inverting the above expression, yields

$$f_T(t) = \mu(1-\rho)e^{-\mu(1-\rho)t} \qquad t > 0$$

$$= (\mu-\lambda)e^{-(\mu-\lambda)t} \qquad t > 0$$

Example 6: cont'd

- This means the total delay is exponentially distributed with mean1/(μ-λ) – Same result as obtained before! (refer to slide 23)
- The waiting time is obtained using

$$\hat{W}(s) = \frac{(1-\rho)s}{s-\lambda+\lambda\hat{\tau}(s)}$$
$$= (1-\rho)\frac{s+\mu}{s+\mu-\lambda}$$
$$= (1-\rho)\left\{1+\frac{\lambda}{s+\mu-\lambda}\right\}$$

Example 6: cont'd

- Therefore the pdf of W is given by

$$f_W(t) = (1-\rho)\delta(t) + \lambda(1-\rho)e^{-\mu(1-\rho)t} \qquad t>0$$

- The δ(t) term indicates there is a ZERO waiting time with probability equal to 1-ρ – i.e. when server is free