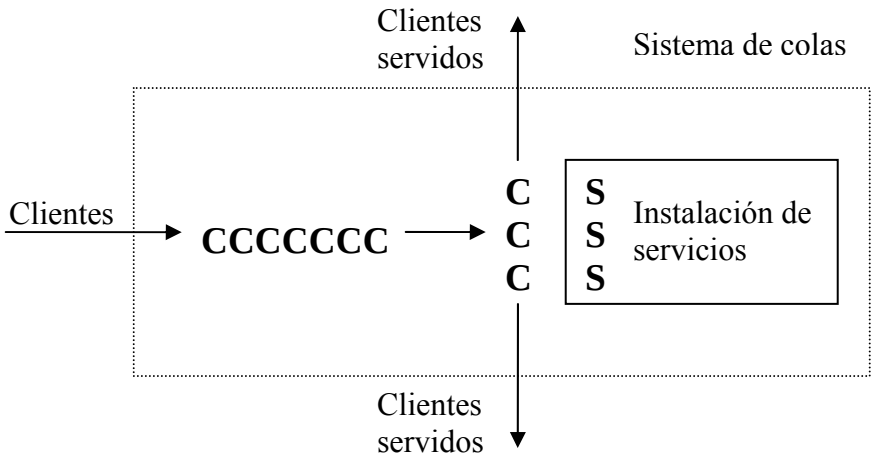
MODELOS DE COLAS

¿Cuándo se producen colas?
Cuando la demanda excede la capacidad de servicio

La teoría de colas proporciona modelos matemáticos que describen el comportamiento estable de las líneas de espera creadas para la adquisición de determinados servicios

El objetivo del analista es establecer un balance entre:

- El costo del servicio
- El costo de la espera de dicho servicio



Diferentes tipos de sistemas que se pueden trabajar con modelos de colas

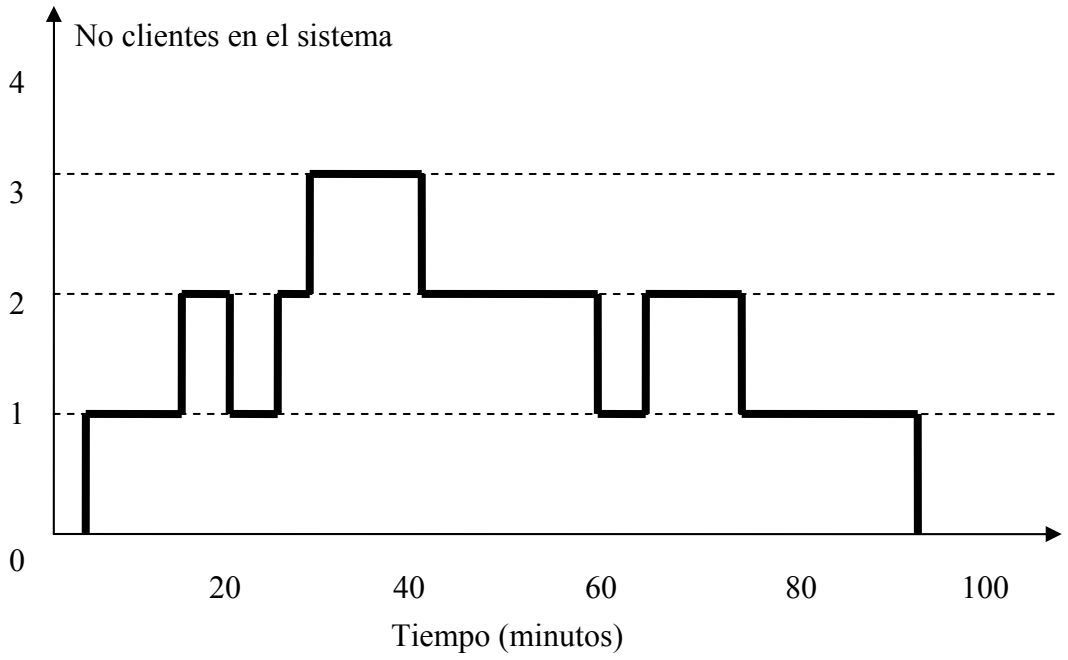
- Servicios comerciales: bancos, supermercados, cafeterías, etc.
- Servicios de transporte: centros de carga y descarga (aeropuertos, puertos, mercados), tráfico (cruces, autopistas), estacionamiento, ascensores, etc.
- Sistemas industriales y de negocios: servicios de mantenimiento, servicios de inspección, almacenes.
- Servicios sociales: hospitales, juzgados.

Ejemplo de un sistema de colas:

La sala de masajes de un reconocido gimnasio, cuenta con una sola camilla para atender a sus clientes. Esta sala abre a las 8:00 a.m. de cada día hábil de la semana. La tabla siguiente muestra los datos referidos a las llegadas, inicio del servicio, tiempo del servicio y finalización del servicio.

Cliente	Tiempo de llegada	Hora inicio servicio	Duración del servicio	Hora finaliza servicio
1	8:03	8:03	17 min.	8:20
2	8:15	8:20	21 min.	8:41
3	8:25	8:41	19 min.	9:00
4	8:30	9:00	15 min.	9:15
5	9:05	9:15	20 min.	9:35
6	9:43	-	-	-

El siguiente gráfico muestra el número de clientes en este sistema de colas durante los 100 primeros minutos.



Al recopilar más datos en la sala de masajes, se encuentra que llegaron 300 clientes durante un período de 100 horas.

Estructuras de los modelos de colas

Llegadas:

Los tiempos que transcurren entre dos llegadas consecutivas a un sistema de colas se llaman **tiempos entre llegadas.**
En los sistemas de colas es común que estos tengan un gran variabilidad en los tiempos entre llegadas, por lo que se hace imposible predecir la llegada del próximo cliente al sistema; sin embargo, después de recopilar muchos datos es posible hacer dos cosas:

a. Estimar el número esperado de llegadas por unidad de tiempo (tasa media de llegadas, denotada por la letra griega λ lambda)

b. Estimar la forma de distribución de probabilidad de los tiempos entre llegadas.

El hecho de que la última llegada no influya en la probabilidad de una llegada en el siguiente minuto se llama propiedad de falta de memoria (o propiedad markoviana). Esta propiedad implica que la distribución de probabilidad del tiempo restante desde ahora hasta que ocurra la siguiente llegada es siempre la misma (llegadas aleatorias). La única distribución de tiempos entre llegadas que se ajusta a las llegadas aleatorias es la distribución exponencial.

Considerando los datos del ejemplo, tenemos que:

λ = tasa media de llegadas, la media es: $1/\lambda$ = tiempo esperado entre llegadas

λ = 300 clientes / 100 horas = 3 clientes por hora

$1/\lambda$ = 1 / 3 horas entre clientes en promedio

Cola:

La cola es donde esperan los clientes antes de ser atendidos (se generan colas solo si la demanda excede la capacidad de servicio).
Existen dos formas de contar los clientes: el número de clientes en la cola (o tamaño de la cola) el cual es el número de clientes que esperan para iniciar el servicio; y el número de clientes en el sistema, el cual es el número de clientes en la cola más el número actualmente en servicio.

La capacidad de la cola es el número máximo de clientes que pueden estar en ella. Existen colas infinitas (teóricamente, para efectos prácticos) y colas finitas. Cuando una cola finita esta llena, cualquier cliente que llegue puede tomar la decisión de marcharse inmediatamente.

Por convención los modelos de colas se suponen que las colas son infinitas.

La disciplina de la cola se refiere al orden en que se seleccionan los miembros de la cola para comenzar el servicio. La más común es primero en entrar, primero en servir.

Servicio:

Para un sistema de colas básico, cada cliente se sirve en forma individual por uno de los servidores. Un sistema con más de un servidor se llama sistema con servidores múltiples, mientras que un sistema de un servidor tiene sólo un servidor.

Cuando un cliente entra a servicio, el tiempo transcurrido de principio a fin del servicio se llama **tiempo de servicio**. Por lo general el tiempo de servicio varía de un cliente al que sigue. El símbolo que se utiliza para la media de la distribución del tiempo de servicio es
$1/\mu$ = tiempo de servicio esperado

donde μ es la letra griega mu, la cual es el número esperado de terminaciones de servicio por unidad de tiempo para un servidor continuamente ocupado. Donde esta cantidad se conoce como la **tasa media de servicio**.

Considerando los datos del ejemplo, tenemos que:

$1/\mu$ = 20 minutos = 1/3 horas por cliente

μ = 3 clientes por hora

Distribuciones de tiempo de servicio

La opción más popular es utilizar la distribución exponencial, pero en sistemas donde el tiempo de servicio es constante (aproximadamente iguales – ejemplo: una taquilla de un banco) se utiliza la distribución degenerada.

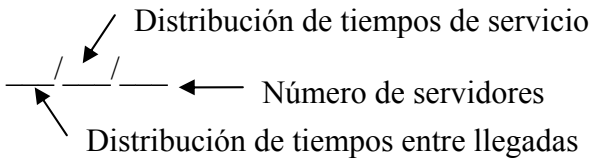
La distribución exponencial y degenerada representan dos casos extremos respecto a la cantidad de variabilidad en los tiempos de servicio. Una medida estándar de la cantidad de variabilidad es la desviación estándar de la distribución (σ letra griega simma).

Para muchos sistemas de colas, la cantidad de variabilidad en los tiempos de servicio está en algún punto entre la de las distribuciones exponenciales y degenerada. Una distribución intermedia es la distribución Erlang, la cual tiene un parámetro k llamado el parámetro de forma.

Para k=1, la distribución es equivalente a la exponencial y cuando tiende a infinito, equivale a la distribución degenerada.

Distribución	**Desviación estándar**
Exponencial	Media (1/μ)
Degenerada (constante)	0
Erlang, cualquier k (k=1,2,3,…)	1/√k media

Formato utilizado para identificar los modelos de colas



Los símbolos utilizados para las distribuciones son:

M = distribución exponencial (markoviana)
D = distribución degenerada (tiempos constantes)
E_k = distribución Erlang (parámetro de forma k)

GI = distribución general independiente de tiempo entre llegadas (permite cualquier distribución arbitraria)
G = distribución general de tiempo de servicio (permite cualquier distribución arbitraria)

Suposiciones generales de un modelo de colas básico

1. Los tiempos de llegadas son independientes e idénticamente distribuidos de acuerdo a una distribución de probabilidad especificada.
2. Todos los clientes que llegan entran al sistema de colas y permanecen ahí hasta que termina el servicio.
3. El sistema de colas tiene una sola cola infinita, de modo que en la cola puede haber un número ilimitado de clientes (para propósito práctico).
4. La disciplina de la cola es primero en entrar, primero en servir.
5. El sistema de colas tiene un número especificado de servidores, donde cada uno es capaz de servir a cualquiera de los clientes.
6. Cada cliente es atendido en forma individual por cualquiera de los servidores.
7. Los tiempos de servicio son independientes e idénticamente distribuidos de acuerdo con la distribución de probabilidad especificada.

Medidas de desempeño de los sistemas de colas

Las medidas de desempeño más utilizadas, responden a las siguientes preguntas:

1. ¿Cuántos clientes suelen esperar en el sistema de colas?
2. ¿Cuánto suelen esperar estos clientes?

Estas dos medidas generalmente se expresan en términos de sus valores esperados (en sentido estadístico), y se deben considerar diferentes, a. el caso donde se mide el número de personas que están en cola esperando por el servicio, o b. mientras se encuentra en cualquier punto del sistema de colas. Estas dos formas de medir el desempeño, genera cuatro medidas:

L = número esperado de clientes en el sistema (incluye a quienes están en el servicio – L proviene de longitud)
L_q = número esperado de clientes en la cola (no incluye a los clientes que están en servicio)
W = tiempo de espera esperado en el sistema (incluye el tiempo en el servicio) para un cliente individual (W proviene de tiempo de espera)
W_q = tiempo de espera esperado en la cola (excluye el tiempo de servicio) para un cliente individual

Estas definiciones suponen que el sistema de colas está en condición de estado estable.

Relaciones entre estas medidas:

$W = W_q + 1/\mu$

Ejemplo:

- W_q = ¾ hora de espera en cola en promedio
- $1/\mu$ = ¼ hora de tiempo de servicio en promedio

W = ¾ hora + ¼ hora = 1 hora de espera en el sistema de colas en promedio

$L = \lambda W$

Ejemplo:

- W = 1 hora de espera en el sistema de colas en promedio
- λ = 3 clientes por hora que llegan en promedio

L = (3 clientes/hora)(1 hora) = 3 clientes en el sistema de colas en promedio

$L_q = \lambda W_q$

Ejemplo:

- W_q = ¾ hora de espera en cola en promedio
- λ = 3 clientes por hora que llegan en promedio

L_q = (3 clientes/hora)(3/4 de hora) = 2 ¼ clientes en la cola en promedio

$L = \lambda W = \lambda (W_q + 1/\mu) = L_q + \lambda/\mu$

Ejemplo:

- L_q = 2 ½, λ = 3 y μ = 4

L = 3 clientes promedio en el sistema