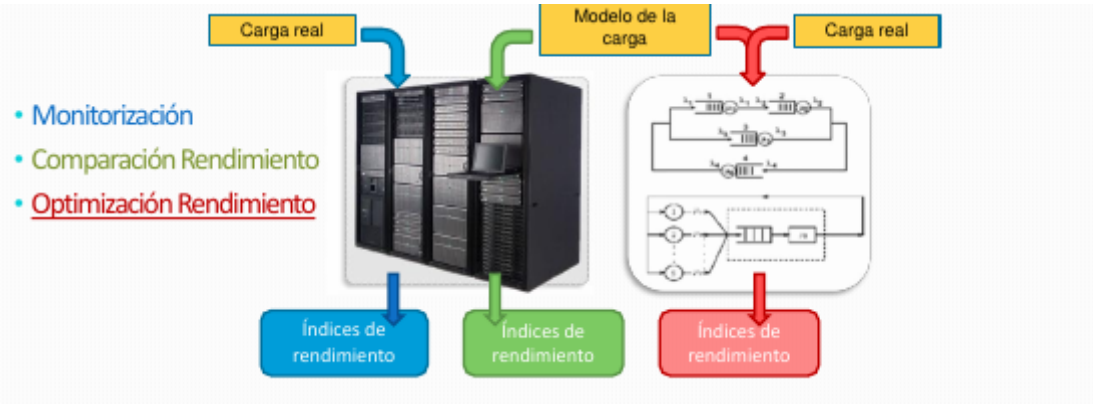
ISE 22/23 - Tema 5:

¿Cómo mejorar el rendimiento de mi servidor?

> *Objetivos:*
>
> - *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.*

5.1. Introducción - Redes de colas de espera

¿Cómo podemos mejorar el rendimiento de un servidor?



El modelo de un sistema informático

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

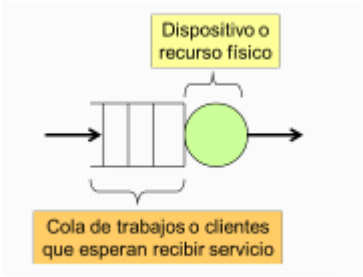
- Dispositivos (resources): núcleos lógicos, unidades de almacenamiento permanentes, tarjetas de red, etc.

- Trabajos (jobs): procesos, accesos, peticiones, etc.

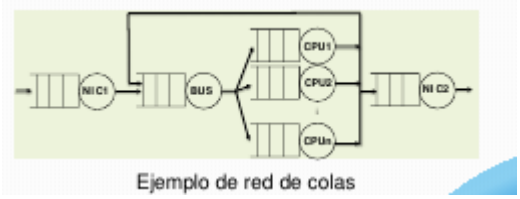
Normalmente un dispositivo o recurso solo puede ser usado por un trabajo a la vez. El resto de trabajos tendrá que esperar.

Modelos basados en redes de colas (queueing networks):

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



Concepto de estación de servicio

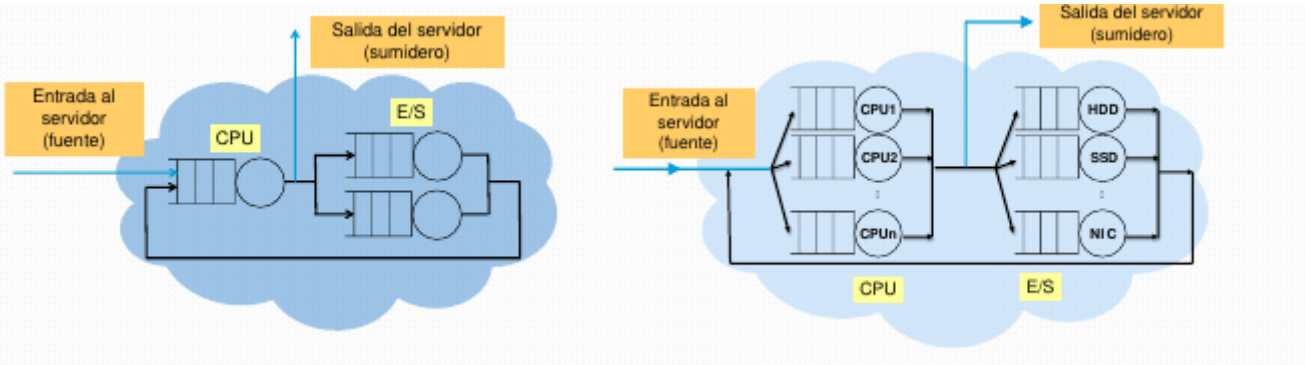


Ejemplo de red de colas

El modelo de servidor central

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

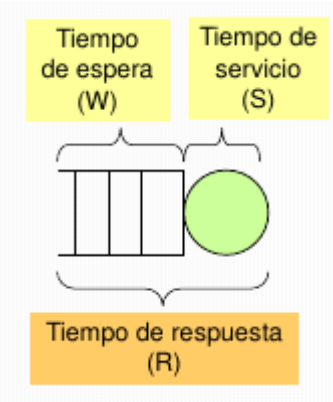
1. Un trabajo que "llega" al servidor comienza utilizando el procesador.
2. Después de "abandonar" el procesador, el trabajo puede:
    1. Terminar (sale del servidor), o bien
    2. Realizar un acceso a una unidad de entrada/salida (discos, red,...).
3. Después de una operación con una unidad de entrada/salida, el trabajo vuelve a "visitar" al procesador.



Algunas variables características a un trabajo, en una estación de servicio, en un instante concreto

- Tiempo de espera en cola (W, waiting time):
  - Tiempo transcurrido desde que el trabajo solicita hacer uso del recurso físico (se pone en la cola) hasta que realmente empieza a utilizarlo.
- Tiempo de servicio (S, service time):
  - Desde que el trabajo accede al recurso físico hasta que lo libera (tiempo que tarda el recurso físico en procesar el trabajo).
- Tiempo de respuesta (R, response time)
  - Suma de los dos tiempos anteriores.

$$R = W + S$$

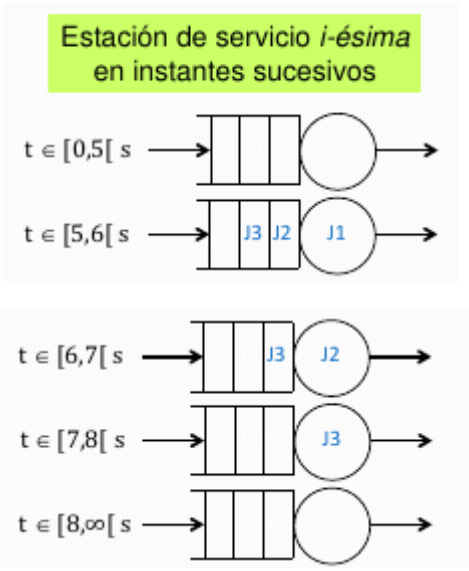


Recopilando estas medidas para múltiples trabajos, obtendremos distribuciones de probabilidad que caracterizan a esa estación de servicio.

Ejercicio

> *Suponga que la estación de servicio i-ésima de una red de colas tiene un tiempo de servicio constante Si=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: J1, J2 y J3 (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.*



Wi(J1) = 0s Ri(J1) = 1s

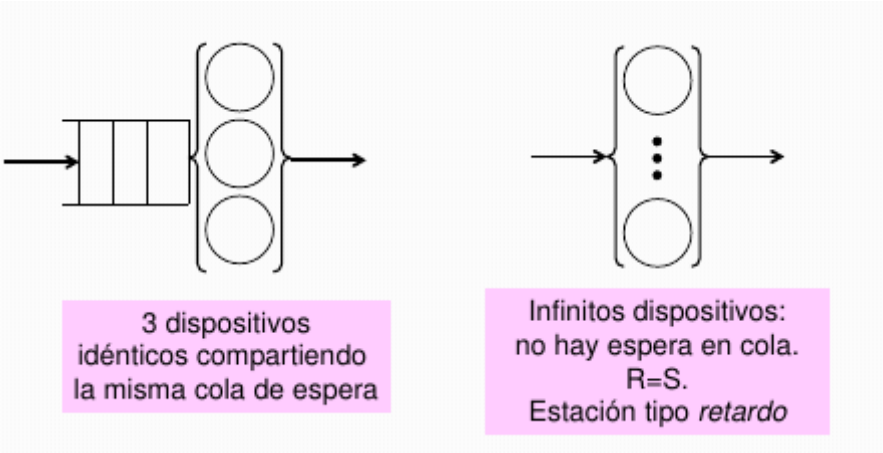
Wi(J2) = 1s Ri(J2) = 2s

Wi(J3) = 2s Ri(J3) = 3s

Valor medio Wi = 1s Valor medio Ri = 2s

Estaciones con más de un servidor

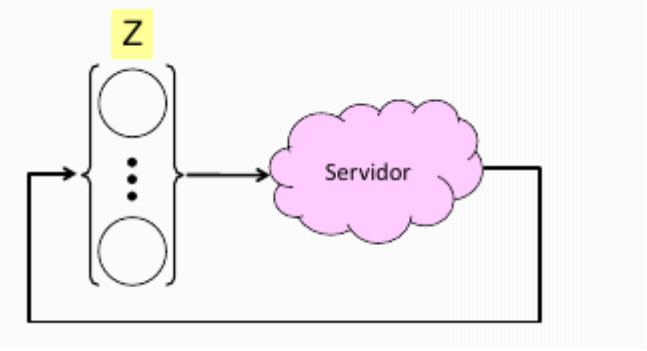
Son capaces de atender a más de un trabajo en paralelo:



El tiempo de reflexión Z (think time)

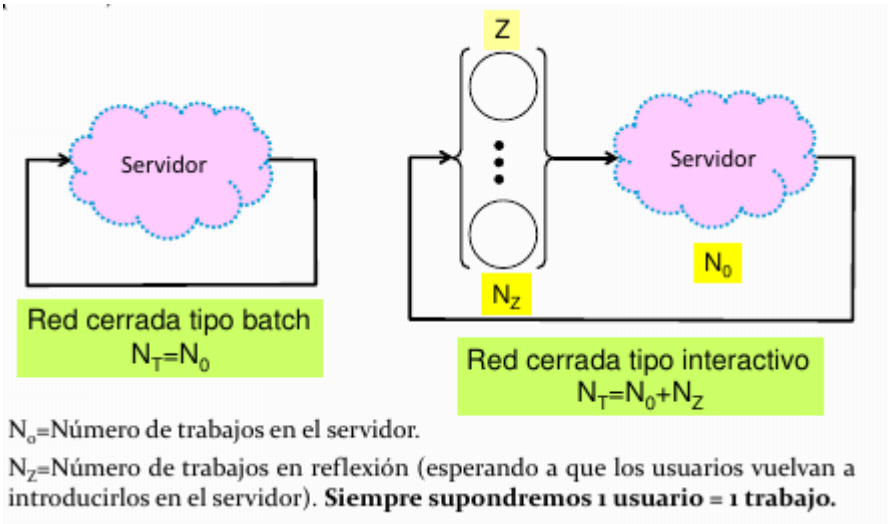
Es un parámetro (Z) que representa el tiempo que requiere el usuario antes de volver a lanzar una petición al servidor tras la respuesta de éste. Se suele modelar mediante una estación de servicio tipo retardo con un tiempo de servicio = Z.

Para ello, realizamos una hipótesis adicional: cada usuario envía un único trabajo al servidor.

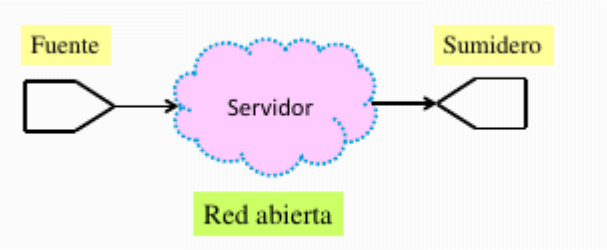


Redes de colas

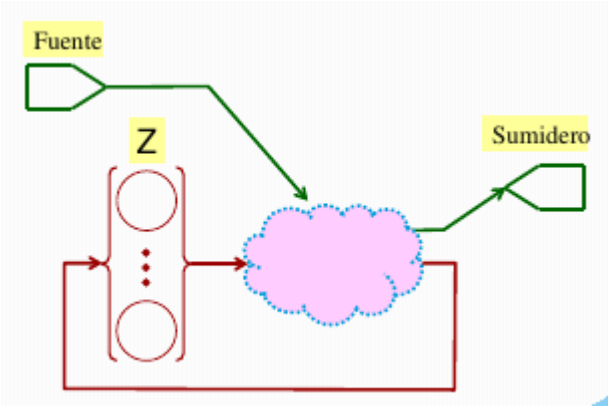
- **Cerradas**
  - Presentan un número constante de trabajos que van recirculando por la red (NT). Dependiendo de si hay o no interacción con usuarios se distingue entre:
    - Redes de tipo batch (por lotes)
    - Redes interactivas



- **Abiertas**
  - Los trabajos llegan a la red a través de una fuente externa que no controlamos. Tras ser procesados, salen de ella a través de uno o más sumideros. No existe realimentación entre sumidero y fuente.
  - El número de trabajos en el servidor (N0) puede variar con el tiempo.



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



5.2. Variables y leyes operacionales

El análisis operacional

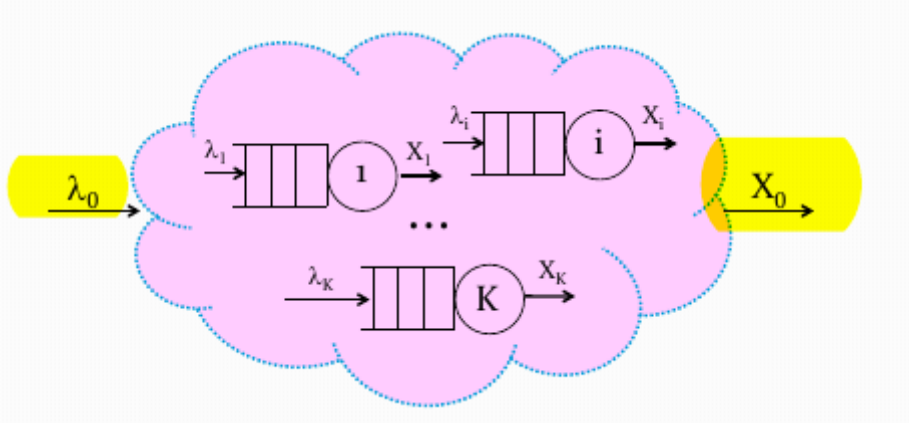
Técnica de análisis de redes de colas basada en valores medios de diferentes variables medibles (variables operacionales) del servidor.

> *Tiene sentido usar valores medios porque son servidores, muchas veces ejecutamos el mismo proceso todo el rato. Si fueran computadores personales NO.*
>
> *En general, se va a omitir el "medio" en las variables, es decir, quitar la barrita que indica que estamos hablando de media.*

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

Variables del servidor y de cada estación de servicio

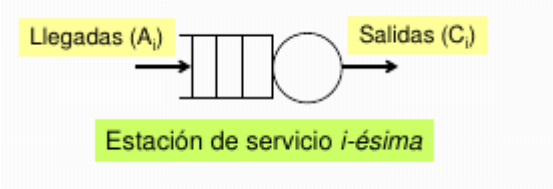
- El servidor contiene K estaciones de servicio (recursos o dispositivos).
- A todo el servidor en su globalidad lo denotamos como dispositivo "cero".



> *λo -- peticiones por segundo*
>
> *Xo -- peticiones completadas por segundo*

Variables operacionales básicas de cada estación de servicio

- Variable global temporal
  - T -- Duración del periodo de medida para el que se extrae el modelo.
- Variables operacionales básicas de la estación de servicio i-ésima, medidas durante el tiempo T:
  - Ai -- Número de trabajos solicitados a la estación (llegadas, ARRIVALS).
  - Bi -- Tiempo que el dispositivo ha estado en uso (=ocupado) (BUSY time).
  - Ci -- Número de trabajos completados por la estación (salidas, COMPLETIONS).

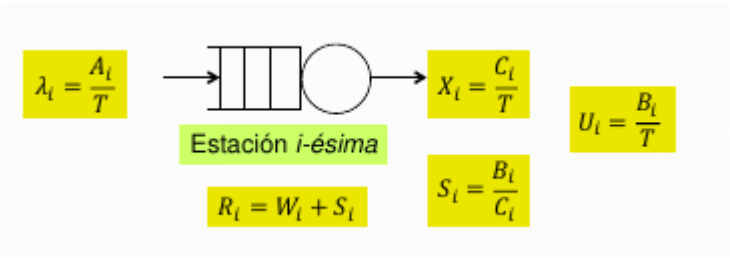


Variables operacionales deducidas

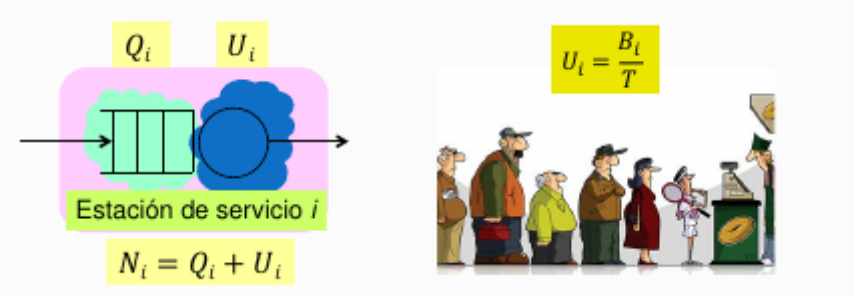
Se deben poder estimar a partir de las variables básicas:

- λi Tasa media de llegada (arrival rate) trabajos/segundo

    - Arrivals/Tiempo -- Ai/T
- Xi Productividad media (throughput) trabajos/segundo
    - Completions/Tiempo -- Ci/T
- Si Tiempo medio de servicio (service time) segundos [/trabajo]
    - Busy/Completions -- Bi/Ci
- Wi Tiempo medio de espera en cola (waiting time) segundos [/trabajo]
- Ri Tiempo medio de respuesta (response time) segundos [/trabajo]
- Ui Utilización media (utilization) sin unidades
    - Busy/Tiempo -- Ui/T
    - Ui = Xi * Si



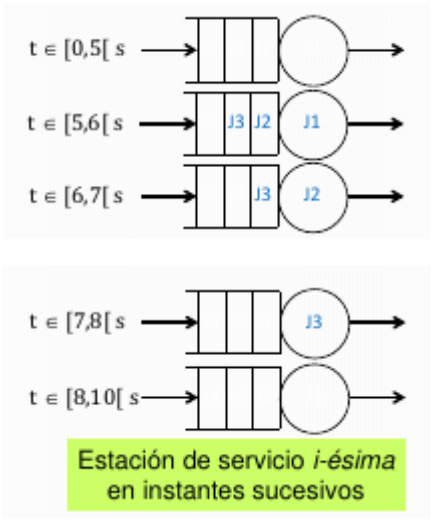
- Ni: Número medio de trabajos en la estación de servicio (cola más recurso).
- Qi : Número medio de trabajos en cola de espera (jobs in queue). Ui: Número medio de trabajos siendo servidos por el dispositivo,
    - $Ui = Ni - Qi$.
    - Coincide numéricamente con la utilización media = proporción de tiempo que el dispositivo ha estado en uso (busy) con respecto al intervalo total de medida (T) (como máximo 1 si Bi=T).



Ejercicio

> *Suponga que la estación de servicio i-ésima de una red de colas tiene un tiempo de servicio constante Si=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 Ai, Bi, Ci, λi, Xi, Ui, Qi, Ni*



Ai - Arrivals, trabajos en total: Ai = 3 trabajos

Bi - Busy, tiempo ocupado: Bi = 3s (porque el tiempo de servicio S es siempre 1 en este caso)

Ci - Completions, trabajos finalizados: Ci = 3 trabajos

λi - Tasa media de llegada: Ai/T = 3/10 = 0.3 trabajos por segundo

Xi - Productividad media: Ci/T = 3/10 = 0.3 trabajos por segundo

Ui - Utilización media: Bi/T = 3/10 = 0.3

Qi - Número medio de trabajos en cola de espera:

- Durante el intervalo 0-5 había 0 trabajos a la cola
- En el intervalo 5-6 habían 2 trabajos a la cola
- En el intervalo 6-7 había 1 trabajo a la cola
- En el intervalo 7-10 había 0 trabajos a la cola
    - (0*5s + 2*1s + 1*1s + 0*3s) / 10s = 0,3 trabajos

Ni - Número medio de trabajos en la estación de servicio (cola más recurso).

- Podemos calcularlo como la Qi pero contando también el trabajo en ejecución o Ni = Qi+Ui = 0.6 trabajos

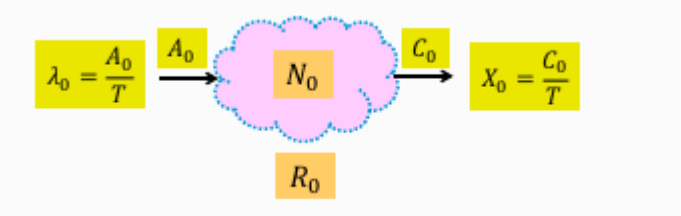
Variables operacionales de un servidor

Variables operacionales básicas de un servidor:

- A0 Número de trabajos solicitados al servidor (arrivals).
- C0 Número de trabajos completados por el servidor (completions).

Variables operacionales deducidas de un servidor:

- λ0 Tasa media de llegada al servidor (arrival rate).
- X0 Productividad media del servidor (throughput).
- N0 Número medio de trabajos en el servidor (#jobs) = N1+N2+...+NK.
- R0 Tiempo medio de respuesta del servidor (response time) ≡ tiempo que tarda, de media, el servidor en procesar una petición.



Razón de visita y demanda de servicio

- Razón media de visita Vi (VISIT ratio).
  - Representa la proporción entre el número de trabajos completados por el servidor y el número de trabajos completados por la estación de servicio i-ésima.
  - Nos indica el número de veces que, de media, un trabajo "visita" la estación de servicio i-ésima antes de abandonar el servidor.

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

- Demanda media de servicio Di (service DEMAND).
  - **Cantidad de tiempo** que, por término medio, el dispositivo de la estación de servicio i-ésima le ha dedicado a cada trabajo que abandona el servidor (= que ha sido procesado por completo por el servidor).

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

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

Ejercicio

> *Después de monitorizar el disco duro de un servidor web durante un periodo de 24horas, se sabe que ha estado en uso (=ocupado) un total de 6 horas.*
>
> *Asimismo, se han contabilizado durante ese periodo un total de 98590 peticiones de lectura/escritura al disco duro y un total de 98591 peticiones completadas. Se ha estimado que cada petición atendida por el servidor web ha requerido una media de 9,5 peticiones de lectura/escritura al disco duro. Calcule, para ese periodo de monitorización:*
>
> *a) La tasa media de llegada y la productividad media del disco duro.*
>
> *b) La utilización media del disco duro.*
>
> *c) El tiempo medio de servicio y la demanda media de servicio del disco duro.*
>
> *d) ¿Cuál es la productividad media del servidor web?*
>
> *Nota: Todas las variables que se usan en este tema son valores medios por lo que, de aquí en adelante, normalmente no se indicará de forma explícita la palabra "medio" al referirnos a ellas.*

- T = 24h
- Bdd = 6h
    - dd == disco duro
- Add = 98590 tr
- Cdd = 98591 tr
    - tr == trabajos, peticiones RW
    - C>A, es posible completar más trabajos de los que llegan porque puede que hubiera cosas pendientes antes de comenzar el monitoreo
- *"Se ha estimado que cada petición atendida por el servidor web ha requerido una media de 9,5 peticiones de lectura/escritura al disco duro."*

  9.5 relaciona C0 y Cdd
    - Si C0=10 >> Cdd=9.5*10
    - Si C0=100>>Cdd=950

  9.5 = Cdd/C0 = Vdd

  Es la razón de visita: por cada trabajo atendido, cuántas veces visita el dd.

a)

a.1) Tasa media de llegada: $\lambda dd = Add/T$

λdd = 4108 tr/24h = //matemagia, factor de conversión y tal... = 1.14 tr/s

a.2) Productividad/Throughput del dd: Xdd = Cdd/t

Prácticamente igual a lo anterior, diferencia nimia porque Add y Cdd solo se llevan 1 de diferencia... así que digamos que Xdd = 1.14 tr/s también

b) Utilización media del dd: Udd = Bdd/T = 0.25, 25% de utilización

c)

c.1) Tiempo medio de servicio: Sdd= Bdd/Cdd = 0.22s

c.2) Demanda media del servicio: Ddd= Bdd/C0 = Vdd* Sdd = 9.5*0.22 = 2.1s

d) Productividad del SERVER ojo:

- X0=C0/T
  - Vdd = 9.5 = Cdd/C0;

    C0 = Cdd/Vdd = 98591/9.5 = 10378tr

    X0 = 10378/24h = 432tr/h = 0.12tr/s

Una buena práctica es volver a traducirlo todo al lenguaje dado en el enunciado, y no dejarlo con el genérico "trabajos":

1. Tasa media de llegada, λdd:1.14 peticiones de RW al dd por segundo

   Throughput del dd, Xdd: 1.14 peticiones de RW al dd por segundo
2. Utilización media del dd, Udd: 25% de utilización
3. Tiempo medio de servicio, Sdd: 0.22s

   Demanda media del servicio del disco duro, Ddd: 2.1s
4. Productividad del servidor, X0: 0.12 peticiones de RW al dd por segundo

Leyes operacionales

El valor de todas las variables utilizadas en el análisis operacional dependerá del intervalo de observación T. **Existen, sin embargo, una serie de relaciones entre algunas variables operacionales que se mantienen válidas para cualquier intervalo de observación** y que no dependen de suposiciones sobre la distribución de los tiempos de servicio o de la forma en la que llegan los trabajos.

Estas relaciones se denominan leyes operacionales.

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

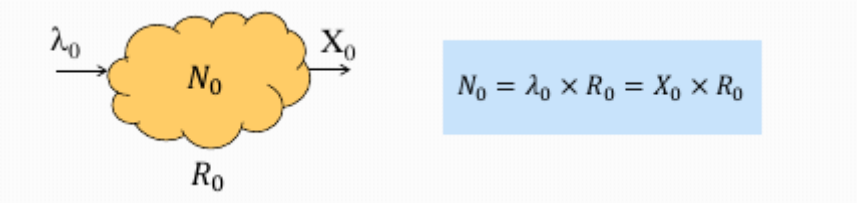
- El número de trabajos que completa el servidor coincide aproximadamente con los solicitados (C0 ≈ A0). Dicho de otra manera, la productividad media coincide aproximadamente con la tasa media de llegada (X0 ≈ λ0).
- El número de trabajos que completa cada estación de servicio coincide aproximadamente con los que se solicitan: (Ci ≈ Ai ->> Xi ≈ λi, ∀i=1...K).

> *Las gallinas que entran por las que salen*

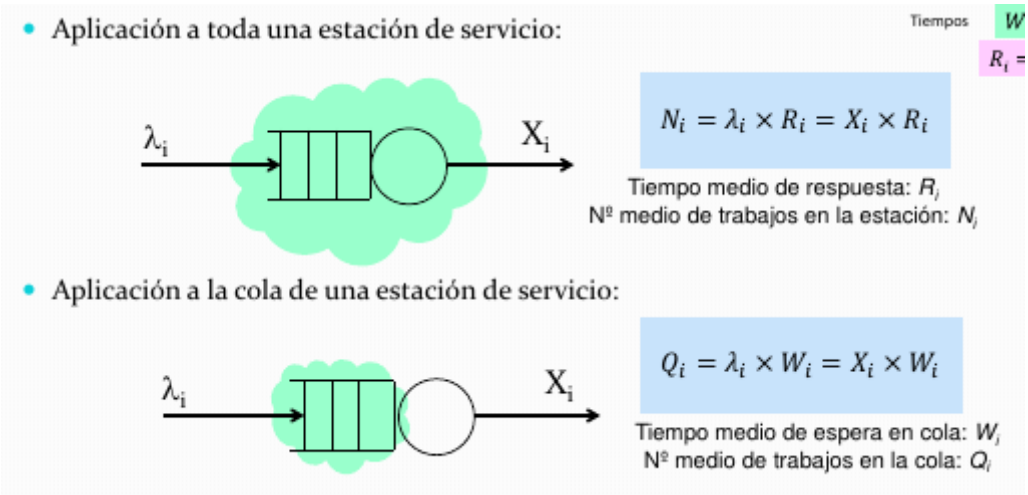
Ley de Little -- N0 = λ0*R0 = X0*R0

> *La más importante - Simulación en SWAD*

Sólo válida si está en equilibrio de flujo. Relaciona la productividad media X0 con el tiempo medio de respuesta R0.

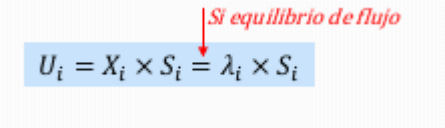


Puede aplicarse a cualquier estación de servicio, no sólo al servidor entero. La podemos aplicar a la estación completa, o a la cola de espera:



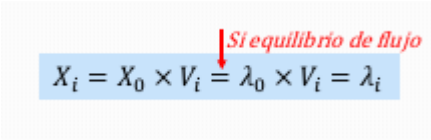
Ley de la utilización -- Ui = Xi*Si = λi*Si

Relaciona la utilización media de un dispositivo λi con el número de trabajos que completa por unidad de tiempo (=su productividad media, Xi) y el tiempo que le dedica, de media, a cada uno de ellos (=su tiempo de servicio medio, Si).



Ley del flujo forzado -- Xi = X0*Vi = λ0*Vi = λi

Las productividades X0, Xi (=flujos de salida) de cada estación de servicio tienen que ser proporcionales a la productividad global del servidor. La ley del flujo forzado relaciona la productividad del servidor con la de cada uno de los dispositivos que integran el mismo:

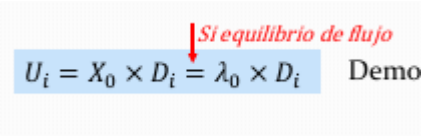


Recordemos que Vi=Ci/C0 = Xi/X0. Vi es el número de accesos del servidor a la estación de servicio por petición.

Relación Utilización-Demanda -- Ui = X0*Di = λ0*Di

Consecuencia de la ley de flujo forzado.

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



Recordemos que Di = Bi/C0 = Ui/X0. Di es la cantidad de tiempo que gasta la estación por cada trabajo que abandona el servidor.

Ejemplo

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

> *λ0 = 120 tr/min = 2tr/s*
>
> *Sdd = 30ms = 0.03s*
>
> *Wdd = 48ms-30ms = 18ms = 0.018s*
>
> *Xdd = 25tr/s*
>
> *Calcule:*
>
> *a) El número medio de peticiones de E/S en la cola de espera del disco duro.*
>
> *b) ¿Cuánto tiempo, de media, consumen los accesos al disco duro por cada consulta que se realiza al servidor?*

a) Qdd (trabajos medios en cola del dd)

LEY DE LITTLE aplicada a la cola de espera de una estación:

Qdd = λ0 * Wdd = **Xdd * Wdd**

Qdd = 25 * 0.018 = 0.45 tr en cola

b) Ddd (tiempo que pasa el dd en cada tr que le manda el servidor)

LEY DE UTILIZACIÓN

Udd = X0*Ddd = **λ0*Ddd**

Udd = Xdd * Sdd = 0.03 * 25 = 0.75

0.75 = 2 * Ddd; Ddd = 0.375 s/tr

Ley general del tiempo de respuesta

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

R0 = V1*R1 + V2*R2...

Ley del tiempo de respuesta interactivo

Un servidor en una red cerrada siempre está en equilibrio de flujo (siempre supondremos que el tamaño de las colas es suficientemente grande, en este caso, ≥ NT).

Al ser una red cerrada, el número total de trabajos (=clientes) en la red (NT=NZ+N0), es constante.

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

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

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

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

5.3. Límites optimistas del rendimiento

5.4. Técnicas de mejora del rendimiento

5.5. Algoritmos de resolución de redes de colas