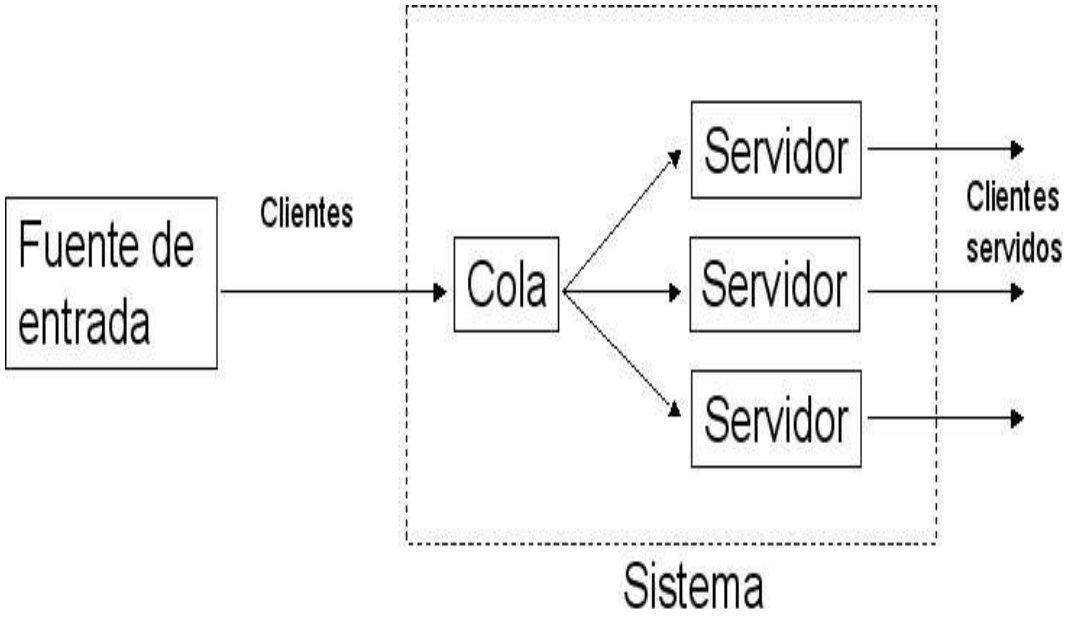
Teoría de Colas

TC: Parte de la Investigación Operativa que estudia el comportamiento de sistemas cuyos elementos incluyen líneas de espera (colas).



Teoría de Colas: ejemplos

- personas esperando por un servicio (bibliotecas, bancos, gasolineras, urgencias en hospital, . . .),
- máquinas esperando por una reparación, piezas de un producto esperando a ser ensambladas,
- programas de ordenador esperando a ser ejecutados por un procesador,
- información de internet esperando en un nodo para ser transferida a su destino,
- aviones esperando a despegar o aterrizar,

Teoría de Colas: historia

Se inició con A. K. Erlang, en la compañía telefónica estatal de Dinamarca (principios del siglo XX).

Se analizaron los tiempos de espera de llamadas a centralitas automáticas (congestión de tráfico).

- Objetivo: satisfacer la demanda incierta en el sistema telefónico con el menor coste para la compañía.

Teoría de Colas

- Introducción.
- Elementos y relaciones en un sistema.
- Modelo $M/M/1$.
- Modelo $M/M/s$.
- Modelo $M/M/1/k$.
- Aplicaciones.

Introducción

Las líneas de espera generan malestar, ineficiencia, retraso y otros problemas, lo que origina un coste 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.

Introducción

- Elementos más importantes en un sistema de colas: **clientes** y **servicio**.

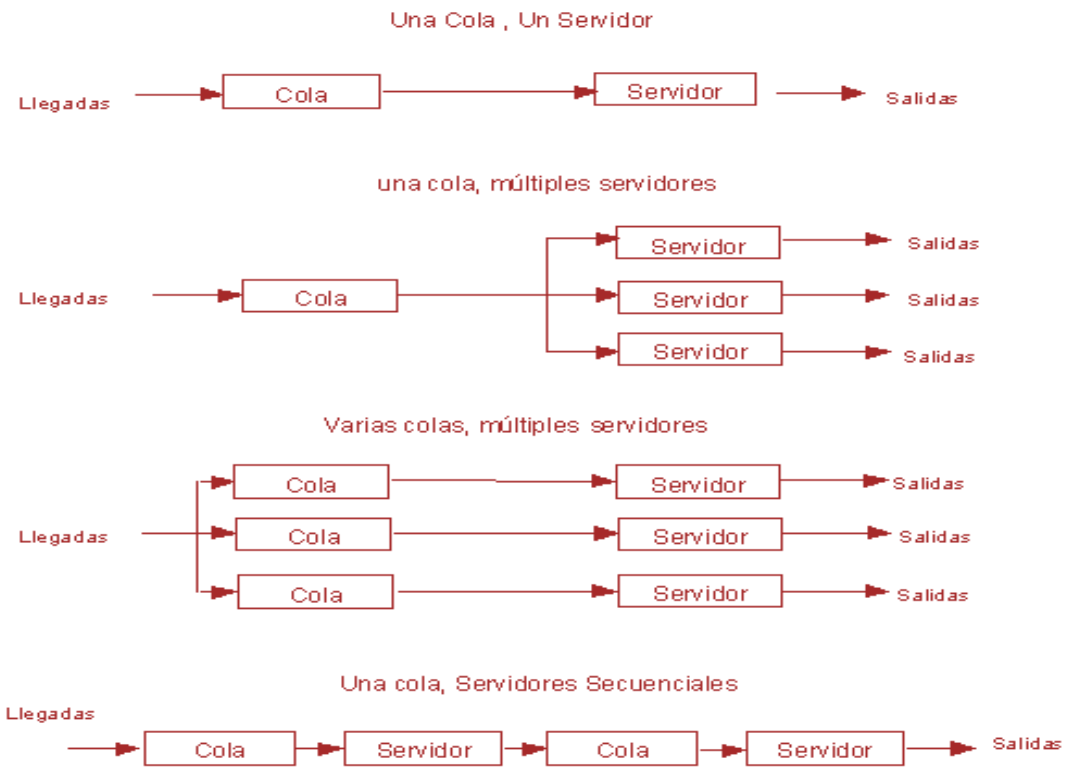
Los clientes se caracterizan por los intervalos de tiempo que separan sus llegadas.

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).

Introducción: tipos de sistemas



Las variaciones en un sistema de colas pueden ser múltiples. Sólo se pueden resolver de forma analítica un conjunto reducido de sistemas.

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: $\lambda \equiv$ número medio de clientes que acceden al sistema por unidad de tiempo.
- Tiempo medio entre llegadas: $\frac{1}{\lambda}$.

Elementos de un sistema: Fuente de entrada

- 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.

Introducción: Clientes

- Pueden ser impacientes.
- Por tanto, los clientes se pueden perder, bien porque no entran en el sistema, 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.

Elementos de un sistema: 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.).

Elementos de un sistema: 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: $\mu \equiv$ número medio de clientes que son atendidos por unidad de tiempo.
- Tiempo medio de servicio: $\frac{1}{\mu}$.

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?.
- ...

Análisis de sistemas de colas

• **Notación de Kendall:** 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

$$M/M/1/\infty/\infty/\text{FCFS}$$

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: $M/M/1/$, $M/M/s/$, $M/M/1/k$.

Distribuciones

En los sistemas de colas normalmente se asume que tanto las llegadas de clientes como los tiempos de servicio son aleatorios.

Es usual suponer que los tiempos entre llegadas y los de servicio se distribuyan de forma exponencial. En este caso, la probabilidad instantánea de ocurrencia de un suceso en las siguientes t unidades de tiempo es:

$$f(t) = \lambda e^{-\lambda t} \quad \text{para } t \geq 0,$$

donde λ denota la tasa de llegadas.

Esta distribución es útil ya que tiene la propiedad de falta de memoria y estacionariedad (el sistema se comporta, transcurrido un plazo, de forma estable e independientemente de las condiciones iniciales).

Distribuciones

Una distribución exponencial de los tiempos entre llegadas implica una distribución de Poisson para las llegadas, es decir, el número de llegadas en el intervalo $(0, t]$ es una Poisson. Una distribución de Poisson describe la probabilidad de que lleguen n clientes en las siguientes t unidades de tiempo:

$$P(X_t = n) = e^{-\lambda t}\frac{(\lambda t)^n}{n!} \quad \text{para } n = 0, 1, \ldots$$

En la práctica, se habla de llegadas Poisson y tiempos de servicio exponencial.

En general se supone que el sistema se encuentra en estado estacionario (estabilidad independiente del tiempo).

Notación universal

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

  - $\lambda \equiv$ tasa de llegadas.
  - $\mu \equiv$ tasa de servicio.
  - $s \equiv$ número de servidores.

Notación universal

se calcula

- $\rho = \frac{\lambda}{s\mu} \equiv$ 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.

- $L \equiv$ valor esperado del número de clientes en el sistema (la variable se denota por N).

- $L_q \equiv$ valor esperado del número de clientes en cola (la variable se denota por N_q).

Notación universal

y

- $W \equiv$ tiempo medio de espera en el sistema (la variable se denota por T).
- $W_q \equiv$ tiempo medio de espera en la cola (la variable se denota por T_q).
- $p_n \equiv$ probabilidad de que n clientes estén en el sistema (en estado estacionario).
- $\bar{c} \equiv$ número medio de clientes en servicio.

Relaciones básicas: Modelo general

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

Además, $W = W_q + \frac{1}{\mu}$.

De estas tres fórmulas se deduce: $L = L_q + \frac{\lambda}{\mu}$.

Relaciones básicas: Modelo general

Se dice que el sistema se encuentra en el estado n si existen exactamente n clientes en el mismo.

- Ecuaciones de balance de flujo (la tasa esperada de llegada al estado n es igual a la tasa esperada de salida del estado n en estado estacionario):

$$
\begin{aligned}
p_0\lambda_0 &= p_1\mu_1 \\
p_0\lambda_0 + p_2\mu_2 &= p_1\lambda_1 + p_1\mu_1 \\
p_1\lambda_1 + p_3\mu_3 &= p_2\lambda_2 + p_2\mu_2 \\
\cdots &= \cdots \\
p_{n-1}\lambda_{n-1} + p_{n+1}\mu_{n+1} &= p_n\lambda_n + p_n\mu_n \\
\cdots &= \cdots
\end{aligned}
$$

Relaciones básicas: Modelo general

Si resolvemos las ecuaciones anteriores para p_i se obtiene

$$
\begin{aligned}
p_1 &= \frac{\lambda_0}{\mu_1} p_0 \\
p_2 &= \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} p_0 \\
p_3 &= \frac{\lambda_2 \lambda_1 \lambda_0}{\mu_3 \mu_2 \mu_1} p_0 \\
\cdots &= \cdots \\
p_n &= \frac{\lambda_{n-1} \cdots \lambda_1 \lambda_0}{\mu_n \cdots \mu_2 \mu_1} p_0.
\end{aligned}
$$

Para calcular p_0 (prob. de que el sistema esté vacío), se utiliza:

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

Modelo $M/M/1$

En este caso, $\lambda_n = \lambda$, $\mu_n = \mu$, $\rho = \frac{\lambda}{\mu} < 1$ 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,

$$L = E(N) = \sum_{n=0}^{\infty} n p_n = \frac{\rho}{1 - \rho} \quad \text{(ejercicio)}.$$

y de la misma forma,

$$L_q = E(N_q) = \sum_{n=1}^{\infty} (n - 1) p_n = \frac{\rho^2}{1 - \rho} \quad \text{(ejercicio)}.$$

Modelo $M/M/1$

Por la fórmula de Little:

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

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

Además, $\bar{c} = L - L_q = \rho$.

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

$$P(N \geq k) = 1 - \sum_{n=0}^{k-1} p_k = 1 - \sum_{n=0}^{k-1} \rho^n(1-\rho) = 1 - (1-\rho)(1-\rho^k)/(1-\rho) = \rho^k.$$

Por tanto,

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

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?

- Datos: $\lambda = 10$ (tasa de llegadas), $\mu = 60/5 = 12$ (tasa de servicio), $s = 1$ (número de servidores). Se suponen distribuciones exponenciales.

- Resultados:

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 $M/M/s$

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

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

y

$$p_n = \frac{(\lambda/\mu)^n p_0}{n!}, \quad \text{si } 0 \leq n \leq s$$

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

Modelo $M/M/s$

Además,

$$L_q = \frac{(\lambda/\mu)^s \; p_0 \; \rho}{s! \; (1-\rho)^2}$$

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

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

$$L = \lambda W = L_q + \frac{\lambda}{\mu}.$$

Prob. de que un nuevo cliente tenga que esperar: $p_w = (\frac{\lambda}{\mu})^s \; \frac{p_0}{s! \; (1-\rho)}$.

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.

- Datos: $\lambda = 60$ (tasa de llegadas), $\mu = 60/2 = 30$ (tasa de servicio), $s = 3$ (número de servidores).

- Resultados:

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 $M/M/1/k$

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,

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

y no existe estado $k + 1$.

Por tanto,

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

Modelo $M/M/1/k$

De la anterior expresión se deduce que

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

$$p_0 = \frac{1}{1+k}, \quad \text{si } \lambda = \mu.$$

y siempre existe una distribución estacionaria (aunque $\lambda > \mu$).

Modelo $M/M/1/k$

Además, se obtienen las siguientes relaciones:

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

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

y

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

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

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

Limitaciones de los sistemas de colas

- La resolución analítica de los sistemas se complica a medida que los sistemas se hacen más complejos. De hecho, para muchos sistemas no existe resolución analítica.

- Ejemplo: un sistema de servidores en paralelo y en serie con múltiples canales y distribuciones generales.

- En sistemas de colas complejos conviene utilizar simulaciones para estudiar su comportamiento.

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 coste total = coste de servicio + coste de espera.

Aplicaciones de Teoría de Colas

- Coste de servicio: coste al aumentar la capacidad de servicio.

La capacidad del servicio se puede aumentar añadiendo más servidores, $s \nearrow$, o haciendo servidores más eficientes, $\mu \nearrow$, 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 coste por unidad de tiempo y unidad de tasa de servicio.

Aplicaciones de Teoría de Colas

- Coste de espera: coste asociado a la espera de los clientes.

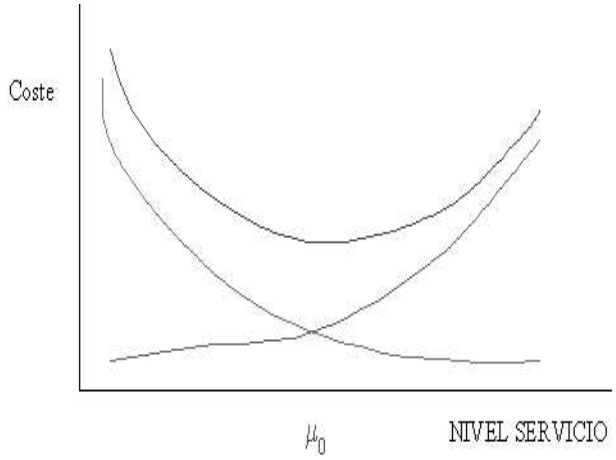
La espera de clientes genera tiempo perdido, pérdida de los mismos, etc.

Habitualmente, la función de coste de espera viene dada por $C_l L(s)$, donde C_l denota el coste 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 $C_w W(\mu)$, donde C_w denota el coste de espera por unidad de tiempo y cliente y $W(\mu)$ 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 costes (en euros por unidad de tiempo):



El coste 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 coste total.

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 coste que le supone abrir una nueva ventanilla es de 6 euros la hora. El coste horario de espera se ha estimado en 18 euros por cliente.

- Datos: $\lambda = 40$ (tasa de llegadas), $\mu = 60/3 = 20$ (tasa de servicio), $s = 3$ (número de servidores), $C_s = 6$, $C_l = 18$.

Ejemplo: ¿cuántos servidores utilizar?

- Resultados:

	$s = 3$	$s = 4$	$s = 5$
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

Por tanto, al banco le interesa abrir sólo una ventanilla más.

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?

- Datos: $\lambda = 50$ (tasa de llegadas), $\mu = 20$ (tasa de servicio) con $s = 3$ (número de servidores), y $\mu = 60$ con $s = 1$.

Ejemplo: ¿un servidor rápido o muchos lentos?

- Resultados:

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

Por tanto, es más conveniente utilizar un ordenador más rápido.