Stochastic Network Modeling (SNM)

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

Parts

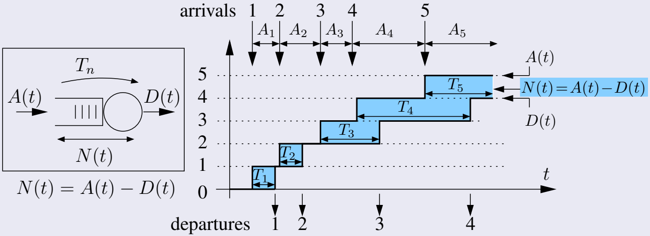
- I Introduction
- II Discrete Time Markov Chains (DTMC)
- III Continuous Time Markov Chains (CTMC)
- IV Queuing Theory

Part IV

Queuing Theory

Outline

Introduction



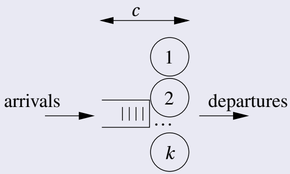
- Queueing theory is the mathematical study of waiting lines, or queues.
- Common notation:
  - $A(t)$: number of arrivals $[0, t]$.
  - A_n: interarrival time between customers n and $n+1$.
  - T_n: time in the system (response time) for customer n.
  - $N(t)$: number in the system at time t.

Kendal Notation

Kendal Notation

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

- 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:

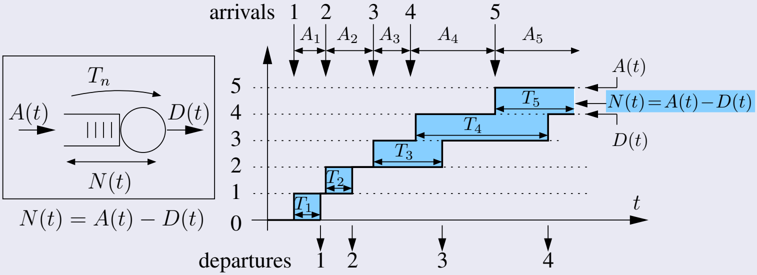
$$N = \lim_{t\to\infty} \frac{1}{t} \int_0^t N(s)\, \mathrm{d}s$$

  - 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).

Graphical proof



- From the graph we have:

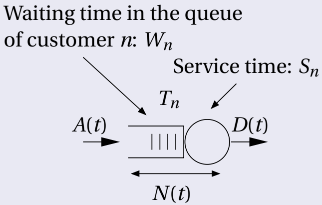
$$\frac{1}{t}\int_0^t N(s)\,ds = \frac{1}{t}\sum_{i=1}^{A(t)} T_i = \frac{A(t)}{t}\,\frac{\sum_{i=1}^{A(t)} T_i}{A(t)}$$

- 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:

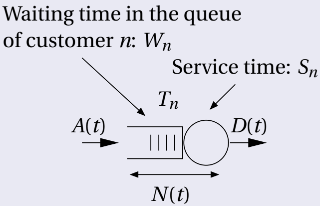


Time in the system:
$T_n = W_n + S_n$
Expected value:
$T = W + S$
where
$T = \mathrm{E}[T_n]$, $W = \mathrm{E}[W_n]$,
$S = \mathrm{E}[S_n]$

- 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



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):

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

- $\rho = N_S = \lambda\, \mathrm{E}[S] = 1 - \pi_0$ 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.

Little Theorem

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.

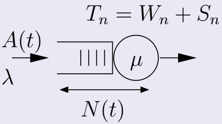
Part IV

Queuing Theory

Outline

The M/M/1 Queue

The M/M/1 Queue



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

$$P\{A_n \le x\} = 1 - \mathrm{e}^{-\lambda x},\ x \ge 0$$

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

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

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

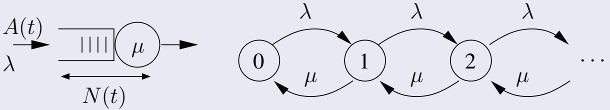
$$P\{S_n \le x\} = 1 - \mathrm{e}^{-\mu x},\ x \ge 0$$

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

The M/M/1 Queue

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

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

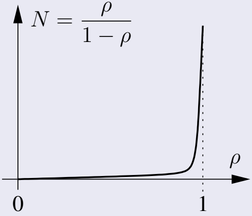
- Mean Number in the queue: $N_Q = \lambda\,W = \dfrac{\rho^2}{1-\rho}$

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

  NOTE: $\pi_0 = 1-\rho$

Stability

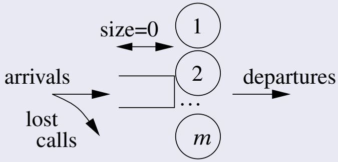
- N and T are proportional to $1/(1-\rho) \Rightarrow$ when $\rho \to 1 \Rightarrow N, T \to \infty$.
- The process $N(t)$ is positive recurrent, null recurrent or transient according to whether $\rho = \lambda/\mu$ is below, equal or greater than 1, respectively.



The M/M/1 Queue

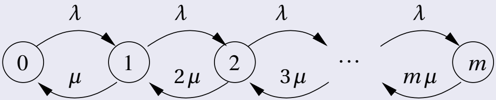
Example: Loss probability in a telephone switching center

- 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.



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 $i\,\mu$:



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. □
- 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

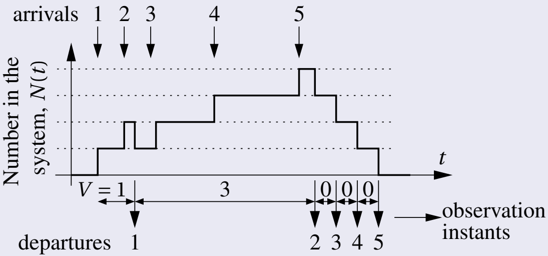
- Let $f_S(x), x \geq 0$ be the service time density function.
- Define the RV $V = \{\text{number of arrivals during a service time}\}$, and the probabilities: $v_i = P\{V = i\}$.
- Conditioning on the service duration:

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

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

M/G/1 Queue

Transition Probability Matrix

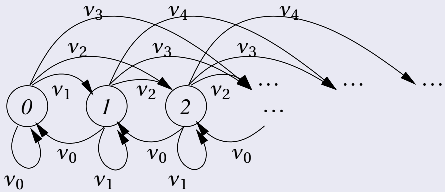


- $v_i = P\{\text{number of arrivals during a service time} = i\} \Rightarrow$

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

M/G/1 Queue

Transition Probability Matrix



$$p_{ij} = \begin{cases} 0, & j < i-1 \\ \nu_j, & i = 0, j \geq 0 \\ \nu_{j-i+1}, & i > 0, j \geq i-1 \end{cases} \Rightarrow \mathbf{P} = \begin{bmatrix} \nu_0 & \nu_1 & \nu_2 & \nu_3 & \cdots \\ \nu_0 & \nu_1 & \nu_2 & \nu_3 & \cdots \\ 0 & \nu_0 & \nu_1 & \nu_2 & \cdots \\ 0 & 0 & \nu_0 & \nu_1 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

- 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:

$$P\{\text{an arriving customer finds } i \text{ in the system}\} = P\{\text{a departing customer leaves } i \text{ in the system}\} \Rightarrow$$

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

- Using PASTA:

$$\pi_i = P\{\text{there are } i \text{ customers in the the system at an arbitrary time}\}$$

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.

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} $= \lim_{t\to\infty} D_i(t)/D(t)$
  - $|A_i(t) - D_i(t)| \le 1$, $N(t) = A(t) - D(t) < \infty$.
  - $\lim_{t\to\infty} A(t) = \infty$, $\lim_{t\to\infty} D(t) = \infty$.
- Thus:

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

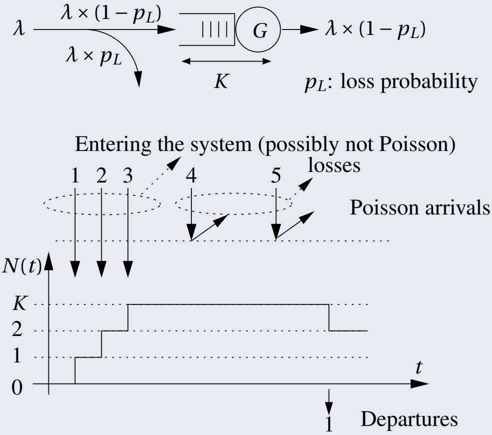
Part IV

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:

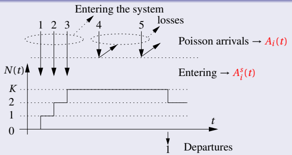
$$\boxed{\pi_i^s = \pi_i^d,\ i = 0, 1, \cdots K-1.}$$

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:

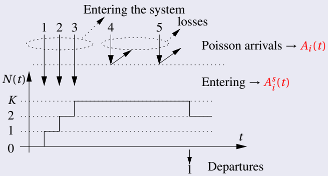
- $A_i^a(t)$: Number of arrivals (lost or not) finding i in the system.
- $A_i^s(t)$: Number of successful arrivals finding i in the system.
- π_i^a, π_i^s the stationary distribution of the embedded Markov chains $A_i^a(t)$, $A_i^s(t)$. By PASTA π_i^a is also the stationary distribution of the continuous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

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

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

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

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.