

Stochastic Network Modeling (SNM)

Queuing Theory

# Stochastic Network Modeling (SNM)

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

#### **Parts**

- Introduction
- Discrete Time Markov Chains (DTMC)
- Continuous Time Markov Chains (CTMC)
- Queuing Theory



Stochastic Network Modeling (SNM)

Queuing Theory

Part IV

# Queuing Theory

#### Outline

- Introduction
- Kendal Notation
- Little Theorem
- PASTA Theorem
- The M/M/1 Queue
- M/G/1 Oueue

- M/G/1/K Queue
- Oueues in Tandem
- Networks of Oueues
- M/G/1 Busy Period
- M/G/1 Delays
- Matrix Geometric Method



## Introduction

Queuing Theory

Introduction

Little Theorem

PASTA

The M/M/

M/G/1 Onen

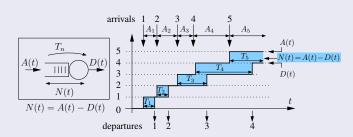
m/G/1 Queuc

Queues in

Networks of

M/G/1 Busy Period

Period



- 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**

Queuing Theory

Introduction

Notation

Little Theore

PASTA Theorem

The M/M/1 Oueue

M/G/1 Queu

M/G/1/K

Queues ii

Networks o Queues

M/G/1 Busy Period

#### **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.

arrivals 2 departures k



### **Kendal Notation**

Queuing Theory

Introduction

Notation

PASTA Theorem

The M/M/I Queue

M/G/1 Queue

Queues in

Networks of Queues

M/G/1 Busy Period

#### 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*):

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

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

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

#### Examples

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



Stochastic Network Modeling (SNM)

Queuing Theory

Part IV

**Queuing Theory** 

Little Theorem

## Outline

- Little Theorem

Llorenç Cerdà-Alabern



Queuing Theory

Introduction

Little Theorem

Graphical proof Application to the waiting line and the server

Mean number in the

PASTA Theorem

The M/M/

M/G/1 Quet

M/G/1/K Queue

Queues in Tandem

Queues in

#### 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:  $T = \lim_{t \to \infty} (\sum_n T_n) / A(t)$
- The following relation follows:

$$N = \lambda T$$

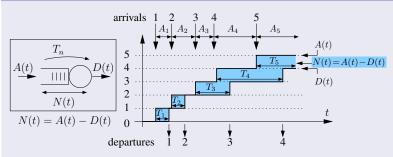
Mnemonic: NAT (Number = Arrivals x Time).



Queuing Theory

Graphical proof

## 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  $t \to \infty$ :  $N = \lambda T$ 



Queuing Theory

Introduction

Little Theorem

Graphical proof

Application to the waiting line and the server

Mean number in the

PASTA Theorem

The M/M/1

M/G/1 Queu

M/G/1/K Queue

Queues in Tandem

Networks of

#### Application to the waiting line and the server

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

Waiting time in the queue of customer n:  $W_n$ Service time:  $S_n$   $A(t) \qquad D(t)$ 

N(t)

Time in the system:  $T_n = W_n + S_n$ Expected value:

$$T = W + S$$
  
where  
 $T = E[T_n], W = E[W_n],$   
 $S = E[S_n]$ 

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

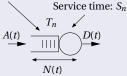


Queuing Theory

Mean number in the

#### Mean number in the Server

Waiting time in the queue of customer n:  $W_n$ 



Time in the system:

$$T_n = W_n + S_n$$
  
Expected value:  
 $T = W + S$ 

where

 $T = E[T_n], W = E[W_n],$  $S = E[S_n]$ 

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

$$\rho = N_S = \mathbb{E}[N_S(t)] = \lambda \, \mathbb{E}[S]$$
  
$$\mathbb{E}[N_S(t)] = 0 \times \pi_0 + 1 \times (1 - \pi_0) = 1 - \pi_0 \Rightarrow \pi_0 = 1 - \rho$$

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



Stochastic Network Modeling (SNM)

Queuing Theory

PASTA Theorem

# Part IV

# **Queuing Theory**

#### Outline

- PASTA Theorem



#### PASTA Theorem

Queuing Theory

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



Queuing Theory

Introductior

. . . . .

Little Theorei

Theorem

Example of PASTA

he M/M/1

M/G/1 Oueue

M/G/1/K Queue

Queues in Tandem

Queues

I/G/1 Busy

#### 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,  $\pi_N$ .
- Thus, the loss probability is  $\pi_N$ .



Stochastic Network Modeling (SNM)

Queuing Theory

Part IV

The M/M/1 Queue

**Queuing Theory** 

#### Outline

- The M/M/1 Queue



Queuing Theory

The M/M/1

Queue

#### The M/M/1 Queue

$$A(t) \xrightarrow{N(t)} T_n = W_n + S_n$$

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

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

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

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

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

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



Queuing Theory

Introduction

ma o da caro.

Little Theorem

PASTA Theorem

The M/N

Q-matrix

Stationary

Distribut Propertie

Example: Loss probability in a telephone switching

M/G/1 Queu

M/G/1/K Queue

Queues

#### Q-matrix

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

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

$$\begin{array}{c|c} A(t) & & \\ \lambda & & \\ \hline & N(t) & \\ \end{array}$$

• Q-matrix:

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



Queuing Theory

Introduction

\_\_\_\_\_

Theorem

The M/I

Queue

Stationary Distribution

Example: Loss probability in a telephone switchin

M/G/1 Que

M/G/1/I Oueue

Queues ir

### **Stationary Distribution**

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

$$\pi_i = (1 - \rho) \rho^i, i = 0, \dots, \infty$$

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



Queuing Theory

Properties

#### **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:  $N_s = N N_O = \rho$ NOTE:  $\pi_0 = 1 - \rho$



Stability

Queuing Theory

• 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.

 $N = \frac{\rho}{1 - \rho}$ 

Stability



Queuing Theory

Introduction

Little Theorem

PASTA

Theorem

Oueue

Queue

Q-matrix

Propertie

Stability
Example: Loss

probability in a telephone switching center

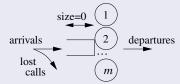
M/G/1 Que

Queue

Queues ir

#### Example: Loss probability in a telephone switching center

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





Queuing Theory

Introductioi

Little Theorem

Theorem

The M/N

Queue

Q-matrix

Properti

Example: Loss probability in a telephone switching

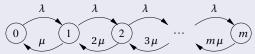
M/G/1 Que

M/G/1/ Oueue

Queues in

#### 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  $i\mu$ :





Queuing Theory

Introduction

...

Little Theorem

PASTA Theoren

The M/M/1

Queue

Distributi Properties

Example: Loss probability in a telephone switching

M/G/1 Queu

M/G/1/K Queue

Queues

#### 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:

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

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

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

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