Queueing Theory

M/M/1

- A simple queueing model (M/M/1) queue:

Queueing system: buffer for queue + server for services

Assumptions

(1) Customers arrive according to a Poisson distribution with parameter λt, where λ is the average number of arrivals per unit time.

- Let X be the number of arrivals in $(0, t)$
  $P(X = k) = (\lambda t)^k \exp(-\lambda t)/k!$, for $k = 0, 1, 2, ...$ Its mean = λt
- Suppose we observe an arrival at t and the next arrival at $t + \tau$, where τ is a random variable, then

$$
\begin{aligned}
P(\tau > x) &= P\big(\text{no arrival occurs in } (t, t+x)\big) \\
&= P(X = 0 \text{ in time } (t, t+x)) \\
&= \exp(-\lambda x), \quad x \geq 0
\end{aligned}
$$

$\Rightarrow F_\tau(x) = P(\tau \leq x) = 1 - P(\tau > x) = 1 - \exp(-\lambda x), \;\; x \geq 0$
$\Rightarrow f_\tau(x) = \lambda \exp(-\lambda x), \;\; x \geq 0$
τ has an exponential distribution with parameter λ (average number of customers arrived per unit time). Its mean $= \frac{1}{\lambda}$

Poisson process with parameter λ: probability of k events in $(0, t)$:

$$P(X = k) = \frac{(\lambda t)^k}{k!} e^{-\lambda t}.$$

In $(0, \Delta t)$, we have:

- Probability of 0 event: $\frac{(\lambda \Delta t)^0}{0!} e^{-\lambda \Delta t} = 1 - \lambda \Delta t + o(\Delta t)$. Here $o(\Delta t)$ means higher order of Δt.
- Probability of 1 event: $\frac{(\lambda \Delta t)^1}{1!} e^{-\lambda \Delta t} = \lambda \Delta t + o(\Delta t)$.
- Probability of two or more events: $o(\Delta t)$.

When $\Delta t \to 0$, we can omit $o(\Delta t)$. So we have at most one event: the probability of no event is $1 - \lambda \Delta t$, and the probability of one event is $\lambda \Delta t$.

Assumptions (cont'd)

(2) The single server takes a random length of time T_s to serve each customer and these times are independent random variables for different customers.

- T_s has an exponential distribution with parameter μ

  $f_{T_s}(t) = \mu \exp(-\mu t), \;\; t \geq 0$

  $F_{T_s}(t) = 1 - \exp(-\mu t), \;\; t \geq 0$

  μ: the average number of customers being served per unit time.

  The mean of T_s is $\frac{1}{\mu}$.

  So if the queue is always full, the departure process is Poisson!

Property of exponential distribution

$$
\begin{aligned}
P(t \le T_s \le t + \Delta t | T_s \ge t) &= \frac{P(t \le T_s \le t + \Delta t)}{P(T_s \ge t)} \\
&= \frac{F_{T_s}(t + \Delta t) - F_{T_s}(t)}{1 - F_{T_s}(t)} \\
&= \frac{\exp(-\mu t)(1 - \exp(-\mu \Delta t))}{\exp(-\mu t)} \\
&= 1 - \exp(-\mu \Delta t)
\end{aligned}
$$

For whatever time duration you have been waiting for the service to be finished, the probability that the service is finished within the next delta t duration is a constant. The constant depends only on delta t.

i.e., the probability of the completion of a service in the next Δt seconds is a constant, independent of how long the service has been going on. The service time has no memory!

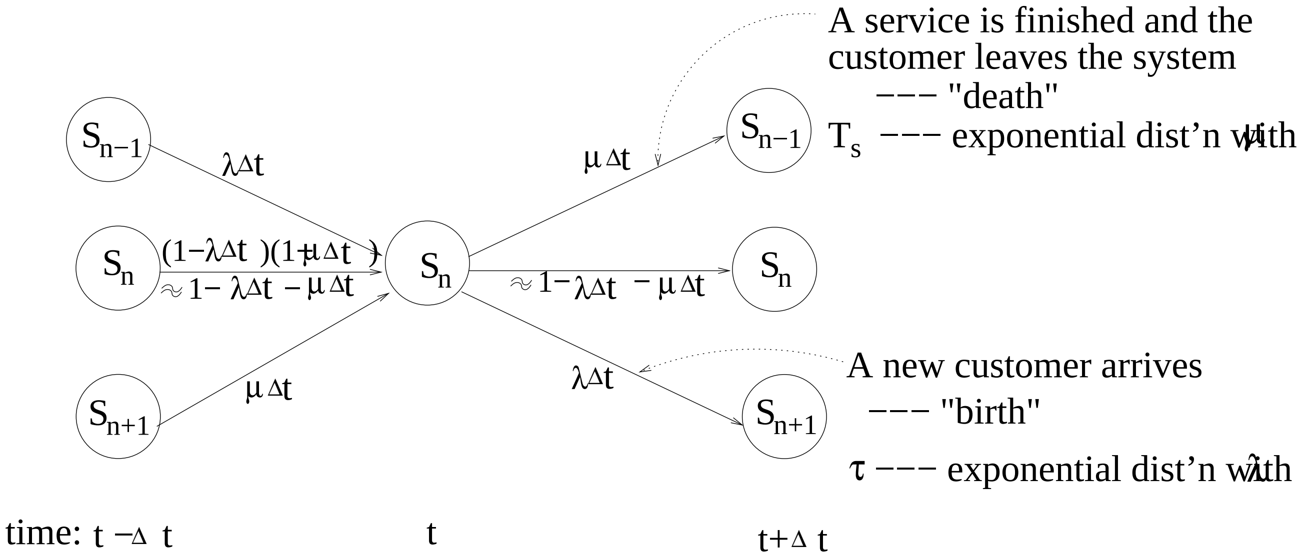
Both random variables τ and T_s have no memory!

M/M/1 queue

- Summary: A single server queue with Poisson arrival and exponential service times is denoted as M/M/1 where M stands for Markov – a process with no memory.

Analysis of M/M/1 queue

- State S_n: there are n customers in the system (including the one being served, if any)



The process is called "birth - death" process.

From state S_n ($n > 0$), within very small Δt, there is at most one birth, and there is at most one death.

Probability of 0 birth: $1 - \lambda\Delta t$.

Probability of 1 birth: $\lambda\Delta t$.

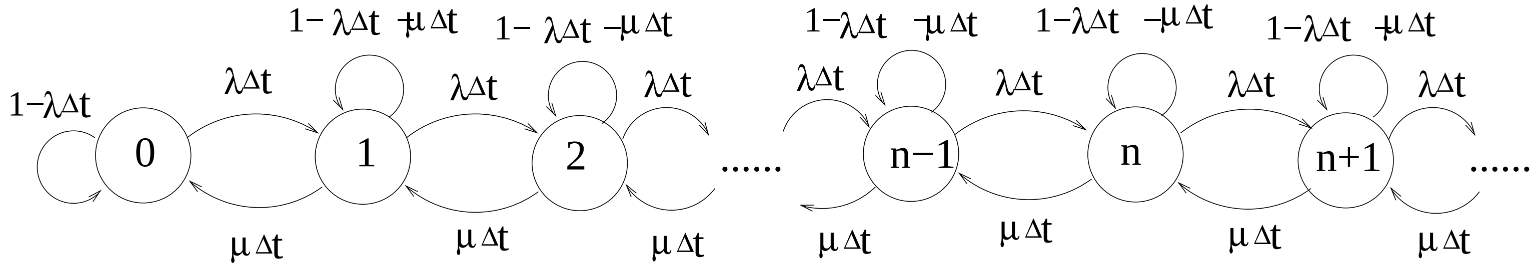
Probability of 0 death: $1 - \mu\Delta t$.

Probability of 1 death: $\mu\Delta t$.

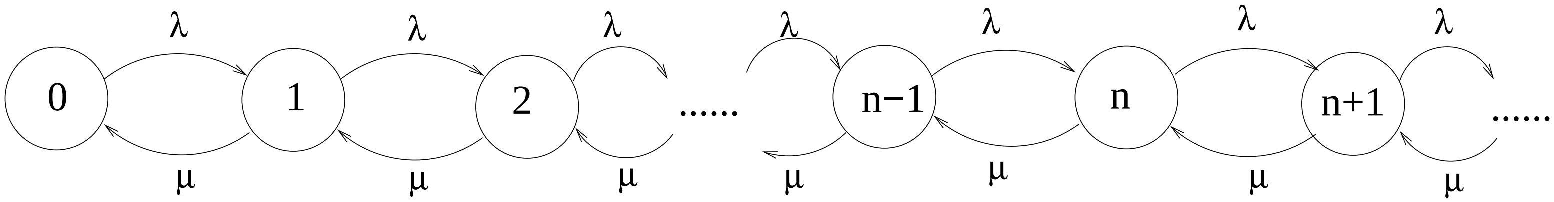
So from S_n:

- To S_{n+1}: 1 birth & 0 death, with probability $\lambda\Delta t(1 - \mu\Delta t) = \lambda\Delta t$.
- To S_{n-1}: 0 birth & 1 death, with probability $(1 - \lambda\Delta t)\mu\Delta t = \mu\Delta t$.
- Remain at S_n: 0 birth & 0 death, or 1 birth & 1 death, with probability $(1 - \lambda\Delta t)(1 - \mu\Delta t) + \lambda\Delta t \cdot \mu\Delta t = 1 - \lambda\Delta t - \mu\Delta t$.

- Discrete-time Markov Chain



- Continuous-time birth-death process



State representation: birth−death process

- Let Y denote the time duration between adjacent state transitions, then $Y = \min(T_s, \tau)$ has an exponential distribution with parameter $(\lambda + \mu)$. (Proof it yourself. Hint: find the prob. of no arrival and no departure during $[t, t + \Delta t)$, given packet arrival and departure are independent processes.)

$$\begin{aligned} P(Y \leq y) &= P(\min(T_s, \tau) \leq y) = 1 - P(\min(T_s, \tau) > y) \\ &= 1 - P(T_s > y)P(\tau > y) = 1 - [1 - P(T_s \leq y)][1 - P(\tau \leq y)]. \end{aligned}$$

- Y has no memory.

- Note: In a Poisson process, as t increases, the number of arrivals (or departures) also increases; in queueing system, as t increases, the number of customers in the system may increase (with new arrivals) or decrease (with new departures); therefore, S_n is not Poisson process.

M/M/1 queue

Let $P_n(t)$ denote the probability that the system is in state n at time t, then

$$\begin{cases} P_n(t+\Delta t) &= P_n(t)(1-\lambda\Delta t-\mu\Delta t)+P_{n-1}(t)\lambda\Delta t+P_{n+1}(t)\mu\Delta t, \;\; n\geq 1 \\ P_0(t+\Delta t) &= P_0(t)(1-\lambda\Delta t)+P_1(t)\mu\Delta t \end{cases}$$

$\Rightarrow$

$$\begin{cases} \frac{P_n(t+\Delta t)-P_n(t)}{\Delta t} &= -(\lambda+\mu)P_n(t)+\lambda P_{n-1}(t)+\mu P_{n+1}(t), \;\; n\geq 1 \\ \frac{P_0(t+\Delta t)-P_0(t)}{\Delta t} &= -\lambda P_0(t)+\mu P_1(t) \end{cases}$$

Let $\Delta t \to 0$, we have

$$\begin{cases} P_n'(t) &= -(\lambda+\mu)P_n(t)+\lambda P_{n-1}(t)+\mu P_{n+1}(t), \;\; n\geq 1 \\ P_0'(t) &= -\lambda P_0(t)+\mu P_1(t) \end{cases}$$

This indicates the fundamental recursive relationship of the “birth-death” process with “birth” parameter λ and “death” parameter μ.

Steady state solution

As $t \to \infty$, $P_n(t) \to P_n$, (not a function of t), i.e., $P_n'(t) = 0$, $P_0'(t) = 0$, (for irreducible Markov process),

$$\Rightarrow \begin{cases} (\lambda + \mu)P_n = \lambda P_{n-1} + \mu P_{n+1}, \quad n \geq 1 \\ \lambda P_0 = \mu P_1 \end{cases} \quad \text{Balance Condition}$$

$$\Rightarrow \begin{cases} P_1 = \frac{\lambda}{\mu} P_0 = \rho P_0 \\ P_2 = \frac{\lambda}{\mu} P_1 = \rho^2 P_0 \\ \ldots \\ P_n = \frac{\lambda}{\mu} P_{n-1} = \rho^n P_0 \end{cases}$$

where $\rho = \frac{\lambda}{\mu}$ is called the traffic intensity or utilization factor. $\rho < 1$ is the stable requirement; otherwise, the queue length will keep on increasing to ∞.

- Find P_0:
  Since $\sum_{i=0}^{\infty} P_i = 1$, we have $P_0(1 + \rho + \rho^2 + ...) = P_0 \frac{1}{1-\rho} = 1$, if $0 < \rho < 1$.
  Thus, $P_0 = 1 - \rho$.
  $P_n = P_0 \rho^n = (1 - \rho)\rho^n$, for $n = 0, 1, 2, ...$ —- a geometric distribution.

- The average number of customers in the *system* is:
  $L = E[N] = \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} n(1 - \rho)\rho^n = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}$

- The average number of customers in the *queue* is:

$$\begin{aligned} E[N_q] &= \sum_{n=1}^{\infty}(n-1)P_n = \sum_{n=1}^{\infty}(n-1)(1-\rho)\rho^n \\ &= (1-\rho)[\sum_{n=1}^{\infty} n\rho^n - \sum_{n=1}^{\infty} \rho^n] = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)} \end{aligned}$$

- The time that a customer must wait in the queue is denoted as T_q:
If there are n customers in the system when a new customer arrives, then the customer will have to wait equal to $T_1 + T_2 + ... + T_n$, where $\{T_i\}$ is a set of independent and identical distributed (i.i.d.) random variables. The conditional PDF of T_q given n can be derived as
$f_{T_q}(t|n) = \frac{\mu^n t^{n-1}}{(n-1)!} \exp(-\mu t), t \geq 0, n \geq 1,$
which is a gamma distribution with parameters n and μ. [Proof is shown in next page.]

Derive $f_{T_q}(t|n)$

Let Y be the number of customers that have departed in $(0, t)$

$$
\begin{aligned}
F_{T_q}(t|n) &= P(T_q \leq t|n) = P(Y \geq n \text{ in}(0,t)) \\
&= \sum_{k=n}^{\infty} \frac{(\mu t)^k}{k!} \exp(-\mu t)
\end{aligned}
$$

$$
\begin{aligned}
f_{T_q}(t|n) &= \frac{\mathrm{d}}{\mathrm{d}t} F_{T_q}(t|n) \\
&= \sum_{k=n}^{\infty} [\mu k \frac{(\mu t)^{k-1}}{k!} - \frac{\mu(\mu t)^k}{k!}] \exp(-\mu t) \\
&= \sum_{k=n}^{\infty} [\mu \frac{(\mu t)^{k-1}}{(k-1)!} - \frac{\mu(\mu t)^k}{k!}] \exp(-\mu t) \\
&= \frac{\mu^n t^{n-1}}{(n-1)!} \exp(-\mu t), \;\; t \geq 0, n \geq 1.
\end{aligned}
$$

To derive the distribution of T_q:

$$
\begin{aligned}
P(T_q > \tau) &= \sum_{n=1}^{\infty} P(T_q > \tau | n) P_n \\
&= \sum_{n=1}^{\infty} \int_{\tau}^{\infty} f_{T_q}(t|n) \mathrm{d}t P_n \\
&= \sum_{n=1}^{\infty} (1-\rho)\rho^n \int_{\tau}^{\infty} \frac{\mu^n t^{n-1}}{(n-1)!} \exp(-\mu t) \mathrm{d}t \\
&= (1-\rho)\rho\mu \int_{\tau}^{\infty} \sum_{n=1}^{\infty} \frac{(\rho\mu t)^{n-1}}{(n-1)!} \exp(-\mu t) \mathrm{d}t \\
&= (1-\rho)\rho\mu \int_{\tau}^{\infty} \exp(\rho\mu t) \exp(-\mu t) \mathrm{d}t \\
&= \rho \exp(-(1-\rho)\mu\tau), \quad \tau \geq 0
\end{aligned}
$$

We can also directly calculate CDF of T_q (the probability of T_q less than or equal to tau). Then the summation of n should be from 0 to infinity. We can get the same CDF expression.

- The PDF of T_q is:

$$f_{T_q}(\tau) \quad = \quad \frac{\mathrm{d}}{\mathrm{d}\tau}[1 - P(T_q > \tau)] = \rho\mu(1-\rho)\exp(-(1-\rho)\mu\tau), \;\; \tau > 0$$

- Average queueing time:

$$\begin{aligned} E[T_q] &= \int_0^\infty \tau f_{T_q}(\tau)\mathrm{d}\tau = \int_0^\infty \tau\rho\mu(1-\rho)\exp(-(1-\rho)\mu\tau)\mathrm{d}\tau \\ &= \frac{\rho}{\mu(1-\rho)} = \frac{\lambda}{\mu(\mu-\lambda)}, \;\; \mu > \lambda \end{aligned}$$

- Let W denote the mean time that a customer has to wait from the moment he arrives until he departures:
$W = E[T_q + T_s] = E[T_q] + \frac{1}{\mu} = \frac{1}{\mu-\lambda}$

- The mean number of customers in the system is
$L = \frac{\lambda}{\mu-\lambda} = \lambda W \Rightarrow$ Little's Law

Little's Law

- "Little's Law" says that for any work-conserving queueing system, the average occupancy of the system, must equal the average delay for the system multiplied by the average arrival rate.

L	ρ
1/3	0.25
1	0.5
3	0.75

$L = \lambda W = \frac{\lambda}{\mu-\lambda} = \frac{\rho}{1-\rho}$. (Draw the L vs. ρ figure yourself.)

- As the utilization increases, the delay increases correspondingly. At $\rho = 0.5$, the average delay is twice the average transmission time, ($E[T_q] = E[T_s]$).

- The probability that the queue exceeds a specified number:
$P(n > N) = \sum_{n=N+1}^{\infty} P_n = (1-\rho) \sum_{n=N+1}^{\infty} \rho^n = \rho^{N+1}$

N	$P(n > N)$	$(\rho = 0.6)$
1	0.36	
3	0.13	
9	0.0061	
19	3.7×10^{-5}	

Example 2.1

- In an M/M/1 queue, customers arrive at the rate of $\lambda = 15$ per hour. What is the minimum server rate to ensure that
  - the server is idle at least 10% of the time?
  - the expected value of the queue length is not to exceed 10?
  - the probability that at least 20 people in the queue is at most 50%?

Solution

- For $P_0 \geq 0.1 \Rightarrow 1 - \rho \geq 0.1 \Rightarrow \lambda/\mu \leq 0.9$. Given $\lambda = 15$ per hour, $\mu \geq 16.67$ per hour.
- $E[N_q] = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)} \leq 10 \Rightarrow \mu \geq 16.375$ per hour.
- $P_n = (1-\rho)\rho^n, \;\; n = 0, 1, 2, ...$
  $P(n \geq 21) = \sum_{n=21}^{\infty}(1-\rho)\rho^n \leq 0.5$
  $\Rightarrow \mu \geq 15 \;\exp(-\ln(0.5)/21)$
  Note: for that at least 20 people in the queue, there are at least 21 people in the system.

Example 2.2

- Consider a packet transmission system whose arrival rate (in packet/sec) is $k\lambda$ ($k > 1$), and departure rate is $k\mu$ (service time is $1/k\mu$). What is the average number of packets in the system? What is the average delay per packet?

Solution

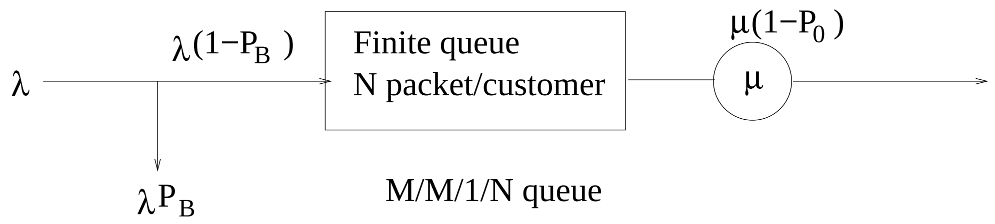
- $L = \frac{k\lambda}{k\mu - k\lambda} = \frac{\lambda}{\mu - \lambda}$
- According to the Little's Law, the average delay per packet is $W = \frac{L}{k\lambda} = \frac{1}{k(\mu - \lambda)}$.
  (Comments: Increasing the arrival and transmission rates by the same factor, the average delay is reduced by the factor, and the average number of customers in the system remains the same.)

M/M/1/N

- In M/M/1/N system, the maximum number of packets in the system is N (with a finite buffer holding at most $N-1$ packets):
  $\sum_{n=0}^{N} P_n = P_0 \sum_{n=0}^{N} \rho^n = P_0[\frac{1-\rho^{N+1}}{1-\rho}] = 1$
  $\Rightarrow P_0 = \frac{1-\rho}{1-\rho^{N+1}}$
  $P_n = \frac{(1-\rho)\rho^n}{1-\rho^{N+1}}$
- If $\rho^{N+1} << 1$, $P_n \approx (1-\rho)\rho^n$ (When (a) $N \to \infty$, or (b) $\rho \to 0$, M/M/1/N $\to$ M/M/1)

Blocking probability of M/M/1/N

- With a finite buffer, the blocking probability $P_B = P_N = \frac{(1-\rho)\rho^N}{1-\rho^{N+1}}$

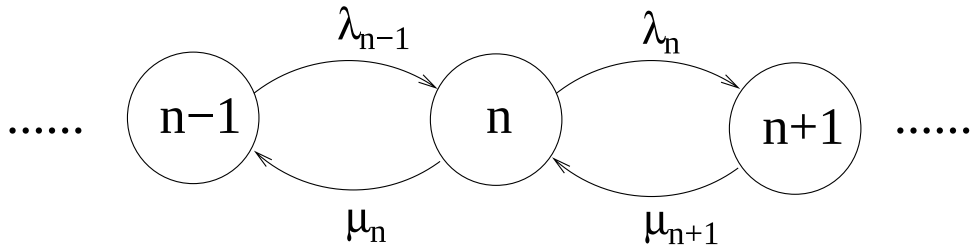


- Arrival rate: λ packet/sec
- Blocking probability: P_B
- Departure rate: μ packet/sec
- Throughput $\lambda(1 - P_B) = \mu(1 - P_0)$

Queues with dependence on state of system

- Multi-server situation
- Customer arrival rate decreasing with queue occupancy to keep the average occupancy down

Queues with dependence on state of system



State representation: birth−death process

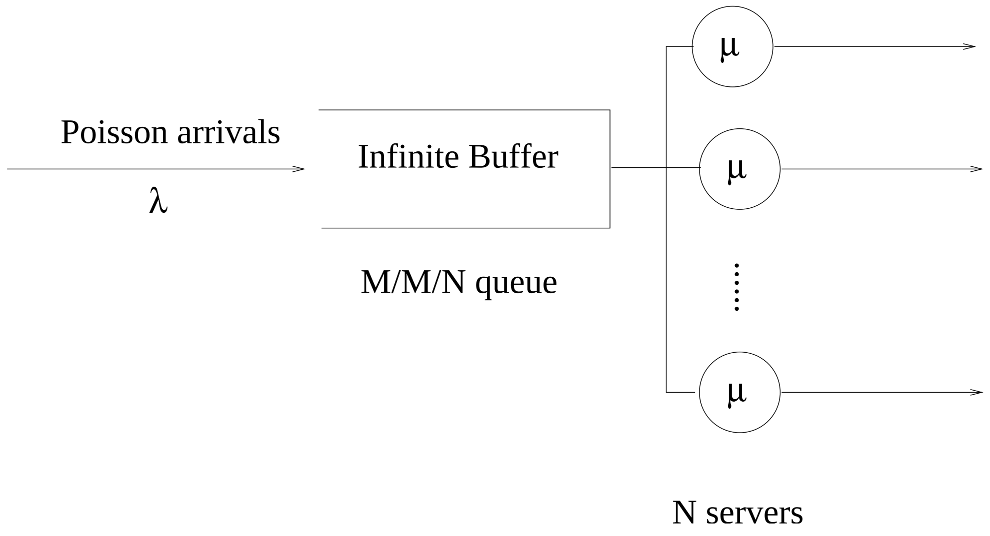
- λ and μ are function of n. The global balance equations for the steady-state probabilities P_n are

  $(\lambda_n + \mu_n)P_n = \mu_{n+1}P_{n+1} + \lambda_{n-1}P_{n-1}, \;\; n \geq 1$

  $\lambda_0 P_0 = \mu_1 P_1$

  $\Rightarrow P_n = \frac{\lambda_0 \lambda_1 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \ldots \mu_n} P_0$

Multi-server case: M/M/N



$$\lambda_n = \lambda, \mu_n = \begin{cases} n\mu, & n \leq N \\ N\mu, & n > N \end{cases}$$

$$P_n = \begin{cases} (\frac{\lambda}{\mu})^n \frac{P_0}{n!}, & n \leq N \\ (\frac{\lambda}{\mu})^n N^{N-n} \frac{P_0}{N!}, & n > N \end{cases}$$

With the condition $\sum_{n=0}^{\infty} P_n = 1$, we obtain

$$\begin{aligned} P_0 &= \left[1 + \sum_{n=1}^{N-1} (\frac{\lambda}{\mu})^n / n! + \sum_{n=N}^{\infty} (\frac{\lambda}{\mu})^n \frac{N^{N-n}}{N!}\right]^{-1} \\ &= \left[1 + \sum_{n=1}^{N-1} \rho^n / n! + \sum_{n=N}^{\infty} \rho^n \frac{N^{N-n}}{N!}\right]^{-1} \end{aligned}$$

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

If $N \to \infty$, $\sum_{n=0}^{\infty} P_n = P_0 \sum_{n=0}^{\infty} \frac{\rho^n}{n!} = P_0 \exp(\rho) = 1$

$\Rightarrow P_0 = \exp(-\rho), P_n = \frac{\rho^n \exp(-\rho)}{n!}$

Queue with discouragement

- With infinite queue size, $\mu_n = \mu$, $\lambda_n = \frac{\lambda}{n+1}$. (As the queue size increase, the arrival rate drops accordingly), it can be derived:

  $P_n = \frac{\rho^n}{n!} P_0 = \frac{\rho^n}{n!} \exp(-\rho)$

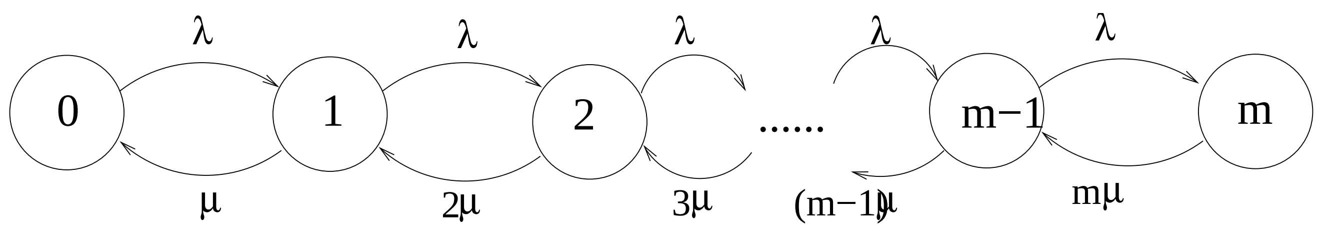
  $P_0 = \exp(-\rho)$

Example

- Consider a M/M/m/m system (This model is in wide use in telephony or circuit switched networks). In this context, customers in the system correspond to active telephone conversations and the m servers represent a single transmission line consisting of m circuits. The principle quantity of interest here is the blocking probability, i.e., the steady-state probability that all circuits are busy, in which case an arriving call is refused service, and the blocked calls are lost.

Solution

State representation:



The balance condition is

$\lambda P_{n-1} = n\mu P_n, n = 1, 2, ..., m.$

$P_n = P_0(\frac{\lambda}{\mu})^n \frac{1}{n!}, \;\; n = 1, 2, ..., m.$

Solving for P_0 in the equation $\sum_{n=0}^{m} P_n = 1$,

$P_0 = [\sum_{k=0}^{m}(\frac{\lambda}{\mu})^k \frac{1}{k!}]^{-1}$

$P_n = (\frac{\lambda}{\mu})^n \frac{1}{n!}[\sum_{k=0}^{m}(\frac{\lambda}{\mu})^k \frac{1}{k!}]^{-1}$

The probability that an arrival will find all m servers busy and will therefore be lost is

$$P_m = (\frac{\lambda}{\mu})^m \frac{1}{m!}[\sum_{k=0}^{m}(\frac{\lambda}{\mu})^k \frac{1}{k!}]^{-1}$$

This equation is known as the **Erlang B formula**, and find wide use in evaluating the blocking probability of telephone system.