Técnicas de modelado analítico

- Introducción
- Conceptos básicos Teoría de Colas
- Análisis Operacional
- Aplicaciones del Análisis Operacional

Introducción

- La predicción cuantitativa implica la existencia de un modelo matemático del modelo observado.
- Modelo matemático: abstracción de un sistema real que pueda ser utilizado para propósitos de predicción y control.
- Debe ser capaz de analizar cuando uno o más cambios en algún aspecto del sistema modelado, puede afectar a otros aspectos del sistema o bien al sistema en su conjunto.
- Los modelos, en general, constan de los siguientes elementos:
  - Variables
    - Estado
    - Entrada (control)
    - Salida
  - Parámetros
  - Relaciones funcionales

Elementos del modelo: Variables

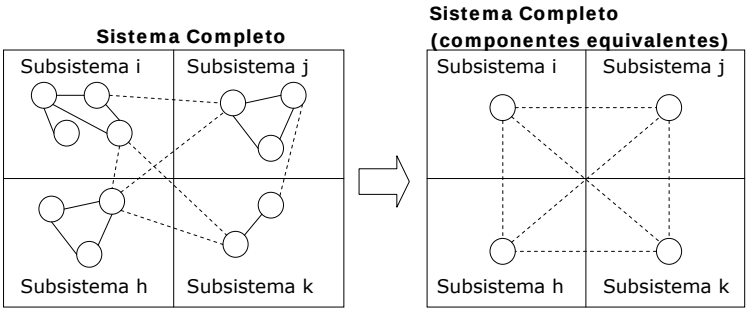
- Estado:
  - Determinan completamente, cuando son conocidas, en que situación se encuentra el sistema representado en el modelo.
  - Variables que describen el estado de un componente del sistema en el tiempo t.
- Entrada:
  - Sirven para modificar la evolución del sistema.
  - Representan los aspectos del sistema que se van a modificar cuando lo utilicemos.
  - Normalmente describen las características de la carga de trabajo y los componentes.
- Salida
  - Tienen observación directa en el sistema representado por el modelo.
  - Valores se determinan por la resolución del modelo
  - Normalmente representan los índices de rendimiento internos y externos del sistema.

Elementos del modelo: Parámetros y Relaciones Funcionales

- Parámetros
  - En este contexto hace referencia a ciertos valores numéricos que se introducen con la finalidad de poder establecer las relaciones que unen las distintas variables.
  - NO CONFUNDIR con la acepción de parámetro utilizada como sinónimo de variable de entrada.
- Relaciones funcionales
  - Conjunto de ecuaciones o inecuaciones de tipo matemático que determinan la evolución del sistema.
  - Describen las interacciones entre las variables de estado y de entrada.
  - Han de ser resolubles por métodos analíticos.
  - Variables no controlables, introducción de simplificaciones.

Descomposición jerárquica

- Enfoque que permite ampliar el conjunto de problemas resolubles analíticamente.
- Se identifican módulos, o subsistemas, contenidos unos en otros.
- Los resultados del análisis de un modulo puede ser utilizados en el análisis del siguiente módulo en la jerarquía que le contiene.
- Partiendo de los subsistemas más elementales, cada subsistema puede ser estudiado de forma aislada.



Teoría de colas

- Agner Krarup Erlang, ingeniero Danés que trabajo para la Copenhagen Telephone Exchange, publicó el primer artículo sobre teoría de colas en 1909.
- David G. Kendall introdujo la notación de colas A/B/C en 1953.
  - A: Llegadas
  - B: Servicio
  - C: Número de servidores
- La teoría de colas o de líneas de espera se puede aplicar a todos aquellos problemas que pueden caracterizarse como problemas de congestión llegada-partida.
- Una cola es una línea de espera.
- Teoría de colas
  - Es el estudio matemático del comportamiento de líneas de espera.
  - Modelos matemáticos que describen sistemas de líneas de espera o de sistemas de colas.
  - Estas se producen cuando un conjunto de clientes llegan a un "sitio" (centro de servicio) para obtener un servicio de un servidor que dispone de cierta capacidad de atención. Si el servidor no está disponible y el cliente decide esperar, entonces se forma la línea de espera.

Modelo de colas aplicados al análisis de Sistemas Computacionales

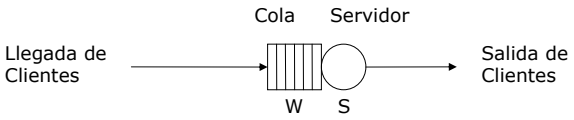
- Los trabajos en los ordenadores comparten los recursos
- En un instante de tiempo solamente un trabajo puede utilizar un recurso. El resto han de esperar en una cola para utilizarlo.
- Determinación del tiempo que los trabajos consumen en las diferentes colas del sistema.
- Clientes llegan al centro de servicio, esperan en cola si es preciso, reciben el servicio del servidor y abandonan el sistema.
- Parámetros
  - Intensidad de carga
  - Demanda de servicio: tiempo medio de servicio requerido por un cliente

Centro de servicio

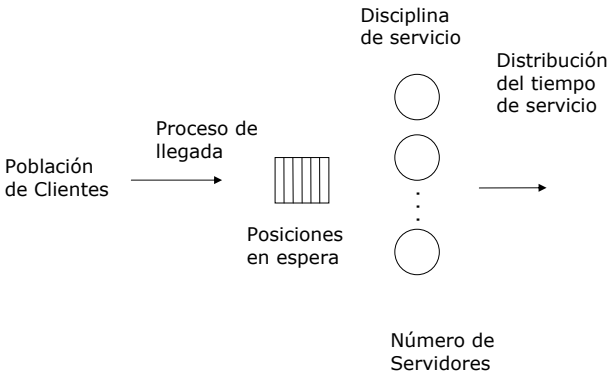
- Es un conjunto compuesto por un servidor y una cola de espera.
- El servidor representa el recurso físico del ordenador
- La cola de espera modela la cola de trabajos que esperan recibir servicio, utilizar el recurso físico.
- Parámetros temporales
  - Tiempo de servicio: tiempo que transcurre desde que un trabajo empieza a utilizar el recurso hasta que lo deja libre.
  - Tiempo de respuesta: tiempo de servicio más el tiempo que el trabajo permanece en la cola de espera.
- Tipos
  - Servidor único, una cola de espera
  - Dos (varios) servidores y una cola de espera
  - Infinitos servidores (no tiene cola de espera): los trabajos que llegan siempre encuentran un servidor disponible. Centro de tipo retardo o demora.

Centro de servicio

- Intensidad de la carga: tasa de llegada de los clientes (λ)
- Tiempo de servicio: tiempo que, por término medio, requiere cada cliente (S)
- Tiempo de espera: tiempo consumido por una petición en espera de acceso al recurso (W)



Componentes de una cola



Notación Colas (Kendall: A/S/m/B/K/SD)

- **A**
  Proceso de llegadas: La distribución de probabilidad de los periodos entre llegadas al centro de servicio.
- **S**
  Tiempo de servicio: La distribución de probabilidad de los periodos de servicio para cada petición en el centro de servicio
- **m**
  Número de servidores
- **B**
  Capacidad del sistema: El número máximo de clientes que pueden estar en el centro de servicio. Capacidad de almacenamiento.
- **K**
  Tamaño de la población: El número total de clientes potenciales que pueden llegar al sistema.
- **SD**
  Disciplina de servicio: La política de servicio (FIFO, LIFO, ….)

Notación Colas (Kendall: A/S/m/B/K/SD)

- La distribución de los tiempos entre llegadas y los tiempos de servicio (parámetros A y S) se denotan de forma general por:
  - M
    Exponencial (la M es por Markov)
  - E_k
    Erlang con parámetro k
  - H_k
    Hiperexponencial con parámetro k
  - D
    Determinista
  - G
    General
- Cuando se supone capacidad infinita de almacenamiento, tamaño infinito de la población y disciplina de servicio FCFS o FIFO los parámetros B, K y SD se suprimen.
  - Colas M/M/1

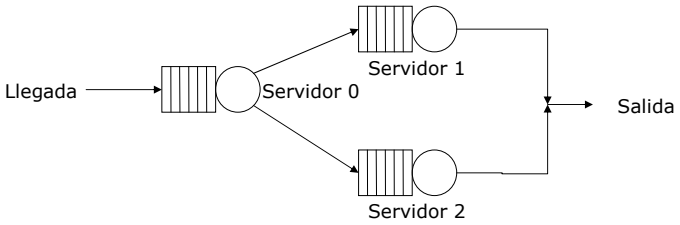
Modelado de sistemas de computación. Modelos de redes de colas.

- Un modelo en el cual los trabajos que salen de una cola (centro de servicio) llegan a otra cola (o posiblemente a la misma cola) se denominan redes de colas.
- Queueing network modeling (QN)
  - El sistema se representa como una red de colas
  - El sistema se puede evaluar analíticamente.
- Modelo de red de colas
  - Colección de centros de servicio interconectados que representan los recursos del sistema
  - Servidor : modelo del recurso del sistema (hardware)
  - Cola: modelo de la cola software asociada al recurso hardware.
  - Clientes (customer): usuarios, transacciones o trabajos en el sistema.

Modelos de redes de colas: Tipos de redes

- Redes abiertas (modelos abiertos).
  - Existencia de, al menos, una fuente de trabajos y uno o más sumideros que absorben los trabajos que salen del sistema.
  - Existe al menos un camino que, a partir de dada nodo, lleve al sumidero.
  - Tienen llegadas y salidas externas
  - El nº de trabajos en el sistema varía a lo largo del tiempo.
  - En los estudios de este tipo de sistemas se parte de productividad conocida. Su valor es igual a la tasa de entrada al sistema.
  - Los clientes que completan su servicio abandonan el sistema.
  - El objetivo es caracterizar el número de trabajos en el sistema y el tiempo de respuesta.
  - Ej: Sistemas transaccionales.

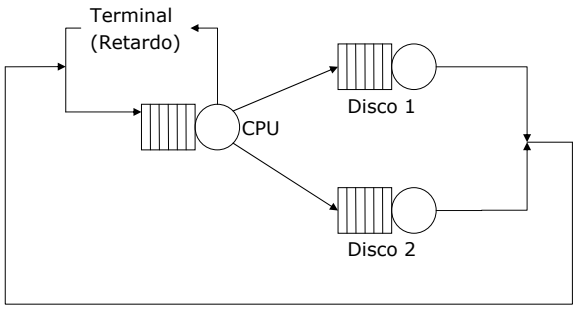
Modelos de redes de colas
Redes Abiertas



Modelos de redes de colas:
Tipos de redes

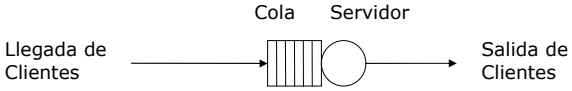
- Redes cerradas (modelos cerrados)
  - No tienen llegadas ni salidas externas.
  - El nº de trabajos o clientes en el sistema permanece constante.
  - Puede verse como un sistema donde su salida (out) se conecta con su entrada (in).
  - El flujo de trabajos saliendo del "out" y entrando por el "in" puede verse como el throughput del sistema cerrado.
  - Se parte del nº de trabajos en el sistema (N).
  - Se intenta determinar el throughput (tasa de finalización de trabajos) y el tiempo de respuesta.
  - Ej: sistemas interactivos.
    - Terminal: estación de retardo
    - N <= que el nº de terminales.

Modelos de redes de colas
Redes Cerradas



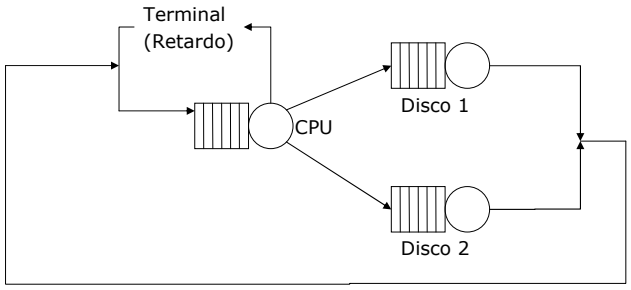
Modelos de redes de colas:
Centros de servicio simples

- [Lazowska 93, pp5] [Fortier & Michel, 2003, pp.236]
- El modelo tiene dos parámetros
  - Intensidad de la carga: tasa a la que llegan los clientes
    - Un cliente cada dos segundos o 0.5 clientes/segundo
  - Demanda de servicio: media del tiempo de servicio requerido por un cliente
    - 1.25 segundos.
- Es posible realizar evaluaciones de rendimiento del tipo:
  - Utilización: proporción de tiempo que el servidor está ocupado
  - Tiempo de residencia: tiempo medio consumido en el centro de servicio
  - Longitud de la cola : número medio de clientes en el centro de servicio, esperando y recibiendo servicio.
  - Throughput: Tasa a la que se procesan las peticiones.

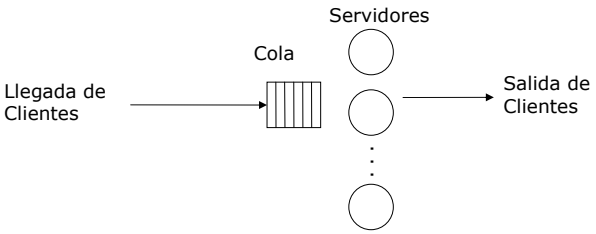


Modelo del servidor central

- Patrón de interconexión entre las estaciones de servicio.
- Introducido por Buzen en 1973.
- Intenta reproducir el comportamiento de los programas cuando se ejecutan en el ordenador.
- No considera la memoria del sistema, únicamente el procesador y las unidades de almacenamiento.



Modelos de redes de colas: Centros de servicio múltiples



Análisis Operacional

- Técnica analítica
- Toma en cuenta las relaciones entre los elementos del sistema.
- Presentado por Buzen y Denning a finales de los 1970.
- Su origen se encuentra en la teoría de colas
- Utiliza variables de más fácil verificación
  - Relaciones entre variables directamente observables en el sistema
  - En este contexto "Operacional" equivale a "Directamente medible"
  - Una hipótesis operacionalmente comprobable es una hipótesis que puede ser verificada por su medida
    - Ej: El número de llegadas de peticiones al sistema es igual al número de finalizaciones
      - … en un tiempo suficientemente largo.

Variables operacionales

- Directamente medibles en un periodo de tiempo finito.
- Ejemplo: Si se observa un dispositivo i de un sistema informático como una caja negra durante un periodo de tiempo T se pueden obtener las siguientes medidas:
  - T (tiempo). Intervalo de observación o de medida del sistema.
  - A_i (trabajos o clientes). Número de llegadas de peticiones observadas durante el intervalo T. (Arrivals)
  - C_i (trabajos o clientes). Número de peticiones completadas o servidas durante el intervalo T. (Completions)
  - B_i (tiempo). Tiempo durante el que el recurso observado (la estación de servicio) ha estado ocupado. (Busy).
- En un tiempo suficientemente largo

$$A_i \approx C_i$$

Variables operacionales deducidas

- Tasa de llegadas

  $\lambda_i = A_i / T$ trabajos por unidad de tiempo

- Productividad o Throughput

  $X_i = C_i / T$ trabajos por unidad de tiempo

- Utilización

  $U_i = B_i / T$

- Tiempo medio de servicio

  $S_i = B_i / C_i$ unidades de tiempo por trabajo

Hipótesis

- Periodo de observación T
- Sistema en estado estable o de equilibrio
  - Hipótesis de flujo de trabajos

    $A_i = C_i$, □ i
- En tiempos de observación grandes

  $A_i - C_i \rightarrow 0$
- Se debe observar que $A_i = C_i$ implica

  $X_i = \lambda_i$

Ley de la Utilización

- La utilización de un dispositivo se puede expresar en función del número de terminaciones mediante la siguiente formula:

$$U_i = \frac{B_i}{T} = \frac{C_i}{T} \times \frac{B_i}{C_i}$$

$$U_i = X_i \times S_i$$

- Esta expresión permite relacionar la productividad de un dispositivo con su tiempo de servicio.

- Si además se cumple la hipótesis de flujo equilibrado de trabajo se obtiene una expresión equivalente a la anterior en función de la tasa de llegada.

$$U_i = \lambda_i \times S_i$$

Ley del Flujo Forzado

- Es de gran importancia.
- Relaciona la productividad del sistema X_o con la productividad de un dispositivo individual X_i.
- En un modelo abierto la productividad está definida por el número de trabajos que abandonan el sistema por unidad de tiempo.
- En un modelo cerrado ningún trabajo abandona el sistema.
- Sin embargo al atravesar el enlace que une la salida con la entrada se comportan como si abandonaran el sistema e inmediatamente reentraran en él.
- La productividad en este último caso viene dada por el número de trabajos que atraviesan este enlace por unidad de tiempo.

Ley del Flujo Forzado

- Supongamos que cada tarea realiza V_i peticiones o visitas al dispositivo i.
- Si el flujo está equilibrado, el número de trabajos que sale del sistema C_o (o atraviesa el enlace exterior) y el número de trabajos que atraviesan el dispositivo i están relacionados por la expresión:

  $C_i = C_0 \times V_i$ y por tanto $V_i = C_i / C_0$
- La variable V_i recibe el nombre de razón de visitas al dispositivo i.
- La productividad del sistema durante el periodo de observación es:

  $X_0 = C_0 / T$
- La productividad del dispositivo i es:

$$X_i = \frac{C_i}{T} = \frac{C_i}{C_0} \times \frac{C_0}{T}$$

Ley del Flujo Forzado

- Finalmente se obtiene una expresión de X_i en función de las variables X_0 y V_i.

  $X_i = X_0 \times V_i$ (Ley del flujo forzado)
- Esta ley establece que el flujo a través de un determinado dispositivo de la red determina el flujo en cualquier otro dispositivo.
- Es válida solo si lo es la hipótesis de flujo equilibrado.
- Combinando este resultado y la ley de utilización se puede obtener la siguiente expresión para el valor de la utilización del dispositivo.

  $U_i = X_i \times S_i = X_0 \times V_i \times S_i = X_0 \times D_i$

  donde $D_i = V_i \times S_i$ recibe el nombre de Demanda de servicio sobre el dispositivo i en todas las visitas que un trabajo realiza al mismo.
- La relación anterior establece que la utilización de cada dispositivo es proporcional a su demanda de servicio.

Ley del Flujo Forzado

- Las razones de visita son otra forma de especificar el encaminamiento de los trabajos a través de la red.
- Otra descripción equivalente se puede realizar mediante la proporción de trabajos, también llamada probabilidad de encaminamiento o de transición.
  - Las probabilidades de encaminamiento p_{ij} indican la proporción de trabajos que salen de la estación i y se dirigen a la estación j.
  - Equivalente: La probabilidad de que un trabajo pase a la estación j después de terminar su servicio en la estación i.
- En este sentido se tendrá que:

$$p_{ij} = C_{ij} / C_i \quad \text{y}$$

- En particular:

$$p_{0j} = C_{oj} / C_0 \quad \text{y}$$
$$p_{i0} = C_{i0} / C_i$$

Ley del Flujo Forzado

- Razones de visita y probabilidades de encaminamiento son equivalentes en el sentido de que a partir de una se obtienen las otras.
- En un sistema con K estaciones de trabajo en que se cumple la hipótesis del flujo equilibrado de trabajos se tiene:

$$C_j = \sum_{i=0}^{K} C_i \times p_{ij}$$

  donde el subíndice 0 representa el exterior del sistema y p_{i0} es la proporción de trabajos que, después de recibir servicio en la estación i, abandonan la red.
- Dividiendo ambos lados de la igualdad por C_0 obtenemos

$$V_j = \sum_{i=0}^{K} V_i \times p_{ij}$$

  que representan las denominadas ecuaciones de razones de visita.
- Como cada visita al mundo exterior corresponde a una terminación de un trabajo, tendremos que siempre se cumplirá la ecuación:

$$V_0 = 1$$

Ley de Little

- Enunciada a principios de la década de 1960.
- La única hipótesis requerida para su aplicación es la del flujo equilibrado de trabajos.
- Si llamamos N_i al número de trabajos y R_i al tiempo de respuesta de la estación de servicio i, la ley de Little establece que:

$$N_i = \lambda_i \times R_i$$

- Al exigirse que se cumpla la hipótesis del flujo equilibrado de trabajos se puede sustituir λ_i por X_i.

$$N_i = X_i \times R_i$$

- Esta ley es de gran interés en el estudio de modelos de colas, ya que combina índices de suma importancia en los estudios de rendimiento:
  - Tiempo de Respuesta y
  - Productividad
- Se puede aplicar a cualquier parte del modelo con la única condición que se cumpla la hipótesis del flujo equilibrado de trabajos.

Ley General del Tiempo de Respuesta

- El número de trabajos en una red de colas formada por K estaciones se puede expresar como

$$N = N_1 + N_2 + \ \ldots \ + N_K$$

- Si se sustituyen los valores de N_i de acuerdo con la Ley de Little se tiene:

$$X_0 \times R = X_1 \times R_1 + X_2 \times R_2 + \ \ldots \ + X_K \times R_K$$

$$X_0 \times R = \sum_{i=1}^{K} X_i \times R_i$$

- Dividiendo ambos miembros de la igualdad por X_0 y aplicando la ley del flujo forzado quedará la expresión

$$R = V_1 \times R_1 + V_2 \times R_2 + \ \ldots \ + V_K \times R_K$$

$$R = \sum_{i=1}^{K} V_i \times R_i$$

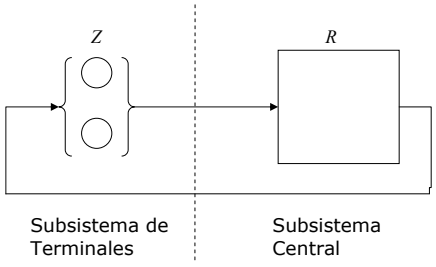
- Esta expresión recibe el nombre de ley general del tiempo de respuesta, y permite ver claramente que el tiempo de permanencia de un trabajo en el sistema depende del número de visitas que realiza a cada dispositivo y del tiempo de respuesta que experimenta en él por cada una de las visitas.

Ley del Tiempo de Respuesta Interactivo

- Todos los modelos con carga interactiva pueden dividirse conceptualmente en dos partes:
  - Subsistema de terminales, modela el tiempo de reflexión.
  - Subsistema central, contiene los dispositivos físicos del computador contemplados por el modelo.
- El tiempo de reflexión (think time), identificado habitualmente mediante la variable Z, es el tiempo que transcurre desde que un trabajo abandona el subsistema central hasta que entra de nuevo en él.
  - Para sistemas interactivos $Z>0$
  - Para sistemas por lotes el valor de Z es cero.
- El tiempo de respuesta del sistema, R, corresponderá al tiempo que un trabajo pasa en el subsistema central.

Ley del Tiempo de Respuesta Interactivo

- El funcionamiento del sistema es el que sigue:
  - Los usuarios generan peticiones desde los terminales que se sirven del subsistema central.
  - Una vez atendidas las peticiones vuelven a los terminales.
- Los terminales están modelados por una estación con infinitos servidores (no hay tiempo de espera en la cola).
- Transcurrido el tiempo de reflexión los usuarios generan la siguiente petición.



Ley del Tiempo de Respuesta Interactivo

- Podemos aplicar la Ley de Little al conjunto de los dos subsistemas (subsistema central y subsistema de terminales).
- El número de trabajos en el conjunto es N.
- El tiempo medio que permanece en el conjunto es igual a $Z+R$.
- Aplicando la Ley de Little se puede escribir:

  $$N = (Z+R) \times X_0$$

  y despejando la variable R obtenemos la expresión de la ley del tiempo de respuesta interactivo.

  $$R = \frac{N}{X_0} - Z$$

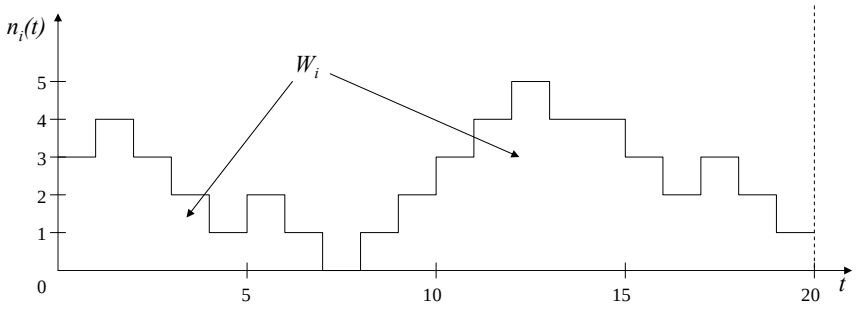
- Nótese que el número de trabajos en los terminales viene dado, empleando la ley de Little, por

  $$Z \times X_0$$

  y el número de trabajos dentro del sistema que compiten por los recursos es:

  $$R \times X_0$$

Ley de Little: Trabajos en el sistema y Tiempo

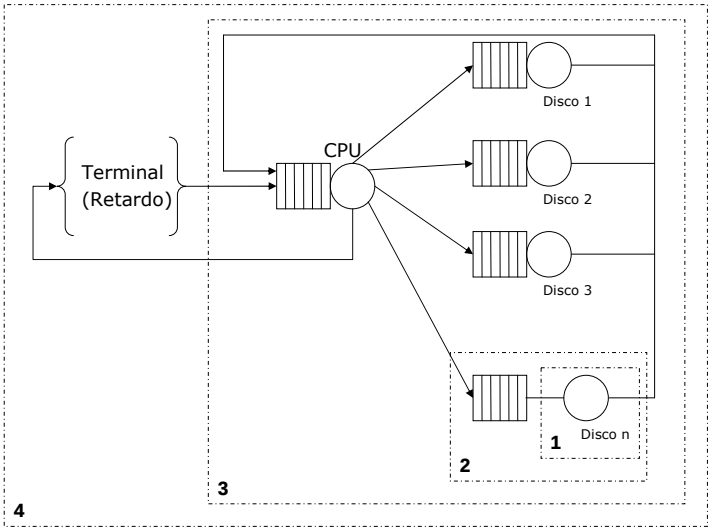


$N_i = W_i/T$

$R_i = W_i/C_{i,}$ considerando que $X = C/T$ se tiene

$\boldsymbol{N_i = X_i \times R_i}$ ***Ley de Little***

Ley de Little: Aplicación a distintos niveles



Ley de Little: Aplicación a nivel 1

- El recurso es utilizado siempre que haya una petición presente.
- La utilización puede interpretarse como el número medio de peticiones del recurso
  - Hay una petición utilizando el recurso durante un U_i% del tiempo
  - Cero peticiones durante (1- U_i)% de tiempo
  - Población
    - Utilización del recurso
    - $N_i \equiv U_i$
  - Productividad
    - Tasa de satisfacción de peticiones: X_i
  - Tiempo de residencia
    - Tiempo medio de servicio requerido por una petición en el recurso R_i
    - $R_i \equiv S_i$

Ley de Little: Aplicación a nivel 2

- Se toman en cuenta el servicio y la cola de tareas
  - Población
    - Número de peticiones en cola y en servicio: N_i
  - Productividad
    - Tasa de satisfacción de peticiones: X_i
  - Tiempo de residencia
    - Tiempo medio de servicio requerido por una petición en el recurso más el tiempo medio en cola : R_i

Ley de Little: Aplicación a nivel 3 (Subsistema central)

- Las peticiones al sistema se tratan como interacciones
  - Población
    - Número de usuarios (N). Interacciones a nivel de sistema. Usuarios que no están pensando
  - Productividad
    - Tasa a la que las interacciones fluyen entre las terminales y el subsistema central: X_i (interacciones segundo)
  - Tiempo de residencia
    - Tiempo de respuesta convencional : R_i
    - Tiempo transcurrido entre que un usuario envía una petición hasta que la respuesta es devuelta al usuario. (seg.)

Ley de Little: Aplicación a nivel 4

- Población
  - Número total de usuarios (N) interactivos
- Productividad
  - Tasa a la que las interacciones fluyen entre las terminales y el subsistema central: X_i (interacciones segundo)
- Tiempo de residencia: R+ Z
  - Tiempo de respuesta más tiempo de reflexión de los usuarios
  - Tiempo de respuesta de un sistema interactivo.

Aplicaciones del Análisis Operacional

- Se aplicarán los instrumentos del análisis operacional para:
  - Estimar el rendimiento de un sistema informático.
  - Cuantificar el efecto en las prestaciones de mejoras a los componentes del sistema.
- Se presentarán dos algoritmos para estimar el tiempo de respuesta y la productividad de un sistema informático.
  - Se considerará una única clase de trabajos.
  - Se supondrá que los tiempos entre llegadas y los tiempos de servicio se distribuyen de forma exponencial.
  - La simplificación es satisfactoria para los tiempos entre llegadas pero no tanto para los tiempos de servicio, aunque es una aproximación aceptable.
- También se calcularán los valores más optimistas del rendimiento.
  - Estos límites reciben el nombre de **límites asintóticos.**
  - Se establecen de manera sencilla.
  - Proporcionan una cota superior del tiempo de respuesta como de la productividad alcanzable.
- Finalmente se verá una forma sencilla de evaluar cuantitativamente el efecto sobre las prestaciones de una mejora en el sistema.

Estimación del rendimiento (1/2)

- Se presentarán dos algoritmos clásicos para resolver modelos de colas sencillos.
- Hipótesis:
  - Si un trabajo está sirviéndose en una estación, el tiempo que le falta para abandonar el servidor es independiente del tiempo que ya lleva en servicio.
  - En un sistema abierto, el tiempo que transcurre hasta la próxima llegada es independiente del instante en que se produjo la última.
- Las hipótesis equivalen a suponer que tanto las distribución del tiempo de servicio como la distribución del tiempo de llegadas en un modelo abierto son exponenciales.
  - Debido a sus propiedades estadísticas se dice que está distribución carece de memoria (*memoryless property*).
- Antes de plantear los algoritmos de resolución introduciremos una expresión para calcular el tiempo de respuesta de una estación de servicio i de tipo cola.

Estimación del rendimiento (2/2)

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

- Un trabajo que llega a la estación i encuentra N_i trabajos en ella y esperará $N_i \times S_i$ unidades de tiempo a que se sirvan, más, S_i para recibir su propio tiempo de servicio.
- Se está utilizando la propiedad de que el tiempo de servicio se distribuye exponencialmente (carece de memoria), y, por tanto no es necesario tener en cuenta el tiempo de servicio ya recibido por el cliente que esta en el servidor cuando se produce la llegada.
- La propiedad de carencia de memoria no puede ser comprobada operacionalmente, por eso no es considerada como una ley operacional.
- Sustituyendo N_i por $X_i \times R_i$ (Ley de Little) podemos relacionar el tiempo de respuesta de una estación i con su tiempo de servicio S_i y su utilización U_i para calcular fácilmente R_i y aplicar las leyes operacionales vistas anteriormente.

$$R_i = (X_i \times R_i + 1)\times S_i \text{, entonces:}$$

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

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

Estimación del rendimiento: Algoritmo para redes abiertas (1/2)

- Los tiempos de servicio y los tiempos entre llegadas están distribuidos exponencialmente.
- Se conocen la razón de visita V_i y el tiempo de servicio S_i de las K estaciones de la red.
- Además se supone conocida la tasa de llegada al λ sistema, la que será igual a la productividad del sistema. (se supone válida la hipótesis de flujo equilibrado de trabajos).
- Se calcularán las variables
  - Para cada estación: X_i , N_i, R_i y U_i.
  - Para toda la red: R y N.
- Demanda de servicio: $D_i = V_i \times S_i$
- Utilizaciones: $U_i = \lambda \times D_i$
- Productividades $X_i = \lambda \times V_i$
- Tiempos de respuesta por estación:
  - Si es tipo cola: $R_i = S_i/(1-U_i)$
  - Si es tipo retardo: $R_i = S_i$

Estimación del rendimiento: Algoritmo para redes abiertas (2/2)

- Finalmente el tiempo de respuesta del sistema se obtiene a partir de los R_i y V_i aplicando la ley general del tiempo de respuesta:

$$R = \sum_{i=1}^{K} V_i \times R_i$$

- El número de trabajos en el sistema se calcula sumando los trabajos contenidos en todas las estaciones del modelo:

$$N = \sum_{i=1}^{K} N_i$$

  o aplicando la ley de Little al sistema completo.

Ejemplo

- Sea una red de colas abierta con dos dispositivos 1 y 2 y una tasa de llegadas de 2 trabajos por segundo. Con los siguientes tiempos de servicio y razones de visita:

	$Dispositivo_1$	$Dispositivo_2$
Razón de Visita	6	7
Tiempo de servicio(seg)	0,01	0,02

- Se calculan las utilizaciones:

$$U_1 = \lambda \times D_1 = \lambda \times V_1 \times S_1 = 2 \times 6 \times 0.01 = 0.12$$
$$U_2 = \lambda \times D_2 = \lambda \times V_2 \times S_2 = 2 \times 7 \times 0.02 = 0.28$$

- Se calculan los tiempos de respuesta de cada estación:

$$R_1 = S_1 / (1-U_1) = 0.01/(1-0,12) = 0.0114 \text{ s}$$
$$R_2 = S_2 / (1-U_2) = 0.02/(1-0,28) = 0.0278 \text{ s}$$

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

$$R = V_1 \times R_1 + V_2 \times R_2 = 6 \times 0.0114 + 7 \times 0.0278 = 0.263 \text{ s}$$
$$N = \lambda \times R = 2 \times 0.263 = 0.526 \text{ trabajos}$$

Estimación del rendimiento: Algoritmo para redes cerradas (1/3)

- También denominado análisis del valor medio (*MVA: Mean-Value Analysis*)
- Se conocen la razón de visita V_i y el tiempo de servicio S_i de las K estaciones de la red, además del tiempo de reflexión Z (que será nulo para un sistema por lotes).
- Las variables a calcular son similares al caso anterior.
  - La diferencia estriba en que no se conoce la productividad del sistema, sino que se ha de estimar.
  - Al tratarse de un modelo cerrado se conoce el número de trabajos N en el sistema.

Estimación del rendimiento: Algoritmo para redes cerradas (2/3)

- Se emplea una ecuación que permite estimar R_i para las estaciones de tipo cola teniendo en cuenta que ahora su valor dependerá del número de trabajos N en el sistema:

$$R_{i(N)} = [\, N_{i(N-1)} + 1 \,] \times S_i$$

  donde $N_{i(N-1)}$ es el número de trabajos en la estación i cuando en la red cerrada hay $(N-1)$ trabajos.
  - Esta relación establece que el estado de la red visto por un trabajo que está en tránsito de una estación a otra (el trabajo ha abandonado una estación, pero aún no se ha incorporado a la siguiente), tiene la misma distribución que el estado que vería un observador aleatorio si el número total de trabajo en la red fuese $N-1$.
  - Esta afirmación es bastante intuitiva, que un trabajo en transito no puede observarse a sí mismo en ninguna estación.
- La ecuación anterior relaciona dos índices de prestaciones, uno para N y otro para $N-1$ dando lugar a un procedimiento de cálculo iterativo.

Estimación del rendimiento: Algoritmo para redes cerradas (3/3)

- Los valores para la primera iteración son fáciles de establecer:
- Para $N=0$ se cumple
  - $N_i = 0$ y por tanto $R_i(1) = S_i$ para $i=1, \ldots, K$
- Para las estaciones de tipo retardo se cumple, además, que $R_i(N) = S_i$, $\forall N$
- Así el algoritmo de resolución tendrá la siguiente forma.

Para n desde 1 hasta N hacer :

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

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

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

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

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

Ejemplo

- Sea una red cerrada con 3 trabajos y dos dispositivos, 1 y 2, que tienen los siguientes tiempos de servicio y razones de visita.

	$Dispositivo_1$	$Dispositivo_2$
Razón de Visita	15	14
Tiempo de servicio(seg)	0,03	0,5

- Se supone que la carga es interactiva con un tiempo de reflexión Z=5.
- Para aplicar el algoritmo hay que aplicar 4 iteraciones, una por cada trabajo presente en el sistema.
- Para cada iteración se calcula:
  - El tiempo de respuesta y la productividad del sistema
  - El tiempo de respuesta, número de trabajos, la productividad y la utilización de cada estación del modelo.

Trabajos: n	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	11,4020	0,1219	0,0569	1,3335
3	0,0317	1,1667	16,8098	0,1376	0,0654	2,2468

- El tiempo de respuesta del sistema, a partir de los datos presentados en la tabla, es de 16.8090 seg., mientras que la productividad es de 0.1376 trabajos por segundo.

Cuellos de botella

- Una consecuencia de la ley del flujo forzado es que las utilizaciones de los dispositivos son proporcionales a las demandas totales de servicios.
- Un cuello de botella es aquel dispositivo con mayor demanda de servicio y por tanto mayor utilización.
  - Su papel resulta determinante en las prestaciones del sistema completo.
  - Cuando la carga incrementa su magnitud el dispositivo que tiende a congestionarse primero es este cuello de botella.
  - Cuando su utilización presenta valores cercanos a 1 se dice que está *saturado*.
- Es deseable que las utilizaciones de los distintos dispositivos sean lo más parecidas posible.
  - Cuando esto ocurre el sistema está *equilibrado* (*balanced system*).
- La mejora del comportamiento del dispositivo cuello de botella redundará en un incremento significativo del rendimiento del sistema completo.
  - La mejora será marginal cuando se haga en los restantes dispositivos.

Límites asintóticos

- Representan una técnica de aplicación sencilla para acotar, desde un punto de vista optimista, los mejores valores de la productividad y el tiempo de respuesta de un sistema informático.
- Se denotará al dispositivo cuello de botella del sistema informático mediante el subíndice b.
- Una vez localizado esté dispositivo se cumpliran las siguientes igualdades:

$$D_b = máx\{D_1, D_2, \ldots D_K\} = V_b \times S_b$$
$$U_b = X_0 \times D_b$$

Límites asintóticos
Sistemas abiertos (1/2)

- Como primera aproximación al establecimiento de límites asintóticos optimistas se considera inicialmente un modelo de colas abierto.
  - El valor máximo de la tasa de llegadas que el sistema puede soportar será aquel que sature completamente el dispositivo cuello de botella.
  - Es decir $U_b = 1$
- Como se cumple la hipótesis de flujo equilibrado de trabajos se puede escribir:

$$\boldsymbol{U_b = X_b \times S_b = X_o \times V_b \times S_b = X_0 \times D_b = \lambda D_b}$$

- Sea λ_{opt} el valor más alto de la tasa de llegadas que el sistema puede aceptar, la cual será equivalente a la productividad del sistema, que denotaremos por X_{opt}.
- Particularizando la expresión anterior para $U_b = 1$ se tiene:

  Si $U_b = 1$ => $X_{opt} \times D_b = 1$ => $X_{opt} = 1 / D_b$
- Cuando la tasa de llegadas al sistema toma el valor $\lambda = X_{opt}$ el sistema satura el cuello de botella. El número de trabajos en el sistema crece de forma indefinida, haciendo que éste se vuelva inestable.

Límites asintóticos
Sistemas abiertos (2/2)

- Si tomamos en cuenta el tiempo de respuesta, el valor optimista del mismo R_{opt} viene dado cuando un trabajo que llega al sistema lo encuentra vacío.
  - El trabajo no habrá de esperar en ningún dispositivo.
  - En este caso, el valor del tiempo de respuesta será equivalente a la suma de las demandas de servicio que haga a los diferentes dispositivos del sistema.
  - Si el modelo tiene K estaciones:

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

- Resumiendo los resultados obtenidos para el modelo abierto se obtienen las siguientes expresiones para los límites asintóticos:

$$\begin{cases} X_{opt} = \dfrac{1}{D_b} \\ R_{opt} = D = \displaystyle\sum_{i=1}^{K} D_i \end{cases}$$

Límites asintóticos
Sistemas cerrados (1/3)

- Se analizará una red de colas cerrada que modela un sistema interactivo
  - Para un sistema por lotes basta con hacer $Z=0$.
- Dado que la carga del sistema viene establecida por N, se consideran dos situaciones:
  - Carga muy baja (sistema vacío: $N=0$)
  - Carga muy alta (sistema saturado, valores de N suficientemente grandes para saturar el cuello de botella)
- Supongamos que el sistema no tiene ningún dispositivo saturado. El valor más optimista para el tiempo de respuesta, R_{opt}, es aquel que experimenta un trabajo que no tiene que esperar en ningún dispositivo.

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

- La particularización de la ley del tiempo de respuesta interactivo para el valor R_{opt} permite obtener una expresión optimista para la productividad:

$$\text{Si} \quad R_{opt} = \frac{N}{X_{opt}} - Z \quad \Rightarrow \quad X_{opt} = \frac{N}{D+Z}$$

Límites asintóticos
Sistemas cerrados (2/3)

- Si ahora se considera un sistema en el cual el cuello de botella está saturado ($U_b=1$) el valor más alto que cabría esperar para la productividad será:

  Si $U_b=1$ => $X_b \times S_b = X_{opt} \times V_b \times S_b = 1$ => $X_{opt} = 1/D_b$

- Particularizando la expresión de la ley del tiempo de respuesta interactivo para este valor de la productividad se tiene:

$$R_{opt} = \frac{N}{X_{opt}} - Z \quad \Rightarrow \quad R_{opt} = N \times D_b - Z$$

- Resumiendo los resultados obtenidos para el modelo cerrado se obtienen las siguientes expresiones para los límites asintóticos:

$$\begin{cases} X_{opt} = mín\left\{ \dfrac{N}{D+Z}, \ \dfrac{1}{D_b} \right\} \\ R_{opt} = máx\left\{ D, \ D_b \times N - Z \right\} \end{cases}$$

Límites asintóticos
Sistemas cerrados (3/3)

- El punto de intersección de las rectas definidas por las anteriores expresiones es:

$$\frac{N}{D+Z} = \frac{1}{D_b} \quad \Rightarrow \quad N = \frac{D+Z}{D_b}$$

- Al valor anterior de N se lo conoce como punto teórico de saturación, ya que desde un punto de vista optimista y asintótico, en el se alcanza la productividad teórica más alta del sistema.
- Si se considera el tiempo de respuesta, a partir de este valor no se puede garantizar el tiempo mínimo establecido por D por que los trabajos en el sistema experimentan esperas en los dispositivos, al menos en el cuello de botella.
- Como el número de trabajos en el sistema viene dado por un número entero, en la práctica el valor anterior se suele expresar como un valor entero.

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

Ejemplo

- Sea un modelo de sistema basado en el servidor central con una carga de tipo interactivo.
- Los trabajos tienen un tiempo medio de reflexión de 6 segundos. La red de colas tiene tres dispositivos: un procesador y dos discos. Las razones de visita y tiempos de servicio se indican en la siguiente tabla.

Dispositivo	Razón de Visita	Tiempo de servicio
Procesador (1)	32	0,0375
Disco (2)	25	0,02
Disco (3)	6	0,05

- Se calculan las demandas de servicio:

$$D_1 = V_1 \times S_1 = 32 \times 0.0375 = 1,2\ s$$
$$D_2 = V_2 \times S_2 = 25 \times 0.02 = 0.5\ s$$
$$D_3 = V_3 \times S_3 = 6 \times 0.05 = 0.3\ s$$

- Es evidente que el cuello de botella es el procesador.
  - Supera con diferencia las demandas de servicio de los discos.

Ejemplo

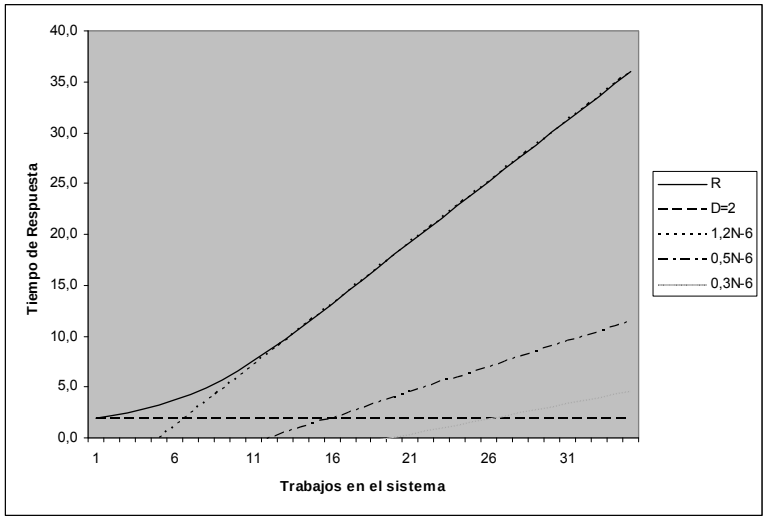
- La demanda total del sistema es la suma de las demandas de los dispositivos

$$D = D_1 + D_2 + D_3 = 1,2\ s + 0,5\ s + 0,3\ s = 2\ s$$

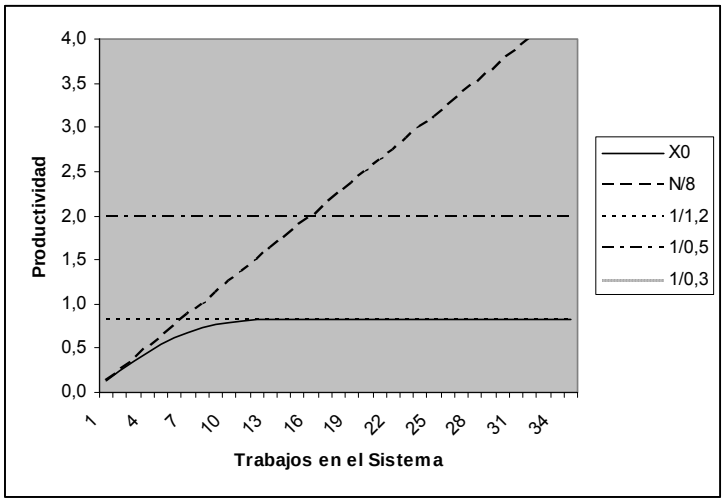
- Si se calculan los límites asintóticos de rendimiento se obtiene:

$$\begin{cases} X_{opt} = mín\left\{\dfrac{N}{D+Z},\ \dfrac{1}{D_b}\right\} = mín\left\{\dfrac{N}{8},\ 0.833\right\} \\ R_{opt} = máx\left\{D,\ D_b \times N - Z\right\} = máx\left\{2,\ 1.2 \times N - 6\right\} \end{cases}$$

Ejemplo



Ejemplo



Mejora del rendimiento

- Debido a la multitud de factores que influyen en el rendimiento de un sistema informático la mejora del rendimiento es una tarea compleja.
- La solución más sencilla al problema de mejora de rendimiento será la localización del cuello de botella del sistema y realizar actuaciones sobre él.
- La primera aproximación consiste en actuar sobre los componentes físicos del mismo, mejorándolos o incrementando su número (*upgradign techniques*).
  - La adición de nuevos componentes puede suponer una gran inversión o el reemplazo puede ser complicado.
- La segunda técnica (ajuste o sintonización: *tunning techniques*) engloba todas las acciones realizadas sobre los programas que se ejecutan en el ordenador para mejorar el uso que hacen de los dispositivos.
  - Su aplicación depende del grado de conocimiento del programa a modificar como del comportamiento e interacción del mismo con los dispositivos físicos del sistema. Siempre y cuando los programas sean susceptibles de modificación o ajuste.

Ejemplo

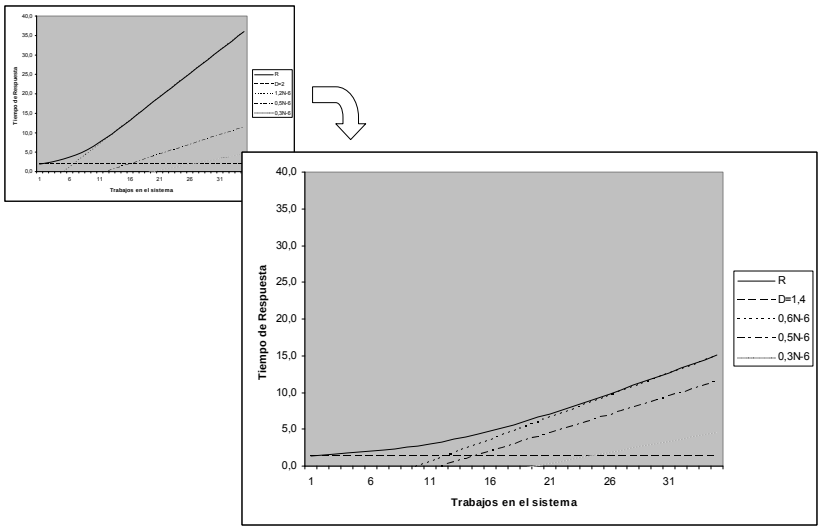
- Para mejorar el sistema del último ejemplo se debe actuar sobre el procesador.
  - ¿Qué pasará si se sustituye por una unidad dos veces más rápida?
- Se reduce el tiempo de servicio de 0,0375 a 0,0175.
- La nueva demanda de servicio es

$$D_1 = V_1 \times S_1 = 32 \times 0.001875 = 0.6\ s$$

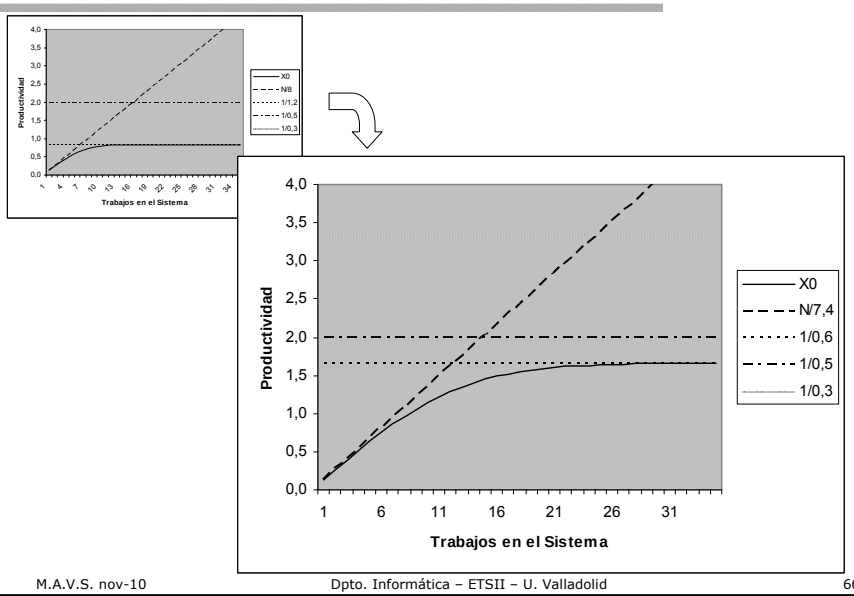
- Aunque se ha reducido a la mitad, aún es el dispositivo con la demanda más alta del sistema.
- Los nuevos límites asintóticos serán:

$$\begin{cases} X_{opt} = mín\left\{\dfrac{N}{D+Z},\ \dfrac{1}{D_b}\right\} = mín\left\{\dfrac{N}{7.4},\ 1.667\right\} \\ R_{opt} = máx\left\{D,\ D_b \times N - Z\right\} = máx\left\{1.4,\ 0.6 \times N - 6\right\} \end{cases}$$

Ejemplo (cont.)



Ejemplo (cont.)



Ejemplo (cont.)

- Se puede observar que el tiempo de respuesta mínimo y la productividad máxima han mejorado. Las asíntotas del cuello de botella se han desplazado.
  - La asíntota D del sistema original ha pasado de 2 a 1,4 segundos.
  - La asíntota del cuello de botella que pasa de $1,2 \times N-6$ a $0,6 \times N-6$ la pendiente pasa a valer la mitad.
  - Si se analiza la evolución de la productividad, cuyo valor máximo ha crecido de $1/1,2=0.833$ a $1/0,6=1.667$.
  - Las asíntotas de los discos han permanecido iguales.
- El punto teórico de saturación ha mejorado sensiblemente pasando de 7 a 13 trabajos.

$$N^* = \left\lceil \frac{D+Z}{D_b} \right\rceil = \left\lceil \frac{1.4+6}{0.6} \right\rceil = \lceil 12.33 \rceil = 13$$

- Si la mejora del procesador hubiese sido de un factor de 3, entonces su demanda hubiera pasado a valer 0.4 segundos, valor por debajo de la demanda del primer disco: Ese disco sería el nuevo cuello de botella.