Modelos de sistemas informáticos

Por R. Puigjaner

Existen numerosas circunstancias en que la evaluación del rendimiento de un sistema informático debe efectuarse mediante la construcción de un modelo. Ello será así siempre que alguno de los elementos hardware o software que componen el sistema no exista, como son los casos de instalación de un nuevo sistema o de reconfiguración (cambio de configuración hardware, introducción de una nueva aplicación, etc.) de uno existente.

No se pretende en este artículo hacer mención de todas las técnicas de construcción y tratamiento de modelos de sistemas informáticos sino exponer tan sólo algunos de los métodos que han demostrado ser especialmente útiles y eficaces. Así, entre otros, dejaremos fuera del ámbito de este artículo los modelos deterministas, dentro de cuyo dominio caen las máquinas de Türing, las redes de Petri, los modelos de grafos de los programas, etc., la física del software (software physics) que ha desarrollado K. Kolence, los métodos de descomposición de P. J. Courtois, el análisis operacional de J. P. Buzen y los métodos de difusión H. Kobayashi y E. Gelenbe. Centraremos nuestro tema en los modelos probabilísticos basados en la teoría de colas y en las redes de colas.

1. INTRODUCCION A LA TEORIA DE COLAS

¿Qué informático no se ha encontrado alguna vez *esperando* a que su listado saliera por la impresora ocupada en emitir los interminables resultados de una explotación o de otro programador?

¿Quién que se haya sentado en un terminal para trabajar en tiempo compartido en un ordenador no ha tenido que *esperar* a que su programa entrara en memoria o utilizara suficientemente la CPU?

¿Qué jefe de explotación no se ha visto en la necesidad de hacer *esperar* programas que no podían ejecutarse por falta de cintas o discos?

Todas estas *esperas* provocan colas dentro o fuera del sistema informático y aun antes de que existieran los ordenadores se había desarrollado la teoría de colas para intentar analizar su comportamiento.

La *teoría de colas* desarrollada con anterioridad y que en sus primeros niveles no es más que un caso particular de un proceso de Markov (que es aquel en que su estado actual depende sólo del anterior y de las entradas que ha recibido y no de los demás estados precedentes; es decir es un proceso sin memoria) se ha utilizado en la construcción de modelos de sistemas informáticos debido a que realmente en ellos aparecen numerosos subsistemas en que para obtener servicio de un elemento es preciso colocarse en cola para poder alcanzarlo, como es el caso de la CPU, discos, canales, etc.

1.1 Componentes de un modelo de colas

Los componentes básicos de un modelo de colas son las *estaciones de servicio*, las *colas* y los *manantiales*. Las estaciones de servicio se usan generalmente para modelizar los recursos solicitados por los trabajos que sometemos a un ordenador. Los trabajos se generan en los manantiales o existen en el modelo desde su creación. Cada estación de servicio puede atender sólo un número limitado de trabajos al mismo tiempo, lo que se conoce como el *número de canales* de la estación de servicio. Estos trabajos cuando encuentran la estación de servicio ocupada esperan hasta que les llega el turno. Cada estación de servicio tiene por lo menos una cola y con frecuencia el concepto estación engloba también la cola. Un trabajo generalmente requiere la atención de una estación de servicio durante un cierto tiempo denominado *tiempo de servicio* y entra en la estación de servicio en un instante denominado *tiempo de llegada* del trabajo.

1.1.1 Características del manantial

a) Su tipo finito o infinito; si un manantial es finito, el número máximo de trabajos generados por él, que un modelo puede contener, tiene una cota finita.

b) La distribución de los intervalos entre la generación de dos trabajos sucesivos.

c) Las demandas de cada trabajo de los servicios de cada estación de servicio del modelo; si las demandas de un determinado tipo de servicio están idénticamente distribuidas para todos los trabajos, es natural considerarlas como una característica de la correspondiente estación de servicio en vez de como una del manantial; sin embargo, puesto que representan demandas de recursos hechas por los trabajos, es más correcto pensar en ellas como características del manantial.

1.1.2 Características de la estación de servicio

a) El número y la capacidad de sus colas; la *capacidad de una cola* es el máximo número de trabajos que puede contener.

b) El número de canales de servicio de cada una.

c) La velocidad de los servidores; es decir el número medio de trabajos que puede atender por unidad de tiempo; también se acostumbra a utilizar su inversa, o sea el tiempo medio de servicio; cuando la velocidad de servicio de la estación es fija y las demandas de servicio están idénticamente distribuidas para todos los trabajos, podemos considerar la distribución de los tiempos de servicio entre las características de la estación de servicio.

d) La disciplina de servicio, que especifica bajo qué condiciones las estaciones de servicio terminan su ser-

vicio a un trabajo, como se selecciona el siguiente trabajo que debe ser servido a partir de la cola de la estación de servicio, y lo que hace un trabajo servido incompletamente.

1.1.3 Interconexiones

El modelo de colas se completa con las interconexiones entre las estaciones de servicio que especifican los caminos existentes entre ellas.

1.2 Casos básicos de sistemas de colas

Los párrafos anteriores nos dejan entrever la gran variedad de casos que pueden surgir combinando las distintas posibilidades consideradas. No obstante, no todos ellos permiten una solución analítica simple. Entre los que teniéndola son de aplicación a los modelos que estamos considerando, se halla el que cumple las siguientes hipótesis:

- manantial infinito.
- los intervalos de tiempo entre dos llegadas consecutivas están distribuidos según una ley exponencial de valor medio $t_m = \frac{1}{\lambda}$, (λ, número de llegadas por unidad de tiempo), cuya probabilidad tiene por expresión

  Prob. $(t \leqslant T) = 1 - e^{T/t_m}$

  lo cual es equivalente a que las llegadas se produzcan según un proceso aleatorio de Poisson, que significa que en un instante dado no pueden producirse dos llegadas, que el número medio de llegadas por unidad de tiempo es λ y que la probabilidad de que por unidad de tiempo se produzcan n llegadas es

  $$p(n) = \frac{e^{-\lambda}\lambda^n}{n!}$$

- el tiempo de servicio es también una variable aleatoria que puede ser desde constante hasta exponencial e hiperexponencial (que puede sustituirse por varias exponenciales en serie) pasando por las leyes de Erlang, cuyo grado de aleatoriedad depende de la relación entre la media y la desviación tipo que se mide generalmente por

  $$c = \frac{\sigma}{m} \quad \text{o} \quad E = \left(\frac{m}{\sigma}\right)^2$$

  y cuyo valor medio es $\frac{1}{\mu}$, siendo μ entonces el número medio de trabajos que puede atender la estación de servicio por unidad de tiempo.

- la disciplina de la cola es FIFO
- para que la cola alcance un régimen estacionario estable, y no tienda a una longitud infinita es preciso que el tiempo medio servicio sea inferior al tiempo medio entre llegadas, o lo que es lo mismo $\mu > \lambda$.
- un solo canal de servicio (fig. 1.1).

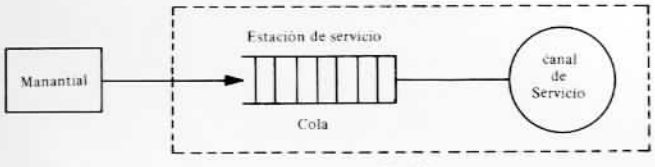


Fig. 1.1.

En estas circunstancias son aplicables las fórmulas de Khintchine-Pollaczeck, en las que interviene el factor de utilización ρ de la estación de servicio y que podemos definir como

$$\rho = \frac{\text{tiempo total con la estación de servicio ocupada}}{\text{tiempo total disponible}}, \text{ o}$$

$$\rho = \frac{\text{carga total de la estación de servicio}}{\text{carga máxima posible de la estación de servicio}}, \text{ o}$$

ρ = número medio de llegadas (λ) x tiempo medio de servicio $\frac{1}{\mu}$ y que representa la probabilidad de que la estación de servicio esté ocupada.

Número medio de trabajos en la estación de servicio:

$$\bar{N} = \rho + \frac{\rho^2}{2(1-\rho)}\left(1 + \frac{1}{E}\right)$$

Tiempo medio de permanencia en la estación de servicio:

$$\bar{t} = \frac{1}{\mu}\left[1 + \frac{\rho}{2(1-\rho)}\left(1 + \frac{1}{E}\right)\right]$$

Estas fórmulas se simplifican cuando el tiempo de servicio es constante, $\sigma = 0$, y por lo tanto $E = \infty$

$$\bar{N} = \frac{(2-\rho)\rho}{2(1-\rho)}$$

$$\bar{t} = \frac{(2-\rho)}{2(1-\rho)}\,\frac{1}{\mu}$$

y cuando la distribución de tiempos de servicio es exponencial, $\sigma = \frac{1}{\mu}$, y, por lo tanto $E = 1$

$$\bar{N} = \frac{\rho}{1-\rho}$$

$$\bar{t} = \frac{1}{1-\rho}\,\frac{1}{\mu}$$

Con frecuencia estas fórmulas se encuentran tabuladas o representadas gráficamente. Asimismo existen fórmulas para determinar las variancias asociadas a los valores medios calculados.

2. MODELOS INDIVIDUALES DE LOS SUBSISTEMAS

La modelización de sistemas informáticos puede enfocarse estableciendo modelos individuales de cada uno de los subsistemas o bien estableciendo un modelo global de todo el sistema. En este apartado analizaremos los modelos individuales de cada uno de los subsistemas, en el siguiente los modelos globales, dejando para el apartado 4 el establecimiento del balance de ambos enfoques.

2.1 Tambores o discos de cabezas fijas

Estos dispositivos sólo permiten tratar una sola entrada/salida simultánea, debido a que la unidad de control y el canal sólo permiten el paso de una de ellas. Por lo tanto, independientemente del número de ejes de que dispongamos, a efectos de su modelización es equivalente a que dispusiéramos de uno solo. El modelo está constituido por una cola que da acceso a la unidad que controla los distintos dispositivos.

El número de accesos por unidad de tiempo (λ) depende de las aplicaciones que se ejecutan en un momento dado.

El tiempo medio de servicio $\frac{1}{\mu}$ y la variancia de los tiempos de servicio se determinan teniendo en cuenta el tiempo de latencia (tiempo de espera hasta que el registro deseado pasa por debajo de la cabeza de lectura/escritura) y el de transferencia de los registros.

Ejemplo: En un canal de I/0 hay tres tambores conectados a una unidad de control. El tiempo medio de transferencia, incluyendo el de latencia, es de 20 ms con una desviación tipo de 10 ms. Los accesos se producen a razón de 30 por segundo. Determinar el tiempo medio de respuesta. ¿Qué sucedería si en vez de 30 accesos/seg. fuesen 40?

$$E = \left(\frac{20}{10}\right)^2 = 4$$

$\lambda = 30$ ac/seg.

$$\frac{1}{\mu} = 20 \text{ ms} = 0{,}02 \text{ seg.}$$

$$\rho = 30 \times 0{,}02 = 0{,}6$$

$$\bar{t} = 0{,}02 \left[1 + \frac{0{,}6}{2(1 - 0{,}6)} \left(1 + \frac{1}{4}\right)\right] =$$

$$= 0{,}039 \text{ seg.}$$

$\lambda = 40$ ac/seg.

$$\frac{1}{\mu} = 20 \text{ ms} = 0{,}02 \text{ seg.}$$

$$\rho = 40 \times 0{,}02 = 0{,}8$$

$$\bar{t} = 0{,}02 \left[1 + \frac{0{,}8}{2(1 - 0{,}8)} \left(1 + \frac{1}{4}\right)\right] =$$

$$= 0{,}07 \text{ seg.}$$

La observación de estos resultados nos permite extraer una consecuencia de tipo general que es debida al término $1 - \rho$ en el denominador, si ρ es pequeño sus variaciones afectan relativamente poco al resultado, pero si ρ se aproxima a 1 su crecimiento provoca aumentos muy importantes en el tiempo de respuesta.

2.2 Discos

De forma simplificada, el proceso de I/0 en un disco se puede descomponer en las siguientes fases:

- El acceso se pone en cola para acceder al disco correspondiente.
- Sale de la cola para lanzar el movimiento del brazo (seek) ocupando durante un tiempo despreciable la unidad de control cuando la unidad de control y el disco correspondiente están libres simultáneamente.
- Una vez posicionado el brazo se trata de iniciar la transferencia a través de la unidad de control, colocándose el acceso en la cola de este dispositivo.
- Una vez se adquiere servicio de la unidad de control, es preciso esperar que el registro llegue debajo de la cabeza de lectura/escritura (tiempo de latencia) para dejar desfilar entonces todo el registro (tiempo de transferencia).
- Una vez acabada la transferencia se liberan a la vez el disco y la unidad de control, que quedan en disposición de atender nuevos accesos.

Este comportamiento lo podemos representar con el diagrama de la figura 2.1.

El cálculo del tiempo de respuesta se desglosa en las siguientes fases:

a) Cálculo para cada archivo del tiempo medio de ocupación de la unidad de control, que es igual al tiempo de latencia, L, más el tiempo de transferencia, X.

El tiempo medio de latencia es igual al tiempo correspondiente a la mitad de una vuelta, ya que es el tiempo medio de espera para que un registro pase por debajo de la cabeza de lectura/escritura.

El tiempo de transferencia se puede considerar igual al tiempo de una vuelta dividido por el número de registros físicos que hay en una pista.

b) Determinación del número medio de accesos, tanto entradas como salidas, a cada archivo, N, que depende de las aplicaciones que se ejecuten en ese instante.

c) Cálculo de la ocupación de la unidad de control provocada por cada archivo y que es $(X + L)N$.

d) Cálculo del factor de utilización del canal, que es igual a la suma de todos los valores calculados en el apartado anterior.

$$\rho_{ch} = \Sigma\,(X + L)N$$

e) Cálculo del tiempo medio del servicio del canal, que es el promedio de los valores calculados en el apartado a).

$$\frac{1}{\mu_{ch}} = \frac{\Sigma\,(X + L)N}{\Sigma\,N}$$

f) Cálculo del tiempo medio de espera en el canal, T_w, que es igual al tiempo medio de estancia en el sistema menos el de servicio. Hay que tener en cuenta que en este caso no son aplicables las fórmulas de Khintchine-Pollaczeck ya que el manantial no es infinito, puesto que esta cola puede tener, como máximo, tantos accesos en espera como ejes haya en el subsistema. Es preciso utilizar fórmulas adecuadas o los gráficos que nos dan ese valor directamente en función de ρ_{ch}, de $\frac{1}{\mu_{ch}}$ y del número de mecanismos de acceso o ejes.

g) Cálculo para cada disco (un archivo puede estar entre varios discos, o, por el contrario, en un disco puede haber varios archivos) del tiempo medio de servicio que es igual al tiempo medio de desplazamiento de brazo, SK, más el tiempo de espera en el canal, T_w, más el tiempo medio de latencia, L, más el tiempo medio

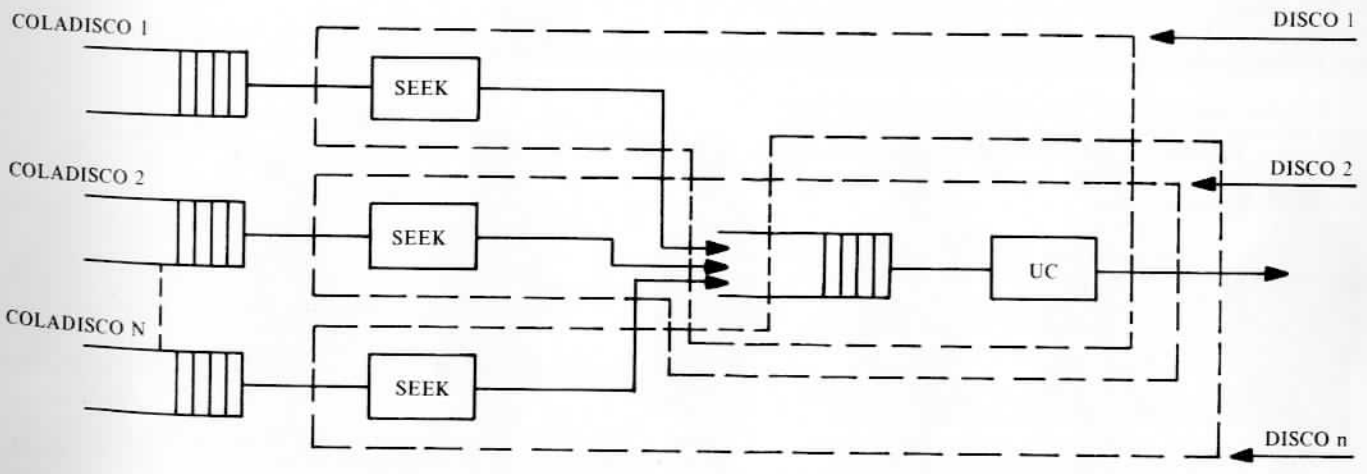


Fig. 2.1.

de transferencia, que depende de los archivos que haya en el disco

$$\frac{1}{\mu_D} = SK + T_w + L + \frac{\Sigma NX}{\Sigma N}$$

El tiempo medio de desplazamiento del brazo depende del tipo de accionamiento del mismo y de la ocupación y distribución de los archivos en el disco. Para discos con un solo archivo, el tiempo medio de desplazamiento del brazo es aproximadamente proporcional al del recorrido de una tercera parte del número total de cilindros y la desviación tipo al del recorrido del número total de cilindros dividido por 4,25.

h) Cálculo del factor de utilización de cada disco

$$\rho_D = \frac{1}{\mu_D} \Sigma N$$

i) Cálculo de la desviación tipo de los tiempos de servicio de cada disco

$$\sigma_D^2 = \sigma_{sk}^2 + \sigma_{T_w}^2 + \sigma_L^2 + \sigma_X^2$$

En el punto *g* ya hemos explicado como puede determinarse σ_{sk}.

σ_{T_w} se hace igual a T_w situándonos en una situación peyorativa, aunque en la mayoría de los casos este término tiene poca importancia.

σ_L^2 es la desviación tipo de una desviación uniforme entre 0 y el tiempo de una vuelta completa y es, por lo tanto,

$$\sigma_L^2 = \frac{(\text{tiempo de una vuelta})^2}{12}$$

σ_X^2 depende de las longitudes de los registros que haya en el disco

$$\sigma_X^2 = \frac{\Sigma NX^2}{\Sigma N} - \left(\frac{\Sigma NX}{\Sigma N}\right)^2$$

Evidentemente si sólo hay un tipo de registros $\sigma_X = 0$.

j) Cálculo del tiempo de respuesta de cada disco por aplicación de la fórmula de Khintchine-Pollaczeck.

k) Al actuar de esta forma no hemos tenido en cuenta que para salir de la cola que da acceso al disco correspondiente es preciso no sólo que el disco esté libre sino que también lo ha de estar la unidad de control. Puede añadirse un tiempo que corrija el tiempo de espera y que puede hacerse igual a

$$\rho_{ch} \cdot \frac{1}{2}\left[\frac{1}{\mu_{ch}} + \frac{\sigma_{ch}^2}{1/\mu_{ch}}\right]$$

y para ser más exactos deben considerarse para cada disco los valores referentes al canal provocados por los restantes discos, pero no él mismo. No obstante, en la mayoría de los casos, esta corrección tiene poca importancia, a no ser que se alcancen elevados factores de utilización de los discos y la unidad de control.

Esta marcha de cálculo nos permite analizar el comportamiento de subsistemas en su funcionamiento más frecuente. Pueden establecerse modelos similares cuando los discos trabajen con posicionamiento angular, con dos unidades de control, etc.

, Ejemplo: Consideremos un subsistema de tres discos donde se hallan los archivos que consulta y actualiza un sistema transaccional de tiempo real. Las transacciones llegan a razón de 30 por segundo. Los discos giran a razón de 3.600 r.p.m.

Al archivo A acceden el 60 % de las transacciones de las cuales el 25 % son de actualización. Hay 10 registros por pista. El disco está totalmente lleno en un 50 %, lo cual hace que la media y la desviación tipo de los tiempos de seek sean 20 ms y 16 ms.

Al archivo B acceden el 40 % de las transacciones sin actualización. Hay 5 registros por pista. El disco está totalmente lleno por lo que la media y la desviación tipo de los tiempos de seek son 30 ms y 24 ms.

Al archivo C acceden el 20 % de las transacciones que hacen en promedio tres accesos. Hay 20 registros por pista. El disco está lleno en un 30 %, por lo que la media y la desviación tipo de los tiempos de seek son 15 ms y 12 ms.

Determinar los tiempos medios de acceso de cada archivo.

$N_A = 30 \cdot 0{,}6 \cdot 1{,}25 = 22{,}5$ accesos/seg

$N_B = 30 \cdot 0{,}4 \cdot 1 = 12$ accesos/seg

$N_C = 30 \cdot 0{,}2 \cdot 3 = 18$ accesos/seg

$$L + X_A = 8{,}33 + \frac{16{,}67}{10} = 10 \text{ mseg}$$

$$L + X_B = 8{,}33 + \frac{16{,}67}{5} = 11{,}67 \text{ mseg}$$

$$L + X_C = 8{,}33 + \frac{16{,}67}{20} = 9{,}17 \text{ mseg}$$

$N_A (L + X_A) = 0{,}225$ seg

$N_B (L + X_B) = 0{,}140$ seg

$N_C (L + X_C) = 0{,}165$ seg

$\rho_{ch} = 0{,}225 + 0{,}140 + 0{,}165 = 0{,}53$

$$\frac{1}{\mu_{ch}} = \frac{0{,}225 + 0{,}140 + 0{,}165}{22{,}5 + 12 + 18} = 0{,}0101 \text{ seg}$$

Para determinar el tiempo de espera en el canal podemos usar el gráfico de la figura 28.6 de la referencia [MART 67], obteniendo

$$T_w = (1{,}45 - 1) \cdot 0{,}0101 = 0{,}00455 \text{ seg}$$

$$\frac{1}{\mu_A} = 20 + 4{,}55 + 10 = 34{,}55 \text{ mseg}$$

$$\frac{1}{\mu_B} = 30 + 4{,}55 + 11{,}67 = 46{,}22 \text{ mseg}$$

$$\frac{1}{\mu_C} = 15 + 4{,}55 + 9{,}17 = 28{,}72 \text{ mseg}$$

$\rho_A = 22{,}5 \times 0{,}03455 = 0{,}777$

$\rho_B = 12 \times 0{,}04622 = 0{,}555$

$\rho_C = 18 \times 0{,}02872 = 0{,}517$

$$\sigma_A^2 = 16^2 + 4{,}55^2 + \frac{16{,}67^2}{12} + 0^2 = 299{,}85 \text{ mseg}^2$$

$$\sigma_B^2 = 24^2 + 4{,}55^2 + \frac{16{,}67^2}{12} + 0^2 = 619{,}85 \text{ mseg}^2$$

$$\sigma_C^2 = 12^2 + 4{,}55^2 + \frac{16{,}67^2}{12} + 0^2 = 187{,}85 \text{ mseg}^2$$

$$\bar{t}_A = 34{,}55 \left[1 + \frac{0{,}777}{2(1 - 0{,}777)} \left(1 + \frac{299{,}85}{34{,}55^2}\right)\right] = 109{,}83 \text{ m}$$

$$\bar{t}_B = 46{,}22 \left[1 + \frac{0{,}555}{2(1 - 0{,}555)} \left(1 + \frac{619{,}85}{46{,}22^2}\right)\right] = 83{,}38 \text{ ms}$$

$$\bar{t}_C = 28{,}72 \left[1 + \frac{0{,}517}{2(1 - 0{,}517)} \left(1 + \frac{187{,}85}{28{,}72^2}\right)\right] = 47{,}70 \text{ ms}$$

La corrección debida a la espera suplementaria en la cola a causa de estar ocupada la unidad de control se calcula como sigue,

$$\rho'_A = 0{,}140 + 0{,}165 = 0{,}305$$

$$\frac{1}{\mu'_A} = \frac{12 \times 11{,}67 + 18 \times 9{,}17}{12 \times 18} = 10{,}17 \text{ mseg}$$

$$\sigma'^2_A = \frac{12\left(11{,}67^2 + \frac{16{,}67^2}{12}\right) + 18\left(9{,}17^2 + \frac{16{,}67^2}{12}\right)}{12 + 18} -$$

$$- 10{,}17^2 = 24{,}652 \text{ mseg}^2$$

$$\Delta t_A = 0{,}305 \cdot \frac{1}{2}\left(10{,}17 + \frac{24{,}652}{10{,}17^2}\right) = 1{,}92 \text{ mseg}$$

$$\rho'_B = 0{,}225 + 0{,}165 = 0{,}39$$

$$\frac{1}{\mu'_B} = \frac{22{,}5 \times 10 \times 18 \times 9{,}17}{22{,}5 + 18} = 9{,}63 \text{ mseg}$$

$$\sigma'^2_B = \frac{22{,}5\left(10^2 + \frac{16{,}67^2}{12}\right) + 18\left(9{,}17^2 + \frac{16{,}67^2}{12}\right)}{22{,}5 + 18} -$$

$$- 9{,}63^2 = 23{,}317 \text{ mseg}^2$$

$$\Delta t_B = 0{,}39 \cdot \frac{1}{2}\left(9{,}63 + \frac{23{,}317}{9{,}63}\right) = 2{,}35 \text{ mseg}$$

$$\rho'_C = 0{,}225 + 0{,}14 = 0{,}365$$

$$\frac{1}{\mu'_C} = \frac{22{,}5 \times 10 + 12 \times 11{,}67}{22{,}5 + 12} = 10{,}58 \text{ mseg}$$

$$\sigma'^2_C = \frac{22{,}5\left(10^2 + \frac{16{,}67^2}{12}\right) + 12\left(11{,}67^2 + \frac{16{,}67^2}{12}\right)}{22{,}5 + 12} -$$

$$- 10{,}58^2 = 23{,}776 \text{ mseg}^2$$

$$\Delta t_C = 0{,}365 \cdot \frac{1}{2}\left(10{,}58 + \frac{23{,}776}{10{,}58}\right) = 2{,}34 \text{ mseg}$$

$$\bar{t}_A = 109{,}83 + 1{,}92 = 111{,}75 \text{ mseg}$$

$$\bar{t}_B = 83{,}38 + 2{,}35 = 85{,}73 \text{ mseg}$$

$$\bar{t}_C = 47{,}70 + 2{,}34 = 50{,}04 \text{ mseg}$$

2.3 Subsistemas secuenciales

La modelización de estos subsistemas no ofrece ninguna dificultad ya que, en general, pueden representarse por la estación de servicio elemental. Además, estos subsistemas no acostumbran a ser los cuellos de botella de un sistema informático y, por lo tanto, su modelización no plantea problemas especiales.

2.4 CPU y memoria

Estos dos elementos de un sistema informático se interaccionan tan íntimamente que su modelización debe llevarse a cabo conjuntamente.

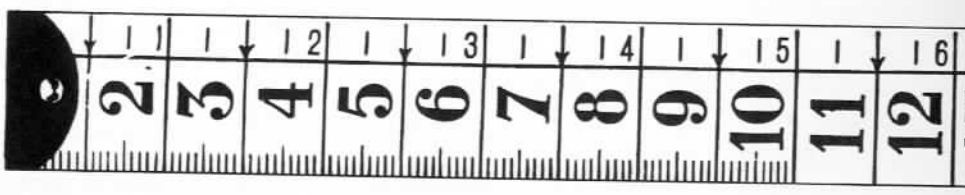
La CPU tiene como modelo una estación de servicio con uno o varios canales de servicio según el número de procesadores existentes. Lo que ya resulta algo más difícil de modelizar correctamente es el número y la política de gestión de las colas ya que depende, por un lado de la política de asignación de procesadores a los trabajos en cola y por otro del régimen de prioridades e interrupciones.

Existen numerosos modelos, pues, para atender a las distintas posibilidades como son los modelos conocidos como de "round robin" o de procesador compartido, de procesador interrumpible, etc., algunos de los cuales aparecerán en el apartado siguiente, pero en primera aproximación es suficiente considerar una cola de capacidad limitada (si quiere considerarse que únicamente pueden acceder a la CPU los programas existentes en memoria) o ilimitada y una estación de servicio con uno o varios canales de servicio de acuerdo con el número de procesadores existentes.

La modelización de la memoria es especialmente importante en el caso de los sistemas de tiempo real e interactivos, mientras que en el caso de los sistemas batch debido a la menor variabilidad a lo largo del tiempo de la ocupación de la memoria es más fácil determinar su mapa.

Centrémonos en el caso de sistemas en tiempo real. Ante todo es preciso conocer el tiempo de permanencia de una transacción en memoria, que es igual a la suma de los tiempos de espera en la cola de la CPU, de uso de CPU y de realización de la entrada/salida. Este tiempo de presencia influye, evidentemente, en la memoria necesaria y el nivel de multiprogramación. A su vez, éste incide en la determinación de la dimensión de la cola existente en la CPU.

Evidentemente hay que considerar además el espacio ocupado por el software fijo (sistema operativo, sistema de base de datos, etc.).



Por otro lado, los sistemas de memoria virtual presentan además las complicaciones adicionales que representan la determinación del conjunto de trabajo y la frecuencia de fallo de página, que, a su vez, también están relacionados.

Las principales dificultades en el establecimiento de de modelos de unidades centrales reside en la obtención de datos suficientemente correctos. Así, respecto a los programas la longitud de código ejecutado, la dimensión de los programas, y en los sistemas de memoria virtual, el conjunto de trabajo y la frecuencia de fallo de página, y respecto al procesador, el número de instrucciones ejecutadas por unidad de tiempo, no son datos fáciles de obtener. Tal vez la forma más segura, y eso aún para cada caso concreto, es a través de medidas realizadas con monitores sobre sistemas ya existentes.

A continuación se expone un sencillo ejemplo de modelo de un conjunto CPU-memoria.

Ejemplo: disponemos de una CPU capaz de ejecutar 500.000 instrucciones por segundo. Este sistema atiende transacciones a razón de 4 por segundo, sabiendo que los accesos de I/0 tienen un tiempo medio de 75 milisegundos. Determinar la memoria necesaria y el factor de multiprogramación del sistema, sabiendo que las características de las transacciones están resumidas en el cuadro que se expone a continuación

	Accesos	Ocupación (K)	Instrucciones x 10^3	Frecuencia %
T1	5	20	80	37,5
T2	6	25	90	32,5
T3	8	40	110	20
T4	9	50	150	10

El factor de ocupación de la CPU es el siguiente:

$$\rho = \frac{(0{,}375 \times 80.000 + 0{,}325 \times 90.000 + 0{,}2 \times 110.000 + }{500.000}$$

$$\frac{+ 0{,}1 \times 150.000)\, 4}{500.000} = 0{,}77$$

Admitiendo en primera aproximación y como hipótesis muy peyorativa que se pueden aplicar a la cola de la CPU las fórmulas de Khintchine-Pollaczeck con los tiempos de servicio distribuidos exponencialmente, tendremos

$$\frac{1}{\mu} = \frac{0{,}375 \times 80.000 + 0{,}325 \times 90.000 + 0{,}2 \times 110.000 + }{500.000}$$

$$\frac{+ 0{,}1 \times 150.000}{500.000} = 0{,}1925 \text{ seg}$$

$$t = \frac{0{,}1925}{1 - 0{,}77} = 0{,}83696 \text{ seg}$$

El tiempo medio de espera en la cola de la CPU será

$$t_w = 836{,}96 - 192{,}5 = 644{,}46 \text{ mseg}$$

Los tiempos de presencia de cada una de las transacciones serán

$$t_1 = \frac{80.000}{500} + 644{,}46 + 5 \times 75 = 1179{,}46 \text{ mseg}$$

$$t_2 = \frac{90.000}{500} + 644{,}46 + 6 \times 75 = 1274{,}46 \text{ mseg}$$

$$t_3 = \frac{110.000}{500} + 644{,}46 + 8 \times 75 = 1464{,}46 \text{ mseg}$$

$$t_4 = \frac{150.000}{500} + 644{,}46 + 9 \times 75 = 1619{,}46 \text{ mseg}$$

El factor de multiprogramación es igual a

$$N = (0{,}375 \times 1{,}17946 + 0{,}325 \times 1{,}27446 + 0{,}2 \times 1{,}46446 + + 0{,}1 \times 1{,}61946) \times 4 = 5{,}325$$

que evidentemente se redondea a uno de los enteros más próximos, por ejemplo, 6.

El espacio de memoria necesario es

$$M = (0{,}375 \times 1{,}17946 \times 20 + 0{,}325 \times 1{,}27446 \times 25 + + 0{,}2 \times 1{,}46446 \times 40 + 0{,}1 \times 1{,}61946 \times 50) = 159{,}26 \text{ K}$$

Si ahora rehiciéramos el cálculo con factor de multiprogramación 6, teniendo en cuenta que la cola es finita encontraríamos (no se detalla al cálculo):

Las probabilidades de tener i elementos en el sistema CPU son

$\rho_0 = 0{,}22973$
$\rho_1 = 0{,}25017$
$\rho_2 = 0{,}22703$
$\rho_3 = 0{,}16483$
$\rho_4 = 0{,}08975$
$\rho_5 = 0{,}03258$
$\rho_6 = 0{,}00591$

lo cual nos da que el número medio de elementos en el sistema es

$$N = 0 \times 0{,}22973 + 1 \times 0{,}25017 + 2 \times 0{,}22703 + 3 \times \times 0{,}16483 + 4 \times 0{,}08975 + 5 \times 0{,}03258 + + 6 \times 0{,}00591 = 1{,}75608$$

y el tiempo medio de permanencia en el sistema es

$$\bar{t} = \frac{1{,}75608}{4} = 0{,}43902 \text{ seg}$$

y, en consecuencia, el tiempo medio de espera será

$$t_w = 0{,}43902 - 0{,}1925 = 0{,}24652$$

Los tiempos medios de cada transacción serán, ahora,

$t_1 = 0{,}78152$ seg
$t_2 = 0{,}87652$ seg
$t_3 = 1{,}06652$ seg
$t_4 = 1{,}22152$ seg

A la vista de estos resultados se podría reconsiderar el factor de multiprogramación, cosa que aquí no se ha hecho.

Ahora bien, para determinar el tiempo de respuesta, es decir desde que una transacción llega hasta que termina su ejecución, hay que sumar a estos tiempos el tiempo de espera que sufren cuando encuentran ocupados los seis espacios de memoria para las transacciones. Esto puede modelizarse mediante una estación con seis canales de servicio, cuya saturación y tiempo medio de servicio son respectivamente,

$$t_m = 0{,}78152 \times 0{,}375 + 0{,}375 + 0{,}87652 \times 0{,}325 + + 1{,}06652 \times 0{,}2 + 1{,}22152 \times 0{,}1 = 0{,}9134 \text{ seg}$$

$$\rho = \frac{4 \times 0{,}9134}{6} = 0{,}6089$$

y utilizando, por ejemplo el gráfico de la figura 26.23 de la referencia [MART 67] tenemos

$$t_w = (1{,}1 - 1) \times 0{,}9134 = 0{,}09134 \text{ seg}$$

con lo que los tiempos de respuesta de cada transacción son

$t_1 = 0{,}78152 + 0{,}09134 = 0{,}87286$ seg

$t_2 = 0{,}87652 + 0{,}09134 = 0{,}96786$ seg

$t_3 = 1{,}06652 + 0{,}09134 = 1{,}15786$ seg

$t_4 = 1{,}22152 + 0{,}09134 = 1{,}31286$ seg

2.5 Red de comunicaciones

La red de comunicaciones puede adoptar numerosas y variadas estructuras, por lo que es difícil indicar una metodología general para su análisis. No obstante en cada caso particular es posible establecer modelos adaptados a la estructura existente basados siempre en las colas que se produzcan en la red.

3. REDES DE COLAS

Diremos que tenemos una red de colas cuando las transacciones que salen de una estación de servicio van a parar:

- O en forma determinista a otro sistema de colas.
- O al exterior del sistema.
- O en forma aleatoria con probabilidades determinadas a uno de entre varios sistemas de colas o al exterior.

En estas redes el número de trabajos que existen en su interior puede ser fijo (red cerrada) o puede producirse un flujo de trabajos que pueden entrar y salir por distintos puntos (red abierta). En ambos casos su estudio es más complejo si se quiere tratar de forma exacta. En muchos casos se puede hallar un tratamiento analítico del conjunto, en otros, sin embargo, es preciso recurrir a métodos de simulación.

Jackson (referencia [JACK 63]) dio una primera forma de tratar redes de colas, que fue ampliada recientemente por Baskett, Chandy, Muntz y Palacios (BCMP) (referencia [BASK 75]) incluyendo todos los casos susceptibles del mismo tratamiento.

3.1 Métodos analíticos (BCMP)

Los sistemas que consideramos contienen un número arbitrario pero finito, N, de estaciones de servicio. Hay un número arbitrario pero finito, R, de clases distintas de trabajos. Es decir, un trabajo de clase r que sale de la estación de servicio i requerirá servicio de la estación j en la clase s con una probabilidad que iniciaremos por $P_{i,r,j,s}$. Esta matriz la designaremos por $P = [P_{i,r,j,s}]$ y sea n_{ir} el número de trabajos de clase r en la estación i.

Las estaciones de servicio pueden ser de los cuatro tipos siguientes

Tipo 1. La estación i tiene un solo canal con tiempo de servicio exponencial de tiempo medio $1/\mu_i\,(n_i)$ idéntico para todas las clases, siendo n_i ($n_i = \sum_r n_{ir}$) el número de trabajos en la estación y la disciplina de la cola, FIFO. (Los discos se asimilan a una estación de este tipo).

Tipo 2. La estación tiene un solo canal, la disciplina de servicio es de procesador compartido (es decir, cuando hay n trabajos en la estación de servicio, cada uno recibe servicio a razón de $1/n$ de segundo cada segundo) y cada clase de trabajo tiene una distribución de tiempos de servicio que puede ser distinta y arbitraria. (Una estación de este tipo es la CPU).

Tipo 3. El número de canales en la estación de servicio es mayor o igual que el número máximo de trabajos que puede haber en la estación en un instante cualquiera y cada clase de trabajo tiene una distribución de tiempos de servicio que puede ser distinta y arbitraria (Una estación de este tipo son los terminales interactivos).

Tipo 4. La estación de servicio tiene un solo canal, la disciplina de cola es LIFO con interrupción del último en llegar y cada clase de trabajo tiene una distribución de tiempos de servicio que puede ser distinta y arbitraria. (Una estación de este tipo es la interrupción de la CPU).

El proceso de llegada a la red sigue una distribución de Poisson de parámetro $\lambda\,(n)$, donde $n = \sum_i n_i$ es el número de trabajos que hay en el sistema representado por la red.

Además es preciso calcular las frecuencias efectivas de llegada de cada clase de trabajo a cada estación de servicio e_{ir}. Ello se logra resolviendo el siguiente sistema de ecuaciones:

$$e_{js} = \sum_{r=1}^{R} \sum_{i=1}^{N} e_{ir}\, P_{i,r,j,s} + q_{js}$$

donde q_{js} es la probabilidad de entrada desde el exterior en la estación j de un trabajo de clase s. Si la red es cerrada, evidentemente todas las q_{js} son nulas.

El enunciado completo del teorema de BCMP lleva a una expresión notablemente compleja, por lo que expondremos aquí sólo las consecuencias que permiten una más fácil comprensión y aplicación.

Definiremos como estado del sistema el número de trabajos de cada clase en cada estación de servicio. Más formalmente el estado del sistema viene dado por $(y_1, y_2, \ldots y_n)$, donde $y_i = (n_{i1}, n_{i2}, \ldots n_{ir})$. Sea $1/\mu_{ir}$ el tiempo medio de servicio de un trabajo de clase r en la estación i. Entonces, para una red de estaciones de servicio que puede ser abierta, cerrada o mixta y en que éstas pueden ser de tipo 1, 2, 3 ó 4, las probabilidades del estado en equilibrio vienen dadas por

$$P(y_1, y_2, \ldots y_n) = Cd(n)\; g_1(y_1)\, g_2(y_2) \ldots\; g_n(y_n)$$

donde:

Si la estación es de tipo 1, entonces

$$g_i(y_i) = n_i! \left(\prod_{r=1}^{R} \frac{1}{n_{ir}!}\; e_{ir}^{\,n_{ir}} \right) \left(\frac{1}{\mu_i} \right)^{n_i}$$

Si la estación es de tipo 2 ó 4, entoncess

$$g_i(y_i) = n_i! \prod_{r=1}^{R} \frac{1}{n_{ir}!} \left(\frac{e_{ir}}{\mu_{ir}} \right)^{n_{ir}}$$

Si la estación es de tipo 3, entonces

$$g_i(y_i) = \prod_{r=1}^{R} \frac{1}{n_{ir}!} \left(\frac{e_{ir}}{\mu_{ir}} \right)^{n_{ir}} ,$$

$$d(n) = \prod_{i=0}^{n-1} \lambda\,(i)$$

si la red es abierta y $d(n) = 1$ si es cerrada.

C es la constante de normalización para lograr que la suma de todas las probabilidades sea igual a 1.

Si además suponemos que la red es abierta y que la llegada no depende del estado del modelo y sólo nos interesa conocer la probabilidad $P_i(n_i)$ de tener n_i trabajos en la estación i, entonces

$P_i(n_i) = (1 - \rho_i)\rho_i^{\,n_i}$ si la estación es de tipo 1, 2 ó 4.

$P_i(n_i) = e^{-\rho_i}\, \rho_i^{\,n_i}/n_i!$ si la estación es de tipo 3 y

donde

$$\rho_i = \sum_{r=1}^{R} \lambda\, e_{ir}/\mu_i \text{ si la estación es de tipo 1.}$$

$$\rho_i = \sum_{r=1}^{R} \lambda\, e_{ir}/\mu_{ir}$$ si la estación es de tipo 2, 3 ó 4.

Ejemplo: Consideremos el ejemplo de la figura 3.1. en que tenemos un sistema cerrado, con dos clases de trabajos y cinco estaciones de servicio, la 1 de tipo 2 (procesador compartido) y las otras cuatro de tipo 1 (discos).

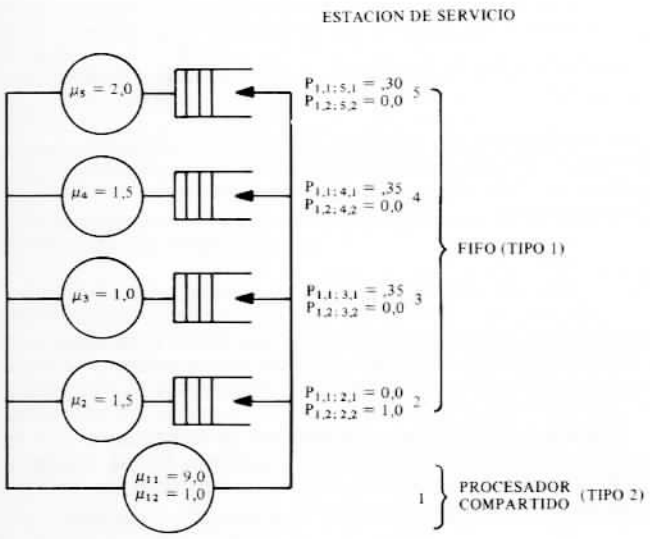


Fig. 3. 1

En la tabla que sigue se exponen los niveles de ocupación de las cinco estaciones de servicio cuando hay un trabajo de clase 2 y un número variable de clase 1. Estos resultados se han obtenido por aplicación directa de la teoría expuesta,

	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5
0	0,6	0,4	0	0	0
1	0,678	0,371	0,384	0,256	0,165
2	0,720	0,352	0,606	0,404	0,260
3	0,744	0,339	0,743	0,495	0,318
4	0,759	0,330	0,831	0,554	0,356
5	0,769	0,324	0,888	0,592	0,381
6	0,775	0,321	0,926	0,617	0,397
7	0,779	0,318	0,951	0,634	0,407

4. BALANCE COMPARATIVO

Acabamos de exponer dos métodos cuyo enfoque, aún partiendo ambos de la teoría de colas, difieren por la consideración parcial o global que se hace del sistema informático.

El primero, de análisis separado de cada subsistema, tiene a su favor la facilidad de ejecución de los cálculos y en su contra:

- el no tener en cuenta la interacción entre los distintos subsistemas
- la no adecuación de las leyes de probabilidad consideradas a la realidad de las frecuencias de llegada o servicio, de las que, en ciertos casos difieren notablemente, pero que se mantienen siempre que no se haya determinado la distribución real, por la facilidad de cálculo a que llevan.

El segundo, de análisis global del sistema, tiene a su favor precisamente la consideración de la interacción de todos los subsistemas que lo componen y una mejor adecuación de las leyes de probabilidad de las frecuencias de llegada y servicio y en su contra:

- la dificultad de la realización de los cálculos
- la limitación de los tipos de estación de servicio que se pueden considerar y que dejan fuera, por ejemplo las estaciones de servicio con varios canales.

Vemos pues que ambos presentan ciertos inconvenientes, que para obviar simultáneamente es preciso recurrir bien sea a métodos aproximados (por ejemplo métodos de difusión) o bien tratamientos por simulación de modelos globales de sistemas informáticos, requiriendo todos ellos grandes cantidades de cálculos.

5. CONCLUSION

Como consecuencia de todo lo expuesto podemos concluir que:

- Los esfuerzos dedicados a modelizar tienen la virtud de obligar a cuantificar los distintos aspectos de un sistema informático y permiten a posteriori analizar las diferencias entre modelo y realidad que conducen siempre a una mejora de los modelos utilizados.
- Es un tema que en general requiere especialistas que aúnen sólidos conocimientos de arquitectura de sistemas informáticos, en sus aspectos hardware y software, con una amplia base de estadística y teoría de probabilidades.
- Se podrá argumentar en contra de estos métodos la dificultad en la obtención de datos suficientemente correctos del hardware (por ejemplo, n.º de instrucciones ejecutadas por unidad de tiempo) y del software (por ejemplo, número de instrucciones ejecutadas en cada programa). No obstante el uso de monitores hardware y software, de medidores de número de instrucciones, etc. permite la obtención de datos suficientemente correctos de todos los parámetros necesarios para la construcción de modelos de sistemas informáticos de los tipos considerados en este artículo.

R. Puigjaner