UNIVERSIDAD ALEJANDRO DE HUMBOLDT

INVESTIGACIÓN DE OPERACIONES

SECCIÓN: 0501AE

PROFESORA: NINOSKA KEY

TEORÍA DE COLAS

Autor:
Matías Martínez 82.071.517

Caracas, 10 de Noviembre de 2004

Indice

“No importa en qué cola se sitúe: La otra siempre avanzará más rápido”
(Primera Ley de Harper)

“Y si se cambia de cola, aquélla en la que estaba al principio empezará a ir más deprisa” (Segunda Ley de Harper)

INTRODUCCIÓN

Las "colas" son un aspecto de la vida moderna que nos encontramos continuamente en nuestras actividades diarias. En el contador de un supermercado, accediendo al Metro, en los Bancos, etc., el fenómeno de las colas surge cuando unos recursos compartidos necesitan ser accedidos para dar servicio a un elevado número de trabajos o clientes.

El estudio de las colas es importante porque proporciona tanto una base teórica del tipo de servicio que podemos esperar de un determinado recurso, como la forma en la cual dicho recurso puede ser diseñado para proporcionar un determinado grado de servicio a sus clientes.

Debido a lo comentado anteriormente, se plantea como algo muy útil el desarrollo de una herramienta que sea capaz de dar una respuesta sobre las características que tiene un determinado modelo de colas.

Definiciones iniciales

La **teoría de colas** es el estudio matemático del comportamiento de líneas de espera. Esta se presenta, cuando los “clientes” llegan a un “lugar” demandando un servicio a un

“servidor”, el cual tiene una cierta capacidad de atención. Si el servidor no está disponible inmediatamente y el cliente decide esperar, entonces se forma la línea de espera.

Una **cola** es una línea de espera y la teoría de colas es una colección de modelos matemáticos que describen sistemas de línea de espera particulares o sistemas de colas. Los modelos sirven para encontrar un buen compromiso entre costes del sistema y los tiempos promedio de la línea de espera para un sistema dado.

Los **sistemas de colas** son modelos de sistemas que proporcionan servicio. Como modelo, pueden representar cualquier sistema en donde los trabajos o clientes llegan buscando un servicio de algún tipo y salen después de que dicho servicio haya sido atendido. Podemos modelar los sistemas de este tipo tanto como colas sencillas o como un sistema de colas interconectadas formando una red de colas. En la siguiente figura podemos ver un ejemplo de modelo de colas sencillo. Este modelo puede usarse para representar una situación típica en la cual los clientes llegan, esperan si los servidores están ocupados, son servidos por un servidor disponible y se marchan cuando se obtiene el servicio requerido.

El problema es determinar qué capacidad o tasa de servicio proporciona el balance correcto. Esto no es sencillo, ya que un cliente no llega a un horario fijo, es decir, no se sabe con exactitud en que momento llegarán los clientes. También el tiempo de servicio no tiene un horario fijo.

Los problemas de “colas” se presentan permanentemente en la vida diaria: un estudio en EEUU concluyó que, por término medio, un ciudadano medio pasa cinco años de su vida esperando en distintas colas, y de ellos casi seis meses parado en los semáforos.

Introducción a la Teoría de Colas

En muchas ocasiones en la vida real, un fenómeno muy común es la formación de colas o líneas de espera. Esto suele ocurrir cuando la demanda real de un servicio es superior a la capacidad que existe para dar dicho servicio. Ejemplos reales de esa situación son: los cruces de dos vías de circulación, los semáforos, el peaje de una autopista, los cajeros automáticos, la atención a clientes en un establecimiento comercial, la avería de electrodomésticos u otro tipo de aparatos que deben ser reparados por un servicio técnico, etc.

Todavía más frecuentes, si cabe, son las situaciones de espera en el contexto de la informática, las telecomunicaciones y, en general, las nuevas tecnologías. Así, por ejemplo, los procesos enviados a un servidor para ejecución forman colas de espera mientras no son atendidos, la información solicitada, a

través de Internet, a un servidor Web puede recibirse con demora debido a congestión en la red o en el servidor propiamente dicho, podemos recibir la señal de líneas ocupadas si la central de la que depende nuestro teléfono móvil está colapsada en ese momento, etc.

Origen:

El origen de la Teoría de Colas está en el esfuerzo de Agner Kraup Erlang (Dinamarca, 1878 - 1929) en 1909 para analizar la congestión de tráfico telefónico con el objetivo de cumplir la demanda incierta de servicios en el sistema telefónico de Copenhague. Sus investigaciones acabaron en una nueva teoría denominada teoría de colas o de líneas de espera. Esta teoría es ahora una herramienta de valor en negocios debido a que un gran número de problemas pueden caracterizarse, como problemas de congestión llegada-salida.

Modelo de formación de colas.

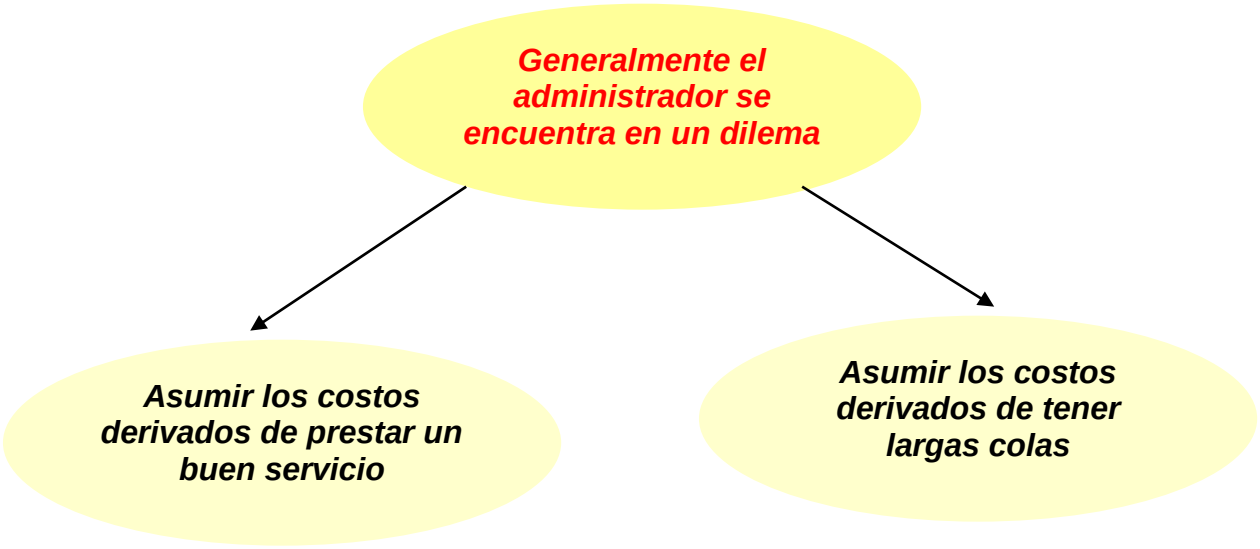
En los problemas de formación de cola, a menudo se habla de clientes, tales como personas que esperan la desocupación de líneas telefónicas, la espera de máquinas para ser reparadas y los aviones que esperan aterrizar y estaciones de servicios, tales como mesas en un restaurante, operarios en un taller de reparación, pistas en un aeropuerto, etc. Los problemas de formación de colas a menudo contienen una velocidad variable de llegada de clientes que requieren cierto tipo de servicio, y una velocidad variable de prestación del servicio en la estación de servicio.

Cuando se habla de líneas de espera, se refieren a las creadas por clientes o por las estaciones de servicio. Los clientes pueden esperar en cola simplemente por que los medios existentes son inadecuados para satisfacer la demanda de servicio; en este caso, la cola tiende a ser explosiva, es decir, a ser cada vez mas larga a medida que transcurre el tiempo. Las estaciones de servicio pueden estar esperando por que los medios existentes son excesivos en relación con la demanda de los clientes; en este caso, las estaciones de servicio podrían permanecer ociosas la mayor parte del tiempo. Los clientes puede que esperen temporalmente, aunque las instalaciones de servicio sean adecuadas, por que los

clientes llegados anteriormente están siendo atendidos. Las estaciones de servicio pueden encontrar temporal cuando, aunque las instalaciones sean adecuadas a largo plazo, haya una escasez ocasional de demanda debido a un hecho temporal. Estos dos últimos casos tipifican una situación equilibrada que tiende constantemente hacia el equilibrio, o una situación estable.

En la teoría de la formación de colas, generalmente se llama sistema a un grupo de unidades físicas, integradas de tal modo que pueden operar al unísono con una serie de operaciones organizadas. La teoría de la formación de colas busca una solución al problema de la espera prediciendo primero el comportamiento del sistema. Pero una solución al problema de la espera consiste en no solo en minimizar el tiempo que los clientes pasan en el sistema, sino también en minimizar los costos totales de aquellos que solicitan el servicio y de quienes lo prestan.

La teoría de colas incluye el estudio matemático de las colas o líneas de espera y provee un gran número de modelos matemáticos para describirlas.



Se debe lograr un balance económico entre el costo del servicio y el costo asociado a la espera por ese servicio

La teoría de colas en sí no resuelve este problema, sólo proporciona información para la toma de decisiones

Objetivos de la Teoría de Colas

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.
- Hay que prestar atención al tiempo de permanencia en el sistema o en la cola: la “paciencia” de los clientes depende del tipo de servicio específico considerado y eso puede hacer que un cliente “abandone” el sistema.

Elementos existentes en un modelo de colas

Fuente de entrada o población potencial: Es un conjunto de individuos (no necesariamente seres vivos) que pueden llegar a solicitar el servicio en cuestión. Podemos considerarla finita o infinita. Aunque el caso de infinitud no es realista, sí permite (por extraño que parezca) resolver de forma más sencilla muchas situaciones en las que, en realidad, la población es finita pero muy grande. Dicha suposición de infinitud no resulta restrictiva cuando, aún siendo finita la población potencial, su número de elementos es tan grande que el número de individuos que ya están solicitando el citado servicio prácticamente no afecta a la frecuencia con la que la población potencial genera nuevas peticiones de servicio.

Cliente: Es todo individuo de la población potencial que solicita servicio. Suponiendo que los tiempos de llegada de clientes consecutivos son $0<t1<t2<...$, será importante conocer el patrón de probabilidad según el cual la fuente de entrada genera clientes. Lo más habitual es tomar como referencia los tiempos entre las llegadas de dos clientes consecutivos: consecutivos: clientes consecutivos: $T\{k\} = tk - tk-1$, fijando su distribución de probabilidad. Normalmente, cuando la población potencial es infinita se supone que la distribución de probabilidad de los Tk (que será la llamada distribución de los tiempos entre llegadas) no depende del número de clientes que estén en espera de completar su servicio, mientras que en el caso de que la fuente de entrada sea finita, la distribución de los Tk variará según el número de clientes en proceso de ser atendidos.

Capacidad de la cola: Es el máximo número de clientes que pueden estar haciendo cola (antes de comenzar a ser servidos). De nuevo, puede suponerse finita o infinita. Lo más sencillo, a efectos de simplicidad en los cálculos, es suponerla infinita. Aunque es obvio que en la mayor parte de los casos reales la capacidad de la cola es finita, no es una gran restricción el suponerla infinita si es extremadamente improbable que no puedan entrar clientes a la cola por haberse llegado a ese número límite en la misma.

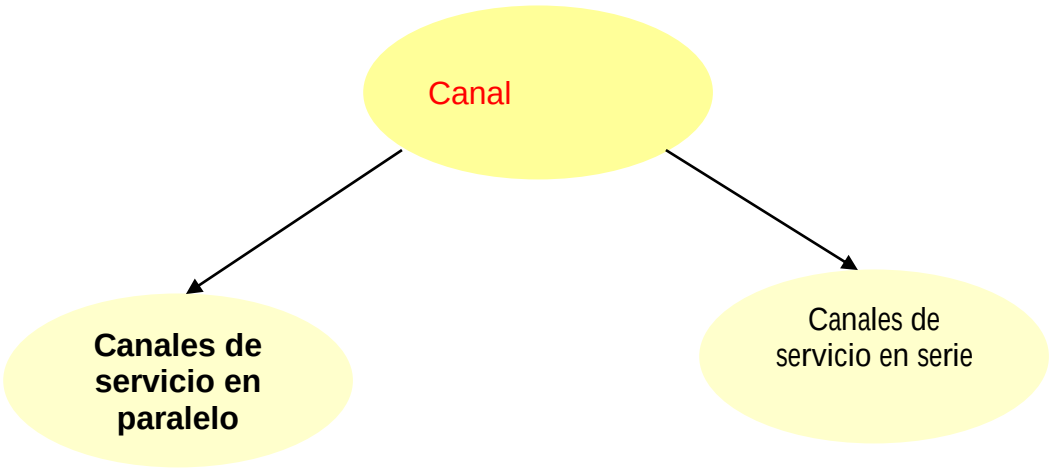
Disciplina de la cola: Es el modo en el que los clientes son seleccionados para ser servidos. Las disciplinas más habituales son:

La disciplina FIFO (first in first out), también llamada FCFS (first come first served): según la cual se atiende primero al cliente que antes haya llegado.

La disciplina LIFO (last in first out), también conocida como LCFS (last come first served) o pila: que consiste en atender primero al cliente que ha llegado el último.

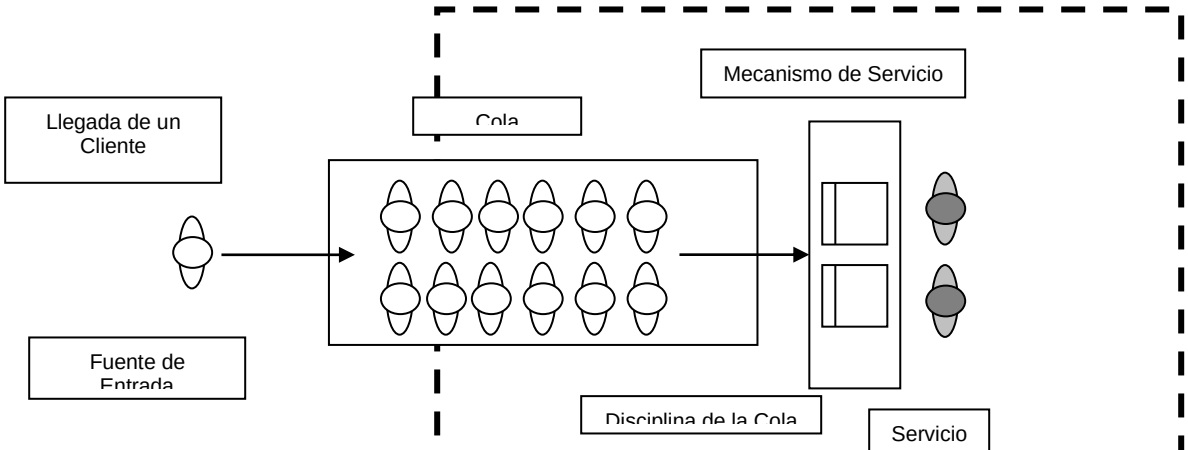
La RSS (random selection of service), o SIRO (service in random order), que selecciona a los clientes de forma aleatoria.

Mecanismo de servicio: Es el procedimiento por el cual se da servicio a los clientes que lo solicitan. Para determinar totalmente el mecanismo de servicio debemos conocer el número de servidores de dicho mecanismo (si dicho número fuese aleatorio, la distribución de probabilidad del mismo) y la distribución de probabilidad del tiempo que le lleva a cada servidor dar un servicio. En caso de que los servidores tengan distinta destreza para dar el servicio, se debe especificar la distribución del tiempo de servicio para cada uno.



La cola, propiamente dicha, es el conjunto de clientes que hacen espera, es decir los clientes que ya han solicitado el servicio pero que aún no han pasado al mecanismo de servicio.

El sistema de la cola: es el conjunto formado por la cola y el mecanismo de servicio, junto con la disciplina de la cola, que es lo que nos indica el criterio de qué cliente de la cola elegir para pasar al mecanismo de servicio. Estos elementos pueden verse más claramente en la siguiente figura:

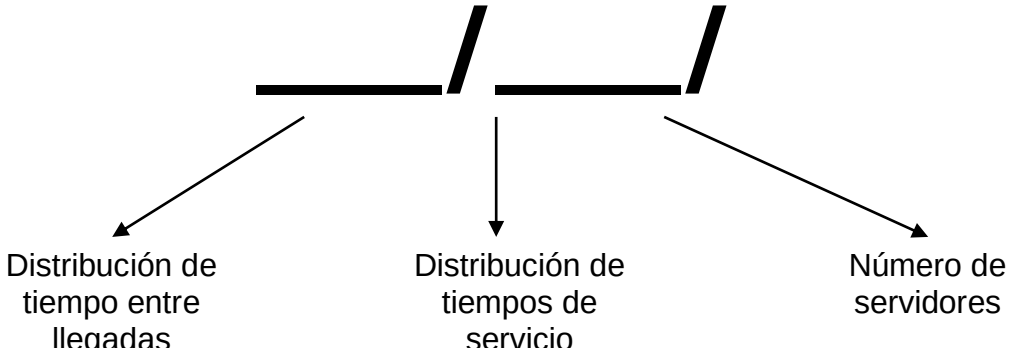


Un modelo de sistema de colas debe especificar la distribución de probabilidad de los tiempos de servicio para cada servidor.

La distribución más usada para los tiempos de servicio es la *exponencia*l, aunque es común encontrar la distribución *degenerada o determinística* (tiempos de servicio constantes) o la distribución *Erlang* (Gamma).

Notación de Kendall

Por convención los modelos que se trabajan en teoría de colas se etiquetan



Las distribuciones que se utilizan son:

- M: Distribución exponencial (markoviana)
- D : Distribución degenerada (tiempos constantes)
- E k : Distribución Erlang
- G : Distribución general

M / M / s : Modelo donde tanto los tiempos entre llegada como los tiempo de servicio son exponenciales y se tienen s servidores.

M / G / 1: Tiempos entre llegada exponenciales, tiempos de servicio general y 1 sólo servidor

Terminología

Usualmente siempre es común utilizar la siguiente terminología estándar:

• **Estado del sistema :** Número de clientes en el sistema.

• **Longitud de la cola:** Número de clientes que esperan servicio.

• **N(t) :** Número de clientes en el sistema de colas en el tiempo t ($t \geq 0$).

• **P_n (t):** Probabilidad de que exactamente n clientes estén en el sistema en el tiempo t, dado el número en el tiempo cero.

• **s :** Número de servidores en el sistema de colas.

• **λ_n :** Tasa media de llegadas (número esperado de llegadas por unidad de tiempo) de nuevos clientes cuando hay n clientes en el sistema.

• **μ_n :** Tasa media de servicio para todo el sistema (número esperado clientes que completan su servicio por unidad de tiempo) cuando hay n clientes en el sistema.

Nota: μ_n representa la tasa combinada a la que todos los servidores ocupados logran terminar sus servicios

λ_n: Cuando λ_n es constante para toda n

μ_n : Cuando μ_n es constante para toda $n \geq 1$

$\frac{1}{\lambda}$ Tiempo entre llegadas esperado

$$\frac{\mathbf{1}}{\mu}$$ Tiempo entre llegadas esperado

Ejemplo:

Sea λ = 3 personas / hora

$$\frac{\mathbf{1}}{\lambda} \quad \frac{1 \text{ hora}}{3} = 20 \text{ minutos}$$

ρ: factor de utilización para la instalación se servicio (fracción esperada de tiempo fue los servidores individuales están ocupados).

$$\rho = \frac{\lambda}{s\mu}$$

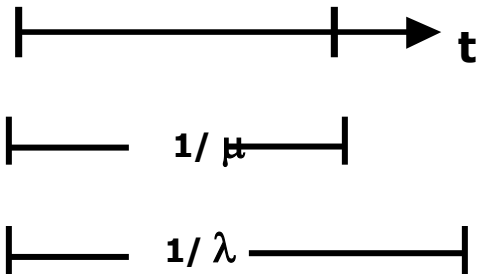
También puede interpretarse como número promedio de personas siendo atendidas

Nota: Para los sistemas de colas que analizaremos haremos la suposición de que el sistema se encuentra en la condición de estado estable.

Demostración

Para s = 1

ρ: fracción esperada de tiempo que los servidores individuales están ocupados).



λ = 12/ hora $\rightarrow$ $1/\lambda = 5$ minutos

μ = 15/ hora $\rightarrow$ $1/\mu = 4$ minutos

El servidor está trabajando 4 de cada 5 minutos, es decir está trabajando el 80% del tiempo

ρ**:** Número promedio de personas siendo atendidas

Número promedio = $0 * P_0 + 1 * P_1$

Número promedio = P_1

Número promedio = $1/\mu \ / \ 1/\lambda$

Número promedio = ρ

La siguiente notación supone la condición de estado estable:

- **P_n :** Probabilidad de que haya exactamente n clientes en el sistema
- **L:** Número esperado de clientes en el sistema.
- **L_q :** Longitud esperada de la cola (excluye los clientes que están en servicio).

• ***W* :** Tiempo de espera en el sistema para cada cliente

• **W :** *E(W*)

• ***W* q:** Tiempo de espera en la cola para cada cliente.

• **W_q:** E *(W_q*)

Relaciones entre L , W , L_q y W_q

Supongamos que λ_n es una constante λ para toda n:

$$L = \lambda W \qquad L_q = \lambda W_q$$

Supongamos que el tiempo medio de servicio es una constante $1/\mu$ para toda $n \geq 1$

$$W = W_q + 1/\mu \qquad L = L_q + \rho$$

Estas relaciones son fundamentales pues permiten determinar las cuatro cantidades fundamentales L, W, L_q, W_q, en cuanto se encuentra analíticamente el valor de una de ellas.

Características claves.

Existen dos clases básicas de tiempo entre llegadas:

Determinístico, en el cual clientes sucesivos llegan en un mismo intervalo de tiempo, fijo y conocido. Un ejemplo clásico es el de una línea de ensamble, en donde los artículos llegan a una estación en intervalos invariables de tiempo (conocido como ciclos de tiempo)

Probabilístico, en el cual el tiempo entre llegadas sucesivas es incierto y variable. Los tiempos entre llegadas probabilísticos se describen mediante una distribución de probabilidad.

En el caso probabilístico, la determinación de la distribución real, a menudo, resulta difícil. Sin embargo, una distribución , la distribución exponencial, ha probado ser confiable en muchos de los problemas prácticos. La función de densidad, para una distribución exponencial depende de un parámetro, digamos λ (letra griega lambda), y está dada por:

$$f(t)=(1/\lambda)e-\lambda^{t}$$

en donde λ (lambda) es el número promedio de llegadas en una unidad de tiempo.

Con una cantidad, T, de tiempo se puede hacer uso de la función de densidad para calcular la probabilidad de que el siguiente cliente llegue dentro de las siguientes T unidades a partir de la llegada anterior, de la manera siguiente:

$$P(\text{tiempo entre llegadas} <= T)=1-e-\lambda^{t}$$

El proceso de servicio.

El proceso de servicio define cómo son atendidos los clientes. En algunos casos, puede existir más de una estación en el sistema en el cual se proporcione el servicio requerido. Los bancos y los supermercados, de nuevo, son buenos ejemplos de lo anterior. Cada ventanilla y cada registradora son estaciones que proporcionan el mismo servicio. A tales estructuras se les conoce como sistemas de colas de canal múltiple. En dichos sistemas, los servidores pueden ser idénticos, en el sentido en que proporcionan la misma clase de servicio con igual rapidez, o pueden no ser idénticos. Por ejemplo, si todos los cajeros de un banco tienen la misma experiencia, pueden considerarse como idénticos.

Al contrario de un sistema de canal múltiple, considere un proceso de producción con una estación de trabajo que proporciona el servicio requerido. Todos los productos deben pasar por esa estación de trabajo; en este caso se trata de un sistema de colas de canal sencillo. Es importante hacer notar que incluso en un sistema de canal sencillo pueden existir muchos servidores que, juntos, llevan a cabo la tarea necesaria. Por ejemplo, un negocio de lavado a mano de automóviles, que es una sola estación, puede tener dos empleados que trabajan en un auto de manera simultánea

Otra característica del proceso de servicio es el número de clientes atendidos al mismo tiempo en una estación. En los bancos y en los supermercados (sistema de canal sencillo), solamente un cliente es atendido a la vez. Por el contrario, los pasajeros que esperan en una parada de autobús son atendidos en grupo, según la capacidad del autobús que llegue.

Otra característica más de un proceso de servicio es si se permite o no la prioridad, esto es ¿puede un servidor detener el proceso con el cliente que está atendiendo para dar lugar a un cliente que acaba de llegar?. Por ejemplo, en una sala de urgencia, la prioridad se presenta cuando un médico, que está atendiendo un caso que no es crítico es llamado a atender un caso más crítico. Cualquiera que sea el proceso de servicio, es necesario tener una idea de cuánto tiempo se requiere para llevar a cabo el servicio. Esta cantidad es importante debido a que cuanto más dure el servicio, más tendrán que esperar los clientes que llegan. Como en el caso del proceso de llegada, este tiempo pude ser determinístico o probabilístico . Con un tiempo de servicio determinístico, cada cliente requiere precisamente de la misma cantidad conocida de tiempo para ser atendido. Con un tiempo de servicio probabilístico, cada cliente requiere una cantidad distinta e incierta de tiempo de servicio. Los tiempos de servicio probabilísticos se describen matemáticamente mediante una distribución de probabilidad. En la práctica resulta difícil determinar cuál es la distribución real, sin embargo, una distribución que ha resultado confiable en muchas aplicaciones , es la distribución exponencial .En este caso, su función de densidad depende de un parámetro, digamos (la letra griega my) y esta dada por

$$s(t)=(1/\mu)e^{-\mu t}$$

en la que:

μ = número promedio de clientes atendidos por unidad de tiempo,

de modo que:

$1/\mu$ = tiempo promedio invertido en atender a un cliente

En general, el tiempo de servicio puede seguir cualquier distribución, pero, antes de que pueda analizar el sistema, se necesita identificar dicha distribución.

Medidas de rendimiento para evaluar un sistema de colas

El objetivo último de la teoría de colas consiste en responder cuestiones administrativas pertenecientes al diseño y a la operación de un sistema de colas. El gerente de un banco puede querer decidir si programa tres o cuatro cajeros durante la hora de almuerzo. En una estructura de producción, el administrador puede desear evaluar el impacto de la compra de una nueva máquina que pueda procesar los productos con más rapidez.

Cualquier sistema de colas pasa por dos fases básicas. Por ejemplo, cuando el banco abre en la mañana, no hay nadie en el sistema, de modo que el primer cliente es atendido de forma inmediata. Conforme van llegando más clientes, lentamente se va formando la cola y la cantidad de tiempo que tienen que esperar se empieza a aumentar. A medida que avanza el día, el sistema llega a una condición en la que el efecto de la falta inicial de clientes ha sido eliminado y el tiempo de espera de cada cliente ha alcanzado niveles bastante estables.

Algunas medidas de rendimiento comunes

Existen muchas medidas de rendimiento diferentes que se utilizan para evaluar un sistema de colas en estado estable. Para diseñar y poner en operación un sistema de colas, por lo general, los administradores se preocupan por el nivel de servicio que recibe un cliente, así como el uso apropiado de las instalaciones de servicio de la empresa. Algunas de

las medidas que se utilizan para evaluar el rendimiento surgen de hacerse las siguientes preguntas:

Preguntas relacionadas con el tiempo, centradas en el cliente, como:

a. ¿Cuál es el tiempo promedio que un cliente recién llegado tiene que esperar en la fila antes de ser atendido?. La medida de rendimiento asociada es el tiempo promedio de espera, representado con Wq
b. ¿Cuál es el tiempo que un cliente invierte en el sistema entero, incluyendo el tiempo de espera y el de servicio?. La medida de rendimiento asociada es el tiempo promedio en el sistema, denotado con W

Preguntas cuantitativas relacionadas al número de cliente, como:

a. En promedio ¿cuántos clientes están esperando en la cola para ser atendidos?. La medida de rendimiento asociada es la longitud media de la cola, representada con Lq
b. ¿Cuál es el número promedio de clientes en el sistema?. La medida de rendimiento asociada es el número medio en el sistema, representado con L

Preguntas probabilísticas que implican tanto a los clientes como a los servidores, por ejemplo:

a. ¿Cuál es la probabilidad de que un cliente tenga que esperar a ser atendido?. La medida de rendimiento asociada es la probabilidad de bloqueo, que se representa por, pw
b. En cualquier tiempo particular, ¿cuál es la probabilidad de que un servidor esté ocupado?. La medida de rendimiento asociada es la utilización, denotada con U. Esta medida indica también la fracción de tiempo que un servidor esta ocupado.
c. ¿Cuál es la probabilidad de que existan n clientes en el sistema?. La medida de rendimiento asociada se obtiene calculando la probabilidad Po de que no haya clientes en el sistema , la probabilidad Pi de que haya un cliente en el sistema, y así sucesivamente. Esto tiene como resultado la distribución de probabilidad de estado, representada por Pn, n=0,1.....
d. Si el espacio de espera es finito, ¿Cuál es la probabilidad de que la cola esté llena y que un cliente que llega no sea atendido?. La medida de rendimiento asociada es la probabilidad de negación del servicio, representada por Pd

Preguntas relacionadas con los costos, como:

a. ¿Cuál es el costo por unidad de tiempo por operar el sistema?
b. ¿Cuántas estaciones de trabajo se necesitan para lograr mayor efectividad en los costos?

El cálculo específico de estas medidas de rendimiento depende de la clase de sistema de colas. Algunas de estas medidas están relacionadas entre sí. Conocer el valor de una medida le permita encontrar el valor de una medida relacionada.

Relaciones entre medidas de rendimiento

El cálculo de muchas de las medidas de rendimiento depende de los procesos de llegadas y de servicio del sistema de colas en específico. Estos procesos son descritos matemáticamente mediante distribuciones de llegada y de servicio. Incluso sin conocer la distribución especifica, las relaciones entre algunas de las medidas de rendimiento pueden obtenerse para ciertos sistemas de colas, únicamente mediante el uso de los siguientes parámetros de los procesos de llegada y de servicio.

λ = número promedio de llegadas por unidad de tiempo

μ = número promedio de clientes atendidos por unidad de tiempo en una sección

Supongamos que una población de clientes infinita y una cantidad limitada de espacio de espera en la fila. El tiempo total que un cliente invierte en el sistema es la cantidad de tiempo invertido en la fila más el tiempo durante el cual es atendido:

Tiempo promedio en el sistema = Tiempo de espera + Tiempo de servicio

El tiempo promedio en el sistema y el tiempo promedio de espera están representados por las cantidades W y Wq, respectivamente. El tiempo promedio de servicio puede expresarse en términos de parámetros de &. Por ejemplo, si & es 4 clientes por hora, entonces , en promedio, cada cliente requiere 1 /4 para ser atendido. En general, el tiempo de servicio es 1/&, lo cual nos conduce a la siguiente relación :

$$W = Wq + 1/\mu$$

Consideremos ahora la relación entre el número promedio de clientes en el sistema y el tiempo promedio que cada cliente pasa en el sistema. Imaginemos que un cliente acaba de llegar y se espera que permanezca en el sistema un promedio de media de hora. Durante esta media hora, otros clientes siguen llegando a una tasa ¿¿digamos doce por hora??. Cuando el cliente en cuestión abandona el sistema, después de media hora, deja tras de sí un promedio de (1/2)*12 = 6 clientes nuevos.

Es decir, en promedio, existen seis clientes en el sistema en cualquier tiempo dado. Entonces:

Tiempo promedio de clientes = Número de llegadas X *Tiempo promedio en el sistema.

de modo que:

$$L = \lambda * W$$

Utilizando una lógica parecida se obtiene la relación entre el número promedio de clientes que esperan en la cola y el tiempo promedio de espera en la fila:

Tiempo promedio de clientes = Número de llegadas X Unidad de tiempo en la cola

de manera que:

$$Lq = \lambda * Wq$$

CONCLUSIÓN

La teoría de las colas es el estudio matemático de las colas o líneas de espera. La formación de colas es, por supuesto, un fenómeno común que ocurre siempre que la demanda efectiva de un servicio excede a la oferta efectiva.

Con frecuencia, las empresas deben tomar decisiones respecto al caudal de servicios que debe estar preparada para ofrecer. Sin embargo, muchas veces es imposible predecir con exactitud cuándo llegarán los clientes que demandan el servicio y/o cuanto tiempo será necesario para dar ese servicio; es por eso que esas decisiones implican dilemas que hay que resolver con información escasa. Estar preparados para ofrecer todo servicio que se nos solicite en cualquier momento puede implicar mantener recursos ociosos y costos excesivos. Pero, por otro lado, carecer de la capacidad de servicio suficiente causa colas excesivamente largas en ciertos momentos. Cuando los clientes tienen que esperar en una cola para recibir nuestros servicios, están pagando un coste, en tiempo, más alto del que esperaban. Las líneas de espera largas también son costosas por tanto para la empresa ya que producen pérdida de prestigio y pérdida de clientes.

La teoría de las colas en si no resuelve directamente el problema, pero contribuye con la información vital que se requiere para tomar las decisiones concernientes prediciendo algunas características sobre la línea de espera: probabilidad de que se formen, el tiempo de espera promedio.

Pero si utilizamos el concepto de "clientes internos" en la organización de la empresa, asociándolo a la teoría de las colas, nos estaremos aproximando al modelo de organización empresarial "just in time" en el que se trata de minimizar el costo asociado a la ociosidad de recursos en la cadena productiva.

BIBLIOGRAFÍA

Arbonas, M.E. Optimización Industrial (I): Distribución de los recursos. Colección Productica No. 26. Marcombo S.A, 1989.

Arbonas, M.E. Optimización Industrial (II): Programación de recursos. Colección Productica No. 29. Marcombo S.A, 1989.

Moskowitz,H. y Wright G.P. Investigación de Operaciones. Prentice_Hall Hispanoamericana S.A. 1991.

Buffa,E: Operations Management: Problems and Models. Edición Revolucionaria,La Habana, 1968.

http://www.eumed.net/

www.gestiopolis.com

www.monografias.com

http://es.wikipedia.org/