Single Queuing Systems

A counting process $\{\underbrace{N(t)}, t \geq 0\}$

representing the
# of events
that have
occurred
up to time t

M/M/1 queuing system

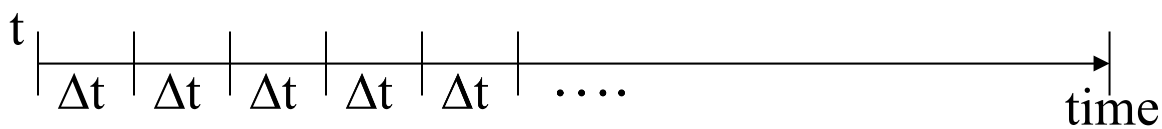
— arrival process is a Poisson process (or the inter-arrival time is exponentially distributed)

— service process is also a Poisson process (or the service time is exponentially distributed)

advantage: a mathematically tractable model with solutions applicable to a wide variety of situations.

Poisson process with an average arrival rate λ:

λ is the proportionality constant



Pr(exactly 1 arrival in [t,t+Δt]) = λΔt

Pr(no arrivals in [t,t+Δt]) = 1 - λΔt

$$\text{pr}\left\{ \begin{matrix} \text{1 event} \\ \text{occurs at } t+\Delta t \end{matrix} \middle| \begin{matrix} \text{event does not} \\ \text{occur at } t \end{matrix} \right\}$$

$$= 1 - e^{-\lambda \Delta t}$$

$$= 1 - \left\{ 1 + \left(-\lambda \Delta t\right) + \frac{\left(-\lambda \Delta t\right)^2}{2!} + \ldots \right\}$$

$$\approx \lambda \Delta t$$

Analogy:

Coin flipping: results of coin flips are independent

Arrivals are also independent

Let $P_n(t) \equiv P$ (# of arrivals = n at time t)

$P_{ij}(\Delta t) \equiv$ the prob. of going from i arrivals to j arrivals in a time interval of Δt seconds

$$\therefore P_n(t+\Delta t) = P_n(t)\overbrace{P_{n,n}(\Delta t)}^{1-\lambda\Delta t} + P_{n-1}(t)\underbrace{P_{n-1,n}(\Delta t)}_{\equiv \lambda\Delta t}$$

$$P_0(t+\Delta t) = P_0(t)\underbrace{P_{0,0}(\Delta t)}_{1-\lambda\Delta t}$$

$$\frac{P_n(t+\Delta t) - P_n(t)}{\Delta t} = -\lambda P_n(t) + \lambda P_{n-1}(t)$$

$$\frac{P_0(t+\Delta t) - P_0(t)}{\Delta t} = -\lambda P_0(t)$$

$$0 \xrightarrow{\lambda} 1 \xrightarrow{\lambda} 2 \xrightarrow{\lambda} 3 \xrightarrow{\lambda} \cdots\cdot$$

Let $\Delta t \to 0$ then

$$\frac{dP_n(t)}{dt} = -\lambda P_n(t) + \lambda P_{n-1}(t) \quad \text{———} \quad ①$$

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t) \text{———} ② \overset{\text{solution}}{\Rightarrow} P_0(t) = e^{-\lambda t}$$

From ①

$$\frac{dP_1(t)}{dt} = -\lambda P_1(t) + \lambda e^{-\lambda t} \quad \therefore P_1 = \lambda t e^{-\lambda t}$$

$$\frac{dP_2(t)}{dt} = -\lambda P_2(t) + \lambda^2 t e^{-\lambda t} \quad \therefore P_2 = \frac{\lambda^2 t^2}{2} e^{-\lambda t}$$

Continuing, by induction,

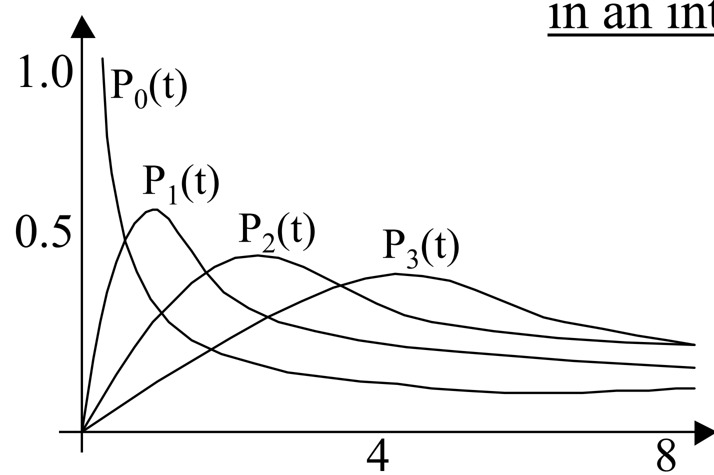
$$\boxed{P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}}$$ — Poisson distribution

meaning: prob. of n arrivals in an interval of t seconds

$$\sum_{n=0}^{\infty} P_n(t) = 1$$

$$\sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} = 1$$

$$\boxed{\sum_{n=0}^{\infty} \frac{X^n}{n!} = e^X}$$



The Poisson distribution

Ex: λ = 100 arrivals/min., what is the prob. of no arrivals in 5 sec.?

$$P_0(5\text{ sec.}) = e^{-100\cdot\left(\frac{1}{12}\right)} = 0.00024$$

* the mean & the variance of the Poisson dist. are both equal to λt.

mean: $$\overline{n(t)} = \sum_{n=1}^{\infty} nP_n(t) = \sum_{n=1}^{\infty} n \cdot \frac{(\lambda t)^n e^{-\lambda t}}{n!} = e^{-\lambda t} \cdot \sum_{n=1}^{\infty} \frac{(\lambda t)^n}{(n-1)!}$$

$$= e^{-\lambda t} \cdot \sum_{n=0}^{\infty} \frac{(\lambda t)^{n+1}}{n!} = e^{-\lambda t} \cdot \lambda t \cdot \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} = e^{-\lambda t} \cdot \lambda t \cdot e^{\lambda t} = \lambda t$$

derivation is based on

$$e^{\lambda t} = \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!}$$

variance: $$[\sigma_n(t)]^2 = \sum_{n=0}^{\infty} n^2 P_n(t) - \left[\overline{n(t)}\right]^2 = \lambda t$$

The inter-arrival time T ← a random variable

Inter-arrival time cumulative dist. function(t) ← cdf (t)

= P (time between arrivals ≤ t)

= 1-$\underbrace{\text{P (time between arrivals > t}}_{P_0(t)}$) ∵ no arrivals in a time interval of t = $P_0(t)$

= 1-$P_0(t)$

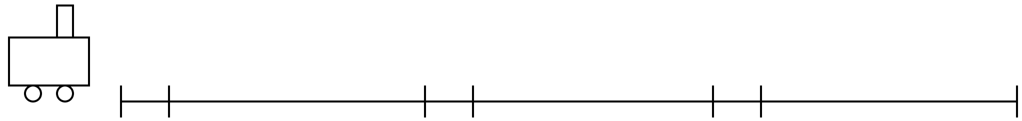
= 1-$e^{-\lambda t}$

∴ inter-arrival time density (t) = $\dfrac{d(1-e^{-\lambda t})}{dt} = \lambda e^{-\lambda t}$ ← pdf (t)

∴ T is an exponentially distributed r.v.

∴ T has a memory-less property

e.g.,



an average of
mean interarrival time = 20 min.

$\frac{1}{\lambda}$ = mean interarrival time

The last train arrived 19 minutes ago.
What is the expected time until the next train arrives?

~~1~~   20 min. ✓

coin flipping explanation in Δt

	prob.
yes	$\lambda \Delta t$
no	$1-\lambda \Delta t$

Memory-less property ≡ <u>Markov property</u>

$$P(T > t_0 + t \mid T > t_0) = P(T > t)$$

* Definition: a <u>Markov chain</u> is a Markov process with a discrete state space.

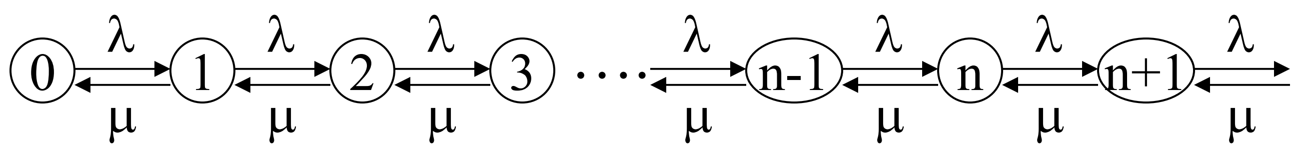
M/M/1

$$P_n(t+\Delta t) = P_n(t)\overbrace{P_{n,n}(\Delta t)}^{(1-\lambda\Delta t)(1-\mu\Delta t)} + P_{n-1}(t)\overbrace{P_{n-1,n}(\Delta t)}^{\lambda\Delta t} + P_{n+1}(t)\overbrace{P_{n+1,n}(\Delta t)}^{\mu\Delta t}$$

$$P_0(t+\Delta t) = P_0(t)\underbrace{P_{0,0}(\Delta t)}_{1-\lambda\Delta t} + P_1(t)\underbrace{P_{1,0}(\Delta t)}_{\mu\Delta t}$$

$$\frac{dP_n(t)}{dt} = -(\lambda+\mu)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t)$$

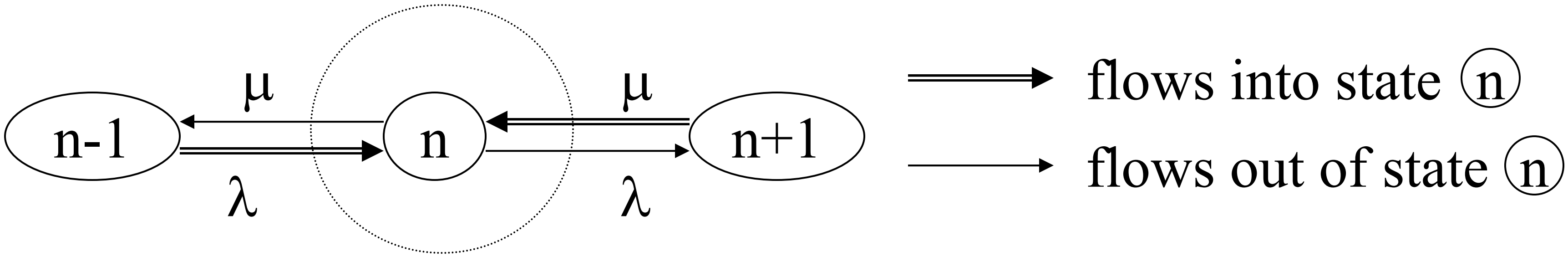
$$\frac{dP_0(t)}{dt} = -\lambda P_0(t) + \mu P_1(t)$$



Probability flux (or flow):

(probability of a state)*(transition rate originating from the state)

physical meaning: # of times per second the event corresponding to the transition occurs.



$$\frac{dP_n(t)}{dt} = \underbrace{-(\lambda + \mu)P_n(t)}_{\text{flows out of state n}} + \underbrace{\lambda P_{n-1}(t) + \mu P_{n+1}(t)}_{\text{flows into state n}}$$

<u>Study:</u>

— transient behavior $\frac{dP_n(t)}{dt} \neq 0$

— equilibrium behavior $\frac{dP_n(t)}{dt} = \frac{dP_{n-1}(t)}{dt} = ... = \frac{dP_0(t)}{dt} = 0$

this yields

<u>Global balance equations</u>: a set of linear equations for t→∞

Equilibrium state probabilities

conservation of probability:

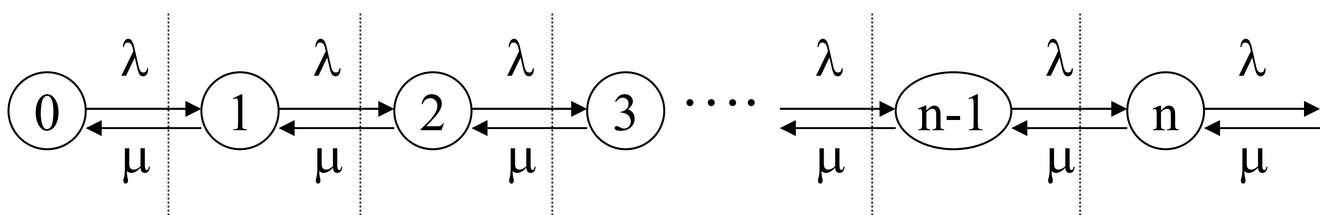
$$\sum_{i=0}^{\infty} P_i(\infty) = 1 \longrightarrow \boxed{\text{normalization equation}}$$

Use local balance equations to solve the global balance equations

1. Local satisfies global
2. Local allows us to relate P_n with a reference state, e.g., P_0

Definition of local balance:

"the probability flow <u>into</u> a state due to an arrival to a queue equals the probability flow <u>out of</u> the same state due to a departure from the same queue"



P_0: P_0

P_1: $P_0\lambda = P_1\mu \Rightarrow \cancel{P_1} = \left(\frac{\lambda}{\mu}\right)P_0$

P_2: $P_1\lambda = P_2\mu \Rightarrow \cancel{P_2} = \left(\frac{\lambda}{\mu}\right)\cancel{P_1}$

$P_{n-1}\lambda = P_n\mu \Rightarrow \cancel{P_n} = \left(\frac{\lambda}{\mu}\right)\cancel{P_{n-1}}$

$$P_n = \left(\frac{\lambda}{\mu}\right)^n P_0$$

applying the normalization equation

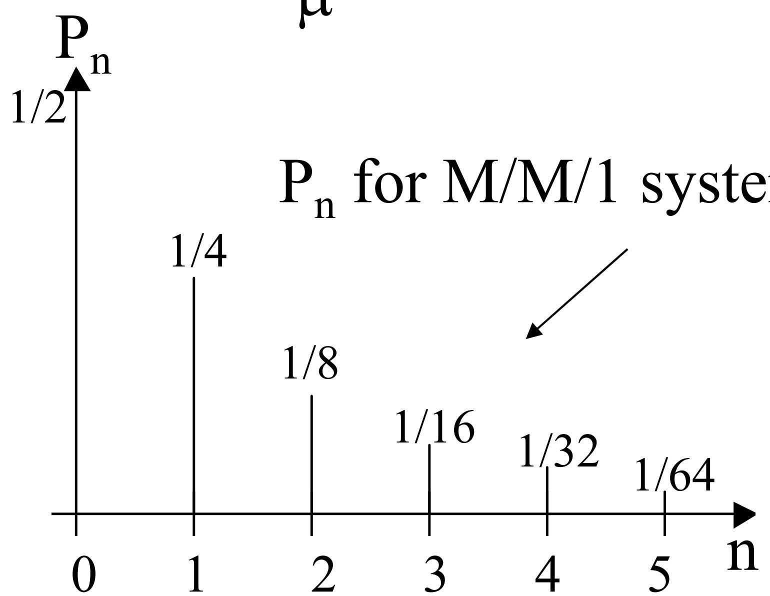
$$\therefore P_0 \sum_{n=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^n = 1 \qquad \sum_{n=0}^{\infty} x^n = \frac{1}{1-x} \quad 0 \le x < 1$$

$$P_0 = \frac{1}{\sum_{n=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^n} = 1 - \left(\frac{\lambda}{\mu}\right) \quad \text{if } 0 \le \frac{\lambda}{\mu} < 1$$

$$\therefore P_n = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right)$$

Utilization: prob. that an M/M/1 queuing system is nonempty

$$\text{Let } \frac{\lambda}{\mu} = \rho, \quad P_n = \rho^n (1-\rho)$$



P_n for M/M/1 system when $\frac{\lambda}{\mu} = \rho = \frac{1}{2}$

	(light) *ρ=0.1	ρ=0.5	(heavy) ρ=0.9
$P(0 \le n \le 3)$	0.9999	0.9375	0.3439
$P(4 \le n \le 7)$	10^{-4}	0.0586	0.2256
$P(8 \le n \le 11)$	10^{-8}	0.0366	0.1480
$P(n \ge 12)$	10^{-12}	0.000249	0.2825

* for a lightly loaded system, there are usually less than 4 customers in the system.

check $\rho = 1 - P_0$?

$$\rho = 1 - P_0 = 1 - (1 - \frac{\lambda}{\mu}) = \frac{\lambda}{\mu}$$

* $\rho \le 1$ otherwise $\lambda > \mu$ and the queuing system would no longer be in equilibrium $\rightarrow$ i.e., unstable.

Q1:throughput?

$$x = \sum_{n=1}^{\infty} P_n * \mu$$

$$= \mu \cdot \sum_{n=1}^{\infty} P_n$$

$$= \mu (1 - P_0)$$

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

because when there is no customer, there is no contribution to throughput.

Q2: Average # of customers in the queuing system?

$$\bar{n} = \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} n \cdot \rho^n (1-\rho) = (1-\rho) \sum_{n=0}^{\infty} n \cdot \rho^n = (1-\rho) \cdot \rho \sum_{n=0}^{\infty} n \cdot \rho^{n-1}$$

$$= (1-\rho) \cdot \rho \frac{d}{d\rho} \sum_{n=0}^{\infty} \rho^n = (1-\rho) \cdot \rho \cdot \frac{d}{d\rho} (\frac{1}{1-\rho}) = (1-\rho) \cdot \rho \cdot \frac{1}{(1-\rho)^2}$$

$$= \frac{\rho}{1-\rho}$$

$$\boxed{\bar{n} = \frac{\rho}{1-\rho}}$$

e.g.,

$$\lambda = \tfrac{1}{2}\mu$$

$$\therefore \rho = \tfrac{\lambda}{\mu} = \frac{\frac{1}{2}\mu}{\mu} = \frac{1}{2}$$

$$\bar{n} = \frac{\frac{1}{2}}{\frac{1}{2}} = 1$$

e.g.,

$$\lambda = \tfrac{2}{3}\mu$$

$$\therefore \rho = \tfrac{2}{3}$$

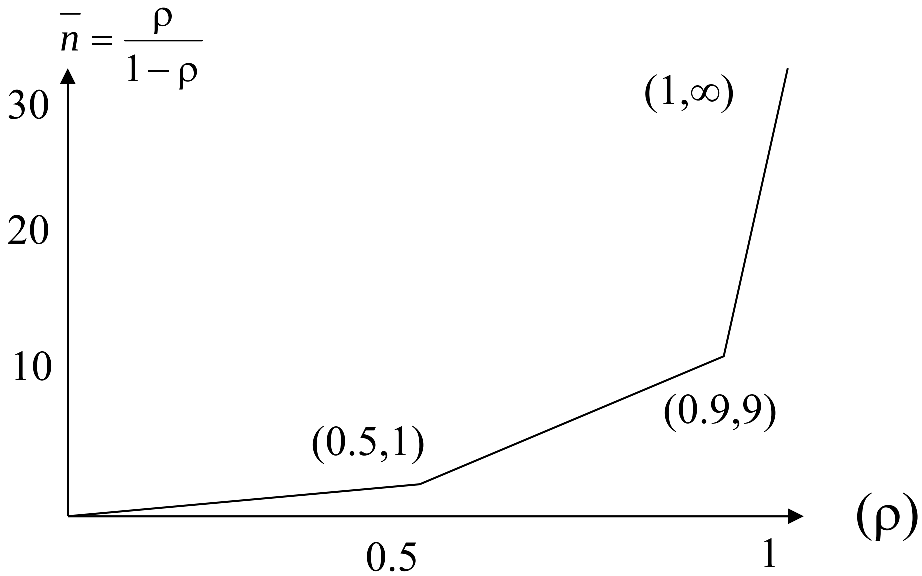
$$\bar{n} = \frac{\frac{2}{3}}{\frac{1}{3}} = 2$$

Let R be the mean response time per customer

Q3: R?

since $\bar{n} = \lambda R$ by little's law (to be discussed later)

$$R = \frac{\bar{n}}{\lambda} = \frac{\rho}{(1-\rho)\lambda} = \frac{1/\mu}{(1-\frac{\lambda}{\mu})}$$ $\longrightarrow$ When ρ=1 system is unstable



M/M/1: average # of customers $\bar{n}$ as a function of ρ

M/M/1

service time

waiting time

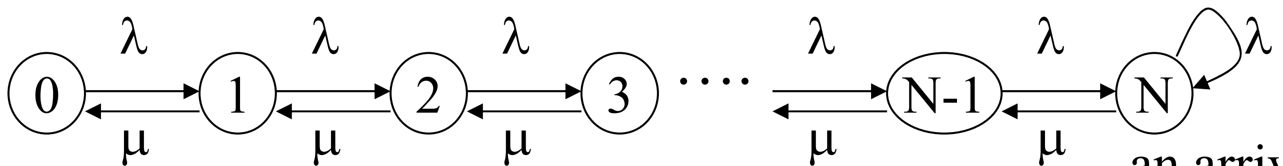
$$R = (1+\bar{n}) \cdot D$$

$$= (1+\bar{n}) \cdot \frac{1}{\mu}$$

$$= \left(1 + \frac{\rho}{1-\rho}\right) \cdot \frac{1}{\mu}$$

$$= \frac{1}{1-\rho} \cdot \left(\frac{1}{\mu}\right)$$

M/M/1/N Queuing system: the finite buffer case



an arriving customer is "lost" or "turned away" when there are already N customers in the system.

Following the previous derivation for M/M/1/∞,

$$P_n = \left(\frac{\lambda}{\mu}\right)^n P_0$$

&

$$P_0 = \frac{1}{\sum_{n=0}^{N}\left(\frac{\lambda}{\mu}\right)^n} = \frac{1}{\dfrac{1-\left(\frac{\lambda}{\mu}\right)^{N+1}}{1-\left(\frac{\lambda}{\mu}\right)}} = \frac{1-\left(\frac{\lambda}{\mu}\right)}{1-\left(\frac{\lambda}{\mu}\right)^{N+1}}$$

<u>no restriction on the range of</u> $\frac{\lambda}{\mu}$

$$\therefore \boxed{P_n = \frac{1-\left(\frac{\lambda}{\mu}\right)}{1-\left(\frac{\lambda}{\mu}\right)^{N+1}} \cdot \left(\frac{\lambda}{\mu}\right)^n}$$

$N \nearrow \quad P_N \swarrow$

$\lambda/\mu \nearrow \quad P_N \nearrow$

Q1: the prob. that the queuing system is full? P_N

Q2: how fast are customers lost? $P_N \times \lambda$

when N=5

$\frac{\lambda}{\mu}$	$P_N = P_5$ (blocking probability)
0.1	$9*10^{-6}$
0.5	0.016
0.75	0.072
1.00	0.166
2.00	0.508
5.00	0.800

applying L'Hopital's rule

$$\lim_{x \to 1} \frac{x^5 - x^6}{1 - x^6} = \lim_{x \to 1} \frac{5x^4 - 6x^5}{-6x^5} = 0.166$$

Q3: population?

$$\bar{n} = \sum_{n=0}^{N} n \cdot P_n$$

Q4: throughput?

$$x = \sum_{n=1}^{N} \mu \cdot P_n = \mu \cdot \sum_{n=1}^{N} P_n = \mu(1 - P_0) = \mu\rho < \lambda$$

ρ is utilization

Q5: Utilization?

$$\rho = 1 - P_0 = 1 - \frac{1 - \frac{\lambda}{\mu}}{1 - \left(\frac{\lambda}{\mu}\right)^{N+1}} < \frac{\lambda}{\mu}$$

Q1: throughput? λ

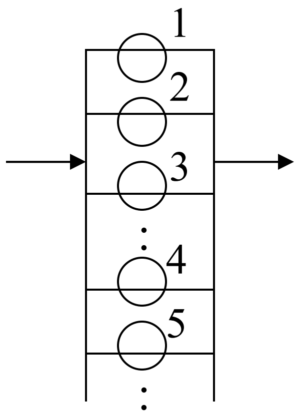
Q2: response time? $1/\mu$

Q3: population? λ/μ

by Little's law

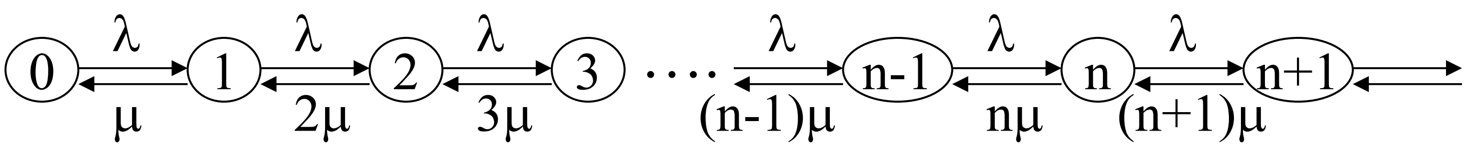
Variations of M/M/1

M/M/∞:



infinite # of servers

$$\boxed{\bar{n} = \sum_{n=1}^{\infty} n \cdot P_n = \sum_{n=1}^{\infty} n \frac{\left(\lambda \cdot \frac{1}{\mu}\right)^n e^{-\lambda \cdot \frac{1}{\mu}}}{n!} = \frac{\lambda}{\mu}}$$



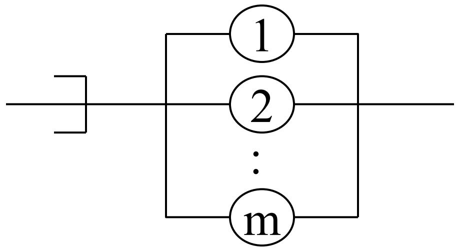
$$P_n = \left(\prod_{i=1}^{n} \frac{\lambda}{i\mu}\right) \cdot P_0 = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n P_0$$

$$\therefore P_0 = \frac{1}{\sum_{n=0}^{\infty} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n} = \frac{1}{e^{+\left(\frac{\lambda}{\mu}\right)}} = e^{-\frac{\lambda}{\mu}}$$

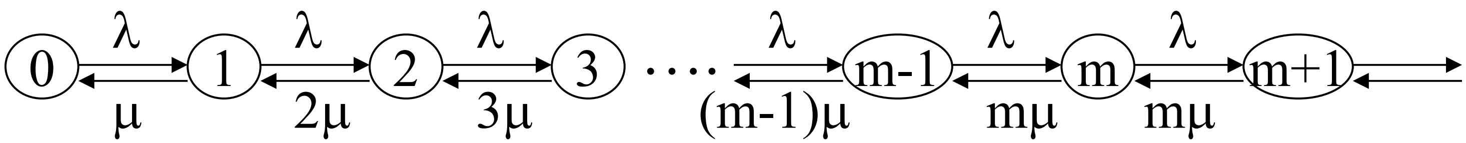
$$\boxed{P_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n e^{-\lambda/\mu}}$$

a Poisson process with mean $\bar{n} = \frac{\lambda}{\mu}$

M/M/m



m servers
e.g., a system with m processors



Solution:

$$P_n = \frac{1}{m!\,m^{n-m}}\left(\frac{\lambda}{\mu}\right)^n \cdot P_0$$

can be obtained by consider 2 cases separately

$$\mu_n = \begin{cases} n\mu & 0 \le n \le m \\ m\mu & n \ge m \end{cases}$$

where $$P_0 = \left[1 + \sum_{n=1}^{m-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{m!}\left(\frac{\lambda}{\mu}\right)^m \left(\frac{1}{1-\rho}\right)\right]^{-1}$$

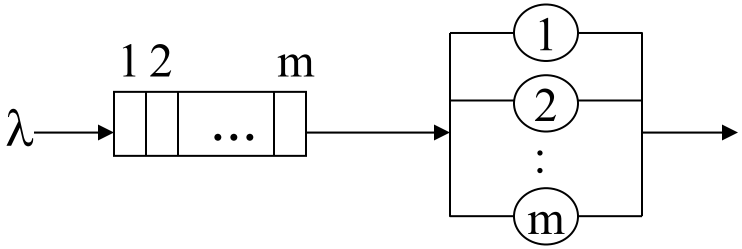
where $\rho = \dfrac{\lambda}{m\mu}$

Q1: what is the probability that all servers are busy? Ans: $\sum_{n=m}^{\infty} P_n$

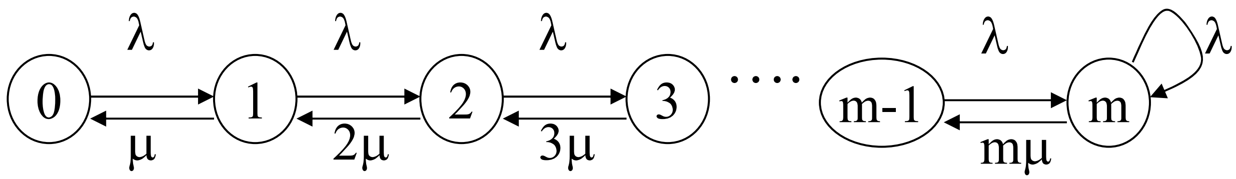
Q2: throughput? Ans: λ

Q3: response time?

M/M/m/m



m servers with a single queue having a buffer space of m (when all servers are busy, a customer walks away), e.g., a telephone switching system.



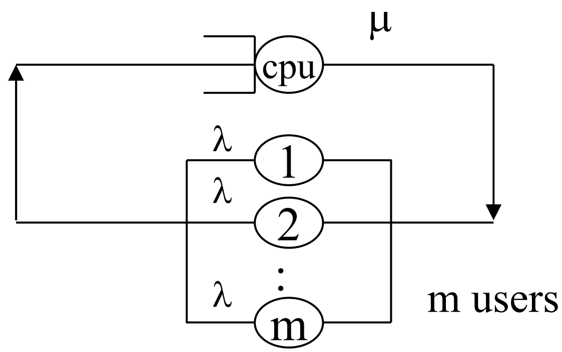
$$P_n = \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n P_0 \qquad \therefore P_0 = \frac{1}{\sum_{n=0}^{m} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n} \quad \& \quad P_n = \frac{\frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n}{\sum_{i=0}^{m} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^i}$$

Q1. Prob. that all m servers are busy (e.g., in a telephone switch company)? P_m ⟵ The expression for P_m is called Erlang's B formula.

Q2. Mean # of calls turned away per time unit? $P_m \times \lambda$

A Client-Server System

Request arrival rate per user : λ



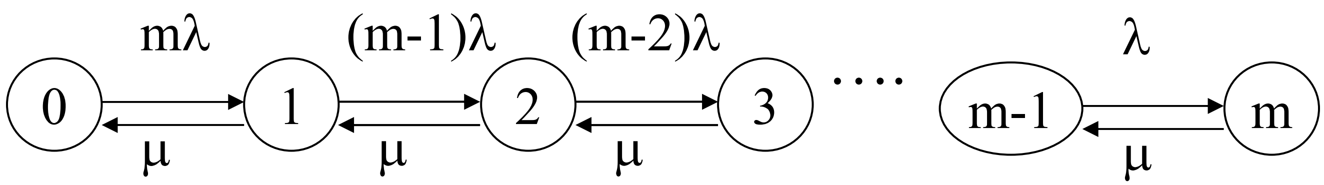
Service rate of the server system with one server: μ

Response time: the time spent by a user at the system between submitting the request & the return of the response

State Description: one state component representation

n: a number representing the # of users in the server system

∴ # of users still thinking (i.e., not issuing requests) = m - n



Recall in M/M/1

$$P_n = \left(\frac{\lambda}{\mu}\right)^n P_0$$

$$\therefore P_n = \left(\prod_{i=1}^{n} \frac{\lambda(m-i+1)}{\mu}\right) P_0$$

$$\text{or } P_n = \left(\frac{\lambda}{\mu}\right)^n \frac{m!}{(m-n)!} P_0$$

$$\text{where } P_0 = \frac{1}{\sum_{n=0}^{m}\left[\left(\frac{\lambda}{\mu}\right)^n \frac{m!}{(m-n)!}\right]}$$

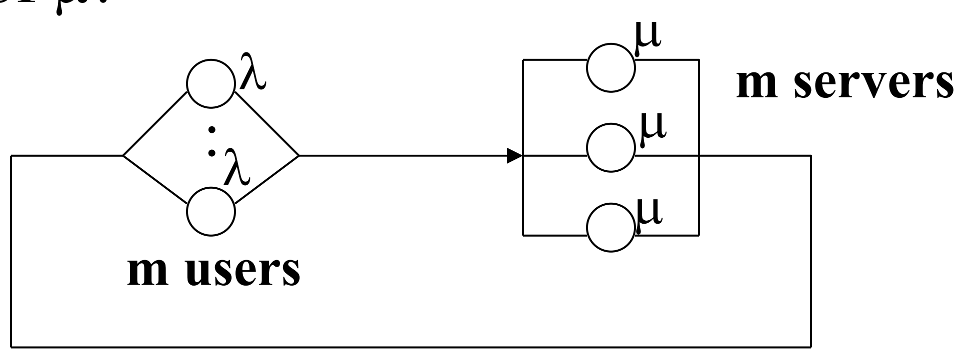
Q1: Avg. # of users in the server system? $\bar{n} = \sum_{n=1}^{m} n \cdot P_n$

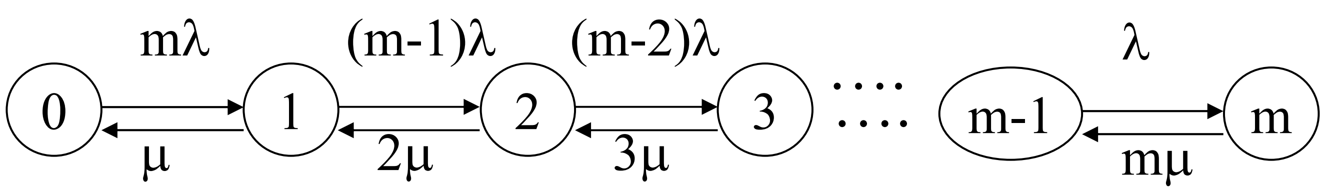
Q2: Avg. # of users still thinking (not issuing requests)? $m - \bar{n}$

Q3: System throughput? $x = \sum_{n=1}^{m} \mu \cdot P_n = (1 - P_0) \cdot \mu$

Q4: Response time per user?

What happens if the server system has m servers, each with a service rate of μ?

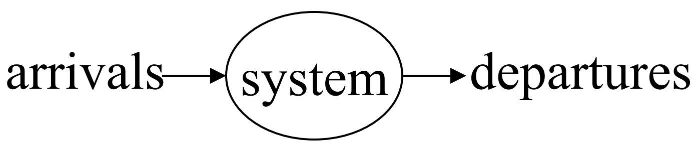


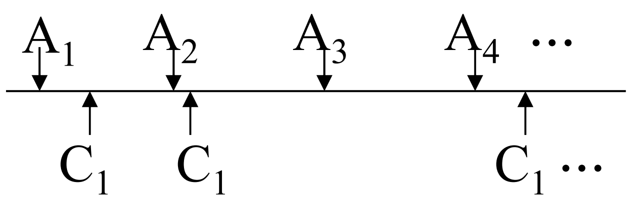


Q1: Throughput? $$x = \sum_{n=1}^{m} \left(P_n * n\mu \right)$$

Q2: Response time?

<u>Fundamental Laws</u>: algebraic relationships among performance measurement quantities.





λ = arrival rate = A/T e.g., $\frac{100\ arrivals}{1\ hr}$

C = # of completions

x = C/T throughput

B = total system busy time

D = B/C average service time per request

ρ = B/T utilization of the system

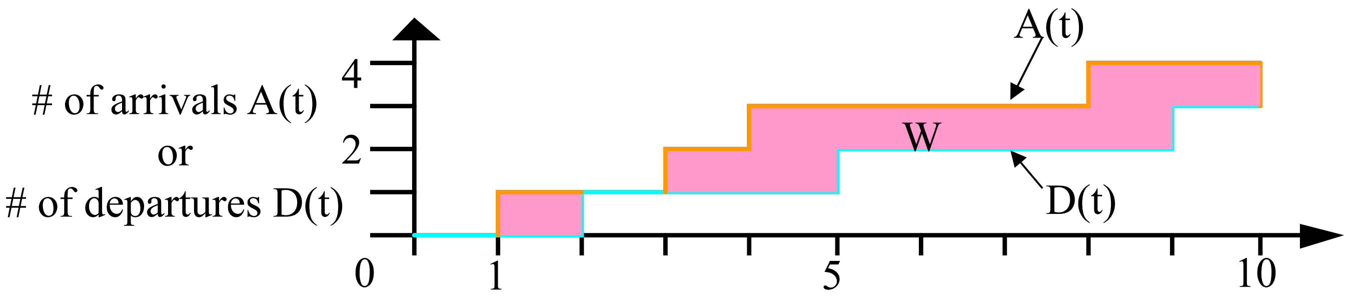
mathematically

$$\frac{B}{T} = \frac{C}{T} * \frac{B}{C} \qquad \boxed{\rho = x * D}$$

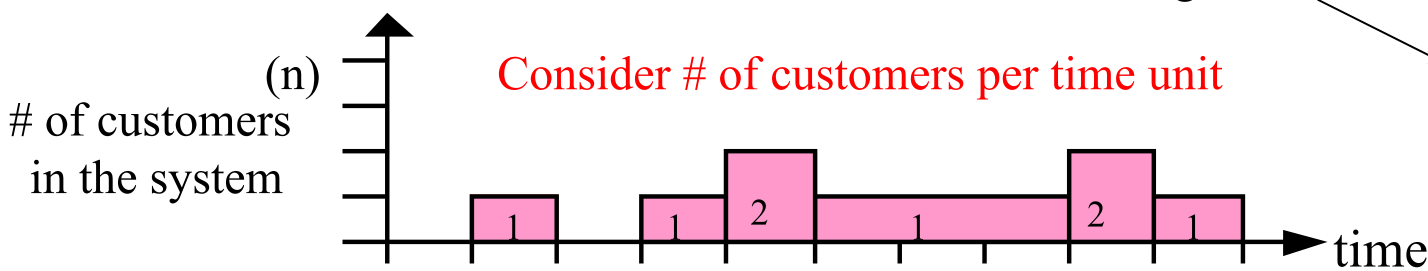
utilization law

Little's law

Consider response time per customer



A	D
1	2
3	5
4	9
8	10



avg. # of customers

the shaded region

$$\bar{n} = \frac{W}{T}$$

observation period

* A meaning of W is the total time spent by all customers in the system. $\therefore$ R = W/C

* Another meaning of W is the total population accumulated (in queue & in service) over T time units. $\therefore \bar{n} = \dfrac{W}{T}$

Algebraically $\quad \bar{n} = \frac{W}{T} = \frac{W}{C} * \overbrace{\frac{C}{T}}^{x} = R * x \qquad \boxed{\therefore \bar{n} = R\ x}$