2 Capítulo 2. ENTORNO DE TRABAJO

2.1 Introducción

En el presente capítulo nos proponemos describir con detalle los principales elementos del sistema de comunicaciones que constituyen el entorno de trabajo en el que se desarrollan las nuevas propuestas realizadas en esta tesis doctoral. Estas nuevas propuestas se refieren tanto a protocolos de acceso como a algoritmos de gestión de los recursos radio.

El capítulo se estructura básicamente en tres partes, que corresponden a los tres niveles estudiados del sistema. En la primera de ellas, correspondiente al punto 2.2, se describe la estructura de la capa física que será utilizada, basada en la tecnología CDMA. Este mecanismo de acceso tiene unas características concretas que deben entenderse para comprender el funcionamiento de las capas superiores. Se ha realizado un breve estudio de los elementos principales del diseño de cualquier sistema basado en este modo de acceso.

En el punto 2.3 se describe la capa de acceso al medio, encargada de organizar los recursos físicos y repartirlos entre los usuarios y las aplicaciones. Se realiza también una descripción del estado del arte en cuanto a protocolos de acceso.

Finalmente, en el punto 2.4 se describe y analiza el problema de la gestión de los recursos radio. También se presenta un estudio del estado del arte, como marco de introducción a las propuestas que serán presentadas en el capítulo 6.

2.2 El acceso CDMA

2.2.1 Introducción

El canal de transmisión en los sistemas de comunicaciones móviles es el aire. Es necesario arbitrar una manera en la que las diferentes comunicaciones puedan compartir este canal radio. La técnica llamada Acceso Múltiple por División en Código, o simplemente CDMA, permite a los usuarios transmitir con la misma frecuencia y de modo simultáneo en el tiempo. Esta técnica es la que se utilizará en los futuros sistemas de comunicaciones móviles de tercera generación.

La separación de los usuarios se hace en base a asignarles a cada uno de ellos un código, de manera que los códigos de diferentes usuarios sean ortogonales entre sí. Cada usuario deberá multiplicar su información a transmitir por su secuencia código asignada. El receptor podrá separar la información de cada usuario haciendo uso del mismo código con el que se transmitió, gracias a la propiedad de ortogonalidad de los mismos. De este modo, todos los usuarios pueden usar todo el ancho de banda disponible durante todo el tiempo.

En los sistemas CDMA convencionales, la estación base dispone de un receptor para cada usuario. En esta situación, todas las transmisiones simultáneas tanto de la propia célula como de las vecinas afectan al resto de conexiones del mismo modo que el ruido. Por tanto, y a diferencia de lo que ocurre con los sistemas TDMA, en la evaluación y diseño del sistema es necesario considerar la interferencia intercelular además de la intracelular. El procedimiento estándar para la igualación del canal y de la señal, de cara a la reducción de las interferencias y la detección de un solo usuario es el llamado receptor RAKE. Este receptor es capaz de

identificar los diferentes caminos de propagación y efectuar un seguimiento de aquellos rayos del canal más significativos.

Por otro lado, las señales CDMA transmitidas se ensanchan en espectro en un factor llamado ganancia de procesado. Este factor es el que en recepción redunda en una mejora de la calidad de la señal cuando se realiza el proceso de desmodulación, en cuanto a relación señal a ruido [11].

En los sistemas CDMA la limitación de capacidad viene definida por las interferencias. Para poder obtener una buena calidad es necesario controlar la interferencia generada por los propios usuarios de una célula (interferencia intracelular). Por tanto es necesario un estricto sistema de control de potencia. El objetivo de este control es que todos los usuarios lleguen a la estación base con el nivel de potencia justo, de modo que ninguno de ellos interfiera más que el resto de forma indebida.

Existe también un tipo de sistemas CDMA con detección multiusuario. En este caso, el sistema aprovecha que conoce todos los códigos con los que los usuarios están transmitiendo su información para tratar de eliminar la interferencia intracelular. Si las células son pequeñas, incluso es posible eliminar parte de la interferencia que generan los usuarios de células vecinas.

Por otro lado, puesto que cada usuario en CDMA utiliza todo el ancho de banda disponible, que es mayor que el estrictamente necesario para sus transmisiones, se puede obtener una mejora significativa en cuanto a la selectividad en frecuencia del canal [12]. Además, el hecho de que la interferencia venga provocada por un gran número de fuentes hace que se reduzca la varianza de la potencia instantánea recibida, vista como un proceso estocástico, lo cual constituye en sí mismo una ventaja de los sistemas CDMA.

Otra característica importante de este modo de acceso es que permite el cambio de la tasa efectiva de transmisión de un modo muy flexible. Tan sólo es necesario cambiar el factor de ensanchamiento espectral de las señales transmitidas. Sin embargo, un aumento de la velocidad, lo que comporta una reducción del factor de ensanchamiento espectral, redunda en una menor protección frente a las interferencias provocadas por los otros usuarios.

La otra vertiente ventajosa de la flexibilidad de los sistemas CDMA es que no se requiere ningún tipo de coordinación entre los usuarios a la hora de realizar sus transmisiones. Cada uno de ellos puede iniciar y finalizar sus transmisiones en cualquier momento. También pueden ajustar de forma individual sus velocidades de transmisión y sus factores de actividad.

Los sistemas CDMA no necesitan ninguna planificación de frecuencias. La separación de los usuarios no se hace mediante la separación en frecuencia y por tanto puede hacerse un reúso completo de las mismas, lo que simplifica su gestión.

Otra característica de los sistemas CDMA es que tienen la llamada capacidad progresiva o *soft-capacity*. Esta propiedad indica que el hecho de añadir nuevos usuarios al sistema es siempre posible, a costa de degradar la calidad de las conexiones en curso, pero no existe un límite físico absoluto al número de usuarios que pueden transmitir de modo simultáneo.

Finalmente, una última ventaja de los sistemas CDMA es que son compatibles y pueden coexistir con los sistemas de banda estrecha, analógicos o digitales, que ya estén operativos. La transmisión de los usuarios CDMA se realiza repartiendo la energía en un gran ancho de

banda, de manera que respecto a los sistemas de banda estrecha, esta transmisión se percibe como una pequeña potencia de ruido añadido.

2.2.2 Principios básicos de CDMA

Para poder compartir un cierto canal de comunicaciones entre diferentes usuarios, es necesario poder separar de algún modo las transmisiones de cada uno de ellos. En términos matemáticos, para poder separar dos señales de información es necesario que exista ortogonalidad entre las citadas señales. Se dice que dos señales $u(t)$ y $v(t)$ son ortogonales cuando se cumple que:

$$\int u(t)v(t)dt = 0 \qquad (2.1)$$

En el caso de señales separadas en tiempo o en frecuencia, la propiedad se cumple de manera evidente. Sin embargo, en las transmisiones CDMA, puesto que las señales de información comparten tiempo y frecuencia, la ortogonalidad se obtiene a base de multiplicarlas por unas ciertas secuencias, llamadas secuencias código, que confieran a las señales resultantes esta propiedad. Normalmente, el ancho de banda de las secuencias utilizadas es muy superior al de la señal de datos, con lo que produce un ensanchamiento espectral. Al este método de codificar las señales se le conoce como DS-CDMA (*Direct Secuense CDMA*).

La modulación utilizada normalmente en los sistemas DS-CDMA es la PSK. Dada una señal de información digital $x(t)$ con período de bit igual a T_b, al multiplicarla por una secuencia código $c(t)$ y modularla en PSK binaria, la señal que se obtiene es de la forma:

$$s(t) = \sqrt{2P}x(t)c(t)\cos(2\pi f_0 t) \qquad (2.2)$$

donde P es la potencia transmitida y f_0 la frecuencia portadora. La **Figura 9** muestra el diagrama de bloques de la generación de la señal a transmitir.

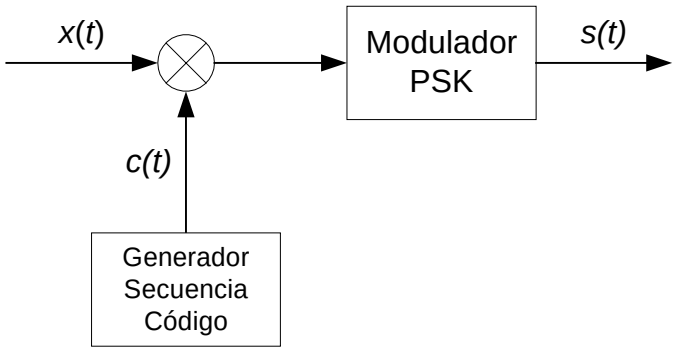

Figura 9. Proceso de transmisión CDMA

Por otro lado, si la secuencia código es una señal digital, a los bits de esta señal se les suele denotar por *chips*, cuyo período es T_c. Normalmente se cumple que $T_c<T_b$ y por tanto el ancho de banda de la señal $s(t)$ se incrementa respecto al de $x(t)$ en un factor llamado ganancia de procesado, o simplemente G_p, que se define por:

$$G_p = \frac{T_b}{T_c} \qquad (2.3)$$

La Figura 10 muestra un ejemplo de las señales que se generan en una transmisión CDMA usando una modulación BPSK y un pulso conformador rectangular sin retorno a cero.

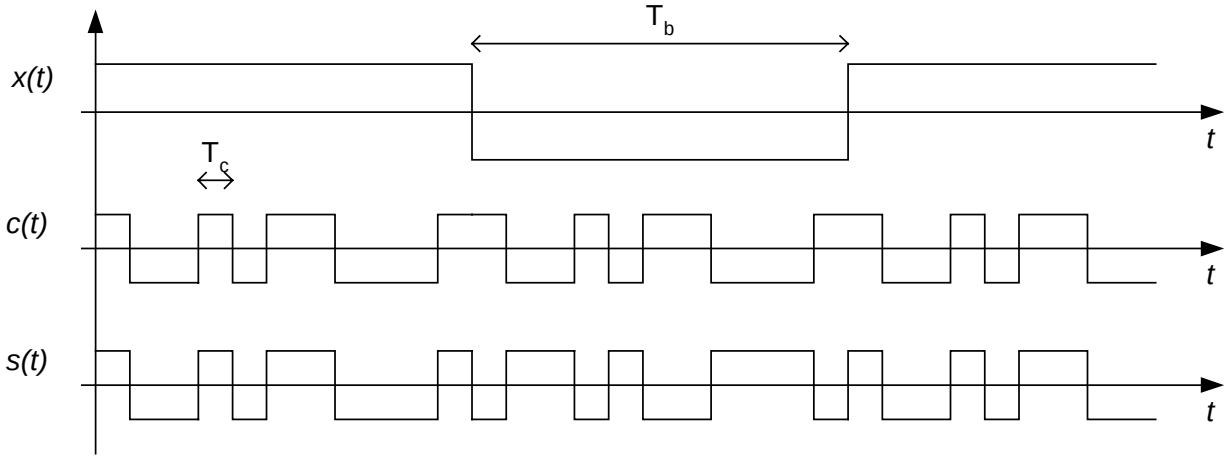

Figura 10. Ejemplo de señales CDMA

Podemos observar en la Figura 11 de manera gráfica el ensanchamiento espectral que se produce entre la señal de información x(t) y la señal que se transmite por el canal s(t). La relación entre un ancho de banda y otro se llama factor de ensanchamiento espectral que se corresponde con la ganancia de procesado.

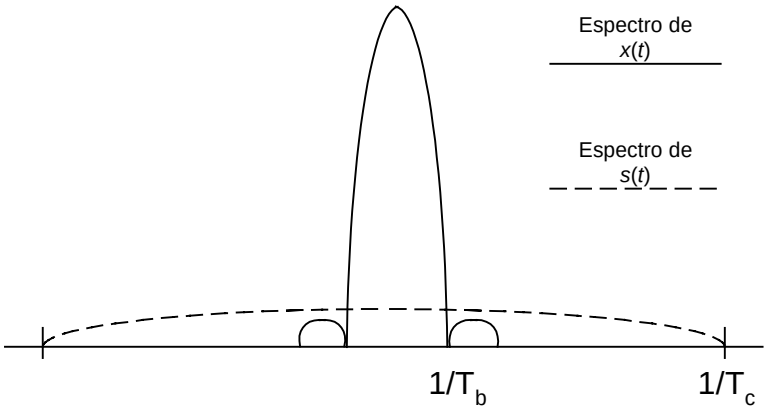

Figura 11. Ensanchamiento espectral en CDMA

Para recuperar las señales originales, el receptor debe multiplicar la señal recibida de nuevo por la misma secuencia código utilizada en transmisión y sumar la señal resultante en un período de bit. La Figura 12 muestra el esquema básico que debe aplicarse.

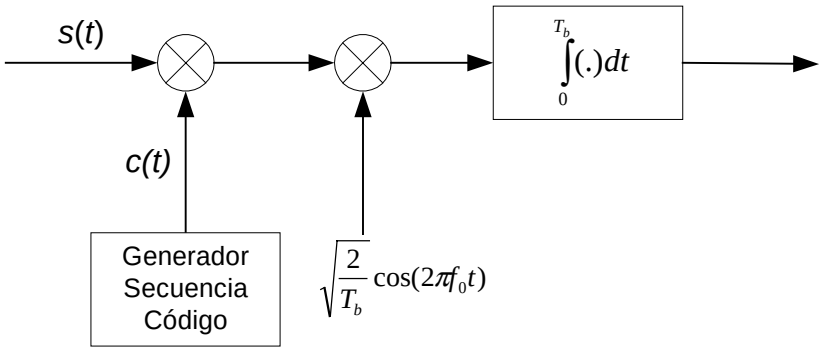

Figura 12. Proceso básico de recuperación de la información en CDMA

En principio, si la ortogonalidad de las secuencias código es perfecta, el receptor es capaz de separar sin ningún error cada una de las señales transmitidas. En la práctica, bien porque los códigos no son totalmente ortogonales, bien porque la respuesta del canal es diferente para las señales de cada usuario, cada una de ellas interfiere al resto en alguna medida [13]. Este grado de interferencia puede estimarse del siguiente modo:

El receptor CDMA concentra el espectro de la señal recibida $s(t)$ en un ancho de banda $1/T_b$. Para la parte correspondiente a la señal interferente, puesto que el código no corresponde con la señal transmitida, el espectro no se concentrará y seguirá repartido en un ancho de banda $1/T_c$. Por tanto, si inicialmente las dos señales se recibían con la misma potencia, ahora la señal interferente afecta únicamente en el nuevo ancho de banda, de manera que se cumple que la relación señal a interferente es:

$$\left(\frac{C}{I}\right) = \frac{T_b}{T_C} = G_p \qquad (2.4)$$

Es decir, el ensanchamiento en frecuencia protege de las interferencias en un grado que depende directamente de la ganancia de procesado.

Si tenemos en cuenta los aspectos físicos de aplicación práctica de un sistema CDMA, como el modelo de propagación del canal, aparecen algunos aspectos que deben estudiarse con más detenimiento. Entre estos elementos, cabe destacar:

- Una señal CDMA será normalmente una señal de banda ancha. En canales de propagación tipo canal móvil, el ancho de banda de coherencia del canal suele ser menor que el de la señal transmitida. Por tanto, no todas las componentes de la señal observarán la misma respuesta del canal (canal selectivo en frecuencia). Este hecho, sin embargo, puede aprovecharse para mejorar la calidad de la transmisión, gracias a que se podrán discriminar los diferentes caminos de propagación. Existe un modelo de receptor, llamado RAKE (ver 2.2.3), que saca partido de este aspecto del sistema.

- La elección de las secuencias código que se asignan a cada usuario puede afectar sustancialmente a la calidad del sistema de transmisión. Es por tanto necesario estudiar con detenimiento el tipo y las características de dichos códigos.

- Un elemento clave en la recuperación de las señales de información es el sincronismo del sistema. La copia local de la secuencia código debe estar perfectamente alineada, con la precisión de un período de *chip*, con la secuencia que se usó en transmisión.

Dado que el período de *chip* suele ser un intervalo relativamente pequeño, esta sincronización puede resultar tecnológicamente difícil en la práctica.

- Otro elemento crucial de los sistemas CDMA es el control de potencia. Es muy importante, para evitar que se produzca el llamado efecto *near-far*, que todos los usuarios lleguen en recepción con la misma potencia. Este efecto se produce cuando la señal de un usuario que está físicamente más cerca de la estación base llega a ésta con una potencia mucho mayor que los que están más lejos. Este hecho hace que la interferencia que produce un usuario en los demás pueda ser de un nivel tan elevado que evite por completo la detección de las señales correspondientes. Es crítico, por tanto, que todas las señales lleguen con la potencia justa, ni menos ni más que la necesaria para detectar correctamente cada una de ellas interfiriendo lo mínimo al resto de usuarios.

Nos detenemos ahora con más detalle en el análisis de cada uno de estos puntos.

2.2.3 Receptor RAKE

Cuando el ancho de banda de una señal es mucho menor que el ancho de banda de coherencia del canal por el que es transmitida, todas las componentes de la misma sufren la misma respuesta. En este caso se dice que el canal no es selectivo en frecuencia. En este tipo de canales, suelen utilizarse técnicas de diversidad para evitar que un desvanecimiento profundo pueda cortar temporalmente la transmisión de una señal. Sin embargo, si se dispone de un ancho de banda de transmisión mucho mayor que el ancho de banda de coherencia del canal, puede aplicarse una técnica de diversidad en frecuencia, de modo que se transmite la misma información situada en diferentes frecuencias portadoras.

Es posible demostrar [14] que con la técnica de transmisión DS-CDMA, que realiza un ensanchamiento espectral de la señal de información, se pueden obtener los mismos resultados que con la técnica de diversidad en frecuencia. Es decir, podemos aprovechar la selectividad en frecuencia del canal para mejorar las prestaciones de la señal recibida CDMA.

Visto desde el punto de vista temporal, la menor duración del tiempo de señalización (tiempo de *chip*) permite poder tener mayor resolución para distinguir copias distintas de la misma señal con diferentes retardos de propagación. El receptor más utilizado que aprovecha estas características de la señal para mejorar su calidad es el llamado receptor RAKE.

Pueden encontrarse referencias en la literatura que describen con detalle el diseño del receptor RAKE, que fue propuesto por Price y Green en 1958 y que se comporta básicamente como un combinador óptimo de máxima ganancia, del tipo MRC (*Maximal Ratio Combining*). En la Figura 13 se muestra un esquema simplificado de este receptor.

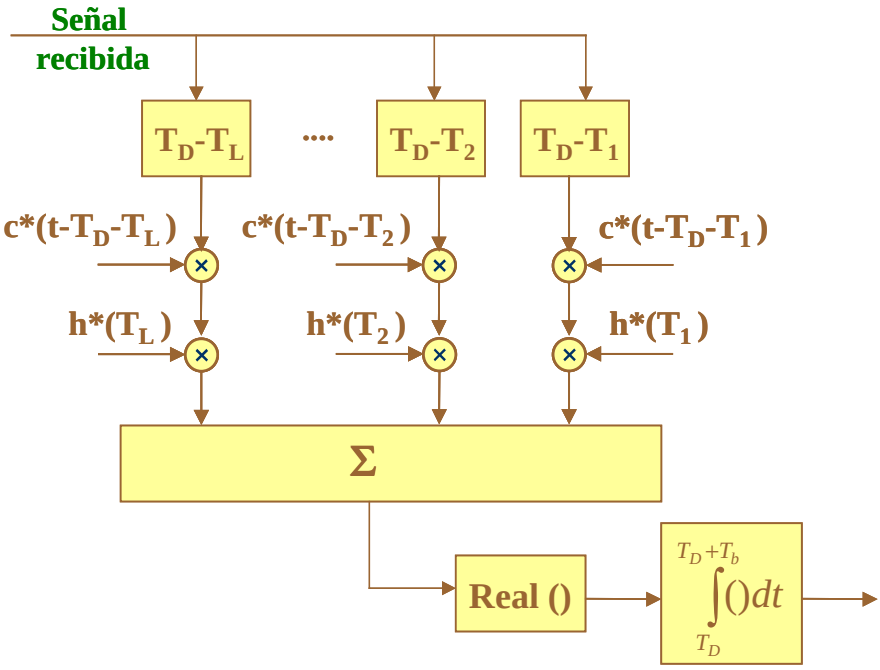

Figura 13. Estructura del receptor RAKE

Básicamente, la idea es que deben calcularse los coeficientes $h(t)$ y los retardos T_i de manera que el receptor sea capaz de sumar de forma coherente las contribuciones de los n caminos de propagación de la señal. El número de etapas, y por tanto el número de caminos que pueden resolverse, es directamente proporcional al ancho de banda disponible, de manera que cuanto mayor sea éste, mayor será la ganancia que se obtiene del receptor RAKE.

2.2.4 Secuencias código

Las secuencias código que se usan en los sistemas CDMA deben cumplir los siguientes requisitos:

- Deben tener naturaleza pseudoaleatoria. Esta propiedad asegura que la señales moduladas tengan apariencia de ruido.
- Deben presentar un buen comportamiento desde el punto de vista de su autocorrelación y de la correlación cruzada con otras secuencias.
- Deben ser sencillas de generar.
- Debe existir un número suficientemente grande de secuencias como para poder asignar a los usuarios las necesarias.

También es necesario observar el entorno de propagación en el que se usará el sistema CDMA. El canal radio tiene unas características específicas concretas que condicionan también la elección de los códigos. Es esencial por tanto encontrar familias de códigos que permitan separar adecuadamente a los usuarios en los entornos de propagación con canal radio móvil. Ello redunda en nuevas propiedades que deben cumplir los códigos:

Dado el conjunto de códigos usados en un sistema, cada uno de ellos debe ser fácilmente distinguible de una versión retardada de sí mismo.

Cada uno de los códigos debe ser fácilmente distinguible de cualquiera de los otros códigos.

Matemáticamente, estas dos últimas propiedades se traducen en el hecho de que la autocorrelación de todos los códigos debe tener valores muy pequeños, excepto para retardo nulo, y que la correlación cruzada de dos códigos diferentes debe tomar siempre valores pequeños.

Las dos principales familias de secuencias código utilizadas en los sistemas CDMA son las llamadas secuencias *m* y las secuencias de Gold.

2.2.4.1 Secuencias *m*

Las secuencias *m* son secuencias generadas por un registro binario de desplazamiento realimentado con una función lineal de su contenido. El término binario se refiere a que el contenido de los elementos del registro son bits. La realimentación se basa en funciones que realizan una suma módulo 2 (función OR-exclusiva). Dado un registro de *n* etapas, es posible tener 2^n estados diferentes del mismo. Puesto que el estado todo ceros no genera ningún cambio a medida que pasa el tiempo, es un estado que se llama degenerado que no se tiene en consideración. Por tanto, un registro de *n* etapas puede generar una secuencia periódica de período máximo 2^n-1.

El período real generado, sin embargo, depende de la lógica que se utilice en la realimentación. Esta lógica de realimentación se representa mediante un polinomio de grado *n* con coeficientes binarios, donde cada coeficiente distinto de cero denota una rama de realimentación. A este polinomio se le denomina polinomio característico. El diagrama de bloques que representa el polinomio x^5+x^2+1 se muestra a título de ejemplo en la Figura 14.

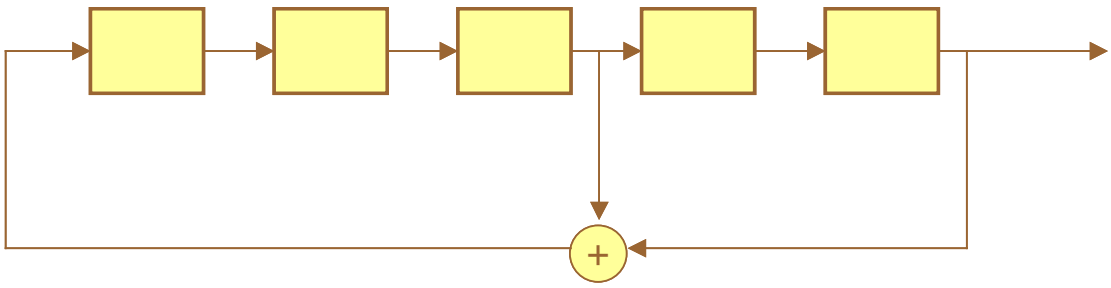

Figura 14. Registro de desplazamiento correspondiente al polinomio x^5+x^2+1

Para que un registro genere una secuencia periódica de longitud máxima es necesario que su polinomio característico sea primitivo. Para definir esta propiedad de los polinomios es necesario primero introducir la definición de otro concepto, el polinomio irreducible:

Se dice que un polinomio de grado *n* es irreducible cuando no es divisible por ningún polinomio de grado no nulo menor que *n.*

Entonces, un polinomio irreducible de grado *n* con coeficientes binarios es primitivo si y sólo si no divide a ningún polinomio de la forma x^m+1 para $m<2^n-1$.

Es matemáticamente posible saber cuántos polinomios de grado *n* son primitivos [15], aunque para saber exactamente cuáles son es necesario hacer la comprobación uno por uno.

Si el polinomio característico de un registro de desplazamiento es primitivo, se obtendrá siempre la misma secuencia de salida, independientemente del valor inicial con el que se carguen los biestables (siempre que no sean todo ceros). Esto es así debido que en la secuencia aparecen todos los 2^n-1 posibles estados diferentes.

A continuación se enumeran algunas de las propiedades más importantes que cumplen las secuencias *m*:

- Cualquiera de las 2^n-1 posibles rotaciones cíclicas de la secuencia puede ser generada cargando el valor inicial adecuado en el registro de desplazamiento.

- Toda secuencia *m* satisface la lógica de recurrencia definida por su polinomio característico.

- En un período completo de cualquier secuencia *m* hay exactamente 2^{n-1} unos y $2^{n-1}-1$ ceros.

- Si una secuencia *m* cualquiera se suma bit a bit (módulo 2) con cualquier rotación cíclica de ella misma, el resultado es otra rotación cíclica de la misma secuencia *m*.

- La autocorrelación de las secuencias *m* sólo toma dos valores distintos: 2^n-1 para desplazamiento nulo, y -1 para cualquier otro desplazamiento.

La función de correlación cruzada de dos secuencias *m* depende de las secuencias consideradas. Sin embargo, se puede encontrar experimentalmente que algunas parejas de secuencias tienen correlaciones periódicas con sólo tres valores diferentes, y además estos valores son pequeños en valor absoluto en comparación con los que aparecen en otras correlaciones. Por el interés que estas secuencias pueden tener en los sistemas CDMA, se investigó la manera de obtener de manera sistemática los polinomios característicos que generasen este tipo de secuencias [16]. Se obtuvieron las siguientes conclusiones:

- Para cualquier valor entero *n* que no es múltiplo de 4, existen parejas de secuencias *m* de período común 2^n-1 para las que la correlación periódica toma solamente tres valores distintos, que son:

$$\left\{-1, 2^{\left\lfloor \frac{n+2}{2} \right\rfloor}-1, -\left(2^{\left\lfloor \frac{n+2}{2} \right\rfloor}+1\right)\right\} \qquad (2.5)$$

donde $\lfloor x \rfloor$ denota el mayor entero menor o igual que *x*. A las parejas que cumplen esta propiedad se les llama parejas preferidas.

- Si *n* es un número múltiplo de 4, existen parejas de secuencias *m* que presentan una correlación periódica que toma sólo 4 valores distintos, cuyo valor absoluto es menor o igual que:

$$2^{\frac{n+2}{2}}-1 \qquad (2.6)$$

- Cuando *n* es par, cualquier secuencia *m* y su copia invertida presentan una correlación periódica con múltiples valores diferentes, pero todos ellos de valor absoluto menor o igual a la misma cota anterior.

Como podemos observar, estas propiedades aplican únicamente a parejas de secuencias. Sin embargo, en un sistema comercial CDMA son necesarias muchas más secuencias código que tengan unas buenas propiedades de correlación. Desafortunadamente, los intentos de ampliar el número de secuencias *m* que tienen buenas propiedades conlleva un aumento de los valores mínimos de sus correlaciones cruzadas. Como consecuencia de este comportamiento, aparece la necesidad de obtener secuencias que, aun teniendo peor comportamiento en cuanto a autocorrelación, sean conjuntos mucho más grandes que mantengan buenas propiedades de correlación cruzada. Una de estas familias son las secuencias de Gold.

2.2.4.2 Secuencias de Gold

Las secuencias de Gold se definen a partir del producto de los polinomios característicos de dos secuencias *m*. En efecto, sean $g(x)$ y $h(x)$ una pareja preferida de polinomios primitivos de grado *n*. El registro de desplazamiento definido mediante el polinomio característico resultado del producto $g(x)\times h(x)$ generará $N+2$ secuencias diferentes de período $N=2^n-1$. Esta familia de secuencias se conoce como una familia de códigos de Gold.

Los códigos de Gold también pueden obtenerse como la suma módulo 2 de las dos secuencias *m* generadas por $g(x)$ y $h(x)$.

Los códigos de Gold presentan una autocorrelación con tres valores diferentes definidos por la expresión (5), mientras que sus correlaciones cruzadas tienen las mismas propiedades que las secuencias *m*. Nótese que estas propiedades las cumplen $N=2^n-1$ secuencias diferentes, con lo que el problema de la cantidad de secuencias disponibles se ve reducido en gran medida.

2.2.5 Sincronismo

El objetivo primordial del sistema de sincronismo es encargarse de alinear la señal recibida con la copia local de la secuencia código que genera el receptor.

Como en cualquier sistema de adquisición de sincronismo, este mecanismo tiene dos fases diferenciadas, el ajuste grueso, o fase de adquisición propiamente dicha, y el ajuste fino o fase de seguimiento.

Se distinguen dos grandes técnicas para la adquisición del sincronismo, la búsqueda serie y la búsqueda en paralelo [17].

La dificultad técnica del sincronismo estriba en la corta duración del período de *chip* de las secuencias código. En un sistema CDMA es necesario que el sincronismo funcione correctamente al menos a nivel de chip.

2.2.6 Control de potencia

Uno de los aspectos más importantes de un sistema CDMA es el control de potencia. Puesto que la capacidad del sistema está limitada por la interferencia que los propios usuarios generan, es crucial que todas las señales de los usuarios lleguen al receptor con la potencia justa.

Recuérdese que un usuario cuya señal llega al receptor con una potencia excesiva está degradando las comunicaciones del resto de usuarios del sistema, tanto de su propia célula como de las células vecinas.

Básicamente existen dos mecanismos para realizar el control de potencia en los terminales móviles:

- Control de potencia en lazo abierto.
- Control de potencia en lazo cerrado.

Vamos a describir en detalle ambas estrategias.

2.2.6.1 Control de potencia en lazo abierto

Consiste en utilizar un Control Automático de Ganancia (CAG) mediante el cual se modifica la ganancia de los amplificadores del transmisor a partir de la medida de la potencia recibida proveniente de la estación base. Esta potencia recibida viene de un canal piloto cuya potencia de transmisión es conocida a priori. De este modo el móvil puede estimar las pérdidas del canal y calcular cuál debe ser la potencia con la que debe transmitir para que la estación base le reciba con la potencia adecuada.

Este mecanismo, sin embargo, tan sólo garantiza que la potencia recibida por la base sea la deseada en valor promedio, pero no en cada realización concreta. Esto ocurre por diversos factores. En primer lugar, la técnica se basa en el hecho de que los canales ascendente y descendente son simétricos, es decir que las pérdidas de propagación son las mismas en ambos sentidos. Esto no siempre es cierto, y menos aún cuando las frecuencias centrales de los canales ascendente y descendente son muy diferentes. Por otro lado, las variaciones rápidas de la respuesta del canal, especialmente presentes en entornos de canal móvil, no pueden ser compensadas por esta técnica.

Para mejorar las prestaciones de esta técnica, aparece la idea del control de potencia en lazo cerrado.

2.2.6.2 Control de potencia en lazo cerrado

Con esta técnica, la estación base va realizando medidas de la potencia que recibe del móvil, y le va enviando una serie de comandos para que este vaya subiendo o bajando la potencia de transmisión según corresponda.

Usualmente, estos mensajes de control consisten en un único bit, cuyo valor indica si el móvil debe aumentar o disminuir su potencia de transmisión en un escalón fijo predefinido de Δ dB.

Con este mecanismo podemos compensar incluso las variaciones rápidas de la respuesta del canal, consiguiendo, idealmente, que la potencia instantánea recibida por la base sea la correcta en todos los instantes de tiempo.

En un sistema donde todos los usuarios tienen los mismos requisitos de calidad y que transmiten con la misma velocidad (misma ganancia de procesado), el escenario óptimo consiste en aquel en el que todos los usuarios llegan a la base con la misma potencia [18].

En caso de que no todos los usuarios transmitan con la misma velocidad, esta consideración ya no es cierta. Se debe recordar que en un sistema digital la calidad de transmisión viene determinada por la probabilidad de error en el bit, y ésta a su vez viene condicionada por la relación energía por bit a ruido e interferencias (E_b/N_0). Si los usuarios que transmiten a diferentes velocidades deben tener la misma calidad, deberán tener la misma $E_b/N_{0,}$ con lo que la potencia instantánea recibida de todos ellos no deberá ser la misma. En particular, esta potencia deberá ir en proporción a la potencia interferente total que recibe la estación base.

2.2.6.3 Hipótesis Gaussiana para el modelado de las interferencias

Una vez descritos los sistemas de control de potencia, y suponiendo que se hace un uso adecuado de los mismos, resulta útil poder obtener expresiones analíticas de la calidad de transmisión de un sistema digital basado en acceso CDMA. El sistema estudiado en la presente tesis está basado en transmisiones por canal común en modo paquete. En un sistema de este tipo, la calidad de transmisión vendrá dada por el caudal efectivo o *thoughput*, medido en bits correctamente recibidos por unidad de tiempo. Para este cálculo se deberá tener en cuenta que cuando cualquiera de los bits de un paquete es erróneo, todos los bits del mismo son descartados. Se pueden encontrar en la literatura estudios que abordan el problema del cálculo de este *throughput* [19], [20].

Una de las ideas intuitivas consideradas es suponer que, si el número de contribuciones a la interferencia que sufre la señal de un usuario es grande, todas ellas son independientes y han sido generadas con secuencias pseudoaleatorias, cada una de estas contribuciones puede verse como un ruido blanco gaussiano respecto a las demás. El teorema central del límite asegura que cuando sumamos un número suficientemente grande de variables aleatorias independientes, sean cuales sean sus distribuciones estadísticas particulares, la variable aleatoria resultante tiende a ser gaussiana.

Si utilizamos la aproximación de que todas las señales se ven como ruido gaussiano para las demás, y que son independientes entre sí, podemos considerar que las señales espurias que ve un cierto usuario son un ruido gaussiano cuya potencia total es la suma aritmética de las potencias de las señales de todos los usuarios, más el ruido térmico. Esta aproximación es la llamada Hipótesis Gaussiana para la evaluación de las interferencias.

Haciendo uso de ella, para usuarios que utilizan control de potencia en lazo cerrado y que llegan a la estación base con la misma potencia, se demuestra en [21] que, si el sistema está sincronizado únicamente a nivel de bit y se puede despreciar el ruido térmico, se cumple para un usuario que:

$$\frac{E_b}{N_0} = \frac{3G_p}{2(K-1)} \tag{2.7}$$

donde G_p es la ganancia de procesado que usan todos los usuarios, y K es el número total de usuarios en el sistema que transmiten de modo simultáneo.

Recordemos que si se utiliza una modulación del tipo BPSK, la probabilidad de error en el bit, P_b, vendrá dada por:

$$P_b = \frac{1}{2} erfc\left(\sqrt{\frac{E_b}{N_0}}\right) \tag{2.8}$$

Con este valor es posible determinar a su vez la probabilidad de tener un bloque de bits erróneo. Esto ocurrirá cuando haya al menos un bit erróneo en el paquete de bits. A esta probabilidad la llamaremos BLER (*Block Error Ratio*). Si los paquetes son de *L* bits y suponemos que la probabilidad de error en el bit se mantiene constante durante la transmisión del paquete, el BLER será:

$$BLER = 1 - (1 - P_b)^L \tag{2.9}$$

En principio, si el número de usuarios es pequeño, esta hipótesis no es aplicable. Sin embargo, en este caso el nivel de interferencia será pequeño, la probabilidad de error también lo será, con lo que puede no resultar relevante la diferencia de valor entre el caso real y el que se obtiene de hacer uso de la hipótesis gaussiana. Podemos concluir entonces que es posible utilizar esta hipótesis de manera razonable en una amplia gama de entornos de trabajo.

De hecho, esta hipótesis ha sido y es ampliamente usada en todos los terrenos tanto de investigación como de aplicación por su equilibrio entre simplicidad y buena precisión. En el presente trabajo será utilizada extensivamente para la evaluación de las probabilidades de error en las transmisiones CDMA.

2.2.6.4 Probabilidad de error en los paquetes

En las transmisiones en modo paquete, la calidad de la transmisión viene definida por la probabilidad de error en cada uno de los bloques o paquetes de bits. Hasta el momento, las expresiones que se han desarrollado en la literatura describen la probabilidad de error en cada uno de los bits. Sin embargo, siendo rigurosos, este valor no puede aplicarse directamente al cálculo de la probabilidad de error en el paquete, puesto que sólo estamos utilizando un valor medio, sin tener en cuenta la estadística de los errores.

Además, todas las expresiones existentes presuponen que todos los usuarios transmiten sus paquetes haciendo uso de la misma técnica de control de potencia. En un sistema real, ocurre en muchas ocasiones que usuarios de diferentes tipos transmiten de forma simultánea usando diferentes modos de control de potencia: algunos de ellos estarán transmitiendo los primeros paquetes de una conexión, probablemente estableciendo la misma o solicitando recursos, y tan sólo podrán realizar un control de potencia en lazo abierto, mientas que por otro lado habrá usuarios que estarán en mitad de una transmisión larga, y les será posible tener un canal descendente de control con el que realizar un control de potencia en lazo cerrado.

En estas situaciones de tráfico heterogéneo, es necesario establecer expresiones de las funciones de densidad de probabilidad (de ahora en adelante *pdf*) de la potencia total recibida, tanto del usuario útil como de los interferentes. Este cálculo constituye una contribución original y novedosa que fue publicada en [22].

Sea un sistema de transmisión por paquetes donde dichos paquetes son de *L* bits y se transmite la información usando secuencias pseudoaleatorias con una ganancia de procesado G_p. Los slots de tiempo de transmisión son de tales que cada paquete de *L* bits se transmite en uno de ellos. Se asume que el tiempo de coherencia del canal [23] es mayor que la duración de los slots, de manera que la respuesta del canal permanece constante durante toda la duración de la transmisión de cada paquete. Por tanto, todos los bits de un paquete sufren la misma atenuación del canal. Analizaremos cuatro escenarios diferentes.

2.2.6.4.1 Todos los usuarios con control de potencia en lazo cerrado

Consideramos un grupo de m+1 usuarios activos. Por tanto, cada usuario tendrá m usuarios interferentes. Todos ellos aplican un control de potencia en lazo cerrado ideal, de manera que todos ellos llegan a la estación base con la misma potencia. En esta situación, tal y como hemos justificado en 2.2.6.3, utilizaremos la hipótesis gaussiana para el modelado de las interferencias, de manera que si despreciamos el ruido térmico y llamamos γ a la relación señal a interferente (SIR), podemos escribir que:

$$\gamma = \frac{3G_p}{2m} \tag{2.10}$$

Asumiendo una modulación BPSK con desmodulación coherente ideal, la probabilidad de error en el bit es:

$$P_b(\gamma) = \frac{1}{2} erfc\left(\sqrt{\gamma}\right) \tag{2.11}$$

En caso de utilizar un código de canal (ver 2.2.7) capaz de corregir hasta t errores en un paquete de L bits, la probabilidad de tener un paquete correcto, que llamaremos $P_c(\gamma)$, vendrá dada por:

$$P_c(\gamma) = \sum_{n=0}^{t} \binom{L}{n} P_b(\gamma)^n \left(1 - P_b(\gamma)\right)^{L-n} \tag{2.12}$$

Finalmente, la probabilidad de error en el paquete, que llamaremos *BLER*, será:

$$BLER = 1 - P_c(\gamma) \tag{2.13}$$

En la expresión (2.12) el valor de t debe corresponder a la capacidad correctora del código de canal utilizado. Este código puede ser de cualquier tipo, tanto bloque como convolucional o turbo código. Por tanto, este cálculo es genérico e independiente del código empleado.

2.2.6.4.2 Todos los usuarios con control de potencia en lazo abierto

Consideramos ahora un conjunto de k+1 usuarios que aplican un control de potencia en lazo abierto. Tendremos por tanto para cada uno de ellos k usuarios interferentes. En este caso, únicamente la potencia media recibida por la base, que llamaremos σ, será la misma para todos los usuarios. Supondremos sin embargo que los desvanecimientos rápidos del canal no pueden ser compensados, y por tanto la potencia instantánea recibida de cada usuario es una variable aleatoria. Modelaremos esta variable a partir de una atenuación del tipo Rayleigh, típica de los entornos de canal móvil donde la señal recibida es la contribución de un número grande de señales independientes. La media de esta variable aleatoria será σ. En estas condiciones, podemos escribir el valor de γ como:

$$\gamma = \frac{3G_p}{2} \frac{\alpha_0^2}{\sum_{i=1}^{k} \alpha_i^2} = \frac{3G_p}{2} \frac{X}{Y} \tag{2.14}$$

Donde hemos llamado $X=\alpha_0^2$ a la potencia recibida por el usuario útil. La interferencia total que observa el usuario útil es $Y=\alpha_1^2+\alpha_2^2+...+\alpha_k^2$. Tanto X como Y son variables aleatorias cuyas funciones densidad de probabilidad son:

$$f_X(x)=\frac{1}{\sigma}e^{-\frac{x}{\sigma}} \qquad x>0 \tag{2.15}$$

$$f_Y(y)=\frac{y^{k-1}}{\sigma^k(k-1)!}e^{-\frac{y}{\sigma}} \qquad y>0 \tag{2.16}$$

Es evidente que ahora γ es una variable aleatoria, que deberá evaluarse matemáticamente para poder derivar el valor de la probabilidad de error en el bloque. Puesto que es una función de dos variables aleatorias (*X, Y*), en el cálculo de su *pdf* deberemos tener en cuenta las *pdf* de estas variables. Como paso intermedio de cálculo, vamos a derivar la expresión de su función de distribución (que llamaremos *PDF*), que nos da la probabilidad de que la variable tome valores inferiores o iguales a cada valor posible que puede tomar. En este caso, la PDF de γ viene dada por:

$$\begin{aligned} F_\gamma(\gamma)=P\left(\frac{3G_p}{2}\frac{X}{Y}\le\gamma\right)=\int_0^\infty e^{-\frac{x}{\sigma}}\int_{\frac{3G_p x}{2\gamma}}^\infty \frac{1}{\sigma^{k+1}}\frac{y^{k-1}}{(k-1)!}e^{-\frac{y}{\sigma}}dydx= \\ =1-\left(\frac{2\gamma}{3G_p}+1\right)^{-k} \qquad \gamma>0 \end{aligned} \tag{2.17}$$

Matemáticamente, la *pdf* de una variable aleatoria se obtiene a partir de la derivada de la *PDF*, con lo que obtenemos que la *pdf* de γ es:

$$f_\gamma(\gamma)=\frac{2k}{3G_p}\left(\frac{2\gamma}{3G_p}+1\right)^{-k-1} \qquad \gamma>0 \tag{2.18}$$

Como reseña interesante, podemos observar que esta distribución es muy similar a la distribución de Pareto. En concreto, se comprueba fácilmente que la variable aleatoria:

$$\beta=\frac{2\gamma}{3G_p}+1 \tag{2.19}$$

sigue exactamente una distribución de Pareto de parámetros (1,*k*). Al igual que ocurre con esta variable, su esperanza es infinita para *k*=1, y su varianza también tanto para *k*=1 como para *k*=2. La variable β, además, tiene un significado físico: es el cociente entre la potencia total recibida a la entra del receptor (contando tanto la potencia útil como la interferente) y la potencia total interferente.

2.2.6.4.3 Usuario útil con control en lazo abierto e interferencia mixta

Consideremos un usuario que transmite utilizando un control de potencia en lazo abierto, junto con dos grupos de usuarios interferentes: *k* usuarios que también utilizan un control de potencia en lazo abierto y *m* usuarios que aplican un control de potencia en lazo cerrado ideal. De nuevo, la atenuación que sufren los usuarios con control en lazo abierto sigue una

estadística tipo Rayleigh, y es independiente entre ellos. En estas condiciones, y siguiendo la misma notación del punto 2.2.6.4.2, tendremos que la relación señal a interferente será:

$$\gamma = \frac{3G_p}{2}\frac{\alpha_0^2}{m\sigma + \sum_{i=1}^{k}\alpha_i^2} = \frac{3G_p}{2}\frac{X}{(m\sigma + Y)} \tag{2.20}$$

donde de nuevo se asume que la potencia media recibida de todos los usuarios con control en lazo abierto, σ, es la misma para todos ellos. Esta hipótesis será válida siempre que todos los usuarios tengan los mismos requisitos de calidad y la misma velocidad de transmisión. Por su parte, la PDF de γ vendrá dada por:

$$F_\gamma(\gamma) = P\left(\frac{3G_p}{2}\frac{X}{Y + m\sigma} \le \gamma\right) = P\left(\frac{3G_p}{2\gamma}X - m\sigma \le Y\right) \tag{2.21}$$

Para poder simplificar la notación, haremos un cambio de variable. Llamaremos *a* a la siguiente expresión:

$$a = \frac{3G_p}{2\gamma} \tag{2.22}$$

En este punto debemos distinguir entre dos intervalos para la variable *x*, que son $x \ge m\sigma/a$ y $x < m\sigma/a$, Podemos desarrollar la expresión (2.21) como sigue:

$$\begin{aligned} P(aX - m\sigma \le Y) &= \int_0^{m\sigma/a} e^{-\frac{x}{\sigma}} \int_0^{\infty} \frac{1}{\sigma^{k+1}}\frac{y^{k-1}}{(k-1)!}e^{-\frac{y}{\sigma}}dydx + \\ &+ \int_{m\sigma/a}^{\infty} e^{-\frac{x}{\sigma}} \int_{ax-m\sigma}^{\infty} \frac{1}{\sigma^{k+1}}\frac{y^{k-1}}{(k-1)!}e^{-\frac{y}{\sigma}}dydx = \\ &= 1 - e^{-\frac{m}{a}} + e^{-\frac{m}{a}}\left(1 - \left(\frac{a}{1+a}\right)^k\right) = 1 - \left(\frac{2\gamma}{3G_p} + 1\right)^{-k} e^{-\frac{2m}{3G_p}\gamma} \quad \gamma > 0 \end{aligned} \tag{2.23}$$

Y por tanto la función que buscamos, la *pdf* de γ, es:

$$f_\gamma(\gamma) = \frac{2}{3G_p}\left(\frac{2\gamma}{3G_p} + 1\right)^{-k}\left[k\left(\frac{2\gamma}{3G_p} + 1\right)^{-1} + m\right]e^{-\frac{2m}{3G_p}\gamma} \quad \gamma > 0 \tag{2.24}$$

Nótese que en realidad la expresión (2.18) es un caso particular de esta última expresión cuando m=0.

2.2.6.4.4 Usuario útil con control en lazo cerrado e interferencia mixta

Consideramos un usuario que aplica un control en lazo cerrado ideal, y que por tanto llega a la base con potencia constante. De nuevo tenemos dos grupos de usuarios interferentes: *k* usuarios que aplican un control en lazo abierto, y *m* usuarios que aplican el control en lazo cerrado. De nuevo, la atenuación que sufren los usuarios con control en lazo abierto sigue una

estadística tipo Rayleigh, y es independiente entre ellos. Ahora, y manteniendo la notación de los puntos anteriores, la relación señal a interferente γ se puede expresar como:

$$\gamma = \frac{3G_p}{2}\frac{\sigma}{m\sigma + \sum_{i=1}^{k}\alpha_i^2} = \frac{3G_p}{2}\frac{\sigma}{(m\sigma + Y)} \tag{2.25}$$

Por tanto, la función de distribución de γ es:

$$F_\gamma(\gamma) = P\left(\frac{3G_p}{2}\frac{\sigma}{Y + m\sigma} \le \gamma\right) = \int_{\left(\frac{3G_p}{2\gamma} - m\right)\sigma}^{\infty} \frac{y^{k-1}}{\sigma^k (k-1)!} e^{-\frac{y}{\sigma}} dy = \tag{2.26}$$
$$= e^{-\left(\frac{3G_p}{2\gamma} - m\right)} \sum_{n=0}^{k-1} \frac{1}{(k-n-1)!}\left(\frac{3G_p}{2\gamma} - m\right)^{k-1-n} \quad 0 < \gamma < \frac{3G_p}{2m} \quad k \ge 1$$

y por tanto su función densidad de probabilidad es:

$$f_\gamma(\gamma) = \frac{3G_p}{2\gamma^2 (k-1)!}\left(\frac{3G_p}{2\gamma} - m\right)^{k-1} e^{-\left(\frac{3G_p}{2\gamma} - m\right)} \tag{2.27}$$
$$0 < \gamma < \frac{3G_p}{2m} \quad k \ge 1$$

Nótese que $f_\gamma(\gamma)=0$ para $\gamma > 3G_p/2m$, lo que significa que la relación señal a interferente está acotada por el valor que corresponde al caso en el que todos los usuarios aplican el control ideal de potencia, que constituye el caso más ideal posible.

2.2.6.4.5 Resultados numéricos

Se han realizado los cálculos de la probabilidad de error en el bloque, *BLER*, para diferentes casos de los mostrados desde el punto 2.2.6.4.1 hasta el punto 2.2.6.4.4. En las siguientes figuras se muestran los resultados obtenidos.

Puesto que γ es una variable aleatoria, el valor *BLER* debe calcularse según la expresión:

$$BLER = 1 - \int_0^\infty P_c(\gamma) f_\gamma(\gamma) d\gamma \tag{2.28}$$

donde $P_c(\gamma)$ se define en (12) y $f_\gamma(\gamma)$ es la que corresponde con las expresiones (2.18), (2.24) y (2.27). Los valores mostrados se han calculado para G_p=64, L=640 bits y con un código de canal cuya capacidad de corrección es de t=3 bits por paquete.

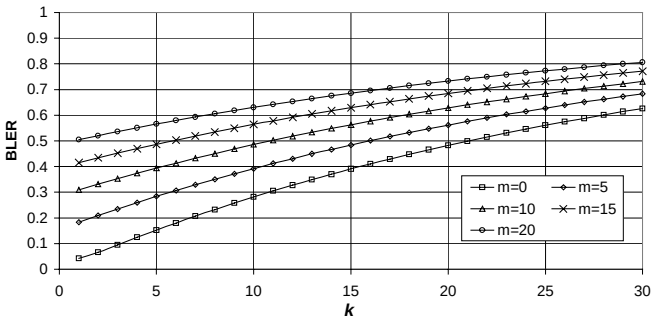

Figura 15. Probabilidad de error en el bloque para los casos de 2.2.6.4.2 (*m*=0) y 2.2.6.4.3

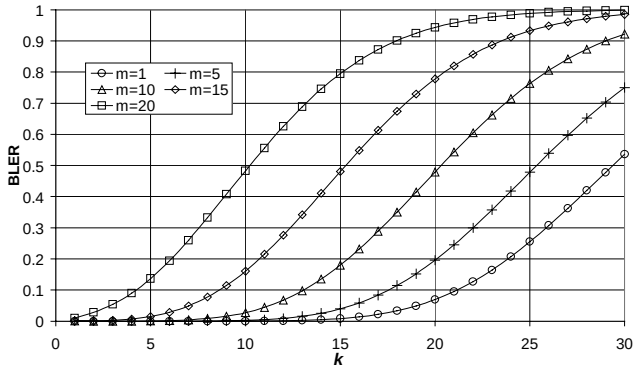

Figura 16. Probabilidad de error en el bloque para el caso de 2.2.6.4.4

Nótese que el *BLER* en el caso correspondiente al punto 2.2.6.4.3 crece más suavemente que el correspondiente al caso 2.2.6.4.4. En efecto, para valores pequeños de *k* el valor del *BLER* es menor para el caso del usuario útil con control en lazo cerrado. A medida que el valor de *k* crece, es el caso con control en lazo abierto el que va presentando un mejor valor del BLER. Este comportamiento se explica por el hecho de que, cuando el nivel de interferencia total es alto, resulta más conveniente dejar que un usuario pueda tener realizaciones diferentes de su potencia recibida, en lugar de mantener este valor constante de manera estricta. Cuando se aplica un control en lazo abierto, la naturaleza Rayleigh de los desvanecimientos puede permitir que algunos paquetes lleguen al receptor con una potencia mayor que la media, gracias a la contribución constructiva de las componentes de señal multicamino de la señal transmitida. De este modo, la probabilidad de que un paquete llegue correctamente puede resultar beneficiada.

2.2.7 Codificación de canal

En los sistemas CDMA puede incorporarse codificación de canal sin penalización en cuanto a ancho de banda. Esto es así gracias a que el ancho de banda ocupado es mucho mayor que el estrictamente necesario para transmitir la información de usuario.

Supongamos que tenemos un sistema sin codificación. La potencia interferente total podrá escribirse como:

$$P_I = N_0 B = N_0 \frac{1}{T_c} \tag{2.29}$$

donde N_0 es la densidad espectral de las interferencias y T_c el tiempo de chip. Al mismo tiempo, la energía por bit de información puede expresarse como $E_b=PT_b$, donde P es la potencia de señal útil recibida y T_b el tiempo de bit. Entonces, la relación señal a interferente podrá escribirse como:

$$\frac{P}{P_I}=\frac{E_b\dfrac{1}{T_b}}{N_0\dfrac{1}{T_c}}=\frac{E_b/N_0}{T_b/T_C}=\frac{E_b/N_0}{G_p} \tag{2.30}$$

En el caso de incorporar un código de canal con una cierta tasa de redundancia r, parte de la energía transmitida no transporta información útil, sino la redundancia, y puede definirse la energía por símbolo transmitido, incluyendo esta redundancia, como $E_s=E_b\cdot r$. Estos símbolos se transmiten en un tiempo $T_s=T_b\cdot r$. La potencia interferente se mantiene sin variación, puesto que las señales transmitidas con otros códigos se mantienen ensanchadas espectralmente. Por tanto, puede escribirse de nuevo la relación señal a interferente como:

$$\frac{P}{P_I}=\frac{E_s\dfrac{1}{T_s}}{N_0\dfrac{1}{T_c}}=\frac{E_s/N_0}{T_s/T_c}=\frac{E_b/N_0}{G_p} \tag{2.31}$$

que coincide con lo obtenido en (2.30) para el caso de no codificar. Para conseguir una misma calidad de transmisión, es decir, una misma probabilidad de error en el bit, es necesario que la E_b/N_0 sea mayor cuando el sistema no usa codificación. Por tanto, la potencia transmitida podrá ser menor en el caso codificado y la interferencia del sistema también. Todo ello redundará en una mayor capacidad del sistema, medida en términos de número de usuarios simultáneos, cuando se utiliza una cierta codificación.

2.3 Protocolos de acceso al medio (MAC)

En los sistemas de comunicaciones, es necesario en ocasiones que varios usuarios independientes compartan un medio común de transmisión. Para gestionar el uso de este recurso común es necesario establecer una estrategia de gestión o protocolo de acceso.

El protocolo de acceso al medio, en inglés *Medium Access Control Protocol* o simplemente MAC, es el encargado de gestionar cómo y cuándo cada uno de los usuarios de un sistema puede utilizar el medio común de transmisión para enviar su información.

La definición de un protocolo MAC eficiente es especialmente crucial en los futuros sistemas de comunicaciones móviles de tercera generación, en los que se pretende poder garantizar una cierta calidad de servicio haciendo uso de transmisiones en modo paquete. Los recursos de transmisión radio son siempre escasos y por tanto su valor es elevado. La gestión de las transmisiones en modo paquete implica en sí mismo una cierta complejidad. A cambio, el aprovechamiento de los recursos puede ser mucho más alto cuando se trabaja con fuentes de tráfico que generan información de tasa variable o a ráfagas.

Otro requisito que deben cumplir siempre los protocolos MAC es la ecuanimidad en el trato de los usuarios del mismo tipo, y en su caso la capacidad de establecer prioridades para aquellos que tengan unos determinados requisitos de transmisión.

Supongamos que el sistema de comunicaciones está formado por un conjunto de usuarios que necesitan transmitir información hacia una estación central o estación base a través de un medio común de transmisión. Podemos hacer las siguientes observaciones:

- El número de usuarios o conexiones que necesitan transmitir información simultáneamente por el medio de transmisión es variable. De hecho, en general es una variable aleatoria que puede tomar diferentes valores.

- Una vez un usuario o conexión inicia la transmisión de su información, el tiempo en el que ocupa los recursos del canal es también una variable aleatoria.

- Cada uno de los usuarios pueden necesitar transmitir su información en momentos no predecibles en el tiempo.

De estas observaciones podemos deducir que si la manera en la que los usuarios acceden al canal es totalmente libre, se producirá con cierta frecuencia un fenómeno llamado colisión. Una colisión se produce cuando dos o más usuarios intentan transmitir su información utilizando los mismos recursos de forma simultánea, de manera que ninguno de ellos puede hacerlo correctamente.

Los protocolos de acceso al medio son los encargados de organizar los accesos de manera que se eviten las colisiones, o en su defecto que los usuarios sepan qué deben hacer cuándo éstas aparecen.

Podemos concluir esta introducción recalcando que los futuros sistemas de comunicaciones móviles requerirán de protocolos MAC potentes y flexibles, que sean capaces de gestionar diferentes tipos de aplicaciones multimedia con características muy diversas. Hasta el momento, una gran cantidad de esfuerzo de investigación se ha dedicado a la definición y estudio de protocolos de acceso al medio para diferentes ámbitos de las telecomunicaciones. En este sentido, la presente tesis doctoral pretende aportar una contribución novedosa en este campo.

2.3.1 Clasificación

Los protocolos de acceso al medio pueden clasificarse de diferentes maneras, según se atienda a unas u otras características de los mismos. Con respecto a su capacidad para evitar o gestionar las colisiones, se puede establecer la siguiente clasificación:

- *Protocolos Sin Contención:* Este tipo de protocolos consiguen eliminar por completo la existencia de colisiones. Para conseguirlo, pueden usar dos estrategias diferentes:

    - Asignación fija de recursos: Los recursos de transmisión del canal se asignan de modo estático a los usuarios, de manera que nunca dos de ellos comparten los mismos recursos. Si un usuario no tiene información para transmitir, sus recursos asignados quedan sin utilizar.

    - Asignación bajo demanda: Los recursos de transmisión se asignan a los usuarios tan sólo cuando lo requieren porque tienen información a transmitir. Sin embargo, cuando algunos recursos les han sido asignados, los utilizan en exclusiva, sin compartirlos con ningún otro usuario.

- *Protocolos Con Contención:* En estos protocolos se permite que se produzcan colisiones, y por tanto existe en ellos un cierto grado de aleatoriedad. Por consiguiente, el protocolo debe establecer algún método para resolver las colisiones. También es posible dividir estos protocolos en:

    - Acceso aleatorio con reserva: La transmisión de datos tiene dos fases. En una primera fase los usuarios acceden de modo aleatorio al canal solicitando permiso para transmitir. En esta fase pueden producirse colisiones. Estas colisiones son resueltas en la llamada fase de contención. Cuando se han resuelto las colisiones, a los usuarios se les asignan unos recursos en exclusividad (reserva) durante un cierto tiempo. En esta segunda fase ya no se producen colisiones.

    - Acceso aleatorio repetitivo: Todos los intentos de transmisión de los usuarios son susceptibles de generar una colisión. Cuando un usuario ha incurrido en una colisión, trata de resolverla a base repetir la transmisión de la información, siempre utilizando las reglas de resolución de colisiones que dicte el protocolo.

2.3.2 Estado del arte

Hasta la fecha, se han propuesto una gran variedad de protocolos de acceso diferentes. Sólo algunos de ellos han sido utilizados en sistemas comerciales de telecomunicaciones. Los primeros de ellos, ALOHA y Slotted-ALOHA, sí han sido ampliamente utilizados como protocolos de acceso aleatorio. Sin embargo, su bajo caudal efectivo (18% y 36% de utilización máxima del canal, respectivamente) y su inestabilidad potencial para cargas elevadas de tráfico han conducido a la aparición de algoritmos eficientes de resolución de colisiones, también llamados CRA (*Collision Resolution Algorithms*). Algunos de ellos se denominan algoritmos en árbol [24], que alcanzan rendimientos de hasta un 56% utilizando información de control ternaria (de la estación base hacia los usuarios) [25]. Algunos protocolos alcanzan rendimientos superiores gracias al uso de unas ranuras especiales de tiempo dedicadas al envío de peticiones de acceso. De entre ellos, el llamado *Announced Arrival Random Access Protocol* (AARA) [26] alcanza el mejor rendimiento de retardo y caudal efectivo (85% usando tres ranuras de acceso por cada una de datos). Sin embargo, para alcanzar rendimientos cercanos al 100% el protocolo AARA necesitaría un número teóricamente infinito de ranuras de acceso, lo cual es evidentemente inviable en la práctica. Otro de los protocolos ampliamente usado que utiliza ranuras de acceso es el *Distributed Queue Request Update Multiple Access* (DQRUMA) [27], que será explicado más adelante. La principal característica de este protocolo es que todo el control de acceso lo ejerce de modo centralizado la estación base. De este modo, es posible que puedan modificarse sobre la marcha los criterios de asignación de recursos de transmisión a los usuarios en función de las necesidades del sistema. Desafortunadamente, dentro del mecanismo de acceso de este protocolo se sigue haciendo uso de una estrategia Slotted-ALOHA, con lo que sigue estando presente el problema de la inestabilidad.

Otros protocolos de acceso al medio han sido específicamente diseñados para redes de cable y por ello han sido ampliamente utilizados en redes informáticas de área local. Entre ellos cabe destacar los llamados *Carrier Sense Múltiple Access* (CSMA) [29] en sus variantes de detección y evasión de colisiones, que se utilizan en algunos estándares de redes como el IEEE 802.11 (Ethernet). También se han utilizado ampliamente los sistemas de paso de testigo, o *Token Ring*, que evitan las colisiones por completo. Para sistemas de comunicaciones móviles también se han desarrollado algunos protocolos que tratan de realizar

tareas equivalentes a la detección del estado del canal que realizan los pensados para red fija. Entre ellos cabe destacar el ISMA (*Inhibit Sense Múltiple Access*) [28], una variante del cual es el utilizado en la primera versión de las especificaciones de UMTS.

Por otro lado, y con la idea de desacoplar la resolución de las colisiones de la transmisión de datos, Xu y Campbell desarrollaron un protocolo llamado DQRAP (*Distributed Queueing Random Access Protocol*) [30], [31]. Este protocolo fue diseñado para la distribución de televisión por cable en un entorno de acceso TDMA. Este protocolo, basado a su vez en otro llamado DQDB (*Distributed Queueing Dual Bus*), que ahora constituye el estándar IEEE 802.6 para redes de área metropolitana, alcanza un rendimiento muy próximo al máximo teórico, que corresponde a un sistema abstracto de colas M/D/1. Otra ventaja muy importante de este protocolo es su comportamiento en cuanto a estabilidad: mantiene un comportamiento estable para cualquier carga de tráfico de entrada. Estas ventajosas características indudablemente inducen a utilizar la filosofía de este protocolo en otros entornos de trabajo, como los sistemas de transmisión por radio en modo paquete.

En los siguientes puntos, vamos a detenernos someramente en la descripción de algunos de los protocolos más significativos mencionados hasta el momento, centrándonos sobretodo en aquellos que se utilizan en entornos similares al que usaremos como marco de trabajo. Esta presentación del estado del arte servirá como antecedente para la presentación de las nuevas propuestas de protocolos realizadas en la presente tesis.

2.3.2.1 ALOHA puro

El protocolo ALOHA puro fue el primero que se utilizó en los sistemas de comunicaciones vía radio. Con él, cuando un usuario tiene datos para transmitir, accede al canal y realiza la transmisión libremente. Si dos o más usuarios realizan una transmisión de modo simultáneo, se produce una colisión. El equipo receptor debe notificar la correcta recepción de la información transmitida con un mensaje de control al efecto. Cuando un usuario que ha transmitido sus datos no ha recibido en un cierto intervalo de tiempo la correspondiente notificación, supone que se ha producido una colisión y reintenta la transmisión previa espera de un tiempo aleatorio. Este tiempo de espera aleatorio es necesario para reducir la probabilidad de que se produzcan nuevas colisiones.

El protocolo ALOHA puro ofrece un buen rendimiento para cargas bajas de tráfico, ya que se minimiza el retardo de propagación. Sin embargo, el problema principal del protocolo aparece cuando la carga de tráfico aumenta. A medida que las colisiones se hacen más frecuentes, el número de retransmisiones aumenta y éstas a su vez incrementan la carga de tráfico ofrecido, lo que genera más colisiones. Este proceso genera una inestabilidad que puede llegar a saturar el sistema y bloquear las transmisiones.

Suponiendo una población infinita de usuarios que generan tráfico con estadística de Poisson, el número medio de paquetes de información correctamente transmitidos por unidad de tiempo, *S*, responde a la expresión:

$$S = Ge^{-2G} \tag{2.32}$$

Donde *G* es la carga ofrecida, incluyendo las retransmisiones. Como puede observarse, el máximo absoluto de paquetes transmitidos correctamente es 1/*e*, que corresponde a una eficiencia en el aprovechamiento del canal de un 18%.

2.3.2.2 Slotted-ALOHA

Surgido como una mejora del ALOHA puro, la diferencia con éste estriba en que el eje de tiempo se divide en segmentos de longitud constante y sólo se permiten las transmisiones en el inicio de cada intervalo. Esta modificación duplica el número máximo de paquetes transmitidos correctamente. De hecho, la expresión de *S* ahora es:

$$S = Ge^{-G} \tag{2.33}$$

En la Figura 17 se muestran las dos expresiones para el caudal efectivo (*S*) en función de *G* tanto para el ALOHA puro como para el Slotted-ALOHA. Se comprueba la mejora que supone el segundo respecto del primero. Sin embargo, el máximo valor de caudal sigue siendo un valor relativamente bajo, el 36%.

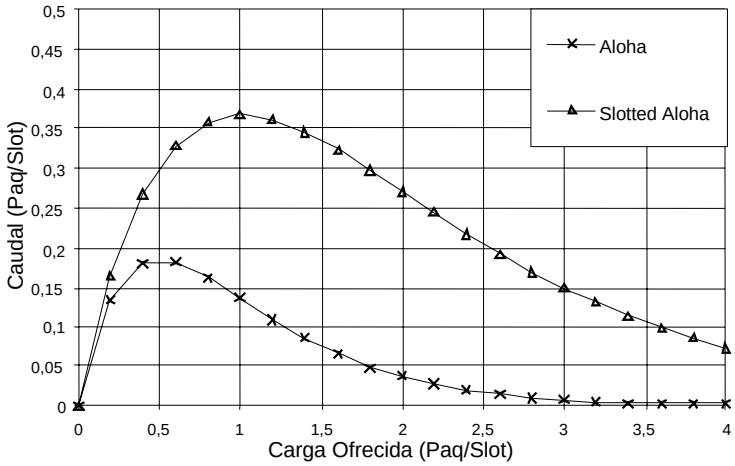

Figura 17. Caudal en función de la carga para ALOHA y Slotted-ALOHA

2.3.2.3 R-ALOHA (ALOHA con reserva)

Este protocolo es a su vez una mejora del Slotted-ALOHA. El eje de tiempos se divide en una sucesión de intervalos, unos destinados a la reserva de recursos y otros a la transmisión de los datos. Cuando un usuario desea transmitir, escoge uno de los intervalos de reserva y envía una petición. El usuario entonces debe esperar a recibir una notificación de la estación base en la que le indique que recibido correctamente su petición y qué intervalos de tiempo tiene reservados para transmitir.

Las colisiones sólo pueden darse en los intervalos reservados para las peticiones. Estos intervalos se definen sustancialmente más cortos que los dedicados a la transmisión de datos, con lo que se aumenta la eficiencia del sistema.

En ocasiones, cuando un intervalo de datos no está reservado para ningún usuario, el intervalo puede transformarse en varios más pequeños dedicados a las peticiones de transmisión.

Por tanto, R-ALOHA [32] es un protocolo con contención donde la asignación de recursos se realiza bajo demanda. Es un esquema centralizado donde la estación base tiene todo el control sobre las transmisiones que se realizan.

2.3.2.4 Packet Reservation Multiple Access (PRMA)

Propuesto inicialmente en 1989 [33], tiene una estructura muy similar al R-ALOHA, pero con la salvedad de que está preparado específicamente para fuentes de información de naturaleza mixta, en particular voz y datos. El diseño está basado en slots de tiempo, unos llamados R dedicados a la reserva de información, otros llamados I para la transmisión de datos y otros llamados A para la transmisión de reconocimientos. La gestión es también centralizada y está específicamente pensada para optimizar el rendimiento cuando el tráfico es una mezcla de voz y datos. Esta optimización se realiza gracias al mecanismo por el cual, cuando un usuario de voz ha realizado con éxito la petición de recursos de transmisión, quedan automáticamente reservados unos ciertos slots de forma periódica hasta que el usuario libera los recursos porque ha terminado su ráfaga de transmisión. Además, se da prioridad siempre a los usuarios síncronos (voz) en detrimento de los de datos, de manera que éstos incluso pueden quedarse en un momento dado sin recursos disponibles.

Este esquema es capaz de alcanzar un incremento de usuarios de voz respecto a una repartición fija de recursos (tipo TDMA) del orden del 60%.

2.3.2.5 Carrier Sense Multiple Access (CSMA)

CSMA y sus variaciones son mejoras basadas en el ALOHA puro. La idea general es que los usuarios tratan de evitar las colisiones en base a escuchar el canal antes de realizar una transmisión. Existen varios tipos de protocolo CSMA, entre los que podemos incluir:

- *1 persistente*:
    - Los usuarios escuchan el canal hasta que lo encuentran desocupado. En ese momento intentan de forma inmediata realizar sus transmisiones. Es evidente que si dos o más usuarios estaban esperando a que el canal quedase libre, con toda seguridad producirán una colisión. Por tanto, esta solución mejora el esquema básico de ALOHA, pero mantiene el problema de la inestabilidad cuando la carga de tráfico es elevada.
- *No persistente*:
    - En esta modalidad, los usuarios esperan un período de tiempo aleatorio entre dos escuchas del canal. En cuanto encuentran el canal libre, inician su transmisión. Este método reduce notablemente la probabilidad de colisión y evita los problemas de inestabilidad. El caudal efectivo no decrece a partir de un cierto valor de la carga de entrada, pero a cambio el comportamiento en retardo es peor que en el caso del 1 persistente.
- *p-persistente*:
    - Los usuarios escuchan permanentemente el canal, y en cuanto lo encuentran libre, realizan su transmisión con probabilidad *p*, o se quedan en silencio con probabilidad 1–*p*. Los usuarios que han permanecido en silencio vuelven a escuchar el canal al cabo de un cierto tiempo y vuelven a aplicar la misma filosofía.

- *Con detección de colisiones (CD)*:

    - Los usuarios comienzan la transmisión de sus datos cuando encuentran en canal libre, pero en cuanto detectan que se ha producido una colisión abortan la su transmisión. La detección de las colisiones se realiza escuchando el canal mientras se transmite, detectando que la señal que hay en el canal no es únicamente la que se está transmitiendo. Debido a la dificultad que tienen los sistemas de transmisión por radio de transmitir y enviar al mismo tiempo con la misma frecuencia de trabajo, este esquema no es adecuado para este tipo de sistemas de comunicaciones.

- *Con elusión de colisiones (CA)*:

    - Es una combinación de modalidades. Cuando un usuario encuentra el canal libre, espera un tiempo aleatorio, sin dejar de escuchar el canal. Cuando detecta que durante un cierto tiempo el canal se ha mantenido libre, inicia la transmisión. Si se detecta una colisión, se aborta la misma.

2.3.2.6 Inhibit Sense Multiple Access (ISMA)

Este protocolo es una variación del CSMA no persistente. Está especialmente pensado para adaptar las ideas de CSMA a entornos de transmisión por radio. Haciendo uso de él, los usuarios no escuchan el canal sino que reciben la información sobre su ocupación de la estación base, a través de un canal de control. Esta información de control consiste en una serie de bits de inhibición, que indican qué recursos de transmisión están ocupados.

Sin embargo, sigue siendo posible que se produzcan colisiones en la transmisión. Esto ocurre cuando dos o más usuarios comienzan al mismo tiempo la transmisión haciendo uso de un recurso indicado por la base como libre.

Este protocolo es especialmente interesante para las comunicaciones móviles. Una variante del mismo es la que se incluye en las primeras especificaciones de la capa física para UMTS. Más adelante, utilizaremos su rendimiento como referencia para comparar con las nuevas estrategias propuestas.

2.3.2.7 Distributed Queueing Request Update Multiple Access (DQRUMA)

DQRUMA es un protocolo contención con asignación bajo demanda. Constituye una extensión del R-ALOHA. Los intervalos de tiempo se dividen en minislots de acceso y en slots de transmisión de datos. Los usuarios deben solicitar recursos de transmisión a través de peticiones transmitidas en los minislots usando la estrategia Slotted-ALOHA. Cada petición lleva información sobre la identificación del usuario, los requisitos de transmisión necesarios y el número de slots requeridos. La estación base recibe las peticiones de todos los usuarios, las coloca en una cola y gestiona esta cola según un cierto algoritmo para asignar los permisos adecuados de transmisión a cada usuario.

Es un esquema de gestión totalmente centralizado que permite a la estación base tener el control absoluto de las transmisiones que realizan todos los usuarios. La aplicación del algoritmo de gestión adecuado puede permitir dar prioridades a unas conexiones respecto a

otras, aunque sigue existiendo el problema de inestabilidad inherente al acceso ALOHA de los minislots.

2.3.2.8 Distributed Queueing Random Access Protocol (DQRAP)

Propuesto por Xu y Campbell en 1994, fue diseñado específicamente para la transmisión de señal de televisión por cable. Se basa en un algoritmo de resolución de colisiones del tipo árbol, que funciona en paralelo y de modo independiente al sistema de transmisión. El eje de tiempos se divide en bloques de tres minislots de control para las peticiones de acceso y un slot para la transmisión de datos. Inspirado en un protocolo llamado DQDB (*Distributed Queue Dual Bus*), que ahora es el estándar IEEE 802.6 para las redes de área metropolitana, tiene un rendimiento muy próximo al máximo teórico constituido por un sistema de colas M/D/1. Alcanza un caudal efectivo muy cercano a la capacidad total de transmisión del canal y mantiene su estabilidad para cualquier carga de tráfico.

El protocolo es un protocolo de acceso libre tipo ALOHA cuando la carga de tráfico es pequeña, con lo que se minimiza el retardo de transmisión. Por otro lado, pasa progresivamente y de forma automática a convertirse en un protocolo de reserva a medida que la carga se va incrementando, colocando todas las transmisiones en una cola. Con esta estrategia se evita el problema de la inestabilidad y se mantiene una alta eficiencia en la utilización del canal.

Se basa en dos colas distribuidas que se encargan de gestionar, respectivamente, el algoritmo de resolución de colisiones y la transmisión de los paquetes de información. Las ideas de este protocolo han sido la motivación base de las nuevas propuestas realizadas y estudiadas en la presente tesis doctoral, pensadas para un entorno de comunicaciones móviles.

*2.4 Algoritmos de gestión de los recursos radio (*scheduling*)*

2.4.1 Introducción

Las nuevas redes de comunicaciones de banda ancha deben ser capaces de ofrecer todo tipo de servicios multimedia. Uno de los requisitos más importantes necesarios para poder ofrecer este tipo de servicios es que la red pueda garantizar, en el momento de establecer una conexión, unos ciertos parámetros de QoS (Calidad de servicio) que deben mantenerse durante el tiempo que dure dicha conexión. En las redes de topología compleja, como son las de comunicaciones móviles, compuestas de múltiples conmutadores y enlaces, puede resultar difícil evaluar los parámetros de calidad del tráfico para poder garantizar, por ejemplo, unas cotas de retardo, variación de retardo (*jitter*) y grado de pérdidas de paquetes.

En particular, la gestión de los recursos del enlace radio, también llamada *Radio Resource Maganegement* (RRM), es una faceta de especial relevancia en el diseño de los sistemas de comunicaciones móviles de tercera generación. Basada en una cierta estructura eficiente para la capa MAC, surge la necesidad de diseñar, estudiar y analizar las distintas disciplinas de servicio que debe tener cada estación base de la red de cara a ser capaz de garantizar que una comunicación extremo a extremo pueda cumplir los criterios de calidad establecidos. Este aspecto será esencial a su vez para poder diseñar correctamente una política de control de admisión y gestión de congestión en la red, lo que permitirá que la eficiencia general del sistema se mantenga en valores aceptablemente altos. Estas técnicas o disciplinas es lo que llamaremos genéricamente algoritmos de gestión de los recursos o algoritmos de *scheduling*.

Al mismo tiempo, los nuevos sistemas de comunicación suscitan unas necesidades añadidas que deben tenerse en cuenta a la hora de diseñar estos algoritmos. En primer lugar, los recursos disponibles deben repartirse de la manera más equitativa posible entre las conexiones que transmiten información. Es decir, que para grados de prioridad iguales (que vendrán definidos por los parámetros que se negocien en el momento del establecimiento de la conexión), todas las conexiones deben recibir el mismo grado de servicio. Por otro lado, debe evitarse que un hipotético comportamiento anómalo (ya sea voluntario o no) de un usuario afecte negativamente al grado de servicio del resto de usuarios. En otras palabras, si una conexión introduce en la red más tráfico del pactado no debe degradarse la calidad de las conexiones que se mantienen dentro de sus límites permitidos.

En este apartado vamos a analizar el marco de trabajo en el que centraremos el problema de la definición de los algoritmos de gestión de recursos. A continuación entraremos a describir los distintos tipos de disciplinas y las implementaciones más relevantes propuestas hasta la actualidad, es decir, el estado del arte. Es interesante reseñar que estos algoritmos analizan el problema de la gestión de los recursos desde el punto de vista de un sistema con acceso TDMA. La presentación de estos algoritmos nos permitirá establecer un marco de referencia que será la base de las nuevas propuestas de gestión presentadas en el capítulo 6, específicas para sistemas móviles con acceso CDMA. Tal y como se ha reseñado anteriormente, esta es la técnica que será utilizada en los sistemas móviles de tercera generación. Para todo ello, es necesario en primer lugar caracterizar los diferentes flujos de información a los que se deberá dar servicio: son los llamados modelos de tráfico.

2.4.2 Modelos de tráfico en servicios garantizados

Existe una infinidad de modelos diferentes para caracterizar las distintas fuentes de tráfico que pueden requerir servicios de una red de comunicaciones. Sin embargo, algunos de ellos se han convertido en tradicionales, en el sentido en que son los usados mayoritariamente para las fuentes de tráfico más comunes. Estos son el modelo de Poisson para fuentes de datos, el modelo *ON-OFF* para fuentes de voz y el modelo MMPP para fuentes de vídeo.

Recientemente se han propuesto nuevos modelos que sugieren acotar el tráfico en lugar de caracterizar el proceso de forma exacta [53]. Estos modelos no proporcionan una descripción a nivel de conexión sino que proporcionan una caracterización del volumen de tráfico general y permiten caracterizar una amplia variedad de fuentes que generan información a ráfagas también llamadas *bursty*.

A continuación describiremos algunos de ellos.

- (σ,ρ). Una conexión de tráfico responde a este modelo si durante un intervalo de tiempo de longitud u, el número de bits que llegan a la conexión en este intervalo es inferior a $\sigma+\rho\, u$. En este modelo σ puede interpretarse como el tamaño máximo de la ráfaga y ρ como la tasa de la fuente.

- $(X_{min}, X_{ave}, I, S_{max})$. Es una variante del modelo anterior. Una conexión de tráfico satisface este modelo si el intervalo de tiempo entre la llegada de dos paquetes consecutivos es siempre superior a X_{min}, si el tiempo medio entre llegadas en un intervalo de tiempo I es siempre superior a X_{ave} y si el tamaño máximo de los paquetes es siempre inferior a $S_{max.}$ A diferencia del descriptor anterior, modela el tráfico a nivel de paquete ya que limita explícitamente el tamaño máximo de paquete y limita la velocidad de pico. El número máximo de bits que una fuente puede enviar en un intervalo de tiempo de longitud t está

acotado por $\left(\frac{t}{I}+2\right)\left\lceil\frac{I}{X_{ave}}\right\rceil S$, donde $\left(\frac{t}{I}+2\right)$ acota el número de intervalos de longitud I en que puede dividirse un intervalo de tiempo t. En cada uno de esos intervalos la fuente puede transmitir $\left\lceil\frac{I}{X_{ave}}\right\rceil S$ bits.

- *(r,T).* El tráfico de una fuente satisface este modelo si en un intervalo de tiempo T no se transmite un número de bits superior a $r \cdot T$.

La caracterización de las cotas puede ser tanto determinista como estocástica. La cota determinista de tráfico define la función de limitación de tráfico. Así pues, diremos que la función monótona creciente $b_j(\cdot)$ es una función determinista de limitación de tráfico para una conexión j si durante un intervalo de tiempo de longitud u el número de bits que llegan a la conexión j no es superior a $b_j(\cdot)$. Es decir si $A_i(t_1,t_2)$ es el número total de bits que llegan a una conexión j en un intervalo de tiempo (t_1,t_2) , $b_j(\cdot)$ es una función de limitación de tráfico para esa conexión si $A_i(t_1,t_2) \leq b_j(\cdot)$.

Para una fuente de tráfico determinada puede haber diferentes funciones limitadoras de tráfico. Cada uno de los modelos de tráfico definidos anteriormente tiene sus correspondientes funciones limitadoras. Por ejemplo, la función de limitación de tráfico para el modelo (σ,ρ) es $\sigma+u\rho$.

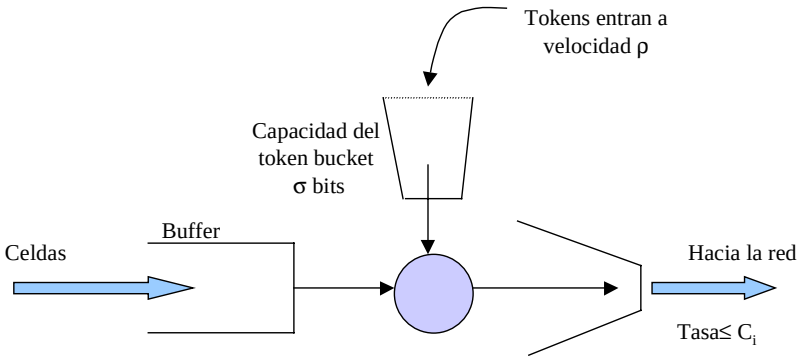

Figura 18. Regulador Leaky Bucket.

Una manera de limitar el tráfico que genera una fuente es el uso de conformadores o suavizadores de tráfico. El conformador llamado *leaky-bucket* (ver Figura 18) regula el descriptor de tráfico (σ,ρ). El objetivo del conformador es igualar el tráfico *bursty* para evitar el desbordamiento del *buffer*. A su vez, le permite a la red garantizar un cierto retardo máximo extremo a extremo. El *leaky bucket* acumula ranuras (*tokens*) de longitud fija en un cubo (*bucket*). Cada ranura permite a la fuente enviar un cierto número de bits hacia la red. Cuando los paquetes llegan al regulador, este envía los paquetes si en el cubo el número de ranuras supera el tamaño del paquete. En caso contrario el paquete ha de esperar a que haya un número suficiente de ranuras para poder ser transmitido. Si esto no ocurre será descartado. Cuando el paquete abandona el regulador, el cubo elimina los *tokens* correspondientes al paquete transmitido. Además, el regulador añade periódicamente *tokens* a una cierta tasa ρ al cubo. Así pues, el *leaky bucket* limita el tamaño de la ráfaga transmitida al tamaño del cubo. Sobre un periodo de tiempo suficientemente grande, la tasa de los paquetes que son transmitidos queda limitada a la tasa a la que ranuras son introducidas en el cubo.

2.4.3 Funcionalidades de los algoritmos

Desde un punto de vista TDMA, dada una conexión que necesita transmitir datos en modo paquete, hay tres tipos de recursos que se le deben asignar:

- Ancho de banda: Velocidad de transmisión con la que se deben transmitir los paquetes.
- Puntualidad: Cuándo deben transmitirse exactamente los paquetes.
- Espacio de los *buffers*: cuántos paquetes se descartarán estadísticamente.

y que afectan directamente a tres parámetros de comportamiento de la conexión, que son:

- Caudal efectivo o *throughput*, medido en bits transmitidos por segundo.
- Retardo medio de cada uno de los paquetes.
- Tasa de perdidas de los paquetes.

Por otro lado, las características que podemos enumerar de un algoritmo de *scheduling*, y que deben tratar de optimizarse, son:

- Eficiencia: Grado de aprovechamiento de los recursos. Un algoritmo es más eficiente que otro si puede conseguir el mismo retardo extremo a extremo cuando la red tiene una mayor carga de tráfico. El uso de un algoritmo eficiente se traduce en una mayor utilización de la red, y por tanto mejor aprovechamiento de los recursos.

- Protección: Robustez ante los usuarios anómalos (*ill–behaving users*). Debe garantizar los requerimientos de cada conexión independientemente de la presencia de usuarios que intenten transmitir violando su contrato de tráfico con la red, de las fluctuaciones de la carga de la red y del tráfico de tipo *best-effort* no acotado.

- Flexibilidad: Capacidad de escalar el sistema y asignar parámetros distintos a cada conexión. Debe permitir asignar diferentes anchos de banda, retardos, potencias de transmisión y tasas de perdidas según las aplicaciones y sus requerimientos de calidad de servicio.

- Simplicidad: Sencillez de implementación. Deben ser analíticamente tratables y fáciles de implementar para permitir su uso a velocidades elevadas y evaluar con sencillez sus prestaciones.

Los algoritmos de *scheduling* se clasifican en dos grandes grupos, según aprovechen siempre la capacidad de transmisión o no:

- ***Estrategias Work-conserving***: Cada uno de los recursos de transmisión disponibles no están nunca en estado ocioso (*idle)* mientras haya algo por transmitir. Se aprovecha siempre la capacidad del enlace cuando hay información a transmitir.

- ***Estrategias Nonwork-conserving***: Un cierto recurso de transmisión puede estar inactivo incluso cuando haya paquetes listos para ser transmitidos. En la sección 2.4.6 se detallan las motivaciones, ventajas e inconvenientes de usar estrategias de este tipo.

Estos dos grandes grupos contienen distintas propuestas que serán analizadas con detenimiento en las secciones 2.4.5 y 2.4.6.

2.4.4 Estado del arte

El principal objetivo de los algoritmos de *scheduling* es asegurar la tasa reservada y las cotas máximas del retardo de los paquetes, del *jitter* del retardo y de las pérdidas de paquetes.

En este punto vamos a citar los algoritmos más relevantes que se han propuesto hasta el momento, para después describirlos con más detalle en las secciones sucesivas. Es interesante reseñar que todas la propuestas han sido diseñadas para entornos del tipo TDMA, pero serán descritas porque constituyen la base de las nuevas estrategias que deben ser definidas para los interfaces radio del tipo CDMA, como las nuevas propuestas presentadas en la presente tesis.

En un entorno TDMA, la forma más simple para conseguir el objetivo de asegurar una tasa de transmisión, cuando todos los paquetes tienen la misma longitud y los mismos requisitos, es servir las conexiones según un esquema *round robin* y transmitir el mismo número de paquetes de cada conexión. Este procedimiento, sin embargo, asigna la misma tasa de transmisión a todas las conexiones, lo cual supone una limitación. Para superar esta limitación existe el algoritmo *weighted round robin* que por un lado asigna distintos pesos a las conexiones en función de la tasa que tienen reservada y por otro sirve las conexiones siguiendo una estrategia *round robin*, pero transmitiendo un número de paquetes igual al peso de cada conexión. Sin embargo, este sistema no es viable cuando las conexiones tienen paquetes de diferente longitud . Una nueva alternativa para solventar este problema es un algoritmo llamado *deficit round robin*.

La principal ventaja de los algoritmos del tipo *round robin* es que son simples de implementar. Sin embargo, garantizan cotas de retardo bastante grandes en términos relativos. Además, dado que el máximo retardo viene especificado por la duración de la rotación de la consulta de las diferentes conexiones, idéntica para todas ellas, no es posible proporcionar cotas de retardo distintas a las diferentes conexiones. Estas limitaciones han hecho recurrir a disciplinas basadas en la asignación dinámica de prioridades.

Los algoritmos de asignación dinámica de prioridades asignan a cada paquete una prioridad y después sirven los paquetes según un orden creciente de prioridades. Si la tasa reservada para una cierta conexión *j* es *r*, se asignan prioridades a los paquetes de dicha conexión tales que el tiempo de partida de cada paquete sea, como máximo, una cierta constante más el tiempo de partida del paquete, suponiendo que las transmisiones se realizan a una tasa *r*. Una posible asignación de prioridades para los paquetes de una conexión es su tiempo de salida con una transmisión a tasa *r*. Esta asignación de prioridades da lugar al algoritmo llamado *Virtual Clock* (ver 2.4.5.2). Sin embargo, dicho algoritmo a pesar de acotar el tiempo de partida, posee dos limitaciones importantes que pueden ilustrarse fácilmente y que dan pie al diseño de otras estrategias:

1) No es equitativo. Esta limitación se observa con un ejemplo: Llamaremos servidor a cada uno de los recursos de transmisión con los que es posible transmitir paquetes a una cierta velocidad. Llamaremos capacidad del servidor a la velocidad con la que se pueden enviar paquetes con ese recurso de transmisión. Supongamos un servidor de capacidad 6 paquetes/seg que debe proporcionar servicio a dos conexiones (1 y 2) que tienen reservada una tasa de 1 paquete/seg. Las conexiones 1 y 2 envían 10 paquetes en los instantes 0 y 1 respectivamente. Entonces, el tiempo de inicio de la transmisión para un paquete *j* (j<10) de la

conexión 1 será *j* en un servidor con tasa 1 paquete/seg. Supongamos que le asignamos una prioridad de valor igual al tiempo de inicio de la transmisión, es decir, *j*. Por otra parte, para los paquetes de esa conexión, el tiempo de finalización de la transmisión será 1+*j* (*j*<10). Ahora bien, teniendo en cuenta que la capacidad del servidor es de 6 paquetes/seg, los 6 primeros paquetes de la conexión 1 serán servidos en el intervalo [0,1]. Después, en el intervalo [1,2], la prioridad de los 6 primeros paquetes de la conexión 2 será mayor que la del séptimo paquete de la conexión 1, con lo cual en ese intervalo sólo se transmitirán paquetes de la conexión 2. Consecuentemente, vemos como la conexión 1 se ve penalizada en el intervalo [1,2] por el hecho de haber utilizado los recursos libres en el servidor durante ese mismo intervalo. Para superar esa limitación se diseñaron los algoritmos equitativos (*Fair Algoritms*). Dichos algoritmos aseguran que si en un intervalo de tiempo dos o más conexiones tienen paquetes en cola, serán servidas en proporción a la tasa que tengan reservada. Pertenecen a este tipo de algoritmos las disciplinas de servicio WFQ (ver 2.4.5.4), SCFQ (ver 2.4.5.6) y FFQ.

2) Asigna únicamente tasa por conexión. Por consiguiente, el tiempo en que un paquete partirá y su correspondiente retardo están asociados a la tasa reservada para la conexión. En el ejemplo anterior, la cota del tiempo de partida y del retardo del primer paquete de la conexión 1 es al menos 1seg. Si la conexión deseara un retardo de 0.5seg, su tasa reservada debería incrementarse a 2 paquetes/seg. Esta relación entre retardo y tasa reservada (ancho de banda asignado) puede derivar en una ineficiente utilización de los recursos.

Para evitar este problema, se han diseñado algoritmos que consiguen una separación entre retardo y tasa reservada. Un ejemplo de ellos es el *Delay-EDD* (ver 2.4.5.7)**.**

Todos los algoritmos considerados hasta el momento sirven paquetes siempre que haya alguno almacenado en las colas del nodo y por tanto son *work–conserving*. En contraposición a este tipo de algoritmos, los *nonwork–conserving* pueden mantener los recursos de transmisión en estado ocioso (*idle*) incluso cuando hay paquetes para transmitir. Ese tipo de algoritmos únicamente sirven a los paquetes de una conexión en instantes predeterminados de tiempo. Pertenecen a este tipo de algoritmos el *Hierarchical Round Robin* (ver 2.4.6.3) y el *Stop-and-Go Queuing* (ver 2.4.6.2). Otros algoritmos *nonwork–conserving* sirven los paquetes únicamente después de que estos pasan a ser *elegibles.* El cálculo del tiempo que debe pasar para que sean elegibles depende del algoritmo en cuestión. Por ejemplo, un método simple de calcular el tiempo de *elegibilidad* puede ser tomar el tiempo en que el paquete habría llegado si la conexión hubiera transmitido sus paquetes a la tasa reservada. Pertenece a este tipo el algoritmo *Rate Controlled Static Priority Queuing* (RCSP, ver 2.4.6.4).

Los algoritmos *nonwork–conserving* pueden repartir únicamente tasa (asignar ancho de banda) o lograr una separación entre reparto de tasa y retardo. Por ejemplo, los algoritmos *Hierarchical Round Robin* y *Stop-and-Go Queuing* únicamente reparten tasa mientras que RCSP y *Jitter* EDD aplican el principio de separación entre reparto de tráfico y retardo. El concepto de ecuanimidad no es aplicable a los algoritmos *nonwork–conserving.*

La principal ventaja de estos algoritmos es que mantienen la estadística del tráfico de la conexión cuando sus paquetes llegan al destino y esto reduce los requerimientos de *buffer* para las conexiones en la red. Sin embargo el hecho de mantener unos recursos en estado *idle* cuando los paquetes están almacenados en la cola incrementa necesariamente el retardo medio de los paquetes. Los algoritmos *work–conserving* consiguen mejor comportamiento en retardo si bien a expensas de aumentar los requerimientos de los *buffers.*

Existen otros algoritmos llamados *best effort*, uno de cuyos principales objetivos es repartir equitativamente el ancho de banda. En este sentido todos los algoritmos *fair* nombrados anteriormente pueden satisfacer este objetivo. Sin embargo, dado que los recursos no son reservados según el esquema de servicio *best-effort*, estos algoritmos asignan el ancho de banda del servidor en proporción al peso (prioridad asignada) y no a las tasas reservadas por las conexiones.

En la Tabla 4 se resume la clasificación de los algoritmos anteriormente presentados.

Tabla 4 Clasificación de lo algoritmos

	RATE ALLOCATION	DELAY ALLOCATION
Work Conserving	Virtual Clock WFQ SCFQ	Delay EDD
Non-work Conserving	Hierarchical Round Robin Stop-and-go Queuing	Jitter EDD RCSP

A continuación profundizaremos un poco más en cada una de estos algoritmos.

2.4.5 Algoritmos Work-conserving

2.4.5.1 First-come-first-served (FCFS)

Es la forma más simple de asignación de prioridades. La marca temporal asignada a cada paquete, que marca su prioridad, es simplemente el tiempo de su llegada. Los paquetes son servidos de acuerdo con sus tiempos de llegada. El principal problema de *FCFS* es que no proporciona ningún tipo de aislamiento. Si una conexión está muy ocupada incrementará el retardo experimentado por el resto de las conexiones, afectando a su QoS. *FCFS* no es capaz de ofrecer una cota determinista del retardo ni del *jitter* del retardo independientemente del estado de la red y de las características del tráfico de las conexiones que compiten por el servicio.

2.4.5.2 Virtual clock (VC)

Propuesto por Zhang [34]. El algoritmo *virtual clock* va encaminado directamente a emular el sistema de multiplexado por división en el tiempo TDMA.

Formalmente, en este algoritmo la marca temporal, llamada $VC^{k}_{i,j,}$ asociada al paquete *k* de la conexión *j* en el servidor *i*, se calcula de manera relativa respecto al tiempo real $a^{k}_{i,j}$ y a la tasa de la conexión, $r_{i,j}$:

1. A la llegada al servidor *i*, el paquete *k*, p_j^k de la conexión *j*, se marca con un parámetro llamado *virtual clock*, $VC^{k}_{i,j,}$ de acuerdo con la expresión:

$$VC^{k}_{i,j} \leftarrow max\left\{VC^{k}_{i,j}, a^{k}_{i,j}\right\}+1/r_{i,j} \qquad (2.34)$$

2. Los paquetes son servidos en orden creciente de sus tiempos virtuales de transmisión.

En la Figura 19 se muestra un ejemplo que permite ilustrar el funcionamiento del *virtual clock*.

Tenemos tres conexiones compartiendo el mismo enlace. Reservan recursos de acuerdo con las características de su tráfico. La conexión 1 tiene un tiempo medio entre llegadas de 2 unidades, las conexiones 2 y 3 tienen un tiempo medio entre llegadas de 5 unidades. Suponemos que los paquetes de todas las conexiones tienen la misma longitud y que el tiempo de transmisión de un paquete es de una unidad de tiempo. Así pues, las conexiones 2 y 3 reservan un 20%, cada una de ellas, del ancho de banda del enlace, mientras la conexión 1 reserva un 50%. La llegada de paquetes a las tres conexiones se muestra en las tres primeras líneas. Las conexiones 2 y 3 envían paquetes a mayor velocidad mientras que la conexión 1 envía los paquetes de acuerdo con lo esperado. La última línea muestra el orden en que serían servidos los paquetes en caso de utilizar un disciplina FCFS. En ese caso, aunque la conexión 1 reserva más recursos, el mal comportamiento de las conexiones 2 y 3 afecta a su comportamiento.

El *virtual clock* asigna a cada paquete un tiempo virtual de transmisión basado en el patrón de llegada de paquetes y en la reserva hecha por la conexión a la cual pertenece. En la cuarta línea se muestra el tiempo virtual de transmisión asignado a cada paquete mientras que en la quinta podemos ver el orden en que son transmitidos los paquetes. Como puede apreciarse, a pesar de que las conexiones 2 y 3 envían paquetes a una velocidad mayor de la especificado, el algoritmo asegura un buen comportamiento de la conexión 1.

El *virtual clock* permite variar la velocidad de transmisión de los paquetes siempre y cuando no se exceda la capacidad del servidor.

Asimismo permite garantizar la tasa y el retardo de cada sesión, independientemente del comportamiento de las otras conexiones.

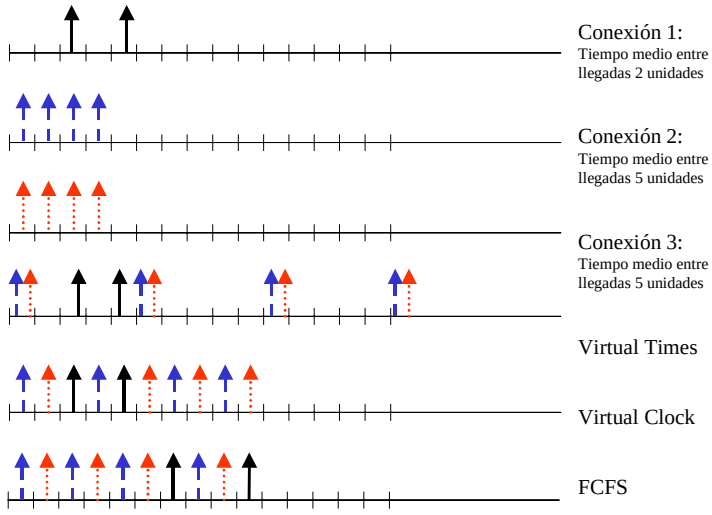

Figura 19. Comparación Virtual Clock y FCFS

2.4.5.3 Generalized Processor Sharing (GPS) o Fluid Fair Queueing (FFQ)

FFQ o GSP es una generalización de la disciplina HOL-PS (*Head Of Line Processor Sharing*) [68].

En HOL-PS se dispone de una cola FIFO para cada una de las conexiones que se reparten el enlace. En cada intervalo de tiempo, supuestas *N* conexiones activas, el servidor da servicio

simultáneamente a los *N* paquetes situados en la cabecera de las colas cada uno de ellos con una tasa de $1/N$ de la velocidad del enlace.

Mientras HOL-PS sirve las colas a la misma velocidad, GPS permite a las diferentes conexiones tener asignadas diferentes velocidades de transmisión.

En GPS cada conexión (cola) tiene asignado un número real positivo, $\phi_1, \phi_2, \phi_3 \ldots \phi_N$. En un instante *t*, la tasa de servicio para los paquetes de una cola *i* será $g_i = \frac{\phi_i}{\sum_{j \in B(t)} \phi_j} C$, donde *B(t)* es el conjunto de colas no vacías y *C* la velocidad del enlace. El parámetro g_i puede interpretarse como el mínimo ancho de banda garantizado a una conexión [35].

GPS es atractivo por varias razones:

1. Supuesto que tenemos una conexión *i* que genera paquetes a una velocidad r_i, siempre que $r_i \leq g_i$ la conexión tendrá un caudal eficaz garantizado independientemente de las peticiones del resto de los usuarios.

2. El retardo de los bits que llegan de una conexión *i* pude acotarse en función de la longitud de la cola de la conexión *i*, independientemente de las colas y de las llegadas de otras conexiones.

3. Variando el parámetro ϕ_i, se dispone de flexibilidad para tratar las conexiones de diferente forma. Por ejemplo, cuando los ϕ_i son iguales para todas las conexiones, el sistema es equivalente al *Uniform Processor Sharing*. Mientras la suma de tasas medias de todas las conexiones sean menor que *C*, cualquier asignación de ϕ_i mantiene estable el sistema. Así pues, si tenemos una conexión con poca sensibilidad a los retardos temporales es posible asignarle una ϕ_i mucho menor que su tasa media, permitiendo en contraposición un mejor tratamiento del resto de las conexiones. Así pues, los retardos experimentados por los paquetes de una sesión pueden reducirse incrementando el valor de ϕ_i para esa sesión. Sin embargo, esa reducción se hará a costa de incrementar el retardo de los paquetes del resto de las sesiones. Este incremento puede no ser importante si la conexión favorecida tiene un tráfico constante [36].

4. Es posible garantizar un retardo máximo cuando las fuentes están regidas por un *Leaky Bucket*. Esto resulta altamente atractivo cuando trabajamos con servicios en tiempo real como voz y vídeo.

5. Puede asegurarse el reparto justo de los recursos de ancho de banda entre todas las conexiones que tienen paquetes en cola sin tener en cuenta si su tráfico está o no limitado. Esta característica es importante para soportar servicios de tráfico *best effort*.

GPS es una disciplina ideal que no transmite los paquetes como entidades. Asume que el servidor puede servir todas las conexiones con paquetes en cola simultáneamente y que el tráfico es infinitamente divisible. En un sistema más realista únicamente una conexión puede recibir servicio en cada instante de tiempo y cada paquete debe ser servido íntegramente antes de que otro paquete pueda ser servido.

2.4.5.4 Packet by packet GPS o Weighted Fair Queueing (WFQ)

PGPS (WFQ) es una buena aproximación de GPS incluso cuando los paquetes son de longitud variable [36]. Su implementación es sin embargo compleja. Un algoritmo más simple aunque menos preciso es el SCFQ [37].

WFQ implica el reparto del ancho de banda disponible entre las conexiones activas de forma proporcional a sus tasas de transmisión de paquetes. Esta idea sólo puede llevarse a la práctica de forma exacta en el algoritmo teórico de *scheduling* GPS. En cualquier algoritmo paquete a paquete, el ancho de banda puede únicamente repartirse con una granularidad definida por el tiempo de transmisión de un paquete

En WFQ, cuando el servidor está listo para transmitir en un instante *t*, escoge de entre todos los paquetes en cola en el sistema en ese momento, aquel que completaría su servicio en primer lugar si no llegaran más paquetes después del instante *t*.

El ejemplo de la Figura 20 muestra las diferencias entre PGPS (WFQ) y *virtual clock*.

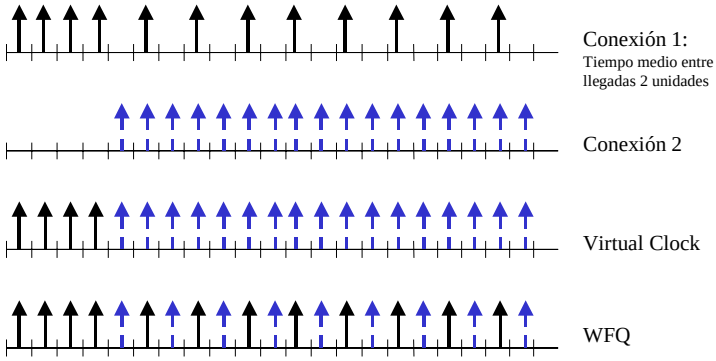

Figura 20. Comparación entre Virtual Clock y WFQ

El ejemplo se explica como sigue: supongamos que hay dos conexiones, ambas con una tasa media especificada de un paquete cada 2 seg. Suponemos además que todos los paquetes tienen una longitud fija y que requieren exactamente 1 seg de servicio. Empezando en el instante cero, llegan 1000 paquetes a la conexión 1 a una velocidad de 1 paquete/seg. En la conexión 2 no empiezan a llegar paquetes hasta el instante 900. En ese momento llegan 450 paquetes a dicha conexión a una velocidad de 1 paquete/seg. Si ambas sesiones son tratadas por igual, el *virtual clock* marcará cada sesión con una tasa 1/2 y por otra parte PGPS asignará un peso de $\phi_1=\phi_2$. Dado que ambas disciplinas son *work conserving*, se servirá a la sesión 1 continuamente en el intervalo [0 900).

En el instante 900^-, no habrá paquetes en la cola de ninguna de las conexiones. En el caso de utilizar el algoritmo *virtual clock*, en el instante 900 el paquete de la sesión 1 tendrá asignado un *VC*=1800 mientras que el primer paquete de la sesión 2 tendrá *VC*=900. Así pues, a medida que los paquetes de la conexión 2 vayan llegando a la cola se les irán asignando los *VC*=900, 902, 904,...., 1798, mientras que los paquetes que vayan llegando a través de la conexión 1 tendrán *VC*=1800,1804,.... Por lo tanto, todos los paquetes de la sesión 2 sean servidos antes que los paquetes de la conexión 1. Como puede apreciarse, los paquetes de la conexión 1 habrán resultado penalizados por el hecho de utilizar el servidor en exclusiva en el intervalo [0 900). Bajo PGPS (WFQ), ambas conexiones serán servidas según un esquema *round robin* desde el instante 900, esto se traducirá en un retardo menor para los paquetes de la conexión 1.

La diferencia entre el comportamiento del *virtual clock* y PGPS viene del hecho de que el tiempo virtual de transmisión en el *virtual clock* es independiente del comportamiento del resto de las conexiones. El retardo de cada paquete depende de la historia completa de la conexión que se refleja en el VC. Sin embargo en WFQ el tiempo virtual del sistema depende también de cuántas conexiones están activas en el sistema.

Virtual clock es una disciplina de servicio más simple que PGPS. Proporciona una cota de comportamiento similar a la de PGPS y garantiza un caudal eficaz medio para cada conexión. Tiene, sin embargo, con respecto a PGPS una desventaja: el caudal eficaz instantáneo puede ser mucho menor que $1/r_i$ durante periodos significativos de tiempo.

2.4.5.5 Worst-case Fair Weighted Fair Queueing (WF^{2}Q)

A pesar de que PGPS (WFQ) ha sido considerado el más próximo al sistema GPS. Sin embargo, se puede comprobar en la práctica que el grado de discrepancia entre los servicios proporcionados por WFQ y GPS es bastante superior a lo esperado. Esto dio lugar a la propuesta de un nuevo algoritmo, llamado *Worst-case Fair Weighted Fair Queueing* (WF^{2}Q) que proporciona un servicio prácticamente idéntico al proporcionado por GPS [69].

A diferencia de WFQ, que únicamente utiliza los tiempos de finalización de transmisión según la referencia de GPS, WF^{2}Q utiliza tanto los tiempos de inicio como los tiempos de finalización de GPS de los paquetes. En WF^{2}Q, cuando en un instante *t* queremos elegir un nuevo paquete para ser servido, en lugar de escoger entre todos los paquetes que se encuentran en el servidor, como en PGPS, el servidor considera únicamente el conjunto de paquetes que habrían empezado a recibir servicio en el sistema GPS correspondiente y selecciona aquel que habría completado antes el servicio.

El ejemplo de la Figura 21 muestra las diferencia entre WFQ (PGPS) y WF^{2}Q.

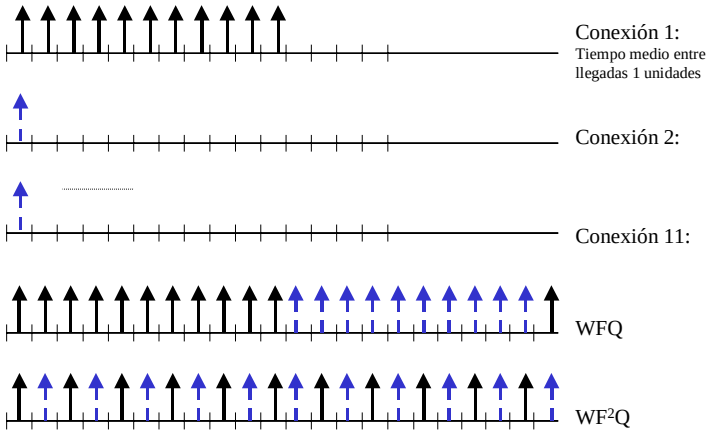

Figura 21. Comparación entre WFQ y WF^{2}Q.

Supondremos que todos los paquetes tienen la misma longitud, una unidad. Suponemos además la existencia de 11 conexiones, la primera de ellas con una tasa garantizada de 0’5 y las otra 10 conexiones con una tasa de 0’05. La conexión 1 empieza a enviar en el instante cero, y envía 11 paquetes a una tasa de 1 paquete/seg. Las otras 10 conexiones envían un único paquete en el instante cero.

Si el servidor es un GPS, atendiendo a la tasa reservada para cada conexión, empleará 2 seg en servir cada uno de los 10 primeros paquetes de la conexión 1, 1 seg en servir el paquete 11 y 20 seg para servir cada uno de los paquetes correspondientes a las otras conexiones. Supuesto el paquete k de la conexión i, p_i^k, en un sistema GPS, lo tiempos de inicio y finalización de la transmisión para dicho paquete serán $2(k-1)$ y $2k$, respectivamente, para p_1^k, con k=1...10. Para k=11 serán 20 y 21. En el caso de los paquetes de las otras conexiones, p_j^1, j=2:11, serán 0 y 20 respectivamente.

Observamos que si los servidores utilizados son WFQ o WF^2Q el orden de servicio será diferente. Si el servidor es WFQ, los 10 primeros paquetes de la conexión 1 tendrán tiempos de finalización inferiores a los paquetes de las otras conexiones, de ahí que estos 10 paquetes sean servidos con anterioridad. A continuación se servirán los 10 paquetes correspondientes a las conexiones i=2..11, ya que todos ellos tienen un tiempo de finalización inferior al del paquete 11 de la conexión 1.

En el caso de que el servidor sea WF^2Q: En el instante 0 todos los paquetes situados en las cabeceras de las colas, p_i^1, i=1,...11, habrían empezado a recibir servicio supuesto un sistema GPS. Entre todos ellos, el paquete p_1^1 es el que tiene un tiempo de finalización menor, por lo que será servido en primer lugar.

En el instante 1 habrá todavía 11 paquetes en las cabeceras de las colas, p_1^2 y p_i^1, i=2,..11. A pesar de que p_1^2 es el que tiene el tiempo de finalización más bajo no empezaría a transmitirse, según un sistema GPS, hasta el instante 2, con lo cual no es elegible para transmisión en el instante 1. El resto de los paquetes si son elegibles para empezar a transmitirse. Se escoge uno de ellos y se le da servicio.

En el instante 3, p_1^2 ya puede ser elegido para transmitir. Dado que es el paquete con menor tiempo de finalización entre todos los paquetes situados en las cabeceras de las colas, será el próximo en transmitirse.

En el ejemplo anterior se constata que:

1. En un instante de tiempo τ, la cantidad de tráfico (en bits) servida bajo ambos sistemas nunca cae por debajo del proporcionado por GPS.

2. Tanto WFQ como WF^2Q proporcionan la misma cota de retardo extremo a extremo.

3. El orden de servicio bajo cada una de las disciplinas es bastante diferente. Así por ejemplo, en el instante 10, 10 paquetes de la conexión 1 se han servido mientras que en el caso de un sistema GPS únicamente 5 habrían recibido servicio. Esta discrepancia entre ambas disciplinas podría ser más acusada a medida que el número de conexiones en el sistema aumentara. En contraposición, WF^2Q no presenta ese problema. Como puede apreciarse, en el instante 10, 5 paquetes de la conexión 1 han recibido servicio. Lo mismo que en el sistema GPS. La diferencia entre el servicio proporcionado por WF^2Q respecto a GPS es siempre inferior a la longitud de un paquete.

Así pues podemos concluir que WF^2Q es la disciplina más próxima a GPS.

2.4.5.6 Virtual Spacing (VS) o Self-Clocked Fair Queueing (SCFQ)

El algoritmo SCFQ [70] es similar a PGPS con la excepción de que ahora el *Virtual Time* se reemplaza por una variable más simple llamada *Spacing Time*. El *Spacing Time* es igual al valor de la marca temporal del último paquete que fue extraído de la cabecera de la cola.

Dado que el *spacing time* no puede ser mayor que la marca temporal de ninguno de los paquetes que están esperando en la cola cuando se actualiza su valor, el algoritmo implica que las marcas temporales de los paquetes almacenados en una conexión están separados por un intervalo de $1/r_i$. El *spacing time* sólo interviene en el cálculo de las marcas temporales de los paquetes que llegan a una conexión cuya cola está vacía y permite a la conexión ser incluida en el lugar apropiado en el algoritmo de *scheduling*.

Computacionalmente es mucho más sencillo que WFQ aunque su grado de separación respecto a GPS es mayor y puede hacer que el comportamiento de SCFQ sea mucho peor que el de WFQ.

La Figura 22 muestra la diferencia de comportamiento entre WFQ y SCFQ.

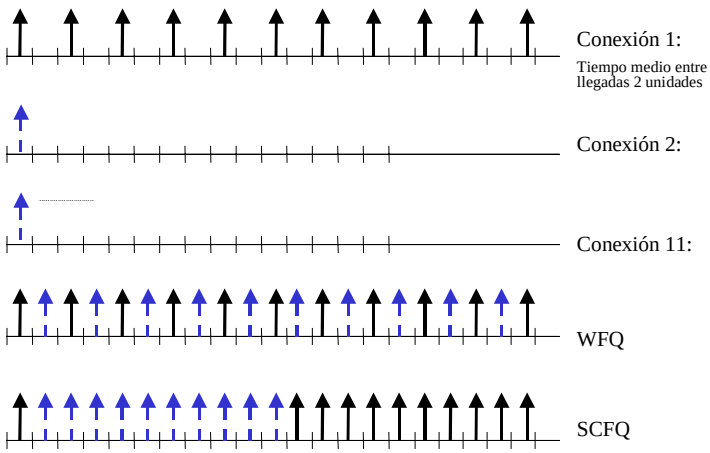

Figura 22. Comparación entre WFQ y SCFQ.

Asumimos que todos los paquetes tienen la misma longitud, una unidad. La velocidad del enlace es de 1 paquete/seg, y la tasa garantizada para la conexión 1 es de 0'5 mientras que la tasa para las otras 10 conexiones es de 0'05. Bajo la disciplina GPS, el tiempo de finalización para todos los paquetes de la conexión 1, p_1^k, k=1...10, será $2k$, y 21 para p_1^{11}. En el caso de los paquetes del resto de las conexiones, p_j^1, k=2...11, será 20. En la cuarta línea podemos ver el orden de transmisión haciendo uso de la disciplina WFQ. Si utilizamos la disciplina SCFQ, en el instante 0, al igual que en WFQ, el paquete p_1^1 es el que tiene el menor tiempo virtual de finalización con lo cual será el primero en ser servido. En el instante 1, todos los paquetes situados en las cabeceras de las colas tienen in un tiempo virtual de finalización de 20.

Entre todos los paquetes que están en las cabeceras de las colas, el paquete p_1^2 es el que tiene la marca temporal más alta, de ahí que no podrá ser servido hasta que todos los paquetes del resto de las conexiones no hayan sido servidos.

Así pues, aunque la conexión 1 envía paquetes de acuerdo con la tasa media especificada, estos sufrirán un significativo retardo.

2.4.5.7 Delay Earliest Due Date (Delay EDD)

Es una extensión de un algoritmo llamado *Earliest-Due-Date-first* (EDD o EDF). Cada paquete de una conexión de tráfico tiene asignado un parámetro llamado *deadline* y los paquetes son enviados en orden creciente de los *deadline*. El servidor negocia un contrato de tráfico con cada fuente. El contrato establece que si una fuente cumple con el contrato de tráfico en cuanto a tasa media y de pico de los paquetes enviados, entonces el servidor le garantiza una cota de retardo. La clave del algoritmo está en la asignación de los *deadlines* a los paquetes. El servidor establece como *deadline* el tiempo en el que los paquetes habrían sido recibidos, de acuerdo con el contrato de tráfico, si se hubieran enviado en ese momento. Es decir, marca los paquetes con el tiempo de llegada del paquete más la cota de retardo establecida por el servidor.

2.4.6 Algoritmos Non-work-conserving

Los algoritmos de *scheduling* que no aplican la idea del mejor esfuerzo, es decir, que no 'conservan el trabajo' de los recursos de transmisión, son muy útiles a la hora de mantener las estadísticas del tráfico entre el transmisor y el receptor de un sistema de comunicaciones. Con este tipo de estrategias, unos ciertos recursos pueden estar ociosos aun cuando existen paquetes esperando para ser enviados, tan sólo porque todavía no les ha llegado su momento para ser transmitidos. De este modo es posible hacer estudios más exactos y por tanto predecir mejor el comportamiento de las conexiones, ya que podemos acotar la variación de la forma del tráfico. Así, será posible definir con precisión las estrategias de control de admisión.

Estos métodos a su vez permiten acotar tanto los retardos extremo a extremo, como entre paquetes de la misma conexión (*jitter*) y controlar así la distorsión de la estadística del tráfico del flujo de información. Por el contrario, la no utilización de los recursos en determinados instantes de tiempo, implica necesariamente una cierta pérdida de eficacia del sistema, aumentando el retardo medio introducido en los paquetes y por tanto el caudal efectivo transportado.

Vamos ahora a analizar las principales propuestas de disciplinas de este tipo que han ido apareciendo en los últimos años, evaluando sus características principales. Es importante reseñar que todos estos algoritmos están pensados para sistemas de comunicaciones donde los paquetes deben ser transmitidos entre varios servidores distintos. Por tanto, son aplicables al interfaz aire de un sistemas de comunicaciones móviles en tanto que representa un enlace más del sistema completo de transmisión.

2.4.6.1 Jitter-Earliest-Due-Date (Jitter-EDD)

Esta estrategia se utiliza en sistemas de comunicaciones en los que los paquetes viajan por diferentes servidores en la red. Consiste en marcar cada uno de los paquetes a los que se da servicio en cada nodo de la red. El servidor coloca en su cabecera el valor correspondiente a la diferencia de tiempo entre el momento teórico máximo (según la cota especificada) en el que debería ser servido (enviado) y el instante real en el que se realiza la transmisión. En el siguiente servidor de la red, un regulador 'retiene' cada paquete durante el tiempo que tiene marcado para que no se sirva antes del tiempo en el que le corresponde. La Figura 23 muestra un esquema de este funcionamiento descrito.

Se introduce aquí de nuevo el concepto de tiempo de elegibilidad. Para poder mantener las características del tráfico que circula por la red, es necesario mantener acotadas las diferencias

de retardo entre paquetes de la misma conexión (*jitter*). Por tanto, si un paquete ha sido enviado un tiempo determinado antes de lo que debería en un servidor, no deberá ser *elegible* hasta que no se espere ese mismo tiempo en siguiente servidor. De ahí que pueda ocurrir que aun teniendo paquetes por servir, un nodo esté en estado ocioso (*idle*)

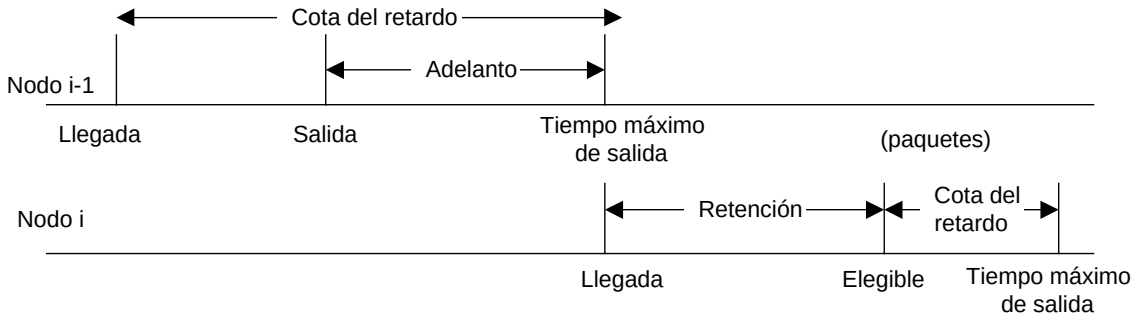

Figura 23. Servicio a los paquetes en *jitter–EDD*

Dado que hay un retardo constante entre los tiempos de elegibilidad de los paquetes de dos nodos adyacentes, se puede acotar el *jitter* del retardo para todo un flujo de paquetes pertenecientes a una conexión. En particular, si se asume que no hay ningún regulador en el servidor de destino, el *jitter* del retado extremo a extremo es el mismo que tiene localmente este último servidor.

2.4.6.2 Stop-and-Go

Esta estrategia se basa en mantener los paquetes de transmisión dentro de unas tramas de período de tiempo *T*. El eje de tiempo en todos los enlaces que llegan o salen de los servidores de trocean en porciones de longitud constante (más adelante comentaremos la variante que consiste en permitir diferentes longitudes de trama). Se definen entonces las tramas de llegada y de salida de los paquetes. Las que corresponden a los enlaces de salida están retrasadas en tiempo una cantidad constante de diseño (llamada normalmente θ). Esta cantidad θ debe mantenerse entre 0 y *T* (longitud de las tramas) para mantener la correspondencia entre cada una de las tramas de entrada con una de salida.

De este modo, todos los paquetes que llegan al servidor en una cierta trama de un enlace, no podrán ser transmitidos hasta que no comience la siguiente trama correspondiente del enlace de salida. De nuevo existe un cierto tiempo de elegibilidad, distinto de cero, que hace que un paquete no pueda ser transmitido hasta un cierto tiempo después de llegar al servidor, con la consiguiente posibilidad de que un servidor esté *idle* incluso teniendo paquetes esperando para ser enviados.

La estrategia Stop–and–Go garantiza que los paquetes que fueron enviados inicialmente en un misma trama de tiempo (los intervalos que se definen), se mantengan juntos en una misma trama durante toda la transmisión a lo largo de toda la red. Por ejemplo, si el tráfico en origen se caracteriza por (*r,T*), lo que quiere decir que no se transmiten más de $r \cdot T$ bits durante un intervalo de tiempo de duración *T*, esta estrategia asegura que se mantiene esta característica a través de toda la red si se definen las tramas del *Stop–and–Go* de duración *T*.

Otra característica importante de esta estrategia es que si cada servidor individual es capaz de asegurar que en su nodo las características del tráfico de entrada de tipo (*r,T*) se mantienen a la salida del mismo con el mismo patrón, podemos garantizar retardos máximos extremo a extremo para cualquier topología arbitraria de la red.

El problema que introduce esta estrategia es un compromiso entre la cota del retardo y la granularidad en la asignación de ancho de banda a cada enlace. Por un lado, para poder reducir el retardo máximo de un paquete es necesario reducir el valor de *T*, pero como este valor también se usa para definir el patrón de tráfico, un valor pequeño de *T* hace que los 'escalones' de ancho de banda que pueden asignarse a cada conexión sean más grandes, de forma que se pierde precisión en la asignación de velocidades de transmisión. Ocurrirá entonces con facilidad que para que una conexión cumpla los requerimientos de tráfico con los que se estableció, se deba sobredimensionar el ancho de banda que se le asigna, desperdiciando parte del mismo de cada enlace.

Para tratar de evitar en lo posible este problema, se propone una modificación de la estrategia en la que se permiten diferentes tamaños de trama temporal. La idea consiste en tener una jerarquía de tramas temporales, de diferentes niveles de prioridad, de tal manera que una trama de nivel *n* contenga *k* tramas de nivel inferior. Cada paquete tiene a su vez un nivel de jerarquía en función de sus requerimientos de retardo y deberá cumplir la condición de tráfico (r,T_n) donde T_n es el tamaño de la trama de nivel *n*. Es decir, los paquetes de un nivel determinado se mantienen dentro de la trama temporal del mismo nivel. De ese modo, los paquetes con menor retardo requerido deben asignarse a niveles donde los tiempo de trama sean pequeños, y los que tengan unas restricciones más suaves pueden viajar en tramas mayores. Por otro lado, para garantizar el buen funcionamiento de esta estrategia multinivel, se debe tener en cuenta que los paquetes que viajan en tramas pequeñas tienen en general prioridad respecto a los que van en niveles superiores y por tanto se genera un cierto grado de gestión por prioridades en los flujos de información, con las implicaciones en complejidad que ellos supone.

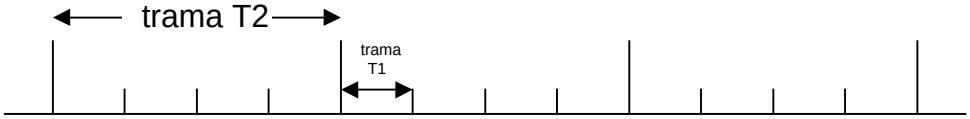

Figura 24. Estructura multitrama de dos niveles con $T_2=4T_1$

En cuanto a la implementación práctica de la estrategia, consiste en parejas de colas FIFO, una para cada nivel de prioridad, y un sistema de conmutación entre ellas. Una se encarga de almacenar los paquetes listos para ser transmitidos, porque ya se cumplió su tiempo de elegibilidad, y otra en la que se guardan los paquetes que deben esperar a que comience la siguiente trama antes de poder ser servidos. Cada tiempo de trama T_n se conmuta entre ambas, indicando que las que estaban en espera ya pueden ser enviadas, porque comenzó su trama correspondiente. Los paquetes de salida se sirven siguiendo el criterio de prioridad antes mencionado en el que los paquetes de niveles inferiores (tramas más pequeñas) se sirven antes. Esta esquema de funcionamiento se muestra en la Figura 25.

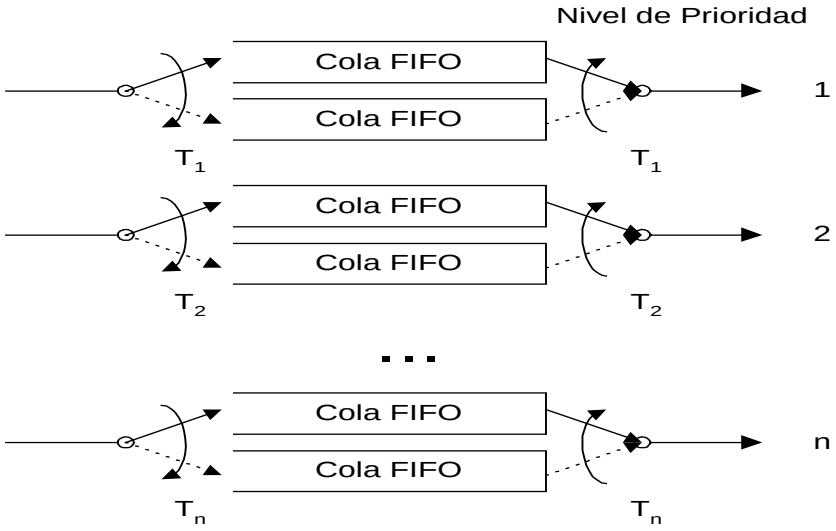

Figura 25. Implementación de la estrategia Stop–and–Go

2.4.6.3 Hierarchical Round Robin (HRR)

HRR es una estrategia similar a *Stop–and–Go* en el sentido en que también hace uso de una estrategia de tramas de tiempo jerárquicas de distintos tamaños. Cada trama (aquí llamadas *slot*) de tiempo puede asignarse bien directamente a una conexión, bien a una trama de nivel inferior. El servidor revisa cíclicamente todas las tramas y paquetes que deben ser enviados. Cuando pasa por un *slot* que ha sido asignado a una conexión, y ésta tiene paquetes pendientes de ser transmitidos, le da servicio. Cuando pasa por un *slot* que pertenece a una trama de nivel inferior, comprueba si ese nivel tiene algún paquete por enviar de las conexiones asignadas y le da servicio en su caso.

Esta estrategia es *nonworking–conserving* puesto que puede darse el caso que el servidor esté revisando un *slot* asignado a una conexión que no tiene paquetes esperando a ser servidos y que el *slot* en cuestión quede vacío, ya que no se aprovecha para enviar paquetes correspondientes a otras conexiones.

Su naturaleza *nonworking–conserving* le confiere también la propiedad de mantener las características del tráfico de la red dentro de unas cotas definibles. Su comportamiento es muy similar a *Stop–and–Go*, salvo en un detalle importante: Los paquetes correspondientes a una conexión que han sido transmitidos en una misma trama de longitud T no tienen necesariamente que mantenerse dentro de la misma trama durante toda la transmisión por la red, cosa que sí ocurre en *Stop–and–Go*). En este caso tan sólo debe mantenerse la condición de que en una trama de duración T no deben transmitirse más de $r \cdot T$ bits, es decir, un número acotado de paquetes. La Figura 26 y la Figura 27 ilustran las diferencias entre ambas estrategias, usando un ejemplo en el que para HRR se mantiene la condición de que no más de 3 paquetes de la conexión se envían en una trama.

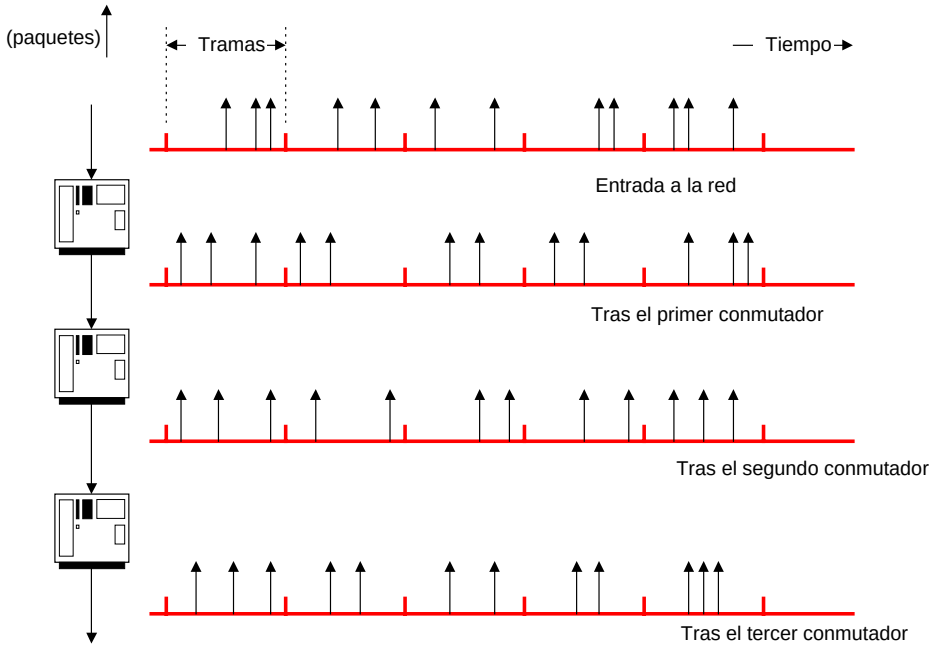

Figura 26. Disciplina de servicio de paquetes en Stop–and–Go

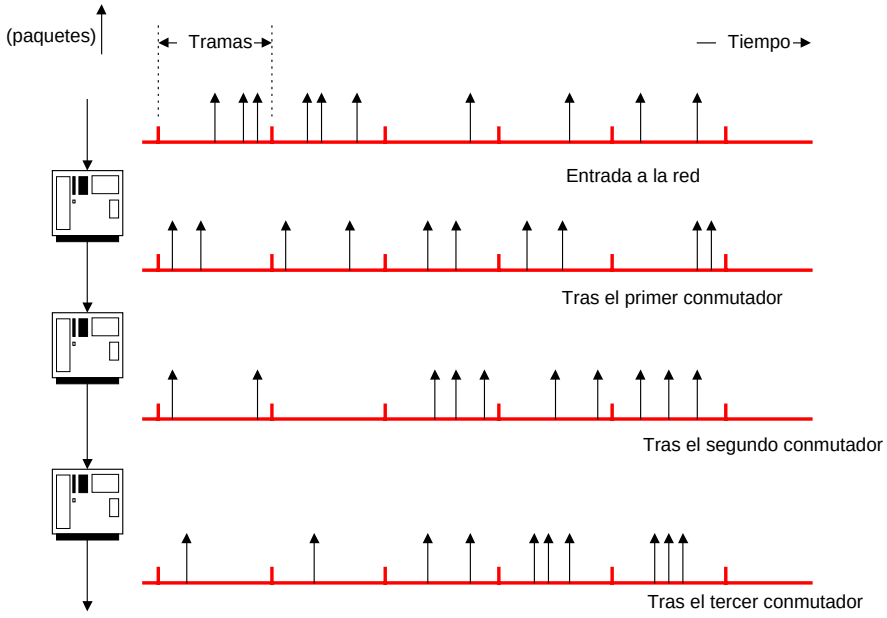

Figura 27. Disciplina de servicio de paquetes HRR

El hecho de usar una estrategia basada en tramas discretas temporales hace que la disciplina HRR mantenga el problema del compromiso entre retardo y granularidad en la asignación del ancho de banda.

2.4.6.4 Rate-Controlled Static Priority (RCSP)

La estrategia RCSP trata de paliar los problemas que supone el uso de tramas temporales, aplicando una técnica de ordenación con prioridades. Su objetivo es conseguir flexibilidad en la asignación de retardos y ancho de banda, al mismo tiempo que mantener una implementación sencilla.

RSCP está formado por dos elementos: un conformador de tráfico y un gestor (*scheduler*) de prioridad estática. El conformador de tráfico consiste en un conjunto de reguladores de tasa, uno para cada conexión entrante en el nodo de conmutación, cada uno de ellos encargado de darle al tráfico de entrada el patrón deseado. En el momento de la llegada de cada paquete, se calcula su tiempo de elegibilidad y se marca el paquete con su valor. El regulador mantiene cada paquete retenido hasta que se cumple este tiempo de elegibilidad antes de pasarlo al gestor.

El gestor controla un conjunto de colas FIFO ordenadas, de modo que siempre se selecciona para dar servicio el paquete a la cabeza de la cola que no este vacía con mayor prioridad. Posee tantas colas FIFO, es decir, tantos niveles de prioridad como diferentes cotas de retardo quieran asignarse al servidor. En el proceso de establecimiento de cada conexión, en función de sus requerimientos de retardos y tráfico, se le asigna a cada una de ellas un grado de prioridad concreto. En el caso de que varias conexiones compartan el mismo grado de prioridad, todas ellas compartirán la misma cola FIFO del nodo de conmutación. La Figura 28 muestra este esquema de funcionamiento.

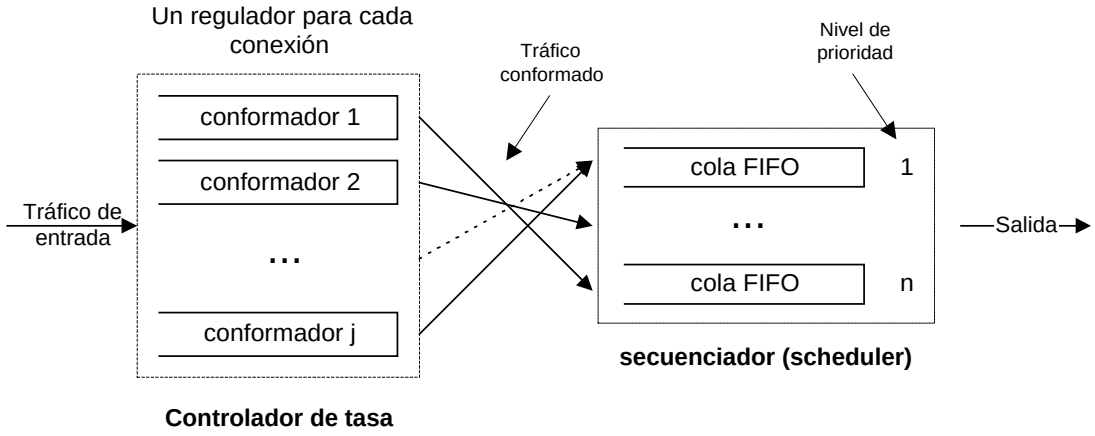

Figura 28. Implementación de RCSP

Existen diferentes formas de calcular el tiempo de elegibilidad de un paquete, en función de los parámetros con los que se caracterice el tráfico de entrada y el que se quiera mantener en la salida. De hecho, las diferentes formas de calcular este tiempo de elegibilidad y los distintos tipos de conformadores de tráfico que pueden usarse en la implementación dan lugar a una amplia variedad de estrategias, entre las cuales incluso pueden incluirse las estudiadas con anterioridad.

2.4.6.5 Resumen

Finalmente, en la Tabla 5 se resumen los algoritmos de *scheduling* descritos en este capítulo, ordenados según su clasificación. Los conceptos surgidos de estos algoritmos han servido de base para la definición de las nuevas estrategias de gestión de los recursos del interfaz radio presentadas en el capítulo 6 para sistemas de comunicaciones móviles con acceso CDMA.

Tabla 5. Algoritmos de *scheduling*

		Un sólo mecanismo		**Dos mecanismos: Regulador de tráfico y scheduler**	
		Prioridad dinámica	**Tramas multinivel**	**Prioridad dinámica**	**Prioridad estática**
Nonwork-conserving (controlan la distorsión)	**Control *jitter* de retardo**		Stop-and-Go	Jitter-EDD	RCSP
	Control *jitter* de tasa		HRR		
Work-conserving (acomodan la distorsión)	**Actualización índices prioridad basada en parámetros de la conexión**	Delay- EDD Virtual Clock		Servidores con reguladores de tráfico con colas *stand-by*	
	Actualización índices prioridad basada en la propia conexión y el resto de conexiones	WFQ SCFQ		WF^2Q	