

# **ISE-TEMA-5.pdf**



**Phantone**



**Ingeniería de Servidores**



**3º Grado en Ingeniería Informática**



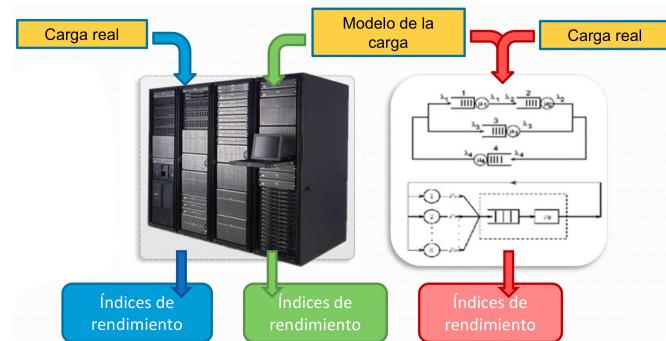
**Escuela Técnica Superior de Ingenierías Informática y de  
Telecomunicación  
Universidad de Granada**

# Tema 5: Optimización del rendimiento de un servidor mediante análisis operacional

## 1. Introducción: Redes de colas de espera

¿Cómo podemos mejorar el rendimiento de un servidor?

- Monitorización
- Comparación de rendimiento
- Optimización del rendimiento



### 1.1 El modelo de un sistema informático

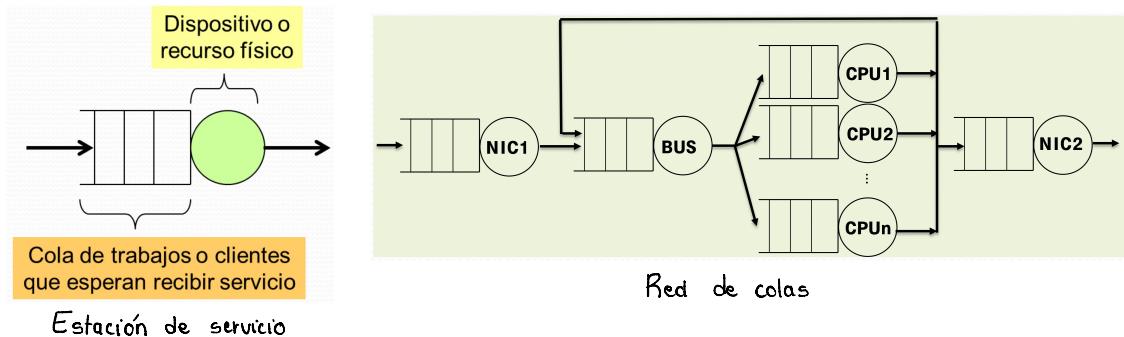
**Abstracción del sistema informático real:** conjunto de dispositivos interrelacionados y trabajos que los usan (carga).

- Dispositivos (resources): núcleos lógicos, unidades de almacenamiento permanentes, tarjetas de red, etc.
- Trabajos (jobs): procesos, accesos, peticiones, etc.

Normalmente un dispositivo o recurso solo puede ser usado por un trabajo a la vez. El resto de trabajos tendrá que esperar.

**Modelos basados en redes de colas** (queueing networks):

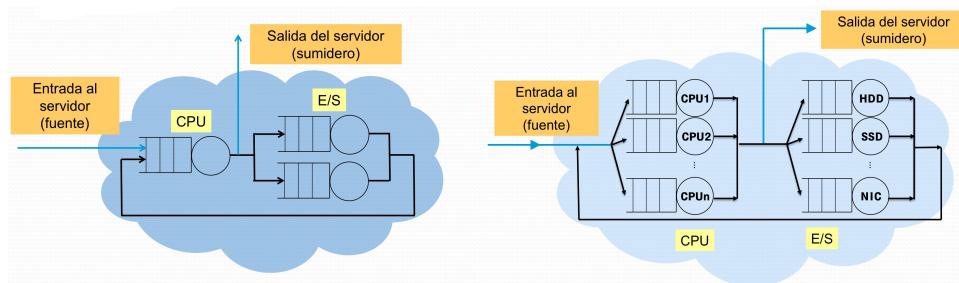
- Una red de colas está formada por un conjunto de estaciones de servicio conectadas entre sí.
- Estación de servicio (service station): Objeto compuesto por un dispositivo (recurso físico) que presta un servicio y una cola de espera para los trabajos (clientes) que demandan un servicio de él.



### 1.2 El modelo del servidor central

Es la red de colas que más se ha utilizado para representar el comportamiento básico de los programas en un servidor de cara a extraer información sobre su rendimiento.

- Un trabajo que "llega" al servidor comienza utilizando el procesador.
- Despues de "abandonar" el procesador, el trabajo puede:
  - terminar (sale del servidor), o bien
  - realizar un acceso a una unidad de entrada/salida (discos, red,...).
- Despues de una operación con una unidad de entrada/salida, el trabajo vuelve a "visitar" al procesador.



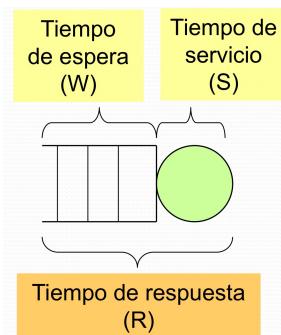
## Algunas variables que caracterizan a un trabajo en una estación de servicio en un instante concreto

• **Tiempo de espera en cola (W, waiting time):** tiempo transcurrido desde que el trabajo solicita hacer uso del recurso físico (=se pone en la cola) hasta que realmente empieza a utilizarlo.

• **Tiempo de servicio (S, service time):** desde que el trabajo accede al recurso físico hasta que lo libera (=tiempo que tarda el recurso físico en procesar el trabajo).

• **Tiempo de respuesta (R, response time):** suma de los dos tiempos anteriores.

Recopilando estas medidas para múltiples trabajos, obtendremos distribuciones de probabilidad que caracterizan a esa estación de servicio.

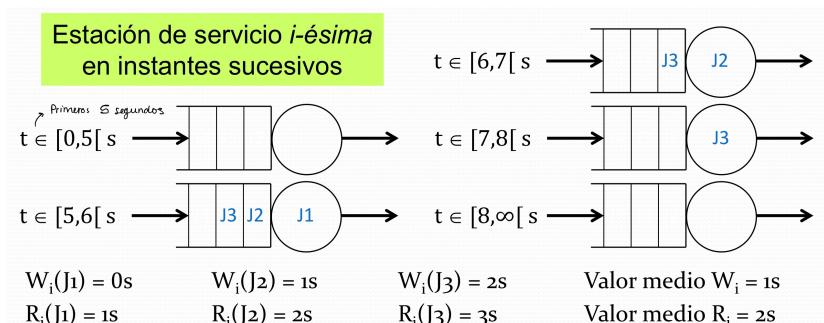


**EJERCICIO:** Suponga que la estación de servicio  $i$ -ésima de una red de colas tiene un tiempo de servicio constante  $S_i=1s$ .

Suponga que los trabajos (jobs) llegan con la siguiente distribución temporal:

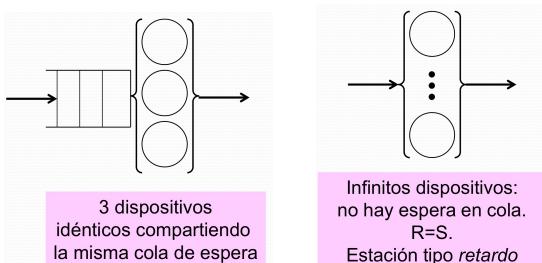
- Durante los primeros 5 segundos no llega ningún trabajo.
- En  $t=5s$  llegan 3 trabajos:  $J_1$ ,  $J_2$  y  $J_3$  (por ese orden).

Calcule los tiempos de espera en la cola y los tiempos de respuesta que experimentan cada uno de los tres trabajos. Calcule finalmente los valores medios de  $W$  y  $R$ .



## 1.3 Estaciones con más de un servidor

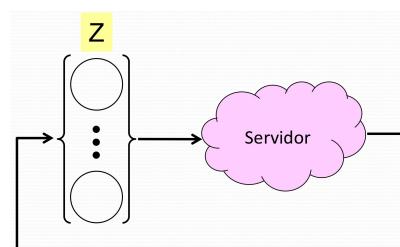
Son capaces de atender a más de un trabajo en paralelo:



## 1.4 El tiempo de reflexión (Z, think time)

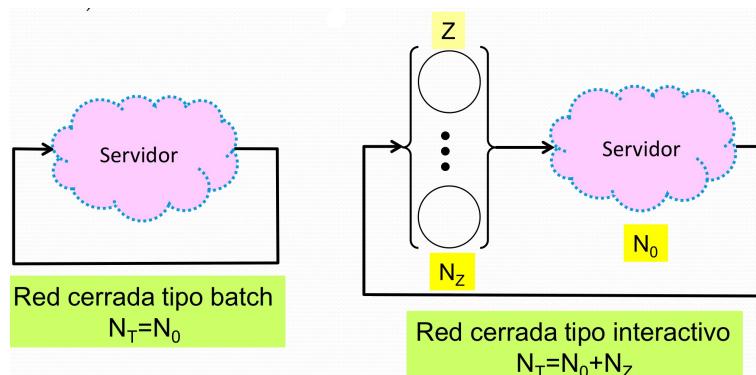
Es un parámetro ( $Z$ ) que representa el tiempo que requiere el usuario antes de volver a lanzar una petición al servidor tras la respuesta de éste.

Se suele modelar mediante una estación de servicio tipo retardo con un tiempo de servicio =  $Z$ . Para ello, realizamos una hipótesis adicional: **cada usuario envía un único trabajo al servidor**.



## 1.5 Redes de colas cerradas

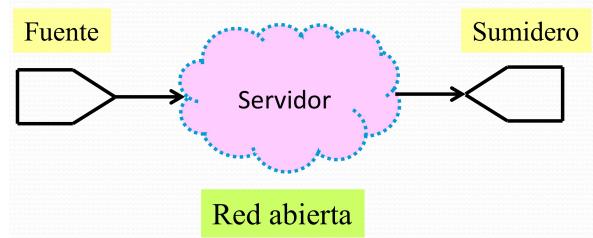
Presentan un número constante de trabajos que van recirculando por la red (NT). Dependiendo de si hay o no interacción con usuarios se distingue entre redes de tipo batch (por lotes) o redes interactivas.



$N_0$ : nº trabajos en el servidor  
 $N_T$ : nº trabajos en el sistema (usuarios)  
 $N_Z$ : nº trabajos en reflexión (esperando a que los usuarios vuelvan a introducirlos en el servidor) Siempre supondremos  $1 \text{ usr} = 1 \text{ trabajo}$ .

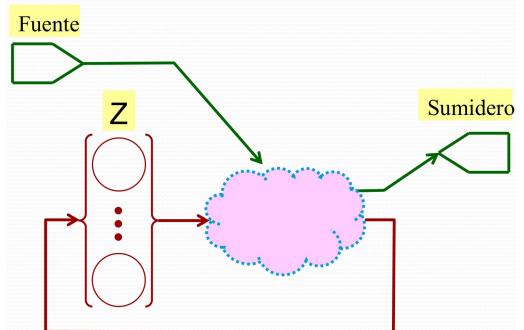
## 1.6. Redes de colas abiertas

Los trabajos llegan a la red a través de una fuente externa que no controlamos. Tras ser procesados, salen de ella a través de uno o más sumideros. No existe realimentación entre sumidero y fuente. El número de trabajos en el servidor ( $N_0$ ) puede variar con el tiempo.



## 1.7 Redes mixtas

Cuando el modelo no corresponde a ninguno de los dos anteriores.



## 2. Variables y leyes operacionales

### 2.1 Análisis operacional

Técnica análisis de redes de colas basada en valores medios de diferentes variables medibles (variables operacionales) del servidor.

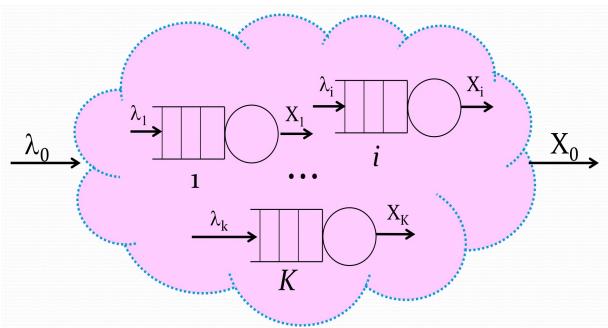


- Nos proporcionará relaciones generales entre las variables operacionales (leyes operacionales).
- Nos permitirá calcular las prestaciones del servidor para los casos de baja y alta carga por medio de cálculos muy sencillos.
- Nos permitirá evaluar los efectos en el rendimiento de diferentes modificaciones en los recursos del servidor.

### 2.2 Variables

#### Variables del servidor y de cada estación de servicio

- El servidor contiene  $K$  estaciones de servicio (recursos o dispositivos).
- A todo el servidor en su globalidad lo denotamos como dispositivo "cero".



#### Variables operacionales básicas de una estación de servicio

##### Variable global temporal:

- $T$  Duración del periodo de medida para el que se extrae el modelo.

##### Variables operacionales básicas de la estación de servicio $i$ -ésima medidas durante el tiempo $T$ :

- $A_i$ : Número de trabajos solicitados a la estación (llegadas, arrivals).
- $B_i$ : Tiempo que el dispositivo ha estado en uso (=ocupado) (busy time).
- $C_i$ : Número de trabajos completados por la estación (salidas, completions).



Estación de servicio  $i$ -ésima

#### Variables operacionales deducidas

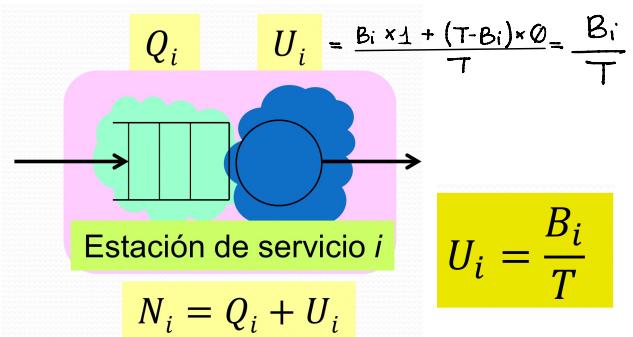
Se deben poder estimar a partir de las variables básicas:

- |   |                |
|---|----------------|
| • $\lambda_i$ Tasa media de llegada (arrival rate)    | - tr/s         |
| • $X_i$ Productividad media (throughput)              | - tr/s         |
| • $S_i$ Tiempo medio de servicio (service time).      | - s [ /tr ]    |
| • $W_i$ Tiempo medio de espera en cola (waiting time) | - s [ /tr ]    |
| • $R_i$ Tiempo medio de respuesta (response time).    | - s [ /tr ]    |
| • $U_i$ Utilización media (utilization)               | - sin unidades |

$$\lambda_i = \frac{A_i}{T} \quad \rightarrow \quad \text{Estación } i\text{-ésima} \quad X_i = \frac{C_i}{T} \quad U_i = \frac{B_i}{T}$$

$$R_i = W_i + S_i \quad S_i = \frac{B_i}{C_i}$$

- **N<sub>i</sub>**: Número medio de trabajos en la estación de servicio (cola más recurso).
- **Q<sub>i</sub>**: Número medio de trabajos en cola de espera (jobs in queue).
- **U<sub>i</sub>**: Número medio de trabajos siendo servidos por el dispositivo,  $U_i = N_i - Q_i$ . Coincide numéricamente con la utilización media = proporción de tiempo que el dispositivo ha estado en uso (busy) con respecto al intervalo total de medida (T) (como máximo 1 si B<sub>i</sub>=T).

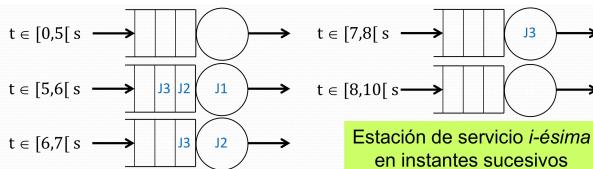


## EJERCICIO

Suponga que la estación de servicio  $i$ -ésima de una red de colas tiene un tiempo de servicio constante  $S_i=1s$ . Suponga que los trabajos llegan con la siguiente distribución temporal:

- Durante los primeros 5 segundos no llega ningún trabajo.
- En  $t=5s$  llegan 3 trabajos: J1, J2 y J3 (por ese orden).

Para el intervalo de medida  $[0, 10[s]$ , calcule A<sub>i</sub>, B<sub>i</sub>, C<sub>i</sub>, λ<sub>i</sub>, X<sub>i</sub>, U<sub>i</sub>, Q<sub>i</sub>, N<sub>i</sub>.



$$A_i = 3 \text{ trabajos}, \quad B_i = 3s, \quad C_i = 3 \text{ trabajos}, \quad \lambda_i = X_i = 3/10 = 0,3 \text{ trabajos/s}$$

Otras alternativas para calcular N<sub>i</sub> y U<sub>i</sub>:

$$N_i = \frac{0 \times 5s + 3 \times 1s + 2 \times 1s + 1 \times 1s + 0 \times 2s}{10s} = 0,6 \text{ tr.}$$

$$U_i = \frac{0 \times 5s + 1 \times 3s + 0 \times 2s}{10s} = 0,3 \text{ tr.}$$

Cálculo de la utilización media (U) y del número medio de trabajos en la cola (Q) y en la estación (N).

$$U_i = \frac{B_i}{T} = \frac{3}{10} = 0,3$$

$$Q_i = \frac{0 \times 5s + 2 \times 1s + 1 \times 1s + 0 \times 3s}{10s} = 0,3 \text{ tr.}$$

$$N_i = Q_i + U_i = 0,6 \text{ tr.}$$

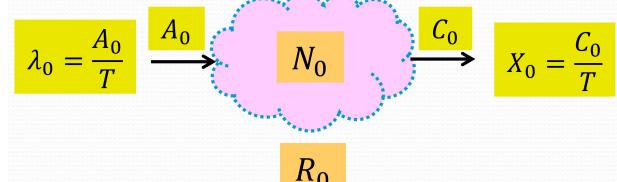
## Variables operacionales de un servidor

### Variables operacionales básicas de un servidor:

- **A<sub>0</sub>** Número de trabajos solicitados al servidor (arrivals).
- **C<sub>0</sub>** Número de trabajos completados por el servidor (completions).

### Variables operacionales deducidas de un servidor:

- **X<sub>0</sub>** Tasa media de llegada al servidor (arrival rate).
- **X<sub>0</sub>** Productividad media del servidor (throughput).
- **N<sub>0</sub>** Número medio de trabajos en el servidor (#jobs) = N<sub>1</sub>+N<sub>2</sub>+...+N<sub>K</sub>.
- **R<sub>0</sub>** Tiempo medio de respuesta del servidor (response time) = tiempo que tarda, de media, el servidor en procesar una petición.



## Razón de visita y demanda de servicio

• **Razón media de visita V<sub>i</sub>** (visit ratio). Representa la proporción entre el número de trabajos completados por el servidor y el número de trabajos completados por la estación de servicio  $i$ -ésima. Nos indica el número de veces que, de media, un trabajo “visita” la estación de servicio  $i$ -ésima antes de abandonar el servidor.

$$V_i = \frac{C_i}{C_0}$$

• **Demanda media de servicio D<sub>i</sub>** (service demand). Cantidad de tiempo que, por término medio, el dispositivo de la estación de servicio  $i$ -ésima le ha dedicado a cada trabajo que abandona el servidor (= que ha sido procesado por completo por el servidor).

$$D_i = \frac{B_i}{C_0} = V_i \times S_i$$

Nótese que la demanda de servicio de una estación no tiene en cuenta la posible espera de un trabajo en su cola.

## EJERCICIO

Después de monitorizar el disco duro de un servidor web durante un periodo de 24 horas, se sabe que ha estado en uso (=ocupado) un total de 6 horas. Asimismo, se han contabilizado durante ese periodo un total de 98590 peticiones de lectura/escritura al disco duro y un total de 98591 peticiones completadas. Se ha estimado que cada petición atendida por el servidor web ha requerido una media de 9,5 peticiones de lectura/escritura al disco duro. Calcule, para ese periodo de monitorización:

- La tasa media de llegada y la productividad media del disco duro.
- La utilización media del disco duro.
- El tiempo medio de servicio y la demanda media de servicio del disco duro.
- ¿Cuál es la productividad media del servidor web?

Nota: Todas las variables que se usan en este tema son valores medios por lo que, de aquí en adelante, normalmente no se indicará de forma explícita la palabra "medio" al referirnos a ellas.

### DATOS

$$T = 24 \text{ h}$$

$$B_{dd} = 6 \text{ h}$$

$$A_{dd} = 98590 \text{ pet. E/S (trabajos)}$$

$$C_{dd} = 98591 \text{ pet E/S (h)}$$

$$V_{dd} = \frac{C_{dd}}{C_0} = 9,5$$

a) ¿ $\lambda_{dd}$  y  $X_{dd}$ ?

$$\lambda_{dd} = \frac{A_{dd}}{T} = \frac{98590}{24} = 4107,92 \text{ tr/h} = 1'14 \text{ tr/s}$$

$$X_{dd} = \frac{C_{dd}}{T} = \frac{98591}{24} = 4107,96 \text{ tr/h} = 1'14 \text{ tr/s}$$

b) ¿ $U_{dd}$ ?  $U_{dd} = \frac{B_{dd}}{T} = \frac{6}{24} = 0,25 \rightarrow 25\%$

c) ¿ $S_{dd}$  y  $D_{dd}$ ?

$$S_{dd} = \frac{B_{dd}}{C_{dd}} = \frac{6h \cdot 3600 \text{ s/h}}{98591} = 0,22 \text{ s/tr}$$

$$D_{dd} = \frac{B_{dd}}{C_0} = S_{dd} \cdot V_{dd} = 0,22 \cdot 9,5 = 2,1 \text{ s/tr}$$

d) ¿ $X_0$ ?

$$X_0 = \frac{C_0}{T} = \frac{10378}{24 \cdot 3600} = 0,12 \text{ s/tr}$$

$$C_0 = \frac{C_{dd}}{V_{dd}} = \frac{98591}{9,5} = 10378 \text{ tr}$$

## 2.3 Leyes operacionales

El valor de todas las variables utilizadas en el análisis operacional dependerá del intervalo de observación T.

Existen, sin embargo, una serie de relaciones entre algunas variables operacionales que se mantienen válidas para cualquier intervalo de observación y que no dependen de suposiciones sobre la distribución de los tiempos de servicio o de la forma en la que llegan los trabajos. Estas relaciones se denominan **leyes operacionales**

Estas leyes son tanto más útiles cuando se cumple la denominada **hipótesis del equilibrio de flujo** que establece que si se escoge un intervalo de observación T suficientemente largo, se cumple que:

- Es capaz de hacer lo que se le pide*
- El número de trabajos que completa el servidor coincide aproximadamente con los solicitados ( $C_0 \approx A_0$ ). Dicho de otra manera, la productividad media coincide aproximadamente con la tasa media de llegada ( $X_0 \approx \lambda_0$ ).
  - El numero de trabajos que completa cada estación de servicio coincide aproximadamente con los que se solicitan: ( $C_i \approx A_i \rightarrow X_i \approx \lambda_i, \forall i=1...K$ ).

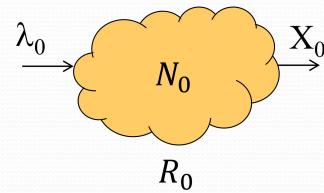
### Ley de Little

"The long-term average number of customers in a stable system is equal to the long-term average arrival rate multiplied by the average time a customer spends in the system"



Aplicada a un servidor, esta ley relaciona las dos variables más importantes que reflejan el rendimiento de un servidor: su productividad ( $X_0$ ) y su tiempo de respuesta ( $R_0$ ).

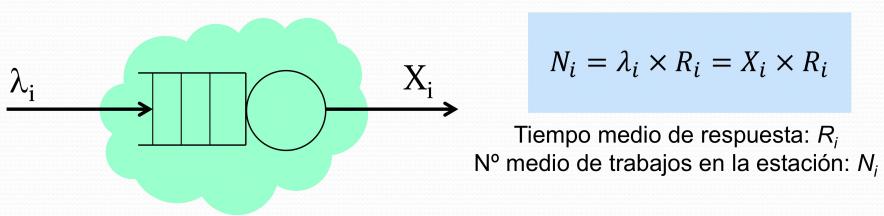
Esta ley solo es válida cuando el servidor está en equilibrio de flujo



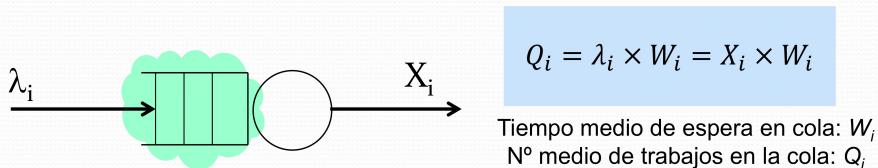
$$N_0 = \lambda_0 \times R_0 = X_0 \times R_0$$

La ley de Little puede ser aplicada no solo al servidor en su totalidad, sino a cada estación de servicio y a cada uno de los diferentes sub-niveles de una estación de servicio.

- Aplicación a toda una estación de servicio:



- Aplicación a la cola de una estación de servicio:



### Ley de la Utilización

Relaciona la utilización de un dispositivo con el número de trabajos que es capaz de realizar por unidad de tiempo (=su productividad) y el tiempo que le dedica a cada uno de ellos (=su tiempo de servicio).

$$U_i = X_i \times S_i = \lambda_i \times S_i \quad \text{Si equilibrio de flujo} \quad \rightarrow \text{Demostración: } S_i = \frac{B_i}{C_i} = \frac{B_i/T}{C_i/T} = \frac{U_i}{X_i}$$

- Una consecuencia inmediata de esta ley es que la productividad media de un dispositivo viene limitada por la inversa de su tiempo de servicio.

$$U_i \leq 1 \quad \Rightarrow \quad X_i \leq \frac{1}{S_i} \quad \forall i = 1, \dots, K$$

### Ley del flujo forzado y relación Utilización-Demanda

Las productividades (=flujos de salida) de cada estación de servicio tienen que ser proporcionales a la productividad global del servidor. La **ley del flujo forzado** relaciona la productividad del servidor con la de cada uno de los dispositivos que integran el mismo:

$$X_i = X_0 \times V_i = \lambda_0 \times V_i = \lambda_i \quad \text{Si equilibrio de flujo} \quad \text{Demostración: } V_i = \frac{C_i}{C_0} = \frac{C_i/T}{C_0/T} = \frac{X_i}{X_0}$$

Como consecuencia de la ley del flujo forzado, las utilizaciones de cada dispositivo son proporcionales a las demandas de servicio del mismo, siendo la constante de proporcionalidad precisamente la productividad global del servidor (**relación Utilización-Demanda de servicio**):

$$U_i = X_0 \times D_i = \lambda_0 \times D_i \quad \text{Si equilibrio de flujo} \quad \text{Demostración: } D_i = \frac{B_i}{C_0} = \frac{B_i/T}{C_0/T} = \frac{U_i}{X_0} \quad \left| \begin{array}{l} U_i \leq 1 \rightarrow X_0 \cdot D_i \leq 1 \rightarrow \\ \rightarrow X_0 \leq \frac{1}{D_i}, \quad i = 1, 2, \dots, K \end{array} \right.$$

### EJEMPLO

Un servidor de base de datos en equilibrio de flujo recibe una media de 120 consultas por minuto. Sabemos que su disco duro tarda, de media, 30ms en atender cada petición de E/S que le llega (48ms si incluimos la espera en la cola) y que su productividad es 25 peticiones de E/S completadas por segundo. Calcule:

- El número medio de peticiones de E/S en la cola de espera del disco duro.
- ¿Cuánto tiempo, de media, consumen los accesos al disco duro por cada consulta que se realiza al servidor?

a) Solución:

$$Q_{dd} = \lambda_{dd} \times W_{dd} = X_{dd} \times (R_{dd} - S_{dd}) = 25 \text{ tr/s} \times 0,018\text{s} = 0,45 \text{ tr.} = 0,45 \text{ pet. de E/S}$$

Solución alternativa:

$$N_{dd} = \lambda_{dd} \times R_{dd} = X_{dd} \times R_{dd} = 25 \text{ tr/s} \times 0,048 \text{ s} = 1,2 \text{ tr.}$$

$$U_{dd} = X_{dd} \times S_{dd} = 25 \text{ tr/s} \times 0,03 \text{ s/tr} = 0,75 \text{ (75%)}$$

$$Q_{dd} = N_{dd} - U_{dd} = 1,2 - 0,75 = 0,45 \text{ tr.} = 0,45 \text{ pet. de E/S}$$

$$b) D_{dd} = \frac{B_{dd}}{C_0} = \frac{B_{dd}/T}{C_0/T} = \frac{U_{dd}}{X_0} = \frac{U_{dd}}{\lambda_0} = \frac{0,75}{120 \text{ tr/min}} = 0,00625 \text{ min} = \boxed{0,375 \text{ s.}}$$

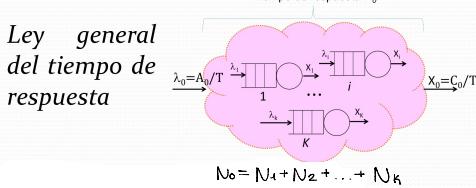
Extra: Por cada consulta al servidor ¿Cuántos accesos se hacen al disco duro?

$$\forall dd = \frac{C_{dd}}{C_0} = \frac{C_{dd}/T}{C_0/T} = \frac{X_{dd}}{X_0} = \frac{25}{2} = \boxed{12,5 \text{ tr/s}}$$

### Ley general del tiempo de respuesta

El tiempo medio de respuesta que experimenta, de media, una petición a un servidor en equilibrio de flujo se puede calcular teniendo en cuenta que cada una de ellas ha tenido que “visitar”  $N_i$  veces al dispositivo  $i$ -ésimo, requiriendo cada visita una media de  $R_i$  segundos:

$$R_0 = V_1 \times R_1 + V_2 \times R_2 + \dots + V_K \times R_K = \sum_{i=1}^K V_i \times R_i$$



Demostración:

$$\begin{aligned} N_0 &= N_1 + N_2 + \dots + N_K \xrightarrow{\text{Ley de Little}} X_0 \times R_0 = X_1 \times R_1 + X_2 \times R_2 + \dots + X_K \times R_K \\ \text{Ley del Flujo Forzado} \quad \Rightarrow X_0 \times R_0 &= X_0 \times V_1 \times R_1 + X_0 \times V_2 \times R_2 + \dots + X_0 \times V_K \times R_K \end{aligned}$$

Nótese que en general:

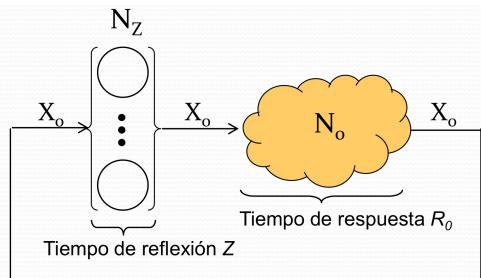
$$R_0 \neq R_1 + R_2 + \dots + R_K = \sum_{i=1}^K R_i$$

### Ley del tiempo de respuesta interactivo

- Una red cerrada siempre está en equilibrio de flujo (si el tamaño de las colas es  $\geq NT$ ).
- Al ser una red cerrada, el número total de trabajos (=clientes) en la red ( $NT=NZ+N0$ ), es constante.

Aplicamos la ley de Little a diversas partes de la red de colas:

- Ley de Little aplicada a los clientes en reflexión:  $N_0 = X_0 \times Z$ , donde  $N_Z$  = Número medio de clientes (=trabajos) en reflexión.
- Ley de Little aplicada al servidor:  $N_0 = X_0 \times R_0$



$$\begin{aligned} N_T &= N_Z + N_0 = X_0 \times Z + X_0 \times R_0 = \\ &= X_0 \times (Z + R_0) \end{aligned}$$

$$R_0 = \frac{N_T}{X_0} - Z$$

## 3. Límites optimistas del rendimiento

### 3.1 Cuello de botella

Al elemento limitador del rendimiento del servidor se le denomina cuello de botella (bottleneck). Además, puede haber más de uno de estos elementos limitadores. La localización del elemento limitador no solo depende del servidor sino también del tipo carga. La única manera de mejorar las prestaciones de un servidor de manera significativa es actuando sobre el cuello de botella.



## Identificación del cuello de botella

El cuello de botella es el dispositivo que primero llegará a **saturarse** (utilización media = 1) cuando aumente la carga ( $\lambda_0$  mayor)

$$U_i = X_0 \times D_i = \lambda_0 \times D_i \quad \xrightarrow{\text{Si equilibrio de flujo}} \quad D_i = \frac{B_i}{C_o} = \frac{B_i/\tau}{C_o/\tau} = \frac{U_i}{X_0} \rightarrow U_i = X_0 \cdot D_i$$

Como  $U_i \propto D_i$  podemos identificar fácilmente el cuello de botella de un servidor simplemente identificando el dispositivo con mayor demanda de servicio o con mayor utilización.

No hace falta llevar el servidor al límite para identificar el cuello de botella.

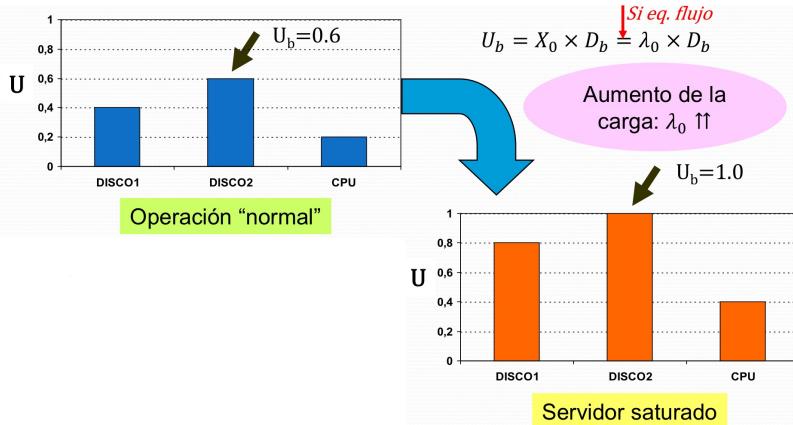
Como  $D_i = V_i \times S_i$  concluimos que la localización del cuello de botella no solo depende de lo rápido que sea el dispositivo ( $S_i$ ) sino también del tipo de carga a la que está sometido ( $V_i$ ).

Denotaremos por “ $b$ ” (bottleneck) al índice del dispositivo cuello de botella. Su demanda de servicio y su utilización vendrán dadas por:

$$D_b = \max_{i=1 \dots K} \{D_i\} = V_b \times S_b \quad U_b = \max_{i=1 \dots K} \{U_i\} = X_0 \times D_b$$

### 3.2 Saturación del servidor

El servidor se satura cuando lo hace el cuello de botella, ya que éste será el primer dispositivo en alcanzar una utilización media = 1 cuando aumente la carga.



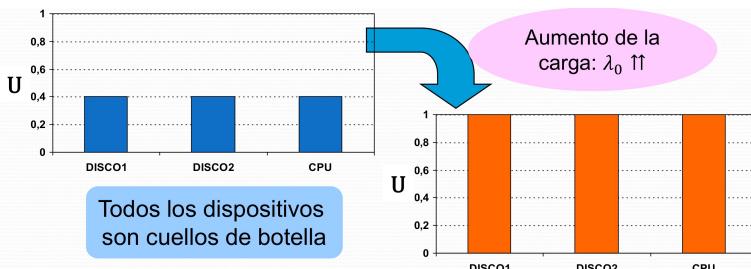
En saturación, el cuello de botella está al máximo de su productividad:

$$1 = U_b = X_b \times S_b \quad \rightarrow \quad X_b = 1/S_b$$

El equilibrio de flujo siempre se cumple mientras el servidor no esté saturado.

### Servidor equilibrado

Servidor en que todos los dispositivos, de media, tienen la misma demanda de servicio y utilización (la carga se absorbe equitativamente):



$$U_i \approx U_j \quad \forall i, j = 1 \dots K$$

$$D_i \approx D_j \quad \forall i, j = 1 \dots K$$

### 3.3 Límites del rendimiento de un servidor

Se trata de estimar las prestaciones límite de un servidor ( $R_0, X_0$ ) en los casos extremos de alta y baja cargas.

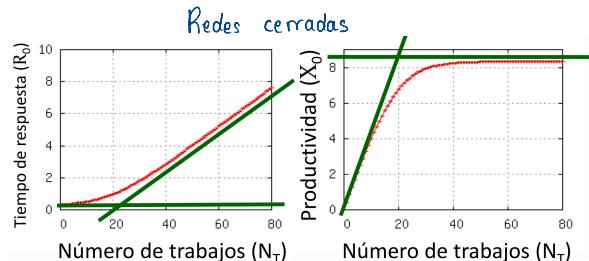
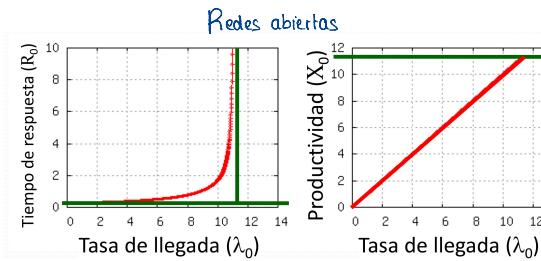
Esencialmente, se trata de estimar una cota superior de la productividad e inferior para el tiempo de respuesta del servidor por lo que a estos límites se les suele denominar límites optimistas del rendimiento. En particular, debemos preguntarnos:

- ¿Cuál es la productividad máxima ( $X_{0\max}$ ) del servidor?
- ¿Cuál es el tiempo de respuesta mínimo ( $R_{0\min}$ ) del servidor?

¿Para qué me sirve esto?:

- Para poder estimar la capacidad del servidor.
- Para poder estimar la mejora potencial de prestaciones que pueden reportar ciertas acciones sobre el servidor.

### 3.3 Localización de los límites de rendimiento



#### Límites optimistas

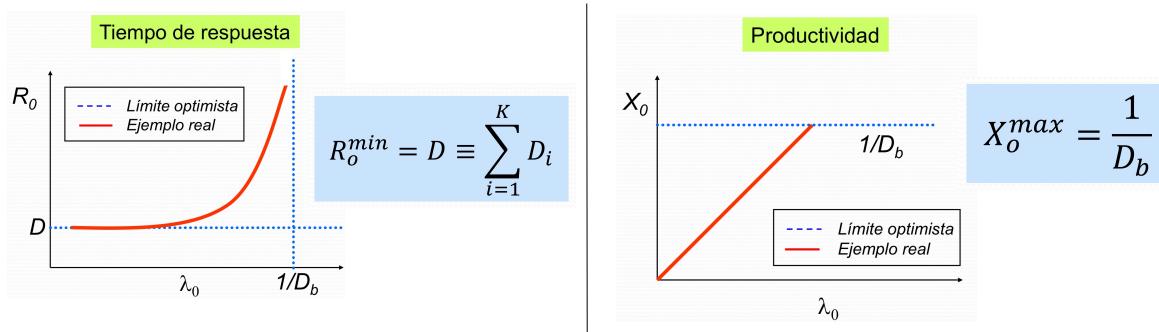
##### Redes abiertas

El valor máximo de la productividad del servidor será aquél producido por una tasa de llegada que sature el dispositivo cuello de botella ( $U_b=1$ )

Una tasa de llegada mayor provocaría un aumento descontrolado de la cola del cuello de botella hasta su desbordamiento final y, por tanto, dejaría de cumplirse la hipótesis del equilibrio de flujo ( $X_0$  ya no podría seguir a  $\lambda_0$ ).  $R_0$  crecería igualmente sin control.

El valor más optimista (= el valor mínimo) del tiempo medio de respuesta del servidor ( $R_{0\min}$ ) será el que experimenta un trabajo cuando llega al servidor sin que haya otros trabajos previamente ( $W_i = 0, \forall i=1\dots K$ ):

$$R_0 = \sum_{i=1}^K V_i \times R_i = \sum_{i=1}^K V_i \times (W_i + S_i) \quad \Rightarrow \quad R_0 \rightarrow R_{0\min} = \sum_{i=1}^K V_i \times S_i = \sum_{i=1}^K D_i \equiv D$$



#### EJEMPLO RED ABIERTA

Dispositivo	$V_i$	$S_i$ (ms)	$D_i$ (s)
CPU	16	10	0,16
DISCO A	7	20	0,14
DISCO B	8	30	0,24

cuello botella

• Tpo. respuesta mínimo:  $R_{0\min} = 0'16 + 0'14 + 0'24 = 0'54$

• Productividad máxima:  $X_0^{\max} = \frac{1}{D_b} = \frac{1}{0'24} = 4'2 \text{ tr/s}$

• Utilización máxima de CPU:  $U_{CPU}^{\max} = X_0^{\max} \cdot D_{CPU} = 0'67 (67\%)$

$$\left. \begin{aligned} \hookrightarrow V_i = \frac{C_i}{C_0} = \frac{C_i/\tau}{C_0/\tau} = \frac{X_i}{X_0} \rightarrow X_i = X_0 \cdot V_i \\ D_i = V_i \cdot S_i \rightarrow V_i = \frac{D_i}{S_i} \end{aligned} \right\} \rightarrow S_i = \frac{B_i}{C_i} = \frac{B_i/\tau}{C_i/\tau} = \frac{V_i}{X_i} \rightarrow U_i = X_i \cdot S_i = X_0 \cdot V_i \cdot S_i = X_0 \cdot D_i / S_i \cdot S_i = X_0 \cdot D_i = U_i$$

•  $U_i$  con  $\lambda_0 = 2 \text{ trabajos/s}$ :

↳ Como  $\lambda_0 < X_0^{\max}$  estamos en equilibrio de flujo, luego:  $X_0 = \lambda_0$ .

•  $U_{CPU}: X_0 \cdot D_{CPU} = 0'32 / \bullet U_A: X_0 \cdot D_A = 0'28 / \bullet U_B: X_0 \cdot D_B = 0'48$

## Redes cerradas

Ley de Little a la red completa ( $N_T = N_0 + Nz$ ):  $N_T = X_0 \times (R_0 + Z)$

$$R_0 = \frac{N_T}{X_0} - Z \quad X_0 = \frac{N_T}{R_0 + Z}$$

### a) Para valores de carga altos ( $N_T$ grande):

- Valor optimista de la productividad: Cuando el dispositivo cuello de botella esté cerca de la saturación:  $U_b = X_0 \times D_b$

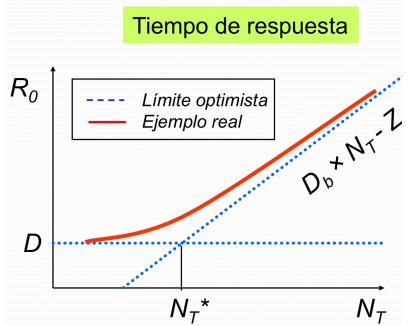
$$\text{Si } U_b \rightarrow 1 \Rightarrow X_0 \rightarrow X_0^{\max} = \frac{1}{D_b}$$

### b) Para valores de carga bajos ( $N_T$ pequeño):

- Valor optimista del tiempo de respuesta: cuando los trabajos siempre encuentran los dispositivos sin ocupar ( $W_i = 0 \forall i=1\dots K$ ):

$$R_0 = \sum_{i=1}^K V_i \times R_i = \sum_{i=1}^K V_i \times (W_i + S_i)$$

$$R_0 \rightarrow R_0^{\min} = \sum_{i=1}^K V_i \times S_i = \sum_{i=1}^K D_i \equiv D$$



$$R_0 \geq \max\{D, D_b \times N_T - Z\}$$

$$D = D_b \cdot N_T^* - Z \rightarrow N_T^* = \frac{D+Z}{D_b}$$

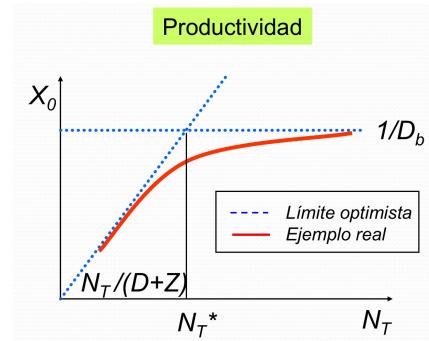
↓  
pto. teórico de saturación

- Valor optimista del tiempo de respuesta, a partir del valor optimista de la productividad (sin más que reemplazar ese valor de  $X_0$  en la ley de Little a la red completa):

$$R_0 \rightarrow \left( \frac{N_T}{X_0^{\max}} \right) - Z = D_b \times N_T - Z$$

- Valor optimista de la productividad a partir del valor optimista del tiempo de respuesta (sin más que reemplazar ese valor de  $R_0$  en la Ley de Little completa):

$$X_0 \rightarrow \frac{N_T}{R_0^{\min} + Z} = \frac{N_T}{D + Z}$$



$$X_0 \leq \min \left\{ \frac{N_T}{D+Z}, \frac{1}{D_b} \right\}$$

$$\frac{N_T^*}{D+Z} = \frac{1}{D_b} \rightarrow N_T^* = \frac{D+Z}{D_b}$$

### Punto teórico de "saturación" (knee point)

Es el valor de  $N_T$  en donde las asíntotas coinciden:  $D = D_b \cdot N_T^* - Z \rightarrow N_T^* = \frac{D+Z}{D_b}$

Propiedades del punto teórico de "saturación"  $N_T^*$ :

- Para un número total de trabajos  $N_T > N_T^*$ , los límites asintóticos vienen impuestos únicamente por el cuello de botella del servidor.
- A partir de  $N_T^*$  trabajos ya no se puede conseguir el tiempo de respuesta mínimo ya que se empiezan a formar colas de espera en, al menos, el dispositivo cuello de botella (en la práctica, esto sucede bastante antes).
- En principio, podría parecer el número ideal de trabajos en la red ya que, al menos teóricamente, para  $N = N^*$  se podría conseguir la productividad máxima y el tiempo de respuesta mínimo absolutos del servidor (en la práctica esto nunca se puede conseguir de forma simultánea):  $N_T^* = X_0^{\max} \times (R_0^{\min} + Z) = (1/D_b) \times (D + Z) = (D + Z)/D_b$

### EJEMPLO RED CERRADA

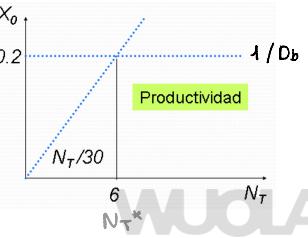
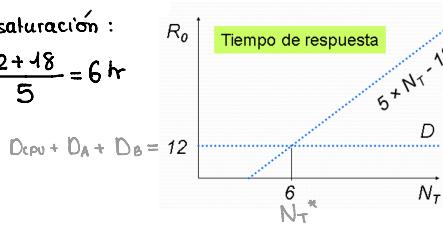
Tiempo de reflexión (Z)		18 s	
Dispositivo	$V_i$	$S_i$ (s)	$D_i$ (s)
CPU	5	1	5
DISCO A	2	2	4
DISCO B	2	1,5	3

• Límites optimistas:  $\begin{cases} R_0 \geq \max \{ D, D_b \cdot N_T - Z \} = \max \{ 12, 5 \cdot N_T - 18 \} \\ X_0 \leq \min \left\{ \frac{N_T}{D+Z}, \frac{1}{D_b} \right\} = \min \left\{ \frac{N_T}{30}, 0.2 \right\} \end{cases}$

• Pto. teórico de saturación:

$$N_T^* = \frac{D+Z}{D_b} = \frac{12+18}{5} = 6 \text{ hr}$$

$$D_{CPU} + D_A + D_B = 12$$



## 4. Técnicas de mejora

Para mejorar las prestaciones de manera significativa hay que actuar sobre el cuello de botella del servidor.

### 4.1 Sintonización o ajuste (tuning)

Optimización del funcionamiento de componentes existentes:

- **Componentes hardware:** parámetros de la placa base (frecuencias, voltajes, etc.)
- **Aplicaciones.** Ficheros de configuración. Profilers (si podemos cambiar el código fuente).
- **Sistema operativo:** políticas de gestión de procesos, memoria, almacenamiento y red: sysctl, /sys, nice, renice, taskset, ulimit, chcpu, tune2fs, ionice, hdparm, blockdev, ethtool, tc, ip, route, stap, kpatch, ...

Algunos inconvenientes:

- Posible alteración de la fiabilidad.
- Hay que conocer muy bien el funcionamiento de los componentes hardware, la aplicación y el S.O.
- Deberíamos realizar tests estadísticos para ver qué factores realmente influyen en las prestaciones.

### 4.2 Actualización y/o ampliación

- Reemplazar dispositivos por otros más rápidos -> Disminuimos el tiempo medio de servicio.
  - Procesador, memoria, disco, conexión de red, ...
- Añadir dispositivos para poder realizar más tareas en paralelo -> Disminuimos la razón media de visita.
  - Ejemplo: multiprocesadores, múltiples DIMM, matrices de discos (RAID),...

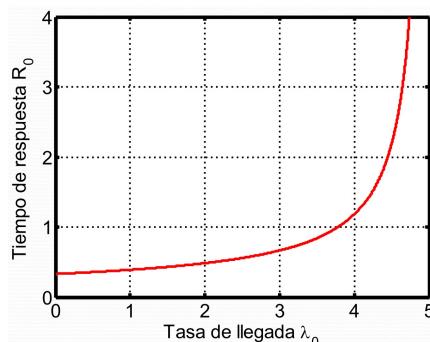
Algunos problemas:

- Facilidad del servidor para dejarse actualizar (extensibilidad/escalabilidad).
- Compatibilidad de los nuevos elementos con los existentes.

#### EJEMPLO

Servidor web modelado mediante una red abierta

Dispositivo	$V_i$	$S_i$ (s)	$D_i$ (s)
CPU	10	0,02	0,2
DISCO A	4	0,02	0,08
DISCO B	5	0,01	0,05



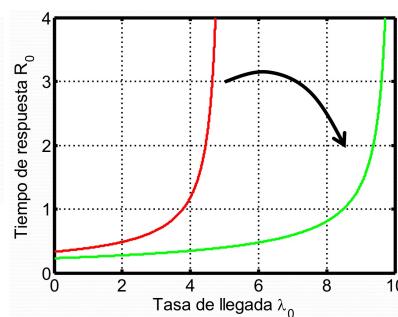
$$R_o^{\min} = 0,33 \text{ s}$$

$$X_o^{\max} = \frac{1}{D_b} = \frac{1}{0,2} = 5 \text{ tr/s}$$

Actualización: CPU doble de rápida

Dispositivo	$V_i$	$S_i$ (s)	$D_i$ (s)
CPU	10	0,01	0,1
DISCO A	4	0,02	0,08
DISCO B	5	0,01	0,05

La CPU se mantiene como cuello de botella pero con menor demanda



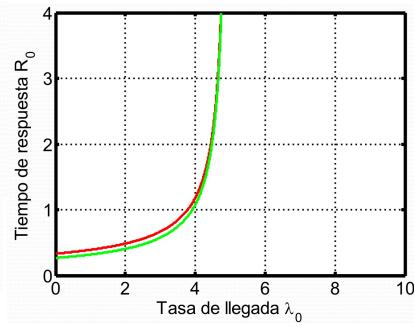
$$R_o^{\min} = 0,23 \text{ s}$$

$$X_o^{\max} = \frac{1}{D_b} = \frac{1}{0,1} = 10 \text{ tr/s}$$

Actualización: discos dobles de rápidos:

Dispositivo	$V_i$	$S_i$ (s)	$D_i$ (s)
CPU	10	0,02	0,2
DISCO A	4	0,01	0,04
DISCO B	5	0,005	0,025

La CPU se mantiene como cuello de botella



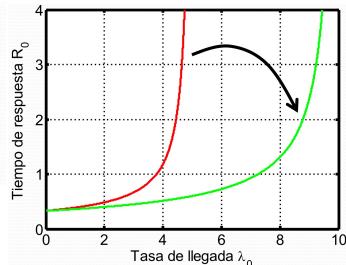
$$R_o^{\min} = 0,265 \text{ s}$$

$$X_o^{\max} = \frac{1}{D_b} = \frac{1}{0,2} = 5 \text{ tr/s}$$

Ampliación: añadimos una segunda CPU.

Dispositivo	$V_i$	$S_i$ (s)	$D_i$ (s)
CPU	5	0,02	0,1
DISCO A	4	0,02	0,08
DISCO B	5	0,01	0,05
CPU 2	5	0,02	0,1

Suponemos que el S.O. equilibra la carga entre ambas CPU



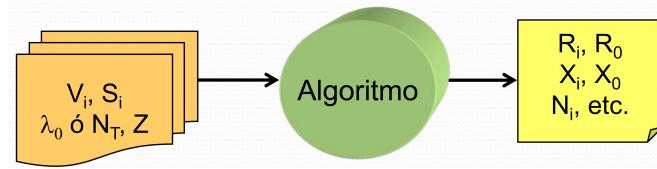
$$R_o^{\min} = 0,33 \text{ s}$$

$$X_o^{\max} = \frac{1}{D_b} = \frac{1}{0,1} = 10 \text{ tr/s}$$

## 5. Algoritmos de resolución de modelos de redes de colas

Supondremos conocido:

- El número de estaciones de servicio (K).
- Por cada estación:
  - Razón de visita medio de cada estación ( $V_i$ ).
  - Tiempo de servicio medio de cada estación ( $S_i$ ).
- Si la red es abierta: Tasa de llegada al servidor ( $\lambda_0$ ).
- Si la red es cerrada:
  - Número total de trabajos en el sistema servidor+clientes (NT).
  - Tiempo medio de reflexión de los clientes (Z).



### 5.1 Redes abiertas: hipótesis de la independencia en la llegada de trabajos

Para redes abiertas en equilibrio de flujo, vamos a suponer que el momento en el que llega un trabajo es independiente de cuándo llegó el trabajo anterior (memoryless property). Esto se puede conseguir si suponemos que todas las distribuciones de probabilidad en la red de colas se rigen por una distribución de tipo exponencial  $P(x) = \lambda e^{-\lambda x}$ .

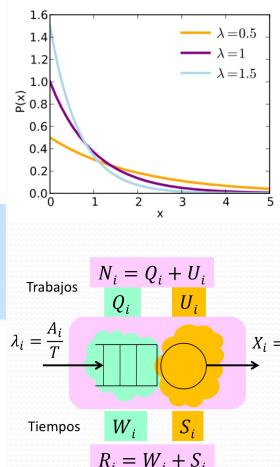
En ese caso, se puede demostrar que cuando un trabajo llega a la estación de servicio  $i$ -ésima tiene que esperar a que se procesen todos los  $N_i$  trabajos que, de media, hay en ese momento en la estación, uno comenzando a ser servido y el resto esperando:

$$W_i = N_i \times S_i$$

Por lo tanto, el tiempo de respuesta medio vendrá dado por:  $R_i = W_i + S_i = N_i \times S_i + S_i$

Aplicando la Ley de Little, ya que estamos en equilibrio de flujo:

$$R_i = X_i \times R_i \times S_i + S_i \Rightarrow R_i = \frac{S_i}{1 - X_i \times S_i} = \frac{S_i}{1 - U_i} = \frac{S_i}{1 - X_0 \times D_i} = \frac{S_i}{1 - \lambda_0 \times V_i \times S_i}$$



Extra ★ Cómo calculo  $R_i$  a partir de  $\lambda_0$ ,  $S_i$  y  $V_i$ ?

En equilibrio de flujo:

$$R_i = W_i + S_i = N_i \cdot S_i + S_i \rightarrow \text{Aplico la Ley de Little } (N_i = X_i \cdot R_i) \rightarrow$$

$$\rightarrow R_i = X_i \cdot R_i \cdot S_i + S_i \rightarrow \text{despejo } R_i \rightarrow R_i = \frac{S_i}{1 - X_i \cdot S_i} = \frac{S_i}{1 - U_i} = \frac{S_i}{1 - X_0 \cdot D_i} = \frac{S_i}{1 - \lambda_0 \cdot V_i \cdot S_i}$$

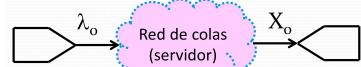
●  $S_i = \frac{B_i}{C_i} = \frac{B_i/T}{C_i/T} = \frac{U_i}{X_i} \rightarrow U_i = X_i \cdot S_i$

●  $D_i = \frac{B_i}{C_0} = \frac{B_i/T}{C_0/T} = \frac{U_i}{X_0} \rightarrow U_i = X_0 \cdot D_i$

● En eq. flujo:  $X_0 \approx \lambda_0$  y  $D_i \approx V_i \cdot S_i$

## 5.2 Algoritmo de resolución de redes de colas abiertas

Suponemos conocidos :  $\lambda_0$  (=X0),  $V_i$  y  $S_i$   $\forall i=1..K$ , y que el servidor está en equilibrio de flujo:



• **Paso 1.-** Calculamos la demanda media de servicio de cada estación:  $D_i = V_i \times S_i$

• **Paso 2.-** Calculamos el tiempo medio de respuesta de cada estación  $R_i = \frac{S_i}{1 - X_i \times S_i} = \frac{S_i}{1 - U_i} = \frac{S_i}{1 - X_0 \times D_i}$   
usando la hipótesis  $W_i = N_i \times S_i$ :

• **Paso 3.-** Calculamos el tiempo medio de respuesta del servidor:  $R_0 = \sum_{i=1}^K V_i \times R_i = \sum_{i=1}^K \frac{V_i \times S_i}{1 - X_0 \times D_i} = \sum_{i=1}^K \frac{D_i}{1 - X_0 \times D_i}$

El resto de variables operacionales ( $X_i$ ,  $U_i$ ,  $N_i$ ,  $N_0$ ,  $W_i$ ,  $Q_i$ ,...) se pueden calcular usando sus expresiones habituales.

### EJEMPLO: resolución de redes abiertas

Recurso	$V_i$	$S_i$ (s)
CPU	9	0,010
DISCO	3	0,020
RED	5	0,016

Suponiendo que la tasa de llegada de peticiones al servidor es de 5 peticiones/s:

- a) Calcule las demandas de servicio de cada recurso.
- b) ¿Qué recurso es el cuello de botella? ¿Cuál es la productividad máxima del servidor? ¿Está el servidor saturado?

Partiendo de la hipótesis de la independencia en la llegada de trabajos:

- c) Calcule el tiempo de respuesta de cada recurso y del servidor.
- d) Calcule el número medio de clientes conectados al servidor (=trabajos en el servidor).
- e) Calcule el tiempo medio de espera en la cola y el número medio de trabajos en la cola de cada recurso.

a)

Recurso	$V_i$	$S_i$ (s)	$D_i$ (s)	$U_i$
CPU	9	0,010	0,09	0,45
DISCO	3	0,020	0,06	0,30
RED	5	0,016	0,08	0,40

b) La CPU es el cuello de botella (el recurso de mayor demanda de servicio).

- Productividad máxima del servidor:

$$X_0^{max} = 1/D_b = 11,1 \text{ tr/s}$$

Como  $\lambda_0 = 5 \text{ tr/s} \leq X_0^{max}$  el servidor está en equilibrio de flujo ( $X_0 = \lambda_0$ ). Calculamos  $U_i = X_0 \times D_i = 5 \frac{\text{tr}}{\text{s}} \times D_i$  y comprobamos que  $U_b < 1$ .

c) Hipótesis de la independencia en las llegadas de trabajos:  $R_i = N_i \times S_i + S_i \Rightarrow R_i = \frac{S_i}{1 - X_0 \times D_i}$

$$R_{CPU} = \frac{S_{CPU}}{1 - X_0 \times D_{CPU}} = \frac{0,01 \text{ s}}{1 - 5 \text{ tr/s} \times 0,09 \text{ s}} = 0,018 \text{ s}$$

Igualmente,  $R_{DISCO} = 0,029 \text{ s}$ ,  $R_{RED} = 0,027 \text{ s}$ .

Finalmente,  $R_0 = V_{CPU} \times R_{CPU} + V_{DISCO} \times R_{DISCO} + V_{RED} \times R_{RED} = 0,38 \text{ s}$

d)  $N_0 = X_0 \times R_0 = 5 \frac{\text{tr}}{\text{s}} \times 0,38 \text{ s} = 1,9 \text{ clientes}$

e)

$R_i$ (s)	$W_i$ (s)	$Q_i$ (tr.)
0,018	0,008	0,37
0,029	0,009	0,13
0,027	0,011	0,27

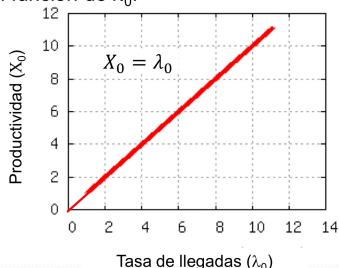
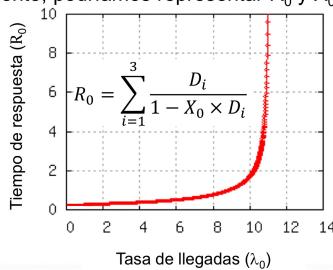
- $R_i = \frac{S_i}{1 - X_0 \times D_i}$
- $W_i = R_i - S_i$
- $Q_i = \lambda_i \times W_i = X_i \times W_i = X_0 \times V_i \times W_i$

$$V_i = \frac{C_i}{C_0} = \frac{C_i/T}{C_0/T} = \frac{X_i}{X_0} \Rightarrow X_i = V_i \cdot X_0$$

Otra forma (sólo si independencia en las llegadas de trabajos):

$$\bullet Q_i = N_i - U_i = \frac{W_i}{S_i} - U_i$$

Adicionalmente, podríamos representar  $R_0$  y  $X_0$  en función de  $\lambda_0$ :



### Resolución con solvenet

Programa muy sencillo que resuelve redes de colas utilizando los algoritmos de esta sección (los parámetros del modelo se indican en la línea de comandos).

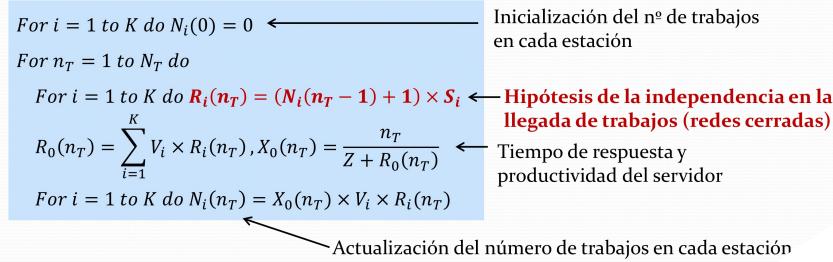
```
Usage: solvenet [0|1] [lambda0| NT Z] K S1 V1...SK VK
With no parameters, shows this message
network: 0 (open) and 1 (closed)
lambda0: arrival rate = throughput (only open networks)
NT:      total number of jobs in the net (only closed nets)
Z:       think time (only interactive closed networks)
K:       number of service stations
S1:      service time of device i
Vi:      ratio visit of device i
```

### 5.3 Resolución de redes cerradas

Suponemos conocidos:  $V_i$ ,  $S_i$ ,  $N_T$  y  $Z$ .

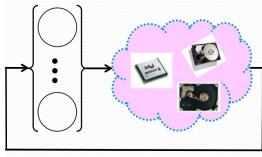
- Método: Debemos ir resolviendo la red para valores incrementales del número de trabajos en la red hasta alcanzar  $N_T$ :  $n_T = 1, \dots, N_T$ .
- Notación:  $N_i(n_T)$ : Número de trabajos en la estación de servicio  $i$ -ésima si en la red hubiese  $n$  trabajos. Ídem para los tiempos de respuesta  $R(n)$  y las productividades  $X(n)$ .
- Hipótesis de la independencia en la llegada de trabajos para redes cerradas:

$$W_i(n_T) = N_i(n_T - 1) \times S_i \Rightarrow R_i(n_T) = (N_i(n_T - 1) + 1) \times S_i$$



#### EJEMPLO: resolución de redes cerradas

T. reflexión (Z)	2 s	
Recurso	$V_i$	$S_i$ (s)
CPU	10	0,01
DISCO1	5	0,02
DISCO2	4	0,03



$$\begin{aligned}
N_{CPU}(0) &= N_{DISCO1}(0) = N_{DISCO2}(0) = 0 \\
n_T &= 1 \\
R_{CPU}(1) &= S_{CPU} = 0,01s \\
R_{DISCO1}(1) &= S_{DISCO1} = 0,02s \\
R_{DISCO2}(1) &= S_{DISCO2} = 0,03s \\
R_0(1) &= 10 \times 0,01 + 5 \times 0,02 + 4 \times 0,03 = 0,32s \\
X_0(1) &= \frac{1}{2 + 0,32} = 0,43 \text{ trabajos/s} \\
N_{CPU}(1) &= 0,43 \times 10 \times 0,01 = 0,043 \\
N_{DISCO1}(1) &= 0,43 \times 5 \times 0,02 = 0,043 \\
N_{DISCO2}(1) &= 0,43 \times 4 \times 0,03 = 0,052 \\
\\
N_{CPU}(1) &= 0,043; N_{DISCO1}(1) = 0,043 \\
N_{DISCO2}(1) &= 0,052 \\
n_T &= 2 \\
R_{CPU}(2) &= (0,043 + 1) \times S_{CPU} = 0,01043s \\
R_{DISCO1}(2) &= (0,043 + 1) \times S_{DISCO1} = 0,0209s \\
R_{DISCO2}(2) &= (0,052 + 1) \times S_{DISCO2} = 0,0316s \\
R_0(2) &= \sum_{i=1}^3 V_i \times R_i(2) = 0,3348s \\
X_0(2) &= \frac{2}{2 + 0,3348} = 0,857 \text{ trabajos/s} \\
N_{CPU}(2) &= 0,857 \times 10 \times 0,01043 = 0,0894 \\
N_{DISCO1}(2) &= 0,857 \times 5 \times 0,0209 = 0,0894 \\
N_{DISCO2}(2) &= 0,857 \times 4 \times 0,03348 = 0,1081 \\
\text{etc. hasta llegar al valor } n_t = N_T \text{ que nos pidan}
\end{aligned}$$

