
CAPÍTULO 1

INTRODUCCIÓN A LA TEORÍA DE COLAS

En los sistemas computadores, muchas tareas comparten los recursos del sistema tales como la CPU, discos y otros dispositivos. Como generalmente sólo una tarea puede usar el recurso en un momento dado, las demás tareas que quieran utilizar el recurso esperarán en colas. Por este motivo, el modelado analítico consiste fundamentalmente en la teoría de colas. La teoría de colas ayuda a determinar el tiempo que la tarea emplea en varias colas dentro del sistema. Estos tiempos se pueden combinar para predecir el tiempo de respuesta, parámetro importante en las prestaciones de un sistema, y que normalmente es la suma de todos los tiempos empleados por la tarea en el sistema.

1.1 Notación de las Colas

Supongamos una sala de terminales como la que aparece en la figura 1.1. La sala tiene un número de terminales dado para ser utilizados por los usuarios. Si todos los terminales están ocupados, los usuarios que llegan esperarán en una cola. En términos de teoría de colas, los usuarios se denominan "clientes". Para analizar el sistema deben especificarse varias características del sistema.

Los valores característicos son:

1. *Proceso de llegada*: Si los usuarios llegan en tiempos t_1, t_2, \dots, t_j , las variables aleatorias $\tau_j = t_j - t_{j-1}$ se llaman tiempos entre llegadas. Se supone generalmente que las τ_j forman una secuencia de variables aleatorias independientes e idénticamente distribuidas (IID). De entre todos los posibles procesos de llegada el más común es el denominado de **Poisson**, lo cual implica que los intervalos de llegada son de tipo IID y están distribuidos exponencialmente.

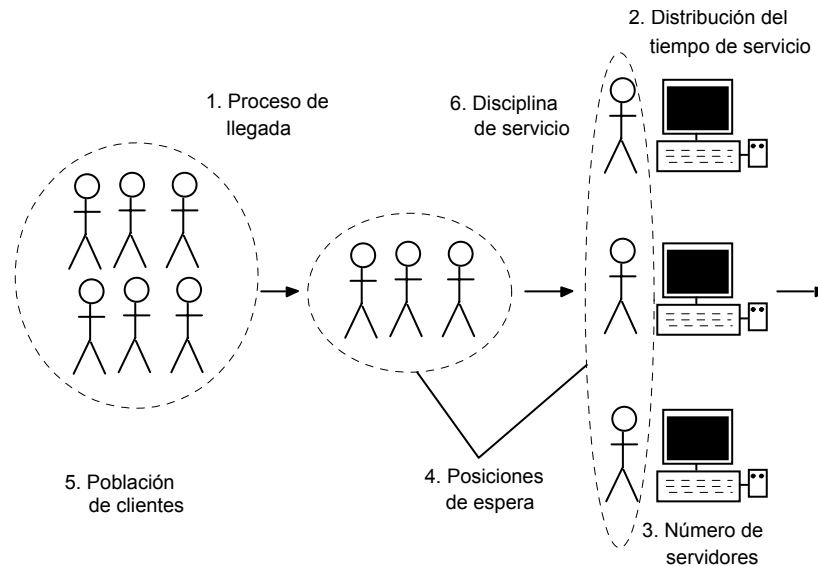


Figura 1.1 Componentes básicos de una cola.

2. *Distribución del tiempo de servicio:* También es necesario conocer el tiempo que cada usuario emplea en el terminal. Esto es lo que se llama tiempo de servicio. Es normal suponer que los tiempos de servicio son variables aleatorias de tipo IID. La distribución empleada con mayor frecuencia es la distribución exponencial.
3. *Número de servidores:* La sala de terminales puede tener uno o más terminales. Todos ellos se consideran como parte del mismo sistema de colas, si todos los terminales son idénticos y pueden asignarse a cualquier usuario. Si alguno no fuera igual debería dividirse el sistema en dos colas, una para cada tipo de sistema.
4. *Capacidad del sistema:* El número máximo de usuarios que pueden estar en la cola puede estar limitado por el espacio disponible, y también para evitar tiempos de espera largos. Este número se denomina capacidad del sistema. En la mayoría de los sistemas, la capacidad es finita. Sin embargo, si el número es grande, se facilita el análisis al suponerlo infinito. La capacidad del sistema incluye tanto a los usuarios en la cola como aquellos que están recibiendo servicio.
5. *Tamaño de la población:* El número total de potenciales usuarios de los terminales recibe el nombre de población. En la mayoría de los sistemas la población es finita. Si la población es grande, se facilitan los cálculos si suponemos que es una población infinita.
6. *Disciplina de servicio:* El orden en el cual se sirven los usuarios se denomina disciplina de servicio. La disciplina más frecuente es primero en llegar, primero en servirse (FCFS). Otras posibles alternativas son último en llegar, primero en servirse (LCFS), pueden admitirse también prioridades, etc.

Existen también lo que se conoce como **servidores infinitos (IS)** o **centros de retardo (delay centers)**, estos nombres se aplican a sistemas con un tiempo de retardo fijo, independientemente del número de usuarios.

Otra posibilidad de servicio se basa en el tiempo requerido, ejemplos de estas disciplinas pueden ser: el de tiempo más corto se procesa primero (SPT), el de tiempo restante más corto se procesa primero (SRPT), también, el más grande primero (BIFS), etc.

Para especificar un sistema de colas, hemos de especificar estos seis parámetros. Se emplea en la teoría de colas una notación abreviada conocida como **notación Kendall**, de la forma: $A/S/m/B/K/SD$, donde cada letra corresponde en orden a los seis parámetros listados anteriormente.

Así, A es el tiempo entre llegadas, S la distribución del tiempo de servicio, m el número de servidores, B es el número de buffers (capacidad del sistema), K es el tamaño de la población y SD es la disciplina de servicio.

Las distribuciones posibles para los tiempos de llegada se denotan por una letra cuyo significado es el siguiente:

M	Exponencial
E_k	Erlang con parámetro k
H_k	Hiperexponencial con parámetro k
D	Determinista
G	General

Una distribución de tipo general significa que la distribución no está especificada y los resultados son válidos para todas las distribuciones. La distribución exponencial, denotada por M, es la única que no tiene memoria, es decir, el tiempo de llegada de un usuario no depende de la llegada del anterior.

Normalmente se emplea una notación abreviada realizando las siguientes simplificaciones: capacidad del sistema infinita, tamaño de la población infinita y disciplina de servicio FCFS, así la notación resultante quedaría del tipo $A/S/m$.

1.2 Reglas para Todas las Colas

En este apartado se tratarán algunas de las variables claves empleadas en el análisis de colas simples y se discutirá la relación entre ellas. La figura 1.2 muestra las variables empleadas en el análisis de colas.

τ = tiempo entre llegadas, es decir, el tiempo transcurrido entre dos llegadas sucesivas.

λ = razón media de llegada = $1/E[\tau]$. En algunos sistemas, puede ser función del estado del sistema. Por ejemplo, puede depender del número de tareas ya en el sistema.

s = tiempo de servicio por tarea.

μ = razón media de servicio por servidor, = $1/E[s]$. La razón de servicio total para m servidores es $m\mu$.

n = número de tareas (trabajos) en el sistema. También se conoce como **longitud de la cola**. Incluye las tareas que están recibiendo servicio y las que esperan en la cola.

n_q = número de tareas esperando para recibir servicio.

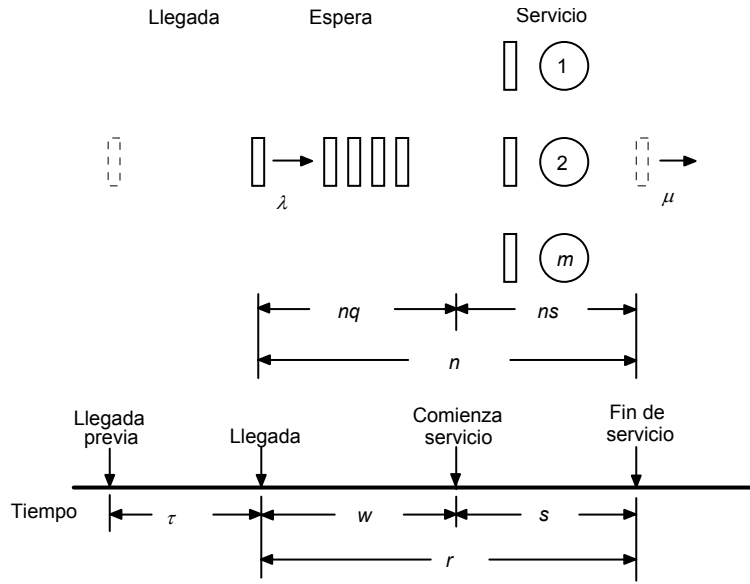


Figura 1.2 Variables comunes empleadas en el análisis de colas.

n_s = número de tareas recibiendo servicio.

r = tiempo de respuesta o tiempo en el sistema. Incluye los dos tiempos, el tiempo de espera por servicio y el tiempo recibiendo servicio.

w = tiempo de espera, es el intervalo de tiempo entre la llegada y el momento en que comienza a recibir servicio.

Todas las variables excepto λ y μ son variables aleatorias. Hay un número de relaciones entre estas variables que se aplican a las colas G/G/m. Como la mayoría de las colas son casos particulares de la anterior, las relaciones son válidas para casi todos los tipos de colas que podamos encontrar.

1. *Condición de estabilidad*: Si el número de tareas en un sistema crece continuamente y tiende a infinito, se dice que el sistema es inestable. Para la estabilidad debe cumplirse que la razón de llegada sea menor que la razón de servicio:

$$\lambda < m\mu$$

Esta condición de estabilidad no es aplicable a poblaciones finitas o sistemas con buffers finitos. En una población finita la longitud de la cola será siempre finita y el sistema nunca será inestable.

2. *Número de elementos en el sistema / número de elementos en la cola*: El número de tareas en el sistema es siempre igual a la suma del número de tareas en la cola y el número de tareas recibiendo servicio.

$$n = n_q + n_s$$

Estas variables son aleatorias, la ecuación anterior conduce a la siguiente relación entre sus medias:

$$E[n] = E[n_q] + E[n_s]$$

Además, si la razón de servicio de cada servidor es independiente del número de elementos en la cola, tendremos:

$$\text{Cov}(n_q, n_s) = 0$$

$$\text{y } \text{Var}[n] = \text{Var}[n_q] + \text{Var}[n_s]$$

3. *Número / tiempo*: Si las tareas no se pierden debido a la falta de buffers, el número medio de tareas en un sistema está relacionado con su tiempo medio de respuesta como sigue:

Número medio tareas en el sistema = Razón de llegada x Tiempo medio de respuesta

(1.1)

Similarmente

Número medio tareas en la cola = Razón de llegada x Tiempo medio de espera

(1.2)

Las ecuaciones 1.1 y 1.2 son conocidas como **ley de Little**. Se abordará esta ley con mayor profundidad en el apartado siguiente.

4. *Tiempo en el sistema y tiempo en la cola*: El tiempo que una tarea pasa en la cola dentro del sistema de colas, es igual a la suma del tiempo de espera en la cola y el tiempo recibiendo servicio:

$$r = w + s$$

Son variables aleatorias, esto conduce a la siguiente relación entre sus medias:

$$E[r] = E[w] + E[s]$$

Si además la razón de servicio es independiente del número de tareas en la cola tendremos:

$$\text{Cov}(w, s) = 0$$

y por tanto

$$\text{Var}[r] = \text{Var}[w] + \text{Var}[s]$$

1.3 Ley de Little

Uno de los teoremas usados con más frecuencia en la teoría de colas es la ley de Little, la cual nos permite relacionar el número de tareas en cualquier sistema con el tiempo medio empleado en el sistema, como sigue:

Número medio en el sistema = Razón de llegada x Tiempo medio de respuesta

Esta relación se aplica a todos los sistemas o partes de sistemas en los cuales el número de tareas entrantes al sistema sea igual al de tareas que completan servicio, es decir, no aparecen nuevas tareas dentro del sistema y tampoco desaparecen tareas en el sistema.

La demostración de esta ley será como sigue: Supongamos que monitorizamos el sistema durante un intervalo de tiempo T y guardamos un registro de los tiempos de

llegada y de partida de cada tarea. Si T es grande, el número de llegadas sería aproximadamente igual al de partidas. Sea ese número N . Entonces:

$$\text{Razón de llegada} = \text{número de llegadas/tiempo total} = N/T$$

Como aparece en la figura 1.3 hay tres formas de representar los datos obtenidos. La figura 1.3a muestra el número total de llegadas y de partidas separadamente en función del tiempo. Si en cada instante restamos la curva de partidas de la curva de llegadas obtendremos el número de tareas en el sistema en cada instante, como aparece en la figura 1.3b. Por el contrario, si se resta el tiempo de llegada del tiempo de partida de cada tarea, obtenemos la figura 1.3c para el tiempo pasado en el sistema.

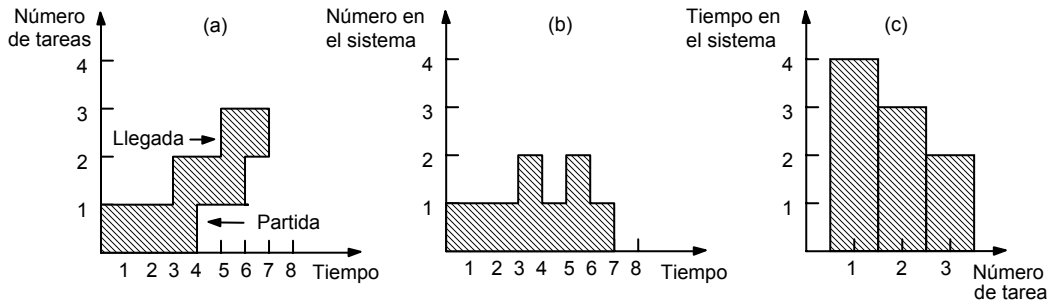


Figura 1.3 Tres formas de representar los tiempos de llegada y partida.

Las áreas rayadas representan el tiempo total pasado en el sistema por todas las tareas. Sea J este área, de la figura 1.3c tenemos:

$$\text{Tiempo medio pasado en el sistema} = J/N$$

de la figura 1.3b

$$\begin{aligned} \text{Número medio en el sistema} &= \frac{J}{T} \\ &= \frac{N}{T} \times \frac{J}{N} \\ &= \text{razón de llegada} \times \text{tiempo medio en el sistema} \end{aligned}$$

La ley de Little puede usarse para sistemas o subsistemas.

Ejemplo 1.1 Una monitorización de un servidor de disco mostró que el tiempo promedio para satisfacer una petición de E/S era de 100 mseg. La razón de E/S era de 100 peticiones por segundo. ¿Cuál fue el número medio de peticiones al servidor de disco?

Usando la ley de Little.

$$\begin{aligned} \text{Número medio en el servidor de disco} &= \text{razón de llegada} \times \text{tiempo de respuesta} \\ &= (100 \text{ peticiones/segundo}) \times (0.1 \text{ segundos}) \\ &= 10 \text{ peticiones} \end{aligned}$$

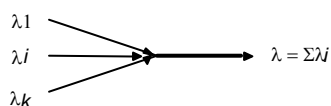
1.4 Tipos de Procesos Estocásticos

En el modelado analítico se emplean no sólo variables aleatorias, sino diferentes secuencias o familias de variables aleatorias que son función del tiempo. Por ejemplo, sea $n(t)$ que denota el número de tareas en la CPU de un sistema computador. Si tomamos varios sistemas idénticos y observamos el número de tareas en la CPU en función del tiempo, encontraríamos que el número $n(t)$ es una variable aleatoria. Para especificar su comportamiento, necesitaríamos especificar la función de probabilidad para $n(t)$ para cada posible valor de t . Tales funciones aleatorias dependientes del tiempo se llaman **procesos estocásticos**. Estos procesos son útiles en la representación del estado de los sistemas de colas. Algunos de los procesos estocásticos más frecuentes usados en la teoría se explican a continuación.

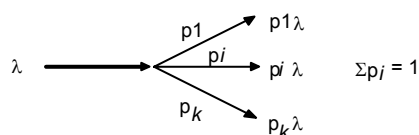
1. *Procesos de estado discreto y estado continuo*: Un proceso se llama de **estado discreto o continuo** dependiendo de los valores que su estado puede tomar.

Si el número de valores posibles es finito o contable, el proceso se denomina de estado discreto. Si no, será un proceso de estado continuo. Un proceso estocástico de estado discreto se llama también cadena estocástica.

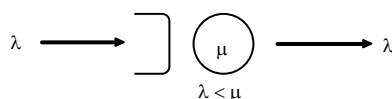
2. *Procesos de Markov*: Si el estado futuro de un proceso es independiente del pasado y depende sólo del presente, el proceso se llama **proceso de Markov**. Estos procesos tienen la ventaja de ser fáciles de analizar. Un proceso de estado discreto de Markov se llama **cadena de Markov**.



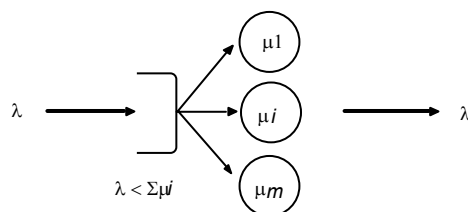
(a) La mezcla de flujos de Poisson es un flujo de Poisson.



(b) Un flujo de Poisson se puede dividir en flujos de Poisson.



(c) La partida de una cola M/M/1 es un proceso de Poisson.



(d) La partida de una cola M/M/m es un proceso de Poisson.

Figura 1.4 Propiedades de los procesos de Poisson.

3. *Procesos de nacimiento-muerte (birth-death)*: Los procesos de Markov en los cuales las transiciones están restringidas a sus estados vecinos se denominan **procesos de nacimiento-muerte**. Para estos procesos, se pueden representar los estados por enteros tales que un proceso en el estado n puede cambiar sólo al

estado $n+1$ ó $n-1$. Por ejemplo, el número de tareas en una cola con un único servidor y llegadas individuales puede representarse como un proceso de nacimiento-muerte.

4. *Procesos de Poisson*: Si los tiempos entre llegadas son IID (independientes e idénticamente distribuidos) y exponencialmente distribuidos, el número de llegadas n en un intervalo dado $(t, t+x)$ tiene una distribución de Poisson, y por tanto, el proceso de llegada se denomina **Proceso de Poisson** o **Flujo de Poisson**. Los flujos de Poisson tienen las siguientes propiedades:

- a) La mezcla de k flujos de Poisson con razón media λ_i resulta un flujo de Poisson con razón media λ dada por:

$$\lambda = \sum_{i=1}^k \lambda_i$$

como se puede ver en la figura 1.4a.

- b) Si un flujo de Poisson se divide en k subflujos tal que la posibilidad de que una tarea vaya al i -ésimo subflujo es p_i , cada subflujo es también de tipo Poisson con una razón media de $p_i\lambda_i$, como aparece en la figura 1.4b.
- c) Si las llegadas a un único servidor con tiempo de servicio exponencial son de tipo Poisson con razón media λ , las partidas son también Poisson con la misma razón λ , como aparece en la figura 1.4c, siempre que se cumpla que la razón de llegadas λ sea menor que la razón de servicio μ .
- d) Si las llegadas a una utilidad con m centros de servicio son Poisson con una razón media λ , las partidas constituyen un flujo de Poisson con la misma razón λ , siempre que la razón de llegadas λ sea menor que la razón de servicio total, $\sum \mu_i$, como vemos en la figura 1.4d.

CAPÍTULO 2

ANÁLISIS DE UNA COLA SIMPLE

El modelo de colas más simple es el que tiene una única cola. Este modelo se utiliza para analizar recursos individuales en sistemas computadores. Por ejemplo, si todas las tareas que esperan por la CPU en el sistema se guardan en una cola, la CPU puede modelarse usando los resultados que se aplican a las colas simples.

2.1 Procesos Nacimiento-Muerte

Un proceso de nacimiento-muerte es útil para modelar sistemas en los cuales las tareas llegan una de cada vez. La llegada de una nueva tarea cambia el estado del sistema al estado $n+1$. Se denomina un nacimiento. De la misma forma, la partida de una tarea cambia el sistema al estado $n-1$. Se denomina una muerte. En la figura 2.1 aparece el diagrama de transiciones de un proceso de nacimiento-muerte.

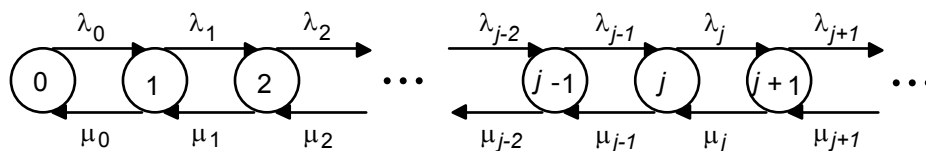


Figura 2.1 Diagrama de transición para un proceso de Nacimiento-Muerte.

Cuando el sistema está en el estado n , tiene n tareas en él. Las nuevas tareas que llegan lo hacen a una razón de λ_n . La razón de servicio es μ_n . Suponemos que ambos tiempos están exponencialmente distribuidos.

La probabilidad de que un proceso de nacimiento-muerte se encuentre en el estado estable n viene dado por el siguiente teorema:

Teorema 2.1. La probabilidad p_n de que un proceso de nacimiento-muerte esté en el estado estable n viene dado por:

$$p_n = \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_0 \mu_1 \cdots \mu_{n-1}} p_0, \quad n = 1, 2, \dots, \infty$$

Donde p_0 es la probabilidad de estar en el estado cero.

Demostración: Supongamos el sistema en el estado j en el tiempo t , existen j tareas en el sistema. En el siguiente intervalo de tiempo de corta duración Δt , el estado se puede mover al estado $j-1$ ó $j+1$ con las siguientes probabilidades:

$$P\{n(t+\Delta t) = j+1 \mid n(t) = j\} = \text{probabilidad de una llegada en } \Delta t = \lambda_j \Delta t$$

$$P\{n(t+\Delta t) = j-1 \mid n(t) = j\} = \text{probabilidad de una partida en } \Delta t = \mu_j \Delta t$$

Si no hay partidas ni llegadas, el sistema permanecerá en estado j , y así:

$$P\{n(t+\Delta t) = j \mid n(t) = j\} = 1 - \lambda_j \Delta t - \mu_j \Delta t$$

Suponemos que Δt es tan pequeño que no es posible que ocurran dos eventos del mismo signo durante el intervalo.

Si $p_j(t)$ denota la probabilidad de estar en el estado j en el tiempo t , podemos escribir las ecuaciones:

$$\begin{aligned} p_0(t + \Delta t) &= (1 - \lambda_0 \Delta t) p_0(t) + \mu_1 \Delta t p_1(t) \\ p_1(t + \Delta t) &= \lambda_0 \Delta t p_0(t) + (1 - \mu_1 \Delta t - \lambda_1 \Delta t) p_1(t) + \mu_2 \Delta t p_2(t) \\ p_2(t + \Delta t) &= \lambda_1 \Delta t p_1(t) + (1 - \mu_2 \Delta t - \lambda_2 \Delta t) p_2(t) + \mu_3 \Delta t p_3(t) \\ &\vdots \\ p_j(t + \Delta t) &= \lambda_{j-1} \Delta t p_{j-1}(t) + (1 - \mu_j \Delta t - \lambda_j \Delta t) p_j(t) + \mu_{j+1} \Delta t p_{j+1}(t) \end{aligned}$$

Para la j -ésima ecuación puede escribirse:

$$\begin{aligned} \lim_{t \rightarrow \infty} \frac{p_j(t + \Delta t) - p_j(t)}{\Delta t} &= \lambda_{j-1} p_{j-1}(t) - (\mu_j + \lambda_j) p_j(t) + \mu_{j+1} p_{j+1}(t) \\ \frac{dp_j(t)}{dt} &= \lambda_{j-1} p_{j-1}(t) - (\mu_j + \lambda_j) p_j(t) + \mu_{j+1} p_{j+1}(t) \end{aligned}$$

En estado estable $p_j(t)$ se aproxima a un valor fijo p_j , por tanto:

$$\lim_{t \rightarrow \infty} p_j(t) = p_j$$

y

$$\lim_{t \rightarrow \infty} \frac{dp_j(t)}{dt} = 0$$

Sustituyendo en la ecuación nos queda:

$$\begin{aligned}
0 &= \lambda_{j-1} p_{j-1} - (\mu_j + \lambda_j) p_j + \mu_{j+1} p_{j+1} \\
p_{j+1} &= \left(\frac{\mu_j + \lambda_j}{\mu_{j+1}} \right) p_j - \frac{\lambda_{j-1}}{\mu_{j+1}} p_{j-1} \quad j = 1, 2, 3, \dots \\
p_1 &= \frac{\lambda_0}{\mu_1} p_0
\end{aligned}$$

La solución a este conjunto de ecuaciones es:

$$\begin{aligned}
p_n &= \frac{\lambda_0 \lambda_1 \cdots \lambda_{n-1}}{\mu_1 \mu_2 \cdots \mu_n} p_0 \\
&= p_0 \prod_{j=0}^{n-1} \frac{\lambda_j}{\mu_{j+1}}, \quad n = 1, 2, \dots, \infty
\end{aligned}$$

Se puede calcular p_0 con la condición de que la suma de todas las probabilidades es igual a 1. Esto nos da:

$$p_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \sum_{j=0}^{n-1} [\lambda_j / \mu_{j+1}]}$$

2.2 Cola M/M/1

La cola M/M/1 es el tipo de cola usado con más frecuencia, se puede usar para modelar sistemas de un único procesador o para modelar dispositivos individuales dentro del sistema computador. Se supone que los tiempos entre llegadas y los tiempos de servicio están distribuidos exponencialmente y existe un único servidor. No existen limitaciones de buffers, ni de población y la disciplina de servicio es FCFS. Para analizar este tipo de cola sólo necesitamos conocer la razón entre llegadas λ y el tiempo medio de servicio μ .

El estado de la cola viene dado por el número de tareas en el sistema. Un diagrama de transición de estados aparece en la figura 2.2. Es similar al proceso nacimiento-muerte con las siguientes correspondencias:

$$\begin{aligned}
\lambda_n &= \lambda & n = 0, 1, 2, \dots, \infty \\
\mu_n &= \mu & n = 0, 1, 2, \dots, \infty
\end{aligned}$$

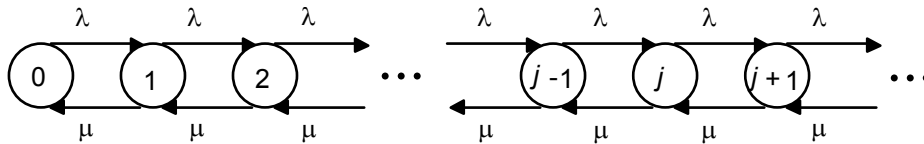


Figura 2.2 Diagrama de transición para una cola M/M/1.

El teorema visto en el apartado anterior nos da una expresión para la probabilidad de tener n tareas en el sistema:

$$p_n = \left(\frac{\lambda}{\mu}\right)^n p_0 \quad n = 1, 2, \dots, \infty$$

La cantidad λ/μ se denomina **intensidad de tráfico** y generalmente se denota con el símbolo ρ . Así:

$$p_n = \rho^n p_0$$

Como la suma de todas las probabilidades debe ser 1, tendremos:

$$p_0 = \frac{1}{1 + \rho + \rho^2 + \dots + \rho^\infty} = 1 - \rho$$

Y sustituyendo en la expresión de p_n :

$$p_n = (1 - \rho)\rho^n, \quad n = 0, 1, 2, \dots, \infty$$

A partir de las expresiones anteriores podemos derivar muchas otras propiedades de las colas M/M/1. Por ejemplo, la utilización del servidor viene dada por la probabilidad de tener una o más tareas en el sistema:

$$U = 1 - p_0 = \rho$$

El número medio de tareas en el sistema viene dado por:

$$E[n] = \sum_{n=1}^{\infty} n p_n = \sum_{n=1}^{\infty} n (1 - \rho) \rho^n = \frac{\rho}{1 - \rho}$$

El tiempo medio de respuesta puede calcularse utilizando la ley de Little, que establece:

$$\text{Número medio en el sistema} = \text{Razón de llegada} \times \text{Tiempo medio de respuesta}$$

Así:

$$E[n] = \lambda E[r]$$

de otra forma:

$$E[r] = \frac{E[n]}{\lambda} = \left(\frac{\rho}{1 - \rho}\right) \frac{1}{\lambda} = \frac{1/\mu}{1 - \rho}$$

Estas expresiones y otras más que se pueden derivar se recogen en la tabla resumen 2.1.

Cuando no hay tareas en el sistema se dice que el servidor está ocioso (*idle*); en otro caso el servidor está ocupado (*busy*). El intervalo de tiempo entre dos estados de inactividad se llama **periodo de trabajo**. El siguiente ejemplo ilustra la aplicación de los resultados obtenidos en el modelo de una pasarela de red.

Ejemplo 2.1 Las medidas efectuadas sobre una pasarela de red (gateway), han mostrado que los paquetes llegan a razón de 125 paquetes por segundo (pps) y la pasarela emplea 2 milisegundos en transmitirlos. Utilizando un modelo M/M/1, analizar la pasarela. ¿Cuál es la probabilidad de desbordar los buffers si la pasarela tiene sólo 13 buffers? ¿Cuántos buffers se necesitarán para perder menos de un paquete por millón?

Razón de llegada, $\lambda = 125$ pps.

Razón de servicio, $\mu = 1/0.002 = 500$ pps.

Utilización de la pasarela, $\rho = \lambda/\mu = 0.25$.

Probabilidad de n paquetes en la pasarela $= (1 - \rho)\rho^n = 0.75(0.25)^n$

Resumen 2.1 Cola M/M/1

1. Parámetros:
 λ = Razón de llegadas, en tareas por unidad de tiempo.
 μ = Razón de servicio, en tareas por unidad de tiempo.
2. Intensidad de tráfico: $\rho = \lambda/\mu$
3. Condición de estabilidad: La intensidad de tráfico, ρ , debe ser menor que 1.
4. La probabilidad de cero tareas en el sistema es: $p_0 = 1 - \rho$
5. Probabilidad de n tareas en el sistema: $p_n = (1 - \rho)\rho^n$, $n = 0, 1, \dots, \infty$
6. Número medio de tareas en el sistema: $E[n] = \rho / (1 - \rho)$
7. Varianza del número de tareas en el sistema: $Var[n] = \rho / (1 - \rho)^2$
8. Probabilidad de k tareas en la cola:

$$P(n_q = k) = \begin{cases} 1 - \rho^2, & k = 0 \\ (1 - \rho)\rho^{k+1}, & k > 0 \end{cases}$$
9. Número medio de tareas en la cola: $E[n_q] = \rho^2 / (1 - \rho)$
10. Varianza del número de tareas en la cola: $Var[n_q] = \rho^2 (1 + \rho - \rho^2) / (1 - \rho^2)$
11. Función de distribución acumulativa del tiempo de respuesta: $F(r) = 1 - e^{-r\mu(1-\rho)}$
12. Tiempo medio de respuesta: $E[r] = (1 / \mu) / (1 - \rho)$
13. Varianza del tiempo de respuesta: $Var[r] = \frac{1 / \mu^2}{(1 - \rho)^2}$
14. q-Percentil del tiempo de respuesta: $E[r] \ln[100/(100-q)]$
15. 90-Percentil del tiempo de respuesta: $2.3E[r]$
16. Función de distribución acumulativa del tiempo de espera: $F(w) = 1 - \rho e^{-\mu w(1-\rho)}$
17. Tiempo medio de espera: $E[w] = \rho \frac{1/\mu}{1-\rho}$
18. Varianza del tiempo de espera: $Var[w] = (2 - \rho)\rho / [\mu^2 (1 - \rho)^2]$
19. q-Percentil del tiempo de espera: $\max\left(0, \frac{E[w]}{\rho} \ln[100\rho / (100 - q)]\right)$
20. 90-Percentil del tiempo de espera: $\max\left(0, \frac{E[w]}{\rho} \ln[10\rho]\right)$
21. Probabilidad de encontrar n o más trabajos en el sistema: ρ^n

Resumen 2.1 Cola M/M/1

22. Probabilidad de servir n tareas en un periodo de trabajo: $\frac{1}{n} \binom{2n-2}{n-1} \frac{\rho^{n-1}}{(1+\rho)^{2n-1}}$

23. Número medio de tareas servidas en un periodo de trabajo: $1 / (1 - \rho)$

24. Varianza del número de tareas servidas en un periodo de trabajo:
 $\rho(1+\rho) / (1-\rho)^3$

25. Duración media del periodo de trabajo: $1 / [\mu(1-\rho)]$

26. Varianza del periodo de trabajo: $1 / [\mu^2(1-\rho)^3] - 1 / [\mu^2(1-\rho)^2]$

$$\text{Número medio de paquetes en la pasarela} = \frac{\rho}{1-\rho} = \frac{0.25}{0.75} = 0.33$$

$$\text{Tiempo medio empleado en la pasarela} = \frac{1/\mu}{1-\rho} = \frac{1/500}{1-0.25} = 2.66 \text{ milisegundos}$$

$$\begin{aligned} \text{Probabilidad de un overflow en el buffer} &= P(\text{más de 13 paquetes en la pasarela}) \\ &= \rho^{13} = 0.25^{13} = 1.49 \times 10^{-8} \\ &\approx 15 \text{ paquetes por billón de paquetes} \end{aligned}$$

Para limitar la probabilidad de pérdida a menos de 10^{-6} ,

$$\rho^n \leq 10^{-6}$$

o bien:

$$n > \log(10^{-6}) / \log(0.25) = 9.96$$

Necesitaremos en este caso 10 buffers.

Los últimos dos resultados son aproximados, pues para obtenerlos deberíamos emplear las fórmulas de las colas de tipo M/M/1/B, sin embargo, debido a la baja utilización los resultados son bastante aproximados.

En la figura 2.3 aparece la representación del tiempo de respuesta en función de la utilización de la pasarela. De la figura se desprende que para que el sistema sea estable, la intensidad de tráfico debe ser menor que uno.

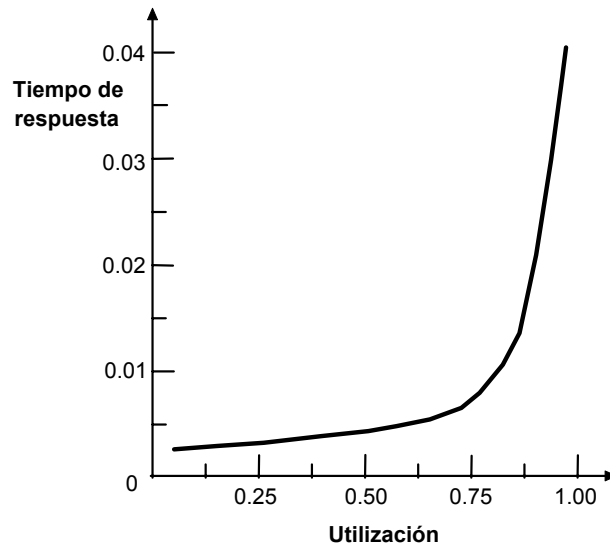


Figura 2.3 Tiempo de respuesta de la pasarela en función la utilización

2.3 Cola M/M/m

La cola M/M/m se puede usar para modelar sistemas multiprocesadores o dispositivos que tienen varios servidores idénticos y todas las tareas esperando por los servidores se mantienen en una misma cola. Se supone que existen m servidores cada uno con una razón de servicio de μ tareas por unidad de tiempo. Si alguno de los m servidores está libre, la tarea que llega se sirve inmediatamente. Si todos los servidores están ocupados, la tarea espera en la cola. El estado del sistema está representado por el número de tareas, n , en el sistema. En la figura 2.4 aparece un diagrama de los estados de transición.

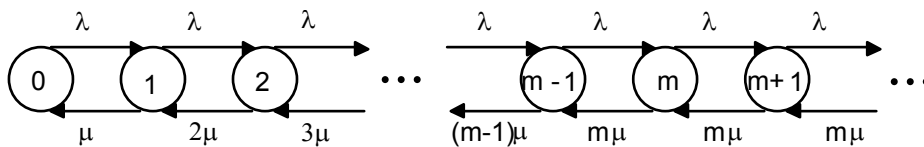


Figura 2.4 Diagrama de transición para una cola M/M/m.

Es fácil observar que el número de tareas en el sistema se asemeja a un proceso de nacimiento-muerte con la siguiente correspondencia:

$$\lambda_n = \lambda \quad n = 0, 1, 2, \dots, \infty$$

$$\mu_n = \begin{cases} n\mu & n = 1, 2, \dots, m-1 \\ m\mu & n = m, m+1, \dots, \infty \end{cases}$$

A partir del teorema 2.1 se obtiene la siguiente expresión para la probabilidad de tener n tareas en el sistema:

$$p_n = \begin{cases} \frac{\lambda^n}{n! \mu^n} p_0 & n = 1, 2, \dots, m-1 \\ \frac{\lambda^n}{m! m^{n-m} \mu^n} p_0 & n = m, m+1, \dots, \infty \end{cases}$$

A partir de esta expresión se pueden derivar otras. En la tabla resumen 2.2 aparecen recogidos estos resultados para la cola M/M/m.

ϕ denota la probabilidad de que una tarea que llega tenga que esperar en la cola. Se conoce como **fórmula C de Erlang**. Es un parámetro importante para las colas M/M/m y se recomienda calcularlo antes de proceder a calcular otros parámetros, pues aparece en varias expresiones.

Ejemplo 2.2 Los estudiantes llegan al centro de cálculo de la universidad siguiendo una distribución de Poisson, con un promedio de 10 por hora. Cada estudiante permanece un promedio de 20 minutos en el terminal y se puede considerar que el tiempo está exponencialmente distribuido. El centro actualmente tiene cinco terminales. Se han recibido quejas de que los tiempos de espera son demasiado largos. Analicemos el uso del centro mediante un modelo de colas.

El centro puede modelarse como una cola M/M/5 con una razón de llegada de $\lambda = 1/6$ por minuto y una razón de servicio de $\mu = 1/20$ por minuto, sustituyendo estos valores en las expresiones que tenemos para esta cola tenemos:

$$\text{Intensidad de tráfico, } \rho = \frac{\lambda}{m\mu} = \frac{0.167}{5 \times 0.05} = 0.67$$

Probabilidad de que todos los terminales estén libres:

$$\begin{aligned} p_0 &= \left[1 + \frac{(5 \times 0.67)^5}{5!(1-0.67)} + \frac{(5 \times 0.67)^1}{1!} + \frac{(5 \times 0.67)^2}{2!} + \right. \\ &\quad \left. + \frac{(5 \times 0.67)^3}{3!} + \frac{(5 \times 0.67)^4}{4!} \right]^{-1} \\ &= 0.0318 \end{aligned}$$

La probabilidad de que todos los terminales estén ocupados es:

$$\phi = \frac{(m\rho)^m}{m!(1-\rho)} p_0 = \frac{(5 \times 0.67)^5}{5!(1-0.67)} \times 0.0318 = 0.33$$

Utilización promedio, $\rho = 0.67$.

Número promedio de estudiantes en el centro:

$$E[n] = m\rho + \frac{\rho\phi}{1-\rho} = 5 \times 0.67 + \frac{0.67 \times 0.33}{1-0.67} = 4.0$$

El número promedio de estudiantes esperando en la cola es:

$$E[n_q] = \frac{\rho\phi}{1-\rho} = \frac{0.67 \times 0.33}{1-0.67} = 0.65$$

Resumen 2.2 Cola M/M/m

1. Parámetros:

λ = Razón de llegadas, en tareas por unidad de tiempo.

μ = Razón de servicio, en tareas por unidad de tiempo.

m = Número de servidores.

2. Intensidad de tráfico: $\rho = \lambda / (m\mu)$

3. Condición de estabilidad: La intensidad de tráfico, ρ , debe ser menor que 1.

4. La probabilidad de cero tareas en el sistema es:

$$p_0 = \left[1 + \frac{(m\rho)^m}{m!(1-\rho)} + \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!} \right]^{-1}$$

5. Probabilidad de n tareas en el sistema: $p_n = \begin{cases} p_0 \frac{(m\rho)^n}{n!}, & n < m \\ p_0 \frac{\rho^n m^m}{m!}, & n \geq m \end{cases}$

6. Probabilidad de estar en cola: $\varphi = P(\geq m \text{ tareas}) = \frac{(m\rho)^m}{m!(1-\rho)} p_0$

7. Número medio de tareas en el sistema: $E[n] = m\rho + \rho\varphi / (1-\rho)$

8. Varianza número de tareas en el sistema: $Var[n] = m\rho + \rho\varphi \left[\frac{1 + \rho + \rho\varphi}{(1-\rho)^2} + m \right]$

9. Número medio de tareas en la cola: $E[n_q] = \rho\varphi / (1-\rho)$

10. Varianza del número de tareas en la cola: $Var[n_q] = \varphi\rho(1 + \rho - \rho\varphi) / (1-\rho^2)$

11. Promedio de utilización de cada servidor: $U = \lambda / (m\mu) = \rho$

12. Función acumulativa del tiempo de respuesta:

$$F(r) = \begin{cases} 1 - e^{-\mu r} - \frac{\varphi}{1 - m + m\rho} e^{-m\mu(1-\rho)r} - e^{-\mu r}, & \rho \neq (m-1)/m \quad r > 0 \\ 1 - e^{-\mu r} - \varphi\mu r e^{-\mu r}, & \rho = (m-1)/m \end{cases}$$

13. Tiempo medio de respuesta: $E[r] = \frac{1}{\mu} \left(1 + \frac{\varphi}{m(1-\rho)} \right)$

14. Varianza del tiempo de respuesta: $Var[r] = \frac{1}{\mu^2} \left[1 + \frac{\varphi(2-\rho)}{m^2(1-\rho)^2} \right]$

15. Función de distribución acumulativa del tiempo de espera:

$$F(w) = 1 - \varphi e^{-m\mu(1-\rho)w}$$

Resumen 2.2 Cola M/M/m

16. Tiempo medio de espera: $E[w] = E[n_q] / \lambda = \varphi / [m\mu(1-\rho)]$
17. Varianza del tiempo de espera: $Var[w] = \varphi(2-\varphi) / [m^2\mu^2(1-\rho)^2]$
18. q-Percentil del tiempo de espera: $\max\left(0, \frac{E[w]}{\varphi} \ln \frac{100\varphi}{100-q}\right)$
19. 90-Percentil del tiempo de espera: $\frac{E[w]}{\varphi} \ln(10\varphi)$

El número promedio de estudiantes usando los terminales es:

$$E[n_s] = E[n] - E[n_q] = 4 - 0.65 = 3.35$$

La media y la varianza del tiempo empleado en el centro son:

$$E[r] = \frac{1}{\mu} \left(1 + \frac{\varphi}{m(1-\rho)} \right) = \frac{1}{0.05} \left(1 + \frac{0.33}{5(1-0.67)} \right) = 24$$

$$Var[r] = \frac{1}{\mu^2} \left(1 + \frac{\varphi(2-\varphi)}{m^2(1-\rho)^2} \right) = \frac{1}{0.05^2} \left(1 + \frac{0.33}{5^2(1-0.67)^2} \right) = 4.79$$

Así, cada estudiante pasa un promedio de 24 minutos en el centro, 20 minutos trabajando y 4 minutos esperando a la cola. Podemos calcular el 90 percentil del tiempo de espera.

$$\max\left\{0, \frac{E[w]}{\varphi} \ln(10\varphi)\right\} = \max\left\{0, \frac{4}{0.33} \ln(10 \times 0.33)\right\} = 14$$

el 10% de los estudiantes tendrán que esperar más de 14 minutos.

Los modelos de colas también se pueden emplear para predecir cómo evoluciona el sistema si se hacen cambios en él. Los siguientes ejemplos ilustran esta situación.

Ejemplo 2.3 Los estudiantes desearían limitar el tiempo de espera a un promedio de 2 minutos y no más de 5 minutos en el 90% de los casos. ¿Es posible? Si lo es, ¿cuántos terminales serían necesarios?

Analicemos el sistema con $m = 6, 7, \dots$ terminales manteniendo las mismas razones de llegada y de servicio, $\lambda = 0.167$ y $\mu = 0.05$ respectivamente.

Con $m = 6$ tendremos:

$$\text{Intensidad de tráfico, } \rho = \frac{0.167}{6 \times 0.05} = 0.556$$

$$\text{Probabilidad de tener todos los terminales libres: } \rho_0 = 0.0346$$

$$\text{Probabilidad de tener todas las terminales ocupadas: } \varphi = 0.15$$

$$\text{Tiempo promedio de espera: } E[w] = 1.1 \text{ minutos}$$

El 90-percentil del tiempo de espera es:

$$\max\left\{0, \frac{1.1}{0.15} \ln(10 \times 0.15)\right\} = \max(0, 3.0) = 3.0$$

Así, con uno o más terminales quedarían cubiertas las demandas.

Una decisión importante que hay que tomar cuando existe más de un servidor idéntico es si se tiene una cola para cada servidor o una cola para todos los servidores. Para llegadas de tipo Poisson y tiempos de servicio exponenciales la primera opción puede modelarse según m colas M/M/1, cada una con razón de llegada de λ/m . La segunda opción puede modelarse usando una cola M/M/ m con una razón de llegada de λ . Es fácil observar que la alternativa de una única cola es mejor, tal como se muestra en el siguiente ejemplo.

Ejemplo 2.4 Considérese que ocurriría si los cinco terminales del ejemplo 2.2 se dispusieran en cinco emplazamientos diferentes, por tanto, se necesitaría una cola para cada uno.

En este caso, el sistema puede modelarse como cinco colas M/M/1 separadas. La razón de llegada para cada terminal sería un quinto del total. Tomando $m = 1$, $\lambda = 0.167/5$ y $\mu = 0.05$ tenemos:

$$\text{Intensidad de tráfico, } \rho = \frac{0.0333}{0.05} = 0.67$$

El tiempo medio empleado en la sala de terminales es:

$$E[r] = \frac{1/\mu}{1-\rho} = \frac{1/0.05}{1-0.67} = 60$$

Y la varianza del tiempo empleado en la sala de terminales:

$$\text{Var}[r] = \frac{1/\mu^2}{(1-\rho)^2} = \frac{1/0.05^2}{(1-0.67)^2} = 3600$$

Comparando estos resultados con los obtenidos en el ejemplo 2.2, se puede ver la diferencia entre ellos y como una única cola es mejor.

En general, si todos los trabajos son idénticos, es mejor tener una única cola que tener múltiples colas.

Un caso especial de la cola M/M/ m es la cola M/M/ ∞ , con infinitos servidores. En esta cola, las tareas nunca tienen que esperar. El tiempo de respuesta es igual al de servicio. El tiempo medio de respuesta es igual al tiempo medio de servicio independientemente de la razón de llegada. Tales centros de servicio se conocen como **centros de retardo (delay centers)**. Un centro de retardo se puede usar para representar recursos dedicados, tales como terminales en sistemas de tiempo compartido. Las propiedades de estas colas pueden derivarse fácilmente de las colas M/M/ m .

2.3 Cola M/M/m/B con Número Finito de Buffers

Una cola M/M/m/B es similar a la cola M/M/m, excepto que el número de buffers B es finito. Una vez que los B buffers están llenos, las siguientes tareas que llegan se pierden. Se supone que B es mayor o igual que m; de otra forma, algunos servidores nunca trabajarían debido a la falta de buffers y el sistema funcionaría como una cola M/M/B/B.

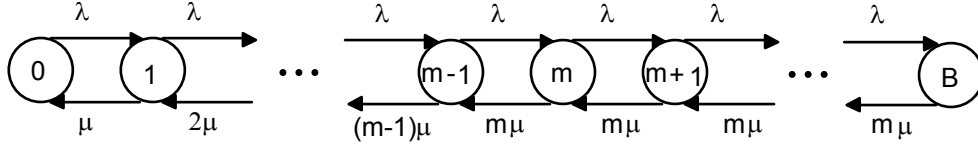


Figura 2.5 Diagrama de transición para una cola M/M/m/B.

En la figura 2.5 aparece el diagrama de transición de estados para una cola M/M/m/B. El sistema puede modelarse como un proceso de nacimiento-muerte usando las siguientes razones de llegada y servicio.

$$\lambda_n = \lambda, \quad n = 0, 1, 2, \dots, B-1$$

$$\mu_n = \begin{cases} n\mu, & n = 1, 2, \dots, m-1 \\ m\mu, & n = m, m+1, \dots, B \end{cases}$$

El teorema 2.1 nos da la siguiente expresión para la probabilidad de tener n tareas en el sistema:

$$p_n = \begin{cases} \frac{\lambda^n}{n! \mu^n} p_0, & n = 1, 2, \dots, m-1 \\ \frac{\lambda^n}{m! m^{n-m} \mu^n} p_0, & n = m, m+1, \dots, B \end{cases}$$

Estas fórmulas se pueden modificar en términos de intensidad de tráfico que se representa, $\rho = \lambda/m\mu$.

Todas las llegadas que ocurren cuando el sistema está en el estado $n = B$ se pierden. La razón de tareas que realmente entra en el sistema se denomina **razón de llegadas efectivas**, es:

$$\lambda' = \sum_{n=0}^{B-1} \lambda p_n = \lambda \sum_{n=0}^{B-1} p_n = \lambda(1 - p_B)$$

La diferencia $\lambda - \lambda' = \lambda p_B$ representa la razón de paquetes perdidos.

Los resultados obtenidos para la cola M/M/m/B se recogen en el resumen 2.3. Para el caso especial de un único servidor, muchos resultados pueden expresarse de forma más sencilla, este caso especial se recoge en el resumen 2.4. El siguiente ejemplo ilustra la aplicación de estos resultados.

Ejemplo 2.5 Considérese la pasarela de red (gateway) del ejemplo 2.1 de nuevo. Analicemos la pasarela suponiendo que sólo tiene dos buffers. La razón de llegadas y la razón de servicio como antes son 125 pps y 500 pps respectivamente. En este caso:

$$\lambda = 125, \quad \mu = 500, \quad m = 1, \quad B = 2.$$

Intensidad de tráfico, $\rho = \frac{\lambda}{m\mu} = \frac{125}{1 \times 500} = 0.25$

Para $n = 1, 2, \dots, B$ los valores de p_n :

$$p_1 = \rho p_0 = 0.25 p_0$$

$$p_2 = \rho^2 p_0 = 0.25^2 p_0 = 0.0625 p_0$$

p_0 se calcula sumando todas las probabilidades:

$$p_0 + p_1 + p_2 = 1 \Rightarrow p_0 + 0.25 p_0 + 0.0625 p_0 = 1 \Rightarrow$$

$$p_0 = \frac{1}{1 + 0.25 + 0.0625} = 0.76$$

Sustituyendo p_0 en p_n tenemos:

$$p_1 = 0.25 p_0 = 0.19$$

$$p_2 = 0.0625 p_0 = 0.0476$$

El número medio de tareas en el sistema es:

$$E[n] = \sum_{n=1}^B n p_n = 1 \times 0.19 + 2 \times 0.0476 = 0.29$$

El número medio de tareas en la cola es:

$$E[n_q] = \sum_{n=m}^B (n-m) p_n = (2-1) \times 0.0476 = 0.0476$$

La razón de llegadas efectivas al sistema es:

$$\lambda' = \lambda(1 - p_B) = 125(1 - p_2) = 125(1 - 0.0476) = 119 \text{ pps}$$

y la razón de paquetes perdidos: $\lambda - \lambda' = 125 - 119 = 6 \text{ pps}$

El tiempo medio de respuesta es:

$$E[r] = \frac{E[n]}{\lambda'} = \frac{0.29}{119} = 2.40 \times 10^{-3} \text{ segundos}$$

Resumen 2.3 Cola M/M/m/B

1. Parámetros:

λ = Razón de llegadas, en tareas por unidad de tiempo.

μ = Razón de servicio, en tareas por unidad de tiempo.

m = Número de servidores.

B = Número de buffers, $B \geq m$.

2. Intensidad de tráfico: $\rho = \lambda/(m\mu)$ 3. Condición de estabilidad: El sistema es siempre estable, $\rho < \infty$.

4. La probabilidad de cero tareas en el sistema es:

$$p_0 = \left[\frac{1 + (1 - \rho^{B-m+1})(m\rho)^m}{m!(1 - \rho)} + \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!} \right]^{-1}$$

Para $m = 1$:

$$p_0 = \begin{cases} \frac{1 - \rho}{1 - \rho^{B+1}}, & \rho \neq 1 \\ \frac{1}{B+1}, & \rho = 1 \end{cases}$$

5. Probabilidad de n tareas en el sistema: $p_n = \begin{cases} \frac{1}{n!}(m\rho)^n p_0, & 0 \leq n < m \\ \frac{m^m \rho^n}{m!} p_0, & m \leq n \leq B \end{cases}$ 6. Número medio de tareas en el sistema: $E[n] = \sum_{n=1}^B n p_n$

Para $m = 1$:

$$E[n] = \frac{\rho}{1 - \rho} - \frac{(B+1)\rho^{B+1}}{1 - \rho^{B+1}}$$

7. Número medio de tareas en la cola: $E[n_q] = \sum_{n=m+1}^B (n-m)p_n$

Para $m = 1$:

$$E[n_q] = \frac{\rho}{1 - \rho} - \rho \frac{1 + B\rho^B}{1 - \rho^{B+1}}$$

8. Tasa de llegada efectiva al sistema: $\lambda' = \sum_{n=0}^{B-1} \lambda p_n = \lambda(1 - p_B)$ 9. Promedio de utilización de cada servidor: $U = \lambda'/(m\mu) = \rho(1 - p_B)$ 10. Tiempo medio de respuesta: $E[r] = E[n] / \lambda' = E[n] / [\lambda(1 - p_B)]$ 11. Tiempo medio de espera: $E[w] = E[r] - 1/\mu = E[n_q] / [\lambda(1 - p_B)]$ 12. La tasa de pérdidas viene dada por λp_B tareas por unidad de tiempo.

Resumen 2.3 Cola M/M/m/B

13. Para una cola M/M/m la probabilidad de que el sistema esté lleno viene dada por:

$$p_m = \frac{(m\rho)^m / m!}{\sum_{j=0}^m \frac{(m\rho)^j}{j!}}$$

Resumen 2.4 Cola M/M/1/B (B buffers)

1. Parámetros:
 λ = Razón de llegadas, en tareas por unidad de tiempo.
 μ = Razón de servicio, en tareas por unidad de tiempo.
 B = Número de buffers, $B \geq m$.
2. Intensidad de tráfico: $\rho = \lambda/\mu$
3. Condición de estabilidad: El sistema es siempre estable, $\rho < \infty$.
4. La probabilidad de cero tareas en el sistema es:

$$p_0 = \begin{cases} \frac{1-\rho}{1-\rho^{B+1}}, & \rho \neq 1 \\ \frac{1}{B+1}, & \rho = 1 \end{cases}$$

5. Probabilidad de n tareas en el sistema:

$$p_n = \begin{cases} \frac{1-\rho}{1-\rho^{B+1}} \rho^n, & \rho \neq 1 \\ \frac{1}{B+1}, & \rho = 1 \end{cases} \quad \begin{matrix} 0 \leq n \leq B \\ n > B \end{matrix}$$

6. Número medio de tareas en el sistema:

$$E[n] = \frac{\rho}{1-\rho} - \frac{(B+1)\rho^{B+1}}{1-\rho^{B+1}}$$

7. Número medio de tareas en la cola:

$$E[n_q] = \frac{\rho}{1-\rho} - \rho \frac{1+B\rho^B}{1-\rho^{B+1}}$$

8. Razón efectiva de llegadas al sistema: $\lambda' = \sum_{n=0}^{B-1} \lambda p_n = \lambda(1-p_B)$

9. Tiempo medio de respuesta: $E[r] = E[n] / \lambda' = E[n] / [\lambda(1-p_B)]$

10. Tiempo medio de espera: $E[w] = E[r] - 1/\mu = E[n_q] / [\lambda(1-p_B)]$

El tiempo medio de espera en la cola es:

$$E[w] = \frac{E[n_q]}{\lambda'} = \frac{0.0476}{119} = 4.0 \times 10^{-4} \text{ segundos}$$

También se puede calcular la varianza y otros estadísticos para el número de tareas en el sistema, puesto que es conocida la función de probabilidad, por ejemplo:

$$Var[n] = E[n^2] - (E[n])^2 = (1^2 \times 0.19 + 2^2 \times 0.0476) - (0.29)^2 = 0.2963$$

2.4 Resultados para Otros Sistemas de Colas

La mayoría de los modelos de colas empleados en el análisis de prestaciones de sistemas computadores suponen tiempos de llegada exponenciales y tiempos de servicio exponenciales. Por tanto, los sistemas M/M/m vistos anteriormente cubren la mayoría de los casos. También se usan en algunas ocasiones sistemas con tiempos de llegada y tiempos de servicio de tipo general. Esto incluye las colas de tipo G/M/1, M/G/1, G/G/1 y G/G/m. Los principales resultados para estos sistemas y para el sistema M/D/1 que es un caso particular del M/G/1 se recogen en los resúmenes 2.5 a 2.10.

Resumen 2.5 Cola M/G/1

1. Parámetros:
 λ = Razón de llegadas, en tareas por unidad de tiempo.
 $E[s]$ = Tiempo medio de servicio por tarea.
 C_s = Coeficiente de variación del tiempo de servicio.
2. Intensidad de tráfico: $\rho = \lambda E[s]$
3. El sistema es estable si la intensidad de tráfico, ρ , es menor que 1.
4. La probabilidad de cero tareas en el sistema es: $p_0 = 1 - \rho$
5. Número medio de tareas en el sistema: $E[n] = \rho + \rho^2 (1 + C_s^2) / [2(1 - \rho)]$
6. Varianza del número de tareas en el sistema:

$$Var[n] = E[n] + \lambda^2 Var[s] + \frac{\lambda^3 E[s^3]}{3(1 - \rho)} + \frac{\lambda^4 (E[s^2])^2}{4(1 - \rho)^2}$$
7. Número medio de tareas en la cola: $E[n_q] = \rho^2 (1 - C_s^2) / [2(1 - \rho)]$
8. Varianza del número de tareas en la cola: $Var[n_q] = Var[n] - \rho + \rho^2$
9. Tiempo medio de respuesta: $E[r] = E[n] / \lambda = E[s] + \rho E[s] (1 + C_s^2) / [2(1 - \rho)]$
10. Varianza del tiempo de respuesta:

$$Var[r] = Var[s] + \lambda E[s^3] / [3(1 - \rho)] + \lambda^2 (E[s^2])^2 / [4(1 - \rho)^2]$$
11. Tiempo medio de espera: $E[w] = \rho E[s] (1 + C_s^2) / [2(1 - \rho)]$

Resumen 2.5 Cola M/G/1

12. Varianza del tiempo de espera: $Var[w] = Var[r] - Var[s]$
13. Distribución del tiempo libre: $F(I) = 1 - e^{-\lambda I}$ (distribución exponencial)
14. Número medio de tareas servidas en un periodo de trabajo: $1/(1-\rho)$
15. Varianza del número de tareas servidas en un periodo de trabajo:

$$\rho(1-\rho) + \lambda^2 E[s^2] / (1-\rho)^3$$
16. Duración media del periodo de trabajo: $E[s] / (1-\rho)$
17. Varianza del periodo de trabajo: $E[s^2] / (1-\rho)^3 - (E[s])^2 / (1-\rho)^2$

Las varianzas cambian para otras disciplinas de colas.

Resumen 2.6 Cola M/G/1 con procesador compartido (PS)

1. Parámetros:
 λ = Razón de llegadas, en tareas por unidad de tiempo.
 $E[s]$ = Tiempo medio de servicio por tarea.
2. Intensidad de tráfico: $\rho = \lambda E[s] < 1$
3. El sistema es estable si la intensidad de tráfico, ρ , es menor que 1.
4. Probabilidad de n tareas en el sistema: $p_n = (1-\rho)\rho^n$, $n = 0, 1, \dots, \infty$
5. Número medio de tareas en el sistema: $E[n] = \rho / (1-\rho)$
6. Varianza del número de tareas en el sistema: $Var[n] = \rho / (1-\rho)^2$
7. Tiempo medio de respuesta: $E[r] = E[s] / (1-\rho)$

Nótese que las expresiones dadas aquí son las mismas que las de la cola M/M/1. Las distribuciones son sin embargo diferentes. El procesador compartido aproxima el comportamiento de un scheduling round-robin con un tamaño de quantum pequeño y una sobrecarga despreciable.

Resumen 2.7 Cola M/D/1

1. Parámetros:
 λ = Razón de llegadas, en tareas por unidad de tiempo.
 $E[s]$ = Tiempo de servicio por tarea, s es constante.
2. Intensidad de tráfico: $\rho = \lambda E[s]$
3. Condición de estabilidad: El sistema es estable si $\rho < 1$.

Resumen 2.7 Cola M/D/1

4. Probabilidad de n tareas en el sistema:

$$p_n = \begin{cases} 1 - \rho, & n = 0 \\ (1 - \rho)(e^\rho - 1), & n = 1 \\ (1 - \rho) \sum_{j=0}^n \frac{(-1)^{n-j} (j\rho)^{n-j-1} (j\rho + n - j) e^{j\rho}}{(n-j)!}, & n \geq 2 \end{cases}$$

5. Número medio de tareas en el sistema: $E[n] = \rho + \rho^2 / E[2(1 - \rho)]$

6. Varianza del número de tareas en el sistema:

$$Var[n] = E[n] + \rho^3 / [3(1 - \rho)] + \rho^4 / [4(1 - \rho)^2]$$

7. Función de distribución acumulativa del tiempo de respuesta:

$$F(r) = p_n \frac{(r - nE[s])}{E[s]} + \sum_{j=0}^{n-1} p_j, \quad r \geq E[s] \quad y \quad n = \left\lfloor \frac{r}{E[s]} \right\rfloor$$

8. Tiempo medio de respuesta: $E[r] = E[s] + \rho E[s] / [2(1 - \rho)]$

9. Varianza del tiempo de respuesta:

$$Var[r] = \rho(E[s])^2 / [3(1 - \rho)] + \rho^2 (E[s])^2 / [4(1 - \rho)^2]$$

10. Número medio de tareas en la cola: $E[n_q] = \rho^2 / [2(1 - \rho)]$

11. Varianza del número de tareas en la cola:

$$Var[n_q] = \rho^2 + \frac{\rho^2}{2(1 - \rho)} + \frac{\rho^3}{3(1 - \rho)} + \frac{\rho^4}{4(1 - \rho)^2}$$

12. Tiempo medio de espera: $E[w] = \rho E[s] / [2(1 - \rho)]$

13. Varianza del tiempo de espera: $Var[w] = Var[r]$

14. Probabilidad de servir n tareas en un periodo de trabajo:

$$P(n) = \frac{(n\rho)^{n-1}}{n!} e^{-n\rho}$$

15. Función de distribución acumulativa del periodo de trabajo:

$$F(b) = \sum_{j=1}^n \frac{(j\rho)^{j-1}}{j!} e^{-j\rho}, \quad n = \left\lfloor \frac{b}{E[s]} \right\rfloor$$

Aquí $\lfloor x \rfloor$ es el mayor entero que excede de x .

Resumen 2.8 Cola M/G/ ∞

1. Parámetros:
 λ = Razón de llegadas, en tareas por unidad de tiempo.
 $E[s]$ = Tiempo medio de servicio por tarea.
2. Intensidad de tráfico: $\rho = \lambda E[s]$
3. El sistema es siempre estable: $\rho < \infty$
4. Probabilidad de cero tareas en el sistema es: $p_0 = e^{-\rho}$
5. Probabilidad de n tareas en el sistema: $p_n = (e^{-\rho} / n!) \rho^n$, $n = 0, 1, \dots, \infty$
6. Número medio de tareas en el sistema: $E[n] = \rho$
7. Varianza del número de tareas en el sistema: $Var[n] = \rho$
8. El número de tareas en la cola es siempre cero: $E[n_q] = 0$
9. Tiempo de respuesta igual al tiempo de servicio. Por tanto tiene la misma distribución que el tiempo de servicio.
10. Tiempo medio de respuesta: $E[r] = E[s]$

Para la cola M/M/ ∞ debería sustituirse, $E[s] = 1/\mu$ en los resultados anteriores.

Resumen 2.9 Cola G/M/1

1. Parámetros:
 $E[\tau]$ = Tiempo medio entre llegadas.
 μ = Razón de servicio por unidad de tiempo.
 $\phi = L_\tau(\mu - \phi\mu) = L_\tau$ Transformada de Laplace de la función de probabilidad.
2. Intensidad de tráfico: $\rho = 1/(E[\tau]\mu)$
3. El sistema es siempre estable si la intensidad de tráfico, ρ , es menor que 1.
4. Probabilidad de cero tareas en el sistema es: $p_0 = 1 - \rho$
5. Probabilidad de n tareas en el sistema: $p_n = \rho \phi^{n-1} (1 - \phi)$, $n = 1, 2, \dots, \infty$
6. Número medio de tareas en el sistema: $E[n] = \rho / (1 - \rho)$
7. Varianza del número de tareas en el sistema: $Var[n] = \rho(1 + \phi - \rho) / (1 - \phi)^2$
8. Función de distribución acumulativa del tiempo de respuesta:

$$F(r) = 1 - e^{(1-\phi)\mu r}, \quad r \geq 0$$
9. Tiempo medio de respuesta: $E[r] = 1 / [\mu(1 - \phi)]$
10. Varianza del tiempo de respuesta: $Var[r] = [1 / \{\mu(1 - \phi)\}]^2$

Resumen 2.9 Cola G/M/1

11. Función de distribución de probabilidad del tiempo de espera:

$$F(w) = 1 - \varphi e^{(1-\varphi)\mu w}, \quad w \geq 0$$

12. Tiempo medio de espera: $E[w] = \varphi / [\mu(1 - \varphi)]$

13. Varianza del tiempo de espera: $Var[w] = (2 - \varphi)\varphi / [\mu^2(1 - \varphi)^2]$

14. q-Percentil del tiempo de respuesta: $E[r] \ln[100 / (100 - q)]$

15. 90-Percentil del tiempo de respuesta: $E[r] \ln[10] = 2.3E[r]$

16. q-Percentil del tiempo de espera: $\max(0, (E[w] / \varphi) \ln[100\varphi / (100 - q)])$

17. 90-Percentil del tiempo de espera: $\max(0, (E[w] / \varphi) \ln(10\varphi))$

18. Probabilidad de encontrar n o más tareas en el sistema: $\varphi^n (1 - \rho) / (1 - \varphi)$

Para las llegadas de tipo Poisson, $\varphi = \rho$, y todas las fórmulas serán idénticas a las de

las colas M/M/1

Resumen 2.10 Cola G/G/m

1. Parámetros:

$E[\tau]$ = Tiempo medio entre llegadas.

λ = Razón de llegadas, en tareas por unidad de tiempo.

$E[s]$ = Tiempo medio de servicio por tarea.

μ = Razón de servicio, $= 1/E[s]$.

2. Intensidad de tráfico: $\rho = \lambda / (m\mu)$

3. El sistema es siempre estable si la intensidad de tráfico es menor que 1.

4. Número medio de tareas en servicio: $E[n_s] = m\rho$

5. Número medio de tareas en el sistema: $E[n] = E[n_q] + m\rho$

6. Varianza del número de tareas en el sistema: $Var[n] = Var[n_q] + Var[n_s]$

7. Tiempo medio de respuesta: $E[r] = E[w] + E[s]$. Alternativamente:

$$E[r] = E[n] / \lambda$$

8. Varianza del tiempo de respuesta: $Var[r] = Var[w] + Var[s]$

9. Tiempo medio de espera: $E[w] = E[n_q] / \lambda$

CAPÍTULO 3

REDES DE COLAS

Una tarea puede recibir servicio en una o mas colas antes de salir del sistema. Tales sistemas se modelan mediante redes de colas. En general, un modelo en el cual las tareas salen de una cola y entran en otra (o en la misma cola) se llama red de colas. En este capítulo se presentan varios conceptos básicos sobre redes de colas.

3.1 Redes de Colas Abiertas y Cerradas

A diferencia de las colas simples, no hay una notación sencilla para especificar el tipo de red de colas. La forma más simple de clasificar una red de colas es como abierta o cerrada. Una **red de colas abierta** tiene entradas y salidas con el exterior como aparece en la figura 3.1. Para analizar un sistema abierto suponemos que la productividad es conocida (igual que la razón de llegadas), y el objetivo es caracterizar la distribución del número de tareas en el sistema.

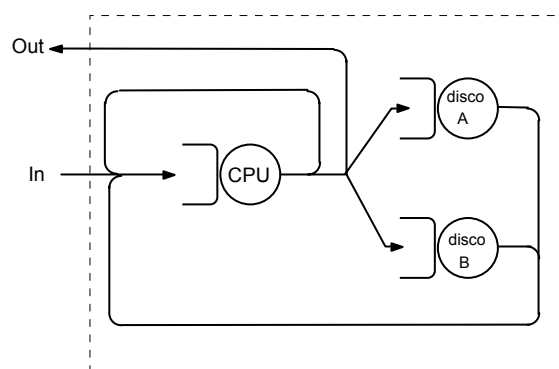


Figura 3.1 Red de colas abierta con entradas y salidas al exterior.

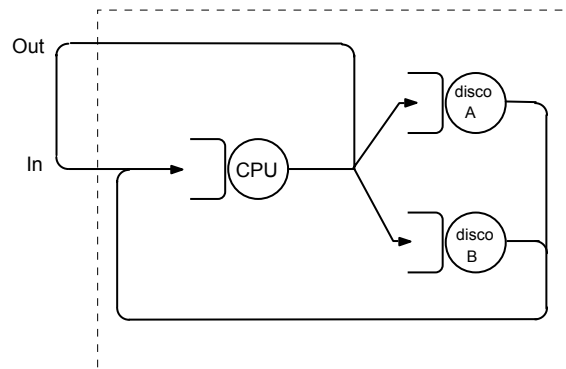


Figura 3.2 Red de colas cerradas sin contacto con el exterior.

Una **red de colas cerrada** no tiene relación con el exterior, como aparece en la figura 3.2. Las tareas se mantienen en el sistema circulando de una cola a la siguiente. El número total de tareas en el sistema es constante. El flujo de tareas en el enlace entre la entrada y la salida determina la productividad del sistema cerrado. Para analizar un sistema cerrado suponemos que el número de tareas viene dado y trataremos de determinar la productividad (razón de tareas que se completan).

También es posible tener una **red de colas mixta**, que será abierta para algunas cargas y cerrada para otras. La figura 3.3 muestra un conjunto de este tipo con dos *clases* de tareas. El sistema es cerrado para tareas interactivas y es abierto para tareas en batch. El término *clase* se refiere al tipo de tarea. Todas las tareas de una misma clase tienen la misma demanda de servicio y probabilidad de transición. Dentro de cada clase, las tareas son indistinguibles.

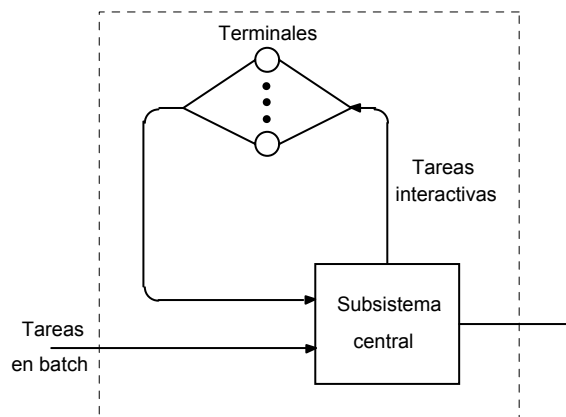


Figura 3.3 Red de colas mixta, abierta para unas tareas, cerrada para otras.

3.2 Redes de Tipo Producto

La red de colas más simple es una serie de M colas de un único servidor con tiempo de servicio exponencial y llegadas de tipo Poisson, como aparece en la figura 3.4. Las tareas que dejan una cola entran inmediatamente a la siguiente cola. Puede verse que cada cola individual puede analizarse independientemente de las otras colas. Cada cola

tiene una razón de llegadas y de partidas λ . Si μ_i es la razón de servicio para el servidor i :

Utilización del servidor i , $\rho_i = \lambda/\mu_i$

Probabilidad de n_i tareas en la i -ésima cola $= (1 - \rho_i) \rho_i^{n_i}$

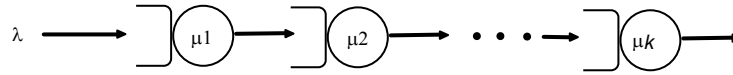


Figura 3.4 Una red de colas simple formada por colas M/M/1.

La probabilidad conjunta de longitud de cola para M colas puede calcularse simplemente multiplicando las probabilidades individuales; por ejemplo:

$$\begin{aligned} P(n_1, n_2, n_3, \dots, n_M) &= (1 - \rho_1) \rho_1^{n_1} (1 - \rho_2) \rho_2^{n_2} (1 - \rho_3) \rho_3^{n_3} \dots (1 - \rho_M) \rho_M^{n_M} \\ &= p_1(n_1) p_2(n_2) p_3(n_3) \dots p_M(n_M) \end{aligned}$$

Esta red de colas es una **red en forma de producto**. En general, el término se aplica a cualquier red de colas en la que la probabilidad tenga la siguiente expresión:

$$P(n_1, n_2, \dots, n_M) = \frac{1}{G(M)} \prod_{i=1}^M f_i(n_i)$$

Donde $f_i(n_i)$ es función del número de tareas en el servidor i , $G(N)$ es una constante de normalización y es función del número total de tareas en el sistema.

Las redes en forma de producto son más fáciles de analizar que aquellas que no son en forma de producto. El conjunto de redes que tienen solución en forma de producto es continuamente aumentado por los investigadores.

En un principio este método se aplicó a redes abiertas con colas de m servidores con tiempos de servicio exponencialmente distribuidos (Jackson 1963). Posteriormente se demostró que cualquier red de colas cerrada de m servidores con tiempos de servicio exponencialmente distribuidos tienen también solución en forma de producto (Gordon y Newell 1967).

Baskett, Chandy, Muntz y Palacios (1975) demostraron que la solución en forma de producto existe para una clase más amplia de redes. Esta clase está formada por las redes que satisfacen los siguientes criterios:

1. *Disciplinas de servicio*: Todos los centros de servicio tendrán una de las cuatro disciplinas siguientes: primero que llega, primero en ser servido (FCFS), procesador compartido (PS), servidores infinitos (IS o delay centers), y último en llegar primero en ser servido con borrado de la prioridad (LCFS-PR).
2. *Clases de tareas*: Las tareas pertenecen a una clase mientras esperan o reciben servicio en un centro de servicio, pero pueden cambiar de clase y de centro de servicio según unas probabilidades fijas cuando se complete una petición de servicio.
3. *Distribución del tiempo de servicio*: En los centros de servicio del tipo FCFS, las distribuciones del tiempo de servicio deben ser idénticas y exponenciales para

todas las clases de tareas. En otros centros de servicio, diferentes clases de tareas pueden tener diferentes distribuciones.

4. *Servicio dependiente del estado*: El tiempo de servicio en un centro FCFS puede depender sólo de la longitud total de la cola en el centro. El tiempo de servicio para una clase en centros PS, LCFS-PR e IS puede depender de la longitud de la cola para cada clase, pero no de la longitud de las colas de otras clases. Además, la razón de servicio global de una subred puede depender del número total de tareas en la subred.
5. *Proceso de llegada*: En redes abiertas, el tiempo entre las llegadas sucesivas de una clase debería estar distribuido exponencialmente. No están permitidas las llegadas en grupo. La razón de llegada puede depender del estado del sistema. Una red puede ser abierta con respecto a alguna clase de tareas y cerrada con respecto a otra clase de tareas.

Las redes que satisfacen estos criterios se conocen como **redes BCMP**, debido a los autores de los criterios.

Denning y Buzen (1978) extendieron las redes en forma de producto a las redes no Markovianas con las siguientes condiciones:

1. *Balance del flujo de tareas*: Para cada clase, el número de llegadas a un dispositivo debe igualar al número de salidas del dispositivo.
2. *Comportamiento de un paso*: Sólo se puede producir un cambio de estado si una tarea entra en el sistema, se mueve entre pares de dispositivos del sistema o sale del sistema. Esta suposición asegura que no se producirán movimientos simultáneos de tareas.
3. *Homogeneidad de los dispositivos*: La razón de servicio de un dispositivo para una clase particular no depende del estado del sistema de ninguna forma, excepto para la longitud total de la cola del dispositivo y la longitud de la cola de la clase dada. Estas suposiciones implican lo siguiente:
 - a) *Posesión de un único recurso*: Una tarea no puede estar presente (esperando o en servicio) en dos o más dispositivos al mismo tiempo.
 - b) *No bloqueo*: Un dispositivo presta servicio siempre que existan tareas; su capacidad de prestar servicio no está controlada por ningún otro dispositivo.
 - c) *Comportamiento independiente de las tareas*: La interacción entre las tareas está limitada a colas para dispositivos físicos; por ejemplo, no debería existir ningún requerimiento de sincronización.
 - d) *Información local*: La razón de servicio de un dispositivo depende sólo de la longitud de la cola local y no del estado del resto del sistema.
 - e) *Servicio adecuado*: Si la razón de servicio difiere por clase, la razón de servicio para una clase depende sólo de la longitud de cola de esa clase en el dispositivo y no de las longitudes de colas de otras clases. Esto significa que los servidores no discriminan entre las tareas de una clase dependiendo de la longitud de cola de las otras clases.
 - f) *Homogeneidad del routing*: La ruta de la tarea debería ser independiente del estado. La ruta significa el camino seguido por la tarea dentro de la red.

3.3 Modelos de Redes de Colas de Sistemas Computadores

Dos de los primeros modelos de colas para sistemas computadores fueron el "modelo de reparación" y el "modelo de servidor central", ambos aparecen en las figuras 3.5 y 3.6 respectivamente.

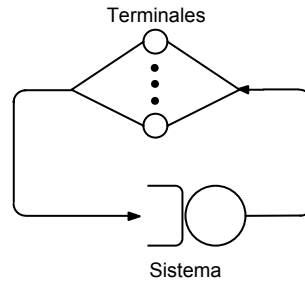


Figura 3.5 Modelo de reparación para un sistema computador.

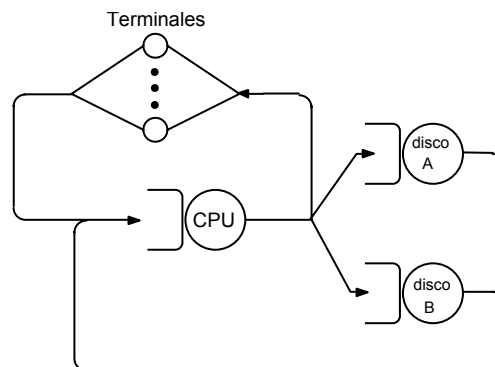


Figura 3.6 Modelo de servidor central para un sistema de tiempo compartido.

El modelo de reparación, como su nombre indica, fue diseñado originalmente para talleres de reparaciones. Se tenía un número de máquinas y un taller con uno o más técnicos. Cada vez que una máquina se averiaba, se ponía a la cola para reparar y se reparaba tan pronto como hubiera un técnico disponible. Este modelo se usó para representar un sistema de tiempo compartido con m terminales.

El modelo de servidor central puede explicarse con el siguiente ejemplo, la CPU es el "servidor central" que gestiona los trabajos de otros dispositivos. Después de un servicio en los dispositivos de E/S, la tarea regresa a la CPU para continuar procesándose y lo deja cuando encuentra la siguiente E/S o cuando la tarea esté completa.

En el modelado de sistemas computadores aparecen tres clases de dispositivos. La mayoría de los dispositivos tienen un único servidor cuyo tiempo de servicio no depende del número de tareas en la cola. Estos dispositivos se llaman **centros de servicio de capacidad fija**. Por ejemplo, la CPU en un sistema puede modelarse como un centro de servicio de capacidad fija.

Existen además dispositivos que no tienen cola, y las tareas emplean el mismo tiempo en el dispositivo independientemente del número de tareas en él. Estos dispositivos pueden modelarse como centros con infinitos servidores y se llaman **centros de retardo** ó **servidores infinitos (delay centers ó IS)**. Un grupo de terminales dedicados pueden modelarse como un centro de retardo.

Por último, el resto de dispositivos se denominan **centros de servicio dependientes de la carga**, puesto que sus razones de servicio pueden depender de la carga o el número de tareas en el dispositivo. Una cola $M/M/m$ (con $m > 2$) es un ejemplo de centro de servicio dependiente de la carga. Su razón de servicio se incrementa tanto más cuantos más servidores se utilicen. Por ejemplo, un grupo de enlaces paralelos entre dos nodos podría ser un ejemplo de este tipo de dispositivos

CAPÍTULO 4

LEYES OPERACIONALES

Un gran número de los problemas cotidianos en el análisis de prestaciones de sistemas computadores pueden resolverse empleando alguna relación sencilla que no requiere hipótesis sobre la distribución de los tiempos de servicio o intervalos de llegada. Son lo que se conoce como **leyes operacionales**.

La palabra *operacional* significa medido directamente. Así, suposiciones *comprobadas operacionalmente* son las suposiciones que se pueden verificar por medición. Por ejemplo, es fácil verificar la suposición de que el número de llegadas es igual al número de trabajos completados en un sistema particular. Esta suposición conocida como **balance del flujo de tareas** es una suposición operacionalmente comprobable.

Las *cantidades operacionales* son las cantidades que se pueden medir directamente durante un periodo de observación finito. Por ejemplo, visto el sistema como una caja negra, si observamos el dispositivo i durante un tiempo finito T , podemos medir el número de llegadas A_i , el número de tareas completadas C_i , y el tiempo de ocupación B_i durante ese periodo. Todas estas son cantidades operacionales. A partir de ellas se pueden derivar las siguientes cantidades operacionales:

$$\text{Razón de llegada } \lambda_i = \frac{\text{número de llegadas}}{\text{tiempo}} = \frac{A_i}{T}$$

$$\text{Productividad } X_i = \frac{\text{número de tareas completas}}{\text{tiempo}} = \frac{C_i}{T}$$

$$\text{Utilización } U_i = \frac{\text{tiempo de ocupación}}{\text{tiempo total}} = \frac{B_i}{T}$$

$$\text{Tiempo medio de servicio } S_i = \frac{\text{tiempo total de servicio}}{\text{número de servicios}} = \frac{B_i}{C_i}$$

Estas cantidades operacionales son *variables* que pueden cambiar de un periodo de observación al siguiente, pero existen ciertas relaciones que se mantienen durante todo

el periodo de observación. Estas relaciones se denominan **leyes operacionales**. A lo largo de este capítulo se verán algunas de estas leyes.

4.1 Ley de Utilización

Dado un número de tareas completas C_i y el tiempo de ocupación B_i de un dispositivo i durante un periodo de observación T , se mantiene la siguiente relación entre estas variables:

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

ó

$$U_i = X_i S_i$$

Esta relación se conoce como **ley de utilización**.

Ejemplo 4.1 Consideremos de nuevo el ejemplo de la pasarela de red del ejemplo 2.1. Los paquetes llegaban a razón de 125 paquetes por segundo (pps) y la pasarela emplea un promedio de 2 milisegundos en atenderlos.

Productividad $X_i = \text{razón de salida} = \text{razón de llegada} = 125 \text{ pps}$

Tiempo de servicio $S_i = 0.002 \text{ segundos}$

Utilización $U_i = X_i S_i = 125 \times 0.002 = 0.25 = 25\%$

Los resultados en este caso son válidos sin necesidad de realizar suposiciones sobre la distribución de las variables.

4.2 Ley de Flujo Forzado

La ley de flujo forzado relaciona la productividad del sistema con las productividades de dispositivos individuales. En un modelo abierto, el número de tareas que dejan el sistema por unidad de tiempo define la productividad del sistema. En un modelo cerrado, la productividad del sistema se define como el número de tareas que atraviesan el enlace entre el último dispositivo y el primero por unidad de tiempo.

Si nuestro periodo de observación T es tal que el número de tareas que llegan a cada dispositivo es el número de tareas completadas, es decir:

$$A_i = C_i$$

podemos decir que el dispositivo satisface la suposición de balance de flujo de tareas. Si el periodo de observación T es largo, la diferencia $A_i - C_i$, es generalmente pequeña comparada con C_i .

Supongamos que cada tarea realiza V_i peticiones para el dispositivo i en el sistema, como aparece en la figura 4.1. Si el flujo de tareas está equilibrado, el número de tareas

C_0 que atraviesan el enlace exterior y el número de tareas C_i que visitan el dispositivo i están relacionados por:

$$C_i = C_0 V_i \quad \text{ó} \quad V_i = \frac{C_i}{C_0}$$

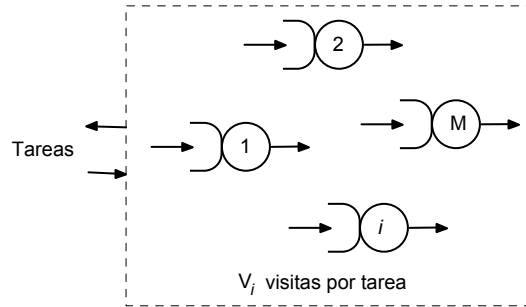


Figura 4.1 Vista interna de un sistema

Así, la variable V_i es la razón de visitas al dispositivo i y el enlace exterior. Se denomina por tanto **razón de visitas**. La productividad del sistema durante el periodo de observación es:

$$\text{Productividad del sistema } X = \frac{\text{tareas completadas}}{\text{tiempo total}} = \frac{C_0}{T}$$

La productividad del dispositivo i y la productividad del sistema están relacionados como sigue:

$$\text{Productividad del dispositivo } X_i = \frac{C_i}{T} = \frac{C_i}{C_0} \times \frac{C_0}{T}$$

En otras palabras:

$$X_i = X V_i \quad (4.1)$$

Esta es la **ley de flujo forzado**, se aplica siempre que la suposición de flujo equilibrado de tareas es cierta.

Combinando la ley de flujo forzado y la ley de utilización, obtenemos:

$$\begin{aligned} \text{Utilización del dispositivo, } U_i &= X_i S_i \\ &= X V_i S_i \end{aligned}$$

ó

$$U_i = X D_i \quad (4.2)$$

Aquí $D_i = V_i S_i$ es la demanda total de servicio en el dispositivo para todas las visitas de una tarea. La ecuación 4.2 establece que la utilización de un dispositivo es proporcional a la demanda total D_i . Así por tanto, el dispositivo con una mayor D_i tiene la más alta utilización y es un **dispositivo cuello de botella**.

Ejemplo 4.2 En un sistema de tiempo compartido, se registra el siguiente perfil de los programas de usuario. Cada programa requiere 5 segundos de tiempo de CPU y hace 80 peticiones de E/S al disco A y 100 peticiones E/S al disco B. El tiempo promedio de reflexión de los usuarios fue de 18 segundos. De las especificaciones de los dispositivos

sabemos que el disco A emplea 50 milisegundos para satisfacer una petición de E/S y el disco B emplea 30 milisegundos por petición. Con 17 terminales activas, se observó que la productividad del disco A es de 15.70 peticiones E/S por segundo. Queremos calcular la productividad del sistema y la utilización de los dispositivos.

Este sistema puede representarse por un modelo de redes de colas. Tenemos los siguientes valores $D_{CPU} = 5$ segundos, $V_A = 80$, $V_B = 100$, $Z = 18$ segundos, $S_A = 0.050$ segundos, $S_B = 0.030$ segundos, $N = 17$, y $X_A = 15.70$ tareas/segundos.

Puesto que las tareas deben visitar la CPU antes de ir a los discos o las terminales, la razón de visitas de la CPU es:

$$V_{CPU} = V_A + V_B + 1 = 181$$

El primer paso en el análisis operacional generalmente es determinar las demandas de servicio total D_i para todos los dispositivos. En este caso:

$$D_{CPU} = 5 \text{ segundos}$$

$$D_A = S_A V_A = 0.050 \times 80 = 4 \text{ segundos}$$

$$D_B = S_B V_B = 0.030 \times 100 = 3 \text{ segundos}$$

Empleando la ley de flujo forzado, los rendimientos serán:

$$X = \frac{X_A}{V_A} = \frac{15.70}{80} = 0.1963 \text{ tareas/segundo}$$

$$X_{CPU} = X V_{CPU} = 0.1963 \times 181 = 35.48 \text{ tareas/segundo}$$

$$X_B = X V_B = 0.1963 \times 100 = 19.6 \text{ tareas/segundo}$$

Empleando la ley de utilización, las utilizaciones de los dispositivos son:

$$U_{CPU} = X D_{CPU} = 0.1963 \times 5 = 98\%$$

$$U_A = X D_A = 0.1963 \times 4 = 78.5\%$$

$$U_B = X D_B = 0.1963 \times 3 = 58.8\%$$

Las razones de visita son una forma de especificar la distribución de las tareas en las redes de colas. Otra forma es especificar las **probabilidades de transición**, p_{ij} de que una tarea se mueva a la cola j después de completar el servicio en la cola i . Las razones de visita y las probabilidades de transición son equivalentes en el sentido de que conocido uno es conocido el otro. En un sistema con flujo de tareas equilibrado:

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

El subíndice 0 denota las visitas al enlace externo. Así, p_{i0} es la probabilidad de que una tarea salga del sistema después de completar el servicio en el dispositivo i . Dividiendo en ambos lados de la ecuación por C_0 , obtenemos:

$$V_j = \sum_{i=0}^M V_i p_{ij} \quad (4.3)$$

Como cada visita al enlace externo se define como la finalización de la tarea, tenemos:

$$V_0 = 1 \quad (4.4)$$

Las ecuaciones (4.3) y (4.4) se conocen como **ecuaciones de razones de visita** y se pueden utilizar para conseguir las razones de visita a partir de las probabilidades de transición.

En modelos de servidor central, después de completar un servicio en las colas, la tarea siempre regresa a la cola de la CPU:

$$\begin{aligned} p_{i1} &= 1 & \forall i \neq 1 \\ p_{ij} &= 0 & \forall i, j \neq 1 \end{aligned}$$

Estas probabilidades se aplican también a la entrada y la salida del sistema ($i=0$). Por tanto las ecuaciones de razones de visita resultan:

$$\begin{aligned} 1 &= V_1 p_{10} \\ V_1 &= 1 + V_2 + V_3 + \dots + V_M \\ V_j &= V_1 p_{1j} = \frac{p_{1j}}{p_{10}}, & j = 2, 3, \dots, M \end{aligned}$$

Ejemplo 4.3 Consideremos la red de colas del ejemplo anterior, las razones de visita son, $V_A = 80$, $V_B = 100$ y $V_{CPU} = 181$.

Es fácil ver en este caso que después de completar el servicio en la CPU, las probabilidades de que una tarea se mueva al disco A, disco B o las terminales son, $\frac{80}{181}$, $\frac{100}{181}$ y $\frac{1}{181}$ respectivamente. Por tanto las probabilidades de transición son: 0.4420, 0.5525 y 0.005525.

Dadas las probabilidades de transición, podemos encontrar las razones de visita dividiendo estas probabilidades por las probabilidades de salida (0.005525):

$$\begin{aligned} V_A &= \frac{0.4420}{0.005525} = 80 \\ V_B &= \frac{0.5525}{0.005525} = 100 \\ V_{CPU} &= 1 + V_A + V_B = 1 + 80 + 100 = 181 \end{aligned}$$

4.3 Ley de Little

La única suposición requerida en este caso es que el número de llegadas sea igual al número de tareas completadas. Esta es la suposición operacionalmente comprobable de flujo equilibrado de tareas.

Podemos aplicar la ley de Little para relacionar la longitud de la cola Q_i y el tiempo de respuesta R_i en el dispositivo i :

Número medio en el dispositivo = razón de llegada x tiempo medio en el dispositivo

Si el flujo de tareas está equilibrado, la razón de llegada es igual a la productividad, y podemos escribir:

$$Q_i = X_i R_i \quad (4.5)$$

Q_i , es la longitud de la cola, que es sinónimo del número de tareas en el dispositivo i . Incluye no sólo las tareas esperando en la cola, sino también las que están recibiendo servicio.

Ejemplo 4.4 La longitud promedio de la cola en el sistema computador del ejemplo 4.2, se observó que era 8.88, 3.19 y 1.40 tareas en la CPU, disco A y disco B respectivamente. ¿Cuáles serán los tiempos de respuesta de estos dispositivos?

En el ejemplo 4.2 las productividades calculadas fueron:

$$X_{CPU} = 35.48, \quad X_A = 15.70, \quad X_B = 19.6$$

Usando la ley de Little, los tiempos de respuesta de los dispositivos son:

$$R_{CPU} = Q_{CPU} / X_{CPU} = 8.88 / 35.48 = 0.250 \text{ segundos}$$

$$R_A = Q_A / X_A = 3.19 / 15.70 = 0.203 \text{ segundos}$$

$$R_B = Q_B / X_B = 1.40 / 19.6 = 0.071 \text{ segundos}$$

4.4 Ley General del Tiempo de Respuesta

Todos los sistemas en tiempo compartido pueden dividirse en dos subsistemas: el subsistema de terminales y el subsistema central que contiene los restantes elementos, incluyendo la CPU, como aparece en la figura 4.2. Existe un terminal por usuario y el resto del sistema es compartido por todos los usuarios.

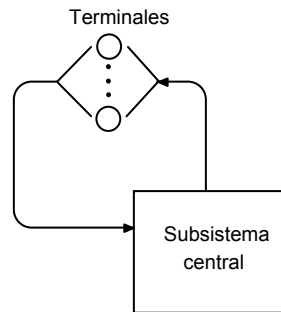


Figura 4.2 Los dos componentes de un sistema en tiempo compartido.

La ley de Little puede aplicarse a cualquier parte del sistema, el único requerimiento es que el flujo de tareas en esa parte esté equilibrado. En particular, puede aplicarse al subsistema central, tendremos:

$$Q = XR$$

Aquí, Q es el número total de tareas en el sistema, R el tiempo de respuesta del sistema, y X la productividad del sistema. Dadas las longitudes de las colas individuales Q_i en los dispositivos, podemos calcular Q :

$$Q = Q_1 + Q_2 + \dots + Q_M$$

Sustituyendo Q_i por la ecuación (4.5) tenemos:

$$XR = X_1 R_1 + X_2 R_2 + \dots + X_M R_M$$

Dividiendo en ambos lados de la ecuación por X y aplicando la ley de flujo forzado tenemos:

$$R = V_1 R_1 + V_2 R_2 + \dots + V_M R_M$$

ó

$$R = \sum_{i=1}^M R_i V_i \quad (4.6)$$

Se conoce esta expresión como **ley general del tiempo de respuesta**. Es posible demostrar que esta ley se cumple aunque el flujo no esté equilibrado.

Ejemplo 4.5 Calculemos el tiempo de respuesta para el sistema de tiempo compartido de los ejemplos 4.2 y 4.4. Para este sistema:

$$\begin{aligned} V_{CPU} &= 181, & V_A &= 80, & V_B &= 100 \\ R_{CPU} &= 0.250, & R_A &= 0.203, & R_B &= 0.071 \end{aligned}$$

El tiempo de respuesta del sistema es:

$$\begin{aligned} R &= R_{CPU} V_{CPU} + R_A V_A + R_B V_B \\ &= 0.250 \times 181 + 0.203 \times 80 + 0.071 \times 100 = 68.6 \end{aligned}$$

4.5 Ley del Tiempo de Respuesta Interactivo

En un sistema interactivo, los usuarios generan peticiones que son servidas por el subsistema central y la respuesta regresa a la terminal. Después de un tiempo de reflexión Z , los usuarios envían la siguiente petición. Si el tiempo de respuesta es R , el tiempo de ciclo total para la petición es $R+Z$. Cada usuario genera $T/(R+Z)$ peticiones en el periodo T . Si existen N usuarios:

$$\begin{aligned} \text{Productividad del sistema } X &= \frac{\text{Número total de peticiones}}{\text{tiempo total}} \\ &= \frac{N[T/(R+Z)]}{T} \\ &= \frac{N}{R+Z} \end{aligned}$$

ó

$$R = (N/X) - Z \quad (4.7)$$

Esta es la **ley del tiempo de respuesta interactivo**.

Ejemplo 4.6 Para el sistema en tiempo compartido del ejemplo 4.2, podemos calcular el tiempo de respuesta usando la ley de tiempo de respuesta interactivo de la forma siguiente:

$$X = 0.1963, \quad N = 17, \quad Z = 18$$

Por tanto,

$$R = \frac{N}{X} - Z = \frac{17}{0.1963} - 18 = 86.6 - 18 = 68.6 \text{ segundos}$$

que es el mismo obtenido en el ejemplo 4.5.

4.6 Análisis de Cuellos de Botella

Una consecuencia de la ley de flujo forzado es que la utilización del dispositivo es proporcional a la demanda total del servicio:

$$U_i = \propto D_i$$

El dispositivo con la mayor demanda de servicio D_i tiene la mayor utilización y se denomina **cuello de botella (bottleneck)**. (Los centros de retardo nunca pueden ser cuellos de botella). Este dispositivo será el factor límite en el rendimiento del sistema, cualquier mejora sobre este dispositivo repercutirá en una mejora global de las prestaciones del sistema, en mayor medida que una mejora en otro dispositivo. Por tanto, *la identificación del dispositivo cuello de botella debería ser el primer paso en cualquier proyecto de mejora de prestaciones*.

Si el dispositivo b es el cuello de botella, implicará que $D_b = D_{max}$ es el mayor entre D_1, D_2, \dots, D_M . La productividad y el tiempo de respuesta del sistema está limitado por:

$$X(N) \leq \min \left\{ \frac{1}{D_{max}}, \frac{N}{D + Z} \right\} \quad (4.8)$$

y

$$R(N) \geq \max \{ D, ND_{max} - Z \} \quad (4.9)$$

Aquí, $D = \sum D_i$ es la suma de las demandas totales de servicio en todos los dispositivos excepto terminales. Las ecuaciones (4.8) y (4.9) se conocen como **fronteras asintóticas**.

Demostración 4.1 Los límites asintóticos están basados en los siguientes puntos:

1. La utilización de cualquier dispositivo no puede exceder de 1. Esto pone un límite a la máxima productividad que se puede obtener.
2. El tiempo de respuesta de un sistema con N usuarios no puede ser menor que el de un sistema con un usuario. Esto establece un límite mínimo al tiempo de respuesta.
3. La fórmula del tiempo de respuesta interactivo se puede usar para convertir los límites de productividad a límites de tiempo de respuesta y viceversa.

A partir de las condiciones anteriores podemos llegar a las expresiones vistas;

$$U_b = XD_{max}$$

Como U_b no puede ser mayor que 1:

$$XD_{max} \leq 1$$

ó

$$X \leq \frac{1}{D_{max}} \quad (4.10)$$

Con una tarea en el sistema no existe cola, luego el tiempo de respuesta será el mínimo:

$$R(1) = D_1 + D_2 + \dots + D_M = D$$

Aquí, D se define como la suma de todas las demandas de servicio. Con más de un usuario existirán colas y el tiempo de respuesta sería mayor, es decir:

$$R(N) \geq D \quad (4.11)$$

Aplicando la ley de tiempo de respuesta interactivo a los límites especificados por las ecuaciones (4.10) y (4.11), tenemos:

$$R(N) = \frac{N}{X(N)} - Z \geq ND_{max} - Z$$

y

$$X(N) = \frac{N}{R(N) + Z} \leq \frac{N}{D + Z}$$

junto con las ecuaciones (4.10) y (4.11) obtenemos las expresiones vistas en (4.8) y (4.9).

En la figura 4.3 aparecen los límites asintóticos para un caso típico. Ambos límites están formados por dos líneas rectas. El punto de intersección se conoce como "rodilla" (knee). El número de usuarios para el que se produce la inflexión N^* viene dado:

$$D = N^* D_{max} - Z \quad \text{ó} \quad N^* = \frac{D + Z}{D_{max}}$$

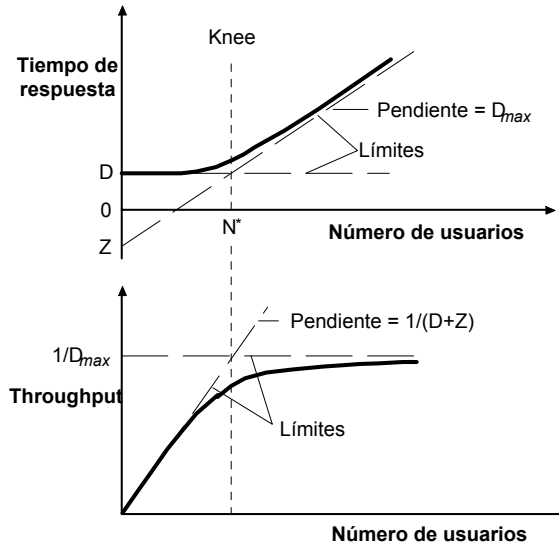


Figura 4.3 Límites asintóticos típicos

Ejemplo 4.7 Para el sistema de tiempo compartido visto en el ejemplo 4.2:

$$D_{CPU} = 5, \quad D_A = 4, \quad D_B = 3, \quad Z = 18$$

$$D = D_{CPU} + D_A + D_B = 5 + 4 + 3 = 12$$

$$D_{max} = D_{CPU} = 5$$

Los límites asintóticos son:

$$X(N) \leq \min \left\{ \frac{N}{D+Z}, \frac{1}{D_{\max}} \right\} = \min \left\{ \frac{N}{30}, \frac{1}{5} \right\}$$

$$R(N) \geq \max \{ D, ND_{\max} - Z \} = \max \{ 12, 5N - 18 \}$$

Estos límites aparecen como líneas de trazos en la figura 4.4. Las líneas continuas muestran los valores reales. El valor del punto de inflexión es:

$$12 = 5N^* - 18$$

ó

$$N^* = \frac{12+18}{5} = \frac{30}{5} = 6$$

Si existen más de 6 usuarios en el sistema existirá una cola en alguna parte.

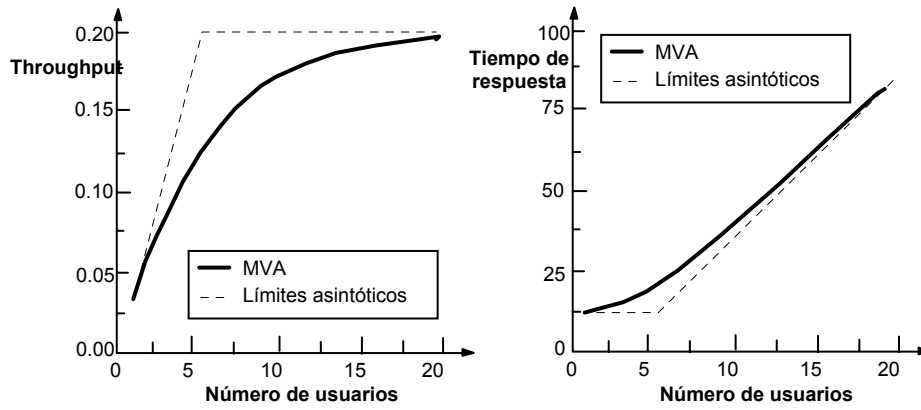


Figura 4.4 Límites asintóticos para la productividad y el tiempo de respuesta

Ejemplo 4.8 ¿Cuántos terminales se pueden soportar en el sistema de tiempo compartido del ejemplo 4.2, si el tiempo de respuesta tiene que ser menor que 100 segundos?

Usando los límites asintóticos para el tiempo de respuesta tenemos,

$$R(N) \geq \max \{ 12, 5N - 18 \}$$

El tiempo de respuesta será mayor que 100 si:

$$5N - 18 \geq 100$$

Es decir:

$$N \geq 23.6$$

El sistema no puede soportar más de 23 terminales si ha de satisfacer un tiempo de respuesta menor de 100 segundos.

En la tabla resumen 4.1 aparecen recogidas las leyes operacionales vistas en el capítulo.

Resumen 4.1 Leyes operacionales

Ley de Utilización	$U_i = X_i S_i = X D_i$
Ley de flujo forzado	$X_i = X V_i$
Ley de Little	$Q_i = X_i R_i$
Ley general del tiempo de respuesta	$R = \sum_{i=1}^M R_i V_i$
Ley del tiempo de respuesta interactivo	$R = (N / X) - Z$
Límites asintóticos	$R \geq \max\{D, N D_{\max} - Z\}$ $X \leq \min\{1 / D_{\max}, N / (D + Z)\}$

Símbolos:

D	Suma de las demandas en todos los dispositivos, $= \sum_i D_i$
D_i	Demanda total por tarea para el dispositivo i , $= S_i V_i$
D_{\max}	Demanda en el dispositivo cuello de botella, $= \max_i \{ D_i \}$
N	Número de tareas en el sistema.
Q_i	Número en el dispositivo i .
R	Tiempo de respuesta del sistema.
R_i	Tiempo de respuesta por visita al dispositivo i .
S_i	Tiempo de servicio por visita al dispositivo i .
U_i	Utilización del dispositivo i .
V_i	Número de visitas por tarea al dispositivo i .
X	Productividad del sistema.
X_i	Productividad del dispositivo i .
Z	Tiempo de reflexión.

ANEXO

LIMITACIONES DE LA TEORÍA DE COLAS

Existen algunos comportamientos de colas en la vida real que no son fáciles de modelar con la teoría de colas. Si el sistema que estamos tratando de modelar tiene alguno de los comportamientos siguientes, será difícil modelarlo empleando la teoría de colas.

1. *Tiempos de servicio no exponenciales*: La mayoría de los resultados obtenidos con modelos de redes de colas requieren la suposición de tiempo de servicio exponencial. A medida que nos apartamos de esta suposición los resultados empeoran. Si la precisión es crítica, debería emplearse una técnica de simulación con la distribución del tiempo de servicio dada.
2. *Llegadas en bloque*: Las llegadas en grupo a un centro de servicio pueden modelarse bajo ciertas condiciones, por ejemplo, si el tiempo entre grupos está distribuido exponencialmente.
3. *Salto y bifurcaciones*: Las sentencias de salto y bifurcaciones se usan en los sistemas computadores para crear y sincronizar subprocesos. Esto hace que el número de tareas en el sistema cambie, lo cual invalida la suposición de independencia entre tareas. Los modelos de colas no son adecuados para analizar estos sistemas.
4. *Llegadas dependientes de la carga*: Las redes de computadores y los sistemas distribuidos tienen políticas inteligentes de equilibrado de la carga, de forma que se distribuya de acuerdo a la carga observada. Estas dependencias de llegadas con la carga son difíciles de modelar.
5. *Bloqueo*: En sistemas computadores, el tamaño excesivo de la cola de un dispositivo puede bloquear otros dispositivos. La teoría actual no permite el análisis de este bloqueo.

6. *Análisis del transitorio*: La mayoría de los resultados de la teoría de colas sólo son válidos durante el estado estable.
7. *Competición*: Las disciplinas empleadas en los modelos de colas son demasiado sencillas para modelar algunos sistemas reales. Por ejemplo, la competición por el canal en una red Ethernet. Estos algoritmos de competición no se pueden modelar con facilidad con la teoría de colas.
8. *Exclusión mutua*: En sistemas distribuidos, varias tareas intentando usar un recurso tienen que seguir un conjunto de reglas de exclusión. Estas reglas pueden no ser fácilmente representables mediante modelos de colas.
9. *Llegadas dependientes del tiempo de respuesta*: Si un paquete o una petición permanece demasiado tiempo en una cola puede ser retransmitido de nuevo por la fuente, y por tanto incrementar aún más la cola. Esto puede conducir al sistema a un estado no estable.
10. *Modelado de memoria*: La memoria virtual permite el intercambio entre memoria física y el número de visitas al dispositivo de paginación. Estos intercambios son difíciles de analizar con la teoría de colas.
11. *Eliminación de las colas*: Un método empleado para evitar la inestabilidad causada por las llegadas dependientes de la respuesta consiste en establecer un tiempo máximo de permanencia en la cola. La petición que está en la cola durante el tiempo máximo se elimina bajo el supuesto de que la fuente ya ha lanzado una nueva petición. Estas eliminaciones son difíciles de analizar usando modelos de colas.
12. *Posesión simultánea de recursos*: En sistemas computadores, una tarea puede estar ejecutándose en dos dispositivos a la vez, ejecutándose en la CPU y a la vez acceder al sistema de E/S (procesamiento en paralelo) Este fenómeno es similar a los saltos y bifurcaciones y es difícil de analizar con la teoría de colas.
13. *Tiempo de reflexión*: En los sistemas actuales, la presencia de sistemas de ventanas y distintos tipos de menús hacen que el concepto de tiempo de reflexión sea menos intuitivo y en todo caso varía su distribución. Esto hace que el factor humano no se pueda representar como un centro de retardo con un tiempo de reflexión.

En la vida cotidiana, las personas tienen un comportamiento incluso más sofisticado que los paquetes o peticiones, lo cual las hace más interesantes pero es difícil de modelar comportamientos tales como el más grande se sirve primero, la charla y el arreglo para conseguir posiciones adelantadas en la cola, desistir antes de entrar en una cola, o saltar entre varias colas.