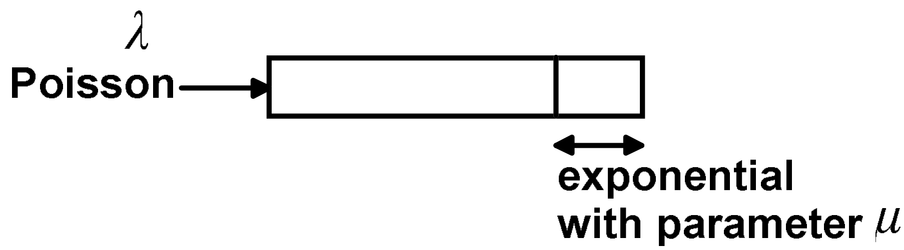
The *M/M/1* Queueing System

M (Poisson arrival process)/M (exponential service times) /1 (1 server)



$$P_r(X_n \geq x) = e^{-\mu x}$$

$$E(x_n) = \frac{1}{\mu}$$

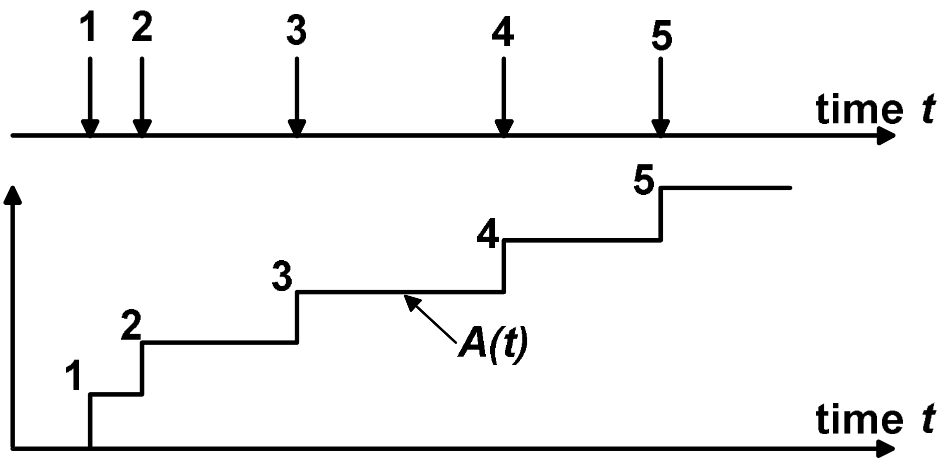
Notation: · / · / · / ·

Arrival process	Service times	# of servers	Max. # of customers in the system
M: Poisson	M: Exponential		
D: Deterministic	D: Deterministic		
G: General	G: General		

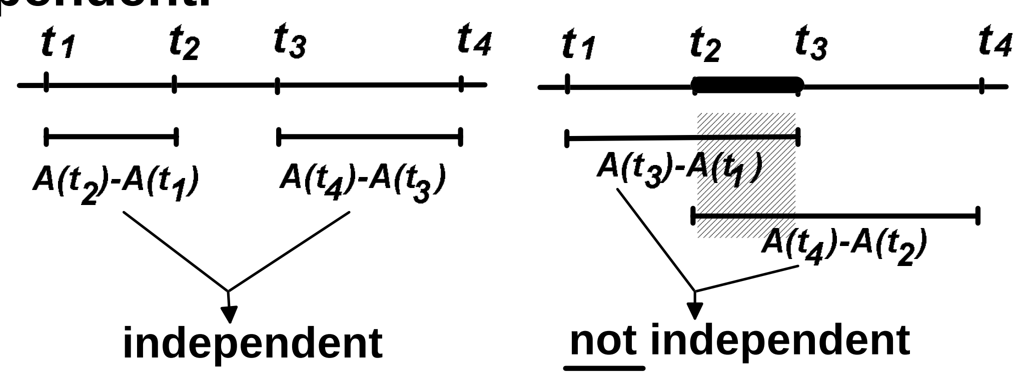
- **We also assume that service times are mutually independent and also independent of all interarrival times.**
- **The server is always serving a customer if any customer is in the system. Assume FCFS service to be specific.**

Poisson Process of Rate λ

A Poisson process $A(t)$ **is a counting process.**
For each $t \geq 0$, $A(t)$ **is a random variable denoting the number of arrivals from** ***0*** **to** ***t*****.**



Number of arrivals in disjoint time intervals are independent.



Number of arrivals in any interval of length τ **is Poisson with parameter** $\lambda \cdot \tau$

$$P\{A(t+\tau) - A(t) = n\} = e^{-\lambda\tau}\frac{(\lambda\tau)^n}{n!}, n = 0, 1, \cdots$$

$$E\{A(t+\tau) - A(t)\} = \lambda \cdot \tau \quad \lambda : \textbf{arrival rate}$$

Properties of Poisson process

Let $t_n =$ **time of** n **th arrival**

$\tau_n = t_{n+1} - t_n =$ **interarrival time**

- $P(\tau_n \geq s) = P\{A(t_n + s) - A(t_n) = 0\} = e^{-\lambda s}$

  (Interarrival times are exponentially distributed with parameter λ **, mean** $\frac{1}{\lambda}$**, variance** $\frac{1}{\lambda^2}$ **)**

- $P(\tau_n \geq r + t \mid \tau_n \geq t) = \frac{P(\tau_n \geq r+t, \tau_n \geq t)}{P(\tau_n \geq t)} = \frac{e^{-\lambda(r+t)}}{e^{-\lambda t}} = e^{-\lambda r} = P(\tau_n \geq r)$

(memoryless)

- **For any *t*, and any (small)** δ**:**

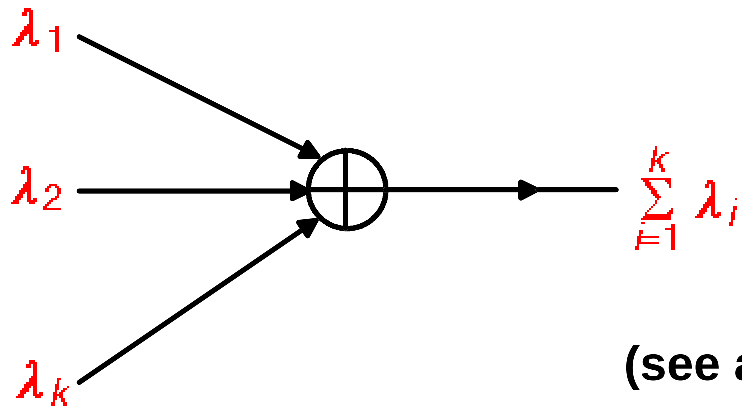
$$P\{A(t+\delta) - A(t) = 0\} = 1 - \lambda\delta + o(\delta)$$
$$P\{A(t+\delta) - A(t) = 1\} = \lambda\delta + o(\delta)$$
$$P\{A(t+\delta) - A(t) \geq 2\} = o(\delta)$$

where $\lim_{\delta \to 0} \frac{o(\delta)}{\delta} = 0$

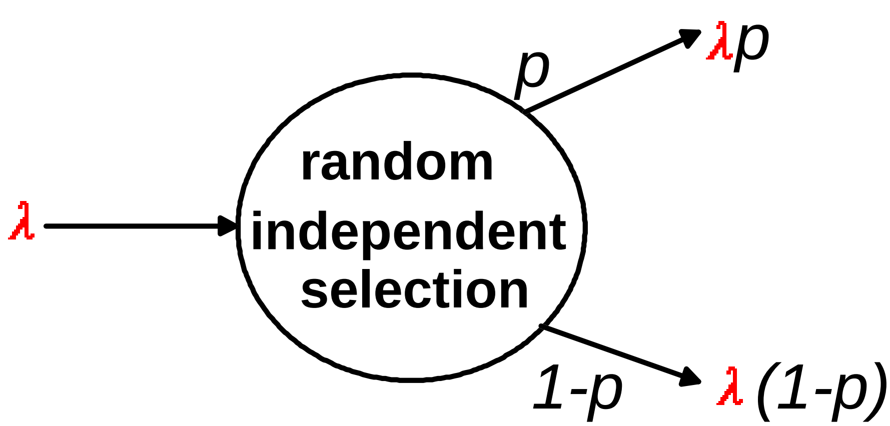
These follow from definition:

$$P\{A(t+\delta) - A(t) = n\} = \frac{e^{-\lambda\delta}(\lambda\delta)^n}{n!}$$

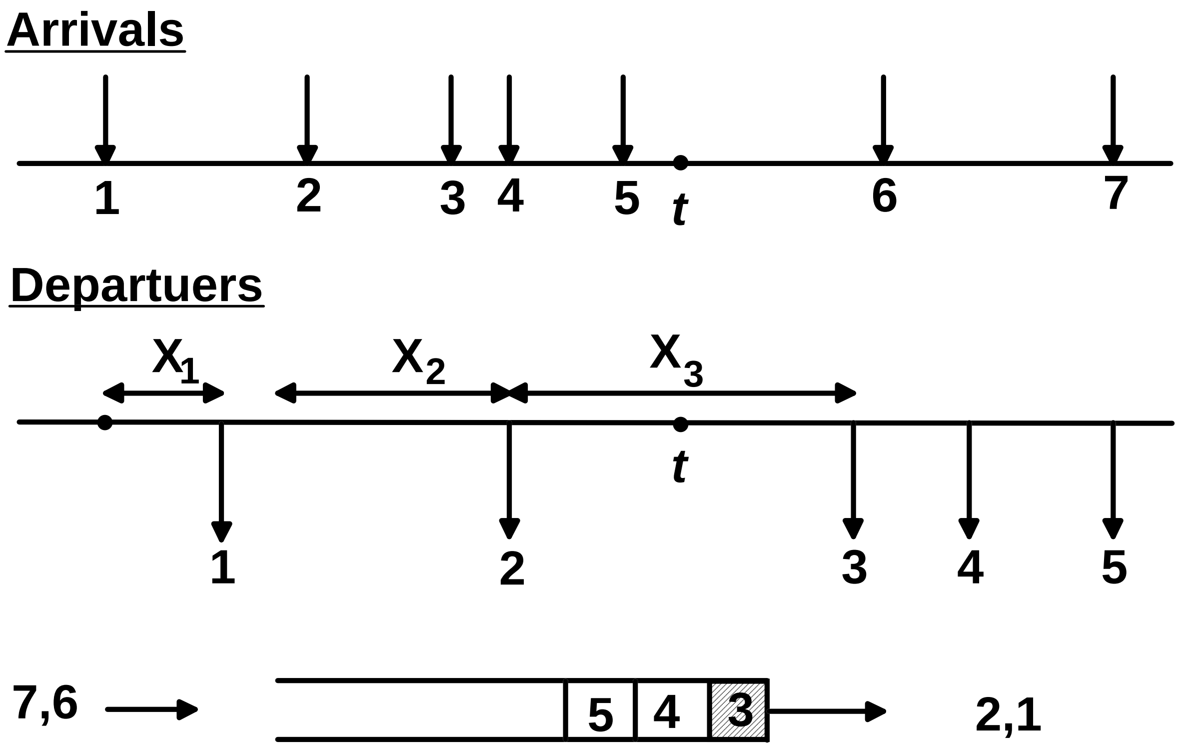
- **If** $A_1(t), A_2(t), \cdots, A_k(t)$ **are independent Poisson processes of rate** $\lambda_1, \lambda_2, \cdots, \lambda_k$**, then** $A_1(t) + A_2(t) + \cdots + A_k(t)$ **is a Poisson process of rate** $\lambda_1 + \lambda_2 + \cdots + \lambda_k$



(see also Ex.3.10c)



If each arrival of a Poisson process is independently sent to system 1 with prob. p **and system 2 with prob.** $1\text{-}p$**, the arrivals to each system are Poisson and independent.(see also Ex.3.11a)**



Situation at time *t*

Starting at a particular time t, the subsequent arrivals do not depend on what has happened in the past, and the subsequent departures depend only on the number *N(t)* of customers in the system at time *t*.

In particular, since service time exponential, it makes no difference, how long the current customer has been in service; the remaining time until departure is still exponential.

Future # of customers in the system depends on past numbers only through the present number *N(t)*.

We focus at times $0, \delta, 2\delta, 3\delta, \cdots\cdots, k\delta, \cdots\cdots$ (δ **small**).

$N_k \stackrel{\text{det}}{=} N(k\delta) =$ **# of customers in the system at time** $k\delta$.

$P_{ij} = P\{N_{k+1} = j \mid N_k = i\}$ **(transition probabilities)**

$P_{00} = 1 - \lambda\delta + o(\delta)$ **(no arrival)**

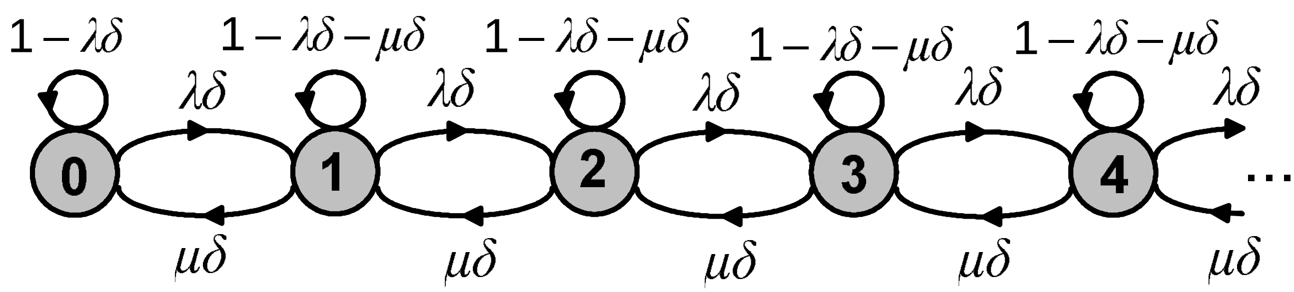
$P_{ii} = 1 - \lambda\delta - \mu\delta + o(\delta),\ i \geq 1$ **(no arrival/departure)**

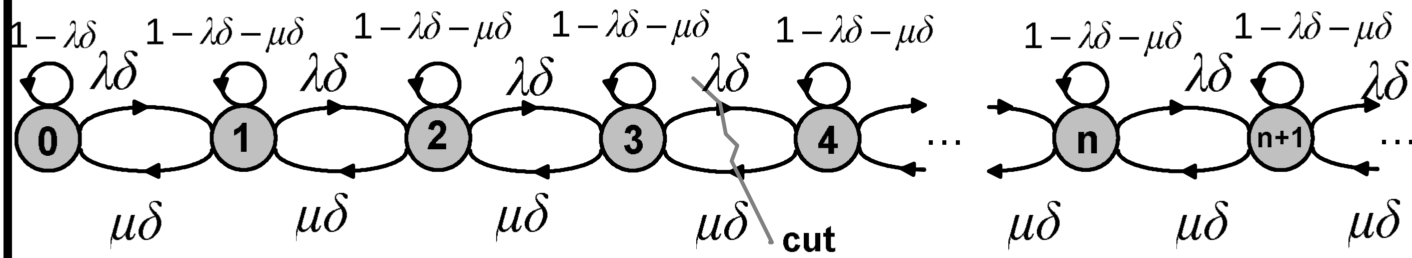
$P_{i,i+1} = \lambda\delta + o(\delta),\ i \geq 0$ **(one arrival)**

$P_{i,i-1} = \mu\delta + o(\delta),\ i \geq 0$ **(one departure)**

[Note: for any state $n \geq 1$**, the server is busy and probability of departure is** $P_r(X \leq \delta) = 1 - e^{-\mu\delta} = \mu\delta + o(\delta)$]

$P_{i,j} = o(\delta),\ j \neq i, i+1, i-1$ **(i.e. the probability of multiple arrivals/departures is negligible.)**





Let p_n be the "steady-state" probability that the system is in state n. [i.e. $p_n = \lim_{t\to\infty} P(N(t) = n)$

Note: over an arbitrarily long period of time, the number of transitions from n to $n+1$ is the same as from $n+1$ to n (plus or minus one).

Thus for any n:

$$p_{n-1}\lambda\delta = p_n\mu\delta \Rightarrow p_n = (\tfrac{\lambda}{\mu})p_{n-1} = (\tfrac{\lambda}{\mu})^2 p_{n-2} = \cdots = (\tfrac{\lambda}{\mu})^n p_0$$

Define

$$\rho = \tfrac{\lambda}{\mu} \text{ (“utilization factor”)}$$

$$\Rightarrow p_n = \rho^n p_0, \ n = 1, 2, \cdots$$

To find p_0:

$$\sum_{n=0}^{\infty} p_n = 1 \Rightarrow \sum_{n=0}^{\infty} \rho^n p_0 = 1 \Rightarrow p_0 \cdot \frac{1}{1-\rho} = 1 \Rightarrow p_0 = 1 - \rho$$

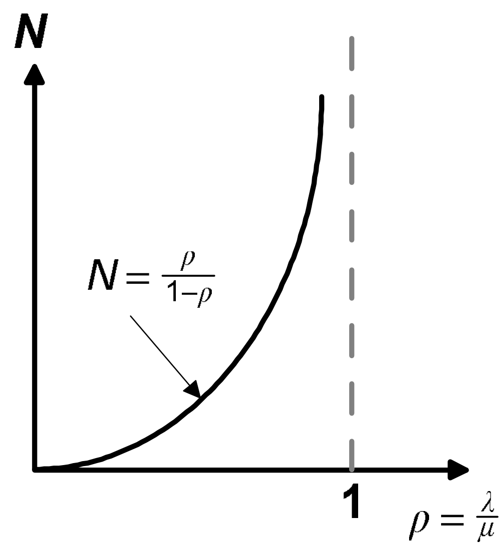
$\Rightarrow p_n = (1-\rho)\cdot\rho^n$, **for** $n \geq 0 \quad \rho = \frac{\lambda}{\mu} < 1$

ρ **= Probability that the system has at least one customer (**$= 1 - p_0$**)**
= Probability server is busy

The expected number N **of customers in the system is**

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

Number in the system blows up as $\rho \to 1$, $N \to \infty$; i.e. as arrival rate λ approaches service rate μ.



From Little's theorem, average customer delay T is

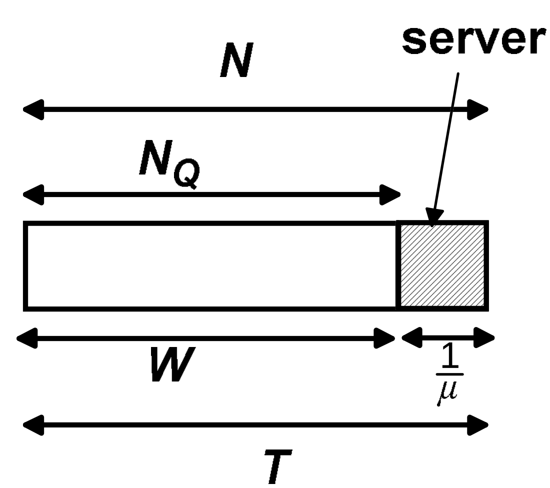
$$T = \frac{N}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu-\lambda} \quad , \lambda < \mu$$

Average time in queue W is

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

Average number of customers in queue N_Q is

$$N_Q = \lambda W = \frac{\rho^2}{1-\rho}$$



Example 1 (Scaling of an *M/M/1* queue)

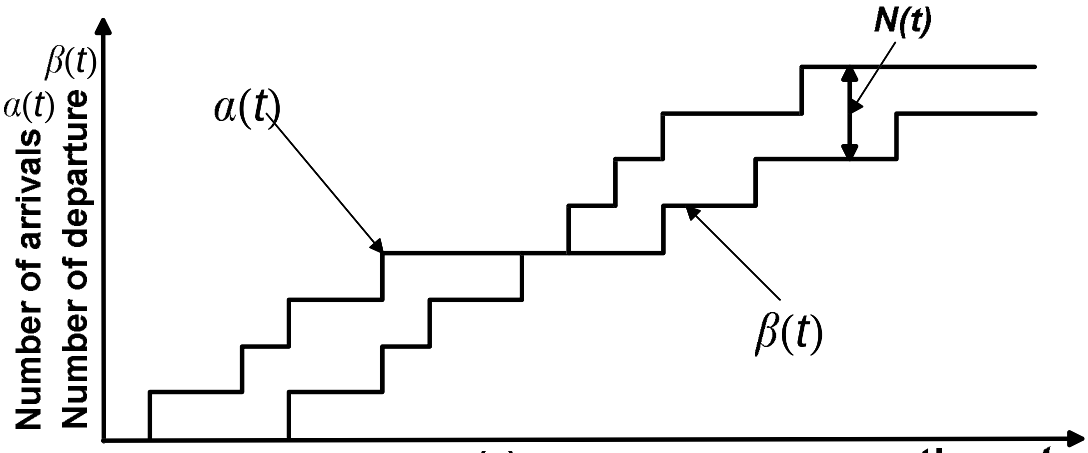
$$N = \frac{\rho}{1-\rho}, p_n = (1-\rho)\rho^n \text{ for } n \geq 0, N_Q = \frac{\rho^2}{1-\rho}$$

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

If one scales the arrival rate λ and the service rate μ, by a constant factor k, N, N_Q, and p_n are unchanged

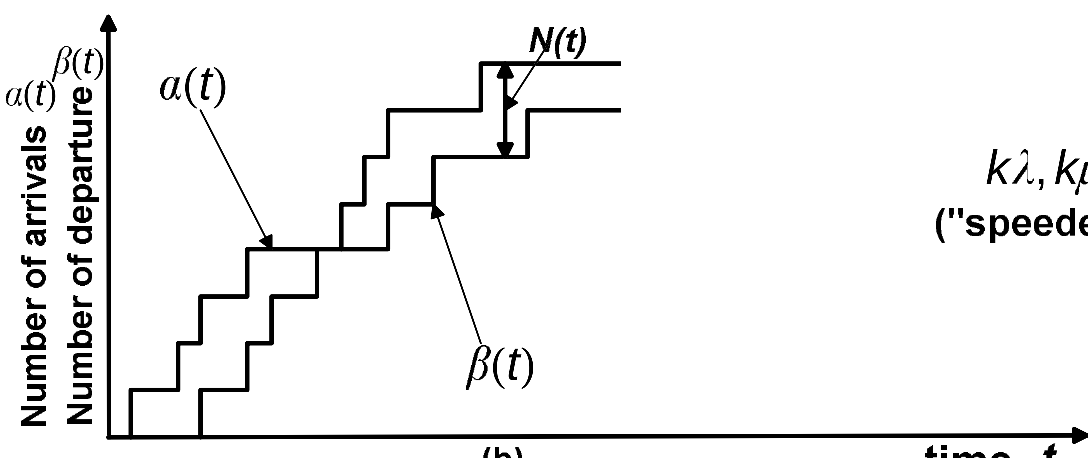
$$T = \frac{1}{\mu-\lambda}, \quad W = \frac{\rho}{\mu-\lambda}$$

System delay T and queueing delay W vary as $\frac{1}{k}$



(a)

λ, μ



(b)

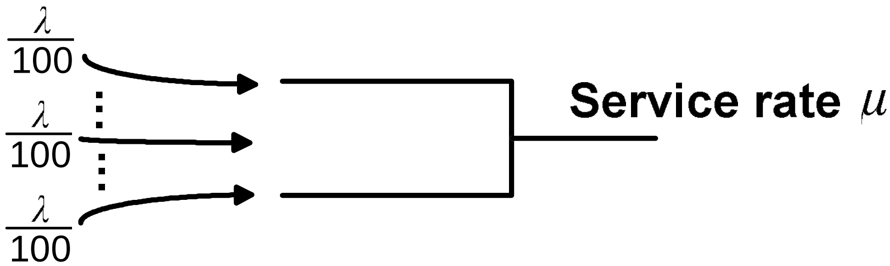
$k\lambda, k\mu, k > 1$

("speeded-up" system

Example 2 (statistical multiplexing verses FDM)

Consider 100 sessions with Poisson arrivals of combined rate λ and exponentially distributed packet lengths sharing a link with service rate μ packets/sec.

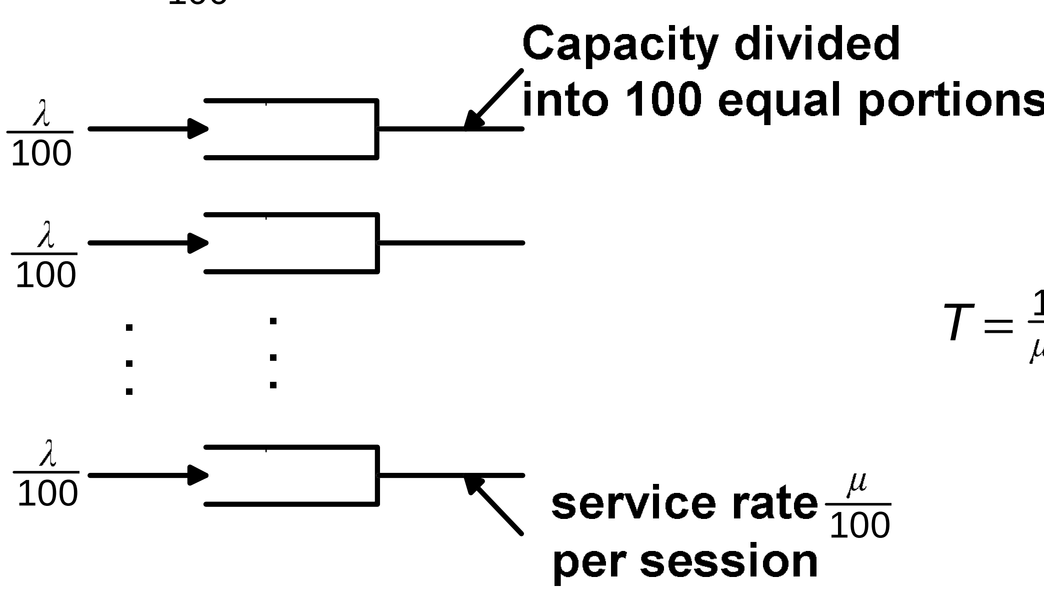
Statistical multiplexing



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

Frequency Division Multiplexing (FDM)

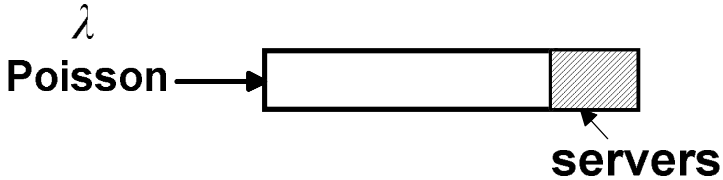
If FDM is used, each session has rate $\frac{\lambda}{100}$ and "sees" service rate $\frac{\mu}{100}$



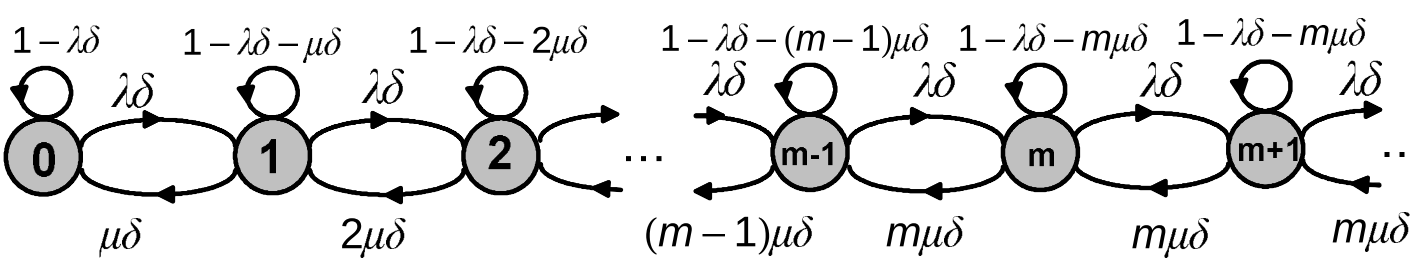
$$T = \frac{100}{\mu - \lambda}$$

M/M/m queue

Poisson arrivals of rate λ
***m* servers, each exponentially distributed with rate μ**



- **Given n customers in the system, $n \le m$, a new arrival will occur in an increment δ with probability $\lambda\delta$. A departure will occur with probability $n\mu\delta$.**
- **For $n > m$ a departure occurs with probability $m\mu\delta$.**



$$\lambda p_{n-1} = n\mu p_n, \quad n \le m$$
$$\lambda p_{n-1} = m\mu p_n, \quad n > m$$

If $n \le m$:

$$p_n = (\frac{\lambda}{n\mu})p_{n-1} = \frac{\lambda^2}{n(n-1)\mu^2}p_{n-2} = \cdots\cdots = \frac{\lambda^n}{n!\mu^n}p_0$$

If $n > m$:

$$p_n = (\frac{\lambda}{m\mu})p_{n-1} = (\frac{\lambda}{m\mu})^2 p_{n-2} = \cdots = (\frac{\lambda}{m\mu})^{n-m} p_m = \frac{\lambda^n p_0}{m^{n-m} m! \mu^n}$$

Let $\rho = \frac{\lambda}{m\mu} < 1 \Rightarrow p_n = \begin{cases} p_0 \frac{(m\rho)^n}{n!}, & n \le m \\ p_0 \frac{m^m \rho^n}{m!}, & n > m \end{cases}$

$$\sum_{n=0}^{\infty} p_n = 1 \Rightarrow \cdots p_0 = [\sum_{n=0}^{m-1} \frac{(m\rho)^n}{n!} + \frac{(m\rho)^m}{m!(1-\rho)}]^{-1}$$

The probability an arriving customer will find all servers busy (and will have to wait) is

$$P\,(\textbf{all servers busy}) = P_Q = \sum_{n=m}^{\infty} p_n = \cdots = \frac{p_0 (m\rho)^m}{m!(1-\rho)}$$

Erlang C formula **used in telephony**

The expected number of customers in queue is

$$N_Q = \sum_{n=0}^{\infty} n p_{m+n} = \sum_{n=0}^{\infty} n p_0 \cdot \frac{m^m \rho^{m+n}}{m!} = \frac{p_0 (m\rho)^m}{m!} \underbrace{\sum_{n=0}^{\infty} n \rho^n}_{\frac{\rho}{(1-\rho)^2}}$$

$$= P_Q \cdot \frac{\rho}{1-\rho} \Rightarrow \frac{N_Q}{P_Q} = \frac{\rho}{1-\rho} \qquad (\rho = \frac{\lambda}{m\mu})$$

Waiting time :

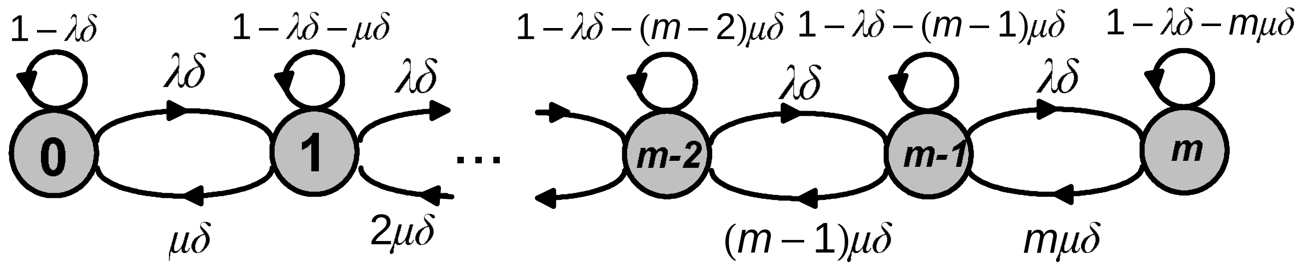
$$W = \frac{N_Q}{\lambda} = \frac{\rho P_Q}{\lambda(1-\rho)}$$

Total time in system:

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

M/M/m/m Queue

Customers arriving when all m servers busy are thrown away, never to return.



$$\lambda p_{n-1} = n\mu p_n \ , \ n = 1, 2, \cdots, m$$
$$p_n = \frac{p_0}{n!} \cdot \left(\frac{\lambda}{\mu}\right)^n, \ n = 1, 2, \cdots, m$$
$$\sum_{n=0}^{m} p_n = 1 \rightarrow p_0 = \left[\sum_{n=0}^{m} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}\right]^{-1}$$

Probability that a customer finds all servers busy is

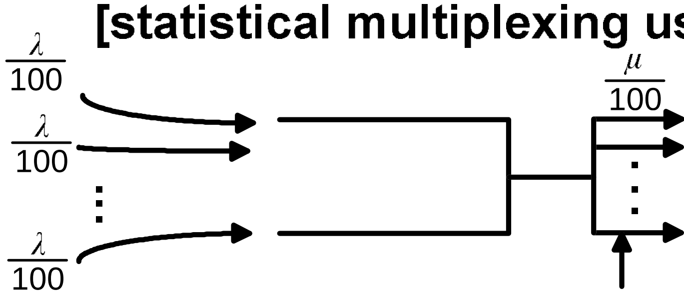
$$p_m = \frac{\frac{(\frac{\lambda}{\mu})^m}{m!}}{\sum_{n=0}^{m} \frac{(\frac{\lambda}{\mu})^n}{n!}}$$

Erlang B formula **(*N*, *T* less than for *M/M/m* but not all customers get served)**

If one uses these models for session arrivals in a voice network, one sees that customer behavior is somewhat between *M/M/m* and *M/M/m/m* behavior—some customers go away if they can't get through and some keep trying.

Example: Assume *m=100* sessions sharing a link. Assume 100 frequency bands, but packets are assigned to any available band. This is an *M/M/100*.

[statistical multiplexing using *m* channels]

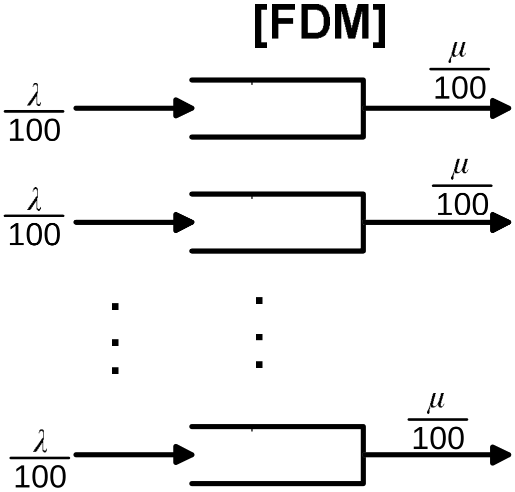


$$T_1 = \frac{100}{\mu} + \frac{p_Q}{\mu-\lambda}$$

100 subchannels with service $\frac{\mu}{100}$ each

Under light load is almost the same as FDM

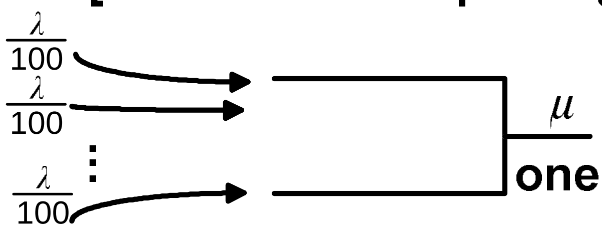
[FDM]



$$T_2 = \frac{100}{\mu-\lambda}$$

Under heavy load delay is almost the same as statistical multiplexing

[statistical multiplexing over the entire channel]



$$T_3 = \frac{1}{\mu - \lambda}$$

one channel with service rate μ

λ **small**: $T_1 \approx T_2 \approx 100T_3$ **,** λ **large** $(\lambda \approx \mu)$: $T_1 \approx T_3 = \frac{T_2}{100}$

$M/M/\infty$ queue.

As in ***M/M/m***, but with $m = \infty$ servers.

Expressions can be found by taking the limit $m \to \infty$ in the ***M/M/m*** expressions:

$$p_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n p_0$$

$$p_0 = \left[1 + \sum_{n=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^n \frac{1}{n!}\right]^{-1} = e^{-\frac{\lambda}{\mu}}$$

$$p_n = \frac{e^{-\frac{\lambda}{\mu}}}{n!}\left(\frac{\lambda}{\mu}\right)^n$$

Average # in the system $N = \frac{\lambda}{\mu}\left(= \sum_{n=0}^{\infty} n p_n\right)$

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