Stochastic Network Modeling (SNM)

Llorenç Cerdà-Alabern
Universitat Politècnica de Catalunya
Departament d'Arquitectura de Computadors
llorenc@ac.upc.edu

Parts

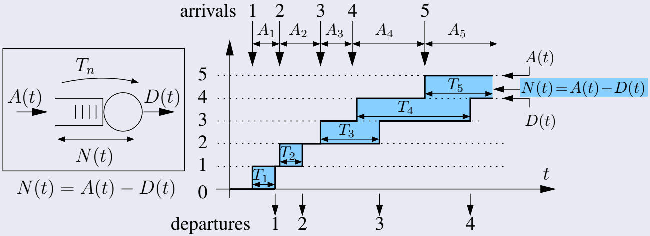
I. Introduction
II. Discrete Time Markov Chains (DTMC)
III. Continuous Time Markov Chains (CTMC)
IV. Queuing Theory

Part IV

Queuing Theory

Outline

Introduction

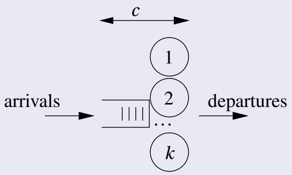


- Queueing theory is the mathematical study of waiting lines, or queues.
- Common notation:
  - $A(t)$: number of arrivals $[0, t]$.
  - A_n: interarrival time between customers n and $n+1$.
  - T_n: time in the system (response time) for customer n.
  - $N(t)$: number in the system at time t.

Kendal Notation

$$\boxed{\text{A/S/k[/c/p]}}$$

- A: arrival process,
- S: service process,
- k: number of servers,
- c: maximum number in the system (number of servers + queue size). Note: some authors use the queue size.
- p: population.
  If "c" or "p" are missing, they are assumed to be infinite.



Kendal Notation

Common arrivals/service processes

- G: general (non specific process is assumed),
- M: Markovian (exponentially or geometrically distributed),
- D: deterministic,
- P: Poisson (discrete RV, N, equal to the number of arrivals exponentially dist. in a time t):

$$P_p(N=n,t)=\frac{(\lambda\, t)^n \,\mathrm{e}^{-\lambda\, t}}{n!},\ n\geq 0, t\geq 0.$$

- Er: Erlang (continuous RV equal to the time t that last n arrivals exponentially dist.):

$$f_e(t)=\lambda\, P_p(N=n-1,t)=\frac{\lambda^n\, t^{n-1}\,\mathrm{e}^{-\lambda\, t}}{(n-1)!},\ t\geq 0, n\geq 1$$

Examples

- M/M/1: M. arr. / M. serv. / 1 server, ∞ queue and population.
- M/G/1: M. arr. / Gen. serv. / 1 server, ∞ queue and population.

Part IV

Queuing Theory

Outline

Little Theorem

Little Theorem

- Define the stochastic processes:
  - $A(t)$: number of arrivals $[0, t]$.
  - T_n: time in the system (response time) for customer n.
  - $N(t)$: number in the system at time t.
- And the mean values:
  - Mean number of customers in the system:

$$N = \lim_{t\to\infty} \frac{1}{t} \int_0^t N(s)\, \mathrm{d}s$$

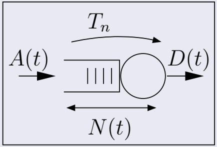
  - Arrival rate: $\lambda = \lim_{t\to\infty} A(t)/t$
  - Mean time in the system: $T = \lim_{t\to\infty} \left(\sum_n T_n\right)/A(t)$
- The following relation follows:

$$\boxed{N = \lambda T}$$

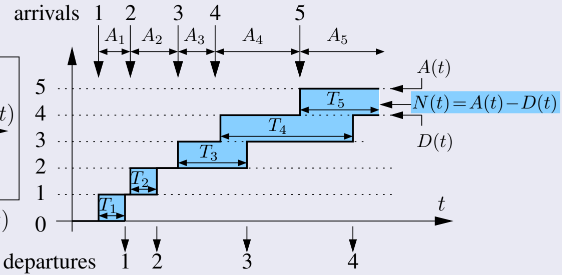
Mnemonic: NAT (Number = Arrivals x Time).

Little Theorem

Graphical proof



$N(t) = A(t) - D(t)$



- From the graph we have:

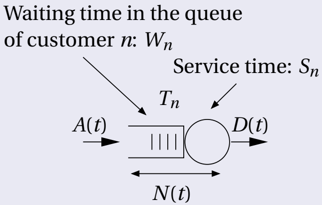
$$\frac{1}{t}\int_0^t N(s)\,ds = \frac{1}{t}\sum_{i=1}^{A(t)} T_i = \frac{A(t)}{t}\,\frac{\sum_{i=1}^{A(t)} T_i}{A(t)}$$

- Taking the limit $t \to \infty$: $\boxed{N = \lambda T}$

Little Theorem

Application to the waiting line and the server

- We can apply the Little theorem to the waiting line and the server:



Time in the system:
$T_n = W_n + S_n$
Expected value:
$T = W + S$
where
$T = \mathrm{E}[T_n]$, $W = \mathrm{E}[W_n]$,
$S = \mathrm{E}[S_n]$

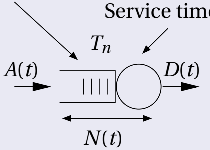
- Mean number of customers in the queue: $N_Q = \lambda W$.
- Mean number of customers in the server: $N_S = \rho = \lambda S$.

Little Theorem

Mean number in the Server

Waiting time in the queue of customer n: W_n

Service time: S_n



Time in the system:
$T_n = W_n + S_n$
Expected value:
$T = W + S$
where
$T = \mathrm{E}[T_n]$, $W = \mathrm{E}[W_n]$,
$S = \mathrm{E}[S_n]$

- In a single server queue (even if not Markovian):

$$\boxed{\begin{array}{c} \rho = N_S = \mathrm{E}[N_S(t)] = \lambda\,\mathrm{E}[S] \\ \mathrm{E}[N_S(t)] = 0 \times \pi_0 + 1 \times (1-\pi_0) = 1 - \pi_0 \Rightarrow \pi_0 = 1 - \rho \end{array}}$$

- $\rho = N_S = \lambda\,\mathrm{E}[S] = 1 - \pi_0$ is the proportion of time the system is busy, in other words, is the server utilization or load.

Part IV

Queuing Theory

Outline

PASTA Theorem

PASTA Theorem: Poisson Arrivals See Time Averages

- The mean time the chain is in state i is $\pi_i \Rightarrow$ using PASTA, the probability that a Markovian arrival see the system in state i is π_i (proof: see [1]).
- The equivalent theorem in discrete time is the arrival theorem, RASTA: Random Arrivals See Time Averages: the probability that a random arrival see the system in state i is π_i.

[1] Ronald W Wolff. "Poisson arrivals see time averages". In: *Operations Research* 30.2 (1982), pp. 223–231.

Example of PASTA

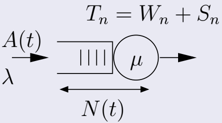
- Assume that a system can have, at most, N customers (e.g $N-1$ in the queue and 1 in service).
- Assume that an arrival is lost when the system is full.
- By PASTA the proportion of Poisson arrivals that see the system full, and are lost, is equal to the proportion of time the system has N in the system, π_N.
- Thus, the loss probability is π_N.

Part IV

Queuing Theory

Outline

The M/M/1 Queue



- Markovian arrivals with rate $\lambda \Rightarrow$ the interarrival time is exponentially distributed with mean $1/\lambda$:

$$P\{A_n \leq x\} = 1 - e^{-\lambda x},\ x \geq 0$$

$\Rightarrow$ $A(t)$ is a Poisson process:

$$P(A(t) = i) = \frac{(\lambda t)^i}{i!} e^{-\lambda t},\ i \geq 0,\ t \geq 0$$

- Markovian Services with rate $\mu \Rightarrow$ service time exponentially distributed with mean $1/\mu$:

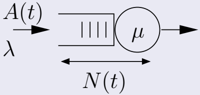
$$P\{S_n \leq x\} = 1 - e^{-\mu x},\ x \geq 0$$

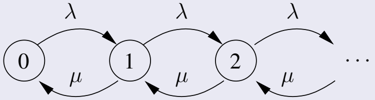
The M/M/1 Queue

Q-matrix

- The process $N(t)$ = {number in the system at time $t \geq 0$} is a CTMC.

  OBSERVATION: for a non Markovian service, the process $N(t)$ would not be a MC! State transition diagram:





- Q-matrix:

$$\mathbf{Q} = \begin{bmatrix} -\lambda & \lambda & 0 & 0 & \cdots \\ \mu & -(\mu+\lambda) & \lambda & 0 & \cdots \\ 0 & \mu & -(\mu+\lambda) & \lambda & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

Stationary Distribution

- Solving the M/M/1 queue using flux balancing (or the general solution of a reversible chain):

$$\boxed{\pi_i = (1-\rho)\,\rho^i,\ i = 0, \cdots, \infty}$$

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

Properties

- Mean customers in the system:

$$N = \lim_{t\to\infty} \frac{1}{t}\int_0^t N(s)\,ds = \sum_{i=0}^{\infty} i\pi_i = \sum_{i=0}^{\infty} i(1-\rho)\rho^i = \frac{\rho}{1-\rho}$$

- Mean time in the system (response time):

  Little: $N = \lambda T \Rightarrow T = \frac{N}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu-\lambda}$

- Mean time in the queue: $W = T - \frac{1}{\mu} = \frac{\rho}{\mu-\lambda}$

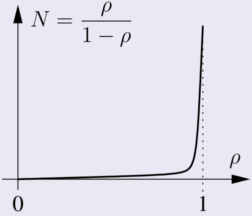
- Mean Number in the queue: $N_Q = \lambda W = \frac{\rho^2}{1-\rho}$

- Mean number in the server: $N_s = N - N_Q = \rho$

  NOTE: $\pi_0 = 1-\rho$

Stability

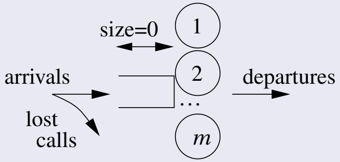
- N and T are proportional to $1/(1-\rho) \Rightarrow$ when $\rho \to 1 \Rightarrow N, T \to \infty$.
- The process $N(t)$ is positive recurrent, null recurrent or transient according to whether $\rho = \lambda/\mu$ is below, equal or greater than 1, respectively.



The M/M/1 Queue

Example: Loss probability in a telephone switching center

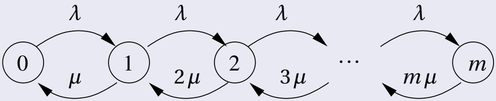
- Hypothesis: Switching center with m circuits and "lost call", infinite population, Markovian arrivals with rate λ and exponentially distributed call duration with mean $1/\mu \Rightarrow$ M/M/m/m queue.



The M/M/1 Queue

Example: Loss probability in a telephone switching center

- Since the minimum of i independent and identically exponentially distributed RV with parameter service time is exponentially distributed with parameter $i\,\mu$:



The M/M/1 Queue

Example: Loss probability in a telephone switching center

- Stationary Distribution of the queue M/M/m/m:
- Solving using the general solution of a reversible chain:

$$\text{Define } \rho_k = \frac{\lambda}{(k+1)\,\mu},\ k = 0, \cdots, m-1$$

$$\pi_0 = \frac{1}{G},\ \pi_i = \frac{1}{G}\prod_{k=0}^{i-1}\rho_k = \frac{1}{G}\left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!},\ 0 < i \leq m \Rightarrow$$

$$\boxed{\pi_i = \frac{1}{G}\left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!},\ 0 \leq i \leq m.\ G = \sum_{k=0}^{m}\left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}.}$$

- Using PASTA Theorem (Poisson Arrivals See Time Average): the loss call probability is the probability that the queue is in state m: π_m, "Erlang B Formula".

Part IV

Queuing Theory

Outline

M/G/1 Queue

M/G/1 Queue

- The process $N(t) = \{\text{number in the system at time } t \geq 0\}$ in general it is not a MC (it is so only if G is Markovian).
- We can build a semi-Markov process observing the system at departure times t_n (note that t_n are also the service completion times). Define the discrete time process:

$$X(n) = \{\text{number in the system at time } t_n \geq 0, n = 0, 1, \cdots\}$$

- Theorem: The process $X(n)$ is a DTMC.
- Proof: $X(n)$ only depends on the number of arrivals in non overlapping intervals. Since arrivals are Markovian, this is a memoryless process. $\square$
- NOTE: Looking at departure times the chain may have self transitions (in contrast to observing at transition times): we can have the same number in the system after a departure.

M/G/1 Queue

Transition Probability Matrix

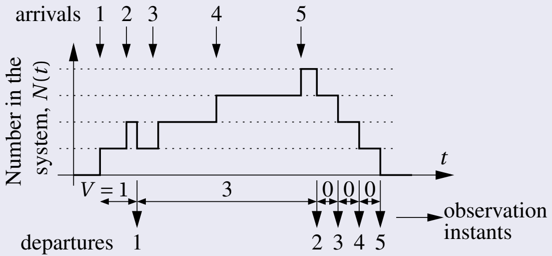
- Let $f_S(x)$, $x \geq 0$ be the service time density function.
- Define the RV $V = \{$number of arrivals during a service time$\}$, and the probabilities: $v_i = P\{V = i\}$.
- Conditioning on the service duration:

$$v_i = \int_{x=0}^{\infty} P\{i \text{ arrivals in time } x \mid S = x\}\, f_S(x)\, dx \Rightarrow$$

$$\boxed{v_i = \int_{x=0}^{\infty} \frac{(\lambda\, x)^i}{i!}\, \mathrm{e}^{-\lambda\, x} f_S(x)\, \mathrm{d}x}$$

M/G/1 Queue

Transition Probability Matrix

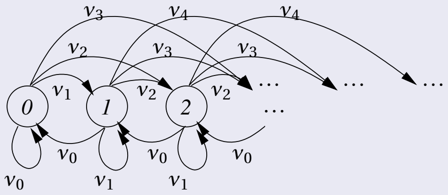


- $v_i = P\{\text{number of arrivals during a service time} = i\} \Rightarrow$

$$
p_{ij} = \begin{cases} 0, & j < i-1 \quad (N(t) \text{ can only be decreased by 1}) \\ v_j, & i = 0, j \geq 0 \quad (i = 0 \rightarrow \text{the queue was empty}) \\ v_{j-i+1}, & i > 0, j \geq i-1 \quad (i > 0 \rightarrow \text{the queue was busy}) \end{cases}
$$

M/G/1 Queue

Transition Probability Matrix



$$p_{ij} = \begin{cases} 0, & j < i-1 \\ v_j, & i = 0, j \geq 0 \\ v_{j-i+1}, & i > 0, j \geq i-1 \end{cases} \Rightarrow \mathbf{P} = \begin{bmatrix} v_0 & v_1 & v_2 & v_3 & \cdots \\ v_0 & v_1 & v_2 & v_3 & \cdots \\ 0 & v_0 & v_1 & v_2 & \cdots \\ 0 & 0 & v_0 & v_1 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

- Stationary distribution: $\pi = \pi\,\mathbf{P}$, $\pi\,\mathbf{e} = 1$.

M/G/1 Queue

Properties of the stationary distribution ($\pi = \pi\,\mathbf{P}, \pi\,\mathbf{e} = 1$)

- Using the "Level Crossing Law" theorem: a queue with unitary arrivals and departures satisfies:

$$\boxed{\begin{aligned} &P\{\text{an arriving customer finds } i \text{ in the system}\} = \\ &\quad P\{\text{a departing customer leaves } i \text{ in the system}\} \Rightarrow \end{aligned}}$$

$$\boxed{\pi_i = P\{\text{an arriving customer find } i \text{ in the system}\}}$$

- Using PASTA:

$$\boxed{\begin{aligned} \pi_i = P\{&\text{there are } i \text{ customers in the} \\ &\text{system at an arbitrary time}\} \end{aligned}}$$

So, in an M/G/1 the stationary distribution of the EMC obtained observing the departures, is the stationary distribution of the continuous time process.

M/G/1 Queue

Proof of the Level Crossing Law Theorem

- Define:
  - $A_i(t)$ ={number of arrivals finding i in the system at $t \geq 0$}
  - $D_i(t)$ ={number of departures leaving i in the system at $t \geq 0$}
  - P{a customer finds i in the system} $= \lim_{t\to\infty} A_i(t)/A(t)$
  - P{a customer leave i in the system} $= \lim_{t\to\infty} D_i(t)/D(t)$
- An arriving customer that finds i in the system produce a transition $i \to i+1$. A customer leaving i in the system produce a transition $i+1 \to i$.
- Since arrivals and departures are unitary, the number of transitions $i \to i+1$ and $i+1 \to i$ can only differ in 1: $|A_i(t) - D_i(t)| \leq 1$. Note that $N(t) = A(t) - D(t)$.
- For a stable queue: $A(t) - D(t) < \infty$

Proof of the Level Crossing Law Theorem

- We have:
  - $A_i(t) =$ {number of arrivals finding i customer in the system}
  - $D_i(t) =$ {number of departures leaving i customers in the system}
  - P{a customer finds i in the system} $= \lim_{t\to\infty} A_i(t)/A(t)$
  - P{a customer leave i in the system} $= \lim_{t\to\infty} D_i(t)/D(t)$
  - $A_i(t) - D_i(t) \in \{0, 1\}$, $N(t) = A(t) - D(t) < \infty$.
  - $\lim_{t\to\infty} A(t) = \infty$, $\lim_{t\to\infty} D(t) = \infty$.
- Thus:

$$\boxed{\begin{aligned}\lim_{t\to\infty}\left\{\frac{A_i(t)}{A(t)} - \frac{D_i(t)}{D(t)}\right\} &= \lim_{t\to\infty}\left\{\frac{A_i(t)}{A(t)} - \frac{D_i(t)}{A(t)} - \left(\frac{D_i(t)}{D(t)} - \frac{D_i(t)}{A(t)}\right)\right\} = \\ &\lim_{t\to\infty}\left\{\frac{A_i(t) - D_i(t)}{A(t)} - \frac{D_i(t)}{D(t)}\frac{A(t) - D(t)}{A(t)}\right\} = 0\end{aligned}}$$

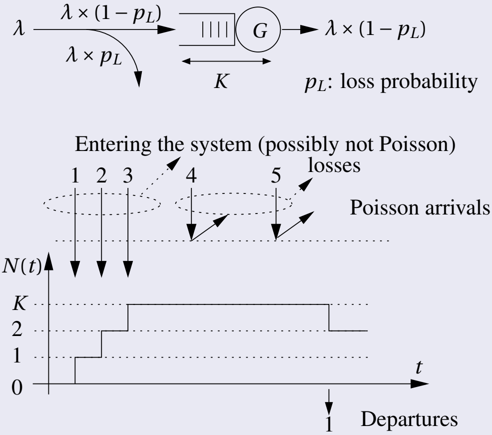
Part IV

Queuing Theory

Outline

M/G/1/K Queue

Problem Formulation



Stationary Distribution

- Using the general solution of an M/G/1/K we obtain the stationary distribution of the number in the system left by a departing customer: π_i^d.
- By the Level Crossing Law this is the stationary distribution of the number in the system found by the successful arrivals:

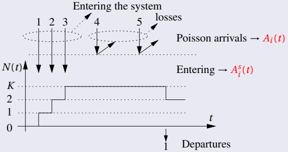
$$\boxed{\pi_i^s = \pi_i^d,\ i = 0, 1, \cdots K-1.}$$

and

$$\boxed{\pi_i^s = P(\text{a customer entering the system finds } i)}$$

- NOTE: a departing customer cannot leave the system full (nor an arrival can enter the system when it is full).

Loss Probability



Define:

- $A_i^a(t)$: Number of arrivals (lost or not) finding i in the system.
- $A_i^s(t)$: Number of successful arrivals finding i in the system.
- π_i^a, π_i^s the stationary distribution of the embedded Markov chains $A_i^a(t)$, $A_i^s(t)$. By PASTA π_i^a is also the stationary distribution of the continuous time process. Thus,

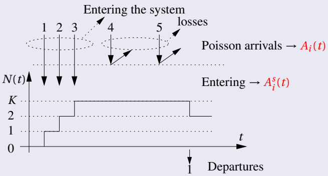
$$\pi_i^s = P(\text{a customer entering the system finds } i),\ i = 0, 1, \cdots K-1 \Rightarrow$$

$$\pi_i^s = \lim_{t\to\infty} \frac{A_i^s(t)}{\sum_{k=0}^{K-1} A_k^s(t)} \frac{\sum_{k=0}^{K} A_k^a(t)}{\sum_{k=0}^{K} A_k^a(t)} = \frac{\pi_i^a}{\sum_{k=0}^{K-1} \pi_i^a} = \frac{\pi_i^a}{1-\pi_K^a} = \frac{\pi_i^a}{1-p_L}, \Rightarrow$$

$$\boxed{\pi_i^a = \pi_i^s(1-p_L) = \pi_i^d(1-p_L),\ i = 0, 1, \cdots K-1}$$

M/G/1/K Queue

Loss Probability



- Applying Little: $\rho_s = \mathrm{E}[N_s] = 1 - \pi_0 = \lambda\,(1-p_L)\mathrm{E}[S] = \rho\,(1-p_L)$. Where $\rho = \lambda\,\mathrm{E}[S]$ and π_0 is the proportion of time the server is empty.
- Using PASTA: $\pi_0 = \pi_0^a$ (Poisson arrivals). Using $\pi_i^a = \pi_i^d\,(1-p_L)$:

$$\left.\begin{array}{l} 1-\pi_0 = 1-\pi_0^a = 1-\pi_0^d\,(1-p_L) \\ 1-\pi_0 = \rho(1-p_L) \end{array}\right\} \Rightarrow \boxed{p_L = \frac{\rho+\pi_0^d-1}{\rho+\pi_0^d},\ \rho = \lambda\,\mathrm{E}[S]}$$

- Where π_0^d is computed using the general solution of an M/G/1/K.

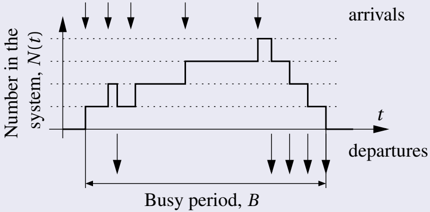
Part IV

Queuing Theory

Outline

M/G/1 Busy Period

Expected Length of a Busy Period

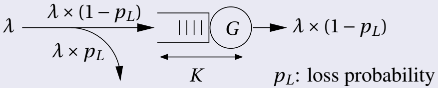


- Define the RV:
  - Busy period, B.
  - Idle period, I. Poisson arrivals with rate $\lambda \Rightarrow \mathrm{E}[I] = 1/\lambda$
- Clearly:

$$\text{System load } \rho = \lambda\,\mathrm{E}[S] = \frac{\mathrm{E}[B]}{\mathrm{E}[I] + \mathrm{E}[B]} \Rightarrow \boxed{\mathrm{E}[B] = \frac{1}{\lambda}\frac{\rho}{1-\rho}}$$

M/G/1 Busy Period

M/G/1/K Busy Period



- Busy period, B.
- Idle period, I. Poisson arrivals with rate $\lambda \Rightarrow$ $\mathrm{E}[I] = 1/\lambda$
- Clearly:

$$\text{System load } \rho_s = \lambda\,(1-p_L)\,\mathrm{E}[S] = \frac{\mathrm{E}[B]}{\mathrm{E}[I]+\mathrm{E}[B]} \Rightarrow$$

$$\boxed{\mathrm{E}[B] = \frac{1}{\lambda}\,\frac{\rho\,(1-p_L)}{1-\rho\,(1-p_L)},\ \rho = \lambda\,\mathrm{E}[S]}$$

- Or, in terms of $\pi_0 = \pi_0^d\,(1-p_L)$:

$$\text{System load } \rho_s = 1-\pi_0 = \frac{\mathrm{E}[B]}{\mathrm{E}[I]+\mathrm{E}[B]} \Rightarrow \boxed{\mathrm{E}[B] = \frac{1}{\lambda}\,\frac{1-\pi_0}{\pi_0}}$$

Part IV

Queuing Theory

Outline

M/G/1 Delays

M/G/1 Mean Time in the Queue

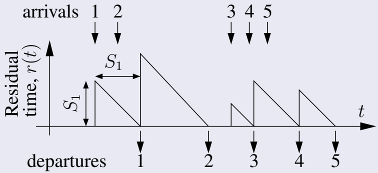
- Method of the moments: Using PASTA, the mean time in the queue (W) for an arriving customer, is the mean time to finish the current service (mean residual time, R) plus the mean time to service the customers in the queue ($\mathrm{E}[S]\, N_Q$):

$$\boxed{W = R + \mathrm{E}[S]\, N_Q}$$

- Using Little for the queue length:

$$\boxed{N_Q = \lambda\, W \Rightarrow W = R + E[S]\, \lambda\, W \Rightarrow W = \frac{R}{1-\rho},\ \rho = \lambda\, E[S].}$$

M/G/1 Mean Time in the Queue



- From the figure (note the right triangles with two equal cathetus), we have:

$$\boxed{R = \frac{1}{t}\int_0^t r(\tau)\,d\tau = \frac{1}{t}\sum_{i=1}^{A(t)}\frac{S_i^2}{2} = \frac{1}{2}\frac{A(t)}{t}\sum_{i=1}^{A(t)}\frac{S_i^2}{A(t)} \underset{t\to\infty}{=} \frac{1}{2}\lambda\,\mathrm{E}[S^2]}$$

M/G/1 Mean Time in the Queue

- For instance, for an M/M/1

$$\mathrm{E}[S^2] = \mathrm{Var}(S) + \mathrm{E}[S]^2 = \frac{1}{\mu^2} + \left(\frac{1}{\mu}\right)^2 = \frac{2}{\mu^2},$$

  thus, the residual time is:

$$R = \frac{1}{2}\lambda\,\mathrm{E}[S^2] = \frac{\lambda}{\mu^2} = \frac{\rho}{\mu}, \rho = \frac{\lambda}{\mu}.$$

- We can check that $\mathrm{E}[R|\text{S idle}] = 0$ and $\mathrm{E}[R|\text{S busy}] = 1/\mu$, thus

$$R = \mathrm{E}[R|\text{S idle}]\,\pi_0 + \mathrm{E}[R|\text{S busy}]\,(1-\pi_0) = \frac{\rho}{\mu}, \rho = 1-\pi_0,$$

  as expected.

M/G/1 Delays

M/G/1 Mean Time in the Queue

- We have:

$$W = \frac{R}{1-\rho}, \rho = \lambda E[S]$$

$$R = \frac{1}{2}\lambda \mathrm{E}[S^2]$$

- Substituting we get the Pollaczek-Khinchin, P-K formula:

$$\boxed{W = \frac{\lambda \mathrm{E}[S^2]}{2(1-\rho)}, \rho = \lambda \mathrm{E}[S]}$$

M/G/1 Delays

M/G/1 Mean Time in the Queue

- Mean time in the system (response time):

$$\boxed{T = \mathrm{E}[S] + W = \mathrm{E}[S] + \frac{\lambda\,\mathrm{E}[S^2]}{2\,(1-\rho)}}$$

- For an M/M/1 queue: $\mathrm{E}[S^2] = \frac{2}{\mu^2} \Rightarrow W = \frac{\rho}{\mu\,(1-\rho)}$
- For an M/D/1 queue: $E[S^2] = \frac{1}{\mu^2} \Rightarrow W = \frac{\rho}{2\,\mu\,(1-\rho)}$
- Observation: The M/D/1 has the minimum value of $E[S^2] \Rightarrow$ is a lower bound of W, T, N_Q and N for an M/G/1.

P-K Formula Does Not Apply to an M/G/1/K Queue

- P-K formula is not applicable to an M/G/1/K queue because the customers entering the system might not be Poisson. Thus, they does not observe the mean residual time.
- Example: Customers entering an M/G/1/1 queue (0 queue size) observe the system always empty. Thus, in an M/G/1/1 queue the expected time in the queue is $W = 0$ (P-K formula does not apply), and the expected time in the system is $T = \mathrm{E}[S]$ (mean service time).
- With an M/G/1/K we can compute $N = \sum_{n=1}^{K} n\pi_n^a$, and use Little: $N = \lambda\,(1 - p_L)\,T$. For instance, for an M/G/1/1 we have $\pi_0^d = 1$, and $N = 0\,\pi_0^a + 1\,\pi_1^a = \pi_1^a = p_L$. Thus, $p_L = \frac{\rho+\pi_0^d-1}{\rho+\pi_0^d} = \frac{\rho}{\rho+1}$, and $T = \frac{N}{\lambda\,(1-p_L)} = \frac{p_L}{\lambda\,(1-p_L)} = \frac{\rho}{\lambda} = \mathrm{E}[S]$, as expected.

Part IV

Queuing Theory

Outline

Queues in Tandem

Burke theorem

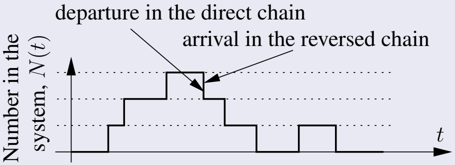
- The departure process in an M/M/m queue, $1 \leq m \leq \infty$, is a Poisson process with the same parameter than the arrival process.
- At each time t, the number of customers in the system is independent of the sequence of departures previous to t.

Burke theorem. Proof (1)

- Relation between the arrival and departure process:

  > The departure process in a reversible queue has the same joint distribution than the arrival process.

- Proof:
  - If the queue is reversible: $q_{ij} = q^r_{ij} \Rightarrow$ the arrival process in the reversed chain has the same distribution than the arrival process in the direct chain,
  - but:

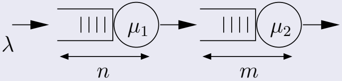


Burke theorem. Proof (2)

- The queue M/M/m is reversible $\Rightarrow$ The departures are Poisson with the same parameter than the arrivals.
- The arrivals in the reversed chain previous to t are Markovian, thus, independent of the number of customers in the system after t. This implies that the departures in the direct chain are independent of the number in the system before t. $\square$

Queues in Tandem

Tandem M/M/m Queues



- Define the chain:

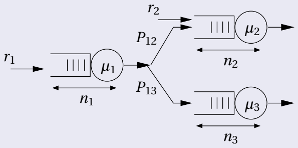
$$X(n,m) = \{n \text{ in the system } 1,\ m \text{ in the system } 2\}$$

- The stationary distribution is the product of the stationary distributions of the isolated queues:

$$\boxed{\pi_{nm} = (1-\rho_1)\,\rho_1^n\,(1-\rho_2)\rho_2^m,\ \rho_1 = \lambda/\mu_1,\ \rho_2 = \lambda/\mu_2}$$

- Proof: Using Burke, the departures of system 1 are Poisson and the number in the system 1 is independent of the previous departures (arrivals to system 2), thus, independent from the number of customers in system 2. $\square$

Feed Forward Queues



- Suppose M/M/1 queues with outside arrivals with rate r_i randomly forwarded with probabilities P_{ij} (see figure).
- The network has the following product form solution:

$$\boxed{\pi(n_1, n_2, \cdots n_K) = (1-\rho_1)\,\rho_1^{n_1}\,(1-\rho_2)\,\rho_2^{n_2}\cdots(1-\rho_k)\rho_k^{n_k}, \quad \rho_i = \lambda_i/\mu_i.}$$

- The rates λ_i are computed solving: $\boxed{\lambda_i = r_i + \sum_j \lambda_j P_{ji}}$.
- Stability condition: $\rho_i < 1$.

Queues in Tandem

Feed Forward Queues

Proof (draft)

- Burke theorem.
- Superposition of Poisson processes with rates λ_i is Poisson with rate $\sum_i \lambda_i$.
- A Poisson process with rate λ randomly split with probabilities p_i, $\sum_i p_i = 1$, produce Poisson processes with rates $p_i \lambda$.

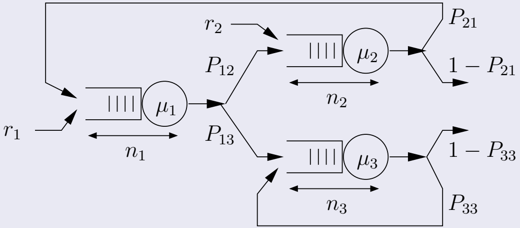
Part IV

Queuing Theory

Outline

Jackson Theorem

- Suppose M/M/m queues. In queue i the customers arrive from outside with rate r_i and depart to queue j with probability P_{ij}, or leave the system with probability $1 - \sum_j P_{ij}$:



Jackson Theorem

- The network has the following product form solution:

$$\boxed{\pi(n_1, n_2, \cdots n_K) = \pi_{n_1} \pi_{n_2} \cdots \pi_{n_k}}$$

  where π_{n_i} is the solution of the queue i with arrival rates λ_i obtained solving:

$$\boxed{\lambda_i = r_i + \sum_j \lambda_j P_{ji}}$$

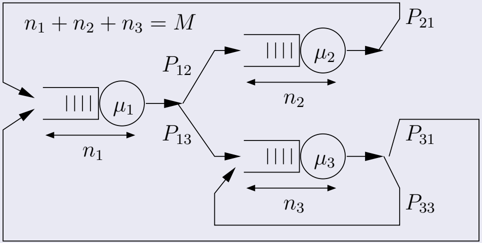
- Stability condition: $\rho_i = \lambda_i / \mu_i < 1$.
- For example, for M/M/1 queues:

$$\boxed{\pi(n_1, n_2, \cdots n_K) = (1-\rho_1)\,\rho_1^{n_1}\,(1-\rho_2)\,\rho_2^{n_2} \cdots (1-\rho_k)\rho_k^{n_k}}$$

- Proof: The solution satisfies the global balance equations.
- NOTE: The proof is different from feed forward queues, since routing loops make arrivals not necessarily Poisson.

Closed Networks of Queues

- M/M/m networks without arrivals and departures to outside of the system:



Jackson Theorem for Closed Networks of Queues

- The network has the following product form solution:

$$\boxed{\pi(n_1, n_2, \cdots n_K) = \frac{1}{G}\rho_1^{n_1}\,\rho_2^{n_2}\cdots\rho_k^{n_k},\ \rho_i = \lambda_i/\mu_i.}$$

- Where the rates λ_i are any solution to the equations:

$$\boxed{\lambda_i = \sum_j \lambda_j P_{ji} \quad \text{(in matrix form: } \boldsymbol{\lambda} = \boldsymbol{\lambda}\mathbf{P}\text{)}}$$

- And the normalization factor is given by:

$$\boxed{G = \sum_{n_1+n_2+\cdots+n_k=M} \rho_1^{n_1}\,\rho_2^{n_2}\cdots\rho_k^{n_k}}$$

- Proof: The solution satisfies the global balance equations.
- NOTE: the equation $n_1 + n_2 + \cdots + n_k = M$ has $\binom{M+k-1}{M} = \binom{M+k-1}{k-1}$ solutions (ways to allocate M items in k boxes).

Part IV

Queuing Theory

Outline

Squared coefficient of variation

- Idea: almost all distributions can be approximated by a mixture of exponentials.
- Squared coefficient of variation, characterization of a distribution variability (for distributions with $E[X] > 0$):

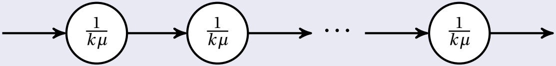
$$\boxed{C_X^2 = \frac{\mathrm{Var}(X)}{E[X]^2} = \frac{E[X^2] - E[X]^2}{E[X]^2} = \frac{E[X^2]}{E[X]^2} - 1}$$

- Deterministic distribution: $C_D^2 = 0$.
- Exponential distribution: $E[X] = 1/\mu$, $\mathrm{Var}(X) = 1/\mu^2$. Thus $C_{\mathrm{exp}}^2 = 1$.
- What if we want a distribution more *deterministic* than an exponential, $C_X^2 < 1$? or with larger variability, $C_X^2 > 1$?

Matrix Geometric Method

$C_X^2 < 1$: Erlang-k

- k stages exponentially distributed with parameter $k\mu$:



$$f_E(t) = \frac{(k\mu)^k \, t^{k-1} \, e^{-k\mu t}}{(k-1)!}, \; t \geq 0, k \geq 1$$

$$E[X] = k\frac{1}{k\mu} = \frac{1}{\mu}$$

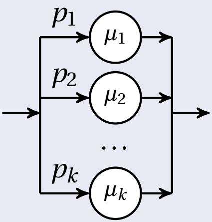
$$\mathrm{Var}(X) = k \times \mathrm{Var}(\exp(k\mu)) = k\frac{1}{(k\mu)^2} = \frac{1}{k\mu^2}$$

$$C_X^2 = \frac{\mathrm{Var}[X]}{E[X]^2} = \frac{1}{k} < 1$$

Matrix Geometric Method

$C_X^2 > 1$: Hyper-exponential

$$f_H(t) = \sum_{i=1}^{k} p_i \mu_i e^{-\mu_i t}, \ \sum_{i=1}^{k} p_i = 1, \ t \geq 0$$



$$\mathrm{E}[X] = \sum_{i=1}^{k} p_i \frac{1}{\mu_i}, \mathrm{E}[X^2] = \sum_{i=1}^{k} p_i \frac{2}{\mu_i^2}$$

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2 =$$

$$\sum_{i=1}^{k} p_i \frac{2}{\mu_i^2} - \left(\sum_{i=1}^{k} p_i \frac{1}{\mu_i}\right)^2 =$$

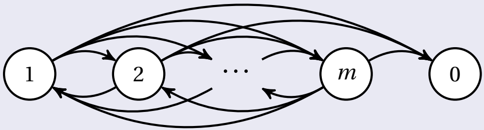
$$\left(\sum_{i=1}^{k} p_i \frac{1}{\mu_i}\right)^2 + \sum_{i=1}^{k} \sum_{j \neq i} p_i p_j \left(\frac{1}{\mu_i} - \frac{1}{\mu_j}\right)^2$$

$$C_X^2 = \frac{\mathrm{Var}[X]}{\mathrm{E}[X]^2} > 1$$

Matrix Geometric Method

Phase type distribution

- General mixture of exponentials.
- The service ends upon reaching the absorbing state.
- Can approximate arbitrary distributions.
- Representation: $PH(\mathbf{a}, \mathbf{T})$.



$$\mathbf{Q}^{m+1\times m+1} = \begin{bmatrix} \mathbf{T}^{m\times m} & \mathbf{c}^{m\times 1} \\ \mathbf{0}^{1\times m} & 0 \end{bmatrix}$$

Initial prob.

$$\begin{bmatrix} \mathbf{a}^{1\times m} & a_0 \end{bmatrix}.$$

$$f_{PH}(t) = \mathbf{a}\, e^{\mathbf{T}\, t}\, \mathbf{c},\ t \geq 0$$

$$E[X^k] = k!\, \mathbf{a}\, (-\mathbf{T}^{-1})^k\, \mathbf{e}$$

where $\mathbf{e}$ is a column vector of 1s.

Quasi Birth Death Processes

- Assume a two dimensional MC with states (n, i) (e.g. an M/PH/1 queue). We call n the level and i the phase. We group the states of the stationary distribution:

$$\boldsymbol{\pi} = \begin{bmatrix} \boldsymbol{\pi}_0 & \boldsymbol{\pi}_1 & \boldsymbol{\pi}_2 & \cdots \end{bmatrix}$$

$$\begin{cases} \boldsymbol{\pi}_0 = \begin{bmatrix} (0,0) & (0,1) & \cdots (0,m') \end{bmatrix} & \text{initial part (level 0)} \\ \boldsymbol{\pi}_i = \begin{bmatrix} (i,1) & \cdots & (i,m) \end{bmatrix}, \; i \geq 1 & \text{repetitive part (level } i \geq 1) \end{cases}$$

$$\mathbf{Q} = \begin{bmatrix} \mathbf{L}_0 & \mathbf{F}_0 & & & \\ \mathbf{B}_0 & \mathbf{L} & \mathbf{F} & & \\ & \mathbf{B} & \mathbf{L} & \mathbf{F} & \\ & & \mathbf{B} & \mathbf{L} & \cdots \\ & & & \cdots & \cdots \end{bmatrix}$$

$$\begin{cases} \mathbf{B} & \text{governs the transitions to previous level} \\ \mathbf{L} & \text{governs the change of phase inside a level} \\ \mathbf{F} & \text{governs the transitions to next level} \end{cases}$$

Matrix Geometric Method

Matrix Geometric Solution

$$\mathbf{Q} = \begin{bmatrix} \mathbf{L}_0 & \mathbf{F}_0 & & & \\ \mathbf{B}_0 & \mathbf{L} & \mathbf{F} & & \\ & \mathbf{B} & \mathbf{L} & \mathbf{F} & \\ & & \mathbf{B} & \mathbf{L} & \cdots \\ & & & \cdots & \cdots \end{bmatrix}, \; [\pi_0 \quad \pi_1 \quad \pi_2 \quad \pi_3 \quad \cdots]\, \mathbf{Q} = 0.$$

- Due to similarity with an M/M/1 ($\pi_i = \pi_0\, \rho^i$) we guess for the repetitive part:

$$\boxed{\boldsymbol{\pi}_{i+1} = \boldsymbol{\pi}_1 \mathbf{R}^i,\ i \geq 0}$$

which gives:

$$\boldsymbol{\pi}_1 \mathbf{F} + \boldsymbol{\pi}_2 \mathbf{L} + \boldsymbol{\pi}_3 \mathbf{B} = \mathbf{0} \Rightarrow$$

$$\boldsymbol{\pi}_1 \mathbf{F} + \boldsymbol{\pi}_1 \mathbf{R} \mathbf{L} + \boldsymbol{\pi}_1 \mathbf{R}^2 \mathbf{B} = \mathbf{0} \Rightarrow$$

$$\boxed{\mathbf{F} + \mathbf{R}\mathbf{L} + \mathbf{R}^2 \mathbf{B} = \mathbf{0}}$$

Matrix Geometric Solution

- Due to similarity with an M/M/1 ($\pi_i = \pi_0 \rho^i$) we guess for the repetitive part:

$$\boxed{\boldsymbol{\pi}_{i+1} = \boldsymbol{\pi}_1 \mathbf{R}^i,\ i \geq 0}$$

which gives:

$$\boxed{\mathbf{F} + \mathbf{R}\,\mathbf{L} + \mathbf{R}^2\,\mathbf{B} = \mathbf{0}}$$

- Isolating **R** we have that it can be found iterating

$$\boxed{\mathbf{R}_{n+1} = -(\mathbf{F} + \mathbf{R}_n^2\,\mathbf{B})\,\mathbf{L}^{-1},}$$

starting e.g. with $\mathbf{R}_0 = \mathbf{I}$.

- Better iterative algorithms can be found in [1].

[1] Guy Latouche and Vaidyanathan Ramaswami. *Introduction to matrix analytic methods in stochastic modeling.* Vol. 5. Siam, 1999.

Queuing Theory

Introduction
Kendal Notation
Little Theorem
PASTA Theorem
The M/M/1 Queue
M/G/1 Queue
M/G/1/K Queue
M/G/1 Busy Period
M/G/1 Delays
Queues in Tandem
Networks of Queues

Basic algorithm with R

- Compute R matrix

```
##
## Basic iterative algorithm to compute the matrix R
##
## B, L, F: repetitive part matrices
##
invL <- -solve(L) # -1/L
C1 <- F %*% invL # -F/L
C2 <- B %*% invL # -B/L
R <- diag(nrow(B))
epsilon <- 1e-15
MaxIter <- 500
IterB <- 1
while(IterB < MaxIter) {
    prev <- R
    R <- C1 + R %*% R %*% C2 # -(F + R^2 B)/L
    if(max(abs(prev-R)) < epsilon) { break }
    IterB = IterB + 1
}
```

Solving $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$

- We have:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{L}_0 & \mathbf{F}_0 & & & \\ \mathbf{B}_0 & \mathbf{L} & \mathbf{F} & & \\ & \mathbf{B} & \mathbf{L} & \mathbf{F} & \\ & & \mathbf{B} & \mathbf{L} & \cdots \\ & & & \cdots & \cdots \end{bmatrix}$$

$$\begin{bmatrix} \boldsymbol{\pi}_0 & \boldsymbol{\pi}_1 & \boldsymbol{\pi}_2 & \boldsymbol{\pi}_3 & \cdots \end{bmatrix} \mathbf{Q} = 0$$

$$\boldsymbol{\pi}_{i+1} = \boldsymbol{\pi}_1 \mathbf{R}^i, \; i \geq 0.$$

- Thus:

$$\begin{aligned} \boldsymbol{\pi}_0 \mathbf{L}_0 + \boldsymbol{\pi}_1 \mathbf{B}_0 &= \mathbf{0} \\ \boldsymbol{\pi}_0 \mathbf{F}_0 + \boldsymbol{\pi}_1 \mathbf{L} + \boldsymbol{\pi}_1 \mathbf{R}\mathbf{B} &= \mathbf{0} \end{aligned} \Rightarrow \boxed{\begin{bmatrix} \boldsymbol{\pi}_0 & \boldsymbol{\pi}_1 \end{bmatrix} \begin{bmatrix} \mathbf{L}_0 & \mathbf{F}_0 \\ \mathbf{B}_0 & \mathbf{L} + \mathbf{R}\mathbf{B} \end{bmatrix} = \mathbf{0}}$$

- and the normalization condition:

$$\boldsymbol{\pi}_0 \mathbf{e}_0 + \sum_{i=0}^{\infty} \boldsymbol{\pi}_1 \mathbf{R}^i \mathbf{e}_1 = 1 \Rightarrow \boldsymbol{\pi}_0 \mathbf{e}_0 + \boldsymbol{\pi}_1 (\mathbf{I} - \mathbf{R})^{-1} \mathbf{e}_1 = 1 \Rightarrow$$

$$\boxed{\begin{bmatrix} \boldsymbol{\pi}_0 & \boldsymbol{\pi}_1 \end{bmatrix} \begin{bmatrix} \mathbf{e}_0 \\ (\mathbf{I} - \mathbf{R})^{-1} \mathbf{e}_1 \end{bmatrix} = 1}$$

where $\mathbf{e}_i$ are column vectors of 1s of appropriate size.

Basic algorithm with R

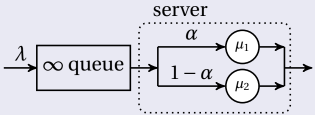
- Compute the initial probabilities

```
##
## Basic algorithm to compute the initial probabilities
##
## B0, L0, F0: initial part matrices
## B, L, F: repetitive part matrices
##
IMRinv <- solve(diag(nrow(B))-R) # 1/(I-R)
M0 <- rbind(cbind(L0, F0), cbind(B0, L + R %*% B))

## Normalization column
NE <- c(rep(1, nrow(L0)), IMRinv %*% rep(1,nrow(IMRinv)))

## solve using the replace 1 equation method
M0 <- cbind(NE, M0[,2:ncol(M0)]) # replace first column of M0 by NE
stat <- solve(t(M0), c(1, rep(0, nrow(M0)-1)))
```

Matrix Geometric Method

Matrix Geometric Method, Example

- Consider an M/G/1 queue where service time is hyper-exponentially distributed:



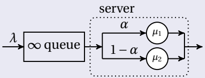
$\lambda = 1$
$\alpha = 1/2$
$\mu_1 = 1$
$\mu_2 = 2$

- Derive the rate matrix, Q, ordering the states lexicographically. Identify the states that form the initial and repetitive part. Identify the submatrices that would be used for a matrix geometric solution: $\mathbf{B}_0$, $\mathbf{L}_0$, $\mathbf{F}_0$, $\mathbf{B}$, $\mathbf{L}$, $\mathbf{F}$.
- Solve the Chain using the matrix geometric method. Compute the number in the system. Check it with the PK formula.

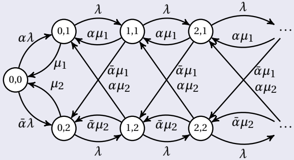
Matrix Geometric Method

Matrix Geometric Method, Example

state: (level, phase)



$\lambda = 1$
$\alpha = 1/2$
$\mu_1 = 1$
$\mu_2 = 2$



$\bar{\alpha} = 1 - \alpha$

$$
\mathbf{P} = \begin{array}{c} \begin{array}{cccccccc} 0,0 & 0,1 & 0,2 & 1,1 & 1,2 & 2,1 & 2,2 & \cdots \end{array} \\
\left[\begin{array}{c|cc|cc|cc|c}
-\lambda & \alpha\lambda & \bar{\alpha}\lambda & 0 & 0 & 0 & 0 & \cdots \\
\hline
\mu_1 & -(\lambda+\mu_1) & 0 & \lambda & 0 & 0 & 0 & \cdots \\
\mu_2 & 0 & -(\lambda+\mu_2) & 0 & \lambda & 0 & 0 & \cdots \\
\hline
0 & \alpha\mu_1 & \bar{\alpha}\mu_1 & -(\lambda+\mu_1) & 0 & \lambda & 0 & \cdots \\
0 & \alpha\mu_2 & \bar{\alpha}\mu_2 & 0 & -(\lambda+\mu_2) & 0 & \lambda & \cdots \\
\hline
0 & 0 & 0 & \alpha\mu_1 & \bar{\alpha}\mu_1 & -(\lambda+\mu_1) & 0 & \cdots \\
0 & 0 & 0 & \alpha\mu_2 & \bar{\alpha}\mu_2 & 0 & -(\lambda+\mu_2) & \cdots \\
\hline
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}\right] \end{array} \quad (1)
$$

$$
\Rightarrow \mathbf{B}_0 = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \mathbf{L}_0 = [-\lambda], \mathbf{F}_0 = [\alpha\lambda \quad \bar{\alpha}\lambda], \mathbf{B} = \begin{bmatrix} \alpha\mu_1 & \bar{\alpha}\mu_1 \\ \alpha\mu_2 & \bar{\alpha}\mu_2 \end{bmatrix}, \mathbf{L} = \begin{bmatrix} -(\lambda+\mu_1) & 0 \\ 0 & -(\lambda+\mu_2) \end{bmatrix}, \mathbf{F} = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}.
$$

Matrix Geometric Method, Example

- Iterating $\mathbf{R}_{n+1} = -(\mathbf{F} + \mathbf{R}_n^2\,\mathbf{B})\,\mathbf{L}^{-1}$ we get:

$$\mathbf{R} = \begin{bmatrix} 5/7 & 1/7 \\ 1/7 & 3/7 \end{bmatrix}$$

- Using $\begin{bmatrix} \boldsymbol{\pi}_0 & \boldsymbol{\pi}_1 \end{bmatrix} \begin{bmatrix} \mathbf{L}_0 & \mathbf{F}_0 \\ \mathbf{B}_0 & \mathbf{L} + \mathbf{R}\,\mathbf{B} \end{bmatrix} = \mathbf{0}$, $\begin{bmatrix} \boldsymbol{\pi}_0 & \boldsymbol{\pi}_1 \end{bmatrix} \begin{bmatrix} \mathbf{e}_0 \\ (\mathbf{I} - \mathbf{R})^{-1}\,\mathbf{e}_1 \end{bmatrix} = 1$ we get:

$$\begin{bmatrix} \boldsymbol{\pi}_0 & \boldsymbol{\pi}_1 \end{bmatrix} = \begin{bmatrix} 1/4 & 3/28 & 1/14 \end{bmatrix}$$

- Number in the system:

$$N = \sum_{n=1}^{\infty} n\,\boldsymbol{\pi}_1\,\mathbf{R}^n\,\mathbf{e}_1 = \boldsymbol{\pi}_1\,(\mathbf{I} - \mathbf{R})^{-2}\,\mathbf{e}_1 = \frac{13}{4}$$

- Using the PK Formula:

$$\mathrm{E}[S] = \frac{\alpha}{\mu_1} + \frac{1-\alpha}{\mu_2} = \frac{1}{4},\ \rho = \lambda\,\mathrm{E}[S] = \frac{3}{4},\ \mathrm{E}[S^2] = \frac{2\,\alpha}{\mu_1^2} + \frac{2\,(1-\alpha)}{\mu_2^2} = \frac{5}{4}$$

thus,

$$T = \mathrm{E}[S] + \frac{\lambda\,\mathrm{E}[S^2]}{2\,(1-\rho)} = \frac{13}{4},\ N = \lambda\,T = \frac{13}{4},\ \text{as expected.}$$