CRONOEJECUCIÓN Y GESTIÓN ADAPTATIVA DE PAQUETES EN EL SISTEMA DE COMUNICACIONES MÓVILES UMTS

José F. Monserrat del Río Ing. de Telecomunicación
Daniel Calabuig Soler Ing. de Telecomunicación
David Gómez Barquero Ing. de Telecomunicación
Narcís Cardona Marcet PhD en Telecomunicación

Investigadores del Grupo de Comunicaciones Móviles
Universidad Politécnica de Valencia - Instituto de Telecomunicaciones y Aplicaciones Multimedia

jomondel@iteam.upv.es dacaso@iteam.upv.es dagobar@iteam.upv.es ncardona@dcom.upv.es
ESPAÑA

Fecha de Recepción del Articulo: 24-03-06 *Fecha de Aceptación del Articulo: 27-04-06*

Articulo Tipo 1

RESUMEN

Fruto del imparable crecimiento que ha experimentado la transmisión de paquetes sobre los sistemas de comunicaciones móviles, ha surgido un creciente interés, tanto por parte de las grandes compañías de telefonía móvil como de los fabricantes, por desarrollar algoritmos eficientes de gestión adaptativa de paquetes. El objetivo principal de la gestión adaptativa de paquetes en los sistemas de comunicaciones móviles es maximizar la capacidad del sistema proporcionando los niveles de cobertura y calidad de servicio deseados, para una cantidad de recursos radio e infraestructura desplegada dadas. En este artículo se presenta un estudio minucioso del proceso de cronoejecución o gestión adaptativa de paquetes en el sistema de comunicaciones móviles UMTS. De esta manera no sólo se describen los conceptos generales relacionados con la cronoejecución sino que también se proporciona una visión práctica de los canales de datos estandarizados y de la implementación real de la cronoejecución en UMTS. Además, se ofrece una comparativa de distintos algoritmos típicos utilizados en la cronoejecución con dos nuevas técnicas, propuestas ambas con el objetivo de minimizar el retardo de los paquetes. Los resultados obtenidos demuestran que las mejores prestaciones del sistema se obtienen realizando conjuntamente la maximización de la tasa binaria transmitida y la minimización del retardo.

PALABRAS CLAVES

Conmutación de paquetes, Cronoejecución, UMTS, Retardo de paquetes

ABSTRACT

Due to the high increase in the packet transmission over mobile communications systems, the big telecommunications operators together with the main vendors have focused their researching effort on the development of efficient packet scheduling algorithms. In the mobile communications systems the adaptive packet scheduling aims at maximizing the overall system capacity, guaranteeing the desired levels in coverage and quality of service, given a certain network infrastructure. This paper performs and in-depth analysis of the scheduling process within the UMTS mobile communications system. Not only the general concepts involved in scheduling are boned up but also a practical overview of the scheduling implementation is provided. Several algorithms are compared with two new proposals, both focused on minimizing the packet delay. The obtained results justifies that the transmitted bit rate maximization carried out together with the delay minimization improves the overall system performances.

KEYWORDS

Packet Switching, Scheduling, UMTS, Packet Delay

INTRODUCCIÓN

En la actualidad las comunicaciones móviles han alcanzado una enorme penetración de mercado en todo el mundo. Estudios recientes señalan que, por ejemplo, nueve de cada diez ciudadanos de la Unión Europea disponen ya de un teléfono móvil. El importante incremento del número de usuarios y la creciente demanda de una nueva variedad de servicios han hecho surgir una serie de necesidades que no pueden ser cubiertas por los sistemas de comunicaciones móviles de segunda generación (GSM o IS-95). Ante tales expectativas se acometió, ya en el año 1999 y a nivel mundial, el desarrollo de una nueva generación de sistemas de comunicaciones móviles, capaces de superar las limitaciones de sus predecesores. Los sistemas de tercera generación, cuya versión europea recibe el nombre de UMTS (*Universal Mobile Telecommunication System*) [1], soportan una mayor capacidad y mayores velocidades de transmisión.

Este aumento en la velocidad de transmisión ha permitido el desarrollo de nuevos servicios móviles, como la videotelefonía, el acceso a Internet o la transmisión de video bajo demanda. Estos nuevos servicios multimedia se han convertido en un producto clave para los operadores móviles. Este hecho junto con el vació existente en las especificaciones, ha incentivado un espectacular interés por parte de la comunidad internacional científica por la investigación y desarrollo de nuevas técnicas de gestión de recursos radio dinámicas [2]-[6].

Como se verá posteriormente, el principal objetivo de los algoritmos de asignación dinámica de recursos es distribuir de manera óptima los recursos compartidos entre los usuarios activos de manera que se maximice su uso a la vez que se minimice el retardo en la transmisión de paquetes. Sin embargo, hasta la fecha muchas de las técnicas propuestas o son incapaces de conseguir una asignación realmente óptima o no pueden trabajar en tiempo real dada su alta complejidad de cómputo.

Por otro lado las redes neuronales realimentadas se han mostrado como una herramienta extremadamente potente para resolver problemas de optimización complejos con tiempos de respuesta de pocos microsegundos. De esta manera en [4]-[6] se trató de resolver el problema de la gestión dinámica de recursos centrándose únicamente en la maximización de uso de los recursos obviando la minimización del retardo. Este artículo mostrará como las prestaciones de los algoritmos de asignación de recursos obtienen mucho mejores prestaciones al introducir esta nueva restricción.

Este artículo describe, en primer lugar, el concepto general de cronoejecución o gestión adaptativa de paquetes, para pasar posteriormente a analizar los distintos métodos de transmisión de paquetes estandarizados en el sistema UMTS. Finalmente se analizan varias técnicas de cronoejecución evaluando las ventajas e inconvenientes de cada una de ellas.

1. CONCEPTO DE CRONOEJECUCIÓN DE PAQUETES

En general los distintos servicios ofrecidos por un operador de telefonía móvil se pueden clasificar en servicios de tiempo real (RT *Real Time*) y servicios de no tiempo real (NRT *Non Real Time*). Los primeros, entre los que se incluyen por ejemplo los servicios de voz y videoconferencia, se caracterizan porque los datos se generan de forma continua y deben ser entregados al receptor por orden de generación y con un retardo máximo garantizado. Para ofrecer una determinada calidad de servicio (QoS *Quality of Service*) en los servicios RT es necesario reservar una serie de recursos fijos durante toda la comunicación; se trata por tanto de servicios que habitualmente se ofrecen por conmutación de circuitos, es decir, reservando un canal exclusivo para cada comunicación. Por otro lado, los servicios NRT (también conocidos como servicios de paquetes), como el correo electrónico, el tráfico web o las descargas de datos, se caracterizan porque el tráfico se genera a ráfagas y porque los paquetes de información han de llegar sin errores pero sin un orden establecido al receptor. En este tipo de servicios tampoco hay una restricción fija de retardo en la transmisión. Por todo esto los servicios NRT se suelen ofrecen por conmutación de paquetes, es decir, varias comunicaciones comparten el mismo canal y una entidad externa decide quien hace uso del canal en cada momento. Si por el contrario los servicios NRT se ofrecieran por conmutación de circuitos habría una clara ineficacia en la utilización de los recursos puesto que al tratarse de tráfico a ráfagas durante grandes periodos de tiempo no se estaría haciendo uso de los recursos reservados.

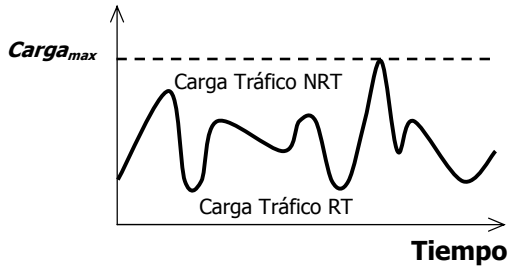
El algoritmo de cronoejecución de paquetes, o *packet scheduler*, es el encargado de gestionar el acceso al canal de todos los servicios de paquetes. La cronoejecución es sin duda uno de los algoritmos básicos en la gestión de recursos radio (RRM o *Radio Resource Management*) dentro de los sistemas de comunicaciones móviles. El algoritmo de *scheduling* decide cuándo se inicia una

transmisión de paquetes así como su velocidad. La cronoejecución tiene, por tanto, un fuerte impacto en las prestaciones de la red, y es uno de los algoritmos fundamentales de gestión de recursos radio en los sistemas móviles que admiten conmutación de paquetes.
Las tareas más importantes del algoritmo de *scheduling* son:

- Determinar los recursos disponibles para servicios de paquetes.
- Repartir los recursos disponibles entre los usuarios.
- Monitorizar la asignación de dichos recursos.
- Monitorizar la carga del sistema, es decir la relación entre recursos consumidos y recursos disponibles en la celda.
- Ejecutar acciones de control de la congestión para los servicios de paquetes cuando corresponda.

Una parte de la cronoejecución de paquetes se realiza en el terminal móvil y otra en la red. El terminal móvil es el encargado de la gestión de los paquetes en el enlace ascendente mediante un proceso FIFO (F*irst Input First Output*). Por otro lado, el móvil provee a la red de todas las medidas necesarias para realizar la cronoejecución en enlace descendente. La parte realizada por la red, en el *packet scheduler*, se encarga de estimar la capacidad disponible para servicios de paquetes y distribuirla entre los usuarios activos.
Desde el punto de vista de la gestión de recursos radio, el sistema se encontrará en un punto de funcionamiento óptimo cuando la carga del sistema sea igual al nivel máximo tolerable, ni más (lo cual provocaría una degradación de la calidad de servicio por bloqueo de llamadas o pérdidas de paquetes), ni menos (lo cual supondría malgastar recursos).
La Figura 1 muestra la relación directa que existe entre los recursos disponibles para la transmisión de paquetes (servicios NRT) y los recursos consumidos por los servicios en tiempo real (RT), prioritarios para el sistema.
Es importante destacar que el tráfico de paquetes que gestiona el algoritmo de scheduling es un tráfico no garantizado, es decir, sin restricciones de retardo y sin una tasa mínima garantizada, en la medida en que la capacidad disponible tampoco está garantizada. Debido a esta restricción, en general no se podrá transmitir toda la información de datos disponible en un determinado momento, por lo que habrá que establecer criterios de priorización en las trasmisiones. Las informaciones referentes a las diferentes clases de servicio (interactive, background) o grados de servicio (premium), serán utilizadas en la definición de dichas prioridades.

Figura 1 Ilustración del papel del *packet scheduling* en la carga.



El objetivo principal de la gestión de paquetes en los sistemas de comunicaciones móviles es maximizar la capacidad del sistema (en términos de número de usuarios activos soportados por la red) proporcionando los niveles de cobertura y calidad de servicio deseados, para una cantidad de recursos radio e infraestructura desplegada dadas.
Los mecanismos de cronoejecución en la red no están estandarizados, por lo que son un elemento diferenciador importante entre equipos de distintos fabricantes. Con ello se persigue que los fabricantes puedan desarrollar y optimizar sus propios algoritmos con el objetivo de obtener una ventaja competitiva respecto a los usuarios finales. Es por tanto de suma importancia en la optimización de los sistemas de comunicaciones móviles actuales el desarrollo de potentes y eficaces algoritmos de *scheduling*.

2. CRONOEJECUCIÓN EN UMTS

El *paquet scheduler* en UMTS se encuentra en el RNC [1] (*Radio Network Controller*, encargado de la gestión de recursos radioeléctricos de la red UMTS) y recibe medidas de la carga de la red de las estaciones base y medidas del tráfico en enlace ascendente de los móviles. El *packet scheduler* trabaja periódicamente en intervalos temporales con valores típicos entre 100 ms y 1 s. El *packet scheduler* utiliza la siguiente información:

- Potencia total utilizada por las estaciones base o Nodos B [1]. Esta información representa una medida de la carga utilizada en enlace descendente.
- Capacidad utilizada por las conexiones en tiempo real. Esta información se puede obtener a partir de la estimación de la carga en enlace ascendente basada en promediado estadístico.

- Umbral máximo para el nivel de carga establecido durante la fase de planificación. Este parámetro define el nivel de interferencia máximo que puede existir en la celda para que las conexiones en tiempo real no vean degradada su QoS.
- Peticiones de incremento de tasa binaria de los móviles.

Los Nodos B proporcionan periódicamente al RNC información sobre la potencia total utilizada así como las potencias dedicadas a cada conexión. Con esta información el *packet scheduler* puede estimar la potencia total utilizada para tráfico no controlable (que consiste en tráfico en tiempo real e interferencias intercelulares). Esta parte de la interferencia no puede ser afectada por el *packet scheduler*. La capacidad restante es compartida por los usuarios de paquetes activos.

2.1 MÉTODOS DE TRANSMISIÓN DE PAQUETES EN UMTS

Existen tres tipos de canales de trasporte en UMTS que pueden utilizarse para transmitir información de paquetes: canales comunes, canales dedicados y canales compartidos. A continuación se describen las propiedades y ventajas de cada tipo de canal para transmitir información de paquetes.

2.1.1 Canales Comunes (RACH/FACH/CPCH)

Los canales comunes son el RACH (*Random Access Channel*) y el CPCH (*Common Packet Channel*) en el enlace ascendente o *uplink* y el FACH (*Forward Access Channel*) en enlace descendente o *downlink*. El RACH y el FACH suelen ser utilizados para transmitir información de señalización, pero también datos. Normalmente sólo suele haber unos pocos canales RACH y FACH por sector (a veces sólo uno).
La principal ventaja de los canales comunes es que no se realiza un proceso de establecimiento de la conexión, por lo que no hay tiempo de establecimiento de la conexión.
Como los canales comunes no tienen un canal de retorno (son unidireccionales), no se puede utilizar el control de potencia rápido característico de UMTS y se utiliza simplemente una potencia fija. Además en estos canales tampoco se pueden realizar traspasos (*handovers*). Por todo ello, el rendimiento a nivel de enlace de los canales comunes es peor que el de los canales dedicados, por lo que producen más interferencias.
Los canales comunes FACH y RACH son adecuados para transmitir paquetes IP pequeños, como, por ejemplo, durante la fase de establecimiento de la conexión de TCP, ya que el establecimiento de la conexión mediante TCP implica la transmisión de varios paquetes IP pequeños (40 bytes).
El CPCH es una versión mejorada del canal RACH en el que se puede utilizar control de potencia rápido (pero no *soft handover*). Además se pueden asignar hasta 64 tramas (640 ms) a un usuario, con lo que se puede transmitir más información que en el RACH. El inconveniente del CPCH es que el tiempo de establecimiento de la conexión es mayor que el del RACH.

2.1.2 Canales Dedicados (DCH)

A pesar de que la transmisión de paquetes siempre es más eficiente mediante conmutación de paquetes, UMTS también permite la posibilidad de transmitir servicios NRT mediante conmutación de circuitos haciendo uso de canales dedicados DCH (*Dedicated Channel*). Los canales dedicados son canales bidireccionales tanto en *uplink* como en *downlink*, y permiten realizar el control de potencia rápido y *soft handovers*, por lo que generan menos interferencia que los canales comunes. El principal inconveniente de los canales dedicados es que es necesario realizar un proceso de establecimiento de la conexión, lo cual implica un mayor retardo comparado con el acceso a los canales comunes.
Los canales dedicados permiten tasas binarias desde unos pocos kb/s hasta 384 kb/s. La tasa binaria de un canal dedicado no es un parámetro fijo, sino que puede variar a lo largo de la conexión. Una vez se termina la información a transmitir, el usuario mantiene el canal dedicado por unos pocos segundos antes de liberar el canal para que pueda ser asignado a otro usuario. Esto implica que los canales dedicados no sean eficientes para tráfico a ráfagas sino mejores para transmitir cantidades de información medias o grandes.
Dada la alta complejidad del *packet scheduler*, en las primeras versiones comerciales de los RNC-UMTS no se incluía su implementación por lo que los servicios NRT siempre se soportaban mediante canales DCH. Desde principios de 2005 las nuevas RNCs incorporan ya los mecanismos de cronoejecución para su aplicación a los canales compartidos.

2.1.3 Canales Compartidos (DSCH)

El DSCH (*Downlink Shared Channel*) se utiliza para transmitir tráfico de paquetes a ráfagas en el enlace descendente. La idea es compartir un único canal físico (es decir, un código) entre muchos usuarios multiplexándolos temporalmente. Si se utilizaran canales dedicados para este tipo de

tráfico, todos los usuarios requerirían un código que vendría determinado por sus tasas binarias máximas. Cuando un usuario termina su información a transmitir, los recursos del DSCH se asignan a otro usuario inmediatamente, incrementando la eficiencia respecto a los canales dedicados. Los canales compartidos se utilizan conjuntamente con canales dedicados de baja tasa binaria, donde se transmite el canal de control físico, la señalización y los comandos de control de potencia. El inconveniente del DSCH es que no admite *soft handovers*.

3. GESTIÓN DE PAQUETES MEDIANTE DSCH

En sentido ascendente, el acceso al medio físico para la transmisión de servicios NRT mediante conmutación de paquetes se realiza por competición por lo que no se aplica ningún algoritmo de cronoejecución. En el enlace descendente (DSCH principalmente) sí que se hace uso de estos algoritmos. En este apartado se analizan en profundidad los distintos procedimientos de gestión necesarios para la transmisión de paquetes mediante DSCH.

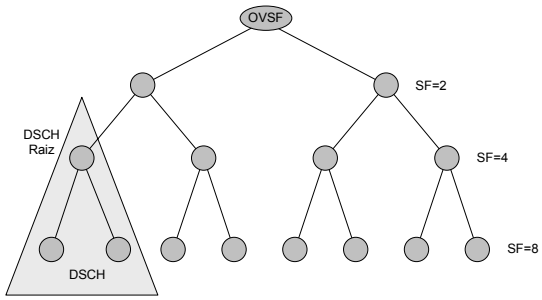
3.1 GESTIÓN DE CÓDIGOS

Como en cualquier sistema CDMA (*Code Division Multiple Access*), en UMTS los distintos usuarios se multiplexan utilizando un conjunto finito de códigos ortogonales. En UMTS estos códigos se denominan códigos OVSF (*Orthogonal Variable Spreading Factor*) y se generan mediante una estructura en forma de árbol como la que se muestra en la Figura 2.

Los árboles OVSF tienen la propiedad de que dos o más códigos pertenecientes a ramas distintas son ortogonales, mientras que códigos que pertenecen a la misma rama no guardan dicha ortogonalidad. De esta manera se puede disponer de 4 códigos con factor de ensanchado (SF o *Spreading Factor*) 4, 8 con SF=8 etc.

Normalmente para la transmisión de paquetes en UMTS se asigna una rama de código OVSF para el canal DSCH, como se aprecia en la Figura 2. La elección de qué ramas de código OVSF se asignan para DSCH se realiza de forma dinámica en función de la carga de los distintos servicios. Por ejemplo, si no hubiera ninguna comunicación RT se podría asignar a transmisión de paquetes hasta 7 de las 8 ramas SF=8 ya que la octava está reservada para la transmisión de los canales comunes necesarios para el funcionamiento de la red (señal de *broadcast*, canal piloto etc.) y para los canales dedicados asociados a cada usuario DSCH.

Figura 2 Rama de código para el canal DSCH



Dado un número *N* de conexiones simultáneas, en [7] se establece una condición necesaria para que esa distribución de recursos sea realizable, es decir, para que se pueda afirmar que hay suficientes códigos para permitir esa asignación.

$$\sum_{i=1}^{N} \frac{1}{SF_i} \leq \frac{1}{SF_{Raiz}} \tag{1}$$

Según esta expresión si se tiene en cuenta todo el árbol de códigos el $SF_{Raiz} = 1$ por lo que

$$\sum_{i=1}^{N} \frac{1}{SF_i} < 1 \tag{2}$$

Si, por otro lado consideramos dos partes del árbol de códigos diferenciadas (los códigos reservados a *M* transmisiones de conmutación de circuitos DCH y los códigos reservados para *L* conexiones por conmutación de paquetes con DSCH) se deberá de cumplir:

$$\begin{aligned} \sum_{i=1}^{L} \frac{1}{SF_i} + \sum_{i=1}^{M} \frac{1}{SF_i} &< 1 \\ \sum_{i=1}^{L} \frac{1}{SF_i} &< 1 - \sum_{i=1}^{M} \frac{1}{SF_i} \end{aligned} \tag{3}$$

O lo que es lo mismo, teóricamente todos los códigos no utilizados por los servicios RT podrán ser utilizados por los servicios NRT. En este sentido, en la asignación de códigos, el comportamiento de UMTS es el reflejado en la Figura 1.

3.2 ASIGNACIÓN DE POTENCIA

Tras asignar el número de códigos disponibles para canales compartidos, y conociendo el número de usuarios que demandan servicio es importante comprobar la disponibilidad de potencia en el Nodo-B. Para ello es necesario encontrar una expresión para la potencia consumida en la estación base en

downlink (DL). La relación de energía de bit a densidad espectral de ruido más interferencia E_b/N_0 de un usuario en DL es:

$$\left(\frac{E_b}{N_0}\right)_{DL,i} = \frac{\frac{P_{DL,i}}{L_{pDL,i}}\frac{W}{R_i}}{P_{NDL} + \chi_i + \rho\frac{P_{DL} - P_{DL,i}}{L_{pDL,i}}} \quad (4)$$

donde $P_{DL,i}$ es la potencia destinada al usuario *i*, *W* la tasa en chips/s del sistema UMTS (3.84 Mchips/s), R_i es la tasa binaria del servicio asociado al usuario *i*, $L_{pDL,i}$ son las pérdidas de propagación, P_{NDL} es el ruido térmico, χ_i es la potencia interferente a la entrada del usuario *i*, ρ es el factor de ortogonalidad ($\rho=0$ para códigos perfectamente ortogonales) y P_{DL} es la potencia total usada por la estación base. Recordar que SF se define exactamente como:

$$SF_i = \frac{W}{R_i} \quad (5)$$

A partir de (4) [8]:

$$P_{DL,i} = \frac{L_{pDL,i}\left(P_{NDL} + \chi_i\right) + \rho P_{DL}}{\rho + \frac{W}{\left(\frac{E_b}{N_0}\right)_{DL,i} R_i}} \quad (6)$$

Con lo que finalmente

$$P_{DL} = \frac{\sum_{i=1}^{N} \frac{L_{pDL,i}\left(P_{NDL} + \chi_i\right)}{\rho + \frac{W}{\left(\frac{E_b}{N_0}\right)_{DL,i} R_i}}}{1 - \sum_{i=1}^{N} \frac{\rho}{\rho + \frac{W}{\left(\frac{E_b}{N_0}\right)_{DL,i} R_i}}} \quad (7)$$

Mediante esta expresión el *packet scheduler* es capaz de estimar la potencia total que consumiría una determinada asignación de recursos. Para ello en UMTS los terminales móviles envían periódicamente lo que se conoce como *Measurement Report* o informe de medidas mediante el que el RNC conoce el nivel de potencia recibido por cada terminal del canal piloto y el nivel de interferencias χ_i experimentado. Como la potencia transmitida por el canal piloto es fija se puede estimar $L_{pDL,i}$. Por último la potencia de ruido térmico es constante por lo que todos los términos de la expresión de la potencia total se conocen con exactitud salvo el factor de ortogonalidad que tiene que haber sido especificado en la fase de planificación y despliegue de la red.

Si el conjunto de recursos solicitados en sentido descendente por todos los usuarios supera la potencia máxima disponible en la estación base se iniciaría el proceso de *scheduling* o de asignación de recursos.

Además de la limitación en la potencia máxima del sistema, en UMTS también existe una limitación máxima en la potencia asignada a una sesión. Esta potencia máxima suele estar comprendida entre 1 y 2 W (30 – 33 dBm).

3.3 PACKET SCHEDULING

Todos los usuarios NRT están vinculados a uno o varios canales DSCH. Cada canal DSCH habilitado estará por tanto compartido por múltiples usuarios. El *packet scheduler* se encarga de seleccionar el uso que hacen los usuario de cada canal DSCH Teniendo en cuenta las anteriores restricciones de potencia y asignación de códigos, el algoritmo de cronoejecución asigna los canales DSCH a los distintos usuarios siguiendo dos principios básicos: la prioridad de los usuarios y la disponibilidad de recursos. La priorización es un proceso fundamental en el que los usuarios se ordenan según un determinado criterio que podrá tener en cuenta la tasa media asignada a cada usuario por contrato, el tamaño de los datos pendientes de transmitir o el retardo acumulado por estos datos. Una vez ordenados los usuarios el siguiente paso consiste en decidir qué tasa de transmisión va a reservarse a cada comunicación definiendo el número de slots asignados a cada usuario en el siguiente periodo de *scheduling* (desde 100 ms hasta 1s). Recorriendo la lista de usuarios priorizados se van asignando recursos comprobando que las limitaciones en códigos y en potencia se satisfacen. Una vez se supere alguna de las dos limitaciones se dará por finalizada la asignación de recursos.

4. TÉCNICAS DE CRONOEJECUCIÓN

Una vez analizada la gestión de paquetes en UMTS mediante el uso del canal compartido DSCH, en esta sección se describen los distintos algoritmos de *packet scheduling* que serán evaluados numéricamente en la sección de resultados.

4.1 ROUND ROBIN (RR)

Esta técnica es sin duda una de las más clásicas dentro de la gestión de recursos radio. Se basa en

que todos los usuarios tienen el mismo nivel de prioridad y por tanto los recursos se van distribuyendo ecuánime y cíclicamente. De esta manera se establece un intervalo de tiempo predeterminado que constituye el intervalo de asignación de recursos. Por otra parte se construye una lista con todos los usuarios activos. En el primer intervalo de tiempo comienzan transmitiendo los *N* primeros usuarios de la lista a la máxima tasa binaria disponible en el sistema estando *N* determinado por las restricciones de potencia y disponibilidad de códigos. Por ejemplo si el hecho de que el usuario *N*+1 también transmita hace que se sobrepase el límite de recursos establecido entonces transmitirán los *N* primeros usuarios. Tras este primer turno transmitirán el usuario *N* +1 y los *M* siguientes, calculando *M* de la misma manera que *N* en el primer turno. Es importante destacar que aunque a los usuarios se les debería asignar la máxima tasa binaria disponible en el sistema esto no siempre es posible por la limitación establecida por la red en la potencia máxima por sesión. En caso de que se supere esta potencia máxima, el usuario conserva su turno pero reduciendo su tasa de transmisión para no sobrepasar la potencia máxima.

4.2 OPTIMUM BIT RATE (OBR)

Este algoritmo definido en [6] tiene en cuenta el retardo en la transmisión de paquetes, una importante métrica de la calidad de servicio de un usuario. La restricción del retardo de paquete o *packet delay* es crítica en servicios RT ya que no se debe exceder el máximo establecido (*deadline*). Para los servicios NRT esta restricción no es tan importante aunque siempre es deseable reducir el *packet delay* para aumentar la QoS.
Para introducir el retardo en la gestión de recursos se define $R_{min,i}$ como la tasa mínima que debe ser asignada al usuario i-ésimo de manera que se garantice una correcta transmisión de paquetes. Por ejemplo, en el servicio de tráfico web donde existen múltiples flujos de datos, uno por cada descarga de página simultánea, la tasa mínima se puede calcular como:

$$R_{\min,i} = P \cdot \max_b \left(\frac{B_b}{t_b} \right) \quad (8)$$

Y para videoconferencia donde solamente existe un flujo de datos:

$$R_{\min,i} = \max_b \left(\frac{\sum_{p=b}^{L_B} B_p}{t_b} \right) \quad (9)$$

Donde B_b y t_b son el número de bits y el tiempo que resta para exceder el máximo retardo de la posición *b*-ésima del buffer, *P* es el número de descargas simultáneas en un escenario web y L_B es la longitud del buffer. Es importante destacar que según este modelo los datos generados en el mismo instante temporal se almacenan en la misma posición del buffer. Por lo tanto el tamaño de cada una de las posiciones del buffer B_b no está fijado a priori. Además se ha supuesto que para el tráfico web, dada una determinada tasa binaria, ésta se divide equitativamente entre todos los flujos de datos, o lo que es lo mismo, entre todas las páginas web descargadas simultáneamente.
Aunque los servicios NRT carecen de un retardo máximo se puede definir un *delay* máximo aconsejado que permitirá obtener una determinada calidad de servicio. Este máximo se podrá exceder sin que ello suponga una pérdida de paquetes pero forzará al sistema a optimizar el servicio.
En el algoritmo OBR se intenta asignar a todos los usuarios la tasa mínima que garantice su *delay* máximo. Tras establecer una lista de usuarios activos, de manera aleatoria se va seleccionando al siguiente usuario al que se le asignará exactamente su tasa binaria $R_{min,i}$. Si llegado un determinado usuario se exceden los recursos máximos del sistema entonces ese usuario se descarta y se intenta dar servicio al resto de usuarios en cola aún no asignados hasta finalizar la lista de usuarios o agotar los recursos disponibles.

4.3 CODE-DIVISION GENERALIZED PROCESSOR SHARING (CDGPS)

Basado en [9] este algoritmo prioriza a los usuarios en función del tiempo que, dada la asignación actual de recursos, un usuario tardaría en vaciar la cola o buffer de datos. A este tiempo se le denomina tiempo virtual de vaciado de cola. Cuanto mayor es este tiempo mayor es la prioridad del usuario. Inicialmente todos los usuarios transmiten a la menor tasa disponible en el sistema. En cada iteración del algoritmo calcula el tiempo virtual de todos los usuarios activos e incrementa la tasa binaria del usuario más prioritario. Si a ese usuario se le ha asignado la máxima tasa binaria disponible en el sistema, si se ha superado la potencia máxima por sesión o si mediante esa nueva asignación se superase la carga máxima del sistema, entonces este usuario se considera definitivamente asignado y se le elimina de la lista de usuarios a tener en cuenta en la siguiente iteración.

4.4 DESCEND BIT RATE (DBR)

Este nuevo algoritmo, llamado DBR, asigna inicialmente la máxima tasa binaria a todos los usuarios e iteración a iteración va reduciendo aleatoriamente la tasa binaria. El objetivo es como siempre alcanzar una asignación de recursos que no sobrepase los recursos máximos disponibles. El proceso de reducción de tasa se divide en dos fases. En la primera, el DBR nunca asigna tasas binarias menores a la $R_{min,i}$ (ecuaciones (8) y (9)) de cada usuario de manera que si no hay recursos suficientes todos los usuarios acaban teniendo asignada su $R_{min,i}$ correspondiente. En la segunda fase, si aún la asignación de recursos sobrepasa el máximo, el algoritmo sigue reduciendo la tasa binaria al azar pero ahora sin un límite inferior definido. De esta manera se garantizar siempre que sea posible la asignación de la $R_{min,i}$ y por tanto no sobrepasar el retardo máximo de los paquetes.

4.5 MINIMUM COST AND RATE OPTIMISATION (MICRO)

Este algoritmo se basa en la minimización de una determinada función de coste global del sistema C y en la conjunta maximización de la tasa binaria global del sistema. C se define como la suma de los costes individuales de cada usuario i, C_i. Cada coste individual depende de la tasa binaria del usuario, R_i, y de su correspondiente $R_{min,i}$. Así, siempre que se cumpla que la tasa binaria asignada a un determinado usuario R_i, sea mayor que la tasa mínima requerida $R_{min,i}$, el coste C_i será decreciente para valores de R_i crecientes.
Para obtener un buen comportamiento sería deseable que la función C_i fuera monótonamente decreciente y que estuviera acotada superior e inferiormente. Además es necesario que haya un importante decremento para tasas binarias asignadas a partir de $R_{min,i}$. Una función monótona, acotada y que presenta un gran crecimiento a partir de un determinado valor es la función sigmoidal:

$$S(x,s,r)=\frac{1}{1+e^{-s(x+r)}} \tag{10}$$

La función de coste queda definida como

$$C=\sum_{i=1}^{N} C_i$$

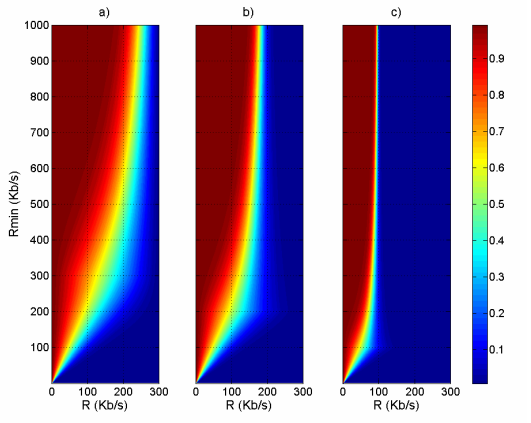
$$C_i=\frac{S\left(\max\left(R_{\max,i}\right),s_i,r_i\right)-S\left(R_i,s_i,r_i\right)}{S\left(\max\left(R_{\max,i}\right),s_i,r_i\right)-S\left(0,s_i,r_i\right)} \tag{11}$$

donde N es el número de usuarios activos y $R_{max,i}$ es la máxima tasa binaria permitida al usuario i. De esta manera, C_i toma valores entre [0,1] para todas las tasas binarias del sistema $\left\{0,\cdots,\max\left(R_{\max,i}\right)\right\}$. Si la tasa binaria asignada es igual a cero entonces la función de coste será máxima e igual a 1. Si por el contrario se asigna la tasa binaria máxima disponible por el usuario el coste individual será mínimo e igual a 0. El valor de s_i y de r_i se ha elegido convenientemente para que la disminución brusca de la función sigmoidal se produzca para valores de $R_i > R_{min,i}$.
La Figura 3 muestra tres ejemplos gráficos de la función de coste estudiando tres distintos tipos de usuario con tasa máxima de 300, 200 y 100 kb/s. Todos los casos tienen el mismo comportamiento si $R_{\min,i}\le R_{\max,i}$.

Es importante destacar que esta función de coste presenta su valor mínimo para la tasa $R_{\max,i}$ por lo que en la misma función de coste se incluye también el proceso de maximización de la tasa binaria del sistema. Además si se sobrepasa el retardo máximo, $R_{\min,i}=\infty$, la función de coste pasa a valer 1 para $R_i < R_{\max,i}$ y 0 si $R_i \ge R_{\max,i}$.

Una vez definido el valor de la función de coste solamente falta optimizar el sistema para los N usuarios activos.
Matemáticamente el problema se puede resolver utilizando diversos métodos numéricos. Un método hardware-implementable de resolución de este tipo de problemas son las redes neuronales realimentadas [10], útiles por su alta velocidad en la obtención del óptimo. En este artículo la optimización de la función de coste se ha realizado haciendo uso de estas redes neuronales.

Figura 3 Función de coste para tasa máxima de 300 (a), 200 (b) y 100 (c) kb/s



5. RESULTADOS NUMÉRICOS

5.1 MODELO DE TRÁFICO

El tráfico se genera en base al modelo ON-OFF propuesto en [11]. Solamente se ha considerado en la simulación el tráfico web al ser éste el más problemático por su comportamiento a ráfagas. Además, los paquetes generados de más de 1MB tampoco se han considerado en las estadísticas mostradas ya que se consideran ficheros de descarga FTP donde el retraso ya no se tiene en cuenta como un parámetro de calidad de la red.

5.2 ESCENARIO DE SIMULACIÓN

En las simulaciones se ha considerado la especificación UMTS Release 99 [8]. El escenario consiste básicamente en 7 celdas de radio 0.5 km, siendo la celda central la sometida a estudio, y el resto interferentes. El número de usuarios por celda con conexión web establecida es de 120 (aunque no todos estarán transmitiendo simultáneamente ya que el tráfico web presenta bastantes periodos de inactividad). La potencia máxima disponible por celda es de 43 dBm mientras que las celdas interferentes se consideran cargadas al 50% (40 dBm de potencia transmitida).
Las pérdidas de propagación para el usuario *i*-ésimo se calculan mediante la siguiente ecuación de pérdidas [12]:

$$L_{P,i}(\mathrm{dB}) = 137.4 + 35.2 \log_{10}(d_i) \qquad (12)$$

donde d_i es la distancia en km entre el usuario *i*-ésimo y la estación base central.
Los usuarios se encuentran en movimiento con una velocidad aleatoria uniformemente distribuida entre 0 y 60 km/h. El nivel de potencia de ruido térmico es de -102 dBm y el ancho de banda de transmisión, *W,* es igual a 3.84 Mchips/s. El algoritmo de *packet scheduling* se ejecuta cada segundo.
El factor de carga máximo, $\eta_{\max}$, se ha establecido en 0.6. El conjunto de posibles tasas binarias es de {256, 128, 64, 32, 16, 0} kb/s y los niveles de relación E_b/N_0 son respectivamente de {5.6, 4.4, 4.62, 4.55, 4.55, 0} dB [13]. El *delay* máximo aconsejado se ha establecido en 60 s.
Dado que no se han tenido en cuenta servicios RT se pueden reservar para la transmisión con DSCH hasta 7 ramas (una rama estará dedicada a los canales de *broadcasting* y comunes) de códigos con factor de ensanchado 8 (256 kb/s) por lo que la limitación en el número de códigos establece que la tasa binaria máxima de una celda es B_T=1792 kb/s.

5.3 COMPARATIVA

Este apartado pretende analizar el comportamiento de los cinco algoritmos de cronoejecución mostrados en el apartado 4. La comparativa se realizará en función de la tasa binaria asignada y el retardo medio.
Para realizar un estudio lo más fiable posible se han realizado 10 simulaciones de 10.000 segundos de duración en las que las condiciones de actividad de los usuarios son idénticas para todos los algoritmos. Los resultados que se muestran se han obtenido promediando las 10 simulaciones.
La Figura 4 muestra la función de distribución de la tasa binaria total del sistema normalizada en función de B_T. Como se puede observar los mejores resultados son los obtenidos por los dos algoritmos propuestos en este artículo. Esta conclusión era bastante predecible puesto que, por su parte, el algoritmo DBR parte de una asignación de tasa máxima y va reduciendo la tasa hasta alcanzar una solución factible.

Figura 4 CDF de la tasa binaria total de la celda

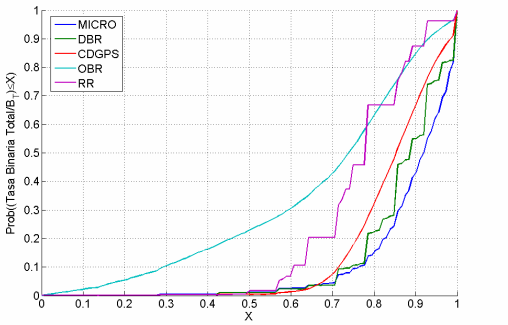


Figura 5 CDF del número de usuarios activos en la celda

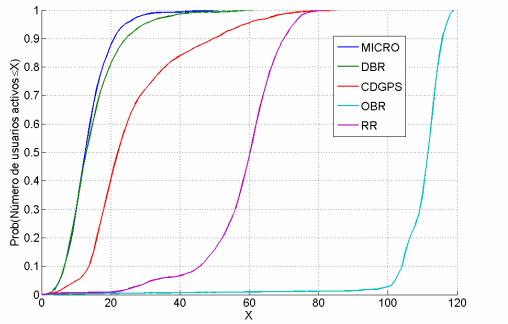
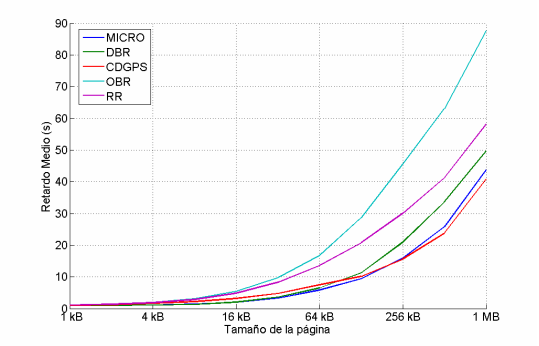


Figura 6 Retardo medio por descarga en función del tamaño de la página



Por otra parte, el algoritmo MICRO es el único que maximiza la tasa binaria total del sistema, lo que justifica su mejor comportamiento respecto a todos los demás. La figura 5 muestra la función de distribución del número medio de usuarios activos por celda. Como se puede observar una mayor tasa global en la celda hace que los usuarios vacíen antes sus colas y por tanto pasen a un estado inactivo en un menor tiempo. Fruto de esto, los algoritmos MICRO y DBR son los que presentan un menor número de usuarios activos y por tanto una mayor tasa media por usuario.

Finalmente la Figura 6 muestra el retardo medio por página web descargada en función del tamaño de la descarga. Como se puede observar, de nuevo el mejor comportamiento lo presenta el algoritmo MICRO salvo para grandes descargas superiores a los 256 kB donde el algoritmo CDGPS se comporta mejor. Esto es debido a que con CDGPS se destinan casi todos los recursos a dar servicio a los usuarios con un buffer muy grande por lo que los usuarios con pequeñas descargas se verán perjudicados. Al ser el número de usuarios con pequeñas descargas mucho mayor, el apoyo a las grandes afectará negativamente al funcionamiento global de la red. Sin embargo, el algoritmo MICRO trata por igual a todos los usuarios, por lo que optimiza el funcionamiento global del sistema a costa de un pequeño empeoramiento del comportamiento de los usuarios con grandes descargas.

6. CONCLUSIONES

Este artículo describe la problemática de la gestión adaptativa de paquetes o *packet scheduling* en UMTS, y en particular mediante el canal compartido DSCH. Diferentes algoritmos han sido comparados para resaltar la importancia de maximizar la tasa binaria total del sistema y de minimizar el retardo simultáneamente.

AGRADECIMIENTOS

Los autores quieren agradecer a la Comisión Interministerial de Ciencia y Tecnología (CICYT) y al Fondo Europeo de Desarrollo Regional la financiación aportada al presente proyecto TIC2005-08211-C02.

7. REFERENCIAS

[1] Especificaciones de UMTS. www.3gpp.org.

[2] MIHAILESCU, C y otros. Performance evaluation of a dynamic resource allocation algorithm for UMTS-TDD systems. En: IEEE Vehicular Technology Conference (VTC) Fall (2000), USA.

[3] FORKEL. L y otros. Dynamic Channel Allocation in UMTS Terrestrial Radio Access TDD Systems. En: IEEE VTC Conference Spring (2001), USA.

[4] LÁZARO, Oscar, Dynamic radio resource management algorithms and traffic models for emerging mobile communication systems. Tesis Doctoral (2001). University of Strathclyde.

[5] AHAN, C.W. y otros. QoS provisioning dynamic connection-admission control for multimedia wireless networks using Hopfield Neural Networks. En: IEEE Transactions on Vehicular Technology (2004), vol. 53, no. 1, pp. 106-117.

[6] GARCÍA, Nuría y otros. A Novel Scheduling Algorithm for Delay-Oriented Services Based on Hopfield Neural Networks Methodology. En: IEEE Wireless Communications and Networking Conference, 2006.

[7] MINN, T. y otros. Dynamic Assignment of Orthogonal Variable Spreading Factor Codes in WCDMA. En: IEEE Journal on Selected Areas in Communications (August 2000), pp. 1429-1440.

[8] PÉREZ ROMERO, Jordi y otros. A Downlink Admission Control Algorithm for UTRA-FDD, En: 4th International Workshop on Mobile and Wireless Communications Network, 2002.

[9] LIANG, Xu y otros. Dynamic bandwidth allocation with fair scheduling for WCDMA systems. En: IEEE Wireless Communications, Vol. 9, pp. 26-32, 2002.

[10] HOPFIELD, J.J y otros. Neural Computation of Decisions in Optimisation Problems. En: Biological Cybernetics (1985), pp 141-152.

[11] STAEHLE, D. y otros. Source Traffic Modeling of Wireless Applications. En: Research Report Series, Institut für Informatik, Universität Würzburg. Report No. 261, 2000.

[12] HOLMA, Harry. WCDMA for UMTS. John Wiley & Sons, 2004.

[13] European Project IST-2000-25133. Advanced Radio Resource Management for Wireless Services (ARROWS).