# SOBRE LOS SISTEMAS DE COLAS

**Emilio Hernández García**
C.U.M. Dpto. de Matemáticas. UEX.
e-mail: ehnandez@unex.es

## 1 Introducción.

Imaginemos situaciones como las siguientes:

- Clientes que esperan ser atendidos en la ventanilla de un banco.
- Automóviles que esperan pasar en un semáforo en rojo.
- Pacientes en lista de espera para un servicio médico.
- Máquinas estropeadas que esperan ser reparadas por un técnico.
- Cartas que esperan ser contestadas por un administrativo.
- Programas que esperan ser procesados por un ordenador (local o remoto).
- Llamadas telefónicas recibidas en espera de ser atendidas en una centralita.

Todas estas situaciones, y otras muchas que sin duda se le ocurrirán al lector, tienen en común el "molesto" fenómeno de la espera, que es el resultado directo de la aleatoriedad de las llegadas de los clientes y/o de los tiempos de operación para atenderlos: se formará una *cola* o *línea de espera* siempre que la demanda actual de un servicio exceda de la capacidad actual de proporcionarlo. Es claro que un incremento de una capacidad de servicio –para que no se forme la cola- no siempre será posible y, en todo caso puede conllevar costes excesivos (por ejemplo el mantenimiento de servicios con frecuencia ociosos); pero también la formación de (largas) colas puede originar costes importantes: costes sociales, pérdida de clientes (incluso por defunción, en el caso de un servicio médico), etcétera.

La *teoría de colas*, que estudia modelos matemáticos para las líneas de espera, no resuelve directamente el problema de conseguir un balance económico entre los costes del servicio y los costes por la espera del mismo, pero sí contribuye de manera fundamental con la información precisa para la resolución del problema. Esta información comprende cuestiones sobre el comportamiento de la cola (si crecerá indefinidamente, si tiende a estabilizarse,...) así como características importantes de la misma o *medidas de desempeño*: número medio de clientes, tiempo medio de espera, etc.

Suele considerarse al danés A.K. Erlang como pionero en la teoría de colas, con sus trabajos sobre telecomunicaciones publicados a partir de 1909. Otros nombres importantes en el desarrollo histórico de la teoría de colas son los de F. Pollaczek , D. Kendall, L. Takács, etc. (consúltese el capítulo 1 de la referencia [2]). Los fundamentos matemáticos se encuentran en la teoría de la probabilidad y los procesos estocásticos, pero aquí no trataremos con estos tópicos, sino que definiremos la nomenclatura estándar sobre colas, expondremos algunos resultados y mostraremos su aplicación en ejemplos sencillos de modelos de ordenadores.

## 2 Conceptos básicos. Algunos modelos de colas.

Una cola simple con un solo servidor aparece esquematizada en la Figura 1; consiste en un *servidor* que realiza los requerimientos de los *clientes*, una línea de espera o *cola* donde los clientes esperan recibir servicios y una *fuente* que genera las llegadas de los clientes al *sistema* (cola(s)+servidor(es)). La *cabecera* de la cola es la primera posición y el *final* la última. En el caso más simple los clientes llegan al final de la cola y son atendidos con una *disciplina* "primero en llegar, primero en ser atendido", que suele denotarse como FIFO (de first in, first out). En un modelo (teórico) de ordenadores el servidor puede corresponder con la CPU que procesa las tareas (clientes). El número máximo (finito o infinito) de clientes que pueden unirse a la cola es la *capacidad de la sala de espera*, o también, en modelos de ordenadores, el *tamaño del buffer*. Típicamente si la sala de espera es finita y llega un cliente cuando la sala está llena, se supone que éste se pierde para el sistema, es decir, como si nunca hubiera llegado.

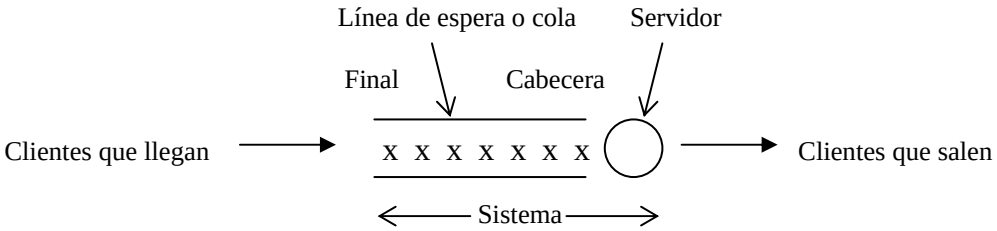


**FIGURA 1.** Una cola con un solo servidor

El modelo simple de colas antes considerado tiene una enorme cantidad de variaciones. Por ejemplo puede haber más de un servidor (*colas con servidores múltiples*); estos servidores podrían atender a los clientes a la misma velocidad (*servidores homogéneos)* o velocidades distintas (*servidores heterogéneos*); los clientes pueden ser atendidos con diferentes disciplinas, que incluyen primero en llegar, último en atender (LIFO), orden aleatorio, orden de prioridades entre los clientes, proceso compartido (PS) donde cada cliente recibe una proporción igual de la capacidad del servidor,... Los clientes podrían llegar en grupos de tamaño fijo o variable, requerir distintos servicios, etc. Es posible que haya *clientes impacientes* que abandonan el sistema con una cierta probabilidad que depende del tamaño de la cola.

Otra variación importante la constituyen las *colas con reintentos*: cuando un cliente llega al sistema y éste está completo, con una cierta probabilidad permanece en una *órbita* y accede al sistema al cabo de una cantidad aleatoria de tiempo (*tiempo de reintento)*; piense el lector cómo funciona el servicio telefónico de "llamadas en espera".

También los servidores pueden dar servicio a grupos de clientes de tamaño variable; por ejemplo un modelo muy estudiado consiste en que el servicio comienza si hay *a (quorum*) o más clientes en la cola, y se atiende simultáneamente a todos los clientes siempre que no superen la *capacidad máxima b*. Si hay más de *b* clientes en la cola, entran en el servidor *b* y el resto esperan en la cola.

Las modificaciones anteriores también pueden combinarse para proporcionar otros modelos de colas o de sistemas de ordenadores. En la Figura 2 se muestra un modelo simple de sistema distribuido en paralelo

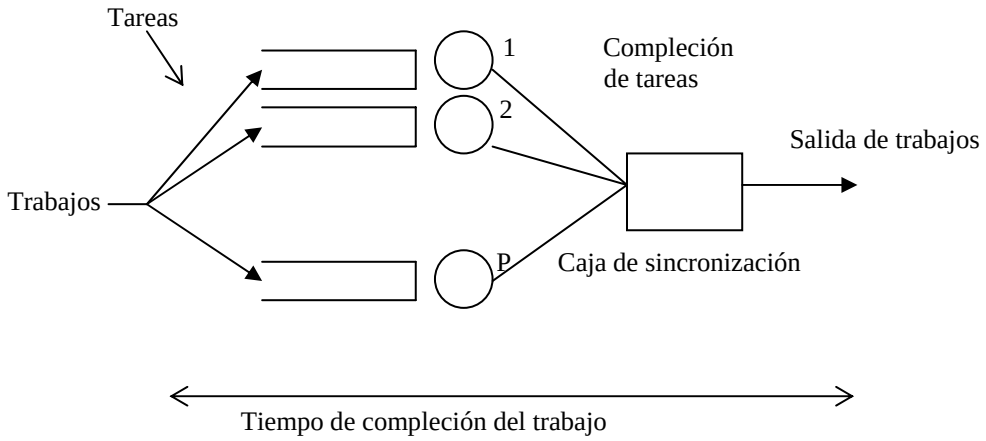


**FIGURA 2.** Un modelo simple en paralelo

Este sistema tiene P procesadores homogéneos cada uno con su propia línea de espera. Los clientes son trabajos que constan de múltiples tareas. Cuando un trabajo llega al procesador se planifica la ejecución de las tareas y se realizan de acuerdo a ese plan. Cuando se completan, las tareas entran en una caja de sincronización como muestra la figura en la que se realiza el ensamblaje. Cuando todas las tareas han finalizado (*punto de sincronización)* el trabajo abandona el sistema.

Una notación que es particularmente adecuada para resumir las características principales de las colas en paralelo se ha estandarizado en el formato:

(a/b/c):(d/e/f)

donde los símbolos a, b, c, d, e y f representan elementos básicos del modelo en la forma siguiente:

a = distribución de llegadas
b = distribución del tiempo de servicio (o salidas)
c = número de servidores en paralelo (1, 2,... $\infty$)
d = disciplina de servicio (por ejemplo LIFO, FIFO, etc.)
e = número máximo de clientes admitidos en el sistema
f = tamaño de la fuente que genera a los clientes

La notación estándar reemplaza los símbolos a y b de llegadas y salidas por códigos como los siguientes:

M = distribución de llegadas o salidas de Poisson (o markoviana), o lo que es equivalente, distribución exponencial entre llegadas o de tiempo de servicio
D = tiempo entre llegadas o de servicio constante o determinista
$E_k$ = distribución de Erlang de parámetro k de tiempo entre llegadas o de servicio
GI = distribución general independiente de llegadas
G = distribución general de salidas, etc.

Para ilustrar la notación, considérese

(M/D/8):(GD/N/∞)

que representa una cola con llegadas de Poisson, tiempo de servicio constante y 8 servidores en paralelo en la instalación. La disciplina de servicio es general (GD) en el sentido de que puede ser cualquiera; el sistema puede alojar como mucho a N clientes y la fuente que genera los clientes es potencialmente infinita.

La notación que se acaba de describir fue ideada originalmente por D.G. Kendall (1951) en la forma (a/b/c); después, A.M. Lee (1966) añadió los símbolos d, e y f. En la literatura se la conoce como *notación de Kendall-Lee* (o *notación de Kendall extendida*). Esta nomenclatura se ha mostrado suficientemente flexible para incluir otros modelos como los antes comentados. Por ejemplo $M^X/M/1$ representa una cola con llegadas en grupo según un proceso de Poisson , donde el tamaño de cada grupo, X, es una variable discreta, existe un único servidor con tiempos de servicio exponenciales, que atiende a los clientes de un mismo grupo de manera individual y aleatoria (por omisión (d/e/f) se supone que es (FIFO/∞/∞)).

## 3 Redes de colas.

Las colas pueden juntarse en *redes* (*de colas*), donde la salida de una cola forma parte de la entrada de la siguiente. Estas redes sirven para modelizar relaciones más complejas de un trabajo en un sistema informático o telemático. Consideremos por ejemplo el esquema simple reflejado en la Figura 3, para una estación de trabajo consistente en una CPU y dos dispositivos de entrada y salida (I/O), digamos un lector/grabador de CD y un disco duro (*centros de servicio)*. Los trabajos se supone que provienen de una fuente externa (*modelo abierto).* Un trabajo consiste en una secuencia alternativa de *requerimientos*, es decir, tiempos de servicio de CPU o I/O (en la figura esta secuencia aparece reseñada como bucle CPU-I/O ). Cuando todos los requerimientos de un trabajo se han completado, el trabajo abandona el sistema. En un momento dado es posible que haya diversos trabajos en el sistema en cualquier centro de servicio. Cuando un servidor finaliza un requerimiento y hay otros esperando en su cola, utiliza una disciplina determinada, por ejemplo FIFO, para determinar cuál es el siguiente a atender.

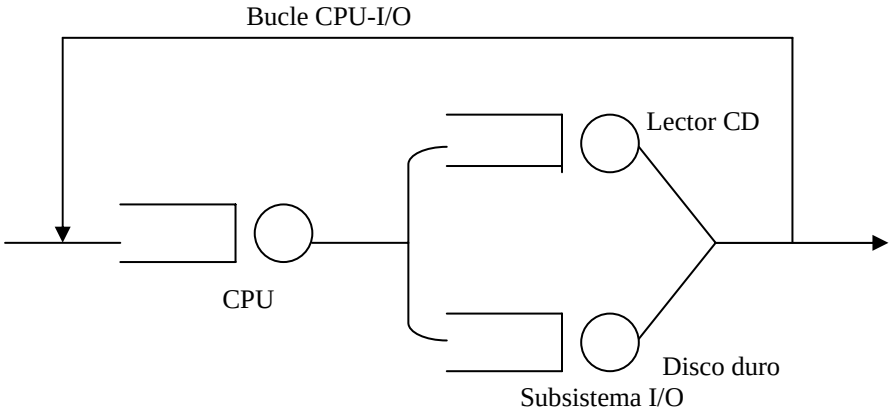


**FIGURA 3.** Una red de colas abierta

De un sistema como éste interesaría conocer cuestiones como el tiempo medio de espera para un trabajo, para un requerimiento al disco duro o al lector de CD, la cantidad máxima de trabajos que puede procesar el sistema, la proporción de tiempo que un centro de servicio está inactivo, etc.

El modelo esquematizado en la Figura 3 es una red abierta porque se supone que las llegadas proceden de una fuente externa. Una versión cerrada de este modelo aparece en la Figura 4. Ahora el número de clientes es fijo y podría ser interpretado como un conjunto de usuarios, que llamaremos *terminales*, que realizan requerimientos de una CPU y de un subsistema de I/O. En un modelo cerrado los trabajos pasan cíclicamente por un tiempo de espera, en el cual se genera un requerimiento, seguido de varios servicios de CPU e I/O. El número de ciclos para completar un trabajo es por lo habitual una cantidad aleatoria. Una vez que un trabajo se completa, se crea un nuevo trabajo que continua siguiendo el mismo ciclo. Este es un ejemplo de sistema cerrado (en cualquier momento hay una cantidad fija de trabajos en el sistema) que, junto con diversas variaciones del mismo, se conoce como *modelos de servidor central*.

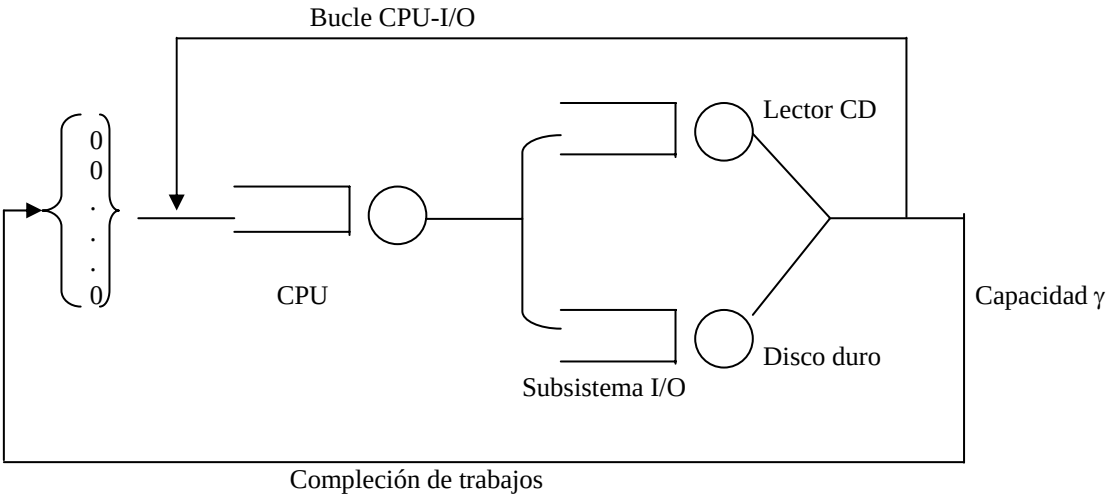


**FIGURA 4.** Una red de colas cerrada

Para estos modelos interesan, entre otras medidas, la *capacidad del sistema* $\gamma$, que es equivalente al número de trabajos que van a la fase de espera por unidad de tiempo, el tiempo medio de respuesta para un trabajo en el sistema E(T), y el tiempo medio de utilización de la CPU, $\rho$. Un famoso resultado conocido como *ley de Little* relaciona estas cantidades, de manera que si $\gamma$ es conocido y M es el número de trabajos en el sistema, entonces $E(T) = M/\gamma$ y $\rho = \gamma E(S)$, donde E(S) es el tiempo medio de servicio de la CPU. Para sistemas cerrados como el del ejemplo precedente si hay K centros de servicio y $n_i$ es el número de clientes en el i-ésimo centro de servicio ($0 \le n_i \le M$), es claro que $M = n_1 + ... + n_K$. Si no hay restricciones en cuanto al número de clientes en cada centro de servicio, entonces el número de posibles distribuciones de clientes en los centros coincide con el número de descomposiciones de M en suma de K enteros no negativos. Ciertamente esto puede originar un número grande de posibilidades, y la complejidad del análisis de redes cerradas es consecuencia de este hecho.

Es fácil pensar (pero no analizar) otros sistemas de colas más complicados que modelicen con más precisión una situación concreta. Por ejemplo en el caso de trabajos que

fluyen por un sistema, estos trabajos podrían ser de distintas clases, que se diferencian por las rutas que pueden seguir y por los requerimientos en cada centro de servicio. En fin, el lector puede comprobar por sí mismo la gran riqueza de posibilidades si se asoma a la literatura especializada.

## 4 Referencias.

Este artículo se ha basado en parte en la sección 1.2 del libro de Nelson (véase la referencia [7]). Algunos textos clásicos sobre colas, a nivel universitario, son los de Cox y Smith ([1]), Gross y Harris ([4]), Kleinrock ([5]), Saaty ([8]) y Takács ([9]), aunque esta lista podría hacerse mucho más amplia. Una referencia muy importante sobre modelos de ordenadores es Lavenberg ([6]). Los dos tomos de Dshalalow ([2] y [3]) muestran un panorama muy completo y actual sobre la investigación en la teoría de colas y sus aplicaciones, escrito por especialistas de primera línea. Todos estos libros contienen extensas bibliografías (en especial [2] y [3]).

[1] COX, D.R. Y SMITH, L.: *Queues.* Chapman and Hall. 1961.
[2] DSHALALOW, J.H. (ed.): *Advances in Queueing.* CRC Press. 1995.
[3] DSHALALOW, J.H. (ed.): *Frontiers in Queueing.* CRC Press. 1997.
[4] GROSS, D. Y HARRIS, C.M.: *Fundamentals of Queueing Theory.* Wiley. 1985.
[5] KLEINROCK, L.: *Queueing Systems.* Vol I y II. Wiley. 1975.
[6] LAVENBERG, S.S.: *Computer Performance Modeling Handbook.* Academic Pr. 1983.
[7] NELSON, R.: *Probability, stochastic processes and queueing theory.* Springer. 1995.
[8] SAATY, T.L.: *Elements of Queueing Theory.* Mc Graw-Hill. 1961.
[9] TAKACS, L.: *Introduction to the Theory of Queues.* Oxford University Press. 1962.