# Capítulo 5: **Estructura de las Redes de Comunicaciones.**

**Objetivos:** Describir las interconexiones de equipos, las desventajas de la interconexión total y la necesidad de redes de comunicaciones con otra conectividad. Establecer las características de las redes conmutadas y las redes de difusión. Explicar las redes conmutadas en sus versiones:de circuitos,de mensajes y de paquetes, esta con sus subclases: circuitos virtuales y datagramas. Describir el congestionamiento en las redes de conmutación de paquetes. Describir la Teoría de colas de espera, su aplicación al tráfico telefónico y a las redes de comunicaciones.

---

## 5.1.- Interconexión de equipos.Orígenes de las redes digitales ó redes de computadores.

Veamos como van evolucionando una serie de equipos que se desean interconectar bidireccionalmente.Obviamente un único equipo no puede establecer ningún tipo de comunicación,el análisis comienza con dos equipos que para intercomunicarse requieren de la instalación de una línea de comunicación entre ellos,y cada equipo debe tener una vía de entrada salida que permita la conexión con esa línea de comunicación,**ver Fig.5.1.**

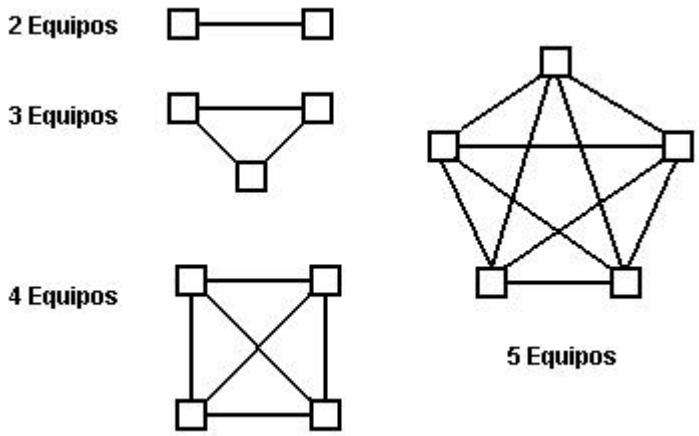



**Figura 5.1.Estructura de conectividad total.**

Si hay un tercer equipo deberíamos instalar dos líneas adicionales, con lo que las líneas totalizan tres y además cada equipo deberá tener dos vías de entrada-salida para conectarse con las dos líneas de interconexión. Sí el número de equipos sube a cuatro, hay que agregar tres líneas y cada equipo deberá tener tres vías de entrada-salida, tal como muestra la **Figura 5.1**. Si continuamos la adición de un enésimo equipo, que deberá tener N-1 líneas de entrada-salida, hace que debamos:

- Tender N-1 líneas de comunicación, una con cada equipo ya instalado.
- Añadir una vía de entrada-salida en cada uno de los N-1 equipos ya instalados.

En consecuencia para **conectividad total**(ó **topología tipo malla**), se requiere:

- Tender N(N-1)/2 líneas de comunicación(cada equipo requiere N-1 líneas, son N equipos, pero cada línea conecta dos equipos, lo que dá el resultado mencionado).
- Que cada equipo tenga N-1 vías de entrada-salida, y como son N equipos en total se requieren N(N-1) vías de entrada-salida.

Por lo tanto la complejidad del equipo de interconexión crece con $N^2$ tanto en líneas de comunicación como en vías de entrada-salida, por otra parte si los equipos no son capaces de establecer más de una comunicación simultánea, o sea comunicaciones en paralelo, por las diversas vías de entrada-salida, se tendrá siempre un estado de ociosidad de los recursos disponibles, por esto la topología malla sólo se usa en circunstancias muy especiales, tales como:

1. Algunas redes de comunicaciones muy especializadas donde existe la posibilidad de transmitir simultáneamente por las N-1 vías instaladas en cada equipo.

2. Redes de telefonía privada ó intercomunicadores, donde la cercanía de los equipos hace que el costo de instalación sea ventajoso respecto de otras tecnologías, además en estos casos la vía de entrada-salida es un sencillo conector telefónico asociado a una tecla.

Por otra parte estadísticamente se ha determinado que un empleado se muda una ó más veces en un año, y que las remodelaciones y reorganizaciones son frecuentes. Modificar ó extender el cableado, que es equipo a equipo, resulta traúmatico y costoso.

En consecuencia la topología de malla ó interconectividad total no se utiliza(salvo los casos especiales mencionados) y en su lugar se instala una **red de comunicaciones** a la que se conectan los diversos equipos y que permite solucionar el problema asociado con el aumento del número de equipos.

La **red de comunicaciones** proporciona las **vías de comunicación** necesarias para establecer las **interconexiones** cuando estas son solicitadas y transportar la información a su destinatario, es claro que estos recursos son **compartidos** entre los usuarios de la red.

La **red de comunicaciones** es entonces **un conjunto organizado de recursos compartidos de comunicación.**

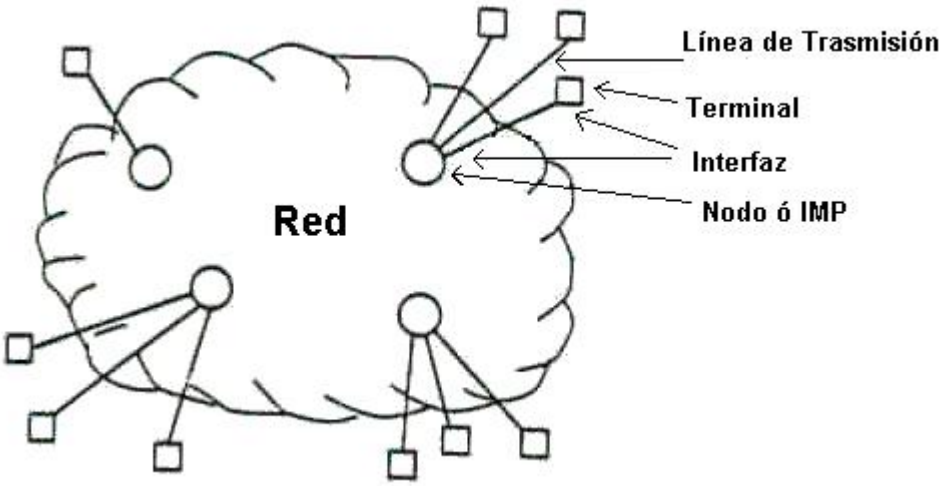



**Figura 5.2.Esquema general de una red de comunicaciones con sus estaciones**

La **Figura 5.2** muestra en forma genérica una red de comunicaciones y sus usuarios(terminales,máquinas de extremo,estaciones de trabajo ó hosts), y en ella observamos,aunque la nomenclatura en las redes no es única,que podemos enumerar los siguientes elementos que las componen:

- **Terminales,máquinas de extremo,estaciones de trabajo ó hosts**, denominación que engloba a cualquier equipo de telecomunicaciones: fax, teléfono, videoteléfonos, computadores, telex, videotext,etc, y es donde se encuentran ubicados los usuarios, se representan mediante un cuadrado en la **Figura 5.2.**

- **nodos ó elementos de conmutación ó central de conmutación ó IMP(Interchange Message Procesors)**,son equipos(generalmente computadores)especializados que se utilizan para conectar dos ó más vías de comunicación.Los datos llegan por una vía de entrada y el elemento de conmutación deberá seleccionar una vía de salida,se representan con un círculo en la **Figura 5.2.**Los nodos conectados **solamente** a otros nodos,que no se muestran en la figura,se denominan **centrales** ó **de tránsito.**Los nodos que tienen conectados terminales se llaman **nodos periféricos.**

- **interfaces**,son los puntos de interconexión de las estaciones ó de los nodos con las vías de comunicación.

- **líneas de transmisión**,son el soporte físico(cables,fibras,enlaces de radio punto a punto,etc) para mover las información(generalmente bits) y por lo tanto hablamos de la implementación genérica de las vías de comunicación,en la **Figura 5.2** solo se muestran,mediante líneas contínuas, las que enlazan las interfaces de las estaciones con las correspondientes de los nodos,las que interconectan nodos entre sí no se detallan pues depende de la estructura específica de esas interconexiones.

- **red de comunicaciones propiamente dicha**,o simplemente **la red**,que ya se definió y que es capaz de transmitir información(analógica ó digital, aunque estas últimas se extienden cada día más)entre interfaces. Algunos autores denominan a la red de comunicaciones **subred de comunicación.** La red se dibuja habitualmente encerrada por una línea que es el **límite de la red,** a veces se considera que el límite de la red llega hasta los interfaces de terminal(la línea terminal-nodo es parte de la red) y otras solo hasta los interfaces de nodo(la línea terminal-nodo **no** es parte de la red).

El diseño de la red de comunicaciones se simplifica si separamos los aspectos puros de comunicación(**la red ó subred)** de los de aplicación(**las estaciones de trabajo)**.

En esta forma de interconectar:

- **cada estación, máquina de extremo, host ó terminal** se conecta a través de su **interfaz** mediante una **línea de transmisión** con la **interfaz** de un **nodo.** Por lo tanto **solo** requiere de **una** vía de entrada salida.

- Los **nodos** se interconectan entre sí por diversos mecanismos que estudiaremos enseguida.

- **la red** no considera ni interpreta el contenido de la información intercambiada, solo se ocupa de transportarla del origen al destino,ese es su **único** fin.

- los recursos de comunicación,entre nodos,están **compartidos.**

**5.2.-Arquitectura de la red y técnicas de transferencia de la información.**

Vamos a estudiar las redes en base a su **arquitectura**(topología de los medios de comunicación y tipo de conmutación) y a las **técnicas de transferencia de información**, y podemos clasificarlas en:

- **Redes conmutadas**,que se ilustran en la **Figura 5.3**,en las que el terminal de origen selecciona un terminal de destino y **la red se encarga de proveer un camino entre ambos,** efectuando las **conmutaciones** necesarias. Dentro de esta categoría existen varias clases:

  - **Redes de conmutación de circuitos.**
  - **Redes de conmutación de mensajes.**
  - **Redes de conmutación de paquetes**, que tiene dos subclases:

    1. **Circuito virtual.**
    2. **Datagramas.**

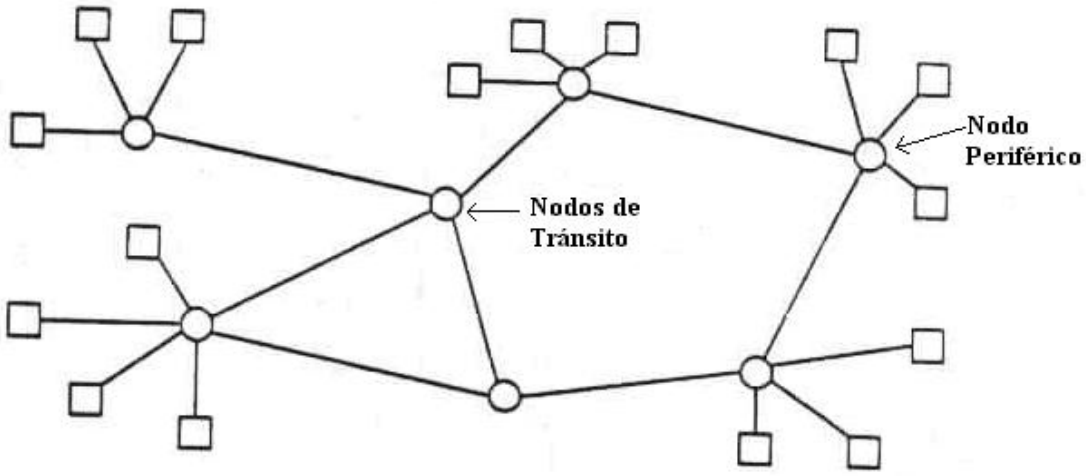



**Figura 5.3 Red conmutada (nodos de tránsito, nodos periféricos)**

- **Redes de difusión**, **Figura 5.4**,en las que el terminal de origen envía la información al medio de transmisión, que es común a todos los usuarios, por lo tanto **todos** la reciben, y uno ó más de ellos seleccionan la información a recibir.

  Aquí también existen varias clases:

  1. **Redes de área local(LANs) cableadas.**
  2. **Redes por satélite.**
  3. **Redes inalámbricas(wireless networks), se iniciaron con de las paquetes por radio(packet radio) y fueron seguidas por las WLAN: wireless LANs estándar 802.11, wireless PAN(Personal Area Network) estándar 802.15 y wireless MAN estándar 802.16.**

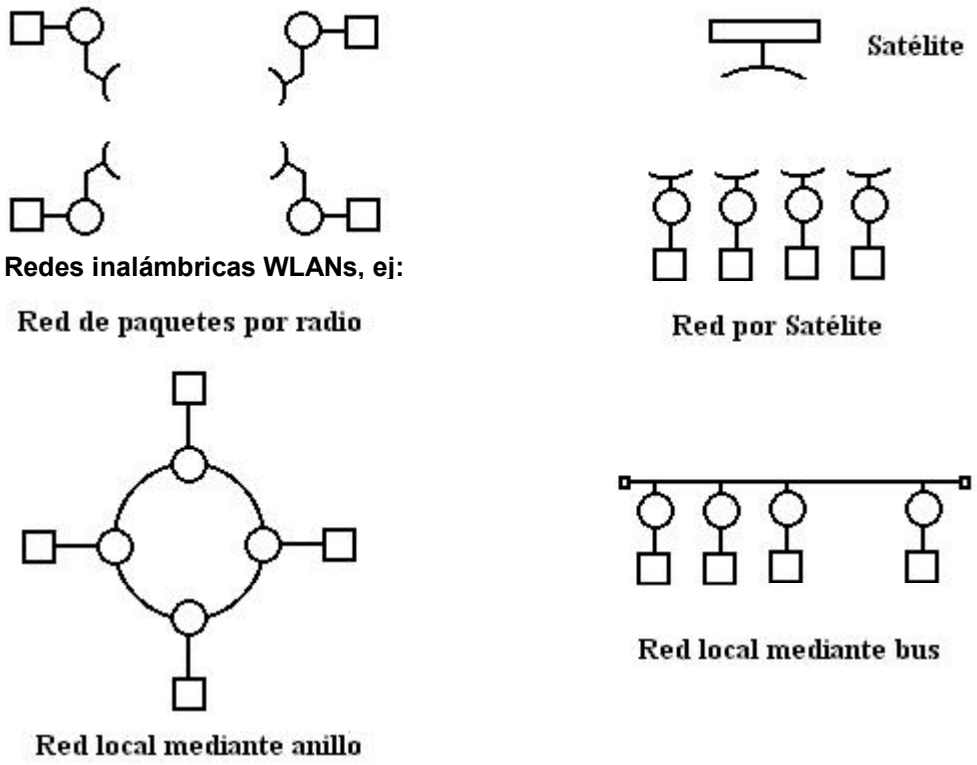



**Figura 5.4.Redes de difusión.**

Se observa que en estas redes de difusión los nodos **no ejercen funciones de conmutación** sino que controlan el **acceso al medio** pues evitan que se envíe información cuando el medio está ocupado por otro terminal,esto no impide que dos terminales accedan a la red **al mismo** tiempo contaminando la información(se denominan **colisiones**) . La topología en anillo evita este problema,común a los tres otros esquemas,pues en lugar de compartir el medio comparte el **transporte** de la información por el medio,aquí el nodo recibe,selecciona lo que le interesa y retransmite.En el Capítulo 6 daremos amplios detalles de estas tecnologías.

### 5.3.-Redes conmutadas.Descripción de las técnicas de conmutación.

#### a.-Conmutación de Circuitos.

La técnica de conmutación de circuitos permite que el terminal emisor se una fisicamente al terminal receptor mediante un circuito único y específico que sólo pertenece a esa unión. **La Figura 5.5** ilustra la conmutacion de circuitos entre dos terminales.

El **circuito** se establece completamente antes del inicio de la comunicación y queda libre (« se libera» en el argot) cuando uno de los terminales involucrados en la comunicación la da por finalizada.

El principal inconveniente de la conmutación de circuitos es la escasa rentabilidad que se obtiene de los circuitos en el caso de que en el proceso de intercambio de información entre los terminales se introduzcan pausas de transmisión motivadas por cualquier circunstancia como por ejemplo, la consulta a una base de datos o la ejecución en interactivo de cualquier programa o utilidad. Para mejorar la rentabilidad de las líneas se multiplexa más de una comunicacion por línea. La multiplexacion es el procedimiento por el cual un circuito o línea de comunicaciones transporta más de una señal, cada una en una localización individualizada que constituye su canal. El sistema desmultiplexor es que permite distinguir las diferentes señales originales.

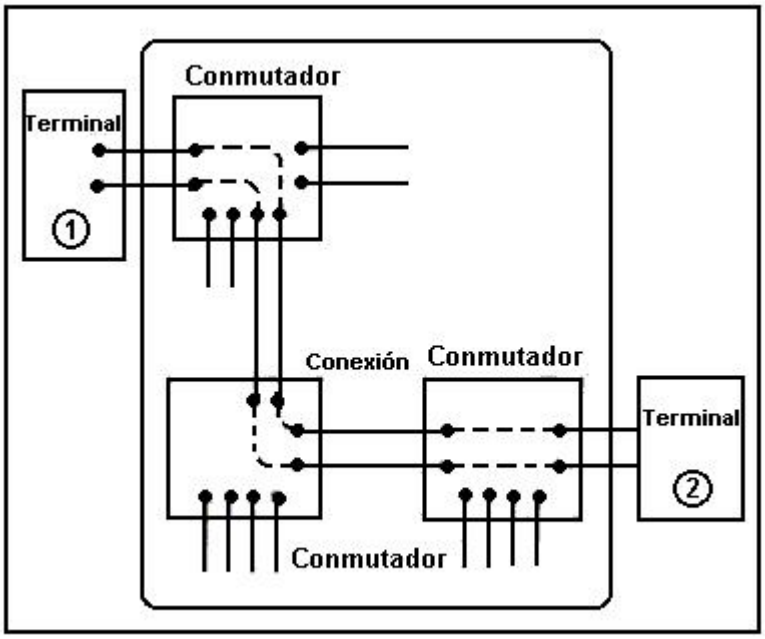



**Figura 5.5.Conmutación de Circuitos.**

**b.-Conmutación de mensajes.**

El **mensaje** es una unidad lógica de datos de usuario, de datos de control o de ambos que el terminal emisor envía al receptor.

El mensaje consta de los siguientes elementos, llamados campos:

_ Datos de usuario. Depositado por el interesado.
_ Caracteres SYN. (Caracteres de sincronía)
_ Campos de dirección. Indican el destinatario de la información.
_ Caracteres de control de la comunicación.
_ Caracteres de control de errores.

Además de los campos citados, el mensaje puede contener una cabecera que ayuda a la identificación de sus parámetros (dirección de destino, enviante, canal a usar, etc. )

La conmutación de mensajes se basa en el envío del mensaje que el terminal emisor quiere transmitir al terminal receptor, a un nodo o centro de conmutación en el que el mensaje es almacenado y posteriormente enviado al terminal receptor o a otro nodo de conmutación intermedio, si es necesario. Este tipo de conmutación siempre conlleva el almacenamiento y posterior envío del mensaje-**store and forward**- lo que origina que sea imposible transmitir el mensaje al nodo siguiente hasta la completa recepción del mismo en el nodo precedente.

**La Figura 5.6** ilustra el caso ideal en el que no existe retraso alguno en la transmisión de la información entre un nodo y el siguiente.

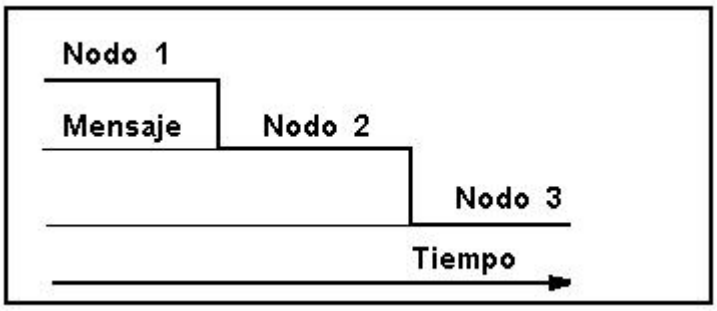



**Figura 5.6.Conmutación de Mensajes suponiendo que es nulo el retraso entre nodos.**

El tipo de funcionamiento hace necesaria la existencia de memorias de masa intermedias en los nodos de conmutación para almacenar la información hasta que esta sea transferida al siguiente nodo. Asimismo se incorporan los medios necesarios para la detección de mensajes erróneos y para solicitar la repetición de los mismos al nodo precedente.

A los mensajes se les une en origen una cabecera que indica el destino de los mismos para que puedan ser correctamente entregados. Los nodos son computadoras encargadas del almacenamiento y posterior retransmisión de los mensajes hacia su destino, con lo que esta técnica resulta atractiva en determinadas condiciones.

La conmutación de mensajes presenta como ventaja relevante la posibilidad de poder transmitir un mismo mensaje a todos los nodos de la red, lo que resulta muy beneficioso en ciertas condiciones.

**c.-Conmutación de paquetes.**

La conmutación de paquetes surge intentando optimizar al utilización de la capacidad de las líneas de transmisión de existentes. Para ello seria necesario disponer de un método de conmutación que proporcionara la capacidad de retransmisión en tiempo real de la conmutación de circuitos y la capacidad de direccionamiento de la conmutación de mensajes.

La conmutación de paquetes se basa en la división de la información que entrega a la red el usuario emisor en **paquetes** del mismo tamaño, que generalmente oscila entre 1000 y 2000 bits.

Los paquetes poseen una estructura tipificada y, dependiendo del uso que la red haga de ellos, contienen información de enlace o información de usuario. **La Figura 5.7** muestra las partes características de un paquete en sus modalidades de enlace y de información de usuario.

La estructura global de los paquetes en la que es dividida la información se compone a su vez de varias entidades individuales, llamados campos. Cada uno de los campos posee su misión especifica.

El campo indicador (Flag) tiene una longitud de ocho bits y su misión es la de indicar el comienzo y el final del paquete.

El campo de dirección (Address) indica cual es el sentido en el que la información debe progresar dentro de la red. Su longitud es de ocho bits.

El campo de secuencia de verificación de trama ( Frame Checking Secuence ) es el encargado de servir como referencia para comprobar la correcta transmisión del paquete. Como

puede verse en la **Figura. 5.7** su posición depende de la naturaleza del paquete que se transmite. Su longitud es de dieciséis bits.

| Indicador | Dirección | Control | FCS | Indicador |
|---|---|---|---|---|
| 01111110 | 8 bits | 8 bits | 16 bits | 01111110 |

a) Paquete con información de enlace solamente.

| Indicador | Dirección | Control | Información | FCS | Indicador |
|---|---|---|---|---|---|
| 01111110 | 8 bits | 8 bits | n-bits | 16 bits | 01111110 |

b) Paquete con la información de enlace y datos de usuario.

**Figura. 5.7 Estructura típica de un paquete de información.**

El campo de la información posee una longitud indeterminada, aunque sujeta a unos márgenes superiores, y es el que contiene la información que el usuario emisor desea intercambiar con el receptor. Además el campo de información incluye otro tipo de datos que son necesarios para el proceso global de la comunicación como el numero de canal lógico que se esta empleando, el numero de orden dentro del mensaje total, etc.

La técnica de conmutación de paquetes permite dos formas características de funcionamiento: **Circuito Virtual y Datagrama,** que describiremos un poco mas adelante.

**5.4.- Redes conmutadas. Sucesión de eventos en las cuatro técnicas de conmutación.**

Las cuatro técnicas de conmutación: de circuitos, de mensajes, de paquetes circuito virtual y de paquetes datagrama, son muy interesantes, y para describirlas en forma similar consideraremos las siguientes etapas:

- Establecimiento de la conexión.
- Transferencia de la información.
- Liberación de la conexión.

Antes de comenzar debemos aclarar que de inmediato veremos que la conmutación de mensajes y la de paquetes requiere nodos de conmutación digital que manipulen(reciban, almacenen y transmitan)estructuras de datos, por ello las redes conmutadas con estos tipos de conmutación **solo** pueden transmitir señales digitales(ó analógicas digitalizadas),en cambio la conmutación de circuitos permitirá conmutar, si se usa la técnica adecuada, tanto señales analógicas como señales digitales.

**La Figura 5.8** muestra en el tiempo los distintos eventos que se suceden en la comunicación por medio de redes conmutadas, debe hacerse notar que la línea de transmisión terminal-nodo es única por terminal, esta permanentemente a la disposición de este y no influye en la conmutación de la red por lo que no se incluye en este estudio.Denominaremos **nodo de origen** el nodo de conmutación del que depende el terminal de origen de la comunicación, y **nodo destino** al nodo de conmutación del que depende el terminal destino de la comunicación.

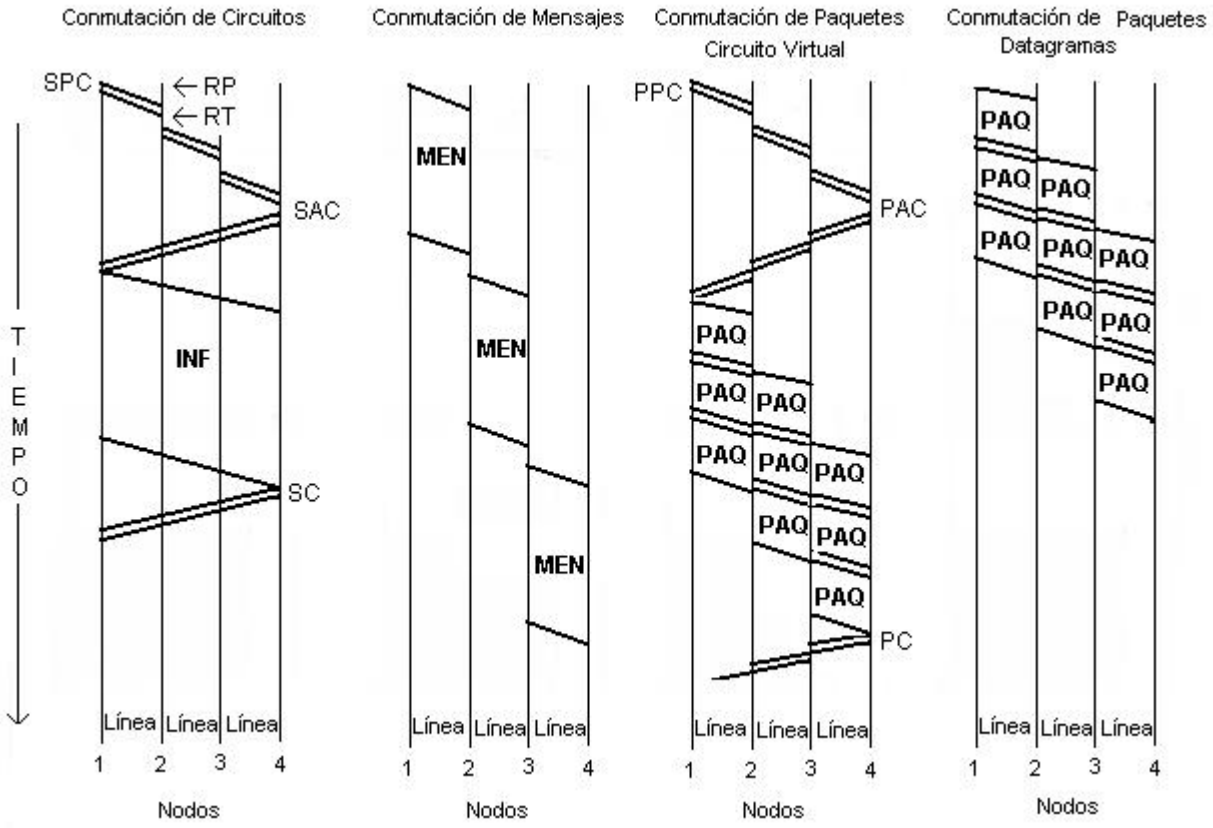



**Figura 5.8 Sucesión de eventos para las cuatro técnicas de conmutación**.

Las claves empleadas son las siguientes:

RP: retardo de propagacion nodo-nodo.
RT: retardo de tratamiento del nodo.
SPC: señal de petición de comunicación.
SAC: señal de aceptación de comunicación.
SC: señal de confirmación.
PPC: paquete de petición de comunicación.
PAC: paquete de aceptación de comunicación.
PC: paquete de confirmación.
INF: información.
MEN: mensaje.
PAQ: paquete.

**1-Conmutación de circuitos.**

- Establecimiento de la conexión.

  - El nodo origen identifica al primer nodo intermedio en el camino hacia el nodo destino y le envia una señal de petición de comunicación.
  - El procedimiento anterior se repite tantas veces en cuantos nodos intermedios sea necesario atravesar hasta arribar al nodo destino.
  - El nodo destino envía una señal de aceptación de comunicación al nodo origen a través de todos los nodos intermedios atravesados durante la petición de comunicación.

- El circuito queda establecido y se mantendrá durante toda la transferencia de la información.

- Transferencia de la información.

  - Se envía toda la información que se desee desde el origen hasta el destino y viceversa, ya que entre los dos terminales existe una línea de transmisión bidireccional dedicada (aunque permanezca inactiva).

- Liberación de la conexión.

  - El nodo origen envía (implícita o explícitamente) al nodo destino la petición de finalización de la comunicación.
  - El nodo destino contesta al nodo origen con una señal de confirmación.
  - El circuito queda liberado.

**2-Conmutación de mensajes.**

- Establecimiento de la conexión.

  - No existe.

- Transferencia de la información.

  - El nodo origen identifica al primer nodo intermedio en el camino hacia el nodo destino y le envía el mensaje completo.
  - Tras recibir el mensaje y almacenarlo, el primer nodo intermedio identifica al segundo nodo intermedio y, cuando la línea de transmisión esté disponible, le envía el mensaje completo.
  - El procedimiento anterior se repite tantas veces cuantos nodos intermedios sea necesario atravesar hasta que el mensaje arribe al nodo destino.

- Liberación de la conexión.

  - No existe.

**3-Conmutación de paquetes mediante circuito virtual.**

- Establecimiento de la conexión.

  - El nodo origen identifica al primer nodo intermedio en el camino hacia el nodo destino y le envia un paquete de petición de comunicación.
  - El procedimiento anterior se repite tantas veces cuantos nodos intermedios sea necesario atravesar hasta arribar al nodo destino.
  - El nodo destino envía un paquete de aceptación al nodo origen a través de todos los nodos intermedios atravesados durante la petición de comunicación.
  - El circuito virtual ( Conexión lógica ) queda establecido. Cada nodo intermedio ha formado una asociación entre el circuito virtual que se ha establecido y la conmutación requerida ( entrada-salida ) en dicho nodo.
  - Todos los paquetes que llevan la información seguirán el mismo camino ( circuito virtual ) desde el origen hasta el destino.

- Transferencia de la información.

  - El nodo origen comienza a enviar paquetes al primer nodo intermedio, usando el circuito virtual. Cada paquete lleva un identificador del circuito virtual.

- El primer nodo intermedio envía los paquetes, según le llegan (sin almacenarlos), al siguiente nodo intermedio del circuito virtual.
- El procedimiento anterior se repite tantas veces cuantos nodos intermedios sea necesario atravesar hasta que todos los paquetes arriben al nodo destino.

- Liberación de la conexión.

  - El nodo origen envía (implícitamente o explícitamente) al nodo destino la petición de finalización de la conmutación.
  - El nodo destino contesta al nodo origen con un paquete de confirmación.
  - Existen dos posibilidades:
    1. El circuito virtual desaparece, no quedando "historia", por lo que cualquier nueva comunicación es como si fuera la primera.
    2. El circuito virtual se mantiene para siempre (es decir, el encaminamiento de los futuros paquetes queda "conocido " por el terminal origen).

**4.-Conmutación de paquetes mediante datagramas.**

- Establecimiento de la conexión

  - No existe.

- Transferencia de la información.

  - El nodo origen comienza a enviar paquetes al primer nodo intermedio. Cada paquete lleva un identificador del nodo y del terminal destino.
  - El primer nodo intermedio envía los paquetes, según le llegan (sin almacenarlos), al siguiente nodo intermedio.
  - El procedimiento anterior se repite tantas veces cuantos nodos intermedios sea necesario atravesar hasta que todos los paquetes arriben al nodo destino.
  - La ruta seguida por cada paquete desde el origen hasta el destino puede ser diferente. Debido a los retardos de propagación los paquetes pueden llegar en orden diferente al que fueron enviados.

- Liberación de la conexión.

  - No existe.

**Encaminamiento.**

Existen dos clases de encaminamiento que se pueden aplicar **a todas** las posibilidades de conmutación consideradas anteriormente.

- **Encaminamiento estático**, en el que cada nodo tiene una tabla fija de rutas que aplica independientemente del estado de la red. Por lo tanto el encaminamiento entre un par de terminales dados será idéntico para cualquier intento de comunicación.

- **Encaminamiento dinámico**, en el que cada nodo primero examina el estado de la red (analiza el tráfico de la misma y la disponibilidad de los enlaces) y despúes decide la mejor ruta. Por lo tanto, en general, el encaminamiento entre un par de terminales dados será diferente para cualquier intento de comunicación.
  Esta decisión puede ser autónoma para cada nodo o puede realizarse por un control central de la red tras considerar el estado global de la misma.

La **Tabla 5.1** efectúa una comparación entre las diversas modalidades de conmutación.

| Conmutación de circuitos | Conmutación de Mensajes | Conmutación de paquete modo datagrama | Conmutación de paquete modo circuito virtual |
|---|---|---|---|
| Circuito de Transmisión dedicado | Circuito de Transmisión no dedicado | Circuito de Transmisión no dedicado | Circuito de Transmisión no dedicado |
| Transmisión continua de los datos | Transmisión de Mensajes | Transmisión de Paquetes | Transmisión de Paquetes |
| Suficientemente rápido para trabajar en interactivo | Demasiado lento para trabajar en interactivo | Suficientemente rápido para trabajar en interactivo | Suficientemente rápido para trabajar en interactivo |
| El camino se establece desde el principio al fin del mensaje | La ruta se establece para cada mensaje | La ruta se establece para cada paquete | La ruta se establece para toda la comunicación |
| Nodos de conmutación electromecánicos o electrónicos | Nodos de conmutación de mensajes con tiempos de almacenamiento elevados | Nodos de conmutación con tiempos de almacenamiento pequeños | Nodos de conmutación con tiempos de almacenamiento pequeños |
| El usuario es responsable de la protección contra pérdida de información | Los mensajes son responsabilidad de la red | La red no se responsabiliza de los paquetes individuales | La red no puede responsabilizarse de las secuencias de paqutes |
| Normalmente no es posible la conversión de códigos o velocidades | Conversión de códigos y velocidades | Conversión de códigos y velocidades | Conversión de códigos y velocidades |
| Transmisión con ancho de banda fijo | Uso dinámico del ancho de banda disponible | Uso dinámico del ancho de banda disponible | Uso dinámico del ancho de banda disponible |
| No hay bits de cabecera después del establecimiento de la comunicación | Bits de Cabecera en cada mensaje | Cabeceras en cada paquete | Cabeceras en cada paquete |
| Envio de señal de ocupado si el receptor no está libre | No existe señal de ocupado | El emisor no es avisado si el paquete no se ha entregado | El emisor es avisado si el receptor no acepta la conexión |
| Retraso en el establecimiento de la llamada. Retraso en la transmisión muy pequeño (propagación) | Retraso en la transmisión del mensaje | Retraso en la transmisión del paquete | Retraso en el establecimiento de la llamada. Retraso en la trasmisión del paquete |
| Los mensajes no son almacenados | Los mensajes se almacenan para recuperarlos posteriormente | Los paquetes pueden almacenarse mientras se toman decisiones de red | Los paquetes se almacenan mientras se toman decisiones de red |

**5.5.-Redes de Comunicaciones.Conmutación de paquetes.Congestionamiento.**

Cuando hay demasiados paquetes en la red(subred)ó en partes de ella hay una degradación del desempeño,si ello sucede hablamos de **congestionamiento**,la **Figura 5.9** muestra esta situación.Cuando la cantidad de paquetes enviados por las estaciones ó hosts esta dentro de la capacidad de conducción de la red,todos los paquetes se entregan(salvo unos pocos afligidos por errorres de transmisión)y la cantidad entregada es proporcional al número enviado,es la parte lineal de la **Figura 5.9.** Sin embargo a medida que aumenta el tráfico los nodos (enrutadores)ya no pueden manejarlo y comienzan a perder paquetes. Esto obliga a retransmisiones que tienden a empeorar las cosas, a muy alto tráfico el desempeño se desploma por completo y casi no hay entrega de paquetes.

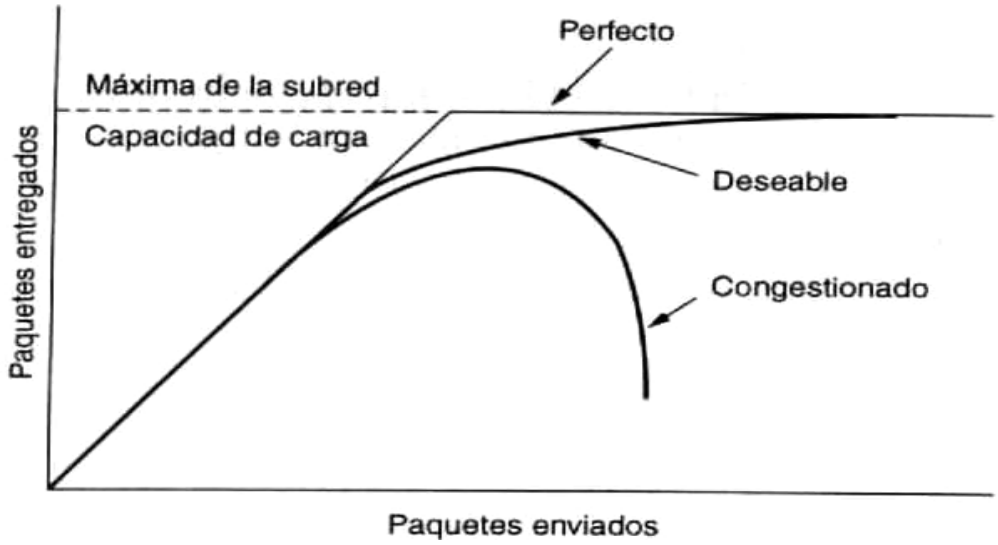



**Figura 5.9.El desempeño de una red se deteriora notablemente al aumentar el tráfico y ocurrir congestionamiento.**

El congestionamiento puede ocurrir por diversas razones,si repentinamente comienzan a llegar paquetes por tres ó cuatro líneas de entrada(tal como explica el trabajo de William Stallings del Apéndice 5 A el tráfico aparece en **ráfagas** y no sigue la distribución de Poisson que se utilizará en la Teoría de Colas de Espera de la Sección 5.6 que clasicamente se ha aplicado) y todos necesitan la misma línea de salida con lo que se generará una **cola**. Sin no hay suficiente memoria para contenerlos a todos se perderán paquetes. La adición de memoria puede ayudar, pero hasta cierto punto,Nagle(1987)descubrió que si los nodos(enrutadores)tienen una capacidad infinita de memoria, el congestionamiento en lugar de mejorar puede empeorar, esto se debe a que cuando los paquetes llegan al principio de la cola su temporización ha terminado(repetidamente)y se han enviado duplicados, todos ellos serán debidamente reenviados al siguiente nodo(enrutador), aumentando la carga en todo el camino hasta el destino.

Los **procesadores lentos** también pueden causar congestionamiento, si los CPU de los nodos son lentos para llevar a cabo las tareas de administración requeridas(buffers de encolamiento, actualización de tablas, etc)pueden alargarse las colas, aún cuando haya capacidad ociosa en las líneas. Del mismo modo **líneas de poco ancho de banda** producen congestionamiento.

La modernización de las líneas sin cambiar los procesadores, o viceversa, generalmente mejoran las cosas un poco, pero desplazan el “cuello de botella” a otra parte. Frecuentemente el problema es un “desequilibrio” entre las diversas partes de la red, y el problema persistirá hasta que todos los componentes estén en equilibrio.

El congestionamiento como se dijo tiende a alimentarse a si mismo y a empeorar, si un nodo(enrutador)no tiene *buffers* libres, debe ignorar los paquetes nuevos, el nodo que se los envió terminará su temporización y como no puede liberar *su buffer* hasta que el paquete sea reconocido, volverá a enviarlo, quizas varias veces, no liberando ese *buffer* que normalmente habría liberado, así el **congestionamiento** se acumula.

Debe mencionarse la diferencia entre **control de congestionamiento** y **control de flujo**, el primero tiene que ver con asegurar que la red sea capaz de manejar el tráfico requerido, es un asunto global, mientras que el control de flujo se relaciona con el tráfico punto a punto entre un transmisor y un receptor dado**[2]**.

El **control de congestionamiento** requiere de políticas de prevención,de supervisión y de corrección que incluyen actividades en las capas de transporte, de red y de enlace, tales como **conformación de tráfico**, **vigilancia de tráfico**, etc**[2]**.

Algunos **algoritmos** de **control de flujo** pueden incorporarse al *hardware* de la interfaz, o al sistema operativo de la estación ó host de manera de moderar las ráfagas de los hosts y reduciendo en buena medida las posibilidades de congestionamiento. Uno de ellos fue propuesto por Turner(1986) y lo describiremos brevemente a continuación.

**Algoritmo de tobo por goteo**

Vamos a imaginar un tobo(balde ó cubeta)con un pequeño agujero en el fondo, tal como muestra la **Figura 5.10a**, sin que importe la rapidez con que el agua entra al tobo el flujo de salida tiene una tasa constante(supongamos que es así y disculpen los especialistas en dinámica de fluídos) cuando hay agua en el tobo, y cero cuando el tobo está vacío.

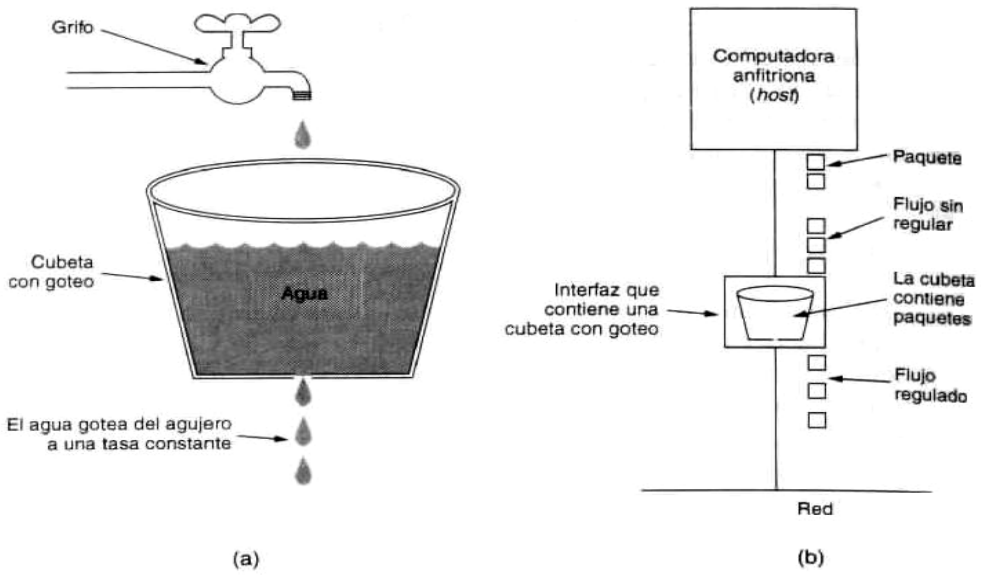



**Figura 5.10 (a)Tobo con goteo,lleno agua,(b)Tobo con goteo,lleno de paquetes.**

Si el tobo está lleno y sigue entrando agua con un flujo superior al de salida el agua comienza a derramarse por los costados y se pierde(es decir no aparece en el flujo por debajo del agujero).

Este mismo concepto se aplica a los paquetes,como indica la **Figura 5.10b**,cada estación ó host está conectado a la red mediante una interfaz que contiene un **tobo de goteo**,para decirlo con más precisión una **cola interna finita**,si llega un paquete cuando la cola está llena simplemente el paquete se descarta,como se dijo este mecanismo,que se llama **algoritmo de tobo por goteo(leaky bucket algorithm)**se incorporará al hardware del interfaz ó al sistema operativo del host.Además el host puede poner en la red un paquete por pulso de reloj,lo que convierte el flujo desigual de paquetes de los procesos de usuario en un flujo contínuo de paquetes hacia la red, regulando las ráfagas y controlando el congestionamiento.

Cuando los paquetes son del mismo tamaño(caso ATM)al algoritmo funciona como se describió,cuando son de tamaño variable lo mejor es transmitir un número fijo de bytes por pulso de reloj, si la regla es 1024 bytes por pulso, podemos transmitir un paquete de 1024 bytes, ó dos de 512 bytes, ó cuatro de 256 bytes, etc, si el número de bytes residuales es demasiado bajo, el siguiente paquete debe esperar al siguiente pulso.

Este algoritmo impone un flujo de salida rígido a la tasa promedio sin importar las ráfagas que tenga el tráfico, en muchas aplicaciones es mejor permitir que la salida se acelere un poco al llegar ráfagas grandes, el **algoritmo de tobo con fichas** atiende a esa idea y lo hace permitiendo que un host inactivo acumule **fichas**, esto es permisos para enviar grupos de datos, en un número n que cubre hasta el tamaño máximo del tobo, de manera que pueden enviarse ráfagas de hasta n paquetes,lo que produce irregularidades en el flujo de salida consecuencia de ráfagas en el flujo de entrada**[2].**

### 5.6.-Introducción a la Teoría de Colas de Espera [1].

Una de las teorías que inicialmente se ha utilizado para efectuar análisis cuantitativos de las redes de computadoras es la teoría de colas de espera que fue aplicada inicialmente para el análisis estadístico de los sistemas de conmutación telefónica y luego ha sido aplicada a resolver problemas de redes de comunicaciones.

Como explica el Apéndice 5A han surgido otras teorías cónsonas con un tráfico con características de ráfagas más que de flujo contínuo,sin embargo la aplicación de la teoría de colas de espera permite describir el comportamiento de muchas redes(entre ellas las LANs que veremos en el siguiente Capítulo)y tener un “feeling” de lo que sucede cuando hay colas de espera.El texto clásico es el descripto con **[3]**aún cuando **[1]** trae un resumen en el esta basado parcialmente lo que sigue.La bibliografía señalada con **[7]** y **[8]** dá un enfoque de última hora en este tema.

#### 5.6.1.-Sistema de colas.

El sistema de colas de espera se utiliza para modelar procesos en los cuales los clientes:

- **van llegando**
- **esperan** su turno para recibir servicio
- **reciben el servicio**
- **se marchan**

Este sistema se encuentra en las colas de los supermercados,bancos,expendio de entradas de cualquier tipo(cine,teatro,beisbol.futbol,basquet),en la sala de espera de los consultorios médicos,etc.Estos sistemas de colas de espera pueden definirse mediante cinco **parámetros**:

1. La función **densidad de probabilidad del tiempo entre llegadas.**
2. La función **densidad de probabilidad del tiempo de servicio.**
3. El **número de servidores.**
4. La **disciplina de ordenamiento** de la cola.
5. El **tamaño máximo** de las colas.

Debe hacerse notar que estamos considerando sistemas con un **numero infinito de clientes**,lo que significa que el hecho de tener un gran número de clientes en cola esperando ser atendidos **no afecta** la velocidad de entradas.

Los **parámetros** enumerados requieren una explicación:

- La **densidad de probabilidad del tiempo entre llegadas**,describe la distribución estadística del tiempo entre llegadas,la situación es simple: estamos ante un proceso aleatorio**[4]**,un observador registra el tiempo transcurrido desde la llegada previa $t_1$, $t_2$ ,$t_3$, $t_4$, etc,estos tiempos se clasifican por ejemplo:tiempo entre llegadas 0.1 seg. $n_1$, tiempo entre llegadas 0.2 seg. $n_2$ ,y así sucesivamente.Si el número total de llegadas es n, $n_i$/n será la **frecuencia relativa** del evento respectivo(i),si realizamos este proceso muchas veces las frecuencias relativas de cada evento tienden a agruparse alrededor de un valor llamado **probabilidad** de ese evento,que como los eventos se han hecho **discretos** es la **densidad discreta de probabilidad[4]**que hemos llamado aquí **densidad de probabilidad del tiempo entre llegadas.**

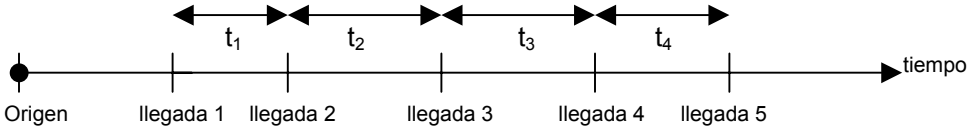



**Figura 5.11**

- Análogamente existirá una **densidad de probabilidad del tiempo de servicio**,ya que el tiempo de servicio de cada usuario es variable,por ejemplo en un supermercado un cliente solo compra un pan,mientras que otro trae ante la cajera un "carrito" lleno de productos,lógicamente los tiempos de servicio serán muy diferentes.

- El **número de servidores** es lo que esa frase indica,sin embargo debe hacerse notar que existen dos modalidades que pueden observarse en los bancos,en algunos hay una sola cola y la persona que encabeza la cola se dirige al primer cajero que se libera,este sistema se llama **de cola multiservidor**,en cambio en otros bancos cada cajero tiene su propia cola,será un sistema **monoservidor** en cada caja.

- La **disciplina de ordenamiento** de la cola se refiere a como son tomados los clientes de la cola de espera,en la sala de emergencia de un hospital los enfermos más graves son atendidos primero,en un mercado el primero en llegar a la caja es el primero en ser atendido.

- El **tamaño máximo** de las colas resalta el hecho de que no todas las colas son infinitas,sucede que cuando en una cola hay un número finito de puestos de espera,si llegan más clientes que puestos algunos son rechazados.

En lo que sigue nos ocuparemos de sistemas: **con un solo servidor,** en el que **el primero en llegar es el primero en ser atendido** y de **capacidad infinita** clientes.

En la literatura referente a este tema se usa la nomenclatura A/B/m, donde **A** es el tipo de densidad de probabilidad del tiempo entre llegadas, **B** es la densidad de probabilidad del tiempo de servicio y **m** el número de servidores.A y B son escogidas de entre:

**M** densidad de probabilidad exponencial(M significa "memoryless" ó Markov).
**D** tiempo fijo de llegada ó de servicio( D es por deterministico)
**G** general,probabilidad arbitraria.

Aquí consideraremos solamente el caso M/M/1,que tiene solución analítica exacta.

Vamos a considerar el caso de la **Figura 5.12** donde la cola es de clientes que pueden ser **paquetes de datos** ó **llamadas** en un sistema telefónico de conmutación de circuitos.

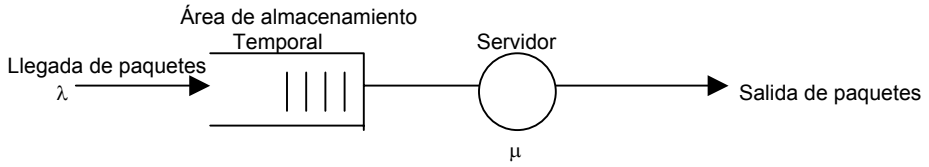



**Figura 5.12 Cola con un solo servidor.**

Los paquetes van llegando al área de almacenamiento temporal ó zona de espera con una **velocidad media** de $\lambda$ paquetes por unidad de tiempo(paq./seg),el servidor va tomando los paquetes en el orden que llegaron y los atiende con una velocidad **media** de $\mu$ paquetes por unidad de tiempo(paq./seg.).Como estamos aplicando esto a redes el **servidor** será el **medio de transmisión**:enlace,línea ó troncal de salida que transmite datos a una velocidad establecida de C bits/seg(la unidad también puede ser caracteres/seg.),si la longitud del paquete(ó longitud media de un grupo de paquetes)es L bits/paquete, tendremos que $\mu$,capacidad de transmisión en paquetes/seg. es:

$$\mu = C/L \quad \text{(paq./seg)} \qquad 5.1$$

En telefonía conmutada $\lambda$ sería el promedio de llamadas efectuadas,en llamadas/seg,y $\mu$ el número de llamadas procesadas por segundo,en llamadas/seg, con lo que $1/\mu$ es la duración promedio de una llamada en segundos.

Recuerdese que tanto $\lambda$ como $\mu$ son valores **medios**, la cola se forma porque la velocidad de llegadas (y la de servicio eventualmente)tiene una distribución estadística.

Denominaremos **intensidad de tráfico** $\rho$ a:

$$\rho = \lambda/\mu \qquad 5.2$$

Si $\lambda$ es mayor que $\mu$ el sistema se bloquea,si $\lambda$ tiende a $\mu$ el sistema se torna inestable,y si $\lambda$ es menor que $\mu$ tenemos estabilidad.

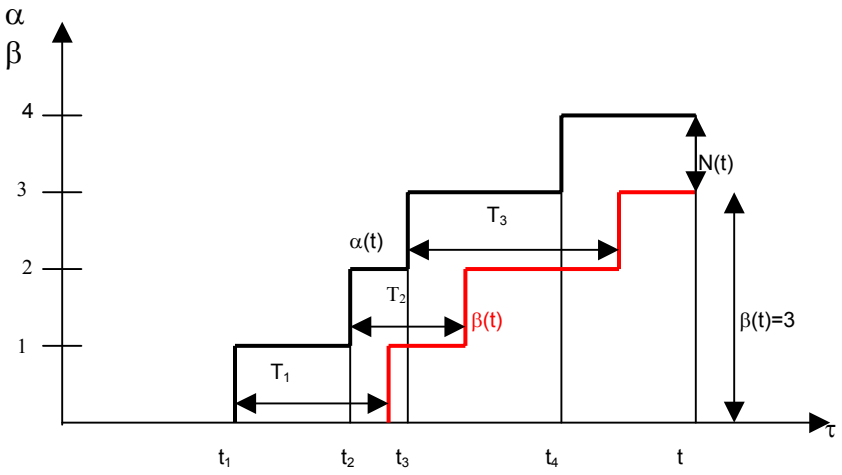
### 5.6.2.-Fórmula de Little.

Dos parámetros importantes en el análisis de colas son,**el tiempo promedio de permanencia de los paquetes en el sistema**,que denominaremos **T (**en **negrita**),y el **número promedio de paquetes en el sistema**,que llamaremos **N** (en **negrita**)**,**nuestro propósito en esta sección es relacionar ambos parámetros.

Para establecerla representemos en la **Figura 5.13,** $\alpha$**(t)** ,que es el **número de paquetes que han llegado** en el intervalo 0 – t,y $\beta$**(t)** ,que es el **número de paquetes que han salido** en el intervalo 0 – t.

Obviamente N(t) = $\alpha$(t) - $\beta$(t) es el **número de paquetes en el sistema**.




$t_i$ tiempo de llegada del paquete i
$T_i$ tiempo de permanencia del paquete i en el sistema

**Figura 5.13.-Diagrama para la demostración de la fórmula de Little**

El área entre los dos diagramas escalonados correspondientes a $\alpha$(t) y a $\beta$(t) viene dada por:

$$\int_0^t N(\tau)d\tau = \sum_{i=1}^{\beta(t)} Ti + \sum_{i=\beta(t)+1}^{\alpha(t)} (t - ti) \quad (5.3)$$

Y el **número promedio de paquetes en el sistema entre 0 y t**, es esa área dividida por t,o sea:

$$Nt = \frac{\int_0^t N(\tau)d\tau}{t} = \frac{\alpha(t)}{t} \cdot \frac{\sum + \sum}{\alpha(t)} \quad (5.4)$$

En la última parte de la expresión 5.4 el primer factor es la **velocidad media de llegadas** $\lambda$ entre 0 y t, mientras que el segundo corresponde al **tiempo promedio de permanencia de los paquetes en el sistema** $\mathbf{T}_t$,también entre 0 y t,en consecuencia si suponemos que el sistema está en estado estacionario $\mathbf{N}_t$ pasa a ser **N** y $\mathbf{T}_t$ se convierte en **T**,con lo que tenemos:

| **N = $\lambda$ T** | Fórmula de Little | 5.5 |
|---|---|---|

**5.6.3.-Aplicación al tráfico telefónico.**

El dimensionamiento de los recursos de la red telefónica conmutada se realiza aplicando la Teoría de Colas. El tráfico ofrecido,necesario para obtener resultados prácticos a partir de la expresiones obtenidas,se estima a partir de la ocupación de los diferentes equipos.

Si $\lambda_a$ es la **velocidad media de llegada de llamadas**(telefónicas)en llamadas/seg y si durante un tiempo T **de observación**, son atendidas todas esas llamadas significa que han sido atendidos $\lambda_a$ T **clientes**. Dado que cada uno de ellos ocupa el servidor que le atiende un **tiempo promedio** E (en segundos), se denomina **tiempo de ocupación del servidor** $\tau$, también en segundos, a:

$$\tau = \lambda_a T E \qquad 5.6$$

Si tenemos un grupo M de servidores y sumamos los tiempos de ocupación de cada uno de ellos tendremos el **volúmen de tráfico** $T_M$ (en segundos), de ese grupo de servidores:

$$T_M = \sum \tau_i \qquad 5.7$$

Se denomina **intensidad de tráfico** $I_T$, cursado por un grupo M de servidores, al cociente el volúmen de tráfico y el tiempo T respectivo de observación:

$$I_T = T_M / T \qquad 5.8$$

La intensidad de tráfico es adimensional pues es, para un grupo de servidores, el cociente entre los **segundos ocupados** en ese grupo de servidores y los **segundos que duró la observación**(segundos observados), y para clarificar de que estamos hablando de intensidad de tráfico se utiliza el **erlang**, o sea se habla de 0.6 erlangs para decir que tenemos una intensidad de tráfico de 0.6.

El tráfico telefónico ofrece notables variaciones:

- Horarias, a lo largo del día.
- Estacionales, según el día de la semana, semana del mes, mes del año, etc.
- Accidentales acontecimientos locales, catástrofes,etc.

En la práctica se suele hablar de la **hora cargada HC**, que se define como el período de 60 minutos consecutivos durante los cuales el volúmen de tráfico, dado por la 5.7, es máximo, la forma de establecerla es tomar lecturas cada 15 minutos y observar la suma máxima de cuatro lecturas consecutivas, el promedio a lo largo del año de esas horas cargadas dá la **hora cargada de referencia**.

El **tiempo ó período de observación** T mencionado no necesariamente es medido en segundos, puede medirse en minutos u horas, sin embargo en la práctica es frecuente oir hablar de períodos de observación de 120 segundos que conduce a una unidad de medida de **volúmen de tráfico** llamada **LLR**(**LL**amada **R**educida), también hay otra llamada **CCS**(**C**entum **C**all **S**econds) en la que el período de observación es 100 segundos.

Por ejemplo si un grupo de servidores ha cursado 2500 llamadas en una hora(pudiera ser la hora cargada, HC) y la ocupación media de cada llamada es 12 segundos,tendremos un **volúmen de tráfico**, según la 5.6 de:

$$T_M = 2.500 \times 12 = 30.000 \text{ seg.} = 250 \text{ LLR} = 300 \text{ CCS}$$

Si efectivamente los datos corresponden a la hora cargada la **intensidad de tráfico** $I_T$ de la 5.8 es:

$$I_T = 30.000/3600 = 8.33 \text{ erlangs} = 250 \text{ LLR}/1 \text{ HC} = 250 \text{ LLR/HC}$$

La intensidad de tráfico media observada depende notablemente de la fuente, la **Tabla 5.2** dá algunos valores típicos.

| Categoría | Tráfico |
|---|---|
| Abonado particular | 0.01-0.04 erlangs |
| Abonado comercial | 0.03-0.06 erlangs |
| Centralita automática | 0.10-0.60 erlangs |
| Teléfono público | 0.05-030 erlangs |

**Tabla 5.2 Intensidades de tráfico media típicas.**

### 5.6.4.-Procesos de Poisson.Colas M/M/1.

Es frecuente modelar la llegada de paquetes a que nos referíamos en relacíon a la **Figura 5.12** como **Procesos de Poisson**,estos procesos están íntimamente relacionados con la estadística exponencial.Esto significa que consideraremos el caso más sencillo de sistema de colas que es el M/M/1,que tiene procesos de llegada de Poisson y tiempos de servicio con distribución exponencial.

Un proceso de Poisson tiene tres postulados básicos:

- La probabilidad de una llegada en el intervalo $\Delta t$ se define como $\lambda\Delta t + 0(\Delta t)$ , $\Delta t << 1$,siendo $\lambda$ una constante de proporcionalidad especificada,**la velocidad media de llegadas.**
- La probabilidad de cero llegadas en $\Delta t$ es $1 - \lambda\Delta t + 0(\Delta t)$
- Las llegada son procesos **sin memoria**,ó sea cada llegada en un intervalo de tiempo es independiente de eventos en intervalos previos ó futuros,un proceso de Poisson se ve como un caso especial de un proceso de Markov.

Con estos postulados es posible demostrar**[1],[4],[5]** que la probabilidad de que lleguen n paquetes(clientes) durante un intervalo de longitud t está dada por:

$$\mathbf{P}_n(t) = [(\lambda t)^n/n!]\, e^{-\lambda t} \qquad 5.9$$

Con este resultado se demuestra **[1]** que la densidad de probabilidad de tiempo entre llegadas a que se refiere la **Figura 5.11** es de tipo exponencial.

En lo que hace al tiempo de servicio es más dificil de justificar que es un proceso de Poisson,sin embargo la función exponencial se aplica pues es la única en la que el tiempo que el cliente lleva recibiendo servicio no importa para el tiempo de servicio restante.

Lo que aquí se ha mencionado es solo para que el lector tenga idea del tipo de problema de que se trata y si le interesa profundizar en este importante tema acuda a la bibliografía respectiva,ahora pasaremos a un tópico de aplicación.

**5.6.5.-La cola de espera M/M/1 en equilibrio.**

El estado de un sistema de colas de espera M/M/1 viene descripto totalmente por la cantidad de clientes que se encuentran en el sistema,contando tanto los que están en la cola como el que se encuentra en servicio,la **Figura 5.14** ilustra la situación.

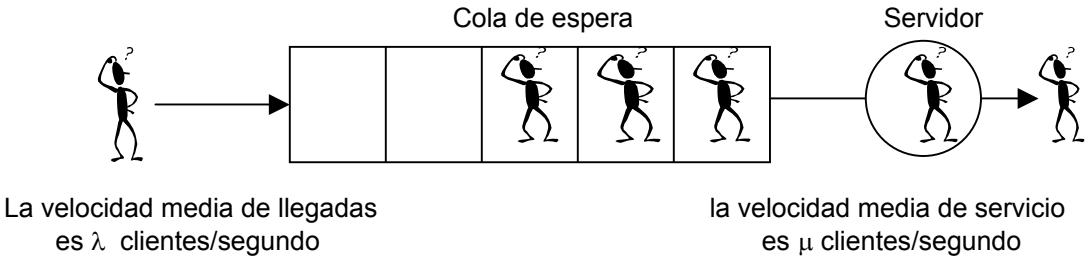



**Figura 5.14.Sistema de colas de espera con un cliente en servicio y tres clientes en cola.**

No es necesario suministrar el tiempo que lleva el cliente en servicio dado que la densidad exponencial no tiene memoria,la probabilidad de que resten t segundos de servicio es independiente del tiempo que lleva siendo servido.

Aún cuando el sistema esté **en equilibrio** pueden ocurrir transiciones,si el sistema se encuentra en el estado 4 ilustrado en la **Figura 5.14**,de tres clientes en cola y uno en servicio, puede ocurrir que llegue un nuevo cliente y pasemos al estado 5,de manera similar cuando un cliente recibe el servicio el sistema se mueve al estado adyacente inferior.

Supongamos que $p_k$ sea la probabilidad **en equilibrio** de que existan exactamente k clientes en el sistema(cola + servidor),si la velocidad media de llegadas de clientes es $\lambda$, el número de transiciones del estado k al k+1 será $\lambda p_k$.

Análogamente si el servidor procesa $\mu$ clientes por segundo la velocidad de transición del estado k+1 al estado k será $\mu p_{k+1}$ , cuando hay **equilibrio** la probabilidad de encontrar al sistema en un estado determinado no varía con el tiempo,en particular la probabilidad de que existan más de k clientes en el sistema es constante,la transición k a k+1 la aumenta y la de k+1 a k la disminuye,por lo tanto estas transiciones deben ocurrir a la misma velocidad,de otro modo el sistema no estaría en equilibrio,la **Figura 5.15** ilustra la situación.

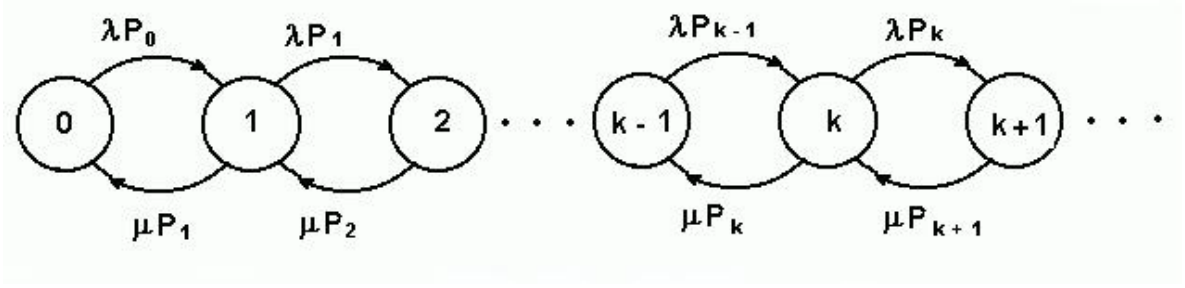



**Figura 5.15.Diagrama de estados de una cola de espera de un solo servidor.**

De la conclusión precedente y de la observación de la **Figura 5.15** concluímos que será:

$$\lambda p_0 = \mu p_1 \qquad 5.10a$$

$$\lambda p_1 = \mu p_2 \qquad 5.10b$$

y en general $$\lambda p_k = \mu p_{k+1} \qquad 5.10c$$

Con la cadena de expresiones 5.10 podemos vincular $p_k$ con $p_0$, recordando que en la expresión 5.2 se definio la intensidad de tráfico(que no tiene nada que ver con la de telefonía de la subsección anterior) $\rho = \lambda/\mu$ resulta,

$$\boxed{p_k = \rho^k p_0} \qquad 5.11$$

Es claro que la sumatoria de todas las probabilidades debe dar 1,o sea:

$$\sum_{i=1}^{infinito} p_i = \sum_{k=1}^{infinito} \rho^k p_0 = 1 \qquad 5.12$$

Como $p_0$ sale fuera del signo de sumatoria,tenemos el caso una suma de serie geométrica,cuyo resultado es bien conocido,

$$\sum_{k=1}^{infinito} \rho^k = 1/(1 - \rho) \qquad 5.13$$

Por lo tanto de 5.12 y 5.13 resulta la expresión de $p_0$,

$$p_0 = 1 - \rho$$

Y con ello resulta aplicando la 5.11 que es,

$$p_k = (1 - \rho)\, \rho^k \qquad 5.14$$

Calculemos ahora el número **promedio** de clientes en el sistema **N** (obsérvese que está en negrita,dado que es un valor medio) basándonos en las probabilidades de estado,

$$\mathbf{N} = \sum_{k=1}^{infinito} k\, p_k = (1 - \rho) \sum_{k=1}^{infinito} k\, \rho^k \qquad 5.15$$

La última sumatoria puede ser calculada observando que es posible derivar ambos miembros de la 5.13 respecto de $\rho$, con lo que obtenemos para **N**,**número promedio de clientes en la cola(paquetes en la cola)**,

$$\boxed{\mathbf{N} = \rho/(1-\rho)} \qquad 5.16$$

Con la Fórmula de Little(expresión 5.5)obtenemos el **tiempo promedio de permanencia de clientes en la cola(paquetes en el sistema) T**,

$$\boxed{\mathbf{T} = \mathbf{N}/\lambda = 1/(\mu - \lambda)} \qquad 5.17$$

Las **Figuras 5.16 y 5.17** muestran la variación de **N** y **T** respectivamente,con $\rho$.

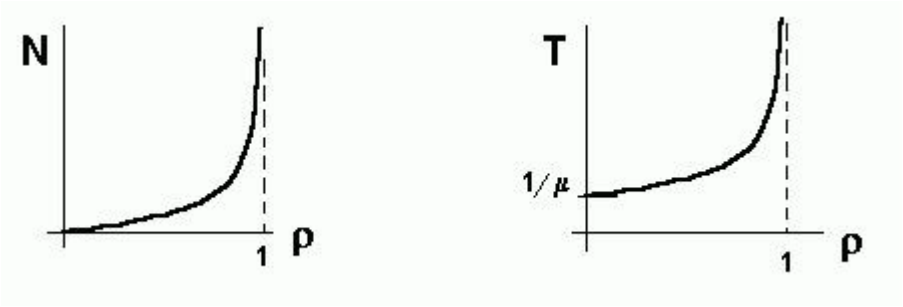



**Figura 5.16.N** en función de $\rho$ **Figura 5.17.T** en función de $\rho$

Estos resultados son importantes en la cuantificación del retardo y numero de clientes en cola de diversos tipos de sistemas. Otros Casos han sido resueltos como el M/G/1 **[1]** y el M/M/1 con cola finita **[6].**

**5.6.6.-Aplicaciones de la Teoría de Colas a las Redes de Comunicaciones.**

El desarrollo efectuado para colas M/M/1 puede ser aplicado al problema de determinar el retardo experimentado por los paquetes en la cola de un nodo de conmutación(ó IMP), sin embargo debemos cambiar ligeramente la notación ya que en un nodo la velocidad de servicio,que es la velocidad de la línea C(bits/seg)es fija, lo que varía estadísticamente es la longitud de los paquetes.

En este caso la función densidad de probabilidad para paquetes de tamaño x (en bits) suponemos que es exponencial, $\nu e^{-\nu x}$ con una media de $\nu$ paquetes/bit, **el valor medio de la longitud de los paquetes** lo llamaremos L en bits/paquete y vale $1/\nu$, la **velocidad media de servicio** pasa a expresarse como $\mu$ = C/L en paquetes/seg..

La expresión 5.17 toma la forma:

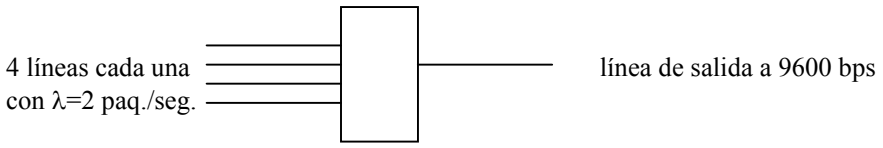
$$\mathbf{T} = 1 / [ (C/L) - \lambda ) ] \qquad 5.18$$

**T** incluye **tanto el tiempo en la cola como el de servicio**,como puede comprobarse al hacer tender $\lambda$ a cero.

Veamos dos ejemplos interesantes dados en **[1],**

**<u>Concentrador de terminales</u>**

Consideremos un concentrador de terminales con cuatro líneas de entrada de 4800 bps pero que entregan una tráfico de Poisson $\lambda_i$ = 2 paquetes/segundo,siendo la longitud media de los paquetes de 1000 bits,y con una línea de salida de 9600 bps.
Se desea conocer **T** el tiempo medio de los paquetes en el concentrador(es el retardo medio de ese concentrador) y **N** ,el número medio de paquetes en el sistema.



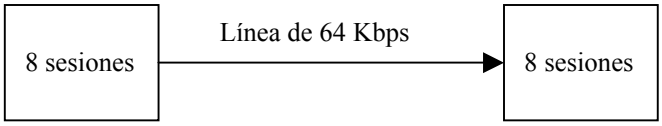

 

Utilizando la expresión 5.18 con:C = 9600 bps,L = 1000 bits/paquete y $\lambda_i = 2$ paquetes/seg., que como son cuatro líneas produce un $\lambda$ de 8 paquetes/seg, obtenemos $\mu$ = 9.6 paquetes/seg., y un **T** de 625 mseg.

Análogamente mediante la 5.16,teniendo en cuenta que $\rho$ resulta 8/9.6 ,resulta **N=**5 paquetes.

**Canales dedicados ó compartidos?**

Tenemos dos máquinas(computadores)conectados con una línea de 64 kbps,hay ocho sesiones(half-duplex) en cada máquina que funcionan al mismo tiempo(paralelamente),cada sesión genera un tráfico de Poisson con un $\lambda_i = 2$ paquetes/seg y paquetes de longitud media 2000 bits/paquete.




Se desea saber si es preferible dar a cada sesión un canal de 8 Kbps (mediante FDM ó TDM) ó bien dejar que todos los paquetes compitan para utilizar el canal compartido de 64 Kbps.

Primera opción: dar a cada sesión un canal de 8 Kbps.

Cada canal tiene un $\lambda_i$ = 2 paquetes/seg,el $\mu_i$ = 4 paquetes/seg. lo que conduce a :

**T** = 500 mseg **N** = 1 paquete **en cada uno de los 8 canales.**

Segunda opción: competencia por la línea de 64 Kbps.

En este caso $\lambda$ = 16 paquetes/segundo $\mu$ = 32 paquetes/segundo lo que conduce a:

**T** = 62.5 mseg **N** = 1 paquete

Se concluye entonces que es mucho más eficiente la segunda opción.Obsérvese que **T** del primer caso es **justamente** 8 veces(el número de canales)el **T** del segundo caso,es sencillo demostrar que esa es la regla **general** entre FDM (ó TDM) y el canal compartido.

El tema es muy amplio varios autores dan muchos ejemplos más,a ellos remitimos al lector interesado.

## BIBLIOGRAFÍA

**[1] Tanenbaum Andrew**, "Redes de Ordenadores", Segunda Edición, Prentice Hall Hispanoamericana.

**[2] Tanenbaum Andrew**, "Redes de Computadoras", Tercera Edición, Prentice Hall Hispanoamericana.

**[3] Kleinrock ,**"Queueing Systems", Volume 2, Computer Applications,John Wiley & Sons, 1976.

**[4] Thomas John B.**,"An Introduction to the STATISTICAL COMMUNICATION THEORY", John Wiley & Sons.

**[5] Papoulis A.**,"Probability,Random Variables,and Stochastic Processes" , 2da edition, Mc Graw-Hill.

**[6] Schwartz Mischa,**"Redes de Comunicaciones",Addison Wesley Iberoamericana.

**[7] "Congestion Control Mechamisms and the Best Effort Service Model"**, Gevros Panos,Crowcroft J,Kirstein P., Bhatti S., IEEE Network, May-June 2001.

**[8] "Traffic Theory and the Internet",** Roberts Jim, IEEE Communications Magazine, January 2001.

**APÉNDICE 5A:**

CARACTERÍSTICAS DEL TRÁFICO

William Stalling: "Self-similarity upsets data traffic assumptions", IEEE Spectrum, January 1997.

In 1993, the field of network performance modeling was rocked by a group of Bellcore and Boston University researchers who delivered a paper at that year's SIGCOMM (Special Interest Group on Data Communications) conference: "On the Self-Similar Nature of Ethernet Traffic," which appeared the following year (February 1994) in the IEEE Transactions on Networking and which is arguably the most important networking paper of the decade.

Although a number of researchers had observed over the years that network traffic did not always obey the Poisson assumptions used in queuing analysis, the paper's authors for the first time provided an explanation and a systematic approach to modeling realistic data traffic patterns.

Simply put, network traffic is more bursty and exhibits greater variability than previously suspected. The paper reported the results of a massive study of Ethernet traffic and demonstrated that it has a self-similar or fractal, characteristic. That means that the traffic has similar statistical properties at a range of time scales: milliseconds,seconds, minutes, hours, even days and weeks. This has several important consequences. One is that you cannot expect that the traffic will "smooth out" over an extended period of time; instead, not only does the data cluster but the clusters cluster. Another consequence is that the merging of traffic streams, such as is done by a statistical multiplexer or an asynchronous-transfer mode (ATM) switch, does not result in a smoothing of traffic. Again, multiplexing bursty data streams tends to produce a bursty aggregate stream.

One practical effect of self-similarity is that the buffers needed at switches and multiplexers must be bigger than those predicted by traditional queuing analysis and simulations. Further, these larger buffers create greater delays in individual streams than originally anticipated.

Self-similarity is not confined to Ethernet traffic or indeed to local-area network traffic in general. The 1994 paper sparked a surge of research in the United States, Europe, Australia, and elsewhere. The results are now in: self similarity appears in ATM traffic, compressed digital video streams, Signaling System Seven (SS7) control traffic on networks based on the integrated-services digital network (ISDN), Web traffic between browsers and servers, and much more.

The discovery of the fractal nature of data traffic should not be surprising. Such self-similarity is quite common in both natural and man-made phenomena; it is seen in natural landscapes, in the distribution of earthquakes, in ocean waves, in turbulent flow, in the fluctuations of the stock market, as well as in the pattern of errors and data traffic on communication channels.

The implications of this new view of data traffic are startling and reveal its importance. For example, the whole area of buffer design and management requires rethinking. In traditional network engineering, it is assumed that linear increases in buffer sizes will produce nearly exponential decreases in packet loss and that an increase in buffer size will result in a proportional increase in the effective use of transmission capacity. With self-similar traffic, these assumptions are false. The decrease in loss with buffer size is far less than expected,and a modest increase in utilization requires a significant increase in buffer size.

Other aspects of network design are also affected. With self-similar traffic, a slight increase in the number of active connections through a switch can result in a large increase in packet loss. In general, the parameters of a network design are more sensitive to the actual traffic pattern than expected. To cope with this sensitivity,designs need to be more conservative. Priority scheduling schemes need to be reexamined. For example, if a switch manages multiple priority classes yet does not en force a bandwidth limitation on the class with the highest priority, then a prolonged burst of traffic from the highest prioritycould keep the other classes from using the network for an extended periodof time.

The explanation for these strange results is that surges in traffic tend to occur in waves. There may be a long period with very little traffic followed by an interval of heavy usage in which traffic peaks tend to cluster, making it difficult for a switch or network to clear the backlog from one peak before the next peak arrives. Therefore, a static congestion control strategy must assume that such waves of multiple peak periods will occur. A dynamic congestion control strategy is difficult to implement. Such a strategy is based on the measurement of recent traffic, and it can fail utterly to adapt to rapidly changing conditions.

Congestion prevention by appropriate sizing of switches and networks is difficult because data network traffic does not exhibit a predictable level of busy period traffic; patterns can change over a period of days, or weeks, or months. Congestion avoidance by monitoring traffic levels and adapting flow control and traffic routing policies is difficult because congestion can ocur unexpectedly and with dramatic intensity.

Finally, congestion recovery is complicated by the need to make sure that critical network control messages are not lost in the repeated waves of traffic hitting the network.

The reason that this fundamental nature of data traffic has been missed up until recently is that it requires the processing of a massive amount of data over a long observation period to detect and confirm this behavior. And yet the practical effects are all too obvious. ATM switch vendors, among others, have found that their products do not perform as advertised once in the field, because of inadequate buffering and the failure to take into account the delays caused by burstiness.

The true nature of high-speed data traffic has now been revealed.
As yet, a consensus on a valid and efficient set of mathematical tools for modeling and predicting such traffic has not emerged. This will be the next step in this important area of research.

William Stallings (SM) is a consultant, lecturer and author of over a dozen books on data communications and computer networking. His "Computer Organization and Architecture" received the award for the the best computer science textbook of 1996 from the Textbook end Academic Authors Association. His latest book is "Data and Computer Communications", Fifth Edition, (Prentice-Hall, Englewood Cliffs, N.J., 1997). He can be reached at ws@shore.net.