Modelado y dimensionado de redes

Desde hace bastantes años la industria de las comunicaciones viene requiriendo un importante esfuerzo de convergencia. Después de todo, lo que podemos encontrar en el corazón de las redes de comunicación es un conjunto de flujos de bits, o sea, unos y ceros. A este nivel de detalle es imposible determinar si esos ceros y unos llevan voz, video, páginas web, o correo electrónico. Por supuesto, si todas estas redes no hacen más que transportar estos flujos de bits, deberíamos ser capaces de hacer converger estos servicios dentro de una red digital de transporte que pueda adecuarse a todos los requerimientos sobre una única plataforma de conmutación y transmisión. Después de todo, esto permitiría reducir los costes de operación e incrementar la funcionalidad de la red.

Esta llamada a la convergencia vino de mano de la arquitectura de red ATM (Asynchronous Transfer Mode), la Red Digital de Servicios Integrados (RDSI) y la arquitectura de red de cable de banda ancha. Después de todos estos esfuerzos de convergencia parecía que la industria se había asentado en una idea común: si el futuro es la convergencia, la plataforma para ello sería Internet. De acuerdo con esta idea, lo que veremos es una Internet multiservicio, donde todas las formas de comunicación (audio, video y datos) son enlazadas en una única plataforma a través de la tecnología de Internet.

Sin embargo, la convergencia no consiste simplemente en unir los distintos flujos de bits. Cada aplicación tiene asociadas una serie de características del servicio. Así, por ejemplo, para realizar una llamada telefónica, necesitamos que la conversación sea fiable, con una tasa de bit constante, un bajo jitter, que el flujo de bits sea simétrico, y hace falta una estructura de control para delimitar el comienzo y el final del flujo de bits. Una operación de descarga de una página web desencadena una serie de interacciones, empezando con un cierto número de intercambios de paquetes individuales de datos con varios servidores de dominio de nombres seguidos de una transferencia de datos desde el servidor web seleccionado, donde la transferencia puede adaptarse a varios niveles de retardo, jitter, e incluso pérdida de paquetes en la red. En este caso, los datos son transmitidos en paquetes y la interacción con la red se puede ver como una interacción basada en paquetes. La convergencia implica el uso de una arquitectura de transmisión y conmutación de bits para la red así como la adaptación de los perfiles de servicio de la red para adaptarse a los requerimientos de cada aplicación.

Hasta el momento, Internet se ha utilizado con un único modelo de servicio: el de envío sin garantía de paquetes donde la unidad de interacción con la red es el paquete y cada paquete es enviado a la dirección de destino independientemente de los paquetes anteriores y posteriores (mediante el protocolo IP). Por encima de este protocolo se dispone de un protocolo de transporte confiable (TCP) cuyos objetivos son en primer lugar proporcionar fiabilidad a red y en segundo lugar maximizar la eficiencia de la transmisión. Internet no realiza una gestión activa de los recursos. Si hay capacidad disponible, la red pondrá a disposición del tráfico de las aplicaciones. Si no hay suficiente capacidad, y cada vez hay más demandas, la red se congestiona. La reacción de las aplicaciones TCP a esta congestión consiste en reducir la tasa de bit para adecuarse a la demanda adicional. En otras palabras, TCP proporciona una respuesta variable, más rápida a baja carga y más lenta a alta carga. Los protocolos de Internet no soportan de forma intrínseca los flujos de datos de tasa de bit constante en los cuales la

transmisión es controlada por un reloj síncrono y la tasa envío de bit constante se corresponde perfectamente con la tasa de bit recibida.

Esto genere una pregunta esencial: ¿Puede Internet ser gestionada de forma que se adapte a las prestaciones requeridas por los distintos servicios? Si es así, entonces debe ser posible emular un circuito temporizado de forma síncrona, al menos con una precisión aceptable. También debe ser posible soportar respuestas constantes de forma que se pueda proporcionar una transacción con un nivel constante de recursos, independientemente de otros niveles de tráfico impuestos a la red. Con tales herramientas de gestión de recursos, debería ser posible controlar la respuesta en prestaciones de la red. Esto debería permitir a su vez la entrada en Intenet de una segunda oleada de servicios, soportando aplicaciones multimedia de alta velocidad, aplicaciones de video y audio con calidad constante así como el uso eficiente de los enlaces de bajo ancho de banda para soportar los PDAs que parecen constituir la próxima oleada de dispositivos de comunicación móviles.

La gestión de loa respuesta de los servicios de una red de Internet suele conocerse como Calidad de Servicio o QoS (Quality of Service) asumiendo que la principal motivación para la gestión de la respuesta de la red es, en última instancia, la mejora de la respuesta de los servicios de la red.

En una red de telecomunicación nos encontramos con aspectos importantes tales como la gestión, el encaminamiento o el control de congestión, entre otros. Para poder tomar decisiones con rigor debemos caracterizar tanto el tráfico de la red, como los distintos elementos del sistema.

La clave para un buen diseño es la habilidad para modelar y estimar correctamente parámetros relacionados con el rendimiento de las redes.

En términos generales podríamos englobar dichos parámetros en el concepto de calidad de servicio, y sería aplicable a las distintas capas de la pila de protocolos.

La explicación del concepto de Calidad de Servicio se va a realizar sobre la capa de aplicación, la capa de red y la capa de enlace de datos. El término de Calidad de Servicio tiene una interpretación diferente en cada una de estas capas. En algunos casos, puede ser incluso cuestionable utilizar el término Calidad de Servicio y tienen que ser definidos términos más concretos para una discusión más consistente.

La Calidad de Servicio en la capa de aplicación corresponde con el grado de satisfacción del usuario. En esta capa depende del grado de servicio percibido por el usuario, y del costo asociado. Por lo tanto, depende directamente de la opinión subjetiva de los usuarios, y es complicado crear parámetros generales que lo caractericen. La mejor herramienta para obtener los requisitos que demandan los usuarios, consiste en realizar tests subjetivos a los usuarios. Estos tests pueden mostrar el grado de satisfacción, por ejemplo, de la calidad percibida de una secuencia de imágenes recibidas en MPEG, y presentadas en una aplicación. El problema es que los resultados varían en función de las personas, y por tanto, pueden estar influenciados por muchos parámetros, que no siempre tienen que estar necesariamente relacionados con la tecnología. El parámetro no tecnológico más importante es el de coste de servicio. Los consumidores tienden a utilizar el servicio más barato, a partir de una determinada

calidad. La variabilidad de este parámetro consiste en que, con el tiempo, un usuario puede cambiar su opinión respecto al grado de Calidad de Servicio percibido o al coste que paga por el mismo.

El principal mecanismo de la capa de aplicación para mantener satisfecho siempre al usuario, es la capacidad de adaptación. Por ejemplo, una conexión de videoconferencia, se podría degradar por problemas de implementación de capas inferiores. La adaptación a poder usar una codificación de vídeo con menor ancho de banda, reducirá la calidad del servicio recibido, pero puede seguir siendo adecuado para el usuario. Normalmente, el usuario tiene que predefinir la tolerancia que considera aceptable como para obtener una calidad de servicio dada.

En la capa de red, el concepto de calidad de servicio es diferente. El principal objetivo para la Calidad de Servicio de la capa de red es soportar la modularidad. Los mecanismos, en esta capa, relacionan la Calidad de Servicio con el manejo de paquetes. En la capa de red, la primera cuestión para una nueva conexión es: "¿Es capaz esta red de soportar una nueva conexión requerida?". Si la red es capaz de soportar la conexión nueva, habrá dos niveles de manejo de la Calidad de Servicio adicionales: las garantías suaves y las garantías duras. Los sistemas con garantías *suaves*, aceptan siempre nuevas conexiones e intentan acomodar el servicio requerido lo mejor que pueden. En los sistemas con garantías *duros,* basándose en la información de las señales de control, las entidades de control de Calidad de Servicio deciden si una nueva conexión puede ser admitida o no. Una conexión no es admitida si existe la posibilidad de poner en peligro la garantía de servicio de cualquier otra conexión que esté en ese instante funcionando.

Las redes conmutadas están basadas en conexiones con señales de control, y en Calidad de Servicio con garantías duras. Después de que la conexión ha sido admitida en la red, la Calidad de Servicio se basa en la asignación de circuitos punto a punto. En las redes de conmutación de paquetes, se puede disponer de calidad de servicio con garantías suaves o duras.

La Calidad de Servicio en la capa de enlace tiene características similares. Los parámetros en la implementación del protocolo incluyen, por ejemplo, medida de valores para el retardo, variación del retardo, o probabilidad de pérdidas de paquetes. Tener suficiente ancho de banda disponible es uno de los elementos de Calidad de Servicio más importantes. En caso contrario, los paquetes se pueden perder o retardar en las colas de la red, esperando que desaparezca la congestión. Si existe alguna variación en la latencia entre paquetes puede hacer más complicado el manejo del tráfico en tiempo real. Si el periodo de latencia es muy largo, puede acabar reflejado en una tasa mayor de pérdidas.

Existen dos mecanismos para la asignación de ancho de banda en la capa de enlace de datos: la conmutación de circuitos y la conmutación de paquetes. La conmutación de circuitos en la capa de enlace de datos ofrece unas buenas garantías de servicio, pero puede desbordar la capacidad del canal. Esto conduciría a la necesidad de redes más caras (abastecimiento del canal y coste relacionado al usuario) que en redes de paquetes conmutadas, donde la capacidad del canal puede ser compartida. Las redes de paquetes de datos, pueden reducir los costes de servicio, haciendo el servicio más atractivo para los usuarios. La desventaja de las redes de conmutación de paquetes es que las conexiones se pueden deteriorar durante instantes de carga elevada si no se realizan

medidas fiables del tráfico. Los principales dispositivos medidores de tráfico son planificadores inteligentes y elementos de descarte de paquetes por peso en función de la conexión.

En las operaciones de las capas inferiores (las capas de red y de enlace de datos) se emplea un determinado tiempo, y por ello, los mecanismos de calidad de servicio deben ser lo más simples posible, para permitir la utilización de tráfico en tiempo real.

Es por tanto imprescindible entender el comportamiento de los paquetes en la red y su temporización para poder dimensionar redes de comunicación que permitan garantizar unos niveles dados de calidad de servicio.

Esto nos lleva al estudio de los procesos estocásticos, y más concretamente al estudio de los distintos modelos matemáticos para sistemas de colas, desde los sencillos modelos con colas infinitas hasta los complejos modelos autorregresivos en los que se tiene en cuenta las dependencias temporales.

Variables aleatorias

Una **variable aleatoria** es una correspondencia entre el conjunto de todos los posibles sucesos del espacio muestral considerado y los números reales. Esto es, una variable aleatoria asocia un número real a cada suceso. Este concepto se expresa a veces en términos de experimento con muchos resultados posibles; una variable aleatoria asigna un valor a cada uno de estos resultados. Por tanto, el valor de una variable aleatoria es una cantidad aleatoria. Vamos a dar la siguiente definición formal:

Una variable aleatoria X es una función que asigna un número a todos los resultados de un espacio muestral, y que satisface las condiciones siguientes:

1. El $\{X \leq x\}$ es un suceso para todo x.
2. $\Pr[X = \infty] = \Pr[X = -\infty] = 0$.

Se dice que una **variable aleatoria** es **continua** si adopta un número incontable infinito de valores diferentes. Una variable es **discreta** si adopta un número finito, o infinito pero contable, de valores.

Funciones de distribución y densidad

Una variable aleatoria continua X se puede describir mediante su **función de distribución** *F(x)* o mediante su **función de densidad** *f(x)*:

Función de distribución:

$$F(x) = \Pr[X \leq x] \qquad F(-\infty) = 0; \qquad F(\infty) = 1$$

Función de densidad:

$$f(x) = \frac{d}{dx} F(x) \qquad F(x) = \int_{-\infty}^{x} f(y)dy; \qquad \int_{-\infty}^{\infty} f(y)dy = 1$$

Para una variable discreta aleatoria, la distribución de probabilidad está caracterizada por:

$$P_x(k) = \Pr[X = k] \qquad \sum_{\forall k} P_x(k) = 1$$

Es frecuente que el objeto de nuestro interés sea una característica de una variable aleatoria, y no toda la distribución, como las que siguen:

Valor medio:
$$\begin{cases} E[X] = \mu_x = \int_{-\infty}^{\infty} x f(x) dx & \textit{caso continuo} \\ E[X] = \mu_x = \sum_{\forall k} k \Pr[x = k] & \textit{caso discreto} \end{cases}$$

Segundo momento:
$$\begin{cases} E[X^2] = \int_{-\infty}^{\infty} x^2 f(x) dx & \textit{caso continuo} \\ E[X^2] = \sum_{\forall k} k^2 \Pr[x = k] & \textit{caso discreto} \end{cases}$$

Varianza: $$Var[X] = E[(X - \mu_x)^2] = E[X^2] = \mu_x{}^2$$

Desviación estándar: $\sigma_x = \sqrt{Var[X]}$

La varianza y la desviación estándar son medidas de la dispersión de valores respecto a la media. Una varianza elevada significa que la variable toma más valores relativamente alejados de la media que en el caso de una varianza reducida.

Distribuciones estadísicas importantes

Existen varias distribuciones que juegan un papel importante en el análisis de colas; se describen a continuación las distribuciones más destacadas.

Distribución exponencial

La distribución exponencial con parámetro $\lambda > 0$ está dada por las figuras a y b y posee las siguientes funciones de distribución y de densidad:

$$F(x) = 1 - e^{-\lambda x} \qquad f(x) = \lambda e^{-\lambda x} \qquad x \geq 0$$

La distribución exponencial posee una propiedad interesante: su media es igual a su desviación estándar:

$$E[X] = \sigma_x = \frac{1}{\lambda}$$

Esta distribución es importante en la teoría de colas porque suele ser posible suponer que el tiempo de servicio perteneciente a un sistema de colas es exponencial. En el caso del tráfico telefónico, el tiempo de servicio es el tiempo durante el cual el abonado hace uso del equipo considerado. En una red de conmutación de paquetes, el tiempo de servicio es el tiempo de transmisión, y, por tanto, es proporcional a la longitud del paquete. El hecho de poder considerar que los tiempos de servicio son exponenciales es muy importante ya que permite simplificar muchísimo el análisis de las colas.

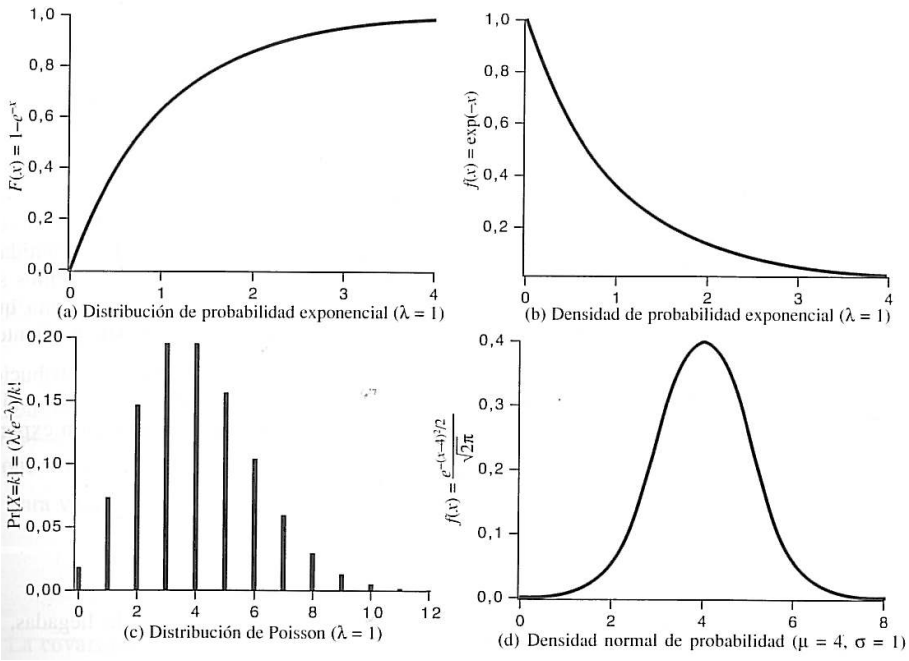


Figura 1 Algunas funciones de probabilidad

Distribución de Poisson

Otra distribución importante es la distribución de Poisson con un parámetro $\lambda > 0$ que toma valores en los puntos situados en 0, 1, 2, …

$$\Pr[X = k] = \frac{\lambda^k}{k!} e^{-\lambda} \qquad k = 0, 1, 2, \ldots$$

$$E[X] = Var[X] = \lambda$$

Si $\lambda < 0$, entonces Pr[X = k] es el máximo para k = 0. Si $\lambda > 0$ pero no es un número entero, entonces Pr[X = k] es máximo para el mayor entero menor que λ; si λ es un entero positivo, entonces hay dos máximos, en $k = \lambda$ y en $k = \lambda - 1$.

La distribución de Poisson también es importante en el análisis de las colas porque debemos suponer que el patrón de llegadas es de Poisson para ser capaces de desarrollar las ecuaciones de colas que veremos en esta asignatura. Afortunadamente, la suposición de llegadas de Poisson suele ser cierta.

La forma en que se puede aplicar la distribución de Poisson al régimen de llegadas es como sigue. Si los elementos llegan a la cola siguiendo un proceso de Poisson, esto se puede expresar en la forma:

$$\Pr[\text{llegan k elementos en el intervalo de tiempo T}] = \frac{(\lambda T)^k}{k!} e^{-\lambda T}$$

E[número de elementos que van a llegar en el intervalo de tiempo T] = λT

Régimen medio de llegadas, en elementos por segundo = λ

Las llegadas que se producen siguiendo un proceso de Poisson suelen recibir el nombre de llegadas aleatorias. Esto se debe a que la probabilidad de llegada de un elemento dentro de un pequeño intervalo es proporcional a la longitud del intervalo, e independiente de la cantidad de tiempo transcurrida desde la llegada del último elemento. Esto es, cuando llegan elementos siguiendo un proceso de Poisson, es tan probable que un elemento llegue en un instante como que llegue en cualquier otro, independientemente de los momentos en que lleguen los demás clientes.

Otra propiedad interesante del proceso de Poisson es su relación con la distribución exponencial. Si se examinan los tiempo entre llegadas de los elementos T_a (que se denominan tiempos entre llegadas), entonces se observa que esta cantidad obedece a la distribución exponencial:

$$\Pr[T_a < t] = 1 - e^{-\lambda t}$$

$$E[T_a] = \frac{1}{\lambda}$$

Por tanto, el tiempo medio entre llegadas es el inverso del régimen de llegadas, como cabía esperar.

Distribución normal

La distribución normal con parámetros $\mu > 0$ y σ tiene la siguiente función de distribución:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{(-x-\mu)^2/2\sigma^2} \qquad F(X) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{X} e^{-(y-\mu)^2/2\sigma^2} dy$$

Donde

$$E[X] = \mu$$

$$Var[X] = \sigma^2$$

Un resultado importante es el teorema del límite, que afirma que la distribución de la media de un elevado número de variables aleatorias independientes será aproximadamente normal, casi independientemente de sus distribuciones individuales. Un requisito clave es que la media y la varianza tienen que ser finitas. El teorema central del límite desempeña un papel clave en la estadística.

Procesos estocásticos

Un proceso estocástico, que a veces se denomina proceso aleatorio, es una familia de variables aleatorias [x(t), $t \in T$], que tiene como índice un parámetro t perteneciente a algún conjunto de índices T, y definidas sobre un espacio Ω, llamado **espacio de estados**. Generalmente, el conjunto índice se interpreta como dimensión temporal, x(t) es una función del tiempo. Otra forma de decir esto es que un proceso estocástico es una variable aleatoria que es una función del tiempo. Un proceso estocástico continuo en el tiempo es aquel en que varía de forma continua, generalmente a lo largo de la recta real no negativa $\{x(t), 0 \leq t \leq \infty\}$, aunque a veces recorre toda la recta real; sin embargo, un proceso estocástico discreto en el tiempo es aquel en que t adopta valores discretos,

generalmente valores enteros positivos {x(t), t=1, 2, 3, …}, aunque en algunos casos el rango son los enteros entre -∞ y +∞.

Los conjuntos Ω y T pueden ser discretos o continuos. En lo que respecta al conjunto T, el parámetro t representa normalmente el tiempo. Dado que T puede tener, bien una naturaleza continua, bien discreta y numerable, suele distinguirse a veces entre **procesos estocásticos** y **sucesiones estocásticas** respectivamente.

Las posibles configuraciones son:

- Ω discreto y T discreto, por ejemplo en el juego de los dados si la variable aleatoria representa el resultado del *n-ésimo* experimento, $\Omega=\{1, 2, 3, 4, 5, 6\}$ y T="nº de tirada". En este caso, el espacio sobre el que está definido el proceso estocástico sería el que forman todas las posibles trayectorias variables en el tiempo.
- Ω continuo y T discreto, por ejemplo, el tiempo de espera en cola del n-ésimo cliente $[\Omega(T \in \mathbf{N}) = t \in \Re]$
- Ω discreto y T continuo, por ejemplo, número de clientes presentes en el tiempo t.
- Ω continuo y T continuo, temperatura en el tiempo t.

Tal y como sucede con cualquier variable aleatoria, x(t) para un valor fijo de t se puede caracterizar mediante una distribución de probabilidad y una densidad de probabilidad. Para los proceso estocásticos de valor continuo, estas funciones adoptan la forma siguiente:

Función de distribución:

$$F(x; t) = \Pr[x(t) \le x] \qquad F(-\infty; t) = 0 \qquad F(+\infty; t) = 1$$

Función de densidad:

$$f(x;t) = \frac{\partial}{\partial x} F(x; t) \qquad F(x; t) = \int_{-\infty}^{x} f(y;t)dy \qquad \int_{-\infty}^{+\infty} f(y;t)dy = 1$$

Para procesos estocásticos de valores discretos,

$$P_{x(t)}(k) = \Pr[x(t) = k] \qquad \sum_{\forall k} P_{x(t)}(k) = 1$$

La naturaleza nos muestra con frecuencia ejemplos de procesos estocásticos de fenómenos, esto es, casos en los que una o varias magnitudes varían con el tiempo de forma difícilmente predecible. Hasta finales del siglo XIX todo el esfuerzo de modelización de tales fenómenos se realizó a través de ecuaciones diferenciales. Los modelos así construidos eran completamente deterministas en su concepción y naturaleza. El máximo exponente de esta línea de pensamiento fue, sin duda, la mecánica clásica, para la cual era posible predecir la trayectoria futura de un cuerpo en movimiento y revelar su pasado si se conoce su estado presente y las fuerzas que obran sobre él.

Sin embargo, es notorio que existen múltiples fenómenos dinámicos que no son predecibles de manera exacta a partir de modelos de ecuaciones diferenciales. En el mundo microscópico dicha impredecibilidad se debe, de acuerdo con la mecánica cuántica, a la naturaleza intrínsecamente aleatoria del comportamiento de las partículas elementales. En el mundo macroscópico, en general, la impredecibilidad se debe a que muchos de los fenómenos que se observan son demasiado complejos como para que su análisis completo pueda reducirse al estudio de unas pocas variables que se introducen en un sistema de ecuaciones diferenciales.

En este sentido, los fenómenos dinámicos macroscópicos podrían clasificarse en tres categorías de impredecibilidad, aunque los límites entre las mismas puedan ser un tanto vagos en algunos casos:

1. Aquéllos en los cuales es posible encontrar y medir un número reducido de variables que determinan en una proporción elevada el comportamiento del fenómeno. En estos casos los modelos deterministas cuentan con un margen de error muy pequeño y constituyen una buena representación de la realidad. Ejemplo de ello son las órbitas planetarias.

2. Aquéllos en los que se conocen diversas variables que intervienen en la evolución del fenómeno; de algunas variables se sabe cómo actúan, pero las otras muestran interacciones tan complejas entre sí y con el fenómeno que resultan difícilmente interpretables o modelizables en términos de claras relaciones causa - efecto. Además existen posiblemente variables desconocidas, con efectos a menor escala, pero que sumados introducen una variabilidad significativa. Ejemplo de esta categoría son los fenómenos climatológicos o los oceanográficos.

3. La última categoría de fenómenos es la que se caracteriza por tener un comportamiento de naturaleza intrínsecamente aleatoria, bien porque las variables que los generan son aleatorias (de comportamiento impredecible), o bien porque su complejidad hace imposible cualquier tipo de modelización causal. Como ejemplo en este caso podemos citar la longitud de una cola de espera.

Es en el contexto de las dos últimas categorías donde surge el concepto de **proceso estocástico**. Ya que no es posible predecir con exactitud la evolución temporal de la variable ó variables de interés, ¿por qué no dar una medida de probabilidad sobre sus posibles cursos de evolución? Supongamos que en el ejemplo del cuerpo en movimiento, conocemos su estado presente, pero además sabemos que las fuerzas que actúan sobre él son tan complejas que llegan a tener la consideración de aleatorias. En tal caso, en lugar de tratar de predecir exactamente su trayectoria futura, tarea obviamente irrealizable, podríamos tratar de construir el conjunto de todas las trayectorias posibles, y asignar a cada una la probabilidad de que sea ella la que efectivamente se produzca. Esta es precisamente la noción de proceso estocástico: no una trayectoria única que ocurre con seguridad, sino todo un conjunto de trayectorias posibles acompañadas de sus respectivas probabilidades de ocurrencia.

La principal aplicación de los procesos estocásticos que vamos a tratar es la modelización de la evolución de las colas de espera que se producen en diversos sistemas de comunicación.

Procesos estocásticos estacionarios

Diremos que un **proceso estocástico** es **estacionario** si su valor esperado es constante y su función de autocorrelación depende únicamente de la diferencia de tiempos:

$$E[x(t)] = \mu$$
$$R(t, t + \tau) = R(t + \tau, t) = R(\tau) = R(-\tau) \text{ para todo } t$$

Siendo la **función de autocorrelación** $R(t_1, t_2)$ el momento conjunto de las variables aleatorias $x(t_1)$ y $x(t_2)$, esto es:

$$R(t_1,t_2)= E[x(t_1)x(t_2)]$$

Una característica importante de $R(\tau)$ es que mide el grado de dependencia de un instante de tiempo de un proceso estocástico con respecto a otros instantes de tiempo. Si $R(\tau)$ tiende a cero exponencialmente cuando τ alcanza valores elevados, entonces hay poca dependencia entre un instante de ese proceso y los instantes de tiempo alejados del mismo. Se dice que este proceso es un **proceso de memoria breve**, mientras que si $R(\tau)$ sigue sin anularse para valores grandes de τ (decrece con una rapidez menor que la exponencial), se dice que el proceso estocástico es un **proceso de memoria larga**.

Procesos de Markov

Diversas leyes de la Física clásica –determinista– especifican que en ciertos sistemas basta conocer su estado actual y las fuerzas que operan sobre él para predecir su comportamiento futuro. La idea subyacente a estas leyes es que lo que le haya sucedido al sistema hasta llegar al estado actual no aporta, en orden a predecir su evolución futura, ninguna información que no esté ya recogida en las variables que definen su estado presente. Es el llamado principio de independencia entre el futuro y el pasado una vez que se conoce el presente. Esta idea puede generalizarse en un contexto probabilístico, y constituye la base de la definición de los procesos de Markov.

Un *proceso de Markov* es un proceso estocástico $\{\xi_t\}_{t\in T}$ tal que:

$$P(\xi_{s+t} \in B| \ \xi_u, u\leq s) = P(\xi_{s+t} \in B| \ \xi_s)$$

En otras palabras es un proceso tal que la distribución condicional del futuro ξ_{s+t} dada toda su evolución hasta el presente ξ_u , $u\leq s$, depende sólo del presente y es independiente del pasado. Si además ocurre que $P(\xi_{s+t} \in B| \ \xi_s)$ es independiente de s, entonces el proceso se dice ***homogéneo***, o dicho con otras palabras, se dice que un proceso es homogéneo cuando sus probabilidades de transición sólo dependen del valor de dicha transición.

Si su función de distribución permanece constante para cualquier transición entonces se trata de un proceso ***estacionario***.

Cuando T es discreto, el proceso de Markov se dice *en tiempo discreto*, y *en tiempo continuo* en otro caso. El conjunto de valores que puede tomar un proceso de Markov

recibe el nombre de *espacio de estados* del proceso. Cuando el espacio de estados es discreto, el proceso suele recibir el nombre de ***cadena de Markov***. Los procesos de Markov en tiempo continuo con espacio de estados continuo reciben el nombre de *procesos de difusión.*

Los procesos de Markov son de gran interés en las aplicaciones prácticas. Constituyen el modelo básico para muchas clases de colas de espera.

Cadenas de Markov en tiempo discreto.

Las cadenas de Markov en tiempo discreto constituyen el modelo de proceso de Markov más simple. El tiempo *T* es discreto (*T=N;* en general se llaman *etapas* a las unidades de tiempo discretas*)* y también es discreto (finito o numerable) el espacio de estados, al que llamaremos *E*. La condición de Markov se expresa en este caso de la forma:

$$P(\xi_{n+1} = j \mid \xi_0=i_0,\ \xi_1=i_1,\ldots,\xi_n=i_n,) = P(\xi_{n+1} = j \mid \xi_n=i_n,)$$

Cuando la cadena de Markov es homogénea, esta condición se transforma en:

$$P(\xi_{n+1} = j \mid \xi_n=i,) = P(\xi_{m+1} = j \mid \xi_m=i,) \quad \forall n,m \in N$$

lo que significa que la probabilidad de pasar del estado *i* al estado *j* en una etapa es independiente de cuál sea esta etapa.

Las cadenas homogéneas presentan en la práctica gran número de aplicaciones. Es fácil observar que una cadena de Markov homogénea queda determinada por:

1. El espacio de estados *E* (finito o numerable).
2. $\forall i,j \in E, \quad P(\xi_{n+1} = j / \xi_n = i) = p_{ij}$ (*probabilidad de transición* del estado *i* al *j*, independiente de *n*). Obviamente $\sum_{j \in E} p_{ij} = 1, \quad \forall i \in E.$
3. $\forall i,j \in E, \quad P(\xi_0 = i) = p_i^{(0)}$ (*distribución inicial de la cadena*).

Suele ser de interés calcular la probabilidad de que la cadena, transcurridas *n* etapas, se encuentre en algún estado *j*, habiendo partido inicialmente del estado *i*, esto es,

$$p_{ij}^{(n)} = P(\xi_n = j / \xi_o = i)$$

Para ello, si definimos las ***matrices de probabilidades de transición*** en una y *n* etapas respectivamente, como:

$$P = \left(p_{ij}\right)_{i,j \in E} \qquad P^{(n)} = \left(p_{ij}^{(n)}\right)_{i,j \in E}$$

puede probarse el siguiente resultado, conocido como ***ecuación de Chapman-Kolmogorov***:

$$P^{(n)} = P^n$$

Si se desea calcular la probabilidad incondicional de que la cadena se encuentre en el estado *j* tras *n* etapas, llamando $p_j^{(n)} = P(\xi_n = j)$, y $p^{(n)} = (p_1^{(n)}, p_2^{(n)}, ...)$, se tiene

$$p^{(n)} = p^{(0)} P^n$$

Otro problema de interés suele ser el cálculo de la probabilidad de que, partiendo del estado *i*, la cadena llegue por primera vez al estado *j* transcurridas *n* etapas. Esta probabilidad suele denotarse como:

$$f_{ij}^{(n)} = P(\xi_n = j, \xi_k \neq j, k < n / \xi_0 = i)$$

Si se definen las funciones generatrices:

$$P_{ij}(s) = \sum_{n=1}^{\infty} p_{ij}^{(n)} s^n \quad , \qquad F_{ij}(s) = \sum_{n=1}^{\infty} f_{ij}^{(n)} s^n$$

puede probarse que:

$$F_{ij}(s) = \frac{P_{ij}(s)}{1 + P_{jj}(s)}$$

La probabilidad de que la cadena alcance alguna vez el estado *j* habiendo partido de *i*, es entonces:

$$f_{ij} = \sum_{n=1}^{\infty} f_{ij}^{(n)} = F_{ij}(1)$$

y el número medio de etapas que emplea en ir de *i* a *j* es, obviamente, $F'_{ij}(1)$.

Por último, otro problema también de gran importancia en las aplicaciones prácticas es el siguiente: transcurrido un número muy grande de etapas ¿cuál es la probabilidad de que la cadena se encuentre en un estado *j* determinado?. Esta pregunta sólo tiene sentido cuando exista la distribución límite:

$$\pi_j = p(\xi = j) = \lim_{n \to \infty} p_j^{(n)}$$

en cuyo caso se dice que la cadena está en ***equilibrio***.

El resultado importante sobre la distribución límite es entonces que si una cadena de Markov es irreducible y ergódica, entonces existe $\pi_j = \lim_{n \to \infty} p_j^{(n)}$ y es independiente de *i*. Además, llamando $\pi = (\pi_1, \pi_2, ...)$, se tiene que π es la única solución de $\boldsymbol{\pi P = \pi}$, con la condición $\sum_{i=1}^{\infty} \pi_i = 1$ (esto es, π es la única distribución estacionaria de la cadena). Otra forma habitual de enunciar este resultado es que, si la cadena es irreducible y aperiódica y existe la distribución estacionaria, entonces la cadena es ergódica y la distribución límite coincide con la distribución estacionaria.

Cadenas de Markov en tiempo continuo.

En este caso $T=[0,\infty)$. La condición de Markov es ahora:

$$P(\xi_{t+s} = j \mid \xi_s=i,\ \xi_u=i_u,\ 0\leq u<s) = P(\xi_{t+s} = j \mid \xi_s=i)$$

Cuando esta probabilidad no depende de *s*, la cadena es homogénea. Las cadenas de Markov homogéneas en tiempo continuo, al igual que sus homólogas en tiempo discreto, tienen también un enorme interés práctico. El ejemplo del sistema de comunicaciones expuesto en el apartado anterior es también válido en este caso cuando el protocolo de comunicación permite que las unidades de información sean transmitidas en cualquier instante *t*, cosa habitual en muchos casos.

Las cadenas de Markov en tiempo continuo homogéneas verifican la siguiente importante propiedad: si llamamos τ_i al tiempo que la cadena permanece en el estado *i* entonces:

$$P(\tau_i \geq s+t \mid \tau_i \geq s) = P(\xi_{t+s} = i \mid \xi_s=i) = P(\xi_t = i \mid \xi_0=i) = P(\tau_i \geq t)$$

esto es, la probabilidad de que el proceso permanezca en el estado *i* aún durante un tiempo t, es independiente del tiempo que el proceso haya permanecido ya en ese estado. Esta propiedad suele expresarse diciendo que la variable τ_i *no tiene memoria*. Ahora bien, puede demostrarse que la única distribución de probabilidad con esta propiedad es la exponencial, y por tanto τ_i tiene una distribución exponencial. Ello nos permite definir un proceso de Markov homogéneo en tiempo continuo como un proceso que, cada vez que alcanza el estado *i:*

(a) La cantidad de tiempo que permanece en ese estado antes de transitar a otro estado sigue una distribución exponencial.

(b) Cuando el proceso abandona el estado *i* entra en otro estado $j\neq i$ con probabilidad p_{ij} que verifica: $p_{ii}=0$, $\sum_{j\in E} p_{ij} = 1, \quad \forall i \in E$.

En otras palabras, una cadena de Markov homogénea en tiempo continuo es un proceso estocástico que se mueve de un estado a otro de acuerdo con una cadena de Markov en tiempo discreto, pero que permanece en cada estado durante un tiempo exponencialmente distribuido (con parámetro dependiente del estado). Además el siguiente estado *j* a que se mueve el proceso es independiente del tiempo que haya permanecido en el estado *i* (si no fuera así se violaría la condición de Markov).

Por último, en analogía con el resultado obtenido para las cadenas en tiempo discreto, a menudo la probabilidad de que la cadena de Markov se encuentre en el estado *j* transcurrido un tiempo muy largo converge a un valor π_j (*probabilidad ergódica*) independiente del estado inicial, esto es:

$$\pi_j = \lim_{t\to\infty} p_{ij}(t)$$

Condición suficiente para ello es que todos los estados intercomuniquen y que sean recurrentes positivos. El valor π_j puede interpretarse como la proporción de tiempo que el sistema permanece en el estado *j*.

Procesos puntuales

Un proceso puntual aleatorio es un proceso estocástico cuyas realizaciones consisten en conjuntos de puntos distribuidos aleatoriamente sobre un cierto espacio continuo. Tales puntos suelen corresponder, en la práctica, a los instantes de tiempo en que han ocurrido algunos sucesos de interés, ó a las localizaciones en el espacio de ciertos objetos. Como ejemplos de procesos puntuales podemos citar el conjunto de instantes en que se producen las llegadas o salidas de clientes de una cola, el conjunto de instantes en que se producen los nacimientos de los individuos de una población, el conjunto de puntos de una zona geográfica en que se encuentra cierta clase de árboles, o el conjunto de localizaciones en que se encuentran los peces de un cardumen en un determinado momento. En los dos primeros ejemplos, el continuo sobre el que se hallan distribuidos los puntos es el tiempo; en los otros dos ejemplos es el espacio físico.

El análisis de los procesos puntuales se realiza habitualmente a través de sus procesos contadores asociados. Dado un proceso puntual definido sobre un espacio continuo X, se define su proceso contador asociado como aquel proceso que a cada subconjunto $A \subset X$ le asigna el número *N(A)* de ocurrencias del proceso puntual localizadas en *A*. Cuando X es el tiempo, el proceso contador suele expresarse como $\{N_t\}_{t\geq 0}$, donde N_t representa el número de sucesos puntuales que han ocurrido en *[0,t]*.

Si $\{N_t\}_{t\geq 0}$ es un proceso contador asociado a un proceso puntual temporal, para $s<t$ el valor de $N_t - N_s$ representa el número de ocurrencias del proceso que han tenido lugar en el intervalo *(s,t]*. Cuando $N_t - N_s$ y $N_v - N_u$ son variables aleatorias independientes $\forall s<t\leq u<v$, el proceso $\{N_t\}_{t\geq 0}$ se dice de ***incrementos independientes***. Asimismo, si la distribución de la variable $N_t - N_s$, $s<t$, depende sólo de la diferencia $t-s$, el proceso se dice de ***incrementos estacionarios***.

Uno de los procesos contadores más importantes es el proceso de Poisson. Es un proceso que se usa extensamente en teoría de colas.

Procesos de Poisson y otros relacionados con el mismo

Cuando se tienen llegadas aleatorias en el tiempo se sigue una distribución de Poisson de forma que:

$$\Pr[\text{llegan k elementos en el intervalo de tiempo T}] = \frac{(\lambda T)^k}{k!} e^{-\lambda T}$$

Se puede definir un proceso de cuenta de Poisson $\{N(t), t \geq 0\}$ de la forma siguiente:

1. N(t) posee incrementos independientes estacionarios
2. $N(0) = 0$
3. Para $0 < t_1 < t_2$, la cantidad $N(t_2) - N(t_1)$ es igual al número de puntos que hay en el intervalo (t_1, t_2) y tiene una distribución de Poisson con una media $\lambda(t_1 - t_2)$.

Conviene señalar el contexto práctico en que surge un proceso contador de Poisson. Puede parecer natural la primera condición (que establece que se empieza a contar en *t=0*), e incluso la segunda, pues no es difícil imaginar casos prácticos en que los incrementos de los procesos contadores sean independientes. Lo que seguramente ya no le resulte tan natural es la tercera condición, pues no es obvio en qué condiciones los incrementos de un proceso van a seguir una distribución de Poisson. Es por ello que resulta de interés introducir la siguiente caracterización alternativa del proceso de Poisson, que ayuda a captar de modo más intuitivo en qué casos el proceso de Poisson puede ser un modelo adecuado.

Un proceso contador $\{N_t\}_{t\geq 0}$ es de *Poisson* si:

1. $P(N_0 = 0) = 1$
2. $\{N_t\}_{t\geq 0}$ es de incrementos independientes.
3. $P(N(t+\Delta t)-N_t = 1) = \lambda_t \Delta t + o(\Delta t)$ y $P(N(t+\Delta t)-N_t > 1) = o(\Delta t)$ donde $o(\Delta t)$ es un infinitésimo de orden superior a Δt.

Se puede probar que estas condiciones son equivalentes a las anteriores cuando existe λ_t. En esta caracterización queda de manifiesto que aquellos procesos de incrementos independientes tales que, durante un tiempo infinitesimal es muy improbable que ocurra más de un suceso seguirán una distribución de Poisson. Como ejemplos en los que esta condición se da frecuentemente, pueden citarse las llegadas de clientes a una cola, o los nacimientos de individuos en una población.

Para los procesos de Poisson se obtienen las siguientes funciones de probabilidad para N(t):

$$\Pr[N(t) = k] = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

$$E[N(t)] = \lambda t$$

Claramente, N(t) no es estacionario, porque su media es función del tiempo. Todas las funciones temporales de este proceso estocástico adoptan la forma de una escalera ascendente, con pasos de longitud 1, que se producen en los puntos aleatorios dados por t_i. La siguiente figura nuestra un ejemplo de proceso de cuenta de Poisson.

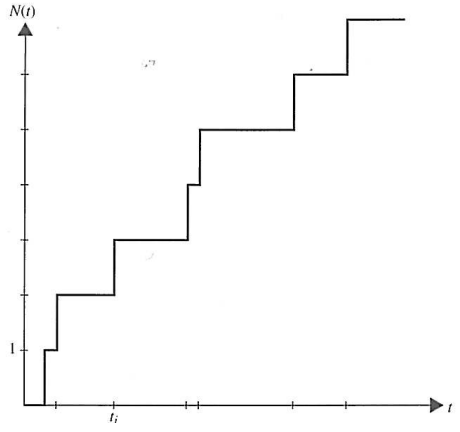


Figura 2 Proceso de cuenta de Poisson

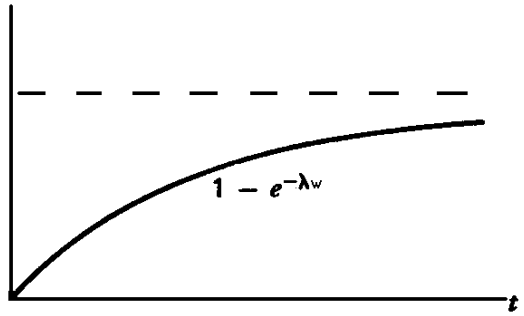
Entre las propiedades del proceso de Poisson destacamos las siguientes, de gran interés en las aplicaciones:

- *El proceso de Poisson es Markoviano*: se sigue directamente del hecho de que sus incrementos son independientes.
- *Distribución de los tiempos entre ocurrencias:* sea $W_n=T_n-T_{n-1}$, entonces:

$$P(W_n \le w | T_{n-1}=t,\ T_{n-2}=t_{n-2},...,T_1=t_1) = 1-e^{-[\Lambda(t+w)-\Lambda(t)]}$$

  Por tanto el tiempo hasta la próxima ocurrencia depende sólo del instante de la última ocurrencia. Como $T_{n-1}=W_1+...+W_{n-1}$, resulta que W_n depende de todos los W_i anteriores. En el caso particular de que el proceso sea homogéneo:

$$P(W_n \le w | T_{n-1} = t) = 1-e^{-[\Lambda(t+w)-\Lambda(t)]}$$



  que, como se observa, sigue una distribución exponencial de parámetro λ, que ahora es independiente de los tiempos de espera anteriores. Puede probarse que el recíproco también es cierto, esto es, si $\{W_n\}_{n\ge1}$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas con distribución $exp(\lambda)$, entonces $\{W_n\}_{n\ge1}$ son los tiempos entre ocurrencias de un proceso de Poisson.

- *Suma y descomposición de Procesos de Poisson:* En muchas aplicaciones nos encontramos con el problema de que las ocurrencias que se observan (por ejemplo, las llegadas de clientes a una cola) son generadas por distintos procesos de Poisson. En tales casos resulta de interés el resultado que indica que la suma de *k* procesos de Poisson independientes $\{N_t^{(1)}\}, \{N_t^{(2)}\}, ..., \{N_t^{(k)}\}$ de intensidades respectivas $\lambda_1(t),\lambda_2(t),...,\lambda_k(t)$ es también un proceso de Poisson con intensidad $\lambda_1(t)+\lambda_2(t)+...+\lambda_k(t)$.

Proceso de incremento de Poisson

Un proceso estacionario relacionado con el proceso de cuenta de Poisson es el **proceso de incremento de Poisson**. Para un proceso de cuenta de Poisson N(t) cuya media es λt, y para una constante L (L > 0), se puede definir un proceso de incremento de Poisson X(t) de la forma siguiente:

$$X(t) = \frac{N(t+L)-N(t)}{L}$$

X(t) es igual a k/L, donde k es el número de puntos del intervalo (t, t + L). En la siguiente figura se muestra el proceso de incremento derivado del proceso de cuenta representado en la figura anterior.

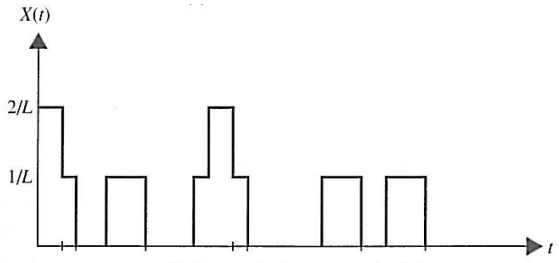


Figura 3 Proceso de incrementos de Poisson

Para el caso de procesos de incrementos de Poisson se tiene la siguiente relación:

$$E(X(t)) = \frac{1}{L} E[N(t+L)] - \frac{1}{L} E[N(t)] = \lambda$$

Con una media constante, X(t) es un proceso estacionario en sentido amplio y, por tanto, tiene una función de autocorrelación de una sola variable, R(τ).

Procesos de nacimiento y muerte

Los procesos de nacimiento y muerte son una subclase importante de las cadenas de Markov en tiempo continuo. Si se supone que en el instante t el estado del sistema es $\xi_x(t)$, viene representado por la magnitud $x = n$, los procesos de nacimiento y muerte serán aquellos que sólo permiten el paso hacia los estados $x = n+1$ *(nacimiento)* ó $x = n-1$ *(muerte)* en una transición.

Procesos de colas

Los sistemas de colas constituyen en la actualidad un campo de investigación y aplicación muy activo. Ello se debe fundamentalmente a que en los modernos sistemas informáticos y de comunicaciones la información circula a través de múltiples dispositivos de procesamiento secuencial, en los cuales las unidades de información, para ser procesadas, deben "esperar su turno" ante cada dispositivo, lo que da lugar a inevitables colas de espera.

Dada la creciente importancia de las tecnologías de comunicaciones en las sociedades modernas, resulta indispensable disponer de modelos adecuados para describir y resolver los problemas causados por la formación de colas. En general, el objetivo de estos modelos es disponer de una representación matemática del sistema en que se producen las colas, a partir de la cual evaluar el comportamiento del mismo, así como diseñar técnicas y estrategias de procesamiento que optimicen su rendimiento. Tal optimización, en la mayoría de los problemas prácticos, consiste básicamente en minimizar los tiempos de espera y maximizar el número de clientes atendidos por unidad de tiempo, aunque en algunos casos se definen otras funciones objetivo de acuerdo con las características y fines de cada sistema concreto.

Los modelos elementales de colas consideran un solo servicio demandado por clientes que forman cola en espera de ser atendidos. Sin embargo en la práctica es frecuente encontrar situaciones mucho más complejas: capacidad finita de los buffers donde se acumulan los clientes hasta ser atendidos; llegadas de clientes en grupo (llegada en

lotes); varios servidores atendiendo a una misma cola; distintos grados de prioridad entre los clientes que acceden al servicio; colas en serie o en paralelo; protocolos que especifican que cada cliente debe pasar varias veces por el mismo procesador; la lista completa de todas las posibilidades sería interminable.

En la práctica es muchas veces imposible encontrar la expresión analítica de resultados de interés (como tiempo medio en cola, distribución del tamaño de la cola, etc.) por lo que la simulación se convierte en una herramienta imprescindible.

Generalidades sobre los procesos de colas.

El término *sistema de colas* se refiere en general a algún sistema compuesto por una o más unidades, llamadas *procesadores* o *servidores*, que se ocupan de realizar las tareas encomendadas por otras unidades, llamadas *clientes*, con la particularidad de que si durante algún intervalo de tiempo la llegada de clientes supera la capacidad de procesamiento del sistema, dichos clientes permanecen en cola hasta que sean servidos.

Describimos a continuación los elementos básicos que permiten caracterizar el funcionamiento de estos sistemas y, por tanto, la formación y evolución de colas en los mismos:

a) El *proceso de llegadas* de los clientes: las llegadas de los clientes se producen, en general, en instantes aleatorios, de ahí que la sucesión de los mismos constituya un proceso puntual. La especificación completa del proceso de llegadas requiere muchas veces determinar si los clientes se agrupan en clases según su procedencia (las llegadas de clientes de dos orígenes distintos se producen según procesos distintos), prioridad o algún otro factor relevante.

b) El *tiempo de servicio:* el tiempo que se emplea en procesar cada tarea que los clientes encomiendan al procesador es, por lo general, también aleatorio.

c) El *número de servidores* de que dispone el sistema: la velocidad con que se reduce el tamaño de una cola depende decisivamente de este número que, en general, no es aleatorio.

d) La *disciplina de la cola*: ésta es la regla que determina el orden en que los clientes son elegidos de la cola para entrar en el servicio (primero en entrar-primero en salir, último en entrar-primero en salir, atención según prioridades, atención en orden aleatorio, etc.); esta disciplina es determinante en la evolución temporal del tamaño de la cola, así como en el tiempo que permanece en la misma cada cliente.

e) La *organización del servicio*: el sistema puede ser tal que cada cliente deba recibir más de un servicio (o incluso recibir cada servicio más de una vez), en cuyo caso es preciso determinar en qué orden deben recibirse los servicios, cuántos servidores hay disponibles para cada uno de ellos y como se produce el tránsito de un servicio a otro.

Con objeto de establecer una clasificación de los distintos modelos de colas en los sistemas con un solo servicio, es útil conocer la notación de Kendall, consistente en un un código tripartito *a/b/c*, donde *a* especifica la distribución de los tiempos entre llegadas, *b* la de los tiempos de servicio y *c* el número de servidores. Los códigos más habituales para estas distribuciones son M (exponencial), G (general), y D

(determinista); en conjunción con éstos puede aparecer el código I, que indica independencia. Así, por ejemplo:

M/M/1: Llegadas según un proceso de Poisson homogéneo (y por tanto Markoviano), tiempos de servicio exponenciales y un solo servidor.

GI/M/m: Tiempos entre llegadas independientes, con distribución general, tiempos de servicio exponenciales y *m* servidores.

G/GI/m: Distribución general para los tiempos entre llegadas y de servicio, siendo éstos últimos independientes; m servidores.

A veces a este código se le añade un cuarto término numérico que representa el tamaño del buffer en caso de que éste sea finito. Un quinto término, también numérico, indicaría que la población de que proceden los clientes es finita y del tamaño consignado.

Los principales procesos de interés en el estudio de un sistema de colas son los siguientes (donde C_n representa al n-ésimo cliente en llegar al sistema):

T_n = : tiempo entre las llegadas de C_n y C_{n-1}.
X_n: tiempo de servicio de C_n.
W_n: tiempo que permanece C_n en cola.
$S_n = W_n + X_n$: tiempo que permanece C_n en el sistema.
$N(t)$: número de clientes en el sistema en el instante t.
$Q(t)$: número de clientes en cola en el instante t.
$V(t)$: tiempo que falta, en el instante t, para que se vacíe el sistema.

En la caracterización de los procesos de colas juega un importante papel el concepto de ***factor de utilización*** ρ, definido como:

$$\rho = \frac{E[X_n]}{mE[T_n]}$$

donde m el número de servidores en el sistema. El significado de ρ es intuitivamente claro si ρ se expresa de la forma:

$$\rho = \frac{t / E[T_n]}{mt / E[X_n]}$$

En un periodo largo de tiempo, de duración t, el numerador de esta expresión representa el número el número medio de tareas que llegan al sistema. Por su parte, el denominador es el número medio de tareas que el sistema puede completar durante ese tiempo. El cociente es, pues, la razón entre el número medio de llegadas y el de servicios en largos periodos de tiempo; o, dicho de otra forma, es la fracción de la capacidad de atención del sistema que se usa por término medio. Es evidente que si $\rho > 1$, el número de llegadas supera a la larga el número posible de servicios, y por tanto la cola crecerá sin límite. Por el contrario, si $\rho < 1$ se producen en media menos llegadas que servicios; ello quiere decir que las colas que puedan formarse terminan vaciándose, dando así lugar a un proceso cíclico de crecimiento - vaciado de la cola, que puede

considerarse estable, ya que da lugar a lo sumo a tiempos finitos de espera y a longitudes finitas de la cola. Se dice entonces que la cola se encuentra en equilibrio.

La tasa con la que entra el *trabajo* al sistema recibe el nombre de ***intensidad o flujo de tráfico, A*** y se expresa normalmente en ***Erlangs***. En el caso de sistemas con un único servidor la intensidad de tráfico coincide con el factor de utilización ρ, mientras que en el caso de varios servidores (por ejemplo, m), la intensidad de tráfico es $m\rho$. Cuando la capacidad de la cola es finita, una parte del tráfico que entra en el sistema puede ser rechazado, en este caso, el factor de utilización se calcula en base al **tráfico cursado**.

En general si $\rho<1$, cuando $t\to\infty$ existen distribuciones límite para las variables aleatorias citadas más arriba, y existe la posibilidad de estudiar el sistema de colas por medio de dichas distribuciones límite, lo cual simplifica notablemente los modelos al eliminar la dependencia del tiempo. Es de destacar que en la práctica, el utilizar las distribuciones límite no impone graves restricciones, pues normalmente los sistemas operan durante un tiempo suficiente para que se alcance el equilibrio.

Algunos resultados interesantes sobre los sistemas de colas en equilibrio son los siguientes: (cuando el sistema está en equilibrio denotamos por T, X, W y S, respectivamente, los tiempos entre llegadas, de servicio, en cola y en el sistema, y por N y Q, el número de clientes en el sistema y en cola respectivamente)

i) $E[S] = E[W] + E[X]$

Este resultado permite interpretar E[W] como el coste (en tiempo) que supone compartir el servicio con otros usuarios, ya que es el tiempo de más que hay que permanecer en el sistema para obtener el servicio.

ii) $E[N] = E[S] / E[T]$ ***(Fórmula de Little)***

Llamando $\lambda = 1/E[T]$ (tasa media de llegada de clientes) la fórmula de Little puede escribirse como $\boldsymbol{E[N] = \lambda E[S]}$, lo que indica que en el equilibrio el número medio de clientes en el sistema coincide con los que, por término medio llegan durante el tiempo de servicio de un cliente.

iii) $E[Q] = E[W] / E[T]$

Este resultado es similar al anterior, y también suele expresarse como $E[Q] = \lambda E[W]$ (el número medio de clientes en cola coincide con los que, por término medio, llegan durante un tiempo de espera)

Como ya hemos indicado, los procesos de nacimiento y muerte constituyen un caso particular de cadena de Markov en tiempo continuo; se caracterizan porque en cualquier intervalo de tiempo infinitesimal sólo son posibles transiciones de un estado E_k al E_{k+1} (nacimiento) o al E_{k-1} (muerte), con independencia de la historia anterior del proceso. En el contexto de las colas, el estado E_k representa que en el sistema hay k clientes; un nacimiento significa una llegada al sistema y una muerte una salida del mismo. Llamemos $p_k(t) = P(N(t)=k)$ y:

λ_k = tasa media de llegadas cuando en el sistema hay k clientes.

μ_k = tasa media de servicio cuando en el sistema hay k clientes.

(esto es, cuando hay k clientes en el sistema el tiempo medio entre llegadas es $1/\lambda_k$ y la duración media del servicio es $1/\mu_k$). Si para algún k_0 se cumple:

$$\lambda_k / \mu_k < 1 \quad \forall k > k_0$$

(esto es, de un k_0 en adelante la tasa media de llegadas es menor que la tasa media de servicio) entonces el sistema alcanza el equilibrio y existe:

$$p_k = \lim_{t \to \infty} p_k(t)$$

Esta probabilidad p_k representa la fracción de tiempo durante la cual el sistema contiene k clientes, una vez alcanzado el equilibrio. Para calcular las p_k , de acuerdo con la definición de proceso de nacimiento-muerte, puede asumirse que en cada intervalo de tiempo de duración infinitesimal Δt sólo puede producirse una llegada, con probabilidad $\lambda_i \Delta t + o(\Delta t)$, o una salida, con probabilidad $\mu_i \Delta t + o(\Delta t)$, siendo i el número de clientes en el sistema al comienzo del intervalo, y o(Δt) un infinitésimo de orden superior a Δt; entonces:

$$p_k(t+\Delta t) = p_{k-1}(t)\lambda_{k-1}\Delta t + p_k(t)(1 - \lambda_k \Delta t - \mu_k \Delta t) + p_{k+1}(t)\mu_{k+1}\Delta t + o(\Delta t), \; k>0$$

$$p_0(t+\Delta t) = p_0(t)(1 - \lambda_0 \Delta t) + p_1(t)\mu_1 \Delta t + o(\Delta t)$$

Reordenando estas ecuaciones y tomando límite cuando $\Delta t \to 0$ se obtiene:

$$p_k'(t) = \lambda_{k-1}\, p_{k-1}(t) - (\lambda_k + \mu_k)\, p_k(t) + \mu_{k+1}\, p_{k+1}(t), \quad k>0$$

$$p_0'(t) = -\lambda_0 p_0(t) + \mu_1\, p_1(t)$$

Cuando se alcanza el equilibrio, las pk no dependen de t, y por tanto las derivadas anteriores se anulan, resultando:

$$(\lambda_k + \mu_k)\, p_k = \lambda_{k-1}\, p_{k-1} + \mu_{k+1}\, p_{k+1}, \quad k>0$$

$$\lambda_0 p_0 = \mu_1\, p_1$$

Estas ecuaciones pueden resolverse recursivamente, obteniéndose:

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}$$

y el valor de p_0 puede calcularse sin más que imponer que $\sum_{k=0}^{\infty} p_k = 1$, de donde:

$$p_0 = \left(1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}\right)^{-1}$$

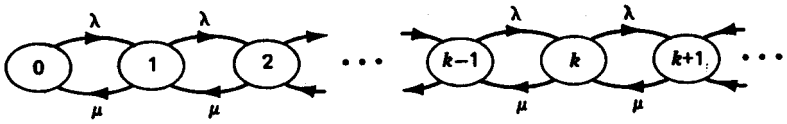
En general, incluso para sistemas *G/G/1*, se cumple que:

$$\rho = 1 - p_0$$

Estos resultados son directamente aplicables a los modelos elementales cuyas características más relevantes comentamos a continuación. Supondremos en todos los modelos que los clientes son atendidos según una política FIFO (*First In First Out*, el primero que llega es el primero en ser atendido).

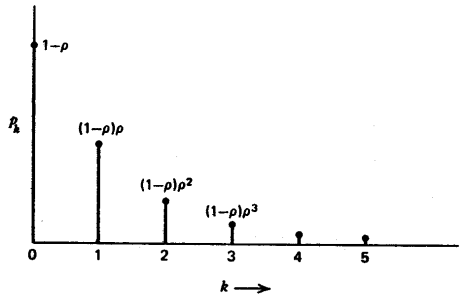
Colas M/M/1

Este modelo es el más sencillo: las llegadas se producen según un proceso de Poisson homogéneo de parámetro λ, los tiempos de servicio son independientes con distribución exponencial de parámetro μ, y hay un solo procesador.



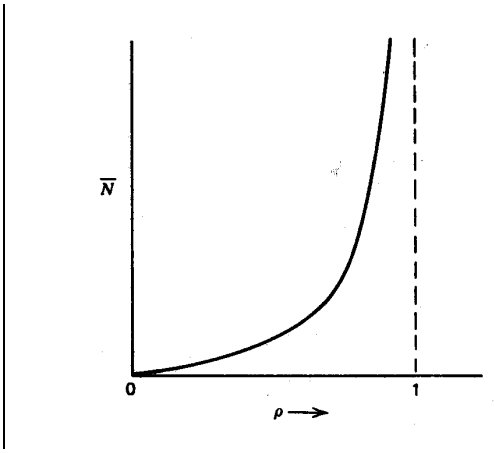
En este caso $\lambda_k = \lambda$ y $\mu_k = \mu \ \forall k$, y dado que $E[X] = 1/\mu$, resulta $\rho = \lambda/\mu$. Utilizando las expresiones de las p_k calculadas en el apartado anterior, se obtiene para el número de clientes en el sistema una **distribución geométrica**:

$$p_k = (1-\rho)\,\rho^k, \quad k=0,1, \ldots$$



de donde,

$$E[N] = \rho/(1-\rho)$$

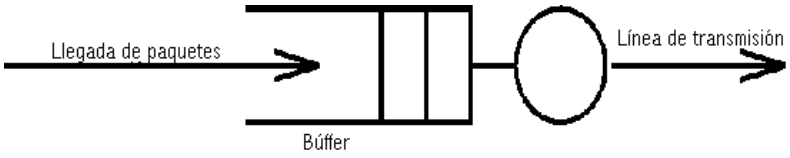


y utilizando la fórmula de Little:

$$E[W] = \frac{\rho/\mu}{1-\rho}, \qquad E[S] = \frac{1/\mu}{1-\rho}$$

Debe señalarse el hecho de que *E[N], E[W]* y *E[S]* son inversamente proporcionales a 1-ρ. Ello significa que a medida que ρ se aproxima a 1 el número medio de clientes en el sistema, así como los tiempos de espera crecen ilimitadamente.

Ejemplo



- Llegada de paquetes, λ= 2000 p/s
- Capacidad del enlace, *C*=1544 Kb/s

Distribución de la longitud del tamaño del paquete, exponencial de media, *L*=515 b/p

- Por tanto, el tiempo de servicio es exponencial de media:

$$\frac{1}{\mu} = \frac{L}{C} = \frac{515\,b/p}{1544\,Kb/s} \approx 0.33 \text{ ms/p}$$

  es decir, los paquetes son atendidos a una tasa de μ= 3000 p/s

- Utilizando las fórmulas de M/M/1:

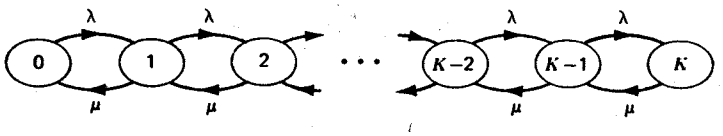
$$\rho = \frac{\lambda}{\mu} = 0.67$$

$$E[N] = \frac{\rho}{1-\rho} = 2.0 \text{ paquetes}$$

$$E[T] = \frac{E[N]}{\lambda} = 1.0 \text{ ms.}$$

Colas M/M/1/K (sistema de capacidad finita K)

Este modelo es análogo al anterior, salvo que se supone que el buffer del sistema tiene capacidad finita K, esto es, en cada instante el sistema puede albergar como máximo K clientes (incluido el que está recibiendo servicio).



Esta situación puede ser tenida en cuenta por el proceso de nacimiento muerte si se definen:

$$\lambda_k = \begin{cases} \lambda \ si k < K \\ 0 \ si k \geq K \end{cases}$$

$$\mu_k = \mu, \quad k=1,2,...,K$$

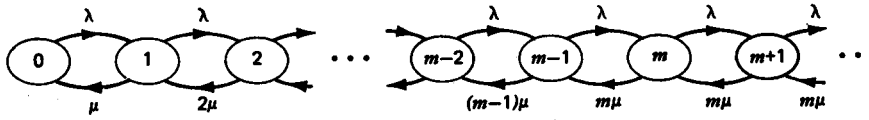
Este sistema siempre alcanza el equilibrio, pues para $k \geq K$ es $\lambda_k / \mu_k = 0$. Sustituyendo en la expresión de los p_k del proceso de nacimiento - muerte se obtiene sin dificultad:

$$p_k = \begin{cases} \dfrac{1-(\lambda/\mu)}{1-(\lambda/\mu)^{K+1}} \left(\dfrac{\lambda}{\mu}\right)^k & 0 \leq k \leq K \\ 0 & k > K \end{cases}$$

Un caso particular interesante se produce cuanto $K=1$.

Colas M/M/m

Constituyen la generalización del modelo *M/M/1* al caso de *m* servidores; supondremos que el cliente que ocupa la cabeza de la cola es atendido por el primer servidor que queda libre.



Como en el caso *M/M/1*, la tasa de llegadas es $\lambda_k = \lambda \ \forall k$, y la tasa de servicio de cada servidor es también constante μ; ahora bien, dado que hay *m* servidores, la tasa de servicio global del sistema es:

$$\mu_k = \begin{cases} k\mu & si \quad 0 \leq k \leq m \\ m\mu & si \quad k > m \end{cases} = min\{k\mu, m\mu\}$$

y la intensidad de tráfico es $\rho = \lambda / m\mu$. Aplicando la fórmula de los p_k del proceso de nacimiento-muerte, se obtiene sin dificultad:

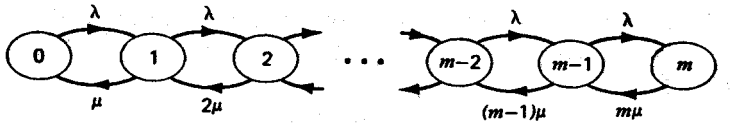
$$p_k = \begin{cases} p_0 \dfrac{(m\rho)^k}{k!} & si \ \ 0 \leq k < m \\ p_0 \dfrac{\rho^k m^m}{m!} & si \ \ k > m \end{cases} \qquad p_0 = \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!(1-\rho)} \right]^{-1}$$

Un resultado de interés, muy útil en telefonía, es la ***fórmula de Erlang C*** que permite calcular la probabilidad de que un nuevo cliente que llega al sistema tenga que hacer cola (corresponde a la probabilidad de mantener una llamada telefónica en espera por encontrar todas las líneas ocupadas):

$$p_Q = \sum_{k=m}^{\infty} p_k = \frac{\frac{(m\rho)^m}{m!}\left(\frac{1}{1-\rho}\right)}{\sum_{k=0}^{m-1}\frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!(1-\rho)}}$$

Colas M/M/m/m

El caso particular de que el sistema rechace a todo cliente que no pueda ser atendido inmediatamente (en el ejemplo de la telefonía ello significa que no se admiten llamadas en espera)



puede modelarse tomando:

$$\lambda_k = \begin{cases} \lambda & si\ \ k < m \\ 0 & si\ \ k \geq m \end{cases}$$

$$\mu_k = k\mu,\ \ k=1,2,...,m$$

con lo que se obtendría:

$$p_k = \frac{(\lambda/\mu)^k / k!}{\sum_{i=0}^{m}(\lambda/\mu)^i / i!}$$

que para el caso de p_m se conoce como ***fórmula de Erlang B,*** $E(m, \lambda/\mu)$ o bien $E_m^{(1)}(\lambda/\mu)$, y fue obtenida por primera vez por Erlang en 1917. Esta fórmula nos da la probabilidad de que el sistema esté ocupado por completo. En general, a la probabilidad de rechazar a un cliente se la conoce como **grado de servicio**.

Ejemplo 1

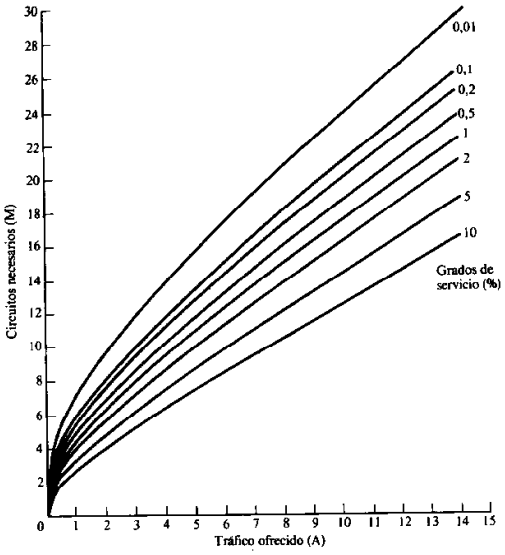
Supongamos 1000 terminales, cada uno de los cuales genera 0.1 Erlang en la hora cargada, conectados a un conmutador capaz de manejar 123 llamadas simultáneas.

Si llamamos A a la intensidad de tráfico:

A= 1000·0.1= 100
m= 123
E(123, 100)= 0.01 ó 1%

Es decir, si el 1% del tráfico se pierde, el flujo medio de tráfico soportado por los 123 enlaces internos del conmutador es 99 Erlangs y cada enlace tiene un factor de utilización de 99/123= 0.8 ó 80%.

En la práctica, el grado de servicio requerido *E(m,A)* y el tráfico ofrecido *A*, son datos y el objetivo es calcular M. Gráficamente:



Las curvas pueden aproximarse por líneas rectas para A grande, por ejemplo:

$M = 5.5 + 1.17 \cdot A$ para $E(M,A) = 1\%$
$M = 7.8 + 1.28 \cdot A$ para $E(M,A) = 0.1\%$

Ejemplo 2

Supongamos un sistema de conmutación de mensajes donde cada mensaje permanece en el nodo 100 ms. con una tasa media de 60 mensajes por segundo. En promedio habrá (A=)6 mensajes en cada momento. Para un grado de servicio de 0.1% necesitaremos reservar espacio para:

$M = 7.8 + 1.28 \cdot 6 = 16$ mensajes