UNIVERSITAT POLITÈCNICA DE CATALUNYA
DEPARTAMENT D'ARQUITECTURA DE COMPUTADORS

MODELOS ANALITICOS PARA LA EVALUACION DE MECANISMOS DE CONTROL DE TRAFICO EN REDES ATM

TESIS DOCTORAL

Autor: Jorge García Vidal
Directora: Olga Casals Torres

Barcelona, Enero 1992

Este trabajo está dedicado a mi familia:

Mamá, papá, Olga, María José, Manolo, Cristina y Sonia.

Indice

Introducción

En este trabajo se presentan una serie de modelos analíticos que son de utilidad para la evaluación de mecanismos de control del tráfico en redes digitales de servicios integrados de banda ancha que usan el modo de transferencia asíncrono (redes ATM, 'Asynchronous Transfer Mode').

Una red ATM es una red en conmutación de paquetes en donde los paquetes, que se denominan 'celdas' o 'células', son de un tamaño fijo y reducido y los protocolos de niveles bajos se han simplificado al máximo para conseguir elevadas velocidades de transmisión. Al usar multiplexación asíncrona tenemos la flexibilidad necesaria para soportar servicios de características dispares. Además, al ser una red de muy alta velocidad podemos incluir servicios de banda ancha y servicios interactivos en tiempo real.

Salvo que se tome alguna medida de control, es fácil que en una red ATM se produzcan situaciones de congestión: el acceso del usuario se produce a gran velocidad, el rango de velocidades de transmisión es muy amplio y las fuentes actúan con gran autonomía. Estos problemas de congestión también se presentan en las redes de conmutación de paquetes convencionales, pero la diferencia en las velocidades de transmisión hace que los métodos que son útiles para controlar la congestión en las redes de transmisión de datos dejen de serlo para el caso de redes ATM.

En este tipo de redes las medidas deben ser preventivas, por lo que la mayoría de esquemas de control de congestión que han sido propuestos se basan en el control de admisión de nuevas conexiones (CAC, 'Conexión Acceptance Control'): Cuando un nuevo usuario quiere establecer una comunicación hace una descripción de las características del tráfico que va a generar. El CAC debe decidir entonces si el nuevo usuario va o no a causar un nivel de congestión excesivo. Aparece, pues, la necesidad de proveer un mecanismo mediante el cual se controle si el tráfico que es emitido por los usuarios de la red corresponde a la descripción que han hecho del mismo. Es lo que se conoce como mecanismo de función de policía o de vigilancia ('Policing Function').

No es realista pensar que exista un mecanismo que pueda eliminar por completo la posibilidad de la aparición de fenómenos de congestión, a no ser que se hiciera un uso muy ineficiente de los recursos de la red. La aparición de una congestión tiene como consecuencia el aumento del retardo que sufre una celda al atravesar la red, y la posibilidad de pérdidas de celdas debidas al desbordamiento en los conmutadores. En las especificaciones de la red deben estar incluidos valores máximos para estos dos parámetros.

En los servicios interactivos en tiempo real una celda tiene fuertes restricciones en el retardo que puede sufrir al atravesar la red. Este hecho ha sido clave para impedir la introducción de ese tipo de servicios en redes de conmutación de paquetes de baja velocidad. Sin embargo la cuestión deja de ser crítica en el caso de redes ATM: la velocidad de transmisión es tan alta que el retardo de propagación a través de la red, que es fijo e incompresible, es mucho mayor que los retardos variables debidos al almacenamiento de la celda en los buffers de los conmutadores, de forma que en general es fácil satisfacer los requerimientos que los usuarios hacen en cuanto al retardo.

La probabilidad máxima de pérdida de celdas requerida por algunos usuarios es extremadamente pequeña (del orden de 1.0 E-10) y necesita un tratamiento más cuidadoso. Una red ATM trabaja a una velocidad muy elevada por lo que es fundamental que el sistema sea tan simple como sea posible. Por ello, en principio, es conveniente usar un único servicio portador para todos los usuarios. Ahora bien, dada la diferencia entre lo requerido por

servicios como la telefonía y la transmisión de datos, que representan una parte muy importante del tráfico total, y lo requerido por servicios como la transmisión de imagen o las señalizaciones internas, la utilización de un único servicio portador, que debería satisfacer las demandas de los servicios más restrictivos, puede suponer un uso muy ineficiente de los recursos de la red.

Una solución a este problema es definir más servicios portadores, de tal modo que aseguremos diferentes valores máximos en la probabilidad de pérdida de celda. Por ejemplo podríamos tener un servicio de alta calidad con una probabilidad de pérdida máxima muy baja y otro servicio portador de mediana calidad con una probabilidad de pérdida máxima de un valor mayor.

Hay varias formas de poder discriminar en los valores máximos de probabilidad de pérdida. Podríamos asignar cada servicio portador a un conjunto distinto de multiplexores con diferentes niveles de carga. Aunque esta posibilidad es de realización simple, tenemos poca flexibilidad en lo que respecta a la gestión de los niveles de prioridad (por ejemplo, para asegurar la secuencia de llegadas de las celdas, todas las celdas de una conexión deberían estar marcadas con el mismo nivel de prioridad). Parece más conveniente tener la posibilidad de marcar cada celda, de forma independiente, con un determinado nivel de prioridad. En este caso los conmutadores deben introducir algún mecanismo de prioridad que discrimine entre los niveles de probabilidad de pérdida de cada celda dependiendo del tipo de servicio portador al que pertenezca. Este tipo de mecanismos se conocen como de prioridad espacial o de pérdida ('Space priority', 'Loss priority').

Otro punto importante en el estudio de las redes ATM es el problema de su evaluación. Los modelos analíticos que se usan en las redes de comunicación convencionales, basados principalmente en el uso de un proceso de Poisson para caracterizar el tráfico, dejan ahora de ser útiles, pues el tráfico tiene características especiales que hacen que las predicciones hechas usando procesos de Poisson se alejen mucho de los valores reales. El problema de la evaluación se hace más complicado (también más interesante) y se deben usar nuevas técnicas de modelización.

De entre ellas las más importantes son aquellas que modelan los diferentes niveles de actividad del tráfico mediante los estados de una cadena de Markov. El análisis del sistema, que suele ser una cola con tiempo de servicio determinista, se puede llevar a cabo mediante un nuevo proceso de Markov (por ejemplo, considerando el sistema en los instantes de salida de un cliente). Para la solución de estos modelos se han usado varias técnicas, entre las que destacan las basadas en métodos analíticos matriciales y en la aproximación de fluido.

En los anteriores modelos surge a menudo un problema de dimensiones, pues si queremos estudiar un sistema realista el número de estados se hace muy elevado. Una forma de eludir este problema es recurrir a una solución aproximada, estudiando, por ejemplo, el comportamiento asintótico del sistema cuando la longitud de cola tiende a infinito. Es de esperar que esta aproximación sea tanto mejor cuanto mayor sea la longitud de cola y menor la probabilidad de pérdida. Esta situación es, de hecho, la que habitualmente aparece en las aplicaciones y, por otra parte, es la que presenta mayores problemas en su solución exacta.

Las principales aportaciones de este trabajo son las siguientes:

Se presenta un modelo aproximado del comportamiento asintótico de la probabilidad de pérdida en una cola de capacidad finita cuando el proceso de entrada es un D-BMAP ('Discrete time Batch Markovian Arrival Process'). Un caso particular de este tipo de proceso es la superposición del tráfico generado por fuentes independientes de dos estados y un tráfico que siga una secuencia periódica de emisión de celdas. Este modelo es usado para evaluar el efecto de la aparición de 'clusters' de celdas que escapan al control de la función de policía cuando el mecanismo usado es de tipo 'pick-up' (por ejemplo, el mecanismo 'Leaky Bucket'). Este fenómeno había sido estudiado anteriormente mediante simulación, por lo que las probabilidades de pérdida que se habían podido medir estaban muy por encima de los valores requeridos en el caso real. Nuestro modelo permite estudiar lo que pasa cuando las probabilidades de pérdidas toman valores muy bajos y con un tráfico de entrada realista.

Se desarrolla un modelo exacto para el mecanismo de prioridad espacial conocido como 'Push-out' cuando el tráfico de entrada es un MAP. Se propone el uso de un umbral y se evalúa el efecto que dicho umbral tendría sobre el rendimiento del sistema. También se desarrollan dos modelos para el mecanismo de prioridad espacial conocido como 'Partial buffer sharing'. En uno de estos modelos el tráfico de entrada es un MAP. El segundo modelo se basa en la aproximación de fluido.

Usando los modelos anteriores se estudian los beneficios derivados del uso de mecanismos de prioridad espacial y se hace una comparación entre ambos mecanismos. Los trabajos que se habían hecho con anterioridad modelaban el tráfico de entrada mediante un proceso de Poisson, por lo que no se podían realizar estudios en condiciones realistas para una red ATM.

La organización en capítulos es la siguiente:

En el capítulo 1 se hace una descripción de los aspectos fundamentales en torno al diseño de redes integradas de alta velocidad. El capítulo 2 hace un repaso de los métodos más importantes de evaluación que se usan en el ámbito de redes ATM.

El capítulo 3 presenta un estudio de las propiedades asintóticas de la probabilidad de pérdida de la cola D-BMAP/D/1. Tenemos como caso particular la multiplexación de fuentes periódicas y fuentes de frecuencia de emisión variable. En el capítulo 4 se abordan los principales problemas que presenta el control de la congestión en una red ATM y se usa el modelo anterior para estudiar la eficiencia del mecanismo de 'Leaky bucket' cuando un usuario envía 'clusters' de celdas.

El capítulo 5 trata del problema de la calidad de servicio y de la introducción de prioridades en una red ATM. En el capítulo 6 se desarrollan varios modelos exactos para dos mecanismos de prioridad espacial, el mecanismo de 'Push-out' y el de 'Partial buffer sharing'. Usando estos modelo estudiamos los beneficios derivados del uso de prioridades y hacemos una comparación entre ambos mecanismos.

En el último capítulo se presentan las principales conclusiones que se sacan de este trabajo y las futuras líneas de investigación que deja abiertas.

Finalmente quería expresar mi más sincero agradecimiento a todas aquellas personas que me han ayudado en la realización de este trabajo. En especial a mi directora de tesis, Olga Casals, por su constante apoyo y orientación. Gracias, Olga. También a José María Barceló y Camil Giralt, que hicieron largos programas de simulación que ayudaron a verificar los modelos que se presentan. A Jordi Domingo y Josep Solé, que me hablaron por primera vez del apasionante tema de redes de banda ancha, y a todos los compañeros del Departament d'Arquitectura de Computadors por prestarme su ayuda siempre que la he necesitado, en especial a José María Cela y a Álvaro Suárez por brindarme su amistad.

Capítulo 1

Redes digitales de servicios integrados de banda ancha

1.1 Introducción [Hui90], [Tob90], [PerFle87]

Un enlace de comunicaciones proporciona posibilidad de comunicación entre sólo dos puntos. Si queremos extender esta capacidad de comunicación a un gran número de usuarios distribuidos geográficamente, se hace obvia la necesidad de construir sistemas con capacidad de conmutación. Además, si queremos compartir los costes de los sistemas de transmisión y conmutación entre estos usuarios, debemos usar algún esquema de multiplexación.

El desarrollo de la tecnología de componentes determina el esquema de multiplexación y el balance entre la conmutación y la transmisión empleada. En los últimos años se han producido avances significativos en las áreas de las tecnologías de comunicaciones y de computadores. En el caso de las tecnologías de comunicaciones, estos avances han sido especialmente importantes en las áreas de los dispositivos de transmisión y de conmutación y han hecho aparecer la posibilidad de introducir nuevos servicios de telecomunicación. Con el desarrollo de la tecnologías de la fibra óptica y del láser, la velocidad de transmisión se ha incrementado de una forma exponencial durante los años 80. En la actualidad se están desarrollando sistemas de transmisión fiables que funcionan a velocidades de Gbps, y se espera que estas velocidades puedan ser incrementadas a decenas o centenares de Gbps mediante el uso de otras tecnologías aún en desarrollo. Estas nuevas tecnologías y servicios se denominan de banda ancha ya que las velocidades de transmisión involucradas están en el rango de 1 Mbps a 100 Mbps y mayores. Es de esperar que su desarrollo introduzca cambios radicales en el diseño de las redes de telecomunicación.

A lo largo de este siglo se han producido grandes cambios en el campo de los sistemas de comunicación. Las primeras centrales de conmutación eran manuales, en donde la función de interconexión era manejada por un operador. Las centrales manuales fueron sustituidas por el conmutador mecánico automático inventado por Strower. Estos medios mecánicos de proveer interconexiones fueron a su vez reemplazados por los conmutadores electrónicos de barras cruzadas.

En paralelo con la evolución de las redes de interconexión, el desarrollo de los sistemas de control programables proporcionó un control y utilización eficiente de los sistemas de conmutación. Todo ello posibilitó la aparición de conmutadores electrónicos de gran tamaño.

Estos primeros conmutadores eran Conmutadores por División de Espacio ('Space Division Switches', SDS). En ellos una comunicación telefónica usaba de forma exclusiva un camino dentro de la red de interconexión durante toda la duración de la llamada. El mecanismo de conmutación se basaba en realizar un contacto entre conductores, ya fuera mecánica o electrónicamente.

Casi en paralelo con el desarrollo de la tecnología de conmutación, las tecnologías de transmisión a larga distancia evolucionaron con el desarrollo de amplificadores analógicos fiables y de la transmisión por microondas o por cables coaxiales. Las señales analógicas de voz eran multiplexadas sobre estos medios de transmisión de gran capacidad mediante Multiplexación por División de Frecuencia (FDM, 'Frecuency Division Multiplexing'). Una vez multiplexadas, estas señales podían ser amplificadas y transmitidas a grandes distancias a un bajo coste.

Un cambio tecnológico fundamental se dio con el desarrollo de la tecnología digital, que representa las señales en un formato binario. Dado este tipo de representación, las diferentes señales son multiplexadas en el dominio del tiempo para formar una única secuencia de bits

que puede ser transmitida a través de un único canal. Esta forma de compartir el medio de transmisión se conoce como Multiplexación por División de Tiempo ('Time Division Multiplexing', TDM), y es la base en la que se apoyan las redes integradas de banda ancha.

Una de las mayores ventajas del uso de TDM sobre el uso de FDM es que el coste de multiplexación por señal decrece cuando la velocidad de multiplexación aumenta. Con los bajos costes y aumento de velocidad de los dispositivos electrónicos digitales, el uso de TDM se ha hecho más rentable que el de FDM. Por lo tanto TDM se ha ido adoptando para los niveles jerárquicos inferiores, aunque FDM predomina para la transmisión a larga distancia, debido al menor coste de transmisión de las señales analógicas.

De hecho, TDM es más que un mecanismo de multiplexación y transmisión. También se puede usar como un mecanismo para conmutar señales, dando lugar a la Conmutación por División de Tiempo ('Time Division Switching', TDS): Diferentes terminales o interfaces distribuidos a lo largo de una línea de transmisión pueden multiplexar su información en el tiempo, y seleccionar la información de la línea que proceda de un determinado terminal o interfaz, de forma que tenemos una serie de comunicaciones conmutadas.

Además de las ventajas de TDM sobre FDM, hay otros beneficios asociados con las comunicaciones digitales, tales como la regeneración fiable de las señales y la facilidad para el procesado de señales digitales.

Los terminales de datos se comunican intercambiando señales binarias, por lo que una red que deba soportarlos debería ser digital. Sin embargo la red telefónica fue originalmente diseñada para la transmisión de telefonía analógica y no se adapta bien a las características de las comunicaciones de datos. Este hecho ha provocado la aparición de redes específicas para la transmisión de datos.

La conmutación de circuitos usada tradicionalmente en la red telefónica tampoco es adecuada para la comunicación de datos, por lo que se desarrolló otro tipo de conmutación, la conmutación de paquetes.

En conmutación de circuitos, se establece un camino de enlaces conectados entre el origen y el destino en el instante en que se establece la comunicación, que permanece dedicado hasta que la comunicación finaliza. El establecimiento de este camino se realiza mediante señalizaciones que circulan por la red desde el origen hasta el destino. Una vez se ha establecido dicho camino, una señal de retorno informa al usuario que origina la comunicación que puede empezar la transmisión. A partir de este momento los conmutadores son virtualmente transparentes a la información transmitida. De hecho todo sucede como si los dos usuarios tuvieran un circuito continuo que los conectara durante la conversación.

En las redes de conmutación de paquetes, la información se transmite en bloques de datos, conocidos como 'paquetes'. En un paquete, además de la información que el usuario quiere transmitir, se añade una cabecera que contiene información de control, tal como la dirección de origen y destino, tipo de mensaje, etc. El paquete se transmite de conmutador a conmutador, siguiendo una ruta puede ser fija a lo largo de una comunicación (circuito virtual) o distinta para cada paquete (datagrama). Cada uno de estos conmutadores tiene una cierta capacidad de procesamiento y de almacenamiento, necesaria para resolver las colisiones que ocurren cuando dos paquetes quieren usar simultáneamente el mismo enlace de salida. De esta forma, no hay un camino completamente dedicado a una comunicación, y se ocupan los recursos de la red sólo cuando se quiere transmitir información.

La conmutación de circuitos se adapta bien a las necesidades de las redes telefónicas convencionales: es rentable cuando existe un flujo continuo de información una vez se establece la comunicación. En el caso de la transmisión de datos, la información tiende a ser transmitida a ráfagas, es decir, los usuarios hacen un uso relativamente escaso de los recursos de la red. Sin embargo cuando utilizan el canal requieren una rápida respuesta. Si cada vez que tenemos una ráfaga de información debemos generar todas las señalizaciones requeridas para el establecimiento de una comunicación en una red en conmutación de circuitos, tendremos retardos en el establecimiento demasiado elevados. Si el circuito se mantiene durante toda la comunicación, la utilización del mismo será demasiado baja.

Las redes de transmisión de video actuales no tienen casi capacidad de conmutación, y se limitan a redes de radiodifusión o CATV.

1.2 Los nuevos servicios de la red de banda ancha [Min89], [Hui90]]

Es inevitable que los componentes de las nuevas tecnologías de transmisión por fibra óptica sean más caros que los de otras tecnologías más maduras, como la del cable de cobre. Sin embargo, su coste se reduce extremadamente cuando son compartidos por muchas llamadas. Por lo tanto, no sería económico reemplazar los antiguos componentes por los nuevos si no se introducen servicios que permitan amortizar la nueva capacidad de ancho de banda.

Estos servicios de banda ancha son necesitados por nuestra sociedad, en donde son cada vez más importantes la transmisión de información y los medios audiovisuales. La disposición de computadores personales facilita el acceso, almacenamiento y manipulación de datos, haciendo necesario la disponibilidad de redes fiables de transmisión de datos. El intercambio de documentos mediante imágenes y el uso de terminales de alta resolución proporciona un modo más natural de interacción entre personas que el que se puede establecer mediante sólo la voz y los datos. Las videoconferencias reducen la necesidad de viajes a larga distancia para participar en reuniones. La introducción gradual de servicios de video de alta definición requiere recursos de comunicación que no pueden ser proporcionados por el sobrecargado espectro radioeléctrico.

Como consecuencia la red de banda ancha debe ser capaz de soportar:

- *Servicios interactivos y de distribución*: La red debe servir como un medio de transporte común a comunicaciones interactivas y no interactivas que pueden incluir audio, video y datos. Los servicios interactivos pueden incluir servicios conversacionales (ejemplo: videotelefonía, videoconferencia, transmisión de datos a alta velocidad), servicios de mensajería (por ejemplo: servicios de tratamiento de mensajes, servicios de correo electrónico para imágenes en movimiento, imágenes de alta resolución e información audio) y servicios de consulta (por ejemplo: servicios de consulta para películas, imágenes de alta resolución, información audio e información de archivos). Los servicios de distribución pueden ser con o sin control de la presentación por parte del usuarios (por ejemplo: radiodifusión de programas de televisión y audio).

- *Tráfico continuo y a ráfagas*: Algunas fuentes proporcionan información a una velocidad constante (por ejemplo: voz o imagen codificados en PCM), mientras que otras fuentes pueden presentar velocidades de transmisión variable (por ejemplo: voz codificada con DSI, video con codificadores VBR, datos).

- *Velocidades de transmisión que pueden estar dentro de un amplio rango*: Las velocidades de transmisión pueden ir de pocos Kbps para servicios de telemetría a cientos de Mbps para video de alta definición.

- *Servicios orientados a la conexión y servicios no orientados a la conexión*: Algunos servicios, por ejemplo, telefonía convencional, videotelefonía, etc, tienen separadas las fases de establecimiento de la conexión, de transmisión de información y de desconexión. Son los servicios orientados a la conexión. Otros servicios, por ejemplo correo electrónico, no separan dichas fases de establecimiento de la conexión y transmisión de la información.

- *Comunicaciones punto a punto y comunicaciones más complejas*: Muchas aplicaciones de banda ancha necesitan conexiones paralelas entre puntos, conexiones entre múltiples usuarios o incluso combinaciones de múltiples conexiones entre múltiples usuarios.

1.3 Las redes de servicios integradas [Hui90]

Tradicionalmente, los servicios mencionados anteriormente son soportados por redes separadas: La voz por la red telefónica conmutada, los datos sobre redes transmisión de datos o redes locales, las videoconferencias sobre redes privadas y la televisión sobre redes de

radiodifusión o redes de transmisión por cable. Cada una de estas redes está diseñada para soportar una aplicación específica y no es adecuada para otras aplicaciones.

Por ejemplo, tal como se ha dicho, la red telefónica tradicional, que usa conmutación de circuitos, es demasiado ruidosa e ineficiente para comunicaciones de datos a ráfagas. Las redes de datos de conmutación de paquetes suelen tener una conectividad limitada, usualmente no disponen del suficiente ancho de banda para transmitir voz y video digitalizados y sufren retrasos inaceptables para servicios en tiempo real. Las redes de radiodifusión de imágenes no suelen tener capacidad de conmutación, o ésta es muy limitada.

A menudo es conveniente tener una única red que sea capaz de soportar todos estos servicios, de forma que se puedan conseguir economías de escala. Esta idea de ahorro ha llevado a la consideración de redes de servicios integrados. La integración evita la necesidad de tener muchas redes que se solapen entre sí, lo que simplifica la gestión de la red y reduce la inflexibilidad en la evolución e introducción de nuevos servicios. La integración de los servicios de banda ancha es posible gracias a los avances en la tecnologías de banda ancha y de procesado de información a altas velocidades.

Sin embargo, la integración de una red puede tener diferentes significados, dependiendo de la parte y de la función de la red que se considere:

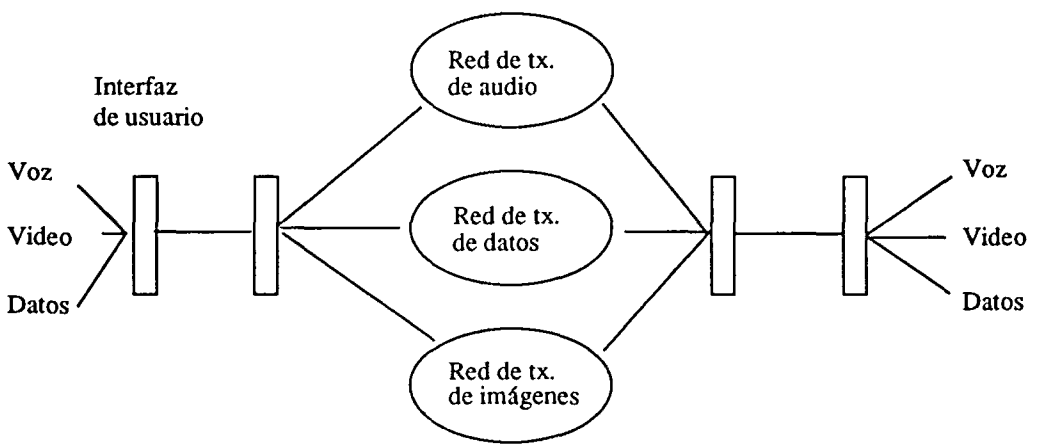


Figura 1.1: Red con acceso integrado

Acceso integrado

El acceso integrado supone que diferentes servicios de un usuario comparten un único interfaz a un único medio de transmisión en el bucle de abonado (figura 1.1). El acceso integrado a la red debería proporcionar la multiplexación flexible de tantos servicios como sea posible. En el caso de servicios de banda ancha, la fibra óptica es un medio aplicable a todos estos servicios, debido a sus excelentes propiedades en la transmisión y a su gran ancho de banda. Dado el abundante ancho de banda de la transmisión óptica, queda pendiente determinar cómo se puede multiplexar de forma flexible y a través de un único interfaz el amplio rango de velocidades de transmisión.

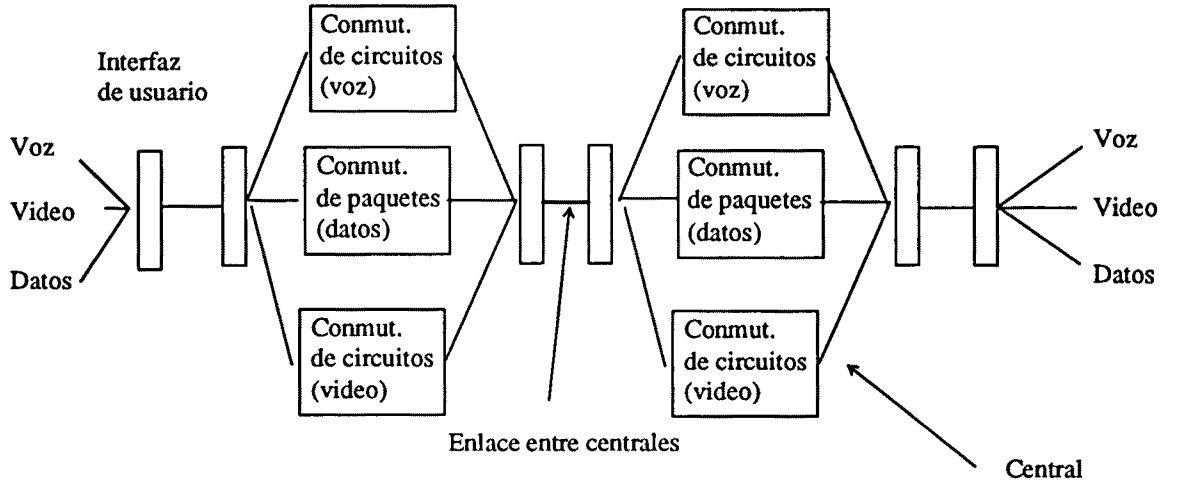


Figura 1.2: Red con transporte integrado.

Transporte integrado

El transporte integrado supone que diferentes servicios de diferentes usuarios comparten los enlaces de gran capacidad de la red (figura 1.2). El transporte integrado evita la segregación de diferentes tipos de tráfico y medios en diferentes enlaces de transmisión.

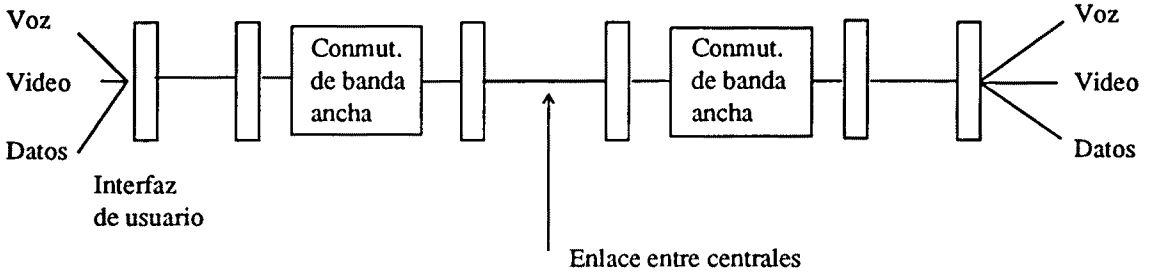


Figura 1.3: Red con conmutación integrada

Conmutación integrada

La conmutación integrada implica que servicios de diferentes velocidades y características son conmutados por la misma central usando una única red de interconexión (figura 1.3). La conmutación integrada evita la necesidad de añadir nuevas redes de interconexión siempre que se introduce un nuevo servicio. Una red con la conmutación integrada debe ser los suficientemente flexible como para satisfacer los requerimientos de retardo y velocidad de transmisión de cada servicio.

Procesamiento integrado de la llamada

El procesamiento integrado de la llamada supone la compartición del software de comunicaciones entre llamadas de diferentes características (figura 1.3). El procesado de las llamadas integrado proporciona una descripción funcional flexible y uniforme de las llamadas, y usa un único procedimiento para mapear los requisitos de recursos de cada llamada con los recursos físicos de la red, lo que presenta mucha ventajas. Proporciona un diseño y desarrollo flexible de nuevos servicios y simplifica la gestión y el mantenimiento de la red. Además una integración del punto de vista lógico de las llamadas separado del punto de vista físico permite que el software de proceso de la llamada sea independiente del sistema físico de conmutación lo que facilita la reutilización de dicho software.

A pesar de las ventajas anteriormente citadas hay razones prácticas por las que el proceso de integración en las redes públicas puede hacerse lento o imposible. Por lo tanto la integración y puesta al día de la nueva red debería ser idealmente compatible con la red ya existente, de forma que podamos reemplazar un terminal o un dispositivo de la red durante el proceso sin que ello implique reemplazar todos los otros terminales o dispositivos de la red.

1.4 Técnicas de multiplexación en redes integradas [Hui90] [Min89]

Como se ha señalado anteriormente, un problema fundamental en el diseño de una red integrada de banda ancha es el establecer un esquema de multiplexación y conmutación capaz de acomodar todos los servicios que pueden estar presentes. A continuación estudiaremos diferentes técnicas de multiplexación y conmutación para estas redes.

El problema de la multiplexación de la transmisión puede ser establecido de la siguiente manera: Tenemos m terminales T_1, T_2, ...,T_m que comparten una misma línea de transmisión de capacidad C bps. Nuestro problema es cómo compartir el ancho de banda de transmisión entre estos terminales. Asumiremos que usamos TDM, de forma que el ancho de banda se comparte transmitiendo información en diferentes instantes de tiempo. También supondremos que hay algún mecanismo (un árbitro), que evita que dos terminales transmitan simultáneamente, pues esto daría lugar a una transmisión ininteligible. (figura 1.4)

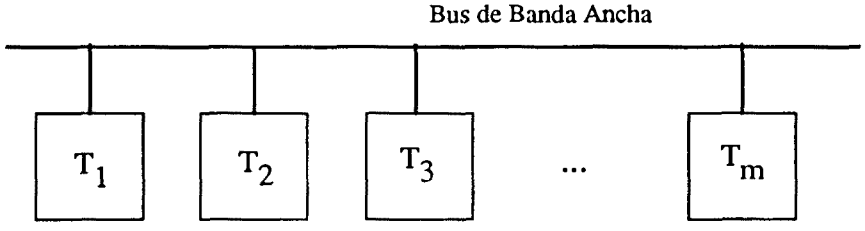


Figura 1.4

Hay dos formas diferentes de asignar estos recursos entre los terminales, conocidos como Modo de Transferencia Síncrona (STM, 'Synchronous Transfer Mode') y Modo de Transferencia Asíncrona(ATM, 'Asynchronous Transfer Mode'). También se pueden usar combinaciones híbridas de ambos.

Synchronous Transfer Mode (STM)

Cuando se usa STM, cada terminal tiene reservados determinados instantes de tiempo durante los cuales puede transmitir información. Para ello, los terminales deben compartir una referencia de tiempo común, que se denomina referencia de trama. El ancho de banda que se reserva para cada terminal se denomina circuito.

Supongamos que los terminales tienen todos la misma velocidad de transmisión, b. La línea puede ser usada como máximo por N = C/b terminales. En este caso, lo más sencillo es tomar una distancia fija entre referencias de trama. Cada uno de estos intervalos se llama trama. Las tramas se dividen en N ranuras. Cada terminal usa una determinada ranura en la que se ha dividido la trama. El número de orden de la ranura que se asigna a un terminal es el mismo a

lo largo de una comunicación y se asigna durante la fase de establecimiento. (figura 1.5).

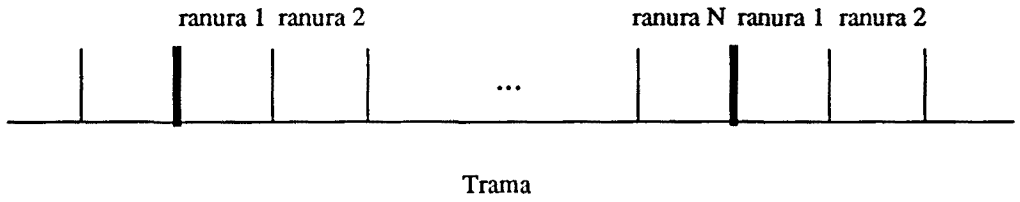


Figura 1.5

Cuando los terminales tienen diferentes velocidades de transmisión, debemos complicar el esquema anterior. Una posibilidad es la siguiente: Agrupamos los terminales en K clases, de forma que los terminales de una clase dada k, tienen una misma velocidad de transmisión b_k. Ahora cada trama se divide en K ventanas y cada ventana en ranuras de diferente tamaño. Dentro de la ventana número k, el tamaño de las ranuras se escogen de acuerdo con la velocidad de transmisión b_k (figura 1.6).

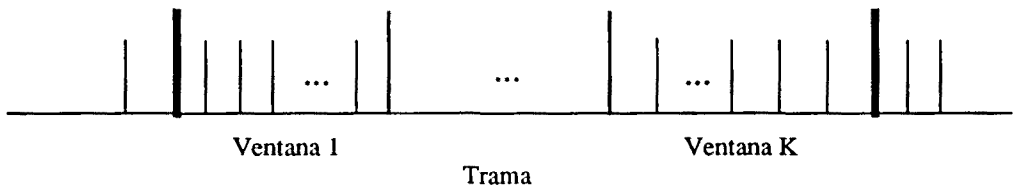


Figura 1.6

Cuando se establece una comunicación se asigna de forma fija una ranura de la ventana adecuada.

Esta forma de multiplexación presenta graves problemas si se quiere usar en una red integrada: Debemos estandarizar las velocidades de transmisión y los tamaños de cada ventana, lo que presenta dificultades a la hora de introducir nuevos servicios en la red. Además, debemos sincronizar los tamaños de las ranuras (que ahora son distintos para cada ventana) y los tamaños de las ventanas.

Se puede pensar en hacer que el tamaño de la ventana cambie de forma adaptativa. Sin embargo, cuando tenemos varias ventanas dentro de una trama, cambiar el tamaño de una ventana supone cambiar muchas divisiones, lo que puede ser muy complejo.

Una forma más flexible de multiplexar terminales de diferentes velocidades de transmisión puede conseguirse cuando se permite que un terminal transmita en más de una ranura de una trama. De esta forma podemos eliminar las ventanas. Para ello se debe escoger una velocidad de transmisión básica. Cuando un terminal transmite a velocidad menor que la básica, sólo ocupa una ranura en cada trama. Si un terminal tiene una velocidad de transmisión

mayor, toma varias ranuras dentro de la trama (figura 1.7).

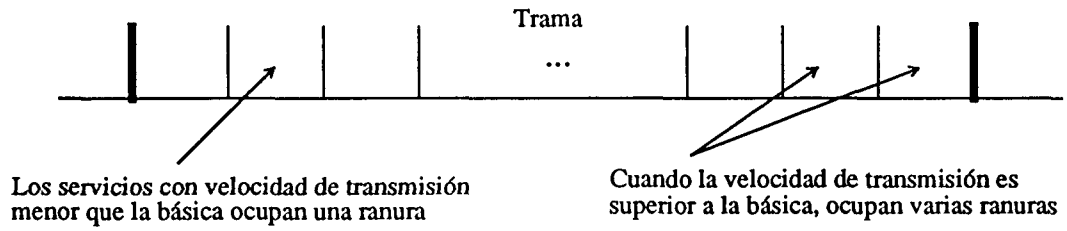


Figura 1.7

El principal problema de este esquema está en la elección de la velocidad de transmisión básica. Si se escoge demasiado pequeña, el número de ranuras de una trama puede crecer demasiado se complica el problema de determinar a qué terminal pertenece cada ranura. Si la velocidad de transmisión es demasiado grande tendremos la posibilidad de que la utilización del ancho de banda sea ineficiente.

Como consecuencia de lo anterior tenemos que, aunque el STM es de uso muy sencillo cuando todos los terminales son del mismo tipo y tienen la misma velocidad de transmisión (por ejemplo, en la red telefónica convencional), en el caso de redes integradas, presenta problemas importantes.

Asynchronous Transfer Mode (ATM)

Vemos que para el caso de una red integrada es difícil escoger una estructura de trama adecuada que cumpla los requerimientos de flexibilidad necesarios. Una solución es eliminar las tramas. En este caso, las ranuras no pertenecen, a priori, a ningún terminal, sino que son tomadas por los terminales según las necesitan. Ahora la información que se transmite en cada ranura debe incluir, además de la información del usuario, un encabezamiento. Este encabezamiento tiene por función primordial identificar las celdas que pertenecen a un mismo canal virtual. Al conjunto de la información de usuario y encabezamiento se denomina celda (o célula) (figura 1.8).

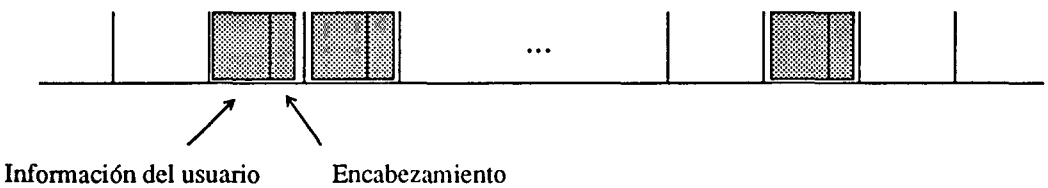


Figura 1.8

En el caso de las redes integradas se suelen usar celdas de tamaño fijo y longitud reducida. Con esto se simplifica la transmisión y se reducen los retardos debidos a la paquetización, es decir, el tiempo que debe esperar la información hasta que el número de bytes sea el suficiente para formar una celda y ser transmitido. En las redes convencionales de transmisión de datos se suelen, en cambio, utilizar celdas mayores y de longitud variable (en ese caso se denominan paquetes).

Las principales ventajas que se derivan del uso de la técnica ATM en una red integrada de banda ancha son:

- Permite obtener la flexibilidad necesaria para el caso de una red integrada.

- Permite una utilización eficiente del ancho de banda cuando el tráfico se produce a ráfagas.

El problema más importante de la multiplexación ATM aparece en las situaciones en que varios terminales quieren ocupar la misma ranura simultáneamente. Evidentemente, sólo uno de ellos puede usarla (el árbitro debe determinar qué terminal es el elegido), mientras que las celdas emitidas por los otros terminales deben esperar. Por lo tanto:

- Deben ser almacenadas, con el consiguiente peligro de que sean perdidas al llenarse las colas de espera.

- Las celdas sufren un retardo variable, que depende de cuántos terminales quieran emitir en ese momento.

Durante el proceso de conmutación de las celdas ATM pueden aparecer estos mismos problemas: los conmutadores deben tener memorias en donde almacenar las celdas cuando se producen conflictos, lo que da lugar a retardos variables y a la posibilidad de pérdidas de celdas.

Se han propuesto esquemas de multiplexación por división de tiempo híbrida ('Hybrid TDM'), en un intento de conseguir las ventajas de los dos esquemas anteriores (STM y ATM). Existen varias posibilidades: Una de ellas es usar una trama con varias ventanas, reservando una de estas ventanas para el tráfico multiplexado mediante ATM. Otra posibilidad consiste en usar un esquema STM en donde una terminal puede ocupar varias ranuras de una trama. Un grupo de estas ranuras se reservan durante el establecimiento de la comunicación (multiplexación STM), mientras el otro grupo de ranuras se reservan bajo demanda (multiplexación ATM).

Estos esquemas híbridos introduce más complejidad a la red, pues se debe distinguir entre varios tipos de tráfico, pero permite introducir la multiplexación ATM de forma que sea compatible con las redes STM ya existentes (por ejemplo, con la red telefónica conmutada).

Durante las primeras fases de discusión sobre la arquitectura de la futura red integrada de banda ancha hubo importantes desacuerdos en cuál era el modo de transmisión adecuado. STM fue inicialmente elegido por muchos como el modo de transferencia ideal. Sin embargo gradualmente la atención se trasladó a ATM, y finalmente en 1989 el CCITT escogió como modo de transferencia el ATM.

En definitiva se puede decir que una red ATM es una red en conmutación de paquetes en donde los paquetes son de reducido tamaño que es fijo y en donde se han simplificado al máximo los protocolos de niveles bajos para conseguir tener elevadas velocidades de transmisión.

1.5 Arquitectura de la red ATM [Toretal90], [Baletal90]

El acceso del usuario a la red

La configuración de referencia para el interfaz usuario-red recomendada por el CCITT es la mostrada en la figura 1.9

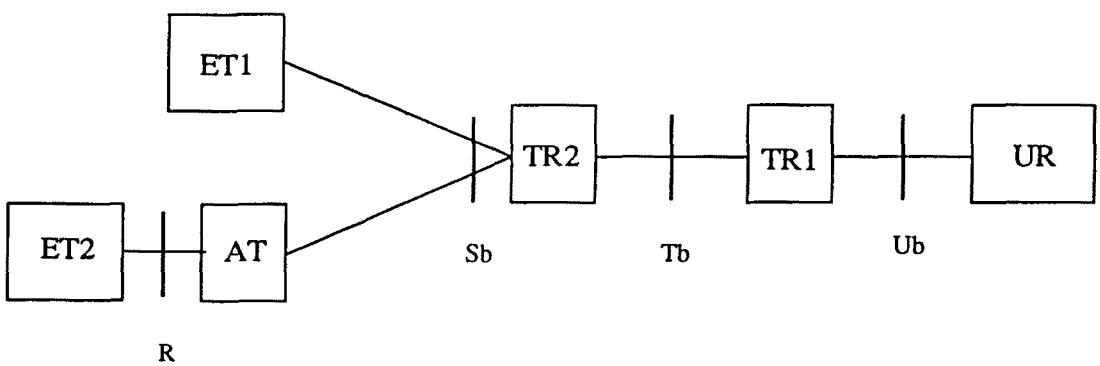


Figura 1.9

Consiste en grupos funcionales separados por los puntos de referencia Rb, Sb, Tb y Ub. Los distintos grupos funcionales son los equipos terminales ET, adaptadores de terminal AT, terminadores de red, TR y la unidad remota UR.

Los equipos terminales se dividen en dos grupos:

- Terminales que siguen las recomendaciones de la red de banda ancha (ET1).
- Terminales que no siguen las recomendaciones (ET2). Este tipo de terminal deberá conectarse a la red a través de un adaptador de terminal.

Los equipos terminales o adaptadores de terminal se conectan con la red a través de los terminadores de red. Los TR1 terminan la línea de transmisión que viene de la central local de conmutación. De entre las funciones que realiza están la regeneración de la señal, el sincronismo de bit, la delimitación de celdas, etc. El TR2 realiza funciones asociadas a las centrales digitales locales y a controladores de terminal. Se prevé que estas funciones sean llevadas a cabo por redes locales de alta velocidad de fibra óptica.

El acceso del usuario a la red se realiza a través de enlaces a 150 Mbps y 600 Mbps. Son pues necesarios medios de transmisión eléctrica u óptica capaces de soportar grandes velocidades de transmisión.

Modelo de protocolo para la red de banda ancha

El modelo de referencia para la interconexión de sistemas abiertos de ISO establece 7 niveles. Los tres primeros, nivel físico, de enlace y de red, corresponden a los llamados servicios portadores. La definición de los teleservicios de banda ancha ATM abarca los 7 niveles.

En el caso de la red ATM se han definido dos nuevos niveles, el nivel ATM y el nivel de adaptación y se distinguen tres planos: El plano de usuario (plano U), para la información entre las aplicaciones de usuario y los protocolos asociados, el plano de control (plano C), para la información de control de la transferencia de información del plano de usuario y una función de gestión de plano, para gestionar la transferencia de la información de usuario y de control (figura 1.10).

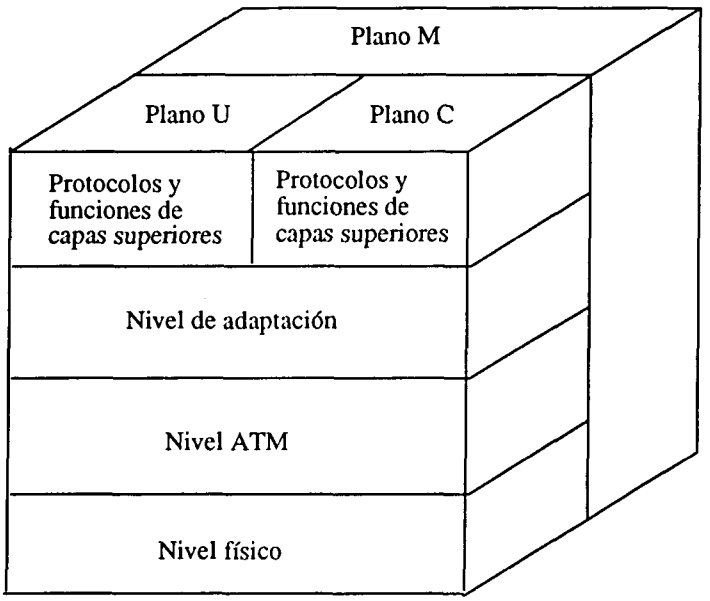


Figura 1.10

El nuevo nivel llamado nivel ATM está situado justo por encima del nivel dependiente del medio físico. El nivel ATM es común a todos los servicios y sus funciones corresponden a

las del encabezamiento de las celdas ATM.

El encabezamiento contiene sólo la información necesaria para transferir el campo de información a través de la red. Las informaciones orientadas a aplicaciones o a servicios no aparecen en el encabezamiento. Entre las funciones que debe asumir el encabezamiento están las siguientes:

- Identificación del canal virtual.
- Detección y tal vez correción de errores en el encabezamiento.
- Indicación de celda no atribuida.
- Identificación de la calidad del servicio (prioridad, etc).
- Numeración secuencial de celdas.
- Identificación del terminal y del equipo de línea.

El nivel de adaptación asegura las funciones del nivel superior de los planos U y C, y soporta las conexiones entre los interfaces ATM y no ATM. Establece una correspondencia entre la información y las celdas ATM. En el extremo emisor, las unidades de información se segmentan o se concentran (por ejemplo: las tramas LAPD se segmentan y las muestras de voz codificada en PCM se concentran). En el extremo receptor la información se recompone a partir de las celdas ATM.

La transmisión en las redes ATM

Las redes ATM pueden trabajar con cualquier jerarquía o sistema de transmisión digital. Sin embargo es conveniente un único estándar en el método de transmisión en el interfaz del usuario con la red. Recientemente el CCITT ha aprobado una nueva jerarquía de transmisión digital, denominada SDH ('Synchronous Digital Hierarchy'), que especifica las velocidades de transmisión, formato, etc, que sirve como estándar para el acceso de los usuarios a la red (recomendaciones G.707, G.708 y G.709).

1.6 El problema de la gestión del tráfico en una red ATM

Un problema fundamental en el diseño de una red ATM es el del control del tráfico. Aunque en los últimos años han aparecido una gran cantidad de propuestas, se puede decir en la actualidad dicho problema aún no ha sido resuelto.

La dificultad está en las especificaciones que se hacen para dicha red: Deben tener acceso a dicha red terminales con velocidades de transmisión dentro de un amplísimo rango, actuando con una gran autonomía, en una red que funciona a gran velocidad.

La imposibilidad de caracterizar de una forma exacta (o incluso aproximada) el tráfico que ofrece cada usuario, la posibilidad de que los usuarios no cumplan las especificaciones fijadas en el momento del establecimiento de la conexión, etc, hacen que sea inevitable la aparición de fenómenos de congestión en la red, con los consiguientes retardos y pérdidas de celdas. La retransmisión de la información pedida es lenta, y no puede usarse en multitud de casos. De ahí se sigue que las especificaciones para la probabilidad de pérdida de celdas tengan para algunos servicios valores extremadamente bajos, lo que hace necesaria una diferenciación en la calidad de la transferencia de la información.

Los controles de congestión que se usan en las redes de paquetes convencionales no son aplicables para el caso de redes ATM debido a la diferencia de velocidades: los mecanismos clásicos tienen un tiempo de respuesta demasiado grande para ser útiles en una red de banda ancha. Se impone pues el desarrollo de nuevos controles de congestión, basados, por ejemplo, en el control de admisión de nuevas comunicaciones, y en el desarrollo de funciones de policía o vigilancia, que debe controlar si el usuario cumple las especificaciones que ha indicado durante el establecimiento de la comunicación.

Capítulo 2

Modelos analíticos de dispositivos de redes ATM

2.1 Cuestiones generales [SauCha81],[MacSau85].

La realización de medidas es el enfoque más directo para evaluar el rendimiento de un sistema. Sin embargo, en el proceso de medida aparecen dos problemas importantes: En primer lugar, no es posible realizar medidas sobre un sistema mientras éste no sea operacional, lo que no ocurre durante las etapas de diseño y desarrollo. En segundo lugar, el proceso de medida suele ser una actividad compleja y cara.

La modelización del sistema es una alternativa para obtener la evaluación de su rendimiento cuando las medidas directas son irrealizables o demasiado costosas. La modelización es un proceso mediante el cual se mapean las propiedades del sistema que se quiere evaluar sobre un sistema que es más sencillo pero que sin embargo retiene las características más relevantes del sistema original. Evidentemente existe un compromiso entre la complejidad del modelo y la exactitud con la que podemos estimar el rendimiento del sistema modelado.

Una vez construido el modelo del sistema en cuestión debemos resolverlo, es decir, debemos encontrar los valores de las magnitudes que nos interesan. Una forma de resolver el modelo es usar un método analítico: representaremos el modelo mediante una serie de ecuaciones, daremos valores a los parámetros del sistema y resolveremos las ecuaciones, de forma que se obtengan las medidas de rendimiento que nos interesan.

A menudo es demasiado complicado hallar las soluciones exactas de las ecuaciones de nuestro modelo analítico. En tales ocasiones dichas ecuaciones se deben resolver de forma aproximada. Evidentemente las soluciones aproximadas son más sencillas de obtener pero introducen un error en la estimación de los valores que queremos medir que muchas veces es difícil de acotar.

Otra alternativa es construir un prototipo (es decir, un sistema que se comporta como nuestro modelo) y tomar medidas directamente sobre él. Este prototipo puede ser un sistema físico (electrónico, mecánico, etc) o un programa de ordenador, es decir, un programa de simulación.

La principal ventaja de los métodos analíticos sobre los basados en la realización de medidas sobre un prototipo es que, en general, son menos costosos. Además, cuando el modelo analítico es resuelto de forma exacta, los valores que proporciona corresponden exactamente a los del funcionamiento del modelo. Este no es el caso de las medidas realizadas sobre prototipos, que están sometidas a errores estadísticos.

La principal ventaja del uso de prototipos sobre los modelos analíticos, es que permiten tratar modelos mucho más complejos y más parecidos al sistema que se

quiere evaluar. Así, y aunque tal como se ha dicho, los modelos analíticos resueltos de forma exacta proporcionan sus resultados de forma exacta, a menudo las aproximaciones que hemos de introducir para obtener un modelo analítico tratable hacen que los valores obtenidos no se acerquen a los valores del sistema original. Además, la exactitud de las medidas obtenidas sobre el prototipo puede hacerse mayor aumentando el periodo de toma de muestras (por ejemplo, en el caso de un programa de simulación, aumentando el tiempo de ejecución del programa).

Las medidas más importantes del rendimiento de un dispositivo de una red ATM son la probabilidad de pérdida de una celda, que pueden alcanzar valores que son del orden de 10^{-10}, y la máxima variación del retardo, muy importante para algunas aplicaciones, como son voz y el video interactivos. Hay otras medidas, como por ejemplo la duración de los periodos con altas tasas de pérdidas, que también pueden ser apropiadas para el caso de una red ATM [WooKos90].

Por lo tanto, los modelos de los dispositivos de este tipo de redes deben ser capaces de caracterizar las distribuciones del retardo y la ocupación de las colas hasta valores extremadamente pequeños, y también deben ser capaces en algunos casos de capturar el comportamiento durante un transitorio. Otro aspecto fundamental de la modelización de dispositivos para redes ATM es la caracterización del tráfico, que será la superposición del tráfico generado por fuentes de características muy diferentes entre sí.

Los métodos de simulación, tal como se ha dicho, permiten abordar de forma flexible y en condiciones realistas el proceso de modelización. Además, son necesarios para validar los métodos analíticos aproximados. Sin embargo, el tiempo de cálculo necesario para obtener estimaciones significativas de valores de probabilidad muy pequeños, tal como se requiere en el caso de las redes ATM, es la mayor parte de las veces excesivamente grande. Por lo tanto si queremos usar simulaciones deben desarrollarse métodos especiales que permitan soslayar este problema [Rob91].

La utilización de prototipos electrónicos permite la representación más detallada de nuestro sistema. Además, en muchos casos pueden funcionar en tiempo real. Requiere el desarrollo de generadores de tráfico que cubran la gran cantidad de posibilidades existentes. Es, sin duda, la forma más costosa de realizar una evaluación.

El problema del uso de modelos analíticos estriba en encontrar métodos de solución que permitan tratar situaciones de mezclas heterogéneas de tráfico y calcular percentiles extremos de las distribuciones del retardo y la ocupación de colas. A lo largo de este capítulo se presentarán diferentes métodos analíticos de resolución que se han usado para la evaluación de dispositivos ATM. Aunque todos estos métodos tienen ya más de una década de historia permanecen abiertas muchas cuestiones que permitan que su utilización sea más sencilla.

2.2 Modelos analíticos de fuentes [Bloetal91].

Dado que estamos estudiando el comportamiento de dispositivos que forman parte de una red de servicios integrados, el tráfico que estará presente en este tipo de redes será la superposición del tráfico generado por fuentes de datos, de voz y de video. Es de esperar que una parte importante de este tráfico sea el resultado de la superposición fuentes de velocidad de transmisión variable, que estará sometido a fluctuaciones a lo largo del tiempo que tendrán un impacto importante en el rendimiento de los sistemas a evaluar y que deberán ser tenidas en cuenta a la hora de modelizar dicho tráfico.

A continuación veremos qué modelos han sido usados para caracterizar el tráfico

generado por distintos tipos de fuentes.

Fuentes con velocidad de transmisión constante (Fuentes CBR).

Las fuentes CBR ('Continuous Bit Rate') son las fuentes que emiten celdas de información a una velocidad constante. Entre ellas están la voz codificada sin supresión de silencio [Srietal91], video codificado con una velocidad de transmisión constante [Veretal88], o el tráfico ofrecido por una red STM.

Aunque la caracterización del tráfico producido por una fuente CBR es muy sencilla, el modelo se hace bastante más complicado cuando se considera la superposición de este tipo de fuentes. Para fuentes CBR idénticas e independientes la superposición es un proceso que no es de renovación y que tiene una naturaleza periódica. En el estudio de la multiplexación de este tipo de fuentes aparece la cola $\Sigma D/D/1$, tratada por ejemplo en [Eck79]. En [RobVir91] se estudia la superposición de fuentes CBR con diferentes periodos.

Fuentes con velocidad de transmisión variable (Fuentes VBR).

Este tipo de tráfico ('Variable Bit Rate') aparece, por ejemplo, en la transmisión de datos, en los casos en donde la voz es codificada de forma que no se generan paquetes durante los silencios, cuando se usan codificadores de velocidad variable para video, y cuando dicho tráfico es generado por otras redes en conmutación de paquetes.

El tráfico proveniente de fuentes de datos suele ser modelado de forma satisfactoria por un proceso de Poisson. Sin embargo, en el caso de estar presentes fuentes de voz o de video los resultados experimentales se apartan de los predichos por los modelos que usan un proceso de Poisson. [SriWit86], [HefLuc86], [Magetal88], [Senetal89].

Ello es debido a que las correlaciones existentes en el número de fuentes activas (caso de la voz) o entre las diferentes tramas de una escena (caso de la imagen), tienen un importante impacto en el funcionamiento del sistema.

El estado de la superposición de este tipo de fuentes puede modelarse mediante una cadena de Markov. Cuando tenemos un solo tipo de fuentes la cadena de Markov que se obtiene es un proceso de nacimiento y muerte, cuyos parámetros dependen de los tiempos de cambio de estado que caracterizan cada fuente individual. En el caso de tener varios tipos de fuentes, la cadena de Markov obtenida es de mayor dimensión [Li90].

En general el modelo del sistema consiste en una cola con tiempo de servicio determinista, debido a la longitud constante de las celdas. La mayor parte de las veces el análisis de tal sistema se puede llevar a cabo mediante un nuevo proceso de Markov (por ejemplo, observando el sistema en los instantes de salida de un cliente de la cola). Para la solución de dichos modelos se han usado varias técnicas. En este capítulo se analizarán con mayor detalle las más importantes:

- Técnicas basadas en métodos analíticos matriciales.
- Técnicas basadas en la aproximación de fluido.

También se han usado otras técnicas de solución, entre las que cabe citar el uso de métodos iterativos para hallar el vector de probabilidad en estado estacionario de la cadena de Markov que describe el estado de la superposición de fuentes y longitud de

cola [Kro91], métodos basados en el estudio de la transformada z de la solucion [Li90], estudio de procesos semi-markovianos [SriWit86], etc.

En general la solución de estos modelos puede presentar dificultades numéricas cuando el número de fuentes es elevado y hay presentes fuentes de diferentes tipos. Se han propuesto posibles soluciones a estos problemas, como el uso de técnicas de descomposición para los casos en que tenemos varios tipos de tráfico [SteAlw91], [Mit88.b], [Li90], el uso de términos asintóticos [SteAlw91], [Bai91] etc.

2.3 Métodos analíticos matriciales

2.3.2 Procesos de llegada markovianos.

El tratamiento de la cola G/G/1 es extremadamente difícil. Sin embargo cuando el proceso de entrada retiene algunas de las propiedades de los procesos markovianos, el análisis de la cola es abordable. Dicha idea es realmente antígua (A.K. Erlang, 1917, D.R. Cox, 1955), aunque hasta recientemente no se han estudiado los problemas numéricos que surgen en el cálculo de las expresiones obtenidos en el análisis de dichas colas [Neu81].

El tratamiento para un amplio conjunto de procesos que retienen propiedades markovianas puede hacerse a través del estudio de un tipo de procesos conocidos como Procesos Markovianos de Llegada (MAP, 'Markovian Arrival Processes'), cuya definición daremos a continuación. En primer lugar consideraremos el caso contínuo.

El proceso de llegada markoviano en lotes de parámetro contínuo (BMAP,'Batch Markovian Arrival Process') [Luc91]

Empezaremos considerando un proceso de Poisson de parámetro λ en donde las llegadas se dan en lotes, siendo p_j la probabilidad de que en un lote lleguen j clientes (j > 0). Sea N(t) el número de llegadas en el intervalo (0,t]. Tenemos pues un proceso de Markov con un espacio de estado i, $0 \leq i$ y con un generador infinitesimal de la forma:

$$F = \begin{vmatrix} d_0 & d_1 & d_2 & d_3 & \cdots \\ 0 & d_0 & d_1 & d_2 & \cdots \\ 0 & 0 & d_0 & d_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{vmatrix}$$

en donde $d_0 = -\lambda$ y $d_j = \lambda\, p_j$ para $j > 0$. Después de estar en un cierto estado i durante un tiempo distribuido exponencialmente, de media λ^{-1}, el proceso salta al estado i+j con probabilidad p_j, en donde dicha transición corresponde a la llegada de un lote de j clientes.

Generalizaremos el proceso anterior permitiendo que el tiempo entre llegadas de lotes no sea exponencialmente distribuido, pero sin perder la estructura markoviana del

proceso. Para ello consideremos un proceso de Markov en dos dimensiones {N(t), J(t)} con un espacio de estados dado por {(i, j); $0 \leq i$ y $1 \leq j \leq m$}. J(t) será el estado de una cadena de Markov que usaremos como substrato en la definición de nuestro proceso. Ahora el generador infinitesimal de {N(t), J(t)} tiene la siguiente estructura:

$$F = \begin{vmatrix} D_0 & D_1 & D_2 & D_3 & \cdots \\ 0 & D_0 & D_1 & D_2 & \cdots \\ 0 & 0 & D_0 & D_1 & \cdots \\ \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \end{vmatrix}$$

en donde ahora D_k $0 \leq k$ son matrices mxm, D_0 tiene en la diagonal elementos negativos y no negativos fuera de la diagonal y D_k $1 \leq k$ son matrices no negativas.

D está definido como

$$D = \sum_{k=0}^{\infty} D_k$$

y es el generador infinitesimal de la cadena de Markov {J(t)}.

Así mismo, definiremos el vector $\bar{\pi}$ como:

$$\bar{\pi} D = \bar{\pi}$$

$$\bar{\pi} \bar{e} = 1$$

Una forma de visualizar el proceso anterior es la siguiente. Consideremos la cadena de Markov que sirve como substrato para la definición del proceso cuyo generador infinitesimal es D. Supongamos que estamos en cierto estado i de dicha cadena de Markov, con tiempo de permanencia en el estado exponencialmente distribuido con parámetro λ_i. Al final de la estancia en el estado i tenemos una transición que puede corresponder o a la llegada de un lote de clientes. Con probabilidad $p_i(0,k)$ $1 \leq k \leq m$, $k \neq i$ habrá una transición al estado k sin llegada de clientes. Con probabilidad $p_i(j,k)$ $1 \leq k \leq m$ habrá una transición al estado k con la

llegada de un lote de j clientes (ver figura 2.1):

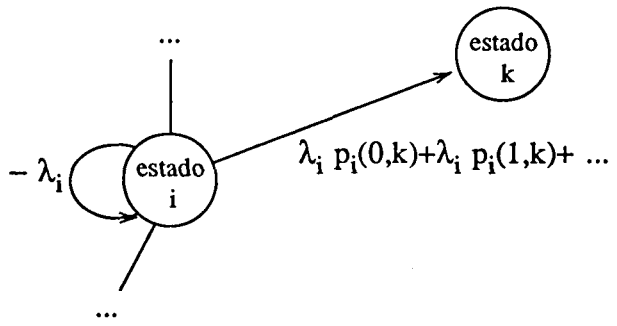


Figura 2.1

Tenemos, pues, para $1 \le i \le m$ que

$$\sum_{\substack{k=1 \\ k \neq j}}^{m} p_i(0,k) + \sum_{j=1}^{\infty} \sum_{k=1}^{m} p_i(j,k) = 1$$

y que

$$(D_0)_{ii} = -\lambda_i \qquad 1 \le i \le m$$
$$(D_0)_{ik} = \lambda_i\ p_i(0,k) \quad 1 \le i,k \le m,\ k \neq i$$
$$(D_j)_{ik} = \lambda_i\ p_i(j,k) \quad 1 \le j\ 1 \le i,k \le m.$$

Por lo tanto la matriz D_0 corresponde a las transiciones en donde no hay a llegadas y D_j corresponde a las transiciones con llegadas de lotes de j clientes.

Sea P(t) la matriz de probabilidad de transición del proceso de Markov {N(t), J(t)}, con generador F. Entonces P(t) satisface las ecuaciones de Chapman-Kolmogorov, teniendo:

$$\frac{d}{dt} P(t) = P(t)\ F \qquad \text{para } t > 0 \tag{2.1}$$
$$P(0) = I$$

Definiremos ahora la matriz mxm P(n,t) con componentes dados por:

$$P_{ij}(n,t) = p\{N(t) = n\ ,\ J(t) = j\ /\ N(0) = 0\ ,\ J(0) = i\}$$

Si partimos P(t) en bloques de mxm elementos, P(n,t) viene dado por el n-esimo

bloque de la primera fila de P(t), por lo que la ecuación (2.1) se transforma ahora en:

$$\frac{d}{dt} P(n,t) = \sum_{j=0}^{n} P(j,t)\, D_{n-j} \qquad 1 \le n \ \text{y}\ t > 0$$

$$P(0,0) = I$$

Si definimos ahora la función generatriz matricial $P^*(z,t)$ como:

$$P^*(z,t) = \sum_{k=0}^{\infty} P(k,t)\, z^k \qquad |z| \le 1$$

obtenemos la ecuación

$$\frac{d}{dt} P^*(z,t) = P^*(z,t)\, D(z)$$

$$P^*(z,0) = I$$

cuya solución viene dada explicitamente por

$$P^*(z,t) = e^{D(z)t} \quad \text{para } |z| \le 1 \text{ y } t > 0$$

en donde $e^{D(z)t}$ es la función exponencial matricial (para la definición y propiedades de dicho tipo de función ver [Bel72]).

La tasa fundamental de llegada viene dada por:

$$\lambda'_1{}^{-1} = \bar{\pi} \sum_{k=1}^{\infty} k\, D_k\, \bar{e}$$

Muchos procesos de llegadas que han sido profusamente usados en las aplicaciones son casos especiales del BMAP:

- El *proceso de llegada markoviano (MAP, 'Markovian Arrival Process')*. El MAP, definido en [Lucetal90], es un BMAP en donde los lotes de clientes tienen siempre longitud 1. Por lo tanto tenemos que $D_j = 0$, $j > 1$. Dentro de casos particulares de MAP están los siguientes:

  - *Proceso de Poisson*: Tomando m=1, con $D_0 = -\lambda$ y $D_1 = \lambda$ tenemos el proceso de Poisson ordinario con parámetro λ.

  - *Procesos de renovación de fase ('PH-renewal process')*. Este tipo de proceso, con representración $(\bar{\alpha}, T)$, (para su definición ver [Neu81]) es un MAP con $D_0 = T$ y $D_1 = -T\, \bar{e}\, \bar{\alpha}$. Dentro de esta clase tenemos la familia Erlang E_k e hiperexponencial H_k, así como combinaciones finitas de dichos

procesos.

- *Proceso de Poisson modulado por una cadena de Markov (MMPP, 'Markov Modulated Poisson Process').* Un MMPP con generador infinitesimal dado por la matriz **R** y la matriz de tasa de llegadas por **L** es un MAP en donde $\mathbf{D_0} = \mathbf{R}-\mathbf{L}$ y $\mathbf{D_1} = \mathbf{L}$.

- *Superposición de procesos de renovación de fase independientes.*

- *Superposición de MAPs independientes.*

- *MAP con llegadas en lotes idénticamente distribuidas e independientes.*

- *Proceso de Poisson con llegadas en lote correladas.*

- *El proceso versátil de Neuts ('Neuts' Versatile Markovian Point Process').* Este proceso fue introducido por Marcel F. Neuts en [Neu79]. Para su definición toma como substrato un proceso de renovación de fasel. Tenemos tres tipos de llegadas asociadas con la evolución del proceso que sirve como substrato: Podemos tener llegadas de Poisson en lotes con una distribución arbitraria dependiente del estado de la cadena de Markov que gobierna el proceso de renovación que sirve de substrato. Además, dicho proceso de Markov puede cambiar de estado con o sin una renovación. A cada cambio de estado podemos tener una llegada de un lote de clientes cuya distribución depende de si se ha dado o no una renovación.

Este proceso es equivalente a un BMAP. Sin embargo la notación usada por el BMAP es mucho más sencilla. La correspondencia entre ambos procesos es la siguiente (se sigue la notación de [Neu79])

$$\mathbf{D_0} = \Delta(\bar{\lambda})\,\Delta(\bar{p}(0)) - \Delta(\bar{\lambda}) + T_{degree}q(0) + T^{o}\alpha_{degree}r(0)$$

$$\mathbf{D_k} = \Delta(\bar{\lambda})\,\Delta(\bar{p}(k)) + T_{degree}q(k) + T^{o}\alpha_{degree}r(k) \quad \text{para } k > 0$$

En definitiva, vemos como una gran cantidad de procesos interesantes pueden ser tratados de una forma unificada a partir del BMAP.

El proceso de llegada markoviano en lotes de parámetro discreto (DBMAP, 'Discrete Batch Markovian Arrival Process').[Blo90]

De forma análoga al proceso BMAP se puede definir un proceso markoviano en donde la cadena de Markov subyacente a la definición es de parámetro discreto; es decir, solo efectua cambios en instantes determinados de tiempo. Dicho proceso es útil para el estudio de sistemas discretos en el tiempo, tal como es el caso de los sistemas ATM. Llamaremos D-BMAP a tales procesos, que han sido definidos en [Blo90]. Las principales diferencias no están tanto en la definición, que es en todo análoga a la hecha para el BMAP, sinó en el análisis de la cola D-BMAP/G/1, que será tratada en detalle en posteriores apartados.

El D-BMAP engloba una serie de procesos que son el equivalente en tiempo discreto a los englobados por el BMAP: proceso de Bernouilli, proceso de Bernouilli modulado por una cadena de Markov, etc [Blo90].

2.3.2 Análisis de colas con procesos de llegada markovianos.

La cola BMAP/G/1 de capacidad infinita.

Empezaremos estudiando la cola BMAP/G/1 con capacidad de buffer infinita. El análisis de dicha cola (en la versión N/G/1) fue tratado por primera vez en [Ram80]. El análisis con la nueva notación BMAP y con nuevos algoritmos numéricos puede hallarse en [Luc91].

Sean T_n los sucesivos tiempos de final de servicio. I_n será el número de clientes en el sistema (incluyendo el servidor) y J_n la fase del proceso de llegada (es decir, el estado de la cadena de Markov que sirve de substrato para su definición) inmediatamente después de T_n. Definiremos $\tau_n = T_n - T_{n-1}$. Es fácil comprobar que la secuencia $\{(I_n, J_n, \tau_n); 0 \leq n\}$ es una secuencia de renovación markoviana.

La matriz de probabilidad de transición esr

$$Q(x) = \begin{vmatrix} B_0(x) & B_1(x) & B_2(x) & B_3(x) & \cdots \\ A_0(x) & A_1(x) & A_2(x) & A_3(x) & \cdots \\ 0 & A_0(x) & A_1(x) & A_2(x) & \cdots \\ 0 & 0 & A_0(x) & A_1(x) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \\ \vdots & \vdots & \vdots & \vdots & \end{vmatrix}$$

$A_n(x)$ y $B_n(x)$ son matrices mxm cuyas componentes están definidas como:

$(A_n(x))_{ij}$ = p{ Dado que hay un final de servicio en el instante 0 y que al menos ha quedado un cliente en el sistema y la fase es i, la próxima salida ocurre antes del instante x con el proceso de llegada en fase j y durante ese servicio han llegado n clientes}

$(B_n(x))_{ij}$ = p{ Dado que hay un final de servicio en el instante 0 y que no ha quedado ningún cliente en el sistema y la fase es i, la próxima salida ocurre antes del instante x con el proceso de llegada en fase j y el sistema queda con n clientes}

De la definición de la matriz $A_n(x)$ es fácil deducir que:

$$A_n(x) = \int_0^x P(n,t)\, dH(u) \qquad \text{para } 0 \leq n \text{ y } x > 0$$

La expresión para $B_n(x)$ es la siguiente (ver figura 2.2):

$$B_n(x) = \sum_{j=1}^{n+1} \int_0^x \int_0^y e^{-D_0 u} D_j \, du \; P(n+1-j, y-u)\, dH(y-u)$$

(y) (u)

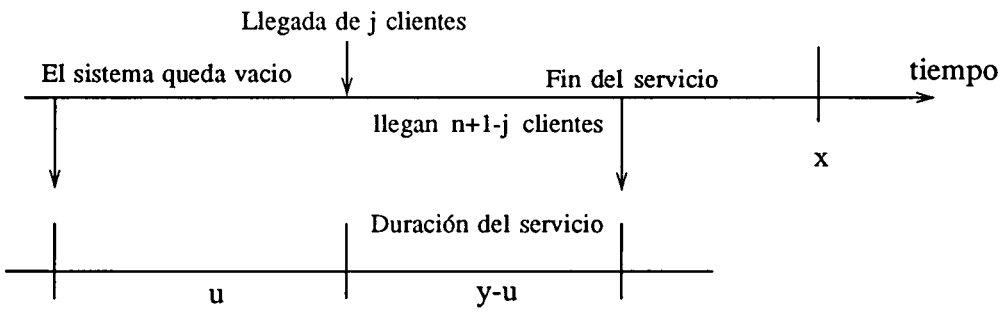


Figura 2.2

A continuación definimos las siguientes transformadas:

$$A_n^*(s) = \int_0^\infty e^{-sx}\, dA_n(x) \qquad B_n^*(s) = \int_0^\infty e^{-sx}\, dB_n(x)$$

$$A^*(z,s) = \sum_{k=0}^{\infty} A_k^*(s)\, z^k \qquad B^*(z,s) = \sum_{k=0}^{\infty} B_k^*(s)\, z^k$$

y también:

$$A_n = A_n^*(0) = A_n(\infty) \qquad B_n = B_n^*(0) = B_n(\infty)$$

$$A = A^*(1,0) \qquad B = B^*(1,0)$$

De lo anterior obtenemos:

$$B^*(z,s) = z^{-1}(sI-D0)^{-1}(D^*(z)-D_0) \quad A^*(z,s)$$
$$B = (I-D_0^{-1})\ A$$
$$B_n = -D_0^{-1} \sum_{j=0}^{n} D_{j+1}A_{n-j}$$

El vector $\bar{\pi}$, ya definido anteriormente, también satisface:

$$\bar{\pi} A = \bar{\pi}$$
$$\bar{\pi}\, \bar{e} = 1$$

y definimos el vector $\bar{\beta}$ como:

$$\bar{\beta} = \frac{d}{dt} A^*(z,0)\Big|_{z=1} \bar{e}$$

Si ρ es la intensidad de tráfico, entonces se cumplirá: $\rho = \lambda'_1\, h = \bar{\pi}\bar{\beta}$.

Las matrices A_n y B_n pueden calcularse siguiendo el algoritmo propuesto por Lucantoni y Ramaswami en [LucRam85]:

En primer lugar se demuestra que P(n,t) puede expresarse como

$$P(n,t) = \sum_{j=0}^{\infty} e^{-\theta t}(\theta t)^j / j!\ \ K_n^{(j)}$$

en donde:

$$\theta = \max_k\ (D_0^{kk})$$

y $K_n^{(j)}$ se define de forma recursiva como:

$$K_0^{(0)} = I; \qquad K_n^{(0)} = 0 \quad \text{para} \quad n > 0$$
$$K_0^{(j+1)} = K_0^{(j)}\ (I + \theta^{-1}D_0)$$
$$K_n^{(j+1)} = \theta^{-1} \sum_{i=1}^{n-1} K_i^{(j)}\ D_{n-i} + K_n^{(j)}\ (I + \theta^{-1}D_0)$$

Substituyendo obtenemos:

$$A_n = \sum_{j=0}^{\infty} \gamma_j K_n^{(j)}; \qquad \gamma_j = \int_0^{\infty} e^{-\theta t} (\theta t)^j / j! \; dH(t)$$

Las matrices B_n pueden ser ahora fácilmente calculadas usando las expresiones dadas anteriormente.

Longitud de cola en los instantes de salida en estado estacionario.

Si ahora consideramos la situación en estado estacionario obtenemos una nueva cadena de Markov cuya matriz de probabilidades de transición vendrá dada por:

$$Q = \begin{vmatrix} B_0 & B_1 & B_2 & B_3 & \cdot & \cdot & \cdot \\ A_0 & A_1 & A_2 & A_3 & \cdot & \cdot & \cdot \\ 0 & A_0 & A_1 & A_2 & \cdot & \cdot & \cdot \\ 0 & 0 & A_0 & A_1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & & \\ \cdot & \cdot & \cdot & \cdot & & & \end{vmatrix}$$

Partiendo el vector estacionario de probabilidad $\bar{x}$ como $\bar{x} = (\bar{x}_0 , \bar{x}_1 , \bar{x}_2 , ...)$ obtenemos el siguiente sistemas de ecuaciones:

$$\bar{x}_i = \bar{x}_0 B_i + \sum_{j=1}^{i+1} \bar{x}_j A_{i+1-j} \qquad \text{para } 0 \le i$$

De donde se deduce que encontrando el valor de $\bar{x}_0$ podemos calcular todos los demas vectores $\bar{x}_i$.$\bar{x}_0$ será calculado a partir de los tiempos de visita al estado 0. Para ello primero necesitamos la expresión para los tiempos de primer paso.

Tiempos de primer paso de un nivel i+1 al nivel i

Empezaremos definiendo $G_{jj'}(r)(k,x)$ como la probabilidad de que el primer paso del estado (i+r,j) al estado (i,j') con $1 \le i$, $1 \le j, j' \le m$ y $r > 0$ ocurra en exactamente k transiciones antes del instante x y que (i, j') sea el primer estado visitado de todos en el nivel i. G(r)(k,x) es la correspondiente matriz mxm. Para el caso r=1 definiremos la

matriz G(k,x). Definiendo la siguiente transformada:

$$G^*(z,s) = \sum_{k=1}^{\infty} \int_0^{\infty} e^{-sx}\, dG(k,x)\, z^k \qquad s > 0 \text{ y } 0 \leq z \leq 1$$

obtenemos la siguiente ecuación no lineal, que juega un papel clave en la solución del sistema:

$$G^*(z,s) = z \sum_{n=1}^{\infty} A_n(s)\, (G^*)^n(z,s) \tag{2.1}$$

Definimos también las matrices:

$$G^*(z) = G^*(z,0)$$

$$G = G^*(1) = \sum_{n=1}^{\infty} A_n\, G^n$$

Cuando $\rho < 1$ tenemos el siguiente vector $\bar{g}$

$$\bar{g}\, G = \bar{g}$$

$$\bar{g}\, \bar{e} = 1$$

y el vector $\bar{\mu}$ definido por:

$$\bar{\mu} = \frac{d}{dt}\, G^*(z,0)\Big|_{z=1} \bar{e}$$

Para el cálculo de la matriz G, que es fundamental para obtener la solución del sistema, se han propuesto varios algoritmos estables desde el punto de vista numérico, que son explicados a continuación.

En primer lugar podemos substituir sucesivamente los términos de una sucesión, cuyo valor inicial sea 0 en la ecuación (2.1), es decir:

$$G_0 = 0$$

$$G_{j+1} = \sum_{n=1}^{\infty} A_n\, G_j^{\,n} \quad \text{para } j > 0.$$

Se puede demostrar que con dicho procedimiento obtenemos una sucesión que tiende

monotónicamente hacia la solución G. Sin embargo, cuando la intensidad de tráfico ρ aumenta la convergencia se hace lenta. Una convergencia más rápida puede obtenerse usando la expresión:

$$G_0 = 0$$

$$G_{j+1} = \sum_{n=0}^{\infty} (I-A_1)^{-1} A_n G_j^n \qquad \text{para } j > 0.$$

En [LucRam91] se ha propuesto otro método de solución. Se demuestra que la matriz G es de la forma:

$$G = \int_0^{\infty} e^{-D(G)x}\, dH(x)$$

de donde se deduce la siguiente expresión para G,

$$\theta = \max_k \; (D_0^{kk})$$

$$G = \sum_{j=0}^{\infty} \gamma_j (I + \theta^{-1} D(G)); \qquad \gamma_j = \int_0^{\infty} e^{-\theta t} (\theta t)^j / j!\; dH(t)$$

y de ahí la recursión:

$$H_{0,k} = I; \quad G_0 = 0$$

$$H_{n+1,k} = ((I + \theta^{-1} D(G_k))H_{n,k} \qquad 0 \le n$$

$$G_{k+1} = \sum_{j=0}^{\infty} \gamma_j H_{n,k}$$

Cálculo de vector $\bar{x}_0$

Definiremos las cantidades $K_{jj'}(k,x)$, $k > 0$, $x > 0$, $1 \le j, j' \le m$ como la probabilidad condicionada de que empezando en el estado (0, j) volvamos por primera vez al nivel 0 en exactamente k transiciones y antes del instante x llendo a parar al estado (0, j'). Definiremos la correspondiente matriz como K(k,x) y su transformada como

$$K^*(z,s) = \sum_{k=1}^{\infty} \int_0^{\infty} e^{-sx}\, dK(k,x)\, z^k \qquad s > 0 \text{ y } 0 \le z \le 1$$

obtenemos:

$$K^*(z,s) = z \sum_{n=1}^{\infty} B_n(s) \, (G^*)^n(z,s)$$

y si definimos $K = K^*(1,0)$ (ver [Luc91]):

$$K = I - D_0^{-1} D(G)$$

Por último, el vector $\bar{x}_0$ puede ser calculado a partir de resultados clásicos de la teoría de procesos de renovación markovianos [Çin72]:

Sea el vector $\bar{k}$ definido como:

$$\bar{k} K = \bar{k}$$
$$\bar{k} \bar{e} = 1$$

y :

$$\bar{k}^* = \frac{d}{dt} K(z,0) \Big|_{z=1} \bar{e}$$

Entonces:

$$\bar{x}_0 = \frac{\bar{k}}{\bar{k} \bar{k}^*}$$

En [Luc91]se demuestra la siguiente expresión para el vector $\bar{x}_0$, que evita el cálculo de las matrices K:

$$\bar{x}_0 = \lambda'_1 \, (1-\rho) \, \bar{g} \, (-D_0)$$

Los vectores $\bar{x}_i$ pueden calcularse de forma sencilla una vez conocido $\bar{x}_0$. Al aplicar la fórmula de forma directa aparecen el problemas numéricos debido a que debemos restar cantidades pequeñas y de magnitud parecida. La solución a este próblema, debida a Ramaswami, es análoga a la propuesta para la cola M/G/1 por P.J.

Burke, obteniendo la siguiente recursión:

$$\tilde{A}_n = \sum_{j=n}^{\infty} A_j G^{j-n} \qquad \tilde{B}_n = \sum_{j=n}^{\infty} B_j G^{j-n} \qquad \text{para } 0 \le n$$

$$\bar{x}_i = (\bar{x}_0 \tilde{B}_{i-1} + \sum_{k=1}^{i-1} \bar{x}_k \tilde{A}_{i+1-k})(I - \tilde{A}_1)^{-1} \qquad \text{para } i > 0$$

estable desde el punto de vista numérico.

Longitud de cola en un instante arbitrario

La relación entre la longitud de cola en los instantes de partida y la longitud de cola en instantes arbitrarios se obtiene con una aplicación del teorema de Blackwell para procesos markovianos de renovación.

Así, $\xi(t)$ y $J(t)$ serán la longitud de cola y la fase de la fuente en un instante arbitrario t. Consideraremos ahora la distribución de la longitud de cola y fase en un instante arbitrario, definiendo:

$$Y(k,j;t) = p(\xi(t) = k , J(t) = j / \xi(0) = k_0 \text{ y } J(0) = j_0)$$

se puede demostrar que los límites en el infinito existen y son independientes de los valores iniciales. Definiremos:

$$y_{kj} = \lim Y(k,j;t) \qquad \text{para } 0 \le k \text{ y } 1 \le j \le m$$

y los vectores de m componentes $\bar{y}_k = (y_{k1}, ...,y_{km})$. El vector $\bar{y}_0$ viene dado por la expresión:

$$\bar{y}_0 = -\lambda'_1{}^{-1} \bar{x}_0 D_0^{-1}$$

de donde se obtiene finalmente:

$$\bar{y}_{i+1} = (\sum_{j=0}^{i} \bar{y}_i D_{i+1-j} -\lambda'_1{}^{-1} (\bar{x}_i - \bar{x}_{i+1})) (-D_0^{-1}) \qquad \text{para } 0 \le i.$$

En [Luc91] se demuestra la siguiente expresión para $\bar{y}_0$

$$\bar{y}_0 = (1-\rho)\,\bar{g}$$

No abordaremos el cálculo de las distribuciones del tiempo de espera en la cola para el caso de longitud infinita. Pueden encontrarse los principales resultados en [Ram80] y [Luc91].

La cola BMAP/G/1 de capacidad finita.

En este caso supondremos que el tamaño del buffer es limitado, de forma que la capacidad del sistema es de N clientes (cola de espera más servidor). La nueva matriz Q tiene ahora la siguiente forma:

$$Q = \begin{vmatrix} B_0 & B_1 & B_2 & B_3 & \cdots & B_{N-2} & \sum_{j=N-1}^{\infty} B_k \\ A_0 & A_1 & A_2 & A_3 & \cdots & A_{N-2} & \sum_{j=N-1}^{\infty} A_k \\ 0 & A_0 & A_1 & A_2 & \cdots & A_{N-3} & \sum_{j=N-2}^{\infty} A_k \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & A_0 & \sum_{j=1}^{\infty} A_k \end{vmatrix}$$

Los argumentos que nos llevaron a calcular $\bar{x}_0$ a partir de los tiempos de primer paso ya no son ahora aplicables, debido a que hemos perdido la homogeneidad espacial. El vector $\bar{x}$ debe ser calculado resolviendo la ecuación de las probabilidades de estado en estado estacionario de forma que, aprovechando la especial extructura de la matriz Q, podamos reducir tanto el espacio necesario para almacenar Q, como el número de operaciones necesaria para resolverlo, obteniendo un algoritmo estable desde el punto de vista numérico.

En [Blo91] se propone el siguiente algoritmo eficiente para calcular el vector $\bar{x}$ que es una generalización a matrices particionadas a bloques del dado en [GraTac85]:

Consideremos la matriz MxM de la forma:

$$X_0 = \begin{vmatrix} x & \bar{b} \\ \bar{a}^t & Y_1 \end{vmatrix}$$

en donde $\bar{a}$ y $\bar{b}$ son vectores de dimensión M-1 y Y_1 es una matriz (M-1)x(M-1). Consideremos ahora la matriz X_1 que se obtiene sumando a las probabilidades de transición desde el estado i al j, con $2 \leq i,j \leq M$, las probabilidades de ir desde i hasta j a través del estado 1, es decir:

$$X_1 = Y_1 + \bar{a}^t (1-x)^{-1} \bar{b}$$

El vector de probabilidad en estado estacionario de X_0 se puede calcular facilmente a partir del correspondiente a X_1.

Aplicando este razonamiento varias veces, obtendremos una sucesión de matrices de dimensión decreciente X_0, X_1, X_2,..., X_{M-1}. La matriz X_{M-1} tiene una dimensión de 2x2, por lo que su vector de probabilidad en estado estacionario puede encontrarse de forma directa. Ahora se calculan los vectores de probabilidad en estado estacionario para X_i, con i = M-2, M-3, ..., 1, y finalmente se obtiene el vector deseado para X_0.

Aplicaremos esta idea a nuestro problema. Queremos encontrar el vector de probabilidad en estado estacionario de una matriz de la forma:

$$Q_0 = \begin{vmatrix} B_0 & B_1 & B_2 & B_3 & \cdots & B_{N-2} & B'_{N-1} \\ A_0 & A_1 & A_2 & A_3 & \cdots & A_{N-2} & A'_{N-1} \\ 0 & A_0 & A_1 & A_2 & \cdots & A_{N-3} & A'_{N-2} \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & A_0 & A'_1 \end{vmatrix}$$

Para ello consideraremos primero un primer paso en la reducción de dimensión. Así definiremos:

$$C_{1,i} = A_i + A_0(I-B_0)^{-1}B_i \quad \text{para } 1 \leq i \leq N-1$$

$$C'_{1,N-1} = A'_{N-1} + A_0(I-B_0)B'_{N-1}$$

con lo que obtenemos la nueva matriz:

$$Q_1 = \begin{vmatrix} C_{1,1} & C_{1,2} & C_{1,3} & \cdots & C_{1,N-2} & C'_{1,N-1} \\ A_0 & A_1 & A_2 & \cdots & A_{N-2} & A'_{N-1} \\ 0 & A_0 & A_1 & \cdots & A_{N-3} & A'_{N-2} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & A_0 & A'_1 \end{vmatrix}$$

Aplicando la recursión k veces llegamos a:

$$Q_k = \begin{vmatrix} C_{k,1} & C_{k,2} & C_{k,3} & \cdots & C_{k,N-k-1} & C'_{k,N-k} \\ A_0 & A_1 & A_2 & \cdots & A_{N-k-1} & A'_{N-k} \\ 0 & A_0 & A_1 & \cdots & A_{N-k-2} & A'_{N-k} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & A_0 & A'_1 \end{vmatrix}$$

en donde

$$C_{k,i} = A_{i-k+1} + A_0(I-C_{k-1,1})^{-1}C_{k-1,i} \quad \text{para } 1 \le i \le N-k$$

$$C'_{k,N-k} = A'_{N-k+1} + A_0(I-C_{k-1,1})C'_{k-1,N-k+1}$$

y finalmente:

$$Q_{N-1} = A'_1 + A_0(I-C_{N-1,1})C'_{N-1,1}$$

De ahí se obtiene el vector $\bar{x}_{N-1}$ que satisface

$$\bar{x}_{N-1}\, Q_{N-1} = \bar{x}_{N-1}$$

y ahora se puede calcular el vector de probabilidad en estado estacionario de Q_{N-2}:

$$(\bar{x}_{N-2}, \bar{x}_{N-1})$$

como

$$\bar{x}_{N-2} = \bar{x}_{N-1} \; A_0(I-C_{N-2,1})^{-1}$$

En general tendremos:

$$\bar{x}_k = \bar{x}_{k+1} \; A_0(I-C_{k,1})^{-1}$$

En la aplicación de este algoritmo cabe destacar lo siguiente (ver los detalles en [Blo91]):

- Solo deben ser almacenadas las dos primeras filas de elementos.
-Todas las cantidades involucradas tienen una interpretación como probabilidades, por lo que son positivas.
- La complejidad del algoritmo es del orden de m^3N^2 en donde m es la dimensión de cada bloque.

Longitud de cola en un instante arbitario

Para la evaluación de las probabilidades en estado estacionario de la longitud de cola en un instante arbitrario, seguiremos el método descrito en ([Blo89]):

En primer lugar calculamos $\bar{y}_0$. Aplicando el teorema de Blackwell [Çin72], obtenemos:

$$y_0^i = \sum_{j=1}^{m} 1/m(0,j) \int_0^{\infty} P_T^{ji}(0,t)\, dt \qquad i = 1,...,m$$

en donde m(0, j) es el tiempo medio de recurrencia del estado (0, j) en Q:

$$m(0,j) = (h - \bar{x}_0 D_0^{-1} \bar{e}) / x_0^j$$

y finalmente:

$$\bar{y}_0 = \bar{x}_0 D_0^{-1} (h - \bar{x}_0 D_0^{-1} \bar{e})^{-1}$$

Ahora las otras probabilidades pueden calcularse así: Consideremos en primer lugar un instante arbitrario de tiempo durante un periodo en donde el servidor esté

ocupado, τ y llamemos G_f (resp. G_b) al tiempo de recurrencia hasta el próximo servicio (resp. desde el anterior servicio)

Definiremos los vectores $\bar{\omega}_n$ (t) dt, n = 1, 2, ..., N con componentes:

ω_n^i (t) dt = p{ En el instante τ el sistema tiene n clientes, el estado de la fuente es i y $t < G_f \leq t+dt$ / El servidor está ocupado en τ }

Para evaluar los anteriores vectores necesitamos la probabilidad conjunta de que haya n llegadas durante G_b y que $t < G_f \leq t+dt$. Sea τ_k el instante en que el servicio actual empezó. Entonces, definimos las siguientes probabilidades:

H_n^{ij} (t) dt = p{ Durante G_b tenemos n llegadas de clientes, $t < G_f \leq t+dt$ y el estado de la fuente en τ_k+G_b es j / el estado $t < G_f \leq t+dt$

Lo que lleva a la siguiente expresión (No incluida en el artículo de BLONDIA y es deducida en el [GarCas90.b]):

$$H_n(t) = (H_n^{ij}(t)) = h^{-1} \int_{0(s)}^{\infty} P(n,s)\,dH(t+s)$$

Con lo que tenemos:

$$\bar{\omega}_n(t) = \bar{x}_0 \sum_{j=1}^{n} U_j\, H_{n-j}(t) + \sum_{j=1}^{n} \bar{x}_j\, H_{n-j}(t) \qquad 0 < n < 1$$

y finalmente:

$$\bar{y}_k = P_{ocup} \int_0^{\infty} \bar{\omega}_k(t)\,dt \quad \text{for} \quad 1 \leq k \leq N-1, \quad \text{y} \quad \bar{y}_N = \bar{\theta} - \sum_{j=0}^{N-1} \bar{y}_j$$

en donde $P_{ocup} = 1-P_{desocup}$ and $P_{desocup} = \bar{y}_0 \bar{e}$.

Para el cálculo de las anteriores integrales se puede usar un razonamiento análogo al hecho en la cola de capacidad infinita, llegando a:

$$\int_0^{\infty} H_n(t)\,dt = h^{-1} \sum_{j=0}^{\infty} \alpha_j K_n^{(j)}; \qquad \alpha_j = \int_{0(t)}^{\infty} \int_{0(u)}^{\infty} e^{-\theta_T u} (\theta_T u)^j / j!\; dH(t+u)$$

Los tiempos de espera

Sea $V_T(t)$ el tiempo que un cliente que entra en el sistema en el instante t debe esperar antes de ser servido. $V_T(t)$ está definido solo para los clientes que no son perdidos. $W_T^j(x)$ se define como:

$$W_T^j(x) = p(\ V_T(t) \leq x,\ J(t) = j)$$

y $W_T^{*j}(s)$ es la transformada de Laplace de la función de densidad correspondiente.

Si en instante t el servidor está ocupado, entonces el tiempo de espera es el tiempo residual de servicio para el cliente que ocupa el servidor más los tiempos de servicio que estás esperando en dicho instante t. Si el servidor está libre, el tiempo de espera es cero.

De lo anterior deducimos [Blo89]:

$$W_T^{*j}(s) = \frac{1}{(1-\bar{y}_N\ \bar{e})}\left(y_0^j + \sum_{n=1}^{N-1} P_{busy}\ \omega_n^{*j}(s)(H^*(s))^{n-1}\right)$$

donde $\omega_n^{*j}(s)$ y $H^*(s)$ son las transformadas de Laplace de $\omega_n^j(x)$ y $H(x)$ respectivamente.

La cola D-BMAP/G/1 de capacidad finita [Blo90].

Ahora estudiaremos el caso de un sistema de tiempo discreto, en donde el proceso de entrada es un D-BMAP y la la cola es de capacidad finita para N clientes, incluyendo el cliente que está siendo servido.

El D-BMAP vendrá caracterizado por matrices mxm D_n, cuyas componentes nos darán las probabilidades de llegadas de n clientes y la fase de la fuente en un ciclo, condicionada a la fase de la fuente en el ciclo anterior. Definimos la matriz D(z) como:

$$D(z) = \sum_{n=1}^{\infty} D_n\ z^n$$

Los tiempos de servicio siguen una distribución dada por $\{g_k\ ;\ k > 0\}$, con transformada:

$$G(z) = \sum_{n=1}^{\infty} g_n\ z^n$$

Introduciremos las matrices mxm $A_n^{(k)}$ cuya componente (i, j) nos da la probabilidad

condicionada de tener n llegadas en el intervalo (0,k] y la fase final es j, dado que la fase inicial era i. Tenemos entonces:

$$A^{(k)}(z) = \sum_{n=0}^{\infty} A_n^{(k)} z^n = \{A^{(1)}(z)\}^k = \{D(z)\}^k$$

De forma análoga a el caso de los sistemas de tiempo contínuo, definimos las matrices mxm A_n cuya componente (i, j) es la probabilidad condicionada de que el número de llegadas durante un servicio sea n y la fase final sea i, dado que la fase inicial era j. Tenemos:

$$A_n = \sum_{k=1}^{\infty} A_n^{(k)} g_k$$

También definiremos A(z) como la transformada z de la anterior secuencia de matrices y A = A(1).

Las matrices B_n se definen de forma que su componente (i, j) nos da la probabilidad condicionada de que en el instante de finalizar el primer servicio la fase es j y han llegado n+1 clientes, dado que el sistema está vacio en el instante inicial y la fase es i. Es inmediato obtener que:

$$B_n = (I-D_0)^{-1} \sum_{j=1}^{n} D_{j+1} A_{n-j}$$

Si definimos B(z) como la transformada z de la anterior secuencia de matrices, llegamos a:

$$B(z) = z^{-1}(I-D_0)^{-1}(D(z)-D_0)A(z)$$

Nuevamente consideraremos el estado del sistema, longitud de cola y fase de la fuente, inmediatamente después de los instantes de salida T_n. Ahora la matriz Q es de la forma:

$$Q = \begin{vmatrix} B_0 & B_1 & B_2 & B_3 & \cdots & B_{N-2} & \sum_{j=N-1}^{\infty} B_k \\ A_0 & A_1 & A_2 & A_3 & \cdots & A_{N-2} & \sum_{j=N-1}^{\infty} A_k \\ 0 & A_0 & A_1 & A_2 & \cdots & A_{N-3} & \sum_{j=N-2}^{\infty} A_k \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \cdots & A_0 & \sum_{j=1}^{\infty} A_k \end{vmatrix}$$

Obtenemos una ecuación que nos permitirá calcular el vector $\bar{x}$ de probabilidades de estado en estado estacionario en los instantes de final de servicio. Los métodos de resolución apuntados para el caso de la cola finita en tiempo contínuo siguen siendo aplicables.

Para calcular el vector de probabilidades de estado en un instante arbitrario, $\bar{y}$ definimos primero E* como el tiempo medio entre salidas del sistema, llegando a [Blo89]:

$$E^* = E\{G\} + \bar{x}_0 (I-D_0)^{-2}(D-D_0)\bar{e}$$

y finalmente:

$$\bar{y}_0 = \bar{x}_0(I - D_0)^{-1} / E^*$$

$$\bar{y}_{n+1} = \sum_{j=0}^{n} \bar{y}_j D_{n+1-j} + (\bar{x}_n - \bar{x}_{n+1})(I-D_0)^{-1}/E^* \quad 0 \le n < N-1$$

$$\bar{y}_N = \bar{\theta} - \sum_{j=0}^{N-1} \bar{y}_j$$

2.4 El método de la aproximación de fluido

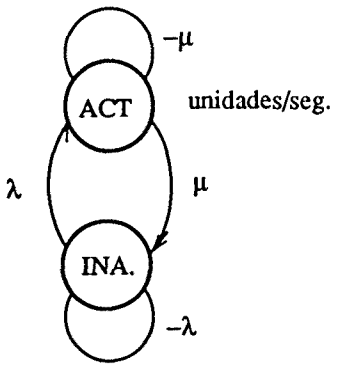
Otro método que ha sido ampliamente usado en la modelización de dispositivos para redes integradas es el que se conoce como método de aproximación de fluido (fluid-flow approximation). Este tipo de aproximación ha sido usada desde hace ya bastante tiempo (ver tomo 2 de [Kle75]). Sin duda los resultados que han facilitado el uso de tal aproximación en la modelización de redes integradas han sido los obtenidos por [Anietal82]. En posteriores trabajos ([Mit88.a], [Mit88.b],[SteAlw91] etc) se han aportado nuevos resultados y se ha estudiado la utilidad del modelo para el estudio de multiplexación de fuentes de voz y video codificados ([Tuc88]).

2.4.1 Análisis de la cola de capacidad infinita.

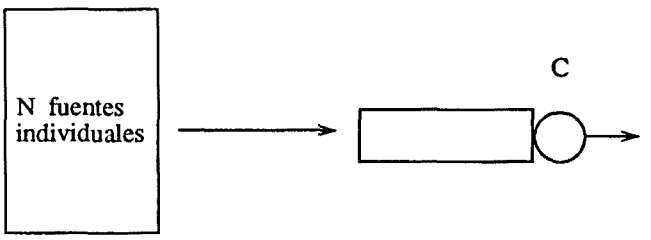
Los resultados expuestos en este apartado pertenecen, en su mayoría, a [Anietal82].

El proceso de entrada consiste en la superposición del tráfico generado por N fuentes de dos estados, estados activo e inactivo. Cuando una fuente está en estado activo emite un flujo contínuo de A unidades de información por unidad de tiempo. Cuando está en estado inactivo no emite. Los tiempos de permanencia en cada estado siguen una distribución exponencial. La media de permanencia en estado activo valdrá

$1/\mu$ y en estado inactivo $1/\lambda$ (ver figura 2.3).



Fuente individual



Cola de capacidad infinita

Figura 2.3

Dicho proceso constituye la entrada a una cola de capacidad infinita, con un servidor capaz de cursar C unidades de información por unidad de tiempo.

Estamos, pues, ante un modelo fluido en el sentido que la información se genera y se cursa de forma contínua (como si estuvieramos llenando y vaciando un recipiente con un líquido). Esto es una buena aproximación al caso real, discreto, cuando los paquetes de información tienen un tamaño reducido frente a los cambios de longitud de cola que se dan entre dos cambios de estado de alguna fuente, y cuando, en el caso de que la cola fuera finita, la longitud de cola no sea muy pequeña.

Estudiaremos la distribución de la longitud de cola en estado estacionario, de donde se puede obtener las probabilidades de pérdida y los retardos en la cola.

Ecuaciones de la distribución de la longitud de cola en estado estacionario

Supongamos que en un cierto instante t tenemos i fuentes activas. Durante un intervalo de duración Δt puede darse el suceso de que una fuente cambie de estado. La probabilidad de que 2 o más fuentes cambien de estado durante dicho periodo de tiempo es de orden $o(\Delta t)$. La probabilidad de que una fuente en estado activo cambie a estado inactivo será $i\mu\Delta t + o(\Delta t)$ (recordemos que los tiempos de permanencia en cada estado siguen una distribución esponencial). La probabilidad de que una fuente en estado inactivo cambie a estado activo vendrá dada por $(N-i)\lambda\Delta t + o(\Delta t)$. La probabilidad de que no haya ningún cambio en el estado de las fuentes será $1-(N-i)\lambda\Delta t-i\mu\Delta t+o(\Delta t)$.

Sean las probabilidades

$$P_i(t,x) = p\{\text{En el instante t hay i fuentes activas, la longitud de cola no excede x / En t=0 había } i_0 \text{ fuentes activas y la longitud de cola no excedía } x_0\}$$

definidas para $0 \leq i \leq N, t > 0, x > 0$.

Si consideramos el vector $\bar{P}(t,x)= (P_0(t,x), ...,P_N(t,x))$, obtenemos el siguiente sistema de ecuaciones:

$$\frac{d}{dt}\bar{P}(t,x) + D\frac{d}{dx}\bar{P}(t,x) = M\bar{P}(t,x)$$

en donde las matrices (N+1)x(N+1) M y D se definen como

$$M = \begin{vmatrix} -N\lambda & \mu & 0 & & & \\ N\lambda & -((N-1)\lambda + \mu) & 2\mu & & & \\ 0 & (N-1)\lambda & -((N-1)\lambda + \mu) & & & \\ & & \ddots & \ddots & & \\ & & & 2\lambda & -(\lambda + (N-1)\mu) & N\mu \\ & & & 0 & \lambda & -N\mu \end{vmatrix}$$

y

$$D = \text{diag}(-c,\ A-c,\ 2A-c,\ldots\ ,NA-c\ ).$$

La solución en estado estacionario vendrá dada considerando el límite de t en el infinito de $P_i(t,x)$, que será llamado $F_i(x)$. Ahora el sistema de ecuaciones diferenciales se transforma en,

$$d\bar{F}(x)/dx = A\ F(x) \qquad x > 0$$

en donde $A = D^{-1}M$.

Tenemos un sistema lineal de N+1 ecuaciones diferenciales ordinarias de coeficientes constantes. La solución puede expresarse como (ver [Bel72]):

$$\bar{F}(x) = \sum_{k=0}^{N} c_k \exp(w_k x)\ \bar{\tau}_k \qquad (2.2)$$

siendo w_k los valores propios por la derecha de la matriz A y $\bar{\tau}_k$ sus correspondientes vectores propios. Los coeficientes c_k se deben determinar a partir de las condiciones de contorno del problema.

Cálculo de los valores y vectores propios de A

En [Anietal82] se calculan los valores y vectores propios de la matriz A de forma explícita. Sea w un valor propio por la derecha de A y $\bar{\tau}$ un vector propio, en donde supondremos que $\tau_N=1$. Tenemos que se cumple

$$w\ D\ \bar{\tau} = M\ \bar{\tau}$$

o lo que es equivalente:

$$w(A^*i-C)\ t_i = \lambda(N+1-i)t_{i-1} - (N + i(\mu-\lambda))t_i + \mu(i+1)t_{i+1} \qquad 0 \le i \le N$$

Llamaremos T(z) a:

$$T(z) = \sum_{j=0}^{N} t_j\, z^j$$

Si multiplicamos las ecuaciones anteriores por z^i y sumamos para todos los valores de i, obtenemos la siguiente ecuación diferencial:

$$\frac{T'(z)}{T(z)} = \frac{\lambda N z + (wC - \lambda N)}{\lambda z^2 + (wA+\mu-\lambda)z - \mu}$$

Las soluciones a la ecuación anterior son de la forma:

$$T(z) = (z-r_1)^{c_1}(z-r_2)^{c_2}$$

en donde:

$$r_{1,2} = \frac{-(wA+\mu-\lambda) \quad ((wA+\mu-\lambda)^2 + 4\mu\lambda)^{1/2}}{2\lambda}$$

y

$$c_2 = N-c_1 \qquad c_1 = \frac{\lambda N(r_1-1)+wC}{\lambda(r_1-r_2)}$$

Ahora debemos hacer una observación que es clave para hallar los valores buscados: Dado que la función generatriz T(z) es un polinomio, y que r_1 y r_2 son distintos, tenemos que c_1 y c_2 deben tomar valores enteros. Sea $c_1=k$, un entero entre 0

y N. Tenemos entonces que la expresión para T(z) viene dada por:

$$T(z) = (z-r_1)^k (z-r_2)^{N-k}$$

Substituyendo k en r_1 y r_2, y teniendo en cuenta la relación entre k y r_1 y r_2, obtenemos la siguiente ecuación para el valor propio w:

$$A(k)\, w^2 + B(k)\, w + C(k) = 0 \quad \text{con } k = 0, 1, ..., N$$

en donde:

$$A(k) = A^2(k-N/2)^2 - (A*N/2 - C)^2$$
$$B(k) = 2\, A(\mu-\lambda)(k - N/2)^2 - N(\mu+\lambda)(A*N/2 - C)$$
$$C(k) = -(\lambda+\mu)^2(N^2/4 - (k - N/2)^2)$$

De las dos valores obtenidos, solo uno es solución de la ecuación (2.2). De ahí es directo encontrar el vector propio asociado, llegando a la siguiente fórmula explícita para dichos vectores propios:

$$t_i = (-1)^{N-i} \sum_{j=0}^{k} \binom{k}{j} \binom{N-k}{i-j} r_1^{k-j}\, r_2^{N-k-i+j} \qquad \text{para } 0 \le i \le N$$

En [Anietal82] se demuestran las siguientes propiedades de los valores propios de A (b es el mayor entero que es menor o igual a C/A):

- Todos los valores propios son reales.

- Hay N-b valores propios negativos, 1 igual a 0 y b positivos.

- El mayor valor propio negativo vale:

$$w = - \frac{\mu + \lambda - \lambda*N*A/C}{A - C/N}$$

A partir de ahora numeraremos los valores propios según se muestra en la figura 2.4:

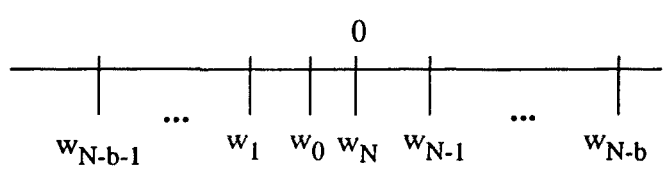


Figura 2.4

Cálculo de los coeficientes

Una vez obtenidos los valores y vectores propios de la matriz A, debemos calcular los coeficientes c_k, que dependen de las condiciones de contorno de nuestro problema.

En el caso de la cola infinita una primera condición puede ser fácilmente establecida: Dado que las funciones $F_j(x)$ han sido definidas como probabilidades, la solución debe estar acotada cuando x tiende a infinito. Ello quiere decir que los términos correspondientes a valores propios positivos deben desaparecer de la expresión (2.2); es decir, los coeficientes c_j con j= N-b a N-1 deben valer 0. Teniendo en cuenta que al tender x a infinito todos los términos con exponente negativo tienden a 0, y que el término correspondiente a w_N es constante, se puede entonces escribir:

$$\bar{F}(x) = \sum_{k=0}^{N-b-1} c_k \exp(w_k x)\, \bar{\tau}_k + F(\infty)$$

:

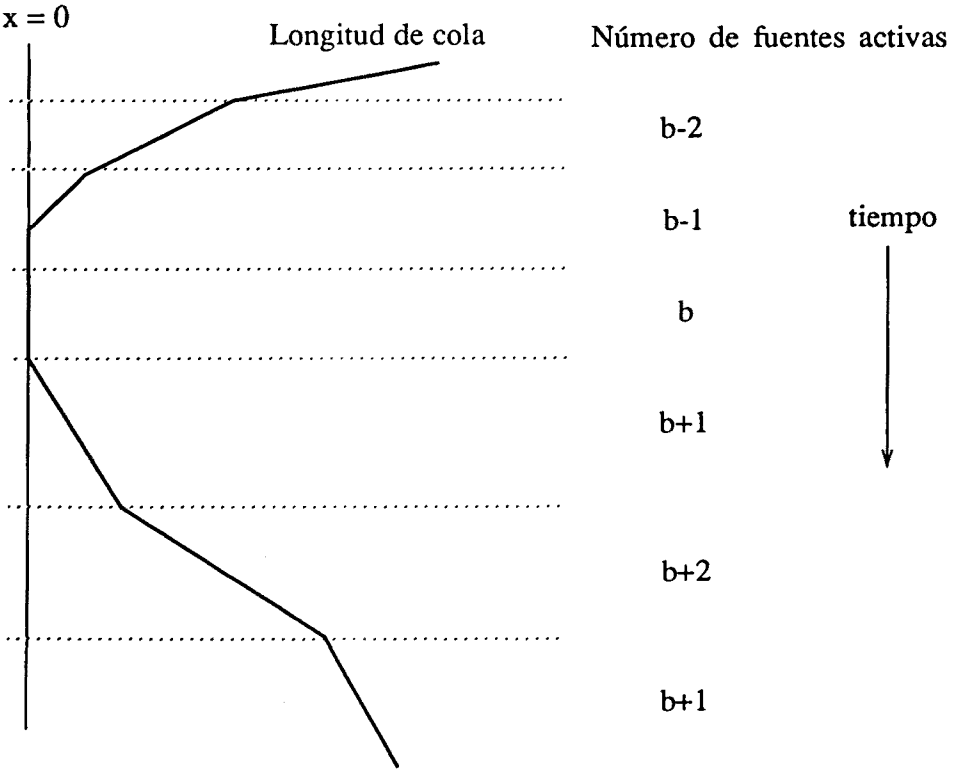


Figura 2.5

Para el cálculo de los otros coeficientes, consideremos la situación en la posición x = 0: Cuando el número i de fuentes activas cumple $i \leq b$, tenemos que la longitud de cola decrece mientras sea positiva. En caso contrario dicha longitud aumenta siempre. En la figura 2.5 tenemos una posible realización del proceso.

Es fácil convencerse de que en los casos en donde el número de fuentes activas

es mayor que b, tendremos que las funciones $F_i(x)$ son contínuas en el origen (figura 2.6),

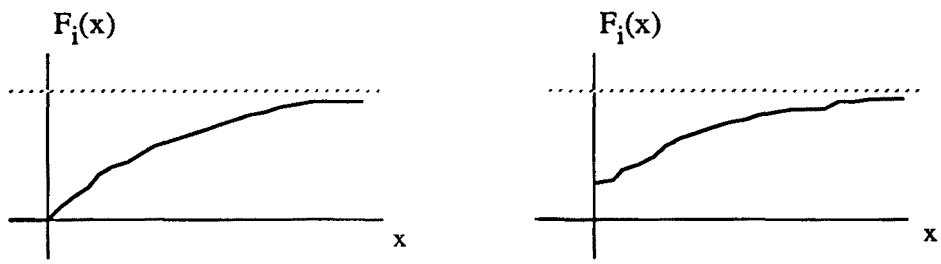


Figura 2.6

Por lo tanto podemos deducir N-b ecuaciones:

$$F_i(0^+) = F_i(0) = 0 \qquad \text{para} \quad i > b$$

de forma que tenemos tantas ecuaciones como incógnitas.

En el caso de la cola infinita la solución al sistema puede ser calculada de forma explícita, siguiendo el siguiente camino:

Teniendo en cuenta la estructura de A, esvemos que la componente i-esima de la derivada j-esima en el origen del vector $\bar{F}(x)$ cumple

$$F_i(x)^{(j)} = (A^j \bar{F}(0))_i = 0 \qquad \text{cuando } b+1+j \le i$$

Si tomamos el caso i = N entonces:

$$F_N(0)^{(j)} = 0 \qquad \text{para } j = 0,1, ..., N-b-1$$

y teniendo en cuenta que las componentes N-esimas de todos los vectores propios elegidos valen 1, llegamos el sistema de ecuaciones (Definimos M = N-b-1):

$$\begin{vmatrix} 1 & 1 & 1 & \cdots & 1 \\ w_0 & w_1 & w_2 & \cdots & w_M \\ w_0^2 & w_1^2 & w_2^2 & \cdots & w_M^2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ w_0^M & w_1^M & w_2^M & \cdots & w_M^M \end{vmatrix} \times \begin{vmatrix} c_0 \\ c_1 \\ c_2 \\ \cdots \\ c_M \end{vmatrix} = \begin{vmatrix} c_N \\ 0 \\ 0 \\ \cdots \\ 0 \end{vmatrix}$$

en donde conocemos que el término c_N vale:

$$c_N = \frac{\lambda^N}{(\mu+\lambda)^N}$$

El sistema obtenido tiene una matriz del tipo Vandermonde, con lo que se encuentra facilmente la solución (ver [Bel72], [GolVan83]), que tiene la forma:

$$c_j = -\frac{\lambda^N}{(\mu+\lambda)^N} \prod_{\substack{i=0 \\ i \text{ DIF } j}}^{M} \frac{w_i}{w_i - w_j} \qquad \text{para } 0 \le j \le M$$

2.4.1 Análisis de la cola de capacidad finita

El análisis de la cola de capacidad finita es totalmente análogo al de la cola de capacidad infinita. Si suponemos que la cola tiene una capacidad de almacenamiento de L unidades de información, nuevamente obtenemos el sistema

$$d\bar{F}(x)/dx = A\,\bar{F}(x) \qquad 0 < x < L$$

en donde la matriz A tiene la misma forma que en el caso anterior, aunque ahora nos restringimos al conjunto (0, L). La solución de la ecuación es:

$$\bar{F}(x) = \sum_{k=0}^{N} c_k \exp(w_k x)\,\bar{\tau}_k \qquad 0 < x < L$$

en donde los valores y vectores propios corresponden a los calculados anteriormente. La única diferencia con el caso de la cola infinita estriba pues en el cálculo de los coeficientes.

Cálculo de los coeficientes

Las condiciones de contorno en el origen siguen siendo las mismas de antes, por lo que nuevamente obtenemos las N-b ecuaciones siguientes:

$$F_i(0^+) = F_i(0) = 0 \qquad \text{para} \quad i > b$$

Las condiciones en el otro extremo de la cola han variado: Ya no podemos asegurar que los términos con valores propios negativos deben anularse. Sin embargo, ahora podemos fijar en el extremo x = L unas condiciones análogas a las establecidas

en el origen: cuando el número de fuentes activas, i, cumpla $i \leq b$, la longitud de cola es siempre decreciente, por lo que no puede permanecer en el valor $x = L$ (ver figura 2.7):

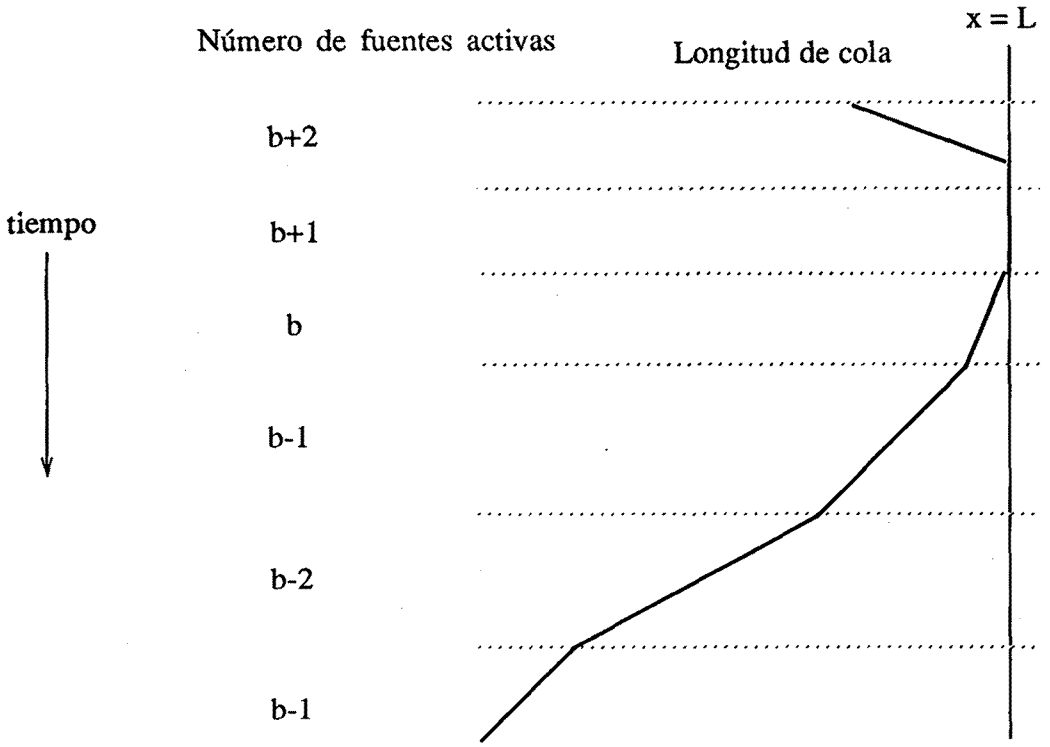


Figura 2.7

y podemos establecer las siguentes condiciones de continuidad (ver figura 2.8):

$$F_i(L^-) = F_i(L) = p\{\text{ Hay i fuentes activas}\} \qquad \text{para} \quad i \leq b$$

de donde obtenemos N+1 ecuaciones para N+1 incógnitas.

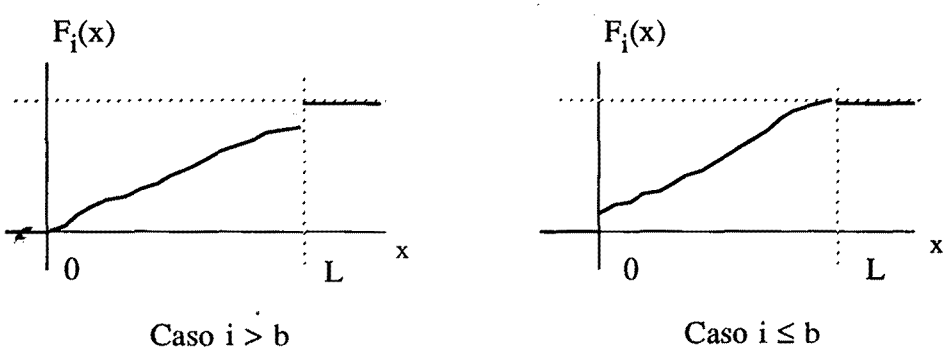


Figura 2.8

2.4 Modelos de fuentes de voz y video.

2.4.1 Modelos de fuentes de voz codificadas.

Tal y como se ha señalado en el capítulo 1, cuando el tráfico de señales vocales constituye una parte significativa del tráfico servido por una red, las técnicas de compresión de voz permiten un incremento considerable del tráfico que es posible cursar. Los métodos de compresión de voz más importantes incluyen algoritmos de modulación adaptativa diferencial de pulsos codificados (ADPCM) e interpolación digital de voz DSI [Srietal91].

Cuando se usan métodos de compresión como el DSI, el tráfico de voz se convierte en un tráfico VBR. A lo largo de los últimos años se han desarrollado modelos análiticos para el estudio del rendimiento de dispositivos en donde se cursa tráfico de voz codificada.

Las diferencias encontradas entre los resultados de los experimentos con trazas reales y los predichos por los modelos analíticos que usan tráfico de Poisson como modelo de la superposición del tráfico de voz son especialmente importantes en sistemas de mediana o baja velocidad (ej. 2Mbps) [HefLuc86], [SriWit86], y son poco importantes para sistemas que trabajan a velocidades muy elevadas, tal y como es el caso de las redes ATM. Así en [Srietal91] se establece que para enlaces a 150 Mbps la hipótesis de tráfico de Poisson es adecuada para cargas de hasta 0.95. Ello es debido que la convergencia hacia un proceso de Poisson de la superposición del tráfico generado por fuentes independientes e idénticamente distribuidas es mejor cuando el número de fuentes es muy elevado.

Cuando no se usa DSI el tráfico generado es determinista, por lo que en su estudio deben usarse los modelos basados en la colas $\Sigma D/D/1$ citados en el apartado 2.2.

Modelo de una única fuente de voz con DSI

Durante una conversación normal, la utilización que cada usuario hace del canal de comunicación es en media inferior al 50%. Por lo tanto, se puede conseguir una importante reducción de los recursos usados por cada usuario si solo se transmite la información durante los intervalos de actividad.

Este hecho fue explotado por los sistemas TASI (Time Assignement Speech Interpolation) en transmisión analógica de señales vocales en cables submarinos. Más recientemente se han desarrollado sistemas digitales con un tipo de funcionamiento similar (Digital Speech Interpolation).

Consideramos que el tráfico generado por una única fuente de voz consiste en una serie de llegadas espaciadas por una distancia T durante los peridodos de actividad de la señal vocal, seguida de un tiempo en donde no se emiten celdas durante los instantes de

silencio. (Figura 2.9)

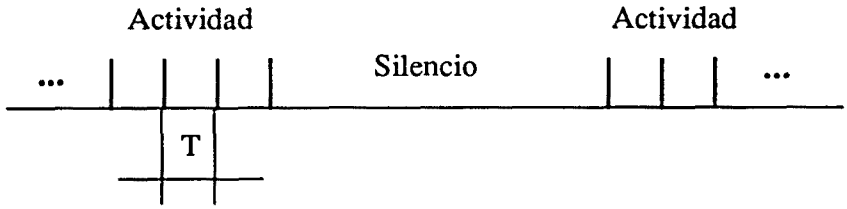


Figura 2.9

Este proceso puede considerarse como un proceso de renovación [HefLuc86], [SriWit86], para el que el tiempo entre llegadas es una variable aleatoria con distribución dada por:

$$F(t) = ((1-\alpha T) + \alpha T(1- \exp(\beta(t-T)))u(t-T)$$

en donde $1/\alpha$ es la media del tiempo de un periodo de actividad, distribuido exponencialmente, y $1/\beta$ es la media de un periodo de silencio. La distribución del tiempo de silencio se considerará también como exponencial, lo que es una buena aproximación cuando el número de fuentes multiplexadas es mayor de 25 [DaiLan86]. Usualmente se toman los siguientes valores:

$1/\alpha = 322$ mseg.

$1/\beta = 650$ mseg.

Superposición de varias fuentes de voz.

El proceso resultante de la superposición de varias fuentes de voz independientes es mucho más complicado que el que describe el tráfico de cada fuente individual. En general deja de ser un proceso de renovación. Además aparecen correlaciones positivas muy grandes, debido a que dicho tráfico es un proceso modulado por el número de fuentes activas que es, a su vez, un proceso correlado.

Para entender esto, consideremos una situación en que estamos multiplexando N fuentes de voz. Si en un momento dado hay i fuentes en estado de actividad, después de un tiempo de, por ejemplo, 1 msec, es muy probable que el número de fuentes activas sea muy próximo a i. Estas correlaciones tienen un efecto muy importante en el comportamiento de los dispositivos, y deben ser modeladas correctamente.

Modelos basados en el Proceso de Poisson Modulado por Markov (MMPP).

Una posibilidad para modelar el tráfico resultante de la superposición, es apróximarlo por un MMPP, que como se ha visto en el apartado 2.3, es un caso particular de BMAP. Este tipo de proceso ha sido usado en [HefLuc86] para determinar el retardo medio que sufre tráfico de voz y datos en un multiplexor estadístico, obteniendo valores muy aproximados a los encontrados mediante simulación. Más recientemente en [Naretal91] se propone un modelo también basado en un MMPP, pero en donde los parámetros del proceso son calculados de forma distinta. Este modelo da buenos resultados a la hora de estimar la probabilidad de prerdida del tráfico de voz en un multiplexor estadístico.

En ambos casos se usa un MMPP de dos estados. Durante la estancia en el

estado 1, el proceso de generación de celdas sigue un tráfico de Poisson de parámetro λ_1. Durante la estancia en el estado 2 el proceso de generación estambién de Poisson, pero de un parámetro distinto, λ_2.(Figura 2.14)

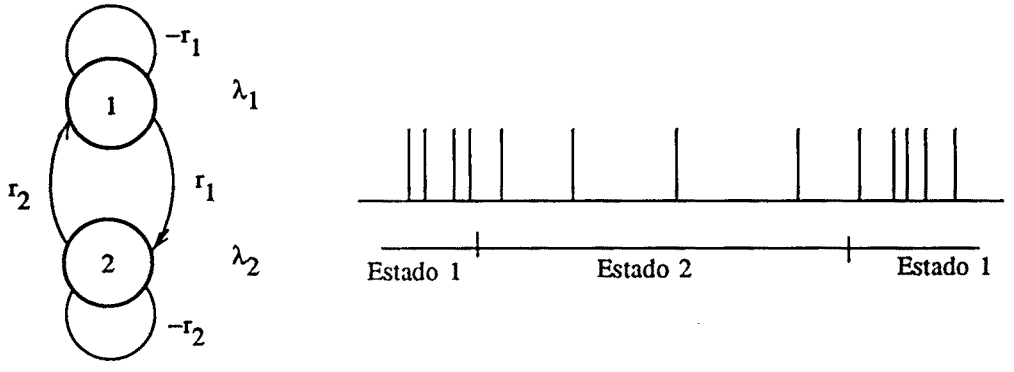


Figura 2.14

Este MMPP es un caso de MAP con matrices D_0 y D_1 de la forma:

$$D_0 = \begin{vmatrix} r_1 - \lambda_1 & -r_1 \\ -r_2 & r_2-\lambda_2 \end{vmatrix} \qquad D_1 = \begin{vmatrix} -\lambda_1 & 0 \\ 0 & -\lambda_2 \end{vmatrix}$$

De lo anterior se deduce que dicho proceso queda determinado por los valores r_1, r_2, λ_1 y λ_2.

Para modelar el tráfico resultante de la superposición debemos dar los valores adecuados a dichos parámetros. Los dos artículos citados anteriormente difieren en qué momentos estadísticos se hacen coincidir entre el tráfico original que es modelado y el MMPP.

Definiremos Ns(0,t) como el número de llegadas en el intervalo (0,t) en el tráfico resultante de la superposición de fuentes de voz independientes.

En [HefLuc86] los momentos que se hacen coincidir son los siguientes:

- El tráfico medio.
- El cociente varianza entre la media del número de llegadas en un intervalo $(0,t_1)$
- El cociente varianza entre la media del número de llegadas en $(0, \infty)$
- El tercer momento del número de llegadas en $(0,t_2)$.

En [HefLuc86] se dan las ecuaciones que relacionan los parámetros del tráfico original y los del MMPP cuando dichos momentos se hacen coincidir. t_1 y t_2 no tienen unos valores definidos analíticamente, pero parece que valores en torno a 500 msec son adecuados.

Los resultados demuestran que el retardo medio queda bien predicho cuando se escoge el anterior conjunto de estadísticas. Sin embargo en [Naretal91] se demuestra

que ese conjunto de valores ya no es adecuado para estimar la probabilidad de pérdida en un multiplexor de capacidad finita.

En [Naretal91] se proponen dos nuevos conjuntos de momentos que se deben hacer coincidir. La elección de uno de estos dos conjuntos depende del tamaño de la cola. Esta diferencia es debido a que en las colas largas o medias las pérdidas que predominan son las se producen a nivel de ráfaga, mientras que el las colas cortas predominan las que se producen a nivel de celda .

A continuación estudiaremos con más detalles el modelo propuesto en [Naretal91]. El estado de la superposición vendrá dado por el número de fuentes activas. Distinguiremos entre los estados para los que es tráfico ofrecido es mayor que el que puede servir el sistema, que llamaremos estados de 'overload' y aquellos en los que el tráfico ofrecido es memor que el que puede servir el sistema, estados de 'underload'. En nuestro MMPP de dos estados haremos corresponder el estado de alta de emisión a la situación de sobrecarga del sistema, y el estado de baja emisión a la de 'underload'. Definiremos las siguientes variables aleatorias:

- $N^s(0,t)$: El numero de llegadas en la superposición en el instante t
- $N_o^s(0,t)$: El número de llegadas en la superposición en el instante t suponiendo que en el instante 0 el proceso estaba en un estado de 'overload'.
- $N_u^s(0,t)$: El número de llegadas en la superposición en el instante t suponiendo que en el instante 0 el proceso estaba en un estado de 'underload'.
- $N^m(0,t)$: El numero de llegadas en el MMPP en el instante t
- $N_h^m(0,t)$: El número de llegadas en el MMPP en el instante t suponiendo que en el instante 0 el proceso estaba en un estado de alta actividad.
- $N_l^m(0,t)$: El número de llegadas en el MMPP en el instante t suponiendo que en el instante 0 el proceso estaba en un estado de baja actividad.

Tanto en el caso del modelo para colas medianas o largas, como para el modelo para colas cortas, hay tres parámetros estadísticos que se hacen coincidir:

- $E(N_o^s(0,t))/t$ y $E(N_h^m(0,t))/t$ para t=0
- Derivada de $E(N_o^s(0,t))/t$ y $E(N_h^m(0,t))/t$ para t=0
- $E(N_o^s(0,t))/t$ y $E(N_h^m(0,t))/t$ para $t=\infty$

Se puede ver que de esta forma conseguimos igualar las funciones $E(N_o^s(0,t))/t$ y $E(N_h^m(0,t))/t$ para cualquier t.

La cuarta ecuación necesaria para determinar todos los parámetros del MMPP depende del tamaño de la cola considerada. Para colas medias o largas (predominan las pérdidas a nivel de ráfaga) la ecuación escogida viene dada por:

- $var(N_o^s(0,t)$ y $var(N_h^m(0,t)$ se hacen coincidir para $t=t_m$.

t_m se escoge como 1 sec. Para colas cortas la ecuación es

- $var(N_u^s(0,t)$ y $var(N_l^m(0,t)$ se hacen coincidir para $t=t_m$.

De los resultados se desprende que este tipo de modelo predice bien las probabilidades de pérdidas en los dos tipos de buffers.

Modelos basados en la aproximación de fluido

Otro tipo de modelos que se han usado para modelar la superposición de fuentes de voz codificadas se basa en la aproximación de fluido explicada anteriormente. Este tipo de modelos son expuestos en, por ejemplo, [DaiLan86], [Tuc88] y [Naretal91]. En este caso, cada fuente individual de la aproximacion fluida representa una fuente de voz codificada. Los resultados obtenidos muestran que este tipo de modelo predice bien las pérdidas que se producen a nivel de ráfaga, con una exactitud del mismo orden que la obtenida con el MMPP [Naretal91]. Sin embargo las predicciones del modelo son peores en el caso de colas cortas (en donde predominan las pérdidas a nivel de celda).

2.4.2 Modelos de fuentes de video codificadas [Bloetal91].

A diferencia de lo que ocurre en el caso de las señales vocales, el tráfico producido por un codificador VBR de video, exhibe variaciones continuas en la velocidad de transmisión, por lo que se han tenido que desarrollar nuevos modelos para su estudio.

En [Magetal88] se presentan dos modelos del comportamiento de un multiplexor estadístico cuya entrada es la superposición de fuentes de video codificadas de forma que solo se transmiten las diferencias entre los pixels de tramas consecutivas, cuando dichas diferencias superan un cierto umbral [Hasetal72].

Modelo markoviano autoregresivo.

El primer modelo propuesto es un modelo markoviano autorregresivo. Si r(n) representa la velocidad de transmisión de una fuente individual durante la trama n-esima, se cumple la siguiente relación recursiva:

$$r(n) = a\ r(n-1) + b\ w(n)$$

donde w(n) es una secuencia de variables aleatorias gausianas independientes, y a y b son constantes, cuyos valores se encuentran a partir de los resultados experimentales. Este modelo se ajusta bien a los resultados experimentales. Sin embargo su tratamiento analítico es muy complejo, por lo que debe resolverse mediante simulación, con las limitaciones que ello supone.

Modelo basado en la aproximación de fluido.

El segundo modelo se basa en la aproximación de fluido. El tráfico generado por la superposición de N fuentes de video se modela con M fuentes de dos estados del tipo visto en el apartado 2.4. Los valores de los tiempos de cambio de estado y de la intensidad de tráfico de cada fuente individual se hallan a partir de resultados

experimentales (ver [Magetal88]). Parece que un valor adecuado para M es M = 20 N.

Aunque este modelo predice con menos exactitud que el anterior el comportamiento del multiplexor, permite el uso las técnicas analíticas presentadas en el apartado 2.4, lo que facilita la obtención de resultados.

Estos modelos son aplicables para casos en donde no hay cambios significativos de escena dentro de la comunicación (por ejemplo, una persona hablando). En [Senetal89] se extiende el modelo anterior a casos en donde hay cambios de escena (y por lo tanto cambios bruscos en la velocidad de transmisión del codificados) o tenemos la multiplexación de fuentes de video de diferentes características. Para ello introduce fuentes individuales de dos estados de diferentes tipos, encontranso el valor de los parámetros del modelo a partir de resultados experimentales.

Modelo basado en el MMPP con llegadas en lotes.

En [Yasetal89] se propone un modelo basado en un MMPP, en donde ocurren llegadas en lotes durantes los cambios de estados. Cada estado representa el tráfico del codificador durante una escena. Los cambios de estado provocan un pico en la velocidad de transmisión, debido a que las diferencias entre tramas se hacen muy grandes. Se verifica la suposición de que los tiempos entre cambios de escena siguen aproximadamente distribuciones exponenciales mediente resultados experimentales. Para la resolución del modelo pueden usarse técnicas analíticas matriciales.

Modelo basado en el D-BMAP

En [Blo91] se estudia un modelo de la superposición de fuentes de video mediante el D-BMAP.

El estado de la superposición (que es equivalente al número de fuentes de dos estados individuales activas en el modelo de [Magetal88]) se cuantifica en M+1 niveles, de forma que obtenemos una cadena de Markov de parámetro discreto con una matriz de transición dada por:

$$D = \begin{vmatrix} 1 - Mb & Mb & 0 & \dots & 0 & 0 \\ a & 1- a -(M-1)b & (M-1)b & \dots & 0 & 0 \\ 0 & 2a & 1-2a-(M-2)b & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & Ma & 1-Ma \end{vmatrix}$$

Cuando estamos en un estado n, la probabilidad de tener un lote de k llegadas vale:

$$c_k(n) = \binom{n}{k} (d-1)^{n-k} d^{-n}$$

en donde 1/d es la probabilidad de que una fuente individual activa genere un cliente. Si

definimos la matriz C_n como:

$$C_n = \text{diag}(c_n(0), c_n(1), ..., c_n(M))$$

y $D_n = C_n D$, obtenemos un D-BMAP del tipo descrito en el apartado 2.2.

Capítulo 3

Estudio del comportamiento asintótico de la probabilidad de pérdida de la cola D-BMAP/D/1.

3.1 Introducción.

En el capítulo anterior se han presentado una serie de métodos exactos de solución de modelos analíticos que son útiles para la evaluación de redes ATM. Sin embargo, en muchas ocasiones, cuando queremos modelar un sistema real el número de operaciones a realizar y los requerimientos de almacenamiento en memoria se hacen excesivamente elevados. Una forma de eludir este problema es recurrir a una solución aproximada, estudiando, por ejemplo, el comportamiento asintótico del sistema cuando la longitud de cola tiende a infinito. Es de esperar que esta aproximación sea tanto mejor cuanto mayor sea la longitud de cola y menor la probabilidad de pérdida. Esta situación es, de hecho, la que habitualmente aparece en las aplicaciones y, por otra parte, es la que presenta mayores problemas en su solución exacta.

Entre los trabajos en los que se hacen estudios similares cabe citar los siguientes: En [Neu86] se estudia el comportamiento asintótico de la cola BMAP/G/1. En [Bai91] se presenta un estudio del comportamiento asintótico de la cola MMPP/G/1, basado en una descomposición espectral del vector $\Pi(z)$, la función generatriz de los vectores de estado en estado estacionario. Muchos de los resultados de este capítulo son una extensión de este trabajo. En [Li90] se da un método de solución similar (aunque no se estudia el comportamiento asintótico del sistema) cuando el proceso de entrada es del tipo BMAP, o superposición de BMAPs independientes. En [BroSim88] se estudia un caso particular de nuestro modelo, un sistema discreto en donde la entrada es la superposición de una fuente determinista y un tráfico caracterizado por un proceso de renovación. También se hace un estudio del comportamiento asintótico de las probabilidades de estado de la cola. En dicho artículo no se usan las propiedades del producto de Kronecker de matrices, sino propiedades sobre la continuidad de las funciones involucradas en el análisis, aunque los resultados obtenidos son totalmente análogos.

3.2 Comportamiento asintótico de la cola D-BMAP/D/1.

Modelaremos el multiplexor mediante una cola cuyo tráfico de entrada es un D-BMAP de N estados. Las probabilidades de emisión durante los cambios de estados están dados por la sucesión de matrices D_i, siendo D(z) la transformada z de dicha sucesión. Tenemos un servidor determinista con tiempo de servicio unidad, aunque la mayor parte de los resultados de este capítulo se pueden extender de forma inmediata al caso de tiempos de servicio con una distribución general. Suponemos que los servicios se producen justo antes de finalizar una ranura de tiempo, mientras que las llegadas se producen justo después de empezar la ranura de tiempo.

La función generatriz de los vectores de probabilidad en estado estacionario

Llamaremos $\pi_S(n)$ al vector cuya componente i-esima es la probabilidad en estado estacionario de tener, justo antes de un servicio, n clientes en la cola y la fuente en estado i,

cuando el tamaño de la cola es S (figura 3.1):

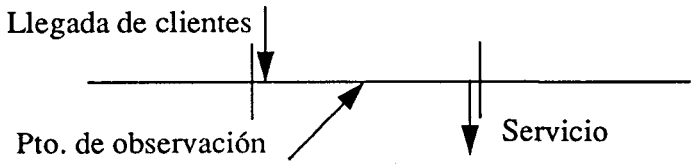


Figura 3.1

Las ecuaciones para las probabilidades de ocupación en estado estacionario son:

$$\pi_S(0)\,D_{i+1} + \sum_{j=1}^{i+1} \pi_S(j)\,D_{i+j-1} = \pi_S(i) \qquad i := 0,1,...S$$

La forma de dichas ecuaciones es independiente de la longitud de cola. De ahí se desprende (ver [Bai91]) que su solución es:

$$\pi_S(i) = \pi_S(0)\,C_i \tag{3.1}$$

en donde la secuencia de matrices C_i no depende de S. Para obtener la transformada z de dicha secuencia, consideraremos el caso de la cola infinita, multiplicaremos cada miembro de la ecuación i-ésima por z^i y sumaremos todas las ecuaciones obteniendo:

$$C(z) = (z-1)\,D(z)\,(zI + D(z))^{-1}$$

Descomposición espectral

Llamaremos x(k,z) al k-ésimo autovalor de la matriz D(z), siendo el vector fila v(k,z) el correspondiente autovector por la izquierda y el vector columna u(k,z) el correspondiente autovector por la derecha. V(z) será la matriz

$$V(z) = \begin{bmatrix} v(0,z) \\ v(1,z) \\ \cdots \\ v(N-1,z) \end{bmatrix}$$

mientras que U(z) será:

$$U(z) = \begin{bmatrix} u(0,z) & u(1,z) & \cdots & u(N-1,z) \end{bmatrix}$$

Los vectores u(k,z) y v(k,z) se eligen de forma que su producto escalar valga 1.

D(z) puede entonces expresarse como:

$$D(z) = U(z)X(z)V(z) = \sum_k x(k,z)\,u(k,z)\,v(k,z)$$

siendo X(z) una matriz diagonal cuyos elementos son los autovalores x(k,z). De lo anterior deducimos:

$$C(z) = U(z)Y(z)V(z) = \sum_k y(k,z)\,u(k,z)\,v(k,z)$$

donde Y(z) es una matriz diagonal cuyos elementos, y(k,z), están definidos como:

$$y(k,z) = \frac{(z-1)x(k,z)}{z-x(k,z)}$$

Los polos de las funciones y(k,z) son las raíces de la ecuación:

$$z = x(k,z) \tag{3.2}$$

Llamaremos ζ_i a los polos de módulo mayor o igual a 1 y η_i a los polos de módulo menor a 1. En el apéndice A de este capítulo se demuestra que uno de los polos ζ_i de módulo mínimo es siempre real y positivo, al que llamaremos ζ_0. Asumiremos que los polos se numeran siguiendo el siguiente orden:

$$\ldots \le |\eta_1| \le |\eta_0| \le 1 < \zeta_0 \le |\zeta_1| \le \ldots$$

En lo que sigue, y para simplificar la notación, supondremos que el resto de polos ζ_i tiene un módulo estrictamente mayor que ζ_0, aunque la extensión al caso más general es directa.

Sea θ es el vector estacionario de la cadena de Markov subyacente a la definición del D-BMAP. De (3.1) tenemos:

$$\pi_S(0)\sum_{k=0}^{S} C_k = \theta$$

La sucesión:

$$L_k = \sum_{i=0}^{k} C_i$$

tiene por transformada:

$$L(z) = \frac{C(z)}{1-z}$$

Si llamamos $Q(\zeta_i)$ y $Q(\eta_i)$ a los residuos de los polos correspondientes, la sucesión L_k podrá expresarse como

$$L_k = \sum_{i=0}^{m} Q(\zeta_i)\,\zeta_i^{-k} + \sum_i Q(\eta_i)\,\eta_i^{-k} + Q(1) + R_k$$

en donde R_k es una sucesión $o(\zeta_m^{-k})$ (ver [Bai91]). En el apéndice B se da una fórmula para los residuos de los polos de L(z).

Comportamiento asintótico de las probabilidades de estado

Cuando la longitud de cola tiende a infinito los términos correspondientes a polos menores a uno también tienden a infinito. De ahí deducimos el siguiente sistema de ecuaciones para la cola de capacidad infinita (ver [Ide88] y [BroSim88]):

$$\pi_\infty(0)\,e = 1 - \rho_{Tot}$$

$$\pi_\infty(0)\,u_i(\eta_i) = 0$$

siendo ρ_{Tot} el tráfico total de entrada.

Si llamamos ζ_0 al autovalor de mínimo módulo mayor que 1, tenemos que los vectores de probabilidad de estado del sistema tienen el siguiente comportamiento asintótico:

$$\pi_\infty(i) = \pi_\infty(0)\,(1-\zeta_0)\,Q(\zeta_0)\,\zeta_0^{-i} + o(\zeta_0^{-i})$$

Comportamiento asintótico de las probabilidad de pérdida

Si queremos hallar el comportamiento asintótico de la probabilidad de pérdida, que llamaremos Pt, se puede usar un razonamiento análogo al usado en [Bai91]:

Tenemos que

$$Pl = \frac{\pi_k(0)\,e + (\rho_{Tot} - 1)}{\rho_{Tot}}$$

con

$$\pi_k(0)\,L_k = 0 \qquad (3.3)$$

Si ahora definimos:

$$L_k^{(a)} = L_k^{-1} det(L_k)$$

llegamos a:

$$Pt = \frac{\theta L_k^{(a)} e + (\rho_{Tot} - 1) det(L_k)}{\rho_{Tot} det(L_k)} \tag{3.4}$$

Además, aplicando la regla de Kramer en (3.4):

$$L_k^{(a)} = \sum_i det(L_k[i])$$

en donde la matriz $L_k[i]$ se obtiene substituyendo la fila i-esima de L_k por el vector θ. Definiendo ahora la matriz G_k como:

$$G_k = L_k + \frac{1}{\rho_{Tot} - 1} e\theta$$

y usando la propiedad P1 del apéndice A de [Bai91] obtenemos:

$$det(G_k) = det(L_k) + \frac{1}{\rho_{Tot} - 1} \sum_i det(G_k[i])$$

y

$$\sum_i det(L_k[i]) = \sum_i det(G_k[i])$$

de forma que (3.4) puede expresarse como

$$Pt = \frac{(\rho_{Tot} - 1) det(G_k)}{\rho_{Tot} det(L_k)}$$

Nuestro objetivo es ahora determinar el comportamiento asintótico de los determinantes de la fórmula anterior cuando la longitud de cola tiende a infinito. Para ello definimos la matriz H_k:

$$H_k = Q(\zeta_0^{-k})\,\zeta_0^{-k} + \sum_i Q(\eta_i)\,\eta_i^{-k}$$

En el apéndice B se demuestra que los residuos Q(.) son matrices de rango 1. Por lo tanto, aplicando la propiedad P2 del apéndice A de [Bai91] llegamos a:

$$det(G_k) = det(H_k) + o(\zeta_0^{-k}\eta_1^{-k}\eta_2^{-k}\ldots)$$

y

$$det(L_k) = det(H_k + \frac{1}{1-\rho_{Tot}} e\theta) + o(\zeta_0^{-k}\eta_1^{-k}\eta_2^{-k}\ldots)$$

Para simplificar las expresiones anteriores, haremos la suposición de que todas las ecuaciones (3.2) tienen una única solución de módulo menor o igual que uno. En el apartado dedicado a la superposición de fuentes VBR y fuentes periódicas veremos cómo tratar el caso en que alguna de estas ecuaciones no tenga solución (es decir, tenemos menos polos de módulo menor o igual a uno que estados tiene la fuente).

Definamos las matrices

$$U_i(z) = \left[u(i,z)\;\; u(1,\eta_1)\;\; \ldots\;\; u(N-1,\eta_{N-1})\right]$$

$$V_i(z) = \begin{bmatrix} v(i,z) \\ v(1,\eta_1) \\ \ldots \\ v(N-1,\eta_{N-1}) \end{bmatrix}$$

y

$$\Phi_k = diag(y(0,\zeta_0)\,\zeta_0^{-k}, y(1,\eta_1)\,\eta_1^{-k}, \ldots, y(N-1,\eta_{N-1})\,\eta_{N-1}^{-k})$$

H_k puede expresarse como:

$$H_k = U_0(\zeta_0)\,\Phi_k V_0(\zeta_0)$$

de forma que

$$det(H_K) = (\prod_i (y(i,\eta_i)\eta_i^{-k}))\ [(y(0,\zeta_0)\zeta_0^{-k})\, det(U_0(\zeta_0)\, V_0(\zeta_0))]$$

y

$$det(H_k + \frac{1}{1-\rho_{Tot}} e\theta) = det(H_k) + \prod_i y(i,\eta_i)\eta_i^{-k}\left[\frac{1}{1-\rho_{Tot}} det(U_0(1)\, V_0(1)) + o(1)\right]$$

Finalmente:

$$P_t = \frac{(1-\rho_{Tot})^2 det(U_0(\zeta_0)\, V_0(\zeta_0))}{-\rho_{Tot} det(U_0(1)\, V_0(1))} y(0,\zeta_0)\zeta_0^{-k} + o(\zeta_0^{-k}) \qquad (3.5)$$

3.3 Multiplexación de fuentes VBR y fuentes periódicas.

A continuación desarrollamos dos modelos analíticos, uno exacto y otro aproximado, para la multiplexación de fuentes de frecuencia de emisión variable (fuentes VBR) y fuentes periódicas (figura 3.2). Estos modelos permiten tratar, por ejemplo, la superposición de tráfico CBR (frecuencia de emisión constante) con tráfico proveniente de fuentes VBR. También nos permitirá estudiar el efecto de la aparición de clusters de celdas sobre la probabilidad de pérdida en un multiplexor. Los modelos desarrollados son modelos discretos. En el capítulo dedicado a los modelos de mecanismos de prioridad espacial se presenta un modelo continuo basado en la aproximación de fluido que permite el estudio de la multiplexación de fuentes VBR y CBR.

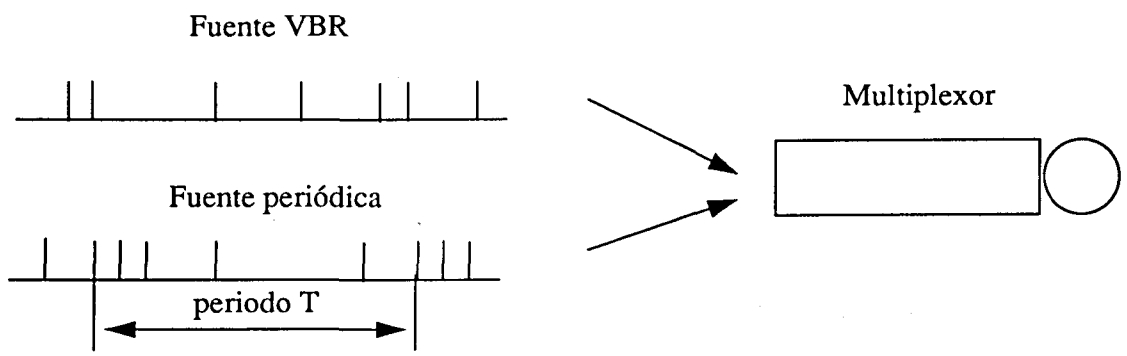


Figura 3.2

Tráfico generado por las fuentes periódicas.

El tráfico generado por una fuente periódica sigue una secuencia de emisión fija que se repite cada T ranuras. Este tráfico puede ser modelado como un D-BMAP cuya cadena de Markov subyacente es periódica con T estados (figura 3.3). Llamaremos b al número total de

celdas emitidas durante un periodo T.

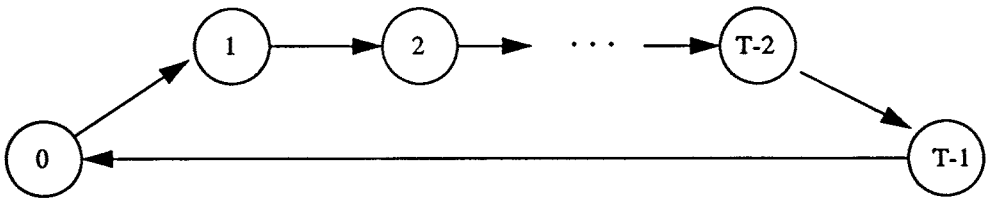


Figura 3.3

Si en las transiciones de un estado i al siguiente tenemos la emisión de k celdas definiremos el término $a_i(z)$ como:

$$a_i(z) = z^k$$

La matriz de tráfico de las fuentes periódicas tendrá ahora la siguiente forma:

$$D_p(z) = \begin{vmatrix} 0 & a_0(z) & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & a_1(z) & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & a_2(z) & \dots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \dots & 0 & a_{T-2}(z) \\ a_{T-1}(z) & 0 & 0 & 0 & \dots & 0 & 0 \end{vmatrix}$$

Para calcular los autovalores de la matriz $D_p(z)$, que llamaremos $x_p(k,z)$, debemos resolver:

$$|D_p(z) - x_p(k,z)\,I| = 0$$

Es fácil comprobar que la expresión del determinante es:

$$|D_p(z) - x_p(k,z)\,I| = (-1)^T x_p(k,z)^T + (-1)^{T-1}(a_0(z)\dots a_{T-1}(z)) =$$

$$= (-1)^T x_p(k,z)^T + (-1)^{T-1} z^b$$

de forma que tenemos T autovalores distintos (k = 0, 1,..., T-1):

$$x_p(k,z) = |z|^{\frac{b}{T}} e^{j\frac{b\varphi}{T}} e^{j\frac{2\pi b}{T}k}$$

en donde φ es el argumento de z. Las expresiones de los autovalores son independientes de la secuencia exacta de celdas emitidas durante un periodo y solo dependen del número total de celdas emitidas durante un periodo.

La expresión de las componentes de los autovectores por la izquierda, $u_p(k,z)$, es:

$$u_p^i(k,z) = u_p^0(k,z)\frac{x_p(k,z)^i}{a_0(z)\ldots a_i(z)}$$

mientras que para los autovectores por la derecha, $v_p(k,z)$, tenemos:

$$v_p^i(k,z) = v_p^0(k,z)\frac{a_0(z)\ldots a_i(z)}{x_p(k,z)^i}$$

Tráfico generado por una fuente de dos estados.

El tráfico generado por las fuentes VBR se modelará como la superposición del tráfico generado por fuentes independientes de dos estados. Sin embargo el método puede extenderse de forma inmediata al caso en donde las fuentes individuales tienen más de dos estados, pues solo necesitamos conocer los autovalores y autovectores de las matrices de tráfico de las fuentes individuales.

Supongamos que tenemos un D-BMAP de dos estados, cuya cadena de markov subyacente se muestra en la figura 3.4

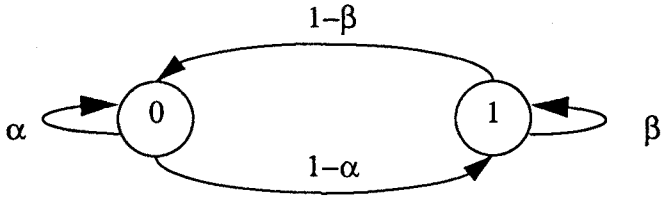


Figura 3.4

Cuando la fuente se encuentra en estado 0 no emite celdas. Cuando está en estado 1 en cada ranura emite k celdas con probabilidad d_k (k = 0, 1,...).

La matriz que nos da las probabilidades de emisión y de cambio de estado para cada fuente individual, $\mathbf{D}_{ind}(z)$, tiene entonces la forma:

$$D_{ind}(z) = \begin{bmatrix} \alpha & 1-\alpha \\ d(z)(1-\beta) & d(z)\beta \end{bmatrix}$$

en donde d(z) vale:

$$d(z) = \sum_{k=0}^{\infty} d_k z^k$$

Tenemos que los autovalores son:

$$x_{ind}(i,z) = \frac{\alpha + d(z)\beta}{2} \pm \sqrt{\left(\frac{\alpha + d(z)\beta}{2}\right)^2 - d(z)\phi} \qquad i = 0,1$$

siendo ϕ definida como:

$$\phi = \alpha + \beta - 1$$

Para los autovectores por la derecha encontramos:

$$u_{ind}(0,z) = \begin{bmatrix} \dfrac{x_{ind}(1,z) - \alpha}{x_{ind}(1,z) - x_{ind}(0,z)} \\ \dfrac{-d(z)\,(1-\beta)}{x_{ind}(1,z) - x_{ind}(0,z)} \end{bmatrix}$$

y

$$u_{ind}(1,z) = \begin{bmatrix} \dfrac{\alpha - x_{ind}(0,z)}{x_{ind}(1,z) - x_{ind}(0,z)} \\ \dfrac{d(z)\,(1-\beta)}{x_{ind}(1,z) - x_{ind}(0,z)} \end{bmatrix}$$

Por la izquierda tenemos

$$v_{ind}(i,z) = \begin{bmatrix} 1 & \dfrac{x_{ind}(i,z) - \alpha}{d(z)\,(1-\beta)} \end{bmatrix} \qquad i = 0,1$$

Autovalores y autovectores para la superposición

Si superponemos N fuentes de dos estados idénticas e independientes obtenemos un D-BMAP cuya cadena de markov subyacente tiene 2^N estados. Los autovalores de la nueva matriz de tráfico valen (el subíndice VBR se referirá al tráfico de la superposición de fuentes de dos estados):

$$x_{VBR}(i,z) = x_{ind}^{N-i}(0,z)\, x_{ind}^{i}(1,z)$$

mientras que los autovectores corresponden a los productos de Kronecker de los autovectores de la cadena de dos estados. Tenemos autovalores repetidos, de forma que los autovectores que corresponden a un mismo número de productos de Kronecker de los autovectores de la fuente individual (aunque evidentemente con distinto orden) tienen asociados los mismos autovalores.

Si agregamos los estados que tienen el mismo número de fuentes activas, obtenemos

una cadena con N+1 estados. Los autovalores son los mismos que los de la cadena de 2^N estados. En el apéndice C encontramos la deducción de las fórmulas de los autovectores correspondientes. Para la componente i-esima del autovector por la derecha asociado al autovalor k-esimo tenemos

$$u^i_{VBR}(k,z) = \frac{\sum_{k=ini}^{fin} \binom{k}{n}\binom{N-k}{i-n} (u^0_{ind}(0,z))^{N-k-i+n} (u^1_{ind}(0,z))^{i-n} (u^0_{ind}(1,z))^{k-n} (u^1_{ind}(1,z))^{n}}{\binom{N}{i}}$$

mientras que para el autovector por la izquierda:

$$v^i_{VBR}(k,z) = \sum_{k=ini}^{fin} \binom{k}{n}\binom{N-k}{i-n} (v^0(0,z))^{N-k-i+n} (v^1_{ind}(0,z))^{i-n} (v^0_{ind}(1,z))^{k-n} (v^1_{ind}(1,z))^{n}$$

con ini = max(0, i-N+k) y fin = min(k, i).

3.4 Modelo exacto para la probabilidad de pérdida de la superposición.

Daremos ahora un modelo exacto para el estudio de la superposición del tráfico generado por las fuentes periódicas y el tráfico generado por las fuentes VBR. Al ser estos dos tráficos independientes tenemos que la superposición es un D-BMAP caracterizado por la matriz:

$$D_T(z) = D_p(z) \otimes D_{VBR}(z)$$

Dicha matriz $D_T(z)$ tiene una dimensión de N*TxN*T y su forma es:

$$D_t(z) = \begin{vmatrix} 0 & a_0(z)D_{VBR}(z) & 0 & \dots & 0 & 0 \\ 0 & 0 & a_1(z)D_{VBR}(z) & \dots & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 & 0 \\ & & \cdot & & & \\ & & \cdot & & & \\ 0 & 0 & 0 & \dots & 0 & a_{T-2}(z)D_{VBR}(z) \\ a_{T-1}(z)D_{VBR}(z) & 0 & 0 & \dots & 0 & 0 \end{vmatrix}$$

Aplicando la metodología para el análisis de la cola D-BMAP/D/1 de capacidad finita expuesta en el capítulo 2, tenemos que, para encontrar los vectores de estado en los instantes de salida en estado estacionario, debemos resolver un sistema de ecuaciones que implica el almacenamiento de 2*S matrices de N*TxN*T elementos, siendo S el tamaño de la cola. Es fácil convencerse que esto supone unos requerimientos de memoria muy grandes cuando S, N b y T tienen valores realistas.

Estas matrices están, en gran medida, vacías, y podemos reordenar los estados de forma que reducimos las necesidades de almacenamiento. Sin embargo aun en ese caso las

necesidades de almacenamiento y cálculo se hacen enseguida intratables.

3.5 Comportamiento asintótico de la probabilidad de pérdida para el caso de la superposición de fuentes VBR con fuentes periódicas.

Hemos visto que un modelo exacto del problema es inabordable debido al gran número de estados que aparecen. En este apartado aplicaremos los resultados obtenidos en 3.2 para el caso particular de la superposición de fuentes periódicas y fuentes VBR.

Tal como se ha señalado anteriormente la matriz de tráfico $D_T(z)$ es el producto de Kronecker de las matrices $D_p(z)$ y $D_{VBR}(z)$, de forma que la descomposición espectral de $D_T(z)$ es:

$$D_T(z) = (U_p(z) X_p(z) V_p(z)) \otimes (U_{VBR}(z) X_{VBR}(z) V_{VBR}(z)) =$$

$$= (U_p(z) \otimes U_{VBR}(z)) (X_p(z) \otimes X_{VBR}(z)) (V_p(z) \otimes V_{VBR}(z))$$

y de ahí deducimos para C(z):

$$C(z) = (U_p(z) \otimes U_{VBR}(z)) (Y_p(z) \otimes Y_{VBR}(z)) (V_p(z) \otimes V_{VBR}(z)) =$$

$$= \sum_i \sum_k y_T(i,k,z)\, u_T(i,k,z)\, v_T(i,k,z)$$

con

$$y_T(k,z) = \frac{(z-1)\, x_T(k,z)}{z - x_T(k,z)}$$

$$x_T(i,k,z) = x_p(i,z)\, x_{vbr}(k,z) = |z|^{\frac{b}{T}} e^{j\frac{b\varphi}{T}} e^{j\frac{2\pi b}{T}i}\, x_{VBR}(k,z)$$

y

$$u_T(i,k,\dot{z}) = u_p(i,z) \otimes u_{VBR}(k,z)$$

$$v_T(i,k,\dot{z}) = v_p(i,z) \otimes v_{VBR}(k,z)$$

Los polos del sistema serán ahora las raíces de la ecuación (i: 0, ..., T-1):

de donde obtenemos (1 : 0, ..., T-b-1):

En el apéndice A de este capítulo se demuestra que hay T-b valores de i para los que la

$$z = |z|^{\frac{b}{T}} e^{j\frac{b\varphi}{T}} e^{j\frac{2\pi b}{T}i} x_{VBR}(k, z) \tag{3.6}$$

$$z = e^{j\frac{2\pi l}{T-b}} x_{VBR}^{\frac{T}{T-b}}(k, z) \tag{3.7}$$

ecuación (3.6) tiene una solución en el disco de radio unidad, mientras que para b valores de i no tenemos soluciones. Por lo tanto por cada valor de k, la transformada

$$x_T(i, k, z)$$

tiene un polo con módulo menor o igual a la unidad para T-b valores distintos de i que será denominado a partir de ahora como η_{ik}. Aquellas transformadas que no tienen polos tienen como transformadas inversas sucesiones que decrecen más rápidamente que una sucesión potencial, siendo por lo tanto despreciables en el desarrollo asintótico del sistema.

Para el cálculo de las soluciones de módulo menor que uno de (3.5) podemos usar varios métodos. En [Li90] se propone, para un problema similar, usar un método iterativo del tipo:

$$z_{n+1} = e^{j\frac{2\pi l}{T-b}} x_{VBR}^{\frac{T}{T-b}}(k, z_n)$$

En los resultados numéricos obtenidos usando este método hemos observado una rápida convergencia hacia la solución. Otra posibilidad ([BroSim88], [Li90]) es usar el desarrollo de Burman-Lagrange

$$z = \sum_{n=1}^{\infty} \frac{w^n d^{(n-1)} \Phi^n}{n! \, du^{n-1}}(0)$$

en donde

$$w = e^{j\frac{2\pi}{T-b}k}$$

y

$$\Phi(z) = x_u(z)^{\frac{T-b}{T}}$$

La convergencia está asegurada cuando el módulo de w es menor que uno. Para el caso que nos interesa, en donde el módulo de w vale 1, la convergencia depende de los valores de los parámetros del sistema.

Finalmente el polo dominante, ζ_0, puede ser hallado también mediante un método iterativo.

Para estudiar el comportamiento asintótico de la ocupación de la cola necesitamos resolver las ecuaciones:

$$\pi_\infty(0)\, e = 1 - \rho_{Tot}$$

$$\pi_\infty(0)\, u(i,k,\eta_{i,k}) = 0$$

El vector incógnita tiene T*N componentes, pero solo disponemos de (T-b)*N ecuaciones independientes. Sin embargo las componentes que corresponden a ranuras para las que para algún entero i se cumple que el número de celdas provenientes de fuentes periódicas que han llegado en las i ranuras anteriores es mayor o igual a i, deben valer cero. De esta forma solo tenemos N*(T-b) incógnitas, que pueden hallarse a partir de las ecuaciones anteriores.

Para encontrar la expresión asintótica de la probabilidad de pérdida en este caso no podemos usar la fórmula dada en general, en la que se había hecho la suposición de que teníamos tantos polos de módulo menor o igual a uno como estados tenía la fuente.

Sea x un vector cualquiera de N*T componentes. Definimos x^0 como aquel vector de N*(T-b) componentes resultante de eliminar las componentes de x en las que el vector $\pi_k(0)$ vale cero. Definiremos, de igual forma, la matriz X^0 como la resultante de eliminar las correspondientes filas y columnas de una matriz X de T*NxT*N.

Tenemos que ahora se cumple:

$$\pi_k^\circ(0)\, L_k^\circ = \theta^\circ$$

Siguiendo los mismos pasos que para el caso en que el número de polos de módulo menor o igual a uno es igual al número de estados de la fuente llegamos a la expresión final

$$P_t = \frac{(1-\rho_{Tot})^2 det(U_0^\circ(\zeta_0)\, V_0^\circ(\zeta_0))}{-\rho_{Tot} det(U_0^\circ(1)\, V_0^\circ(1))} y(0,\zeta_0)\, \zeta_0^{-k} + o(\zeta_0^{-k})$$

3.5 Comparación entre el modelo exacto y el modelo aproximado.

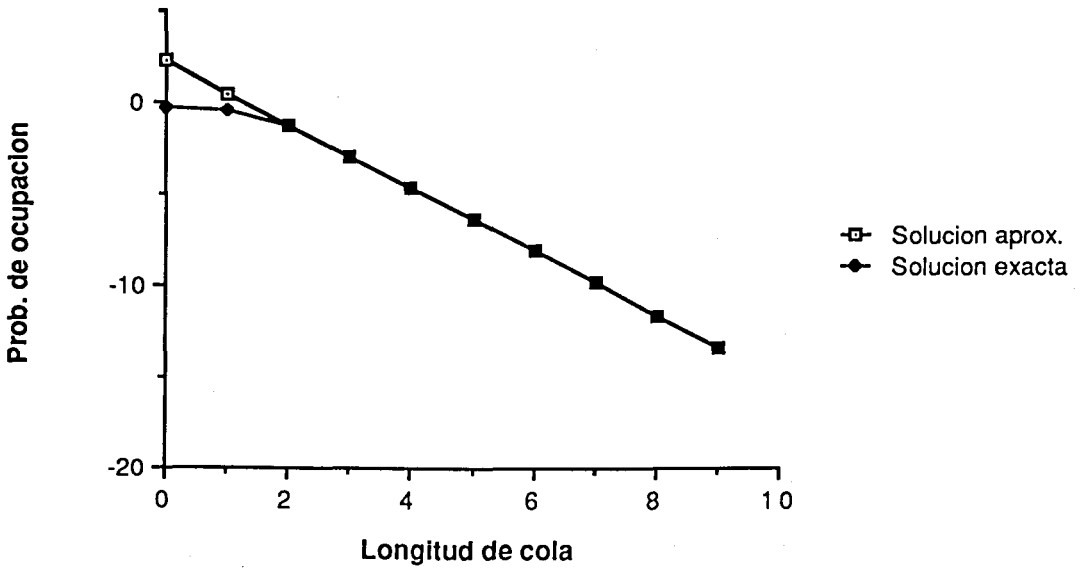


Figura 3.5 Probabilidades de ocupación en función de la longitud de cola en el caso de multiplexar una fuente CBR de 50 Mbps con tráfico proveniente de una fuente de dos estados con tiempos medios de permanencia en ambos estados de 3.5 segundos y tráfico de pico de 37.5 Mbps. La velocidad del multiplexor es de 150 Mbps.

A continuación haremos una comparación entre los resultados obtenidos con nuestro modelo exacto y con nuestro modelo aproximado. Para ello calcularemos las probabilidades de ocupación que se obtienen cuando multiplexamos una fuente de dos estados con tráfico periódico para una cola de longitud S = 10 (Figura 3.5). Observamos que incluso para colas pequeñas los resultados obtenidos con nuestra aproximación son extraordinariamente parecidos a los valores exactos.

Apéndice A

En primer lugar demostraremos lo siguiente:

Propiedad 1:

Sea D_i una sucesión de matrices no negativas con transformada D(z):

$$D(z) = \sum_i D_i z^i$$

y sea x(0,z) el autovalor de módulo máximo de la matriz D(z). Tenemos que se cumple:

$$|x(0,z)| \leq x(0,|z|)$$

Para demostrar esto usaremos un razonamiento similar al usado por J.F.C. Kingman ([Neu89]) para demostrar la convexidad del logaritmo de dicho autovalor para z real.

Sea G el conjunto de funciones g tales que:

$$|g(z)| \leq |g(|z|)|$$

Entonces si f y g pertenecen a G, también pertenece a G su suma y lo mismo sucede con su producto. Además si g_n son funciones pertenecientes a G tenemos que la función g definida como:

$$\lim_{sup} g_n = g$$

también pertenece a G. Los elementos (i, j) de la matriz D(z) son funciones pertenecientes a la clase G:

$$|D^{ij}(z)| = \left|\sum_k D_k^{ij}(z)\, z^k\right| \leq \sum_k |D_k^{ij}(z)\, z^k| = \sum_k D_k^{ij}(z)\, |z|^k = D^{ij}(|z|)$$

y sabemos que

$$x(0,z) = \lim_{sup} (traza(D^n(z)))^{\frac{1}{n}}$$

es decir, que x(0,z) se puede expresar como sumas, productos y límites de funciones de la clase G, por lo que pertenece también a dicha clase.

De lo anterior es inmediato ver que todas las soluciones de módulo mayor que 1 de la ecuación

$$z = x(k,z)$$

tienen un módulo mayor o igual que la solución obtenida para k = 0 y cuando z es real.

En el caso particular de la superposición de fuentes VBR y fuentes periódicas tenemos que resolver las ecuaciones:

$$z = |z|^{\frac{b}{T}} e^{j\frac{b\phi}{T}} e^{j\frac{2\pi b}{T}i} x_{VBR}(k,z)$$

que pueden ser transformadas en:

$$z = e^{j\frac{2\pi l}{T-b}} x_{VBR}^{\frac{T}{T-b}}(k,z)$$

Los autovalores para el tráfico VBR cumplen la desigualdad:

$$\left| x_{VBR}^{\frac{T}{T-b}}(k,z) \right| < \left| x_{VBR}^{\frac{T}{T-b}}(0,z) \right| \qquad k > 0$$

por lo que:

$$\left| x_{VBR}^{\frac{T}{T-b}}(k,z) \right| < \left| x_{VBR}^{\frac{T}{T-b}}(0,z) \right| < 1 \qquad k > 0, |z| = 1, z \neq 1$$

Si tomamos $z = 1 + \varepsilon e^{j\phi}$, $-\pi/2 < \phi < \pi/2$ tenemos:

$$\frac{\left| x_{VBR}^{\frac{1}{T-b}}(0,z) \right|}{|z|} = \frac{1 + \frac{T}{T-b} \left| x'_{VBR}(0,1) \right| \varepsilon \cos(\phi) + o(\varepsilon)}{1 + \varepsilon \cos(\phi) + o(\varepsilon)} =$$

$$= 1 + \left(\frac{T}{T-b}\left|x'_{VBR}(0,1)\right| - 1\right)\varepsilon\cos(\phi) + o(\varepsilon)$$

Si el sistema es estable (ρ_{VBR} es el tráfico generado por las fuentes VBR):

$$\rho_{orT} = \rho_{VBR} + \frac{b}{T} < 1$$

o lo que es lo equivalente:

$$\frac{T}{T-b}\rho_{VBR} < 1$$

con lo que tenemos que, para ε suficientemente pequeño, se cumple la desigualdad:

$$\left|x_{VBR}^{\frac{T}{T-b}}(k,z)\right| < \left|x_{VBR}^{\frac{T}{T-b}}(0,z)\right| < 1$$

en un contorno del tipo dibujado en la figura 3.A.1.

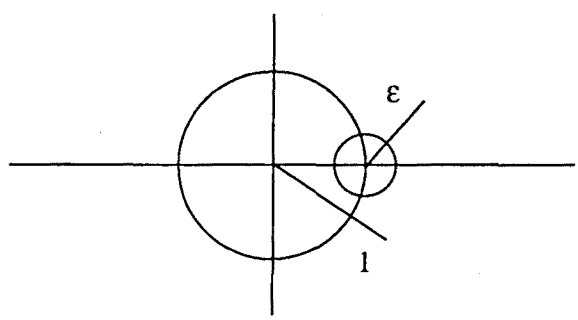


Figura 3.A.1

Aplicando el teorema de Rouche, vemos que nuestra ecuación tiene una única solución de módulo menor que uno en el caso k > 0, y que para k=0 tenemos una única solución en z = 1.

Apéndice B

A continuación veremos cómo calcular los residuos de L(z):

Sea α un polo de L(z), es decir, α es una raíz de la ecuación z=x(i,z) para un cierto autovalor x(i,z) de la matriz D(z). Para hallar el residuo en dicho polo debemos calcular:

$$Q(x) = \lim_{z \to \alpha} (1 - \frac{z}{\alpha}) L(z) = y(i, \alpha) u(i, \alpha) v(i, \alpha)$$

con

$$y(i, \alpha) = \lim_{z \to \alpha} (1 - \frac{z}{\alpha}) \frac{x(i, z)}{x(i, z) - z}$$

obteniendo

$$y(i, z) = \frac{1}{1 - \frac{d}{dz} x(i, z)}$$

Sea ρ el tráfico de entrada a la cola y θ el vector de probabilidad en estado estacionario de la cadena de Markov subyacente a la definición del D-BMAP. En [Neu89] se demuestra que:

$$\frac{d}{dz} x(0, 1) = \rho$$

y que se pueden escoger los vectores u(0,1) y v(0,1) de forma que.

$$u(0, 1) = e$$

$$v(0, 1) = \theta$$

con lo que para el residuo en el polo α=1 tenemos:

$$Q(1) = \frac{1}{1 - \rho} e\theta$$

Ahora calcularemos las derivadas de los autovalores. Tenemos:

$$D(z) u(i, z) = x(i, z) u(i, z)$$

derivando con respecto a z:

$$D'(z) u(i, z) + D(z) u'(i, z) = x'(i, z) u(i, z) + x(i, z) u'(i, z)$$

multiplicando por v(i,z) por la derecha:

$$v(i,z)\,D'(z)\,u(i,z) + v(i,z)\,D(z)\,u'(i,z) = x'(i,z)\,v(i,z)\,u(i,z) + x(i,z)\,v(i,z)\,u'(i,z)$$

de aquí

$$v(i,z)\,D'(z)\,u(i,z) + x(i,z)\,v(i,z)\,u'(i,z) = x'(i,z) + x(i,z)\,v(i,z)\,u'(i,z)$$

y finalmente

$$v(i,z)\,D'(z)\,u(i,z) = x'(i,z)$$

Apéndice C.

Cuando superponemos N fuentes independientes de 2 estados obtenemos una cadena de Markov de 2^N estados. Cada uno de estos estados puede ser visto como un vector de N componentes cuya componente i-ésima indica el estado de la fuente correspondiente. e[j,k] será un vector con j fuentes activas. El número de vectores con j fuentes activas vendrá dado por las combinaciones de N elementos tomados de j en j.

Si agregamos todos los estados con el mismo número de fuentes activas (es decir, los vectores e[j,k] con misma j), obtendremos una cadena de Markov de N+1 estados. En general, el estado resultante de agregar los estados e[j,k] se denominará estado j.

En primer lugar encontraremos relaciones entre las probabilidades de transición entre estados agregados y sin agregar. La propiedad clave que usaremos es la siguiente:

Propiedad 1.

Sea T una permutación entre elementos de un vector de N componentes. Entonces, la probabilidad de pasar de un estado cualquiera e[i,n] a otro estado e[j,k], p(e[j,k]|e[i,n]), cumple:

$$p(e[j,k] \mid e[i,n]) = p(T(e[j,k]) \mid T(e[i,n]))$$

La demostración de que lo anterior es cierto es muy sencilla, pues basta considerar que una permutación de los elementos del vector e[j,k] equivale de hecho a una renumeración de las fuentes.

De aquí se deduce también de forma inmediata que para cualquier valor de k:

$$\sum_n p(e[j,k] \mid e[i,n]) = \sum_n p(e[j,0] \mid e[i,n])$$

Apliquemos la permutación T tal que T(e[j,k]) = e[j,0] a todos los vectores e[i,n] del sumatorio. Evidentemente los términos del segundo sumatorio no son más que los del primero cambiados de orden, y de ahí obtenemos la igualdad.

Puesto que todos los sumatorios en n del primer miembro son iguales deducimos que:

$$\sum_k \sum_n p(e[j,k] \mid e[i,n]) = \binom{N}{j} \sum_n p(e[j,0] \mid e[i,n]) = \sum_n p(j \mid e[i,n])$$

Por último tenemos

Propiedad 2.

$$\sum_n p(e[j,0] \mid e[i,n]) = \frac{\binom{N}{i}}{\binom{N}{j}} p(j \mid i)$$

Basta con notar que:

$$\sum_n p(j|\, e[i,n]) = \sum_n \frac{p(j, e[i,n])}{p(e[i,n])} = \binom{N}{i} \sum_n \frac{p(j, e[i,n])}{p(i)} = \binom{N}{i} (p(j|\, i))$$

Pasaremos ahora a encontrar relaciones entre los autovalores y autovectores de las matrices para los procesos sin agregar y agregado.

Sea $D_{sa}(z)$ la matriz para el proceso sin agregar. La emisión de las celdas no depende del estado futuro, sino solo del número de fuentes activas del estado actual. Por lo tanto:

$$D(z)_{sa}^{e[i,n],e[j,k]} = d(z,i)\, p(e[j,k]|\, e[i,n])$$

en donde d(z,i) es una función que solo depende de z y de i.

Para la matriz $D_a(z)$, que corresponde a la cadena en donde hemos agregados los estados con igual número de fuentes activas tenemos:

$$D(z)_a^{i,j} = d(z,i)\, p(j|\, i)$$

Sea H un autovector por la derecha de la matriz $D_{sa}(z)$, asociado al autovalor x(z). Numeraremos sus componentes según los estados de la cadena. Evidentemente se cumple:

$$\sum_j \sum_k d(z,i)\, p(e[j,k]|\, e[i,n])\, H_{e[j,k]}(z) = x(z)\, H_{e[i,n]}(z)$$

Sumando las anteriores ecuaciones obtenemos:

$$\sum_n \sum_j \sum_k d(z,i)\, p(e[j,k]|\, e[i,n])\, H_{e[j,k]}(z) = x(z) \sum_n H_{e[i,n]}(z)$$

cambiando de orden los sumatorios

$$\sum_j \sum_k \sum_n d(z,i)\, p(e[j,k]|\, e[i,n])\, H_{e[j,k]}(z) = x(z) \sum_n H_{e[i,n]}(z)$$

y aplicando la propiedad 2

$$\sum_j d(z,j)\, p(j|\, i) \frac{\sum_k H_{e[j,k]}(z)}{\binom{N}{j}} = x(z) \frac{\sum_n H_{e[i,n]}(z)}{\binom{N}{i}}$$

Tenemos N+1 autovalores distintos, por lo que deducimos que los autovalores de la matriz

$\mathbf{D_a}(z)$ son los mismos que los de la matriz $\mathbf{D_{sa}}(z)$. Si h es un autovector por la derecha de la matriz $\mathbf{D_a}(z)$ asociado al autovalor x(z), sus componentes cumplen:

$$hj = \frac{\sum_k H_{e[j,k]}(z)}{\binom{N}{j}}$$

Usando un razonamiento totalmente análogo para los autovectores por la izquierda vemos que si M es un autovector por la izquierda asociado a x(z) de la matriz $\mathbf{D_{sa}}(z)$ y m lo es para la matriz $\mathbf{D_a}(z)$ sus componentes cumplen:

$$mj = \sum_k M_{e[j,k]}(z)$$

La deducción de la fórmula final para $u_T(k,z)$ y $v_T(k,z)$ es ahora inmediata.

Capítulo 4

El control de la congestión en redes ATM

4.1 Cuestiones generales [Boy90],[Hui88].

En una red ATM los conmutadores deben servir un tráfico que procede de la multiplexación de fuentes independientes y de características dispares. Dicho tráfico fluctúa con el tiempo como consecuencia de las variaciones del tráfico de las fuentes (aparición de ráfagas de celdas [Sol91]), de la admisión de nuevas llamadas, o de la desconexión de llamadas ya establecidas.

Estas fluctuaciones pueden ser absorbidas por los conmutadores sin que se produzcan pérdidas de información siempre y cuando el tráfico instantáneo de entrada supere la capacidad de tratamiento del nodo durante sólo breves periodos de tiempo. Sin embargo, cuando estas situaciones de sobrecarga se dan durante periodos de tiempo demasiado largos, se produce el desbordamiento de las colas de espera en los conmutadores, con la consiguiente pérdida de celdas y aumento de los retardos.

En una red ATM es fácil que se produzcan estas situaciones de congestión, a no ser que se adopte alguna medida de control: El acceso del usuario se produce a gran velocidad (150 Mbps), el rango de velocidades de transmisión es muy amplio y las fuentes actúan con gran autonomía.

De hecho no es realista pensar que exista un control de congestión que elimine por completo la posibilidad de que se produzcan pérdidas de celdas (salvo, tal vez, al precio de tener un uso muy ineficiente de los recursos de la red), por lo que las pérdidas de celdas deben estar contempladas en las especificaciones de la red. Este punto es estudiado más en profundidad en el capítulo dedicado a la calidad de servicio en las redes ATM.

En las redes convencionales de conmutación de paquetes deben también proveerse métodos de control de congestión. Sin embargo, debido a la diferencia de velocidades de transmisión, los métodos que son válidos para las redes de transmisión de datos dejan de serlo para el caso de redes ATM, por lo que deben pensarse nuevas formas de control de congestión.

En este tipo de redes deben tomarse medidas preventivas. Por eso, en la mayoría de esquemas para el control de congestión que han sido propuestos, el control de admisiones de nuevas conexiones (CAC, 'Conexion Acceptace Control') juega un papel esencial: Cuando un nuevo usuario quiere acceder a la red debe hacer una descripción de las características del tráfico que va a generar, de forma que el mecanismo de CAC pueda decidir si el nuevo usuario va o no a causar un nivel de congestión excesivo.

Es, por lo tanto, esencial tener un mecanismo mediante el cual la red controle si el tráfico que es realmente emitido por los usuarios que han sido admitidos corresponde a la descripción que han hecho del mismo: es lo que se conoce como mecanismo de función de policía o de vigilancia ('Policing Function').

Cuando analizamos un mecanismo de control de congestión debemos hacernos las siguientes preguntas:

- ¿ Evitará dicho mecanismo la congestión de la red?.
- ¿ Va a ser capaz un usuario de proporcionar los parámetros que se le exigen en el momento de la admisión de la nueva comunicación?.
- ¿ Es posible hacer una vigilancia efectiva de dichos parámetros?.
- Si se usa dicho mecanismo de control, ¿ Tendremos un uso eficiente de los recursos

de la red?.

Como veremos la mayor parte de los mecanismos de control propuestos presentan problemas a la hora de responder las anteriores preguntas. Podemos concluir, pues, que en la actualidad el problema de definir un control de congestión eficiente para una red ATM aún no ha sido resuelto.

4.2 El control de admisión de conexiones (CAC, 'Conexion Acceptance Control') [Boy90], [RAC90]

El CAC es un procedimiento que decide si una nueva conexión que quiere establecerse puede ser o no aceptada por la red. Esta decisión debe ser hecha a partir de los recursos ocupados por las otras conexiones ya establecidas, de los recursos que necesitará la nueva conexión, y de la calidad de servicio requerida por los usuarios de la red. El usuario debe indicar a la red, mediante mensajes de señalización:

- Las características del tráfico emitido por la fuente.
- El grado de servicio requerido,
- La dirección del destinatario (o destinatarios) de la comunicación.

Se debe establecer entonces un procedimiento que, basándose en los datos sobre el tráfico de la red y los proporcionados por el usuario que quiere acceder, decida si el nivel de congestión que introduciría la nueva fuente es o no excesivo.

Se han propuesto varios procedimientos para el CAC. Entre ellos cabe citar los siguientes:

Reserva del tráfico de pico

Los usuarios deben especificar el tráfico de pico de su comunicación. La nueva comunicación es admitida si la suma de tráficos de pico de las conexiones ya establecidas y el de la nueva comunicación es menor a la capacidad de transmisión del canal. Un control de congestión basado solamente en un CAC de este tipo asegura que nunca se producen congestiones, es de realización muy sencilla y sólo requiere determinar y vigilar el valor del tráfico de pico de la comunicación. Sin embargo, cuando el tráfico producido es variable, la utilización del canal se haría demasiado baja, pues dependiendo del tipo de fuente pueden conseguirse valores muy elevados de ganancia de multiplexación [Lou91].

'Linear Connection Acceptance Control'

En este método se asigna a cada fuente un ancho de banda efectivo, que representa el ancho de banda requerido para transportar el tráfico generado por la fuente sobre el enlace de transmisión con un determinado nivel de probabilidad de pérdida. Este ancho de banda efectivo se calcula a partir de las características de la fuente, del ancho de banda del enlace y de la probabilidad de pérdida admisible.

Una nueva conexión es admitida cuando el ancho de banda efectivo del resto de conexiones ya establecidas sumado al ancho de la nueva fuente no excede el ancho de banda del enlace.

'Two Moment Allocation Scheme'

En este método se parte de la suposición de que:

- El ancho de banda de una conexión se puede caracterizar a partir de su media y varianza.
- La distribución del ancho de banda total requerido por el conjunto de todas las llamadas que comparten un enlace puede ser aproximada mediante una distribución Gaussiana.

Partiendo de estas hipótesis, y con sólo conocer los dos momentos del ancho de banda de cada fuente (que caracterizan por completo la distribución gaussiana), se calcula el ancho de banda total requerido. Cuando el resultado es tal que garantiza la calidad de servicio de todos los usuarios, se admite la nueva comunicación.

'Source Independent CAC procedures'

En [RasSor89] se da una cota superior de la probabilidad de bloqueo que causa una fuente, basándose en el conocimiento de su tráfico de pico y su tráfico medio, de forma que se usan estos dos parámetros para determinar si una nueva fuente es admitida o no en la red.

Partiendo de este resultado, se calcula el ancho de banda total requerido por las conexiones ya establecidas y por la nueva conexión que quiere establecerse. Cuando el resultado cumple las especificaciones necesarias para garantizar la Calidad de Servicio de todos los usuarios, la nueva comunicación es admitida.

Estos mecanismos de control de congestión basados en un CAC presentan graves problemas:

- Tal como se ha señalado, el procedimiento de reserva del tráfico de pico es demasiado pesimista: En muchos casos el tráfico generado por las fuentes tiene un comportamiento estadístico bien conocido y no presenta fluctuaciones pronunciadas (por ejemplo, es el caso de muchos codificadores de video). De forma que la congestión que causan estas fuentes puede ser bien predicha sin necesidad de hacer una reserva de tráfico de pico, es decir, se puede conseguir una gran **ganancia de multiplexación** [Lou91], [GarCas91.b]. Ahora bien, para otros tipos de fuentes, en los que el comportamiento estadístico no es bien conocido o en los que la variabilidad del tráfico es tan grande que no se puede obtener ganancia de multiplexación, se debería hacer una reserva del tráfico de pico.

- Los tres últimos esquemas soslayan la inexistencia de modelos de tráfico para muchos tipos de fuente, demandando una serie de parámetros de tráfico a la fuente, que son utilizados para caracterizar la superposición del tráfico total. Sin embargo, en muchos casos, el usuario no podrá precisar los parámetros necesarios. Por ejemplo, **el tráfico medio es un parámetro que en muchos casos no se puede conocer a priori**, y además, el exigir a un usuario que fije cuál va a ser el tráfico medio de la conexión, puede suponer una violación de la libertad de comunicación (se impone restricciones al usuario no sobre el tipo de servicio que va a usar, sino sobre qué uso va a hacer de dicho servicio).

En vista de lo anterior se concluye que, si queremos tener una utilización eficiente de los recursos de la red, no hay un procedimiento de control de congestión aplicable a cualquier tipo de fuente.

'Source Dependent CAC procedure'

Recientemente ([Boy90], [Lou91], [Tra91], [BoyTra91]) se ha propuesto el uso de un control de congestión que diferencia entre clases de fuentes. Los terminales deben dar a la red información sobre:

- Identificación del terminal.
- Tipo de fuente de tráfico.
- Señales de activación y desactivación de los terminales.
- Valor del tráfico de pico.

Las fuentes se agrupan en cuatro clases:

- Clase A: Son las fuentes CBR.

- Clase B: Fuentes esporádicas, para las que se puede distinguir entre varios estados de actividad. El cambio de estado se produce por iniciativa de la fuente. Sin embargo la

fuente acepta la eventualidad de que deba esperar antes de cambiar de estado ('Negociated Stepwise VBR sources'). Su tráfico de pico debe ser pequeño (del orden de pocos Mbps).

- Clase C: Fuente VBR pero con un variabilidad reducida y un tráfico de pico también reducido. El comportamiento estadístico de estas fuentes debe estar bien caracterizado.

- Clase Z: Fuentes que no pertenecen a ninguna de las clases anteriores.

El control de congestión actúa diferenciando la clase a que pertenece cada fuente. Para las clases A y Z se debe usar una reserva del tráfico de pico. Las fuentes de clase C pueden ser controladas simplemente mediante un mecanismo de CAC: Al ser conocidas sus características estadísticas, se puede calcular de antemano qué congestión van a causar en la red [Lou91]. Las congestión causada por las fuentes de clase B se controla, además de mediante el uso de un CAC, usando un protocolo que negocia el acceso de la fuente a la red cada vez que se produce un cambio de estado (FRP, 'Fast Reservation Protocol') [BoyTra91], [Tra91].

El FRP se basa en el envío, por parte del terminal, de una celda de señalización cada vez que se produce un cambio de estado de la fuente. La red debe decidir entonces si el nuevo tráfico va a causar una congestión aceptable y responder dando o no dando al terminal permiso de acceso a la red.

4.3 La función de policía. [BoyGui91], [RatThe90] [Sanetal90]

Sea cual sea el mecanismo de control de congestión que se use, se debe proveer de un mecanismo que se encargue de vigilar que los parámetros y características que han sido dados por la fuente que ha establecido una nueva conexión no cambien a lo largo de la comunicación, pues de otro modo deberíamos tomar alguna medida para evitar la congestión de la red. Este mecanismo se engloba dentro de la función de policía (también llamada de vigilancia).

Se han propuestos una serie de mecanismos [Tur86], [Lel89] en los que el tráfico generado por la fuente no es modificado, salvo que se detecte una violación de los parámetros contratados. En este caso se toman medidas encaminadas a evitar la congestión de la red, que pueden ser, por ejemplo, eliminar las nuevas celdas que llegan hasta que las condiciones del contrato vuelven a ser cumplidas, o marcar estas nuevas celdas con un bajo nivel de prioridad [Ecketal89].

Entre estos mecanismos tenemos el 'Leaky Bucket' (LB), el 'Moving Window', el 'Jumping Window' y 'Exponential Weighted Moving Average'. Son conocidos, genéricamente, como mecanismos de 'pick-up'.

Los mecanismos de 'pick-up' han recibido numerosas críticas, pues parece que no son suficientes para eliminar el riesgo de congestión (ver apartado 4.4). Se han propuesto otro tipo de mecanismos entre los que destaca el 'Spacer-Controller' [Boy90], en donde se modifica la secuencia de celdas generadas por el terminal, forzando así el cumplimiento de las condiciones del contrato de tráfico ('Traffic Shaphing').

A continuación haremos una descripción de estos mecanismos de policía [Sanetal90], [Boy90]:

'Leaky Bucket'

El mecanismo de 'Leaky Bucket' fue propuesto en [Tur86] como una candidato para llevar a cabo la función de policía en una red ATM. Su funcionamiento se basa en simular, mediante un contador, una cola con tiempo de servicio determinista, que se llena con el flujo de celdas emitido por la fuente. Cuando la cola está llena, y se produce la llegada de celdas procedentes de la fuente que está siendo controlada, se toma alguna medida de control. Por ejemplo estas celdas se pierden, o se marcan con un bajo nivel de prioridad.

Aunque el funcionamiento del mecanismo LB puede ser modelado mediante una cola

de espera, es importante remarcar que el flujo de celdas no es alterado por este mecanismo, salvo cuando se activan las medidas de control. Los parámetros que permiten controlar el tráfico son el tiempo de servicio de dicha cola, es decir, el tiempo de decrementación del contador, y la longitud de la cola de espera, es decir, el número de créditos del contador.

El 'Leaky Bucket' puede ser usado para controlar el tráfico medio emitido por la fuente y el tráfico de pico. Para hacerlo los parámetros que definen el mecanismo deben ser dimensionados de forma adecuada.

'Moving Window'

Este mecanismo permite la llegada de un número determinado de celdas en un tiempo arbitrario T. Va sacando celdas del sistema de policía una a una, un tiempo fijo T después de la celda aceptada más antigua. Se debe, pues, almacenar los instantes de tiempo en que se producen las llegadas de celdas.

'Jumping Window'

El mecanismo es muy parecido al de 'Moving Window', con la diferencia de que los instantes de tiempo T no tienen relación con los instantes de llegada de celdas. Es decir, se divide el tiempo en intervalos de longitud T y no se permite que el número de celdas que llega dentro de cada intervalo supere un cierto valor.

'Exponentially Weighted Moving Average'

Es una versión del 'Jumping Window' en donde se tiene en consideración lo sucedido durante los intervalos de tiempo previos. En este algoritmo, el número de celdas aceptado en cada intervalo se determina mediante la siguiente fórmula:

$$S_n = (1-\alpha)X_n + \alpha S_{n-1}$$

En donde X_n es el número de celdas ya aceptadas en el intervalo n-ésimo y α es una constante menor que 1. En cada intervalo debemos cumplirse que S_n sea menor que un valor S_{max}. Si se toma $\alpha = 0$ tenemos simplemente el mecanismo de 'Jumping Window'.

El 'Spacer-Controller'

Como alternativa a los problemas que presentan los mecanismos de 'pick-up' (que son tratados en la sección 4.4), se ha propuesto un nuevo mecanismo [Boy90], conocido como 'Spacer-Controller' para ejercer la función de policía. Conceptualmente, este mecanismo se comporta como un LB dimensionado para controlar el tráfico de pico, pero con la diferencia de que se provee de un lugar físico de espera para las celdas. De este modo, se actúa de forma activa sobre las características del tráfico que está siendo controlado, asegurando que la distancia mínima entre celdas es la negociada en el momento de establecerse la conexión.

Ahora la longitud de la cola debe escogerse de forma que un usuario que emite según su contrato no pierda celdas debido a las variaciones en la distancia entre celdas introducidas por las etapas de multiplexación.

4.4 La eficiencia de los mecanismos de 'pick-up' para el control de la congestión.

Recientemente ([BoyGui91]) se han realizado serias críticas al uso de mecanismos de 'pick-up' para la realización de la función de policía de una red ATM. El problema que presentan estos mecanismos es que pueden ser 'engañados' de forma relativamente fácil: Un usuario puede pactar un cierto tipo de conexión y después modificar los parámetros del tráfico, de

forma que la violación del contrato no es detectada pudiendo causar congestión en la red. Tomaremos como ejemplo el mecanismo de 'pick-up' más popular, el 'Leaky Bucket'. Sin embargo las conclusiones que obtendremos se pueden extender al resto de mecanismos.

El dimensionamiento del mecanismo LB consiste esencialmente en fijar los tiempos de decrementación y el número de créditos. Si queremos controlar el tráfico medio, se toma como tiempo de decrementación la distancia media entre llegadas. De esta forma, si una fuente emite con un tráfico medio mayor que el contratado, el mecanismo LB se comportará como una cola con una carga mayor que 1, es decir, será un sistema inestable, y una gran parte de las celdas emitidas serán perdidas.

Sin embargo, aparece ahora un grave problema: Si queremos que probabilidad de pérdida causada por el mecanismo de policía en las fuentes que respetan el contrato sea muy baja, el tamaño de la cola (osea, el número de créditos del contador) deberá ser muy grande [RatThe90] por lo que:

- Tardaremos mucho en detectar una violación del contrato.
- Se permite el acceso a la red de largos trenes de celdas ('clusters') emitidos a la máxima velocidad del enlace, sin que sean cortados.

Una posible solución a estos problemas es sobredimensionar el tiempo de servicio de la cola. De esta forma, para el tráfico pactado, la carga del sistema será menor que 1 y se pueden conseguir probabilidades de pérdida aceptables con menores longitudes de cola. Sin embargo, esto también provocará que fuentes con un tráfico medio superior al pactado, no sean recortadas y puedan acceder a la red.

Pondremos un ejemplo tomado de [RatThe90]:

Si el tráfico producido por la fuente es un proceso de Poisson, queremos que el nivel de pérdida introducido por la función de policía sea menor que 1.0 E-10, y dimensionamos el tiempo de decrementación como el 90% del tiempo medio entre llegadas, necesitamos un contador con 110 créditos. Si el tiempo de decrementación fuese del 80%, harían falta 50 créditos. Cuando se toma un tráfico más variable, los valores de los créditos requeridos son incluso mayores.

Si se quiere controlar el tráfico de pico, el dimensionamiento del LB es diferente. Ahora el tiempo de decrementación del contador se debe tomar igual a la distancia mínima entre celdas, y el número de créditos del contador debe ser igual a 0. De esta forma sólo se perderían las celdas que violen el contrato.

Sin embargo, aparece otro problema: Por cuestiones de seguridad, el mecanismo LB deberá estar situado, como mínimo, detrás del interfaz Tb. Ello quiere decir que cuando las celdas de una comunicación llegan al punto en donde son controladas, ya han sido multiplexadas con tráfico de otros terminales (por ejemplo, en el TR_2). Durante este proceso de multiplexación se producen perturbaciones en el tráfico [Rob89] lo que pueden provocar que celdas emitidas con una separación mayor que la mínima pactada, sean detectadas como celdas violadoras del contrato por la función de policía. Por lo tanto, una vez más debemos dar un número de créditos al contador que puede ser muy elevado.

En [BoyGui90] se da un ejemplo de dimensionamiento del mecanismo de Leaky Bucket para el control del tráfico de pico cuando se producen perturbaciones en la multiplexación en el TR_2. El tráfico ya perturbado se modela mediante un proceso de Erlang, y se pide que la probabilidad de pérdida introducida por la función de policía sea menor a 1.0 E-10. De sus cálculos se desprende un dimensionamiento de unos 40 créditos.

¿Que efectos tiene sobre la congestión de los conmutadores este sobredimensionamiento del número de créditos?. En [BoyGui90] se estudia el caso siguiente: Un usuario negocia una comunicación como si el tráfico que va a emitir fuera tráfico CBR a una velocidad μ. El LB está dimensionado con b créditos. El usuario entonces emite un tráfico a ráfagas, consistente en b celdas consecutivas, seguidas de un silencio de longitud $b(\Lambda/\mu-1)$, en donde Λ es la máxima capacidad del enlace. Este tipo de tráfico puede atravesar el RT_2 sin

sufrir excesivos cambios, atravesar el punto de vigilancia sin ser detectado, y llegar a los conmutadores, con el consiguiente peligro de congestión (figura 4.1).

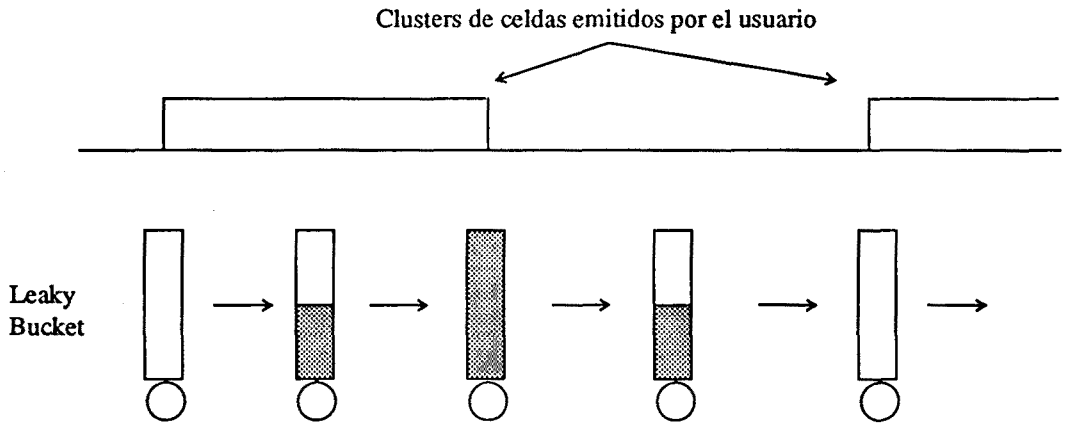


Figura 4.1

En [BoyGui90] no se da un modelo analítico que permita obtener, en condiciones realistas, el nivel de congestión que introduce este tráfico. Sin embargo, mediante simulación, se estudia un caso en donde los niveles de probabilidad de requeridos pérdida son mayores, obteniendo que dicho tráfico puede empeorar considerablemente las probabilidades de pérdida de los demás usuarios. Así supone una cola con capacidad para 16 celdas, el tráfico que modela al resto de usuario es Poisson con un valor medio de 0.81, y un nuevo usuario negocia el establecimiento de una conexión de tráfico CBR con una carga de 0.04. Suponiendo que este tráfico CBR con el resto de usuarios se comporta como tráfico de Poisson, la probabilidad de pérdida obtenida es de 1.18 E-3. Si b=10 la probabilidad de pérdida es de 1.2 E-2. Cuando b=20, tenemos 2.4 E-2 y si b=30, la probabilidad de pérdida es de 3.0 E-2. Vemos, pues, que incluso para valores de b pequeños, el valor de la probabilidad de pérdida empeora considerablemente.

A continuación vamos a usar el modelo desarrollado en el capitulo 3 para estudiar este fenómeno en condiciones de tráfico realistas. Para ello modelaremos el tráfico de entrada de un multiplexor mediante la superposición de dos fuentes de dos estados, siendo el tráfico de pico de cada fuente de 40.2 Mbps y los tiempos medios de permanencia en los estados de actividad y de inactividad de 700 msec. Un usuario que emite un tráfico CBR a 50 Mbps quiere establecer una nueva conexión. El multiplexor trabaja a 150 Mbps, siendo el tamaño de cola de 32 celdas. Supondremos que la máxima probabilidad de perdida admisible es de 1.0 E-10.

En las anteriores condiciones la probabilidad de pérdida en el multiplexor es de 9.021 E-11, por lo que esa nueva conexión podría ser admitida en la red.

Supondremos que el nuevo usuario está siendo controlado con un Leaky Bucket dimensionado con 10 celdas (como se ha visto este es un caso muy optimista). El nuevo usuario aprovecha esta circunstancia para emitir 'clusters' de b celdas consecutivas seguidas de 2b silencios, de forma que su tráfico medio emitido sigue siendo de 50 Mbps.

En la figura 4.2 vemos cómo se altera la probabilidad de pérdida total en el multiplexor en función de los valores de b. Observamos que incluso en este caso que puede ser considerado como favorable, la probabilidad de pérdida se ve considerablemente afectada, aumentado casi en un orden de magnitud. Es de esperar que con tamaños de b mayores el efecto sería aún más pronunciado

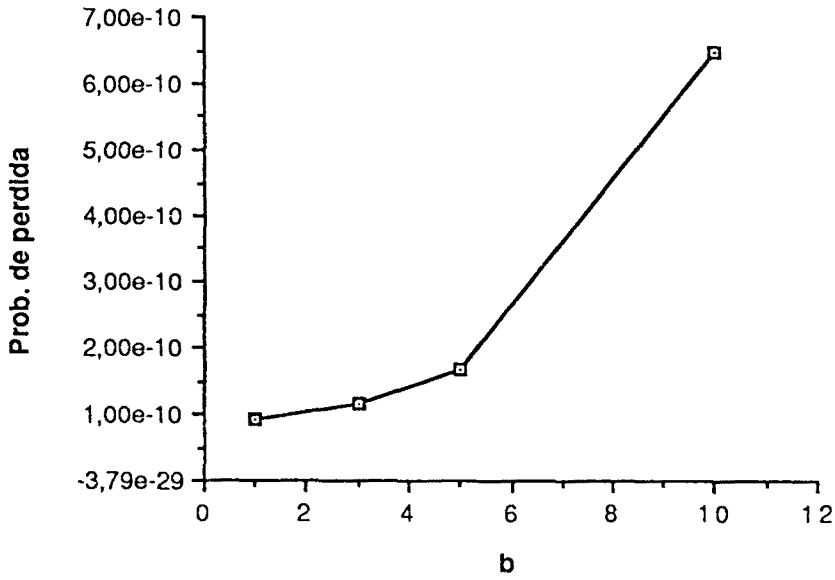


Figura 4.2

En la figura 4.3 se muestra cómo varía la distribución de la ocupación de la cola para el caso b = 20. Como es de esperar de los resultados del capítulo anterior las dos distribuciones son paralelas (el polo dominante sólo depende del valor medio del tráfico) aunque los valores de la ocupación para valores grande de cola están aumentados en casi dos órdenes de magnitud.

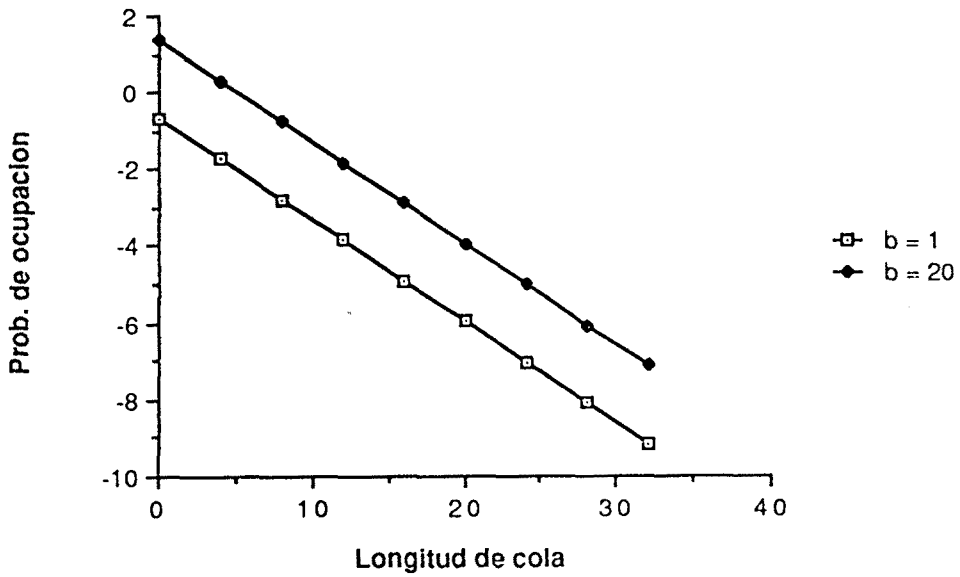


Figura 4.3

Capítulo 5

Calidad de servicio y prioridades en redes ATM

5.1 Cuestiones generales.

La calidad de servicio es definida por el CCITT como el efecto global de las características de servicio que determinan el grado de satisfacción de un usuario. El grado de servicio (GDS) de una red es un conjunto de variables de tráfico que permiten tener una medida de la aptitud de la red para satisfacer una cierta calidad de servicio a sus usuarios.

Los dos parámetros más importantes a la hora de determinar el GDS de una red ATM durante la fase de transmisión de la información son la probabilidad de pérdida de una celda y el retardo que sufre una celda al atravesar la red [YokKod89], [WooKos90], [Kroetal91]. La misma red es compartida por usuarios de diferentes servicios, para los que los valores mínimos necesarios del GDS son distintos. Este hecho es característico de las redes integradas y es en el que centraremos nuestra discusión.

El retardo que sufre una celda al atravesar una red ATM es debido al tiempo de ensamblado de la celda, al retardo de propagación, al retardo de transmisión, al tiempo de espera en las colas y al tiempo de compensación de la variación del retardo ('Jitter').

En los servicios interactivos en tiempo real, como es el caso de las comunicaciones de voz y video interactivo, una celda tiene fuertes restricciones en el retardo que puede sufrir al atravesar la red (figura 5.1). Para las redes de conmutación de paquetes este retardo es variable y no es conocido de antemano. Este hecho ha sido clave para impedir la introducción de servicios interactivos en tiempo real en redes de conmutación de paquetes de baja velocidad.

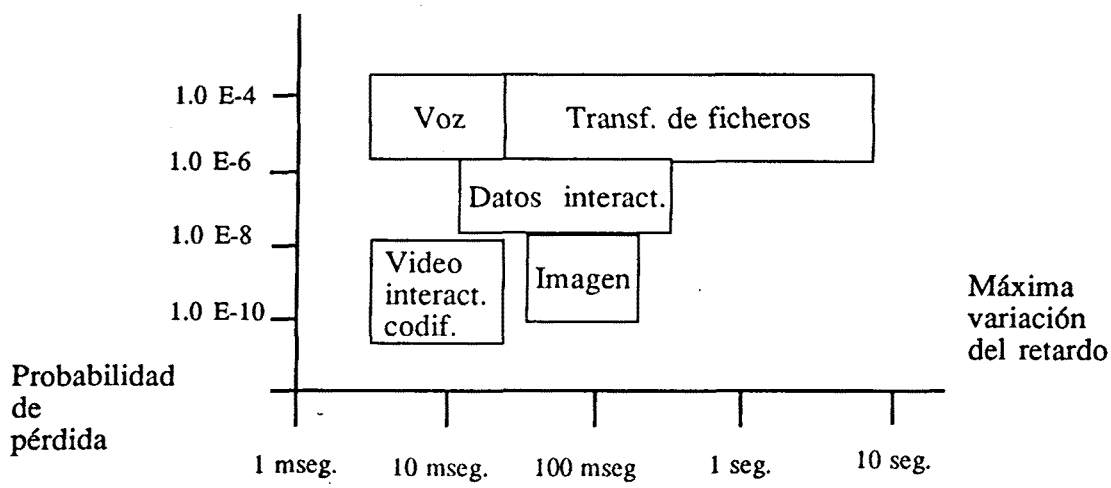


Figura 5.1

Sin embargo, la cuestión deja de ser crítica cuando la velocidad de transmisión es tan elevada como es en el caso de redes ATM: el retardo de propagación de la celda a través de la red es mucho mayor que el tiempo de transmisión, por lo que la parte variable del retardo supone, aun en el peor caso, una pequeña fracción del total. En consecuencia, nuestra red se comporta de forma muy parecida a una red STM.

Aclararemos lo anterior con un ejemplo. Supongamos que dos usuarios situados a una distancia de 1.200 Km están unidos mediante un enlace de fibra óptica con una velocidad de transmisión de 600 Mbps. Supondremos que la velocidad de propagación de las señales a lo largo del enlace es de 5 μseg/Km. y que el tamaño de las celdas es el recomendado por el CCITT para el caso de redes ATM: 56 octetos. El tiempo de transmisión de una celda es, pues, de unos 0.7 μseg, mientras que el tiempo de propagación es 6 msec, es decir, aproximadamente unas 9.000 veces mayor. Si el enlace fuera a 60 Kbps la situación cambiaría radicalmente: El tiempo de propagación sería el mismo (6 mseg.), pero el tiempo de transmisión aumentaría 10.000 veces, es decir, sería ahora de 7 mseg.

En [Tra89] se hace un estudio de las variaciones del tiempo que tarda una celda en atravesar la red para el caso de una red ATM. Para ello se consideran tres escenarios distintos: acceso directo a la red de 600 Mbps, acceso a través de un TR2 a 150 Mbps y a través de una TR2 a 2 Mbps. En el caso de acceder a 150 Mbps, o a 2 Mbps, el interfaz es síncrono (figura 5.2).

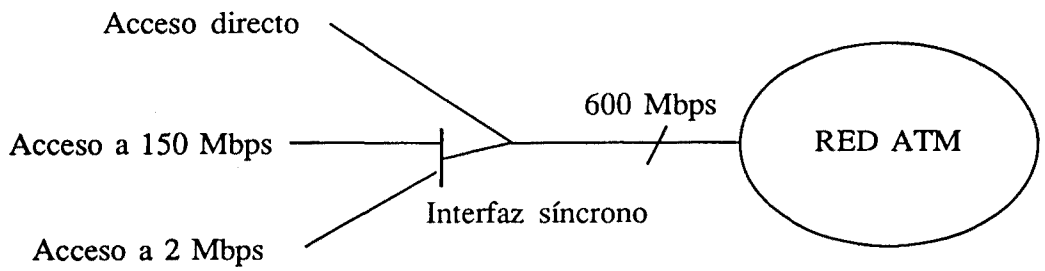


Figura 5.2

Se suponen 30 etapas a lo largo del recorrido de la red y una configuración simétrica en el otro extremo. La carga de los conmutadores de la red ATM se supone de 0.9, la distancia entre los usuarios es de 1.200 Km y en el enlace es de fibra óptica.

La variación en el tiempo de atravesar la red puede producirse en las etapas de multiplexación y demultiplexación asíncrono de las TR2 de los extremos y en los conmutadores internos de la red. Si t es el tiempo en las colas de espera de los multiplexores, demultiplexores y conmutadores, se calcula el valor t_{max} que cumple:

$$p(t > t_{max}) = 1.0 \text{ E-9}$$

Para el caso de acceso directo a 600 Mbps, t_{max} es menor a 250 μseg. En el caso de acceso a 150 Mbps, con una carga en el TR2 menor a 0.9, se cumple que t_{max} < 400 μseg. Si añadimos los tiempos de propagación, de alineamiento en la entrada, etc, obtenemos, para el caso del acceso a 150 Mbps, que el cuantil a 1.0 E-9 del tiempo

total es de 6.5 mseg. que es un valor perfectamente tolerable para los usuarios humanos. En el caso del acceso a 2 Mbps, para una carga del TR2 menor de 0.9, se cumple t_{max} < 5 mseg. El cuantil a 1.0 E-9 del tiempo total sigue siendo aceptable para el usuario humano: 13 mseg.

Como consecuencia se ve que las variaciones del tiempo que una celda tarda en atravesar la red son pequeñas frente a los tiempos físicos de propagación y son tolerables para el usuario humano.

Las pérdidas de celdas en una red ATM son debidas a los errores de transmisión de la cabecera de la celda, a las pérdidas causadas por la función de policia y a los desbordamientos en las colas de los conmutadores.

Los errores en la transmisión son extremadamente bajos cuando se usan técnicas de transmisión ópticas. Las pérdidas producidas por la función de policia y los desbordamientos en las colas de espera de los conmutadores deben hacerse pequeñas mediante un dimensionado adecuado de la red.

En los servicios en que no se requiere una transmisión en tiempo real (por ejemplo, transmisión de datos) las celdas perdidas pueden ser recuperadas mediante un protocolo de extremo a extremo. Sin embargo, para servicios en tiempo real, como es el caso de comunicaciones de voz y video interactivo, estos mecanismos de retransmisión no son utilizables.

En el caso de la comunicación vocal, una celda transporta varias muestras de la señal digitalizada. Por ejemplo, si usamos PCM, una celda transporta 47 muestras codificadas con 8 bits, es decir, unos 6 mseg. de señal vocal. Si una celda se pierde se puede producir en el receptor un 'click' audible. Sin embargo, la probabilidad de pérdida tolerada para la calidad requerida por el servicio telefónico, es relativamente alta [Srietal91] (ver figura 5.1).

Los servicios de video son mucho más sensibles a pérdidas de celdas. Debido a que la velocidad de transmisión es más alta, las pérdidas ocurren con una mayor frecuencia. Por ejemplo: si la probabilidad de pérdida de una celda es de 1.0 E-8, el tiempo medio entre pérdidas para un servicio como HDTV es de varios minutos, mientras que para voz a 64 Kbps es de varios dias. Además, en el caso del video, la trama tiene una duración de unos 40 msec, por lo que es transportada en varias celdas ATM. Una pérdida en una celda puede corromper la información del resto de la trama o incluso de tramas consecutivas ([Veretal88], [Gha89]), lo cual tiene efectos visibles en la imagen. Como consecuencia, la probabilidad de pérdida de celda tolerada por estos servicios tiene un valor mucho menor que en el caso de la voz (figura 5.1).

De los anterior se deduce que:

- Los requerimientos de los diferentes usuarios en lo que concierne al retardo que sufre una celda al ser transmitida por la red son fáciles de satisfacer.

- Sin embargo, las probabilidades máximas de pérdidas de celdas que requieren algunos usuarios tienen valores muy pequeños, y requieren un tratamiento más cuidadoso.

En el diseño de redes de alta velocidad se debe partir de la premisa de que el sistema debe ser lo mas simple posible. Por ello es en principio deseable usar un único servicio portador para todos los usuarios. Sin embargo, la probabilidad de desbordamiento en las colas de espera de los conmutadores está directamente relacionada el tráfico presente en la red. Debemos llegar, pues, a un compromiso entre

la utilización de los recursos de la red y la calidad de servicio exigida por los usuarios. Dada la diferencia entre lo requerido por servicios como la telefonía y la transmisión de datos, que por otra parte representarán una parte muy importante del tráfico total, y lo requerido por servicios como la transmisión de video o las señalizaciones internas, la utilización de un único servicio portador, que debería satisfacer las demandas de los servicios más restrictivos, puede suponer un uso muy ineficiente de los recursos de la red.

Una solución a este problema es definir más servicios portadores, que asegurarían diferentes valores mínimos en la probabilidad de pérdida o de retardo que una celda sufre al atravesar la red. El número de servicios portadores debería mantenerse reducido (por ejemplo, solo dos o tres) y los mecanismos que permitan su introducción en la red deberían ser tan sencillos como sea posible.

Se han realizado diferentes propuestas en lo que respecta a la introducción de varios servicios portadores. Por ejemplo en [Tur86], [Boy88] se propone el uso de dos servicios portadores con diferentes valores de probabilidad de pérdida. En [WooKos90] se propone el uso de tres servicios portadores, introduciendo diferencias en los valores de probabilidad de pérdida y de retardo. El CCITT recomienda el uso de un bit de prioridad espacial en la cabecera de la celda ATM.

En este trabajo estudiaremos la propuesta que parece más atractiva para el caso de redes ATM: el uso de dos servicios portadores que aseguran diferentes probabilidades de pérdidas. De esta forma unas celdas serían marcadas como prioritarias, y se les asegurarían unos niveles de pérdidas muy bajos mientras que las otras celdas serían marcadas como ordinarias y tendrían valores de pérdidas mayores.

En el caso de usar diferentes servicios portadores, los codificadores deberían introducir diferentes niveles de prioridad en cada celda. Una posibilidad sería marcar con el mismo nivel de prioridad todas las celdas pertenecientes a una misma comunicación. Sin embargo parece deseable el poder introducir diferencias entre las celdas de una misma comunicación [Kroetal91].

En [Srietal91], [Yinetal90] se propone, para el caso de tráfico de voz, enviar en celdas prioritarias los bits más significativos de las muestras codificadas y los menos significativos en celdas ordinarias. Con ello se consigue que las pérdidas de celdas ordinarias tengan un impacto reducido sobre la calidad de la comunicación. Aunque este tipo de codificación aumentaría el retardo de ensamblado de celda, parece que los retardos introducidos son tolerables. En [Tur86] se propone la utilización de un mecanismo análogo para el tráfico de video.

Otro método de codificación considerado en [Srietal91], [Yinetal90] consiste en enviar las muestras pares en celdas de una clase y las muestras impares en celdas de otra clase, de forma que las muestras perdidas puedan reconstruirse mediante interpolación. Sin embargo este método es eficiente para voz codificada con PCM, pero no para el caso de estar codificada en DPCM [JayChr81], por lo que el primer método parece más eficaz.

En [Yinetal90] se propone un método de codificación de la voz basado en el uso de 2 umbrales de detección de la energía de la señal. Si la energía de un segmento de voz es menor que el umbral más bajo, se asume que es un silencio y no se transmite. Si está entre los dos umbrales se transmite en celdas de baja prioridad (semisilencio). Si supera el umbral más alto se transmite en celdas de alta prioridad (actividad).

En [Gha89] se propone la utilización de un codificador de video con dos niveles. El nivel 1 transmite la información que es vital para la reconstrucción de la escena con

un mínimo nivel de calidad, tal como las señales de sincronismo, parte de la información de video, etc, y es transmitido en celdas prioritarias. El segundo nivel transmite el resto de la información y es transmitido en celdas ordinarias. De esta forma se consigue que las pérdidas de celdas ordinarias tengan un efecto reducido sobre la imagen.

Se supone que el tráfico prioritario sería solo una pequeña fracción del tráfico total (por ejemplo un 20%, [Rigetal90], [Kroetal91]).

5.2 Propuestas de mecanismos de prioridad espacial para redes ATM

Los sistemas de colas con prioridades, en donde se separan en clases a los clientes de una cola, dando un trato preferente en algún sentido a los clientes de una clase sobre los demás clientes, han sido ampliamente estudiados.

El caso más común consiste en sistemas en donde se usa un mecanismo que da prioridad de acceso al servidor a los clientes de una determinada clase. Este tipo de mecanismos se denominan mecanismos de prioridad temporal, porque en ellos se intenta, en general, minimizar el tiempo que los clientes de una clase deben esperar antes de recibir servicio. Un estudio de diferentes mecanismos propuestos, leyes de conservación, etc. puede encontrarse en la mayoria de libros sobre sistemas de colas (por ejemplo, [Kle75], [Coo72]).

Cuando tratamos con sistemas en donde hay pérdidas, podemos pensar en otro tipo de mecanismos de prioridades, en donde los clientes de una clase tienen prioridad de acceso a las posiciones de espera del sistema. Este tipo de mecanismos se denominan mecanismos de prioridad espacial ('Space priorities mechanisms'), pues el recurso que se reparte entre las diferentes clases es el espacio de la cola. También se denomínan mecanismos de prioridad de pérdida ('Loss priorities mechanisms'), pues lo que se pretende es controlar la probabilidad de pérdida de las diferentes clases de clientes.

Los mecanismos de prioridad espacial han sido menos estudiados que los de prioridad temporal. Sin embargo, tal como se ha dicho, parecen los más adecuados en el caso de las redes ATM.

Recientemente se han propuesto dos mecanismos de prioridad espacial, que son especialmente interesantes en su aplicación a redes ATM y en cuyo estudio se centra la segunda parte de este trabajo: Son los mecanismos de 'Push-Out' y de 'Partial buffer sharing' [Kroetal91]. En las definiciones de estos dos mecanismos supondremos que tenemos dos tipos de celdas: prioritarias, con un bajo nivel de pérdidas garantizado, y ordinarias, con un nivel de pérdidas que puede ser mayor.

El mecanismo de 'Push-Out' (PO).

El mecanismo conocido como de 'Push-Out', que nosotros abreviaremos como mecanismo PO, garantiza el acceso al sistema a cualquier celda, siempre y cuando la cola no esté llena. Cuando la cola está llena y llega una nueva celda, pueden ocurrir dos situaciones. Si la celda que llega es de baja prioridad, es decir es una celda ordinaria, dicha celda se pierde. En cambio, si la celda que llega es de alta prioridad, es decir una celda prioritaria, el sistema se comporta de la siguiente forma:

- Si hay alguna celda ordinaria esperando a ser servida, dicha celda ordinaria es expulsada de la cola, y la celda prioritaria que llega al sistema puede entrar en él.

- Si no hay ninguna celda ordinaria esperando a ser servida, es decir, la cola está completamente llena de celdas prioritarias (no se tiene en cuenta la que está siendo servida) la celda prioritaria que llega se pierde.

Hay varias políticas distintas en cuanto se referiere a cómo se escoge la celda ordinaria que va a ser reemplazada: Así esta podría ser escogida al azar (política de reemplazamiento RANDOM), se podría escoger la que lleva más tiempo en la cola (política de reemplazamiento FIFO), o la que lleva menos tiempo en la cola (política de reemplazamiento LIFO). Nos restringiremos al caso de política LIFO. De los estudios realizados para tráfico de Poisson parece que hay solo hay ligeras diferencias en los resultados obtenidos para las diferentes políticas [Kro90] y, en el caso de redes ATM, la política LIFO minimiza la complejidad del manejo del la cola.

Un punto importante hace referencia a la disciplina de servicio en la cola. Las celdas que llegan al multiplexor deben ser servidas en el orden de llegada. Por la tanto la disciplina de servicio debe ser FIFO y cuando una celda prioritaria reemplaza a una ordinaria no puede substituirla en la misma posición del buffer de donde sacamos la celda ordinaria, sino que debe situarse en la última posición de la cola. Ello implica que las celdas situadas entre la celda substituida y el final de la cola deben ser desplazadas una posición hacia delante.

Construir una cola relativamente grande (por ejemplo de 64 posiciones) en la que deben realizarse deslizamientos del tipo anteriormente descrito y que trabaje a altas velocidades es considerablemente complejo. Una posibilidad, propuesta y estudiada en este trabajo, es dotar a la cola de un umbral. El mecanismo de reemplazamiento solo se realizaría en las posiciones comprendidas entre el umbral y el final de la cola, de forma que una celda prioritaria que encontrase un sistema completo buscaría si hay celdas prioritarias en esta última zona de la cola. Si las hubiera, esa celda prioritaria podría entrar en el sistema. En caso contrario se perdería. Obviamente esta modificación supone un empeoramiento en las probabilidades de pérdidas de las celdas prioritarias (y una mejora en la obtenida para las ordinarias), de forma que el umbral debería fijarse para alcanzar las probabilidades de pérdidas requeridas (figura 5.3).

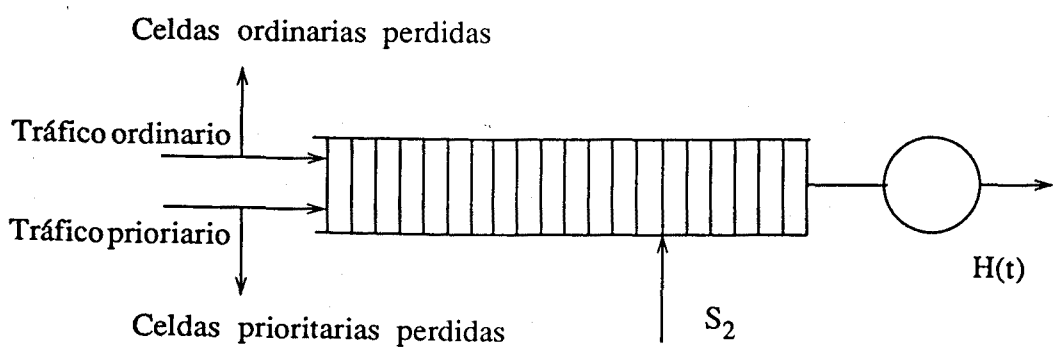


Figura 5.3. Mecanismo PO con umbral

Un mecanismo similar al PO fue propuesto en [DosHef86] para el control de congestión de una cola M/M/1. En este caso, sin embargo, no se distinguían entre

clases de celdas, pues no era en realidad un mecanismo de prioridad espacial.

Así puede considerarse que este mecanismo de prioridad espacial fue propuesto por primera vez en [SumOza89], en donde se dan leyes de conservación para las probabilidades de pérdida, número medio de clientes y tiempo medio de espera en la cola. En dicho artículo no se da un método que permita calcular de forma exacta las probabilidades de pérdida para cada clase que es sin duda el parámetro más relevante en nuestro caso.

En [HebGra89], se da un método para calcular las probabilidades de pérdida cuando este mecanismo se usa en una cola M/M/1 y M/D/1. Sin embargo el algoritmo usado presenta problemas numéricos para buffers de un tamaño mayor que 10.

En [Kro90] se propone el mecanismo PBS, que será explicado a continuación, y se da un algoritmo mejor desde el punto de vista numérico para el análisis del mecanismo PO.

En [Kroetal91] se da un modelo aproximado para el análisis cuando se tienen fuentes con dos estados de actividad.

En [Niletal90] se analiza un sistema IBP/GEO/1/K con prioridad espacial (mecanismo PO) y temporal (mecanismo HOL, ver, por ejemplo [Kle72]).

El mecanismo de 'Partial buffer sharing' (PBS)

El mecanismo the 'Partial buffer sharing', o simplemente mecanismo PBS, fue propuesto en [Kro90] como una alternativa al mecanismo PO. Su objetivo es obtener unos resultados similares a los que se pueden obtener con el mecanismo PO, pero de forma que la realización del sistema sea más sencilla.

En este mecanismo fijamos un umbral en la cola. Cuando la longitud de cola está por debajo de dicho umbral, todas las celdas que llegan son admitidas. Sin embargo, cuando la longitud de cola supera el umbral, las celdas ordinarias son rechazadas. Si el sistema está completamente lleno, también se pierden las celdas prioritarias.

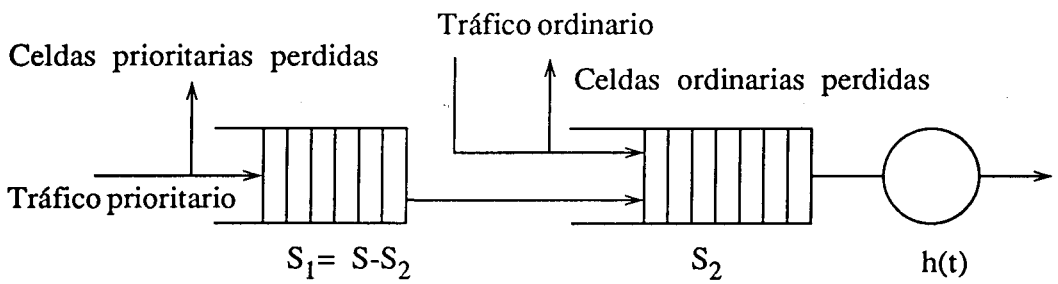


Figura 5.4. El mecanismo PBS

Varios mecanismos de compartición de una cola basados en umbrales han sido estudiados. Entre ellos cabe citar los siguientes:

En [Irl78] se estudian diferentes mecanismos de compartición de una cola finita entre N flujos de paquetes en un nodo de una red de conmutación de paquetes. Para cada uno de estos tráficos, que siguen una estadística de Poisson, tenemos un servidor exponencial.

En [Yinetal90] se estudia un modelo basado en la aproximación de fluido de un mecanismo similar al PBS. Sin embargo el análisis es solo válido cuando el tráfico máximo de celda prioritarias es menor que el tráfico servido por la red, por lo que la longitud de cola no sobrepasa el valor S_2, y el modelo se reduce a una cola de longitud S_2 con dos clases de clientes.

5.4 Los modelos propuestos para el estudio de mecanismos de prioridad espacial

Modelo con tráfico de entrada de Poisson [Kro90] ·

Los modelos exactos para mecanismos de prioridad espacial estudiados en [Kro90] se restringen a sistemas en donde el tráfico de entrada es de Poisson.

El proceso de entrada se considera como la superposición de dos procesos de Poisson independientes de parámetros λ_p y λ_{np} (para las celdas prioritarias y ordinarias, respectivamente) de forma que el proceso total es también de Poisson de parámetro $\lambda = \lambda_p + \lambda_{np}$. B_p es la probabilidad de pérdida de las celdas prioritarias y B_{np} la de las celdas ordinarias.

El análisis de la cola con mecanismo PO sigue los siguientes pasos.

En primer lugar en [SumOza89] se establece la siguiente ley de conservación:

B_T = Probabilidad de pérdida total.

$$B_T \lambda = B_p \lambda_p + B_{np} \lambda_{np} \qquad (5.1)$$

La probabilidad de pérdida total es la misma que la de un sistema M/G/1/N, que puede ser obtenida facilmente. Por lo tanto solo es necesario calcular la probabilidad de pérdida para una de las dos clases de celdas.

La probabilidad de pérdida de las celdas ordinarias puede ser calculada a partir de la expresión:

$$B_{np} = \sum_{k=0}^{N} p_k \ (1 - P(\text{servida}/k))$$

donde p_k son las probabilidades de estado vistas por un cliente que llega al sistema (que coinciden con las que ve un observador arbitrario, debido a la propiedad PASTA

[Woo82]).

Para evaluar P(servida/k), que es la probabilidad de que una celda ordinaria sea servida cuando llega a la cola en posición k, se emplea el siguiente algoritmo:

Definimos A(n) como la probabilidad de que durante un servicio lleguen n clientes prioritarios al sistema, teniendo entonces:

$$A(n) = \int_0^\infty e^{-\lambda_p u} \frac{(\lambda_p u)^n}{n!} dH(u)$$

Así mismo, definimos $A_k(n)$ como la probabilidad de que durante el tiempo residual de servicio de un cliente prioritario que encuentra la cola con k clientes, lleguen n clientes al sistema:

$$A_k(n) = \int_0^\infty e^{-\lambda_p u} \frac{(\lambda_p u)^n}{n!} r_k(t)\, du$$

donde $r_k(t)$ es la función de densidad del tiempo residual de servicio de un cliente prioritario que encuentra la cola con k clientes. Su expresión es

$$r_k(t) = \int_0^\infty \lambda\, h(u+t)\, e^{-\lambda u} \left(p_0 \frac{(\lambda u)^{k-1}}{(k-1)!} + \sum_{j=0}^{k-1} p_{k-j} \frac{(\lambda u)^j}{j!} \right) du \qquad (5.2)$$

Cuando una celda ordinaria entra en el sistema en la posición k (para k= S_2+1,...,S), se mueve hacia la posición k-1 cuando el servicio acaba, siempre y cuando durante el tiempo residual hasta el final de dicho servicio hayan llegado menos de S-k celdas prioritarias. En general, si dicha celda llega a la posición k-f > S_2, será desplazada a la posición k-f-1 si durante el tiempo residual inicial y los f siguientes servicios llegan menos de S-k+f celdas prioritarias.

De lo anterior se puede establecer el siguiente algoritmo [Kro90]:

C_f (k,n) = p{La celda ordinaria que está siendo observada se mueve desde la posición k-f a la k-f-1 y n celdas prioritarias llegan desde que ha entrado en la cola}

Paso 1:

$$C_0(k,n) = \begin{cases} A_k(n) & \text{si } 0 \le n \le S\text{-}k \\ 0 & \text{en otro caso} \end{cases} \qquad (5.3)$$

Paso f: $1 \le f \le k\text{-}1$

$$C_f(k,n) = \begin{cases} C_{f-1}(k,n) \otimes A(n) & \text{si } 0 \le n \le S\text{-}k+f \\ 0 & \text{en otro caso} \end{cases}$$

el operador $\otimes$ es la convolución en n.

Finalmente:

$$P(\text{servida/k}) = \sum_{n=0}^{S-1} C_{k-1}(k,n)$$

El mecanismo PBS puede ser analizado, siguiendo los siguientes pasos:

Considerando la longitud de cola en los instantes de servicio, obtenemos la siguiente ecuación para la longitud de cola que en estado estacionario observa un cliente que abandona el sistema:

$$\bar{x}Q = x^{-} \qquad (5.4)$$
$$\bar{x}\,\bar{e} = 1$$

La componente i del vector $\bar{x}$ es la probabilidad en estado estacionario de que un cliente que abandona el sistema deje k clientes esperando. La matriz Q tiene la

siguiente forma:

$$Q = \begin{vmatrix}
A_1(0) & A_1(1) & \dots & A_1(S2\text{-}1) & A_{12}(0) & A_{12}(1) & \dots & A_{12}(S\text{-}S2\text{-}1) & \sum_{j=S-S_2}^{\infty} A_{12}(j) \\
A_1(0) & A_1(1) & \dots & A_1(S2\text{-}1) & A_{12}(S2,0) & A_{12}(S2,1) & \dots & A_{12}(S2,S\text{-}S2\text{-}1) & \sum_{j=S-S_2}^{\infty} A_{12}(S2,j) \\
0 & A_1(0) & \dots & A_1(S2\text{-}2) & A_{12}(S2\text{-}1,0) & A_{12}(S2\text{-}1,1) & \dots & A_{12}(S2\text{-}1,S\text{-}S2\text{-}1) & \sum_{j=S-S_2}^{\infty} A_{12}(S2\text{-}1,j) \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \dots & A_1(2) & A_{12}(2,0) & A_{12}(2,1) & \dots & A_{12}(2,S\text{-}S2\text{-}1) & \sum_{j=S-S_2}^{\infty} A_{12}(2,j) \\
0 & 0 & \dots & A_1(1) & A_{12}(1,0) & A_{12}(1,1) & \dots & A_{12}(1,S\text{-}S2\text{-}1) & \sum_{j=S-S_2}^{\infty} A_{12}(1,j) \\
0 & 0 & \dots & 0 & A_2(0) & A_2(1) & \dots & A_2(S\text{-}S2\text{-}1) & \sum_{j=S-S_2}^{\infty} A_2(j) \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \dots & 0 & 0 & 0 & \dots & A_2(1) & \sum_{j=2}^{\infty} A_2(j) \\
0 & 0 & \dots & 0 & 0 & 0 & \dots & A_2(0) & \sum_{j=1}^{\infty} A_2(j)
\end{vmatrix}$$

en donde las expresiones de los componentes de la matriz Q son las siguientes:

$$A_1(n) = \int_0^{\infty} e^{-\lambda u} \frac{(\lambda u)^n}{n!} \, dH(u)$$

$$A_2(n) = \int_0^{\infty} e^{-\lambda p u} \frac{(\lambda p u)^n}{n!} \, dH(u)$$

y

$$A_{12}(n,m) = \sum_{j=n+m}^{\infty} A_1(j) \binom{j-n}{m} \frac{\lambda_p{}^m \lambda_{np}{}^n}{\lambda^{j-m-n}}$$

Resolviendo la ecuación (5.4) obtenemos las probabilidades de ocupación que observan los clientes que abandonan el sistema, que son las mismas que las que observan los clientes que entran en el sistema. De ahí es fácil obtener las probabilidades de ocupación que observan los clientes que llegan al sistema y finalmente obtener las probabilidades de pérdida para cada clase de cliente.

Modelo aproximado con fuentes de dos estados [Kroetal91]

Supondremos que tenemos Ni fuentes de clase i (i = p, np). Cada clientes de clase i puede estar en dos estados: activo e inactivo. Cuando un cliente de clase i está en estado activo, emite celdas cada T_{bi} segundos. El número de celdas emitidas en estado de actividad sigue una distribución geométrica de media n_{bi}. En el estado inactivo no se emiten celdas y tiene una duración distribuida exponencialmente de media T_{si} (figura 3.5).

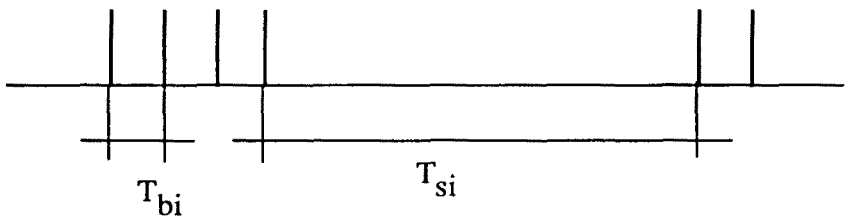


Figura 5.5

La probabilidad de que un cliente de clase i esté activo viene dada por:

$$p_{ai} = \frac{n_{bi} T_{bi}}{n_{bi} T_{bi} + T_{si}}$$

La probabilidad de que x_i clientes de clase i estén simultáneamente en estado activo valdrá:

$$p(x_i) = \binom{n_{bi}}{x_i} p_{ai}^{x_i} (1-p_{ai})^{N_i - x_i}$$

Consideraremos que nuestro sistema es un sistema de pérdidas, es decir, supondremos que el sistema no es capaz de almacenar las celdas cuando el tráfico de entrada es mayor que el servido, y que nunca se producen pérdidas si el tráfico de entrada es menor que el tráfico servido. Por lo tanto, si tenemos x_p y x_{np} fuentes activas, el tráfico perdido vendrá dado por:

$$\max(0, x_p/T_{bp} + x_{np}/T_{bnp} - 1/h)$$

y las pérdidas totales del sistema serán:

$$B = \frac{1}{N_p p_{ap}/T_{bp} + N_{np} p_{anp}/T_{bnp}} \sum p(x_p)\, p(x_{np}) \max(0, x_p/T_{bp} + x_{np}/T_{bnp} - 1/h)$$

En el caso de las celdas prioritarias, consideraremos que ven el sistema por entero a su disposición, por lo que su probabilidad de pérdida vendrá dada por:

$$B_p = \frac{1}{N_p \, p_{ap}/T_{bp}} \sum p(x_p) \max(0, x_p/T_{bp} - 1/h)$$

Para calcular la probabilidad de pérdida de las celdas ordinarias aplicaremos la ley de conservación (5.1).

Esta aproximación dará valores cercanos a los exactos cuando las duraciones de los estados activo e inactivo sean grandes, y cuando las longitudes de las colas sean suficientemente grandes como para no presentar pérdidas a nivel de celda.

Capítulo 6

Modelos analíticos para la evaluación de políticas de prioridad espacial

6.1 Introducción

Como se ha señalado en el capítulo 2, una evaluación de los mecanismos de prioridad espacial para redes ATM, requiere el uso de procesos de entrada que sean capaces de capturar características del tráfico (correlaciones, existencia de ráfagas), que tienen un importante impacto en el comportamiento del sistema. El proceso de Poisson, que ha sido ampliamente usado en el estudio de las redes de comunicación, no es capaz de modelizar dichas situaciones.

En este capítulo se presentan modelos de dos mecanismos de prioridad espacial en donde el tráfico de entrada es un proceso complejo que puede caracterizar una situación realista de una red ATM. Estos modelos se resuelven usando técnicas analíticas matriciales y la aproximación de fluido.

Esto nos permitirá hacer una comparación entre ambos mecanismos y estudiar sus principales características.

6.2 Modelo de un multiplexor con política PBS y MAP como entrada ([GarCas90.a], [GarCas90.b], [GarCas91.a]).

A continuación vamos a dar un modelo de un multiplexor en donde se incorpora la política PBS en la gestión de la cola, suponiendo que el tráfico de entrada es un MAP. Dicho modelo se puede extender facilmente a el caso en donde tenemos como entrada un BMAP y a los casos discretos.

El multiplexor se modela como una cola con capacidad finita para S clientes, incluyendo al que está siendo servido. La parte de la cola que es compartida puede acoger S_2 clientes. El tiempo de servicio sigue una distribución general, con función de distribución H(.) y media h. El proceso de entrada será un MAP, que genera dos tipos de celdas, prioritarias y ordinarias.

El proceso de entrada

El MAP ya ha sido descrito en detalle en el capítulo 2. A continuación veremos como generamos, a partir de dicho proceso, dos flujos de celdas de diferente prioridad.

En nuestro modelo de fuente asumiremos lo siguiente: Cuando el proceso está en estado i y realiza una transición al estado j, de forma que emite una celda, dicha celda es

prioritaria con una probabilidad ρ_p^{ij}, y es ordinaria con probabilidad ρ_{np}^{ij} ($\rho_p^{ij} + \rho_{np}^{ij} = 1$).

C_T, D_T, C_p, D_p, C_{np} y D_{np} son las matrices asociadas al MAP que corresponden al tráfico total, al tráfico de celdas prioritarias y al de celdas ordinarias, respectivamente (recordemos que el MAP es un caso particular de BMAP en dondes los lotes tienen siempre longitud 1, y que llamábamos $C = D_0$ y $D = D_1$). Definimos las matrices P_T (n,t), P_p (n,t) y P_{np} (n,t) cuyas componentes nos dan las probabilidades condicionadas de tener n llegadas de clientes de cada clase en el intervalo (0,t) y la fase final de la fuente dada la fase inicial.

Longitud de cola en los instantes de salida

Consideraremos el estado del sistema inmediatamente después de τ_n, los instantes de salida de un cliente. I_n y J_n serán la longitud de cola y la fase del proceso de entrada en el instante τ_n+. $\{(I_n, J_n, \tau_{n+1}-\tau_n);\ 0 \le n\}$ forma una secuencia semimarkoviana con espacio de estados $\{0,..,S\}x\{1,..,m\}$.

Llamaremos Q a la matriz de probabilidad de transición en estado estacionario, siendo $\bar{x}$ el vector de probabilidad que cumple:

$$\bar{x}\,Q = \bar{x} \qquad (6.1)$$
$$\bar{x}\,\bar{e} = 1$$

El vector $\bar{x}$ puede ser partido como $\bar{x} = (\bar{x}_0, ..,\bar{x}_S)$ en donde $\bar{x}_k$ es un vector con componentes $\bar{x}_k = (x_k^1, ..., x_k^m)$, siendo x_k^i la probabilidad de que un cliente que abandona el sistema después de ser servido, deje k clientes en la cola y la fuente en estado i.

Los componentes de la matriz Q tienen diferentes expresiones dependiendo de los niveles entre los que se realiza las transiciones:

1) Transiciones de un nivel $0 < k \le S_2$ a un nivel $j < k-1$

Al ser el número de clientes servidos siempre igual a 1, no se pueden dar transiciones desde un nivel k a un nivel $j < k-1$. Por lo tanto todas las componentes de Q correspondientes valdrán 0.

2) Transiciones de un nivel $0 < k \le S_2$ a un nivel $j \le S_2$

Todas las celdas que han llegado, sea cual sea su tipo, pueden entrar en el sistema. Por lo tanto estas transiciones están gobernadas por los parámetros del tráfico total. Consideraremos las matrices mxm $A_1(n_1)$, definidas como:

$$A_1(n_1,x) = \int_0^x P_T(n_1,t)\, dH(t); \qquad A_1(n_1) = A_1(n_1,\infty)$$

La componente i-j de esta matriz es la probabilidad condicionada de que n_1 celdas de cualquier tipo lleguen entre dos instantes de final de servicio y la fase final de la fuente es j, dado que la fase inicial era i.

3) Transiciones de un nivel $k > S_2$ a un nivel $j > S_2$

En este tipo de transiciones estamos siempre en la parte no compartida de la cola. Las celdas ordinarias que llegan son simplemente perdidas, y solo se permite el paso a las celdas prioritarias. Por lo tanto, los parámetros de tráfico relevantes en este tipo de transiciones son solo los del tráfico de celdas prioritarias. Consideraremos ahora las matrices $\mathbf{A}_2(n_2)$ definidas como

$$\mathbf{A}_2(n_2,x) = \int_0^x \mathbf{P}_p(n_2,t)\, dH(t); \qquad \mathbf{A}_2(n_2) = \mathbf{A}_2(n_2,\infty)$$

La componente i-j de esta matriz es la probabilidad condicionada de que n_2 celdas prioritarias lleguen entre dos finales de servicio y la fase final de la fuente es j, dado que la fase inicial de la fuente era i.

4) Transiciones de un nivel $k \leq S_2$ a un nivel $j > S_2$

En estas transiciones partimos de un nivel perteneciente a la parte común de la cola y llegamos a un nivel que pertenece a la parte de la cola no compartida. Por lo tanto, y mientras no se llena la parte común, todas las celdas que llegan son admitidas independientemente de su clase. Una vez llena la parte común del buffer, solo las celdas prioritarias son admitidas. Por lo tanto aquí son necesarios tanto los parámetros del tráfico total, como los del tráfico de celdas prioritarias. Para la evaluación de las correspondientes componentes de la matriz **Q**, definimos las matrices mxm $\mathbf{A}_{12}(n_1,n_2)$ como:

$$\mathbf{A}_{12}(n_1,n_2,x) = \int_0^x \int_0^\xi \mathbf{P}_T(n_1\text{-}1,t)\, d\mathbf{P}_T/dt\ (1,0)\ \mathbf{P}_p(n_2,\xi\text{-}t)\, dt\, dH(\xi);$$

$$\mathbf{A}_{12}(n_1,n_2) = \mathbf{A}_{12}(n_1,n_2,\infty)$$

La componente i-j de esta matriz es la probabilidad de que tengamos n_1 ($n_1 > 0$) llegadas de celdas de cualquier clase, seguidas de n_2 llegadas de celdas prioritarias entre dos instantes de final de servicio, y la fase final sea j, dado que la fase inicial era i (ver

inicial era i (ver figura 6.1)

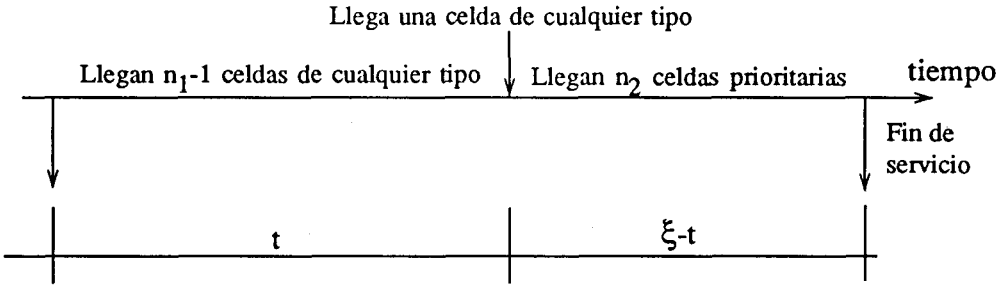


Figura 6.1

La derivada en el origen de $P_T(n,t)$ para n = 1 vale $dP_T/dt(1,0) = D_T$.

2) Transiciones de un nivel k = 0 a un nivel $j \leq S$

En este caso tenemos que el cliente que ha sido servido deja el sistema vacio. Definimos las matrices $B_1(n_1)$ y $B_{12}(n_2)$ como:

$$B_1(n_1,x) = U(x) * A_1(n_1,x) \qquad B_1(n_1) = B_1(n_1,\infty)$$

$$B_{12}(n_2,x) = U(x) * A_{12}(S_2,n_2,x) \qquad B_{12}(n_2) = B_{12}(n_2,\infty)$$

en donde * denota la convolución matricial y U(x) es definido como:

$$U(x) = \int_0^x P_T(0,t)\, D_T\, dt \qquad U = U(\infty)$$

La componente i-j de la matriz U(x) es la probabilidad de que la primera llegada de una celda de cualquier clase ocurra antes de un tiempo x con la fuente en fase j, dado que la fuente estaba originalmente en fase i.

Las expresiones para $B_1(n_1)$ y $B_{12}(n_2)$ son:

$$B_1(n_1) = -C_T^{-1} D_T\, A_1(n_1) \; ; \qquad B_{12}(n_2) = -C_T^{-1} D_T\, A_{12}(S_2,n_2)$$

Con todas las anteriores definiciones tenemos ahora que la matriz Q viene dada por:

$$Q = \begin{vmatrix}
B_1(0) & B_1(1) & \dots & B_1(S_2\text{-}1) & B_{12}(0) & B_{12}(1) & \dots & B_{12}(S\text{-}S_2\text{-}1) & \sum_{j=S-S_2}^{\infty} B_{12}(j) \\
A_1(0) & A_1(1) & \dots & A_1(S_2\text{-}1) & A_{12}(S_2,0) & A_{12}(S_2,1) & \dots & A_{12}(S_2,S\text{-}S_2\text{-}1) & \sum_{j=S-S_2}^{\infty} A_{12}(S_2,j) \\
0 & A_1(0) & \dots & A_1(S_2\text{-}2) & A_{12}(S_2\text{-}1,0) & A_{12}(S_2\text{-}1,1) & \dots & A_{12}(S_2\text{-}1,S\text{-}S_2\text{-}1) & \sum_{j=S-S_2}^{\infty} A_{12}(S_2\text{-}1,j) \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \dots & A_1(2) & A_{12}(2,0) & A_{12}(2,1) & \dots & A_{12}(2,S\text{-}S_2\text{-}1) & \sum_{j=S-S_2}^{\infty} A_{12}(2,j) \\
0 & 0 & \dots & A_1(1) & A_{12}(1,0) & A_{12}(1,1) & \dots & A_{12}(1,S\text{-}S_2\text{-}1) & \sum_{j=S-S_2}^{\infty} A_{12}(1,j) \\
0 & 0 & \dots & 0 & A_2(0) & A_2(1) & \dots & A_2(S\text{-}S_2\text{-}1) & \sum_{j=S-S_2}^{\infty} A_2(j) \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & \dots & 0 & 0 & 0 & \dots & A_2(1) & \sum_{j=2}^{\infty} A_2(j) \\
0 & 0 & \dots & 0 & 0 & 0 & \dots & A_2(0) & \sum_{j=1}^{\infty} A_2(j)
\end{vmatrix}$$

Para el cálculo de las matrices $A_1(n_1)$, $A_2(n_2)$ y $A_{12}(n_1,n_2)$ podemos usar un algoritmo análogo al usado en la cola BMAP/G/1 sin prioridades, y descrito en el apartado 2.3.2:

Tenemos que:

$$P_i(n,t) = \sum_{j=0}^{\infty} e^{-\theta_i t} (\theta_i t)^j / j! \; K_n^{i(j)} \qquad \text{para} \quad i: T, p, np$$

en donde:

$$\theta_i = \max_k \; (C_i^{kk})$$

y $K_n^{i(j)}$ está definido de forma recursiva por

$$K_0^{i(0)} = I; \qquad K_n^{i(0)} = 0 \quad \text{para} \quad n > 0$$

$$K_0^{i(j+1)} = K_0^{i(j)} \; (I + \theta_i^{-1} C_i)$$

$$K_n^{i(j+1)} = \theta_i^{-1} K_{n-1}^{i(j)} \; D_i + K_n^{i(j)} \; (I + \theta_i^{-1} C_i)$$

Substituyendo ahora las definiciones de las matrices en cuestión se obtiene:

$$A_1(n_1) = \sum_{j=0}^{\infty} \gamma_j^T K_{n1}^{T(j)}; \qquad \gamma_j^T = \int_0^{\infty} e^{-\theta T t} (\theta_T t)^j / j!\ dH(t)$$

$$A_2(n_2) = \sum_{j=0}^{\infty} \gamma_j^p K_{n2}^{p(j)}; \qquad \gamma_j^p = \int_0^{\infty} e^{-\theta p t} (\theta_p t)^j / j!\ dH(t)$$

$$A_{12}(n_1,n_2) = \sum_{j=0}^{\infty} \sum_{i=0}^{\infty} \gamma_{ij} K_{n1-1}^{T(j)} D_T K_{n2}^{p(i)}$$

$$\gamma_{ij} = \int_0^{\infty} dt \int_0^{\xi} e^{-\theta m \xi} t^j (\xi - t)^i \theta_m^{i+j} / (i+j)!\ dt\ dH(\xi); \qquad \theta_m = \max(\theta_T, \theta_p)$$

La resolución de las ecuaciones (6.1) proporciona la distribución de la longitud de cola que ven los clientes que dejan el sistema. La resolución de este sistema de ecuaciones puede realizarse con un método análogo al expuesto para la cola BMAP/G/1 de capacidad finita.

Nuestro interés se centra en calcular las probabilidades de pérdida de las celdas prioritarias y ordinarias, B_p y B_{np} respectivamente. Para ello debemos evaluar la distribución de longitud de cola que observan los clientes que llegan al sistema. Así, si $p_k^{j\,i}$ es la probabilidad de que un cliente de clase j (j : p, np) observe la cola con k clientes y la fuente en fase i y definimos p_k^{j} como la suma de las anteriores probabilidades para todas las fases, tenemos.

$$B_p = p^p{}_{S+1} \qquad B_{np} = \sum_{k=S2}^{S+1} p^{np}{}_k$$

Hasta ahora, nuestro modelo ha seguido pasos análogos a los necesarios para evaluar las probabilidades de pérdida cuando el proceso de entrada era de Poisson. Sin embargo ahora aparece una diferencia importante.

Tal como se ha indicado en el apartado 5.4, cuando el proceso de entrada es de Poisson, las probabilidades de estado que observa un cliente que llega al sistema se pueden relacionar de forma directa con las que observa un cliente que abandona el sistema. Si definimos como π_k la probabilidad de que un cliente que llega observe k clientes en la cola y x_k la probabilidad de que un cliente que es servido deje k clientes en la cola (al ser la entrada de Poisson no tenemos en cuenta la fase de la fuente), tenemos que:

$$x_k = \pi_k$$

Una vez conocidas las probabilidades π_k es fácil relacionarlas con las probabilidades que observa un cliente que llega, y que debido a la propieded PASTA son de hecho las que tenemos en un instante arbitrario.

La situación en nuestro caso es más complicada. Se sigue cumpliendo que las probabilidades de ocupación que observa un cliente que abandona el sistema son las mismas que las que observa uno que entra (esta es una propiedad de la cola G/G/1). Sin embargo si tenemos en cuenta las fases del sistema, esta relación deja ya de cumplirse: Las fases de la fuente en los instantes en que los clientes que entran observan el sistema con un determinado número de clientes no se pueden relacionar de forma sencilla con las fases de la fuente en los instantes en que los clientes que abandonan el sistema lo dejan con dicho número de clientes (ver figura 6.2).

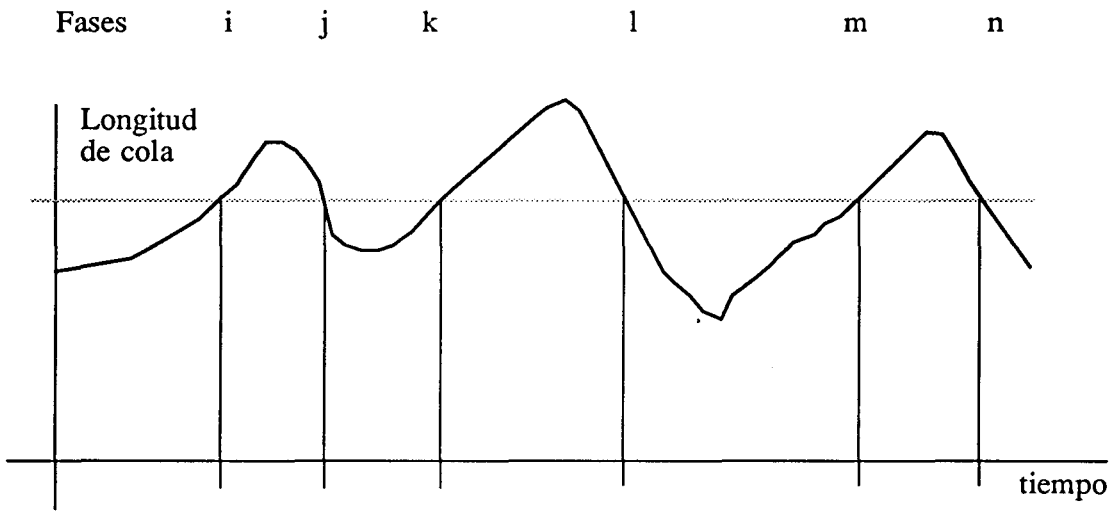


Figura 6.2

Por lo tanto, en nuestro caso, podemos saber de forma directa las probabilidades de ocupación que observa un cliente que entra en el sistema, pero no podemos saber en qué fase observa la fuente. De ahí que no podamos conocer la probabilidad de que el cliente que llega sea de una determinada clase, y por lo tanto no podemos calcular las probabilidades de pérdida.

Hay un caso en que lo anterior deja de ser cierto: Si la probabilidad de emisión de una celda de un determinado tipo no depende del estado de la fuente (es decir, si las proporciones de celdas prioritarios y ordinarias son las mismas para cualquier fase), podemos calcular la probabilidad de que un cliente que encuentra la cola en cierto estado sea de una determinada clase sin necesidad de conocer en qué fase se halla la fuente. Este caso fue analizado en una primera etapa del desarrollo del modelo [GarCas90.a], y se obtiene lo siguiente:

Ahora definimos el vector $\bar{\pi}_k$, cuyas componentes nos darán las probabilidades de que un cliente que entra observe la cola con k clientes y la fuente en una determinada fase. También definiremos los vectores $\bar{\lambda}_i = (\lambda_i^0 , \lambda_i^1 , ...)$ para i = p, np y T, cuyas componentes nos dan la probabilidad de emisión de una celda de clase i cuando estamos en una determinada fase.

Para cada componente de los vectores $\bar{p}_k$ y $\bar{\pi}_k$ tenemos:

$$\pi_k^i = (p_k^i \ \bar{\lambda}_T \bar{\theta}) / (\bar{\lambda}_T \bar{\theta} - \bar{\lambda}_p \bar{p}_{S+1} - \bar{\lambda}_{np} \sum_{j=S2+1}^{S} \bar{p}_j) \qquad k= 0 \ldots S_2$$

$$\pi_k^i = (\lambda_p^i / \lambda_T^i) (p_k^i \ \bar{\lambda}_T \bar{\theta}) / (\bar{\lambda}_T \bar{\theta} - \bar{\lambda}_p \bar{p}_{S+1} - \bar{\lambda}_{np} \sum_{j=S2+1}^{S} \bar{p}_j) \qquad k= S_2+1..S$$

Sabemos que el número medio de clientes que entran en la cola debe ser igual al de clientes que abandonan el sistema, que es el inverso del valor medio del tiempo entre salidas. Cuando la cola no está vacía, este tiempo vale h. Cuando la cola queda vacía, debemos tener en cuenta el tiempo medio entre llegadas al sistema. De ahí que si definimos Um_{ij} como la media de la variable aleatoria cuya función de distribución viene dada por $U_{ij}(x)$, podamos establecer que:

$$\lambda_T \ \bar{\theta} - \bar{\lambda}_p \bar{p}_{S+1} - \bar{\lambda}_{np} \sum_{j=S2+1}^{S} \bar{p}_j = 1 / (h + x_0 U_m \ \bar{e}^t)$$

El vector $\bar{\theta}$ es el vector de probabilidad de estado estacionario para la cadena de Markov que sirve como substrato en la definición del MAP.

Finalmente obtenemos, para el caso en que las proporciones de celdas de cada tipo sean constantes en todos los estados, lo siguiente:

$$p_k = x_k / ((h + \bar{x}_0 U_m \ \bar{e}^t \ ) \ \bar{\lambda}_T \ \bar{\theta}) \qquad k= 0..S_2$$

$$p_k = x_k / ((h + \bar{x}_0 U_m \ \bar{e}^t \ ) \ \bar{\lambda}_p \bar{\theta}) \qquad k= S_2+1..S$$

$$p_{S+1} = 1 - 1 / ((h + \bar{x}_0 U_m \ \bar{e}^t \ ) \ \bar{\lambda}_T \ \bar{\theta}) \ - (\bar{\lambda}_{np} \ \bar{\theta} / \ \bar{\lambda}_T \ \bar{\theta}) \sum_{j=S2+1}^{S} p_j$$

y:

$$Bp = p_{S+1} \qquad Bnp = \sum_{j=S2}^{S+1} p_j$$

En el caso general en el que la proporción de celdas de una determinada clase depende de la fase de la fuente debemos seguir otro camino. Relacionaremos las probabilidades de estado que ven los clientes que llegan con las que vería un observador arbitrario. Una vez hecho esto encontraremos esa última probabilidad a partir de las probabilidades de estado que observan los clientes que abandonan el sistema, obteniendo el resultado deseado.

En primer lugar demostraremos lo siguiente:

$$p_k^{j\,i} = \sum_{t=1}^{m} D_j^{it}\; y_k^i \;/\; \bar{\theta}\, D_j\, \bar{e} \qquad (6.2)$$

en donde y_k^i es la probabilidad de que un observador arbitrario observe la cola con k clientes y la fuente en fase i.

Demostración:

$$p_k^{j\,i} = p\{\, I(\tau) = k,\, J(\tau) = i \,/\, \text{una celda de clase j llega en el instante } \tau \,\} =$$

$$= \frac{p\{\, \text{Una celda de clase j llega en } \tau \,/\, I(\tau) = k,\, J(\tau) = i \,\}\; p\{\, I(\tau) = k,\, J(\tau) = i \,\}}{p\{\, \text{Una celda de clase j llega en } \tau \,\}} =$$

Sin embargo, la probabilidad de que en un instante dado llegue un cliente de clase j solo depende de la fase de la fuente en dicho instante, con lo que la anterior expresión es igual a:

$$\frac{p\{\, \text{Una celda de clase j llega en } \tau \,/\, J(\tau) = i \,\}\; p\{\, I(\tau) = k,\, J(\tau) = i \,\}}{p\{\, \text{Una celda de clase j llega en } \tau \,\}} =$$

$$= \sum_{t=1}^{m} D_j^{it}\; y_k^i \;/\; \bar{\theta}\, D_j\, \bar{e}.$$

A continuación calcularemos las probabilidades de estado en un instante arbitrario.

Distribución de la longitud de cola en un instante arbitrario

Para calcular las probabilidades de estado en un instante arbitrario seguiremos un camino análogo al usado en el análisis de la cola BMAP/G/1 de capacidad finita.

En primer lugar calcularemos la probabilidad de que el la cola esté vacía, es decir $\bar{y}_0$. Aplicando los mismos argumentos a los usados en el capítulo 2, obtenemos:

$$\bar{y}_0 = \bar{x}_0\, C_T^{-1} (h - \bar{x}_0\, C_T^{-1} \bar{e}\,)^{-1}$$

Ahora las otras probabilidades pueden calcularse así:

Sea τ un instante arbitrario de tiempo durante un periodo en donde el servidor esté ocupado, y llamemos G_f (resp. G_b) al tiempo de recurrencia hasta el próximo servicio (resp. desde el anterior servicio).

Definiremos los vectores $\bar{\omega}_n$ (t) dt, n = 1, 2, ..., N con componentes:

ω_n^i(t) dt = p{ En el instante τ el sistema tiene n clientes, el estado de la fuente es i y $t < G_f \leq t+dt$ / El servidor está ocupado en τ }

Para evaluar los anteriores vectores necesitamos la probabilidad conjunta de que haya n llegadas durante G_b y que $t < G_f \leq t+dt$. Sea τ_k el instante en que empezó el servicio actual. Entonces definimos las siguientes probabilidades:

$H_n^{T\ ij}$ (t) dt = p{ Durante G_b tenemos n llegadas de clientes de cualquier clase, $t < G_f \leq t+dt$ y el estado de la fuente en τ_k+G_b es j / el estado de la fuente en τ_k era i }

$H_n^{p\ ij}$ (t) dt = p{ Durante G_b tenemos n llegadas de clientes prioritarios, $t < G_f \leq t+dt$ y el estado de la fuente en τ_k+G_b es j / el estado de la fuente en τ_k era i }

y también:

$H_{n1,n2}^{ij}$ (t) dt = p{ Durante G_b hay primero n_1 llegadas de celdas de cualquier clase seguidas de n_2 llegadas de celdas prioritarias, $t < G_f \leq t+dt$ y el estado de la fuente en τ_k+G_b es j / el estado de la fuente en τ_k es i }

Las expresiones que se obtienen son:

$$H_n^T(t) = (H_n^{T\ ij}(t)) = h^{-1}\int_0^\infty P_T(n,s)h(t+s)\,ds \qquad (6.3)$$

$$H_n^p(t) = (H_n^{p\ ij}(t)) = h^{-1}\int_0^\infty P_p(n,s)h(t+s)\,ds$$

$$H_{n1,n2}(t) = (H_{n1,n2}^{ij}(t)) = h^{-1}\int_0^\infty h(t+s)\,ds \int_0^s P_T(n_1-1,r)\,dP_T/dt\,(1,0)\,P_p(n_2,s-r)\,dr$$

Demostración:

De las definiciones tenemos:

$H_n^{T\ ij}$ (t) dt = $\int_0^\infty$ p{ Durante G_b llegan n celdas de cualquier clase, $J(\tau_k+Gb) = j$ / $G_f = t$, $G_b = s$, $J(\tau_k) = i$ } p{ $G_f = t$ / $G_b = s$, $J(\tau_k) = i$ } p{ $G_b = s$ / $J(\tau_k) = i$ } dsdt.

G_b es independiente del estado de la fuente en τ_k, por lo que

$$p\{ G_b = s / J(\tau_k) = i \} = p\{ G_b = s\} = h^{-1}(1-H(s))$$

$$p\{ G_f = t / G_b = s, J(\tau_k) = i \} = p\{ G_f = t / G_b = s \} = h(t+s)/(1-H(s))$$

$$p\{ \text{Durante } G_b \text{ hay n llegadas de celdas de cualquier clase, } J(\tau_k+G_b) = j / G_f = t, G_b = s, J(\tau_k) = i \} = p\{ \text{Durante } G_b \text{ hay n llegadas de celdas de cualquier clase, } J(\tau_k+G_b) = j / G_b = s, J(\tau_k) = i \} = P_T^{ij}(n,s)$$

y finalmente:

$$H_n^T(t) = h^{-1} \int_0^\infty P_T(n,s)h(t+s)\ ds$$

Las otras matrices se obtienen mediante argumentos análogos. Esta expresión es una alternativa a la dada en [Blo89].

Ahora:

$$\bar{\omega}_n(t) = \bar{x}_0\ UH_{n-1}^T(t) + \sum_{j=1}^{n} \bar{x}_j\ H_{n-j}^T(t) \qquad 0 < n < S_2+1$$

y

$$\bar{\omega}_n(t) = \bar{x}_0\ UH_{S2,n-S2-1}(t) + \sum_{j=1}^{S2} \bar{x}_j\ H_{S2+1-j,n-S2-1}(t) + \sum_{j=S2+1}^{n} \bar{x}_k\ H_{n-j}^p(t) \qquad S_2+1 \le n \le S$$

Finalmente, los vectores $\bar{y}_k$ pueden ser obtenidos como:

$$\bar{y}_k = P_{ocup} \int_0^\infty \bar{\omega}_k(t)\ dt \quad \text{para} \quad 1 \le k \le S; \quad \bar{y}_{S+1} = \bar{\theta} - \sum_{j=0}^{S} \bar{y}_j$$

en donde $P_{ocup} = 1-P_{desoc}$ y $P_{desoc} = \bar{y}_0 \bar{e}$.

El cálculo de las anteriores integrales puede hacerse como:

$$\int_0^\infty H_n^T(t)\,dt = h^{-1} \sum_{j=0}^{\infty} \alpha_j^T K_n^{T(j)}; \qquad \alpha_j^T = \int_0^\infty dt \int_0^\infty e^{-\theta_T u} (\theta_T u)^j / j!\; h(t+u)du$$

$$\int_0^\infty H_n^p(t)\,dt = h^{-1} \sum_{j=0}^{\infty} \alpha_j^p K_n^{p(j)}; \qquad \alpha_j^p = \int_0^\infty dt \int_0^\infty e^{-\theta_p u} (\theta_p u)^j / j!\; h(t+u)du$$

$$\int_0^\infty H_{n1,n2}(t)\,dt = h^{-1} \sum_{j=0}^{\infty} \sum_{i=0}^{\infty} \alpha_{ij} K_{n1-1}^{T(j)} D_T K_{n2}^{p(i)}$$

$$\alpha_{ij} = \int_0^\infty dt \int_0^\infty du \int_0^u e^{-\theta_m u} \xi^j (u-\xi)^i \theta_m^{i+j} / (i+j)!\; h(t+u)\,d\xi; \qquad \theta_m = \max(\theta_T, \theta_p)$$

La distribución del retardo para el mecanismo PBS se verá, junto con la del mecanismo PO, en el próximo apartado.

6.3 Modelo de un multiplexor con política PO y MAP como entrada ([GarCas90.b], [GarCas91.a]).

En el caso de una cola con el mecanismo PO, el número total de celdas en el sistema es el mismo que para la cola MAP/G/1/S+1 en donde el tráfico de entrada tiene las mismas características que el tráfico total de nuestro sistema. Ello es debido a que se produce una pérdida cuando una celda que llega encuentra el sistema lleno, y solo en ese caso. De ahí, y del hecho de que todas las celdas reciben el mismo tipo de servicio, se deduce la siguiente ley de conservación [SumOza89]:

B_T = Probabilidad de pérdida total.

$$B_T = B_p (\bar{\theta}\, D_p\, \bar{e}) / (\bar{\theta}\, D_T\, \bar{e}) + B_{np} (\bar{\theta}\, D_{np} \bar{e}) / (\bar{\theta}\, D_T\, \bar{e}) \qquad (6.4)$$

En el capítulo 2 se ha estudiado las características de la cola finita sin prioridades lo que nos permitirá calcular B_T. Por lo tanto es suficiente evaluar la probabilidad de pérdida para una clase de celdas, pues la probabilidad para la otra clase puede ser deducida de la ley de conservación anterior.

Encontraremos la probabilidad de pérdida para las celdas ordinarias. Para ello observaremos una celda ordinaria desde el momento en que llega al sistema hasta que alcanza el umbral S_2. Dicha celda puede ser perdida en el instante de llegada, si

encuentra la cola llena, o puede ser perdida al ser reemplazada por una celda prioritaria:

$$B_{np} = p\{ \text{La celda que llega en } \tau \text{ se pierde / En } \tau \text{ llega una celda ordinaria}\} =$$

$$= \sum_{k=0}^{S+1} \sum_{i=1}^{m} p\{ \text{La celda que llega en } \tau \text{ se pierde, hay k clientes en el sistema en } \tau \text{ y el estado de la fuente es i / En } \tau \text{ llega una celda ordinaria}\}$$

Ahora, para $k > S_2$ definimos:

$$P^{ij}_{(servida/k)} = p\{ \text{La celda que llega en } \tau \text{ es servida y el estado de la fuente cuando esta celda llega a la posición } S_2 \text{ es j / En } \tau \text{ llega una celda ordinaria, hay k clientes en el sistema y el estado de la fuente es i }\}$$

$$P_{(servida/k)} = (P^{ij}_{(servida/k)}) \qquad i,j = 1,...,m$$

y, como en el apartado anterior,

$$p_k^{np\ i} = p\{ \text{Un celda ordinaria que llega encuentra el sistema con k clientes y la fuente está en estado i }\};$$

$$\bar{p}_k^{np} = (p_k^{np\ 1},..., p_k^{np\ m})$$

Al igual que sucedía en el análisis de la cola con mecanismo PBS, no podemos aplicar la propiedad PASTA, que facilitaba el análisis en el caso del sistema con llegadas de tipo Poisson. Por lo tanto usamos la expresión (6.2), deducida en el apartado anterior

$$p_k^{np\ i} = \sum_{j=1}^{m} D^{ij}_{np}\ z_k^i \ / \bar{\theta}\ D_{np}\ \bar{e} \qquad i = 1,...,m$$

en donde z_k^i son las probabilidades de estado en un instante arbitrario para la cola sin prioridades. De ahí se deduce:

$$B_{np} = \sum_{k=S2+1}^{S+1} \bar{p}_k^{np} (\bar{e} - P_{(servida/k)}\ \bar{e})$$

Ahora nos falta calcular las matrices $P_{(servida/k)}$, para lo que usaremos las

siguientes expresiones

$A_k^{ij}(n)$ = p{ Hay n llegadas de celdas prioritarias durante G_f y $J(\tau+G_f)=j$ / En τ llega una celda ordinaria, $I(\tau)=k$ y $J(\tau)=i$ }

$A_k(n) = (A_k^{ij}(n))$ $\quad i,j = 1,...,m$

$A^{ij}(n)$ = p{ Hay n llegadas de celdas prioritarias durante un servicio y el estado final de la fuente es j / el estado inicial de la fuente es i }

$A(n) = (A^{ij}(n))$ $\quad i,j = 1,...,m$

De lo anterior es inmediato obtener:

$$A(n) = \int_0^\infty P_p(n,t)\, dH(t)$$

La expresión para $A_k(n)$ es la siguiente:

$$A_k(n) = \int_0^\infty r_k(t) P_p(n,t) dt$$

en donde $r_k(t)$ está definido como una matriz diagonal, cuyos elementos en la diagonal son:

$r_k^{ii}(t)$ = p{ $G_f = t$ / una celda ordinaria llega en τ y hay k clientes en el sistema y el estado de la fuente es i }

Ahora demostraremos que:

$$r_k^{ii}(t) = P_{busy}\, \omega_k^i(t) / z_k^i \qquad (6.5)$$

Demostración:

$r_k^{ii}(t)$ = p{ $G_f = t$ / una celda ordinaria llega en τ, $L(\tau) = k$, $J(\tau) = i$ } =

$$\frac{p\{ \text{una celda ord. llega en } \tau / G_f = t,\ L(\tau) = k,\ J(\tau) = i \}\ p\{ G_f = t,\ L(\tau) = k,\ J(\tau) = i \}}{p\{ L(\tau) = k,\ J(\tau) = i / \text{una celda ord. llega en } \tau \}\ p\{ \text{una celda ordinaria llega en } \tau \}} =$$

$$= \sum_{j=1}^{m} D_{np}^{ij} P_{busy} \omega_k^i(t) / \bar{\theta} D_{np}\, \bar{e}\ p_k^{np\, i} = P_{busy}\, \omega_k^i(t) / z_k^i$$

Esta expresión generaliza para la cola MAP/G/1 la obtenida en [Kro90] para la cola M/G/1.

Definimos $\omega_k^i(t)$ y $H_n^T(t)$ de igual forma que en el apartado anterior. Ahora tenemos:

$$\bar{\omega}_k(t) = \bar{z}_0\, UH_{k-1}^T(t) + \sum_{j=1}^{k} \bar{z}_j\, H_{k-j}^T(t) \qquad 0 < k \le S \tag{6.6}$$

$$\bar{\omega}_{S+1}(t) = \bar{z}_0 \sum_{j=S+1}^{\infty} UH_{j-1}^T(t) + \sum_{j=1}^{S} \bar{z}_j \sum_{k=S+1-j}^{\infty} H_k^T(t)$$

Substituyendo (6.5) y (6.6) en la expresión de $A_k(n)$ y denotando como $d(\bar{a})$ la matriz diag(a^1, ..., a^m) para cualquier vector $\bar{a}$, obtenemos las siguientes integrales que pueden ser calculadas usando el algoritmo dado en el capítulo 2:

$$\int_0^{\infty} d(\bar{z}_0\, UH_{k-1}^p(t))\, H_n^T(t)\, dt = \sum_{j=0}^{\infty} \sum_{i=0}^{\infty} \beta_{ij}\, d(\bar{z}_0\, UK_{k-1}^{p(j)})\, K_n^{T(i)}$$

$$\int_0^{\infty} d(\bar{z}_f\, H_{k-f}^p(t))\, H_n^T(t)\, dt = \sum_{j=0}^{\infty} \sum_{i=0}^{\infty} \beta_{ij}\, d(\bar{z}_f\, UK_{k-f}^{p(j)})\, K_n^{T(i)}$$

$$\beta_{ij} = \int_0^{\infty} dt \int_0^{\infty} e^{-\theta_m(t+\xi)}\, t^j\, \xi^i\, \theta_m^{\,i+j} / (i!\, j!)\; h(\xi+t)d\xi; \qquad \theta_m = \max(\theta_T, \theta_p)$$

Asumiendo que la celda ordinaria que estamos observando entra en el sistema en la posición k (para k= S_2+1,...,S), se mueve hacia la posición k-1 al final del servicio siempre y cuando durante el tiempo residual hasta el final de dicho servicio han llegado menos de S-k celdas prioritarias. En general, si dicha celda llega a la posición k-f > S_2, será desplazada a la posición k-f-1 si durante el tiempo residual inicial y los f siguientes servicios llegan menos de S-k+f celdas prioritarias. Finalmente, la probabilidad conjunta de ser servida y del estado de la fuente es la probabilidad conjunta de que la celda sea capaz de llegar hasta la posición S_2 y el estado de la fuente en dicho instante (ver figura 6.3).

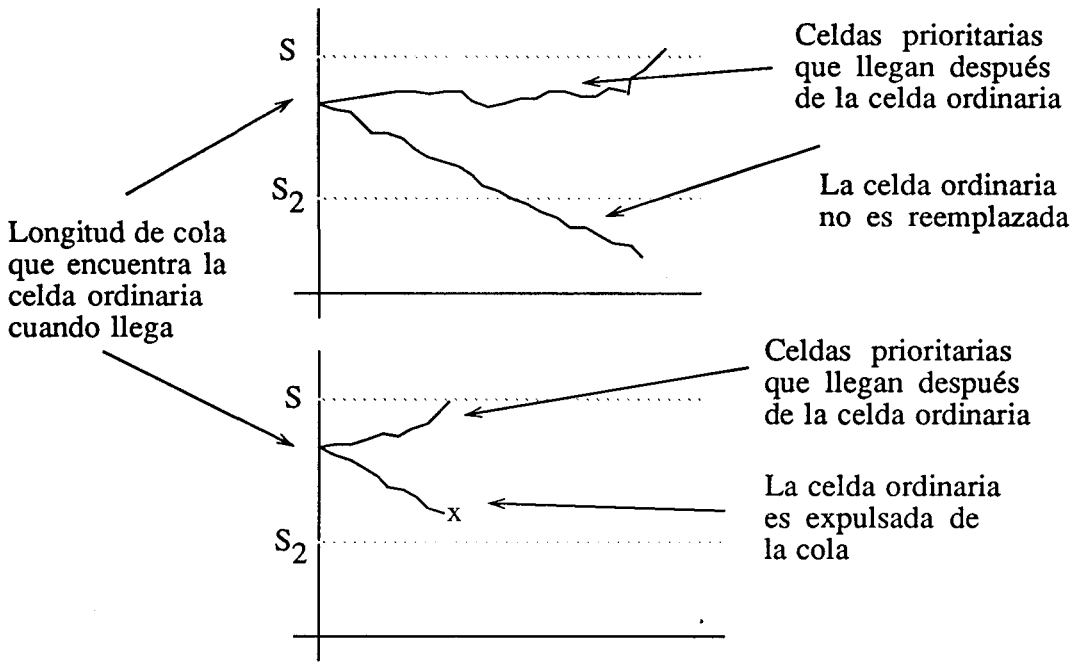


Figura 6.3

De ahí se puede establecer el siguiente algoritmo, que es una extensión del usado en [Kro90] ver (5.3):

$C_f^{ij}(k,n) = p\{$ La celda ordinaria que está siendo observada se mueve desde la posición k-f a la k-f-1 y n celdas prioritarias llegan desde que ha entrado en la cola y el estado final de la fuente es j / el estado inicial de la fuente era i $\}$

$C_f(k,n) = (C_f^{ij}(k,n))$ $\quad i,j = 1,\ldots,m$

Paso 1:

$$C_0(k,n) = \begin{cases} A_k(n) & \text{si } 0 \le n \le S\text{-}k \\ 0 & \text{en otro caso} \end{cases}$$

Paso f: $1 \le f \le k\text{-}S_2\text{-}1$

$$C_f(k,n) = \begin{cases} C_{f-1}(k,n) \otimes A(n) & \text{si } 0 \le n \le S\text{-}k+f \\ 0 & \text{en otro caso} \end{cases}$$

el operador $\otimes$ denota la operación de convolución en n.

Finalmente tenemos:

$$P_{(servida/k)} = \sum_{n=0}^{S-S2-1} C_{k-S2-1}(k,n)$$

La distribución del retardo para los mecanismos PBS y PO

Sea $V_T(t)$ el tiempo que un cliente de cualquier clase que entra en el sistema en el instante t debe esperar antes de ser servido. $V_T(t)$ está definido solo para los clientes que no son perdidos. También definiremos $W_T^j(x)$ como

$$W_T^j(x) = p\{ V_T(t) \leq x, J(t) = j\}$$

y $W_T^{*j}(s)$ es la transformada de Laplace de la función de densidad correspondiente.

De forma análoga se puede definir $W_p^j(x)$ para el tiempo de espera para las celdas prioritarias y $W_{np}^j(x)$ para las ordinarias. $W_p^{*j}(s)$ y $W_{np}^{*j}(s)$ denotan las transformadas de Laplace de las funciones de densidad correspondientes.

Tenemos:

$$W_i^j(x) = p\{ V(t) \leq x, J(t) = j \,/\, \text{Llega en t una celda de clase i y será servida } \} =$$

$$\frac{\sum_{k=0}^{S} p\{ V(t) \leq x, J(t) = j, L(t)= k, \text{llega en t una celda de clase i y será servida } \}}{p\{ \text{Llega en t una celda de clase i y será servida } \}}$$

en donde

$$p\{ \text{Llega en t una celda de clase i y será servida } \} = (1-B_i)\,\bar{\theta}\,D_i\,\bar{e}$$

y:

$$\sum_{k=0}^{S} p\{ V(t) \leq x, J(t) = j, L(t) = k, \text{llega en t una celda de clase i y será servida } \} =$$

$$\sum_{k=0}^{S} p\{ V(t) \leq x \,/\, J(t) = j, L(t) = k, \text{llega en t una celda de clase i y será servida } \}.$$

$$p\{ \text{La celda que llega será servida} \,/\, \text{en t llega una celda de clase i, } J(t) = j, L(t) = k \}$$

$$p\{ \text{en t llega una celda de clase i} \,/\, J(t) = j \}\; p\{ J(t) = j, L(t) = k \}.$$

Razonamos igual que en el apartado 2.3 : Si en instante t el servidor está ocupado, entonces el tiempo de espera es el tiempo residual de servicio para el cliente que ocupa el servidor más los tiempos de servicio que estás esperando en dicho instante t. Si el servidor está libre, el tiempo de espera es cero.

De esta forma obtenemos, para el PBS:

$$W_{np}^{*j}(s) = \frac{\sum_{i=1}^{m} D_{np}^{ji}}{(1-B_{np})\,\bar{\theta}\,D_{np}\,\bar{e}}\left(y_0^j + \sum_{n=1}^{S2} P_{ocup}\,\omega_n^{*j}(s)(H^*(s))^{n-1}\right)$$

$$W_p^{*j}(s) = \frac{\sum_{i=1}^{m} D_p^{ji}}{(1-B_p)\,\bar{\theta}\,D_p\,\bar{e}}\left(y_0^j + \sum_{n=1}^{S} P_{ocup}\omega_n^{*j}(s)(H^*(s))^{n-1}\right)$$

La expresión de $W_{np}^{*j}(s)$ para el mecanismo PO es:

$$W_{np}^{*j}(s) = \frac{\sum_{i=1}^{m} D_{np}^{ji}}{(1-B_{np})\,\bar{\theta}\,D_{np}\,\bar{e}}\ x$$

$$\left(z_0^j + \sum_{n=1}^{S2} P_{busy}\,\omega_n^{*j}(s)(H^*(s))^{n-1} + \sum_{n=S2+1}^{S} \left(P_{(served/n)\bar{e}} \right)_j P_{busy}\,\omega_n^{*j}(s)(H^*(s))^{n-1}\right)$$

La expresión del retardo para las celdas prioritarias para el mecanismo PO parece difícil de obtener.

Finamente, para ambos mecanismos tenemos:

$$W_T^{*j}(s) = \frac{(1-B_{np})\,\bar{\theta}\,D_{np}\,\bar{e}\,W_{np}^{*j}(s) + (1-B_p)\,\bar{\theta}\,D_p\,\bar{e}\,W_p^{*j}(s)}{(1-B_{np})\,\bar{\theta}\,D_{np}\bar{e} + (1-B_p)\,\bar{\theta}\,D_p\,\bar{e}}$$

6.4 Modelo de un multiplexor con política PBS usando la aproximación de fluido [GarCas91.c].

A continuación vamos a estudiar un modelo de un multiplexor estadístico usando la aproximación de fluido, presentada en el capítulo 2.

El proceso de entrada

Los dos tipos de tráfico, de celdas ordinarias y prioritarias, son incorporados a nuestro modelo del siguiente modo:

El proceso de entrada consiste en la superposición del tráfico generado por M fuentes de dos estados, estados *activo* e *inactivo*. Cuando una fuente está en estado *activo* emite un flujo contínuo de información a un ritmo de A_T unidades de información por unidad de tiempo. De este tráfico, A_p corresponde al tráfico de celdas prioritarias y A_{np} al de celdas ordinarias. Cuando está en estado *inactivo* no emite información. Los tiempos de permanencia en cada estado siguen una distribución exponencial. La media de permanencia en estado *activo* valdrá $1/\mu$ y en estado *inactivo* $1/\lambda$,ver figura (6.4))

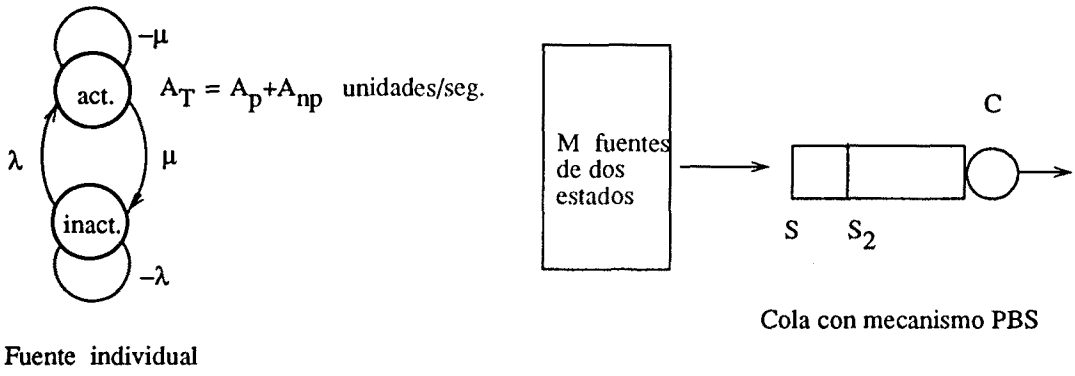


Figura 6.4

Dicho proceso constituye la entrada a una cola de capacidad finita, de longitud S gestionada por un mecanismo de prioridad espacial PBS, siendo S_2 la longitud de la cola compartida, con un servidor capaz de cursar C unidades de información por unidad de tiempo.

La longitud de cola en estado estacionario

Al igual que en el capítulo 2 , consideramos la probabilidad conjunta de la longitud de cola y número de fuentes activas en estado estacionario. De una forma totalmente análoga al caso de la cola finita, encontramos las siguientes ecuaciones:

$$d\bar{F}(x)/dx = A_T \bar{F}(x) \quad 0 < x < L_2 \qquad (6.6)$$

$$d\bar{F}(x)/dx = A_p \bar{F}(x) \quad L_2 < x < L$$

en donde las matrices A_i cumplen:

$$A_i = D_i^{-1} E \qquad (i = t,p)$$

con

$$E=\begin{vmatrix} -M\lambda & \mu & & & & \\ M\lambda & -((M-1)\lambda+\mu) & 2\mu & & & \\ & (M-1)\lambda & -((M-2)\lambda+2\mu) & & & \\ & & \ddots & \ddots & & \\ & & & 2\lambda & -(\lambda+(M-1)\mu) & M\mu \\ & & & & \lambda & -M\mu \end{vmatrix}$$

y

$$D_i = \text{diag}(-C,\ A_i-C,\ 2A_i-C,\ldots,MA_i-C).$$

Definimos los valores y vectores propios de A_t por la derecha como z_k y $\bar{\phi}_k$. Los de la matriz A_p serán y_k and $\bar{\theta}_k$. Estas matrices son idénticas a las encontradas en el análisis de las colas sin prioridades, de forma que las expresiones de los valores y vectores propios son ya conocidas (apartado 2.4).

La solución del sistema de ecuaciones diferenciales anterior es de la forma:

$$\bar{F}(x) = \sum_{k=0}^{M} a_k \exp(z_k x)\,\bar{\phi}_k \qquad \text{si } 0 < x < L_2$$

$$\bar{F}(x) = \sum_{k=0}^{M} b_k \exp(y_k x)\,\bar{\theta}_k \qquad \text{si } L_2 < x < L$$

Cálculo de los coeficientes

Los coeficientes a_k y b_k deben ser calculados imponiendo condiciones de contorno en los puntos x=0, L_2 y L. Para cada componente $F_i(x)$ del vector $\bar{F}(x)$ podemos establecer las siguientes condiciones:

- Si $A_T*i > c$ entonces $F_i(0^+) = F_i(0) = 0$.
- Si $A_p*i < c$ entonces $F_i(L^-) = F_i(L) = p\{$ Hay i fuentes activas $\}$.
- Si $A_T*i > c$ y $A_p*i > c$ entonces $F_i(L_2^-) = F_i(L_2^+)$.
- Si $A_T*i < c$ y $A_p*i < c$ entonces $F_i(L_2^-) = F_i(L_2^+)$.

El razonamiento que nos lleva a establecer las condiciones anteriores en x=0 y x=L son las mismas que las usadas en el análisis de la cola finita. En las figuras (6.5 y 6.6) se muestran dos ejemplos que corresponden a las condiciones de contorno en L_2.

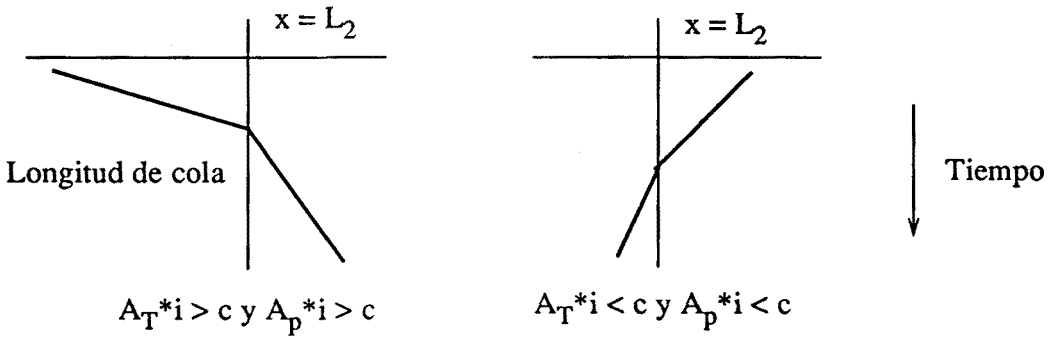


Figura 6.5

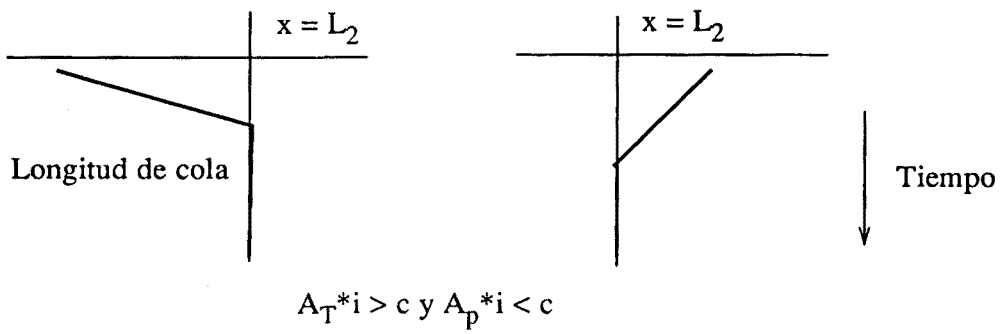


Figura 6.6

De aquí deducimos que las funciones $F_i(x)$ deben tener la siguiente forma:

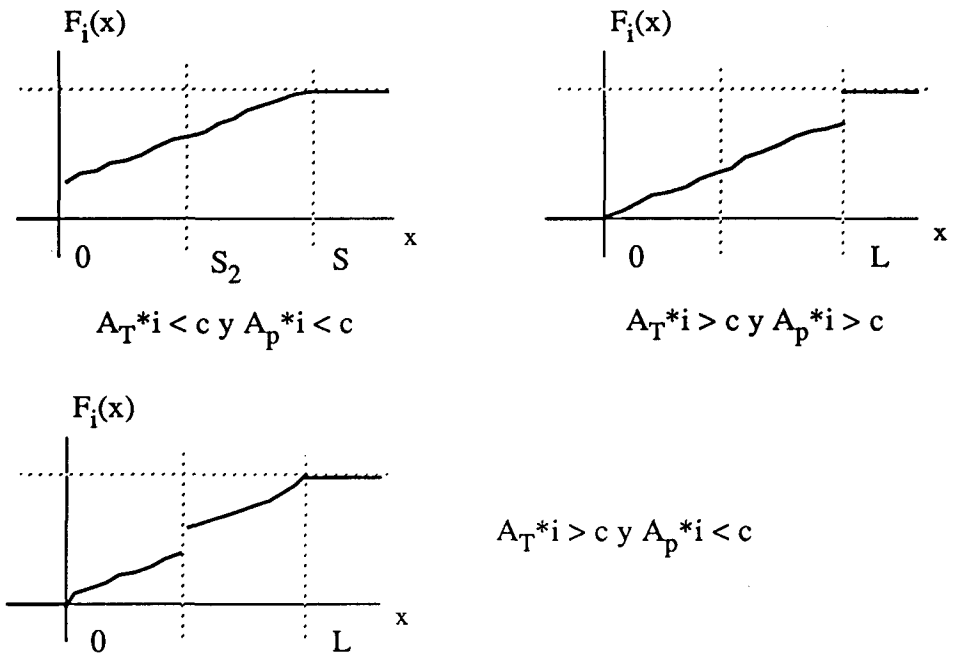


Figura 6.7

de modo que por cada función $F_i(x)$ tenemos dos ecuaciones independientes. Resolviendo el sistema lineal de 2(M+1) ecuaciones, encontramos a_k and b_k.

La probabilidad de pérdida para el tráfico prioritario viene dado por.

$$Bp = (1/m_p) \sum_{k=m}^{M} (Ap*k-c)(F_k(L^+) - F_k(L^-))$$

donde m es el mínimo valor de k para el que la expresión A_p*k-C es positiva.

En la probabilidad de pérdida del tráfico ordinario tenemos dos contribuciones: Las celdas ordinarias pueden perderse cuando la longitud de cola es mayor que L_2 o cuando es precisamente L_2. En la primera situación todas las celdas que llegan son pérdidas. En la segunda situación parte de las celdas ordinarias pueden entrar en la cola.

$$Bnp = (1/m_{np}) \sum_{k=0}^{M} Anp*k*(F_k(L^+) - F_k(L_2^+)) +$$
$$(1/m_{np})*(Anp/At) \sum_{k=n}^{M} (At*k- c)*(F_k(L_2^+) - F_k(L_2^-))$$

donde n es el mínimo valor de k para el que la expresión A_T*k-C es positiva.

Inclusión de tráfico generado por fuentes CBR

El tráfico generado por fuentes CBR puede ser incluido facilmente en nuestro modelo si consideramos que un flujo continuo de celdas llega hasta nuestra cola independientemente del estado de las fuentes que componen el tráfico VBR.

Si el tráfico total generado por las fentes CBR es At^{cbr} y el tráfico prioritario es Ap^{cbr}, nuiestro modelo es equivalente, en cuanto se refiere a la longitud de cola, a considerar que el fñlujo de salida se modifica, de forma que vale $C_1 = C - At^{cbr}$ cuando la longitud de cola es menor a L_2 y $C_2 = C - Ap^{cbr}$ cuando dicha longitud es mayor a L_2.

Las condiciones de contorno son ahora:

- Si $At^*i > c_1$ entonces $F_i(0^+) = F_i(0) = 0$.
- Si $Ap^*i < c_2$ entonces $F_i(L^-) = F_i(L) = p$(Hay i fuentes activas).
- Si $At^*i > c_1$ y $Ap^*i > c_2$ entonces $F_i(L_2^-) = F_i(L_2^+)$.
- Si $At^*i < c_1$ y $Ap^*i < c_2$ entonces $F_i(L_2^-) = F_i(L_2^+)$.

y las probabilidades de pérdida son, para las fuentes VBR,las siguientes:

$$Bp^{vbr} = (1/m_p) \sum_{k=0}^{M} Ap^*k(1-c\ /(Ap^*k + Ap^{cbr}))^*(F_k(L^+) - F_k(L^-))$$

$$Bnp^{vbr} = (1/m_{np}) \sum_{k=0}^{M} Anp^*k^*(F_k(L^+) - F_k(L_2^+)) + $$

$$+(1/m_{np})(Anp/(At+At^{cbr}))^* \sum_{k=s}^{M} (At^*k - c1)(F_k(L_2^+) - F_k(L_2^-))$$

donde s es el mínimo valor de k para el que la expresión A_T*k-C1 es positiva.

Para las fuentes CBR tenemos:

$$Bp^{cbr} = \frac{\sum_{k=0}^{M} Ap^{cbr}(1-c\ /(Ap^*k + Ap^{cbr}))(F_k(L^+) - F_k(L^-))}{Ap^{cbr}}$$

$$Bnp^{cbr} = \sum_{k=0}^{M} Anp^{cbr}\ ^*(F_k(L^+) - F_k(L_2^-))/\ Anp^{cbr}$$

6.5 El uso de mecanismos de prioridad espacial

De la introducción de prioridades espaciales en la red cabe esperar:

- Un incremento de la carga admisible por la red,
- Un ahorro en el tamaño de la memoria requerida por los conmutadores.

Por contra, tendremos:

- Un aumento en la complejidad de los conmutadores,
- Un aumento en la complejidad de los terminales.

Para estudiar si es o no conveniente introducir prioridades en la red debemos hacer una valoración de los incrementos de carga y ahorros de espacio frente al incremento de coste introducido por el uso de prioridades espaciales. En la determinación del incremento de coste entran en juego factores tecnológicos (coste de las memorias de alta velocidad, etc) que no han sido abordados en este trabajo. Nos restringimos, pues, a hacer una estimación de los beneficios que cabe esperar del uso de prioridades, a hacer una comparación entre dos mecanismos propuestos, y a estudiar sus características más relevantes.

Para hacer un estudio de este tipo es esencial tener en cuenta el tráfico que deben servir los conmutadores. La caracterización del tráfico presente en una red ATM en condiciones reales de funcionamiento es un problema muy complejo, que aún no ha sido resuelto [KUH91]. Por lo tanto las conclusiones que se pueden obtener al estudiar el sistema variando una serie de parámetros que definen nuestro modelo de fuente deben ser consideradas con precaución.

En una primera aproximación al problema, simplificaremos nuestro sistema usando un modelo aproximado: supondremos que el mecanismo usado es el de push-out, y que la probabilidad de pérdida de las celdas prioritarias es despreciable (es decir, consideraremos que $B_p = 0$). De esta forma tenemos que la probabilidad de pérdida de las celdas ordinarias se puede calcular facilmente a partir de la de una cola sin prioridades, usando la ley de conservación (6.4). En condiciones normales de funcionamiento esta aproximación será muy buena, pues cabe esperar que B_p tome valores muchos ordenes de magnitud por debajo de B_{np} (ver figura 6.22). Más tarde usaremos los modelos exactos desarrollados para obtener una información más precisa y confirmar las conclusiones generales que se derivan de esta primera aproximación.

Durante nuestro estudio, consideraremos tres posibles escenarios en donde el tráfico tendrá diferentes niveles de variabilidad. No intentaremos dar una definición precisa de 'nivel de variabilidad' pero para fijarlo serán factores muy importantes la duración de los periodos de sobrecarga y la diferencia entre el tráfico de pico y el tráfico medio. A este último cociente lo denominaremos 'burstiness'. Por lo tanto diferenciaremos entre tres tipos de tráfico:

- *Tráfico de Poisson.* Es el resultado de la superposición de un número elevado de fuentes con niveles de variación moderados.

- *Tráfico con niveles de variabilidad pequeños.* Por ejemplo, sería el resultado de la superposición de un número elevado de fuentes de dos estados con unos niveles de actividad moderados o medios, o la superposición de fuentes VBR con bajos niveles de variación (por ejemplo, codificadores de video). Cuando usemos un modelo basado en un MMPP de dos estados, consideraremos que los tiempos de permanencia en cada estado son pequeños (menores a 100 μseg., es decir, con r_1 y

r_2 mayores a 1.0 E-2). También usaremos los resultados obtenidos del modelo aproximado de [Kroetal91] y del modelo basado en la aproximación de fluido para el mecanismo PBS de [GarCas91.c]. En este caso, superpondremos fuentes de dos estados en donde el tráfico en estado de actividad de cada fuente individual no sea muy grande frente al tráfico medio, y cuyo tráfico medio sea pequeño frente a la velocidad del enlace.

- *Tráfico con niveles de variabilidad grandes*: Tendremos grandes variaciones entre los estados de actividad y de silencio y durante un tiempo elevado, por ejemplo del orden de mseg., se producen situaciones en donde el tráfico instantáneo supera en mucho al tráfico medio. Cuando usemos un MMPP de dos estados consideraremos que los tiempos de estancia en cada estado son grandes (p.e. r_1 y r_2 mayores a 1.0 E-4) con importantes diferencias entre el tráfico en estado de alta actividad y el tráfico medio. También usaremos los modelos para la superposición de fuentes de dos estados suponiendo fuentes individuales con tráficos medios elevados y grandes diferencias entre estados de actividad y tráfico medio.

El ahorro en espacio de cola

En las figuras 6.8, 6.9 y 6.10, se muestra la probabilidad de pérdida frente al tamaño de cola, para diferentes valores de carga del conmutador. Se han señalado los niveles de probabilidad de pérdida de 1.0 E-6 y 1.0 E-10.

En todos los casos se observa un comportamiento de tipo exponencial para la probabilidad de pérdida en función del espacio. Además, para el caso de colas muy pequeñas, se observa una confluencia de las curvas (siempre y cuando la carga no tome valores extremadamente bajos). De ahí se deduce que:

- **El ahorro en tamaño va a ser en gran medida independiente de las condiciones en que hagamos el estudio.** El tanto por ciento de tamaño de cola que ahorraremos al usar prioridades vendrá fijado, principalmente, por las diferencias entre las probabilidades mínimas exigidas para cada tipo de celda. Si estos valores son de 1 E-6 y 1 E-10, el ahorro en el tamaño de cola que cabe esperar es de un 40%.

- **El porcentaje de tráfico prioritario sobre el total va a influir poco en los tamaños de cola.** Solo será importante cuando tome valores próximos al 100 %.

Se sabe de otros estudios [Kroetal90], que el MMPP de dos estados no captura el fenómeno de pérdidas a nivel de ráfaga. En ese caso las curvas tienen un codo y su pendiente disminuye a partir de un cierto valor del tamaño del buffer. Sin embargo el comportamiento hasta ese nivel es también aproximadamente exponencial, por lo que las conclusiones anteriores siguen siendo ciertas.

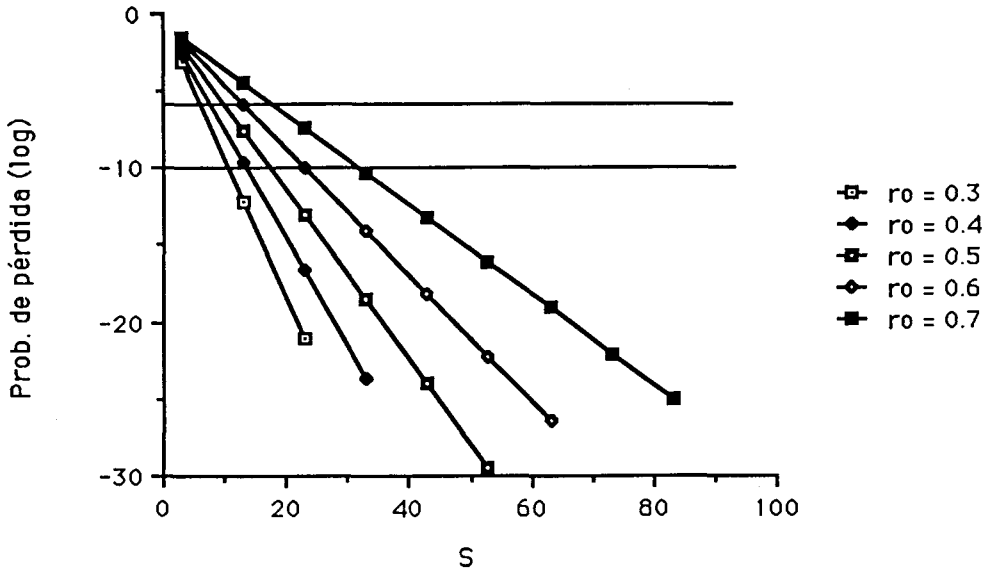


Figura 6.8 : Probabilidad de pérdida frente al tamaño de cola para diferentes valores de carga. Tráfico de Poisson.

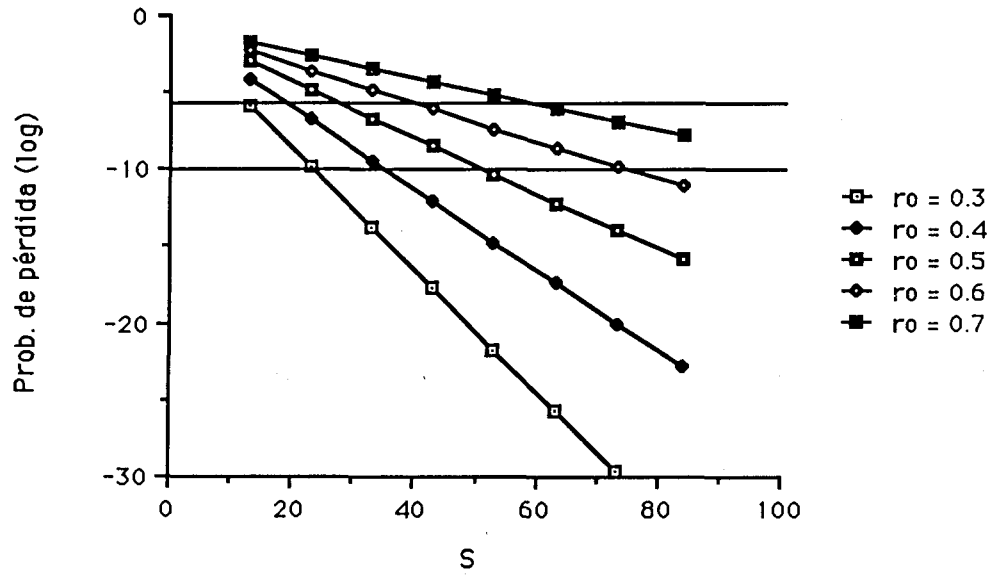


Figura 6.9 : Probabilidad de pérdida frente al tamaño de cola para diferentes valores de carga. MMPP de dos estados, con r_1 = 1.0 E-1, r_2 = 5.0 E-1, burstiness = 3. El estado de alta actividad es el estado 2.

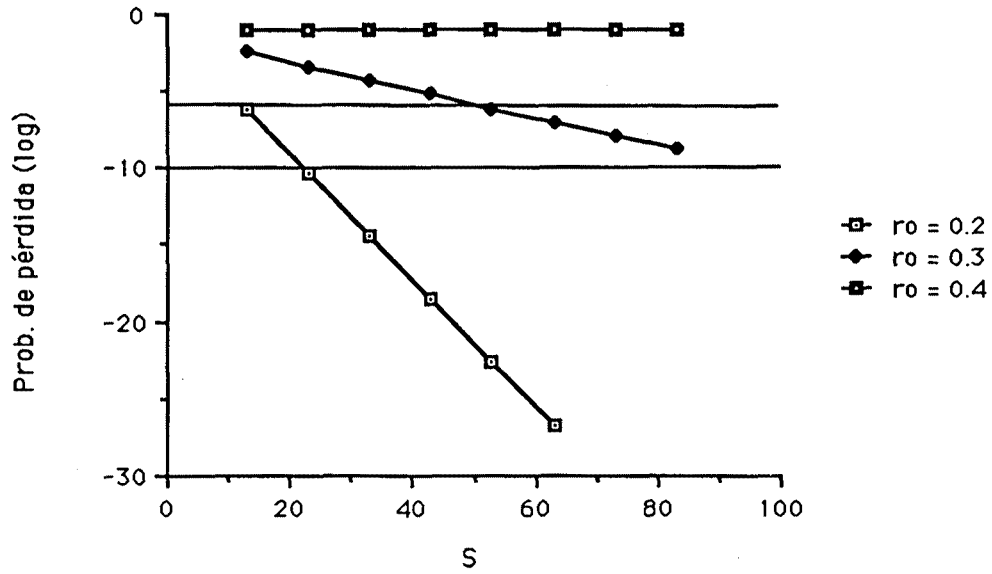


Figura 6.10 : Probabilidad de pérdida frente al tamaño de cola para diferentes valores de carga. MMPP de dos estados, con r_1 = 1.0 E-5, r_2 = 5.0 E-5, burstiness = 3. El estado de alta actividad es el estado 2.

El incremento de la carga admisible

El incremento de la carga admisible por el multiplexor parece muy dependiente de las condiciones de tráfico:

- **Para tráfico de Poisson, el incremento de la carga admisible está en el rango del 10-15 %** [Kroetal91].

- **Al aumentar la variabilidad del tráfico este incremento de carga se hace más importante**, alcanzando valores en torno al 40-50% o incluso mayores.

- **Al hacerse grande la variabilidad del tráfico, este incremento vuelve a disminuir**, llegando a tomar valores en torno al 10% o inferiores.

- En todos los casos el incremento es mayor si partimos de longitudes de cola menores.

El comportamiento anterior puede explicarse a partir de las figuras 6.11-6.16:

- Para la situación del tráfico de Poisson (figura 6.11), y tráficos de baja variabilidad (figuras 6.12 y 6.13) se observan curvas en donde no hay un codo claro, sino un descenso suave desde el valor de probabilidad de pérdida para una carga igual a 1 hasta la asíntota vertical a menos infinito cuando la carga tiende a 0. Tenemos ganancias importantes para los casos de tráfico de baja variabilidad, que aumentan cuando la pendiente de estas curvas va disminuyendo.

- Sin embargo esta tendencia se invierte cuando la curva adquiere un aspecto de codo, que corresponde al tráfico de alta variabilidad (figuras 6.14 y 6.15).

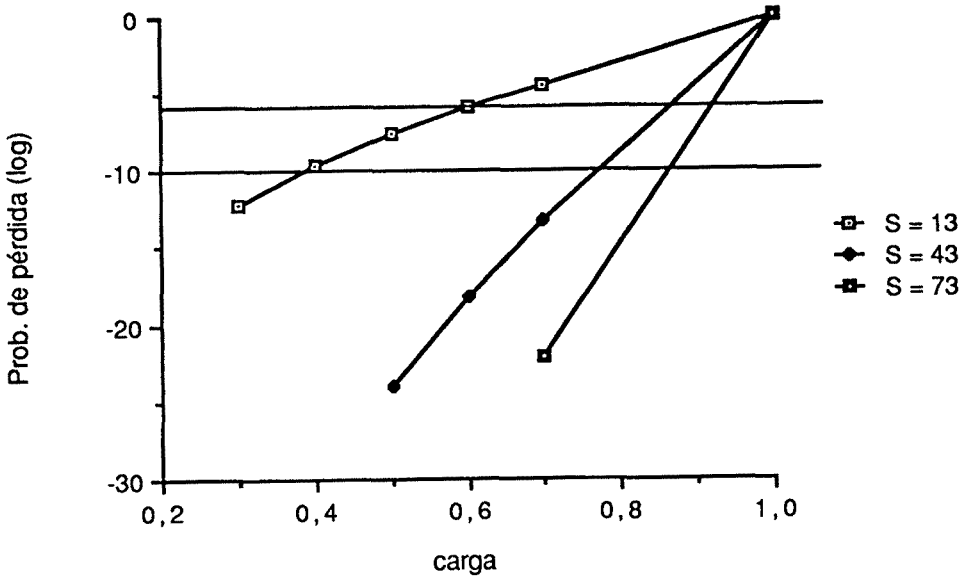


Figura 6.11 : Probabilidad de pérdida frente a la carga del multiplexor, para diferentes tamaños de cola. Tráfico de Poisson.

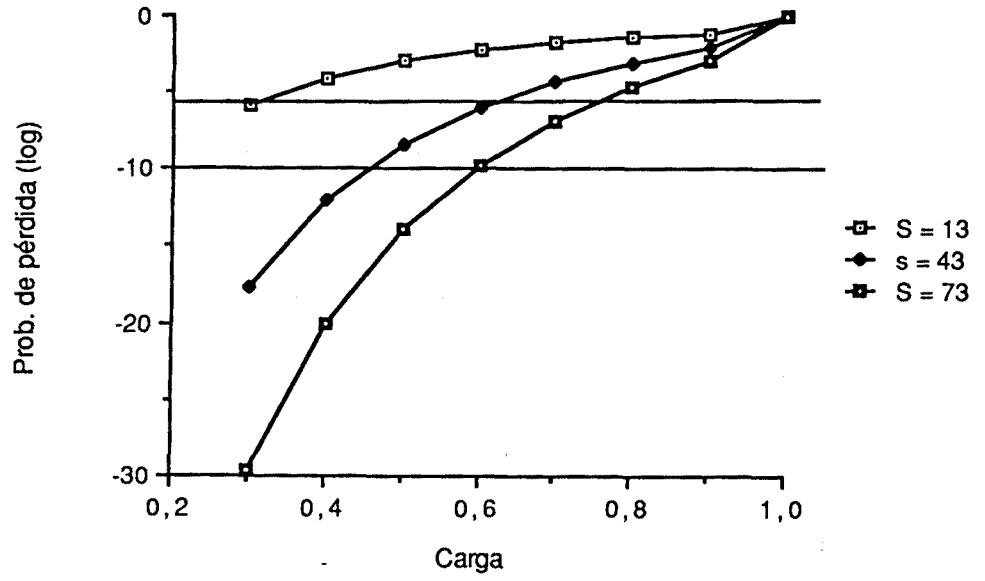


Figura 6.12 : Probabilidad de pérdida frente a la carga del multiplexor, para diferentes tamaños de cola. MMPP de dos estados, con r_1 = 1.0 E-1, r_2 = 5.0 E-1, burstiness = 3. El estado de alta actividad es el estado 2.

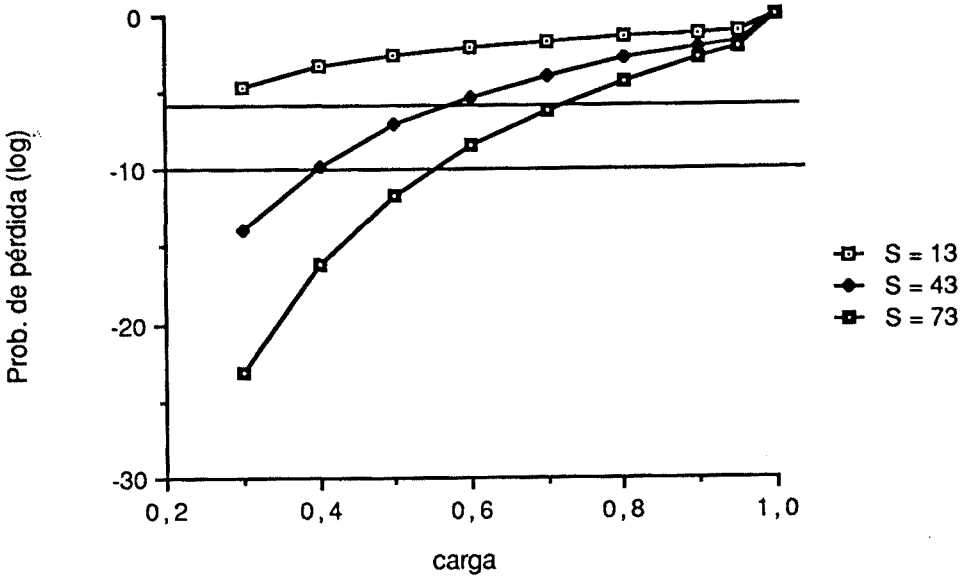


Figura 6.13 : Probabilidad de pérdida frente a la carga del multiplexor, para diferentes tamaños de cola. MMPP de dos estados, con r_1 = 1.0 E-1, r_2 = 10.0 E-1, burstiness = 5. El estado de alta actividad es el estado 2.

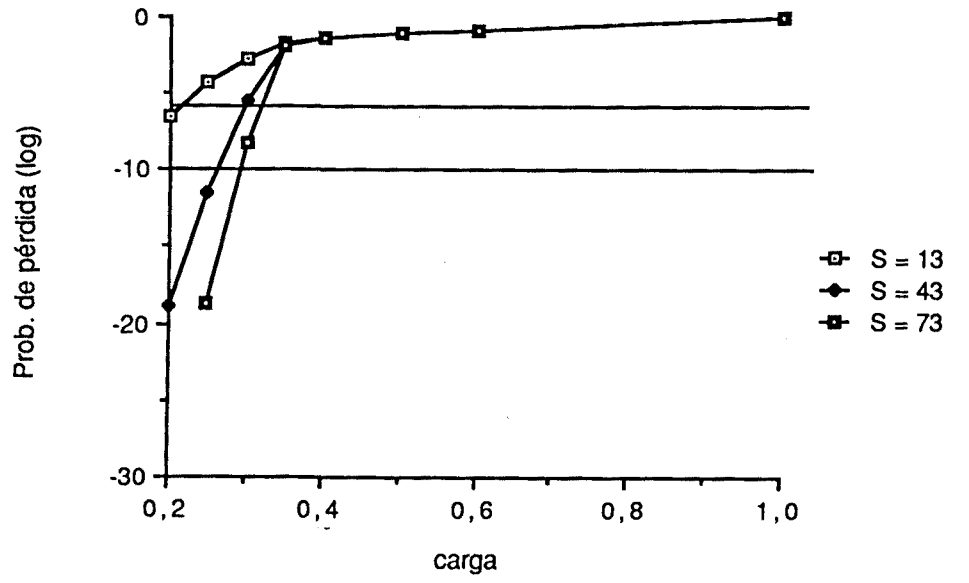


Figura 6.14 : Probabilidad de pérdida frente a la carga del multiplexor, para diferentes tamaños de cola. MMPP de dos estados, con r_1 = 1.0 E-5, r_2 = 5.0 E-5, burstiness = 3. El estado de alta actividad es el estado 2.

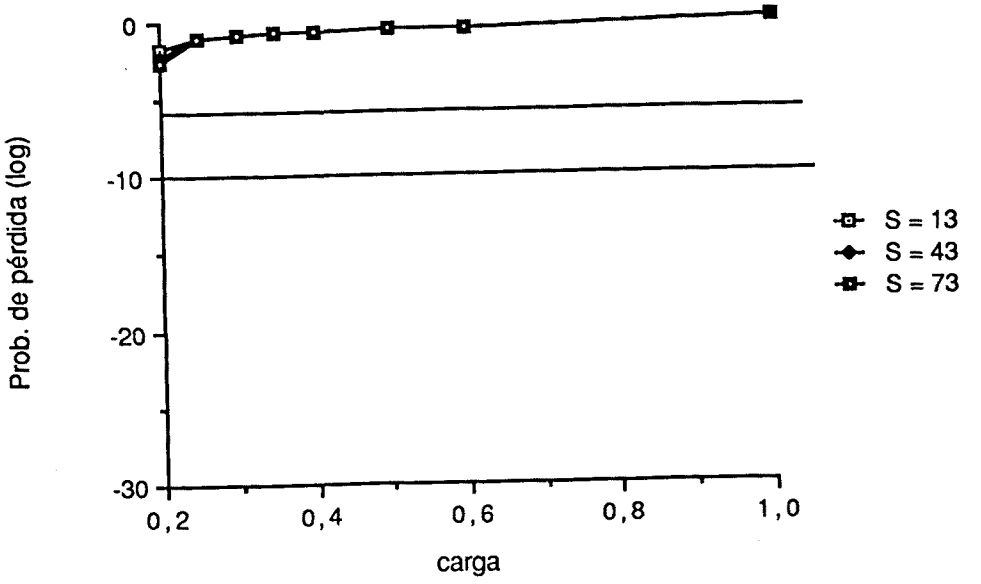


Figura 6.15 : Probabilidad de pérdida frente a la carga del multiplexor, para diferentes tamaños de cola. MMPP de dos estados, con r_1 = 1.0 E-5, r_2 = 10.0 E-5, burstiness = 5. El estado de alta actividad es el estado 2.

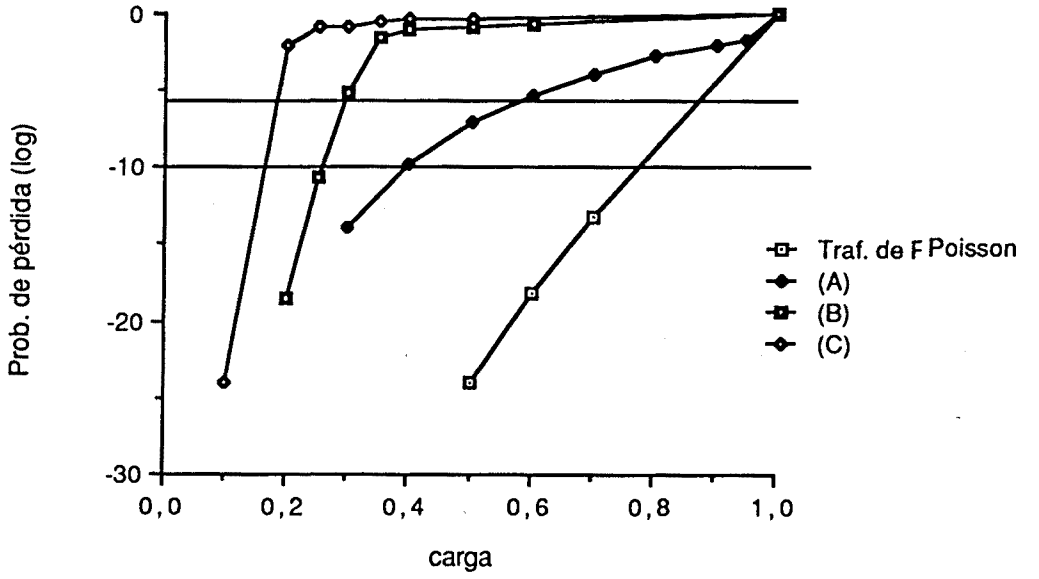


Figura 6.16 : Probabilidad de pérdida frente a la carga del multiplexor, para diferentes tipos de tráfico. La longitud de cola es S = 43.
(A): MMPP de dos estados, con r_1 = 1.0 E-1, r_2 = 10.0 E-1, burstiness = 5.
(B): MMPP de dos estados, con r_1 = 1.0 E-1, r_2 = 5.0 E-1, burstiness = 3.
(C): MMPP de dos estados, con r_1 = 1.0 E-5, r_2 = 5.0 E-5, burstiness = 5.
En todos los casos, el estado de alta actividad es el estado 2.

En la figura 6.16 se muestra, para una longitud fija de cola, la probabilidad de pérdida en función de la carga, para diferentes tipos de tráfico, observándose claramente el fenómeno señalado.

Si modelamos el tráfico como la superposición de fuentes de dos estados obtenemos conclusiones parecidas: En [Kroetal91] se dan resultados a partir de un modelo aproximado para la superposición de fuentes de dos estados, obteniendo unos incrementos de la carga admisible en torno al 30 %. En [GarCas90.b] se obtienen para un modelo basado en la aproximación de fluido resultados similares: En la figura 6.17 se muestra la ganancia en carga que se puede obtener en el caso de usar prioridades para el mecanismo PBS. Observamos que dicha ganancia es pequeña para tráficos de muy poca variabilidad, luego aumenta para volver a disminuir cuando el tráfico sufre grandes fluctuaciones.

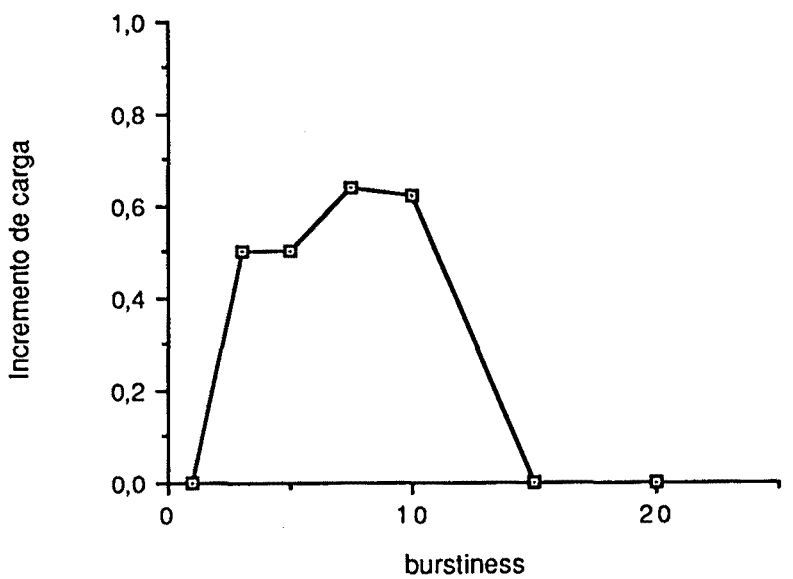


Figura 6.17 : Incremento de carga debida al uso de prioridades espaciales. El mecanismo usado es el PBS. El tráfico medio de las fuentes individuales es de 2 Mbps, y el tráfico servido por el multiplexor es de 150 Mbps. El tiempo en estado de alta actividad es de 100 mseg. El burstiness se calcula para las fuentes individuales. La proporción de tráfico prioritario sobre el total es de un 20 %.

Es de suponer que estos resultados son ciertos en general. Por ejemplo, si el tráfico de entrada es Erlang-n, se obtiene, cuando el servidor es exponencial, curvas de probabilidad de perdida semejantes cuando aumenta la variabilidad del tráfico [Klei72].

6.6 Comparación entre las políticas PO y PBS ([GarCas90.b], [GarCas91.a]).

En este apartado usaremos los resultados obtenidos en [GarCas90.b] y [GarCas91.a], para hacer una comparación entre los dos mecanismos de prioridad

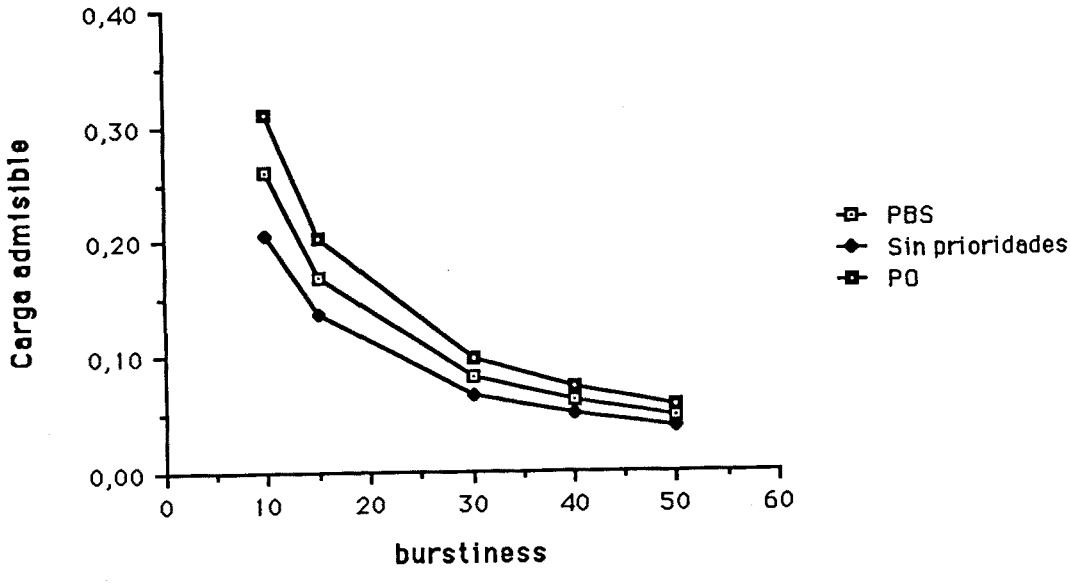


Figura 6.19 : Carga admisible en función del burstiness para diferentes políticas de gestión de la cola. El tamaño de la cola es S= 63. En el caso de la política PBS, S_2 = 57. La proporción de tráfico prioritario sobre el total es de un 20 %. r_1 = 1.0 E-2 y r_2 = 50.0 E-2. El estado de alta actividad es el estado 2.

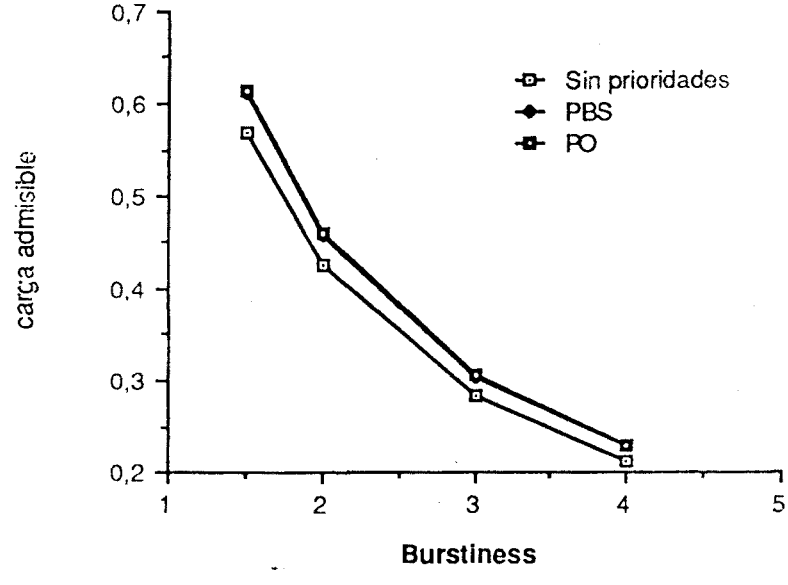


Figura 6.20 : Carga admisible en función del burstiness para diferentes políticas de gestión de la cola. El tamaño de la cola es S= 63. En el caso de la política PBS, S_2 = 57. La proporción de tráfico prioritario sobre el total es de un 20 %. r_1 = 1.0 E-5 y r_2 = 5.0 E-5. El estado de alta actividad es el estado 2.

El ahorro de espacio

La figura 6.21 muestra la carga admisible por el multiplexor frente al tamaño de la cola. De [Kroetal91] obtenemos para el caso del tráfico de Poisson los siguientes valores: . Se observa que.

- El ahorro de espacio para el mecanismo PO es superior al obtenido cuando se usa el mecanismo PBS. Sin embargo estas diferencias no son muy grandes. Los valores obtenidos están en todos los casos en torno al 40% y al 30%.

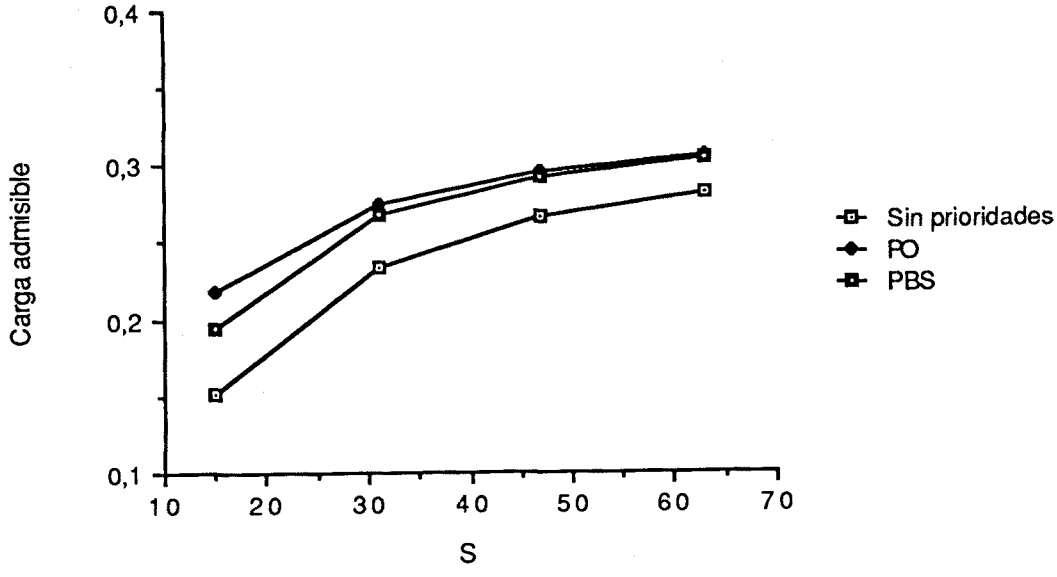


Figura 6.21 : Carga admisible en función del tamaño de cola para diferentes políticas de gestión de la cola. El burstiness es de 3. S_2 es siempre el óptimo. Para el PO, ΔS = 8. La proporción de tráfico prioritario sobre el total es de un 20 %. r_1 = 1.0 E-5 y r_2 = 5.0 E-5. El estado de alta actividad es el estado 2.

6.5 Estudio de las características más relevantes de cada mecanismo ([GarCas90.b], [GarCas91.a], [GarCas91.b] y [GarCas91.c])

El mecanismo de PO

En la definición original del mecanismo PO, S_2 era igual a 0. Sin embargo, como se ha señalado, parece que la introducción del mecanismo de reemplazamiento en una zona limitada de la cola podría reducir el coste de su realización. En la figura 6.19 se muestra la variación de las probabilidades de pérdida para las dos clases de celdas cuando se varía el umbral S_2. Se observa que cundo S_2 vale 0, la probabilidad de pérdida de las celdas prioritarias es extremadamente baja (siempre que la proporción de tráfico prioritario no sea próxima al 100 %). De ahí que para obtener un nivel de pérdidas de celdas prioritarias inferior a 1.0 E-10, el valor $\Delta S = S\text{-}S_2$ pueda tener un valor pequeño.

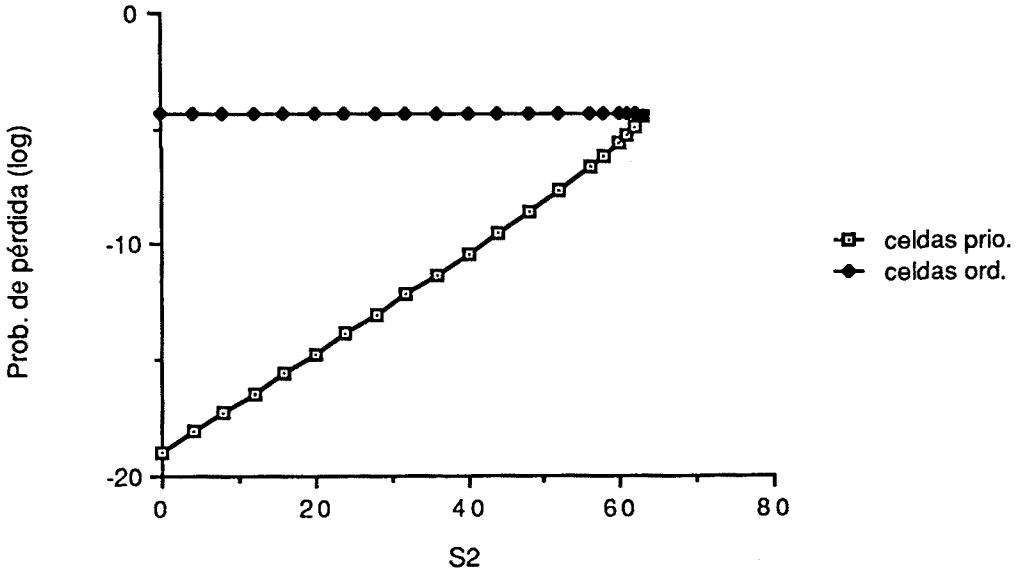


Figura 6.22 : Probabilidad de pérdida en función de S_2 para el mecanismo de PO. El burstiness es de 1.4. La proporción de tráfico prioritario sobre el total es de un 20 %. La carga es de 0.8. r_1 = 1.0 E-2 y r_2 = 3.0 E-2. El estado de alta actividad es el estado 2.

A continuación estudiaremos las características del comportamiento de estos mecanismos cuando varían los parámetros de la fuente:

- Variación de la carga:

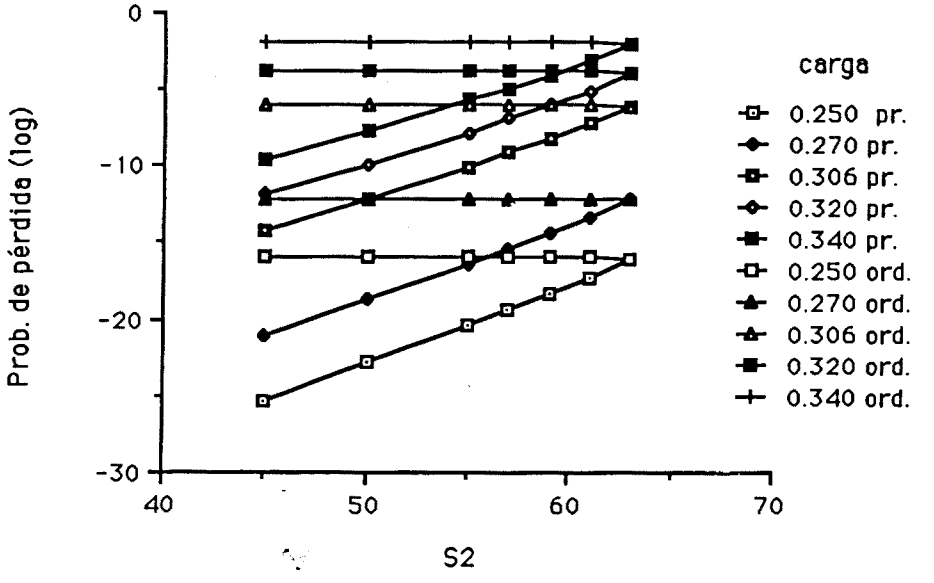


Figura 6.23 : Probabilidad de pérdida frente a S_2 para el mecanismo de PO para diferentes valores de carga. S = 63. El burstiness vale 3. r_1 = 1.0 E-5 y r_2 = 5.0 E-5. La proporción de tráfico prioritario sobre el tráfico total es de un 20 %. El estado de alta actividad es el estado 2.

En la figura 6.23 se muestran las variaciones de B_{np} y B_p frente a S_2 para diferentes valores de carga. Observamos que:

- Ambas probabilidades son fuertemente dependientes de la carga total.
- Las curvas son casi paralelas, por lo que ΔS es independiente de la carga total.

- Variación del tiempo de estancia en cada estado:

En la figura 6.24 se muestran las variaciones de B_{np} y B_p frente a S_2 para diferentes valores del parámetros r1, el inverso del tiempo de permanencia en estado de baja actividad. Se observa que:

- Para valores pequeños del tiempo de permanencia en cada estado, ambas probabilidades son muy sensibles a las variaciones.
- Las curvas son casi paralelas, por lo que ΔS es independiente de la carga total.

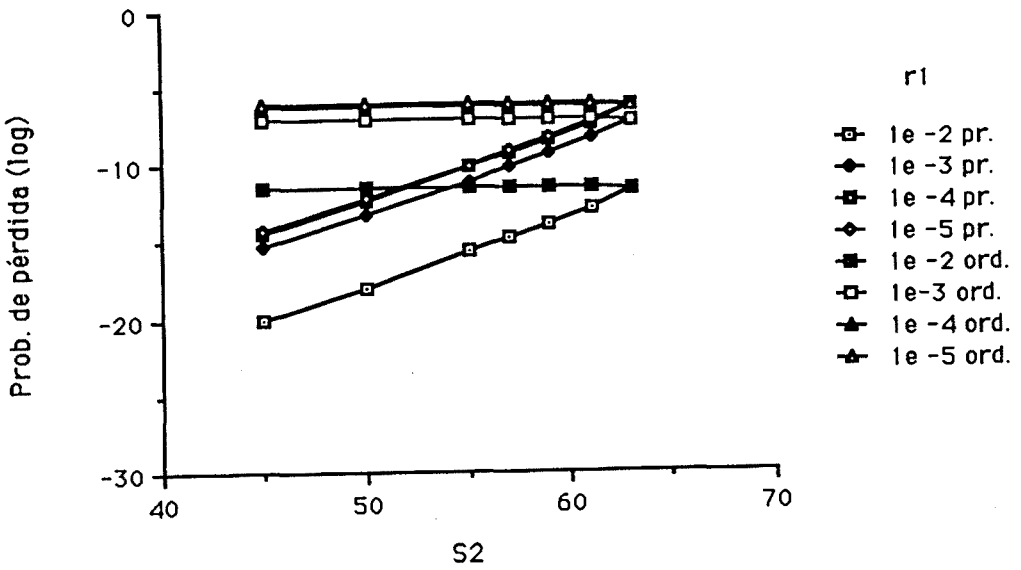


Figura 6.24 : Probabilidad de pérdida frente a S_2 para el mecanismo de PO para diferentes valores de r_1. S = 63. El burstiness vale 3. r_2 = 5 r_1. La carga es de 0.306. La proporción de tráfico prioritario sobre el tráfico total es de un 20 %. El estado de alta actividad es el estado 2.

- Variación con el burstiness:

En la figura 6.25 se muestran las variaciones de B_{np} y B_p frente a S_2 cuando varía el valor del burstiness. Se observa que:

- Ambas probabilidades son muy sensibles al valor del burstiness.
- Las curvas son casi paralelas, por lo que ΔS no depende del burstiness.

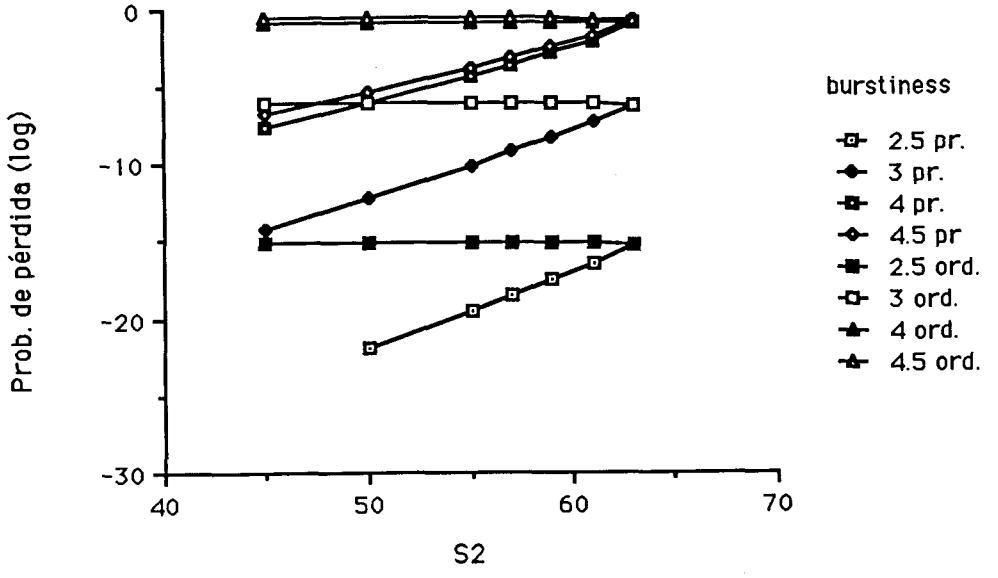


Figura 6.25 : Probabilidad de pérdida frente a S_2 para el mecanismo de PO para diferentes valores de burstiness. S = 63. r_1 = 1.0 E-5, r_2 = 5.0 E-5. La carga es de 0.306. La proporción de tráfico prioritario sobre el tráfico total es de un 20 %. El estado de alta actividad es el estado 2.

- Variación con la proporción de tráfico prioritario sobre el total.

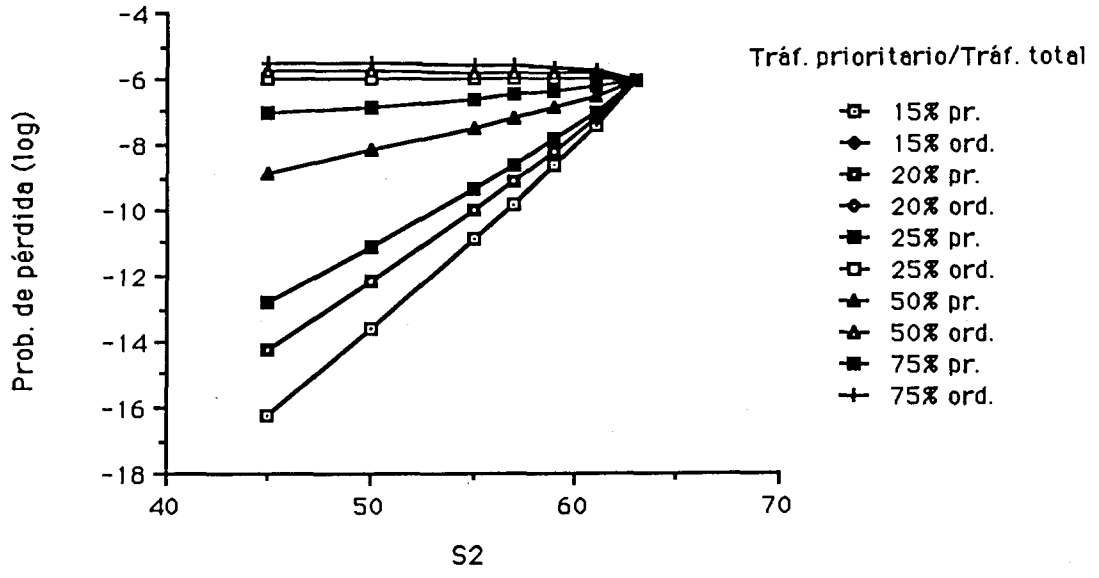


Figura 6.26 : Probabilidad de pérdida frente a S_2 para el mecanismo de PO para diferentes proporciones de tráfico prioritario respecto al total. S = 63. r_1 = 1.0 E-5, r_2 = 5.0 E-5. La carga es de 0.306. El burstiness vale 3. El estado de alta actividad es el estado 2.

En la figura 6.26 se muestran las variaciones de B_{np} y B_p frente a S_2 cuando varía

el valor de la proporción del tráfico de celdas prioritaria sobre el total. Se observa que:

- La probabilidad B_{np} es casi insensible a las variaciones de dicha proporción, salvo que tome valores próximos al 100%. La probabilidad B_p es muy sensible a variaciones de dicha proporción. Dicha sensibilidad aumenta con el incremento de ΔS. (figura 6.27)

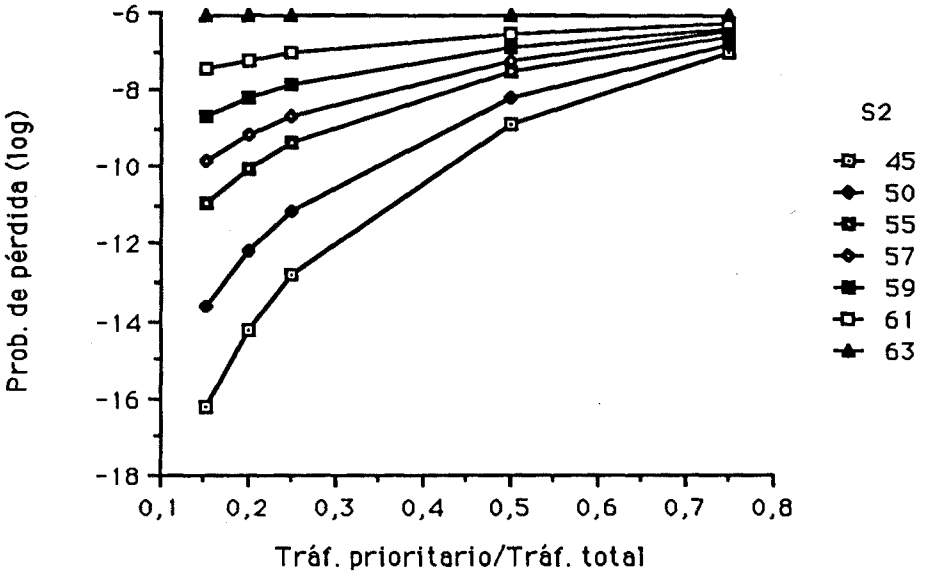


Figura 6.27 : Probabilidad de pérdida de las celdas prioritarias frente a diferentes proporciones de tráfico prioritario respecto al total para diferentes valores de S_2 para el mecanismo PO. S = 63. r_1 = 1.0 E-5, r_2 = 5.0 E-5. La carga es de 0.306. El burstiness vale 3. El estado de alta actividad es el estado 2.

De los resultados anteriores se pueden sacar las siguientes conclusiones:

- **El dimensionado de la longitud total de cola solo depende de las caracterícticas del tráfico total.**

- **El dimensionado de ΔS depende de la proporción de tráfico prioritario sobre el total, y no de las características del tráfico total.**

El mecanismo de PBS

- Variación de la carga:

En la figura 6.28 se muestran las variaciones de B_{np} y B_p frente a S_2 para diferentes valores de carga. Observamos que:

- Ambas probabilidades son fuertemente dependientes de la carga total.
- Las curvas son casi paralelas, por lo que ΔS es independiente de la carga total.

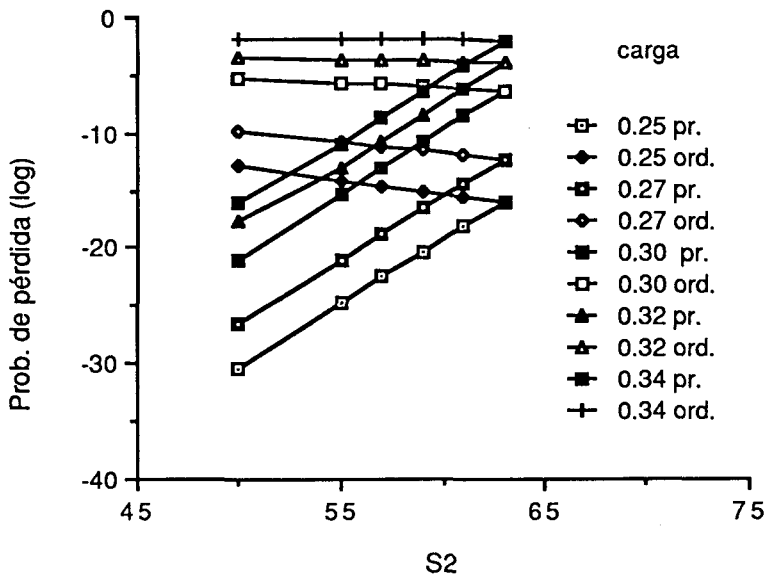


Figura 6.28 : Probabilidad de pérdida frente a S_2 para el mecanismo de PBS para diferentes valores de carga. S = 63. El burstiness vale 3. r_1 = 1.0 E-5 y r_2 = 5.0 E-5. La proporción de tráfico prioritario sobre el tráfico total es de un 20 %. El estado de alta actividad es el estado 2.

- Variación del tiempo de estancia en cada estado:

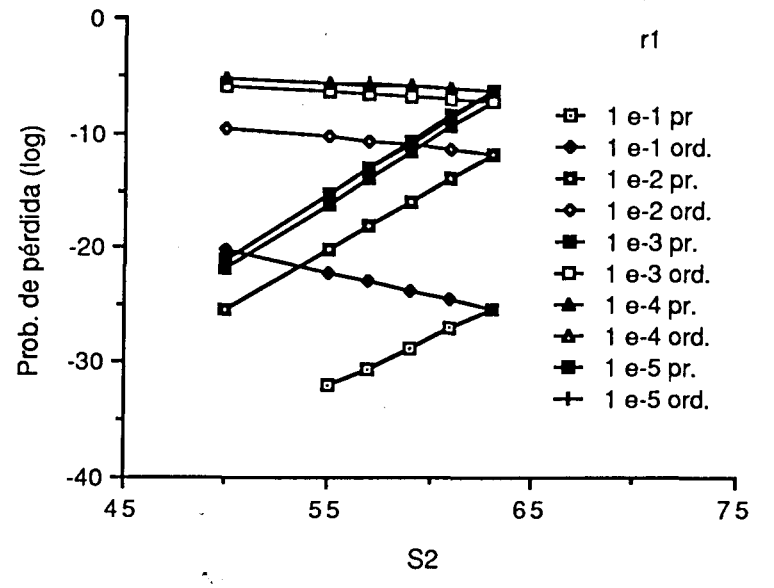


Figura 6.29 : Probabilidad de pérdida frente a S_2 para el mecanismo de PBS para diferentes valores de r_1. S = 63. El burstiness vale 3. r_2 = 5 r_1. La carga es de 0.306. La proporción de tráfico prioritario sobre el tráfico total es de un 20 %. El estado de alta actividad es el estado 2.

En la figura 6.29 se muestran las variaciones de B_{np} y B_p frente a S_2 para diferentes valores del parámetros r_1, el inverso del tiempo de permanencia en estado de baja actividad. Se observa que:

- Para valores pequeños del tiempo de permanencia en cada estado, ambas probabilidades son muy sensibles a las variaciones.
- Las curvas son casi paralelas, por lo que ΔS es independiente de la carga total.

- Variación con el burstiness:

En la figura 6.30 se muestran las variaciones de B_{np} y B_p frente a S_2 cuando varía el valor del burstiness. Se observa que:

- Ambas probabilidades son muy sensibles al valor del burstiness.
- Las curvas son casi paralelas, por lo que ΔS no depende del burstiness.

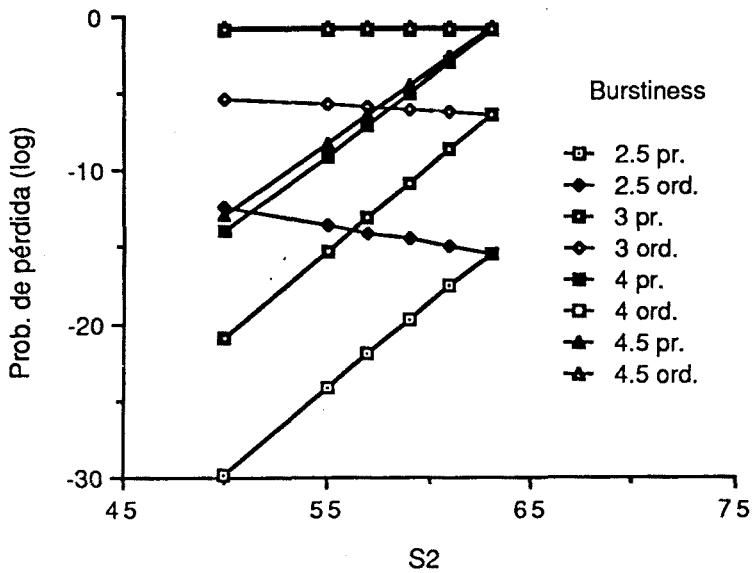


Figura 6.30 : Probabilidad de pérdida frente a S_2 para el mecanismo de PBS para diferentes valores de burstiness. S = 63. r_1 = 1.0 E-5, r_2 = 5.0 E-5. La carga es de 0.306. La proporción de tráfico prioritario sobre el tráfico total es de un 20 %. El estado de alta actividad es el estado 2.

- Variación con la proporción de tráfico prioritario sobre el total.

En la figura 6.31 se muestran las variaciones de B_{np} y B_p frente a S_2 cuando varía el valor de la proporción del tráfico de celdas prioritaria sobre el total. Se observa que:

- La probabilidad B_{np} es casi insensible a las variaciones de dicha proporción, salvo que tome valores próximos al 100%. La probabilidad B_p es muy sensible a variaciones de dicha proporción. Dicha sensibilidad aumenta con el incremento de

ΔS. (figura 6.32)

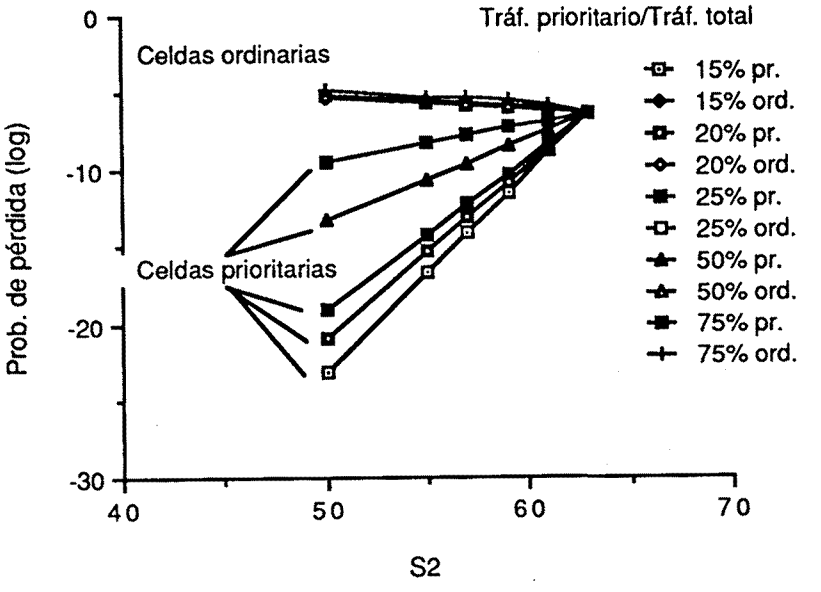


Figura 6.31 : Probabilidad de pérdida frente a S_2 para el mecanismo de PBS para diferentes proporciones de tráfico prioritario respecto al total. S = 63. r_1 = 1.0 E-5, r_2 = 5.0 E-5. La carga es de 0.306. El burstiness vale 3. El estado de alta actividad es el estado 2.

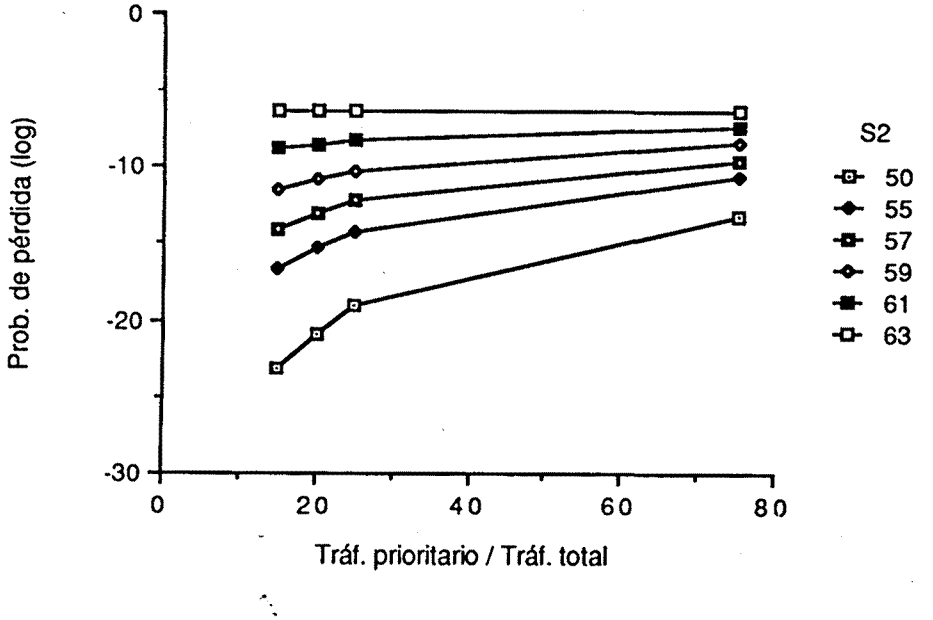


Figura 6.32 : Probabilidad de pérdida de las celdas prioritarias frente a diferentes proporciones de tráfico prioritario respecto al total para diferentes valores de S_2 para el mecanismo PBS. S = 63. r_1 = 1.0 E-5, r_2 = 5.0 E-5. La carga es de 0.306. El burstiness vale 3. El estado de alta actividad es el estado 2.

- Influencia de la correlación entre el tráfico ordinario y prioritario:

En todos nuestros resultados hemos supuesto una correlación positiva entre los dos tipos de celdas. Es decir, los estados de alta actividad para ambos tipos de celdas coincidian, al igual que los tiempos en cada estado, etc. En la figura 6.33 se muestran las variaciones de B_{np} y B_p frente a S_2 cuando tenemos una fuente correlada (caso que ha sido tratado hasta ahora) y cuando se considera que el tráfico ordinario y el prioritario son independientes entre sí. En este último caso obtenemos un MMPP de cuatro estados. Se observa que:

- Las probabilidades de pérdida son ligeramente inferiores para el caso de tráficos independientes, como era de esperar. **Sin embargo las diferencias no son importantes y el comportamiento general es el mismo.**

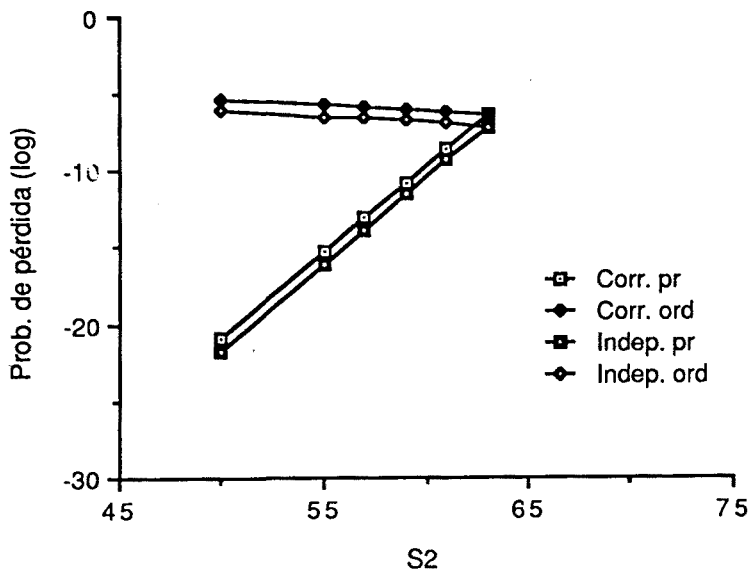


Figura 6.33 : Probabilidad de pérdida frente a S_2 para el mecanismo de PBS cuando el tráfico de celdas prioritarias y ordinarias es totalmente correlado y cuando es independiente. S = 63. La carga es de 0.306. La proporción de tráfico prioritario sobre el tráfico total es de un 20 %.

- Conclusiones

De los resultados anteriores se pueden sacar las siguientes conclusiones:

- El dimensionado de la longitud total de cola solo depende de las caracterícticas del tráfico total.

- El dimensionado de ΔS depende de la proporción de tráfico prioritario sobre el total, y no de las características del tráfico total.

6.8 Conclusiones finales.

De los resultados presentados anteriormente se puede concluir:

- La introducción de mecanismos de prioridad espacial supone un ahorro de espacio de un 40% para el mecanismo PO, y de un 30% para el PBS. Parece que estos ahorros son independientes de las características de la fuente.

- La introducción de mecanismos de prioridad espacial supone un incremento en la carga admisible que es fuertemente dependiente de las características del tráfico:

  - Para tráfico de Poisson, dicha ganancia está en torno al 10%.
  - Cuando el tráfico se hace más variable aumenta hasta el 50% o más.
  - Para tráfico muy variable vuelve a descender hasta el 10% o menos.

- Las diferencias entre los mecanismos PO y PBS parecen pequeñas y son dependientes del tráfico de entrada. El coste del mecanismo PO parece mayor que el del mecanismo PBS. Sin embargo este coste puede reducirse considerablemente con la introducción de un umbral.

- Para ambas políticas, el dimensionamiento de la longitud de cola total es dependiente solo de las características del tráfico total, siempre que la proporción del tráfico prioritario sobre el total no sea próxima al 100 %. El valor ΔS depende fuertemente de dicha proporción, pero no de las características del tráfico total.

Sin embargo, de estas conclusiones no puede esperarse dar la respuesta sobre si es o no conveniente la introducción de un segundo servicio portador en la RDSI-BA, pues las cuestiones tecnológicas sobre el incremento de coste que ello supondría no han sido abordadas. Además, debe tenerse en cuenta que las características del tráfico real no son aun conocidas con exactitud y los resultados parecen muy dependientes del tipo de tráfico.

Conclusiones

En este trabajo se han presentado varios modelos analíticos, exactos y aproximados, que pueden ser usados para la evaluación de ciertos mecanismos de control de congestión en redes ATM.

Las especificaciones de una red ATM hacen del control de la congestión un problema muy difícil de resolver: debemos permitir que casi cualquier tipo de tráfico pueda tener acceso a la red, pero somos muy exigentes en los valores que permitimos a ciertos parámetros que definen el grado de servicio de la red, tales como el retardo y la probabilidad de pérdida de una celda.

En las propuestas iniciales para el control de congestión en una red ATM los mecanismos que se usaban eran extremadamente simples: El control de la congestión se basaba en un CAC ('Conexion Acceptance Control') común a todos los usuarios. En el momento de establecer una conexión, cada usuario debía dar a la red una descripción de las características del tráfico que iba a emitir. Esta descripción debía ser sencilla, pero lo suficientemente completa como para que el mecanismo de CAC pudiera predecir, a partir de ella, el nivel de congestión que iba a causar ese nuevo usuario. Estos parámetros que describían el tráfico debían ser fácilmente controlables por un mecanismo de función de policía. Los mecanismos de función de policía propuestos tenían una mínima influencia sobre el tráfico emitido, actuando sólo cuando existía una violación del contrato.

En cuanto a las exigencias de grado de servicio de los diferentes usuarios se proveía de un único servicio portador, es decir, un único tipo de celdas, que debía acomodarse a las exigencias de los servicios más restrictivos.

Sin embargo parece que las anteriores propuestas eran demasiado optimista: No podemos usar mecanismos de control de congestión tan sencillos, a no ser que hagamos un uso muy ineficiente de los recursos de la red. Parece más realista hacer una clasificación de las fuentes, de forma que cada clase de fuente recibe un tratamiento diferente. La función de policía debe conformar de forma activa los parámetros del tráfico emitido. Debemos distinguir entre los parámetros del grado de servicio para diferentes tipos de servicios.

El precio a pagar por este mejor control de la congestión es que aumentamos la complejidad de la red. La decisión final sobre si estos nuevos mecanismos de control propuestos son o no adecuados vendrá de una comparación entre los beneficios que se derivan de su uso y el coste que supondría su introducción. Con nuestro trabajo se ha pretendido contribuir a dar respuesta a este problema.

En lo que concierne al estudio de mecanismos de control de policía se demuestra la ineficiencia de los mecanismos de 'pick-up' para controlar la congestión, dando un ejemplo de cómo un usuario puede burlar la vigilancia de un mecanismo de función de policía de 'pick-up', dañando considerablemente la calidad de servicio vista por el conjunto de usuarios de la red. En los trabajos anteriores no se disponía de un modelos analítico adecuado para estudiar este fenómeno, por lo que hemos usado un nuevo modelo desarrollado en el capítulo 2.

También se lleva a cabo un estudio del problema de la introducción de mecanismos de prioridad en una red ATM: Se evalúan los beneficios, tanto en el aumento de carga de la red como en el ahorro en memoria de alta velocidad en los conmutadores, derivados de la introducción de dos servicios portadores con diferentes niveles de probabilidad de pérdida máxima. Después se hace un estudio comparativo entre dos mecanismos que han sido propuestos para la introducción de prioridades espaciales, el mecanismo de 'Push-out' y el de 'Partial buffer sharing', contemplando el uso de un umbral en el mecanismo de 'Push-Out'

que reduce la complejidad de su realización. Los modelos que se había usado con anterioridad no permitían la evaluación de estos mecanismos en condiciones realistas de tráfico, por lo que se han desarrollado varios modelos exactos, basados en técnicas analíticas matriciales y en la aproximación de fluido.

Otro problema importante que aparece en el estudio de las redes ATM es el de su evaluación. El tráfico presente en este tipo de redes tiene características que complican los modelos que debemos usar. Además, los valores extremadamente bajos de los parámetros a medir dificultan el uso de simulaciones.

El tráfico puede modelarse mediante un proceso markoviano: Se considera que dicho tráfico varía entre diferentes estados de actividad que pueden ser vistos como los estados de una cadena de Markov. Dentro de cada nivel de actividad debemos buscar un modelo de la forma de emisión, con parámetros dependientes del estado.

Numerosos estudios muestran que estos modelos permiten predecir de forma bastante aproximada los valores reales de los parámetros a evaluar. Sin embargo, cuando queremos dar valores realistas a los modelos de tráfico, el número de estados de la cadena de Markov que describe el sistema se hace enormemente grande, dificultando la resolución exacta del modelo

Hemos intentado contribuir a la solución de este problema desarrollando una aproximación asintótica al estudio de la multiplexación de fuentes descritas mediante un D-BMAP, que nos permite incluir fuentes de diferentes características (por ejemplo, fuentes que siguen una secuencia periódica de emisión junto con fuentes VBR) obteniendo resultados extremadamente exactos con poco esfuerzo computacional. Este modelo ha permitido evaluar la eficiencia de un mecanismo de función de policía en condiciones que no podrían ser abordadas con un modelo exacto y puede ser usado, por ejemplo, para estudiar la superposición de fuentes de video con diferentes niveles de actividad, o para estudiar el 'jitter' introducido durante el proceso de multiplexación de una fuente periódica con tráfico VBR.

Lineas abiertas

Quedan aún numerosos aspectos que deben ser estudiados dentro del área del control de la congestión y de las técnicas de evaluación. En el ámbito del control de la congestión tenemos:

- Debemos hacer una estimación del incremento de carga que permiten obtener los protocolos de acceso a la red basados en el FRP ('Fast Reservation Protocol') frente a la complejidad que supone su introducción.

- Debemos evaluar la eficiencia en el control de la congestión de los mecanismos de función de policía en donde se conforma el tráfico emitido ('Traffic saphing').

- El modelo asintótico de la probabilidad de pérdida puede ser aplicado al estudio de otros problemas: 'jitter' que introduce la multiplexación de una fuente determinista con fuentes VBR, multiplexación de fuentes de video con diferentes niveles de actividad, etc.

En cuanto a las técnicas de evaluación de dispositivos ATM también quedan interesantes cuestiones pendientes de ser resueltas:

- Cuando estamos tratando con sistemas con una sola cola, se deben desarrollar métodos computacionales sencillos que permitan tratar problemas de mayor dimensión a un bajo coste.

- Es de especial importancia tener una caracterización del proceso de salida de un multiplexor. Permitiría, por ejemplo, estudiar la eficiencia de los mecanismos de 'Traffic saphing'.

- Además poco se ha hecho para abordar el importante problema que surge al considerar una red de colas en donde el tráfico de entrada a cada nodo es un proceso con fuertes correlaciones, tal como se espera que sea el caso de una red ATM.

Referencias

[Anietal82] D. Anick, D. Mitra y M. M. Sondhi, "Stochastic Theory of a Data-Handling System with Multiple Sources". Bell System Tech. J., Vol 61, 1982.

[Bai91] A. Baiocchi, "Asymptotic Behaviour of the Loss Probability of the MM/G/1 Queue". Enviado para su publicación. 1991.

[Baietal91] A. Baiocchi, N. Blefari y A. Roveri, "Buffer Dimensioning Criteria for an ATM Muktiplexer Loaded with Homogeneous On-Off Sources". Queueing, Performance and Control in ATM, (ITC-13). Ed. Elsevier Science Publishers, 1991.

[Baletal90] R. Balcer, J. Eaves, J. Legras, R. McLintock y T. Wright, "An Overview of Emerging CCITT Recommendations for the Synchronous Digital Hierarchy: Multiplexers, Line Systems, Management, and Network Aspects". IEEE Comm. Magazine, Agosto 1990.

[Bel72] R. Bellman, "Introduction to Matrix Analysis". McGraw-Hill, New York, 1972.

[Blo89] C. Blondia, "The N/G/1 Finite Capacity Queue". Stochastic Models, Vol. 5, 1989.

[Blo90] C. Blondia, "A Discrete-Time Batch Markovian Arrival Process". RACE Document, PRLB_123_0028_CD_CC, Diciembre 1990.

[Blo91] C. Blondia, "Analytical Models for VBR Video Sources". RACE document, PRLB_123_0032_CD_CC, Abril 1991.

[Bloetal91] C. Blondia, U. Briem y O. Casals, "Analytical Source Models". RACE Document PRLB_123_0029_CD_CC / UST_123_0027_CD_CC / UPC_123_0029_CD_CC., Enero 1991.

[BloThe89] C. Blondia y T. Theimer, "A Discrete-Time Model for ATM Traffic". RACE Document, PRLB_123_0018_CD_CC / UST_123_002_CD_CC, Octubre 1989.

[Boy88] P. Boyer, "Priorities in an ATM Network". RACE Document, CNET-123-035-CD-CC. Noviembre 1988.

[Boy90] P. Boyer, "Definition d'un Contrôle de Congestion pour un Réseau Temporel Asynchrone". CNET, Note technique LAA/RSM/165, Abril 1990.

[BoyGui90] P. Boyer, F. Guillemin, "ATM-Based Network Congestion". RACE document, CNET_123_08_037_CD_CC, 1990.

[BoyTra91] P. Boyer, D. Tranchier, "Specification of the FRP/DT". RACE ATM Network Planning and Evolution workshop, Londres, Abril 1991.

[BroSim88] P. Brown y A. Simonian, "Perturbation of a Periodic Flow in a Synchronous Server". PERFORMANCE'87, P.-J. Courtois y G. Latouche Ed. Elsevier Science Publishers, 1988.

[CCI_I121] CCITT RECOMMENDATION I121, "On broadband aspects of ISDN". Study Group XVIII, Ginebra, Junio 1988.

[CCI_I.361] CCITT DRAFT RECOMMENDATION I.361, "ATM Layer Specification for B-ISDN". Study Group XVIII, Ginebra, Enero 1990.

[Cin75] E. Cinlar, "Introduction to Stochastic Processes", Ed. Englewood Cliffs, New Jersey, 1975.

[Coo72] R. B. Cooper, "Introduction to Queueing Theory". Ed. MacMillan, New York/London, 1972.

[DaiLan86] J. N. Daigle y J. D. Langford, "Models for Analysis of Packet Voice Communications Systems". IEEE J. Select. Areas Commun., Vol. SAC-4, No. 6, Septiembre 1986.

[DosHef86] B. T. Doshi y H. Heffes, "Overload Performance of Several Processor Queueing Disciplines for the M/M/1 Queue". IEEE Trans. Commun., Vol. COM-34, No. 6, Junio 1986.

[Eck79] A. E. Eckberg, "The Single Server Queue with Periodic Arrival process and Deterministic Service Times". IEEE Trans. Commun., Vol. COM-27, No. 3, 1979.

[Eckatal89.a] A. E. Eckberg, D. T. Luan y D. M. Lucantoni, "Bandwidth Management: A Congestion Control Strategy for Broadband Packet Networks - Characterizing the Throughput-Burstiness Filter". ITC Specialist Seminar, Adelaide, Septiembre 1989.

[Ecketal89] A. E. Eckberg, D. T. Luan, D. M. Lucantoni, "Meeting the Challenge: Congestion and Flow Control Strategies for Broadband Information transport". IEEE GLOBECOM'89, Dallas, Noviembre 1989.

[Fil89] J. Filipiak, "Structured Systems Analysis Methodology for Design of an ATM Network Architecture". IEEE J. Select. Areas Commun., Vol. SAC-7, No. 8, Octubre 1989.

[GarCas90.a] J. García y O. Casals, "Priorities in ATM Networks". High-Capacity Local and Metropolitan Area Networks, Ed. G. Pujolle. Springer-Verlag, Berlin, 1991.

[GarCas90.b] J. García y O. Casals, "Stochastic Models of Space Priority Mechanisms with Markovian Arrival Processes". Aceptado para su publicación en Annals of Operations Research.

[GarCas91.a] J. García y O. Casals, "Space Priority Mechanisms with Bursty Traffic". Proceedings of the International Conference on the performance of Distributed Systems and Integrated

Communications Networks, Kyoto, Septiembre 1991.

[GarCas91.b] J. García y O. Casals, "Performance Evaluation of Source Dependent Congestion Control Procedures in ATM Networks". SICON'91. Singapur, Septiembre 1991.

[GarCas91.c] J. García y O. Casals, "Statistical Multiplexing Gain Using Space Priority Mechanisms". IEEE GLOBECOM'91. Phoenix, AZ, Diciembre 1991.

[Gha89] M. Ghambari, "Two-Layer Coding of Video Signals for VBR Networks". IEEE J. Select. Areas Commun., Vol. SAC-7, No. 5, Junio 1989.

[Graetal85] W. K. Grassman, M. Taksar y D. Heyman, "Regenerative Analysis and Steady State Distributions for Markov Chains". Operations Research, Vol. 33, 1985.

[Graetal90] A. Gravey, J.-R. Louvion y P. Boyer, 'On the Geo/D/1 and Geo/D/1/n Queues". Performance Evaluation, Vol. 11, 1990.

[GolVan88] G.H. Golub, C. F. Van Loan, "Matrix Computations", Ed. Johns Hopkins University Press, Baltimore, 1988.

[HebGra89] G. Hebuterne y A. Gravey, "A Space Priority Queueing Mechanism for Multiplexing ATM Channels". ITC Specialist Seminar, Adelaide, Septiembre 1989.

[HefLuc86] H. Heffes, D. M. Lucantoni, "A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer performance". IEEE J. Select. Areas in Commun., Vol. SAC-4, No. 6, Septiembre 1986.

[Hui88] J. Y. Hui, "Resource Allocation for Broadband Networks". IEEE J. Select. Areas in Commun., Vol. SAC-6, No. 9, 1988.

[Hui90] J. Y. Hui, "Switching and Traffic Theory for Integrated Broadband Networks". Ed. Kluwer Academic Publishers, Boston, 1990.

[Irl78] M. I. Irland, "Buffer Management in a Packet Switch". IEEE Trans. Commun., Vol. COM-26, No. 3, Marzo 1978.

[Ide88] I. Ide, "Superposition of interrupted Poisson processes and its aplications to packetized voice multiplexers", 12th ITC, Torino, 1988.

[JayChr81] N. S. Jayant y S. W. Christensen, "Effect of packet losses in waveforms coded speech and improvement due to an add-even sample interpolation procedure". IEEE Trans. Commun. Vol. COM-29, No. 2, Febrero 1981.

[KawSai90] K. Kawashima y H. Saito, "Teletraffic Issues in ATM Networks". Computer Networks and ISDN Systems, Vol. 20. 1990.

[Kle75] L. Kleinrock, "Queueing Systems". Vol. I y II, Wiley Interscience, New York, 1975.

[Kra87] M. Kramer, "Computational Methods for Markov Chains Ocurring in the Queueing Theory". Messung, Modellierung und Bewertung

von Rechensystemen, eds. U. Herzog y M. Paretok, Informatik-Fachberichte 154, Springer, 1987.

[Kro90] H. Kröner, "Comparative Performance Study of Space Priority Mechanisms for ATM Networks". IEEE INFOCOM'90, San Francisco, Junio 1990.

[Kro91] H. Kröner, "Statistical Multiplexing of Sporadic Sources - Exact and Approximate Performance Analysis". Preprint, 1991

[Kroetal90] H. Kröner, T. H. Theimer, U. Briem, "Queueing Models for ATM systems - A Comparison". Proceedings of the 7th. ITC Seminar, Morristown, Octubre 1990.

[Kroetal91] H. Kröner, G. Hebuterne, P. Boyer y A. Gravey, "Priority Mangement in ATM Switching Nodes". IEEE J. Select. Areas Commun, Vol. SAC-9, No. 3, Abril 1991.

[Lel89] W. E. Leland, "Window-based Congestión Management in Broadband ATM Networks: The Performance of Three Acces-Control Policies". IEEE GLOBECOM'89, Dallas, Noviembre 1989.

[Li89.a] S.-Q. Li, "Study of Information Loss in Packet Voice Systems". IEEE Trans. Commun., Vol. COM-37, No. 11, Noviembre 1989.

[Li89.b] S.-Q. Li, "Overload Control in a Finite Message Storage Buffer". IEEE Trans. Commun. Vol. COM-37, No. 12, Diciembre 1989.

[Li90] S. Q. Li, "A General Solution Technique for Discrete Queueing Analysis of Multimedia Traffic on ATM". Proceedings of the IEEE GLOBECOM'90, San Diego, CA.,1990.

[LiMar88] S. Q. Li, S. W. Mark, "Traffic Characterization for Integrated Services". IEEE INFOCOM'88, New Orleans, Marzo 1988.

[Lou91] J.-R. Louvion, "Rate-Based Multiplexing of Periodic On-Off Sources in ATM Networks". RACE ATM Network Planning and Evolution workshop, Londres, Abril 1991.

[Louetal90] J.-R. Louvion, J. Boyer y J.-B. Gravereaux, "Statistical Multiplexing of VBR Sources in ATM Networks". 3rd. IEEE CAMAD, Torino, September 1990.

[Luc91] D. M. Lucantoni, "New Results on the Single Server Queue with Batch Markovian Arrival Process". Stochastic Models, Vol. 7, 1991.

[Lucetal90] D. M. Lucantoni, K. S. Meier-Hellstern y M. F. Neuts, "A Single Server Queue of Non-Renewal Arrival Processes". Adv. Appl. Prob., Vol. 22, 1990.

[LucRam85] D. M. Lucantoni y V. Ramaswami, "Efficient Algorithms for Solving the Non-Linear Matrix Equations Arising in Phase Type Queues". Stochastic Models, Vol 1. 1985.

[MacSau85] E. A. MacNair, C. H. Sauer, "Elements of Practical Performance Modeling". Ed. Prentice Hall, New Jersey, 1985.

[Magetal88] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson y J. D. Robbins, "Performance Models of Statistical Multiplexing in Packet Video

Communications". IEEE Trans. Commun., Vol. COM-36, No. 7, Julio 1988.

[Min89] S. E. Minzer, "Broadband ISDN and Asynchronous Transfer Mode (ATM)". IEEE Comm. Magazine, Septiembre 1989.

[Mit88.a] D. Mitra, "Stochastic Theory of a Fluid Model of producers and Consumers Coupled by a Buffer". Adv. Appl. Prob., Vol. 20, 1988.

[Mit88.b] D. Mitra, "Stochastic Fluid Models". PERFORMANCE'87, P. J. Courtois y G. Latouche Ed. Elsevier Science Publishers, 1988.

[Musetal85] H. G. Musmann, P. Pirsch y H.-J. Grallet, "Advances in Picture Coding". Proceedings of the IEEE. Vol. 73, No. 4, Abril 1985.

[Naretal91] R. Narajan, J. F. Kurose y D. Towsley, "Approximation Techniques for Computing Packet Loss in Finite-Buffered Voice Multiplexers". IEEE J. Select. Areas Commun. Vol. SAC-9, N0. 3, Abril 1991.

[NetLim80] A. N. Netravali y J. O. Limb, "Picture Coding: A Review". Proceedings of the IEEE, Vol. 68, N0. 3, Marzo 1980.

[Neu79] M. F. Neuts, "A Versatile Markovian Point Process". J. Appl. Prob., Vol. 16, 1979.

[Neu81] M. F. Neuts, "Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach". The Johns Hopkins University Press, Baltimore, 1981

[Neu86] M. F. Neuts, "The Caudal Characteristic Curve of Queues". Adv. Appl. Prob., Vol. 18. 1986.

[Neu89] M. F. Neuts, "Structured Stochastic Matrices of M/G/1 Type and Their Applications". Marcel Dekker, New York, 1989.

[Niletal90] A. A. Nilsson, F. Lai y H. G. Perros, "A Queueing Model of a Bufferless Synchronous Clos ATM Switch with Head-Of-Line Priority and Push-Out". Research Report., Comp. Sci. Dept. North Carolina State Univ., 1990.

[PerFle87] S. D. Personick y W. O. Fleckenstein, "Communications Switching - From Operators to Photonics". Proceedings of the IEEE, Vol. 75, No. 10, Octubre 1986.

[Ram80] V. Ramaswami, "The N/G/1 Queue and its Detailed Analysis". Adv. Appl. Prob., Vol. 12, 1980.

[RAC90] "The relationship of the Telecommunications Management Network to Connection Acceptance Control and Source Policing", RACE Project 1022, 1990.

[RasSor89] C. Rasmussen, J. Sorensen, "A Simple Call Acceptance Algorithm for an ATM Network". ITC Specialist Seminar, Adelaide, Septiembre 1989.

[RatThe90] E. P. Rathgeb, T. H. Theimer, "The Policing Function in ATM Networks". XIII International Switching Symposium, Estocolmo, Junio 1990.

[Rigetal90] G. Rigolio, P. Vaccari y L. Verri, "Use of priority in ATM Networks: Efficiency Evaluations". RACE Document, 1990.

[Rob89] J. W. Roberts, "Jitter Due to an ATM Multiplex - Application to Peak Rate Policing". COST 224, Agosto 1989.

[Rob91] J. W. Roberts, "Performance Evaluation and Design of Multiservice Networks". COST 224, Final Report. Paris, 1991.

[RobVir91] J. W. Roberts y J. T. Virtamo, "The Superposition of Periodic Cell Arrival Stream in an ATM Multiplexer". IEEE Trans. Commun., Vol. COM-39, No. 2, Febrero 1991.

[Sanetal90] G. Santos, J. Solé y O. Casals, "Comunicaciones en banda ancha, gestión de red". Mundo Electrónico, Septiembre 1990.

[SauCha81] C. H. Sauer, K. M. Chandi, "Computer Systems Performance Modeling". Ed. Prentice-Hall, New Jersey, 1981.

[Senetal89] P. Sen, B. Maglaris, N.-E. Rikli y D. Anastassiou, "Models for Packet Switching of Variable-Bit-Rate Video Sources". IEEE J. Select. Areas Commun., Vol. SAC-7, No. 5, Junio 1989.

[Sol91] J. Solé, "Estudi i proposta d'esquemes d'avaluació per a dispositius ATM". Tesis Doctoral, Barcelona, 1991.

[SriWhi86] K. Sriram y W. Whitt, "Characterizing Superposition Arrival processes in Packet Multiplexers for Voice and Data". IEEE J. Select. Areas Commun., Vol. SAC-4, No. 6, Septiembre 1986.

[Srietal91] K. Sriram, R. S. McKinney y M. H. Sherif, "Voice Packetization and Compression in Broadband ATM Networks". IEEE J. Select. Areas Commun. Vol. SAC-9, N0. 3, Abril 1991.

[SteElW91] T. E. Stern y A. I. ElWalid, "Analysis of Separable Markov-Modulated Rate Models for Information-Handling Systems". Adv. Appl. Prob., Vol. 23, 1991.

[SumOza89] S. Sumita y T. Ozawa, "Achievability of Performance Objectives in ATM Switching Nodes". Proceedings of the International Seminar on Performance of Distributed and Parallel Systems, Kyoto, Diciembre 1988.

[Tan88] A. Tannembaum, "Computer Networks". Ed. Prentice Hall, 1988.

[Tob90] F. Tobagi, "Fast Packet Switch Architectures for Broadband Integrated Services Digital Networks". Proceedings of the IEEE, Vol. 79, No. 1, Enero 1990.

[Toretal90] N. Torralba, J. Domingo y J. García, "Comunicaciones en banda ancha, arquitectura de red y acceso de usuario". Mundo Electrónico, Mayo 1990.

[Tra89] D. Tranchier, "Etude de la gigue dans les réseaux ATM: Temps de traversée des files d'attente". CNET, Note Technique NT/LAA/ RSM/163, Diciembre 1989.

[Tra91] D. Tranchier, "The Fast Reservation Protocol / DT". RACE ATM Network Planning and Evolution Workshop, Londres, Abril 1991.

[Tuc88] R. C. Tucker, "Accurate Method for Analysis of a Packet-Speech Multiplexer with Limited Delay". IEEE Trans. Commun. VOL 36, No 4, Abril 1988.

[Tur86] J. S. Turner, "New Directions in Communications (or Which Way to the Information Age)". IEEE Comm. Magazine, Vol. 24, No. 10, Octubre 1986.

[Veretal88] W. Verbiest, L. Pinoo y B. Voeten, "The Impact of the ATM Concept on Video Coding". IEEE J. Select. Areas Commun., Vol. SAC-6, No. 9, Diciembre 1988.

[VirRob89] J. Virtamo y J. Roberts, "Evaluating Buffer Requirements in an ATM Multiplexer". IEEE GLOBECOM'89, Dallas 1989.

[Woo82] R. W. Wolff, "Poisson Arrivals See Time Averages". Operations Research, Vol. 30, No. 2, Abril 1982.

[WooKos90] G. M. Woodruff y R. Kositpaiboon, "Multimedia Traffic Management Priciples for Guaranteed ATM Network Performance". IEEE J. Select. Areas Commun., Vol. SAC-8, No. 3, Abril 1990.

[Yasetal89] Y. Yasuda, H. Yasuda, N. Ohta y F. Kishino, "Packet Video Transmission Through ATM Networks". IEEE GLOBECOM'89, Dallas, 1989.

[Yinetal90] N. Yin, S.-Q. Li, T. E. Stern, "Congestion Control for Packet Voice by Selective Packet Discarding". IEEE Trans. Commun., Vol. COM-38, No. 5, Mayo 1990.

[YokKod89] T. Yokoi y K. Kodaira, "Grade of Service in the ISDN Era". IEEE Comm. Magazine, Abril 1989.

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Curs: 1991

ACTA DEL GRAU DE DOCTOR

El/la doctorand/a Jorge García Vidal

nascut/uda el dia 17 de juny del 19 55, a El Ferrol

província de La Coruña amb el DNI núm. [illegible] i amb estudis de

Enginyer de Telecomunicació

cursats a la Universitat Politècnica de Catalunya

amb el títol d' Enginyer de Telecomunicació

lliurat en data 14 d'octubre de 19 88 (resguard del títol)

admès/esa al programa de doctorat Arquitectura i tecnologia de computadors

de[1] Dept. d'Arquitectura de Computadors

amb el projecte de tesi acceptat en data 8 d'octubre de 19 91 i amb la Suficiència Investigadora lliurada en data 25 de novembre de 19 91 té aprovats els 32 crèdits que exigeix el RD 185/1985.

Barcelona, 19 de febrer de 19 92

Vist i plau
El/la Vice-rector/a d'investigació

Reunit el Tribunal designat a l'efecte, el/la doctorand/a exposa el tema de la seva TESI DOCTORAL: Modelos Analíticos para la Evaluación de Mecanismos de Control de Tráfico en Redes ATM

Acabada la lectura i després de donar resposta a les qüestions formulades pels membres del Tribunal, aquest atorga a la tesi esmentada la qualificació d' APTO "CUM LAUDE" POR UNANIMIDAD

Barcelona, 28 de febrero de 19 92

El/la President/a

NOM Mateo Valero Cortés

El/la Secretari/a

NOM Ramon Ferrús Ferré [illegible]

VOCAL

NOM [illegible]

VOCAL

NOM [illegible]

VOCAL

NOM [illegible]