Introducción a la

Teoría Básica de Teletráfico

Sr. H. Leijon, UIT

UNION INTERNATIONALE DES TELECOMMUNICATIONS
INTERNATIONAL TELECOMMUNICATION UNION
UNION INTERNACIONAL DE TELECOMUNICACIONES

Teoría Básica de Teletráfico (T)

INTRODUCCION

Contenido

1. Antecedentes

2. Alcance y naturaleza de la Teoría de Teletráfico

   - Proceso de entrada

   - Mecanismo de servicio

   - Disciplina de la disposición en cola (queue discipline)

   - Conservación del flujo

3. Modelo matemático

1. Antecedentes

El desarrollo de la teoría de tráfico telefónico comenzó a principios de este siglo. Los primeros logros en este campo se deben a Dane A. K. Erlang, cuyos trabajos se publicaron entre 1909 y 1928. Entre aquéllos que continuaron con las ideas de Erlang, debemos mencionar al sueco Conny Palm, cuyos aportes durante el período 1936 - 1946 (1957) contribuyeron a dar a la teoría de trafico su actual rigor. Muchas otras personas de diversas nacionalidades han contribuido también al desarrollo de la presente teoría.

La teoría de tráfico que puede ser aplicada a casos prácticos, se basa en el supuesto del equilibrio estadístico, lo que implica que ésta sólo puede tratar con casos sujetos a condiciones estacionarias.

Aún no se ha inventado ningún método práctico de cálculo para las condiciones no estacionarias. No obstante, el antecedente teórico para abordar tales casos se presentó en la tesis doctoral de Palm, en 1943, en la cual hizo un estudio de variaciones en la intensidad de llamadas. Ahora, con la ayuda de simulaciones computarizadas, ya es posible tratar con casos de tráfico no estacionario. Sin embargo, las teorías aquí consideradas se limitarán a las condiciones estacionarias.

Las teorías existentes usan diferentes combinaciones de supuestos y las derivaciones desde 1909 a la fecha surgen desde diferentes niveles de conocimientos y usan, en parte, diferente terminología. Un repaso directo de las derivaciones para diferentes casos originalmente presentados no proveerían, por tanto, un claro escrutinio de la habilidad de la teoría para describir los diferentes casos que ocurren en la práctica. Por eso, se ha preferido presentar la teoría de tráfico de una manera más general, de la cual puedan derivarse varios casos particulares.

Por ello, en la primera sección se presentarán las características comunes a diferentes métodos de agrupación, tanto para sistemas de pérdida como de retardo. Luego, será preferible tratar separadamente con el grupo de disponibilidad total en un sistema de pérdida y con el grupo de disponibilidad total en un sistema de retardo. Las teorías para interconexiones graduales (gradings) y sistemas de enlace serán tratadas en los capítulos posteriores, en los cuales se comprenderá más fácilmente cómo la teoría general se aplica a estos casos.

2. Alcance y naturaleza de la Teoría de Teletráfico

La teoría de teletráfico puede considerarse como la teoría de la disposición en cola aplicada a los sistemas de telecomunicaciones. El concepto general de esta teoría tiene que ver con el análisis matemático de sistemas sujetos a demandas, cuyas ocurrencias y duraciones pueden, en general, especificarse sólo probabilísticamente. Por ejemplo, considere un sistema telefónico cuya función es proveer trayectos de comunicación entre pares de aparatos telefónicos (clientes) de acuerdo a la demanda. La provisión de un trayecto de comunicación permanente entre cada par de aparatos telefónicos sería astronómicamente caro y tal vez imposible. En respuesta a este problema, se proveen en un solo grupo común, las facilidades necesarias para establecer y mantener un trayecto de conversación entre un par de aparatos telefónicos, para ser usadas por una llamada cuando sean requeridas, y retornadas al grupo cuando ya no sean necesarias. Esto introduce la posibilidad de que el sistema sea incapaz de establecer una llamada demandada por carencia de equipo disponible en ese momento. Así, surge inmediatamente la pregunta: ¿cuánto equipo debe proveerse para que la proporción de llamadas experimentando retardos esté por debajo de un nivel aceptable específico?
Preguntas similares a esa surgen en el diseño de muchos sistemas bastante diferentes de un sistema telefónico: cuántas camas debería proveer un hospital?; ¿cuántos terminales de datos pueden ser conectados a un servicio de computadora de tiempo compartido ?
Estas preguntas tienen una característica común: no pueden preverse en cada caso las veces en que los requerimientos de servicio ocurran y las cantidades de tiempo en que estos requerimientos emplearán facilidades, excepto en un sentido estadístico. Aunque estos sistemas son usualmente muy complejos, con frecuencia es posible abstraer de la descripción del sistema, un modelo matemático cuyo análisis brinde información útil.

Considere el siguiente modelo. Los clientes piden el uso de un tipo particular de equipo (servidor). Si un servidor está disponible, el cliente que llega lo tomará y lo mantendrá por un tiempo, después del cual el servidor estará inmediatamente disponible para otro cliente que llega o está en espera. Si el cliente entrante no encuentra servidor disponible, entonces toma una acción específica tal como esperar o retirarse. En consecuencia, el modelo se define en términos de 3 características: el proceso de entrada, el mecanismo de servicio y la disciplina de la disposición en cola.

El proceso de entrada describe la secuencia de pedidos de servicio. Con frecuencia, por ejemplo, el proceso de entrada se especifica en términos de la distribución de las duraciones de tiempo entre los instantes de llegada de clientes consecutivos. El mecanismo de servicio es la categoría que incluye características tales como el número de servidores y la duración del tiempo en que los clientes retienen los servidores. Por ejemplo, los clientes pueden ser procesados por un

solo servidor, cada cliente reteniendo el servidor por la misma duración de tiempo. La disciplina de disposición en cola especifica la disposición de los clientes bloqueados (clientes que encuentran todos los servidores ocupados). Por ejemplo, podría asumirse que los clientes bloqueados dejen el sistema inmediatamente o que esperen en cola por el servicio y sean servidos por orden de llegada.

Ahora considere el siguiente modelo. Dos ciudades están interconectadas por un grupo de n troncales telefónicas (servidores). Suponga que las llamadas que ingresan encuentran todas las troncales ocupadas y no esperan en línea sino que se retiran del sistema. (Técnicamente no hay espera en cola). ¿Qué proporción de llamadas entrantes (clientes) no podrán encontrar una troncal disponible (y así se perderán)?.

Queremos derivar una fórmula que prediga la proporción de llamadas perdidas como una función de la demanda; es decir, deseamos derivar una fórmula que permita estimar el número de troncales requeridas para cumplir un criterio de servicio predeterminado desde un estimado de la carga de tráfico telefónico que se genera entre las dos ciudades. El gran valor práctico de cualquier modelo que lleve a tal fórmula es obvio.

Ahora daremos una derivación heurística de la fórmula requerida, usando un concepto de gran importancia en ciencia e ingeniería, aquél de la conservación de flujo. La siguiente derivación es heurística, de modo que nadie debe esperar comprenderla totalmente; la "derivación" es un argumento de plausibilidad y es correcta en ciertas circunstancias.

Cuando el número de clientes en el sistema es *j*, se dice que el sistema está en estado E_j (j = 0, 1, ... n). Sea P_j la proporción de tiempo en que j troncales están ocupadas; P_j es la proporción que de tiempo que el sistema gasta en el estado E_j. Llamemos λ a la tasa de llegada de la llamada; λ es el número promedio de solicitudes de servicio por unidad de tiempo. Considere primero el caso j < n. Ya que las llamadas llegan con una velocidad total λ, y ya que la proporción de tiempo que el sistema gasta en el estado E_j es P_j, la velocidad a la que ocurre la transición $E_j \rightarrow E_{j+1}$ (el número promedio de tales transiciones por unidad de tiempo) es, por tanto, λP_j. Ahora, consideremos el caso en el que j = n. Ya que el estado E_{n+1} representa un estado físicamente imposible (sólo hay n troncales), la transición $E_n \rightarrow E_{n+1}$ es cero. Por lo tanto la tasa a la cual la transición ascendente $E_j \rightarrow E_{j+1}$ ocurre es λP_j cuando j = 0,1,.....n -1; y es Cero cuando j = n.

Ahora consideremos las transiciones descendentes.

$E_{j+1} \rightarrow E_j$ (j = 0, 1, ..., n - 1)

Suponga que el tiempo medio de ocupación (el tiempo promedio en que una llamada retiene una troncal) es τ, entonces si una troncal simple está ocupada, el número promedio de llamadas terminando durante un tiempo transcurrido τ es 1; la tasa de terminación para una sola llamada es entonces, $1/\tau$. De modo similar, si dos llamadas están en progreso simultáneamente y el promedio de duración de una llamada es τ, el número promedio de llamadas terminando durante un tiempo transcurrido τ es 2; por tanto, la tasa de terminación para dos llamadas simultáneas es $2/\tau$. Por medio de este razonamiento, entonces, la tasa de terminación para j+1 llamadas simultáneas es $(j+1)/\tau$. Como el sistema está en estado E_{j+1} una proporción de tiempo P_{j+1}, concluimos que la transición descendente $E_{j+1} \rightarrow E_j$, ocurre a la velocidad $\frac{j+1}{\tau} \cdot P_{j+1}$ transiciones por unidad de tiempo (j = 0, 1, ... n-1).

Ahora aplicamos el principio de conservación de flujo. Si el sistema ha de estar en equilibrio estadístico, es decir, si la proporción relativa de tiempo que el sistema gasta en cada estado ha de ser una cantidad estable, entonces la transición ascendente $Ej \rightarrow Ej_{+1}$ debe ocurrir con la misma velocidad que la transición descendente $Ej+1 \rightarrow Ej$. Así es como tenemos las llamadas ecuaciones de balance del equilibrio estadístico.

$$\lambda \ P_j = (j+1) \ . \ \tau^{-1} \ . \ P_{j+1} \qquad \text{(j = 0, 1, j....., n - 1)} \qquad \text{(TIN 2.1)}$$

Estas ecuaciones pueden resolverse recursivamente; el resultado, que expresa cada P_j en términos del valor P_o es:

$$P_j = \frac{(\lambda \cdot \tau)^j}{j!} \cdot P_o \qquad \text{(j = 0, 1, ..., n)} \qquad \text{(TIN 2.2)}$$

Ya que los números $\{P_j\}$ representan todos los posibles estados, la suma de ellos debe ser igual a la unidad:

$$P_o + P_1 + \ \ldots\ldots\ P_n = 1 \qquad \text{(TIN 2.3)}$$

Usando la ecuación de normalización (TIN 2.3) conjuntamente con la ecuación (TIN 2.2), podemos determinar P_o :

$$P_o = (\sum_{k=0}^{n} \frac{(\lambda \cdot \tau)^k}{k!})^{-1} \qquad \text{(TIN 2.4)}$$

Así, para la proporción P_j del tiempo en que j troncales están ocupadas, obtenemos la fórmula:

$$P_j = \frac{(\lambda.\tau)^j / j!}{\sum_{k=0}^{n} (\lambda.\tau)^k / k!} \qquad \text{(TIN 2.5)}$$

Debe hacerse una observación importante con respecto a la fórmula (TIN 2.5) y ella es que las proporciones $\{P_j\}$ dependen de la tasa de llegada λ y el tiempo medio de ocupación τ sólo a través del producto $\lambda\ \tau$. Este producto es una medida de la demanda hecha al sistema; con frecuencia ésta es llamada la carga ofrecida y se le da el símbolo A, $A = \lambda \cdot \tau$. Los valores numéricos de A son expresados en unidades llamadas erlangs (erl), por el matemático danés A. K. Erlang, quien fue el primero en publicar la fórmula (TIN 2.5) en 1917. Cuando j = n en la fórmula (TIN 2.5), tenemos la bien conocida fórmula de pérdida de Erlang, simbolizada en Europa por $E_{1n}(A)$:

$$E_{1n}(A) = \frac{A^n / n!}{\sum_{k=0}^{n} A^k / k!} \qquad \text{(TIN 2.6)}$$

Posteriormente, derivaremos estos resultados más detenidamente. El punto aquí es que algunos resultados matemáticos potencialmente útiles, han sido derivados sólo usando razonamiento heurístico. La pregunta que ahora debemos hacernos es: ¿bajo qué condiciones son válidos estos resultados?

Más precisamente, ¿qué suposiciones acerca del proceso de entrada y mecanismo de servicio se requieren para la validación de las fórmulas (TIN 2.5) y (TIN 2.6)?. ¿Puede justificarse la aseveración de que la velocidad del tránsito descendente es proporcional al tiempo medio de ocupación? ¿Cuál es la relación entre la proporción P_j de tiempo que j llamadas están en progreso y la proporción Π_j, es decir, de llamadas que ingresan y que encuentran j otras llamadas en progreso? ¿Qué tan ampliamente es aplicable el análisis de conservación de flujo? ¿Cómo manejar los procesos para los cuales este tipo de análisis es inaplicable?

A veces, preguntas de esta naturaleza requieren argumentos matemáticos altamente sofisticados. Sin embargo, aquí tomaremos una posición intermedia con respecto al uso de matemáticas avanzadas. El material debe ser accesible a un estudiante que entiende la teoría de probabilidades aplicada y las áreas afines de las matemáticas.

3. Modelo matemático

De lo anterior comprendemos que la teoría de tráfico consiste en el modelo matemático de un sistema de telecomunicaciones (o de alguna parte de él) y su comportamiento cuando las demandas son hechas en él o por medio de él. Toda esta teoría es un ejercicio en base a un modelo: no podemos establecer y examinar un sistema idéntico y por eso construimos una versión (hipotética) simplificada, con entradas bien definidas y la analizamos. Por eso, la validez y utilidad de la teoría que desarrollamos descansa enteramente sobre la respuesta a la pregunta : ¿Cuán satisfactorio es el modelo? Si tenemos poca confianza en el modelo, entonces sin importar cuán sofisticadas sean nuestras matemáticas, los resultados finales de la teoría serán poco confiables.

A fin de establecer nuestro modelo, tenemos que considerar cuidadosamente un número de puntos:

1) ¿Exactamente en qué parte del sistema estamos interesados? ¿Podemos separar la sección relevante y verla independientemente, o debemos tratarla dentro de un todo.

2) ¿Cuál es el comportamiento técnico preciso de esta sección por la cual nos hemos decidido, en términos de tiempos operativos, limitaciones en los accesos, tiempos muertos, respuesta detallada a una demanda, etc.

3) ¿Cómo se comporta el flujo entrante de demandas?

4) ¿Qué información queremos de nuestro modelo, y cuán exacta debe ser?

Todos los puntos son interdependientes. Tal vez es más fácil comenzar por el flujo entrante de demanda, e inquirir cómo ésta está estructurada: lo cual por sí mismo involucrará algunas suposiciones de que las fuentes de demanda - por ej. abonados - están reaccionando al comportamiento "usual" del sistema de la manera "usual". Típicamente este flujo de demandas está ampliamente regido por el azar, de modo que los métodos de procesos estocásticos serán apropiados. Entonces, necesitamos saber:

a) ¿cuál es (en lenguaje probabilístico) el *proceso de llegada* de demandas?; y,

b) ¿cuál es la distribución del trabajo que ellas traen?

La descripción del proceso de llegada podrá requerir mayor o menor detalle. En un sistema de baja congestión de tráfico nuevo ofrecido por una multitud de abonados independientes, podemos asumir con una alta precisión que (a cualquier velocidad sobre períodos de tiempo no muy largos) las llegadas son *puro azar*; es decir, que forman un proceso poissoniano. Si examinamos un sistema al que se ofrece tráfico de desbordamiento, necesitaremos una descripción más compleja; y más compleja aún si esperamos una proporción significativa de intentos repetidos.

Consideremos ahora el segundo punto: la distribución del trabajo que trae una sola demanda. Sin duda esto puede tener una considerable variación cualitativa: si nos referimos a la ocupación de circuitos, el "trabajo" consiste en un solo tiempo continuo de ocupación; mientras que si estamos modelando el control común de un sistema procesador complejo, puede ser una secuencia de tareas separadas, muy diferentes en tipos y duraciones. Sin embargo, hay dos casos particularmente comunes e importantes y son aquéllos donde el tiempo de ocupación tiene una distribución exponencial negativa y donde éste es efectivamente determinístico, esto es, una constante.

Ahora volvamos nuestra atención al punto (4) precedente: qué información queremos del modelo. Esto naturalmente requiere, a cualquier velocidad, algún conocimiento del comportamiento del sistema (por ejemplo, están bloqueadas las llamadas perdidas o ellas están en fila de espera) y una comprensión del proceso de entrada. Ya que tal entrada es estocástica, la salida de nuestro modelo será probabilística, y puede consistir de probabilidades de pérdida, retardos promedio, percentiles de ocupación del procesador, o cantidades similares. Incluso puede ser ésta sólo la primera etapa en la construcción de un modelo más complejo, en cuyo caso, podemos necesitar saber cómo completar detalles de las distribuciones de retardos, de llamadas desbordadas, o sobre algún otra cantidad que afecta el resto del sistema. En esta etapa, se considera el punto (1).

Ahora estamos listos para atender los detalles de ingeniería del sistema, punto (2) precedente; y es en esta etapa, en la que nuestro modelo matemático toma forma y se sabe si es posible un tratamiento analítico. Finalmente, todo el proceso se repite hasta que tengamos la confianza que es consistente y que el comportamiento del sistema es sin duda compatible con los supuestos del flujo de demanda entrante y viceversa.

Asumimos entonces que un modelo ha sido o puede ser establecido. Este necesariamente será, de una manera u otra, aproximado y debemos estimar qué efecto tendrá esta aproximación en los resultados del modelo. Si la respuesta es demasiado efecto, debemos dar nueva forma al modelo. El último requerimiento es siempre para un número o grupo de números, esto es, para un cálculo numérico: cambios tan útiles con frecuencia pueden hacerse entre modelos aproximados y cálculos aproximados.

De hecho, puede no ser posible especificar los datos de entrada del comportamiento del abonado con el detalle que se necesita, porque las cantidades relevantes son desconocidas o, inclusive, inmensurables. En tales circunstancias son razonables las aproximaciones: decidir qué haría El Hombre Razonable y tomar una decisión sobre qué es lo que en el sistema debe ser dimensionado (por supuesto, con las salvaguardas apropiadas al sistema mismo) o analizar varios modelos, que difieran sólo en el comportamiento del abonado, y presentar un rango de resultados para la toma de decisiones final basados en otros datos.

Afortunadamente la mayor parte de cantidades de interés, como la salida desde los modelos matemáticos, son notablemente sólidas con respecto a las variaciones en los procesos de entrada (por supuesto siempre que ciertos parámetros críticos, como el tráfico total ofrecido, se mantengan constantes); y modelos tan simples dan resultados tan exactos y útiles. Sin embargo, nunca debe olvidarse que no importa que tan elaborada sea la solución matemática,

UN MAL MODELO SIGNIFICA RESULTADOS NO CONFIABLES