Investigación Operativa

Operations Research

comillas.edu

Queueing Theory

Departamento de Organización Industrial

September 1, 2023

Learning outcomes

- Represent any queueing system by the birth-death process
- Analytical expressions of steady-state performance measures for the most common queueing systems
- Make optimal decisions in queueing systems

1

What QT is for?

What Queueing Theory (QT) is for?

- When does queueing happen?
  - When there is a temporary surge in demand that cannot be quickly handled with the available service capacity
  - Waiting in lines is usual at different systems (restaurants, bank branches, elevators, etc.)
- QT sets mathematical models to estimate the steady-state performance of waiting lines for different types of queueing systems
- Queue Analyst should balance optimally between costs of:
  - Investments in System Design
  - Wasted time cost of waiting and idle periods

Common Queueing Systems

❶ Business

- Bank branches
- Grocery stores, supermarkets, malls, ...
- Fast food and standard restaurants

❷ Transportation

- Load Centre (Airports, ports, ...)
- Traffic (roads, traffic lights, ...)
- Parking
- Elevators

❸ Industrial Systems

- Maintenance
- Control Systems
- Storage Centre

❹ Social services

- Hospitals
- Courts

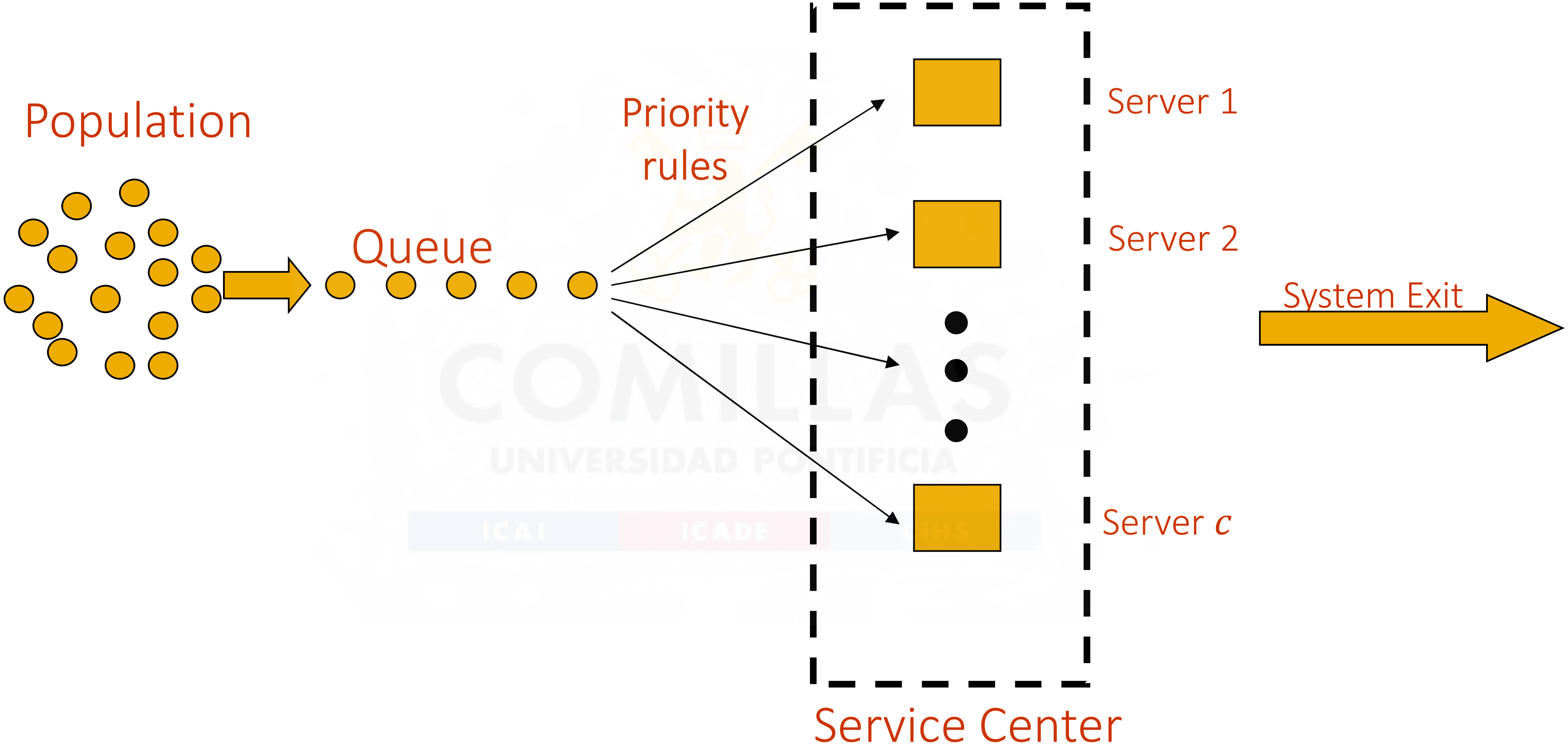
...

Clients	Service	Servers
Clients of a shop	Selling a product	Shop assistant
Clients of a bank	Banking process	Clerk
Clients of a supermarket	Payment	Cash register
Cars	Fill the tank	Pump
Cars	Fix the breakdown	Workers
Planes	Landing / takeoff	Runway
Phone calls	Conversation	Call center
Patients	Medical care	Medical doctor
Cases	Transportation	Storage robot
Trials	Trial	Judge

2

Elements of waiting lines

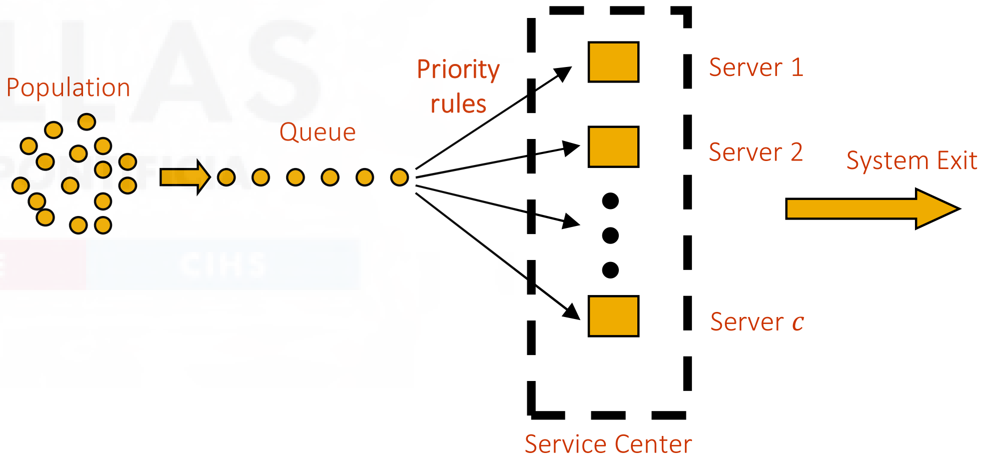
Elements of Waiting Lines



Elements of Waiting Lines (Cont'd)

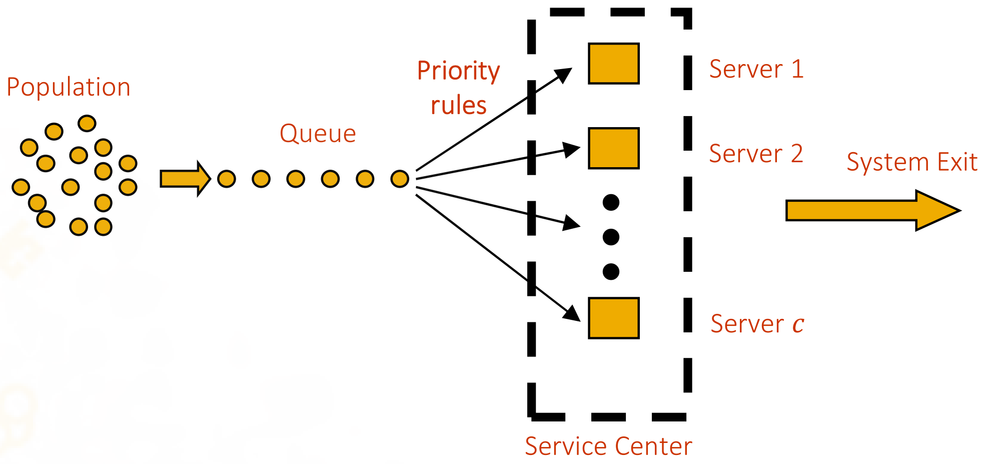
- Population:
  - Finite or infinite population
    - Usually, infinite
  - Interarrival pattern
    - Usually, exponential distribution

- Queue:
  - Infinite or finite capacity
    - Usually, infinite
  - Simultaneous queue lines



- Priority rules:
  - FIFO (First In – First Out)
  - LIFO (Last In – First Out)
  - Others:
    - By types of product
    - By customers
    - By priority

Elements of Waiting Lines (Cont'd)



Service Center:

- Number of servers:
  - Single server
  - Multiserver
- Service time pattern:
  - Such as Exponential, Gamma, and constant

Client actions:

- Balking (not to enter the waiting line)
- Reneging (first enters, before being served leaves)
- Jockeying (to change from one line to another)

Elements of Waiting Lines (Cont'd)

KENDALL'S NOTATION

Number of servers

M / M / 2 / ∞

Interarrival time pattern

Service time pattern

System capacity

- M: Exponential
- D: Constant
- Ek: Erlang
- G: General

SYMBOLS

λ Mean arrival rate (average number of customers arriving per unit of time)

μ Mean service rate (average number of customers that can be served per unit of time)

t Time

c Number of servers in the system

k System capacity

$1/\lambda$ Mean time between consecutive arrivals

$1/\mu$ Mean service time

ρ Utilization ratio ($\rho = \lambda/c\mu$)

Performance of Waiting Lines

n	system state, clients in the System (waiting or being served)	
P_n	probability of n customers in the System at any given time	
L	average number of customers in the System	$L = E[n]$
n_q	queue length, number of customers waiting in line	
L_q	average number of clients waiting in line	$L_q = E[n_q]$
t	total time spent in the System, including service time	
W	average total time spent in the System	$W = E[t]$
t_q	waiting time in the queue	
W_q	average waiting time in the queue	$W_q = E[t_q]$

Performance of Waiting Lines (Cont'd)

Little's laws:

Definition: The steady state of a Queueing System is reached when the number of customers' probability distribution is the same over time

1. The average number of customers in the System/Queue = Arrival rate x Average time spent in the System/Queue

$$L = \lambda W \qquad\qquad L_q = \lambda W_q$$

2. Average total time spent in the System = Average time waiting in line + Average service time

$$W = W_q + 1/\mu$$

3. The average number of customers in the System = The average number of customers waiting in line + The average number of customers being served

$$L = L_q + \lambda/\mu$$

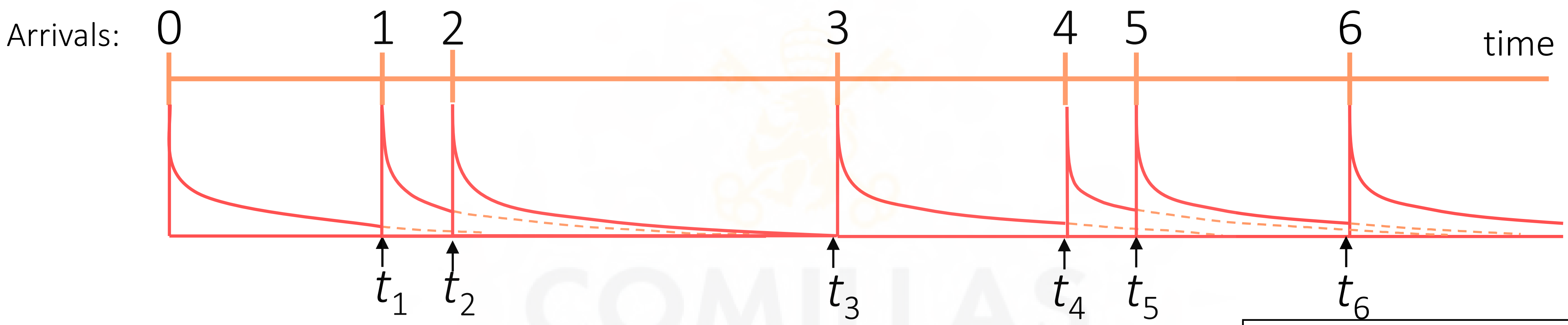
These formulas can't be used if there are different service rates.

3

Poisson process

Poisson Process

- Generally, QT assumes that the arrival process follows a Poisson process
  - A Poisson process sets interarrival time as an Exponential distribution



$t_i - t_{i-1}$: Interarrival time

Prob. of an arrival at time t

Exponential density function:

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases}$$ PDF

$$P\{T \leq t\} = 1 - e^{-\lambda t}$$ CDF

$$P\{T > t\} = e^{-\lambda}$$

$$E(T) = \frac{1}{\lambda}$$

$$Var(T) = \frac{1}{\lambda^2}$$

Arrival rate: λ

Poisson Process (Cont'd)

Properties of *Exponential* distribution as interarrival pattern

1 Lack of memory:

The arrival has the same probability to occur in a specific time interval regardless of previous spent time

$$\boxed{P\{T > t + \Delta t | T > \Delta t\} = P\{T > t\}}$$

2 Minimum of $\boldsymbol{n}$ Exponential variables:

This minimum is distributed as an Exponential distribution whose λ is:

$$\lambda = \sum_{i=1}^{n} \lambda_i$$

3 Number of arrivals during period $\boldsymbol{t}$:

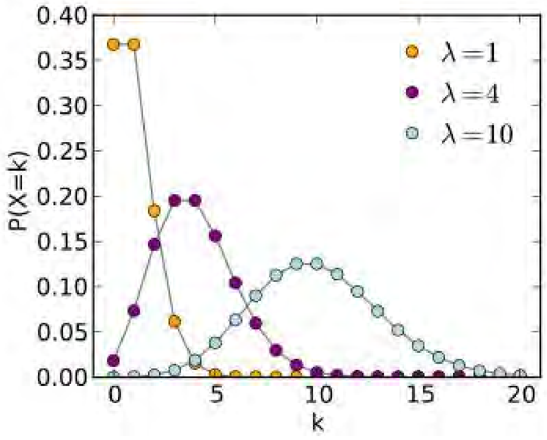
Such variable N is distributed as a Poisson variable with parameter $\lambda \boldsymbol{t}$

$$\boxed{P\{N(t) = n\} = \frac{(\lambda t)^n e^{-\lambda}}{n!}}$$

$$n = 0,1,\ldots$$

$$\boxed{P\{N(t) = 0\} = e^{-\lambda t} = P\{T > t\}}$$

$$\boxed{E[N(t)] = \lambda t}$$



Poisson Process (Cont'd)

Properties of the *Poisson* process

4 Arrival probability during a small period Δt

If Δt is sufficiently small, its arrival probability is approximately $\lambda \, \Delta t$

$$\boxed{P\{T \leq t + \Delta t | T > t\} \approx \lambda \Delta t}$$

Two or more simultaneous arrivals have a negligible probability

5 Compound m Poisson processes

Compounding arrival Poisson processes obtains another arrival Poisson process whose arrival rate λ is:

$$\lambda = \sum_{i=1}^{m} \lambda_i$$

6 Decomposing a Poisson process

A Poisson process of arrival rate λ can be decomposed into i Poisson processes with probabilities p_i and arrival rates λ_i

$$\boxed{\sum_{i=1}^{I} p_i = 1} \qquad \boxed{\lambda_i = \lambda p_i}$$

4

Birth-death process

Birth-Death Process

- Most queueing models emulate customer arrivals like *births* and customer departures like *deaths.*

- Main characteristics of this kind of modeling are:

1. System state is given by the number of customers in the System: $\boldsymbol{N(t)}$

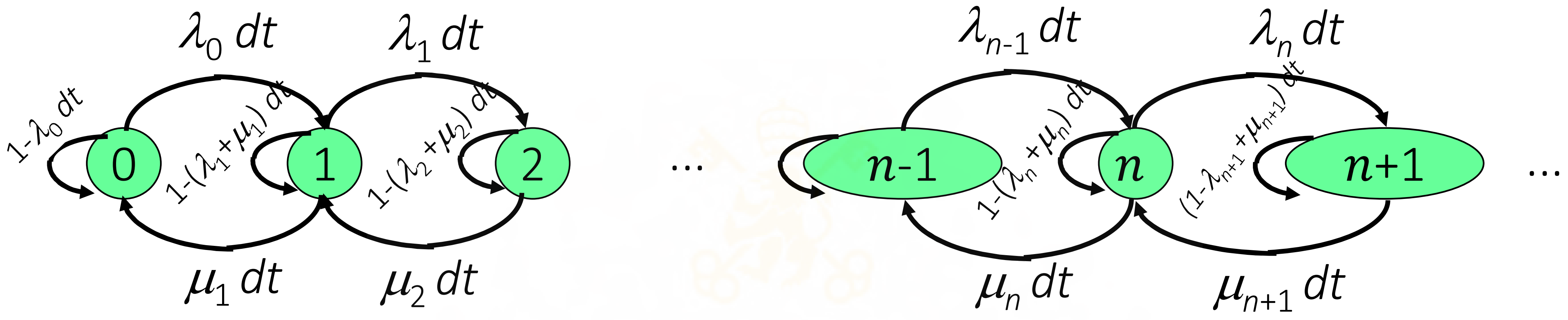
2. Given $N(t) = n$, time between two consecutive births is distributed as an Exponential probability function with rate λ_n

3. Given $N(t) = n$, time between two consecutive deaths is distributed as an Exponential probability function with rate μ_n

4. Birth and death events are independent

Birth-Death Process (Cont'd)

TRANSITION DIAGRAM



TRANSITION MATRIX

State at $t+dt$

State at t	0	1	2	3	4
0	$1-\lambda 0\,dt$	$\lambda 0\,dt$	0	0	0
1	$\mu 1\,dt$	$1-(\lambda 1+\mu 1)\,dt$	$\lambda 1\,dt$	0	0
2	0	$\mu 2\,dt$	$1-(\lambda 2+\mu 2)\,dt$	$\lambda 2\,dt$	0
3	0	0	$\mu 3\,dt$	$1-(\lambda 3+\mu 3)\,dt$	$\lambda 3\,dt$
4	0	0	0	$\mu 4\,dt$	$1-(\lambda 4+\mu 4)\,dt$

Birth-Death Process (Cont'd)

Computation of steady-state probabilities

$$\left.\begin{aligned} &\frac{d}{dt}p_0(t) = \frac{p_0(t+dt)-p_0(t)}{dt} = -\lambda_0 p_0(t) + \mu_1 p_1(t) = 0 \\ &\frac{d}{dt}p_n(t) = \lambda_{n-1}p_{n-1}(t) - (\lambda_n+\mu_n)p_n(t) + \mu_{n+1}p_{n+1}(t) = 0 \end{aligned}\right\}$$

At steady state, all derivatives are identically zero

$$\boxed{p_i(t) = p_i(t+dt) = p_i}$$

$$\left.\begin{aligned} &\lambda_0 p_0 = \mu_1 p_1 \\ &(\lambda_1+\mu_1)p_1 = \lambda_0 p_0 + \mu_2 p_2 \quad \Rightarrow \quad \lambda_1 p_1 = \mu_2 p_2 \\ &\dots \\ &(\lambda_{n-1}+\mu_{n-1})p_{n-1} = \lambda_{n-2}p_{n-2} + \mu_n p_n \quad \Rightarrow \quad \lambda_{n-1}p_{n-1} = \mu_n p_n \end{aligned}\right\}$$

$$\boxed{p_n = \frac{\lambda_0\lambda_1\lambda_2\cdots\lambda_{n-1}}{\mu_1\mu_2\cdots\mu_n}p_0}$$

Birth-Death Process (Cont'd)

$$p_n = \frac{\lambda_0 \lambda_1 \lambda_2 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \ldots \mu_n} p_0 = C_n p_0$$

$$\sum_{i=0}^{\infty} p_i = 1$$

To solve this set of equations, one equation should be removed to get n independent equations

Therefore, its solution determines all the steady-state probabilities

$$p_0 = [C_0 + C_1 + \ldots + C_n + \ldots]^{-1}$$

$$p_n = C_n p_0$$

Birth-Death Process (Cont'd)

GENERAL PERFORMANCE MEASURES

$$L = \sum_{n=0}^{\infty} n p_n$$

Average number of customers in the System

$$\mathrm{L_q} = \sum_{\mathrm{n=c+1}}^{\infty} (\mathrm{n} - \mathrm{c})\mathrm{p_n}$$

Average number of customers waiting in a c-Server System

$$\bar{\lambda} = \sum_{n=0}^{\infty} \lambda_n \, p_n$$

Average arrival rate

$$W = \frac{L}{\bar{\lambda}}$$

Average total time spent in the System

$$W_q = \frac{L_q}{\bar{\lambda}}$$

Average waiting time

5

Standard models

COMILLAS UNIVERSIDAD PONTIFICIA ICAI ICADE CIHS

Standard Models

- Standard Models emulate frequent queueing systems based on birth-death process

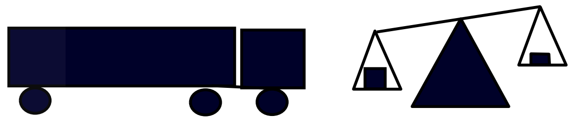
<u>Infinite</u> population

M/M/1: Single server with Exponential interarrival and service times

For example: a unique entrance *parking*

M/M/1/*k*: Same as previous with a k system capacity

For example: a truck *scale*



M/G/1: Single server with a general service time distribution

For example: a *car washing machine*

M/M/*c*: c multiserver

For example: a *petrol station*

M/M/*c*/*k*: c multiserver with finite system capacity

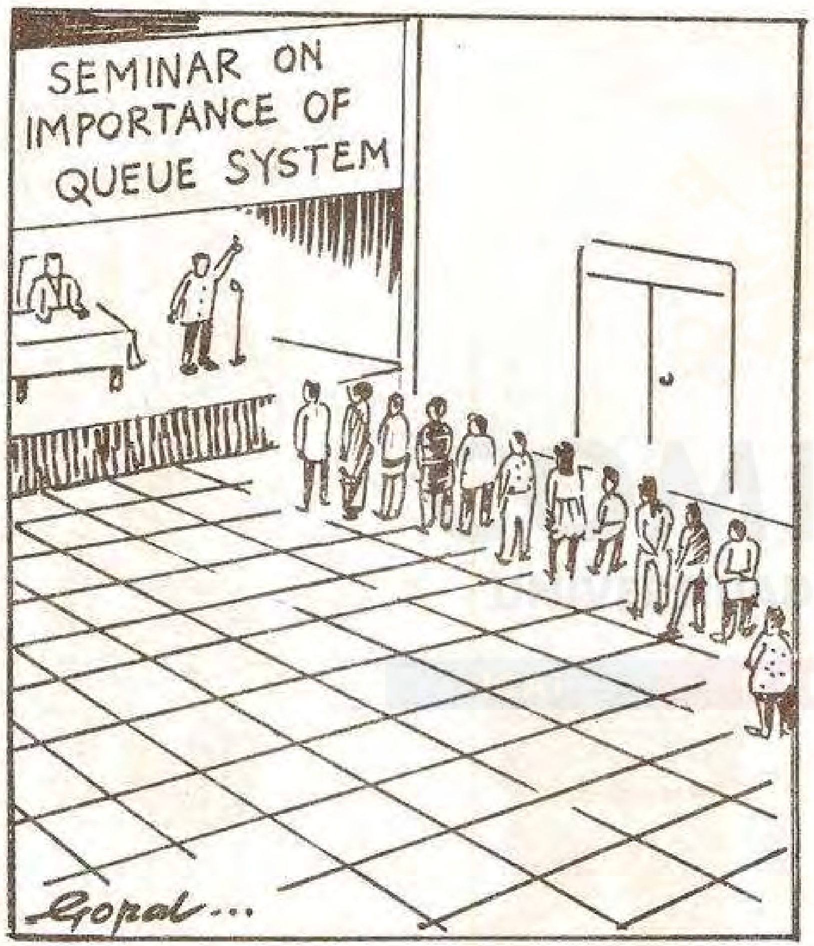
For example: a *call center* with a limited waiting line

<u>Finite</u> population (Closed Systems)

M/M/1: For example: the fax with a reduced office staff

M/M/*c*: For example: maintenance personnel for a finite number of machines

Cartoon Question?



QUESTIONS

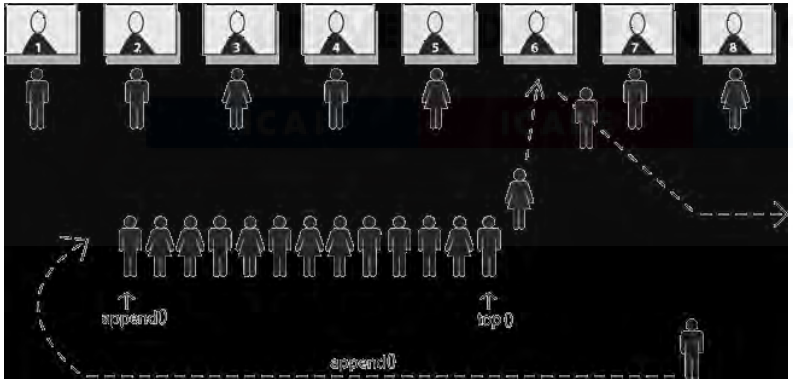
What type of model may be running?

Which one should be chosen?

What queueing system is more effective?

- 8 servers with 8 queues

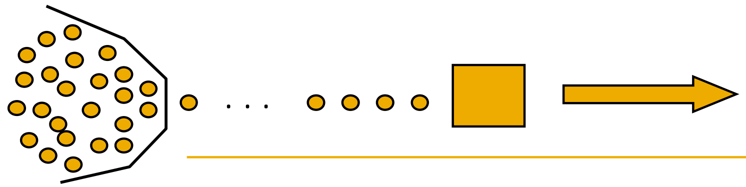
- 8 servers supplied by 1 queue



6

Infinite population

Model M/M/1



Infinite population

- Interarrival time is statistically distributed as an Exponential of rate λ
- Service time is statistically distributed as an Exponential of rate μ

- The average utilization of the server ρ is:

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

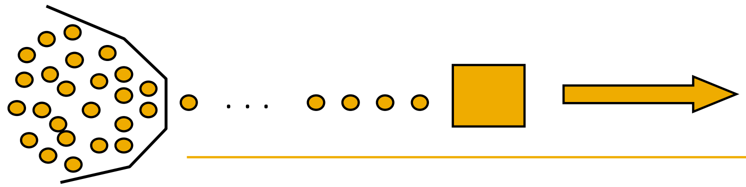
  - $\boldsymbol{\rho < 1}$ to reach a stable queueing System
  - Steady-state probabilities based on general formulas:

$$C_n = \rho^n \Rightarrow p_n = \rho^n p_0$$

$$\sum_{n=0}^{\infty} p_n = p_0 \to \sum_{n=0}^{\infty} \rho^n = p_0 \frac{1}{1-\rho} = 1 \Rightarrow p_0 = 1 - \rho$$

$$\left.\begin{array}{l} p_n = \rho^n p_0 \\ p_0 = 1-\rho \end{array}\right\} p_n = (1-\rho)\rho^n$$

Model M/M/1 (Cont'd)



PERFORMANCE MEASURES

Infinite population

$$L = \sum_{n=0}^{\infty} n p_n = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}$$

Average number of customers in the System

$$L_q = \sum_{n=2}^{\infty} (n-1) p_n = \frac{\rho^2}{1-\rho} = \rho L$$

Average number of customers waiting in line

$$W = \frac{L}{\lambda} = \frac{1}{\mu(1-\rho)}$$

Average total time spent in the System

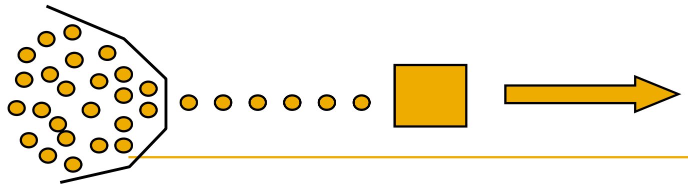
$$W_q = W - \frac{1}{\mu} = \frac{\rho}{\mu(1-\rho)}$$

Average waiting time in line

$$\overline{c} = L - L_q = \rho = 1 - p_0$$

Average server utilization

Model M/M/1/k



Infinite population

- As System capacity is limited, customers are lost if the System is full
- State dependent arrival and service rates:

$$\lambda_n = \begin{cases} \lambda & n < k \\ 0 & n \geq k \end{cases} \qquad \mu_n = \mu \quad \forall n$$

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

$$p_n = \begin{cases} \rho^n p_0 & n \leq k \\ 0 & n > k \end{cases}$$

- Steady-state probabilities:

$$(\rho \neq 1) \quad p_0 = \frac{1-\rho}{1-\rho^{k+1}}, \quad p_n = \begin{cases} \rho^n p_0 & n \leq k \\ 0 & n > k \end{cases}$$

$$(\rho = 1) \quad p_n = \frac{1}{k+1} \quad n = 0,1,\ldots,k$$

- Customer effective entrance rate:

$$\lambda_{EF} = \lambda(1 - p_k)$$

- Customer loss rate:

$$\lambda_{LOSS} = \lambda p_k$$

Model M/M/1/*k* (Cont'd)

PERFORMANCE MEASURES

Infinite population

$$L = \sum_{n=1}^{k} n\rho^n p_0$$

Average number of customers in the System

$$L_q = L - (1 - p_0)$$

Average number of customers waiting in line

$$W = \frac{L}{\lambda_{EF}}$$

Average total time spent in the System

$$W_q = W - \frac{1}{\mu}$$

Average waiting time in line

Model M/G/1

Infinite population

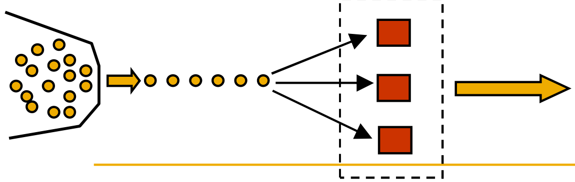
Assumptions: Exponential interarrival time and General service time whose mean and variance are:

$$\boxed{E[S] = \frac{1}{\mu}} \qquad \boxed{V[S] = \sigma^2}$$

Pollaczek-Khintchine Formula:

$$\boxed{L = \rho + \frac{\rho^2 + \lambda^2\sigma^2}{2(1-\rho)}} \qquad \boxed{\rho = \frac{\lambda}{\mu}}$$

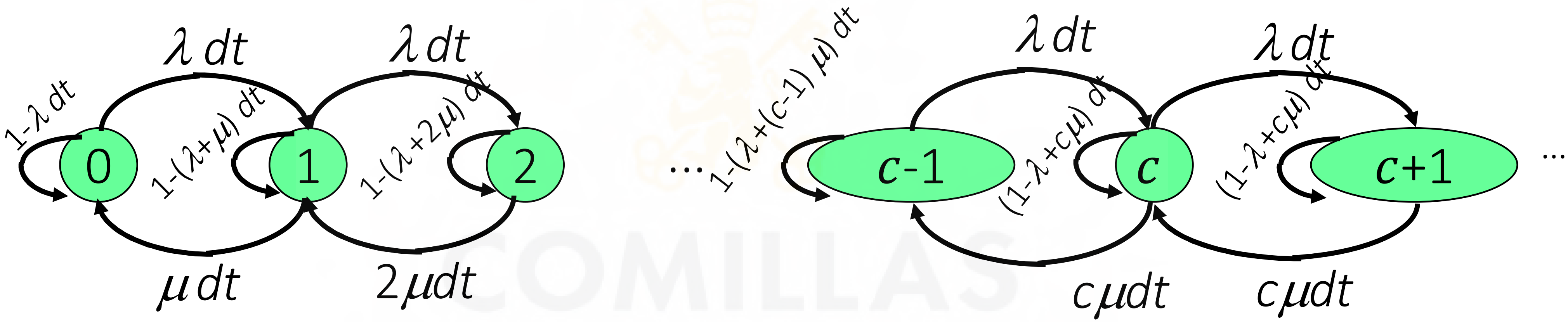
Model M/M/*c*



λ arrival rate for <u>infinite</u> population and μ service rate <u>per server</u> (***c*** servers)

TRANSITION DIAGRAM

Infinite population



TRANSITION MATRIX

State at $t + dt$

State at t

	0	1	2	…	c-1	c	c+1
0	1-λdt	λdt	0				
1	μdt	1-(λ+μ)dt	λdt	0			
2	0	2μdt	1-(λ+2μ)dt	λdt	0		
…	…	…	…	…	…	…	…
c-1					1-(λ+(c-1)μ)dt	λdt	0
c					$c\mu dt$	1-(λ+$c\mu$)dt	λdt
c+1					0	$c\mu dt$	1-(λ+$c\mu$)dt

Model M/M/*c* (Cont'd)

Infinite population

Steady State

- Steady-state Condition ($\rho < 1$)

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

- Steady-state probabilities after setting all derivatives to zero:

$$p_0 = \frac{1}{\frac{(c\rho)^c}{c!\,(1-\rho)} + \sum_{n=0}^{c-1} \frac{(c\rho)^n}{n!}}$$

$$p_n = \begin{cases} \dfrac{1}{n!}\left(\dfrac{\lambda}{\mu}\right)^n p_0 = \dfrac{c\rho}{n} p_{n-1} & 1 \le n \le c \\ \dfrac{1}{c!\,c^{n-c}}\left(\dfrac{\lambda}{\mu}\right)^n p_0 = \rho p_{n-1} & n \ge c \end{cases}$$

Model M/M/c (Cont'd)

PERFORMANCE MEASURES

Infinite population

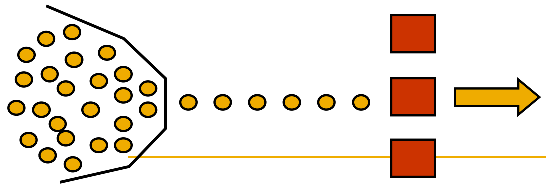
$$L = \sum_{n=0}^{\infty} n p_n = \frac{(c\rho)^c \rho}{c!\,(1-\rho)^2} p_0 + c\rho$$

Average number of customers in the System

$$L_q = L - c\rho = \frac{(c\rho)^c \rho}{c!\,(1-\rho)^2} p_0$$

Average number of customers waiting in line

Model M/M/*c*/*k*



Infinite population

- When System capacity is full, customers are lost
- State dependent arrival and service rates:

$$\lambda_n = \begin{cases} \lambda & n < k \\ 0 & n \geq k \end{cases} \qquad \mu_n = \begin{cases} n\mu & n < c \\ c\mu & n \geq c \end{cases}$$

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

- Steady-state probabilities:

$$(\rho \neq 1) \quad p_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!} p_0 & 1 \leq n \leq c \\ \dfrac{(\lambda/\mu)^n}{c!\, c^{n-c}} p_0 & c \leq n \leq k \end{cases}$$

$$\sum_{n=0}^{k} p_n = 1 \Rightarrow p_0$$

- Customer effective entrance rate:

$$\lambda_{EF} = \lambda(1 - p_k)$$

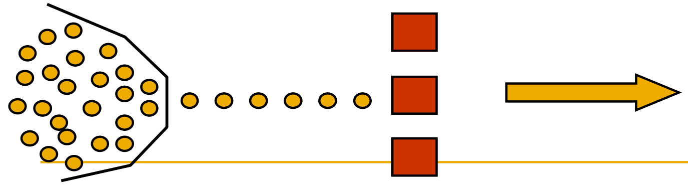
- Customer loss rate:

$$\lambda_{LOSS} = \lambda p_k$$

- Average service rate:

$$\bar{\mu} = \sum_{n=1}^{c} n\mu p_n + \sum_{n=c+1}^{k} c\mu p_n$$

Model M/M/*c*/*k* (Cont'd)



PERFORMANCE MEASURES

Infinite population

$$(\rho = 1) \quad L_q = \frac{(c\rho)^c (k-c)(k-c+1)}{2c!} p_0$$

$$(\rho \neq 1) \quad L_q = p_0 \frac{(c\rho)^c \rho}{c!\,(1-\rho)^2} \left[1 - \rho^{k-c+1} - (k-c+1)(1-\rho)\rho^{k-c}\right]$$

Average number of customers waiting in line

7

Finite population

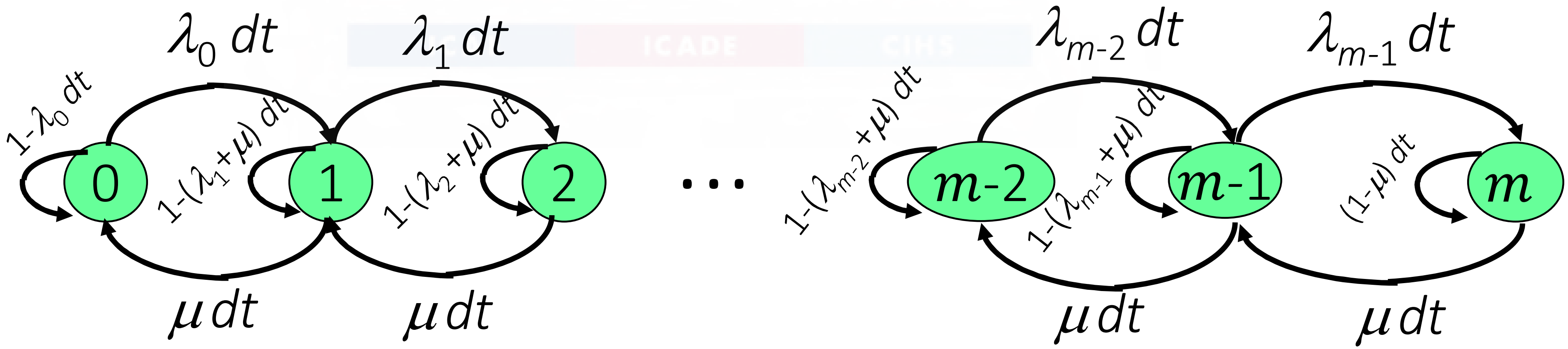
Finite Population Models (Closed Systems)

FINITE population

- Arrival rate changes based on the number of customers in the system
- Population size m
- Individual arrival rate λ
- Service rate μ

Arrival rate when n customers are in the system: $\boxed{\lambda_n = (m-n)\lambda}$

TRANSITION DIAGRAM FOR A SINGLE SERVER:



Model M/M/1 (Closed System)

Steady-State PERFORMANCE MEASURES

FINITE population

$$p_0 = \left(1 + \sum_{n=1}^{m} \frac{m!\,\rho^n}{(m-n)!}\right)^{-1}$$

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

Probability of not having customers in the system

$$p_n = \frac{m!}{(m-n)!}\rho^n p_0 = (m-n+1)\rho p_{n-1} \qquad 0 < n \leq m$$

$$p_n = 0 \qquad n > m$$

Steady-state probabilities

$$L = \sum_{n=1}^{m} n p_n = m - \frac{1-p_0}{\rho}$$

$$L_q = \sum_{n=2}^{m} (n-1) p_n = m - \frac{1+\rho}{\rho}(1-p_0)$$

$$\lambda_{\mathrm{EF}} = (m-L)\lambda$$

$$W_q = \frac{L_q}{(m-L)\lambda} = \frac{1}{\mu}\left[\frac{m}{1-p_0} - \frac{1+\rho}{\rho}\right]$$

$$W = \frac{L}{\lambda_{EF}}$$

Model M/M/c (Closed System)

FINITE population

$$\lambda_n = \begin{cases} (m-n)\lambda & 0 \le n \le m \\ 0 & n > m \end{cases}$$

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

$$p_n = \begin{cases} \binom{m}{n}\left(\frac{\lambda}{\mu}\right)^n p_0 & 1 \le n \le c \\ \binom{m}{n}\frac{n!\,(\lambda/\mu)^n}{c!\,c^{n-c}} p_0 & c \le n \le m \end{cases} \Rightarrow \sum_{n=0}^{m} p_n = 1 \Rightarrow p_0 \rightarrow \begin{cases} L \\ L_q \\ W \\ W_q \end{cases}$$