Tema 5

Introducción a la Teoría de Colas

A groso modo, podemos describir un sistema de colas (o sistema de líneas de espera) como un sistema al que los clientes llegan para recibir un servicio, si el servicio está ocupado esperan en cola a recibir el mismo, y una vez que han sido servidos salen del sistema.

Las líneas de espera son parte de la vida diaria. Hacemos cola en el supermercado para obtener diferentes productos o para pagar los artículos adquiridos, en el banco para gestionar nuestras cuentas, en la cantina universitaria para tomar un café o en el médico esperando a ser atendidos. También encontramos colas en un proceso de producción en el que los artículos esperan a ser manipulados por la siguiente maquinaria, o en un servicio de reparaciones en el que los distintos aparatos esperan su turno para ser revisados por los técnicos, o en determinados sistemas informáticos en el que las tareas esperan a ser procesadas por los computadores. El tiempo que la población pierde al esperar en las colas es un factor importante tanto en la calidad de vida como en la eficiencia de la economía.

La teoría de colas es el estudio de los sistemas de líneas de espera en sus distintas

modalidades. El estudio de estos modelos sirve para determinar la forma más efectiva de gestionar un sistema de colas. Proporcionar demasiada capacidad de servicio llevaría consigo excesivos costos, pero no contar con la suficiente capacidad de servicio supondría aumentar los tiempos de espera y las posibilidades de rechazo del servicio. Los modelos de colas permiten encontrar un equilibrio adecuado entre el coste del servicio y los tiempos de espera.

5.1. Estructura básica de los modelos de colas

La **fuente de entrada** la forman los clientes que desean acceder al sistema (de alguna forma es como la población de clientes potenciales). Su tamaño es el número total de clientes que pueden requerir el servicio en un determinado momento. El tamaño puede ser finito o infinito, de modo que se dice que la fuente de entrada es limitada o ilimitada. Como los cálculos son mucho más sencillos para el caso infinito, esta suposición se hace aún cuando el tamaño real sea un número finito relativamente grande. El caso finito es más difícil analíticamente, pues el número de clientes en cola afecta al número de clientes potenciales fuera del sistema.

Los **clientes** entran al sistema cada cierto tiempo y se unen a una cola. Se debe especificar el patrón estadístico mediante el cual los clientes entran al sistema. La suposición habitual es que los clientes acceden al sistema según un proceso de Poisson, lo que significa que los clientes que llegan en un intervalo determinado de tiempo siguen una distribución Poisson, con tasa media fija y sin importar cuántos clientes ya están en el sistema (con lo cual la fuente de entrada sería ilimitada). Una suposición equivalente es que los tiempos entre dos llegadas consecutivas (tiempo entre llegadas) es exponencial. También pueden considerarse otras suposiciones acerca del comportamiento de los clientes cuando llegan al sistema, como por ejemplo que un cliente rehúse acceder al

servicio porque la cola es demasiado larga.

Cuando los clientes entran al sistema se unen a una **cola**. La cola es donde los clientes esperan a ser servidos. Una cola se caracteriza por el número máximo permisible de clientes que puede admitir. La suposición de una cola infinita es más fácil de manejar analíticamente y es por ello la suposición estándar en la mayoría de los modelos, incluso si existe una cota superior los suficientemente grande que es difícil alcanzar. Sin embargo, si el tamaño máximo de la cola es tan pequeño que se alcanza con cierta frecuencia, sería necesario suponer una cola finita. También pueden considerarse otras suposiciones acerca del comportamiento de los clientes cuando llegan al sistema, como por ejemplo que un cliente rehúse acceder al servicio porque la cola es demasiado larga. Igualmente, en ciertos sistemas de colas puede ocurrir que, por limitaciones de espacio, no sea posible admitir más clientes en cola a partir de una cierta **capacidad del sistema**. En este caso, los clientes que llegan a la cola cuando el sistema alcanza su capacidad máxima son rechazados y abandonan inmediatamente el sistema.

En un determinado momento se selecciona un miembro de la cola, mediante alguna regla conocida como **disciplina de servicio**. La disciplina de servicio se refiere al orden en el que se seleccionan los clientes de la cola para recibir el servicio. Puede ser:

- FIFO (primero en entrar, primero en salir). La suposición más habitual.
- LIFO (último en entrar, primero en salir). Aplicable a sistemas de inventarios.
- Aleatoria
- Procedimiento de prioridad. Aplicable en servicios de emergencia.

Cuando un cliente es seleccionado de la cola, accede al mecanismo de servicio. El mecanismo de servicio puede consistir en una secuencia de instalaciones de servicio en

serie que el cliente debe pasar para completar el servicio. Cada instalación de servicio estará formada por varios canales de servicio paralelos, llamados servidores. En una instalación dada, el cliente entra en un servidor que le presta el servicio completo relativo a dicha instalación. Un modelo de colas debe especificar el número de instalaciones de servicio en serie y el número de servidores paralelos en cada una de ellas. Los modelos más comunes suponen una única instalación con uno o varios servidores disponibles.

En cada instalación, el tiempo que transcurre desde el inicio del servicio hasta su fin en dicha instalación se llama **tiempo de servicio**. El modelo de colas debe especificar la distribución de probabilidad del tiempo de servicio de cada servidor, y quizás de cada tipo de cliente, aunque lo común es que todos los servidores sigan las misma distribución. La suposición más habitual es que este tiempo de servicio es exponencial. Otras distribuciones de servicio importantes son la degenerada y la Erlang. En modelos más complejos, la tasa de servicio podría depender del número clientes esperando en cola.

Por último una vez que el cliente finaliza su servicio, abandona el sistema.

5.2. Análisis de un sistema de colas

Los modelos de sistemas de colas se pueden utilizar para responder preguntas tales como:

- qué fracción de tiempo está libre cada servidor
- cuál es el número esperado de clientes en cola
- cuál es el número esperado de clientes en el sistema
- cuál es el tiempo promedio que pasa un cliente en cola

- cuál es el tiempo promedio que pasa un cliente en el sistema

Generalmente, el análisis de un sistema de colas puede ir dirigido a obtener medidas de efectividad de un sistema dado, o bien a encontrar un diseño óptimo del sistema. Para esto último, si el analista fuese capaz de cuantificar los costes y beneficios asociados al sistema (costes de espera de clientes en cola, beneficio que proporciona cada cliente, coste de mantener un servidor sin utilizar, etc.), se podrían determinar de forma óptima diversas características del sistema como el número adecuado de servidores, el tamaño apropiado del sistema, etcétera.

5.3. Terminología de un modelo de colas

Siguiendo la notación de Kendall (1953), los modelos de colas habitualmente se etiquetan como sigue:

$$1/2/3/4/5/6$$

1. Distribución del tiempo entre llegadas (se asumen independientes)

   M Exponencial

   D Determinística

   E_k Erlang(k)

   G Distribución arbitraria.

2. Distribución del tiempo de servicio (se asumen independientes)

   M Exponencial

   D Determinística

E_k Erlang(k)

G Distribución arbitraria.

3. Número de servidores en paralelo
4. Disciplina de la cola
   - FIFO (first in, first out)
   - LIFO (last in, first out)
   - SIRO (service in random order)
   - PRI (priority)
   - GD (general discipline)
5. Capacidad del sistema
6. Tamaño de la fuente de entrada

Usualmente, si no se especifica nada, la disciplina de la cola se asume FIFO, y la capacidad del sistema y el tamaño de la fuente de entrada se asumen ilimitadas, así que en la mayoría de los casos sólo se utilizan los tres primeros índices.

5.4. Modelo de colas determinístico D/D/1/-

Consideremos el caso más elemental de un modelo de colas, cuando las llegadas se producen a una tasa constante, a un sistema de un solo servidor que tiene un tiempo de servicio constante. Los individuos son atendidos según una disciplina FIFO.

La terminología estándar es la siguiente:

- Estado del sistema = número de clientes en el sistema
- Longitud de la cola = número de clientes en cola = Estado del sistema - número de clientes en servicio
- $n(t)$: número de clientes en el sistema en el instante t
- λ: tasa de llegadas.
- μ: tasa de servicios.
- $W_q^{(n)}$: tiempo de espera en cola del n-ésimo cliente

Asumiremos que en el instante $t = 0$, no hay clientes en el sistema. Sea λ el número de llegadas por unidad de tiempo, i.e., $\frac{1}{\lambda}$ el tiempo constante entre dos llegadas consecutivas. Igualmente, sea μ el número de servicios por unidad de tiempo (cuando el sistema está ocupado), y por lo tanto, $\frac{1}{\mu}$ el tiempo que se tarda en realizar un servicio. En nuestro análisis distinguiremos dos casos, según si la tasa de llegadas es mayor o menor igual que la de servicios.

5.4.1. Caso $\lambda > \mu$

En esta situación, puesto que llegan más clientes por unidad de tiempo que los que son servidos, el estado del sistema crecería indefinidamente. Por ello, supondremos que la capacidad máxima del sistema es K. Así pues, analizaremos un modelo $M/M/1/K$.

Suponiendo que tan pronto como acaba un servicio empieza el siguiente, el número de clientes en el sistema en el instante t se puede determinar como:

$$n(t) = |\text{llegadas en } (0,t]| - |\text{servicios en } (\frac{1}{\lambda},t]| = \left[\frac{t}{1/\lambda}\right] - \left[\frac{t-(1/\lambda)}{1/\mu}\right] = [\lambda t] - \left[\mu t - \frac{\mu}{\lambda}\right],$$

con $n(0) = 0$. Esta ecuación sólo es valida hasta que ocurra el primer rechazo. El primer rechazo ocurre en el instante $\bar{t}$, donde $\bar{t}$ es el menor número real que verifique $n(\bar{t}) = K + 1 = [\lambda \bar{t}] - [\mu \bar{t} - \frac{\mu}{\lambda}]$. En el intervalo entre el instante $\bar{t}$ y el instante en el que se concluya el siguiente servicio $n(t)$ permanece en K. Cuando concluye el siguiente servicio, pueden pasar dos cosas: si en ese mismo instante se produce una nueva llegada, entonces $n(t)$ permanece en K; en caso contrario, $n(t)$ baja a $K - 1$ y volverá a K cuando se produzca la siguiente llegada que ocurrirá, por hipótesis, antes del siguiente servicio. En resumen, $n(t)$ nunca bajará ya de $K - 1$, y permanecerá siempre en K si el tiempo de servicio es múltiplo del tiempo entre llegadas.

5.4.1.1. Subcaso $\frac{1}{\mu} = m\left(\frac{1}{\lambda}\right)$

Si $\frac{1}{\mu} = m\left(\frac{1}{\lambda}\right)$, para algún entero $m \geq 1$, $n(t)$ toma la siguiente expresión

$$n(t) = \begin{cases} 0 & t < \frac{1}{\lambda} \\ [\lambda t] - [\mu t - \frac{\mu}{\lambda}] & \frac{1}{\lambda} \leq t < \bar{t} \\ K & \bar{t} \leq t \end{cases}$$

En cuanto a los tiempos de espera en cola, se observa que, independientemente de la distribución de los tiempos de llegada y de servicio, en cualquier sistema con un único servidor los tiempos de espera en cola de dos clientes consecutivos están relacionados por la siguiente expresión:

$$W_q^{(n+1)} = \begin{cases} W_q^{(n)} + S^{(n)} - T^{(n)} & \text{si } W_q^{(n)} + S^{(n)} - T^{(n)} > 0 \\ 0 & \text{si } W_q^{(n)} + S^{(n)} - T^{(n)} \leq 0 \end{cases}$$

donde $S^{(n)}$ es el tiempo de servicio del n-ésimo cliente y $T^{(n)}$ es el tiempo transcurrido entre las llegadas de los clientes n-ésimo y $(n+1)$-ésimo.

Observemos que $S^{(n)} = \frac{1}{\mu}$ y si $n < \lambda \bar{t}$, entonces $T^{(n)} = \frac{1}{\lambda}$. Consecuentemente,

$W_q^{(n+1)} = W_q^{(n)} + \left(\frac{1}{\mu} - \frac{1}{\lambda}\right)$, y por lo tanto, $\Delta W_q^{(n)} = \left(\frac{1}{\mu} - \frac{1}{\lambda}\right) n + C$. Puesto que el primer cliente no espera nada, $W_q^{(1)} = 0$, se tiene que $C = -\left(\frac{1}{\mu} - \frac{1}{\lambda}\right)$, y por lo tanto, $W_q^{(n)} = \left(\frac{1}{\mu} - \frac{1}{\lambda}\right)(n-1)$, para todo $n < \lambda\bar{t}$.

Por otro lado, cuando $n \geq \lambda\bar{t}$, un cliente entra al sistema justo cuando sale otro. Por lo tanto, el cliente que entra se encuentra con $(K-2)$ clientes en cola delante de él y un cliente empezando el servicio. Consecuentemente, el tiempo de espera en cola para este nuevo cliente será $(K-1)\frac{1}{\mu}$.

En definitiva, tenemos que

$$W_q^{(n)} = \begin{cases} \left(\frac{1}{\mu} - \frac{1}{\lambda}\right)(n-1) & \text{si } n < \lambda\bar{t} \\ (K-1)\frac{1}{\mu} & \text{si } n \geq \lambda\bar{t} \end{cases}$$

Ejemplo 5.1. Analizar un modelo de colas determinístico con un único servidor, en el que se producen exactamente 1 llegada cada cuatro minutos, el tiempo de servicio es exactamente de 8 minutos y la capacidad máxima del sistema es de 4 individuos.

5.4.1.2. Subcaso $\frac{1}{\lambda} \neq m\frac{1}{\mu}$

Este caso es más difícil alcanzar resultados generales, debido a que ocasionalmente el número de clientes en el sistema disminuirá a $K-1$. En cualquier caso, los ejemplos específicos son fáciles de analizar.

Ejemplo 5.2. Analizar un modelo de colas determinístico con un único servidor, en el que se producen exactamente 1 llegada cada cuatro minutos, el tiempo de servicio es exactamente de 6 minutos y la capacidad máxima del sistema es de 4 individuos.

Ejemplo 5.3. Analizar un modelo de colas determinístico con un único servidor, en el que se producen exactamente 1 llegada cada tres minutos, el tiempo de servicio es exactamente de 7 minutos y la capacidad máxima del sistema es de 4 individuos.

5.4.2. Caso $\frac{1}{\lambda} \geq \frac{1}{\mu}$

En este caso la situación es muy simple. Puesto que el tiempo entre llegadas es mayor o igual que el tiempo de servicio, cuando llega un cliente es servido y sale del sistema antes de que llegue el siguiente (en el peor de los casos, justo cuando llegue el siguiente si el tiempo entre llegadas es igual al tiempo de servicio). Por lo tanto, si consideramos que en el instante cero no hay individuos en el sistema, el número de clientes en el sistema siempre variará entre 0 y 1 y el tiempo de espera en cola de cualquier cliente es cero.

Una situación interesante para este caso consiste en suponer que cuando el sistema se inicia ya hay un número de clientes C esperando en el mismo.

5.4.2.1. Subcaso $\frac{1}{\lambda} = \frac{1}{\mu}$

Si el tiempo entre llegadas coincide con el tiempo de servicio, entonces el análisis es trivial, puesto que siempre habrá C clientes en el sistema y el tiempo de espera en cola del n-ésimo cliente toma la siguiente expresión:

$$W_q^{(n)} = \begin{cases} (n-1)\frac{1}{\mu} & 1 \leq n \leq C \\ (C-1)\frac{1}{\mu} & n \geq C \end{cases}$$

5.4.2.2. Subcaso $\frac{1}{\lambda} > \frac{1}{\mu}$

Si el tiempo entre llegadas es mayor que el tiempo de servicio, entonces habrá un primer instante $\bar{t}$ en el que el número de clientes en el sistema sea 0, $n(\bar{t}) = 0$. Hasta ese momento, el número de clientes en el sistema en cada instante será

$$n(t) = C - \left(\left[\frac{t}{\frac{1}{\mu}} \right] - \left[\frac{t}{\frac{1}{\lambda}} \right] \right] = C - ([\mu t] - [\lambda t])$$

Consecuentemente $\bar{t}$ será el menor real positivo tal que $C = [\mu t] - [\lambda t]$.

Inmediatamente después del instante $\bar{t}$ el sistema permanece vacío hasta que se produce la siguiente llegada, la cual se produce en en instante $t_1 = [\lambda t]\frac{1}{\lambda} + \frac{1}{\lambda}$. En el instante t_1 un nuevo cliente entra al sistema y como no hay nadie más, entra directamente al servicio, del que saldrá en el instante $t_1 + \frac{1}{\mu}$. Así pues el estado del sistema tomará el valor 1 entre t_1 y $t_1 + \frac{1}{\mu}$. El siguiente cliente llegará en el instante $t_2 = t_1 + \frac{1}{\lambda} = [\lambda t]\frac{1}{\lambda} + \frac{2}{\lambda}$, después de haber salido el cliente anterior, luego el sistema volverá a estar vacío en el intervalo $[t_1 + \frac{1}{\mu}, t_2)$ y volverá a tomar valor 1 en $[t_2, t_2 + \frac{1}{\mu})$, y así consecutivamente. Por lo tanto,

$$n(t) = \begin{cases} C - ([\mu t] - [\lambda t]) & 0 \geq t < \bar{t} \\ 0 & \bar{t} \leq t < t_1, t_k + \frac{1}{\mu} \geq t < t_k + \frac{1}{\lambda} \\ 1 & t_k \geq t < t_k + \frac{1}{\mu} \end{cases}$$

donde $t_k = [\lambda \bar{t}]\frac{1}{\lambda} + \frac{k}{\lambda}$.

En cuanto al tiempo de espera en cola, el análisis para los primeros C clientes es similar al subcaso anterior. Por otro lado, si un cliente llega después del instante $\bar{t}$, entonces no espera nada nada, pues, tal y como acabamos de ver, a su llegada se encontraría el sistema vacío. Veamos qué ocurre con los restantes clientes, es decir aquellos nuevos clientes que llegan antes de que el sistema se haya vaciado por primera vez.

El primer nuevo cliente que llega al sistema, el cliente $C + 1$, lo hace en el instante $\frac{1}{\lambda}$, y el último servicio de los C clientes previos acabará en el instante $M\frac{1}{\mu}$, luego su tiempo de espera en cola es $M\frac{1}{\mu} - \frac{1}{\lambda}$. El cliente $C + 2$ llegará en el instante $\frac{2}{\lambda}$, y dado que el cliente anterior acabará su servicio en el instante $(M + 1)\frac{1}{\mu}$, el tiempo de espera para este segundo nuevo cliente será $(M + 1)\frac{1}{\mu} - \frac{2}{\lambda}$. Continuando el análisis de modo similar, podemos concluir que el cliente $C + k$, siempre que llegue antes de $\bar{t}$, es decir, si

$n \leq [\lambda \bar{t}]$, tendrá un tiempo de espera en cola igual a $(M+k-1)\frac{1}{\mu} - \frac{k}{\lambda}$ (observemos que de nuevo se verifica la relación $W_q^{(n+1)} = W_q^{(n)} + S^{(n)} - T^{(n)}$). En definitiva, tenemos la siguiente expresión para el tiempo de espera en cola:

$$W_q^{(n)} = \begin{cases} (n-1)\frac{1}{\mu} & 1 \leq n \leq C \\ (C+k-1)\frac{1}{\mu} - \frac{k}{\lambda} & n = C+k, 0 \leq k \leq [\lambda \bar{t}] \\ 0 & n = C+k, k > [\lambda \bar{t}] \end{cases}$$

Ejemplo 5.4. Analizar un modelo de colas determinístico con un único servidor, en el que se producen exactamente 1 llegada cada tres minutos, el tiempo de servicio es exactamente de un minuto y suponiendo que hay 7 individuos esperando cuando se inicia el sistema.