King Fahd University of Petroleum & Minerals Computer Engineering Dept

COE 540 – Computer Networks

Term 112

Dr. Ashraf S. Hasan Mahmoud

Rm 22-420

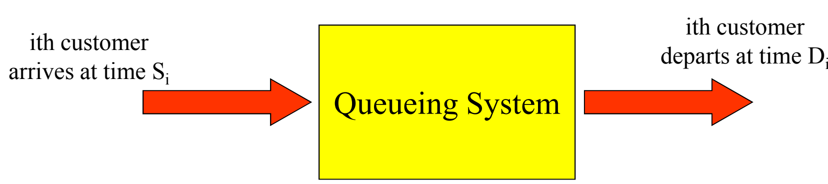
Ext. 1724

Email: ashraf@kfupm.edu.sa

Queuing Model

- **Consider the following system:**

A(t) $N(t) = A(t) - D(t)$ D(t)



$T_i = D_i - A_i$

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

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

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 2:

- **<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 2: 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

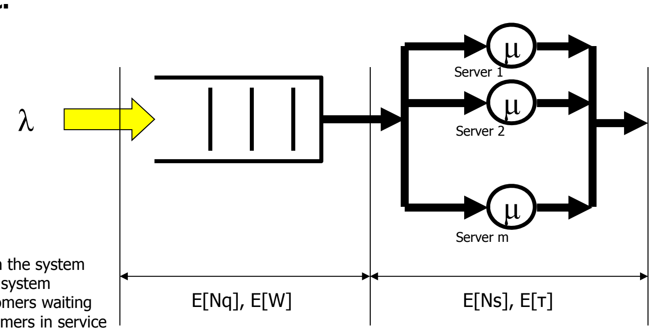
$$\rho = \lambda E[\tau]$$

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

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

Queue System and Parameters

- **Queueing system with m servers**
  - **When m = 1 – single server system**
- **Input: arrival statistics (rate λ), service statistics (rate μ), number of customers (m), buffer size**
- **Output: E[N], E[T], E[Nq], E[W], Prob[buffer size = x], Prob[W<w], etc.**



E[N] = mean # of customers in the system
E[T] = mean time spent in the system
E[Nq] = mean number of customers waiting
E[Ns] = mean number of customers in service
E[W] = mean waiting time for a customer
E[τ] = mean service time for a customer

E[N], E[T]
E[N] = E[Nq] + E[Ns],
E[T] = E[W] + E[τ]

The M/M/1 Queue

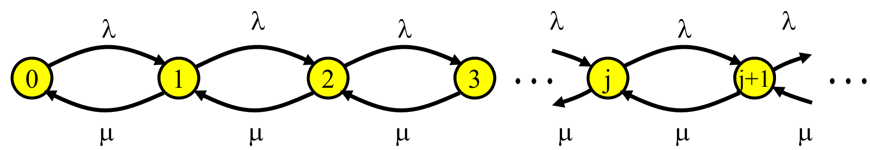
- **Consider m-server system where customers arrive according to a Poisson process of rate λ**
  - **➔ inter-arrival times are iid exponential r.v. with mean 1/λ**
- **Assume the service times are iid exponential r.v. with mean 1/μ**
- **Assume the inter-arrival times and service times are independent**
- **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**

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

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

The M/M/1 Queue – Mean Customer Delay

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

$$E[T] = E[N]/ \lambda$$
$$= \rho /[\lambda (1- \rho)]$$
$$= 1/\mu (1-\rho)$$
$$= 1/(\mu - \lambda)$$

The M/M/1 Queue – Mean Queueing Time

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

$$E[W] = E[T] - E[\tau]$$
$$= \rho / (1-\rho) \quad E[\tau]$$

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

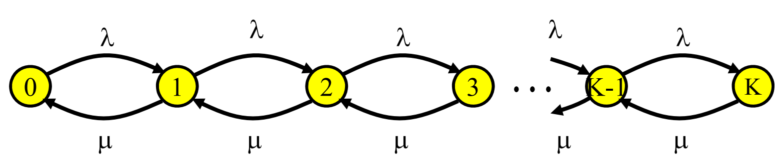
All previous quantities E[N], E[T], E[W], and E[Nq] $\rightarrow \infty$ as $\rho \rightarrow 1$

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)!**

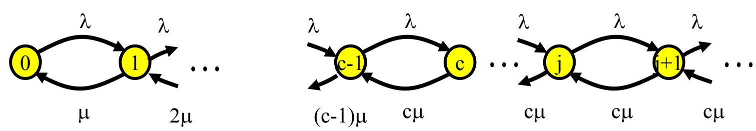
M/M/1/K – Finite Capacity Queue – cont'd

- **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}**



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**
  - **We can show that the service time for a customer finding k servers busy is exponentially distributed with mean 1/(kμ)**



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** $j= c, c+1, \ldots$

Note this distribution is the same as that for M/M/1 when you set c to 1.

➔

$p_j = a^j/j!\ p_0$ **(for j=1, 2, …, c) and**

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

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

- **From this we note that the probability of system being in state c, pc, is given by**

$$p_c = a^c/c!\ p_0$$

Variations of M/M/1 queue

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

Question: What is Prob[W>0] for M/M/1 system?

Variations of M/M/1 queue

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 5:

- **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.**
- **What is the average waiting time for an incoming call?**

Example 5: cont'd

- **Solution:**

  $\lambda = 1/2$, $1/\mu = 4$, $c = 4$ ➔ $a = \lambda/\mu = 2$

  ➔ $\rho = a/c = 1/2$

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

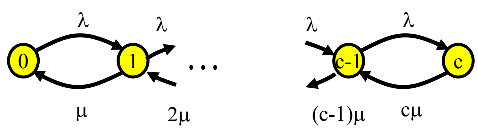
$p_c = a^c/c!\ p0$
$= 2^4/4!\ X\ 3/23$

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

(2) To find E[W], find E[Nq] …
E[Nq] = $\rho/(1-\rho)$ * Prob[W>0] = 0.1739
E[W] = E[Nq]/ λ = 0.35 min

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**



Variations of M/M/1 queue

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 $\sum p_j = 1$**

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

Variations of M/M/1 queue

Erlang-B Formula

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

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

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:**

Why?

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

- **Average number of customers:**

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

Example 6:

- **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 BLOCKS the incoming call and generates a busy signal.**
- **Find the probability of being blocked.**

Example 6:

- **Solution:**
  $\lambda = 1/2$, $1/\mu = 4$, $c = 4 \Rightarrow a = \lambda/\mu = 2$
  $\Rightarrow \rho = a/c = 1/2$

$$p_c = \frac{a^c/c!}{1 + a + a^2/2! + a^3/3! + a^4/4!}$$

$$= \frac{2^4/4!}{1 + 2 + 2^2/2! + 2^3/3! + 2^4/4!} = 9.5\%$$

Therefore, the probability of being blocked is 0.095.

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**

Mean Waiting Time in M/G/1

- **Main result**

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

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

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

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 7:

- **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

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

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

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

Example 7: 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

Example 8:

- **Problem: Assume traffic is arriving at the input port of a router according to a Poisson arrival process of rate λ = 100 packets/sec. If the traffic distribution is as follows:**

  30% of packets are 512 Bytes long,

  50% of packets are 1024 Bytes long,

  20% of packets are 4096 Bytes long

  If the transmit speed of the router output port is 1.5 Mb/s

a) What is the average packet transmit time?

b) What is the average packet waiting time before transmit?

c) What is the average buffer size in the router?

Example 8: cont'd

- **Solution:**

a) Average packet size,

$E[L] = 0.3x512 + 0.5x1024 + 0.2x4096$

$= 1484.8$ Bytes

average transmit time = $E[L]/R = 1484.8x8/1.5x10^6 = 0.0079$ sec

b) $E[L^2] = 0.3x(512x8)^2 + 0.5x(1024x8)^2 + 0.2x(4096x8)^2 = 2.5334e+008$ Bits^2

$E[\tau^2] = E[L^2]/R^2 = 1.1259e-004$ sec^2

$\rho = \lambda E[\tau] = 0.7919$

$E[W] = 0.5 \lambda E[\tau^2] / (1-\rho)$

$= 0.0271$ sec

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

$= 2.705$ packet