881201

UNIVERSIDAD ANAHUAC

ESCUELA DE ACTUARIA

CON ESTUDIOS INCORPORADOS A LA

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

UNIVERSIDAD ANAHUAC

VINCE IN BONO MALUM

TEORIA DE COLAS Y SUS APLICACIONES EN SISTEMAS DE COMPUTO

T E S I S

QUE PARA OBTENER EL TITULO DE

A C T U A R I O

P R E S E N T A

PATRICIA MARCELA SUÑOL FRANCES

MEXICO, D. F. 1987

INDICE

INTRODUCCION

Teoría de Colas (o líneas de espera), se basa en describir los patrones de llegada y/o partida (servicio) por medio de distribuciones de probabilidad apropiadas. Después se derivan características operativas de una situación de colas utilizando la teoría de probabilidades. Ejemplos de estas características son los tiempos promedio de espera hasta que se completa el servicio de un cliente ó el porcentaje de tiempo en que está ocioso el servidor. Tales medidas permiten a los analistas hacer inferencias acerca de la operación del sistema. Los parámetros del sistema (tales como la tasa de servicio) pueden entonces ajustarse para asegurar una utilización más efectiva desde los puntos de vista tanto del cliente como del servidor.

A menudo es deseable basar decisiones concernientes a la eficiencia de una situación de colas en algún análisis de costos. Por ejemplo, un aumento en el número de servidores en el sistema disminuiría eventualmente el tiempo promedio de espera pero aumentaría el costo de servicio. Por otro lado, una disminución en el número de servidores aumentaría el tiempo promedio de espera pero disminuiría el costo de servicio. Por lo tanto, si uno puede expresar el tiempo

promedio de espera en valores monetarios, es posible seleccionar el número óptimo de servidores (o tasa de servicio) que minimiza la suma de los costos de servicio y espera. Aunque esto es teóricamente posible, generalmente surgen dificultades en la práctica al estimar el costo por unidad de tiempo de espera. En la mayoría de los casos, estos costos son tan sutiles que una estimación confiable es prácticamente imposible. En tales casos, uno debe buscar otro criterio como longitud de la cola, tiempo promedio de respuesta, etc. para tomar una decisión.

La Teoría de Colas provee varios modelos matemáticos para describir distintas situaciones y resultados matemáticos que predicen algunas de las características de la cola para estos modelos.

Las colas son comunes en sistemas de cómputo. Así, hay colas de personas esperando para usar una terminal, colas de instrucciones para ser procesadas por el sistema, colas de listados para ser impresos, etc.

En cualquier tipo de sistema de cómputo que puede ser modelado por un sistema de colas existen puntos a favor y en contra que deben ser considerados.

Por otro lado, si todas las personas deben unirse a una cola y las terminales están rara vez libres, puede existir insatisfacción y posiblemente gente que ya no quiera utilizar la computadora. Teoría de Colas, en muchos casos, permite al diseñador de sistemas de computo, asegurarse que se alcance el nivel deseado de servicio en términos de los requerimientos de tiempo de respuesta (tiempo de respuesta es la suma del tiempo de espera y el tiempo de servicio) y al mismo tiempo evitando costos excesivos. El diseñador puede hacer esto considerando varias alternativas de sistemas y evaluándolos mediante modelos analíticos de Teoría de Colas.

El desempeño futuro de un sistema existente puede predecirse de tal forma que pueda mejorarse el sistema periodicamente. Por ejemplo, un modelo analítico de un sistema puede indicar que dentro de dos años la capacidad no será suficiente; el modelo puede permitir evaluar distintas alternativas para incrementar su capacidad, tales como agregar más memoria, obtener un CPU más rápido, proveer más almacenaje auxiliar, cambiar discos suaves por duros, etc.

Esta tesis tiene como objeto el de introducir al lector a la Teoría de Colas y mostrar las aplicaciones prácticas que ésta tiene en el análisis y evaluación de sistemas de cómputo.

El capítulo I introduce al lector a Teoría de Colas. Contiene un ejemplo en el que se muestra una aplicación de la Teoría y se introducen los elementos de Teoría de Colas que se explican con mas detalle posteriormente en el mismo capítulo.

El capítulo II trata los dos modelos con que pueden solucionarse los problemas de líneas de espera, el matemático y el de simulación. Se habla también acerca de las medidas de eficiencia y de la evaluación de desempeño que son muy importantes para lo que se plantea en el último capítulo de la tesis y finalmente se dan algunos temas en los que se han hecho aplicaciones de Teoría de Colas.

En el capítulo III se tratan tres aplicaciones de Teoría de Colas en sistemas de cómputo. Se ve como se pueden aplicar los modelos teóricos para resolver problemas en tres sistemas diferentes y se utilizan las medidas de eficiencia y el análisis de desempeño para lograr ciertas mejoras en ellos.

CAPITULO I

TEORIA DE COLAS

1.1 INTRODUCCION.

La Teoría de Colas es el estudio de los fenómenos de las líneas de espera.

En un sistema abierto de colas existe una población de clientes potenciales siendo que el término cliente significa una unidad que desea alguna clase de servicio -la transmisión de un mensage, procesamiento de algun tipo de información, servicio de I/O- de un local de servicio en el cual hay uno o mas servidores, que son unidades que proveen el servicio requerido por los clientes. Si todos los servidores están ocupados cuando un cliente entra al sistema, el cliente se une a la cola hasta que un servidor esté disponible siempre y cuando haya espacio en la sala de espera.

Una cola ocurre en cualquier sistema cuando en un momento dado, el número de clientes que desean servicio excede la capacidad de las instalaciones de servicio. (Esto puede deberse a que no pueden predecirse las llegadas en un momento dado o a otros factores de los que se hablara mas tarde).

Es claro que la capacidad promedio de la instalación de servicio debe ser suficiente para atender al número promedio de llegadas de clientes; sin embargo, debido a las variaciones en los intervalos de tiempo entre llegadas y a la duración variable de los tiempos de servicio, ocurren las colas. Las colas se forman aunque la proporción del tiempo que los servidores están ocupados con el tiempo en el que están desocupados sea pequeña.

1.2 UN EJEMPLO.

Con el objeto de dar a conocer la terminología empleada en la Teoría de Colas se desarrolló el siguiente ejemplo:

Considérese la operación de un servicio de autos en el cual los automóviles esperan en una línea para ser lavados al pasar a través de un sistema de chorros de agua, cepillos, tubos de aire y aspiradoras. Se les da servicio a los autos en el orden en que llegan (primero en llegar, primero en ser servido). Supóngase que el tiempo de lavado de un auto es constante. Por lo tanto el tiempo de servicio de todos los autos tiene la misma duración. Por medio de muestras de las llegadas, es

posible estimar la distribución del número de llegadas por unidad de tiempo o la distribución del tiempo entre llegadas. De ahí que el número de autos esperando en la línea varía. Después de esperar en la línea, los autos son lavados a una tasa constante m (autos por unidad de tiempo). También dejan el sistema a una tasa constante mientras haya cola. El dueño de la empresa puede estar interesado en el promedio del número de autos en la línea de espera para poder proveer espacio suficiente para los autos y para determinar si el sistema es adecuado para manejarlos. Un cliente puede impacientarse y decidir irse porque hay demasiados autos frente a él; multiplicando el número de autos en la cola frente a él por el tiempo de servicio obtiene un intervalo de tiempo mayor al que desea esperar. El dueño, que ha visto que existen clientes que se van por impaciencia, decide abrir otro canal de servicio de tal modo que los dos canales cooperan para servir a los autos que están esperando. Por lo tanto la longitud promedio de la cola puede reducirse considerablemente utilizando dos canales paralelos independientes. La probabilidad de que un cliente se impaciente y se marche puede, por lo tanto, reducirse.

Si el dueño deseara incluir un servicio de lubricación tal que cada auto después de ser lavado pueda esperar en una cola para obtener servicio adicional, el canal de lavado y el de lubricación operan en serie. Cada uno puede tener una cola frente a él. Pueden asignarse prioridades a clientes, si estos desean pagar más por ser servidos antes.

El dueño puede decidir a qué cliente servir lanzando un dado de tal manera que cada individuo sea igualmente probable de ser elegido (elección aleatoria para servicio). En particular en este negocio, esto puede significar la quiebra.

Así, como se indicó anteriormente, el conocimiento de la distribución de las llegadas y tasa de servicio les permite tanto al dueño como al cliente tomar decisiones muy útiles.

1.3 QUE ES UNA LINEA DE ESPERA.

Una linea de espera o cola es el resultado de las siguientes condiciones:

a) Unidades que requieren servicio (por ejemplo clientes) deben esperar debido a la escasez de instalaciones que puedan atenderlos. La escasez puede deberse a la falta de instalaciones (pocas cajas, exceso

de tiempo de servicio dedicado a unidades, etc.) y puede deberse a una planeación insuficiente de las instalaciones existentes.

b) Las instalaciones de servicio permanecen desocupadas (empleados esperando clientes). Este tiempo de ocio puede ser causa no solo de la falta de clientes en cantidad, sino del espaciamiento de tiempo entre llegadas, y de si los clientes no se marchan antes de ser atendidos.

En cualquiera de estas situaciones, o una combinación de ellas, se forma una línea de espera. En a) las unidades de entrada están en espera de servicio. En b) las unidades de servicio están en espera de clientes.

1.4 QUE ES UN PROBLEMA DE LINEAS DE ESPERA.

Un problema de líneas de espera envuelve el carácter cambiante de las unidades que llegan, o de las instalaciones de servicio, o ambas. Para afectar estos cambios, se requiere manipular o controlar los siguientes factores: tasa de llegada, orden de servicio, número de instalaciones de servicio, etc. El propósito de estas manipulaciones es el de obtener un resultado más eficiente y económico. Por ejemplo, si en

una fábrica hay pocos mecánicos disponibles para reparar máquinas, entonces la cantidad de trabajo de reparación se amontona de tal forma que las operaciones de la fábrica se ven en serios problemas. En esta situación, como en muchas otras, el costo debido a la espera (cola de máquinas que necesitan reparación) varia en proporción inversa al número de mecánicos disponibles, mientras que el costo debido a unidades de servicio desocupadas (mecánicos) varia directamente con el número de mecánicos. La suma de estos dos costos será alto cuando hayan muchos mecánicos, pero también cuando haya muy pocos. El problema es arreglar el proceso, si es posible, para minimizar el costo total.

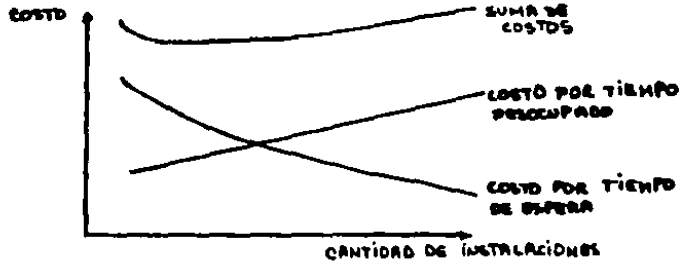


Figura 1

El desarrollo de la Teoría de Colas y la derivación de fórmulas para expresar algunas partes del proceso tales como longitud de la cola y tiempo de

espera promedio, en términos de otras como la tasa promedio de llegada y la tasa promedio de servicio, depende de la naturaleza de las entradas y salidas. Algunos hechos concerniendo el tiempo entre llegadas y el tiempo requerido para servicio pueden ser determinados (o supuestos). Por ejemplo, al estudiar una situación particular, las observaciones sobre un periodo de tiempo pueden indicar que los clientes llegan para servicio de acuerdo a la siguiente tabla:

Tiempo entre llegadas de clientes	% de llegadas	% acumulativo de llegadas
0 - 4.9 min.	63	63
5 - 9.9 min.	23	86
10 - 14.9 min.	9	95
15 - 19.9 min.	4	99
20 min. o mas	1	100

Tabla 1

Esta tabla representa la distribución de tiempo entre llegadas, y sería una de las fuentes de información importantes en el análisis de un proceso de colas en el cual los clientes con estas características son los elementos de entrada. Muestra que, en promedio, un periodo de tiempo de 15 min. o menos precede la

llegada de 95 clientes de cada 100. Información similar del tiempo de servicio a estos clientes sería necesaria, y se considerarían otros puntos como prioridad de servicio, número de puntos de servicio (mostradores), qué hacen los clientes cuando deben esperar, variaciones por temporadas, etc.

Como una breve ilustración de un problema de colas, supongamos la siguiente información acerca de una gasolinera atendida por una sola persona y que está en servicio 24 hrs. (un empleado está en turno por cada periodo de 8 hrs.).

llegadas

1. Autos llegan para servicio a una tasa promedio de uno cada 2.5 min. En un intervalo dado de 5 min., pueden haber cualquier número de llegadas; pero en un periodo más largo de tiempo, se espera un promedio de 2 llegadas en cada intervalo de 5 min.
2. El tiempo de llegada es aleatorio, i.e. el tiempo exacto de una llegada es impredecible y no está influenciado por otra llegada.
3. Cada auto toma su sitio en la cola en orden de llegada y le toca servicio cuando el auto frente a él se va.

servicio

1. Al empleado le toma un promedio de 1 2/3 min. el atender un auto, i.e. en promedio, 3 autos requieren un total de 5 min. Este tiempo incluye el requerido para saludar al nuevo cliente después de terminar con el anterior.

2. El tiempo de servicio para cada auto es aleatorio y completamente independiente de el del auto anterior. Algunos autos requieren muy poco tiempo de servicio, mientras que otros requieren más del tiempo promedio; pero sobre un periodo de tiempo se espera un promedio de 3 autos servidos en un total de 5 min. cuando hayan autos esperando servicio.

En esta situación, se desea investigar si tener solo un empleado en turno es adecuado para cumplir con la demanda de servicio. Si se tienen autos esperando mucho tiempo, existe un costo que resulta de los clientes insatisfechos; si se reduce el tiempo de espera a un mínimo mediante la instalación de más empleados, existe un aumento en el costo de mano de obra requerido para atender a tal rapidez. El problema es saber la cantidad y tipo de servicio para que la suma de estos dos costos sea mínima. Esta es una situación muy común en líneas de espera en la que se supone completa

aleatoriedad tanto en llegadas como en servicio. Cuando estas condiciones son reales, el número promedio de autos en la línea en cualquier momento, y el tiempo promedio que un cliente debe esperar, pueden expresarse en términos de las tasas promedio de llegadas y de servicio. (Ver sección 2.2)

1.5 ELEMENTOS.

Cada problema de colas tiene su propio conjunto de características; sin embargo, todos los problemas de colas tienen las siguientes características fundamentales o especificaciones:

1.5.1 Proceso de llegadas.

Los clientes pueden llegar a la estación de servicio a una tasa promedio constante; sin embargo, los intervalos de tiempo entre llegadas sucesivas son generalmente variables aleatorias independientes de las cuales se supone tienen una distribución que puede aproximarse por observación o experiencia previa con situaciones similares. La tasa promedio de llegadas de clientes se denota por λ, que es igual al recíproco del tiempo promedio entre llegadas sucesivas.

Suponemos que los clientes llegan al sistema en los tiempos t0 < t1 < t2 < ... < tn. Las variables aleatorias Tk = tk - tk-1 (k > 1) son los tiempos entre llegadas. Suponemos que las Tk forman una secuencia de variables aleatorias independientes e identicamente distribuidas.

El patrón de llegadas mas común en Teoría de Colas es el aleatorio o proceso de llegadas Poisson. Esto significa que la distribución del tiempo entre llegadas es Exponencial, esto es, $P(T<t) = 1-e^{-\lambda t}$ para cada tiempo entre llegadas, y la probabilidad de n llegadas en cualquier intervalo de tiempo de longitud t es

$$\frac{e^{-\lambda t}(\lambda t)^n}{n!} \qquad n = 0,1,2,...$$

donde λ es la tasa promedio de llegadas, y el número de llegadas por unidad de tiempo tiene una distribución Poisson.

1.5.1.1 Supuestos de la distribución de llegadas Poisson.

Para tener un patrón de llegadas Poisson, deben satisfacerse las condiciones siguientes:

1. El número total de llegadas durante un intervalo de tiempo dado es independiente al número de llegadas que ocurrieron antes del comienzo del intervalo.

2. Para cada intervalo (t, t+dt), la probabilidad de que exactamente una llegada ocurra es λdt, λ es una constante.

1.5.2 Tiempos de servicio.

De la misma forma se tomará el supuesto de que tiempos de servicio individuales son variables aleatorias independientes que tienen una distribución estadística que puede aproximarse mediante la observación de la realidad. Se denota como μ a la tasa promedio en la cual un cliente es servido, esto es igual al recíproco del promedio del tiempo de servicio.

La distribución mas común del tiempo de servicio en Teoría de Colas es la Exponencial, la cual define el servicio aleatorio. La función de distribución del servicio aleatorio está dada por $1-e^{-\mu t}$ con $t > 0$.

1.5.2.1 Supuestos de la distribución de tiempo de servicio Exponencial.

Si un canal está ocupado en el momento t, la probabilidad de que se desocupe durante el siguiente intervalo de tiempo dt es μdt donde μ es una constante. De allí que la función de frecuencia de los tiempos de

servicio es $\mu e^{-\mu t}$ donde la duración media del servicio es $1/\mu$ ya que el valor esperado de t es:

$$E(t) = \mu \int_0^\infty t e^{-\mu t} = 1/\mu$$

1.5.3 Población.

La población de unidades expuestas a cierta probabilidad de demanda de servicio puede ser finita o infinita. En la práctica, si la población es finita pero grande de tal forma que la tasa de llegadas no es afectada por la disminución de la población causada por aquellas unidades que esperan servicio o que están siendo servidas, entonces se considera infinita. Cuando este supuesto no es válido, debe tomarse en cuenta la disminución de la población.

Un sistema cuya población es finita no puede tener un cola arbitrariamente larga y el número de clientes en el sistema afecta la tasa de llegadas. Para un sistema cuya población es infinita, la cola para servicio es limitada y la tasa de llegadas no se ve afectada por el número de clientes ya presentes en el sistema.

Los cálculos son mucho mas simples para modelos infinitos.

1.5.4 Disciplina de la Cola.

La disciplina de la cola, es la regla para seleccionar al siguiente cliente.

Si no se especifica de otra manera, se supone que el orden en el cual los clientes son servidos es primero en llegar, primero en ser servido. En muchos casos este es el orden en el cual los clientes son servidos. Otras disciplinas, tales como selección aleatoria existen en casos como los de sistemas telefónicos y mantenimiento de máquinas.

Cuando se asignan prioridades a los clientes que llegan en base a tiempo esperado de servicio o costo de la espera, los parámetros de interés del sistema son afectados. Estos parámetros son: número promedio en la cola, probabilidad de que al llegar deba esperar para obtener servicio, etc.

1.5.5 Capacidad máxima del sistema.

En algunos sistemas de colas, la capacidad se supone infinita. Esto es, a cada cliente que llega se le permite esperar hasta que pueda dársele servicio. Otros sistemas llamados sistemas de pérdida, tienen capacidad de espera cero. Esto es, si un cliente llega

cuando las instalaciones de servicio están siendo totalmente utilizadas, no se le permite entrar al sistema. Otros sistemas tienen capacidad positiva pero no infinita.

1.5.6 Número de servidores.

El sistema mas simple de colas es el de un solo servidor. El sistema de servidores múltiples tiene c servidores idénticos.

1.5.7 Intensidad de tráfico.

La intensidad de tráfico u es la proporción entre el tiempo de servicio promedio y el tiempo promedio entre llegadas. Este es un parámetro muy importante de los sistemas de colas y está definido por:

$$u = \lambda / \mu.$$

La intensidad de tráfico determina el número mínimo de servidores que se requieren para poder mantener el flujo de clientes.

1.5.8 Utilización de los servidores.

Otro parámetro importante es la intensidad de tráfico por servidor λ/c que es llamado utilización del servidor cuando el tráfico está igualmente distribuido entre los servidores. La utilización de los servidores es la probabilidad de que un servidor esté ocupado y por lo tanto es la fracción aproximada de tiempo que cada servidor está ocupado. Para sistemas de un solo servidor, debe notarse que la intensidad de tráfico es igual a la utilización del servidor.

1.5.9 Probabilidad de que n clientes estén en el sistema en el tiempo t.

Esta probabilidad, $P_n(t)$, depende no solo de t, sino también de las condiciones iniciales del sistema de colas, esto es, el número de clientes presentes cuando comienza el funcionamiento del sistema y de las distribuciones y parámetros antes mencionados. Para los sistemas mas útiles, cuando t aumenta, $P_n(t)$ se acerca a un valor estable de P_n, el cual es independiente tanto de t como de las condiciones iniciales. Se dice entonces que el sistema está estable.

La Teoría de Colas provee medidas estadísticas del funcionamiento de sistemas de colas. Entre estas medidas estadísticas se encuentran el tiempo promedio de

espera en la cola W, el tiempo promedio de espera en el sistema R (que es la suma del tiempo promedio de espera en la cola mas el tiempo de servicio), el número promedio de clientes en la cola L y el número promedio en el sistema N (que es el numero promedio en la cola mas el numero de clientes que estan siendo atendidos).

Las siguientes fórmulas conocidas como la Ley de Little (Little's Law) son muy útiles para relacionar estas cuatro medidas:

$$L = \lambda W \qquad N = \lambda R$$

1.6 NOTACION.

Una notación abreviada y ya generalizada es la notación de Kendall. Esta notación fue desarrollada para especificar sistemas de colas y tiene la forma A/B/c/K/N/Z. Donde A especifica la distribución del tiempo entre llegadas, B la distribución del tiempo de servicio, c el número de servidores, K la capacidad del sistema, N el número en la población y Z la disciplina de la cola.

Usualmente se utiliza la notación A/B/c cuando no existe límite en la linea de espera, la población es infinita y la disciplina de la cola es FCFS (primero en llegar, primero en ser servido tambien conocido como FIFO).

CAPITULO II

SOLUCION DE PROBLEMAS DE COLAS Y SUS APLICACIONES

2.1 SOLUCION DE PROBLEMAS DE LINEAS DE ESPERA.

La solución de un problema de líneas de espera puede llevarse a cabo por medio de modelos: matemático y de simulacion.

a) Un modelo matemático, es aquel en el cual se toman supuestos con respecto a las llegadas y servicio y posteriormente se escriben ecuaciones que representan partes del proceso. Por ejemplo, puede comenzar por suponerse que el porcentaje de llegadas está relacionado al tiempo entre llegadas por una ley matemática representada en la siguiente figura:

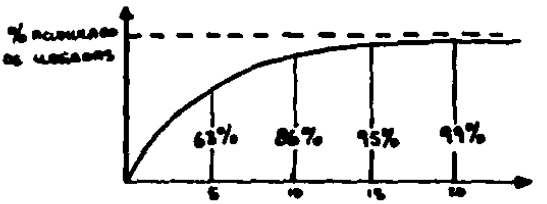


Figura 2

La ecuación de la curva que representa la distribución de tiempo entre llegadas, se usa entonces

junto con otra ecuación que expresa la distribución del tiempo de servicio, para llegar a relaciones quedescriban el proceso. Estas relaciones se resuelven para obtener el tiempo promedio de espera, longitud esperada de la cola, etc. Cuando pueden asignarse costos del tiempo de espera y de los servidores desocupados, además de costos de operación, pueden determinarse las condiciones para minimizar el costo total.

b) Un modelo de simulación, es aquel en la cual se duplican, mecánicamente, estadísticas de llegadas y servicio (Tabla 2). Este método, conocido como el de Monte Carlo, resuelve el problema jugando ya sea con las distribuciones reales, o con las supuestas. Variando las estadísticas y duplicando miles de llegadas, puede estudiarse el efecto de un cambio en las condiciones sin tener que esperar los datos reales en un periodo largo de tiempo. Este método es particularmente valioso si se cuenta con una computadora.

1. Se toman muestras aleatorias de:
distr. de llegadas y distr. de servicio
2. Se observan:
Tasa de llegadas — Tasa de servicio
Tiempo de espera — Tiempo de ocio
Longitud de la cola
(clientes)
3. Se repite el muestreo (en papel) un número grande de veces para cada variación de factores (Tasa de servicio, Tasa de llegadas, etc.) que pueden controlarse.
4. Se asignan costos al tiempo de espera y al tiempo de ocio.
5. Se determina la política para obtener el mínimo costo total.

Tabla 2

2.2 UN MODELO MATEMATICO.

Para desarrollar fórmulas de la longitud esperada de la cola en el ejemplo de la gasolinera, se comienza escribiendo hechos acerca de las probabilidades de que haya un número dado de autos en la linea de espera bajo varias condiciones de llegada y de servicio posibles. Esto lleva a ecuaciones que pueden resolverse para elementos particulares del proceso. Para el caso de la gasolinera, se tomaran como supuestos que la distribucion de llegadas es Poisson y la de tiempo de servicio es Exponencial. Se tiene que si λ es el número de llegadas por unidad de tiempo (tasa promedio de llegadas) y μ es el número de unidades que pueden ser servidas por unidad de tiempo (tasa promedio de servicio), entonces la longitud promedio de la cola L, (incluyendo el auto que está siendo atendido), está dada por la fórmula

$$L = \lambda / (\mu - \lambda)$$

Esta fórmula de la longitud promedio de la cola, muestra que cuando la tasa promedio de servicio se acerca a la de llegadas, la longitud de la cola se incrementa rápidamente.

A la proporción λ/μ se le llama intensidad de tráfico. Por lo tanto, si se desea evitar mucho tráfico, es necesario plantear las instalaciones de servicio de tal forma que λ/μ sea menor que 1. Aunque este resultado puede ir en contra de la intuición, debe recordarse que λ y μ son tasas promedio, y los tiempos de llegadas y de servicio son aleatorios. Si λ se acerca mucho a μ, el tiempo "perdido" es difícil de recuperar una vez que la cola se comienza a formar. La longitud promedio de la cola y el tiempo de espera tienden a infinito, a menos que se interrumpa la actividad de la estación.

Si la tasa promedio de servicio es igual a la tasa promedio de llegadas, la cola crece fuera de toda frontera.

En el caso de la gasolinera, la tasa promedio de llegadas es un auto cada 2.5 min. Entonces, con base a un intervalo de 5 min., $\lambda=2$. La tasa promedio de servicio es 3 autos en un intervalo de 5 min. De ahí que $\mu=3$ y $\lambda/\mu=2/3$. Sustituyendo en la fórmula, L=2. Esto significa que si los supuestos acerca de las llegadas y servicios se cumplen, debe esperarse que en promedio, hayan 2 autos en la cola esperando servicio. En esta formulación del problema, cuando un auto está siendo servido, se incluye como miembro de la cola.

Para este tipo de problema (de un solo servidor), tambien es posible derivar una fórmula para el tiempo promedio de espera W, de los autos en la cola. La fórmula es:

$$W = \lambda / (\mu(\mu - \lambda))$$

(Esto no incluye el tiempo requerido de servicio). En el ejemplo, el tiempo promedio de espera es 2/3 min. o 40 segundos.

Por lo tanto, parece que con un empleado en turno es suficiente sin arriesgarse a una cola grande y sin ocasionar largas esperas a los clientes.

En muchos casos, las distribuciones que son resultado de las observaciones directas no pueden expresarse por fórmulas matemáticas y entonces se recurre a los métodos de Monte Carlo o de Simulación.

2.3 UN MODELO DE SIMULACION.

Para ilustrar el método de Monte Carlo en la solución del problema de la gasolinera supondremos que se anotan los tiempos de llegadas de autos y el número de minutos requeridos de servicio para cada auto en un intervalo de media hora, de 8:00 a 8:30. Esta información podria obtenerse ya fuera por muestreo de las distribuciones teóricas o supuestas, o por muestreo

de las operaciones actuales en un periodo de tiempo. Los resultados de un muestreo se encuentran en la siguiente tabla.

Tiempo lleg.	Tiempo serv.	Tiempo com.	Tiempo term.	Long. cola	Tiempo espera	Tiempo ocio
8:00	1	8:00	8:01	1	0	0
8:03	1	8:03	8:04	1	0	2
8:04	2	8:04	8:06	1	0	0
8:05	1	8:06	8:07	2	1	0
8:06	4	8:07	8:11	2	1	0
8:08	1	8:11	8:12	2	3	0
8:09	1	8:12	8:13	3	3	0
8:14	1	8:14	8:15	1	0	1
8:19	2	8:19	8:21	1	0	4
8:21	2	8:21	8:23	1	0	0
8:26	1	8:26	8:27	1	0	3
8:28	2	8:28	8:30	1	0	1
Total	19			17	8	11
Promedio	1:58			1:42	0.67	0.92

Tabla 3

En base a esta muestra (el procedimiento debe repetirse varias veces), deben esperarse, en promedio, 1 o 2 autos en la cola en un momento dado, el tiempo promedio de espera para un auto es de 40 seg., el empleado tiene un promedio de 55 seg. desocupados entre servicios, y emplea un poco mas de 1.5 min. en promedio en cada cliente.

Si el procedimiento de muestreo se repite un número grande de veces, los promedios de los resultados obtenidos pueden tomarse como valores representativos del proceso. Una ventaja de este método es que puede utilizarse aun con las distribuciones de llegadas y servicio que no pueden anotarse explícitamente. En este caso, el muestreo puede tomarse de las operaciones reales con el supuesto de que los datos son representativos del proceso. Otra ventaja es la manipulación que puede hacerse de los factores que están sujetos al control. Por ejemplo, puede investigarse el efecto de dar más o menos tiempo para servicio, o de incrementar o disminuir el número de unidades de servicio, de cambiar las tasas de llegada si es posible, etc. La investigación puede llevarse a cabo alimentando las estadísticas necesarias en una computadora duplicando así, en unos cuantos segundos, miles de llegadas y servicios.

Por medio del siguiente ejemplo se ilustra el método de Monte Carlo como ayuda para tomar una decisión de la forma óptima de acomodar instalaciones de servicio para manejar un tipo de llegadas y servicio en una sola estación.

Supóngase que a los empleados de una gasolinera se les paga $4000.00 al día por 8 horas de trabajo. Los autos llegan para servicio (gasolina, aceite, llantas, etc.) a una tasa promedio de un auto cada 5 min. El tiempo de servicio en promedio es de 4 min. por auto con un empleado, 3 min. por auto con dos empleados y 2 min. por auto con tres empleados. La utilidad promedio por cliente se estima en $500.00. Debido a la localización de la gasolinera, se ha notado que el número de autos en cola en cualquier momento, no afecta la decisión de otro cliente de unirse a ella. Sin embargo, cuando la cola excede los 3 autos, algunos clientes se impacientan y se marchan sin servicio. El promedio de "perdidas" es de 50% con un empleado, 20% con dos y 10% con tres. Se supone que no existen otros factores que afecten la situación. El problema es decidir el número de empleados que deben tenerse en un turno para minimizar el costo de mano de obra más el costo debido a pérdidas de clientes en un periodo de 8 horas.

El método de Monte Carlo consiste en tomar repetidamente muestras aleatorias de la distribución de llegadas y de cada una de las tres distribuciones alternativas de servicio. Llegadas, tiempos de servicio y clientes perdidos pueden duplicarse muchas veces en una computadora sin tener que esperar mucho tiempo para obtener datos reales. El costo total puede entonces determinarse para cada condición de servicio, y la alternativa cuyo costo total sea el menor, será considerada óptima.

Los datos resumidos despues de varias muestras para un turno de 8 horas son los siguientes:

No. empleados	1	2	3
Tasa prom. llegadas en 5 min.	1	1	1
Tasa prom. servicio en 5 min.	1 1/4	1 2/3	2 1/2
No. autos unen a cola de + de 3	30	15	10
% clientes perdidos	50%	20%	10%
No. promedio clientes perdidos	15	3	1
Costo por clientes perdidos	7500	1500	500
Costo por empleados	4000	8000	12000
Costo total	11500	9500	12500

De los datos puede verse que el mínimo costo total es de $9500 cuando 2 empleados están en servicio. En este caso no se están tomando en cuenta otros costos como el de clientes insatisfechos por el servicio. Debe notarse también que los costos totales obtenidos son costos "esperados" o costos promedio que se obtendrán después de un largo periodo de servicio bajo las condiciones supuestas.

Los ejemplos aquí expuestos son muy simples. En la aplicación de Teoría de Colas, pueden tomarse en cuenta variaciones en las tasas de llegada y servicio por día de semana, hora del día o estación del año. Una vez que se han reunido suficientes datos que describan las condiciones, los efectos de tales factores como descompostura de equipo, mal clima, tiempos extras, etc. y los costos totales resultantes de la operación, pueden estudiarse utilizando la técnica de Monte Carlo.

1.4 MEDIDAS DE EFICIENCIA.

Al buscar cantidades cuantificables que sirvan como criterios para estudiar un sistema de colas, inmediatamente viene a la mente el estudio de promedios como posibles medidas de eficiencia. Aproximaciones a la teoría han llevado a medidas de eficiencia que no

solo estan basadas en la teoría sino que han probado ser medidas efectivas para proveer remedios a problemas prácticos.

A continuación se darán algunos ejemplos de medidas de eficiencia utilizadas en Teoría de Colas.

El conocimiento de la longitud de la línea de espera permite tomar decisiones acerca de si hay espacio suficiente para los clientes que deban hacer cola para esperar servicio.

La probabilidad de tener n clientes esperando y en servicio en el tiempo t, dadas las condiciones iniciales del sistema es necesaria para calcular el número promedio de clientes en el sistema.

La probabilidad de esperar menos que un tiempo dado T es necesaria para calcular el tiempo promedio de espera.

Si la probabilidad de esperar mas de T es grande, los clientes pueden desanimarse y ya no harán cola; por lo tanto esta probabilidad puede utilizarse como una medida de eficiencia.

El tiempo promedio de espera puede utilizarse, por ejemplo, para determinar si es más barato aumentar el número de canales o estaciones de servicio o mejorar el sistema disminuyendo el tiempo promedio de servicio. Así, puede tomarse una decisión, tomando en cuenta los costos, acerca de cual mejora debe llevarse a cabo.

Las medidas mas comunes y mayormente utilizadas de un sistema de colas son: utilización de servidores, intensidad de tráfico, tiempo de respuesta o de estancia en el sistema, tiempo de espera en la cola, tiempo de servicio, longitud de la cola y número promedio de unidades en el sistema. Otra medida muy útil es el percentil: encontrar el valor de x tal que y% del tiempo el número de unidades en el sistema sea menor que x. (Mientras mayor sea y, menos probabilidad de congestionar el sistema).

2.5 EVALUACION DE DESEMPEÑO:

Con ayuda de las medidas de eficiencia analizadas en la sección anterior, es factible y al mismo tiempo fácil, analizar y evaluar el desempeño de un sistema.

Debe tenerse en cuenta que se necesitan comprender, intuir e incluir un número de parámetros claves para poder controlar un sistema.

Esto puede ayudar a evitar un serio problema como es el que durante el desarrollo de un sistema no se considere el desempeño y que al terminarse sea totalmente inaceptable. En este caso, las dos opciones existentes son: abandonar el sistema o rediseñarlo y volverlo a desarrollar hasta que sea aceptable.

Cualquiera de estas opciones es mucho más cara que el diseño y desarrollo de un sistema que desde el principio considere desempeño explícitamente.

Algo que se debe tener en cuenta porque beneficia en última instancia el desempeño de un sistema es compartir recursos. El ejemplo clásico es el de un sistema operativo multiprogramable, cuyo objetivo es tener varios programas utilizando el procesador y las unidades de entrada y salida al mismo tiempo de tal forma que cada uno progrese en forma similar a como lo haría si estuviera solo en la máquina. Este compartimiento de recursos reduce el costo atribuido a cada programa, i.e., si un solo programa utiliza la máquina, debe cargársele el costo del tiempo que esté desocupada y del que esté ocupada. En el sistema de multiprogramación ideal, a los programas se les carga solo el tiempo que utilizan los recursos. Sin embargo, el compartirlos es causa inherente de competir por ellos; si dos programas necesitan el procesador, uno debe esperar. En un sistema bien diseñado, la ganancia por compartir es mayor que la pérdida por competir. Pero esta competencia es generalmente un factor muy significativo de desempeño y uno de los mas difíciles de cuantificar.

El análisis del desempeño de un sistema de cómputo es una base para planear el futuro.

2.6 ALGUNAS CONSIDERACIONES IMPORTANTES.

2.6.1 Familia de problemas de Colas.

Situaciones de colas difieren entre sí en el número de canales de servicio disponibles y en las estadísticas del proceso de servicio. Pueden existir otras diferencias como un patrón de llegadas que no sea Poisson o prioridades en la forma de servicio pero estas diferencias son poco frecuentes. Es por esto que una familia de dos parámetros de problemas de Colas corresponde a la mayor parte de las situaciones de líneas de espera mas comunes. Estos parámetros son: el número de canales de servicio y la intensidad de tráfico y sus componentes.

2.6.2 Salida de un sistema de colas.

Para un sistema de colas con llegada Poisson, una sola línea de espera sin deserciones y tiempos de servicio Exponenciales, la distribución de equilibrio del número de clientes atendidos en un intervalo de

tiempo arbitrario es la misma que la distribución de llegadas, es decir Poisson, para cualquier número de servidores.

Este resultado tiene aplicaciones en problemas de colas de colas.

Un ejemplo es el de clientes en una tienda que deben ser atendidos primero por un vendedor y después por un cajero o alguien que envuelva su mercancía. Otro ejemplo, mas complicado es el de una llamada telefónica que pasa por un sistema de redes de teléfonos.

Intuitivamente está claro que en los procesos de colas de colas del tipo mencionado, si la distribución de salida de cada etapa es tal que el sistema formado por la segunda etapa es compatible para ser analizado, entonces la cola de colas puede analizarse por etapas. Este tipo de análisis por etapas es mucho mas sencillo que si se trata de analizar todo el sistema simultáneamente. Afortunadamente, bajo las condiciones citadas a continuación, es verdad que la salida tiene la simplicidad requerida para tratar cada etapa individualmente.

- <u>El modelo</u>

"La salida de un sistema de colas con llegadas Poisson y tiempos de servicio Exponencial es Poisson." Los detalles de esta hipótesis son los siguientes:

Suponemos una cola de una sola etapa con llegadas aleatorias. El intervalo promedio entre llegadas tiene longitud $1/\lambda$. Esto es, la probabilidad de una llegada durante un intervalo de longitud t es λt.

Existen c servidores (canales) cada uno con una distribución de tiempo de servicio Exponencial con promedio $1/\mu$. Los tiempos de servicio son completamente independientes de todas las condiciones. Por lo tanto la probabilidad de que un cliente que está recibiendo servicio al principio de un intervalo t termine durante el intervalo es μt.

Bajo estos supuestos y la condición de que $c > \lambda/\mu$, existe una distribución de equilibrio del número de clientes en el sistema. Más aún, esta distribución es la misma que la de los estados encontrados por los clientes al llegar al sistema.

Todos los clientes se quedan en el sistema hasta haber recibido servicio. De otra forma, la disciplina de la cola, u orden de servicio es irrelevante, ya que la distribución de salida y no la de retraso es la de interés.

La prueba de este teorema puede encontrarse en : "The output of a Queuing System" de Paul J. Burke, OR 1956.

- Un ejemplo

Para ilustrar una aplicación de este resultado, se considerará la idealización de la situación con vendedores y cajeros antes mencionada. Se supondrá que los clientes tienen acceso a cualquiera de los vendedores que esté libre y que los vendedores tienen igual acceso a la mercancía. Después de ser atendido por un vendedor, un cliente procede a los cajeros y tiene acceso a cualquiera de ellos sin importar la mercancía que desea comprar. Existe una sola cola frente a los vendedores y una sola cola frente a los cajeros. El servicio dado por vendedores y cajeros es en orden de llegada.

Se supondrá además que existe evidencia de que los tiempos de servicio de los vendedores pueden aproximarse satisfactoriamente por una distribución Exponencial con promedio de 1.5 min., mientras que el tiempo de servicio de los cajeros es casi constante en 1 min. El problema es determinar las cantidades de vendedores y cajeros necesarias para que las probabilidades de que un cliente deba esperar mas de 3 min. frente a los vendedores o mas de 2 min. frente a los cajeros sean cada una de menos de 0.05 por un periodo de varias horas durante las cuales las llegadas de los cliente son aleatorias (Poisson) con un promedio de 2 por minuto.

Debido a que los tiempos de servicio de los vendedores son Exponenciales, puede inferirse del teorema, que las llegadas a los cajeros son Poisson con un promedio de 2 por minuto sin importar el número de vendedores, siempre y cuando excedan de 3, y por lo tanto el número de cajeros puede determinarse independientemente del número de vendedores.

El número de vendedores requeridos es: Para 5 vendedores, el tiempo en el sistema es 0.60 ya que las llegadas son 3 por tiempo de servicio; y la probabilidad de un retraso de mas de 2 tiempos de servicio (3 min) es .0047, lo cual está dentro de los límites del criterio. Para 4 vendedores sin embargo, el tiempo en el sistema es 0.75 y la probabilidad de un retraso de mas de 2 tiempos de servicio es 0.07 lo cual no está en los límites del criterio. Se necesitan, por lo tanto, 5 vendedores.

De la misma forma si hay 3 cajeros, el tiempo en el sistema es de 0.67 y la probabilidad de retraso mayor a 2 tiempos de servicio será menos de 0.01. Por lo tanto, se necesitan 5 vendedores y 3 cajeros para cumplir con los criterios del sistema.

2.7 APLICACIONES.

Entre las aplicaciones de Teoría de Colas de mayor interés (aparte del trabajo en teléfonos), se encuentran analisis de los retrasos en el tráfico de las casetas en túneles y puentes de puertos, la obtencion del número óptimo de empleados que deben asignarse a los talleres en uso de fabricas de aviones. Se han hecho aplicaciones industriales importantes a problemas de reparación de máquinas descompuestas. Una compañía debe asignar mecánicos para reparar máquinas de tal forma que se minimicen pérdidas de producción causadas por la falta de máquinas en buen estado. Esencialmente, las máquinas forman una linea de espera para ser reparadas por los mecánicos que les dan servicio. Existe un punto en el cual los salarios de los mecánicos asignados a estar listos en caso de una falla en alguna máquina son mayores que la pérdida de producción potencial. De nuevo aqui, se busca el número óptimo de mecánicos utilizando Teoria de Colas.

En el gran número de los procesos que valen la pena de ser estudiados por analistas de operaciones, una gran parte puede describirse muy bien por un modelo de colas. La teoría sería valiosa para obtener información similar a la discutida en el problema de lavado de autos

descrito en la Sección 1.2 y en la sección de medidas de eficiencia.

Otras aplicaciones de Teoria de Colas son: conversaciones telefonicas, aterrizaje de aviones, carga y descarga de barcos, citas de pacientes en clinicas, el paso de gente en la aduana, descompistura y reparacion de maquinas, moviemiento de aeronaves en pistas, lavado de autos, servicio en un restaurante, gente en la parada de autobus, clientes impacientes en una tienda, radiocomunicaciones, supermercados, boletos de cine y teatro, flujo de produccion, etc.

CAPITULO III

MODELOS DE COLAS EN SISTEMAS DE COMPUTO.

Hasta ahora se ha hablado de Teoría de Colas y se han mencionado los sistemas de cómputo varias veces, a continuación se desarrollan varias aplicaciones más de ésta Teoría en dichos sistemas.

Un sistema de cómputo puede representarse como una red de colas y ser evaluado analiticamente. Los recursos del sistema son los centros de servicio, y los usuarios o transacciones los clientes.

La creciente popularidad de Teoría de Colas para modelar sistemas de cómputo se debe principalmente a cuatro razones:

1) Estos modelos capturan los factores más importantes de los sistemas. Por ejemplo mecanismos independientes con colas y jobs moviéndose de un mecanismo al siguiente.

2) Los supuestos del análisis son realistas. Las distribuciones Exponenciales del tiempo de servicio se manejan en muchos mecanismos.

3) Los algoritmos que resuelven las ecuaciones del modelo están disponibles como paquetes muy eficientes de evaluación de redes de colas.

4)Con los modelos de colas pueden estudiarse medidas de desempeño tales como utilizaciones, longitud promedio de la cola y tiempo promedio de respuesta en sistemas reales.

3.1 APLICACIONES DE UN MODELO ABIERTO DE COLAS.

El sistema de colas M/M/c es un modelo simple y abierto que puede ser utilizado para modelar, al menos aproximadamente, muchos sistemas de cómputo. Se considera abierto porque los clientes llegan al sistema desde afuera, reciben servicio y salen del sistema. Los sistemas cerrados, en los que los clientes nunca salen del sistema, serán considerados después. Las ecuaciones de este sistema de colas, así como su demostración, se encuentran en el Apéndice A de ésta tesis.

Esta aplicación tiene base en una empresa en la que los empleados dicen tener que esperar mucho tiempo para poder utilizar la computadora. La lógica indica que el número de terminales que actualmente se tienen son suficientes para que no deban esperar demasiado. Puede Teoría de Colas ayudar a la toma de una decisión en este caso?

Ejemplo 1. Una empresa tiene una terminal en línea conectada a un sistema central de cómputo 8 horas al día. El patrón de la gente que llega a la oficina a utilizarla es Poisson con un promedio de 10 personas al día. La distribución del tiempo que pasa una persona en la terminal es aproximadamente Exponencial con un promedio de 1/2 hora. Por lo tanto la terminal es utilizada 5/8 (10 x 1/2 = 5 horas de 8 posibles en promedio). El gerente recibe quejas de los usuarios por la cantidad de tiempo que muchos de ellos tienen que esperar para utilizar la terminal. Al gerente no le parece razonable poner otra terminal cuando la de ahora sólo se utiliza 5/8 del tiempo, en promedio. Cómo puede Teoría de Colas ayudar al gerente?

Solución. El sistema de colas M/M/1 es un modelo razonable de este sistema con u = 5/8. Usando las ecuaciones del Apéndice A para el modelo M/M/c con c = 1, pueden calcularse las medidas de desempeño.

W = 50 min. Tiempo promedio de espera en la cola.

R = 80 min. Tiempo promedio en el sistema.

Pq(90) = 146.61 min. 90 percentil de tiempo en la cola.

Como λ = 10 personas por cada 8 horas = 10/(8x60) = 1/48 personas/min.

L = 1.0417 Número promedio de gente en la cola.

n = 1.667 personas Número promedio de gente para usar terminal.

Estas estadísticas muestran que más de una persona al día se pierde haciendo cola para utilizar la terminal.

En el problema del próximo ejemplo se mostrará como el modelo M/M/c puede utilizarse para ayudar a obtener la información necesaria para tomar una decisión para resolver el problema.

Ejemplo 2. Aunque la terminal se utiliza sólo 62.5 %, el tiempo promedio en la cola son 50 min. con 10% teniendo que esperar mas de 146.61 min. El problema no podía resolverse con horarios de utilización para los usuarios asi es que debían agregarse una o más terminales. Se especificó que el tiempo promedio en la cola debia ser menor a 10 min. con 90% sin exceder 15 min. El gerente pensó que si el tiempo promedio en la cola con una terminal era de 50 min., con dos terminales seria de 25 min. y que por lo tanto hacían falta 5 terminales. Cuántas hacen falta?

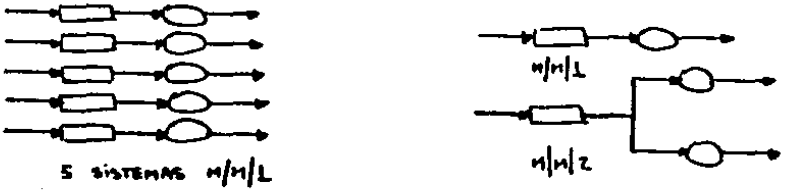


Figura 3

Solución. La solución depende de cómo se utilicen las terminales i.e. puede utilizarse el sistema de colas M/M/c, o varios sistemas M/M/1. Primero se buscará la solución con el sistema M/M/c para c=2.

$u = .625/2 = .3125$

$C(2, \lambda/\mu) = C(2, .625) = .1488$

$W = 3.247$ min.

$Pq(90) = 8.67$

Por lo tanto basta otra terminal para satisfacer todos los requerimientos.

Si la segunda terminal se utiliza de tal forma que el tráfico se reparta igual entre las dos, i.e. con 2 sistemas M/M/1, cada una con $u = .3125$ entonces:

$W = 13.64$ min.

$Pq(90) = 49.72$ min.

Pq(90)	terminales	sistema	u	W
146.61	1	M/M/1	.624	50
8.67	1	M/M/2	.3125	3.25
49.72	2	M/M/1	.3125	13.64
15.86	4	M/M/1	.1562	5.55
7.65	5	M/M/1	.125	4.29

Tabla 4

Por lo tanto harían falta cinco terminales para cumplir con los requisitos. La Tabla 4 muestra los valores para varios sistemas.

3.2 MODELOS DE COLAS CON POBLACION FINITA DE SISTEMAS DE COMPUTO INTERACTIVOS.

En los dos ejemplos anteriores se utilizaron modelos abiertos con un número infinito de clientes. En la realidad, hay muy pocos sistemas que sean realmente abiertos e infinitos; sin embargo, muchos sistemas pueden ser aproximados por dichos modelos. Se considerará ahora un modelo de colas más realista que es cerrado (ningún cliente entra o sale del sistema) y tiene una población finita de clientes.

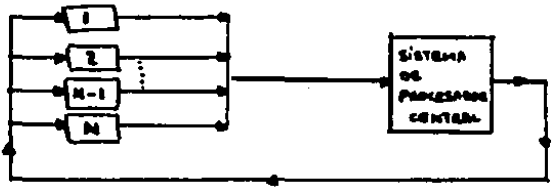


Figura 4

En la figura se muestra un modelo de colas con población finita que podría representar un sistema de cómputo interactivo. El sistema consta de una cola para el procesador central, un CPU, algunos mecanismos de I/O, y un sistema de colas asociado a estas unidades de servicio. Los clientes (usuarios) interactúan con el sistema por medio de N terminales. Cada cliente (usuario) está exactamante en uno de tres estados en cualquier instante del tiempo: (1) "pensando" en la terminal, (2) haciendo cola para un tipo de servicio, o (3) recibiendo servicio. El tiempo en que está pensando, incluye todo el tiempo que pasa entre la finalización del servicio, hasta que se hace un pedido para otra interacción. Un usuario en una terminal no puede pedir servicio de CPU hasta que el anterior servicio ha sido terminado. Las ecuaciones válidas para todos los modelos de este tipo están en el Apéndice A.

Este modelo puede aplicarse en un problema práctico de un sistema de cómputo como el que sigue.

Ejemplo 3. Un sistema de cómputo interactivo finito tiene 20 terminales activas, un tiempo promedio de pensamiento de 3 seg. una tasa promedio de servicio del CPU de 500,000 instrucciones por segundo, y requisitos promedio de 100,000 instrucciones. Se desea saber el tiempo promedio de respuesta R y la utilización del CPU. Se desea saber también cómo cambiarían estos datos si se agregan 10 terminales i.e. con un total de 30 terminales.

Solución. Como cada interacción requiere de la ejecución de 100,000 instrucciones de CPU en promedio,

$$\lambda/\mu = 500{,}000/100{,}000 = 5 \text{ seg.}$$

La probabilidad de que el CPU esté desocupado es

$$Po = .045593216$$

El tiempo promedio de respuesta es:

$$R = 1.191 \text{ seg.}$$

La utilización del CPU es:

$$\rho = .9544$$

Las interacciones por unidad de tiempo son:

$$\lambda t = 4.722 \text{ interacciones por segundo}$$

Si el número de terminales se aumenta a 30, entonces los resultados son:

$$Po = .00022118$$

$$\rho = .99977882$$

R = 3.00 seg.

λ t = 5 interacciones por seg.

Por lo tanto un aumento del 50% en el número de terminales aumentó las salidas en solo 5.8% mientras que el tiempo de respuesta aumentó en un 151.9%!!

Este ejemplo muestra el concepto de saturación de un sistema. Si hay una terminal, no hay cola. Para pocas terminales, los usuarios interfieren muy poco mutuamente. Generalmente cuando un usuario quiere utilizar el CPU, los otros están pensando así que hay poca cola.

Existen casos en los que un número elevado de usuarios ocasiona que interfieran entre sí i.e. que compitan por los recursos en vez de compartirlos.

3.3 EL MODELO DE SERVIDOR CENTRAL DE MULTIPROGRAMACION.

Este modelo es más complejo (con mayor detalle) que los modelos considerados previamente. El modelo está representado en la figura siguiente:

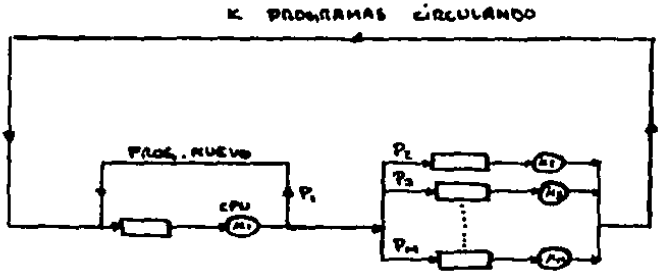


Figura 5

Es un modelo cerrado ya que contiene un número fijo de programas k que circulan en el sistema interminablemente. Sin embargo, cada vez que un programa completa un ciclo del CPU al CPU, se supone que se ha completado una ejecución de un programa y uno nuevo entra al sistema.

Existen M-1 mecanismos de I/O, cada uno con su propia cola y cada uno con tiempos de servicio distribuidos Exponencialmente con tasas promedio de servicio μi (i=2,3,...,M). Se supone que el CPU también provee servicio Exponencial (con tasa promedio de $\mu 1$). Al terminarse una ejecución en el CPU, el job regresa al CPU con probabilidad p1 o requiere de un mecanismo de I/O con probabilidad pi (i=2,3,...,M). Al salir del mecanismo de I/O, regresa a la cola del CPU para otro ciclo. Sea K = (k1,k2,...,kM) el estado del sistema, con K el número de jobs en la i-ésima cola (en cola o en

servicio), entonces las probabilidades p(k1,k2,...,kM) de que el sistema esté en el estado K están dadas por:

$$P(k_1,k_2,\ldots,k_M) = \frac{1}{G(K)} \prod_{i=1}^{M} \left(\frac{\mu_i p_i}{\mu_i}\right)^{k_i}$$

para (k1,k2,...,kM) tales que $\sum_{i=1}^{M} k_i = K$

donde G(K) se define de tal forma que las probabilidades sumen 1.

El método para calcular G(0)=1, G(1),G(2),...,G(K) está dado a continuación:

Dados μ_i, pi para i=1,2,...,M, el algoritmo genera G(K),G(K-1),...G(1),G(0)=1

1. asignar valores a xi:

   x1=1 xi=$\mu_1$pi/μ_i i=2,3,...,M

2. asignar valores iniciales:

   g(k,1) = 1 para k = 0,1,...,K

   g(0,h) = 1 para h = 1,2,...,M

3. inicializar k:

   k = 1

4. calcular el k-ésimo renglón:

   g(k,h) = g(k,h-1) + xhg(k-1,h) h = 2,3,...,M

5. incrementar k:

   k = k+1

6. algoritmo finalizado?

   Si k < K ir a 4. De otra forma, terminar el algoritmo.

$g(n,M) = G(n)$ para $n = 0,1,\ldots,K$

Las ecuaciones que describen este modelo se encuentran en el Apendice A.

A continuación se muestra una aplicación en la que se hace una evaluación de desempeño y se obtiene una predicción, por medio del cálculo de medidas de eficiencia, sobre los efectos que tendrian dos cambios en un modelo de este tipo.

Ejemplo 4. Se desea modelar un sistema en el cual el sistema de procesador central es el modelo del servidor central. Durante el periodo más ocupado del día, se tiene un promedio de 100 terminales activas con un tiempo promedio de pensamiento de 13 seg. El tiempo promedio de respuesta observado es de 6.41 seg.

$\lambda t = 5.15$ interacciones por min.

Los parámetros de la parte del modelo que es servidor central son:

$M = 3$ Un CPU y dos mecanismos de I/O

$K = 4$ Nivel de multiprogramación

$\mu 1 = 100$

$\mu 2 = 25$ tasas de servicio

$\mu 3 = 40$

p1 = .1

p2 = .2 probabilidades de ramificación

p3 = .7

Valores observados de utilización:

ρ_1 = 0.521 utilización del CPU

ρ_2 = 0.409 utilización del primer mecanismo I/O

ρ_3 = 0.911 utilización del segundo mecanismo

λt = 5.18 interacciones/seg.

R = 6.41 seg. tiempo promedio de respuesta

Si el modelo se ajusta razonablemente, se desean investigar los efectos de dos posibles mejoras de hardware. La primera mejora considerada es la de dar memoria principal adicional suficiente para aumentar el nivel de multiprogramación de 4 a 15. La segunda mejora es la de hacer cambios en hardware y software que aumenten las velocidades de los dos mecanismos de I/O en un 25% dejando el nivel de multiprogramación en 4.

Solución. Con el algoritmo antes mostrado (algoritmo de Buzen) se calcula:

x1 = 1, x2 = 0.8 y x3 = 1.75

Continuando con el algoritmo:

	x1	x2	x3	
	1	.8	1.75	
0	1	1	1	= G(0)
1	1	1.8	3.55	= G(1)
2	1	2.44	8.6525	= G(2)
3	1	2.952	18.093875	= G(3)
4	1	3.3616	35.02588125	= G(4) = K

Sustituyendo en las ecuaciones:

$\rho 1$ = .51658586 utilización del CPU

$\rho 2$ = .413268688 utilización del 1er. mecanismo

$\rho 3$ = .904025256 utilización del 2do. mecanismo

λt = 5.1658586 salidas

R = 6.357866 tiempo promedio de respuesta

Estos valores son muy cercanos a los medidos para validar el modelo.

Cálculos similares muestran que si el nivel de multiprogramación se incrementa de 4 a 15, entonces:

$\rho 1$ = .571282414

$\rho 2$ = .457025931

$\rho 3$ = .999744225

λt = 5.71282

R = 4.505 seg.

Por lo tanto las salidas promedio se han incrementado en un 10.6% y el tiempo promedio de respuesta ha disminuido en un 29.2%.

Si el nivel de multiprogramación se mantiene en 4 y se incrementa la velocidad de los mecanismos en un 25%, entonces:

$$\rho_1 = .616300129$$

$$\rho_2 = .394432082$$

$$\rho_3 = .86282018$$

$$\lambda t = 6.163$$

$$R = 3.226 \text{ seg.}$$

La segunda mejora parece ser más favorable que la primera; mejora las salidas promedio en un 19.3% y disminuye el tiempo promedio de respuesta en un 49.3% y no sobrecarga los mecanismos tanto como la primera mejora.

Si se hicieran las dos mejoras, λt sería 7.123 interacc./seg. y R sería 1.039 seg.

Esto muestra un incremento del 15.6% en salidas promedio y una disminución en el tiempo de respuesta en un 67.8% sobre la segunda mejora. Comparado al sistema original representa un incremento en λt de 24.7% y una disminución del 76.9% en tiempo promedio de respuesta.

Si el primer mecanismo del sistema original es reemplazado por uno con la misma velocidad que el segundo y la carga de los dos se balancea de tal forma que la probabilidad de ramificación de cada uno sea de .45, entonces λt sería 6.14 interacc./seg. y R sería 3.285 seg.

Este modelo se utiliza para modelar sistemas de cómputo de multiprogramación.

Se ha mostrado, con breves explicaciones y varios ejemplos, como la Teoría de Colas puede ser utilizada para crear modelos analíticos de sistemas de cómputo. Estos modelos son eficientes y fáciles de utilizar, y requieren muy poco esfuerzo para obtener predicciones razonablemente exactas del desempeño de un sistema.

APENDICE A

EL PROCESO DE NACIMIENTO Y MUERTE.

En el contexto de Teoría de Colas, un nacimiento es la llegada de un cliente al sistema de colas, y el término muerte se refiere a una salida de un cliente que ya ha sido servido.

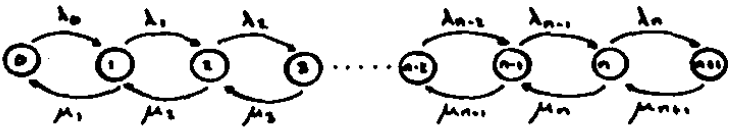
Los supuestos de un proceso de nacimiento y muerte son los siguientes.

Supuesto 1. Dado que el número en el sistema es n, la distribución de probabilidad del tiempo que queda para que suceda el siguiente nacimiento (llegada) es exponencial con parámetro λ_n $(n = 0, 1, 2, \ldots)$.

Supuesto 2. Dado que el número en el sistema es n, la distribución de probabilidad del tiempo que queda para que suceda la siguiente muerte (finalización de servicio) es exponencial con parámetro μ_n $(n = 0, 1, 2, \ldots)$.

Supuesto 3. Sólo un nacimiento o muerte puede ocurrir en un momento dado.

Como λ_n y μ_n son las tasas promedio, podemos resumir estos supuestos en el siguiente diagrama de transición.



Las flechas en este diagrama muestran las transiciones posibles en el sistema y junto a cada flecha se muestra la tasa con la que pasa de un estado a otro.

Considérese cualquier estado del sistema n (n = 0,1,2,...). Supóngase que se empieza a contar el número de veces que el proceso entra en este estado y el número de veces que sale de este estado. Como los dos tipos de incidentes deben alternar, estos dos números deben ser iguales o diferir solo por 1. Esta posible diferencia de 1 causaría solo una muy pequeña diferencia en las tasas promedio. Por lo tanto estas dos tasas deben ser

iguales en el largo plazo. Esto nos lleva al siguiente principio.

> TASA DE ENTRADA = TASA DE SALIDA. Para cualquier estado del sistema, la tasa promedio con la que ocurren los incidentes de entradas debe ser igual a la tasa promedio con que ocurren los incidentes de salida.

La ecuación que expresa este principio se llama ecuación de balance para el estado n. Después de construir las ecuaciones de balance para todos los estados en términos de las probabilidades desconocidas Pn (Probabilidad de que estén n en el sistema), este sistema de ecuaciones puede resolverse para encontrar dichas probabilidades.

Para ilustrar una ecuación de balance, considérese el estado 0. El proceso entra a este estado solo del estado 1. Por lo tanto la probabilidad de estar en el estado 1 (P1) representa la proporción de tiempo que sería posible para que el proceso entre en el estado 0. Dado que el proceso está en el estado 1, la tasa promedio de entrar al estado 0 es μ_1 . (En otras palabras, para cada unidad de tiempo que el proceso está en el estado 1, el número esperado de veces que saldría del estado 1 para entrar al estado 0 es μ_1 .) De cualquier otro estado, esta tasa promedio es 0. De ahí que la tasa promedio a la que el proceso sale de su estado actual para entrar al estado 0 es

$$\mu_1 P_0 + o(1 - P_1) = \mu_1 P_1$$

Con el mismo razonamiento, la tasa promedio de salidas debe ser $\lambda_0 P_0$, por lo que la ecuación de balance para el estado 0 es

$$\mu_1 P_0 = \lambda_0 P_0$$

Para cada otro estado existen dos posibles transiciones una de entrada y otra de salida del estado. Por lo tanto cada lado de las ecuaciones de balance para estos estados representa la suma de las tasas promedio para las dos transiciones. A parte de esto, el razonamiento es exactamente igual al que se hizo para el estado 0. Esta ecuaciones de balance se encuentran en la siguiente tabla.

Estado	TASA ENTRADA	=	TASA SALIDA
0	$\mu_1 P_1$	=	$\lambda_0 P_0$
1	$\lambda_0 P_0 + \mu_2 P_2$	=	$(\lambda_1 + \mu_1) P_1$
2	$\lambda_1 P_1 + \mu_3 P_3$	=	$(\lambda_2 + \mu_2) P_2$
⋮			⋮
n-1	$\lambda_{n-2} P_{n-2} + \mu_n P_n$	=	$(\lambda_{n-1} + \mu_{n-1}) P_{n-1}$
n	$\lambda_{n-1} P_{n-1} + \mu_{n+1} P_{n+1}$	=	$(\lambda_n + \mu_n) P_n$
⋮			⋮

Resolviendo este sistema de ecuaciones.

Estado

$$0 \quad : \quad P_1 = \frac{\lambda_0}{\mu_1} P_0$$

$$1 \quad : \quad P_2 = \frac{\lambda_1}{\mu_2} P_1 + \frac{1}{\mu_2}(\mu_1 P_1 - \lambda_0 P_0) = \frac{\lambda_1}{\mu_2} P_1 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} P_0$$

$$2 \quad : \quad P_3 = \frac{\lambda_2}{\mu_3} P_2 + \frac{1}{\mu_3}(\mu_2 P_2 - \lambda_1 P_1) = \frac{\lambda_2}{\mu_3} P_2 = \frac{\lambda_2 \lambda_1 \lambda_0}{\mu_3 \mu_2 \mu_1} P_0$$

$$\vdots$$

$$n \quad : P_{n+1} = \frac{\lambda_n}{\mu_{n+1}} P_n + \frac{1}{\mu_{n+1}}(\mu_n P_n - \lambda_{n-1} P_{n-1}) = \frac{\lambda_n}{\mu_{n+1}} P_n = \frac{\lambda_n \lambda_{n-1} \cdots \lambda_0}{\mu_{n+1} \mu_n \cdots \mu_1} P_0$$

Para simplificar la notación, sea

$$C_n = \frac{\lambda_{n-1} \lambda_{n-2} \cdots \lambda_0}{\mu_n \mu_{n-1} \cdots \mu_1} \qquad n = 1, 2, \ldots$$

Por lo tanto las probabilidades son

$$P_n = C_n P_0 \qquad n = 1, 2, \ldots$$

El requerimiento de que

$$\sum_{n=0}^{\infty} P_n = 1$$

implica que

$$\left[1 + \sum_{n=1}^{\infty} C_n\right] P_0 = 1$$

y así,

$$P_0 = \frac{1}{1 + \sum_{n=1}^{\infty} C_n}$$

Dada esta información

$$N = \sum_{n=0}^{\infty} n P_n$$

También, como el número de servidores c representa el número de clientes que pueden ser servidos simultaneamente (y así salen de la cola),

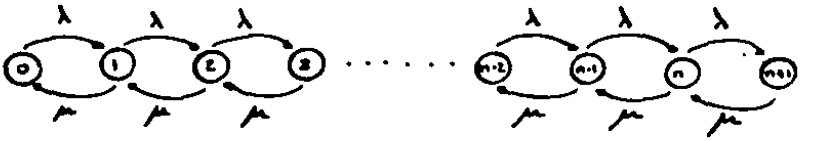
$$L = \sum_{n=c}^{\infty} (n-c)P_n$$

Estos resultados fueron obtenidos bajo el supuesto de que λ_n y μ_n tienen valores tales que el proceso llega a una condición estable. Este supuesto se mantiene si $\lambda_n = 0$ para algún valor de n, por lo que solo un número finito de estados es posible. También se mantiene siempre si $\rho = \lambda/c\mu < 1$.

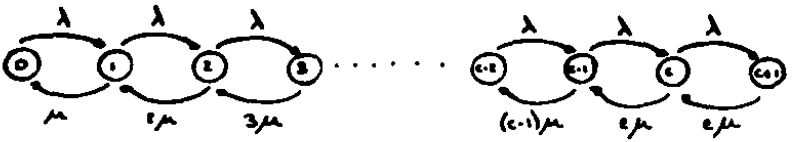
EL MODELO BASICO (TASAS DE LLEGADA Y DE SERVICIO CONSTANTES.

Es muy común que en un sistema de colas la tasa promedio de llegadas y la tasa promedio de servicio por servidor ocupado sean constantes. De ahí que el modelo básico tome este supuesto. Cuando el sistema tiene sólo un servidor (c=1), esto implica que los parámetros para el proceso de nacimiento y muerte son $\lambda_n = \lambda$ (n=0,1,2,.) y $\mu_n = \mu$ (n=0,1,2,..). Los diagramas de transición se muestran a continuación.

a) Un sólo servidor (c=1) $\lambda_n = \lambda$, $\mu_n = \mu$



b) Servidores múltiples (c>1) $\lambda_n = \lambda$, $\mu_n = \begin{cases} n\mu & n = 1, 2, \ldots, c \\ c\mu & n = c, c+1, c+2, \ldots \end{cases}$



Sin embargo cuando el sistema tiene servidores múltiples (c>1), la tasa μ_n no puede expresarse en forma tan sencilla. Manténgase en mente que μ_n representa la tasa promedio de servicio para todo el sistema (i.e. la tasa promedio con la cual ocurren las terminaciones de los servicios para que los clientes dejen el sistema) cuando existen n clientes en el sistema. Cuando la tasa promedio por servidor ocupado es μ, la tasa de servicio para todo el sistema con n servidores ocupados es $n\mu$. Por lo tanto $\mu_n = n\mu$ cuando $n \leq c$, mientras que $\mu_n = c\mu$ cuando $n \geq c\mu$ es decir cuando todos los servidores están ocupados.

Cuando la máxima tasa promedio de servicio ($c\mu$) excede la tasa promedio de llegadas (λ), esto es cuando

$$\rho = \frac{\lambda}{c\mu} < 1$$

El sistema alcanzará eventualmente una condición estable. En esta situación los resultados obtenidos para la condición estable se pueden aplicar directamente como se muestra a continuación.

Resultados para c=1

$$C_n = \left(\frac{\lambda}{\mu}\right)^n = \rho^n \qquad n = 1, 2, \ldots$$

Por lo tanto

$$P_n = \rho^n P_0 \qquad n = 1, 2, \ldots$$

donde

$$P_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \rho^n}$$

$$= \left(\sum_{n=1}^{\infty} \rho^n\right)^{-1}$$

$$= \left(\frac{1}{1-\rho}\right)^{-1}$$

$$= 1 - \rho$$

De ahí que,

$$P_n = (1-\rho)\rho^n \qquad n = 0, 1, 2, \ldots$$

Consecuentemente,

$$N = \sum_{n=0}^{\infty} n(1-\rho)\rho^n$$

$$= (1-\rho)\rho \sum_{n=0}^{\infty} \frac{d}{d\rho}(\rho^n)$$

$$= (1-\rho)\rho \frac{d}{d\rho}\left(\sum_{n=0}^{\infty} \rho^n\right)$$

$$= (1-\rho)\rho \frac{d}{d\rho}\left(\frac{1}{1-\rho}\right)$$

$$= \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}$$

En forma similar,

$$L = \sum_{n=1}^{\infty} (n-1)P_n$$

$$= N - 1(1-P_0)$$

$$= \frac{\lambda^2}{\mu(\mu-\lambda)}$$

Resultados para c>1.

$$C_n = \frac{(\lambda/\mu)^n}{n!} \qquad n = 1, 2, \ldots, c$$

$$C_n = \frac{(\lambda/\mu)^c}{c!}\left(\frac{\lambda}{c\mu}\right)^{n-c} = \frac{(\lambda/\mu)^n}{c!\,c^{n-c}} \qquad n = c, c+1, \ldots$$

Consecuentemente, si $\lambda < c\mu$, entonces

$$P_0 = 1/\left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!}\sum_{n=c}^{\infty}\left(\frac{\lambda}{c\mu}\right)^{n-c}\right]$$

$$= 1/\left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!}\frac{1}{1-(\lambda/c\mu)}\right]$$

y

$$P_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!} P_0 & 0 \le n \le c \\ \dfrac{(\lambda/\mu)^n}{c!\, c^{n-c}} P_0 & n \ge c \end{cases}$$

Usando la notación $\rho = \lambda/\mu c$

$$L = \sum_{n=c}^{\infty} (n-c) P_n$$

$$= \sum_{j=0}^{\infty} j P_{c+j}$$

$$= \sum_{j=0}^{\infty} j \frac{(\lambda/\mu)^c}{c!} \rho^j P_0$$

$$= P_0 \frac{(\lambda/\mu)^c}{c!} \rho \sum_{j=0}^{\infty} \frac{d}{d\rho}(\rho^j)$$

$$= P_0 \frac{(\lambda/\mu)^c}{c!} \rho \frac{d}{d\rho}\left(\sum_{j=0}^{\infty} \rho^j\right)$$

$$= P_0 \frac{(\lambda/\mu)^c}{c!} \rho \frac{d}{d\rho}\left(\frac{1}{1-\rho}\right)$$

$$= \frac{P_0 (\lambda/\mu)^c \rho}{c!\,(1-\rho)^2}$$

$$W = \frac{L}{\lambda}$$

$$R = W + \frac{1}{\mu}$$

$$N = \lambda\left(W + \frac{1}{\mu}\right)$$

$$= L + \frac{\lambda}{\mu}$$

Con el método descrito se deducen las fórmulas de los sistemas de colas.

Las fórmulas para el sistema M/M/1 son:

$U = \lambda / \mu$

$P(n) = (1 - U)U^{n}$ $U(1$

$N = U / (1 - U)$ No. promedio en sistema

$W = U / \mu(1 - U)$ Tiempo de espera

$L = U / (1 - U)$ Longitud promedio cola

$R = 1 / \mu(1 - U)$ Tiempo de respuesta

$P(\%) = R \text{ Log } (1-\%)U$ Precentil de espera

Las fórmulas para el sistema M/M/c son:

$U = \lambda / \mu$

$\rho = U/c$

$Po = c!(1 - \rho)C(c,U) / U^{c}$

$$P_n = \begin{cases} \dfrac{U^n}{n!} P_o & n = 0, 1, \ldots, c \\ \dfrac{U^n}{c! c^{n-c}} P_o & n > c \end{cases}$$

$$C(c,U) = (U^c/c!) \; / \; [(U^c/c!) + (1-\rho) \sum_{n=0}^{c-1} U^n/n!]$$

Todos los servidores ocupados.

$$N = L + U$$

$$W = C(c,U) \; / \; c\mu(1-\rho)$$

$$L = U\, C(c,U) \; / \; c(1-\rho)$$

Las fórmulas para el sistema de población finita de un modelo de cómputo interactivo son:

$E[t] = 1/\alpha$ es el tiempo promedio de pensamiento.

$$P_o = \left[\sum_{n=0}^{N} \frac{N!}{(N-n)!} \left(\frac{\alpha}{\mu} \right)^n \right]^{-1}$$

La utilización del CPU es:

$$\rho = 1 - P_o$$

El tiempo promedio de respuesta es:

$$R = \frac{N}{\mu(1-P_o)} - \frac{1}{\alpha}$$

Las salidas promedio son:

$$\lambda_t = \rho\mu = (1-p_0)\mu$$

Las fórmulas para el modelo del servidor central de multiprogramación son:

Después de calcular G(0), G(1),...,G(k) con el algoritmo de Buzen, las utilizaciones de los servidores son:

$$\rho_i = \begin{cases} G(K-1) \,/\, G(K) & i = 1 \\ \dfrac{\mu_1 \rho_1 P_i}{\mu_i} & i = 2,3,\ldots,M \end{cases}$$

Las salidas promedio están dadas por:

$$\lambda_i = \mu_i \rho_i P_i$$

El tiempo promedio de respuesta es:

$$R = \frac{N}{\lambda_t} - \frac{1}{\mu} = \frac{k}{\lambda_t}$$

COMENTARIO FINAL

En esta tésis se trató de hacer una exposición detallada de Teoría de Colas para posteriormente mostrar algunas aplicaciones que ésta puede tener en sistemas de cómputo.

Cabe recalcar la gran importancia que tiene el derivar medidas de desempeño utilizando los modelos de Teoría de Colas que, como se mostró en los ejemplos, proveen información vital para diseñar eficientemente sistemas de colas que alcancen un balance apropiado entre el costo de proveer un servicio y el costo asociado a esperar por ese servicio.

BIBLIOGRAFIA

Arnold O. Allen. Probability, Statistics and Queuing Theory. Academic Press. 1978.

Arnold O. Allen. Queuing Modelos of Computer Systems. IBM Systems Science Institute.

Arnold O. Allen. Capacity Planning for Management. IBM Systems Management Institute.

Paul J. Burke. The Output of a Queuing System. Operations Research 4 (1956).

Cox & Smith. Queues. London (1961).

Hillier, Lieberman. Operations Research. Second Edition. Holden Day.

M. Chandy, C. Saver. Computer Systems Performance Modeling. Prentice Hall 1981.

Philip M. Morse, H.N. Garber. A Family of Queuing Problems. Operations Research 2 (1954).

Philip M. Morse. Queues, Inventories & Maintenance. Wiley, 1958.

Taha. Operations Research. Collier Macmillan. 1976.

Thomas L. Saaty. Resume of Useful Formulas in Queuing Theory. Operations Research 5 (1957).

W.R. Van Voorhis. Waiting Line Theory as a Management Tool.