

ISE-T5-Apuntes.pdf



marinamuca01



Ingeniería de Servidores



3º Grado en Ingeniería Informática



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

**QUIERES
CONSEGUIR
ISE??**

→ TRÁENOS A TU
CRUSH DE APUNTES

ANTES DE QUE LOS QUEME



WUOLAH

**QUIERES
CONSEGUIR
15€??**

TRÁENOS A TU
CRUSH DE APUNTES
ANTES DE QUE
LOS QUEME



TEMA 5

Optimización del rendimiento de un servidor mediante análisis operacional

Introducción: Redes de colas de espera

Modelo de un sistema informático

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

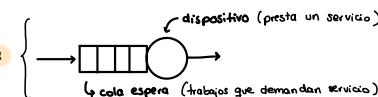
Dispositivos (resources) → núcleos lógicos, unidades almacenamiento, ...
Trabajos (jobs) → procesos, accesos, peticiones...

Normalmente 1 dispositivo (recurso) sólo puede ser usado por 1 trabajo a la vez. El resto tienen que esperar.

Modelos basados en redes de colas

red colas formada por conjunto estaciones servicio conectadas entre sí.

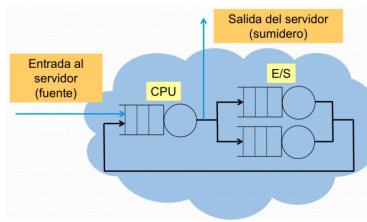
Estación servicio = dispositivo + cola espera



Modelo de servidor central

Red colas + utilizada para representar comportamiento básico de programas en un server para extraer info sobre su rendimiento.

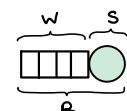
- Trabajo "llega" al server → comienza utilizando CPU.
- Tras "abandonar" el CPU puede finalizar realizar acceso a una ud. E/S.
- Tras E/S vuelve a "visitar" la CPU



si
consigues
que suba
apuntes, te
llevas 15€ +
5 Wuolah
Coins para
los sorteos

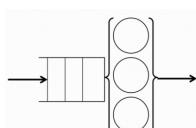
Algunas variables que caracterizan a un trabajo en una estación de servicio en un instante

- $T_{espera} \equiv W$ (waiting time) → t desde que trabajo solicita uso recurso (se pone en cola) hasta que empieza a utilizarlo.
- $T_{servicio} \equiv S$ (service time) → t desde que trabajo accede al recurso físico hasta que lo libera. ($= t$ en procesar trabajo × dispositivo)
- $T_{respuesta} \equiv R$ (response time) → Suma $W + S$. $R = W + S$

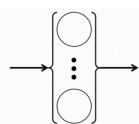


Recopilando estas medidas para múltiples trabajos, obtenemos distribuciones de probabilidad que caracterizan a esa estación de servicio.

Estaciones con más de un servidor



3 dispositivos
idénticos compartiendo
la misma cola de espera



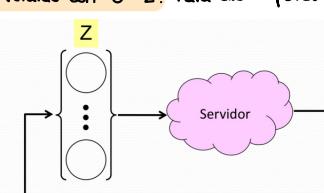
Infinitos dispositivos:
no hay espera en cola.
 $R=S$.
Estación tipo retardo

Son capaces de atender a + de un trabajo en paralelo

Tiempo de reflexión (Z, think time)

$Z \rightarrow t$ que requiere usuario antes de volver a lanzar una petición al servidor tras la respuesta de éste.

Modelo → estación servicio tipo retardo con $S = Z$. Para ello hipótesis adicional = cada usuario envía 1 trabajo al servidor.

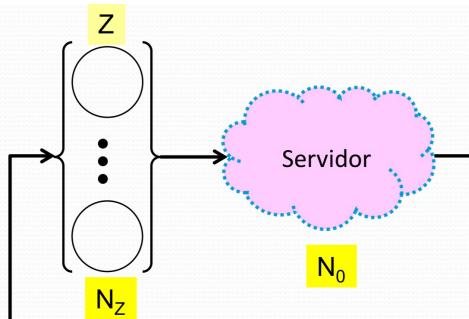
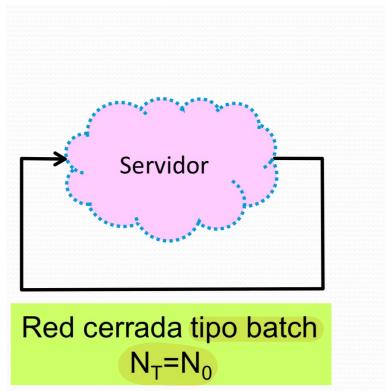


WUOLAH

Redes de colas cerradas

$N_T = \text{Nº constante de trabajos que van recirculando por la red}$

Dependiendo de si hay o no interacción con usuarios se distingue entre
 tipo Batch (por lotes)
 redes interactivas.



Siempre supondremos 1 usuario = 1 trabajo

n° trabajos = n° clientes \leftrightarrow n° trabajos en reflexión \rightarrow n° trabajos en servidor (esperando a que usuarios vuelvan a introducirlos)

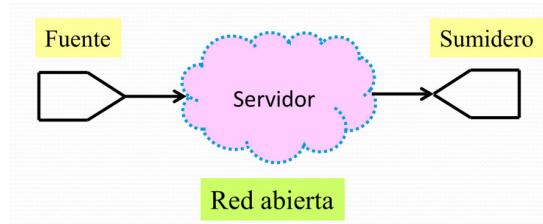
Redes de colas abiertas

Trabajos llegan a través de fuente externa (no controlada).

Tras ser procesados, salen de ella a través de 1 o + sumideros.

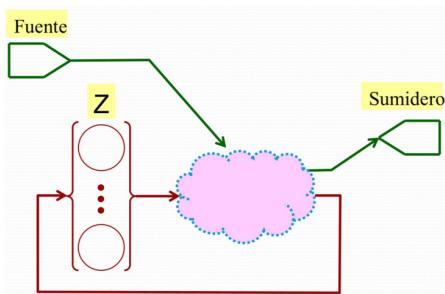
No existe realimentación entre sumidero y fuente.

Nº trabajos en el servidor (N_s) puede variar con el tiempo.



Redes mixtas

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



Variables y leyes operacionales

NOTA: A partir de ahora usaremos la misma variable para simbolizar la media: $\bar{z} = z$

Análisis operacional

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



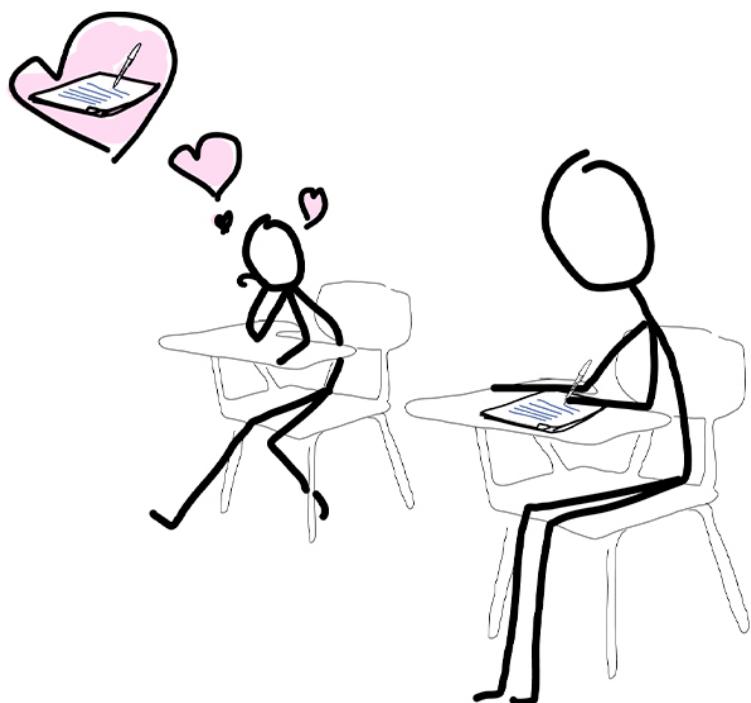
Nos proporcionará relaciones generales entre las variables operacionales (leyes operacionales)

Nos permite calcular prestaciones del servidor para los casos de baja y alta carga por medio de cálculos muy sencillos.

Nos permite evaluar los efectos en el rendimiento de diferentes modificaciones en los recursos del servidor.

QUIERES
CONSEGUIR
15€ ??

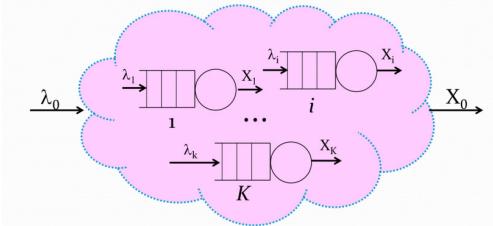
TRAÉNOS A TU
CRUSH DE APUNTES
ANTES DE QUE
LOS QUEME

si consigues que suba apuntes, te llevas 15€ + 5
Wuolah Coins para los sorteos

WUOLAH

Variables del servidor y de cada estación de servicio



Servidor contiene K estaciones de servicio
Servidor completo \equiv dispositivo "cero".

ESTACIÓN DE SERVICIO

Variable global temporal:

$T \rightarrow$ duración periodo de medida para el que se extrae el modelo.

Variables operacionales básicas medidas durante el tiempo T :

$A_i \rightarrow$ nº trabajos solicitados a la estación (llegadas, arrivals)

$B_i \rightarrow$ tiempo que dispositivo ocupado (busy time)

$C_i \rightarrow$ Nº trabajos completados por la estación (salidas, completions)

Variables operacionales deducidas:

$\lambda_i \rightarrow$ tasa media de llegada (arrival rate)

$$\lambda_i = \frac{A_i}{T}$$

tr/s

$X_i \rightarrow$ productividad media (throughput)

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

tr/s

$S_i \rightarrow$ tiempo de servicio (service time)

$$S_i = \frac{B_i}{C_i}$$

s/[tr]

$W_i \rightarrow$ tiempo medio de espera en cola (waiting time)

s/[tr]

$R_i \rightarrow$ tiempo medio de respuesta (response time)

$$R_i = W_i + S_i$$

s/[tr]

$U_i \rightarrow$ Utilización media (utilization)

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

sin ud.

$N_i \rightarrow$ nº medio trabajos en la estación de servicio (cola + recurso)

$$N_i = Q_i + U_i$$

$Q_i \rightarrow$ nº medio trabajos en cola de espera (jobs in queue)

$U_i \rightarrow$ nº medio trabajos siendo servidos por el dispositivo

$$U_i = N_i - Q_i$$

\hookrightarrow coincide numéricamente con utilización media \equiv proporción tiempo que dispositivo ha estado busy con respecto al total de medida (como máx 1 si $B_i = T$)

SERVIDOR ($\text{No hay } B_0, W_0, U_0$)

Variable operacionales básicas de un servidor:

$A_0 \rightarrow$ nº trabajos solicitados al servidor (arrivals)

$C_0 \rightarrow$ Nº trabajos completados por servidor (completions)

Variables operacionales deducidas de un servidor:

$\lambda_0 \rightarrow$ tasa media de llegada al servidor (arrival rate)

$$\lambda_0 = \frac{A_0}{T}$$

tr/s

$X_0 \rightarrow$ productividad media del servidor (throughput)

$$X_0 = \frac{C_0}{T}$$

tr/s

$N_0 \rightarrow$ nº medio trabajos en servidor (#jobs) $= N_1 + N_2 + \dots + N_K \Rightarrow N_0 = \sum_{i=1}^K N_i$

$R_0 \rightarrow$ tiempo medio de respuesta del servidor (response time) \equiv tiempo que tarda de media el servidor en procesar una petición

Razón de visita y demanda de servicio

$V_i \rightarrow$ Razón media de visita (visit ratio) \equiv proporción entre nº trabajos completados por el servidor y el nº de trabajos completados por la estación i -ésima. \Rightarrow Nº medio veces que un trabajo visita la estación i antes de abandonar el servidor.

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

$D_i \rightarrow$ Demanda media de servicio (es un tiempo) \equiv Cantidad de tiempo que, por término medio, el dispositivo de la estación i -ésima le ha dedicado a cada trabajo que abandona el servidor (= procesado por completo).

\hookleftarrow no tiene en cuenta la espera de un trabajo en su cola.

QUIERES CONSEGUIR 15€??

TRÁENOS A TU
CRUSH DE APUNTES
ANTES DE QUE LOS QUEME



Leyes operacionales

Valor de variables operacionales depende de intervalo de observación T.

• Relaciones entre variables operacionales que se mantienen válidas para cualquier intervalo de observación y no dependen de suposiciones sobre distribución de los tiempos de servicio o forma en la que llegan los trabajos. ⇒ Leyes Operacionales

Son + útiles si se cumple hipótesis equilibrio de flujo = Si se recoge un intervalo T suficientemente largo se cumple:

- nº trabajos completados x servidor ≈ a los solicitados ⇒ $C_0 \approx A_0 \Rightarrow X_0 \approx \lambda_0$ (prod. media ≈ tasa media llegada)
- nº trabajos completados x estación ≈ a los solicitados ⇒ $C_i \approx A_i \Rightarrow X_i \approx \lambda_i \quad \forall i, 1 \dots k$

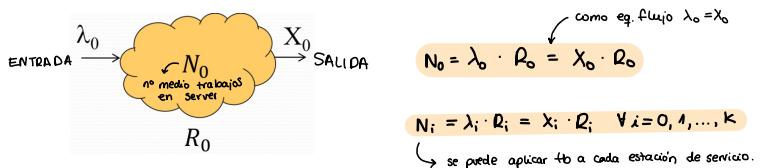
En resumen:

$$C_i \approx A_i \quad \forall i, 1 \dots k \quad \xleftarrow[\text{serv. est. servicio}]{\text{server}} X_i \approx \lambda_i \quad \forall i, 1 \dots k$$

LEY DE LITTLE

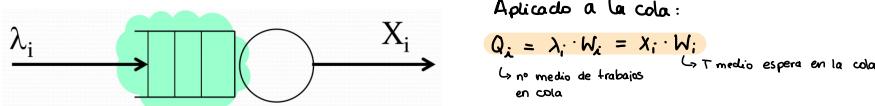
El nº medio de clientes a largo plazo en un sistema estable es igual a la tasa de llegadas a largo plazo por el tiempo medio que un cliente pasa en el sistema.

Aplicado a server → relaciona las variables más importantes que reflejan rendimiento del servidor: productividad (X_0) y respuesta (R_0).



Solo es válida si equilibrio de flujo

También se puede aplicar a cada uno de los diferentes sub-niveles de una estación de servicio



LEY DE LA UTILIZACIÓN

Relaciona la utilización de un dispositivo con el nº de trabajos que es capaz de realizar por vd. de tiempo y el tiempo que le dedica a cada uno.

$$S_i = \frac{B_i}{C_i} = \frac{B_i / T}{C_i / T} = \frac{U_i}{X_i} \Rightarrow U_i = X_i \cdot S_i = \lambda_i \cdot S_i$$

dividimos por T arriba y abajo
si eq. flujo

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

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

si $X_i \leq \frac{1}{S_i} \Rightarrow i \equiv \text{cuello de botella}$

LEY DE FLUJO FORZADO

Las productividades de cada estación de servicio tienen que ser proporcionales a la productividad global del servidor.

Ley flujo reforzado relaciona la productividad del server con la de cada uno de los dispositivos que lo integran:

$$V_i = \frac{C_i}{C_0} = \frac{C_i / T}{C_0 / T} = \frac{X_i}{X_0} \Rightarrow X_i = X_0 \cdot V_i = \lambda_0 \cdot V_i = \lambda_i$$

dividimos por T
si equilibrio flujo

RELACIÓN UTILIZACIÓN - DEMANDA

Consecuencia ley flujo forzado = utilizaciones de cada dispositivo proporcionales a demanda servicio

$$D_i = \frac{B_i}{C_0} = \frac{\text{divido por } \bar{T}}{C_0/\bar{T}} = \frac{U_i}{X_0} \Rightarrow U_i = X_0 \cdot D_i \quad \begin{cases} \text{si eq. fluiro} \\ \downarrow U_i \leq 1 \Rightarrow X_0 \cdot D_i \leq 1 \Rightarrow X_0 \leq \frac{1}{D_i} \end{cases}$$

LEY GENERAL DEL TIEMPO DE RESPUESTA

T medio respuesta que experimenta de media una petición a un servidor en eq. flujo se puede calcular teniendo en cuenta que cada una de ellas ha tenido que visitar V_i veces al dispositivo i , requiriendo en cada visita una media de R_i segs.

$$R_0 = V_1 \cdot R_1 + V_2 \cdot R_2 + \dots + V_k \cdot R_k = \sum_{i=1}^k V_i \cdot R_i$$

Demostración:

$$N_0 = N_1 + N_2 + \dots + N_K \Rightarrow X_0 \cdot R_0 = X_1 \cdot R_1 + X_2 \cdot R_2 + \dots + X_K \cdot R_K \Rightarrow X_0 \cdot R_0 = X_0 \cdot V_1 \cdot R_1 + X_0 \cdot V_2 \cdot R_2 + \dots + X_0 \cdot V_K \cdot R_K \quad (\text{dividimos por } X_0)$$

Ley little

flujo forzado

LEY DEL TIEMPO DE RESPUESTA INTERACTIVO

Una red cerrada siempre está en eq. flujo (si tamaño de colas es $\geq N_T$)

Como red cerrada $N_T = N_0 + N_Z$ cte.

Aplicamos ley little a diversas partes de la red:

- Clientes en reflexión: $N_2 = x_0 \cdot Z$
 - Server: $N_0 = x_0 \cdot R_0$

$$\text{Server : } N_0 = x_0 \cdot R_0$$

desprezo R_0

$$N_T = N_Z + N_0 \Rightarrow N_T = x_0 \cdot Z + x_0 \cdot R_0 = x_0 \cdot (Z + R_0) \Rightarrow R_0 = \frac{N_T}{x_0} - Z$$

-Límites optimistas del rendimiento

Limitaciones en rendimiento: cuello de botella

Todo server → limitación en rendimiento

$$V_i \cdot S_i = D_i = \frac{B_i}{C_0} = \frac{B_i / T}{C_0 / T} = \frac{U_i}{X_0} \Rightarrow U_i = X_0 \cdot D_i$$

Elemento limitador = cuello botella (puede haber + de 1). Depende del serv y de la carga.

Única forma mejorar las prestaciones de un servidor → ACTUAR SOBRE CUELLO BOTELLA

IDENTIFICAR CUELLO BOTELLA

Cuello botella = disp que primero se satura ($U_i=1$) cuando aumenta λ_0 .

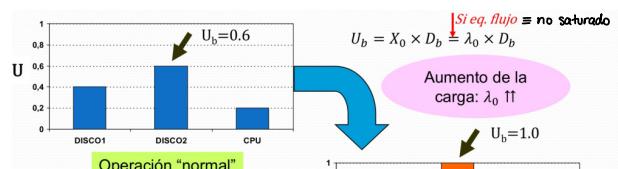
Como U_i es proporcional a $D_i \Rightarrow$ cuello botella = dispositivo con $> D_i$ o $> U_i$

No hace falta llevar servir al límite para identificarlo.

$D_i = V_i \cdot S_i \rightarrow$ localización cuello botella depende de rapidez dispositivos (S_i) y tipo carga (V_i)

$$b = \text{indice disp. cuello botella.} \Rightarrow D_b = \max_{i=1 \dots k} \{D_i\} = V_b \cdot S_b$$

$$V_b = \max_{i=1 \dots k} \{V_i\} = x_0 \cdot D_b$$



SATURACIÓN DEL SERVIDOR

Servir saturando cuando se satura el cuello botella. Es el 1º en alcanzar $U_i = 1$.

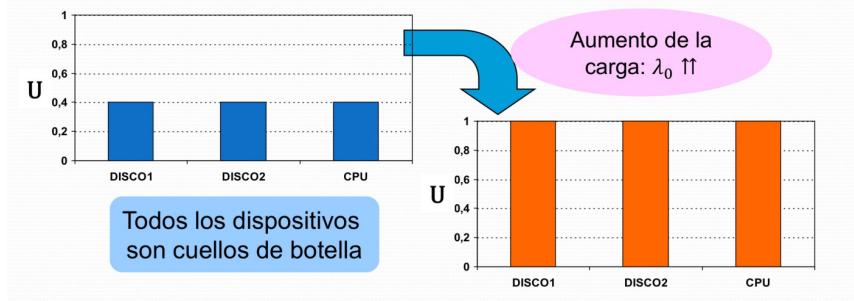
$$\text{En saturación, cuello botella = máximo de su productividad.} \Rightarrow 1 = U_b = X_b \cdot S_b \Rightarrow X_b = \frac{1}{S_b}$$



SEVIDOR EQUILIBRADO

Todos los dispositivos, de media, tienen la misma demanda de servicio y utilización

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

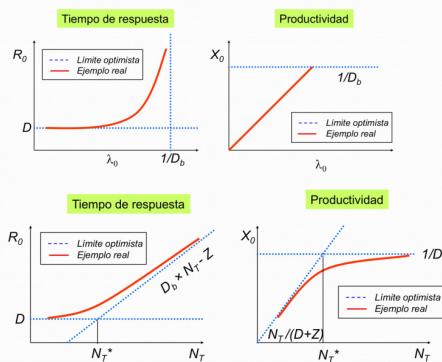


LIMITE DE RENDIMIENTO DEL SERVIDOR

Estimar prestaciones límite de un servidor (R_0, X_0) en los casos extremos de altas y bajas cargas. \Rightarrow Cotas superior e inferior \equiv límites optimistas.
¿ X_0^{\max} ? ¿ R_0^{\min} ?

LOCALIZACIÓN LÍMITES DE RENDIMIENTO

- ReDES ABIERTAS
- ReDES CERRADAS



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

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

$$R_0 \geq \max \{ D, D_b \cdot N - Z \}$$

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

REDES ABIERTAS (Demostración)

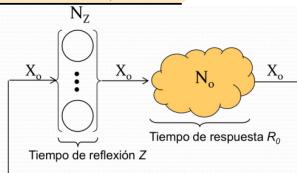
$$X_0^{\max} \equiv \text{tasa llegada que satura el server} \quad U_b = X_0 \cdot D_b = \lambda_b \cdot D_b \quad \text{Si } U_b = 1 \Rightarrow X_0^{\max} = \frac{1}{D_b}$$

mayor tasa llegada \rightarrow provoca aumento cola cuello botella hasta desbordamiento final \Rightarrow deja de cumplirse eq. flujo $\Rightarrow R_0$ crece sin control

$R_0^{\min} \equiv$ cuando trabajo llega a server sin que haya otros $W_i = 0 \quad \forall i = 1 \dots k$

$$R_0 = \sum_{i=1}^k v_i \cdot R_i = \sum_{i=1}^k v_i (W_i + S_i) \Rightarrow R_0^{\min} = \sum_{i=1}^k v_i (W_i^0 + S_i) = \sum_{i=1}^k v_i \cdot S_i = \sum_{i=1}^k D_i \equiv D$$

REDES CERRADAS (Demostración)



Ley de Little a la red completa ($N_T = N_o + N_z$):

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

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

a) Valores de carga altos (N_T grande)

$$X_0^{\max} \equiv \text{cuello botella cerca de saturación} \Rightarrow U_b = X_0 \cdot D_b \Rightarrow \text{Si } U_b \rightarrow 1 \Rightarrow X_0 \rightarrow X_0^{\max} = \frac{1}{D_b}$$

$$R_0^{\text{optim.}} \text{ a partir de } X_0^{\max} \Rightarrow \text{reemplazar en ley little} \Rightarrow R_0 \rightarrow \left(\frac{N_T}{X_0^{\max}} \right) - Z = D_b \cdot N_T - Z$$

b) Valores carga bajos (N_T pequeño)

$$R_0^{\min} \equiv \text{trabajos encuentran dispositivos sin ocupar. } (W_i = 0) \quad R_0 = \sum_{i=1}^k v_i \cdot R_i = \sum_{i=1}^k v_i (W_i + S_i) \Rightarrow R_0^{\min} = \sum_{i=1}^k v_i \cdot S_i = \sum_{i=1}^k D_i \equiv D$$

$$X_0^{\text{opt.}} \text{ a partir de } R_0^{\min} \Rightarrow \text{reemplazar en ley little} \quad X_0 \rightarrow \frac{N_T}{R_0^{\min} + Z} = \frac{N_T}{D+Z}$$

**QUIERES
CONSEGUIR
15€??**

TRÁENOS A TU
CRUSH DE APUNTES
ANTES DE QUE
LOS QUEME



PUNTO TEÓRICO DE SATURACIÓN (KNEE POINT)

Valor de N_T donde las asíntotas coinciden (N_T^*)

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

Propiedades $\begin{cases} \text{Para } N_T > N_T^* \rightarrow \text{límites asíntoticos impuestos por cuello botella.} \\ \text{A partir de } N_T^* \text{ ya no se puede conseguir el } R_o^{\min} \text{ ya que se empiezan a formar colas de espera en el cuello de botella.} \\ \text{Nº ideal de trabajos en red ya que teóricamente para } N_T = N_T^* \text{ se podría conseguir } x_0^{\max} \text{ y } R_o^{\min} \text{ absolutos del servidor.} \\ \hookrightarrow \text{en la práctica no se consigue de forma simultánea} \end{cases}$

$$N_T^* = x_0^{\max} (R_o^{\min} + Z) = \frac{D+Z}{D_b}$$



Técnicas de mejora de prestaciones

mejora significativa → Activar sobre cuello botella $\begin{cases} \text{Sintonización / ajuste} \\ \text{Actualización y/o ampliación.} \end{cases}$

Sintonización o ajuste (tunning)

Optimizar funcionamiento componentes existentes $\begin{cases} \text{HW} \rightarrow \text{parámetros planta base (frecuencias, voltajes, etc.)} \\ \text{APPS} \rightarrow \text{ficheros config., profilers.} \\ \text{S.O.} \rightarrow \text{políticas gestión procesos, memoria, almacenamiento y red.} \end{cases}$

Inconvenientes $\begin{cases} \text{Alteración fiabilidad} \\ \text{Conocer bien funcionamiento componentes HW, la APP y el S.O.} \\ \text{Realizar tests estadísticos para ver qué factores influyen en prestaciones.} \end{cases}$

Actualización y/o ampliación

Reemplazar dispositivos por otros más rápidos \Rightarrow disminuye S_i
Añadir dispositivos para poder realizar más tareas en paralelo \Rightarrow disminuye V_i

Problemas $\begin{cases} \text{facilidad servir para dejarse actualizar (extensibilidad / escalabilidad)} \\ \text{Compatibilidad nuevos elementos con existentes.} \end{cases}$

Algoritmos de resolución de modelos de redes de colas

Metodología para resolver redes de colas. Suponemos conocido:

- Nº estaciones servicio (k)
- Para cada estación $\rightarrow V_i$ y S_i
- Si red abierta $\rightarrow \lambda_0$
- Si red cerrada $\rightarrow N_T$ y Z

Redes abiertas: hipótesis de independencia en llegada de trabajos

En eq. flujo \rightarrow Suponemos que trabajo llega en \neq instante que anterior. \rightarrow Se consigue suponiendo que todas las distribuciones de probabilidad se rigen por una de tipo exponencial $P(x) = \lambda e^{-\lambda x}$ de Poisson.

En ese caso se puede demostrar que cuando un trabajo llega a la estación i -ésima tiene que esperar a que se procesen todos los N_i que hay en ese momento en la estación (uno comenzando a ser servido, el resto esperando).

$W_i = N_i \cdot S_i$ → **NO ES LEY OPERACIONAL!** Solo válido si hipótesis de independencia de carga de trabajo = True

Por tanto $R_i = W_i + S_i = N_i \cdot S_i + S_i$ Aplicando Ley de Little $\rightarrow N_i = X_i \cdot R_i$

EXAMEN

$$R_i \text{ en función de } V_i, S_i, \lambda_0 \quad \text{factor común } R_i$$

$$R_i = X_i \cdot R_i \cdot S_i + S_i \Rightarrow 1 = X_i \cdot S_i + S_i / R_i \Rightarrow R_i = \frac{S_i}{1 - X_i \cdot S_i} = \frac{S_i}{1 - V_i} = \frac{S_i}{1 - \lambda_0 \cdot D_i} = \frac{S_i}{1 - \lambda_0 \cdot V_i \cdot S_i}$$

Si la necesitamos
hay que demostrarla

$$\hookrightarrow D_i = \frac{S_i}{C_0} = \frac{B_i / T}{C_0 / T} = \frac{U_i}{\lambda_0} \Rightarrow V_i = \lambda_0 \cdot D_i$$

Redes abiertas: algoritmo resolución

Suponemos conocidos $\lambda_0 (= \lambda_0)$, V_i , S_i $\forall i=1..K$ y servidor en equilibrio flujo.

① Calcular $D_i = V_i \cdot S_i \quad \forall i=1..K$

② Calcular R_i usando hipótesis $W_i = N_i \cdot S_i \Rightarrow R_i = \frac{S_i}{1 - X_i \cdot S_i} = \frac{S_i}{1 - V_i} = \frac{S_i}{1 - \lambda_0 \cdot D_i}$

③ Calculamos R_0 $R_0 = \sum_{i=1}^K V_i \cdot R_i = \sum_{i=1}^K \frac{V_i \cdot S_i}{1 - \lambda_0 \cdot D_i} = \sum_{i=1}^K \frac{D_i}{1 - \lambda_0 \cdot D_i}$

Resto variables se calculan usando sus expresiones habituales.

Redes cerradas: algoritmo resolución

Suponemos conocidos V_i, S_i, N_T y Z .

Método → ir resolviendo la red para valores incrementales del nº de trabajos hasta alcanzar N_T : $n_T = 1..N_T$.

Notación → $N_i(n_T) \equiv$ nº trabajos en estación de servicio i ésima si en la red hubiese n_T trabajos. Idem para $R_i(n_T)$ y $X_i(n_T)$

Hipótesis de independencia en la llegada de trabajos para redes cerradas: $W_i(n_T) = N_i(n_T-1) \cdot S_i \Rightarrow R_i(n_T) = (N_i(n_T-1) + 1) \cdot S_i$