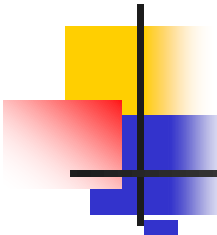
Modelos de cola

http://academia.utp.ac.pa/humberto-alvarez



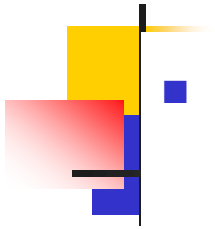
Las colas…

- Las colas son frecuentes en la vida cotidiana:
  - En un banco
  - En un restaurante de comidas rápidas
  - Al matricular en la universidad
  - Los autos en un lava-autos

Las colas…

- En general, a nadie le gusta esperar
- Cuando la paciencia llega a su límite, la gente se va a otro lugar
- Sin embargo, un servicio muy rápido tendría un costo muy elevado
- Es necesario encontrar un balance adecuado

¿Qué son las colas o filas?



- Una cola es una colección ordenada de ítems donde la adición de nuevos ítems tiene lugar en uno de los extremos, denominado "final", y la remoción de ítems existentes ocurre en el otro extremo, comúnmente llamado "frente".
- Un elemento ingresa a la cola por el final y espera hasta el momento que un ítem sea eliminado para avanzar hacia el frente.
- Generalmente, el ítem más recientemente agregado en la cola debe esperar al final de la colección. El ítem que ha permanecido más tiempo en la colección está en el frente.

Costos de un sistema de colas

1. Costo de espera: Es el costo para el cliente al esperar
   - Representa el costo de oportunidad del tiempo perdido
   - Un sistema con un bajo costo de espera es una fuente importante de competitividad

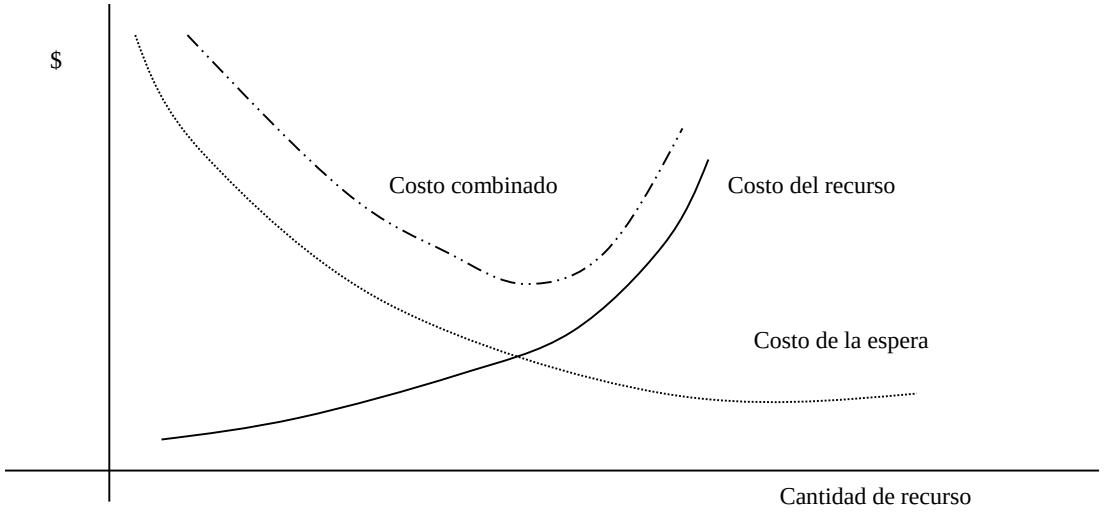
Costos de un sistema de colas

2. Costo de servicio: Es el costo de operación del servicio brindado

- Es más fácil de estimar
  - El objetivo de un sistema de colas es encontrar el sistema del costo total mínimo

Objetivo

- Analizar el costo de proveer un servicio o asignar un recurso y el costo o inconveniencia de la espera



Psicología de las colas:

- El tiempo desocupado parece más largo que el ocupado
- Las esperas previas y posteriores al proceso parecen más largas que las que se producen dentro del mismo
- La ansiedad hace que las esperas parezcan más largas
- Las esperas inciertas son más largas que las conocidas
- Las esperas no explicadas parecen más largas que las explicadas
- Las esperas injustas son más largas que las equitativas
- Cuanto más valioso sea el servicio, más espera la gente
- Las esperas en solitario parecen más largas que acompañado
- Las esperas incómodas parecen más largas que las cómodas
- Las esperas no familiares parecen más largas que las familiares

Teoría de colas

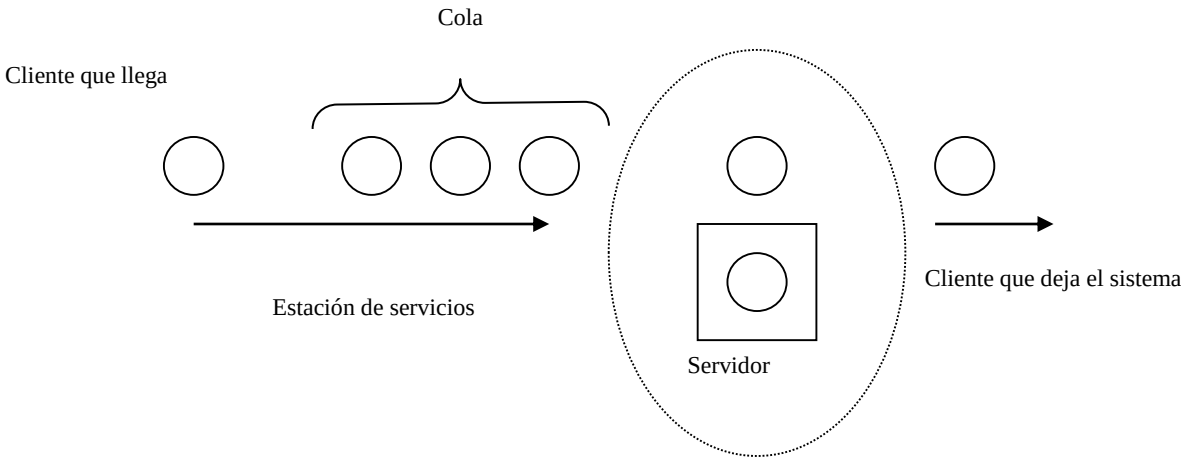
- Una cola es una línea de espera
- La teoría de colas es un conjunto de modelos matemáticos que describen sistemas de líneas de espera particulares
- El objetivo es encontrar el estado estable del sistema y determinar una capacidad de servicio apropiada

Teoría de colas

- Existen muchos sistemas de colas distintos
- Algunos modelos son muy especiales
- Otros se ajustan a modelos más generales
- Se estudiarán ahora algunos modelos comunes
- Otros se pueden tratar a través de la simulación

Definición

- Una **cola** es una línea de "clientes" que **esperan** por un servicio o recurso proveído por un **servidor**. Estos servidores se pueden considerar **estaciones** individuales donde cada cliente recibe un servicio específico.



Sistemas de colas: modelo básico

- Si cuando el cliente llega no hay nadie en la cola, pasa de una vez a recibir el servicio
- Si no, se une a la cola
- Es importante señalar que la cola no incluye a quien está recibiendo el servicio

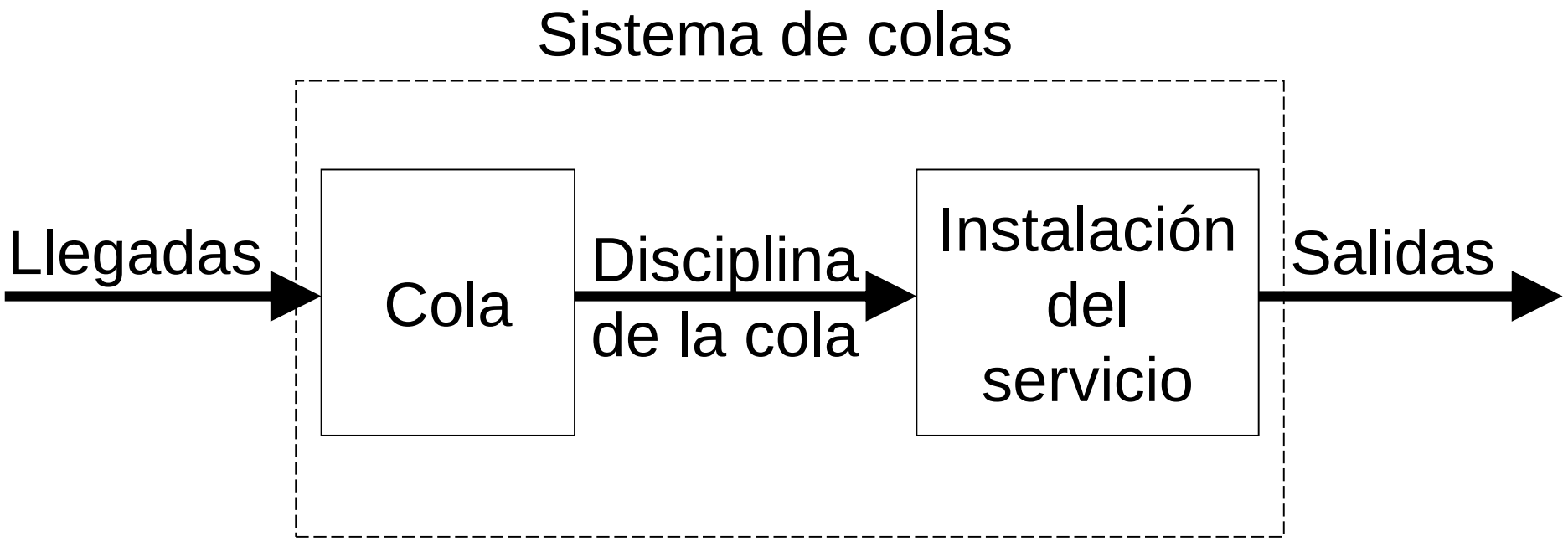
Sistemas de colas: modelo básico

- Las llegadas van a la instalación del servicio de acuerdo con la disciplina de la cola
- Disciplina: PEPS, prioridades, tiempo de procesamiento, fecha de entrega
- Pero pueden haber otras reglas o colas con prioridades

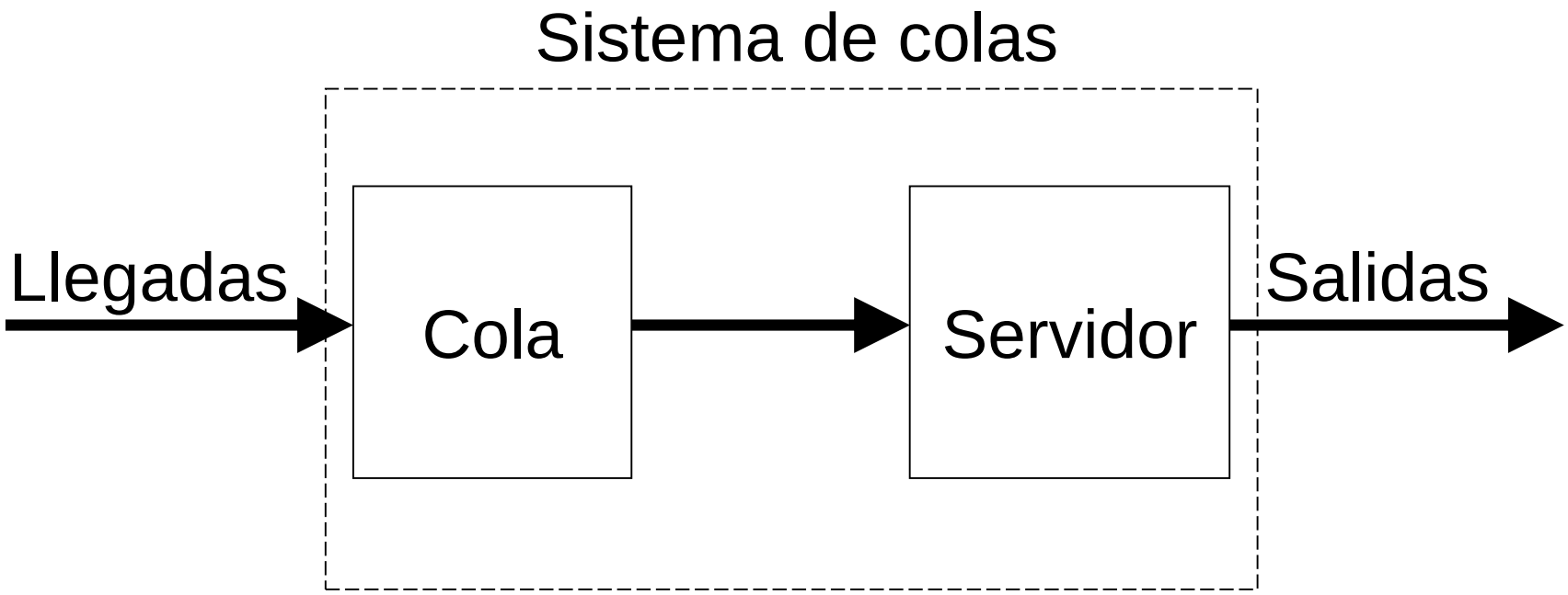
Configuraciones de colas

- Tienen que ver con el número de colas, su ubicación, sus requerimientos espaciales y el comportamiento del cliente

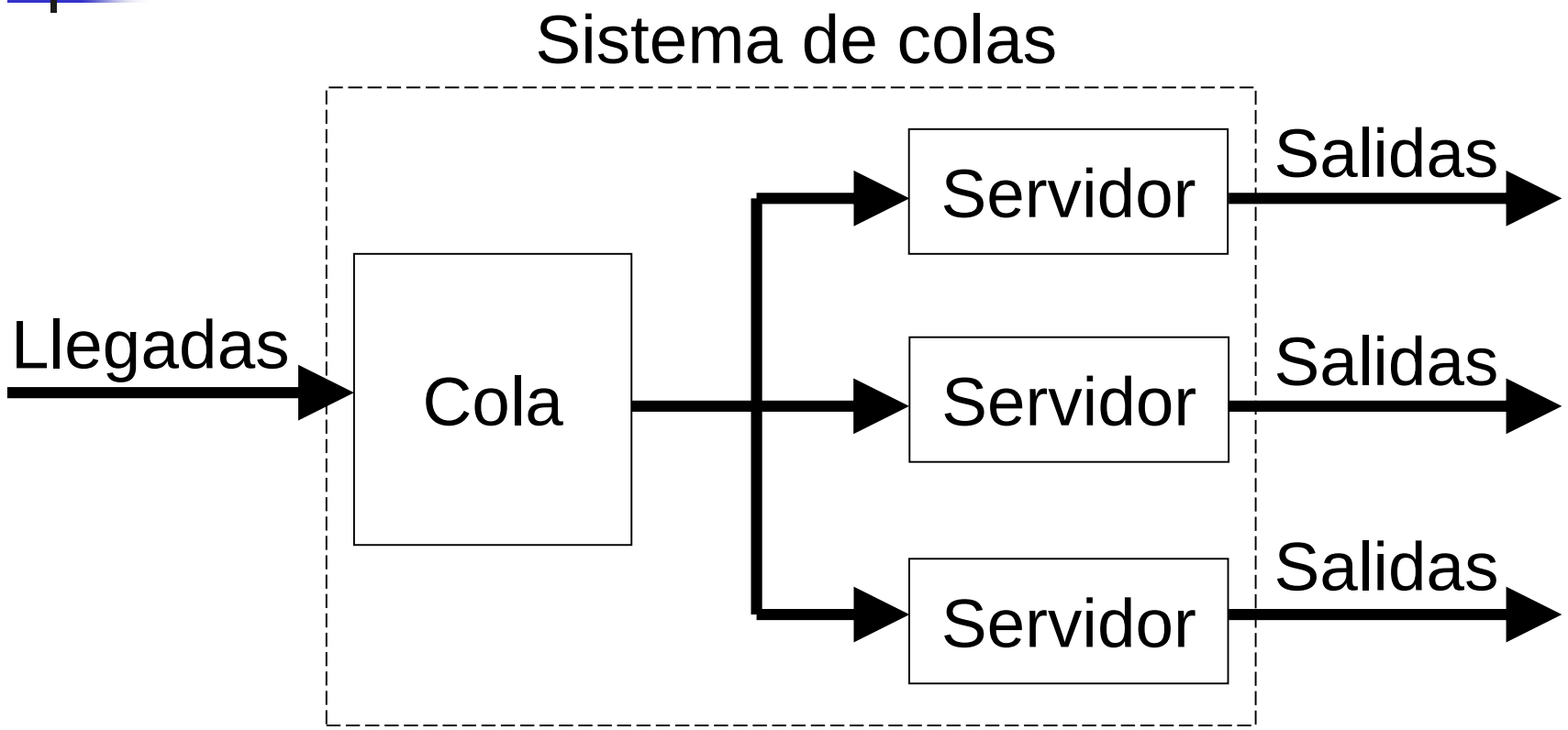
Sistemas de colas: modelo básico



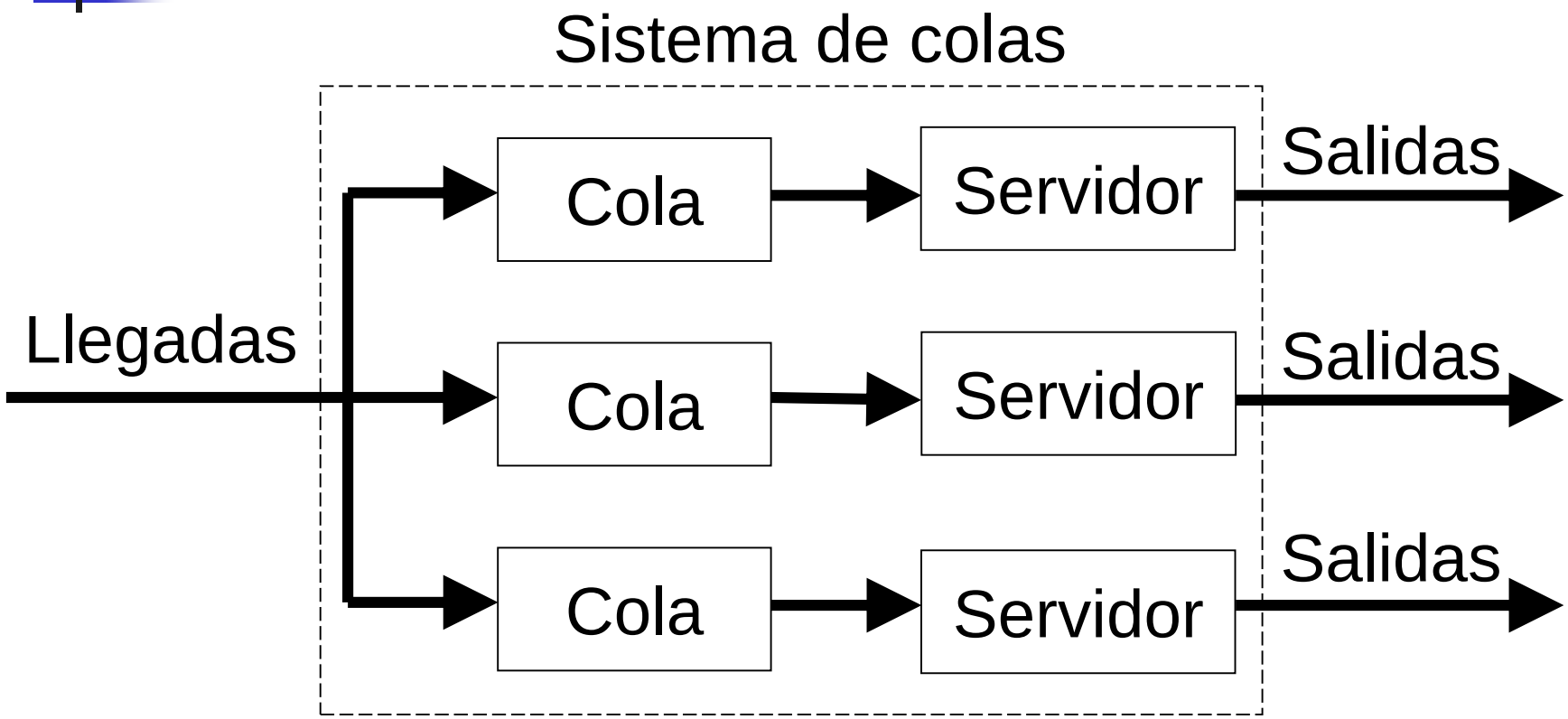
Estructuras típicas de sistemas de colas: una línea, un servidor



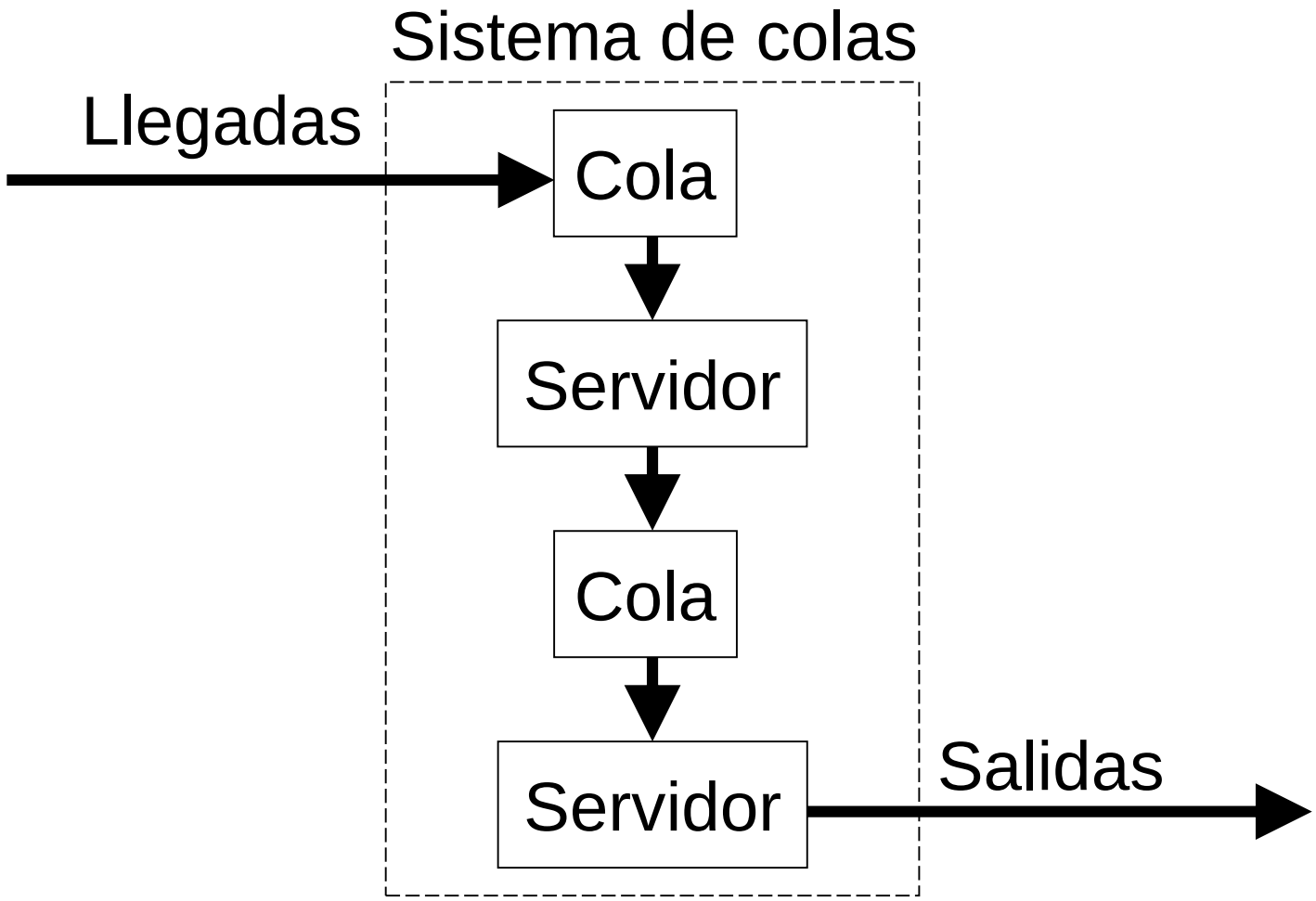
Estructuras típicas de sistemas de colas: una línea, múltiples servidores



Estructuras típicas de colas: varias líneas, múltiples servidores



Estructuras típicas de colas: una línea, servidores secuenciales



Sistemas de colas: Las llegadas

- El tiempo que transcurre entre dos llegadas sucesivas en el sistema de colas se llama tiempo entre llegadas
- El tiempo entre llegadas tiende a ser muy variable
- El número esperado de llegadas por unidad de tiempo se llama tasa media de llegadas (λ)

Sistemas de colas: Las llegadas

- El tiempo esperado entre llegadas es $1/\lambda$
- Por ejemplo, si la tasa media de llegadas es $\lambda = 20$ clientes por hora
- Entonces el tiempo esperado entre llegadas es $1/\lambda = 1/20 = 0.05$ horas o 3 minutos

Sistemas de colas: Las llegadas

- Además es necesario estimar la distribución de probabilidad de los tiempos entre llegadas
- Generalmente se supone una distribución de Poisson
- Esto depende del comportamiento de las llegadas

Sistemas de colas: Las llegadas - Distribución de Poisson

- Es una distribución discreta empleada con mucha frecuencia para describir el patrón de las llegadas a un sistema de colas
- Para tasas medias de llegadas pequeñas es asimétrica y se hace más simétrica y se aproxima a la binomial para tasas de llegadas altas

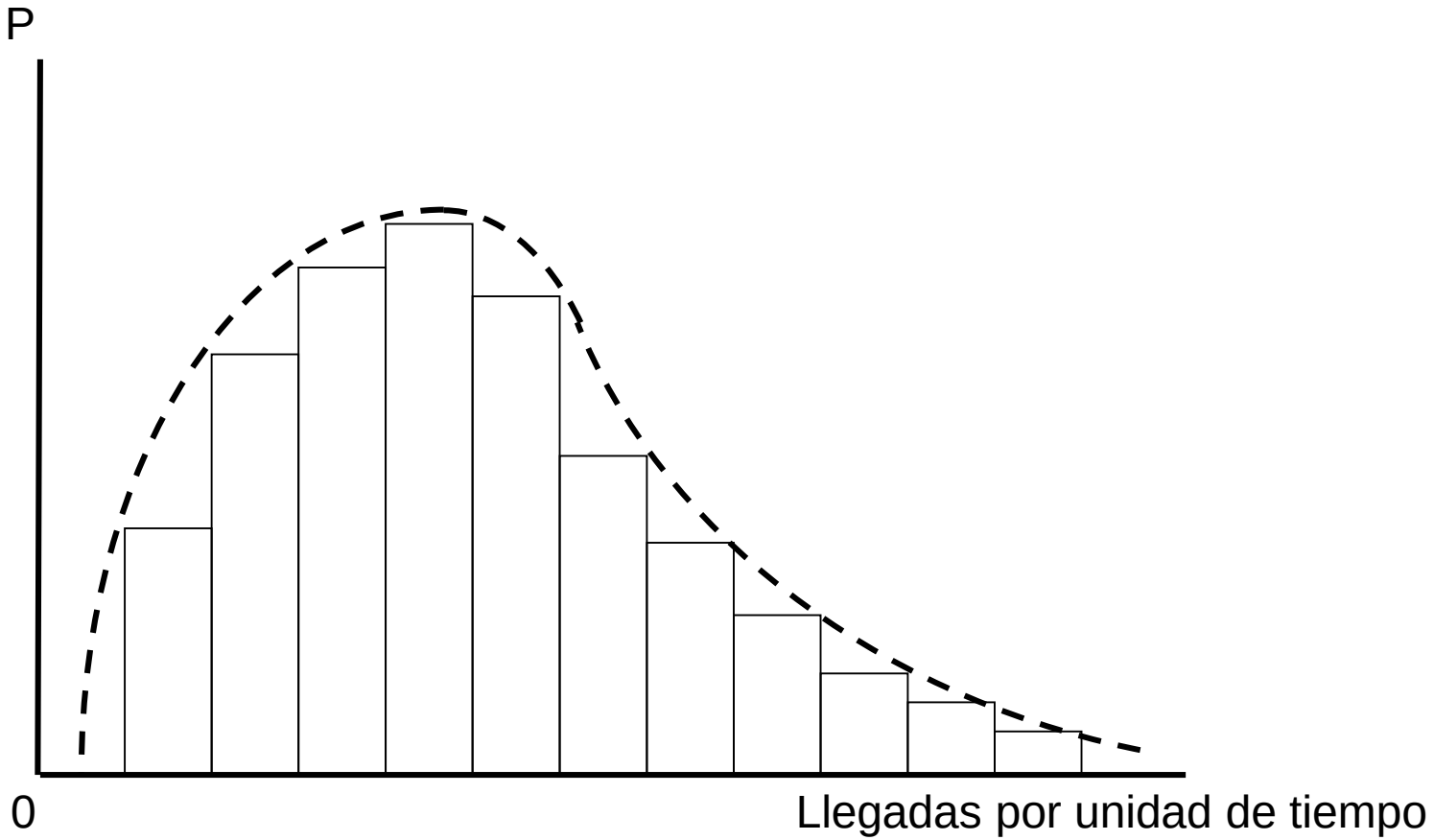
Sistemas de colas: Las llegadas - Distribución de Poisson

- Su forma algebraica es:

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- Donde:
  - ***P(k)*** : probabilidad de k llegadas por unidad de tiempo
  - λ : tasa media de llegadas

Sistemas de colas: Las llegadas - Distribución de Poisson



Sistemas de colas: La cola

- El número de clientes en la cola es el número de clientes que esperan el servicio
- El número de clientes en el sistema es el número de clientes que esperan en la cola más el número de clientes que actualmente reciben el servicio

Sistemas de colas: La cola

- La capacidad de la cola es el número máximo de clientes que pueden estar en la cola
- Generalmente se supone que la cola es infinita
- Aunque también la cola puede ser finita

Sistemas de colas: La cola

- La disciplina de la cola se refiere al orden en que se seleccionan los miembros de la cola para comenzar el servicio
- La más común es PEPS: primero en llegar, primero en servicio
- Puede darse: selección aleatoria, prioridades, UEPS, entre otras.

Sistemas de colas: El servicio

- El servicio puede ser brindado por un servidor o por servidores múltiples
- El tiempo de servicio varía de cliente a cliente
- El tiempo esperado de servicio depende de la tasa media de servicio (μ)

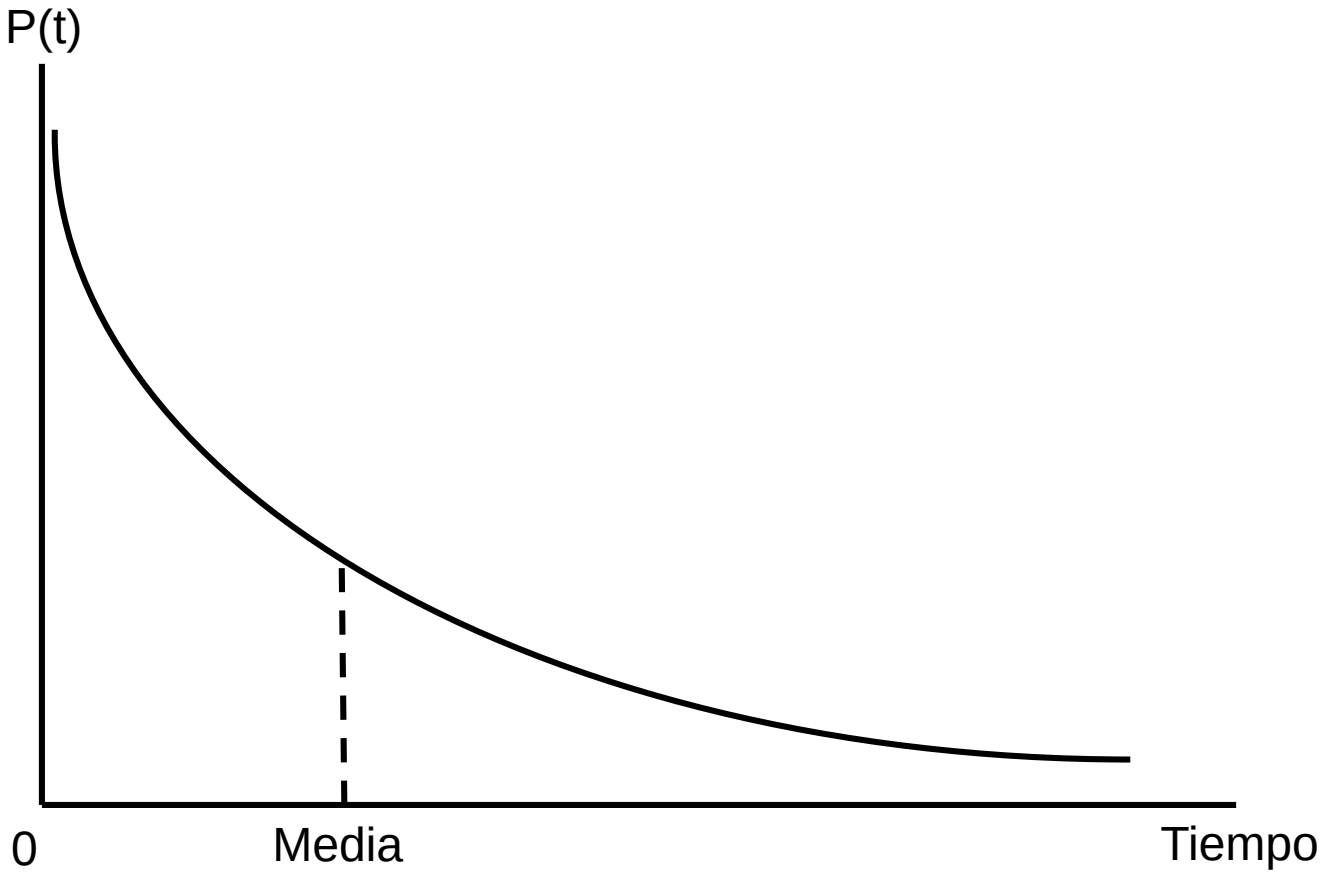
Sistemas de colas: El servicio

- El tiempo esperado de servicio equivale a $1/\mu$
- Por ejemplo, si la tasa media de servicio es de 25 clientes por hora
- Entonces el tiempo esperado de servicio es $1/\mu = 1/25 = 0.04$ horas, o 2.4 minutos

Sistemas de colas: El servicio

- Es necesario seleccionar una distribución de probabilidad para los tiempos de servicio
- Hay dos distribuciones que representarían puntos extremos:
  - La distribución exponencial (σ=media)
  - Tiempos de servicio constantes (σ=0)

Sistemas de colas: El Servicio – Distribución exponencial



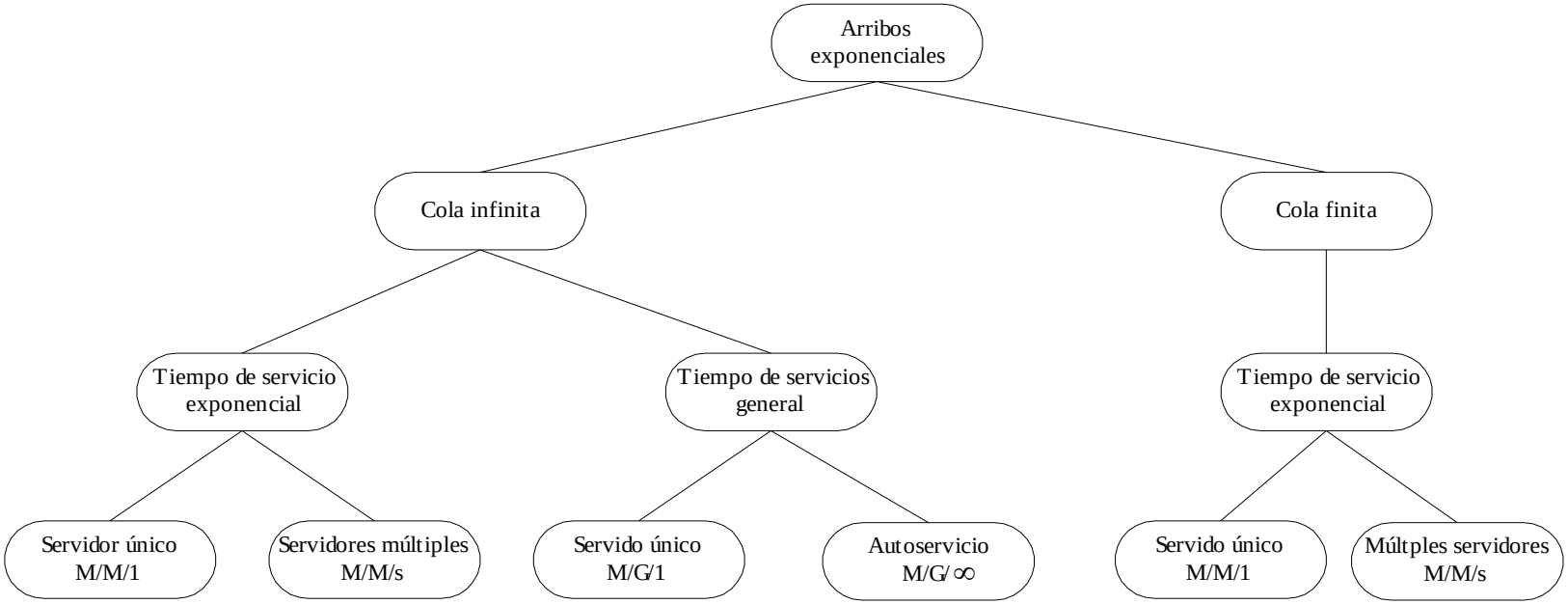
Sistemas de colas: El servicio – Distribución exponencial

- La distribución exponencial supone una mayor probabilidad para tiempos entre llegadas pequeños
- En general, se considera que las llegadas son aleatorias
- La última llegada no influye en la probabilidad de llegada de la siguiente

Características

- La notación Kendall A/B/c
  - A: Naturaleza de las llegadas – M, D
  - B: Naturaleza de los procesos – M, D, G
  - c: Número de servidores
    - M: distribución exponencial
    - D: distribución degenerada
    - G: distribución general
    - E_k: distribución Erlang

Taxonomía


Arribos exponenciales
Cola infinita
Cola finita
Tiempo de servicio exponencial
Tiempo de servicios general
Tiempo de servicio exponencial
Servidor único M/M/1
Servidores múltiples M/M/s
Servido único M/G/1
Autoservicio M/G/∞
Servido único M/M/1
Múltples servidores M/M/s

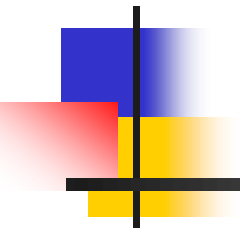
Terminología de los modelos

- λ : Tasa promedio de arribos
- μ : Tasa promedio de servicio o capacidad del servidor
- ρ : Factor de intensidad o utilización de un servidor $\frac{\lambda}{\mu}$, $0<\rho<1$
- U : Intensidad de tráfico del sistema $\frac{\lambda}{s\mu}$, también llamado factor de utilización

  mide el porcentaje de tiempo que el sistema, con **s** servidores está ocupado.
- N : Número máximo de clientes o transacciones permitidas en el sistema
- P_0 : Probabilidad de que no exista ninguna transacción en el sistema
- P_n : Probabilidad de que el siguiente arribo tenga que esperar
- P_n : Probabilidad de que existan exactamente n transacciones en el sistema
- L_q : Número promedio de clientes esperando en la cola
- L : Número promedio de transacciones en el sistema
- L_b : Número promedio de transacciones en la cola para un sistema ocupado
- W : Tiempo promedio total en el sistema
- W_q : Tiempo promedio de espera en la cola
- W_b : Tiempo de espera promedio en cola para un sistema ocupado

Suposiciones en los modelos de colas

- El sistema bajo estudio está en estado estable
- La información es conocida
- Los sistemas son estáticos



Algunos modelos típicos

El Modelo M/M/1

- El tamaño de la cola es infinitamente grande
- Todos los arribos puedan ser admitidos al sistema y esperar a ser atendidos

$$P_0 = 1 - \lambda/\mu \qquad W_q = W - \frac{1}{\mu}$$

$$P(n \geq k) = (\lambda/\mu)^k \qquad L = \lambda W$$

$$P_n = P_0(\lambda/\mu)^n \qquad L_q = \lambda W_q$$

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

El Modelo M/M/1 con capacidad finita

- El tamaño de la cola es finito y conocido
- Todos los arribos no puedan ser admitidos al sistema y se descartan

$$P_0 = \begin{cases} \dfrac{1-\rho}{1-\rho^{N+1}} & \text{para } \lambda \neq \mu \\ \dfrac{1}{N+1} & \text{para } \lambda = \mu \end{cases} \qquad L = \begin{cases} \dfrac{\rho}{1-\rho} - \dfrac{(N+1)\rho^{N+1}}{1-\rho^{N+1}} & \text{para } \lambda \neq \mu \\ \dfrac{N}{2} & \text{para } \lambda = \mu \end{cases}$$

$$P(n>0) = 1 - P_0 \qquad L_q = L - (1-P_0)$$

$$P(n \leq N) = P_0 \rho^n \qquad L_b = \frac{L_q}{1-P_0}$$

$$W = \frac{L_q}{\lambda(1-P_N)} + \frac{1}{\mu} \qquad W_q = W - \frac{1}{\mu} \qquad W_b = \frac{W_q}{1-P_0}$$

El Modelo M/G/1

- El tamaño de la cola es infinito
- Servidor no está exponencialmente distribuido

$$P_0 = 1 - \rho \qquad L_q = \frac{\rho^2 + \lambda^2 \sigma^2}{2(1-\rho)}$$

$$P_w = \rho \qquad L = L_q + \rho$$

$$W_q = \frac{L_q}{\lambda} \qquad W = W_q + \frac{1}{\mu} \qquad W_b = \frac{L_q}{\lambda}$$

Donde σ^2 es la varianza del tiempo de servicio

El Modelo M/G/∞

- Modelo de un sistema de autoservicio
- Servidor no está exponencialmente distribuido

$$P_n = \frac{e^{-\rho}}{n!}\rho^n \quad \text{para n} \geq 0$$

$$\mathrm{L} = \rho$$

$$\mathrm{W} = \frac{1}{\mu}$$

Ejemplo 1

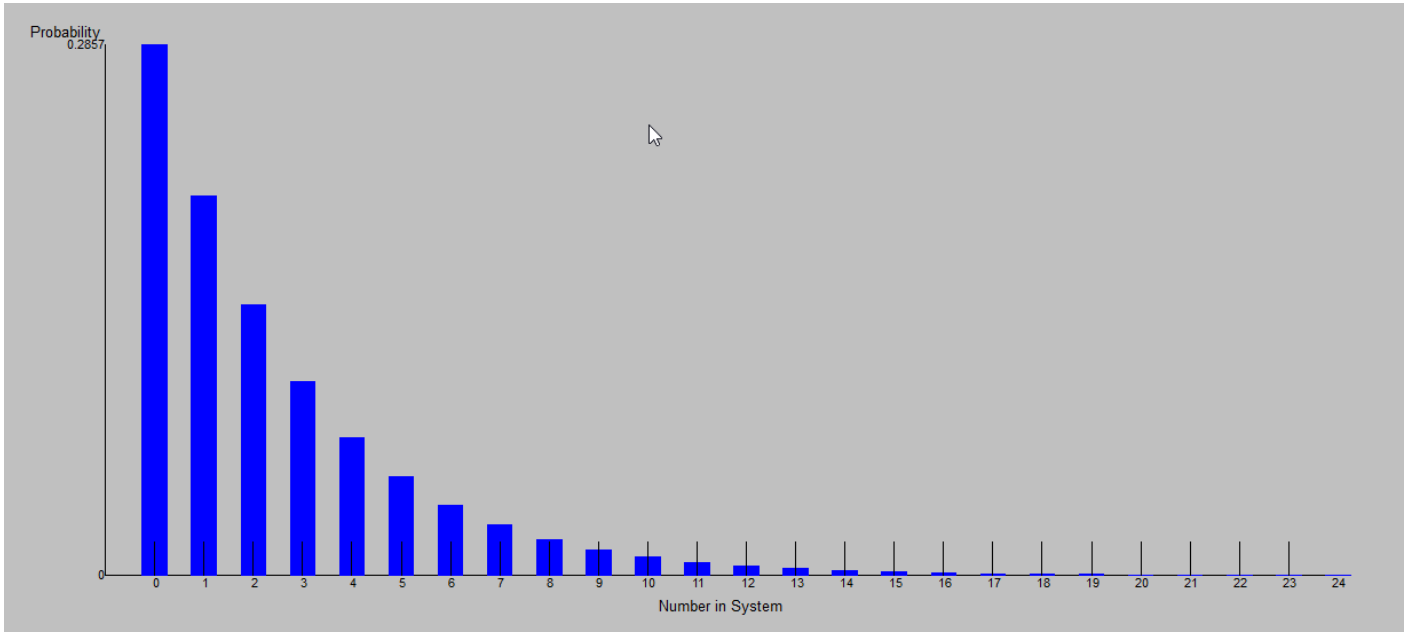
- Un centro de servicio al cliente tiene solamente un técnico que atiende llamadas. Las llamadas llegan de manera aleatoria con una media de de cinco llamadas por hora y siguen una distribución de Poisson. El técnico puede atender y servir las solicitudes a una tasa de siete llamadas por hora, distribuidas de manera exponencial.
- Existen quejas de que los clientes son puestos en espera antes que atiendan las llamadas por lo que se quiere tanto tanto el comportamiento del sistema, como la cantidad óptima de técnicos necesarios.

Definiendo los parámetros del modelo

Parámetro	Valor
Servidores	1
μ	7 por hora
λ	5 por hora

La solución en QM

Parameter	Value	Parameter	Value	Minutes	Seconds
Single-channel system		Average server utilization	0.7143		
Arrival rate(lambda)	5.	Average number in the queue(Lq)	1.7857		
Service rate(mu)	7.	Average number in the system(Ls)	2.5		
Number of servers	1.	Average time in the queue(Wq)	0.3571	21.4286	1,285.714
		Average time in the system(Ws)	0.5	30.	1,800.



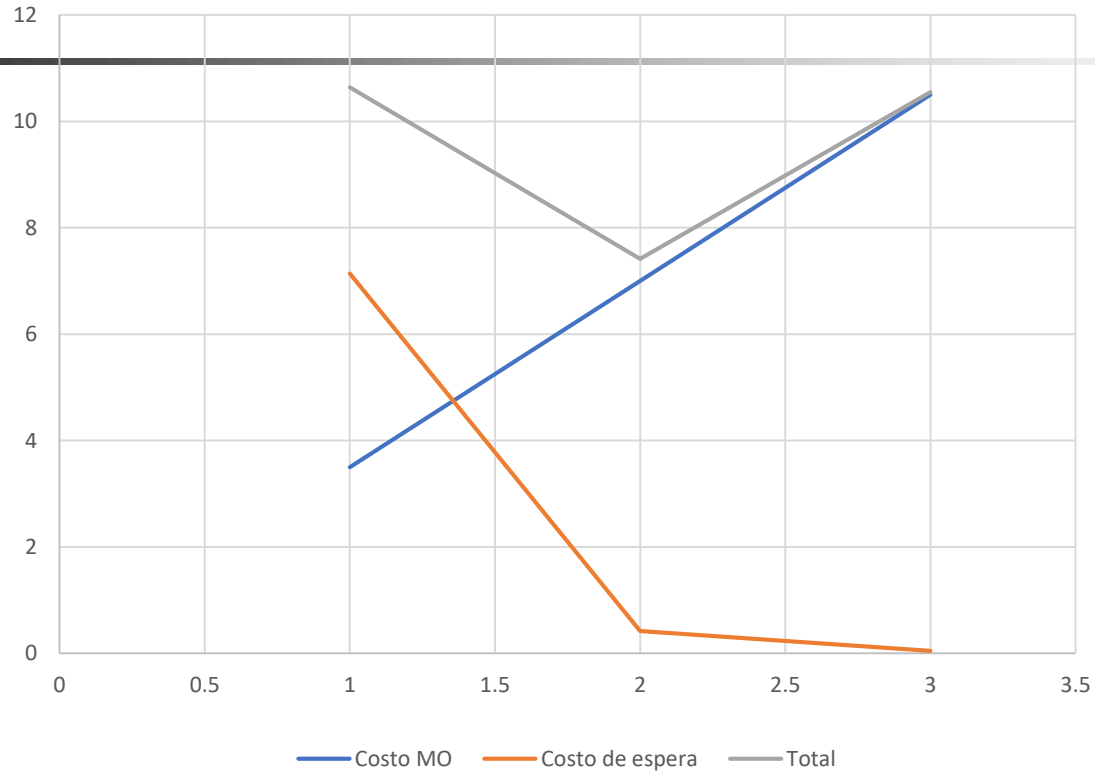
Otros supuestos para el análisis óptimo

- Costo de un técnico B/.3.50/hora
- Costo estimado de espera de un cliente B/.4.00/hora

Parameter	Value	Parameter	Value	Minutes	Seconds
Single-channel system		Average server utilization	0.7143		
Arrival rate(lambda)	5.	Average number in the queue(Lq)	1.7857		
Service rate(mu)	7.	Average number in the system(Ls)	2.5		
Number of servers	1.	Average time in the queue(Wq)	0.3571	21.4286	1,285.714
Server cost $/time	3.5	Average time in the system(Ws)	0.5	30.	1,800.
Waiting cost $/time	4.	Cost (Labor + # waiting*wait cost)	10.6429		
		Cost (Labor + # in system*wait cost)	13.5		

Parameter	Value	Parameter	Value	Minutes	Seconds
M/M/s		Average server utilization	0.3571		
Arrival rate(lambda)	5.	Average number in the queue(Lq)	0.1044		
Service rate(mu)	7.	Average number in the system(Ls)	0.8187		
Number of servers	2.	Average time in the queue(Wq)	0.0209	1.2531	75.188
Server cost $/time	3.5	Average time in the system(Ws)	0.1637	9.8246	589.4737
Waiting cost $/time	4.	Cost (Labor + # waiting*wait cost)	7.4177		
		Cost (Labor + # in system*wait cost)	10.2749		

Parameter	Value	Parameter	Value	Minutes	Seconds
M/M/s		Average server utilization	0.2381		
Arrival rate(lambda)	5.	Average number in the queue(Lq)	0.0122		
Service rate(mu)	7.	Average number in the system(Ls)	0.7264		
Number of servers	3.	Average time in the queue(Wq)	0.0024	0.1459	8.7535
Server cost $/time	3.5	Average time in the system(Ws)	0.1453	8.7173	523.0392
Waiting cost $/time	4.	Cost (Labor + # waiting*wait cost)	10.5486		
		Cost (Labor + # in system*wait cost)	13.4058		


12
10
8
6
4
2
0
0
0.5
1
1.5
2
2.5
3
3.5
Costo MO
Costo de espera
Total

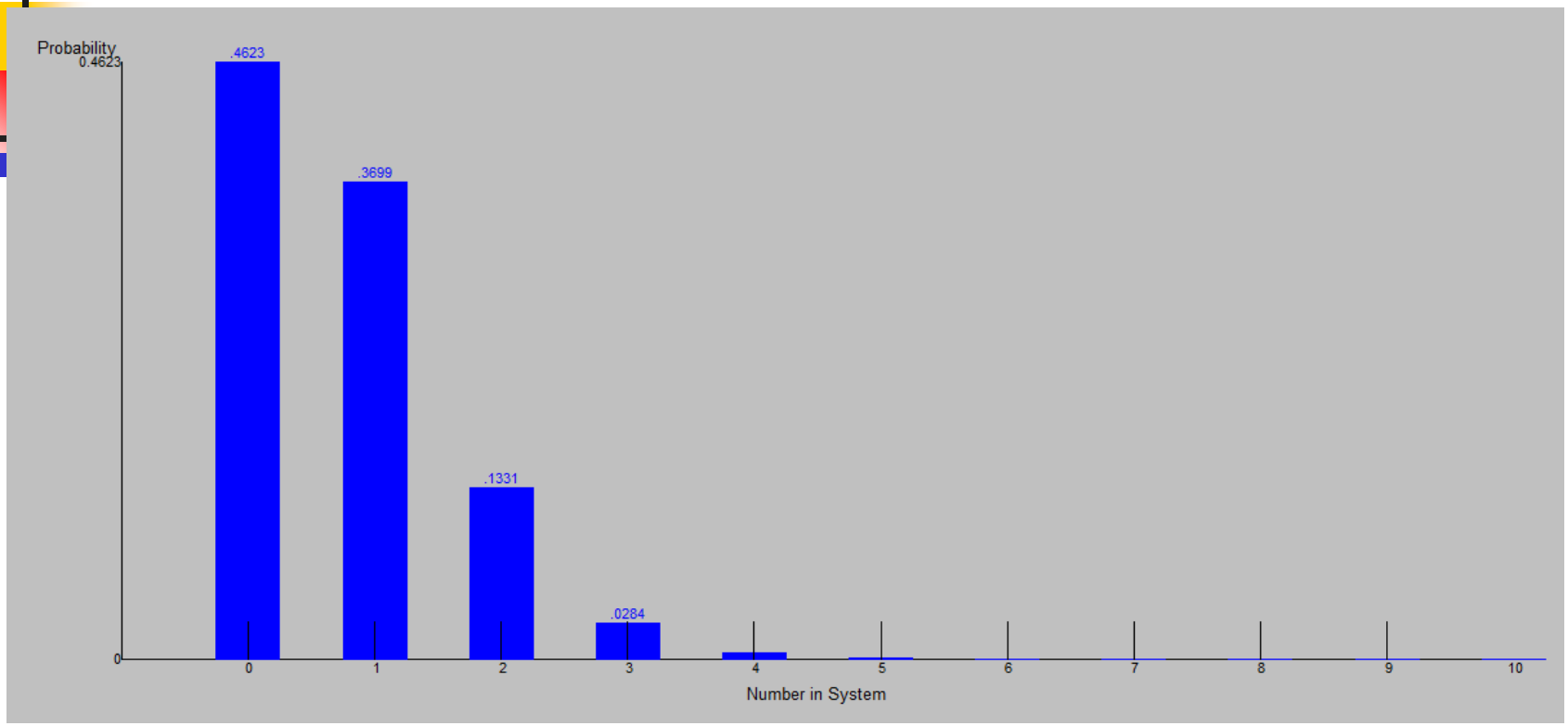
Ejemplo 2

- La División Industrial de la ACP tiene 10 máquinas idénticas que utiliza en sus talleres. Las máquinas se dañan con un MTBF de 1 daño cada 100 horas. La ACP a estimado un costo de B/.100 por cada hora que una máquina no trabaja.
- La división tiene un mecánico asignado para reparar las máquinas cada vez que se dañan, siendo el MTBR de 8 horas cada vez. Si al técnico se le pagan B/.20 por hora, analizar la factibilidad del sistema.

Parámetros del Modelo

- En este caso se conoce el MTBF ($1/\lambda$) que es el tiempo medio entre fallas y el MTBR ($1/\mu$) o tiempo medio entre reparaciones.

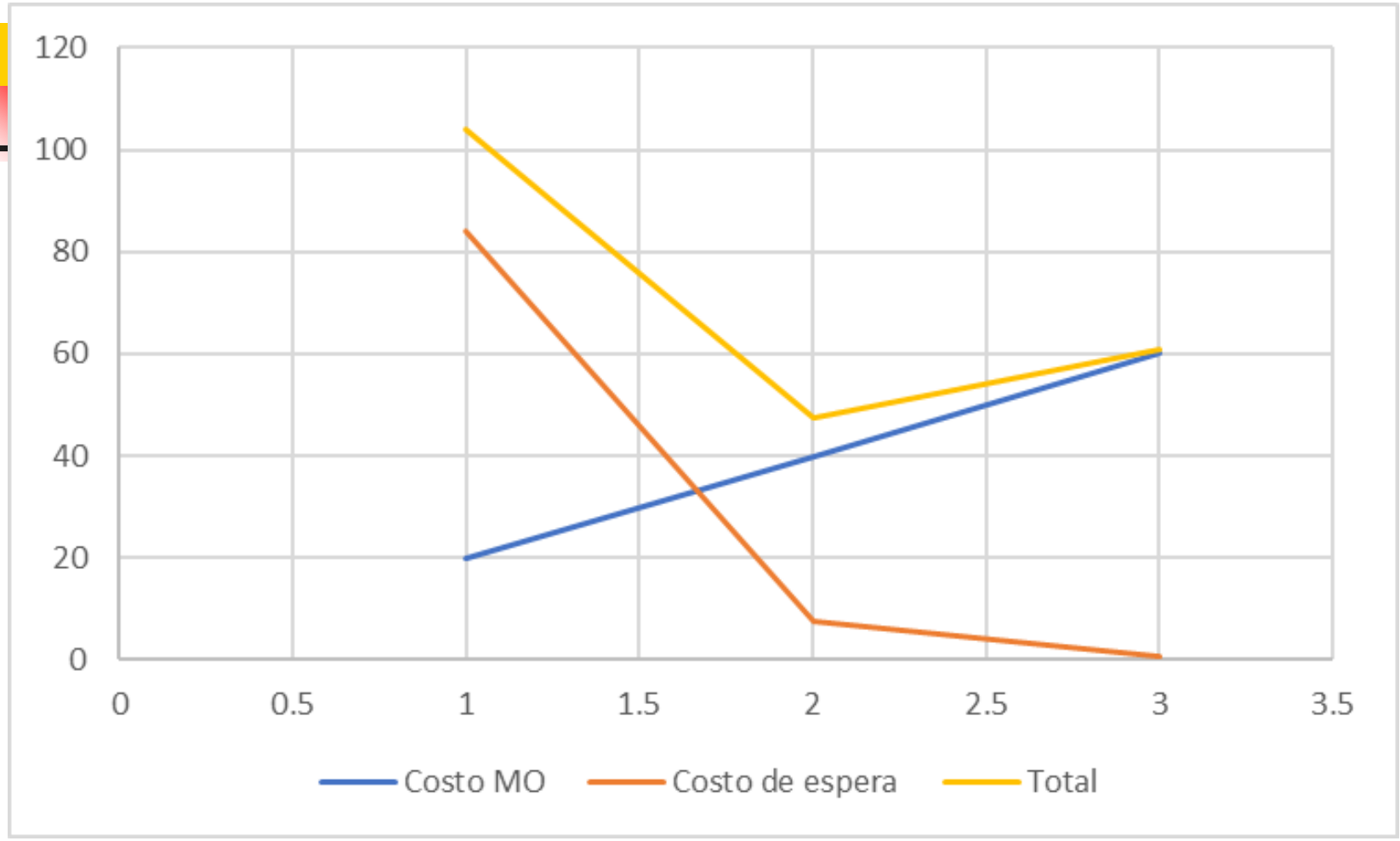
Parámetro	Valor
MTBF	100 horas
λ	0.01 por hora
MTBR	8 horas
μ	0.125 horas
Servidores	1
Máquinas	10


Probability
0.4623
0
.4623
.3699
.1331
.0284
0
1
2
3
4
5
6
7
8
9
10
Number in System

Parameter	Value	Parameter	Value	Minutes	Seconds
M/M/s with a finite population		Average server utilization	0.678		
Arvl rt PER CUSTOMER	0.01	Average number in the queue(Lq)	0.8463		
Service rate(mu)	0.125	Average number in the system(Ls)	1.5244		
Number of servers	1.	Average time in the queue(Wq)	9.9856	599.1388	35,948.33
population size	10.	Average time in the system(Ws)	17.9857	1,079.139	64,748.33
Server cost $/time	20.	Effective Arrival Rate	0.0848		
Waiting cost $/time	100.	Probability that customer waits	0.6201		
		Cost (Labor + # waiting*wait cost)	104.6344		
		Cost (Labor + # in system*wait cost)	172.4393		

Parameter	Value	Parameter	Value	Minutes	Seconds
M/M/s with a finite population		Average server utilization	0.3676		
Arvl rt PER CUSTOMER	0.01	Average number in the queue(Lq)	0.0761		
Service rate(mu)	0.125	Average number in the system(Ls)	0.8112		
Number of servers	2.	Average time in the queue(Wq)	0.8282	49.6893	2,981.36
population size	10.	Average time in the system(Ws)	8.8282	529.6893	31,781.36
Server cost $/time	20.	Effective Arrival Rate	0.0919		
Waiting cost $/time	100.	Probability that customer waits	0.1544		
		Cost (Labor + # waiting*wait cost)	47.6098		
		Cost (Labor + # in system*wait cost)	121.1201		

Parameter	Value	Parameter	Value	Minutes	Seconds
M/M/s with a finite population		Average server utilization	0.2467		
Arvl rt PER CUSTOMER	0.01	Average number in the queue(Lq)	0.0074		
Service rate(mu)	0.125	Average number in the system(Ls)	0.7476		
Number of servers	3.	Average time in the queue(Wq)	0.0799	4.7934	287.6013
population size	10.	Average time in the system(Ws)	8.0799	484.7934	29,087.6
Server cost $/time	20.	Effective Arrival Rate	0.0925		
Waiting cost $/time	100.	Probability that customer waits	0.0254		
		Cost (Labor + # waiting*wait cost)	60.7392		
		Cost (Labor + # in system*wait cost)	134.7585		


120
100
80
60
40
20
0
0
0.5
1
1.5
2
2.5
3
3.5
Costo MO
Costo de espera
Total