TEORÍA DE COLAS Y SU APLICACIÓN EN EL DISEÑO DE CENTRALES TELEFÓNICAS

(Tails Ttheory and its Application in the Telephonic Central Design)

Hermán Fernández González

Escuela de Ingeniería Electrónica, Uptc-Sogamoso, Grupo de Investigación GINTEL,
hfernandez@sogamoso.uptc.edu.co

(Recibido Mayo 15 de 2007 y aceptado Septiembre 28 de 2007)

Resumen
Desarrolla los modelos matemáticos que describen el comportamiento de los sistemas de teletráfico y su aplicación en el diseño de centrales telefónicas en función del grado de servicio y la carga del sistema.

Palabras clave: Erlang-B, GOS, Kendall.

Abstract
This article develops the mathematical models that describe the behavior of the tele-traffic systems and its applications in the telephonic centrals design, functional to the service grade and to the load of the system.

Key Words: Erlang-B, GOS, Kendall.

1. INTRODUCCIÓN

En el diseño de los sistemas de comunicaciones es importante el análisis de tráfico y capacidad del sistema, buscando optimizar el número de canales necesarios para el intercambio de información. El objetivo principal de este artículo es analizar los sistemas de teletráfico mediante modelos matemáticos que permitan describir las características y el comportamiento de aquellos.

2. TEORÍA DE COLAS

Una cola es un modelo probabilístico (estocástico) en el cual los usuarios arriban a algún tipo de servicio de alguna manera; una vez llegados al servicio esperan en cola hasta que les corresponda su turno para ser atendidos; luego de ser atendidos, generalmente se asume que abandonan el servicio. Para este tipo de modelos generalmente se está interesado en determinar, entre otros resultados, los siguientes:

- El número de usuarios en el sistema, (N_s)
- El numeró de usuarios en cola o en línea de espera, (N_w)
- El tiempo promedio de un usuario en el sistema, (T_s)
- El tiempo promedio de un usuario en cola o en línea de espera, (T_w)

2.1 Tipos de colas

Sistemas o colas con pérdidas: cuando al llegar un usuario, si todos los servidores están ocupados, el usuario abandona el sistema y no es atendido.

Sistemas o colas con línea de espera: cuando al llegar un usuario, si todos los servidores están ocupados, el usuario entra en una línea de espera.

2.2 Notación Kendall para los sistemas de colas

Es utilizada para identificar las características de una cola, desde la distribución de probabilidad de entrada hasta la disciplina del servicio. La notación Kendall consta de seis elementos ubicados de la siguiente manera:

- 1/2/3/4/5/6

Donde 1 es la distribución de probabilidad de entrada, identifica una variable aleatoria discreta que mide el número de elementos que ingresan al sistema en un intervalo de tiempo; 2 es la distribución de probabilidad de tiempo de servicio, identifica una variable aleatoria continua que mide el tiempo que dura el servicio; 3 es el número de servidores en paralelo, identifica el número de servidores que prestan el servicio en el sistema en forma indistinguible; 4 es la capacidad del sistema, es el número de elementos máximo que admite el sistema en servicio y en cola; 5 es el tamaño de la población, es el número de elementos que potencialmente requieren el servicio que presta el sistema, y 6 es la disciplina del servicio, cuando la disciplina es FIFO, se deja la casilla en blanco, en otro caso se indica la disciplina de servicio.

Por ejemplo, la cola M/MN/N indica que la entrada o arribo es Poisson, la atención es exponencial, se tiene N servidores en paralelo, la capacidad del sistema es N, la población es infinita y la disciplina del servicio es FIFO. Este tipo de colas es muy útil para el diseño de centrales de conmutación y de estaciones base en sistemas de telefonía móvil celular.

3. MEDIDAS DE DESEMPEÑO DE UN SISTEMA DE COLAS

Cuando se formula un tratamiento matemático para un sistema de colas su finalidad es evaluar algunas medidas de desempeño.

Una de las principales medidas de desempeño es el nivel de congestión. Y la medida más simple de desempeño para medir la congestión es la **intensidad de tráfico** (A) o **tráfico ofrecido**. Donde la intensidad de tráfico está dada por:

$$A = \frac{1/\mu}{1/\lambda} = \frac{\lambda}{\mu} \qquad (1)$$

Con $1/\mu$ como el tiempo medio de servicio, $1/\lambda$ el tiempo medio entre arribos, λ la tasa de arribos y m la tasa de servicio. Si $A > 1$, los usuarios llegan más rápido de lo que el servicio puede atenderlos. Si existen K servidores en paralelo, cada servidor recibe λ/K usuarios por unidad de tiempo; por lo tanto los K servidores pueden manejar una intensidad de tráfico hasta K.

3.1 Utilización del servicio o factor de utilización

Este parámetro adimensional, denotado por ρ, representa la fracción del tiempo que un servidor está ocupado. Si consideramos un intervalo de tiempo lo suficientemente grande, en un servicio con K servidores en paralelo, se espera que lleguen en promedio $\lambda T/K$, usuarios por servidor, asumiendo que el tráfico se distribuye uniformemente entre los K servidores. Cada uno de los usuarios requiere en promedio un tiempo de servicio $1/\mu$. Así que el tiempo total que se espera que el servidor esté ocupado es:

$$\frac{\lambda T}{K} * \frac{1}{\mu} \qquad (2)$$

Dividiendo por T tenemos:

$$\rho = \frac{\lambda T}{K} * \frac{1}{\mu} * \frac{1}{T} \qquad (3)$$

$$\rho = \frac{\lambda}{K\mu} = \frac{A}{K} \qquad (4)$$

Dado que es físicamente imposible para el servidor estar ocupado más del 100% del tiempo, el factor de utilización máximo por servidor es uno ($\rho_{max} = 1$). Por lo tanto, la expresión correcta para el factor de utilización es:

$$\rho = \min\left\{\frac{\lambda}{K\mu}, 1\right\}; \text{ para K servidores} \qquad (5)$$

$$\rho = \min\left\{\frac{\lambda}{\mu}, 1\right\}; \text{ para un solo servidor} \qquad (6)$$

En este caso $\rho = A$; para $K = 1$.

3.2 Caudal

Es una medida de desempeño frecuentemente usada en sistemas de colas. El caudal se define como el número promedio de usuarios que terminan su servicio por unidad de tiempo.

$$Caudal = K\mu\rho = \min\,\{\lambda, K\mu\}; \qquad (7)$$

Es decir, el caudal es equivalente a la tasa de arribo λ, mientras esta tasa permanezca inferior a la máxima de servicio que es $K\mu$. Para tasas de arribo superiores a este valor, el caudal se fija a un valor de saturación de $K\mu$.

3.3 Tiempo de espera en cola

El tiempo de espera en cola y el tiempo en el sistema o tiempo de respuesta son las medidas de desempeño más importantes desde el punto de vista del usuario, para un sistema de colas. El tiempo de espera en cola (T_w) para el usuario es el tiempo que el usuario permanece en cola antes de ser atendido. Y el tiempo en el sistema (T_s) para el usuario es el tiempo total que el usuario permanece en el sistema.

$$T_{sj} = T_{wj} + E[s] \qquad (8)$$

Con E[s] como el tiempo medio del servicio. $\{T_{wj}; j = 1,2,\ldots\}$ y $\{T_{sj}; j = 1,2,\ldots\}$ son procesos estocásticos, por lo que normalmente se adoptan sus valores esperados T_w y T_s.

3.4 Tamaño de la cola

Sea un proceso aleatorio $\{Q(t), t >= 0\}$ = número de usuarios en la cola en el tiempo t. E igualmente sea el proceso $\{N(t), t >= 0\}$ = número de usuarios encontrados en cola en el tiempo t en el sistema, ya sea en servicio o en cola. En un sistema de K servidores estos dos procesos se relacionan de la siguiente manera:

$$Q(t) = \max\,\{0, N(t) - K\} \qquad (9)$$

La información relacionada con los procesos Q(t) y N(t) es útil en muchas situaciones. Por ejemplo, si se desea calcular el tamaño de buffer requerido para acomodar usuarios en cola de espera, se debe conocer la distribución de Q(t).

4. COLAS MULTISERVIDORES

Existen varios modelos de colas multiservidores (M/M/N/N, M/M/N/∞, M/M/N/K/M, etc.); en este trabajo se describe el modelo M/M/N/N, que es utilizado en el diseño de centrales telefónicas en función del GOS (grado de servicio del sistema).

Cola M/M/N/N. Se caracteriza por:

- Modelo de arribo: Proceso de Poisson con tasa λ.
- Un usuario que arriba entra al sistema, si existe por lo menos un servidor disponible.

- Permanece una cantidad de tiempo exponencialmente distribuida con tasa μ siendo servido.

Un sistema de pérdida es un sistema de colas en el cual los arribos que encuentran todos los servidores ocupados no entran al sistema sino que se pierden.

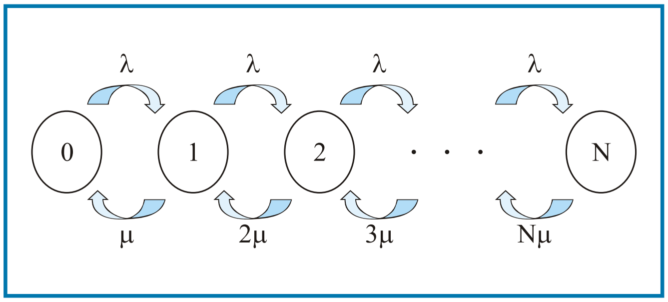

Figura 1. Modelo de colas M/M/N/N.

En la figura 1 se ve que si el sistema se encuentra en el estado n, donde n <= N, quiere decir que tenemos un usuario en el servidor y n-1 usuarios en línea de espera "cola", luego tenemos n usuarios en el sistema. En este caso no pueden existir más de N usuarios en el sistema, ya que este tiene una capacidad finita N. De acuerdo con esto, si se produce un arribo cuando ya existen N usuarios presentes en el sistema, el que llega no entra en él. Además, el modelo tiene la restricción de que los usuarios salen en diferente intervalo de tiempo del servicio.

4.1 Análisis del modelo

Tabla 1. Estados del sistema.

Estado	Salidas	Entradas
0	λP_0	μP_1
1	$(\lambda + \mu)P_1$	$2\mu P_2 + \lambda P_0$
2	$(\lambda + 2\mu)P_2$	$3\mu P_3 + \lambda P_1$
3	$(\lambda + 3\mu)P_3$	$4\mu P_4 + \lambda P_2$
$0 < i < N$	$(\lambda + i\mu)P_i$	$(i+1)\mu P_{i+1} + \lambda P_{i-1}$
N	$N\mu P_N$	λP_{N-1}

De la tabla 1 tenemos, tomando el estado 0:

$$P_1 = \left(\lambda/\mu\right)P_0 \qquad (10)$$

Tomando el estado 1 tenemos:

$$P_2 = \frac{(\lambda + \mu)P_1 - \lambda P_0}{2\mu} \qquad (11)$$

$$P_2 = \frac{\lambda P_1 + \mu P_1 - \lambda P_0}{2\mu} \qquad (12)$$

$$P_2 = \lambda/2\mu\, P_1 + 1/2\, P_1 - \lambda/2\mu\, P_0 \qquad (13)$$

Tomando la ecuación (10) y reemplazando en (13) se tiene:

$$P_2 = \lambda/2\mu\, P_1 \qquad (14)$$

Tomando nuevamente la ecuación (10) y reemplazando en (14) tenemos:

$$P_2 = \frac{\left(\lambda/\mu\right)^2}{2} P_0 \qquad (15)$$

Tomando el estado 2 tenemos:

$$P_3 = \frac{(\lambda + 2\mu)P_2 - \lambda P_1}{3\mu} \qquad (16)$$

$$P_3 = \frac{\lambda P_2 + 2\mu P_2 - \lambda P_1}{3\mu} \qquad (17)$$

$$P_3 = \lambda/3\mu\, P_2 + 2/3\, P_2 - \lambda/3\mu\, P_1 \qquad (18)$$

Tomando la ecuación (14) y reemplazando en (18) se tiene:

$$P_2 = \lambda/3\mu\, P_2 \qquad (19)$$

Tomando la ecuación (15) y reemplazando en (19) tenemos:

$$P_3 = \frac{\left(\lambda/\mu\right)^3}{3!} P_0 \qquad (20)$$

De igual forma para el estado N-1 tenemos:

$$P_{N-1} = \frac{\left(\lambda/\mu\right)^{N-1}}{(N-1)!} P_0 \qquad (21)$$

De la tabla 1, tomando el estado N, tenemos:

$$P_N = \left(\lambda/N\mu\right)P_{N-1} \qquad (22)$$

Tomando la ecuación (21) y reemplazando en (22)

$$P_N = \frac{\left(\lambda/\mu\right)^N}{N!} P_0 \qquad (23)$$

Si tiene finitos estados, el sistema se encuentra en un estado cualquiera con certeza. Luego:

$$\sum_{i=0}^{N} P_i = 1 \qquad (24)$$

De la ecuación (23) y con $P_N = P_i$ tenemos:

$$\sum_{i=0}^{N} \frac{\left(\lambda/\mu\right)^i}{i!} P_0 = 1 \qquad (25)$$

$$P_0 = \frac{1}{\sum_{i=0}^{N} \frac{\left(\lambda/\mu\right)^i}{i!}} \qquad (26)$$

De la ecuación (23) y con P_N = Pn tenemos:

$$P_n = \frac{\left(\lambda/\mu\right)^n}{n!} \frac{1}{\sum_{i=0}^{N} \frac{\left(\lambda/\mu\right)^i}{i!}} \qquad (27)$$

La ecuación (27) varía con los diferentes valores de n, para n = 0, 1, 2, 3,…, N.

Con A = λ/μ, donde A es la intensidad de tráfico tenemos:

$$P_n = \frac{(A)^n}{n!} \frac{1}{\sum_{i=0}^{N} \frac{(A)^i}{i!}} \qquad (28)$$

La ecuación (28) define la probabilidad de que el sistema esté en cualquier estado (n-esimo) si n < N, donde N es el número máximo de servidores. En los sistemas de comunicaciones se necesita es la probabilidad de bloqueo, esto quiere decir, cuando el sistema se encuentra en el estado N debido a que todos los arribos se pierden. Si me encuentro en este estado, la ecuación (28) pasa a ser la expresión de las tablas erlang-B con N como el número de circuitos de interconexión, luego:

$$P_n = \frac{(A)^N}{N!} \frac{1}{\sum_{i=0}^{N} \frac{(A)^i}{i!}} \qquad (29)$$

En la ecuación (29) P_n pasa a ser P_B (Probabilidad de bloqueo) con P_B = Erlang- B = $E_{1,N}$ (A) o Erlang-B de primera clase.

4.2 Resultados

De la ecuación (29) se puede concluir la importancia que tiene la probabilidad de bloqueo en el diseño de centrales telefónicas.

4.3 Análisis y discusión de resultados

La ecuación (29) muestra que en la medida que se aumenta el tráfico del sistema (A), la probabilidad de que las llamadas en progreso sean bloqueadas aumenta, a no ser que se aumente el número de enlaces o canales (N) del sistema.

5. CONCLUSIONES

Para el diseño de centrales telefónicas se requiere un análisis estadístico que permita generar los valores de la tasa de entrada l y la tasa de salida m, con los cuales se pueda calcular el tráfico real del sistema.

La optimización de los recursos de comunicaciones en el diseño de centrales telefónicas es dependiente de la calidad con la que se quiera ofrecer el servicio (GOS).

Una vez se cuente con los datos que se generaron del modelo matemático, el proveedor de los servicios de comunicaciones debe realizar un análisis costo-beneficio.

AGRADECIMIENTOS

A la Universidad Pedagógica y Tecnológica de Colombia por brindar los espacios académicos que permiten la socialización de conocimientos.

BIBLIOGRAFÍA

Camerano R. (1997). Teoría de colas. Bogotá: Universidad Distrital Francisco José de Caldas.

Daigle J. (1991). Queueing Theory for Telecommunications. Addison Wesley Publishing.

Robertazzi T. (2000). Computer Networks and Systems: Queueing Theory and Performance Evaluation. 3 ed. New York Springer.

Schwartz M. (1994). Redes de telecomunicaciones. Addison-Wesley.

Tanenbaum A. (2003). Computer Networks. 4 ed. New Jersey: Prentice Hall.

Molnar S., Vidacs A. (1998). How to Characterize HurstyTraffic? Departamento de Telecomunicaciones y Telemática. Universidad Técnica de Budapest.