APLICACIONES DEL ANÁLISIS OPERACIONAL

Administradores y diseñadores

¿Cómo conocer la capacidad residual de un sistema informático?

CONTENIDO

1. Introducción
2. Algoritmos de resolución de redes de colas
   - Tiempo de respuesta de una estación
   - Análisis de redes abiertas
   - Análisis de redes cerradas
3. Límites optimistas del rendimiento
   - Concepto de cuello de botella
   - Ecuaciones para el tiempo de respuesta y la productividad
4. Técnicas de mejora
   - Actualización
   - Sintonización

1. INTRODUCCIÓN

Limitaciones del análisis operacional

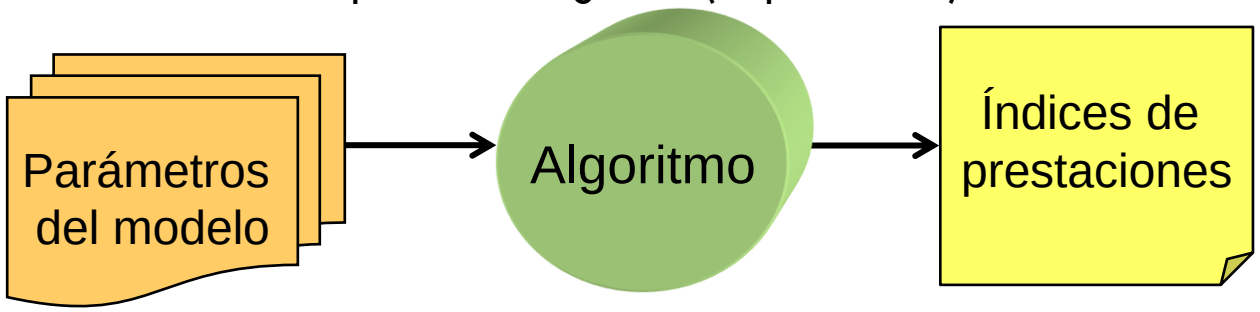
ANALISIS OPERACIONAL: LIMITACIONES

Aplicando únicamente análisis operacional, no se puede aplicar ninguna distribución estadística, particularmente para calcular:

- Número medio de trabajos en una estación
- Tiempo medio de respuesta en una estación

Los algoritmos de resolución obtienen estos resultados en base a unas hipótesis de partida sobre:

- La distribución del tiempo de servicio (exponencial)
- La distribución del tiempo entre llegadas (exponencial)



HIPÓTESIS DE PARTIDA

La distribución estadística de los tiempos de los clientes:

- Si un trabajo está siendo atendido en una estación, el tiempo que le queda antes de abandonar la misma es independiente del tiempo que ya lleva en servicio
- En un sistema abierto el tiempo que queda hasta la próxima llegada es independiente del instante en que se produjo la llegada anterior

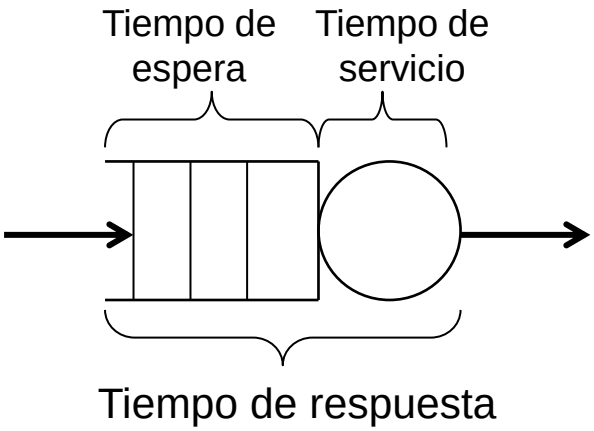
Estas hipótesis se cumplen cuando las distribuciones del tiempo entre llegadas y el tiempo de servicio son exponenciales (*memoryless property*)

TIEMPO DE RESPUESTA DE UNA ESTACIÓN

En la estación hay N_i trabajos y el tiempo de servicio es exponencial de media S_i

Cuando llega un trabajo a la estación:

- Espera a que se procesen los N_i trabajos
- Luego, cumple su propio tiempo de servicio



$$R_i = (N_i + 1)\, S_i = (X_i R_i + 1)\, S_i \Rightarrow$$

$$R_i = \frac{S_i}{1 - X_i S_i} = \frac{S_i}{1 - U_i}$$

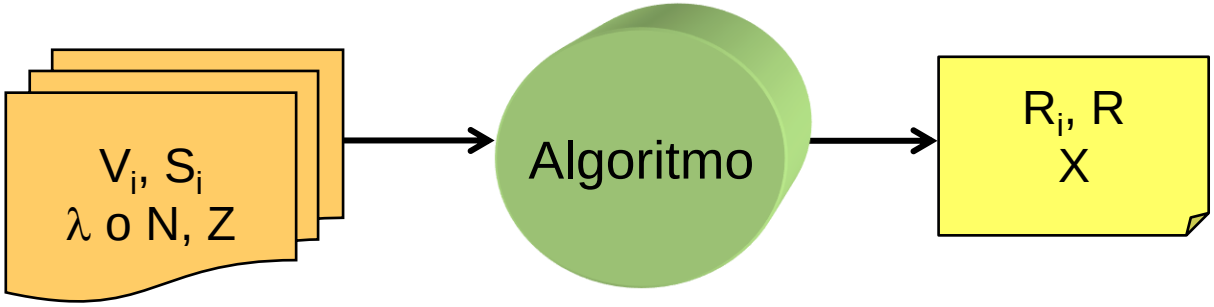
$$R_i = (N_i + 1)\, S_i$$

No es una relación operacional

ESQUEMA DE RESOLUCIÓN

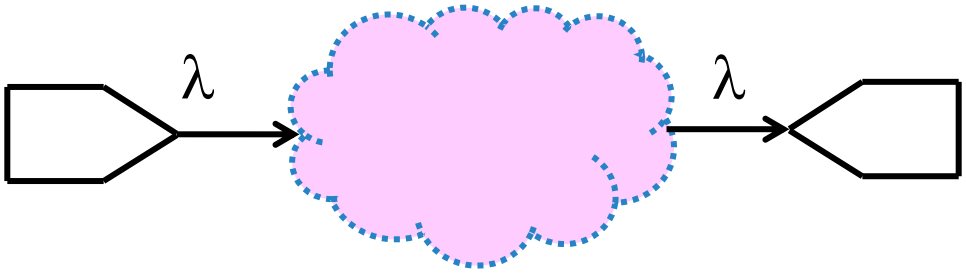
Datos del modelo con K estaciones

- Si la red es abierta
  - Tasa de llegadas al sistema (λ)
- Si la red es cerrada
  - Número de trabajos en el sistema (N)
  - Tiempo de reflexión de los usuarios (Z)
- Por cada estación
  - Razón de visita de cada estación (V_i)
  - Tiempo de servicio de cada estación (S_i)



RESOLUCIÓN DE REDES ABIERTAS

Sistemas abiertos (suponemos conocidos: λ, V_i y S_i)

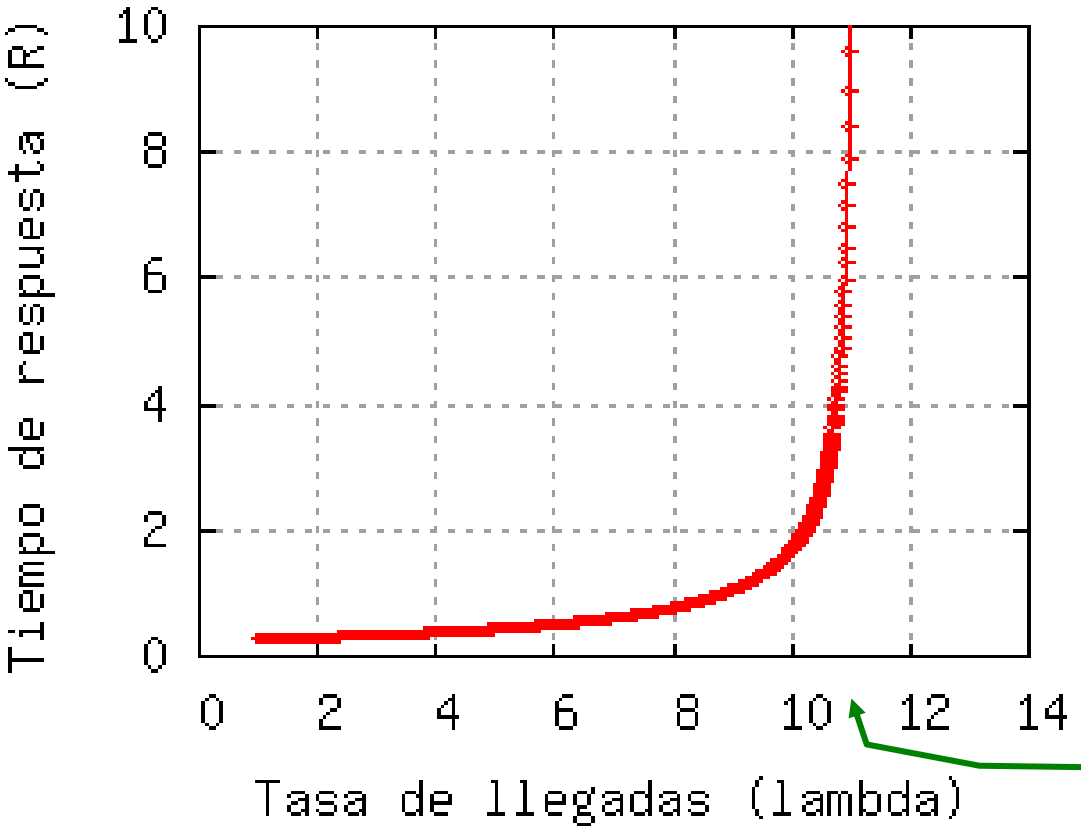


$$U_i = X_i S_i = \lambda V_i S_i$$ ← Utilización de cada estación

$$R_i = \frac{S_i}{1 - U_i}$$ ← Tiempo de respuesta de cada estación

$$R = \sum_{i=1}^{k} V_i R_i = \sum_{i=1}^{k} \left(\frac{V_i S_i}{1 - U_i} \right)$$ ← Tiempo de respuesta del sistema

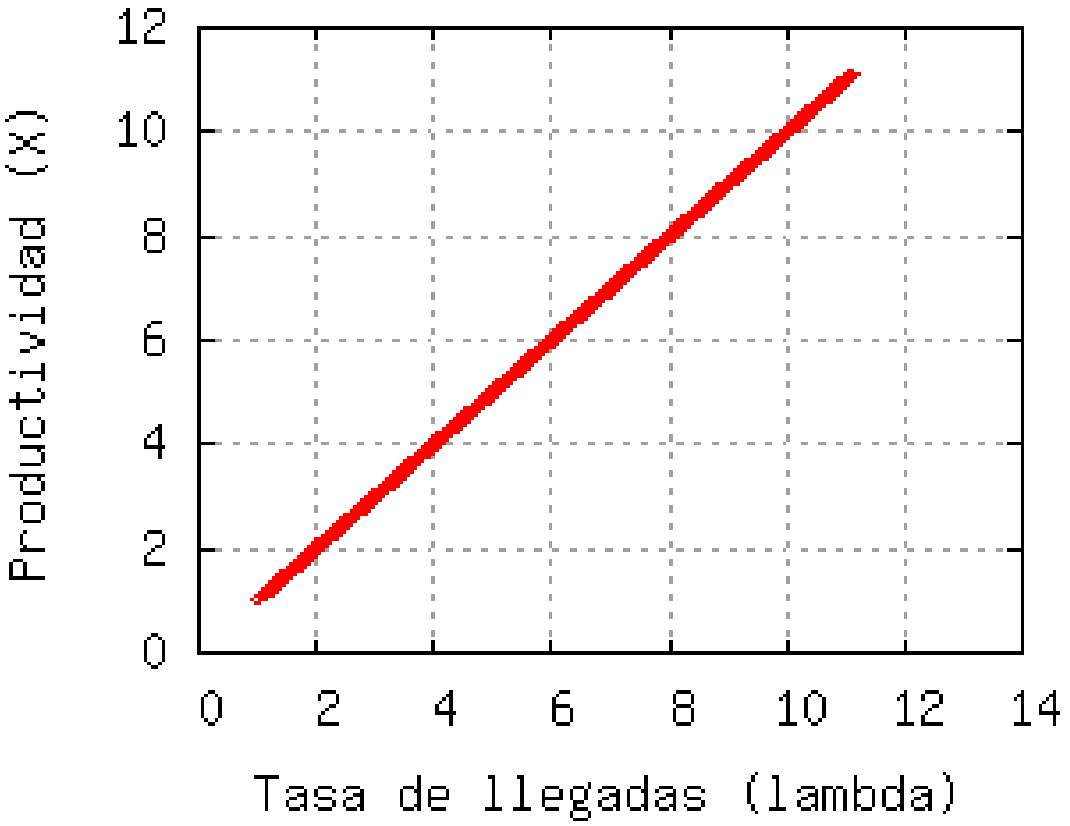
TIEMPO DE RESPUESTA EN REDES ABIERTAS



Dispositivo	V_i	S_i
CPU	9	0.01
DISCO	4	0.02
DISCO	4	0.02

Hay un valor máximo (capacidad del sistema)

PRODUCTIVIDAD EN REDES ABIERTAS



Dispositivo	V_i	S_i
CPU	9	0.01
DISCO	4	0.02
DISCO	4	0.02

No aporta ninguna información, ya que $X=\lambda$

EJEMPLO (RED ABIERTA, λ=2 TRABAJOS/S)

Dispositivo	Razón de visita	Tiempo de servicio (s)
1	6	0,01
2	7	0,02

En primer lugar se pueden calcular las utilizaciones:

$$\begin{aligned} U_1 &= \lambda \times D_1 = \lambda \times V_1 \times S_1 = 2 \times 6 \times 0{,}01 = 0{,}12 \\ U_2 &= \lambda \times D_2 = \lambda \times V_2 \times S_2 = 2 \times 7 \times 0{,}02 = 0{,}28 \end{aligned}$$

En este momento ya se pueden calcular los tiempos de respuesta de cada estación utilizando la ecuación 5.2:

$$\begin{aligned} R_1 &= \frac{S_1}{1 - U_1} = \frac{0{,}01}{1 - 0{,}12} = 0{,}0114 \text{ s} \\ R_2 &= \frac{S_2}{1 - U_2} = \frac{0{,}02}{1 - 0{,}28} = 0{,}0278 \text{ s} \end{aligned}$$

Finalmente, el tiempo de respuesta del sistema y el número de trabajos contenidos en él se calculan utilizando las relaciones:

$$\begin{aligned} R &= V_1 \times R_1 + V_2 \times R_2 = 6 \times 0{,}0114 + 7 \times 0{,}0278 = 0{,}263 \text{ s} \\ N &= \lambda \times R = 2 \times 0{,}263 = 0{,}526 \text{ trabajos} \end{aligned}$$

RESOLUCIÓN DE REDES CERRADAS

Sistemas cerrados (suponemos conocidos: N, Z y V_i ,S_i)

- Algoritmo del valor medio (*MVA, mean value analysis*)
- Iterativo con $n = 1,2,...,N$
- La productividad X(n) no se conoce, se debe de estimar

$$\text{For } i = 1 \text{ to } K \text{ do } N_i(0) = 0$$

$$\text{For } n = 1 \text{ to } N \text{ do}$$

$$R_i(n) = (N_i(n-1)+1)\, S_i$$

$$R(n) = \sum_{i=1}^{K} V_i R_i(n), \quad X(n) = \frac{n}{Z + R(n)}$$

$$N_i(n) = X(n)\, V_i R_i(n)$$

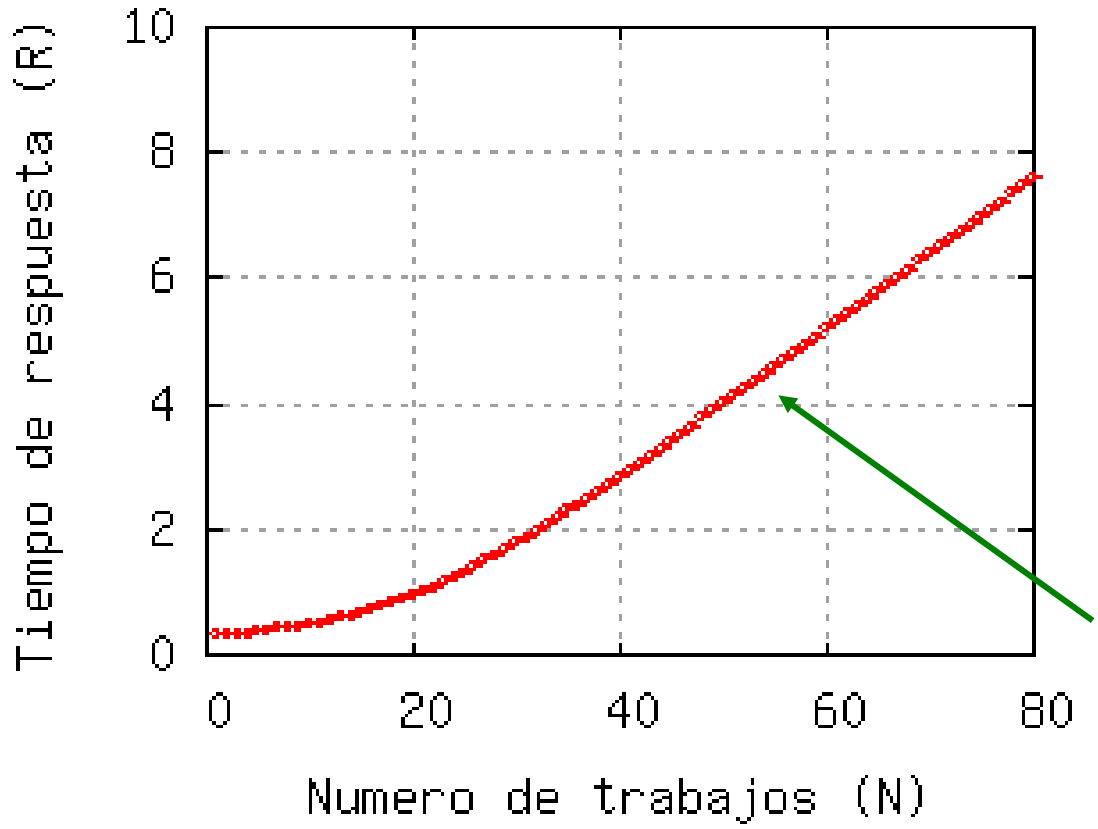
← Inicialización de trabajos en las de estaciones

← Tiempo de respuesta de cada estación

← Tiempo de respuesta y productividad del sistema

← Actualización del número de trabajos en cada estación

TIEMPO DE RESPUESTA EN REDES CERRADAS

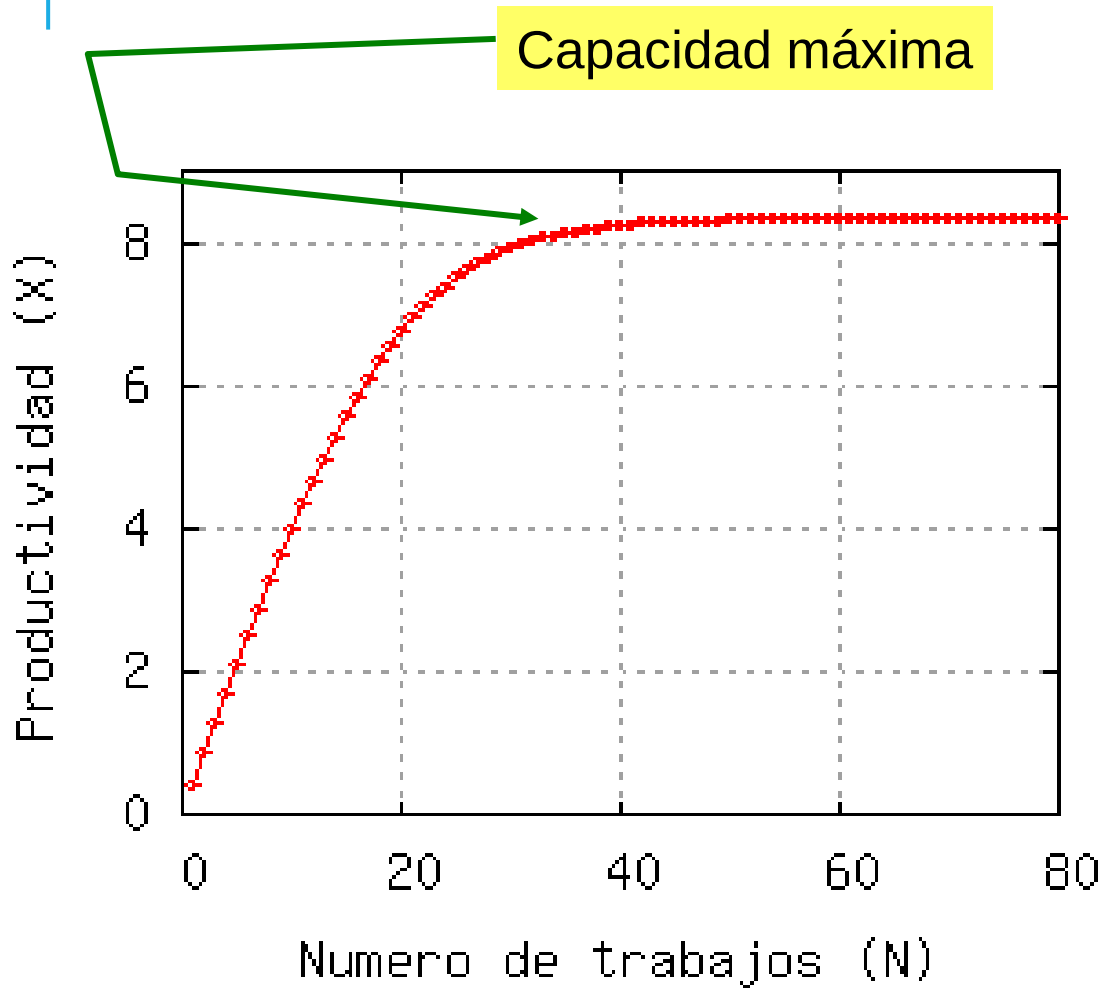


Tiempo de reflexión	2

Dispositivo	V_i	S_i
CPU	10	0.01
DISCO	5	0.02
DISCO	4	0.03

El incremento es lineal

PRODUCTIVIDAD EN REDES CERRADAS



Tiempo de reflexión	2

Dispositivo	V_i	S_i
CPU	10	0.01
DISCO	5	0.02
DISCO	4	0.03

EJEMPLO (RED CERRADA, Z = 5s)

Para n desde 1 hasta N hacer:

$$R_i(n) = (N_i(n-1)+1) \times S_i, \text{ con } N_i(0) = 0$$

$$R(n) = \sum_{i=1}^{K} V_i \times R_i(n), \quad X(n) = \frac{n}{Z + R(n)}$$

$$N_i(n) = X(n) \times V_i \times R_i(n)$$

$$X_i(n) = X(n) \times V_i$$

$$U_i(n) = X(n) \times V_i \times S_i$$

Como ejemplo sencillo de aplicación de este algoritmo, supongamos una red de colas cerrada con tres trabajos y dos dispositivos, 1 y 2, que tienen los tiempos de servicio y razones de visita expresados en la siguiente tabla:

Dispositivo	Razón de visita	Tiempo de servicio (s)
1	15	0,03
2	14	0,5

EJEMPLO (REDES CERRADAS)

Supondremos que la carga es interactiva con un tiempo de reflexión $Z = 5$ segundos. Para aplicar el algoritmo habrá que hacer un total de 3 iteraciones, una por cada trabajo presente en el sistema. Para cada iteración calcularemos, en primer lugar, el tiempo de respuesta de cada estación; a continuación se calculará el tiempo de respuesta del sistema y su productividad, y finalmente se podrán determinar el número de trabajos, la productividad y la utilización de cada estación del modelo. La Tabla 5.1 muestra los datos obtenidos en cada iteración del algoritmo (no se muestra ni las utilizaciones ni la productividades individuales):

Trabajos	R_1	R_2	R	X_0	N_1	N_2
1	0,0300	0,5000	7,4500	0,0803	0,0361	0,5622
2	0,0311	0,7811	11,4920	0,1219	0,0569	1,3335
3	0,0317	1,1667	16,8090	0,1376	0,0654	2,2468

Tabla 5.1: Ejemplo de aplicación del algoritmo del valor medio

El tiempo de respuesta del sistema, a partir de los datos presentados en la tabla, es de 16,8090 segundos, mientras que la productividad es de 0,1376 trabajos por segundo.

3. LÍMITES OPTIMISTAS DEL RENDIMIENTO

Concepto de cuello de botella

Ecuaciones para R y X

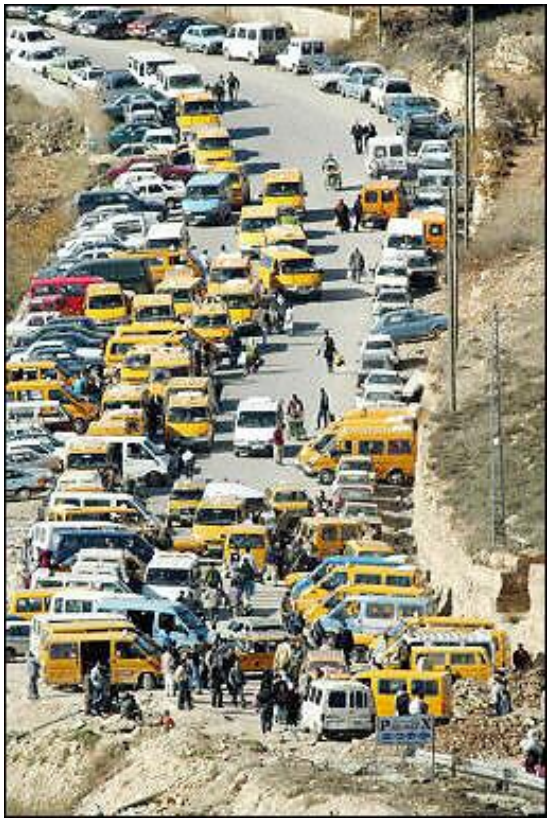
LIMITACIONES EN EL RENDIMIENTO

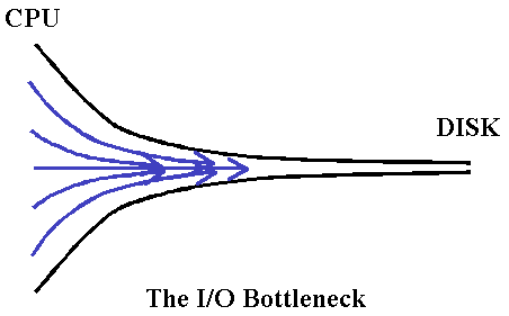
Todo sistema presenta alguna limitación en su rendimiento

- Causas: hardware, software, organización del sistema, etc.

La localización del elemento limitador depende del sistema y de la carga

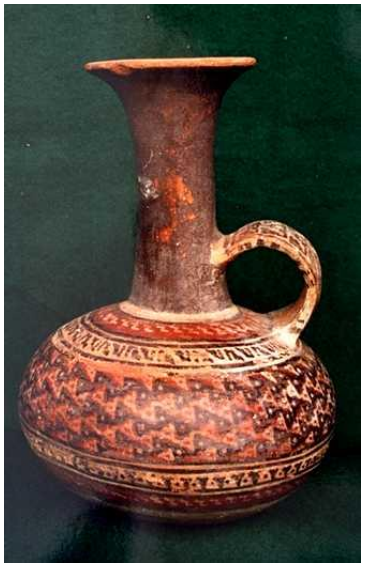
- Puede haber uno o varios de estos elementos limitadores





The I/O Bottleneck

¿QUÉ ES UN CUELLO DE BOTELLA?



Un cuello de botella (*bottleneck*) es un elemento limitador del rendimiento del sistema

Las prestaciones globales del sistema dependen del dispositivo cuello de botella

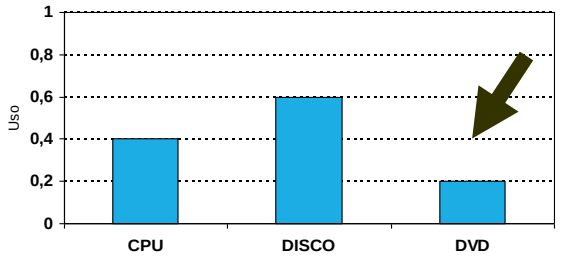
- La única manera de mejorar las prestaciones de manera significativa es actuando sobre este dispositivo
- El cuello de botella es el dispositivo con la mayor demanda de servicio (o utilización más alta)

$$U_i = X_i S_i = X V_i S_i = X D_i \Rightarrow U_i \propto D_i$$

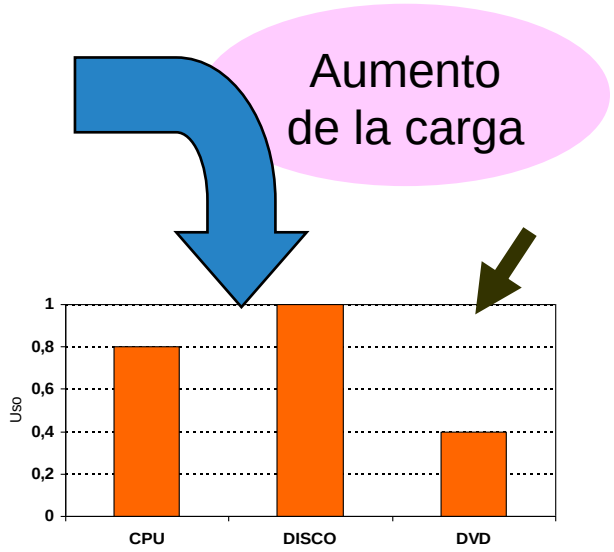
SATURACIÓN DEL SISTEMA

El sistema se satura cuando lo hace el cuello de botella

- Es el primer dispositivo en saturarse



Operación “normal”



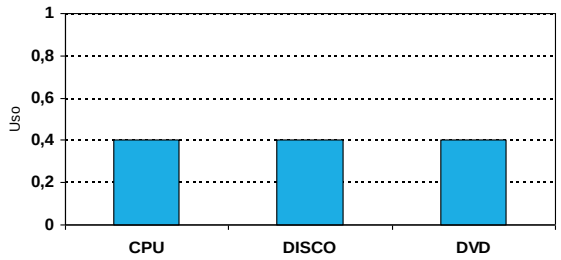
Sistema saturado

El sistema alcanza
su productividad máxima

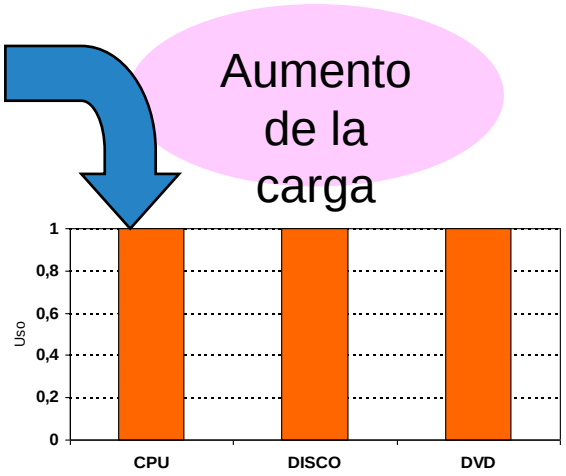
SISTEMA EQUILIBRADO (*BALANCED SYSTEM*)

Sistema en que todos los dispositivos tienen la misma demanda o utilización (la carga se absorbe equitativamente)

$$U_i = U_j, \forall i, j \Rightarrow D_i = D_j, \forall i, j$$



Todos los dispositivos son cuellos de botella



LÍMITES OPTIMISTAS DEL RENDIMIENTO

Cota superior de la productividad (X) e inferior del tiempo de respuesta (R)

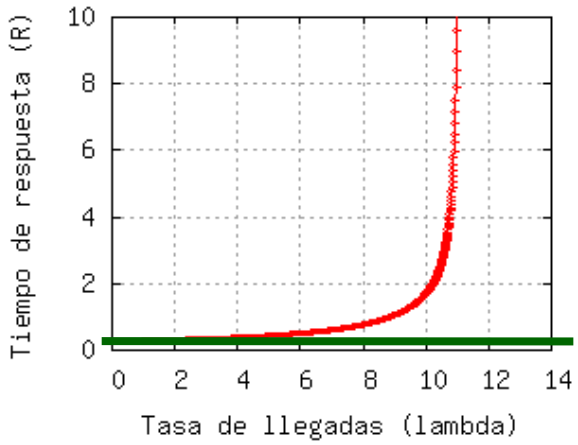
- ¿Cuál es la productividad máxima?
- ¿Cuál es el tiempo de respuesta mínimo?

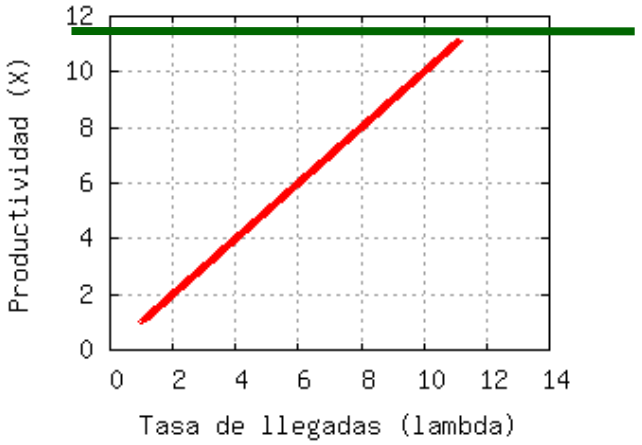
Campos de aplicación

- Estudios preliminares: consideración de un gran número de configuraciones candidatas
- Estimación de la mejora potencial de prestaciones que pueden reportar acciones sobre el sistema
- Planificación de la capacidad (*capacity planning*)

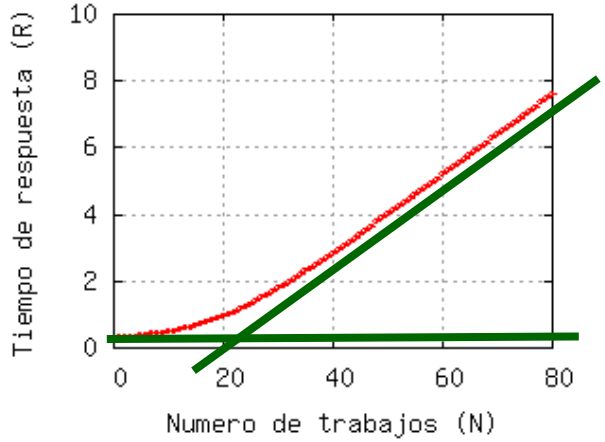
LOCALIZACIÓN DE ASÍNTOTAS

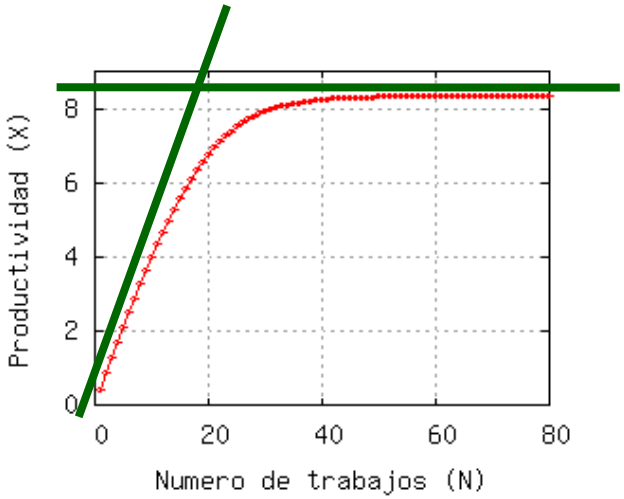
Modelos abiertos





Modelos cerrados





DATOS DE PARTIDA

Localización del cuello de botella (dispositivo b)

- D_b: demanda de servicio del dispositivo cuello de botella (la máxima del sistema)

$$D_b = \max_{i=1,2,...,K} \{D_i\} = \max_{i=1,2,...,K} \{V_i S_i\} = V_b S_b$$

$$U_b = \max_{i=1,2,...,K} \{U_i\} = \max_{i=1,2,...,K} \{X_i S_i\} = X_b S_b = X V_b S_b = X D_b$$

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

D: Demanda total de servicio

Z: Tiempo de reflexión (sistemas interactivos)

LÍMITES OPTIMISTAS: SISTEMAS ABIERTOS

El valor máximo de la tasa de llegada λ será aquél que sature completamente el dispositivo cuello de botella

Valor optimista de la productividad

$$U_b = X_b S_b = X D_b = \lambda D_b$$

$$\text{Si } U_b = 1 \Rightarrow \lambda D_b = 1 \Rightarrow \lambda = \frac{1}{D_b}$$

$$X_{opt} = \frac{1}{D_b}$$

LÍMITES OPTIMISTAS: SISTEMAS ABIERTOS

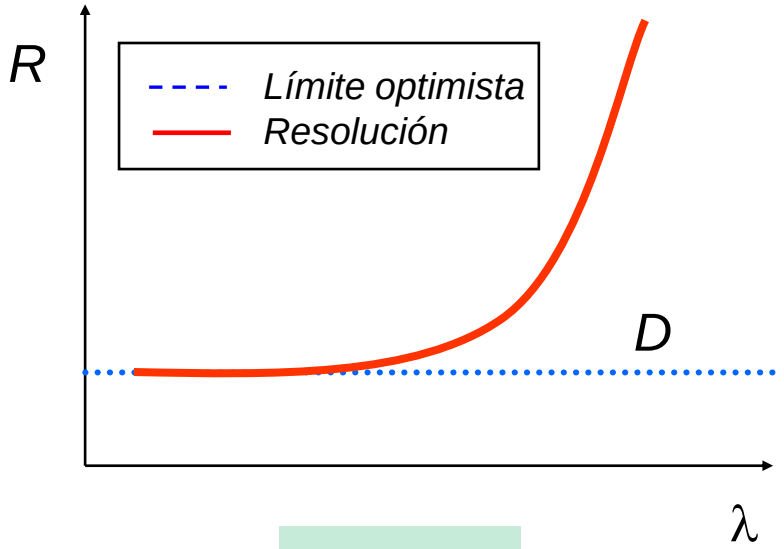
Valor optimista del tiempo de respuesta

- Cuando $\lambda = 1/D_b$ el número de trabajos en el sistema crece indefinidamente
- El valor mínimo del tiempo de respuesta será el que experimente un trabajo a solas en el sistema

$$R_{opt} = \sum_{i=1}^{K} D_i = D$$

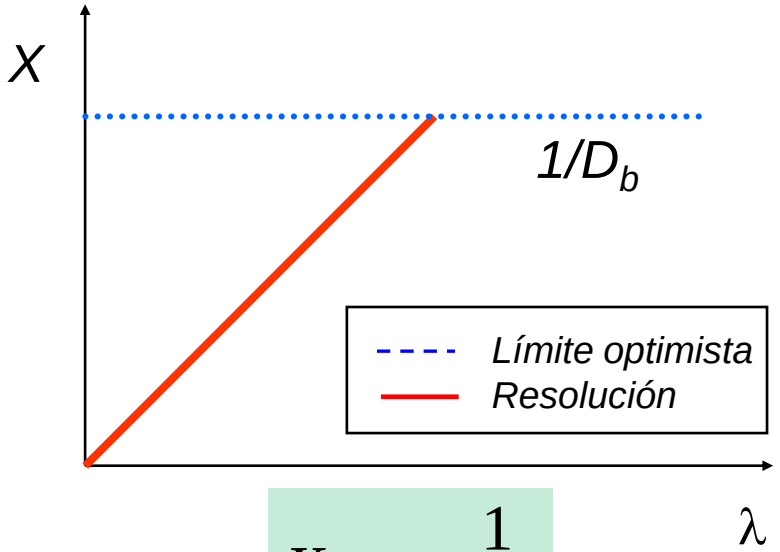
RESUMEN SISTEMAS ABIERTOS

Tiempo de respuesta



$$R_{opt} = D$$

Productividad



$$X_{opt} = \frac{1}{D_b}$$

EJEMPLO DE SISTEMA ABIERTO

Dispositivo	V_i	S_i (ms)	D_i (ms)
CPU	16	10	160
DISCO A	7	20	140
DISCO B	8	30	**240**

Tiempo de respuesta mínimo

- $D = 160+140+240 = 540$ ms

Productividad máxima

- $\lambda_{máx} = 1/D_b = 1/240 = 0.0042$ trabajos/ms $= 4.2$ trabajos/s

Utilización máxima del disco A

- $U_A = \lambda_{máx} V_A S_A = 0.5833$
- $U_B = 1$

U_i con $\lambda = 0.002$ trabajos/ms

- $U_{CPU} = \lambda D_{CPU} = 0.32$
- $U_A = \lambda D_A = 0.28$
- $U_B = \lambda D_B = 0.48$

LÍMITES OPTIMISTAS: SISTEMAS CERRADOS

Sistema sin dispositivos saturados

- Valor optimista del tiempo de respuesta
  - Los trabajos siempre encuentran los dispositivos sin ocupar

$$R_{opt} = \sum_{i=1}^{k} D_i = D$$

- Valor optimista de la productividad
  - Se puede obtener a partir del valor optimista del tiempo de respuesta

$$R_{opt} = \left(\frac{N}{X_{opt}} \right) - Z \Rightarrow X_{opt} = \frac{N}{D+Z}$$

LÍMITES OPTIMISTAS: SISTEMAS CERRADOS

Sistema con el dispositivo cuello de botella saturado

- Valor optimista de la productividad

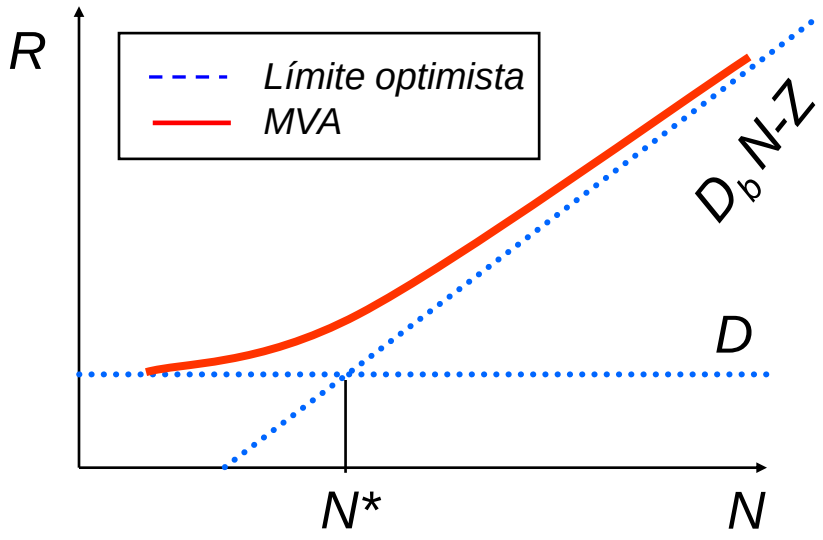
$$\text{Si } U_b = 1 \Rightarrow X_b S_b = X V_b S_b = 1 \Rightarrow X_{opt} = \frac{1}{D_b}$$

- Valor optimista del tiempo de respuesta
  - Se puede obtener a partir del valor optimista de X

$$R_{opt} = \left(\frac{N}{X_{opt}} \right) - Z \Rightarrow R_{opt} = N D_b - Z$$

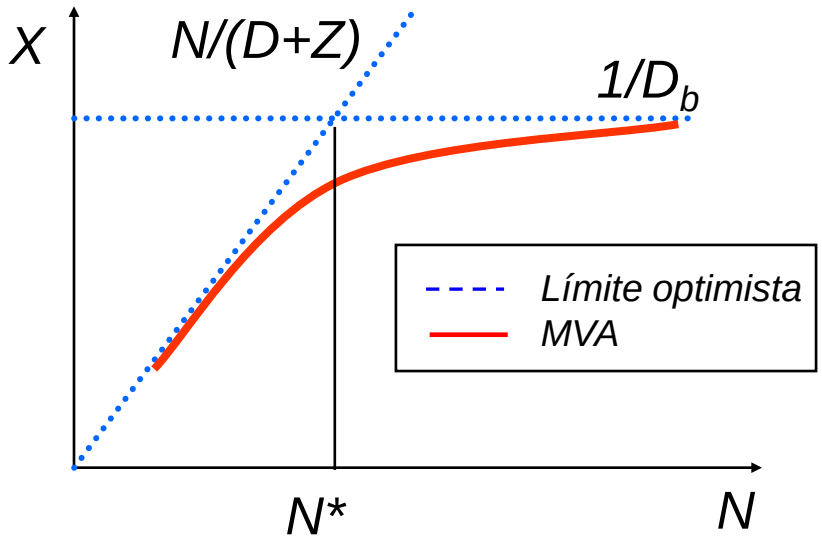
RESUMEN SISTEMAS CERRADOS

Tiempo de respuesta



$$R_{opt} = \text{máx}\{D, D_b N - Z\}$$

Productividad



$$X_{opt} = \text{mín}\left\{\frac{N}{D+Z}, \frac{1}{D_b}\right\}$$

PUNTO TEÓRICO DE SATURACIÓN

Propiedades del punto teórico de saturación N*

- Se consigue la productividad teórica máxima
- No se puede mejorar el tiempo de respuesta mínimo porque a partir de este valor se empiezan a formar colas de espera en al menos el dispositivo cuello de botella

$$D = ND_b - Z \Rightarrow N = \frac{D+Z}{D_b}$$

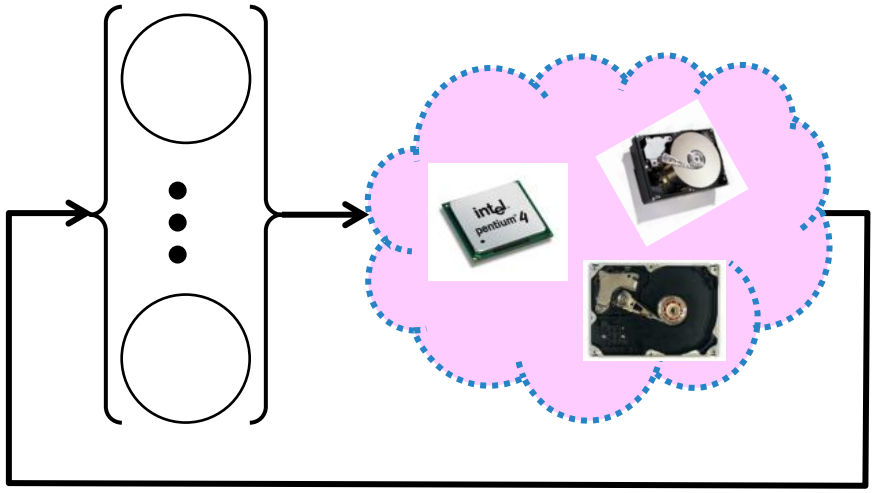
- Se suele expresar como un número entero

$$N^* = \left\lceil \frac{D+Z}{D_b} \right\rceil$$

EJEMPLO DE SISTEMA CERRADO

Tiempo de reflexión	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



Tiempo de respuesta mínimo

- $D = 5 + 4 + 3 = 12$ s

Productividad máxima

- $X_{máx} = 1/D_b = 1/5 = 0.2$ trabajos/s

Punto teórico de saturación

- $N^* = \lceil (D+Z)/ D_b \rceil = 6$ trabajos

Asíntotas

$$R_{opt} = \text{máx}\{12, 5N - 18\}$$

$$X_{opt} = \text{mín}\left\{\frac{N}{30}, 0.2\right\}$$

Máximo número de trabajos que permiten que $R_{opt} \leq 100$ s

- $5N - 18 \leq 100 \Rightarrow N \leq 23.6$

LÍMITES ASINTÓTICOS

ACTIVIDAD TEMA 5.1 (VOLUNTARIA)

Según la teoría de colas y los modelos que hemos visto para los sistemas ¿Cuál es el mínimo tiempo de respuesta de un sistema?¿y el máximo?

¿Cuál es la máxima productividad de un sistema?¿y la mínima?

¿Qué significa el punto teórico de saturación?

4. TÉCNICAS DE MEJORA

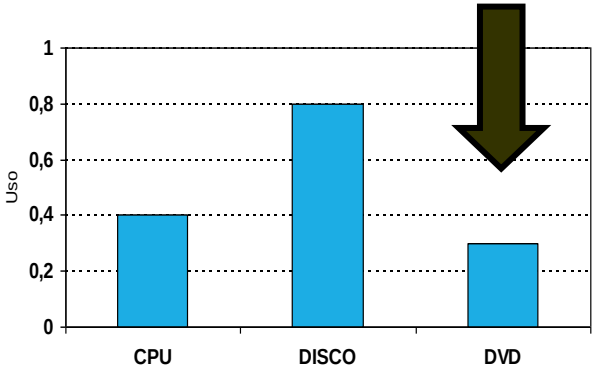
Actualización de componentes

Sintonización del sistema

TÉCNICAS PARA MEJORAR LAS PRESTACIONES

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

- Actualización y sintonización



ACTUALIZACIÓN O REPOSICIÓN (*UPGRADING*)

Reemplazar dispositivos por otros más rápidos

- Procesador, memoria, placa base

Añadir más dispositivos para poder realizar más tareas en paralelo

- Ejemplo: biprocesadores, matrices de discos (RAID)

Algunos problemas

- Compatibilidad de los nuevos elementos con los existentes
- Facilidad del sistema para dejarse actualizar

SINTONIZACIÓN O AJUSTE (*TUNING*)

Tratan de optimizar el funcionamiento de todos los componentes (hardware y software)

- Componentes hardware
- Sistema operativo
- Aplicación

Muchos ajustes se hacen en el sistema operativo

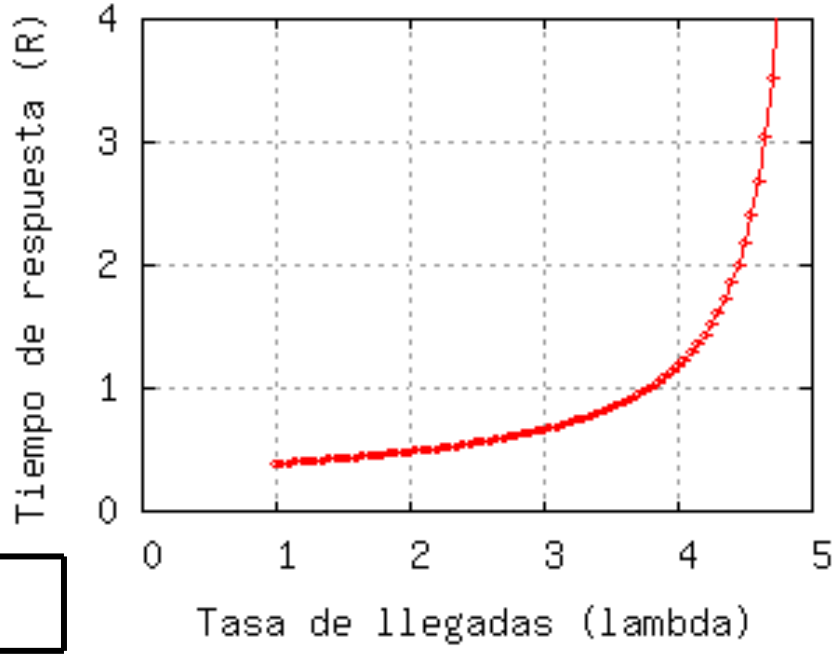
- Políticas de gestión de procesos y memoria virtual
- Distribución de la información entre discos

Algunos problemas

- Hay que conocer muy bien el sistema operativo y el funcionamiento de los componentes hardware
- Posible alteración de la fiabilidad

EJEMPLO: 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

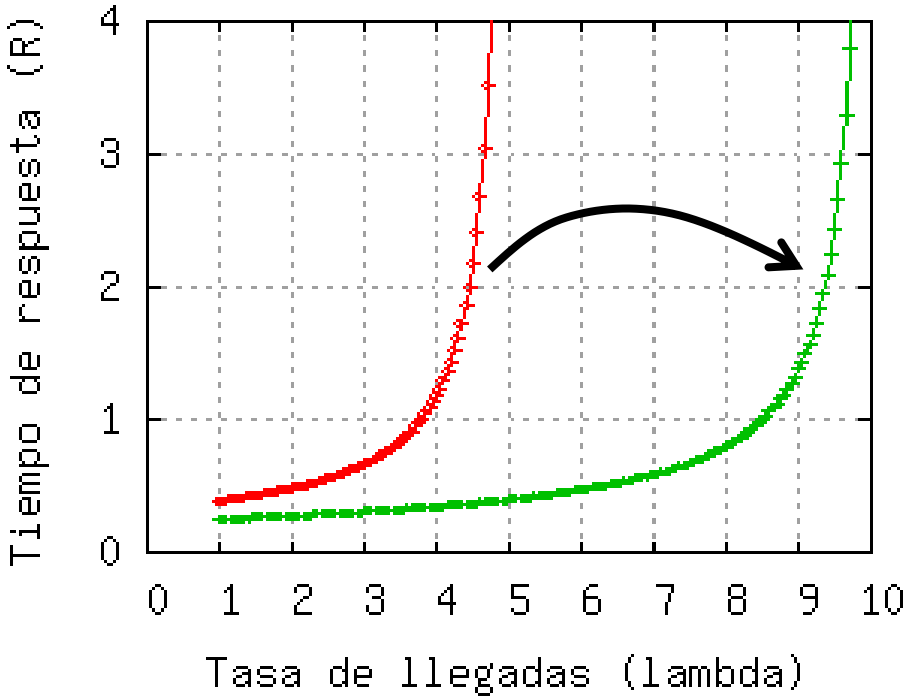


$$D = 0.33\,\text{s} \quad \lambda_{máx} = \frac{1}{D_b} = \frac{1}{0.2} = 5.0\ \text{pet/s}$$

ACTUALIZACIÓN: CPU DOBLE 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

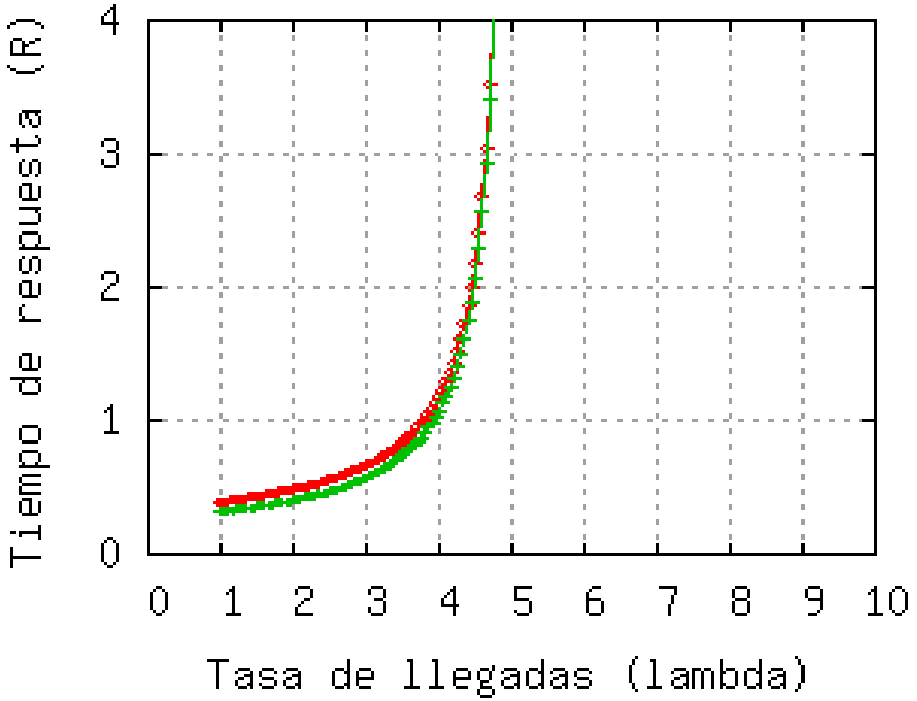


$$D = 0.23\,\text{s} \quad \lambda_{máx} = \frac{1}{D_b} = \frac{1}{0.1} = 10.0\ \text{pet/s}$$

ACTUALIZACIÓN: DISCOS DOBLE 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 pero con la misma demanda

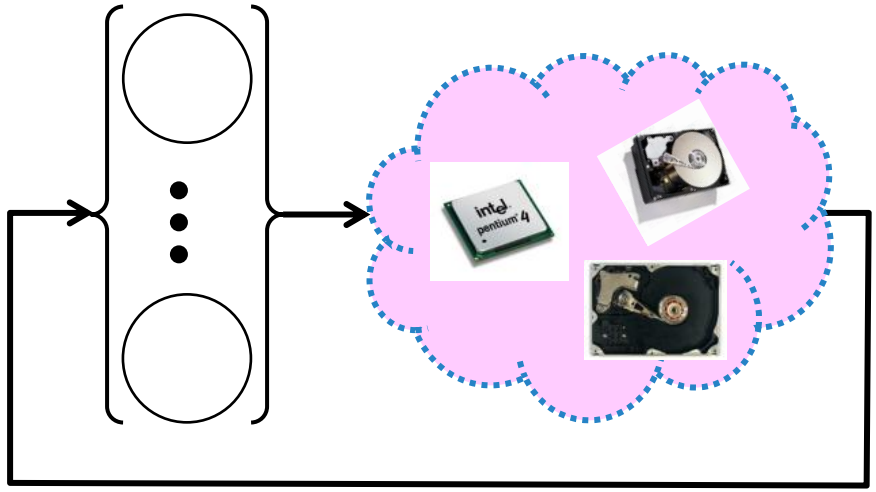


$$D = 0.265 \text{ s} \quad \lambda_{máx} = \frac{1}{D_b} = \frac{1}{0.2} = 5.0 \text{ pet/s}$$

EJEMPLO: SERVIDOR DE FICHEROS

Tiempo de reflexión	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

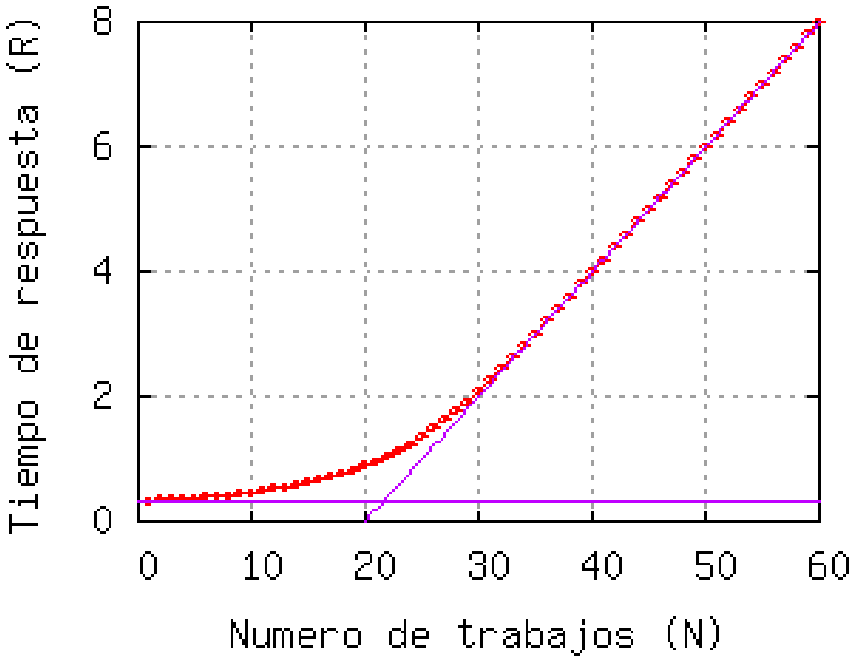


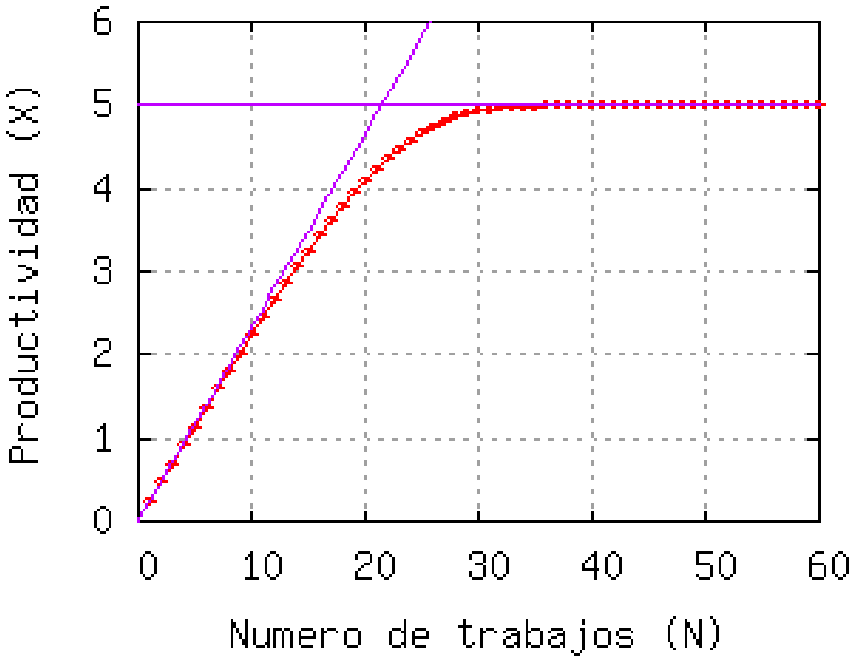
$$D = 0.33\ \text{s} \qquad X_{máx} = \frac{1}{D_b} = \frac{1}{0.2} = 5.0\ \text{pet/s}$$

$$R_{opt} = \text{máx}\{0.33, 0.2N - 4\}$$

$$X_{opt} = \text{mín}\left\{\frac{N}{4.3}, 5\right\}$$

RENDIMIENTO DEL SISTEMA ORIGINAL

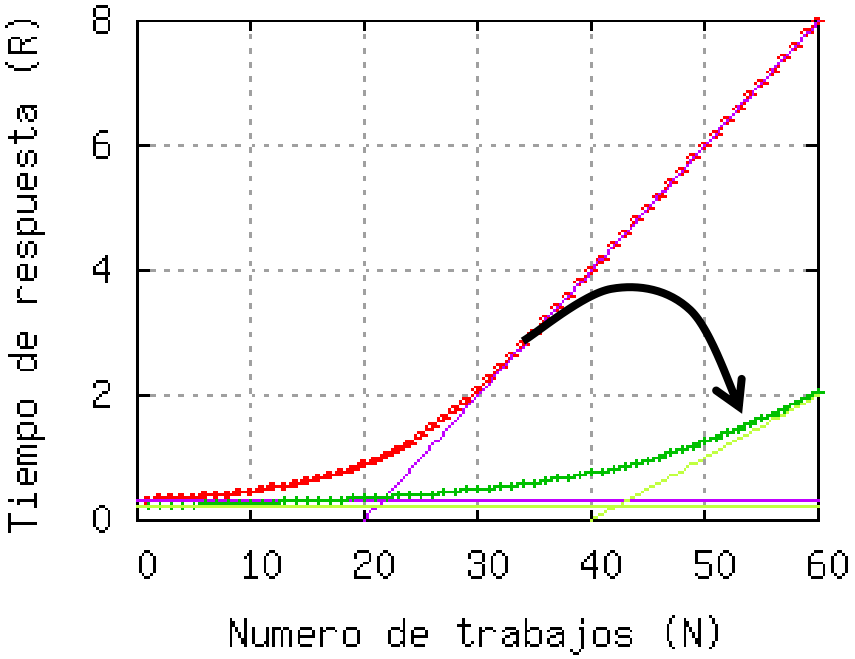


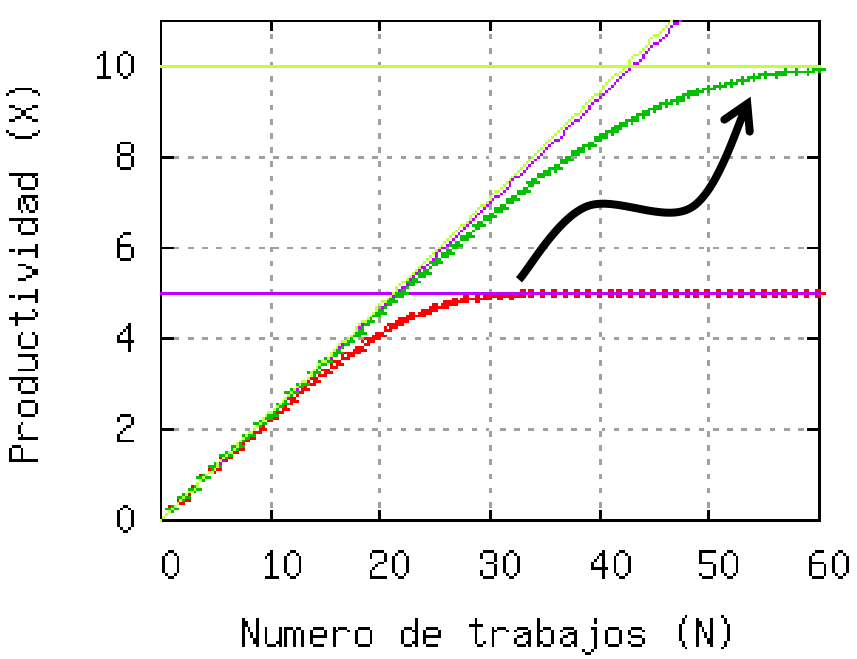


$$N^* = \left\lceil \frac{D+Z}{D_b} \right\rceil = 22$$

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

ACTUALIZACIÓN: CPU DOBLE RÁPIDA





$$N^* = \left\lceil \frac{D+Z}{D_b} \right\rceil = 43$$

$$D = 0.23\,\text{s} \qquad X_{máx} = \frac{1}{D_b} = \frac{1}{0.1} = 10.0\ \text{pet/s}$$

LÍMITES ASINTÓTICOS

ACTIVIDAD TEMA 5.2 (**OBLIGATORIA**)

¿Para un administrador de sistemas es mejor un sistema parcialmente saturado o un sistema equilibrado? ¿Por qué?

Una vez eliminado el cuello de botella de un sistema ¿ya no hay cuello de botella? ¿Por qué?