I N T R O D U C C I O N

Este informe tiene como finalidad presentar una teoría operacional sobre la *Teoría de Colas,* la cual incluye el estudio matemático de las colas o líneas de espera, siendo la de mayor aplicación potencial y sin embargo es la más difícil de aplicar. Los fenómenos de espera para recibir servicio son cosas de la vida diaria; por ejemplo, esperar en una cola para pagar el teléfono o en el supermercado. No obstante, la espera no solo se limita a personas sino a procedimientos o ensamblados de máquinas, por lo tanto en esta unidad se describen modelos matemáticos aplicables a cualquier situación donde se forme una cola.

No pretendo incluir en un solo tema todo lo que necesita saber el estudiante, sino ofrecer un marco de los conocimientos básicos presentados en forma clara y precisa.

La formación de líneas de espera es, por supuesto, un fenómeno común que ocurre siempre que la demanda actual de un servicio excede a la capacidad actual de proporcionarlo. Con frecuencia, en la industria y en otros sitios, deben tomarse decisiones respecto a la cantidad de capacidad que debe proporcionarse. Sin embargo, muchas veces es imposible predecir con exactitud cuándo llegarán las unidades que buscan el servicio y/o cuánto tiempo será necesario para dar ese servicio; es por esto que esas decisiones suelen ser difíciles. Proporcionar demasiado servicio implica costos excesivos. Por otro lado, carecer de la capacidad de servicio suficiente causa colas excesivamente largas en ciertos momentos. Las líneas de espera largas también son costosas en cierto sentido, ya sea por un costo social, por un costo causado por la pérdida de clientes, por el costo de empleados ociosos o por algún otro costo importante. Entonces, la meta final es lograr un balance económico entre el costo de servicio y el costo asociado con la espera por ese servicio. La teoría de colas en sí no resuelve directamente este problema, pero contribuye con información vital que se requiere para tomar las decisiones concernientes prediciendo algunas características sobre la línea de espera como el tiempo de espera promedio.

La teoría de colas proporciona un gran número de modelos matemáticos para describir una situación de línea de espera. Con frecuencia se dispone de resultados matemáticos que predicen algunas de las características de estos modelos.

Como ejemplo prototipo expondré la sala de emergencia del Hospital General, el cual proporciona cuidados médicos rápidos a los casos de emergencia que llegan en ambulancia o vehículos particulares. En cualquier momento se cuenta con un doctor de guardia. No obstante, a causa de la mala situación económica que vive nuestro país existe una creciente tendencia a usar estas instalaciones para casos de emergencia en lugar de ir a una clínica privada, es por ello que el hospital ha venido experimentando un aumento continuo en el número de pacientes anuales que llegan a la sala de emergencia. Como resultado, es bastante común que los pacientes que llegan durante las horas pico (temprano en la tarde) tengan que esperar turno para recibir el tratamiento del doctor. Por esto, se ha hecho una propuesta para asignar un segundo doctor a la sala de emergencia durante esas horas pico, para que se puedan atender dos casos de emergencia al mismo tiempo. Se ha pedido al ingeniero administrador del hospital que estudie esta posibilidad.

El ingeniero comenzó por reunir los datos históricos pertinentes y por hacer una proyección de estos datos al siguiente año. Reconoció que la sala de emergencia es un sistema de líneas de espera y aplicó algunos modelos de teoría de colas para predecir las características de la espera en el sistema con uno y dos doctores como veremos a continuación:

ESTRUCTURA BÁSICA DE LOS MODELOS DE COLAS

- **Proceso básico de colas**

El proceso básico supuesto por la mayor parte de los modelos de colas es el siguiente. Los *clientes* que requieren un servicio se generan a través del tiempo en una *fase de entrada*. Estos *clientes* entran al *sistema* y se unen a una *cola*. En determinado momento se selecciona un miembro de la cola, para proporcionarle el servicio, mediante alguna regla conocida como *disciplina de servicio*. Luego, se lleva a cabo el servicio requerido por el cliente en un *mecanismo de servicio*, después de lo cual el cliente sale del sistema de colas. En la siguiente figura se da un esquema de este proceso.

- **Fuente de entrada (población potencial)**

Una característica de la fuente de entrada es su tamaño. El *tamaño* es el número total de clientes que pueden requerir servicio en determinado momento, es decir, el número total de clientes potenciales distintos. Esta población a partir de la cual surgen las unidades que llegan se conoce como **población de entrada**. Puede suponerse que el tamaño es *infinito* o *finito* (de modo que también se dice que la fuente de entrada es *ilimitada* o *limitada*). Como los cálculos son mucho más sencillos para el caso infinito, esta suposición se hace muy seguido aun cuando el tamaño real sea un número fijo relativamente grande, y deberá tomarse como una suposición implícita en cualquier modelo que no establezca otra cosa. El caso finito es más difícil analíticamente, pues el número de clientes en la cola afecta el número potencial de clientes fuera del sistema en cualquier tiempo; pero debe hacerse esta suposición finita se la tasa a la que la fuente de entrada genera clientes nuevos queda afectada en forma significativa por el número de clientes en el sistema de líneas de espera.

También se debe especificar el patrón estadístico mediante el cual se generan los clientes a través del tiempo. La suposición normal es que se generan de acuerdo a un *proceso Poisson*, es decir, el número de clientes que llegan hasta un tiempo específico tiene una distribución Poisson. En nuestro caso corresponde a aquel cuyas llegadas al sistema ocurren de manera aleatoria pero con cierta tasa media fija y sin importar cuántos clientes están ya ahí (por lo que el *tamaño* de la fuente de entrada es *infinito*). Una suposición equivalente es que la distribución de probabilidad del tiempo que transcurre entre dos llegadas consecutivas es *exponencial*. Se hace referencia al tiempo que transcurre entre dos llegadas consecutivas como **tiempo entre llegadas**.

- **Cola**

Una cola se caracteriza por el número máximo permisible de clientes que puede admitir. Las colas pueden ser *finitas* o *infinitas*, según si este numero es finito o infinito. La suposición de una *cola infinita* es la estándar para la mayor parte de los modelos, incluso en situaciones en las que de hecho existe una cota superior (relativamente grande) sobre el número permitido de clientes, ya que manejar una cota así puede ser un factor complicado para el análisis. Los sistemas de colas en los que la cota superior es tan pequeña que se llega a ella con cierta frecuencia, necesitan suponer una cola *finita*.

- **Disciplina de la cola**

La disciplina de la cola se refiere al orden en el que se seleccionan sus miembros para recibir el servicio. Por ejemplo, puede ser: primero en entrar, primero en salir, aleatoria, de acuerdo a algún procedimiento de prioridad o a algún otro orden. La que suponen como normal los modelos de colas es la primero en entrar, primero en salir, a menos que se establezca otra cosa.

- **Mecanismo de servicio**

El mecanismo de servicio consiste en una o más *instalaciones de servicio*, cada una de ellas con uno o más *canales paralelos de servicio*, llamados **servidores**. Si existe más de una instalación de servicio, puede ser que sirva al cliente a través de una secuencia de ellas (*canales de servicio en serie*). En una instalación dada, el cliente entra en uno de estos canales y el servidor le presta el servicio completo. Un modelo de colas debe especificar el arreglo de las instalaciones y el número de servidores (canales paralelos) en cada una. Los

modelos más elementales suponen una instalación, ya sea con un servidor o con un número finito de servidores.

El tiempo que transcurre desde el inicio del servicio para un cliente hasta su terminación en una instalación se llama **tiempo de servicio** (o *duración del servicio*). Un modelo de un sistema de colas determinado debe especificar la distribución de probabilidad de los tiempos de servicio para cada servidor (y tal vez para los distintos tipos de clientes), aunque es común suponer la *misma* distribución para todos los servidores.

- **Un proceso de colas elemental**

Como ya se ha sugerido, la teoría de colas se aplica a muchos tipos diferentes de situaciones. El tipo que más prevalece es el siguiente: una sola línea de espera (que puede estar vacía en ciertos lapsos de tiempos) se forma frente a una instalación de servicio, dentro de la cual se encuentran uno o más servidores. Cada cliente generado por una fuente de entrada recibe servicio de uno de los servidores, quizá después de esperar un poco en la cola (línea de espera). En la figura se da un esquema del sistema de colas elemental del que se habla (cada cliente se indica por una *C* y cada servidor por una *S*).

Observe que el proceso que ilustramos en el ejemplo al inicio es de este tipo. La fuente de entrada genera clientes en la forma de casos de emergencia que requieren cuidado médico. La sala de emergencia es la instalación de servicio y los doctores son los servidores.

Un servidor no tiene que ser un solo individuo; puede ser un grupo de personas, por ejemplo, una cuadrilla de reparación que combina fuerzas para realizar, de manera simultánea, el servicio que solicita el cliente. Aún más, los servidores ni siquiera tienen que ser personas. En muchos casos puede ser una máquina o una pieza de equipo, como un cargador frontal que presta el servicio cuando se requiere (tal vez con la ayuda de un operador). Con esta misma línea de ideas, los clientes en la cola no tienen que ser personas. Por ejemplo, pueden ser unidades que esperan ser procesadas en una cierta máquina, o pueden ser carros que esperan pasar por una caseta de cobro.

No es necesario que de hecho se forme físicamente una línea de espera delante de una estructura física que constituye la instalación de servicio; es decir, los miembros de la cola pueden estar dispersos en un área mientras esperan que el servidor venga a ellos, como las máquinas que esperan reparación. El servidor o grupo de servidores asignados a un área constituyen la instalación de servicio para esa área. De todas maneras, la teoría de colas da un número promedio de clientes en espera, el tiempo promedio de espera, etc. pues es irrelevante si los clientes esperan agrupados o no. El único requisito esencial para poder aplicar la teoría de colas es que los cambios en el número de clientes que esperan un servicio ocurran como si prevaleciera la situación física que se describe en la figura anterior (o una contraparte valida).

Muchos de los modelos para la teoría de colas hacen la suposición de que todos los *tiempos entre llegadas* y todos los *tiempos de servicio* son independientes e idénticamente distribuidos. Por ejemplo, el modelo *M/M/s* supone que tanto los tiempos entre llegadas como los de servicio tienen una distribución exponencial y que el número de servidores es *s* (cualquier entero positivo). El modelo *M/G/1* supone que los tiempos entre llegadas siguen una distribución exponencial pero no pone restricciones sobre la distribución de los tiempos de servicio, mientras que el número de servidores está restringido a exactamente *1*.

EJEMPLOS DE SISTEMAS DE COLAS REALES

Puede parecer que la descripción de los sistemas de colas pueden parecer más o menos abstracta y sólo es aplicables en situaciones prácticas bastante especiales. Por el contrario, los sistemas de colas ocurren con sorprendente frecuencia en una amplia variedad de contextos. Para ampliar el horizonte sobre la aplicabilidad de la teoría de colas, se mencionarán brevemente varios ejemplos reales de sistemas de colas.

Una clase importante de sistemas de colas que se encuentran en la vida es el **sistema de servicio comercial**, en donde los clientes externos reciben un servicio de una organización comercial. Muchos de estos sistemas incluyen un servicio de persona a persona en una localidad fija, como una peluquería (los peluqueros son los servidores), es servicio de una cajera de banco, las cajas de cobro en un supermercado y una cola en una cafetería (canales de servicio en serie). Muchos otros sistemas son de tipo diferente, como la reparación de aparatos domésticos (el servidor va hacia el cliente), una maquina de monedas (el servidor es una máquina) y una gasolinera (los clientes son automóviles).

Otra clase importante es la de **sistemas de servicio de transporte**. Para algunos de estos sistemas los vehículos son los clientes, como los automóviles que esperan pasar por una caseta de cobro o un semáforo (el servidor), un camión de carga o un barco que esperan que una cuadrilla les dé el servicio de carga o descarga y un avión que espera aterrizar o despegar en una pista (el servidor). (Un estacionamiento es un ejemplo poco usual de este tipo, en el que los carros son los clientes y los espacios son los servidores, pero no existe una cola porque si el estacionamiento está lleno, los clientes se van a otro lado a estacionarse). En otros casos, los vehículos son los servidores, como los taxis, los camiones de bomberos y los elevadores.

En los últimos años, tal vez la teoría de colas se ha aplicado más a los **sistemas de servicio interno en la industria y en los negocios**, en donde los clientes que reciben el servicio son *internos* o parte de la organización. Los ejemplos incluyen sistemas de manejo de materiales, en donde las unidades de manejo de materiales (los servidores) mueven cargas (los clientes); sistemas de mantenimiento, en donde las brigadas de mantenimiento (los servidores) reparan máquinas (los clientes) y puestos de inspección en los que los inspectores de control de calidad (los servidores) inspeccionan artículos (los clientes). Las instalaciones para empleados y los departamentos que dan servicio a empleados también entran en esta categoría. Además, las máquinas se pueden ver como servidores cuyos clientes son los trabajos que se están procesando. Un ejemplo relacionado muy importante es un centro de cómputo en el que la computadora se puede ver como el servidor.

Es del reconocimiento general que la teoría de colas también se puede aplicar a **sistemas de servicio social**. Por ejemplo, un sistema judicial es una red de colas, en donde las cortes son las instalaciones de servicio, los jueces (o los jurados) son los servidores y los casos que esperan el proceso son los clientes. Un sistema legislativo es una red de colas parecida, en el que los clientes son los asuntos que el congreso va a tratar. Algunos sistemas de salud pública son sistemas de colas. Al inicio se vio un ejemplo (la sala de emergencia de un hospital), pero también las ambulancias, las máquinas de rayos X y las camas del hospital pueden jugar el papel de servidores en sus propios sistemas de colas. En forma parecida, las familias en espera de viviendas de interés social u otros servicios sociales se pueden concebir como clientes de un sistema de colas.

Aun cuando éstas son cuatro clases amplias de sistemas de colas, la lista todavía no se agota. De hecho, la teoría de colas comenzó a principios de siglo con aplicaciones a ingeniería telefónica (el fundador de la teoría de colas, A.K. Erlang, era un empleado de la Danish Telephone Company en Copenhague), y la ingeniería telefónica constituye todavía una importante aplicación. Lo que es más, cada individuo tiene sus propias líneas de espera personales: tareas, libros que leer, etc. Estos ejemplos son suficientes para sugerir que los sistemas de colas sin duda ocurren con toda frecuencia en muchas áreas de la sociedad.

PROCESO DE NACIMIENTO Y MUERTE

La mayor parte de los modelos elementales de colas suponen que las entradas (llegada de clientes) y las salidas (clientes que se van) del sistema ocurren de acuerdo al *proceso de nacimiento y muerte*. Este importante proceso de teoría de probabilidad tiene aplicaciones en varias áreas. Sin embrago en el contexto de la teoría de colas, el término **nacimiento** se refiere a *llegada* de un nuevo cliente al sistema de colas y el término **muerte** se refiere a la *salida* del cliente servido. El *estado* del sistema en el tiempo t (t 0), denotado por N (t), es el número de clientes que hay en el sistema de colas en el tiempo t. El proceso de nacimiento y muerte describe en *términos probabilísticos* cómo cambia N (t) al aumentar t. En general, dice que los nacimientos y muertes *individuales* ocurren *aleatoriamente*, en donde sus tasas medias de ocurrencia

dependen del estado actual del sistema. De manera más precisa, las suposiciones del proceso de nacimiento y muerte son las siguientes:

SUPOSICIÓN 1. Dado $N(t) = n$, la distribución de probabilidad actual del tiempo que falta para el próximo *nacimiento* (llegada) es *exponencial* con parámetro (n=0,1,2,.).

SUPOSICIÓN 2. Dado $N(t) = n$, la distribución de probabilidad actual del tiempo que falta para la próxima *muerte* (terminación de servicio) es *exponencial* con parámetro (n=1,2,.).

SUPOSICIÓN 3. La variable aleatoria de la suposición 1 (el tiempo que falta hasta el próximo nacimiento) y la variable aleatoria de la suposición 2 (el tiempo que falta hasta la siguiente muerte) son mutuamente independientes.

Como consecuencia de las suposiciones 1 y 2, el proceso de nacimiento y muerte es un tipo especial de *cadena de Markov de tiempo continuo*. Los modelos de colas que se pueden representar por una cadena de Markov de tiempo continuo son mucho más manejables analíticamente que cualquier otro.

Excepto por algunos casos especiales, el análisis del proceso de nacimiento y muerte es complicado cuando el sistema se encuentra en condición *transitoria*. Se han obtenido algunos resultados sobre esta distribución de probabilidad de $N(t)$ pero son muy complicados para tener un buen uso práctico. Por otro lado, es bastante directo derivar esta distribución *después* de que el sistema ha alcanzado la condición de *estado estable* (en caso de que pueda alcanzarla).

C O N C L U S I O N

Los sistemas de colas son muy comunes en la sociedad. La adecuación de estos sistemas pueden tener un efecto importante sobre la calidad de vida y la productividad.

Para estudiar estos sistemas, la teoría de colas formula modelos matemáticos que representan su operación y después usa estos modelos para obtener medidas de desempeño. Este análisis proporciona información vital para diseñar de manera efectiva sistemas de colas que logren un balance apropiado entre el costo de proporcionar el servicio y el costo asociado con la espera por ese servicio.

En este informe se realizo un resumen de algunos modelos básicos de teoría de colas para los que se tienen resultados particularmente útiles. Se hubiera podido considerar muchos otros modelos interesantes si el espacio lo hubiera permitido. De hecho, han aparecido en la literatura técnica *varios miles* de artículos de investigación que formulan y/o analizan modelos de colas, y ¡cada año se publican mucho más!

La *distribución exponencial* juega un papel fundamental en la teoría de las colas para representar la distribución de los tiempos entre llegadas y de servicio, ya que esta suposición permite representar un sistema de colas como una *cadena de Markov de tiempo continuo*. Por la misma razón, son de gran utilidad las *distribuciones tipo fase* como la *distribución Erlang*, en donde se desglosa el tiempo total en fases individuales que tienen distribuciones exponenciales. Haciendo algunas suposiciones adicionales, se han obtenido importantes resultados analíticos sólo para un pequeño número de modelos de colas.

Los modelos de disciplina de prioridades son útiles para la situación común en la que se da prioridad a algunas categorías de clientes sobre otras para recibir el servicio.

En otra situación común los clientes deben recibir servicio en distintas estaciones o instalaciones. Los modelos de redes de colas se usan cada vez más en estas situaciones. Esta es una área especialmente activa en la investigación actual.

Cuando no se dispone de un modelo manejable que proporcione una representación razonable del sistema bajo estudio, un enfoque usual es obtener los datos de desempeño pertinentes mediante el desarrollo de un programa de computadora para simular la operación del sistema.

La teoría de colas ha demostrado ser una herramienta muy útil y se pronostica que su uso seguirá ampliándose conforme crezca el reconocimiento de los beneficios de los sistemas de colas.

REPUBLICA BOLIVARIANA DE VENEZUELA

UNIVERSIDAD SANTA MARIA

FACULTAD DE INGENIERÍA

CÁTEDRA: INVESTIGACIÓN DE OPERACIONES

TEORÍAS DE COLAS

Caracas, Febrero de 2.000.

Clientes

Clientes

servidos

Mecanismo

de servicio

Cola

Fuente de

entrada

Sistema de colas

Clientes servidos

Clientes servidos

C

C

C

C

S

S

S

S

C C C C C C C

Clientes

Mecanismo

de servicio

Sistema de colas