Kendall Notation:

A/D/S/C/SD

A - arrival process
D - departure process
S - # of servers
C - buffer capacity
SD - service discipline

Little's Law:

$E(L) = \lambda E(S)$
$E(L_q) = \lambda E(W)$
$E(B) = \lambda E(\text{service time})$

M/M/1:

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

$\pi_0 = 1 - \rho$

$\bar{S} = \frac{1}{\mu(1-\rho)}$

$\bar{L_q} = \frac{\rho^2}{1-\rho}$

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

$\bar{W} = \frac{\rho}{\mu \cdot \lambda}$

$E(T) = \frac{1}{\lambda}$

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

DBE: $\pi_n \lambda = \pi_{n+1} \mu$

$P(X = i) = \frac{\lambda^i e^{-\lambda}}{i!}$

$P(T \leq t) = 1 - e^{-\lambda t}$

M/M/1/K:

$$\Pi_0 = \frac{1-\rho}{1-\rho^{K+1}}$$

$$\Pi_n = \frac{\rho^n(1-\rho)}{1-\rho^{K+1}}$$

$$P_{Loss} = \Pi_K = \frac{\rho^K(1-\rho)}{1-\rho^{K+1}}$$

Π_K is the state of packet loss.

M/M/c:

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

$$\Pi_n = \frac{(c\rho)^n}{n!}\Pi_0$$

$$\Pi_{c+n} = \rho^n \Pi_c = \rho^n \frac{(c\rho)^c}{c!}\Pi_0$$

$$\Pi_0 = \left[\sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{c!}\cdot\frac{1}{1-\rho}\right]^{-1}$$

queueing prob: prob. an incoming packet has to wait in the buffer

$$\Pi_w = \Pi_c + \Pi_{c+1} \cdots = \frac{\Pi_c}{1-\rho} = \frac{(c\rho)^c}{c!}\frac{1}{1-\rho}\Pi_0$$

mean queue length

$$E(L_q) = \sum_{n=1}^{\infty} n \cdot \pi_{c+n} = \pi_w \frac{\rho}{1-\rho}$$

mean waiting time

$$E(W) = \frac{E(L_q)}{\lambda} = \pi_w \frac{1}{1-\rho} \frac{1}{c\mu}$$

Continuous Time Markov Chain (CTMC)

Def (CTMC): A stochastic process $\{X(t): t \geq 0\}$ with discrete state space S is called CTMC if

$$Pr\{X(t)=j \mid X(h)=i, X(t_{n-1})=i_{n-1}, \ldots, X(t_1)=i_1\}$$

$$= Pr\{X(t)=j \mid X(h)=i\}$$

where $0 \leq t_1 \leq t_1 \leq \ldots \leq t_{n-1} \leq h \leq t$ and

$i_1, i_2 \ldots i_{n-1}, i, j \in S$, $\forall n \geq 1$.

Again, given the current state, the future state is independent from the past state.

Time Homogenity: says CTMC is time homogeneous if For any $h \leq t$ and any state $i, j \in S$.

$$Pr[X(t)=j \mid X(h)=i] = Pr[X(t+h)=j \mid X(0)=i] = P_{ij}(t-h)$$

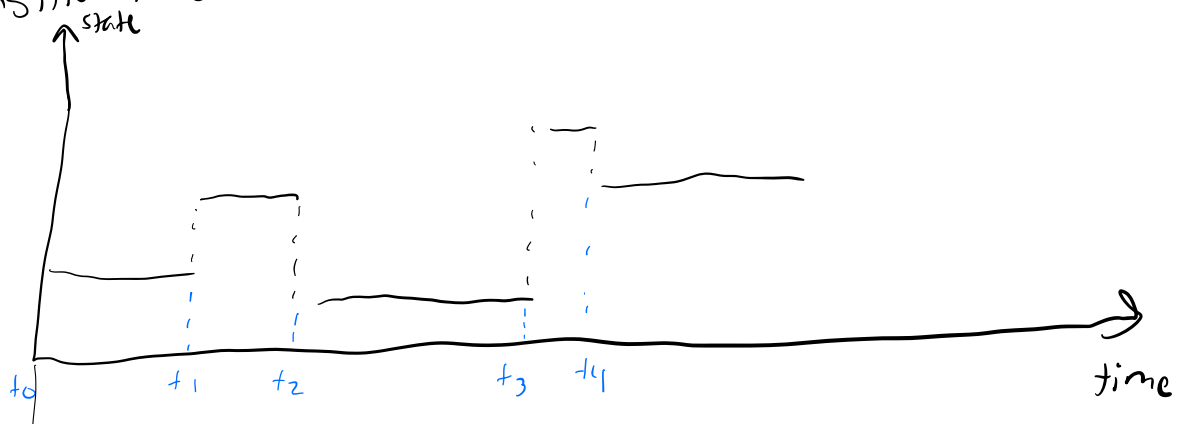
↑ transition probability

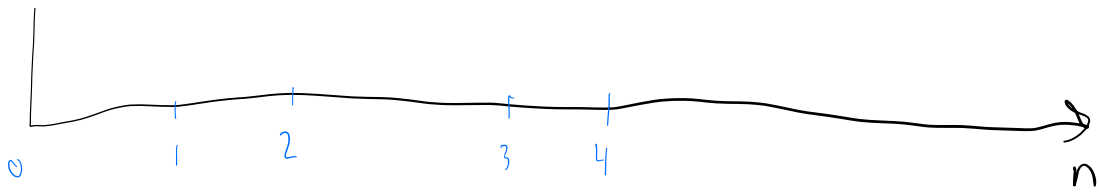
10/6/25

By time homogenity, whenever the process enters state i, the way it evolves probablisticly from that point is the same as if the process started in the state i at time 0. When the process enters state i, the time it spends there before it leaves state i is called the **holding time** in State i. The Markove property dictates that the holding time must have memoryless property and thus are exponetially distributed. Since an exponential distribution is completely determined by its rate

$\therefore$ we can conclude that for each $i \in S$, there must exist a constant rate $a_i > 0$, such that the Chain, when entering state i remains there for an amount of time $T_i \sim exp(a_i)$.

Embedded DTMC: Discretize the CTMC by the moments when the process leaves a state and the state right after the transition is defined as the state of the DTMC.





Every time the CTMC leaves state i, its transition to a state j is governed by a DTMC according to a transition probability matrix $[P_{ij}]$. ($P_{ii} = 0$)

Now, let's consider the first-order derivative of the continuous time transition probability, $P_{ij}(t)$, evaluated at $t = 0$, i.e., $P_{ij}'(t)\big|_{t=0} = P_{ij}'(0)$

i.e., the infinitesimal transition rate =

$$P_{ij}'(0) = \lim_{h \to 0} \frac{P_{ij}(h) - P_{ij}(0)}{h}$$

(1) Case 1: when $i \neq j$ $\quad P_{ij}(0) = 0$

$$P_{ij}'(0) = \lim_{h \to 0} \frac{P_{ij}(h)}{h} = \lim_{h \to 0} \frac{\Pr\{N_i(h) = 1\} \cdot P_{ij}}{h}$$

cdf of exp. dist. $\quad T_n \leq h$

$$\Pr\{N_i(h) = 1\} = \Pr\{T_i \leq h\} = 1 - e^{a_i h}$$

Do the Taylor expansion

$$\Pr\{N_i(h) = 1\} = a_i h + O(h)$$

$$P'_{ij}(0) = \lim_{h \to 0} \frac{[a_i h + O(h)] \cdot P_{ij}}{h} = a_i \cdot P_{ij} \text{ for } i \neq j.$$

only need to know result

② Case 2: when $i = j$

$$P_{ii}(0) = 1$$

$$P'_{ij}(0) = \lim_{h \to 0} \frac{P_{ii}(h) - 1}{h}$$

$$P_{ii}(h) = 1 - Pr\{Pr(x/h) \neq i \mid X(0) = i\}$$

$$= 1 - [1 - a_i h + O(h)]$$

$$= a_i h + O(h)$$

$$P'_{ii}(0) = \lim_{h \to 0} \frac{P_{ii}(h) - 1}{h}$$

10/13/25

The matrix $Q = \{P'_{ij}(0)\}$ is called the transition rate matrix, or infinitesimal generator of the CTMC.

Ex: if $S = \{0,1,2,3,4\}$, then,

$$Q = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} -a_0 & q_{01}a_0 & q_{02}a_0 & q_{03}a_0 & q_{04}a_0 \\ & -a_1 & & & \\ & & -a_2 & & \\ & & & -a_3 & \\ & & & & -a_4 \end{pmatrix} \end{array}$$

Note that each row adds to 0.

★ how to compute $P_{ij}(t)$ ⟶ Kolmogorov backward equation

THM: For CTMC with transition rate matrix $Q = P'(0)$, where $P(t) = \{P_{ij}(t)\}$ the following set of linear differential equation is satisfied by $P(t)$.

$$P'(t) = Q\,P(t), \text{ where } t \geq 0, \quad P(0) = I.$$

that is
$$P'_{ij}(t) = -a_i P_{ij}(t) + \sum_{k \neq i} q_{ik} P_{kj}(t)$$
$$= -a_i P_{ij}(t) + \sum_{k \neq i} a_i P_{ik} P_{kj}(t), \quad i,j \in S, \; t > 0$$

The solution is thus of the exponential form.

$$P(t) = e^{qt}, \quad t > 0$$

where for square matrix M.

don't have to calculate this

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

Stationary distribution: Let $\{x(t) : t \geq 0\}$ be CTMC with state space S, generator Q and transition probability matrix $P(t)$. A row vector $\vec{\pi} = (\pi_i)_{i \in S}$ with $\pi_i \geq 0$ for $\forall i \in S$ and $\sum_{i \in S} \pi_i = 1$.

is a stationary distribution of the CTMC if $\vec{\pi} = \vec{\pi} P(t)$ for all $t \geq 0$.

$\vec{\pi}$ is a stationary distribution $\iff \vec{\pi} = \vec{\pi} P(t)$, $\forall t \geq 0$

$$\iff \vec{\pi} = \vec{\pi} \sum_{n=0}^{\infty} \frac{(tQ)^n}{n!} \quad \text{for all } t \geq 0$$

$$\iff \vec{\pi} = \vec{\pi} + \vec{\pi} \sum_{n=1}^{\infty} \frac{(tQ)^n}{n!}$$

$$\iff \vec{0} = \vec{\pi} \sum_{n=1}^{\infty} \frac{(tQ)^n}{n!}$$

$$\therefore \vec{0} = \sum_{n=1}^{\infty} \frac{t^n}{n!} \vec{\pi} Q^n \quad \text{for all } t \geq 0$$

$$\therefore \vec{0} = \vec{\pi} Q^n \quad \text{for all } n \geq 1$$

$\ldots \quad \vec{0} = \vec{\pi} Q$

Expanded form: The j^{th} equation in $\vec{0} = \vec{\pi} Q$ is:

$$0 = -q_j \pi_j + \sum_{i \neq j} q_{ij} \pi_i$$

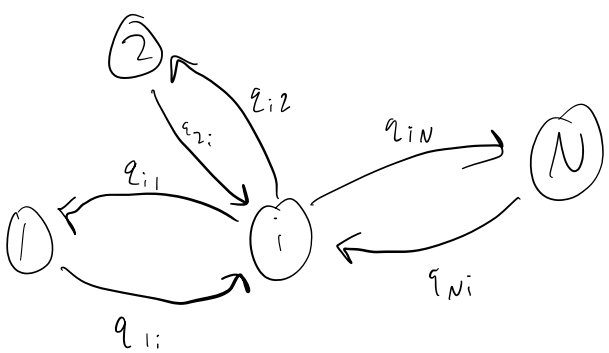
or

$$\pi_j q_j = \sum_{i \neq j} \pi_i q_{ij}$$

LHS (left): the rate leaving state j weighted by the probably of state j.

RHS: q_{ij} is the transition rate going into state j from i weighted by the stationary prob. of state i (π_i).

This is called global balance equation. The long term rate out of state j has to be equivalent of the long term rate going into state j.



at stationary $\pi_i (q_{i1} + q_{i2} + \ldots q_{iN}) = \pi q_{1i} + \pi q_{2i} + \pi q_{Ni}$

leaving entering

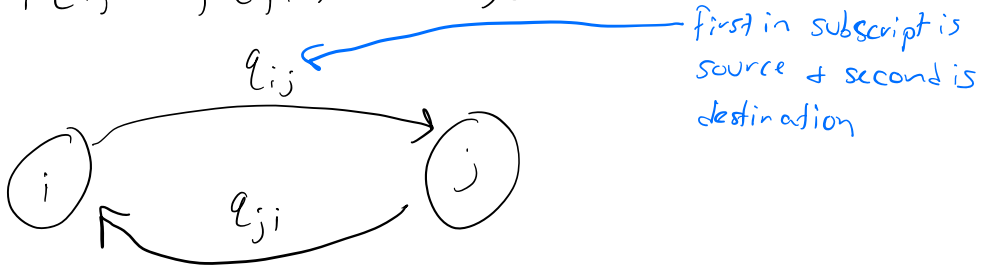
Normalization condition $\sum_{i \in S} \pi_i = 1$.

10/15/25

Detailed Balance Equation:

For a CTMC, if the solution $\vec{\pi}$ to the following equation set exists

$$\pi_i q_{ij} = \pi_j q_{ji}, \text{ for all } i, j \in S$$



Then $\vec{\pi}$ is the stationary distribution of the states.

The above equation set is called detailed balance equation.

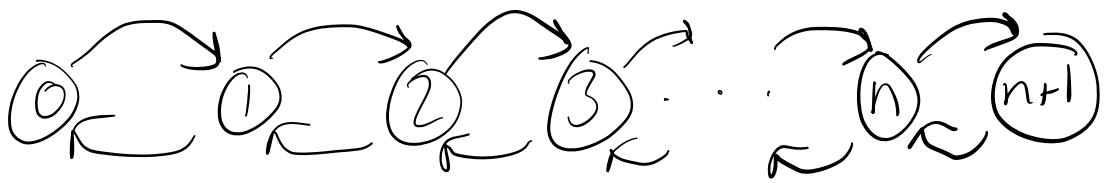
It can be shown that the solution to the detailed balance equation must be the solution to the global balance equation, but not vice versa, which means that a general CTMC should always satisfy the global balance equation but may not satisfy the detailed balance equation.

Detailed is sufficient but not necessary.
Always there are Global.

It can be shown that in birth-death process satisfies the DBE.

Birth-death process : A markov chain with transition only between adjacent/neighboring states

$q_{ij} = 0$ if $|i-j| > 1$



In this case instead of looking at each state (node), we look at the edges and apply the DBE.

$$\pi_n q_{n\,n+1} = \pi_{n+1} q_{n+1\,n}$$

DBE is much easier to solve than GBE.

Basic Queueing Theory

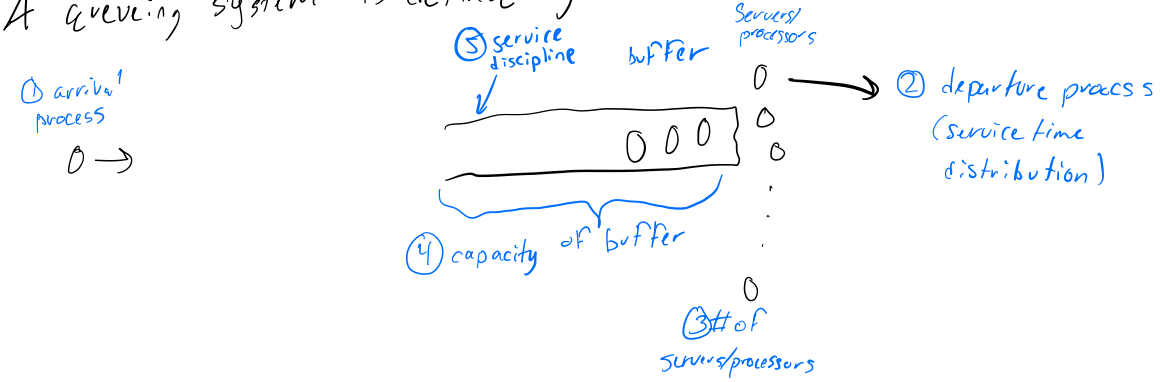
* Kendall notation

* Basic types of queues

* How to compute important performance matrix

* Queueing networks

A queueing system is defined by 5 elements.



Kendall notation: A/D/S/C/SD

A: arrival process (Poisson)

D: departure process (service time distribution)

S: # of servers

C: buffer capacity

SD: service discipline (FIFO, FILO)

M/M/1/∞

If no SD then it is FIFO

Also don't have to put ∞

More concise: M/M/1

S → M = Markovian → Poisson with rate, λ

D → M = Markovian → Exponential service time with rate μ

# of server = 1

queue capacity = ∞

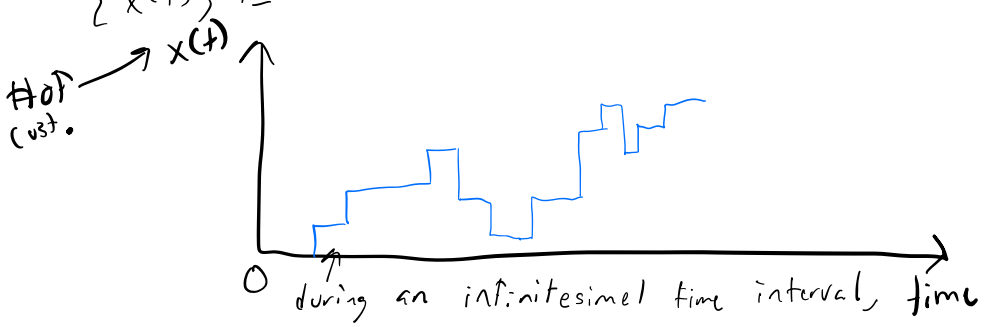
SD = FIFO (first in first out)

10/17/25

M/M/1

Define the # of customers (packets/jobs) in the system (including the customers that are waiting in the queue and the one being serviced) as the state of the stochastic process. Because the customer interarrival times and the service times are all independent exponentially distributed r.v.s, the process

$\{X(t), t \geq 0\}$ is a CTMC with state space $S = \{0, 1, 2 \cdots \infty\}$



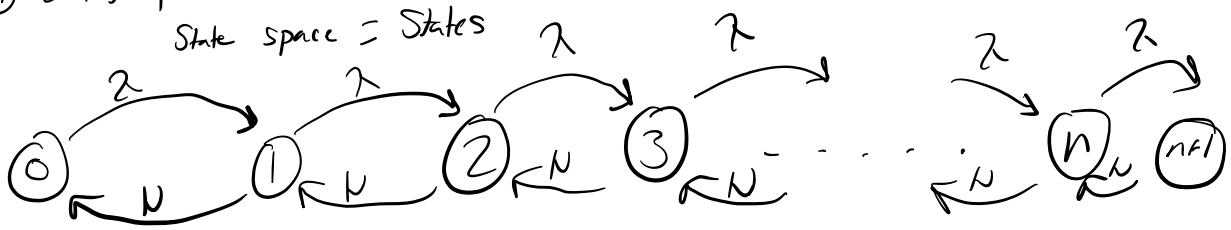
only 1 event (arrival or departure) can happen, multiple event cannot happen at the same time.

To solve for the stationary distribution for $X(t)$

① Let's plot the state transition diagram.

this represents # of customers

State space = States



Cannot have more than 1 event happening at 1 time

λ = arrival rate

μ = departure rate

$Pr\{$ a new arrival within $h\}$

$= Pr\{\tau < h\}$

$= \lambda h + O(h)$

$$\lim_{h \to 0} \frac{\lambda h + O(h)}{h} = \lambda$$

② Apply the GBE

$$\pi_0 \cdot \lambda = \pi_1 \cdot \mu \Rightarrow \pi_1 = \frac{\lambda}{\mu}\pi_0$$

cancel each other

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

$\pi_2 = \frac{\lambda}{\mu} \pi_1$

$\pi_2 = \left(\frac{\lambda}{\mu}\right)^2 \pi_0$

$\vdots$

$\pi_n (\lambda+\mu) = \pi_{n-1} \lambda + \pi_{n+1} \mu$

$\pi_3 = \left(\frac{\lambda}{\mu}\right)^3 \pi_0$

$\pi_n = \left(\frac{\lambda}{\mu}\right)^n \pi_0$

Let $\rho \triangleq \frac{\lambda}{\mu}$ we have $\pi_n = \rho^n \pi_0 \quad \ldots n = 1, 2 \ldots$

$0 \le \rho \le 1$

$\pi_0 + \pi_1 \ldots = 1$

$\pi_0 (1 + \rho + \rho^2 + \ldots) = 1$

$\pi_0 \frac{1}{1-\rho} = 1 \Rightarrow \pi_0 = 1 - \rho.$

Alternatively, apply DBE:

$\pi_0 \lambda = \pi_1 \mu \Rightarrow \pi_1 = \frac{\lambda}{\mu} \pi_0$

$\pi_1 \lambda = \pi_2 \mu \Rightarrow \pi_2 = \frac{\lambda}{\mu} \pi_1 = \left(\frac{\lambda}{\mu}\right)^2 \pi_0$

$\pi_2 \lambda = \pi_3 \mu \Rightarrow \pi_3 = \left(\frac{\lambda}{\mu}\right)^3 \pi_0$

$\pi_{n-1} \lambda = \pi_n \mu \Rightarrow \pi_n = \left(\frac{\lambda}{\mu}\right)^n \pi_0$

It's clear that the stationary distribution exists if and only if $\rho < 1$.

ρ is also called the load of the system (or occupation rate or service utilization).

Let's calculate a set of important performance matrix:

① Average # of customers in the system $\quad n \cdot \rho^{n-1} = (\rho^n)'$

$$E(L) = \sum_{n=0}^{\infty} n\pi_n = \sum_{n=0}^{\infty} n(1-\rho)\cdot\rho^n = \frac{\rho}{1-\rho}$$

② Whats the prob. that the server is idle?

$\pi_0 = 1-\rho$ so server utilization $= 1-\pi_0 = \rho$.

10/20/25

arrivale rate - λ

service rate - μ

$\pi_0 = 1-\rho \quad \pi_n = \rho^n \pi_0 = \rho^n(1-\rho)$.

average L $\quad E(L) = \frac{\rho}{1-\rho}$

average number of customers being serviced: B

$$E(B) = 0\cdot\pi_0 + 1\cdot(1-\pi_0) = \rho \qquad E(B) = \rho$$

Queue length: L^q

$$L = L^q + B \Rightarrow E(L) = E(L^q) + E(B)$$

$$\Rightarrow E(L^q) = E(L) - E(B) = \frac{\rho^2}{1-\rho}$$

Waiting time: time spent in the queue → from the moment of arrival (entering the buffer) to the moment of leaving the queue to get service. (W).

Sojourn time: the waiting time plus service time → from the moment of arrival to the moment of leaving the system (departure). (S)

(Also called delay, packet gets to router to leaving the router)

Little's Law: specifies the relationship between $E(L)$ and $E(S)$.

$$E(L) = \lambda E(S).$$

The above applies to any arrival and / departure process and any service order and even subqueues.

$$E(L^q) = \lambda E(W)$$

$$E(B) = \lambda E(\text{service time}) = E(B) = \lambda \frac{1}{\mu} = \rho$$

So for M/M/1 queue:

$$E(S) = \frac{E(L)}{\lambda} = \frac{1}{\mu(1-\rho)}.$$

PASTA property: For queueing system with Poisson arrivals (M/./. System), the special property holds that arriving customers finds on average the same situation in the queueing system as an outside observer looking at the system at an arbitrary point in time.

Priorities: Consider a M/M/1 queueing system servicing 2 types of user (packets) type 1 & type 2. Type 1 & Type 2 customers arrival according to independent Poisson processes with rates λ_1 & λ_2 respectively. The service times of all customers are iid exponential with mean $\frac{1}{\mu}$. We assume that $\rho_1 + \rho_2 < 1$, where $\rho_i := \frac{\lambda_i}{\mu}$. Type 1 has higher priority than Type 2.

1) Preemptive Priority:

Def. In preemptive-resume rule, when type 2 job is in service, and a type 1 arrives, the type 2 service is interrupted (suspended) the server proceeds with the type 1 job. Once there is no more type 1 jobs in the system, the server resumes the service of the type 2 job at the point when its interrupted.

Let L_i denote the number of type i jobs in the system and S_i denote the sojourn time of a type i job. Please calculate the expectation of $E(L_i)$ and $E(S_i)$ for $i = 1, 2$.

10/22/25

M/M/1 - queue 2 types of jobs

Type 1 > Type 2

Poisson rates λ_1, λ_2

Service time μ

① Preemptive Priority

L_i , S_i , $E(L_i)$, $E(S_i)$

For type 1, the type 2 jobs do not exist. So we have:

$$E(L_1) = \frac{\rho_1}{1-\rho_1} \qquad E(S_1) = \frac{1}{\mu(1-\rho_1)}$$

↑ Average delay for type 1 job

Since the service time is exponential, the total # of jobs in the system does not depend on the service order. So

$\rho_i = \frac{\lambda_i}{\mu}$

$$E(L_1+L_2) = \frac{\rho_1+\rho_2}{1-\rho_1-\rho_2}$$

$$E(L_1+L_2) = E(L_1)+E(L_2) \Rightarrow E(L_2) = E(L_1+L_2) - E(L_1)$$

$$= \frac{\rho_1+\rho_2}{1-\rho_1-\rho_2} - \frac{\rho_1}{1-\rho_1}$$

$$= \frac{\rho_2}{(1-\rho_1)(1-\rho_1-\rho_2)}$$

↑ average # of type 2

Applying Little's law on type 2 traffic:

$$E(L_2) = \lambda_2 \cdot E(S_2) \Rightarrow E(S_2) = \frac{E(L_2)}{\lambda_2} = \frac{1/\mu}{(1-\rho_1)(1-\rho_1-\rho_2)}$$

② Non preemptive priority

Type 1 job is not allowed to interrupt the service of type 2 job.

there is 1 $\quad$ there is 0

$$p_2 \frac{1}{\mu} \quad + \quad (1-p_2)\, 0$$

$$E(S_1) = E(L_1)\cdot\frac{1}{\mu} + \frac{1}{\mu} + p_2\frac{1}{\mu}$$

- $E(L_1)\cdot\frac{1}{\mu}$: service for those type-1 jobs in front of me
- $\frac{1}{\mu}$: service time for myself
- $p_2\frac{1}{\mu}$: average service time for the prob. type 2 job

By Little's Law: $E(L_1) = \lambda_1 E(S_1)$

$$E(L_1) = \frac{(1+p_2)p_1}{1-p_1} \qquad E(S_1) = \frac{1+p_2}{\mu(1-p_1)}$$

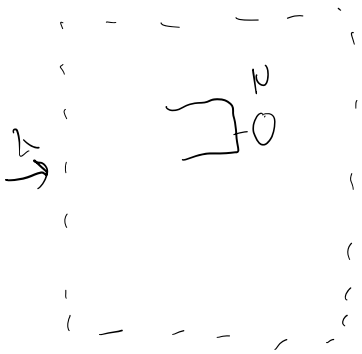
$$E(L_2) = E(L_1+L_2) - E(L_1)$$

$$= \frac{p_1+p_2}{1-p_1-p_2} - \frac{(1+p_2)p_1}{1-p_1} = \frac{[1-p_1(1-p_1-p_2)]p_2}{(1-p_1)(1-p_1-p_2)}$$

$$E(S_2) = \frac{E(L_2)}{\lambda_2} = \frac{[1-p_1(1-p_1-p_2)]}{\mu(1-p_1)(1-p_1-p_2)}$$

Let's compare the following 2 architecture:

A_1

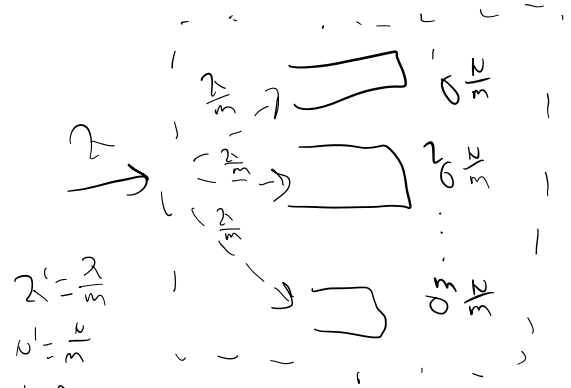


$$\bar{L} = \frac{\rho}{1-\rho} = \frac{\lambda}{N-\lambda}$$

$$\bar{S} = \frac{\bar{L}}{\lambda} = \frac{1}{N-\lambda}$$

$$\bar{W} = \frac{\rho}{N-\lambda}$$

A_2



$\lambda' = \frac{\lambda}{m}$

$N' = \frac{N}{m}$

$\rho' = \rho$

$$\bar{L}' = \frac{\rho'}{1-\rho'} = \bar{L}$$

$$\bar{S}' = \frac{\bar{L}'}{\lambda'} = \frac{\bar{L}}{\frac{\lambda}{m}} = m\bar{S}$$

$$\bar{W}' = \frac{\rho'}{N'-\lambda'} = m\bar{W}$$

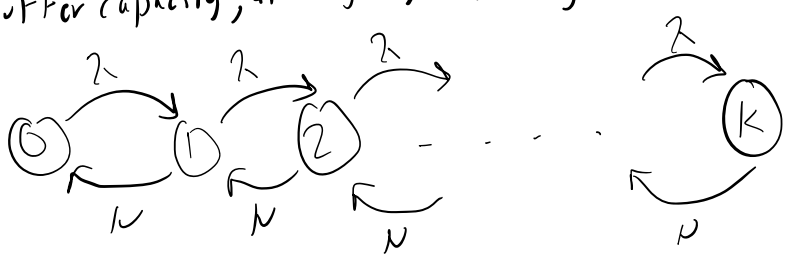
Key observation: A_2 slows down the arrival & service rate by a factor of m ⇒ same queue length, but delay is m fold longer.

10/24/25

* M/M/1/K queue (at most K packets in the system)

K - buffer capacity, # of jobs, capacity

birth/death b/c M/M/1



λ - arrival rate

μ - service rate

Exactly the same as in M/M/1 queue, but state space is truncated!

$\pi_n = \rho^n \pi_0, \quad n = 1, 2, \dots K$ Finite # of π_n

normalization condition: $\sum_{n=0}^{K} \pi_n = 1 \Rightarrow \pi_0 = \frac{1-\rho}{1-\rho^{K+1}}$

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

Now there will be packet loss b/c overflow can happen

packet loss rate / buffer overflow rate $P_{loss} = \pi_K = \frac{\rho^K (1-\rho)}{1-\rho^{K+1}}$

$E = \lambda S$ — M/M/1

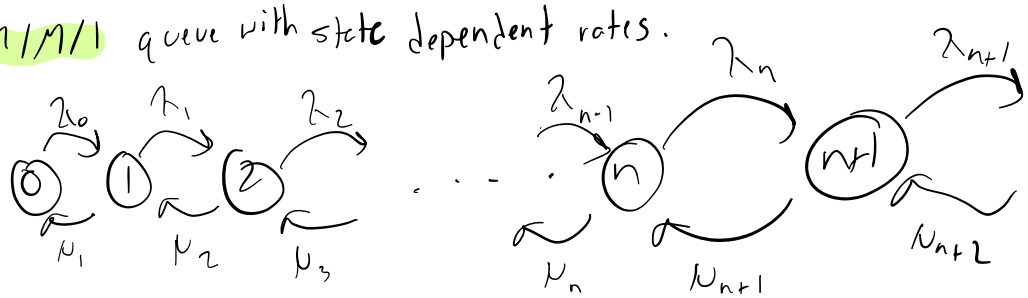
↑ average # of customers in system

$E = \lambda_{eff} S$ — M/M/1/K

$\lambda_{eff} = (1 - P_{loss}) \lambda$

↑ total arrival

* M/M/1 queue with state dependent rates.



the rates now depend on the underlining state of the system

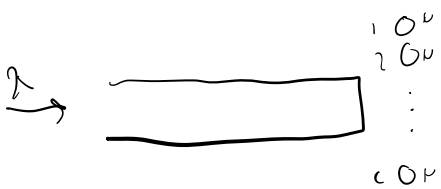
still birth/death process, so we use DBE:

$$\pi_n = \pi_0 \prod_{i=0}^{n-1} \frac{\lambda_i}{\mu_{i+1}}, \quad n \geq 1$$

$$\pi_0 = \left[1 + \sum_{n=1}^{\infty} \prod_{i=0}^{n-1} \frac{\lambda_i}{\mu_{i+1}}\right]^{-1}.$$

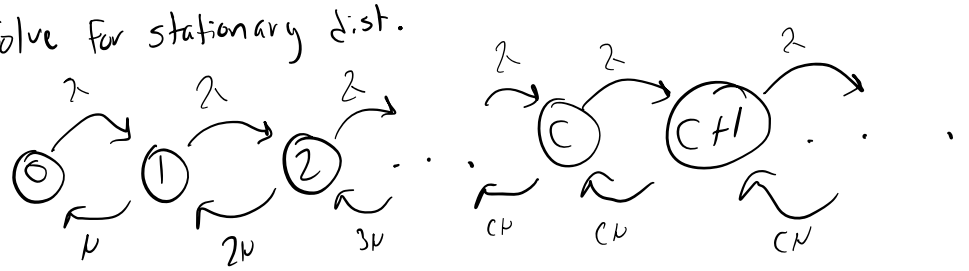
M/M/c queue

exponential interarrival times with rate λ, exponential service times with rate μ for each server, and there are c parallel identical servers.



The total occupation rate/load $\rho = \frac{\lambda}{c\mu} < 1$.

Solve for stationary dist.



state transition diagram

We then get

$$\pi_n = \frac{(cp)^n}{n!}\pi_0, \quad n = 0, 1 \ldots c$$

and

$$\pi_{c+n} = \rho^n \pi_c = \rho^n \frac{c^c p^c}{c!}\pi_0, \quad n = 0, 1 \ldots \infty$$

π_0 is calculated using the normalization condition

$$\pi_0 = \left[\sum_{n=0}^{c-1} \frac{(cp)^n}{n!} + \frac{(cp)^c}{c!} \cdot \frac{1}{1-\rho} \right]^{-1}$$

Queueing probability (the probability than an incoming packet/job/customer has to wait in the buffer).

$$\pi_W = \pi_c + \pi_{c+1} \cdots = \frac{\pi_c}{1-\rho} = \frac{(cp)^c}{c!}\frac{1}{1-\rho}\pi_0$$

mean queue length:

$$E(L^q) = \sum_{n=1}^{\infty} n \cdot \pi_{c+n} = \pi_W \frac{\rho}{1-\rho}$$

$n=1$: ignore $n=0$

start at $c+1$ b/c now have backlogs

mean waiting time:

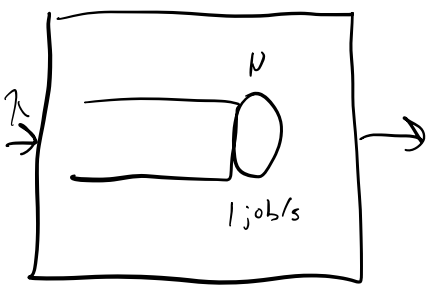
$$E(U) = \frac{E(L^q)}{\lambda} = \pi_W \frac{1}{1-\rho}\frac{1}{c\mu}$$

10/27/25

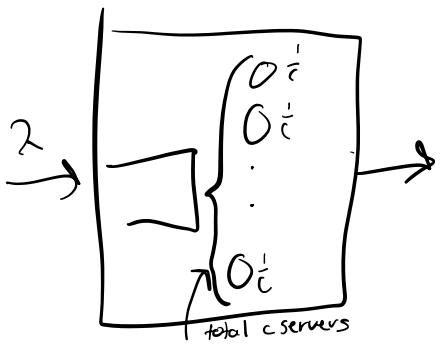
Mid 2 11/4

M/M/c

A1



A3



① Let's fix the incoming traffic rate $\lambda = 0.9$ arrivals per second. and the load $\rho = 0.9$ (ratio arrival/service, no rate). We will change the service rate of the servers in A3 with c. i.e., using less # of higher performance servers vs. using more # of lower performance servers ⟷ you have a choice of the type of servers.

# of servers — avg. waiting time — (Little's law) — avg queue length — avg. # of customers in the system — avg. sojourn time, total delay

c	E(W)	$E(L^q)$	E(L)	E(S)	
1	9	8.1	9	10	choose this
2	8.53	7.67	9.47	10.52	
5	7.63	6.86	11.36	12.63	
10	6.69	6.02	15.02	16.69	
	decreasing	decreasing	increasing	increasing	

Clearly, putting more under performing servers is not desirable as the total delay will be higher. We prefer higher performance server.

② Let's fix μ of each server. Say $\mu = 1$. We fix $\rho = 0.9$. (same kind of server and the same traffic load). Means we change λ with c.

prob of queueing

most important factor

c	Π_w	E(W)	$E(L^q)$	E(L)	E(S)
1	0.9	9	8.1	9	10
2	0.85	4.26	7.67	9.47	5.26
5	0.76	1.83	6.86	11.36	2.52
10	0.67	0.67	6.02	15.02	1.67
20	0.55	0.28	4.96	22.96	1.24
	decreases	decreases	decreases	increases	decreases, smaller delay

Would rather have bigger team

Clearly, putting more servers leads to smaller queueing delay and overall delay. We prefer more servers.

Overall, the higher the performance of the server the better. The more servers you can have (when you cannot choose the type) the better.

Erlong-B formula:

Consider a circuit switching telephone network e.g. satellite phone network. There are m circuit (lines) provided by the satellite. So at most m calls can be supported by the satellite at the same

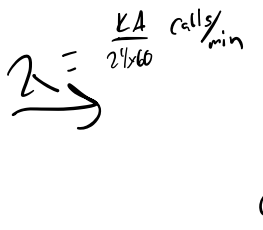
time. When these m lines are all being occupied, a dialog caller will be blocked, i.e., the caller will have a busy time when dials. What's the blocking probability of a user, suppose in total in the service area we have K users each user on average makes A calls a day and each call on average lasts B minutes?

0 buffer

circuit is servers (n)

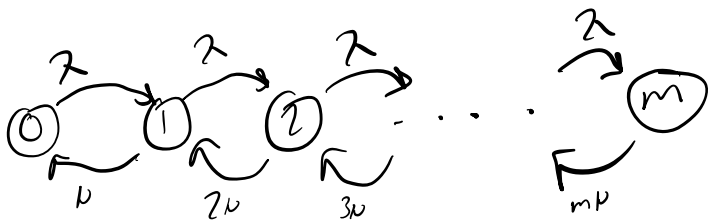
M/M/m/0, the 0 is representing the buffer size

$\lambda = \frac{KA}{24 \times 60}$ calls/min $\quad \mu = \frac{1}{B}$ calls/min



DBE

birth/death process



$$\pi_n = \frac{\pi_0 \left(\frac{\lambda}{\mu}\right)^n}{n!} \quad \text{for } n = 0, 1, \ldots m$$

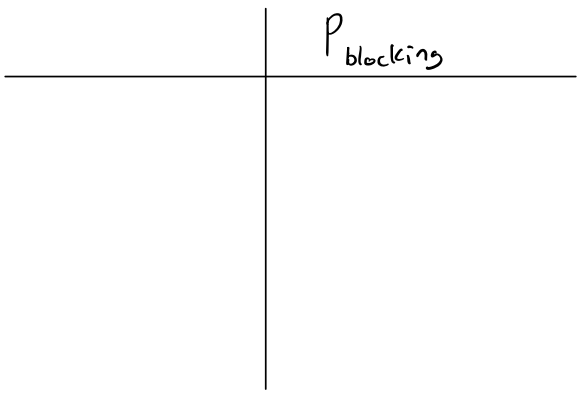
$$\pi_0 = \left[\sum_{n=0}^{m} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} \right]^{-1}$$

$$P_{blocking} = \pi_m = \frac{\left(\frac{\lambda}{\mu}\right)^n / n!}{\sum_{n=0}^{m} \left(\frac{\lambda}{\mu}\right)^n / n!} \leftarrow \text{Erlong-B}$$

Ex. $\lambda = 4$ calls/min

$\mu = \frac{1}{3} \Rightarrow \frac{\lambda}{\mu} = 12$

Finish notes



10/29/25

Midterm 2: 11/10

M/Er/1

arrival - Poisson

service time - Erlang distributed

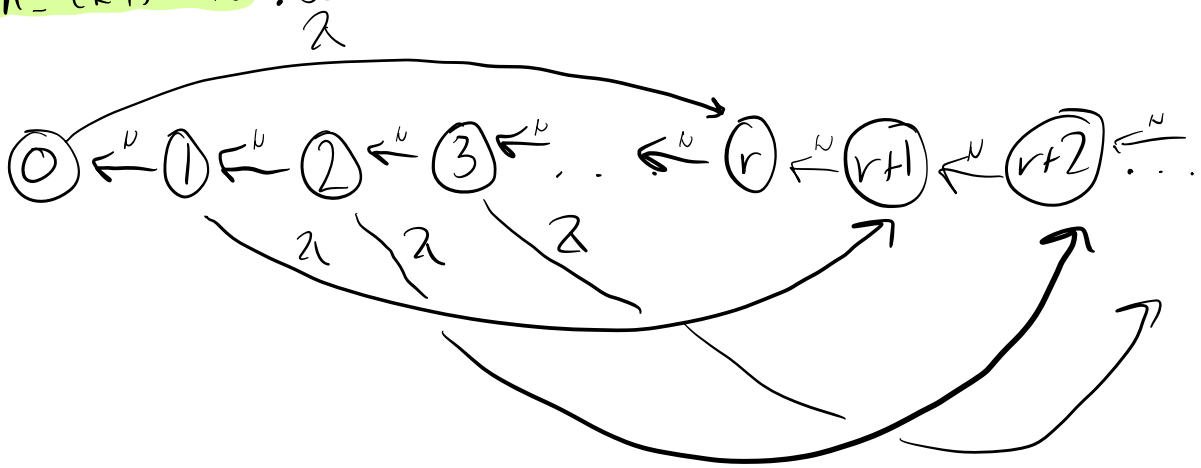
To service a job/packet, the server has to go through r stages sequentially, where each stage takes an exponentially distributed time ($\frac{1}{\mu}$). The job arrival process is Poisson with rate λ. The load of the system $\rho = \frac{r\lambda}{\mu}$.

$\frac{r}{\mu} \rightarrow$ service rate $= \frac{\mu}{r}$

$\rho = \frac{\text{arrival rate}}{\text{service rate}}$

Definition of the state: the natural way to describe the state of nonempty system is by the pair (K, ℓ), where K denotes the # of jobs in the system and ℓ denotes the remaining # of stages of the job that is in service. This is a 2-dimensional MC. An alternative 1-D way is to describing the state by counting the total # of uncompleted stages that the jobs in the system still need to finish. There is a one-to-one mapping between this number (say n) and the pair (K, ℓ).

$n = (K-1) \cdot r + \ell$. So we can use the following 1-D model.



Let π_n denote the stationary prob. of n stages in the system. Based on GBE, the flow going out of state i is equivalent to total flows going into that state i.

$$\pi_0 \lambda = \pi_1 \mu$$

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

$$\pi_n (\lambda + \mu) = \pi_{n+1} \mu \quad \text{for } n = 1, \ldots, r-1.$$

$$\pi_n (\lambda + \mu) = \pi_{n-r} \lambda + \pi_{n+1} \mu \quad \text{for } n = r, r+1 \ldots$$

Below is not tested on

If can be shown that

$$\pi_i = \sum_{k=1}^{r} c_k x_k^i, \text{ where } i = 0, 1, 2 \ldots$$

where $x_1 \ldots x_r$ where the r are the distinct roots to the following equation:

$$(\lambda + \mu) \cdot x^r = \lambda + \mu x^{r+1} \quad \text{if } |x| < 1.$$

the coefficients

$$c_k = \frac{1 - \rho}{\prod_{j \neq k} \left(1 - \frac{x_j}{x_k}\right)}, \quad k = 1, \ldots r$$

Distribution of # of jobs: K $\qquad q_k \triangleq Pr\{k \text{ jobs}\}$

$$q_0 = \pi_0$$
$$q_1 = \pi_1 + \ldots + \pi_r$$
$$\vdots \qquad \vdots$$
$$q_i = \sum_{n=(i-1)r+1}^{ir} \pi_n .$$

M/G/1 queue: the jobs arrive according to Poisson process with rate λ, the service time is iid with CDF $F_B(T)$ and pdf $f_B(\tau)$. To be stable, we require $\rho = \lambda E(B) < 1$.

10/31/25

Generating function:

Let the probability mass function of a discrete nonnegative r.v.

X be $P_r(X=n) = p(n)$. where $n = 0, 1, 2, \ldots$. Then the moment generating function of $P_x(z)$ of

X is defined as:

$$P_x(z) = E(z^x) = \sum_{n=0}^{\infty} z^n p(n).$$

(Recap : z- transform of $P(n)$ is $P(z) = \sum_{n=0}^{\infty} p(n) z^{-n}$)

Properties: (a) $|P_x(z)| \leq 1$ for all $|z| \leq 1$.

(b) $P_x(0) = p(0)$, $P_x(1) = 1$.

(c) $P_x'(z)\big|_{z=1} = E(x)$.

$$P_x'(z) = \sum_{n=0}^{\infty} n z^{n-1} p(n)$$

$$\text{So } P_x'(1) = \sum_{n=0}^{\infty} n p(n) = E(x)$$

$$P_x^{(2)}(z) = \sum_{n=0}^{\infty} n(n-1) z^{n-2} p(n)$$

$$P_x^{(2)}(1) = \sum_{n=0}^{\infty} n(n-1) p(n) = E(x(x-1)) = Ex^2 - Ex$$

So in general: $P_x^k(1) = E[x(x-1)\cdots(x-k+1)]$

M/G/1 queue The service time is i i d with CDF $F_B(\tau)$, and pdf $f_B(\tau)$.

The moment generating function of the number of jobs/products in the system:

$$P_L(z) = \frac{(1-\rho)\tilde{B}(\lambda-\lambda z)(1-z)}{\tilde{B}(\lambda-\lambda z)-z} \quad \text{where } \rho = \lambda E(B)$$

$\tilde{B}(s)$ is the Laplace transform of the service time distribution.

$$\tilde{B}(s) = \int_0^\infty f_B(T)\, e^{-sT}\, dT.$$

Distribution of the sojourn time:

$$\tilde{S}(s) = \frac{(1-\rho)\,\tilde{B}(s)\cdot s}{\lambda\tilde{B}(s)+s-\lambda}$$

Distribution of the waiting time:

$$\tilde{W}(s) = \frac{(1-\rho)\cdot s}{\lambda\tilde{B}(s)+s-\lambda}$$

All above are the Pollaizek-Khinchin formula

① mean waiting time in the buffer.

$$\bar{W} = \frac{\lambda E(x^2)}{2(1-\rho)} \qquad Var(X) = Ex^2 - (Ex)^2$$
$$\Rightarrow Ex^2 = Var(X) + (\bar{E}x)^2$$

$$= \frac{\rho + \lambda N\, Var(X)}{2(N-\lambda)}$$

② mean sojourn time (service time)

$$\bar{S} = \bar{W} + E(x) = \frac{1}{N} + \frac{\rho + \lambda N\, Var(x)}{2(N-\lambda)}$$

(3) mean queue length

$$\bar{L}_q = \lambda \bar{W} = \frac{\lambda^2 \overline{x^2}}{2(1-\rho)}$$

$$= \frac{\rho^2 + \lambda^2 Var(x)}{2(1-\rho)}$$

(4) mean number of packets in the system

$$\bar{L} = \rho + \bar{L}_q = \rho + \frac{\rho^2 + \lambda^2 Var(x)}{2(1-\rho)}$$

Example: M/D/1 queueing system: Poisson arrival with rate λ, deterministic service time $\frac{1}{\mu}$ (↑ constant)

$$\bar{L} = \rho + \frac{1}{2}\frac{\rho^2}{1-\rho}$$

$$\bar{L}_q = \frac{1}{2}\frac{\rho^2}{1-\rho}$$

M/M/1 queue same λ, μ

$$\bar{L}_q = \frac{\rho^2}{1-\rho}$$

Not required

$$\bar{S} = \frac{1}{\mu} + \frac{\rho}{2\mu(1-\rho)}$$

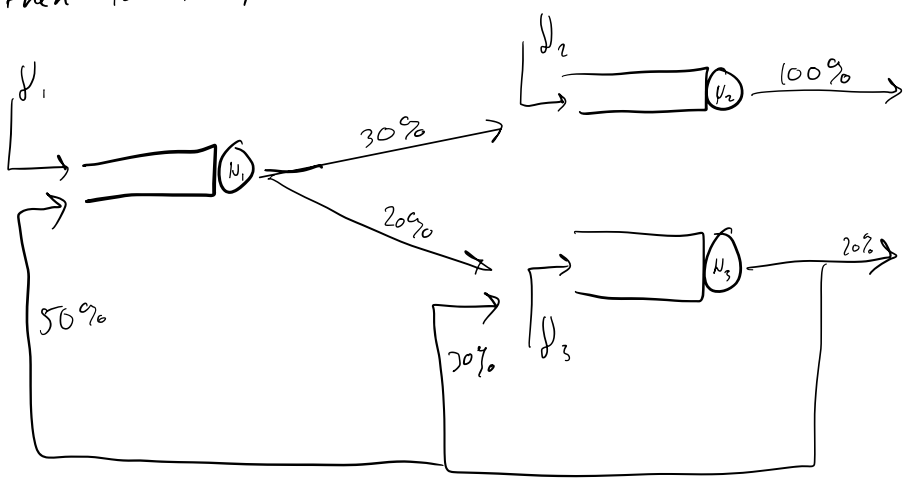
$$\bar{W} = \frac{\rho}{2\mu(1-\rho)} \leftarrow \text{half of } \bar{W} \text{ for M/M/1}$$

Queueing networks: a set of interconnected queues.

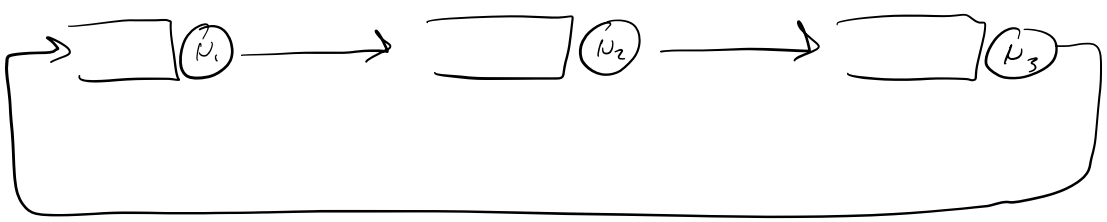
Types of queueing networks

- Open networks – external traffic
- Closed networks
- Hybrid networks

open networks: traff comes from outside the system, and are serviced and then leave/depart from the network.

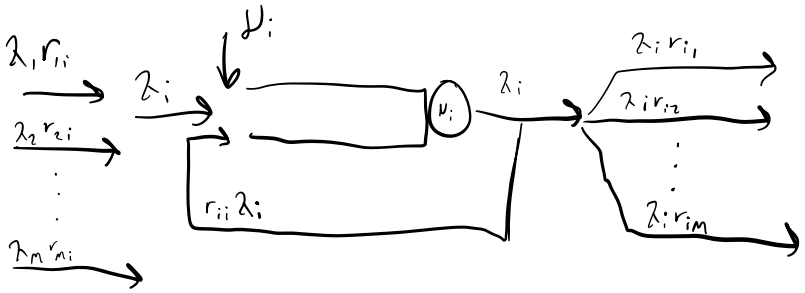


closed networks: a fixed number of customers (k) are trapped in the system and circulate among the queues.



11/3/25 Only worry about open networks

Consider an open network; assume an arbitrary network of M queues with infinite buffer capacity. Jobs arriving from outside the system are serviced and then depart. Note that a job may visit several queues before departing, including possibly visiting some queues more than once. Service time of queue is generally distributed with rate μ_i for queue i. Arrivals from outside to queue i occur according to general iid process with mean rate γ_i. the total mean job arrival rate to queue i is denoted by λ_i.



r_{ij} – routing probability, fraction of packets, that a job completing service at queue i goes to queue j.

r_{iM+1} – routing probability that a job completing service at queue i leaves the network (goes to outside).

traffic conservation condition: $\sum_{j=1}^{M+1} r_{ij} = 1$.

Let λ_i be the total mean arrival rate of queue i.

$$\lambda_i = \gamma_i + \sum_{j=1}^{M} \lambda_j r_{ji}$$

Denote $\vec{\lambda} \triangleq [\lambda_1 \ \lambda_2 \dots \lambda_M]$

$\vec{\gamma} \triangleq [\gamma_1, \gamma_2 \dots \gamma_M]$

routing matrix

$\vec{R} \triangleq [r_{ij}]$ $(1 \le i \le M, \ 1 \le j \le M)$

The flow conservation equation can be written in matrix form

$$\vec{\lambda} = \vec{\nu} + \vec{\lambda}\vec{R} \Rightarrow \vec{\lambda}[I - \vec{R}] = \vec{\nu}$$

$$\vec{\lambda} = \vec{\nu}[I - \vec{R}]^{-1}$$

(I: identity matrix; $\vec{R}$: routing matrix; $\vec{\nu}$: external traffic)

It relates external arrival rates and routing to determine the total flow rate at each queue.

A special type of open network is the **Jackson network**.

- This is a network of M queues
- There is only one class of jobs in the network
- All service times are exponentially distributed with service rate μ_i at queue i. The service discipline at all queues is FCFS (First come first serve). All external arrival processes are Poisson with rate ν_i at queue i.

Because the merge of independent Poisson processes is Poisson with rate equal to the sum of the individual rates. The departure process of an $M/M/1$ queue is also Poisson with rate equal to the input rate of the queue (i.e., λ_i), and probabilistic splitting of Poisson process results in Poisson processes for each of the sub flows. Therefore, the input and output processes of each queue i in the network is a Poisson process.

Solution for Stationary Distribution: Let $\tilde{n}_i(t)$ be the number of jobs at queue i at time t. The state of the network is defined by the vector $(\tilde{n}_1(t), \tilde{n}_2(t), \ldots \tilde{n}_m(t))$. The $\{(\tilde{n}_1(t), \tilde{n}_2(t), \ldots \tilde{n}_m(t)), t \geq 1\}$.

is a M-dimensional Markov process.

The stationary distribution, $\vec{n} = (n_1, n_2 \dots n_m)$

$$\Pi(\vec{n}) = \lim_{t \to \infty} \text{Prob}\{\tilde{n}_1(t) = n_1, \tilde{n}_2(t) = n_2 \dots \tilde{n}_m(t) = n_m\}$$

It can be shown that the stationary state probability has the following form:

$$\Pi(\vec{n}) = C \rho_1^{n_1} \rho_2^{n_2} \dots \rho_m^{n_m} = C \prod_{i=1}^{m} \rho_i^{n_i} \quad \text{where } \rho_i = \frac{\lambda_i}{\mu_i}$$

and C is some constant

normalization condition: $\sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \dots \sum_{n_m=0}^{\infty} C \rho_1^{n_1} \rho_2^{n_2} \dots \rho_m^{n_m} = 1$

$$\Rightarrow C = \prod_{i=1}^{M} (1-\rho_i)$$

Therefore, your $\Pi(\vec{n}) = \prod_{i=1}^{M} (1-\rho_i) \rho_i^{n_i}$ ← Stationary distribution for M/M/1 queue.

recall that for an M/M/1 queue $\Pi_n = (1-\rho)\rho^n$

$\therefore \Pi(\vec{n}) = \prod_{i=1}^{M} \Pi_{n_i}$ ← the product of the stationary probability of M independent M/M/1 queues.

11/5/25

Jackson's Theory: If in an open Jackson network $\lambda_i < \mu_i$ holds for all the queue $i = 1 \dots M$, then,

(1) The arrival rates λ_i can be computed by $\vec{\lambda} = \vec{\nu}(I - \vec{R})^{-1}$

(2) The stationary probability of the network can be expressed as the product of the stationary probability of the individual queues.

$$\Pi(k_1, k_2 \dots, k_M) = \Pi_1(k_1) \Pi_2(k_2) \dots \Pi_M(k_M)$$

(3) The nodes of the network can be considered as independent M/M/1 queues with arrival rate (aggregate) λ_i and service rate μ_i, respectively.

Performance matrix: Because each queue i in M/M/1 with ρ_i

$$\overline{L_i} = \frac{\rho_i}{1-\rho_i}, \quad \pi_{n_i} = (1-\rho_i)\rho_i^{n_i}, \quad \overline{S} = \frac{1}{\mu_i - \lambda_i}$$

↑ average # of packets in the system

↑ average sojourn time

$$\overline{W} = \frac{\rho_i}{\mu_i - \lambda_i}$$

↑ average waiting time in buffer

For the network as a whole

L_N - average # of packets in the network

W_N - average delay for a packet to go through the network

$$L_N = \sum_{i=1}^{M} \overline{L_i}$$

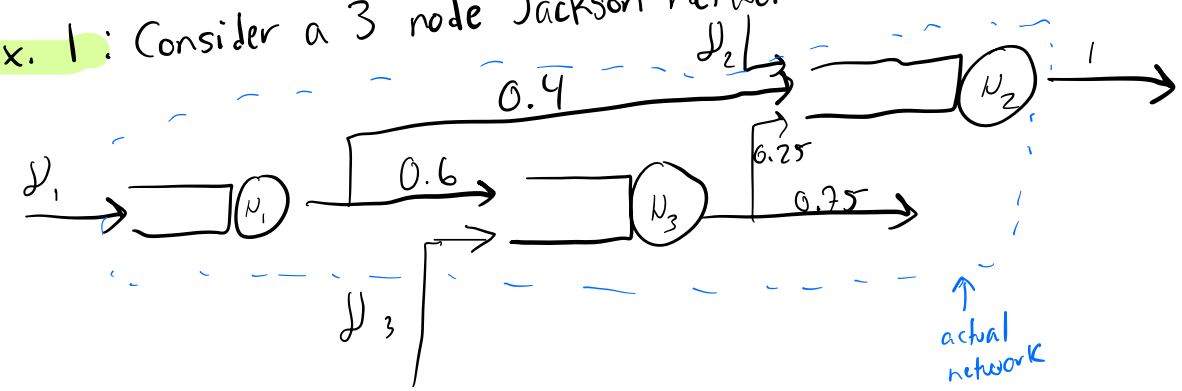
Use Little's Law to Find W_n

γ_N - total external traffic arrival rate to the network

$$\gamma_N = \sum_{i=1}^{M} \gamma_i$$

$$\overline{W}_N = \frac{\sum_{i=1}^{M} \frac{\rho_i}{1-\rho_i}}{\sum_{i=1}^{M} \gamma_i}$$

Ex. 1: Consider a 3 node Jackson network



Poisson external arrival with rates $\gamma_1 = 0.5$, $\gamma_2 = 0.25$, $\gamma_3 = 0.25$.

Exponential service times at each queue/server: $\mu_1 = 1$, $\mu_2 = 1$, $\mu_3 = 1$.

Average delay?

The routing matrix: $r_{12} = 0.4$ $r_{11} = 0$ $r_{13} = 0.6$

$r_{21} = 0$ $r_{22} = 0$ $r_{23} = 0$

$r_{31} = 0$ $r_{32} = 0.25$ $r_{33} = 0$

$r_{14} = 0$ ← don't need

$r_{24} = 1$

$r_{34} = 0.75$

$$\vec{R} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} = \begin{bmatrix} 0 & 0.4 & 0.6 \\ 0 & 0 & 0 \\ 0 & 0.25 & 0 \end{bmatrix}$$

$$\vec{\lambda} = \vec{\gamma}(I-R)^{-1} \Rightarrow \vec{\lambda} = [0.5 \quad 0.5875 \quad 0.55]$$

$$\rho_1 = \frac{\lambda_1}{\mu_1} = 0.5 \qquad \rho_2 = \frac{\lambda_2}{\mu_2} = 0.5875 \qquad \rho_3 = \frac{\lambda_3}{\mu_3} = 0.55$$

$$\gamma_N = \gamma_1 + \gamma_2 + \gamma_3 = 0.5 + 0.25 + 0.25 = 1$$

$$\bar{W}_N = \frac{\bar{L}_N}{\gamma_N} = \frac{\quad}{1} = 3.646 \text{ sec.}$$

$$\frac{0.5}{1-0.5} + \frac{0.5875}{1-0.5875} + \frac{0.55}{1-0.55} = 1 + \frac{0.5875}{1-0.5875} + \frac{0.55}{0.45}$$