

Tema 5: Análisis operacional en servidores

¿Cómo modelar el rendimiento y predecir la capacidad de mi servidor?

Analistas, administradores y diseñadores



Objetivos del tema

- Proporcionar un modelo analítico de comportamiento de un sistema informático como punto de partida para obtener índices de rendimiento.
- Entender la importancia de los cuellos de botella como limitadores del rendimiento de los sistemas informáticos.
- Saber aplicar las leyes operacionales en ejemplos sencillos para obtener índices de rendimiento.
- Saber interpretar los límites optimistas del rendimiento que establece el análisis operacional.
- Saber evaluar de forma cuantitativa el efecto de diferentes terapias de mejora o estrategias de diseño sobre el rendimiento de un servidor.

2

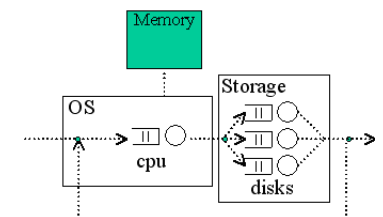
Bibliografía

- *Evaluación y modelado del rendimiento de los sistemas informáticos*. X. Molero, C. Juiz, M. Rodeño. Pearson Educación, 2004. Capítulos 4 y 5.
- *The art of computer system performance analysis*. R. Jain. John Wiley & Sons, 1991. Capítulos 30, 32, 33 y 34.
- *Measuring computer performance: a practitioner's guide*. David J. Lilja, Cambridge University Press, 2000. Capítulo 11.

3

Contenido

- [Introducción: Redes de colas de espera.](#)
- [Variables y leyes operacionales.](#)
- [Límites optimistas del rendimiento.](#)
- [Técnicas de mejora del rendimiento.](#)
- [Algoritmos de resolución de redes de colas.](#)



4

No os despistéis...

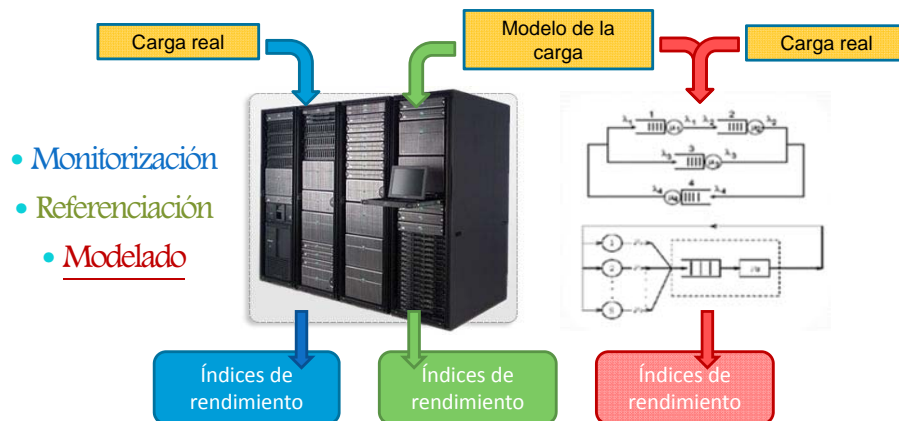


5

5.1. Introducción: Redes de Colas de Espera

Como decíamos al final del tema 2...

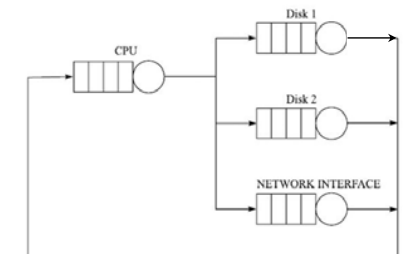
- ¿Cómo podemos evaluar/analizar/estudiar el rendimiento de un servidor?



7

El modelo de un sistema informático

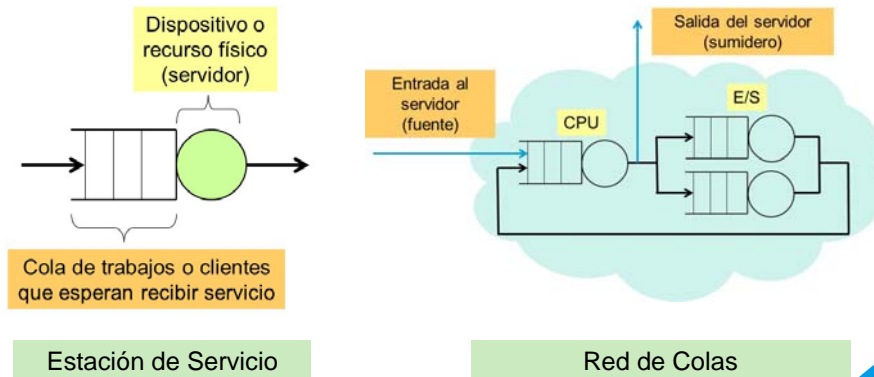
- Abstracción del sistema informático real.
 - Conjunto de dispositivos interrelacionados y trabajos que los usan (carga).
 - Dispositivos (*resources*): procesador, discos, memoria, red, etc.
 - Trabajos (*jobs*): procesos, peticiones, accesos, etc.
 - Normalmente un dispositivo o recurso solo puede ser usado por un trabajo a la vez. El resto de trabajos habrá de esperar.
- Modelos basados en redes de colas (*queueing networks*).
 - Introducidos en la década de 1950.
 - Objetivo básico: cálculo del tiempo de respuesta que experimenta un trabajo procesado por el sistema.
 - Aproximación estadística.
- Otros modelos: redes de Petri, cadenas de Markov, ...



8

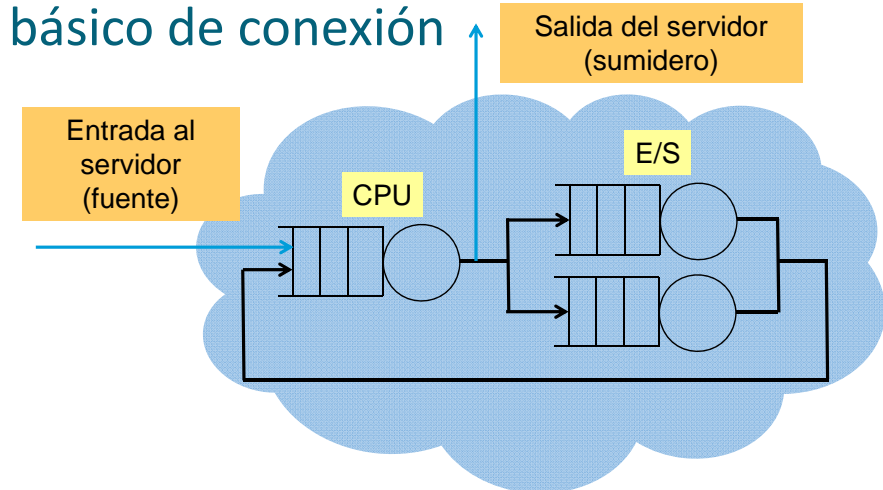
Red de colas: concepto

- Conjunto de estaciones de servicio conectadas entre sí.
- Estación de servicio (*service station*): Objeto abstracto 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.



9

Modelo de servidor central: Diagrama básico de conexión



- Propiedad de este modelo: Cada trabajo que abandona el servidor (=proceso ejecutado) ha "visitado" la CPU tantas veces como la suma de sus "visitas" a cada uno los dispositivo de E/S más 1.

10

El modelo de servidor central

- Es el modelo de redes de colas que más se ha utilizado para representar el comportamiento de los programas en un servidor de cara a extraer información sobre su rendimiento.
- ¿Cuál es este comportamiento?
 - Un trabajo que "llega" al servidor comienza utilizando el procesador.
 - Después de "abandonar" el procesador, el trabajo puede:
 - terminar (sale del servidor), o bien
 - realizar un acceso a una unidad de entrada/salida (discos, red,...).
 - Después de una operación con una unidad de entrada/salida, el trabajo vuelve a "visitar" al procesador.
- Recursos considerados
 - Procesador o procesadores.
 - Entrada/salida: unidades de disco, red, etc.

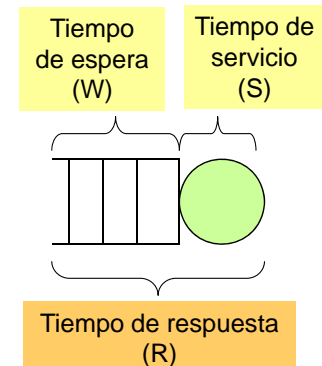
11

Variables temporales de un trabajo en una estación de servicio

- Tiempo de espera en cola (*W, waiting time*)
 - Tiempo transcurrido desde que el trabajo solicita hacer uso del recurso físico hasta que realmente empieza a utilizarlo.
- Tiempo de servicio (*S, service time*)
 - Tiempo transcurrido desde que el trabajo hace uso del recurso físico hasta que lo libera.
- Tiempo de respuesta de la estación de servicio (*R, response time*)
 - Suma de los dos tiempos anteriores.

$$R = W + S$$

- En general, para múltiples trabajos, se trata de magnitudes estadísticas (distribuciones de probabilidad).



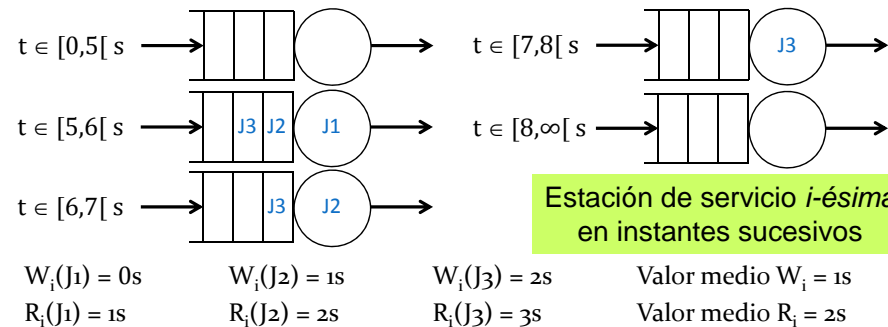
12

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).

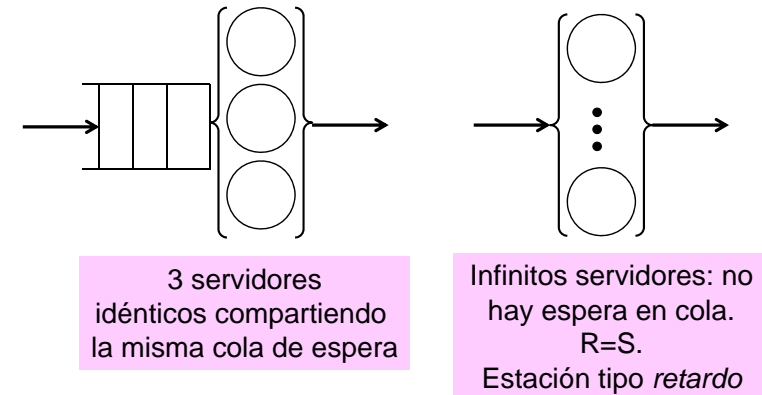
Calcule los tiempos de espera en la cola y los tiempos de respuesta que experimentan **cada uno** de los tres trabajos. Calcule finalmente los valores medios de W y R .



13

Estaciones con más de un servidor

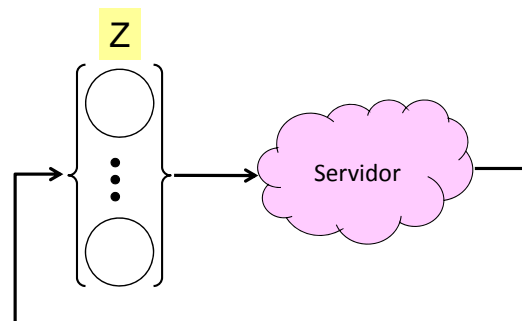
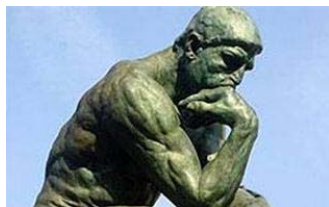
- Son capaces de atender a más de un trabajo en paralelo:



14

El tiempo de reflexión (Z , *think time*)

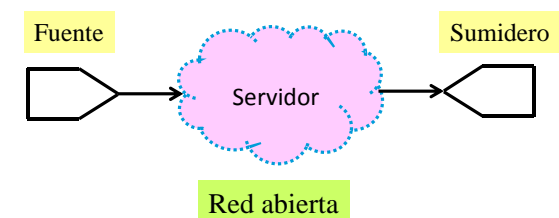
- Es un parámetro (Z) que representa el tiempo que requiere el usuario antes de volver a lanzar una petición al servidor tras la respuesta de éste.
- Se suele modelar mediante una estación de servicio tipo retardo con un tiempo de servicio = Z .



15

Redes de colas abiertas

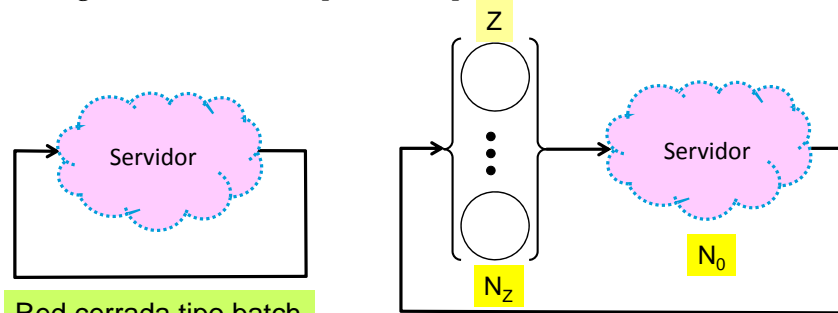
- Los trabajos llegan a la red a través de una fuente externa que no controlamos. Tras ser procesados, salen de ella a través de uno o más sumideros.
- El número de trabajos en el servidor (N_o) puede variar con el tiempo.



16

Redes de colas cerradas

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



Red cerrada tipo batch
 $N_T = N_0$

Red cerrada tipo interactivo
 $N_T = N_0 + N_Z$

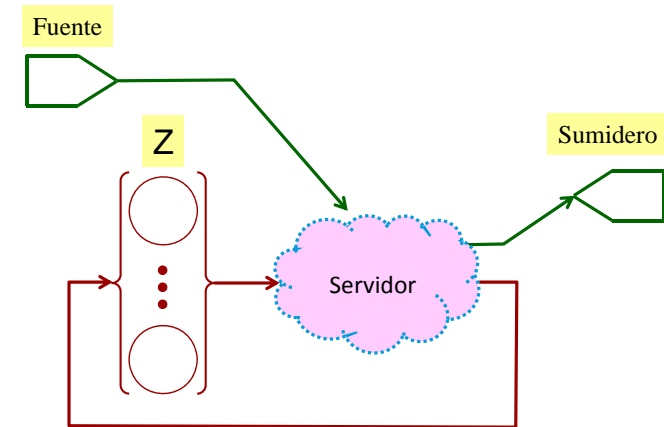
N_0 = Número de trabajos en el servidor.

N_Z = Número de trabajos en reflexión (esperando a que los usuarios vuelvan a introducirlos en el servidor). **Siempre supondremos 1 usuario = 1 trabajo.**

17

Redes mixtas

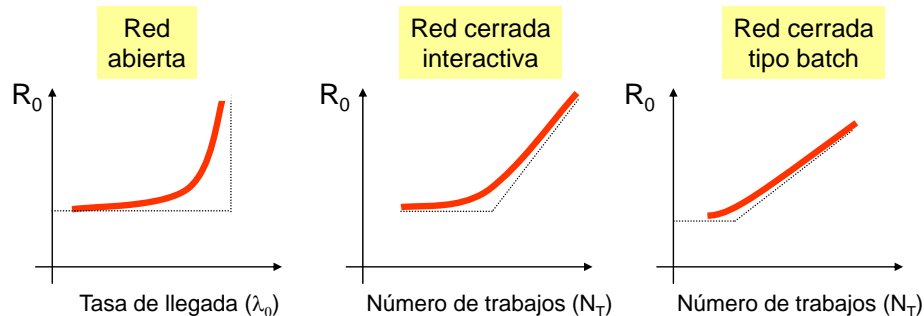
- Cuando el modelo no corresponde a ninguno de los dos anteriores.



18

Evolución del tiempo medio de respuesta del servidor (R_0) según el tipo de red

- R_0 = Tiempo que transcurre desde que el trabajo entra al servidor hasta que lo abandona = tiempo que tarda el servidor en procesar una petición.



19

5.2. Variables y leyes operacionales

El análisis operacional

- Técnica de análisis de redes de colas presentada por Denning y Buzen en 1978.
- Basada en valores medios de *variables operacionales* (=magnitudes medibles) del sistema informático.

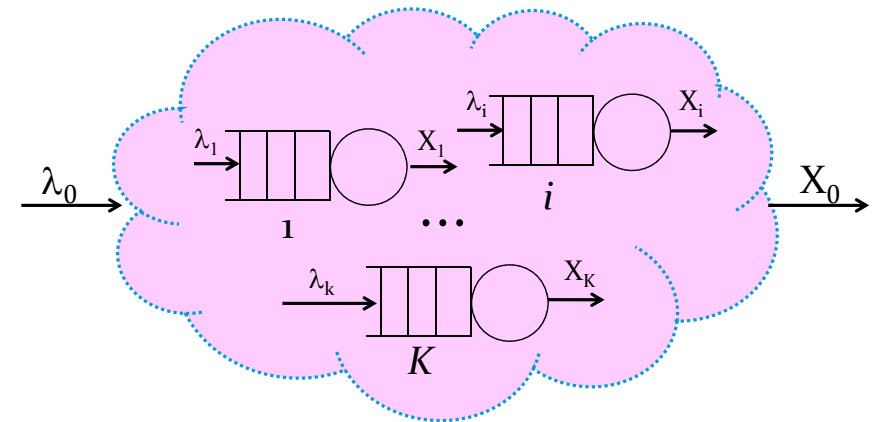


- Leyes operacionales: relaciones entre las variables operacionales.
- Límites optimistas de las prestaciones del servidor por medio de cálculos muy sencillos.

21

Variables: servidor vs. estación de servicio

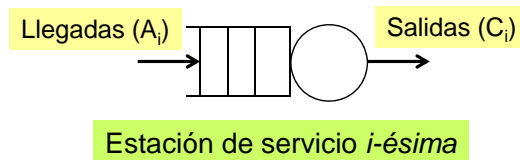
- El servidor contiene K estaciones de servicio (recursos o dispositivos).
- A todo el servidor en su globalidad lo denotamos como dispositivo “cero”.



22

Variables operacionales básicas

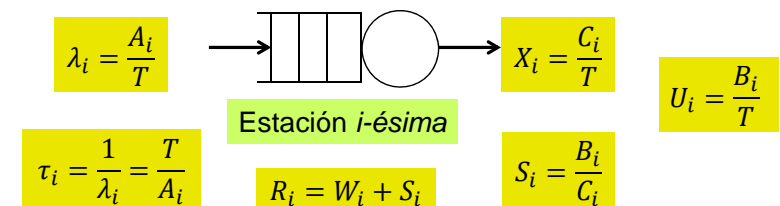
- Variable global temporal:
 - T Duración del periodo de medida para el que se extrae el modelo.
- Variables operacionales básicas de la estación de servicio i -ésima medidas durante el tiempo T :
 - A_i Número de trabajos solicitados a la estación (llegadas, *arrivals*).
 - B_i Tiempo que el dispositivo está ocupado (*busy time*).
 - C_i Número de trabajos completados por la estación (salidas, *completions*).



23

Variables operacionales deducidas

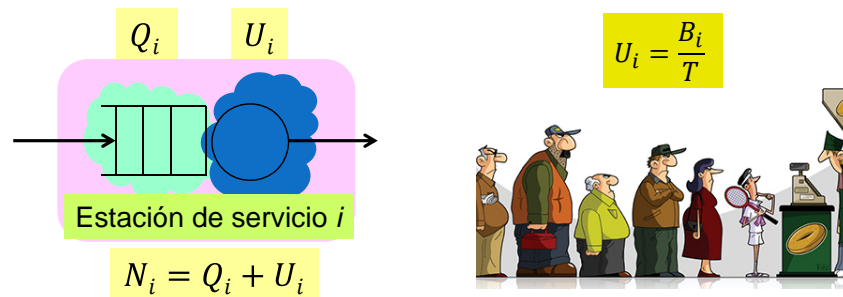
- Se deben poder estimar a partir de las variables básicas:
 - λ_i Tasa media de llegada (*arrival rate*) trabajos/segundo
 - τ_i Tiempo medio entre llegadas (*interarrival time*) segundos [/trabajo]
 - X_i Productividad media (*throughput*) trabajos/segundo
 - S_i Tiempo medio de servicio (*service time*) segundos [/trabajo]
 - W_i Tiempo medio de espera en cola (*waiting time*) segundos [/trabajo]
 - R_i Tiempo medio de respuesta (*response time*) segundos [/trabajo]
 - U_i Utilización media (*utilization*) sin unidades



24

Variables operacionales deducidas (II)

- N_i : Número medio de trabajos en la estación de servicio (cola más recurso).
- Q_i : Número medio de trabajos en cola de espera (*jobs in queue*).
- U_i : 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 $B_i=T$).



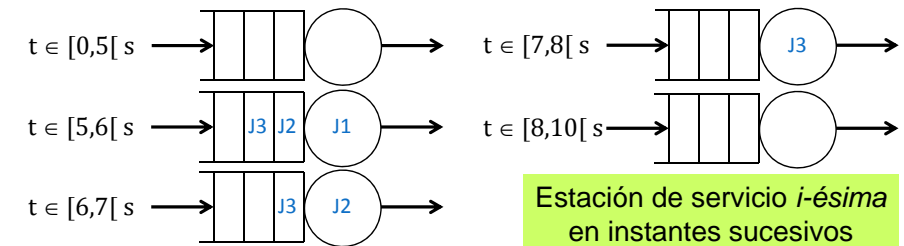
25

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_i , B_i , C_i , λ_i , τ_i , X_i , Q_i , N_i , U_i .

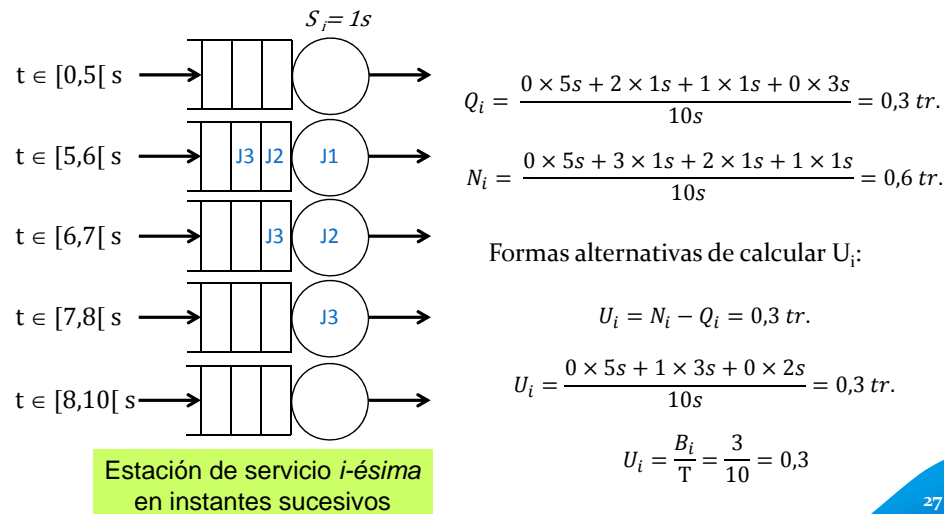


$A_i = 3$ trabajos, $B_i = 3s$, $C_i = 3$ trabajos, $\lambda_i = X_i = 3/10 = 0,3$ trabajos/s, $\tau_i = 3,3$ s.

26

Ejercicio (cont.)

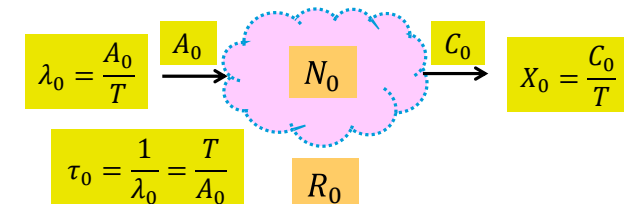
Cálculo de la utilización media (U_i) y del número medio de trabajos en la cola (Q_i) y en la estación (N_i).



27

Variables globales del servidor

- Variables básicas
 - A_0 Número de trabajos solicitados al servidor (*arrivals*).
 - C_0 Número de trabajos completados por el servidor (*completions*).
- Variables deducidas
 - λ_0 Tasa media de llegada al servidor (*arrival rate*).
 - τ_0 Tiempo medio entre llegadas al servidor (*interarrival time*).
 - X_0 Productividad media del servidor (*throughput*).
 - R_0 Tiempo medio de respuesta del servidor (*response time*).
 - N_0 Número medio de trabajos en el servidor (*#jobs*) = $N_1 + N_2 + \dots + N_K$.



28

Razón de visita y demanda de servicio

- Razón de visita media V_i (*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 (dispositivo). Es como si el trabajo tuviera que “visitar” el dispositivo i -ésimo una media de V_i veces antes de poder abandonar el servidor.

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

- Demanda de servicio media D_i (*service demand*). Cantidad de tiempo que, por término medio, el dispositivo i -ésimo dedica a cada trabajo que abandona el servidor.

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

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

29

Ejercicio

Después de monitorizar el **disco duro** de un servidor web durante un periodo de **30** segundos, se sabe que ha estado en funcionamiento un total de **27** segundos. Asimismo, se han contabilizado durante ese periodo un total de **74** peticiones de lectura/escritura al disco duro y un total de **72** peticiones completadas. Se ha estimado que cada consulta atendida por el servidor web ha requerido una media de 4 accesos de E/S al disco duro. Calcule:

- ¿Cuál es la tasa de llegada al disco duro?
- ¿Cuál es la productividad del disco duro?
- Determinése la utilización del disco duro, su tiempo de servicio y su demanda de servicio.
- ¿Cuál es la productividad 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.

30

Leyes operacionales

- Casi todas las variables utilizadas en el análisis operacional son **valores medios** para el intervalo de monitorización T . Por tanto, el valor de dichas variables 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 en un servidor **no saturado**, si se escoge un intervalo de observación suficientemente largo, se cumple que:
 - El número de trabajos que llegan al servidor coincide aproximadamente con el que sale ($A_0 \approx C_0$). Dicho de otra manera, la tasa media de llegada coincide aproximadamente con la productividad media ($\lambda_0 \approx X_0$).
 - El número de trabajos que llegan a cada estación de servicio coincide aproximadamente con el que sale: ($A_i \approx C_i \Rightarrow \lambda_i \approx X_i, \forall i=1..K$).

31

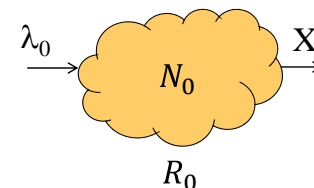
Ley de Little

“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, relaciona el número medio de trabajos en el servidor (N_0) con el tiempo medio de respuesta (R_0) y su tasa de llegada (λ_0).

$$N_0 = \lambda_0 \times R_0$$

- Bajo la hipótesis del equilibrio de flujo (servidor no saturado), 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).



Si no sat.

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

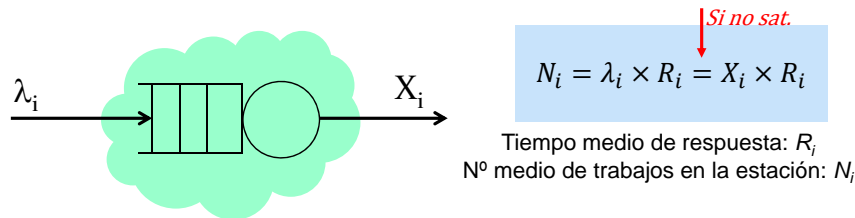
Ley de Little aplicada a un servidor no saturado

32

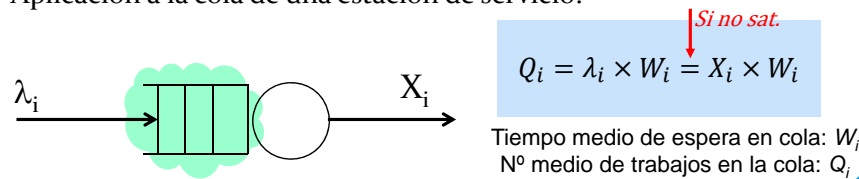
Ley de Little (cont.)

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:



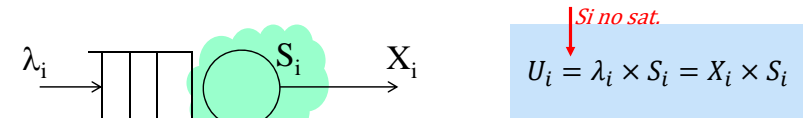
33

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):

$$U_i = \frac{B_i}{T} = \frac{C_i}{T} \times \frac{B_i}{C_i} = X_i \times S_i \quad \text{Ley de la Utilización}$$

- Para un sistema en equilibrio de flujo (servidor no saturado), se podía haber deducido también la ley de la utilización sin más que aplicar la ley de Little al recurso físico (dispositivo) de una estación:



Ley de la Utilización
(servidor no saturado)

34

Ley del flujo forzado

- Las productividades (flujos) a diferentes niveles del servidor 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:

$$V_i = \frac{C_i}{C_0} = \frac{X_i}{X_0} \quad \Rightarrow \quad X_i = X_0 \times V_i \quad \text{Ley del Flujo Forzado}$$

- 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:

$$U_i = X_i \times S_i = X_0 \times V_i \times S_i = X_0 \times D_i \quad \text{Relación Utilización-Demanda de Servicio}$$

35

Ejemplo de aplicación

- Un servidor de base de datos **no saturado** recibe una media de 120 consultas por minuto. Sabemos que el tiempo medio de respuesta del disco duro principal del servidor es 48ms, su tiempo medio de servicio es 30ms y su productividad 25 peticiones de E/S completadas por segundo. Calcule:

a) El número medio de trabajos en la cola de espera del disco duro.

- Solución alternativa 1:

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

- Solución alternativa 2:

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

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

$$Q_{dd} = N_{dd} - U_{dd} = 1,2 - 0,75 = 0,45 \text{ tr.}$$

➔ Cuestión: ¿Por qué el tiempo de respuesta (48 ms) es bastante mayor (60%) que el tiempo de servicio (30 ms) si la utilización del disco es bastante menor que 1 (0,75)?

b) ¿Cuánto tiempo, de media, consumen los accesos al disco duro por cada consulta que se realiza al servidor?

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

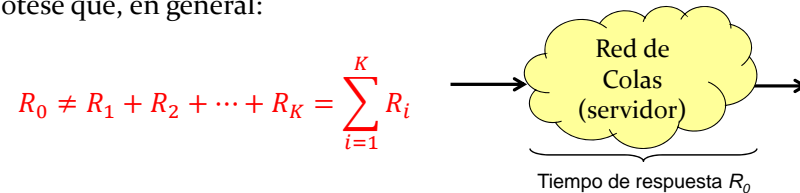
36

Ley general del tiempo de respuesta

- Si el servidor no está saturado y teniendo en cuenta que el número medio de trabajos dentro de un servidor no es más que la suma del número medio de trabajos que hay en cada estación de servicio $N_0 = N_1 + N_2 + \dots + N_K$, y que $N_i = X_i \times R_i = X_0 \times V_i \times R_i$, es fácil demostrar que:

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

expresión conocida como “Ley general del tiempo de respuesta”.
Nótese que, en general:

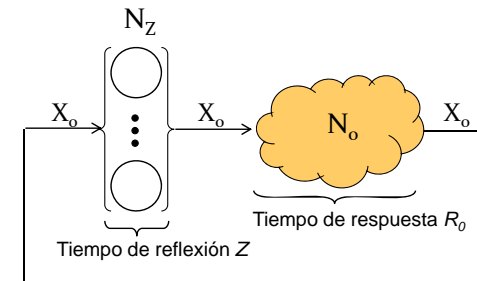


37

Ley del tiempo de respuesta interactivo

- Se obtiene mediante la aplicación de la ley de Little a una red cerrada de tipo interactivo.
- Una red cerrada nunca se satura (si el tamaño de las colas es suficiente).

Recordemos: N_Z = Número medio de trabajos (=clientes) en reflexión.



$$N_Z = X_0 \times Z; \quad N_0 = X_0 \times R_0$$

$$N_T = N_Z + N_0 = X_0 \times Z + X_0 \times R_0 = X_0 \times (Z + R_0)$$

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

que se conoce como la *Ley del tiempo de respuesta interactivo*.

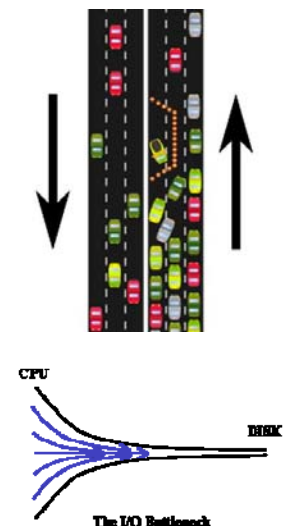
Nótese que al ser una red cerrada, el número total de trabajos (=clientes) en la red cerrada ($N_T = N_Z + N_0$), es constante.

38

5.3. Límites optimistas del rendimiento

Limitaciones en el rendimiento: cuello de botella

- Todo servidor presenta alguna limitación en su rendimiento.
- En esta sección veremos que la localización del elemento limitador no solo depende del servidor sino también de la carga.
 - Al elemento limitador del rendimiento del servidor se le denomina cuello de botella (*bottleneck*).
 - Además, puede haber más de uno de estos elementos limitadores.
- Veremos que la única manera de mejorar las prestaciones de un servidor de manera significativa es actuando sobre el cuello de botella.



40

Identificación de un cuello de botella

- El cuello de botella es el dispositivo con **mayor utilización** ya que será el primero que llegue a saturarse ($U_i=1$) cuando aumente la carga (λ_0 mayor).

$$U_i = X_0 \times D_i = \lambda_0 \times D_i$$

↓ Si no sat.

- Podemos identificar fácilmente el cuello de botella de un servidor simplemente buscando el dispositivo con **mayor demanda de servicio** ($U_i \propto D_i$).
- No hace falta llevar el servidor al límite para identificar el cuello de botella.
- Denotaremos por “b” al índice del dispositivo cuello de botella. Su demanda de servicio y su utilización vendrán dadas por.

$$D_b = \max_{i=1 \dots K} \{D_i\} = V_b \times S_b$$

$$U_b = \max_{i=1 \dots K} \{U_i\} = X_0 \times D_b$$

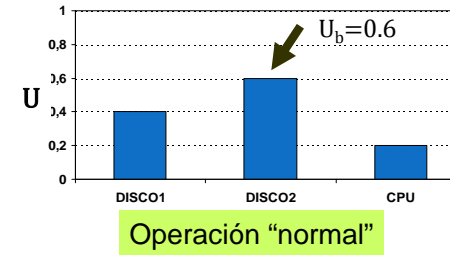
41

Saturación del servidor

- El servidor se satura cuando lo hace el cuello de botella ya que, por definición, éste será el primero en alcanzar una utilización = 1 cuando aumente la carga.

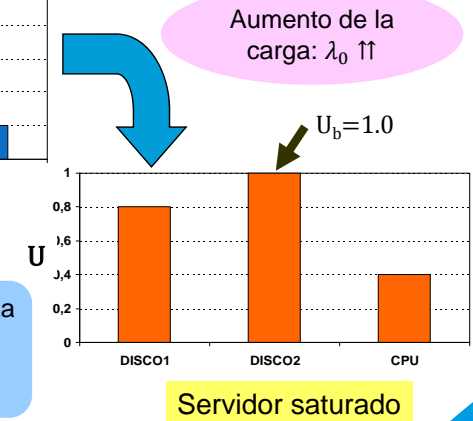
$$U_b = X_0 \times D_b = \lambda_0 \times D_b$$

↓ Si no sat.



En saturación, el servidor alcanza su productividad máxima:

$$X_0^{max} = 1/D_b$$

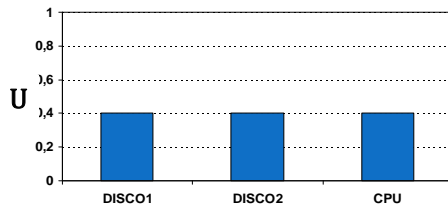


42

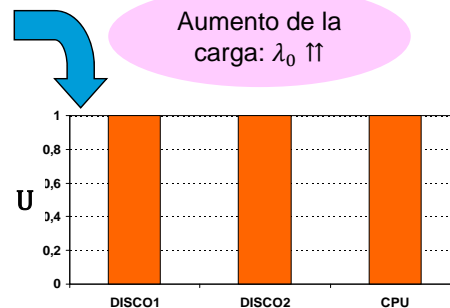
Servidor equilibrado (*balanced system*)

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

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



Todos los dispositivos son cuellos de botella



43

Límites Optimistas del Rendimiento

- Se trata de encontrar una cota superior de la productividad e inferior para el tiempo de respuesta del servidor. Son, por tanto, **límites optimistas** del rendimiento. En particular, sabremos:

- ¿Cuál es la productividad máxima (X_0^{max}) del servidor?
- ¿Cuál es el tiempo de respuesta mínimo (R_0^{min}) del servidor?

- Campos de aplicación:

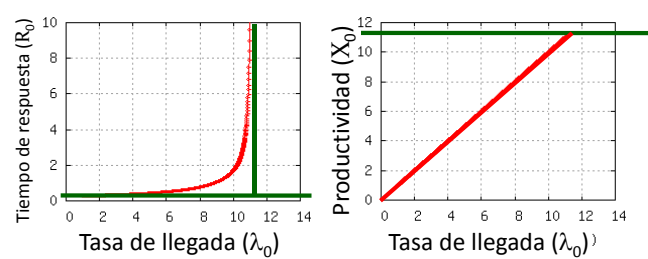
- Planificación de la capacidad (*capacity planning*).
- Estimación de la mejora potencial de prestaciones que pueden reportar ciertas acciones sobre el servidor.



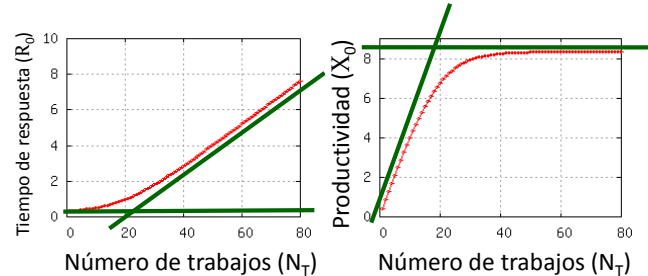
44

Localización de los límites (asíntotas)

Redes abiertas

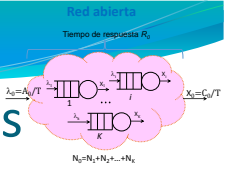


Redes cerradas



45

Límites optimistas: redes abiertas



- El valor máximo de la tasa de llegada del servidor, será aquél que sature completamente el dispositivo cuello de botella ($U_b=1$):

$$U_b = X_0 \times D_b = \lambda_0 \times D_b \quad \text{Si no sat.} \quad \text{Si } U_b = 1 \Rightarrow X_0^{max} = \frac{1}{D_b}$$

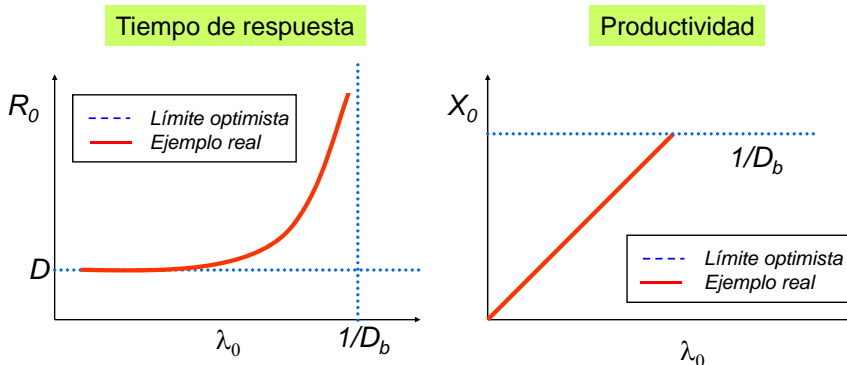
A partir de ese momento, una tasa de llegada mayor provocaría la saturación del servidor y, por tanto, dejaría de cumplirse la hipótesis del equilibrio de flujo (X_0 ya no podría seguir a λ_0).

- 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 ($W_i = 0 \rightarrow R_i^{min} = S_i \forall i=1..K$):

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

46

Resumen: Límites optimistas en redes abiertas



$$R_0^{min} = D \equiv \sum_{i=1}^K D_i$$

$$X_0^{max} = \frac{1}{D_b}$$

47

Ejemplo de red abierta

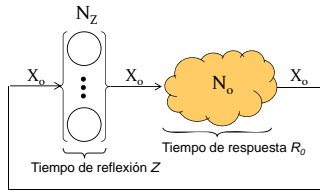
Dispositivo	V_i	S_i (ms)	D_i (s)
CPU	16	10	0,16
DISCO A	7	20	0,14
DISCO B	8	30	0,24

- Tiempo de respuesta mínimo:
 - $R_0^{min} = 0,16 + 0,14 + 0,24 = 0,54s$.
- Productividad máxima:
 - $X_0^{max} = \frac{1}{D_b} = \frac{1}{0,24} = 4,2 \text{ tr/s}$.
- Utilización máxima de la CPU:
 - $U_{CPU}^{max} = X_0^{max} \times D_{CPU} = 0,67$.
- Productividad máxima de la CPU:
 - $X_{CPU}^{max} = X_0^{max} \times V_{CPU} = 67 \text{ tr/s}$.
- U_i con $\lambda_0 = 2$ trabajos/s ($=X_0$):
 - $U_{CPU} = X_0 \times D_{CPU} = 0,32$.
 - $U_A = X_0 \times D_A = 0,28$.
 - $U_B = X_0 \times D_B = 0,48$.



48

Límites optimistas: redes cerradas



Ley de Little a la red completa
($N_T = N_Z + N_O$):

$$N_T = X_0 \times (R_0 + Z)$$

- 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 (aunque nunca llegue realmente a saturarse):

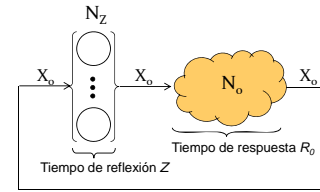
$$U_b = X_0 \times D_b \quad \text{Si } U_b \rightarrow 1 \Rightarrow X_0 \rightarrow X_0^{max} = \frac{1}{D_b}$$

- 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):

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

49

Límites optimistas: redes cerradas (II)



Ley de Little a la red completa
($N_T = N_Z + N_O$):

$$N_T = X_0 \times (R_0 + Z)$$

- b) Para valores de carga bajos (N_T pequeño):
 - Valor optimista del tiempo de respuesta: cuando los trabajos siempre encuentran los dispositivos sin ocupar ($W_i = 0$, por lo que $R_i = S_i \forall i=1..K$):

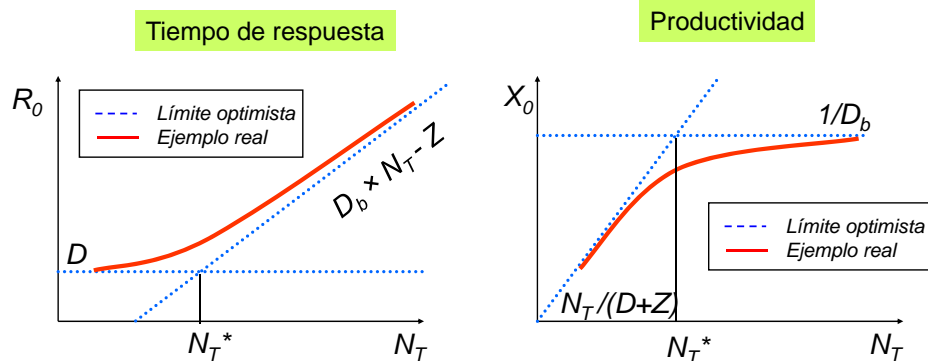
$$R_0 \rightarrow R_0^{min} = \sum_{i=1}^K V_i \times R_i^{min} = \sum_{i=1}^K V_i \times S_i = \sum_{i=1}^K D_i \equiv D$$

- 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 a la red completa):

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

50

Resumen: límites optimistas en redes cerradas



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

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

51

Punto teórico de "saturación" (knee point)

- Es el valor de N_T en donde las asíntotas coinciden:

$$D = D_b \times N_T^* - Z \Rightarrow N_T^* = \frac{D + Z}{D_b}$$

- Propiedades del punto teórico de "saturación" N_T^* :
 - Para un número total de trabajos $N_T > N_T^*$, los límites asíntóticos 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 $N_T = N_T^*$ 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): $N_T^* = X_0^{max} \times (R_0^{min} + Z) = \frac{D + Z}{D_b}$

52

Ejemplo de red cerrada

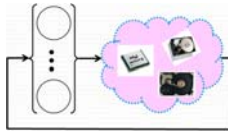
Tiempo de reflexión (Z)		18 s	
Dispositivo	V _i	S _i (s)	D _i (s)
CPU	5	1	5
DISCO A	2	2	4
DISCO B	2	1.5	3

- Asíntotas:

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

$$= \max\{12, 5 \times N_T - 18\}$$

$$X_0 \leq \min\left\{\frac{N_T}{D+Z}, \frac{1}{D_b}\right\} = \min\left\{\frac{N_T}{30}, 0.2\right\}$$

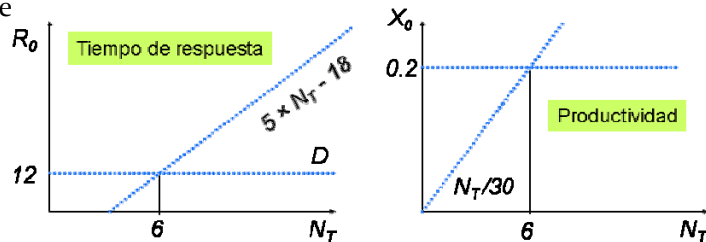


- Punto teórico de saturación:

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

$$= \frac{12+18}{5}$$

$$= 6 \text{ trabajos}$$

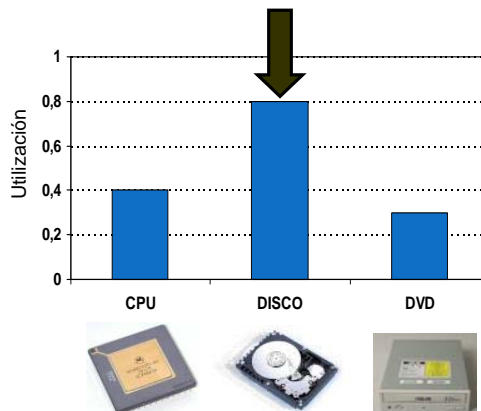


53

5.4. Técnicas de mejora

Técnicas para mejorar las prestaciones

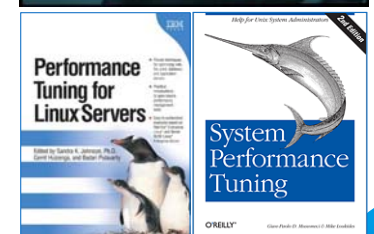
- Para mejorar las prestaciones de manera significativa hay que actuar sobre el cuello de botella del servidor.
 - Sintonización o ajuste (tuning).
 - Actualización o ampliación (upgrading).



55

Sintonización o ajuste (*tuning*)

- Optimización del funcionamiento de componentes existentes:
 - Componentes hardware: frecuencias, voltajes, parámetros de la placa, ...
 - Aplicaciones. Usamos profilers.
 - Sistema operativo: políticas de gestión de procesos y memoria virtual, distribución de la información entre discos,...
- Algunos problemas:
 - Posible alteración de la fiabilidad.
 - Hay que conocer muy bien el sistema operativo y el funcionamiento de los componentes hardware.
 - Deberíamos realizar tests estadísticos para ver qué factores realmente influyen en las prestaciones.



56

Actualización o ampliación (upgrading)

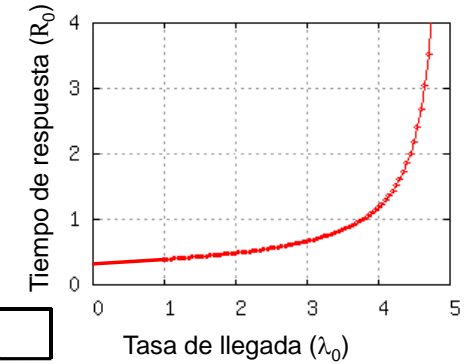
- Reemplazar dispositivos por otros más rápidos → Disminuimos el tiempo medio de servicio.
 - Procesador, memoria, placa base, disco...
- Añadir dispositivos para poder realizar más tareas en paralelo → Disminuimos la razón de visita.
 - Ejemplo: multiprocesadores, matrices de discos (RAID),...
- Algunos problemas:
 - Compatibilidad de los nuevos elementos con los existentes.
 - Facilidad del servidor para dejarse actualizar (extensibilidad/escalabilidad).



57

Ejemplo: red abierta (servidor web)

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



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

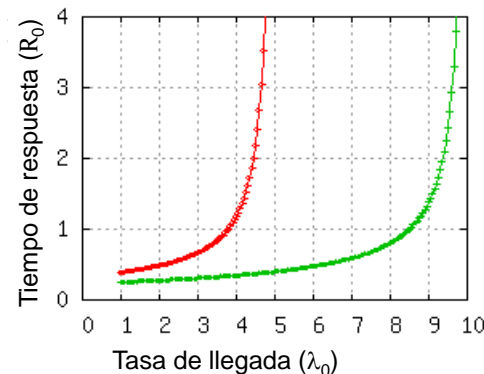
$$X_o^{\max} = \frac{1}{D_b} = \frac{1}{0,2} = 5 \text{ tr/s}$$

58

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



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

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

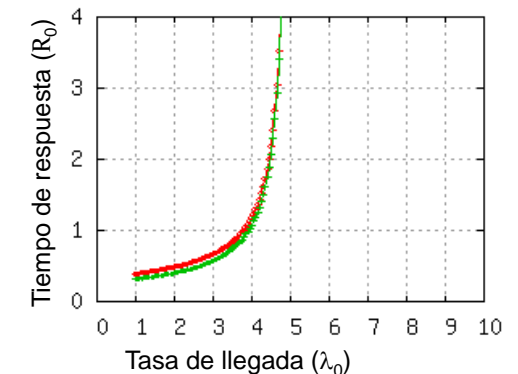


59

Actualización: discos doble 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



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

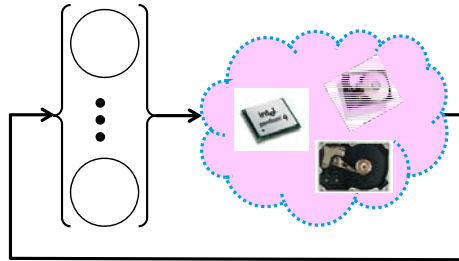
$$X_o^{\max} = \frac{1}{D_b} = \frac{1}{0,2} = 5 \text{ tr/s}$$



60

Ejemplo: red cerrada interactiva (servidor de streaming)

Tiempo de reflexión (Z)		4 s		
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	



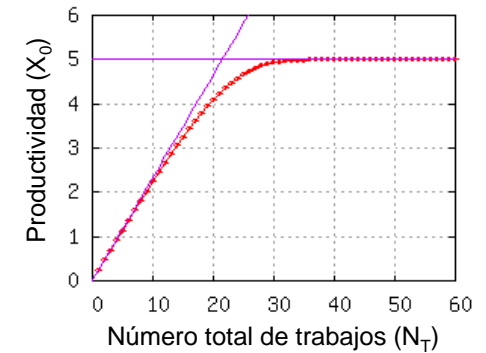
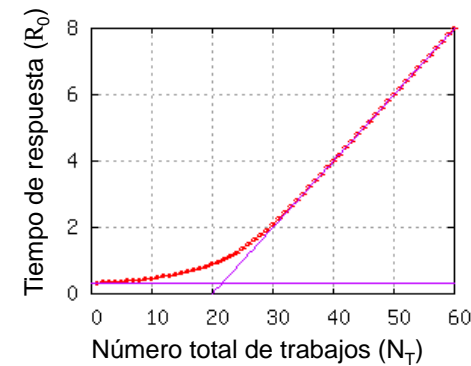
Asíntotas:

$$R_0 \geq \max\{D, D_b \times N_T - Z\} = \max\{0,33, 0,2 \times N_T - 4\}$$

$$X_0 \leq \min\left\{\frac{N_T}{D+Z}, \frac{1}{D_b}\right\} = \min\left\{\frac{N_T}{4,33}, 5\right\}$$

61

Rendimiento del servidor original

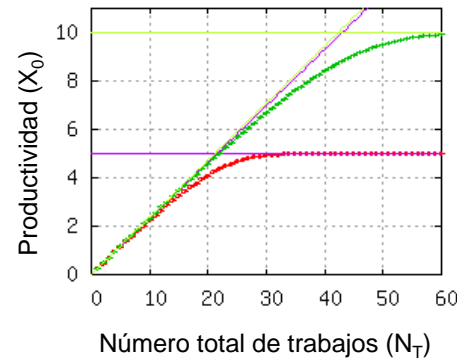
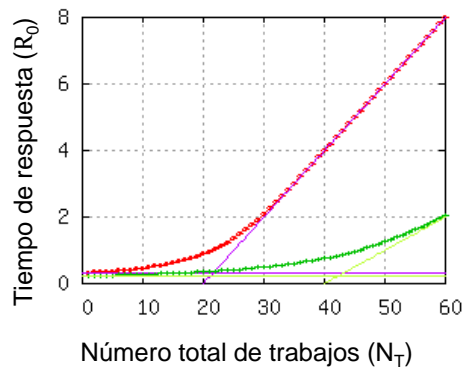


Punto teórico de saturación:
 $N_T^* = \frac{D+Z}{D_b} = 21,7$ trabajos

Los valores óptimos de ambos índices para N_T altos están determinados por el cuello de botella (CPU)

62

Actualización: CPU doble de rápida



Punto teórico de saturación:
 $N_T^* = \frac{D+Z}{D_b} = 42,3$ trabajos

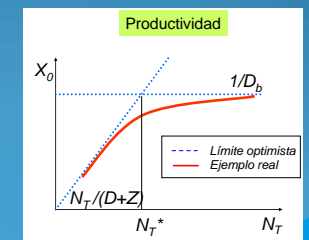
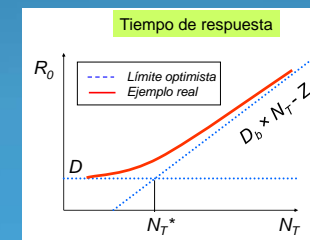
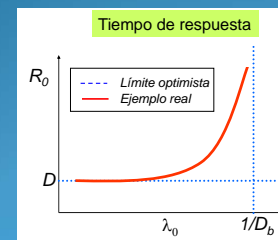
Asíntotas:

$$R_0 \geq \max\{D, D_b \times N_T - Z\} = \max\{0,23, 0,1 \times N_T - 4\}$$

$$X_0 \leq \min\left\{\frac{N_T}{D+Z}, \frac{1}{D_b}\right\} = \min\left\{\frac{N_T}{4,23}, 10\right\}$$

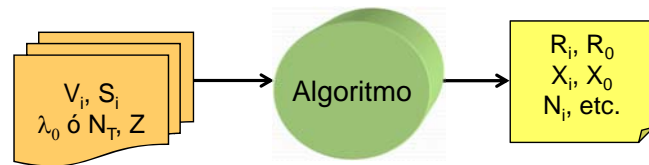
63

5.5. Algoritmos de resolución de modelos de redes de colas



Algoritmos de resolución de redes de colas

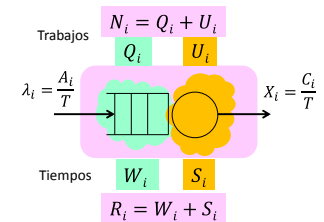
- En este apartado vamos a proporcionar una metodología (algoritmo) para resolver modelos de redes de colas. Supondremos conocido:
 - El número de estaciones de servicio (K).
 - Por cada estación:
 - Razón de visita medio de cada estación (V_i).
 - Tiempo de servicio medio de cada estación (S_i).
 - Si la red es abierta: Tasa de llegada al servidor (λ_0).
 - Si la red es cerrada:
 - Número total de trabajos en la red (N_T).
 - Si la red es interactiva: Tiempo medio de reflexión de los usuarios (Z).



65

Hipótesis del “peor escenario posible” para redes abiertas

- Cuando un trabajo llega a la estación de servicio i -ésima de una **red de colas abierta** supondremos que para poder acceder al dispositivo 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, en este peor escenario posible, por: $R_i = W_i + S_i = N_i \times S_i + S_i$

Aplicando que $N_i = \lambda_i \times R_i$:

$$R_i = \lambda_i \times R_i \times S_i + S_i \Rightarrow$$

$$R_i = \frac{S_i}{1 - \lambda_i \times S_i} = \frac{S_i}{1 - \lambda_0 \times V_i \times S_i} = \frac{S_i}{1 - \lambda_0 \times D_i}$$

↑
si no saturado

66

Resolución de redes de colas abiertas

- Modelos abiertos (suponemos conocidos: λ_0, V_i y $S_i \forall i=1..K$).



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:

$$R_i = \frac{S_i}{1 - \lambda_i \times S_i} = \frac{S_i}{1 - \lambda_0 \times D_i}$$

Paso 3.- Calculamos el tiempo medio de respuesta del servidor:

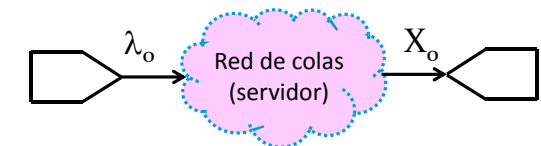
$$R_0 = \sum_{i=1}^K V_i \times R_i = \sum_{i=1}^K \frac{V_i \times S_i}{1 - \lambda_0 \times D_i} = \sum_{i=1}^K \frac{D_i}{1 - \lambda_0 \times D_i}$$

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

67

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:

- Calcule las demandas de servicio de cada recurso.
- ¿Qué recurso es el cuello de botella? ¿Cuál es la productividad máxima del servidor? ¿Está el servidor saturado?
- Calcule el tiempo de respuesta, para el peor caso posible, de cada recurso y del servidor.
- Calcule el nº medio de clientes conectados (=trabajos) en el servidor.
- Calcule la utilización, el tiempo medio de espera en la cola y el número de medio de trabajos en la cola de cada recurso.

68

Solución del ejemplo

a)

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

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 $\lambda_0 = 5 \text{ tr/s} < X_0^{max}$ el servidor no está saturado.

c) Peor caso posible: $R_i = N_i \times S_i + S_i \Rightarrow R_i = \frac{S_i}{1 - \lambda_0 \times D_i}$

$$R_{CPU} = \frac{S_{CPU}}{1 - \lambda_0 \times D_{CPU}} = \frac{0,01 \text{ s}}{1 - 5 \text{ tr/s} \times 0,09 \text{ s}} = 0,018 \text{ s}$$

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

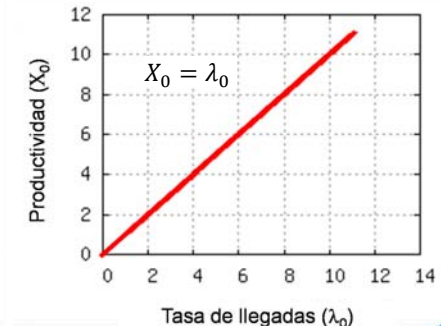
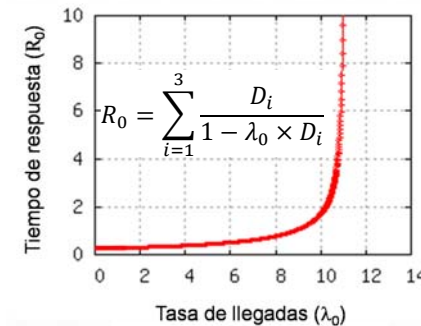
69

Solución del ejemplo (cont.)

e) $U_i = X_0 \times D_i = 5 \text{ tr/s} \times D_i$; $W_i = R_i - S_i$
 $Q_i = \lambda_i \times W_i = X_i \times W_i = X_0 \times V_i \times W_i = 5 \text{ tr/s} \times V_i \times W_i$

Recurso	V_i	S_i (s)	D_i (s)	R_i (s)	U_i	W_i (s)	Q_i (tr.)
CPU	9	0,010	0,09	0,018	0,45	0,008	0,37
DISCO	3	0,020	0,06	0,029	0,30	0,009	0,13
RED	5	0,016	0,08	0,027	0,40	0,011	0,27

Adicionalmente, podríamos representar R_0 y X_0 en función de λ_0 :



70

Resolución con solvenet

- Programa muy sencillo que resuelve redes de colas utilizando los algoritmos de esta sección.
- Disponible el código fuente en lenguaje C (SWAD).
- 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
Si:      service time of device i
Vi:      ratio visit of device i
```

71

Resolución con solvenet: redes abiertas

Recurso	V_i	S_i (s)	D_i (s)
CPU	9	0.010	0.09
DISCO	3	0.020	0.06
RED	5	0.016	0.08

λ_0 5.0 trabajos/s

`solvenet 0 5.0 3 0.01 9 0.02 3 0.016 5`



NAME	U_i	N_i	R_i	X_i	W_i
DEV 1	0.4500*	0.8182*	0.0182*	45.0000*	0.0082*
DEV 2	0.3000*	0.4286*	0.0286*	15.0000*	0.0086*
DEV 3	0.4000*	0.6667*	0.0267*	25.0000*	0.0107*

72

Resolución con solvenet: redes abiertas (II)

Recurso	V_i	S_i (s)	D_i (s)
CPU	9	0.010	0.09
DISCO	3	0.020	0.06
RED	5	0.016	0.08

λ_o	5.0 trabajos/s
-------------	----------------

solvenet 0 5.0 3 0.01 9 0.02 3 0.016 5



```

*****
*   NAME   *   Vi   *   Si   *   Di   *   Qi   *
*****
*         *         *         *         *
* DEV 1   *   9.0000*   0.0100*   0.0900*   0.3682*
*         *         *         *         *
* DEV 2   *   3.0000*   0.0200*   0.0600*   0.1286*
*         *         *         *         *
* DEV 3   *   5.0000*   0.0160*   0.0800*   0.2667*
*         *         *         *         *
*****
    
```

73

Resolución con solvenet: redes abiertas (III)

Recurso	V_i	S_i (s)	D_i (s)
CPU	9	0.010	0.09
DISCO	3	0.020	0.06
RED	5	0.016	0.08

λ_o	5.0 trabajos/s
-------------	----------------

solvenet 0 5.0 3 0.01 9 0.02 3 0.016 5



```

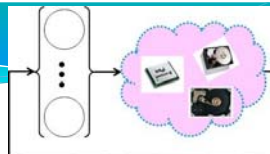
*****
*   SYSTEM VARIABLES   *
*****
* #JOBS IN SYSTEM (N0) *   1.9134*
*         *         *
* RESPONSE TIME (R0)   *   0.3827*
* MINIMUM RESPONSE TIME *   0.2300*
*         *         *
* THROUGHPUT (X0)      *   5.0000*
* MAXIMUM THROUGHPUT   *   11.111*
*         *         *
*****
    
```

```

*****
*   ASIMPTOTIC BOUNDS   *
*****
* R0_min =   0.2300      *
* X0_max =   11.1111     *
*         *         *
*****
    
```

74

Resolución de redes cerradas



- Modelos cerrados (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_T trabajos. Ídem para los tiempos de respuesta $R_i(n_T)$ y las productividades $X_i(n_T)$.

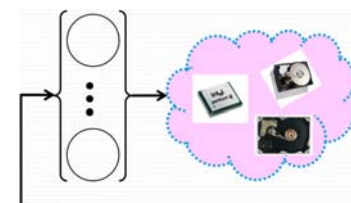
```

For i = 1 to K do  $N_i(0) = 0$  ← Inicialización del nº de
                                trabajos en cada estación
For  $n_T = 1$  to  $N_T$  do
    For i = 1 to K do  $R_i(n_T) = (N_i(n_T - 1) + 1) \times S_i$  ← Hipótesis del peor caso
                                                                para redes cerradas
     $R_0(n_T) = \sum_{i=1}^K V_i \times R_i(n_T)$ ,  $X_0(n_T) = \frac{n_T}{Z + R_0(n_T)}$  ← Tiempo de respuesta
                                                                y productividad del
                                                                servidor.
    For i = 1 to K do  $N_i(n_T) = X_0(n_T) \times V_i \times R_i(n_T)$ 
                                                                Actualización del número de
                                                                trabajos en cada estación.
    
```

75

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



$$N_{CPU}(0) = N_{DISCO1}(0) = N_{DISCO2}(0) = 0$$

$$n_T = 1$$

$$R_{CPU}(1) = S_{CPU} = 0.01s$$

$$R_{DISCO1}(1) = S_{DISCO1} = 0.02s$$

$$R_{DISCO2}(1) = S_{DISCO2} = 0.03s$$

$$R_0(1) = 10 \times 0.01 + 5 \times 0.02 + 4 \times 0.03 = 0.32s$$

$$X_0(1) = \frac{1}{2 + 0.32} = 0.43 \text{ trabajos/s}$$

$$N_{CPU}(1) = 0.43 \times 10 \times 0.01 = 0.043$$

$$N_{DISCO1}(1) = 0.43 \times 5 \times 0.02 = 0.043$$

$$N_{DISCO2}(1) = 0.43 \times 4 \times 0.03 = 0.052$$

76

Ejemplo: resolución de redes cerradas (II)

$$N_{CPU}(1) = 0.043; N_{DISCO1}(1) = 0.043$$

$$N_{DISCO2}(1) = 0.052$$

$$n_T = 2$$

$$R_{CPU}(2) = (0.043 + 1) \times S_{CPU} = 0.01043s$$

$$R_{DISCO1}(2) = (0.043 + 1) \times S_{DISCO1} = 0.0209s$$

$$R_{DISCO2}(2) = (0.052 + 1) \times S_{DISCO2} = 0.0316s$$

$$R_0(2) = \sum_{i=1}^3 V_i \times R_i(2) = 0.3348s$$

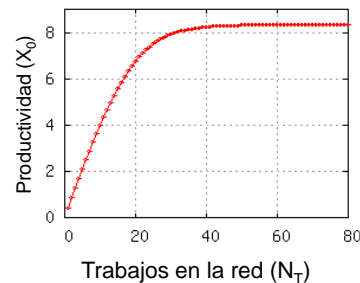
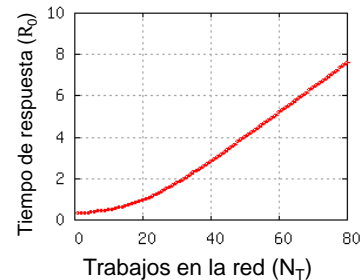
$$X_0(2) = \frac{2}{2 + 0.3348} = 0.857 \text{ trabajos/s}$$

$$N_{CPU}(2) = 0.857 \times 10 \times 0.01043 = 0.0894$$

$$N_{DISCO1}(2) = 0.857 \times 5 \times 0.0209 = 0.0894$$

$$N_{DISCO2}(2) = 0.857 \times 4 \times 0.0316 = 0.1081$$

etc. hasta llegar al valor $n_i = N_T$ que nos pidan



77

Resolución con solvenet: redes cerradas

Recurso	V_i	S_i (s)
CPU	10	0.01
DISCO1	5	0.02
DISCO2	4	0.03

T. reflexión (Z)	2 s
Nº Trabajos en la red cerrada (N_T)	2

solvenet 1 2 2 3 0.01 10 0.02 5 0.03 4



NAME	U_i	N_i	R_i	X_i	W_i
DEV 1	0.0857*	0.0894*	0.0104*	8.5659*	0.0004*
DEV 2	0.0857*	0.0894*	0.0209*	4.2830*	0.0009*
DEV 3	0.1028*	0.1081*	0.0316*	3.4264*	0.0016*

78

Resolución con solvenet: redes cerradas (II)

Recurso	V_i	S_i (s)
CPU	10	0.01
DISCO1	5	0.02
DISCO2	4	0.03

T. reflexión (Z)	2 s
Nº Trabajos en la red cerrada (N_T)	2

solvenet 1 2 2 3 0.01 10 0.02 5 0.03 4



NAME	V_i	S_i	D_i	Q_i
DEV 1	10.0000*	0.0100*	0.1000*	0.0037*
DEV 2	5.0000*	0.0200*	0.1000*	0.0037*
DEV 3	4.0000*	0.0300*	0.1200*	0.0053*

79

Resolución con solvenet: redes cerradas (III)

Recurso	V_i	S_i (s)
CPU	10	0.01
DISCO1	5	0.02
DISCO2	4	0.03

T. reflexión (Z)	2 s
Nº Trabajos en la red cerrada (N_T)	2

solvenet 1 2 2 3 0.01 10 0.02 5 0.03 4



SYSTEM VARIABLES	
#JOBS IN SYSTEM (N0)	0.2868*
#INTERACTIVE USERS (NZ)	1.7132*
#JOBS IN THE NET (NT)	2*
SATURATION POINT (N*)	20*
RESPONSE TIME (R0)	0.3348*
MINIMUM RESPONSE TIME	0.3200*
THROUGHPUT (X0)	0.8566*
MAXIMUM THROUGHPUT	8.3333*

ASIMPTOTIC BOUNDS	
$R_0 \geq \max\{0.32, 0.12 \cdot NT - 2.00\}$	
$X_0 \leq \min\{NT/2.32, 8.33\}$	

80