Stochastic Network Modeling (SNM)

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

Parts

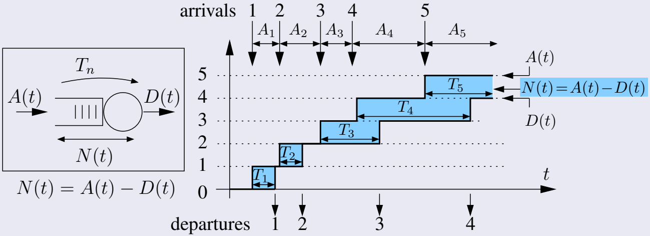
I. Introduction
II. Discrete Time Markov Chains (DTMC)
III. Continuous Time Markov Chains (CTMC)
IV. Queuing Theory

Part IV

Queuing Theory

Outline

Introduction



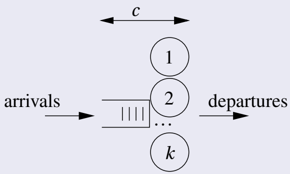
- Queueing theory is the mathematical study of waiting lines, or queues.
- Common notation:
  - $A(t)$: number of arrivals $[0, t]$.
  - A_n: interarrival time between customers n and $n+1$.
  - T_n: time in the system (response time) for customer n.
  - $N(t)$: number in the system at time t.

Kendal Notation

Kendal Notation

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

- 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:

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

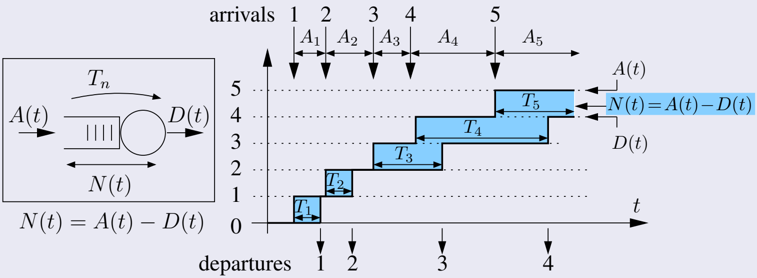
  - Arrival rate: $\lambda = \lim_{t\to\infty} A(t)/t$
  - Mean time in the system: $T = \lim_{t\to\infty} \left(\sum_n T_n\right)/A(t)$
- The following relation follows:

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

Mnemonic: NAT (Number = Arrivals x Time).

Little Theorem

Graphical proof



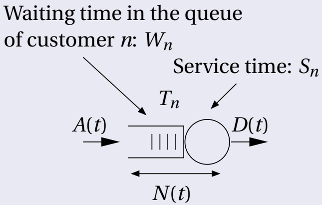
- From the graph we have:

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

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

Application to the waiting line and the server

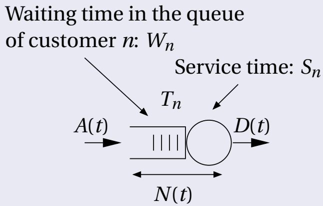
- We can apply the Little theorem to the waiting line and the server:



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: $N_S = \rho = \lambda S$.

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):

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

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

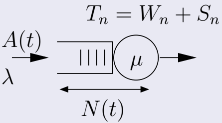
- Assume that a system can have, at most, N customers (e.g $N-1$ in the queue and 1 in service).
- Assume that an arrival is lost when the system is full.
- By PASTA the proportion of Poisson arrivals that see the system full, and are lost, is equal to the proportion of time the system has N in the system,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

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

$\Rightarrow$ $A(t)$ 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 $1/\mu$:

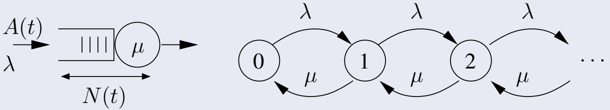
$$P\{S_n \le x\} = 1 - \mathrm{e}^{-\mu x},\ x \ge 0$$

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:

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

The M/M/1 Queue

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

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

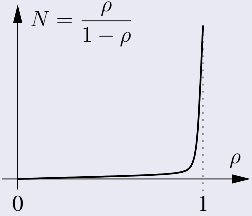
- Mean Number in the queue: $N_Q = \lambda W = \dfrac{\rho^2}{1-\rho}$

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

  NOTE: $\pi_0 = 1-\rho$

Stability

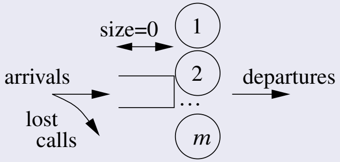
- N and T are proportional to $1/(1-\rho) \Rightarrow$ when $\rho \to 1 \Rightarrow N, T \to \infty$.
- The process $N(t)$ is positive recurrent, null recurrent or transient according to whether $\rho = \lambda/\mu$ is below, equal or greater than 1, respectively.



The M/M/1 Queue

Example: Loss probability in a telephone switching center

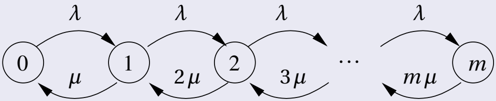
- Hypothesis: Switching center with m circuits and "lost call", infinite population, Markovian arrivals with rate λ and exponentially distributed call duration with mean $1/\mu \Rightarrow$ M/M/m/m queue.



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 $i\mu$:



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:

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

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

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

- 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”.