16.36 Communication Systems Engineering
Spring 2009

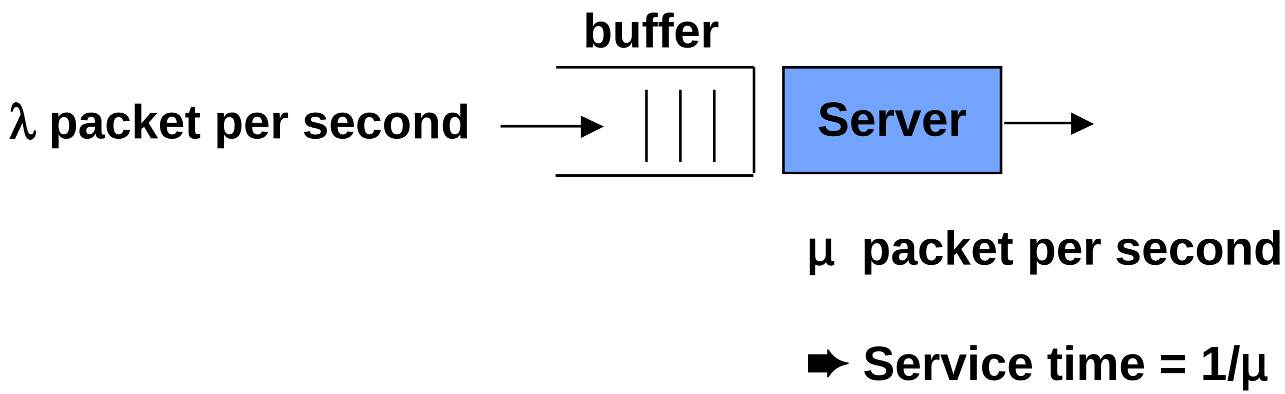
16.36: Communication Systems Engineering

Lecture 20: Delay Models for Data Networks

Part 2: Single Server Queues

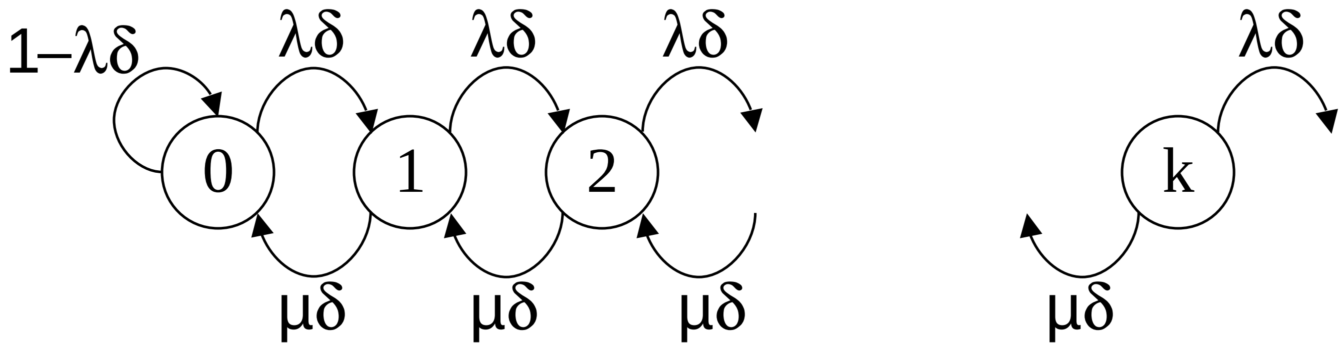
Eytan Modiano

Single server queues



- **M/M/1**
  - Poisson arrivals, exponential service times
- **M/G/1**
  - Poisson arrivals, general service times
- **M/D/1**
  - Poisson arrivals, deterministic service times (fixed)

Markov Chain for M/M/1 system



- **State k $\Rightarrow$ k customers in the system**
- **P(i,j) = probability of transition from state I to state j**
  - **As $\delta \Rightarrow 0$, we get:**

    $P(0,0) = 1 - \lambda\delta$, **$P(j,j+1) = \lambda\delta$**

    $P(j,j) = 1 - \lambda\delta - \mu\delta$ **$P(j,j-1) = \mu\delta$**

    $P(i,j) = 0$ for all other values of I,j.
- **Birth-death chain: Transitions exist only between adjacent states**
  - **$\lambda\delta, \mu\delta$ are flow rates between states**

Equilibrium analysis

- **We want to obtain P(n) = the probability of being in state n**
- **At equilibrium $\lambda P(n) = \mu P(n+1)$ for all n**
  - **$P(n+1) = (\lambda/\mu)P(n) = \rho P(n)$, $\rho = \lambda/\mu$**
- **It follows: $P(n) = \rho^n P(0)$**
- **Now by axiom of probability:**

$$\sum_{i=0}^{\infty} P(n) = 1$$

$$\Rightarrow \sum_{i=0}^{\infty} \rho^n P(0) = \frac{P(0)}{1-\rho} = 1$$

$$\Rightarrow P(0) = 1 - \rho$$

$$P(n) = \rho^n (1-\rho)$$

Average queue size

$$N = \sum_{n=0}^{\infty} nP(n) = \sum_{n=0}^{\infty} n\rho^n(1-\rho) = \frac{\rho}{1-\rho}$$

$$N = \frac{\rho}{1-\rho} = \frac{\lambda/\mu}{1-\lambda/\mu} = \frac{\lambda}{\mu-\lambda}$$

- **N = Average number of customers in the system**
- **The average amount of time that a customer spends in the system can be obtained from Little's formula (N=λT ⇒ T = N/λ)**

$$T = \frac{1}{\mu-\lambda}$$

- **T includes the queueing delay plus the service time (Service time = D_{TP} = 1/μ)**
  - **W = amount of time spent in queue = T - 1/μ ⇒**

$$W = \frac{1}{\mu-\lambda} - \frac{1}{\mu}$$

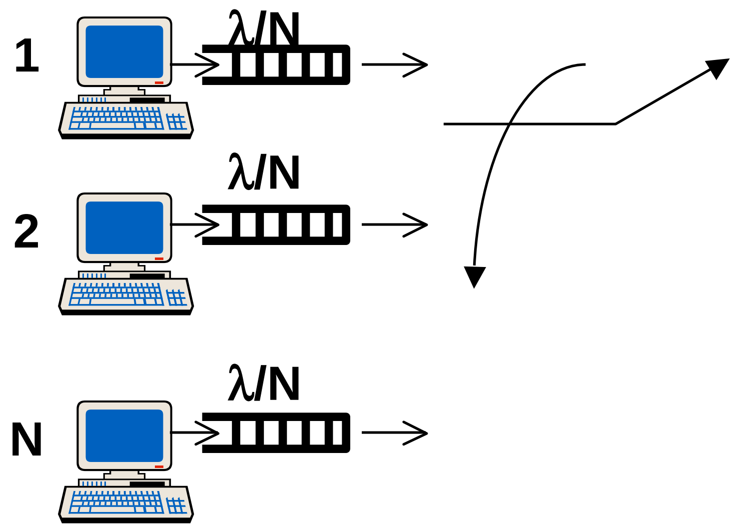
- **Finally, the average number of customers in the buffer can be obtained from little's formula**

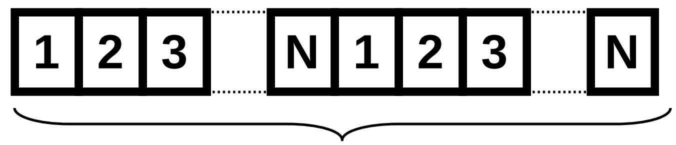
$$N_Q = \lambda W = \frac{\lambda}{\mu-\lambda} - \frac{\lambda}{\mu} = N - \rho$$

Example (fast food restaurant)

- **Customers arrive at a fast food restaurant at a rate of 100 per hour and take 30 seconds to be served.**
- **How much time do they spend in the restaurant?**
  - **Service rate = μ = 60/0.5=120 customers per hour**
  - **T = $1/\mu - \lambda$ = 1/(120-100) = 1/20 hrs = 3 minutes**
- **How much time waiting in line?**
  - **W = T - $1/\mu$ = 2.5 minutes**
- **How many customers in the restaurant?**
  - **N = λT = 5**
- **What is the server utilization?**
  - $\rho = \lambda/\mu$ = **5/6**

Packet switching vs. Circuit switching



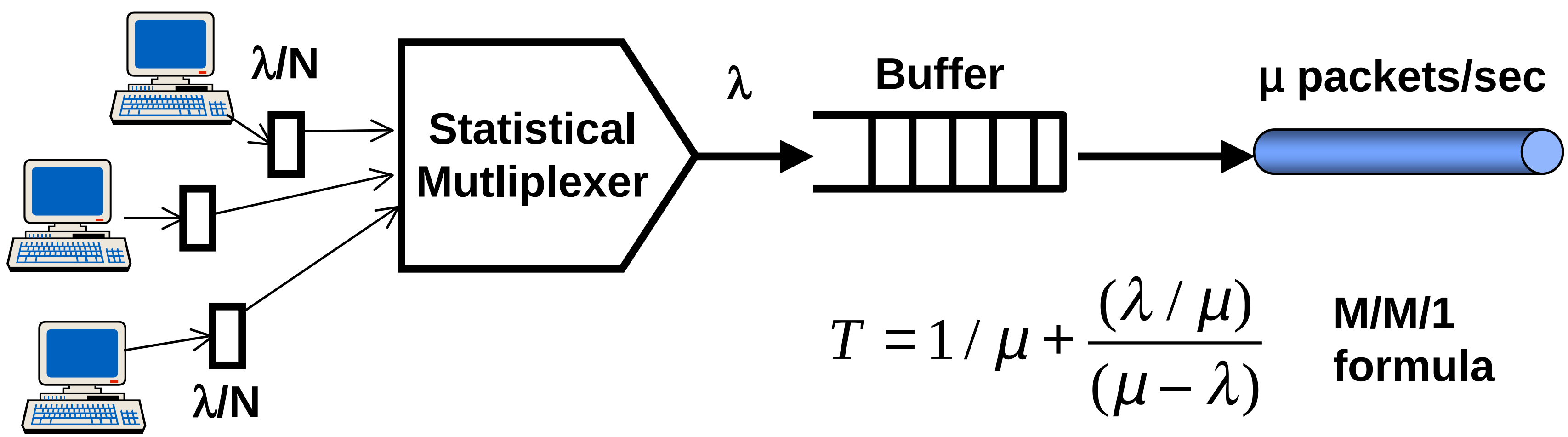


TDM, Time Division Multiplexing
Each user can send μ/N packets/sec and has packet arriving at rate λ/N packets/sec

$$T = N/\mu + \frac{N(\lambda/\mu)}{(\mu-\lambda)}$$

M/M/1 formula

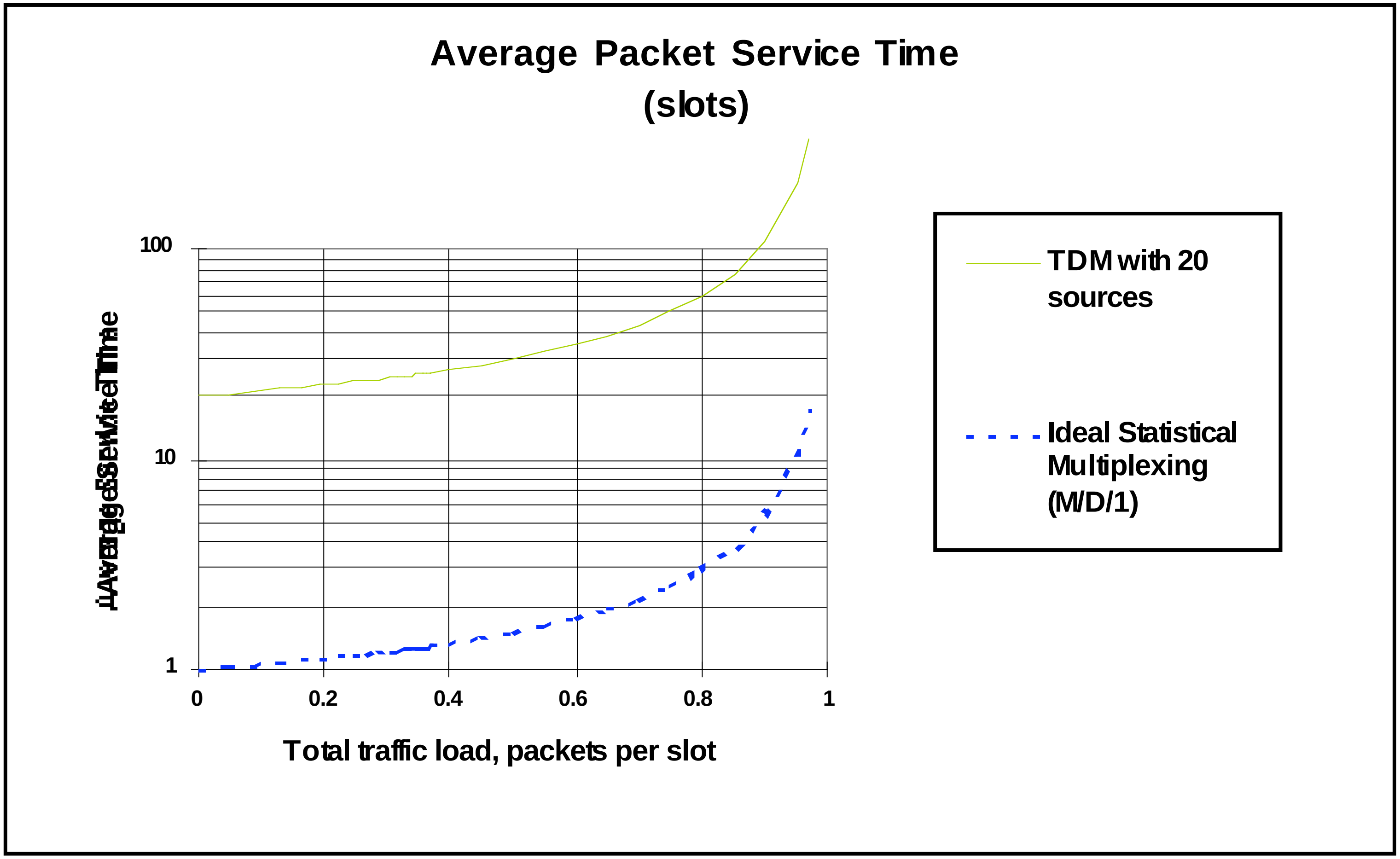
Packets generated at random times



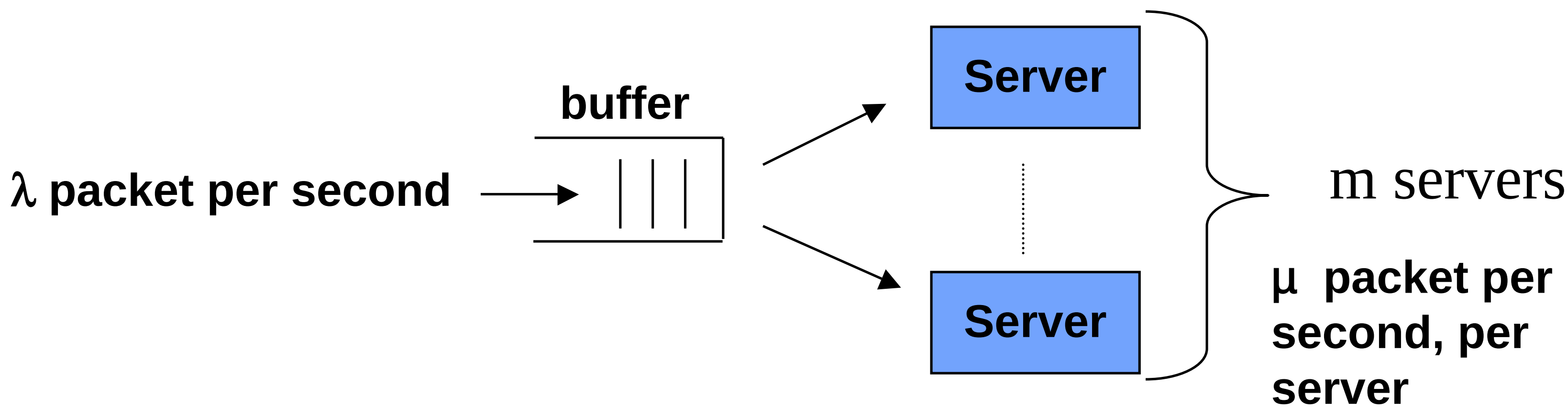
$$T = 1/\mu + \frac{(\lambda/\mu)}{(\mu-\lambda)}$$

M/M/1 formula

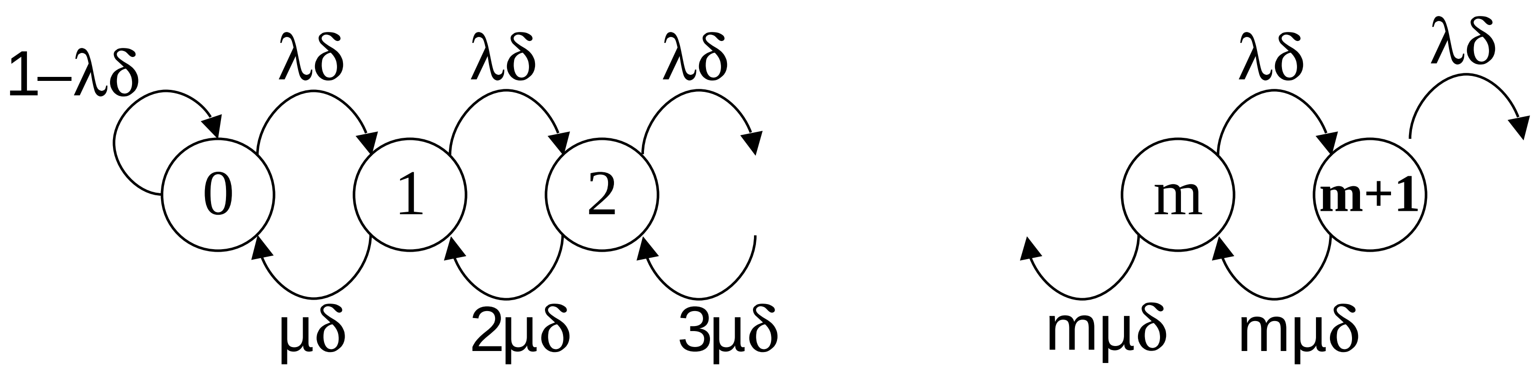
Circuit (TDM/FDM) vs. Packet Switching



Multi-server systems: M/M/m



- Departure rate is proportional to the number of servers in use
- Similar Markov chain:



M/M/m queue

- **Balance equations:**

$$\lambda P(n-1) = n\mu P(n) \quad n \le m$$

$$\lambda P(n-1) = m\mu P(n) \quad n > m$$

$$P(n) = \begin{cases} P(0)(m\rho)^n / n! & n \le m \\ P(0)(m^m \rho^n) \ / m! & n > m \end{cases}, \qquad \rho = \frac{\lambda}{m\mu} \le 1$$

- **Again, solve for P(0):**

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

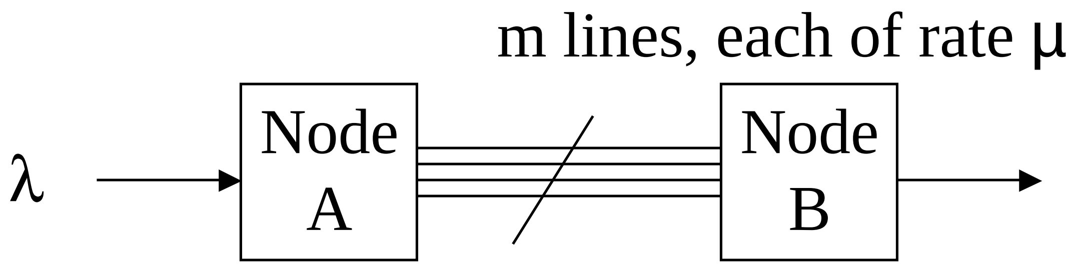
$$P_Q = \sum_{n=m}^{n=\infty} P(n) = \frac{P(0)(m\rho)^m}{m!(1-\rho)}$$

$$N_Q = \sum_{n=0}^{n=\infty} nP(n+m) = \sum_{n=0}^{n=\infty} nP(0)\left(\frac{m^m \rho^{m+n}}{m!}\right) = P_Q\left(\frac{\rho}{1-\rho}\right)$$

$$W = \frac{N_Q}{\lambda}, \ T = W + 1/\mu, \ N = \lambda T = \lambda/\mu + N_Q$$

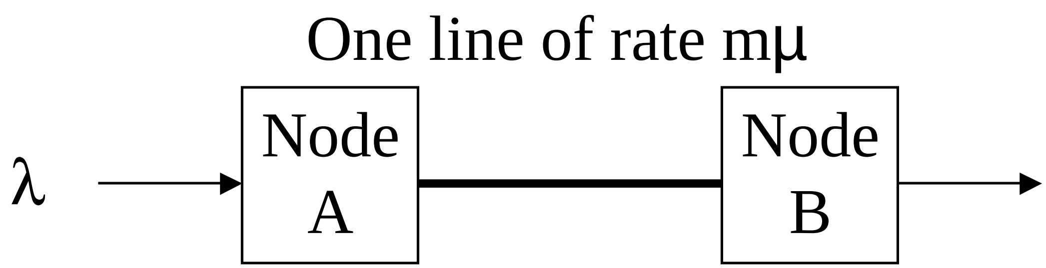
Applications of M/M/m

- **Bank with m tellers**
- **Network with parallel transmission lines**



Use M/M/m formula

VS



Use M/M/1 formula

- **When the system is lightly loaded, PQ~0, and Single server is m times faster**
- **When system is heavily loaded, queueing delay dominates and systems are roughly the same**

Blocking Systems
(circuit switched networks)

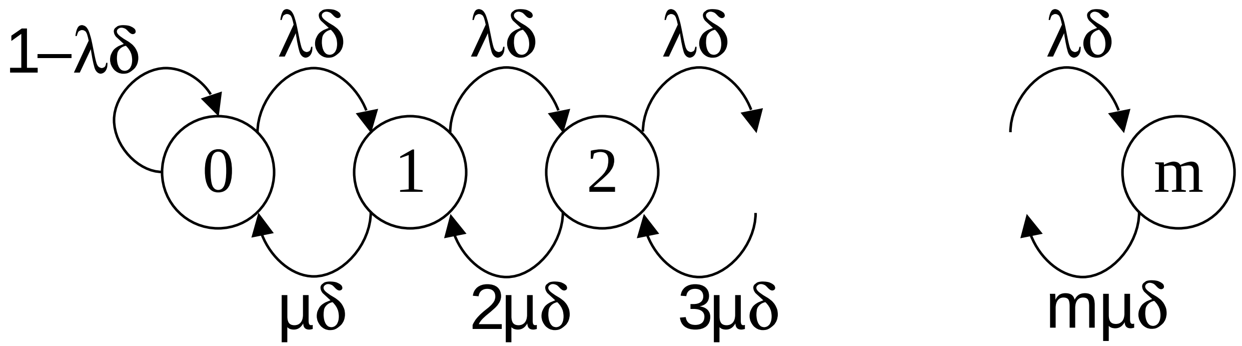
- **A circuit switched network can be viewed as a Multi-server queueing system**
  - **Calls are blocked when no servers available - "busy signal"**
  - **For circuit switched network we are interested in the call blocking probability**

- **M/M/m/m system**
  - **m servers $\Rightarrow$ m circuits**
  - **Last m indicated that the system can hold no more than m users**

- **Erlang B formula**
  - **Gives the probability that a caller finds all circuits busy**
  - **Holds for general call arrival distribution (although we prove Markov case only)**

$$P_B = \frac{(\lambda/\mu)^m / m!}{\sum_{n=0}^{m} (\lambda/\mu)^n / n!}$$

M/M/m/m system: Erlang B formula



$$\lambda P(n-1) = n\mu P(n),\ 1 \le n \le m,\ \Rightarrow P(n) = \frac{P(0)(\lambda/\mu)^n}{n!}$$

$$P(0) = \left[\sum_{n=0}^{m} (\lambda/\mu)^n / n!\right]^{-1}$$

$$P_B = P(Blocking) = P(m) = \frac{(\lambda/\mu)^m / m!}{\sum_{n=0}^{m} (\lambda/\mu)^n / n!}$$

Erlang B formula

- **System load usually expressed in Erlangs**
  - **A= λ/μ = (arrival rate)*(ave call duration) = average load**
  - **Formula insensitive to λ and μ but only to their ratio**

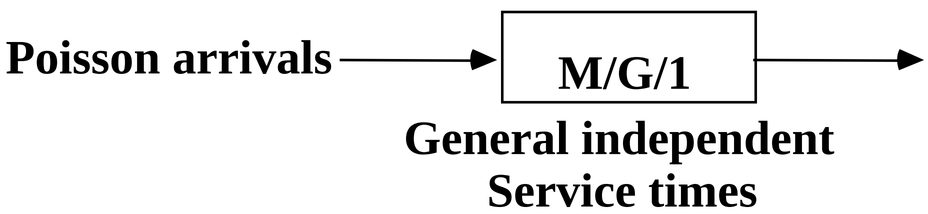
$$P_B = \frac{(A)^m / m!}{\sum_{n=0}^{m} (A)^n / n!}$$

- **Used for sizing transmission line**
  - **How many circuits does the satellite need to support?**
  - **The number of circuits is a function of the blocking probability that we can tolerate**
    - **Systems are designed for a given load predictions and blocking probabilities (typically small)**

- **Example**
  - **Arrival rate = 4 calls per minute, average 3 minutes per call ⇒ A = 12**
  - **How many circuits do we need to provision?**
    - **Depends on the blocking probability that we can tolerate**

Circuits	P_B
20	1%
15	8%
7	30%

M/G/1 QUEUE



- **Poisson arrivals at rate λ**

- **Service time has arbitrary distribution with given E[X] and E[X^2]**
  - **Service times are independent and identically distributed (IID)**
  - **Independent of arrival times**
  - **E[service time] = 1/μ**
  - **Single Server queue**

Pollaczek-Khinchin (P-K) Formula

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

where $\rho = \lambda/\mu = \lambda E[X]$ = line utilization

From Little's Theorem,

$$N_Q = \lambda W$$

$$T = E[X] + W$$

$$N = \lambda T = N_Q + \rho$$

M/G/1 EXAMPLES

- **Example 1: M/M/1**

  $E[X] = 1/\mu$; $E[X^2] = 2/\mu^2$

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

- **Example 2: M/D/1 (Constant service time $1/\mu$)**

  $E[X] = 1/\mu$; $E[X^2] = 1/\mu^2$

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

Delay Formulas (summary)

- **M/G/1**

$$T = \bar{X} + \frac{\lambda \overline{X^2}}{2(1-\lambda/\mu)}$$

- **M/M/1**

$$T = \bar{X} + \frac{\lambda/\mu}{\mu-\lambda}$$

- **M/D/1**

$$T = \bar{X} + \frac{\lambda/\mu}{2(\mu-\lambda)}$$

<u>Delay components:</u>

Service (transmission) time (LHS)

Queueing delay (RHS)

Use Little's Theorem to compute N, the average number of customers in the system