Universidad de Cantabria

Departamento de Ingeniería de Comunicaciones

TESIS DOCTORAL

Contribución al Desarrollo de Herramientas Estratégicas para el Diseño, Dimensionado y Evaluación de Redes de Telecomunicación de Banda Ancha

Alberto Eloy García Gutiérrez

2009

Universidad de Cantabria

Departamento de Ingeniería de Comunicaciones

TESIS DOCTORAL

Contribución al Desarrollo de Herramientas Estratégicas para el Diseño, Dimensionado y Evaluación de Redes de Telecomunicación de Banda Ancha

Autor: *Alberto Eloy García Gutiérrez*
Director: *Klaus Dieter Hackbarth*

Santander, julio de 2009

A Mariví, Javi y Marijose
Os fuisteis antes de que
pudiera deciros adiós

Agradecimientos

Con este documento se cierra un ciclo (demasiado tiempo abierto) que ha marcado mi vida por muchas razones. Algunos dirían profesionalmente, y es cierto, durante este tiempo me he dado cuenta cuánto me gusta investigar, pero sobre todo, cuánto me gusta enseñar. Sin embargo, lo que más me ha marcado es la calidad humana del grupo de personas que un día me acogieron en su grupo y me dieron esta oportunidad.

Empezando por mi Director de Tesis, el Dr. Klaus Hackbarth por toda la paciencia que ha tenido conmigo, por su ayuda y su apoyo incondicional, pero sobre todo, porque en todos estos años no he tenido nunca la oportunidad de darle las gracias por haberme metido, a base de funciones de distribución, el gusanillo de la Telemática en el cuerpo y además haber sido mi Tutor en mucho más que el sentido académico de la palabra.

Por supuesto que tengo que agradecer a toda la plantilla telemática su amistad, y no doy nombres, porque todos los que me habéis apoyado ya sabéis que tenéis vuestro sitio en mi corazón y no hay más o menos.

A Juanito y Tita por todo lo que me han dado por nada: la vida.

A mis hermanos, los que están, a los que vinieron y a los que se fueron, por todo lo que me han dado, me dan y, espero, me darán.

A toda mi familia, de un lado, y del otro, porque lo que les doy, me lo devuelven con creces.

Y finalmente, a mi mujer, Inés, que está en todo y para todo conmigo,

A todos,

gracias.

Resumen

El rápido desarrollo de las tecnologías de redes de banda ancha, sobre todo en la parte de agregación, ha supuesto unos sigificativos cambios cualitativos en el mundo de las comunicaciones, pues ha acercado al usuario altos anchos de banda y servicios que hasta ahora le eran inaccesibles. Internet ha supuesto el despegue de las redes de conmutación de paquetes, las ha convertido en un medio al que todo el mundo quiere acceder, y ha obligado a replantear todas las previsiones que manejaban los operadores sobre el tráfico que iban a soportar sus redes en cada una de sus partes: el acceso, la agregación y la dorsal. Precisamente, la evaluación de estas y otras situaciones deben estar incluidas en el proceso de diseño y planificación de cualquier red, en lo que se ha denominado planificación estratégica. La previsión de las expectativas de los usuarios, los servicios existentes y futuros soportados por la red y, sobre todo, la estimación de la demanda de tráfico sobre cada elemento de la misma son algunos de los objetivos principales que se acometen en el proceso de planificación.

La estimación de la demanda de tráfico resulta especialmente importante porque permite al operador anticiparse a situaciones de congestión y bloqueo de la red. Utiliza herramientas de modelado del tráfico, bien para su cálculo analítico o para su estudio mediante simulación, y al ser una herramienta fundamental, existe un extenso catálogo de modelos de tráfico: genéricos y particularizados, de fuentes individuales y de redes completas, de elementos de interconexión y de enlaces.

Los modelos de tráfico de fuente son la opción más interesante, puesto que consideran el comportamiento del flujo de bits que se genera desde los propios elementos, a partir de sus parámetros fundamentales y constituyen el elemento básico para la estimación del tráfico multifuente. En esta Tesis, se han analizado los diferentes modelosexistentes y se ha desarrollado un nuevo modelo que tiene en cuenta el comportamiento del tráfico desde su preparación por parte de la capa de aplicación, hasta su inserción en la capa física. Su ventaja frente a los modelos existentes es que permite un cálculo rápido sin reducir la exactitud, siempre dentro de los límites de la planificación estratégica. El modelo propuesto parte de la solución ON-OFF tradicional, aunque aplicado a tres niveles diferentes (conexión, sesión y ráfaga), de forma que cada estado ON se encuentra modulado por el modelo correspondiente a la capa inferior. Como alternativa, se ha considerado que, el estudio de los valores máximos de tráfico, se puede llevar a cabo de forma determinista, mediante la observación de flujos reales, tal como propone la teoría del *Network Calculus*, y que ha sido evaluada y aplicada en este trabajo con ejemplos prácticos que se han incluido en el documento.

Sin embargo, cada flujo individual recorre la red, atravesando diferentes nodos de interconexión, encontrando otros flujos en su camino, y entonces, su comportamiento va a verse, posiblemente, modificado. Como alternativa más simple, surgen los modelos multifuente, que intentan introducir el efecto de la interacción

entre las diferentes fuentes, también conocido como agregación. Para considerar este caso, se ha realizado un estudio de modelos específicos para su aplicación a puntos de agregación, especialmente los situados en la red de acceso, al ser ésta la parte de la red que más alto coste presenta en su implementación. Partiendo del modelo ON-OFF para una sola fuente, se ha propuesto una variante multifuente, que aprovecha las propiedades de la función de distribución binomial para realizar el cálculo del tráfico agregado por un número determinado de fuentes. Al mismo tiempo, se ha estudiado también la opción planteada por el *Network Calculus*.

Finalmente, la aplicación de estos modelos depende fuertemente de la elección correcta del conjunto de parámetros de entrada que definen el comportamiento del tráfico. El estudio realizado en esta Tesis concluye que el tráfico de fuente está condicionado a tres premisas fundamentales: el usuario, el servicio y la tecnología de acceso. En este trabajo, se ha desarrollado un nuevo concepto que permite especificar de forma sistemática y ordenada todas las variaciones, en cuanto al tipo de fuentes que se puedan encontrar en una determinada red de acceso, que constituye todo el conjunto un escenario de red concreto. A esta nueva idea se la ha denominado CASUAL (Cubo de Acceso/Servicios/Usuarios de Asignación Libre) y ha sido aplicada en una herramienta realizada a su efecto, junto con los modelos de tráfico de fuente propuestos. Esta aplicación muestra un ejemplo básico de herramienta de planificación. Tras seleccionar previamente el tipo de usuarios, servicios y tecnologías de acceso, se configuran los parámetros que serán utilizados por los modelos de tráfico en cada combinación (tripleta), directamente con sus valores, o bien mediante capturas reales del tráfico. Posteriormente, se define el escenario a partir de los puntos de acceso y de los usuarios conectados. Los resultados pueden ser consultados en cada punto de la red, obteniendo las estimaciones de tráfico individual y agregado en cada caso.

Abstract

The rapid development of access network technologies has led to a revolution in the telecommunications realm, enabling users to access services to which they were not able to access so far. Internet has been the catalyst of the development of packet switched networks, which have become the resource which everyone wants to access. As a consequence, the operators need to recalculate all their forecasts about the traffic to be transported by their networks. In fact, the evaluation of these and other situations should be included within the design and planning phases of any network, in the so called strategic planning overall process. Anticipating user expectations, considering both existing and future supported services; and, above all, estimation of traffic demand for each network element are some of the main objectives of the planning process.

Traffic demand estimation is particularly interesting, since it allows the operator to anticipate bottlenecks and blocking situations of the network. It uses traffic modelling tools, either for their analytical or simulation-based study. Since they have become an essential tool, there is a wide range of traffic models: generic and particularized, for individual sources and complete networks, for interconnection elements and links.

Source traffic models are the most interesting ones, since they consider the behaviour of the bit stream which is generated from the network elements, by means of their basic parameters. In this work a new model has been developed; it takes into account the behaviour of traffic starting from its preparation at the application layer, to its insertion into the physical layer. The novel proposal is based on a traditional ON-OFF model, but applied to three different levels (connection, session and burst level). Each ON state is modulated by the model located at the lower layer. Alternatively, the study of peak traffic values can be carried out on a deterministic way, by means of the observation of real flows, as proposed by the Network Calculus theory, which has been evaluated and applied in practical examples.

However, each individual flow travels through the network across different interconnection nodes, finding other streams in its way, and thus, its behaviour might be modified. As a simpler alternative, it is worth mentioning multisource models, which introduce the effect of interaction between different sources, also known as aggregation. In this case, the thesis presents a study of specific models and their application to aggregation points within the access network. Based on the aforementioned ON-OFF model, a multisource variation is proposed, it exploits the properties of the binomial distribution function to calculate the aggregated traffic for a specified number of sources. At the same time, the Network Calculus based option is also analyzed.

Finally, the application of these models entirely depends on the appropriate choice of the set of input parameters which define the traffic behaviour. This study concludes

that source traffic is subject to three fundamental conditions: the user, the service and the access technology. This work introduces a new concept which brings about the systematic and ordered specification of all the different types of sources which can be found in a given access network, resulting, the whole scenario, in a specific network scenario. This new paradigm has been called CASUAL (Cube for Access/Services/Users with Free Allocation) and was implemented in a tool specifically deployed for that purpose, together with the proposed source traffic models. The application can be seen as a basic example of a planning tool. The first step is to select the type of users, services and access technologies to be used. Then, the parameters which will be employed by the traffic models are configured for each of the combinations, either straightforwardly (with their own values) or by means of real traffic captures. Afterwards, the scenario is defined as a number of access points and the users connected to them. Finally, results can be examined at every point within the network, obtaining the estimations of both individual and aggregated traffic in each case.

Índice

Indice de Figuras

Índice de Tablas

1. **Introducción y objetivos de la Tesis**

> **"*internet.***
>
> *1. amb. Red informática mundial, descentralizada, formada por la conexión directa entre computadoras u ordenadores mediante un protocolo especial de comunicación.*
>
> *ORTOGR. Escr. t. con may. inicial.***"**
>
> Real Academia de la Lengua: Diccionario de Lengua Española – Avance de la vigésima tercera edición

Nunca hasta ahora, una sola palabra había dado tanto significado a tantas cosas sin necesidad de tener ni idea de ninguna de ellas. Cada vez que una persona la pronuncia, automáticamente surge en su mente esa imagen tan manida, una tela de araña en la que todo es posible, la cual, permite que delante de un ordenador cualquier lugar esté cerca. Atrás quedaron los tiempos en los que los Ingenieros de Telecomunicación eran esos "pirados" que estaban delante del computador, nadie sabía muy bien para qué, obsesionados con términos raros, como *Hercio*, *Byte* o *Kilobit*. Ahora, un gran número de la población conoce lo que es un *Giga*, el *ADSL*, o cómo *WIFI* es sinónimo de "Internet gratis" a costa del vecino, y que, si no estás conectado, no tienes un perfil en *Facebook* o no descargas películas de estreno prácticamente no existes. Se ha pasado de un

escenario en el que la comunicación era algo accesorio, a otro en el que es el usuario el que pide, exige y puede ser un actor más que interactúa en un sector más de la denominada Sociedad de la Información. Cada día los operadores ofertan nuevos y mejores servicios, lo que hace imprescindible que Internet evolucione y se adapte a las nuevas necesidades.

Este proceso se ha venido repitiendo desde que Alexander Graham Bell descolgara por primera vez un teléfono (…¿O tendría que decir Antonio Meucci?). Una silenciosa maquinaria trabaja día a día para que, conforme crece la tela de araña, ésta no acabe rota por su propio peso: la planificación. Cada vez que un nuevo usuario contrata su acceso a Internet, un operador oferta un nuevo servicio o se desarrolla una nueva tecnología de acceso, puede producirse un "efecto mariposa", que afectara a todo Internet, si no fuera porque cada una de esas acciones ha sido prevista y/o estimada anteriormente. La planificación es un proceso que está presente desde que se diseña la red de comunicación, la observa, controla y gestiona durante su despliegue y desarrollo, y finalmente, la prepara y acomoda para evolucionar ante cualquier cambio.

Como tal, la planificación y el diseño de redes de comunicación se convierten en unas tareas complejas, que abarca desde el punto terminal de conexión de los usuarios hasta los nodos de conmutación, las interfases para la interconexión con otras redes y las interfases a grandes granjas de servicio y almacenes de datos y contenidos. En este contexto, se presentan múltiples frentes abiertos:

- Los usuarios conectados: Debe establecerse una clasificación de los modelos de potenciales usuarios que van a hacer uso de la red, conocer su situación y número real, además de predecir o estimar el aumento de la demanda de puntos de conexión.

- Los servicios ofertados por la red: Al igual que para los usuarios, es preciso conocer las características de todos los servicios de Telecomunicación actuales y futuros, su comportamiento y modelización dentro de la red (lo que implica un gran esfuerzo investigador en el desarrollo de modelos y su correspondientes algoritmos), así como, su índice de implantación en la sociedad y, en concreto, respecto a cada tipo de usuario individual.

- El comportamiento de los enlaces de intra e interconexión entre nodos de la red, o diferentes redes: Es preciso predecir cuellos de botella o desaprovechamientos de los enlaces, posibilitando calcular el ancho de banda por enlace que debe establecerse entre dos puntos, o detectar situaciones críticas y puntos de mal funcionamiento (con un alto peso en el desarrollo de algoritmos de cálculo).

- La evaluación de los costes procedentes de la introducción de nuevos servicios, el mantenimiento de los existentes, la conexión de nuevos usuarios, las inversiones de equipos, tanto al nivel de clientes finales(minoristas), como al de prestación a otros operadores (mayoristas).

- La evaluación de la tarificación de los distintos servicios y su relación con la política desreguladora de las Telecomunicaciones.

El análisis conjunto de todas estas variables solamente es posible mediante herramientas creadas específicamente para la resolución de problemas, la optimización y el dimensionado de todos y cada uno de los elementos de la red de comunicación, entre las que se encuadran las herramientas para la planificación estratégica (planificación a un horizonte más largo). Todas ellas, se basan en métodos predictivos y/o estimativos mediante los cuales es posible evaluar anticipadamente el comportamiento de la red en general, así como, de cada uno de los elementos de red en particular. Dentro del extenso conjunto de herramientas existentes destacan las técnicas de predicción y modelado de tráfico, aspectos de los que se va a presentar en esta Tesis un estudio pormenorizado, poniendo especial énfasis primero en el modelado de tráfico de fuente, y posteriormente, en su aplicación en el problema de la agregación de tráfico procedente de múltiples fuentes. A lo largo de este documento, se exponen las ideas que dieron lugar al desarrollo de esta Tesis, los estudios que se llevaron a cabo y, finalmente, todas sus aportaciones y propuestas, enmarcadas dentro del desarrollo de herramientas dirigidas hacia la planificación estratégica de redes de comunicación de banda ancha, principalmente basadas en el protocolo IP.

1.1. Motivación y objetivos

Desde su creación, el Grupo de Ingeniería Telemática del Departamento de Ingeniería de Comunicaciones de la Universidad de Cantabria viene desarrollando, dentro de una de sus líneas de investigación, herramientas informáticas encaminadas a facilitar y mejorar los procesos de diseño y dimensionado de redes. Las diferentes aplicaciones, sus sucesivas versiones y adaptaciones, abarcan la planificación tanto a nivel de redes de Área Local, como de Área Extensa y Corporativa, sobre tecnologías cableadas e inalámbricas. Las utilidades que incorporan estos programas los encuadran dentro de los objetivos de la planificación estratégica, y como tal, obtienen resultados específicos dentro de cada uno de los aspectos que la constituyen:

- Clasificación de usuarios: Fundamentada en su conocimiento desde el punto de vista del uso que hacen de las tecnologías de red y de las necesidades y resultados que exigen a las mismas.
- Clasificación y descripción de los servicios de telecomunicación: Cada arquitectura de red presenta características que la hacen idónea para uno u otro servicio, por eso resulta fundamental identificar mecanismos y soluciones que faciliten el diseño de nuevos servicios y la adaptación de los ya existentes.
- Descripción del tráfico de fuente asociado a servicios de telecomunicación: La estructura y funcionamiento de un determinado servicio es estudiado desde el punto de vista del tráfico que genera, aunque estratégicamente supone la necesidad de anticipar estos resultados incluso antes de que la red, y por supuesto el servicio, se encuentren aún disponibles.
- Determinación de la distribución y agregación del tráfico de fuente: El conocimiento del tráfico generado por una fuente debe ser ampliado para anticipar el comportamiento tanto de la red como del servicio a medida que éste atraviesa los diferentes nodos de interconexión, encontrándose, en su camino, con otros flujos de tráfico procedentes de otros usuarios, servicios o redes.
- Evaluación económica: A partir de las características de los servicios, y del conocimiento del tipo de usuario y el uso que hace de ellos, se desarrollan esquemas de coste y su correspondiente tarificación, ajustados a la política de desregulación de las telecomunicaciones, con el objetivo de establecer la correspondiente evaluación de costes que justifique cada una de las acciones que se lleven a cabo, pasadas, presentes, y si es posible, futuras.

Todos ellos presentan interrelaciones que hacen que cada solución tenga que ser analizada previamente desde el punto de vista de la influencia que pueda presentar respecto al resto de etapas del proceso de planificación. En la Fig. 1.1

puede observarse cómo al estudiar dichas relaciones, las diferentes etapas pueden clasificarse en dos niveles diferenciados según el tipo de evaluación que se realiza: técnica y económica.

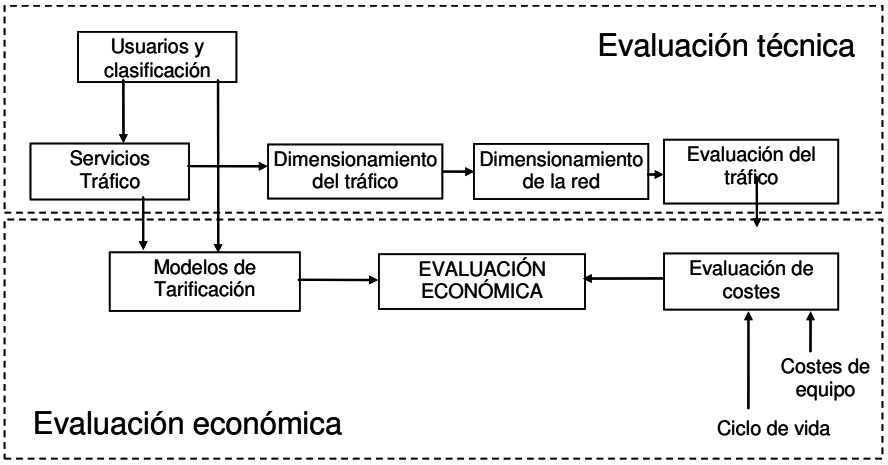


Fig. 1.1: El proceso de planificación de una red de comunicaciones

Esta Tesis se centra en las características específicas de los métodos utilizados durante el proceso de evaluación técnica, sobre las que se asientan la mayor parte de las aplicaciones desarrolladas por el grupo de investigación comentado anteriormente, en donde se han llevado a cabo todos los estudios, análisis e implementaciones incluidas en este documento. Más en detalle, se estudian los diferentes modelos de tráfico generado por los servicios aplicados por los usuarios, su idoneidad para la planificación estratégica y las mejoras de estos modelos para tal fin. Se pone especial interés en los puntos de accesso, en donde se produce la primera agregación, ya que es el punto más critico: por su alta distribución en la geografía cubierta por la red y por sus elevados costes de implementación, en comparación con puntos superiores de agregación.

A partir del conocimiento adquirido durante el desarrollo y aplicación de las herramientas que existían hasta ese momento, se identificaron todos aquellos puntos susceptibles de mejora y se plantearon nuevos desarrollos que permitieran adaptarlas a las tecnologías de acceso que comenzaban a imponerse. De esta manera, se fijó una serie de objetivos que han marcado el ritmo de todos los estudios realizados durante este trabajo, y que a continuación se pasa a enumerar desde los diferentes puntos de vista de la planificación:

- De la clasificación de usuarios: Los índices de penetración de Internet confirman que la Red es universal y que el uso de las redes de comunicaciones no es exclusivo del mundo empresarial. Como consecuencia, los usos y el comportamiento del tráfico también, por lo que es necesario acotar las nuevas condiciones asociadas a cada caso particular.

- De la clasificación de servicios: "*A rey muerto, rey puesto*", dice el saber popular. Un servicio es algo vivo que nace, se utiliza, y si no se adapta a los nuevos tiempos, muere. Los usuarios hacen uso de aquellos servicios que les aportan lo que ellos buscan, y no al revés. Por ello, si el uso depende del usuario, habrá que tenerlo en cuenta y por tanto, es necesario acotar las características asociadas a los diferentes servicios en función de sus usuarios y sus futuros desarrollos.

- De la descripción del tráfico de fuente: Internet añade un problema más a los dos anteriores, ya que introduce una alta diversidad de tecnologías de acceso y de redes, propiamente dichas, que forman parte de su entramado. Las características del tráfico (generado por los servicios, utilizados por los usuarios) suele presentar comportamientos bien definidos, pero que dependen de dicha tecnología. En estas condiciones, los modelos tradicionales de tráfico de fuente han tenido que evolucionar también, por lo que, para poder adaptar las herramientas existentes, es necesario determinar nuevas técnicas para la estimación de los valores de dicho tráfico y, a poder ser, establecer las características del tráfico que genera cada tipo de servicio en función del tipo de usuario y de la tecnología de acceso mediante la que se conecta.

- De la distribución del tráfico de fuente: La rígida estructura jerárquica de las redes tradicionales también se ha visto transformada. El tráfico de un servicio se junta, ya desde la propia red de acceso, al del resto de usuarios de su misma área de distribución. El comportamiento de cada flujo de paquetes individual se ve así condicionado al del resto, y plantea un nuevo problema, el de la agregación de tráfico. De nuevo, los modelos tradicionales también han tenido que evolucionar, por lo que es necesario establecer nuevos métodos que permitan obtener estimaciones válidas tanto de los tráficos individuales como de los agregados.

A estas cuatro grandes líneas maestras, sobre las que se ha desarrollado todo el trabajo que se presenta en los siguientes capítulos, se une una quinta y última, que ya desde los primeros estudios, cobraba especial importancia: la aplicabilidad de todas las propuestas y soluciones que fueran planteadas, en futuras aplicaciones estratégicas.

1.2. Estructura de la Tesis

A continuación, se comentan brevemente los contenidos de este documento, dividido en cinco capítulos, incluido este primero, de introducción, al que pertenece este apartado.

El capítulo 2 proporciona un amplio repaso de los principales conceptos teóricos sobre los cuales se fundamentan los estudios y propuestas que se van a presentar, estando organizado en cuatro apartados diferenciados. En el primero se desarrolla el concepto de planificación estratégica, aplicado al diseño de redes de comunicación, sus características, objetivos y, finalmente los métodos y herramientas que permiten llevarla a cabo. Precisamente, en el segundo apartado se estudia una parte importante de estas herramientas, el modelado de tráfico fuente, aplicado al caso concreto de Internet. Se realiza un recorrido por todos los modelos teóricos existentes y su aplicación en la estimación del mismo. Sin embargo, como se indicaba entre los objetivos, este modelado queda reducido cuando se produce la mezcla de múltiples fuentes, momento en el que se aplican modelos multifuente para la estimación de tráfico agregado, que se estudia en el tercer apartado de este capítulo. Por último, en el cuarto apartado, se muestra una alternativa basada en un modelo determinístico denominado *Network Calculus*, un nuevo marco teórico utilizado para el análisis de las limitaciones que los diferentes elementos imponen al tráfico de fuente, y que se presenta como una posible alternativa para el cálculo del tráfico agregado.

En el tercer capítulo se plantean y desarrollan las propuestas y aportaciones deque esta Tesis, repartidas en cinco apartados, donde, en el primero, se hace una descripción del escenario de trabajo, Internet y su tráfico, sobre el que se desarrollan y aplican el resto. En el segundo apartado se presenta un nuevo modelo conceptual, desarrollado a lo largo de esta Tesis, para la definición de escenarios de red de acceso, mediante el cual llevar a cabo la clasificación de usuarios, servicios y tecnologías requerida por la planificación estratégica. Para la aplicación del modelo anterior ha sido necesario relacionar cada elemento del escenario con sus correspondientes técnicas de modelado de tráfico de fuente, para lo que se ha desarrollado un nuevo modelo ON-OFF multinivel, que se presenta en el tercer apartado. A continuación, se propone una alternativa a los modelos matemáticos basada en el uso de curvas de llegada y optimización a partir de los límites del tráfico. Finalmente, en el quinto apartado se presentan dos variaciones sobre las propuestas anteriores, adaptadas al cálculo de tráfico agregado. La primera se ha desarrollado sobre un nuevo modelo de agregación de fuentes ON-OFF mediante aproximaciones basadas en binomiales. Por último, en el segundo caso, se propone de nuevo el uso de la metodología del *Network Calculus* a partir de curvas de llegada.

En el cuarto capítulo se propone un esquema para la aplicación conjunta de las ideas y resultados presentados en el capítulo anterior mediante la implementación de una herramienta estratégica. Para ello, el primer apartado describe las características fundamentales de dicha aplicación, así como su estructura y modo de funcionamiento. El segundo de los apartados muestra, a modo de ejemplo, la aplicación del modelo ON-OFF multinivel al caso concreto del servicio *Web*, para lo que se realiza una breve descripción previa del mismo, así como de los modelos de tráfico que han sido utilizados tradicionalmente.

Por último, en el capítulo 5 se hace una recapitulación de las ideas y soluciones presentadas, y, finalmente, se proponen algunas ideas que abren nuevas líneas de investigación sobre el tema.

2. **Conceptos Teóricos**

Este capítulo desarrolla la teoría sobre la que se asientan los estudios, ideas y resultados abordados en esta Tesis y que se exponen en este documento. Los avances en las comunicaciones interpersonales están íntimamente ligados al continuo desarrollo de las redes de comunicación existentes, la implantación de nuevas tecnologías y la implementación de nuevas alternativas de red. Es por ello que los operadores de redes de comunicación dan una especial importancia a la planificación y diseño de todas sus soluciones, mediante el uso de ingentes recursos técnicos y humanos. El uso de herramientas especializadas resulta así clave, y de esta manera el estudio y desarrollo de más y mejores mecanismos de cálculo, estimación y predicción pasa a ser prioritario. En esta Tesis se realiza un estudio de varias soluciones para la estimación de parámetros específicos en el proceso de planificación. Concretamente, el conocimiento a priori de los valores de tráfico, tanto de fuente como agregado, es parte fundamental de cualquier herramienta de aplicación, ampliamente estudiado pero con infinidad de soluciones en la mayoría de los casos de elevadísima complejidad computacional. Como alternativa, el uso de algoritmos más simples, basados en aproximaciones y simplificaciones de los modelos más complejos, son utilizados en algunas de las fases del proceso de planificación. En este documento se plantean dos métodos diferenciados para la obtención de estimaciones del tráfico de fuente y agregado, basados en modelos de tráfico de fuente y en una nueva teoría especializada en el análisis del tráfico denominada *Network Calculus*. La aplicación de estos métodos ha sido utilizada para el desarrollo de un nuevo concepto, presentado también en esta Tesis, el modelo

CASUAL, cuya implementación permite definir escenarios de red complejos y estimar las figuras de tráfico asociadas.

Cuatro son entonces las ideas básicas que se desarrollan en este capítulo:

- En la primera parte se expone una descripción detallada del concepto de planificación, sus variantes y fases de desarrollo, la importancia de definir un modelo de planificación de capacidades y la necesidad de implementar herramientas especializadas y adaptadas a las necesidades del operador y de la red.

- En el apartado 2.2 se define lo que es un modelo de tráfico, como mecanismo fundamental para la estimación de parámetros básicos de la planificación. Se parte del concepto general de modelo, su aplicación al caso concreto del tráfico de red y al modelado del tráfico de fuente, particularizado al diseño y planificación de redes y servicios en Internet.

- La mezcla de tráfico que se produce en cualquier punto intermedio de la red se conoce como agregación, concepto que se desarrolla en el apartado 2.3, presentando modelos de tráfico especializados, la mayoría derivados de los modelos de tráfico de fuente.

- Finalmente, en el apartado 2.4 se define y describe la teoría fundamental del *Network Calculus*, así como su aplicación en el modelado de tráfico, tanto de fuente como agregado.

2.1. Planificación y Diseño de Redes de Comunicación

Planificar es hacer que ocurran cosas que, de otro modo, no hubiesen sucedido, según expone W. E. Goetz[1] en. Por su parte Fremon Kast y James E. Rosenzweig definen planificación como *un proceso para decidir de antemano qué hacer, y cómo*. Por su parte R. Anthony completa la definición asegurando que *la planificación permite fijar las diferentes posibilidades, estableciendo los procedimientos a emplear y su secuencia de actuación, durante el tiempo e intensidad suficientes que aseguren los objetivos establecidos.*[2]

Al aplicar las definiciones anteriores a la planificación de redes y/o servicios de telecomunicación, el resultado es un proceso fundamental que debe ser completado antes del establecimiento de dicha red/servicio, mediante el cual se asegura que todos sus requerimientos, condiciones y prestaciones cumplen con todas las expectativas previstas por el operador y esperadas por el cliente. Este proceso define y estima las necesidades del cliente, estableciendo las capacidades de tráfico asociadas. De acuerdo al plan de negocio del operador, establece las capacidades totales del sistema y define todos los requerimientos asociados a cada elemento individual de la red. Todo el proceso es completado mediante el modelado del comportamiento, tanto de los usuarios como de la red en sí, aportando la realimentación necesaria para la corrección y ajuste de todos los parámetros fundamentales de la red.

2.1.1. El proceso de planificación

Desde el punto de vista de las Redes de Comunicaciones, el proceso de Planificación comienza con la localización geográfica de puntos de presencia (POP) del Operador de la Red, así como la previsión de tráfico asociada a cada nodo[3]. De acuerdo con estas previsiones, el objetivo fundamental de dicho proceso es realizar el diseño y dimensionado de la red basándose en las arquitecturas correspondientes que realicen el correspondiente transporte del tráfico de usuario, materializado en la elección previa de arquitectura, tanto

[1] W. E. Goetz: "*Management planning and context*", McGraw–Hill, 1949
[2] R. Anthony: "Planning & control systems: a framework for analysis", Boston, MA: Harvard University Press, 1965
[3] Como ejemplo, en la RPT/RDSI se observa que estos puntos constituyen el entramado de distribución (MDF) y se sitúan en el primer conmutador/concentrador local, donde termina el bucle de usuario.

para los nodos como la red en sí, así como de la estrategia de protección y recuperación (ante fallos) correspondiente, tal como se indica en [Vigo, 2000] y [Arijs, 2000]. De esta manera, el diseño de la red va a facilitar el enrutamiento de todo el tráfico inter-nodal, teniendo en cuenta aspectos tales como el mantenimiento de la calidad de transmisión (a nivel físico), o la consideración de diferentes escenarios de error. A menudo, la fase de diseño suele fundamentarse en un pequeño conjunto de escenarios de error, por ejemplo, fallos en enlaces individuales. Sin embargo, cuanto mayor sea el conjunto de escenarios, mayor fiabilidad presentará el diseño final.

Con todo, el principal objetivo de la fase de diseño es la minimización de los costes, especialmente los asociados a la estructura de la red que generalmente puedan y deban ser presupuestados. Esta minimización está limitada por las condiciones de fiabilidad y restauración en caso de fallos.

Una vez establecido el diseño, se desarrolla la planificación en sí, entendida como la optimización, en muchos casos iterativa, del diseño original en función de los escenarios de red establecidos. El principal objetivo de esta fase es definir e inicializar el conjunto de parámetros asociados con cada elemento de red. De esta forma la planificación puede ser considerada, tal como se muestra en la Fig. 2.1, desde el punto de vista estratégico, es decir, utilizar el conjunto de escenarios para realizar la estimación de capacidades y de los correspondientes costes asociados, o bien desde el punto de vista táctico, centrándose en los elementos individuales de la red, la distribución de los diferentes recursos y su configuración.

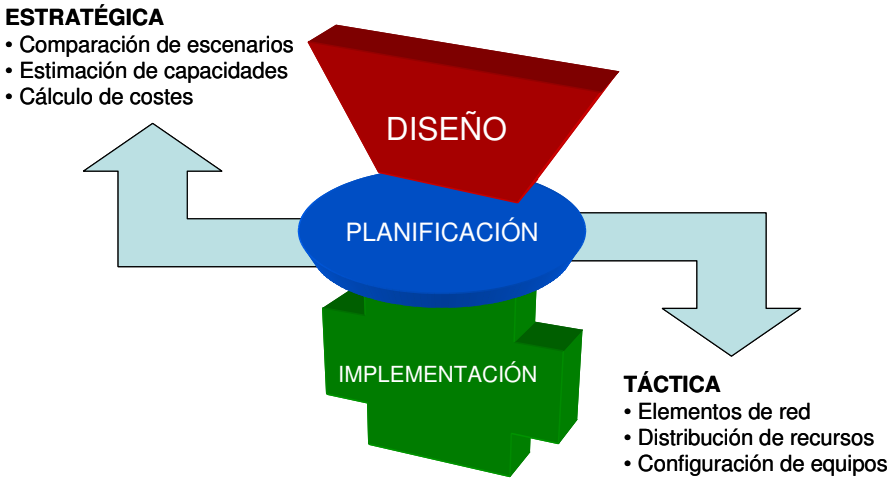


Fig. 2.1: El proceso de planificación desde el punto de vista del Operador de la Red

En la fase posterior de evaluación de la planificación, la red diseñada debe ser evaluada mediante los correspondientes análisis de su comportamiento ante escenarios de error inesperados y condiciones dinámicas de tráfico, como medida de su sensibilidad ante fluctuaciones de los parámetros de entrada. Esta

fase de evaluación hace las veces de etapa de realimentación para la mejora del proceso de diseño, adaptando la solución final de forma iterativa.

De acuerdo con todo lo anterior, la planificación puede ser considerada tanto desde los puntos de vista estratégico y táctico, como desde el punto de vista meramente operativo, estando relacionados todos ellos como etapas diferenciadas pero dependientes, tal como muestra la Fig. 2.2.

Fig. 2.2: Desarrollo del ciclo de planificación

- El ciclo de Planificación Estratégica lleva a cabo el estudio de la evolución de las demandas requeridas por los servicios soportados por la red y los objetivos del operador, como por ejemplo la reducción de los costes, el mantenimiento de la calidad o la maximización de la fiabilidad. Normalmente se desarrolla en períodos más o menos extensos de tiempo (de 3 a 5 años), dada la necesidad de establecer y definir muy concretamente los objetivos a largo plazo.
- El ciclo de Planificación Táctica estudia el comportamiento actual y analiza todas las posibles alternativas. Para ello se vuelve a hacer uso del diseño y modelado de redes, siendo su duración de 1 a 3 años.
- El ciclo de planificación operativa es una de las partes integradas dentro del proceso de mantenimiento. Su duración debe coincidir con el ciclo de vida operativo de la red. Básicamente consiste en chequeos, normalmente anuales, mediante el uso de herramientas de gestión y control de redes (*NMC-Network Management & Control tools*).

De esta forma la planificación es un proceso que se desarrollará durante todo el ciclo de vida de la red: establece sus orígenes, determina su crecimiento y evolución, y por último, decide su eliminación (migración a nuevas tecnologías, topologías o estructuras).

2.1.1.1. **Planificación Estratégica**

Es el ciclo de mayor duración dentro del proceso de planificación. Entre sus finalidades destaca asegurar el cumplimiento de los objetivos y metas del operador. Para ello es fundamental que permita mantener su competitividad frente al resto de operadores y redes coexistentes, sin perder su flexibilidad para poder evolucionar de acuerdo a las exigencias impuestas por dicha competencia.
Esto supone el estudio constante de las nuevas tecnologías, de la evolución de las ya existentes y de los estándares y recomendaciones relacionadas, analizando todas sus posibles aplicaciones, e incluso su compatibilidad dentro de la estructura global de la red.

Teniendo en cuenta todas estas consideraciones el grupo de planificación debe asegurar la consolidación de las tecnologías y estructuras utilizadas, e incluso ser capaz de llegar a determinar, si las circunstancias así lo requieren, la correspondiente migración hacia nuevas tecnologías, infraestructuras o topologías.

2.1.1.2. **Planificación Táctica**

Una vez que se ha definido de forma clara y contundente cuáles son las vías de migración a largo plazo, la compañía debe definir el conjunto de soluciones viables a seguir. En este nuevo ciclo se determina de forma precisa aquellas soluciones que posibiliten la consolidación de una determinada tecnología o estructura de la red ya en uso, o por el contrario, la propuesta de nuevas alternativas.

Durante este período se cuantifican las relaciones existentes entre los productos o servicios previstos, así como el comportamiento real de la red. Estos análisis aseguran la viabilidad de una determinada tecnología, o la integración de sistemas a varios niveles, de acuerdo con los requerimientos de productividad deseados, es decir, la mejor relación coste/producto/ servicio.

El proceso de planificación táctica sigue un esquema de desarrollo más o menos fijo:

1. Modelado de sistemas: ya sea de los sistemas existentes o de los nuevos sistemas. Tiene en cuenta tanto los equipos, como los servicios, las aplicaciones y los usuarios, e incluso la relación de todos ellos en la productividad de la red.
2. Determinación del ciclo de vida del modelo de tráfico para cada servicio/ aplicación.

3. Valoración de nuevas tecnologías y estándares asociados a cada servicio.
4. Determinación de diseños viables, ya sea para una o varias estructuras de red, con sus correspondientes modelos de costes y comportamiento.
5. Dimensionado de acuerdo con los modelos aplicados, de forma que se asegure un servicio ininterrumpido para todos los usuarios.
6. Por último, selección de la solución más deseable, o en su defecto, de menor riesgo, poniendo especial atención a las posibilidades de gestión de la misma.

Uno de los grandes aliados de la planificación táctica es el continuo desarrollo de sistemas y herramientas para el diseño, análisis y modelado de dispositivos, protocolos e incluso redes completas. El uso de la simulación en el diseño suele ser complicado, ya que ésta solamente resulta válida en regímenes permanentes, y requiere gran cantidad de recursos computacionales. Sin embargo, los avances informáticos han ido posibilitando la aparición de nuevas herramientas que permiten modelar redes de nueva generación en tiempo real.

2.1.1.3. **Planificación Operacional**

Este desarrollo se extiende a lo largo del ciclo de vida de la misma red. Básicamente consiste en el chequeo periódico de la misma, dirigido a asegurar la correcta interconexión de los equipos y mantener la calidad de los servicios provistos. En definitiva, la red queda monitorizada por un sistema de control y gestión denominado "Gestión de Red".

De igual forma que en el proceso de planificación táctica, este ciclo queda determinado por una secuencia de operación fija:

1. Gestión de la configuración, relacionada con la composición, topología y estructura en tiempo real de la red, haciendo uso generalmente de bases de datos relacionales.
2. Gestión de fallos, encargada de la supervisión del estado de la red, el testeo de extremo a extremo, el diagnóstico de errores, notificación de alarmas, realización de copias de seguridad, reconfiguración, etc.
3. Gestión de operación, se ocupa de la monitorización y definición del comportamiento en general, las tendencias y umbrales en tiempo real, etc.
4. Gestión de contabilidad, relacionada con la planificación de los presupuestos relacionados, determinación de costes hacia los usuarios finales, o la verificación de los objetivos del sistema.
5. Gestión de seguridad, lleva a cabo el establecimiento y mantenimiento de los criterios de gestión de los accesos, así como el control de reconfiguración para cualquier partición definida.
6. Como resultado de la gestión combinada, se lleva a cabo la planificación en sí de la red, que consiste en la optimización periódica

de la misma, el modelado de la recuperación de errores ante contingencias, y el diseño de estrategias de ayuda al usuario para la consecución de sus objetivos, tanto a corto como a largo plazo.
7. Gestión del funcionamiento, se ocupa de asegurar la eficiencia de todos los centros de operación, incluyendo las tareas de provisión de personal, entrenamiento y control del flujo de información.
8. Facilidades software, en relación con la gestión y actualización de los paquetes de software existentes, así como el diseño de nuevos paquetes optimizados.
9. Control integral de sistemas, es decir, la capacidad de gestionar y controlar redes heterogéneas de forma transparente desde cualquier consola de usuario.

Como resultado, la planificación operativa intenta mantener el coste efectivo de las subredes frente a cualquier perturbación dentro de la red global, ya sea motivada por el tráfico, los servicios o los fallos y averías.

2.1.1.4. **Modelo de planificación de capacidades**

El continuo desarrollo de las tecnologías de red y de los servicios de telecomunicación supone un continuo cambio en las necesidades de ancho de banda exigido. Para determinar los recursos necesarios de los sistemas implicados se utilizan modelos de planificación de capacidades, en base a la calidad de las aplicaciones tal y como son percibidas por el usuario.

En [Wu, 2004] y [Wu, 2003] se define la planificación de capacidades como el proceso de diseño y dimensionado de redes en función de la demanda esperada, de forma que la red soporte el crecimiento de las demandas de usuario manteniendo un servicio satisfactorio, es decir, unas determinadas calidades de servicio. Clientes, servicios y operadores establecen a través de la red una relación singular, representada en la Fig. 2.3, incluso antes de que el servicio o la red hayan sido implementados.

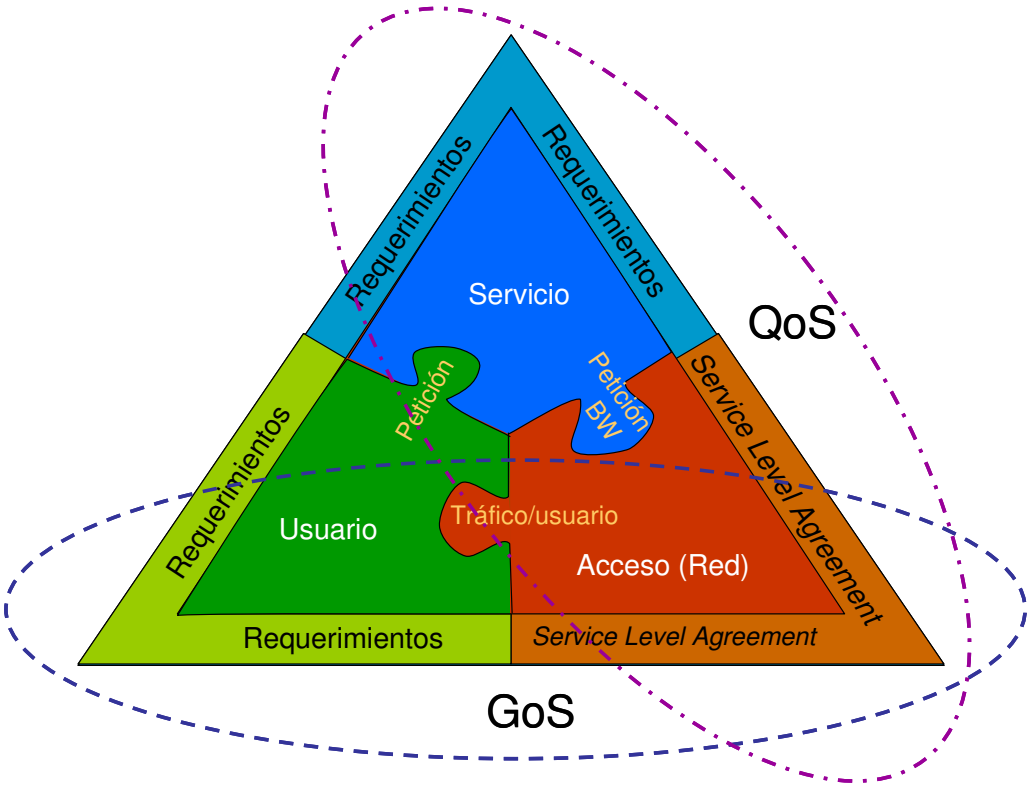


Fig. 2.3: Interrelación Usuario/Servicio/Operador desde el punto de vista del modelo de capacidades

De esta manera, los modelos utilizados tienen por objetivo principal estimar las cargas y demandas de tráfico para todos los servicios, asegurando la optimización de costes dentro de los parámetros de calidad esperados. El usuario se encuentra íntimamente ligado a ellos, de la misma forma que lo están los servicios en sí, las necesidades particulares de cada cliente, o las necesidades del sistema. Sin embargo, en la mayoría de los casos, el operador acota dicha influencia ofertando escenarios de servicio con parámetros pre-establecidos a los cuales el usuario se acoge. Adicionalmente, estos modelos consideran las influencias de la arquitectura y del hardware implementado.

De forma general, el modelo de capacidades considera seis variables fundamentales:

1. Los usuarios, porque son los que interactúan con la red, establecen necesidades e imponen prioridades. Es fundamental conocer su número y localización dentro de la red, y siempre que sea posible, su distribución en la misma (en forma de ley que permita generalizar en grupos de usuarios)

2. Los servicios y aplicaciones, es decir, el conjunto de servicios provistos por la red, y el tipo de aplicaciones que hacen uso de los mismos.
3. El ancho de banda, entendido como la necesidad del usuario al imponer una serie de prioridades a cada servicio, lo cual se traduce en requerimientos de recursos diferentes para cada tipo de servicio y aplicación. Un servicio activado por un usuario causa una demanda de ancho de banda a los elementos de red, y que éste proporciona en relación con su capacidad total.
4. La calidad y el grado de servicio (QoS + GoS), como medidas directas de la aceptación por parte del usuario de los servicios y aplicaciones provistos. Están directamente relacionadas con la sincronización y cuantificación de la cadena de procesos directos que componen el desarrollo de cualquier servicio.
5. Arquitectura de red, esto es, la estructura de distribución de los sistemas implicados en la provisión de los servicios. Impone en la mayoría de los casos la necesidad de distribuir equipos concretos en puntos determinados de la red.
6. La tecnología, entendida como el hardware específico, seleccionado de acuerdo con los requerimientos del sistema global y de la arquitectura elegida. Debe asegurar la consistencia entre todas las capas intermedias.

La obtención de estos parámetros no va a depender solamente del grado de conocimiento de la red por parte del planificador, sino que es necesario apoyarse en fuentes externas que aporten dicha información, ya sean operadores, fabricantes, reguladores, etc. Algunos investigadores justifican la existencia de fuentes internas y externas al proceso de planificación [Arana, 1997]. Se puede entender como internas todas aquellas relacionadas con la obtención de las características del tráfico asociado, las cuales son inferidas mediante la correspondiente caracterización y modelado. De forma paralela, el análisis económico de las infraestructuras existentes determina los criterios de actuación a la hora de establecer diseños o estructuras específicas. Por su parte, determinados parámetros, como por ejemplo los asociados con la aparición de nuevos servicios, el cambio de comportamiento de los usuarios (reflejado en la demanda y los parámetros de calidad) o la evolución de la tecnología, solamente pueden ser obtenidos mediante la consulta o colaboración entre operadores, fabricantes y administraciones.

En cualquier caso, según [Arijs, 2000], la planificación de capacidades se resuelve mediante métodos de optimización, utilizando para ello técnicas propias de la Investigación Operativa, como por ejemplo la programación lineal, como en [Wu, 2003], la programación no lineal, en [Liu, 2003], o las técnicas de búsqueda heurísticas (de forma alternativa, la heurística constructiva con soluciones basadas en el sentido común pueden ofrecer resultados a tener en consideración). Dependiendo de la complejidad del problema y del grado de optimalidad deseado, la planificación de capacidades puede hacer uso de las diferentes tareas, ya sea secuencialmente o incluso de forma integrada.

Teniendo en cuenta que las soluciones secuenciales presentan más bajos requerimientos computacionales frente a resultados más robustos de las soluciones integradas, es precisamente en estos casos cuando la resolución puede hacer uso de la segmentación en sub-problemas, bien mediante un algoritmo integrado, o bien mediante un algoritmo que internamente realice la descomposición del problema en múltiples fases perfectamente realimentadas y resolubles mediante procesos iterativos.

En [Cho, 2001a], se desarrolla un modelo parametrizado aplicado a la planificación de redes de conmutación de circuitos. Como se puede observar en la Fig. 2.4, se compone de dos blucles, uno interno que optimiza los costes dentro de una topología determinada y, otro externo, que optimiza los costes a base de modificar la topología. Su adaptación a otros tipos de redes, incluidas las de conmutación de paquetes, es prácticamente inmediata, sin más que sustituir las estimaciones de número de circuitos por anchos de banda requeridos.

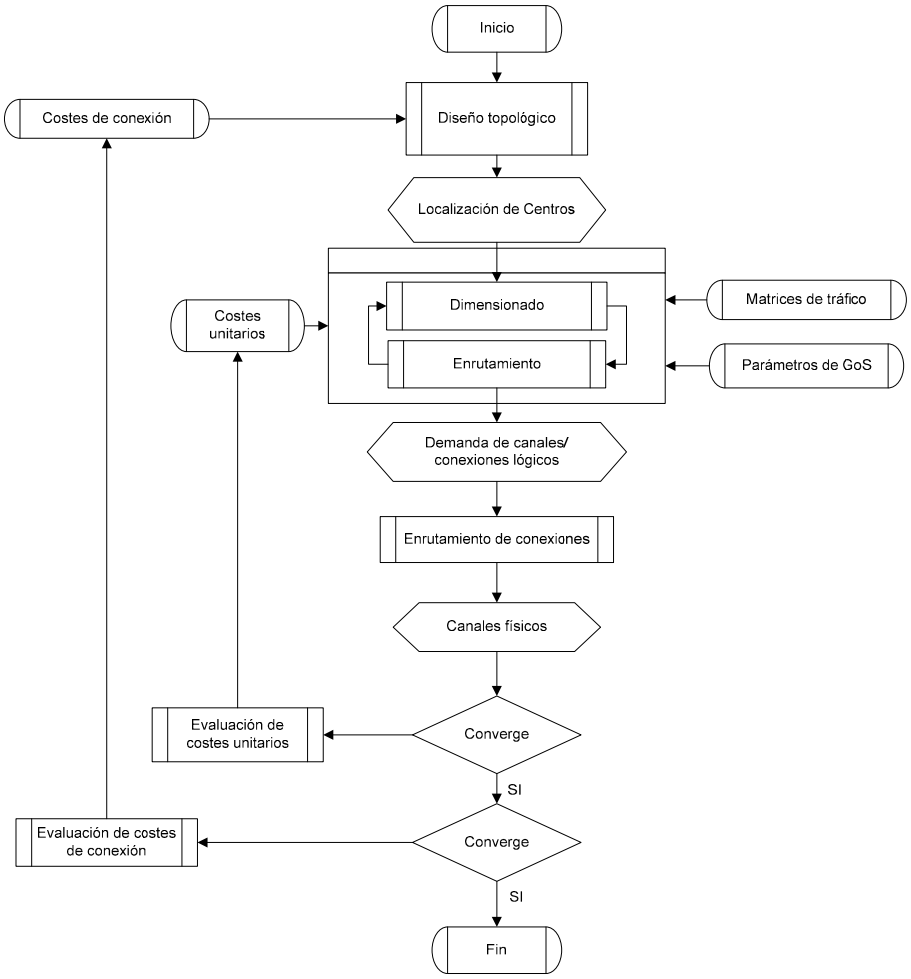


Fig. 2.4: Parametrización del proceso de planificación según [Cho, 2001b]

2.1.2. Herramientas para la planificación

Dentro del proceso de planificación se llevan a cabo tres acciones fundamentales: el análisis para la predicción del comportamiento del sistema, asumiendo situaciones ideales, como pueden ser distribuciones de tiempo entre llegadas discretas o tiempos de respuesta fijos; el diseño que provea un determinado nivel de servicio; y finalmente el modelado para su emulación hasta obtener resultados válidos para la elección de las arquitecturas y topologías correctas.

De todo el proceso parece determinante la elección correcta de la topología y arquitectura de la red. Esta es realizada como un proceso iterativo que se inicia en el mismo momento en que comienza la planificación de la red. De acuerdo con esto y siguiendo cualquiera de las metodologías de diseño, impuestas normalmente por las distintas topologías y técnicas de acceso (véase [Arijs, 2000]), se obtiene una configuración acorde a la tecnología elegida. A partir de este diseño inicial, se evalúan todas las soluciones posibles modificando cada uno de los parámetros correspondientes, e incluso cambiando de tecnología, hasta encontrar una solución aceptable, momento en el cual se procede a su correcta justificación y documentación (necesaria para el gestor de red).

De esa manera, según [Nurminen, 2003b], una herramienta de planificación consta de cuatro partes diferenciadas:

- Los algoritmos, que son la parte computacional, se dividen entre aquellos dedicados a tareas de diseño, y los encaminados al análisis. Los primeros son capaces de sintetizar nuevas soluciones u optimizar aquéllas ya existentes. Los segundos son capaces de estimar el resultado de un plan prefijado y definir el comportamiento en función de parámetros claves de su ejecución. Junto a los modelos y los algoritmos numéricos, puede aparecer el procesamiento simbólico mediante, por ejemplo, bases de conocimiento y motores de inferencia.
- Lal interfaz de usuario que interactúa con el mismo, bien de forma gráfica o textual, y define los ámbitos de trabajo de la máquina y el humano.
- La base de datos que lleva a cabo el almacenamiento de toda la información, de todas las decisiones de diseño, así como, la interfaz hacia sistemas externos que frecuentemente actúan como fuentes de información de entrada. Normalmente se distinguen dos apartados, las estructuras de datos de la memoria principal que implementa el modelo de datos en tiempo de ejecución, y los soportes de almacenamiento permanente, tales como el sistema de ficheros o los sistemas de gestión de bases.

- Las utilidades en forma de interfaces hacia el sistema operativo, como por ejemplo librerías gráficas o accesos a sistemas de ficheros, ofrecidos como servicios al resto de partes de la herramienta.

De cualquier modo, al partir del modelado del entorno de red existente, de las necesidades impuestas por el usuario y del incremento de tráfico asociado, un análisis de las distintas alternativas establece las bases de un correcto dimensionado. Las opciones tecnológicas válidas son analizadas económicamente hasta encontrar la solución óptima que será implementada. A partir de aquí, la planificación operativa toma el control, y con sus chequeos del comportamiento del sistema realimenta el proceso de mantenimiento de la red, dirigido mediante las correspondientes herramientas de diseño y análisis. Cabe destacar tres tipos de herramientas:

- Simulación física: requiere de la disponibilidad de todo el sistema final, sobre el que se realizan cargas de tráfico típicas. Dicha disponibilidad supone un gran inconveniente, traducido en elevados costes tanto económicos como temporales.
- Simulación computacional: requiere de algoritmos especializados para todas y cada una de las operaciones analizadas, que son ejecutados hasta alcanzar un equilibrio estadístico en sus resultados. Su principal inconveniente estriba en el elevado coste temporal que supone el desarrollo de los códigos especializados. Por el contrario, si se dispone de módulos de simulación estándar, su coste resulta muy económico aunque solamente en el caso de redes de dimensiones reducidas. Este supuesto es poco realista comparado con el tamaño de las redes reales.
- Evaluación analítica: hace uso de técnicas analíticas testadas para el modelado y predicción del comportamiento de la red, o de sus elementos individuales. Resulta el método más elegante, rápido y barato, aunque su precisión dependa casi exclusivamente del método analítico utilizado. La evaluación del sistema debe ir dirigida a cuatro puntos de vista distintos:

    1. Costes: es el más crítico ya que determina el coste directo de los servicios sobre el usuario. En una red típica, se considera que cerca del 60% de los gastos son generados por el transporte en sí de la información, mientras que el resto se los reparten, por igual tanto la parte física como los sistemas de gestión y control.
    2. *Throughput*, es decir, el número de transacciones soportadas por la red, ya sea en número de paquetes, *bitrate*, etc.
    3. Calidad de servicio (QoS), que expresa los valores típicos de rendimiento de una conexión física o virtual, como son el retardo, *jitter*, etc.
    4. Grado de servicio (GoS), que expresa la disponibilidad del servicio tanto en situaciones temporales normales, como de alto

tráfico (en la hora cargada), de buen funcionamiento, como bajo situaciones de fallo y averías.

2.1.2.1. **Diseño de una herramienta de planificación**

Además de los requisitos indicados en el apartado anterior, una herramienta de planificación debe ser diseñada siguiendo un esquema jerárquico o en capas, de las cuales la tarea fundamental será precisamente el modelado funcional. El hecho de poder obtener el número de nodos para cada nivel, teniendo en cuenta que los niveles vacíos pueden ser simplificados, permite realizar la planificación de forma casi manual y pasar directamente al análisis del comportamiento de la red. Esta primera capa es, por sí sola, el elemento único y suficiente para determinadas herramientas de planificación, como indica [Nurminen, 2003a].

La segunda capa está constituida por mecanismos específicos, algoritmos de análisis para una evaluación más intensa del funcionamiento de la red, y algoritmos de planificación para la automatización de determinados procesos de la planificación. Los de análisis permiten reducir el tiempo de desarrollo realimentando mediante información específica al usuario, por ejemplo, mediante la evaluación del coste de la red. Los de planificación permiten realizar de forma automática tareas repetitivas o la eliminación de errores, por ejemplo, durante el enrutamiento de tráfico. La principal función de esta segunda capa es la de reducir las necesidades de planificación manual, acelerando el proceso de planificación y reduciendo el número de errores.

La tercera capa la constituyen los modelos de optimización, que permiten obtener resultados óptimos o en su defecto cercanos a ellos. Son un suplemento al resto de algoritmos de la herramienta de planificación, pudiendo hacer uso incluso de ellos para solucionar problemas secundarios. Como estos modelos toman en consideración varios aspectos del problema simultáneamente, parten de la suposición de su idoneidad, y además pueden obtener resultados en tiempos razonables y mejores a los que pueda realizar un planificador experimentado.

En [Arana, 1997] se propone un modelo de capas alternativo, mostrado en la Fig. 2.5, según el cual el proceso de planificación comienza con la medida o estimación de las demandas de tráfico, mediante las cuales se establece la arquitectura de red, incluido el enrutamiento. Posteriormente, el dimensionado de dicha estructura y el análisis y optimización de cada uno de los procesos anteriores permite fijar los parámetros de calidad de la red.

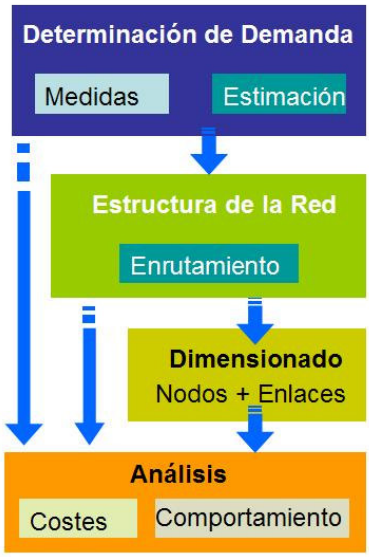


Fig. 2.5: Modelo de capas para la planificación

Sin embargo en [ITU-T, 2002] se propone un modelo más complejo, mucho más cercano a la definición comentada en el apartado 2.1.1, que tiene en cuenta que la planificación estratégica se realiza bajo un entorno liberalizado que significa tener en cuenta aspectos de competitividad con otros operadores. El primer paso sería el análisis tanto del contexto físico de la red, como de las necesidades de negocio, para lo que se establecen escenarios de trabajo adecuados. El conjunto de resultados es utilizado para la estimación de las demandas de tráfico y el diseño y configuración de los elementos de la red, así como para la definición, mediante el análisis económico, del plan de inversión y de negocio. Cada uno de los procesos anteriores utiliza herramientas y procedimientos específicos que a su vez pueden formar parte de una herramienta de planificación genérica. La estructura completa se muestra en la Fig. 2.6.

Desde el punto de vista del desarrollo de las herramientas, la estructura de capas debe facilitar la definición del proyecto sobre dicha organización, de forma que solamente las funciones básicas de las capas más importantes aparecen en los estados iniciales del proyecto, y así por ejemplo, la interfaz gráfica de usuario y el modelo funcional se sitúan como los únicos elementos obligatorios para cualquier sistema de planificación.

A partir de la estructura básica de la herramienta se pueden empezar a incluir algoritmos de análisis y planificación, preferiblemente simples si hay interés en maximizar la relación beneficio/coste. Su facilidad de implementación no supone un alto riesgo frente a su comportamiento y utilidad.

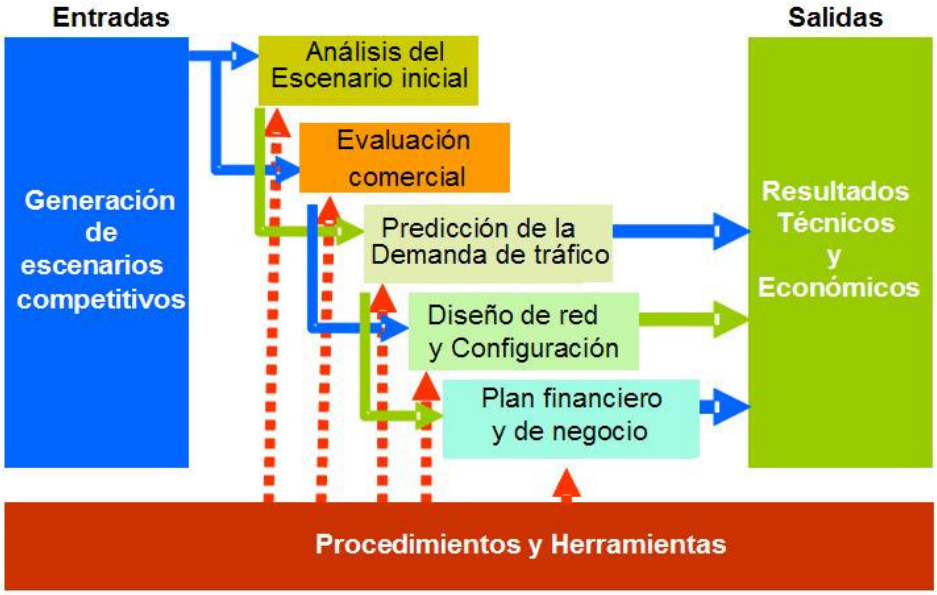


Fig. 2.6: Modelo de planificación según la ITU (Fuente: [ITU-T-2002]

2.1.2.2. Clasificación de las herramientas de planificación

La clasificación expuesta al principio del apartado 2.1.2 puede resultar demasiado genérica. De hecho, atendiendo al criterio de la ITU, las herramientas de planificación se diferencian en función del grado de detalle de la misma (y su correspondiente análisis) y del número de capas que conlleva el proceso. Tal como se observa en la Fig. 2.7, ambos criterios son contrapuestos pero permiten establecer cuatro tipos fundamentales:

Fig. 2.7: Tipos de herramientas de planificación según el criterio de la ITU [ITU-T-2002]

- De evaluación tecno-económica: No tienen un alto grado de detalle puesto que su objetivo es evaluar las características generales de toda la red. Así por ejemplo, los estudios de redes multicapa (especialmente las redes ópticas) o basadas en múltiples tecnologías de interconexión pueden ser consideradas dentro de esta categorías.
- De diseño y optimización: Con un grado de detalle medio, obtienen resultados detallados para entidades específicas, como por ejemplo en segmentos de red.
- De optimización y mejora de diseño: Dado que su grado de detalle es alto, se centran al nivel de entidades, de las cuales obtienen sus características específicas.
- De análisis y simulación: Entrando en el mayor detalle posible, este tipo de herramientas considera el comportamiento real de cada elemento de red.

Sin embargo, esta clasificación considera herramientas genéricas, y probablemente independientes entre sí, siempre en función del grado de detalle y/o su ámbito de aplicación. La combinación de ambos términos permite obtener una definición más extensa: el objetivo de la planificación es realizar un estudio pormenorizado del dimensionado de los mecanismos asociados a la red, desde el punto de vista temporal y funcional. De acuerdo con esta idea la planificación de redes podría ser considerada como un sistema tridimensional en el que sus ejes relacionan el desarrollo de la red temporal, geográfica y, por supuesto, funcional. Cada eje presenta valores característicos, tal y como muestra la Tabla 2.I.

Planificación Temporal	Planificación geográfica	Planificación funcional
Estratégica	SANP (Subscriber Access NP)	Lógica
Táctica	ANP (Aggregation Network Part)	Física
Operacional	CNP (Core Network Part)	Topológica

Tabla 2.I: Modelo tridimensional del proceso de planificación de redes

a) Temporal:

Como se comentó en el apartado 2.1.1, la planificación operacional estudia el diseño y desarrollo tanto de la red como de los servicios asociados. Sin embargo, dicha operación cubre períodos de tiempo relativamente cortos, lo que permite establecer criterios concretos que permiten mantener los requerimientos de GoS y QoS asociados a cada servicio. Para ser llevada a cabo, es necesario un conocimiento detallado de la red implementada, así como de los posibles mecanismos de gestión por lo que suele ser tarea, casi exclusiva, de los operadores sobre sus propias redes.

Por su parte, la planificación táctica considera redes ya implementadas para el análisis y estudio de la evolución a medio plazo del comportamiento de las mismas ante la introducción de nuevos servicios y/o el incremento del tráfico asociado a los ya existentes.

Finalmente la planificación estratégica es la tarea más abstracta puesto que realiza el estudio y caracterización de la red y todos sus servicios a medio y largo plazo, esto es, desde la implementación en sí, hasta la introducción de nuevas tecnologías y/o servicios. La mayoría de las veces el estudio comienza por la emulación completa de la red, tanto si existe previamente (*Scorched Node Approach*), como si es de nueva implementación (*Greenfield Approach*). A diferencia de las planificaciones táctica y operativa, la planificación estratégica no es exclusiva de los operadores de red, sino que dado su carácter estratégico, puede implicar a instituciones y organismos reguladores tanto a nivel nacional como internacional, vease [ITU-T, 2002], y el ejemplo de [GMBH, 2007]

b) Geográfica:

Atendiendo a la cobertura geográfica del proceso de planificación, tradicionalmente, las redes de telecomunicación se encuentran divididas según tres áreas diferenciadas:

- El acceso a la red, tradicionalmente denominada SANP (*Subscriber Access Network Part*): Normalmente la parte más costosa debido a que cubre una extensión geográfica limitada de forma densa, pero con concentraciones de tráfico bajas. También recibe el nombre de Red de Distribución.
- El acceso a la red dorsal, denominada ANP (*Aggregation Network Part*) o Red de Agregación. A diferencia de la anterior, su extensión es menos limitada y cubre normalmente entornos regionales. Su concentración de tráfico puede oscilar en función de dicha extensión, de tal forma que frecuentemente se considera su sub-división en LANP (*Low ANP*) y UANP (*Up ANP*).
- Por último, la red dorsal o CNP (*Core Network Part*). Jerárquicamente hablando, cubre la parte alta, a nivel nacional e internacional. La concentración del tráfico pasa a ser secundaria frente al problema que plantea la resolución del enrutamiento, y la correspondiente protección y el respaldo frente a errores.

c) Funcional:

Las funciones asociadas pueden ser consideradas de dos formas distintas: de acuerdo con el modelo de capas OSI, o de acuerdo a las características del tráfico, sea éste de datos de usuario, datos de control de la red o datos de gestión de la misma. Dicha subdivisión es representada como un modelo estratificado, en el que el tráfico de datos del terminal de usuario se

sitúa en las capas superiores, y los equipos de la red se sitúan en las capas más bajas. La Fig. 2.8 muestra el modelo correspondiente, deducido a partir del modelo de la ITU para las redes de banda ancha ATM [ITU-T, 1996].

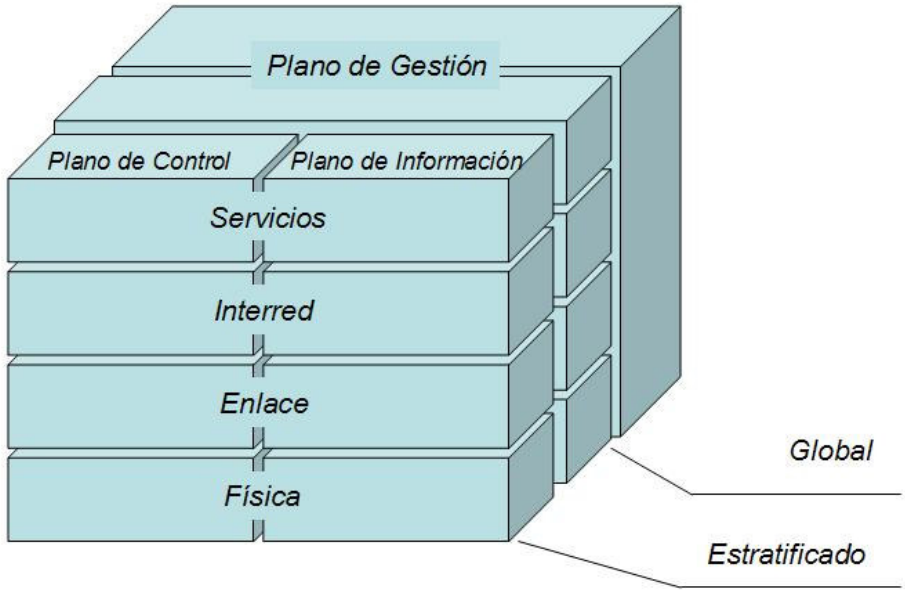


Fig. 2.8: Modelo funcional

Las capas dos y tres proveen la red de capa lógica, mientras que la capa uno la capa física. Actualmente, la planificación parte del tráfico de usuario situado en el plano de información, teniendo en cuenta la sobrecarga de control generada por los otros dos planos (control y gestión).

La subdivisión de tareas de planificación, tal y como se han venido explicando a lo largo de este capítulo, sobre los tres ejes temporal, geográfico y funcional, permite establecer al menos teóricamente, 27 combinaciones que gráficamente se asemeja al conocido cubo *Rubik* (vease[EURONGI, 2005]) y que la Fig. 2.9 muestra con detalle.

Sin embargo, en función de la arquitectura que se considere, no todas las combinaciones representadas por dicho cubo tienen aplicación práctica. Por ejemplo, la SANP de ADSL sobre el bucle de abonado de PSTN/ISDN, es una red pasiva donde solamente tienen sentido las subcapas inferiores de la red física, como se puede observar en la Tabla 2.II.

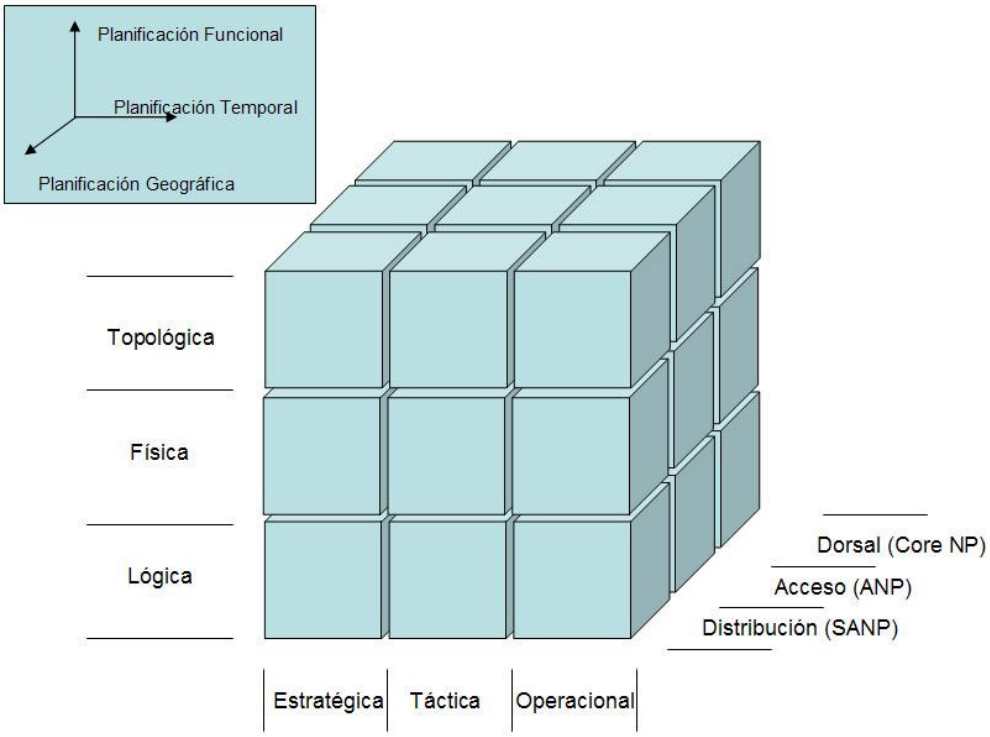


Fig. 2.9: El cubo de Rubik de la Planificación de Red

:

	SANP	ANP	Core NP
Lógica	---	X	X
Física	---	X	X
Topológica	(X)	X	X

Tabla 2.II: Relación entre el eje geográfico y el eje funcional para el caso de redes PSTN/ISDN y redes de Banda Ancha sobre DSL

2.2. Modelado de tráfico en redes Internet

Internet es el mayor compendio de sistemas específicos para comunicaciones. Su heterogeneidad es claramente apreciable con solo fijarnos en sus componentes principales, los enlaces, desde los simples modems de PSTN hasta los complejos enlaces dorsales, desde las tecnologías cableadas hasta las inalámbricas. Desde un punto de vista macroscópico Internet puede ser considerado como una red de Sistemas Autónomos que, incluso individualmente, mantienen la misma heterogeneidad que su conjunto. Las características de dichos sistemas, estando todos ellos interrelacionados, los hace independientes de forma que cualquier estudio puede ser particularizado a cada SA individual. Sin embargo, el comportamiento del tráfico es homogéneo,

independientemente del Sistema Autónomo, y solamente diferenciable de forma cuantitativa.

Múltiples estudios han concluido con la definición de la autosimilaridad del tráfico de Internet, véase [Leland, 1993a / Paxson, 1995 / Wilson, 1994]. Según esta idea, el tráfico correspondiente a un determinado enlace se comporta de idéntica forma con independencia de la escala temporal en la que es observada. De esta manera es posible determinar las características estadísticas asociadas con la dinámica temporal del tráfico, en lo que se ha denominado "autosimilitud de segundo orden asintótica", o de forma equivalente la dependencia de amplio rango (*Long Range Dependence* – LRD). De acuerdo con estas ideas, la autocorrelación del tráfico Internet decae de forma potencial en períodos de tiempo suficientemente amplios, lo que contrasta con la suposición tradicional, según la cual, decaería exponencialmente, comportándose incluso como un ruido blanco.

Internet está fundamentado en una estructura de capas interrelacionadas por los servicios y funciones que se ofrecen unas a otras, y que dialogan cuando los equipos y sistemas comparten dicha estructura. Si bien las capas básicas pueden ser de muy diferente naturaleza, IP es el nexo común para la interconexión de sistemas heterogéneos, estableciendo los mecanismos de direccionamiento tanto a nivel local (*Autonomous Systems* - AS) como global (Internet). Los datagramas IP son las unidades de datos básicas de Internet, y si bien dependen de las capas superiores (transporte y aplicación), establecen un patrón de tráfico a ráfagas, de anchura proporcional al tamaño medio de dichos bloques. La capa inmediatamente superior es la que caracteriza a su vez cada una de las ráfagas, ya sea en conexiones reales (TCP), como en sesiones abiertas (UDP). La generación de los datagramas IP va a ser modulada por las diferentes conexiones o secuencias gobernadas por TCP o UDP respectivamente. Precisamente, este comportamiento es similar y establece patrones de ráfaga como secuencias consecutivas de paquetes y silencios. Sin embargo, el término "tráfico Internet" suele ser sinónimo de tráfico TCP, y como tal, ha sido el referente tradicional en el modelado de tráfico Internet, no solo por sus características concretas, sino por su importancia cuando se refiere a aplicaciones Internet. Los servicios de correo (SMTP), *Web* (HTTP) y FTP son las aplicaciones de Internet por excelencia, todas ellas basadas en TCP. De todas las formas, en los últimos tiempos, la necesidad de incorporar aplicaciones de tiempo real ha obligado a adaptar viejas aplicaciones sobre modos de transporte más flexibles, como los que UDP ofrece con el modo datagrama. Así el *streaming* de audio/vídeo o la videoconferencia se han ido abriendo también camino en la *Red de Redes*.

Es precisamente la capa de aplicación la que acota el tráfico tal como se entiende, estableciendo sesiones para cada aplicación, las cuales pueden ser caracterizadas por su número (tiempo entre llegadas), duración y tamaño (bytes). A su vez, estas últimas características van a ser definidas por la capa de transporte correspondiente. Así por ejemplo, TCP establece conexiones

durante cada sesión (en algunos casos individuales, por ejemplo Telnet o e-mail, en otros casos múltiples conexiones, por ejemplo *Web* o FTP). De esta manera la sesión queda adicionalmente caracterizada mediante el número de conexiones (tiempo entre conexiones), su duración y tamaño. Cada conexión TCP es soportada por múltiples flujos IP, de forma que el tráfico es considerado como una agregación de paquetes IP generados extremo a extremo. Un flujo es una ráfaga en sí, siendo cada conexión el conjunto de flujos que la soportan.

Es así cómo la red Internet presenta una complejidad no siempre evidente, que hace de su estudio todo un reto, teniendo en cuenta que el tráfico generado por un usuario va a tener características muy diferentes a las de otros individuos por el simple hecho de hacer uso de servicios diferentes pero coexistentes. La planificación de la red Internet pasa entonces por la necesidad de conocer el comportamiento de los clientes frente a los diferentes servicios, estimar el tráfico de fuente con él asociado y analizar su comportamiento a lo largo de la red de acceso y la dorsal hasta que finalmente llegue a su destino. La estimación por usuario y servicio es resuelto mediante la elaboración de modelos de tráfico de fuente de diferente complejidad. Sin embargo, el problema más importante que plantea Internet es la necesidad de compartir todos los recursos al nivel de agregación y dorsal, lo que se traduce en la obligación de conocer, no solo los comportamientos de las fuentes individuales, sino también el comportamiento de su mezcla, es decir, modelar la agregación en puntos discretos de la red.

2.2.1. Modelos de Tráfico: Definición

Estadísticamente un modelo describe un proceso estocástico utilizado para la predicción o estimación del comportamiento de un flujo real. Idealmente, un modelo de tráfico representará todas las propiedades estadísticas relevantes correspondientes al flujo original, lo cual puede suponer la definición de modelos enormemente complejos, con múltiples variables dependientes e incluso diferentes niveles de abstracción que completen su comportamiento a diferentes escalas temporales.

De hecho, durante los últimos 30 años se han venido desarrollando diferentes técnicas y modelos cuyo principal objetivo ha sido siempre la obtención del proceso estocástico que idealice, de una forma simple pero exacta, el comportamiento del tráfico en las redes de comunicación. Tradicionalmente, los mayores esfuerzos en materia de modelado han estado centrados en el caso concreto de la conmutación de circuitos. Sin embargo la evolución de las redes, y la aparición de las técnicas de conmutación de paquetes han dado paso al desarrollo de nuevos modelos que en muchos casos han redefinido los tradicionales, y que han llegado ha plantear la necesidad de elaborar nuevas técnicas de cálculo e incluso de programación.

En cualquier caso, pese a cambiar las técnicas e incluso la algorítmica, el modelo sigue manteniendo las mismas características que antes: un conjunto limitado de parámetros que hagan coincidir los dos primeros momentos estadísticos, incluida la autocovarianza, del proceso estocástico y del tráfico real, de forma lo más manejable posible.

Según lo anterior, el éxito de un determinado modelo puede depender casi exclusivamente de la correcta selección de sus parámetros fundamentales, véase[Kode, 2001]. Además este modelado requiere en la mayoría de los casos, combinar las medidas reales y el análisis de la red, además del entendimiento de la aplicación que lo genera. De la misma manera, obtener el modelo correspondiente al tráfico generado por una determinada aplicación requiere conocer perfectamente el comportamiento de la aplicación en sí.

En [Nurminen, 2003a] se identifican dos tipos de modelos aplicables directamente: los de programación matemática y los funcionales.

2.2.1.1. **Modelo de Programación Matemática**

La técnica de programación matemática clásica formula el problema de planificación de la red como un modelo de optimización donde una función de coste dada es minimizada (o maximizada) en base a unos criterios o restricciones establecidos. Normalmente mediante un único modelo se pretende capturar todos los aspectos más relevantes del problema, y una vez obtenida una solución óptima, los valores de las variables clave permiten decidir la mejor acción a considerar.

Según esta idea, un único modelo, más o menos complejo, constituye el módulo principal de la herramienta. El resto de módulos son meras funcionalidades adicionales, como por ejemplo la entrada de datos, la parametrización, la generación del modelo o la representación de los resultados. De todo ello resulta una aproximación dirigida donde sus necesidades y capacidades definen la funcionalidad final de la herramienta. Ejemplos de estas soluciones aparecen continuamente en el diseño de redes, como por ejemplo en el despliegue de dorsales, la configuración de redes existentes o la planificación de capacidades de reserva, véase [Bley, 2002 / Heckmann, 2002 / Wu, 2004].

Idealmente, la programación matemática obtiene aproximaciones cercanas al modelo buscado, siempre que el problema sea formulado de forma rigurosa. Cuando el modelo contempla los aspectos más relevantes del enunciado, el resultado será preciso y óptimo. Normalmente, el coste de este tipo de modelo radica en la necesidad de establecer óptimos globales que no hagan incrementar el tiempo invertido en su obtención.

Las soluciones utilizadas en los problemas de planificación suelen ser típicamente no lineales, combinacionales y en muchos casos NP-completos.

En este caso, suelen ser tan complejos que la resolución de problemas de tamaño real suele ser demasiado costoso. De esta forma, la formulación del enunciado y sus soluciones no pueden obtenerse de forma separada, el modelo debe incluir el suficiente grado de abstracción sobre el problema en sí, y derivar una implementación de acuerdo a la solución heurística más correcta. Encontrar el balance adecuado entre precisión y manejabilidad es la clave del desarrollo del mismo[4]. Este siempre será una aproximación del problema real y no podrá representar todos y cada uno de los detalles relevantes del problema. Algunos aspectos serán ignorados para poder mantener así su simplicidad; otros, sin embargo, no será posible cuantificarlos, y cada caso concreto de planificación podrá contar con su propio conjunto de objetivos y restricciones. La imposibilidad de cuantificar determinados factores es un problema dada la necesidad de utilizar funciones de evaluación puramente numéricas, y además, se plantea el problema de decidir en función de los diferentes objetivos en un entorno de multiplicidad de criterios de decisión.

El uso de la heurística para la resolución rápida de problemas difícilmente tratables mediante otros métodos presenta un inconveniente, ya que el sistema puede volverse vulnerable ante necesidades cambiantes. Así por ejemplo, los avances tecnológicos pueden suponer la necesidad de realizar pequeñas modificaciones en la formulación del modelo, de forma que esos cambios pueden hacer que un algoritmo perfectamente adaptado, pase a ser completamente inútil. Además, los algoritmos heurísticos constan de parámetros que deben ser ajustados para obtener un correcto comportamiento por lo que, en el peor de los casos, debería realizarse dicho ajuste para cada nueva planificación.

Estos modelos tampoco son aconsejables si lo que se desea es una alta interacción con el usuario. Como requisito previo, el usuario debe conocer antes cómo trabaja el modelo y cómo deben utilizarse sus parámetros para poder así mantener el control sobre sus resultados. Esto puede resultar bastante complicado teniendo en cuenta que los usuarios, normalmente, van a ser expertos en determinadas áreas, no necesariamente matemáticas. Por otro lado la parametrización del modelo, mediante la asignación de pesos asociados a cada criterio de optimización, es realmente difícil y la viabilidad de la experimentación con diferentes valores es directamente proporcional a la velocidad de resolución del algoritmo.

[4] El ejemplo clásico consiste en la aplicación de un modelo de Erlang para modelar el tráfico de un número de fuentes evitando así el uso del modelo de Engset, cuya aplicación resulta costosa.

Los algoritmos estocásticos, como por ejemplo el *simulated annealing*[5], o los algoritmos genéticos son una alternativa para la resolución de determinados tipos de problemas de optimización de forma escalonada y sobre espacios de búsqueda particulares y/o discontinuos, obteniendo mejores soluciones globales que los algoritmos clásicos de programación matemática. Así por ejemplo, *simulated annealing* a partir de un mínimo local realiza un proceso de búsqueda en un espacio de soluciones con alta probabilidad de ser óptimos globales. Sin embargo el proceso resulta excesivamente lento y es sustituido por otros mecanismos de búsqueda de soluciones locales.

Una característica a destacar de los algoritmos estocásticos, es que pueden ser utilizados incluso sin un modelo algebraico previo del comportamiento del sistema. Éste es tratado como una caja negra, siempre y cuando exista una función de evaluación que permita comparar todas las soluciones candidatas. Así, su implementación resulta simple y únicamente la función de evaluación es la que debe ser adaptable a los cambios.

2.2.1.2. **Modelado funcional**

Estos modelos definen el comportamiento del sistema. Contrariamente a los de programación matemática, un modelo funcional (como por ejemplo una simulación) no tiene un objetivo explícito y solamente representa el comportamiento del sistema ante diferentes circunstancias. De esta forma, el planificador puede probar diferentes alternativas y analizar los resultados de los diferentes escenarios. Es precisamente el planificador el que asume el papel principal, ya que el modelo asume las decisiones de planificación y simula sus efectos. Puede ser complementado mediante funcionalidades añadidas, como por ejemplo algoritmos de análisis de diferente complejidad a partir de los cuales poder realizar un estudio pormenorizado de los resultados. Mediante la prueba de diferentes alternativas y el uso de sus resultados, puede tomar conciencia del problema y tomar las mejores decisiones, e incluso, obtener soluciones que no habían sido previstas cuando las herramientas fueron desarrolladas.

---

[5] Simulated annealing (SA) es una meta-heurística para problemas de optimización global, es decir, encontrar una buena aproximación al óptimo global de una función en un espacio de búsqueda grande. El nombre e inspiración viene del proceso de templado (annealing en inglés) en metalurgia, una técnica que incluye calentar y luego enfriar controladamente un material para aumentar el tamaño de sus cristales y reducir sus defectos. El calor causa que los átomos se salgan de sus posiciones iniciales (un mínimo local de energía) y se muevan aleatoriamente; el enfriamiento lento les da mayores probabilidades de encontrar configuraciones con menor energía que la inicial.

El modelo solamente simula el comportamiento de la red, aunque el planificador se responsabiliza de la toma de decisiones de la planificación. Además, solamente obtiene las consecuencias de dichas decisiones, no provee soluciones por sí mismo.

En cuanto a la implementación, un modelo funcional no necesita especificar funciones objetivo, por lo que evita tener múltiples criterios o parámetros no cuantificables. Además, las técnicas de programación orientada a objetos, tan importantes en la ingeniería de software, están especialmente indicadas para su implementación. Son fáciles de comprender y existe un salto conceptual muy pequeño entre el mundo real y el modelo, reducido aún más mediante el uso de interfaces gráficas de usuario adecuadas.

2.2.1.3. Algoritmos de propósito especial

Este tipo de algoritmos resuelven los problemas de forma particular. Algunos de estos algoritmos, como por ejemplo el de *Kruskal* del *minimum spanning tree*, permiten obtener soluciones óptimas. Sin embargo otros algoritmos de este tipo, como por ejemplo el TSP (*Travelling Salesman Problem*) son soluciones heurísticas, que han demostrado ser cercanas a la óptima en un elevado número de ejemplos reales, véace[Adler, 2001].

Suelen ser utilizados junto con los modelos funcionales, los cuales representan los datos que serán manipulados por el algoritmo. Como normalmente cada uno resuelve un determinado problema, puede ser necesario el uso de varios que cooperen entre sí. De esta manera la función de los algoritmos es la de proveer herramientas que pueden ser aplicadas bajo el control del planificador o por algún mecanismo de nivel superior. Si se permite su control, también se puede seleccionar cuál ejecutar, sobre qué parte del problema utilizarlo y realizar modificaciones en cualquier punto intermedio. El uso de soluciones más rápidas permite a éste estudiar diferentes alternativas, entender el comportamiento de la red y elaborar los mejores planes.

Dejar al planificador todo el peso de la herramienta implica la necesidad de que conozca cómo realizar la planificación de la red. Los algoritmos de planificación le permiten abordar problemas computacionalmente complejos, aunque en la mayoría de los casos, no proveen la asistencia suficiente como para afrontar las tareas de planificación más complejas. Los planificadores poco experimentados prefieren que determinadas decisiones del proceso de planificación sean tomadas directamente desde los algoritmos.

Otra deficiencia es que el uso de algoritmos de óptimos locales junto con criterios que no permiten obtener un óptimo global en tiempo razonable, podría ser simplificado en un único paso con resultados mejores. Sin embargo, el coste computacional es el que determina su viabilidad.

Los algoritmos de propósito general pueden ser también utilizados como bloques constitutivos de otros más complejos. Por ejemplo, los algoritmos de camino corto, que enrutan el tráfico entre una pareja de nodos, pueden ser utilizados cambiando el valor correspondiente según los resultados de cada iteración. Un problema heurístico simple de ordenación de los pares de nodos, en base al tráfico existente entre ellos, se transforma en un algoritmo para el enrutamiento de tráfico en redes con limitaciones de capacidad. Es por ello, que estos algoritmos deben ser computacionalmente rápidos.

La mayoría de estos algoritmos son fáciles de implementar, y los problemas más comunes aparecen referenciados y solucionados en libros de texto y librerías software. Como los algoritmos suelen estar definidos de forma genérica, no suelen verse influenciados por los cambios tecnológicos en las redes, y así estos mismos pueden ser utilizados en diferentes escenarios mediante mínimas modificaciones.

2.2.2. Modelado y estimación de tráfico

QoS y GoS son parámetros íntimamente ligados con las medidas del tráfico entre cada par de nodos de la red. El GoS describe la accesibilidad a un servicio, típicamente en la hora cargada (*Bussiness Hour* – BH), mientras que el QoS los parámetros intrínsecos al uso del servicio, como por ejemplo el MOS (*Mean Opinion Rate*) en servicios de telefonía o el PLR (*Packet Loss Rate*), PER (*Packet Error Rate*) o el retardo en redes de paquetes. De acuerdo con las distintas técnicas de análisis y diseño, el modelado y estimación de las cargas dentro de la red resultan determinantes a la hora de planificarla y dimensionarla. Sin embargo, esta tarea puede llevarse a cabo de tres formas distintas, dependiendo de las características del tráfico que se desee estimar:

- Estimación de cargas de tráfico en nuevos sistemas: Su objetivo es la estimación de las medidas de intensidad de tráfico para cada línea de usuario, de forma que se considera que cada aplicación y cada usuario determina una intensidad de tráfico concreta. Para llevar a cabo dicha tarea, es necesario definir las intensidades de tráfico previstas en la hora cargada, medidas éstas en *milierlangs* o bits por segundo para cada localización geográfica de los usuarios. A continuación, se definen los terminales o equipos utilizados en dichas localizaciones, se aplican los correspondientes modelos, se calculan las estadísticas en la hora cargada para cada nodo y usuario, y por último, se calculan las intensidades de tráfico en cada nodo y línea de acceso en concordancia con la topología de red seleccionada.
- Estimación de cargas de tráfico en sistemas existentes: El objetivo es obtener las medidas de intensidad para cada línea de usuario, las cuales van a poder ser obtenidas a partir de la propia observación de la red. Para ello se obtienen las intensidades de tráfico asociadas a la hora cargada en base a los datos reales correspondientes a cada servicio. A continuación, se calculan las cargas de tráfico en la hora cargada

mediante el uso de los modelos que se corresponden a los datos obtenidos estadísticamente. Finalmente se calculan las intensidades para cada localización y línea de acceso para cada caso concreto en la evolución de la red, pudiendo incluso realizar estimaciones dentro de las nuevas situaciones internas y/o externas a la misma.

- Estimación de tráfico dorsal: Su objetivo es la obtención de las medias de intensidad de tráfico para la red dorsal, ya sea a partir de los flujos entre cada pareja de nodos, o bien asumiendo que el caudal por cada fuente individual está distribuido hacia todos los destinos en función de sus cargas y de su distancia. Para ello es necesario calcular el tráfico total en tránsito para cada nodo conmutador de la red, a continuación se calcula el total, tanto entre cada pareja de nodos como para cada nodo en particular. Por último, se realiza el cálculo del tráfico en cada línea troncal que une cada pareja de nodos.

Sin embargo, todas estas estimaciones están basadas en el conocimiento del supuesto comportamiento del tráfico que genera cada elemento de la red. Este conocimiento ha tenido que adaptarse a la velocidad del desarrollo de las redes que modelaban, véase[Sarvotham, 2001]. Los modelos de cálculo clásicos basados en los de tráfico simples, la mayoría de las veces bajo el influjo Markoviano, han dado paso a correcciones y nuevas hipótesis que permitan modelar el comportamiento del tráfico de una forma más fidedigna, sin aumentar el coste computacional relacionado. Esto ha dado lugar a la aparición de nuevos modelos de tráfico que permiten simular/emular este comportamiento desde el mismo punto en el que ha sido generado. Son precisamente estos modelos los que forman parte del mecanismo fundamental a utilizar en el proceso de planificación.

2.2.3. **Modelos de tráfico de fuente**

El tráfico observado en un determinado enlace de una red puede ser tenido en cuenta como una secuencia de peticiones de servicio o de observaciones, en general, son interpretadas de muy diferente forma en función del grado de detalle y el tipo de características estadísticas consideradas. Así por ejemplo, desde el punto de vista del usuario, el tráfico es tratado como una secuencia de acciones, o en el caso más general, de sesiones. Desde el punto de vista de la red, o incluso de las aplicaciones, el tráfico es considerado como una secuencia de paquetes de tamaño variable. Independientemente de su significado, la secuencia de observaciones es modelada mediante un conjunto de variables aleatorias con funciones de distribución de probabilidad definida y en la misma escala temporal de la secuencia, véase[Marot, 1999]. Precisamente, las características temporales permiten distinguir entre procesos discretos, si las observaciones se obtienen en intervalos finitos de tiempo, y continuos en el

tiempo, si son obtenidos en cualquier momento dentro de intervalos de tiempo finitos o infinitos.

$$S = ..., X(t_{n-1}), X(t_n), X(t_{n+1}), ... = X(t_i) \begin{cases} \{X_n\}_{n=0}^{\infty} & t_{discreto} \\ \{X(t)\}_{t=0}^{\infty} & t_{continuo} \end{cases} \quad (2.1)$$

De esta manera, los procesos que describen las llegadas de entidades individuales (paquetes, peticiones, etc.) en una secuencia discreta de tiempos de llegada con origen en cero, reciben el nombre de proceso puntual. Cuando dicha secuencia se desarrolla de forma continua en el tiempo con intervalos independientes e idénticamente distribuidos, el proceso describe el número de peticiones/llegadas hasta el momento, recibiendo el nombre de proceso contador. En la Fig. 2.10 se puede observar la relación que existe entre ambos conceptos:

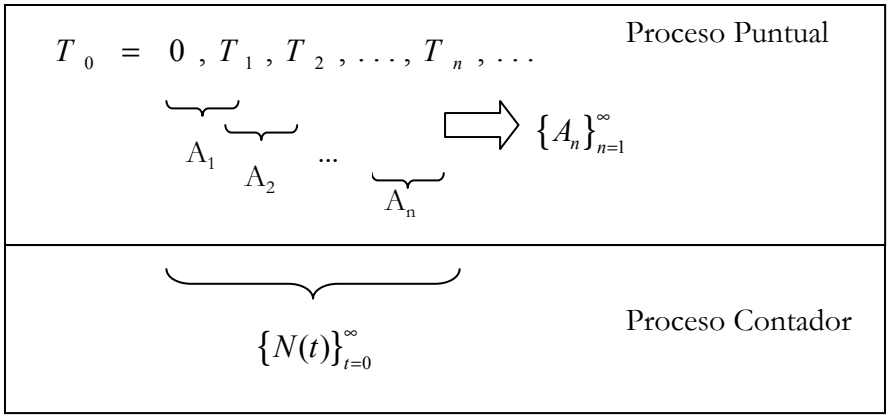


Fig. 2.10: Relación entre procesos puntuales y procesos contadores

Los procesos puntuales pueden también ser definidos mediante el proceso que modela el tiempo entre llegadas $\{A_n\}_{n=1}^{\infty}$, donde $A_n = T_n - T_{n-1}$ es el intervalo que separa la llegada n-ésima de la anterior[6].

Ésta, como otras, es una de las bases fundamentales sobre la que se sostienen la mayor parte de los modelos de tráfico de fuente desarrollados hasta el

[6] La notación utilizada distingue entre el caso continuo y el caso discreto:
N_n cadena discreta (en tiempo)
N(t) cadena continua (en tiempo)
X_n proceso discreto (en tiempo)
X(t) proceso continuo (en tiempo)

momento. Todos ellos parten de la teoría estocástica, no siendo más que aplicaciones directas de la misma, definiendo casos particulares que permiten establecer modelos individuales y a su vez, permiten ser combinados obteniendo nuevos modelos. De hecho, y pese al elevado número de modelos y combinaciones existentes actualmente, en [Jagerman, 1998] se atreven a hacer una amplia clasificación de mecanismos concretos para el modelado de fuentes de tráfico. La Tabla 2.III enumera las más importantes:

1. Renewal traffic models
Poisson Processes
Bernoulli
Phase-Type
2. Markov based traffic models
Markov Modulated Processes
Markov Modulated Poisson Processes
Transition Modulated Processes
3. Fluid traffic models
4. Autoregressive –type traffic models
Linear Autoregressive (AR) Processes
Moving Average (MA) Processes
Autoregressive Moving Average (ARMA) Processes
Autoregressive Integrated Moving Average (ARIMA) Processes
5. TES Traffic Models
5.1. TES Processes
5.2. Empirical TES Modelling

Tabla 2.III: Modelos para la caracterización de tráfico de fuente según [Jagerman, 1998]

Sin embargo, [Kode, 2001]hace una clasificación cualitativa de dichos modelos, acorde con las técnicas de modelado vistas en el apartado anterior. De esta forma el modelado de tráfico de fuente puede ser solucionado mediante los siguientes métodos:

- Modelos deterministas: Para el modelado del tráfico de fuente utilizan el análisis de su comportamiento en el peor caso, con lo que los retardos en sistemas de espera son fácilmente calculables. Estos modelos suponen la sobreestimación de los recursos necesarios para cursar el tráfico y, por lo tanto, su infrautilización. Es recomendable conocer a priori los patrones de tráfico, facilitando el proceso de selección de los valores para cada uno de los parámetros que componen el modelo. Entre estos métodos destacan el modelo de *leaky-bucket* y el modelo de la envolvente empírica: Sea A el tráfico generado por una determinada fuente en función del tiempo t, y $A[\tau, \tau+t]$ dicha función en un intervalo dado. Dicha función está limitada por otra función $A^*[t]$ tal que cualquier valor de $A[\tau, \tau+t]$ se encuentra siempre por debajo de los valores de $A^*[t]$, cualquiera que sea t:

$$A[\tau, \tau + t] < A^*[t] \quad \forall \tau, t \geq 0 \tag{2.2}$$

  Existen infinitas funciones límite, como por ejemplo considerar los valores máximos de la tasa de tráfico, pero interesa encontrar funciones límite ajustadas que minimicen la sobreestimación de ancho de banda. Matemáticamente, la función límite ajustada a una determinada fuente de tráfico se representa por:

$$E^*(t) = \sup_{\tau>0} A[\tau, \tau + t] \quad \forall \tau, t > 0 \tag{2.3}$$

donde "sup" representa el límite superior ajustado, o el mínimo límite superior de un determinado conjunto. Este método suele utilizarse para el modelado del comportamiento de múltiples elementos en cascada, como por ejemplo la concatenación de *leaky-buckets*.

- Modelos estadísticos: En este caso se hace uso de la aproximación para llegar a la caracterización de una determina fuente de tráfico. Cada fuente es representada mediante un modelo analítico y, para su inclusión en un agregado de tráficos, es necesario analizar el comportamiento del QoS del agregado antes de tener en cuenta la nueva fuente. Se plantean varias dificultades: es difícil determinar un modelo de tráfico a priori que se ajuste en todo momento al comportamiento de una fuente, salvo que se realicen constantes correcciones sobre los valores de los correspondientes parámetros. Ello obliga a realimentar los modelos con medidas que permitan actualizarlos. La resolución en tiempo real se hace, de esta manera, muy difícil. Además, sería recomendable realizar los análisis en base al comportamiento extremo a extremo, lo cual también resulta complicado.

- Modelos analíticos: Describen matemáticamente la función de distribución de la variable aleatoria modelada. Estos modelos resultan muy simples, aunque las características del tráfico modelado son de menor calidad frente a los modelos empíricos. En general, los modelos analíticos predicen una forma aproximada de la función de distribución modelada, y solamente si los parámetros que componen el modelo son conocidos. La ventaja radica en que dichos valores pueden ser obtenidos de forma empírica.

- Modelos de Poisson: Tradicionalmente el más utilizado para el modelado de tráfico. Al ser la interllegada una distribución exponencial, los procesos de llegada de Poisson generan tráficos suaves conforme aumenta la escala de observación. Además, aún estando en niveles bajos, el uso de este modelo para tráfico a ráfagas está limitado conforme se agregan fuentes de estas características. La suma de fuentes tiende a suavizar el tráfico agregado.

- Modelado de tráfico autosimilar: Los modelos tradicionales requieren gran número de parámetros para conseguir capturar el comportamiento del tráfico a ráfagas, tal y como se entiende en los servicios Internet. Existen algunas soluciones simples que evitan este problema:
  - *Fractional Brownian Motion* FBM: Es el método utilizado para describir la agregación de varios flujos de tráfico. Básicamente obtiene la cantidad de tráfico que se espera obtener en un tiempo dado. Solamente presenta tres parámetros básicos:

- Tasa media *m*: cantidad de tráfico o grado de utilización de los recursos.
- Relación varianza-media *a*: mide el grado de fluctuación respecto de la media.
- *Hurst*: tasa de decremento de la correlación del tráfico. También indica la dependencia a largo plazo (LRD *Long Range Dependence*).

Los dos últimos parámetros caracterizan el comportamiento a ráfagas del tráfico. Este método, puede ser utilizado para definir lo que se denomina el *Fractional Leaky Bucket* (FLB), caso especial de FBM. El *network calculus* también puede ser aplicado en la concatenación de FLB para el dimensionado de recursos, el cálculo del retardo extremo a extremo y el multiplexado de fuentes de tráfico.

- Mapas caóticos: Las carácterísticas de autosimilitud del tráfico Internet están relacionadas directamente con la teoría del caos. Este tipo de modelos hacen uso de funciones iterativas, mediante las cuales mapean el comportamiento caótico de sistemas dinámicos, entre los que se encuentra Internet.

- Superposición de fuentes ON-OFF: Consiste en la generación de tráfico mediante multiplexación de varias fuentes ON-OFF. Cada fuente de dos estados genera paquetes, de forma que al utilizar distribuciones de "cola pesada" se consigue un tráfico con características autosimilares, normalmente mediante distribuciones de Pareto.

A continuación se desarrollan los modelos más interesantes desde el punto de vista del modelado de tráfico Internet.

2.2.3.1. **Procesos de Renovación**[7]

En los procesos de renovación $X(t_i)$ la función de distribución de probabilidad de interllegadas entre dos nacimientos es independiente e idénticamente distribuida (*i.i.d.*), cualquiera que sea su función de distribución, por lo que su autocorrelación es cero. Dentro de este tipo de procesos hay que destacar dos casos particulares:

[7] Renewal Processes

- Procesos de Poisson: Las llegadas de las observaciones son proporcionales al intervalo de tiempo entre cada una, de tal forma que el tiempo entre llegadas n-ésimo queda descrito mediante una función de distribución exponencial de media λ, siendo éste su parámetro fundamental.
- Procesos de Bernoulli: Es la particularización de los procesos de Poisson en el dominio discreto de tiempo, con una separación entre observaciones k. Con esta condición, la probabilidad de que haya llegadas en cada k, es independiente, y sigue una función de distribución binomial, mientras que el tiempo entre ellas sigue una geométrica, siendo el parámetro fundamental la probabilidad de llegadas p.

La utilidad de estos modelos es muy limitada, dada la simplicidad de sus principios, por lo que solamente son utilizados para representar el proceso de llegadas de paquetes, siempre y cuando el tráfico observado tenga autocorrelación nula, o modelar sesiones o secuencias de comandos al nivel de aplicación independientes entre sí.

2.2.3.2. **Modelos de Markov**

Sin embargo, $X(t_i)$ no siempre es independiente, por lo que es necesario modelar precisamente dicha dependencia, utilizando los denominados procesos de Markov, según los cuales cada observación depende solamente de la observación anterior, véase [Akimaru, 1999]. De esta manera, se definen cadenas de Markov como una secuencia de variables aleatorias $X(t_n)$, de forma que la probabilidad de que el próximo valor observado (estado) sea $X(t_{n+1})=j$ sólo depende del estado actual $X(t_n)=i$, siendo su valor p_{ij}. En este caso el tiempo de permanencia en un determinado estado es independiente del resto de estados, y en función de si la cadena es discreta o continua, podrá ser caracterizado mediante una función de distribución geométrica o exponencial negativa respectivamente.

Una variante de los modelos de Markov son los procesos de renovación de Markov, según los cuales tanto las observaciones, como los tiempos, dependen exclusivamente de la última observación. Sin embargo, en este caso, y al contrario que en las cadenas de Markov, el tiempo entre dos llegadas puede estar distribuido de forma arbitraria.

Aunque más elaborados que los procesos de renovación, su utilidad también está limitada al modelado de tráficos con autocorrelación baja o incluso nula, o de comportamientos de usuario con cierta interacción.

2.2.3.3. **Modelos modulados de Markov**

El comportamiento de la cadena de Markov, que modela el tráfico, puede ser controlado a su vez por otro proceso de Markov, de forma que cada estado está dirigido a su vez por los estados del proceso auxiliar. El caso más sencillo consiste en "modular" un proceso de Poisson mediante una cadena de Markov, de forma que el estado actual de la cadena define el parámetro λ_k del proceso de Poisson, y conforme se lleven a cabo las transiciones de estado de la cadena, la caracterización del proceso cambiará. La obtención de todos los posibles estados del sistema, así como las probabilidades de transición entre ellos, permite definir los modelos denominados Procesos de Poisson Modulados por Markov (MMPP – *Markov Modulated Poisson Processes*).

Este tipo de modelos permite recrear el comportamiento de los usuarios, y por tanto de su tráfico generado, o en general de tráficos con diferentes comportamientos. Sus características permiten tener en cuenta cierto grado de autocorrelación, más cercano con determinados tipos de tráfico.

Entre estos modelos también se encuentran variantes discretas. En [Lombardo, 2004] se expone el denominado Modelo SBBP (*Switched Batch Bernoulli Process*), el cual modela un proceso estocástico doble, en el que una cadena de Markov, de N estados, modula a otro proceso de lotes caracterizado por una función densidad de probabilidad dependiente de cada estado de los *N* en los que se encuentre. De esta manera, el proceso SBBP queda definido mediante la matriz de transición de la cadena de *N* estados, y una matriz con las distribuciones correspondientes al proceso de llegadas asociada a cada estado de los *N*. Para el modelado de la agregación de varios flujos, es posible considerar un sistema *SBBP/Sc/1/K*, donde el proceso SBBP es ajustado a partir de la función densidad de probabilidad y autocorrelación deseadas, siendo *Sc* la curva de servicio ofrecida, y *K* la longitud del *buffer*, tal como se muestra en la Fig. 2.11.

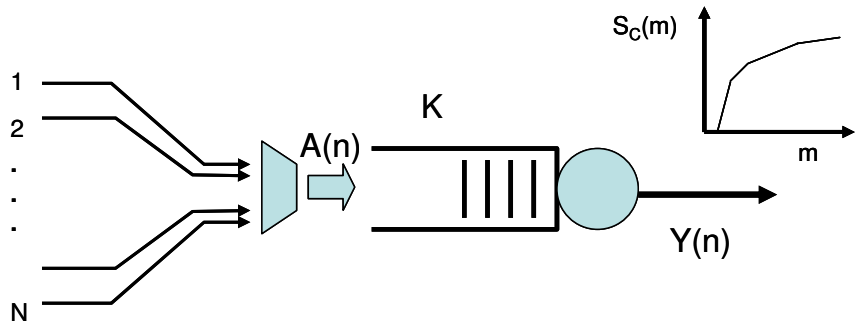


Fig. 2.11: Simplificación de múltiples fuentes mediante el modelo SBPP

Además estos modelos, pese a ser simples, no han perdido su validez, incluso cuando han sido cuestionados tras demostrarse la autosimilitud del tráfico Internet. Así en [Robert, 1997] se demuestra cómo el uso de cadenas moduladas de Markov también permite el modelado de tráfico con características muy cercanas a las del tráfico autosimilar, en este caso sobre líneas Ethernet.

2.2.3.4. **Modelos de fluidos**

Estos modelos consideran procesos en los que el número de observaciones es tan grande que la información que aporta cada una de ellas es insignificante, pudiendo ser tratadas todas ellas como un flujo constante en el que solamente los cambios determinan la variación de los parámetros de dicho flujo. Es precisamente la señalización de dichos cambios la que lo modela. Es por ello que estos modelos son utilizados bien ante aplicaciones y servicios de tasa constante, donde dichos parámetros son también prácticamente constantes como es el caso de ATM, o en las transferencias de datos que llevan a cabo algunas aplicaciones TCP/IP.

2.2.3.5. **Modelos estocásticos lineales**

También denominados auto-regresivos, definen cada variable aleatoria X_n como una función explícita de las variables aleatorias anteriores de acuerdo a una ventana temporal que incluye un número determinado de dichas variables. La distribución de X_n se denomina distribución marginal, y la función de autocorrelación establece, para cada intervalo *k,* el correspondiente coeficiente de correlación entre X_n y X_{n-k}. La ventaja de estos modelos es que pueden ser aplicados en tráficos con dependencia a corto plazo, lo cual no se ajusta a la realidad del tráfico Internet.

En el caso de considerar procesos estocásticos discretos, si las variables aleatorias pueden ser definidas mediante una distribución normal multivariante, este proceso se denomina de Gauss, siendo su distribución marginal otra distribución normal.

En cualquier caso, el conjunto de variables aleatorias asociadas a la forma general de un modelo estocástico lineal es de la forma:

$$X_n = \alpha_0 + \sum_{r=1}^{\infty} (\alpha_r X_{n-r} - \beta_r \varepsilon_{n-r}) + \varepsilon_n \qquad (2.4)$$

Siendo *n* mayor que cero, α_r y β_r constantes reales, y ε_n variables aleatorias i.i.d. de media cero e independientes de X_n. La variación de los diferentes componentes permite clasificar los modelos lineales en:

- Modelos auto-regresivos (*AR(p)*) de orden *p*: La expresión anterior se reduce a:

$$X_n = \sum_{r=1}^{p} (\alpha_r X_{n-r}) + \varepsilon_n \quad (2.5)$$

Su simplicidad permite calcular los parámetros de dicho modelo fácilmente, pero sin embargo su autocorrelación decae exponencialmente, por lo que no son la elección adecuada para el modelado de señales con autocorrelaciones con caídas menores. Algunas variaciones proponen el uso de modelos de Markov que modulen procesos AR, o incluso la combinación de varios modelos AR.

- Modelos de medias móviles *MA(q)* de orden *q*: En este caso la expresión reducida resulta:

$$X_n = \sum_{r=1}^{q} (\beta_r \varepsilon_{n-r}) + \varepsilon_n \quad (2.6)$$

Su utilidad en el modelado de señales es tan limitada, que normalmente son desechados a favor de los autorregresivos de media móvil (*ARMA(p,q)*), con AR de orden *p* y MA de orden *q*:

$$X_n = \sum_{r=1}^{\infty} (\alpha_r X_{n-r} - \beta_r \varepsilon_{n-r}) + \varepsilon_n \quad (2.7)$$

Aunque algo mejores que los simples *AR(p)*, los *ARMA(p,q)* presentan una autocorrelación con caída exponencial que tampoco los hace ser demasiado útiles salvo en casos muy concretos. Hay que sumar la complejidad adicional que supone el cálculo de los parámetros del modelo, solamente resoluble mediante ecuaciones diferenciales no lineales, y que toma su máxima expresión en su extensión, denominada procesos autorregresivos de media móvil integrada (*ARIMA(p,d,q)*). En éstos, el polinomio de coeficientes autorregresivos presenta al menos *d* raíces unitarias, y son utilizados en el modelado de series temporales homogéneas y no estacionarias. El caso ARIMA para *d=0* se corresponde al caso ARMA. De hecho, algunas simplificaciones, con un conjunto de parámetros limitado, son utilizadas para comparar y evaluar el comportamiento de los modelos con dependencia a largo plazo, como es el caso de los modelos autosimilares.

Un ejemplo es el caso de los modelos autorregresivos discretos (*DAR (p)*), donde *p* indica el número de grados de libertad, o de autocorrelaciones comparadas. Así por ejemplo el *DAR(1)* presenta la siguiente expresión:

$$X_n = \alpha_n X_{n-1} + (1-\alpha_n)\varepsilon_n \quad (2.8)$$

Y es utilizado en el modelado de tráfico VBR [Heyman, 1996].

Sin embargo, todos estos modelos asumen funciones de distribución con varianza finita, cuando muchos estudios han demostrado que el tráfico Internet no se comporta dentro de esos límites, sino que, en general, su varianza tiende a infinito, véase [Crovella, 1998]. Dicho comportamiento puede ser modelado partiendo de trazas de tráfico reales. Esto es lo que hacen los modelos TES (*Transform-Expand-Sample*), los cuales obtienen una función de distribución aproximada mediante la transformación de una secuencia uniformemente distribuida (mediante el uso de histogramas). Su capacidad para capturar las principales características de la función de autocorrelación de la secuencia original, les ha hecho interesantes para su uso en modelos específicos de tráfico de VBR, Ethernet y Web, véase [Jelenkovic, 1996] y [Chen, 1995] .

Cuando el análisis de trazas reales no es posible, o la obtención del correspondiente modelo TES se complica, puede ser razonable el uso de funciones de distribución específicas que presenten varianza infinita o al menos muy grande. Estas funciones de distribución han recibido el nombre de distribuciones de cola-pesada (*heavy-tailed*), en las cuales la correspondiente variable aleatoria cumple lo siguiente:

$$\Pr[X > x] \sim x^{-\alpha} \tag{2.9}$$

Siendo $x \to \infty \quad y \quad 0 < \alpha < 2$.

Esto significa que, sin tener en cuenta los valores bajos, la forma asintótica de la distribución es hiperbólica. Así por ejemplo, la función de distribución de cola pesada más simple es la de Pareto, la cual es hiperbólica en todo su rango. En [Addie, 1998] se presenta un ejemplo de aplicación, el modelo M/Pareto, el cual supone un proceso de Poisson con λ ráfagas superpuestas con tasa de bits fija r y duración conforme a una función de Pareto. Según esta, la probabilidad de que la duración T de una ráfaga supere un tiempo t es:

$$\Pr\{T > t\} = \begin{cases} \left(\frac{t}{\delta}\right)^{-(3-2H)} & \forall \quad t > \delta \\ 1 & resto \end{cases} \tag{2.10}$$

Con δ>0 y 0.5 < H < 1.

2.2.3.6. **Modelos autosimilares**

Precisamente, el comportamiento de cola pesada fue una de las razones por las cuales en [Leland, 1993a] se asegura que el tráfico Ethernet, y posteriormente en [Paxson, 1995] con el tráfico Internet, presentan una naturaleza fractal, esto es, la estadística de su autocorrelación es prácticamente la misma independientemente de la escala de tiempo a la que se realice su observación, diciéndose que el tráfico es autosimilar (*self-similar*). Además, esta característica hace que la varianza decaiga de forma lenta, y que estos sistemas presenten una dependencia a largo plazo (LRD – *Long Range Dependence*).

Según Mandelbrot, un objeto es autosimilar o auto-semejante cuando cada una de sus partes presenta formas o estructuras iguales a todo él, aunque a diferentes escalas e incluso con pequeñas variaciones, permitiendo establecer hasta tres grados de autosimilitud (o fractalidad), véase[Mandelbrot, 2004]:

- Exacta: Es la más restrictiva al conservar forma y estructura a diferentes escalas. Normalmente está asociada a sistemas de funciones iteradas.
- Cuasiautosimilitud: La forma y estructura sufre distorsiones conforme se modifica la escala de observación. Normalmente asociada a sistemas con relaciones de recurrencia.
- Estadística: El objeto al modificar la escala mantiene numérica y estadísticamente sus valores. Está relacionado con la aleatoriedad de ciertos procesos, entre los que se pueden encuadrar la mayor parte de los relacionados con el tráfico Internet.

Matemáticamente, partiendo de un conjunto de series de medias nulas y estacionarias en el tiempo de la forma:

$$X = \{X_n\}_{n=0}^{\infty} \tag{2.11}$$

La agregación en bloques no solapados de tamaño m es posible sin más que realizar la suma de las diferentes series temporales:

$$X^{(m)} = \left\{X_k^{(m)}\right\}_{k=0}^{\infty} \tag{2.12}$$

Se dice que X es autosimilar, si para todos los valores de m, $X^{(m)}$ presenta la misma distribución que X, con un factor de escala de m^{-H} resultando:

$$X_n = m^{-H} \sum_{i=(n-1)m+1}^{nm} X_i = m^{-H} X^{(m)} \tag{2.13}$$

Para todo $m \in N$.

A partir de su función de autocovarianza $K_{XX}(k)$, definida como:

$$K_{XX}(k) = E\left[(X_t - \mu_t)(X_{t+k} - \mu_{t+k})\right] = E\left[X_t X_{t+k}\right] - \mu^2 = R_{XX}(k) - \mu^2 \tag{2.14}$$

Para todo t y k, y $\mu_t = \mu_{t+k} = \mu$.

A $R_{XX}(k)$ se le denomina función de autocorrelación (ACF), aunque suele aparecer normalizada por la varianza de la serie:

$$\rho_{XX}(k) = \frac{R_{XX}(k)}{\sigma^2} \tag{2.15}$$

Se dice que X es asintóticamente autosimilar si:

$$\rho_{XX}(k) \sim c_\rho k^{-(2-2H)} \quad , k \to \infty \tag{2.16}$$

Donde $c_\varrho > 0$. Es decir, la autocorrelación decae hiperbólicamente y no exponencialmente, lo que implica que la ACF no es sumable, y se dice que X presenta dependencia a largo plazo (LRD).

El exponente *H* (que también aparecía anteriormente en el modelo *M*/Pareto) recibe el nombre de parámetro de Hurst, y proporciona una aproximación para la medida del "grado" de autosimilitud del proceso, lo que algunos autores han denominado "grado de influencia del presente sobre el futuro", y toma como referencia el fenómeno conocido como "movimiento browniano" o "del caminante aleatorio". En el estudio de la autosimilitud, su valor puede oscilar entre *0.5* y *1*, correspondiéndose respectivamente a procesos no autosimilares y autosimilares exactos en ambos límites. Si la igualdad anterior solamente se verifica para la varianza y la función de autocorrelación, se dice que el proceso es autosimilar de segundo orden. Matemáticamente, los valores de *H* permiten clasificar a un sistema en:

- Persistente (correlación positiva): Para $H > 0.5$, por ejemplo, si $H = 0.7$, entonces existe una probabilidad de 70% de que la siguiente muestra de la serie exhiba la misma tendencia que la de la muestra actual.
- Aleatorio (correlación nula o "ruido blanco"): Para $H = 0.5$.
- Antipersistente (correlación negativa): Para $H < 0.5$.

Para la estimación del parámetro de Hurst existen varias alternativas, véase [Sheluhin, 2007], entre los métodos más sencillos destacan:

- Varianza de los agregados (gráfico varianza-tiempo): Se basa en la propiedad según la cual en los procesos autosimilares la varianza decae de forma lenta. Para ello se representa en escalas logarítmicas la varianza de la serie en función del número de muestras. Una pendiente $-b$ mayor que -1 indica autosimilitud, teniendo en cuenta que $H = 1-b/2$.
- Gráfico de Reescalado (Gráfico R/S): Grafica en escalas logarítmicas la variación de la relación entre el rango y la desviación estándar en función del número de muestras consideradas. La pendiente determina el valor estimado de *H*.
- Estimación de la función de autocorrelación: Describe la propiedad de LRD mediante el uso de la autocorrelación, ya que ésta no debe presentar tendencia a cero. Al representar los coeficientes de correlación en escalas logarítmicas, su tendencia de caída da una aproximación de *H*.

Aunque la característica de autosimilitud aparece una y otra vez a lo largo de los diferentes modelos estudiados en los apartados anteriores, existen casos específicos que generan directamente procesos autosimilares exactos, como es el caso del movimiento browniano o del sistema M/G/∞, o al menos cuasi-perfectos mediante mapas caóticos o fuentes ON-OFF.

2.2.3.7. **Fraccional Brownian Motion (fBm)**

El movimiento browniano fraccional es un proceso Gaussiano $B_H(t)$ de media cero con autosimilitud perfecta, con parámetro H que, a diferencia del modelo M/Pareto, se extiende entre *0* y *1*, y que presenta una función de correlación de la forma:

$$E\left[B_H(t)B_H(s)\right] = \frac{1}{2}\left(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right) \qquad (2.17)$$

El movimiento browniano en sí, es aquel con valor de H igual a *0.5*. En [Kode, 2001] recibe el nombre de FLB (*Fractional Leacky Bucket*), siendo utilizado en el dimensionado de recursos, el cálculo del retardo extremo a extremo y el multiplexado de fuentes de tráfico.

El valor de H determina, por ejemplo, los tiempos de muestreo necesarios para la estimación de los parámetros de tráfico a partir de medidas experimentales. Así en [Awerbuch, 1998] se demuestra cómo el tiempo de muestreo debe ser siempre inferior al minuto salvo para aquellos casos con valores de H comprendidos entre *0.5* y *0.8*. Concretamente en [Caglar, 1998] se especifica que para valores comprendidos entre *0.85* y *1*, el tiempo de muestreo no debe superar el segundo.

Los resultados del modelo fBm son también obtenidos mediante casos particulares de modelos autorregresivos del tipo ARIMA, con d un número no entero. Surgen así los denominados modelos FARIMA (Fraccional ARIMA) ó ARFIMA, utilizados por ejemplo en [Krunz, 1998] para el modelado del tráfico de vídeo.

Todos los modelos indicados anteriormente presentan una complejidad directamente relacionada con las características autosimilares del tráfico, intentando emular las dependencias, tanto a corto como a largo plazo. Utilizando el concepto de la transformación tiempo-frecuencia, surgen los modelos basados en *wavelets*. Estos modelos transforman el proceso autosimilar del dominio temporal al de la frecuencia, con la esperanza de que el proceso resultante presente características de autosimilitud menos marcadas y así sea posible aplicar modelos más simples para, una vez resueltos, retornar al dominio temporal. La transformada *wavelet* no es sino una representación tiempo-frecuencia similar al análisis armónico, existiendo variantes discretas, normalmente utilizadas en la codificación de señal, y continuas, utilizadas en el análisis de señales. En [Sheng, 2001] se compara y complementa el uso de los *wavelets* con modelos *fBm* y FARIMA.

2.2.3.8. **El Modelo M/G/∞**

Tradicionalmente, la caracterización del tráfico IP, desde el punto de vista del conjunto de la pila de protocolos, suele ser simplificada mediante un único

sistema de colas M/G/∞, véase [Pieda, 1999] , justificado por el hecho de que este modelo permite generar tráfico asintóticamente autosimilar. El modelo M/G/∞, aplicado por ejemplo al caso del tráfico telefónico, considera que las peticiones de servicio al nivel de sesión siguen el modelo clásico de Poisson. Sin embargo, las características intrínsecas del tráfico de ráfagas que se observa en el caso de IP, son muy diferentes a las del tráfico telefónico tradicional, dando paso a trazas con valores elevados de autocorrelación sobre largos períodos de observación. Esto ha hecho necesario utilizar las denominadas funciones de distribución de cola pesada como solución para el modelado de tráfico IP cuya correlación a largo plazo presente valores diferentes de cero conforme aumenta el período de observación, comportamiento completamente opuesto al de los modelos de Poisson. Tomando como referencia el caso de la correlación a largo plazo, diferentes correcciones de los modelos de Markov han ido adaptando su comportamiento a la autosimilitud, adoptando figuras de autocorrelación con marcada caída hiperbólica, como se muestra en [Leland, 1993b] y [Mondragon, 2001]. Para ello, algunos de estos modelos parten de considerar procesos de llegadas de Poisson, siempre y cuando el tiempo de servicio presente características autosimilares. La mayoría de las veces la duración del proceso de servicio es modelada por una función de distribución de Pareto o de Weibull [Paxson, 1995]. Sin embargo, en función de la aplicación y del tipo de servicio, el uso de colas pesadas puede ser simplificado considerando distribuciones clásicas en casos particulares, como por ejemplo las exponenciales negativas en el caso de tráfico de VoIP.

Sin embargo, el uso de sistemas M/G/∞ no es del todo útil cuando es el propio proceso de llegadas el que presenta características autosimilares. Este comportamiento implica la necesidad de realizar el modelado del tráfico en toda la escala temporal, y no sólo al nivel de llamada, como es el caso contemplado por los modelos de cola markovianos.

2.2.3.9. **Mapas caóticos**

Se define caos como el fenómeno mediante el cual un sistema dinámico determinístico no lineal, descrito mediante un conjunto limitado de variables y parámetros, presenta un comportamiento complejo y aparentemente aleatorio. Las variables del modelo evolucionan en el tiempo tomando valores que solamente pueden ser relacionados con sus estados previos de acuerdo a una o varias leyes dinámicas no lineales, como por ejemplo ecuaciones diferenciales, [Gleick, 1988]. Pese a ser un sistema dinámico basado en leyes dinámicas deterministas, su comportamiento resulta inherentemente impredecible, propiedad que recibe el nombre de SIC (*Sensitive dependence on Initial Conditions*), es decir, pequeños cambios en las condiciones iniciales conducen a grandes discrepancias en los resultados. De esta manera dos trayectorias arbitrariamente similares de un mismo sistema caótico difieren una de otra exponencialmente. Matemáticamente la propiedad SIC toma la expresión:

$$\left| f^N(x_0+\varepsilon) - f^N(x_0) \right| = \varepsilon e^{N\lambda(x_0)} \tag{2.18}$$

Donde $f^N(\cdot)$ representa las N iteraciones de un mapa caótico definido por la expresión $x_{n+1} = f(x_n)$, x_0 y $x_0 + \varepsilon$ son las condiciones iniciales de dos trayectorias similares, y $\lambda(x_0) > 0$ se denomina exponente de Liapunov[8] (caracteriza el grado de separación de dos trayectorias infinitesimalmente cercanas).

Todas y cada una de las trayectorias así generadas, pese a divergir unas de otras, convergen hacia estados concretos, denominados "atractores", que suelen ser representados a partir de su "Espacio de Fases", es decir, la representación coordenada de sus variables independientes. En estos sistemas caóticos, es fácil encontrar trayectorias de movimiento no periódico, pero cuasi-periódicas. Los "atractores" denominados extraños suelen tener formas geométricas caprichosas y, en muchos casos, parecidos o similitudes a diferentes escalas. En este caso, a estas formas que son iguales a sí mismas en diferentes escalas, reciben el nombre de fractales.

Son precisamente estos comportamientos los que interesan al estudio de la caracterización del tráfico Internet, no en vano una de sus principales características diferenciadoras, la autosimilitud, es representada como una estructura recurrente de ráfagas dentro de ráfagas a diferentes escalas temporales.

Dependiendo de la tecnología, aplicación e incluso la escala de tiempo el tráfico de paquetes puede ser constante, periódico o aleatorio, presentando correlación a diferentes escalas de tiempo. De hecho, en escalas cortas el tráfico de paquetes presenta cierto determinismo al estar formado por trenes de paquetes equiespaciados y sus sucesivas superposiciones. Este comportamiento se obtiene fácilmente mediante el uso de mapas caóticos simples, como se hace en [Erramilli, 1994 / Mondragon, 1999 / Pruthi, 1995 / Samuel, 1999]. En el caso más general y sencillo, un mapa caótico de una sola dimensión es aquel cuya variable de estado x_n evoluciona en el tiempo de acuerdo a un mapeo no lineal del tipo:

---

[8] Mayor Exponente de Lyapunov (L): Es una estimación de la máxima razón de divergencia entre dos trayectorias del Espacio de Fase cuyas condiciones iniciales difieren infinitesimalmente. Las unidades son bits por unidad de tiempo (en base 2) y se calcula con el algoritmo de Wolf.
Posibilidades:
- $L \leq 0$: serie periódica
- $L > 0$: serie caótica
- L tiende a infinito : serie aleatoria

$$\begin{cases} x_{n+1} = f_1(x_n), & y_n = 0, \quad 0 < x_n \le d \\ x_{n+1} = f_2(x_n), & y_n = 1, \quad d < x_n < 1 \end{cases} \tag{2.19}$$

El mapa anterior será caótico siempre y cuando $f_1(\cdot)$ y $f_2(\cdot)$ cumplan la condición SIC, y como cada condición inicial determina una trayectoria diferente, se comporta de forma análoga a la realización de un proceso estocástico. En estas condiciones, la fuente de tráfico se comporta como la ON-OFF, en función de que x_n tome valores inferiores ó superiores a *d* respectivamente, siendo y_n el proceso que describe la llegada de paquetes.

Así el caso más simple de mapas caóticos es aquel en el que $f_1(\cdot)$ y $f_2(\cdot)$ son lineales, destacando el denominado *Bernoulli Shift*:

$$x_{n+1} = \begin{cases} \dfrac{x_n}{(1-\lambda)} & 0 < x_n \le (1-\lambda) \\ \dfrac{x_n - (1-\lambda)}{\lambda} & (d \equiv 1-\lambda) < x_n < 1 \end{cases} \tag{2.20}$$

El proceso de llegadas y_n resultante es independiente e idénticamente distribuido, de forma que en cada iteración la probabilidad de una nueva llegada es λ. De esta forma los estados ON y OFF están geométricamente distribuidos. El efecto de las ráfagas puede ser obtenido mediante segmentos lineales adicionales, y así por ejemplo, mediante tres segmentos (dos parámetros) se obtienen secuencias similares a las generadas por procesos IPP. Sin embargo, los mejores resultados se obtienen cuando se combinan segmentos no lineales. Es el caso del denominado *Intermittency Map*, extensión directa del *Bernoulli Shift* al modificar el primer segmento y obtener el siguiente mapa:

$$x_{n+1} = \begin{cases} \varepsilon + x_n + c x_n^m & 0 < x_n \le d \\ \dfrac{x_n - d}{1-d} & d < x_n < 1 \end{cases} \quad donde \quad c = \frac{1-\varepsilon-d}{d^m} \tag{2.21}$$

La principal característica de este modelo es que permite generar ráfagas con similar estructura a niveles temporales diferentes, con lo que se captura el efecto de cola pesada que actualmente caracteriza al proceso de inter-llegadas del tráfico Internet. Para ello solamente se requiere fijar valores de ε bastante menores que *d*.

El efecto de cola pesada puede ser intensificado haciendo uso de dos segmentos no lineales, mapa que es denominado *Double Intermittency Map*, de acuerdo con la siguiente expresión:

$$x_{n+1} = \begin{cases} \varepsilon_1 + x_n + c_1 x_n^{m_1} & 0 < x_n \le d \\ 1 - \varepsilon_1 - (1 - x_n) - c_2 (1 - x_n)^{m_2} & d < x_n < 1 \end{cases} \quad donde \quad \begin{matrix} c_1 = \dfrac{1 - \varepsilon_1 - d}{d^{m_1}} \\ c_2 = \dfrac{\varepsilon_2 - d}{(1-d)^{m_2}} \end{matrix} \; y \quad \varepsilon_1, \varepsilon_2 \ll d$$

(2.22)

Precisamente la caracterización autosimilar de los procesos generados puede ser programada, o dicho de otra manera, estos modelos permiten obtener fuentes de tráfico autosimilar con valores concretos del parámetro de Hurst, teniendo en cuenta que este depende exclusivamente de la porción no lineal. Para el *intermitency map H* puede ser aproximado por:

$$H = \frac{3m - 4}{2(m-1)} \tag{2.23}$$

Y en el caso del *double intermittency map*, se evalúa de acuerdo a las siguientes condiciones:

$$H = \begin{cases} \dfrac{3m_1 - 4}{2(m_1 - 1)} & m_1 > m_2 \ge \dfrac{3}{2} \\ \dfrac{3m_2 - 4}{2(m_2 - 1)} & m_2 > m_1 \ge \dfrac{3}{2} \\ const. & \dfrac{3}{2} > m_1, m_2 > 1 \end{cases} \tag{2.24}$$

En [Samuel, 1999] se hace un estudio exhaustivo del comportamiento de *H* en función de los diferentes parámetros utilizados en estos modelos, así como su aplicación para el modelado de tráfico agregado.

2.2.3.10. **Modelos ON-OFF**

La alternativa más simple para el modelado de tráfico, y que es aplicada más adelante como elemento básico de este trabajo, ha estado liderada por modelos muy sencillos, aplicados inicial y tradicionalmente a fuentes de voz, denominados en general como ON-OFF. En [Adas, 1997] se hace un estudio amplio de estos modelos y que posteriormente han sido aplicados para el modelado del tráfico de Internet, como por ejemplo en el caso del servicio Web, tal como se expone en [Barford, 1998]. Como su nombre indica, este tipo de modelos parten de la idea de que la transmisión de voz se produce de forma discontinua, con la alternancia entre sonidos y silencios. Por tanto, la fuente de voz oscila entre dos estados, un estado de actividad (ON), que se corresponde con la transmisión, y un estado inactivo (OFF) o de silencio. El paso al estado de actividad sigue una función de distribución invariante en el tiempo que, en el caso de la duración de los diferentes períodos, sigue una exponencial negativa. Así, por ejemplo, cuando el tiempo entre llegadas en el estado ON es de tipo Poisson, el modelo recibe el nombre de IPP (*Interrupted Poisson Proccess*), siendo el caso más simple el de los modelos MMPP comentados anteriormente.

Aparte de su simplicidad, una de las principales restricciones del modelo ON-OFF, tal como se apunta en [Sohraby, 1993], es que el comportamiento modelado suele estar más cercano al de fuentes de tráfico con velocidades de transmisión equivalentes a las máximas correspondientes al enlace físico. Es decir, la fuente en estado ON transmite a velocidad máxima, lo cual no es cierto, ya que la tasa media de transmisión de la fuente siempre es inferior a la del enlace.

Además, los modelos ON-OFF tradicionales, asumen distribuciones del tiempo de permanencia en cada estado con varianza finita, modificando los estados ON y OFF mediante funciones de distribución diferentes a la exponencial negativa, como por ejemplo expone[Navickas, 2006], donde se extiende a distribuciones de Weibull ó Lognormal. Sin embargo, esto supone que la agregación de un número suficiente de fuentes de este tipo no presenta prácticamente correlación salvo a muy corto plazo. La extensión natural de este tipo de fuentes considera distribuciones con varianza infinita (como aplicación del efecto Noah[9]), con lo que su agregación presenta una gran dependencia a

[9]Efecto Noah: también llamado "síndrome de la varianza infinita", según el cual determinadas series temporales presentan desviaciones típicas puntuales muy superiores a las esperadas, tendentes a infinito.

largo plazo (consecuencia del efecto Joseph[10]), comportamientos expuestos en [Adas, 1997]. Según esto, cuando las transiciones entre el estado ON y OFF siguen una función de distribución "de cola pesada", como una Pareto con parámetros a_1 y a_2, el proceso estocástico observado es similar a un ruido gaussiano fraccional con parámetro de Hurst $H=(3-\min(a_1, a_2))/2$. Así, por ejemplo, [Schwefel, 1999] plantea el uso de Fuentes ON-OFF con períodos OFF distribuidos exponencialmente, y períodos ON con distribución de Pareto, es decir, una evolución de los procesos MM (*Markov Modulated*), resuelto mediante un sistema semimarkoviano del tipo SM/M/1 y que los autores denominan modelo ON-OFF "1-Burst".

2.2.3.11. **Modelos basados en el comportamiento del tráfico**

Todos los modelos expuestos anteriormente, consideran a la fuente de tráfico como un sistema normalmente complejo, con comportamientos dependientes, en su mayor parte, de los usuarios y de los datos intercambiados. Algunos autores defienden que, ante todo, las fuentes de tráfico dependen básicamente del comportamiento de la red, que es la que al final "modela" los flujos de tráfico resultantes de un determinado usuario haciendo uso del correspondiente servicio. Tal es la importancia de la red que, en algunos casos, ésta puede ser modelada en síi misma como un único elemento, cuya función es la de acomodar, "conformar", el tráfico generado por los usuarios.

Este concepto es ampliamente utilizado para describir determinadas políticas de control de acceso, como las denominadas *shaping* y/o *scheduling*,. De hecho, dos de las técnicas básicas de *shaping* son también utilizadas como base de modelos de tráfico de fuente básicos. Es el caso de los modelos de *leaky bucket* y *token bucket*. Muchas veces, ambos mecanismos son erróneamente confundidos como uno solo, si bien sus propiedades y aplicaciones son bastante diferentes. Su principal diferencia radica en que, mientras el *leaky bucket* impone un límite estricto para la tasa de bit de la fuente, el *token bucket* solamente limita su valor medio.

[10] Efecto Joseph: considera que las variaciones que sufren las series temporales siguen ciclos y tendencias fijas pese a la aleatoriedad de las mismas, especialmente en períodos largos.

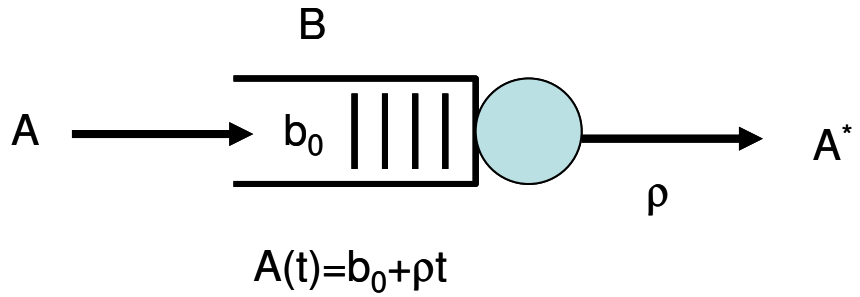


Fig. 2.12: Representación conceptual de un *Leaky bucket* (B, ϱ)

Básicamente, el modelo de *leaky bucket*, representado en la Fig. 2.12., considera a la fuente de tráfico como un flujo de tasa de bit constante, siempre y cuando el tráfico de usuario no supere un umbral máximo, representado éste como un *buffer*. Así, por ejemplo, en [Kode, 2001] el tráfico de tasa variable (VBR) es modelado mediante un *leaky bucket* de tasa de bit *p* y tamaño del *buffer B*, y representado como *(B, p)*. En un determinado intervalo de tiempo *T*, la cantidad de datos generados están limitados por *(B+ϱT)*. Es decir, en un momento dado, las ráfagas de datos no superan nunca los *B* bytes, siendo la tasa de servicio *ϱ*.

Los bits que llegan al *buffer* son automáticamente encolados y servidos. Sin embargo, en Internet, el flujo de bits no es siempre constante puesto que la unidad de datos mínima es el paquete. Es por ello que el *token bucket* define una variante del *leaky bucket*, en la que la transmisión está condicionada a la posibilidad de enviar unidades de datos completas, ya sean paquetes o bloques de bytes, denominadas *token*. Según [Sharafeddine, 2003] un *token bucket* queda modelado por el conjunto de parámetros *S=(r, b, p, m, M)*, siendo *b* el tamaño máximo de ráfaga, *r* la tasa de generación de *token*, *p* la tasa máxima del tráfico entrante, *m* el tamaño mínimo de datos y *M* el tamaño máximo de datos (paquete/datagrama).

Sin embargo, tal como se indica en [Tang, 1999], el IETF en su *IntServ Working Group* define el *token bucket* como un contador que acumula *tokens* a una velocidad constante *r* hasta que alcanza un máximo *b*. A medida que se reciben paquetes, estos solamente serán cursados si el contador contiene al menos suficientes *tokens* como bytes contenga. En caso contrario, el paquete es descartado, o bien encolado si el *token bucket* incluye un *buffer*. En este último caso se obtiene un efecto de "suavizado" (*smothing*) del tráfico a costa de la introducción de retardo. De esta forma, el modelo queda definido mediante el par *(r, b)* y el tamaño de la cola *q*, tal como se muestra en la Fig. 2.13.

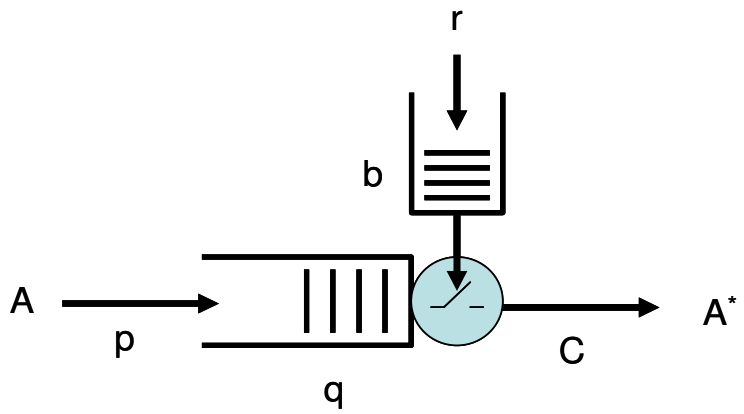


Fig. 2.13: Conceptualización de un *Token bucket*

En el caso de un *token bucket* ideal, sin pérdidas ni retardo, los valores límite de *r* y *b* quedan determinados. Así, por ejemplo, si se considera un flujo P_n tal que, en un intervalo de tiempo $[t_o, t_n]$ genera paquetes de tamaño $[p_0, p_1,\ldots p_n]$, la velocidad del contador *r* debe ser lo suficientemente rápida como para que el *token bucket* permita cursar cualquier paquete p_i. De acuerdo con esto los límites resultantes son:

$$r_{\min} = \max_{1\le k\le n}\left[\frac{\sum_{i=1}^{k} p_i - c_0}{t_k - t_0}\right] \tag{2.25}$$

siendo c_0 el valor inicial del *token bucket* y $k>0$

$$r_{\max} = \max_{2\le k\le n}\left[\frac{p_1 - c_0}{t_1 - t_0}, \frac{p_k}{t_k - t_{k-1}}\right] \tag{2.26}$$

y por su parte para *b*:

$$b_{\min} = \max_{1\le k\le n}\left[p_k, c_0\right] \tag{2.27}$$

$$b_{\max}(r) = \max_{1\le i\le n}\left[c_i(r)\right] \tag{2.28}$$

donde

$$c_k(r) = c_0 + r(t_k - t_0) - \sum_{i=1}^{k-1} p_i \tag{2.29}$$

Tal es la importancia del modelo de *token bucket*, que suele ser utilizado recursivamente en los estudios relacionados, bien en su forma original, como en variaciones basadas en el encadenado de múltiples módulos *Leaky* y *token*. Así, por ejemplo, en [Dovrolis, 1997 / Schmitt, 1999 / Schmitt, 2001], se

identifica el comportamiento de un enlace Internet como dos *token bucket* en serie, estructura que denominan *(b, r, p, M)-regulator*, y que se muestra en la Fig. 2.14:

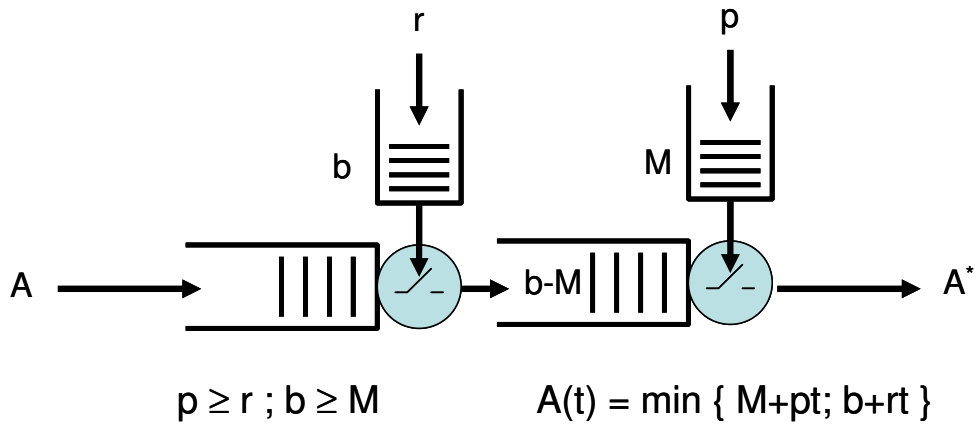


Fig. 2.14: Enlace Internet desde el punto de vista del IETF (Servicios Garantizados)

El IETF define la denominada Clase de Servicio Garantizado, según la cual se asegura un ancho de banda sin pérdidas y con retardo máximo limitado. El tráfico asociado a este tipo de servicio recibe el nombre de *Tspec* y suele ser modelado mediante un *(b, r, p, M)-regulator*, siendo sus parámetros fundamentales:

- la capacidad del *bucket b* (bytes)
- la tasa de servicio del *token bucket r* (bytes/seg)
- la tasa máxima *p* (bytes/seg)
- el tamaño máximo de los paquetes *M* (bytes)
- la mínima unidad de datos supuesto que exista control del tráfico *m* (bytes)

El hecho de establecer límites claros para el retardo extremo a extremo, permite caracterizar, mediante este modelo, a la mayor parte de los tráficos de fuente relacionados con las redes Internet. Es un modelo que surge como aplicación de los estudios realizados a partir de la obtención de curvas de llegada y de servicio en lo que se ha venido denominando *Network Calculus* y que se explica con más detalle en el apartado 2.4. Según este método, el tráfico *Tspec* está caracterizado por una curva de llegada denominada *Tspec*(*r, b, p, M*) de la forma de la Fig. 2.15, siendo sus expresiones correspondientes:

$$a(t) = \min\left(M + pt, b + rt\right) \tag{2.30}$$

para la curva de llegada y

$$c(t) = R\left(t - V\right)^{+} \tag{2.31}$$

para la curva servicio, con

$$V = \frac{C}{R} + D \tag{2.32}$$

Donde R es la tasa de servicio, y los términos C y D dependen casi exclusivamente del tipo de política de control de acceso implementado (p.e. en el caso de PGPS : $C=M$ y $D=M'/c$, siendo M el tamaño máximo de paquete, M' la MTU y c la velocidad del enlace).

De acuerdo con este comportamiento, la capacidad que es necesario reservar en un determinado trayecto para poder asegurar un retardo máximo d_{max} se calcula como:

$$R = \begin{cases} p \dfrac{b-M}{p-r} + M + C \\ \hline d_{\max} + \dfrac{b-M}{p-r} - D \end{cases} \quad p \geq R \geq r \qquad \frac{M+C}{d_{\max} - D} \quad R \geq p \geq r \tag{2.33}$$

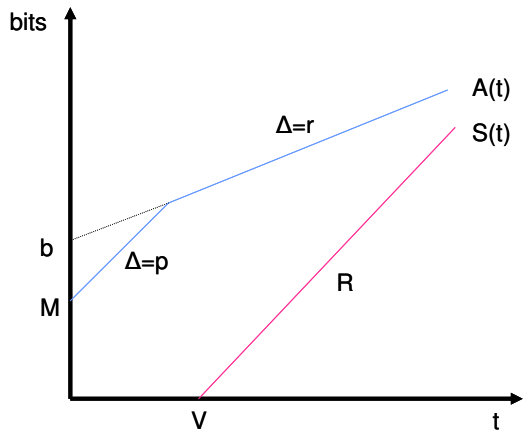


Fig. 2.15: Representación de curvas *TSpec*: A(t) - Curva de llegada y S(t) - Curva de servicio

Esta aproximación presenta dos problemas:

- No puede ser aplicada más que en los puntos en los que se realice la agregación, de forma que los nodos contienen recursos proporcionales al número de fuentes que dependen de él, no siendo éste siempre el caso cuando se trata de la red dorsal.
- Mantener los criterios de bajo retardo supone la necesidad de gran cantidad de recursos, por lo que la agregación de fuentes individuales sólo aumenta el sobredimensionado del tráfico agregado.

2.2.4. **Estimación del tráfico de fuente**

Desde el punto de vista de la planificación estratégica de redes, el conocimiento ó estimación del valor asociado a parámetros específicos del tráfico relacionado es fundamental. El uso de aproximaciones es siempre bien recibido, frente a métodos de cálculo más exactos, pero con complejidad de cálculo creciente. El uso de estimaciones basadas en aproximaciones obtenidas a partir de los modelos de fuente de tráfico es entonces más importante aún. La mayoría de las veces reducen el cálculo a los estadísticos fundamentales de un conjunto reducido de parámetros, como puede ser la tasa de bit, el retardo y/o la capacidad del enlace.

En la literatura, el concepto de capacidad toma diferentes significados, aunque a nivel de estimación de tráfico destacan especialmente dos acepciones: Capacidad efectiva frente a capacidad equivalente de una fuente, como se expone en [Chen, 2007 / Jaising, 2002]. La capacidad efectiva se define como la cantidad mínima de ancho de banda que asegura que un flujo de tráfico pueda ser cursado dentro de unos parámetros de QoS dados, esto es, sin superar ni un retardo máximo, ni en el peor de los casos, un porcentaje máximo de paquetes perdidos. La capacidad efectiva no es un valor constante, sino que depende de las características de la red en general y del enlace en concreto. Por su parte, la capacidad equivalente se define como la cantidad mínima de ancho de banda que asegura que el flujo de tráfico pueda ser cursado, considerando un sistema de cola finito, sin superar un número máximo de paquetes perdidos. En este caso, su valor depende de cada elemento (en general de la red), del tamaño del *buffer* y de la carga soportada, esto es, de la existencia de más fuentes activas en el mismo elemento.

Ambos conceptos aparecen de forma reiterada en líneas de estudio diversas, aunque convergentes, especialmente en todos los temas relacionados con los conceptos: Servicios Garantizados, Control de Acceso a la Conexión (CAC) y *Network Calculus*. Precisamente este último campo permite establecer de forma gráfica la diferencia fundamental entre ambos conceptos, tal y como se describe en [Boudec, 1996] y se muestra en la Fig. 2.16.

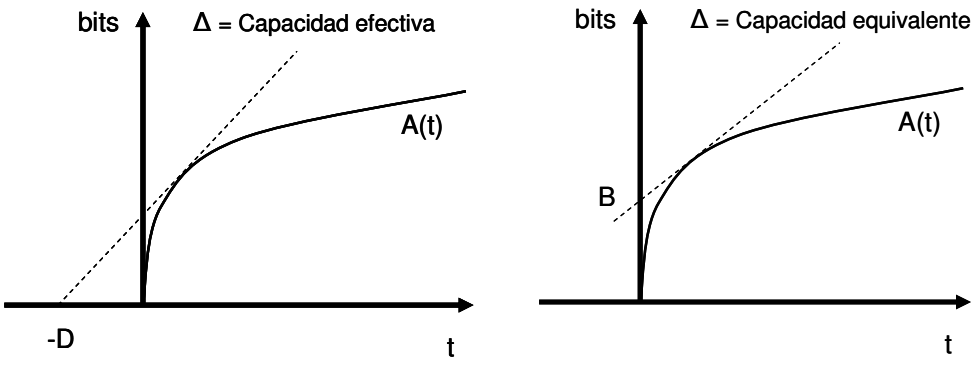


Fig. 2.16: La capacidad efectiva como función del retardo; la capacidad equivalente como función del *buffer*

Sin embargo, existen otros autores que definen ambas capacidades como la misma, en función ya no del tamaño del *buffer* y/o del retardo, sino en términos absolutos a partir de las características de la fuente y/o la probabilidad de pérdida. El caso más sencillo, desarrollado en [Courcoubetis, 1995] calcula su valor haciendo uso de un modelo de fluidos basado en cadenas de Markov de dos estados con tasas 0 y *r* respectivamente, y probabilidades distribuidas según una exponencial de parámetros λ y μ, tal como se indica en la figura siguiente:

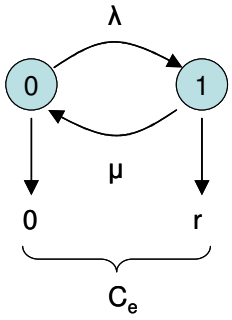


Fig. 2.17: Modelo de fluidos de dos estados

Su ancho de banda efectivo puede aproximarse por:

$$C_e = \frac{\lambda r}{\lambda + \mu} + \frac{\delta \lambda \mu r^2}{(\lambda + \mu)^3} + o(\delta) \tag{2.34}$$

Siendo normalmente δ bajo $(1/8)$ y $o(\delta)$ despreciable en la mayoría de los casos.

La expresión se simplifica sin más que elegir un valor de *r* múltiplo del tamaño de una cola de un proceso M/M/∞ modulado. Este valor se obtiene a partir del límite del producto $N\lambda$, cuando el número de fuente *N* tiende a ∞ (λ tiende a 0 y $N\lambda$ a una constante). El ancho de banda efectivo se expresa ahora como:

$$C_e = \frac{r}{\mu} + \frac{r^2}{\mu^2}\delta + o(\delta) \tag{2.35}$$

El mismo caso, pero teniendo en cuenta las pérdidas, se desarrolla en [Lee, 1995], que toma como referencia el modelo de tráfico ON-OFF con parámetros (λ, μ, r) y calcula el ancho de banda efectivo, como la capacidad requerida para servir a dicha fuente dentro de los términos de la probabilidad de pérdidas P_p, donde λ y μ modelan la duración de los estados ON y OFF respectivamente, *r* la tasa media de bits y *B* el tamaño del *buffer*:

$$c_e = \frac{\lambda + \mu + rz + \sqrt{(\lambda + \mu + rz)^2 - 4\lambda rz}}{2z} \tag{2.36}$$

Siendo:

$$z = \frac{1}{B}\log\left(\frac{1}{P_l}\right) \tag{2.37}$$

Esta expresión, que parte de funciones de distribución exponenciales, es generalizada en [Nahas, 2002] para funciones de distribución del tipo $E\left[e^{\theta x(t)}\right]$, haciendo uso del denominado "límite de Chernoff" [11] para determinar un valor de θ a partir del cual calcular la capacidad efectiva de la fuente como:

$$C_e = \lim_{t\to\infty}\frac{1}{t}\frac{1}{\theta}\ln\left(E\left[e^{\theta A(t)}\right]\right) \tag{2.38}$$

Siendo:

$$\theta = \frac{1}{x}\ln\left(\frac{1}{P_l}\right) \tag{2.39}$$

donde x es el tamaño máximo de una ráfaga. Así, por ejemplo, en el caso de una fuente de Poisson la capacidad efectiva es aproximada por:

$$C_e = \lambda\left(\frac{e^{\theta}-1}{\theta}\right) \tag{2.40}$$

Otra versión muy similar es la aproximación que hace[Courcoubetis, 2002], aunque en este caso en función de la carga total producida por la fuente X en un intervalo de tiempo [*0, t*]:

$$\alpha(s,t) = \frac{1}{st}\log E\left[e^{sX[0,t]}\right] \tag{2.41}$$

siendo t (ms) el tiempo transcurrido hasta bloquear el *buffer* y s (1/Kb) el grado de multiplexado dependiente, entre otros, de la relación entre la tasa de bit máxima y la capacidad del enlace. En particular, para capacidades del enlace muy superiores a las tasas máximas individuales, s tiende a 0 y el ancho de banda efectivo se acerca al valor medio de la fuente. Cuando la capacidad del enlace es fija (determinada respecto a los máximos de las fuentes y entonces independiente de los estados de la cadena), s crece y hace que el ancho de banda efectivo se aproxime a los valores máximos de X/t.

[11] Para una variable aleatoria x(t) con función de distribución $E\left[e^{\theta x(t)}\right]$, el límite de Chernoff se define como $P[q > x] \le e^{-\theta x}E\left[e^{\theta q}\right]$, siendo q el tamaño del buffer del switch y θ siempre positivo.

Precisamente, es el comportamiento conjunto de varias fuentes que comparten un determinado elemento de la red el que más ha sido estudiado. En los estudios más simples, como por ejemplo en [Awerbuch, 2001 / Courcoubetis, 1995], un elemento de red al que confluyen varias fuentes de tráfico *N*, clasificadas éstas en función del tipo y características del tráfico en *M* clases diferentes, solamente podrá cursar todo el tráfico resultante si su capacidad *C* cumple:

$$\sum_{i=1}^{M} N_i B_i \leq C \tag{2.42}$$

siendo B_i el ancho de banda efectivo de una fuente de tráfico de clase *i*, a su vez menor o igual a la tasa de bit máxima de la fuente.

Las fuentes de tráfico a ráfagas presentan valores de ancho de banda efectivo mayores a su media, aunque no necesariamente igual a sus tasas máximas. Sin embargo, al considerar un conjunto de fuentes independientes entre sí, se demuestra que el agregado presenta un suavizado de los anchos de banda globales, lo que justifica la utilización del multiplexado estadístico. De hecho en [Wang, 1998], partiendo de esta idea, se propone el cálculo de la capacidad equivalente de una agrupación de fuentes C_e (mediante multiplexación estadística) a partir de los momentos característicos de las mismas:

$$C_e = \min\left\{\sum_{i=1}^{N}\rho_i + \alpha\sqrt{\sum_{i=1}^{N}\sigma_i^2}, \sum_{i=1}^{N} R_i\right\} = \min\left\{\sum_{i=1}^{N}\rho_i + \alpha\sqrt{\sum_{i=1}^{N}\rho_i\left(R_i-\rho_i\right)}, \sum_{i=1}^{N} R_i\right\} \tag{2.43}$$

siendo ρ_i y σ_i^2 la tasa de bit media y la varianza para la fuente i-ésima, R_i las tasas de bit máximas para cada fuente *i,* y *a* un factor de ponderación en función de la tasa de errores esperada ε:

$$\alpha = \sqrt{-2\ln(\varepsilon) - \ln(2\pi)} \tag{2.44}$$

En [Jaising, 2002], para cada agrupación homogénea de fuentes del mismo tipo, la expresión de la capacidad equivalente es calculada como:

$$C_i = (m_i + \alpha\sigma_i)R_i \tag{2.45}$$

siendo *m* y σ la media y la desviación estándar del número de fuentes activas respectivamente.

Con todo, asumiendo el caso más simple, esto es, considerando fuentes moduladas de Markov, los *M* tipos de fuentes estarían superpuestos, y puede calcularse el límite superior para la agregación como:

$$C_e = \sum_{i=1}^{M} C_i \tag{2.46}$$

donde C_i sería la capacidad equivalente de las fuentes de tipo *i.*

2.3. Agregación

Internet presenta una estructura compleja compuesta en su nivel más alto de elementos interrelacionados de diferentes maneras, pero independientes entre sí, denominados Sistemas Autónomos (AS). Estas entidades están normalmente asociadas a operadores y/o proveedores de servicio y conforman la red tal y como se conoce, como una gran dorsal de cobertura mundial. Los AS's pueden ser considerados a su vez, como un conjunto de sistemas interconectados por su propia estructura de encaminadores dorsales (*core routers*), como muestra la Fig. 2.18. Las relaciones entre los diferentes AS's y las redes de distribución y de acceso se establecen a través de *routers* frontera ("*edge routers*"). Por su parte, el usuario cuelga de dichas redes, que encaminan todo el tráfico Internet a través de los POP(*Point of Presence*) de los operadores, nexo de unión entre la dorsal y la red de acceso.

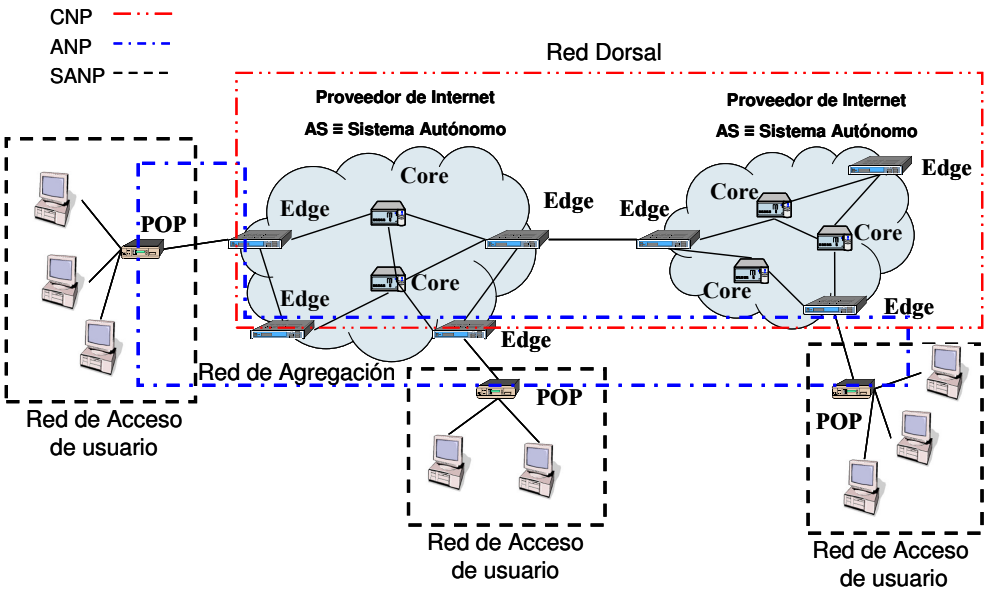


Fig. 2.18: Estructura de la red Internet: *Core Network Part* + *Aggregation Network Part* + *Subscriber Network Part*

De esta manera el tráfico de usuario y, en definitiva, el tráfico de fuente, fluye desde los equipos de usuario hacia los elementos que dan acceso a la red de distribución y de acceso. Las aportaciones individuales correspondientes a cada equipo de usuario van ascendiendo en su estructura, encontrándose con otros flujos y compartiendo el ancho de banda de los elementos que atraviesan. Es entonces cuando los flujos individuales son agrupados/multiplexados/ agregados, y sus características comienzan a alejarse del comportamiento esperado por los modelos de tráfico de fuente.

La agregación de flujos de tráfico se lleva a cabo a lo largo de toda la red, aunque cabe destacar cuatro puntos singulares en los cuales esta acción es más clara y, por tanto, pueden ser tomados como puntos de referencia a la hora de realizar estimaciones de sus características y comportamiento y se muestra en la Fig. 2.19, tal como se propone en [COST-257, 2000]:

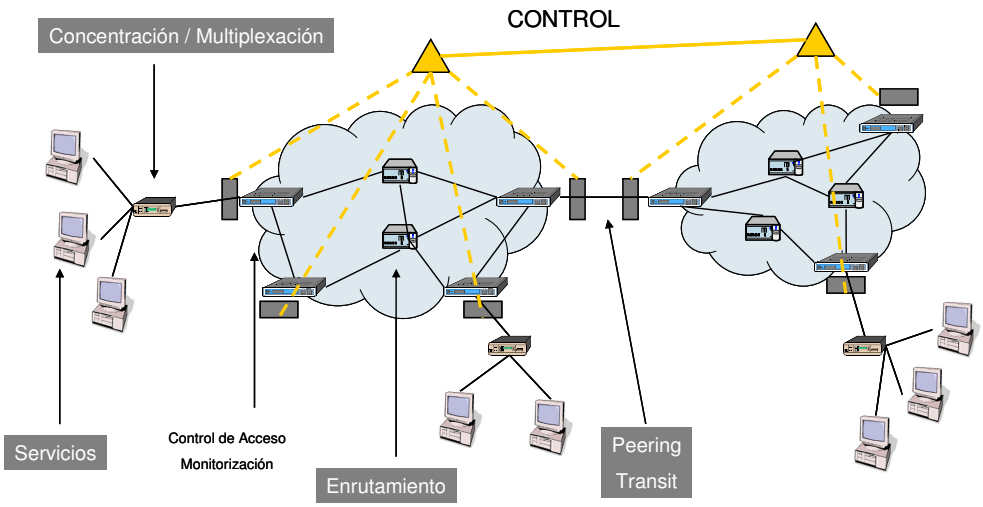


Fig. 2.19: Puntos singulares donde se realiza la agregación de flujos de tráfico, por variedad de servicios, concentración, enrutamiento e intercambio de tráfico

- El equipo de usuario: Cada aplicación o servicio que utiliza el usuario, genera un flujo característico de tráfico que es agregado en la línea de acceso, compartiendo el ancho de banda correspondiente.
- El POP: La red de distribución y de acceso hace uso de elementos específicos para la agregación de los flujos correspondientes a todos o parte de los usuarios pertenecientes a uno o varios áreas de cobertura.
- El *Edge Router*: Los flujos agregados en los correspondientes POP son encaminados hacia la red dorsal, en la que el ancho de banda compartido se corresponde con el de los enlaces troncales. Por otro lado, si el *Edge* realiza la interconexión entre AS's, el tráfico procedente de la dorsal con destino a otro AS (tráfico de *Peering*/*Transit*)[12] será encaminado y agregado en el enlace correspondiente.

[12] *Peering* es la interconexión voluntaria entre operadores con el fin de permitir intercambiar tráfico entre sus clientes. Dos AS's (AS1 y AS2) intercambian su tráfico con origen en AS1 y destino AS2 y viceversa. Típicamente se AS1 y AS2 se encuentran en el mismo nivel de la jerarquía de red (Tier 2 o Tier 1)

- El *Core Router*: Como elemento de interconexión principal, entre diferentes *edge routers* y otros *core routers*, el tráfico es encaminado a través de los enlaces troncales, sufriendo sucesivas agregaciones y extracciones en función de sus destinos.

Adicionalmente, el plano de control de la red, introduce flujos de tráfico adicionales que, además de ser agregados, pueden establecer mecanismos y criterios a aplicar en cualquiera de los procesos de agregación indistintamente de los elementos de red existentes.

El proceso de planificación y dimensionado se ve así, directamente influenciado por el efecto de la agregación de los diferentes flujos, haciendo necesario el modelado y estimación del tráfico resultante. La importancia de la agregación es tal, que la modificación en el comportamiento del tráfico, prácticamente llega a invalidar la mayor parte de los modelos de fuente, obligando a su revisión y corrección, y a la aparición incluso de nuevos modelos, muchos de ellos complejos y computacionalmente costosos. Precisamente, esta Tesis plantea, entre uno de sus objetivos, establecer métodos que permitan realizar una estimación del tráfico agregado de forma simple pero efectiva, asegurando el cálculo rápido de aproximaciones dentro de unos límites establecidos a priori.

2.3.1. **Definición y antecedentes**

La agregación de tráfico consiste en la combinación y transporte de múltiples flujos sobre enlaces compartidos. En el caso de Internet, la capacidad de los enlaces es compartida de forma dinámica entre todas las fuentes, sin reservas ni ajustes previos, aunque con la necesidad de establecer límites estadísticos siempre y cuando se pretenda establecer un nivel mínimo de garantía de servicio para cada flujo. Tradicionalmente se distinguen dos mecanismos de agregación típicos:

- Agregación geográfica (algunos autores la denominan división): Consiste en el agrupamiento por contigüidad de unidades geográficas en unidades de nivel de agregación superior (por ejemplo, tráfico de una red de distribución hacia la de acceso). El modo y el nivel de agregación geográfica de las unidades de referencia, pueden modificar las relaciones existentes entre las propiedades medidas por éstas. Si

---

Transit es la interconexión a través de terceros operadores o redes (otros SA's). AS2 acepta paquetes desde AS1 dirigidos a otros AS. Normalmente AS1 se encuentra en niveles inferiores a AS2.

bien este método es importante, desde el punto de vista físico de la planificación estratégica de redes, no resulta útil si no se incorporan criterios de QoS concretos. Algiunos ejemplos son [Awerbuch, 2001 / Cho, 2001b / Korkmaz, 2000]

- Agregación estadística: Consiste en el agrupamiento de unidades semejantes en función de atributos específicos. Los métodos de clasificación jerárquica utilizan criterios de agregación claros, como por ejemplo "la maximización del momento centrado de orden 2", lo que significa asegurar una varianza máxima entre cada grupo, y mínima dentro de uno mismo. Estos métodos identifican niveles de agregación que traducen la heterogeneidad (la varianza intra-clase) asociada a una cierta partición. Algunos ejemplos de aplicación aparecen en [Boorstyn, 2000 / Caglar, 1998 / Wang, 1998].

Precisamente, las características de la agregación estadística son las que, según [Clark, 1999], permiten establecer límites asociados a los requerimientos necesarios para mantener un grado de servicio aceptable, de forma que no se sobredimensionen ineficientemente los enlaces compartidos (caso de utilizar estimaciones sobre los valores máximos de tasa de bits), ni que se obtengan utilizaciones máximas a costa de degradación del servicio de flujos concretos (caso de utilizar estimaciones basadas en los valores medios a largo plazo del *bitrate*). El efecto resultante combina el comportamiento de los picos de tráfico individuales, por usuario, a lo largo del tiempo sobre el enlace compartido. De esta forma se asegura tanto el comportamiento medio de todos los flujos agregados, como los picos de ancho de banda correspondientes a flujos individuales.

2.3.2. **Modelado de la agregación**

La agregación estadística, solamente es posible gracias al comportamiento del tráfico Internet con ráfagas normalmente muy cortas, aunque un número menor de ellas puedan ser muy largas, lo que ha venido a denominarse "comportamiento de cola pesada" (*heavy-tailed*). Para modelar el comportamiento de las ráfagas Internet se hace uso de distribuciones de Pareto, según la cual, la probabilidad de que una ráfaga supere un tamaño de x bytes es $(1/x)^{\alpha}$, tomando para a valores superiores a *1* (por ejemplo, las ráfagas generadas por el tráfico Web presentan un valor de a de *1.1*).

Sin embargo, en [Vutukury, 2003] se plantean varias alternativas para el modelado de dicho comportamiento, aplicadas tradicionalmente en la agregación de tráfico, normalmente de tipo telefónico.

Partiendo de un enlace de capacidad C que da servicio a N fuentes ON-OFF con parámetros λ, a y V, la multiplexación estadística es, en sí, un sistema de colas, como se representa en la Fig. 2.20 que permite agregar fuentes de forma que C es incluso inferior a N.

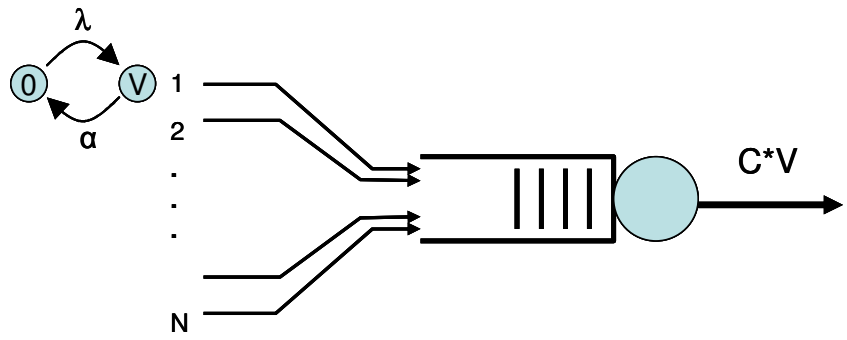


Fig. 2.20: Multiplexación estadística según el modelo ON-OFF

Según este modelo, la probabilidad de que una fuente esté activa es $\lambda/(a+\lambda)$, con lo que su contribución a largo plazo a la tasa de llegadas media es $\lambda V/(a+\lambda)$. Entonces las llegadas correspondientes a las N fuentes son $\lambda VN/(a+\lambda)$, y el *buffer* será dimensionado de forma que las llegadas no superen el producto CV:

$$\rho \equiv \lambda N / \left[(\alpha + \lambda) C \right] < 1 \tag{2.47}$$

Por otro lado, el número de fuentes activas, presenta las mismas características que un proceso de Markov (nacimiento y muerte), como el de la Fig. 2.21:

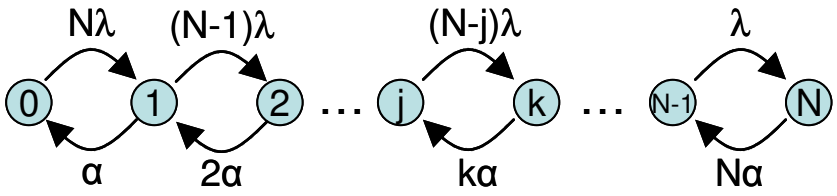


Fig. 2.21: Cadena de nacimiento y muerte de fuentes activas

O bien, considerando que $\rho<1$, el número de fuentes activas en el estado ergódico, queda modelado mediante una variable aleatoria binomial, tal que

$$\Pr\left[k\_activas\right] = \left[\lambda/(\lambda+\alpha)\right]^{k} \left[\alpha/(\lambda+\alpha)\right]^{(N-k)} \tag{2.48}$$

Existen dos aproximaciones:

- Flujo de fluidos, en el que el flujo de paquetes es considerado como un líquido, y el sistema de colas resultante es el de la Fig. 2.22:

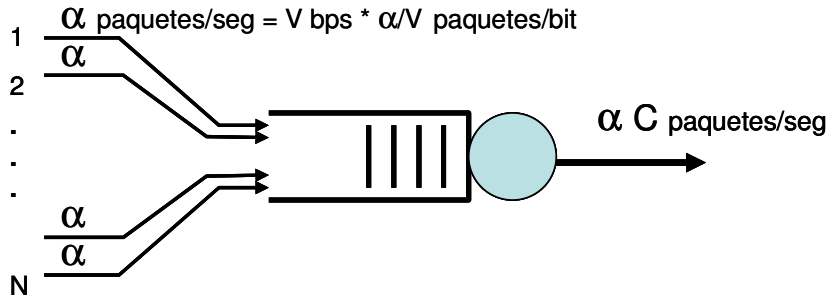


Fig. 2.22: Multiplexación estadística según el modelo de fluidos

Cuando *i* fuentes están activas, existen a_i unidades de información (número de paquetes esperados durante el estado activo V/a) por segundo que entran al *buffer*.

- Procesos MMPP, en el que el flujo de paquetes es un proceso de Poisson. Los paquetes se generan en el estado activo de acuerdo a un proceso de Poisson de tasa β, siendo el tiempo de servicio otra variable aleatoria con distribución exponencial (y tasa de bits del enlace V), tal como se muestra en el sistema de la

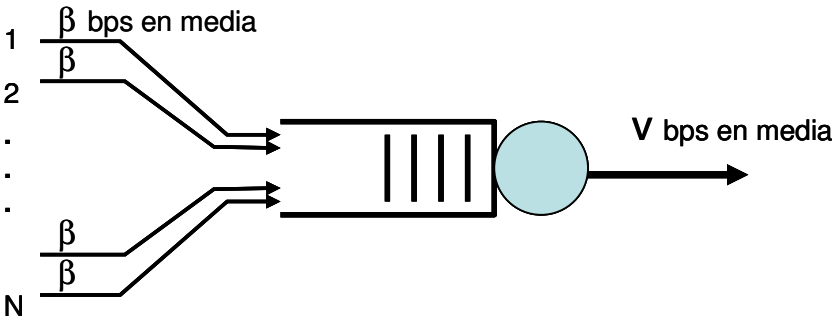


Fig. 2.23:

Fig. 2.23: Multiplexación estadística según los modelos Markovianos

2.3.2.1. **Modelos de fluidos**

De la misma forma que en el caso del modelado de fuentes individuales, el problema de su agregación es resuelto mediante mecanismos que emulan el comportamiento *self-similar* del tráfico Internet, como es la teoría de fluidos o el movimiento browniano. Dentro de este tipo de soluciones, destaca la adaptación del modelo de Engset (conocido para la simulación de sistemas de

pérdidas para tráfico telefónico) que se aplica en [Bonald, 2003], mediante el cual se modelan flujos de tráficos en general, aunque dando especial importancia al tráfico elástico sobre TCP. La consideración del flujo agregado como un fluido se sustenta bajo la suposición de que la capa de transporte (TCP) realiza el reparto de anchos de banda entre los diferentes flujos. De esta forma, en un enlace de capacidad *C* compartido por *N* usuarios con tasas de acceso individuales *c* (1 Mbps por ejemplo en ADSL), los flujos activos van a compartir la capacidad del enlace de forma equitativa. Tomando como referencia la tasa de llegadas *d*, y el tráfico ofrecido *a,* la capacidad necesaria para asegurar a *N* usuarios es:

$$C = \frac{N}{1/a + 1/d - 1/c} \tag{2.49}$$

para $d<c$, o bien $C>Na$ si d=c.

Estas expresiones son aplicadas para el modelado de agregación de fuentes homogéneas, independientemente de las funciones de distribución de llegadas y del *thinking-time*. Precisamente, es esta característica la que los hace especialmente recomendables para el modelado jerárquico de tráfico de datos, donde los flujos son generados a nivel de sesión. Al analizar en detalle el modelo, desde el punto de vista estocástico, no deja de ser una red de colas con procesamiento compartido.

Partiendo del comportamiento a ráfagas del tráfico Internet, cada flujo se corresponde con una sucesión de transferencias de documentos (páginas Web, correos, secuencias de video) y períodos de inactividad, que normalmente reciben el nombre de *thinking-time*, tal como se puede observar en la Fig. 2.24:

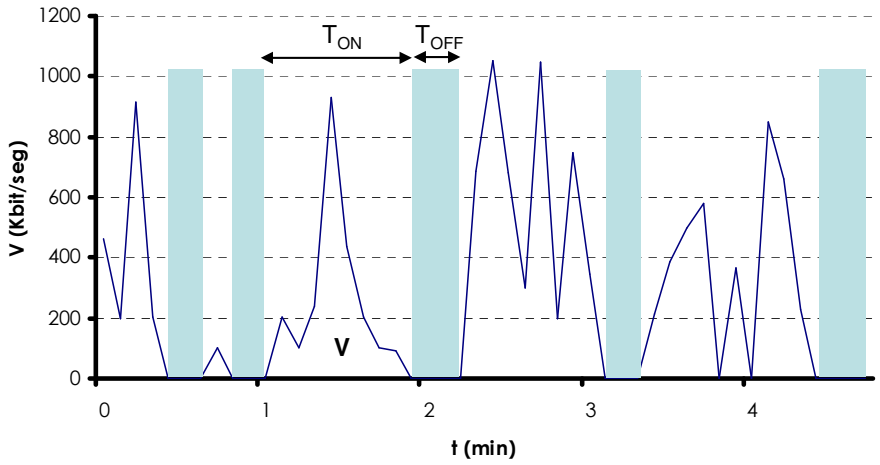


Fig. 2.24: El comportamiento del tráfico a ráfagas según el modelo de fluidos

Tomando V como el volumen medio de datos, T_{ON} el tiempo medio de transmisión y T_{OFF} el *thinking-time*, el valor de d se calcula como $d=V/T_{ON}$.

Si la tasa de acceso es estable, el tráfico ofrecido se calcula como:

$$a = \frac{V}{V/c + T_{OFF}} \tag{2.50}$$

y el tráfico cursado como:

$$b = \frac{V}{T_{ON} + T_{OFF}} \tag{2.51}$$

En [Bonald, 2003] también se obtiene una modificación del modelo anterior para su uso en agregación de tráfico heterogéneo. Para ello considera K clases de tráfico, donde la clase k tiene N_k usuarios con tráfico ofrecido a_k para todo k entre 1 y K resultando:

$$a_k = \frac{V_k c}{V_k + T_{OFF}^{(k)} c} \tag{2.52}$$

La tasa útil es aproximadamente la misma para todos los usuarios y cercana a la tasa media útil d (cociente del volumen medio y la duración media para todos los usuarios)

Cuando el tráfico ofrecido $\sum_{k=1}^{K} N_k a_k$, es menor que la capacidad del enlace C, la tasa útil media d tiende a ser igual a la tasa de acceso c.

Cuando el tráfico ofrecido $\sum_{k=1}^{K} N_k a_k$, es mayor que la capacidad del enlace C, el tráfico cursado $\sum_{k=1}^{K} N_k b_k$ tiende a C, con lo que $C \approx \sum_{k=1}^{K} \frac{N_k}{\frac{1}{a_k} + \frac{1}{d} - \frac{1}{c}}$

Otra aproximación interesante se presenta en [Adas, 1997], según la cual, los modelos ON-OFF utilizados hacen uso de distribuciones de Pareto para el modelado del estado ON, en general de media a_τ y varianza infinita, mientras que para el estado OFF se utiliza una distribución genérica de media a_θ. M fuentes independientes de este tipo presentarán una tasa de tráfico total Y_t:

$$E[Y_t] = MR\frac{a_\tau}{a_\tau + a_\theta} \tag{2.53}$$

siendo R la tasa media durante el estado ON.

De esta forma, el proceso agregado resultante es dependiente a largo plazo y asintóticamente autosimilar con parámetro de Hurst $H=(3-\beta)/2>0.5$, aunque el número de fuentes en cola puede ser obtenido mediante un sistema M/G/1. Conforme aumenta el número de fuentes agregadas el tráfico tiende a ser más

dependiente a largo plazo, tendiendo hacia el comportamiento de Ruido Gaussiano Fraccional.

2.3.2.2. **Modelos basados en Markov**

Teniendo en cuenta que un enlace de acceso es aquel, cuya capacidad puede ser utilizada íntegramente por una de sus fuentes, el *throughput* solamente se verá limitado por la capacidad *C* del enlace. Según[Bonald, 2000], la capacidad del mismo es dividida en función del número de fuentes activas (*1/n*). Este comportamiento puede ser modelado de forma simplista mediante un sistema M/G/1. Partiendo de la carga ofrecida de cada fuente, el número de fuentes con servicio sigue una distribución geométrica:

$$\forall k \geq 0, \quad \Pr[n=k] = \rho^k (1-\rho) \quad con\ \rho < 1 \tag{2.54}$$

Siendo:

$$E[n] = \frac{\rho}{1-\rho} \tag{2.55}$$

Sin embargo, en un enlace dorsal, los flujos presentan una limitación del *throughput*, normalmente constante ($r < C$) y relacionada con la capacidad del enlace de acceso, ya sea de origen o de destino. Cada flujo de los *n* activos reciben su tasa de acceso *r* o, como mínimo, *C/n*. Si el enlace presenta una capacidad múltiplo de las capacidades de los enlaces de acceso, el sistema puede modelarse como una cola M/G/m, con servidores $r=C/m$:

$$\Pr[n=k] = \begin{cases} (1-\rho)\dfrac{m!}{k!}(\rho m)^{k-m} f(m,\rho) & \text{para } k < m \\ (1\text{-}\rho)\,\rho^{k-m} f(m,\rho) & \text{para } k \geq m \end{cases} \tag{2.56}$$

Siendo

$$f(m,\rho) = \frac{\dfrac{(\rho m)^m}{m!}}{\dfrac{(\rho m)^m}{m!} + (1-\rho)\displaystyle\sum_{k=0}^{m-1}\frac{(\rho m)^k}{k!}} \tag{2.57}$$

En general, el uso de sistemas M/G/∞ permite modelar tráfico TCP con comportamientos muy cercanos al del tráfico *self-similar* real. Sin embargo, según [Paxson, 1995] hay que tener en cuenta, que si bien en tiempos de escala bajos el comportamiento de las fuentes modeladas mediante distribuciones de Pareto presentan un comportamiento claramente *self-similar*, la agregación de las mismas no se comporta de la misma manera, por lo que el uso del modelo M/G/∞ suele reservarse para tiempos de escala del orden de segundos e incluso minutos.

Asumiendo que TCP se comporta de forma ideal, un enlace que transporte exclusivamente flujos de tráfico elástico, puede ser modelado como una cola de procesamiento compartido del tipo M/G/R-PS, como la que se muestra en la Fig. 2.25, y que es utilizada intensivamente en [Riedl, 2000]. Este modelo

asume que la capacidad del enlace es compartida por todos los flujos activos en cada instante, comportamiento muy cercano al de TCP, en el que las diferentes conexiones comparten el mismo canal de forma concurrente. En cualquier caso, cada flujo está limitado a utilizar su tasa máxima r_{peak} dentro del enlace, el cual se comporta como un sistema de R servidores, siendo $R=C/r_{peak}$. Esto no significa que el enlace no pueda transportar más de R flujos, pero sí supondría reducir las capacidades individuales asignadas a cada flujo.

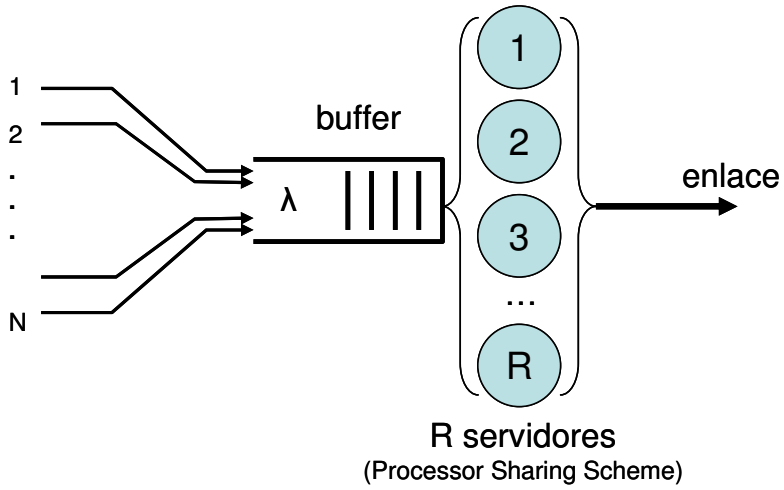


Fig. 2.25: Agregación de flujos de tráfico según el Modelo M/G/R-PS

Cuando la tasa máxima por flujo supera la capacidad del enlace, el comportamiento del sistema se simplifica, pasando de un modelo M/G/R-PS a otro más simple M/G/1-PS. De acuerdo con [Bartoli, 2001], el tiempo medio de transmisión de un objeto de tamaño V se obtiene como:

$$E[T]=\frac{V}{C(1-\rho)} \tag{2.58}$$

donde C es la capacidad del enlace y ρ la carga media ofrecida.

De la misma forma, se define el ancho de banda medio disponible como:

$$B=C(1-\rho) \tag{2.59}$$

La principal ventaja de este tipo de sistemas es que el tiempo medio en el sistema es independiente de la función de distribución del tiempo de servicio o del tamaño de los datos, de forma que la transmisión de grandes ficheros no penaliza a los ficheros de menor tamaño.

El *throughput* durante la transferencia de un fichero es:

$$D=\frac{r_{peak}}{\left(1+\frac{E_2(R,R\rho)}{R(1-\rho)}\right)} \tag{2.60}$$

donde $\rho = \lambda_e x_{mean} / C$ es el factor de utilización del enlace de un flujo de tasa de llegadas λ_e y tamaño medio de fichero x_{mean}, y E_2 es la segunda fórmula de Erlang (Erlang-C) con $A=R\rho$:

$$E_2(R,A) = \frac{\frac{A^R}{R!}\cdot\frac{R}{R-A}}{\sum_{l=0}^{R-1}\frac{A^l}{l!}+\frac{A^R}{R!}\cdot\frac{R}{R-A}} \tag{2.61}$$

Sin embargo, estos modelos sólo representan el efecto de un solo flujo de tráfico, no teniendo en cuenta la interacción de los diferentes flujos dentro del mismo enlace. Una alternativa sería considerar la adición de flujos, obteniendo el valor límite del tiempo de transmisión como la suma de los tiempos individuales correspondientes a cada uno de los flujos existentes. Este cálculo se corresponde con el peor de los casos, ya que es equivalente a un sistema *store&forward*, según el cual, cada flujo debe esperar a que finalice la transmisión de cualquiera de los que comparten el enlace. Este mecanismo de aditividad puede ser aplicado independientemente del modelo individual de fuentes, sin más que tener en cuenta:

Los requerimientos de QoS deben estar expresados en función del ancho de banda disponible para cada objeto *B*.

Como parámetro fundamental, el retardo extremo a extremo debe incluir, tanto los tiempos de transmisión como los de servicio, para lo que se considerará un tamaño arbitrario del objeto de forma que:

$$E[T] = \frac{L}{B} \tag{2.62}$$

La optimización se realizará sobre el retardo extremo a extremo para cada enlace, utilizando como función objetivo el valor medio del retardo *D*.
Cada enlace es dimensionado de acuerdo a sus requerimientos de retardo.

En el caso del tráfico de voz, incluido el de VoIP, el nivel de llamada puede ser modelado mediante una llegada con distribución de Poisson. Por su parte, a nivel de paquetes, se realiza el dimensionado en función de los requerimientos de QoS. Además, hay que tener en cuenta que, si bien las fuentes individuales dependen del tipo de codificación utilizada, la superposición de estas fuentes sigue una distribución de Poisson en función directa del número de fuentes agregadas. De acuerdo con ello, el modelo M/D/1 podría ser válido al nivel de paquetes. Sin embargo, cuando se considera la superposición de fuentes distintas, sólo es factible la consideración del caso general modelado mediante M/M/1.

La agregación de flujos heterogéneos es solucionada en [Bartoli, 2001] con el método que denomina "dimensionado por enlace". Considerando cada clase de servicio como flujos aislados unos de otros, y partiendo de valores conocidos de tráfico ofrecido (B_i), tamaño de paquetes (L_i), retardo medio D_i

(en el caso de tráfico de datos en general), o el retardo máximo aceptable (D_{max}) y el porcentaje de paquetes que superan ese valor (ε), el ancho de banda asociado a cada clase de servicio se calcula, en el caso de tráfico de datos en general, como:

$$c_i = B_i + \frac{L_i}{D_i} \tag{2.63}$$

o bien, en el caso de considerar el modelo M/M/1, como:

$$C_i = B_i + \frac{L_i}{D_{\max}} \cdot \ln\left(\frac{1}{1-\varepsilon}\right) \tag{2.64}$$

Y en el caso del modelo M/D/1:

$$C_i = B_i + \frac{L_i}{2D_{\max}} \cdot \ln\left(\frac{1}{1-\varepsilon}\right) \tag{2.65}$$

Para el cálculo de la capacidad total del enlace, habría que considerar el modo de agregación de cada clase de servicio. Así, por ejemplo, en el caso concreto de mecanismos *Diff-Serv*, como por ejemplo WFQ, la capacidad total del enlace queda:

$$C = \sum_i B_i + \frac{\sum_i w_i L_i}{\sum_i w_i D_i} \tag{2.66}$$

Tomando:

$$w_i = \frac{B_i}{\sum_i B_i} \tag{2.67}$$

En [Pandit, 2002] cada flujo, con retardos diferentes (aunque una vez agregados son tratados como uno solo), asume que los retardos individuales deben ser siempre iguales al menor de todos los flujos agregados. Según esto, su ancho de banda efectivo resulta:

$$r_D = \frac{B_D}{\min_i\{d_i\}} \tag{2.68}$$

Donde

$$B_D = \sum_i B_i \tag{2.69}$$

Además, el enlace puede ser considerado como un único elemento conformador, puesto que la suma de flujos resultante se comporta como uno solo. En el caso de modelar el enlace como un *token bucket* (B_S, r_S), el retardo mínimo será:

$$d_S = \frac{B_S}{r_D} = \frac{B_S}{B_D}\min_i\{d_i\} \le \min_i\{d_i\} \tag{2.70}$$

con lo que la capacidad equivalente del enlace podrá adaptarse a esta condición, siempre y cuando se asegure la suma de las tasas individuales:

$$r_S = \max\left\{\frac{B_S}{\min_i\{d_i\}}, \sum_i r_i\right\} = \max\left\{r_D - \left(\frac{B_D - B_S}{\min_i\{d_i\}}\right), \sum_i r_i\right\} \qquad (2.71)$$

Sin embargo, tal como se ha comentado para el caso del modelado de fuentes individuales, los modelos M/M/1 y M/D/1 difícilmente permiten modelar el comportamiento *self-similar* del tráfico Internet, pese a que la agregación de varios flujos suavice dicho comportamiento. Es por ello que, en[Ni, 1996], se hace uso de versiones moduladas, en concreto el modelo D-MMDP (*Discrete-time Markov Modulated Deterministic Process*). Éste permite modelar la agregación de fuentes ON-OFF, utilizando como proceso modulador otra cadena de Markov discreta en el tiempo. De esta manera, un sistema D-MMDP de (*M+1*) estados define un proceso de llegadas cuya tasa es controlada por una cadena de Markov discreta $\{\tilde{X}_n, n \geq 0\}$de M estados $R = \{R_0, R_1, ..., R_M\}$. Además, el sistema puede ser generado a partir de la superposición de M fuentes D-MMDP de dos estados (con tasas R_{ON} y R_{OFF}), resultando un sistema D-MMDP/D/1/K. La probabilidad de que existan *i* fuentes de dos estados en el estado ON es una binomial de la forma[13]:

$$\Pr\{\tilde{Y}_n = i\} = b\left(M, i, p_{ON}/\left(p_{ON} + p_{OFF}\right)\right) \qquad (2.72)$$

para *i* desde *0* hasta *M*.

Fijando la probabilidad de pérdida asociada a la agregación sobre un determinado enlace, la capacidad máxima del mismo se calcula como:

$$C_{\max} = \frac{\sum_{i=v}^{M} p_i\left(iR_{ON} + (M-i)R_{OFF}\right) - P_{LOSS}\sum_{i=0}^{M} p_i\left(iR_{ON} + (M-i)R_{OFF}\right)}{\sum_{i=v}^{M} p_i} \qquad (2.73)$$

con:

$$p_i = b\left(M, i, p_{ON}/\left(p_{ON} + p_{OFF}\right)\right) \qquad (2.74)$$

$$v = \min\{i, C_{\max}\} \qquad (2.75)$$

siendo $i = 0,1,...,M$ y

$$C_{\max} \leq \left(iR_{ON} + (M-i)R_{OFF}\right) \qquad (2.76)$$

---

[13] $b(n,k,\theta)$ es una función de agrupamiento de probabilidades binomial de la forma:

$$b(n,k,\theta) = \begin{cases} \binom{n}{k}\theta^k(1-\theta)^{n-k} & 0 \leq k \leq n \\ 0 & resto \end{cases}$$

Esta estimación puede ser utilizada frente a las tradicionales aproximaciones Gaussianas, de tasa máxima y de ancho de banda efectivo.

Como se expone en [Duffield, 1999], la aproximación Gaussiana asume que las tasas asociadas a fuentes individuales están distribuidas de forma normal e independiente. Partiendo del valor medio y la varianza de la tasa del tráfico agregado y una probabilidad de pérdida dada, el ancho de banda requerido se calcula como:

$$C = \mu + \sigma\sqrt{\left(-\ln\left(2\pi\right) - 2\ln P_l\right)} \tag{2.77}$$

De esta forma, la aproximación a partir de los máximos permite una estimación para tráfico agregado libre de pérdidas, muy similar a la obtenida mediante los modelos D-MMPP. Sin embargo, el uso de aproximaciones basadas en distribuciones normales, para el caso de agregación de tráfico, sólo puede ser aplicado a escalas temporales relativamente altas. Así por ejemplo, en [Norros, 1996] se indican las limitaciones asociadas a funciones normales, y en [Kilpi, 2002] se especifica el caso concreto de tráficos relativamente bajos, asociados normalmente a los canales *upstream*. Según esta fuente, para estos casos sería necesario hacer uso de distribuciones asimétricas, como por ejemplo la log-normal, mientras que para los canales *downstream* sería posible hacer uso de la aproximación gaussiana en la mayor parte de los casos.

2.3.2.3. **Modelos ON-OFF**

Cuando el tráfico que se desea modelar se corresponde a la agregación de varios flujos independientes, observado a la salida de cualquier elemento multiplexor, concentrador o de agregación, su comportamiento se traduce en una secuencia de ráfagas separadas por períodos de silencio. El modelado de la agregación, para este tipo de tráfico, pasa entonces por la necesidad de modelar la correlación de las sucesivas llegadas, siendo el modelo ON-OFF uno de los más firmes candidatos. Sin embargo, tanto los períodos de actividad/inactividad, como los sucesivos ciclos ON-OFF en realidad no están correlacionados. De hecho, este modelo solamente es aplicable cuando la autocorrelación sólo es observable dentro de cada ráfaga (la autocorrelación entre ráfagas diferentes es despreciable frente a la que se presenta dentro de cada ráfaga). Es por ello, que en [Hao, 1998] se modifica el modelo básico añadiendo varios estados ON y OFF, con tamaños de ráfaga determinados. Las transiciones entre cada par de estados ON y OFF están definidas en función del tamaño de la ráfaga, de forma que el modelo resultante es el de la Fig. 2.26.

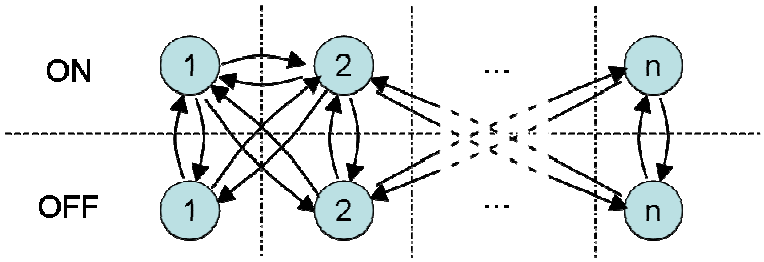


Fig. 2.26: Modelo ON-OFF multiestado para tráfico agregado

La dificultad de este modelo radica en la necesidad de determinar el número de estados ON/OFF necesarios, así como las probabilidades de transición resultantes, para lo cual es fundamental caracterizar completamente el tráfico agregado mediante la observación de trazas reales.

La complejidad de los modelos basados en múltiples estados adicionales llevó a [Taqqu, 1997] a proponer un modelo ON-OFF basado en dos niveles que, posteriormente en [Liu, 2004] es utilizado para el modelado de una sesión TCP como la de la Fig. 2.27. En el primer nivel, un proceso ON-OFF modela tanto la duración total de la sesión ($T_{sesión}$) y el tiempo entre sesiones (T_{is}). A su vez, dentro del $T_{sesión}$, otro proceso ON-OFF captura el comportamiento del protocolo TCP, modelando la duración de las ráfagas (T_{ON}) y el tiempo entre ráfagas consecutivas (T_{OFF}), siempre bajo el supuesto de que la tasa de paquetes es siempre constante y de valor *B* durante el estado de actividad. De acuerdo con este modelo, el usuario establece una conexión TCP de $T_{sesión}$ segundos de duración, para posteriormente estar T_{is} segundos inactivo hasta comenzar una nueva conexión TCP. Dentro de cada conexión, se suceden una serie de ráfagas, de duración T_{ON}, asumiendo que el RTT (Round Trip Time) es la suma de T_{ON} y T_{OFF}. Para ello se propone que los tiempos de actividad, $T_{sesión}$ y T_{ON} sean variables aleatorias con función de distribución *Pareto(K, a)* [14], mientras que los tiempos de inactividad (T_{is} y T_{OFF}) presentan funciones de distribución exponenciales, siendo todas ellas variables estadísticamente independientes unas de otras, tal como posteriormente se recoge en [Sheluhin, 2007]. Este modelo es una corrección de otro más sencillo presentado en [Deng, 1996] en el que las funciones de distribución de $T_{sesión}$ y T_{OFF} son aproximadamente Weibull, mientras que T_{is} sigue una Pareto. La mejora que presentan estos modelos, frente a los modelos ON-OFF comentados en el apartado 2.2.3.9 radica, precisamente, en que toman en consideración la fuerte dependencia del tráfico, motivada por la variación del RTT y del tiempo de actividad de las

[14] $\Pr[T > t] = \begin{cases} (K/t)^{\alpha} & ;t \geq K \\ 1 & ;0 < t < K \end{cases}$

ráfagas, pero sin embargo, quedan limitados al no capturar las interdependencias a largo plazo entre sesiones de un mismo usuario.

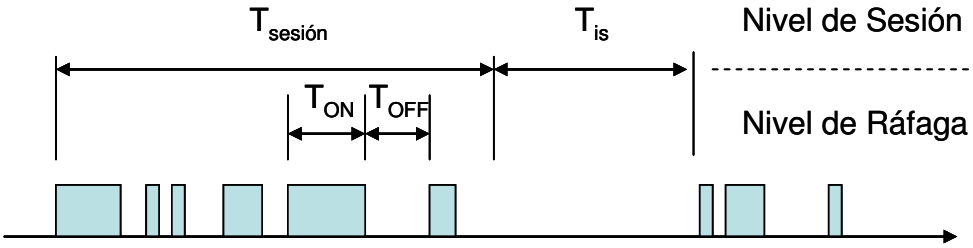


Fig. 2.27: Modelo de tráfico de una conexión TCP

Precisamente en [Rolland, 2006], la interdependencia entre sesiones es aproximada mediante la introducción de niveles adicionales, atendiendo a las características concretas de la aplicación Internet, utilizando como referencia el protocolo HTTP. Así, por ejemplo, definen cuatro niveles atendiendo a lo que denominan entidades de tráfico, distinguiendo entre sesión, página, ráfaga y paquetes. Un número de sesiones $N_{sesión}$ cada T_{is} segundos, hacen uso de los recursos de la red para acceder a varias páginas, N_{pagina}, separadas por el tiempo utilizado en su lectura, T_{OFF}. A su vez, cada página está compuesta por N_{objeto} objetos, siendo la diferencia de tiempos entre el primer objeto y los sucesivos IA_{obj}. Por último, cada objeto es transferido como secuencias de paquetes separados IA_{paq}. En la Fig. 2.28 se observan los cuatro niveles y sus respectivos parámetros característicos.

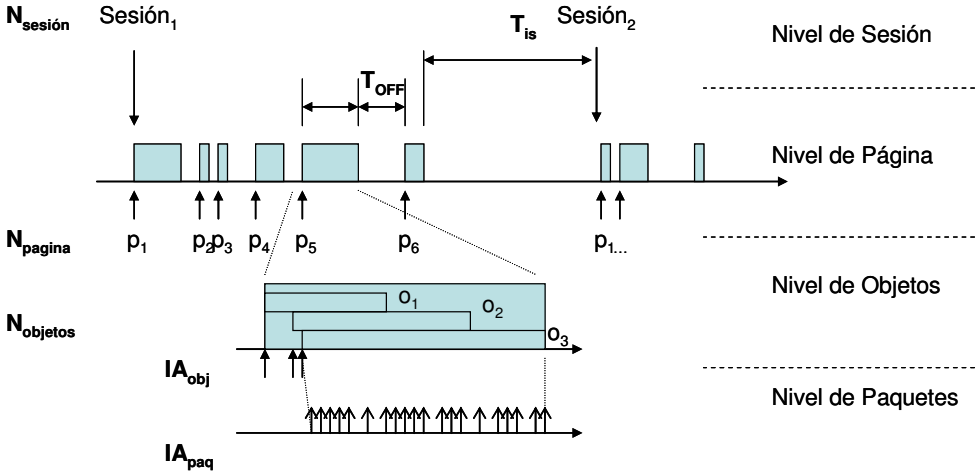


Fig. 2.28: Modelo multinivel de tráfico Web

El modelo se completa mediante la definición de funciones de distribución ajustadas a las variables aleatorias identificadas. Así, en el nivel de sesión, $N_{sesión}$ queda definida por una función log-normal y T_{is} por una exponencial. En el nivel de página, tanto N_{pagina} como T_{OFF} se ajustan a funciones de Pareto. En el

nivel de ráfaga, N_{objeto} es una Pareto e IA_{obj} una distribución de Weibull. Por su parte, la variable IA_{paq} queda modelada por una distribución log-normal.

El uso de modelos basados en el comportamiento del tráfico a diferentes niveles temporales ha ido tomando fuerza, siendo una de las mejores opciones para el modelado de tráfico Internet, ya que permiten definir las interrelaciones entre los diferentes niveles mediante los cuales capturar, no solo el comportamiento autosimilar del tráfico a ráfagas, sino también el efecto de los mecanismos de control de tráfico y de provisión de QoS/GoS. Precisamente, la posibilidad de capturar el comportamiento general del tráfico agregado, es la razón por la cual en esta Tesis se plantea un nuevo modelo de tráfico, presentado y desarrollado en el capítulo siguiente, basado en fuentes ON-OFF moduladas a diferentes niveles temporales, de forma similar a los modelos presentados anteriormente.

2.4. Network Calculus

La teoría del *Network Calculus* es un reciente desarrollo que proporciona una visión profunda de los problemas relacionados con el flujo de tráfico, que aparecen cuando se interconectan varias redes de ordenadores. La principal diferencia con respecto a la teoría de flujos en las redes tradicionales estriba en la definición de un nuevo álgebra, por la que, entre otras cosas, la suma pasa a transformarse en un cómputo de mínimos, y el producto pasa a convertirse en suma, véase [Boudec, 2002]. Esta metodología ha sido aplicada con éxito en temas tan dispares como el "suavizado" de vídeo, el control de ventana de flujo, y por supuesto, problemas de *schedulling* o de análisis del retardo en redes de paquetes, ver [Altman, 2002 / Malaney, 1999]. Sin embargo, dicho concepto no sólo abarca los problemas de tráfico propiamente dicho, sino que también encuentra su aplicación en la teoría de grafos para la resolución de problemas de diseño y dimensionado de topologías de red, véase [Starobinski, 2003].

Esta última aplicación es la que se toma en consideración en esta Tesis, la aplicación de una aritmética específica para obtener, de forma simple, estimaciones fiables del tráfico de fuente correspondiente al comportamiento natural de los usuarios, analizar la agregación de flujos procedentes a diferentes servicios y usuarios, y proyectar dichos resultados a lo largo de los diferentes elementos, desde la red de agregación, hasta la dorsal. Esta metodología no se toma en consideración de forma absoluta, sino que será aplicada en función de las características del estudio realizado. En principio, la utilidad del concepto *Network Calculus* se revela interesante para el análisis del tráfico a partir de trazas reales, la validación de trazas simuladas, y el uso combinado con otros mecanismos de modelado.

2.4.1. Definición y antecedentes

Las aproximaciones analíticas obtienen resultados con tendencia hacia la linealidad, con simplificaciones normalmente representadas por figuras de tráfico muy simples. La teoría del *Network Calculus* utiliza precisamente la simplificación de las figuras de tráfico, aunque con soluciones más generalistas, definiendo a partir de flujos reales, figuras límite, representadas como curvas de llegadas y curvas de servicio. La aplicación de dichas curvas no es sino una aproximación genérica al comportamiento real de fuentes individuales sobre diferentes elementos de la red, véase [Liebeherr, 2001]. De hecho se distinguen dos variantes de *Network Calculus* claramente diferenciadas:

- Determinista: Solamente considera aquellas figuras límite correspondientes a los casos más pesimistas, sin tener en cuenta las ventajas asociadas a, por ejemplo, la multiplexación estadística de varios flujos de tráfico sobre un único enlace.
- Estadístico: Tiene en consideración las características específicas de Garantía de Servicio para la agregación de los flujos (curvas de servicio deterministas), y para cada flujo individualmente (curvas de servicio efectivas).

Una curva de servicio efectiva representa el límite más probable al que tiende un servicio correspondiente a un determinado flujo de tráfico, definición análoga a la utilizada para los conceptos de capacidad y/o ancho de banda efectivo expuesto en el apartado 2.2.4. El uso de estas curvas permite establecer tres posibles aproximaciones para la estimación de los requerimientos de Garantía de Servicio:

- Estimación de tasa de bit máxima: Si la curva de servicio *j* presenta la forma $S_j(t) = P_j t$, siendo P_j la tasa de bit máxima para el flujo *j*, esta curva obtiene el límite máximo correspondiente a los recursos utilizados por cada flujo *j*, garantizando completamente los requerimientos de retardo y de pérdida de paquetes, véase [Ishii, 2006].
- Estimación de tasa de bit media: Si la curva de servicio es de la forma $S_j(t) = \rho_j t$, siendo ρ_j la tasa media para el flujo *j*, es posible obtener un límite mínimo de los recursos utilizados por cada flujo *j*. Por ejemplo, el modelo LBAP (*Linear Bounded Arrival Processes*), véase [Garroppo, 2001], considera cada fuente de tráfico como un token bucket (b,ρ), siendo *b* su capacidad y ρ la tasa de bit, con curva de llegada $A(t) \le b + \rho t \qquad \forall t > 0$
- Estimación determinista: Este método considera la mejor curva de servicio de acuerdo a la reserva de recursos para cada flujo *j,* asegurando las condiciones de retardo extremo a extremo.

Una de las bases del *network calculus* es la simplificación e idealización de los sistemas y elementos de red, caracterizados por una curva de servicio mediante la cual el flujo de entrada es acomodado a un determinado flujo de salida. El caso más sencillo se denomina conformador. Básicamente, el elemento de red acomoda cualquier señal de entrada *x(t)* a un determinado flujo de salida *y(t)* de

acuerdo a los límites de una función σ (curva de servicio, también denominada curva de conformado) aún a costa de introducir retardo. Si los flujos de entrada y salida son representados mediante sus correspondientes funciones acumulativas, es decir, la cantidad de información total para cada señal observada en un intervalo [0, *t*], *x* e *y* pueden ser relacionadas mediante la siguiente expresión:

$$y(t) = (\sigma \otimes x)(t) = \inf_{s:0 \le s \le t} \{\sigma(t-s) + x(s)\} \tag{2.78}$$

Esta definición tiene una gran similitud con la del operador convolución tradicional, sin más que sustituir el operador inf{} por la integral. De hecho, el *Network Calculus* denomina a esta operación "convolución min-plus $\otimes$".

En la Fig. 2.29 se observa como la curva *y(t)* representa el límite inferior de las posibles curvas de salida, tras atravesar la curva *x(t)* el elemento con curva de servicio σ.

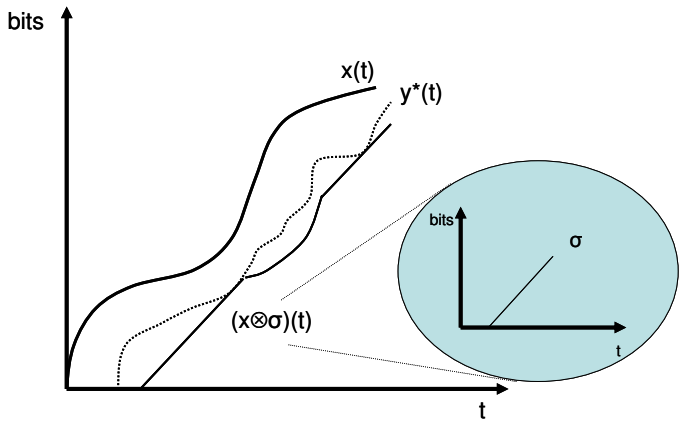


Fig. 2.29: Relación entre la curva de llegadas x(t), la curva de servicio σ, la curva de salida real y*(t) y la curva de salida límite (x⊗σ)(t)

De acuerdo con las consideraciones anteriores, observando simultáneamente tanto la entrada como la salida del sistema, podrá existir un número dado de bits que aún se encuentren en el interior del mismo (por ejemplo, en el caso de un sistema de cola, la longitud de la misma). En general, el número de bits en tránsito dentro del sistema en cuestión recibe el nombre de *backlog*[15], y es igual a *B(t) = y(t)-x(t)*. A su vez, un determinado bit sufrirá un retardo desde que

[15] Backlog: es la cantidad de paquetes sin transmitir, pero almacenados en el buffer. En el caso de sistemas conservativos la transmisión de paquetes se lleva a cabo mientras el backlog sea positivo.

entra al sistema, hasta que todos los bits precedentes sean servidos y pueda salir del mismo. A este retardo se le denomina retardo virtual y puede ser expresado como:

$$d(t) = \inf\{\tau \geq 0 : x(t) \leq y(t+\tau)\} \tag{2.79}$$

De esta manera, la llegada de paquetes en un intervalo $[t_1, t_2]$ puede ser acotada por una función no decreciente, de forma que $x(t_2) - x(t_1) \leq A(t_2 - t_1)$, expresión denominada "restricción de ráfaga" (*burstiness constraint*). En este caso se dice que *x* está "suavizada por *A*", donde *A(t)* es la "curva de llegada".

2.4.2. Aplicación al modelado de fuentes

Los parámetros básicos de un flujo de tráfico condicionan la forma de la curva de llegada correspondiente, de tal forma que una lectura detenida de las características de la curva de llegada también permite determinar los valores límite de dichos parámetros, sin más que observar puntos singulares de la curva, tal como se puede apreciar en la Fig. 2.30.

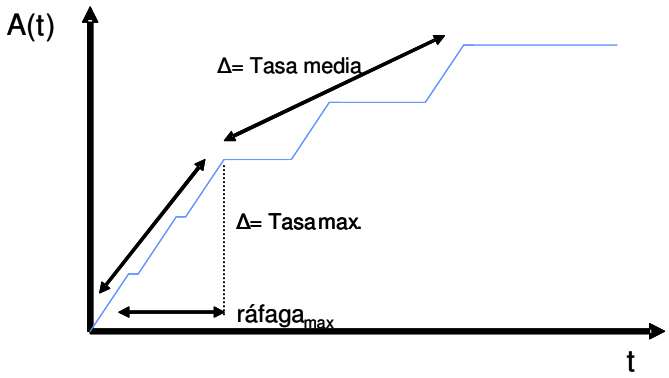


Fig. 2.30: Ejemplo de Curva de Llegada genérica y su relación con las características propias del flujo de entrada

Aunque la curva de llegadas más simple sería *A(t)=rt*, la cual describe un flujo ajustado a una tasa máxima *r*, y que normalmente recibe el calificativo de flujo de tasa constante (CBR – *Constant Bit Rate*), la curva por excelencia es $A(t)=\sigma+\rho t$, en el que σ representa la máxima cantidad de tráfico asociado a una ráfaga, y ρ es el límite superior de la tasa media a largo plazo. Según esta curva, una fuente puede enviar σ bits de una vez, pero no más de ρ bits/seg en un período de tiempo continuado. Suele utilizarse la notación $A \approx (\sigma, \rho)$, y en este caso se dice que el flujo de entrada está suavizado por (σ, ρ), tal y como se describe en [Cruz, 1995]. Por su parte en [Cruz, 1991] se dice que el flujo de entrada está "regulado", cuando su envolvente, para todo Δt puede ser aproximada por:

$$\int_{\Delta t} x(t)dt \le \sigma + \rho(\Delta t) \tag{2.80}$$

El tamaño del *backlog* en un sistema conservativo no puede superar en ningún momento el valor de σ. En el caso de dos flujos regulados, agregados en un sistema conservativo de capacidad de transmisión *C*, el *backlog* de dicho sistema no será superior a la suma de los valores máximos de las ráfagas individuales.

En general, cualquier sistema conservativo, dado un retardo máximo asociado *D*, permite establecer la siguiente relación entre las tasas de un determinado flujo de entrada regulado R_{in} y su correspondiente salida regulada R_{out}:

$$\forall \quad R_{in}(\sigma,\rho) \quad y \quad \mathrm{D} < \infty \quad \Rightarrow R_{out}(\sigma + \rho D, \rho) \tag{2.81}$$

Para el caso de una línea que introduzca un retardo constante (como máximo *D*), su *backlog* estará limitado por $\sigma + \rho D$, y $R_{out} = R_{in}$.

En el caso de un *buffer*, éste puede ser considerado como un sistema con tasa de transmisión finita a su entrada, e infinita a su salida. La salida del *buffer* se produce una vez el paquete ha llegado completamente, por lo que introduce un retardo máximo de *L*/*C*, siendo *L* el límite del *backlog*:

$$R_{in}(\sigma,\rho) \Rightarrow R_{out}(\sigma + \rho L / C, \rho) \tag{2.82}$$

En el caso de un demultiplexor, un único enlace debe ser dividido en múltiples enlaces de salida, aunque no se considerará este retardo. En el caso del multiplexor el procedimiento es el inverso.

En el caso concreto de los Servicios Integrados definidos por el IETF (RFC 1633) se hace uso de las curvas T-Spec, de la forma $A(t)=\min\{M+pt,\ rt+b\}$. En este caso *M* representa el tamaño máximo de paquete, *p* la tasa máxima, *b* la tolerancia de ráfaga y *r* la tasa sostenible. Hay que tener en cuenta que tanto la tasa máxima, como la sostenible están reguladas, y que en el caso de que sólo estuviera regulada la tasa media, *p* tendería a infinito.

Además con el álgebra Min-Plus se define el operador deconvolución como:

$$(x \oslash \sigma)(t) = \sup_{u \ge 0}\{x(t+u) - \sigma(u)\} \tag{2.83}$$

que permite realizar el cálculo de la curva de llegada mínima correspondiente a un determinado flujo como $(x \oslash x)(t)$.

Si se considera un elemento concreto de una red, cualquier flujo que lo atraviese caracterizado por una función R_{in} a su entrada, y la correspondiente función R_{out} a su salida, asumiendo que inicialmente el elemento de red está vacío, el número de paquetes que se encuentran aún en el mismo al finalizar el slot *t* viene dado por la función $B(t)= R_{in}(t)-R_{out}(t)>=0$. A su vez, R_{in} y R_{out} están relacionados mediante $R_{in}(t)=R_{out}(t+d(t))$ para todo *t*, donde *d(t)* es el retardo asociado a un paquete en un tiempo dado *t*. La relación entre R_{in} y R_{out} permite definir las llamadas "curvas de servicio", de forma que un enlace/sistema ofrece una curva de servicio *S(t)* a un determinado flujo de entrada, si para

cualquier *t*, existe un $s \leq t$, tal que el *backlog* en *s* sea cero ($B(s)=0$), y que $R_{out}(t) \geq S(t-s)$.

En términos de la aritmética *Min-plus*, puede expresarse como:

$$y(t) \geq \inf_{0 \leq s \leq t} \{S(t-s) + x(s)\} = (S \otimes x)(t) \tag{2.84}$$

Una curva de servicio del tipo $S(t) = R \cdot t$ va a garantizar que cada flujo sea servido a una tasa mínima de *R* bits/seg durante el período ocupado, y normalmente se corresponde con sistemas con política de servicio GPS (*Generalized Processor Sharing*). En general, una curva de servicio de la forma $S(t) = [-\sigma+\rho t]^+$ con σ y $\rho >= 0$, garantiza un servicio (σ, ρ). Si dicha política se corresponde con una cola FIFO, esta función representa la curva de servicio del tráfico agregado. Este tipo de curvas son las que el IETF define para servicios basados en RSVP, y reciben el nombre de curvas *rate-latency*.

Al considerar el caso típico de flujos VBR, tal como se presenta en [Boudec, 2000a / Boudec, 2000b], modelados mediante curvas T-Spec(*M, p, r, b*), que atraviesan un nodo con curva de servicio *rate-latency* $R[t-T]^+$, donde *R* es, al menos, del orden de la tasa sostenible *r*, la máxima cantidad de datos procedentes de un determinado flujo que existirán en un momento dado en su interior estará limitada por:

$$w_{\max} = b + r \cdot \max\left(\frac{b-M}{p-r}, T\right) \tag{2.85}$$

Y el retardo que sufrirán dichos datos será como máximo:

$$d_{\max} = \frac{M + \frac{b-M}{p-r}(p-R)^+}{R} + T \tag{2.86}$$

Gráficamente, dichos valores vienen representados por las distancias máximas entre las curvas de llegada y de servicio en los ejes Y y X respectivamente, como se indica en la Fig. 2.31.

A su vez, el álgebra *Min-Plus* permite obtener la curva a la salida del elemento de red como la curva de llegada del flujo saliente correspondiente, en función de las respectivas curvas de entrada y de servicio, esto es:

$$\sigma^+ = \sigma \oslash \mathrm{S} \tag{2.87}$$

El límite del *backlog* según [Liebeherr, 2001] será entonces:

$$\mathrm{b}_{\max} = \sigma^+ \oslash \mathrm{S}(0) \tag{2.88}$$

Y el del retardo:

$$d_{\max} = \inf\left\{d \geq 0 \middle| \forall t \geq 0 : \sigma^+(t-d) \leq S(t)\right\} \tag{2.89}$$

En general, cualquier flujo de tráfico arbitrario puede ser transformado en su variante "suavizada" mediante la introducción de mecanismos de control de tráfico, como por ejemplo, los denominados "reguladores". Estos dispositivos tienen por función almacenar paquetes el tiempo necesario para obtener una

R_{out} "suavizada" por (σ, ρ), donde $\sigma \geq 1$ y $0<\rho \leq 1$. La importancia de este tipo de elementos es tal, que incluso en [Chang, 1998] se relaciona con el concepto de *leaky bucket* y se establece el procedimiento de cálculo de los parámetros del mismo para un determinado flujo, considerado éste como el regulador "máximo" para dicho flujo.

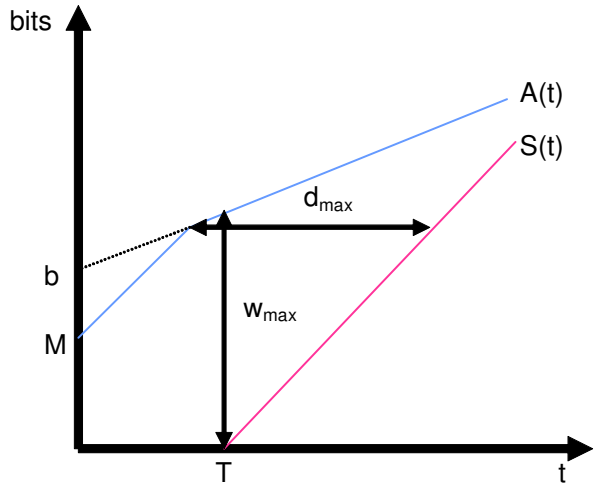


Fig. 2.31: Estimación del retardo y *buffer* mediante curvas de llegadas y de servicio

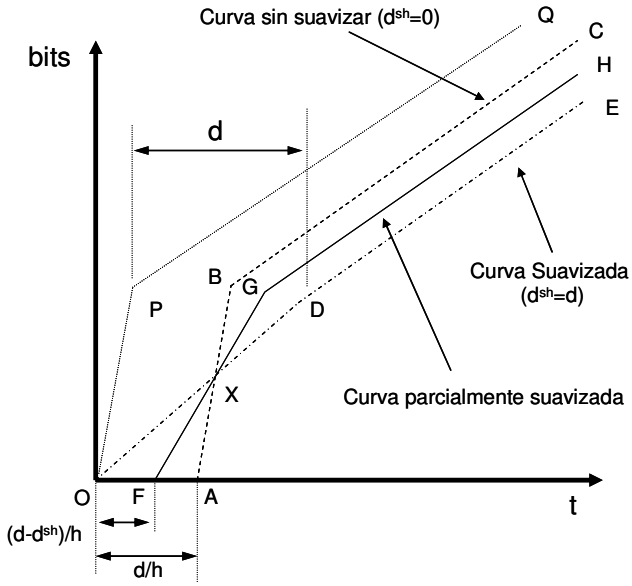


Fig. 2.32: Efecto del "suavizado" en función del retardo introducido por un elemento de red a un flujo de tráfico que lo atraviesa

Precisamente en[Sivaraman, 2000], se relaciona con mayor detalle la relación entre las curvas TSpec y su aplicación en el *Network Calculus*. Para ello, se amplía la idea del "suavizado", tomando no un *leaky bucket*, sino que la envolvente de un flujo de llegada a lo largo de una determinada ruta puede ser representada mediante un doble *leaky bucket* (OPQ) con parámetros fundamentales tanto el límite de retardo *d* como el número de saltos *h*, tal y como se observa en la Fig. 2.32.

Según esto, un elemento de la red que actúe como conformador (*shaper*), introduce un retardo d^h, que puede ser *0* , con lo que *A(t)* solamente es desplazado *d/h* (determinado por los puntos ABC). Por su parte, un elemento *scheduler* no introduce retardos (puntos ODE), mientras que en el otro extremo se produce el *smoothing* completo, en el que se añade un *buffer* que permite introducir un retardo de $d^h=d$.
El doble *leaky bucket* equivalente, que satisface este requerimiento de retardo, se aproxima por (p',σ',ρ) (puntos FGH):

$$p' = \frac{p}{1 + d^{sh}\left(p - \rho\right)/\sigma'} \tag{2.90}$$

$$\sigma' = \sigma \frac{\left(1 - \rho/p'\right)}{\left(1 - \rho/p\right)} \tag{2.91}$$

tomando como límite

$$d^{sh} = \min\left\{ d\left(1 - \frac{1}{h}\right), \frac{\sigma}{\rho} \right\} \tag{2.92}$$

La aplicación más interesante es su uso para el modelado de todos aquellos servicios que pueden ser regulados o conformados ya en la red de acceso. Según esta teoría, dado un flujo de tráfico reglado, éste puede ser cuantificado en su desplazamiento a través de la red elemento a elemento. Para ello se hace uso de las denominadas "curvas de llegadas", las cuales describen los flujos de entrada a un determinado elemento de la red, de tal forma que un flujo *x(t)* está representado por una curva de llegadas *A(t)*, solo si para todo momento $s \leq t$ se cumple que $x(t)-x(s) \leq A(t-s)$, o de forma equivalente:

$$x(t) \leq \inf_{0 \leq s \leq t} \left\{ A\left(t - s\right) + x\left(s\right) \right\} \tag{2.93}$$

En el caso de que el elemento de red soporte hasta *N* flujos de tráfico distintos, se obtiene la tasa de llegadas agregada como la suma de los parámetros de tasa y de ráfaga:

$$A(t) = \min\left(M_{\max} + \sum_{i=1}^{N} p_i t\,, \quad \sum_{i=1}^{N} r_i t + \sum_{i=1}^{N} b_i \right) \tag{2.94}$$

siendo

$$M_{\max} = \max_{i=1..N} \left\{ M_i \right\} \tag{2.95}$$

2.4.3. **Aplicación a la agregación de tráfico**

El comportamiento del tráfico no es independiente del número de fuentes, y en la mayoría de los casos, el problema se traslada a la obtención de una función límite que determine los requerimientos necesarios para garantizar la QoS del conjunto de flujos de tráfico. Para ello es posible utilizar los conceptos del *Network Calculus* para la sub-aditividad. Así, en [Boorstyn, 2000] se justifica el cálculo de funciones límite, tanto individuales como la agregación de varias de estas funciones. De acuerdo con este estudio, un flujo de tráfico cualquiera puede ser representado mediante una función determinista basada en valores límite del mismo en determinados intervalos, que además es sub-aditiva, esto es, el comportamiento general de la fuente siempre tendrá como límite la suma de las funciones correspondientes a cada intervalo:
Siendo $A_j(t,t+\tau)$ la función de llegada del flujo *j* en el intervalo *[t, t+τ]*, su función límite será A_j^* tal que

$$A_j(t,t+\tau) \le A_j^* \tag{2.96}$$

Que además es sub-aditiva, de forma que

$$A_j^*(\tau_1+\tau_2) \le A_j^*(\tau_1) + A_j^*(\tau_2) \qquad \forall \tau_1,\tau_2 \ge 0 \tag{2.97}$$

siempre bajo los supuestos de que la fuente cumple los requisitos de estacionariedad e independencia. En el caso de varias fuentes *N* con límites determinísticos A_j^* para cada $j \in N$, el flujo agregado presenta una función determinista límite H_N en todo *t* de forma que:

$$P\left[A_N(t,t+\tau) \le H_N(\tau,\varepsilon)\right] \ge 1-\varepsilon \tag{2.98}$$

Siendo:

$$A_N(t,t+\tau) = \sum_{j=1}^{N} A_j(t,t+\tau) \tag{2.99}$$

El método de cálculo de H_N es importante, pudiendo calcularse en momentos discretos (por ejemplo a partir de $A_j(0,\tau)$ y $A_N(0,\tau)$, utilizar la función de Chernoff) La función límite para una variable aleatoria *Y* puede ser aproximada por:

$$P[Y \ge y] \le e^{-sj} E\left[e^{sY}\right] \qquad \forall s \ge 0 \tag{2.100}$$

o bien utilizar soluciones geométricas.

Por su parte, atendiendo a la naturaleza de los elementos de red, los reguladores de tráfico típicos consisten en *leaky buckets* con limitadores de tasa máxima, con lo que el tráfico de cada flujo *j* puede ser definido en función de sus valores máximos y medios de tasa de bits y por el tamaño medio de la ráfaga (P_j, ρ_j, σ_j). En [Oottamakom, 2001] se resuelve el cálculo de su envolvente como:

$$A_j^*(\tau) = \min\{P_j\tau; \sigma_j + \rho_j\tau\} \quad \forall \tau \geq 0 \quad y \quad P_j \geq \rho_j \qquad (2.101)$$

En general, para *K* doble *leaky buckets*, la envolvente del flujo *j* puede aproximarse por:

$$A_j^*(\tau) = \min_{k=1..K}\{\sigma_{jk} + \rho_{jk}\tau\} \qquad (2.102)$$

Sin embargo, el uso del *Network Calculus* no significa el rechazo de los modelos de Markov y similares. Así por ejemplo, en [Lombardo, 2004] se defiende el uso conjunto de modelos de Markov y Curvas de Servicio. Según los autores, su sistema permite modelar el comportamiento de *routers* individuales, pero también indican su aplicación para el modelado de la transmisión de múltiples flujos de paquetes IP a través de un trayecto completo, constituido por múltiples *routers* y enlaces. El conjunto de paquetes que es reenviado y procesado de la misma forma y a través del mismo camino recibe el nombre de FEC (*Forwarding Equivalence Class*), y el conjunto de *routers* que atraviesa un FEC se denomina LSP (*Label Switched Path*), tal como se define en MPLS (*Multiprotocol Label Switching*). El sistema resultante puede ser obtenido a partir de un modelo SBBP equivalente, en el que la curva de servicio total puede ser calculada aplicando la aritmética definida por el *Network Calculus*, como la convolución de todas las curvas de servicio asociadas a cada *router* del FEC. Incluso en [Starobinski, 2000] se simplifica aún más el cálculo haciendo uso de sumas de exponenciales como aproximación de límites de funciones de distribución más complejas o indefinidas. En realidad, los autores lo presentan como una mejora al método EBB (*Exponentially Bounded Burstiness*), según el cual un flujo de tráfico de tasa instantánea *R(t)* está limitado por una tasa máxima ρ y una función límite $Ae^{-a\sigma}$ de forma que para todo $\sigma \geq 0$ y $t > s \geq 0$

$$\Pr\left\{\int_s^t R(u)\,du \geq \rho(t-s) + \sigma\right\} \leq Ae^{-\alpha\sigma} \qquad (2.103)$$

Permitiendo el modelado del *buffer* del enlace mediante exponenciales. El método presentado se denomina SBB (*Stochastically Bounded Burstiness*, y se define más abajo) modelando la función límite mediante una suma de exponenciales, definiendo además un *Network Cálculus* específico para la obtención de las curvas SBB.

Pero quizá la mayor aplicación del *Network Calculus* estriba en la posibilidad de establecer y calcular límites bien definidos para el retardo extremo a extremo. Esta característica hace que, mediante expresiones más o menos sencillas, se pueda caracterizar a la mayor parte de los tráficos de fuente relacionados con las redes Internet, y que el IETF lo haya modelado mediante curvas *TSpec*, tal como se ha comentado en el apartado 2.2.3.11.

Por su parte en [Schmitt, 1999] y [Schmitt, 2001] se da una solución para el dimensionado de acuerdo con la agregación de varios flujos de tráfico. Así, al considerar la agrupación de varias fuentes *TSpec*, de acuerdo con las RFC 2212

y RFC 2216, la suma de *n* flujos *TSpec* genera otro flujo *TSpec* , denominado *Summed-TSpec* de la forma:

$$\sum_{i=1}^{n} TSpec\left(r_i, b_i, p_i, M_i\right) = TSpec\left(\sum_{i=1}^{n} r_i, \sum_{i=1}^{n} b_i, \sum_{i=1}^{n} p_i, \max\left(M_i\right)\right) \quad (2.104)$$

Esta función representa la suma de *N* flujos de entrada, y permite dimensionar el sistema para realizar una reserva compartida por los *N* diferentes flujos, siendo deseable aunque no imprescindible que la función obtenga la menor suma posible.

En este caso la reserva de ancho de banda *R* se calcula como:

$$R = \begin{cases} \dfrac{\sum_{j=1}^{n} p_j \dfrac{\sum_{J=1}^{n} b_j - M}{\sum_{j=1}^{n} p_j - \sum_{j=1}^{n} r_j} + M + C}{d_{\max} + \dfrac{\sum_{j=1}^{n} b_j - M}{\sum_{j=1}^{n} p_j - \sum_{j=1}^{n} r_j} - D} & para \quad \sum_{j=1}^{n} p_j \geq R \geq \sum_{j=1}^{n} r_j \\ \dfrac{M + C}{d_{\max} - D} & para \quad R \geq \sum_{j=1}^{n} p_j \end{cases} \quad (2.105)$$

Por otra parte, de acuerdo con la teoría del *Network Cálculus*, se puede obtener la suma exacta de cada uno de los flujos representados por cada curva *TSpec*, que se demuestra será inferior o igual a la *TSpec Suma*. Para ello se obtiene la curva de llegada para el conjunto de *n* flujos, como la concatenación de (*n*+1) *token buckets* (operación ⊗ del *Network Calculus*) denominada *Cascaded-TSpec*. Recalculando las expresiones anteriores:

$$R = \begin{cases} \dfrac{\sum_{l=1}^{k-1}\left(b_l - M_l\right) + M + \left(\sum_{j=k}^{n} p_j + \sum_{l=1}^{k-1} r_l\right)\left(\dfrac{b_k - M_k}{p_k - r_k}\right) + C}{d_{\max} + \dfrac{b_k - M_k}{p_k - r_k} - D} & para \quad \sum_{j=k}^{n} p_j + \sum_{l=1}^{k-1} r_l \geq R \geq \sum_{j=k+1}^{n} p_j + \sum_{l=1}^{k} r_l \\ \dfrac{M + C}{d_{\max} - D} & para \quad R \geq \sum_{j=1}^{n} p_j \end{cases}$$

(2.106)

Con lo que, en definitiva, la curva *Cascaded-TSpec* es otra *TSpec* de la forma:

$$TSpec\left(\sum_{j=k+1}^{n} p_j + \sum_{l=1}^{k} r_l, \sum_{l=1}^{k}\left(b_l - M_l\right) + M, \sum_{j=k}^{n} p_j + \sum_{l=1}^{k-1} r_l, \sum_{l=1}^{k-1}\left(b_l - M_l\right) + M\right) \quad (2.107)$$

En la Fig. 2.33 se puede observar la relación existente entre las estimaciones realizadas mediante las curvas agregadas *Summed-TSpec* y *Cascaded-TSpec* y el tráfico total agregado.

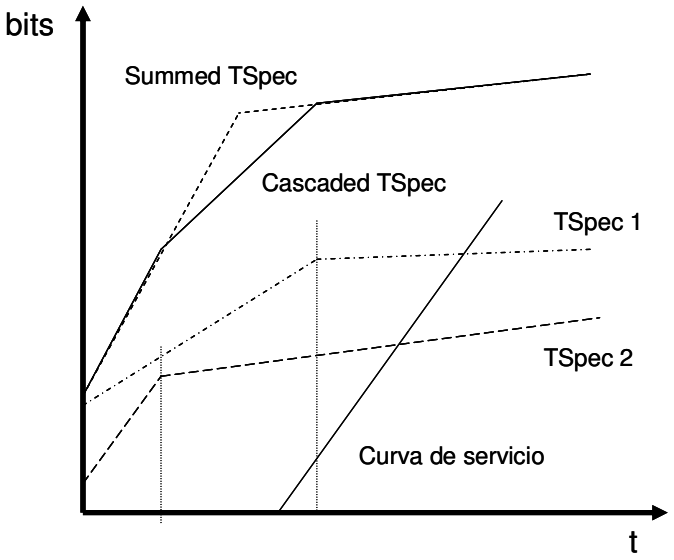


Fig. 2.33: Comparación de curvas de tráfico agregado calculadas a partir de curvas de llegada *TSpec*

2.5. Conclusiones

Planificación, modelado, agregación y *Network Calculus* son palabras clave que definen a grandes rasgos el contenido de este documento. En este capítulo se ha hecho un ineludible repaso sobre el significado de estos conceptos, sus definiciones, variaciones, virtudes y aplicaciones.

El concepto de planificación, aplicado a las redes de comunicación, ha sido tratado desde el punto de vista de su desarrollo y ejecución. En la primera parte de este capítulo se define y describe la estructura del proceso de planificación para, a partir de los conceptos fundamentales, justificar la necesidad de desarrollar e implementar herramientas específicas, adaptadas a problemáticas concretas. Cada caso requiere su propia solución y la planificación estratégica necesita adelantarse a cualquier solución mediante la estimación y predicción basada en el modelado, la observación y la simulación.

En esta Tesis se abordan los dos primeros temas por lo que, en la segunda parte de este capítulo, se ha descrito lo que es un modelo de tráfico, y cómo se desarrolla y aplica, hasta llegar al caso concreto de los modelos de tráfico de fuente más importantes. Se hace especial mención a los modelos autosimilares y a los modelos ON-OFF, por su adaptación a las características del tráfico de Internet. De la misma forma, en la tercera parte de este capítulo, se exponen las adaptaciones de los modelos de fuente única para la resolución del

problema de la estimación del tráfico multi-fuente. En el siguiente capítulo, se volverá a incidir en el tema al proponerse un nuevo modelo de agregación basado en fuentes ON-OFF y funciones de distribución binomiales, basados en todos los conceptos anteriores.

Para finalizar, se ha expuesto la teoría del *Network Calculus*, y concretamente, su aplicación en la estimación de los parámetros asociados a los flujos de paquetes, como es el caso del tráfico en Internet. Las ideas y soluciones propuestas por esta teoría han sido utilizadas en este estudio, y finalmente se han aplicado para completar los mecanismos de estimación de tráfico explicados aquí.

3. **Aplicación práctica para la planificación de tráfico Internet**

El incremento en la demanda de tráfico de banda ancha se debe, fundamentalmente, a la introducción de accesos de banda ancha "real" en el segmento residencial y PYME, al uso de aplicaciones específicas con cada vez mayores requerimientos de ancho de banda (como por ejemplo las aplicaciones P2P) a los propios servicios emergentes, relacionados tanto con el comercio electrónico, como a los servicios de difusión/distribución de TV, audio y vídeo, y al uso generalizado del acceso *wireless*. El diseño y dimensionado de la red se convierte en algo "vivo" que los operadores afrontan introduciendo nuevas herramientas y mecanismos, los cuales, faciliten la adaptación de sus actuales infraestructuras, de una forma lo más dinámica posible y con el menor esfuerzo tecnológico y económico.

Los modelos presentados en el capítulo 2 permiten implementar herramientas utilizadas, recursivamente, en cada una de las evoluciones sufridas por la red Internet. Se basan en algoritmos y fundamentos que obtienen soluciones, en el peor de los casos, muy cercanas a la exactitud. Precisamente, ésta es la que cobra un elevado precio en forma de recursos hardware / software, y tiempo de cálculo, lo que en muchos casos les hace inviables para el estudio y modelado de redes a nivel nacional. En estos casos, se plantea la necesidad de utilizar modelos y algoritmos más sencillos que, a costa de perder precisión, obtienen soluciones en tiempos razonables, o bien dentro de los límites

asociados a las diferentes figuras de tráfico. Estos valores son suficientes para llevar a cabo estudios previos y estimaciones del comportamiento de tráfico en redes relativamente grandes.

Este es el caso del modelo ON-OFF multinivel para la estimación de tráfico de fuentes IP y su agregación basada en el *Network Calculus*, estudiadas a lo largo de toda esta Tesis y cuya aplicación se expone ahora. En este capítulo se presenta un nuevo modelo mediante el cual, se aplican los conceptos indicados para el desarrollo de herramientas informáticas para el cálculo y estimación del tráfico agregado en redes de acceso a Internet. Para ello, en este capítulo, se hace una descripción de Internet, su estructura y la definición de diferentes configuraciones, para después realizar un estudio de las características del tráfico relacionado, sobre trazas reales. A continuación, se presenta el modelo para la caracterización de escenarios de acceso denominado CASUAL (Cubo de Accesos/Servicios/Usuarios de Asignación Libre). Su uso permite establecer criterios específicos para la aplicación de los modelos de tráfico a través de los diferentes servicios existentes en Internet, y se aborda tanto la problemática desde el punto de vista del tráfico de fuente, como desde el del agregado. Para ello se proponen dos mecanismos para la estimación del tráfico requerido por un usuario haciendo uso de un determinado servicio y acceso. La primera, desarrollada junto al modelo CASUAL a lo largo de esta Tesis, está basada en un modelo ON-OFF multinivel. La segunda propuesta hace uso del *Network Calculus* para la obtención de dicho valor a partir de medidas de tráfico reales, tras adaptar un algoritmo de optimización genérico reconocido, que ha sido adaptado al caso concreto del problema de estimación de tráfico de fuente. Para finalizar, se presentan las respectivas propuestas para la estimación del tráfico agregado: el uso de binomiales junto con fuentes ON-OFF multinivel, y el uso de aritmética *Max-Min* en el caso del *Network Calculus*.

3.1. Definición del escenario: la red Internet

Internet, actualmente, está organizada como una jerarquía en la que los usuarios se conectan a través de miles de proveedores de acceso (ISP) sea al nivel privado, local, regional, o incluso nacional, y que acceden a las redes y usuarios de otros ISP's gracias a acuerdos meramente comerciales. Conforme se asciende en la jerarquía, los grandes proveedores nacionales concentran sus inversiones en la denominada *Tier-2*, establecen su interconexión con otros ISP's a su mismo nivel mediante acuerdos de *peering*, y a los proveedores internacionales mediante acuerdos de tránsito (*transit peering*). Estos últimos pueden ocupar o no la parte más alta de la jerarquía, denominada *Tier-1*. En este punto las interconexiones son similares a las realizadas en el nivel inferior.

Esta estructura, pese a su rigidez, establece una diversificación de ISP's tal, que se hace necesario el establecimiento de gran número de puntos de intercambio de tráfico, tanto para la realización de acuerdos de *peering* como de *transit*. Estos

puntos reciben el nombre de IXP (Internet Exchange Points), y son las localizaciones concretas en las que los diferentes niveles de la red se interconectan entre sí, cuyo número depende directamente del número de ISP's existentes en cada nivel. Precisamente este número se ha visto incrementado en los últimos años por la proliferación de nuevos proveedores, así como por la alta disgregación de la cobertura de los mismos, incluso por debajo del ámbito local. Es por ello, que lo que en un principio se limitaba a localizaciones a nivel *Tier-1* y *Tier-2*, ha dado paso a numerosos IXP regionales. Para más información acerca del problema de la interconexión véase [Elixmann, 2002].

Esta situación se ha visto magnificada por la incorporación de nuevas redes gracias a la convergencia de tecnologías que propone el concepto NGN (*Next Generation Networks*). Según esta idea, Internet deja de ser exclusiva de tecnologías cableadas, gracias a la incorporación de las soluciones inalámbricas, 2G y 3G (WIFI, WIMAX, GPRS, UMTS, LMDS), y de las comunicaciones vía satélite.

3.1.1.1. **El escenario global**

La organización y estructura general de Internet es, a todos los efectos, tradicional desde el punto de vista de su planificación, diseño e implementación. Tal como se expone en el capítulo 2, los usuarios acceden a ella a través de redes de distribución (la SANP pertenecientes a un operador y/o ISP) interconectadas en un primer nivel por un acceso a la red (ANP) y a su vez enlazadas por una o varias dorsales (CNP). Por su parte, el tráfico generado por los usuarios, recorre dicha estructura de acuerdo a las políticas de agregación definidas por ISP's y operadores. En el caso más simple, los diferentes servicios utilizados por el internauta son tratados como flujos de tráfico hasta cierto punto independientes, que comparten el acceso físico de forma transparente para el usuario. El conjunto de aplicaciones genera así un único flujo de tráfico independiente para cada cliente, que hace uso de las capacidades ofrecidas por la red de distribución. Conforme los diferentes flujos ascienden desde el acceso hasta la dorsal, los recursos y capacidades de la red deben ser compartidos. Su interfaz también recibe el nombre de red de agregación, precisamente porque el conjunto de flujos individuales, procedentes de las redes de distribución pasan a compartir capacidades y recursos para ascender hasta la dorsal. La agregación de los diferentes flujos no es exclusiva de la ANP, puede llevarse a cabo a niveles bajos, en la red de distribución, en los denominados puntos de presencia POP, como paso previo al acceso a los *routers* de acceso a la dorsal, también denominados *Edge routers* (*border routers* o *routers* de frontera de área). A niveles superiores, aunque posible, la agregación se produce de forma puntual, en los denominados *core routers* (*routers* dorsales). En la Fig. 3.1 se muestra un ejemplo típico de distribución de los elementos de conmutación y enrutamiento desde el punto de vista de la red dorsal.

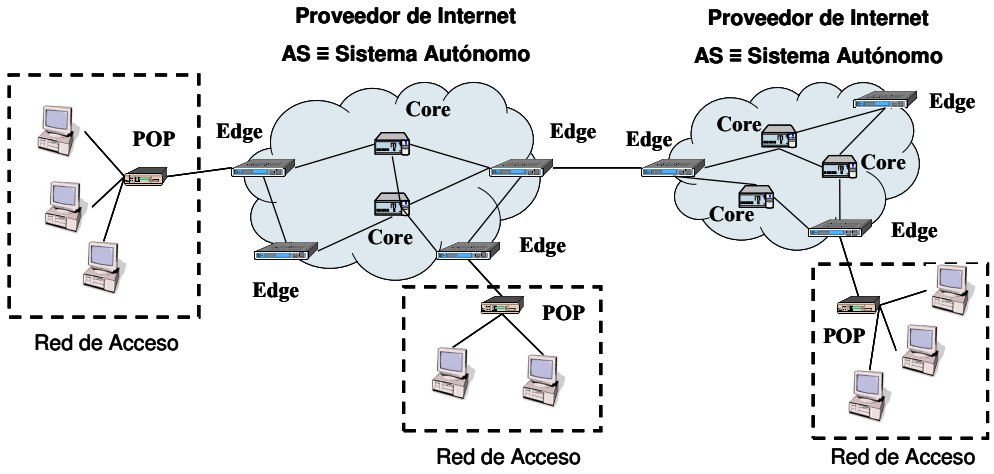


Fig. 3.1: Estructura típica de interconexión de usuarios y proveedores de servicio en Internet

Como se indicó en el apartado 2.3.1.2, la evolución de las redes actuales viene dirigida precisamente por aspectos directamente relacionados con un modelo tridimensional basado en términos geográficos, lógicos y funcionales, como pueden ser el nivel de acceso, el de transporte y el plano de control y servicios.

Si bien los usuarios hacen uso de accesos cableados e inalámbricos individuales, la red cuenta con centros o nodos de acceso (POP) que actúan como concentradores del tráfico de los usuarios, los cuales se conectan a la dorsal (*Edge routers*) mediante enlaces de alta capacidad. Es aquí donde las redes de nueva generación realizan la extensión de IP, para ofrecer QoS y gestión diferenciada de servicios y adecuar el ancho de banda de la red de transmisión, de transporte o dorsal. Los nodos de la parte dorsal (*core routers*), emplean GSR (*Gigabit Switch Routers*) de gran capacidad o bien TSR (*TeraSwitch Routers*) con capacidades específicas para enrutamiento a muy alta velocidad. De esta manera, el plano de control se distribuye a lo largo de toda la red, llevando a cabo todas las tareas específicas para la provisión de QoS y el control del tráfico.

Tradicionalmente, el reenvío en Internet se realiza de forma equitativa, todos los paquetes reciben la misma calidad de servicio, utilizando mecanismos de encolado FIFO. Sin embargo, hay situaciones en las que es deseable que el *router* provea diferentes perfiles de QoS para cada flujo, de acuerdo al modelo de servicio, tal como se indica en [Braden, 1994a]. El conjunto de funciones que permiten al *router* proveer calidades de servicio diferentes, son agrupadas en lo que se denomina "control de tráfico", destacando los siguientes elementos:

- *Packet scheduler*: Gestiona los reenvíos de los diferentes flujos de paquetes haciendo uso de un conjunto de colas y, en su caso, incluso temporizadores. Está localizado en el punto en el que los paquetes son

encolados, al nivel de dispositivos de salida de un Sistema Operativo típico, correspondiéndose con el protocolo de la capa de enlace. La elección del mecanismo de *scheduling* está condicionada al tipo de medio de salida utilizado. Adicionalmente puede aparecer otro dispositivo, el estimador, implementado como un algoritmo de medida de las propiedades del flujo de tráfico de salida y que obtiene las estadísticas necesarias para realimentar los procedimientos de control del *scheduling* y de admisión. Este dispositivo suele ser considerado parte integrada del *scheduler*.

- Clasificador: Con el objetivo de controlar el tráfico, cada paquete entrante es mapeado en su clase correspondiente, de forma que, todos los paquetes pertenecientes a una determinada clase reciban el mismo tratamiento en el *scheduler*. La definición de cada clase se realiza en función de los contenidos de las cabeceras de los paquetes y, en algunos casos, mediante identificadores específicos añadidos a cada paquete, de forma que una determinada clase puede coincidir con una extensa categoría de flujos, como por ejemplo todos los correspondientes a un determinado origen. En el extremo, una clase puede representar a un único flujo. Hay que tener en cuenta que el concepto abstracto de clase depende, localmente, de cada *router* individual, es decir, el mismo conjunto de paquetes son clasificados de forma diferente en puntos discretos de su trayecto. Así por ejemplo, los *core routers* mapean un número grande de flujos en unas pocas clases agregadas, mientras que los *edge routers*, donde la agregación es bastante menos importante, pueden separar cada flujo en clases diferentes.

Por su parte, el conjunto de funciones, que permiten al *router* mantener las calidades dependiendo del comportamiento real del tráfico, se denomina "control de recursos", entre los que cabe destacar:

- Control de admisión: Implementa el algoritmo de decisión que los *routers* o los equipos utilizan para determinar si un nuevo flujo puede ser servido con la QoS necesaria sin perjuicio de los ya garantizados. Es invocado en cada nodo para realizar localmente la decisión de aceptación o de rechazo, en el mismo momento que se solicita un servicio de tiempo real sobre un determinado trayecto Internet. El algoritmo responsable debe ser consistente con el modelo de servicio, y forma parte del proceso de control de recursos.
- Monitorización (*policing*): Situada en el *edge* de la red asegura, paquete a paquete, que un equipo no supere sus características de tráfico previstas. Suele ser considerada como parte del funcionamiento del *scheduler*. Se lleva a cabo en el nivel IP, por contra del control de admisión que se realiza en el de conexión.

Por último, el conjunto de funciones que permiten al *router* asegurar el cumplimiento de las restricciones de QoS y GoS extremo a extremo, se agrupan en funciones que se han denominado "ingeniería de tráfico", mediante

el uso de mecanismos de optimización de recursos y del enrutamiento. Sin embargo, este término está definido por la UIT como el conjunto de mecanismos necesarios para la planificación, diseño, dimensionado, gestión y supervisión de una red de comunicación en condiciones óptimas de acuerdo a la demanda de servicios, la QoS y el entorno regulatorio y comercial [UIT, 1993], por lo que estrictamente hablando realmente engloba a todos lo mecanismos anteriores.

En la tabla Tabla 3.I se muestran ejemplos de mecanismos y tecnologías asociados, así como un resumen de sus funcionalidades más importantes:

Categoría	Mecanismo	Tecnología	Implicaciones
Control de Tráfico	Clasificación + reenvío (*forwarding*) *Buffering*, *schedulling* *Shaping*	WFQ, WF2Q, WRR Colas de prioridad simples	Particionado estricto del ancho de banda Priorizado del tráfico
Control de Recursos	Control de adminisión (CAC) Monitorización de clases en el edge	*IntServ* + RSVP *DiffServ* + *Bandwidth Broker* Gestión de recursos	CAC→Bloqueo de flujos Monitorización→Límites predefinidos de tráfico + degradación del servicio
Ingeniería de Tráfico	Uso optimizado de recursos Optimizado del enrutamiento	OSPF, IS-IS, EIGRP MPLS *Multipath*	Enrutamiento basado en: Fuente/destino *Single*/*Multipath* Fijo/adaptativo

Tabla 3.I: Tecnologías y mecanismos utilizados por el plano de control de una red de Acceso a Internet

En [Cisco, 1999 / Lolischkies, 2003] se identifica la aplicación de los diferentes mecanismos a cada parte específica de la red, tal como se muestra en la Fig. 3.2.

Teniendo en cuenta que cada operador puede usar redes con diferentes arquitecturas, la agregación puede particularizarse en función de las mismas, y especialmente de las técnicas utilizadas para la provisión de QoS, donde las más importantes son:

- *DiffServ:* Permite al operador establecer mecanismos simples y escalables mediante los cuales clasificar y gestionar todo el tráfico de la red IP con garantías reales de QoS. De esta forma, el tratamiento del tráfico se realiza de forma diferente según sea considerado su uso, o lo que es lo mismo, el tipo de servicio al que pertenezca, distinguiendo en general entre servicios garantizados (en QoS) y servicios *best-effort* (con bajos requerimientos de QoS). Así, los servicios diferenciados permiten, a su vez, modificar el tratamiento específico del tráfico en función del tipo de usuario o de acceso a Internet en el que se está

haciendo uso de dicho servicio, y por supuesto el criterio de agregación, tal como se describe en [Ferrari, 2000].

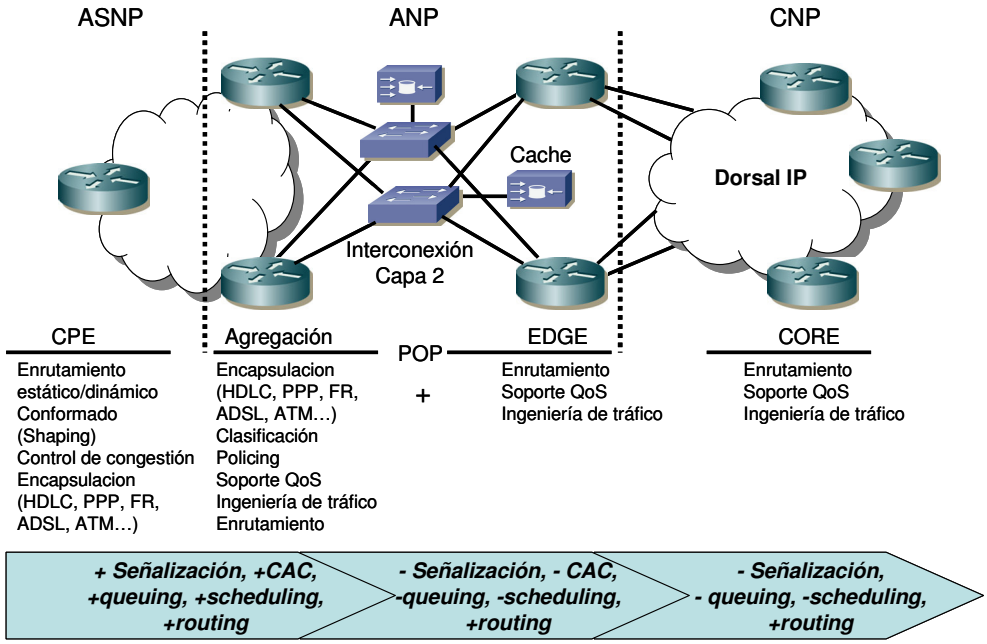


Fig. 3.2: Distribución de funciones del plano de control en una red de Acceso a Internet

- *Int-Serv* : Cada elemento de la red realiza el control de la QoS de forma individual, por lo que cada flujo debe obtener previamente privilegios suficientes como para asegurar la mínima QoS requerida, véase [Fu, 2003]. Así, el tráfico presenta unas especificaciones tanto de forma, como de necesidades, denominadas TSPEC y RSPEC respectivamente. De esta forma la agregación se realiza totalmente a medida, permitiendo garantizar la QoS de cada agregado independientemente, a costa de introducir mecanismos específicos para la señalización y control de cada conjunto de parámetros.
- RSVP: Aunque este protocolo es precisamente el mecanismo sobre el que se implementa *Int-Serv*, hay quien lo considera como una arquitectura independiente, en la que cada flujo IP establece sus propias reservas de recursos a lo largo de toda la red hasta completar el trayecto hasta su destino. Sin llegar al extremo de *Int-Serv*, los flujos pueden ser agregados de acuerdo a pautas arbitrarias, tanto más restrictivas cuanto mayor QoS se desee.
- Sobredimensionado: Tradicional e independientemente de la arquitectura de red, es una de las soluciones más utilizadas por los operadores. Provee la QoS, cualesquiera que sean las características, volumen de tráfico y el sobredimensionado de los enlaces, de forma que todos los flujos que converjan sobre él tengan recursos suficientes para llevar a cabo la comunicación. Todos son entonces agregados,

prácticamente sin señalización específica, con lo que los recursos son compartidos en igualdad de condiciones. Precisamente, esta es la razón por la que la QoS queda limitada por la cantidad de recursos disponibles en cada instante.

- Túneles virtuales MPLS: Cada flujo es tratado de forma independiente, pero lógicamente, mediante señalización específica. Suele aparecer en combinación de *DiffServ* e *IntServ*.

En la Fig. 3.3 se representan los diferentes mecanismos de provisión de QoS en función del grado de agregación y de la complejidad en su señalización. Conforme aumenta el control y se individualiza la agregación, la eficiencia de la solución mejora aunque a costa de una mayor complejidad que puede traducirse de diferentes maneras.

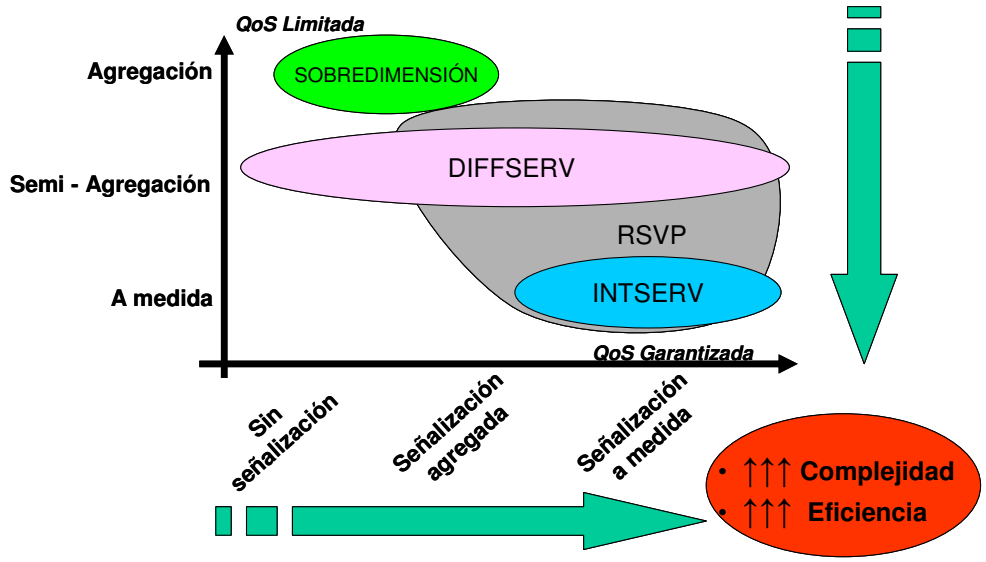


Fig. 3.3: Tecnologías de ingeniería de tráfico: Relación entre las necesidades de señalización en el plano de control y el nivel de agregación de tráfico obtenido

Precisamente, la provisión por parte de los ISP's y operadores de servicios específicos, como por ejemplo, la distribución de vídeo bajo demanda o de televisión digital, implican la necesidad de intensificar el uso de técnicas de ingeniería de tráfico como las anteriormente indicadas. En el entorno de las redes de nueva generación, toma fuerza el concepto de agregación de servicios como mecanismo fundamental para la provisión de los niveles y requisitos de conectividad necesarios, que hagan posible el soporte de servicios avanzados de cara al usuario final, con niveles de calidad y confianza suficientes. Si bien, originalmente la agregación de servicios consistía en la adición natural de diferentes flujos de tráfico, correspondientes a distintas actividades llevadas a cabo por los usuarios finales (aplicaciones y servicios TCP/UDP), actualmente las necesidades del mercado, y en especial el desarrollo de los conceptos de empresa y hogar digital, han obligado a los operadores e ISP's a desarrollar y

ofertar plataformas tecnológicas que realizan dicha agregación con los requerimientos de GoS y QoS correspondientes a cada tipo de servicio ofertado, y de acuerdo a las condiciones propias de cada usuario en particular. De esta forma, la agregación de servicios pasa a ser el soporte a través del cual, los desarrolladores y proveedores de contenidos y servicios hacen llegar su oferta a los usuarios finales. Como resultado, la tipificación de los usuarios en función del paquete de servicios contratados, establece esquemas de agregación de tráfico específicos e independientes en función de servicios concretos, del tipo de acceso contratado, o incluso de la clase de usuario final.

3.1.1.2. **El tráfico en Internet**

A lo largo de esta Tesis se proponen mecanismos específicos para la estimación de tráfico de Internet. En concreto, uno de ellos, basado en el *Network Calculus*, permite obtener una aproximación de dicho valor a partir de observaciones reales de dicho tráfico. Para ilustrar los diferentes modelos y algoritmos se ha optado por dos vías de actuación. Por un lado, el testeo de la bondad de los algoritmos propuestos mediante trazas de tráfico sintéticas, obtenidas a partir de modelos y funciones de distribución teóricas. Por otro lado, la comprobación de la utilidad de dichas estimaciones sobre capturas reales de tráfico Internet. En este apartado se hace un análisis de las trazas de tráfico utilizadas en este trabajo, desde el punto de vista estadístico, en cuanto a su autosimilitud y su aplicabilidad como muestras de referencia.

Múltiples estudios, entre los que destacan especialmente los de [Crovella, 1998 / Leland, 1991 / Paxson, 1995], han puesto en entredicho la concepción tradicional acerca de las características del tráfico en redes de comunicación. Sin embargo, es lógico tenida cuenta que sus resultados parten del análisis del comportamiento de una arquitectura totalmente diferente a la utilizada por las redes telefónicas y a las primeras redes de datos basadas en conmutación de paquetes (como X.25).

Tal como se presentó en el apartado 2.2, Internet hace uso de un modelo de capas prácticamente gobernado por el nivel de transporte, es decir, por los protocolos TCP y UDP, que a su vez determinan características específicas al flujo de datagramas IP que generan. Así un flujo Internet presenta una estructura básica formada por secuencias de paquetes, más o menos agrupadas en el tiempo en lo que se ha venido denominando ráfagas. Hasta aquí las coincidencias con el modelo tradicional (basado en las redes telefónicas) en el que el concepto de ráfaga está más bien limitado. Si bien el modelado mediante procesos de Poisson tanto para el proceso de llegadas como para la duración de las llamadas podría ser considerado como válido para el tratamiento de fuentes individuales a escalas de tiempo mayores, la agregación de varias de ellas provoca el suavizado del tráfico, desapareciendo el efecto de las ráfagas. [Leland, 1991] demuestra que en el tráfico Internet el efecto de las ráfagas puede ser observado a diferentes escalas de tiempo, y que la correlación entre las mismas abarca períodos de tiempo mayores que en el tráfico telefónico.

Estadísticamente, el valor de la varianza decae muy lentamente y la correlación persiste a lo largo del tiempo, propiedades que coinciden con las de los procesos autosimilares, descritos en el apartado 2.2.3.6.

El análisis de trazas de tráfico reales permite confirmar muchas de las características atribuidas a las comunicaciones sobre Internet. Es por ello que desde el año 96 el ACM SIGCOMM pone a disposición de la Comunidad Científica una biblioteca de información de tráfico Internet denominada *The Internet Traffic Archive*[16]. Aquí se reúne un extenso conjunto de trazas de tráfico obtenidas en diferentes momentos y redes, disponibles para su uso y estudio. Desde su creación han sido utilizadas con diversos objetivos, desde el estudio y análisis del tráfico de redes, la caracterización de usos de Internet o de los patrones de tráfico, hasta la definición de generadores de tráfico basados en dichas medidas. Este es el caso de las colecciones dedicadas a tráfico WAN en general, y que han sido utilizadas en sus investigaciones en [Fowler, 1991 / Leland, 1991 / Paxson, 1994a / Paxson, 1994b / Paxson, 1995 / Wilson, 1994][17]

Sin embargo, las capturas más numerosas corresponden a tráfico *Web* de servidores, como por ejemplo las de la EPA (*Environmental Protection Agency*), el *Research Triangle Park* en Carolina del Norte, el *San Diego Supercomputer Center* de California, el Departamento de Ciencias de Computación de la Universidad de Calgary en Canadá, *ClarkNet* del *Metro Baltimore-Washington DC*, el *NASA Kennedy Space Center* de Florida, vease [Arlitt, 1996] o la Web del Campeonato Mundial de Futbol del año 1998[18].

Este último caso ha sido uno de los que se tomaron inicialmente como referencia a lo largo de este trabajo, puesto que representa un ejemplo de escenario, específico, completo y perfectamente caracterizado, correspondiente a un subconjunto de equipos haciendo uso de Internet, como no podía ser de otra forma, de un servicio *Web*. Ya en el análisis realizado en [Jin, 1999] se desprende que desde el punto de vista del comportamiento a ráfagas resultante, en escalas de tiempo elevadas el tráfico resulta hasta cierto punto predecible. En esta Tesis se va a determinar hasta que punto esto es cierto y puede ser

---

[16] http://www.acm.org/sigs/sigcomm/ITA/

[17] [Leland, 1994], [Fowler, 1991], [Leland, 1994]: Tráfico LAN y WAN *Ethernet* del *Bellcore Morristown Research*, tomadas en el año 1989 y que incluye 1 millón de trazas, entre flujos Ethernet y de tránsito WAN.
[Paxon, 1995]: Tráfico WAN de la red de *Digital Equipment Corporation* (DEC), tomado en uno de los puntos primarios de acceso a Internet de la compañía durante el mes de marzo de 1995.
[Paxon, 1994] y [Paxon, 1994b]: captura TCP WAN de la red del *Lawrence Berkeley Laboratory* (LBL), tomadas desde el año 1993 al 1995.

[18] http://ita.ee.lbl.gov/html/contrib/WorldCup.html

utilizado en el proceso de planificación, no sólo desde el punto de vista del comportamiento del tráfico agregado, tal como se expone en el estudio original, sino lo que es más importante, la parametrización de los flujos individuales.

Más tarde, se han incorporado capturas más modernas ya que la evolución de Internet, de sus servicios y de las redes de acceso, han modificado enormemente tanto el uso de la red, como el comportamiento de los usuarios. En Internet, precisamente, se encuentran varias fuentes disponibles, entre las que destacan la Web del proyecto *Passive Measurement and Analysis* de la Universidad de California en San Diego, EEUU[19], la *Web* del proyecto WIDE – *Widely Integrated Distributed Environment*[20] y los resultados del proyecto M2C - *Measuring, Modelling and Cost Allocation*, mantenidos por la Universidad de Twente[21], la más interesante, al ser la única con medidas en una red a la europea.

A continuación se muestran algunos de los resultados del análisis previo que se hizo a la captura del Mundial'98, y a dos de las trazas de la Universidad de Twente (de 2002 y 2007). En este análisis se pretende observar el carácter autosimilar del tráfico Internet, y la posibilidad de filtrado de la información que permita utilizar parte de las capturas en estudios más específicos, por ejemplo, el comportamiento del tráfico de clientes individuales.

3.1.1.3. Caso 1: *Website* del Mundial de futbol

En concreto, todas las medidas se realizaron sobre el *Website* de la Organización del Campeonato del Mundo de Futbol celebrado en Francia en 1998, cuyo soporte contaba con 30 servidores distribuidos en cuatro localizaciones (4 en Francia y el resto en EEUU, 10 en Virginia, 10 en Texas y 6 en California). La información obtenida incluye todos los datos básicos de cada uno de los accesos a cada servidor durante los meses de Mayo y Junio de 1998, puesto que todas las peticiones eran distribuidas automáticamente manteniendo el balance de cargas entre todos los servidores. En estas condiciones el *Website* se comporta como un sistema de colas de hasta 30 servidores, y la captura refleja la secuencia de peticiones de servicio cursadas por el conjunto. Desde el punto de vista del tráfico, los datos permiten caracterizar el flujo total resultante de la misma manera que si hubiera sido obtenido en un único punto de agregación.

---

[19] http://pma.nlanr.net/
[20] http://tracer.csl.sony.co.jp/mawi/
[21] http://traces.simpleweb.org/

En la Fig. 3.4 se representa el total de bytes solicitados por una selección de 28 clientes con mayor actividad en este servicio *Web* durante la hora cargada, que en este caso se encontraba centrada alrededor de las 20:00 (dentro de la franja que iba desde las 18:00, momento en el que se iniciaban los partidos, hasta las 22:00 en que aproximadamente finalizaban). Hay que tener en cuenta que esta captura solamente representa una pequeña parte del tráfico generado por todo el conjunto de servidores (según [Jin, 1999] se obtuvo una media de 10500 peticiones por segundo), pero recoge cada una de las conexiones completas realizadas por los 28 clientes seleccionados y cuyos tráficos individuales agregó el sistema (aunque en diferentes puntos de la red).

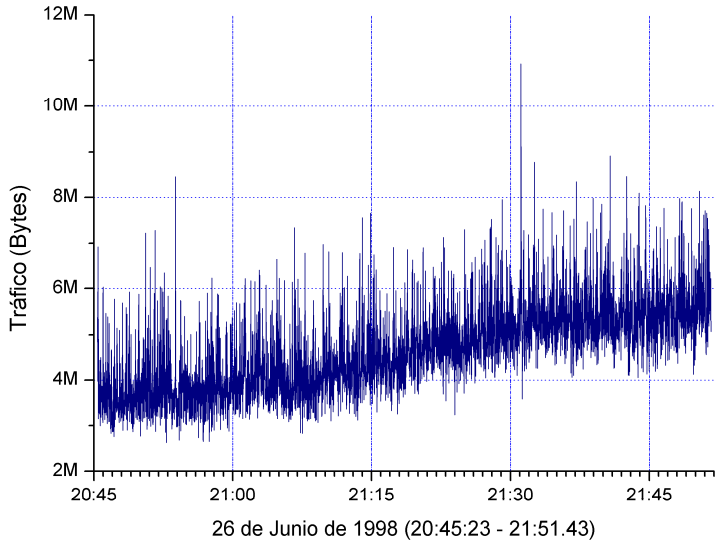


Fig. 3.4: Website del Mundial de Futbol 1998: Distribución del tráfico intercambiado en una hora punta

Se observa cómo el tráfico aumenta conforme los encuentros se van desarrollando y hasta su finalización. En la Fig. 3.5 se muestran dos detalles de 8 y 2 minutos de duración respectivamente.

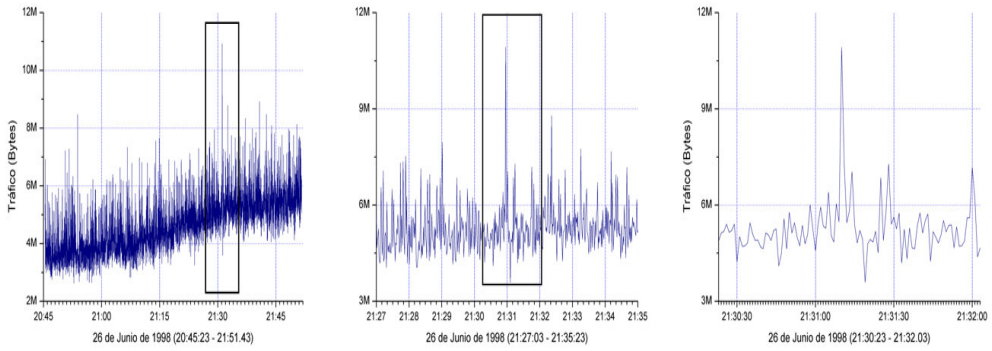


Fig. 3.5: Web del Mundial'98: Detalle de la captura, de izquierda a derecha, 1 hora, 8 minutos y 2minutos

El incremento del tráfico se produce a largo plazo, al menos 10 minutos, manteniéndose en niveles similares en todo momento. Al incrementar el período de observación de 1 segundo a 2, 5 y 10 segundos se obtienen los resultados de la.Fig. 3.6.

El incremento en el período de muestreo permite observar cómo el tráfico mantiene su estructura y forma, incluso a diferente escala de tiempo. Al menos estadísticamente, el proceso definido por el tráfico agregado del ejemplo es autosimilar. Para confirmar esta característica se suele recurrir al uso de períodogramas (también llamados correlogramas) y a la estimación del parámetro de Hurst.

Un períodograma va a representar la función de autocorrelación. Para ello, partiendo de una determinada serie de tamaño N, se calculan los coeficientes de autocorrelación de orden k, siendo k el desplazamiento o retardo. En cualquier caso, se recomienda que $N > k+1$, $N \geq 50$ y $k \leq N/4$. La estimación de los coeficientes de autocorrelación se obtiene mediante la autocovarianza, mientras que para la varianza se utiliza el *estadístico de Barlett*:

$$\hat{V}(r_k) \cong \frac{1}{N} \sum_{-(K-1)}^{K-1} r_i^2 \Rightarrow \begin{cases} V(r_k) \cong \dfrac{1}{N} & k \geq K,\ K = 1 \\ V(r_k) \cong \dfrac{1}{N}\left(1 + 2\sum_{1}^{K-1} r_i^2\right) & k \geq K,\ K > 1 \end{cases} \tag{3.1}$$

Junto con la estimación de los coeficientes, se suele representar el intervalo $\pm 2S(r_k)$, donde $S(\cdot)$ es la desviación estándar, denominado intervalo de no significación del coeficiente de autocorrelación, y que representa el conjunto de valores que puede tomar r_k para afirmar que no existe correlación entre valores separados un tiempo k, con una probabilidad de acierto del 95%. En la Fig. 3.7 se muestra el correlograma correspondiente a la traza de tráfico estudiada.

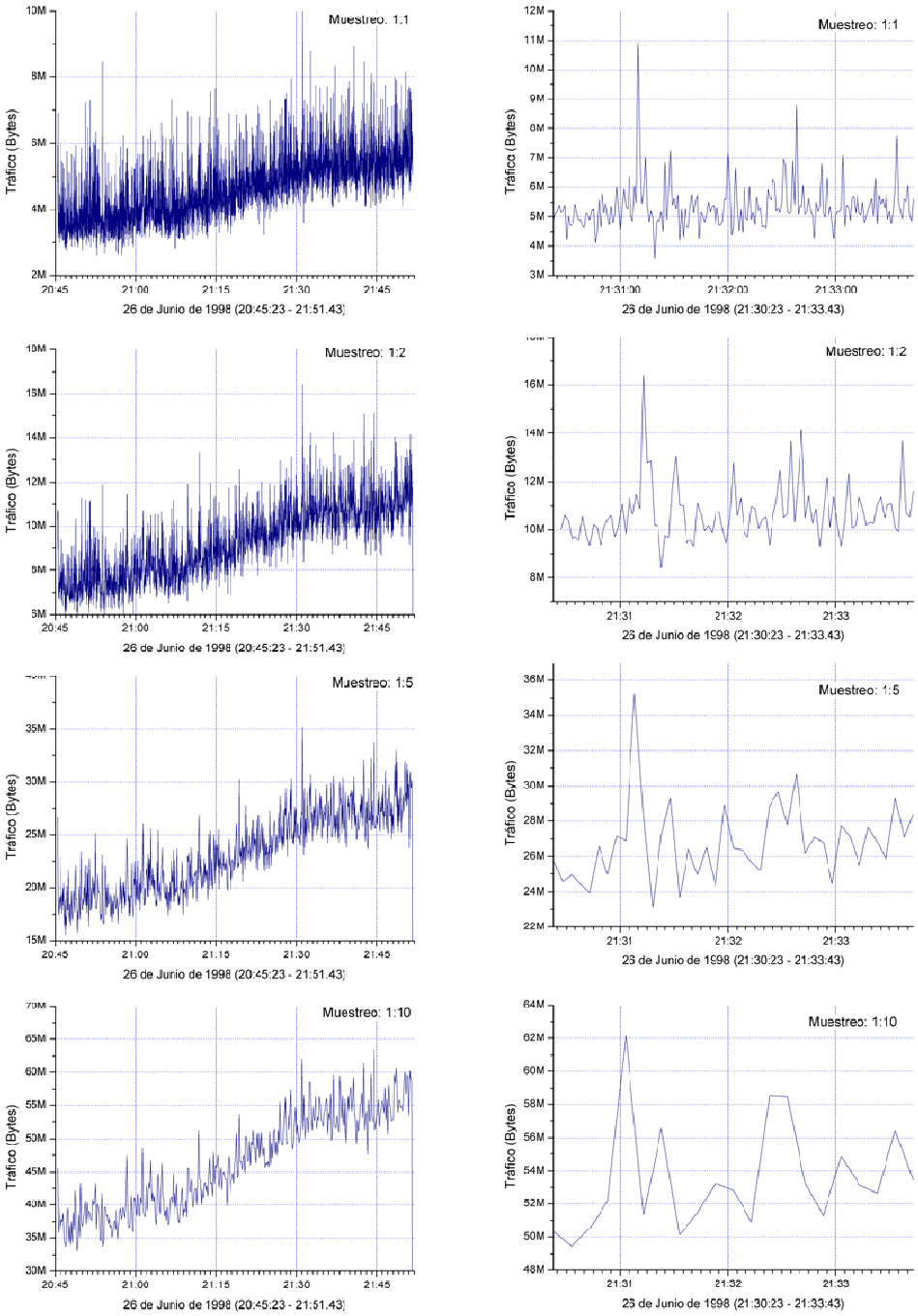


Fig. 3.6: Web del Mundial'98: Efecto del cambio de escala en la estructura y forma de la captura, a la izquierda sobre una observación de 1 hora y a la derecha solamente 3 minutos. De arriba a abajo, dichas observaciones, pero incrementando el tiempo de muestreo de 1 a 2, 5 y 10 segundos.

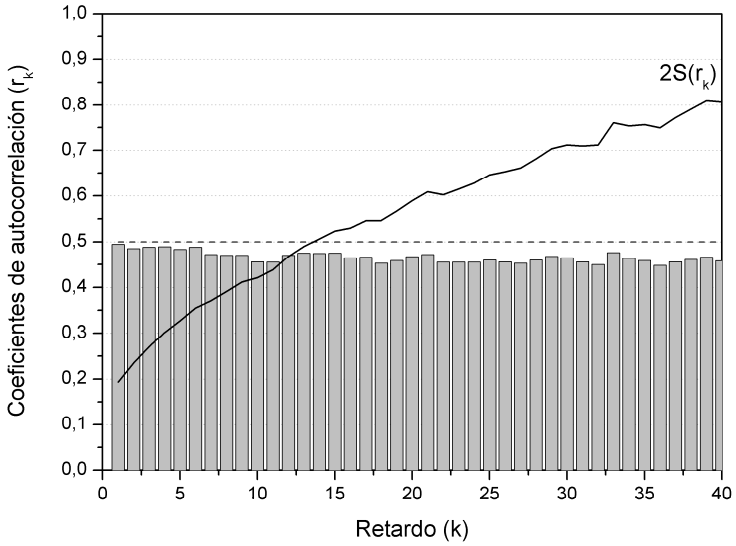


Fig. 3.7: Website del Mundial de Futbol 1998- Coeficientes de autocorrelación e intervalo de no significación

Los coeficientes de correlación muestran la falta de estacionalidad de la serie, y que la correlación solamente es apreciable a partir de 12 segundos.

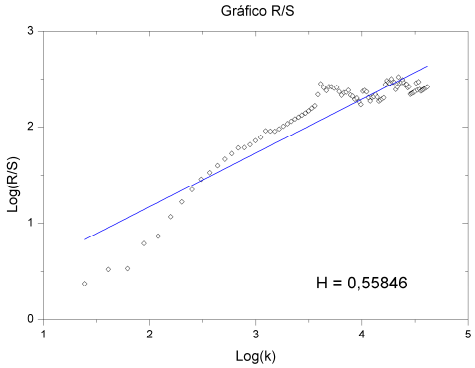
Por su parte, para el cálculo del parámetro de Hurst, se utiliza el método del Gráfico de Re-escalado (R/S), haciendo uso de la *addin* para *Excel* de *Kaotix*[22]. En la Fig. 3.8 se muestra dicho cálculo, junto con la gráfica de la varianza para diferentes períodos de observación.

Se observa cómo, al aumentar el tiempo de observación, el parámetro de Hurst se incrementa, pasando de la aleatoriedad a corto plazo ($H = 0,56$ para 2 minutos), a la autosimilitud exacta a largo plazo (entre 30 y 60 minutos). Para tiempos superiores el parámetro de Hurst toma valores mayores que uno, fuera de los límites de los procesos autosimilares. Algunos autores, como por ejemplo en [Giorgi, 2008], achacan estos valores al efecto de muestras erróneas y a la no estacionalidad de la serie.

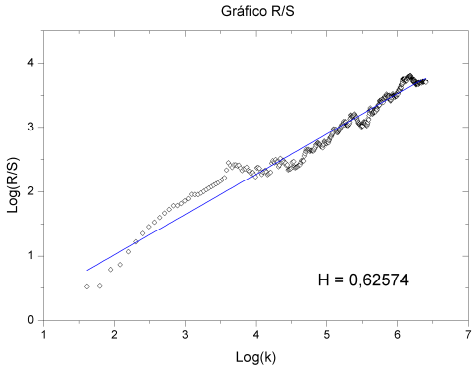
Si bien esta captura no se corresponde realmente al tráfico que cabría esperar en un punto de agregación de la red, cada uno de sus componentes, esto es, cada una de las trazas pertenecientes a los clientes individuales se corresponden a sesiones Web completas, obtenidas en la parte del servidor, por lo que serán

[22] http://www.xlpert.com/rescaled.htm

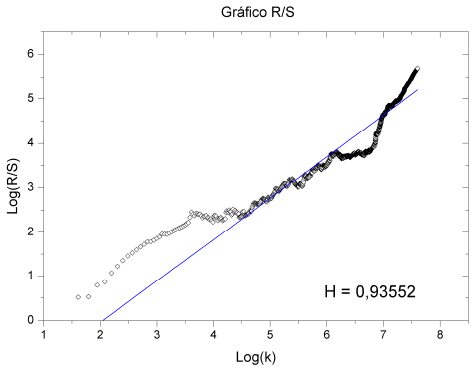
utilizadas como ejemplo de tráfico de clientes individuales en los estudios que se realizan en los apartados siguientes.



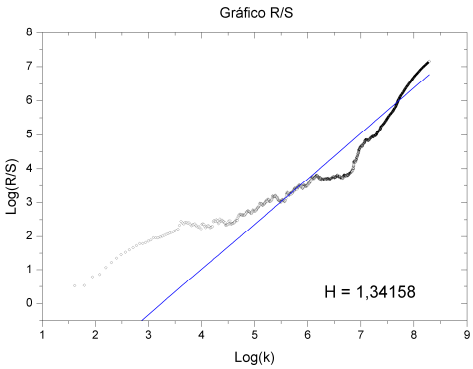
a) 2 minutos



b) 10 minutos



c) 30 minutos



d) 1 hora

Fig. 3.8: Website del Mundial de Futbol 1998 – Estimación del parámetro de Hurst para secuencias de 2, 10, 30 y 60 minutos

3.1.1.4. **Caso 2: Red de distribución en el Campus de la Universidad de Twente**

Uno de los objetivos del proyecto M2C[23] era comprender las características del tráfico Internet. Para ello se realizan medidas detalladas del tráfico generado en redes de distribución diferentes. Las capturas se realizan en los puntos de acceso a Internet de dichas redes, esto es, el conjunto de usuarios accede a la red de distribución que a su vez se conecta a Internet en dicho punto, siendo el lugar donde se produce la agregación del tráfico perteneciente a fuentes individuales.

Esta captura corresponde a una zona residencial de la universidad, con su propia infraestructura de red, en la que unos 2000 estudiantes acceden con interfaces Fast-Ethernet a 100 Mbps que se interconectan mediante *switchs*, y estos entre si mediante enlaces de 100 y 300 Mbps. El punto de acceso a la red de la universidad (y a Internet) se compone de una dorsal de 300 Mbps (3x100 Mbps).

A continuación se muestra un fragmento 15 minutos de la captura, realizada el martes 28 de mayo de 2002, a media mañana:

[23] http://traces.simpleweb.org/

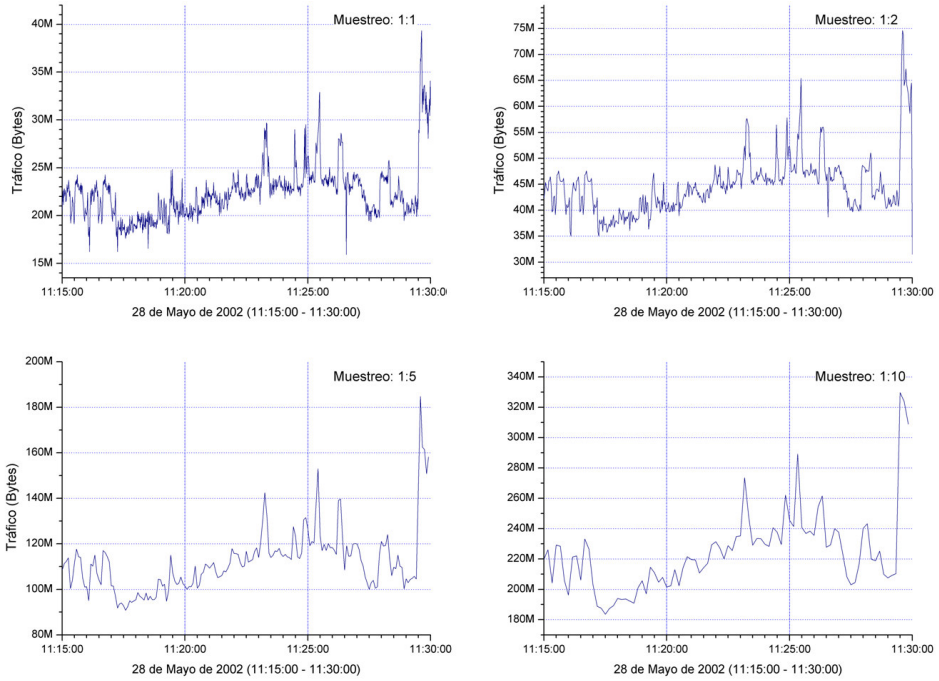


Fig. 3.9: Campus Univ. Twente – Efecto del cambio de escala para 1, 2 5 y 10 segundos

Al modificar el período de muestreo la figura del tráfico mantiene su estructura y forma, por lo que puede considerarse estadísticamente autosimilar. Por su parte la autocorrelación tiene la forma de la Fig. 3.10:

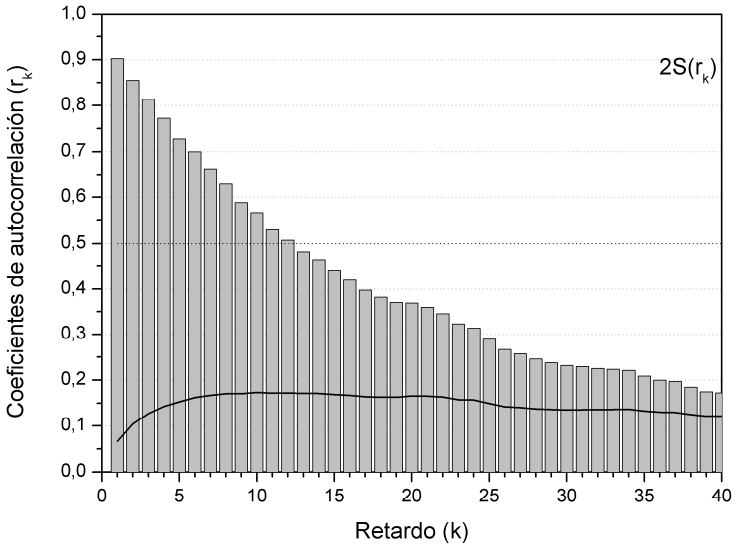


Fig. 3.10: Campus Univ. Twente - Coeficientes de autocorrelación e intervalo de no significación

En este caso, las evidencias de autosimilitud son más claras al añadir que el proceso no presenta ninguna estacionalidad, no hay correlación alguna (aunque la figura sólo muestra los coeficientes hasta un retardo de 40 segundos), y la autocorrelación no decae exponencialmente. Si se suma el gráfico R/S de la Fig. 3.11:

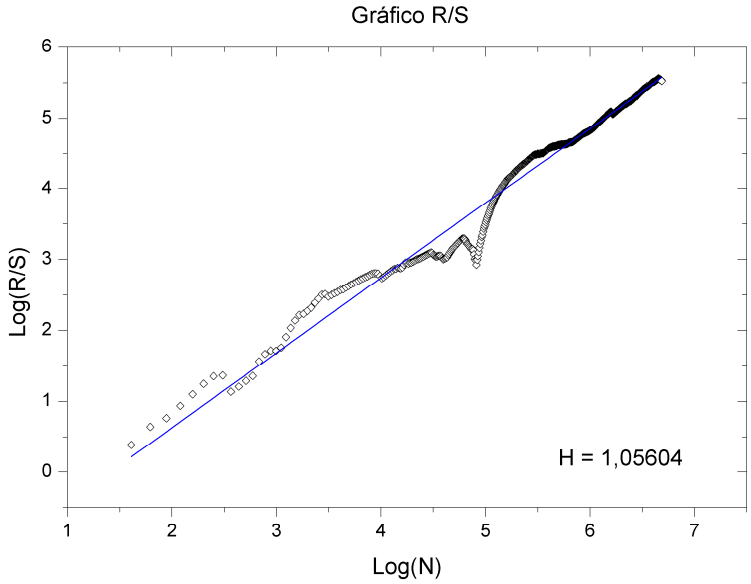


Fig. 3.11: Campus Univ. Twente – Estimación del parámetro de Hurst

El parámetro de Hurst es prácticamente 1 y, a diferencia del caso anterior, se mantiene constante con independencia de la escala de observación.

3.2. Modelo de caracterización del tráfico INTERNET: El modelo CASUAL

Actualmente, la caracterización del tráfico Internet, puede afrontarse siguiendo tres puntos de vista distintos pero íntimamente relacionados, con dependencias muy claras en determinados casos y algo difusas en otros. Por un lado el tráfico va a depender directamente del tipo de servicio, de las aplicaciones y de los protocolos que lo implementan. A su vez, un determinado tipo de servicio va a ser utilizado de distintas formas, con diferentes finalidades y usos, pudiendo modificar el comportamiento del tráfico resultante, en función de las características del usuario, tanto morfológicas (ej. por franja de edad), como ambientales (ej. uso residencial frente a uso profesional) o circunstanciales (ej. modas y novedades). Por último, el comportamiento de un determinado usuario no va a ser el mismo a la hora de utilizar un servicio concreto, en función de las características físicas de su acceso a Internet, modificando radicalmente las características del tráfico de fuente resultante.

Con todo ello, la caracterización del tráfico Internet es una función multivariable, definida a lo largo de tres variables fundamentales (tipo de servicio, de usuario y de acceso). Cada uno de los puntos de la función representa un conjunto de parámetros fundamentales, que caracterizan el tráfico de fuente resultante, así como el valor de éste, específico para cada posible combinación de las tres variables fundamentales. La obtención del valor, o en su defecto de la expresión (ó función) correspondiente, del tráfico asociado a cada punto de la función, depende directamente del conjunto de parámetros definidos en el mismo, así como de los mecanismos de modelado vistos en los capítulos anteriores.

En esta Tesis se propone un sistema de referencia tridimensional para la identificación de las características y el valor del tráfico de fuente asociado a cada combinación de los valores de los ejes de referencia. Cada tripleta va a ser considerada de forma totalmente independiente, y puesto que los tres ejes presentan variables categóricas, van a poder ser representadas como cubos dentro de una gran malla cúbica. La existencia de cada tripleta depende, casi exclusivamente, de que la combinación correspondiente sea considerada o no por el observador. Teóricamente, todas las combinaciones son posibles, aunque la práctica permite eliminar situaciones contradictorias(ej. videoconferencia con acceso *modem* a 1600 bps) y marginales o despreciables (ej. HDTV en una PYME).

Precisamente, el desarrollo y aplicación de este concepto es una de las aportaciones principales de esta Tesis, y ha sido utilizado para la definición del modelo bautizado como CASUAL, acrónimo de Cubo de Accesos, Servicios y Usuarios de Asignación Libre.

El modelo CASUAL permite definir escenarios de red completos, bien al nivel de nodos concretos o al de la red completa en sí. Así, por ejemplo, un determinado nodo de acceso POP, o incluso toda la red de distribución, quedará caracterizado por el conjunto de tripletas que representan a todos los servicios utilizados por cada tipo de usuario conectado, en función del acceso utilizado. La observación de cada tripleta de forma individual permite caracterizar el tráfico de fuente asociado, y aplicar la agregación de tráfico allí donde sea necesario y de acuerdo al mecanismo de provisión de QoS seleccionado. Dicha agregación se afronta de forma individual en cada tripleta (agregación de tráfico homogéneo) o bien a lo largo de uno, dos o los tres ejes, por ejemplo, la agregación del tráfico correspondiente a tripletas a lo largo del eje de servicios, permite obtener el tráfico total por cada tipo de usuario. Tomando como referencia el eje de accesos se obtiene el tráfico total por cada tipo de servicio, mientras que si es el de usuarios, se obtiene el total por cada acceso. De la misma manera, la combinación de dos o más ejes obtiene figuras de tráfico agregado diferentes, tal como se muestra en la Fig. 3.12.

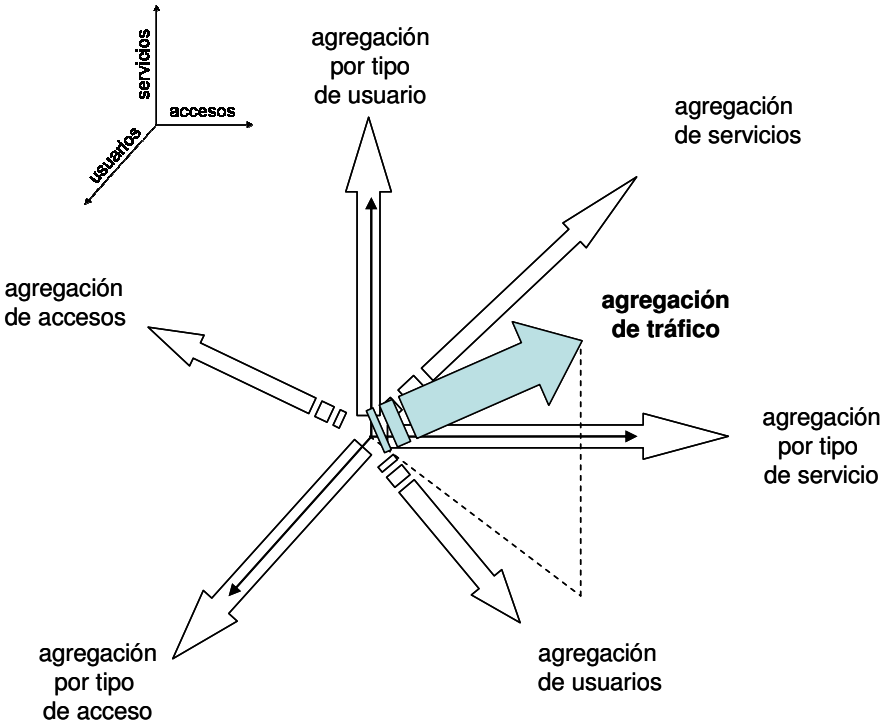


Fig. 3.12: Ejemplo de caracterización multivariante aplicada por el modelo CASUAL

La definición del escenario será tanto más completo cuanto mayor y mejor sea la categorización de los diferentes ejes de referencia. En los siguientes apartados se hace una breve descripción del significado de cada uno de ellos, y los correspondientes valores que pueden ser utilizados en una aplicación típica.

3.2.1. El eje USUARIO

El distinto comportamiento de los usuarios, sin dejar de ser algo totalmente subjetivo, puede ser relativizado hasta cierto punto si tienen en cuenta la afinidad de costumbres y necesidades relacionadas en un grupo de clientes. Estas son aún más claras cuando se observan comportamientos asociados con determinados estatus económicos y laborales.

Tradicionalmente, los usuarios de telecomunicación han sido categorizados en función del porcentaje de recursos utilizados, y por tanto contratados. Así, a grandes rasgos, los operadores distinguen entre Gran Empresa (GE), Pequeña y Mediana Empresa (PYME) y usuarios residenciales. Dichas categorías presentan grandes diferencias, no solo en la inversión en redes y servicios de telecomunicación, sino en el grado de uso, los requerimientos de seguridad y de calidad, en el tipo de servicios demandados, y por supuesto, en el tipo de tráfico generado:

- Gran Empresa: Requiere/ofrece redes IP alternativas, en forma de intranets, extranets seguras y todo ello reduciendo los costes. Actualmente la tendencia es la implementación ó contratación de Redes Privadas Virtuales (VPN – *Virtual Private Networks*), y en cualquier caso, la exigencia de cumplimiento por parte del operador de unas condiciones técnicas, económicas y de disponibilidad, recogidas en lo que se denomina SLA (*Service Level Agreement*). En cuanto al tráfico generado, mayoritariamente se reduce a la transferencia de grandes volúmenes de información, asociadas a aplicaciones de bases de datos, gestión de información, o copias de respaldo y de seguridad.
- PYMES: Las necesidades de las empresas conforme se reduce su volumen se acerca cada vez más a las del entorno residencial. Sin embargo, en este caso el acceso a redes corporativas resulta fundamental, intranets y VPNs que permitan la comunicación de la empresa con sus distribuidores y clientes, e incluso facilitar el tele-trabajo. Los costes tienden a reducirse pero los requerimientos de SLA siguen siendo en la mayoría de los casos críticos.
- Residencial: El usuario final también demanda servicios y redes de telecomunicación concretas, aunque con fines y gustos muy diferenciados frente a los usuarios de empresa. La rapidez del acceso prima por encima de la calidad de los servicios, aunque en la actualidad, la demanda de nuevos servicios multimedia hace que los operadores también hagan su propia clasificación de usuarios mediante paquetes de acceso concretos. Al contrario que en las dos categorías anteriores, actualmente no se contratan accesos con un SLA concreto, sino una tarificación basada en tarifa plana con limitaciones mensuales en el volumen de bytes intercambiados con la red.

En la Fig. 3.13 se muestra el número total de líneas con acceso a Internet existentes en España en el período 2004-2006:

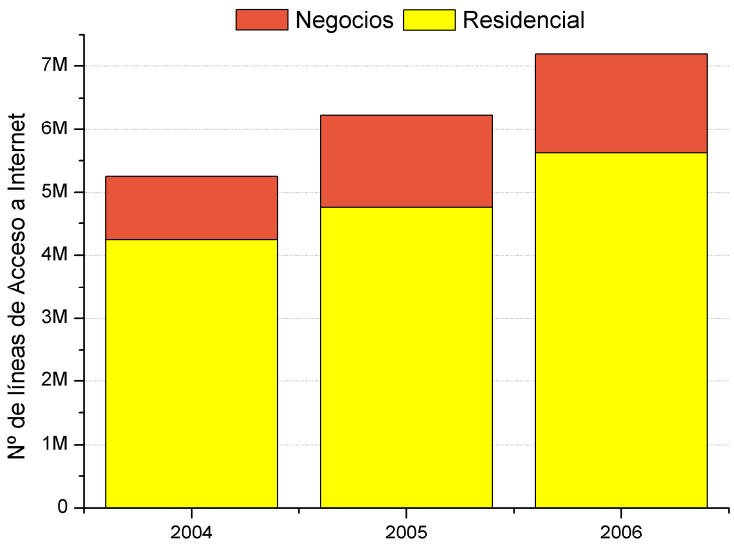


Fig. 3.13: Número de líneas con servicio de acceso a Internet en España (fuente: CMT)

Según la estadística de [CMT, 2006], la proporción de líneas residenciales es, en media, del 78,5 % del total de líneas con acceso a Internet, aunque el incremento de líneas anual, tanto en uno como en otro segmento, es de alrededor del 18%.

En cuanto al uso de dichas líneas, el comportamiento de los usuarios también ha evolucionado, aumentando significativamente la frecuencia de acceso a Internet, como se desprende del estudio realizado por el INE (Instituto Nacional de Estadística) y que se muestra en la Fig. 3.14:

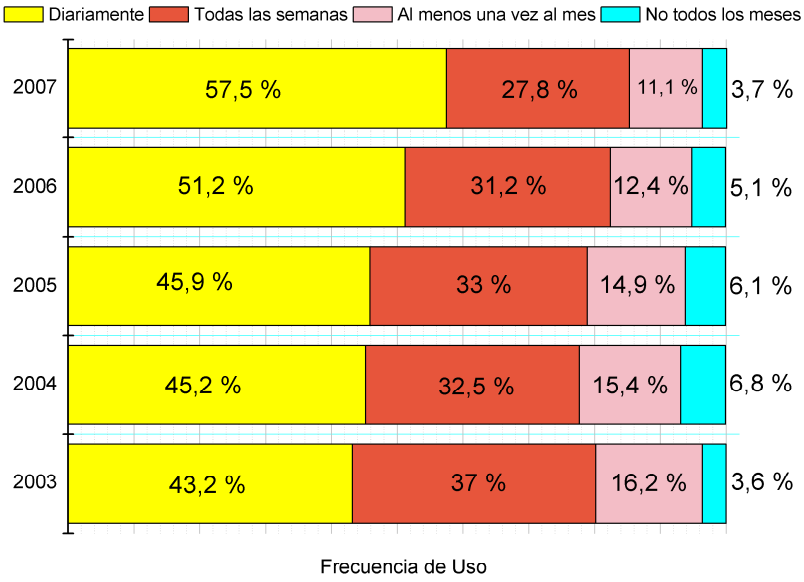


Fig. 3.14: Evolución de la frecuencia de uso del acceso a Internet en España (fuente:[Telefónica, 2007a])

3.2.2. El eje ACCESO

La evolución de las tecnologías de acceso ha venido motivada, fundamentalmente, por el gran incremento en la demanda de ancho de banda por parte de los usuarios, la aparición de nuevos servicios y la evolución de los existentes y a la necesidad por parte de los operadores de ampliar su área de cobertura sin penalizar al usuario con un ancho de banda menor.

Básicamente, el acceso a Internet se clasifica en cuatro grandes grupos de tecnologías:

- Acceso cableado (par trenzado): Tradicionalmente asociado al servicio telefónico, es el medio físico con mayor implantación en la red, especialmente en la red de distribución. Aprovechando dicha coyuntura estructural, las diferentes tecnologías han evolucionado prácticamente en un único sentido, aumentar el ancho de banda reduciendo el coste de infraestructuras. Así, cronológicamente aparecieron tres soluciones diferenciadas:
  - *Modems* en banda vocal (VBD – *Voiceband Data*), implementados inicialmente en las redes de telefonía básicas, aprovechan el ancho de banda vocal para la transmisión de datos hasta velocidades de 56 Kbps. Actualmente es una de las opciones más baratas para el acceso a Internet, especialmente en entornos residenciales, incluida por todos los operadores

con servicios de telefonía, independientemente de la tecnología de red empleada.

- RDSI (Red Digital de Servicios Integrados), también denominada DSL (*Digital Subscriber Line*). Surge como evolución natural de los *modems* de banda vocal. En este caso, el uso de mayor ancho de banda y de mecanismos de codificación más eficientes, permite alcanzar velocidades múltiplos de 2x64 kbps (2 canales B a los que hay que añadir 16 kbps ó canal D, de señalización fuera de banda). Si bien esta tecnología no ha llegado a tener la aceptación esperada, las calidades ofrecidas la convirtió en elección natural de muchas empresas, y en casos, de usuarios residenciales "avanzados".
- xDSL, quizá la gran impulsora de la generalización de los accesos a Internet, tanto en entornos residenciales como de empresa. La evolución de los *modems* de banda vocal y de RDSI, da paso a múltiples soluciones que hacen que la red de distribución pueda ser considerada como de "alta velocidad", aunque siempre limitados por la distancia desde los usuarios hasta el punto de acceso a la red del operador (*Central Office* ó punto de distribución). Prueba de su enorme desarrollo, es la cantidad de variantes que han surgido, como se resume en la Tabla 3.II:

Tecno	V max. Descarga	V max. Subida	Distancia	ITU-standard
ADSL	8 Mbps	1 Mbps	2 Km	ANSI T1.413 Issue 2 ITU G.992.1 (G.DMT) ITU G.992.2 (G.Lite)
ADSL2	12 Mbps	2 Mbps	2,5 Km	ITU G.992.3/4 ITU G.992.3 Annex J ITU G.992.3 Annex L
ADSL2+	24 Mbps	5 Mbps	2,5 Km	ITU G.992.5 ITU G.992.5 Annex L ITU G.992.5 Annex M
HDSL	T1/E1 (sobre 2 pares)	T1/E1(sobre 2 pares)	20 Km (con rep.)	ITU G.991.1
HDSL2	T1/E1 (sobre 1 par)	T1/E1 (sobre 1 par)	20 Km (con repetidores)	
IDSL	128 Kbps	128 kbps		
MSDSL	2 Mbps	2 Mbps	8,8 Km	
PDSL	30 Mbps	30 Mbps	50 Km (con repetidores)	
RADSL	Hasta 8 Mbps	Hasta 1 Mbps		
SDSL	2,3 Mbps	2,3 Mbps	3 Km	
SHDSL	6 Mbps	6 Mbps	2,7 Km	ITU G.991.2
UDSL	Hasta 200 Mbps	Hasta 200 Mbps		
VDSL	30 Mbps	5 Mbps	1,6 Km	ITU G.993.1
VDSL2	100 Mbps	100 Mbps	0,5 Km	ITU G.993.2

Tabla 3.II: Principales tecnologías xDSL

- Acceso híbrido (fibra óptica/coaxial): Su origen se remonta a las redes de cable coaxial para la distribución de TV. De la misma manera que en el caso del par trenzado, su adaptación para la transmisión de datos, hizo que cronológicamente aparecieran tres variantes diferenciadas:
    - Híbrida de fibra y coaxial (HFC): Originalmente con capacidad exclusiva para bajada de hasta 30 Mbps mediante el uso de un *cablemodem*, y que implementa el canal de subida mediante VBD ó RDSI.
    - HFC bidireccional: Evolución de la red de TV de cable, que manteniendo los 30 Mbps de bajada, implementa un canal de subida basado en técnicas de contienda, por lo que limita su velocidad hasta los 10 Mbps
    - Sistemas SDB (*Switched Digital Broadband*): Clasificados como sistemas digitales en banda base, con canal de bajada de hasta 50 Mbps. Al igual que en el caso anterior, el canal de subida utiliza técnicas de contienda, con velocidades máximas de 1,5 Mbps, aunque con un mínimo de 16 Kbps por usuario. La diferencia estriba en que la fibra llega prácticamente hasta grupos reducidos de usuarios (54 a 60 viviendas frente a 400-2000 del HFC bidireccional.
- Acceso *Wireless*: Conceptos como la movilidad o la cobertura total han motivado la necesidad de abandonar las técnicas de transmisión cableadas, al menos en el acceso a Internet. Existen cuatro grandes variantes:

    - Sistemas Celulares: De igual manera que los sistemas cableados evolucionaron desde la red de telefonía, los sistemas *wireless* han evolucionado a partir de las redes de telefonía móvil, desde capacidades mínimas de transmisión de datos de las primeras redes móviles, hasta las últimas tecnologías (3.5G), tal como se muestra en la Tabla 3.III:

Tecnología	Descarga (Kbit/s)	Subida (Kbit/s)
GSM	9,6	9,6
CSD	9.6	9.6
HSCSD	28.8 - 43.2	14.4
GPRS	80.0 / 60.0	20.0 / 40
EGPRS	236.8 / 177.6	59.2 / 118.4
UMTS	144 Kbps / 384 Kbps / 17.2 Mbps En movimiento / En espacio abierto / Estático	
HSPA	14.4 Mbps	2 Mbps

Tabla 3.III: Tecnologías de redes móviles de comunicaciones

- Sistemas de Difusión Terrestre: Tienen su origen también en los sistemas de distribución de TV. En este caso son tecnologías que utilizan la banda de microondas y establecen un número dado de canales independientes, cada uno de los cuales permite transmisiones en sentido de descarga de unos 10 Mbps y hasta 27 Mbps, mientras que en el sentido ascendente hacen uso de líneas telefónicas. Este es el caso de MMDS (*Multichannel Multipoint Distribution Service*), en la bada de 2-3 GHz, y de LMDS (*Local Multipoint Distribution Service*), en la banda de 26-29 GHz. Este último permite distancias de hasta 8 Km entre repetidores, pero cuadriplicando el ancho de banda de descarga.
- Sistemas de Difusión Satélite: Similares a los anteriores, en este caso el repetidor es un satélite en órbita con canales de descarga de hasta 1 Mbps, mientras que el canal de subida se implementa sobre líneas telefónicas, fijas ó móviles (GPRS). Como alternativa, aunque muchísimo más costosa, existen sistemas bidireccionales, como por ejemplo el ofrecido por INMARSAT, en el que tanto la descarga, como la subida llegan a valores de hasta 500 Kbps sin necesidad de equipos excesivamente sofisticados.
- WLAN (*Wireless Local Area Network*): Las últimas tecnologías incorporadas al conjunto de accesos a Internet no fueron inicialmente concebidas con tal fin. Es el caso de Bluetooth y los enlaces mediante infrarrojos (IR), que actualmente permiten establecer conexiones desde unos pocos centímetros hasta unas pocas decenas de metros. Una alternativa para su aplicación en entornos LAN, denominada por el IEEE como 802.11, más conocida como WIFI, ha evolucionado desde velocidades simétricas de 1 Mbps en decenas de metros, hasta velocidades de 108 Mbps a distancias de hasta 1 Km (con técnicas MIMO - *Multiple-Input and Multiple-Output*). Como alternativa a los mecanismos de acceso cableados surge también WIMAX (*Worldwide Interoperability for Microwave Access*), también conocida como 802.16 y WMAN (*Wireless Metropolitan Area Network*). Este mecanismo, basado en microondas, aunque obtiene tasas de transmisión de hasta 70 Mbps / 50 Km, en entornos semi-urbanos cae hasta 2 Mbps / 10 Km, y en entornos puramente urbanos a 10 Mbps / 2 Km, aunque eso sí, totalmente simétricas.

- Acceso Óptico (todo fibra): La evolución de la red dorsal, desde los sistemas cableados hacia los sistemas ópticos, también se ve reflejada en la parte de acceso. La fibra óptica cada vez se encuentra más cerca del usuario final, por lo que muchos operadores apuestan directamente a la unificación de criterios mediante la implantación de redes con toda su tecnología óptica. Existen dos alternativas claras:

- PON (*Passive Optical Network*): En este caso una única fibra óptica da servicio a múltiples usuarios, hasta los que llegan sendas terminaciones ópticas desde un dispositivo totalmente pasivo (denominado *splitter*). Actualmente existen varias especificaciones, que se muestran en la Tabla 3.IV:

Tecnología	Implementación	Estándar
APON		ITU-T G.983
BPON	TDM-PON WDM-PON	ITU-T G.983
GPON		ITU-T G.984
EPON / GEPON	DOCSIS-PON RF-PON or RFOG	IEEE 802.3ah
10GEPON	TDM-PON WDM-PON DOCSIS-PON RF-PON or RFOG	IEEE 802.3av

Tabla 3.IV: Passive Optical Network: Principales tecnologías y estándares.

- SONET / SDH (*Synchronous Optical NETworking* / *Synchronous Digital Hierarchy*). La mayor parte de la red dorsal Internet está formada por anillos de fibra óptica basados en esta tecnología. Su extensión o la implementación de anillos dedicados en la parte del acceso, parece una evolución natural para los grandes operadores. Esta tecnología trabaja en modo circuito y TDM (*Time Division Mulplexing*), por lo que cada flujo presenta una tasa de transmisión y un retardo constantes. El ancho de banda toma valores múltiplos enteros de 64 Kbps, de acuerdo con los valores máximos indicados en la Tabla 3.V.

SONET Portadora Optica	Trama	SDH Trama	Payload (Kbps)	Línea (Kbps)
OC-1	STS-1	STM-0	48,960	51,840
OC-3	STS-3	STM-1	150,336	155,520
OC-12	STS-12	STM-4	601,344	622,080
OC-24	STS-24	-	1,202,688	1,244,160
OC-48	STS-48	STM-16	2,405,376	2,488,320
OC-96	STS-96	-	4,810,752	4,976,640
OC-192	STS-192	STM-64	9,621,504	9,953,280
OC-768	STS-768	STM-256	38,486,016	39,813,120
OC-1536	STS-1536	-	76,972,032	79,626,120
OC-3072	STS-3072	STM-1024	153,944,064	159,252,240

Tabla 3.V: Anchos de banda y correspondencias SONET / SDH

La mayor parte de estos mecanismos, pese a presentar características y capacidades tan variadas, van abriéndose camino entre los usuarios con muy diversos resultados. La fidelización de los mismos es una de las grandes preocupaciones de los ISP's, en un mercado en el que, la mayoría de los usuarios no duda en obtener las mejores ofertas de acuerdo a su percepción de la relación velocidad de descarga / precio, muchas veces asociada a servicios adicionales en forma de paquetes *double-play* (TV + Internet, Teléfono + Internet) ó *triple-play* (Teléfono + TV + Internet), e incluso *four-play* (TV + Internet + Teléfono fijo + Teléfono Móvil).

Actualmente en España, el mercado de redes de acceso a Internet está claramente dominado por las diferentes versiones de ADSL y de Modems de Cable, copando entre ambos más del 80 % del mercado, tal como se puede observar en la Fig. 3.15.

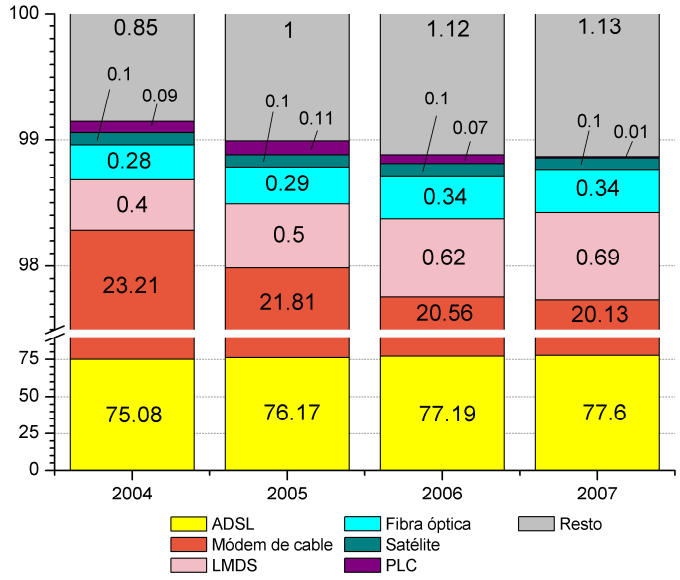


Fig. 3.15: Porcentaje de Cuota de Mercado de las tecnologías de banda ancha más utilizadas en España (fuente:[Telefónica, 2007b])

El resto de tecnologías no dejan de resultar soluciones ocasionales, bien por circunstancias de falta de cobertura de las principales, como es el caso de los sistemas satélite y LMDS, por poca implantación de la tecnología, como es el caso de PLC (*Power-Line Communication*), o por ser tecnologías todavía en fase de implantación, como es el caso de todas las tecnologías *wireless* agrupadas en el concepto "Resto".

La tendencia en España no difiere en gran medida de la que presentan el resto de los países europeos, salvo por razones estructurales en la mayoría de los casos. Así por ejemplo, en España la introducción tardía de las tecnologías de banda ancha y el interés de los usuarios por Internet, hizo que se generalizara el uso de accesos *dial-up*, tal como se puede observar en las Fig. 3.16 y Fig. 3.17.

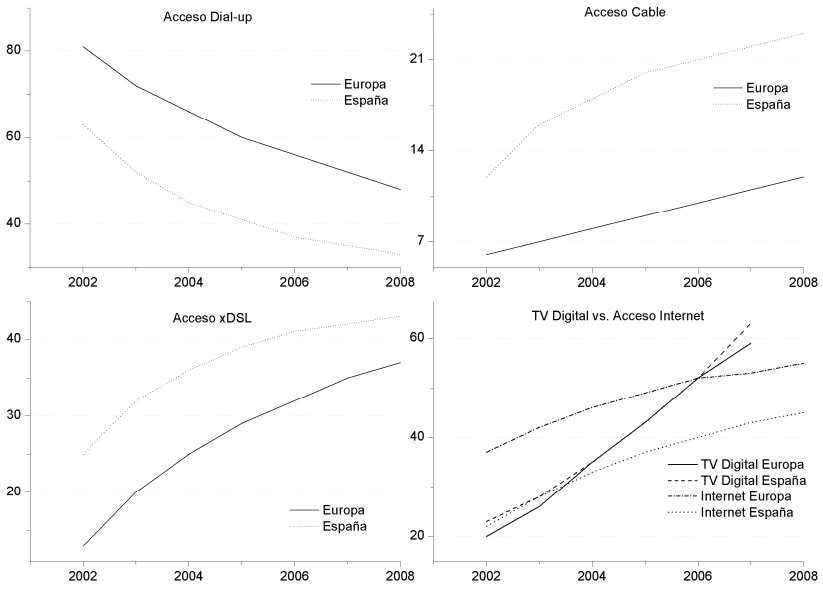


Fig. 3.16: Evolución de la penetración de las tecnologías de acceso más importantes y comparación entre el caso de la Televisión Digital y el Acceso a Internet en el entorno de usuarios residenciales (fuente:[Telefónica, 2007a])

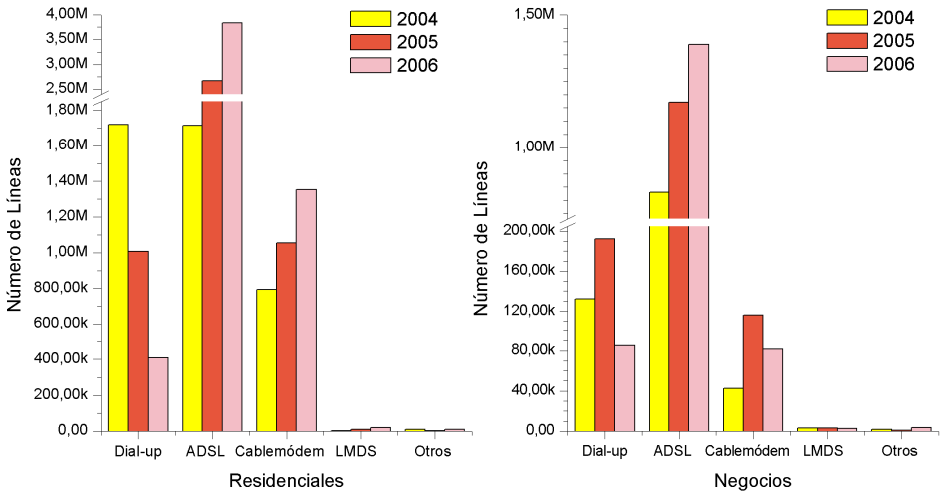


Fig. 3.17: Evolución de número de líneas con servicio de Acceso a Internet en el período 2004-2006, en función del tipo de tecnlogía empleada (fuente:[CMT, 2006])

Conforme se ha generalizado la implantación de dichas tecnologías, este tipo de accesos han dado paso a las tecnologías de cable y xDSL, siendo adoptadas por los antiguos usuarios de forma generalizada, incluso en mayor medida a la media europea. Sin embargo, la incorporación de nuevos internautas sigue siendo menor en España, quizá debido a las diferencias existentes en el coste del acceso a Internet (desde el punto de vista de la relación ancho de banda de descarga/coste), ya que para el acceso a la TV digital, en igualdad de condiciones tecnológicas pero sin coste adicional para el usuario, se iguala a la media europea.

A modo de resumen, la Tabla 3.VI muestra las tecnologías de acceso consideradas actualmente en el modelo CASUAL, con las velocidades actuales típicas, en el lado del usuario, tanto para el sentido de descarga (*downstream*), como para el sentido de subida (*upstream*).

Tecnología de Acceso	Velocidad de acceso	
	Downstream	Upstream
Dial-Up	56 Kbps	33 Kbs
RDSI	128 Kbps hasta 2 Mbps	128 Kbps hasta 2 Mbps
xDSL	20 Mbps	1 Mbps
Cable	27 Mbp	10 Mbps
PLC	2,5 Mbps	2,5 Mbs
LMDS	4 Mbs	4 Mbps
Satélite	8 Mbps	2 Mbps
TDT	3 Mbps	33 Kbps (vía RTC)
WLAN	54 Mbps	54 Mbps
UMTS	2 Mbps	384 Kbps

Tabla 3.VI: Principales tecnologías de acceso a Internet. Datos según oferta de los operadores en 2007

3.2.3. El eje SERVICIOS

En la teoría económica, un servicio es un conjunto de actividades que buscan responder a necesidades de un cliente. Concretamente, en la especificación[ISO, 2007], se define como el resultado de llevar a cabo necesariamente al menos una actividad en la interfaz[24] entre el proveedor y el cliente. Desde el punto de vista de las telecomunicaciones, es el conjunto de

[24] En este caso la definición utiliza el concepto interfaz de forma genérica, como la vía mediante la cual se establece la comunicación entre el cliente y el proveedor.

capacidades y mecanismos específicos que provee un determinado sistema de telecomunicación y que permiten la transferencia de información entre usuarios (humanos y/o máquinas),véase [ITU-T, 1988a].

Para [Shenker, 1995] existen dos componentes claramente diferenciados que definen el conjunto de servicios ofrecidos por una red:

- Modelo de servicio: Es el conjunto básico de servicios ofertados por la red, y es por tanto, la máxima expresión de la arquitectura utilizada. Está basado en los requerimientos fundamentales de las aplicaciones utilizadas, entre los que destacan los modelos para servicios dorsales y para redes de acceso. Desde el punto de vista de flujos individuales se prevén dos tipos de servicios en tiempo real, los predictivos y los garantizados, cada uno de los cuales puede presentar múltiples niveles de servicios elásticos.
- Especificación de servicio: Es la parametrización detallada del modelo de servicio. Incluye la caracterización del servicio ofrecido a cada flujo (por ejemplo los límites de retardo). Detalla tanto la caracterización de la red a partir de cada flujo (por ejemplo *token bucket*, *leaky bucket*, *peak rate*, etc) como la forma en que dicha caracterización se lleva a término (p.e. mediante descarte o retardo de paquetes en los POP o en los *routers*, etc). Mientras el modelo de servicio se obtiene a partir de las características generales, la especificación requiere el conocimiento de un número determinado de opciones detalladas.

De acuerdo con lo anterior, la arquitectura *IntServ* en la RFC1633 considera tres tipos de servicios diferentes, véase [Braden, 1994b]:

- De enlace compartido: Están presumiblemente controlados mediante la correspondiente interfaz de gestión de la red. Puesto que el plano de control no puede ser accedido por las aplicaciones más comunes, no impone restricciones importantes para su coexistencia con el resto de servicios.
- Elásticos: Este tipo de aplicaciones no requieren de mecanismos de control de admisión, ya que no es necesaria la previsión de recursos. Las características de la comunicación son especificadas directamente en la interfaz del protocolo de transporte, para su entrega en la capa de red mediante el marcado de los correspondientes campos de cabecera.
- De tiempo real: Las aplicaciones, en estos casos, deben realizar explícitamente la petición de provisión de dicho servicio, y realizar las reservas de recursos que permitan mantener un flujo de tráfico suficiente, por lo que también reciben el nombre de servicios de *streaming*.

El procedimiento de reserva se realiza en dos pasos, primero la aplicación invoca las interfaces correspondientes del sistema operativo para solicitar la

reserva, y entonces un protocolo de inicialización o de reserva reenvía las solicitudes a la red, la cual devuelve una respuesta. Los *routers* interactúan no con las aplicaciones directamente, sino con el protocolo de reserva. Por su parte, los protocolos de inicialización presentan su propio modelo de servicio para cada configuración de los estados de reserva que proveen.

Por su parte, según [CISCO, 2005], la arquitectura *DiffServ* considera tres tipos fundamentales de servicios:

- *Premium*, denominados *Expedited Forwarding Services* (EF), en los que las características de tráfico exigen requerimientos extremo a extremo más restrictivos (bajo retardo, perdidas reducidas y mínimo *jitter*), normalmente relacionados con comunicaciones en tiempo real, como voz o vídeo. Son directamente controlados por el Control de Acceso y las correspondientes técnicas de ingeniería de tráfico, como las vistas en el apartado 3.1.1.
- Asegurados (*Assured Forwarding Services* - AF): Gracias al marcado individual de cada flujo de tráfico, estos servicios garantizan una mayor fiabilidad y seguridad para aquellos flujos considerados de alta prioridad, siendo su comportamiento predecible frente a los de baja prioridad. A su vez, existen subclases en función del grado de prioridad y de los recursos disponibles en la red.
- *Best-Effort*: Aunque *DiffServ* no los considera como servicios específicos, engloban a todos aquellos flujos que no presentan requerimientos preestablecidos y a todos aquellos Asegurados que no cumplen las características de tráfico contratadas. Actualmente, la mayor parte de las aplicaciones que hacen uso de Internet, generan tráfico considerado como de este tipo.

Sin embargo, es posible categorizar los diferentes usos de Internet, y por tanto el tráfico generado, de forma general e independientemente del tipo de arquitectura de ingeniería de tráfico utilizada, tomando como referencia la aplicación o aplicaciones que le dan origen. En la Fig. 3.18 se muestran la relación existente entre los servicios definidos por las técnicas *IntServ* y *DiffServ*, así como las aplicaciones Internet asociadas a cada clase de tráfico.

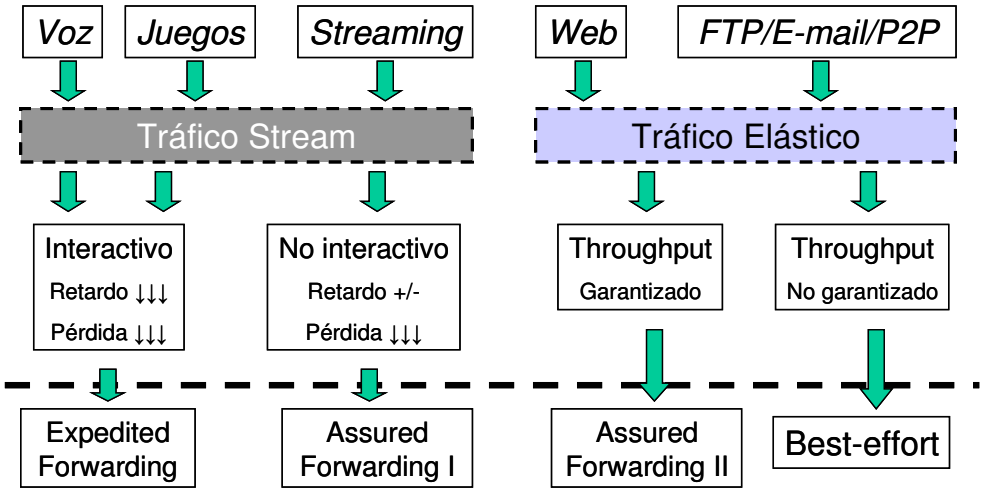


Fig. 3.18: Servicios Internet y su clasificación según IntServ y DiffServ

Así, los servicios de tiempo real, que *IntServ* engloba como *Streaming*, son equivalentes a los EF y parte de los AF de *DiffServ*, mientras que los elásticos aglutinarían a los AF menos restrictivos y a todos los servicios *Best-effort*. Además, su estudio, desde el punto de vista del tráfico generado, puede generalizarse como un conjunto de flujos independientes, tal como se indica en la Fig. 3.19, con tratamientos específicos en función del tipo de servicio y de los mecanismos de ingeniería de tráfico aplicados, especialmente del Control de Admisión y Monitorización, los cuales controlan el reparto correspondiente de las capacidades totales del enlace.

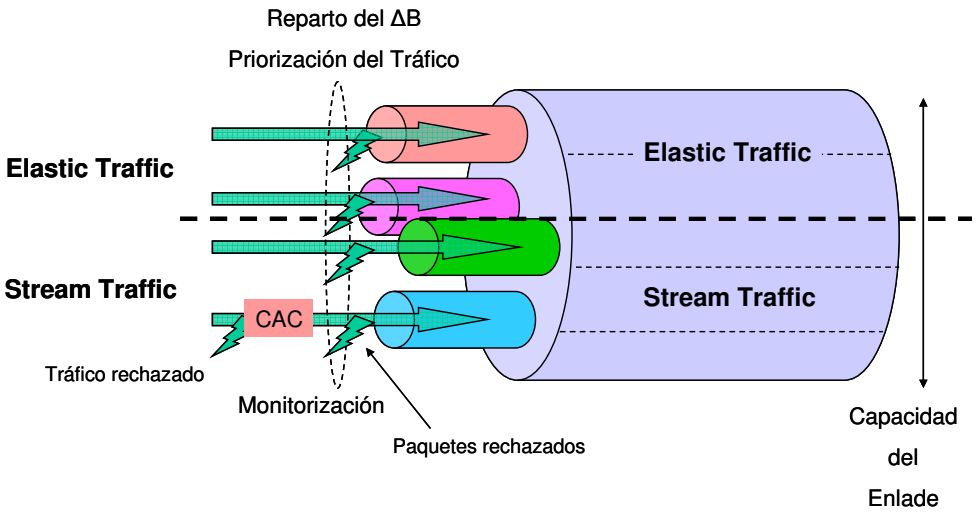


Fig. 3.19: Modelo genérico de reparto de capacidades del enlace Internet.

La relación entre la proporción de tráfico de cada tipo dentro de un mismo enlace depende, en gran medida, del tipo de aplicaciones utilizadas por los usuarios, y por las características concretas de su uso por parte de los mismos. Aquí es donde se establece una relación directa con el tipo de usuario, que selecciona diferentes aplicaciones, y con el acceso, ya que su comportamiento puede variar ampliamente en función de las capacidades que disponga (normalmente ancho de banda).

Así, por ejemplo, las necesidades de los usuarios residenciales van a ser muy diferentes de las de los de negocios, dado que sus intereses son normalmente muy dispares. En [CMT, 2006] se muestra una estimación de los intereses por parte de los internautas en general, y de los internautas de empresas en concreto, y que se resumen en la Fig. 3.20 y Fig. 3.21.

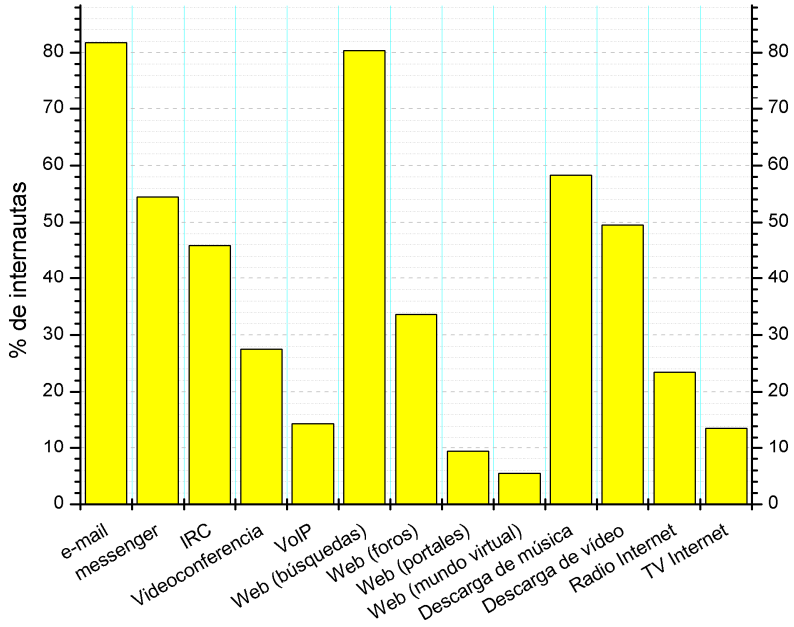


Fig. 3.20: Principales servicios utilizados por los internautas en España (fuente: Telefónica. Datos de 2007)

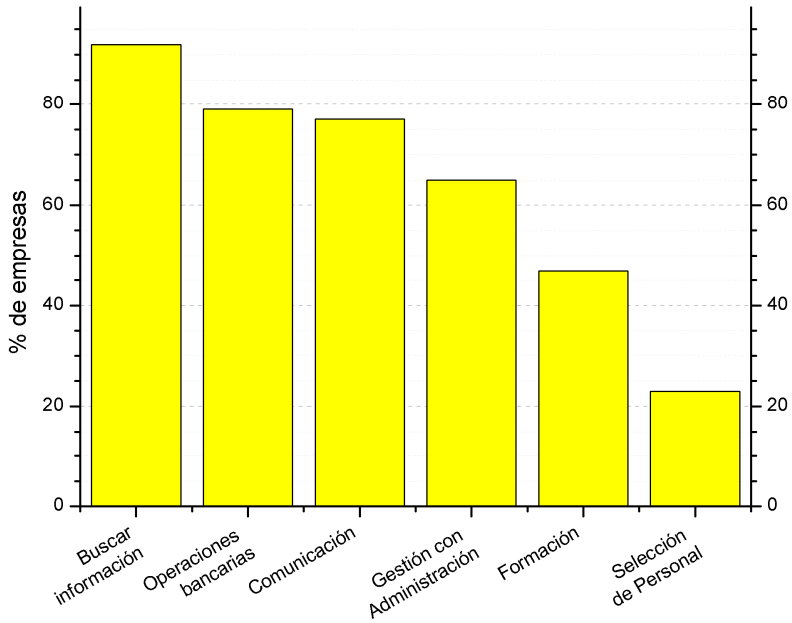


Fig. 3.21: Realización de tareas a través de Internet en las empresas conectadas en España (fuente AETIC/everis, datos de 2006)

El uso lúdico de Internet queda patente en el porcentaje de usuarios que realizan habitualmente operaciones de descarga de archivos multimedia (imágenes, audio, vídeo, etc.), las cuales se ven incrementadas si se tiene en cuenta que, gran parte de las operaciones de búsqueda de información (navegación *Web*) incluyen descargas de este tipo. Por su parte en los entornos empresariales, las tareas realizadas suelen ser una composición de varios servicios Internet, siendo las comunicaciones interpersonales las que más variados mecanismos utilizan (e-mail, videoconferencia, mensajería instantánea o *messenger*, VoIP y *Web*)

Las características del tráfico generado, sin embargo, pese a presentar una tendencia similar, están claramente dominadas por los contenidos, tal como se desprende del estudio del tráfico de la NFSNET en Norteamérica y que se resume en la Fig. 3.22:

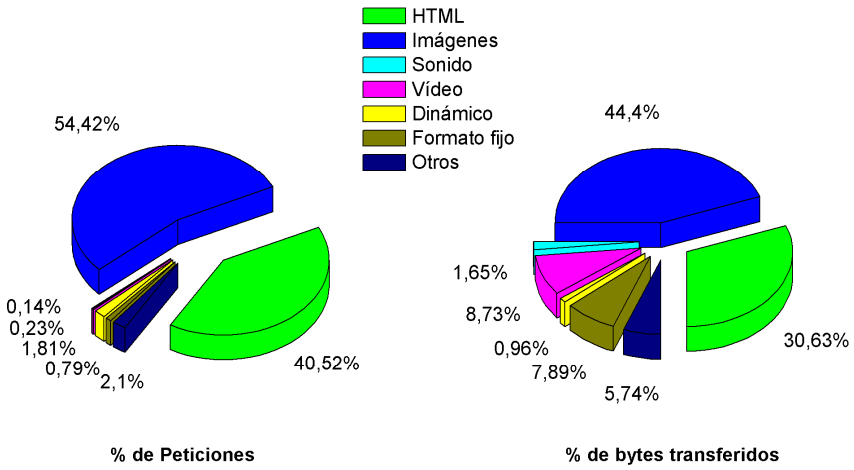


Fig. 3.22: Contenidos asociados a los servicios Internet, tanto desde el punto de vista de número de peticiones, como del volumen de información transferido. (Fuente: NSFNET. Datos de 2006)

Aunque cerca del 95% de las peticiones de contenido realizadas sobre Internet son casi exclusivamente de archivos HTML e imágenes, frente a sólo (todavía) un 0,23% de vídeos, el tráfico generado por aquellas, se reduce hasta el 71% del total, frente a un 8,73% de contenidos de vídeo. Sin embargo, las aplicaciones que generan dicho tráfico, mostradas en la Fig. 3.23, se reducen prácticamente a tres, el tráfico Web (HTTP), el *Peer-To-Peer* (P2P) y el de mensajes (NEWS y correo electrónico), sumando el 92 % del total en la red.

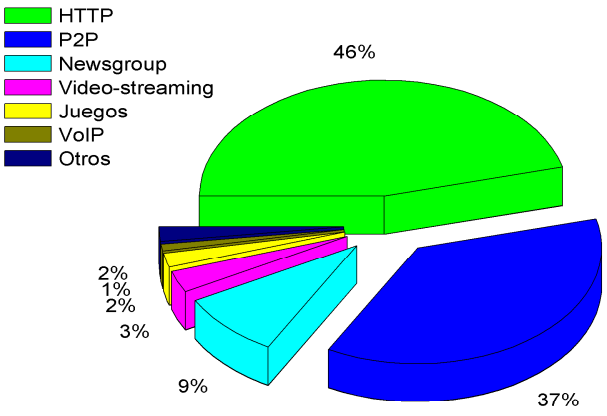


Fig. 3.23: Porcentaje de tráfico de banda ancha asociado a aplicaciones de Internet en Norteamérica. (fuente: Ellacoya Networks / Arbor Networks. Datos de 2007)

El auge del tráfico P2P, como mecanismo de intercambio de contenidos multimedia entre internautas, ha supuesto un reto a los operadores, que han visto cómo sus redes han pasado a duplicar y hasta triplicar el ancho de banda utilizado por sus clientes. A esto hay que sumar lo que ya se ha denominado "Efecto *YouTube*", según el cual se ha disparado el intercambio de contenido multimedia de carácter personal, a través de portales Web, y que explica que los intercambios HTTP sigan siendo superiores a los intercambios P2P, tal como se indica en [Anderson, 2007].

3.3. Modelado de una fuente de tráfico IP mediante modelos ON-OFF multinivel

Tal como se ha explicado a lo largo del capítulo 2, los métodos tradicionales de caracterización del tráfico IP consideran la pila de protocolos TCP/IP como un único sistema de cola, siendo $M/G/\infty$ el más utilizado, véase[Mata, 1994], manteniendo así un modelo de inter-llegadas de Poisson. Sin embargo, en las redes IP, los tiempos entre llegadas no siguen las tradicionales distribuciones exponenciales negativas, ni las fuentes asociadas a los servicios IP siguen el modelo clásico de fuentes de voz modeladas mediante Poisson. Precisamente, la correlación a largo plazo del tráfico IP hace que, al contrario de los modelos basados en Poisson, la autocorrelación sea distinta de cero conforme aumenta el período de observación. Esto es debido a que las características del tráfico a ráfagas y los parámetros estadísticos de utilización de los servicios, son muy diferentes a los de la telefonía tradicional, el cual genera trazas de tráfico con altos valores de autocorrelación para tiempos de observación elevados, y solamente modelables mediante funciones de distribución de cola pesada.

De esta manera, las correcciones a los modelos de Markov, de acuerdo con el concepto de correlación a largo plazo, han evolucionado hacia el modelo autosimilar, es decir, aquellos que independientemente del período temporal de observación tienen un comportamiento semejante. El tráfico IP presenta figuras de autocorrelación que siguen un claro decrecimiento hiperbólico, tal y como se muestra en [Berkeley]. La corrección basada en sistemas $M/G/\infty$, modela la duración asociada al proceso de servicio mediante distribuciones de cola pesada, como Pareto o Weibull[Lombardo, 1998], mucho más cercanas a la naturaleza autosimilar del sistema global. En función del tipo de servicio asociado a la aplicación modelada, el uso de colas pesadas se simplifica mediante exponenciales negativas en los casos más sencillos (de forma similar a los modelos de tráfico telefónico).

Sin embargo, el uso de un sistema $M/G/\infty$ plantea muchas dudas a la hora de valorar su idoneidad, tenida cuenta de la necesidad de modelar un comportamiento autosimilar, tanto a corto como a largo plazo. Dicha característica implica el modelado a lo largo de toda la escala temporal, y no solamente al nivel de llamadas, como es el caso de los modelos de colas

markovianas tradicionales. En general, los autores coinciden en la posibilidad de establecer diferentes puntos de referencia temporal, que tradicionalmente han sido aplicados en el modelado de servicios de transmisión de voz, y que actualmente establecen un marco de referencia viable para el modelado de servicios aplicados a las redes IP, como lo demuestran los modelos ON-OFF multinivel indicados en el apartado 2.3.2.3. En [Rubin, 1995] ya se definen tres escalas temporales de referencia aplicables al caso de la transmisión de voz digital, y que puede ser adaptado al caso de IP, de la misma forma que se hace en [Deng, 1996 / Mah, 1997]:

- Conexión/llamada: Modela el comportamiento entre dos accesos consecutivos a Internet. En el caso de los accesos basados en la red telefónica y/o RDSI, la conexión a Internet comienza con el establecimiento de una llamada al proveedor de servicios Internet. La asociación usuario-proveedor así establecida suele impedir el establecimiento de múltiples conexiones de forma simultánea, salvo excepciones a niveles altos de la red, como es el caso del *bonding* o *bundle*[25]. En el caso de los accesos de Banda Ancha (cable, DSL ó *Ethernet*), la conexión prácticamente comienza en el momento que el PC accede a la red, estableciendo múltiples asociaciones con el ISP. La duración de la conexión ha evolucionado de acuerdo con el desarrollo tecnológico de Internet, en general, y los métodos de acceso al servicio y la situación sociocultural de forma particular, hasta llegar a un escenario de usuarios *always-on*, en los que la conexión se mantiene siempre activa, con continuas sesiones correspondientes a aplicaciones de intercambio de archivos, solamente. Estas sesiones actúan igual que un ruido de fondo sobre el que destacan el resto de servicios iniciados con el usuario delante del ordenador.

- Sesión: Se considera como ejemplo de sesión, la descarga de una única página Web, una conversación de voz, una videoconferencia, etc. En este nivel se modela el tiempo entre dos sesiones/llamadas consecutivas y la duración de cada una. El concepto de agregación de tráfico, simplifica el modelado del escenario multiservicio que supone el acceso a Internet de un único usuario, dado que las características del tráfico asociado a una sesión depende del tipo de servicio y aplicación.

[25] *Bonding*: Técnica de agregación de ancho de banda según la cual se combinan varios circuitos/enlaces físicos independientes de forma combinada, aumentando el ancho de banda total disponible. www.vicomsoft.com/knowledge/reference/bondteam.html

- Ráfaga: Es el nivel de referencia más bajo, donde se modelan los patrones de tráfico dentro de una única sesión, como el tiempo de interllegada de los objetos o ráfagas pertenecientes al servicio. Según los autores, este nivel puede ser desglosado en función de la unidad básica de datos considerada en nivel de paquetes y/o de bit.

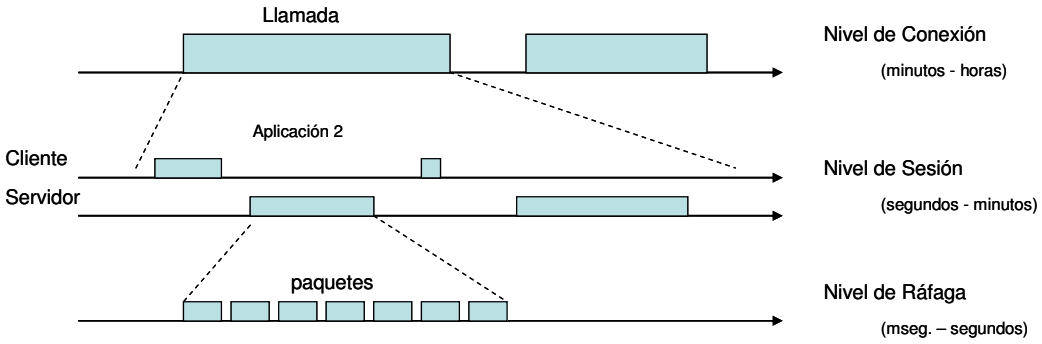


Fig. 3.24: Escalado temporal de tráfico

La Fig. 3.24 muestra, conceptualmente, los tres niveles temporales descritos en [Rubin, 1995] y que fue ampliado por [Charzinski, 2002] al adaptar el concepto a las redes IP y aplicar el conocimiento del comportamiento de los usuarios, el modo de funcionamiento de los servicios utilizados, y los mecanismos asociados a los protocolos correspondientes. De esta manera se identifican hasta seis niveles temporales diferentes, tal como muestra la Fig. 3.25.

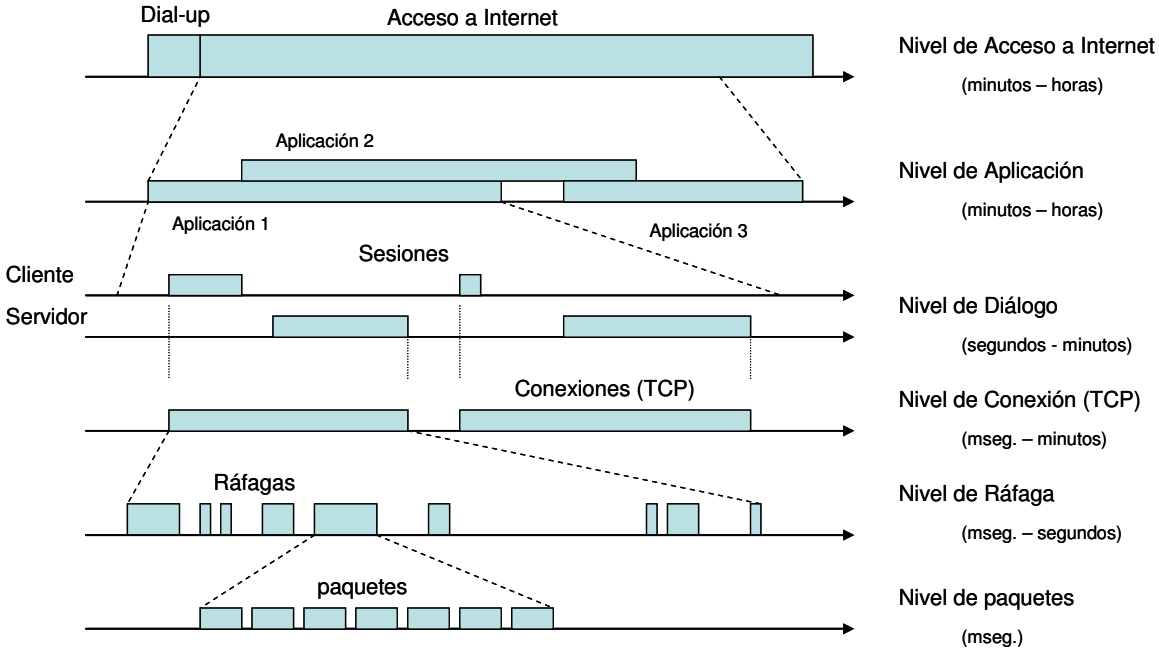


Fig. 3.25: Escalado temporal aplicado al tráfico Internet

El comportamiento del usuario determina la duración de la conexión (nivel de acceso a Internet), eligiendo diferentes servicios y aplicaciones (nivel de aplicación). En función del servicio/aplicación seleccionado, el comportamiento del tráfico será diferente, pudiendo establecerse diálogos diferenciados (nivel de diálogo), a su vez convertidos en conexiones de transporte concretas (nivel de conexión TCP). Por último, en función de las capas de transporte y de red, el intercambio de información se desarrolla con ráfagas de unidades de datos concretas (niveles de ráfaga y de paquete). Sin embargo, la mayoría de los autores han optado finalmente por el uso de modelos temporales más sencillos, volviendo al modelo de tres capas aunque manteniendo las características básicas del modelo de 6 capas. Este es el caso de los modelos de escalado temporal utilizados para la definición de fuentes de tráfico Web (HTTP), véase [Staehle, 2003]. En el nivel de conexión, el servicio Web accede a Internet a través del ISP, ya sea de forma temporal o permanente. Una vez conectado al ISP, en el nivel de sesión, el usuario "navega" realizando diferentes peticiones de páginas Web, cada una de las cuales incluyen la transmisión de múltiples objetos (imágenes, sonidos, texto, scripts, etc). A su vez, en el nivel de ráfaga, cada página es transferida en forma de ráfagas de paquetes IP.

Esta visión temporal es la misma que se expone en el apartado 2.3.2.3, como mecanismo aplicado para el modelado de la agregación de tráfico IP mediante el uso de fuentes ON-OFF multinivel. En este estudio se propone el uso de estas fuentes para el modelado de flujos individuales de tráfico IP, introduciendo todas las características asociadas en cada uno de los puntos temporales correspondientes. Desde el punto de vista del comportamiento del tráfico generado por una sola fuente, se considera de forma independiente cada nivel temporal y, de esta manera, se modelan los tres mediante el uso de modelos de tráfico más o menos simples. La solución propuesta hace uso de tres fuentes ON-OFF interrelacionadas a través de las correspondientes escalas temporales. Partiendo desde el nivel de sesión, el comportamiento del usuario y de la conexión Internet a medio y largo plazo es modelado mediante una fuente, en la que el estado ON está a su vez modulado por el comportamiento del usuario y de la aplicación a corto y medio plazo. A su vez, el estado ON del nivel de sesión se encuentra a su vez modulado por otra fuente correspondiente al nivel de ráfaga, y por tanto, el comportamiento de la pila de protocolos TCP/IP.

El concepto de escalabilidad temporal y la aplicación de modelos ON-OFF para cada nivel de escala se resume de acuerdo con la Fig. 3.26:

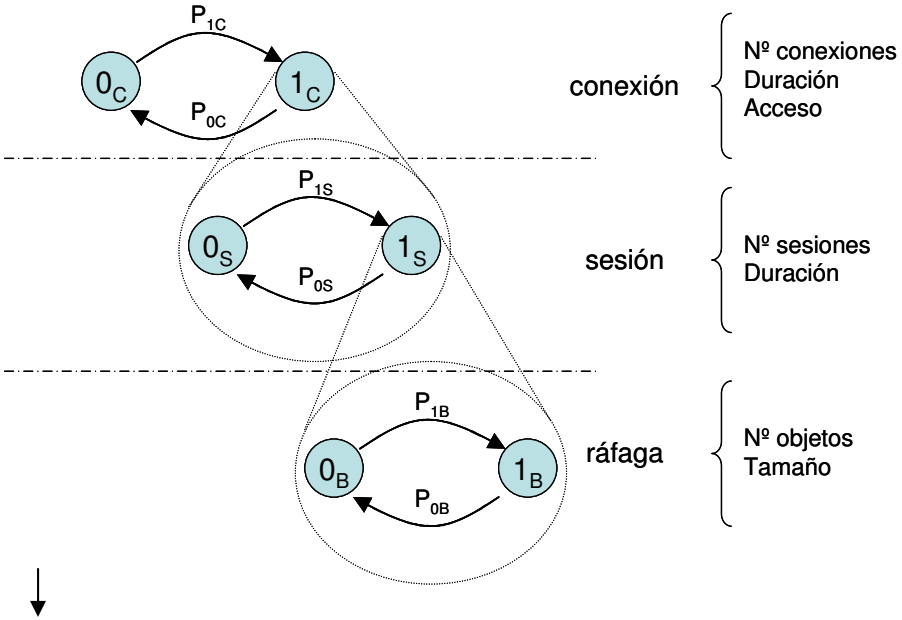


Fig. 3.26: Modelo ON-OFF multinivel

- Conexión: La fuente modela el acceso a Internet a través del ISP, con lo que los estados 0 y 1 representan respectivamente la desconexión/conexión al ISP. Los parámetros fundamentales que permiten modelar la fuente son: el tiempo entre peticiones de conexión y la duración de las mismas. Ambos valores presentan una variedad muy elevada en función del comportamiento asociado a determinados usuarios, y especialmente al tipo de tecnologías de acceso utilizadas por estos.

- Sesión: Modela el conjunto de peticiones y transferencias, así como los tiempos de espera durante cada descarga, por ejemplo, de una página Web. Los estados 0 y 1 representarían los períodos de inactividad entre sesiones y el conjunto de peticiones y transferencias asociadas a cada sesión, comunicación de voz, página Web, descarga FTP, etc. Los parámetros fundamentales del modelo son el tiempo entre sesiones, también denominado en algunos casos como *thinking time*, así como la duración de cada sesión. En este caso, la influencia del comportamiento de los usuarios es menos importante, siendo el tipo de servicio/aplicación el factor determinante. El tipo de aplicación determina fuertemente el tamaño de los objetos transferidos y, por tanto, la longitud y duración de las ráfagas de tráfico resultantes.

- Ráfaga: La fuente modela cada una de las peticiones que componen cada sesión. Los estados 0 y 1 representan respectivamente el denominado tiempo *idle* o latencia entre paquetes y el tiempo de transferencia de cada uno. En este caso los parámetros fundamentales

del modelo son el tiempo entre paquetes, determinado normalmente mediante confirmaciones, y el tiempo de servicio asociado a cada uno.

Hay que observar que, solamente el estado ON de las ráfagas, se genera una carga efectiva de paquetes IP, denominada flujo, que es la que se transmite a través de una conexión virtual, por ejemplo mediante MPLS.

El cálculo de las capacidades efectivas, asociadas a una conexión (correspondiente a un servicio utilizado por un determinado tipo de usuario conectado con una tecnología de acceso dada), se resuelve sin la necesidad de conocer el número de transiciones de un estado a otro para cada nivel temporal, es decir, el número de "objetos" (en general: conexiones, sesiones, ráfagas) intercambiados, sino que basta con conocer, estimar o fijar las probabilidades asociadas a cada estado. Estos valores son determinados a partir de los tiempos totales medios de permanencia en cada estado.

Sin embargo, para poder llevar a cabo dicho cálculo, es preciso resolver nivel a nivel el comportamiento de cada una de las fuentes ON-OFF consideradas. Para ello, se debe realizar un análisis *Top-Down* del modelo multinivel, tomando como punto de partida el usuario y su comportamiento, esto es, desde el nivel de conexión. A partir de aquí se establecen las relaciones y dependencias entre el comportamiento interno de la conexión y el nivel de sesión. Una vez resueltas, se repite el proceso, pero esta vez para establecer las dependencias entre el nivel de sesión y el de ráfaga.

En los apartados siguientes se desarrolla todo el proceso anterior, en el que se ha tenido en cuenta que para cada nivel temporal se obtiene, en cada caso, las figuras de tráfico asociadas, expresado por las velocidades medias correspondientes a cada nivel, así como sus correspondientes varianzas. Además, el cálculo de las velocidades asociadas a un determinado nivel, por ejemplo el de conexión, se realiza sin más que conocer las probabilidades de activación y desactivación de la fuente ON-OFF correspondiente.

En el caso de un servicio en el que predomina la descarga de datos (*download*), para obtener los valores de capacidad correspondientes, es necesario resolver primero las velocidades correspondientes al nivel temporal inmediatamente inferior, esto es, los estadístico del nivel de conexión solo pueden ser obtenidos a partir de los del de sesión, que a su vez dependen del de ráfagas. Es por ello que, si bien el desarrollo del modelo es *top-down*, la resolución del mismo es *bottom-up*, siendo el punto de partida la velocidad física del acceso al punto de agregación, a partir de ahora abreviado como v_{ac}, que se toma como referencia del estado ON del nivel de ráfaga. En función del tipo de acceso y de los protocolos de enlace implicados, incluso cuando no se está transmitiendo directamente datos (ON), puede incluirse tráfico de señalización, control, o simplemente de sincronización, por lo que, aparece un parámetro v_{mi}, es decir, la velocidad mínima asociada al estado OFF del nivel de ráfaga. Sin embargo, en la mayoría de las tecnologías de acceso el valor de v_{min} es despreciable o cero,

por lo que no ha sido considerado a lo largo del desarrollo que sigue más adelante.

En el caso contrario, en el de los servicios con predominio del tráfico de subida (*upload*), o, como en el caso de la VoIP, con simetría en ambos flujos, el nivel determinante es el de sesión, que es el que establece la velocidad de generación del flujo de bytes y el tamaño máximo de los paquetes donde se encapsula.

Por otro lado, desde el punto de vista del usuario, un servicio genera dos flujos de tráfico, uno en cada sentido de la comunicación (comúnmente denominados *upstream*, para el que procede del propio usuario, y *downstream*, para el que recibe desde Internet). Según las figuras Fig. 3.24 y Fig. 3.25, esta diferenciación se produce en el nivel de sesión o de diálogo respectivamente. Esto significa que, aunque las características de la conexión son comunes a ambos flujos, el comportamiento durante la sesión y de las ráfagas no tiene por que coincidir. Tal es el caso de los servicios con diálogos asimétricos, donde el cliente realiza peticiones de objetos (páginas Web, ficheros, mensajes, etc.), de tamaño variable pero, en comparación con la información que recibe, de tamaño reducido. En estos casos, las capacidades requeridas para el *upstream* y el *dowstream* también son diferentes, por lo que, el servicio asociado a un usuario determinado debiera modelarse mediante dos fuentes ON-OFF multinivel con estadísticas simétricas al nivel de conexión, pero con valores específicos en los de sesión y ráfaga, en función del tipo de servicio (de la asimetría del diálogo) y del tipo de acceso (si es asimétrico, como en ADSL, o simétrico, como en RDSI o SDSL). En muchos casos, que el ancho de banda del acceso físico es simétrico, o bien la existencia de una marcada asimetría entre ambos flujos, el cálculo se reduce a la estimación del más restrictivo, es decir, el de mayores requerimientos. En general, dada las características de la mayoría de las aplicaciones en Internet, basadas en la filosofía cliente-servidor, una fuente correspondiente a un cliente queda caracterizada por su capacidad efectiva en el *downstream*, mientras que, en el caso de un servidor, por la del *upstream*.

Sin embargo, la caracterización del upstream/dowstream resulta mucho más compleja si se tienen en cuenta las características particulares de cada servicio, y que modifican el comportamiento del modelo ON-OFF multinivel, haciendo necesario definir mecanismos distintos para su aplicación. En concreto, se identifican dos tipos de comportamiento diferenciado: el de las aplicaciones controladas por el nivel de conexión, y por tanto, por el comportamiento del usuario, y el de las que la aplicación toma el control del nivel de sesión. En los siguientes apartados se desarrolla el análisis, nivel a nivel, de dos fuentes ON-OFF multinivel particularizadas para modelar dos servicios genéricos que presenten dichos comportamientos.

3.3.1. Caso 1: Servicios controlados por el nivel de conexión

El tráfico generado por la gran mayoría de los servicios Internet depende casi exclusivamente del uso que el usuario hace de ellos, ya que los niveles de sesión y de ráfaga actúan como mero transporte de la información. El usuario decide cuándo activar el servicio, la duración y el tipo y tamaño de los datos intercambiados. Las aplicaciones (ya sean clientes o servidores) se limitan a pasar los datos a TCP/IP, que realiza su empaquetado y posterior transmisión por la correspondiente interfaz. En esta situación, la velocidad asociada al nivel de conexión es calculada utilizando las probabilidades de activación correspondiente a la fuente ON-OFF y la velocidad total del nivel de sesión (v_s).

$$v_{mc} = p_{1c} v_s \tag{3.2}$$

Y su desviación típica:

$$\sigma_{conexion} = \sqrt{p_{1c} v_s^2 - v_{mc}^2} \tag{3.3}$$

Entonces, la velocidad total del nivel de conexión se calcula como:

$$v_c = \min\left(v_{mc} + \sigma_{conexion}\gamma, v_s\right) \tag{3.4}$$

siendo γ un factor de ponderación seleccionado de acuerdo al intervalo de confianza deseado. La elección de este valor debe realizarse previamente, y muy posiblemente apoyada por resultados obtenidos mediante simulación. Sin embargo, la mayor parte de los autores coinciden en utilizar este factor, siempre con valor 2 ó 3[26].

De la misma forma, en el nivel de sesión, la velocidad media se calcula a partir de las probabilidades de activación del estado ON y la velocidad total del nivel de ráfaga v_r:

$$v_{ms} = p_{1s} v_r \tag{3.5}$$

$$\sigma_{sesion} = \sqrt{p_{1s} v_r^2 - v_{ms}^2} \tag{3.6}$$

La velocidad total del nivel de sesión será entonces:

[26] En el caso de la distribución normal, la constante γ toma los valores de 1, 2 ó 3 para asegurar rangos de confianza del 68, 95 y 99.73% respectivamente. En el caso de distribuciones genéricas, de acuerdo con el Teorema de Tchebysheff, el grado de confianza alrededor del valor medio toma el valor $100(1-1/\gamma^2)$%, esto es, el 75 y el 89 % para valores de γ de 2 y 3 respectivamente.

$$v_s = \min\left(v_{ms} + \sigma_{sesion}\gamma, v_r\right) \tag{3.7}$$

Una vez en el nivel de ráfaga, se vuelve a utilizar la probabilidad de activación del estado ON inferior. La velocidad en dicho estado coincide, bien con la velocidad máxima del enlace (v_{ac}), o bien la velocidad fijada por el conjunto de protocolos que definen el servicio:

$$v_{mr} = p_{1r} v_{ac} \tag{3.8}$$

$$\sigma_{rafaga} = \sqrt{(p_{1r} v_{ac}^2 - v_{mr}^2)} \tag{3.9}$$

Y finalmente, la velocidad total al nivel de ráfaga será:

$$v_r = \min\left(v_{mr} + \sigma_{rafaga}\gamma, v_{ac}\right) \tag{3.10}$$

Este cálculo, aplicado a una sola fuente, aunque didáctico, no tiene una aplicación práctica clara, ya que el cálculo de las velocidades a partir de los dos primeros momentos solamente se justifica cuando se modelan múltiples fuentes, con lo que su comportamiento se aproxima a una distribución normal, comportamiento asociado al uso del factor *γ*. En cualquier caso, para el modelado de una fuente individual se debe utilizar exclusivamente una aproximación comprendida entre los valores medios y máximos.

Para el cálculo de la velocidad total requerida en la conexión se comienza a resolver, entonces, en el nivel de ráfaga, tal como se indicó al principio.

Una vez calculada la vc para una sola fuente, la capacidad total estimada para un agregado de M fuentes sería:

$$V_T = M v_c + \gamma\sqrt{M \operatorname{var}(v_c)} \tag{3.11}$$

Esa misma expresión puesta en función del coeficiente de variación de la velocidad de conexión queda:

$$V_T = M \cdot v_c \left[1 + \frac{C(v_c)}{\sqrt{M}}\gamma\right] \tag{3.12}$$

de donde la capacidad equivalente para una fuente se calcula como:

$$V_{eq} = \frac{V_T}{M} = v_c \left[1 + \frac{C(v_c)}{\sqrt{M}}\gamma\right] \tag{3.13}$$

Es decir, una vez conocidos los dos primeros momentos del nivel de sesión de una fuente individual y el número de fuentes a agregar, es posible calcular la capacidad total de un conjunto de fuentes idénticas y la capacidad equivalente asociada a cada una de ellas.

A modo de ejemplo de aplicación, se realiza el cálculo sobre una fuente genérica que en la hora cargada realiza una conexión durante la cual descarga una media de unos 4 objetos (n_{obj}) de 41 KBytes[27] (L_{obj}) cada uno. El tiempo entre las diferentes peticiones, durante el cual la fuente se encuentra inactiva (tiempo de *thinkingtime*) es de unos 60 segundos, la velocidad del acceso al punto de agregación es de V_{acc}=150 Mbps (aproximadamente la capacidad al nivel de aplicación de un STM-1). En estas condiciones se desea estudiar la tendencia de la capacidad efectiva de la fuente en función de la capacidad de agregación del proveedor. Bajo el supuesto de que la fuente pertenezca a un servicio típico de descarga, TCP/IP realiza el empaquetado de los objetos de aplicación ajustados a la MTU (*Maximum Transfer Unit*) del acceso, en el lado del usuario, o del punto de agregación en el lado del proveedor. Esto significa que el tamaño de los datagramas será típicamente de 1500 bytes (L_{IP}), de los cuales 1460 corresponden a la aplicación L_{pac} y 40 corresponden a sobrecarga (20 de TCP y 20 de IP, o bien 28 si se utiliza UDP en vez de TCP). Con estos datos las probabilidades de activación de la fuente ON-OFF al nivel e ráfaga se calcula como:

$$p_{1r} = \frac{\frac{L_{IP}}{V_{acc}}}{\frac{L_{IP}}{V_{acc}} + \frac{RTT \cdot L_{pac}}{L_{obj}}} \tag{3.14}$$

donde el RTT (*Round Trip Time*) representa el tiempo durante el cual la fuente espera como mínimo antes de mandar una nueva petición (bien por intercambio de confirmaciones, o por temporización). Aunque este valor depende de múltiples factores, en [Huffaker, 2002] se concluye que su valor en función de la distancia entre cliente y servidor está en el rango de 100 a 200 milisegundos (de 5000 a 15000 kilómetros). En este ejemplo se toma 200 milisegundos para eliminar la distancia como factor en los resultados. La velocidad del nivel de ráfaga será:

$$v_{mr} = p_{1r} V_{acc} \tag{3.15}$$

Para una fuente al nivel de sesión resulta:

$$v_{ms} = p_{1r} V_{acc} = \frac{\frac{L_{IP}}{V_{acc}} + \frac{RTT \cdot L_{pac}}{L_{obj}}}{T_{think} \frac{L_{obj}}{L_{pac}} + \frac{L_{IP}}{V_{acc}} + \frac{RTT \cdot L_{pac}}{L_{obj}}} \cdot v_{mr} \tag{3.16}$$

y en en el nivel de conexión:

[27] Aunque en la definición de los parámetros, todos los tamaños de objetos, bloques y paquetes se dan en Bytes, el valor utilizado en las expresiones se da en bits

$$v_{mc} = p_{1c} V_{acc} = \frac{n_{obj}}{3600}\left(T_{think} + \frac{L_{IP}}{V_{acc}} + \frac{RTT \cdot L_{Pac}}{L_{obj}} \right) \cdot v_{ms} \tag{3.17}$$

Si ahora se aplican los resultados para un agregado de M fuentes idénticas, el ancho de banda agregado será:

$$V_T = \min\left(M v_{mc} + \gamma \sqrt{M \ \mathrm{var}(v_{mc})}, M V_{acc} \right) \tag{3.18}$$

Con los datos de la fuente descrita anteriormente, al variar el número de fuentes desde 1 hasta 1000 se obtienen los resultados de la Tabla 3.VII.

Esos mismos resultados aparecen reflejados en la Fig. 3.27, donde se ha completado el ejemplo al repetir los cálculos para diferentes valores de γ, y se han añadido, a modo de referencia los valores medios de la fuente para cada uno de los niveles contemplados.

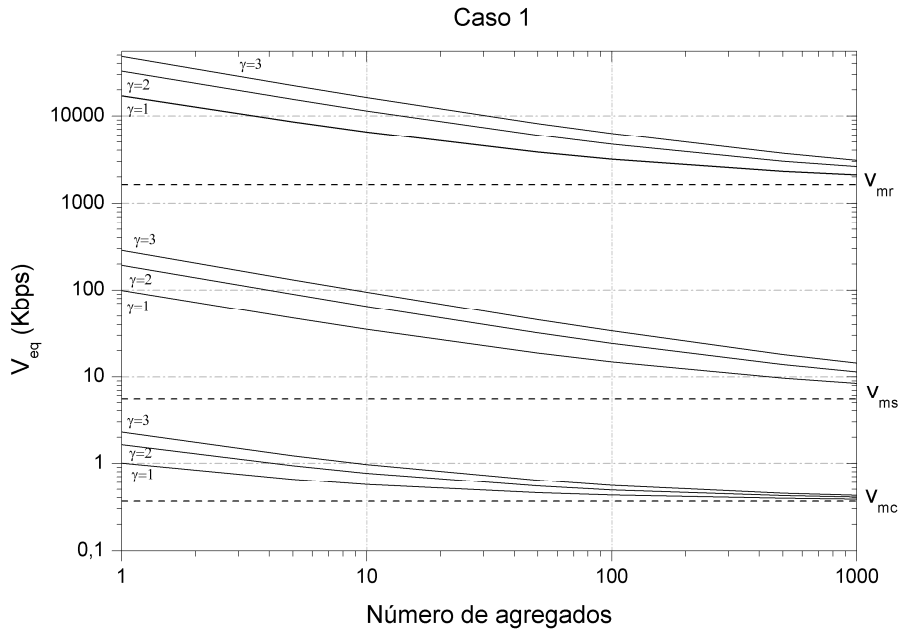


Fig. 3.27: Caso 1 - Estadística de la capacidad equivalente en función del número de fuentes agregadas y del factor γ

Nivel de Ráfaga (γ=3)

Nº de fuentes	v_{mr}	$var(v_{mr})$	V_T	V_{eq}	$C(v_{mr})$
1	1626,02	2,41E+08	48223,51	48223,51	9,55
5	8130,08	1,21E+09	112325,25	22465,05	4,27
10	16260,16	2,41E+09	163614,38	16361,44	3,02
50	81300,81	1,21E+10	410794,86	8215,90	1,35
100	162601,63	2,41E+10	628576,58	6285,77	0,96
500	813008,13	1,21E+11	1854959,81	3709,92	0,43
1000	1626016,26	2,41E+11	3099558,45	3099,56	0,30

Velocidades en Kbps

Nivel de Sesión (γ=3)

Nº de fuentes	v_{ms}	$var(v_{ms})$	V_T	V_{eq}	$C(v_{ms})$
1	5,46	8849,93	287,68	287,68	17,23
5	27,31	44249,66	658,37	131,67	7,70
10	54,61	88499,32	947,08	94,71	5,45
50	273,05	442496,58	2268,66	45,37	2,44
100	546,10	884993,15	3368,33	33,68	1,72
500	2730,52	4424965,76	9041,21	18,08	0,77
1000	5461,05	8849931,52	14385,70	14,39	0,54

Velocidades en Kbps

Nivel de Conexión (γ=3)

Nº de fuentes	v_{mc}	$var(v_{mc})$	V_T	V_{eq}	$C(v_{mc})$
1	0,37	0,41	2,29	2,29	1,75
5	1,83	2,05	6,13	1,23	0,78
10	3,65	4,11	9,73	0,97	0,55
50	18,26	20,53	31,86	0,64	0,25
100	36,53	41,07	55,75	0,56	0,18
500	182,65	205,34	225,64	0,45	0,08
1000	365,30	410,67	426,09	0,43	0,06

Velocidades en Kbps

Tabla 3.VII: Caso 1: Resultados del modelo ON-OFF multinivel

Se observa como el efecto de la agregación hace que, conforme aumenta el número de fuentes, la capacidad equivalente converja rápidamente hacia el valor medio, especialmente en el nivel de conexión.

El ejemplo presentado toma para modelar sus parámetros valores típicos de servicios reales. En concreto, la fuente genérica presentada podría corresponderse perfectamente con una aplicación *Web*, concretamente con el flujo *downstream* medida en el punto de agregación.

Finalmente mencionar que la capacidad equivalente al nivel de conexión permite en una planificación maximizar el número de usuarios que se pueden agregar conocida la velocidad de acceso, o a la inversa, dado el número de usuarios, determinar la velocidad de acceso requerida. Los valores al nivel de sesión y ráfaga sirven sobre todo para el análisis de rendimiento (*performance analysis*), como por ejemplo, el cálculo del retardo y el *jitter*, o el dimensionado de las colas.

3.3.2. Caso 2: Servicios controlados por el nivel de aplicación

En el otro lado del conjunto de servicios Internet, se encuentran las aplicaciones interactivas y multimedia. La transmisión de voz, vídeo y los juegos en red implican un intercambio continuo de información entre ambos extremos de la comunicación. Precisamente la naturaleza de esa información, hace que, en muchas ocasiones, las aplicaciones tengan que acomodar los datos a una corriente de tráfico con propiedades específicas. El usuario no tiene control directo sobre las características del tráfico que genera, es la aplicación la que toma el control antes incluso de que TCP/IP o UDP/IP conforme el correspondiente flujo de paquetes. Esta situación se da en todos los servicios con requerimientos de tiempo real o, al menos, restricciones en el valor del retardo, como es el caso de VoIP. La aplicación consta de un digitalizador de audio que comprime y empaqueta la voz hasta obtener un flujo de tráfico de velocidad prácticamente constante, donde los paquetes presentan estructuras y tamaños fijos. En este caso, el tráfico no puede modelarse inicialmente desde nivel de conexión, ya que es el de sesión el que toma el control sobre el nivel de ráfaga. El nivel de sesión solamente hace uso de la corriente de paquetes que el de sesión le ofrece. En esta situación, el cálculo de los parámetros del modelo ON-OFF multinivel deben realizarse siguiendo una secuencia distinta a la del caso genérico, explicado en el apartado anterior.

Igual que antes, se utiliza un ejemplo para hacer más transparente la aplicación del modelo ON-OFF multinivel en este tipo de casos. Sea una fuente de voz comprimida, con una velocidad de compresión de 16 Kbps, períodos de actividad y silencio, de 200ms cada uno, y que empaqueta la voz en bloques de tamaño fijo, 100 bytes y 40 bytes de cabeceras. En estas condiciones, el nivel de sesión está dirigido por la generación de paquetes de voz, por lo que su velocidad media se puede calcular como:

$$v_{ms} = p_{1s} \frac{L_{IP}}{L_{app}} V_{com} \tag{3.19}$$

Ahora, el nivel de ráfaga presenta una velocidad igual a:

$$v_{mr} = \frac{L_{IP}}{L_{app}} V_{acc} \tag{3.20}$$

Y para el de conexión resulta nada más que:

$$v_{mc} = v_{mr} \alpha T \tag{3.21}$$

Conocido el número de llamadas en la hora cargada (α=2), y su duración media *T*=2 minutos.

Al estudiar la agregación de fuentes de este tipo, resulta la

Velocidades en Kbps

Tabla 3.VIII:

Nivel de Ráfaga (γ=3)

Nº de fuentes	v_{mr}	$var(v_{mr})$	V_T	V_{eq}	$C(v_{mr})$
1	22,40	992280,58	2000,00	2000,00	44,47
5	112,00	4961402,88	6794,26	1358,85	19,89
10	224,00	9922805,76	9674,15	967,41	14,06
50	1120,00	49614028,80	22251,17	445,02	6,29
100	2240,00	99228057,60	32123,98	321,24	4,45
500	11200,00	496140288,00	78022,62	156,05	1,99
1000	22400,00	992280576,00	116901,46	116,90	1,41

Velocidades en Kbps

Nivel de Sesión (γ=3)

Nº de fuentes	v_{ms}	$var(v_{ms})$	V_T	V_{eq}	$C(v_{ms})$
1	11,20	2,56	16,00	16,00	0,14
5	56,00	12,80	66,73	13,35	0,06
10	112,00	25,60	127,18	12,72	0,05
50	560,00	128,00	593,94	11,88	0,02
100	1120,00	256,00	1168,00	11,68	0,01
500	5600,00	1280,00	5707,33	11,41	0,01
1000	11200,00	2560,00	11351,79	11,35	0,00

Velocidades en Kbps

Nivel de Conexión (γ=3)

Nº de fuentes	v_{mc}	$var(v_{mc})$	V_T	V_{eq}	$C(v_{mc})$
1	1,12	54,79	23,33	23,33	6,61
5	5,60	273,96	55,26	11,05	2,96
10	11,20	547,92	81,42	8,14	2,09
50	56,00	2739,61	213,02	4,26	0,93
100	112,00	5479,22	334,07	3,34	0,66
500	560,00	27396,10	1056,55	2,11	0,30
1000	1120,00	54792,19	1822,23	1,82	0,21

Velocidades en Kbps

Tabla 3.VIII: Caso 2: Resultados del modelo ON-OFF multinivel

Al igual que antes la Fig. 3.28 se muestra la capacidad equivalente en función del número de fuentes agregadas y del valor de γ para cada uno de los tres niveles:

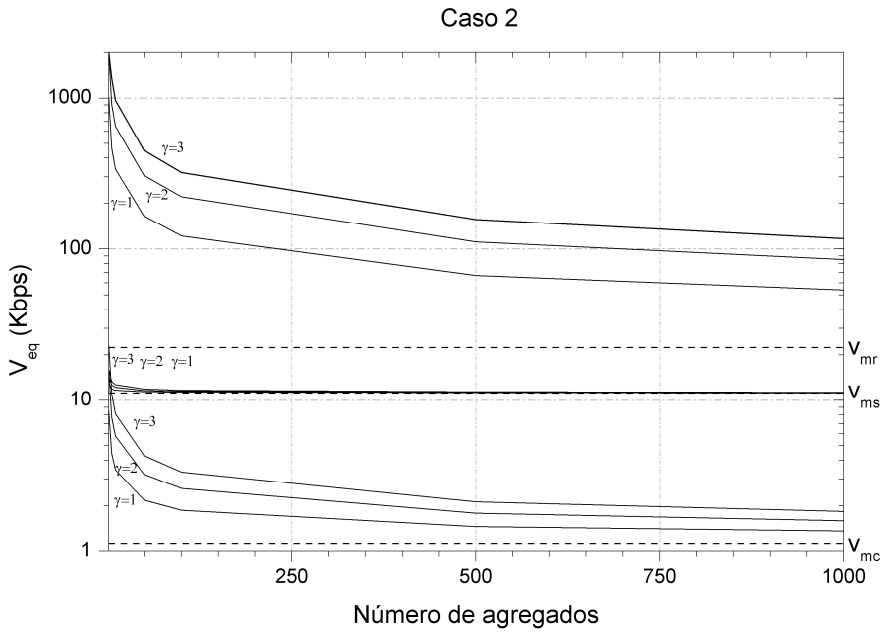


Fig. 3.28: Caso 2 – Estadística de la capacidad equivalente en función del número de fuentes agregadas y del factor γ

Se observa como, en este caso, la aplicación controla el flujo de paquetes, de tal manera que la capacidad equivalente en el nivel de sesión es prácticamente igual que la tasa de generación. Además, al igual que en las fuentes del caso 1, en el nivel de conexión el efecto de la agregación hace que el valor de la capacidad equivalente se acerque rápidamente hasta el valor medio conforme aumenta el número de fuentes. Este no es el caso del nivel de ráfaga, debido a que aquí no solamente se produce una agregación de tráfico multifuente para un solo servicio, sino multifuente multiservicio, que no ha sido considerada en estos ejemplos.

En cuanto al factor γ, en general, conforme aumenta el número de fuentes, las diferencias también se reducen. Desde el punto de vista de un operador, la elección de su valor no se realiza en función del grado de confiabilidad de la estimación. Al utilizar estas herramientas en estudios estratégicos, sus intereses se centran en los ratios entre los costes que supone proveer esos valores de ancho de banda y los que aseguran un grado de servicio suficiente, aceptable u óptimo, desde el punto de vista del usuario.

3.3.3. **Modelado de una fuente de tráfico IP mediante curvas de llegada**

En general, el flujo que puede transportar un determinado enlace puede ser caracterizado por tres parámetros fundamentales: su tasa máxima, la tasa máxima sostenible y el tamaño mínimo de ráfaga. Según [Boudec, 2000a],

"cuando se tienen en consideración valores de retardo determinísticos, siempre existe un valor de tasa máxima, del que se puede deducir el denominado ancho de banda efectivo, para cualquier tipo de tráfico de llegada. Además, siempre y cuando la función de coste[28] asociada al enlace sea lineal, es posible encontrar un algoritmo explícito para la obtención de las características del mismo". Desde el punto de vista de una fuente IP, sus características se ven modificadas por los parámetros básicos del enlace, o expresado de acuerdo con la teoría del *Network Calculus* del apartado 2.4, las características del enlace definen una curva de servicio a la que, las diferentes curvas de llegada, van a acomodarse. Supuesto un retardo máximo D, la capacidad mínima para la que el enlace ofrezca una curva de servicio que no altere las características de la curva de llegada procedente de una fuente IP, sigue la condición $C \geq C_D$, donde C_D es el ancho de banda efectivo de una fuente IP genérica, de forma que el ancho de banda efectivo correspondiente a la agregación de N fuentes sea igual a $N \cdot C_D$.

Tal como se muestra en la Fig. 3.29, dada una fuente IP con una curva de llegada $\alpha(t)$, un enlace de capacidad C ofrece una curva de servicio $\lambda_t(t)=Ct$, que asegura que $\alpha(t) \leq C(t+D)$ para todo $t \geq 0$, esto es, la capacidad cumple la siguiente condición:

$$C \geq \sup_{t \geq 0} \left\{ \frac{\alpha(t)}{t+D} \right\} = e_D(\alpha) \tag{3.22}$$

Siendo $e_D(\alpha)$ el ancho de banda efectivo de la curva de llegada α.

[28] En el Network Calculus se utiliza como factor de coste el ratio entre los costes reales para aumentar la capacidad del enlace y el backlog frente a los beneficios de relajar los requerimientos de retardo.

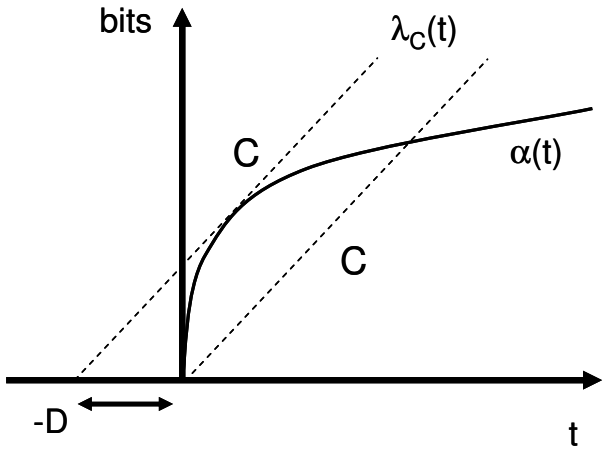


Fig. 3.29: Determinación de la curva de servicio de un enlace que asegure que un flujo α no sufra un retardo máximo de D

Según [Boudec, 1998], en el caso de que $\alpha(t)$ fuera diferenciable, e_D es calculada como la pendiente de la recta tangente a la curva de llegada que pasa por el punto *(-D, 0)*. En estas condiciones, la tasa sostenible *m* y la tasa máxima *p* se calculan como:

$$m = \inf_{t \to +\infty} \frac{\alpha(t)}{t} \tag{3.23}$$

$$p = \sup_{t \geq 0} \frac{\alpha(t)}{t} \tag{3.24}$$

de forma que $m \leq e_D \leq p$, y si además $\alpha(t)$ es cóncava:

$$\lim_{D \to +\infty} e_D(\alpha) = m \tag{3.25}$$

Fijada la capacidad *C*, el enlace garantiza un b*acklog* máximo de *B* bits si:

$$C \geq \sup_{t \geq 0} \left\{ \frac{\alpha(t) - B}{t} \right\} = f_B(\alpha) \tag{3.26}$$

donde $f_B(\alpha)$ es la capacidad equivalente del enlace.

Para el caso concreto de curvas de llegada del tipo *T-Spec*, el ancho de banda efectivo toma, según [Boudec, 2000b], la expresión:

$$e_D = \max \left\{ \frac{M}{D}, r, p \left(1 - \frac{D - \frac{M}{p}}{\frac{LM}{p - r}} \right) \right\} \tag{3.27}$$

Además, el ancho de banda efectivo, correspondiente a un flujo agregado, siempre y cuando los retardos máximos se mantengan, es menor o igual a la suma de los anchos de banda efectivos de los flujos individuales. La diferencia

recibe el nombre de ganancia de multiplexado (G_m) e indica la cantidad de capacidad reutilizable cuando un enlace es utilizado por varios flujos.

$$G_m = \sum_i e_D(\alpha_i) - e_D\left(\sum_i \alpha_i\right) \tag{3.28}$$

En el caso más general, un enlace con tasa máxima *P*, tasa máxima sostenible *S* y tolerancia de ráfaga (*backlog*) *B*, su curva de servicio δ debe cumplir:

$$\delta(t+D) \geq \alpha(t) \tag{3.29}$$

Para todo $t \geq 0$, siendo:

$$\delta(t) = \min\{Pt, St + B\} \tag{3.30}$$

En estas condiciones, la curva de servicio ideal cumple que:

$$\forall t \geq 0 \begin{cases} (t+D)P \geq \alpha(t) \\ (t+D)s + B \geq \alpha(t) \end{cases} \tag{3.31}$$

En la Fig. 3.30 se muestra la curva de servicio ideal para una curva de llegada $\alpha(t)$:

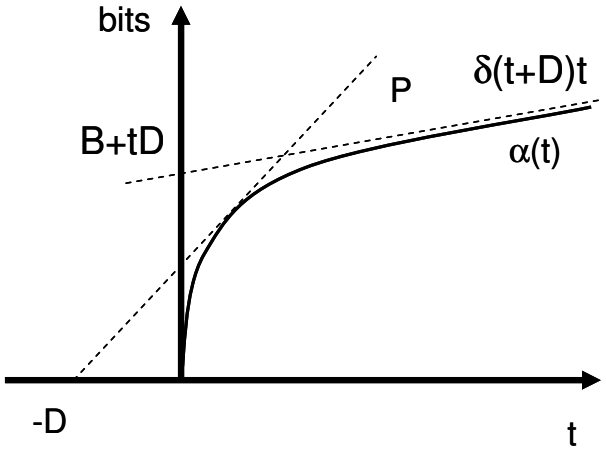


Fig. 3.30: Límites de la curva de servicio mínima para un retardo máximo *D*

De donde se deduce que:

$$P \geq e_D(\alpha) = P_0 \tag{3.32}$$

De esta forma, por ejemplo, en el caso de flujos de tráfico de tasa constante (CBR) el enlace presenta un ancho de banda efectivo igual a P_0.

Para obtener los valores óptimos de *S* y *B,* [Boudec, 1998] propone un método para determinar el área de optimización fijando *P* a P_0., de forma que sea posible determinar los parámetros fundamentales de la curva de servicio óptima, y entonces, el parametrizado de un flujo de tráfico IP pueda ser realizado a partir de sus medidas reales. De esta forma, conocida una determinada curva de llegada, es posible determinar el área de optimización

dentro del período de observación de la misma y calcular los valores óptimos de B_0, P_0 y S_0. Estos valores a su vez determinan los límites superiores de la curva de llegada en dicho período y, si por ejemplo, las muestras se corresponden con medidas en la hora cargada de un adaptador de red, los límites del tráfico de usuario o de fuente, y así estimar las curvas de servicio asociadas al elemento de interconexión.

El método planteado en [Boudec, 1996 / Boudec, 1998] , ha sido utilizado y adaptado para el cálculo de dichos valores con el fin de determinar figuras de tráfico límite basadas en funciones *TSpec*, a partir de las curvas $P_0(t+D)$ y $S_0(t+D)+B_0$, siendo D el retardo máximo asumible por el servicio/red/usuario. El principal problema que plantea este algoritmo es que su definición se realizó con vistas a algorítmica de programación lineal. Como alternativa, a continuación se propone una modificación del mismo, fácilmente programable, con la única limitación que la precisión de las estimaciones empeora en los extremos de la serie. El desarrollo completo, a partir de la propuesta original, ha sido incluido en el Apéndice B, y aquí se muestra la adaptación realizada y programada para su aplicación en los ejemplos que se muestran más adelante.

Sea $\alpha(t)$ una curva de llegada medida en un determinado elemento de red cuyos parámetros fundamentales son S_{max} y B_{max}, que representan tasa máxima sostenible (relacionada con la velocidad física de los enlaces) y la tolerancia de ráfaga. Como parámetros restrictivos, D y S_{max} son parte fundamental en los costes asociados al elemento de red. Es por ello que se define una medida del coste atribuido a S_{max} y B_{max}, representada por u, de forma que si $u<D$ entonces $S_0=S_{max}$:

Antes de aplicar directamente este método, se optó por realizar algunas pruebas de funcionamiento sobre sistemas teóricos conocidos. A partir de librerías C++ específicas para la generación de números pseudoaleatorios (Newran[29], fSeries[30] y ToolPak[31]), se desarrolló un generador de trazas con funciones de distribución conocidas, para los tiempos entre llegadas y el tamaño de las peticiones. Así, el ejemplo que se presenta a continuación, se basa en una fuente markoviana tradicional, con funciones de Poisson y exponencial negativa respectivamente.

[29] http://www.robertnz.net/nr02doc.htm
[30] http://phase.hpcc.jp/mirrors/stat/R/CRAN/doc/packages/fSeries.pdf
[31] http://www.business.ualberta.ca/aingolfsson/QTP/download.htm

$$
\begin{array}{l}
P_0 = \max_{s\geq 0}\left\{\dfrac{\alpha(s)}{s+D}\right\} \\
S_{\max} = \min\{P_0, S_{\max}\} \\
\texttt{si } B_{\max} \geq \max_{s\geq 0}\{\alpha(s)-(s+D)S_{\max}\} \texttt{ entonces} \\
\quad \texttt{si } u < D \texttt{ entonces} \\
\quad\quad S_0 = S_{\max} \\
\quad \texttt{si no} \\
\quad\quad S_e = \max_{s\geq 0}\left\{\dfrac{\alpha(s)-B_{\max}}{s+D}\right\} \\
\quad\quad \overline{S} = x \texttt{ tal que minimice } (u-D)x-\breve{\alpha}(x) \\
\quad\quad \texttt{si } \overline{S} > S_{\max} \texttt{ entonces} \\
\quad\quad\quad S_0 = S_{\max} \\
\quad\quad \texttt{si no} \\
\quad\quad\quad S_0 = \max\{S_e, \overline{S}\} \\
\quad B_0 = \max_{s\geq 0}\{\alpha(s)-(s+D)S_0\} \\
\texttt{si no "No hay solución"}
\end{array}
$$

Para este ejemplo se modela una fuente con un tiempo medio entre peticiones de 3 segundos, y un tamaño medio de las mismas de 41 KBytes (son valores típicos correspondientes a una sesión *Web*, tal como se estudia en el capítulo siguiente). En la Fig. 3.31 se muestra la curva de llegada de la traza resultante, y a su derecha la curva *a(t)*, en la que se representa el número de bytes total acumulados en función del tiempo. Por su parte, en la Fig. 3.32 se presenta el gráfico R/S, para la estimación de *H*, y el correlograma correspondiente.

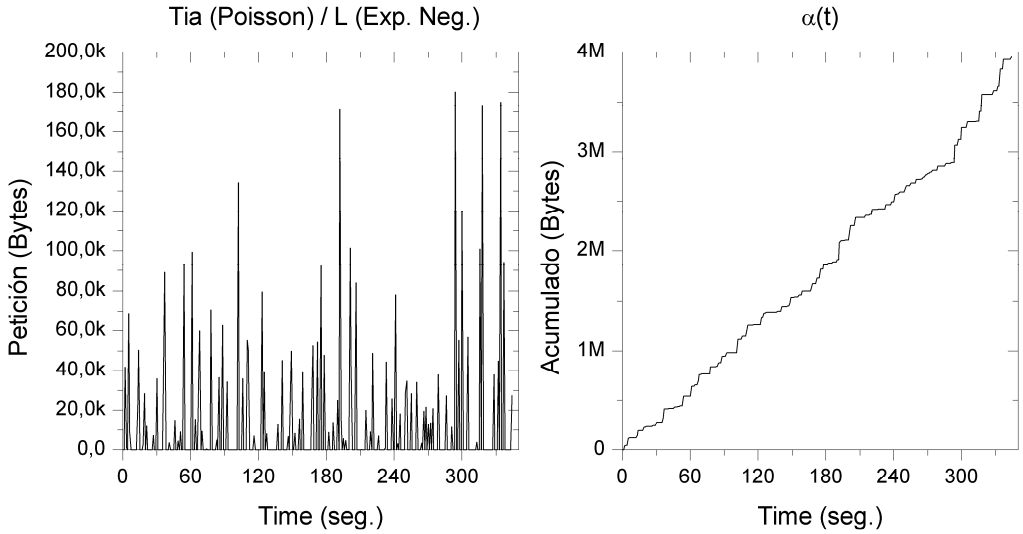


Fig. 3.31: Secuencia generada por una fuente con tiempo entre llegadas Poisson y tamaño de peticiones exponencial negativa y su correspondiente curva de llegadas α(t)

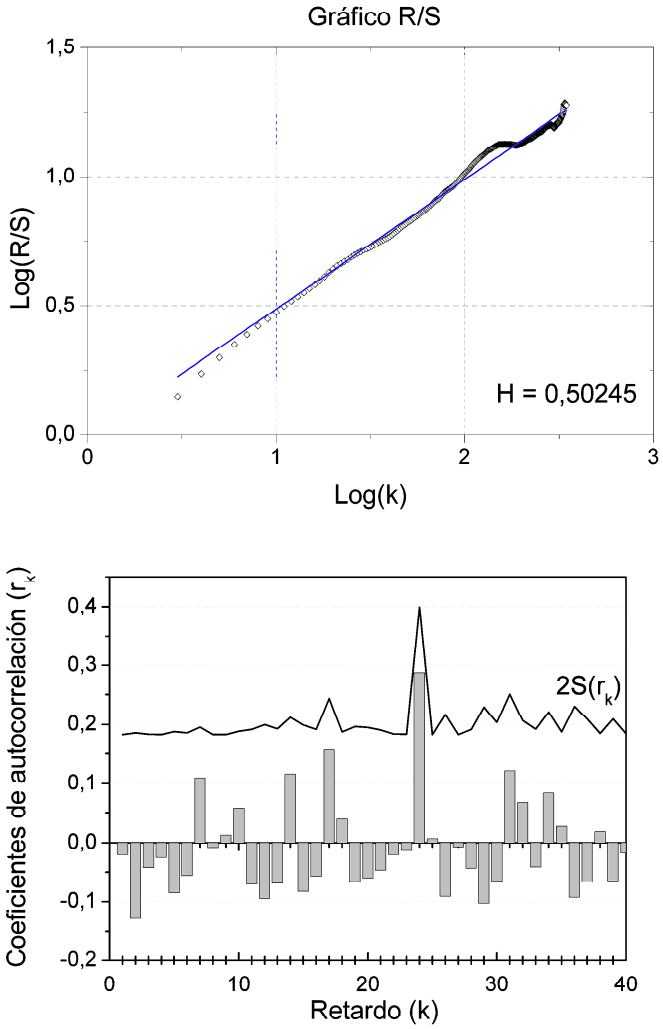


Fig. 3.32: Fuente Poisson/ExpNeg - Estimación del parámetro de Hurst y coeficientes de autocorrelación.

Se observa cómo una fuente markoviana presenta características autosimilares muy limitadas. El parámetro de Hurst toma valores alrededor del 0.5, aunque, como era de esperar, no existe ninguna correlación entre muestras ni estacionalidad.

Supuesto que la traza obtenida se corresponde a un usuario del servicio *Web*, todavía falta caracterizar el acceso físico para poder llevar a cabo el cálculo de (P_0, S_0, B_0). Para este ejemplo se considera un usuario que accede a Internet mediante un modem telefónico a 33 Kbps, con retardos máximos de 1 segundo y un buffer de línea de 256 KBytes. La curva límite resultante es igual a $A(t) = min\{ P_0(t+D) , S_0(t+D)+B_0 \}$, con ($P_0$ = 18633,45 Bps, S_0 = 7705,44 Bps, B_0 = 262144 Bytes), se muestra en la figuraFig. 3.33:

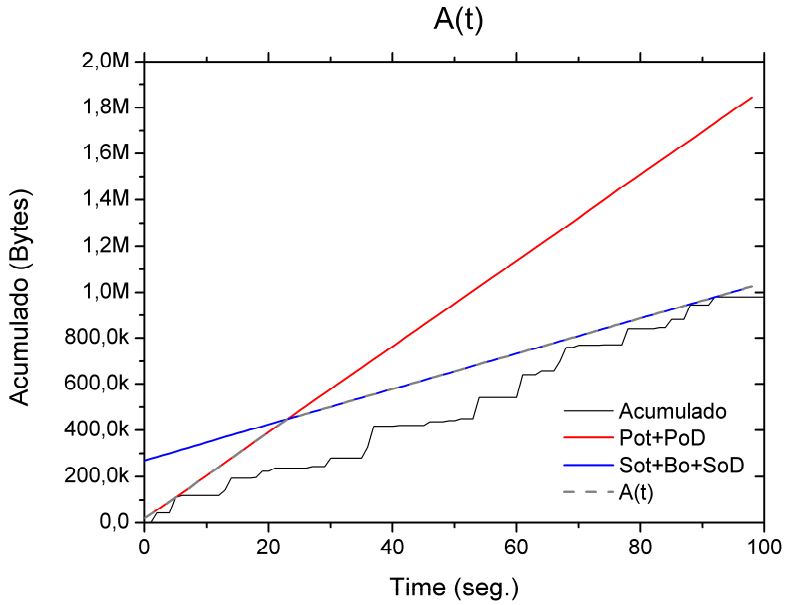


Fig. 3.33: Fuente Poisson/ExpNeg. – Curva límite A(t) = min{ $P_0(t+D)$, $S_0(t+D)+B_0$ }

En estas condiciones, sería imposible que el usuario accediese al servicio, ya que para asegurarse de que no existieran pérdidas, necesitaría un ancho de banda mínimo de 149 Kbps, al menos los primeros 24 segundos. Con un acceso a 33 Kbps sería necesario aumentar el buffer para dar cabida a al menos 586 KBytes pero asumiendo retardos superiores a 2 minutos y medio (146 segundos).

Sin embargo, gracias a las características del algoritmo de optimización utilizado, el caso en el que $u>=D$ (por ejemplo, en el supuesto que modificar B_{max} y S_{max} es más costoso que asumir mayor retardo), los valores de (P_0, S_0, B_0) se mantienen constantes, por lo que desde el punto de vista del usuario permite determinar las características específicas del acceso físico para soportar el servicio, y desde el punto de vista del operador, el ancho de banda mínimo que le va a requerir cada usuario que utilice dicho servicio.

En el caso contrario, para $u<D$ (por ejemplo, cuando las propias características del servicio imponen restricciones específicas en el retardo D), el aumento del ancho de banda relaja las condiciones del *backlog*, y velocidades inferiores a la teórica (P_0) pueden ser todavía válidas. Así, por ejemplo, en las condiciones de la fuente anterior, la Fig. 3.34 muestra los valores de B_0 en función del retardo máximo para diferentes velocidades de acceso.

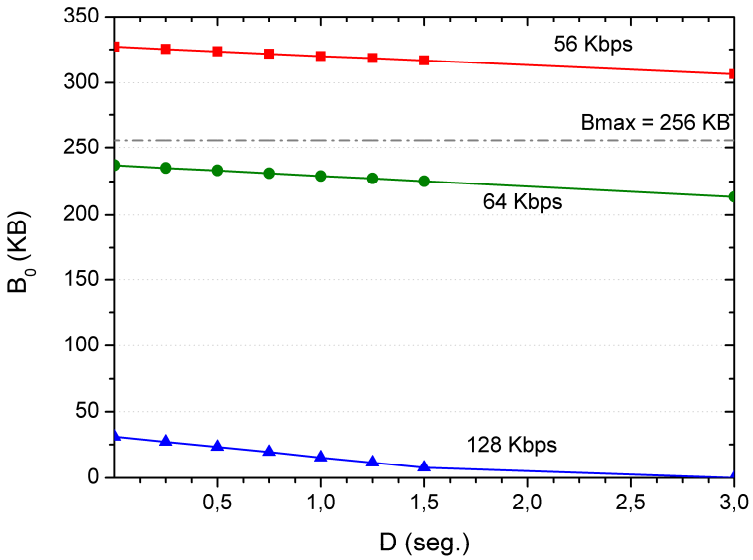


Fig. 3.34: Fuente Poisson/ExpNeg. – Evolución de la tolerancia de ráfaga (B_0) en función del retardo y de la velocidad de acceso

Se observa cómo, un acceso a 64 Kbps (p.e. un canal B de RDSI) puede ser suficiente incluso en valores de retardo mínimos.

A continuación se muestra un ejemplo de aplicación sobre una captura real, en concreto, correspondiente a un determinado cliente del Website del Mundial del 98. En la Fig. 3.35 se muestra la secuencia de llegadas original, y en la Fig. 3.36 la curva de llegadas límite, calculada a partir de este método, con (P_0 = 5597,76 Bps, S_0 = 752 Bps, B_0 = 262144 Bytes). Para la optimización se ha considerado un retardo máximo de 1 segundo y un buffer de 256 Kbytes.

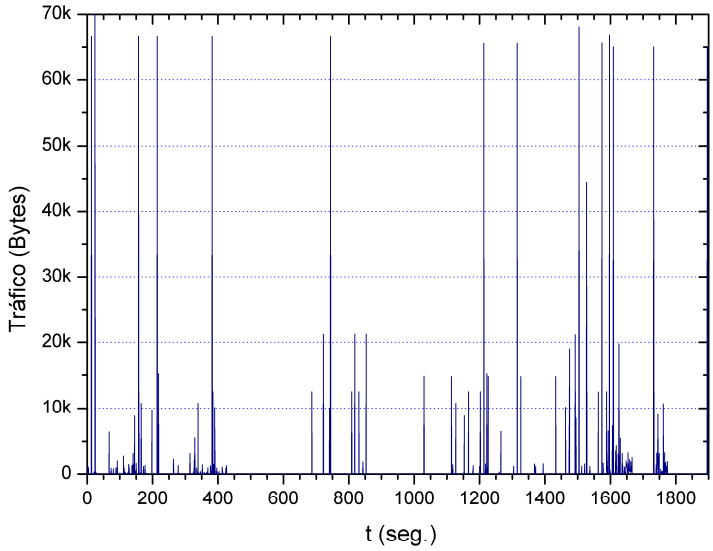


Fig. 3.35: Secuencia de llegadas de un cliente Web, extraída de la captura "Mundial98" del apartado 3.1.1.3

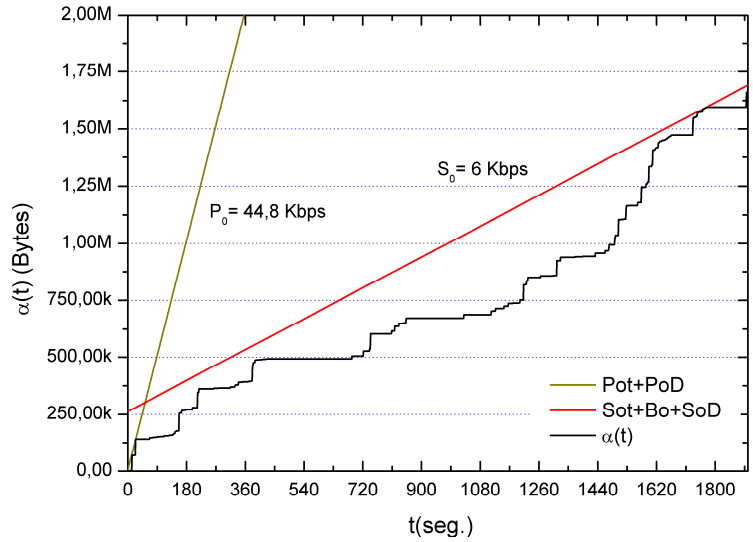


Fig. 3.36: Curva de llegadas real y composición de la posible curva de llegadas optima

Sin embargo, la utilidad de este método se limita a la parametrización determinista de cualquier curva de llegada con el fin de estimar curvas de servicio ajustadas a restricciones de retardo máximo. Las curvas de servicio óptimas en estos casos coinciden con tasas permanentes de valor P_0.

Sin embargo, aún cuando se da la condición $u<D$, si el flujo ha sido generado por un servicio *best-effort*, su curva puede ser caracterizada con la única limitación del *backoff*, con la consiguiente reducción en el valor de la capacidad equivalente, y por tanto, de los costes relacionados. Esta es precisamente la adaptación que se ha introducido al algoritmo de optimización anterior, como alternativa a la aproximación propuesta bajo la condición $u<D$.

Se calcula la tasa máxima distribuida $P_0^*(t)$ a lo largo de la secuencia en bloques de h muestras y se obtiene su media $\overline{P}_0^*$.

$$\overline{P}_0^* = E\left[\max_{t\le s\le t+h}\left(\frac{\alpha(s)}{s+D}\right)\right] \qquad con\ t \ge 0 \tag{3.33}$$

La curva de servicio mínima tiene la expresión:

$$P_{BE\min}(t) = \overline{P}_0^*(t+D) - \min_{s>0}\left(\overline{P}_0^*(s+D) - \alpha(s)\right) = \breve{P}_{BE}t + B \qquad B \le B_{\max} \tag{3.34}$$

Donde $\breve{P}_{BE}$ es la pendiente mínima, aunque en condiciones de retardo máximo al no aprovechar todo B_{max},,. Por ello, hay que calcular aquella pendiente que obtenga un compromiso entre retardo y *backlog*, es decir, se busca un valor de la pendiente $\breve{P}_{BE}^*$ que obtenga una curva de servicio que reduzca el retardo medio y siempre dentro de los límites del buffer. A partir de la siguiente expresión:

$$C_s(n) = \min\left\{\alpha(n), \breve{P}_{BE}^* + C_s(n-1)\right\} \qquad n \ge 0\ y\ \breve{P}_{BE}^* \ge \breve{P}_{BE} \tag{3.35}$$

Se busca optimizar el valor de $\breve{P}_{BE}^*$ bajo la condición:

$$\max\left(\alpha(t) - C_s(t)\right) \le B_{\max} \tag{3.36}$$

La curva de servicio será $\breve{P}_{BE}^* t + B_{BE}$, con $B_{BE} \le B_{max,}$ que se muestra en la Fig. 3.37, donde se compara con la curva de llegada real y la curva de llegada óptima. Así, $\breve{P}_{BE}$ y $\breve{P}_{BE}^*$ definen el rango de valores entre los que puede encontrarse la capacidad equivalente de un servicio *best-effort* en condiciones de retardo y *backlog* flexibles.

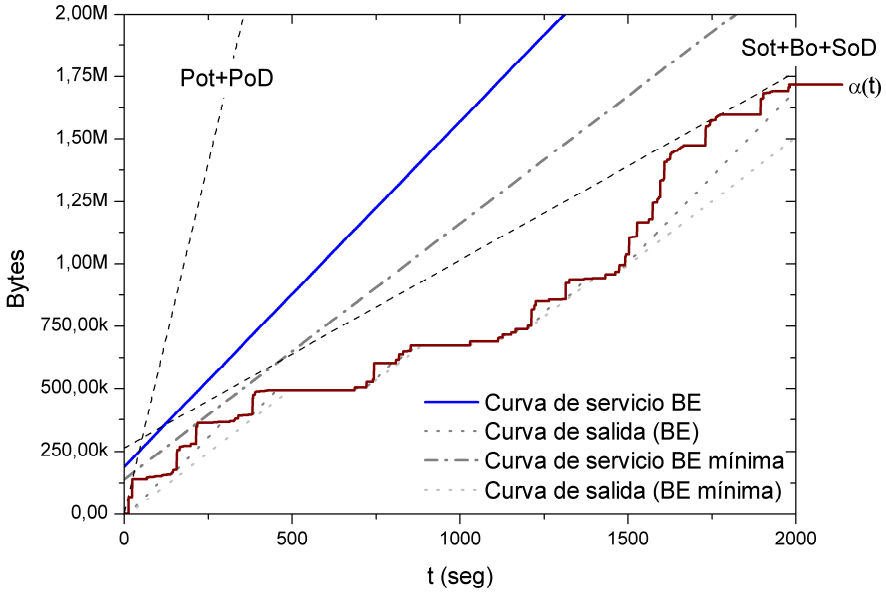


Fig. 3.37: Estimación de Curvas de servicio para servicios con restricciones flexibles en el retardo

Hasta aquí se ha comprobado la utilidad del cálculo de curvas de llegada mediante un algoritmo de optimización, y cómo es posible estimar una curva de tráfico ajustada a las condiciones límite que caracterice a un determinado flujo, de forma estricta, o con cierta flexibilidad. Ha sido uno de los objetivos de esta Tesis, estudiar su uso en la caracterización del tráfico de fuente de servicios en función de las condiciones externas del mismo, esto es, el tipo de acceso y el comportamiento de diferentes tipos de usuarios, así como su manejo para el cálculo del tráfico agregado, tal como se expone en los apartados siguientes.

3.4. Modelado del tráfico agregado

Una vez conocida la capacidad efectiva utilizada por una fuente, es posible determinar la figura de tráfico normalizado correspondiente expresada en *Erlangs*. Sin embargo, el cálculo presentado anteriormente, aunque altamente didáctico, tiene su desventaja a la hora de considerar el efecto de la mezcla de tráfico, tal como se explicó en el apartado 2.3.2. El modelo anterior, solamente describe el comportamiento de una sola fuente con una sola aplicación, por lo que todavía queda sin resolver el problema del tráfico multi-fuente que es el que genera una verdadera mezcla de tráfico.

De acuerdo con el modelo ON-OFF multinivel presentado, la agregación de fuentes de este tipo puede ser llevada acabo mediante tres aproximaciones

diferentes: la multiplexación estadística, los modelos basados en procesos modulados de Markov y las aproximaciones binomiales.

La aproximación que realiza la Multiplexación Estadística permite calcular la capacidad equivalente asociada a *N* fuentes multiplexadas como:

$$C = \min\left\{\sum_{i=1}^{N}\rho_i + \gamma\sqrt{\sum_{i=1}^{N}\sigma_i^2}, \sum_{i=1}^{N}R_i\right\} = \min\left\{\sum_{i=1}^{N}\rho_i + \gamma\sqrt{\sum_{i=1}^{N}\rho_i\left(R_i-\rho_i\right)}, \sum_{i=1}^{N}R_i\right\} \quad (3.37)$$

Con ϱ_i y σ_i^2 la tasa de bit media y la varianza de la fuente i-ésima, R_i la tasa de bit máxima. El factor γ en este caso es ajustado en función error ε, utilizado. En [Wang, 1998] se propone la expresión:

$$\gamma = \sqrt{-2\ln(\varepsilon) - \ln(2\pi)} \quad (3.38)$$

Por su parte, las fuentes moduladas de Markov consideran la superposición de las *N* fuentes, cuyo límite superior es

$$C = \sum_{i=1}^{N} C_i \quad (3.39)$$

siendo C_i la capacidad equivalente de la fuente i.

Tal como se explica en el apartado 2.3.2.2, los modelos basados en fuentes moduladas de Markov, como por ejemplo D-MMDP, modelan la agregación de fuentes ON-OFF utilizando una cadena discreta de Markov como proceso modulador. De acuerdo con esta idea, un sistema D-MMDP define un proceso de llegadas cuya tasa de bit está controlada por la probabilidad de que dos fuentes se encuentren activas, modelada ésta por una función de distribución binomial. Sin embargo, asumiendo la aproximación gaussiana (fuentes individuales normales e independientes), el ancho de banda requerido se calcula conocidos los dos primeros momentos de tráfico agregado. El valor obtenido es sólo una aproximación del límite superior, basada en valores máximos y que representa una estimación de ancho de banda libre de errores. Por el contrario, el modelo D-MMDP, utiliza una aproximación basada en el ancho de banda efectivo, aunque sin embargo, el resultado es muy similar.

Por último, la mezcla de tráfico suele aproximarse mediante funciones de distribución binomiales y que, por extensión, permiten establecer el correspondiente modelo de agregación, como se expone en [Parkinson, 2002] y, dentro de las publicaciones realizadas durante esta Tesis, en [Garcia, 2002]. Es precisamente este método, el que parece adecuarse mejor a la configuración de fuentes ON-OFF multinivel.

3.4.1. Aplicación de la aproximación binomial en el modelado de fuentes ON-OFF agregadas

En las condiciones planteadas por el modelo ON-OFF multinivel, resultaría interesante poder llevar a cabo la estimación del tráfico equivalente correspondiente a la agregación de varias fuentes de este tipo. Para ello se va a partir del caso general a partir de fuentes ON-OFF simples. De esta forma, N fuentes independientes son agregadas tal como se muestra en la Fig. 3.38, en las que cada fuente genera v_p bits/seg en su estado activo:

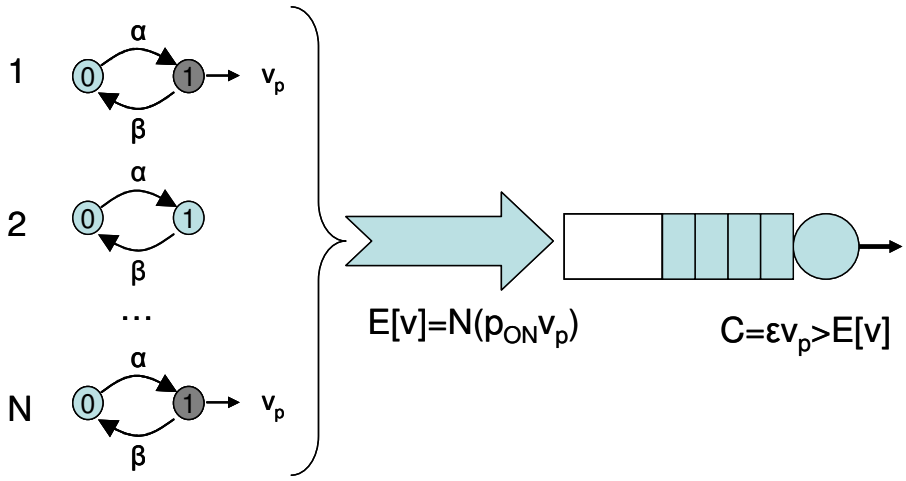


Fig. 3.38: Modelo de Agregación de N fuentes ON-OFF

En el caso más sencillo, que todas las fuentes sean CBR($p_{OFF}=0; p_{ON}=1$), la tasa de datos agregada será de Nv_p. En el caso de que existan facilidades adicionales propias de determinados tipos de servicios, como por ejemplo introducir la compresión de silencios en el caso de VoIP, la ganancia que introduce la multiplexación estadística permite calcular la capacidad del servidor C como $C = \varepsilon v_p$, siendo ε el número de fuentes CBR equivalentes. Si k fuentes independientes están activas, con una probabilidad p_{ON}, $(N-k)$ fuentes están inactivas con probabilidad $p_{OFF} = (1-p_{ON})$. Por otro lado hay N sobre k posibilidades de seleccionar k elementos fuera de los N. De esa manera la probabilidad de que haya k fuentes activas sigue una función de distribución binomial, esto es, en media $E(k) = N \cdot p_{ON}$ fuentes están activas de forma que la tasa de datos media generada por N fuentes resulta $E(v) = v_p\ N\ p_{ON}$. En consecuencia ε debe cumplir que $N \geq \varepsilon > E(k) = N\,p_{ON}$.

La condición de sobrecarga se produce cuando la capacidad del servidor es menor que la máxima tasa de datos de la fuente. Bajo esta condición el *buffer* del servidor se encontrará lleno y se producirá la pérdida de paquetes, de acuerdo con la expresión:

$$P(k) = \binom{N}{k} \cdot {p_{ON}}^{k} \cdot (1 - p_{ON})^{N-k} = \binom{N}{k} \cdot \left(\frac{\alpha}{\alpha+\beta}\right)^{k} \cdot \left(\frac{\beta}{\alpha+\beta}\right)^{N-k} \quad (3.40)$$

Esta misma situación puede también ser derivada a partir de una cadena de Markov obtenida a partir del modelo binomial M(N)/M/N y mostrada en la Fig. 3.39. Si el servidor provee capacidad para N fuentes y $\varepsilon < s_o$, todos los estados desde s_o hasta N producirán sobrecarga, la probabilidad de sobrecarga podrá ser calculada como la probabilidad de que haya más fuentes activas que capacidad disponible:

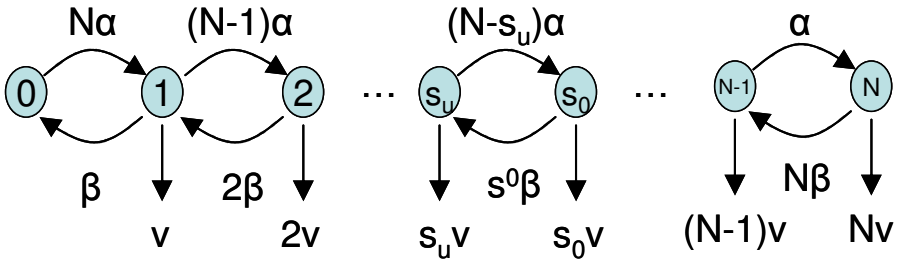


Fig. 3.39: Cadena de nacimiento y muerte para el modelo M(N)/M/N

$$P_{ol} = P(k > C) = \sum_{i=C+1}^{N} \binom{N}{i} \cdot {p_{ON}}^{i} \cdot (1 - p_{ON})^{N-i} \quad (3.41)$$

Donde C expresa la capacidad del servidor (de la dorsal) en número de fuentes, k el número de fuentes activas y N el número total de fuentes. Cabe destacar que, en el caso de sistemas de pérdida pura sin *buffer*, la probabilidad de sobrecarga coincide con la probabilidad de pérdida del sistema. Es por ello que a partir de ahora se asume el caso de sistemas de pérdida pura, caso de los niveles de conexión con mecanismos CAC.

Para tener en cuenta la probabilidad de pérdida como una función de p_{on} y del número de fuentes N, la capacidad del servidor debe ser modificada mediante un factor corrector γ que en este caso depende tanto de P_B como de p_{ON} y de N. Este valor γ es similar al ya visto anteriormente, actuando como un múltiplo de la desviación estándar del flujo de datos agregado. Aunque la selección de su valor suele ser confirmada mediante simulación, se suelen utilizar los valores típicos asumiendo. La correspondiente formulación permite calcular la capacidad total para N fuentes binomiales agregadas en *bps,*

$$C = E(v) + \gamma(P_B, N, p_{ON}) \cdot \sigma(v) = (N \cdot p_{ON} + \gamma(P_B, N, p_{ON}) \cdot \sqrt{N \cdot p_{ON} \cdot (1 - p_{ON})}) \cdot v_p \quad (3.42)$$

el número de circuitos equivalentes para N fuentes en unidades de v_p ,

$$N_{eq} = \frac{C}{v_p} = E(N) + \gamma(P_B, N, p_{ON}) \cdot \sigma(N) = N \cdot p_{ON} + \gamma(P_B, N, p_{ON}) \cdot \sqrt{N \cdot p_{ON} \cdot (1 - p_{ON})} \quad (3.43)$$

y la capacidad equivalente de una fuente binomial en unidades de v_p:

$$v_{eq} = \frac{C}{C_{CBR}} = \frac{C}{v_p \cdot N} = \frac{N_{eq}}{N} = p_{ON} v_p + \gamma(P_B, N, p_{ON}) \cdot \frac{\sqrt{p_{ON} \cdot (1 - p_{ON})}}{\sqrt{N}} v_p \quad (3.44)$$

Todo ello asumiendo la limitación del comportamiento de la capacidad equivalente v_{eq}:

$$\lim_{N \to \infty} v_{eq} = \lim_{N \to \infty} (p_{ON} v_p + \frac{const}{\sqrt{N}} v_p) = p_{ON} v_p \quad (3.45)$$

Sin embargo, el uso del modelo de agregación binomial debe realizarse teniendo en cuenta sus limitaciones. Esta aproximación, dependiendo del signo de la distribución utilizada, desplaza la tendencia de las estimaciones realizadas hacia la sobrestimación en el caso de la distribución binomial negativa, o bien a la infravaloración en el caso de la binomial positiva. En la Fig. 3.40 se muestra una comparativa entre el modelo de agregación binomial aplicado a agregados de fuentes de Voz sobre IP y los resultados obtenidos mediante simulaciones realizadas en el simulador OPNET[32]. En ella se observa cómo la aproximación realiza una sobreestimación de la capacidad efectiva requerida por el agregado de fuentes, tanto menor cuanto más se incrementa el volumen de agregación.

[32] http://www.opnet.com/

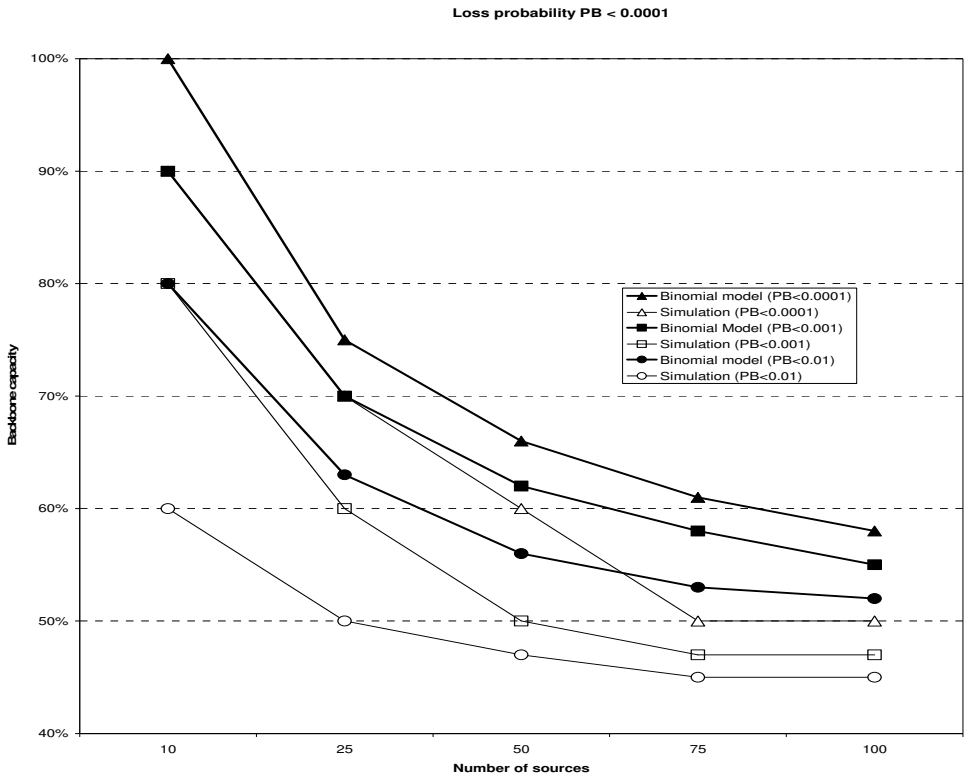


Fig. 3.40: Comparativa del sobredimensionado introducido por el modelo de agregación binomial en función del número de fuentes agregadas y de la probabilidad de bloqueo.

Todos los detalles del estudio, realizado durante esta Tesis, dentro de un Proyecto de Colaboración con la Universidad de Dresde (Alemania) fueron publicados en la revista WSEAS Trans. on Communications en 2002, y presentados en los congresos internacionales y nacionales [TELEC, 2002], [WSEASCSCC, 2002], [JITEL, 2001] y [TELECOM I+D, 2001]. Las referencias completas se han incluido en el apéndice Apéndice E.

Con todo ello, las distribuciones binomiales son utilizadas para obtener el número de ocurrencias simultáneas dentro de un grupo de procesos estadísticos independientes. Sin embargo, gracias a las características del tráfico a ráfagas, la agregación de varias fuentes de este tipo puede ser modelada utilizando una función de distribución binomial negativa. Este es el caso que se da al realizar la agregación entre niveles temporales consecutivos. Como es lógico, los niveles inferiores (el de ráfaga y el de sesión), si el servicio que modelan presenta esa estructura de ráfagas, la agregación será modelada

mediante binomiales negativas, mientras que si el servicio presenta total independencia entre el nivel de sesión y las ráfagas, las funciones seleccionadas serán positivas[33]. Esta es la razón por la cual la estructura de tres niveles propuesta, hace uso de un modelo mixto que aplica, por ejemplo, binomiales negativas en el nivel de ráfaga y positivas en el nivel de conexión. Así, para calcular el número de usuarios en el nivel de sesión, en caso de utilizar binomiales negativas y conexiones permanentes, resulta en media y varianza:

$$n_{ms} = pop\frac{P_{1s}}{1-P_{1s}} \tag{3.46}$$

$$v_s = var[n_{ms}] = pop\frac{P_{1s}}{(1-P_{1s})^2} \tag{3.47}$$

Siendo *pop* el número de usuarios potenciales y P_{1s} la probabilidad del estado de actividad en el nivel de sesión, con lo que el número de usuarios que están transmitiendo ráfagas en un instante dado se calcula como:

$$n_{mr} = (n_{ms} + \gamma v_s)\frac{P_{1r}}{1-P_{1r}} \tag{3.48}$$

Siendo n_{mr} el número medio de usuarios en el nivel de ráfaga, n_{ms} el número medio de usuarios con sesiones activas, v_s la varianza para este nivel y P_{1r} la probabilidad del estado de actividad a nivel de ráfaga. Por otro lado la varianza del número de usuarios activos en el nivel de ráfaga se calcula como:

$$v_r = var\left[n_{mr}\right] = (n_{ms} + \gamma v_s)\frac{P_2}{(1-P_2)^2} \tag{3.49}$$

con valores de γ entre 1 y 3, asegurando un rango de confianza desde 65 hasta 99,73%.

En el caso de considerar el efecto del establecimiento de conexiones, y de la misma manera que se calculan en los dos niveles inferiores, el número de conexiones activas puede ser calculado aplicando esta misma idea, salvo que, generalmente, el comportamiento del nivel de conexión recomienda hacer uso

[33] El tráfico a ráfagas tiene la característica de que, cuando llega un bloque de datos, es muy probable que le sigan más bloques de datos, lo mismo que en el caso de no recibir nada. En una binomial negativa este efecto es simulado dividiendo la probabilidad de activación entre la probabilidad de no activación, de tal manera que el resultado crezca muy rápidamente al aumentar la probabilidad de activación (se simula así la llegada de un chorro de información siempre que nos llegan datos). Sin embargo, el nivel de conexión, en principio, no tiene una naturaleza de ráfaga, ya que la aparición de una llamada, no solo no indica la aparición de varias consecutivas, sino que más bien la invalida, puesto que una vez conectado al ISP no se realiza otra conexión en un tiempo elevado.

de binomiales positivas en vez de negativas. En este caso, el número de conexiones activas se calcula como:

$$n_{mc} = popP_{1c} \tag{3.50}$$

$$v_c = pop\frac{P_{1c}}{1-P_{1c}} \tag{3.51}$$

Siendo P_{1c} la probabilidad de que un usuario establezca una conexión. En este caso, las expresiones (3.46) y (3.47) quedarían como:

$$n_{ms} = (n_{mc} + \gamma v_c)\frac{P_{1s}}{1-P_{1s}} \tag{3.52}$$

$$v_s = var[n_{ms}] = (n_{mc} + \gamma v_c)\frac{P_{1s}}{(1-P_{1s})^2} \tag{3.53}$$

En cualquier caso, las tasas de bit son calculadas a partir de los dos primeros momentos del número de usuarios con ráfagas activas:

$$V_{avg} = n_{mr}v_r \tag{3.54}$$

$$\mathrm{var}\left[V_{avg}\right] = var[n_{mr}]v_r^{\,2} \tag{3.55}$$

Siendo precisamente la capacidad requerida por un grupo de usuarios para utilizar un determinado servicio.

Para analizar el comportamiento del modelo de agregación propuesto, se toman como referencia las fuentes ON-OFF multinivel utilizadas en los ejemplos del 3.1.1.2. En este caso se modela un punto de agregación teórico con 10000 usuarios, conectados a través de una red de distribución genérica, cuyas características individuales coinciden con las de las del caso 1. Como se ha indicado anteriormente, las características del tráfico observado en cada nivel, permiten decidir el tipo de binomial utilizado en la agregación. El nivel de ráfaga en Internet, como su nombre indica, con un comportamiento idéntico en todos los servicios, justifica la agregación de fuentes mediante el uso de una binomial negativa. Por su parte, el nivel de sesión, depende directamente del tipo de servicio, con lo que la agregación de las fuentes podrá ser realizada por binomiales negativas o positivas según el caso. A modo de comparación se aplican ambas variantes al ejemplo, y se realiza el estudio de la capacidad equivalente que el punto de agregación provee a cada usuario, en función del ancho de banda total de que dispone. La Fig. 3.41 muestra los resultados obtenidos hasta una capacidad de 1 STM-1:

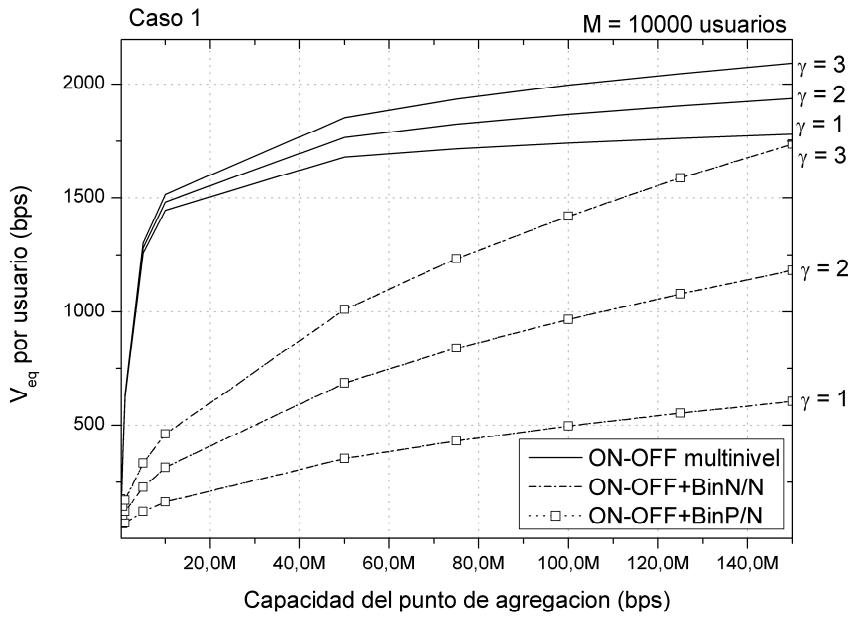


Fig. 3.41: Agregación ON-OFF multinivel mediante binomiales – Capacidad equivalente por cada fuente individual (Caso 1) respecto al ancho de banda del punto de agregación

En el gráfico se muestran tres grupos de estimaciones del ancho de banda requerido por cada usuario: mediante el modelo ON-OFF multinivel del apartado 3.3, el modelo ON-OFF multinivel con agregación mediante binomiales negativas (tanto en el nivel de ráfaga como en el de sesión – Bin N/N), y este mismo pero aplicando una binomial positiva en el nivel de sesión (Bin P/N). Al variar el factor γ se observa como los valores de las estimaciones binomiales están muy por debajo de la velocidad media del servicio, y como la elección de dicho valor resulta importante dada la varianza de las estimaciones. Además, en las condiciones del ejemplo, el efecto de la agregación binomial es el mismo independientemente de la opción en el nivel de sesión.

Al repetir las estimaciones haciendo uso de las fuentes del caso 2, se obtienen los resultados de Fig. 3.42:

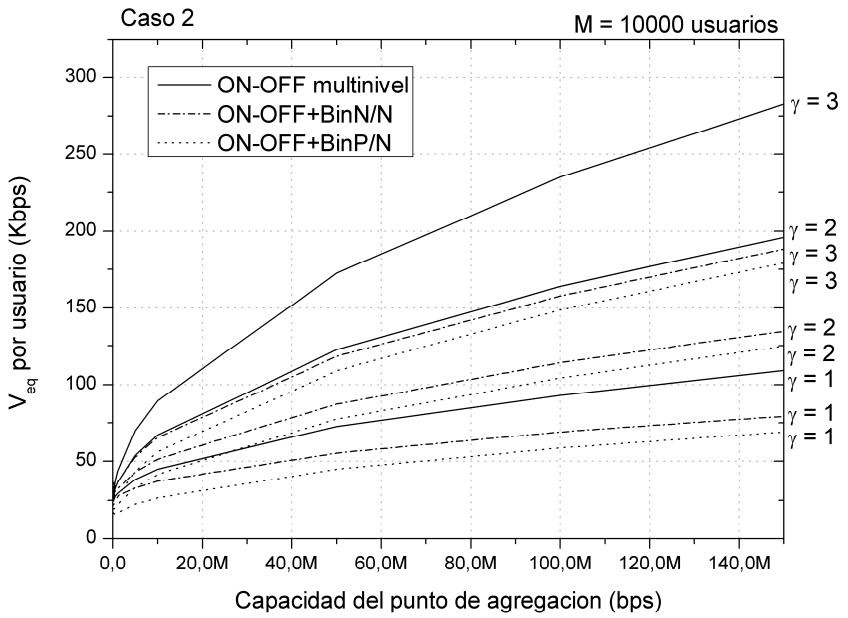


Fig. 3.42: Agregación ON-OFF multinivel mediante binomiales – Capacidad equivalente en fuentes del caso 2 en función del ancho de banda del punto de agregación

En este caso, los modelos se encuentran más a la par. Las estimaciones binomiales se mantienen en todo momento más cerca del valor medio del ancho de banda del servicio. Los resultados del modelo ON-OFF con factor $\gamma = 2$, son prácticamente equivalentes al caso más pesimista de la opción binomial. Además, tal como era de esperar, la opción N/N está siempre por encima del a Bin P/N. En la Fig. 3.43 se comparan los tráficos agregados con el de acceso, para lo que se representa el grado de ocupación del acceso físico, por usuario y en cada caso, en función γ.

Si bien en media el ancho de banda utilizado en el acceso físico puede resultar incluso inferior al requerido en el punto de agregación, al considerar el conjunto de fuentes se produce una suavización de la varianza con la consiguiente reducción de la estadística del tráfico en torno a la media. Al menos teóricamente, ambas estimaciones podrían considerarse como los límites entre los que se encontrará el ancho de banda requerido por cada usuario para dicho servicio.

Si además, tal como se explica en el apartado anterior, se tuviera en cuenta el efecto del establecimiento de conexiones, y se modelara mediante una binomial positiva, los resultados, en este ejemplo concreto, se ven incrementados en un 1.5, 0.8 y 0.23 % para valores de γ de 1, 2 y 3 respectivamente, ya que el número de conexiones activas se ve incrementado por el efecto de la varianza.

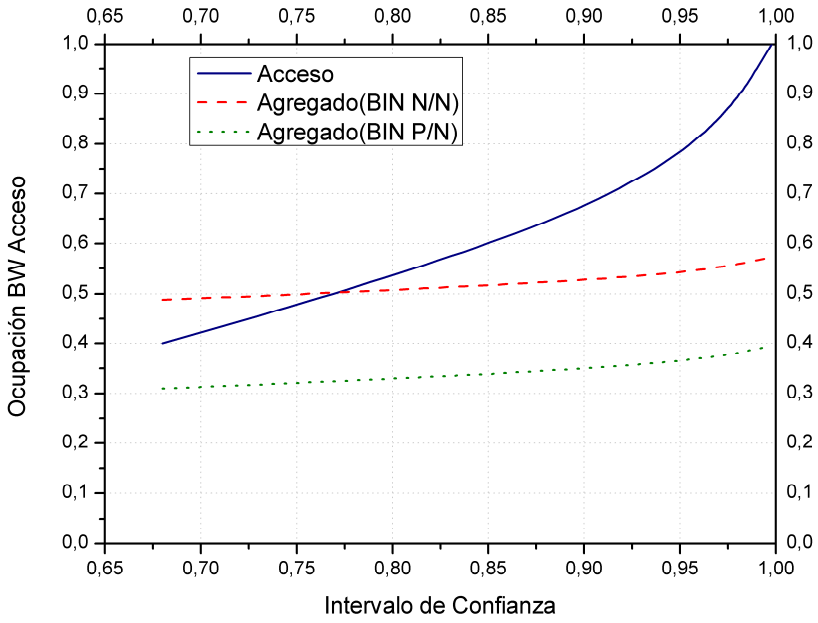


Fig. 3.43: Agregación ON-OFF multinivel – Comparación del efecto ocasionado al aplicar binomial positiva o negativa en la agregación multinivel (BIN N/N = Binomial Negativa en nivel de sesión / Binomial Negativa en el nivel de ráfaga; BIN P/N = Binomial Positiva en nivel de sesión / Binomial Negativa en el nivel de ráfaga)

Otro análisis que resulta interesante es comprobar cómo afecta el comportamiento de la fuente al ancho de banda por usuario en el punto de agregación. En este ejemplo se modifica el comportamiento al nivel de sesión. Para ello basta con modificar el volumen de descarga de la fuente, con lo que se mantienen todos los parámetros iniciales de la fuente original, salvo los tiempos de silencio y actividad por sesión. Al utilizar dos binomiales negativas (niveles de ráfaga y sesión), se obtienen los resultados de las Fig. 3.44 y Fig. 3.45, en la que se representa el ancho de banda por usuario, para diferentes velocidades de acceso, en función del volumen de descarga medio por conexión, y en el que el área en gris aparece ampliada para una mejor observación.

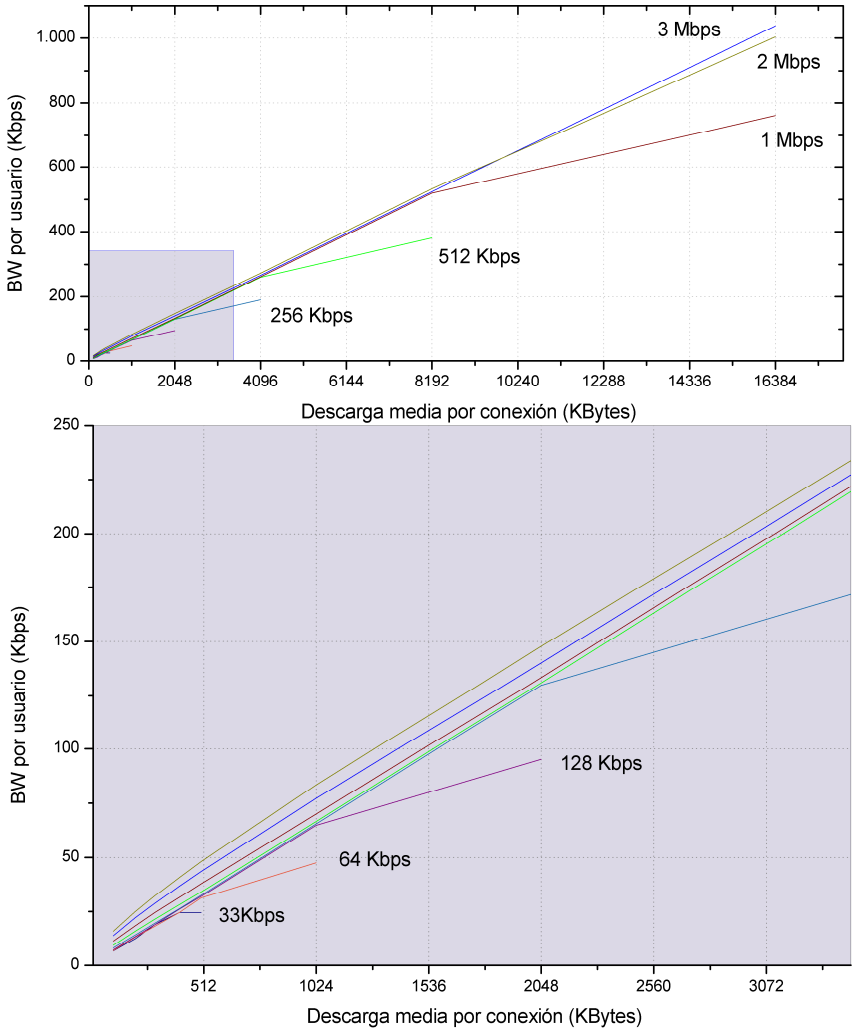


Fig. 3.44: Agregación ON-OFF multinivel – Ancho de banda por usuario en función del volumen de descarga y la velocidad de acceso en caso de utilizar agregación Bin N/N

En todos los casos, el ancho de banda por usuario se incrementa de forma lineal y prácticamente constante hasta alcanzar aproximadamente la mitad de la capacidad total del acceso. A partir de esa cifra, el incremento del ancho de banda permanece constante, pero con una pendiente menos acusada, aproximadamente entre un 30 y un 40 % inferior. Por el contrario, en el caso de utilizar la binomial positiva en el nivel de sesión, esta tendencia es mucho menos marcada.

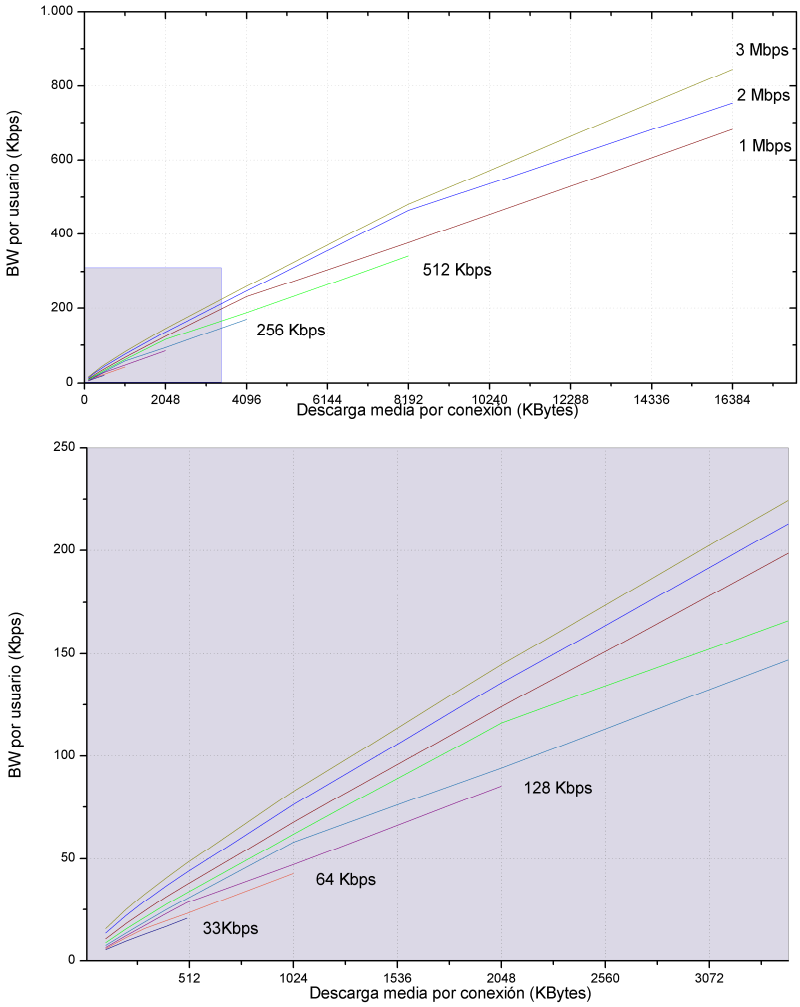


Fig. 3.45: Agregación ON-OFF multinivel – Ancho de banda por usuario en función del volumen de descarga y la velocidad de acceso en caso de utilizar agregación Bin P/N

3.4.2. Aplicación de la aditividad de curvas de servicio para el cálculo de tráfico IP agregado

Como alternativa al modelo de agregación anterior, en el apartado 2.4.3 se describe una solución al problema de agregación mediante el uso de algunos de los principios del *Network Calculus*. La posibilidad de determinar un límite determinista al comportamiento de un determinado flujo IP, se ve reforzada al tener en cuenta la posibilidad de utilizar el mismo cálculo, directamente, al

agregado de varios flujos. Sin embargo, cuando solo es posible obtener los resultados de flujos individuales, las soluciones basadas en curvas *TSpec* asumen un sobredimensionado del tráfico agregado estimado. Su límite máximo, resulta al considerar la agregación, como la suma de flujos byte a byte, de forma que, a su vez, ésta puede ser considerada como uno solo, y determinada su *TSpec Suma*. Sin embargo, como se expone en el apartado mencionado, el *Network Calculus* permite obtener la suma exacta sobre los flujos individuales.

En este apartado se muestra un ejemplo de aplicación de la agregación de curvas de llegada para la estimación de la curva de llegada, correspondiente a un conjunto de múltiples flujos agregados. Para ello, se parte de las capturas comentadas en el apartado 3.1.1.3, al corresponderse con sesiones completas del servicio *Web*, y se hace una selección de nueve sesiones, correspondientes a clientes con alto volumen de descarga. En la Fig. 3.46 se muestra la distribución del volumen de descarga del total de clientes de la traza considerada, y que ha permitido realizar la selección de los flujos más interesantes para el estudio:

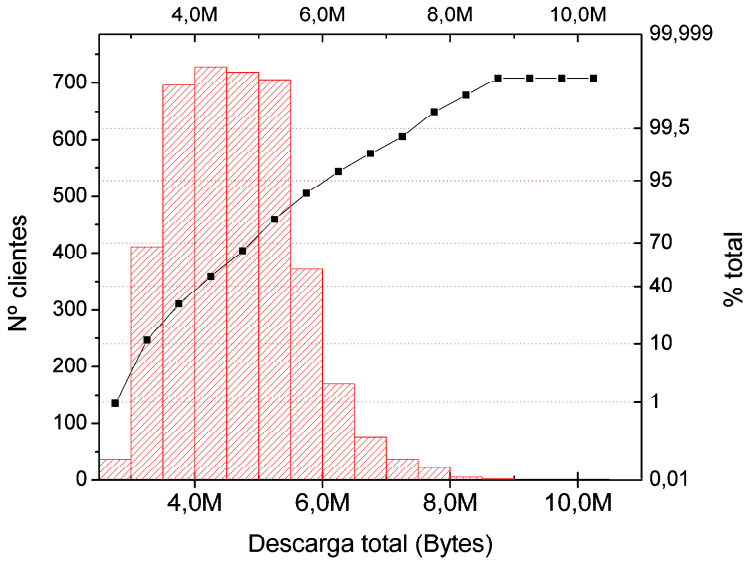


Fig. 3.46: Distribución del volumen de descarga por cliente durante una hora en la traza de tráfico del Mundial 98

Como se puede observar, casi el 80% de los clientes (identificados por su dirección IP) observados durante la captura, presentan volúmenes de descarga entre 4 y 6 MBytes. Para este ejemplo se realizado un filtrado de la captura total, en el que se han seleccionado los nueve clientes con mayor volumen de descarga, al ser éstos los que están requiriendo mayor cantidad de recursos a la red de acceso. El análisis se centra en un fragmento de unos ocho minutos de duración, durante los cuales, se incluyen todas las sesiones correspondientes a los clientes seleccionados. La captura resultante se muestra en la Fig. 3.47:

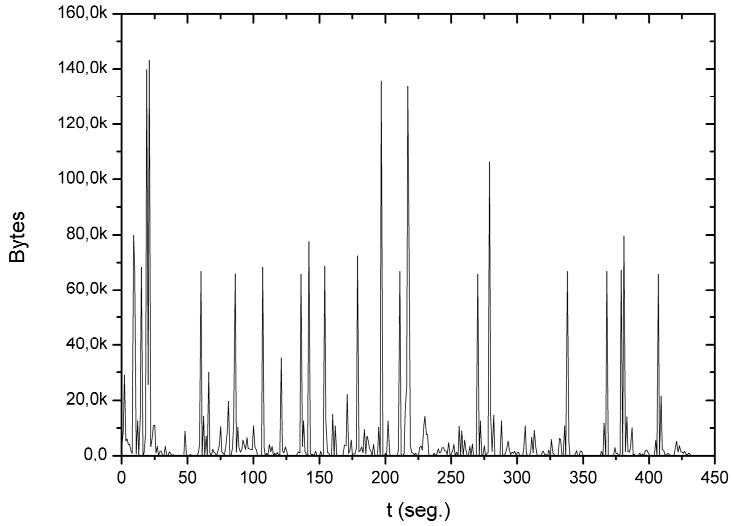


Fig. 3.47: Tráfico agregado capturado (9 clientes)

De la figura anterior se puede deducir una alta varianza del volumen de datos por ráfaga es muy elevada, y hace prever requerimientos de ancho de banda elevados al principio de las sesiones.

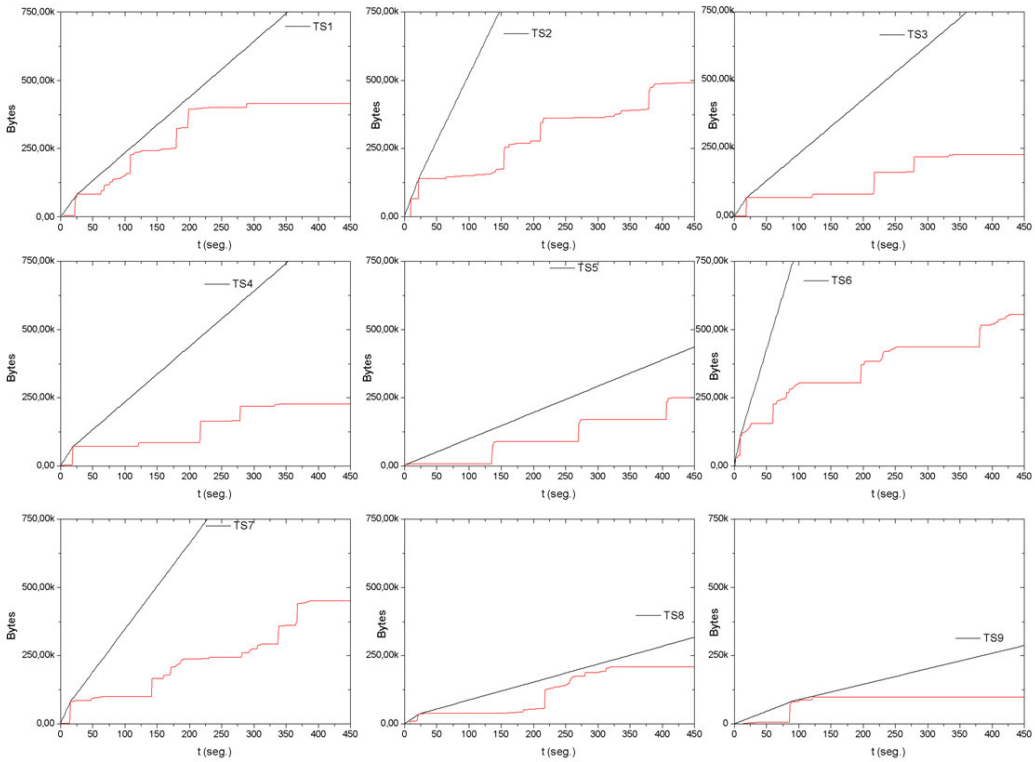


Fig. 3.48: Curvas de llegada óptimas (período de observación 8 minutos, retardo máximo 1 segundo, tolerancia de ráfaga 32 KBytes)

En la figura anterior, la Fig. 3.48, se muestra el resultado de la aplicación de los cálculos de optimización del apartado 3.3.3, sobre cada una de las curvas de llegada de los nueve clientes analizados.

Tal como se puede observa en la mayoría de los casos, el primer tramo de las *TSpec* sufre un efecto de sobredimensionado provocado por los primeros segundos de la sesión, frente al desarrollo real de la demanda (la propia curva de llegada). El alto ancho de banda inicial, afecta al ajuste posterior, y como consecuencia, la estimación obtenida dispara el sobredimensionado a partir del primer minuto. Tras aplicar el cálculo de la *Summed TSpec* tal como se describe en el apartado 2.4.3, los curvas resultantes se muestran en la Fig. 3.49:

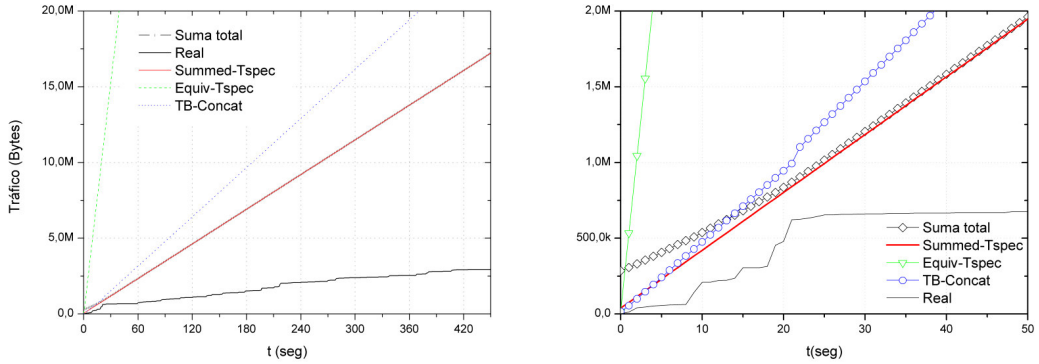


Fig. 3.49: Curvas agregadas: Suma real, *TSpec* equivalente, *Summed-TSpec* y concatenación de *Tocken-bucket*

Se observa como el sobredimensionado, respecto al tráfico real, es tanto más importante cuanto mayor es el tiempo de las sesiones, especialmente cuando estas suelen estar marcadas por grandes descargas en los primeros segundos, como es este caso. Este resultado es totalmente lógico al sumarse todas las contribuciones de cada curva individual. Sin embargo, tal como se observa en la ampliación del primer minuto (gráfica de la derecha), la estimación realizada mediante la curva *Summed-TSpec* no resulta tan mala, y sigue siendo inferior a la suma aritmética de todas las *TSpec* individuales. Por su parte, la opción *Cascaded-TSpec* (que a aparece en la gráfica como *TB-concat*) sólo se mantiene dentro de los márgenes del resto durante unos 20 segundos. En definitiva, y a modo de conclusión, en el ejemplo presentado, la validez de la estimación de la capacidad equivalente de un agregado, obtenida mediante curvas *TSpec* se limita a períodos de observación cortos, del orden de segundos. En este caso, las *TSpec* individuales pueden ser sustituidas por la *Summed-TSpec*, ya que esta mejora el resultado de la suma. Además, el uso de la *TSpec* equivalente, en función de los datos de las curvas individuales, queda fuera de toda utilidad.

En el ejemplo anterior, y en varias pruebas más realizadas modificando los clientes y el período de observación, la curva *Cascaded-TSpec* resultaba peor que la de la *Summed-TSpec*, comportamiento totalmente contrarío al esperado. Por

ello se repitió el cálculo, aunque solamente con dos clientes de los anteriores, se calcularon sus *TSpec* , y se aplicaron de nuevo las versiones *Summed* y *Cascaded*. Las curvas de llegada y sus correspondientes *TSpec* se muestran en la Fig. 3.50:

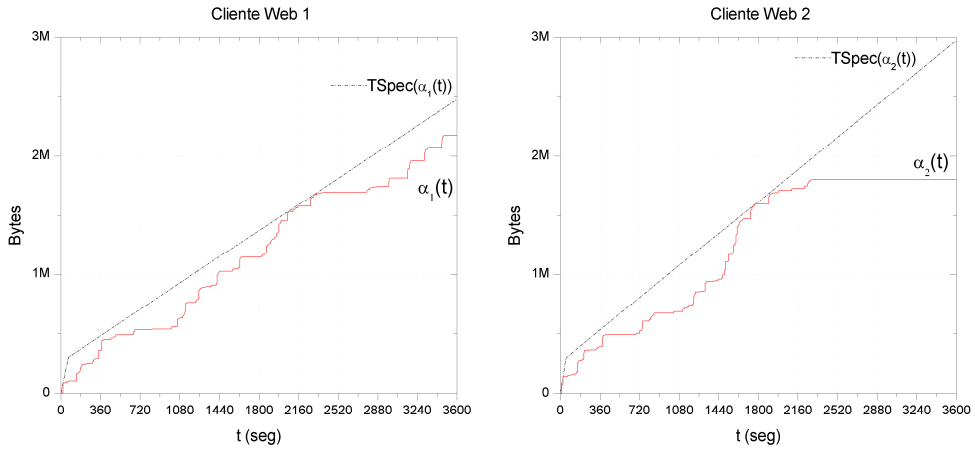


Fig. 3.50: Curvas de llegada reales y *TSpec* calculadas (período de observación 1 hora, retardo máximo 1 segundo y tolerancia de ráfaga de 256 KBytes)

Tras realizar la optimización correspondiente, las *TSpec* correspondientes a los dos clientes indicados y su suma equivalente toman los valores de la Tabla 3.IX:

	R	***b***	***p***	***m***
TSpec ***($\alpha_1(t)$)***	**6**	**256**	**45**	**5,5**
TSpec *($a_2(t)$)*	**5**	**256**	**35**	**4,3**
SummedTS	**11**	**512**	**78**	**5,5**
	Kbps	**Kbytes**	**Kbps**	**Kbytes**

Tabla 3.IX: Parámetros de las *TSpec* calculadas con las fuentes del Mundial 98

Los resultados para ambas fuentes son bastante similares, sobre todo si se comparan junto con las curvas de llegada, como se muestra en la Fig. 3.51.

Ambas curvas de llegada presentan carácterísticas muy parecidas, estando a la par casi todo el rato. Esto se puede tener su explicación en que las sesiones Web fueron realizadas sobre el mismo servidor, por lo que, los contenidos a los que acceden los clientes son, con una alta probabilidad los mismo, y de esa forma, el desarrollo de la sesión también será similar. En cuanto al ajuste de las curvas calculadas individualmente, es mucho mejor que el que alcanzan las *TSpec* del agregado, lo cual es normal, tenida cuenta que se basan en

estimaciones de los límites de las curvas (tal como se explicó en el apartado 2.4.2), por tanto, se parte de un sobredimensionado de base.

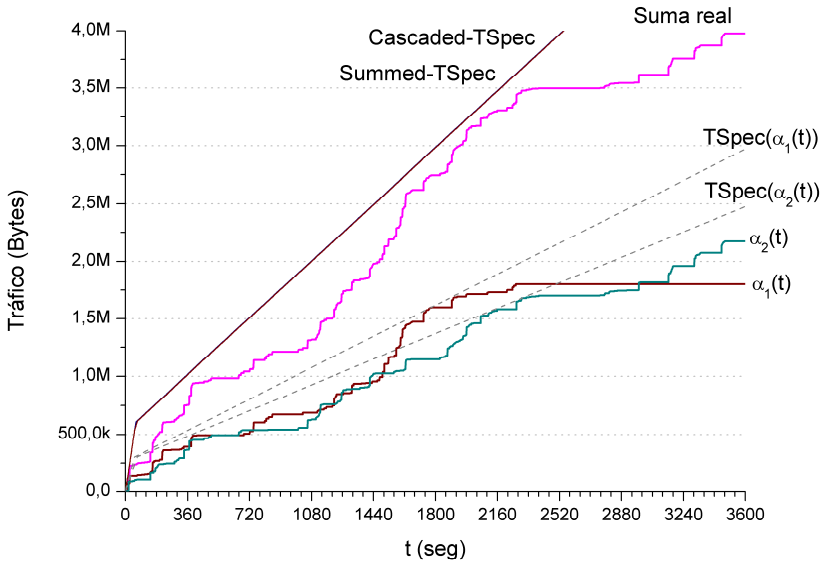


Fig. 3.51: Escenario completo del ejemplo – Curvas de llegada con sus *TSpec* asociadas, Curva de llegada agregada, *TSpec-Suma* y *Cascaded-TSpec*

Además, los resultados de las aproximaciones para los agregados, no presentan inicialmente grandes diferencias las unas de las otras. En el caso de la Summed-TSpec, como su cálculo se realiza a partir de los parámetros correspondientes a cada TSpec individual, y estas utilizan el mismo valor de 256KBytes en el buffer, va acoincidir con la suma de ambas. Por su parte, la Cascaded-TSpec parte de los mismos parámetros que la Summed, pero sus resultados teóricamente dependen del número de curvas agregadas, que en este caso es 2. En la Fig. 3.52 se muestra un detalle de dichas curvas, especialmente en la zona donde resulta más clara la diferencia entre ellas.

En este ejemplo tan concreto, en el que solamente se han agregado dos curvas, las diferencias entre las aproximaciones son mínimas. Sin embargo, es posible obtener algunas conclusiones, como por ejemplo, que la solución basada en *Summed-TSpec* obtiene mejores resultados que si se sumaran las *TSpec* individuales una a una, tal como se indica en el 2.4.3. Por su parte, la *Cascaded-TSpec* también obtiene una aproximación "inferior" a la opción *Summed*. De hecho, ambas curvas se siguen muy de cerca una de otra. En cuanto al ajuste de las aproximaciones con respecto a las curvas agregadas reales, éste mejora conforme aumenta el número de flujos agregados, al introducirse nuevos vértices en la opción *Summed*, o bien más token-bucket en la opción *Cascaded*. Sin embargo, conforme aumenta dicho ajuste, también puede incrementarse el sobredimensionado total, al sumarse las aportaciones de cada curva individual.

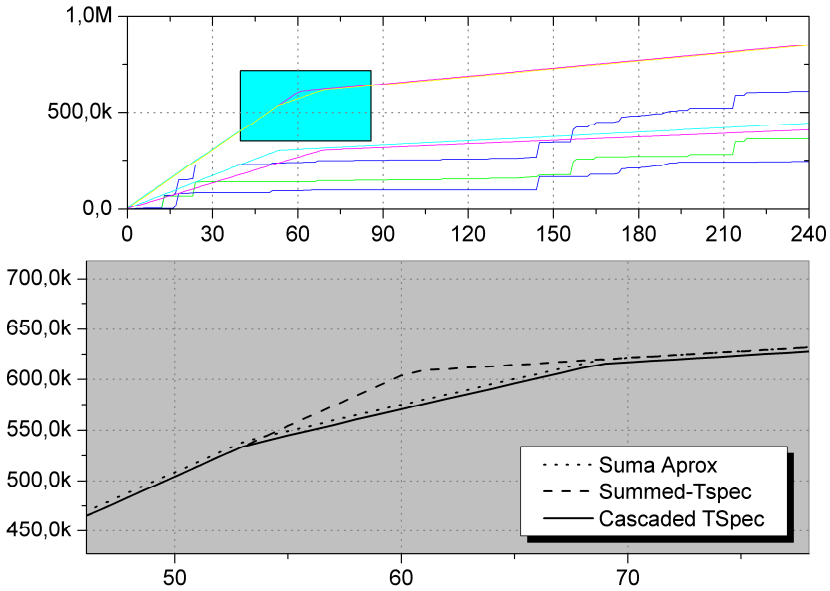


Fig. 3.52: Detalle del "codo" de las estimaciones del agregado: *Summed*, *Cascaded* y la suma de las *TSpec* individuales

Por último, solo queda realizar una pequeña comparación del método gráfico, presentado a lo largo de este apartado, y el modelo teórico, basado en fuentes ON-OFF multinivel expuesto, al principio de este capítulo.

Tal como se indicó, el ejemplo de fuente correspondiente al primer caso (servicio Web), utilizaba parámetros típicos de un servicio Web. Por su parte, las capturas estudiadas en este subapartado, también pertenecen a dicho servicio. Para establecer una comparación entre ellos, se procedió al modelado de una fuente ON-OFF multinivel equivalente, basada en el servicio Web del ejemplo, sin más que variar los parámetros de velocidad de acceso. En la **Tabla 3.X** y en la Fig. 3.53 se muestran los resultados correspondientes a dicho ejercicio.

	TSpec	ON-OFF multinivel
$a_1(t)$	P_0 = 45 Kbps	V_{eq}=54.16 Kbps ≈ Vac = 56 Kbps
$\alpha_2(t)$	P_0 = 35 Kbps	V_{eq}=32.35 Kbps ≈ Vac = 33 Kbps

Tabla 3.X: Resultados de los modelos comparados y posibles equivalencias con accesos comerciales

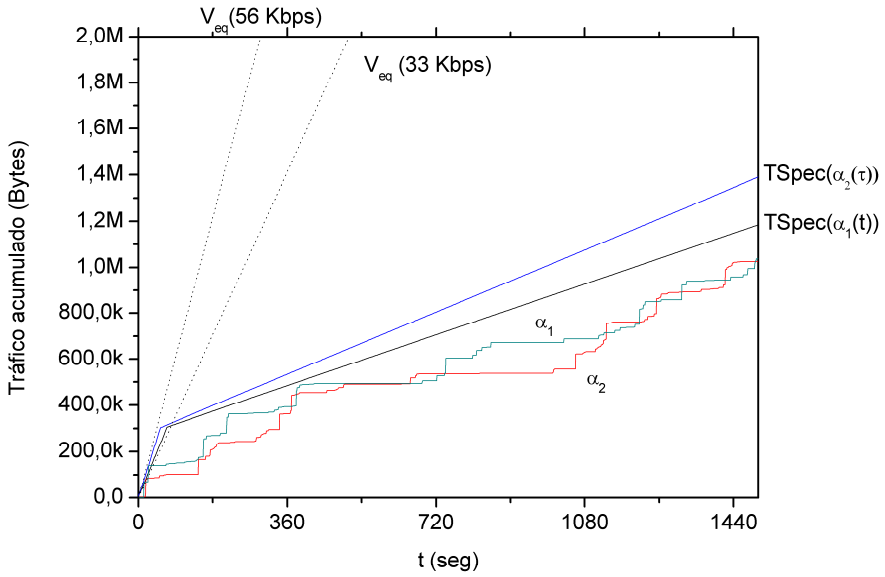


Fig. 3.53: Comparación de estimaciones obtenidas mediante *Network Calculus* y del modelo ON-OFF multinivel

Las fuentes ON-OFF han sido modeladas con valores de anchos de banda de accesos comerciales, sobre acceso telefónico. Son valores bajos si se comparan con los que existen actualmente, pero eran las velocidades típicas del año 1998, del que datan las capturas utilizadas. Al comparar las capacidades equivalentes resultantes, con las tasas de servicio P_0 de las *TSpec* calculadas, se observa como, en ambos casos, sus valores andan muy cercanos, con diferencias del 17 y 8 % respectivamente. Aparentemente, los resultados de ambas propuestas coinciden, por lo que podrían ser utilizados indistintamente. Sin embargo, en ambos casos, su uso debe de realizarse teniendo en cuenta las deficiencias que también presentan.

Aún así, las ventajas del método gráfico, presentado en este apartado, se encuentran en la posibilidad de hacer las estimaciones sobre fuentes independientes y, posteriormente, obtener los valores para diferentes agregados, sin necesidad de obtener capturas específicas del mismo. En un estudio estratégico puede resultar muy interesante para el dimensionado inicial de puntos de agregación sobre tecnologías o servicios no implantados, a partir de figuras de tráfico teóricas, como las vistas hasta ahora.

3.5. Conclusiones

En este capítulo, se han presentado las siguientes cuatro propuestas para el dimensionado de fuentes de tráfico Internet y sus correspondientes agregados, mediante estimaciones basadas en dos puntos de vista muy diferentes:

- Fuente ON-OFF multinivel: Estudiar el comportamiento de la fuente en los niveles de ráfaga, sesión y conexión, que modela mediante una fuente ON-OFF en cada nivel.
- Fuente ON-OFF multinivel binomial: Aplicar la agregación de fuentes ON-OFF multinivel introduciendo el comportamiento multifuente haciendo una estimación del número de fuentes activas en cada nivel mediante el uso de la aproximación binomial.
- Estimación de la capacidad equivalente de una fuente mediante curvas de llegada: Utilizar la observación de curvas de tráfico real acumulado para obtener los límites determinísticos del ancho de banda requerido.
- Estimación del ancho de banda agregado mediante curvas de llegada: Utilizar las estimaciones a nivel individual de las fuentes para obtener una curva límite equivalente a la de la agregación.

Los dos modelos teóricos (ON-OFF multinivel y ON-OFF multinivel binomial) que se han propuesto, basados en el mismo razonamiento: considerar el servicio como una fuente ON-OFF de tres niveles, ráfaga, sesión y conexión se pueden percibir bastante académicos desde el punto de vista de la planficicación estratégica real, pero permiten obtener una primera estimación sobre las capacidades equivalentes de fuentes en agregados homogéneos. Por ello, la segunda propuesta estudia el modelo ON-OFF multinivel desde el punto de vista de la agregación, realizando la estimación del número de fuentes activas en cada nivel mediante el uso de la aproximación binomial.

Por su parte, se ha sugerido el uso de métodos ya conocidos, basados en el *Network Calculus*, como la estimación basada en límites deterministas a partir de la observación directa de tráfico. La propuesta hace uso de un algoritmo de optimización que ha sido adaptado para la estimación de las capacidades equivalentes, y que ha dado origen a una variante aplicable a servicios con requerimientos bajos de calidad de servicio, denominados *best-effort*. El tráfico agregado también ha sido estudiado, siguiendo la línea comentada anteriormente, aunque solamente se ha comprobado el uso de las aproximaciones, ya existentes, al combinarlas con los resultados de las curvas individuales. Su uso en la planificación estrategica se limita por el hecho que, tanto el tipo de los servicios, como su aplicación por parte de los usuarios, no son estáticos y cambian de un año para otro. Aún así pueden ser utilizados para contrastar los resultados de los modelos ON-OFF multinivel, usando en cada caso medidas de tráfico actualizado

4. Aplicación del modelo CASUAL para la estimación del tráfico agregado en escenarios de redes de acceso

Equation Chapter (Next) Section 1 Tal como se definió en el capítulo anterior, el modelo CASUAL propone la configuración de escenarios de red mediante el uso de un eje de referencia tridimensional en base a tres características fundamentales: usuarios, accesos y servicios. Desde el punto de vista teórico, este concepto parte de la idea general de que el tráfico Internet tiene su origen en el uso de determinados servicios (aplicaciones) que, a priori, presentan características bien definidas por la arquitectura de protocolos TCP/IP. Sin embargo, aunque la estructura de los diferentes protocolos se mantiene constante, su comportamiento estadístico va sufrir considerables variaciones en función de los parámetros físicos de la red, y que en el caso de las redes de acceso, pueden resumirse en el ancho de banda, el retardo y el jitter. Puesto que los dos últimos pueden venir fijados por la propia aplicación (por ejemplo en aplicaciones de VoIP), el ancho de banda en el acceso actúa como modulador del comportamiento del tráfico, y de sus propiedades, pudiendo determinar la idoneidad de un tipo de acceso determinado. Así, la relación que se establece entre los diferentes servicios, la velocidad física del acceso y el tipo de acceso puede resumirse en la Fig. 4.1

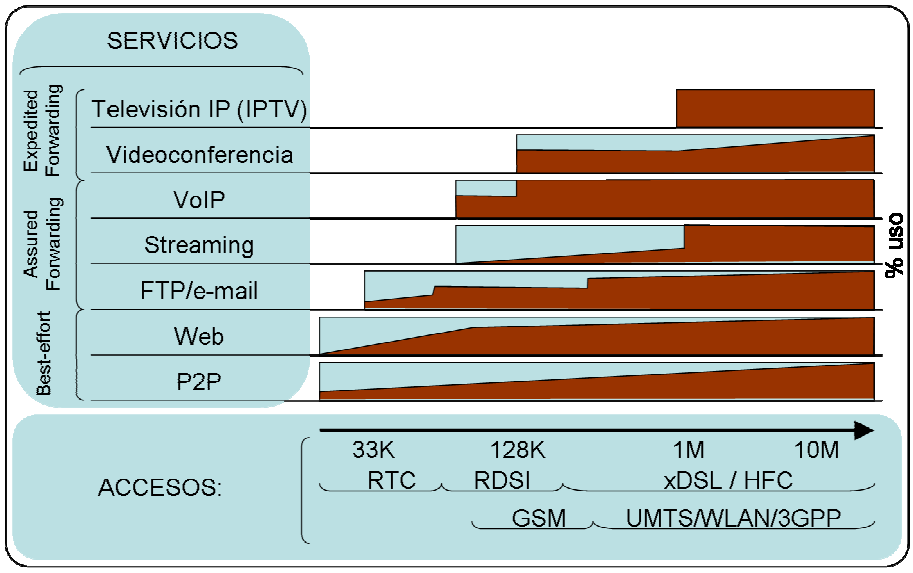


Fig. 4.1: Utilización de servicios Internet en función del ancho de bada del acceso

Existe una relación directa entre el tipo de acceso contratado y el de servicios que soporta, como se intenta representar en la fiura anterior, donde la parte sombreada representa el porcentaje de uso del acceso en función de su tipo y del servicio utilizado. En función del ancho de banda disponible, y conforme aumenta, el comportamiento del servicio mejora, y sus requerimientos de ocupación del mismo disminuyen.

Por otro lado, los usuarios son los que tienen la última palabra, tanto en el tipo de acceso que contratan, como en el uso que hacen de su ancho de banda. Aquí intervienen, fundamentalmente, el poder adquisitivo y las necesidades específicas, todo ello polarizado, por supuesto, por los gustos personales. Sin embargo, de todos esos factores, destaca la división de usuarios en residenciales y de negocios, con todas sus posibles subdivisiones en función de otros factores. Las empresas utilizan Internet para usos muy concretos, fundamentalmente para sus comunicaciones internas/externas, por lo que requieren cierto grado de servicio y optan por accesos específicos de "gama alta", o como denominan los operadores, *Premium*, oro, línea empresas, etc, normalmente asociadas a tecnologías de banda ancha. Por su lado, el usuario residencial se deja llevar por aquellas opciones más económicas que le aseguren percibir subjetivamente un mínimo de calidad de servicio, lo que actualmente se denomina QoE(*Quality of Experience*). La Fig. 4.2 resume esta idea:

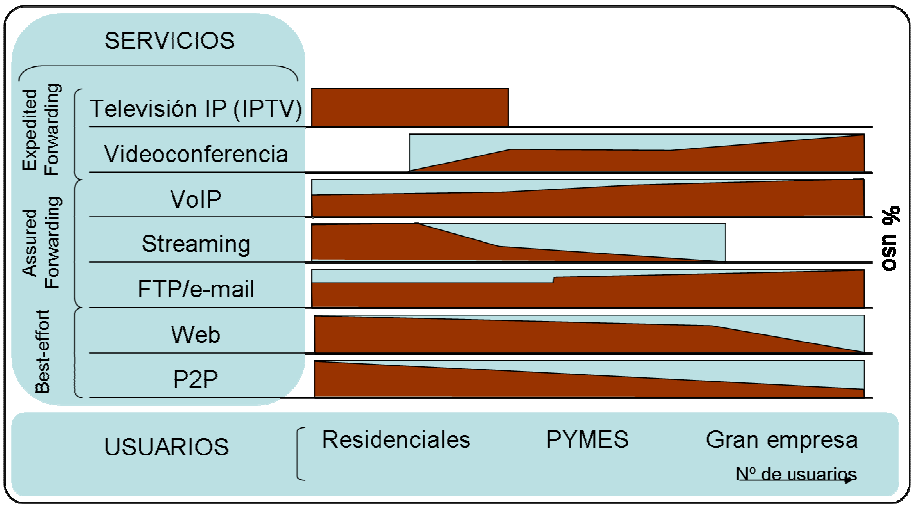


Fig. 4.2: Empleo de servicios Internet en función del tipo de usuario

Al final, y hasta cierto punto, el usuario residencial hace una valoración más simple, pero muy similar, de la relación servicio/acceso. Para ello, tomas sus aplicaciones favoritas como referencia, es decir, basa su decisión en la experiencia (personal o de otros usuarios). Por supuesto, el usuario de negocios sigue basando sus decisiones en la mejor relación productividad/coste, especialmente para aquellas aplicaciones que le suponen algún beneficio (los grandes ejemplos son la reducción de costes por asistencia a reuniones que ha posibilitado la videoconferencia, o la posibilidad del teletrabajo, que supone en muchos casos una productividad 24/7 – 24 horas los 7 días de la semana).

Precisamente, su predilección frente a determinadas aplicaciones hace que los usuarios escojan de acuerdo a los requerimientos específicos de las mismas, que solamente cumplen determinados tipos de acceso y, por tanto, resultan elegidas como se resume en la Fig. 4.3.

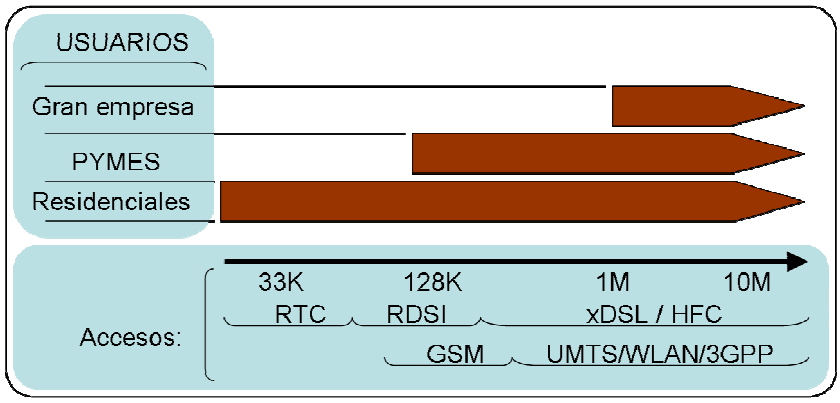


Fig. 4.3: Preferencias de tecnologías de acceso en función del tipo de usuario

Cualquiera de las tres opciones anteriores es utilizada como referencia para clasificar el tráfico Internet y establecer conjuntos de parámetros (comunes y específicos),

como se muestra en la Fig. 4.4, y cuyos valores dependen de la combinación correspondiente servicio/acceso, servicio/usuario, usuario/ acceso).

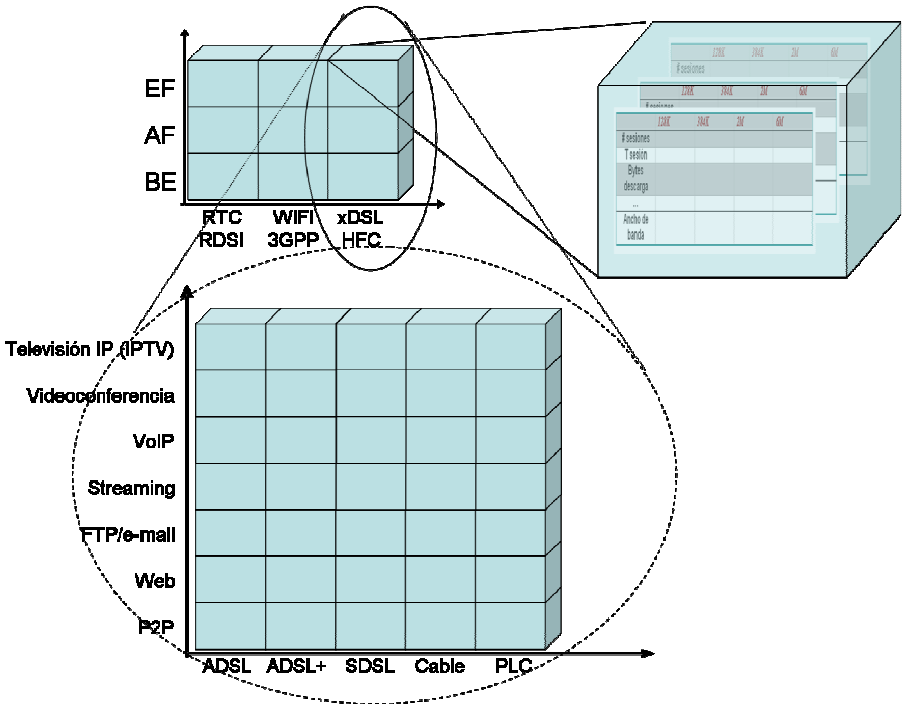


Fig. 4.4: Clasificación de servicios en función de tecnologías de acceso. Cada par (servicio, acceso) presenta parámetros con valores diferenciados, útiles para su dimensionado

Las variables comunes, independientemente del par combinado, permiten establecer un grado más de complejidad, al permitir comparar dos a dos cada relación. La Fig. 4.5 muestra cómo las relaciones dos a dos se relacionan también entre sí. De esa manera sus valores pueden ser identificados mediante ternas, resultado de la combinación de los tres ejes.

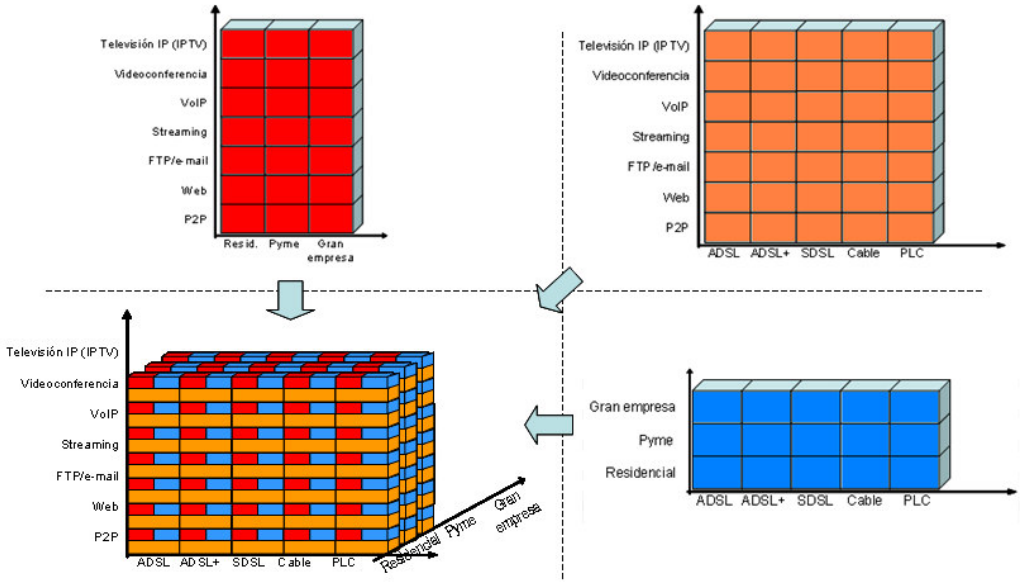


Fig. 4.5: CASUAL utiliza las clasificaciones bidimensionales para crear una clasificación tridimensional. Cada tripleta (servicio, usuario, acceso) presentan valores diferenciados en sus parámetros

Como resultado, cada terna representa a un conjunto de valores, que a su vez describen un escenario concreto (véase Fig. 4.6) y la suma de todas las tripletas define todos los posibles casos particulares dentro del escenario completo de una red. En algunos casos especificados al detalle (por ejemplo usuarios individuales), en otros mediante agrupación de los casos más comunes (todos los usuarios de un determinado área con el mismo tipo de acceso), o incluso, directamente, sobre los elementos de concentración y agregación (por ejemplo un DSLAM con soporte para diferentes velocidades de ADSL/SDSL/VDSL).

La mayoría de las combinaciones son comunes en la mayoría de redes de acceso existentes, por lo que además facilita la reutilización de todos los datos y valores, previamente calculados, a medida que se definen nuevos escenarios y se completan nuevas tripletas. De esta manera la aplicación de CASUAL, en una herramienta informática, permite generar una biblioteca de escenarios y sub-escenarios de referencia.

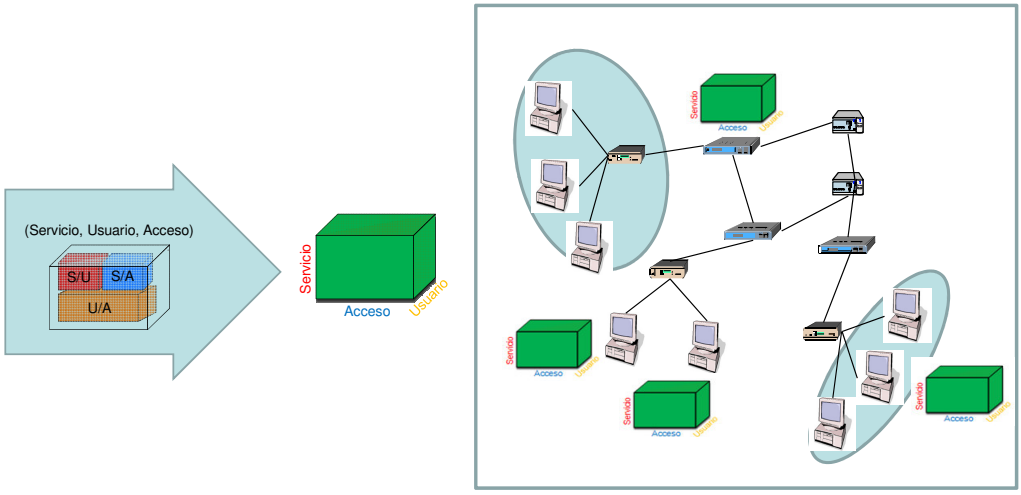


Fig. 4.6: Definición de escenarios mediante el modelo CASUAL. Cada elemento de red, terminal de usuario, concentrador o punto de agregación puede ser caracterizado mediante un conjunto de tripletas que definan el tráfico que lo atraviesa, tanto individual como agregado

4.1. Modelado de servicios Internet según el modelo CASUAL

A lo largo de esta Tesis se ha desarrollado un ejemplo de aplicación, no sólo del modelo CASUAL, sino también de los métodos de estimación de tráfico propuestos, con el objetivo de implementar una herramienta que permita la estimación del tráfico agregado en una red de acceso. La idea fundamental es que el tráfico que generan los usuarios conectados puede ser estimado de forma individual en función de los servicios que utilizan, y posteriormente, calcular tanto el tráfico agregado por servicio y usuario, como el que atraviesa por el primer punto de agregación. Tal como se ha indicado anteriormente, la aplicación del modelo puede realizarse a lo largo de uno, dos o los tres ejes, de forma totalmente independiente. Sin embargo, la influencia de los ejes de tipo de usuario y de acceso son meros modificadores del comportamiento del tráfico correspondiente a los diferentes servicios. Es por ello que, para el modelado, se propone el eje de servicio como principal, tomando como referencia un conjunto mínimo de servicios, considerados como los más importantes actualmente en Internet:

- Servicio *Web*: Entendido como la navegación y búsqueda de información, que puede incluir cualquier tipo de contenido, aunque con ciertas limitaciones en el caso de contenidos multimedia, ya que en la mayoría de los casos la transferencia se realiza haciendo uso de otro tipo de servicios más específicos.
- Servicio de VoIP: Mecanismo de comunicación alternativo a la red telefónica tanto fija como móvil. Servicio emergente que está siendo

adoptado por los usuarios de negocio, inicialmente como solución intra-empresa.

- Servicio de Videoconferencia: Mecanismo de comunicación avanzado, muy desarrollado en el entorno empresarial como alternativa de medio de comunicación inter-empresas.
- Servicio de *Streaming*: Tanto en su versión de audio, como de vídeo. Es el mecanismo más utilizado para el acceso a contenidos multimedia a través de internet, incluida la transmisión en tiempo real. Puede estar asociado al servicio Web, sobre todo para el acceso al contenido multimedia de Portales.
- Servicio de Transferencia de Ficheros: Aunque inicialmente ha sido considerado como un servicio independiente (FTP), actualmente suele estar asociado al servicio Web, y más recientemente al servicio P2P.
- Servicio de correo electrónico: Lo mismo que en el caso de FTP, el correo electrónico está cada vez más asociado al servicio Web, dada la facilidad que supone el uso de dicho servicio como mecanismo de acceso a las cuentas de usuario.
- Servicio P2P: Aunque puede parecer una combinación de varios servicios de los anteriores, el intercambio de contenidos *peer-to-peer* presenta características únicas, por lo que el tráfico generado, junto con el volumen de peticiones, lo hace característico.

Cada uno de estos servicios tiene que ser tratado de forma individual. Era preciso proponer mecanismos específicos para el cálculo y estimación del tráfico generado a partir de los modelos y aproximaciones vistos en los capítulos 2 y 3. Por ello, se ampliaron los estudios de modelos de tráfico particularizados en cada servicio en concreto. En este capítulo se incluye el estudio completo realizado en el caso del servicio *Web*, por ser uno de los más populares y estudiados. El resto de servicios, en concreto los servicios de VoIP, videoconferencia y *streaming* también fueron objeto de estudio, y por ello se incluyen en los anexos.

4.1.1. Estructura y modo de operación de la aplicación

La herramienta puede definirse como una base de datos de escenarios, mediante la cual identificar casos particulares y obtener así, las figuras de tráfico agregado tras aplicar los métodos y soluciones relacionados. El núcleo principal está constituido por una base de datos relacional, con soporte SQL, en donde se mantiene el Cubo de Servicios de CASUAL. Una interfaz gráfica permite el acceso a formularios de consulta, típicos de cualquier base, y a un módulo de definición de escenarios, dividido a su vez en un gestor de tripletas y el gestor del escenario. Los gestores tienen acceso a módulos de cálculo, desarrollados a partir de los modelos propuestos en la Tesis, cuatro en total, dos para el cálculo de capacidades equivalentes individuales, y dos para la estimación de tráfico agregado. En la Fig. 4.7 se observa esta estructura.

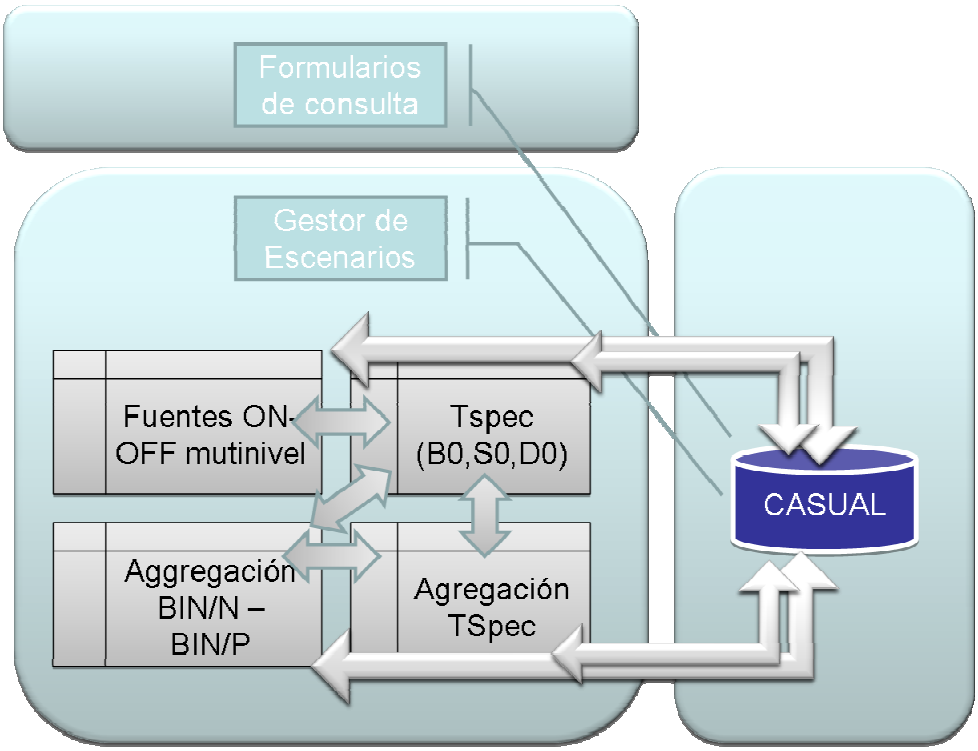


Fig. 4.7: Estructura de una herramienta de planificación basada en el modelo CASUAL

Básicamente, el modo de operación es el siguiente:

1. Definición tripletas específicas: Se parte de un servicio concreto y se especifican los accesos y tipos de usuario que hacen uso de él. Para ello, el gestor de tripletas introducirá datos de dos formas diferentes:
    a. Parametrizado: Se introducen todos los valores correspondientes a los parámetros del modelo que define al tráfico que genera dicho servicio para cada caso en particular.
    b. Captura: Se introducen trazas reales de tráfico, correspondientes al servicio y a las condiciones deseadas.
2. Definición de escenario: Mediante el gestor se seleccionan las tripletas que caracterizan a cada elemento de red, usuario, acceso, o servicio, de los cuales se desea realizar la estimación correspondiente.
3. Cálculo del escenario: A partir de los datos recopilados se procede a aplicar los modelos correspondientes. Existen dos opciones:
    a. Opción 1: Partiendo del parametrizado, se hace uso de los modelos de servicios basados en fuentes ON-OFF multinivel para el cálculo de las capacidades equivalentes, en función del tipo de usuario o de acceso (*a1*), o bien, se aplica el modelo de agregación binomial para la estimación del tráfico agregado total del servicio (*a2*).
    b. Opción 2: Se utilizan capturas reales para, a partir del *Network Calculus,* realizar estudios individuales para la estimación del tráfico equivalente por usuario (*b1*), o bien, mediante trazas de tráfico agregado obtener el tráfico agregado por servicio o usuario (*b2*).

4. La combinación de ambas opciones también devuelve resultados interesantes. Si tras calcular tráficos individuales (*a1* y *b1*) se aplica el cálculo de *TSpec* y *Summed TSpec*, el resultado es la estimación de tráfico agregado por grupo de usuarios/servicios/accesos.

Sin embargo, todo el sistema está soportado por el modelado individual de cada servicio, a partir del cual establecer el conjunto de parámetros fundamental, y los criterios de cálculo específicos a aplicar. A continuación se expone, como ejemplo ilustrativo, el estudio realizado para el modelado del servicio *Web* y la definición del conjunto de parámetros que componen la tripleta correspondiente.

4.2. Modelado de tráfico Web

Es sin duda el servicio por excelencia de Internet, utilizado por todo tipo de usuarios y con unos requerimientos de ancho de banda relativamente bajo, aunque la proliferación de contenidos multimedia (especialmente distribución de música y/o vídeo) ha hecho fundamental el desarrollo y utilización de redes de acceso de cada vez mayor capacidad.
Desde sus orígenes, el servicio *Web,* ha seguido fielmente la filosofía cliente/servidor impuesta por la arquitectura TCP/IP. Los servidores *Web* almacenan un conjunto de contenidos (texto, imágenes, audio, vídeo, secuencias de comandos, etc.) que son entregados según se aceptan las peticiones cursadas por los correspondientes clientes, los conocidos navegadores o *browsers*.

Para ello, el cliente establece una conexión TCP con el servidor en el puerto 80, a través de la cual, el protocolo HTTP realiza el intercambio de la petición y la correspondiente respuesta, de acuerdo con el denominado modelo de Hipertexto, según el cual el propio texto permite establecer enlaces que referencian documentos situados en servidores remotos.

Una vez establecida la conexión TCP, el cliente envía una petición HTTP en la que se incluye un comando de operación (generalmente GET) y la dirección URL[1] del objeto solicitado, en la mayoría de los casos lo que normalmente se conoce como página *Web*.

Los objetos intercambiados mediante HTTP pueden ser prácticamente de cualquier naturaleza, aunque normalmente se hace uso de lenguajes de alto nivel que

[1] URL : *Uniform Remote Location*. Esta notación expresa de manera uniforme los distintos recursos que podemos acceder con el cliente *Web* (RFC 1738 y RFC 1808). Consta de varios campos mediante los cuales se indica el protocolo de intercambio (HTTP en el caso web), el servidor destino y el objeto solicitado (página *Web* / HTML, texto, imagen, etc)

permiten organizar, dar formato y establecer enlaces hacia los diferentes objetos intercambiados. El lenguaje Web por excelencia es HTML (*HyperText Markup Language*), aunque actualmente existen gran número de variaciones y alternativas, como se puede observar en la Fig. 4.8, destacando XML y XHTML:

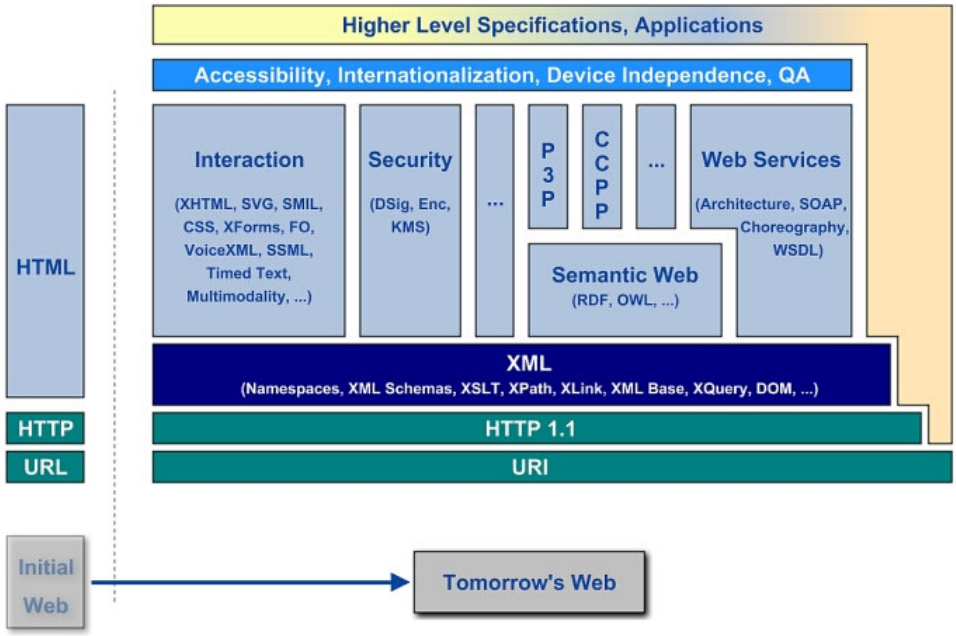


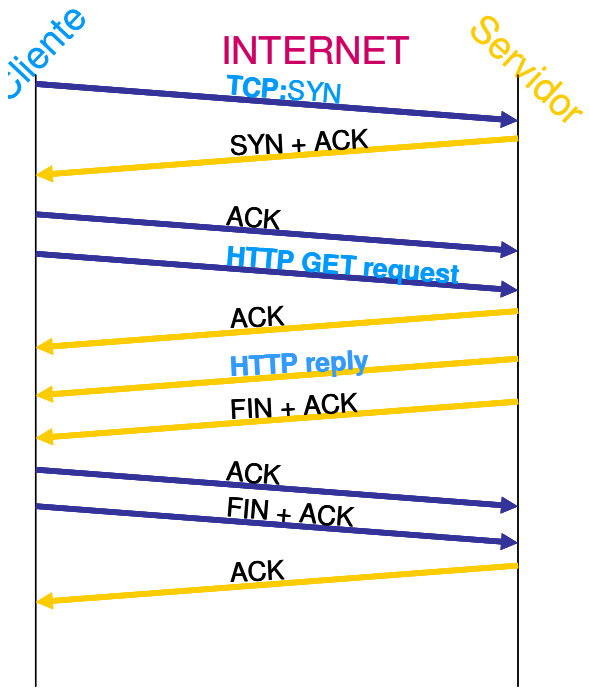
Fig. 4.8: Protocolos y estándares relacionados con el servicio Web (fuente: W3C)

- **XML (*eXtensible Markup Language*):** Considerado precursor de la Nueva Generación *Web*. Integra lenguajes diseñados a medida en una infraestructura XML común.
- **X)HTML(*eXtended HyperText Markup Language*):** Como su nombre indica es una evolución de HTML que utiliza la sintaxis de XML. Se basa en los denominados *XHTML profiles*, los cuales pueden ser adaptados en función del tipo de dispositivo cliente utilizado.

Independientemente del uso de HTML, XML, etc., una página Web típica consiste en un **objeto principal** que contiene el documento HTML / XML y que es el primero en ser descargado. Este documento puede hacer referencia directa a otros objetos que serán descargados tras finalizar la descarga del principal, y que se denominan ***in-line***. Suelen ser imágenes, sonidos, módulos de código, o incluso objetos embebidos.

Precisamente, la forma en que se obtienen los objetos *in-line* es la principal diferencia entre las diferentes versiones del protocolo HTTP, definidas en las RFC 1945 y RFC 2068, versiones 1.0 y 1.1 respectivamente. Originalmente, tras la petición GET, el servidor realizaba el envío del objeto principal, el documento HTML por ejemplo, cerrando la conexión TCP. Si el objeto principal hacía referencia a objetos *in-line*, el cliente debería establecer una nueva conexión TCP con el servidor para cada uno de los objetos *in-line* referidos. El comportamiento de la versión 1.0 de HTTP se muestra en la Fig. 4.9:

Fig. 4.9: Establecimiento de la conexión TCP mediante HTTP 1.0



Este comportamiento resulta totalmente ineficaz, especialmente conforme aumenta el número de objetos *in-line* que componen la página *Web*. De acuerdo con el protocolo TCP, cada nueva conexión introduce el correspondiente retardo asociado al saludo a tres vías (*three-way-handsaking*) en su establecimiento. A esto hay que añadir el aumento en el número de segmentos intercambiados durante la transferencia de cada objeto, salvo que el TCP desactivara la opción de arranque lento (*slow-start*). Como solución aparece la versión 1.1 de HTTP, que aprovecha la conexión TCP inicial siempre que haya objetos *in-line* posteriores que transmitir (este método recibe el nombre de conexión persistente). De esta forma, tras recibir el objeto principal, el cliente puede enviar el comando de petición del primer objeto *in-line*. Esta técnica elimina los retardos de establecimiento de múltiples conexiones, aunque requiere finalizar cada intercambio de objetos *in-line* antes de solicitar el siguiente. Por ello, HTTP v.1.1 introduce un modo de funcionamiento *pipelining*, según el cual, las peticiones de objetos *in-line* no tienen porque realizarse de forma secuencial, pudiendo ser solicitados simultáneamente sin necesidad de esperar el final de cada transferencia, tal como muestra la Fig. 4.10. Sin embargo, estas mejoras solamente son útiles siempre y cuando el tiempo de persistencia de las

conexiones no sea elevado, y en el caso de permitirse varias conexiones que aprovechen el *pipelining*, que sean finalizadas según termine la transferencia de la página *Web* completa [Liu, 1999].

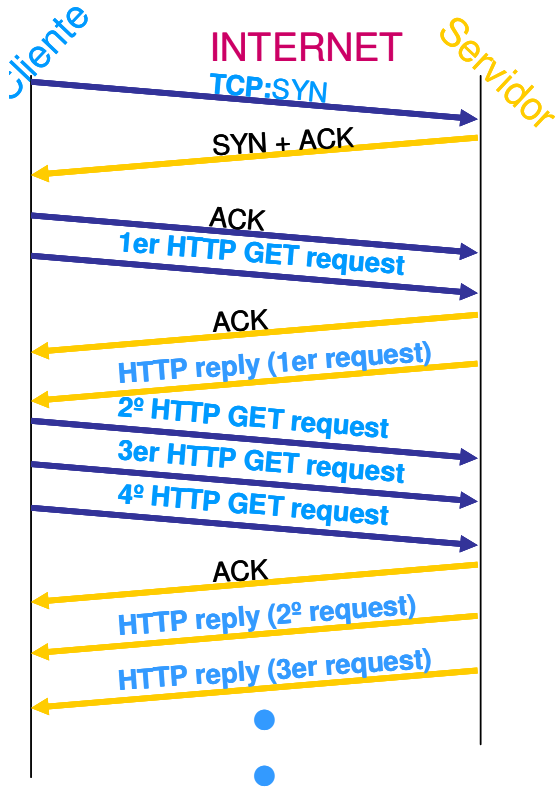


Fig. 4.10: Establecimiento de una conexión TCP mediante HTTP 1.1

En la actualidad el servicio *Web* combina el uso de varias conexiones persistentes sobre las que se realiza el *pipelining*, como se indica en la Fig. 4.11 y se explica en [Staehle, 2003], dependiendo casi exclusivamente del tipo de navegador empleado (así por ejemplo *Explorer* utiliza 2 conexiones frente a las 4 de *Netscape*). En cualquier caso, el procedimiento seguido podría resumirse de la siguiente manera:

1. El cliente abre una conexión TCP.
2. Petición de página por parte del cliente. (*HTTP Request*).
3. Respuesta del servidor, enviando el objeto principal en uno o varios paquetes.
4. Recepción en el cliente del objeto principal, análisis y llamada (*HTTP Request*) de los objetos *in-line* referenciados en el objeto principal, y siempre que no estén previamente en caché, en la misma conexión o/y en conexiones paralelas.
5. Respuesta del servidor, que envía los objetos *in-line*.
6. Cierre de las conexiones TCP.
7. Visionado de la página.

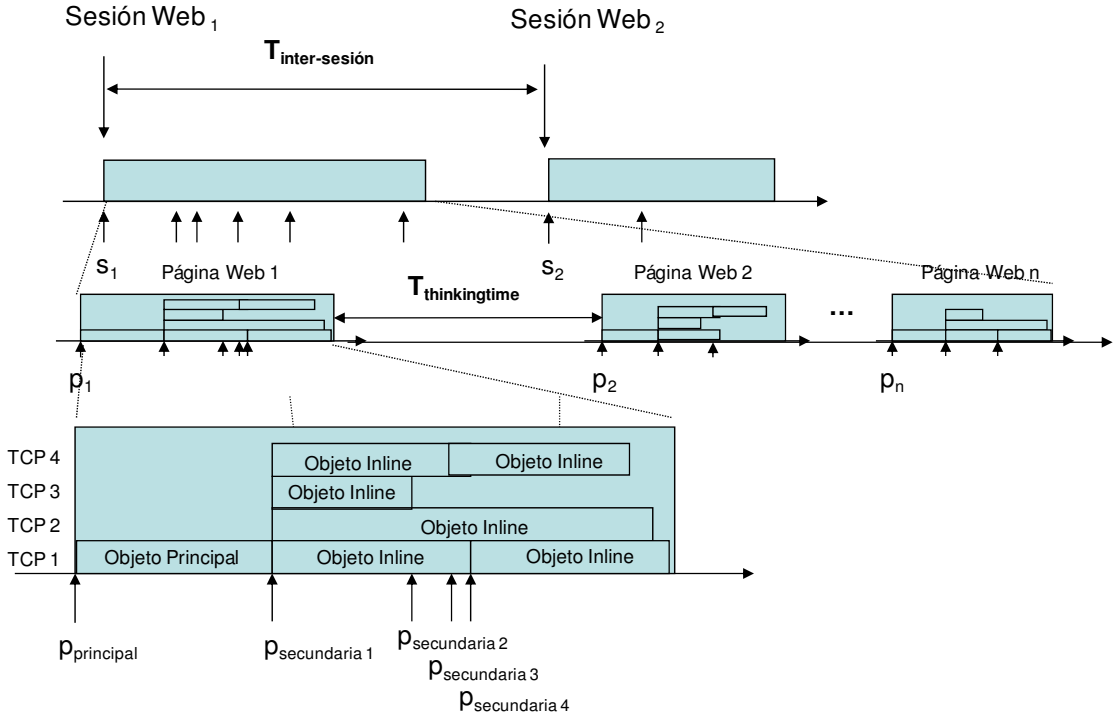


Fig. 4.11: Representación del comportamiento del servicio Web a diferentes escalas temporales (nivel de sesión, nivel de página y nivel de objetos)

Sin embargo, hoy en día existe un problema adicional ya que todas las paginas de un sitio *Web* no tienen porque residir en un mismo servidor (esto se ha venido a llamar aprovisionamiento de contenidos heterogéneos).

Un sitio *Web* se puede configurar en esta manera debido a la naturaleza de los datos y a su tamaño, además de permitir a los usuarios recordar el servidor con un solo nombre (razones comerciales). Esto se puede deber a consideraciones administrativas y a que un sitio Web puede contener datos de múltiples fuentes y diferente temática. El mecanismo de *HTTP Redirect* se usa para mantener este tipo de contenido heterogéneo. Sin embargo una redirección de un servidor a otro puede llevar, normalmente, una comunicación entre servidores de alrededor de 9-10 paquetes TCP, lo cual se podría hacer con tan solo dos paquetes UDP. De hecho, en muchos casos el servidor *Web* no es más que un intermediario, un punto de acceso que recibe todas las peticiones y utiliza un gran número de fuentes para responder a las mismas. Por ejemplo, si un fichero pedido no esta en el sistema local del servidor, este realiza la búsqueda en un sistema de ficheros de red (probablemente algo parecido a una base de datos distribuida).

Para los puntos donde existe un amplio número de usuarios de HTTP 1.0, en algunas ocasiones, se usa un *proxy* HTTP para evitar los diálogos TCP y lo sustuye por UDP (un esquema híbrido UDP-TCP) sin modificar los navegadores de los

clientes, véase[Touch, 1998]. Se opta por usar un sistema híbrido TCP-UDP, con el primero para las transferencias grandes y UDP en las pequeñas. En el caso de atravesar una red colapsada, con alto número de errores, o si el navegador no lo soporta se pasaría a utilizar TCP.

Este tipo de casos particulares no serán tratados a la hora de modelar el cálculo de este servicio, debido a la imposibilidad de conocer el estado de la red y las múltiples variantes, totalmente impredecibles, en cuanto a la posibilidad de servidores distribuidos.

4.2.1. Modelado del servicio Web

El comportamiento explicado del servicio *Web* ha sido estudiado ampliamente en [Choi, 1999] y [Liu, 1999] dadas las similitudes con los modelos de fuente ON-OFF clásicos. De esta forma, se toma como referencia una sesión, entendida esta como la "navegación" de un usuario a través de una determinada URL (esto significa que el usuario visita el *Web-site* observando páginas relacionadas directamente entre ellas, esto es, pertenecientes al mismo *Web-Server* o como mínimo pertenecientes al mismo dominio), siguiendo una pauta similar a la de la Fig. 4.12:

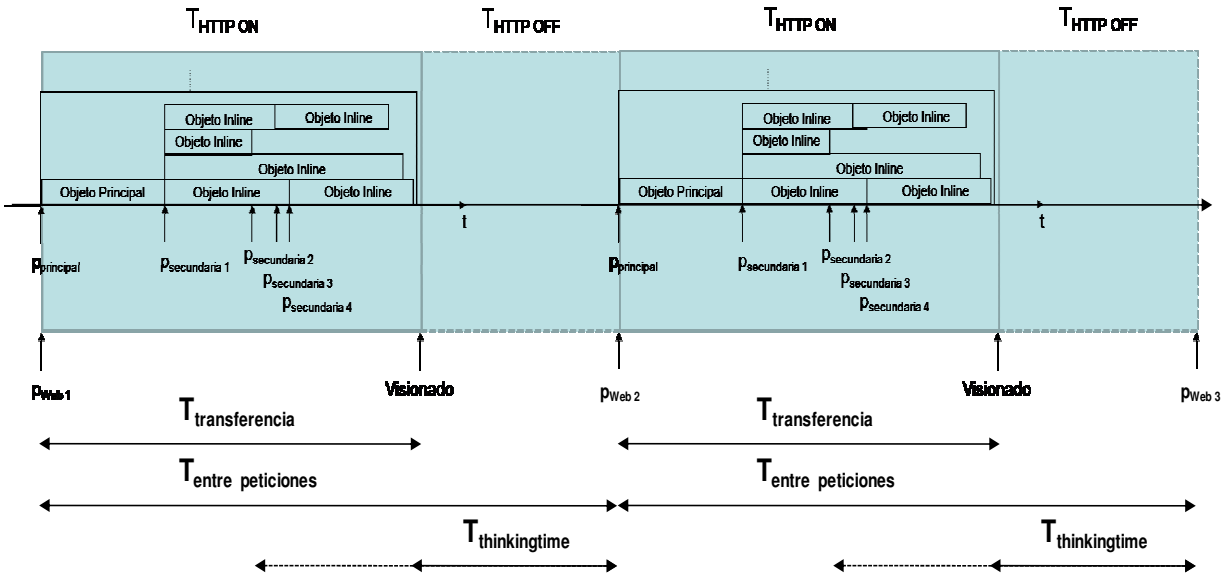


Fig. 4.12: Parametrizado de tiempos relacionados con el desarrollo de una sesión *Web*

Desde el punto de vista de una conexión, la fuente ON-OFF quedaría definida por conjuntos de parámetros fundamentales: el número de "clics" por sesión, el tiempo entre clics consecutivos (que suele denominarse *thinking time*), y las características de la transferencia y procesamiento de los objetos descargados, teniendo en cuenta que según [Barford, 1999] el comportamiento de una conexión *Web* es sensiblemente dependiente del componente software utilizado, tanto en el lado servidor, como en

el lado cliente, así como de la carga de peticiones soportada por los servidores, véase [Hava, 2000] y [Muntean, 2001].

Sin embargo, el estudio detallado de cada transferencia de datos, es decir, de cada objeto intercambiado, permite establecer un tercer nivel de análisis que se corresponde con el comportamiento en sí del tráfico IP, y en concreto, en el caso del servicio *Web* de TCP. Este nivel es el que se denomina nivel de ráfaga, y que es utilizado en [Liu, 2004] definiendo el denominado modelo ON-OFF jerárquico, comentado en el apartado 2.3.2.3, o en [Staehle, 2003], utilizando solamente los dos niveles inferiores de análisis, en lo que denomina modelo de trafico de fuente estocástico.

La Tabla 4.I muestra un recopilatorio de los primeros estadísticos de los parámetros fundamentales identificados en el servicio *Web*, así como las funciones de distribución relacionadas tal y como se propone en [Choi, 1999 / Liu, 1999 / Staehle, 2003]:

	Distribución	Media	Desviación Estándar
Nº de clicks	Gauss Inversa LogNormal Exp-normal	4.6 - 12	1.28 – 21.6
Thinking-time	Weibull Weibull-Gauss Inversa LogNormal	33.4 seg.-39.45	92.57 seg.
Tiempo TX	LogNormal Gauss Inversa – Gauss Inversa	7.03 seg	4.3 seg.
Tamaño Petición	Lognormal	360.4 B	106.52
Tamaño objeto	Lognormal	7757 – 10710 B	25032 – 126168
Nº de objetos	Gamma	5.55	11.35
Tiempo entre objetos	Gamma	0.86 seg.	2.15 seg.

Tabla 4.I: Parámetros fundamentales de una sesión *Web*

Adicionalmente en [Liu, 1999] también se plantea el uso de distribuciones de Pareto ($0.4 \leq \alpha \leq 0.63$; $k \leq 21$ Kbytes) para el modelado del tamaño de los objetos, aunque en [Mah, 1997] se distingue entre objetos individuales, modelados mediante una Pareto con $1.04 \leq \alpha \leq 1.14$), y objetos compuestos, a su vez divididos entre objetos principales (marcos) y secundarios (objetos *in-line*), modelados respectivamente por sendas distribuciones de Pareto con ($0.85 \leq \alpha \leq 0.97$) y ($1.12 \leq \alpha \leq 1.39$) respectivamente.

La aplicación de modelos tradicionales, como los utilizados en el caso del tráfico de voz, da lugar a modelos ON-OFF muy sencillos, aunque con baja resolución estadística en sus resultados. Es el caso del denominado modelo *WebUser,* véase

[Jaeger, 2000]. Este modelo ON-OFF toma cada petición *Web* como punto de partida del estado activo. De esta forma, el modelo queda definido por tres variables fundamentales: el tiempo entre peticiones dentro del período ON, duración del período OFF y duración del período ON.

Esta simplificación parte de una idea de agregación básica, esto es, considerar que el ancho de banda del sistema es compartido por todas las conexiones TCP, teniendo en cuenta que, un número determinado de nodos, genera el mismo número de conexiones TCP sin tener en cuenta su naturaleza estadística. El comportamiento de dichas conexiones también resulta muy básico, ya que los datos transferidos no tienen en cuenta los retardos ni por procesamiento en el servidor, ni por retransmisiones (la comunicación se considera libre de errores), el tamaño de los paquetes es fijo y máximo y los enlaces y puntos de agregación se consideran dimensionados idealmente.

De acuerdo con todas esas consideraciones, los parámetros fundamentales del modelo pueden ser obtenidos mediante funciones de distribución de Weibull[1] (tiempo entre llegadas y duración del ON) y Pareto[2] (duración del OFF y tamaño medio de paquetes) tal como se indica en la Tabla 4.II.

	Distribución		α	β	Media	Varianza
Período ON	Weibull	Mañana:	0.909	$e^{4.4}$	85.2 s	93.9 s
		Tarde:	0.769	$e^{4.6}$	116 s	146 s
Período OFF	Pareto	Mañana:	0.577	52,7		
		Tarde:	0.903	62,8		
Tiempo entre llegadas	Weibull		0.5	1.4	2.8 s	6.26 s
Tamaño de paquetes	Pareto		1.06	2500	40.6 KB	

Tabla 4.II: Valores típicos de los parámetros fundamentales del modelo *WebUser*

La necesidad de caracterizar estadísticamente el comportamiento asimétrico del servicio *Web* hace que, en [Sousa, 2000] se plantee un modelo ON-OFF más complejo. Para ello establece dos estados ON, uno para cada sentido de la comunicación, ON/U (*upstream* o navegador-*Webserver*) y ON/D (*downstream* o *Webserver*-navegador), como se muestra en la Fig. 4.13:

[1] $f_{WEIBULL}(x) = \alpha\beta^{-\alpha}x^{\alpha-1}e^{-\left(\frac{x}{\beta}\right)^{\alpha}}, \forall \quad \alpha > 0 \quad \beta > 0$

[2] $f_{PARETO}(x) = \alpha\beta^{\alpha}x^{-(\alpha+1)}$

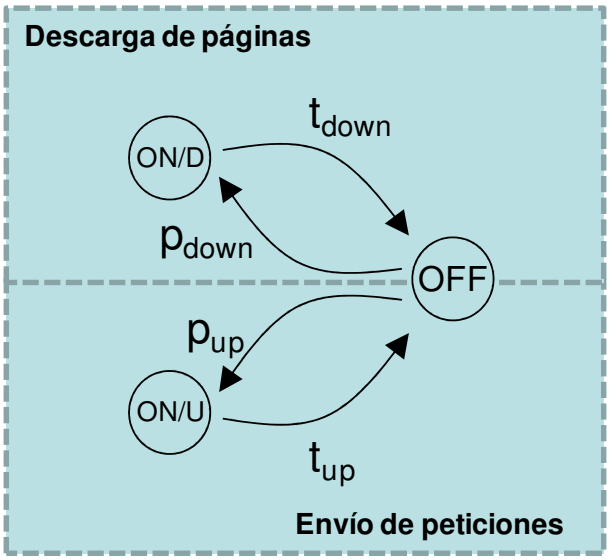


Fig. 4.13: Modelo de ON-OFF de doble canal: *Downstream/ Upstream*

Se puede obtener el correspondiente modelo de Markov y obtener el tiempo medio de inactividad como:

$$t_{OFF} = \left(p_{down} t_{down} + p_{up} t_p \right) \Big/ (1 - \Phi_F) \quad \forall \quad \Phi_F = t_{XN} t_{on} \Big/ N_f \qquad (4.1)$$

siendo p_{down} y p_{up} las probabilidades de transición del estado OFF a *downstream* o a *upstream* respectivamente, t_{down} y t_{up} los tiempos medios de permanencia en los estados de transmisión en bajada y subida, t_{XN} el número medio de aparición de nuevas fuentes activas por segundo, t_{ON} el tiempo medio de actividad de la fuente, y N_f el número medio de fuentes Web activas (tanto en *upstream* como *downstream*).

Este modelo tiene en cuenta el comportamiento autosimilar del tráfico a ráfagas al utilizar funciones de distribución de cola pesada. Sin embargo, mantiene la filosofía de los modelos ON-OFF tradicionales, aunque realmente sea uno para cada sentido de la comunicación. Por lo demás, solamente se tiene en cuenta el efecto del tiempo entre peticiones y la función de distribución del tamaño de la sesión, dejando de lado otros parámetros igual de importantes. El modelo ON-OFF multinivel propuesto resulta más complejo, pero tiene en consideración más factores que afectan al tráfico, por lo que sigue siendo la opción a considerar para llevar a cabo el modelado del servicio *Web*.

4.2.2. **Modelo CQN (Closed Queueing Network)**

Este es uno de los últimos modelos, desarrollado en [Berger, 2000], para la emulación de servicios *Web* y tráfico elástico en general, sobre enlaces con garantías de calidad, como puede ser MPLS sobre TCP/IP ó ABR sobre VPC en ATM. En este caso, el servicio no es modelado partiendo del flujo ofrecido, sino a partir del tamaño de datos intercambiados, entendidos estos como el total de la sesión. Para ello, el modelo utiliza dos bloques servidores: un servidor infinito que representa el conjunto de fuentes, y un servidor PS (*Processor Sharing*) para modelar la cola y transferencia de paquetes, como se muestra en la Fig. 4.14. Cada transición entre el servidor IS y el servidor PS representaría el comienzo de una sesión, siendo la única restricción considerada, los valores de dimensionado del enlace. El número de fuentes activas viene representado por el número total de trabajos (sesiones) existentes en el servidor PS.

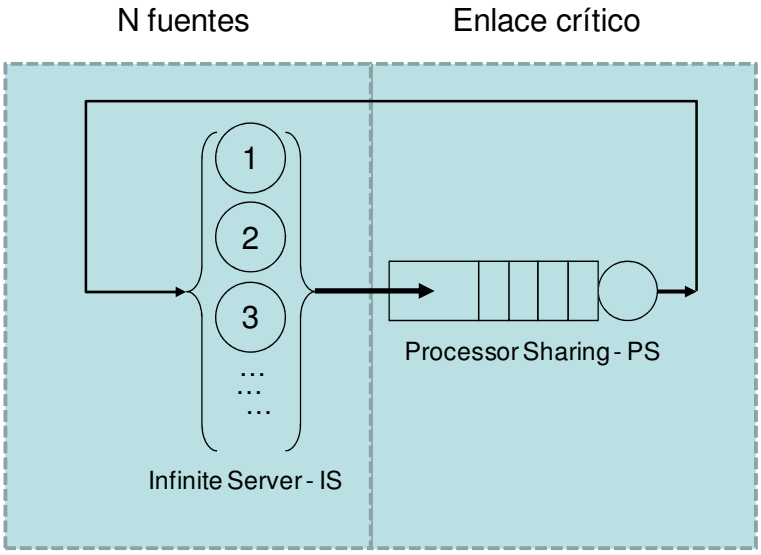


Fig. 4.14: Representación conceptual del modelo CQN (*Closed Queueing Network*)

Los parámetros de entrada del modelo son básicamente tres: el número de fuentes *N*, la tasa de servicio del servidor PS ($1/\mu$), y el tiempo medio entre transmisiones en el servidor IS ($1/\lambda$). Dado que $1/\mu$ representa el tiempo medio de transmisión durante una sesión, puede ser aproximado mediante la relación:

$$\frac{1}{\mu} = \frac{f}{B} \tag{4.2}$$

siendo *f* el total de datos intercambiados durante una sesión, y *B* la capacidad del canal.

De acuerdo con este modelo, [Berger, 2000]plantea dos soluciones para el cálculo del ancho de banda efectivo utilizado por un conjunto de *N* fuentes *Web*, con

ancho de banda medio *b*. La primera de ellas plantea el comportamiento "normal" de las fuentes, considerando que el ancho de banda utilizado por cada fuente cumple la media, es decir $B_s \geq b$, según lo cual:

$$B = Nh \qquad siendo \quad h = \left(\frac{1}{b} + \frac{1}{\lambda f} \right)^{-1} \tag{4.3}$$

con $b \in \left[10^4, 10^6\right]$

La segunda solución, tiene en cuenta el comportamiento en los límites de las fuentes, considerando que $\Pr(B_s < b) < \alpha$:

$$B = h\left[N + \gamma + \sqrt{2\gamma N + \gamma^2} \right] \quad siendo \quad \gamma = \frac{1}{2} q_\alpha^2 \frac{h}{\lambda f} \tag{4.4}$$

y q_a el (1-α) quantile de una normal de media 0 y varianza 1 ($\sqrt{N}\,{}^{\lambda f}\!/_{b} > q_\alpha$ y $B < N\lambda\overline{f}$)

Este modelo resulta muy sencillo aunque plantea algunas deficiencias, como por ejemplo que en ningún momento se tienen en cuenta ni las pérdidas ni los retardos de los paquetes. Al igual que en el modelo *WebUser*, solamente tiene en cuenta las estadísticas totales al nivel de sesión, sin considerar los efectos del nivel de ráfaga y simulando la agregación mediante un sistema de cola con prioridades, que no se corresponde con la realidad del servicio, en el que no existe la priorización de los clientes y sí varios servidores de respaldo. El modelo ON-OFF multinivel, hasta el momento, continúa siendo la opción más interesante.

4.2.3. **Modelado microscópico**

En [Heidemann, 1997 / Touch, 1998] se plantea la posibilidad de observar el tráfico Internet al nivel de bit. Para ello se parte del modelado de aplicaciones Internet en base al comportamiento, tanto de las aplicaciones en si, como de los protocolos de transporte implicados. Dicho comportamiento, es tomado como referencia para la estimación de los tiempos de sesión y de ráfaga, partiendo de un valor de tiempo de transferencia dependiente del tamaño del fichero, el tamaño de la unidad de transferencia de datos y de la ventana de transmisión, y que debe ser corregido por el comportamiento concreto del protocolo de transporte utilizado. Así por ejemplo, en el caso del servicio FTP, tras abrirse una conexión TCP de control, se realiza la petición de descarga/subida del fichero que genera una nueva apertura de conexión TCP, una para cada fichero intercambiado como datos. Ambos procesos de conexión están sujetos al mecanismo de *three way handsaking*, como se muestra en la Fig. 4.15. Es fundamental tener en cuenta que el proceso de transferencia está también controlado por el mecanismo de *slow-start*, lo cual se traduce en un retardo en los tiempos totales de transferencia, y por tanto de ráfaga por cada petición de archivo, y de sesión con cada conexión de control. En el caso más sencillo de HTTP, en el que solamente se establece una conexión (HTTP

persistente), el mecanismo *slow-start* introduce un transitorio en la transferencia de datos hasta el momento en el que, el mecanismo de *congestion avoidance* estabiliza la transferencia, lo que permite considerarla constante a partir de dicho momento. Es precisamente en este proceso en el que los parámetros fundamentales de TCP pueden ser utilizados para el modelado del tráfico generado durante la transferencia, como se muestra en la Fig. 4.16 y en la Tabla 4.III.

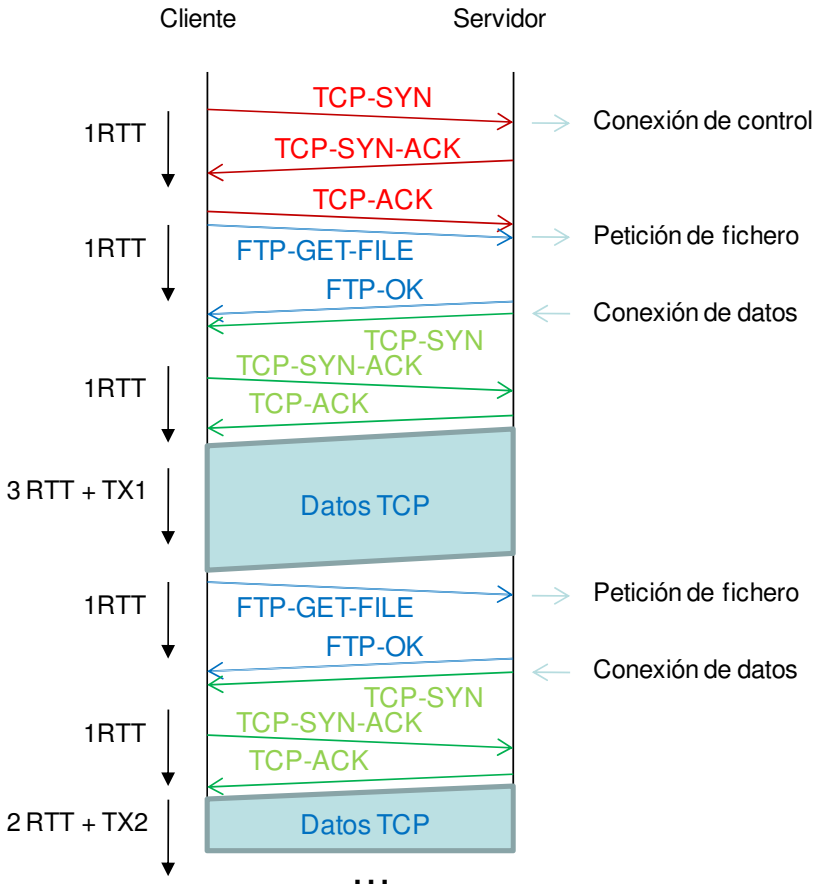


Fig. 4.15: FTP y HTTP

Parámetro		Equivalencia
RTT	Round Trip Time	
BW	Bandwidth	
MSS	Maximum Segment Size	
STT	Segment Transmission Time	STT = MSS/BW
MUWS	Maximum useful Window Size	MUWS = RTT/STT

Tabla 4.III: Parámetros fundamentales de una conexión TCP.

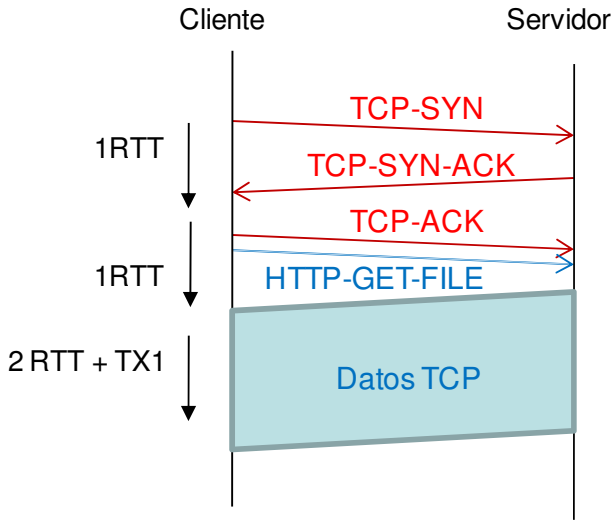


Fig. 4.16: TCP y HTT

Así por ejemplo en [Touch, 1998] se aproxima el tiempo asociado a una sesión *Web* genérica mediante la siguiente expresión:

$$S_{\min}(sesiónWeb) = RTT + \sum_{i=1}^{n}(T_{\min}(i) - RTT) \tag{4.5}$$

donde *n* es el número de objetos (peticiones) independientes, *RTT* el *Round Trip Time* y T_{min} el tiempo mínimo de transmisión. Este último puede aproximarse por:

$$\begin{aligned} T_{\min} &= conexión + petición_{\min} + procesamiento + respuesta_{\min} \\ T_{\min} &= RTT + \frac{L_{petición}}{BW} + t_{proceso} + \frac{L_{respuesta}}{BW} \end{aligned} \tag{4.6}$$

$$\begin{aligned} T_{\min} &= conexión + petición_{\min} + procesamiento + respuesta_{\min} \\ T_{\min} &= RTT + \frac{L_{petición}}{BW} + t_{proceso} + \frac{L_{respuesta}}{BW} \end{aligned} \tag{4.7}$$

Sin embargo, al introducir comportamientos específicos de TCP, como es el caso del *slow-start*, la expresión anterior quedaría como:

$$T_{TCP} = 2RTT + \frac{L_{peticion}}{BW} + t_{proceso} + [t_{slow-start} + \frac{L_{respuesta}}{BW}] \tag{4.8}$$

Con lo que la duración de una sesión HTTP sobre TCP, en la que se lleven a cabo *n* peticiones independientes, puede ser aproximada por:

$$S_{TCP} = RTT + \sum_{i=1}^{n}[T_{TCP}(i) - RTT] = S_{\min} + nRTT + \sum_{i=1}^{n}[t_{slow-start}(i)] \tag{4.9}$$

Existen implementaciones de servicios *Web* que utilizan UDP como protocolo de transporte, como por ejemplo ARDP (*Asynchronous Reliable Delivery Protocol*), expuesto en [M16]. En estos casos, aunque se elimina el efecto del *three-way-*

handshaking, se mantiene el mecanismo de *slow-start* como técnica de control de congestión, con lo que el tiempo mínimo de transmisión queda ahora como:

$$T_{UDP} = T_{TCP} - RTT \tag{4.10}$$

Y la duración de una sesión con *n* peticiones independientes puede aproximarse mediante la expresión:

$$S_{UDP} = RTT + \sum_{i=1}^{n}\left[T_{UDP}(i) - RTT\right] = S_{\min} + \sum_{i=1}^{n}\left[t_{slow-start}(i)\right] \tag{4.11}$$

Tomando como referencia los valores de [Mah, 1997] para el tamaño medio de una petición *Web* de 240 Bytes (con máximo de 320 Bytes), para el tamaño medio de un objeto *in-line* de 1600 bytes, y supuestos unos tiempos medios de 200ms tanto para el *RTT* como para el tiempo de procesado (*server thinking time*) se obtiene la duración media de la transferencia HTTP, en función del número de peticiones y el ancho de banda utilizado en la conexión.

Este es quizá el método que permite el modelado más detallado. Se tiene en cuenta el comportamiento real de las aplicaciones implicadas en el servicio, al nivel de paquete IP e introduciendo el control de ventana deslizante de TCP. Sin embargo, para su resolución se parte de sesiones tipo, y por tanto, resultaría correcto si la varianza del tamaño de los objetos in-line no fuera grande, cosa que no es cierta. De esta manera, resulta complicado obtener un conjunto de parámetros de entrada que permita obtener una estimación fiable sobre una sesión Web genérica.

4.2.4. **Parámetros asociados al tráfico Web**

Finalmente, vistos los modelos específicos implementados hasta el momento, la opción del modelo ON-OFF multinivel se sitúa como firme candidato para ser utilizado en el modelo CASUAL. Para ello, es preciso determinar el conjunto de parámetros a partir de los cuales poder aplicar dicho modelo. A lo largo de los apartados anteriores, se han especificado parámetros concretos y, en su caso, mecanismos para la estimación de los valores correspondientes a cada uno de los niveles temporales estudiados:

En el Nivel de Conexión:

- Número de conexiones por hora (α): Número de conexiones por hora y en un día normal que realiza un usuario del tipo contemplado.
- Duración de la conexión (T_s): Tiempo que el usuario permanece conectado al ISP de manera continuada.
- Velocidad del acceso (*Vac*): Ancho de banda de la línea de acceso a Internet.

Para el Nivel de Sesión:

- Tiempo de transferencia (T_{pagina}): Tiempo que tarda una página Web solicitada en ser descargada.
- Tiempo de visionado ($T_{thinking}$): Es el tiempo durante el cual el usuario no realiza otra petición de página.
- Tiempo entre sesiones: Tiempo que pasa entre dos sesiones consecutivas pedidas por el usuario. Normalmente suele ser igual a $T_{pagina}+T_{thinking}$

Finalmente, al Nivel de Ráfaga:

- Tiempo entre ráfagas (T_{idle}): Tiempo que pasa entre la descarga de dos objetos consecutivos pertenecientes a una página Web.
- Duración de ráfaga (T_{on}): Tiempo de descarga del objeto.

4.2.5. Estimacion del tráfico generado por fuentes *Web*

A continuación, se hace una demostración de la aplicación del modelo ON-OFF multinivel binomial para el modelado de una tripleta de las definidas en CASUAL en el caso del servicio *Web*. El modelado del nivel de sesión queda definido prácticamente, por el número de peticiones de páginas que realiza el usuario. Dicho número es necesario para el cálculo del tiempo de duración de la sesión, aunque no de forma determinante ya que va a variar enormemente en función de: la velocidad del acceso a Internet, del tamaño de las páginas solicitadas, e incluso del estado de la red.

Por otro lado, en el nivel de ráfaga se realiza la descarga de cada uno de los objetos *in-line* pertenecientes a cada uno de los "marcos de página" principales, y que representan en definitiva a la página *Web*.

Ambos niveles son semejantes y puede modelarse su agregación mediante binomiales negativas, ya que los dos niveles presentan el comportamiento de ráfagas de naturaleza muy similar, lo cual implica la consabida característica de autosimilaridad de este tipo de tráfico.

Para su cálculo, se va a partir del número de usuarios de un determinado tipo, que hacen un uso similar del servicio *Web*, y que, con muy alta probabilidad, en media presentan valores muy similares en el T_{idle} (tiempo entre ráfagas) y el T_{on} (duración de la ráfaga). Es importante notar que no se necesita conocer el tiempo exacto de duración de las ráfagas y de los tiempos de espera, sino la relación entre ellos, por lo que se puede utilizar sus sumas dentro de una sesión para calcular las probabilidades correspondientes (utilizando T_{pagina} y $T_{thinking}$).

El desarrollo matemático es idéntico3.3.1, por lo que aquí solo se va a mostrar la deducción de los valores correspondientes a los diferentes parámetros, comenzando

por las probabilidades de activación de cada uno de los tres niveles Al nivel de sesión se calculan como:

$$P_{1s} = \frac{T_{pagina}}{T_{thinking} + T_{pagina}} \tag{4.12}$$

$$P_{0s} = \frac{T_{thinking}}{T_{thinking} + T_{pagina}} \tag{4.13}$$

Y para el nivel de ráfaga:

$$P_{1r} = \frac{T_{on}}{T_{idle} + T_{on}} \tag{4.14}$$

$$P_{0r} = \frac{T_{idle}}{T_{idle} + T_{on}} \tag{4.15}$$

La probabilidad de conexión se obtiene como:

$$P_{1c} = \frac{\alpha \bullet T_s}{3600} \tag{4.16}$$

A partir de estas probabilidades, se calcula el número de usuarios que están en una conexión multiplicando el valor de la población por el de P_{1c}:

$$N_c = pop \bullet P_{1c} \tag{4.17}$$

Una vez se tiene ese valor, se calcula el número de usuarios que están en una sesión simultáneamente:

$$N_s = N_c \frac{P_{1s}}{(1-P_{1s})} + \gamma \sqrt{N_c \frac{P_{1s}}{(1-P_{1s})}} \tag{4.18}$$

A continuación se calcula el número de usuarios en una ráfaga:

$$N_r = N_s \frac{P_{1r}}{(1-P_{1r})} + \gamma \sqrt{N_c \frac{P_{1r}}{(1-P_{1r})}} \tag{4.19}$$

Finalmente, conocida la velocidad de ráfaga (la velocidad del acceso), el ancho de banda agregado correspondiente al servicio *Web* se obtiene como:

$$BW_{WEB} = N_r \cdot V_{ac} \tag{4.20}$$

Si por ejemplo, se asignan los valores de:

- T_{on}=4
- T_{idle}=6
- T_{pagina}=10
- $T_{thinking}$=15
- γ=2
- Población=10000
- V=33 Kbps

Si se considera un nivel de conexión permanente (todos los usuarios permanecen conectados), el nivel de caudal medio de cada uno de los usuarios sería de 15.7 Kbps. Por el contrario, para una probabilidad de conexión de 0.1, este valor aumenta hasta obtener un caudal de 18 Kbps por usuario conectado. Este aumento es lógico debido a la peor multiplexación estadística producida al tener menos usuarios en el sistema.

4.3. Conclusiones

En este capítulo se ha llevado a cabo un ejercicio práctico de aplicación de las propuestas presentadas en esta Tesis. El modelo CASUAL permite establecer pautas concretas para la aplicación ordenada y sistemática de los mecanismos de modelado de tráfico de fuente, así como, de agregación en escenarios de red de acceso complejos. Se ha demostrado cómo, la definición y establecimiento de criterios específicos para la selección de los parámetros de entrada de la red, sigue una parametrización a lo largo de tres ejes diferenciados, según el tipo de usuario, servicio y tecnología de acceso. De esta manera, la definición de los diferentes elementos de red utiliza este sistema de referencia, a partir del cual, es posible seleccionar los métodos de modelado más adecuados según el caso. Como ejemplo se ha propuesto el diseño de una aplicación estratégica para la estimación de valores de tráfico agregado a diferentes niveles de una red de acceso especificada mediante CASUAL. Se ha presentado la estructura de esta herramienta, los diferentes elementos que la componen y las funcionalidades de que dispone.

5. Conclusiones

En esta Tesis se han presentado varias técnicas para el modelado de tráfico de fuente y sus correspondientes agregados, así como sus posibles aplicaciones. En el presente epígrafe, se resumen las conclusiones y principales aportaciones, así como, las posibles líneas de investigación que darán continuidad a los estudios y propuestas presentados.

5.1. Conclusiones y aportaciones

La estructura del proceso de planificación de redes de comunicación justifica el desarrollo e implementación de herramientas especializadas, adaptadas a la resolución del dimensionado de elementos de red. Al ser una acción de tipo estratégico, se basa en la estimación y predicción de los parámetros sobre tres ejes principales, usuarios, servicios y tecnología de accesso y sus correspondientes interrelaciones, líneas de trabajo que en esta Tesis, se han abordado desde el punto de vista de las soluciones basadas en el modelado y la observación.

En el segundo capítulo se ha desarrollado el concepto de los modelos de tráfico, como base teórica fundamental de las propuestas desarrolladas a lo largo de este trabajo. Se ha concentrado, como base fundamental, en el caso particular de los modelos de tráfico de fuente y agregado, y finalmente se ha proporcionado una breve descripción de un modelo determinístico basado en el *Network Calculus*, como posible alternativa a los modelos de tráfico

tradicionales basados en costosos modelos estocásticos. Las principales aportaciones de dicho capítulo son:

- El estudio realizado sobre los modelos de tráfico de fuente, incluido en este documento, describe el estado del arte de las principales técnicas de modelado, desde los procesos de renovación, hasta los modelos autosimilares, de fluidos y fractales.
- Se ha expuesto el problema de la agregación desde el punto de vista de los modelos de tráfico de fuente, y cómo éstos han evolucionado hasta modelos multifuente adaptados.
- El *Network Calculus* es una teoría formulada en los años 90, que ha sido retomada en este estudio para su aplicación al problema de la agregación. Las primeras aplicaciones que utilizaron sus principios, sólamente llegaron a modelar fuentes *DiffServ*.

Destacan las siguientes conclusiones:

- La planificación estratégica de redes de comunicación de banda ancha, evolución de la actual Internet, requiere nuevos modelos de tráfico para adaptar el proceso de planificación y diseño a los patrones de tráfico que generan los nuevos servicios.
- El modelado del tráfico a ráfagas que se observa en Internet requiere soluciones adaptadas, entre las que destacan, especialmente, los modelos autosimilares y ON-OFF multinivel, como alternativas más cercanas.
- El modelado del tráfico de Internet agregado plantea nuevas necesidades que los actuales modelos multifuente tampoco resuelven. Los resultados reales quedan bastante lejanos de la simple multiplexación estadística.
- El cálculo mediante técnicas de análisis del tráfico real es una opción más al margen de la idoneidad de los modelos de tráfico fuente.

El capítulo 3 concentra las principales propuestas presentadas en este trabajo, todas ellas centradas en el dimensionado de las capacidades requeridas por las fuentes, con su correspondiente tráfico y sus agregados, a partir de mecanismos y métodos diferentes:

- El modelo CASUAL: Este concepto ha sido desarrollado por completo en esta Tesis. Propone una nueva forma de interrelacionar los parámetros fundamentales del tráfico con las condiciones externas, esto es, el comportamiento del usuario, las características específicas de los servicios, y la influencia de la tecnología de acceso en todos ellos.
- El modelo ON-OFF multinivel: Un nuevo modelo de tráfico de fuente basado en estudio del comportamiento de los servicios Internet a diferentes niveles de observación: Al nivel de ráfaga, en el de sesión y el de conexión, modelados cada uno de ellos por una fuente ON-OFF.

- El cálculo de la capacidad equivalente de flujos de Internet individuales a partir de la observación de las curvas de llegadas de tráfico. Se ha propuesto una curva de servicio optimizada para servicios *best-effort*.
- El modelo ON-OFF multinivel multifuente: A partir del modelo de tráfico de fuente, se propone aplicar el mismo concepto al problema de la agregación, la cual se resuelve nivel a nivel mediante aproximaciones binomiales.
- El cálculo de curvas de tráfico agregado mediante curvas de llegada *TSpec*. En este caso solamente se ha realizado el estudio de su aplicación como alternativa al modelo anterior.

Tras desarrollar dichas aportaciones e ilustrarlas con ejemplos de aplicación, se pueden extraer las siguientes conclusiones:

- El modelo ON-OFF multinivel sobre fuentes individuales no tiene aplicación práctica de forma directa, pero sin embargo, constituye la base para ser aplicado en agregados de fuentes homogéneas, sobre los que obtiene una primera aproximación al valor de la capacidad equivalente por fuente.
- La versión multifuente del modelo ON-OFF multinivel mejora la aproximación anterior. Su principal virtud es que el modelo es adaptable a las características específicas del servicio sin más que modificar las características de función de agregación, en cualquiera de los tres niveles, aunque en la planificación estratégica normalmente sólo afecte al de conexión, mientras que los dos niveles inferiores son considerados en la planificación táctica y la operación y mantenimiento (OAM) de la red.
- La estimación a partir de curvas de llegadas permite obtener resultados de una forma muy rápida, con el inconveniente que el método parte de principio sobre la base de la sobreestimación y requiere medidas de tráfico, aspecto difícil de proporcionar dentro de la planificación estratégica. Sin embargo, los resultados, en el caso de fuentes individuales se quedan cerca de los que obtienen el resto de modelos y estudiar su aplicación en la planificación táctica y el análisis de rendimiento dentro de la fase de OAM de la red puede resultar interesante.
- La estimación del tráfico agregado mediante curvas de llegadas, tiene el inconveniente de multiplicar el efecto de sobredimensionado conforme aumenta el número de fuentes agregadas. En los ejemplos planteados con dos y nueve fuentes respectivamente, las soluciones estudiadas permiten obtener directamente la estimación de la curva suma a partir de las características básicas de las curvas individuales.

Por último, el capítulo 4 ha presentado un ejemplo de integración de los modelos propuestos mediante la aplicación del modelo CASUAL. Las principales propuestas han sido:

- La aplicación del concepto CASUAL en una herramienta como alternativa para la introducción de los parámetros de modelado de los diferentes elementos de red.
- Se ha propuesto una estructura y un modo de funcionamiento adecuado a la aplicación de los modelos de tráfico de fuente y agregado estudiados.

Como conclusiones finales cabe destacar que:

- El modelo CASUAL permite establecer pautas concretas para la aplicación ordenada y sistemática de los mecanismos de modelado de tráfico de fuente y de agregación en escenarios de red de acceso complejos.
- El uso del sistema de referencia propuesto por CASUAL en la definición de los diferentes elementos de red admite hacer una selección de los métodos de modelado más adecuados según el caso.
- La aplicación del concepto CASUAL permite facilitar el desarrollo de una herramienta estratégica para la planificación y diseño de redes de acceso a Internet.

5.2. Líneas futuras de investigación

A la vista de los resultados obtenidos en esta Tesis, a continuación se proponen algunas de las líneas de investigación que se abren tras sus conclusiones:

- **Aplicación de funciones de distribución de "cola pesada" en las fuentes del modelo ON-OFF multinivel**: La integración del efecto autosimilar, no sólo a una fuente sino a los tres niveles puede permitir modelar comportamientos del tráfico que hasta ahora ningún modelo ha considerado.

- **Extensión del modelo ON-OFF multinivel para el modelado de fuentes multiservicio**: Actualmente el modelo solamente resuelve el problema de agregación multifuente y su agregación (multifuente-multiservicio) y estudiar su aplicación e influencia en la planificación estratégica. Actualmente el modelo sólamente resuelve el problema de la agregación multifuente sobre un único servicio, lo que se corresponde con el modelado de túneles virtuales separados para cada clase de servicio. Aprovechando la definición de parámetros y siguiendo el modelo propuesto por CASUAL, la estimación del tráfico de fuente correspondiente a un usuario quedaría completada en un solo cálculo.

- **Ampliación del método de estimación a partir de la optimización de curvas de llegada**: El problema del sobredimensionado puede ser

resuelto mediante la introducción de vértices adicionales a las *TSpec* calculadas. Esto requiere redefinir el método de optimización.

- **Ampliación de la definición del modelo CASUAL para su aplicación en el resto de tareas relacionadas con la planificación estratégica**: El uso de este concepto puede estar especialmente indicado para su aplicación en herramientas implicadas en las tareas de la evaluación económica para, por ejemplo, para el desarrollo de esquemas de tarificación.

- **Desarrollo de nuevos modelos de coste para multiservicios considerando aspectos de GoS y QoS**. Los modelos actuales, sobre todo el modelo de coste incremental a largo plazo (Long Run Incremental Cost - LRIC) y su aplicación a cada elemento de la red (Total Element LRCI –TELRIC), se basan exclusivamente en el valor de ancho de banda medio y no consideran diferentes clases de servicios con sus correspondientes parámetros GoS y QoS.

- **Estudio de la influencia de la política de ingeniería de tráfico a los costes de un servicio**, aplicado en la operación de una red de banda ancha basada en el protocolo IP, y su influencia en le modelo TELRIC.

Apéndices

Apéndice A.

Abreviaturas y Acrónimos

ADSL.	Assimetric Digital Subscriber Loop
AF.	Assured Forwarding
ANP.	Aggregation Network Part
APON.	ATM Passive Optical Network
AR.	AutoRegressive
ARFIMA.	AutoRegressive Fractionally Integrated Moving Average
ARIMA.	AutoRegressive Integrated Moving Average
ARMA.	AutoRegressive Moving Average
AS.	Autonomous System
ATM.	Asynchronous Transfer Mode
BH.	Bussiness Hour
BPON.	Broadband Passive Optical Network
CAC.	Connection Admission Control
CASUAL.	Cubo de Accesos/Servicios/Usuarios de Asignación Libre
CBR.	Constant Bit Rate
CMT.	Comisión del Mercado de las Telecomunicaciones
CNP.	Core Network Part
CSD.	Circuit Switched Data
DAR.	Discrete AutoRegressive
D-MMDP.	Discrete Markov Modulated Deterministic Process
DOCSIS.	Data Over Cable Service Interface Specification
DSL.	Digital Subscriber Loop
EBB.	Exponentially Bounded Burstiness
EF.	Expedited Forwarding
EGPRS.	Enhanced Data rates for GSM of Evolution General Packet Radio Service
EPON /	GEPON. Ethernet /Gigabit-Ethernet Passive Optical Network
FARIMA.	Fractional ARIMA
FBM.	Fractional Brownian Motion
FEC.	Forwarding Equivalence Class
FIFO.	First In First Out
FLB.	Fractional Leaky Bucket
FTP.	File Transfer Protocol
GoS.	Grade of Service
GPON.	Gigabit Passive Optical Network
GPRS.	General Packet Radio Service
GPS.	Generalized Processor Sharing
GSM.	Global System for Mobile

GSR.	Gigabit Switch Router
HDSL.	High bit-rate Digital Subscriber Loop
HDTV.	High Definition TV
HFC.	Hybrid Fibber - Coax
HSCSD.	High Speed Circuit Switched Data
HSPA.	High-speed Packet Access
HTML.	HyperText Markup Language
HTTP.	HyperText Transfer Protocol
IDSL.	ISDN Digital Subscriber Loop
IEEE.	Institute of Electrical and Electronics Engineers
IETF.	Internet Engineering Task Force
INE.	Instituto Nacional de Estadística
IP.	Internet Protocol
IPP.	Interrupted Poisson Process
ISDN.	Integrated Service Digital Network
ISP.	Internet Service Provider
ITU.	International Telecomunication Union
IXP.	Internet EXchange Points
LAN.	Local Area Network
LANP.	Low Aggregation Network Part
LBAP.	Linnear Bounded Arrival Process
LMDS.	Local Multipoint Distribution Service
LRD.	Long Range Dependence, Long Range Dependence
LSP.	Label Switching Path
MA.	Moving Average
MIMO.	Multiple Input and Multiple Output
MM.	Markov Modulated
MMDS.	Multichannel Multipoint Distribution Service
MMPP.	Markov Modulated Poisson Process
MOS.	Mean Opinion Score
MPLS.	Reservation Specification, MultiProtocol Label Switching
MSDSL.	Multirate Symmetric Digital Subscriber Loop
MTU.	Maximum Transfer Unit
NFSNET.	National Foundation Science Network
NGN.	Next Generation Network
NMC.	Network Management & Control
OAM.	Cost, Operations, Administration, and Maintenance
OC.	Optical Carrier
OSI.	Open System Interconnection
P2P.	Peer To Peer
PDSL.	Personal Digital Subscriber Loop
PER.	Packet Error Rate
PGPS.	Packetized Generalized Processor Sharing
PLC.	Power-Line Communication
PLR.	Packet Loss Rate
PON.	Passive Optical Network
POP.	Point Of Presence

PSTN.	Public Switched Telephone Network
PYME.	Pequeña Y Mediana Empresa
QoS.	Quality of Service
RADSL.	Rate Adaptive Digital Subscriber Loop
RF.	Radio Frequency
RFOG.	Radio Frequency Over Glass
RSPEC.	Reservation Specification
RSVP.	Resource Reservation Protocol
RTT.	Round Trip Time
SA.	Simulated Annealling
SANP.	Subscriber Access Network Part
SBB.	Stochastically Bounded Burstiness
SBPP.	Switched Batch Bernoulli Process
SDB.	Switched Digital Broadband
SDH.	Synchronous Digital Hierarchy
SDSL.	Symmetric Digital Subscriber Loop
SHDSL.	Symmetric High bit-rate Digital Subscriber Loop
SIC.	Sensitive dependence on Initial Conditions
SLA.	Service Level Agreement
SMTP.	Simple Mail Transfer Protocol
SONET.	Synchronous Optical NETworking
STS.	Synchronous Transport Signal
TCP.	Transmission Control Protocol
TDM.	Time Division Multiplex
TDT.	Terrestrial Digital Television
TES.	Transform Expand Sample
TSP.	Travelling Salesman Problem
Tspec.	Traffic Specification
TSR.	Tera Switch Router
UDP.	User Datagram Protocol
UDSL.	Unidirectional Digital Subscriber Loop
UMTS.	Universal Mobile Telecommunication System
UANP	Upper Aggregation Network Part
VBD.	Voice Band Data
VBR.	Variable Bit Rate
VDSL.	Very High bit-rate Digital Subscriber Loop
VoIP.	Voice Over Internet Protocol
VPN.	Virtual Private Network
WDM.	Wavelength Division Multiplexing
WFQ.	Weighted Fair Queueing
WIFI.	WIreless FIdelity
WI-MAX.	Worlwide Interoperability for Microwave Access
WLAN.	Wireless Local Area Network
WMAN.	Wireless Metropolitan Area Network
XHTML.	Extensible HyperText Markup Language
XML.	Extensible Markup Language

Apéndice B.

Optimización de los valores de S y B a partir de la ancho de banda efectivo (P0)

Para obtener los valores óptimos de *S* y *B* se puede utilizar una función de coste *C(P,S,B)* fijando el valor de *P* a P_0:

$$C(P_0,S,B)-C(P_0,0,0)=uS+vB \tag{B .1}$$

Fijando $v=1$, y los límites de *S* y *B* como:

$$S_{\max}\le P_0 \Rightarrow 0\le S\le S_{\max} \tag{B .2}$$

$$B_{\max} \Rightarrow 0\le B\le B_{\max} \tag{B .3}$$

la expresión a minimizar es $uS+B$ con $B+(s+D)S-\alpha(s)\ge 0$.

Para que todo ello sea cierto se debe cumplir:

$$B_{\max}\ge \sup_{s\ge 0}\{\alpha(s)-(s+D)S_{\max}\} \tag{B .4}$$

Haciendo un cambio de variable:

$$\begin{cases} x=S \\ y=B+sD \end{cases} \tag{B .5}$$

La expresión a optimizar ahora queda como:

$$(u-D)x+y\begin{cases} 0\le x\le S_{\max} \\ 0\le y-Dx\le B_{\max} \\ y\ge -\hat{\alpha}(x) \end{cases} \tag{B .6}$$

Donde:

$$\hat{\alpha}(x)=\inf_{s\ge 0}\{xs-\alpha(s)\} \tag{B .7}$$

es la conjugada cóncava de α, convexa, y por tanto decreciente en sentido amplio.

En estas condiciones aparecen dos áreas susceptibles de ser óptimas, tal como se muestra en la Fig. 0.1:

$$R_1\begin{cases} 0\le x\le S_{\max} \\ 0\le y-Dx\le B_{\max} \end{cases} \tag{B .8}$$

$$R_2 \{ y \geq -\hat{\alpha}(x) \tag{B .9}$$

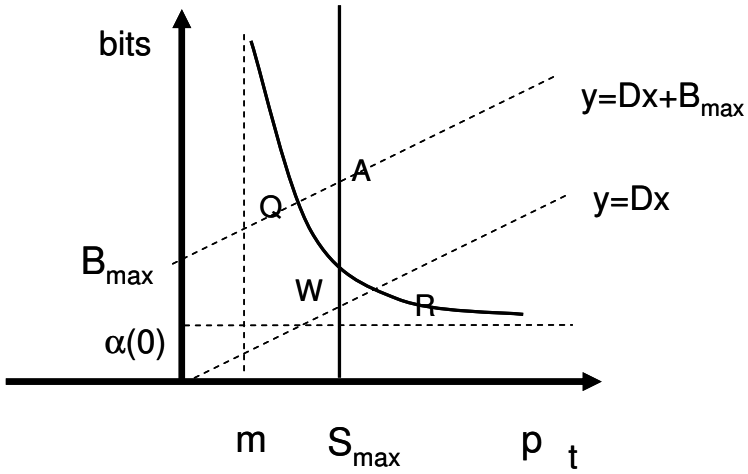


Fig. 0.1: Determinación del área de optimización de los parámetros de la curva de servicio

Si α es subaditiva y se la suponen dos valores límites:
Tasa sostenible:

$$m = \inf_{s \to \infty} \frac{\alpha(s)}{s} \tag{B .10}$$

Tasa máxima:

$$p = \sup_{s \to \infty} \frac{\alpha(s)}{s} \tag{B .11}$$

El área total queda delimitada por los límites de la conjugada cóncava:

$$\hat{\alpha} : \begin{cases} si & x \leq m \Rightarrow -\hat{\alpha}(x) = \infty \\ si & x \geq p \Rightarrow -\hat{\alpha}(x) = \alpha(0) \end{cases} \tag{B .12}$$

Se buscan los puntos de corte entre:

$$\begin{cases} y = Dx + B_{\max} \\ y = Dx \\ x = S_{\max} \\ \hat{\alpha}(x) \approx R_2 \end{cases} \Rightarrow \begin{cases} Q(x_Q, y_Q) \\ R(x_R, y_R) \\ A(x_A, y_A) \\ W(x_W, y_W) \end{cases} \tag{B .13}$$

Para Q:

$$\begin{cases} x_Q = \sup_{s\geq 0} \dfrac{\alpha(s) - B_{\max}}{s + D} \\ y_Q = Dx_Q + B_{\max} \end{cases} \tag{B .14}$$

Para R:

$$\begin{cases} x_R = \sup_{s\geq 0} \dfrac{\alpha(s)}{s + D} = P_0 \qquad x_Q \leq S_{\max} \leq x_R \\ y_R = Dx_R \end{cases} \tag{B .15}$$

Para W:

$$\begin{cases} x_W = S_{\max} \\ y_W = -\hat{\alpha}(S_{\max}) \leq DS_{\max} + B_{\max} = y_A \end{cases} \tag{B .16}$$

Y para A:

$$\begin{cases} x_A = S_{\max} \\ y_A = DS_{\max} + B_{\max} \end{cases} \tag{B .17}$$

En función de un punto genérico x_0, se identifican tres casos:

$$x_Q \leq x_0 \leq S_{\max} : \text{El mínimo es } x = x_0 \tag{B .18}$$

$$x_0 > S_{\max} : \text{El mínimo es } x = S_{\max} \tag{B .19}$$

$$x_0 < x_Q : \text{El mínimo es } x = x_Q \tag{B .20}$$

Apéndice C.

Modelado de tráfico de VoIP

El servicio de voz sobre IP ha tomado en los últimos tiempos un mayor auge. En primer lugar para integrar en una red IP los servicios tradicionales de la RTC/RDSI y, en segundo lugar, para hacer un uso más eficiente del ancho de banda existente, debido a la posibilidad de una multiplexación estadística gracias, sobre todo, a la compresión de silencios que se hace en VoIP. De esta forma y de cara al usuario, independizar las llamadas de la distancia ha reducido los costes y las tarifs correspondientes a las llamadas a larga distancia, a nivel internacional (Varios operadores ofrecen tarifas planas para la voz al nivel nacional dentro de los paquetes *triple-play*).

Este es un tipo de servicio en tiempo real, en el que múltiples parámetros influyen, aunque no tanto como lo hace el retardo. En las comunicaciones de voz sobre IP se identifican tres tipos de retardos diferentes:

- El retardo intrínseco a la conmutación de paquetes debido a la necesidad de codificar, comprimir y empaquetar la voz y al retardo de conmutación en la red. Además tiene la desventaja de ser un retardo variable generalmente denominado como *jitter*. Este es un parámetro critico cara a la calidad de servicio y que viene agravado por los dos retardos restantes.
- El eco que se puede producir debido a la reflexión en el terminal remoto, y que obliga a implementar canceladores de eco.
- El retardo puro que implica que los dos hablantes esperen o hablen simultáneamente. Este parámetro empieza a producirse cuando el retardo es superior a 250 milisegundos.

El aprovechamiento del ancho de banda en este servicio se basa en parte a la compresión tanto de la voz como de los silencios. Actualmente se considera que el 56% de la llamada está formada por silencios. El 50% correspondiente a que se escucha al interlocutor y un 6% correspondiente al intervalo entre palabras o sílabas.

Definición del servicio desde el punto de vista de CASUAL

Aunque existen varias implementaciones, se suele considerar el modelo de capas genérico siguiente, muy cercano al estándar de video sobre IP(H.323) y que aparece en la Fig. 0.2.

Codificador de audio G.729
RTP/UDP
IP
PPP, ethernet, ATM (AAL-1)
ADSL, RDSI...

Fig. 0.2: Pila de protocolos del servicio de VoIP

La razón de por qué usar el codificador de voz G.729 es que hoy en día, la voz sobre IP esta ligada al estándar H.323, que busca una solución de compromiso entre la calidad de sonido y el poco ancho de banda disponible, utilizando una compresión 1/8. Este estándar define varios codificadores posibles a usar como son el G.711, el G.729 o el G.723.1 y se ha optado por utilizar en CASUAL el G.729 como ejemplo. H.225 es utilizado en la señalización la llamada y H.245 para el control de la señalización (como fija H.323).

Modelado del servicio de VoIP

Con el codificador de audio mencionado anteriormente solo se transmiten datos a la red cuando se detecta actividad de voz. Durante esa actividad un número fijo de tramas de audio se introduce en un solo paquete RTP (*Real Time Protocol*), consiguiendo paquetes de longitud determinística.

En cuanto a los *overheads* que son introducidos, al utilizar el codificador de voz G.729, con una codificación de audio de 8 Kbps, resultan dentro del compromiso entre una buena calidad de sonido y una tasa binaria aceptable (tasa de compresión 1/8). En la Fig. 0.3 se muestra la estructura de los paquetes RTP generados, teniendo en cuenta que dos tramas G.729 (que transportan 20 ms de audio) se empaquetan dentro de un solo paquete RTP.

PPP,*ethernet*	IP	UDP	RTP	G.729
18 bytes	20 bytes	8 bytes	12 bytes	20 bytes

Fig. 0.3: Formato completo de un frame de una aplicación de vídeo

Como se puede observar, para enviar 20 bytes de datos, estos se encapsulan en un total de 78 bytes, esto es, un overhead de 58 bytes, es decir, más del 74% de la comunicación.

A pesar del derroche que pueda representar este valor, el resultado no es tan caótico si se compara con los valores que se tendrían transmitiendo voz sin comprimir, ya en 20 ms se tendría que transmitir 160 bytes de voz PCM sin

comprimir. Pese a todo con G.729 se ahorran 140 bytes de envío en cada paquete IP.

El modelado del servicio de voz puede afrontarse teniendo en cuenta toda la problemática del *buffer* necesario para asegurar la calidad del tráfico IP generado, con el objetivo de minimizar el impacto de las problemáticas de restricción temporal, de retardo y de *jitter*. En principio, el servicio de voz sobre IP se basa en estándares que permiten conocer la naturaleza de las ráfagas generadas, permitiendo conocer con detalle cómo se comportan las fuentes, siendo de esta manera mucho más fácil de analizar que el tráfico *Web*.

Así, el nivel de sesión puede ser dimensionado de la misma manera que se modelan las llamadas telefónicas sobre RTC, mediante un modelo de Erlang-B con proceso de entrada de Poisson. De esta forma, dada una población, es posible calcula la probabilidad de que se produzcan un numero determinado de llamadas simultáneas. Esto justifica el uso de binomiales clásicas para el cálculo de la mezcla de tráfico de diferentes usuarios.

En cuanto al nivel de ráfaga, este tráfico presenta claras características de tráfico a ráfagas (bursty) y por lo tanto los modelos de poisson no pueden analizarlo completamente debido a que su correlación cae de forma muy lenta para periodos largos de tiempo y por lo tanto dependencia a largo plazo. Por ello el modelado de este tipo de tráfico recurre a los procesos de Markov modulados (MMP), comentados en el capítulo 3.

Parámetros asociados al tráfico VoIP

Los parámetros, de acuerdo con la filosofía del modelo ON-OFF multinivel, para cada uno de los tres niveles temporales son:

Nivel de Conexión:

- Número de conexiones por hora (α): número de llamadas por hora y en un día normal que realiza un usuario del tipo contemplado.
- Duración de una conexión ($t_{conexión}$): Normalmente coincide con la duración de cada llamada, como en el caso de telefonía.
- Velocidad del acceso (Vac): ancho de banda de la línea d acceso a Internet.
- En el caso de los usuarios de negocios, personal asignable al tipo de usuario seleccionado ($N_{personal}$).y que por ejemplo, una empresa de 1000 empleados no genera el mismo tráfico que una de 20.

Nivel de Sesión:

- Tiempo entre sesiones ($T_{silencio}$): tiempo que pasa entre dos llamadas consecutivas pedidas por el usuario.

- Duración de la sesión ($T_{llamada}$): tiempo que dura la conversación solicitada.

Nivel de Ráfaga:

- Tiempo entre ráfagas (T_{idle}): tiempo entre dos ráfagas de voz consecutivas.
- Duración de la ráfaga (T_{on}): tiempo de transmisión de la ráfaga de voz.

Estimación del tráfico generado por fuentes de VoIP

La cadena de estados utilizada, como ya se vió anteriormente, será una fuente ON-OFF por cada nivel temporal considerado, realizando la mezcla de tráfico mediante una binomial negativa. La fuente ON-OFF presenta estados de actividad junto a estados de silencio cuyos tiempos de duración están distribuidos exponencialmente para el nivel de sesión y de manera constante para el nivel de ráfaga. Durante el tiempo de ráfaga se generan paquetes con longitud fija a intervalos de tiempo constantes Δt debido al sistema de codificación de voz utilizado. Experimentalmente se ha demostrado que valores de tiempos de actividad de 0.4 segundos y silencio de 0.6 segundos son típicos.

Las probabilidades de activación o silencio al nivel de ráfaga son entonces:

$$P_{1r} = \frac{T_{on}}{T_{idle} + T_{on}} \qquad P_{0r} = \frac{T_{idle}}{T_{idle} + T_{on}} \tag{5.21}$$

Mientras que para el nivel de sesión serán:

$$P_{1s} = \frac{T_{llamada}}{T_{silencio} + T_{llamada}} \qquad P_{0s} = \frac{T_{silencio}}{T_{silencio} + T_{llamada}} \tag{5.22}$$

La probabilidad de conexión se obtiene, como en el caso *Web*:

$$P_{1c} = \frac{\alpha \bullet T_s}{3600} \tag{5.23}$$

A partir de estas probabilidades se calcula el número de usuarios que están en una conexión multiplicando el valor de la población por el de P_{1c}, para posteriormente calcular el número de usuarios que están en una sesión simultáneamente:

$$N_s = N_c P_{1s} + \gamma \sqrt{N_c \frac{P_{1s}}{(1 - P_{1s})}} \tag{5.24}$$

Y el número de usuarios con ráfagas activas será:

$$N_r = N_s \frac{P_{1r}}{(1 - P_{1r})} + \gamma \sqrt{N_c \frac{P_{1r}}{(1 - P_{1r})}} \tag{5.25}$$

Una vez se dispone de este valor solo hay que multiplicarlo por la velocidad de la ráfaga (Velocidad del acceso a Internet) para obtener el ancho de banda asociado al servicio de VoIP.

Así, a modo de ejemplo, se va comparar cómo afecta el cambio de la distribución binomial respecto al caso del servicio *Web*, para lo cual se utilizan los mismos valores e los diferentes parámetros. El valor final resultante no es real, puesto que la velocidad de acceso mínima para ese servicio es de 64 Kbps.

- T_{on}=4
- T_{idle}=6
- $T_{llamada}$=10
- $T_{silencio}$=15
- γ=2
- Población=10000
- V=33 Kbps (no sostenible porque ya, solamente para este servicio, el codificador envía una trama de 78 bytes cada 20 ms que nos da un régimen de 31.2 Kbps para cada interlocutor).

Para usuarios permanentemente conectados el ancho de banda medio para cada usuario sería de 9.08 Kbps sensiblemente inferior al calculado para el servicio, puesto que cuando no se considera el efecto de las ráfagas en alguno de los niveles temporales suaviza el tráfico apreciablemente. Si se considera una probabilidad de conexión de 0.1, este valor se reduce a 0.918 Kbps, que puede ser corregido mediante el factor de ponderación para la hora cargada (*A*).

En la realidad, si se consideran las probabilidades del nivel de ráfaga con valores de 0.4 y 0.6 respectivamente (valores reales para este servicio), una duración de llamada de 180 segundos y un tiempo entre llamadas de 1800 segundos, el ancho de banda obtenido por cada conexión activa es de 2.4 Kbps por usuario, para conexiones permanentes. Si se introduce el nivel de conexión con probabilidad de 0.1, el valor obtenido caería hasta 3.25 Kbps. Para obtener el valor de tráfico basta con multiplicar este ancho de banda por la población y por el factor de ponderación deseado en el nivel de conexión. Finalmente, sólo queda normalizarlo al ancho de banda de la línea deseada (en principio un valor típico suele ser 1 Mbps).

Apéndice D.

Servicios de transmisión de Vídeo

El servicio de video a través de la red IP es un servicio relativamente nuevo y novedoso ya que incluye la transmisión de sonido e imagen en un tiempo real o semirreal, con una transferencia de gran cantidad de información en un tiempo corto. A modo de ejemplo, una secuencia de vídeo de 25 segundos transmitida haciendo uso del programa y codecs de Quick-time a una resolución baja de 320x240 *pixels* generaría cerca de 2.3 MB, lo cual implica un ancho de banda medio aproximado de 96 Kbps, que no se podía realizar mediante el *modem* tradicional.

Por otro lado este tipo de servicios presentan unas características singulares ya que generan tráfico, normalmente de valor variable, que debe ser transmitido inmediatamente, con restricciones importantes en los retardos máximos y unos requerimientos de QoS altos y directamente percibidos por el usuario. Así por ejemplo, mientras que la recepción de audio en el oído humano permite retardos de hasta 250 ms, en vídeo este parámetro es inferior, no por la tolerancia del usuario final, sino por la variabilidad de la cantidad de información a transmitir en cada instante.

Aunque existen varias implementaciones, el estándar de facto H.323 [Turaga, 2000c] es uno de los más utilizados, permitiendo la transmisión de audio y video en tiempo real sobre redes de paquetes, especialmente IP, especificando los protocolos y procedimientos multimedia.

Para la transmisión de vídeo sobre la red Internet, H323 hace uso de la capa de transporte en modo datagrama (UDP). El uso de UDP es común a todas las aplicaciones en tiempo real, incluida la VoIP vista en el apartado anterior. Esto es debido a que las retransmisiones que introduce el control de flujo del protocolo TCP introducen una redundancia que en la mayoría de los casos resultan inútiles al superar los límites del retardo máximo. Además se da la circunstancia que la codificación de vídeo introduce por si la suficiente redundancia como para tolerar pérdidas dentro de límites aceptables en función de la calidad percibida y tolerada por el usuario.

Otra razón muy importante es que la información correspondiente al vídeo codificado es empaquetada por el protocolo de sesión RTP, el cual aporta servicios y facilidades adicionales, entre las que destacan la reconstrucción temporal, la detección de paquetes perdidos, el control de seguridad e identificación de contenido, y el soporte para comunicaciones multicast, imprescindible por ejemplo para la provisión de servicios de distribución de

contenidos. Precisamente, gran parte de estos servicios incluyen funciones típicas de la capa TCP, por lo que se duplicarían.

Una característica importante de RTP es que separa audio y vídeo en dos conexiones separadas, haciendo uso de un plano de control implementado por los protocolos RTCP (*Real Time Control Protocol*) y RTSP (*Real Time Streaming Protocol*). La separación física de ambas fuentes se debe a que utilizan estándares de codificación diferentes, como por ejemplo G.711, G.721, G.722 ó G.728 en el aso del audio, y CellB, JPEG, H261, MPEG I-II en el caso del vídeo.

Definición de Fuentes de Vídeo

A partir de ahora se van a considerar exclusivamente las fuentes de vídeo como fuentes de tráfico cuya característica fundamental es una tasa de bits variable, es decir, son fuentes de tráfico VBR (*Variable Bit Rate*). Esto no es estrictamente cierto, puesto que determinados servicios pueden generar flujos CBR (*Constant Bit Rate*) mediante mecanismos de adaptación en tiempo real o debido a la naturaleza del vídeo / audio transmitido. Sin embargo, estos casos pueden ser considerados como casos particulares de las fuentes VBR.

Aunque la caracterización de una fuente de tráfico VBR es complicada, en el caso de las fuentes de vídeo dicha caracterización viene prácticamente supeditada al tipo de esquemas de codificación y compresión utilizados, y que en el caso de Internet suele reducirse a los estándares correspondientes tanto de la ISO como de la ITU, así como del uso específico asociado al servicio de comunicación para el que está siendo utilizado.

Estándares de codificación / compresión de vídeo

Actualmente existen dos líneas de desarrollo de estándares de codificación de vídeo, los estándares MPEG dependientes del ISO/IEC, y las recomendaciones H.xxx del ITU-T para la definición e implementación de sistemas audiovisuales y multimedia.

El MPEG Group (Motion Pictures Expert Group) es la comisión encargada del desarrollo de estándares para la codificación y compresión tanto de vídeo como de audio digital. Desde su creación en 1988 tres son los estándares fundamentales que han visto la luz, véase [MPEG_Group, 2007]:

- ISO/IEC 11172: Aparece en 1991 y recibe el nombre de MPEG-1. Define el estándar para la codificación eficiente de vídeo para su utilización sobre soportes de almacenamiento específicos (por ejemplo CD-ROM), siendo la tasa de transferencia asociada de entre 1.2 y 1.8 Mbps.

- ISO/IEC 13818: En 1994 aparece el estándar MPEG-2, diseñado específicamente para la codificación de vídeo de alta calidad, como es el caso de la televisión digital (SDTV – Estándar Definition Television y HDTV – High Definition Television). La tasa de transferencia en este caso varía desde 4 – 8 Mbps para calidades equivalentes a NTSC, PAL y SECAM hasta los 20 – 80 Mbps para HDTV.
- ISO/IEC 14496: De 1999 y denominado MPEG-4. Define un método de codificación basada en objetos, permitiendo su escalabilidad y una elevada interactividad, gracias a la integración de contenidos de diferente naturaleza sobre un mismo soporte. Su elasticidad frente a errores hace que el rango de la tasa de transferencia resultante oscile desde los 5 Kbps hasta los 4 Mbps.

Aunque existen otros estándares MPEG posteriores, estos no definen nuevos esquemas de codificación distintos a los ya comentados, sino que pretenden establecer criterios y recomendaciones para el diseño de contenidos adecuados a los mecanismos de codificación existentes. Así por ejemplo, cabe destacar MPEG-7 y MPEG-21. El primero denominado "Interfaz para la descripción de contenidos multimedia" data de 2001 y provee mecanismos y herramientas para la creación y definición de elementos y contenidos multimedia. Por su parte, "MPEG-21 Multimedia Framework" aparece en 2002, y define el conjunto de necesidades asociadas a cualquier reproductor de contenidos multimedia para que posibilite el acceso e intercambio de contenidos digitales de forma eficiente y totalmente transparente.

De forma paralela el ITU-T ha venido definiendo estándares similares, vease [ITU-T, 1988b / Turaga, 2000b]:

- H.261: Define un esquema de codificación de vídeo para la transmisión de servicios audiovisuales a velocidades desde los 8 Kbps, y especialmente múltiplos de 64 Kbps (hasta 2Mbps), por lo que está especialmente indicado para su uso en RDSI. Recibe la denominación ITU-T3615 y fue aprobada en 1993.
- H.263: En este caso, partiendo de la recomendación H.262, se provee de funcionalidades adicionales, especialmente mejora la calidad y compensación del movimiento. Su referencia es ITU-T3515LBC, siendo aprobada en 1998.
- H.26L: Ha sido aprobada en 2003, y es el estándar para la codificación de vídeo avanzada para servicios audiovisuales genéricos. Su denominación es H.264, y entre sus mejoras destaca su robustez para la provisión de transmisiones de vídeo de bajo retardo, tanto en soportes telefónicos tradicionales como en los móviles de tercera generación (UMTS).

El desarrollo de ambas líneas de estandarización ha provocado que en ocasiones hayan convergido, como por ejemplo en el caso de la recomendación H.262, la cual incluso utiliza como referencia la denominación IS 13818

(MPEG-2). En la Fig. .1 se observa el desarrollo histórico de este proceso. Ambas líneas presentan múltiples puntos de convergencia, de forma que algunos estándares han sido basados a su vez en sus homólogos:

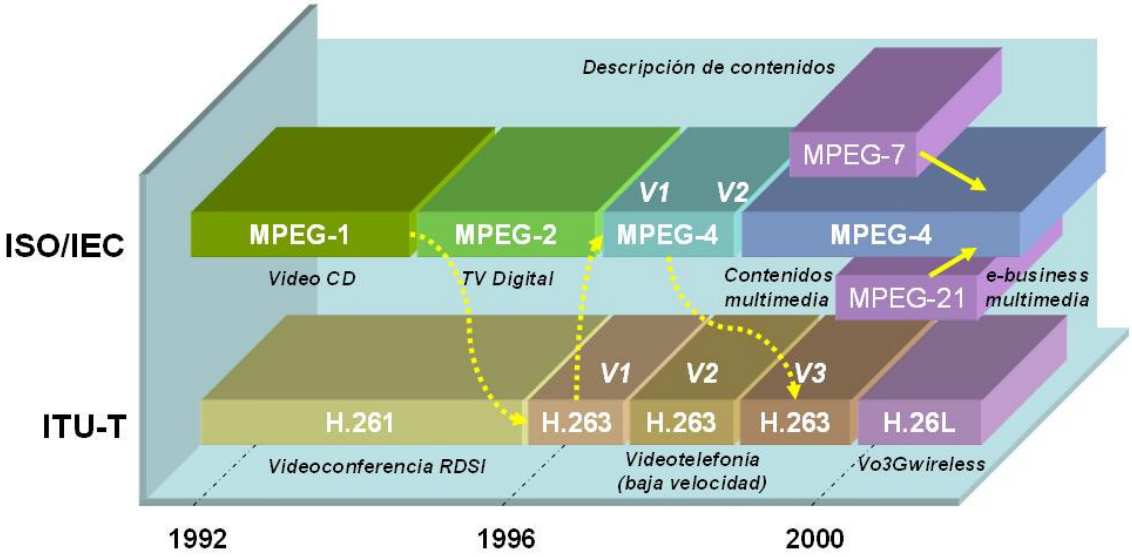


Fig. .1: Evolución temporal de los estándares de codificación de vídeo [Morisse, 2001]

Este paralelismo se hace todavía más patente si se observan los principales entornos de aplicación de dichos estándares, como se muestran en la Tabla D .I:

Aplicación	Ancho de banda	Estándar
Televisión Digital	2 – 6 Mbps	MPEG-2
HDTV	(10 – 20 Mbps)	
Vídeo DVD	6 – 8 Mbps	MPEG-2
Vídeo Internet Streaming	20 – 200 Kbps	Sistemas Propietarios, H.263, MPEG-4, H.26L
Videoconferencia Videotelefonía	20 – 320 Kbps	H.261, H.263
Vídeo sobre Wireless 3G	20 – 100 Kbps	H.263, MPEG-4, H.26L

Tabla D .I: Entornos de aplicación de los estándares de codificación de vídeo [Girod, 2001]

Fundamentos de la compresión de vídeo

El vídeo analógico digitalizado (Recomendación UIT-T BT-601) contempla velocidades de transmisión de hasta 270 Mbps. Esta elevada tasa explica la necesidad de utilizar los estándares de compresión indicados anteriormente, como se expone en [Mavrogeanes, 2001].

La compresión de vídeo parte de dos características fundamentales de la señal original: su resolución y su tasa de "frames", es decir, el número de fotogramas por segundo que son representados:

- La resolución de vídeo hace referencia al número de elementos de imagen, los que se conocen como *pixels*, que componen un fotograma, lo que repercute tanto en la calidad del mismo, como en la velocidad de transmisión requerida. Hay que tener en cuenta que la resolución en las señales televisivas se encuentra en el rango de 352x240 a 720x480, PAL y NTSC respectivamente, frente a los 1024x768 de un ordenador.
- La tasa de frames tiene valores también dependientes del sistema, te tal manera que, por ejemplo, las secuencias cinematográficas solamente muestran 24 fotogramas por segundo, frente a los 25 de PAL y los 30 de NTSC (29.97)

El proceso de compresión comienza con una primera reducción en la resolución de vídeo, hasta alcanzar un máximo de 352x240 pixels, es decir prácticamente una reducción de ancho de banda requerido de 4:1 frente a la señal NTSC original. A partir de aquí los fotogramas son transformados, de forma que se permita reducir toda aquella información redundante entre fotogramas consecutivos, esto es en pocas palabras, se procesa únicamente la información relativa al movimiento mediante la división del fotograma en bloques (a su vez estos bloques sufren una transformación basada en algoritmos DCT – *Discrete Cosine Transform* facilitando secuencias fácilmente comprimibles)

Puesto que la información procesada no es ni mucho menos la totalidad del fotograma, todos los estándares definen un modelo de codificación basado en la superposición de capas, de forma que en función del número y composición de las mismas se asegura la total escalabilidad tanto temporal (número de fotogramas) como espacial (resolución), como se muestra en la Fig. D.2:

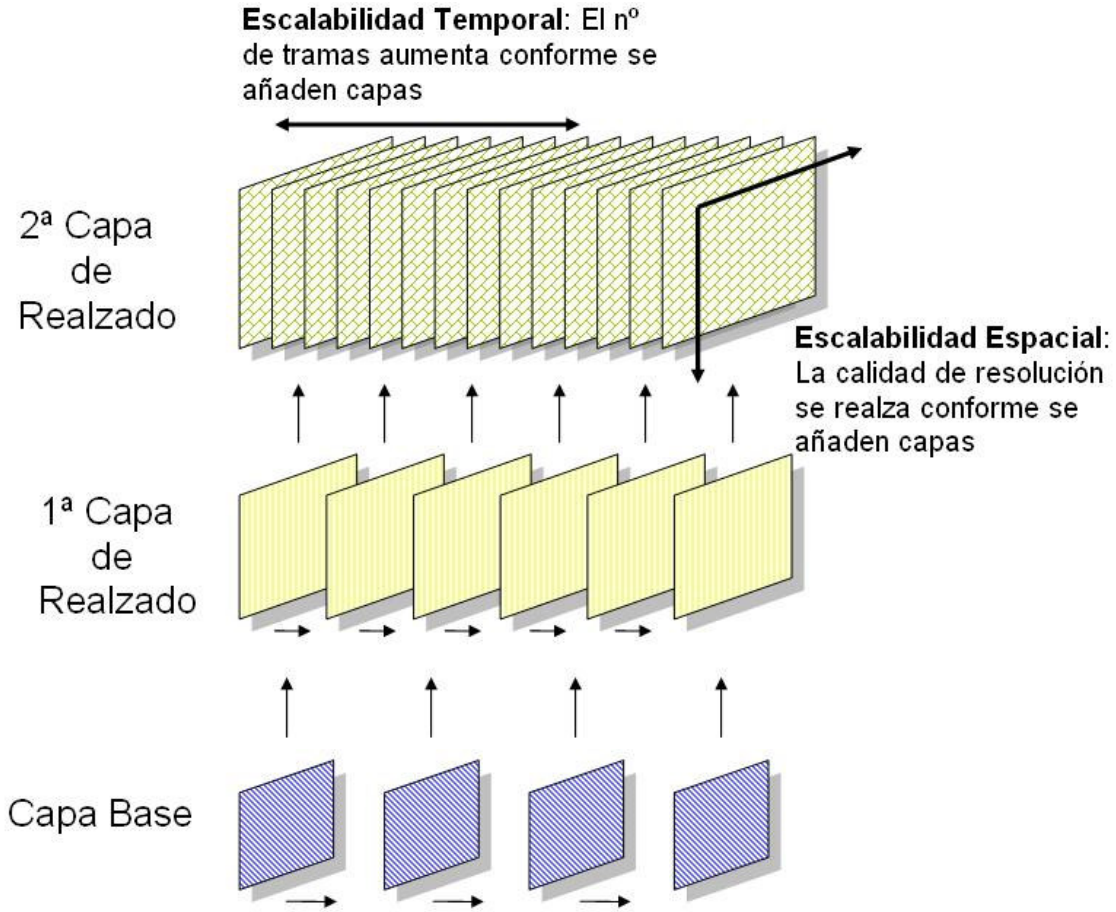


Fig. D .2: Modelo de capas utilizado en codificación de vídeo[Girod, 2001].

Aunque el número de capas totales puede ser variable, prácticamente todos los esquemas de compresión contemplan hasta tres: una básica y dos de mejora o realzado. De acuerdo con esta idea, tanto el estándar MPEG como H.263 definen varios tipos de fotograma [Tanenbaum, 2003], en función de la capa a la que corresponden:

- Fotogramas I (*intra-coded*): Constituyen la primera de las capas, denominada capa base, y no son más que imágenes estáticas de tipo JPEG a la máxima resolución definida por la secuencia de vídeo. Estas imágenes son totalmente independientes unas de otras y aparecen periódicamente para así asegurar una mínima recuperación frente a errores en la transmisión, así como funcionalidades avanzadas, como

por ejemplo el avance hacia delante o hacia atrás, o el comienzo en puntos concretos o aleatorios de la secuencia de vídeo. Típicamente el número de fotogramas I es de 1 ó 2 por segundo.

- Fotogramas P (*Predictive*): Constituyen la primera de las capas de "realzado". Codifican las diferencias respecto del fotograma anterior, bloque a bloque[37]. Para ello cada bloque es buscado en el fotograma anterior, tras lo que pueden suceder dos cosas: si el bloque es encontrado se codifica tomando las diferencias respecto a su situación anterior, es transformado mediante el DCT y finalmente es codificado como JPEG; si el bloque no es encontrado directamente es codificado como JPEG. Aunque este es el procedimiento básico, depende casi exclusivamente de las implementaciones y su aplicación, ya que el proceso puede resultar costoso y poco recomendable por ejemplo en el caso de transmisiones interactivas en tiempo real (como por ejemplo una videoconferencia)
- Fotogramas B (*Bidirectionally predictive*): Constituyen la tercera de la capas, segunda de las de "realzado". Son muy similares a los P, solo que ahora cada bloque se compara tanto con el fotograma anterior como con el posterior, por lo que supone un elevado coste de procesamiento y por ello algunas implementaciones no los soportan (p.ejm. MPEG-1). El efecto visual se traduce en una compensación del movimiento, especialmente cuando objetos de la escena cambian de plano.
- Fotogramas DC (*DC-coded*): Solamente se usan en aplicaciones con opciones de visionado rápido, siendo imágenes de baja resolución, representación de los valores medios de cada bloque. Suelen situarse justo a continuación de los fotogramas I, permitiendo así, por ejemplo, la búsqueda de escenas.

Tal como se muestra en la Fig. D.3, la superposición de todas las capas compone el "stream" de vídeo final, constituido por secuencias consecutivas de fotogramas I, P y B, que pueden ser agrupadas en lo que se conoce como GOP (*Group of Pictures*). Un fotograma I siempre inicia el GOP, marcando el límite del GOP anterior. Tras él pueden aparecer uno o varios fotogramas P, entre los cuales pueden insertarse varios fotogramas B. Estos últimos pueden ser eliminados en función de la escala temporal definida en la transmisión.

[37] El fotograma es dividido en bloques de 8×8 pixels (10×10 en MPEG-2), sobre cada uno de los cuales se aplica el algoritmo DCT.

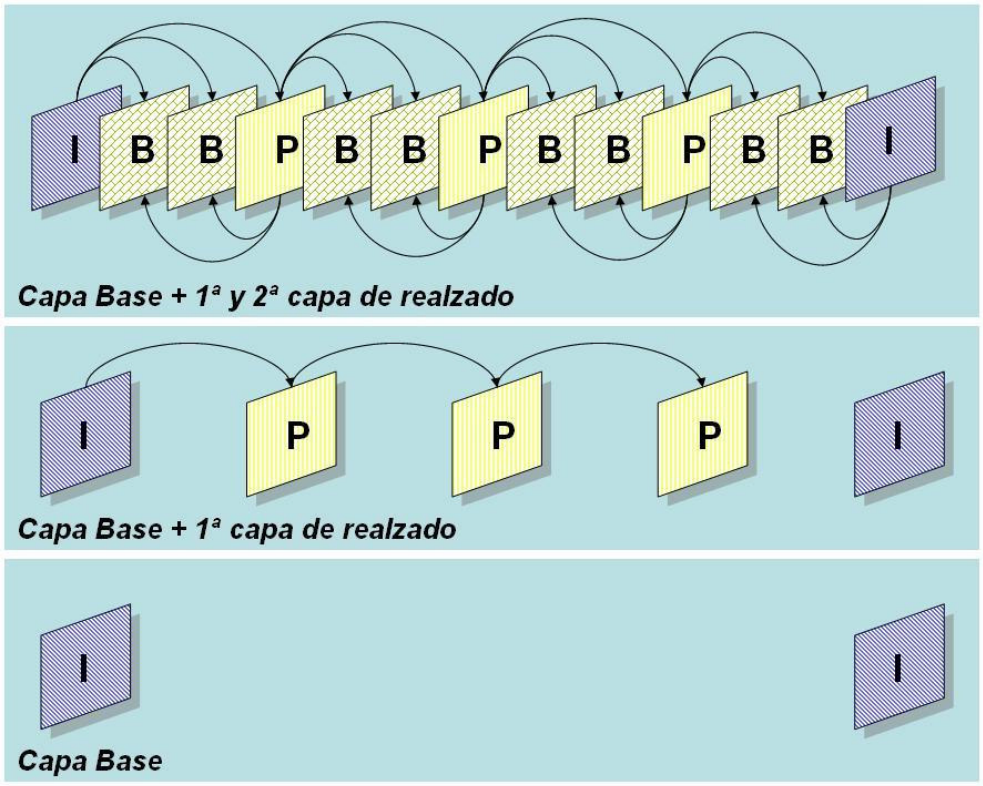


Fig. D .3: Estructura del GOP: Superposición de capas e interdependencias entre fotogramas [Girod, 2001]

Pese a seguir pautas concretas para la generación del GOP, el tamaño y composición de los fotogramas del mismo es muy variable, dependiendo, por ejemplo, de la naturaleza de los contenidos y de la calidad subjetiva deseada. De esta forma, un GOP de baja longitud (Ej. I-P) resulta mucho más robusto frente a un error en la transmisión, de forma que la pérdida de fotogramas en el GOP se traduce en pequeños saltos en la secuencia. Por el contrario, un GOP largo supone cortes más ostensibles ante errores de fotogramas, pero sin embargo aporta una mayor compresión y por tanto, para una tasa de transferencia igual, mejora enormemente la calidad de la imagen.

Modelado de fuentes de vídeo

Como es de suponer, prácticamente todos los modelos de tráfico VBR se encuentran basados en mayor o menor medida en el modelo de generación de la secuencia de fotogramas (GOP). Hay que tener en cuenta que si desde el punto de vista de la aplicación, la transmisión de vídeo no es sino una sucesión de secuencias de vídeo, desde el punto de vista de la fuente una secuencia es una sucesión de GOPs, por lo que este suele ser considerado como unidad de datos básica en la fuente de vídeo, como puede verse en la Fig. D.4. A partir de aquí, y desde el punto de vista del servicio, los modelos consideran el GOP como una sucesión de fotogramas (I, P y B), cada uno de los cuales es

empaquetado y transmitido de forma totalmente dependiente de la red de distribución, mediante células en el caso de redes ATM ó paquetes en redes IP ó Internet.

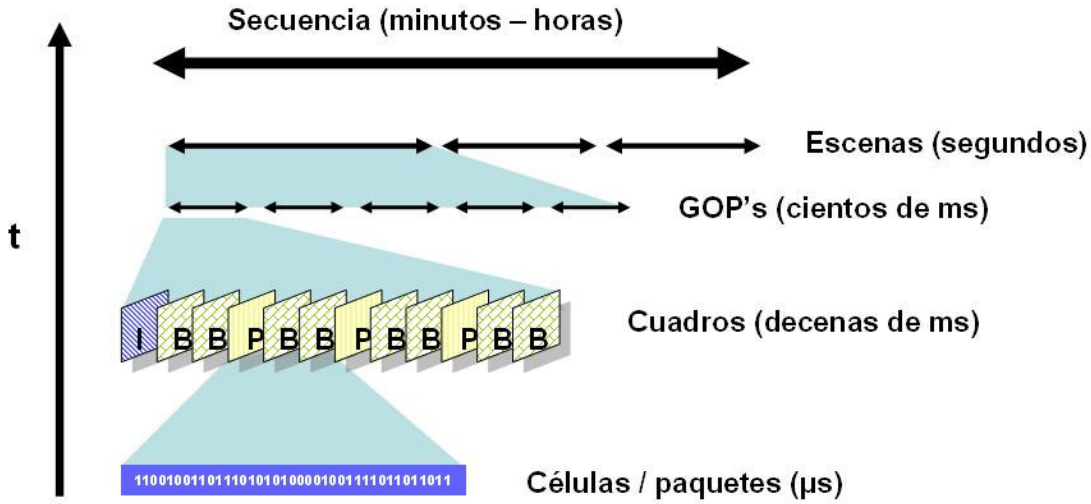


Fig. D .4: Modelo de capas para la transmisión de señales de vídeo MPEG.

A partir de esta idea, en la literatura se encuentran multitud de implementaciones y estudios de modelos para fuentes VBR, aunque pueden ser agrupadas en torno a tres de las líneas indicadas en el capítulo 3, los modelos de Markov, autorregresivos y autosimilares o fractales, así como sus correspondientes combinaciones, véase [Izquierdo, 1999].

De forma genérica, en la mayoría de las implementaciones se utilizan de forma puntual aproximaciones similares en la estimación de ciertos parámetros, como por ejemplo una distribución gamma para el modelado de los tamaños de cuadro y de GOP (aunque no obtiene una figura de correlación entre los fotogramas y el patrón del GOP, este parámetro no influye en la pérdida de paquetes).

Modelos de Markov

Estos modelos hacen uso intensivo de las teorías markovianas. Aunque tradicionalmente Markov ha sido utilizado como patrón fundamental para el modelado del proceso de generación de fotogramas, sus limitaciones a la hora de obtener una figura de autocorrelación similar a la del caso real, ha llevado a la aparición de multitud de implementaciones, tantas como combinaciones del número de estados de la cadena asociada (la autocorrelación mejora conforme aumenta la longitud de la cadena). A continuación se presentan brevemente algunas de las variantes más utilizadas [Rose, 1997]. En [Lombardo, 1998 /

Mata, 1994] pueden encontrarse otros modelos similares que toman cadenas de Markov como parte fundamental de su implementación.

Modelo del Histograma

Este modelo es el más simple ya que se basa en una cadena de Markov de orden 0. Para ello, a partir de GOP's reales se obtienen tantos histogramas como número de estados de la cadena de Markov. De acuerdo con esto, la probabilidad de entrar al estado *i (Pr$_i$)* es igual a n_i/N, siendo n_i el número de tamaños del GOP en el intervalo *i*. El proceso modelado se genera obteniendo muestras a partir de un vector de probabilidades {*Pr$_i$*}. Cabe destacar que el proceso resultante no presenta características exclusivas de los procesos markovianos, ya que cada histograma se obtiene de forma totalmente independiente de cualquier estado anterior, véase [Mashat, 1997]. No es un método muy utilizado debido a que por sus características no permite obtener ninguna información asociada con la correlación.

Modelo simple de Markov

En este caso la cadena de Markov es de orden 1. El número de estados M viene determinado por la relación G_{max}/σ_G ,siendo G_{max} el tamaño del GOP más largo y σ_G la desviación estándar del tamaño del GOP, véase [Rose, 1997] Las entradas de la matriz de probabilidades de transición {*Pr$_{ij}$*}. Pueden

$$\mathrm{Pr}_{ij} = \frac{n_{ij}}{\sum_{k=1}^{M} n_{ik}}$$

estimarse fácilmente mediante:
Donde n_{ij} representa el número de transiciones en el intervalo (i, j).

Frente al modelo del histograma mejora en tanto que permite obtener la correlación inmediata, es decir entre GOP's consecutivos. Sin embargo la estimación de la autocorrelación continúa siendo muy pobre.

Modelo orientado a la escena

Este modelo controla el proceso de cambio de escena mediante una cadena de Markov principal, y a continuación hace uso de una o varias cadenas de Markov adicionales, del tipo del modelo simple [Rose, 1997 / Rose, 1994]. Mediante estas cadenas se generan las secuencias de GOP's correspondientes a cada clase de escena, tenida cuenta que el número de clases se corresponde con el número de cambios de escena. La ventaja de este modelo es que no necesita determinar empíricamente los niveles de escena, sino solamente analizar el tamaño del GOP para cada clase de escena.

El funcionamiento de este modelo puede resumirse como sigue: La cadena principal determina el nuevo estado de escena. Si el estado permanece constante, se utiliza la misma cadena empleada por el estado anterior de entre el conjunto de cadenas adicionales para generar el GOP. Para ello se calcula el tamaño del nuevo GOP haciendo uso del tamaño del GOP anterior. En el caso de que el nuevo estado cambiara respecto al anterior, se determina cuál de las cadenas adicionales resulta adecuada para generar el nuevo GOP, de acuerdo a la clase de escena y tomando aleatoriamente un nuevo estado inicial.

Entre las ventajas de este modelo cabe destacar tanto su buena estimación de los tamaños del GOP como una de las mejores aproximaciones de la autocorrelación. En cuanto a sus desventajas, aunque los cálculos matriciales individualmente son simples, la complejidad del sistema completo (cadena principal junto con cadenas de clase de escena) hace que este modelo solamente sea tenido en cuenta para el modelado de fuentes individuales.

Modelos Autorregresivos

Aunque la mayoría abandona el modelo markoviano, los más avanzados vuelven a él aunque mediante el uso de cadenas dobles y sistemas modulados. A continuación se describen los tipos de modelos autorregresivos más utilizados.

Modelos orientados al GOP

Bajo este epígrafe se engloban modelos que consideran tanto el proceso de generación de escena, como el de creación del GOP, siendo este último fundamental, ya que el patrón del mismo también se genera. Estos modelos suelen ser combinación de varios métodos, aunque en la mayoría de los casos una cadena de Markov es la que se encarga de modelar los cambios de escena, dejando para la generación del GOP los modelos autorregresivos y fractales. Se distinguen dos clases de modelos, aquellos que consideran patrones fijos del GOP y los que modelan GOP's de tamaño y estructura variable.

Modelo de Doulamis

Este modelo asume un GOP de estructura fija [Turaga, 2000c], con la trama I seguida de 12 tramas B y P (I-B-B-P-B-B-P-B-B-P-B-B-P). Su funcionamiento es muy similar al modelo orientado a la escena, de hecho cada GOP sigue una clasificación basada en la actividad de la escena de acuerdo con la tasa media asociada a cada fotograma. El valor correspondiente al GOP puede ser calculado a partir del conocimiento de dicha tasa y de parámetros básicos

como el número de fotogramas que lo constituyen (*L*), así como el número total de fotogramas que componen la escena(*N*):

$$x_{GOP}(j) = \frac{1}{L}\sum_{i=0}^{L-1} x_{frame}(j \cdot L + i) \qquad siendo \quad j \in \left[0, \frac{N}{L}\right]$$

donde $x_{frame}(k)$ es la tasa media del fotograma *K*.

En función de este valor se determinan una serie de umbrales, TH y TL, a partir de los cuales se realiza la clasificación de actividades de escena:

$$Si \quad x_{GOP} \begin{cases} > T_H & Actividad \quad H \\ T_L \le x_{GOP} \le T_H & Actividad \quad M \\ < T_L & Actividad \quad L \end{cases}$$

Conocida la clasificación de actividades, el modelo implementa una cadena de Markov de tantos estados como niveles de actividad existan (H, M y L), modelando el proceso de decisión del nivel de actividad del GOP.

Una vez se ha decidido qué nivel de actividad tendrá el GOP, el modelo hace uso de varios procesos tipo AR(1), uno por cada tipo de fotograma (I, P ó B), para generar el correspondiente fotograma. Los parámetros asociados a cada proceso AR(1) varían en función del tipo de fotograma y nivel de actividad asociada, de tal forma que incluso, en casos de actividad baja los fotogramas P y B pueden ser considerados como constantes, como se indica en la Fig. D .5:

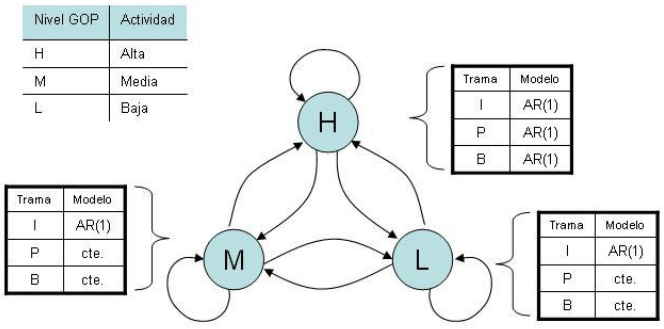


Fig. D .5:Modelo de Doulamis de tres estados de actividad y generación de GOP simplificada basada en procesos AR(1)

El modelo básico de Doulamis se aplica en dos variantes diferentes, que se exponen a continuación.

4.1.1.1 Modelo de dos estados (I + P)

Existe una variante a caballo entre el modelo de Doulamis y las cadenas de Markov que aunque simple puede resultar suficiente [Turaga, 2000c]. Este consiste en una cadena de Markov de dos estados, correspondientes a la generación de fotogramas I ó P, sin restricciones de estructuras de GOP fijas. Dichos procesos de generación de fotogramas se modelan mediante procesos

AR(1) con distribuciones gaussianas de acuerdo con el modelo de la Fig. D.6. El proceso AR(1) de generación de fotogramas P se reinicia cada vez que se produce una transición del estado I al P. De esta forma se consigue emular figuras de correlación adecuadas, esto es, el estado I asegura la correlación a largo plazo, y el estado P la correlación a corto plazo.

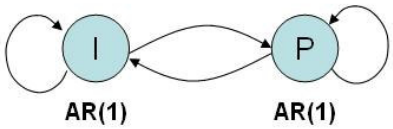


Fig. D .6: ***Modelo de dos estados (I+P)***

Si tiene en cuenta que es poco probable la existencia de GOPs de dimensión 1, el estado I puede ser simplificado al eliminar la probabilidad de permanencia en el mismo, con lo cual el modelo se reduce al cálculo de las tres probabilidades restantes, pudiendo ser modeladas mediante un único proceso AR(1), como se indica en [Chandra, 1995a / Chandra, 1995b]. Este modelo tan simple no tiene en cuenta la posibilidad de cambios de escena, por lo que solamente resulta adecuado para el modelado de vídeo con actividad más o menos fija, por ejemplo, en una videoconferencia.

4.1.1.2 Modelos Dobles de Markov (Doubly Markov Model)

Cuando el nivel de actividad de la secuencia de vídeo no es "fijo", se hace necesario retomar la idea del modelado de los mismos. El modelo doble de Markov está basado en la combinación del modelo de dos estados junto con el modelo de Doulamis. Para ello este modelo hace uso de una doble cadena de Markov, una para el modelado del proceso de selección del tipo de fotograma (I ó P) y otra para el de selección del nivel de actividad, véase [Turaga, 2000a]. En función de cuál de las dos cadenas se procesa antes aparecen las diferentes variantes de este modelo:

- Tipo I: La cadena de Markov principal es la que modela las transiciones del estado I al P y viceversa. Una vez en cada estado, la segunda cadena decide el nivel de actividad (L, M ó H) de forma idéntica a como era realizado en el modelo de Doulamis, tal como se muestra en la Fig. D .7. Al contrario del modelo de dos estados, dicho proceso se reinicializa cada vez que se accede a cualquiera de los estados I ó P, salvo que se repita de estado caso en el que se memoriza el valor anterior. Una vez decidido el nivel de actividad y el tipo de fotograma, este es generado mediante el correspondiente proceso AR(1).

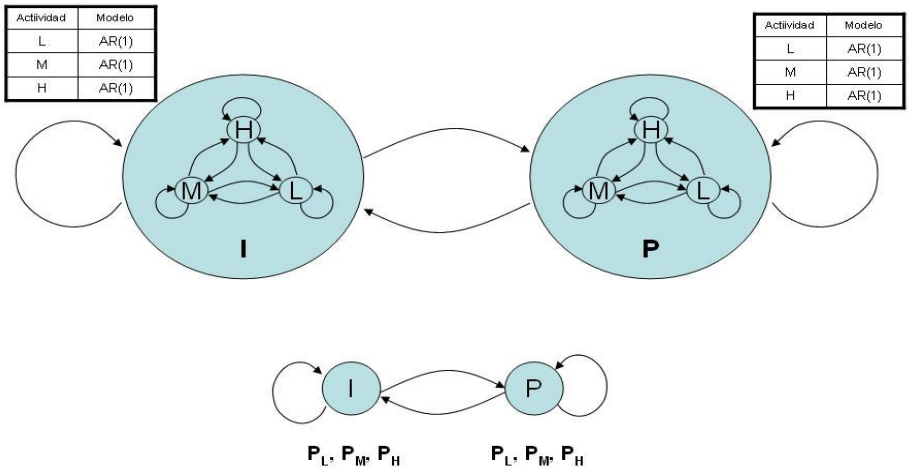


Fig. D .7: ***Modelo doble de Markov tipo I y modelo simplificado.***

La cadena interior, la decisión del nivel de actividad, puede ser simplificada tenida cuenta que las probabilidades de transición entre los tres estados pueden ser consideradas como incondicionales. Según esta idea, la cadena puede ser sustituida por el conjunto de probabilidades de selección de un nivel L, M ó H. Incluso existen desarrollos que solamente consideran esta idea para el estado I, manteniendo la segunda cadena del estado P, aunque sustituyen los procesos simples AR(1), por la modulación o multiplexación de varios de ellos, en lo que se conoce como procesos autorregresivos "pinchados" (*punctured*), como en [Chen, 2002].

- Tipo II: La cadena principal modela las transiciones entre los estados L, M y H, siendo cada uno de estos estados el que decide posteriormente el tipo de fotograma a generar. Para ello, una vez seleccionado un nivel de actividad se inicializa la generación de fotogramas, mediante el correspondiente proceso AR(1), empezando siempre por un fotograma I. A partir de aquí se generan transiciones al estado P hasta llegar a otro I, momento en el que se decide un nuevo nivel de actividad. Como se puede observar en la Fig. D .8, este modelo es muy similar al de Doulamis, salvo por el GOP de tamaño variable.

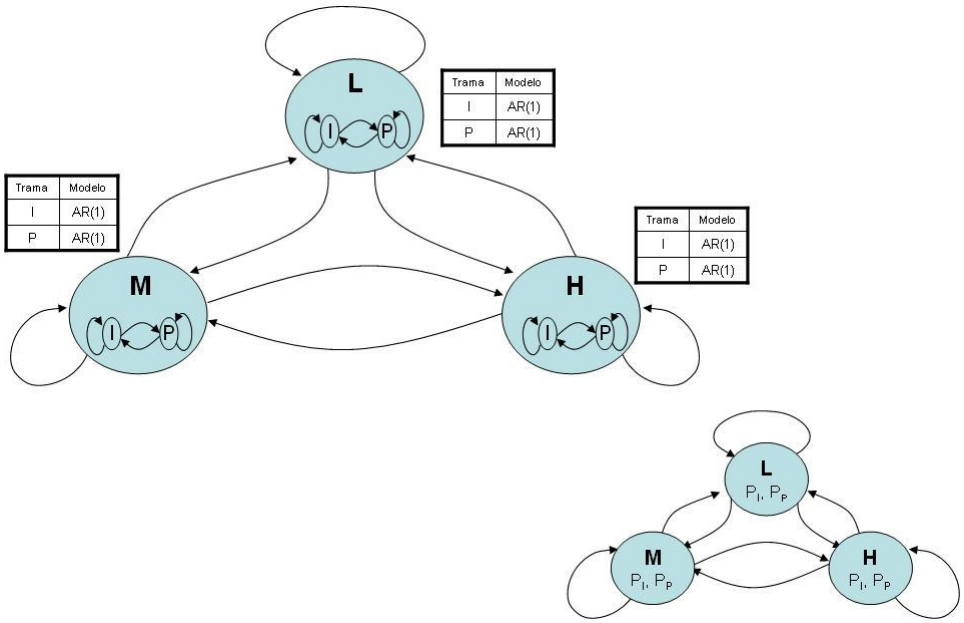


Fig. D .8: ***Modelo doble de Markov tipo II y modelo simplificado.***

Al igual que el tipo I, este modelo puede ser simplificado considerando las probabilidades incondicionales de transición al estado I ó P. Sin embargo en este caso el nivel de actividad debe ser decidido para cada fotograma, siendo por tanto variable dentro incluso del mismo GOP.

4.1.1.3 Modelos de GOP realista

La composición del GOP, aunque variable, sigue ciertas pautas en su generación, como es el hecho de que el codificador de vídeo, siempre que pueda, va a seguir un patrón fijo de dos fotogramas B seguidos de un fotograma I. Teniendo en cuenta esta puntualización, los modelos de dos estados han sido completados mediante la introducción de un tercer estado, el correspondiente a los fotogramas B. Dada la posibilidad cierta de que aparezcan parejas de fotogramas B, los modelos introducen un cuarto estado correspondiente a este caso, véase [Turaga, 2001], e incluso, en el caso más genérico, tantos estados como fotogramas B consecutivos se deseen considerar, véase [Casilari, 1998]. Con todo ello, el modelo resultante se indica en la Fig. D0.9, en el que se han introducido algunas restricciones razonables, como por ejemplo que no existen transiciones entre los estados B y BB, que no hay permanencia en los estados B, BB y P, y que al estado B se puede llegar desde el estado P, aunque no transita hacia el estado I.

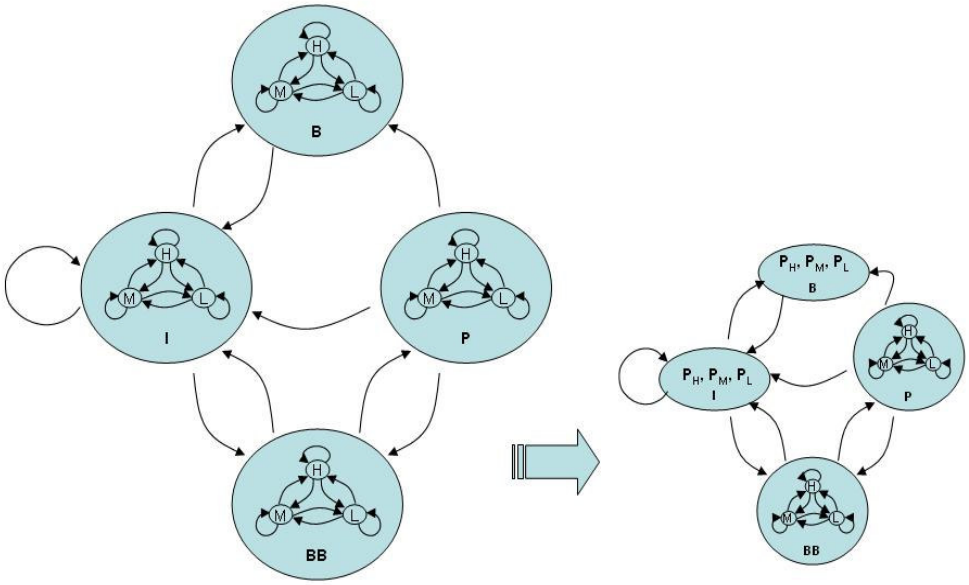


Fig. D .9: Modelo de GOP realista (El GOP puede tener formas como IBBPBBP, IBBPB, IB ó solo I) y modelo simplificado.

Al igual que en el modelo de dos estados de tipo I, este modelo puede ser simplificado haciendo uso de las probabilidades incondicionales asociadas a las transiciones entre los diferentes niveles de actividad. Para ello se sustituye la cadena de Markov interna por las correspondientes probabilidades incondicionales de generar un fotograma de acuerdo con un determinado nivel de actividad. Sin embargo, esto solamente va a ser cierto para el caso de los fotogramas I y B, puesto que aparecen con menor frecuencia y siempre en GOP diferentes, con intervalos relativamente altos entre ambos.

De la misma manera aparece la correspondiente adaptación del modelo de dos estados Tipo II, que se muestra en la Fig. D .10, teniendo en cuenta que se en este caso se debe asumir que todo el GOP presentará el mismo nivel de actividad, lo cual no es totalmente cierto:

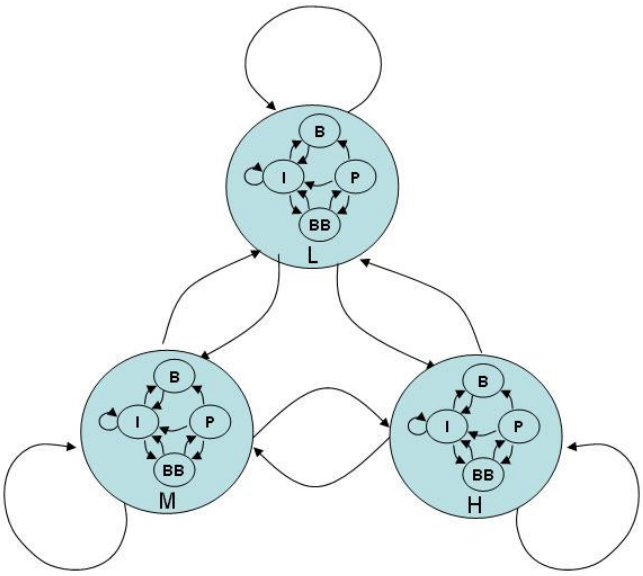


Fig. D .10: Modelo de GOP realista basado en el modelo de dos estados tipo II

Esta última variante no puede ser simplificada dada la gran dependencia del método de entrenamiento del modelo, especialmente en cuanto a la decisión inicial de los tipos de fotograma.

La complejidad de los modelos utilizados para el modelado de vídeo, hace que sean difíciles de aplicar en estudios de estimación de tráfico, dado que plantean un número demasiado elevado de parámetros, cuyos valores puede ser muy difícil obtener. Así, el modelo ON-OFF multinivel binomial se presenta como alternativa de bajo coste computacional para la estimación del ancho de banda asociado a los servicios de transmisión de vídeo. A continuación se plantean dos servicios de este tipo, con carácterísticas diferencias, como son el servicio de videoconferencia (servicio de tiempo real) y el de *streaming* de vídeo (con variantes desde *Expedited Forwarding* al *Best-effort*).

Fuentes de vídeo simétricas: Videoconferencia

En videoconferencia, según [Turaga, 2000c] existe la ventaja de tener cambios de escena sólamente en situaciones puntuales (p.e. al cambiar de orador en una multiconferencia) siendo, además, cambios regulares provocados por el movimiento suaves y lentos.

El tiempo de interllegada de frames es el inverso de la tasa de frames (cuadros) que puede considerarse constante. Para sistemas PAL el periodo interframe es de 40 ms, transmitiendo con una cadencia de hasta 25 frames/s, configurable por el usuario, ya que este valor produce un "stream" de video de alta calidad que requiere por tanto un ancho de banda elevado. Precisamente, como los

objetos/personas mostrados por pantalla suelen encontrarse estáticos o realizan movimientos lentos y cortos (típicos de una conversación), dicha tasa de transmisión puede ser reducida, utilizando las técnicas de codificación y compresión correspondientes, para producir vídeo de calidad aceptable con tasas de transmisión de 5 a 15 frames/s (valor típico de 12 frames/s).

El parámetro fundamental a calcular es entonces el tamaño de dichas frames. Este valor depende tanto de la calidad de la imagen, como del tamaño y detalle de la misma. Una vez fijadas esas características por el usuario, este parámetro suele ser modelado mediante una distribución Gamma como la siguiente:

$$f(t) = \frac{\lambda(\lambda t)^{s-1}}{\Gamma(s)} \cdot e^{-\lambda t} \tag{5.26}$$

Con valores típicos de λ=0.02353 y s=3.066, donde t es el tamaño del frame en bytes y λ y s los parámetros de forma con los valores indicados y $\Gamma(s)$ una función de la forma:

$$\Gamma(s) = \int_0^\infty t^{s-1} e^{-t} dt \tag{5.27}$$

En cuanto al valor medio es muy dependiente del ancho de banda disponible o del tipo de servidor y aplicación que se estén utilizando, así como del sistema de compresión con el que se traten los datos. De hecho tanto en los sistemas de videoconferencia como en "video-streaming" es usual que la aplicación al ser instalada consulte al usuario sobre el tipo de acceso del que dispone (MODEM, ADSL...), e incluso la calidad de imagen y/o vídeo deseadas.

El valor de referencia para el tamaño e un frame típico es de 1.5 KB, con lo que para una tasa de generación de 12 frames por segundo, el ancho de banda requerido por la conexión es de 144 Kbps. Este valor es aceptable para accesos que disponen de alrededor de 256 Kbps, en donde tanto en transmisión como en recepción se trabaja con este flujo de bits. Para accesos de menor capacidad, como por ejemplo la RDSI a 128 Kbps, el tamaño del frame (y por lo tanto la calidad de la imagen) debe reducirse para poder acomodarlo al flujo de información que este tipo de acceso puede soportar.

Identificación de parámetros:

Como se puede observar este tipo de servicio tiene un gran parecido con el tráfico de voz comprimido sobre IP con la diferencia de que el flujo de bits es mayor y que se suele dar una flexibilidad dentro de los estándares con el objetivo de encontrar una solución de compromiso entre la calidad deseada y el ancho de banda disponible. Es por ello que los parámetros que intervienen en el servicio de videoconferencia sean similares los del servicio de VoIP. Sin embargo, van a existir diferencias directamente relacionadas con la naturaleza del vídeo y de los mecanismos de codificación y compresión, especialmente en nivel de ráfaga.

En el nivel de Conexión se consideran:

- Numero de conexiones al ISP (α) por hora o día: número de conexiones que hace en media y en una hora de un día normal un usuario del tipo evaluado.
- Duración de la conexión (T_s): tiempo que está conectado, en media y de manera consecutiva, un usuario del tipo evaluado.
- Velocidad de la conexión (V): velocidad del acceso físico considerado.
- Personal que existe para ese tipo de usuario (solo en el caso de usuarios de empresa).

En el nivel de Sesión:

- Tiempo entre sesiones ($T_{silencio}$): tiempo que pasa entre dos sesiones consecutivas pedidas por el usuario, en este servicio indica el tiempo entre dos videoconferencias consecutivas.
- Duración de la sesión ($T_{llamada}$): tiempo que dura la videoconferencia solicitada.

Y en el Nivel de ráfaga:

- Tiempo de transmisión del frame (T_{frame}): duración de la transmisión de cada una de las frames que forman la secuencia.
- Tiempo entre frames (T_{off}): tiempo de llegada entre cada frame.

Estos dos parámetros dependen directamente del ancho de banda del acceso a Internet, del tamaño medio del frame, y de la tasa de generación de frames.

Estimación del tráfico generado por fuentes de videoconferencia

El modelo ON-OFF multinivel resultante es muy similar al estudiado para el servicio de voz, por lo que CASUAL proporciona la estimación del tráfico generado para este tipo de fuentes se calcula del mismo modo que se describió en el apartado 4.3.2.3.

Utilizando una línea ADSL que con una velocidad mínima de 128 Kbps (en subida ó upstream, es decir desde el usuario hacia el DSLAM), con un valor típico del tamaño del frame de alrededor de 5 Kb para 12 frames/s, la transmisión del frame toma un valor de alrededor de 0.04 segundos, mientras que el tiempo entre frames es de 0.043 segundos.

Presuponiendo una probabilidad de estado activo de sesión de 0.1 (esto depende del número y duración de la sesión estudiada, y por lo tanto del tipo de usuario considerado), se obtiene un caudal binario por cada usuario conectado de 14.8 Kbps.

Aunque este valor parece bajo considerando que se está hablando de una videoconferencia, hay que tener en cuenta que se está suponiendo una probabilidad de efectuar esa videoconferencia baja pero lógica, y una calidad de imagen mínima.

Fuentes de vídeo asimétricas: Video-streaming

Este servicio se basa en la transmisión en tiempo "peudo-real" de información de audio y video simultáneamente, ya que para poder transmitir este tipo de información es necesario utilizar en el receptor un *buffer* que almacena previamente parte de la información para asegurar una calidad de servicio aceptable, provocando así, un retardo en la transmisión mayor o menor, dependiendo generalmente de factores como el ancho de banda y los recursos hardware/software del terminal (siendo más critico el primero de estos dos parámetros). Actualmente casi todos los servicios de distribución de contenido multimedia utilizan técnicas de "streaming", incluso en el caso de distribución de vídeo en tiempo real (TV y HDTV), con la única diferencia de que para este tipo de servicios los requerimientos de retardo y ancho de banda son lógicamente más estrictos.

Si se compara este tipo de tráfico con el de videoconferencia se observa que en este caso las exigencias de ancho de banda y retardo pueden ser mayores debido a que la naturaleza de los contenidos multimedia son muy variables (vídeo dinámico con gran cantidad de cambios de escena, objetos en movimiento dentro de a misma escena, etc.).

Para el estudio de este servicio, se va a considerar la transmisión de información codificada bajo el estándar MPEG, puesto que la infinidad de codificaciones posibles van a comportarse de una manera similar por lo que los valores de tráfico no serán excesivamente diferentes los obtenidos en este caso.

De acuerdo con el estándar MPEG, los tres diferentes tipos de tramas van a caracterizar el nivel de ráfaga dentro de la transmisión. Experimentalmente todos los modelos expuestos anteriormente tienen mejores resultados conforme los tamaños de los *buffers* de recepción aumentan ya que se aprovechan los efectos de la autocorrelación a más largo plazo del modelo de escena, siendo una clara ventaja a la hora de realizar un dimensionado. Aunque el tamaño de dicho *buffer* puede ser un parámetro directamente relacionado con los protocolos RTP y RTSP, suele ser normal que su valor sea estimado en tiempo real en función del ancho de banda disponible y de la información a transmitir, permitiendo así almacenar varios segundos de tramas, mediante los

cuales "suavizar" los efectos provocados por la congestión de la red (traducido en el aumento del retardo o del jitter) sin una perdida de calidad de servicio apreciable para el usuario.

Algunos valores típicos correspondientes a las características de un archivo de vídeo para su transmisión mediante video-streaming se resumen en la Tabla D .II. Dichos valores han sido extraídos extraídos de [Turaga, 2000a]:

Estadistica	Valor medio
Duración	2 horas
Frames de video	171000
Resolución	480 lineas x 504 pixels
Resolución del píxel	8 bits/píxel
Frame rate	24 f/s
Velocidad media	5.34 Mbps
Compresión media	8.7

Tabla D .II: Valores típicos de un archivo de vídeo (película) para su difusión mediante video-streaming

Estos valores se corresponden con un archivo de vídeo catalogado como película de acción, a calidad media, con muchos cambios de escena y elevada actividad de los objetos incluidos en las mismas. Sin embargo, estos valores van a variar enormemente de unos archivos de vídeo a otro según la naturaleza de los contenidos, e incluso dentro de un mismo archivo en función de los cambios de escena y la dinámica de las mismas. El establecimiento de valores típicos no es una tarea simple, por lo que normalmente se recomienda el dimensionado para los casos más restrictivos, esto es vídeo de calidad normal o alta con escenas de gran dinamismo.

Identificación de parámetros para su modelado en CASUAL:

Al igual que para el resto de servicios, los niveles de conexión y sesión van a resultar muy similares salvo por los valores concretos de cada parámetro. Los parámetros que intervienen en videostreaming son:

Al nivel de conexión:

- Numero de conexiones al ISP por hora o día (α): número de conexiones que hace en media y en una hora de un día normal un usuario del tipo evaluado.
- Duración de la conexión (T_s): tiempo que está conectado, en media y de manera consecutiva, un usuario del tipo evaluado.
- Velocidad de la conexión (V): velocidad del acceso físico considerado.

- Personal que existe para ese tipo de usuario. Solo para el caso de usuarios de empresa, aunque este tipo de servicios solamente va a ser empleado de forma marginal

Para el modelado del nivel deSesión:

- Tiempo entre sesiones (T_{espera}): tiempo que pasa entre dos sesiones consecutivas pedidas por el usuario, en este servicio indica el tiempo entre dos sesiones de videostreaming consecutivas.
- Duración de la sesión (T_{video}): tiempo que dura la sesión de video-streaming solicitada.

Por último, y tal como se indicó en el apartado anterior, el nivel de ráfaga va ser modelado por los diferentes tipos de frames que van a generarse:

- Tiempo de duración de la transmisión de cada uno de los frames I,P,B, supuesta una secuencia de GoP típica. Normalmente estos valores se sustituyen mediante el uso de las probabilidades asociadas a la generación de cada tipo de frame y a los tamaños medios de cada tipo de trama.

Estimación del tráfico generado por fuentes de video-streaming

Para el tráfico de videoconferencia se utilizó una metodología muy semejante a la del tráfico de voz ya su funcionamiento es prácticamente igual modificando valores de parámetros como el caudal de la ráfaga o frame, su duración, o las duraciones de las sesiones. Por ello la metodología de fuentes ON-OFF multinivel, una por nivel, resulta válida, de forma que la estimación del tráfico agregado hace uso de binomiales para los niveles de conexión y sesión, los cuales no se consideran como tráfico a ráfagas, mientras que en el nivel de ráfaga se hace uso de binomiales negativas.

En el caso del tráfico de *video-streaming* se introduce una ligera modificación puesto que a nivel de ráfaga no existe una única posibilidad de transmitir frames o no, sino que dicha posibilidad está modulada por la probabilidad de que la trama sea I, P ó B con características diferentes. Para su modelado se va a sustituir el modelo ON-OFF del nivel de ráfaga por tres fuentes ON-OFF simples e independientes, que modelarán la generación de frames de cada tipo, y que se muestra en la Fig. D .11. La independencia entre cada fuente ON-OFF se justifica por el hecho de que se va a considerar un GOP determinista (de estructura fija), con los que las probabilidades de transición van a ser calculadas a priori.

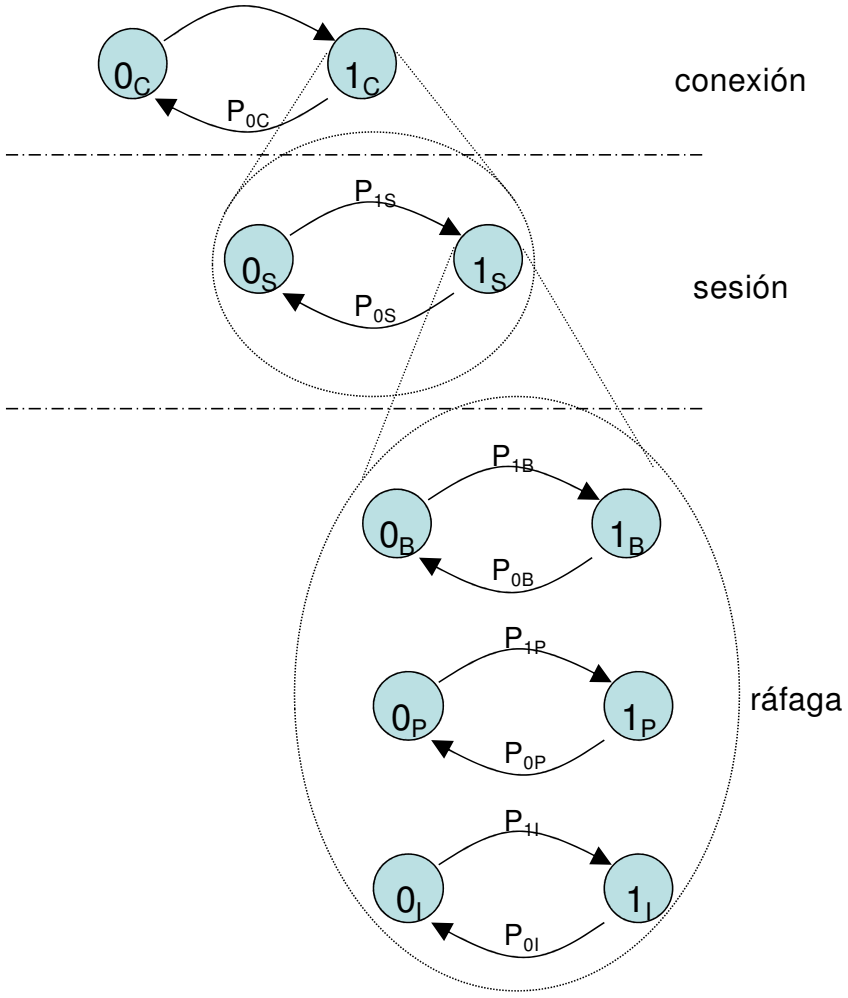


Fig. D .11: Modelo ON-OFF multinivel para generación de tráfico de vídeo

Como se puede observar las fuentes ON-OFF del nivel de ráfaga modelan la generación de cada tipo de frame, mientras que los niveles superiores, conexión y sesión, se comportan como ya se explicó en el tráfico IP genérico.

Para la agregación se consideran binomiales clásicas para los dos niveles superiores, mientras que para el nivel de ráfaga es más conveniente realizar la agregación mediante binomiales negativas.

Supuestos valores de activación de cada una de las tres tramas I, P, B de 0.16, 0.427, 0.42 respectivamente, una probabilidad de activación de este servicio de 0.1 y una velocidad de 128 kbps, en el caso de una población de 10000 habitantes (supuestos todos conectados) se obtiene un valor de régimen binario por usuario de 23.6 Kbps. Si por el contrario la probabilidad de conexión por usuario es del 0.1, dicho valor será de 39.13 Kbps por usuario.

Apéndice E.

Publicaciones relacionadas con la Tesis

Revistas

1. J.A. Portilla, Klaus D. Hackbarth, Alberto E. García: "New Trends in design for mobile networks planning tools", The Journal of the Institution of British Telecommunication Engineers,Vol. 2 Issue 3 (ISBN: 1470-5826) pp. 112-115, Julio 2001

2. Dieter Elixman, Mark Scanlan, Alberto E. García, Klaus D. Hackbarth, Annette Hillebrand: "The Economics of IP Networks - Market, Technical and Public Policy Issues Relating to Internet Traffic Exchange", Study for the European Commission, Ed. WIK-Consult GmbH / European Commission, Vol. 1, pp. 1-255, Mayo 2002

3. Alberto. E. García, Klaus D. Hackbarth: "Analytical Model for Voice over IP Traffic Characterization", WSEAS Trans. on Communications (ISBN 1109-2742), pp. 59-65, Julio 2002

4. A.E. García, K.D. Hackbarth: "Next Generation IP Access Networks Planning: Approximated Methods", WSEAS Trans. on Communications, Vol. 5, Issue 3 (ISSN 1109-2742), pp. 535-542, Marzo 2006

5. Alberto E. García, Klaus D. Hackbarth: "Approximation to a Behavioral Model for Estimating Traffic Aggregation Scenarios", Journal of Universal Computer Science, Vol. 14, pp. 731-744, Abril 2008

Congreso

Internacional

1. Alberto E. García, Klaus D. Hackbarth: "Estimación de la Matriz de Tráfico y la Estructura de una RDSI-BA ATM", YUFORIC'97 IEEE Computer Society y Universitat Politécnica de Cataluña, Abril 1997

2. Alberto E. García, Klaus D. Hackbarth: "Título: NETrix: A Procedure for the Calculation of the Network Structure and Traffic Matrix in

ATM Networks", IEEE International Conference on ATM ICATM'98, pp. 359-364 (ISBN 0-7803-4982-2 IEEE Catalog Number: 98EX190), Junio 1998

3. Alberto E. García, Klaus D. Hackbarth: "SWINET: Switched Network Emulation: Logical Network Structure Design and Access/Backbone Topology Planning", 2nd International Conference on Telecomm. and E-Commerce (ICTEC), Noviembre 1999

4. Alberto E. García, Klaus D. Hackbarth, J. A. Portilla: "Methods and Solutions for Network Planning Tools", Mobile Comm. Summit 2000, Information Society Technologies (IST), Noviembre 2000

5. J.A. Portilla, Klaus D. Hackbarth, Alberto E. García: "INEDAC project: A tool to calculate interconnection tariff based on a bottom-up method", 1st European Conference on Universal Multiservice Network ECUMN'2000 (ISBN 0-7803-6419-8 IEEE, Catalog Number 00EX423), pp. 319-327, Noviembre 2000

6. Klaus D. Hackbarth, Alberto E. García, Roberto Ortiz: "Design of Telecommunication Network Planning Tools with Functional Objects and their Application in Tele-Education", Workshop on Network Traffic Engineering, Febrero 2002

7. Klaus D. Hackbarth, Alberto E. García, J.A. Portilla: "INEDAC: a bottom up model for cost calculation of telecommunication network interconnection", 9th International Conference on Telecommunication Systems, Febrero 2001

8. J.A. Portilla, Klaus D. Hackbarth, Alberto E. García: "New Trends in design for mobile networks planning tools", 40 Th European Telecommunication Congress FITCE, Agosto 2001

9. Alberto E. Garcia, Klaus D. Hackbarth, Andreas Brand, Ralph Lehnert: "Analytical Model for Voice over IP traffic characterization", 6th WSEAS International Multiconference CSCC (ISBN-960-8052-63-7), Julio 2002

10. Alberto E. García, Klaus D. Hackbarth, Ralph Lehnert, Andreas Brand: "Validación mediante simulación de los modelos analíticos para la estimación de tráfico en redes multimedia: Servicios de Voz sobre IP", International Conference TELEC 2002 (ISBN-84-8138-506-9), Julio 2002

11. A.E. García, K.D. Hackbarth: "Simplified methods for Next Generation IP Access Networks planning", 5th WSEAS International Conference on ELECTRONICS, HARDWARE, WIRELESS and

OPTICAL COMMUNICATIONS (EHAC '06 ISSN: 1790-5117 ISBN: 960-8457-41-6), Febrero 2006

12. A.E. García, K.D. Hackbarth: "Approximation to a behavioural model for estimating traffic aggregation scenarios", Workshop on Socio-Economic Aspects of NGI in Relation with its Architecture, Design & Dimensioning, Junio 2006

13. García Gutiérrez, Alberto E., Hackbarth, Klaus, Rodríguez de Lope, Laura: "Approximation to a Behavioural Model for Estimating", Workshop on Socio-Economics Aspects of Next Generation Networks, Junio 2007

14. Laura Rodríguez, Alberto E. García, Klaus D. Hackbarth: "Influence of the Traffic Engineering Scheme and QoS in the Dimensioning of Broadband Access Networks", 8th WSEAS International Conf. on Distance Learning, Multimedia and Video Technologies (ISBN 978-960-474-005-5 / ISSN 1790-5109), Septiembre 2008

15. Laura Rodríguez, Alberto E. García, Klaus D. Hackbarth: "Cost models for Next Generation Networks with Quality of Service parameters", NETWORKS 2008 13th International Telecommunications Network Strategy and Planning Symposium, Noviembre 2008

Nacionales

1. Alberto E. García, Klaus D. Hackbarth: "Planificación Estratégica de la Estructura Lógica de Redes de Banda Ancha, Proceedings del JITEL'97 (ISBN 84-89654-04-2), pp. 5-13, Septiembre 1997

2. Alberto E. García: "GIT/DICOM: Herramientas de Planificación de Redes", II Congreso Nacional de Ingeniería de Telecomunicación, Junio 1998

3. Alberto E. García, Klaus D. Hackbarth: "PLATON: Planificación de Topologías por Niveles", TELECOM 99 Univ. Politécnica de Madrid / Barcelona, Septiembre 1999

4. Klaus D. Hackbarth, Alberto E. García: "La influencia de la interconexión en el diseño y dimensionamiento de redes de comunicación", TELECOM 99 Univ. Politécnica de Madrid / Barcelona, Septiembre 1999

5. Alberto E. García, Klaus D. Hackbarth: "Generación de Escenarios: Planificación de la Estructura Lógica y de las Topologías de Acceso y

Backbone (SWINET)", JITEL 99 (ISBN 84-89315-14-0), Septiembre 1999

6. Alberto E. García, Klaus D. Hackbarth, Andreas Brand, Ralph Lehnert: "Modelos analíticos para tráfico de Voz sobre IP", III Jornadas de Ingeniería Telemática (JITEL 2001 ISBN 84-7653-783-2), pp. 477-484, Septiembre 2001

7. A.E. García, K.D. Hackbarth: "Métodos simplificados para la planificación del acceso en Redes IP de Próxima Generación", VI Jornadas de Ingenieria Telemática (JITEL 2007), Septiembre 2007

8. A.E. García, K.D. Hackbarth, L. Rodríguez de Lope: "Modelo analítico para el cálculo de coste de servicios Bitstream con criterios de QoS", VI Jornadas de Ingeniería Telemática (JITEL 2007), Septiembre 2007

Otras publicaciones

1. J. Alvarez, A. E. García, K.D.Hackbarth, R.Ortiz: "Applications of Security Protocols for Tele-education Environments: Virtual Exams", International Conference on Education IADAT-e2004: Innovation, Technology and Research in Education (ISBN:84-933971-0-5), pp. 397-401, Julio 2004

2. Klaus D. Hackbarth, A.E.García, Carlos Díaz: "Modelling and Simulation of a Dynamic QoS Management in a Satellite Communication System", 2nd International Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks HET-NETs'04 (ISBN:0-9540151-6-9), Julio 2004

3. A.E.García, K.D.Hackbarth, J.A.Portilla, R.Ortiz: "Collaborative Environment for Tool Sharing in the Framework of Euro-NGI Network of Excellence", 2nd International Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks HET-NETs'04 (ISBN: 0-9540151-6-9), Julio 2004

4. A.E. García, K.D. Hackbarth, R. Ortiz: "Web-Based Service for Remote Execution: NGI Network Design Application", 1st EURONGI Conference on Next Generation Internet Networks NGI'05 (ISBN 0-7803-8901-8), Abril 2005

5. A.E. García, K.D. Hackbarth: "WeBaSeReX: Next Generation Internet Network Design Using Web Based Remote Execution Environments", IEEE International Conference on Services Computing ISC'05 (ISBN 0-7695-2408-7), Julio 2005

6. Alberto E. García, Laura Rodríguez, Klaus D. Hackbarth, Miguel Faro: "3GPP towards IMS: Quality of Service and Charging", 8th WSEAS International Conf. on Distance Learning, Multimedia and Video Technologies (ISBN 978-960-474-005-5 / ISSN 1790-5109), Septiembre 2008

7. A.E. García, K.D. Hackbarth, R. Ortiz: "WeBaSeReX (Web Based Service for Remote Execution): Implementación de Servicios de Acceso Seguro a Aplicaciones Compartidas", XV Jornadas TELECOM I+D (ISBN: 84-689-3794-0), Noviembre 2005

8. Alberto E. García, Andreas Berl, Karin A. Hummel , Roman Weidlich, Amine Houyou, Klaus D. Hackbarth, Herman de Meer, Helmut Hlavacs,: "An Economical Cost Model for Fair Resource Sharing in Virtual Home", 4th EURO-NGI Conference on Next Generation Internet Networks (IEEE Catalog Number CFP08596-CDR),Abril 2008

Apéndice F.

Bibliografía

[Adas, 1997] A. Adas, "Traffic Models in Broadband Networks," in *IEEE Communications Magazine*. vol. 35, 1997, pp. 82-89.

[Addie, 1998] R. Addie, M. Zukerman, and T. Neame, "Broadband Traffic Modelling: Simple Solutions to Hard Problems," in *IEEE Communications Magazine*, 1998, pp. 88-95.

[Adler, 2001] M. Adler, T. Bu, R. K. Sitaraman, and D. Towsley, "Tree Layout for Internal Network Characterizations in Multicast Networks," in *Third International COST264 Workshop on Networked Group Communication*, 2001, pp. 189-204.

[Akimaru, 1999] H. Akimaru and K. Kawashima, *Teletraffic: Theory and Applications. 2nd Edition*: Springer, 1999.

[Altman, 2002] E. Altman, K. Avrachenkov, and C. Barakat, "*TCP Network Calculus: The case of large delay-bandwidth product*," in *IEEE Infocom 2002*, New York, USA, 2002.

[Anderson, 2007] N. Anderson, "The YouTube effect: HTTP traffic now eclipses P2P," Ars Technica: The Art of Technology Portal, 2007.

[Arana, 1997] F. S. Arana, A. Bartolomé, G. Paramés, and F. Fermández, "La planificación de la red de conmutación y control en la telefonía móvil," *Comunicaciones de Telefónica I+D,* Dic 1997 1997.

[Arijs, 2000] P. Arijs, B. V. Caenegem, P. Demester, P. Lagasse, W. V. Parys, and P. Achten, "Design of Ring and Mesh based WDM Transport Network," in *Optical Networks Magazine SPIE/Baltzer Science Publishers*, 2000, pp. 25-40.

[Arlitt, 1996] M. F. Arlitt and C. L. Williamson, "Web Server Workload Characterization: The Search for Invariants," *ACM SIGMETRICS Performance Evaluation Review,* vol. 24, pp. 126-137, May 1996 1996.

[Awerbuch, 1998] B. Awerbuch, Y. Du, B. Khan, and Y. Shavitt, "Routing Through Networks with Hierarchical Topology Aggregation," in *3rd IEEE Symposium on Computers & Communications (ISCC'98)*, Athens, Greece, 1998, pp. 406-412.

[Awerbuch, 2001] B. Awerbuch and Y. Shavitt, "Topology Aggregation for Directed Graphs," *IEEE/ACM Transaction on Networking,* vol. 9, pp. 82-90, February 2001 2001.

[Barford, 1999] P. Barford, A. Bestavros, A. Bradley, and M. Crovella, "Changes in Web Client Access Patterns: Characteristics and Caching Implications," *World Wide Web, Special Issue on Characterization and Performance Evaluation,* vol. 2, August 1999 1999.

[Barford, 1998] P. Barford and M. Crovella, "Generating representative Web workloads for network and server performance evaluation," *ACM SIGMETRICS Performance Evaluation Review,* vol. 26, pp. 151-160, 1998.

[Bartoli, 2001] M. Bartoli and P. Castelli, "Dimensioning Differentiated Service Networks," CSELT2001.

[Berger, 2000] A. W. Berger and Y. Kogan, "Dimensioning Bandwidth for Elastic Traffic in High-Speed Data Networks," *IEEE/ACM Transactions on Networking,* vol. 8, pp. 643-654, Oct 2000 2000.

[Bley, 2002] A. Bley and T. Koch, "Integer programming approaches to access and backbone IP-network planning." vol. 2003: Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB), 2002.

[Bonald, 2003] T. Bonald, P. Olivier, and J. Roberts, "Dimensioning high speed IP access networks," in *18th International Teletraffic Congress*, Berlin, Germany, 2003, pp. 241-250.

[Bonald, 2000] T. Bonald and J. W. Roberts, "Performance of Bandwidth Sharing Mechanisms for Service Differentiation in the Internet," in *ITC Specialist Seminar* Monterey, USA, 2000.

[Boorstyn, 2000] R. Boorstyn, A. Burchard, J. Liebeherr, and C. Oottamakorn, "Effective Envelopes: Statistical Bounds on Multiplexed Traffic in Packet Networks," in *IEEE INFOCOM 2000*, 2000.

[Boudec, 1996] J.-Y. L. Boudec, "Network Calculus Made Easy," Laboratoire de Réseaux de Communication Ecole Polytechnique Fédérales de Laussane (EPFL), Lausanne, Switzerland, Technical Report EPFL-DI 96/218, December 1996 1996.

[Boudec, 1998] J.-Y. L. Boudec, "Application of Network Calculus to Guaranteed Service Networks," *IEEE Transactions on*

Information Theory, vol. 44, pp. 1087-1096, May 1998 1998.

[Boudec, 2000a] J.-Y. L. Boudec and P. Thiran, "A Short Tutorial on Network Calculus I: Fundamental Bounds in Communication Networks," in *IEEE International Symposium on Circuits and Systems (ISCAS'00)*, Geneva, Switzerland, 2000a, pp. 93-96.

[Boudec, 2000b] J.-Y. L. Boudec, P. Thiran, and S. Giordano, "A Short Tutorial on Network Calculus I: Fundamental Bounds in Communication Networks," in *IEEE International Symposium on Circuits and Systems (ISCAS'00)*, Geneva, Switzerland, 2000b.

[Boudec, 2002] J. Y. L. Boudec and P. Thiran, *Network Calculus: A theory of deterministic Queuing Systems for the Internet* vol. LNCS 2050: Springer Verlag 2002.

[Braden, 1994a] R. Braden, D. Clark, and S. Shenker, "RFC-1633: Integrated Services in the Internet Architecture: an Overview," Netwrok Working Group, RFC 1633, June 1994 1994a.

[Braden, 1994b] R. Braden and D. C. S. Shenker, "Integrated Services in the Internet Architecture: an Overview," in *RFC 1633*, IETF, Ed., 1994b.

[Caglar, 1998] M. Caglar, K. R. Krishnan, and I. Saniee, "Estimation of Traffic Parameters in High-Speed Data Networks," in *16th International Teletraffic Congress (ITC'98)*, Edinburgh, UK, 1998, pp. 867-876.

[Casilari, 1998] E. Casilari, A. Reyes, A. Diaz-Estrella, and F. Sandoval, "Characterisation and Modellinf of VBR Video Traffic," *Electronic Letters,* vol. 34, pp. 968-969, May 1998 1998.

[Cisco, 1999] Cisco, "Dedicated Internet Access," C. Systems, Ed., 1999.

[CISCO, 2005] CISCO, "Diffserv-The Scalable End-to-end Quality of Service Model," Cisco System Whitepaper Series, 2005.

[Clark, 1999] D. Clark, W. Lehr, and I. Liu, "Provisioning for Bursty Internet Traffic: Implications for Industry and Internet Structure," in *MIT ITC Workshop on Internet Quality of Service*, 1999.

[CMT, 2006] CMT, "Informe Anual 2006," 2006.

[COST-257, 2000] COST-257, "Impacts of New Services on the Architecture and Performance of Broadband

Networks," COST-257, Würzburg, Germany, Final ReportSeptember 2000 2000.

[Courcoubetis, 1995] C. Courcoubetis and R. Weber, "Effective Bandwidth for Stationary Sources," *Probability in Enginering and Informational Sciences,* vol. 9, pp. 285-294, 1995 1995.

[Courcoubetis, 2002] C. A. Courcoubetis, A. Dimakis, and G. D. Stamoulis, "Traffic Equivalence and Substitution in a Multiplexer With Applications to Dynamic Available Capacity Estimation," *IEEE/ACM Transactions on Networking,* vol. 10, pp. 217-231, April 2002 2002.

[Crovella, 1998] M. Crovella, M. Taqqu, and A. Bestavros, "Heavy-Tailed Probability Distributions in the World Wide Web," in *A Practical Guide To Heavy Tails*, C. Hall, Ed. New York, 1998, pp. 3-26.

[Cruz, 1991] R. L. Cruz, "A Calculus for Network Delay, Part II: Network Analysis," *I EEE Transactions on Information Theory,* vol. 37, pp. 132-141, January 1991 1991.

[Cruz, 1995] R. L. Cruz, "Quality of Service Guarantees in Virtual Circuit Switched Networks," *IEEE Journal on Selected Areas in Communications,* vol. 13, pp. 1048-1056, August 1995 1995.

[Chandra, 1995a] K. Chandra and A. R. Reibman, "Modeling traffic and statistical gains for multimedia applications," in *IEEE/ACM Conf. on Community Networking*, 1995a, pp. 171-178.

[Chandra, 1995b] K. Chandra and A. R. Reibman, "Modeling Traffic and Statistical Gains for Multimedia Applications," in *2nd International IEEE/ACM Conf. on Community Networking*, 1995b, pp. 171-178.

[Chang, 1998] C.-S. Chang, "On Deterministic Traffic Regulation and Services Guarantees: A Systematic Approach by Filtering," *I EEE Transactions on Information Theory,* vol. 44, pp. 1097-1110, May 1998 1998.

[Charzinski, 2002] J. Charzinski, "Internet Traffic - Characteristics, Performance and Models Tutorial," ICNP, Paris, FranceNov. 2002 2002.

[Chen, 2007] T. M. Chen, "Network Traffic Modelling," in *The Handbook of Computer Networks, Volume 3: Distributed Networks, Network Planning, Control, Management, New Trends and Applications*, H. Bidgoli, Ed., 2007.

[Chen, 2002] T. P.-c. Chen and T. Chen, "Markov Modulated Punctured Autoregressive Processes for Traffic and

Channel Modeling," in *International Packet Video Workshop 2002*, Pittsburgh, USA, 2002.

[Chen, 1995] Y. Chen, Z. Deng, and C. Williamson, "A model for self-similar ethernet LAN traffic: design implementation and performance implications," University of Saskatchewan, Canada1995.

[Cho, 2001a] K. Cho, R. Kaizaki, and A. Kato, "Aguri: An Aggregation-Based Traffic Profiler," in *COST 263: Proceedings of the Second International Workshop on Quality of Future Internet Services*, 2001a, pp. 222-242.

[Cho, 2001b] K. Cho, R. Kaizaki, and A. Kato, "Aguri: An Aggregation-based Traffic Profiler," in *Quality of Future Internet Services, 2nd COST 263 International Workshop (QofIS)*, Coimbra, Portugal, 2001b.

[Choi, 1999] H.-K. Choi and J. O. Limb, "A behavioral Model of Web Traffic," in *IEEE International Conference on Network Protocols ICNP 1999*, Toronto, Canada, 1999, pp. 327-334.

[Deng, 1996] S. Deng, "Empirical Model of WWW Document Arrivals at Access Link," in *IEEE International Conference on Communications (ICC'96)*, Dallas, USA, 1996, pp. 1796-1802.

[Dovrolis, 1997] K. Dovrolis, M. P. Vedam, and P. Ramanathan, "The selection of the token bucket parameters in the IETF guaranteed service class," Tech. Rep., University of Wisconsin-Madison1997.

[Duffield, 1999] N. G. Duffield, P. Goyal, A. G. Greenberg, P. P. Mishra, K. K. Ramakrishnan, and J. E. v. d. Merive, "A Flexible Model for Resource Management in Virtual Private Networks," in *ACM SIGCOMM'99*, Cambridge, Massachusetts, USA, 1999, pp. 95-108.

[Elixmann, 2002] G. Elixmann, M. Scanlan, A. E. García, K. D. Hackbarth, A. Hillebrand, G. Kulenkampff, and A. Metzler, "The economics of IP networks - Market, technical and public policy issues relating to Internet Traffic Exchange," wik-Consult, Bad Honnef, GermanyMay 2002 2002.

[Erramilli, 1994] A. Erramilli, R. P. Singh, and P. Pruthi, "Chaotic Maps as Models of Packet Traffic," in *14th International Teletraffic Congress ITC 14*, Amsterdam, Netherland, 1994, pp. 329-338.

[EURONGI, 2005] EURONGI, "Survey Describing Current Internet Design Tools," Technical Report, Ref. D.JRA.3.4.1, 2005.

[Ferrari, 2000] T. Ferrari, "End-To-End performance analysis with Traffic Aggregation," *Computer Networks,* vol. 34, pp. 905-914, Dec 2000 2000.

[Fowler, 1991] H. J. Fowler and W. E. Leland, "Local Area Network Traffic Characteristics, with Implications for Broadband Network Congestion Management," *IEEE JSAC,* vol. 9, pp. 1139-1149, September 1991 1991.

[Fu, 2003] H. Fu and E. W. Knightly, "A Simple Model of Real Time Flow Aggregation," in *IEEE/ACM Transactions on Networking*. vol. 11, 2003, pp. 422-435.

[Garcia, 2002] A. E. Garcia, K. D. Hackbarth, A. Brand, and R. Lehnert:, "Analytical Model for Voice over IP traffic characterization," *WSEAS Transactions on Communications,* vol. 1, pp. 59-65, 2002.

[Garroppo, 2001] R. G. Garroppo, S. Giordano, S. Niccolini, and F. Russo, "DiffServ Aggregation Strategies of Real Time Services in a WF2Q+ Schedulers Network," *Lecture Notes in Computer Sciences,* vol. 2170, pp. 481-491, 2001.

[Giorgi, 2008] G. Giorgi and C. Narduzzi, *Rate-interval curves - A tool for the analysis and monitoring of network traffic*: Elsevier, 2008.

[Girod, 2001] B. Girod, "Video Coding Standards," in *EE398 Image Communication II, 2000/2001 Course*, U. o. Stanford, Ed., 2001.

[Gleick, 1988] J. Gleick, *Caos : la creación de una ciencia*, 1988.

[GMBH, 2007] W.-C. GMBH, "Specification of the strategic network planning tool GSM-CONNECT for implementing the WIK-MNCM," Specification Report for the Australian Competition and Consumer Commission, 2007.

[Hao, 1998] F. Hao, I. Nikolaidis, and E. W. Zegura, "Efficient simulation of ATM networks with accurate end-to-end delaystatistics," in *IEEE International Conference on Communications (ICC 98)*, 1998, pp. 1799-1804.

[Hava, 2000] C. Hava, S. Holban, J. Murphy, and L. Murphy, "Initial Toll for Monitoring Performance of Web Servers," in *4th International Conference on Technical Informatics (CONTI'2000)*, Timisoara, Romania, 2000.

[Heckmann, 2002] O. Heckmann, J. Schmitt, and R. Steinmetz., "Multi-Period Resource Allocation at System Edges," in *10th International Conference on Telecommunication Systems Modelling and Analysis (ICTSM10)*, Monterey, USA, 2002, pp. 1-25.

[Heidemann, 1997] J. Heidemann, K. Obraczka, and J. Touch, "Modelling the Performance of HTTP Over Several Transport Protocols," *IEEE/ACM Transactions on Networking,* vol. 5, pp. 616-630, October 1997 1997.

[Heyman, 1996] D. P. Heyman and T. V. Lakshman, "Source Models for VBR Broadcast-Video Traffic," *IEEE/ACM Transactions on Networking,* vol. 4, pp. 40-48, February 1996 1996.

[Huffaker, 2002] B. Huffaker, M. Fomenkov, D. J. Plummer, D. Moore, and K. Claffy, "Distance Metrics in the Internet," in *IEEE International Telecommunications Symposium (ITS)*, 2002.

[Ishii, 2006] D. Ishii and S. Shioda, "Real-time bandwidth-requirement estimation using a queue simulation function," *Electronics and Communications in Japan (Part I: Communications),* vol. 89, pp. 75-87, Dec 2006 2006.

[ISO, 2007] ISO, "ISO 9000 - Quality management," 11 ed, 2007.

[ITU-T, 1988a] ITU-T, "Definiciones relativas a los servicios de telegrafía, de telemática y de transmisión de datos," in *Suplemento 1 a las Recomendaciones UIT-T de la serie F, Recommendation F.Sup1 (11/88)*, 1988a.

[ITU-T, 1988b] ITU-T, "H. ITU-T Recommendations: Audiovisual and Multimedia Systems," 1988b.

[ITU-T, 1996] ITU-T, "Asynchronous Transfer Mode (ATM) Management of the Network Element View," in *ITU-T Rec. 1.751*, ITU-T, Ed., 1996.

[ITU-T, 2002] ITU-T, "Network Planning," in *ITU/BDT-COE workshop* Bangkok, 2002.

[Izquierdo, 1999] M. R. Izquierdo and D. S. Reeves, "A Survey of Statischical Source Models for Variable Bit Rate Compressed Video," *Multimedia Systems,* vol. Springer Verlag, pp. 119-213, 1999.

[Jaeger, 2000] R. Jaeger, "Video & Interactive Internet Access in a DVB Network," in *1st European Conference on Universal Multiservice Networks (ECUMN'2000)* Colmar, France, 2000, pp. 439-445.

[Jagerman, 1998] D. L. Jagerman, B. Melamed, and W. Willinger, "Stochastic Modelling of Traffic Processes," Frontiers in queueing: models and applications in science and engineering ed: CRC Press, Inc, 1998, pp. 271-320.

[Jaising, 2002] R. Jaising, "Measurement Based Connection Admission Control." vol. Master Thesis on Computer Science: North Carolina State University, USA, 2002.

[Jelenkovic, 1996] P. R. Jelenkovic, A. A. Lazar, and N. Semret, "Multiple Time Scales and Subexponentiallity in MPEG video streams," in *International IFIP-IEEE Conference on Broadband Communications* Montreal, Canada, 1996.

[Jin, 1999] M. A. T. Jin, "Workload Characterization of the 1998 World Cup Web Site," Internet Systems and Applications Laboratory HP Laboratories Palo AltoSeptember 1999 1999.

[Kilpi, 2002] J. Kilpi and I. Norros, "Testing the Gaussian Approximation of Aggregate Traffic," in *2nd ACM/SIGCOMM Workshop on Internet Measurement Workshop*, Marseille, France, 2002, pp. 49-61.

[Kode, 2001] S. Kode, M. Nandwani, J. Maheswary, and S. Suresh, "Traffic Characterization for Heterogeneous Applications," Virginia Tech. University, Technical ReportMay 2001 2001.

[Korkmaz, 2000] T. Korkmaz and M. Krunz, "Source-Orientes Technology Aggragation with Multiple QoS Parameters in Hierarchical Networks," *ACM Transaction on Modelling and Computer Simulation,* vol. 10, pp. 295-325, October 2000 2000.

[Krunz, 1998] M. Krunz and A. Makowski, "Modeling Video Traffic Using M/G/infinity Input Processes: A Compromise Between Markovian and LRD Models," *IEEE Journal on Selected Areas in Communications,* vol. 16, pp. 733-748, June 1998 1998.

[Lee, 1995] E. W. Lee and J. W. Mark, "Capacity Allocation in Statistical Multiplexing of ATM Sources," *IEEE/ACM Transactions on Networking,* vol. 3, pp. 139-151, April 1995 1995.

[Leland, 1993a] W. E. Leland, M. S. Taqq, W. Willinger, and D. V. Wilson, "On the self-similar nature of {Ethernet} traffic," in *ACM SIGCOMM Conference on Communications Architectures*, San Francisco, California, 1993a, pp. 183-193.

[Leland, 1993b] W. E. Leland, M. S. Taqq, W. Willinguer, and D. V. Wilson, "On the self-similar nature of {Ethernet} traffic," in *ACM SIGCOMM Conference on Communications Architectures*, San Francisco, California, USA, 1993b.

[Leland, 1991] W. E. Leland and D. V. Wilson, "*High time-resolution measurement and analysis of LAN traffic: Implications for LAN interconnection,* ," in *IEEE INFOCOM '91*, 1991, pp. 1360-1366.

[Liebeherr, 2001] J. Liebeherr, S. Patek, and A. Burchard, "A Calculus for End-to-End Statistical Service Guarantees," University of Virginia, Charlottesville, USA CS-2001-19, June 2001 2001.

[Liu, 2004] N. X. Liu and J. S. Baras, "Long-Run Performance Analysis of a Multi-Scale TCP Traffic Model," *IEE Proceedings Communications,* vol. 151, pp. 251-257, June 2004 2004.

[Liu, 2003] X. Liu, "Network Capacity Allocation for Traffic with Time Priorities," *International Journal of Network Management,* vol. John Wiley&Sons Ltd., pp. 411-417, 2003.

[Liu, 1999] Z. Liu, N. Niclausse, C. Jalpa-Villanueva, and S. Barbier, "Traffic Model and Performance Evaluation of Web Servers," Institut National de Recherche en Informatique et en Automatique (INRIA) 3840, Dec. 1999 1999.

[Lolischkies, 2003] J. Lolischkies and J. Krumrei, "Next Generation Networks (NGN) From PSTN to IP-based NGN: Optimal Migration by Strategic Partnerships," DETECON, Opinion PaperOctober 2003 2003.

[Lombardo, 1998] A. Lombardo, G. Morabito, and G. Schembra, "An Accurate and Treatable Markov Model of MPEG-Video Traffic," in *IEEE INFOCOM'98*, 1998, pp. 217-224.

[Lombardo, 2004] A. Lombardo, G. Morabito, and G. Schembra, "A Novel Analytical Framework Compounding Statistical Traffic Modelling and Aggregate-Level Service Curve Disciplines: Network Performance and Efficiency Implications," *IEEE/ACM Transaction on Networking,* vol. 12, pp. 443-455, February 2004 2004.

[Mah, 1997] B. A. Mah, "An Empirical Model of HTTP Network Traffic," in *IEEE INFOCOM'97*, Kobe, Japan, 1997, pp. 592-600.

[Malaney, 1999] R. A. Malaney and G. Rogers, "Network Calculus and Service Curve Scherduling in Heterogeneous Networks," in *IEEE International Conference on Networks (ICON'99)*, Brisbane, Australia, 1999, pp. 250-256.

[Mandelbrot, 2004] B. B. Mandelbrot, G. Evertsz, and M. C. Gutzwiller, *Fractals and chaos : the mandelbrot set and beyond*. New York, 2004.

[Marot, 1999] M. Marot and G. Kotsis, "BISANTE Deliverable 1.1: Application and User Behaviour Characterisation," BISANTE ESPRIT Project EP28425, Deliverable BISANTE/DEL11, September 1999 1999.

[Mashat, 1997] A. Mashat and M. Kara, "Statistical Analysis and Modelling of MPEG Sources for Workload," in *V IFIP Workshop on Performance Modelling and Evaluation of ATM Networks*, 1997, pp. 16.01 - 16.10.

[Mata, 1994] J. Mata, S. Sallent, J. Balsells, J. Zamora, and A. v. d. Kolk, "Statistical Models for MPEG Video Standard," in *IEE EUSIPCO'94*, 1994, pp. 624-627.

[Mavrogeanes, 2001] R. Mavrogeanes, "Introduction to MPEG Video: For the Digital Network Professional," 2001.

[Mondragon, 2001] R. J. Mondragon, D. K. Arrowsmith, J. M. Griffiths, and J. M. Pitts, "Chaotic Maps for Network Control: Traffic Modelling and Queuing Performance Analysis," *Performance Evaluation,* vol. 43, pp. 223-240, 2001.

[Mondragon, 1999] R. L. Mondragon, "A Model of Packet Traffic Using a Random Wall Model," *Journal of Bifurcation and Chaos,* vol. 9, pp. 1381-1392, July 1999 1999.

[Morisse, 2001] K. Morisse, "Principles and Standards for Video Coding," in *Socrates IP Summerschool* Chania - Kreta, Greece, 2001.

[MPEG_Group, 2007] MPEG_Group, "MPEG Standards," The Moving Picture Experts Group (MPEG) Website, 2007.

[Muntean, 2001] C. H. Muntean, J. McManis, and J. Murphy, "The Influence of Web Page Images on the Performance of Web Servers," in *Lecture Notes in Computer Science 2093*, Springer-Verlag, Ed., 2001, pp. 821-828.

[Nahas, 2002] M. Nahas, "Stochastic bounds using effective capacity," V. University, Ed., 2002.

[Navickas, 2006] A. Ž. Z. Navickas and R. Rindzevičius, "Bursty Traffic Simulation by ON - OFF Model," *ELEKTRONIKA IR ELEKTROTECHNIKA - ELECTRONICS AND*

ELECTRICAL ENGINEERING, vol. 6, pp. 68-73, 2006.

[Ni, 1996] J. Ni, T. Yang, and D. H. K. Tsang, "Source Modelling, Queuing Analysis and Bandwidth Allocation for VBR MPEG-2 Video Traffic in ATM Networks," *IEE Proceedings on Communications,* vol. 143, pp. 197-205, 1996.

[Norros, 1996] I. Norros and P. Pruthi, "On the Applicability of Gaussian Traffic Models," in *13th Nordic Teletraffic Seminar*, 1996, pp. 37-50.

[Nurminen, 2003a] J. K. Nurminen, "Models and Algorithms for Network Planning Tools - Practical Experiences," Systems Analysis Laboratory Helsinki University of Technology, Helsinki, Series E - Electronic Reports E14, May 2003 2003a.

[Nurminen, 2003b] J. K. Nurminen, "Modelling and Implementation Issues in Circuit and Network Planning Tools," in *Engineering Physics and Mathematics* Espoo, Finland: Helsinki University, 2003b, p. 41.

[Oottamakom, 2001] C. Oottamakom and D. Bushmitch, "A DiffServ Measurement-Based Admission Control Utilizing Effective Envelopes and Services Curves," in *IEEE International Conference on Communications (ICC 2001)*, Helsinki, Finland, 2001, pp. 1187-1195.

[Pandit, 2002] K. Pandit, J. Schmitt, and R. Steinmetz., "Aggregation of Heterogeneous Real-Time Flows with Statistical Guarantees," in *International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS'02)*, San Diego, USA, 2002, pp. 57-64.

[Parkinson, 2002] R. Parkinson, "Traffic Engineering Techniques in Telecommunications." vol. 2005: Infotel System Corporation, 2002.

[Paxson, 1994a] V. Paxson, "*Empirically-Derived Analytic Models of Wide-Area TCP Connections*," *IEEE/ACM Transactions on Networking,* vol. 2, pp. 316-336, August 1994 1994a.

[Paxson, 1994b] V. Paxson, "*Growth Trends in Wide-Area TCP Connections* " *IEEE Network,* vol. 8, pp. 8-17, July 1994 1994b.

[Paxson, 1995] V. Paxson and S. Floyd, "Wide-Area Traffic: The Failure of Poisson Modelling," *IEEE/ACM Transactions on Networking,* vol. 3, pp. 226-244, June 1995 1995.

[Pieda, 1999] P. Pieda, N. Seddigh, and B. Nandy, "The Dynamics of TCP and UDP Interaction in IP QoS Differentiated Services Networks," in *3rd Canadian Conference on Broadband Research*, Ottawa, Canada, 1999.

[Pruthi, 1995] P. Pruthi and A. Erramilli, "Heavy tailed on/off source behaviour and self-similar traffic," in *ICC '95*, Seattle, Washington, 1995, pp. 445-450.

[Riedl, 2000] A. Riedl, M. Perske, T. Bauschert, and A. Probst, "Dimensioning of IP Access Networks with Elastic Traffic," in *Networks 2000*, Toronto, Canada, 2000.

[Robert, 1997] S. Robert and J.-Y. L. Boudec, "New Models for Pseudo Self-Similar Traffic," *Performance Evaluation Journal,* vol. 30, pp. 57-68, 1997.

[Rolland, 2006] C. Rolland, J. Ridoux, and B. Baynat, "Hierarchical Models for Web Traffic on CDMA-1xRTT Networks," Technical Report University of Pierre et Marie Curie, France2006.

[Rose, 1997] O. Rose, "Simple and Efficient Models for Variable Bit Rate MPEG Video Traffic," *Performance Evaluation,* vol. 30, pp. 69-85, July 1997 1997.

[Rose, 1994] O. Rose and M. Frater, "A Comparison of Models for VBR Video Traffic Sources in B-ISDN," in *Broadband Communication*, ISBN 0-444-81834-0 ed, 1994, pp. 275-287.

[Rubin, 1995] I. Rubin, K. K. Chang, and G. Forcina, "Performance Analysis of Load-Adaptive/TDMA Systems with Applications to Integrated- Services Satellite Networks," *International Journal on Satellite Communications,* vol. 13, pp. 427-440, 1995.

[Samuel, 1999] L. G. Samuel, "The Application of Non-Linear Dynamics to Teletraffic Modelling," in *Department of Electronic Engineering* London, UK: Queen Mary and Westfield College, University of London, 1999, p. 165.

[Sarvotham, 2001] S. Sarvotham, R. Riedi, and R. Baraniuk, "Connection-level Analysis and Modelling of Network Traffic," in *ACM SIGCOMM Internet Measurement Workshop*, San Francisco, USA, 2001, pp. 99-103.

[Schmitt, 1999] J. Schmitt, M. Karsten, and R. Steinmetz, "Aggregation of Guaranteed Service Flows," in *7th IEEE/IFIP International Workshop on Quality of Service (IWQoS'99)*, London, UK, 1999, pp. 147-155.

[Schmitt, 2001] J. Schmitt, M. Karsten, and R. Steinmetz, "On the Aggregation of Deterministic Service Flows," *Computer Communications,* vol. 24, pp. 2-18, January 2001 2001.

[Schwefel, 1999] H. P. Schwefel and L. Lipsky, "Analytic models of traffic in telecommunication systems, based on multiple ON-OFF sources with self-similar behavior," in *16th International Teletraffic Congress, ITC-16*, Edimburg, UK, 1999.

[Sharafeddine, 2003] S. Sharafeddine, A. Riedl, J. Glasmann, and J. Totzke, "On Traffic Characteristics and Bandwidth Requirements of Voice over IP Applications," in *8th IEEE International Symposium on Computers and Communications*, Kemer-Antalya, Turkey, 2003, pp. 1324-1332.

[Sheluhin, 2007] O. I. Sheluhin, S. M. Smolskiy, and A. V. Osin, *Self-similar Processes in Telecommunications*: Wiley, 2007.

[Sheng, 2001] M. Sheng and J. Chuanyi, "Modeling heterogeneous network traffic in wavelet domain," *IEEE/ACM Transactions on Networking,* vol. 9, pp. 634-649, 2001.

[Shenker, 1995] S. Shenker, "Service Models and Pricing Policies for an Integrated Services Internet," in *Public access to the Internet* Cambridge, USA: MIT Press, 1995.

[Sivaraman, 2000] V. Sivaraman and F. Chiusi, "Providing End-to-End Statistical Delay Guarantees with Earliest Deadline First Scheduling and Per-Hop Traffic Shaping," in *IEEE INFOCOM 2000*, Tel Aviv, Israel, 2000, pp. 631-640.

[Sohraby, 1993] K. Sohraby, "On the theory of general on-off sources with applications in high-speed networks," in *IEEE INFOCOM '93*, 1993, pp. 401-410.

[Sousa, 2000] J. M. d. Sousa, "Planning Multi-Service Networks," in *ITS 2000*, Buenos Aires, Argentina, 2000.

[Staehle, 2003] D. Staehle, K. Leibnitz, and K. Tsipotis, "QoS of Internet Access with GPRS," *ACM Wireless Networks,* vol. 9, pp. 213-222, May 2003 2003.

[Starobinski, 2000] D. Starobinski, "Stochastically Bounded Burstiness for Communication Networks," *IEEE TRANSACTIONS ON INFORMATION THEORY,* vol. 46, pp. 206-212, January 2000 2000.

[Starobinski, 2003] D. Starobinski, M. Karpovski, and L. A. Zakrevski, "Application of Network Calculus to General Topologies Usin g Turn-Prohibition," *IEEE/ACM*

Transactions on Networking, vol. 11, pp. 411-421, June 2003 2003.

[Tanenbaum, 2003] A. S. Tanenbaum, *Computer Networks*, 4 ed., 2003.

[Tang, 1999] P. P. Tang and T.-Y. C. Tai, "Network Traffic Characterization Using Token Bucket Model," in *IEEE INFOCOM'99*, New York, 1999, pp. 51-62.

[Taqqu, 1997] M. Taqqu, W. Willinger, and R. Sherman, "Proof of a fundamental result in self-similar traffic modeling," *Computer Communications,* vol. 26, pp. 5-23, 1997.

[Telefónica, 2007a] F. Telefónica, "La Sociedad de la Información en España 2007," 2007a.

[Telefónica, 2007b] F. Telefónica, "Libro Blanco del Hogar Digital y las Infraestructuras Comunes de Telecomunicaciones," Publicaciones de Telefónica 2007b.

[Touch, 1998] J. Touch, J. Heidemann, and K. Obraczka, "Analysis of HTTP Performance," University of Southern California /ISI, Research Report 98-463, Dec 1998 1998.

[Turaga, 2000a] D. Turaga and T. Chen, "Activity-adaptive modelling of Dynamic Multimedia Traffic," in *IEEE International Conference on Multimedia*, New York, 2000a.

[Turaga, 2000b] D. Turaga and T. Chen, "Fundamentals of Video Coding: H.263 as an example," in *Compressed Video Over Networks: The Signal Processing Series*, M. Dekker, Ed., 2000b.

[Turaga, 2000c] D. Turaga and T. Chen, "Modeling of Dynamic Video Traffic," in *International Symposium on Circuits and Systems (ISCAS'2000)*, Geneva, Switzerland, 2000c.

[Turaga, 2001] D. Turaga and T. Chen, "Hierarchical Modeling of Variable Bit Rate Video Sources," in *International Packet Video Workshop 2001*, Kyongju, Korea, 2001.

[UIT, 1993] UIT, "Términos y Definiciones de Ingeniería de Tráfico," in *Recomendación UIT-T E.600*, 1993.

[Vigo, 2000] E. Vigo, "Next Generation Networks(NGN): Planificación de Redes Basadas en IP," D.-D. D. Consultant, Ed., 2000.

[Vutukury, 2003] S. Vutukury and J. J. G. Luna-Aceves, "WiNN: An Efficient Method for Routing Short-Lived Flows," in *10th International Conference on Telecommunications (ICT'2003)*, Tahiti, 2003, pp. 1008-1013.

[Wang, 1998] S. Wang, H. Zheng, and J. A. Copeland, "Video Multiplexing with QoS Constraints," in *IEEE SPIE Conference on Internet Routing and QoS*, 1998, pp. 81-91.

[Wilson, 1994] W. E. L. M. T. W. W. D. Wilson, "On the Self-Similar Nature of Ethernet Traffic (Extended Version)," *IEEE/ACM Transactions on Networking,* vol. 2, pp. 1-15, February 1994 1994.

[Wu, 2003] K. Wu and D. S. Reeves, "Link Dimensioning and LSP Optimization for MPLS Networks Supporting DiffServ EF and BE Traffic Classes," in *18th International Teletraffic Congress*, Berlin, Germany, 2003.

[Wu, 2004] K. Wu and D. S. Reeves, "Capacity Planning of DiffServ Networks with Best-Effort and Expedited Forwarding Traffic," *Telecommunication System,* vol. 25, pp. 193-207, March 2004 2004.