# INTRODUCCIÓN A LOS MODELOS DE TRÁFICO PARA REDES DE BANDA ANCHA

*David Rincón Rivera*

*Estudiante de doctorado*
*Departament de Matemàtica Aplicada i Telemàtica (UPC)*
*drincon@mat.upc.es, http://www-mat.upc.es/~drincon*

## 1. INTRODUCCIÓN

La necesidad de disponer de redes que integren diversos servicios de telecomunicaciones motivó la aparición del concepto ISDN (RDSI, Red Digital de Servicios Integrados), que consiste en el uso de una única infraestructura para el transporte de datos, voz e imágenes. A mediados de la década de los 80, el CCITT (actual ITU-T, Unión Internacional de Telecomunicaciones) comenzó la estandarización de su modelo de RDSI de banda ancha (*Broadband ISDN*), que permitiría la transmisión de información a altas velocidades. Un punto importante en este proceso fue la elección del mecanismo de transferencia, que se podría definir como el conjunto de mecanismos de multiplexación y conmutación que usa la red. La elección recayó en ATM (*Asynchronous Transfer Mode*).

Las redes ATM permiten multiplexar eficientemente diversas fuentes. La unidad fundamental de transferencia es la celda ATM, de 53 bytes (5 de cabecera y 48 de información producida por las capas superiores). Mediante el concepto de *multiplexación estadística* de fuentes, se consigue una gran eficiencia en el uso de los recursos de la red [Led94]. Supongamos que disponemos de diversas fuentes que emiten a una tasa variable. Para la transmisión de dichas fuentes, las redes "tradicionales" reservan un ancho de banda igual a la tasa de pico, malgastando el ancho sobrante cuando la tasa era inferior. Esta ineficiencia es la que se trata de evitar en ATM, multiplexando las fuentes sobre enlaces con capacidad menor que la suma de tasas de pico. El precio a pagar es que, al ser un método estadístico, la posibilidad de que los picos coincidan no es nula. En este caso, se perderán celdas por *overflow*, con el consiguiente perjuicio en la calidad de servicio (QOS). En la figura 1 se pueden observar las diferencias entre la multiplexación en base a la tasa de pico y la multiplexación estadística.

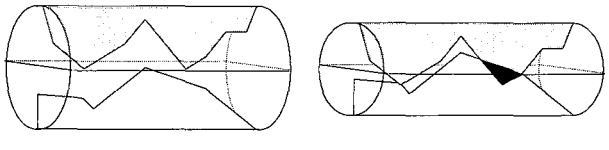

***Figura 1.*** *Izquierda: cada fuente dispone de un ancho fijo igual a su tasa de pico, malgastando recursos de transmisión. Derecha: al multiplexar estadísticamente las fuentes sobre un enlace de capacidad menor que la suma de los picos, se produce un conflicto (zona negra), pero se ahorran recursos (se ocupa menos ancho de banda).*

Para minimizar estos efectos de pérdida y mantener la QOS tanto de las nuevas llamadas como de las que ya se están cursando, se necesita disponer de controles de admisión de conexiones (CAC) que decidan si se puede admitir una nueva llamada, así como de mecanismos de control de la congestión. Para ello son necesarias matemáticas que describan el comportamiento de las fuentes; es decir, necesitamos **modelos matemáticos de tráfico**. En este artículo se intentará ofrecer una visión de algunos de los modelos de tráfico propuestos para redes de banda ancha, sin ánimo de ser exhaustiva. Se hará un breve resumen sobre las escalas temporales que se distinguen en el tráfico ATM, se presentarán los modelos de tráfico clásicos y los nuevos modelos basados en la dependencia a largo plazo. El artículo finaliza con algunas reflexiones sobre el impacto de los nuevos modelos y las implicaciones de su aparición en el campo del modelado de fuentes.

## 2. ¿CÓMO SE COMPORTAN LAS FUENTES?

Para llegar a generar buenos modelos, es necesario estudiar y entender el comportamiento estadístico de las fuentes de información multimedia (voz, audio y vídeo) que serán transportadas por las redes ATM.

**Fuentes de datos**: los generadores de datos más habituales serán los servidores de Web, las transferencias de ficheros (FTP), las conexiones remotas (Telnet) y los *chats* (IRC). Es decir, casi todos los servicios presentes hoy por hoy en las redes TCP/IP (Internet). Estos servicios tienen comportamientos muy diferentes. El web, por ejemplo, se distingue por presentar una gran cantidad de transferencias, habitualmente de reducido tamaño (a menos que se esté navegando por una página con ficheros de sonido e imágenes grandes). FTP, por el contrario, presenta largas transferencias de información.

**Fuentes de audio**: lo más habitual es que la señal transmitida sea de voz, por lo que se aprovechan las características del habla humana, que se presenta a ráfagas (*talk spurts*) con silencios intercalados entre palabras y entre frases. Por ello lo más habitual es que los codificadores de voz (GSM, por ejemplo) incorporen detectores de silencios, durante los cuales no transmiten información.

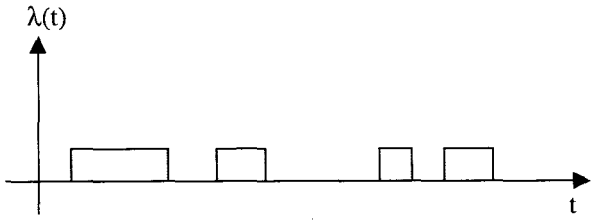


***Figura 2.*** *Posible traza de una fuente de voz.*

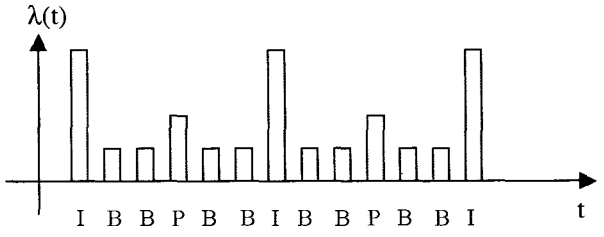


***Figura 3.*** *Posible traza de un codificador*

**Fuentes de vídeo de tasa variable (VBR)**: son muy dependientes del tipo de codificador usado. En general, presentan una cantidad no despreciable de correlaciones, a diferentes niveles (línea, bloque, cuadro, escena). Los codificadores más modernos, como los MPEG 1 y 2, aprovechan la alta redundancia espacial y temporal de las señales de vídeo para conseguir factores de compresión elevados. Por ello las señales de vídeo de estos codificadores son de tasa variable, y repiten un esquema en el que aparecen imágenes codificadas en el modo *intra* (que contienen toda la información del cuadro, denominadas *I*), e imágenes codificadas a partir de las *intra* por predicción (hacia delante o hacia atrás, denominadas *P* y *B*). Los cuadros tipo *intra* son los que provocan picos en la tasa de transmisión, ya que contienen mucha más información que los cuadros *P* y *B*.

## 3. MODELOS DE TRÁFICO PARA REDES ATM

Como hemos comentado, la principal característica de las redes ATM (y el principal quebradero de cabeza para los implementadores) es su capacidad de multiplexar estadísticamente las fuentes. Otra peculiaridad importante es la jerarquía temporal de tres niveles que presentan sus transmisiones [COST242].

Las redes ATM son **orientadas a conexión**. Eso quiere decir que cada vez que se quiere transferir información se debe establecer un circuito virtual desde el terminal origen al destino. Dentro de esta llamada o sesión, se producen ráfagas (*bursts*) en los que se transmiten celdas, seguidos de intervalos de silencio. A un nivel todavía menor, dentro de cada ráfaga las celdas no han de ser necesariamente transmitidas de modo uniforme. Por tanto, ATM presenta un comportamiento diferenciado en cada una de las tres escalas de tiempo consideradas: llamada, ráfaga y celda (figura 4). Esta distinción es importante si queremos generar modelos de tráfico precisos.

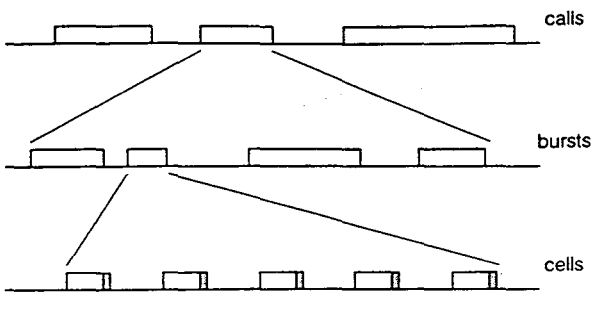


***Figura 4.*** *Escalas de tiempo en ATM [COST242].*

Las escalas de celda y ráfaga son las que nos permiten dimensionar los buffers de los conmutadores ATM, así como los mecanismos de admisión (CAC) y los de uso de parámetros (UPC). Por ello son los que han recibido más atención. Vamos a ver las características de cada uno de ellos [CasBlo]:

### Nivel de celda

- El tráfico está compuesto por entidades discretas: las celdas.
- La mayoría de las fuentes ATM son localmente periódicas a esta escala.
- La aleatoricidad del tráfico agregado proviene de la independencia de las fases de los flujos (localmente periódicos), y no de las fluctuaciones de dichos flujos.
- Las fluctuaciones a escala de celda deben ser suavizadas, por lo que se presenta un problema de dimensionado de buffers.
- La distribución de la longitud de las colas cuando el sistema trabaja próximo a su capacidad máxima es crucial para el dimensionado.

### Nivel de ráfaga (burst)

- Sólo es relevante para fuentes de tasa variable (VBR, variable bit rate), ya que si la tasa es constante (CBR, constant bit rate) no se producen ráfagas.
- Los posibles problemas provienen de la posibilidad de que el tráfico agregado exceda la capacidad de salida del sistema.
- No es importante conocer el momento exacto de llegada de las celdas.
- Esencialmente, el sistema se comporta como un sistema con pérdidas. Los *overflows* se pueden evitar mediante el uso de mecanismos de control de admisión.
- La prevención de pérdidas provocadas por ráfagas implica forzar una tasa media de entrada muy reducida para el tráfico de alta variabilidad (*bursty traffic*).

Ya que vamos a presentar diferentes modelos de tráfico, sería interesante disponer de algún criterio que permitiera valorar la bondad de cada modelo. Esta valoración vendría a medir lo "próximo" que es el modelo a una fuente real de tráfico. Los parámetros de bondad de un modelo son la capacidad de capturar la autocorrelación de las fuentes, las distribuciones marginales (las *colas* de

la distribución), y la precisión en el cálculo de retardos y probabilidad de pérdidas [Ada97].

Los modelos de tráfico pueden ser **estacionarios** o **no estacionarios**. Los estacionarios se dividen en dependientes a corto y largo plazo (**short range dependent**, **long range dependent**). Entre los primeros encontramos los procesos de Markov y y los modelos regresivos. Estos modelos presentan correlación sólo en intervalos cortos de tiempo. En cambio, los modelos *long range dependent*, como el F-ARIMA o el Movimiento Browniano Fractal presentan correlaciones incluso en intervalos de tiempo grandes.

## 4. LOS MODELOS CLÁSICOS

Vamos a denominar como *clásicos* a toda una familia de modelos cuya principal característica es que no contemplan la posibilidad de que las trazas de tráfico procedentes de las fuentes presenten dependencia a largo plazo.

¿Qué significa esto y qué implicaciones tiene de cara al modelado? La autocorrelación de una serie temporal (en nuestro caso, la serie temporal es la tasa o *throughput* de la fuente medido en bits/segundo o equivalentemente, en celdas/segundo) indica cuál es la dependencia que existe entre los datos de la serie separados en una cierta cantidad de muestras. Es decir, qué grado de similitud existe entre las muestras analizadas en un instante $t_0$ y las muestras emitidas en $t_1 = t_0 + \tau$. Si la autocorrelación de la serie generada por una fuente decae rápidamente con la separación $\tau$, las tasas emitidas por la fuente tendrán poco que ver con los emitidos $\tau$ unidades antes. Si, en cambio, la autocorrelación decae lentamente, existirá una dependencia a largo plazo.

### 4.1 Modelos para fuentes de datos

Tradicionalmente, se ha asumido que en las redes de conmutación de paquetes se daban las condiciones necesarias para suponer que la generación de celdas sigue un proceso de Poisson o de Bernouilli. Sin embargo, como ya se ha mencionado antes al mencionar diversos servicios de Internet, cada aplicación presenta una tasa y una distribución de generación de celdas diferente, que puede ir desde transmisiones esporádicas y cortas hasta largas transferencias de informacion (*bulk transfers*).

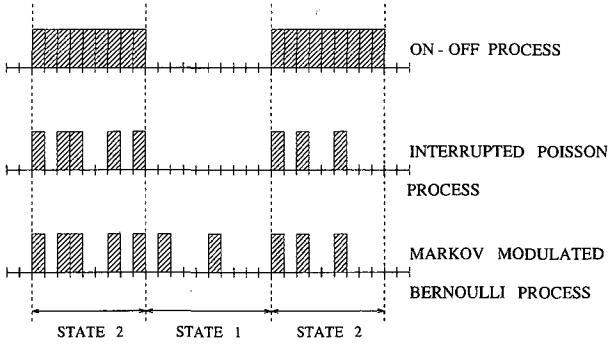


***Figura 5.*** *Modelos de tráfico a tiempo discreto para fuentes de voz y datos [Led94].*

También es un factor a tener en cuenta el hecho de que los paquetes de las redes de área local (LAN) son mucho más grandes que las celdas ATM, lo que provoca la aparición de pequeñas ráfagas de celdas ATM cada vez que una trama LAN se introduzca en la red. Por ello, los modelos más utilizados son los procesos on-off markovianos, y los interrumpidos de Poisson o Bernouilli (figura 5).

### 4.2 Modelos para fuentes de voz

#### Modelos on-off

Los procesos que más se aproximan al fenómeno *talk spurt* son los que se basan en una cadena de Markov de dos estados (*on* y *off*). Estos modelos, ampliamente utilizados, intentan describir una fuente que emite información a ráfagas, de manera que en el estado on se generan paquetes de voz y en el estado off hay silencio. Dependiendo de la distribución del proceso de llegadas cuando la fuente se encuentra en estado activo, se pueden considerar dos tipos de procesos: Bernouilli Interrumpido o Poisson Interrumpido.

Si suponemos que las llegadas siguen se producen según Poisson, el tiempo de estancia en los estados está exponencialmente distribuido con medias $\alpha^{-1}$ y $\beta^{-1}$, respectivamente. Los procesos de Poisson interrumpidos son una particularización del caso anterior, donde el estado activo se corresponde con llegadas distribuidas exponencialmente con media l. Como ejemplo, en [Led94] se presenta el caso de un codificador ADPCM a 32 Kbit/s, con tiempos de paquetización de 12 ms., $\alpha^{-1} \approx 650$ ms. y $\beta^{-1} \approx 350$ ms. La agregación de fuentes de voz se puede modelar como un proceso de nacimiento y muerte con N+1 estados, que resulta de encadenar N procesos on-off.

#### Procesos de Poisson modulados por cadenas de Markov (MMPP, Markov Modulated Poisson Processes)

Empecemos presentando los procesos de Markov modulados, también llamados procesos doblemente estocásticos, que usan una cadena de Markov que define (según el estado en el que nos encontremos) la distribución de probabilidad del tráfico [Sch96]. Es decir, la cadena "modula" el proceso de generación. Un MMPP usa un proceso de Poisson como proceso modulado, es decir, el que cambia el ritmo de generación de llegadas según el estado en el que nos encontremos (a un estado $s_k$ le corresponde un proceso de Poisson con media $\lambda_k$).

Mediante los MMPP, se puede conseguir que el modelo de tráfico sea analíticamente abordable. Para ello, se cuantiza la tasa de llegada en diferentes niveles, que se corresponden con los estados de la cadena de Markov. Las probabilidades de transición del estado i al j, $q_{ij}$, se pueden estimar a partir de los datos empíricos

mediante el cálculo de la fracción de la cantidad de veces que se produce dicha transición, respecto a la cantidad total de transiciones. Es fácil observar que un MMPP de M+1 estados se puede formar a partir de la superposición de M procesos IPP independientes e idénticamente distribuidos. Los MMPP se han aplicado en algunas ocasiones para modelar mezclas de tráfico de voz y datos. Para ello, consideramos que el MMPP modela el tráfico de voz, de manera que el estado k se corresponde con un tráfico de voz de tasa $\lambda_k$, mientras que el tráfico de datos se considera poissoniano con media $\lambda_d$. Así, el tráfico total generado es $\lambda_k+\lambda_d$.

## 4.3 Modelos para fuentes de vídeo

### Modelos de fluidos modulados por cadenas de Markov (Markov Modulated Fluid Models)

Los modelos de fluidos se caracterizan por modelar el tráfico como un flujo continuo, siendo especialmente indicados cuando las unidades de tráfico (paquetes o celdas) son muy pequeñas comparadas con el tráfico total. De cara a la simulación, estos modelos son mucho más tratables computacionalmente que los modelos que distinguen cada celda. En los modelos de fluido modulados por Markov, el estado de la cadena determina la tasa del fluido (a un estado $s_k$ le corresponde una tasa constante $\lambda_k$). Este modelo es usado en fuentes de vídeo del tipo VBR (tasa variable).

En [Mag88], la tasa continua es cuantizada a un número finito de niveles discretos (equivalente al número de estados de la cadena de Markov) y es muestreada temporalmente en puntos aleatorios escogidos según una distribución de Poisson. En este modelo, la tasa en el estado i es iA, donde A se define como el paso de cuantización (figura 6). El modelo no captura exactamente la autocorrelación de los datos empíricos, ya que sólo se permiten transiciones entre estados adyacentes, pero si tuviéramos que tomar en consideración todas las transiciones (necesarias cuando se produce un cambio brusco en la tasa, de más de un nivel), la complejidad de cálculo se haría prohibitiva.

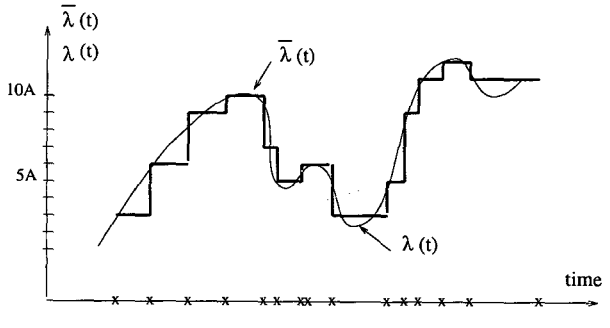


***Figura 6.*** *Muestreo de Poisson y discretización de la tasa de la fuente. Se permiten saltos de más de un nivel [Led94].*

Para mejorar la respuesta del modelo ante los saltos bruscos, en [Mag88] se propone una extensión bidimensional de la cadena, de manera que existen dos pasos de cuantificación (uno grande, $A_h$, y otro pequeño, $A_l$. Para identificar los estados del modelo, se definen unos índices i y j, $0 \le i \le M$, $0 \le j \le N$, de manera que la tasa total generada es $jA_h + iA_l$. En [Sen89] se analiza el comportamiento del modelo. Con M=1, todavía es analíticamente tratable, y presenta una autocorrelación que decae exponencialmente. A medida que se introducen nuevos niveles, el coste computacional se dispara.

### Modelos regresivos

Los modelos regresivos se caracterizan por predecir la siguiente variable aleatoria de la secuencia, a partir de los valores anteriores, que se encuentran dentro de una "ventana móvil", y de ruido blanco. A continuación se presentan algunos modelos regresivos.

#### *- Modelos autoregresivos (AR)*

El modelo autoregresivo de orden p, denominado AR(p), tiene la forma:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \varepsilon_t$$

donde $\varepsilon_t$ es ruido blanco, $\phi_i$ son números reales (coeficientes autoregresivos) y $X_t$ son los valores de la secuencia de variables aleatorias. Si $\varepsilon_t$ es ruido blanco gaussiano aditivo de varianza $\sigma_\varepsilon^2$, las $X_t$ son variables aleatorias con distribución normal. Si definimos un operador de retardo B como $X_{t-1} = BX_t$, y el polinomio $\phi(B) = 1-\phi_1 B - \ldots - \phi_p B^p$, entonces podemos expresar el proceso AR(p) como

$$\phi(B)\, X_t = \varepsilon_t$$

Se puede demostrar que la autocorrelación del proceso AR(p) será, en general, una exponencial, lo que indica que los modelos autoregresivos presentan dependencia a corto plazo, que como veremos más adelante es una característica que puede marcar su campo de aplicación [Ada97]. Los modelos AR(p) son especialmente adecuados en el caso de videoconferencia, ya que dicho servicio genera imágenes con poca variabilidad (no suele haber cambios de escena). Aunque es fácil estimar los parámetros del modelo AR, el hecho de que la autocorrelación decaiga exponencialmente no permite capturar funciones de autocorrelación que lo hagan a un ritmo menor, como es el caso de las señales de televisión digital a tasa variable (donde los cambios de escena y los picos de tasa asociados son abundantes).

#### *- Modelos autoregresivos de media móvil (ARMA)*

Los modelos autoregresivos de media móvil de orden p y q, denominados ARMA(p, q), tienen la forma:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q}$$

o alternativamente,

$$\phi(B)\, X_t = \theta(B)\varepsilon_t \quad \text{donde } \theta(B) = 1-\theta_1 B - \ldots - \theta_q B^q$$

Esto es equivalente a filtrar ruido blanco (el proceso $\varepsilon_t$) a través de un filtro lineal y causal (desplazado temporalmente), con p polos y q ceros, de la forma:

$$H(z)=\frac{B_q(z)}{A_P(z)}=\frac{1-\sum_{k=0}^{q}\theta_k z^{-k}}{1-\sum_{k=0}^{p}\phi_k z^{-k}}$$

Se demuestra que la autocorrelación de los procesos ARMA(p,q) decae exponencialmente, lo que implica que este modelo presenta dependencia a corto plazo.

Los modelos ARMA son capaces de capturar mejor que los AR los picos debidos a los cambios de escena, pero desde el punto de vista práctico, la estimación de los coeficientes $\phi_k$ implica la resolución de un sistema de ecuaciones no lineales, hecho que complica enormemente su obtención.

***- Modelos autoregresivos de media móvil integrada (ARIMA)***

Los modelos autoregresivos de media móvil integrada de orden p, d y q, denominados ARIMA(p, d, q), son una extensión de los ARMA (p,q). Se obtienen cuando se permite que el polinomio $\phi(B)$ tenga d raíces unitarias, mientras que el resto permanecen fuera del circulo unidad. La forma general es

$$\varphi(B)\,\nabla^d X_t = \theta(B)\,\varepsilon_t$$

donde $\nabla$ (operador diferencia) se define como $\nabla X_t = X_t - X_{t-1}$, y $\varphi(B)$, $\theta(B)$ son polinomios en B. Los ARIMA(p, d, q) se usan para modelar series temporales homogéneas y no estacionarias.

## 5. NUEVOS MODELOS CON DEPENDENCIA A LARGO PLAZO

Hemos visto, hasta ahora, modelos de tráfico que tenían en común que la autocorrelación decaía exponencialmente; es decir, que la dependencia de la señal consigo misma sólo se da en un margen corto de tiempo. Sin embargo, medidas realizadas sobre tráfico real han llevado a cuestionar la validez de estos modelos [Lel94], y se ha detectado que la autocorrelación decae a un ritmo menor. Por ello se han adoptado nuevos enfoques y modelos. Vamos a introducir las herramientas matemáticas necesarias para entenderlos.

Definamos $\{X_t\}$, t = 0, 1, 2... como un proceso estocástico estacionario en sentido amplio, es decir, un proceso con:

- una media estacionaria $\mu = E[X_t]$
- una varianza estacionaria y finita, $\nu = E[(X_t-\mu)^2]$
- una función de autocovarianza estacionaria, $\gamma_k = E[(X_t-\mu)(X_{t+k}-\mu)]$, que sólo depende de k y no de t.

Nótese que $\nu = \gamma_0$. Si definimos la autocorrelación de $\{X_t\}$ en el instante k como $\rho_k$, tenemos que por definición, $\rho_k = \gamma_k/\gamma_0$.

Para toda m, definimos $\{X_j^{(m)}\}$ como una nueva serie obtenida calculando la media de la serie original $\{X_t\}$ sobre bloques de longitud m, sin superposición:

$$X_j^{(m)} = (1/m)\,(X_{jm-m+1} + \ldots + X_{jm})$$

La varianza de $\{X_j^{(m)}\}$, $\nu_m$, es

$$\nu_m = E[(1/m)\,(X_{jm-m+1} + \ldots + X_{jm})]^2$$

En el caso de que el proceso sea ruido blanco, las variables aleatorias estarán incorreladas y $\rho_k = 0$ para k>0, y $\nu_m = \nu m^{-1}$. Si m (longitud de los bloques a promediar) es grande, se puede aproximar la expresión de la autocorrelación por:

$$\nu_m = \nu\left[2\sum_{k=1}^{m}\rho_k\right]m^{-1}$$

Si $\rho_k \to 0$ y $\Sigma_k \rho_k < \infty$, la varianza decae asintóticamente a cero de manera proporcional a $m^{-1}$, es decir, $\nu_m = \nu c m^{-1}$ (donde c es una constante). Procesos como los ARMA o Markov verifican esta propiedad.

Las medidas de tráfico real [Lel94, Ber95] han demostrado que la varianza $\nu_m$ decae a un ritmo menor que $m^{-1}$. Por ello se introduce un factor de corrección en el modelo, haciendo la varianza proporcional a $m^{-\alpha}$ con $\alpha \in (0,1)$. Esto implica que

$$\sum_{k=-\infty}^{\infty}\rho_k \to \infty$$

Por tanto, la autocorrelación decae lentamente y no es sumable. Un ejemplo de este tipo de funciones es:

$$\rho_k = C\rho|k|^{-\alpha} \qquad \text{para k elevada.}$$

### 5.1 Autosimilitud

Es un concepto muy relacionado con la idea de los fractales, introducida por Mandelbrot. Se dice que un proceso $\{X_t\}$ es exactamente autosimilar (*self-similar*) si la estadística de la autocorrelación $\rho_k^{(m)}$ se mantiene en diferentes escalas de tiempo. El proceso se denomina

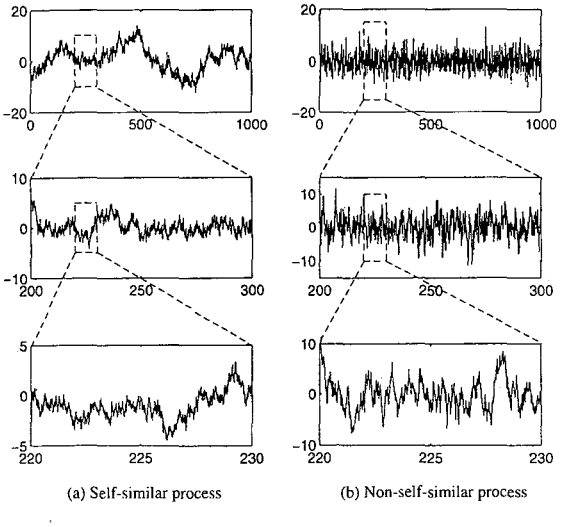


***Figura 7.*** *Comparación de un proceso autosimilar y otro no autosimilar. Se puede apreciar que el primero mantiene su apariencia mientras que el segundo tiende a comportarse como ruido blanco cuando está agregado (arriba) [Sta98].*

asintóticamente similar si la condición sólo se cumple para $m \to \infty$ y $k \to \infty$. Esto se expresa matemáticamente de la siguiente manera:

$$\text{Dist}\{X_{at}\} = \text{Dist}\, a^H\{X_t\}$$

H es el parámetro de Hurst, que da una medida de "lo autosimilar" que es el proceso. Puede parecer que la definición de autosimilitud es poco rigurosa, al basarse en la apariencia de las gráficas de la función de autocorrelacion. Es lo que se llama la "prueba visual" de la autosimilitud (figura 7). Hurst hizo unos estudios hidrológicos que le llevaron a descubrir que muchas series naturales presentan autosimilitud, es decir, que el hecho de cambiar la escala de observación del proceso agregado no cambia la estadística.

### 5.2 Dependencia a corto (*short range*) y largo (*long range*) plazo

Se dice que un proceso $\{X_t\}$ presenta dependencia a corto plazo si su autocorrelación es sumable:

$$\sum_{k=-\infty}^{\infty} \rho_k < \infty$$

Equivalentemente, la varianza $v_m$ decae proporcionalmente a $m^{-1}$, la densidad espectral de potencia tiene un valor finito en el cero, y el proceso promediado $\{X_j^{(m)}\}$ tiende a comportarse como ruido blanco a medida que $m \to \infty$. Los procesos con autocorrelaciones que decaen exponencialmente se denominan dependientes a corto plazo (*short range dependent*).

Se dice que un proceso $\{X_t\}$ muestra dependencia a largo plazo (*long range dependence*) si su autocorrelación no es sumable:

Equivalentemente, $v_m$ decae a un ritmo menor que

$$\sum_{k=-\infty}^{\infty} \rho_k \to \infty$$

$m^{-1}$, la densidad espectral de potencia tiene una singularidad en cero y el proceso promediado tiene la misma estadística que el proceso incial (autosimilitud). Los procesos con $\rho_k \sim Ck^{-a}$ (para k grande), presentan dependencia a largo plazo.

A continuación veremos dos casos de procesos autosimilares: uno exactamente autosimilar (Fractional Gaussian Noise) y otro asíntoticamente autosimilar (Fractional ARIMA).

#### Fractional ARIMA

Los procesos F-ARIMA (*Fractional Autoregressive Integrated Moving Average*) dependen de tres parámetros: p, d y q, como ya vimos en los ARIMA y se definen como una extensión de éstos:

$$\varphi(B)\, \nabla^d X_t = \theta(B)\, \varepsilon_t \qquad \text{con } 0 < d < 0.5.$$

Expresando el operador $\nabla^d$ en forma binomial, para el caso F-ARIMA (0, d, 0) con $0<d<0.5$ (d puede tomar valores no enteros, de aquí el adjetivo *fraccional*) y $k \to \infty$, se obtiene que $\rho^k \sim k^{2d-1}$. De aquí obtenemos $d = (1-\alpha)/2 = H-0.5$. Los procesos F-ARIMA pueden modelar dependencia a largo y corto plazo, según el valor de d. Por eso se les considera flexibles y potentes, especialmente en el caso del modelado de fuentes de vídeo VBR. Desgraciadamente, la estimación de los parámetros p, d y q, y de los coeficientes es costosa.

El tema de la estimación de d a partir de datos reales es de gran importancia. Uno de los métodos más eficientes es el de la gráfica varianza-tiempo. En este método, $v_m = \text{Var}(\{X^{(m)}\})$ es representada respecto a m, en escalas logarítmicas. La pendiente asintótica es $-\alpha$, y de ahí se puede estimar d.

#### Fractional Brownian Motion

El movimiento Browniano es un proceso estocástico $\{B_t\}$, para $t \geq 0$, con las siguientes características:

- los incrementos $B_{t+to} - B_{to}$ siguen una distribución normal de media 0 y varianza $\sigma^2 t$.
- los incrementos en los intervalos temporales no solapados $[t_1, t_2]$ y $[t_3, t_4]$, $B_{t4} - B_{t3}$ y $B_{t2} - B_{t1}$, son variables aleatorias independientes.
- $B_0 = 0$ y $B_t$ es continuo en $t = 0$.

El Movimiento Browniano fraccional, $\{fB_t\}$ es un proceso autosimilar gaussiano con $H \in [0.5,1)$. La diferencia respecto al Movimiento Browniano es que los incrementos tienen una varianza igual a $\sigma^2 t^{2H}$. Se demuestra que, para el caso discreto, la autocorrelación normalizada de la secuencia de incrementos (llamada también ruido gaussiano fraccional) es proporcional a $k^{2H-2}$ cuando $k \to \infty$ y, por tanto, presenta dependencia a largo plazo.

El parámetro H se puede estimar, como en el caso de los F-ARIMA, mediante la gráfica varianza-tiempo.

## 6. VALORACIÓN DE LOS MODELOS AUTOSIMILARES

### 6.1 Los estudios pioneros

La voz de alarma llegó en 1994, cuando un equipo de investigadores encabezados por Taqqu, Leland, Willinger y Wilson analizó trazas provenientes de una red Ethernet y constataron que el tráfico seguía un comportamiento autosimilar, en contradicción con los modelos clásicos markovianos (figura 8).

Los mismos autores efectuaron un análisis a nivel de fuentes aisladas, y concluyeron que el tráfico Ethernet se comporta como la agregación de muchas fuentes on-off con distribuciones de alta variabilidad (Pareto, *heavy tails*) [Will97].

Estas pruebas de que el tráfico real no coincide con el predicho por los modelos con dependencia a corto plazo provocaron la sensación de que todo el trabajo

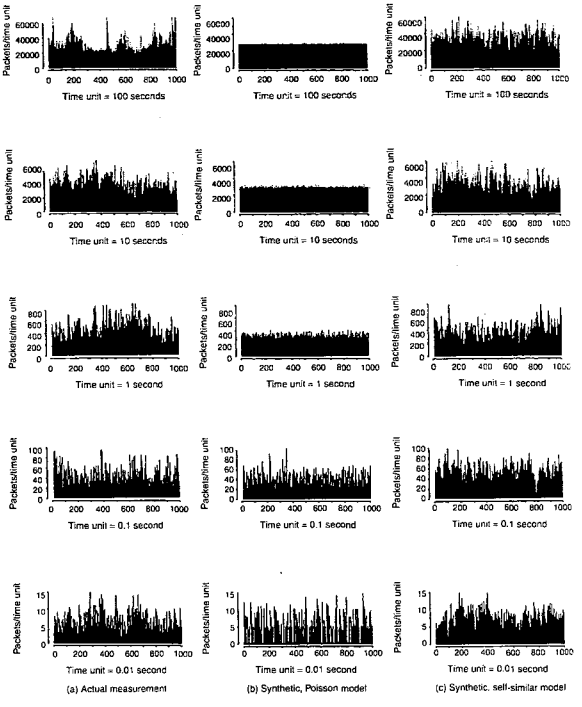


***Figura 8.*** *Constatación de la autosimilitud del tráfico Ethernet. A la izquierda, las trazas capturadas. En el centro, un modelo clásico. A la derecha, una traza sintetizada a partir de un modelo autosimilar [Will97].*

realizado hasta la fecha en el campo del modelado de tráfico se venía abajo. Los estudios sobre dimensionado de buffers, capacidades equivalentes, o control de acceso, parecían quedar invalidados por los nuevos modelos autosimilares. Si bien no se llegó a este extremo, y todavía se está discutiendo si es necesario abandonar los modelos clásicos en todos los casos, sí es cierto que se ha revitalizado la investigación sobre los modelos de fuentes. A continuación expondremos algunos de los resultados obtenidos a partir de los nuevos modelos.

## 6.2 Análisis de diferentes clases de tráfico

Desde los primeros artículos aparecidos en 1994, se ha demostrado que en muchas situaciones el tráfico se puede modelar con procesos autosimilares.

· **HTTP** (World Wide Web). Se puede modelar como la superposición de fuentes on/off de tipo Pareto, con varianza infinita. La justificación se basa en la distribución de los tamaños de los ficheros HTML y multimedia [Sta98].

· **SS7** (señalización). Se demuestra que si se utilizan los modelos poissonianos, el tráfico de señalización se subestima, ya que dicho tráfico presenta distribuciones con colas del tipo *heavy tail*, que indican un fuerte componente de dependencia a largo plazo [Sta98].

· **TCP, FTP, Telnet**. Paxson [Pax95] presenta un estudio exhaustivo sobre los servicios de Internet. Las principales conclusiones son:

– Los modelos poissonianos subestiman la variabilidad (*burstiness*) de las transferencias de paquetes TCP.

– Telnet es bien representado per Poisson a nivel de conexión, pero no a nivel de paquete, donde se detecta autosimilitud.

– Las transferencias en bloque de FTP se desvían mucho del modelo poissoniano, excepto en el nivel de sesión. La distribución del número de bytes en cada paquete sigue un modelo *heavy tailed* (un ejemplo es la distribución de Pareto).

· **Vídeo VBR** (*Variable Bit Rate*). Es uno de los temas más estudiados, dada la importancia que este tipo de tráfico tendrá en las redes de banda ancha (es el que más recursos de transmisión necesita). Garrett [Garr94] y Beran et al. [Ber95] presentan trabajos en los que se analizan diferentes películas y programas de televisión, y sus conclusiones son las siguientes:

– Se demuestra que la traza es autosimilar (presenta variaciones lentas de la componente continua) y que la distribución es *heavy-tailed.*

– Se destaca la presencia de componentes con dependencia a corto plazo, no despreciables (por eso son tan interesantes los modelos F-ARIMA, ya que capturan los dos comportamientos).

– La dependencia a largo plazo constituye una característica inherente a la estructura de las películas (planteamiento-nudo-desenlace, con diferentes requerimientos de ancho de banda para cada fase) y a la jerarquía plano-escena-secuencia (las diferentes escalas en las que se da la autosimilitud).

· **Buffers de conmutadores ATM**. Dado que las redes de banda ancha se basarán en la conmutación de celdas ATM, es de vital importancia prever el comportamiento que tendrán los *buffers* de los conmutadores cuando se les inyecte tráfico autosimiliar. Es uno de los campos en los que se ha generado más polémica, ya que mientras unos autores destacan que los modelos poissonianos subestiman la varianza del tráfico y pueden llevar a desbordamientos y pérdidas de celdas, otros autores argumentan que en la escala de tiempo a la que trabajan los conmutadores la autosimilitud del tráfico no tiene consecuencias graves.

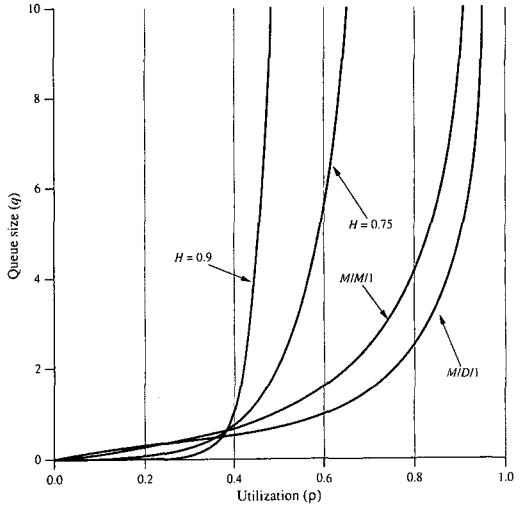


***Figura 9.*** *Tamaño de cola. Se presentan curvas de tráfico autosimilar (H=0.9, H=0.75) y los modelos clásicos (M/M/1 y M/D/1) [Nor94].*

– Norros [Nor94, COST242] analiza el comportamiento de un buffer bajo tráfico autosimilar, inyectando un proceso FBM a un buffer infinito con tiempos de servicio constantes. Concluye que se encuentran grandes discrepancias con los modelos M/M/1 i M/D/1, que se necesitan buffers mucho más grandes que los previstos cuando se usan dichos modelos (figura 9), y que también hay diferencias en el retardo de las celdas.

– Heyman et al. [Hey96] argumenta que la LRD puede tener una influencia mínima en el dimensionado de los *buffers* si el factor de Hurst es reducido y la parte SRD es importante. Asimismo, se cuestiona la influencia de la LRD en el cálculo de las pérdidas y se destaca el "*truncating effect of finite buffers*", que disminuye los efectos debidos a la dependencia a largo plazo (ya que se pierden celdas y la dependencia se reduce).

### 6.3 Reflexiones sobre los modelos autosimilares

Una vez revisado el trabajo realizado en este campo, y después de constatar que hay aspectos todavía abiertos, aparecen lagunas y surgen muchas preguntas:

– ¿Es realmente autosimilar el tráfico de las redes de banda ancha? Parece una buena aproximación, pero podría ser que el tráfico fuera simplemente no estacionario, y no necesariamente autosimilar (que es uno de los casos particulares de no estacionariedad).

– ¿La autosimilitud es debida a las fuentes o a su agregación? Depende del servicio. Por ejemplo, el tráfico LAN es autosimilar sólo cuando hay agregación, mientras que en el caso del vídeo VBR el efecto aparece en el nivel de fuentes individuales.

– ¿Hasta qué punto es la agregación y la propia dinámica de la red la que provoca la autosimilitud? Todavía está por estudiar a fondo la influencia de los protocolos en el comportamiento del tráfico. Por ejemplo, el protocolo TCP introduce una serie de mecanismos de control (ventana adaptativa, lucha per el ancho de banda, control de congestión). ¿Qué influencia podría tener ATM y su dinámica?

– ¿Qué papel juega la presencia *residual* de dependencia a corto plazo (SRD) en las fuentes de vídeo VBR?. ¿Cómo incluir sus efectos? ¿Es suficiente con los modelos F-ARIMA, o es necesaria una nueva familia de procesos que permitan una estimación de parámetros más fácil?

– Si el tráfico es realmente autosimilar, ¿son los modelos propuestos una buena solución para modelarlo? Ciertamente, son mejores que los modelos poissonianos, y proporcionan una explicación física del fenómeno (al menos en el caso del tráfico Ethernet, a través de la agregación de terminales que se comportan según modelos on-off con distribuciones de Pareto). Pero incluso los modelos autosimilares no son capaces de capturar al 100% el comportamiento de las trazas reales, por la presencia de SRD y periodicidades (como el caso de las fuentes MPEG).

– ¿Son realmente importantes los efectos autosimilares en la práctica? Es un tema abierto, y se ofrecen conclusiones contradictorias. No está claro que a las escalas temporales en que trabajan los buffers aparezca LRD, aunque se habla de importantes errores de subestimación. Stallings [Sta98] ha propuesto el estudio a dos niveles: **aplicación** (autosimilitud inherente a la fuente) y **red** (influencia de la dinámica de los protocolos como el TCP). A nivel de aplicación, la autosimilitud tendría influencia en el control de acceso (CAC) y la asignación de recursos de transmisión, mientras que a nivel de red, influiría en la congestión, el dimensionado de los buffers, etc. Tampoco queda claro qué influencia tiene cada nivel.

Vemos, por tanto, que pese a que parece bastante claro que los modelos autosimilares son prometedores, hace falta mucho más trabajo en este campo para llegar a conclusiones sólidas y útiles.

## 7. REFERENCIAS


[Ada97] ADAS, A. "Traffic Models in Broadband Networks", IEEE Communications Magazine, Vol. 35, No. 7, July 1997.

[Ber95] BERAN, J., SHERMAN, R., TAQQU, M., WILLINGER, W., "Long-range Dependence in Variable-bit-rate Video Traffic.", IEEE Transactions on Communications, February 1995

[CasBlo] CASALS, O. AND BLONDIA, C., "Traffic Management in ATM Networks: an overview", pendiente de publicación en la revista Computer Networks & ISDN Systems.

[COST242] "Methods for the performance evaluation and design of broadband multiservice networks. The COST 242 Final Report. Part III. Traffic models and queueing analysis", COST 242 Management Committee, 1996.

[Led94] LEDUC, J.-P. "Digital Moving Pictures – Coding and Transmission on ATM Networks", Elsevier, 1994.

[Lel94] LELAND, W., TAQQU, M., WILLINGER, W., AND WILSON, D., "On the Self-similar Nature of Ethernet Traffic (Extended Version).", IEEE/ACM Transactions on Networking, February 1994.

[Mag88] B. MAGLARIS ET AL., "Performance Models of Statistical Multiplexing in Packet Video Communications", IEEE Trans. Communications, vol. 36, July 1988.

[Nor94] NORROS, I., "A Storage Model with Self-similar Input", Queueing Systems, Vol. 16, 1994.

[Hey96] HEYMAN, D. AND LAKSHMAN, T., "What Are the Implications of Long-Range Dependence for VBR-Video Traffic Engineering?", IEEE/ACM Transactions on Networking, vol. 4, no.3, June 1996

[Pax95] PAXSON, V., AND FLOYD, S., "Wide Area Traffic: The failure of Poisson Modelling", IEEE/ACM Transactions on Networking, June 1995.

[Sen89] P. SEN ET AL., "Models for Packet Switching of Variable-Bit-Rate Video Sources", IEEE JSAC, vol. 7, no. 5, 1989.

[Sch96] M. SCHWARTZ, Broadband Integrated Networks, Prentice Hall, 1996.

[Sta98] STALLINGS, W., High-speed networks. TCP/IP and ATM design principles, Prentice Hall, 1998.

[Will97] WILLINGER, W., TAQQU, M., SHERMAN, R., WILSON, D., "Self-similarity through High Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level.", IEEE/ACM Transactions on Networking, February 1997.