PROBABILIDAD Y ESTADÍSTICA II

TEMA 3.1 ESTRUCTURA DE UNA COLA

CONTENIDOS

1. Introducción

2. Elementos básicos de un modelo de colas: población, régimen de llegadas, número de servidores, régimen de servicio, capacidad del sistema, disciplina de la cola, fases de servicio

3. Notación Kendall: A / B / c / k / m / Z

4. Medidas de comportamiento: intensidad de tráfico, utilización del servidor, paso a través del sistema (productividad). Sistema estable y sistema saturado

5. Ecuaciones de coste y fórmula(s) de Little: relación entre el número medio de clientes en sistema (cola, servicio) y tiempo medio de un cliente en sistema (cola, servicio)

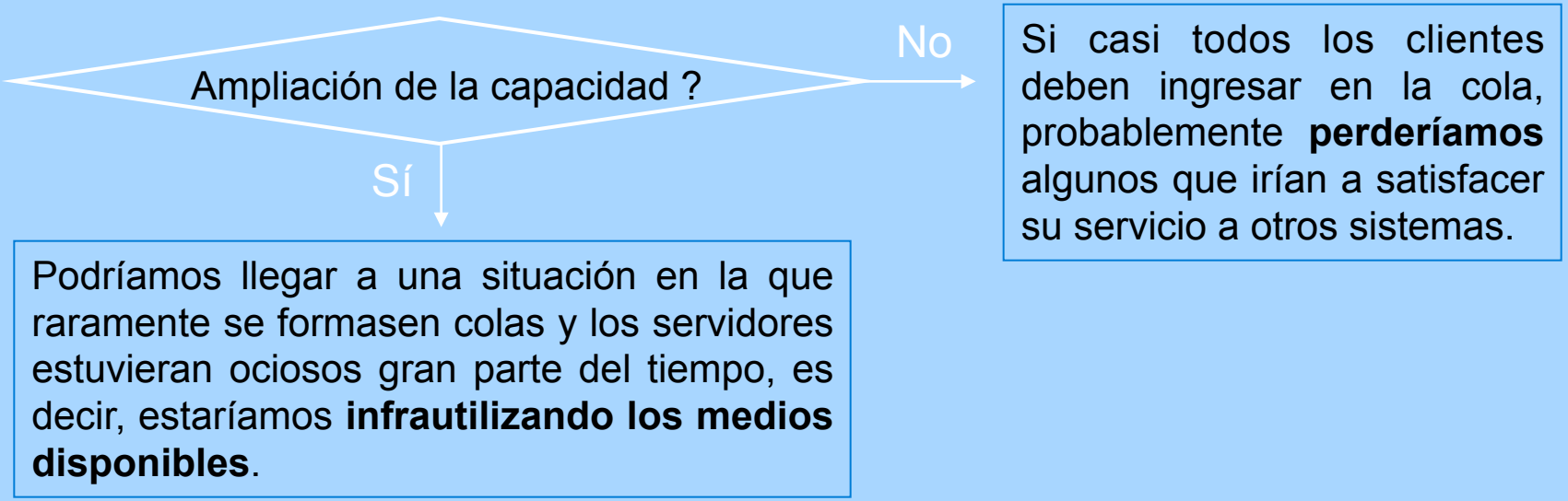
6. Comportamiento de transición y estacionario. Probabilidades límites. Propiedad PASTA

1. Introducción

En numerosas situaciones de nuestra vida diaria esperamos en una cola o línea de espera, como para comprar el billete del metro o la entrada de cine, para cobrar un cheque en el banco, para pagar en el supermercado o la cafetería, para obtener una mesa en un restaurante, para ser operado o atendido en un hospital, para echar gasolina o pagar el peaje, para desplazarnos en un atasco de tráfico, etc.

También en los sistemas informáticos son frecuentes los fenómenos de espera. Así, puede haber colas de personas esperando a usar un terminal, colas de solicitudes de entrada/salida (E/S), mensajes o paquetes de datos o programas informáticos que esperan para ser procesados por un sistema central o llamadas telefónicas esperando una línea desocupada para completar la conexión.

La espera se produce porque hay **más demanda de servicio que el disponible**. Sin embargo, ampliar esta capacidad de servicio no siempre es la solución adecuada.



Por lo tanto, se trata de compensar un nivel adecuado de servicio con unos gastos no excesivos.

Para llegar a una solución, el analista del sistema necesita conocer las respuestas a preguntas como:

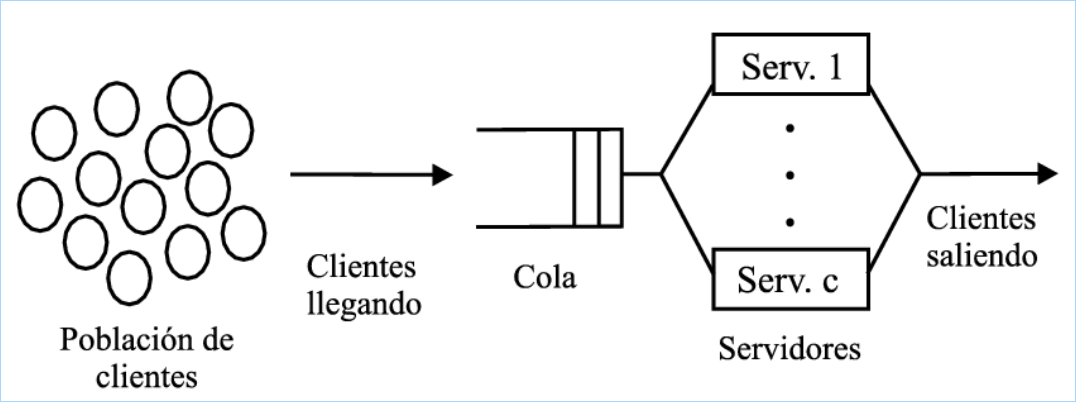
- ¿cuánto tiempo debe esperar un cliente?
- ¿cuántos clientes se acumularán en la cola? o
- ¿cuántos clientes llegan por unidad de tiempo?

A partir de ellas podrá considerar varios sistemas alternativos y tratar de evaluar su funcionamiento.

La **teoría de colas** proporciona al planificador diversos modelos que dan respuesta a estas cuestiones y, en particular, ha demostrado ser una de las áreas más fructíferas de la Teoría de la Probabilidad aplicada a la Informática.

2. Elementos básicos de un modelo de colas

La siguiente figura esquematiza los elementos de un sistema de colas. Los clientes, que provienen de una población o fuente llegan al sistema para recibir algún tipo de servicio.



El dispositivo de servicio del sistema ofrece un conjunto (limitado) de servidores o recursos, a veces llamados canales, para satisfacer las peticiones de los clientes. Si cuando el cliente llega al sistema, todos los servidores están ocupados, deberá esperar en cola antes de empezar a recibir servicio. Una vez que el cliente recibe el servicio demandado abandona el sistema.

Una descripción más precisa de un sistema de colas requiere especificar en detalle siete características básicas:

1. **Población o fuente de clientes**
2. **Modelo de llegadas**
3. **Modelo de servicio de cada servidor**
4. **Número de servidores o canales**
5. **Número de etapas de servicio**
6. **Capacidad del sistema**
7. **Disciplina de la cola**

Población o fuente de clientes

La población o fuente de clientes potenciales puede ser finita o infinita.

Infinita → conduce a sistemas con descripciones matemáticas más sencillas,
Finita → el número de clientes en el sistema afecta a la tasa de llegadas, que será cero si todos los clientes están en el sistema.

Si la población es finita pero suficientemente grande, se asume que es infinita para simplificar el análisis del modelo.

Modelo de llegadas

Describe el patrón de llegadas de los clientes al sistema.

Si es **determinista** (llegadas están igualmente espaciadas en el tiempo) bastará caracterizarlo midiendo el número medio de llegadas por unidad de tiempo o el tiempo medio entre llegadas consecutivas.

Denotaremos con λ a la tasa media de llegadas o velocidad media, siendo por tanto $1/\lambda$ el tiempo medio entre llegadas.

En general, habrá **incertidumbre** en el modelo de llegadas y habrá que especificar la ley de probabilidad que rige el comportamiento aleatorio de las llegadas.

Suponemos que los tiempos de llegada de los clientes son

$$0 = t_0 < t_1 < t_2 < \ldots < t_n < \ldots$$

La observación del sistema comienza en el instante 0 y t_k es el instante en el que llega el cliente k-ésimo. Las variables aleatorias $T_k = t_k - t_{k-1}$, k =1,2, 3,... representan los tiempos entre llegadas.

Normalmente, supondremos que el proceso estocástico continuo de tiempo discreto $T_1, T_2, \ldots$ es una secuencia de variables aleatorias independientes e idénticamente distribuidas (v.a.i.i.d.) con distribución T, con $E(T)=1/\lambda$.

El patrón de llegadas queda especificado dando la distribución de probabilidad de los tiempos de llegada o equivalentemente de los tiempos entre llegadas $P(T \leq t)$, que es la que suele utilizarse.

Los descriptores usuales son:

• *M*: tiempo entre llegadas es exponencial (el proceso de llegadas es de Poisson); la letra *M* proviene de la propiedad Markoviana de la exponencial;

• *D*: tiempo entre llegadas con patrón determinista o constante;

• E_k: tiempo entre llegadas con distribución de Erlang de *k* etapas;

• H_k: tiempo entre llegadas con distribución hiperexponencial de *k* etapas;

• *G*: tiempo entre llegadas sigue una distribución general o arbitraria.

Si el patrón de llegadas no cambia con el tiempo, es decir, la forma y los valores de los parámetros de la distribución son siempre iguales con el paso del tiempo, el modelo de llegadas es **estacionario**.

Posibilidad de que se produzcan llegadas en lotes o en masa al sistema, en lugar de un cliente cada vez. (Ej. llegada de familias a la consulta de un dentista). En este caso, el tamaño del lote es otra variable aleatoria del sistema.

Reacción del cliente al llegar al sistema:

• Un cliente puede decidir ingresar en la cola sin importarle la longitud de ésta, o si la considera demasiado larga puede decidir no entrar.

• También es posible que después de estar cierto tiempo en la cola decida marcharse.

• Por último, en el caso de disponer de dos o más colas simultáneamente, los clientes pueden decidir cambiarse de una a otra.

Estas situaciones son ejemplos de colas con clientes impacientes, y pueden considerarse llegadas dependientes del estado (o congestión) del sistema.

Modelo de servicio en cada servidor

Describe el tiempo de servicio que emplea un servidor en atender a un cliente.

En el caso **determinístico**, el patrón de servicio quedará descrito mediante el número de clientes servidos por canal por unidad de tiempo o mediante el tiempo requerido para servir a un cliente.

En caso contrario, será necesario especificar la distribución de probabilidad de la variable aleatoria *s*.
Supondremos que la secuencia de los tiempos de servicio de clientes sucesivos $s_1, s_2, \ldots$ también es un conjunto de v.a.i.i.d. con $E(s) = 1/\mu = W_s$.

Además, los procesos de llegadas y de servicio suelen considerarse independientes entre si.

Sin embargo, existe una diferencia importante entre ambos procesos. Cuando hablamos de tasa de servicio o tiempo de servicio, los términos están condicionados a que el sistema no esté vacío (el servidor esté ocupado).

La tasa de servicio es la capacidad intrínseca del servidor para satisfacer las necesidades de los clientes que, en general, es distinta de la tasa de salidas o número de clientes que dejan el canal de servicio una vez satisfechas sus necesidades.

Es decir, la tasa media de servicio es la tasa media a la que el servidor procesa a los clientes si estuviese ocupado el 100% del tiempo.

La notación para los patrones de servicio: M, D, E_k, H_k, G.

Así, M significa que la variable aleatoria s es exponencial de parámetro μ, que es la más usada. Por la propiedad de pérdida de memoria de la exponencial, el tiempo restante hasta completar el servicio de un cliente es independiente del tiempo que este cliente lleve en el canal.

El servicio puede ser **estacionario** o no respecto al tiempo (Ej. Un servidor que va aprendiendo y se vuelve más eficiente según adquiere experiencia).

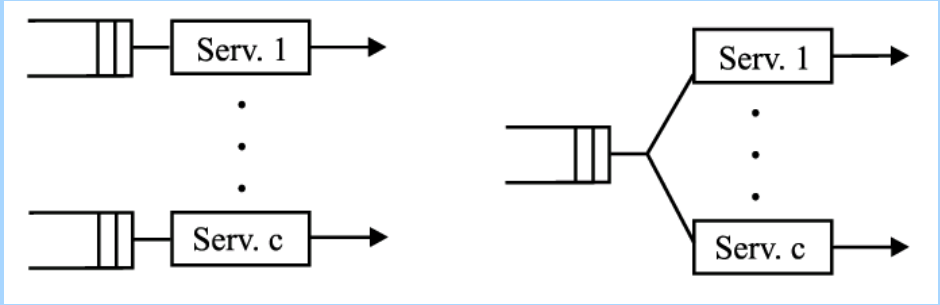
La dependencia del tiempo no debe confundirse con **dependencia del estado**, es decir, del número de clientes que hay en el sistema. (Ej. trabajar más rápido según ve que se incrementa el número de clientes en la cola).

Puede haber situaciones en las que varios clientes sean atendidos simultáneamente por el mismo servidor, es decir, servicio por lotes o en masa, como un ordenador con procesamiento paralelo o turistas en una visita guiada.

Número de servidores o canales

El sistema de colas más sencillo tiene un único servidor, que atiende a un solo cliente cada vez. Un sistema multicanal o multiservicio dispone de *c* canales paralelos y puede dar servicio a *c* clientes a la vez.

La siguiente figura muestra dos variaciones de sistemas multicanal, cada canal tiene su propia línea de espera (cajas de supermercados o pasar la ITV de un coche), y una sola cola para todos los canales (turno en la peluquería).



Normalmente, se supone que los servidores son idénticos y funcionan de forma independiente unos de otros.
En el caso extremo de infinitos servidores, utilizado a veces como aproximación, cada cliente que llega es atendido inmediatamente.

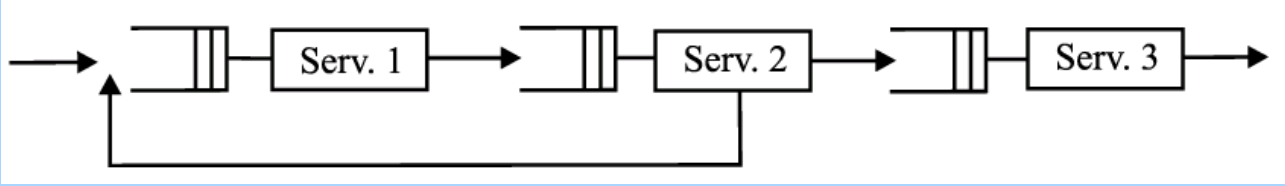
Número de etapas de servicio

A veces existen varias etapas de servicio por las que debe pasar el cliente, como en una cadena de producción o de montaje. Variantes:

- cada etapa acepta un cliente una vez que ha terminado el servicio del anterior (línea de sistemas de espera),

- la primera etapa sólo acepta un nuevo cliente cuando el anterior ha abandonado la última etapa (servidor constituido por diversas etapas).

La siguiente figura ilustra un sistema de colas con la primera variante, y además, presenta reciclado o retroalimentación, típico en procesos de fabricación con inspecciones para control de calidad en ciertas etapas.
Un artículo que no cumple ciertas normas de calidad debe ser reprocesado.



Capacidad del sistema

En algunos sistemas de colas hay limitación física sobre el número máximo *K* de clientes que puede haber en el sistema. Cuando la cola alcanza cierta longitud, cualquier cliente que llega es rechazado hasta que se dispone de sitio, por completarse algún servicio.

El estado del sistema influye en la tasa de entradas al sistema λ_e, que es igual a la tasa de llegadas al sistema λ menos el número medio de clientes que no entran al mismo.

Un caso extremo son los llamados sistemas de pérdidas, que no admiten colas, como algunos sistemas de comunicación telefónica. En otros sistemas, sin embargo, puede suponerse que su capacidad es infinita y todo cliente que llega puede esperar hasta que se le proporcione servicio.

Disciplina de la cola

La disciplina de gestión de la cola o estrategia de servicio es la forma en que se seleccionan los clientes que aguardan en la cola para entrar en el dispositivo de servicio.

- FIFO (first-in, first-out) o FCFS (first-come, first-served). La más corriente es la de "el primero en llegar es el primero en entrar“. Es la que se supondrá por defecto.

- LIFO (last-in-first-out) o LCFS . El primero en entrar es el último en llegar. Usada en muchos sistemas de inventario en los que las unidades no son perecederas y resulta más sencillo tomar las unidades más cercanas, que se almacenaron más tarde.

- SIRO (service in random order). Servicio en orden aleatorio, independientemente del instante de llegada a la cola.

- SIFO (shortest-in, first-out) o SJF (shortest job first). Se sirve primero al cliente que demanda un tiempo de servicio menor.

- RR (round robin, turno robado). Reparte el tiempo del servidor equitativamente entre todos los clientes que esperan. Si el cliente no termina su servicio al final de la rodaja de tiempo que le corresponde utilizar, retorna a la cola, que se gestiona mediante disciplina FIFO. Esto se repite hasta que el cliente termina su servicio.

- PS (processor sharing) o de compartición del procesador. Disciplina RR en la que las rodajas de tiempo son infinitamente pequeñas. Es como si todos los clientes fueran servidos simultáneamente y sus tiempos de servicio incrementados de la misma forma.

Esquemas de prioridades → dan trato preferencial a ciertos clientes sobre otros. Los clientes están divididos en clases de prioridades. Los de prioridades más altas son servidos antes que los de prioridades más bajas, independientemente de su instante de llegada al sistema. Las prioridades pueden ir variando con el paso del tiempo.

- prioridad expulsiva, o con desalojo o apropiación (preemptive). Se interrumpe el servicio, el cliente recién llegado comienza a ser servido, y el cliente cuyo servicio ha sido interrumpido vuelve a la cabeza de la cola de su clase. Cuando el cliente desalojado reanude su servicio, éste comenzará desde el principio o desde el punto de interrupción, dependiendo del tipo de sistema.

- prioridad sin desalojo, el cliente recién llegado espera hasta que el cliente siendo servido complete su servicio.

3. Notación Kendall: A / B / c / k / m / Z

Todos estos elementos básicos que describen un sistema de colas se representan mediante una notación abreviada estándar, denominada notación de Kendall (en honor a David Kendall).

Escribiremos *A/B/c/K/m/Z*, donde:

A indica la distribución del tiempo entre llegadas,
B la distribución del tiempo de servicio,
c el número de canales de servicio (c≥1),
K la capacidad del sistema,
m el tamaño de la población y
Z la disciplina de la cola.

Por ejemplo, *D*/*M*/3/40/∞/LIFO.

En muchos casos no hay límite sobre la capacidad del sistema, la fuente de clientes es infinita y la disciplina es FIFO. En estas situaciones suelen omitirse los tres símbolos finales. Así, M/G/4 es lo mismo que M/G/4/∞/∞/ FIFO.

4. Medidas de rendimiento del sistema

Serán útiles tanto para el propio cliente que ingresa en el sistema como para el planificador, gestor o analista del mismo.

Como estos modelos representan sistemas dinámicos, los valores de estas medidas varían con el tiempo. Sin embargo, analizaremos los resultados que se obtienen cuando el sistema está en **equilibrio**, es decir, el comportamiento transitorio ha finalizado, está en estado estacionario, el sistema se ha normalizado y los valores de las medidas de comportamiento son independientes del tiempo.

Entonces, se verifica que la tasa a la que los clientes llegan al sistema es igual a la tasa a la que salen del sistema. A este sistema también se le denomina sistema estable.

Las soluciones transitorias sólo están disponibles en forma cerrada para sistemas muy simples, y en casos más generales, habrá que recurrir a técnicas de cadenas de Markov (*Procesos Estocásticos*).

Las medidas más importantes del rendimiento de un sistema de colas son:

1. **Probabilidad de que haya *n* clientes en el sistema**
2. **Trabajo**
3. **Intensidad de tráfico**
4. **Utilización o uso del servidor**
5. **Productividad del sistema**
6. **Tiempos de permanencia en el sistema y en la cola**
7. **Número medio de clientes en el sistema y en la cola**

Probabilidad de que haya *n* clientes en el sistema

Los valores medios de muchas otras medidas podrán deducirse de ella:

$$\pi_n = P \text{ (hay } n \text{ clientes en el sistema)} = \lim_{t\to\infty} p_n(t)$$

Por ejemplo, $\pi_0 = 0.4$ indica que a largo plazo el sistema estará vacío el 40% del tiempo.

Trabajo

La llegada de cada cliente supone al sistema una cantidad media de trabajo, que suele medirse en tiempo y que coincide con su tiempo medio de servicio $W_s = 1/\mu$.

Intensidad de tráfico

Se define como:

$$r = \lambda/\mu = \lambda\, E(s) = E(s) / E(T)$$

Si λ es el número medio de llegadas por unidad de tiempo, la cantidad total de trabajo que tiene el sistema en media por unidad de tiempo es $r = \lambda\, W_s = \lambda/\mu$.

Por ejemplo, si el tiempo entre llegadas fuese siempre constante e igual a 30 segundos y el de servicio 15 segundos, entonces el servidor estaría ocupado la mitad del tiempo: $r = 15/30 = 0.5$.
Si este servidor fuese reemplazado por otro más lento, que emplea 45 segundos en dar servicio, entonces $r = 45/30 = 1.5$. Es decir, necesitaría dar 45 segundos de servicio cada 30 segundos, lo que es imposible, a no ser que se añadiese otro servidor.

Es, por tanto, una medida del número mínimo de canales que se necesitan para atender el flujo de clientes que llegan y que hace que el sistema sea estable.

Por ejemplo, si $\lambda = 9$ clientes por día y $\mu = 2$ clientes por día, entonces $r = 4.5$ y necesitaríamos al menos 5 canales para poder satisfacer las necesidades de los clientes.

Esto es, el número de canales requeridos es el menor entero positivo c tal que $r / c < 1$. Conocido r, el gestor tiene las opciones de aumentar el número de canales si no son suficientes o bien aumentar su velocidad de servicio (es decir, μ) para disminuir r.

Utilización o uso del servidor

El uso o (factor de) utilización de un servidor si hay varios (c) en el sistema, es la fracción media de servidores activos (proporción media de tiempo que cada servidor está ocupado).

Por tanto, como $c\mu$ es la tasa global de servicio, suponiendo que el tráfico está igualmente repartido entre todos los servidores,

$$p = r / c = \lambda / c\mu$$

representa también la cantidad media de trabajo que recibe cada servidor.

Si sólo hay un servidor, p es la proporción de tiempo que está ocupado, es decir, $p = r$, siempre que no haya límite sobre la capacidad del sistema.

p es una medida de la congestión del sistema y puede usarse para formular la condición de comportamiento estable mencionada antes, $p < 1$ → para que el sistema soporte el nivel de demanda, en media debe ser menor el número de clientes que llegan en una unidad de tiempo que el número de clientes que pueden ser atendidos.

Si no, el número de clientes almacenados en la cola crecerá sin límite con el paso del tiempo. Por eso llamaremos situación de congestión a $p \geq 1$.

La situación de igualdad, $p = 1$, da lugar a congestión salvo en contadas excepciones, como en una cola $D/D/c$ con tráfico no aleatorio.

Productividad del sistema

La productividad del sistema o caudal o paso a través del sistema, $\boldsymbol{\lambda}$, es el número medio de clientes cuyo servicio se completa en una unidad de tiempo, es decir, es la tasa de salida del sistema.

Sistema con capacidad ilimitada, $\boldsymbol{\lambda} = \min\{\lambda, c\mu\}$.
Sistema congestionado, $\boldsymbol{\lambda} = c\mu$.
Sistema estable y sin pérdidas, como la tasa de salida coincide con la tasa de llegadas, se tiene que $\boldsymbol{\lambda} = \lambda = \rho c\mu$

Recordemos de nuevo que si la capacidad del sistema es finita, la productividad puede ser diferente a la tasa externa de llegadas y será $\boldsymbol{\lambda} < \min\{\lambda, c\mu\}$.

Tiempos de permanencia en el sistema (*w*) y en la cola (*q*)

Las medias de dichos tiempos las denotaremos como $W = E(w)$ y $W_q = E(q)$, respectivamente. Por tanto, $w = q+s$ y tomando esperanzas, $W = W_q+E(s) = W_q+1/\mu = W_q+ W_s$.

A veces utilizaremos las funciones de distribución de estas tres variables aleatorias, denotadas como $F_w(t)$, $F_q(t)$, $F_s(t)$, respectivamente.

Número de clientes en el sistema (*N*) y en la cola (N_q)

La media de la variable aleatoria *N*, que toma los valores 0,1,2..., será

$$L = E(N) = \sum_{n=1}^{\infty} n\pi_n$$

Podemos decir que $N_q = \max\{0, N-c\}$. Su esperanza matemática la denotaremos con $L_q = E(N_q)$.

Si N_s indica el número de clientes siendo servidos, con media L_s, entonces $N = N_q + N_s$ y $L = L_q + L_s$.

5. Ecuaciones de coste y fórmulas de Little

Fórmulas de Little:

$$L = \lambda W$$
$$L_q = \lambda W_q$$
$$L_s = \lambda W_s$$

El teorema de Little es válido bajo condiciones muy generales de un sistema estable, cualquier número de servidores y para todas las disciplinas de colas.

Intuitivamente puede explicarse de la siguiente forma:

Un cliente que acaba de llegar saldrá del sistema, en promedio, después de un tiempo W. Cuando salga, quedarán en el sistema L clientes en promedio. Cada uno de estos clientes ha llegado, en promedio, tras un tiempo $1/\lambda$. El tiempo que han tardado en llegar estos L clientes es $L \times (1/\lambda)$ y ha de ser igual al tiempo que nuestro cliente ha pasado en el sistema, W.

El razonamiento es análogo para la cola y el servicio.

Otra explicación se basa en la idea de que los clientes han de pagar por estar en el sistema. En este caso, esperamos que se cumpla

$$\left(\begin{array}{c}\text{Tasa media de}\\ \text{ingresos del sistema}\end{array}\right) = \lambda \times \left(\begin{array}{c}\text{Cantidad media}\\ \text{pagada por cliente}\end{array}\right)$$

Entonces, si cada cliente paga 1 euro por cada unidad de tiempo que pasa en el sistema, la igualdad anterior se convierte en $L = \lambda W$; si paga 1 euro por cada unidad de tiempo que pasa en la cola, se obtiene $L_q = \lambda W_q$ y si paga 1 euro por cada unidad de tiempo que pasa en el servidor, se verifica $L_s = \lambda W_s$

Si multiplicamos por λ la igualdad $W = W_q + E(s)$, obtenemos $\lambda W = \lambda W_q + \lambda E(s)$. Por las fórmulas de Little, resulta

$$L = L_q + r$$

que sirve para cualquier modelo $G/G/c$ y nos dice que el número medio de clientes en el sistema es el número medio de clientes en cola más el número medio de clientes en los servidores.

6. Comportamiento de transición y estacionario. Probabilidades límites. Propiedad PASTA

Nótese que siempre asumimos que el sistema ha alcanzado el equilibrio, pues si estuviera en estado transitorio habría que considerar la dependencia del tiempo que el sistema lleva funcionando y de las condiciones iniciales.

Hay dos variantes de las probabilidades π_n:

- proporción a_n ($n \geq 0$) de clientes que encuentran n en el sistema al llegar
- proporción b_n ($n \geq 0$) de clientes que dejan n en el sistema al salir.

a_n corresponde a lo que una llegada observa,
b_n a lo que una salida observa y
π_n a lo que observaría alguien desde fuera.

Las tres probabilidades no tienen por qué coincidir.

Proposición. En un sistema en el que los clientes llegan de uno en uno y se sirven de uno en uno, $a_n = b_n, \forall n$.

Demostración. Cuando el sistema pasa de tener n clientes a n+1 es cuando una llegada ve n clientes en el sistema. De igual modo, cuando pasa de n +1 a n es cuando una salida deja n en el sistema.

En cualquier intervalo de tiempo, el número de transiciones de n a n +1 se diferencia en 1 del número de transiciones de n +1 a n.

Por tanto, las tasas de transiciones de n a n +1 y de n +1 a n coinciden, o equivalentemente, también lo hacen la tasa a la que las llegadas encuentran n clientes y la tasa a la que las salidas dejan n clientes.
Se llega así al resultado deseado porque las tasas de llegadas y salidas globales deben coincidir.

En media, las llegadas y salidas ven siempre el mismo número de clientes, aunque en general no ven tiempos medios. Hace falta que las llegadas sean de Poisson para que coincidan las tres probabilidades a_n, b_n y π_n.

Esta **propiedad** se denomina **PASTA** (Poisson Arrivals See Time Averages): Las llegadas de Poisson siempre ven tiempos medios, es decir, $\pi_n = a_n, \forall n$.

Demostración. Si $A(t, t+\Delta t)$ denota la llegada de un cliente en el intervalo $(t, t+\Delta t)$, se tiene

$$\begin{aligned} a_n(t) &= \lim_{\Delta t \to 0} P(N(t) = n \mid A(t, t+\Delta t)) \\ &= \lim_{\Delta t \to 0} \frac{P(A(t, t+\Delta t) \mid N(t) = n) P(N(t) = n)}{P(A(t, t+\Delta t))} \\ &= \lim_{\Delta t \to 0} \frac{P(A(t, t+\Delta t)) P(N(t) = n)}{P(A(t, t+\Delta t))} = P(N(t) = n) = p_n(t). \end{aligned}$$

La tercera igualdad se debe a la propiedad de pérdida de memoria de la distribución exponencial, al ser las llegadas de Poisson.
Hemos obtenido que la distribución de lo que una llegada en el tiempo t observa es la misma que la distribución del estado del sistema en el tiempo t. De ahí, a largo plazo,

$$a_n = \lim_{t \to \infty} a_n(t) = \lim_{t \to \infty} p_n(t) = \pi_n.$$

Por tanto, la distribución de probabilidad que observarán los clientes al llegar será la distribución de probabilidad a lo largo del tiempo.

La tarea del analista de sistemas de colas consiste en caracterizar adecuadamente el modelo que va a utilizar para analizar el sistema real bajo estudio y todas las variables implicadas, pero también los costes asociados.

Una vez realizado este estudio, si el analista pudiera encontrar una relación a optimizar, de acuerdo a algún criterio, llegaría a determinar un sistema óptimo.

Sin embargo, no siempre es fácil llegar a este tipo de relación. En sistemas complejos nos deberemos conformar con formular diversas soluciones alternativas y obtener cuál de ellas es la mejor según ciertos criterios prefijados.

Cuando el problema no pueda resolverse por medios analíticos, recurriremos a métodos numéricos o a la simulación.