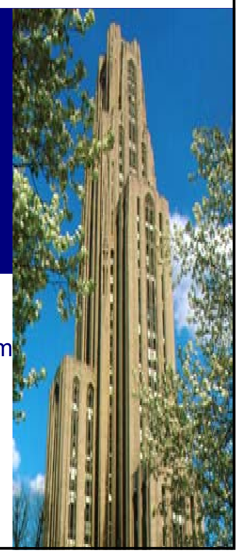
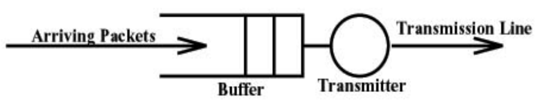
TELCOM 2130
Basic Queues



David Tipper
Associate Professor
Graduate Telecommunications and Networking Program
University of Pittsburgh

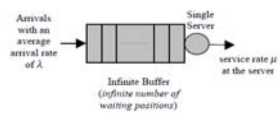
Nomenclature of a Queueing System



- The input process – how customers arrive
- The system structure – waiting space – number of servers, etc.
- The service process
- Kendall's Notation 1/2/3/4/5/6
  - Shorthand notation to describe a queueing system
  - 1 : Customer arriving pattern (Interarrival times distribution).
  - 2 : Service pattern (Service-times distribution).
  - 3 : Number of parallel servers.
  - 4 : System capacity.
  - 5 : Queueing discipline.
  - 6: Customer Population

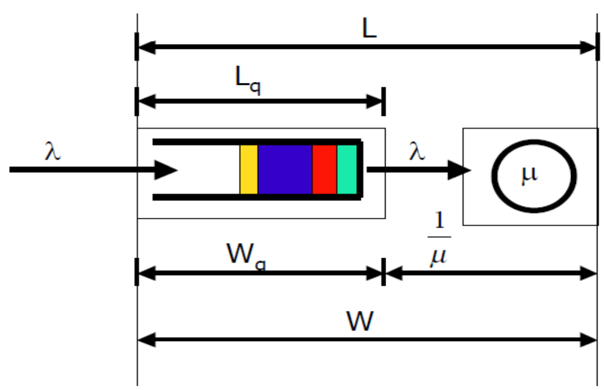
Nomenclature

- Standard notation
  - λ mean arrival rate of customers/time unit
  - μ mean service rate in customers/time unit
  - *n(t)* – number of customers in the system at time t
  - $\pi_i = \lim_{t \to \infty} P\{n(t) = i\}$
  - $\rho = \lambda/\mu$ is server utilization remember $\rho < 1$ for stability
  - *L* – Average number of customers in systems
  - L_q - Average number of customers in the queues
    know $L = L_q + \rho$
  - *W* – Average delay in system (includes server + queue)
  - W_q – Average delay in queue
    know $W = W_q + 1/\mu$



Nomenclature

- Standard notation – relationships



- What can one say in general about a queueing system?????

Resource Utilization and Traffic Intensity

- Resource Utilization is basically a measure of how busy the resource is.
  - It represents the fraction of time a server is engaged in providing service. Utilization usually denoted by ρ

$$\rho = \frac{Time\ \ a\ server\ is\ occupied}{Time\ available}$$

  - ρ is dimensionless and should be less than one in order for a serve to cope with the service demand and for the system to be stable.

Flow Conservation Law

- Flow Conservation Law states that

  "For a stable queueing system, the rate of customers entering the system should equal to the rate of customers leaving the system if we observe it for a sufficiently long period of time "

$$\lambda_{in} = \lambda_{out}$$

- This notion of flow conservation is useful when we wish to calculate system throughput or study networks of queues

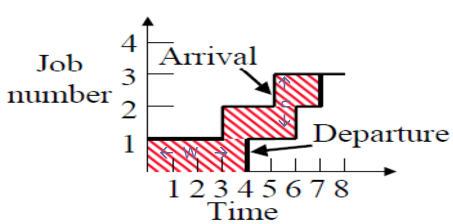
Little's Theorem

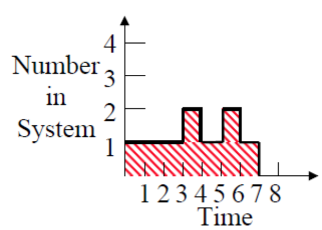
- Consider a general queueing system (G/G/..
- Little's theorem relates performance metrics in general
  - Let L denote the average number of customers in a steady-state queueing system.
  - Let λ represent the average arrival rate of the customers.
  - Further W is the average time a customer spends in the system.
- The theorem states :

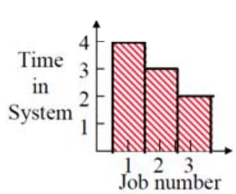
$$L = \lambda W$$

  - Proof in book

Little's Law







Little's Theorem (2)

- Important facts:
  - It does not assume any specific distribution for the arrival as well as the service process
  - It does not assume any queueing discipline.
  - It does not depend upon the number of parallel servers in the system.
- The theorem can be applied to all types of queueing systems including priority queueing and multi-server systems.
- The theorem holds for any queueing disciplines as long as the servers are kept busy when the system is not empty.
  - Working conserving discipline.
- More specific analysis of queues (i.e., how to find W or L) depends on arrival and service process and stochastic processes analysis

9

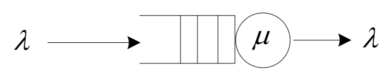
Single Queue Analysis

- Consider single queue case – G/G/C doesn't have a closed form solution - will consider approximations later
- First focus on basic models widely used in network performance analysis
  - Data networks and database systems
    - M/M/1
    - M/M/1/K
  - Telephony
    - M/M/C $\Rightarrow$ Erlang C
    - M/M/C/C $\Rightarrow$ Erlang B
  - All are Markovian queues, study using Birth Death process CTMC

TELCOM 2110 10

Single Queue Analysis (M/M/1)

- Most basic Markovian queue is the M/M/1/∞/FIFO/∞ queue



  - Customers arrive according to a Poisson process with exponentially distributed interarrival times (IAT)
    P{ IAT ≤ t} = 1 – $e^{-\lambda t}$, mean interarrival time = 1/ λ
  - Customers are served by a single server with exponential service time distribution P(service time < t) = 1 – $e^{-\mu t}$
    mean service time = 1/ μ
  - The arrival rate (λ) and service rate (μ) do not depend upon the number of customers in the system or time
  - Consider behavior of n(t) – number of customers in the system at time t
    ➔ forms a Markov Process

M/M/1 Analysis

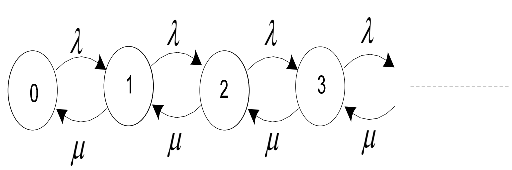
- Consider n(t) behavior over small time interval Δt

  P{ more than one arrival} ≈ 0,

  P { more than one service completion} ≈ 0

  P{ arrival and a service completion} ≈ 0

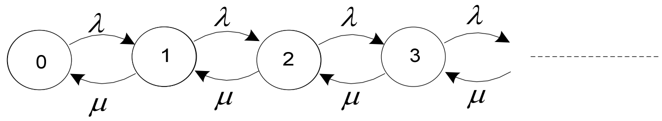
Get birth-death state transition diagram and generator matrix



$$Q=\begin{bmatrix} -\lambda & \lambda & 0 & \cdots & & & \\ \mu & -(\lambda+\mu) & \lambda & 0 & \cdots & & \\ 0 & \mu & -(\lambda+\mu) & \lambda & 0 & \cdots & \\ \vdots & & \ddots & \ddots & \ddots & & \end{bmatrix}$$

M/M/1 Queue

- From state transition diagram flow balance or $\pi Q = 0$ get the equations to solve for the steady state probabilities



flow out state *j* = flow in state *j*

$$\lambda\pi_0 = \mu\pi_1 \qquad j = 0$$

$$(\lambda + \mu)\pi_j = \lambda\pi_{j-1} + \mu\pi_{j+1} \qquad j > 0$$

Also use Normalization equation $\sum_{\forall i} \pi_i = 1$

M/M/1 Queue

- Solving the equations for π_i

$$\lambda\pi_0 = \mu\pi_1 \quad \Rightarrow \quad \pi_1 = \rho\pi_0$$

- *Consider the general equation for j = 1*

$$(\lambda + \mu)\pi_1 = \lambda\pi_0 + \mu\pi_2$$

- *Substituting for* $\pi_1 = \rho\pi_0$

$$(\lambda + \mu)\rho\pi_0 = \lambda\pi_0 + \mu\pi_2 \quad \Rightarrow \quad \pi_2 = \rho^2\pi_0$$

- *By induction* $\pi_n = \rho^n\pi_0$

$$\sum_{i=0}^{\infty} \pi_i = 1 \quad \Rightarrow \sum_{i=0}^{\infty} \rho^i \pi_0 = 1 \quad \Rightarrow \quad \pi_0 = 1 - \rho$$

$$\Rightarrow \quad \pi_n = \rho^n (1 - \rho)$$

M/M/1 Continued

$$\pi_n = \rho^n (1-\rho)$$ Geometric distribution

where $\left(\rho = \frac{\lambda}{\mu}\right) < 1$ Stability Condition

Mean Number in System L

$$L = \sum_i i \times \pi_i = \frac{\rho}{(1-\rho)}$$

Mean Delay W = L/λ

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

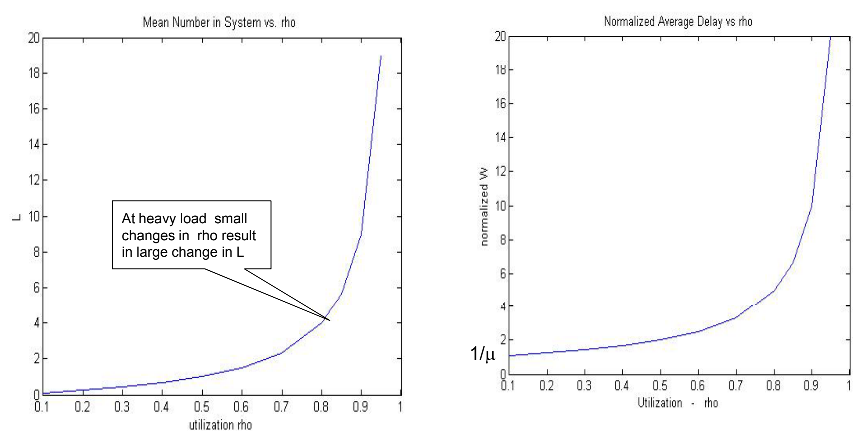
Variance of number in system σ_L

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

Variance of Delay σ_W

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

M/M/1 Queue - Mean Behavior



M/M/1

- M/M/1 is often used as a first cut model of computer network equipment (e.g. router based networks) with the following assumptions:
  - Packets arrive to a network link according to a Poisson process with rate λ packets / time unit.
  - Infinite capacity queues with FIFO service
  - Capacity of the link is fixed at C bps.
  - The length of the packets (thus, a service time) is exponentially distributed with average length $1/\mu$
    - Mean service rate = μC packets / time units
    - Link Utilization = $\rho = \lambda / (\mu C)$

M/M/1 Example

- Consider a concentrator that receives messages from a group of terminals and transmits them over a single transmission line.
- The packets arrive according to a Poisson process with one packet every 2.5 ms and the packet transmission times are exponentially distributed with a mean of 2 ms. That is the arrival rate = 1 packet/2.5 ms = 400 packets/sec
- Service rate = 1packet/2ms = 500 packets/sec
  - Find the average delay through the system
    - Utilization = $\rho = 400/500$ = .8
  - Delay W = 1/(500 – 400) = .01 secs = 10 msecs
  - Mean Queue Length L = $\rho/(1-\rho)$ = 4

M/M/1 Example: Economy of Scale (1)

- Consider a company that has K terminal rooms. Each terminal room is identical containing as set of terminals/workstations connected by a concentrator to a network.
  - Each set of terminals generates messages to be sent over the concentrator according to a Poisson process with rate λ
  - Each message requires an exponentially distributed amount of time to be sent by the concentrator with a rate of μ .
- The company is considering replacing the set of K rooms and K concentrators with one large room and a concentrator that is K times faster.

Economy of Scale (2)

- Comparing two **options** :
- *K independent rooms*
  - Each room can be modeled as multiple M/M/1 queues with arrival rate λ and service rate μ.
  - Average delay at any room (W) = $1/(\mu - \lambda) = (1/\mu) / (1-\rho)$
- *Single large room*
  - It can be modeled as single M/M/1 queue with arrival rate $K\lambda$ and service rate $K\mu$.
  - Average delay (W) = $1/ (K\mu - K\lambda) = (1/K\mu) / (1-\rho)$
- The combined system is K time faster !

Statistical Multiplexing

- There are m independent Poisson data streams, each supplying packet at rate λ/m, arriving at a common "concentrator" where they are mixed into a single data stream of combined rate λ.
- Packet lengths are independent and exponentially distributed with mean transmission time $1/\mu$.
- The concentrator can be viewed as **M/M/1** system which is *"statistically multiplexes"* the independent data streams into a single data stream.
  - The average number of packets at the concentrator

    $$L_{SM} = \lambda / (\mu - \lambda)$$

  - The average delay per packet

    $$W_{SM} = 1 / (\mu - \lambda)$$

Time/Frequency Division Multiplexing (TDM/FDM)

- In TDM and FDM, transmission capacity is divided equally over m data stream so that each data stream effectively sees a dedicated line with service rate μ / m.
- TDM and FDM can be modeled as ***m* (M/M/1)** systems operating in parallel.
  - Each M/M/1 queue observes packet arrival rate of λ/m and service rate of μ / m.
  - The average number of packets at the concentrator

    $$L_{TDM} = (\lambda/m) / (\mu/m - \lambda/m) = \lambda / (\mu - \lambda) = L_{SM}$$

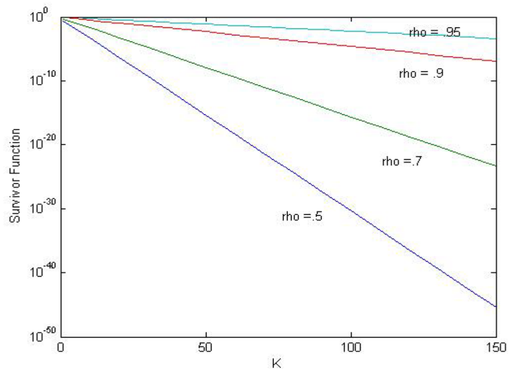
  - The average delay per packet

    $$W_{TDM} = 1 / (\mu/m - \lambda/m) = m / (\mu - \lambda) = m \cdot W_{SM}$$

M/M/1 Queue Other Metrics

- Survivor Function $SF_k = P\{n(t) > k\}$
- often used to dimension buffer space or estimate loss rates for finite queues

$$SF_k = P\{n(t) > k\} = \sum_{i=k}^{\infty} \pi_i = \sum_{i=k}^{\infty} \rho^i (1-\rho) = \rho^{k+1}$$



M/M/1 Queue Other Metrics

- Sojurn Time Distribution (i.e., distribution of time in system)

P{st < t} - find by conditioning on the number in system *n(t)*

$$P\{st \le t\} = \sum_{k=0}^{\infty} P\{st \le t \mid n(t) = k\} P\{n(t) = k\}$$

P{st ≤ t |n(t) =k} is sum of k+1 exponential service times ➔ Erlang K+1

$$P\{st \le t\} = \sum_{k=0}^{\infty} \left(\int_0^t \frac{(\mu x)^k}{k!} \mu e^{-\mu x} dx \right) (1-\rho)\rho^k$$

Interchanging the order of summation and integration one gets

$$P\{st \le t\} = 1 - e^{-(\mu - \lambda)t}$$ ← exponential distribution

M/M/1 Queue Other Metrics

- Let t_p be the pth percentile of Sojurn Time Distribution

  $t_p = P\{st < t\} = 1 - e^{-(\mu-\lambda)t}$

  Solving for t_p results in

  $$t_p = \frac{-1}{\mu - \lambda} \ln(1 - p) = -W \ln(1 - p)$$

- A similar analysis for the distribution of the waiting time in the queue yields

  $$P\{waiting\ \ time\ \ in\ \ queue \le t\} = 1 - \rho e^{-(\mu-\lambda)t}$$

- The pth percentile tp of the waiting time in queue is

  $$t_p = \frac{-1}{\mu - \lambda} \ln \frac{(1 - p)}{\rho} = -W \ln \frac{(1 - p)}{\rho}$$

Example-III

- The average response time on a database system is 3 seconds. During a one minute observation interval, the idle time on the system was measured to be 10 seconds. Using M/M/1 model for the system determine the following:
- System utilization

  The fraction of time server is busy = ρ = 50/60 = 0.8333
- Average service time per query

  Mean delay (W) = 3 = $1/(\mu - \lambda) = (1/\mu) / (1-\rho)$

  Substitute ρ = 0.8333, therefore $1/\mu$ = 0.5
- Number of queries completed during the observation interval

  60 λ = 60 μ ρ = 100
- Average number of jobs in the system

  L = ρ / (1- ρ) = 5

Example-III

- Probability of number of jobs in the system being greater than 10

  $SF_{10} = \rho^{11} = 0.135$

- The 90th percentile of the response time (time in the system)

  $t_p = -W \ln(1-p)$

  p = 0.9

  $t_{0.9} = 6.9$

- The 90th percentile of the waiting time (time in the system)

  $t_p = -W \ln\{(1-p) / \rho\}$

  p = 0.9

  $t_{0.9} = 6.36$

Markovian Queues Analysis

- Develop state transition diagram
  - System state is indicated by the number of customers in the system at time $t \Rightarrow \{n(t), t \geq 0\}$
- Flow Balance Equations
  - Derive steady state probability (π_i)

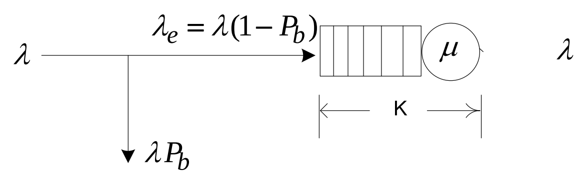
$$\sum flow\,in = \sum flow\,out$$

$$\sum_{\forall i} \pi_i = 1$$

- Apply Little's theorem to obtain mean performance metrics.

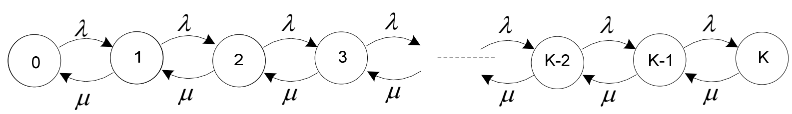
M/M/1/K

- The system has a finite capacity of size K.



- The state space will be truncated at state K.

M/M/1/K (2)



Flow Balance Equations

$$\lambda \pi_0 = \mu \pi_1 \qquad j = 0$$

$$(\lambda + \mu)\pi_j = \lambda \pi_{j-1} + \mu \pi_{j+1} \qquad 1 \le j < K$$

$$\mu \pi_j = \lambda \pi_{j-1} \qquad j = K$$

Also use Normalization equation $\sum_{\forall i} \pi_i = 1$

M/M/1/K

Solving equations yields $\pi_n = \rho^n \pi_0 \quad n \le K$

where

$$\rho = \frac{\lambda}{\mu}$$ ← Normalized offered load

$$0 < \rho < \infty$$

Solving normalization equation one gets

$$\pi_n = \frac{(1-\rho)\rho^n}{1-\rho^{K+1}} \qquad \rho \neq 1$$

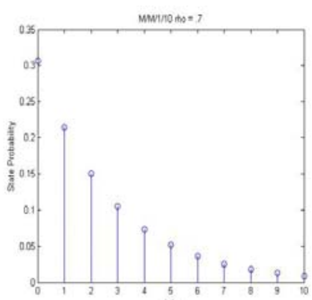
From L'Hopital' rule get

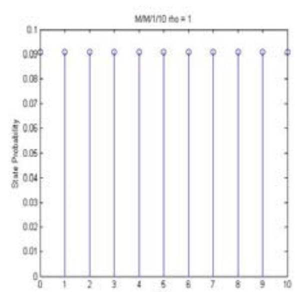
$$\pi_n = \frac{1}{K+1} \qquad \rho = 1$$

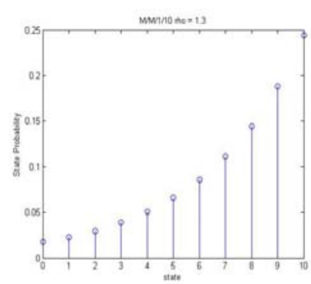
M/M/1/K

- Behavior of state probabilities π_i with ρ

$\rho < 1$ ➔ π_0 largest $\rho = 1$ ➔ π_i discrete uniform $\rho > 1$ ➔ π_K largest







M/M/1/K

Probability of blocking (P_b) = Loss Rate

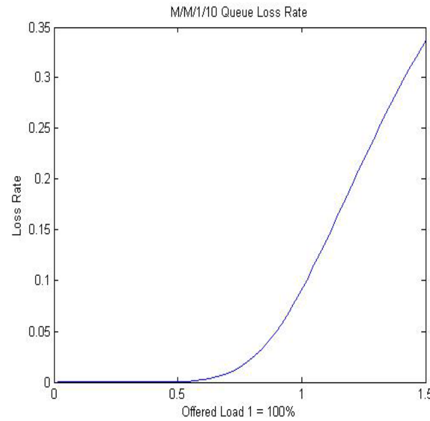
$$P_b = \pi_k = \frac{(1-\rho)\rho^K}{1-\rho^{K+1}} \qquad \rho \neq 1$$

$$P_b = \pi_k = \frac{1}{K+1} \qquad \rho = 1$$

Portion of traffic dropped/rejected = $\lambda \cdot P_b$

Effective throughput of the system

$\lambda_e = \lambda(1-P_b)$ ⇐ effective arrival rate



Example M/M/1/10
Notice how it is nonlinear

M/M/1/K

Effective server utilization : (actual utilization of the system)

$$\rho_e = \frac{\lambda(1-P_b)}{\mu} = \frac{\lambda_e}{\mu}$$

Average number in the system

$$\rho \neq 1 \qquad L = \sum_{i=0}^{K} i\pi_i = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}$$

$$\rho = 1 \qquad L = \sum_{i=0}^{K} i\pi_i = \sum_{i=0}^{k} i\left(\frac{1}{K+1}\right)$$

$$= \left(\frac{1}{K+1}\right)\sum_{i=0}^{k} i$$

$$= \frac{K}{2}$$

M/M/1/K

Other performance measures

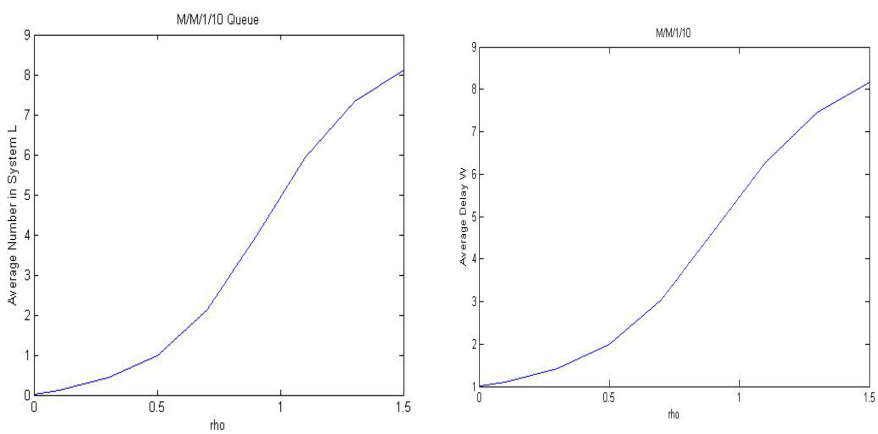
Mean Delay $W = \frac{L}{\lambda_e}$

Mean Queueing Delay $W_q = W - \frac{1}{\mu}$

Mean Number in Queue $L_q = L - \rho_e$

M/M/1/K

- Compare with M/M/1/∞ results on slide 10



M/M/1/K Example

- Consider the queue at an output port of router. The transmission link is a T1 line (1.544Mbps), packets arrive according to a Poisson process with mean rate λ = 659.67 packets/sec, the packet lengths are exponentially distributed with a mean length of 2048 bits/packet.
- If the system size is 16 packets what is the packet loss rate?

  model as M/M/1/16 queue with

  $\lambda = 659.67$, μ= 1.544 Mbps/2048 bits per packet = 753.9 packets/sec

  $\rho = \lambda/\mu = 0.875$

  Thus the packet loss rate = blocking probability

$$P_b = \frac{(1-\rho)\rho^K}{1-\rho^{K+1}} = \frac{(1-.875).875^{16}}{1-.875^{17}} = 0.0165$$

Example (M/M/1 and M/M/1/K)

- A data concentrator has 40 terminals connected to it. During the busiest time of day each terminal is occupied and produces packets which are exponentially distributed with a mean of 1000 bits. The link connecting the concentrator to the campus network carries traffic at 1.552 Mbps. The arrival process of packets to the concentrator forms a Poisson process with ten of the terminals producing on average 1 packet per 10 msec, twenty of the terminals producing on average 1 packet per 50 msec, and ten of the terminals producing on average 1 packet per .5 second.

(a) Determine the utilization of the concentrator.

(b) Assuming the buffer at the concentrator is infinite, determine the average delay in the queue.

(c) If the concentrator has a system capacity of 20 packets, determine the packet loss rate.

Example (M/M/1 and M/M/1/K)

(a) Determine the utilization of the concentrator.

mean service rate μ = 1.552 *10^6 bps/1000 bits per packet = 1552 packets/sec

mean arrival rate λ = 10*(1packet/10msec) + 20*(1packet/50msec) + 10*(1packet/.5sec) = 1420 packets/sec

Thus ρ = 1420/1552 = .9149

(b) Assuming the buffer at the concentrator is infinite, determine the average delay in the queue.

$W_q = W - 1/\mu$ = 6.93 msec or

$W_q = \rho / (\mu - \lambda)$ = 0.9149 / (1552-1420) = 6.93 msec.

(c) If concentrator has a system capacity of 20 packets, find the packet loss rate.

The system is now modeled as a M/M/1/K queue - from the M/M/1/K results

The blocking probability P_B = packet loss rate (use K = 20, ρ = .9149)

$$P_B = \frac{(1-\rho)\rho^K}{1-\rho^{K+1}} = 0.017 \Rightarrow 1.7\%$$

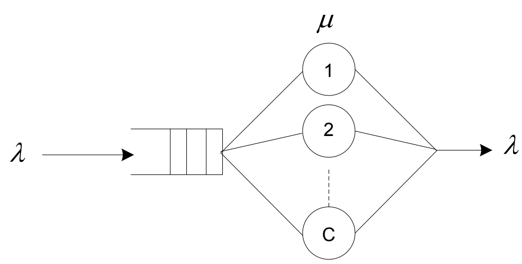
Therefore, packet loss rate = λP_B = 1420 * 0.017 = 24.14 packets/sec.

Summary

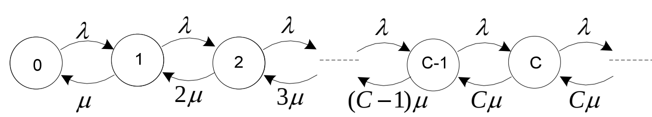
- Basic single server Markovian queueing systems
- Analysis using Birth Death process state diagram
  - M/M/1
  - M/M/1/K
- Applications in examples and homework

M/M/C

- C identical servers processes customers in parallel.
- Infinite system capacity.



M/M/C (2)



$$\lambda \pi_0 = \mu \pi_1 \qquad j = 0$$

$$(\lambda + j\mu)\pi_j = \lambda \pi_{j-1} + (j+1)\mu \pi_{j+1} \qquad 1 \le j < C$$

$$(\lambda + C\mu)\pi_j = \lambda \pi_{j-1} + C\mu \pi_{j+1} \qquad j \ge C$$

M/M/C (3)

$$\pi_j = \frac{\lambda}{j\mu}\pi_{j-1} \qquad j < C$$

$$\pi_j = \frac{\lambda}{C\mu}\pi_{j-1} \qquad j \geq C$$

M/M/C (4)

The server utilization (ρ)

$$\rho = \frac{\lambda}{C\mu}$$

The traffic intensity (*a*) $\Leftarrow$ offered load (Erlangs)

$$a = \frac{\lambda}{\mu}$$

The stability requirement

$$\rho = \frac{a}{C} < 1 \qquad \Rightarrow \quad a < C$$

With traffic intensity *a* Erlangs, *C* is the minimum number of servers requirement.

M/M/C (5)

$$\pi_0 = \frac{1}{\sum_{n=0}^{C-1}\frac{a^n}{n!} + \frac{a^c}{(c-1)!(c-a)}}$$

$$\pi_i = \frac{a^i}{i!}\pi_0 \qquad ;1 \le i < C$$

$$\pi_i = \frac{a^i}{c!c^{i-c}}\pi_0 \qquad ;i \ge C$$

M/M/C (6)

Probability of a customer being delayed *C(c,a)*

$$C(c,a) = \sum_{j=c}^{\infty}\pi_j = \frac{\frac{a^c}{(c-1)!(c-a)}}{\sum_{n=0}^{c-1}\frac{a^n}{n!} + \frac{a^c}{(c-1)!(c-a)}}$$

C(c,a) ⇐ Erlang's *C* formula
Erlang's delay formula
Erlangs second formula

In the telephone system,
C(c,a) represents a blocked call delayed (BCD).

Erlang C model

Erlang C Traffic Table

Maximum Offered Load Versus B and N
B is in %

N/B	0.01	0.05	0.1	0.5	1.0	2	5	10	15	20	30
1	.0001	.0005	.0010	.0050	.0100	.0200	.0500	.1000	.1500	.2000	.3000
2	.0142	.0319	.0452	.1025	.1465	.2103	.3422	.5000	.6278	.7403	.9390
3	.0860	.1490	.1894	.3339	.4291	.5545	.7876	1.040	1.231	1.393	1.667
4	.2310	.3533	.4257	.6641	.8100	.9939	1.319	1.653	1.899	2.102	2.440
5	.4428	.6289	.7342	1.065	1.259	1.497	1.905	2.313	2.607	2.847	3.241
6	.7110	.9616	1.099	1.519	1.758	2.047	2.532	3.007	3.344	3.617	4.062
7	1.026	1.341	1.510	2.014	2.297	2.633	3.188	3.725	4.103	4.406	4.897
8	1.382	1.758	1.958	2.543	2.866	3.246	3.869	4.463	4.878	5.210	5.744
9	1.771	2.208	2.436	3.100	3.460	3.883	4.569	5.218	5.668	6.027	6.600
10	2.189	2.685	2.942	3.679	4.077	4.540	5.285	5.986	6.469	6.853	7.465

M/M/C (7)

Other performance measures

$$L_q = \left(\frac{a}{c-a}\right) \cdot C(c,a)$$

$$L = L_q + a$$

$$W_q = \frac{L_q}{\lambda} = \frac{\frac{1}{\mu} C(c,a)}{c-a}$$

$$W = W_q + \frac{1}{\mu}$$

M/M/C (8)

Distribution of the waiting time in the queue

$$P\{w_q \le t\} = 1 - C(c,a) \cdot e^{-c\mu(1-\rho)t}$$

The *p*th percentile of the time spent waiting in the queue t_p

$$t_p = \frac{-\ln\left(\frac{1-p}{C(c,a)}\right)}{c\mu(1-\rho)}$$

Note: $p > 1 - C(c,a)$

Traffic Engineering Example

- A service provider receives unsuccessful call attempts to wireless subscribers at a rate of 5 call per minute in a given geographic service area. The unsuccessful calls are processed by voice mail and have an average mean holding time of 1 minute. When all voice mail servers are busy – customers are placed on hold until a server becomes free.
- Determine the minimum number of servers to keep the percentage of customers placed on hold < or equal to 1%

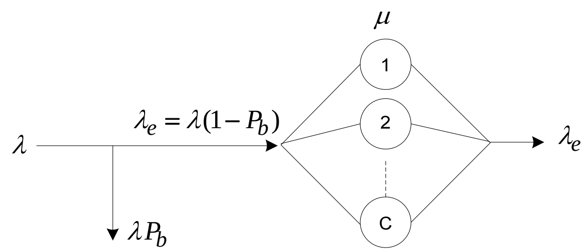
  The offered load is a = 5 call per minute x 1 minute/call = 5

  From the Erlang C tables 13 servers are needed.
- Determine the .995% of the delay in access the voice servers
- With p = .995, C(c,a) = .01, c = 13, and μ = 1

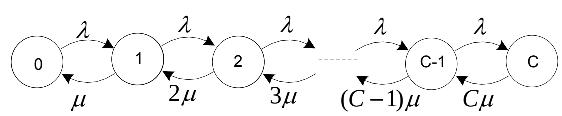
$$t_p = \frac{-\ln\left(\frac{1-p}{C(c,a)}\right)}{c\mu(1-\rho)}$$ yields t_p = .0866 minute = 5.2 secs

M/M/C/C

- C identical servers processes customers in parallel.
- The system has a finite capacity of size C.



M/M/C/C (2)



$$\lambda \pi_0 = \mu \pi_1 \qquad j = 0$$

$$(\lambda + j\mu)\pi_j = \lambda \pi_{j-1} + (j+1)\mu \pi_{j+1} \qquad 1 \le j < C$$

$$(C\mu)\pi_c = \lambda \pi_{c-1} \qquad j = C$$

M/M/C/C (3)

$$\pi_0 = \frac{1}{\sum_{n=0}^{c} \frac{a^n}{n!}}$$

$$\pi_i = \frac{a^i}{i!}\pi_0 = \frac{\frac{a^i}{i!}}{\sum_{n=0}^{c} \frac{a^n}{n!}} \qquad \forall i = 1,2,\ldots c$$

M/M/C/C (4)

Probability of a customer being blocked *B(c,a)*

$$B(c,a) = \pi_c = \frac{\frac{a^c}{c!}}{\sum_{n=0}^{c} \frac{a^n}{n!}} \qquad \Leftarrow \text{Valid for M/G/c/c queue}$$

B(c,a) ⇐ Erlang's *B* formula
Erlang's blocking formula
Erlangs first formula

In the telephone system,
B(c,a) represents a blocked call cleared (BCC).

M/M/C/C (5)

Erlang B formula can be computed from the recursive formula

$$B(c,a) = \frac{a \cdot B(c-1,a)}{c + a \cdot B(c-1,a)}$$

Erlang B formula can be used to computer Erlang C formula

$$C(c,a) = \frac{c \cdot B(c,a)}{c - a \cdot (1 - B(c,a))}$$

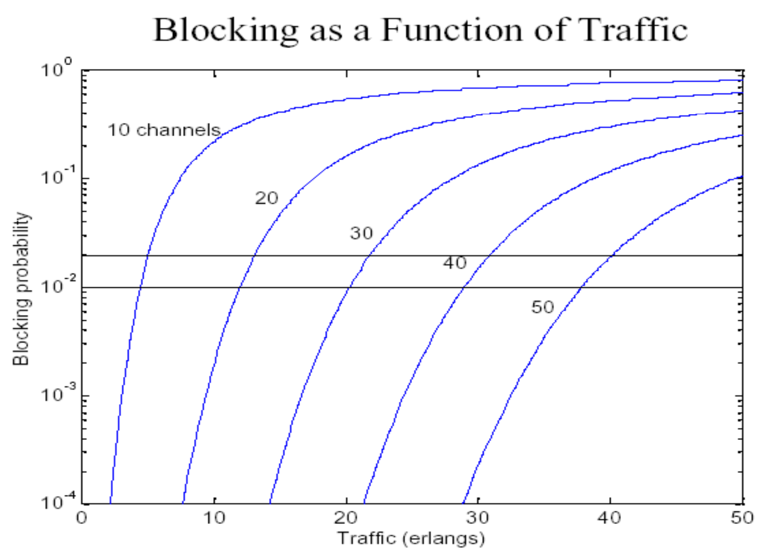
Note : C(c,a) > B(c,a)

Traffic Engineering Erlang B table

Appendix 1.1
Blocked-Calls-Cleared
(Erlang *B*)

	A, erlangs												
	B												
N	1.0%	1.2%	1.5%	2%	3%	5%	7%	10%	15%	20%	30%	40%	50%
1	.0101	.0121	.0152	.0204	.0309	.0526	.0753	.111	.176	.250	.429	.667	1.00
2	.153	.168	.190	.223	.282	.381	.470	.595	.796	1.00	1.45	2.00	2.73
3	.455	.489	.535	.602	.715	.899	1.06	1.27	1.60	1.93	2.63	3.48	4.59
4	.869	.922	.992	1.09	1.26	1.52	1.75	2.05	2.50	2.95	[illegible]	5.02	6.50
5	1.36	1.43	1.52	1.66	1.88	2.22	2.50	2.88	3.45	4.01	5.19	6.60	8.44
6	1.91	2.00	2.11	2.28	2.54	2.96	3.30	3.76	4.44	5.11	6.51	8.19	10.4
7	2.50	2.60	2.74	2.94	3.25	3.74	4.14	4.67	5.46	6.23	7.86	9.80	12.4
8	3.13	3.25	3.40	3.63	3.99	4.54	5.00	5.60	6.50	7.37	9.21	11.4	14.3
9	3.78	3.92	4.09	4.34	4.75	5.37	5.88	6.55	7.55	8.52	10.6	13.0	16.3
10	4.46	4.61	4.81	5.08	5.53	6.22	6.78	7.51	8.62	9.68	12.0	14.7	18.3
11	5.16	5.32	5.54	5.84	6.33	7.08	7.69	8.49	9.69	10.9	13.3	16.3	20.3
12	5.88	6.05	6.29	6.61	7.14	7.95	8.61	9.47	10.8	12.0	14.7	18.0	22.2
13	6.61	6.80	7.05	7.40	7.97	8.83	9.54	10.5	11.9	13.2	16.1	19.6	24.2
14	7.35	7.56	7.82	8.20	8.80	9.73	10.5	11.5	13.0	14.4	17.5	21.2	26.2
15	8.11	8.33	8.61	9.01	9.65	10.6	11.4	12.5	14.1	15.6	18.9	22.9	28.2
16	8.88	9.11	9.41	9.83	10.5	11.5	12.4	13.5	15.2	16.8	20.3	24.5	30.2
17	9.65	9.89	10.2	10.7	11.4	12.5	13.4	14.5	16.3	18.0	21.7	26.2	32.2
18	10.4	10.7	11.0	11.5	12.2	13.4	14.3	15.5	17.4	19.2	23.1	27.8	34.2
19	11.2	11.5	11.8	12.3	13.1	14.3	15.3	16.6	18.5	20.4	24.5	29.5	36.2
20	12.0	12.3	12.7	13.2	14.0	15.2	16.3	17.6	19.6	21.6	25.9	31.2	38.2

Erlang B Charts



M/M/C/C (5)

The carried load

$$\lambda_e = \lambda \cdot (1 - B(c,a))$$ $\Leftarrow$ Effective throughput of the system

Mean server utilization $$\rho_e = \frac{a}{c} \cdot (1 - B(c,a))$$

Mean number in the system $$L = \frac{a}{\mu} \cdot (1 - B(c,a))$$

Average delay in the system $$W = \frac{1}{\mu}$$

Traffic Engineering Example

- Consider a single analog cell tower with 56 traffic channels, when all channels are busy calls are blocked. Calls arrive according to a Poisson process at a rate of 1 call per active user an hour. During the busy hour 3/4 the users are active. The call holding time is exponentially distributed with a mean of 120 seconds.
- (a) What is the maximum load the cell can support while providing 2% call blocking?

  From the Erlang B table with c= 56 channels and 2% call blocking the maximum load = 45.9 Erlangs
- (b) What is the maximum number of users supported by the cell during the busy hour?

  Load per active user = 1 call x 120 sec/call x 1/3600 sec = 33.3 mErlangs

  Number active users = 45.9/(0.0333) = 1377

  Total number users = 4/3 number active users = 1836
- Determine the utilization of the cell tower ρ

  $\rho = \alpha/c$ = 45.9/56 = 81.96%

Summary

- Basic Single server Markovian queueing systems
- M/M/1
- M/M/1/K
- M/M/C
- M/M/C/C
- Applications in homework and examples

M/G/1

- The customer arrival process is Poisson with mean rate λ.
- The customer service times are i.i.d with non-negative distribution

$$F_S(t) = P\{\text{ service time} \le t\ \}$$

  - The mean service time = $1/\mu$
  - The square coefficient of variation of service time distribution = C_s^2

M/G/1 (2)

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

$$L_q = L - \rho$$

$$W = \frac{L}{\lambda}$$

$$W_q = \frac{L_q}{\lambda}$$

Note that the mean behavior of M/G/1 depends only on the first two moment Of the service time distribution, not the entire distribution.

M/G/1 (3)

- Variety of service time distribution
  - Deterministic (M/D/1) $\Leftarrow C_s^2 = 0$
  - Exponential distribution (M/M/1) $\Leftarrow C_s^2 = 1$
  - Erlang-K distribution ($M/E_K/1$) $\Leftarrow C_s^2 = 1/K$
  - Hyperexponential distribution ($M/H_K/1$)

$$C_s^2 = -1 + 2 * \sum_{i=1}^{K} \frac{p_i}{\mu_i^2} \bigg/ \left(\sum_{i=1}^{K} \frac{p_i}{\mu_i} \right)^2$$

  - Gamma distribution (M/Gamma/1) $\Leftarrow C_s^2 = 1/\alpha$

Example V

- Fixed length of packets arrive to a modem according to a Poisson process with mean rate of 7 packets/sec. The time required to transmit the packet across the line and to receive an acknowledgement of correct reception is a fixed 0.1 sec/packet.
- If a packet is acknowledged then the modem proceeds with the next packet.
- If a negative acknowledgement of a packet occurs due to transmission errors then the modem retransmits the packet.
- Assuming that transmission errors are independent and occur with probability p=0.015.
- Determine the average delay of a packet through the modem line.

G/G/1

- The customer interarrival times are i.i.d with non-negative distribution

  $F_a(t)$ = P{ *service time* $\leq t$ }

- The customer service times are i.i.d with non-negative distribution

  $F_s(t)$ = P{ *service time* $\leq t$ }

- Exact closed form results for G/G/1 queue are unavailable.
- A widely used approximation : KLB approximation.

G/G/1 (2)

- KLB approximation based on two moments of the arrival and service time distributions.

- This approximation is often used to determine the effects of increased utilizations on systems where measurement data is available to determine C_s^2 and C_s^2

$$L \approx \rho + \frac{\rho^2 (C_a^2 + C_s^2) J}{2(1-\rho)}$$

where

$J = scaling\ \ factor$

$$J = \begin{cases} e^{\frac{-2(1-\rho)(1-C_a^2)^2}{3\rho(C_a^2 + C_s^2)}} & ; C_a^2 \leq 1 \\ e^{\frac{-(1-\rho)(C_a^2 - 1)}{(C_a^2 + 4C_s^2)}} & ; C_a^2 > 1 \end{cases}$$