Contribución a la Mejora del Control de Flujo en Redes de Acceso Inalámbrico

Tesis Doctoral

Director:
Juan José Alcaraz Espín
juan.alcaraz@upct.es

Departamento de las
Tecnologías de la Información
y las Comunicaciones

Gaspar Pedreño López
gaspar.pedreño@upct.es

Octubre 2015

Agradecimientos

Si hay alguien a quien tenga que agradecer particularmente el haber llevado a cabo esta tesis, es a mi director, Juan José Alcaraz. Perseverante, ingenioso, conciliador y, sobre todo, trabajador. Durante estos años has sido el espejo en el que me he mirado, mi ejemplo a seguir. Me siento afortunado de tenerte como director y como amigo. Esta tesis es más tuya que mía.

Gracias a mi familia, que tan orgullosos están de mí. Gracias por darme la mejor educación, todas las oportunidades del mundo y un cariño infinito. Va por vosotros.

Por último, y no menos importante, a ti Yolanda. Como tú bien dices, aunque solo sea por aguantarme te tengo que dar las gracias. Gracias por apoyarme y animarme a terminar este trabajo. ¡Qué grande eres!

Abstract

The framework of this thesis is the backhaul of radio access networks (RANs). We refer to backhaul as the infrastructure connecting the base stations of a cellular networks to either the radio network controller (RNCs) nodes or the core network nodes. In particular, this thesis addresses the issues associated to the congestion of the backhaul and the control algorithms that manage the resources of this infrastructure.

The potential problems caused in the RAN by the backhaul congestion are of different nature depending on the RAN technology. We will provide a historical overview of the evolution of the RAN technologies over the last two decades, focusing on the functionalities relying on the backhaul, and how each generation imposes different and somewhat more stringent demands on the backhaul infrastructure. Along this work we focus chronologically on a specific RAN generation, starting with 3G, in particular Universal Mobile Telecommunication System (UMTS), then 3.5, High Speed Data Access (HSPA) and 4G Long Term Evolution (LTE). On each one, we study the impact of backhaul congestion on the RAN performance and propose a strategy to optimally control and manage the resources of the system.

The objectives of this thesis can be formulated as follows:

1. To develop and evaluate effective control schemes to improve the performance of the transport channel synchronization functionality under backhaul congestion. The technological framework of this first objective is 3G (UMTS).
2. To develop and evaluate flow control mechanisms that jointly consider the radio interface and the backhaul. The technological framework of this second objective is 3.5G (HSPA).
3. To study the impact of backhaul congestion on downlink scheduling over the radio interface. The technological framework of this second objective is 3.5G (HSPA).
4. To develop and evaluate mechanisms for coordinate resource allocation at both the radio interface and the backhaul.
5. To address previous objective with proposals that are compliant with the technical specifications of the involved protocols.

With regard to this last objective, we consider that it is not fully realistic to come up with algorithms implying changes in systems that are already standardized and widely deployed. However, 3GPP standards do not pretend to define

every algorithm involved, since the 3GPP's main objective is to clearly specify the interfaces to allow the interoperability between operators and different vendors' devices. The internal mechanisms contained in different layers of the protocol stack to accomplish each task are generally left open to operator or vendor choice. Especially resource management functions (such as scheduling or congestion control). This approach leverages innovation and research within the industry and the academia, and allows that mechanisms as the ones presented in this thesis be feasible within existing standardized cellular networks.

The main objective of transport channel synchronization in UMTS is to make sure that the frames sent by the RNC arrive on time to the Nodes B (NBs) to be transmitted over the radio interface. For this task, the 3GPP specifies an algorithm known as timing adjustment (TA) that controls the delay suffered by the frames over the interface (Iub) connecting each NB with its corresponding RNC. The TA can add or subtract a certain quantity to the transmission delay. We show that the typical mechanism reacts too slowly in situations where the Iub delay increases abruptly, e.g. under transient congestion of the Iub. Besides, this classic algorithm shows potential instability issues in scenarios of very high delay. We address this problem using the tools of discrete-time control theory, which allows us to propose a new scheme that assures stability under any circumstances and improves the classical mechanism. The performance evaluation is carried out by means of simulation and considering realistic traffic scenarios for the Iub.

With the introduction of HSPA (3.5G) in UMTS, the scheduling function was moved from the RNC to the Node B, imposing the inclusion of new data buffers at the NBs. Additionally, this redistribution of the data storage function between the RNCs and the NBs created the need of a flow control mechanism regulating the data transfers on the Iub. We model and analyze this mechanism as a quadratic optimization problem, and exploit this approach to propose a new flow control scheme that minimizes the end-to-end delay over the RAN by considering no only the situation of the NB buffer, but also of the RNC buffers. Our approach is a backhaul-aware optimal flow control system for RAN.

Finally, in LTE (4G) systems, the peak transmission rates achievable by a user over the radio interface have boosted compared to previous generations. For the first time operators, vendors and the academic community agree on the need to optimize the backhaul resources, not only the radio resources. In fact, our point is that backhaul congestion impacts the performance of radio resource allocation and this function should be redesigned taking into consideration the capacity lim-

itation of the backhaul. We use network utility maximization (NUM) techniques to address the problem of joint radio-backhaul scheduling. We use a dual decomposition approach to propose a low-complexity, distributed mechanism, that can make resource allocation decisions in a subframe basis. Finally, we incorporate the optimal control policy for tande queues in our mechanism, improving even more the performance compared to non-coordinated schedulers.

Resumen

Esta tesis se enmarca en el estudio de la red de transporte de las redes de acceso radio (radio access networks, RANs), tambien denominado backhaul. En particular esta tesis aborda los problemas asociados a la congestión en el backhaul mediante una aproximación basada en teoría de control y la optimización matematica. los problemas causados por la congestión en el backhaul son de distinta naturaleza dependiendo de la tecnología RAN a la que da soporte. En esta tesis elaboramos un recorrido histórico por la evolución de las tecnologías RAN durante las dos últimas décadas, en el que nos centramos en las funcionalidades que dependen del backhaul, y en las diferentes y en general, más exigentes, demandas que éstas imponen en la infraestructura de backhaul. Iniciamos nuestro recorrido en las redes 3G (Universal Mobile Telecommunication System, UMTS), proseguimos con 3.5G (High Speed Data Access, HSDA) y finalizamos con 4G (Long Term Evolution, LTE). Describimos el impacto que la congestión del backhaul tiene en cada una de ellas y se proponen mecanismos de control para contrarrestar sus efectos negativos.

Los objetivos de esta tesis son los siguentes:

1. Proponer y evaluar mecanismos de control que mejoren el rendimiento de la sincronización de canal de transporte en FP en situaciones de congestión del backhaul. El marco tecnológico de este objetivo son las redes 3G (UMTS).
2. Proponer y evaluar mecanismos de control de flujo que consideren conjuntamente el interfaz radio y en el backhaul. El marco tecnológico de este objetivo es 3.5G (HSPA).
3. Proponer y evaluar un scheduling radio que considere conjuntamente los recursos del interfaz radio y del backhaul. El marco tecnológico de este objetivo es 4G (LTE).
4. Proponer y evaluar mecanismos de asignación coordinada de recursos en el interfaz radio y el backhaul. El marco tecnológico de este objetivo es 4G (LTE).
5. Abordar los objetivos anteriores dentro de la compatibilidad con las especificaciones técnicas de los protocolos implicados.

Con respecto al último objetivo, consideramos que no resulta realista plantear algoritmos que impliquen un cambio en sistemas ya estandarizados y de amplio despliegue. Sin embargo, las especificaciones del 3GPP no tienen vocación de definir el funcionamiento de todos los algoritmos involucrados ya que su objetivo último es determinar claramente las interfaces para facilitar la interconexión

entre dispositivos y equipos de distintos fabricantes y operadores. Los mecanismos que internamente emplean muchos protocolos para realizar ciertas tareas, en particular algoritmos de gestión de recursos o scheduling, se dejan abiertos a la implementación de los fabricantes, de forma que éstos pueden diferenciarse tecnológicamente unos de otros en un entorno de competencia. Esto permite que nuevos mecanismos como los propuestos en este trabajo tengan cabida en los sistemas considerados.

El objetivo principal de la Sincronización de Canal de Transporte en redes UMTS es que las tramas enviadas por la RNC lleguen a tiempo a los Nodos B para su transmisión a través del interfaz radio. Para ello, el 3GPP especifica un algoritmo conocido como *Timing Adjustment* que se encarga de controlar el retraso que experimentan las tramas en el interfaz Iub sumando o restando una cantidad constante. Este algoritmo reacciona con demasiada lentitud ante variaciones abruptas del retardo del interfaz Iub y, además, se puede volver inestable en escenarios de alta demora. Son situaciones que pueden estar propiciadas, entre otros motivos, por la congestión en el backhaul. Aplicando teoría de control en tiempo discreto, proponemos un nuevo mecanismo que garantiza la estabilidad en cualquier situación y mejora el rendimiento del algoritmo clásico. Las medidas de rendimiento se realizan mediante un simulador de la red UTRAN (UMTS RAN) teniendo en cuenta condiciones reales de tráfico en el interfaz Iub.

Con la incorporación de HSPA en las redes UMTS la función de *scheduling* se ha desplazado desde la RNC hasta el Nodo B, generando la necesidad de unos *buffers* en el Nodo B. A su vez, esta nueva distribución de la capacidad de almacenamiento entre la RNC y el Nodo B requiere de un mecanismo de control de flujo que regule la transferencia de datos entre ambos. En esta tesis realizamos un detallado estudio analítico de este control de flujo abordándolo como un problema de optimización cuadrática. Partiendo de este análisis desarramos un nuevo algoritmo de control de flujo que consigue minimizar el retardo extremo a extremo gracias a que incorpora como parámetro la ocupación de los buffers de la RNC (además de la ocupación en el Nodo B), algo que no se había considerado en algoritmos anteriores. Se trata, por tanto, de un control de flujo consciente del backhaul (*backhaul-aware*).

Desde la implantación de LTE (4G), la tasas máximas de transmisión alcanzables en el interfaz radio se han disparado con respecto a las anteriores generaciones de sistemas móviles. Por primera vez operadores, fabricantes y comunidad académica coinciden en la necesidad de optimizar el uso los recursos del backhaul además de los recursos radio. En esta tesis estudiamos el impacto que tiene

el backhaul en el algorimo de gestión de recursos radio, o scheduling. Para ello consideramos un escenario sencillo compuesto por una sola celda y una infreastructura backhaul consistente en un enlace punto-a-punto de capacidad C. Mostramos que cuando el backhaul es el cuello de botella, el rendimiento del scheduler radio con consideraciones de backhaul es claramente superior al scheduler convencional. Haciendo uso de técnicas de optimización de la utilidad de la red (NUM), abordamos de forma conjunta la gestión de recursos radio y del backhaul. Empleando descomposición dual, proponemos un mecanismo distribuido y de baja carga computacional que permite generar decisiones de asignación de recursos en el backhaul y en el interfaz radio, de forma coordinada y subtrama a subtrama. Finalmente, incorporamos en nuestro algoritmo el control óptimo de colas tandem, mejorando aún más el rendimiento respecto a los schedulers no coordinados.

Contents

CHAPTER 1

Introducción

1.1 Introducción

Esta tesis aborda los problemas asociados a la congestión en la infraestructura de transmisión que da soporte a la red de acceso radio (RAN) de las redes celulares. En el ámbito de los operadores y fabricantes de equipos, esta infraestructura se ha venido denominando red de transporte, aunque recientemente se ha adoptado ampliamente el término backhaul tanto en la industria como en el mundo académico. Como veremos, los problemas causados por la congestión en el backhaul son de distinta naturaleza dependiendo de la tecnología RAN a la que da soporte. Por ejemplo, en los estándares del Third Generation Partnership Project (3GPP) para redes Universal Mobile Telecommunication System (UMTS), a las que nos referiremos como redes de tercera generación o 3G, la congestión puede dar lugar a incrementos abruptos del retardo que, a su vez, pueden causar problemas de sincronización a nivel de trama entre las estaciones base (Nodos B o NBs) y las estaciones de control (RNCs). Este problema, sin embargo, no existe en las redes Long Term Evolution (LTE), que denominaremos 4G, ya que en ellas todas las funcionalidades radio se concentran en las estaciones base (conocidas como evolved Nodes B, o eNBs), y por tanto no implementan ningún mecanismo de sincronización de trama entre RNC y eNBs. En esta tesis se ponen de manifiesto las diferencias entre las redes 3G (UMTS), 3.5G (HSDPA) y 4G (LTE), se describen los retos que la congestión del backhaul plantea en cada una de ellas y se proponen mecanismos de control para contrarrestar sus efectos negativos.

Hasta cierto punto resulta sorprendente que la congestión en la red de backhaul haya sido un aspecto de las redes celulares prácticamente ignorado por la

comunidad científica hasta la introducción de 4G a principios de esta década. Sólo entonces han comenzado a aparecer ciertas contribuciones en las que la posible escasez de recursos en el backhaul era tomada en cuenta. Históricamente toda la atención académica se dirigía al interfaz radio. Posiblemente este cambio de actitud ha sido propiciado porque los operadores han de destinar una proporción cada vez mayor de su inversión al backhaul para poder garantizar las tasas de transmisión de las nuevas estaciones 4G. Al verse obligados a realizar un dimensionamiento más ajustado, resulta más necesario un control eficaz de la congestión para poder garantizar la calidad de servicio que los usuarios esperan de las nuevas redes. La necesidad de desarrollar métodos eficaces para controlar la congestión y reducir sus efectos se ha trasladado a los fabricantes y, finalmente, a la comunidad científica. Esta tesis, no obstante, se inició mucho antes de que el backhaul RAN fuera considerado un asunto de interés científico. Tal como siempre creímos, el tiempo nos ha dado la razón.

1.2 Motivación

En las redes 3G uno de los efectos potenciales de la congestión en el backhaul de la RAN es la variación abrupta del retardo en el interfaz Iub, que conecta los NBs con las RNCs. En dicho interfaz se define un protocolo de capa 2 denominado Frame Protocol (FP), descrito en las especificaciones TS 25.427 [1] y TS 25.402 [2]. Estas fluctuaciones de retardo afectan directamente a una de las funcionalidades esenciales del FP, la sincronización de canal de transporte.

El protocolo FP encapsula en tramas los bloques de datos (Transport Blocks) intercambiados entre la RNC y el usuario, que son de tamaño variable. Una de las funciones de FP es indicar al destino el formato que en el que se envían los bloques de transmisión enviados y recibidos. Este formato viene determinado por el tipo de canal de transporte, el tipo y la cantidad de datos que se transmiten y si el canal de transporte es dedicado o compartido. En sentido downlink, FP desempeña otra función de gran importancia: regular el instante de envío de las tramas de forma que éstas lleguen al Nodo B dentro de una ventana de recepción. Mediante este procedimiento, denominado sincronización de canal de transporte, se pretende que las tramas lleguen al Nodo B a tiempo para ser transmitidas por el interfaz radio y al mismo tiempo, que el tiempo de encolamiento en el Nodo B sea pequeño. La transmisión por el interfaz radio se realiza con una temporización fija, cada Transmisión Time Interval (TTI). La duración de cada TTI (tTTI) puede ser de 10, 20, 40 u 80 ms. El objetivo fundamental es que el encolamiento de las

tramas se produzca en la RNC, puesto que es en este elemento donde se ubican los algoritmos de scheduling y gestión de recursos, y no en el Nodo B. Además, no es conveniente que el Nodo B almacene muchas tramas, ya que éstas serán enviadas aunque el usuario abandone la celda, causando una interferencia innecesaria.

El único método documentado ([3], [4]) de sincronización de canal de transporte en FP es un mecanismo de ajuste del instante de envío de las tramas FP desde la RNC basado en sumar o restar un valor constante al momento planificado de envío. Pese a su sencillez, este mecanismo basa su funcionamiento en varios parámetros configurables que tienen un efecto directo en la tasa de pérdida de tramas y el tráfico de señalización en el Iub y cuya configuración tampoco se ha abordado de manera rigurosa. Además, la velocidad de respuesta del mecanismo clásico es muy lenta ante situaciones de incremento brusco del retardo en el Iub.

Naturalmente, existen otros causas, aparte de la congestión del backhaul, para que el Iub experimente una variación abrupta de retardo, por ejemplo:

- Porque mecanismos de gestión de prioridades en la red de transporte retengan durante un tiempo determinados flujos de datos FP de menor prioridad.
- Cuando los sistemas de ingeniería de tráfico o de restauración de rutas encaminan los flujos Iub de un Nodo B por un trayecto físico distinto.
- Si la fuente ha estado inactiva durante un tiempo suficiente para que el retardo experimente una deriva considerable, incapaz de ser seguida por el mecanismo de sincronización DCH.

Para dar una idea del orden de magnitud de las diferencias de retardo entre diferentes "ramas" Iub, si consideramos los ejemplos proporcionados en TR 25.853, una rama puede tener un retardo de 12 ms y otra de un retardo de 31 ms. En estos ejemplos, el servicio considerado es el servicio de voz y la red de transporte es ATM. Si se consideran servicios de datos y red de transporte IP, los retardos pueden ser mayores.

El problema con el algoritmo clásico es, por un lado, que su reacción es lenta en relación a la magnitud del incremento del retardo. Esto provoca que en una situación de incremento de retardo se pierdan muchas tramas FP consecutivas, afectando al rendimiento de las capas superiores. Si la situación es de reducción del retardo, se producirá un encolamiento excesivo en el Nodo B, reduciendo la efectividad de los algoritmos de scheduling y de gestión de recursos radio. Por otro lado, dependiendo de la configuración de parámetros y de la magnitud del

retardo en el Iub, el algoritmo de ajuste temporal puede sufrir un comportamiento no estable.

En 3.5G y 4G desaparece el mecanismo de sincronización de canal de transporte, pero el incremento de tráfico por usuario y de la capacidad de transmisión del interfaz radio hacen que el backhaul se convierta en el cuello de botella con más frecuencia. De hecho, ha sido el aumento de capacidad del interfaz radio de las estaciones la causa de que tanto la industria como el mundo académico comienzan a prestar una atención creciente al backhaul. En el caso de 4G, el aumento de capacidad viene acompañado por la creciente densificación de la red, con el despliegue de celdas pequeñas o small-cells, que permiten reducir la distancia del enlace radio aumentando así su capacidad. Además, las small-cells dan servicio a un número reducido de usuarios, por lo que cada usuario puede alcanzar tasas medias mayores. En resumen, el aumento de capacidad de las redes conlleva un aumento de capacidad y capilaridad en el backhaul y por tanto, el coste del mismo comienza a convertirse en una parte cada vez mayor del gasto en inversión (Capital Expenditure, CAPEX), y en el caso de los despliegues de small-cells, superando el coste de las propias estaciones.

Como el propio 3GPP indica en TR 36.932 [5], coexisten dos tipos de infraestructura en el backhaul: el tipo 1, considerado el caso ideal, y basado en fibra óptica y el tipo 2, no ideal, y compuesto por líneas DSL, y enlaces microondas. Para este segundo caso la gestión eficiente del ancho de banda en el backhaul es un aspecto crítico. Este problema enlaza directamente con uno de los temas recurrentes en investigación de redes inalámbricas: el scheduling o gestión de recursos en el interfaz radio. Existe una extensísima bibliografía en relación al scheduling en el interfaz radio, pero hasta el momento, la gestión de recursos en el backhaul y el scheduling radio se han considerado temas disjuntos. Sin embargo, puesto que el cuello de botella del sistema puede estar ubicado indistintamente en el interfaz radio o en el backhaul, una de las principales motivaciones de este trabajo es que se deben proponer y analizar mecanismos de asignación de recursos que consideren conjuntamente los recursos de interfaz radio y del backhaul.

1.3 Objetivos

Dados los precedentes y motivaciones expuestas en el anterior apartado, los objetivos de esta tesis son los siguientes:

1. Proponer y evaluar mecanismos de control que mejoren el rendimiento de la

sincronización de canal de transporte en FP en situaciones de congestión del backhaul.

2. Proponer y evaluar mecanismos de control de flujo que consideren conjuntamente el interfaz radio y el backhaul.
3. Estudiar el impacto de la congestión del backhaul en el scheduling radio.
4. Proponer y evaluar mecanismos de asignación coordinada de recursos en el interfaz radio y el backhaul.
5. Abordar los objetivos anteriores dentro de la compatibilidad con las especificaciones técnicas de los protocolos implicados.

Con respecto al último objetivo, consideramos que no resulta realista plantear algoritmos que impliquen un cambio en sistemas ya estandarizados y desplegados. Sin embargo, las especificaciones del 3GPP no tienen vocación de definir el funcionamiento de todos los algoritmos involucrados ya que su objetivo último es determinar claramente las interfaces para facilitar la interconexión entre dispositivos de distintos fabricantes y operadores. Los mecanismos que internamente emplean muchos protocolos para realizar ciertas tareas, en particular algoritmos de gestión de recursos o scheduling, se dejan abiertos a la implementación de los fabricantes, de forma que estos pueden diferenciarse tecnológicamente unos de otros en un entorno de competencia. Esto permite que nuevos mecanismos, como los propuestos en este trabajo, tengan cabida en los sistemas considerados.

1.4 Metodología

La metodología general empleada para abordar cada objetivo consta de las siguientes fases:

1. **Planteamiento del problema**. Incluye documentación y revisión del estado de la técnica (gestión de recursos backhaul y herramientas matemáticas).
2. **Construcción de la hipótesis**. La hipótesis es el mecanismo propuesto. El planteamiento se lleva cabo en dos fases.
   a) Formulación matemática del problema en el marco de la herramienta matemática empleada.
   b) Resolución del problema mediante un algoritmo de control.
3. **Experimentación**.
   a) Desarrollo de software de simulación. En esta fase se tendrán en cuenta las especificaciones del 3GPP en cada ámbito considerado.

b) Evaluación numérica de las propuestas. Esta fase se compone a su vez de tres etapas: i) Implementación de los mecanismos a evaluar, incluyendo los benchmark; ii) simulación y obtención de resultados numéricos estadísticamente significativos; iii) revisión de la hipótesis de partida en base a los resultados (vuelta al punto 2) o finalización del proceso.

Dada la naturaleza del mecanismo de sincronización del canal de transporte, resulta posible modelarlo como un sistema de control en tiempo discreto con ciertas asunciones que, como se verá, tienen poco impacto en la veracidad de los resultados. Este tipo de modelado es novedoso en este ámbito y permite plantear y analizar de una forma metódica y rigurosa nuevos algoritmos para FP empleando técnicas clásicas de sistemas de control en tiempo discreto. Para la validación de los resultados derivados del análisis numérico, se ha desarrollado un simulador que implementa el protocolo FP de acuerdo a las especificaciones del 3GPP, empleando el lenguaje C++ y haciendo uso de la librería de simulación por eventos discretos OMNeT++.

Para el diseño de mecanismos de control de flujo que contemplen conjuntamente el interfaz radio y el backhaul se emplearán también herramientas matemáticas de control, en este caso basadas en programación dinámica y optimización convexa. Uno de los retos que plantea el uso de estas herramientas es encontrar la formulación adecuada que permita que el algoritmo resultante tenga una baja carga computacional.

El estudio de la influencia de los recursos de backhaul en el scheduling radio está fundamentado en la técnica de maximización de la utilidad de red (*network utility maximization*, NUM), que emplea conceptos de la teoría de utilidad para un reparto justo de los recursos y técnicas de optimización convexa. En particular empleamos el método basado en las condiciones Karush-Kunh-Tucker (KKT) para encontrar, en ciertos casos, expresiones cerradas para la asignación de recursos.

Finalmente, para el diseño de un mecanismo scheduling coordinado para radio y backhaul, volvemos a hacer uso de la metodología NUM. Su implementación distribuida se apoya en técnicas de descomposición dual. Usamos técnicas de control óptimo de redes de colas para mejorar el algoritmo obtenido con NUM. En particular, adaptamos la política de control óptima de colas en tándem, basada en un modelo fluido de colas, para incorporarla en el mecanismo distribuido.

1.5 Contenido de la Tesis

El resto de esta memoria se organiza en los siguientes capítulos:

Capítulo 2, en el que se repasa la evolución del backhaul desde las primeras fases de 3G hasta las actuales redes LTE-Advanced, dedicando especial atención a los problemas asociados a la congestión en el backhaul.

Capítulo 3. En este capítulo se presenta nuestra propuesta de mecanismo de ajuste temporal mediante técnicas de control en tiempo discreto, tal como estipula el objetivo 1. El escenario de aplicación de este capítulo son las redes 3G (UMTS).

Capítulo 4. Este capítulo desarrolla una nueva metodología de control de flujo en el backhaul que tiene en cuenta el estado de los buffers tanto del backhaul como del interfaz radio. El escenario de aplicación son las redes 3.5G (HSPA).

Capítulo 5. En este capítulo se estudia el efecto de considerar la limitación de recursos del backhaul en el scheduling radio. El escenario de aplicación son las redes 4G (LTE).

Capítulo 6. En este capítulo se desarrollan y evalúan mecanismos para la asignación coordinada de recursos radio y backhaul. El escenario de aplicación son las redes 4G (LTE).

Capítulo 7, donde se resumen las conclusiones de esta tesis.

1.6 Contribuciones de esta Tesis

Esta tesis ha dado lugar, hasta el momento de su redacción, a *dos* artículos publicados en revistas litadas en el Journal Citation Reports (JCR) [6], [7], a *tres* contribuciones en congresos internacionales [8], [9], [10] y *tres* contribuciones en congresos nacionales [11], [12], [13].

CHAPTER 2

El Backhaul de las Redes Celulares

Resumen: En este capítulo realizamos un recorrido histórico a través de la evolución que ha experimentado, durante las dos últimas décadas, la infraestructura de interconexión de las estaciones radio. Esta infraestructura, conocida actualmente como backhaul, es una parte fundamental de la red de acceso radio (RAN), y su evolución ha sido paralela a la de las redes celulares. Iniciamos el recorrido en las redes de tercera generación (3G), en particular Universal Mobile Telecommunications System (UMTS), continuamos con 3.5G, High Speed Data Access (HSPA) y acabamos en 4G, Long Term Evolution (LTE) y LTE-Advanced, tecnología a la que prestamos una especial atención por ser la más actual y la que ha propiciado que tando la industria de las telecomunicaciones como el mundo académico dirijan una mayor atención al backhaul. En este capítulo repasamos las principales funcionalidades a las que el backhaul da soporte en cada una de las generaciones mencionadas, y los requerimentos que la red móvil impone en el backhaul. Discutimos por qué el diseño del backhaul y de los mecanismos de control que lo gestionan son tareas cada vez más relevantes y suponen un reto técnico de mayor nivel.

2.1 Introducción

Todas las redes de acceso radio necesitan, en general, algún tipo de infraestructura de cable o inalámbrica para conectar sus estaciones radio con los nodos de control

y conmutación. A esta infraestructura se la ha denominado historicamente como red de transporte, aunque actualmente se emplea más habitualmente el término *backhaul*. El conjunto formado por las estaciones radio y el backhaul que las interconecta se conoce como RAN (*radio access network*). En este capítulo realizamos un recorrido histórico por la evolución del backhaul desde las redes 2G (GSM) hasta las actuales redes 4G (LTE), deteniéndonos en los aspectos de interés en esta tesis: funcionalidad del backhaul, especificaciones de sus interfaces, requerimientos de ancho de banda y mecanismos de control de calidad de servicio.

Durante las últimas dos décadas, las redes backhaul han evolucionado desde un red de capa física de conmutación de circuitos basada en tecnologías de multiplexación por división en el tiempo (TDM), a una red de conmutación de paquetes en donde los servicios y aplicaciones de datos consumen ya la mayor parte de los recursos. Esta evolución tecnológica ha estado siempre acompañada de un incremento constante en los requerimientos de ancho de banda, especialmente con las redes LTE recientemente desplegadas. Puesto que la capacidad y complejidad de los enlaces del backhaul determinan una parte cada vez mayor del coste operativo y de capital (OPEX y CAPEX), los operadores móviles están poniendo un énfasis cada vez mayor en optimizar el backhaul RAN. El diseño de la red de backhaul se ha convertido en una parte importante del proceso de planificación de red en las etapas de despliegue y optimización de las redes móviles. Este mayor interés por parte de los operadores ha propiciado una mayor atención al backhaul también por parte de los fabricantes de equipos, que han aumentado sus inversiones en el desarrollo de tecnologías que dan respuesta a una mayor variedad de requerimientos y de modelos de negocio. Este esfuerzo conjunto de la industra ha encontrado, por fin, respuesta por parte de la comunidad académica, que tradicionalmente ha estado centrada casi en exclusiva en el interfaz radio de las redes móviles y que en los últimos cinco años ha comenzado a incrementar el número de contribuciones en las que se tiene en cuenta el backhaul y los retos que éste plantea.

2.2 Evolución del Backhaul

Tradicionalmente, ha sido el tráfico de voz el que ha establecido los requerimientos de diseño de las redes backhaul. Sin embargo, la evolución de los terminales de usuario y las aplicaciones durante la última década ha impulsado un mayor énfasis en los servicios de datos. En particular, las redes backhaul tuvieron que adaptarse para satisfacer los mayores requerimientos de ancho de banda asociados a la introducción de la tecnología high speed packet access (HSPA) y posteriormente,

a unos requerimientos aún mayores de long term evolution (LTE) de más reciente implantación, y su futura actualización LTE-Advanced.

Cada una de estas evoluciones tecnológicas ha venido acompañada de un incremento en las tasas máximas de transmisión de al menos un orden de magnitud con respecto a la tecnología anterior. Estos incrementos implican un aumento del gasto dedicado al transporte de datos, especialmente si éste continúa basado en las tecnologías convencionales de time division multiplexing (TDM) y conmutación de circuitos. Por tanto, cada vez existe un mayor interés en reducir estos costes por parte de la industria, introduciendo agregación a nivel de paquetes en distintos puntos intermedios de la RAN. Este mayor interés en el backhaul en general ha encontrado también su respuesta en el ámbito académico, en el que la problemática asociada al backhaul ha pasado de ser un tema marginal a convertirse en un *hot topic*.

En la tecnología 2G el tráfico originado en una estación base (BTS) era del orden de unos pocos Mbps, por lo que, aunque se empleaban dispositivos de agregación de tráfico (basados siempre en conmutación de circuitos), no se solía planificar un emplazamiento específico para ubicarlos. Fue con la llegada de la tecnología 3G cuando las funciones de concentración de tráfico (*traffic grooming*) y de optimización de la carga (*payload*) comenzaron a convertirse en esenciales y la red de acceso empezó a incorporar nodos específicos para estas funciones.

La evolución del *backhaul* se ha producido en varios niveles: requerimientos de ancho de banda por estación, interfaz físico y protocolos de enlace y/o red. El consorcio 3GPP apostó por la tecnología *asynchronoous transfer mode* (ATM) como nivel de enlace entre los Nodos B y la *radio network controller* (RNC). En las primeras fases de desarrollo de los estándares 3G aún no se consideraba IP (Internet Protocol) como una tecnología capaz de proporcionar la fiabilidad y calidad de servicio que TDM era capaz de garantizar en las redes GSM. No obstante, esta consideración fué evolucionando de esa percepción incial de servicio *best effort* no orientado a conexión, para ser considerado cada vez más como un tecnología asociada a la calidad de servicio. Por ello, los estándares más recientes del 3GPP especifican IP como la tecnología de referencia en redes de acceso.

Los interfaces físicos, por su parte, no están especificados en los estándares, y su elección se considera una decisión de diseño. Debido a la amplia disponibilidad de redes de acceso de *ultima milla* basada en tecnología PDH (*plesiochronous digital hierarchy*) y SDH (*synchronous digital hierarchy*), los interfaces más comunes en 2G, y en gran medida también en 3G, son los E1 (T1 en ciertos países). Sin

Table 2.1: Evolución de los interfaces de las estaciones.

	GSM	3G	3G (HSDPA)	3G (HSPA)	4G (LTE)
Ancho de banda (Mbps)	2-4	4-6	9-12	18-24	100-300
Interfaz capas 2/3	N/A	ATM	ATM	IP	IP
Interfaz físico	PDH	PDH	PDH/SDH	Ethernet	Ethernet

embargo, las estaciones 4G demandan anchos de banda generalmente muy superiores al proporcionado por los interfaces E1 y los interfaces de mayor capacidad que SDH puede soportar se consideran economicamente menos efectivos que el transporte basado en Ethernet. Por tanto, la percepción común es que, una vez que el pico de tráfico de una estación supera la capacidad de una portadora E1, la opción más viable es Ethernet. La tabla 2.1 resume la evolución de los aspectos esenciales del backhaul a través de los distintos avances tecnológicos experimentados por las redes celulares en las últimas dos décadas. Las estimaciones de ancho de banda se realizan, como en [14], para una estacion de volumen intermedio de tráfico. En todos los casos se consideran estaciones con tres sectories. Para cada sector, en el caso de GSM, se consideran 5 portadoras de 200 KHz. En 3G se consideran una portadora de 5 MHz con 60 códigos, para HSDPA se consideran 2 portadoras con tasa de pico de 1.8 Mbps, y en HSPA, dos portadoras por sector con tasa de pico de 3.6 Mbps.

Se dan por tanto dos tendencias de cambio: 1) Los interfaces físicos evolucionan de E1 (PDH y SDH) hacia Ethernet y 2) Los interfaces de capas 2 y 3 convergen a IP. La Figura 2.1 muestra distintas fases de cambio en el backhaul partiendo de la arquitectura 3G hacia 4G. Estas dos tendencias están relacionadas con un cambio fundamental en la arquitectura RAN de las redes 4G con respecto a las redes 2G y 3G: las redes 4G desplazan todas las funcionalidades del interfaz radio de las RNC a las estaciones, como veremos en la siguiente sección.

2.3 El Backhaul en LTE

Dependiendo de la categoría del terminal de usuario, el interfaz radio de LTE puede proporcionar, en una banda de 20 MHz, tasas máximas de transmisión de datos de entre 100 Mbps y 300 Mbps (de bajada), y una latencia en el plano de usuario inferior a 5 ms. Desde una perspectiva de backhaul, LTE introduce importantes cambios estructurales con respecto a UTRAN. Desde el punto de vista de arquitectura de protocolos, todas las funciones del *Radio Link Control* RLC se trasladan a los eNode N, es decir, el backhaul no proporciona transporte

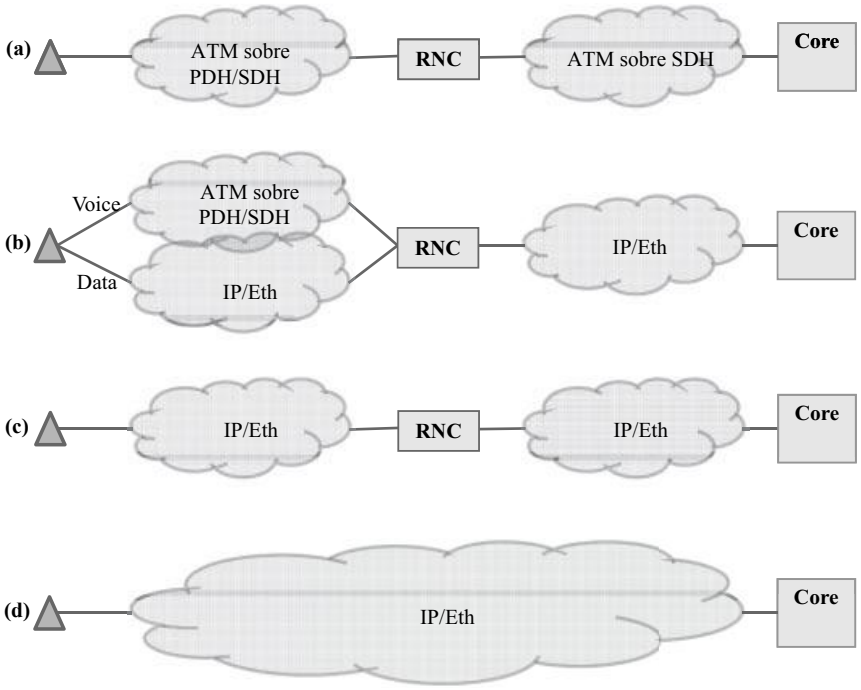

Figure 2.1: Distintas fases del backhaul: (a) Red 3G con transporte ATM sobre PDH/SDH: (b) red de transporte basada en conmutación de paquetes desplegada en paralelo con una red de tipo TDM; (c) todo la red de transporte pasa a conmutación de paquetes; (d) red de transporte para LTE basada en conmutación de paquetes.

a protocolos de capa radio ya que estos terminan en el eNode B. La Figura 2.2 resume las entidades e interfaces más relevantes de la red de acceso LTE. Los eNode B se conectan, a través del interfaz S1-u, a los Serving Gateways (S-GW) que le dan acceso a la red nucleo (*core*) en el plano de usuario. Se introduce un nuevo interfaz, el X2, que permite la conexión directa entre eNBs. Este nuevo interfaz permite, por ejemplo, realizar un handoff entre eNBs, traspasando el tráfico del plano de usuario a través del interfaz X2, mientras que los mensajes asociados de gestión de movilidad (plano de control) se enviarían a la entidad de control de movilidad (MME) por el interfaz S1-c.

A diferencia de 3G, el backhaul de LTE está concebido con una arquitectura todo IP, o *flat* IP. Junto con la adopción de interfaces IP abiertos, Ethernet se ha convertido en el estándar de facto para capa 2 tanto en el backhaul de la red de acceso como en la red core. Esto también es aplicable en la infraestructura 3G de más reciente implantación, HSPA (también denominada 3.5G), pero no en la generación previa de estándares 3G, en los que el soporte de interfaces IP se especificó como opcional. Este aspecto es importante, ya que el ritmo de adopción de nuevas

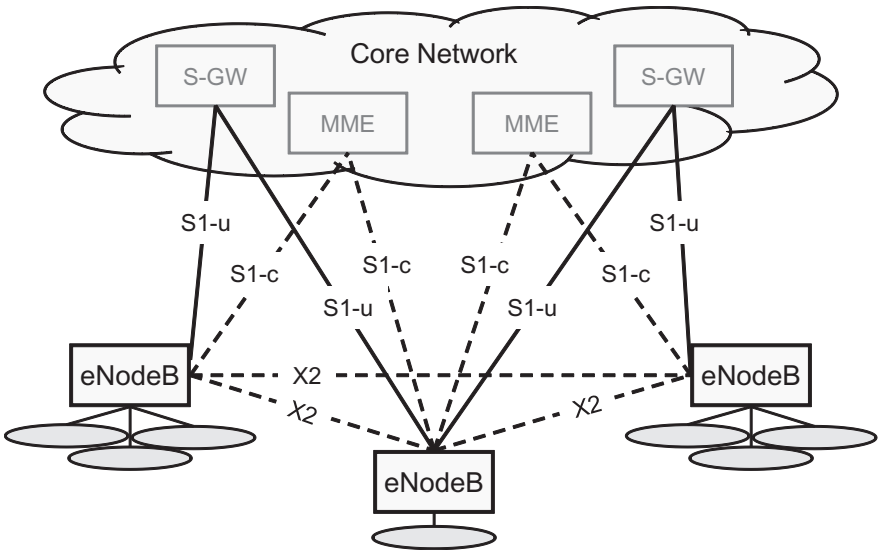

Figure 2.2: Resumen de las entidades e interfaces definidos en la red de acceso LTE.

tecnologías por parte de los usuarios es más lento que el ritmo de implantación de las nuevas redes. Por tanto, el despliegue de redes móviles de cuarta generación no implica el desmantelamiento de las redes de tercera generación, al igual que el despliegue 3G no implicó en su momento el apagado de las redes 2G. Los operadores han garantizar la coexistencia de varias redes también en el backhaul, lo que impone un requerimiento adicional al equipamiento de red: el soporte de interfaces para *legacy* networks, en particular ATM y PDH. Existen fundamentalmente dos opciones para facilitar la coexistencia: diseño integrado o diseño overlay. En el primer caso, una red de transporte de nueva generación (IP/Ethernet) constituye la infraestructura backhaul subyacente, y los routers en los extremos de la red ofrecen a las estaciones radio y a los nodos del core network, interfaces TDM, ATM o Ethernet, en función de sus especificaciones. Los routers transportan estos flujos heterogeneos mediante el servicio *pseudowire emulation* de IP/multi protocol label switching (MPLS). En el diseño overlay el backhaul LTE se diseña como una red overlay sobre la infraestructura backhaul existente para 2G/3G, por ejemplo SDH. La elección entre las dos opciones de coexistencia es esencialmente una decisión estratégica de los operadores, determinada por factores como la inversión realizada en infraestructura backhaul y el grado de penetración de LTE respecto a las redes de generaciones anteriores.

El interfaz radio de LTE está concebido para operar con diversos anchos de

banda espectrales, configuraciones de antena y modulaciones. Todo ello influye en los requerimientos de ancho de banda del backhaul. Por ejemplo, considerando una estación con 3 sectores, operando en un canal de 10 MHz y con una eficiencia espectral de hasta 8.64 b/s/Hz, da como resultado una tasa de pico de $8.64 \times 10 \times 3 = 259$ Mbps. No obstante el factor de reutilización de frecuencia en LTE (que puede llegar a ser tan bajo como 1) hace que sea muy improbable que se alcance la máxima eficiencia espectral en todas las subportadoras de todas las celdas simultáneamente, por lo que desde una perspectiva práctica resulta más realista asumir una tasa de pico de 200 Mbps para la estación considerada. Naturalmente, el dimensionamiento de los enlaces donde se agrupa tráfico se ha de aplicar un factor de multiplexación estadística para optimizar el coste de la red. El diseño de este factor ha de tener en cuenta que gran parte del tráfico LTE es de datos y por tanto del tipo *non-real-time*. Los requerimientos del tráfico *real-time* deberán ser garantizados por los mecanismos de calidad de servicio proporcionados por DiffServ o técnicas similares. La consecuencia conocida de un diseño ajustado por un factor de multiplexación estadística es la probabilidad no nula de congestión en la red, asociado a incrementos abruptos del retardo (*jitter*) y a la pérdida de paquetes en el backhaul. Estos efectos ya se daban en las redes 3G, pero la introducción de LTE y LTE-A ha dirigido la atención de la industria en esta problemática, posiblemente porque la inversión en backhaul supone un porcentaje mucho mayor que en el pasado.

El impacto de las Small Cells

Para poder garantizar la cobertura y la calidad de servicio deseada a todos los terminales, incluyendo los que están en los límites de alcance de las macro cells, las redes más avanzadas adoptan una estructura más flexible y escalable caracterizada por la introducción de las celdas pequeñas (small cells). El término small cell se refiere a estaciones con un alcance de transmisión inferior al de las estaciones macro (300-2000m), y que en función de su alcance se pueden denominar femtoceldas (10-30m), pico-celdas (30-100m), micro-celdas (100-200m). La cobertura de las small cells se solapa con la de las macro cell (y potencialmente con otras small cells) creando una arquitectura con distintos niveles (multi-tier) denominada genéricamente red heterogenea (HetNet). La Figura 2.3 ilustra la evolución hacia redes heterogéneas, en un proceso denominado *densificación de la red*. Son varias las ventajas que aporta esta arquitectura: 1) una mayor reutilización espacial del espectro radioeléctrico, aumentando así su eficiencia, 2) aumenta la probabilidad de que un terminal encuentre una estación cercana, mejorando su calidad de ser-

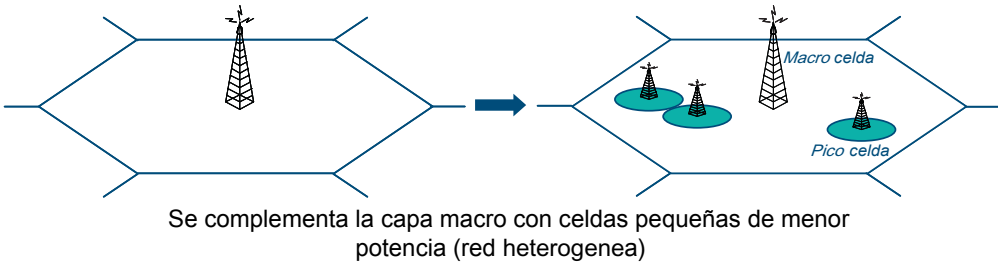

Figure 2.3: Ejemplo de densificación de la red para aumentar la capacidad del sistema y proporcionar mayores tasas de transmisión.

vicio y reduciendo el consumo de batería, 3) libera recursos de la macro cells.

Al contrario de lo que sucede con las macro cells, el despliegue de small cells (especialmente al nivel de pico y femto-celda), no está planificado, pudiendo ser los propios usuarios los que instalan de forma autónoma pico o femto-celdas en su propiedad. Por ejemplo, el organismo de estandarización 3GPP, integrado por fabricantes y operadores, denomina Home e-Node B (HeNB) a las femto-celdas domésticas en su informe TR 36.932. Para dotar de conectividad a estas small cells existen diversas opciones, que podríamos clasificar en tres tipos: el tipo 1, considerado el caso ideal, y basado en fibra óptica; el tipo 2, no ideal, y compuesto por líneas DSL, y enlaces microondas; y el tipo 3, para casos en los que no existe conectividad cableada ni tampoco linea de vista (LOS) para implementar un enlace de microondas. En este caso se contempla la posiblidad de dotar de un enlace backhaul NLOS mediante el propio interfaz radio LTE.

2.4 Funcionalidades del Backhaul

Redes 3G

En redes 3G se emplea el Frame Protocol para encapsular otros protocolos a través del interfaz Iub, como muestra la Figura 2.4, y también en el Iur. Este protocolo también se encarga de la señalización del control de potencia en lazo abierto (outer loop power control), y del alineamiento temporal. Esta última funcionalidad es objeto de estudio en esta tesis. El alineamiento temporal permite ajustar el tiempo de recepción de una trama FP downlink a la temporización del interfaz radio. Mediante este procedimiento, el Nodo B puede indicar a la RNC que adelante o retrase el envío de tramas FP. En el Nodo B se define una ventana temporal dentro

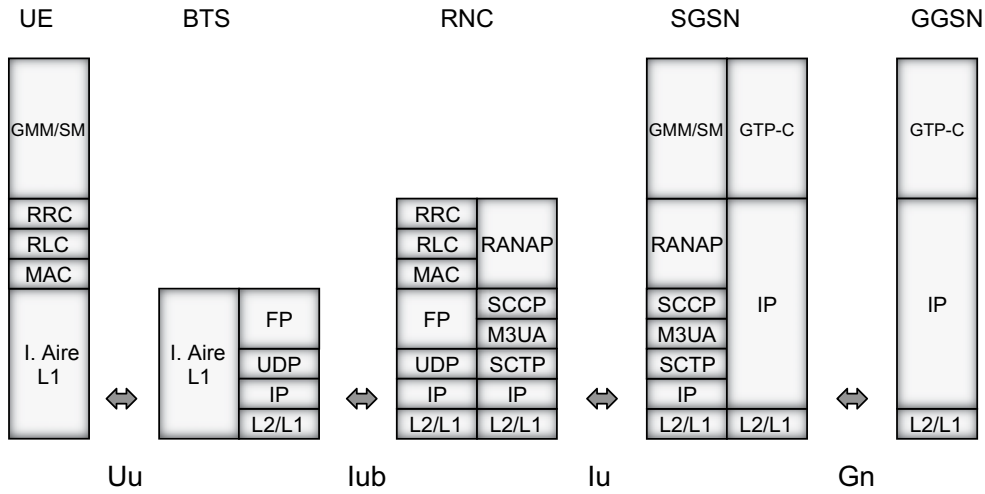

Figure 2.4: Pila de protocolos para servicios de conmutación de paquetes en redes 3G UMTS.

de la cual se espera que lleguen las tramas FP procedentes de la RNC. Las tramas que lleguen más tarde son descartadas.

En HSDPA, la capa FP incorpora la funcionalidad de petición y asignación de capacidad (control de flujo HSDPA). El Nodo B determina la cantidad de datos que la RNC puede enviar a través del Iub. Esta capacidad se especifica en los siguientes términos: cantidad de datos o número de PDUs, intervalo de tiempo durante el cual se ha de transmitir esta cantidad de datos, y tamaño máximo de las PDUs. La Figura 2.5 resume la pila de protocolos y las novedades introducidas en HSDPA con respecto a las versiones 3G anteriores.

Redes 4G

En LTE se simplifican las funcionalidades asociadas al backhaul en comparación a los sistemas 3G. El servicio extremo a extremo consiste en una portadora EPS (establecida entre el terminal de usuario y el PGW) y una portadora externa (por ejemplo Internet). A su vez, la portadora EPS se compone de la portadora E-RAB y el interfaz S5/S8. La portadora E-RAB se establece entre el terminal de usuario el SGW, y comprende la portadora radio y el interfaz S1, y es el ámbito de estudio en esta tesis. El diagrama de la Figura 2.6 muestra los nodos y los niveles en los que se establece cada portadora en LTE.

Cada portadora EPS está caracterizada por unos parámetros de calidad de servicio. Estos parámetros son el Quality of sevice Class Indicator (QCI), el Allocation and Retention Priority (ARP), el Guaranteed Bit Rate (GBR) y el Maximum

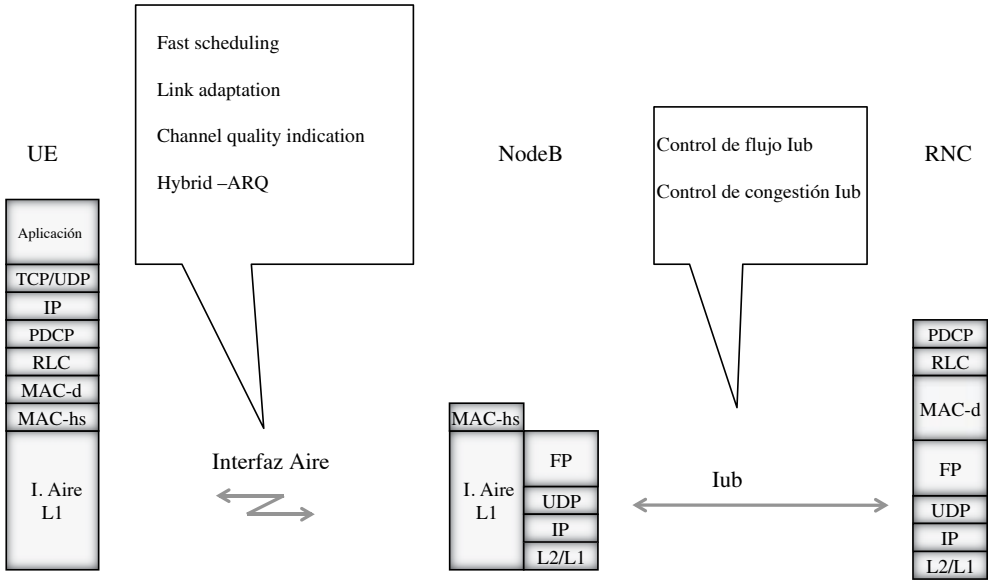

Figure 2.5: Novedades introducidas por HSDPA en la pila de protocolos 3G.

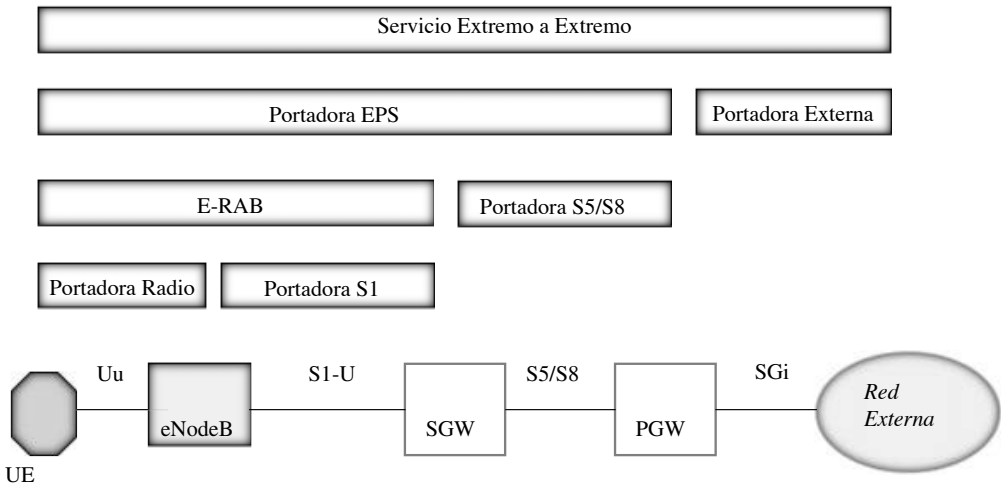

Figure 2.6: Esquema resumen de las portadoras de 4G LTE.

Bit Rate (MBR). Estos parámetros determinan la política de scheduling, gestión de colas y configuración de RLC que experimentará el tráfico mapeado en dicha portadora EPS. De acuerdo a estas características de QoS, el establecimiento de portadoras está sujeto a control de admisión en todos los nodos (eNode B, SGW, PGW). En el plano de usuario del E-RAB, únicamente es obligatorio garantizar los parámetros QCI y ARP.

Las distintas clases de QCI están estandarizadas por el 3GPP en TS23.203 [15],

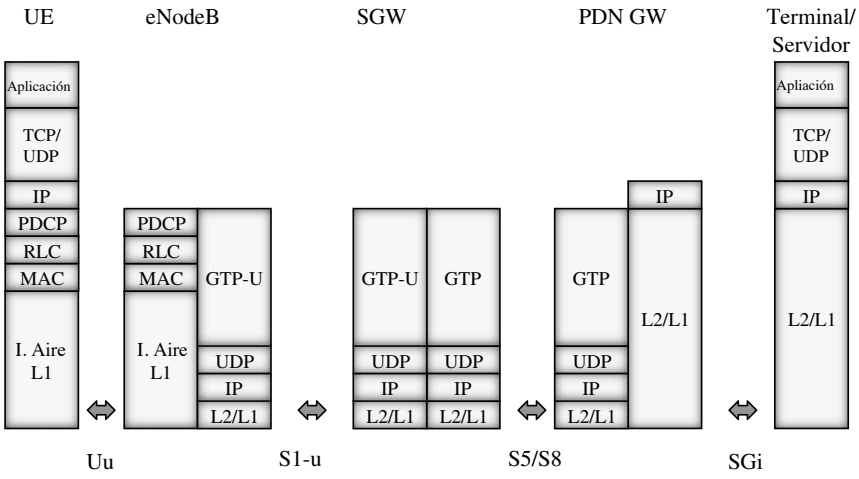

Figure 2.7: Pila de protocolos en LTE (4G).

y están asociadas, entre otros aspectos a tolerancias máximas de retardo, niveles de prioridad y tasas máximas de pérdidas de paquete, o Packet Error Loss Rate (PELR). El valor de PELR asume que las pérdidas de paquetes en backhaul (en el interfaz S1) son despreciables. Por tanto, para que el sistema no supere este límite, el backhaul debe perder una cantidad prácticamente nula de paquetes, lo que implica que el backhaul esté bien dimensionado y libre de congestión. No obstante, la congestión en la red de backhaul puede causar pérdidas por descarte de paquetes en las colas de salida de los nodos RAN o en los routers intermedios. También pueden producirse pérdidas de paquetes si se pierde temporalmente la conectividad backhaul, por ejemplo, debido a caídas de enlaces o reconfiguración de la red.

Al contrario que los interfaces de acceso de las redes 2G y 3G, el interfaz S1 no transporta información de protocolos de capas radio, ya que todos ellos terminan en el eNode B, como se muestra en la Figura 2.7 Los requerimentos de retardo, variación de retardo, y pérdidas de paquetes en dicho interfaz derivan de los objetivos de calidad extremo a extremo comprometidos con el usuario, y no de restricciones impuestas por la propia arquitectura del sistema móvil, como era el caso de las generaciones anteriores.

Debido a lo anterior, ante la pérdida de un paquete del plano de usuario, el interfaz S1 reacciona de forma muy distinta al interfaz Iub (asumiendo un Iub en el que la capa RLC opera con reconocimiento). Los paquetes perdidos en el interfaz Iub son retransmitidos por la capa RLC mientras que, en LTE, la capa RLC termina en el eNode B. Esto implica que una pérdida de paquete en S1 es

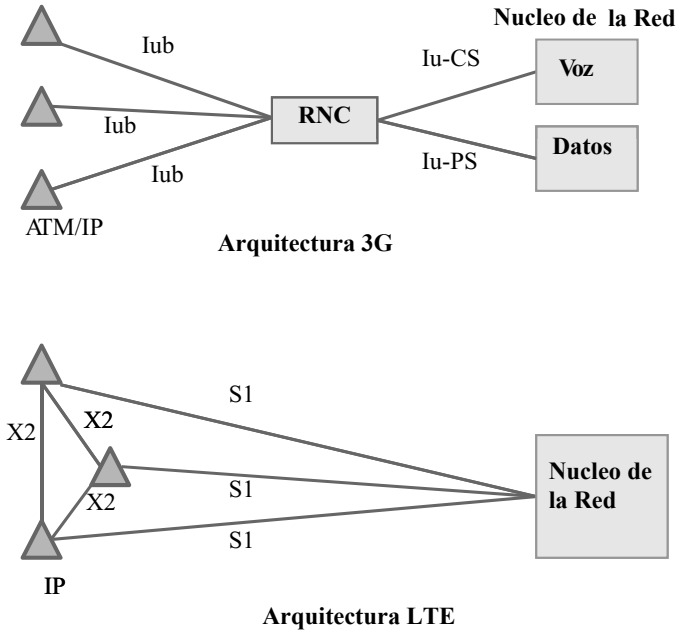

Figure 2.8: Esquemas de las arquitecturas de red de acceso en 3G y en 4G.

visible a las capas superiores, ya que no hay ningún protocolo que enmascare estas pérdidas mediante retransmisiones. En este aspecto el interfaz S1 es comparable al interfaz Iu de 3G. Si la aplicación emplea una capa de transmisión fiable como TCP, una pérdida en S1 causaría una retransmisión TCP. Las diferencias en la arquitectura de las redes de acceso 3G y 4G se observan de forma esquemática la Figura 2.8

2.5 Problemas Abiertos

LTE ha simplificado notablemente la arquitectura RAN al concentrar todas las funcionalidades radio en los eNBs y al imponer interfaces IP abiertos. Los interfaces S1 transportan datos de usuario mediante túneles GTP (GPRS Tunneling Protocol) individuales, por lo que la gestión del ancho de banda y la QoS en el backhaul se puede realizar a nivel de flujo de datos de usuario. Este grado tan alto de granularidad permite mejorar la experiencia de usuario, aunque desde la industria aún es necesario estudiar el compromiso entre complejidad computacional y calidad de servicio para decidir sobre su viabilidad comercial. Debido a esto, esta tesis pone un énfasis especial en el desarrollo de mecanismos de control con una baja carga computacional.

La nueva evolución de 4G, LTE-Advanced introduce novedades que deberán tenerse en cuenta en el diseño del backhaul: 1) Tasas de pico de hasta 1 Gbps en el downlink y 500 Mbps en el uplink, 2) transmisión y recepción empleando Coordinated Multipoint (CoMP), y 3) relays de capa 3 denomidados relays de auto-backhauling. El aumento de la tasa de pico afectará al dimensionamiento y la densidad de despliegue de los nodos que componen la RAN y su backhaul. Dependiendo del tipo de funcionalidad CoMP que se implemente, puede ser necesario que el procesamiento de la señal en banda base está separado físicamente de las antenas, por lo que habría que proporcionar también conectividad backhaul desde las antenas al Nodo B donde se realiza el procesado en banda base. El concepto de relay de capa 3 se introduce en la especificación TR 36.806. En contraste con los repetidores pasivos empleados en las redes móviles actuales, los relays de capa 3 operan de manera coordinada con los eNb solo cuando es necesario para aumentar la tasa disponible en el terminal. Estos relays operan con un principio de auto-backhaul, en el que el interfaz radio LTE se emplea para transportar el trafico entre el relay y el eNb. Esta estrategia tiene la ventaja de que no requiere ningún medio de backhaul adicional, pero consume los recursos espectrales de la red.

CHAPTER 3

Sincronización de Canal de Transporte

Resumen:

El objetivo principal de la Sincronización de Canal de Transporte en redes UMTS es que las tramas enviadas por la RNC lleguen a tiempo a los Nodos B para su transmisión a través del interfaz radio. Para ello, el 3GPP especifica un algoritmo conocido como *Timing Adjustment* que se encarga de controlar el retraso que experimentan las tramas en el interfaz Iub sumando o restando una cantidad constante. Como se demostrará en este capítulo, este algoritmo reacciona con demasiada lentitud ante variaciones abruptas del retardo del interfaz Iub y, además, se puede volver inestable en escenarios de alta demora. Aplicando teoría de control en tiempo discreto a este problema, se propone un nuevo mecanismo que garantiza la estabilidad en cualquier situación y mejora el rendimiento del algoritmo clásico. Las medidas de rendimiento se realizan mediante un simulador de la red UTRAN teniendo en cuenta condiciones realistas de tráfico en el interfaz Iub.

3.1 Introducción

Arquitectura de la red UMTS

En esta sección se pretende dar una visión general de la red de acceso radio de UMTS, UTRAN, puesto que es el escenario concreto donde se aplicarán las técnicas de control discreto estudiadas en este trabajo. A continuación se explicarán las

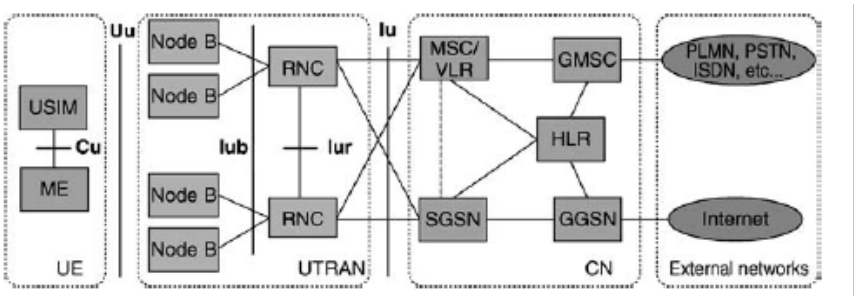

Figure 3.1: Arquitectura Global de una red UMTS.

características principales de UTRAN las cuales se describen en las especificaciones del 3GPP, organismo internacional de estandarización que aglutina a numerosos fabricantes y operadores.

La infraestructura de una red de telefonía móvil de tercera generación UMTS consta de tres partes principales, tal y como se puede observar en la figura 3.1:

1. El equipo de usuario UE, constituido por el terminal móvil de usuario.
2. La red de acceso UTRAN, que es la parte de la red que permite la conexión de los usuarios al sistema a través del interfaz radio, denominado Uu, y donde se ubican las entidades que gestionan los mecanismos que permiten a los usuarios acceder a los servicios.
3. La red troncal CN, donde se realiza la conmutación y enrutado de las llamadas y conexiones de datos desde y hacia redes externas. Además, se gestionan los servicios y funcionalidades de la red, como la tarificación, la autentificación y la movilidad.

En la figura 3.1 se observa que la red de acceso está formada por un conjunto de subsistemas RNS conectados al CN a través del interfaz Iu. Cada subsistema RNS está formado a su vez por un controlador RNC y una serie de Nodos B, que dan soporte al interfaz radio, Uu, a través del cual se conectan los terminales de usuario, UE. La RNC es responsable de todas las funciones relacionadas con la asignación y gestión de recursos radio y del mantenimiento de las condiciones de calidad de servicio, QoS. El Nodo B, por su parte, realiza los procedimientos de la capa física, tales como la sincronización y el control de potencia.

Los Nodos B y las RNC se conectan mediante el interfaz Iub, por el que intercambian información de control y transportan el tráfico de los usuarios. Los subsistemas RNS están conectados entre sí mediante los interfaces Iur, que permiten la movilidad del usuario entre distintos subsistemas. Los interfaces Iu, Iub

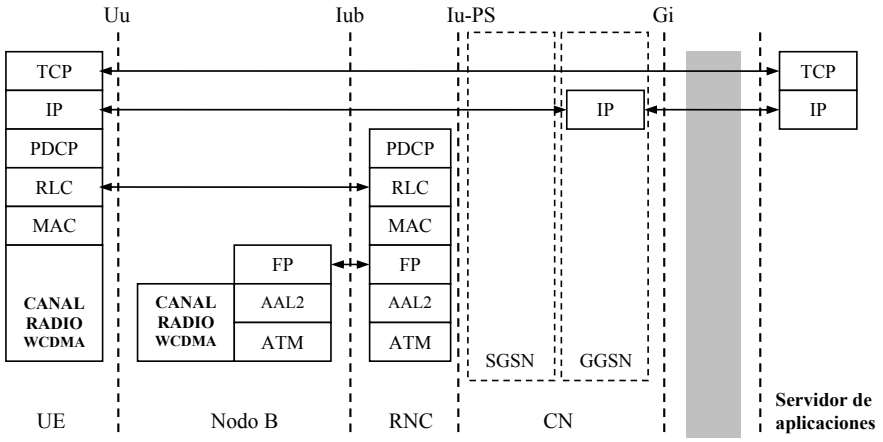

Figure 3.2: Pila de protocolos del plano de usuario involucrados en UTRAN y el terminal.

e Iur son interfaces lógicos, lo cual quiere decir que pueden estar soportados tanto por un enlace físico directo como por una red de transporte adecuada. En las primeras versiones de la estandarización del sistema, la tecnología subyacente de esta red de transporte se especificó como ATM pero a partir de las especificaciones *Release 5* se introduce IP como alternativa de transporte, de cara a una futura evolución hacia una red todo IP.

Respecto a los protocolos, en UMTS se definen dos pilas de protocolos en paralelo, una para el plano de usuario y otra para el plano de control. En el plano de usuario se definen todos los protocolos involucrados en la transferencia de tráfico, y en el plano de control los protocolos involucrados con la gestión de recursos, señalización, etc. El plano de usuario se puede presentar, a su vez, en dos versiones, una correspondiente a los servicios de conmutación de circuitos (por ejemplo servicio de voz) y otra para los servicios de conmutación de paquetes. En la Figura 3.2 puede observarse la pila de protocolos del plano de usuario para servicios de conmutación de paquetes que constituye el ámbito en el que se desarrolla el estudio realizado en este trabajo.

En la Figura 3.2 centraremos nuestra atención en la pila de protocolos ubicada en la RNC para la comunicación con el terminal de usuario y el Nodo B. Con el terminal de usuario se establecen tres niveles de protocolos: PDCP, RLC y MAC. Por su parte, la comunicación entre RNC y Nodo B se realiza mediante un protocolo propio de UTRAN: FP, por debajo del cual se ubican los protocolos de la red de transporte que, para este ejemplo concreto, se ha supuesto de tecnología ATM. En este caso, los flujos de tramas FP con tráfico de usuario se encapsulan

en celdas ATM con adaptación AAL2. Los protocolos PDCP, RLC, MAC y FP se definen a continuación:

- PDCP multiplexa flujos de tráfico de un usuario, procedentes de protocolos de nivel de red, para ser transmitidos de forma transparente sobre la red de acceso UTRAN. Además, el protocolo PDCP puede mejorar la eficiencia de la transmisión mediante la aplicación de mecanismos de compresión de cabeceras. Este protocolo está especificado en [16].
- RLC realiza la transferencia de paquetes procedentes de la capa superior a través del enlace radio. El tipo de servicio ofrecido es configurable, dando lugar a tres modos de operación de RLC: 1) RLC TM, correspondiente al Modo Transparente (Transparent Mode), en el que la información de la capa superior es transmitida sin añadir ninguna cabecera del protocolo RLC, 2) RLC UM, Modo No Reconocido (Unacknowledged Mode) en el que las tramas contienen información adicional que permite la detección de errores, pero dichos errores no se recuperan, y 3) RLC AM, Modo Reconocido (Acknowledged Mode) donde se implementa un mecanismo de recuperación de errores de tipo SR ARQ. Los tres modos de operación están descritos en la especificación [17].
- MAC se encarga de controlar la cantidad de información que puede transmitir la entidad RLC en cada instante de transmisión del canal radio, TTI. Este control se realiza a partir de las indicaciones de un protocolo de nivel superior ubicado en el plano de control, el protocolo RRC encargado del control de los recursos de UTRAN. El protocolo RRC recoge medidas de carga de tráfico y calidad del canal radio para cada usuario, y a partir de esas medidas especifica la tasa de transferencia máxima permitida para cada usuario en cada TTI. A partir de estas medidas y teniendo en cuenta la cantidad de datos a transmitir por el usuario y los mecanismos de planificación o scheduling establecidos, se escoge una cantidad concreta de tramas RLC transferibles a la subcapa MAC. Esta cantidad de tramas, junto con la información de señalización disponible, determina el conjunto de bloques de transporte, TBS, que la subcapa MAC intercambia con el nivel físico cada TTI (10ms). Es importante resaltar que las transferencias entre la subcapa MAC y la capa física, se realizan a través de lo que se conoce como canal de transporte. Hay dos grandes grupos de canales de transporte: los canales dedicados, de los que sólo hay un tipo, DCH, que transportan información a un sólo usuario, y los canales compartidos, de los que hay varios tipos, como

por ejemplo el canal compartido ascendente, el canal compartido descendente y el canal de acceso aleatorio, entre otros.

- FP es el protocolo encargado, entre otras funciones, de transferir los TBS entre la RNC y el Nodo B a través del interfaz Iub. Debe tenerse en cuenta que la trama radio existe únicamente entre el terminal de usuario y el Nodo B. Por tanto, en sentido downlink, los TBS deben transferirse entre la subcapa MAC de la RNC y la capa física del Nodo B (canal radio). En sentido uplink, el Nodo B decodifica la trama radio del usuario y extrae el TBS que debe llegar a la subcapa MAC de la RNC. El protocolo FP se describe en las especificaciones [1] y [2].

En el ámbito de este trabajo se asume un servicio de transferencia de datos dentro del dominio de conmutación de paquetes, ofrecido sobre un canal de transporte dedicado DCH.

Problema analizado: Sincronización de Canal de Transporte

Debido al jitter de la red de transporte del interfaz Iub, ya sea ATM o IP, los flujos de tramas a través de canales dedicados DCH entre RNC y Nodos B encuentran retardos aleatorios que hacen necesaria una sincronización entre ambos para que las tramas lleguen al Nodo B a tiempo de ser enviadas en su correspondiente TTI. Este proceso se conoce como Sincronización de Canal de Transporte (*Transport Channel Synchronization*), se aplica sólo en sentido downlink (de la RNC al Nodo B) y de él se encarga el protocolo FP (especificado por el 3GPP en [1] y [2]) que controla la comunicación a nivel de enlace entre la RNC y los Nodos B. Este protocolo trabaja de forma independiente para cada canal de transporte DCH, encapsulando los bloques de transporte TB generados por la capa MAC en tramas FP y enviando éstas a través de la red de transporte del Iub.

El procedimiento para conseguir la Sincronización de Canal de Transporte se denomina *Timing Adjustment* y consiste en ajustar el instante de envío de las tramas de datos desde la RNC al Nodo B con el fin de hacerlas llegar dentro de una ventana de recepción predefinida. Como se puede observar en la figura 3.3, esta ventana de recepción se define independientemente por cada canal de transporte DCH mediante dos parámetros: TOAWS y TOAWE. La duración temporal de esta ventana debe ser suficientemente pequeña con el fin de minimizar el tiempo de almacenamiento de las tramas en los buffers del Nodo B ya que este almacenamiento debe realizarse en la RNC, donde están localizados los mecanismos de

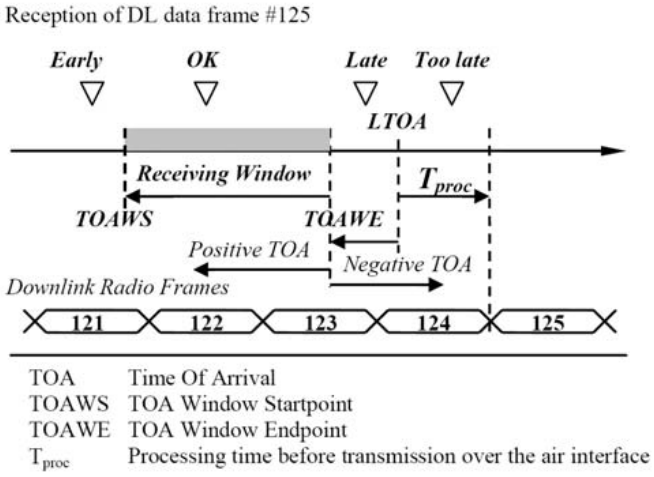

Figure 3.3: Instantes de recepción de tramas en el Nodo B.

scheduling y de gestión de los recursos radio. Por otra parte, reducir el tamaño de la ventana incrementa el tráfico de senalización. Por tanto, la configuración de estos dos parámetros influye en el rendimiento del sistema.

La figura 3.3 muestra los posibles instantes de llegada de tramas y las referencias temporales. Cuando una trama llega fuera de su ventana, el Nodo B responde con una trama de ajuste de tiempo TA que contiene el TOA de la trama recibida. La RNC utiliza esta información como realimentación para ajustar el instante de envío de las siguientes tramas, tratando de ajustar el tiempo de llegada de las futuras tramas dentro de la ventana. Por lo tanto, por cada trama TA, la RNC modifica la estimación del retardo (*offset*) del correspondiente canal DCH. Si una trama llega después del instante LTOA, el Nodo B no será capaz de procesarla antes del TTI de dicha trama y, por tanto, será descartada. Se puede encontrar una explicación más detallada de este proceso de señalización en [2].

El algoritmo original especificado por el 3GPP para realizar este procedimiento de *Timing Adjustment* consiste en sumar o restar una cantidad constante a la estimación del retardo (*offset*) que hace la RNC en función del valor del parámetro TOA que recibe en las tramas TA:

$$\begin{aligned} t_{n+1} &= t_n + t_{TTI} - K \quad if \quad \text{TOA} < 0 \\ t_{n+1} &= t_n + t_{TTI} + K \quad if \quad \text{TOA} > 0 \end{aligned} \tag{3.1}$$

donde t_{TTI} es la duración de un TTI y t_n es el instante de transmisión de la trama n. En el resto del capítulo, nos referiremos a este algoritmo como *algoritmo clásico*. En [4], se evalúa el efecto de los parámetros de la ventana sobre la capacidad del

enlace Iub para tráfico de voz considerando este algoritmo pero no se propone ninguna mejora sobre el mismo. En cambio, el trabajo presentado en [3] ofrece un método alternativo de ajuste considerando una modificación del algoritmo clásico según el cual el *offset* se decrementa lentamente en ausencia de tramas TA pero cuando llega una de ellas el *offset* se incrementa de forma pronunciada (principio decremento-aditivo, incremento-multiplicativo). Sin embargo, en un escenario con cambios bruscos de retardo, su funcionamiento es análogo al algoritmo clásico.

En este capítulo, se propone como mecanismo de *Timing Adjustment*, un algoritmo de seguimiento proporcional, PTA. Esta propuesta se fundamenta en la corrección del *offset* proporcionalmente a la distancia entre el TOA y el centro de la ventana. Como se mostrará a lo largo de este capítulo, PTA mejora sensiblemente el rendimiento (bajos niveles en los buffers de los Nodos B, menos pérdidas y menos tramas de señalización), añade robustez y simplifica la configuración en comparación con el algoritmo clásico. Esta nueva propuesta no requiere ningún cambio en las especificaciones 3GPP.

El resto del capítulo se organiza de la siguiente forma:

La sección 3.2 describe el algoritmo PTA como un sistema de control en tiempo discreto realizando su correspondiente análisis de estabilidad. La sección 3.3 muestra los resultados del estudio de simulación realizado sobre la configuración y evalucación de ambos algoritmos. Finalmente, en la sección 3.4 se resumen las aportaciones y conclusiones de este trabajo.

3.2 Modelo de Control en Tiempo Discreto

En esta sección se presenta un algoritmo alternativo basado en un modelo de control clásico para llevar a cabo la Sincronización de canal de transporte: el algoritmo de seguimiento proporcional PTA.

Como se explica en la Sección 3.1, cuando llega una trama de datos fuera de la ventana de recepción, el nodo B envía una trama de señalización TA que indica el TOA de esa trama de datos. Hasta ahora ningún algoritmo hacía uso de esta información. El algoritmo PTA utiliza este TOA para programar la transmisión de las siguientes tramas de datos desde la RNC. El objetivo de PTA es ajustar el tiempo de llegada de las tramas al centro de la ventana. Se elige este instante porque minimiza la señalización y reduce el tiempo de almacenamiento en el Nodo B. De acuerdo a este principio, tras la llegada de una trama TA, la RNC realiza la siguiente corrección en el instante de envío de la siguiente trama:

$$t_{n+1} = t_n + t_{TTI} + K \cdot (\text{TOA} - \text{TOAWS}/2) \tag{3.2}$$

es decir, el envío de la siguiente trama se adelantara o se atrasara K veces la diferencia entre el centro de la ventana y el instante de llegada de la anterior trama. Si la trama llega después del final de la ventana, TOA − TOAWS/2 es negativo y, por tanto, el envío de la siguiente trama se adelantará. Este avance implica un incremento similar en la estimación del desplazamiento realizado en la RNC. Por otro lado, si la trama llega antes del inicio de la ventana, el envío de la siguiente trama será retrasado. Mientras las tramas lleguen dentro de la ventana, no se envía señalización y, por tanto, la RNC continuará transmitiendo con la misma sincronización.

La figura 3.4 ilustra el funcionamiento de PTA con los retardos que participan en el proceso. Se define el RTT del sistema de control de FP como el tiempo que transcurre entre el instante de envío de una trama y el momento en el que su correspondiente trama de corrección TA se aplica al envío de otra trama. Representando el número de TTIs que ocurren dentro de un RTT como $r = \lceil RTT/t_{TTI} \rceil$ y definiendo j y m como los retardos de bajada y de subida respectivamente medidos en TTIs, $r = j + m$, se puede expresar 3.2 como:

$$\text{offset}_{n+1} = \text{offset}_n + K \cdot e_n \tag{3.3}$$

donde e_n es el error entre el retardo (offset) aplicado a la trama $n-r$ (offset_{n-r}) y el retardo medido en el slot temporal $n - m$ (delay_{n-m}), así

$$e_n = \text{delay}_{n-m} - \text{offset}_{n-r}. \tag{3.4}$$

El valor de r tiene un fuerte impacto sobre la estabilidad y las prestaciones del sistema, como veremos más adelante.

Análisis de la estabilidad del algoritmo

En esta sección se hace uso de la teoría de de control en tiempo discreto para estudiar la estabilidad del algoritmo propuesto PTA así como para ajustar la ganancia K de acuerdo con el rendimiento deseado a la respuesta escalón (que modela un aumento brusco del retardo). Los resultados de la simulación mostrados en la sección 3.3 confirman las mejoras de PTA en comparación con el algoritmo clásico.

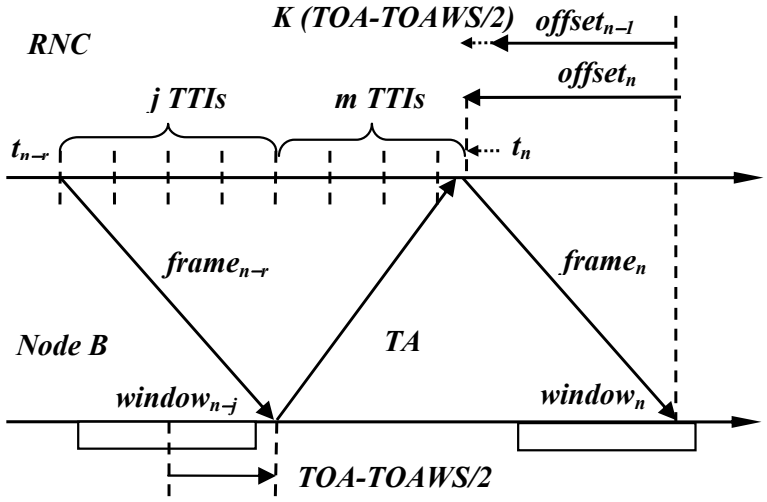

Figure 3.4: Diagrama de tiempos del mecanismo de ajuste temporal.

Tanto PTA como el algoritmo clásico son sistemas donde las señales involucradas, el retardo, el error y el desplazamiento, toman un valor por trama. Por lo tanto, podemos modelar este algoritmo como un sistema de control en tiempo discreto realizando las siguientes asunciones:

1. La RNC transmite una trama por TTI, es decir, no hay momentos sin tráfico de datos en el periodo de tiempo bajo estudio.
2. La RNC dispone de la información sobre el TOA en cada slot temporal. Esto implica que el tamaño de la ventana (TOAWS) es $w = 0$ (ver figura 3.4).
3. El tiempo entre envíos de tramas desde la RNC es constante e igual a t_{TTI}. Por tanto, el modelo omite la variación de la duración del TTI provocada por la corrección del offset en cada slot temporal.

La figura 3.5 muestra un diagrama del sistema de control del algoritmo PTA, donde la frecuencia de muestreo es $1/t_{TTI}$ y los bloques $1/z^j$ y $1/z^m$ representan respectivamente los retardos de bajada y subida. La señal $u(n)$, se corresponde con delay_n y actúa como la señal de control, mientras que $x(n)$ es el offset_n.

Expresando 3.3 en terminos de las señales de la figura 3.5 obtenemos:

$$x(n) = x(n-1) + K(u(n-m) - x(n-r)) \tag{3.5}$$

Tomando la transformada Z en 3.5 conseguimos:

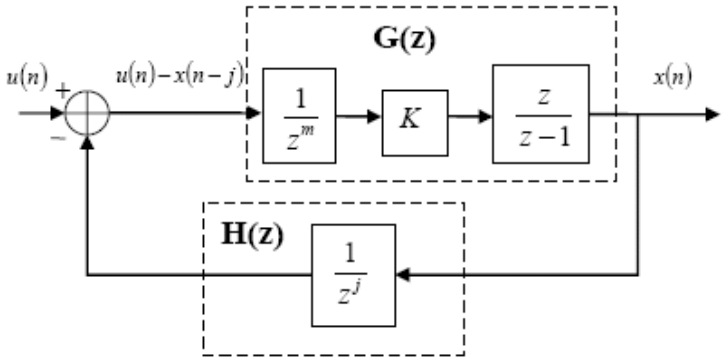

Figure 3.5: Modelo de control en tiempo discreto del algoritmo PTA.

$$X(z) = z^{-1}X(z) + K \cdot z^{-m}U(z) - K \cdot z^{-r}X(z) \tag{3.6}$$

Resultando la siguiente función de transferencia:

$$\frac{X(z)}{U(z)} = \frac{K \cdot z^{j}}{z^{r} - z^{r-1} + K} \tag{3.7}$$

Con el fin de ajustar la ganancia K, usamos un parámetro habitual en teoría de control [18], el porcentaje de *overshoot* del primer pico de la respuesta, M_p. Considerando una entrada escalón de la forma $u(n) = C$ para $n > 0$, y una salida cuyo primer pico es igual a A_{max}, la definición es:

$$M_p(\%) = \frac{A_{max}}{C} \cdot 100 \tag{3.8}$$

La ecuación anterior está relacionado con el ratio de amortiguación (ζ) porcentaje

$$\frac{A_{max}}{C} = e^{-\zeta\pi/\sqrt{1-\zeta^2}} \tag{3.9}$$

Usando 3.9 obtenemos el ratio (ζ) para un *overshoot* dado. La ganancia K se obtiene resolviendo la ecuación característica $1 + G(z) \cdot H(z) = 0$, la cual se puede expresar como $F(z) = -1$, donde $F(z) = G(z) \cdot H(z)$. Identificando $G(z)$ y $H(z)$ con el sistema de la figura 3.5 se obtiene:

$$F(z) = \frac{K \cdot z}{z^{r}(z-1)} = -1 \tag{3.10}$$

Table 3.1: Valores de la ganancia K para diferentes porcentajes de overshoot.

r	2%	5%	10%	15%	20%	$\zeta \to \infty$
2	0.313	0.346	0.389	0.428	0.464	1
3	0.186	0.207	0.233	0.257	0.280	0.618
4	0.133	0.140	0.167	0.184	0.200	0.445
5	0.103	0.114	0.129	0.143	0.156	0.347
6	0.084	0.094	0.106	0.117	0.127	0.285
7	0.071	0.079	0.089	0.099	0.108	0.241

Si $\zeta < 0$, los polos del sistema se pueden escribir como $z = exp(-2\pi\zeta\omega/\sqrt{1-\zeta^2})$ de acuerdo a [18], donde $\omega = \omega_d/\omega_s$ y ω_s es la frecuencia de muestreo en rad/s. En nuestro caso $\omega_s = 2\pi/t_{TTI}$, donde t_{TTI} está fijado a 10 ms en los canales considerados. Dado que $F(z)$ es una magnitud compleja, 3.10 se puede dividir en dos ecuaciones, una para la fase de $F(z)$, igual a π radianes, y la otra para el módulo que debe ser igual a 1. La condición establecida para la fase conduce a la ecuación:

$$(r-1)2\pi\omega + arg(C^{\omega}e^{j2\pi\omega} - 1) = \pi \tag{3.11}$$

Donde $C = e^{-2\pi\zeta/\sqrt{1-\zeta^2}}$. A partir de 3.11 se obtiene ω y aplicando este resultado a la condición del módulo se obtiene K:

$$K = C^{\omega(r-1)} \left| C^{\omega}e^{j2\pi\omega} - 1 \right| \tag{3.12}$$

Para el caso particular de $\zeta = 0$, se obtiene el valor crítico de K, es decir, la ganancia por encima de la cual el sistema se vuelve inestable. Estos valores se pueden comprobar mediante el test de estabilidad de Jury [18] para sistemas de tiempo discreto. En la tabla 3.1 se resumen los valores de K para distintos porcentajes de *overshoot* y distintos valores de retardo (r, expresado en número de TTIs).

A partir de este análisis teórico se deduce que la ganancia K debe ajustarse en función del offset medido en la RNC para que el algoritmo se mantenga estable. Por tanto, es necesario que K se actualice automáticamente de acuerdo a la estimación del retardo del Iub en sentido *downlink* realizado en la RNC. Esta operación requiere que el protocolo FP guarde un conjunto de valores precalculados

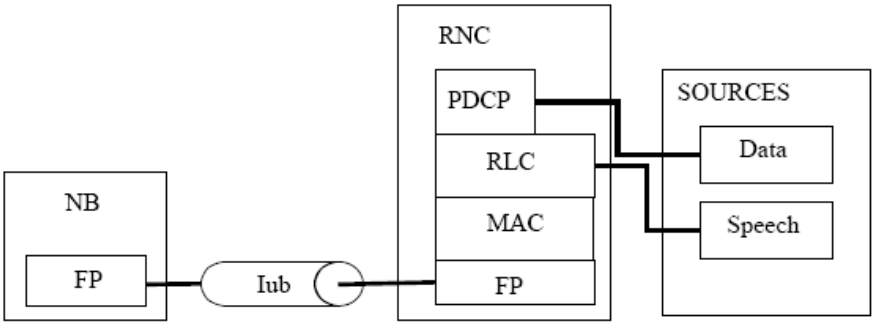

Figure 3.6: Topología de la red simulada.

de K. En la implementación final se opta por el conjunto de valores correspondientes al 10% de *overshoot*. El valor de K es seleccionado automáticamente de este conjunto en función del retardo estimado para el Iub por la RNC (*offset*), resultando de esta forma en un ajuste dinámico.

3.3 Resultados de simulación

Con el fin de evaluar las prestaciones de ambos algoritmos sobre canales dedicados DCH, se ha desarrollado una detallada implementación del protocolo FP, de acuerdo a las especificaciones del 3GPP [2] y [1], usando OMNeT++, una librería de simulación en C++ orientada a eventos [19].

En la figura 3.6 se puede observar la topología de la red que se ha simulado. Se trata de un enlace punto a punto simulando el interfaz Iub entre RNC y Nodo B cuyo ancho de banda es de 1.7 Mbit/s, aproximadamente la tasa de un enlace E1 PDH restando un 10% para tráfico de control y señalización.

Las simulaciones se han realizado teniendo en cuenta tanto tráfico de voz como de datos. Las fuentes de voz están basadas en codecs AMR mientras que cada fuente de datos representa el tráfico de una aplicación web. Se puede encontrar una descripción detallada de las características de ambos tipos de fuentes en la recomendación del 3GPP TR 25.933 [20]. Cada una de estas fuentes utiliza un canal DCH a una tasa constante de 384 Kbps.

Respuesta a cambios abruptos de retardo

En este apartado se estudia el comportamiento de ambos algoritmos en una situación de incremento abrupto del retardo del Iub. En este escenario, los resultados de

simulación muestran que PTA reacciona más rápido que el algoritmo clásico evitanto pérdidas excesivas, reduciendo la señalización y al mismo tiempo manteniendo los niveles de los buffers del Nodo B suficientemente bajos. Además, el algoritmo clásico puede presentar problemas de estabilidad dependiendo del valor de la ganancia K, requiriendo una configuración muy precisa. Estos cambios bruscos en el retardo del Iub se corresponden con situaciones reales que pueden ser causadas por varias razones:

- Una conexión/desconexión simultánea de varias fuentes en un mismo Nodo B provocando un fuerte incremento/decremento de tráfico hacia ese Nodo B.
- Un nodo intermedio de la red de transporte UTRAN que esté manejando una gran cantidad de tráfico.
- Cuando se produce un re-encaminamiento de las conexiones que circulan por el Iub. Ver [21] y sus referencias.

En los ejemplos expuestos en [22], el retardo en un solo sentido de diferentes caminos/ramas del interfaz Iub varían desde 12 hasta 31 ms considerando tráfico de voz y transporte ATM. Esto nos da una idea de las diferencias de retardo que se pueden alcanzar en la red que conecta los nodos de acceso 3G. Si se consideran servicios de datos y transporte IP, estos retardos pueden ser mucho mayores. En este trabajo, se han considerado cambios de retardo entre 10 ms y 60 ms.

En esta sección, se evalúan dos escenarios donde el retardo de la interfaz Iub experimenta un cambio brusco: un incremento y un decremento. En estos experimentos se considera una fuente de voz activa y los parámetros *TOAWS* y *TOAWE* se han fijado a sus valores típicos [4], 10 ms y 5 ms respectivamente.

En la figura 3.7 se observa la reacción de PTA y del algoritmo clásico con $K = 1$ ms y $K = 3$ ms en un escenario de incremento del retardo. También se muestra la variación del parámetro TOA de la conexión FP. La interfaz Iub se configura inicialmente con un retardo de propagación de 10 ms. En el instante de simulación $t = 50$ ms, dicho retardo del Iub cambia a 50 ms. Este cambio implica que las tramas empiezan a llegar al Nodo B después de la ventana de recepción provocando la generación de tramas TA (*signaling events*) y el descarte de las tramas que llegan demasiado tarde (*loss events*), como se muestra en la figura 3.7. En consecuencia, el buffer del Nodo B permanece vacío hasta que el instante de envío de nuevas tramas se ajusta en la RNC para que lleguen a tiempo de ser transmitidas a través de la interfaz radio.

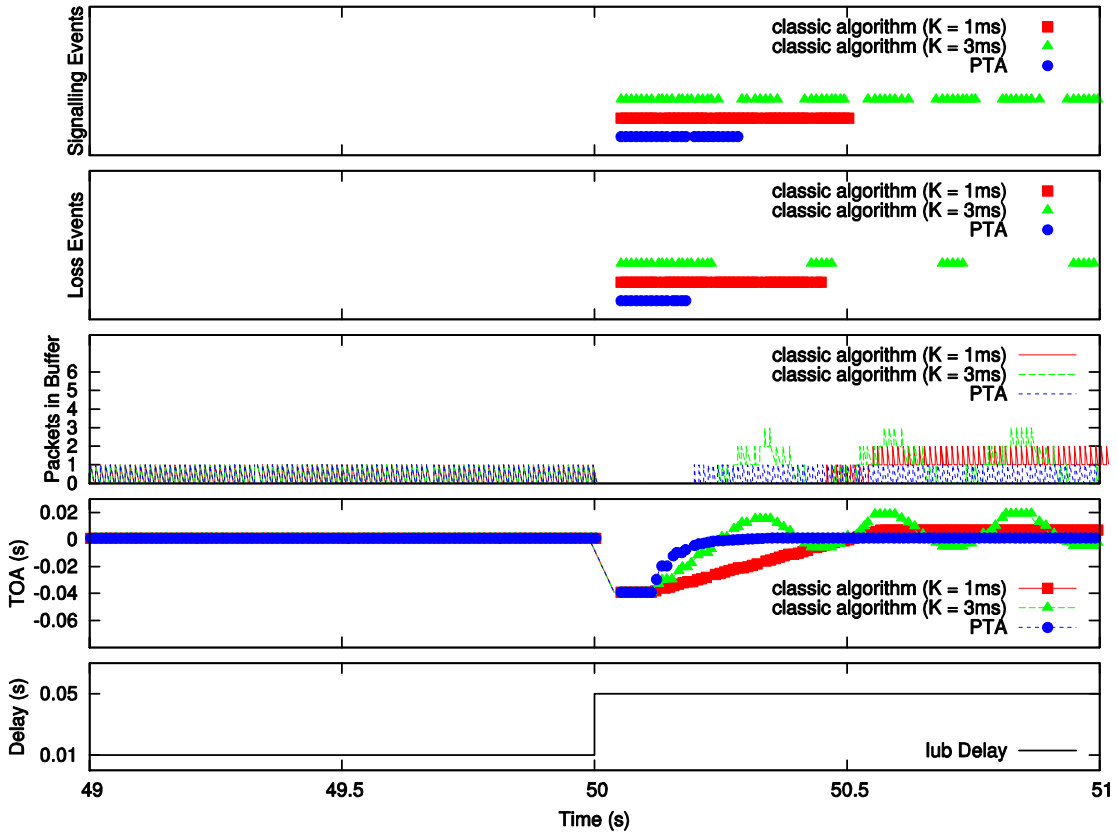

Figure 3.7: Reacción de ambos algoritmos ante un incremento del retardo.

Estas gráficas muestran que PTA se comporta claramente mejor y se estabiliza mucho más rápido que el algoritmo clásico para cualquiera de sus configuraciones. Con $K = 3$ ms (apropiado para un retardo de 10 ms en el Iub) el mecanismo clásico pierde la estabilidad tras el aumento del retardo y, con $K = 1$ ms (apropiado para un retardo del Iub de 50 ms), la respuesta del algoritmo clásico es considerablemente más lenta si se compara con la de PTA.

Los resultados del experimento de disminuir el retardo se ilustran en la figura 3.8. El interfaz Iub se configura inicialmente con un retardo de propagación de 50 ms y, en el instante de simulación $t = 50$ ms, se reduce a 10 ms. Después de este cambio, las tramas comienzan a llegar antes de la ventana de recepción y, por tanto, generan señalización aunque no se descartan. Sin embargo, una disminución del retardo tiene un efecto negativo: el llenado de los buffers del Nodo B. Por esta razón, el algoritmo de ajuste de tiempo debe reaccionar lo más rápido posible para evitar un almacenamiento excesivo, pero, al mismo tiempo, asegurando una respuesta estable del sistema. Como se aprecia en la figura 3.8, el

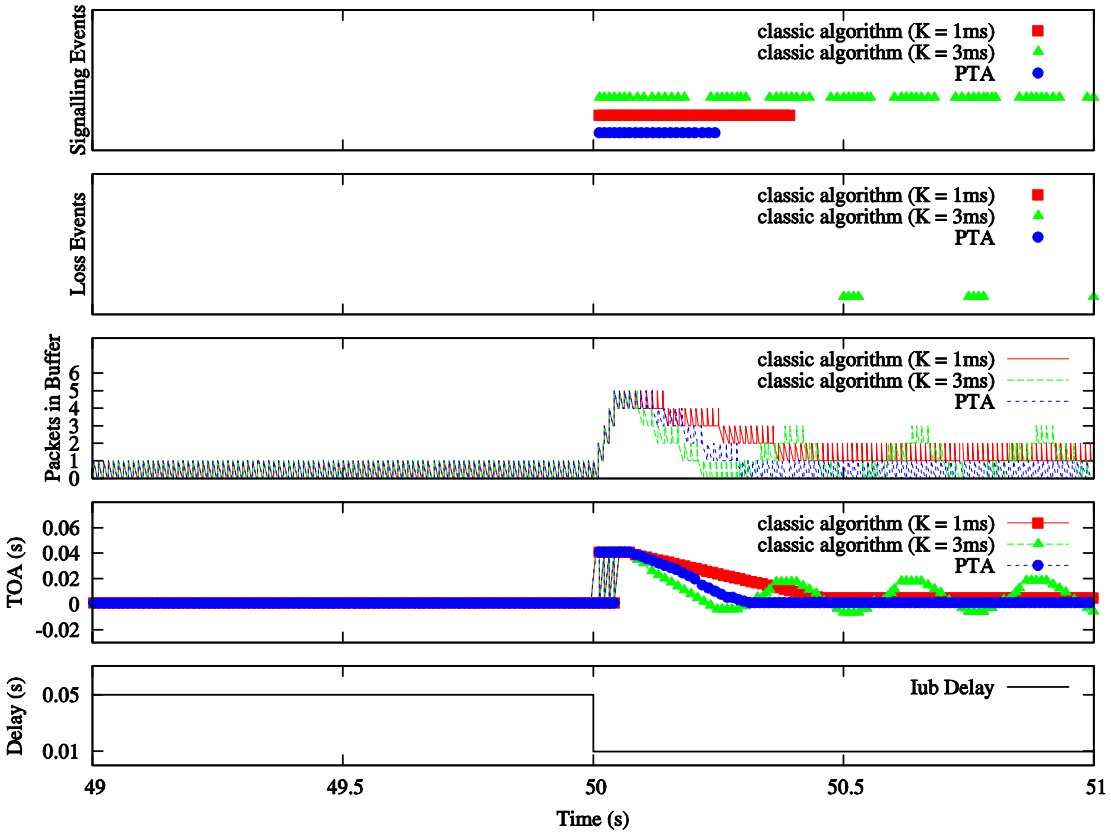

Figure 3.8: Reacción de ambos algoritmos ante un decremento del retardo.

algoritmo clásico con $K = 3$ ms es el más rápido a la hora de vaciar el buffer, pero se vuelve inestable. Por otro lado, el algoritmo clásico con $K = 1$ ms se mantiene estable pero su respuesta es demasiado lenta. PTA presenta el comportamiento más equilibrado puesto que es capaz de seguir la variación del retardo de forma rápida gracias a la estrategia proporcional mientras que la variación dinámica de K (basado en los valores de la tabla 3.1 para un *overshoot* del 10%) asegura la estabilidad. En comparación con el algoritmo clásico, PTA muestra un mejor comportamiento para cualquier situación de cambio en el retardo del Iub: mantiene bajos los niveles de los buffers del Nodo B, reduce las pérdidas y genera menos tramas de la señalización.

Comparación de prestaciones en condiciones de tráfico realista

En esta sección, ambos algoritmos son evaluados en condiciones de tráfico que reflejen un escenario práctico (30 fuentes de voz y 6 fuentes de datos). Con el

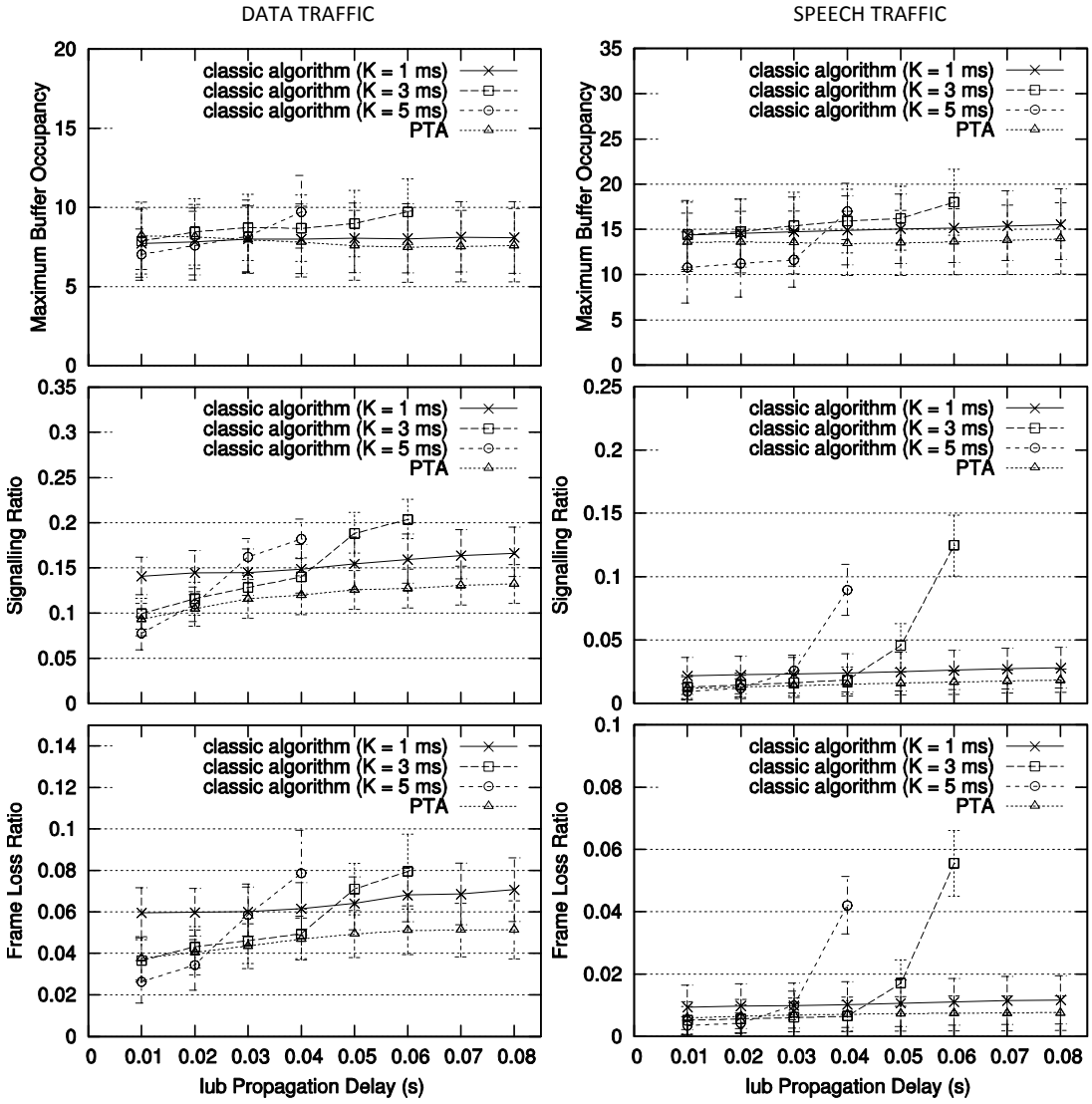

Figure 3.9: Evaluación de prestaciones para distintos valores de retardo constante en el Iub.

fin de comparar su rendimiento se miden el ratio de tramas perdidas, el ratio de tramas de señalización y el nivel máximo del buffer en el Nodo B. El ratio de señalización es el cociente entre el número de tramas TA enviadas por el Nodo B y el número total de tramas de datos recibidas en el nodo B. El ratio de perdidas es la relación entre el número de tramas que son descartadas (llegan "demasiado tarde") y el número total de tramas recibidas en el nodo B. Ambos ratios están expresadas en porcentaje, mientras que los niveles del buffer se expresan en número de tramas.

En la primera prueba se comparan ambos algoritmos dados diferentes retardos

de propagación en el Iub. Cada valor de retardo permanece constante durante el tiempo que dura cada simulación. Los parámetros *TOAWS* y *TOAWE* se han establecido a 20 ms y ms 5, respectivamente. Los resultados de la simulación se muestran en la figura 3.9. Cada uno de estos valores se obtienen promediando los resultados de diez ejecuciones de la simulación y se representan con un intervalo de confianza del 95%.

En la figura 3.9, se puede observar que el algoritmo clásico con altos valores de K se comporta de manera similar a PTA para bajos retardos del Iub (10 o 20 ms), pero se vuelve inestable con retardos mayores. Nótese que aunque el retardo de propagación permanezca constante, los distintos lujos de datos experimentan variaciones de retardo por el hecho de compartir recursos de transmisión. Por otro lado, el algoritmo clásico con $K = 1$ ms se mantiene estable para un amplio rango de retardos en el Iub pero su rendimiento es menor que el de PTA en todos los escenarios, debido a su lenta respuesta. La única forma de acelerar esta respuesta en el algoritmo clásico es aumentar K, lo cual provoca inestabilidad en situaciones de altos retardos. Este hecho pone de manifiesto el principal inconveniente del algoritmo clásico: el compromiso entre una rápida respuesta y la robustez frente a las oscilaciones.

En el siguiente escenario el retardo del Iub experimenta un cambio repentino en una situación de tráfico idéntica al escenario anterior. La figura 3.10 ilustra las prestaciones de ambos algoritmos para distintos aumentos (signo positivo) y decrementos (signo negativo) del retardo del Iub, que van desde 10 ms a 40 ms. El *TOAWS* se establece a 10 ms y *TOAWE* a 5 ms. Cada valor de rendimiento se mide durante los 60 segundos posteriores al cambio de retardo. Los valores medios y los intervalos de confianza (95%) se obtienen a partir de 10 ejecuciones de cada simulación.

Las prestaciones recogidas en la figura 3.10 confirman los resultados observados en la sección 3.3, donde se consideraba una sola fuente de voz y, por tanto, un único flujo FP. Se demuestra que bajo condiciones realistas tráfico, PTA también obtiene mejores resultados de rendimiento que el algoritmo clásico bajo incrementos y decrementos del retardo en el Iub. Se observa que la ocupación de los buffers del Nodo B en algunas ocasiones es mayor para PTA que para el algoritmo clásico. Esto es debido a la menor tasa de perdida de tramas de PTA, por lo que el nodo B tiene que almacenar tramas que con el algoritmo clásico se habrían perdido.

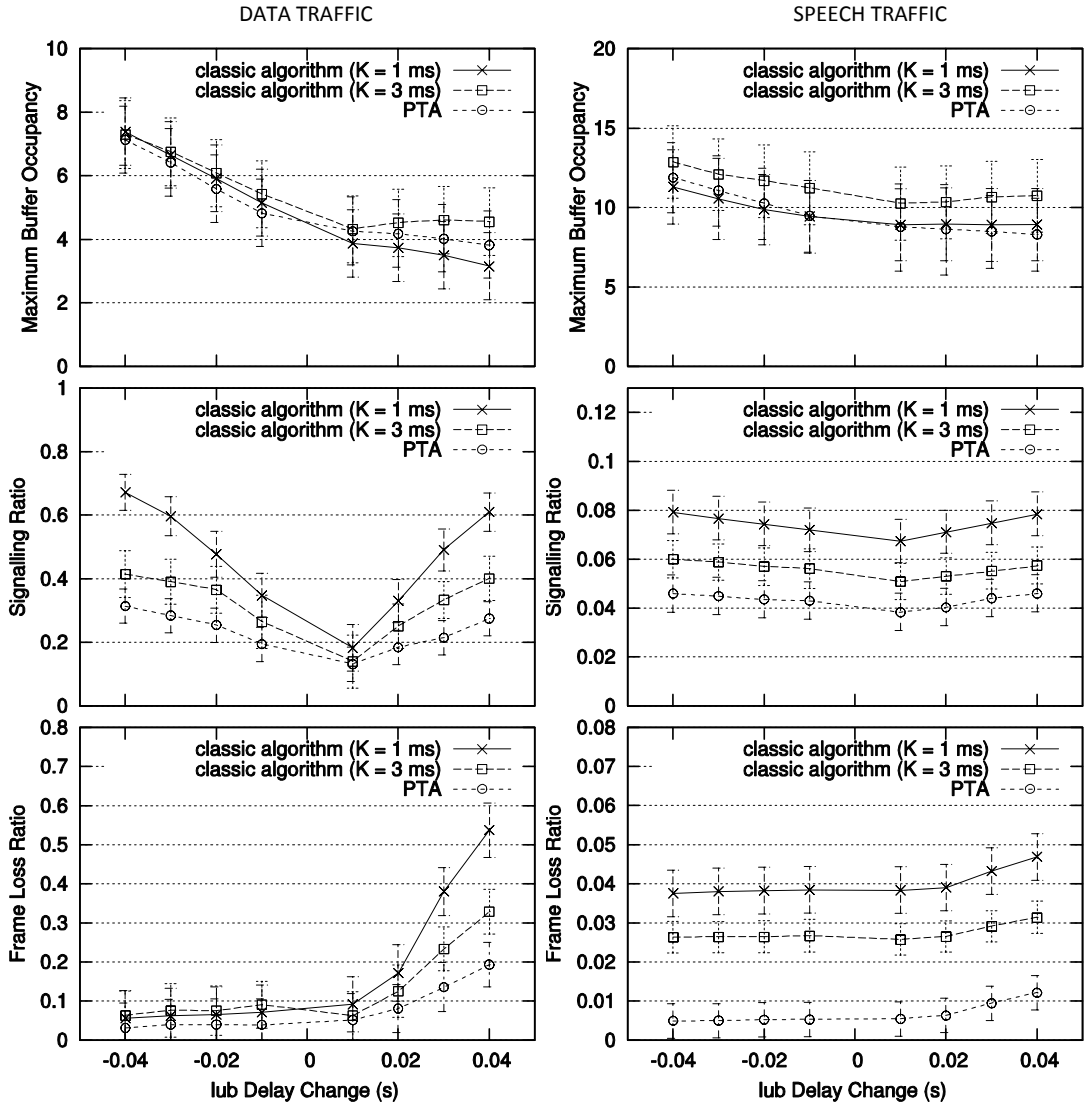

Figure 3.10: Evaluación de prestaciones ante un cambio brusco del retardo en el Iub.

3.4 Conclusiones

En este capítulo se ha propuesto un nuevo algoritmo de *timing adjustment* para el mecanismo de sincronización de canal de transporte en la red de acceso radio de UMTS. La estabilidad de este algoritmo, conocido como PTA, ha sido analizada mediante teoría de control en tiempo discreto y sus prestaciones se han evaluado a través de simulaciones en condiciones realistas de tráfico donde ha demostrado un mejor rendimiento que el algoritmo clásico. A nivel de resultados, las conclusiones más importantes son que PTA tiene una respuesta más rápida a los cambios

abruptos en el retardo del Iub que el algoritmo clásico y que su funcionamiento es estable independientemente del retardo del Iub.

Un algoritmo como el aquí presentado supone una mejora de la QoS experimentada por el usuario ya que consigue una sincronización más rápida y eficiente en la comunicación entre RNC y Nodos B, mejora que será aún más notable con la sustitución gradual de la red de transporte ATM a una red todo IP.

CHAPTER 4

Control de Flujo en el Backhaul

Resumen: Con la incorporación de HSDPA en las redes UMTS la función de *scheduling* se ha desplazado desde la RNC hasta el Nodo B, generando la necesidad de unos nuevos buffers en el Nodo B. A su vez, esta nueva distribución de la capacidad de almacenamiento entre la RNC y el Nodo B requiere de un mecanismo de control de flujo que regule la transferencia de datos entre ambos. En este capítulo se realiza un detallado estudio analítico de este control de flujo abordándolo como un problema de optimización cuadrática. Partiendo de este análisis se propone un nuevo algoritmo de control de flujo que consigue minimizar el retardo extremo a extremo gracias a que este algoritmo tiene en cuenta un parámetro como la ocupación de los buffers de la RNC que hasta ahora no se había considerado en algoritmos anteriores.

4.1 Introducción

HSDPA

La evolución del mercado de la telefonía móvil ha supuesto una fuerte demanda de sistemas de mayor capacidad así como de tasas de transferencia más altas. En este contexto, el 3GPP extendió la especificación de WCDMA en la *Release 5* con el concepto HSDPA [23]. HSDPA, diseñado específicamente para tráfico a ráfagas (servicios de tipo *interactive*, *streaming* y *background*), tiene como principales objetivos aumentar la velocidad de transmisión de datos en sentido *downlink*, mejorar la QoS percibida por el usuario y lograr un menor coste por bit entregado.

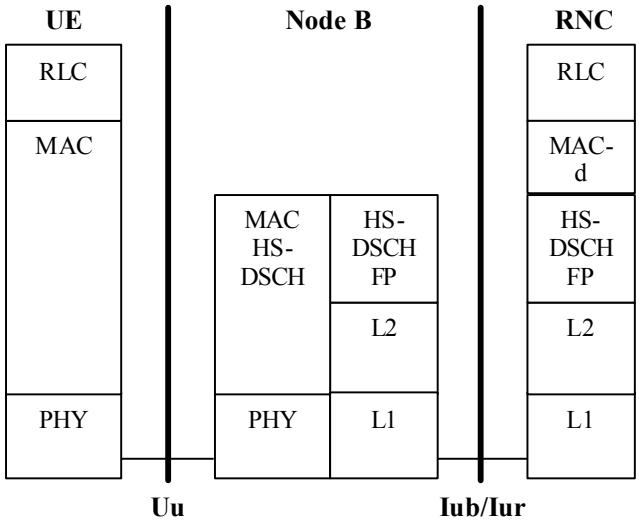

Figure 4.1: Estructura de protocolos para la interfaz radio de HSDPA.

El concepto HSDPA se basa en la implementación de un nuevo canal de transporte compartido en sentido *downlink*, denominado HS-DSCH [23], y en la transferencia de algunas funcionalidades de la capa MAC desde la RNC hasta el Nodo B. HSDPA presenta múltiples novedades con respecto a WCDMA, entre las que destacan la codificación y modulación adaptativas (*fast link adaptation*), HARQ, planificación rápida de paquetes (*fast packet scheduling*) y una reducción del intervalo de tiempo de transmisión (TTI) a 2 ms [24]. Estas características son soportadas en una nueva sub-capa MAC que se introduce en el Nodo B y que se conoce como MAC-hs (*Medium Access Control-high speed*). El resto de capas/protocolos de la red de acceso radio que se encuentran por encima de la capa MAC son los mismos que en la arquitectura que se presentó en *Release 99* para UMTS (ver figura 3.2), sin modificación alguna. No obstante, el protocolo RLC (*Radio Link Control*) sólo puede operar en modo con o sin reconocimiento, pero no en modo transparente debido al cifrado. La figura 4.1 ilustra la estructura de protocolos de la interfaz radio para una red UMTS con funcionalidad HSDPA.

En comparación con la arquitectura de *Release 99*, el desplazamiento del planificador de paquetes (*scheduler*) desde la RNC hasta el Nodo B es el cambio más notable. La motivación de este cambio es la necesidad de información reciente sobre el estado del canal radio por parte de los mecanismos de adaptación al canal (*link adaptation*) y del propio *scheduler*, a fin de poder realizar un seguimiento instantáneo de las condiciones radio de cada usuario. Ahora, al tomarse las decisiones de *scheduling* en la capa MAC-hs del Nodo B, se requieren nuevos buffers en

el Nodo B para almacenar los datos de cada usuario que esperan ser transmitidos a través de la interfaz radio. Estos buffers se denominan colas de prioridad y cada uno recibe información de un solo flujo de datos desde la RNC. Puede haber hasta 8 colas de prioridad para cada usuario (UE) [23].

Motivación

La distribución de la capacidad de almacenamiento entre la RNC y el Nodo B implica la necesidad de un mecanismo de control de flujo que regule la transferencia de datos (tramas RLC) desde la RNC al Nodo B. El objetivo de este mecanismo es mantener los buffers del Nodo B a un nivel tal que, por un lado, la capacidad del canal radio no sea malgastada y, por otro, el retardo de encolamiento en el Nodo B no sea demasiado alto. Es decir, el buffer debe estar lo suficientemente lleno como para que nunca falten datos a la hora de transmitir por la interfaz radio (*starvation*). Por otro lado, dichos buffers no deben estar demasiado llenos ya que, en caso de producirse un handover (cambio de celda/Nodo B por parte del terminal móvil), los datos que estaban almacenados en el Nodo B inicial son eliminados y la RNC debe volver a transmitirlos hacia el nuevo Nodo B que controle ahora al usuario.

El mecanismo de control de flujo recogido en las especificaciones del 3GPP para el canal HS-DSCH [25] es el mismo que se propuso para los canales dedicados (*Dedicated Channels*, DCH) en la *Release '99* y se conoce como un sistema basado en créditos. Sin embargo, las especificaciones básicamente se limitan a definir los formatos de las tramas de señalización *HS-DSCH Capacity Request* y *HS-DSCH Capacity Allocation*, y de datos *HS-DSCH Data Frames* que se utilizan durante el proceso [25]. Por último, también se especifica que dicho control de flujo se debe realizar de forma independiente para cada flujo de datos con el fin de proporcionar un trato equitativo a todos los flujos.

Partiendo de estas premisas son varios los controles de flujo para HSDPA aparecidos durante los últimos años en la literatura científica. Un esquema que se utiliza habitualmente consiste en observar y controlar directamente el nivel del buffer en el Nodo B para cada flujo de datos de forma que el tiempo de encolamiento no supere un valor previamente definido [26, 27]. Este esquema será descrito y analizado en detalle en la siguiente sección desde un punto de vista de optimización matemática. En [28] los autores proponen un algoritmo de asignación de recursos para la interfaz Iub basado en una estimación del throughput de HSDPA a través de la interfaz aire mediante una cadena de Markov. Otro esquema de control de

flujo para HSDPA es mostrado en [29]. Este esquema previene el desbordamiento de los buffers del Node B fijando un parámetro de control que determina el nivel máximo de ocupación en dichos buffers.

En este trabajo se realiza un estudio analítico del control de flujo de HSDPA. Dicho control se plantea como un problema de optimización cuadrática que se resuelve mediante técnicas de programación dinámica DP [30]. Además, se propone un nuevo algoritmo de control de flujo basado en estas mismas técnicas que minimiza el retardo extremo a extremo. Esto se consigue gracias a que el nuevo algoritmo tiene en cuenta también la ocupación de los buffers de la RNC que hasta ahora no se había considerado en algoritmos anteriores. Por último, dicho algoritmo es puesto a prueba en distintas situaciones mediante simulación computacional.

El resto del capítulo está organizado de la siguiente forma:

En la sección 4.2 se describirá el modelo matemático utilizado para caracterizar el sistema de control de flujo de HSDPA y se resolverá el problema de optimización derivado de dicho modelo matemático. Partiendo del estudio anterior, en el apartado 4.3 se presentará un nuevo algoritmo que consigue minimizar el retardo extremo a extremo. En la sección 4.4, se muestran los resultados obtenidos mediante simulación para ambos algoritmos. Por último, la sección 4.5 recoge las conclusiones de este trabajo.

4.2 Control de Flujo en HSDPA

Modelo del sistema

El escenario considerado para realizar nuestro estudio es una celda 3G con funcionalidad HSDPA (ver figura 4.2), donde varios usuarios (*User Equipments*, UEs) se conectan al Nodo B a través del canal *High Speed Downlink Shared Channel* (HS-DSCH) en sentido *downlink* y mediante un canal dedicado (DCH) en sentido *uplink*. La interfaz Iub, la cual conecta el Nodo B y la RNC, se supone de capacidad constante para el tráfico HSDPA. Asimismo, se considera que la RNC se conecta directamente a Internet a través de un enlace de capacidad constante.

Consideramos un flujo MAC-d por usuario, es decir, una sóla conexión de datos por usuario. Cada uno de estos flujos está asociado a una cola de prioridad del Nodo B, que almacena los datos que están a punto de ser transmitidos por la interfaz radio, y a un buffer en la RNC donde llegan nuevos paquetes de dicho flujo

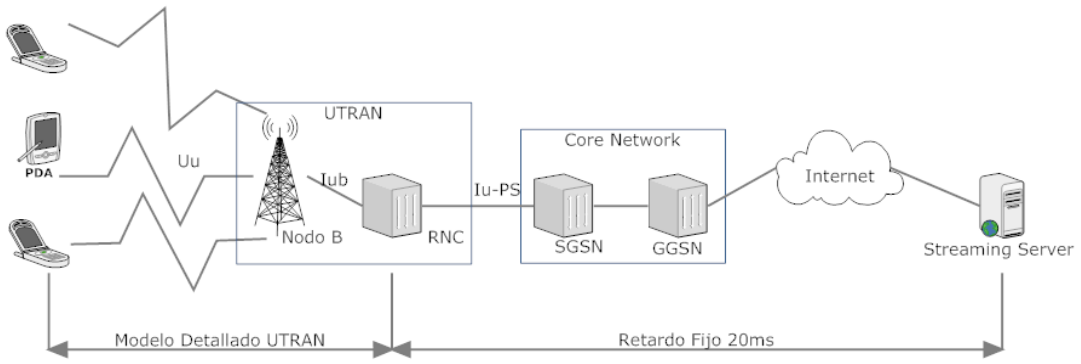

Figure 4.2: Arquitectura de la red HSDPA considerada.

desde Internet. Por tanto, la función del control de flujo consistirá en determinar la cantidad de información a transmitir para cada conexión de datos desde la RNC al Nodo B en un periodo de tiempo de longitud fija (cada conexión de datos tiene su propio proceso de control de flujo independiente del resto, tal y como describe el 3GPP en [25]). Dicho periodo es lo que comúnmente llamamos time-slot. Con el fin de obtener matemáticamente una política que determine las decisiones del control de flujo de HSDPA, cada conexión de datos entre la RNC y el Nodo B se modela como un sistema lineal en tiempo discreto. Cada time-slot se corresponde con la unidad de tiempo mínima en nuestro sistema en tiempo discreto, T_u.

Denotaremos por $q_i[n]$ la longitud en tramas RLC de la cola del Nodo B para el usuario i al inicio del time-slot n, y por $r_i[n]$ la tasa de transferencia de tramas RLC que son transmitidas por primera vez a través del canal radio para el usuario i durante el time-slot n (tasa efectiva). La variable $v_i[n]$ denota el número de *créditos*/tramas RLC concedidas por la RNC al usuario i para el time-slot n. La evolución de la cola en el Nodo B para cada usuario i viene dada por la ecuación lineal:

$$q_i[n+1] = q_i[n] + v_i[n] - r_i[n] \cdot T_u \tag{4.1}$$

Dada la variabilidad de la calidad del canal radio, la tasa de transmisión efectiva $r_i[n]$ puede cambiar de manera impredecible siendo desconocido su valor al inicio del time-slot n. En nuestro modelo matemático la representaremos como:

$$r_i[n] = \overline{r}_i[n] + \xi_i[n] \tag{4.2}$$

dónde $\overline{r}_i[n]$ es el valor esperado de $r_i[n]$ y $\xi_i[n]$ es una variable aleatoria de media cero y varianza finita que consideraremos como un error en la estimación de $r_i[n]$.

Análisis Matemático

Genéricamente, la mayor parte de algoritmos diseñados para llevar a cabo el control de flujo en HSDPA regulan la transferencia de datos desde la RNC hasta el Nodo B con el objetivo de conseguir un tiempo de encolamiento predefinido T_w para los buffer del Nodo B $q_i[n]$ [26, 27]. En consecuencia, el control de flujo trata de mantener el nivel del buffer para cada flujo a un valor objetivo $Q_i[n]$:

$$Q_i[n] = r_i[n] \cdot T_w \tag{4.3}$$

Para cada usuario i, cada time-slot el control de flujo trata de compensar la diferencia entre el nivel deseado $Q_i[n]$ y el nivel real $q_i[n]$. En este trabajo se formula el problema de control de flujo en HSDPA como un problema de optimización cuadrática. Así, la función objetivo a minimizar para cada usuario es:

$$E\{\sum_{n=0}^{K}[(q_i[n] - Q_i[n])^2 + (v_i[n] - r_i[n] \cdot T_u)^2]\} \tag{4.4}$$

dónde el primer término representa la penalización por desviarse del valor objetivo de la cola del Nodo B y el segundo término es una medida de la calidad con la que la tasa de transferencia de la RNC (*créditos*) sigue la tasa efectiva del canal. K es el número de etapas o time-slots sobre los que se minimiza.

Introduciendo los siguientes cambios de variable:

$$\begin{array}{rcl} x_i[n] & := & q_i[n] - Q_i[n] \\ u_i[n] & := & v_i[n] - \overline{r}_i[n] \cdot T_u \end{array} \tag{4.5}$$

y usando notación vectorial

$$\begin{array}{rcl} \mathbf{x}_n & = & (x_1[n], \ldots, x_M[n])' \\ \mathbf{u}_n & = & (u_1[n], \ldots, u_M[n])' \\ \mathbf{w}_n & = & (\xi_1[n] \cdot T_u, ..., \xi_M[n] \cdot T_u)' \end{array} \tag{4.6}$$

la ecuación del sistema (4.1) con M usuarios en la celda se puede expresar como:

$$\mathbf{x}_{n+1} = \mathbf{A}_n\mathbf{x}_n + \mathbf{B}_n\mathbf{u}_n + \mathbf{w}_n = \mathbf{x}_n + \mathbf{u}_n + \mathbf{w}_n \tag{4.7}$$

donde $\mathbf{x}_n$ es el vector de estados, $\mathbf{u}_n$ es el vector de control y $\mathbf{w}_n$ es el vector de perturbaciones. En nuestro caso, tanto $\mathbf{A}_n$ como $\mathbf{B}_n$ son matrices identidad de orden $M \times M$.

Por su parte la función de coste o función objetivo (4.4) se puede expresar ahora de la siguiente forma:

$$E\left\{\mathbf{x}'_K\mathbf{Q}_K\mathbf{x}_K + \sum_{n=0}^{K-1}\left(\mathbf{x}'_n\mathbf{Q}_n\mathbf{x}_n + \mathbf{u}'_n\mathbf{R}_n\mathbf{u}_n\right)\right\} = E\left\{\mathbf{x}_K{}^2 + \sum_{n=0}^{K-1}\left(\mathbf{x}_n{}^2 + \mathbf{u}_n{}^2\right)\right\} \tag{4.8}$$

donde las matrices $\mathbf{Q}_n$ y $\mathbf{R}_n$ son matrices identidad de orden $M \times M$. Las primeras representan el coste asociado a la ocupación de los buffers y las segundas el coste asociado a los vectores de control. K es el número de etapas sobre las que se desea optimizar el sistema. En nuestro caso, estamos interesados en reducir los niveles de los buffers al final de cada time-slot y, por tanto, minimizaremos sobre una sola etapa aplicando, por tanto, una política *one-step lookahead* [30].

La minimización de la función de coste se llevará a cabo mediante programación dinámica. Básicamente lo que haremos es ir calculando etapa a etapa cuál es la combinación de controles que consigue minimizar el estado siguiente (nivel buffers), empezando en orden inverso desde la última etapa (n+1 en nuestro caso) hasta llegar a la etapa inicial. Así, en la etapa n, nuestro objetivo será minimizar el coste del control $\mathbf{u}_n$ y del estado $\mathbf{x}_{n+1}$. El coste de la última etapa sería:

$$J_{n+1}(\mathbf{x}_{n+1}) = \mathbf{x}_{n+1}{}^2 \tag{4.9}$$

El objetivo es encontrar el vector de control $\mathbf{u}_n$ con el que se obtenga el mínimo coste desde la etapa n hasta la $n+1$. En la etapa n, el coste vendría dado por:

$$J_n(\mathbf{x}_n) = \min_{\mathbf{u}_n} E\left\{\mathbf{x}_n{}^2 + \mathbf{u}_n{}^2 + J_{n+1}(\mathbf{x}_{n+1})\right\} \tag{4.10}$$

aplicando (4.9) y sustituyendo $\mathbf{x}_{n+1}$ por su valor en (4.7) obtenemos:

$$\begin{aligned} J_n(\mathbf{x}_n) &= 2\mathbf{x}_n{}^2 + E\left\{\mathbf{w}_n{}^2 + 2\mathbf{x}'_n\mathbf{w}_n\right\} \\ &+\min_{\mathbf{u}_n}\left\{2\mathbf{x}'_n\mathbf{u}_n + 2\mathbf{u}_n{}^2 + E\left\{2\mathbf{w}'_n\mathbf{u}_n\right\}\right\} \end{aligned} \tag{4.11}$$

Teniendo en cuenta que la esperanza de $\mathbf{w}_n$ es cero (sus componentes $\xi_i[n]$ son variables aleatorias de media cero) el último término de la expresión anterior desaparece. El vector de control que minimiza $J_n(\mathbf{x}_n)$ puede obtenerse directamente derivando (4.11) con respecto a $\mathbf{u}_n$ e igualando el resultado a cero, obteniendo la siguiente ley de control:

$$\mathbf{u}_n = -\frac{1}{2}\mathbf{x}_n \tag{4.12}$$

Deshaciendo los cambios de variable realizados en (4.5), obtenemos el número de créditos óptimo que debe enviar la RNC para cada flujo i:

$$v_i[n] = -\tfrac{1}{2}\left(q_i[n] - Q_i[n]\right) + r_i[n]\cdot T_u \tag{4.13}$$

Curiosamente, en [26] se propone esta misma expresión desde un planteamiento intuitivo.

4.3 Algoritmo Propuesto

Con el objetivo de mejorar las prestaciones del algoritmo anterior, proponemos un nuevo esquema de control de flujo que no sólo tiene en cuenta el estado de los buffers del Nodo B a la hora de tomar sus decisiones sino también el de los buffers de la RNC. Con este nuevo algoritmo se consigue minimizar el retardo extremo a extremo de cada conexión de datos, pero con la contrapartida de aumentar ligeramente la ocupación de los buffers del Nodo B.

Denotaremos por $y_i[n]$ la ocupación del buffer de la RNC para el usuario i al inicio del time-slot n, y $\lambda_i[n]$ la tasa de llegada de tramas RLC para el usuario i a la RNC durante el time-slot n. La evolución de las colas en el Nodo B y la RNC para cada usuario i viene dado por el siguiente sistema de ecuaciones lineales:

$$\begin{aligned} q_i[n+1] &= q_i[n] + v_i[n] - r_i[n] \cdot T_u \\ y_i[n+1] &= y_i[n] - v_i[n] + \lambda_i[n] \cdot T_u \end{aligned} \tag{4.14}$$

Tanto $r_i[n]$ como $\lambda_i[n]$ son variables aleatorias cuyo valor es desconocido al inicio del time-slot n y, por tanto, tenemos que aproximarlas por sus valores esperados, $\bar{r}_i[n]$ y $\bar{\lambda}_i[n]$ respectivamente. El sistema se reescribe ahora como:

$$\begin{aligned} q_i[n+1] &= q_i[n] + v_i[n] - (\bar{r}_i[n] + \xi_i[n]) \cdot T_u \\ y_i[n+1] &= y_i[n] - v_i[n] + (\bar{\lambda}_i[n] + \delta_i[n]) \cdot T_u \end{aligned} \tag{4.15}$$

dónde ξ_i y δ_i son variables aleatorias de media cero que representan los errores asociados a las predicciones de r_i y λ_i respectivamente.

Al igual que en el anterior algoritmo, introducimos el cambio de variable $x_i[n] := q_i[n] - Q_i[n]$ siendo $Q_i[n] = r_i[n]T_w$ el valor objetivo para la longitud del buffer (T_w tiempo de encolamiento predefinido). Con este cambio el nuevo sistema queda de la siguiente forma:

$$\begin{aligned} x_i[n+1] &= x_i[n] + v_i[n] - (\bar{r}_i[n] + \xi_i[n]) \cdot T_u \\ y_i[n+1] &= y_i[n] - v_i[n] + (\bar{\lambda}_i[n] + \delta_i[n]) \cdot T_u \end{aligned} \tag{4.16}$$

Por simplicidad expresaremos el sistema en notación matricial considerando M usuarios activos en la celda HSDPA. Así, definimos los siguientes vectores:

$$\begin{aligned} \mathbf{z}_n &= (x_1[n], ..., x_M[n], y_1[n], ..., y_M[n])' \\ \mathbf{v}_n &= (v_1[n], ..., v_M[n], v_1[n], ..., v_M[n])' \\ \mathbf{w}_n &= (-\bar{r}_1[n] - \xi_1[n], ..., -\bar{r}_M[n] - \xi_M[n], \\ &\quad \bar{\lambda}_1[n] + \delta_1[n], ..., \bar{\lambda}_M[n] + \delta_M[n])' \end{aligned} \tag{4.17}$$

donde $\mathbf{z}_n$ es el vector de estados, $\mathbf{v}_n$ es el vector de control y $\mathbf{w}_n$ es el vector de perturbaciones. El sistema lineal (4.16) puede expresarse ahora en notación matricial como:

$$\mathbf{z}_{n+1} = \mathbf{A}_n\mathbf{z}_n + \mathbf{B}_n\mathbf{v}_n + T_u \cdot \mathbf{w}_n \tag{4.18}$$

dónde $\mathbf{A}_n$ es una matriz identidad de dimensión $2M \times 2M$ y $\mathbf{B}_n$ tiene la siguiente forma:

$$B = \begin{pmatrix} I_N \\ -I_N \end{pmatrix} \tag{4.19}$$

con I_N matriz identidad de dimensión $M \times M$.

El objetivo de nuestro algoritmo es minimizar la ocupación total del sistema y, por tanto, minimizar el retardo extremo a extremo. Para ello formularemos este problema de minimización como un problema de programación dinámica [30], asociando una función de coste al sistema anteriormente descrito, la cual deberemos minimizar. Asumiendo un coste cuadrático, dicha función de coste presenta el siguiente aspecto:

$$E\left\{\mathbf{z}'_K\mathbf{Q}_K\mathbf{z}_K + \sum_{n=0}^{K-1}\left(\mathbf{z}'_n\mathbf{Q}_n\mathbf{z}_n + \mathbf{v}'_n\mathbf{R}_n\mathbf{v}_n\right)\right\} \tag{4.20}$$

donde las matrices $\mathbf{Q}_n$ son simétricas semidefinidas positivas y las matrices $\mathbf{R}_n$ son simétricas definidas positivas. K es el número de etapas sobre las que se desea optimizar el sistema. Al igual que en el algoritmo anterior, nos interesa reducir los buffers al final de cada slot y, por tanto, minimizaremos sobre una sola etapa.

Como vemos, minimizar la función de coste supone minimizar dos factores cuadráticos: por un lado, la ocupación de los buffers del Nodo B y RNC ($\mathbf{z}_n$) y, por otro, los vectores de control o créditos ($\mathbf{v}_n$). El primer factor se corresponde con nuestro objetivo inicial mientras que el segundo impone una restricción sobre los créditos que evita que obtengamos como solución al problema la solución trivial, es decir, asignar infinitos recursos.

Las matrices $\mathbf{Q}_n$ representan el coste asociado a la longitud de los buffers y pueden ser expresadas como:

$$\mathbf{Q}_n = \begin{pmatrix} \mathbf{Q}_{00n} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{11n} \end{pmatrix} \tag{4.21}$$

con todas las submatrices de dimensión $M \times M$ y valor:

$$\begin{aligned} \mathbf{Q}_{00n} &= \alpha_x[n]\mathbf{I} \\ \mathbf{Q}_{11n} &= \alpha_y[n]\mathbf{I} \end{aligned} \tag{4.22}$$

donde $\mathbf{I}$ es las matriz identidad. $\alpha_x[n]$ y $\alpha_y[n]$ son los factores de ponderación aplicados a los buffers de Nodo B y RNC respectivamente, es decir, el coste o penalización asociada a la ocupación de dichos buffers. El ajuste de estos pesos se comentará más adelante.

Por su parte, $\mathbf{R}_n$ es una matriz identidad de orden $2M \times 2M$ y puede interpretarse como el coste asociado a los vectores de control $\mathbf{v}_n$ (créditos).

En general, las matrices $\mathbf{Q}_n$ y $\mathbf{R}_n$ pueden depender del índice pero como vamos a minimizar etapa a etapa nos referiremos a ellas como $\mathbf{Q}$ y $\mathbf{R}$. En la etapa n, nuestro objetivo será minimizar el coste del control $\mathbf{v}_n$ y del estado $\mathbf{z}_{n+1}$. El coste de esta última etapa sería:

$$J_{n+1}(\mathbf{z}_{n+1}) = \mathbf{z}'_{n+1}\mathbf{Q}\mathbf{z}_{n+1} \tag{4.23}$$

El objetivo es encontrar el vector de control $\mathbf{v}_n$ con el que se obtenga el mínimo coste desde la etapa n hasta la última etapa ($n+1$ en nuestro caso). En la etapa n, el coste vendría dado por:

$$J_n(\mathbf{z}_n) = \min_{\mathbf{v}_n} E\left\{\mathbf{z}'_n\mathbf{Q}\mathbf{z}_n + \mathbf{v}'_n\mathbf{R}\mathbf{v}_n + J_{n+1}(\mathbf{z}_{n+1})\right\} \tag{4.24}$$

aplicando (4.23) y sustituyendo $\mathbf{z}_{n+1}$ por su valor en (4.18) obtenemos:

$$\begin{aligned} J_n(\mathbf{z}_n) &= \mathbf{z}'_n(\mathbf{A}'\mathbf{Q}\mathbf{A} + \mathbf{Q})\mathbf{z}_n + E\left\{\mathbf{w}'_n\mathbf{Q}\mathbf{w}_n + 2\mathbf{z}'_n\mathbf{A}'\mathbf{Q}\mathbf{w}_n\right\} \\ &+\min_{\mathbf{v}_n}\left\{2\mathbf{z}'_n\mathbf{A}'\mathbf{Q}\mathbf{B}\mathbf{v}_n + \mathbf{v}'_n(\mathbf{B}'\mathbf{Q}\mathbf{B} + \mathbf{R})\mathbf{v}_n + E\left\{2\mathbf{w}'_n\mathbf{Q}\mathbf{B}\mathbf{v}_n\right\}\right\} \end{aligned} \tag{4.25}$$

Teniendo en cuenta la linealidad del operador esperanza, el último término de la expresión anterior se puede escribir como $2\overline{\mathbf{w}}'_n\mathbf{Q}\mathbf{B}\mathbf{v}_n$, donde

$$\overline{\mathbf{w}}_n = (-\overline{r}_1[n], ..., \overline{r}_M[n], \overline{\lambda}_1[n], ..., \overline{\lambda}_M[n]). \tag{4.26}$$

El vector de control que minimiza $J_n(\mathbf{z}_n)$ puede obtenerse directamente derivando (4.25) con respecto a $\mathbf{v}_n$ e igualando el resultado a cero

$$(\mathbf{B}'\mathbf{Q}\mathbf{B} + \mathbf{R})\mathbf{v}_n + (\mathbf{z}'_n\mathbf{A}'\mathbf{Q}\mathbf{B} + \overline{\mathbf{w}}'_n\mathbf{Q}\mathbf{B})' = 0 \tag{4.27}$$

Despejando el valor de $\mathbf{v}_n$, obtenemos que la ley de control resultante es:

$$\mathbf{v}_n = -\left(\mathbf{R} + \mathbf{B}'\mathbf{Q}\mathbf{B}\right)^{-1}\mathbf{B}'\mathbf{Q}\mathbf{A}\left(\mathbf{z}_n + \mathbf{A}^{-1}\bar{\mathbf{w}}_n\right) \tag{4.28}$$

Sustituyendo en (4.28) las matrices por sus valores correspondientes obtenemos que el control aplicado para cada flujo de datos i debería ser:

$$v_i[n] = \tfrac{1}{\alpha_x[n]+\alpha_y[n]}\left(\alpha_y[n]\left(y_i[n] + \lambda_i[n]\cdot T_u\right) - \alpha_x[n]\left(x_i[n] - r_i[n]\cdot T_u\right)\right) \tag{4.29}$$

y deshaciendo el cambio de variable:

$$v_i[n] = \frac{\alpha_y[n]}{\alpha_x[n]+\alpha_y[n]} \left(y_i[n] + \lambda_i[n] \cdot T_u\right) - \frac{\alpha_x[n]}{\alpha_x[n]+\alpha_y[n]} \left(q_i[n] - Q_i[n] - r_i[n] \cdot T_u\right) \tag{4.30}$$

Como vemos los créditos concedidos dependen de dos factores, el primero relacionado con el estado de las colas de la RNC $y_i[n]$, y el segundo con las colas del Nodo B $q_i[n]$. La configuración de los factores $\alpha_x[n]$ y $\alpha_y[n]$ se ha de ajustar de acuerdo al objetivo buscado. El ajuste puede llevarse a cabo offline, mediante técnicas de simulación combinadas con aprendizaje automático, o puede llevarse a cabo online mediante tecnicas de aprendizaje óptimo como multi-armed-bandits [31], o response surface methodology [32]. En nuestro caso, se han empleado simulaciones para el ajuste de los valores, y se ha obtenido el siguiente heurístico:

$$\begin{array}{rcl} 0 < \frac{q_i[n]}{Q_i[n]} < 1 & \Rightarrow & \alpha_x[n] = 1, \alpha_y[n] = 1 \\ 1 \leq \frac{q_i[n]}{Q_i[n]} < 2 & \Rightarrow & \alpha_x[n] = \frac{q_i[n]}{Q_i[n]} + 1, \alpha_y[n] = 1 \\ \frac{q_i[n]}{Q_i[n]} \geq 2 & \Rightarrow & \alpha_x[n] = 1, \alpha_y[n] = 0 \end{array} \tag{4.31}$$

Si el nivel del buffer del Nodo B aún no ha llegado al nivel objetivo ($0 < \frac{q_i[n]}{Q_i[n]} < 1$) se ponderan por igual ambos factores (0.5), tanto el que depende del Nodo B como el de la RNC. De esta forma, se incrementa la tasa de trasnmisión de la RNC para que el buffer del Nodo B alcance lo antes posible el valor objetivo. Una vez alcanzado/superado el nivel objetivo ($1 \leq \frac{q_i[n]}{Q_i[n]} < 2$), se toma $\alpha_x[n]$ mayor que $\alpha_y[n]$ de forma que se reduzca la importancia del factor relativo a los buffers de la RNC para frenar el envío de créditos. En caso de superar el umbral de $2Q_i[n]$ se reduce drásticamente la tasa, dejando de considerar la ocupación de la RNC.

4.4 Evaluación de Prestaciones

Modelo de Simulación

Para la evaluación de prestaciones de los algoritmos presentados se ha implementado un modelo detallado de celda HSDPA que incluye los protocolos RLC, MAC-d y MAC-hs. El simulador ha sido desarrollado en OMNeT++, herramienta de simulación en C++ orientada a eventos [19].

Para la implementación de la capa física (canal radio) se ha optado por un modelo basado en curvas BLER, el cual está completamente descrito en [33]. El modelo elegido es un *ITU-T Pedestrian A* con velocidad de 3 km/h. Los equipos de usuario (UE's) informan al Nodo B del estado del canal radio por medio del CQI con un retardo constante de 6 ms [33]. Los formatos de transporte (TF) son

elegidos en la capa MAC-hs a partir de estos informes de forma que el BLER sea inferior al 10%. El Nodo B se conecta a la RNC a través de un enlace punto a punto de ancho de banda constante e igual a 6 Mbps. Por otra parte, se supone que Internet y el *core network* introducen un retardo constante de 20 ms en cada dirección.

Todos los experimentos se realizan con tráfico streaming video. Dicho tráfico se modela mediante unas fuentes de tasa constante que transmiten a 312 Kbps con un tamaño de paquete de 576 bytes, similares a las utilizadas en [34]. En consecuencia, al tratarse de tráfico en tiempo real la capa RLC funciona en modo sin reconocimiento (UM). El número máximo de retransmisiones en la capa MAC-hs es de 3 y el tamaño de la ventana de MAC-hs se fija en 12. Se considera que los terminales móviles son de categoría 6, lo que implica que la máxima velocidad a la que pueden recibir datos desde el canal HSDPA (HS-DSCH) es 3.6 Mbps. No se implementa multiplexación en código. Los enlaces ascendentes DCHs operan a 64 Kbit/s. El tamaño de las tramas RLC se ajusta a su valor típico, 320bits.

El tiempo de espera deseado en los buffers del Nodo B para cada flujo de datos se establece en $T_w = 100$ ms. El tiempo de actualización T_u tiene que ser lo suficientemente pequeño como para permitir al flujo de control seguir con precisión las variaciones del canal de radio pero a su vez tiene que ser lo suficientemente grande como para mantener la carga de señalización entre RNC y Nodo B a un nivel razonable. Teniendo en cuenta este compromiso se ha fijado $T_u = 100$ ms.

Resultados

En primer lugar se ha realizado un experimento para comparar las prestaciones de los dos algoritmos expuestos en este trabajo frente a la situación sin control de flujo bajo tres schedulers radio diferentes: *Round Robin*, *Maximum C/I* y *Proportional Fair*. El primero de ellos elige el usuario que debe transmitir de forma equitativa siguiendo un orden secuencial. Con *Maximum C/I*, en cada TTI trasmite siempre el usuario que experimenta las mejores condiciones radio (mayor ratio portadora a interferencia), siendo el esquema más injusto. Por último, *Proportional Fair* es una combinación de los dos anteriores pues, pretende ser equitativo pero tiene en cuenta también el estado del canal radio a la hora de seleccionar al usuario. En [35] se describe detalladamente este *scheduler*. Los resultados mostrados en la Fig. 4.3 corresponden al promedio de diez réplicas de cien segundos de duración, usando un intervalo de confianza del 95%. Se considera que hay doce usuarios activos en la celda.

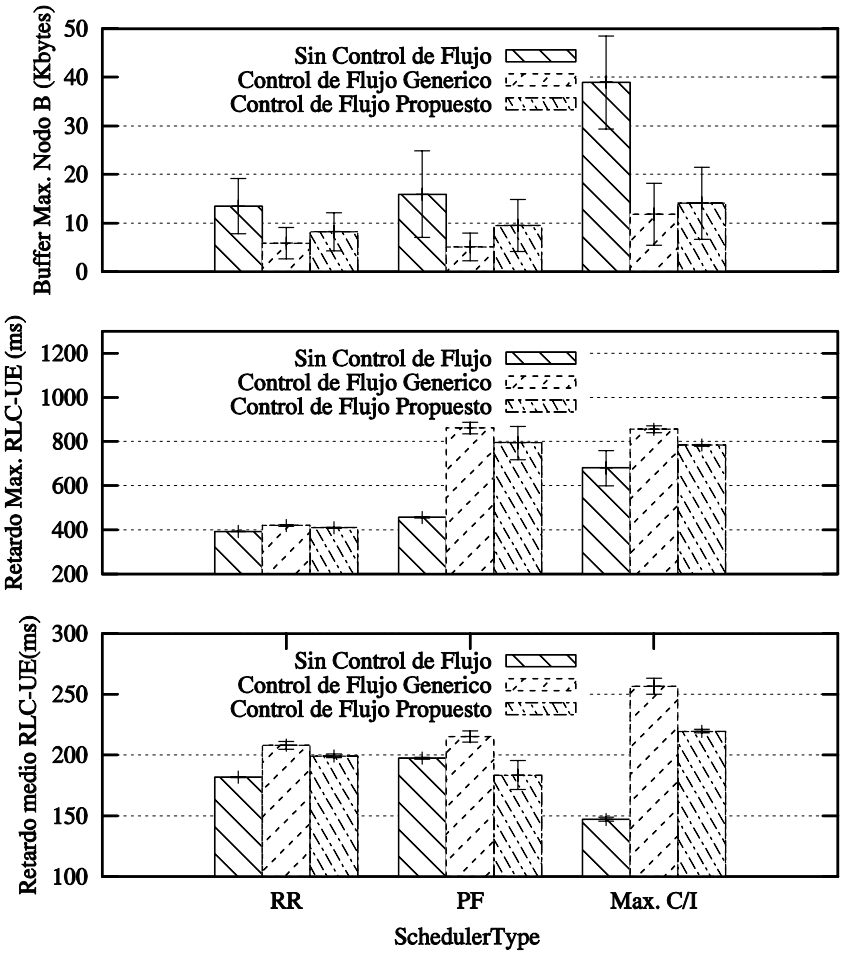

Figure 4.3: Prestaciones frente a distintos schedulers

Como es lógico, en ausencia de control de flujo se consiguen los mejores resultados en cuanto a retardo para cualquier *scheduler* puesto que los datos atraviesan la RNC sin espera alguna, siendo transmitidos directamente hacia el Nodo B. Sin embargo, esta situación no es deseable puesto que los buffers del Nodo B no están controlados. En ausencia de control los buffers pueden desbordarse o, en el caso de que ocurra un handover, todos los datos que estuvieran almacenados en el Nodo B original serían eliminados y la RNC debe volver a transmitirlos hacia el nuevo Nodo B que controlase la celda donde se hubiera desplazado el usuario. En lo que a los controles de flujo se refiere, se observa como nuestro algoritmo consigue mejorar significativamente el retardo extremo a extremo con respecto al algoritmo genénico.

Por último se comparan los algoritmos frente al número de usuarios activos en el

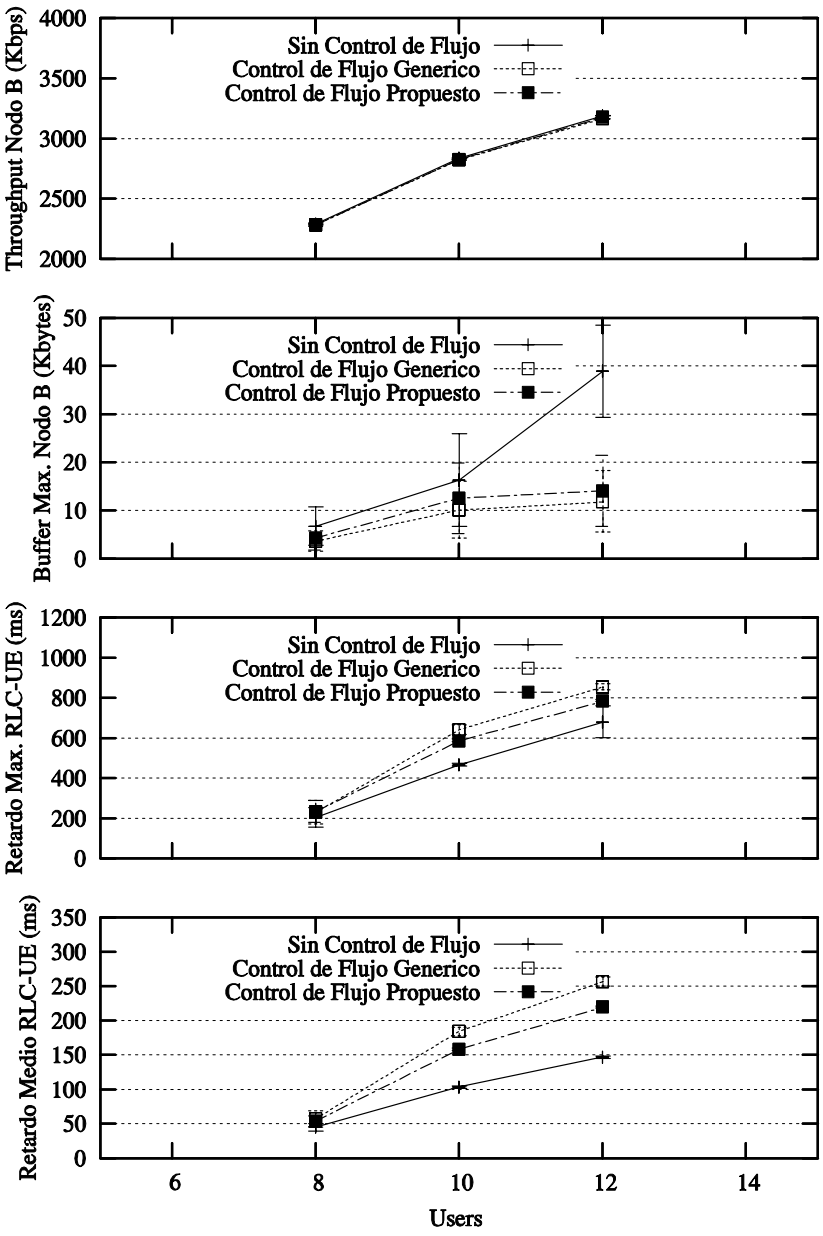

Figure 4.4: Prestaciones en función del número de usuarios

sistema con un *scheduler Maximum C/I*. En la Fig. 4.4 se muestran los resultados de este experimento. Al igual que en la anterior gráfica, los datos corresponden al promedio de diez réplicas de cien segundos de duración usando un intervalo de confianza del 95%.

Como vemos, los resultados de este experimento están en línea con los anteriores. El algoritmo propuesto consigue mejores figuras de retardo (a costa de

una mayor ocupación del buffer del Nodo B) que el algoritmo genérico. Por otro lado, cuánto mayor es el número de usuarios mayores diferencias se aprecian entre el rendimiento de los algoritmos. En el caso de los doce usuarios vemos como el throughput del Nodo B se acerca ya a los 3.6Mbit/s que, para usuarios de categoría 6, sería la tasa máxima a la que podría transmitir.

4.5 Conclusiones

En este capítulo se ha abordado el problema del control de flujo de HSDPA desde una perspectiva diferente a los trabajos realizados hasta la fecha. Se ha planteado el mecanismo de control de flujo como un problema de optimización cuadrática, cuya solución se puede obtener de forma directa y relativamente sencilla empleando programación dinámica. El resultado coincide con una política de control previamente recogida en la literatura, pero cuya obtención se basaba en criterios intuitivos, por lo que el resultado obtenido permite dotar de base teórica un mecanismo ya existente. Empleando la metodología propuesta, hemos podido desarrollar un nuevo algoritmo de control de flujo que minimiza el retardo extremo a extremo desde la RNC hasta el UE.

CHAPTER 5

Asignación de Recursos Radio Considerando el Backhaul

Resumen: Desde la implantación de LTE (4G), la tasas máximas de transmisión alcanzables para el usuario en el interfaz radio se han disparado con respecto a las anteriores generaciones de sistemas móviles. Por primera vez operadores, fabricantes y comunidad académica coinciden en la necesidad de optimizar los recursos del backhaul además de los recursos radio. En este capítulo estudiamos el impacto que tiene el backhaul en el algorimo de gestión de recursos radio, o scheduling. Para ello consideramos un escenario sencillo compuesto por una sola celda y una infraestructura backhaul consistente en un enlace punto a punto de capacidad C. En este escenario planteamos el problema de optimización correspondiente a un scheduler proportional fair, incluyendo la restricción impuesta por los límites de capacidad del backhaul. Aplicando las condiciones Karush-Kunh-Tucker se ha obtenido la solución, en algunos casos cerrada, del problema de optimización. Las evaluaciones numéricas muestran que cuando el backhaul es el cuello de botella el rendimiento del scheduler radio con consideraciones de backhaul es claramente superior al scheduler convencional.

5.1 Introducción

Dependiendo del ancho de banda espectral asignado a una celda radio LTE (4G), las tasas máximas de transmisión en el interfaz radio pueden llegar a superar los 300 Mbps. En comparación con los sistemas anteriores (3G y 3.5G), estas tasas

suponen un notable aumento de velocidad de transmisión. Este hecho, junto con una mayor demanda de datos por parte de los usuarios, ha incrementado sustancialmente la carga de tráfico que ha de soportar la infraestructura de interconexión entre las estaciones base y el núcleo de la red, infraestructura a la que denominamos backhaul.

El aumento de tráfico en el backhaul conlleva la necesidad de realizar mayores inversiones para adecuar esta infraestructura de red a las nuevas necesidades. No obstante, el *upgrade* del backhaul se enfrenta a a una dificultad fundamental: dotar a todas las celdas de una conectividad sin limitaciones de ancho de banda puede ser técnicamente inabordable en un tiempo razonable o economicamente inviable. Pensemos por ejemplo en llevar cable de fibra óptica a zonas rurales o espacios turísticos a decenas de kilómetros de distancia del nodo óptico más cercano. Por tanto, como el propio 3GPP indica TR36.932 [5], coexisten dos tipos de infraestructura en el backhaul: el tipo 1, considerado el caso ideal, y basado en fibra óptica y el tipo 2, no ideal, compuesto por líneas DSL y enlaces microondas. Para este segundo caso la gestión eficiente del ancho de banda es un aspecto crítico.

Históricamente, la investigación en comunicaciones inalámbricas y en particular en redes celulares ha prestado una gran atención a la gestión de recursos del interfaz radio, el *scheduling*, pero sólo recientemente se empieza a apreciar un cierto interés en el backhaul. En este capítulo estudiamos el impacto que un backhaul con ancho de banda limitado tiene en el rendimiento del scheduling, y mostramos cómo un scheduler radio que tenga en cuenta estas limitaciones puede mejorar este rendimiento. Ha habido algunos trabajos previos como [36], donde se estudia el impacto en LTE del backhaul limitado en relación a la reducción de la capacidad de cooperación de las estaciones a través del interfaz X2. Otros trabajos [37] y [38] se centran en el impacto del ancho de banda backhaul en la formación de clusters de estaciones para reducir la comunicación a través del backhaul. Otro funcionalidad básica del interfaz radio que se puede ver afectada por las limitaciones del backhaul es la asociación usuario-estación. Este tema está presente en [39] y [40].

A continuación explicamos los detalles técnicos de LTE relevantes en el problema estudiado y describimos la estrategia que planteamos para abordarlo.

Scheduling Radio en LTE

La trama radio en LTE tiene una duración de 10 ms, y está dividida a su vez, en 10 subtramas de 1 ms, como vemos en la Figura 5.1. La tecnología radio

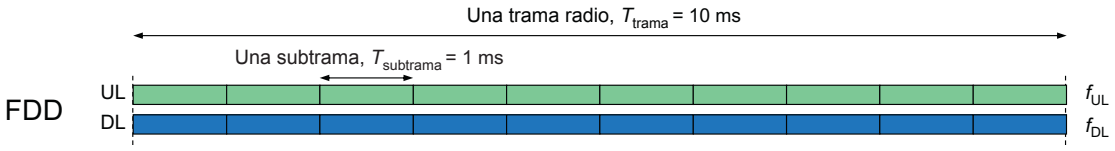

Figure 5.1: Diagrama de la trama radio LTE FDD.

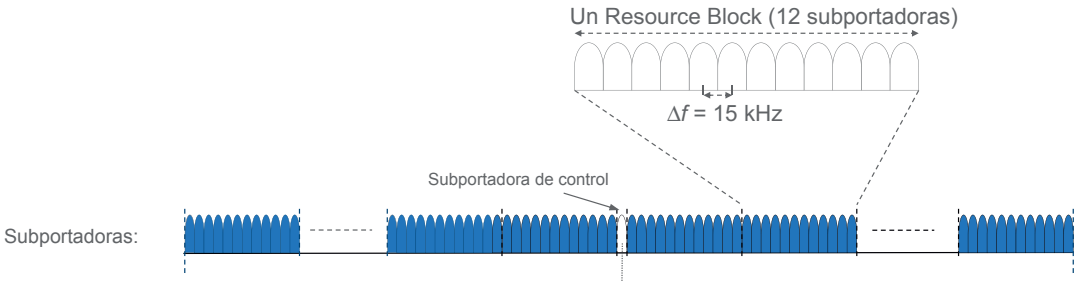

Figure 5.2: Subportadoras y Resource Blocks.

es Orthogonal Frequency Division Multiplexing (OFDM). En el dominio de la frecuencia, la subtrama está dividida en subportadoras de 15 MHz. A su vez, estas subportadoras se agrupan en conjuntos de 12, constituyendo lo que se denomina *resource blocks* (RBs), como se muestra en la Figura 5.2. Dentro de la trama OFDM, un RB ocupa 180 KHz en frecuencia y 0.5 ms en tiempo (correspondiente a la duración de media subtrama, también denominado *slot*).

En cada subtrama, los terminales de usuario informan de la calidad del canal radio, especificando la tasa de transmisión máxima por RB que permitiría al receptor decodificar correctamente (o con una tasa de errores aceptable, generalmente inferior al 10 %) la señal procedente de la estación base. Esta información se denomina *channel quality information* (CQI) y puede realizarse de forma periódica o aperiódica. El CQI puede estar referido a un conjunto de subportadoras (subbanda) o a todas las subportadoras en su conjunto. En este capítulo asumiremos que el transmisor conoce en todo momento el CQI de cada receptor, y que este CQI está referido a todas las subportadoras de la banda en la que se transmite la trama.

El problema del scheduling radio consiste en determinar en cada subtrama, y teniendo en cuenta el CQI de cada usuario, cuántos RB de la subtrama se asignarán a cada usuario activo en la celda. Esta decisión puede tener también en cuenta los datos almacenados en los buffers de cada usuario en el eNB o la tasa de transmisión

alcanzada hasta el momento por cada usuario. Generalmente se define un criterio global como maximizar la tasa de transmisión total de la celda, la tasa media por usuario, el retardo medio, etc. Un criterio ampliamente adoptado es el de maximización de la función de utilidad de la red (*Network Utility Maximization*, NUM), definida a su vez como la suma de las funciones de utilidad de los usuarios, que veremos a continuación. El criterio NUM permite introducir el concepto de justicia (fairness) en el reparto de recursos cuando se maximiza un criterio global como la tasa de transmisión de la celda.

Nótese que la tasa de transmisión de cada RB es variable en función del CQI de cada usuario, por lo que cada decisión de asignación en el canal radio implica, en general, una carga distinta en el interfaz S1. Este aspecto es crucial en cuanto a la carga del backhaul si éste tiene recursos limitados y por tanto debe tenerse en cuenta en la formulación del problema. En este capítulo la asignación de recursos en el canal radio se aborda a una escala temporal mayor que la trama radio, lo que permite considerar la tasa media que cada usuario puede obtener en dicho interfaz durante un periodo en el que la potencia media recibida se mantiene estable. En función del modelo de movilidad y de fading lento, este periodo puede oscilar entre unos pocos segundos y varios minutos. El resultado será por tanto una asignación a *alto nivel* de los recursos radio, interpretable como el porcentaje de recursos radio (RBs) que deberán asignarse a cada usuario durante el periodo considerado. Estos porcentajes podrían aplicarse subtrama a subtrama, trama a trama o a escalas mayores, en función de los objetivos de retardo establecidos, empleando para ello otro algoritmo que funcione a una escala temporal menor. Los motivos por los que abordamos el problema a una escala temporal superior a la duración de una trama son:

- Porque dota de sentido a las funciones de utilidad de los usuarios, ya que estos perciben el throughput a escalas mayores a la trama, al menos, del orden de segundos.
- Porque permite considerar un sistema estático en el que el problema de optimización NUM es más factible de analizar que en un sistema estocástico, en el que habría que recurrir a procesos de decisión de Markov mucho más complejos de abordar tanto analítica como numéricamente (*maldición de la dimensionalidad*).
- Porque aún a alto nivel esperamos obtener resultados que demuestren que el scheduling radio debe tener en cuenta las limitaciones de recursos en el backhaul.

5.2 Modelo del Sistema

El modelo considerado consiste en una única celda LTE, en la que nos centramos únicamente en la dirección downlink de transmisión. Como recursos radio (en downlink), la celda dispone de M subportadoras OFDM por las que transmite con una potencia fija. Como se ha explicado anteriormente, se considera un periodo de tiempo (del orden de unos pocos segundos) que denominaremos periodo de scheduling, durante el cual se realizan las siguientes asunciones:

1. El número de usuarios N en la celda permance fijo. El conjunto de usuarios en la celda es $\mathcal{N} = \{1, 2, \ldots, N\}$.
2. En cada subtrama en la que la estación transmite hacia un usuario, le asigna a este usuario todas las subportadoras.
3. Se emplea el modelo de tráfico *full buffer*, uno de los recomendados por el 3GPP en TR 36.814 [41], y que consiste en asumir que los usuarios activos en la celda siempre tienen datos disponibles en el buffer downlink.
4. La potencia media recibida por los usuarios permance constante, así como la interferencia media recibida de otras celdas. Como consecuencia, durante el periodo de scheduling a cada usuario puede alcanzar una tasa media de transmisión $R_i > 0$ en el canal radio, y este valor es conocido por la estación transmisora al principio del periodo.

En el backhaul consideramos únicamente el interfaz S1, y asumimos que durante el periodo de scheduling los recursos disponibles para el interfaz S1 de la celda estudiada están caracterizados por una tasa de transmisión C, conocida por la estación. Para garantizar los objetivos de calidad de servicio comprometidos, este valor C no podrá superarse durante el periodo de scheduling. El diagrama de la Figura 5.3 ilustra el sistema considerado.

El objetivo es determinar la proporción de recursos radio que se asignará a cada usuario durante el periodo de scheduling. Considerando que en cada subtrama se asignan todas las subportadoras a un solo usuario, el scheduling determina, para cada usuario $i \in \mathcal{N}$ la fracción de tiempo β_i que la estación transmite a ese usuario. Dado que se está considerando que el periodo de scheduling es varios órdenes de magnitud más largo que la duración de una subtrama, podemos asumir que la fracciones de tiempo $\{\beta_i\}_{i\in\mathcal{N}}$ son números reales no negativos.

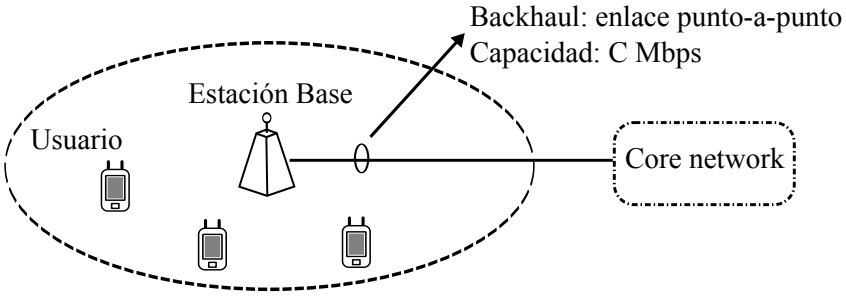

Figure 5.3: Diagrama del sistema modelado.

5.3 Análisis del Problema

Partimos de la noción de α-fairness, introducida en [42], y que se ha empleado extensamente para maximización de throughput con formulaciones NUM [43], [44]. Si x es el throughput ofrecido a un usuario en concreto, la utilidad obtenida por este usuario viene dada por $U_\alpha(x) = \frac{x^{1-\alpha}}{1-\alpha}$ si $\alpha > 1$ y $\alpha \neq 1$, y viene dada por $U_\alpha(x) = \log(x)$ si $\alpha = 1$. Este último caso es de especial interés, ya que corresponde al scheduling *proportional fair* (PF), que es el que se asumirá en el análisis que desarrollamos a continuación.

Es un resultado conocido [42] que maximizar la suma de las funciones de utilidad proportional fair es equivalente a maximizar la media geométrica del throughput de los usuarios:

$$\overline{T}(\{x_i\}_{i\in\mathcal{N}}) = \left(\prod_{i\in\mathcal{N}} x_i\right)^{\frac{1}{N}} \tag{5.1}$$

El scheduling debe determinar $\{\beta_i\}_{i\in\mathcal{N}}$, donde β_i es la fracción de tiempo que la estación base transmite al usuario i. Como datos de entrada, la estación dispone de la tasa de transmisión media R_i de cada usuario al inicio del periodo de scheduling. En el caso de scheduling α-fair, los valores de $\{\beta_i\}_{i\in\mathcal{N}}$ maximizan $\sum_{i\in\mathcal{N}} U_\alpha(\beta_i R_i)$, donde $\beta_i R_i$ es el throughput obtenido por el usuario i durante el periodo de scheduling. Por tanto, el scheduling α-fair con recursos limitados en backhaul se obtiene

resolviendo el siguiente problema de optimización:

$$\max_{\{\beta_i\}_{i\in\mathcal{N}}} \sum_{i\in\mathcal{N}} U_\alpha\left(\beta_i R_i\right)$$

s.t.

$$\sum_{i\in\mathcal{N}} \beta_i \leq 1 \tag{5.2}$$

$$\sum_{i\in\mathcal{N}} \beta_i R_i \leq C \tag{5.3}$$

$$\beta_i \geq 0 \; \forall i \in \mathcal{N} \tag{5.4}$$

El siguiente resultado nos da una propiedad importante de la solución para el caso de scheduling PF.

Proposición 1. Dadas unas tasas $R_i > 0$ para todo $i \in \mathcal{N}$, el scheduling PF ($\alpha = 1$) con backhaul limitado $C > 0$ tiene una solución óptima única con $\beta_i > 0$ para todo $i \in \mathcal{N}$.

Prueba. Dado que, para $\alpha = 1$, $U_\alpha\left(\beta_i R_i\right) = \log\left(\beta_i R_i\right)$, si existe una solución al problema, ésta debe satisfacer $\beta_i R_i > 0$ para todo i, y por tanto $\beta_i > 0$ para todo i. El conjunto de soluciones factibles (que cumplen todas las restricciones) y además cumplen $\beta_i > 0$ es no vacío, ya que todas las restricciones (5.2), (5.3) y (5.4) pueden cumplirse con todas las β_i estrictamente positivas. En consecuencia sabemos que existe al menos una solución óptima al problema con $\beta_i > 0$, y puesto que la función objetivo es estrictamente cóncava y las restricciones son convexas, sabemos que hay una única solución. ■

Planteamos el lagrangiano del problema de optimización:

$$L(\beta;\lambda,\mu,\nu) = -\sum_{i\in\mathcal{N}} U_\alpha\left(\beta_i R_i\right) + \lambda\left(\sum_{i\in\mathcal{N}} \beta_i - 1\right) + \mu\left(\sum_{i\in\mathcal{N}} \beta_i R_i - C\right) - \sum_{i\in\mathcal{N}} \nu_i \beta_i \tag{5.5}$$

Donde $\beta = (\beta_1, \ldots, \beta_N)$ y $\nu = (\nu_1, \ldots, \nu_N)$. Empleando el lagrangiano, L, podemos obtener las condiciones Karush-Kuhn-Tucker (KKT), necesarias para la

optimalidad [45]:

$$\frac{\partial L}{\partial \beta_i} = 0 \Longrightarrow \beta_i = \frac{1}{\mu R_i + \lambda - \nu_i} \; \forall i \in \mathcal{N} \tag{5.6}$$

$$\lambda \left(\sum_{i \in \mathcal{N}} \beta_i - 1 \right) = 0 \tag{5.7}$$

$$\mu \left(\sum_{i \in \mathcal{N}} \beta_i R_i - C \right) = 0 \tag{5.8}$$

$$\nu_i \beta_i = 0 \; \forall i \in \mathcal{N} \tag{5.9}$$

$$\lambda \geq 0; \; \mu \geq 0; \nu_i \geq 0 \; \forall i \in \mathcal{N} \tag{5.10}$$

Puesto que el problema primal consiste en optimizar una función cóncava definida en un conjunto convexo, sabemos que cualquier tupla de variables primales y duales (β, λ, μ, ν) que cumplan las condiciones KKT, contiene el óptimo primal β y el óptimo dual (λ, μ, ν), y el *duality gap* es cero [45]. Además, sabemos por la proposición 1, que el óptimo primal existe y es único, y que $\beta_i > 0$ para todo i. Por tanto, por la condición (5.9), $\nu_i = 0$ para todo i, por lo que en adelante prescindiremos de las variables duales ν_i. La condición de primer orden (5.6) queda así

$$\beta_i = \frac{1}{\mu R_i + \lambda} \; \forall i \in \mathcal{N} \tag{5.11}$$

que además impone la condición $\mu R_i + \lambda \neq 0$, por lo que ($\mu = 0$, $\lambda = 0$) no es una combinación posible. Por tanto quedan tres posibles combinaciones de variables duales ($\mu = 0$, $\lambda > 0$), ($\mu > 0$, $\lambda = 0$) y ($\mu > 0$, $\lambda > 0$).

El sistema resultante de las condiciones KKT podría resolverse numéricamente para obtener la solución PF óptima. Sin embargo, como veremos, el sistema se puede simplificar y, en ciertos casos, obtener soluciones cerradas.

Abordaremos las soluciones para cada posible combinación de variables duales.

Caso 1 ($\mu = 0$, $\lambda > 0$). Aplicando las condiciones KKT, esta hipótesis implica

$$\beta_i = \frac{1}{\lambda} \text{ y } \sum_{i \in \mathcal{N}} \beta_i = 1 \Rightarrow \beta_i = \frac{1}{N} \; \forall i \in \mathcal{N} \Rightarrow \lambda = N \tag{5.12}$$

Para que la solución primal sea factible, debe cumplir las restricciones del problema primal. En particular, la restricción (5.3), implica que $\sum_{i \in \mathcal{N}} \beta_i R_i \leq C$. Es decir, la capacidad del backhaul debe ser $C \geq \frac{1}{N} \sum_{i \in \mathcal{N}} R_i$. En este caso, la tupla ($\beta_i = \frac{1}{N}$ $\forall i$, $\mu = 0$, $\lambda = N$) cumple las condiciones KKT, y nos dan la solución primal que por la proposición 1 sabemos que es única.

Caso 2 ($\mu > 0$, $\lambda = 0$). Aplicando nuevamente las condiciones KKT tenemos que

$$\beta_i = \frac{1}{\mu R_i} \text{ y } \sum_{i \in \mathcal{N}} \beta_i R_i = C \Rightarrow \mu = \frac{N}{C} \Rightarrow \beta_i = \frac{C}{N R_i} \; \forall i \in \mathcal{N} \tag{5.13}$$

Al realizar la comprobación de las restricciones del problema primal, observamos que para cumplir la restricción (5.2), $\sum_{i \in \mathcal{N}} \frac{C}{N R_i} \leq 1$, y por tanto la capacidad del backhaul debe cumplir $C \leq \frac{N}{\sum_{i \in \mathcal{N}} \frac{1}{R_i}}$. En este caso, la tupla ($\beta_i = \frac{C}{N R_i} \; \forall i$, $\mu = \frac{N}{C}$, $\lambda = 0$) cumple las condiciones KKT, y hemos obtenido la solución primal única.

Caso 3 ($\mu > 0$, $\lambda > 0$). En este caso las condiciones (5.7) y (5.8) imponen que las restricciones primales (5.2) y (5.3) se cumplan con igualdad:

$$\sum_{i \in \mathcal{N}} \beta_i = 1 \text{ y } \sum_{i \in \mathcal{N}} \beta_i R_i = C \tag{5.14}$$

Sustituyendo β_i en estas igualdades por su expresión (5.11), obtenemos el siguiente sistema de ecuaciones no lineales

$$\sum_{i \in \mathcal{N}} \frac{R_i}{\mu R_i + \lambda} = C \tag{5.15}$$

$$\sum_{i \in \mathcal{N}} \frac{1}{\mu R_i + \lambda} = 1 \tag{5.16}$$

Que han de resolverse para $\mu > 0$ y $\lambda > 0$.

En conclusión, hemos encontrado dos valores críticos de la capacidad del backhaul:

$$c^* = \frac{N}{\sum_{i \in \mathcal{N}} \frac{1}{R_i}} \tag{5.17}$$

$$C^* = \frac{1}{N} \sum_{i \in \mathcal{N}} R_i \tag{5.18}$$

A partir de los cuales se pueden identificar tres casos en el sistema, cada uno de ellos caracterizado por una expresión para el scheduling PF óptimo:

- Caso 1 ($C \geq C^*$). Scheduling PF óptimo: $\beta_i = \frac{1}{N} \; \forall i$.
- Caso 2 ($C \leq c^*$). Scheduling PF óptimo: $\beta_i = \frac{C}{N R_i} \; \forall i$.
- Caso 3 ($c^* < C < C^*$). Scheduling PF óptimo: $\beta_i = \frac{1}{\mu R_i + \lambda} \; \forall i$, donde $\mu > 0$ y $\lambda > 0$ son las soluciones de

$$\sum_{i \in \mathcal{N}} \frac{R_i}{\mu R_i + \lambda} = C \tag{5.19}$$

$$\sum_{i \in \mathcal{N}} \frac{1}{\mu R_i + \lambda} = 1 \tag{5.20}$$

Benchmark: Scheduling PF sin Considerar el Backhaul

Uno de los objetivos de este capítulo es poner de manifiesto la necesidad de considerar los recursos del backhaul en el diseño de algoritmos de scheduling en el interfaz radio. Por tanto consideraremos como benchmark el scheduling PF en el que no se consideran restricciones en el backhaul, o lo que es lo mismo, asumiendo que $C = \infty$. Todos las propuestas existentes de scheduling en el interfaz radio que no consideran las limitaciones de backhaul están haciendo implícitamente esta asunción. En el modelo empleado, la solución del scheduling PF para $C = \infty$ es la del caso 1: $\beta_i = \frac{1}{N}\ \forall i$. Es decir, se obtiene un scheduler que reparte por igual el tiempo de transmisión de la trama entre todos los usuarios. Por este motivo, este scheduler se denomina *equal time*, y si la capacidad del backhaul es lo bastante grande, es un scheduler PF óptimo.

Sin embargo, si $C < C^*$, cuando el scheduler equal time trata de darles a todos los usuarios el mismo tiempo en el periodo de scheduling ($\beta_i = \beta_j$ para todo $i, j \in \mathcal{N}$), el backhaul se satura provocando un efecto de *starvation* en el interfaz radio. Matemáticamente, la congestión en el backhaul implica igualdad en la restricción (5.2) $\sum_{i\in\mathcal{N}} \beta_i R_i = C$, ya que el interfaz radio no puede transmitir más datos que los que se entregan por el backhaul. Si $\beta_i = \beta^*$ para todo $i \in \mathcal{N}$, tenemos que $\beta^* = \frac{C}{\sum_{i\in\mathcal{N}} R_i}$. Por tanto, podemos formalizar el scheduler equal time de la siguiente forma:

$$\beta = \min\left\{\frac{1}{N}, \frac{C}{\sum_{i\in\mathcal{N}} R_i}\right\} \tag{5.21}$$

En el siguiente apartado compararemos el scheduler equal time con el scheduler PF que considera el backhaul. Los resultados nos permitirán valorar la relevancia del backhaul en el scheduling radio.

5.4 Resultados Numéricos

En este apartado evaluaremos numéricamente el rendimiento del scheduler PF consciente del backhaul descrito en el apartado anterior, y lo compararemos con el benchmark, el scheduler equal-time, o lo que es lo mismo, un scheduler PF no-consciente del backhaul. Para las simulaciones consideramos un sistema LTE FDD, es decir basado en OFDMA, en el que cada resource block (RB) consiste en 12 subportadoras OFDM de 15 KHz, y 0.5 milisegundos en los que se transmiten 7 símbolos. La duración de un RB se denomina time-slot, y un periodo de 2 time-slots se denomina subtrama. La trama radio contiene 10 subtramas. Por tanto, en una trama cada subportadora puede transmitir $7 \times 2 \times 10 = 140$ símbolos.

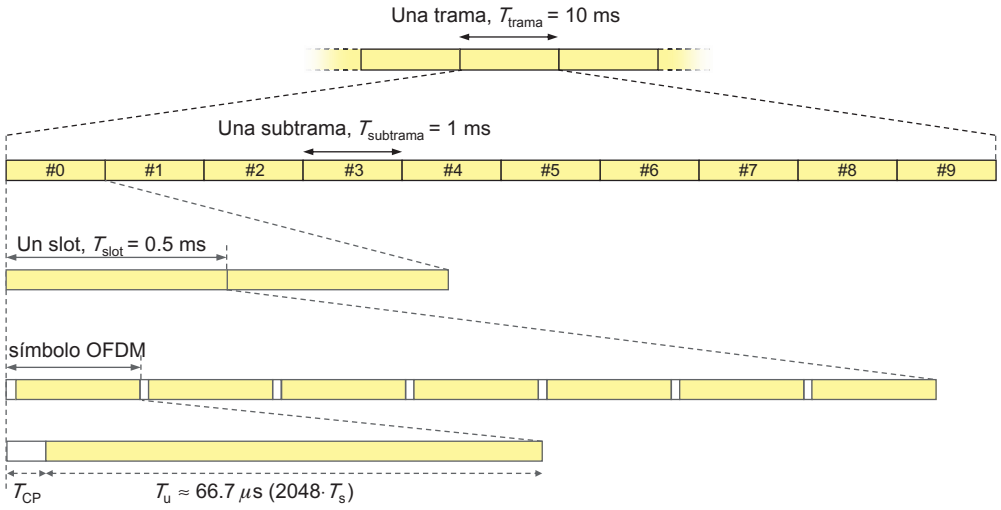

Figure 5.4: Desglose de la trama radio LTE.

La Figura 5.4 muestra un desglose en tiempo de la trama radio hasta el nivel de símbolo. El ancho de banda asignado a la celda determina el número de RBs por time-slot. Existen anchos de banda preestablecidos, por ejemplo 3 MHz, que corresponden a 15 RBs, 5 MHz son 25 RBs, 10 MHz son 50 RBs, etc. Consideramos dos casos: 10 MHz y 20 MHz, que reflejan dos situaciones en cuanto a los recursos disponibles en el interfaz radio. Las diferencias entre ambos escenarios se señalan en la Tabla 5.1.

Las tasas máxima y mínima que un usuario puede alcanzar se determinan de acuerdo a la especificación 3GPP 36.213 [46]. Nótese que dichas tasas corresponden a la asignación de todos los recursos de la celda a un solo usuario. Según este documento, en función de la calidad del canal downlink se selecciona el *modulation and coding scheme* MCS que, junto con el número de RBs asignadas por subtrama, permiten determinar el tamaño, en bits, del *transport block size* (TBS) correspondiente a datos de usuario. De acuerdo a la recomendación del 3GPP 36.814 [41], en ambos escenarios el número de usuarios activos durante el periodo de scheduling es 30, y el modelo de tráfico considerado es *full buffer*.

El parámetro de rendimiento evaluado es la media geométrica del throughput de los usuarios de la celda $\overline{T}$, de acuerdo a la expresión (5.1), ya que es la medida optimizada por el scheduling PF. Para estimar el valor de $\overline{T}$ se han promediado los resultados de 5000 simulaciones para cada valor de C, donde cada simulación correspondería a un *periodo de scheduling*. En cada simulación se han generado aleatoriamente las tasas de transmisión de cada usuario, empleando una

Table 5.1: Diferencias entre los escenarios considerados.

	Escenario 1	Escenario 2
Ancho de banda	10 MHz	20 MHz
RBs por time-slot	50	100
Transmisión	2x2 MIMO	4x4 MIMO
Tasa mínima (Mbps)	2.7	11.2
Tasa máxima (Mbps)	73.4	301.4

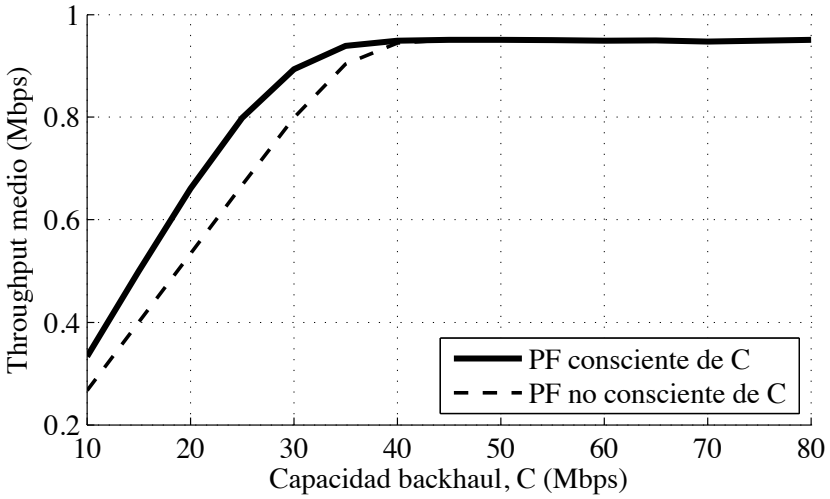

Figure 5.5: Media geométrica del throughput para el escenario 1.

distribución normal truncada a los valores de tasa mínima y máxima de cada escenario, y se ha empleado una desviación típica de 20 en el escenario 1 y de 50 en el escenario 2. Los resultados obtenidos en cada escenario se pueden observar en la Figura 5.5 y 5.6 respectivamente.

Como era de esperar, el throughput medio es mayor en el escenario 2 que en el 1, ya que el interfaz radio dispone de más recursos en el 2 que en el 1 para un mismo número de usuarios. No obstante, en ambos escenarios se observa la misma tendencia: por debajo de un cierto ancho de banda en el backhaul, el scheculer PF que no considera el backhaul es claramente subóptimo. Por ejemplo, en el escenario 1, con un ancho de banda C de 20 Mbps, el scheduler PF consciente del backhaul alcanza un throughput por usuario de 660 Kbps en el periodo de scheduling, mientras que el scheduler PF que no considera el backhaul quedaría en 533 Kbps. Es decir, no considerar el backhaul implicaría una pérdida del 20% de throughput por usuario. En el escenario 2, por ejemplo, si $C = 40$ Mbps el

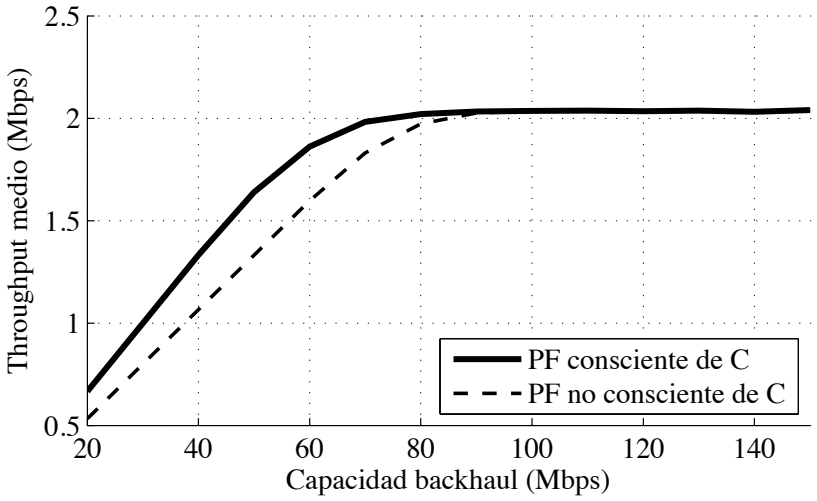

Figure 5.6: Media geométrica del throughput para el escenario 2.

scheduler consciente del backhaul alcanza 1.33 Mbps por usuairo, y el no consciente se queda en 1.06 Mbps. La pérdida de rendimiento en este caso es superior al 25%. En ambos escenarios, una vez superado un cierto valor de C, el backhaul ya no supone el cuello de botella del sistema y por tanto no se observan diferencias entre considerar o no el backhaul en el scheduler PF. Estos resultados prueban nuestra hipótesis de que es necesario replantear los algoritmos de asignación de recursos radio en redes en las que el backhaul puede convertirse eventualmente en el cuello de botella.

5.5 Conclusiones

En este capítulo hemos estudiado la influencia de los recursos del backhaul en el diseño de algoritmos de scheduling para el interfaz radio. Para ello planteamos un scheduling downlink a alto nivel, en el que se determina el porcentaje de recursos radio asignados a cada usuario en un periodo superior a la duración de una trama (típicamente del orden de segundos). Consideramos una sola celda conectada con un enlace backhaul punto-a-punto de capacidad limitada C. En la formulación del problema de optimización hemos incluido la restricción impuesta por los límites de capacidad del backhaul. El objetivo del problema es la maximización de la utilidad de la red, NUM, particularizando a una función de utilidad del tipo proportional fair, PF, para la que hemos obtenido la solución, en algunos casos cerrada, aplicando las condiciones Karush-Kunh-Tucker. Finalmente, hemos

llevado a cabo evaluaciones numéricas en dos escenarios realistas de LTE, empleando para ello las especificaciones 3GPP para capa física (interfaz radio). Los resultados obtenidos muestran que cuando el backhaul es el cuello de botella, si el scheduler radio no es consciente de la capacidad C, los usuarios pueden sufrir unas pérdidas de throughput superiores al 20 %.

CHAPTER 6

Asignación Coordinada de Recursos Radio y Backhaul

Resumen: En este último capítulo abordamos de forma conjunta la gestión de recursos radio y del backhaul, formulándola como un unico problema de maximización de la utilidad de red (NUM) en el que incorporamos también el control de tasa de las fuentes de datos. El problema ha de resolverse de forma distribuida en el eNB, el S-GW y las fuentes, para lo que empleamos una estrategia de descomposición dual. La estructura de la solución se traslada a un mecanismo basado en cálculo de índices de baja carga computacional, lo que permite generar decisiones de asignación de recursos subtrama a subtrama. A igualdad de throughput, evaluamos la eficiencia de los algoritmos por el retardo medio obtenido. Para mejorar este aspecto, aplicamos la estructura de la política de control óptima para dos colas en tándem. La estrategia de cálculo de índices permite generalizar el control óptimo a N parejas de colas en tándem y encajarlo de forma natural en la estructura distribuida del mecanismo. Los resultados numéricos verifican que aplicando el control óptimo de colas tándem, se obtiene el mínimo retardo tanto cuando las fuentes de datos ajustan su tasa como cuando transmiten a una tasa fija.

6.1 Introducción

En este último capítulo abordamos de forma conjunta la gestión de recursos radio y del backhaul. Los resultados de los capítulos anteriores nos han mostrado la ventaja de contar con información relativa al interfaz radio para el diseño de

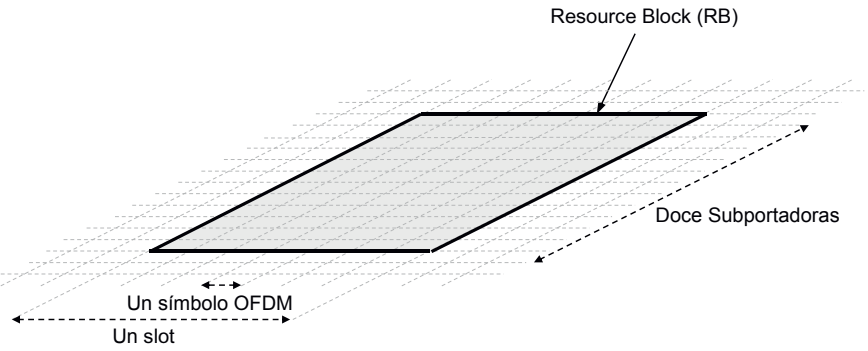

Figure 6.1: Resource Block (RB) del interfaz radio LTE.

algoritmos de control de flujo en el backhaul (Capítulo 4), de la misma forma que también es beneficioso considerar las limitaciones del backhaul para diseñar el scheduling radio (Capítulo 5). Esto nos lleva a plantear en un único problema la asignación de todos los resursos del sistema, incluyendo también el control de la tasa de transmisión de las fuentes de datos. Este último aspecto, que supone una novedad con respecto a los capítulos anteriores, hace que el modelo sea más general y completo, y aporta nuevos resultados.

El escenario estudiado es similar al del Capítulo 5, y consta de una estación base radio (enhanced Node B, eNB), un Service Gateway (S-GW) y N terminales. Se considera únicamente la transmisión en sentido *downlink*, y a cada terminal le corresponde un flujo de datos downlink. En el eNB consideramos una única celda LTE FDD, cuyo interfaz radio emplea tecnología OFDM. El interfaz S1, que conecta el eNB con el S-GW, está soportado por una red de transporte (backhaul) con ancho de banda limitado. El sistema de control debe determinar, en cada etapa de decisión, cuántos recursos del interfaz radio y del backhaul se debe asignar a cada uno de los N flujos. Como se vió en el capítulo anterior, la unidad mínima de asignación de recursos en LTE es el resource block (RB), representado en la Figura 6.1. Como en capítulos anteriores, cuando nos referimos al backhaul, los términos *asignación de recursos*, *scheduling* y *control de flujo* son sinónimos en este capítulo.

El objetivo es diseñar un algoritmo capaz de tomar decisiones en tiempo real, en el sentido en que las etapas de decisión coinciden con las subtramas dentro de la trama OFDM, de 1 ms de duración, tal como estipula el 3GPP [47]. Por tanto, a diferencia del Capítulo 5, la solución obtenida no es una asignación de los recursos *en promedio* a una escala temporal superior la subtrama, sino que se

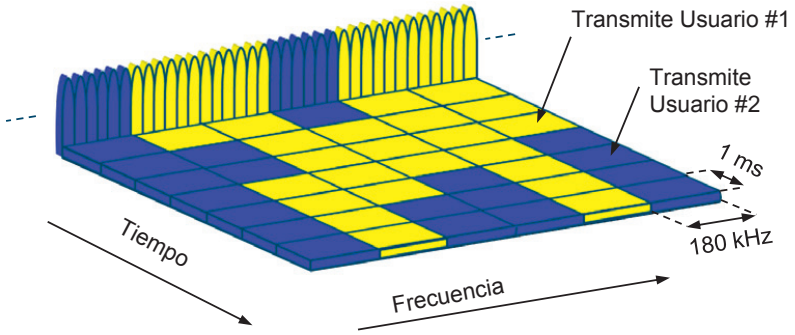

Figure 6.2: Ejemplo de asignación de recursos radio en OFDM para dos usuarios.

propone un mecanismo que decide en cada subtrama cuántos recursos se asignan a cada flujo en función del estado del sistema al inicio de la subtrama. La Figura 6.2 muestra un ejemplo de asignación de recursos radio. El estado del sistema varía aleatoriamente de una subtrama a la siguiente, y está compuesto por la longitud de las N colas de datos en el eNB, las N colas del S-GW, la calidad del canal radio para cada terminal y la tasa de transmisión de cada fuente, consituyendo un complejo proceso estocástico.

Partimos de la formulación de un problema de maximización de la utilidad de la red (NUM). En este capítulo no consideramos una función de utilidad específica, sino genérica. Dado que la formulación engloba a todo el sistema, el problema no puede resolverse en un solo nodo, sino de forma distribuida entre el eNB, el S-GW y las fuentes de datos. Para ello, la estrategia de búsqueda de soluciones hace uso de la *descomposición dual* del problema. El análisis matemático basado en descomposición del problema NUM nos proporciona la estructura de la solución, que trasladamos a un mecanismo basado en índices, y de baja complejidad (lineal en N).

Dado que cualquier algoritmo capaz de estabilizar el tráfico entrante de datos alcanza el mismo throughput, para discriminar la eficiencia de los algoritmos empleamos el *retardo medio por usuario*. Con esta perspectiva proponemos una mejora adicional que consiste en incorporar, en el mecanismo obtenido mediante NUM, la estructura de la política de control óptimo para dos colas en tándem. El objetivo de esta política es minimizar el retardo, y se obtiene mediante el análisis de un modelo fluido de colas [48]. Como veremos, la estrategia de cálculo de índices permite generalizar el control óptimo a N parejas de colas en tándem, y encajarlo de forma natural en la estructura distribuida del mecanismo.

Resumiendo, las contribuciones de este capítulo son:

1. Diseñamos un mecanismo que asigna en cada subtrama los recursos radio y backhaul. Este mecanismo es distribuido, de baja carga computacional y orientado a la maximización de la utilidad de la red.
2. Incorporamos en dicho mecanismo la política de control óptimo de dos colas en tándem, generalizándola a N parejas de colas. La evaluación numérica verifica que esta modificación consigue minimizar el retardo, manteniendo el throughput.
3. Para ilustrar la importancia de que el scheduling radio y backhaul esté coordinado, comparamos las dos propuestas con otras soluciones óptimas en throughput pero que operan localmente en cada nodo.

El resto del capítulo se organiza de la siguiente forma: El apartado 6.2 detalla el modelo del sistema. En el apartado 6.3 se plantea el problema NUM y se desarrolla el algoritmo de asignación distribuido. En el apartado 6.4 se modifica el algoritmo incorporando la política de control óptima de dos colas tándem. Los resultados numéricos se muestran en el apartado 6.5 para fuentes de tasa de transmisión fija y variable. Finalmente el apartado 6.6 resume las conclusiones de este capítulo.

6.2 Modelo del Sistema

El escenario estudiado consiste en una celda LTE de un eNB, un S-GW, y N terminales. Entre el eNB y el S-GW se define el interfaz S1, que asumimos soportado por una red de transporte (backhaul) con un ancho de banda C limitado. La dirección de la transmisión es downlink, de forma que a cada terminal le corresponde un flujo de datos. Cada fuente de datos i transmite a una tasa x_i. Para cada flujo i hay una cola de datos $q_i^{(1)}$ en el S-GW y una cola $q_i^{(2)}$ en el eNB. La Figura 6.3 ilustra el sistema considerado, donde r_i denota la tasa de transmisión radio entre la estación y el terminal i.

El modelo empleado implica las siguientes suposiciones:

- La tasa de transmisión r_i se selecciona de un conjunto de posibles valores para los cuales la tasa de error en recepción se puede asumir igual a 0. Esta suposición se apoya en el mecanismo existente en LTE por el que los terminales informan al eNB de la calidad de su enlace radio.
- El número de usuarios N en la celda permance fijo. El conjunto de usuarios en la celda es $\mathcal{N} = \{1, 2, \ldots, N\}$.

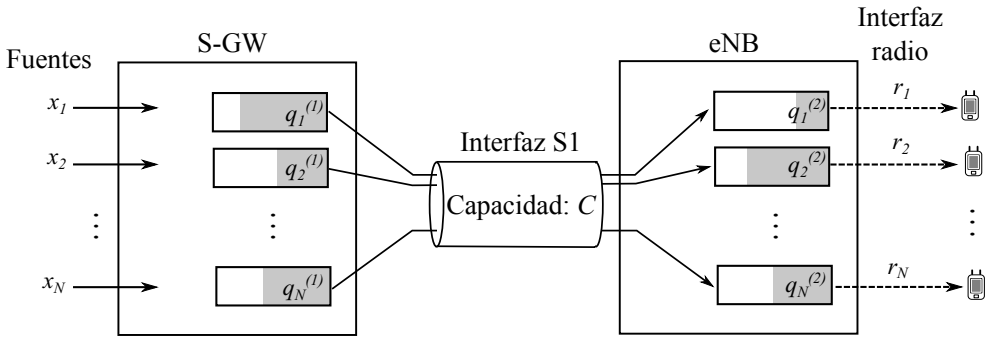

Figure 6.3: Esquema del sistema estudiado.

- La calidad del enlace radio para cada terminal i, varía de una subtrama a otra, pero en cada subtrama la calidad es la misma para todas las subportadoras radio.
- El retardo de propagación en el backhaul es inferior a la duración de una subtrama.

En cada subtrama t, el estado de las colas de cada flujo de datos i se representa como $q_i^{(1)}(t)$ y $q_i^{(2)}(t)$. Análogamente, $x_i(t)$ denota los bits que llegan al S-GW procedentes de la fuente i. El algoritmo de asignación debe decidir, en cada subtrama t, cuantos bits $b_i^{(1)}(t)$ y $b_i^{(2)}(t)$ se deben transmitir del S-GW y del eNB, para todo $i \in \mathcal{N}$. Las ecuaciones que caracterizan la evolución de las colas de i son:

$$q_i^{(1)}(t+1) = q_i^{(1)}(t) + x_i(t) - b_i^{(1)}(t) \tag{6.1}$$

$$q_i^{(2)}(t+1) = q_i^{(2)}(t) + b_i^{(1)}(t) - b_i^{(2)}(t) \tag{6.2}$$

Los conjuntos de valores $\left\{b_i^{(1)}(t)\right\}_{i\in\mathcal{N}}$ y $\left\{b_i^{(2)}(t)\right\}_{i\in\mathcal{N}}$ están sujetos a restricciones. En primer lugar, $b_i^{(j)}(t) \leq q_i^{(j)}(t)$ para $j = 1, 2$. En el caso del S-GW $\sum_{i\in\mathcal{N}} b_i^{(1)}(t) \leq C\tau$, donde τ es la duración en segundos de una subtrama (1 ms para LTE). En el caso del eNB, determinar $b_i^{(2)}(t)$ es equivalente a seleccionar la tasa de transmision $r_i(t)$ para el terminal i durante la subtrama t. A su vez, $\{r_i(t)\}_{i\in\mathcal{N}}$ debe estar dentro del conjunto de valores admitidos en el interfaz radio, como se explica en la siguiente subsección.

Región de Capacidad del Interfaz Radio

La calidad de la señal recibida en los terminales depende del fading y la interferencia, y por tanto varía de una subtrama a otra. Típicamente, la calidad de la señal

es su SINR (Signal to Interference and Noise Ratio). Para cada terminal, denominamos **calidad del enlace** a la calidad de la señal recibida en dicho terminal. La calidad de un enlace en una subtrama puede tomar, en principio, cualquier valor continuo no negativo. Sin embargo, desde un punto de vista práctico, se emplea una variable discreta, denominada **estado del enlace**, para representar la calidad del enlace. El estado del enlace está expresado en términos de tasa máxima a la que el terminal puede recibir datos de la estación con una baja tasa de error.

En cada subtrama, el estado de cada enlace toma valores, aleatoriamente, de un conjunto discreto de M' estados, y cada terminal informa a la estación del valor de dicho estado mediante el mecanismo CQI (Channel Quality Information). Denominamos **estado del interfaz radio** al vector (en $\mathbb{R}^N$) que contiene los estados de todos los N usuarios conectados a la estación. El estado del interfaz radio toma, por tanto, valores de un conjunto discreto con $M = M'^N$ posibles estados. Gracias al mecanismo CQI, la estación conoce, en cada subtrama, el estado del interfaz radio.

Denominamos r al vector de posibles tasas de transmisión de datos, donde r_i es la tasa de transmisión de la estación al terminal $i \in \mathcal{N}$. Los recursos del interfaz radio se han de repartir en fracciones discretas (RBs, Resource Blocks). El CQI de cada usuario determina la tasa de transmisión asociada a cada RB asignado a dicho usuario. Por tanto, dado que las posibles asignaciones de RBs a usuarios es un conjunto discreto, el vector r toma valores de un conjunto discreto $\mathcal{R}_m$ cuando el canal está en el estado m (que determina los CQIs de cada usuario). Nótese que a un usuario se le pueden asignar 0 RBs, por lo que dado un $r \in \mathcal{R}_m$, cualquier vector r' en el que algunos elementos r'_i sean iguales a r_i y otros sean iguales 0 también pertenece a $\mathcal{R}_m$.

Consideremos un elemento r del conjunto $\mathcal{R}_m$. Definimos $\alpha_{m,r}$ como la fracción de tiempo que la estación emplea el vector r cuando el intefaz radio está en el estado m. El conjunto de posibles tasas medias de transmisión alcanzables en el estado m viene dado por la envolvente convexa de $\mathcal{R}_m$:

$$Co(\mathcal{R}_m) = \left\{ \mu : \mu = \sum_{r \in \mathcal{R}_m} \alpha_{m,r} r, \text{ tal que } \alpha_{m,r} \geq 0 \text{ y } \sum_{r \in \mathcal{R}_m} \alpha_{m,r} = 1 \right\}. \quad (6.3)$$

El estado del interfaz radio varia de una subtrama a otra. Denominamos π_m a la fracción de tiempo que el interfaz se encuentra en el estado m. Dado que hay M posibles estados, tenemos que

$$\sum_{1\leq m\leq M} \pi_m = 1. \tag{6.4}$$

Definimos la región de capacidad $\mathcal{C}$ del interfaz radio como

$$\mathcal{C} = \left\{ \mu : \mu = \sum_{1\leq m\leq M} \pi_m r_m, \text{ tal que } r_m \in Co(\mathcal{R}_m) \right\}. \tag{6.5}$$

Una asignación de recursos radio es factible o realizable si el vector de tasas medias se encuentra contenido en la región de capacidad, $\mu \in \mathcal{C}$. Dada una política de asignación de recursos α, la tasa media de un usuario i viene dada por

$$\mu_i = \sum_{1\leq m\leq M} \pi_m \sum_{r\in\mathcal{R}_m} \alpha_{m,r} r_i. \tag{6.6}$$

Dado un vector de tasas de llegada λ, se dice que es *sostenible*, o también *estable en estado estacionario*, si existe un algoritmo de scheduling bajo el que

$$\lim_{Q\to\infty} \lim_{t\to\infty} P\left(|q(t)| \geq Q\right) = 0, \tag{6.7}$$

donde $|q(t)|$ es la suma de las longitudes de todas las colas de la red en el instante t.

Es importante recordar que es imposible encotrar un algoritmo de scheduling para el que $\lambda \notin \mathcal{C}$ sea sostenible. Éste es un resultado clásico y fundamental en el diseño de algoritmos de asignación de recursos. Vease por ejemplo [[49], Teorema 5.3.1].

6.3 Algoritmo Basado en NUM

En esta sección planteamos el problema NUM (maximización de la utilidad de la red) específico para el escenario descrito en el apartado anterior, que incluye las fuentes de datos, el reparto de recursos en el backhaul y el reparto de recursos radio. Para cada terminal $i \in \mathcal{N}$, denominamos x_i a la tasa media a la que transmite su fuente de datos, y denotamos por $\mu_i^{(1)}$ y $\mu_i^{(2)}$ a las tasas medias obtenidas por i en el backhaul y en el interfaz radio respectivamente. Estas variables deben cumplir las siguientes restricciones:

$$x_i \leq \mu_i^{(1)} \tag{6.8}$$

$$\mu_i^{(1)} \leq \mu_i^{(2)} \tag{6.9}$$

Estas restricciones nos indican que la tasa de transmisión de datos de la fuente i no puede superar la tasa de transmisión hacia el usuario i en el backhaul, y a su vez, la tasa de llegada de datos a la estación radio procedentes de i, no puede superar a la tasa de transmisión entre la estación y el terminal i. Como vemos, esencialmente se trata de restricciones de estabilidad.

Para hacer el sistema más general consideramos que las fuentes pueden ajustar su tasa de transmisión, y por tanto x_i, $i \in \mathcal{N}$ también son variables del problema. Asociando a cada fuente una función de utilidad cóncava $U_i(x_i)$, podemos formular el problema NUM de la siguiente forma:

$$\max_{x,\alpha^{(1)},\alpha^{(2)}\geq 0} \sum_{i\in\mathcal{N}} U_i\left(x_i\right)$$

s.t.

$$x_i \leq \mu_i^{(1)} \text{ para } i \in \mathcal{N} \tag{6.10}$$

$$\mu_i^{(1)} \leq \mu_i^{(2)} \; \forall i \in \mathcal{N} \tag{6.11}$$

$$\mu_i^{(1)} = \alpha_i^{(1)} C \; \forall i \in \mathcal{N} \tag{6.12}$$

$$\sum_{i\in\mathcal{N}} \alpha_i^{(1)} \leq 1 \tag{6.13}$$

$$\alpha_i^{(1)} \geq 0; \forall i \in \mathcal{N} \tag{6.14}$$

$$\mu_i^{(2)} = \sum_{1\leq m\leq M} \pi_m \sum_{r\in\mathcal{R}_m} \alpha_{m,r}^{(2)} r_i \; \forall i \in \mathcal{N} \tag{6.15}$$

$$\sum_{r\in\mathcal{R}_m} \alpha_{m,r}^{(2)} = 1 \; \forall m \tag{6.16}$$

$$\alpha_{m,r}^{(1)} \geq 0 \; \forall m, \text{ y } \forall r \in \mathcal{R}_m \tag{6.17}$$

Observamos que las restricciones (6.15), (6.16) y (6.17) nos indican esencialemente que el vector de tasas medias de transmisión en el interfaz radio $\mu^{(2)}$ debe encontrarse en la región de capacidad de dicho interfaz, que denominaremos $\mathcal{C}^{(2)}$. Por tanto, podemos expresar estas restricciones de forma más compacta:

$$\mu^{(2)} \in \mathcal{C}^{(2)}, \tag{6.18}$$

donde $\mathcal{C}^{(2)}$ se define como en 6.5.

La región de capacidad del interfaz S1 (backhaul) se define directamente como

$$\mathcal{C}^{(1)} = \left\{ \mu^{(1)} : \sum_{i\in\mathcal{N}} \mu_i^{(1)} \leq C, \text{ siendo } \mu_i^{(1)} \geq 0, \forall i \right\}. \tag{6.19}$$

Y permite expresar las restricciones (6.12), (6.13) y (6.14) como

$$\mu^{(1)} \in \mathcal{C}^{(1)}, \tag{6.20}$$

El problema NUM inicial podemos plantearlo así:

$$\max_{x,\mu^{(1)},\mu^{(2)}\geq 0} \sum_{i\in\mathcal{N}} U_i\left(x_i\right)$$

s.t.

$$x_i \leq \mu_i^{(1)} \text{ para } i \in \mathcal{N} \tag{6.21}$$

$$\mu_i^{(1)} \leq \mu_i^{(2)} \; \forall i \in \mathcal{N} \tag{6.22}$$

$$\mu^{(1)} \in \mathcal{C}^{(1)} \tag{6.23}$$

$$\mu^{(2)} \in \mathcal{C}^{(2)}. \tag{6.24}$$

A continuación desarrollamos la estrategia de resolución del problema. Para cada i definimos dos multiplicadores de Lagrange, $\phi_i^{(1)}$ y $\phi_i^{(2)}$, asociados a las restricciones (6.21) y (6.22) respectivamente. Añadimos estas restricciones al objetivo con sus respectivos multiplicadores y obtenemos

$$\max_{x,\mu^{(1)},\mu^{(2)}\geq 0} \sum_{i\in\mathcal{N}} U_i\left(x_i\right) - \sum_{i\in\mathcal{N}} \phi_i^{(1)}\left(x_i - \mu_i^{(1)}\right) - \sum_{i\in\mathcal{N}} \phi_i^{(2)}\left(\mu_i^{(1)} - \mu_i^{(2)}\right) \tag{6.25}$$

s.t.

$$\mu^{(1)} \in \mathcal{C}^{(1)} \tag{6.26}$$

$$\mu^{(2)} \in \mathcal{C}^{(2)}. \tag{6.27}$$

Para valores fijos de $\phi_i^{(1)}$ y $\phi_i^{(2)}$, la maximización se puede realizar de forma independiente para cada una de las variables x, $\mu^{(1)}$ y $\mu^{(2)}$, quedando el problema descompuesto en los siguientes tres problemas de optimización:

$$\textbf{sub-problema 1:} \sum_{i\in\mathcal{N}} \max_{x_i\geq 0}\left(U_i\left(x_i\right) - \phi_i^{(1)} x_i\right), \tag{6.28}$$

$$\textbf{sub-problema 2:} \max_{\mu^{(1)}\in\mathcal{C}^{(1)}} \sum_{i\in\mathcal{N}} \left(\phi_i^{(1)} - \phi_i^{(2)}\right)\mu_i^{(1)}, \tag{6.29}$$

$$\textbf{sub-problema 3:} \max_{\mu^{(2)}\in\mathcal{C}^{(2)}} \sum_{i\in\mathcal{N}} \phi_i^{(2)}\mu_i^{(2)}, \tag{6.30}$$

Si se conocen los multiplicadores, cada uno de estos problemas se pueden resolver localmente en cada uno de los nodos implicados:

- Subproblema 1: Cada fuente $i \in \mathcal{N}$ resuelve localmente $\max_{x_i\geq 0}\left(U_i\left(x_i\right) - \phi_i^{(1)} x_i\right)$, para lo que debe conocer el multiplicador correspondiente $\phi_i^{(1)}$.

- Subproblema 2: Se resuelve en el nodo que gestiona la dirección downlink del interfaz S1 (el S-GW). Este nodo debe conocer todos los multiplicadores $\phi^{(1)}$ y $\phi^{(2)}$.
- Subproblema 3: Se resuelve en la estación radio (el eNB), que solo debe conocer los multiplicadores $\phi^{(2)}$.

En este punto, asumiendo que disponemos de un mecanismo para estimar $\phi^{(1)}$ y $\phi^{(2)}$ y distribuirlos convenientemente entre las fuentes de datos, el S-GW y el eNB, el algoritmo de asignación conjunta de recursos radio y backhaul queda definido con la resolución de los subproblemas 1, 2 y 3 en las entidades correspondientes. No obstante, para llegar a un algoritmo de interés práctico, aún quedan por resolver dos aspectos no triviales: Primero, debemos encontrar un mecanismo para computar o estimar adecuadamente $\phi^{(1)}$ y $\phi^{(2)}$, y distribuirlos conforme a los requerimientos de cada subproblema. Segundo, al igual que en capítulo 5, la solución del problema NUM nos proporciona los valores óptimos en cuanto al reparto de recursos *en promedio*. Por ese motivo en dicho capítulo considerabamos un periodo de scheduling largo en relación a la duración de una subtrama. Sin embargo, para que el algoritmo tenga un valor práctico, debemos ofrecer un mecanismo que tome decisiones de asignación subtrama a subtrama, tal como especifica el 3GPP.

Abordemos en primer lugar la estimación de los multiplicadores de Lagrange $\phi^{(1)}$ y $\phi^{(2)}$. Para ello, comenzaremos considerando el problema de optimización (6.25),(6.26),(6.27), en el que añadíamos las restricciones a la función objetivo (relajación lagrangiana), y al que aplicaremos *descomposición dual*. La función dual se define como

$$D(\phi^{(1)}, \phi^{(2)}) = \sup_{x \geq 0, \mu^{(1)} \in \mathcal{C}^{(1)}, \mu^{(2)} \in \mathcal{C}^{(2)}} L\left(x, \mu^{(1)}, \mu^{(2)}, \phi^{(1)}, \phi^{(2)}\right) \tag{6.31}$$

donde

$$L\left(x, \mu^{(1)}, \mu^{(2)}, \phi^{(1)}, \phi^{(2)}\right) = \sum_{i \in \mathcal{N}} U_i\left(x_i\right) - \sum_{i \in \mathcal{N}} \phi_i^{(1)}\left(x_i - \mu_i^{(1)}\right) - \sum_{i \in \mathcal{N}} \phi_i^{(2)}\left(\mu_i^{(1)} - \mu_i^{(2)}\right)$$

es la función objetivo(6.25), denominada lagrangiano, o de forma más precisa lagrangiano parcial, ya que (6.26),(6.27) quedan fuera del mismo. El problema dual consiste en minimizar $D(\phi^{(1)}, \phi^{(2)})$ con respecto a $\phi^{(1)}$ y $\phi^{(2)}$. El algoritmo de descomposición dual consiste, en este caso, en repetir la siguiente secuencia:

1. Resolver los subproblemas 1, 2 y 3 (en paralelo), obteniendo los valores de x, $\mu^{(1)}$ y $\mu^{(2)}$ que maximizan $L\left(x, \mu^{(1)}, \mu^{(2)}\phi^{(1)}, \phi^{(2)}\right)$ y por tanto proporcionan la función dual $D(\phi^{(1)}, \phi^{(2)})$.

2. Actualizar el valor de las variables duales $\phi^{(1)}$ y $\phi^{(2)}$, mediante un mecanismo de *descenso del gradiente*:

$$\phi^{(1)}(t+1) = \phi^{(1)}(t) - \epsilon \nabla_{\phi^{(1)}} D(\phi^{(1)}(t), \phi^{(2)}(t)) \tag{6.32}$$
$$\phi^{(2)}(t+1) = \phi^{(2)}(t) - \epsilon \nabla_{\phi^{(2)}} D(\phi^{(1)}(t), \phi^{(2)}(t)) \tag{6.33}$$

donde ϵ es un factor de ajuste para garantizar la convergencia.

Considerando los valores $x(t)$, $\mu^{(1)}(t)$ y $\mu^{(2)}(t)$ que optimizan (6.31) en la iteración t, obtenemos

$$\nabla_{\phi^{(1)}} D(\phi^{(1)}(t), \phi^{(2)}(t)) = -(x(t) - \mu^{(1)}(t)) \tag{6.34}$$
$$\nabla_{\phi^{(2)}} D(\phi^{(1)}(t), \phi^{(2)}(t)) = -(\mu^{(1)}(t) - \mu^{(2)}(t)) \tag{6.35}$$

Por lo que la iteración de descenso del gradiente, para cada $i \in \mathcal{N}$ queda como sigue:

$$\phi_i^{(1)}(t+1) = \phi_i^{(1)}(t) + \epsilon(x_i(t) - \mu_i^{(1)}(t)) \tag{6.36}$$
$$\phi_i^{(2)}(t+1) = \phi_i^{(2)}(t) + \epsilon(\mu_i^{(1)}(t) - \mu_i^{(2)}(t)) \tag{6.37}$$

Los valores $\mu^{(1)}(t)$ y $\mu^{(2)}(t)$ están definidos como tasas promedio. Sin embargo buscamos las asignaciones para cada subtrama t. Como explicábamos en la sección 6.2, el algoritmo debe determinar los bits transmitidos de cada flujo i en el backhaul y en el interfaz radio, durante la subtrama t ($b_i^{(1)}(t)$ y $b_i^{(2)}(t)$ para $i \in \mathcal{N}$). Recordemos las ecuaciones (6.1) (6.2) que caracterizan la evolución de las colas. Para la cola i-ésima en el S-GW tenemos

$$q_i^{(1)}(t+1) = q_i^{(1)}(t) + x_i(t) - b_i^{(1)}(t) \tag{6.38}$$

y en el eNB

$$q_i^{(2)}(t+1) = q_i^{(2)}(t) + b_i^{(1)}(t) - b_i^{(2)}(t) \tag{6.39}$$

Como vemos, la evolución de las colas $q_i^{(1)}$ y $q_i^{(2)}$ es muy similar a la de los multiplicadores de Lagrange. Una aproximación habitual ([49]) y conveniente en este caso es sustituir los multiplicadores de Lagrange por la longitud de las colas.

Veamos a continuación cómo determinar los valores de $b^{(1)}(t)$ y $b^{(2)}(t)$ en cada subtrama. Comencemos por el interfaz radio. En cada subtrama t, el eNB conoce el estado de todas las colas downlink $q^{(2)}(t)$, y el CQI de cada usuario. El CQI permite al eNB saber cuántos bits de i puede transmitir por el RB k, $r_i^{(k)}(t)$, si k se asigna a i. Supongamos que sólo se pudiera seleccionar un usuario en cada

subtrama. En este caso, si se selecciona i, los bits transmitidos serían $b_i^{(2)}(t) = \sum_{1\leq k\leq K} r_i^{(k)}(t)$, y $b_j^{(2)}(t) = 0$ para $j \neq i$. El problema se reduciría a seleccionar, en cada subtrama, el usuario al que asignar todo el interfaz radio. Consideremos por tanto el subproblema 3, sustituyendo $\phi^{(2)}$ por $q^{(2)}(t)$ y $\mu^{(2)}$ por $b_i^{(2)}(t)$. La política *greedy* (la que busca maximizar la función de utilidad en cada etapa t) para este problema es la siguiente:

En cada t, el eNB selecciona el usuario i tal que

$$i \in \arg\max_j q_j^{(2)}(t) b_j^{(2)}(t), \tag{6.40}$$

escogiendo aleatoriamente en caso de empate.

Esta política es conocida como *MaxWeight*. Fué estudiada por primera vez en [50], y generalizada en múltiples artículos posteriores [51], [52], [53], [54]. Se sabe que esta política es óptima en throughput (vease por ejemplo [49], Teorema 5.3.2). Esencialmente es una política que resuelve de forma sencilla (complejidad $O(N)$) un problema de programación dinámica estocástica, sorteando la *maldición de la dimensionalidad* que afecta a este tipo de problemas. La estructura de la solución es una estructura tipo índice (muy similar la regla c-μ, o índices de Klimov [48]).

No obstante, en nuestro caso no estamos asumiendo que todos los RBs hayan de ir a un solo usuario en cada subtrama. Por el contrario, la decisión de asignación consiste precisamente en decidir cuántos RBs se asignan a cada usuario en cada subtrama, por lo que debemos decidir cómo adaptar MaxWeight a este caso. Contemplamos dos opciones:

Primera opción: En cada t aplicar MaxWeight directamente, seleccionando el i obtenido en (6.40). Si este usuario tuviera menos datos en cola que los que se pueden transmitir por el interfaz radio, se le asignan los resursos necesarios de acuerdo a

$$k_i \in \arg\min_{k'} \sum_{1\leq k\leq k'} r_i^{(k)}(t) \tag{6.41}$$

s.t.

$$\sum_{1\leq k\leq k'} r_i^{(k)}(t) \geq q_i^{(2)}(t). \tag{6.42}$$

Tras asignar los k_i RBs resultantes de (6.41) al usuario i, se selecciona el siguiente usuario j con mayor índice MaxWeight, al que se le pueden asignar, como máximo, $K - k_i$ RBs, y así sucesivamente. El proceso se repite hasta asignar los K RBs.

Segunda opción: Dividimos cada etapa de decisión t en K rondas de asignación (una por RB en la subtrama). En cada ronda k se selecciona el usuario i con mayor

índice MaxWeight de la siguiente forma

$$i \in \arg\max_j q_j^{(2)}(t) r_j^{(k)}(t), \tag{6.43}$$

y actualizamos su cola de acuerdo a

$$q_i^{(2)}(t) = \left(q_i^{(2)}(t) - r_i^{(k)}(t) \right)^+. \tag{6.44}$$

De esta forma los índices van cambiando cada vez que se asigna un RB.

Estas dos opciones fueron estudiadas en [55], donde los autores muestran que la segunda opción (a la que denominan *server side greedy*, SSG), es superior a la primera ya que es óptima en throughput (ver Teorema 5 de [55]) y garantiza una menor longitud media de las colas. La complejidad de este algoritmo es $O(KN)$.

Para la asignación de recursos en el backhaul, nos centramos en el subproblema 2. Al sustituir $\phi^{(1)}$ por $q^{(1)}$ y $\phi^{(2)}$ por $q^{(2)}$, obtenemos un problema análogo al subproblema 3, por lo que podemos adoptar una estrategia análoga a SSG para resolverlo. Dividimos en H bloques de c bits ($c = C\tau/H$), la cantidad de datos que el backhaul puede transmitir durante un intervalo t. Estos H bloques pueden ser considerados como una unidad de datos encapsulable en los paquetes o tramas de la red de transporte que da soporte al backhaul. Por tanto, en cada t, tenemos H rondas de decisión en las que el S-GW selecciona la cola i con mayor índice

$$i \in \arg\max_j \left(q_j^{(1)}(t) - q_j^{(2)}(t) \right), \tag{6.45}$$

y actualizamos su cola de acuerdo a

$$q_i^{(1)}(t) = \left(q_i^{(1)}(t) - c \right)^+. \tag{6.46}$$

Es decir, como en el caso anterior aplicamos una política de tipo índice en cada ronda de decisión. En este caso los índices son $(q_i^{(1)}(t) - q_i^{(2)}(t))$ para cada $i \in \mathcal{N}$. A esta política se la conoce como *Backpressure*, [49], y ha sido ampliamente estudiada e incluso llevada a la práctica en redes inalámbricas malladas [56], [57], [58].

Por tanto, aplicar la formulación NUM al problema de la asignación conjunta de recursos radio y backhaul da como resultado un algoritmo en el que el eNB emplea MaxWeight y el S-GW emplea Backpressure. La tabla 6.1 resume el algoritmo resultante.

Resulta de especial interés resaltar cuál es el intercambio de información requerido para el funcionamiento de este algoritmo. Al final de cada etapa t:

Table 6.1: Algoritmo basado en NUM

En cada etapa t los siguientes algoritmos se ejecutan en paralelo en las entidades indicadas.

Fuente i-ésima

Input: $q_i^{(1)}(t)$.

La tasa de transmisión $x_i(t)$ de la fuente i se obtiene resolviendo el siguiente problema:

$$x_i(t) \in \arg \max_{0 \le x_i \le x_{\max}} (U_i(x_i) - q_i^{(1)}(t)x_i),$$

y a continuación genera $a_i(t)$ paquetes, donde $a_i(t)$ es una v.a. entera tal que la tasa media de datos generados sea $x_i(t)$.

S-GW

Input: $q^{(1)}(t)$, $q^{(2)}(t)$, $a(t)$.

1) Para cada $h \in \{1, \ldots, H\}$:
   - selecciona $i \in \arg\max_j \left(q_j^{(1)}(t) - q_j^{(2)}(t)\right)$,
   - actualiza cola i: $q_i^{(1)}(t) = \left(q_i^{(1)}(t) - c\right)^+$,
   - actualiza los datos enviados: $b_i^{(1)}(t) = \min\left(q_i^{(1)}(t), b_i^{(1)}(t) + c\right)$.
2) Actualiza las colas: $q^{(1)}(t+1) = q^{(1)}(t) + a(t)$.

eNB

Input: $q^{(2)}(t)$, $b^{(1)}(t)$, y $r_i^{(k)}(t)$ para todo $i \in \mathcal{N}$ y $0 \le k \le K$.

1) Para cada $k \in \{1, \ldots, K\}$:
   - selecciona $i \in \arg\max_j q_j^{(2)}(t) r_j^{(k)}(t)$,
   - actualiza cola i: $q_i^{(2)}(t) = \left(q_i^{(2)}(t) - r_i^{(k)}(t)\right)^+$,
   - actualiza los datos enviados:
   $$b_i^{(2)}(t) = \min\left(q_i^{(2)}(t), b_i^{(2)}(t) + r_i^{(k)}(t)\right).$$
2) Actualiza las colas: $q^{(2)}(t+1) = q^{(2)}(t) + b^{(1)}(t)$.

- El eNB debe enviar al S-GW el estado de las colas $q^{(2)}(t)$.
- El S-GW debe enviar a las fuentes el estado de las colas $q^{(1)}(t)$.

El primer intercambio implica usar un mecanismo de señalización similar al empleado en HSDPA para que el NB asigne créditos a la RNC. Lo que este modelo nos dice, en contraposición al modelo del capítulo 5, es que para maximizar el throughput no es imprescindible que el interfaz radio sea consciente del backhaul, ya que este objetivo puede alcanzarse haciendo que el backhaul sea consciente del interfaz radio, es decir, con un adecuado control de flujo en el interfaz S1.

6.4 Algoritmo Basado en el Control Óptimo de Colas Tándem

Sabemos que MaxWeight y Backpressure son algoritmos óptimos en throughput [49]. No obstante, esto no quiere decir que sean también óptimos en cuanto a retardo. Por ejemplo, consideremos de nuevo las dos opciones de implementación de MaxWeight en el interfaz radio (subproblema 3 del apartado anterior). La primera opción consistía en aplicar MaxWeight a todos los RBs como si fueran un sólo recurso radio, y el segundo aplicaba iterativamente MaxWeight en cada RB. Ambas opciones son óptimas en throughput, pero se ha comprobado que la segunda opción tiene un retardo medio inferior a la primera [55]. Por este motivo se ha adoptado la versión iterativa tanto para MaxWeight en el eNB como para Backpressure en el S-GW. No obstante, para el segundo caso, los resultados de [55] no son directamente extrapolables, y se plantea la cuestión de si es posible mejorar la política empleada en el S-GW en términos de retardo. El objetivo de este apartado es encontrar un algoritmo de baja complejidad para el S-GW, orientado a minimizar el retardo.

En el apartado anterior planteamos el problema NUM conjunto y los descomponíamos en tres subproblemas (fuentes, S-GW y eNB). En este apartado optamos por una estrategia de descomposición distinta: considerar por separado cada pareja de colas $q_i^{(1)}$ y $q_i^{(2)}$, para un flujo $i \in \mathcal{N}$. El sistema de dos **colas en tándem** es bien conocido y ha sido estudiado ampliamente en [59], [60], [61], [48].

Se sabe que la estructura de la política óptima, en términos de retardo medio o longitud media de las colas, es del tipo función de conmutación (o *switching curve*) [59]. Definiendo el coste instantáneo como una función lineal de $q_i^{(1)}$ y $q_i^{(2)}$, la política óptima consiste en transmitir con la tasa máxima en la cola 1 si $q_i^{(2)} < S(q_i^{(1)})$, y no transmitir en la cola 1 si $q_i^{(2)} \geq S(q_i^{(1)})$, donde la función S es lo que se denomina *switching curve*. En este apartado adoptamos la metodología de S. Meyn [48] para computar la *switching curve*, basada en modelos fluidos de colas, y esta función nos servirá de base para redefinir unos nuevos índices alternativos a (6.45). El motivo de emplear esta metodología en vez de MDPs (procesos de decisión de Markov), es que los modelos fluidos no requieren una caracterización estocástica de los procesos de llegadas y servicio en las colas, sino tan solo una estimación de los valores medios, simplificando enormemente la obtención de unas políticas que, para el caso de las colas tándem, se aproximan en gran medida a las políticas MDP [48].

El modelo fluido de colas consiste en una ecuación diferencial ordinaria (ODE) que puede entenderse como una descripción del comportamiento de los valores *medios* del proceso estocástico. Con un cierto abuso de notación, consideremos que $q_i^{(1)}$ y $q_i^{(2)}$ representan las longitudes de dos colas en tándem **en una representación determinista**, correspondientes a un flujo i. Su comportamiento determinista responde a la siguiente ecuación:

$$\begin{bmatrix} q_i^{(1)}(t) \\ q_i^{(2)}(t) \end{bmatrix} = \begin{bmatrix} y_i^{(1)} \\ y_i^{(2)} \end{bmatrix} + \begin{bmatrix} -\mu_i^{(1)} & 0 \\ \mu_i^{(1)} & -\mu_i^{(2)} \end{bmatrix} \begin{bmatrix} z_i^{(1)}(t) \\ z_i^{(2)}(t) \end{bmatrix} + \begin{bmatrix} \lambda_i^{(1)} \\ 0 \end{bmatrix} t \tag{6.47}$$

donde $y_i^{(1)}$ y $y_i^{(2)}$ representan el estado inicial de las colas en $t = 0$ ($y_i^{(1)} = q_i^{(1)}(0)$ y $y_i^{(2)} = q_i^{(2)}(0)$), los valores de $\mu_i^{(1)}$ y $\mu_i^{(2)}$ indican las tasas medias de servicio de cada cola, y $\lambda_i^{(1)}$ representa la tasa de llegada de datos al sistema (en particular a la cola 1). Las variables $z_i^{(1)}(t)$ y $z_i^{(2)}(t)$ representan el tiempo total que se han servido las colas 1 y 2 respectivamente hasta el instante t. Estas variables representan el **proceso acumulativo de asignación**. Si una cola j se sirve durante un periodo de tiempo $(t_1 - t_0)$ entonces $\left(z_i^{(j)}(t_1) - z_i^{(j)}(t_0)\right) = (t_1 - t_0)$. Para cualquier periodo de tiempo y cualquier asignación siempre se cumple que $\left(z_i^{(j)}(t_1) - z_i^{(j)}(t_0)\right) \leq (t_1 - t_0)$. La ecuación (6.47) se puede expresar de forma más compacta:

$$q_i(t) = y_i + B_i z_i(t) + \lambda_i t \tag{6.48}$$

Se define el vector de tasa de asignación como

$$\zeta_i(t) = \frac{d}{dt} z_i(t), \text{ para } t \geq 0. \tag{6.49}$$

Usando esta notación, el modelo fluido se puede expresar como una ODE

$$\frac{d}{dt} q_i(t) = B_i \zeta_i(t) + \lambda_i \tag{6.50}$$

Nótese que una cola puede atenderse a una tasa máxima de una unidad de tiempo por unidad de tiempo, y por tanto $0 \leq \zeta_i^{(j)}(t) \leq 1$.

El objetivo es definir una regla de control que a partir del estado $q_i(t)$ proporcione el vector de tasa de asignación $\zeta_i(t)$. Existen diversos criterios de control (*non-idling*, óptimo en tiempo, coste óptimo con horizonte infinito con o sin descuento, *greedy*, y *greedy* óptima en tiempo [48]). En nuestro caso, optamos por el control de coste óptimo con horizonte infinito, que se define como aquél que, dada una condición inicial $q(0) = y$, minimiza la siguiente **función de coste total**

$$J(x) = \int_0^\infty g(q(t))dt \tag{6.51}$$

donde $g(q(t))$ es una **función de coste**, y $q(t)$ está descrita por una ODE del tipo (6.48). Puesto que nuestro control óptimo debe minimizar el retardo, la función de coste empleada es

$$g(x) = |x| = \sum x_i, \tag{6.52}$$

denominada función de inventario total.

Para un vector de tasa de asignación ζ y un vector de longitudes de cola y, se define el tiempo de drenaje $T(y)$ como el mínimo valor de T tal que $y + (B\zeta + \lambda)\,T = 0$, es decir $T(y)$ representa el tiempo que tarda el sistema descrito por (6.48) en llegar al estado en el que todas las colas son iguales a 0. Consideremos el sistema de colas tándem. Dado $y = q_i(t)$, cuando el cuello de botella es la cola 1, el tiempo de drenaje es

$$T^{(1)}\left(q_i(t)\right) = \frac{q_i^{(1)}(t)}{\mu_i^{(1)} - \lambda_i^{(1)}}, \tag{6.53}$$

y si el cuello de botella es la cola 2 el tiempo de drenaje es

$$T^{(2)}(q_i(t)) = \frac{q_i^{(1)}(t) + q_i^{(2)}(t)}{\mu_i^{(2)} - \lambda_i^{(1)}}. \tag{6.54}$$

Para la función de coste (6.52), el control óptimo para horizonte finito de colas tándem coincide con el control óptimo en tiempo (que minimiza el tiempo de drenaje) [48]. En particular, la estructura de dicho control óptimo es la siguiente:

$$\zeta_i^{(1)}(t) = 1, \text{ si y solo si } T^{(2)}(q_i(t)) \leq T^{(1)}\left(q_i(t)\right) \text{ y } q_i(t) \neq 0, \tag{6.55}$$

y $\zeta_i^{(2)}(t) = 1$ siempre que $q_i^{(2)}(t) \neq 0$. Es decir, el control óptimo tiende a equilibrar el tiempo de drenaje de las dos colas. Aplicando (6.53) y (6.54), podemos expresar el control (6.55) como sigue

$$\zeta_i^{(1)}(t) = 1, \text{ si y solo si } q_i^{(2)}(t) \leq \frac{\mu_i^{(2)} - \mu_i^{(1)}}{\mu_i^{(1)} - \lambda_i^{(1)}} q_i^{(1)}(t) \text{ y } q_i(t) \neq 0, \tag{6.56}$$

Como vemos, la *switching curve* es una función lineal de $q_i^{(1)}(t)$.

A continuación extendemos este control óptimo a un conjunto de colas tándem que comparten recursos. En primer lugar observamos que, para una cola del interfaz radio $q_i^{(2)}(t)$, el control óptimo prescribe $\zeta_i^{(2)}(t) = 1$ siempre que dicha cola no esté vacía. Asumimos por tanto que las colas de este interfaz están siempre atendidas, a la tasa $\mu_i^{(2)}$ que les garantice el algoritmo MaxWeight iterativo visto en el apartado anterior.

Para las colas del backhaul, empleamos la misma metodología iterativa basada en índices del apartado anterior. En cada iteración, el control (6.56) se puede reformular como un índice que indique el grado de prioridad con el que se debe asignar a cada cola $i \in \mathcal{N}$ el bloque de c bits ($\zeta_i^{(1)}(t) = 1$ sobre dicho recurso), de tal forma que el recurso se asigne a

$$i \in \arg\max_j \left(\frac{\mu_j^{(2)} - \mu_j^{(1)}}{\mu_j^{(1)} - \lambda_j^{(1)}} q_j^{(1)}(t) - q_j^{(2)}(t) \right) \tag{6.57}$$

que puede expresarse también como

$$i \in \arg\max_j \left(T^{(1)}\left(q_i(t)\right) - T^{(2)}(q_i(t)) \right). \tag{6.58}$$

La interpretación de este índice es que, de todas las colas del S1, se comienza a drenar con más prioridad aquella con mayor diferencia entre el tiempo de drenaje de la cola del S1 y el de la cola del interfaz radio. Puesto que el control óptimo trata de igualar estos dos valores, el control conjunto trata de acercar las colas tándem a este punto de funcionamiento. En comparación con el algoritmo Backpressure, (6.57) emplea mucha más información. Los valores de $\mu_i^{(1)}$, $\mu_i^{(2)}$ y $\lambda_j^{(1)}$ deben ser estimados por el S-GW. Nos referiremos a las estimaciones de estos parámetros en cada estapa de decisión t como $\tilde{\mu}_i^{(1)}(t)$, $\tilde{\mu}_i^{(2)}(t)$ y $\tilde{\lambda}_j^{(1)}(t)$. Empleando estas estimaciones, el algoritmo de asignación coordinada de recursos radio y backhaul queda como se muestra en la tabla 6.2. Con respecto al algoritmo NUM, únicamente ha cambiado la parte correspondiente al S-GW, aunque para una mayor claridad se muestra también la parte ejecutada por las fuentes y el eNB.

Como en el algoritmo NUM, resumimos a continuación el intercambio de información requerido para el funcionamiento de este algorimo.

- El eNB debe enviar al S-GW el estado de las colas $q^{(2)}(t)$ y la estimación $\tilde{\mu}_i^{(2)}(t)$.
- El S-GW debe enviar a las fuentes el estado de las colas $q^{(1)}(t)$.

Además, el S-GW debe realizar las estimaciones $\tilde{\mu}_i^{(1)}(t)$ y $\tilde{\lambda}_i^{(1)}(t)$. Como se verá en siguiente apartado, unas estimaciones sencillas de estos parámetros, con bajo o nulo coste computacional, resultan igualmente efectivas.

Table 6.2: Algoritmo basado en control óptimo de colas tándem

En cada etapa t los siguientes algoritmos se ejecutan en paralelo en las entidades indicadas.

Fuente i-ésima

Input: $q_i^{(1)}(t)$.

La tasa de transmisión $x_i(t)$ de la fuente i se obtiene resolviendo el siguiente problema:

$$x_i(t) \in \arg \max_{0 \le x_i \le x_{\max}} (U_i(x_i) - q_i^{(1)}(t)x_i),$$

y a continuación genera $a_i(t)$ paquetes, donde $a_i(t)$ es una v.a. entera tal que la tasa media de datos generados sea $x_i(t)$.

S-GW

Input: $q^{(1)}(t)$, $q^{(2)}(t)$, $a(t)$, $\tilde{\mu}_j^{(1)}(t)$, $\tilde{\mu}_j^{(2)}(t)$, $\tilde{\lambda}_j^{(1)}(t)$.

1) Para cada $h \in \{1, \ldots, H\}$:

- selecciona $i \in \arg\max_j \left(\dfrac{\tilde{\mu}_j^{(2)}(t) - \tilde{\mu}_j^{(1)}(t)}{\tilde{\mu}_j^{(1)}(t) - \tilde{\lambda}_j^{(1)}(t)} q_j^{(1)}(t) - q_j^{(2)}(t) \right)$,
- actualiza cola i: $q_i^{(1)}(t) = \left(q_i^{(1)}(t) - c \right)^+$,
- actualiza los datos enviados: $b_i^{(1)}(t) = \min \left(q_i^{(1)}(t), b_i^{(1)}(t) + c \right)$.

2) Actualiza las colas: $q^{(1)}(t+1) = q^{(1)}(t) + a(t)$.

eNB

Input: $q^{(2)}(t)$, $b^{(1)}(t)$, y $r_i^{(k)}(t)$ para todo $i \in \mathcal{N}$ y $0 \le k \le K$.

1) Para cada $k \in \{1, \ldots, K\}$:

- selecciona $i \in \arg\max_j q_j^{(2)}(t) r_j^{(k)}(t)$,
- actualiza cola i: $q_i^{(2)}(t) = \left(q_i^{(2)}(t) - r_i^{(k)}(t) \right)^+$,
- actualiza los datos enviados:
$$b_i^{(2)}(t) = \min \left(q_i^{(2)}(t), b_i^{(2)}(t) + r_i^{(k)}(t) \right).$$

2) Actualiza las colas: $q^{(2)}(t+1) = q^{(2)}(t) + b^{(1)}(t)$.

6.5 Evaluación de Rendimiento

Metodología

En este apartado evaluamos numéricamente el rendimiento de los algoritmos mostrados en los dos apartados anteriores. Para distinguirlos, nos referiremos a ellos por el mecanismo de scheduling empleado en el S-GW, que es donde se diferencian. Nos referiremos al primero como *Backpressure* y al segundo como *Switching Curve*. Además los comparamos con dos benchmarks:

- **Benchmark 1**: El eNB y el S-GW aplican cada uno, por separado, un

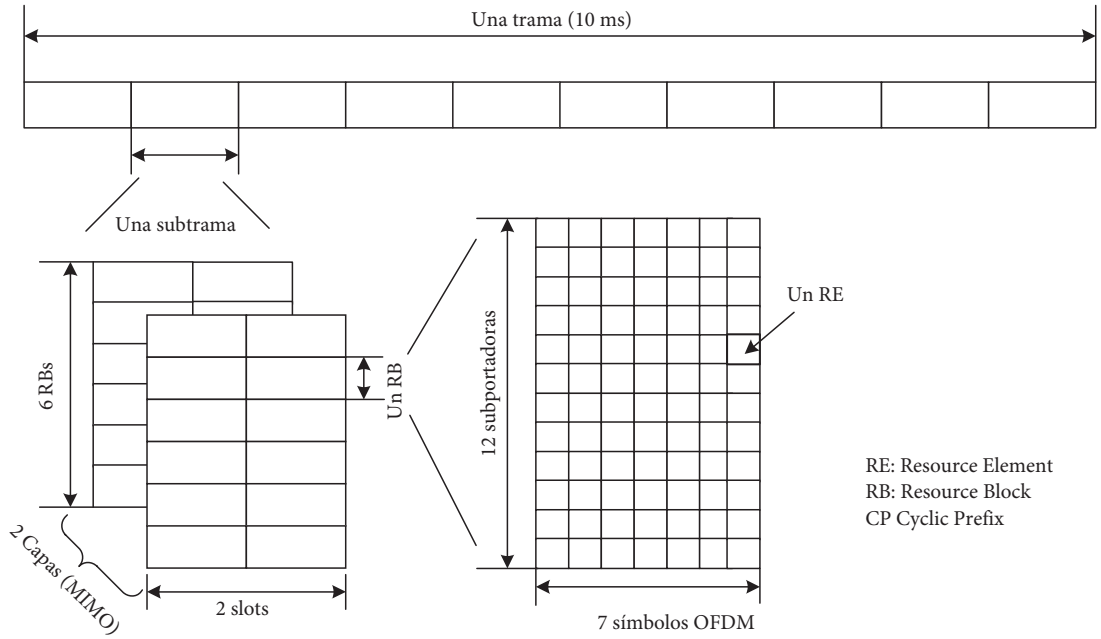

Figure 6.4: Desglose de la trama radio LTE para un interfaz radio de 1.4 MHz de ancho de banda (6 RBs), el mínimo especificado por el 3GPP.

scheduling Proportional Fair, en sus respectivos interfaces.

- **Benchmark 2**: El eNB y el S-GW aplican cada uno, por separado, un scheduling MaxWeight, en sus respectivos interfaces.

Como en el capítulo 5, las simulaciones consideran una celda LTE FDD, cuyo interfaz radio emplea tecnología OFDM. Cada resource block (RB) consiste en 12 subportadoras OFDM de 15 KHz, y 0.5 milisegundos en los que se transmiten 7 símbolos. Cada etapa de decisión t corresponde a una subtrama, cuya duración temporal es de 1 ms. Cada subtrama se divide en 2 slots de 0.5 ms, que coinciden con la duración de 2 RBs. La trama radio contiene 10 subtramas. Por tanto, en una trama cada subportadora puede transmitir $7 \times 2 \times 10 = 140$ símbolos. La Figura 6.4 muestra un ejemplo de trama radio desglosada en frecuencia y tiempo para un ancho de banda de 6 RBs. El escenario simulado considera un interfaz radio con ancho de banda igual a 10 MHz, correspondiente a 50 RBs, por lo que en cada subtrama el algoritmo de scheduling debe asignar $50 \times 2 = 100$ RBs. El número de usuarios activos en la celda es $N = 30$.

Las tasas máxima y mínima que un usuario puede alcanzar se determinan de acuerdo a la especificación 3GPP 36.213 [46]. Nótese que dichas tasas corresponden a la asignación de todos los recursos de la celda a un solo usuario. Según

Table 6.3: Configuración de los parámetros radio en las simulaciones.

Parámetro	Configuración
Ancho de banda	10 MHz
RBs por subtrama	100
Transmisión	2x2 MIMO
Tasa mínima (Mbps)	2.7
Tasa máxima (Mbps)	73.4

este documento, en función de la calidad del canal downlink se selecciona el *modulation and coding scheme* MCS que, junto con el número de RBs asignadas por subtrama, permiten determinar el tamaño, en bits, del *transport block size* (TBS) correspondiente a datos de usuario. Existen 27 posibles combinaciones de MCS, correspondientes a los 27 niveles de CQI que puede señalizar el terminal. Estos niveles varían aleatoriamente de una subtrama a otra. En el simulador se emplea un modelo de Markov para la generación de niveles similar a [62]: Para cada terminal i, si en t el estado del canal es $e_i(t) \in \{1, \ldots, 27\}$, el estado siguiente $e_i(t+1)$ se determina generando una v.a. discreta $\delta_i(t)$ con distribución uniforme entre -2 y 2, de forma que

$$e_i^+(t+1) = \max\left(e_i(t) + \delta_i(t), 1\right) \tag{6.59}$$

$$e_i(t+1) = \min\left(e_i^+(t+1), 27\right) \tag{6.60}$$

Ante la previsible igualdad en throughput de los algoritmos propuestos, el parámetro de rendimiento que centra nuestro interés es el **retardo medio** desde el S-GW hasta el terminal de usuario. Para el cómputo de este retardo se asume que un paquete enviado por el S-GW en t estará disponible en el eNB para su envío por el interfaz radio en $t+1$, es decir, el backhaul introduce un retardo de 1 ms (la duración de subtrama). El retardo medio $\bar{D}$ se compone por tanto de tres factores:

$$\bar{D} = \bar{D}^{(1)} + 1 + \bar{D}^{(2)}, \tag{6.61}$$

donde $\bar{D}^{(1)}$ es el retardo medio en las colas del S-GW, y $\bar{D}^{(2)}$ es el retardo medio en las colas del eNB.

Fuentes de Tasa de Transmisión Constante

El primer escenario de simulación incorpora fuentes que **transmiten datos con una tasa fija** de 800 Kbps. Esto supone una tasa agregada de 24 Mbps, que

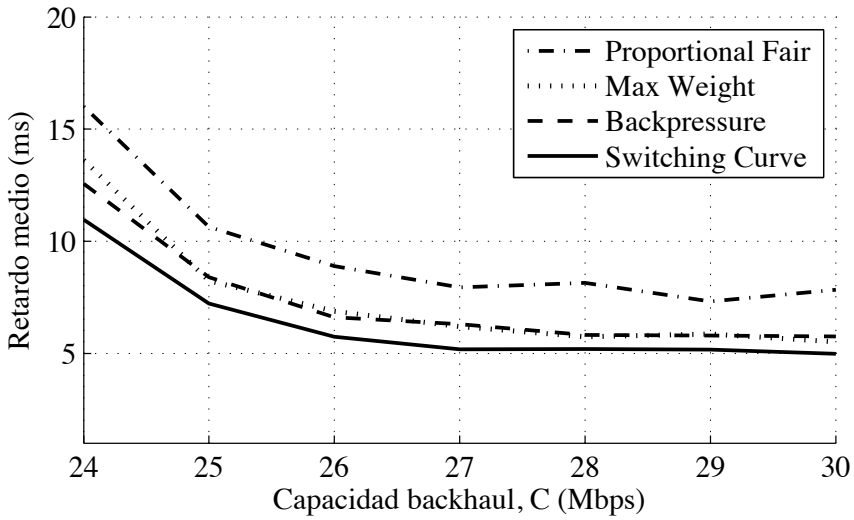

Figure 6.5: Retardo medio obtenido con cada algoritmo considerando fuentes de tasa constante.

está dentro del margen que puede transmitirse de forma estable por el interfaz radio (dado que su capacidad media ronda los 38 Mbps). Se asume que el S-GW conoce la tasa de cada fuente ($\tilde{\lambda}_i^{(1)}(t) = 800$ Kbps para todo t). Como estimación $\tilde{\mu}_i^{(2)}(t)$ se emplea la tasa instantánea anunciada por el terminal mediante el CQI, y la estimación $\tilde{\mu}_i^{(1)}(t)$ es igual a la capacidad del interfaz S1 (C) repartida proporcionalmente entre las colas $q^{(1)}(t)$ no vacías. Variando la capacidad C del backhaul desde el mínimo valor aceptable (24 Mbps) hasta 30 Mbps, obtenemos, para cada algoritmo (Proportional Fair, MaxWeight, Backpressure y Switching Curve) los retardos mostrados en la Figura 6.5. En todos los casos el backhaul es, en promedio, el cuello de botella del sistema. Para cada valor de C, los resultados se han obtenido promediando 100 ejecuciones de 20 segundos cada una, siendo el intervalo de confianza inferior al 3% con un grado de confianza del 95%.

En esta figura se observa que el algoritmo Switching Curve es el que alcanza sistemáticamente un menor retardo. Las medidas de throughput no son informativas ya que todos los algoritmos alcanzan el mismo valor (800 Kbps por usuario como cabría esperar). Es sorprendente hasta cierto punto que emplear Backpressure en el S-GW no mejore el retardo con respecto a MaxWeight. No obstante, recordemos que Backpressure se enmarcaba en una formulación NUM que incluía fuentes que ajustaban su tasa de transmisión. En el segundo escenario consideramos fuentes de este tipo para mostrar la mejora que aporta este algoritmo.

Fuentes Elásticas

En el segundo escenario se consideran fuentes elásticas, es decir, fuentes de tasa de transmisión variable y que ajustan su tasa de transmisión x_i con el objetivo de maximizar su función de utilidad $U_i(x_i)$, tal como se consideran en el apartado 6.3 y cuyo algoritmo (incluido en en las tablas 6.1 y 6.2) reproducimos a continuación:

La tasa de transmisión $x_i(t)$ de la fuente i se obtiene resolviendo el siguiente problema:

$$x_i(t) \in \arg \max_{0 \leq x_i \leq x_{\max}} (U_i(x_i) - q_i^{(1)}(t)x_i), \tag{6.62}$$

y a continuación genera $a_i(t)$ paquetes, donde $a_i(t)$ es una v.a. entera tal que la tasa media de datos generados sea $x_i(t)$.

La función de utilidad empleada por la fuentes es

$$U_i(x_i) = -\frac{2}{T_i^2 x_i}, \tag{6.63}$$

donde T_i representa el retardo de ida y vuelta o *Round Trip Time* (RTT). Esta función de utilidad modela el control de tasa empleado por TCP-Reno [49].

La Figura 6.6 muestra el retardo obtenido con cada algoritmo para distintos valores de C. En este caso la diferencia entre emplear Backpressure o Switching Curve en el S-GW y emplear los otros dos algoritmos de benchmark si que es muy significativa: los algorimos propuestos mantienen el retardo constante en valores muy bajos mientras que el retardo de los algoritmos de benchmark aumenta conforme aumenta la capacidad del backhaul.

Este comportamiento se debe a que las fuentes aumentan su tasa proporcionalmente a $\left(T_i\sqrt{q_i^{(1)}}\right)^{-1}$. Por tanto, si no se controla adecuadamente el flujo en el S-GW, cuanto mayor es C, más se acerca $q_i^{(1)}$ a 0, lo que produce un aumento de carga en el sistema y en consecuencia un mayor retardo para los algoritmos de benchmark. Para ilustrar este hecho, la Figura 6.7 muestra el thrughput alcanzado en cada caso. Como vemos, los algoritmos propuestos también obtienen valores de throughput similares entre sí.

Volviendo por un momento a la Figura 6.6 vemos que el retardo de Backpressure y Switching Curve se mantiene constante, y en un valor que mejora los resultados obtenidos con fuentes de tasa constante (Figura 6.5). La conclusión es que el algoritmo Backpressure trabaja para optimizar la suma de las utilidades de las fuentes, por lo que cuando estas utilidades se orientan a minimizar el retardo, el resultado es que Backpressure alcanza el retardo mínimo. Por su parte, el

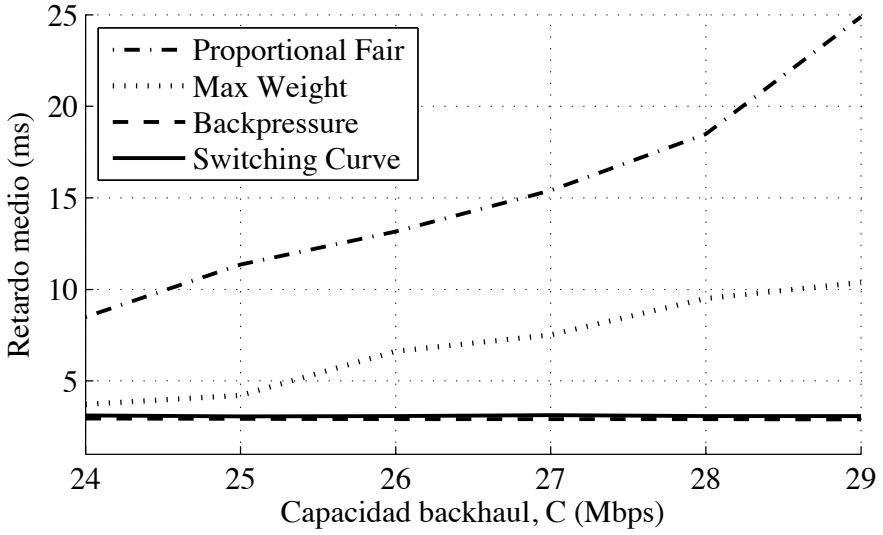

Figure 6.6: Retardo medio obtenido con cada algoritmo considerando fuentes elásticas.

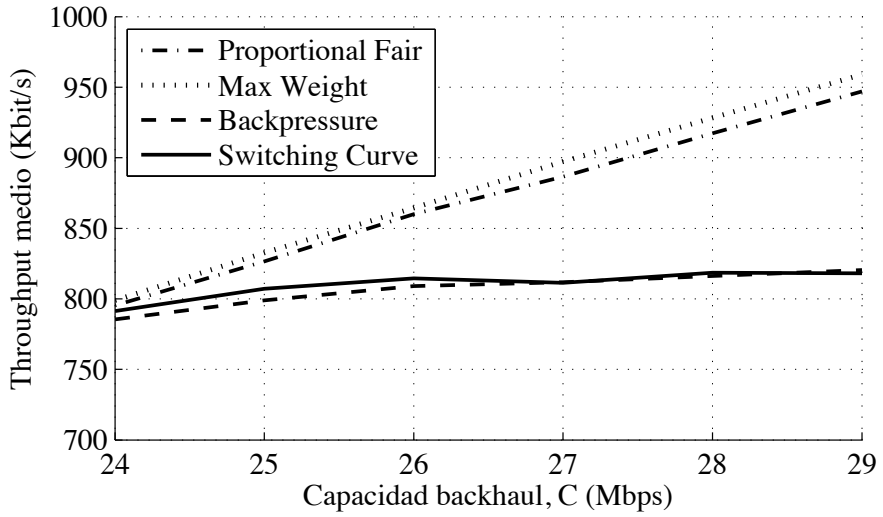

Figure 6.7: Throughput medio obtenido con cada algoritmo considerando fuentes elásticas.

algoritmo Switching Curve busca minimizar el retardo independientemente de cuál sea el objetivo de las fuentes y, por tanto, obtiene el retardo mínimo tanto para fuentes fijas como para fuentes elásticas.

6.6 Conclusiones

En este capítulo hemos abordado la asignación de recursos radio y backhaul de forma conjunta. Se ha formulado el problema NUM incluyendo también el control de tasa de las fuentes, y se ha aplicado descomposición dual para resolverlo de forma distribuida. El eNB, el S-GW y las fuentes ejecutan parte del algoritmo e intercambian la información necesaria para que el mecanismo global sea efectivo. Pero además se emplea otro nivel de descomposición del problema, al dividir los recusos de cada interfaz en bloques discretos. Para asignar cada bloque se aplica una la política basada en el cálculo de índices, lo que reduce enormemente el coste computacional y facilita que las decisiones se tomen en cada subtrama. Los índices empleados son MaxWeight en el eNB y Backpressure en el S-GW. En contraste con el Capítulo 5, ahora es el S-GW el que es consciente del interfaz radio, ya que Backpressure requiere conocer el estado de las colas de eNB. Con el objetivo de minimizar el retardo medio, proponemos unos nuevos índices para el S-GW basados en la política de control óptima para redes en tándem, que es del tipo switching curve (SC). Los resultados muestran que con los índices SC se alcanza menor retardo (e igual throughput) que con Backpressure tanto para fuentes de tasa fija como fuentes de tasa variable. Esta mejora se logra en parte porque el control óptimo de colas en tándem explota más información que Backpressure, pero conlleva el inconveniente de que hay que incluir más información en la señalización del eNB hacia el S-GW. Un problema aún abierto es la robustez de los algoritmos frente a errores o retardos en la información intercambiada entre los nodos.

Conclusiones

Mecanismos de control para sincronización de canal de transporte

El objetivo 1 fijado en este trabajo era proponer y evaluar mecanismos de control que mejoren el rendimiento de la sincronización de canal de transporte en FP. Los resultados de este objetivo los presentamos en el capítulo 3 de esta tesis. El mecanismo propuesto se encarga de adaptar el envío de tramas FP a los cambios de retardo en el Iub. Hemos mostrado cómo esta funcionalidad tiene una influencia directa en las pérdidas de tramas FP, la carga de señalización uplink en el Iub y el overbuffering en el Nodo B. También hemos estudiado la estabilidad y la reacción del algoritmo ante cambios bruscos de retardo. Para ello optamos por una metodología novedosa en este ámbito: modelar el procedimiento de sincronización como un sistema de control en tiempo discreto. El estudio del mecanismo clásico (CA) ha mostrado que se trata de un algoritmo inherentemente metaestable, que se estabiliza gracias al uso de la ventana de recepción. Para mejorar el rendimiento del timing adjustment proponemos un algoritmo sencillo (PTA) basado en un sistema de seguimiento proporcional, y analizamos el impacto del retardo en la estabilidad del sistema y en los parámetros de calidad de la respuesta a los incrementos del retardo como el incremento de la carga señalización y las pérdidas de paquetes de datos.

Mecanismos de control de flujo en backhaul

El objetivo 2 de este trabajo era proponer y evaluar mecanismos de control de flujo en el backhaul que consideren tanto el interfaz radio como el backhaul. Este objetivo lo llevamos a cabo en el capítulo 4, en el que abordamos el problema del control de flujo en HSDPA. La metodología que empleamos es novedosa con respecto a los trabajos realizados hasta la fecha, y se basa en la formulación de

un problema de optimización cuadrática que resolvemos mediante programación dinámica. Por un lado, esta aproximación dota de una base teórica a mecanismos ya existentes y por otro lado, nos permite plantear una mejora consistente en considerar la información tanto de los buffers del interfaz radio (NB) como de los del Iub (RNC) en el algoritmo de control. Esta nueva estrategia consigue minimizar el retardo extremo a extremo desde la RNC hasta el UE.

Mecanismos de scheduling radio conscientes del backhaul

El objetivo 3 de este trabajo era estudiar el impacto de la congestión del backhaul en el scheduling radio. El capítulo 5 de esta tesis contiene los resultados relativos a este objetivo, centrado en la tecnología LTE. En dicho capítulo analizamos un scheduler proportional fair, PF, incluyendo en su formulación la restricción impuesta por los límites de capacidad del backhaul. Aplicando las condiciones Karush-Kunh-Tucker obtenemos la solución, en algunos casos cerrada, del problema de optimización. Hemos llevado a cabo evaluaciones numéricas en dos escenarios realistas de LTE, empleando las especificaciones 3GPP para caracterizar la capa física (interfaz radio). Los resultados obtenidos nos muestran que cuando el backhaul es el cuello de botella, si el scheduler radio no es consciente de la capacidad del backhaul, los usuarios pueden sufrir unas pérdidas de throughput superiores al 20 %.

Mecanismos coordinados de scheduling radio y control de flujo backhaul

El objetivo 4 de este trabajo era proponer y evaluar mecanismos que asignasen de forma coordinada los recursos del interfaz radio y del backhaul. Este objetivo se ha desarrollado en el capítulo 5 de esta tesis, centrado en la tecnología LTE. La asignación de recursos del interfaz radio y del backhaul se plantea como un unico problema de maximización de la utilidad de red (NUM) en el que incorporamos también el control de tasa de las fuentes de datos. Para resolverlo de forma distribuida en el eNB, el S-GW y las fuentes, empleamos una estrategia de descomposición dual. El mecanismo resultante aplica iterativamente MaxWeight en el eNB, y Backpressure en el S-GW. Su sencilla estructura basada en índices permite tomar decisiones de asignación en cada subtrama, como estipula el 3GPP. Como

mejorar adicional, proponemos sustituir Backpressure por una política basada en el control óptimo de colas en tandem, lo que permite reducir el retardo medio tanto con fuentes de tasa fija como con fuentes elásticas (tipo TCP).

Bibliography

[1] 3GPP, "Iub/iur interface user plane protocol for DCH data streams," 3GPP TS 25.427, Tech. Rep., 2006.

[2] ——, "Synchronization in UTRAN stage 2," 3GPP TS 25.402, Tech. Rep., 2006.

[3] M. Sagfors *et al.*, "Transmission offset adaptation in UTRAN," in *Proc. IEEE VTC*, vol. 7, 2004, pp. 5240–5244.

[4] S. Abraham *et al.*, "Effect of timing adjust algorithms on iub link capacity for voice traffic in W-CDMA systems," in *Proc. IEEE VTC*, vol. 1, 2002, pp. 311–315.

[5] 3GPP, "Scenarios and requirements for small cell enhancements for E-UTRA and E-UTRAN," 3GPP TR 36.932, Tech. Rep., 2013.

[6] J. J. Alcaraz, G. Pedreno, and F. Cerdan, "A control-based approach to transport channel synchronization in UTRAN," *IEEE communications letters*, vol. 11, no. 7, p. 595, 2007.

[7] J. J. Alcaraz, G. Pedreño, F. Cerdán, J. García-Haro, and F. García-Sánchez, "Discrete-time control analysis of transport channel synchronization in 3g radio access networks," *Computer Communications*, vol. 32, no. 13, pp. 1505–1514, 2009.

[8] G. Pedreño, J. J. Alcaraz, and F. Cerdán, "Using design patterns in a HSDPA system simulator," in *Wireless Communication Systems, 2006. ISWCS'06. 3rd International Symposium on.* IEEE, 2006, pp. 679–683.

[9] J. J. Alcaraz, G. Pedreño, F. Cerdán, and J. García-Haro, "Simulation of 3G DCHs supporting TCP traffic: design, experiments and insights on parameter tuning," in *Proceedings of the 1st international conference on Simulation tools and techniques for communications, networks and systems & workshops.* ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2008, p. 75.

[10] G. Pedreño, J. J. Alcaraz, F. Cerdán, and J. García-Haro, "Improving the performance of DCH timing adjustment in 3G networks," in *NETWORKING 2008 Ad Hoc and Sensor Networks, Wireless Networks, Next Generation Internet*. Springer, 2008, pp. 768–779.

[11] G. P. López, J. J. A. Espín, and F. C. Cartagena, "Nuevo algoritmo de control de flujo para el interfaz Iub en HSDPA," *I Jornadas de introducción a la investigación de la UPCT*, no. 1, pp. 19–22, 2008.

[12] G. Pedreño López, J. J. Alcaraz Espín, and J. F. Cerdán Cartagena, "Modelado matemático del control de flujo en HSDPA," *II Jornadas de introducción a la investigación de la UPCT*, 2009.

[13] ——, "Sincronización del canal de transporte en UTRAN," *IV Teleco Forum*, pp. 79–81, 2006.

[14] H. Raza, "A brief survey of radio access network backhaul evolution: part I," *Communications Magazine, IEEE*, vol. 49, no. 6, pp. 164–171, June 2011.

[15] 3GPP, "Policy and charging control architecture, v13.5.1," 3GPP TR 23.203, Tech. Rep., 2015.

[16] ——, "Packet data convergence protocol (PDCP) specification, v6.9.0," 3GPP TS 25.323, Tech. Rep., 2007.

[17] ——, "RLC protocol specification, v6.0.0," 3GPP TS 25.322, Tech. Rep., 2006.

[18] K. Ogata, *Discrete Time Control Systems*. Prentice Hall, 1994.

[19] A. Vargas, *OMNeT++ 3.2 User Manual and API Reference*, www.omnetpp.org, 2003.

[20] 3GPP, "IP transport in UTRAN," 3GPP TS 25.933, Tech. Rep., 2003.

[21] M. C. C. Tian Bu and R. Ramjee, "Connectivity, performance, and resiliency of IP-based CDMA radio access networks," *IEEE Trans. on Mobile Computing*, vol. 5, no. 8, 2006.

[22] 3GPP, "Delay budget within the access stratum," 3GPP TS 25.853, Tech. Rep., 2004.

[23] ——, "UTRA high speed downlink packet access (HSDPA);overall description," 3GPP TS 25.308, Tech. Rep., 2006.

[24] ——, "UTRA high speed downlink packet access," 3GPP TS 25.950, Tech. Rep., 2006.

[25] ——, "HSDPA - Iub/Iur protocol aspects," 3GPP TS 25.877, Tech. Rep., 2006.

[26] M. C. Necker and A. Weber, "Impact of Iub flow control on HSDPA system performance," in *2nd International Symposium on Wireless Communication Systems*, 2005.

[27] Z. N. Szilveszter Nadas, Sandor Racz and S. Molnar, "Providing congestion control in the Iub transport network for HSDPA," in *Globecom*, 2007.

[28] J. Y. K. Xinzhi Yan and B. Jones, "An adaptive resource management technique for a HSDPA network," in *IFIP International Conference on Wireless and Optical Communications Networks*, 2007.

[29] P. M. Hong Wei, Chen Shuping and W. Wembo, "An efficient Iub flow control algorithm in HSDPA," in *International Symposium on Microwave, Antenna, Propagation and EMC Technologies for Wireless Commnication*, 2007.

[30] D. P. Bertsekas, *Dynamic Programing and Optimal Control.* Prentice Hall, 1987, vol. 1.

[31] R. W. John Gittins, Kevin Glazebrook, *Multi-armed Bandit Allocation Indices.* 2011 John Wiley and Sons, Ltd, 2011, vol. 1.

[32] C. M. A.-C. Raymond H. Meyers, Douglas C. Montgomery, *Response Surface Methodology.* 2011 John Wiley and Sons, Ltd, 2009, vol. 1.

[33] N. Whillans, "End-to-end network model for enhanced UMTS. 1st SEACORN project deliverable d3.2v2," SEACORN, Tech. Rep., 2003.

[34] M. C. Necker and A. Weber, "Protocol interference between up- and downlink channels in HSDPA," in *International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2006.

[35] T. E. Kolding, "Link and system performance aspects of proportional fair scheduling in WCDMA/HSDPA," in *Vehicular Technology Conference*, vol. 2, 2003, pp. 492–499.

[36] Somekh *et al.*, "On the impact of limited-capacity backhaul and inter-users links in cooperative multicell networks," in *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*, 2008, pp. 776–780.

[37] H. Xu and P. Ren, "Joint user scheduling and power control for cell-edge performance improvement in backhaul-constrained network MIMO," in *Personal Indoor and Mobile Radio Communications (PIMRC), 2013 IEEE 24th International Symposium on*, 2013, pp. 1342–1346.

[38] W. Yu, T. Kwon, and C. Shin, "Joint scheduling and dynamic power spectrum optimization for wireless multicell networks," in *Information Sciences and Systems (CISS), 2010 44th Annual Conference on*, 2010, pp. 1–6.

[39] H. Galeana-Zapien and R. Ferrus, "Design and evaluation of a backhaul-aware base station assignment algorithm for OFDMA-based cellular networks," *Wireless Communications, IEEE Transactions on*, vol. 9, no. 10, pp. 3226–3237, 2010.

[40] E. Bernal-Mor, V. Pla, J. Martinez-Bauset, and L. Guijarro, "Performance analysis of two-tier wireless networks with dynamic traffic, backhaul constraints and terminal mobility," *Vehicular Technology, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.

[41] 3GPP, "Evolved universal terrestrial radio access (EUTRA); further advancements for E-UTRA physical layer aspects," 3GPP TR 36.814, Tech. Rep., 2010.

[42] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *Networking, IEEE/ACM Transactions on*, vol. 8, no. 5, pp. 556–567, 2000.

[43] D. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *Selected Areas in Communications, IEEE Journal on*, vol. 24, no. 8, pp. 1439–1451, 2006.

[44] M. Chiang, S. Low, A. Calderbank, and J. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 255–312, 2007.

[45] L. V. Stephen Boyd, *Convex Optimization*. Cambridge University Press, 2004.

[46] 3GPP, "Evolved universal terrestrial radio access (E-UTRA); physical layer procedures," 3GPP TS 36.213, Tech. Rep., 2010.

[47] E. Dahlman, *4G LTE/LTE-Advanced for Mobile Broadband*. Academic Press - Elsevier, 2011.

[48] S. Meyn, *Control Techniques for Complex Networks*. Cambridge University Press, 2008.

[49] S. Srikant and L. Ying, *Communication Networks*. Cambridge University Press, 2014, vol. 1.

[50] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *Automatic Control, IEEE Transactions on*, vol. 37, no. 12, pp. 1936–1948, Dec 1992.

[51] A. Eryilmaz, R. Srikant, and J. Perkins, "Stable scheduling policies for fading wireless channels," *Networking, IEEE/ACM Transactions on*, vol. 13, no. 2, pp. 411–424, April 2005.

[52] R. Leelahakriengkrai and R. Agrawal, "Scheduling in multimedia wireless networks," *Teletraffic Science and Engineering*, vol. 4, pp. 285–298, 2001.

[53] S. Meyn, "Stability and asymptotic optimality of generalized maxweight policies," *SIAM Journal on Control and Optimization*, vol. 47, no. 6, pp. 3259–3294, 2009.

[54] M. J. Neely, E. Modiano, and C. E. Rohrs, "Dynamic power allocation and routing for time-varying wireless networks," *Selected Areas in Communications, IEEE Journal on*, vol. 23, no. 1, pp. 89–103, 2005.

[55] S. Bodas, S. Shakkottai, L. Ying, and R. Srikant, "Low-complexity scheduling algorithms for multichannel downlink wireless networks," *Networking, IEEE/ACM Transactions on*, vol. 20, no. 5, pp. 1608–1621, Oct 2012.

[56] A. Sridharan, S. Moeller, B. Krishnamachari, and M. Hsieh, "Implementing backpressure-based rate control in wireless networks," in *Information Theory and Applications Workshop, 2009.* IEEE, 2009, pp. 341–345.

[57] A. Warrier, S. Janakiraman, S. Ha, and I. Rhee, "DiffQ: Practical differential backlog congestion control for wireless networks," in *INFOCOM 2009, IEEE.* IEEE, 2009, pp. 262–270.

[58] A. Warrier, L. Le, and I. Rhee, "Cross-layer optimization made practical," in *Broadband Communications, Networks and Systems, 2007. BROADNETS 2007. Fourth International Conference on.* IEEE, 2007, pp. 733–742.

[59] Z. Rosberg, P. P. Varaiya, and J. C. Walrand, "Optimal control of service in tandem queues," *Automatic Control, IEEE Transactions on*, vol. 27, no. 3, pp. 600–610, 1982.

[60] B. Hajek, "Optimal control of two interacting service stations," *Automatic Control, IEEE Transactions on*, vol. 29, no. 6, pp. 491–499, Jun 1984.

[61] S. Stidham and R. Weber, "A survey of markov decision models for control of networks of queues," *Queueing systems*, vol. 13, no. 1, pp. 291–314, 1993.

[62] J. Razavilar, K. Liu, and S. Marcus, "Jointly optimized bit-rate/delay control policy for wireless packet networks with fading channels," *Communications, IEEE Transactions on*, vol. 50, no. 3, pp. 484–494, Mar 2002.