M/M/c Queues

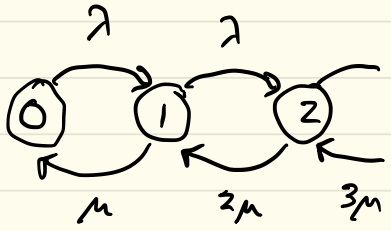
M/M/m queues
- c servers rather than 1
- Exponential inter-arrivals and service times

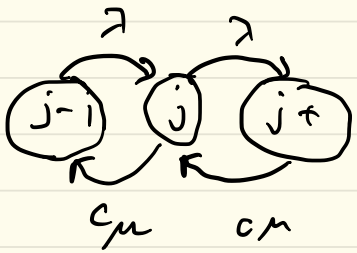
Examples : multiprocessors
multiple servers

Basic Assumptions
- c servers, identical service rate μ
- single queue for waiting customers
- arrival rate λ

Birth-Death process applies

$$\lambda_n = \lambda \quad \text{for all } n$$





$j \gg c$

$$\mu_n = \begin{cases} n\mu & n = 0, 1, \ldots, c \\ m\mu & n > c \end{cases}$$

M/M/c Queues

From Birth-Death process solution

$$\pi_n = \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} \qquad n = 1, 2, \ldots \infty$$

with $\rho = \lambda/\mu$ (note: useage of ρ varies between $\rho = \frac{\lambda}{\mu}$ and $\rho = \frac{\lambda}{c\mu}$)

Gives

$$\pi_n = \begin{cases} \dfrac{\rho^n}{n!} \cdot \pi_0 & n < c \\ \dfrac{\rho^n}{c!\, c^{n-c}} \cdot \pi_0 & c \le n \end{cases}$$

where

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

Erlang C formula

phone system modeled as M/M/c queue

gives probability an arriving call waits

$$E(c,\rho) = \sum_{n=c}^{\infty} \pi_n = \pi_0 \cdot \sum_{n=c}^{\infty} \frac{\rho^n}{c!\, c^{n-c}} = \frac{\pi_0 \rho^c}{c!} \sum_{n=0}^{\infty} \left(\frac{\rho}{c}\right)^n$$

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

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

M/M/c Queues

mean number $\bar{n}$ in system $\sum_{n=0}^{\infty} n \cdot \pi_n$

$$\bar{N} = \rho + \frac{\rho}{c} \cdot \frac{E(c,\rho)}{\left(1 - \frac{\rho}{c}\right)}$$

Expected number of jobs in service

$$\bar{n}_s = \sum_{n=1}^{c-1} n\pi_n + \sum_{n=c}^{\infty} c \cdot \pi_n = c\frac{\lambda}{\mu}$$

mean response time, use Little's Law

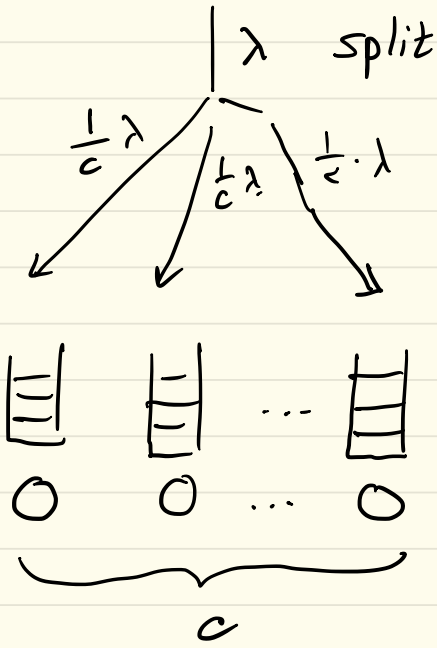
$$\therefore \quad \frac{\bar{n}}{\lambda}$$

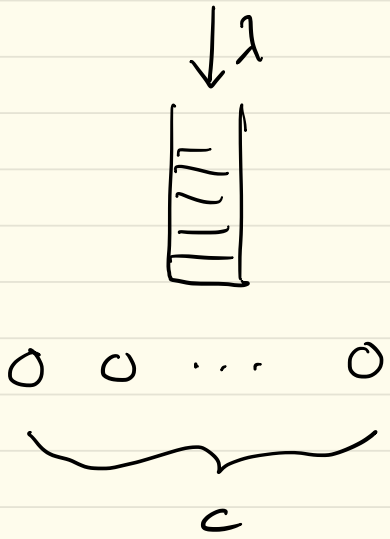
So

$$\bar{R} = \frac{1}{\lambda} + \frac{1}{\mu c} \cdot \frac{E(c,\rho)}{\left(1 - \frac{\rho}{c}\right)}$$

Comparison between M/M/1 & M/M/c

Consider 2 ways of organizing servers





- is there a difference, e.g. in time in system?

YES!

note it is possible on left to have free server and job in queue implies ....?

Comparison between M/M/1 & M/M/c

mean response time

c M/M/1 queues

$$\bar{R} = \frac{1}{\mu - \lambda/c}$$

1 M/M/m queue

$$\bar{R} = \frac{1}{\lambda} + \frac{1}{\mu c} \cdot \frac{E(c,\rho)}{\left(1 - \frac{\rho}{c}\right)}$$

Use of formulae to assess system behavior

Question:

- What performance for a given set of parameters?
- What parameters needed to achieve particular performance goals?

Examples

- What happens to mean response time if λ doubles?
- what happens to mean response time if we double # servers
- how many servers do we need for given λ, μ to achieve target mean response time

Example

Observation of jobs arriving to server with $c = 10$ CPUs is that

- $\lambda = 5$
- $1/\mu = 1$

Use M/M/c formula to estimate mean number in Q and mean response time

$N_Q = 0.036$ $\quad \bar{R} = 1.00$

virtually instantaneous response

Suppose arrival rate increases to $\lambda = 8$ after a promotion campaign

$N_Q = 1.63$ $\quad \bar{R} = 1.20$

marked relative increase in N_Q, only 20% increase in $\bar{R}$

Example $\mu = 1$, $\lambda = 3.9$

Find smallest number of servers c s.t. time in Queue is less than 20% of time in service

c	R_Q	$1/\mu$
4	9.45	1
5	0.47	1
[6	0.12	1]

JUST EVALUATE AT DIFFERENT c's

M/M/c/B

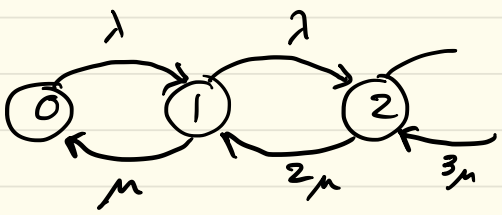
Finite buffers, at most B jobs in system

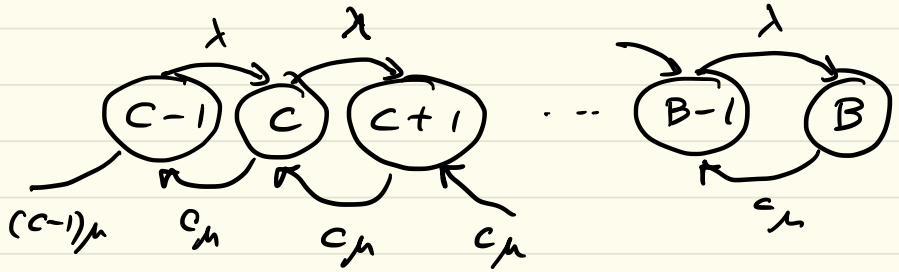
Using birth-death process framework

$$\lambda_n = \lambda \quad \text{for} \quad n = 0, 1, \ldots, B$$

$$\mu_n = \begin{cases} n\mu & \text{for } n < c \\ c\mu & \text{for } n = c, c+1, \ldots, B \end{cases}$$

NOTE: only makes sense for $B \geq c$ else servers idled





M/M/c/B

State occupancy probs, $\rho = \lambda/\mu$

$$\pi_n = \begin{cases} \frac{\rho^n}{n!} \cdot \pi_0 & n = 1, 2, \ldots, c-1 \\ \frac{\rho^n}{c! c^{n-c}} \pi_0 & n = c, c+1, \ldots, B \end{cases}$$

with

$$\pi_0 = \left[1 + \frac{\left(1 - \left(\frac{\rho}{c}\right)^{B-c-1}\right) \rho^c}{m!(1-\rho/c)} + \sum_{n=1}^{c-1} \frac{\rho^n}{n!} \right]^{-1}$$

use these to compute
- mean response time
- mean queue lengths
- "effective" arrival rates

M/M/c/B

mean queue length $\bar{n} = \sum_{n=0}^{B} n \cdot \pi_n$

mean number in queue $\bar{n}_q = \sum_{n=c}^{B} (n-c)\pi_n$

Given λ, "effective" arrival rate is rate of arrivals when they are accepted — i.e., not in state B

$$\tilde{\lambda} = \lambda \cdot (1 - \pi_B)$$

Use effective arrival rate with Little's Law to get mean response time

$$\bar{r} = \frac{\bar{n}}{\tilde{\lambda}} = \frac{\bar{n}}{\lambda(1-\pi_B)}$$

Utilization of a server is $\frac{\tilde{\lambda}}{\mu} = \rho(1-\pi_B)$

Other Queues of Interest

Exact solutions for

- $M/E_r/1$ — E_r is sum of r exponentials
- $M/M/1$ — with bulk arrivals
- $M/D/1$
- $G/M/1$
- $M/G/1$
- $G/G/1$

M/G/1

Problem – at instant when job arrives, a job in service has STATE
- unlike exponential, distribution of remaining service depends on length of time in service already

FINESSE
- define state transitions at embedded points where there is no memory

DEPARTURE INSTANTS

IF job has been in service t time, how many arrivals from Markovian arrival process?

Poisson (λt)

consider embedded DTMC

M/G/1

Consider # jobs left in queue at the point of a job's departure, This will be state of DTMC

From state $i > 0$

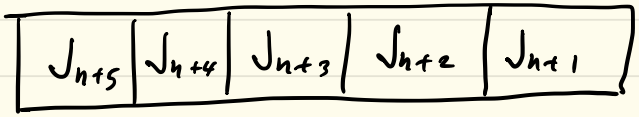
- transition to state j i.e. $j - i + 1$ arrivals during the service time of the customer entering service

example: At departure of n^{th} job



3

At departure of $(n+1)^{st}$ job



5

$\rightarrow$

J_{n+3}, J_{n+4}, J_{n+5} arrived

From state $i = 0$

- transition to j if j arrivals during service time of next arrival

M/G/1

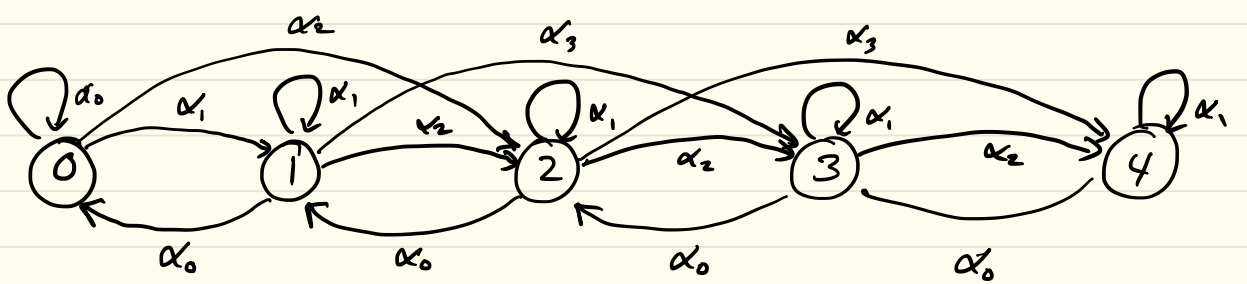
α_i = Prob{ i arrivals in a service time}

Given service time s, probability of i arrivals is

Prob a Poisson (λs) has value i

so $$\alpha_i = \int_0^\infty g(s) \frac{(\lambda s)^i}{i!} e^{-\lambda s} \, ds$$

where $g(s)$ is pdf of service time distribution



etc.

M/G/1

Solution of α_i gives distribution of queue length at job departure

mean queue length

$$\tilde{q} = \sum_{i=0}^{\infty} i \cdot \alpha_i$$

magic bit.

Limiting mean queue length at departures is limiting mean queue length!

M/G/1

Pollaczek–Khinchin formula for mean queue length

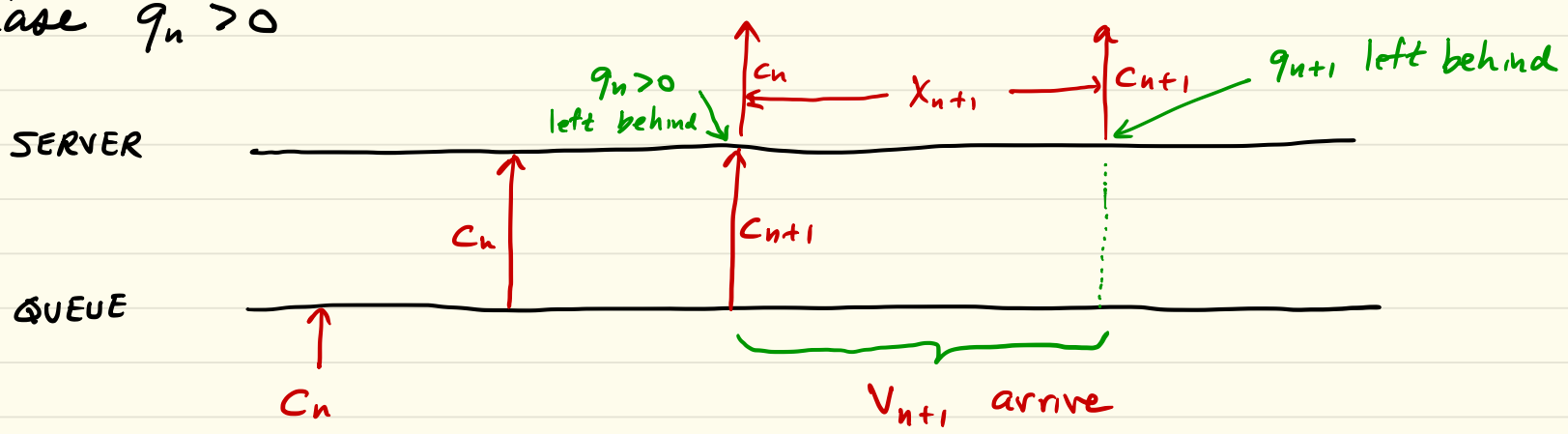
Define

C_n	n^{th} customer to enter system
T_n	arrival time of n^{th} customer
t_n	$T_n - T_{n-1}$
X_n	service time for n^{th} customer
q_n	number left in queue by departure of C_n
V_n	number of customers that arrive while C_n is receiving service

M/G/1

Formula for q_{n+1}

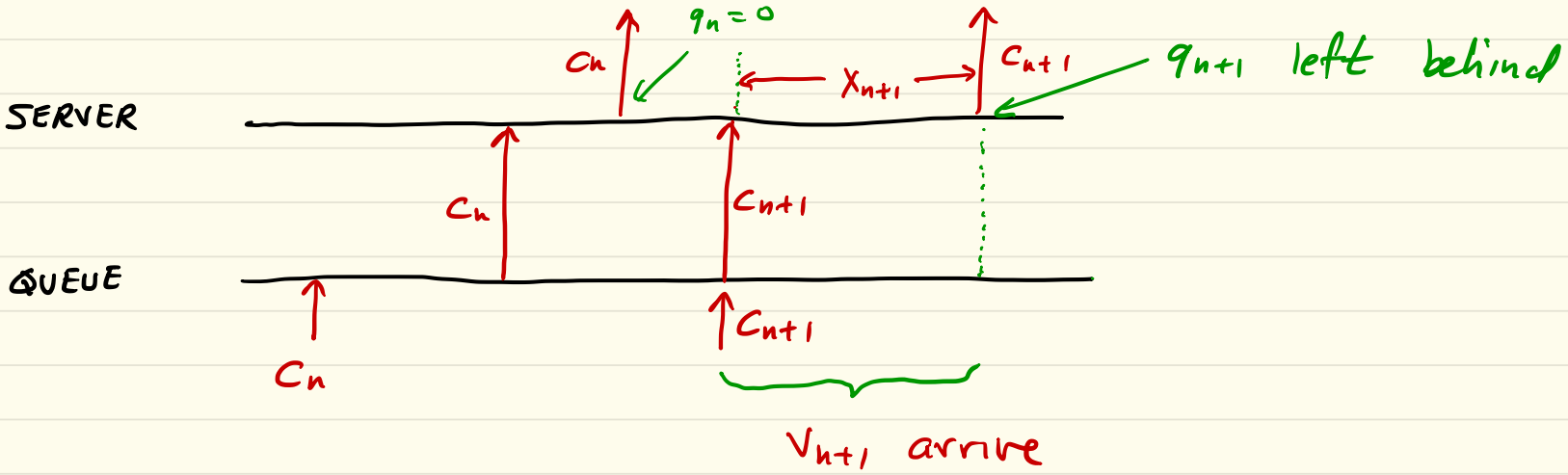
Case $q_n > 0$



$$q_{n+1} = q_n - 1 + V_{n+1} \quad \text{when } q_n > 0$$

M/G/1

Case $q_n = 0$



$q_{n+1} = V_{n+1}$ $q_n = 0$

M/G/1

Combine expressions

$$q_{n+1} = \begin{cases} q_n - 1 + V_{n+1} & \text{if } q_n > 0 \\ V_{n+1} \end{cases}$$

Define SHIFTED DISCRETE STEP FUNCTION

$$\Delta_k = \begin{cases} 1 & \text{for } k = 1, 2, \ldots \\ 0 & \text{for } k = 0 \end{cases}$$

and re-write

$$q_{n+1} = q_n - \Delta_{q_n} + V_{n+1}$$

Note: in limit as $n \to \infty$, $E[q_{n+1}] = E[q_n]$

M/G/1

$$q_{n+1} = q_n - \Delta_{q_n} + V_{n+1}$$

$$\Rightarrow \quad E[q_{n+1}] = E[q_n] - E[\Delta_{q_n}] + E[V_{n+1}]$$

$$\Rightarrow \quad \lim_{n\to\infty} E[q_{n+1}] = \lim_{n\to\infty} E[q_n] - E[\Delta_{q_n}] + E[V_{n+1}]$$

$$\Rightarrow \quad \lim_{n\to\infty} E[\Delta_{q_n}] = \lim_{n\to\infty} E[V_{n+1}]$$

We can compute

$$\lim_{n\to\infty} E[V_{n+1}] = \int_0^\infty g(s) \cdot \underbrace{\sum_{k=0}^{\infty} k \cdot \frac{(\lambda s)^k}{k!} \cdot e^{-\lambda s}}_{\text{mean of Poisson } \lambda s} ds$$

$$= \int_0^\infty g(s)\, \lambda \cdot s$$

$$= \lambda \cdot E[\text{service time}]$$

M/G/1

$$\lim_{n\to\infty} E[\Delta_{q_n}] = E[\Delta_{\tilde{q}}]$$

$$= \sum_{k=0}^{\infty} \Delta_k P[\tilde{q}=k]$$

$$= \Delta_0 P[\tilde{q}=0] + \Delta_1 P[\tilde{q}=1] + \cdots$$

recall definition of Δ_i

$$= 0 \cdot P(\tilde{q}=0) + 1 \cdot P(\tilde{q}>0)$$

$$= P(\tilde{q}>0)$$

$$= P(\text{busy system})$$

$$= \frac{\lambda}{E[\text{service time}]}$$

$$= \rho$$

M/G/1

BUT we didn't get to $E[\tilde{q}]$!

From $$q_{n+1} = q_n - \Delta_{q_n} + V_{n+1}$$

derive
$$q_{n+1}^2 = (q_n - \Delta_{q_n} + V_{n+1})^2$$

$$= q_n^2 + \Delta_{q_n}^2 + V_{n+1}^2 - 2q_n\Delta_{q_n} + 2q_nV_{n+1} - 2\Delta_{q_n}V_{n+1}$$

Note that $\Delta_i^2 = \Delta_i$ and $q_n\Delta_n = q_n$

which brings us to

$$E[q_{n+1}^2] = E[q_n^2] + E[\Delta_{q_n}] + E[V_{n+1}^2] - 2E[q_n] + 2E[q_nV_{n+1}] - 2E[\Delta_{q_n}V_{n+1}]$$

note that q_n and V_{n+1} are independent
Δ_{q_n} and V_{n+1} are independent

$\Rightarrow$

$$E[q_{n+1}^2] = E[q_n^2] + E[\Delta_{q_n}] + E[V_{n+1}^2] - 2E[q_n] + 2E[q_n]E[V_{n+1}] - 2E[\Delta_{q_n}]E[V_{n+1}]$$

leads to

$$0 = E[\Delta_{\tilde{q}}] + E[V^2] - 2E[\tilde{q}] + 2E[\tilde{q}]E[V] - 2E[\Delta_{\tilde{q}}]E[V]$$

but we've seen already that $E[\Delta_{\tilde{q}}] = E[V] = \rho$

which brings us to $$E[\tilde{q}] = \rho + \frac{E[V^2] - \rho}{2(1-\rho)}$$

the only unknown is $E[V^2]$

Aside: 2^{nd} moment of Poisson λs is $(\lambda s)^2 + \lambda s$
(from variance = second moment − mean^2

then $$E[V^2] = \int_0^\infty g(s) \cdot E[A(\lambda,s)^2]\, ds$$

$A(\lambda,s)^2$ 2^{nd} moment of Poisson λs

$$= \int_0^\infty g(s)\left[(\lambda s)^2 + \lambda s\right] ds$$

$$= \lambda^2 E[s^2] + \lambda E[s]$$

s the service time

M/G/1

$$E[\hat{q}] = \rho + \frac{E[V^2] - \rho}{2(1-\rho)}\rho = \rho + \frac{\lambda E[S^2] + \lambda E[S] - \rho}{2(1-\rho)}$$

let C_b be coefficient of variation of service time dist.

$$C_b = \frac{\text{var}(S)^{1/2}}{E[S]} \qquad \text{var}(S) = E[S^2] - E[S]^2$$

BUNCH OF ALGEBRA LATER :

Pollaczek – Khinchin mean value Formula

$$\bar{q} = \rho + \rho^2 \frac{1 + C_b^2}{2(1-\rho)}$$

M/G/1

$$\tilde{q} = \rho + \rho^2 \frac{(1 + C_b^2)}{2(1-\rho)}$$

EXAMPLE: Exponential service time: $C_b = 1$

$$\tilde{q} = \rho + \rho^2 \frac{2}{2(1-\rho)}$$

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

Deterministic: $C_b^2 = 0$

$$\tilde{q} = \rho + \frac{\rho^2}{2(1-\rho)} = \frac{\rho}{1-\rho} - \frac{\rho^2}{2(1-\rho)}$$

So M/D/1 has fewer customers on average than M/M/1