Capítulo 6

Teoría de Colas

Capítulo 6

Teoría de Colas

Los objetivos de este capítulo son los siguientes:

- Conocer la clasificación de los sistemas de colas[1], sus características y disciplina de manejo.
- Comprender la teoría de colas y la importancia de su correcta aplicación en el campo de la simulación.
- Familiarizarse con las técnicas matemáticas aplicadas en la resolución de los sistemas de colas.

6.1 Introducción

Cuando la demanda de un servicio es demasiado grande para la capacidad de prestación del servicio se forman las colas o filas de espera, un problema muy común en la vida diaria.

Es, precisamente, este problema el que da origen a la teoría de colas la cual fue presentada en 1909 por Agnar K. Erlang quien publicó un trabajo acerca de la congestión en el tráfico telefónico.

En este capítulo veremos los principales aspectos de la Teoría de colas y su relación con la simulación. Expondremos en forma general la estructura básica de los sistemas de colas, la clasificación o disciplinas de manejo, las técnicas matemáticas aplicadas en su resolución, la simulación y aplicación de las teorías de colas en la solución de sistemas reales.

6.2 Descripción General

Las filas de espera o colas las encontramos diariamente en problemas relacionados con el transporte, diseño de sistemas, comunicaciones, trabajos en la computadora, así como, en los supermercados, bancos, gasolineras, etc.

Las filas de espera se originan porque no se puede atender simultáneamente a todos los clientes debido a que no hay suficientes servidores. Además, no es económicamente factible tener demasiados servidores. Los clientes pueden llegar al azar, pero esto no garantiza que no se formen filas, así es que debe existir un número de servidores que permita que las filas de espera no sean demasiados largas.

Las líneas de espera son tan comunes en la vida real que nos sorprendería que la mayoría de los problemas encontrados en el modelaje o la simulación de una operación involucre colas.

La teoría de colas es un estudio matemático que permite aislar factores tales como:

---

[1] Se conoce también como *filas* o *líneas de espera*.

- Promedio de longitud de las líneas de espera.
- Promedio de tiempo que un elemento, persona, máquina, suceso debe esperar en la cola antes de ser atendido.
- Número de elementos que se calcula del sistema total.
- Tiempo que se calcula que un elemento esté afectando al sistema.

Esta teoría la podemos dividir a su vez en dos estudios matemáticos. Uno trata de distribuciones de tipo específico, de la cual se derivan fórmulas matemáticas, como lo son la distribución exponencial y de Poisson. El otro trata de distribución clásicas empíricas[2] que se analizan mediante métodos de simulación.

La teoría de colas tiene importancia y debe estudiarse porque las filas de espera desorganizan muchos aspectos de la vida. Un problema básico de formación de colas tiene tres partes diferenciadas:

- Una fuente de clientes.
- La fila de espera.
- La instalación de servicio (puede tener 1 o más servidores).

Las teorías de la formación de colas encontraron aplicaciones para los procesos en que las llegadas se producían en desorden y el servicio era limitado; pero la complejidad matemática dificultó sus esfuerzos. Sin embargo, hay un método más sencillo para resolver los problemas y se conoce con el nombre de Técnica de Montecarlo, el cual veremos en detalle en el capítulo 7.

6.3 Terminología y Características de las Colas

En esta sección se presenta la estructura de un sistema de colas, así como las definiciones de los componentes del sistema y las variables que se utilizarán en los diferentes modelos. Esta sección es importante porque enfatiza el hecho de que se cuenta con estándares para poder entender los diferentes modelos de colas.

6.3.1 Estructura Básica de los Modelos de Cola

El proceso básico que utilizan los modelos de cola es el siguiente:

- Los clientes que pueden solicitar un servicio forman una "fuente de entrada" (conocida como población).
- Los clientes entran al sistema y se unen a una cola o fila de espera mientras esperan ser atendidos.
- Se selecciona en diversos momentos, mediante una regla conocida como disciplina de servicio[3], un cliente para que reciba el servicio.
- El cliente es servido por el mecanismo de servicio.

---

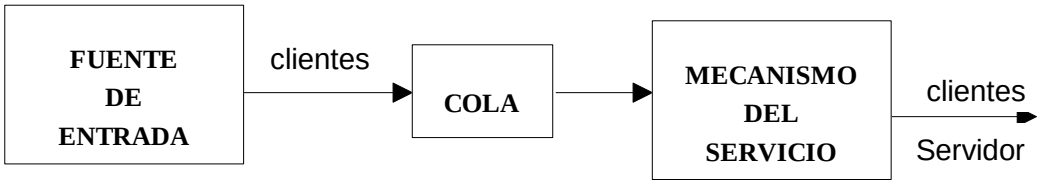
[2] Datos de una muestra.

[3] Este tema se trata más adelante en este capítulo.

- Luego de ser atendido el cliente sale del sistema.

Veamos con un esquema el proceso en la figura 6.1.

Figura 6.1:



Sistema de Colas

6.3.2 Definiciones

A continuación, se definen los términos más comunes dentro de un sistema cliente servidor:

Fuente de entrada: Está formada por clientes que pueden ser personas, partes, procesos, trabajos, máquinas, etc. La fuente puede ser finita o infinita. Se considera finita si está formada por un número pequeño y contable de clientes y en este caso todos los sucesos son dependientes. Se considera infinita si está formada por un número ilimitado de clientes que exigen un servicio y en este caso todos los sucesos son independientes de otros. Los clientes tendrán una tasa de arribo que es la tasa a la cual llegan los clientes para ser atendidos.

Patrón de arribo: Se refiere a la distribución de arribo (llegada) de los clientes, si a los clientes se les permite ir sin recibir el servicio y si los clientes llegan solos o en grupos. Un patrón de arribo es aquel que es determinístico o carece de toda incertidumbre. Generalmente los patrones incluyen alguna incertidumbre.

Cola o Fila de Espera: Es el número de clientes que espera recibir el servicio. La teoría de colas nos proporciona el número promedio de clientes que espera, el tiempo promedio de espera, y otros factores que veremos más adelante.

Proceso de Servicio: Considera factores como la distribución del tiempo requerido, si los clientes son atendidos solos o en grupos, si el nivel de servicio permanece constante o cambia a medida que se forma la fila. El servidor puede cambiar el servicio dependiendo del largo de la línea de espera que se está formando o de la demanda requerida.

Disciplina en el Servicio: Se refiere a la distancia en la cola. Es la técnica por la cual los clientes son seleccionados de la cola para ser atendidos. El más común es el FIFO el cual supone que el primero en llegar es el primero en ser atendido. Otras disciplinas incluyen LIFO (el último en llegar es el primero en salir), SIRO (servicio en orden aleatorio) y PRI (prioridad).

Mecanismo del Servicio: Consiste en uno o más medios de servicio. Si existe más de un medio, el cliente puede recibir el servicio por uno de los diferentes medios de servicio

(canales[4] paralelos) o por una sucesión de estos medios (canales en serie). Estos conceptos los veremos más adelante. El tiempo o duración del servicio es el tiempo que transcurre para un cliente, desde que se inicia el servicio hasta que se termina en uno de los medios. Comúnmente se supone la misma distribución de probabilidad de los tiempos de servicio para todos los servidores. Cada servidor posee su propio tiempo promedio de servicio, aunque a veces se asume el mismo valor para todos.

Capacidad del Sistema: En la mayoría de los sistemas la capacidad es finita y causa un efecto pronunciado en la operación del sistema. Por consiguiente, al modelar se debe tomar en cuenta la capacidad máxima del sistema. Por ejemplo, en una barbería, el largo de la cola estará limitado por el número máximo de sillas de espera disponibles.

Tasa de Servicio[5]: Cantidad de clientes por unidad de tiempo que puede atender un servidor.

Costo de Espera. Esperar significa desperdicio de algún recurso activo que bien se puede aprovechar en otra cosa y está dado por:

Costo total de espera = $\mathbf{C_w L}$

Donde $\mathbf{C_w}$ = costo de espera por hora (en dólares) por llegada por unidad de tiempo y **L** = longitud promedio de la línea.

Costo de Servicio. Este en la mayoría se trata de comprar varias instalaciones de servicio, en estos casos solo se ocupan los costos comparativos o diferenciales.

Sistema de costo mínimo. Aquí hay que tomar en cuenta que, para tasas bajas de servicio, se experimenta largas colas y costos de espera muy altos. Conforme aumenta el servicio, disminuyen los costos de espera, pero aumenta el costo de servicio y el costo total disminuye, sin embargo, finalmente se llega a un punto de disminución en el rendimiento. Entonces el propósito es encontrar el balance adecuado para que el costo total sea el mínimo.

6.3.3 Variables

En esta sección se presentan las variables más utilizadas en los diferentes modelos de colas que se presentan más adelante[6]:

λ: Tasa de llegada de clientes. Rata de arribo de clientes. Promedio de llegadas de clientes en una unidad de tiempo.

---

[4] Canal es sinónimo de servidor.

[5] Rata de servicio.

[6] *Sistema* se refiere al tiempo que está un cliente en cola y en el servidor.

μ: Tasa de servicio a los clientes. Rata de servicio a los clientes. Promedio de servicios prestados a clientes en una unidad de tiempo por canal.
L: Número esperado o estimado[7] de clientes en el sistema. Conocida también como L_s.
L_q: Número esperado o estimado de clientes que están en la cola.
W: Tiempo esperado o estimado en el sistema. Conocida también como W_s.
W_q: Tiempo esperado o estimado en la cola.
P_0: Probabilidad de que no existan clientes en el sistema.
P_n: Probabilidad de tener n clientes en el sistema (n clientes en el servidor).
ρ: Utilización del sistema o factor de utilización para el sistema. Probabilidad de que la facilidad de servicio esté siendo utilizada. Es igual a $1 - P_0$.
C: Número de canales o servidores.
K: Tamaño de la población finita.

Para poder utilizar estas variables en cualquier modelo de colas, es obligatorio que λ y μ tengan la misma unidad de tiempo. También es importante denotar que:

Tiempo promedio en el sistema	=	Tiempo promedio en la cola	+	Tiempo promedio en el servidor

6.3.4 Medidas de rendimiento para evaluar un sistema de colas

Existen muchas medidas de rendimiento diferentes que se utilizan para evaluar un sistema de colas en estado estable.

Para diseñar y poner en operación un sistema de colas, por lo general, los administradores se preocupan por el nivel de servicio que recibe un cliente, así como el uso apropiado de las instalaciones de servicio de la empresa. Algunas de las medidas que se utilizan para evaluar el rendimiento surgen de hacerse las siguientes preguntas:

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

  ¿Cuál es el tiempo promedio que un cliente recién llegado tiene que esperar en la fila antes de ser atendido? La medida de rendimiento asociada es el tiempo promedio de espera, representado por Wq.

  ¿Cuál es el tiempo promedio que un cliente invierte en el sistema entero, incluyendo el tiempo de espera y de servicio? La medida de rendimiento asociada es el tiempo promedio en el sistema, representado por Ws.

- Preguntas cualitativas pertenecientes al número de clientes, como:

  En promedio, ¿Cuántos clientes están esperando en la cola para ser atendidos? La medida de rendimiento asociada es la longitud media de la cola, representado por Lq.

  ¿Cuál es el número promedio de clientes en el sistema? La medida de rendimiento asociada es el número medio en el sistema, representado por Ls.

---

[7] Promedio.

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

  ¿Cuál es la probabilidad de que un cliente que llegue tenga que esperar a ser atendido? La medida de rendimiento asociada es la probabilidad de bloqueo, representada por pw.

  En cualquier tiempo particular, ¿cuál es la probabilidad de que un servidor esté ocupado? La medida de rendimiento asociada es la utilización, denotada con ρ. Esta medida indica también la fracción de tiempo que un servidor está ocupado.

  ¿Cuál es la probabilidad de que existan n clientes en el sistema? La medida de rendimiento asociada se obtiene calculando la probabilidad P_0 de que no haya clientes en el sistema, la probabilidad P_1 de que haya un cliente en el sistema, y así sucesivamente. Esto tiene como resultado la distribución de probabilidades de estado, representada por P_n, n = 0,1...

  Si el espacio de espera es finito, ¿Cuál es la probabilidad que la cola esté llena y que un cliente que llegue no sea atendido? La medida de rendimiento asociada es la probabilidad de negación de servicio, representada por pd.

- Preguntas relacionadas con los costos, como:

  ¿Cuál es el costo promedio por unidad de tiempo para operar el sistema?

  ¿Cuántas estaciones de trabajo se necesitan para lograr la mayor efectividad de costos?

6.4 Clasificación de las Colas

Como hemos visto anteriormente, una línea de espera o cola está constituida básicamente por clientes que esperan ser atendidos por un servidor(es).

Las líneas de espera las podemos clasificar de acuerdo con:

- La fuente de entrada que genera los clientes que requieren de un servicio, la cual puede ser finita o infinita.
- Los números de clientes que esperan en la cola, los cuales pueden ser finitos o infinitos.
- La forma como esperan los clientes ya sea en una o varias colas o con opción a cambiarse o no de cola.
- El tiempo transcurrido entre la llegada de un cliente y el inmediatamente anterior. Este tiempo se conoce como *tiempo entre arribo*. Este lapso de clientes puede ser una constante o una variable aleatoria independiente cuyo comportamiento puede o no conocerse.
- El tiempo de servicio, el cual es un intervalo de tiempo que puede ser una constante o variable aleatoria dependiente o independiente cuya distribución de probabilidad puede o no conocerse. Este tiempo es dependiente cuando varía por factores de presión del sistema (por ejemplo, las quejas de la gente que espera) y es

independiente cuando la duración del servicio no se afecta por este tipo de presiones.

- La disciplina de la cola que depende de la política de servicio que se utilice. Puede ser FIFO, LIFO, SIRO, etc.[8].
- El número de servidores o canales de servicio.
- La estructura de las estaciones de servicio las cuales pueden estar en serie, paralelas o mixtas.
- La estabilidad del sistema que puede ser transitoria o estable. El primero se refiere al estado inicial del sistema de tiempo que ha transcurrido desde su inicio. El segundo se refiere al estado del sistema después que ha pasado bastante tiempo, el cual se vuelve independiente del estado inicial.

Todos los ejemplos de líneas de espera que se han presentado incluyen personas, pero éste no siempre es el caso. Las llegadas pueden ser cartas, carros, incendios, ensambles intermedios en una fábrica, etc. En la tabla 1 se muestran ejemplos de varios sistemas de colas.

[8] Este tema se cubre en la sección 6.5

Tabla 1: Ejemplos de Sistemas de Cola

Situación	Llegadas	Cola	Mecanismo de servicio
Aeropuerto	Aviones	Aviones en carreteo	Pista
Aeropuerto	Pasajeros	Sala de espera	Sala de espera
Departamento de bomberos	Alarmas de incendio	Incendios	Departamento de bomberos
Compañía de Teléfonos	Números	Llamadas	Conmutador
Lavado de autos	Autos	Autos sucios	Mecanismo de lavado
La corte	Casos	Casos atrasados	Juez
Carga de Camiones	Camiones	Camiones en espera	Muelle de carga
Oficina de correos	Cartas	Buzón	Empleados de correo
Fotocopias	Pedidos de copias	Trabajos	Copiadoras
Hospital	Pacientes	Personas enfermas	Hospital

Nótese que en cada situación sólo fluye un tipo de artículo a través del sistema.

Dicho de otra manera, las llegadas son homogéneas o vienen de la misma población. Esta es una limitación importante de la teoría de colas. Cuando una instalación de servicio, como un aeropuerto, maneja diferentes tipos de llegadas, éstas se deben tratar por separado.

Se reconocen diferencias en la estructura de los sistemas mostrados en la tabla 1.

Por ejemplo, los bancos casi siempre tienen más de un cajero, cada uno con una línea de espera separada. Con frecuencia los aeropuertos tienen más de una pista de aterrizaje. La oficina postal maneja el correo en base a prioridades.

Permitiendo que varíen el número de colas y el número de servidores, pueden hacerse los diagramas de los cuatro tipos de sistemas.

De acuerdo con la clasificación, veremos algunos ejemplos de tipos de sistemas de colas que se presentan en la vida.

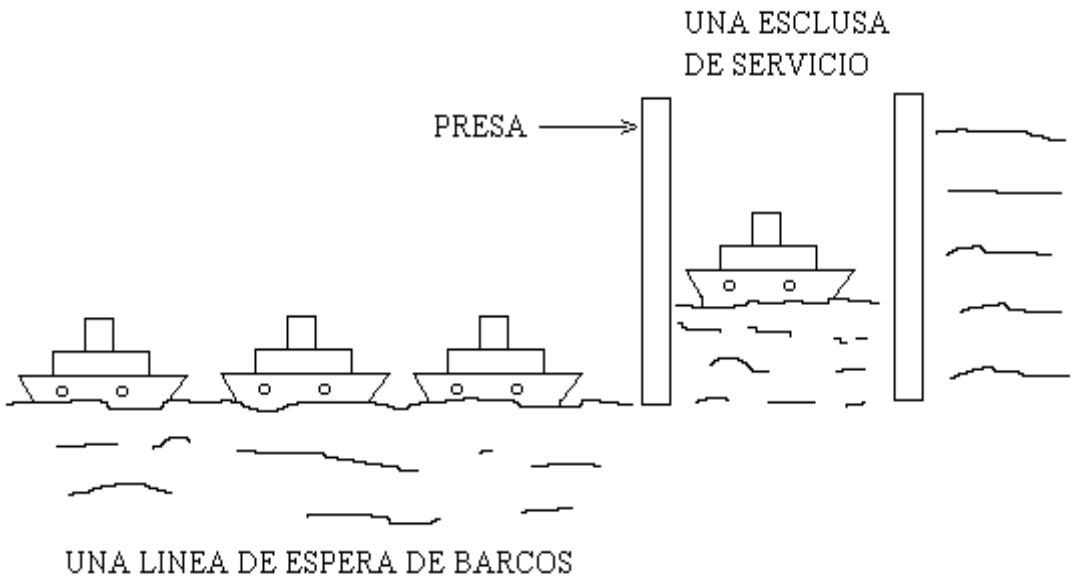
6.4.1 Una Cola - un Servidor

Los clientes esperan ser atendidos de acuerdo con un orden riguroso de llegada. Al finalizar el servicio a un cliente, se le brinda al que sigue en la cola.

Ejemplos:

- La taquilla de un cine en donde se venden boletos de acuerdo con la llegada de los espectadores.
- Los barcos que esperan ser admitidos para cruce de una represa de río. En este caso los barcos que esperan ser atendidos en orden de llegada representan la cola y la esclusa única utilizada para elevar o bajar los barcos representa al servidor, también existe la disciplina de darle servicio al primero en llegar, el cual será el primero en salir.

Figura 6.2 Una Esclusa del Canal de Panamá: Una Cola – Un Servidor



6.4.2 Una Cola – Múltiples Servidores en Paralelo

Existen varias estaciones de servicio para una sola fila de clientes. Al terminar de brindar el servicio independientemente de cuál sea el servidor, el cliente que ocupe la primera posición en la fila pasará a ese servidor.

Ejemplos:

- Una fila en la sección de embutidos y quesos de un supermercado. Los clientes, a medida que llegan, van tomando un tiquete enumerado y esperan ser atendidos por cualquiera de los servidores.

- Una peluquería con cinco sillones, donde se atiende a los clientes por orden de llegada. Los clientes representan la fila y los cinco sillones con los peluqueros representan el servicio múltiple.

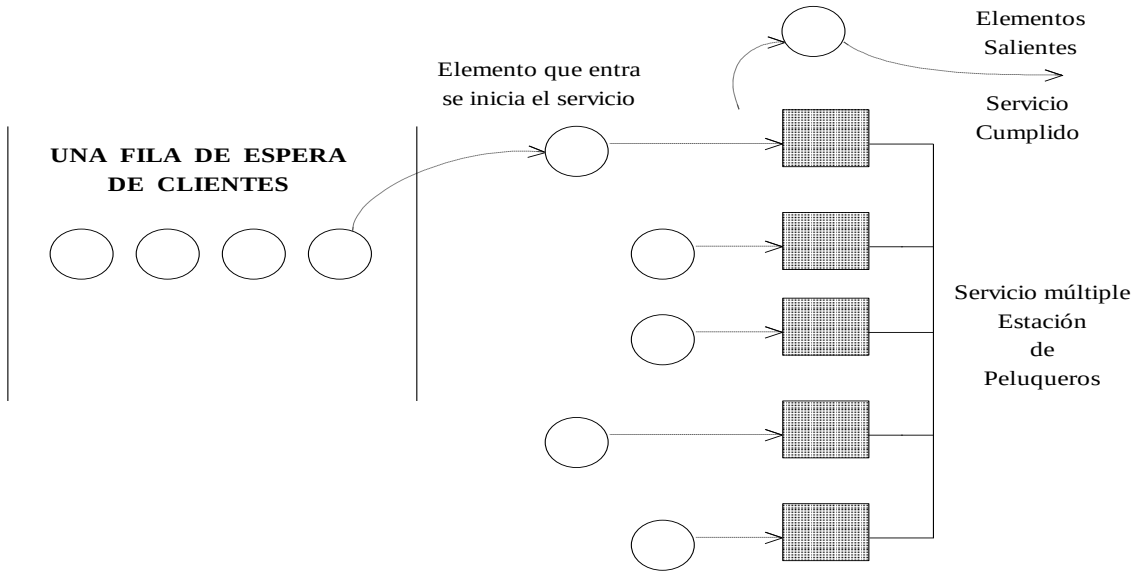


Figura 6.3 Una Cola – Múltiples Servidores

6.4.3 Una Cola - Múltiples Servidores en Serie

Existen varias estaciones de servicio, pero una sola fila de clientes y cada cliente pasa por cada uno de los servidores en serie.

Ejemplo:

- En una embotelladora, los servicios serían una esterilizadora, una máquina de llenado, una encorchadora, una etiquetadora y una empaquetadora. Las botellas van en fila y cada una recibe el servicio múltiple de esterilizar, llenar con líquido, encorchar, etiquetar y luego empaquetar.

6.4.4 Múltiples Colas – Múltiples Servidores en Paralelo sin Cambio de Colas

Existen varias filas y varios servidores en paralelo los cuales atienden clientes específicos. En el momento que un cliente se coloca en una cola, él no se puede cambiar.

Ejemplos:

- El establecimiento de pagos de cheques a jubilados. Aquí hay varias colas y varios servidores. Para agilizar el proceso, cada servidor atiende números específicos de seguro social, por ejemplo, en la fila 1 se atiende números de S.S. del 01 al 10 y, en la fila 2 los números de S.S. del 11 al 20 y así sucesivamente. Todos deben atenderse en sus filas respectivas, pues allí es donde recibirán su cheque.

- En la oficina de pago de préstamos y becas a los beneficiarios donde existen cuatro ventanillas de servicio de acuerdo con la inicial del apellido paterno.

6.4.5 Múltiples Colas – Múltiples Servidores en Paralelo con Opción a Cambiar de Cola

Existen varios servidores y varias filas donde se permite cambio de fila.

Ejemplos:

- En un banco donde hay cierto número de cajas y los clientes se forman en la cola que más le convenga, con opción a cambiarse de cola.
- Un hospital donde hay tres ascensores. Las personas se agrupan en torno a cada uno de ellos con la esperanza de ser trasladados a su destino en menor tiempo. Si el ascensor número 3 llega primero, algunos de las que ocupan puestos más periféricos en los grupos 1 y 2 se colocaran en el 3, confiando que esto los ayudará a trasladarse más rápido.

6.5 Notación Kendall

Existe un código de clasificación para los diferentes tipos de líneas de espera desarrollado por Kendall. La forma general de la notación es la siguiente[9]:

A/B/C/D/E/F

Los símbolos representan las características de los sistemas de colas donde:

A: Tipo de distribución aplicada para los patrones de arribo (tiempo de arribo).
B: Distribución de los tiempos de servicio o atención.
C: Número de servidores.
D: Disciplina en la cola.
E: Capacidad del sistema[10].
F: Población.

A, B, C, D, E pueden tomar cualquiera de los siguientes valores:

Valores para A:

GI: Corresponde a una distribución de arribos independientes con tiempo promedio entre arribos.
D: Corresponde a una entrada determinística con tiempo promedio constante entre arribos.
M: Corresponde a una distribución Random o de Poisson en las cuales el tiempo de la siguiente llegada (arribo) es independiente de la llegada

[9] La literatura presenta diferentes formatos de esta notación.

[10] Esta opción se descarta en muchos casos y no se utilizará en los modelos que se presentan a continuación.

anterior. Existe un número promedio de unidades de llegada que requieren el servicio por unidad de tiempo.

E_k: Corresponde a una distribución tipo Erlang en donde los datos se agrupan más estrechamente alrededor de la media.

Valores para B:

G: Corresponde a una distribución general de los tiempos de servicio.
D: Corresponde a un tipo de distribución determinístico.
M: Utiliza una distribución exponencial para generar los tiempos de servicio.
E_k: Ce compone de k número de tareas donde cada una tiene un servicio exponencial idéntico.

Valores para C:

Toma el valor del número de unidades de servicio, es decir, el número de facilidades que ofrece el servicio.

Valores para D:

Puede aplicarse cualquiera de las distintas disciplinas LIFO, FIFO, PRI, SIRO o GD.

Valores para E:

Toma el valor de números de clientes admisibles en el sistema. Si es una constante se pone k y si no tiene restricciones de capacidad se pone ∞.

Valores de F:

Es el número total de clientes que pueden requerir servicio en determinado momento.

6.6 Disciplinas de Colas

La disciplina de la cola consiste en determinar cómo se selecciona el siguiente cliente de una cola para proceder a servirlo. Las disciplinas más comunes son las siguientes:

- **FIFO**: Es una disciplina de servicio en la que el primero que entra es el primero en salir (first in, first out). También se abrevia PEPS en español. Ocurre cuando los clientes que llegan se reúnen en el tiempo en que llegan, y el servicio se ofrece enseguida a la entidad que ha esperado el máximo tiempo. Por ejemplo, esta disciplina es adoptada en los diferentes lugares en donde para ser atendidos se toma un cupo, así, el primero en llegar es el primero en salir.

- **LIFO**: Es una disciplina de servicio en la que el último en llegar es el primero en salir (last in, first out). También se abrevia UEPS en español. Ocurre cuando los clientes

forman una cola en el orden en que llegan, pero se ofrece el servicio al que llegó más recientemente en donde las personas que entraron de último son los primeros en salir. Por ejemplo, los platos limpios de un restaurante de auto servicio rápido son colocados en una pila y el último en la pila es el primero en tomarse para ser utilizado.

- **SIRO**: Es una disciplina aleatoria (service in random order) en donde se hace una selección entre todas los clientes que esperan en el momento que se ofrece el servicio. A no ser que se especifique de otra manera, el término aleatorio implica que todos los clientes que esperan tienen igual oportunidad de ser seleccionados. Por ejemplo, en los sorteos de la Lotería, las balotas están dentro del ánfora y luego de haber girado, son escogidas al azar con igual probabilidad.

- **PRI**: Es una disciplina de prioridad que ocurre cuando un cliente tiene derecho a ser atendido antes que otros clientes, los cuales tienen un nivel de prioridad menor. En el caso que haya clientes con disciplinas de “primeros que entran son los primeros que salen”, la prioridad es un atributo del cliente y depende de la cola y no del servicio. Por ejemplo, los trabajos en una computadora que esperan en la cola de trabajo para ser ejecutados, algunos tienen prioridad mayor y son atendidos primeros que otros.

Existen dos clases de disciplinas de prioridad que son:

1. **Prioridad No Asegurada**: El cliente que está recibiendo el servicio no puede ser desplazado por un cliente de mayor prioridad que entre al sistema. Si el servidor queda libre, entonces se selecciona al cliente con mayor prioridad.

2. **Prioridad Asegurada**: Ocurre cuando un cliente de prioridad inferior que está recibiendo el servicio es desplazado (sacado del servidor) siempre que entra al sistema un cliente de prioridad superior. Por consiguiente, se libera un servidor para empezar a dar servicio inmediatamente a una nueva llegada. Cuando un servidor tiene éxito en terminar un servicio, el nuevo cliente, a quien empezará a atender, se selecciona de igual forma como se describió anteriormente, de modo que un cliente desplazado, normalmente regresará una vez más a recibir el servicio, y después de un número suficiente de intentos, terminará de recibirlo.

- **Disciplina de Retiro de la Cola**: Ocurre cuando los clientes abandonan la cola y se debe especificar las reglas para retirarse. El retiro depende de la longitud de la cola o del tiempo en que ha esperado el cliente. Frecuentemente se da una función de probabilidad para determinar el punto en que se retira. Por ejemplo, una persona se cansa de esperar y se sale de la cola.

- **Disciplina de Sondeo:** Ocurre cuando se forma más de una cola esperando el mismo servicio, es decir, se comparte el servicio entre dos colas. Al especificar esta disciplina se da el orden en que se sirven las colas, la cantidad de clientes servidos en cada ofrecimiento del servicio y el tiempo en la transferencia de servicio entre colas. Por ejemplo, una computadora que escudriña una cantidad de terminales de entrada para detectar la presencia de mensajes de entrada.

6.7 Clasificación de los Sistemas de Colas

Antes de entrar a los distintos sistemas de colas tenemos que hacer notar la importancia de las probabilidades[11] en la simulación.

Para modelar un sistema de colas, se hace necesario dar ciertas funciones probabilísticas a dos características de las colas que son:

- Patrón de arribo.
- Tiempo de servicio.

Los patrones de arribo se refieren al intervalo entre arribo sucesivos, y la probabilidad interviene si los arribos varían de una forma estocástica, definiéndose una función de los tiempos entre arribos. En esta intervienen el tiempo medio entre arribos y la taza de arribos media. Según sea el caso, se tiene distintos tipos de distribución. El patrón de arribo más utilizado es el de Poisson. Este nos indica que el tiempo entre arribos está distribuido exponencialmente. La teoría de colas permite la resolución de distintos casos por medio de la suposición de que un tiempo de arribo es independiente del arribo anterior.

Los tiempos de servicio deben describirse también mediante una función de probabilidades para aquellos casos que varíen estocásticamente. Si se considera el tiempo de servicio completamente aleatorio, se puede representar mediante distribuciones Exponenciales, Erlang o Hiperexponencial.

6.7.1 Sistema de Canal Simple

El sistema de colas de Poisson para un canal simple se define así: Existen clientes que llegan a una instalación para pedir servicio y si este está vacío, un cliente entra y recibe servicio, si no, el cliente que llega debe tomar su lugar al final de la fila y esperar su turno para ser atendido, utilizando, por ejemplo, la disciplina de que el primero que llega es el primero en salir.

En este sistema la distribución del tiempo entre arribos sucesivos es exponencial como la distribución de tiempo para servir a una unidad, es decir, si el tiempo entre eventos sucesivos se distribuye exponencialmente, la distribución del número de eventos que tienen lugar en cualquier intervalo de tiempo sigue una distribución de Poisson.

Un sistema de colas de canal simple se representa en la vida diaria al ir al cine, a un lava-autos, etc.

Para construir un modelo de este sistema simple, se deben considerar los siguientes puntos:

- El tiempo promedio que se pasa una unidad en el sistema y la distribución de frecuencia de ese tiempo.
- La distribución promedio de las instalaciones de servicio.

---

[11] Por eso, este libro presenta un resumen de los temas más importantes de las probabilidades.

- El número promedio de tiempo que se pasa una unidad en la cola y su distribución de frecuencias.

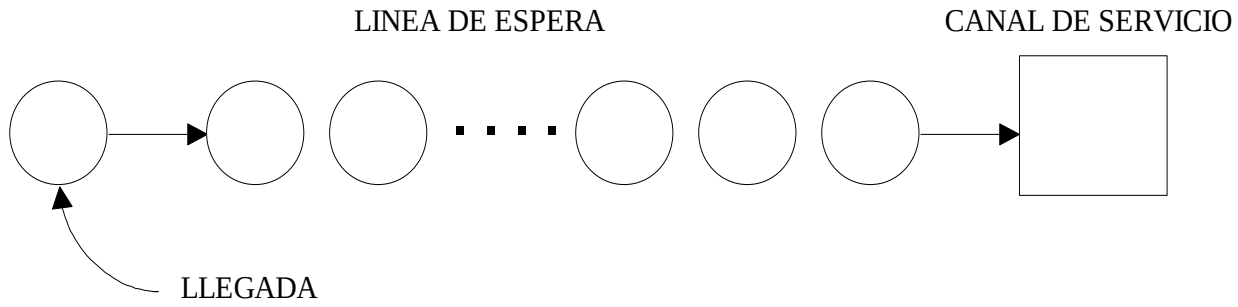


Figura 6.4 Sistema de Clientes Servidor

El objetivo del programa de simulación es proporcionar información respecto al funcionamiento de un sistema de colas de Poisson de canal simple. En este sistema, las unidades llegan en forma aleatoria a una instalación de servicio y esto es lo que hay que lograr que suceda en la computadora. El sistema debe servir a esa unidad por orden de llegada y es posible que algunas unidades tengan que esperar en la cola hasta que llegue el turno para recibir el servicio.

En este sistema las instalaciones de servicio pueden tener dos estados:

- Ocupadas, es decir, el servicio a la siguiente unidad en espera comienza inmediatamente.
- Inactivas, es decir, ninguna unidad espera para recibir servicio.

Si una unidad entra al sistema para recibir servicio, también estará en uno de los estados posibles: se encuentra en línea de espera o recibe servicio.

Las condiciones que pueden existir en el sistema (los estados en la unidad y en las instalaciones de servicio) y los resultados para la unidad cuando existen esas condiciones deben incluirse en el modelo. Para observar los eventos y resultados en un tiempo simulado se adopta la idea de relojes para vigilar lo que debe ocurrir en el modelo. En la mayoría de los sistemas el tiempo se considera como una variable aleatoria.

Un modelo de simulación por computadora ofrece un método único para estudiar y analizar cualquier estudio de simulación, de si se debe hacer o no un análisis más profundo, después de obtener la salida inicial.

Una vez construido el modelo de simulación en una computadora será una manifestación del sistema dado y se podrá usar para un análisis más profundo de ese sistema, según lo exijan los objetivos. Por esto, una vez construido el modelo de canal simple, se podrá usar para

estudiar sistemas proyectados o propuestos en los que se determinan los procesos de tiempo de servicio y arribo mediante algún método menos matemático.

6.7.2 Sistema de Canales en Serie

El sistema de colas de canales múltiples en serie es aquel en el que cada unidad que llega al sistema debe pasar por cada canal de servicio.

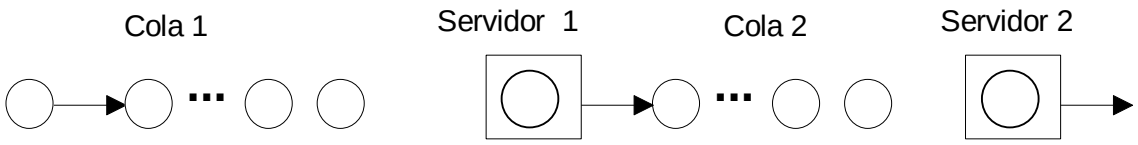


Figura 6.5 Sistema de Canales en Serie

El sistema incluye "n" canales de servicio o instalaciones y una unidad ante cada canal. La distribución de tiempo entre arribos para unidades que entran en el sistema tiene una media y se supone que el tiempo de servicio para cada canal también tiene una media.

Una unidad que intente entrar en la cola inicial cuando esté llena, será eliminada del sistema. Una unidad saldrá de cualquier canal de servicio sólo cuando la cola siguiente no esté llena. Si la cola que sigue está llena, la unidad permanecerá en el canal de servicio hasta que pueda entrar en la siguiente cola, impidiendo de esta forma que el canal dé servicio.

Si se supone que en todas las colas hay una disciplina de servicio por orden de arribo, se puede observar que el proceso de arribo de las unidades a cada uno de los canales de dos a "n", está regido no sólo por el proceso de arribo a la instalación inicial, sino también por la distribución del tiempo de servicio a cada una de las instalaciones anteriores.

6.7.3 Sistema de Canales en Paralelo

Nuestros sistemas físicos de instalación de servicios se manifiestan como un sistema de colas, con canales múltiples de servicio en paralelo. En este sistema las unidades llegan en forma aleatoria y toman su lugar de la línea de espera, cuando no pueden obtener servicio inmediatamente.

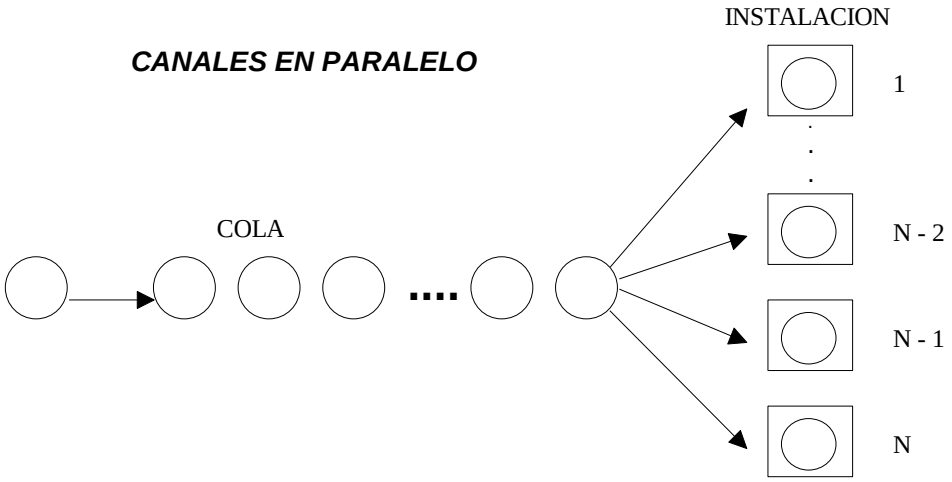


Figura 6.6 Sistema de Canales en Paralelo

La primera unidad en la línea de espera entra en el primer canal de servicio que se encuentra disponible. Una unidad que intente entrar en la cola cuando esté llena, tendrá que salir del sistema. Desde el punto de vista de las instalaciones de servicio el sistema de colas de canal simple es casi idéntico, la única diferencia es que cada canal en paralelo debe competir para obtener unidades de la cola. Cuando la cola está vacía y hay por lo menos un canal inactivo, entrará una unidad en el canal de servicio inmediatamente, en cuanto ingrese al sistema, si existen otras condiciones, la unidad abandonará el sistema o entrará en la cola para esperar el servicio. Si se encuentra disponible más de un canal cuando la unidad se dispone a entrar al servicio, irá al primer canal disponible. Esto implica que no hay una relación de preferencia entre los canales y que la elección es suficientemente aleatoria. Pueden existir preferencias basadas en la rapidez del servicio, es decir, que la unidad irá al canal con el menor tiempo esperado de servicio.

6.8 Modelos Matemáticos de Sistemas de Colas

Muchos de los sistemas han sido modelados exitosamente por un modelo de colas en donde las distribuciones de entre arribo y de servicio son distribuidas exponencialmente. La distribución de entre arribo exponencial implica que el proceso de arribo es Poisson. Para analizar el modelo revisaremos las propiedades de las distribuciones Exponencial y de Poisson.

1. La probabilidad de que un cliente llegue al sistema en un intervalo $\Delta t = \lambda \Delta t + \theta(\Delta t)$ donde:
   λ = valor medio de la distribución
   $\theta(\Delta t)$ se asume que es cero
2. La probabilidad de que lleguen dos o más clientes al sistema en $\Delta t = \theta(\Delta t) \approx 0$, de manera que siempre suponemos que en Δt llega uno o cero clientes.

3. El número de arribos en un intervalo es independiente del número de arribos en un intervalo anterior.

4. La probabilidad del tiempo de entre arribo es igual a la probabilidad de que no haya ningún arribo durante t minutos o segundos:

$$P(T > t) = P(\text{cero arribos en tiempo t}) = e^{-\lambda t}$$

donde λ = Cantidad de clientes por hora
T = Tiempos de entre arribos sucesivos.

5. La función de distribución de la forma exponencial:

$$F(t) + P(T \leq t) = 1 - P(T > t) = 1 - e^{-\lambda t}$$

es una forma de demostrar que si los arribos tienen distribución de Poisson entonces el tiempo de entre arribo tiene distribución exponencial negativa.

6. La distribución de Poisson también posee una propiedad que es de gran valor en el análisis de muchos sistemas de colas: La agregación y disgregación. Esta nos dice que si tenemos un proceso con varias fuentes de entrada ($\lambda_1, \lambda_2, \lambda_3$) se pueden reunir en una sola variable la cual sería la razón de llegada al sistema.

Esta distribución sería:

$$\lambda = \lambda_1 + \lambda_2 + \lambda_3 + \cdots\cdots + \lambda_n$$

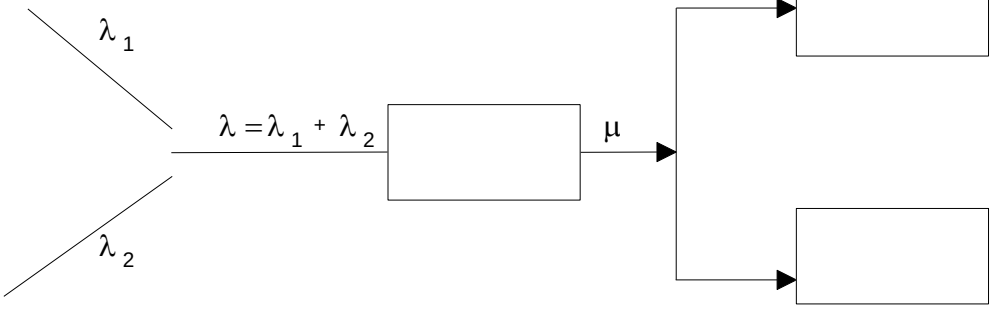


Figura 6.7

Distribución de Poisson

Después de haber mencionado algunas de las propiedades de la distribución exponencial de Poisson, estaremos listos para analizar los modelos de colas. Asumiremos que las entradas al sistema de colas siguen una distribución de Poisson. También asumiremos que el arribo ocurre durante el intervalo (0, t). Por consiguiente, el momento exacto del arribo sigue una distribución uniforme, esto significa que el arribo ocurre aleatoriamente en el intervalo.

6.8.1 Sistema M/M/1/∞/FIFO (Canal Simple con Población Infinita)

Este sistema tiene un solo servidor, cuyos tiempos de entre arribo y de servicio están distribuidos exponencialmente con parámetros $1/\lambda$ y $1/\mu$ respectivamente. No hay restricción

en la capacidad del sistema y la disciplina en la cola es la del primero en llegar será el primero en salir.

Es importante para el análisis de cualquier sistema de colas el número de clientes en el sistema. Denotamos S_n como el estado del sistema, donde n son los clientes presentes, para $n \geq 0$. $P_n(t)$ denota la probabilidad del estado S_n en un tiempo t. El sistema estará en el estado S_n en el tiempo $t + \Delta t$ si y sólo si uno de los siguientes eventos mutuamente exclusivos ocurre:

El sistema está en el estado S_{n-1} en el tiempo t, y ocurre un arribo (llega un cliente) pero no una salida (no sale ningún cliente) durante un intervalo $(t, t + \Delta t)$[12].
El sistema está en el estado S_n en el tiempo t, y no ocurre ningún arribo ni ninguna salida (no sale ni llega ningún cliente) durante el intervalo $(t, t + \Delta t)$.
El sistema está en el estado S_{n+1} en el tiempo t y no hay ningún arribo(no llega cliente) pero si hay una salida (sale un cliente durante el intervalo $(t, t + \Delta t)$.

Vemos que la probabilidad de un solo arribo durante el intervalo $(t, t + \Delta t)$ es $\lambda\Delta t + \theta(\Delta t)$ mientras que la probabilidad de una sola salida durante el intervalo $(t, t + \Delta t)$ es $\mu\Delta t + \theta(\Delta t)$. La probabilidad de múltiples arribos o salidas durante el intervalo es insignificante. Entonces:

$$P_n(t + \Delta t) = P_{n-1}(t)(\lambda\Delta t)(1 - \mu\Delta t) + P_n(t)(1 - \lambda\Delta t)(1 - \mu\Delta t) + P_{n+1}(t)(1 - \lambda\Delta t)(\mu\Delta t) \quad \text{para } n = 1, 2, \ldots$$

Simplificando y reordenando los términos tenemos:

$$\frac{P_n(t + \Delta t) + P_n(t)}{\Delta t} = \lambda P_{n-1}(t) - (\lambda - \mu)P_n(t) + \mu P_{n+1}(t) \qquad n = 1, 2, \ldots$$

si se toma el límite de ambos lados donde $\Delta t \to 0$ tenemos:

$$P_n'(t) = \lambda P_{n-1}(t) - (\lambda + \mu)P_n(t) + \mu P_{n+1}(t) \qquad n = 1, 2, \ldots$$

Esta ecuación está limitada para n = 1, 2, ... El caso en que n = 0 debe manejarse separadamente, pues en este caso S_{n-1} no es posible. Utilizando el mismo procedimiento tenemos:

$$P_0'(t) = -\lambda P_0(t) + \mu P_1(t)$$

Estas ecuaciones se pueden resumir como un conjunto de ecuaciones diferenciales las cuales nos dan la solución de la distribución del número de clientes en el sistema:

$$P_0'(t) = -\lambda P_0(t) + \mu P_1(t)$$

$$P_n'(t) = \lambda P_{n-1}(t) - (\lambda + \mu)P_n(t) + \mu P_{n+1}(t) \qquad n \geq 1$$

---

[12] Tiempo que transcurre entre t y $t + \Delta t$.

Una vez que el sistema esté en estado estable se usarán estas ecuaciones. Suponiendo que el sistema está estable, las probabilidades no varían apreciablemente en el tiempo, pero el sistema si varía (es estocásticamente estable). Veamos las ecuaciones:

$$P_n = \frac{\lambda}{\mu} P_0$$

$$P_{n+1} = \frac{\lambda - \mu}{\mu} P_n - \frac{\lambda}{\mu} P_{n-1} \qquad j \geq 1$$

En sistemas de colas más complejos puede que sea necesario utilizar derivaciones alternas de las ecuaciones mostradas anteriormente. Usando las ecuaciones en una forma iterativa tenemos que:

$$P_n = \left(\frac{\lambda}{\mu}\right)^n P_0$$

Para completar la solución de la ecuación de estado estable, se necesita encontrar P_0, usando la definición:

$$P_n = \rho^n P_0 \qquad n = 1,2,... \quad \text{donde} \quad \rho = \frac{\lambda}{\mu}$$

Definiendo previamente P_n, para n = 0, 1, ..., como la probabilidad que se encuentren n clientes en ese sistema, tenemos:

$$\sum_{n=0}^{\infty} P_n = \sum_{n=0}^{\infty} \rho^n P_0 = 1 \quad \Rightarrow \quad P_0 = \frac{1}{\sum_{n=0}^{\infty} \rho^n}$$

donde $\sum_{n=0}^{\infty} \rho^j$ es una serie geométrica. Esta serie converge sí y solo sí $\frac{\lambda}{\mu} = \rho < 1$. Cuando converge:

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

Si asumimos $\rho < 1$, la condición necesaria para que el sistema logre un estado estable será que $P_o = 1 - \rho$ y la solución para la ecuación de un estado estable está dada por:

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

Una vez resueltas las ecuaciones de estado estable para este sistema de colas, la distribución del número de clientes en el sistema se conoce, por lo menos en este estado. Esta distribución

se utilizará para calcular varias medidas que se podrán usar para caracterizar el comportamiento del sistema. Estas medidas son:

Utilización del sistema[13]**:** ρ denota el porcentaje de uso del sistema. Mientras más cerca esté ρ de 1, más cargado está el sistema, lo que genera colas más largas y tiempos de espera más grandes:

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

Probabilidad de n clientes en el sistema: La X denota la variable aleatoria que cuenta el número de clientes en el sistema, o sea, en cola o atendiéndose. Cuando el sistema está en estado estable:

Probabilidad de 0 Clientes en el Sistema:

$$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho$$

Probabilidad de n Clientes en el Sistema:

$$P(X = n) = P_n = (1 - \rho)\rho^n \qquad x = 0, 1, 2, ...$$

El valor esperado X se calcula:

$$E(x) = \sum_{x=0}^{\infty} xP_x = (1-\rho)\sum_{x=0}^{\infty} x\rho^x$$

Ahora:

$$\sum_{x=0}^{\infty} x\rho^x = \rho\sum_{x=1}^{\infty} x\rho^{x-1} = \rho\frac{d}{d\rho}\left[\sum_{x=0}^{\infty}\rho^x\right]$$

sí $\rho < 1, \quad \sum_{x=0}^{\infty}\rho^x = \frac{1}{(1-\rho)}$ entonces

$$\sum_{x=0}^{\infty} x\rho^x = \rho\frac{d}{d\rho}\left[\frac{1}{1-\rho}\right] = \frac{\rho}{(1-\rho)^2}$$

Entonces:

[13] Conocida también como *intensidad de tráfico*.

$$E(x) = (1-\rho)\left(\frac{\rho}{(1-\rho)^2}\right) = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}$$

Número de clientes en el sistema: L denota el número de clientes que esperan más los que se atienden en el servidor:

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

Número de clientes en la cola: Para calcular el largo promedio de la cola tenemos que denotar Q como la variable aleatoria que cuenta los clientes que esperan en la cola para recibir un servicio. Entonces:

$$\begin{aligned} E(Q) &= 0P_0 + \sum_{j=1}^{\infty}(j-1)P_j \\ &= \sum_{j=1}^{\infty} jP_j - \sum_{j=1}^{\infty} P_j \\ &= L - \sum_{j=1}^{\infty} P_j \end{aligned}$$

donde $\sum_{j=0}^{\infty} P_j = 1$, así es que $\sum_{j=1}^{\infty} P_j = 1 - P_0$.

Entonces E(Q) = L-(1 - Po), donde $P_0 = 1-\rho$, así es que,

$$E(Q) = L - \rho = \frac{\rho}{1-\rho} - \rho = \frac{\rho^2}{1-\rho}$$

Esta cantidad está denotada como L_q. Así, el largo promedio de la cola está dada por:

$$L_q = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

Tiempo esperando en el sistema: Es el tiempo promedio que un cliente pasa en el sistema, esperando y recibiendo el servicio. Denotamos W como el tiempo esperado en el sistema, L el número esperado en el sistema y λ es la rata de arribo de los clientes para la facilidad de servicio. Entonces $L = \lambda W$.

Recordando que $L = \frac{\rho}{(1-\rho)} = \frac{\lambda}{(\mu-\lambda)}$, tenemos:

$$W = \frac{L}{\lambda} = \frac{1}{\mu-\lambda}$$

Tiempo esperando en la cola: El tiempo promedio de espera un cliente en la cola W_q se relaciona con el tiempo promedio en el sistema por:

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

Así,

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

ó

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)}$$

Relacionando el tiempo promedio de espera en la cola con el largo promedio de la cola, tenemos que:

$$L_q = \lambda W_q$$

Ejemplo 1:

Considere un sistema de comunicación cuya función es retransmitir mensajes y tiene 5 terminales. La unidad de control tiene un almacenamiento disponible infinito, así es que la capacidad del sistema no es un factor significante. La velocidad con que cada terminal transmite los mensajes se expresa mediante una distribución de Poisson con los siguientes valores:

$\lambda_1 = 2$, $\lambda_2 = 0.5$, $\lambda 3 = \lambda_4 = 1$, $\lambda_5 = 1.5$ todas a $\left[\frac{\text{mensajes}}{\text{minuto}}\right]$. Asumir que el tiempo para procesar mensajes sigue una distribución exponencial con un tiempo promedio de servicio de 4 segundos.

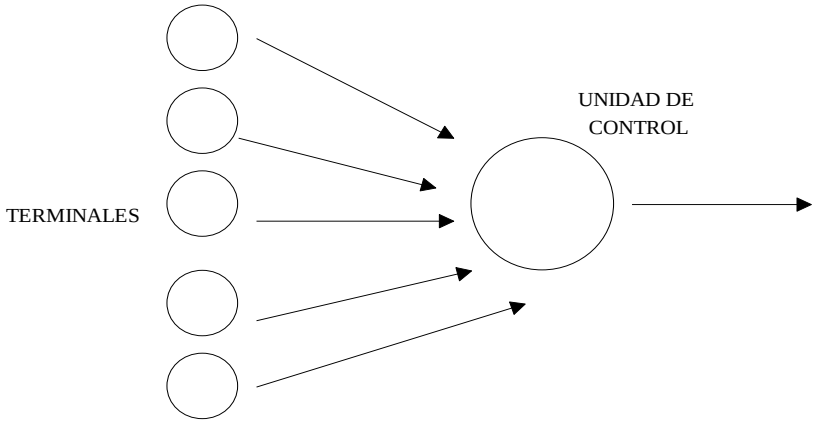


Figura 6.8 Sistema de Retransmisión de Mensajes

Asumir el estado estable y determine:

La probabilidad de que haya cinco o menos mensajes en el sistema: P(x < 5)
El número promedio de mensajes en el sistema procesados y transmitidos: L y L_q
El tiempo promedio empleado por cada mensaje en el sistema: W
El tiempo promedio empleado por cada mensaje antes de ser retransmitido: W_q

Solución:

$$\frac{4}{60} = \frac{1}{15}\text{min. por mensaje} \Rightarrow \mu = 15 \text{ mensajes/min.}$$

$$\rho = \frac{\lambda_T}{\mu} = \frac{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5}{\mu} = \frac{6}{15} = 0.4$$

$$\begin{aligned} P(x < 5) &= P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) + P(x = 4) \\ &= P_0 + P_1 + P_2 + P_3 + P_4 \\ &= (1-\rho) + (1-\rho)\rho + (1-\rho)\rho^2 + (1-\rho)\rho^3 + (1-\rho)\rho^4 \\ &= (1-\rho)(1+\rho+\rho^2+\rho^3+\rho^4) \\ &= (1-0.4)[1+0.4+(0.4)^2+(0.4)^3+(0.4)^4] \\ &= (0.6)(1+0.4+0.16+0.064+0.0256) \\ &= (0.6)(1.6496) \\ &= 0.98976 \end{aligned}$$

$$L = \frac{\rho}{1-\rho} = \frac{0.4}{1-0.4} = 0.6667$$

$$L_q = \frac{\rho^2}{1-\rho} = \frac{(0.4)^2}{1-0.4} = \frac{0.16}{0.6} = 0.2667$$

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

$$W_T = \frac{1}{\mu - \lambda_T} = \frac{1}{15-6} = \frac{1}{9} = 0.1111 \text{ min.}$$

$$W_T = 6.667 \text{ seg.}$$

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)}$$

$$W_{q_T} = \frac{\lambda_T}{\mu(\mu - \lambda_T)}$$

$$W_{q_T} = \frac{6}{15(15-6)} = \frac{6}{15(9)} = \frac{6}{135} = 0.0444 \text{ min.}$$

$$W_{q_T} = 2.667 \text{ seg.}$$

En la sección 6.10 de este capítulo se presenta un resumen de todas las fórmulas de este modelo y de los otros que se presentan a continuación.

Ejemplo 2:

En un Hospital los pacientes llegan con una tasa de llegada de 3 clientes por hora, y son atendidos a una tasa de 9 clientes por hora. Determine el tiempo de utilización, Probabilidad de que no haya clientes en el sistema, número esperado de clientes en el sistema, tiempo esperado de clientes en el sistema.

$$\rho = \frac{\lambda}{\mu} = \frac{3}{9} = 0.3333$$

$$P_0 = 1 - \rho = 1 - 0.333 = 0.6667$$

$$L = \frac{\lambda}{\mu - \lambda} = \frac{3}{9-3} = 0.5$$

$$W = \frac{1}{\mu - \lambda} = \frac{1}{3-9} = 0.1667$$

Ejemplo 3:

Existe una máquina que falla de acuerdo con un proceso de Poisson a una razón de 5 máquinas / horas. Si el costo por cada hora de ocio (no uso) de cada máquina es de $10.00. El gerente de este sistema debe decidir entre dos servicios de reparación. Un taller de reparación. Un taller de reparación cobra $5.00 por hora y puede reparar 6 máquinas por hora. El otro taller cobra $6.00 por hora y puede reparar 8 máquinas por hora.

Suponiendo que los tiempos de servicio están distribuidos exponencialmente. ¿Qué tipo de taller debe contratar el gerente?

Solución:

Asumimos que el sistema es M/M/1/∞/FIFO, es decir, hay un sólo servidor, la capacidad del sistema no está restringida y la disciplina en la cola es la del primero en llegar es el primero en salir.

Falla = tiempo de llegada

λ = 5 máquinas / hora
λ = 6 máquinas / hora
λ = 8 máquinas / hora

Costos C_1 = 5$/hora ===> taller 1
C_2 = 6$/hora ===> taller 2
C_o = 10$/hora ===> lo que cuesta tenerla dañada

W = tiempo desde que se manda una máquina al taller hasta que llegó a la fábrica y la reparamos.

L = máquinas que están fuera de uso (total de máquinas llevadas al taller) costo total promedio = (L)(W) (Co)

$$\rho = \lambda / \mu$$

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

$L_1 = \frac{\lambda}{\mu_1 - \lambda} = \frac{5}{6 - 5}$ = 5 máquinas

$L_2 = \frac{\lambda}{\mu_2 - \lambda} = \frac{5}{8 - 5}$ = 1.66 máquinas

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

$W_1 = \frac{L_1}{\lambda} = \frac{5}{5} = 1$ hora

$W_2 = \frac{5/3}{5} = \frac{1}{3} = 0.33$ hora

Taller 1:

costo $t_1 = (L_1)(W_1)(C_0) = (5)(1)(10) = 50/9 = 50$ \$/hora

Taller 2:

costo $t_2 = (L_2)(W_2)(C_0) = (5/3)(1/3)(10) = 50/9 = 5.55$ \$/hora

Costo Total:

costo T_1 = costo $t_1 + C_1 = 50 + 5 = \$55.00$/hora
costo T_2 = costo $t_2 + C_2 = 5.55 + 6 = \$11.55$/hora

6.8.2 Sistema M/M/1/K/FIFO (Canal Simple con Población Finita)

Este modelo es casi igual al M/M/1/∞/FIFO, sólo que el sistema está restringido con un máximo de K clientes que pueden presentarse en cualquier tiempo dado. El número máximo de clientes que se admiten en el sistema es N (longitud máxima de la línea de espera es igual a N -1). Esto significa que cuando haya N clientes e el sistema, todas las nuevas llegadas se eluden o bien no se les permite unirse al sistema. El resultado es que la tasa efectiva de llegadas en la instalación se vuelve menor que la tasa a la cual se generan llegadas desde la fuente. Este es un modelo más real. El número en el sistema debe ser menor que K. Para n = K tenemos que:

$$P_K(t + \Delta t) = P_K(t)(1 - \mu\Delta t) + P_{K-1}(t)(\lambda\Delta t)(1 - \mu\Delta t)$$

Esta ecuación incluye que el estado S_{k+1} no es posible. Un cliente que arribe es sacado del sistema si está lleno. Si el sistema se asume en equilibrio, las ecuaciones de estado son:

Probabilidad de 0 clientes en el sistema:

$$P_o = \frac{1}{\sum_{n=0}^{k}\left(\frac{\lambda}{\mu}\right)^n} = \frac{1-\rho}{1-\rho^{k+1'}}$$

Probabilidad de n clientes en el sistema:

$$P_n = \left(\frac{\lambda}{\mu}\right)^n P_0 = \rho^n P_0, \quad \text{Para n = 0,1,2...K.}$$

El número esperado en el sistema denotado por L:

$$L = \sum_{n=0}^{K} nP_n$$

$$L = \frac{\rho}{1-\rho} - \frac{(k+1)\rho^{k+1}}{1-\rho^{k+1}}$$

El número esperado en la cola denotada por L_q:

$$L_q = L - (1 - P_0)$$

El tiempo esperado en el sistema denotado por W:

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

El tiempo esperado en el sistema denotado por W_q:

$$W_q = \frac{L_q}{\lambda'}$$

donde $\lambda' = \sum_{n=0}^{\infty} \lambda_n P_n = \lambda \left(1 - P_k\right)$

Ejemplo 1:

Utilizamos el mismo ejemplo anterior, cuya función es la de retransmitir mensajes. Asumimos que la unidad de control es capaz de almacenar un máximo de 10 mensajes. Calcular la probabilidad de que en el sistema existan 5 o menos mensajes, el número de mensajes en el sistema L y el tiempo promedio empleado por cada mensaje en el sistema W.

Solución:

$k = 10$

$\lambda = 6$ mensajes/min.

$\mu = 15$ mensajes/min.

$\rho = 0.4$

Paso 1: Calcular P_0

Utilizando la fórmula $P_o = \frac{1-\rho}{1-\rho^{k+1}}$

$$P_0 = \frac{1 - 0.4}{1 - (0.4)^{10+1}} = \frac{0.6}{0.999958} = 0.60$$

Paso 2: Calcular P_n

Utilizando la fórmula $P_n = \left(\frac{\lambda}{\mu}\right)^n P_0 = \rho^n P_0$

$$P(n \le 5) = P_0 + P_1 + P_2 + P_3 + P_4 + P_5$$
$$P_1 = (0.4)^1 (0.6) = 0.236$$
$$P_2 = (0.4)^2 (0.6) = 0.0944$$
$$P_3 = (0.4)^3 (0.6) = 0.0377$$
$$P_4 = (0.4)^4 (0.6) = 0.0151$$
$$P_5 = (0.4)^5 (0.6) = 0.006$$
$$P(n \le 5) = P_0 + P_1 + P_2 + P_3 + P_4 + P_5 = 0.9792 = 0.98$$

Paso 3: Calcular L

$$L = \frac{\rho}{1-\rho} - \frac{(k+1)\rho^{k+1}}{1-\rho^{k+1}}$$
$$= \frac{0.4}{1-0.4} - \frac{(10+1)0.4^{10+1}}{1-0.4^{10+1}}$$
$$= 0.67$$

Paso 4: Calcular L_q

$$L_q = L - (1 - P_0)$$
$$L_q = .67 - (1 - .6)$$
$$= 0.27$$

Ejemplo 2:

Un centro de impresión y copiado tiene una tasa de llegada de los clientes 9 clientes por hora y procesan a los clientes con una tasa de servicio de 20 por hora si su población es de K = 30, determine la utilización, probabilidad de que haya menos de 5 clientes y el número esperado de clientes en el sistema y en la cola.

$\lambda = 9$

$\mu = 20$

$K = 30$

$$\rho = \frac{\lambda}{\mu} = \frac{9}{20} = 0.45$$

$$P(x \leq 5) = P_0 + P_1 + P_2 + P_3 + P_4 + P_5$$

$$P(x \leq 5) = 0.98155$$

$$Lq = \frac{\lambda + \mu}{\lambda}(1 - Po) = 0.8182$$

$$L = Lq + (1 - Po) = 0.3682$$

6.8.3 Sistema M/M/C/∞/FIFO (Canales Múltiples con Población Infinita)

Este sistema considera C número de servidores o canales de servicio, cada uno con una distribución de tiempo de servicio exponencial distribuido independientemente a una rata μ. El efecto final de utilizar C servidores en paralelo es el de acelerar la tasa de servicio en comparación con el caso de un servidor, permitiendo que se dé servicio a un máximo de C clientes al mismo tiempo. El proceso de arribo se asume que es Poisson. Primero consideremos la medida de la rata de servicio del sistema. Si hay más de C clientes en el sistema, todos los servidores estarán ocupados, de aquí, la medida de la rata de servicio es $C\mu$. Si hay menos que C clientes en el sistema, digamos k, algunos de los servidores estarán ociosos, significando esto que la medida de la rata de servicio es $K\mu$.

Probabilidad de 0 clientes en el sistema:

El cálculo de P_o para este sistema es más complicado que en los anteriores porque la relación para P_n es más compleja. Usando la condición de que $\sum_{j=0}^{\infty} P_j = 1$, tenemos:

$$P_0 = \left[\sum_{n=0}^{C-1} \frac{\lambda^n}{n!\,\mu^n} + \sum_{n=C}^{\infty} \frac{\lambda^j}{C^{n-C} C!\,\mu^n} \right] = 1$$

Definiendo $r = \lambda / \mu$ y $\rho = \lambda / C\mu$ esta relación se escribe:

$$P_0 \left[\sum_{j=0}^{C-1} \frac{r^j}{j!} + \sum_{j=C}^{\infty} \frac{r^j}{C^{j-C}\, C!} \right] = 1$$

Finalmente:

$$P_0 = \left[\sum_{n=0}^{C-1} \frac{r^n}{n!} + \frac{C r^c}{C!(C-r)} \right]^{-1}$$

ó

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

El requisito para el estado estable es que $\lambda / C\mu < 1$ mejor dicho que $\lambda / \mu < 1$.

La segunda fórmula también se puede replantear así:

$$P_0 = \left[\frac{(\lambda/\mu)^C}{C! \left(1 - \frac{\lambda/\mu}{C}\right)} + 1 + \frac{(\lambda/\mu)^1}{1!} + \frac{(\lambda/\mu)^2}{2!} + ... + \frac{(\lambda/\mu)^{C-1}}{(C-1)!} \right]^{-1}$$

donde $\rho = \frac{\lambda}{C\mu}$ tal como se dijo anteriormente.

Probabilidad de n clientes en el sistema:

$$P_n = \begin{cases} \frac{\lambda^n}{n! \mu^n} P_0 & 0 \le n \le C \\ \frac{\lambda^n}{C^{n-C} C! \mu^n} P_0 & n > C \end{cases} \quad \text{ó} \quad P_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!} P_0 & n \le C \\ \frac{(\lambda/\mu)^n}{C! C^{n-C}} P_0 & n > C \end{cases}$$

Número esperado de clientes en la cola denotado por Lq:

$$Lq = \left[\frac{(\lambda/\mu)^{C+1}}{C \cdot C! \left(1 - \frac{\lambda/\mu}{C}\right)^2} \right] P_0 \quad \text{ó} \quad Lq = \frac{P_0 \left(\frac{\lambda}{\mu}\right)^c \rho}{C!(1-\rho)^2}$$

Número esperado de clientes en el sistema denotado por L:

$$L = \frac{\lambda}{\mu} + \left[\frac{(\lambda/\mu)^{C+1}}{C \cdot C! \left(1 - \frac{\lambda/\mu}{C}\right)^2} \right] P_0 = \frac{\lambda}{\mu} + \left[\frac{\mu \left(\frac{\lambda}{\mu}\right)^C}{(C-1)!(C\mu - \lambda)^2} \right] P_0 = \frac{\lambda}{\mu} + L_q$$

Tiempo esperado en la cola denotado por Wq:

$$Wq = \frac{Lq}{\lambda} = W - \frac{1}{\mu}$$

Tiempo esperado en el sistema denotado por W:

$$W = \frac{1}{\mu} + W_q = \frac{1}{\mu} + \left[\frac{(\lambda/\mu)^C \mu}{(C-1)!(C\mu - \lambda)^2} \right] P_0$$

Ejemplo 1:

Utilizando el mismo ejemplo del sistema M/M/1/∞/FIFO, sólo que, en vez de un servidor, se utilizarán 4 servidores, es decir, habrá cuatro líneas idénticas unidas al centro de control, cada una dando el servicio de mensajería en un promedio de 4 segundos. El sistema se analiza como M/M/4/∞/FIFO. Calcular la probabilidad de que hayan 5 o menos mensajes retransmitidos, el tiempo esperado en el sistema (L).

Solución:

$$C = 4$$

$$\mu = \frac{4}{60} = \frac{1}{15}, \quad \mu = 15 \text{ mensajes/min.}$$

$$\lambda_T = 6 \qquad \lambda/\mu = 6/15 = 0.4$$

$$P_0 = \left[\frac{(\lambda/\mu)^C}{C!\left(1 - \frac{\lambda/\mu}{C}\right)} + 1 + \frac{(\lambda/\mu)^1}{1!} + \frac{(\lambda/\mu)^2}{2!} + \ldots + \frac{(\lambda/\mu)^{C-1}}{(C-1)!} \right]^{-1}$$

$$P_0 = \left[\frac{(0.4)^4}{4!\left(1 - \frac{0.4}{4}\right)} + 1 + \frac{(0.4)^1}{1!} + \frac{(0.4)^2}{2!} + \frac{(0.4)^3}{3!} \right]^{-1}$$

$$P_0 = (0.00119 + 1 + 0.4 + 0.08 + 0.01067)^{-1}$$

$$P_0 = 0.670$$

$$P(x \le 5) = P_0 + P_1 + P_2 + P_3 + P_4 + P_5$$

$$= P_0\left(1+\frac{\lambda}{\mu} + \frac{1}{2}\left(\frac{\lambda}{\mu}\right)^2 + \frac{1}{6}\left(\frac{\lambda}{\mu}\right)^3 + \frac{1}{24}\left(\frac{\lambda}{\mu}\right)^4 + \frac{1}{120}\left(\frac{\lambda}{\mu}\right)^5\right)$$

$$= 0.670\left(1 + 0.4 + \frac{1}{2}(0.4)^2 + \frac{1}{6}(0.4)^3 + \frac{1}{24}(0.4)^4 + \frac{1}{120}(0.4)^5\right)$$

$$= 0.670\left(1 + 0.4 + 0.08 + 0.01067 + 1.0667\times10^{-3} + 8.5\times10^{-5}\right)$$

$$= 0.9995$$

$$L = \frac{\lambda}{\mu} + \left[\frac{(\lambda/\mu)^{C+1}}{C\cdot C!\left(1-\frac{\lambda/\mu}{C}\right)^2}\right]P_0$$

$$= 0.4 + \left[\frac{(0.4)^5}{4(4!)\left(1-\frac{0.4}{4}\right)^2}\right]0.67$$

$$= 0.40$$

6.8.4 Sistema M/M/C/K/FIFO (Canales Múltiples con Población Finita)

Este modelo es parecido al anterior, sólo que la población es finita. Para efecto de las fórmulas, k = m. Se asume que el número de servidores es menor o igual a la población finita:

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

$$P_0 = \left[\sum_{j=0}^{C-1}\frac{1}{j!}\left(\frac{\lambda}{\mu}\right)j + \frac{1}{C!}\left(\frac{\lambda}{\mu}\right)^C\left(\frac{C\mu}{C\mu-\lambda}\right)\right]^{-1}$$

$$P_n = \begin{cases} P_0\left(\frac{m!}{(m-n)!n!}\right)\left(\frac{\lambda}{\mu}\right)^n & 0 \le n < C \\ P_0\left(\frac{m!}{(m-n)!C!C^{n-C}}\right)\left(\frac{\lambda}{\mu}\right)^n & C \le n \le m \end{cases}$$

$$L = \sum_{n=0}^{C-1} nP_n + \sum_{n=C}^{m}(n-C)P_n + C\left(1-\sum_{n=0}^{C-1}P_n\right)$$

$$L = \frac{\lambda}{\mu} + \left[\frac{(\lambda/\mu)^C \ \lambda\mu}{(C-1)!\ (C\mu-\lambda)^2} \right] P_0$$

$$L = \frac{\lambda}{\mu} + \left[\frac{(\lambda/\mu)^{C+1}}{C \cdot C! \left(1 - \frac{\lambda/\mu}{C}\right)^2} \right] P_0$$

$$L_q = \sum_{n=C}^{m} (n-C) P_n$$

W_q y W igual que en el modelo M/M/C/∞/FIFO.

Ejemplo:

En una oficina se desea decidir si contratar 1 o 2 secretarias. Si se contrata 1, μ = 40 documentos/día; si se contratan 2, μ = 20 documentos/día. Suponga λ=38 documentos/día. Calcule el número promedio de documentos en el sistema, el tiempo medio de cada documento en ser procesado, el tiempo medio de espera de cada documento en ambos casos y decida qué hacer.

Solución:

Con una secretaria: Sistema M/M/1/∞/FIFO

$$\mu = 40 \text{ doc / dia}$$
$$\lambda = 38 \text{ doc / dia}$$

Número promedio de documentos en el sistema: L

$$L = \frac{\lambda}{\mu - \lambda} = \frac{38}{40 - 38} = 19$$

Tiempo medio de cada documento en ser procesado en días = W - Wq donde Wq es el tiempo de espera en cada documento.

$$W = \frac{1}{\mu - \lambda} = \frac{1}{40 - 38} = \frac{1}{2} = 0.5$$

$$Wq = \frac{1}{\mu - \lambda} - \frac{1}{\mu} = \frac{1}{2} - \frac{1}{40} = 0.475$$

$$W - Wq = 0.5 - 0.475 = 0.025 \text{ días}$$

Con dos secretarias: Sistema M/M/C/∞/FIFO

$$\mu = 20 \text{ doc / dia}$$
$$\lambda = 38 \text{ doc / dia}$$

Número promedio de documentos en el sistema: L

$$L = \frac{\lambda}{\mu} + \left[\frac{(\lambda / \mu)^C \ \lambda\mu}{(C-1)! \ (C\mu - \lambda)^2} \right] P_0$$

$$P_0 = \left[\sum_{j=0}^{C-1} \frac{1}{j!} \left(\frac{\lambda}{\mu} \right) j + \frac{1}{C!} \left(\frac{\lambda}{\mu} \right)^C \left(\frac{C\mu}{C\mu - \lambda} \right) \right]^{-1}$$

$$P_0 = \left[\frac{1}{0!} \left(\frac{38}{20} \right)^0 + \frac{1}{1!} \left(\frac{38}{20} \right)^1 + \frac{1}{2!} \left(\frac{38}{20} \right)^2 \left(\frac{2(20)}{2(20) - 38} \right) \right]^{-1}$$

$$P_0 = \left[1 + 1.9 + 36.1 \right]^{-1}$$

$$P_0 = \frac{1}{39} = 0.0256$$

$$L = \frac{38}{20} + \left[\frac{(38 / 20)^2 \ (38) \ (20)}{1! \left(2 \ (20) - 38 \right)^2} \right] 0.0256$$

$$L = 19.459$$

$$\lambda / \mu = 38 / 20 = 1.9$$

$$P_0 = \left[\frac{(\lambda/\mu)^C}{C!\left(1 - \frac{\lambda/\mu}{C}\right)} + 1 + \frac{(\lambda/\mu)^1}{1!} + \frac{(\lambda/\mu)^2}{2!} + ... + \frac{(\lambda/\mu)^{C-1}}{(C-1)!} \right]^{-1}$$

$$= \left[\frac{(1.9)^2}{2!\left(1 - \frac{1.9}{2}\right)} + 1 + \frac{(1.9)^1}{1!} \right]^{-1}$$

$$= (36.1 + 1 + 1.9)^{-1}$$

$$= 0.0256$$

$$L = \frac{\lambda}{\mu} + \left[\frac{(\lambda/\mu)^{C+1}}{C \cdot C!\left(1 - \frac{\lambda/\mu}{C}\right)^2} \right] P_0$$

$$= 1.9 + \left[\frac{(1.9)^3}{2(2!)\left(1 - \frac{1.9}{2}\right)^2} \right] 0.0256$$

$$= 19.459$$

Tiempo medio de cada documento en ser procesado en días = W - Wq donde Wq es el tiempo de espera de cada documento

$$W = \frac{1}{\mu} + \left[\frac{(\lambda/\mu)^C \ \mu}{(C-1)!\ (C\mu - \lambda)^2} \right] P_0$$

$$W = \frac{1}{20} + \left[\frac{(38/20)^2 \ (20)}{1!\left(2(20) - 38\right)^2} \right] 0.0256$$

$$W = 0.512$$

$$W_q = \left[\frac{(\lambda/\mu)^2 \ (20)}{1!\left(2(20) - 38\right)^2} \right] 0.0256$$

$$W_q = 0.462$$

$$W - W_q = 0.512 - 0.462 = 0.05 \text{ dias}$$

Usando otro juego de fórmulas:

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

$$= \frac{19.459}{38}$$

$$W = 0.512$$

$$W_q = \frac{L_q}{\lambda}$$

$$= \left[\frac{(\lambda / \mu)^{C+1}}{C \cdot C! \left(1 - \frac{\lambda / \mu}{C}\right)^2} \right] P_0 * \frac{1}{\lambda}$$

$$= \left[\frac{(1.9)^3}{2 \cdot 2! \left(1 - \frac{1.9}{2}\right)^2} \right] 0.5256 * \frac{1}{38}$$

$$W_q = 0.462$$

$$W - W_q = 0.512 - 0.462 = 0.05 \text{ dias}$$

Conclusión: Resulta más ventajoso el sistema de una secretaria puesto que la fila del sistema es más corta y el tiempo de proceso es más rápido.

6.9 Fórmulas

Modelo de Canal Simple con Población Infinita: (M/M/1/∞/FIFO)

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

$$P_0 = 1 - \frac{\lambda}{\mu} = 1 - \rho$$

$$P_n = \left(\frac{\lambda}{\mu}\right)^n P_0$$

$$L = \frac{\lambda}{(\mu - \lambda)}$$

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

$$W_q = \frac{\lambda}{\mu(\mu - \lambda)}$$

$$W = W_q + \frac{1}{\mu} = \frac{1}{\mu - \lambda} = \frac{1}{\mu(1 - \rho)}$$

Modelo de Canal Simple con Población Finita (M/M/1/K/FIFO)

k = población n = número de clientes en el sistema

$$P_o = \frac{1}{\sum_{n=0}^{k}\left(\frac{\lambda}{\mu}\right)^n} = \frac{1 - \rho}{1 - \rho^{k+1}} \quad \text{ó} \quad P_0 = \begin{cases} \frac{1-\rho}{1-\rho^{K+1}} & \rho \neq 1 \\ \frac{1}{K+1} & \rho = 1 \end{cases}$$

$$P_n = \left(\frac{\lambda}{\mu}\right)^n P_0 = \rho^n P_0 \quad \text{ó} \quad P_n = \begin{cases} \frac{(1-\rho)\rho^n}{1-\rho^{K+1}} & \rho \neq 1 \\ \frac{1}{K+1} & \rho = 1 \end{cases} \quad n = 0, 1, 2, \ldots, K$$

$$L = m - \frac{\mu}{\lambda}(1 - P_0)$$

$$L_q = m - \frac{\lambda + \mu}{\lambda}(1 - P_o)$$

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

$$W_q = \frac{L_q}{\lambda'}$$

donde $\lambda' = \sum_{n=0}^{\infty} \lambda_n P_n = \sum_{n=0}^{\infty} (m-n)\lambda P_n = \lambda(m-L)$

$$P_o = \frac{1}{\sum_{n=0}^{m}\left(\frac{m!}{(m-n)!}\left(\frac{\lambda}{\mu}\right)^n\right)}$$

$$P_0 = \begin{cases} \dfrac{1-\rho}{1-\rho^{K+1}} & , \quad \rho \neq 1 \\ \dfrac{1}{K+1} & , \quad \rho = 1 \end{cases}$$

$$P_n = \frac{m!}{(m-n)!}\left(\frac{\lambda}{\mu}\right)^n P_0$$

$$P_n = \begin{cases} \dfrac{(1-\rho)\rho^n}{1-\rho^{K+1}} & , \quad \rho \neq 1 \\ \dfrac{1}{K+1} & , \quad \rho = 1 \end{cases} \qquad n = 0, 1, 2, ..., K$$

$$L = \begin{cases} \dfrac{K}{2} & , \quad \rho = 1 \\ \dfrac{\rho\left[1-(K+1)\rho^K + K\rho^{K+1}\right]}{\left(1-\rho^{K+1}\right)\left(K-\rho\right)} & , \quad \rho \neq 1 \end{cases}$$

$$L = m - \frac{\mu}{\lambda}\left(1 - P_0\right)$$

$$L_q = m - \frac{\lambda+\mu}{\lambda}(1-P_o)$$

$$L_q = L - (1 - Po)$$

$$W = \frac{L}{\lambda'} = \frac{L}{\lambda\,(1-P_K)}$$

$$W_q = W - \frac{1}{\mu} = \frac{L_q}{\lambda'} = \frac{L_q}{\lambda\,(1-P_K)}$$

Modelo de Canales Múltiples con Población Infinita: (M/M/C/∞/FIFO)

C = números de canales

$$P_0 = \left[\frac{(\lambda/\mu)^C}{C!\left(1-\frac{\lambda/\mu}{C}\right)} + 1 + \frac{(\lambda/\mu)^1}{1!} + \frac{(\lambda/\mu)^2}{2!} + ... + \frac{(\lambda/\mu)^{C-1}}{(C-1)!} \right]^{-1}$$

$$P_0 = \left[\sum_{j=0}^{C-1} \frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^j + \frac{1}{C!}\left(\frac{\lambda}{\mu}\right)^C \left(\frac{C\mu}{C\mu-\lambda}\right) \right]^{-1}$$

$$P_n = \begin{cases} \frac{(\lambda/\mu)^n}{n!} P_0 & , \quad n \le C \\ \frac{(\lambda/\mu)^n}{C!\, C^{n-C}} P_0 & , \quad n > C \end{cases}$$

$$P_j = \begin{cases} \frac{\lambda^j}{j!\,\mu^j} P_0 & , \quad 1 \le j \le C \\ \frac{\lambda^J}{C^{j-C}\, C!\,\mu^j} P_0 & , \quad j > C \end{cases}$$

$$\rho = \frac{1}{C\mu}$$

$$L = \frac{\lambda}{\mu} + \left[\frac{(\lambda/\mu)^{C+1}}{C \cdot C!\left(1-\frac{\lambda/\mu}{C}\right)^2} \right] P_0$$

$$L = \frac{\lambda}{\mu} + \left[\frac{(\lambda/\mu)^C \ \lambda\mu}{(C-1)!\,(C\mu-\lambda)^2} \right] P_0$$

$$Lq = \left[\frac{(\lambda/\mu)^{C+1}}{C \cdot C!\left(1-\frac{\lambda/\mu}{C}\right)^2} \right] P_0$$

$$Lq = \left[\frac{(\lambda/\mu)^C \ \lambda\mu}{(C-1)!\,(C\mu-\lambda)^2} \right] P_0$$

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

$$W = \frac{1}{\mu} + \left[\frac{(\lambda/\mu)^C \mu}{(C-1)!\,(C\mu-\lambda)^2}\right] P_0$$

$$Wq = \frac{Lq}{\lambda}$$

$$Wq = \left[\frac{(\lambda/\mu)^C \mu}{(C-1)!\,(C\mu-\lambda)^2}\right] P_0$$

Modelo de Canales Múltiples con Población Finita: (M/M/C/K/FIFO)

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

$$P_n = \begin{cases} P_0\left(\frac{m!}{(m-n)!\,n!}\right)\left(\frac{\lambda}{\mu}\right)^n & , \quad 0 \le n < C \\ P_0\left(\frac{m!}{(m-n)!\,C!\,C^{n-C}}\right)\left(\frac{\lambda}{\mu}\right)^n & , \quad C \le n \le m \end{cases}$$

$$L = \sum_{n=0}^{C-1} n P_n + \sum_{n=C}^{m}(n-C)\,P_n + C\left(1 - \sum_{n=0}^{C-1} P_n\right)$$

$$L_q = \sum_{n=C}^{m}(n-C)\,P_n$$

6.10 Resumen

A diario hacemos cola en diferentes lugares sin que nos demos cuenta, por ejemplo, esperando en el semáforo, en la caja de un almacén, en un cajero automático, etc. Inclusive, dentro de la computadora se forman muchas colas por las transacciones internas en proceso de ejecución. Las colas se originan debido a que la demanda de un servicio es mayor que el volumen de atención que puede dar el servidor de la cola, siendo entonces la fila y el servidor los elementos de un sistema cliente servidor. La teoría de colas es un enfoque analítico para el estudio del comportamiento de sistemas que involucran colas. Se han desarrollado un conjunto de ecuaciones para estimar el tamaño de las colas, tiempos de esperas, probabilidades de demora, y otros parámetros. La teoría de colas nos permite analizar la eficiencia de un servidor, como también nos puede ayudar a simular o modelar el comportamiento de un sistema de colas. Con la teoría de colas podemos optimizar el servicio que brindan los servidores.

Entre los diferentes tipos de colas tenemos:

- una cola, un servidor
- una cola, servidores múltiples en paralelo
- una fila, servidores múltiples en serie
- filas múltiples, servidores múltiples en paralelo sin cambio de fila
- filas múltiples, servidores múltiples en paralelo con opción a cambiar de fila.
- y otros.

Entre las disciplinas de cola más comunes tenemos:

- FIFO ó PEPS (Primero en entrar, primero en salir)
- LIFO ó UEPS (Ultimo en entrar primero en salir)
- SIRO o servicio aleatorio
- PRI o servicio de prioridad de más alta
- Disciplina de retiro
- Disciplina de sondeo

Para simular un sistema cliente servidor, es necesario dar las características de arribo a la cola y de servicio. Con estos datos, y las fórmulas dadas, podemos entonces, hacer un estudio del sistema, y así optimizar el servicio que se brinda.

6.11 Bibliografía

Teoría de Colas, http://lovecraft.die.udec.cl/Redes/disc/trabajos/colas/redes.htm

[PAPO] A. Papoulis, Probability, Random Variables, and Stochastic Processes, 2' ed., Nueva York, McGraw-Hill, 1984.

[COX] D.R. Cox y H.D. Miller, The Theory of Stochastic Processes, Londres, Methuen, 1965.

Teoría de Colas, http://tarwi.lamolina.edu.pe/~leojeri/Colas%20en%20el%20Gato.doc

Teoría de Colas o Líneas de Espera,
http://www.unapvic.cl/academicos/nzencovich/admiprod/Unidad%20V/Unidad%20V.html

Características de un Sistema de Cola,
http://www.um.es/~geloca/gio/ampliacion/node3.html

Introducción a la Investigación de Operaciones, Frederick S. Hillier y Gerald J. Lieberman, V Edición, Editorial McGraw Hill