UNIVERSIDAD DE COSTA RICA

Modelos de las de espera y Teoría de colas

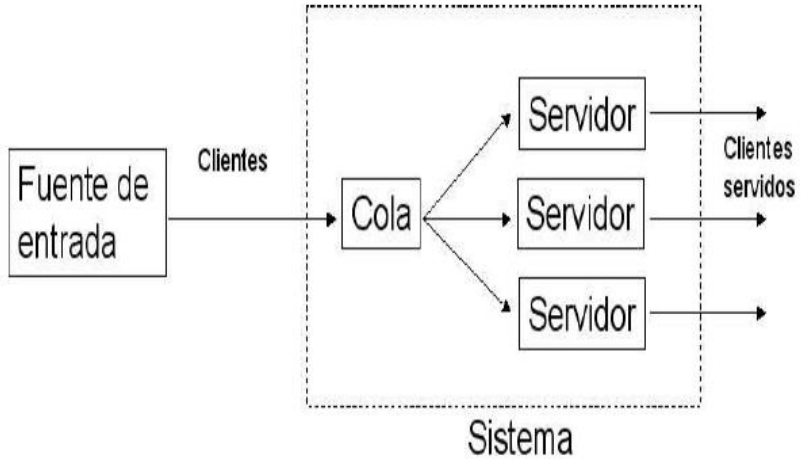
Ing. Luis Delgado Lobo MBA

Teoría de Colas y Líneas de espera

- El estudio de líneas de espera, llamado teoría de colas, es una de las técnicas de análisis cuantitativo más antiguas y que se utilizan con mayor frecuencia.
- Desarrollado por Anger Krarup Erlang entre 1909 to 1929 para la Copenhagen Telephone Company (CTC).
- Erlang desarrolló una solución del problema de determinar cuántos circuitos eran necesarios para proveer un servicio telefónico aceptable en centrales telefónicas automáticas.

Sistema de Colas

- Las líneas de espera generan malestar, ineficiencia, retraso y otros problemas, lo que origina un costo de tiempo y económico.
- Es muy importante evaluar el balance entre el aumento del nivel de servicio y el tamaño de las colas de espera.
- Por tanto, es necesario entender la relación entre el número de servidores en un sistema (o eficacia de los mismos) y la cantidad de tiempo gastado en la cola (o cantidad de clientes en la misma).
- En sistemas de colas sencillos dichas relaciones se pueden encontrar analíticamente. En sistemas más complejos se pueden analizar mediante simulación.



Sistema de Colas

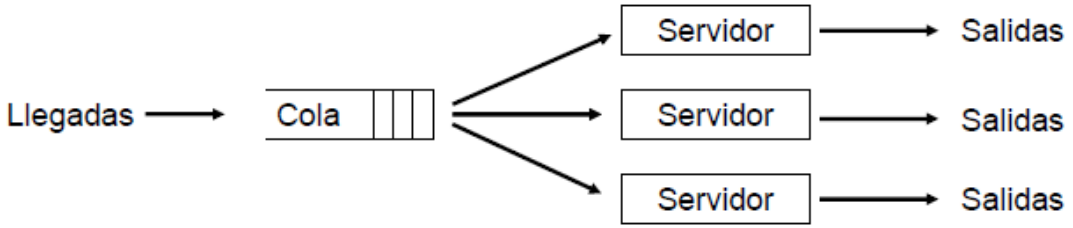
- Elementos más importantes en un sistema de colas: **clientes y servicio.**
- Los clientes se caracterizan por los intervalos de tiempo que separan sus llegadas.
- El servicio se caracteriza por el tipo y tiempo de servicio, además de por el número de servidores. El tipo de servicio o disciplina representa el orden en el que los clientes se seleccionan de la cola.
- Las llegadas de clientes pueden ser deterministas o aleatorios (en este caso se modelan mediante una distribución estadística).
- Los tiempos de servicio también pueden ser deterministas o aleatorios (distribución estadística).

Tipos de sistemas

Una cola, un servidor

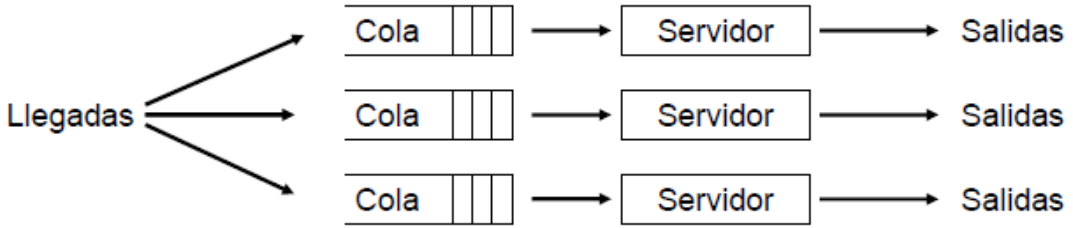
auto mc

Una cola, múltiples servidores



Banco

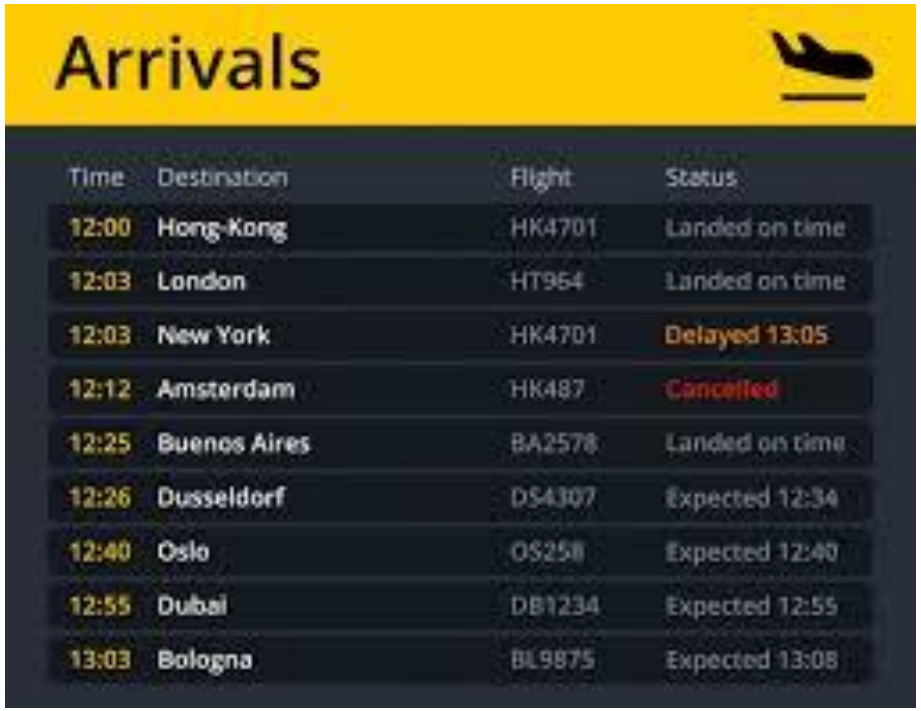
Múltiples colas, múltiples servidores



Super

Elementos de un sistema: Llegadas

- Pueden existir una o varias fuentes.
- Se suele asumir independencia entre llegadas.
- Intervalos entre llegadas: deterministas o aleatorios.
- Tasa de llegadas: λ = número medio de clientes que acceden al sistema por unidad de tiempo.
- Tiempo medio entre llegadas: $1/\lambda$



Elementos de un sistema: Fuente de entrada

Bancos Servidor

- Puede ser **infinita** o **finita** (sistemas abiertos o cerrados, respectivamente).
- **Ejemplo de sistema abierto**: un banco, ya que es prácticamente imposible que todos los posibles clientes coincidan en su llegada.
- **Ejemplo de sistema cerrado**: un servidor de internet con un número relativamente pequeño de usuarios autorizados (es posible que en un momento determinado se conecten todos los usuarios al servidor).
- Si la fuente es **finita**, entonces el número de clientes en la cola afecta al número de clientes fuera del sistema.
- La llegada puede ser en bloque o de forma unitaria. Frecuentemente el bloque se trata como un solo cliente.

Clientes

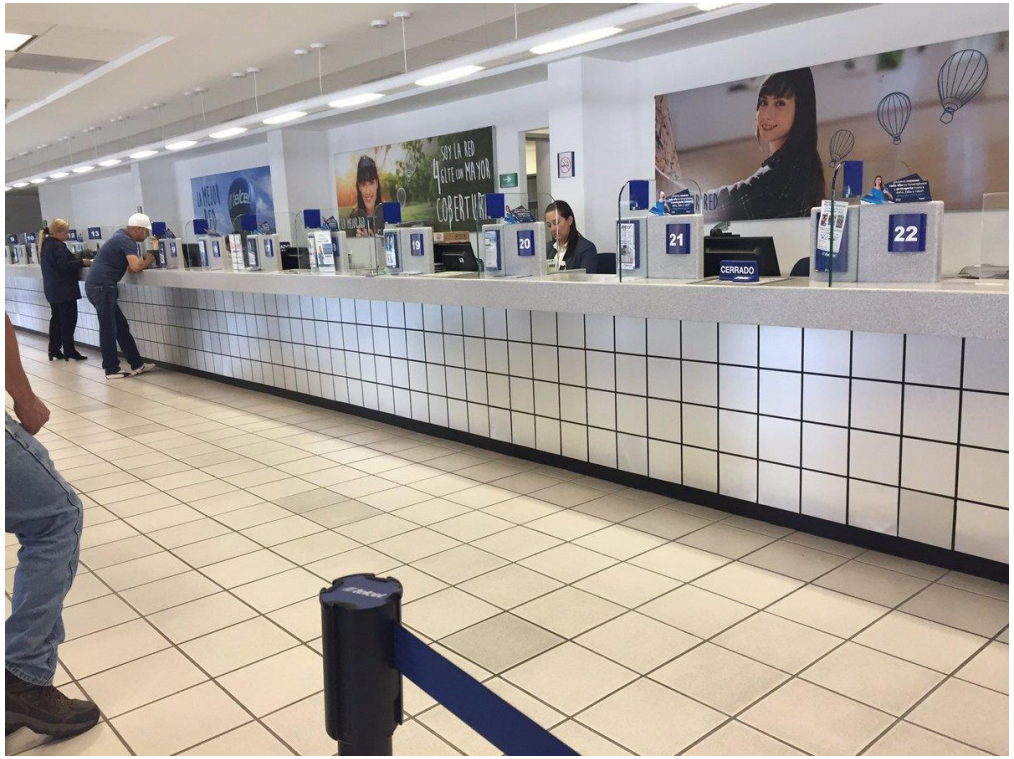
- Pueden ser pacientes o impacientes.
- Por tanto, los clientes se pueden perder, bien porque no entran en el sistema, o bien porque abandonan tras un tiempo en el sistema.
- También, los clientes pueden percibir un ritmo más acelerado en una cola distinta y por tanto decidir cambiarse

Cola o canal de espera

- Puede ser de uno o varios canales.
- Puede existir interferencia entre canales.
- Puede ser de capacidad limitada.
- Disciplina de la cola: orden de selección en el servicio (FIFO, LIFO, aleatorio, orden de prioridad, etc.)

Servicio

- Pueden existir uno o varios servidores.
- Se suele asumir independencia entre tiempos de servicio.
- Duración de los servicios: deterministas o aleatorios.
- Tasa de servicio: μ = número medio de clientes que son atendidos por unidad de tiempo.
- Tiempo medio de servicio: 1/μ.



Análisis de sistemas de colas

- Una vez caracterizado el sistema, se pueden contestar a las siguientes preguntas:
- ¿Qué proporción de tiempo están los servidores desocupados?
- ¿Cuál es el tiempo medio de espera para un cliente?, ¿es éste un tiempo razonable?, ¿se pierden clientes por tiempos de espera largos?
- ¿Es conveniente añadir más servidores para reducir el tiempo medio de espera?
- ¿Cuál es el número medio de clientes en cola?
- ¿Cuál es la probabilidad de que la espera sea mayor que una determinada longitud en un tiempo determinado?.

Notación de Kendall

- **D. G. Kendall** desarrolló en 1953 una notación ampliamente aceptada para especificar el patrón de las llegadas, la distribución del tiempo de servicio y el número de canales en un modelo de colas.
- Con frecuencia esta notación se encuentra en el software de modelos de colas.
- Las características del sistema se especifican por los símbolos:

A/B/s/k/t/d/

donde **A** y **B** denotan las distribuciones de los tiempos entre llegadas y de servicio, respectivamente **s** denota el número de servidores en paralelo o canales, **k** denota la capacidad del sistema, **t** denota el tamaño de la fuente de entrada, y **d** es la disciplina de la cola.

Análisis de sistemas de colas

La distribución puede ser

- M Exponencial
- D Constante o determinista
- E_k Erlang de parámetro k
- G Genérica e independiente

La disciplina puede ser

- FCFS First come, first served
- LCFS Last come, first served
- SIRO Service in random order
- GD General discipline

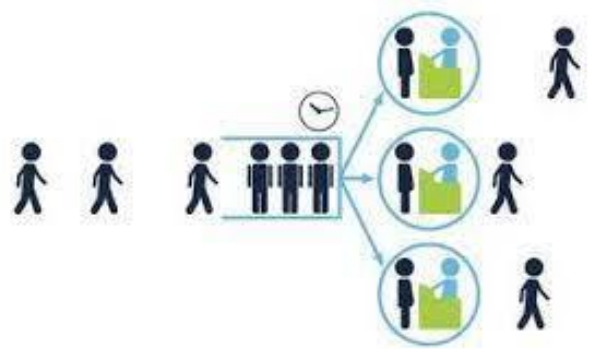
Análisis de sistemas de colas

- Por ejemplo, un sistema que se describe como

$$M/M/1/\infty/\infty/FCFS$$

- denota un sistema abierto que contiene un único servidor con tiempos de llegada y servicio exponenciales, capacidad infinita y disciplina primero que entra, primero que se sirve.
- Sólo un número pequeño de sistemas se puede resolver analíticamente.

- Modelos sencillos: M/M/1/, M/M/s/, M/M/1/k.



Notación universal

Objetivo: dados los siguientes parámetros (se suelen estimar estadísticamente)

λ = tasa de llegadas. μ =asa de servicio. s = número de servidores.

- $\rho = \lambda/s\mu$ = factor de utilización del sistema o intensidad de tráfico (proporción de tiempo esperado en el que los servidores están ocupados). Si $\rho < 1$ entonces el sistema se estabiliza. En otro caso el número de clientes en el sistema se incrementa sin límite.
- L = valor esperado del número de clientes en el sistema
- Lq = valor esperado del número de clientes en cola
- W = tiempo medio de espera en el sistema
- Wq = tiempo medio de espera en la cola
- pn = probabilidad de que n clientes estén en el sistema
- $\bar{c}$ = número medio de clientes en servicio.

Relaciones básicas: Modelo general

Fórmula de Little: $L = \lambda W$ y $L_q = \lambda W_q$.

Además, $W = W_q + 1/\mu$.

De estas tres fórmulas se deduce: $L = L_q + \lambda/\mu$

Modelo M/M/1

- En este caso, $\lambda_n = \lambda$, $\mu_n = \mu$, $\rho = \lambda/\mu < 1$ para todo n, Entonces,

$$P_n = \rho^n p_0, \quad p_0 = 1 - \rho,$$

- Por lo que $p_n = \rho^n(1-\rho)$.
- Por tanto.
- $L = E(N) = \sum_{n=1}^{\infty} np_n = \frac{\rho}{1-\rho}$
- $L_q = E(N_q) = \sum_{n=1}^{\infty}(n-1)pn = \frac{\rho^2}{1-\rho}$
- $W = E(T) = L/\lambda = \frac{1}{\mu(1-\rho)}$
- $W_q = E(T_q) = W - \frac{1}{\mu} = \frac{\rho}{\mu(1-\rho)}$
- Además $\bar{c} = L - Lq = \rho$

La probabilidad de que haya más de *k* clientes en el sistema es:

$$P(N \geq k) = \rho^k$$

y

$$P(N < k) = 1 - \rho^k$$

Modelo M/M/1: Ejemplo

- La tasa de llegadas de estudiantes al mostrador de una biblioteca es de 10 por hora. En el mostrador existe una sola persona y atiende con una tasa de 5 minutos por persona. ¿Cuáles son las medidas de comportamiento del sistema?

L	5	p_0	0.16
L_q	4.16	p_1	0.14
W	0.5	p_2	0.11
W_q	0.42	p_3	0.09
ρ	0.83	p_4	0.08

Modelo M/M/s

Dos servidores

- En sistemas con múltiples servidores (s > 1), la tasa de servicio depende del número de clientes en el sistema. En este caso, $\rho = \lambda/s\mu < 1$, y se puede probar que

$$p_0 = \frac{1}{\sum_{n=0}^{s-1} \frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda}{\mu}\right)^s}{s!\,(1-\rho)}}$$

y

$$p_n = \frac{\left(\frac{\lambda}{\mu}\right)^n p_0}{n!}, \text{ si } 0 \le n \le s$$

$$p_n = \frac{\left(\frac{\lambda}{\mu}\right)^n p_0}{s!\,s^{n-s}}, \text{ si } n > s$$

Modelo M/M/s

- Además

$$L_q = \frac{\left(\frac{\lambda}{\mu}\right)^s p_0 \rho}{s!\,(1-\rho)^2}$$

$$W_q = L_q / \lambda$$

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

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

La probabilidad de que un nuevo cliente tenga que esperar es

$$p_w = \left(\frac{\lambda}{\mu}\right)^s \frac{p_0}{s!(1-\rho)}$$

Modelo M/M/s: Ejemplo

- Un banco dispone de 3 ventanillas de atención. Los clientes llegan al banco con tasa de 1 por minuto. El tiempo de servicio es de 2 minutos por persona.

L	2.89	p_0	0.11
L_q	0.89	p_1	0.22
W	0.049	p_2	0.22
W_q	0.015	p_3	0.15
ρ	0.67	p_4	0.10

Modelo M/M/1/k

limita

- En este caso, si el sistema está lleno (la capacidad es k) no se permite la entrada de nuevos clientes al sistema. Por tanto, la tasa de llegada efectiva no es constante y varía con el tiempo (en función de si el sistema está lleno o no):

$$\lambda ef = \lambda(1 - p_k).$$

En este caso,

$$p_n = \rho^n p_0, \text{ para } n = 0, 1, \ldots, k$$

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

y no existe estado k + 1.

Por tanto,

$$p_0 + p_1 + p_2 + \cdots + p_k = 1.$$

$$p_0 = \frac{1-\rho}{1-\rho^{k+1}}, \text{ si } \lambda \neq \mu$$

$$p_0 = \frac{1}{1-k}, \text{ si } \lambda = \mu$$

Modelo M/M/1/k

- Además, se obtienen las siguientes relaciones:

$$L = \frac{\rho(1-(k+1)\rho^{k}+k\rho^{k+1})}{(1-\rho)(1-\rho^{k+1})} \qquad \text{si } \lambda \neq \mu$$

$$L = \frac{k}{2}, \text{ si } \lambda = \mu$$

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

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

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

Modelo M/M/1/k: Ejemplo

- Se está haciendo planes para abrir una pequeña estación para lavar automóviles y debe decidirse cuánto espacio dejar para automóviles que esperan. Se estima que los clientes llegarían aleatoriamente con una tasa media de uno cada 4 minutos, a menos que el área de espera esté llena, en cuyo caso el cliente se llevaría su automóvil a otra parte. El tiempo que puede atribuirse al lavado de un automóvil tiene una distribución exponencial con una media de 3 minutos. Compárese la fracción esperada de clientes potenciales que se perderán, debido a un espacio de espera inadecuado, si se tuviera a) Cero b) dos o c) Cuatro espacios (sin incluir el del automóvil que se está lavando)

1 - 2 1 - 4

Espacios Totales	K=1	K=3	K=5
λef	8.57	12.69	13.92

Aplicaciones de Teoría de Colas

- Se pueden usar los resultados de Teoría de Colas para la toma de decisiones:
- ¿Cuántos servidores emplear en el sistema?
- ¿Es mejor usar un único servidor rápido o muchos servidores más lentos?
- ¿Es mejor usar servidores idénticos o servidores específicos?
- Objetivo: minimizar el costo total = costo de servicio + costo de espera.

Aplicaciones de Teoría de Colas

- Costo de servicio: costo al aumentar la capacidad de servicio.
- La capacidad del servicio se puede aumentar añadiendo más servidores, $s \uparrow$, o haciendo servidores más eficientes, $\mu \uparrow$, etc.
- Habitualmente, la función de coste de servicio viene dada por $C_s S$, donde C_s representa el coste por unidad de tiempo y servidor.
- También se utiliza $C_\mu \mu$, donde C_μ representa el costo por unidad de tiempo y unidad de tasa de servicio.

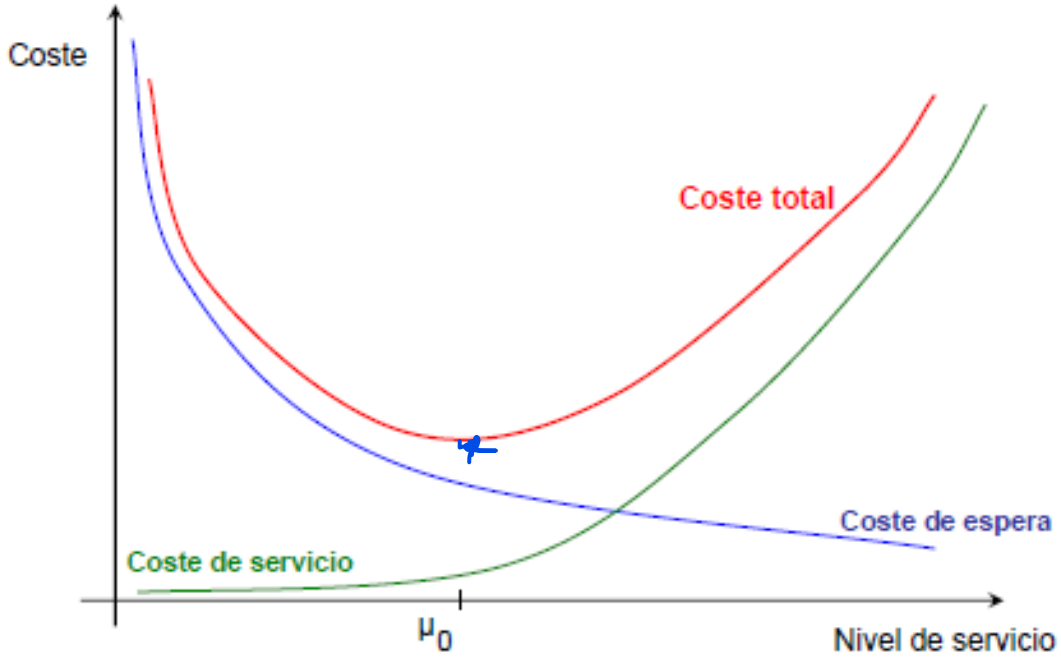
Aplicaciones de Teoría de Colas

- Costo de espera: coste asociado a la espera de los clientes.
- La espera de clientes genera tiempo perdido, pérdida de los mismos, etc.
- Habitualmente, la función de costo de espera viene dada por ClL(s), donde Cl denota el costo de espera por unidad de tiempo y cliente y L(s) es el valor esperado del número de clientes en el sistema para s servidores.
- También se utiliza CwW(μ), donde Cw denota el costo de espera por unidad de tiempo y cliente y W(μ) es el valor esperado del tiempo medio de espera en el sistema para una tasa de servicio de μ unidades.

Aplicaciones de Teoría de Colas

- La siguiente figura representa un modelo típico de costos
- El costo del servicio aumenta con el incremento en el nivel del servicio pero el coste por espera disminuye con el nivel.
- Hay que buscar el nivel de servicio que minimiza el costo total.

$$CT=Cl*L + Cs*S$$



Ejemplo: ¿cuántos servidores utilizar?

- Un banco dispone de 3 ventanillas de atención. Los clientes llegan al banco a una tasa de 40 por hora. El tiempo de servicio es de 3 minutos por persona.
- El banco se plantea si le conviene aumentar el número de ventanillas para satisfacer mejor a los clientes.
- El costo que le supone abrir una nueva ventanilla es de 6 dólares la hora. El costo horario de espera se ha estimado en 18 dólares por cliente.

	$s = 3$	$s = 4$	$s = 5$
L	2.88889	2.17391	2.03980
Coste de servicio	18.00	24.00	30.00
Coste de espera	52.00	39.13	36.72
Coste total	70.00	63.13	66.72

Ejemplo: ¿un servidor rápido o muchos lentos?

- En un servidor de Internet existen 3 nodos que atienden peticiones a razón de 50 por minuto. El tiempo medio de servicio de cada nodo es de 3 segundos por petición.
- En el servidor se plantean la posibilidad de instalar un único nodo con tiempo de servicio de 1 segundo por petición. ¿Es conveniente esta opción para reducir el tiempo medio de espera en el sistema?

	$s = 3$	$s = 1$
W	0.1202	0.1000