TEORÍA DE COLAS

Nota Técnica de Apoyo nº 1.1

1ª Ley de Harper

No importa en qué cola se sitúe: la otra siempre avanzará más rápido

2ª Ley de Harper

Y si se cambia de cola, aquélla en que estaba al principio empezará a ir más deprisa

1. Introducción

La Teoría de Colas es un formulación matemática para la optimización de sistemas en que interactúan dos procesos normalmente aleatorios: un proceso de llegada de clientes y un proceso de servicio a los clientes, en los que existen fenómenos de acumulación de clientes en espera del servicio, y donde existen reglas definidas (prioridades) para la prestación del servicio.

La Teoría de Colas es una aproximación matemática potente para la optimización del problema, y tiene aplicaciones (crecientes) en sistemas donde las llegadas y el servicio admiten una representación matemática (probabilística); en problemas que no admiten esta representación existen otras técnicas, como muestra la tabla siguiente:

Menos Complejidad / Coste de análisis Más

Heurísticos	Modelos de aproximación lineal	Teoría de Colas: modelos analíticos	Simulación (benchmarking)

Heurísticos: reglas derivadas de la experiencia que no tratan de optimizar el problema sino de dar directrices simples para una aplicación razonablemente eficaz. Por ejemplo, se suele utilizar una regla heurística en el pre−diseño de circuitos electrónicos con canales de E/S, consistente en dimensionar los canales E/S de forma que su ocupación no supere el 35% en aplicaciones con acceso on−line a sistemas de almacenamiento externo, o del 40% si las aplicaciones son batch.

Modelos lineales: obtienen valores medios de representación de las variables, y bajo las hipótesis de comportamiento lineal ante variaciones en los flujos se extrapolan los niveles de capacidad / utilización.

Simulación: construcción de modelos detallados de simulación por ordenador. En particular, donde el modelo se basa en datos obtenidos de aplicaciones reales se habla de técnicas de benchmarking: por ejemplo, en el diseño de las oficinas bancarias.

Colas: concepto intuitivo de línea de espera, equivalente al británico (queues) y al americano (waiting lines).

Origen de la Teoría de Colas: trabajos de A. K. Earlang (Dinamarca, 1.905) estudiando el problema de dimensionamiento de líneas y centrales de conmutación telefónica para el servicio de llamadas.

Los problemas de Colas se presentan permanentemente en todas las aplicaciones de la vida diaria: un estudio de EE.UU. concluyó que un ciudadano medio pasa 5 años de su vida esperando en distintas Colas, y de ellos

casi 6 meses parado en los semáforos. Problemas típicos de Teoría de Colas son:

Programación de actividades de despegue / aterrizaje en un aeropuerto

Sistema de consulta médica

Piezas en un taller donde pasan por diferentes máquinas en el proceso de mecanizado

Sistema de cajas en una oficina bancaria

2. Costes asociados a un sistema de Colas

¿Por qué es necesario contar con herramientas de optimización para los problemas de Colas?

Normalmente en cualquiera de estos sistemas existen dos familias de costes:

a) Los costes asociados a la espera de los clientes

Por ejemplo, el valor del tiempo perdido o la gasolina malgastada en los atascos o los semáforos, o las horas perdidas en las Colas de las urnas electorales (valor normalmente estimado).

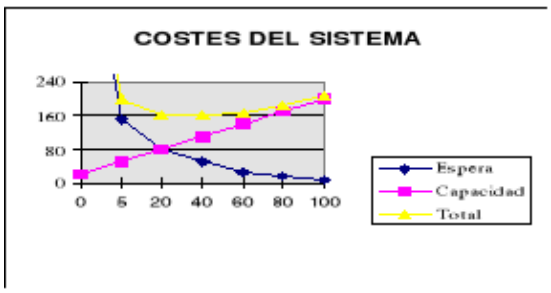
La hipótesis natural establece que estos costes de la espera decrecen conforme aumenta la capacidad de servicio del sistema: por ejemplo, conforme aumenta el número de médicos de cabecera en un ambulatorio más corto es el tiempo de espera de los pacientes, y el coste de oportunidad del tiempo perdido decrece.

b) Los costes asociados a la expansión de la capacidad de servicio

Contra la reducción anterior de costes de espera, es también normal que el coste asociado a incrementar la capacidad de servicio crezca con alguna proporcionalidad en relación a esta capacidad; en el ejemplo anterior, los costes de salarios, despachos, enfermeras ayudantes, etc. ligados al aumento del número de médicos son casi directamente proporcionales al número de médicos (o con una parte fija y otra directamente proporcional).

c) Los costes totales del sistema de servicio

La suma de los dos costes anteriores da una función de costes totales del sistema en función de la capacidad, que tendrá una forma similar a la siguiente:



3. Objetivos de la Teoría de Colas

Dada la función de costes anterior, los objetivos de la Teoría de Colas consisten en:

Identificar el nivel óptimo de capacidad del sistema que minimiza el coste global del mismo.

Evaluar el impacto que las posibles alternativas de modificación de la capacidad del sistema tendrían en el coste total del mismo.

Establecer un balance equilibrado (óptimo) entre las consideraciones cuantitativas de costes y las cualitativas de servicio.

El objetivo de coste es claro; entre los objetivos de servicio se suelen plantear aspectos medibles (llamados medidas duras) y hay otros (medidas blandas) que hay que valorar externamente; algunas medidas duras típicas de sistemas de Colas son:

Tasa de ocupación de las estaciones de servicio: las ocupaciones admisibles dependen del tipo de sistema, y es claro que no son estándares: la ocupación permanente de un sistema automático como una barrera de aparcamiento no puede ser la misma que la de un médico en una consulta.

Número de clientes en el sistema o en la Cola: hay límites (en ocasiones hasta físicos) al tamaño de una Cola, que también dependen del tipo de servicio. En casos en que hay restricciones al tamaño de la Cola (por ejemplo en una gasolinera en el centro de la ciudad) una medida importante será la proporción de clientes servidos en relación a los potenciales (llegados al sistema).

Tiempo de permanencia en el sistema o en la Cola: la paciencia de los clientes depende del tipo de servicio específico considerado.

4. Elementos de un Sistema de Colas

El Sistema de Colas incluye tanto las estaciones de servicio y los clientes en ellas, como la propia cola y los clientes en espera:

Población Cola Servidores

SISTEMA DE COLAS

Los elementos del Sistema de Colas son los siguientes:

a) Población

La población puede clasificarse (y las técnicas de Colas difieren) en función de su tamaño relativo, como finita o infinita: será infinita cuando el número de clientes potenciales es muy grande en relación a la capacidad del sistema; en caso contrario, será finita.

La importancia de la diferenciación entre población finita e infinita radica en que, en poblaciones finitas, las probabilidades de llegada de un cliente (o de ocurrencia de un suceso) varían según el estado del sistema: por ejemplo, si hay seis máquinas en un servicio de mantenimiento y una de ellas está rota (en reparación) la probabilidad de rotura de otra es diferente.

b) Proceso de llegada de los clientes

Las llegadas de clientes al sistema son en la mayoría de las ocasiones controlables: por ejemplo, hay sistemas que juegan con los precios, o con la capacidad / comodidad, o con ofertas; en casos hipotéticamente

incontrolables como las llegadas de urgencias a una UVI se toman acciones previas sobre el sistema de ambulancias para comunicar el estado / la saturación de las instalaciones y desviar pacientes a otros hospitales.

Normalmente la Teoría de Colas opera sobre los tiempos entre llegadas consecutivas de clientes: modelos típicos son el teórico de llegadas a intervalos fijos iguales, o los que consideran diferentes distribuciones de probabilidad.

Asimismo, las llegas pueden ser individuales (un único cliente en cada llegada) o múltiples (varios clientes en una misma llegada).

c) Línea de espera o Cola

Como se ha dicho, la Cola viene definida en primer lugar por la forma de llegada de los clientes (con / sin distribución conocida, perfil de la distribución).

Por otra parte el Sistema se define también por la conducta del cliente potencial ante la Cola; los tipos de cliente en relación a la conducta se denominan:

Impaciente	Si hay Cola abandona el Sistema
Paciente / rechazo	Si la Cola supera un límite definido para cada cliente, abandona el Sistema
Paciente / abandono	Aguanta la Cola durante un cierto tiempo
Paciente / Permanencia	Aguanta hasta ser atendido

d) Capacidad de la Cola

El caso teórico más simple es el de cola de capacidad infinita; existen múltiples casos de Colas de longitud acotada (por ejemplo un restaurante drive−in, o un taller mecánico). Un enfoque matemático simplificador consiste en tratar los Sistemas con capacidad finita como si fueran de capacidad infinita cuando se evalúa la probabilidad de llenado de la capacidad de la Cola como muy baja.

e) Proceso de servicio

Se caracteriza la distribución de tiempos de duración del servicio; los modelos más utilizados emplean una distribución exponencial (luego se discutirá).

f) Reglas de servicio

Las reglas más utilizadas son:

FIFO (primero en llegar, primero en ser servido). Se percibe como la más justa en los sistemas de Colas más habituales.

LIFO: por ejemplo en productos perecederos en que se consulta la fecha de caducidad.

Existen otras reglas que se caracterizan por la ruptura de la disciplina de Cola: por ejemplo casos en que hay clientes privilegiados (urgencias hospitalarias) donde se puede situar al cliente prioritario como primero de la Cola (prioridad débil), o incluso sustituir al cliente en servicio actual si es de prioridad inferior como en las UVIs (prioridad fuerte).

Otros modelos más sofisticados contemplan estaciones de servicio específicas para determinados segmentos de clientes, o puestos reservados, etc.

g) Número de estaciones de servicio

En función del número de estaciones (canales) de servicio y de las fases del proceso de servicio, tenemos los siguientes tipos de problemas de Colas:

Canales	**Fases**	**Ejemplos típicos**
Uno	Una	Kiosco de prensa con un empleado
Uno	Varias	Lavado / secado de coches
Varios	Una	Oficina bancaria con varios cajeros
Varios	Varias	Centro de servicios radiológicos de hospital

5. Denominación de los problemas de Teoría de Colas

La denominación simplificada de los problemas se basa en tres códigos:

Texto 1 / Texto 2 / Número

donde Texto 1 define el proceso de llegada (M aleatorio de Markov, G distribución genérica), Texto 2 define el proceso de servicio (igual M o G), y Número define el número de puestos de servicio. En esta denominación se supone que todos los puestos son idénticos y operan en una sola fase (paralelo).

6. Procesos de Poisson

La definición de la tasa de llegada de clientes al Sistema de Colas se realiza describiendo la probabilidad de ocurrencia de un acontecimiento (una llegada) en un intervalo de tiempo determinado; la distribución de Poisson caracteriza una gran parte de fenómenos reales que cumplen las siguientes condiciones:

a) Las llegadas ocurren distribuidas en el tiempo de forma discreta.

b) Las llegadas son independientes: el hecho de que se haya producido una llegada en un instante no condiciona las llegadas en instantes posteriores.

c) Dada una duración de intervalo de tiempo fija, la probabilidad de ocurrencia de una llegada durante este intervalo es constante: así, dado un período de 2 minutos, la probabilidad de una llegada entre las 09:00 y las 09:02 es la misma que entre las 21:17 y las 21:19. Esta característica, que puede parecer poco realista para largos períodos de tiempo (por ejemplo un día) sí es representativa para intervalos menores (2 horas punta de un servicio hospitalario): los problemas deben representarse durante períodos de tiempo en que esta condición se cumpla.

d) Para intervalos de tiempo suficientemente cortos, la probabilidad de una llegada en el intervalo es proporcional a la duración del intervalo, y la probabilidad de más de una llegada en el intervalo es despreciable.

La distribución de Poisson describe el número de llegadas en un intervalo de tiempo, bajo las condiciones anteriores. La distribución de Poisson se define por medio del parámetro , que mide el número esperado de llegadas en el plazo total de tiempo del experimento. Dada , la función de distribución de Poisson viene dada por: