Universidad Politécnica de Valencia

Departamento de Comunicaciones

Tesis Doctoral

Contribución a la Gestión de Recursos en Redes de Acceso Celulares. Modelos Analíticos y Evaluación de Prestaciones

Autor: Vicent Pla Boscà
Ing. de Telecomunicación
Director: Vicente Casares Giner
Dr. Ing. de Telecomunicación

Valencia
Junio 2005

The young specialist in English Lit, (. . .) lectured me severely on the fact that in every century people have thought they understood the universe at last, and in every century they were proved to be wrong. It follows that the one thing we can say about our modern "knowledge" is that it is wrong.
(. . .)
My answer to him was, "John, when people thought the Earth was flat, they were wrong. When people thought the Earth was spherical they were wrong. But if you think that thinking the Earth is spherical is just as wrong as thinking the Earth is flat, then your view is wronger than both of them put together."

Isaac Asimov

The Relativity of Wrong

A Teresa, a mos pares i
a les meues germanes

Agradecimientos

A través de estas líneas quiero expresar mi agradecimiento a todas aquellas personas que de una forma u otra han contribuido a que hoy pueda aspirar a obtener el título de Doctor, título que de algún modo culmina, aunque no cierra, una larga etapa de formación académica.

Mencionar expresamente a todos a los que va dirigido este agradecimiento es casi imposible y correría el riesgo de cometer olvidos injustos. No obstante, no quiero dejar de nombrar a algunos de ellos cuya relación con este trabajo es más directa o está más próxima en el tiempo: a Vicente por dirigirme y apoyarme en la realización de esta tesis; a Jorge, José Ramón, Luis y Pablo, por su inestimable ayuda y apoyo en la realización de este trabajo y en los múltiples quehaceres que me han deparado estos años de relación con el mundo universitario; a David, José Manuel y María José por comprender mi falta de dedicación durante estos últimos meses a nuestros trabajos conjuntos; a Ana y Antonio por ser unos magníficos compañeros de trabajo, de asignatura y de despacho.

Quiero también hacer una mención especial del Colegio Mayor San Juan de Ribera y de la gente con la que allí tuve la suerte de convivir, su influencia durante los decisivos años de estudiante universitario ha sido sin duda determinante para que ahora pueda estar escribiendo esto.

Finalmente, en la medida en la que la realización de la tesis doctoral es el colofón de mi formación académica, ésta no sería posible sin todo aquello sobre lo que descansa. Por ello, quiero expresar mi agradecimiento a todos los docentes que a lo largo de mi vida académica han intervenido en mi formación.

Abstract

Since its commence more than twenty years ago, mobile cellular telephony has experienced an enormous growth and important transformations and, even today, it is immersed in a phase of growth and change.

Unlike fixed networks, where the task of efficient capacity management is pushed into the background by the enormous capacity of optical fiber, the scarcity of the spectrum, which is the physical medium at the radio interface in mobile networks, makes efficient resource management an important issue. Although technological progress has broaden the frequency range employed by communications and significantly increased the spectral efficiency, the traffic growth and the higher diversity of mobile services keep radio resource management (RRM) crucial.

Admission Control (AC) is an important mechanism for RRM. Despite the major boost that the study of AC techniques received within the context of the *Broadband Integrated Services Digital Network* B-ISDN, the specific features of mobile networks (spectrum limitations, stochastic behavior of the radio channel and especially terminal mobility) make the AC in such networks more complex. This work studies AC in connection with mobility: in order to admit a new session there must exist some guarantees —at least in a statistical sense— that there will be enough resources available to maintain the compromised *Quality of Service* (QoS) to all the ongoing sessions and to the new one. Such guarantee must be provided considering that during a session life-time a mobile terminal can change its location and the required resour-

ces must be available wherever it goes to. Although this topic has drawn the attention of the research community for several years it is still alive since in order to meet the growing capacity demand there is a trend to reduce the cell size. Cell size reduction increases the handover frequency and also the number of handovers that a session must undergo and these, along with the higher diversity of services in next generation networks, have a negative impact on the efficiency of resource utilization. Hence, the interest of devising efficient RRM mechanisms grows accordingly.

This thesis aims at developing analytical models for the study and performance evaluation of RRM in mobile cellular networks. In a somewhat more specific manner this work pursues to contribute to the field by developing models, methods and algorithms to: study a family of algorithms that give priority to handover requests, analyze certain aspects related to the terminal sojourn in the handover area and its repercussions on the evaluation and design of RRM and study the AC from an optimization perspective.

Resum

Des dels seus inicis fa poc més de vint anys, la telefonia mòbil cel·lular ha experimentat un enorme creixement i importants transformacions i, encara hui, s'hi troba en una fase de canvi i creixement.

A diferència del que ocorre amb una xarxa fixa, on l'enorme capacitat de transmissió que aporta la fibra òptica relega a un segon pla la gestió eficient d'eixa capacitat, el medi de transmissió a la interfície ràdio de les xarxes mòbils, l'espectre radioelèctric, és un bé escàs. Tot i que els avanços tecnològics permeten ampliar el rang de freqüències utilitzables i aconseguir una major eficiència espectral, l'augment del tràfic i l'aparició de nous usos i serveis fan que la gestió dels recursos ràdio continue sent de gran importància

El *control d' admissió* (CA) és un mecanisme important per a la gestió dels recursos ràdio. L'estudi de les tècniques de CA va rebre un impuls important dins del context de la *xarxa digital de serveis integrats de banda ampla* (RDSI-BA), però les xarxes mòbils tenen certes característiques que fan que el CA siga més complex: les ja referides limitacions de l'espectre, les característiques pròpies del canal ràdio que resulten en un comportament aleatori i, sobretot, la mobilitat dels terminals. En aquest treball s'estudia el CA en relació amb la mobilitat: per admetre una nova sessió deu haver certes garanties —almenys en un sentit estadístic— de que la xarxa tindrà recursos suficients per mantindre, amb la qualitat de servei requerida (QoS), tant eixa nova sessió com les que ja existien, i això tenint en compte que durant la vida d'una connexió, aquesta —el terminal— pot canviar d'àrea de cobertura (cèl·lula), i els re-

cursos hauran d'estar disponibles allà on vaja el terminal. Tot i que l'interés per aquest tema no és nou, continua vigent, ja que per atendre la demanda creixent de capacitat una de les solucions passa per reduir la grandària de les cèl·lules, i això implica un augment de la freqüència amb la que es produeixen els traspassos (handovers) i del nombre d'aquestos que ocorren durant una sessió. Aquest augment dels handovers i la major diversitat de serveis de les xarxes de nova generació influeixen de forma negativa en l'eficiència d'utilització del recursos, per la qual cosa també creix la importància de buscar mecanismes més eficients.

Aquest treball pretén ser una contribució al desenvolupament de models analítics per estudiar i avaluar la gestió dels recursos radio en xarxes mòbils cel·lulars. De forma quelcom més concreta, les contribucions d'aquesta tesi consistixen en el desenvolupament de models, algorismes i mètodes per a: analitzar una família d'algorismes que prioritzen les peticions de handover; analitzar diversos aspectes relacionats amb la permanència del terminal mòbil a l'àrea de handover i les seues repercussions en el disseny i avaluació de la gestió de recursos; i, estudiar el control d'admissió des d'una perspectiva basada en l'optimització.

Resumen

Desde sus inicios hace algo más de veinte años, la telefonía móvil celular ha experimentado un enorme crecimiento e importantes transformaciones y, todavía hoy, se encuentra en una fase de cambio y crecimiento.

A diferencia de lo que ocurre en una red fija, en la que la enorme capacidad de transmisión que aporta la fibra óptica relega a un segundo plano la gestión eficiente de esta capacidad, el medio de transmisión en la interfaz radio de las redes móviles es un bien escaso. Aunque los avances tecnológicos permiten ampliar el rango de frecuencias utilizables y conseguir una mayor eficiencia espectral, el aumento del tráfico junto a la aparición de nuevos usos y servicios hacen que la gestión eficiente de los recursos radio continúe siendo de gran importancia.

El *control de admisión* (CA) es un mecanismo importante para la gestión de los recursos radio. Aunque el estudio de las técnicas de CA recibió un impulso importante en el contexto de la *red digital de servicios Integrados de banda ancha* (RDSI-BA), las redes móviles tienen ciertas características específicas que hacen que el CA sea más complejo: las ya referidas limitaciones del espectro, las características propias del canal radio que resultan en un comportamiento aleatorio y, sobre todo, la movilidad de los terminales. En este trabajo se estudia el CA en relación con la movilidad: para admitir una nueva sesión se deben tener ciertas garantías —al menos en un sentido estadístico— de que la red tendrá recursos suficientes para mantener, con la calidad de servicio (QoS) requerida, tanto esa nueva sesión como las ya existentes en ese

momento, y ello teniendo en cuenta que durante la vida de una sesión, ésta —el terminal— puede cambiar de área de cobertura (célula), y los recursos deberán estar disponibles allí donde vaya el terminal. Aunque el interés por este tema no es nuevo, continúa vigente, pues para atender la creciente demanda de capacidad, una de las formas pasa por reducir el tamaño de las células, y esto implica un aumento de la frecuencia con la que se producen los traspasos (handovers) y del número de handovers que se producen durante una sesión. Este aumento de los handovers y la mayor diversidad de servicios de las redes de nueva generación influyen de forma negativa en la eficiencia de la utilización de los recursos, por lo que crece también la importancia de buscar mecanismos más eficientes.

Este trabajo pretende ser una contribución al desarrollo de modelos analíticos para el estudio y la evaluación de la gestión de los recursos radio en redes móviles celulares. De forma algo más concreta, las contribuciones de esta tesis consisten en el desarrollo de modelos, algoritmos y métodos para: analizar una familia de algoritmos que priorizan las peticiones de handover; analizar diversos aspectos relacionados con la permanencia del terminal móvil en el área de handover y sus repercusión en el diseño y evaluación de la gestión de recursos; y estudiar el control de admisión desde una perspectiva basada en la optimización.

Índice general

Índice de figuras

Indice de Tablas

Capítulo 1

Introducción

A diferencia de lo que ocurre en las redes fijas donde la enorme capacidad de transmisión que aporta la fibra óptica relega a un segundo plano la gestión eficiente de esta capacidad, el medio de transmisión en la interfaz radio de las redes móviles es el espectro radioeléctrico, un bien escaso. Aunque los avances tecnológicos en este campo permiten ampliar el rango de frecuencias utilizables y conseguir una mayor eficiencia espectral, el enorme crecimiento del número de usuarios así como la introducción de nuevos servicios, que además son más exigentes por lo que respecta al consumo de capacidad, hacen que la gestión eficiente de los recursos radio continúe siendo de gran importancia.

El *control de admisión* (CA) es un mecanismo importante para la gestión de los recursos radio. Aunque el estudio de las técnicas de CA recibió un impulso importante en el contexto de la *red digital de servicios Integrados de banda ancha* (RDSI-BA), las redes móviles tienen ciertas características específicas que hacen que el CA sea más complejo: las ya referidas limitaciones del espectro, las características propias del canal radio que resultan en un comportamiento aleatorio y, sobre todo, la movilidad de los terminales. En este trabajo se estudia el CA en relación con la movilidad: para admitir una nueva sesión se deben tener ciertas garantías —al menos en un sentido estadístico—

de que la red tendrá recursos suficientes para mantener, con la calidad de servicio (QoS) requerida, tanto esa nueva sesión como las ya existentes en ese momento, y ello teniendo en cuenta que durante la vida de una sesión, ésta —el terminal— puede cambiar de célula, y los recursos deberán estar disponibles allí donde vaya el terminal. Aunque el interés por este tema cuenta con más de dos décadas de historia, éste continúa vigente a juzgar por el número de trabajos que todavía encontramos en la literatura especializada, lo que a nuestro juicio podría explicarse, al menos en parte, por la siguiente razón: para atender la creciente demanda de capacidad una de las formas pasa por reducir el tamaño de las células y esto implica un aumento de la frecuencia con la que se producen los handovers[1] y del número de handovers que se producen durante una sesión. Este aumento de los handovers y la mayor diversidad de servicios de las redes de nueva generación influyen de forma negativa en la eficiencia de la utilización de los recursos [VC00], por lo que crece también la importancia de buscar mecanismos más eficientes.

Este trabajo pretende ser una contribución al desarrollo de modelos analíticos para el estudio y la evaluación de la gestión de recursos radio en redes móviles celulares. Aun siendo conscientes de que las limitaciones de los modelos analíticos impiden capturar todos los detalles de sistemas de gran complejidad, creemos que es igualmente importante disponer de modelos analíticos cuya aplicación complemente a la de los modelos de simulación, los prototipos y las medidas sobre un sistema real. En general, los modelos analíticos son más adecuados para descubrir tendencias generales y proporcionar resultados cualitativos que facilitan una mejor comprensión del funcionamiento del sistema. Además, este conocimiento es útil como guía para diseñar los experimentos adecuados a realizar mediante simulación o la construcción de un prototipo, experimentos que en general son más costosos en tiempo y recursos.

[1] Aunque el término *handover* no está recogido en el Diccionario de la lengua española y el significado del término *traspaso (traslado de algo desde un lugar a otro; cesión a favor de otra persona del dominio de algo)* [Esp03] describiría bastante bien lo que es un handover, en este trabajo se ha preferido utilizar el anglicismo —y su plural— por ser más específico y por estar bastante extendido, dentro del ámbito del lenguaje técnico, entre los hispanohablantes.

Estructura de la tesis

En la última parte de este capítulo de introducción se hace una rápido recorrido a través de la evolución de las comunicaciones móviles. En el capítulo 2 se realiza una revisión no exhaustiva del estado del arte en el campo del CA en redes celulares desde una perspectiva, fundamentalmente, analítica. El capítulo 3 propone y analiza de forma unificada una familia de algoritmos para asignar recursos en redes celulares otorgando prioridad al tráfico de peticiones de handover. Además de un método de análisis del tipo geométrico-matricial se desarrolla también un análisis basado en técnicas espectrales matriciales. En el capítulo 4 se estudian diversos aspectos relacionados con la permanencia del terminal móvil en el área de handover y sus repercusiones en los modelos de evaluación de la gestión de recursos. Se desarrolla un método analítico-numérico para caracterizar estadísticamente el tiempo de residencia en el área de handover y de ocupación de los recursos mientras el móvil está en esta zona. Asimismo se evalúa la sensibilidad de los parámetros de medida de prestaciones frente a la variabilidad del tiempo de residencia en el área de handover. Por último, se desarrolla un modelo de colas que permite evaluar la aplicación de distintas disciplinas de servicio (FIFO, LIFO o SIRO), y de gestión del espacio de almacenamiento, a las peticiones de handover que por no poder ser atendidas inmediatamente esperan mientras el terminal permanece en el área de handover. El capítulo 5 aborda el diseño del control de admisión en redes celulares como un problema de optimización, proponiendo distintos criterios de optimización para redes multiservicio que dan lugar a la formulación de un problema de programación lineal. También se desarrolla un algoritmo para optimizar la política de control de admisión dentro de un conjunto más restringido, el de las políticas del tipo *trunk reservation* y se evalúa su eficiencia computacional. Por último, se estudia la ganancia que puede obtenerse dotando al proceso de optimización con una predicción sobre la posible llegada de peticiones de handover. Por último, el capítulo 6 presenta un resumen del trabajo realizado y sus conclusiones, y se sugieren de manera genérica posibles líneas de

trabajo futuras.

Evolución histórica de las comunicaciones móviles

Podemos situar el inicio de las radiocomunicaciones en la década de 1860 con la publicación de los trabajos de James Clerck Maxwell: las célebres ecuaciones de Maxwell que gobiernan las radiaciones electromagnéticas. En 1888 Heinrich Hertz demuestra en el laboratorio la generación y detección de ondas radio, pero no es hasta la década siguiente cuando Nikola Tesla, y posteriormente Aleksandr Stepanovich Popov y Gugliemlmo Marconi, utilizan la capacidad de las radiocomunicaciones para transmitir mensajes [Soc02]. En 1916 ingenieros de Bell System realizan el primer ensayo de radiotelefonía entre barcos y en 1921, en la comisaría de policía de Detroit (EEUU), se instala el primer sistema móvil de envío de mensajes (*dispatch*) [ZTH+00]. En 1947, D.H. Ring, de Bell Labs, propone el primer sistema celular [ZTH+00] y en 1979 la operadora japonesa NTT despliega en la ciudad de Tokio el primer sistema de telefonía celular del mundo [Wes02]. Durante los años siguientes los países más tecnificados introducen distintos sistemas celulares analógicos e incompatibles entre sí —la primera generación de sistema de telefonía celular—. En el ámbito europeo se crea en 1982 el *Groupe Spécial Mobile* (GSM) con el propósito de desarrollar un sistema paneuropeo de telefonía celular, que en 1989 dará lugar a la publicación del estándar GSM por el *European Telecommunications Standars Institute* (ETSI). El lanzamiento comercial de GSM se produce entre 1991 y 1992 [VLLX02]. GSM es el principal representante de los sistemas de telefonía celular denominados de segunda generación (2G), cuya diferencia principal con los sistemas de primera generación consiste en la utilización de tecnología digital. Como indicación del éxito y del enorme crecimiento de GSM en particular, y de la telefonía móvil digital en general, sirva que a finales del año 2004 el número total de usuarios de telefonía móvil digital estaba muy próximo a los 2000 millones, y de los cuales más del 75 % utilizan GSM. Actualmente existen más de 600 redes que utilizan GSM a lo largo de más de 200 países [GSM]. En una fase posterior se han intro-

ducido actualizaciones en la interfaz radio de GSM para aumentar el caudal del tráfico de datos, bien manteniendo la conmutación de circuitos —*High-Speed Circuit-Swithced Data* (HSCSD)—, o bien introduciendo la conmutación de paquetes —*General Packet Radio Service* (GPRS)— [Wes02]. La introducción de nuevos esquemas en la capa física —*Enhanced Data rates for Global Evolution* (EDGE)— ha aumentado todavía más las tasas binarias que pueden lograrse con HSCSD y GPRS [VLLX02, Wes02]. Actualmente nos encontramos en el inicio del despliegue de las redes de tercera generación 3G cuyo proceso de estandarización se llevo a cabo durante la década de los noventa [MK00, ZAB00]. Las principales características que aportan las redes 3G son el aumento de la capacidad, debido principalmente a la utilización de una interfaz radio basada en *Wideband Code Division Multiple Access* (W-CDMA), estar concebidas desde un principio para cursar tráfico de datos mediante la conmutación de paquetes —sin descartar el tráfico de tiempo real ni la conmutación de circuitos— y la movilidad (itinierancia) global [Wes02, VLLX02]. El gran éxito de Internet y la aparición de nuevas tecnologías de acceso inalámbrico, fundamentalmente redes de área local inalámbricas (WLAN), ha hecho que se hable ya de la cuarta generación (4G) de redes móviles: un conjunto heterogéneo de redes de acceso inalámbricas que ofrecerían un acceso global e integrado a Internet mediante la utilización de los protocolos IP, aunando así la ubicuidad del acceso a Internet con una alta capacidad en emplazamientos estratégicos [Sal04] y de forma que el usuario final acceda a la red en cada momento y en cada lugar de la forma más adecuada (*Always Best Connected*) [GJ03].

Capítulo 2

La gestión de recursos en redes celulares y su estudio analítico

En este capítulo se revisan, de una forma no exhaustiva, los estudios previos que configuran el estado del arte en el campo de la gestión de recursos en redes celulares desde una perspectiva, fundamentalmente, analítica. De una parte se examinan los distintos modelos utilizados en los estudios y sus hipótesis, y de otra, los diferentes esquemas que se han propuesto para realizar el CA. Como ya hemos comentado, la lista de referencias que aparecen en esta revisión no es exhaustiva —algo que sería prácticamente imposible dada la ingente cantidad de ellas—. Por otra parte, su estructuración no es desde luego la única, ni mucho menos la mejor, sino que se ha realizado atendiendo a aquellos aspectos que se han abordado en esta tesis.

2.1. Modelo del sistema

Las características principales del sistema objeto de estudio, y que por tanto constituyen las características básicas de cualquier modelo del mismo, son las siguientes:

1. La zona de servicio de la red está divida en células y en cada una de ellas existe una estación base con la que se conectan los terminales móviles vía radio.

2. Los terminales presentan periodos de actividad que denominaremos sesiones y a lo largo de uno de estos periodos se ocupan una cierta cantidad de recursos de la interfaz radio. En redes de conmutación de circuitos una sesión se corresponderá con una conexión y los recursos empleados se refieren a la cantidad fija de éstos que se ocupan de forma permanente, en cambio, en redes de conmutación de paquetes la cantidad de recursos consumidos por una sesión debemos entenderlo en un sentido estadístico que estaría en la línea de, por ejemplo, el ancho de banda efectivo [EE99].

3. Los terminales son móviles por lo que durante el transcurso de una sesión el terminal puede cambiar de célula.

4. Con el fin de asegurar la continuidad de la comunicación, existe un solape entre las zonas de cobertura de células vecinas.

5. Cada célula dispone de una cantidad fija de recursos radio. En algunos sistemas celulares también es posible emplear una asignación dinámica (*Dynamic Channel Allocation*, DCA) y existe un buen número de estrategias [TJ91, KN96] que utilizan la idea de asignación dinámica para conseguir una gestión más eficiente de los recursos, aunque, naturalmente, la complejidad asociada a este tipo de estrategias también es significativamente mayor.

2.1.1. Modelos de pérdidas y de espera, abandonos y reintentos

En el estudio del tipo de sistema que nos ocupa surgen diferentes modelos de colas. En general, cuando llega una petición al sistema (la célula), el control de admisión debe decidir si esta petición se acepta o se rechaza. Para

las peticiones que son bloqueadas, es decir, las que no se aceptan inmediatamente tras su llegada, existen distintos tratamientos posibles: pueden esperar a que se liberen los recursos necesarios para poder ser admitidas, o bien ser rechazadas. En este último caso —no hay cola de espera—, puede producirse un reintento transcurrido un tiempo o bien desistir de forma definitiva. En caso de que las peticiones bloqueadas esperen en una cola a que se liberen los recursos necesarios, puede darse la situación en la que dicha espera se prolongue hasta que finalmente se acepta la petición, en cuyo caso diremos que el cliente (la petición) tiene una paciencia infinita, o bien abandonar la cola si la petición no se admite antes de un determinado plazo, en este caso se dice que el cliente es impaciente.

Todos los escenarios anteriores y combinaciones de ellos han sido estudiados en la literatura que trata el control de admisión en sistemas celulares. El uso de una cola para los handovers (peticiones de handover) bloqueados es un mecanismo de priorización de los handovers ampliamente propuesto (véase [PG85, HR86, TJ91, TJ92, KN96, TRV98, PCG02b, LJCP03, XT04] y sus referencias). Cuando se emplea esta técnica el handover puede permanecer en la cola mientras el móvil esté en la zona en la que recibe suficiente potencia de ambas células (la de origen y la de destino), pero, si el móvil deja de recibir suficiente potencia de la célula de origen antes de que se liberen los recursos necesarios en la célula de destino, se producirá una terminación forzosa de la sesión en curso. Aunque la mayoría de los trabajos que consideran la existencia de una cola para los handovers asumen que esta cola se sirve según una disciplina FIFO, existen también algunos ejemplos en los que se considera una disciplina de servicio distinta que tiene en cuenta el fenómeno de la impaciencia, de forma que trata de servir primero a aquellos handovers que previsiblemente abandonarán antes la zona de solape [TJ92, ET99, DRFG99a, DRFG99b, Fan00, LJCP03, XT04]. La utilización de una cola para las peticiones de establecimiento de nuevas sesiones también se ha propuesto con el fin de incrementar el tráfico cursado total aunque, en este caso el fenómeno de la impaciencia no siempre se incluye en el modelo [Gué88, DJ92] ya que la tolerancia al tiempo de espera de este tipo de

peticiones es mucho menor. Algunos estudios contemplan la posibilidad de espera para ambos tipos de peticiones, por ejemplo [CSC94, McM95, LJCP03]. Los modelos de sistemas que integran tráfico de voz (con requisitos de tiempo real) y de datos (sin requisitos de tiempo real) [ZA02, WZA03] utilizan colas distintas para los dos tipos de servicio y, en algunos casos, se permite que un handover de una sesión de voz desaloje a una sesión de datos o que tome parte del ancho de banda utilizado por el tráfico de datos [SW04].

En [TGM97, AL02] se modela el fenómeno del reintento de las peticiones de sesiones nuevas bloqueadas. Tran-Gia y Mandjes [TGM97] modelan el tiempo entre dos reintentos consecutivos mediante una variable aleatoria exponencial mientras que Alfa y Li [AL02] consideran una distribución más general: del tipo *phase type* (PH). Ajmone Marsan et al. [MMDCL+01] estudian un sistema con reintentos tanto de la peticiones de sesiones nuevas como para las de handover y, dada la complejidad del modelo, proponen una aproximación para su análisis en la que la descripción del estado del modelo no incluye el número de peticiones en la órbita (peticiones que han sido bloqueadas y van a reintentar) sino sólo si la órbita está vacía o no.

2.1.2. Superposición de micro y macrocélulas

La superposición de micro y macrocélulas para formar una estructura jerárquica es otra idea en la que se basan un grupo de estrategias encaminadas también a conseguir una gestión más eficiente de los recursos radio. La idea básica consiste en utilizar las microcélulas para dar cobertura a áreas con una alta intensidad de tráfico y las macrocélulas cubrirían las áreas con una intensidad de tráfico menor y el tráfico de desbordamiento de la áreas con alta intensidad de tráfico [Rap94]. Otra ventaja de este esquema, que está más relacionada con la gestión de recursos a la escala temporal en la que opera el CA, es que la tasa de handovers puede reducirse si a los terminales que se mueven con mayor velocidad sólo se les asigna recursos de las macrocélulas [MM00].

Los modelos analíticos que se han utilizado para evaluar esta clase de sistemas son de tipo markoviano. En [Rap94] se utiliza un proceso de nacimiento y muerte multidimensional, cuyo análisis resulta ser bastante complejo. Debido a esta complejidad otros autores han utilizado distintas aproximaciones como, por ejemplo, modelar el flujo de peticiones de desbordamiento como un proceso de Poisson modulado con un proceso de Markov (MMPP), o simplemente con un proceso de Poisson si tanto el número de microcélulas como la intensidad de tráfico son altos [MM00]. En [MM00] los autores consideran también distintos tipos de servicio.

2.1.3. Suposiciones e hipótesis del modelo

Célula aisladas frente a grupo de células

En el estudio del CA en sistemas celulares es bastante común considerar que: el tráfico es homogéneo a lo largo de toda el área de servicio, es decir, que las tasas de llegada de sesiones nuevas y de handovers son idénticas en todas las células; las células están todas en equilibrio estadístico y son independientes entre sí. Con estas suposiciones las prestaciones del sistema pueden estudiarse considerando únicamente una célula aislada. Sin embargo, estas suposiciones pueden no ser apropiadas en determinados escenarios, en cuyo caso un modelo multicélula sería más adecuado [BBP01]. Por otra parte, considerar un modelo u otro también repercute en el proceso utilizado para describir el tráfico de handovers, pero este asunto se trata más adelante (página 13). Otra ventaja de los modelos multicélula es que permiten incorporar patrones de movilidad basados, por ejemplo, en el trazado de las calles o carreteras. Por otro lado la principal desventaja de emplear un modelo multicélula es la complejidad computacional. Mantener una descripción detallada y conjunta del estado de cada célula del modelo se convierte en intratable para un número de células no muy alto. Una solución intermedia entre un modelo multicélula completo y otro de células homogéneas independientes, consiste en considerar un escenario de células heterogéneas (en cuanto a car-

ga y cantidad de recursos) en el que el proceso de llegada de handovers a una célula es un proceso de Poisson que no depende en cada momento del estado de las células circundantes, sino del flujo medio de handovers salientes de dichas células. Para determinar las tasas de handover se utiliza método iterativo de punto fijo [McM91, BBP01, MS01] aunque su precisión es cuestionable cuando la movilidad es alta [MS01]. Determinar los parámetros que optimizan la política de CA de forma global en un escenario multicélula también reviste una complejidad importante. En [BBP01] se aplica una búsqueda Tabu para un escenario con un único servicio y en [MS01] se estudia un escenario multiservicio y, para reducir la complejidad, únicamente se consideran políticas que producen una solución en forma de producto.

Teletráfico y variables aleatorias

En los modelos de redes celulares se utilizan variables aleatorias para describir las magnitudes siguientes: procesos de llegada (tiempo entre peticiones), duración de una sesión, duración residual de una sesión después de un handover, tiempo de residencia en una célula, tiempo de residencia en la zona de solape entre células, tiempo de ocupación de los recursos de una célula por una sesión, ... La tendencia general en la mayoría de estudios es suponer que todas estas variables aleatorias siguen una distribución exponencial y, la razón principal para hacer esta suposición es mantener la tratabilidad del modelo, sobre todo cuando se trata de un modelo analítico. Los estudios que encontramos en la literatura que persiguen validar esta hipótesis arrojan resultados en ocasiones contradictorios (véase [RT01] y sus referencias).

La utilización de un proceso de Poisson (tiempo entre llegadas exponencial e independiente entre sí) para el flujo de llegada de peticiones de nuevas sesiones es tal vez la hipótesis menos cuestionada. No obstante, el trabajo de Barceló y Bueno [BB97], basado en medidas de campo de un sistema móvil, aunque no celular, concluye que la suposición subyacente de población infinita puede estar bastante lejos de ser adecuada. Por otra parte, la medidas presentadas en [BB97] revelan que el tiempo entre intentos de llamada sigue

una distribución que es más suave que la exponencial, esto es, su coeficiente de variación es inferior a la unidad ($CV < 1$). [1] En [BS99] se aplica un método similar a un sistema celular y se observa que el proceso agregado de llegadas (sesiones nuevas más peticiones de handover) es también más suave que uno de Poisson.

En [AL02], Alfa y Li desarrollan un modelo analítico en el que no se supone que el proceso de llegadas sea de Poisson, sino que toman algo mucho más general y versátil como el MAP (*Markovian Arrival process*), que incluye como un caso particular al MMPP. Sin embargo, la complejidad computacional para obtener resultados numéricos de este modelo es extremadamente alta y de hecho los autores no proporcionan ningún resultado numérico.

La suposición de un proceso poissoniano para el proceso de llegada de las peticiones de handover ha sido más estudiada aunque, las conclusiones respecto a su validez no son unánimes. Por otra parte, una buena parte de los trabajos que analizan el proceso de llegadas de handovers —en general todos los que no utilizan medidas de campo— se basan en la suposición de que la llegada de peticiones de nuevas sesiones sigue un proceso de Poisson y en un cierto modelo de movilidad, por lo que la validez de las conclusiones estará sujeta, al menos, a la validez de estos dos modelos: el del proceso de llegadas de peticiones nuevas y el de movilidad. Chlebus y Ludwin [CL95] han sido probablemente los primeros en cuestionar la suposición de tráfico de handover poissoniano, en su artículo estos autores demuestran que si las sesiones nuevas llegan según un proceso de Poisson, el bloqueo que éstas sufren hace que el tráfico de handovers no sea de Poisson y, utilizando una caracterización de dos momentos del tiempo entre llegadas, concluyen que el tráfico de handover es suave. Sin embargo, la comparación de los resultados analíticos —suponiendo que el trafico de handover es de Poisson— con los resultados obtenidos mediante simulación —donde no se hace esa suposición— demuestra que la aproximación de tráfico de Poisson es razonablemente buena, en especial para una carga del sistema de baja a moderada.

[1] El coeficiente de variación de una variable aleatoria X se define como $CV_X = \sigma_X / E[X]$.

Sidi y Starobinski [SS97], concluyen que la suposición de tráfico de Poisson es válida cuando el tráfico es homogéneo a lo largo de un número elevado de células mientras que no lo es si se considera una red con pocas células o con tráfico heterogéneo. Rajaratnam y Takawira [RT00, RT01] demuestran también la condición suave del tráfico de handover y que la aproximación poissoniana para este tráfico podría no ser válida cuando durante el transcurso de una sesión el terminal atraviesa un número alto de células, esto es, cuando la movilidad es alta. Mediante un análisis aproximado que utiliza los dos primeros momentos, demuestran que la aproximación poissoniana sobreestima la probabilidad de bloqueo de los handovers. En [vDT03], Doorn y Ta deducen formalmente los principales resultados que Rajaratnam y Takawira habían obtenido mediante un estudio numérico o empírico. Orlik y Rappaport [OR01] abordan este asunto comparando tres escenarios distintos: 1) un grupo de siete células (una célula central, que es la célula en estudio, y las seis del anillo circundante) al que la llegada de handovers desde la periferia del grupo sigue un proceso de Poisson; 2) una célula aislada a la que los handovers llegan según MMPP de dos estados ajustado a partir del escenario anterior; 3) una célula central a la que llegan handovers según un proceso de Poisson. A diferencia de otros trabajos, en este caso los resultados demuestran que no existen diferencias significativas entre los tres escenarios, principalmente cuando la carga es alta, lo que contrasta con las conclusiones de [CL95]. El efecto que la distribución del tiempo de residencia en una célula tiene en la distribución del tiempo entre llegadas de handover se trata en [ZC99] considerando tanto el escenario sin bloqueo como el escenario en que sí hay bloqueo. La conclusión principal es que si el tiempo de residencia en una célula sigue una distribución Gama —suposición que está avalada por diferentes estudios como se verá mas adelante— con una varianza alta, el tráfico de handover no se puede caracterizar mediante un proceso de Poisson. Motivados por los trabajos antes citados, en [DTL03] los autores estudian un sistema en el que el tiempo entre llegadas de los handovers se modela mediante una distribución de Erlang o hiperexponencial. Los parámetros de interés para las prestaciones del sistema se comparan con

los obtenidos cuando el tráfico de handover es de Poisson, manteniendo en ambos casos la misma tasa de llegadas. Si el tiempo entre llegadas sigue una distribución erlanguiana no hay diferencias importantes respecto al caso en el que la distribución es exponencial, mientras que si se considera la distribución hiperexponencial sí las hay. No obstante, para los valores concretos de los parámetros de la distribución hiperexponencial que se toman en este artículo, el tráfico de handover no es suave sino más bien lo contrario (el coeficiente de variación es $CV = 1.22 > 1$), por lo que según los resultados de este artículo, interpretados a la luz de los resultado anteriores que afirman que el tráfico de handover es suave, no descartan sino que más bien reforzarían la hipótesis de que describir el tráfico de handover con un proceso de Poisson es una buena aproximación.

La caracterización estadística del tiempo de residencia en una célula (*cell residence time*, CRT) y del tiempo que una sesión ocupa los recursos de una célula (*channel holding time*, CHT) ha sido abordado en un buen número de estudios. En los trabajos de Hong y Rappaport [HR86] y de Guérin [Gué87] se concluye que para células con una geometría circular el CHT puede aproximarse satisfactoriamente mediante una distribución exponencial. En [Sch03] Schweigel obtiene la distribución de CRT para células de geometría rectangular y evalúa el error de ajustar esta distribución mediante una distribución exponencial. En [ZD97], se emplea un modelo de simulación que es más general que el de los trabajos anteriores y se obtiene que el CRT se ajusta una distribución Gama generalizada y, si se supone que la duración de la sesión está distribuida exponencialmente, el CHT podría aproximarse mediante una distribución exponencial. En sendos estudios basados en medidas [JL96, BJ00], Jedrzycki y Leung [JL96], y Barceló y Jordán [BJ00] proponen utilizar una distribución log-normal o una mezcla (combinación convexa) de éstas para describir el CHT. En una serie de estudios empíricos llevados a cabo por Hidaka et al. (véase [HSSK01, HSSK02] y sus referencias) se ha descubierto que el CRT y el CHT presentan características de *autosemejanza*. Distintos autores han estudiado cuál es el error que introduce en los parámetros de interés la suposición de que el CHT tiene una distribución exponencial si se

compara con los resultados obtenidos utilizando las distribuciones que sugieren los trabajos previos. En [KZ97, XT03] se alcanza la conclusión de que la aproximación puede considerarse buena, mientras que en [LC97, HSSK02] se presentan algunas situaciones en las que aparecen divergencias importantes.

El tiempo de residencia en la zona de solape entre células ha sido mucho menos estudiado y la conclusión general es que no se ajusta a una distribución exponencial [RGS98, PCG02a]. No obstante, los resultados de [PCG02b] sugieren que considerar una distribución exponencial es en muchos casos una buena aproximación y, cuando no, una opción conservadora.

Recientemente, Machihara [Mac05] ha publicado una generalización de la fórmula de Erlang a un escenario celular. Según este resultado, de indudable interés teórico, el sistema presenta la interesante propiedad de ser insensible respecto a la distribución de la duración de una sesión y del CRT. Sin embargo, dicha insensibilidad se ha obtenido cuando las sesiones nuevas y los handovers son tratados del mismo modo, es decir, no se da ningún tipo de prioridad a las peticiones de handover.

2.2. Políticas de CA en redes celulares

De una forma genérica podemos definir una política de *control de admisión de la conexión* (CA) como el criterio aplicado para decidir sobre la admisión o no de una petición de conexión.[2] En esta sección se intenta dar una perspectiva general de las propuestas en el campo de las políticas para el CA en redes celulares. Esta revisión de la literatura se organiza contemplando o agrupando las políticas de CA y su diseño desde tres puntos de vista distintos: según el tipo de información en el que se basa la decisión de admisión, según la forma general de la política, y según el método empleado para ajustar los parámetros de una de estas formas generales de política.

[2] En nuestro caso de una petición de una nueva sesión o de una petición de handover.

2.2.1. Información de estado versus información predictiva

La mayoría de las políticas de CA propuestas basan la decisión de admitir la petición en el estado de la célula sobre la que opera. Esta información de estado local, en general, consiste en: el número de sesiones en curso en dicha célula —bien de forma agregada o detallada por tipos de usuario/servicio— o la cantidad total de recursos ocupados —de forma agregada o detallada—. Existe otro grupo de propuestas, aunque bastante menos numeroso, que trata de aprovechar el hecho de que en las redes celulares es posible disponer con antelación de cierta información sobre las peticiones de handover antes de que éstas se produzcan. En estos casos se utiliza información sobre el estado de las células vecinas, la velocidad y trayectoria de los móviles, el trazado de calles o carreteras, patrones de movilidad,... para predecir la llegada de peticiones de handover. En [NS96] y [WWL02] se utiliza la información de estado (número de llamadas en curso) de la célula objetivo (sobre la que se aplica el CA) y de sus vecinas, para predecir la probabilidad de bloqueo en una ventana temporal futura (de duración $T = 20$ s). Esta predicción se realiza mediante una aproximación gaussiana en [NS96] y resolviendo las ecuaciones del régimen transitorio en [WWL02]. En un mayor número de artículos se proponen métodos para predecir, a partir de información de movimiento actual e histórica, hacia qué célula se desplazará un determinado móvil que tiene una sesión activa y cuándo ocurrirá esto [LAN97, CB00, YL02, CS02, SK04]. En estos estudios, al método de predicción de handovers propuesto se le asocia una política de CA basada en consideraciones más o menos heurísticas o intuitivas que mejoran las prestaciones respecto al caso en que no se dispone de la información predictiva. Frente a esta utilización heurística de la información predictiva, en [YR97] y [PGGMCG04] se utiliza un enfoque basado en optimización.

2.2.2. Familias de políticas de CA

Políticas de reserva de recursos

Realmente, bajo el título de *reserva de recursos* se podría incluir a cualquier tipo de política de CA pues en el fondo, cuando se rechaza una petición habiendo recursos suficientes para atenderla lo que se está haciendo es reservar estos recursos para otras peticiones de mayor valor o prioridad. No obstante, tradicionalmente se ha empleado este nombre —u otros semejantes como *trunk reservation*, *guard channel* (GC) o *reserva de canales*— para aquellas políticas en las que la reserva aparece de forma explícita.

Por otra parte, este tipo de políticas son el método más utilizado de cuantos encontramos en la literatura para dar prioridad a los handovers. La idea básica en esta familia de políticas consiste en reservar una cierta cantidad de recursos para los handovers y admitir peticiones de sesiones nuevas únicamente si los recursos disponibles en ese momento superan la cantidad reservada.

La aplicación de esta técnica al CA en redes celulares fue introducida a mediados de los ochenta [PG85, HR86] y desde entonces han aparecido un gran número de variantes, generalizaciones, extensiones o mejoras. La política *Fractional Guard Channel* (FGC) [RNT97] introduce la posibilidad de reservar una cantidad no necesariamente entera de la unidad básica de recursos, mediante decisiones de admisión probabilísticas. La ventaja del carácter fraccionario es que posibilita un ajuste más fino de los parámetros de la política. Para un escenario de un único servicio, podemos encontrar en [McM95, Sch90, CH96, CG01, KA99, PC01, DS02, DS04] distintas extensiones de las políticas GC y FGC. En [CC97, LLC98] encontramos una generalización de la política GC a un escenario multiservicio y, del mismo modo, en [HUCPOG03a] se generaliza la política FGC.

La cantidad de tráfico que puede cursarse mediante las políticas tratadas en este punto es mayor que la que puede conseguirse aplicando una política de las que tienen solución en forma de producto (véase el punto si-

guiente, 2.2.2). Sin embargo, frente a esta superioridad de las políticas del tipo *trunk reservation* cabe hacer dos observaciones: la capacidad superior se obtiene para el valor nominal de los parámetros de teletráfico para los que se optimizan los parámetros de la política; y segunda, la complejidad computacional del análisis es bastante superior y para sistemas grandes el diseño o ajuste de la política podría ser intratable. Para tratar el problema de la complejidad computacional se han desarrollado métodos aproximados de análisis que reducen significativamente el esfuerzo computacional [BS97, BM98, CPVAOG04]. Por otra parte, la política virtual partitioning (VP) [BM98, YMW+04] puede verse como una generalización de la política *trunk reservation* que se adapta automáticamente cuando la carga de un determinado servicio supera la prevista y, de este modo, consigue un buen equilibrio entre eficiencia y equidad.

Políticas con solución en forma de producto

Este tipo de políticas se caracterizan porque en un escenario multiservicio, cuando se utiliza un proceso de Markov para modelar sistema, las probabilidades de estado de este proceso pueden calcularse como el producto de las distribuciones marginales para cada servicio, multiplicadas por una constante de normalización. La familia de políticas de CA denominada *convexa coordinada* [Ros95] está íntimamente ligada con la familia de políticas cuya solución tiene forma de producto. Todas las políticas que son convexas coordinadas tienen solución en forma de producto pero existen algunas políticas cuya solución tiene forma de producto que no encajarían en la definición de política convexa coordinada. Una política convexa coordinada limita los estados posibles del sistema a un conjunto convexo coordinado[3] y acepta una petición si y sólo si después de aceptar la petición el sistema permanece dentro del conjunto permitido de estados. Por tanto, la petición se acepta o se

[3]Sea $S := \{(x_1, \ldots, x_N) : x_i \in \mathbb{N},\ i = 1, \ldots, N\}$ se dice que $\Omega \subseteq S$ ($\Omega \neq \emptyset$) es un conjunto convexo coordinado si cumple la propiedad siguiente: si $(x_1, \ldots, x_i, \ldots, x_N) \in \Omega$ y $x_i > 0$, entonces $(x_1, \ldots, x_i - 1, \ldots, x_N) \in \Omega$ [Ros95].

rechaza dependiendo del estado que resultaría de aceptar la petición pero sin importar el estado de procedencia. Esto último no se cumple para las políticas del tipo *trunk reservation*, en las que sí afecta el estado en el que llega la petición, o lo que es lo mismo, se tiene en cuenta el tipo de petición y no únicamente el estado que resultará de aceptarla. Salvo los casos triviales, las políticas del tipo *trunk reservation* no tienen una solución en forma de producto.

En sistemas multiservicio, el número de estados del proceso de Markov puede llegar a ser extraordinariamente alto para sistemas con un número de servicios y una cantidad de recursos razonable (véase [CC97, sección IV.B] para más detalles sobre este aspecto). El hecho de tener una solución en forma de producto reduce considerablemente la complejidad computacional necesaria y hace posible estudiar sistemas que de otro modo no sería posible. En [Ros95] se puede encontrar un análisis comparativo de distintos métodos para calcular las probabilidades de estado en sistemas cuya solución tiene forma de producto. Cuando el número de unidades básicas de recurso no es excesivamente alto el *algoritmo de convolución* [Ive02] permite resolver el sistema de una manera efectiva.

Lea y Altawaya [LA95] proponen una metodología para definir políticas con solución en forma de producto, utilizando umbrales de ocupación por servicio tanto enteros como fraccionarios. En [JV94] se da una caracterización de la política que es óptima dentro del conjunto de las políticas convexas coordinadas .

Finalmente en [GMP04] y [BS97] se compara la capacidad que puede obtenerse mediante este tipo políticas con la que puede obtenerse cuando se aplican políticas del tipo *trunk reservation*. A cambio de obtener una menor complejidad computacional, las políticas con solución en forma producto reducen la capacidad. Para los escenarios estudiados en [GMP04] esta reducción está entorno al 10 %. Biswas y Sengupta [BS97] caracterizan un tipo de escenarios en los que el beneficio de utilizar una política *trunk reservation* es mayor: cuando existe un tipo de servicio cuya tasa de llegadas es baja, la

cantidad de recursos que consume una sesión de ese servicio es alta y la probabilidad de bloqueo que demanda es baja. En [BF01] se comparan también distintas políticas atendiendo a su respuesta frente a situaciones de sobrecarga severa y se concluye que, desde este punto de vista, la política denominada *upper limit* UL, que es convexa coordinada, presenta ventajas importantes.

Políticas estacionarias aleatorizadas

Aunque en general las políticas GC, FGC y sus correspondientes extensiones para el caso multiservicio permiten alcanzar unas prestaciones relativamente buenas, al menos dentro de las políticas estáticas, existen escenarios en los que no son la mejor alternativa; la política estacionaria óptima no es del tipo *trunk reservation*. En un escenario monoservicio se ha demostrado [RNT97, Bar01] que, para ciertos criterios, la política óptima es del tipo GC o FGC. Este resultado, sin embargo, no es generalizable al escenario multiservicio [BC02, PCG04]. En este caso, la política óptima pertenece a la familia de las políticas estacionarias (para un cierto tipo de petición la decisión de admitirla o no depende del estado actual del sistema expresado como el número de sesiones en curso de cada servicio) o al de las estacionarias aleatorizadas (la decisión depende del estado actual del sistema y de un factor aleatorio) [Ros70]. Nótese que en las políticas *trunk reservation*, que constituyen un subconjunto de las políticas estacionarias, para decidir sobre la admisión de una petición únicamente se considera la cantidad total de recursos en uso, y no el desglose de éstos por cada servicio.

A modo de síntesis en la figura 2.1 se representan los distintos conjuntos de políticas y las relaciones de pertenencia e inclusión que existen entre ellos.

2.2.3. Alternativas para el diseño de la política de CA

De forma general podemos decir que las políticas de CA en redes celulares tratan de resolver eficientemente el compromiso que existe entre el bloqueo de nuevas sesiones y la terminación forzosa de sesiones en curso

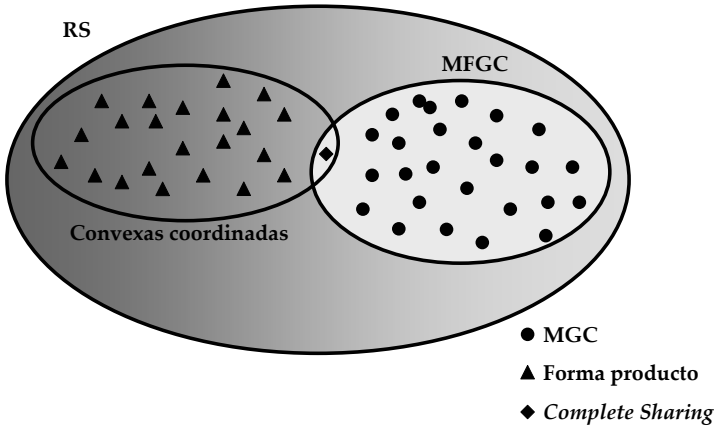


Figura 2.1: Conjuntos de políticas de control de admisión.

causada por handovers fallidos. Para lograr este objetivo se han utilizado fundamentalmente dos enfoques distintos. El primero y más común consiste en proponer una nueva política de CA y evaluarla para demostrar que consigue mejores resultados o presenta alguna ventaja —al menos bajo ciertas condiciones— que las propuestas previas. En el segundo enfoque, el problema de CA se formula como un problema de optimización de un proceso de decisión: ante la llegada de una petición ha de decidirse si aceptarla o rechazarla. Esta decisión dependerá del estado del sistema y del tipo de petición por lo que el problema consiste en encontrar la correspondencia óptima entre cada posible situación y la decisión asociada; óptima en el sentido de que maximiza (minimiza) una función de mérito (coste) de las prestaciones del sistema, sujeta a una serie de restricciones. Este problema de optimización por lo general se formula utilizando el marco de los *procesos de decisión markovianos* (MDP) o semimarkovianos (SMDP) [Ros70], o bien se utiliza cualquier procedimiento de optimización que trabaje con una función de la que no se tiene una forma explícita, pero puede evaluarse numéricamente (esta evaluación puede comportar un coste computacional alto) y de la que puede que se conozcan propiedades de monotonicidad. La formulación basada en MDP

(o SMDP) tiene la ventaja de contar con un marco teórico y un conocimiento más o menos desarrollado pero, por otra parte, las otras alternativas, aunque más desasistidas en este sentido gozan de una mayor flexibilidad para definir la función a optimizar y el conjunto de políticas sobre el que se realiza la optimización.

El enfoque basado en optimización ha sido utilizado en [RNT97] y [Bar01] para el escenario monoservicio utilizando distintos criterios. Yener y Rose [YR97] también emplean este enfoque pero dada la complejidad del problema de optimización resultante utilizan un algoritmo genético para hallar una política cuasióptima. Este enfoque también se ha aplicado en el caso multiservicio en el que la dimensionalidad del problema entraña complejidad importante para realizar la optimización. Para tratar esta complejidad se han examinado distintos métodos. En [BC02, PCG04] se utiliza *programación lineal* para obtener la solución del MDP. Xiao et al. [XCW01] emplean la misma metodología en un contexto en el que consideran aplicaciones multimedia adaptativas y en [XCW00], estos mismos autores proponen la utilización de una algoritmo genético debido a las limitaciones computacionales de plantear el problema como un SMDP. García et al. [GMP04, GMP05] utilizan un método basado en ascenso (*hill climbing*). En [EAYH01a, EAYH01c, EAYH01b, EAYH01d] y también en [PGGMCG04] los autores utilizan un método de aprendizaje automático basado en la teoría del *aprendizaje por refuerzos* [SB98]. Este método presenta la interesante propiedad de no necesitar un modelo del sistema y de poder adaptarse a cambios en los parámetros de entrada.

Capítulo 3

Sistemas monoservicio

Gran parte de la investigación sobre el control de admisión en redes celulares considera únicamente un servicio distinguiendo entre sesiones nuevas y peticiones de handover, bien porque la red ofrece un único servicio —redes de primera o segunda generación— o bien porque aun ofreciendo varios servicios existe una separación de los recursos asignados a los distintos servicios.

La literatura sobre propuestas de mecanismos de prioridad para sistemas con dos tipos de tráfico y su estudio es extensísima, incluso cuando nos limitamos a la aplicación en redes celulares (véase por ejemplo [Bar04, DS04] y sus referencias). En [Bar04], Barceló propone un marco genérico de análisis de una familia de mecanismos de control del admisión, SRS (*State-dependent Rejection Scheme*), en los que la probabilidad de aceptar una nueva sesión depende del número de canales ocupados en ese momento. Por otra parte, existe otro grupo de mecanismos [CH96, KA99, CG01, PC01, DS04] que en principio no pertenecen a la familia SRS ya que no todos los canales son considerados equivalentes, y la probabilidad de aceptar una nueva sesión depende no sólo del número total de canales ocupados sino también de su tipo. No obstante, al menos desde una perspectiva teórica es importante señalar que, según [DS04, Proposición 5.1], para los mecanismos de este otro grupo,

si no se considera la posibilidad de espera para las peticiones de baja prioridad, es posible encontrar un mecanismo *equivalente* que sí pertenecería a la clase SRS.

Cuando se adoptan las hipótesis habituales de llegadas de Poisson y tiempo de ocupación de los recursos distribuido exponencialmente, los esquemas de la familia SRS pueden describirse mediante un proceso markoviano de nacimiento y muerte. Si los canales no son indistinguibles como en SRS, o aun siéndolo las llamadas nuevas no son tratadas según un modelo puramente de pérdidas, el proceso que describe el sistema ya no es de nacimiento y muerte sino cuasi de nacimiento y muerte (Quasi-Birth-and-Death, QBD).

Mientras que el análisis numérico de un proceso de nacimiento y muerte, en general, no reviste ninguna dificultad, en los procesos QBD la complejidad numérica puede llegar a ser considerable y también pueden aparecer problemas de precisión [KI95, Mit95].

En este capítulo se analiza una familia de mecanismos de control de admisión para sistemas con dos tipos de tráfico de distinta prioridad. Estos algoritmos y su análisis son una compilación, y en cierto modo una generalización, de una serie de propuestas que han aparecido en la literatura especializada. El capítulo está dividido en dos partes: en la primera de ellas se desarrolla un modelo analítico para la evaluación de prestaciones basado en el método geométrico-matricial [Neu81, LR99]; y en la segunda se evalúan las ventajas de aplicar un método de análisis basado en técnicas espectrales como las utilizadas en [KI95]. A lo largo de este capítulo se conserva la terminología propia de una red de telefonía celular en la que en su interfaz radio se emplea una tecnología de acceso basada en la división en canales. No obstante, los algoritmos de este capítulo y su análisis seria extensible a un entorno más general en el que todas las sesiones consumen la misma cantidad de recursos.[1] En este caso el término llamada se referiría a una sesión, y canal a los recursos utilizados por una sesión.

[1] El definición del término sesión y de los recursos que ésta utiliza se han introducido en el punto 2 de la página 8.

3.1. Algoritmos de prioridad con dos flujos de tráfico

3.1.1. Descripción de los algoritmos

Para la especificación de los algoritmos partimos de una descripción alternativa a la habitual del mecanismo FGC [RNT97] en la que además se considera la posibilidad de espera para ambos tipos de peticiones. A partir de esta descripción alternativa el resto de mecanismos aparecen como una extensión más o menos natural del primero.

Los canales se reparten en tres grupos: grupo primario, grupo secundario y canal parcialmente reservado (grupo de un único canal). La idea general en todos los algoritmos es que los canales del grupo primario pueden ser asignados a cualquier tipo de petición —llamada nueva o handover— los del grupo secundario se reservan para las peticiones de mayor prioridad —handover— y el canal parcialmente reservado puede asignarse a una petición de llamada nueva sólo a veces —con cierta probabilidad—. C representa el número total de canales y el parámetro t ($t \in \mathbb{R}$, $0 < t < C$) establece el número de canales de los grupos primario y secundario y la probabilidad con la que el canal parcialmente reservado puede utilizarse para las llamadas nuevas. Así, si m representa el número de canales en el grupo primario, n en el grupo secundario y f la probabilidad de que una llamada nueva acceda al canal parcialmente reservado, se tiene que

$$m = \lfloor t \rfloor \qquad f = t - m \quad \text{y} \quad n = C - (m + 1).$$

De este modo el número de canales no reservados es, en media, $m + f \cdot 1 = t$. Por tanto, mediante este mecanismo probabilístico se puede reservar, en media, una cantidad no entera de canales lo cual permite un ajuste más fino del control de admisión.

Para cada tipo de petición existe una cola de espera para aquellas peticiones que no pueden ser atendidas en el momento de su llegada. Ambas

colas son de capacidad finita y las llegadas a una cola llena se pierden. En el modelo se considera que las peticiones de ambos tipos tienen una paciencia limitada por lo que pueden abandonar la cola antes de recibir servicio si el tiempo de espera excede un determinado límite.

En cada algoritmo existen tres tipos de evento que disparan la realización de una acción por parte del sistema de asignación de recursos. Los eventos posibles son: la llegada de una petición de handover, la llegada de una llamada nueva y la liberación de un canal. A continuación se describe, para cada algoritmo, la acción o acciones a realizar por el sistema cuando se produce alguno de estos eventos.

FGC *Fractional Guard Channel*

Llegada de una petición de handover.

Intentar la asignación de un canal o de una posición en la cola, siguiendo la secuencia:

1. Canal del grupo primario.
2. Canal parcialmente reservado.
3. Canal del grupo secundario.
4. Posición en la cola de handovers.

Si ninguna de estas asignaciones es posible se rechaza la petición.

Llegada de una llamada nueva.

Intentar la asignación de un canal o de una posición en la cola, siguiendo la secuencia:

1. Canal del grupo primario.
2. Con probabilidad f, canal parcialmente reservado.
3. Posición en la cola de llamadas nuevas.

Si ninguna de estas asignaciones es posible se rechaza la petición.

Liberación de un canal.

Si hay alguna petición de handover esperando en la cola, asignar el canal a la primera de ellas. Si no,

- reetiquetar los canales[2] e
- intentar la asignación del canal libre a la primera petición de la cola de llamadas nuevas siguiendo el mismo criterio que cuando llega una llamada nueva.

F-HOPSWR ***Fractional-Handovers Overflow from Primary to Secondary With Rearrangement***

Llegada de una petición de handover.

Igual que en el algoritmo FGC.

Llegada de una llamada nueva.

Igual que en el algoritmo FGC.

Liberación de un canal.

Realizar la primera de la acciones siguientes que sea posible:

1. Si hay alguna petición de handover esperando en la cola, asignar el canal a la primera de ellas.
2. Si hay alguna petición de llamada nueva en la cola, intentar la asignación del canal libre a la primera de ellas siguiendo el mismo criterio que cuando llega una llamada nueva.
3. Reetiquetar los canales.

[2] Se comienza por los canales ocupados y se les asignan las etiquetas en el orden siguiente: grupo primario, canal parcialmente reservado y grupo secundario. Las etiquetas sobrantes se asignan a los canales desocupados. De este modo primero se ocupan todos los canales del grupo primario, después el canal parcialmente reservado y por último los canales del grupo secundario.

F-HOPS ***Fractional-Handovers Overflow from Primary to Secondary***

Llegada de una petición de handover.
Igual que en el algoritmo FGC.

Llegada de una llamada nueva.
Igual que en el algoritmo FGC.

Liberación de un canal.
Realizar la primera de la acciones siguientes que sea posible:

1. Si hay alguna petición de handover esperando en la cola, asignar el canal a la primera de ellas.
2. Si hay alguna petición de llamada nueva en la cola, intentar la asignación del canal libre a la primera de ellas siguiendo el mismo criterio que cuando llega una llamada nueva.

F-HOSP ***Fractional-Handovers Overflow from Secondary to Primary***

Llegada de una petición de handover.
Intentar la asignación de un canal o de una posición en la cola, siguiendo la secuencia:

1. Canal del grupo secundario.
2. Canal parcialmente reservado.
3. Canal del grupo primario.
4. Posición en la cola de handovers.

Si ninguna de estas asignaciones es posible se rechaza la petición.

Llegada de una llamada nueva.
Igual que en el algoritmo FGC.

Liberación de un canal.
Igual que en el algoritmo F-HOPS.

3.1.2. Antecedentes

Los algoritmos de control de admisión con dos niveles de prioridad que se analizan aquí son una compilación, y en algunos casos una generalización, de una familia de mecanismos que se aparecen en la literatura especializada.

El algoritmo denominado *Guard Channel* (GC) —entre muchas otras formas— fue introducido como una técnica de CA en redes celulares a mediados de los ochenta [PG85, HR86]. En [RNT97] se introduce la técnica *Fractional Guard Channel* (FGC) que es una generalización del GC en la que el número de canales reservados puede ser fraccionario. Schehrer [Sch90] y posteriormente McMillan [McM95] proponen y analizan una extensión del GC que incorpora un ciclo de histéresis que gobierna la cantidad de recursos reservados. En [McM95] se considera la posibilidad de espera y abandono por impaciencia para los dos tipos de llegada. Casares y Holtzman [CH96] y Casares [CG01] estudian toda una familia de extensiones al mecanismo GC en el contexto de un sistema de *trunking*. Del análisis de [CH96, CG01] destaca el hecho de que se consideran tasas de servicio distintas para los dos tipos de llegada; en nuestro análisis, y en el resto de trabajos que aquí citamos, se supone que el tiempo medio de ocupación de los recursos en una célula es el mismo independientemente de si se trata de una llamada nueva o una llamada que se traspasó de otra célula. Los algoritmos que consideramos aquí son un subconjunto de los estudiados por Casares, sin embargo, los que se analizan aquí son más generales por cuanto incorporan posibilidad de espera para las peticiones de alta prioridad, el parámetro de configuración de cada algoritmo puede ser fraccionario y en ambas colas se considera la posibilidad de abandono por impaciencia. En un trabajo anterior [PC01] ya habíamos considerado estos mecanismos en los que además se incorporaba un ciclo de histéresis como el de [Sch90, McM95]. No obstante, en [PC01] no existe abandono para las llamadas nuevas que esperan y el parámetro de configuración de los algoritmos ha de ser entero. Kulavaratharasah y Aghvami [KA99] estudian mediante simulación los mecanismos HOPS, HOSP y un tercero que es una combinación probabilística de los dos primeros. En [KA99]

el número de canales reservados es entero y no hay espera para ninguno de los dos tipos de petición. Daley y Servi [DS02, DS04] introducen la reserva fraccionaria en los mecanismos HOPS y HOSP, sin embargo, aunque al final de [DS04] se sugiere la posibilidad de considerar la espera y el abandono por impaciencia, estos aspectos no se incluyen en los resultados del artículo.

3.1.3. Descripción del modelo

A la célula llegan peticiones de llamadas nuevas y de handover con unas tasas λ_n y λ_h, respectivamente, y $\lambda = \lambda_n + \lambda_h$ representa la tasa agregada. Ambos flujos de llegada siguen un proceso de Poisson. Los tiempos listados a continuación se describen mediante variables aleatorias exponenciales cuyo parámetro se indica entre paréntesis: duración de una llamada (μ_c), tiempo de permanencia en una célula (μ_r), tiempo de permanencia en el área de handover (μ'_r) y tiempo máximo de espera en cola de una llamada nueva (η). Por tanto, las variables aleatorias siguientes también seguirán una distribución exponencial, con los parámetros indicados: tiempo de ocupación de recursos ($\mu = \mu_c + \mu_r$) y tiempo máximo de espera en cola de una petición de handover ($\gamma = \mu_c + \mu'_r$). El número de posiciones en la cola de llamadas nuevas es Q_n y en la de peticiones de handover Q_h. Anteriormente se han introducido los parámetros C y t: C representa el número total de canales, t ($0 < t < C$) establece el número de canales de los grupos primario y secundario, y la probabilidad con la que el canal parcialmente reservado puede utilizarse para las llamadas nuevas. Así, $m = \lfloor t \rfloor$ es el número de canales en el grupo primario, $n = C - (m + 1)$ en el grupo secundario y $f = t - m$ la probabilidad de que una llamada nueva acceda al canal parcialmente reservado.

Las prestaciones del sistema se cuantifican mediante las probabilidades de bloqueo y abandono, cuya representación es: probabilidad de bloqueo (abandono) de una llamada nueva P_b^n (P_a^n), probabilidad de bloqueo (abandono) de una petición de handover P_b^h (P_a^h). Por tanto, la probabilidad de pérdida de una petición —porque es bloqueada o abandona por impaciencia, y sólo una de las dos— será $P^n = P_b^n + P_a^n$, para las llamadas nuevas y $P^h = P_b^h + P_a^h$,

para las peticiones de handover.

La descripción del estado del sistema no es la misma para todos los algoritmos sino que utilizamos dos diferentes: una para los algoritmos FGC y F-HOPSWR, y otra para los algoritmos F-HOPS y F-HOSP. En todos los casos la representación elegida da lugar a un modelo que es un proceso *cuasi de nacimiento y muerte* (QBD), finito y no homogéneo [Neu81]. A continuación se describen las dos alternativas:

Algoritmos FGC y F-HOPSWR. El estado del sistema se representa mediante la terna de números enteros,

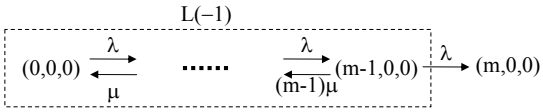
$$(k, i, j) \qquad 0 \leq k \leq C, \quad 0 \leq i \leq Q_n, \quad , 0 \leq j \leq Q_h$$

donde k es el número de canales ocupados, i es el número de peticiones en la cola de llamadas nuevas y j es el número de peticiones en la cola de handovers. El *nivel* y la *fase* de cada estado no se corresponden directamente con ninguna de sus coordenadas —k, i o j— sino que se sigue el criterio siguiente. Si $L(l_0)$ es el conjunto de estados del nivel l_0,

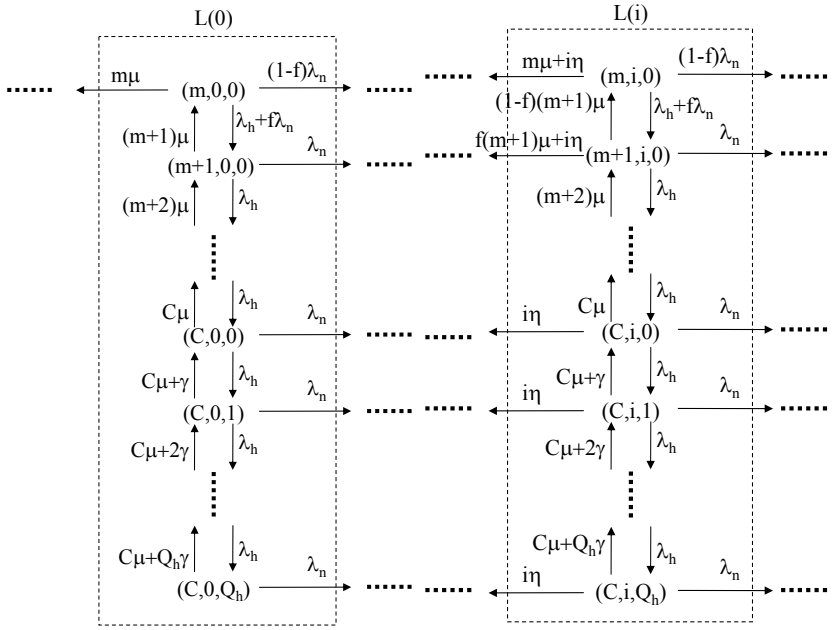
$$\begin{aligned} L(-1) &= \{(k, 0, 0) : k < m\} \\ L(l_0) &= \{(k, l_0, j) : k \geq m, 0 \leq j \leq Q_h\} \qquad l_0 = 0, \ldots, Q_n \end{aligned}$$

Esto es, en el nivel -1 se agrupan todos los estados en los que el número de canales ocupados es inferior a m — todavía quedan canales libres en el grupo primario — y cuando el número de canales ocupados es igual o superior a m el nivel está determinado por el número de llamadas nuevas en la cola. En el nivel -1 el número de canales ocupados k representa la fase. En el resto de niveles ($l_0 = 0, \ldots, Q_n$) la fase de un estado se corresponde con el número de canales ocupados que excede de m más el número de peticiones de handover en la cola, es decir, la fase del estado (k, l_0, j) es $k - m + j$.

En la figura 3.1 se representa en el diagrama de transiciones del algoritmo FGC y en la figura 3.2 el del algoritmo F-HOPSWR. En cada nivel solo se han representado las transiciones internas y las de salida de ese nivel.



(a) Nivel −1.



(b) Nivel 0. (c) Nivel *i*.

Figura 3.1: FGC: diagrama de transiciones.

Algoritmos F-HOPS y F-HOSP. El estado del sistema se representa mediante la quintupla de números enteros

$$(i,j,r,k,l) \qquad 0 \le i \le m; \quad 0 \le j \le n; \quad r = 0,1; \quad 0 \le k \le Q_n; \quad 0 \le l \le Q_h$$

donde i es el número de canales del grupo primario ocupados, j es el número de canales del grupo secundario ocupados, r indica si el canal parcialmente reservado está ocupado ($r = 1$) o no ($r = 0$), k es el número de peticiones en la cola de llamadas nuevas y l es el número de peticiones en la cola de

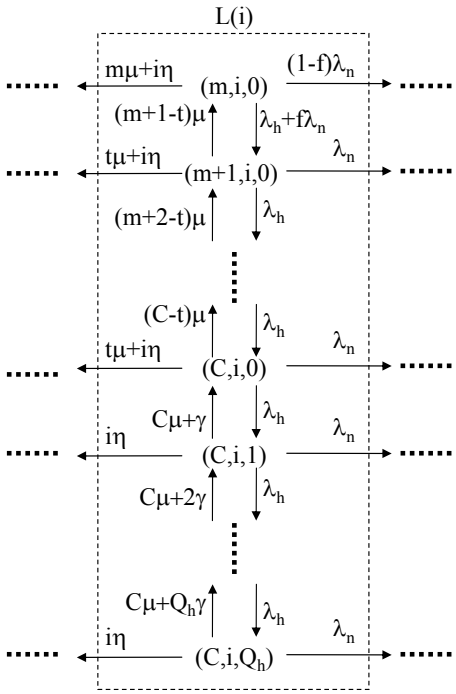


Figura 3.2: F-HOPSWR: diagrama de transiciones nivel i. Las transiciones de los niveles -1 y 0 son las de la figura 3.1

handovers.

La agrupación de los estados en niveles es del siguiente modo

$$L(l_0) = \{(i,j,r,k,l) : i - m + k = l_0\} \qquad l_0 = -m, \ldots, Q_n.$$

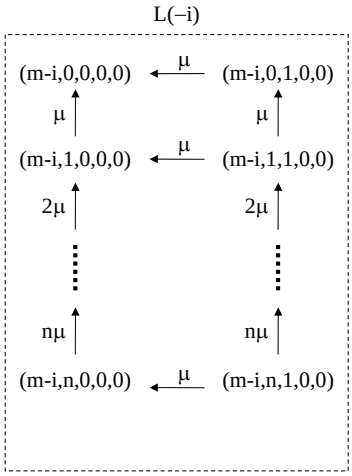
Dependiendo del número de estados (fases) podemos distinguir dos tipos de niveles

$$L(l_0) = \{(l_0 + m, j, r, 0, 0) : 0 \leq j \leq n; r = 0, 1\} \qquad l_0 = -m, \ldots, -1$$

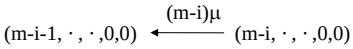
$$L(l_0) = \{(m,j,r,k,l) : 0 \leq j \leq n; r = 0,1; 0 \leq k \leq Q_n; 0 \leq l \leq Q_h\}$$
$$l_0 = 0, \ldots, Q_n.$$

Al igual que antes, los niveles cuyo índice es negativo se corresponden con aquéllos en los que todavía quedan canales libres para acomodar la llegada de una llamada nueva. Los niveles del $-m$ al -1 tienen por tanto $2(n+1)$ fases y los niveles del 0 al Q_n tienen $2(n+1)+Q_h$ fases. Así, el nivel del estado (i,j,r,k,l) es $i-m+k$ y, su fase $j+l+r(n+1)$.

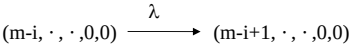
En la figura 3.3 se representa en el diagrama de transiciones del algoritmo F-HOPS y en figura 3.4 el del algoritmo F-HOSP.



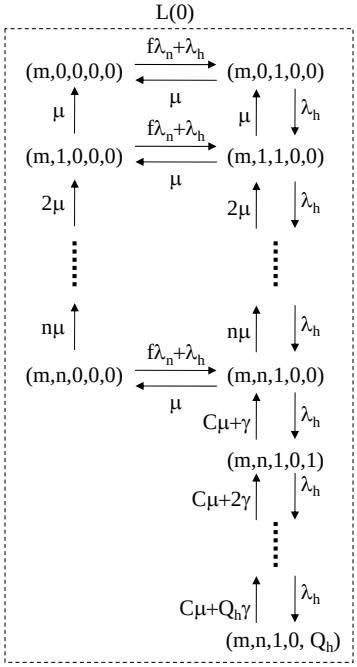
(a) Nivel $-i$, $m \geq i > 0$



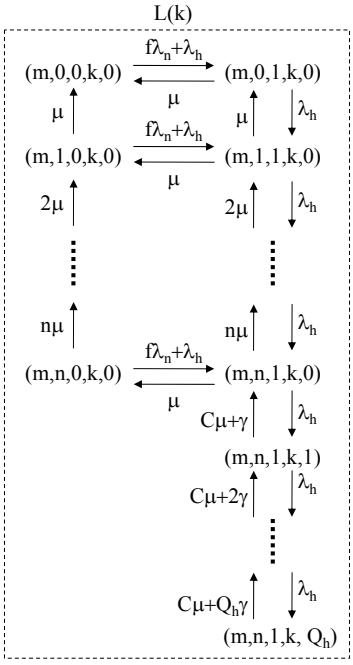
(b) $L(-i-1) \longleftarrow L(-i)$, $m > i \geq 0$



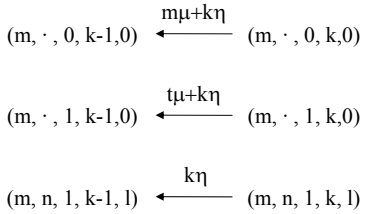
(c) $L(-i) \longrightarrow L(-i+1)$, $m \geq i > 0$



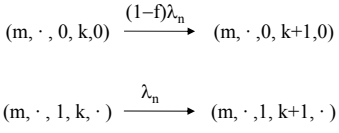
(d) Nivel 0



(e) Nivel k, $0 \leq k \leq Q_n$

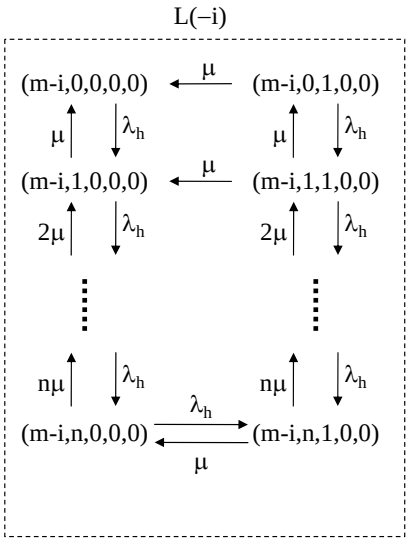


(f) $L(k-1) \longleftarrow L(k)$, $0 < k \leq Q_n$, $0 < l \leq Q_h$

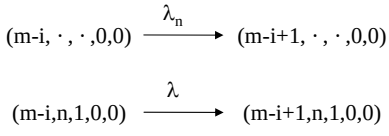


(g) $L(k) \longrightarrow L(k+1)$, $0 \leq k < Q_n$

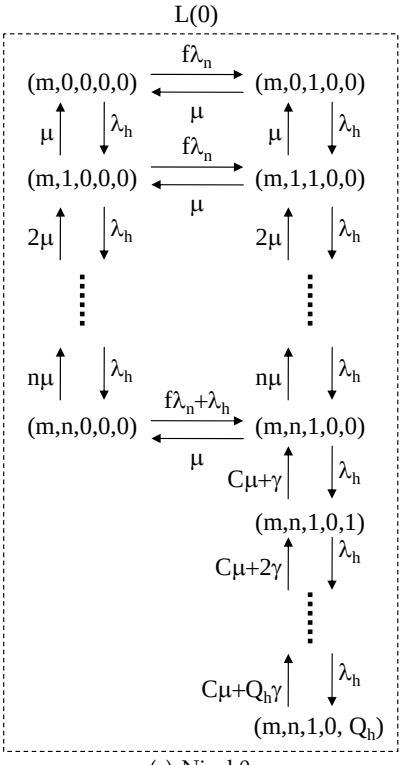
Figura 3.3: F-HOPS: diagrama de transiciones.



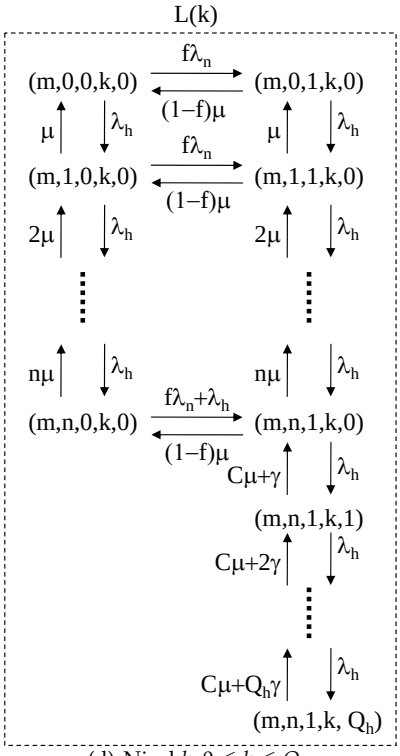
(a) Nivel $-i$, $m \geq i > 0$



(b) $L(-i) \longrightarrow L(-i+1)$, $m \geq i > 0$



(c) Nivel 0



(d) Nivel k, $0 \leq k \leq Q_n$

Figura 3.4: F-HOSP: diagrama de transiciones. Aquellas transiciones no recogidas en esta figura coinciden con las de la figura 3.3.

3.1.4. Análisis

Sea $\boldsymbol{\pi}$ el vector de las probabilidades estacionarias del proceso. Del mismo modo que con los estados, dividimos $\boldsymbol{\pi}$ en otros vectores más pequeños $\boldsymbol{\pi}^{(l_0)}$ correspondiendo cada uno de ellos a las probabilidades de estado de un nivel, por lo que el vector $\boldsymbol{\pi}^{(l_0)}$ tendrá tantas componentes como fases en el nivel l_0. El proceso que describe el comportamiento de cualquiera de los cuatro algoritmos es un proceso QBD pues únicamente existen transiciones entre estados del mismo nivel o de dos niveles adyacentes y, en consecuencia, el generador infinitesimal del proceso tiene una estructura tridiagonal a bloques

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{A}_1^{(-J)} & \boldsymbol{A}_0^{(-J)} & & & & & \\ \boldsymbol{A}_2^{(-J+1)} & \boldsymbol{A}_1^{(-J+1)} & \boldsymbol{A}_0^{(-J+1)} & & & & \\ & \cdots & \cdots & \cdots & & & \\ & & \boldsymbol{A}_2^{(0)} & \boldsymbol{A}_1^{(0)} & \boldsymbol{A}_0^{(0)} & & \\ & & & \cdots & \cdots & \cdots & \\ & & & & \boldsymbol{A}_2^{(Q_n)} & \boldsymbol{A}_1^{(Q_n)} & \end{bmatrix}, \tag{3.1}$$

siendo $J = 1$ para los algoritmos FGC y F-HOPSWR, y $J = m$ para los algoritmos F-HOPS y F-HOSP. En el apéndice C se puede encontrar la estructura de los distintos bloques de Q para cada algoritmo.

Las convenciones utilizadas en la notación son las habituales (véase el apéndice A).

Las probabilidades de estado $\boldsymbol{\pi}$ se obtienen de la resolución del sistema de ecuaciones lineales

$$\boldsymbol{\pi}\boldsymbol{Q} = \boldsymbol{0}^t, \quad \boldsymbol{\pi}\boldsymbol{e} = 1.$$

Si $\boldsymbol{Q}$ es una matriz de dimensiones finitas, como es nuestro caso, este sistema en principio puede resolverse mediante cualquiera de los métodos estándar del álgebra lineal. Sin embargo, parece conveniente —sobre todo si el tamaño del sistema es grande— aprovechar la estructura y la naturaleza de $\boldsymbol{Q}$, que es un generador infinitesimal tridiagonal por bloques. Aquí hemos utilizado el

algoritmo *Linear Level Reduction* [LR99, GJL84], que se aplica a la resolución de procesos QBD finitos y no homogéneos:

$U \leftarrow \boldsymbol{A}_1^{(Q_n)}$
$R^{(Q_n)} \leftarrow \boldsymbol{A}_0^{(Q_n-1)} (-U)^{-1}$
for $l = Q_n - 1, Q_n - 2, \ldots, 0, -J$ **do**
  $U \leftarrow \boldsymbol{A}_1^{(l)} + R^{(l+1)} \boldsymbol{A}_2^{(l+1)}$
  $R^{(l)} \leftarrow \boldsymbol{A}_0^{(l-1)} (-U)^{-1}$
end for

solve $\boldsymbol{\pi}^{(-J)}$ **from** $\{\boldsymbol{\pi}^{(-J)} U = \mathbf{0}^t;\ \boldsymbol{\pi}^{(-J)} \boldsymbol{e} = 1\}$
for $l = -J+1, \ldots, 0, \ldots, Q_n$ **do**
  $\boldsymbol{\pi}^{(l)} = \boldsymbol{\pi}^{(l-1)} R^{(l)}$
end for

A partir de las probabilidades de estado los parámetros de medida de prestaciones se obtienen del siguiente modo

$$P_b^n = \boldsymbol{\pi}^{(Q_n)} \cdot \begin{cases} \left[(1-f) \quad \overbrace{1 \quad \cdots \quad 1}^{n+1+Q_h} \right]^t, & \text{FGC, F-HOPSWR} \\ \left[\overbrace{(1-f) \quad \cdots \quad (1-f)}^{n+1} \quad \overbrace{1 \quad \cdots \quad 1}^{n+1+Q_h} \right]^t, & \text{F-HOPS, F-HOSP} \end{cases}$$

$$P_a^n = \frac{1}{\lambda_n} \sum_{r=1}^{Q_n} \boldsymbol{\pi}^{(r)} r \eta \boldsymbol{e}$$

$$P_b^h = \sum_{r=0}^{Q_n} \boldsymbol{\pi}^{(r)} \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix}^t$$

$$P_a^h = \frac{1}{\lambda_h} \sum_{r=0}^{Q_n} \boldsymbol{\pi}^{(r)} \mu_r' \cdot \begin{bmatrix} 0 & \cdots & 0 & 1 & 2 & \cdots & Q_h \end{bmatrix}^t.$$

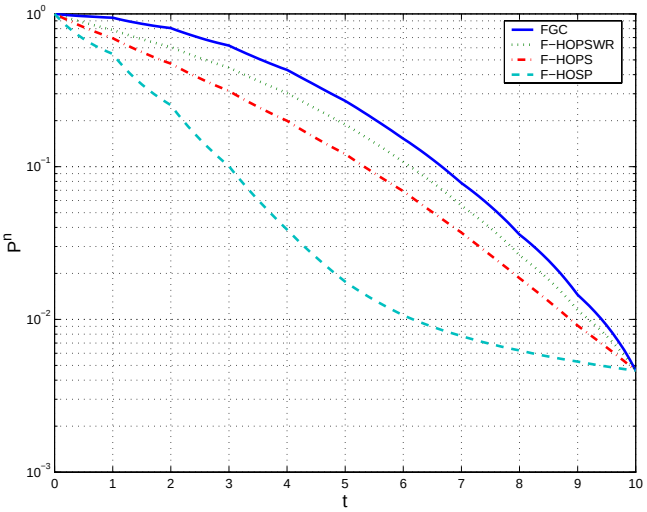


Figura 3.5: Probabilidad de fallo de llamada nueva P^n.

3.1.5. Resultados numéricos

En los ejemplos numéricos mostrados a continuación se ha considerado la siguiente configuración: $\mu_r/\mu_c = 2$, $\mu_c + \mu_r = 1$ llamadas/s, $\lambda_n = 1.5$ llamadas/s, $\eta/\mu_c = \mu'_r/\mu_r = 10$, $C = 10$, $t = 8$, $Q_n = Q_h = 10$. En los distintos ejemplos se ha ido variando el valor de algunos parámetros de esta configuración básica. En las figuras 3.5 y 3.6 se muestra la variación en función de t de las probabilidades de pérdida de una petición nueva $P^n = P^n_b + P^n_a$, que puede deberse a la perdida inmediata por falta de espacio de almacenamiento (P^n_b) o al abandono por impaciencia (P^n_a), y de terminación forzosa $P^{ft} = \mu_r/\mu_c P^h/(1 + \mu_r/\mu_c P^h)$ donde $P^h = P^h_b + P^h_a$ es la probabilidad de fallo de un handover, que también puede estar causado por el bloqueo por falta de espacio de almacenamiento o el abandono por impaciencia. Estas dos probabilidades varían de forma monótona y continua con el valor de t.

En las figuras de la 3.7 a la 3.10 se estudia el impacto del tamaño de las colas (Q_n y Q_h) sobre P^n y P^{ft}. El signo de este impacto es el esperado: un

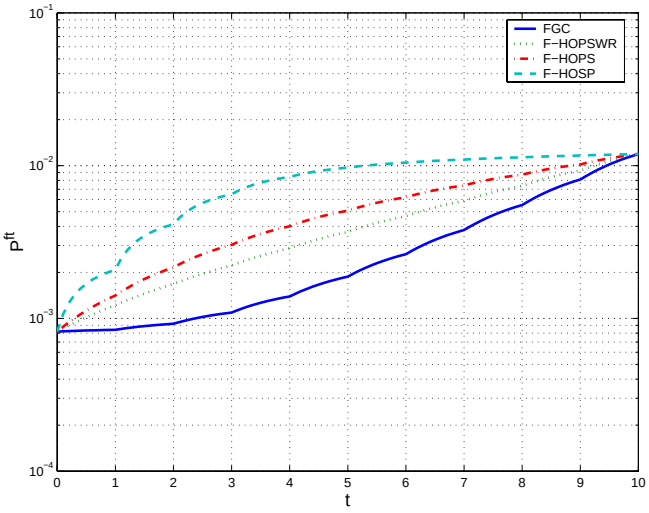


Figura 3.6: Probabilidad de terminación forzosa P^{ft}.

aumento de Q_n influye positivamente en P^n y negativamente en P^{ft} mientras que el aumento de Q_h produce el efecto contrario. Sin embargo, lo más llamativo es que este efecto es prácticamente despreciable para tamaños de cola por encima de unas pocas unidades. En las figuras 3.11– 3.14 se muestran los resultados utilizando unas tasas de impaciencia menores $\eta/\mu_c = \mu'_r/\mu_r = 2$ y aunque inicialmente el impacto de aumentar el tamaño de las colas es mayor, éste se atenúa rápidamente y de nuevo deja de ser perceptible a partir de un tamaño de muy pocas unidades.

Finalmente, en la figura 3.15 se comparan las prestaciones de los diferentes algoritmos fijando un objetivo de QoS en función de la probabilidad de terminación forzosa $P^{ft} \leq 0.005$ y viendo cuál sería el valor de P^n en cada caso. Las curvas de la figura 3.15(a) representan el valor máximo de t para el que se cumple el objetivo y en la gráfica de la figura 3.15(b) se representa P^n para el valor calculado de t. Según este criterio el algoritmo FGC es superior al resto, hecho este que ha sido demostrado para el caso particular en el que no hay colas ($Q_n = Q_h = 0$) en [DS04].

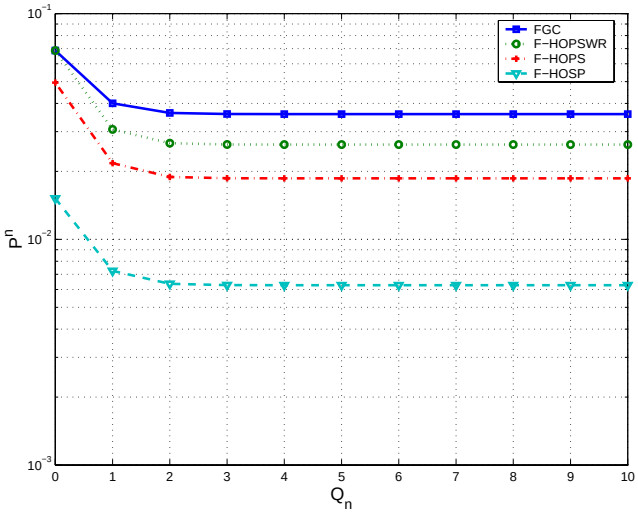


Figura 3.7: Influencia del valor de Q_n en P^n.

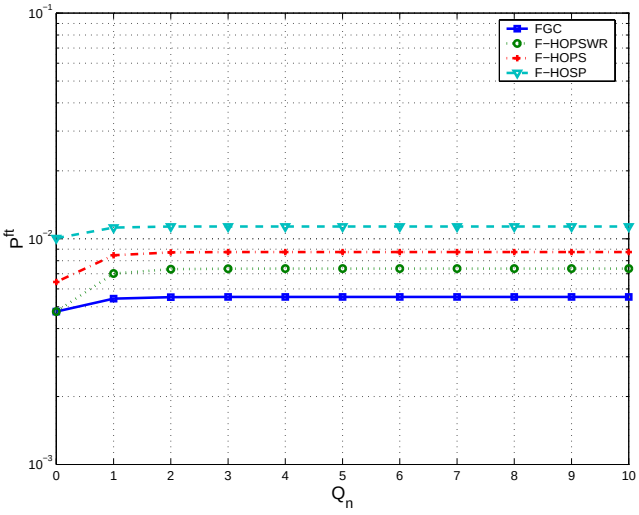


Figura 3.8: Influencia del valor de Q_n en P^{ft}.

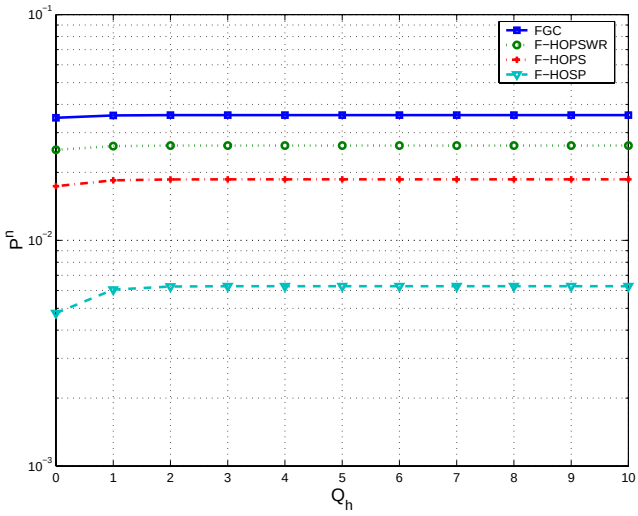


Figura 3.9: Influencia del valor de Q_h en P^n.

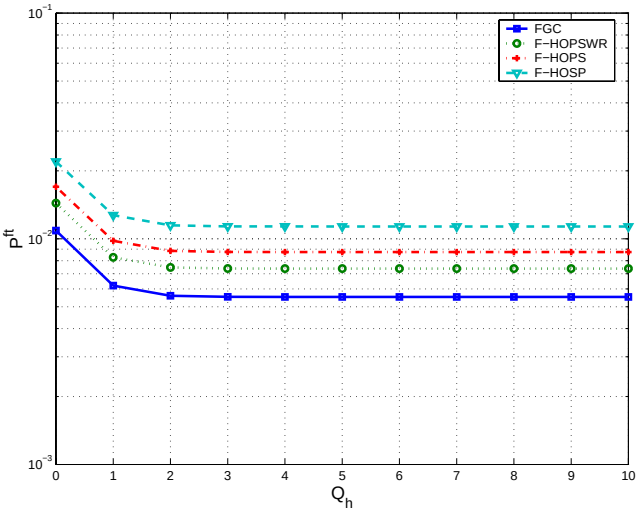


Figura 3.10: Influencia del valor de Q_h en P^{ft}.

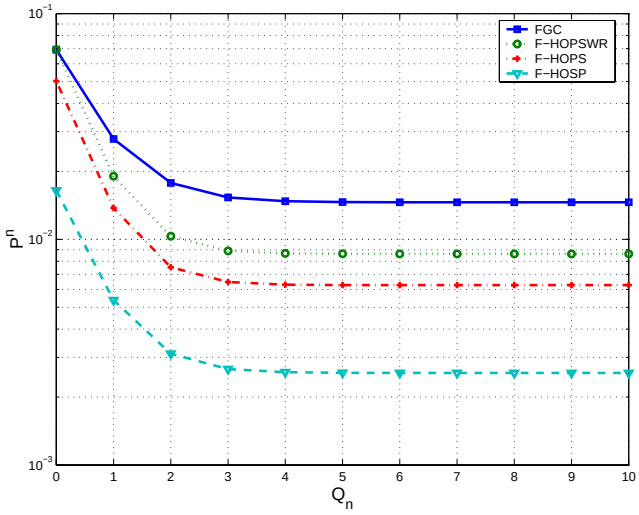


Figura 3.11: Influencia del valor de Q_n en P^n; $\eta/\mu_c = \mu'_r/\mu_r = 2$.

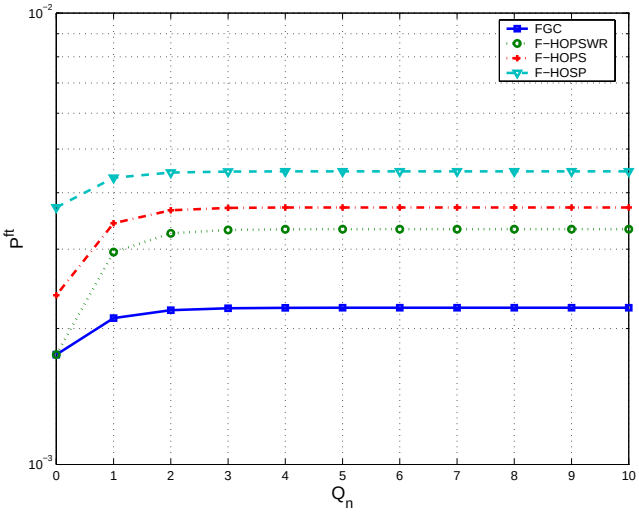


Figura 3.12: Influencia del valor de Q_n en P^{ft}; $\eta/\mu_c = \mu'_r/\mu_r = 2$.

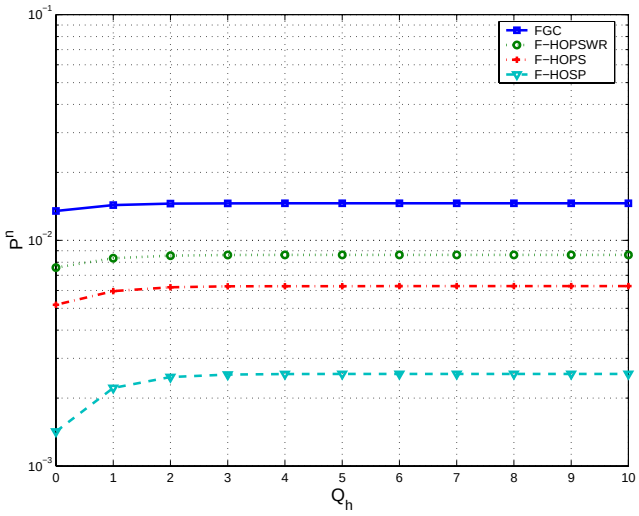


Figura 3.13: Influencia del valor de Q_h en P^n; $\eta/\mu_c = \mu_r'/\mu_r = 2$.

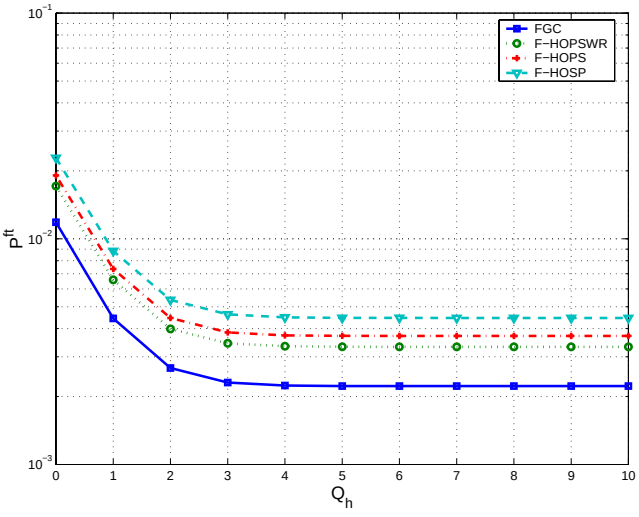
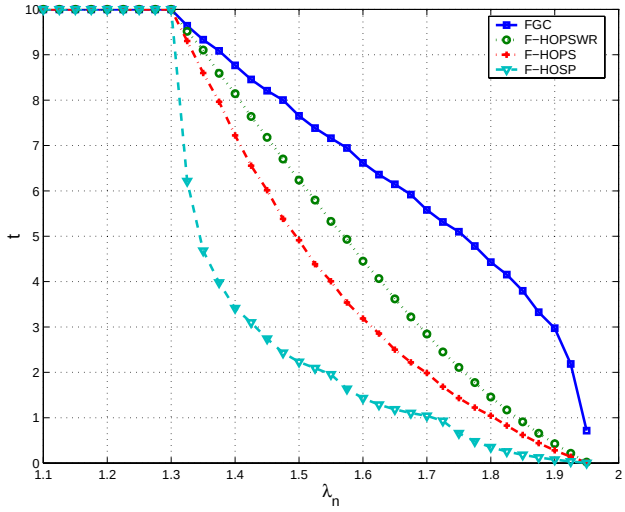
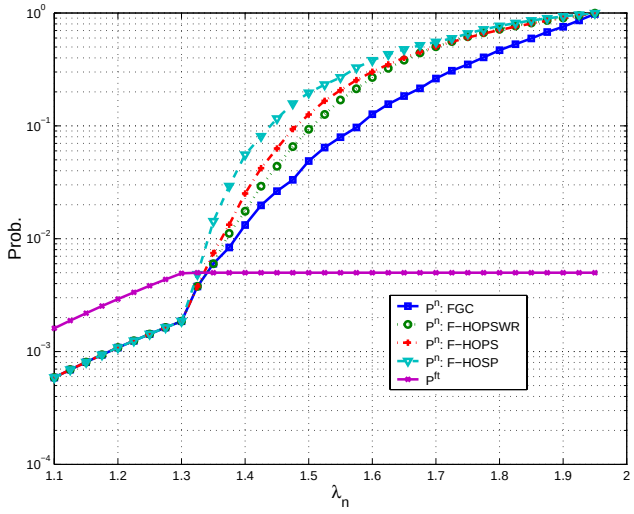


Figura 3.14: Influencia del valor de Q_h en P^{ft}; $\eta/\mu_c = \mu_r'/\mu_r = 2$.



(a) valor de t



(b) valor de P^n y P^{ft}

Figura 3.15: Ajuste del parámetro t para que $P_{ft} \leq 0.005$.

3.2. Aspectos numéricos: método espectral

En los algoritmos de asignación de recursos que estamos considerando, cuando no basta una única dimensión para definir el estado del sistema —cosa que ocurre por ejemplo cuando existe espera para las peticiones de bajas prioridad—, el modelo markoviano del sistema ya no es un proceso de nacimiento y muerte sino un proceso QBD. La complejidad del análisis de un proceso QBD es, frente al de un proceso de nacimiento y muerte, superior tanto desde un punto de vista algebraico como numérico. El algoritmo *Guard Channel*(GC) con espera para las llamadas nuevas se ha analizado en [Gué88, DJ92, McM95, KI95] aplicando distintos métodos. De estos métodos, aquí nos centramos en el empleado por Keilson e Ibe [KI95], que está basado en la utilización de la función generatriz del proceso subyacente junto con técnicas espectrales matriciales. Este método se fundamenta en los mismos principios que la técnica propuesta por Mitrani [Mit95], por lo que comparten las mismas ventajas en la evaluación numérica: menor complejidad computacional y mayor precisión.

En esta sección extendemos el análisis de [KI95] —que se aplica a una versión menos general del algoritmo FGC descrito en 3.1.1— a los otros tres algoritmos —F-HOPSWR, F-HOPS y F-HOSP, aplicándoles las mismas particularizaciones que al FGC— y posteriormente comparamos la eficacia de esta técnica de análisis con la del método geométrico-matricial.

3.2.1. Descripción de los algoritmos y su modelo

Los algoritmos que se analizan en esta sección son una particularización de los tratados en la sección anterior en los que: el tamaño de la cola de llamadas nuevas es infinito ($Q_n = \infty$) y estas peticiones no presentan impaciencia ($\eta = 0$); no existe cola para las peticiones de handover ($Q_h = 0$); el parámetro t sólo toma valores enteros por lo que no existirá canal parcialmente reservado, el número de canales en el grupo primario es $m = \lfloor t \rfloor = t$ y en el secundario $n = C - m$. Estas versiones particularizadas de los algoritmos

se denotarán eliminando la letra F de los acrónimos introducidos en 3.1.1, es decir: GC, HOPSWR, HOPS y HOSP.

Con estas particularizaciones se puede utilizar una descripción del estado del sistema más simple, con dos componentes únicamente (r, k) cuyo significado es como sigue:

HOPSWR r es el número de canales ocupados y k es el número de peticiones en la cola de llamadas nuevas.

HOPS,HOSP r es el número total de llamadas nuevas en el sistema (en curso o esperando en la cola) y k el número de canales ocupados que corresponden a peticiones de handover.

3.2.2. Análisis

En primer lugar se describe con detalle el análisis del algoritmo HOPSWR, a continuación se analiza el algoritmo HOPS omitiendo aquellos puntos que sean una repetición de lo anterior y, finalmente, el análisis del algoritmo HOSP se describe haciendo referencia al del HOPS y destacando únicamente las diferencias entre ambos.

Algoritmo HOPSWR

Si $p_{r,k}$ representa la probabilidad estacionaria del estado (r, k), definimos

$$\boldsymbol{P}_k^t = [p_{m,k}, p_{m+1,k}, \ldots, p_{m+n,k}]$$

y sea $\boldsymbol{Q}$ el generador infinitesimal del proceso resultante de considerar únicamente los estados de fase $k \geq 1$, $(\cdot, k)$ y las transiciones entre ellos

$$\boldsymbol{Q} = \begin{bmatrix} * & \lambda_h & 0 & \cdots\cdots & 0 & 0 \\ \mu & * & \lambda_h & \cdots\cdots & 0 & 0 \\ 0 & 2\mu & * & \cdots\cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots\cdots & * & \lambda_h \\ 0 & 0 & 0 & \cdots\cdots & n\mu & * \end{bmatrix}, \tag{3.2}$$

donde los asteriscos $(*)$ toman el valor necesario para que la suma de los elementos de una fila sea cero. Definimos también $\boldsymbol{Q}_B = \boldsymbol{Q} + m\mu\boldsymbol{D}$ $(k = 0)$, donde

$$\boldsymbol{D} = \begin{bmatrix} 0 & 0 & 0 & \cdots\cdots & 0 & 0 \\ 1 & -1 & 0 & \cdots\cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots\cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots\cdots & -1 & 0 \\ 0 & 0 & 0 & \cdots\cdots & 1 & -1 \end{bmatrix}.$$

A partir de las ecuaciones de balance globales obtenemos las siguientes ecuaciones en forma vectorial

$$-\lambda_n \boldsymbol{P}_0^t + \boldsymbol{P}_0^t \boldsymbol{Q}_B + m\mu \boldsymbol{P}_1^t = \boldsymbol{0}^t \tag{3.3}$$

$$\lambda_n \boldsymbol{P}_{k-1}^t - (\lambda_n + m\mu)\boldsymbol{P}_k^t + \boldsymbol{P}_k^t \boldsymbol{Q} + m\mu \boldsymbol{P}_{k+1}^t = \boldsymbol{0}^t, \qquad k > 0 \tag{3.4}$$

donde se ha utilizado que $\lambda p_{m-1,0} = m\mu p_{m,0}$.

Introducimos ahora la función generatriz

$$\boldsymbol{P}^t(z) = \sum_{k \geq 0} \boldsymbol{P}_k^t z^k,$$

y de (3.3) y (3.4) se sigue que

$$\boldsymbol{P}^t(z)\Big(\big(\lambda_n(1-z) + m\mu(1-z^{-1})\big)\boldsymbol{I} - \boldsymbol{Q}\Big) = m\mu \boldsymbol{P}_0^t\Big(\boldsymbol{D} + (1-z^{-1})\boldsymbol{I}\Big).$$

Por tanto,

$$\boldsymbol{P}^t(z) = \lim_{w \to z} m\mu \boldsymbol{P}_0^t \Big(\boldsymbol{D} + (1 - z^{-1})\boldsymbol{I} \Big) \boldsymbol{M}(w)^{-1} \tag{3.5}$$

donde

$$\boldsymbol{M}(z) := \Big(\lambda_n(1-z) + m\mu(1-z^{-1}) \Big) \boldsymbol{I} - \boldsymbol{Q}.$$

En (3.5) se ha introducido el límite ya que existen algunos valores en $|z| \leq 1$ para los que la matriz $\boldsymbol{M}(z)$ es singular y por tanto $\boldsymbol{M}(z)^{-1}$ no está definida. Introduciendo la representación espectral de $\boldsymbol{M}(z)$ y teniendo en cuenta que [Kei79]

$$f(\boldsymbol{Q}) = \sum_{j=1}^{n+1} f(\gamma_j) \boldsymbol{J}_j$$

obtenemos

$$\boldsymbol{P}^t(z) = \lim_{w \to z} \left[m\mu \boldsymbol{P}_0^t \Big(\boldsymbol{D} + (1 - w^{-1})\boldsymbol{I} \Big) \cdot \sum_{j=1}^{n+1} \frac{\boldsymbol{J}_j}{\lambda_n(1-w) + m\mu(1-w^{-1}) - \gamma_j} \right], \tag{3.6}$$

donde γ_j son los autovalores de $\boldsymbol{Q}$, y las matrices $\boldsymbol{J}_j$ se definen como

$$\boldsymbol{J}_j = \frac{\boldsymbol{u}_j \boldsymbol{v}_j^t}{\boldsymbol{v}_j^t \boldsymbol{u}_j} \qquad j = 1, 2, \ldots, n+1$$

siendo $\boldsymbol{u}_j$ y $\boldsymbol{v}_j^t$ los autovectores de $\boldsymbol{Q}$ por la derecha y la izquierda respectivamente. Los autovalores γ_j cumplen la propiedad enunciada en el siguiente teorema.

Teorema 1. *Los autovalores*

$$\gamma_j \,,\, j = 1, \ldots, n+1$$

de la matriz $\boldsymbol{Q}$ son todos reales, distintos y no positivos ($\gamma_j \leq 0$).

Demostración. Sea

$$\boldsymbol{D} = \operatorname{diag} \left\{ 1, \quad \sqrt{\frac{\lambda_h}{\mu}}, \quad \sqrt{\frac{\lambda_h^2}{2\mu^2}}, \quad \ldots, \quad \sqrt{\frac{\lambda_h^n}{n!\mu^n}} \right\},$$

es inmediato comprobar que la matriz

$$\tilde{Q} = DQD^{-1}$$

es simétrica y por lo tanto sus autovalores son reales. Además, $\tilde{Q}$ es tridiagonal y los elementos de la subdiagonal son distintos de cero, por lo que sus autovalores además de reales son todos distintos (véase [Wil78, sección 5.37]). Dado que Q y $\tilde{Q}$ son equivalentes y tienen los mismos autovalores, concluimos que los autovalores de Q son reales y distintos; por tanto queda por demostrar que son no positivos. Según el Teorema de Gershgorin [Wil78, sección 2.13] los autovalores de Q están en el interior de la región

$$\mathbf{\Gamma} := \left\{ \lambda \ : \ |\lambda - q_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n+1} |q_{ij}|, \ i = 1, 2, \ldots, n+1 \right\}$$

por lo que sustituyendo los valores de los elementos de Q, obtenemos

$$\mathbf{\Gamma} := \left\{ \lambda \ : \ |\lambda - q_{ii}| \leq -q_{ii}, \ i = 1, 2, \ldots, n+1 \right\}$$

y, dado que $q_{ii} < 0$ es fácil ver que $\mathbf{\Gamma}$ está contenida en el semiplano $\mathbf{Re}\, z \leq 0$. □

En lo sucesivo consideraremos que los autovalores γ_j están numerados de manera descendente, es decir,

$$0 = \gamma_1 > \gamma_2 > \cdots > \gamma_{n+1}.$$

Por otra parte, dado que Q es un generador infinitesimal se cumple que

$$\boldsymbol{u}_1 = \boldsymbol{e} \quad y \quad \boldsymbol{v}_1 = \boldsymbol{\theta}$$

donde $\boldsymbol{\theta}^t = [\theta_0, \ldots, \theta_n]$ es un vector cuyos elementos son las probabilidades estacionarias del proceso de Markov asociado a Q, esto es

$$\theta_l = \frac{(\lambda_h/\mu)^l}{l!} \left[\sum_{i=0}^{n} \frac{(\lambda_h/\mu)^i}{i!} \right]^{-1}, \qquad l = 0, \ldots, n.$$

Por tanto,

$$\boldsymbol{J}_1 = \boldsymbol{e}\boldsymbol{\theta}^t.$$

Ecuaciones complementarias. La ecuación (3.6) por sí misma no proporciona una expresión para $\boldsymbol{P}(z)$ de la que puedan obtenerse la probabilidades de estado ya que por una parte en la ecuación aparece $\boldsymbol{P}_0^t$ y, por otra, las probabilidades $p_{0,0}, \ldots, p_{m-1,0}$ no están comprendidas en $\boldsymbol{P}(z)$. Sin embargo, estas últimas pueden relacionarse fácilmente con el primer elemento del vector de probabilidades $\boldsymbol{P}_0^t$,

$$p_{r,0} = \frac{m!}{r!} \frac{p_{m,0}}{(\lambda/\mu)^{m-r}} \qquad r = 0, \ldots, m-1. \tag{3.7}$$

En total, por tanto, falta determinar $n+1$ probabilidades para resolver el sistema. Para calcular estas probabilidades se utilizan las condiciones siguientes:

Singularidades de $\boldsymbol{M}(z)$: n ecuaciones

Una condición necesaria para que exista el límite de (3.6) es que cada raíz z ($|z| \leq 1$) de la ecuación

$$\lambda_n(1-z) + m\mu(1-z^{-1}) - \gamma_j = 0, \qquad j = 1, \ldots, n+1 \tag{3.8}$$

sea también una raíz de

$$\boldsymbol{P}_0^t \Big(\boldsymbol{D} + (1-z^{-1})\boldsymbol{I} \Big) \boldsymbol{u}_j = 0, \qquad j = 1, \ldots, n+1. \tag{3.9}$$

Esto es también una condición suficiente en virtud de la proposición 2.

Proposición 2. *Para cada autovalor de $\boldsymbol{Q}$ ($\gamma_j \quad j = 1, \ldots, n+1$) la ecuación (3.8) tiene exactamente una raíz dentro del circulo de radio unidad $|z| \leq 1$, cuya expresión es*

$$z_j = \frac{1}{2\lambda_n} \left(m\mu + \lambda_n - \gamma_j - \sqrt{(m\mu + \lambda_n - \gamma_j)^2 - 4m\mu\lambda_n} \right). \tag{3.10}$$

Demostración. La demostración se divide en dos partes: para $j = 1$ y para $j > 1$.

Si $j = 1$ las raíces de (3.8) son $z_1 = 1$ y $z_2 = m\mu/\lambda_n$. Es obvio que $|z_1| = 1$ y por otra parte, si el sistema es estable $|z_2| = m\mu/\lambda_n > 1$. Por tanto si $j = 1$ (y el sistema es estable) hay exactamente una raíz dentro del círculo unidad.

Ahora consideramos el resto de casos, $j = 2, \ldots, n+1$. Si introducimos las funciones

$$f(z) = -\left(\frac{m\mu}{\lambda_n} + 1 - \frac{\gamma_j}{\lambda_n}\right) z \qquad \text{y} \qquad g(z) = z^2 + \frac{m\mu}{\lambda_n}$$

podemos reescribir la ecuación (3.8) del siguiente modo

$$z^2 - \left(\frac{m\mu}{\lambda_n} + 1 - \frac{\gamma_j}{\lambda_n}\right) z + \frac{m\mu}{\lambda_n} = f(z) + g(z) = 0, \qquad j = 2, \ldots, n+1.$$

Dado que γ_i es negativo (teorema 1) tenemos que

$$|f(z)|_{|z|=1} = \frac{m\mu}{\lambda_n} + 1 - \frac{\gamma_j}{\lambda_n} > \frac{m\mu}{\lambda_n} + 1 \geq \left|z^2 + \frac{m\mu}{\lambda_n}\right|_{|z|=1} = |g(z)|_{|z|=1}$$

y aplicando el *Teorema de Rouche* [CB86] concluimos que $f(z)$ y $f(z) + g(z)$ tienen el mismo número de raíces en $|z| < 1$. Es evidente que $f(z)$ tiene exactamente una raíz ($z = 0$) en $|z| < 1$ por lo que concluimos que $f(z) + g(z)$ tiene exactamente una raíz en $|z| < 1$ y obtener su expresión es inmediato al ser $f(z) + g(z)$ un polinomio de segundo grado. □

Por tanto, para que exista el límite de (3.6) deben cumplirse las igualdades siguientes

$$\boldsymbol{P}_0^t\Big(\boldsymbol{D} + (1 - z_j^{-1})\boldsymbol{I}\Big)\boldsymbol{u}_j = 0 \qquad j = 1, \ldots, n+1$$

donde los valores de z_j están definidos en (3.10). Cuando $j = 1$, $\gamma_1 = 0$, $\boldsymbol{u}_1 = \boldsymbol{e}$, $z_1 = 1$ y la ecuación correspondiente sería $\boldsymbol{P}_0^t\boldsymbol{0} = 0$, que obviamente no proporciona ninguna información acerca de $\boldsymbol{P}_0^t$. Por tanto en total hemos obtenido un total de n ecuaciones ($j = 2, \ldots, n+1$) a partir de las singularidades de $\boldsymbol{M}(z)$.

Ecuación de normalización: 1 ecuación.

Lema 3. *Sea* $P_r(z) := \boldsymbol{P}^t(z)\boldsymbol{e}$ *entonces*

$$P_r(z) = \frac{m\mu}{m\mu - z\lambda_n}\boldsymbol{P}_0^t\boldsymbol{e}$$

Demostración.
De (3.6) y observando que $\boldsymbol{J}_j \cdot \boldsymbol{e} = \boldsymbol{0}$ si $j \neq 1$, $\gamma_1 = 0$ y $\boldsymbol{J}_1 \boldsymbol{e} = \boldsymbol{e}\boldsymbol{\theta}^t \boldsymbol{e} = \boldsymbol{e}$ se sigue que

$$P_r(z) = \boldsymbol{P}^t(z)\boldsymbol{e} = \lim_{w \to z} \frac{m\mu \boldsymbol{P}_0^t \Big(\boldsymbol{D} + (1 - w^{-1})\boldsymbol{I} \Big) \boldsymbol{e}}{\lambda_n(1-w) + m\mu(1-w^{-1})}$$

y observando que $\boldsymbol{D} \cdot \boldsymbol{e} = \boldsymbol{0}$ finalmente obtenemos

$$P_r(z) = \lim_{w \to z} \frac{m\mu(1-w^{-1})}{\lambda_n(1-w) + m\mu(1-w^{-1})} \boldsymbol{P}_0^t \boldsymbol{e} = \frac{m\mu}{m\mu - z\lambda_n} \boldsymbol{P}_0^t \boldsymbol{e}.$$

□

La ecuación de normalización es

$$1 = \sum_{j=0}^{m-1} p_{0,j} + \sum_{j \geq 0} \sum_{i=m}^{m+n} p_{i,j} = \sum_{j=0}^{m-1} p_{0,j} + \sum_{j \geq 0} \boldsymbol{P}_j^t \boldsymbol{e} = \sum_{j=0}^{m-1} p_{0,j} + P_r(1)$$

y aplicando (3.7) y el lema 3 obtenemos

$$1 = \sum_{j=0}^{m-1} \frac{m!}{j!} \Big(\frac{\mu}{\lambda} \Big)^{m-j} p_{m,0} + \frac{\boldsymbol{P}_0^t \boldsymbol{e}}{1 - \lambda_n/(m\mu)} = \boldsymbol{P}_0^t \left[\boldsymbol{e}_1 \sum_{j=0}^{m-1} \frac{m!}{j!} \Big(\frac{\mu}{\lambda} \Big)^{m-j} + \frac{1}{1 - \lambda_n/(m\mu)} \boldsymbol{e} \right]. \quad (3.11)$$

De las n ecuaciones de las singularidades de $\boldsymbol{M}(z)$ y la ecuación de normalización se obtienen los $n+1$ valores del vector $\boldsymbol{P}_0^t$. No hemos demostrado que las $n+1$ ecuaciones lineales son independientes pero esta condición se cumplía en todos lo ejemplos numéricos que se han probado.

Parámetros de interés para la evaluación de prestaciones. A continuación se obtienen las expresiones para el cálculo de parámetros que pueden ser de interés en la evaluación de prestaciones del sistema. Este desarrollo se ha estructurado en proposiciones con una proposición por parámetro.

Proposición 4. *El tiempo medio de espera para los clientes de baja prioridad es*

$$W_n = \frac{m\mu}{(m\mu - \lambda_n)^2}\, \boldsymbol{P}_0^t \boldsymbol{e}$$

Demostración. Para obtener W_n se calcula primero el número medio de clientes de baja prioridad en la cola (L_n) y luego aplicamos la formula de Little [Lit61].

$$L_n = \lim_{z\to 1} \frac{d}{dz} P_r(z) = \frac{m\mu\lambda_n}{(m\mu - \lambda_n)^2}\, \boldsymbol{P}_0^t \boldsymbol{e}$$

y por tanto

$$W_n = \frac{L_n}{\lambda_n} = \frac{m\mu}{(m\mu - \lambda_n)^2}\, \boldsymbol{P}_0^t \boldsymbol{e}. \tag{3.12}$$

□

Como puede observarse, para que existan los valores de L_n y W_n debe cumplirse que $\lambda_n < m\mu$, que es la condición de estabilidad del sistema.

Lema 5.

$$\sum_{j=2}^{n+1} \gamma_j^{-1} \boldsymbol{J}_j = \boldsymbol{e}\boldsymbol{\theta}^t - (\boldsymbol{e}\boldsymbol{\theta}^t - \boldsymbol{Q})^{-1} \tag{3.13}$$

Demostración. Primero demostramos que cuando $i \neq j$

$$\boldsymbol{v}_i^t \boldsymbol{u}_j = 0. \tag{3.14}$$

Por definición $\boldsymbol{Q}\boldsymbol{u}_j = \gamma_j \boldsymbol{u}_j$ y multiplicando a ambas partes del igual por $\boldsymbol{v}_i^t$ obtenemos

$$\begin{aligned} \boldsymbol{v}_i^t \boldsymbol{Q}\boldsymbol{u}_j &= \gamma_j\, \boldsymbol{v}_i^t \boldsymbol{u}_j \\ \gamma_i\, \boldsymbol{v}_i^t \boldsymbol{u}_j &= \gamma_j\, \boldsymbol{v}_i^t \boldsymbol{u}_j \\ (\gamma_i - \gamma_j)\, \boldsymbol{v}_i^t \boldsymbol{u}_j &= 0 \end{aligned}$$

y dado que $\gamma_i \neq \gamma_j$ (teorema 1) debe cumplirse que $\boldsymbol{v}_i^t \boldsymbol{u}_j = 0$.

Por otra parte, utilizando el hecho de que $\gamma_1 = 0$, escribimos

$$\boldsymbol{Q} = \sum_{j=1}^{n+1} \gamma_j \boldsymbol{J}_j = \sum_{j=2}^{n+1} \gamma_j \boldsymbol{J}_j$$

de donde se sigue que

$$\boldsymbol{J}_1 - \boldsymbol{Q} = \boldsymbol{J}_1 - \sum_{j=2}^{n+1} \gamma_j \boldsymbol{J}_j, \tag{3.15}$$

y aplicando (3.14) se obtienen las igualdades siguientes,

$$\begin{aligned} (\boldsymbol{J}_1 - \boldsymbol{Q})\,\boldsymbol{e} &= \boldsymbol{e} & \qquad \boldsymbol{\theta}^t\,(\boldsymbol{J}_1 - \boldsymbol{Q}) &= \boldsymbol{\theta}^t \\ (\boldsymbol{J}_1 - \boldsymbol{Q})\,\boldsymbol{u}_j &= -\gamma_j \boldsymbol{u}_j & \qquad \boldsymbol{v}_j^t\,(\boldsymbol{J}_1 - \boldsymbol{Q}) &= -\gamma_j \boldsymbol{v}_j^t. \end{aligned}$$

Luego $\{1, \gamma_2, \ldots, \gamma_{n+1}\}$ son los autovalores de $\boldsymbol{J}_1 - \boldsymbol{Q}$, y sus autovectores por la derecha (izquierda) son $\{\boldsymbol{e}, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{n+1}\}$ ($\{\boldsymbol{\theta}^t, \boldsymbol{v}_2^t, \ldots, \boldsymbol{v}_{n+1}^t\}$). Por tanto, el término de la izquierda de (3.15) es la expansión espectral de $\boldsymbol{J}_1 - \boldsymbol{Q}$ por lo que podemos expresar su inversa como

$$(\boldsymbol{J}_1 - \boldsymbol{Q})^{-1} = \boldsymbol{J}_1 - \sum_{j=2}^{n+1} \gamma_j^{-1} \boldsymbol{J}_j.$$

Finalmente sustituyendo $\boldsymbol{J}_1 = \boldsymbol{e}\boldsymbol{\theta}^t$ llegamos al resultado deseado. □

Lema 6.

$$\boldsymbol{P}^T(1) = m\mu \boldsymbol{P}_0^t \left(\frac{1}{m\mu - \lambda_n}\, \boldsymbol{e}\boldsymbol{\theta}^t + \boldsymbol{D}\left(\boldsymbol{e}\boldsymbol{\theta}^t - \boldsymbol{Q}\right)^{-1} \right)$$

Demostración. Manipulando (3.6) y recordando que $\gamma_1 = 0$ y $\boldsymbol{J}_1 = \boldsymbol{e}\boldsymbol{\theta}^t$ se obtiene que

$$\boldsymbol{P}^T(1) = \frac{m\mu}{m\mu - \lambda_n}\, \boldsymbol{P}_0^t \boldsymbol{e}\boldsymbol{\theta}^t - m\mu \boldsymbol{P}_0^t \boldsymbol{D} \sum_{j=2}^{n+1} \gamma_j^{-1} \boldsymbol{J}_j,$$

y aplicando el lema 5 llegamos al resultado deseado. □

Proposición 7. *La probabilidad de que haya $m + k$ $(0 \leq k \leq n)$ servidores ocupados cuando un cliente de baja prioridad se coloca en la cola es*

$$P(m+k) = m\mu \boldsymbol{P}_0^t \left(\frac{\theta_k}{m\mu - \lambda_n}\, \boldsymbol{e} + \boldsymbol{D}\left(\boldsymbol{e}\boldsymbol{\theta}^t - \boldsymbol{Q}\right)^{-1} \boldsymbol{e}_{k+1} \right), \tag{3.16}$$

donde $\boldsymbol{e}_i$ $(i = 1, 2, \ldots)$ representa un vector columna cuyos elementos son todos cero salvo el de la i-ésima posición que es uno, y $\theta_k = \boldsymbol{\theta}^t \boldsymbol{e}_{k+1}$.

Demostración. Dado que $P(m+k) = \boldsymbol{P}^T(1)\boldsymbol{e}_{k+1}$ basta con aplicar el lema 6. □

Comentario. El término $(\boldsymbol{e}\boldsymbol{\theta}^t - \boldsymbol{Q})^{-1}\boldsymbol{e}_{k+1}$ de (3.16) puede calcularse como la solución $\boldsymbol{x}$ del sistema lineal $(\boldsymbol{e}\boldsymbol{\theta}^t - \boldsymbol{Q})\boldsymbol{x} = \boldsymbol{e}_{k+1}$, cuyo coste computacional es inferior al de invertir una matriz.

Proposición 8. *La probabilidad de bloqueo P_b de los clientes de alta prioridad es*

$$P_b = P(m+n) = m\mu \boldsymbol{P}_0^t \left(\frac{\theta_n}{m\mu - \lambda_n}\, \boldsymbol{e} + \boldsymbol{D}\left(\boldsymbol{e}\boldsymbol{\theta}^t - \boldsymbol{Q}\right)^{-1} \boldsymbol{e}_{n+1} \right) \tag{3.17}$$

Proposición 9. *El tiempo medio de permanencia en estado de congestión*[3] *es*

$$E[T_n^c] = \frac{1}{m\mu - \lambda_n}\, \frac{\boldsymbol{P}_0^t \boldsymbol{e}}{\boldsymbol{P}_0^t \boldsymbol{e}_1} \tag{3.18}$$

Demostración. El sistema está en congestión cuando el proceso está en alguno de los estados $\{(r,k) \;:\; r \geq 0, m \leq k \leq m+n\}$. En un intervalo $[0,T]$ el tiempo que el sistema estará en congestión es $P_r(1)T + O(T)$ y el número de veces que en este intervalo el sistema entra en estado de congestión es $\lambda p_{m-1,0} T + O(T)$, por lo que el tiempo que dura en media cada periodo de congestión será

$$\frac{P_r(1)T + O(T)}{\lambda p_{m-1,0} T + O(T)}.$$

Tomando límites cuando $T \to \infty$, sustituyendo $P_r(1)$ por la expresión del lema 3 y utilizando (3.7) para sustituir $p_{m-1,0}$, se obtiene el resultado deseado. □

[3] Llamamos congestión a la situación en la que los clientes de baja prioridad tienen que esperar tras su llegada para recibir servicio.

Algoritmo HOPS

Para el análisis de este algoritmo redefinimos algunas variables de la siguiente manera

$$\begin{aligned} \boldsymbol{P}_r^t &= [p_{r,0}, p_{r,1}, \ldots, p_{r,n}], \\ \boldsymbol{P}^t(z) &= \sum_{r \geq m} \boldsymbol{P}_r^t z^r, \end{aligned}$$

$$\boldsymbol{Q}_B = \begin{bmatrix} * & 0 & 0 & \cdots\cdots & 0 & 0 \\ \mu & * & 0 & \cdots\cdots & 0 & 0 \\ 0 & 2\mu & * & \cdots\cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots\cdots & * & 0 \\ 0 & 0 & 0 & \cdots\cdots & n\mu & * \end{bmatrix},$$

mientras que $\boldsymbol{Q}$ mantiene la definición de (3.2). Así, de las ecuaciones de balance globales obtenemos

$$\begin{aligned} -\lambda \boldsymbol{P}_0^t + \boldsymbol{P}_0^t \boldsymbol{Q}_B + \mu \boldsymbol{P}_1^t &= \boldsymbol{0}^t && && (3.19) \\ \lambda \boldsymbol{P}_{r-1}^t - (\lambda + r\mu)\boldsymbol{P}_r^t + \boldsymbol{P}_r^t \boldsymbol{Q}_B + (r+1)\mu \boldsymbol{P}_{r+1}^t &= \boldsymbol{0}^t, && 0 < r < m && (3.20) \\ \lambda \boldsymbol{P}_{m-1}^t - (\lambda_n + m\mu)\boldsymbol{P}_m^t + \boldsymbol{P}_m^t \boldsymbol{Q} + m\mu \boldsymbol{P}_{m+1}^t &= \boldsymbol{0}^t && && \\ \lambda_n \boldsymbol{P}_{r-1}^t - (\lambda_n + m\mu)\boldsymbol{P}_r^t + \boldsymbol{P}_r^t \boldsymbol{Q} + m\mu \boldsymbol{P}_{r+1}^t &= \boldsymbol{0}^t, && m < r && \end{aligned}$$

y aplicando el mismo procedimiento que antes derivaríamos

$$\boldsymbol{P}^t(z) = \lim_{w \to z} \left[\left(\lambda \boldsymbol{P}_{m-1}^t - m\mu w^{-1} \boldsymbol{P}_m^t \right) \cdot \sum_{j=1}^{n+1} \frac{\boldsymbol{J}_j}{\lambda_n (1-w) + m\mu(1-w^{-1}) - \gamma_j} \right]. \tag{3.21}$$

Ecuaciones complementarias. La ecuaciones vectoriales de (3.19) y (3.20) correspondientes a los niveles frontera aportan un total de $m(n+1)$ ecuaciones escalares.

Siguiendo un procedimiento análogo al utilizado en los lemas 3 y 6 obtendríamos, respectivamente,

$$P_r(z) = \frac{m\mu}{m\mu - \lambda_n z}\boldsymbol{P}^t{}_m\boldsymbol{e}$$
$$\boldsymbol{P}^T(1) = \frac{m\mu}{m\mu - \lambda_n}\,\boldsymbol{P}^t{}_m\boldsymbol{e}\boldsymbol{\theta}^t + (\lambda\boldsymbol{P}^t{}_{m-1} - m\mu\boldsymbol{P}^t{}_m)(\boldsymbol{e}\boldsymbol{\theta}^t - \boldsymbol{Q})^{-1}.$$

La ecuación de normalización tomaría la forma siguiente

$$1 = \sum_{j=0}^{m-1}\boldsymbol{P}^t_j\boldsymbol{e} + \frac{m\mu}{m\mu - \lambda_n}\boldsymbol{P}^t{}_m\boldsymbol{e}. \tag{3.22}$$

Imponer la existencia del límite de (3.21) cuando $|z| \leq 1$ aporta el resto de ecuaciones necesarias, que son

$$\left(\lambda\boldsymbol{P}^t{}_{m-1} - m\mu z_j^{-1}\,\boldsymbol{P}^t{}_m\right)\boldsymbol{u}_j = 0 \qquad j = 2,\ldots,n+1$$

donde los z_j son los definidos en (3.10).

De forma totalmente paralela a las demostraciones de las proposiciones 4, 7, 8 y 9 se obtendrían los resultados siguientes

$$W_n = \frac{m\mu}{(m\mu - \lambda_n)^2}\,\boldsymbol{P}^t{}_m\boldsymbol{e}$$
$$P(m+k) = \frac{m\mu}{m\mu - \lambda_n}\,\boldsymbol{P}^t{}_m\boldsymbol{e}\theta_k + (\lambda\boldsymbol{P}^t{}_{m-1} - m\mu\boldsymbol{P}^t{}_m)(\boldsymbol{e}\boldsymbol{\theta}^t - \boldsymbol{Q})^{-1}\boldsymbol{e}_{k+1}$$
$$P_b = P(m+n)$$
$$E[T_n^c] = \frac{m\mu}{\lambda(m\mu - \lambda_n)}\,\frac{\boldsymbol{P}^t{}_m\boldsymbol{e}}{\boldsymbol{P}^t{}_{m-1}\boldsymbol{e}}.$$

Las expresiones para estos mismos parámetros en el algoritmo HOPSWR son (3.12), (3.16), (3.17) y (3.18).

Algoritmo HOSP

Para el análisis del algoritmo HOSP mantenemos la misma notación y definiciones que para el algoritmo HOPS. Las ecuaciones de balance por niveles

para el nivel de frontera $r = m$ y para los niveles homogéneos ($r > m$) son

$$\lambda_h(\boldsymbol{P}^t_{m-1}\boldsymbol{e}_{n+1})\boldsymbol{e}^t_{n+1} + \lambda_n\boldsymbol{P}^t_{m-1} - (\lambda_n + m\mu)\boldsymbol{P}^t_m + \boldsymbol{P}^t_m\boldsymbol{Q} + m\mu\boldsymbol{P}^t_{m+1} = \boldsymbol{0}^t$$
$$(r > m) \qquad \lambda_n\boldsymbol{P}^t_{r-1} - (\lambda_n + m\mu)\boldsymbol{P}^t_r + \boldsymbol{P}^t_r\boldsymbol{Q} + m\mu\boldsymbol{P}^t_{r+1} = \boldsymbol{0}^t$$

de donde se obtiene que

$$\boldsymbol{P}^t(z) = \lim_{w\to z}\left[\left(\lambda_h(\boldsymbol{P}^t_{m-1}\boldsymbol{e}_{n+1})\boldsymbol{e}^t_{n+1} + \lambda_n\boldsymbol{P}^t_{m-1} - m\mu w^{-1}\boldsymbol{P}^t_m\right)\cdot \right.$$
$$\left. \cdot \sum_{j=1}^{n+1}\frac{\boldsymbol{J}_j}{\lambda_n(1-w) + m\mu(1-w^{-1}) - \gamma_j}\right]. \quad (3.23)$$

El resto de ecuaciones que corresponden a los niveles no homogéneos son

$$\lambda_h\boldsymbol{P}^t_{m-1}\boldsymbol{e}_{n+1} + \lambda_n\boldsymbol{P}^t_{r-1}\boldsymbol{e} - r\mu\boldsymbol{P}^t_r\boldsymbol{e} = 0 \qquad r = 1, 2, \ldots m,$$

que suponen un total de $m(n+1)$ ecuaciones escalares. Siguiendo un procedimiento análogo al utilizado en los lemas 3 y 6 obtendríamos, respectivamente

$$P_r(z) = \frac{m\mu\boldsymbol{P}^t_m\boldsymbol{e}}{m\mu - \lambda_n z}$$
$$\boldsymbol{P}^T(1) = \frac{m\mu\boldsymbol{P}^t_m\boldsymbol{e}\boldsymbol{\theta}^t}{m\mu - \lambda_n} + \left(\lambda_h(\boldsymbol{P}^t_{m-1}\boldsymbol{e}_{n+1})\boldsymbol{e}^t_{n+1} + \lambda_n\boldsymbol{P}^t_{m-1} - m\mu\boldsymbol{P}^t_m\right)(\boldsymbol{e}\boldsymbol{\theta}^t - \boldsymbol{Q})^{-1}.$$

La ecuación de normalización resulta ser la misma que la del algoritmo HOPS (3.22). Y al igual que antes, exigir la existencia del límite que aparece en (3.23) cuando $|z| \leq 1$, proporciona el resto de ecuaciones necesarias

$$\left(\lambda_h(\boldsymbol{P}^t_{m-1}\boldsymbol{e}_{n+1})\boldsymbol{e}^t_{n+1} + \lambda_n\boldsymbol{P}^t_{m-1} - m\mu w^{-1}\boldsymbol{P}^t_m\right)\cdot\boldsymbol{u}_j = 0, \qquad j = 2, \ldots, n+1$$

donde las z_j son las definidas en (3.10).

De forma totalmente paralela a las demostraciones de las proposiciones 4,

7, 8 y 9 se obtendrían los resultados siguientes

$$W_n = \frac{m\mu}{(m\mu - \lambda_n)^2}\, \boldsymbol{P}_m^t \boldsymbol{e}$$

$$P(m+k) = \frac{m\mu \boldsymbol{P}_m^t \boldsymbol{e}\theta_k}{m\mu - \lambda_n}\left(\lambda_h(\boldsymbol{P}_{m-1}^t \boldsymbol{e}_{n+1})\boldsymbol{e}_{n+1}^t + \lambda_n \boldsymbol{P}_{m-1}^t - m\mu \boldsymbol{P}_m^t\right)(\boldsymbol{e}\boldsymbol{\theta}^t - \boldsymbol{Q})^{-1}\boldsymbol{e}_{k+1}$$

$$P_b = P(m+n)$$

$$E[T_n^c] = \frac{m\mu}{m\mu - \lambda_n}\, \frac{\boldsymbol{P}_m^t \boldsymbol{e}}{\boldsymbol{P}_{m-1}^t(\lambda_h \boldsymbol{e}_{n+1} + \lambda_n \boldsymbol{e})}$$

La expresiones para estos mismos parámetros en el algoritmo HOPSWR son (3.12), (3.16), (3.17) y (3.18).

3.2.3. Evaluación numérica

Para cuantificar las ventajas que el método espectral ofrece en términos de eficiencia computacional y precisión se ha realizado una evaluación numérica de los algoritmos HOPSWR, HOPS y HOSP utilizando el método espectral y el geométrico-matricial, y ambos se comparan observando la precisión y el coste computacional.

En resultados que se presentan en esta sección la tasa de llegadas de clientes de alta prioridad (handovers) se ha calculado considerado que en la célula existe un equilibrio estadístico entre el flujo entrante y saliente por lo que se verifica que [Jab96]

$$\lambda_h = \frac{\lambda_n}{\mu_c/\mu_r + P_b}. \tag{3.24}$$

Por otra parte P_b depende a su vez de λ_h por lo que (3.24) es una ecuación no lineal de la que puede obtenerse λ_h aplicando un método iterativo de punto fijo como el descrito en [HR86, LMN94]. Para los parámetros correspondientes a la duración de la llamada (μ_c) y el tiempo de residencia en una célula (μ_r) se han utilizado los valores $\mu_c = 2/3\ s^{-1}$ y $\mu_r = 1/3\ s^{-1}$.

Precisión

Los resultados obtenidos corroboran que, tal y como se sostiene en [Mit95], el método geométrico matricial aplicado al análisis de nuestros algoritmos, presenta problemas de precisión cuando la carga del sistema se aproxima a la zona de inestabilidad, mientras que no ocurre así con el método espectral.

Las figuras 3.16–3.18 muestran la estimación del error relativo[4] de los parámetros P_b y W_n cuando se calculan con el método geométrico matricial. Para estimar el error relativo se ha considerado que el método espectral devuelve el valor exacto. Para validar esta suposición se ha comparado el valor del parámetro W_n en carga alta obtenido mediante el método espectral con el del tiempo medio de espera del modelo Erlang-C $M/M/m$ (recuérdese que m el número máximo de canales en el grupo primario). En esta comparación el error relativo del método espectral —tomando el del módelo Erlang-C como exacto— ha resultado ser lo suficientemente bajo (inferior a 10^{-3}) para poder considerarlo como exacto para nuestro propósito, mientras que el error relativo del método geométrico-matricial es significativo. Los resultados mostrados en estas figuras se han calculado para un sistema de parámetros $m = 8$, $n = 2$.

Los resultados muestran una ventaja importante del método espectral con respecto al geométrico-matricial. Sin embargo, aunque esta superioridad podría ser de interés en algunas aplicaciones, sus consecuencia prácticas en el contexto del CA en sistemas celulares son más bien limitadas pues la superioridad del método espectral únicamente es apreciable en situación de congestión severa.

Eficiencia computacional

Para comparar la complejidad computacional de ambos métodos se ha medido el número de operaciones en coma flotante que emplea cada método para cada algoritmo y para distintos tamaños del sistema.

[4] Si x es el valor exacto y $\hat{x}$ un valor aproximado, su error relativo se calcula como $|(\hat{x} - x)/x|$.

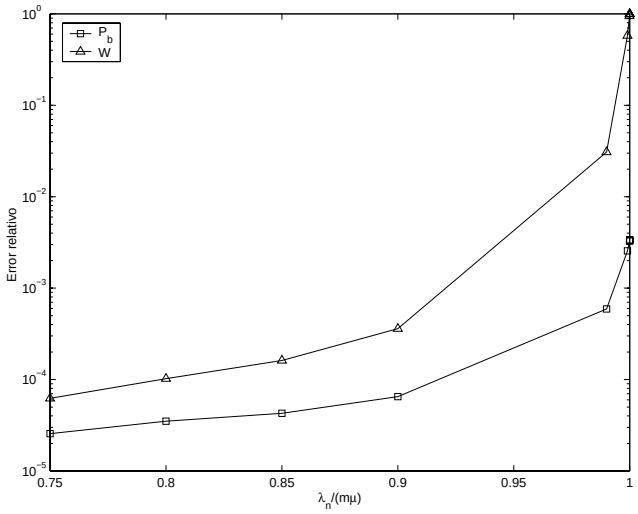


Figura 3.16: HOPSWR: error relativo de los parámetros de QoS

En ambos métodos, la resolución numérica del sistema consta de dos partes:

1. Niveles no homogéneos. En ambos métodos, las probabilidades correspondientes a los estados de estos niveles se obtienen, salvo normalización, resolviendo el sistema lineal de ecuaciones que proporcionan las ecuaciones de equilibrio correspondientes a los niveles en los que el sistema no está congestionado.

2. Niveles homogéneos. En esta parte es en la que difieren ambos métodos: el método geométrico-matricial se basa el cálculo de la *matriz de tasas* $\boldsymbol{R}$, que se obtiene como solución a una ecuación matrical cuadrática ($\boldsymbol{R}^2\boldsymbol{A}_2 + \boldsymbol{R}\boldsymbol{A}_1 + \boldsymbol{A}_0 = \overline{\overline{\boldsymbol{0}}}$), y posteriormente las probabilidades de estado se obtienen a partir de las potencias de la matriz $\boldsymbol{R}$; en el método espectral, la función generatriz $\boldsymbol{P}^t(z)$ se expresa en función de los autovalores y autovectores de la matriz $\boldsymbol{Q}$, ecuación (3.6).

En la tabla 3.1 se muestra el tamaño del problema numérico que hay que

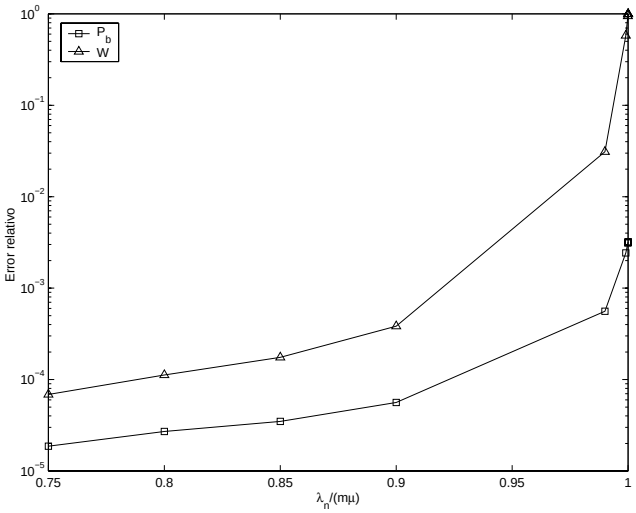


Figura 3.17: HOPS: error relativo de los parámetros de QoS

Tabla 3.1: Tamaño del problema para cada parte.

Algoritmo	**Parte 1**	**Parte 2**
GC, HOPSWR	$O(m+2n)$	$O(n)$
HOPS, HOSP	$O(m \cdot n)$	$O(n)$

resolver en cada una de las partes. Cabe destacar que el tamaño del problema es el mismo para ambos métodos y también que en ambos métodos la parte primera del problema se resuelve de la misma forma, por lo que la diferencia entre los dos métodos radica en la eficiencia númerica con la que se resuelve la segunda parte. En las figuras de la 3.19 a la 3.21 se representa el coste computacional en función de la cantidad de recursos $C = m + n$, siendo $m = 0.8C$ y $n = 0.2C$, y tomando una carga $\lambda_n = 0.5m\mu$. Y en las figuras de la 3.22 a la 3.24 se representa el coste computacional relativo del método geométrico-matricial con respecto al espectral. De la observación de las gráficas y la tabla 3.1, pueden extraerse las conclusiones siguientes: cuando el tamaño de ambas partes es del mismo orden (valores de C bajos), el coste

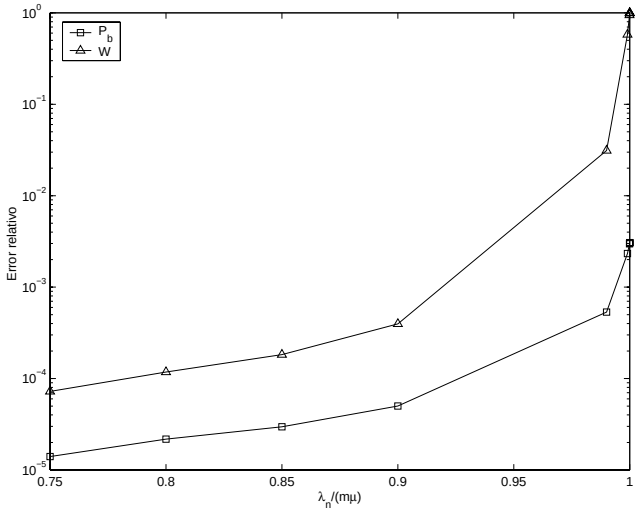


Figura 3.18: HOSP: error relativo de los parámetros de QoS

computacional de la parte 1 (resolver un sistema lineal) es inferior al de la parte 2 (calcular autovalores y autovectores o resolver una ecuación matricial cuadrática), por lo que el peso específico de la parte 2 es superior y el método espectral resulta ser más eficiente. Sin embargo, para los algoritmos HOPS y HOSP el tamaño de la parte 1 crece de forma más rápida que el de la parte 2 —crecimiento cuadrático frente a lineal— por lo que llega un punto que el coste computacional de la parte 1 predomina sobre el de la parte 2, por lo que en estos algoritmos cuando C crece el coste computacional de ambos métodos tiende a igualarse.

3.3. Conclusiones

En este capítulo hemos analizado una familia de mecanismos de control de admisión para sistemas con dos tipos de tráfico de distinta prioridad, como pueden ser las redes celulares monoservicio. Esta familia de algorit-

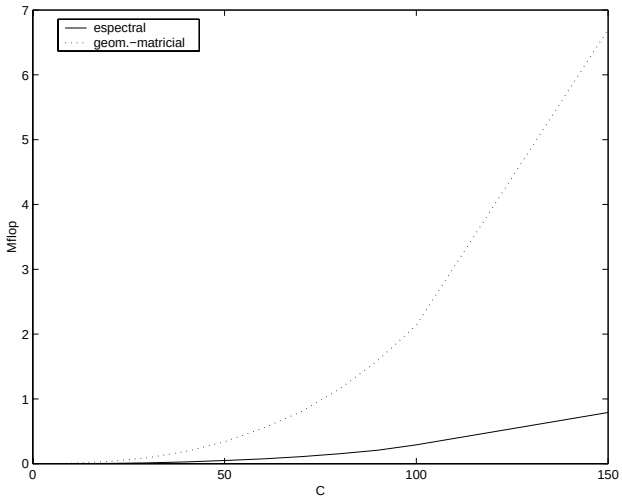


Figura 3.19: HOPSWR: coste computacional.

mos de asignación de canales se basa en separar los canales disponibles en dos tipos: los que pueden ser utilizados por cualquier tipo de petición y los reservados que sólo pueden ser utilizados por las peticiones de alta prioridad (handovers). Ambos flujos de tráfico son tratados según un modelo de espera con cola finita y, en el modelo, se contempla el abandono por impaciencia para ambos tipos de petición. En todos los algoritmos el parámetro que establece la cantidad de canales reservados puede variar de forma continua entre cero y el total de canales disponibles.

El análisis de los algoritmos se basa en la metodología geométrico-matricial. Además se desarrolla también un análisis utilizando una técnica más reciente, que está basada en la utilización de la función generatriz del proceso subyacente junto con técnicas espectrales matriciales. Esta técnica alternativa presenta la ventaja de una mayor estabilidad numérica cuando la carga del sistema es elevada y, en algunos casos, una menor carga computacional.

La especificación y el análisis de estos algoritmos supone una unificación y una extensión de varios estudios y propuestas aparecidos en la literatura

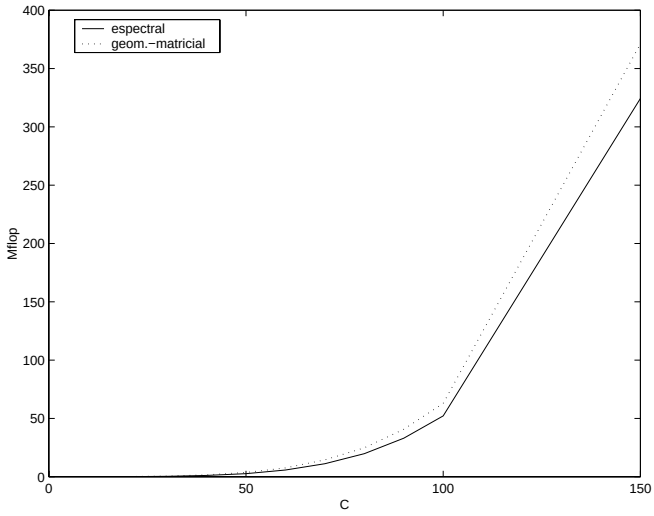


Figura 3.20: HOPS: coste computacional.

especializada.

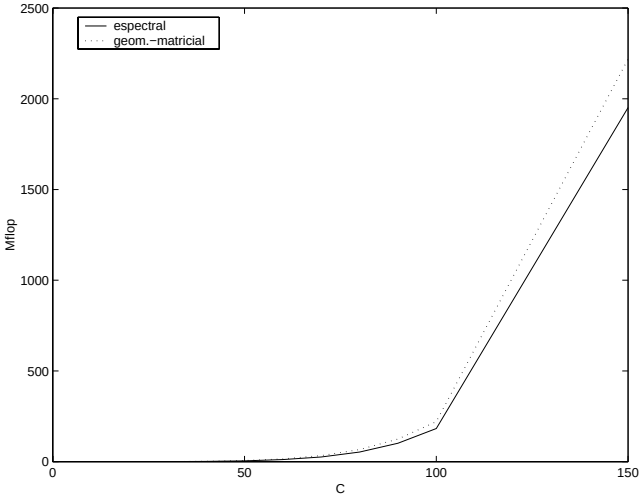


Figura 3.21: HOSP: coste computacional.

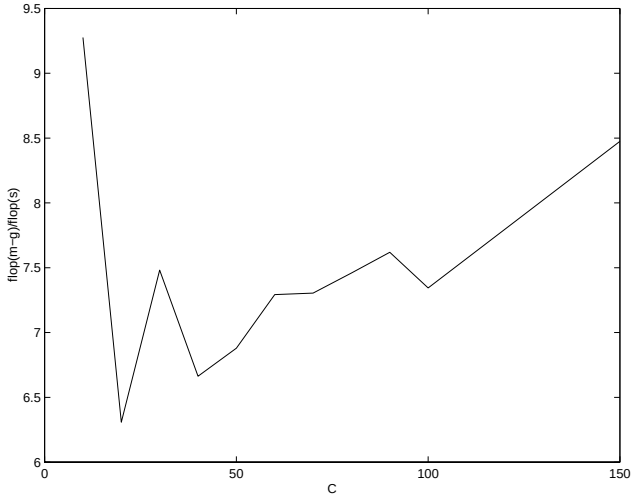


Figura 3.22: HOPSWR: coste computacional relativo.

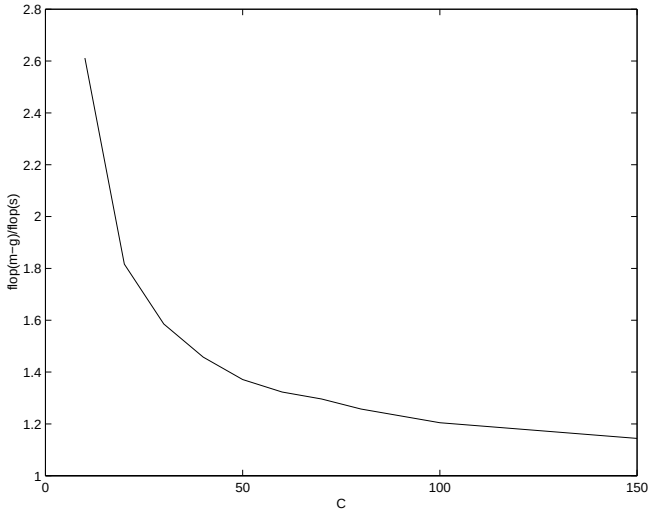


Figura 3.23: HOPS: coste computacional relativo.

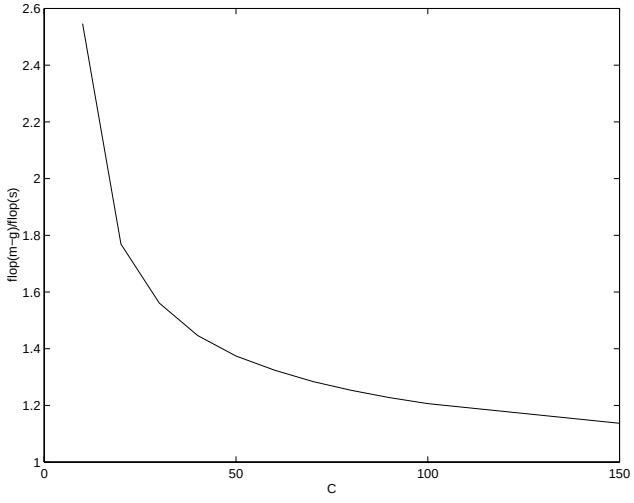


Figura 3.24: HOSP: coste computacional relativo.

Capítulo 4

Área de *handover* y clientes impacientes

Con el fin de asegurar la continuidad de la comunicación, normalmente existe un solape entre las zonas de cobertura de células vecinas. La existencia de esta zona en la que un terminal recibe potencia suficiente de varias estaciones base, hace posible que en el curso del cambio de célula exista un cierto margen temporal para atender la petición. Este margen temporal abarca desde que se detecta que comienza a recibirse potencia suficiente de la nueva célula y se solicita el handover, hasta que el móvil deja de recibir suficiente potencia de la estación base a la que está conectado. Si al final de este intervalo todavía no se ha producido el handover —no se han asignado recursos en la nueva célula— la sesión deberá terminar de manera forzosa. A esta zona, dentro de la cual puede producirse el handover sin que tenga que abortarse la sesión en curso, se la denomina *área de handover* y al tiempo que un terminal permanece en ella *intervalo de degradación* [HR86, TJ91]. El hecho de poder demorar durante un cierto tiempo la ejecución del handover si éste no puede llevarse a cabo en el momento en que se solicita ha sido aprovechado en un buen número propuestas para otorgar prioridad a las peticiones de handover y hacer una gestión más eficiente de los recursos (véase [HR86, TJ91, TJ92, KN96, TRV98, PCG02b, LJCP03, XT04] y sus referencias). En los modelos, esta técnica se refleja incorporando una cola

para las peticiones de handover, bien porque efectivamente existe una implementación de esta cola, o bien porque las peticiones de handover que no pueden ser atendidas en un primer momento lo reintentan de forma continuada mientras están en el área de handover [Bar04]. En cualquiera de los dos casos los clientes —las peticiones de handover— pueden esperar en la cola un tiempo máximo que se corresponde con el intervalo de degradación.

Este tipo de cola en la que los clientes tienen un tiempo máximo de espera y, trascurrido este tiempo la abandonan, se conoce con el nombre de *cola con clientes impacientes* [BdW94]. Aparte de la aplicación que acabamos de referir, las colas con clientes impacientes tienen numerosas aplicaciones tanto en el campo de las TIC como en otras áreas. De entre estas aplicaciones merece mención especial, por el interés que ha despertado en los últimos años, la de los centros de atención de llamadas (*call centers*) [Man04].

La gran mayoría de los modelos que contemplan una cola para las peticiones de handover —mayoría que se convierte en la práctica totalidad si se trata de un modelo analítico— describen el tiempo de residencia en el área de handover (HART) con una variable aleatoria exponencial. Sin embargo, la validez de esta suposición no ha sido suficientemente estudiada, por lo que es pertinente plantearse: primero, qué distribución o distribuciones de probabilidad caracterizan adecuadamente el HART; y segundo, en el caso de que la respuesta a la pregunta anterior resulte ser que la distribución exponencial no es buen modelo, para cada aplicación, y en particular para el modelado del CA, habrá que evaluar en qué medida afecta a los resultados el utilizar una distribución exponencial.

Por otra parte, la existencia de clientes impacientes cuyo tiempo de impaciencia no sigue una distribución exponencial hace que la disciplina de servicio influya en las prestaciones del sistema, por lo que algunos autores han planteado la utilización de otras disciplinas de servicio [TJ92, ET99, DRFG99a, DRFG99b, Fan00, XT04]. Además, en el caso de que no exista realmente una cola, sino que ésta resulta del reintento continuado de los terminales que están en el área de handover, dicha cola virtual se serviría con una

disciplina que podría aproximarse por una disciplina de servicio en orden aleatorio (SIRO) [Bar04].

En la sección 4.1 de este capítulo se expone un método analítico-numérico para la caracterización estadística del HART. Posteriormente se aplica este método a un tipo concreto de escenario y se estudia la distribución resultante y la aproximación de ésta por algunas distribuciones conocidas. En la sección 4.2 se desarrollan sendos modelos analíticos, uno exacto y otro aproximado, para la evaluación de un mecanismo de CA en una red celular cuando el HART sigue un distribución cualquiera. Estos modelos se utilizan para evaluar cómo influye, en los parámetros de interés, utilizar una aproximación exponencial para la distribución del HART. En la sección 4.3 se desarrolla un modelo analítico para un sistema multiservidor con cola finita, clientes impacientes cuyo tiempo de espera máximo sigue una distribución del tipo PH, y tres disciplinas de servicio distintas: FIFO, LIFO y SIRO. Finalmente en 4.4 se hace un resumen del capítulo y se presentan las conclusiones.

4.1. Caracterización estadística del tiempo de permanencia y de ocupación de recursos en el área de handover

Aunque la cantidad de trabajos que tratan de algún modo el estudio del tiempo de permanencia y/o de ocupación de recursos en el área de handover, es más bien reducido, pueden encontrarse algunos precedentes en la literatura [PSS96, KS97, KS99, DRFG99b, RGS98, MEBC02]. En [PSS96, KS97] los autores proponen un modelo analítico para el tiempo de ocupación de recursos en el área de handover. Este modelo es para un escenario de tipo Manhattan en el que el movimiento se produce únicamente en dos direcciones (horizontal y vertical) y la velocidad de los terminales sigue una distribución uniforme. En estos trabajos, sin embargo, no encontramos resultados numéricos ni una caracterización del tipo de distribución resultante de aplicar el

modelo propuesto. Kim y Sung [KS99] utilizan una distribución general para el tiempo residual de permanencia en el área de handover con el fin de obtener algunos parámetros de importancia para un sistema con *soft-handover*. Sin embargo, esta distribución general es únicamente un dato de entrada para el desarrollo posterior y no se da ninguna indicación acerca de sus características. Del Re et al. [DRFG99b] elaboran un modelo que permite obtener la distribución del HART para un sistema móvil que utiliza satélites de órbita baja (LEO). En este escenario el movimiento relativo del terminal respecto al satélite está provocado fundamentalmente por el movimiento del satélite y es, por tanto, de naturaleza determinista. En [MEBC02], se utiliza un modelo de simulación para caracterizar el tiempo transcurrido entre la petición del handover y la ejecución de éste, o la terminación forzosa en caso de que no llegue a ejecutarse. Este retardo se correspondería con el tiempo de ocupación de recursos mientras el terminal está en el área de handover. El artículo de Ruggieri et al. [RGS98] es tal vez el que más se aproxima, en objetivos y en metodología, al estudio que presentamos en esta sección. Lo más destacado de este artículo es que incluye el *shadow fading* del canal radio en el modelo. En este trabajo se concluye que el HART se ajusta a una distribución gaussiana truncada. Para fijar los límites del área de handover el modelo de [RGS98] únicamente considera la potencia que se recibe desde la célula de origen. Sin embargo, de entre los múltiples métodos empleados para decidir el inicio del proceso de handover, y por tanto la entrada en el área de handover, prácticamente todos consideran no sólo la potencia recibida de la estación base actual sino también la potencia recibida desde la célula de destino [Pol96]. Otro trabajo que desde un punto de vista metodológico guarda relación con el contenido de esta sección es [CKS+98], donde propone un método numérico para obtener el CRT a partir de la distribución de la distancia recorrida por el móvil en el interior de la célula y de la distribución de la velocidad.

Aquí proponemos un modelo para obtener una caracterización estadística del HART y del tiempo de ocupación de recursos en el área de handover. En nuestro modelo tomamos dos células (la de origen y la de destino) y únicamente se consideran las pérdidas de potencia debido a la distancia ya que

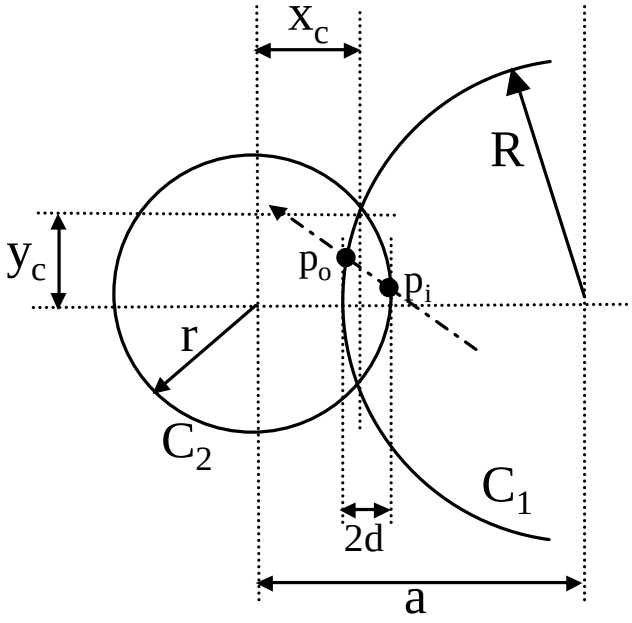


Figura 4.1: Diagrama general.

suponemos que las fluctuaciones rápidas se eliminan mediante un promediado temporal [Pol96]. Primero, mediante un procedimiento analítico-numérico obtenemos un muestreo de la distribución de la distancia recorrida dentro del área de handover y de aquí, aplicando un método similar al de [CKS+98], obtenemos la distribución del HART. También se derivan las expresiones que permiten obtener la distribución del tiempo de ocupación de recursos a partir de la distribución del HART. En ambos casos —HART y tiempo de ocupación de recursos— se evalúa el ajuste que se consigue mediante distribuciones conocidas.

4.1.1. Descripción del modelo y la metodología

En un principio nos limitaremos al caso en el que las células tienen forma circular y posteriormente se describe como podría adaptarse el modelo para aplicarlo a geometrías más generales que pudieran aparecer en escenarios reales. Consideremos una situación de handover en la que un terminal se desplaza desde la célula (círculo) de origen C1 hacía la célula de destino C2, cuyos centros están a una distancia a (véase la fig. 4.1). Cuando el terminal está a una distancia r de C2 se produce la petición de handover y la conexión con C1 podrá mantenerse mientras la distancia a su centro sea inferior a R; generalmente se cumplirá que $R \geq r$. Si suponemos que no se producen cambios de dirección mientras el móvil está dentro del área de handover, la trayectoria que sigue el terminal es el segmento comprendido entre p_i y p_o.

Aunque la distancia recorrida dentro del área de handover está completamente determinada mediante p_i y p_o introducimos las variables φ y θ (véase la figura 4.2) para simplificar la notación del desarrollo posterior. En las figuras 4.1 y 4.2 se define el significado geométrico de la variables x_c, y_c, d, φ_{max}, θ_{min} y θ_{max}, cuyas expresiones son

$$\begin{aligned} x_c &= \frac{a^2 + r^2 - R^2}{2a} \\ y_c &= \sqrt{r^2 - x_c^2} \\ \varphi_{max} &= \arctan\left(\frac{y_c}{x_c}\right) \qquad (4.1) \\ \theta_{min}(\varphi) &= \varphi + \arctan\left(\frac{r\cos\varphi - x_c}{r\sin\varphi + y_c}\right) \qquad (4.2) \\ \theta_{max}(\varphi) &= \varphi + \pi - \arctan\left(\frac{r\cos\varphi - x_c}{y_c - r\sin\varphi}\right). \qquad (4.3) \end{aligned}$$

Por otra parte, la superficie del área de handover vale

$$S_{ha} = \arctan\left(\frac{y_c}{x_c}\right) r^2 + \arctan\left(\frac{y_c}{a - x_c}\right) R^2 - a y_c. \qquad (4.4)$$

Para unos valores determinados de R, r y a el punto p_i está definido por el ángulo φ y, fijado este último, el ángulo θ define el punto p_o. Por tanto,

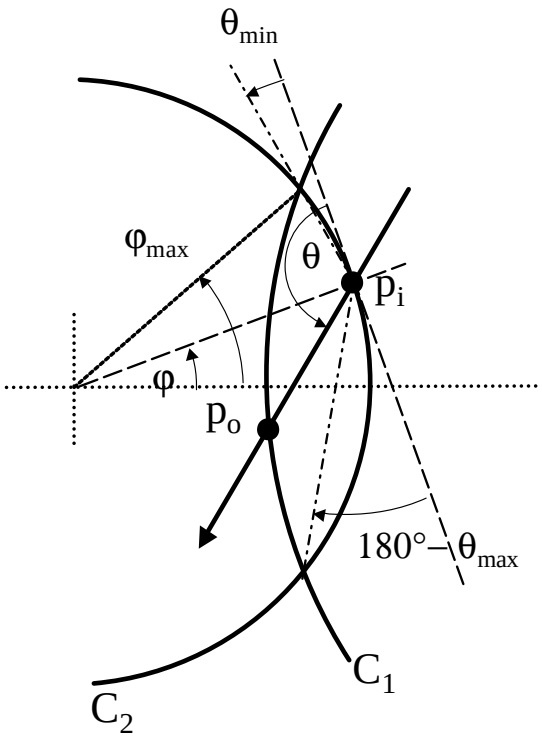


Figura 4.2: Diagrama de ángulos y dominio de φ.

fijados R, r y a, la distancia entre p_i y p_o puede expresarse como una función de los dos ángulos φ y θ, como

$$Z = dist(p_i, p_o) = f(\varphi, \theta).$$

Una vez se tiene el valor de la distancia recorrida Z, el tiempo de permanencia se calcula como

$$T_d = \frac{Z}{V},$$

donde V es la velocidad media del terminal en el área de handover.

Para obtener la distribución de T_d primero calculamos la distribución de

Z y, a partir de ésta, se calcula la distribución de T_d mediante una versión ligeramente modificada del método empleado en [CKS+98].

Distribución de la distancia recorrida

Supongamos que las distribuciones de los ángulos son conocidas y sus funciones de distribución son

$$F_\varphi(\varphi) \qquad y \qquad F_\theta^\varphi(\theta).$$

Obsérvese que la distribución de θ depende del valor de φ. Sea

$$I_\varphi = \{\varphi_1, \ldots, \varphi_n\}$$

un conjunto de n valores para la variable aleatoria φ convenientemente distribuidos y, de idéntico modo, definimos

$$I_\theta(\varphi_i) = \{\theta_1^{\varphi_i}, \ldots, \theta_m^{\varphi_i}\}.$$

Los valores en estos conjuntos pueden obtenerse de la forma siguiente

$$I_\varphi = F_\varphi^{-1}(U) \qquad y \qquad I_\theta(\varphi) = \left(F_\theta^\varphi\right)^{-1}(U),$$

donde U representa un conjunto — del tamaño apropiado — cuyos elementos están uniformemente distribuidos en el intervalo $[0, 1]$, por ejemplo si N es el tamaño de U, entonces:

$$U = \left\{0, \frac{1}{N}, \ldots, \frac{N-1}{N}\right\}.$$

Para cada pareja de valores $(\varphi_i, \theta_j^{\varphi_i})$ obtenemos la distancia correspondiente $z_{ij} = f(\varphi_i, \theta_j^{\varphi_i})$ y formamos el conjunto

$$I_Z = \left\{z_{ij} \setminus i = 1, \ldots, n;\ j = 1, \ldots, m\right\}.$$

Como resultado, si n y m son lo suficientemente altos, el conjunto I_Z será una muestra representativa de la variable aleatoria Z y, por tanto, a partir de los valores en I_Z obtenemos un muestreo discreto de las funciones de distribución y densidad de la variable Z, $F_Z(z)$ y $f_Z(z)$.

Distribución del tiempo de permanencia

El procedimiento anteriormente descrito da como resultado una versión numérica de $f_Z(z)$. Por otra parte, dado que suponemos que la distribución de la velocidad media es conocida, se dispone también de $f_V(v)$. Utilizando que $T_d = Z/V$, mediante integración numérica calculamos $f_{dt}(t)$ [CKS+98] evaluando

$$f_{dt}(t) = \int_{v_{min}}^{\min(v_{max}, \frac{2y_c}{t})} v f_Z(vt) f_V(v) \mathrm{d}v, \tag{4.5}$$

o bien

$$f_{dt}(t) = \frac{1}{t^2} \int_{tv_{min}}^{\min(tv_{max}, 2y_c)} z f_Z(z) f_V\left(\frac{z}{t}\right) \mathrm{d}z. \tag{4.6}$$

Desde un punto de vista analítico las dos expresiones anteriores son equivalentes, aunque en su evaluación numérica puede haber diferencias. En [CKS+98] los autores argumentan que (4.6) es preferible a (4.5). Aquí utilizamos una expresión u otra dependiendo del valor de t: los valores bajos de t resultan en un muestreo más fino de $f_V(vt)$ en (4.5) que el muestreo que resultaría de $f_V(z/t)$ en (4.6), y viceversa.

Otras geometrías

Si las células no tienen una forma circular, la geometría del área de handover ya no será la intersección de dos círculos. En este caso el método que hemos descrito puede modificarse de la siguiente manera (véase como ejemplo la figura 4.3):

1. Tomar círculos del radio adecuado para que el área de handover quede completamente contenida dentro de la intersección de los círculos.

2. Realizar el barrido de φ y θ de la forma descrita y, calcular la distancia recorrida considerando el área de handover auténtica, es decir, tomar z_1 en vez de z (figura 4.3).

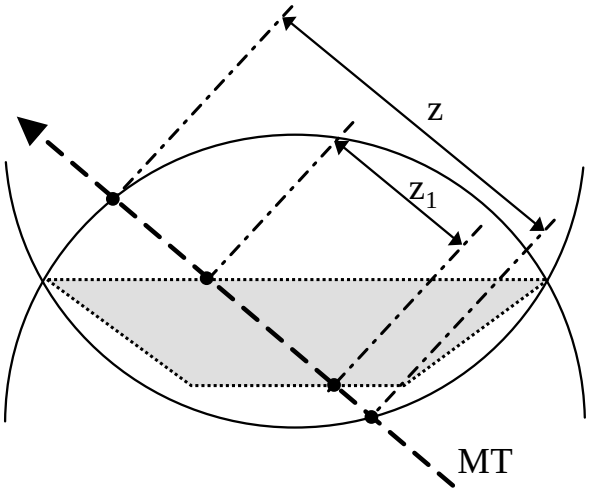


Figura 4.3: Geometría del área de handover irregular.

4.1.2. Evaluación numérica

El método antes descrito es válido para cualesquiera distribuciones de φ, θ y V. Aquí vamos a realizamos un estudio numérico para unas distribuciones particulares que se derivan de un determinado modelo de movilidad.

Distribución de φ. El modelo de movilidad que empleamos asume que los terminales se mueven de manera independiente unos de otros, que están uniformemente distribuidos por toda la zona de servicio y que cualquier dirección de movimiento tiene la misma probabilidad. En consecuencia, cualquier punto de la frontera de una célula tiene la misma probabilidad de ser atravesado por un terminal, por lo que la variable aleatoria φ seguirá una distribución uniforme en el intervalo $[-\varphi_{max}, \varphi_{max}]$. Además, dada la simetría existente entre $[-\varphi_{max}, 0]$ y $[0, \varphi_{max}]$ podemos realizar el análisis utilizando únicamente uno de los dos intervalos.

Distribución de θ. En un principio, siguiendo el mismo razonamiento que para φ llegaríamos a la conclusión de que θ también sigue una distribución uniforme. Sin embargo, debemos tener en cuenta que la variable aleatoria θ —al igual que ocurrirá con la velocidad media V— está aplicada a un terminal del que sabemos que está cruzando la frontera entre células por lo que es más preciso aplicar la transformación del *muestreo sesgado* (*biased sampling*) [XG93, ZD97]:

$$f_\theta(\theta) = \begin{cases} \frac{1}{2}\sin(\theta), & 0 \leq \theta \leq \pi \\ 0, & \text{en otro caso} \end{cases}$$

Además, truncamos esta distribución de modo que $\theta \in [\theta_{min}, \theta_{max}]$. De esta forma se excluyen las trayectorias que atravesarían el área de handover para después volver a la célula de origen. Por tanto, finalmente tenemos que

$$f_\theta(\theta) = \begin{cases} \frac{1}{\cos\theta_{min} - \cos\theta_{max}}\sin(\theta), & \theta_{min} \leq \theta \leq \theta_{max} \\ 0, & \text{en otro caso} \end{cases}$$

Distribución de V. Al igual que en [TJ92, ZD97, RGS98] supondremos que la velocidad media de un terminal cualquiera sigue una distribución gaussiana (en realidad una gaussiana truncada para evitar las velocidades negativas), y para los terminales que están cruzando las fronteras aplicamos la transformación del *muestreo sesgado* [XG93, ZD97]

$$f_{V^\star}(v) = \frac{v f_V(v)}{\overline{v}}. \tag{4.7}$$

Obsérvese la similitud que existe entre (4.7) —y el razonamiento que llevaría hasta esta expresión— y su correspondiente en la idea de la *densidad de los intervalos seleccionados* en [Kle75, sección 5.2].

La figura 4.4 muestra la función de densidad de probabilidad $f_Z(z)$ para distintos valores de d y valores de los radios $R = r = 1$ km. Los valores utilizados para el parámetro d (50 m, 100 m y 200 m) se corresponden con superficies del área de handover que representan, respectivamente, un 8 %, 22 % y 50 % de la superficie total de cobertura de una célula.

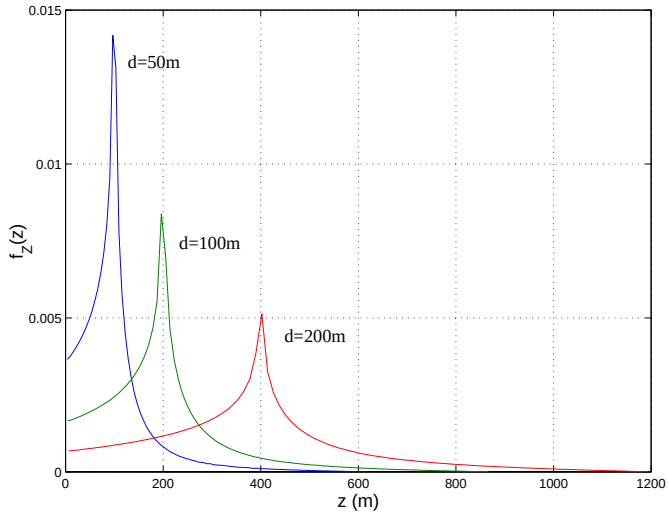


Figura 4.4: Densidad de probabilidad de la distancia recorrida, $R = r = 1$ km.

A partir de $f_Z(z)$ y $f_{V^\star}(v)$, calculamos la distribución del HART $f_{dt}(t)$ mediante la evaluación numérica de (4.6) y (4.5). La figura 4.5(a) muestra los resultados obtenidos cuando $\overline{v} = 50$ km/h y $\sigma_v = 10$ km/h (estos estadísticos están referidos a la distribución para cualquier terminal). En la figura 4.5(b) se ha fijado el valor $d = 100$ m y se muestra el efecto de variar el coeficiente de variación (CV) de V.

4.1.3. Ajuste de la distribución del tiempo de permanencia en el área de handover

La aplicación del método descrito en 4.1.1 proporciona una versión numérica y discreta de la función de densidad de probabilidad del HART, ahora se examina el ajuste de estos valores numéricos mediante algunas distribuciones conocidas y se evalúa la bondad de dicho ajuste. Las distribuciones que se consideran son: *exponencial, exponencial doble, Erlang-k, Erlang-jk, hiper-*

Tabla 4.1: Distribuciones candidatas. $t, \alpha, \beta, \gamma, \beta_1, \beta_2, \delta > 0$ and $0 \leq p \leq 1$.

Distribución	**pdf**
Exponencial	$f(t) = \frac{1}{\beta} e^{-\frac{t}{\beta}}$
Erlang-k	$f(t) = \frac{t^{k-1}}{\beta^k \ (k-1)!} e^{-\frac{t}{\beta}}$
Erlang-jk	$f(t) = p \frac{t^{j-1}}{\beta^j \ (j-1)!} e^{-\frac{t}{\beta}} + (1-p) \frac{t^{k-1}}{\beta^k \ (k-1)!} e^{-\frac{t}{\beta}}$
Hiper-Erlang-jk	$f(t) = p \frac{t^{j-1}}{\beta_1^j \ (j-1)!} e^{-\frac{t}{\beta_1}} + (1-p) \frac{t^{k-1}}{\beta_2^k \ (k-1)!} e^{-\frac{t}{\beta_2}}$
Exponencial doble	$f(t) = \frac{1}{\beta \ (2 - e^{-\frac{d}{\beta}})} e^{-\frac{\lvert t-\delta \rvert}{\beta}}$
Gamma generalizada	$f(t) = \frac{\gamma}{\beta^{\alpha\gamma}\Gamma(\alpha)} \ t^{\alpha\gamma - 1} \ e^{-\left(\frac{t}{\beta}\right)^{\gamma}}$

Erlang-jk y *gamma generalizada*. La expresión de la función de densidad para estas distribuciones se muestra en la tabla 4.1.

La distribución exponencial se emplea comúnmente por la manejabilidad que aporta su *memoria nula*, y su adecuación o no para describir el HART es uno de los aspectos que se pretende validar. Las distribuciones del tipo Erlang —en general las de tipo *phase type* (PH)— aunque no tienen memoria nula pueden verse como una composición de distintas etapas con memoria nula en cada una de ellas, por lo que su tratamiento analítico continúa siendo factible en algunos casos. Las distribuciones de tipo Erlang que se consideran son casos particulares de la distribución hiper-Erlang[1], que se ha considerado como una buena aproximación para aspectos relacionados con la movilidad en redes inalámbricas (véase [Fan01] y sus referencias). La distribución gamma generalizada ha demostrado ser un buen ajuste para el tiempo de permanencia en una célula [ZD97] y, además, incluye como casos particulares a las distribuciones *gamma, lognormal* y *weibull*, que también aparecen con cierta frecuencia como modelo para variables aleatorias de tipo temporal. Por último la elección de la distribución *exponencial doble* como candidata está

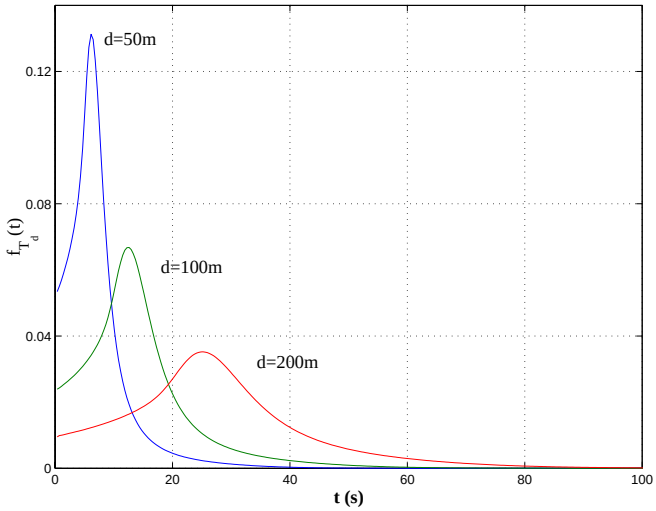
[1] Empleamos el término *hiper-Erlang* para referirnos a una combinación convexa de un número cualquiera de distribuciones Erlang, y *hiper-Erlang-jk* para referirnos a una combinación convexa de dos distribuciones Erlang de órdenes j y k.

Tabla 4.2: Bondad del ajuste (G). $R = r = 1$ km, $d = 100$ m, $\overline{v} = 50$ km/h, $\sigma_v = 10$ km/h

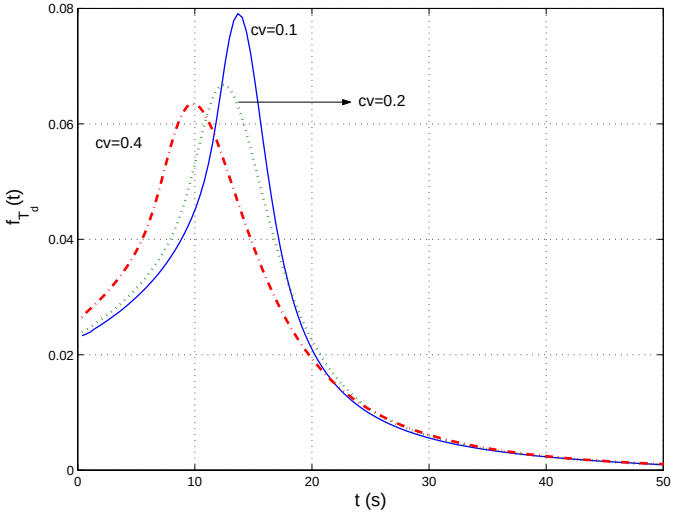
Distribución	***G***	**Parámetros**
Hiper-Erlang-8,1	**0.0120**	β_1=1.7665, β_2=14.128, p=0.55701
Exponencial doble	0.0166	β = 8.1887, δ = 11.296
Gamma generalizada	0.0285	α = 10.491, β = 0.10836, γ = 0.49206
Erlang-4,1	0.0302	β = 3.7669, p=0.87587
Erlang-3	0.0331	β = 4.6647
Exponencial	0.1113	β = 11.647

motivada por la forma de la curva de $f_{dt}(t)$. En la distribuciones Erlang-jk e hiper-Erlang-jk uno de los parámetros discreto k (o j) se fijará manualmente a $k = 1$ ya que $f_{dt}(0) > 0$.

El grado de semejanza entre la distribución original y las obtenidas del ajuste puede evaluarse visualmente en las figuras 4.6 y 4.7. Los parámetros de las distribuciones ajustadoras se muestran en la tabla 4.2. El procedimiento para el cálculo de estos parámetros se comenta más adelante. En la figura 4.7 se representan gráficos de *curvas de probabilidad* que se utilizan también para evaluar de una forma visual la semejanza entre distribuciones de probabilidad. En estos gráficos se ha representado la función de la distribución ajustadora respecto a la original: si $\widehat{F}(t)$ es la función de distribución ajustadora y F_{dt} la función de distribución original, $(x, y) = (p, \widehat{F}(F_{T_d}^{-1}(p)))$, $p \in [0, 1[$ es una representación paramétrica de las curvas de probabilidad. La calidad del ajuste será mayor cuanto más próximas estén las curvas de probabilidad a la línea recta cuya representación paramétrica es (p, p).

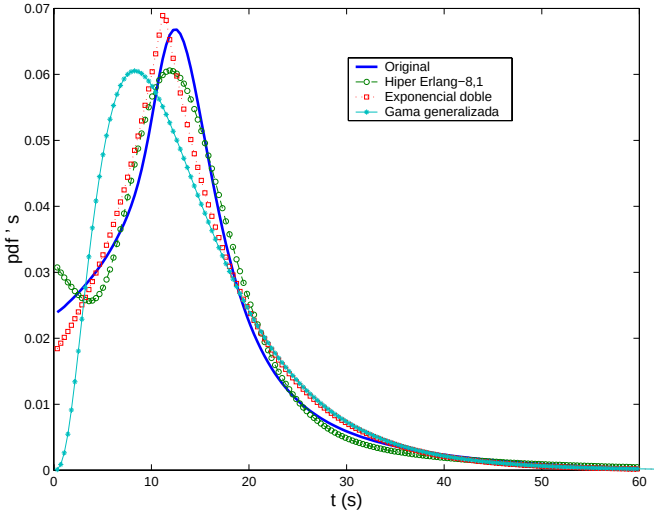


(a) Influencia de la anchura del área de handover.

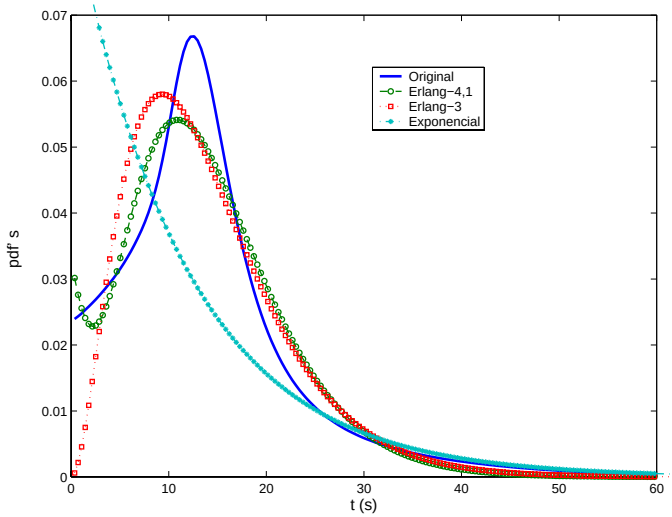


(b) Influencia de la dispersión en la velocidad media.

Figura 4.5: Densidad de probabilidad del HART, $R = r = 1$ km,$d = 100$ m, $\overline{v} = 50$ km/h.

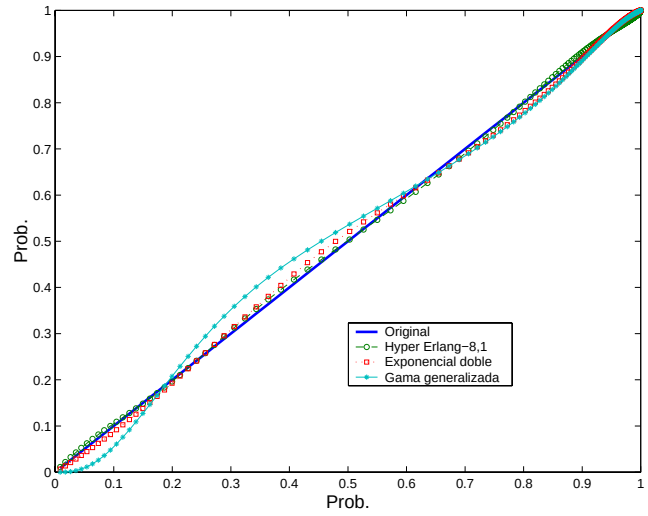


(a)

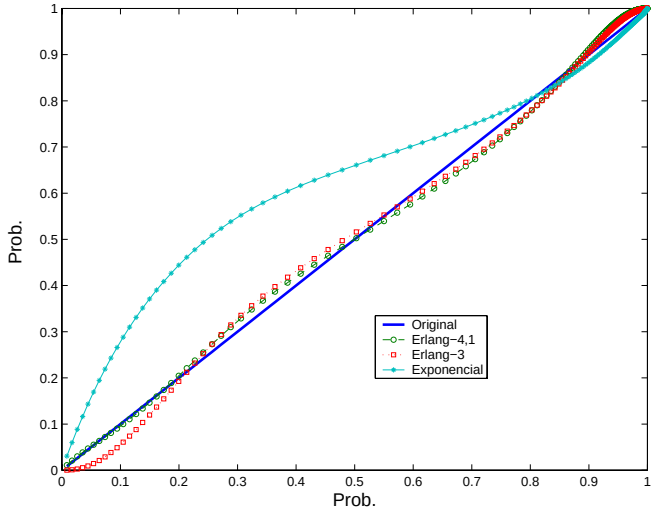


(b)

Figura 4.6: Ajuste de la distribución del HART mediante algunas distribuciones conocidad, $R = r = 1$ km, $d = 100$ m, $\overline{v} = 50$ km/h, $\sigma_v = 10$ km/h.



(a)



(b)

Figura 4.7: Gráficos de probabilidad de los ajustes. $R = r = 1$ km, $d = 100$ m, $\overline{v} = 50$ km/h, $\sigma_v = 10$ km/h.

Tabla 4.3: Bondad de del ajuste (G), ($\overline{v} = 50$ km/h, $R = r = 1$ km)

	$\sigma_v = 10$ km/h			$d = 100$ m	
Distribución	$d = 100$ m	$d = 50$ m	$d = 200$ m	$\sigma_v = 5$ km/h	$\sigma_v = 20$ km/h
Hiper-Erlang-j,1	**0.0042**	**0.0120**	0.0117	**0.0166**	**0.0100**
Exponencial doble	0.0288	0.0166	**0.0072**	0.0218	0.0249
Gamma generalizada	0.0377	0.0285	0.0211	0.0351	0.0288
Erlang-j,1	0.0423	0.0302	0.0178	0.0331	0.0447
Erlang-k	0.0462	0.0331	0.0241	0.0376	0.0396
Exponencial	0.0927	0.1113	0.1434	0.1229	0.0813

De la observación de las curvas se concluye que la distribuciones *hiper-Erlang-j,1* y *exponencial doble* ofrecen un ajuste más preciso mientras que el ajuste de la *exponencial* es demasiado holgado. Otros escenarios que no se han representado conducen a la misma conclusión.

Se ha utilizado también el indicador cuantitativo G ([HR86, RGS98]),

$$G = \frac{\int_0^\infty |F^c_{T_d}(t) - \widehat{F}^c(t)|\mathrm{d}t}{2\int_0^\infty F^c_{T_d}(t)\mathrm{d}t}$$

para evaluar la bondad del ajuste. Un valor de $G = 0$ indica un ajuste completamente exacto y $G = 1$ indica que no hay ninguna correlación [HR86]. El valor de G se ha utilizado también para calcular los parámetros de las distribuciones ajustadoras. Primero se calcula un valor inicial de estos parámetros mediante ajuste de momentos y el valor calculado se utiliza para inicializar un algoritmo numérico de minimización de G.

En la tabla 4.2 se muestran el valor de los parámetros y el valor de G para el ejemplo de la figura 4.6. Los valores de la tabla 4.3 cuantifican el efecto de variar la anchura del área de handover d y la varianza de la velocidad V. Los valores de G que se han obtenido confirman las conclusiones que se habían obtenido de la inspección visual de las gráficas. En particular, la distribución *hiper-Erlang-j,1* obtiene el mejor resultado (valor en negrita) en todos los casos salvo en uno ($d = 200$ m), e incluso en este caso el ajuste de la *hiper-Erlang-j,1* podría considerarse aceptable ($G = 0.0117$). La exponencial doble ocuparía el segundo lugar y la exponencial siempre el último.

4.1.4. Tiempo de ocupación de recursos en el area de handover

Aquí derivamos la distribución del tiempo de ocupación de recursos T_{cht} para el caso en que la llegada de peticiones de nuevas sigue un proceso de Poisson y la duración de una sesión está distribuida exponencialmente. Para ello distinguimos dos casos dependiendo de si la sesión se inicia antes de que

el terminal penetre en la zona de handover o lo hace una vez dentro. Primero se realiza el análisis para cada uno de estos casos por separado y después se calcula la probabilidad de cada una de estas situaciones.

Sesiones iniciadas fuera del área de handover

Si T_c denota el tiempo residual de una sesión cuando el terminal entra en el área de handover, se cumplirá que

$$T_{cht} = \text{mín}\{T_{dt}, T_c\}$$

y por tanto

$$\begin{aligned} F^o_{cht}(t) &= 1 - P(T_{cht} > t) \\ &= 1 - P(T_{dt} > t)P(T_c > t) \\ &= 1 - (1 - F_{dt}(t))\,(1 - F_c(t))\,. \end{aligned} \tag{4.8}$$

Dado que suponemos que la duración de una sesión está distribuido exponencialmente, y esta distribución tiene memoria nula, sabemos que T_c sigue exactamente la misma distribución, $F_c(t) = 1 - e^{-\mu_c t}$. Luego

$$F^o_{cht}(t) = 1 - (1 - F_{dt}(t))\, e^{-\mu_c t} \tag{4.9}$$

y de aquí obtenemos que

$$f^o_{cht}(t) = [\mu_c\,(1 - F_{dt}(t)) + f_{dt}(t)]\, e^{-\mu_c t}. \tag{4.10}$$

Sesiones iniciadas dentro del área de handover

Introduzcamos la variable aleatoria $\widehat{Z}$ que representa la distancia entre el punto en el que está el terminal cuando se inicia la sesión —que estará en el interior del área de handover— y el punto por el que el terminal sale del área de handover p_o. La variable aleatoria $\widehat{Z}$ puede verse como la vida residual de Z observada en el momento que comienza la sesión y, como suponemos

que la llegada de sesiones nuevas sigue un proceso de Poisson, se cumplirá que [Kle75, Eq. 5.10]

$$f_{\widehat{Z}}(z) = \frac{1 - F_Z(z)}{E[Z]} \tag{4.11}$$

Si reemplazamos $f_Z(z)$ por $f_{\widehat{Z}}(z)$ en (4.5) y (4.6) se obtiene la función de densidad del tiempo de permanencia en el área de handover contado desde el momento en el que se inicia la sesión $f_{\widehat{dt}}(t)$) y procediendo del mismo modo que para obtener (4.10), llegamos a

$$f^i_{cht}(t) = \left[\mu_c \left(1 - F_{\widehat{dt}}(t)\right) + f_{\widehat{dt}}(t)\right] e^{-\mu_c t} \tag{4.12}$$

En este caso $f_{\widehat{dt}}$ y $F_{\widehat{dt}}$ deben obtenerse utilizando la distribución general de la velocidad y no la versión del muestreo sesgado, pues aquí se están considerando los terminales que inician la sesión dentro del área de handover y, para un cierto terminal, la probabilidad de que esto ocurra es independiente de su velocidad.

La función de densidad descondicionada la podemos escribir como una suma ponderada de $f^o_{cht}(t)$ y $f^i_{cht}(t)$, en la que los pesos son las probabilidades de que el terminal haya iniciado la sesión fuera (P_o) o dentro (P_i) del área de handover:

$$f_{cht}(t) = P_i f^i_{cht}(t) + P_o f^o_{cht}(t). \tag{4.13}$$

A continuación se obtienen las expresiones para las probabilidades P_i y P_o. Sean λ_n y λ_h, respectivamente, las tasas cursadas de peticiones nuevas y de handover en una célula. De igual modo, definimos λ^{ha}_n y λ^{ha}_h como las tasas cursadas en el área de handover. Si consideremos que los terminales están espacialmente distribuidos de manera uniforme se cumplirá que

$$\lambda^{ha}_n = \frac{S_{ha}}{S_{cell}} \lambda_n \tag{4.14}$$

$$\lambda^{ha}_h = 2\frac{2\varphi_{max}}{2\pi}\lambda_h = \frac{2\varphi_{max}}{\pi}\lambda_h, \tag{4.15}$$

donde φ_{max} puede calcularse según la expresión (4.1) (véanse también las figuras 4.1 y 4.2), S_{cell} es la superficie de la célula y S_{ha} la superficie del área

de handover, que puede calcularse mediante (4.4). Luego,

$$P_i = \frac{\lambda_n^{ha}}{\lambda_n^{ha} + \lambda_h^{ha}} = \frac{3S_{ha}\lambda_n}{3S_{ha}\lambda_n + S_{cell}\lambda_h} \tag{4.16}$$

$$P_o = \frac{\lambda_h^{ha}}{\lambda_n^{ha} + \lambda_h^{ha}} = \frac{S_{cell}\lambda_h}{3S_{ha}\lambda_n + S_{cell}\lambda_h}. \tag{4.17}$$

Es importante destacar que puesto que λ_n y λ_h representan tasas cursadas su valor dependerá de las probalidades de bloqueo correspondientes que, a su vez, dependen del esquema de CAC (véase por ejemplo [Jab96]).

4.1.5. Ajuste de la distribución del tiempo de ocupación de recursos en el área de handover

Procediendo de una forma totalmente análoga a la de la sección 4.1.3 ajustamos la distribución del tiempo de ocupación de recursos en el área de handover mediante las distribuciones de la tabla (4.1) y evaluamos la bondad del ajuste, utilizando las mismas configuraciones que en la sección 4.1.3.

La tabla 4.4 muestra unos valores representativos de los resultados que evalúan la bondad del ajuste mediante el indicador G. En las figuras 4.8, 4.9 y 4.10 se han representado las curvas correspondientes a los tres mejores ajustes de las columnas primera, tercera y quinta de la tabla 4.4. De forma general puede decirse que para las distribuciones que se han probado el ajuste mejora cuando aumenta el valor de P_i. Por otra parte, la distribución *hiper-Erlang-j,1* continúa siendo la mejor ajustadora y la *exponencial* la peor. Sin embargo, la *exponencial doble* ya no ocupa de forma tan clara la segunda posición sino que rivalizaría con la *gamma generalizada* y la *Erlang-j,1*.

Tabla 4.4: Bondad del ajuste (G). Influencia de $P_i = 1 - P_o$. $\overline{v} = 50$ km/h, $R = r = 1$ km, $\mu_c^{-1} = 100$ s.

Distribution	$P_i = 0.1$	$P_i = 0.3$	$P_i = 0.5$	$P_i = 0.7$	$P_i = 0.9$
Hiper-Erlang-j,1	**0.0077**	**0.0063**	**0.0050**	**0.0039**	**0.0027**
Exponencial doble	0.0219	0.0238	0.0234	0.0203	0.0156
Gamma generalizada	0.0284	0.0261	0.0223	0.0178	0.0129
Erlang-j,1	0.0245	0.0217	0.0192	0.0166	0.0145
Erlang-k	0.0307	0.0261	0.0271	0.0357	0.0388
Exponencial	0.1016	0.0872	0.0719	0.0557	0.0388

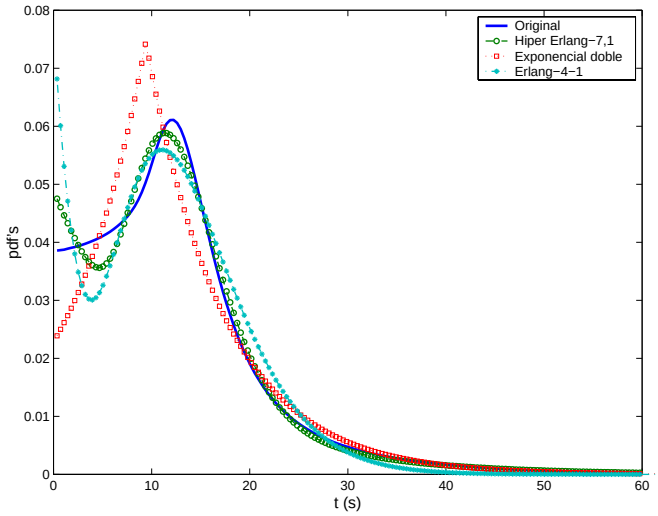


Figura 4.8: $P_i = 1 - P_o = 0.1$; $R = r = 1$ km, $d = 100$ m, $\overline{v} = 50$ km/h, $\sigma_v = 10$ km/h, $\mu_c^{-1} = 100$ s.

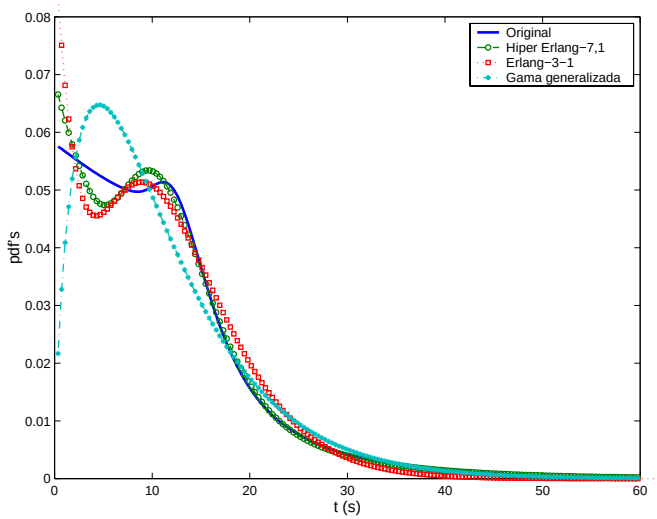


Figura 4.9: $P_i = 1 - P_o = 0.5$; $R = r = 1$ km, $d = 100$ m, $\overline{v} = 50$ km/h, $\sigma_v = 10$ km/h, $\mu_c^{-1} = 100$ s.

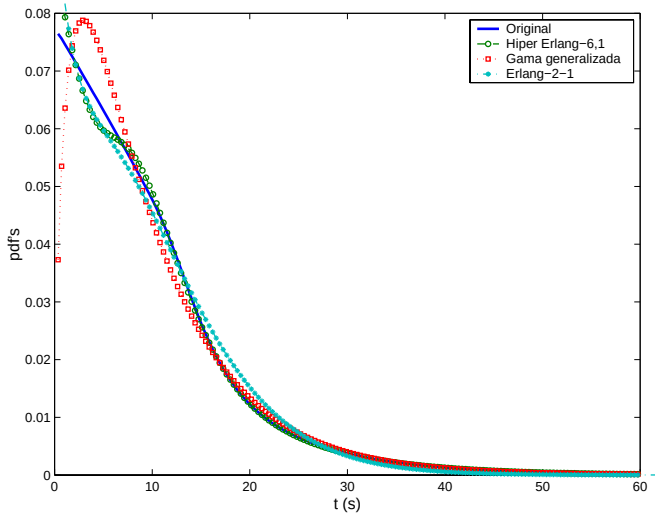


Figura 4.10: $P_i = 1 - P_o = 0.9$; $R = r = 1$ km, $d = 100$ m, $\overline{v} = 50$ km/h, $\sigma_v = 10$ km/h, $\mu_c^{-1} = 100$ s.

4.2. Distribución del tiempo de permanencia en área de handover y prestaciones del CA

Los resultados de la sección anterior junto con los precedentes encontrados en la literatura [RGS98, MEBC02] y la ausencia de resultados que avalen la hipótesis contraria, hacen que la suposición habitual de que el tiempo de permanencia en el área de solape (HART) siga una distribución exponencial sea por lo menos cuestionable. Esto no quiere decir que, de entrada, esta hipótesis no sea válida en ningún caso, sino que para cada aplicación, y en particular para el modelado del CAC, habrá que evaluar en qué medida afecta a los resultados el utilizar una aproximación exponencial.

Esta misma cuestión se ha planteado para los tiempos de residencia (CRT) y de ocupación de recursos (CHT) en toda la célula. Distintos estudios ([JL96, ZD97, BJ00, HSSK01, HSSK02] e indirectamente también [Mac05]) proponen distribuciones distintas de la exponencial para describir los tiempos CRT y CHT. A la vista de estos resultados otros estudios [KZ97, LC97, HSSK02, XT03] han evaluado el error que introduce en los parámetros de interés aproximar el CRT o el CHT mediante variables aleatorias exponenciales, o han desarrollado modelos para la evaluación de prestaciones en los que se relaja esta suposición [JGA01, AL02].

El trabajo que presentamos a continuación pretende ser una contribución en la segunda fase del estudio del modelado estadístico del HART. Primero se desarrolla un modelo analítico de un mecanismo de CAC en una red celular cuando el HART sigue un distribución cualquiera y, después, se presenta un modelo analítico aproximado cuya evaluación numérica es más eficiente que la del modelo exacto. Con estos modelos se evalúa el impacto de utilizar una aproximación exponencial para la distribución del HART sobre los parámetros de interés. También se comparan el coste computacional y la precisión del modelo aproximado con los del modelo exacto.

4.2.1. Descripción del modelo

El modelo considera una célula que dispone de una cantidad total de recursos N y cuyos flujos de entrada y salida están en equilibrio estadístico. A la célula llegan peticiones de sesiones nuevas y de handover con unas tasas λ_n y λ_h, respectivamente. Ambos flujos de llegada siguen un proceso de Poisson. La duración de una sesión y el tiempo de permanencia en la célula se describen mediante variables aleatorias exponenciales de parámetros μ_c y μ_r, respectivamente, por lo que el tiempo de ocupación de recursos seguirá también una distribución exponencial de parámetro $\mu = \mu_c + \mu_r$. Para priorizar las peticiones de handover se emplea un CAC del tipo *Guard Channel* (GC) con espera para las peticiones de handover bloqueadas y con pérdida para las sesiones nuevas [PG85, HR86]. La cantidad de recursos reservada exclusivamente para peticiones de handover es n. Las sesiones nuevas sólo se admiten cuando la cantidad de recursos libres supera las n unidades y se rechazan en caso contrario. Las peticiones de handover se aceptan siempre que haya suficientes recursos libres y en caso contrario esperan su turno en la cola correspondiente que es atendida siguiendo una disciplina FIFO según se van liberando los recursos necesarios. Una petición de handover puede permanecer en la cola mientras el terminal esté en el área de handover, si el terminal sale del área de handover antes de que se haya liberado una cantidad suficiente de recursos para cursar la petición, ésta abandonará la cola sin recibir servicio y se producirá una terminación forzosa de la sesión. El tiempo de permanencia en el área de solape se describe mediante la variable aleatoria T_{ha} de media μ_{ha}^{-1} y cuya función de densidad $f_{ha}(t)$ es una función cualquiera.

4.2.2. Análisis

Aquí se analiza el modelo del sistema y se obtienen expresiones analíticas para la probabilidad de bloqueo de las sesiones nuevas P_b y para la probabilidad de fallo del handover P_h. El análisis se basa en los resultados existentes

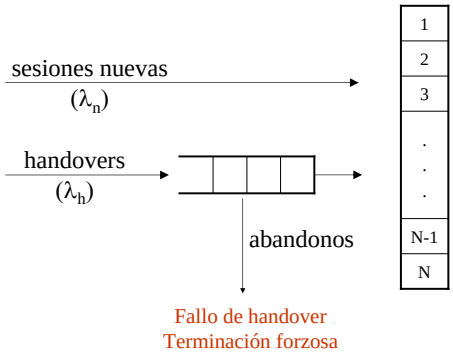


Figura 4.11: Modelo de la gestión de recursos en la célula.

para colas multiservidor con clientes impacientes [BdW94].

Definimos las probabilidades en estado estacionario

$$p_i \equiv Prob.(i \text{ unidades de recursos ocupadas}) \quad 0 \leq i \leq N.$$

Obsérvese que p_N incluye todos los posibles estados de ocupación de la cola de handovers. Aplicando los resultados de [BdW94, Ecuaciones (2.3)–(2.7)] se obtiene que:

$$p_i = \begin{cases} \rho^i / i!\, p_0 & 1 \leq i \leq N-n \\ \rho^{N-n} \rho_h^{i-(N-n)} / i!\, p_0 & N-n+1 \leq i < N \\ \lambda_h J p_{N-1} & i = N \end{cases} \tag{4.18}$$

donde $\rho_n = \lambda_n/\mu$, $\rho_h = \lambda_h/\mu$, $\rho = \rho_n + \rho_h$, p_0 se calcula a partir de la condición de normalización

$$p_0 = \left[\sum_{i=0}^{N-n} \frac{\rho^i}{i!} + \left(\frac{\rho}{\rho_h}\right)^{N-n} \cdot \left(\sum_{i=N-n+1}^{N-1} \frac{\rho^i}{i!} + \lambda_h J \frac{\rho_h^{N-1}}{(N-1)!} \right) \right]^{-1} \tag{4.19}$$

y

$$J = \int_0^\infty e^{\left(\lambda_h \int_0^x F_{ha}^c(u)\mathrm{d}u - N\mu x\right)} \mathrm{d}x; \tag{4.20}$$

donde $F_{ha}^{c}(t) = 1 - \int_0^t f_{ha}(u)\mathrm{d}u$. Además, la probabilidad de que una petición abandone la cola antes de ser admitida, es decir que el handover falle, vale

$$P_h = \left[\left(1 - \frac{N}{\rho_h}\right)\lambda_h J + 1\right] p_{N-1} \tag{4.21}$$

y la probabilidad de bloqueo de una sesión nueva es

$$P_b = 1 - \sum_{i=0}^{N-n-1} p_i = 1 - p_0 \sum_{i=0}^{N-n-1} \frac{\rho^i}{i!}. \tag{4.22}$$

Aplicando la condición de equilibrio entre el flujo de handovers entrante y saliente obtenemos [Jab96]

$$\lambda_h = \frac{1 - P_b}{\mu_c/\mu_r + P_h}\lambda_n. \tag{4.23}$$

Las expresiones de la (4.18) a la (4.23) forman un sistema no lineal de ecuaciones de cuya solución se obtienen los valores de P_b y P_h. La probabilidad de terminación forzosa P_{ft} se relaciona con la probabilidad de fallo de handover a través de la expresión [Jab96]

$$P_{ft} = \frac{P_h}{\mu_c/\mu_r + P_h}.$$

4.2.3. Evaluación numérica

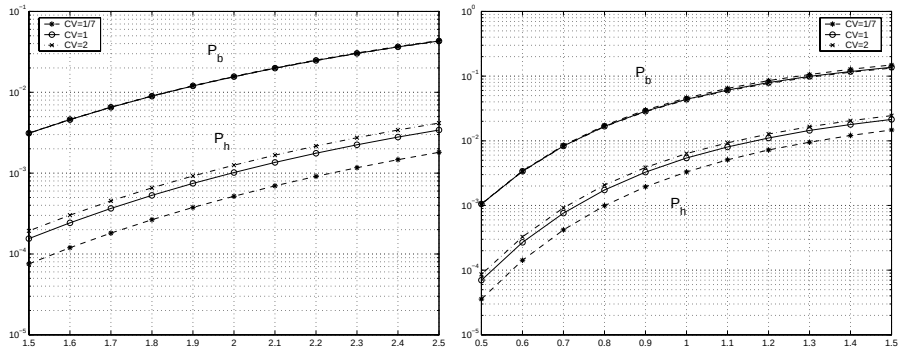
La evaluación numérica que aquí se presenta pretende por una parte mostrar la aplicabilidad del modelo desarrollado y por otra evaluar la influencia de la variabilidad de T_{ha} en las prestaciones del sistema.

Además de las definiciones introducidas anteriormente utilizaremos la notación y definiciones siguientes:

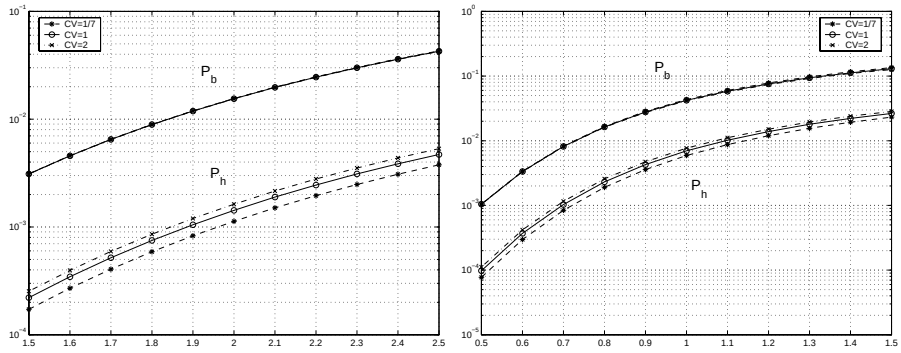
- $\alpha = \mu_{ha}/(N\mu)$.
- La movilidad de los terminales se cuantifica mediante μ_r/μ_c, que corresponde al número medio de handovers por sesión en un escenario con infinitos recursos.

- La variabilidad de T_{ha} se cuantifica a través de su coeficiente de variación $CV = \sqrt{E[T_{ha}^2] - E[T_{ha}]^2}/E[T_{ha}]$.

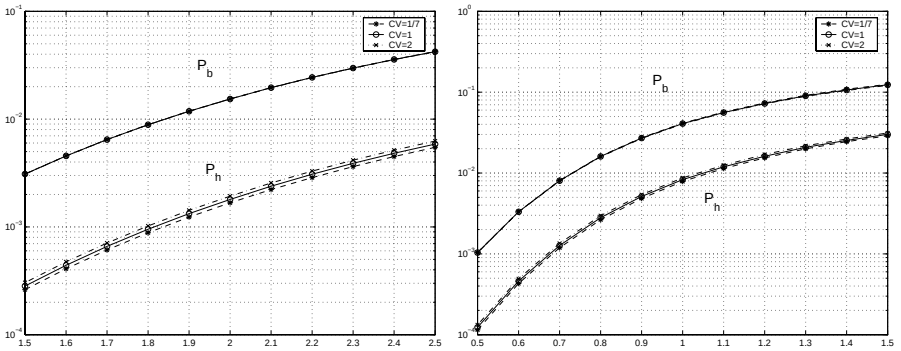
En las distintas gráficas de la figura 4.12 se representa el impacto del CV sobre P_h y P_b para diferentes valores de la movilidad μ_r/μ_c y del parámetro α. En todas las gráficas la cantidad total de recursos es $N = 10$ y la reserva para las peticiones de handover es $n = 1$. Para evaluar el impacto de la variabilidad de T_{ha} de ha tomado como referencia la distribución exponencial, para la que se cumple que $CV = 1$. Para conseguir una variabilidad menor ($CV < 1$) se ha utilizado una distribución de Erlang, y una distribución hiperexponencial de dos ramas para conseguir una variabilidad mayor ($CV > 1$). Los parámetros de la distribución se han ajustado para conseguir el CV deseado mientras que la media se determina a partir del valor de α. En los tres casos se ha tomado la misma configuración, $N = 10$, $n = 1$. De las curvas se observa que una variabilidad inferior a la de referencia ($CV < 1$) resulta en un valor menor de P_h, y una variabilidad mayor ($CV > 1$) da lugar a valores superiores de P_h, aunque este último efecto es algo menos pronunciado. Por otra parte, la influencia del CV sobre P_b es despreciable (en cada gráfica las tres curvas correspondientes a P_b se superponen). Se observa también que a valores más altos de la movilidad y del parámetro α se reduce el impacto del CV, mientras que la carga ρ_n no parece tener un impacto significativo. Para evaluar la influencia del CV en las prestaciones del sistema se ha considerado también la capacidad de este último, entendiendo como capacidad la máxima carga ρ_n que puede soportar el sistema para mantener una determinada *calidad de servicio* (QoS): $P_b < P_b^{max}$, $P_{ft} < P_{ft}^{max}$. La figura 4.13 (línea de trazo continuo) representa la variación de la capacidad respecto al valor del CV.



(a) $\alpha = 0.5$, $\mu_r/\mu_c = 1$ (b) $\alpha = 0.5$, $\mu_r/\mu_c = 4$



(c) $\alpha = 1$, $\mu_r/\mu_c = 1$ (d) $\alpha = 1$, $\mu_r/\mu_c = 4$



(e) $\alpha = 2$, $\mu_r/\mu_c = 1$ (f) $\alpha = 2$, $\mu_r/\mu_c = 4$

Figura 4.12: Probabilidades en función de ρ_n; $N = 10$, $n = 1$.

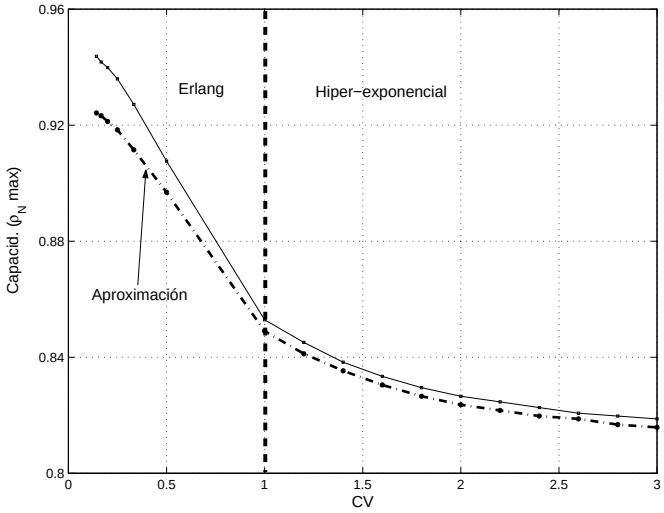


Figura 4.13: Capacidad del sistema frente al CV de T_{ha}; $P_b^{max} = 0.05$, $P_{ft}^{max} = 0.01$, $\mu_r/\mu_c = 4$, $\alpha = 0.5$, $N = 10$, $n = 1$.

4.2.4. Modelo aproximado

Una solución analítica y cerrada de (4.20) únicamente es posible para un número más bien reducido de distribuciones de T_{ha} (por ejemplo, determinista o exponencial [BdW94]), para el resto es necesario recurrir a la integración numérica con el fin de evaluar J. El coste compuacional de la evaluación numérica de las dos integrales anidadas de (4.20) no es despreciable, además el método la resolución —también numérico— del sistema no lineal (4.18)–(4.23) necesitará de varias iteraciones y en cada una de ellas hay que evaluar J.

Para reducir la carga computacional de la evaluación del modelo exacto se propone un modelo aproximado. Esta aproximación se basa en sustituir el eventual abandono de la cola por la "decisión" de no unirse a la cola desde un principio: si una petición finalmente abandona la cola sin recibir servicio, anticipar el abandono al momento de su llegada no debería suponer ninguna

diferencia respecto al número de peticiones atendidas o que abandonan. Para incorporar esta observación al modelo, se calcula la probabilidad de que una petición de handover pueda ser admitida antes de que expire su tiempo máximo de espera, en función del número de peticiones en la cola en el momento de su llegada. En el modelo aproximado se toma esta probabilidad como la probabilidad de admitir o no la petición y, una vez admitidas, las peticiones permanecen en el sistema hasta que reciben el servicio, es decir, no hay abandonos. El modelo al que se llega de esta forma es sólo aproximado y no exacto porque continuamos suponiendo que en cada estado las peticiones que entran al sistema siguen un proceso de Poisson, suposición que realmente no se cumple.

Sea w_i la probabilidad de que una petición de handover que ocupa la posición i-ésima en la cola sea atendida antes de que ésta tenga que abandonar la cola. El tiempo de espera para avanzar una posición en la cola sigue una distribución exponencial de tasa $\mu_w = N\mu$. Luego w_i es la probabilidad de que T_{ha} sea mayor que la suma de i variables aleatorias independientes e idénticamente distribuidas (de forma exponencial), por lo que

$$\begin{aligned} w_i &= \int_0^\infty \left(\int_0^u \frac{\mu_w^i}{(i-1)!} t^{i-1} e^{-\mu_w t}\, dt \right) f_{ha}(u)\, du \\ &= \frac{\mu_w^i}{(i-1)!} \int_0^\infty \int_t^\infty t^{i-1} e^{-\mu_w t} f_{ha}(u)\, du\, dt \\ &= \frac{\mu_w^i}{(i-1)!} \int_0^\infty t^{i-1} e^{-\mu_w t} \left(1 - \int_0^t f_{ha}(u)\, du \right) dt \\ &= \frac{\mu_w^i}{(i-1)!} \int_0^\infty t^{i-1} e^{-\mu_w t} F_{ha}^c(t)\, dt \end{aligned} \tag{4.24}$$

Utilizando la *transformada de Laplace* de $f_{ha}(t)$ encontramos una expresión alternativa para (4.24)

$$\begin{aligned} w_i &= \frac{\mu_w^i}{(i-1)!} (-1)^{i-1} \frac{d^{i-1}}{ds^{i-1}} \left(\mathrm{LT}\left[F_{ha}^c(t) \right](s) \right) \Big|_{s=\mu_w} \\ &= \frac{\mu_w^i}{(i-1)!} (-1)^{i-1} \frac{d^{i-1}}{ds^{i-1}} \left(\frac{1 - \mathrm{LT}\left[f_{ha}(t) \right](s)}{s} \right) \Big|_{s=\mu_w} . \end{aligned} \tag{4.25}$$

Si llamamos q_i a la probabilidad estacionaria de que haya i clientes en el sistema (en servicio más en espera), de las ecuaciones de balance globales se sigue que

$$q_i = \begin{cases} \frac{\rho^i}{i!}\, q_0 & 1 \leq i \leq N-n \\ \frac{\rho^{N-n}\rho_h^{i-(N-n)}}{i!}\, q_0 & N-n+1 \leq i \leq N \\ \frac{\rho^{N-n}\rho_h^n}{N!}\left(\frac{\rho_h}{N}\right)^{i-N} \prod_{j=1}^{N-1}(w_j)\, q_0 & i > N \end{cases},$$

donde q_0 se calcula a partir de la condición de normalización. A partir del valor de las probabilidades q_i calculamos

$$P_b = 1 - \sum_{i=0}^{N-n-1} q_i = 1 - q_0 \sum_{i=0}^{N-n-1} \frac{\rho^i}{i!}.$$

y

$$P_h = \sum_{i\geq 0} q_{N+i}(1 - w_{i+1}).$$

La evaluación del modelo aproximado no está completamente libre de utilizar integración numérica. Sin embargo, existen algunas diferencias con respecto al modelo exacto que hacen que la evaluación del modelo aproximado sea computacionalmente menos costosa:

- La integral de (4.24) puede resolverse analíticamente en más casos que la integral de (4.20).
- En (4.20) aparecen dos integrales anidadas por lo que el número de veces que hay que evaluar el integrando es el número de veces que habría que hacerlo en una integral "normal" al cuadrado.
- Las probabilidades w_i no dependen del valor de λ_h mientras J sí depende. Por tanto, si se quieren obtener resultados para distintas configuraciones habrá que recalcular más veces el valor de J. Además, para determinar el equilibrio de flujos de handovers hay resolver un sistema no lineal que requiere evaluar J en cada iteración mientras que los valores de las w_i únicamente habrá que calcularlos una vez.

Tabla 4.5: Comparación del coste computacional; $f_{ha}(t)$ Erlang / Hiper-exponencial, $N = 10$, $n = 1$.

		Mflop [a]		Tiempo [b] (s)	
α	$\frac{\mu_r}{\mu_c}$	Exacto	Aprox.	Exacto	Aprox.
0.5	1	17	0.1	8	0.8
2	4	46	0.3	21	1.2

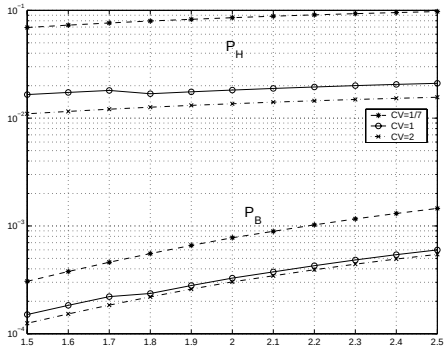
[a] 1 Mflop $= 10^6$ operaciones en coma flotante.

[b] Pentium® III (450) MHz ejecutando MATLAB ®.

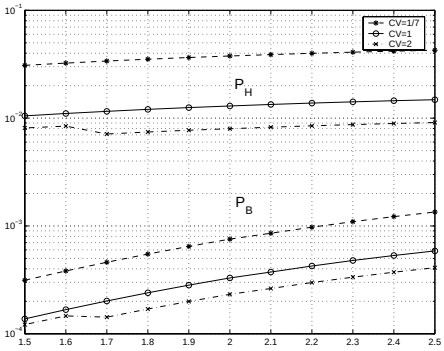
Tabla 4.6: Comparación del coste computacional; $f_{ha}(t)$ Gaussiana, $N = 10$, $n = 1$.

		Mflop		Tiempo (s)	
α	$\frac{\mu_r}{\mu_c}$	Exacto	Aprox.	Exacto	Aprox.
0.5	1	584	34	487	12
2	4	1707	19	1428	7

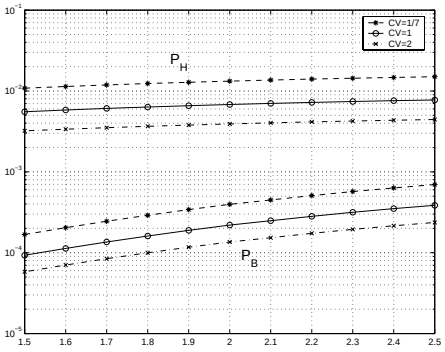
En las figuras 4.13 y 4.14 se comparan los resultados obtenidos mediante el modelo exacto y el aproximado, y en las tablas 4.5 y 4.6 se compara el coste computacional de la evaluación numérica de los dos modelos. Los valores de cada tabla se corresponden con los costes de obtener los valores de una gráfica como la de la figura 4.12. La tabla 4.5 muestra los costes cuando T_{ha} sigue una distribución erlanguiana o hiper-exponencial —dependiendo del valor de CV— y la tabla 4.6 cuando T_{ha} sigue una distribución gaussiana. En el primer caso es posible resolver analíticamente la integral interna de (4.20) y la integral de (4.24), mientras que en el segundo caso hay que recurrir a la integración numérica en todos los casos. Consecuentemente, los costes son notablemente superiores en el segundo caso.



(a) $\alpha = 0.5$



(b) $\alpha = 1$



(c) $\alpha = 2$

Figura 4.14: Error relativo en las probabilidades en función de ρ_n; $\mu_r/\mu_c = 1$, $N = 10$, $n = 1$.

4.3. Sobre la cola $M/M/C/K/(FIFO, LIFO, SIRO) + PH$

Aunque los sistemas con clientes impacientes han suscitado cierto interés desde hace algunas décadas, la cantidad de publicaciones sobre este tema puede decirse que es relativamente moderada. En particular es difícil encontrar estudios en los que no se utilice alguna o varias de las suposiciones siguientes:

- El tiempo máximo de espera está distribuido exponencialmente.
- La cola es infinita.
- La disciplina de servicio es FIFO.

Concretamente, en el modelo desarrollado en la sección anterior se utilizaban las dos últimas suposiciones.

Eliminar o relajar las dos primeras hipótesis permitiría disponer de un modelo que capturase mejor la realidad que éste pretende describir en un mayor número de situaciones. Además, existen situaciones con clientes impacientes en las que la mejor disciplina de servicio no es FIFO, sino LIFO [TP92], o bien en las que por las características del sistema —no por ser la mejor opción— el orden con el que se atienden las peticiones es aleatorio (SIRO) [Bar04].

En esta sección desarrollamos un modelo en el que la cola es finita, el tiempo máximo de espera de los clientes sigue una distribución del tipo PH y la disciplina de servicio puede ser FIFO, LIFO o SIRO. Además, cuando se produce una llegada y la cola está llena se elige entre las siguientes opciones: perder la petición que acaba de llegar, desalojar la petición que ocupa la cabecera de la cola, o elegir de forma aleatoria entre las dos anteriores.

De entre los trabajos previos relacionados con este tema es interesante mencionar, al menos, los siguientes. Zhao y Alfa [ZA95] analizan de forma

aproximada un sistema en el que la disciplina de servicio es LIFO y el tiempo de paciencia es determinista. Doshi y Heffes [DH86] estudian un sistema en el que aunque no existen cliente impacientes —al menos en el sentido de la impaciencia que aquí utilizamos— sí guarda relación con este tipo de sistemas. En el modelo de [DH86], después de un cierto tiempo las peticiones "caducan" y aunque no abandonan el sistema, y por tanto acaban recibiendo servicio, no se contabilizan en el trabajo cursado. En [DH86] se consideran las disciplinas FIFO y LIFO, y la duración en "buen estado" de la peticiones sigue una distribución cualquiera. Pese a las afinidades entre el modelo de [DH86] y el nuestro, existen diferencias sustanciales que hacen que el método de análisis que allí se utiliza no sea aplicable en nuestro caso. Movaghar [Mov98] analiza una cola finita con clientes impacientes cuyo tiempo de impaciencia sigue un distribución general y en el que la disciplina de servicio es FIFO.

Por otra parte, en el contexto de la aplicación de modelos de colas con clientes impacientes al estudio de sistemas celulares, existen trabajos en los que se propone la utilización de medidas para estimar el tiempo de paciencia de cada petición de handover e implantar una disciplina de servicio que se base en esas estimaciones. En esta línea, Xhafa y Tonguz [XT04] analizan un sistema con dos colas finitas de distinta prioridad. A su llegada la peticiones se colocan en una de las dos colas basándose en una estimación de su tiempo de paciencia. La prioridad de las peticiones en la cola de menor prioridad varía dinámicamente y éstas pueden promocionar a la cola de mayor prioridad. Primero se atienden las peticiones de la cola de mayor prioridad y dentro de cada cola se sigue una disciplina FIFO. En este modelo, todas las variables temporales están distribuidas exponencialmente. Tekinay y Jabbari [TJ92] proponen una disciplina de servicio para la cola de peticiones de handover que denominan MBPS (*Measurement-Based Priority Scheme*). En esta disciplina, en el momento que se va a admitir una de las peticiones que están esperando, se elige la de aquel terminal que en ese momento recibe una señal más débil. La evaluación de este esquema se realiza mediante simulación y la distribución del tiempo máximo de espera es gaussiana. Ebersman y Tonguz [ET99] proponen un esquema como el de [TJ92] pero en el que la

asignación de prioridad a las peticiones se basa no sólo en la potencia recibida por los terminales, sino también en la tasa de cambio de ésta.

El modelo que aquí desarrollamos se diferencia de los anteriores en que no se supone una distribución exponencial del tiempo de paciencia, la cola es finita y además de la disciplina FIFO permite analizar también las disciplinas LIFO y SIRO. El análisis del modelo se realiza mediante el método geométrico-matricial [Neu81, LR99].

4.3.1. Descripción del modelo y analísis

Las características del modelo que se estudia son las siguientes:

- Las llegadas siguen un proceso de Poisson.
- El tiempo de servicio sigue una distribución exponencial de parámetro μ.
- El número de servidores es C.
- La cola es de capacidad finita y tiene N posiciones.
- El tiempo máximo de espera de una petición sigue una distribución del tipo PH cuya representación [LR99] es $(\boldsymbol{\beta}, \boldsymbol{T})$; el parámetro m representa el número de fases y por tanto la dimensión de $\boldsymbol{\beta}$ es $1 \times m$, y la de $\boldsymbol{T}$ es $m \times m$.

Utilizando la extensión de la notación de Kendall que es habitual en los sistemas con clientes impacientes —al menos desde [BH81]— la representación del sistema que analizamos aquí sería $M/M/C/(C+N) + PH$.

El modelo del sistema es un proceso *cuasi de nacimiento y muerte* (QBD), finito y no homogéneo [Neu81]. En el desarrollo siguiente utilizamos las conveciones y notación matricial habituales (véase el apéndice A).

Sea $\{X(t) : t > 0\}$ el proceso estocástico que describe el estado del sistema que toma valores en el espacio bidimensional S,

$$S = \Big\{(l,k) : 0 \leq l \leq N;\, 0 \leq k \leq m^l - 1\Big\} \bigcup \Big\{(-1,k) : 0 \leq k \leq C-1\Big\}.$$

La primera coordenada de un estado de S representa el *nivel* y la segunda la *fase*. El conjunto de estados en un mismo nivel lo representamos como

$$L(l_0) = \{(l,k) : l = l_0;\, (l_0,k) \in S\}$$

y por tanto S puede representarse también como

$$S = \bigcup_{l=-1}^{N} L(l).$$

El nivel -1 agrupa aquellos estados en los que alguno de los servidores está libre. Para los estados de este nivel la fase representa el número de clientes en el sistema, que están siendo servidos. El nivel l ($0 \leq l \leq N$) agrupa aquellos estados en los que todos los servidores están ocupados y hay l clientes esperando. En este caso la fase de un estado representa la fase del tiempo de paciencia de cada uno de los l clientes que están esperando. La correspondencia entre la fases de los tiempos de paciencia y la fase k del estado es de la siguiente forma. Si la l-tupla,

$$(k_1, \ldots, k_l) \qquad 1 \leq k_i \leq m, \quad 1 \leq i \leq l$$

denota las fases asociadas a las distribuciones PH del tiempo de paciencia de los clientes que esperan, siendo k_i la fase del cliente que está en la i-ésima posición, entonces la fase del estado es

$$k = \sum_{1 \leq i \leq l} k_i m^{l-i},$$

es decir, las l-tuplas se numeran siguiendo un orden lexicográfico desde $(1, \ldots, 1)$ a $(m, \ldots, m)$ y este número es la fase.

Sea $\boldsymbol{\pi}$ el vector de las probabilidades estacionarias del proceso. Del mismo modo que con los estados, siguiendo los niveles dividimos $\boldsymbol{\pi}$ en los vectores

$\boldsymbol{\pi}^{(l)}$, $-1 \leq l \leq N$, donde $\boldsymbol{\pi}^{(-1)}$ tiene C componentes y $\boldsymbol{\pi}^{(l)}$ $(l \geq 0)$ m^l componentes. Por ser tratarse de un proceso QBD únicamente existen transiciones entre estados del mismo nivel o de dos niveles adyacentes y, en consecuencia, el generador infinitesimal del proceso tiene una estructura tridiagonal a bloques

$$\boldsymbol{Q} = \left[\begin{array}{c|ccc} \boldsymbol{A}_1^{(-1)} & \boldsymbol{A}_0^{(-1)} & \boldsymbol{0}^t & \cdots \\ \hline \boldsymbol{A}_2^{(0)} & & & \\ \boldsymbol{0} & & \boldsymbol{Q}_p & \\ \vdots & & & \end{array}\right]$$

donde

$$\boldsymbol{Q}_p = \begin{bmatrix} \boldsymbol{A}_1^{(0)} & \boldsymbol{A}_0^{(0)} & & \\ \boldsymbol{A}_2^{(1)} & \boldsymbol{A}_1^{(1)} & \boldsymbol{A}_0^{(1)} & \\ & & \ddots & \\ & & \boldsymbol{A}_2^{(N)} & \boldsymbol{A}_1^{(N)} \end{bmatrix} \tag{4.26}$$

Los bloques que implican al nivel de frontera -1 ($\boldsymbol{A}_1^{(-1)}$, $\boldsymbol{A}_0^{(-1)}$, $\boldsymbol{A}_2^{(0)}$) no siguen el procedimiento general de construcción que se utilizará para el resto de bloques. Las estructura de estos bloques es

$$\boldsymbol{A}_1^{(-1)} = \begin{bmatrix} * & \lambda & & & & \\ \mu & * & \lambda & & & \\ & 2\mu & * & \lambda & & \\ & & & & & \\ & & & \ddots & & \\ & & & & & \\ & & & & (C-1)\mu & * \end{bmatrix},$$

donde los elementos de la diagonal de $\boldsymbol{A}_1^{(-1)}$, que se han representado mediante asteriscos, toman el valor necesario para que las filas de $\boldsymbol{Q}$ sumen cero,

es decir, $\boldsymbol{A}_1^{(-1)}\boldsymbol{e} + \boldsymbol{A}_0^{(-1)}\boldsymbol{e} = \boldsymbol{0}$;

$$\boldsymbol{A}_0^{(-1)} = \begin{bmatrix} 0 & 0 & \cdots & \lambda \end{bmatrix}^t;$$

$$\boldsymbol{A}_2^{(0)} = \begin{bmatrix} 0 & 0 & \cdots & C\mu \end{bmatrix}$$

Hasta este punto la descripción del modelo es común para todas las disciplinas de servicio. Sin embargo, la construcción de los bloques $\boldsymbol{A}_{0,1,2}^{(l)}$ depende de la disciplina de servicio y del esquema de gestión de memoria empleado. A continuación se detalla el procedimiento para construir los distintos bloques para cada una de las disciplinas (FIFO, LIFO y SIRO) cuando la gestión de memoria empleada es la que rechaza una petición que llega a una cola llena y, posteriormente, se describen los cambios que habría que hacer cuando el método de gestión de memoria es otro más general.

Disciplina FIFO

Matrices $\boldsymbol{A}_0^{(l)}$, $0 \leq l \leq N-1$. Estas matrices se corresponden con las transiciones del nivel $L(l)$ al nivel $L(l+1)$, $(0 \leq l \leq N-1)$, que representan la llegada de un nuevo cliente que ocupará la $(l+1)$-ésima posición en la cola. El tiempo de paciencia de este cliente comenzará en la fase i-ésima con probabilidad β_i.

$$\boldsymbol{A}_0^{(l)} = \boldsymbol{I}_{m^l} \otimes \lambda\boldsymbol{\beta} \tag{4.27}$$

Matrices $\boldsymbol{A}_2^{(l)}$, $1 \leq l \leq N$. Estas matrices se corresponden con las transiciones del nivel $L(l)$ al nivel $L(l-1)$, $(1 \leq l \leq N)$, que representan la marcha de un cliente, la cual puede deberse a un abandono de la cola por impaciencia o por haber finalizado el servicio. El primer tipo de transición (por impaciencia) corresponde a la matriz $\boldsymbol{U}_1^{(l)}$, y el segundo (por finalización del servicio) a la matriz $\boldsymbol{U}_2^{(l)}$. Esto es,

$$\boldsymbol{A}_2^{(l)} = \boldsymbol{U}_1^{(l)} + \boldsymbol{U}_2^{(l)} \tag{4.28}$$

donde

$$\boldsymbol{U}_1^{(l)} = \begin{cases} \boldsymbol{\tau}, & l = 1 \\ \boldsymbol{\tau} \otimes \boldsymbol{I}_{m^{l-1}} + \boldsymbol{I}_m \otimes \boldsymbol{U}_1^{(l-1)}, & 1 < l \leq N \end{cases}$$

siendo $\boldsymbol{\tau} = -\boldsymbol{T}\boldsymbol{e}$, y

$$\boldsymbol{U}_2^{(l)} = C\mu \boldsymbol{e}_m \otimes \boldsymbol{I}_{m^{l-1}}. \tag{4.29}$$

Nótese que en (4.29) se ha utilizado el hecho de que la disciplina de servicio es FIFO.

Matrices $\boldsymbol{A}_1^{(l)}$, $0 \leq l \leq N$. Estas matrices se corresponden con las transiciones dentro del nivel $L(l)$, $(0 \leq l \leq N)$, que representan cambios en la fase de las distribuciones del tiempo de paciencia de los clientes que esperan.

Como paso intermedio para el cálculo de $\boldsymbol{A}_1^{(l)}$ se introduce $\boldsymbol{D}^{(l)}$, que coincide con $\boldsymbol{A}_1^{(l)}$ salvo en la diagonal principal. En primer lugar obtenemos la expresión de $\boldsymbol{D}^{(l)}$ sin preocuparnos de los elementos de su diagonal principal y posteriormente se obtiene $\boldsymbol{A}_1^{(l)}$ como

$$\boldsymbol{A}_1^{(l)} = \boldsymbol{D}^{(l)} - \operatorname{diag}\left\{ \boldsymbol{A}_2^{(l)}\boldsymbol{e} + \boldsymbol{D}^{(l)}\boldsymbol{e} + \boldsymbol{A}_0^{(l)}\boldsymbol{e} \right\}. \tag{4.30}$$

La matriz $\boldsymbol{D}^{(l)}$ puede construirse recursivamente como

$$\boldsymbol{D}^{(l)} = \begin{cases} \boldsymbol{T}, & l = 1 \\ \boldsymbol{T} \otimes \boldsymbol{I}_{m^{l-1}} + \boldsymbol{I}_m \otimes \boldsymbol{D}^{(l-1)}, & 1 < l \leq N \end{cases} \tag{4.31}$$

Nótese que para $\boldsymbol{D}^{(N)}$ se está utilizando el hecho de que los clientes que llegan cuando el proceso está en el nivel N —cuando la cola está llena—, se pierden. Aplicando la proposición 1, que se demuestra a continuación, podemos reescribir (4.30) como

$$\boldsymbol{A}_1^{(l)} = \begin{cases} \boldsymbol{D}^{(l)} - (C\mu + \lambda)\boldsymbol{I}_{m^l}, & l < N \\ \boldsymbol{D}^{(N)} - C\mu \boldsymbol{I}_{m^N}, & l = N \end{cases}. \tag{4.32}$$

Lema.

$$(A \otimes B)\,\boldsymbol{e} = (A\boldsymbol{e}) \otimes (B\boldsymbol{e}) \tag{4.33}$$

Proposición 1. *Para* $1 \leq l \leq N$ *se verifican las igualdades siguientes*

$$\operatorname{diag}\left\{A_0^{(l)}\boldsymbol{e}\right\} = \lambda \boldsymbol{I}_{m^l} \tag{4.34}$$

$$\operatorname{diag}\left\{A_2^{(l)}\boldsymbol{e} + \boldsymbol{D}^{(l)}\boldsymbol{e}\right\} = C\mu \boldsymbol{I}_{m^l} \tag{4.35}$$

Demostración. La ecuación (4.34) se obtiene inmediatamente de la aplicación del lema (4.27) y de la observación de que $\boldsymbol{\beta e} = 1$ y $\operatorname{diag}\{\boldsymbol{e}\} = \boldsymbol{I}$.

Para probar (4.35) primero observamos que

$$A_2^{(l)}\boldsymbol{e} + \boldsymbol{D}^{(l)}\boldsymbol{e} = \boldsymbol{U}_1^{(l)}\boldsymbol{e} + \boldsymbol{U}_2^{(l)}\boldsymbol{e} + \boldsymbol{D}^{(l)}\boldsymbol{e} \tag{4.36}$$

y aplicando el lema a (4.29) se obtiene que

$$\boldsymbol{U}_2^{(l)}\boldsymbol{e} = C\mu \boldsymbol{e}_{m^l}. \tag{4.37}$$

Por otra parte

$$\begin{aligned}
&\boldsymbol{U}_1^{(l)}\boldsymbol{e}_{m^{l-1}} + \boldsymbol{D}^{(l)}\boldsymbol{e}_{m^l} \\
&\quad = (\boldsymbol{\tau} \otimes \boldsymbol{I}_{m^{l-1}} + \boldsymbol{I}_m \otimes \boldsymbol{U}_1^{(l-1)})\boldsymbol{e} + (\boldsymbol{T} \otimes \boldsymbol{I}_{m^{l-1}} + \boldsymbol{I}_m \otimes \boldsymbol{D}^{(l-1)})\boldsymbol{e} \\
&\quad = (\boldsymbol{\tau} \otimes \boldsymbol{e}_{m^{l-1}} + \boldsymbol{e}_m \otimes (\boldsymbol{U}_1^{(l-1)}\boldsymbol{e}_{m^{l-2}})) \\
&\qquad + (\boldsymbol{T}\boldsymbol{e}_m \otimes \boldsymbol{e}_{m^{l-1}} + \boldsymbol{e}_m \otimes (\boldsymbol{D}^{(l-1)}\boldsymbol{e}_{m^{l-1}})) \\
&\quad = \boldsymbol{e}_m \otimes (\boldsymbol{U}_1^{(l-1)}\boldsymbol{e}_{m^{l-2}} + \boldsymbol{D}^{(l-1)}\boldsymbol{e}_{m^{l-1}})
\end{aligned} \tag{4.38}$$

y aplicando (4.38) de forma recursiva obtenemos

$$\begin{aligned}
\boldsymbol{U}_1^{(l)}\boldsymbol{e}_{m^{l-1}} + \boldsymbol{D}^{(l)}\boldsymbol{e}_{m^l} &= \boldsymbol{e}_m \otimes (\boldsymbol{U}_1^{(l-1)}\boldsymbol{e}_{m^{l-2}} + \boldsymbol{D}^{(l-1)}\boldsymbol{e}_{m^{l-1}}) \\
&= \boldsymbol{e}_{m^2} \otimes (\boldsymbol{U}_1^{(l-2)}\boldsymbol{e}_{m^{l-3}} + \boldsymbol{D}^{(l-2)}\boldsymbol{e}_{m^{l-2}}) \\
&\quad \vdots \\
&= \boldsymbol{e}_{m^{l-1}} \otimes (\boldsymbol{U}_1^{(1)} + \boldsymbol{D}^{(1)}\boldsymbol{e}_m) \\
&= \boldsymbol{e}_{m^{l-1}} \otimes (\boldsymbol{\tau} + \boldsymbol{T}\boldsymbol{e}_m) = \boldsymbol{e}_{m^{l-1}} \otimes \boldsymbol{0} = \boldsymbol{0}
\end{aligned} \tag{4.39}$$

Por tanto, de (4.37) y (4.39) se sigue que $A_2^{(l)}\boldsymbol{e} + \boldsymbol{D}^{(l)}\boldsymbol{e} = C\mu \boldsymbol{e}_{m^l}$, y aplicando el operador $\operatorname{diag}\{\cdot\}$ a ambos lados de esta igualdad obtenemos el resultado buscado. □

Disciplina LIFO

El hecho de cambiar la disciplina de servicio únicamente afecta a la forma en que se selecciona un cliente de la cola cuando uno de los que estaba recibiendo servicio termina. Por tanto, sólo habrá que modificar la expresión de $\boldsymbol{U}_2^{(l)}$, que en el caso de que la disciplina sea LIFO toma la siguiente forma

$$\boldsymbol{U}_2^{(l)} = \boldsymbol{I}_{m^{l-1}} \otimes C\mu \boldsymbol{e}_m \tag{4.40}$$

Por otra parte, es fácil comprobar que (4.37) sigue cumpliéndose por lo que la expresión de $\boldsymbol{A}_1^{(l)}$ no variará.

Disciplina SIRO

En este caso $\boldsymbol{U}_2^{(l)}$ toma la forma siguiente

$$\boldsymbol{U}_2^{(l)} = \begin{cases} C\mu \boldsymbol{e}_m, & l = 1 \\ l^{-1}\left[C\mu \boldsymbol{e}_m \otimes \boldsymbol{I}_{m^{l-1}} + \boldsymbol{I}_m \otimes (l-1)\boldsymbol{U}_2^{(l-1)}\right], & 1 < l \leq N \end{cases}$$

La expresión de $\boldsymbol{A}_1^{(l)}$ tampoco variará ya que la igualdad (4.37) continúa cumpliéndose. En este caso (4.37) puede demostrarse aplicando inducción sobre l.

Gestión de memoria

Hasta ahora hemos supuesto que cuando la cola está llena y llega una petición ésta se rechaza. Ahora vamos a ver qué cambios habrá que realizar para considerar un mecanismo más general en el que, cuando llega un cliente y la cola está llena, se elige de forma aleatoria e independiente para cada cliente entre rechazar este cliente o expulsar al cliente que ocupa la primera posición, que es el que lleva más tiempo de espera. La componente aleatoria de este mecanismo es ajustable a través del parámetro q, que representa la probabilidad con la que se elige la opción de expulsar al cliente que lleva más tiempo esperando.

La acción que se realice cuando se produce una llegada y la cola está llena solo repercute en las transiciones dentro del nivel $L(N)$ y, por tanto, sólo se modificará la matriz $\boldsymbol{D}^{(N)}$. Esta modificación es la siguiente

$$\boldsymbol{D}^{(N)} \longleftarrow \boldsymbol{D}^{(N)} + q\,(\boldsymbol{e}_m \otimes \boldsymbol{I}_{m^{N-1}} \otimes \lambda\boldsymbol{\beta})$$

luego (4.31) se convierte en

$$\boldsymbol{D}^{(l)} = \begin{cases} \boldsymbol{T}, & l = 1 \\ \boldsymbol{T} \otimes \boldsymbol{I}_{m^{l-1}} + \boldsymbol{I}_m \otimes \boldsymbol{D}^{(l-1)}, & 1 < l < N \\ \boldsymbol{T} \otimes \boldsymbol{I}_{m^{N-1}} + \boldsymbol{I}_m \otimes \boldsymbol{D}^{(N-1)} + q(\boldsymbol{e}_m \otimes \boldsymbol{I}_{m^{N-1}} \otimes \lambda\boldsymbol{\beta}), & l = N \end{cases} \tag{4.41}$$

y de la misma forma que se obtuvo (4.32) llegamos a

$$\boldsymbol{A}_1^{(l)} = \begin{cases} \boldsymbol{D}^{(l)} - (C\mu + \lambda)\boldsymbol{I}_{m^l}, & l < N \\ \boldsymbol{D}^{(N)} - (C\mu + q\lambda)\boldsymbol{I}_{m^l}, & l = N \end{cases}. \tag{4.42}$$

Análisis

Las probabilidades de estado $\boldsymbol{\pi}$ se obtienen de la resolución del sistema de ecuaciones lineales

$$\boldsymbol{\pi}\boldsymbol{Q} = \boldsymbol{0}^t, \quad \boldsymbol{\pi}\boldsymbol{e} = 1$$

Si $\boldsymbol{Q}$ es una matriz de dimensiones finitas, como es nuestro caso, este sistema en principio puede resolverse mediante cualquiera de los métodos estándar del álgebra lineal. Sin embargo, parece conveniente —sobre todo si el tamaño del sistema es grande— aprovechar la estructura y la naturaleza de $\boldsymbol{Q}$, que es un generador infinitesimal tridiagonal por bloques. Aquí hemos utilizado el algoritmo *Linear Level Reduction* [LR99, GJL84], que se aplica a la resolución de procesos QBD finitos y no homogéneos:

$U \leftarrow A_1^{(N)}$
$R^{(N)} \leftarrow A_0^{(N-1)} (-U)^{-1}$
for $l = N-1, N-2, \ldots, 0, -1$ **do**
  $U \leftarrow A_1^{(l)} + R^{(l+1)} A_2^{(l+1)}$
  $R^{(l)} \leftarrow A_0^{(l-1)} (-U)^{-1}$
end for

solve $\boldsymbol{\pi}^{(-1)}$ **from** $\{\boldsymbol{\pi}^{(-1)} U = \mathbf{0}^t;\ \boldsymbol{\pi}^{(-1)} \boldsymbol{e} = 1\}$
for $l = 0, 1, \ldots, N$ **do**
  $\boldsymbol{\pi}^{(l)} = \boldsymbol{\pi}^{(l-1)} R^{(l)}$
end for

Distribución del número de clientes en el sistema. Si $p_k (0 \leq k \leq N + C)$ denota la probabilidad de que haya k clientes en el sistema, tenemos que

$$p_k = \begin{cases} \boldsymbol{\pi}_k^{(-1)}, & k < C \\ \boldsymbol{\pi}^{(0)}, & k = C \\ \boldsymbol{\pi}^{(k-C)} \boldsymbol{e}, & C < k \leq C + N \end{cases} \tag{4.43}$$

De igual modo la distribución del número de clientes en la cola es

$$q_k = \begin{cases} \sum_{i=0}^{C} p_k, & k = 0 \\ p_{k+C}, & 0 < k \leq N \end{cases} \tag{4.44}$$

Probabilidades de bloqueo, expulsión y abandono. Dado que las llegadas son de Poisson, aplicando la propiedad *PASTA* [Wol82] tenemos que la probabilidad de que un cliente encuentre la cola llena es p_{C+N}. Por tanto, las probabilidades de que se bloquee a este cliente (P_b) o de que se expulse al más antiguo (P_e) son, respectivamente,

$$P_b = (1 - q) p_{C+N} = (1 - q) \boldsymbol{\pi}^{(N)} \boldsymbol{e} \tag{4.45}$$

$$P_e = q p_{C+N} = q \boldsymbol{\pi}^{(N)} \boldsymbol{e} \tag{4.46}$$

La probabilidad de abandono P_r se calcula dividiendo la tasa de abandonos entre la tasa de llegadas

$$P_r = \frac{1}{\lambda} \sum_{l=1}^{N} \boldsymbol{\pi}^{(l)} \boldsymbol{U}_1^{(l)} \boldsymbol{e}. \tag{4.47}$$

Tiempo medio en congestión y en bloqueo. Diremos que el sistema está congestionado si los clientes que llegan tienen que esperar, es decir si el proceso está en $\bigcup_{l=0}^{N} L(l)$, y denotaremos mediante la variable aleatoria T_c el tiempo de permanencia en esta situación. De igual modo, diremos que el sistema está en situación de bloqueo cuando está lleno —el proceso está en el nivel $L(N)$—, y la variable aleatoria T_b representa el tiempo de permanencia en esta situación.

La situación de congestión comienza cuando el proceso entra en el nivel $L(0)$ y termina cuando el proceso vuelve por primera vez al nivel $L(-1)$. Durante este periodo el proceso irá visitando los estados de $\bigcup_{l=0}^{N} L(l)$ y en todos ellos el tiempo de residencia es exponencial, por lo que la distribución de T_c es del tipo PH y su representación es $PH(\boldsymbol{\beta}^{(c)}, \boldsymbol{T}^{(c)})$ donde

$$\boldsymbol{\beta}^{(c)} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \end{bmatrix} \quad \text{y} \quad \boldsymbol{T}^{(c)} = \boldsymbol{Q}_p.$$

La matriz $\boldsymbol{Q}_p$ está definida en (4.26).

Para calcular el valor medio de T_c utilizamos un argumento probabilístico pues hacerlo utilizando su distribución implicaría invertir la matriz $\boldsymbol{T}^{(c)}$. El razonamiento se basa en tomar un periodo de observación suficientemente largo y dividir el tiempo total que el sistema está en congestión durante este periodo de observación entre el número de veces que el sistema entra en congestión. Así,

$$\overline{T}_c = \lim_{t_0 \to \infty} \frac{\left(\sum_{k=C}^{C+N} p_k \right) t_0 + o(t_0)}{\lambda p_{C-1} t_0 + o(t_0)} = \frac{1}{\lambda p_{C-1}} \sum_{k=C}^{C+N} p_k$$

Análogamente, deducimos que T_b sigue una distribución de tipo PH cuya

representación es $PH(\boldsymbol{\beta}^{(b)}, \boldsymbol{T}^{(b)})$, siendo

$$\boldsymbol{\beta}^{(b)} = \frac{1}{\boldsymbol{\pi}^{(N-1)}\boldsymbol{A}_0^{(N-1)}\boldsymbol{e}}\boldsymbol{\pi}^{(N-1)}\boldsymbol{A}_0^{(N-1)}$$
$$T^{(b)} = \boldsymbol{A}_1^{(N)}.$$

Por lo que su valor medio es [LR99, Eq. (2.13)]

$$\overline{T}_b = \boldsymbol{\beta}^{(b)}\left(-\boldsymbol{T}^{(b)}\right)^{-1}\boldsymbol{e}. \tag{4.48}$$

Aplicando el lema de la página 112 a (4.27) se sigue que

$$\boldsymbol{A}_0^{(N-1)}\boldsymbol{e} = \lambda\boldsymbol{e}$$

y, por otra parte, se tiene que

$$\boldsymbol{\pi}^{(N-1)}\boldsymbol{A}_0^{(N-1)} + \boldsymbol{\pi}^{(N)}\boldsymbol{A}_1^{(N)} = \boldsymbol{0}^t.$$

Finalmente, aprovechando las dos igualdades anteriores podemos reescribir (4.48) como

$$\overline{\boldsymbol{T}}_b = \frac{\boldsymbol{\pi}^{(N)}\boldsymbol{e}}{\lambda\boldsymbol{\pi}^{(N-1)}\boldsymbol{e}}.$$

4.3.2. Ejemplo numérico

Aquí se presentan unos ejemplos de aplicación del análisis del modelo que se ha desarrollado anteriormente. En estos ejemplos consideramos un sistema con diez servidores y cinco posiciones en la cola ($C = 10$, $N = 5$). Para el mismo sistema se evalúan cuatro combinaciones de disciplina de servicio y gestión de memoria:

FIFO disciplina FIFO y gestión de memoria con parámetro $q = 0$.

FIFOpo disciplina FIFO y gestión de memoria con parámetro $q = 1$.

SIRO disciplina SIRO y gestión de memoria con parámetro $q = 0.5$

LIFO disciplina LIFO y gestión de memoria con parámetro $q = 0$.

El tiempo medio de servicio vale una unidad ($\mu = 1$), lo que es equivalente a considerar que la tasa de llegadas λ y las tasas de transición del tiempo de paciencia $\boldsymbol{T}$ están normalizadas respecto a μ. Para el tiempo de paciencia se consideran dos distribuciones: una cuya tasa de abandono[2] es creciente con el tiempo (Erlang) y otra para la que es decreciente (hiper-exponencial). Esta elección está motivada por los resultados de [TP92] según los cuales cabe esperar un comportamiento relativo distinto para las diferentes disciplinas de servicio en cada uno de los dos casos. La distribuciones Erlang e hiper-exponencial son dos casos particulares de distribuciones del tipo PH. La representaciones PH de las distribuciones concretas que hemos utilizado son: para la erlanguiana

$$\boldsymbol{\beta} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{T} = \begin{bmatrix} -3 & 3 & 0 \\ 0 & -3 & 3 \\ 0 & 0 & -3 \end{bmatrix};$$

y para la hiper-exponencial

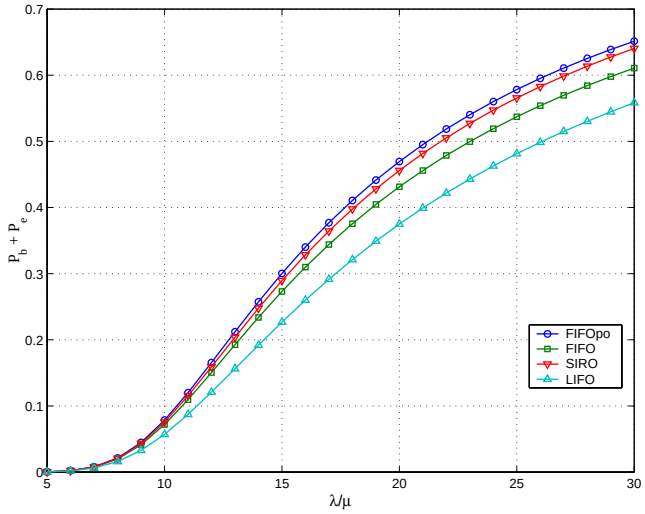
$$\boldsymbol{\beta} = \tfrac{1}{3} \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}, \quad \boldsymbol{T} = \tfrac{-56}{150} \operatorname{diag} \left\{ \begin{bmatrix} 50 & 10 & 1 \end{bmatrix} \right\}.$$

Para que los resultados puedan ser comparables las dos distribuciones se han ajustado para que su media sea la misma.

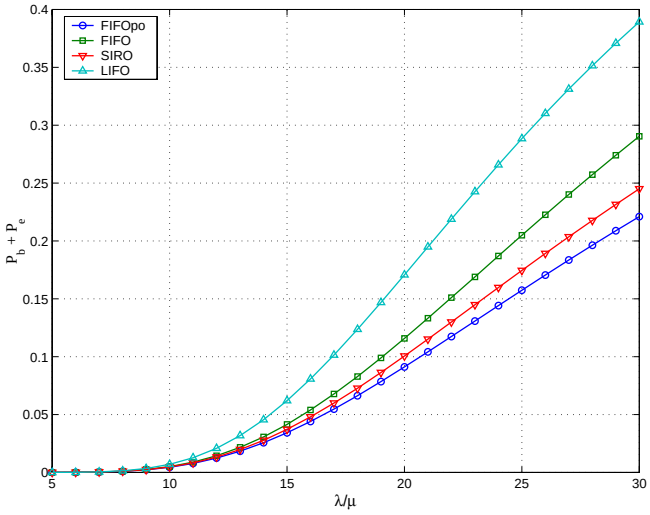
En la figura 4.15 se representa la suma de las probabilidades de bloqueo y expulsión en función del tráfico ofrecido, y en la figura 4.16 la probabilidad de abandono.

[2] La tasa de abandono es la función de riesgo de la distribución del tiempo de paciencia; si la función de densidad de la distribución es $f(t)$ la tasa de abandono será

$$h(t) = \frac{f(t)}{1 - F(t)}$$

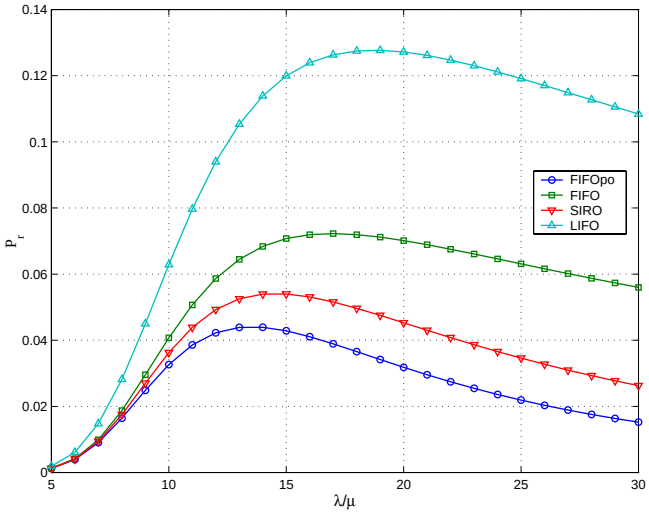


(a) Erlang

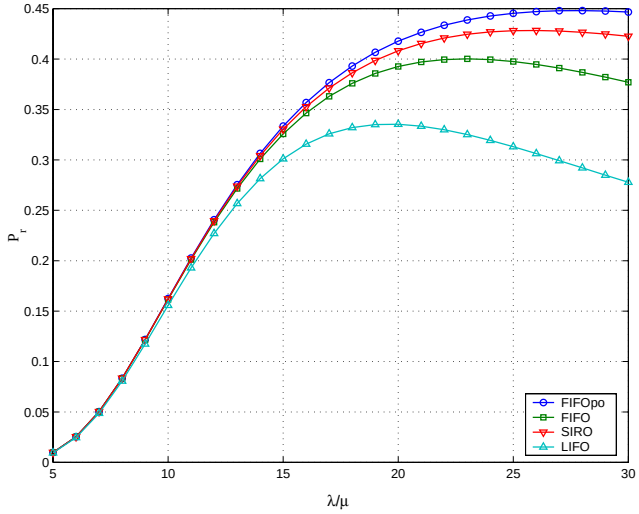


(b) Hiper-exponencial

Figura 4.15: Suma de las probabilidades de bloqueo y expulsión en función del tráfico ofrecido



(a) Erlang



(b) Hiper-exponencial

Figura 4.16: Probabilidad de abandono en función del tráfico ofrecido

4.4. Conclusiones

En este capítulo se ha presentado un método analítico-numérico para obtener la caracterización estadística del tiempo que un terminal móvil permanece en el área de handover y también del tiempo de ocupación de recursos mientras está en esta zona. Este método se ha aplicado a una estructura de células circulares y un modelo de movilidad sin direcciones privilegiadas (no contempla la existencia de carreteras o calles) en la que la velocidad de los terminales sigue una distribución de tipo gaussiano. El resultado de este análisis indica que los tiempos de permanencia y de ocupación de los recursos en el área de solape pueden ajustarse de forma satisfactoria por una distribución de tipo hiper-erlanguiano, y que modelar el tiempo de permanencia mediante una distribución exponencial —suposición común en los modelos de sistemas celulares— no es, en principio, una buena aproximación; si la aproximación exponencial es buena o no es algo que dependerá de cada modelo particular y de lo que se pretenda evaluar con él.

Concretamente, para estudiar la influencia de utilizar una distribución u otra del tiempo de permanencia en el área de handover, cuando lo que se quiere es estudiar la gestión de recursos de un sistema celular se ha utilizado un modelo en el que el tiempo de permanencia en el área de handover sigue una distribución cualquiera. Los resultados demuestran que en este caso suponer una distribución exponencial puede proporcionar una aproximación de una precisión razonable o, al menos una estimación conservadora: si el tiempo de residencia en el área de solape estuviese distribuido de la forma que resulta de nuestro estudio previo (distribución de tipo hiper-erlanguiano) y lo aproximáramos mediante una exponencial, las prestaciones del sistema calculadas con la aproximación serían peores que las exactas. Además del modelo exacto se ha propuesto también un modelo aproximado cuya resolución numérica es de una complejidad computacional significativamente inferior.

Por último se ha desarrollado un modelo analítico de una cola de capacidad finita con clientes impacientes cuyo tiempo de impaciencia sigue una

distribución de tipo PH. La principal aportación de este modelo consiste en que, además de cuando la disciplina de servicio es FIFO, permite realizar el análisis cuando la disciplina es LIFO o SIRO.

Capítulo 5

Optimización del control de admisión

En el capítulo 3 se han analizado distintos métodos para priorizar las peticiones de *handover* que pueden englobarse dentro de la familia de políticas de control de admisión basadas en la idea de *trunk reservation* (*Guard Channel* (GC) [PG85, HR86] y *Fractional Guard Channel* (FGC) [RNT97]) y sus variantes. Estas políticas se aplican a sistemas con un único servicio —tradicionalmente voz— y para decidir sobre la admisión de una petición se basan en el estado de ocupación de la célula en la que están operando. En este marco, y con las suposiciones habituales relativas a los procesos de llegada y tiempo de ocupación de los recursos, dichas políticas han demostrado ser óptimas para diferentes criterios [PG85, RNT97, Bar01, DS04]. En este capítulo se aborda la búsqueda de políticas óptimas en un marco más general en el que por una parte se considera la existencia de múltiples servicios y, por otra, la información utilizada por el control de admisión incorpora la predicción sobre la llegada de handovers a corto plazo. Como se verá, aunque en el escenario multiservicio se pierde el carácter óptimo de las políticas del tipo *trunk reservation*, la política denominada *Multiple Fraccional Guard Channel*, que es del tipo *trunk reservation*, permite obtener una capacidad del sistema que sin ser óptima en el más caso general, se acerca razonablemente a ésta. Esta razón ha motivado el estudio de un algoritmo para ajustar los parámetros de una po-

lítica MFGC y la capacidad máxima que en un sistema dado puede cursarse con una política del tipo MFGC.

El resto del capítulo está estructurado del siguiente modo. En la sección 5.1 se estudia el diseño de políticas de control de admisión optimas en el escenario multiservicio y en la sección 5.2 se propone un algoritmo para buscar la configuración óptima de una política MFGC y se evalúa su coste computacional. En la sección 5.3 se estudia el efecto que ejerce sobre las prestaciones del sistema utilizar la predicción de handovers en el control de admisión. Finalmente en 5.4 se resume este capítulo y se presentan las conclusiones.

5.1. Políticas de control de admisión en sistemas celulares multiservicio

La gestión de recursos radio en redes celulares monoservicio es un tema que ha sido, e incluso continua siendo, profusamente estudiado. Por otra parte, la aparición de la *Red Digital de Servicios Integrados de Banda Ancha* (RDSI-BA) propició en su momento el interés por el estudio de las redes fijas multiservicio [Ros95]. Sin embargo, la interacción de estos dos elementos, movilidad y multiservicio, es un materia que no ha recibido mucha atención hasta hace relativamente poco. En [LLC98], Li et al. proponen una generalización del mecanismo *Guard Channel* (GC) [HR86] para un escenario en el que existen varios servicios. En esta propuesta las llamadas nuevas de cada tipo dejan de admitirse a partir de un determinado nivel de ocupación del sistema, que puede ser distinto para cada servicio, mientras que la peticiones de handover de todos los servicios se admiten siempre que haya suficientes recursos libres. Bartollini y Chlamtac [BC02] consideran una política de admisión más general que la anterior en la que las peticiones de handover (salvo las del servicio más prioritario) también tienen asociado un umbral de ocupación a partir del cual no son admitidas. Recientemente, Heredia et al. [HUCPOG03c, HUCPOG03b, HUCPOG03a] plantean una extensión del

caso anterior en la que los umbrales pueden ser números no enteros, o lo que es lo mismo, la generalización al caso multiservicio del también conocido *Fractional Guard Channel* (FGC) [RNT97]. Finalmente, en [BC02] se demuestra que la política de admisión óptima, en cuanto que minimiza una cierta función de coste, no es en general de ninguno de los tipos anteriores, sino que pertenece al grupo más amplio de las políticas *estacionarias*[1] [Ros95].

Naturalmente, cuanto más general sea una política de admisión mayores serán sus posibilidades de cumplir unos determinados objetivos de QoS pero, por otra parte, esto implica más grados de libertad, es decir, más parámetros que han de ser ajustados adecuadamente para concretar esa potencialidad mayor. Desde un punto de vista teórico, el análisis de este tipo de sistemas y políticas no reviste mayor dificultad, pero desde una vertiente más ingenieril está lejos de ser una cuestión trivial por dos razones: cuando la cantidad de recursos disponibles y/o el número de servicios crecen mínimamente nos encontramos con el problema de la explosión de estados, lo que dificulta su resolución numérica; estamos ante un problema que no es exactamente de análisis sino de síntesis o diseño, nuestro problema no es evaluar una política concreta con unos parámetros concretos, sino encontrar la política o el valor de los parámetros que satisfagan un cierto QoS, y para esto es inviable una solución basada en un mero tanteo o búsqueda exhaustiva. Por otra parte, este problema de síntesis ha de resolverse en dos escalas temporales distintas: la de planificación de la red (¿cuál es la cantidad mínima de recursos necesaria para atender el tráfico ofrecido en el caso peor ?) y la de operación (si no estoy en el caso peor, para el tráfico ofrecido en ese momento, ¿cuál es la política de admisión que mejor satisface un determinado objetivo?).

En nuestro estudio consideramos las siguientes familias de políticas de control de admisión: *acceso total* (o *Complete Sharing*) (CS) [Ros95], *Multiple Guard Channel* (MGC) [BC02], *Multiple Fractional Guard Channel* (MFGC) [HUCPOG03c, HUCPOG03b] y *Radomized Stationary* (RS) [Ros95]. Estas políticas se comparan desde el punto de vista de la capacidad y se demuestra que

[1] Una política estacionaria es aquella en la que la decisión depende únicamente del estado actual del sistema.

la aplicación de la teoría de los *procesos markovianos de decisión* (MDP) [Ros70, Ros95] junto con técnicas de *programación lineal* constituye una herramienta versátil y eficiente para el diseño de las políticas RS, tanto en la fase de planificación como en la de operación.

5.1.1. Descripción del modelo

Consideramos un sistema de una célula que dispone de una cantidad total de recursos C y atiende peticiones (llamadas o conexiones) de N tipos distintos de usuarios. Cada uno de estos tipos de usuario tendrá unas características distintas y unos requisitos de calidad de servicio diferentes. Además, para cada tipo de tráfico tenemos que distinguir entre las peticiones de establecimiento de conexiones nuevas y las peticiones de establecimiento fruto de un traspaso (handover). Por las conocidas razones de tratabilidad matemática del modelo, supondremos que los procesos de llegada son poissonianos y que el tiempo de ocupación de los recursos por una sesión está distribuido exponencialmente. En virtud de estas suposiciones, el sistema tendrá la deseable propiedad de *memoria nula* por lo que su estado estará completamente representado por el vector $\boldsymbol{x} = (x_1, \ldots, x_N)$, donde x_i es el numero de llamadas del tipo i que están siendo cursadas independientemente de si accedieron al sistema como una llamada nueva o un handover. Por su parte, cada tipo de tráfico i ($i = 1, \ldots, N$) estará caracterizado por los siguientes parámetros: b_i, número de recursos necesarios para cursar una petición del servicio i; $1/\mu_i$, tiempo medio de ocupación de los recursos por la sesión, nótese que el tiempo de ocupación de los recursos no tiene porqué coincidir con la duración de la sesión ya que esta última comprende la utilización de recursos en una o varias células; λ_i^n, tasa de llegada de llamadas nuevas a una célula; λ_i^h, tasa de llegadas de de peticiones de handover a una célula; P_i^n, probabilidad de que una llamada nueva no sea admitida; P_i^h, probabilidad de que una petición de handover no sea admitida.

5.1.2. Políticas de Control de Admisión

En está sección se describe el funcionamiento de las políticas de control de admisión consideradas. En general, todas ellas basan la decisión de aceptar o rechazar una petición en el estado de ocupación del sistema en el momento que llega la petición y en el tipo de petición. El tipo de petición está determinado por el tipo de servicio solicitado y la distinción dentro de cada servicio entre llamadas nuevas y llamadas de handover. Así, el número total de tipos de petición será $2N$. Obsérvese que no se considera la historia pasada del sistema, cosa que, por otra parte, es lógica si tenemos en cuenta la *memoria nula* de nuestro modelo.

Las políticas o familias de políticas consideradas son, de menos a más general, las siguientes:

CS (Complete Sharing)
: Se acepta cualquier petición siempre que haya suficientes recursos disponibles.

MGC (Multiple Guard Channel)
: A cada tipo de petición se le asigna un umbral t, que es un número entero comprendido entre 1 y C. Tendremos por tanto un total de $2N$ umbrales $(t_1^n, \ldots, t_N^n, t_1^h, \ldots, t_N^h)$. Cuando llega una petición (por ejemplo del tipo i) se compara la cantidad de recursos ocupados si se admitiera la petición ($b(\boldsymbol{x}) + b_i = \sum_{j=1}^{N} x_j b_j + b_i$) con el umbral correspondiente y, en caso de que sea igual o menor, se acepta la petición. La política anterior (CS) es un caso particular de ésta en el que $t_i^n = t_i^h = C$ (i=1,. . . ,N).

MFGC (Multiple Fractional Guard Channel)
: Se trata de una generalización de la política anterior en la que se permite que los umbrales sean números no enteros ($0 < t_i^{n,h} \leq C$). Si t es el umbral asociado a una petición cualquiera, r la cantidad de recursos necesaria para atender la petición y $b(\boldsymbol{x})$ el número de recursos utilizados, el criterio seguido para decidir sobre la admisión sería:

- si $b(\boldsymbol{x}) + r \leq \lfloor t \rfloor$, aceptar.
- si $b(\boldsymbol{x}) + r = \lfloor t \rfloor$+1, aceptar con probabilidad $t - \lfloor t \rfloor$.
- si $b(\boldsymbol{x}) + r > \lfloor t \rfloor + 1$, rechazar.

RS (Randomized Stationary)
Una política estacionaria aleatorizada es aquella en la que la decisión depende únicamente del estado actual del sistema y de un componente aleatorio [Ros70]. En nuestro caso esto se podría formalizar del siguiente modo: a cada estado del sistema $\boldsymbol{x}$ se le asocian dos N-tuplas $\boldsymbol{\alpha}^n(\boldsymbol{x}), \boldsymbol{\alpha}^h(\boldsymbol{x}) \in [0,1]^N$; una petición de una llamada nueva (handover) del tipo i que llega al sistema en estado $\boldsymbol{x}$ se aceptará con probabilidad $\alpha_i^n(\boldsymbol{x})$ ($\alpha_i^h(\boldsymbol{x})$). Nótese que los grupos de políticas anteriores están incluidos dentro de éste, pero existen políticas RS que no son de ninguno de los tipos anteriores.

5.1.3. Análisis y diseño

Para el análisis de las políticas CS, MGC y MFGC el enfoque aplicado es el tradicional. Fijados los parámetros del sistema y de la política de admisión (valor de los umbrales $t_i^{n,h}$), se plantean las ecuaciones de balance globales del proceso de Markov de donde se obtienen el valor para los parámetros de QoS. En el caso particular de la política CS el proceso de Markov es reversible y, por tanto, las probabilidades de estado tienen forma de producto [Ros95], lo que permite utilizar el *algoritmo de convolución* [Ive02] para calcular éstas de forma más eficiente. En los otros dos casos, en los que no se da esta condición, el sistema de ecuaciones se ha resuelto empleando un método iterativo (*Gauss-Seidel*) que aprovecha el caracter disperso del sistema de ecuaciones.

En cualquier caso, con independencia del método que se utilice para resolver el sistema de ecuaciones, lo que se tiene es una función ($\Phi_1(C)$) que devuelve el valor de los parámetros de QoS ($P_i^{n,h}$) en función del valor de los parámetros de la política de admisión ($t_i^{n,h}$), de las características del tráfico

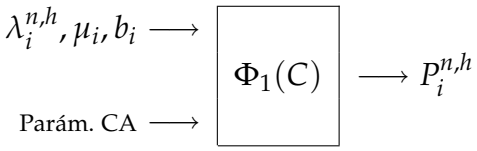


Figura 5.1: Diagrama del procedimiento de análisis

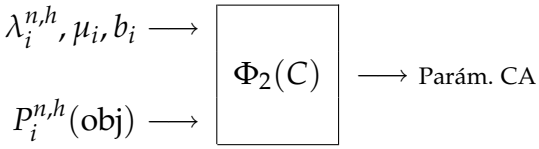


Figura 5.2: Diagrama del procedimiento de diseño

$(\lambda_i^{n,h}, \mu_i, b_i)$ y de los recursos del sistema (C); véase el diagrama de la figura 5.1. Esto es justamente lo que se necesita para analizar un sistema con una política de admisión concreta. Sin embargo, en el diseño o ajuste de una política de admisión el problema es el inverso: dados los parámetros del trafico y dados unos valores (o valores límite) para los parámetros de QoS, obtener el valor adecuado para los parámetros de control de admisión; a la función que realiza este cálculo la hemos llamado $\Phi_2(C)$ según se muestra en el diagrama de bloques de la figura 5.2 .

El método de análisis anteriormente descrito permite evaluar Φ_1 pero no Φ_2. Por otra parte, se podría obtener numéricamente Φ_2 a partir de Φ_1 mediante una cierta "busqueda inversa". Sin embargo, el carácter multidimensional hace que sea un problema difícil en sí mismo, además de costoso computacionalmente. El coste computacional de evaluar Φ_1 puede llegar a ser muy elevado y el uso de métodos iterativos para realizar esta búsqueda lo aumentaría todavía más. Encontrar algoritmos que realicen la búsqueda es complejo pues si bien P_i^n (P_i^h) decrece cuando el umbral correspondiente t_i^n (t_i^h) aumenta, el comportamiento con el valor del resto de los umbrales —al contrario de lo que en un principio podría parecer intuitivo— no siempre es monótono. En la figura 5.3 se representa un ejemplo de esto último para la configuración: $\lambda_1^n = 5$, $\lambda_2^n = 1$, $\lambda_1^h = 0.01$, $\lambda_2^h = 0$; $\mu_1 = 5$, $\mu_2 = 15$, $b_1 = 1$,

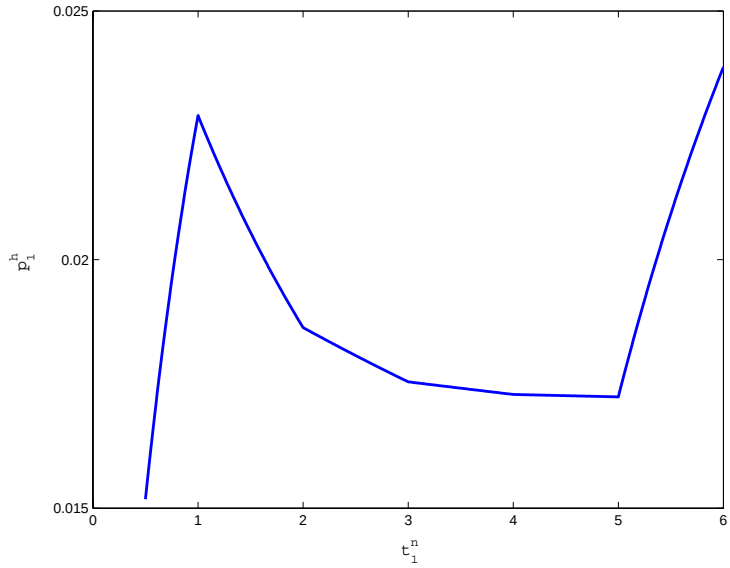


Figura 5.3: Comportamiento no monótono de P_1^h con t_1^n

$b_2 = 5$; $t_1^h = t_2^{n,h} = C = 6$. Una explicación intuitiva del comportamiento observado en la gráfica sería como sigue. Por una parte las llamadas nuevas del servicio 1 compiten por los recursos del sistema con las llamadas de handover del mismo servicio, por lo que al aumentar el umbral asociado a las llamadas nuevas del servicio 1 (t_1^n) se admitirán más llamada de este tipo lo que empujaría hacia un aumento de la probabilidad de bloqueo de las llamadas de handover del servicio 1; pero por otra parte, las llamadas del servicio 2 (tanto las nuevas como las de handover) también sufren el efecto del aumento de t_1^n por lo que cabría esperar una reducción en el número de llamadas del servicio 2 en el sistema lo que a su vez puede influir positivamente en las llamadas de handover del servicio 1 al dejar más recursos disponibles. Tenemos por tanto que al aumentar t_1^n se producen dos efectos opuestos sobre P_1^h, de forma que la curva es creciente cuando predomina el primero de ellos y decreciente cuando lo hace el segundo.

Debido a la existencia de estos inconvenientes se considera otra familia de políticas (RS). Al tratarse de una familia de políticas que contiene a las anteriores, teóricamente pueden obtenerse mejores prestaciones. Además, y tal vez más importante, al considerar esta clase más amplia de políticas es posible aplicar unas herramientas que permiten resolver el problema de diseño de una forma más adecuada; se dispone de algoritmos que se ajustan al paradigma de la figura 5.2.

Políticas RS

En el análisis de las políticas RS vamos a utilizar el marco teórico de los *procesos markovianos de decisión* (MDP) [Ros70, Ros95] junto con técnicas de *programación lineal* para la resolución de los problemas de optimización asociados. A continuación se hace un resumen de la notación y de los conceptos del ámbito de los MDP que van a emplearse.

El espacio de estados es

$$S := \left\{ \boldsymbol{x} : \sum_{i=1}^{N} x_i b_i \leq C;\ x_i \in \mathbb{N} \right\}. \tag{5.1}$$

Cada estado tiene asociado un conjunto de acciones posibles $A(\boldsymbol{x}) \subseteq A$, donde A es el conjunto de todas las acciones posibles,

$$A := \{ \boldsymbol{a} = (a_1, \ldots, a_N) :\ a_i = 0, 1, 2 \}. \tag{5.2}$$

El elemento a_i de una acción $\boldsymbol{a}$ representa el tratamiento que se les da a las peticiones del servicio i; los valores de a_i tienen el significado siguiente: $a_i = 0$, no se admiten llamadas nuevas ni peticiones de handover; $a_i = 1$, no se admiten llamadas nuevas pero sí peticiones de handover; $a_i = 2$, se admiten llamadas nuevas y peticiones de handover. Se supone que para un mismo servicio las peticiones de handover siempre tendrán mayor prioridad que las llamadas nuevas por lo que no se ha considerado la opción de admitir llamadas nuevas y no admitir peticiones de handover. El número total de acciones

en A será 3^N, aunque el numero de acciones posibles en un estado puede ser menor al no haber recursos disponibles; por ejemplo, para un estado en el que todos los recursos están ocupados ($b(\boldsymbol{x}) = C$) el conjunto de acciones posibles en ese estado sólo tiene un elemento, $A(\boldsymbol{x}) = \{(0, \ldots, 0)\}$. En una política RS cada vez que el proceso visita el estado $\boldsymbol{x}$ se elige una de las acciones posibles $A(\boldsymbol{x})$ de manera aleatoria según la distribución condicional de probabilidad $p_x(\boldsymbol{a})$, $\boldsymbol{a} \in A(\boldsymbol{x})$. Las probabilidades asociadas a las acciones y las probabilidades de admisión de cada tipo de petición se relacionan del siguiente modo:

$$\alpha_i^n(\boldsymbol{x}) = \sum_{a_i=2} p_x(\boldsymbol{a}), \quad \alpha_i^h(\boldsymbol{x}) = \sum_{a_i=1,2} p_x(\boldsymbol{a}). \tag{5.3}$$

En cada estado $\boldsymbol{x}$, la elección de una acción $\boldsymbol{a}$ determina la tasas de transición a los otros estados, $r_{\boldsymbol{xy}}(\boldsymbol{a})$. En nuestro caso tenemos dos tipos de transición, llegadas y salidas del sistema. Si $\boldsymbol{e}_i$ represent un vector en el que todas sus entradas son 0 salvo la i-ésima que vale 1, entonces las tasas correspondientes a cada tipo de transición son: para las llegadas ($\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{e}_i$)

$$r_{\boldsymbol{xy}}(\boldsymbol{a}) = \begin{cases} 0 & \text{si } a_i = 0 \\ \lambda_i^h & \text{si } a_i = 1 \\ \lambda_i^n + \lambda_i^h & \text{si } a_i = 2 \end{cases} \tag{5.4}$$

y para las salidas ($\boldsymbol{y} = \boldsymbol{x} - \boldsymbol{e}_i,\ x_i > 0$)

$$r_{\boldsymbol{xy}}(\boldsymbol{a}) = x_i \mu_i, \tag{5.5}$$

donde i denota el tipo de servicio al que pertenece la llegada o salida y

Para formular nuestro problema como un MDP necesitamos transformar el proceso de Markov en una cadena de Markov equivalente (*uniformizar* el proceso [Wol89]). Esto es posible pues se puede encontrar una cota superior Γ para la tasa total de salida de cada uno de los estados,

$$\sum_{\boldsymbol{y} \in S} r_{\boldsymbol{xy}}(\boldsymbol{a}) < \Gamma \quad \forall \boldsymbol{x} \in S,\ \boldsymbol{a} \in A(\boldsymbol{x}), \tag{5.6}$$

donde

$$\Gamma = \sum_{i=1}^{N} (\lambda_i^n + \lambda_i^h + C\mu_i). \tag{5.7}$$

La probabilidades de transición de la cadena de Markov resultante son,

$$p_{\boldsymbol{xy}}(\boldsymbol{a}) = \frac{r_{\boldsymbol{xy}}(\boldsymbol{a})}{\Gamma} \quad \text{si } \boldsymbol{y} \neq \boldsymbol{x} \tag{5.8}$$

donde $r_{\boldsymbol{xy}}(\boldsymbol{a})$ esta definido por las expresiones (5.4) y (5.5) según se trate de una llegada o una salida, respectivamente. Además, fruto de la transformación de *uniformización* aparecen autolazos (transiciones de un estado a sí mismo), cuya probabilidad es

$$p_{\boldsymbol{xx}}(\boldsymbol{a}) = 1 - \sum_{\boldsymbol{y} \in S} p_{\boldsymbol{xy}}(\boldsymbol{a}). \tag{5.9}$$

Se definen además las siguientes funciones de coste[2]

$$c_i^n(\boldsymbol{x}, \boldsymbol{a}) = \begin{cases} 1 & \text{si} \quad a_i = 0, 1 \\ 0 & \text{si} \quad a_i = 2 \end{cases} \tag{5.10}$$

$$c_i^h(\boldsymbol{x}, \boldsymbol{a}) = \begin{cases} 1 & \text{si} \quad a_i = 0 \\ 0 & \text{si} \quad a_i = 1, 2 \end{cases} \tag{5.11}$$

de modo que el promedio temporal de cada uno de los costes coincide con la probabilidad de bloqueo correspondiente, es decir,

$$P_i^{n,h} = \lim_{k \to \infty} \frac{E\left[\sum_{t=0}^{k} c_i^{n,h}(\boldsymbol{x}(t), \boldsymbol{a}(t))\right]}{k+1} \tag{5.12}$$

donde $(\boldsymbol{x}(t), \boldsymbol{a}(t))$ representa el estado y la acción en el instante t.

Si $p(\boldsymbol{x})$ representa la probabilidad estacionaria del estado $\boldsymbol{x}$, definimos $p(\boldsymbol{x}, \boldsymbol{a}) = p(\boldsymbol{x}) p_x(\boldsymbol{a})$ por lo que se cumple que

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{a} \in A(\boldsymbol{x})} p(\boldsymbol{x}, \boldsymbol{a}). \tag{5.13}$$

[2]Para las funciones de coste se ha respetado la notación que se utiliza en el caso más general en el que las funciones de coste dependen del estado y la acción aunque en nuestro caso únicamente dependen de la acción.

Conjuntos de restricciones. A continuación se definen distintos conjuntos de restricciones que posteriormente se utilizan para plantear diferentes criterios de diseño.

R0

$$\sum_{\boldsymbol{a}\in A(\boldsymbol{x})} p(\boldsymbol{x},\boldsymbol{a}) = \sum_{\substack{\boldsymbol{y}\in S \\ \boldsymbol{a}\in A(\boldsymbol{y})}} p(\boldsymbol{y},\boldsymbol{a}) p_{yx}(\boldsymbol{a}), \qquad \boldsymbol{x}\in S$$

$$\sum_{\substack{\boldsymbol{x}\in S \\ \boldsymbol{a}\in A(\boldsymbol{x})}} p(\boldsymbol{x},\boldsymbol{a}) = 1$$

$$p(\boldsymbol{x},\boldsymbol{a}) \geq 0, \qquad \boldsymbol{x}\in S,\ \boldsymbol{a}\in A(\boldsymbol{x})$$

Las restricciones de **R0** provienen de las ecuaciones correspondientes a la cadena de Markov asociada al proceso de decisión, por lo que estas restricciones serán aplicables en todos los criterios de diseño.

R1 ($i = 1, \ldots, N$)

$$\sum_{\substack{\boldsymbol{x}\in S \\ \boldsymbol{a}\in A(\boldsymbol{x})}} p(\boldsymbol{x},\boldsymbol{a}) c_i^n(\boldsymbol{x},\boldsymbol{a}) \leq P_i^n(max)$$

$$\sum_{\substack{\boldsymbol{x}\in S \\ \boldsymbol{a}\in A(\boldsymbol{x})}} p(\boldsymbol{x},\boldsymbol{a}) c_i^h(\boldsymbol{x},\boldsymbol{a}) \leq P_i^h(max)$$

En **R1** se han introducido los parámetros de diseño $P_i^{n,h}(max)$ que corresponden a los valores máximos para las probabilidades de bloqueo (QoS mínima).

R2 ($i = 1, \ldots, N$)

$$\sum_{\substack{\boldsymbol{x}\in S \\ \boldsymbol{a}\in A(\boldsymbol{x})}} p(\boldsymbol{x},\boldsymbol{a}) c_i^n(\boldsymbol{x},\boldsymbol{a}) \leq \frac{1}{n_i} z \tag{5.14}$$

$$\sum_{\substack{\boldsymbol{x}\in S \\ \boldsymbol{a}\in A(\boldsymbol{x})}} p(\boldsymbol{x},\boldsymbol{a}) c_i^h(\boldsymbol{x},\boldsymbol{a}) \leq \frac{1}{h_i} z \tag{5.15}$$

En **R2** se han introducido los pesos n_i, h_i que se utilizarán para ponderar las probabilidades de bloqueo de los distintos tipos de petición. También se ha introducido la variable auxiliar z, que junto con los pesos, se utilizará para aplicar un criterio de equidad del tipo *minimax* ponderado, es decir, optimizar (minimizar) la peor probabilidad de bloqueo ponderada (la más alta), o de otro modo,

$$\text{mín}\left\{\ \text{máx}\left\{n_i P_i^n,\ h_i P_i^h;\quad i=1,\ldots,N\right\}\ \right\}. \tag{5.16}$$

Con esto se busca una distribución equilibrada y equitativa de la probabilidad de bloqueo entre los distintos tipos de llamada (ponderadas mediante el coeficiente correspondiente) ya que en la aplicación de esta restricción será el tipo de llamada que recibe una peor QoS la que determine la QoS global del sistema.

Criterios de Diseño (CD). Los criterios de diseño que se contemplan están formados por una función objetivo a minimizar más uno o varios de los conjuntos de restricciones anteriores. Como tanto la funciones objetivo como las restricciones son lineales, el problema de diseño se convierte en un problema de programación lineal.

CD1

- Minimizar:
$$\sum_{\substack{i=1\\ \boldsymbol{x}\in S\\ \boldsymbol{a}\in A(\boldsymbol{x})}}^{N} p(\boldsymbol{x},\boldsymbol{a})\left(n_i c_i^n(\boldsymbol{x},\boldsymbol{a}) + h_i c_i^h(\boldsymbol{x},\boldsymbol{a})\right)$$
- Sujeto a: **R0**

Este criterio de diseño es el utilizado en [BC02] y encuentra la política óptima en tanto que minimiza el valor de

$$\sum_{i=1}^{N}\left(n_i P_i^n + h_i P_i^h\right).$$

Este criterio tiene la particularidad de que la solución es siempre una política estacionaria pura, es decir no aleatorizada: en cada estado siempre se elige la misma acción ($\forall \boldsymbol{x} \in S,\ \exists\ \boldsymbol{a} \in A(\boldsymbol{x}) : p(\boldsymbol{x}, \boldsymbol{a}) = 1$) lo que podría suponer una ventaja para la implementación. Sin embargo, este criterio tiene el inconveniente de que al hacer optimización global puede ocurrir que sacrifique excesivamente las probabilidades de bloqueo con menor peso dando lugar a problemas de equidad entre distintos servicios. Para solucionar este problema se introducen los criterios siguientes.

CD2

- Minimizar:
$$\sum_{\substack{i=1 \\ \boldsymbol{x}\in S \\ \boldsymbol{a}\in A(\boldsymbol{x})}}^{N} p(\boldsymbol{x}, \boldsymbol{a}) \left(n_i c_i^n(\boldsymbol{x}, \boldsymbol{a}) + h_i c_i^h(\boldsymbol{x}, \boldsymbol{a}) \right)$$
- Sujeto a: **R0, R1**

De este modo se limita el valor máximo que puede alcanzar la probabilidad de bloqueo de cada tipo de petición. En este caso la solución ya no es una política estacionaria pura, aunque puede demostrarse [Ros89] que, si $n_a(\boldsymbol{x})$ es el número de acciones entre las que se elige en el estado $\boldsymbol{x}$, se verifica que

$$\sum_{\boldsymbol{x}\in S} \left(n_a(\boldsymbol{x}) - 1 \right) \leq 2N.$$

Desde un punto de vista práctico es importante señalar que el problema asociado a **CD2** puede no tener solución si C no es lo suficientemente alto. Encontrar el valor mínimo de C para que el problema tenga solución, o su dual, el valor máximo del trafico ofrecido para que el problema tenga solución con un valor dado de C, son problemas propios de la fase de planificación o dimensionado de la red en los que puede aplicarse el **CD2**.

Durante la operación de la red, en una situación en la que el sistema está en congestión el problema de **CD2** no tiene solución puesto que no es posible cumplir las restricciones **R1**. No obstante, en este caso podría ser importante

que el deterioro de QoS se repartiese de forma equitativa entre los distintos servicios. Para este propósito se plantea el siguiente criterio.

CD3

- Minimizar: z
- Sujeto a: **R0, R2**
  Obsérvese que z es una variable auxiliar que se introduce a través del conjunto de restricciones **R2** (ecuaciónes (5.14) y (5.15)).

Este criterio también puede aplicarse en el supuesto contrario al anterior: cuando durante la operación de la red el tráfico ofrecido está por debajo del que el sistema puede cursar manteniendo los requisitos de QoS de **R1**, puede interesar repartir el margen de mejora de QoS de forma equitativa entre los servicios.

Por último, el criterio **CD4** permite definir dos modos de operación distintos dentro de una situación de no congestión: carga normal y carga alta.

CD4

- Minimizar: z
- Sujeto a: **R0, R1, R2**

Este criterio se obtiene añadiendo el conjunto de restricciones **R1** al criterio **CD3** de manera que se consigue el efecto siguiente. Existe un valor para la carga del sistema (tasa total de llegadas de llamadas nuevas) por debajo del cual ninguna de las restricciones de **R1** estaría activa, esto es, el óptimo del programa lineal definido por **CD4** se alcanza en un punto para el que las desigualdades de **R1** se cumplen de forma estricta. A la zona de funcionamiento por debajo de este valor la denominamos *modo de carga normal* y por

encima de este valor *modo de carga alta*. En el modo de carga normal el funcionamiento de **CD4** sería por tanto idéntico al de **CD3**: se minimiza el valor de todas las probabilidades de forma equilibrada manteniedo una proporcionalidad entre ellas según los pesos n_i, h_i. Sin embargo, en la zona de carga alta no es posible mantener esta proporcionalidad sin violar una o varias de las restricciones de **R1** por lo que estás restricciones entran en juego para limitar el valor máximo de las probabilidades de bloqueo. Esto se ilustrará mediante un ejemplo numérico en la sección siguiente, la representación gráfica de dicho ejemplo puede verse en la figura 5.8.

Aplicación de los diferentes criterios

Aquí se dan unas líneas generales sobre los casos o situaciones en los que pueden emplearse cada uno de los distintos criterios descritos anteriormente. Como ya se ha mencionado, el criterio **CD2** es adecuado para la fase de planificación o dimensionado de la red en la que a partir de las previsiones del tráfico ofrecido, se determina la cantidad mínima de recursos necesaria para ofrecer una cierta calidad de servicio. Por su parte, el criterio **CD3**, y el **CD4**, que es una versión más sofisticada de este último, son adecuados para ajustar dinámicamente la política durante la fase de operación pues las previsiones del tráfico que se emplean en la fase de planificación corresponden a los valores máximos que deberá soportar la red, mientras que durante el funcionamiento normal, el tráfico ofrecido en cada momento irá variando, tomando valores que pueden ser bastante distintos de los que se emplearon en la fase de planificación. Finalmente, el criterio **CD1** es el menos versátil de todos pues sólo permite especificar una importancia relativa entre los distintos tipos de llamada pero sin ofrecer garantías individuales para cada tipo ni siquiera garantías de equidad. Sin embargo, el problema de optimización asociado a este criterio es el menos complejo de todos. Estas características hacen que el criterio **CD1** sea apto para entornos en los que no se requieren garantías individuales para los diferentes tipos de llamadas pero, y concurren una o varias de las situaciones siguientes: el tamaño del sistema es elevado

Tabla 5.1: Parámetros de las configuraciones

	Configuración				
	A	B	C	D	E
b_1 (recursos)	1	1	1	1	1
b_2 (recursos)	2	4	2	2	2
f_1	0.8	0.8	0.2	0.8	0.8
f_2	0.2	0.2	0.8	0.2	0.2
$P_1^n(max)$ (%)	5	5	5	1	1
$P_2^n(max)$ (%)	1	1	1	2	1
	A,B,C,D,E				
$P_i^h(max)$	$0.1P_i^n(max)$				
λ_i^n (llamadas/s)	$f_i\lambda$				
λ_i^h (llamadas/s)	$0.5\lambda_i^n$				
μ_1 (s^{-1})	1				
μ_2 (s^{-1})	3				

(en número de recursos y/o servicios), la capacidad de computo es reducida, la frecuencia con la que se ha de actualizar la política es elevada.

5.1.4. Ejemplos de Aplicación y Resultados Numéricos

En esta sección se consideran varios casos de estudio a los que se les aplica las técnicas y criterios que se han descrito anteriormente. En primer lugar se evalúa la capacidad del sistema para las distintas políticas de admisión. Definimos la capacidad del sistema como el valor máximo de la tasa total de llegadas de llamadas nuevas ($\lambda = \sum_{i=1}^{N} \lambda_i^n$) para el que se cumplen los requisitos de QoS (valores máximos para $P_i^{n,h}$). La capacidad se ha evaluado para las distintas configuraciones recogidas en la tabla 5.1 y los valores de la capacidad se muestran en la tabla 5.2. Como es lógico, con una políticas más general se obtiene mayor capacidad, aunque la ganancia relativa que se obtiene es menor cuando la cantidad de recursos es más alta ($C = 40$). En particular, *RS* y *MFGC* tienden a igualarse cuando C aumenta, mientras que para *MGC* se observa la misma tendencia aunque de forma más lenta. La diferencia entre cualquiera de estas tres políticas y *CS* también se reduce,

Tabla 5.2: Capacidad (λ_{max} en llamadas/s)

Conf.	C	CS	MGC	MFGC	RS
A	10	1.54	1.88	2.05	2.07
	20	5.61	7.07	7.35	7.38
	40	15.7	19.4	19.7	19.8
B	10	0.36	0.40	0.42	0.44
	20	2.77	3.35	3.46	3.48
	40	10.3	12.5	12.7	12.8
C	10	1.36	1.51	1.65	1.67
	20	5.77	6.91	6.98	7.00
	40	17.6	20.1	20.4	20.5
D	10	1.74	1.97	2.02	2.04
	20	6.04	6.82	6.93	6.94
	40	16.5	18.2	18.4	18.4
E	10	1.5	1.7	1.8	1.8
	20	5.6	6.3	6.4	6.5
	40	15	17	17	17

aunque aquí el efecto es bastante menor y en todos los casos se mantiene una diferencia superior al 10 %.

A modo de ejemplo, el ajuste de las diferentes políticas para el caso $C = 10$ de la configuración A sería el siguiente: MGC, $t_1^n = 7$, $t_2^n = 9$, $t_1^h = 8$, $t_2^h = 10$; MFGC, $t_1^n = 6.43$, $t_2^n = 8.76$, $t_1^h = 8.19$, $t_2^h = 10$; RS, véase la tabla 5.3. Para ajustar la política RS se ha aplicado el **CD2**. Como puede verse la política resultante no es estacionaria pura aunque sólo está aleatorizada en 3 estados. Si se quiere obtener una política estacionaria pura se puede aplicar el **CD1**. Si se aplica el **CD1** utilizando como pesos el valor inverso de la máxima probabilidad de bloqueo correspondiente, es decir, $n_i = 1/P_i^n(max)$ y $h_i = 1/P_i^h(max)$, se obtiene una capacidad $\lambda_{max} = 1.81$. Esta capacidad es un 4 % menor que con MGC, un 12 % que con MFGC y un 13 % que con RS. Para mejorar este valor se han ajustado los pesos de la función objetivo empleando un algoritmo iterativo heurístico. Como resultado se obtiene una capacidad $\lambda_{max} = 2.03$, para los siguientes valores de los pesos $n_1 = 0.9697$, $n_2 = 2.265$, $h_1 = 3.991$, $h_2 = 53.355$. Por tanto, manteniéndose dentro de la familia de las políticas estacionarias puras se consigue una capacidad que es un 1 % y un 2 % menor que la que se obtiene con MFGC y RS, respectivamente, y es un 8 % superior a la capacidad que se obtiene con MGC.

En los ejemplos anteriores se han comparado las diferentes familias de políticas desde el punto de vista de la capacidad. Esto es útil en la fase de planificación de la red cuando se necesita conocer el tráfico que puede cursar con una cierta cantidad de recursos cumpliendo unos requisitos de QoS. Para este tipo de cálculos el operador debe considerar el caso peor para el tráfico ofrecido, es decir, una previsión del tráfico durante la hora cargada. Sin embargo, si la previsión es adecuada, la mayor parte del tiempo el tráfico ofrecido estará por debajo del valor previsto para la hora cargada y, aunque con menor frecuencia, podría ocurrir también que se superase este valor. En el primero de los casos habrá una cierta holgura en el cumplimiento de los requisitos de QoS y en el segundo se violará uno o varios de estos requisitos. Cabría plantearse por tanto, si es posible repartir la mejora o el empeoramiento de manera equitativa entre los distintos servicios. Para lograr esto

Tabla 5.3: Ejemplo de política RS

estado	recursos ocupados	probabilidades de admisión		acción	prob. de de acción
$\boldsymbol{x}$	$b(\boldsymbol{x})$	$\boldsymbol{\alpha}^n(\boldsymbol{x})$	$\boldsymbol{\alpha}^h(\boldsymbol{x})$	$\boldsymbol{a}$	$p_x(\boldsymbol{a})$
$(*,*)$	$0,\ldots,5$	$(1,1)$	$(1,1)$	$(2,2)$	1
$(6,0)$	6	$(0,1)$	$(1,1)$	$(1,2)$	1
$(4,1)$	6	$(0.75,1)$	$(1,1)$	$(1,2)$	0.25
				$(2,2)$	0.75
$(2,2)$	6	$(1,1)$	$(1,1)$	$(2,2)$	1
$(0,3)$	6	$(1,1)$	$(1,1)$	$(2,2)$	1
$(7,0)$	7	$(0,0)$	$(1,1)$	$(1,1)$	1
$(5,1)$	7	$(0,0.83)$	$(1,1)$	$(1,1)$	0.17
				$(1,2)$	0.83
$(3,2)$	7	$(0,1)$	$(1,1)$	$(1,2)$	1
$(1,3)$	7	$(1,1)$	$(1,1)$	$(2,2)$	1
$(8,0)$	8	$(0,0)$	$(0,1)$	$(0,1)$	1
$(6,1)$	8	$(0,0)$	$(0,1)$	$(0,1)$	1
$(4,2)$	8	$(0,0)$	$(1,1)$	$(1,1)$	1
$(2,3)$	8	$(0,0.86)$	$(1,1)$	$(1,1)$	0.14
				$(1,2)$	0.86
$(0,4)$	8	$(0,1)$	$(1,1)$	$(1,2)$	1
$(7,1)$	9	$(0,0)$	$(1,0)$	$(1,0)$	1
$(5,2)$	9	$(0,0)$	$(1,0)$	$(1,0)$	1
$(3,3)$	9	$(0,0)$	$(1,0)$	$(1,0)$	1
$(1,4)$	9	$(0,0)$	$(1,0)$	$(1,0)$	1
$(8,1)$	10	$(0,0)$	$(0,0)$	$(0,0)$	1
$(6,2)$	10	$(0,0)$	$(0,0)$	$(0,0)$	1
$(4,3)$	10	$(0,0)$	$(0,0)$	$(0,0)$	1
$(2,4)$	10	$(0,0)$	$(0,0)$	$(0,0)$	1
$(0,5)$	10	$(0,0)$	$(0,0)$	$(0,0)$	1

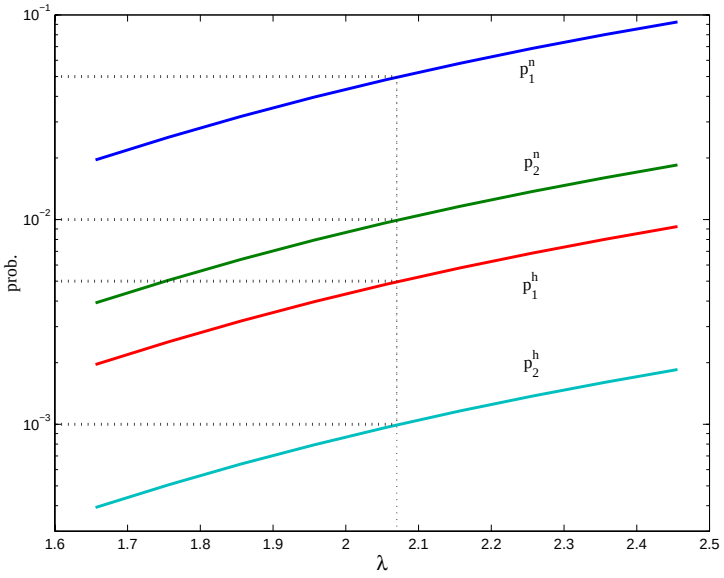


Figura 5.4: CD3: Probabilidades de bloqueo para distintas cargas.

se necesita reajustar los parámetros de la política en tiempo de operación a partir de los valores del tráfico estimados mediante medida, y aplicar el algoritmo que se deriva del **CD3** utilizando $n_i = 1/P_i^n(max)$ y $h_i = 1/P_i^h(max)$. Como ejemplo ilustrativo se ha utilizado el caso $C = 10$ de la configuración A. Se ha tomado el valor máximo de la capacidad ($\lambda_{max} = 2.07$) y se ha variado la tasa total ofrecida (λ) desde un 10 % por debajo de la capacidad a un 10 % por encima. En la figura 5.4 se representa la evolución de las distintas probabilidades de bloqueo cuando se emplea una política RS y se realiza el reajuste de la misma aplicando el **CD3**. En esta gráfica puede observarse como la distancia entre las diferentes curvas se mantiene constante. Esto mismo puede observarse con mayor exactitud en la figura 5.5, donde se representa el cociente entre cada probabilidad de bloqueo y su valor máximo; como consecuencia del reparto equitativo las cuatro curvas se solapan en una única. En la figura 5.6 y la figura 5.7 se muestra el mismo tipo de

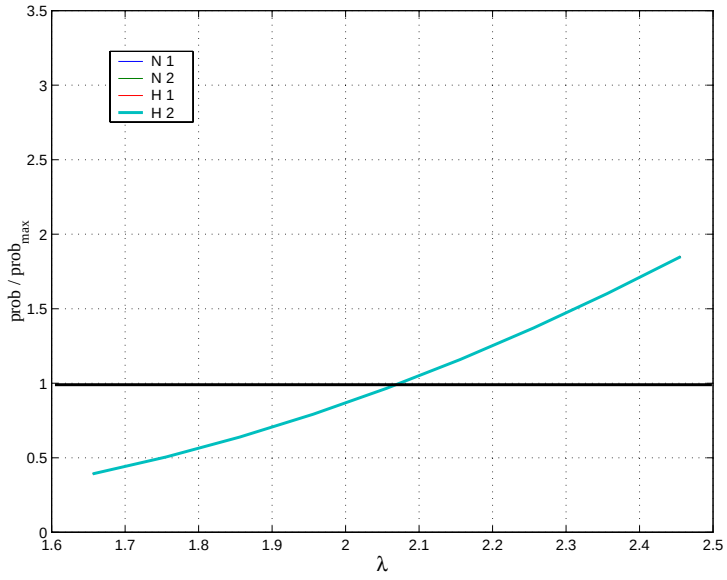


Figura 5.5: RS: valores relativos respecto a la especificación de QoS.

representación pero para las políticas MFGC y MGC, en las que no es posible el reajuste. En estos casos se observa que el aumento o el deterioro de la QoS no se distribuye equitativamente.

Por último, se presenta un ejemplo (figura 5.8) en que se utiliza el **CD4** para ajustar en tiempo de operación un sistema en el que hay dos modos de funcionamiento dependiendo de la carga, a saber, carga normal y carga alta. En el modo de carga normal el objetivo de QoS es $P^n_{1,2} \leq 0.01$ y $P^h_{1,2} \leq 0.001$, mientras que en el modo de carga alta se permite un deterioro de la QoS del servicio 1 ($P^n_1 \leq 0.05, P^h_1 \leq 0.005$) manteniendo la misma exigencia para el servicio 2 ($P^n_2 \leq 0.01, P^h_2 \leq 0.001$). De nuevo se ha utilizado el caso $C = 10$ de la configuración A. Para ajustar la política se ha aplicado el **CD4** con los parámetros siguientes: $P^n_1(max) = 0.05$, $P^n_2(max) = 0.01$, $P^h_1(max) = 0.005$, $P^h_2(max) = 0.001$, $n_1 = n_2 = 1$, $h_1 = h_2 = 10$.

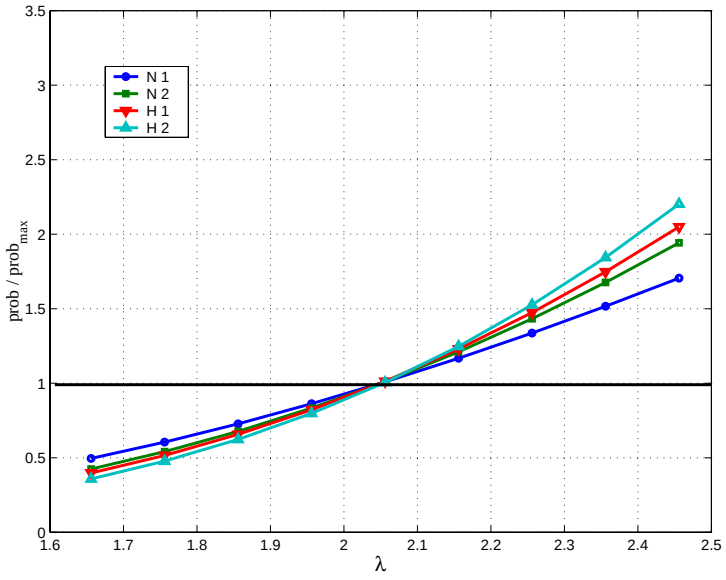


Figura 5.6: MFGC: valores relativos respecto a la especificación de QoS.

5.2. Algoritmo para la optimización de la política *Multiple Fraccional Guard Channel*

En la sección anterior se ha visto que, en general, la política de control de admisión óptima no es del tipo MFGC, sin embargo, según estos mismos resultados, la mejor política dentro de la familia MFGC consigue una capacidad que sin ser la óptima está muy próxima al valor óptimo, característica ésta que se acentúa todavía más al aumentar los recursos del sistema (véase la tabla 5.2 en la página 142). Además, las políticas del tipo MFGC presentan la ventaja frente a las del tipo RS de necesitar un menor número de parámetros para describir la política. Sin embargo, los métodos de optimación que se han presentado, basados en la formulación de un programa lineal, no sirven para optimizar la política de control de admisión dentro de la familia MFGC. Aunque este tipo de política ha recibido una atención relativamente

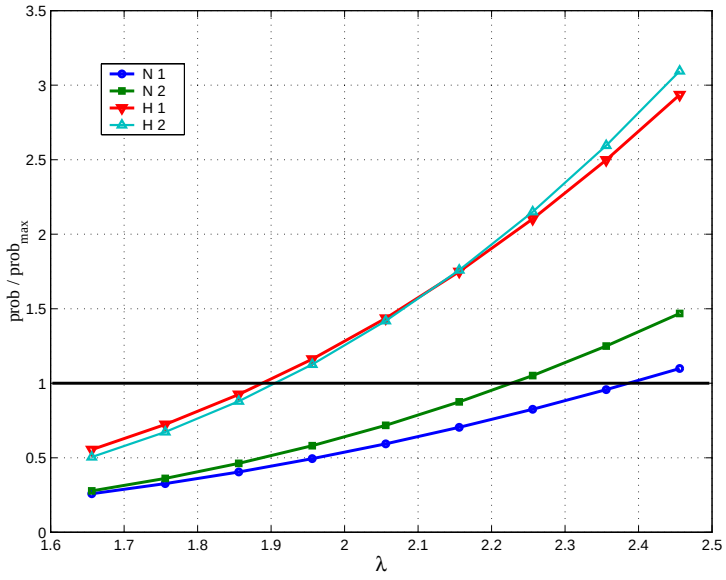


Figura 5.7: MGC: valores relativos respecto a la especificación de QoS.

importante en la literatura, sólo hemos sido capaces de encontrar una propuesta de algoritmo para realizar el ajuste óptimo de los parámetros de una política MFGC [HUCPOG03c, HUCPOG03b, HUCPOG03a]. En esta sección proponemos un algoritmo que realiza la misma función que el propuesto por Heredia et al. pero de una forma computacionalmente más eficiente.

5.2.1. Descripción del modelo

El modelo analítico del sistema que se utiliza en esta sección es esencialmente el mismo que se ha descrito en 5.1.1. No obstante, para poder demostrar una de las mejoras que introduce nuestro algoritmo, las tasas de llegada de las peticiones de handover no se toman como un dato, sino que éstas se calculan a partir de la condición de equilibrio entre los flujos de entrada y salida a la célula. Para ello necesitamos introducir sendas variables aleatorias

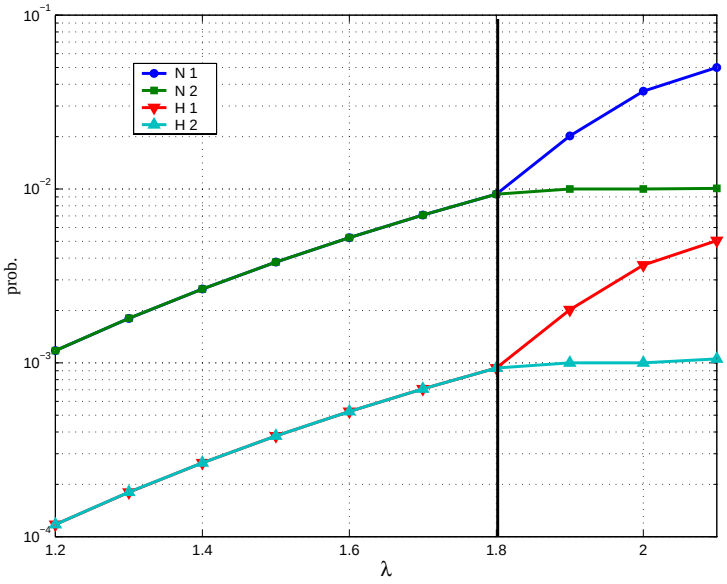


Figura 5.8: Doble modo de operación: carga normal y carga alta.

para el tiempo de residencia en una célula y la duración de la sesión, que supondremos que siguen una distribución exponencial de parámetros μ_i^r y μ_i^c (i=1, ..., N), respectivamente. Además, para poder comparar directamente nuestros resultados con los del algoritmo que se pretende mejorar se introduce la probabilidad de terminación forzosa para cada servicio (P_i^{ft}) como un parámetro en la especificación de la QoS. Si asumimos que tenemos una red de células homogéneas, las probabilidades de fallo de handover (P_i^h) y las de terminación forzosa están relacionas a través de la expresión

$$P_i^{ft} = \frac{P_i^h}{\mu_i^c/\mu_i^r + P_i^h} \qquad i = 1, \ldots, N.$$

Para hacer más compacta la notación denotaremos como flujo i a las llegadas de llamadas nuevas del servicio i, y como flujo $N+i$ a las llegadas de peticiones de handover del servicio i. De este modo el vector $\boldsymbol{p} = (P_1, \ldots, P_{2N})$ representará las probabilidades de bloque de los $2N$ flujos de llegada, donde

$P_i^n = P_i$, $P_i^h = P_{N+i}$.

Además de las razones habituales que respaldan la suposición de la distribución de probabilidad exponencial, es importante destacar que la contribución que se presenta en esta sección es el algoritmo que permite ajustar de forma óptima los parámetros de una política MFGC que es independiente del método particular que se emplee para calcular las probabilidades que caracterizan la QoS (vector $\boldsymbol{p}$), por lo que dicho método podría ser sustituido por otro —por ejemplo para trabajar con distribuciones no exponenciales— sin que ello afectase al algoritmo propuesto.

5.2.2. Análisis del modelo

El modelo del sistema es un proceso de nacimiento y muerte multidimensional cuyo conjunto de estados visitables es

$$S := \left\{ \boldsymbol{x} :\ x_i \in \mathbb{N};\ \sum_{i=1}^{N} x_i b_i \leq C;\ x_i b_i \leq \lfloor \max(t_i, t_{i+N}) \rfloor + 1, 1 \leq i \leq N \right\}.$$

Si $r_{\boldsymbol{xy}}$ es la tasa de transición del estado $\boldsymbol{x}$ al $\boldsymbol{y}$

$$r_{\boldsymbol{xy}} = \begin{cases} \alpha_i^n(\boldsymbol{x})\lambda_i^n + \alpha_i^h(\boldsymbol{x})\lambda_i^h & \text{if } \boldsymbol{y} = \boldsymbol{x} + \boldsymbol{e}_i \\ x_i \mu_i & \text{if } \boldsymbol{y} = \boldsymbol{x} - \boldsymbol{e}_i \\ 0 & \text{en otro caso} \end{cases}$$

donde $\boldsymbol{e}_i$ denota un vector cuyas entradas son todas nulas excepto la i-ésima que es 1. Los coeficientes $\alpha_i^n(\boldsymbol{x})$ y $\alpha_i^h(\boldsymbol{x})$ representan las probabilidades de aceptar, respectivamente, una llamada nueva y un handover del servicio i. Para una configuración particular de la política de admisión $(t_1, \ldots, t_{2N})$ estos coeficientes se pueden obtener como

$$\alpha_i^n(\boldsymbol{x}) = \begin{cases} 1 & \text{if } b(\boldsymbol{x}) + b_i \leq \lfloor t_i \rfloor \\ t_i - \lfloor t_i \rfloor & \text{if } b(\boldsymbol{x}) + b_i = \lfloor t_i \rfloor + 1 \\ 0 & \text{if } b(\boldsymbol{x}) + b_i > \lfloor t_i \rfloor + 1 \end{cases}$$

y

$$\alpha_i^h(\boldsymbol{x}) = \begin{cases} 1 & \text{if } b(\boldsymbol{x}) + b_i \leq \lfloor t_i \rfloor \\ t_{N+i} - \lfloor t_{N+i} \rfloor & \text{if } b(\boldsymbol{x}) + b_i = \lfloor t_{N+i} \rfloor + 1 \\ 0 & \text{if } b(\boldsymbol{x}) + b_i > \lfloor t_{N+i} \rfloor + 1 \end{cases}$$

De lo anterior se derivan las ecuaciones de balance globales

$$p(\boldsymbol{x}) \sum_{\boldsymbol{y} \in S} r_{\boldsymbol{x}\boldsymbol{y}} = \sum_{\boldsymbol{y} \in S} r_{\boldsymbol{y}\boldsymbol{x}} p(\boldsymbol{y}) \qquad \forall \boldsymbol{x} \in S. \tag{5.17}$$

Donde $p(\boldsymbol{x})$ es la probabilidad estacionaria del estado $\boldsymbol{x}$ que se obtienen a partir de (5.17) y de la ecuación de normalización. Para resolver este sistema de ecuaciones puede emplearse cualquier método estándar para resolver sistema de ecuaciones lineales. En nuestro caso se ha empleado el método de *Gauss-Seidel*. A partir de los valores de $p(\boldsymbol{x})$ se calculan las probabilidades de bloqueo

$$P_i = P_i^n = \sum_{\boldsymbol{x} \in S} \left(1 - \alpha_i^n(\boldsymbol{x})\right) p(\boldsymbol{x}) \qquad P_{N+i} = P_i^h = \sum_{\boldsymbol{x} \in S} \left(1 - \alpha_i^h(\boldsymbol{x})\right) p(\boldsymbol{x}).$$

Cuando el sistema está en equilibrio estadístico la tasas de llegada de las peticiones de handover pueden expresarse en función de las tasas de llegada de las llamadas nuevas y de las probabilidades de bloqueo (P_i) a través de la expresión [Jab96]

$$\lambda_i^h = \lambda_i^n \frac{1 - P_i^n}{\mu_i^c / \mu_i^r + P_i^h}. \tag{5.18}$$

Las probabilidades de bloqueo a su vez dependen de las tasas de llegada de handover por lo que (5.18) constituye un sistema de ecuaciones no lineales que puede resolverse aplicando un método iterativo de punto fijo como el descrito en [HR86, LMN94].

5.2.3. Algoritmo

El algoritmo aquí propuesto busca la configuración de la política MFGC que maximiza la capacidad del sistema y el valor de esta capacidad. Por

capacidad del sistema entendemos la máxima carga que el sistema puede soportar cumpliendo con los requisitos de QoS. Estos requisitos de QoS están especificados como cotas superiores para las probabilidades de bloqueo de las llamadas nuevas (B_i^n) y las probabilidades de terminación forzosa (B_i^{ft}). Si llamamos $\lambda^T = \sum_{1 \leq i \leq N} \lambda_i^n$ a la tasa agregada de llegada de llamadas nuevas y f_i a la fracción de λ^T que corresponde al servicio i, esto es $\lambda_i^n = f_i \lambda^T$, entonces podemos expresar de una manera formal el problema de optimización de la capacidad como sigue

Dados: $C, b_i, f_i, \mu_i^c, \mu_i^r, B_i^n, B_i^{ft}$; $i = 1, \ldots, N$

Maximizar: λ^T

buscando los parámetros de MFGC apropiados t_i; $i = 1, \ldots, 2N$

Sujeto a: $P_i^n \leq B_i^n, P_i^{ft} \leq B_i^{ft}$; $i = 1, \ldots, N$

Nuestra propuesta consisten en un algoritmo para resolver este problema de optimización de capacidad cuando se emplea una política MFGC. Nuestro algoritmo se compone de una parte principal (Algoritmo 1 `capacity`) desde la que se llama al procedimiento `solveMFGC` (Algoritmo 2) el cual, a su vez, llama a otro procedimiento (`MFGC`) para el cálculo de las probabilidades de bloqueo. Con el propósito de simplificar la notación se introduce el vector de cotas superiores para las probabilidades de bloqueo $\boldsymbol{p_{max}} = (B_1^n, \ldots, B_N^n, B_1^h, \ldots, B_N^h)$, en el que el valor de B_i^h viene dado por

$$B_i^h = \frac{\mu_i^c}{\mu_i^r} \frac{B_i^{ft}}{1 - B_i^{ft}} \tag{5.19}$$

Las figuras 5.9 a la 5.11 muestran un ejemplo de traza que ilustra el funcionamiento básico del algoritmo. En este ejemplo se ha tomado un escenario bastante simple (únicamente dos tipos de llegada) para permitir la representación gráfica (en dos dimensiones) de las trazas. Cada una de las figuras representa una ejecución del algoritmo `solveMFGC` para un valor fijo (λ^T). En cada una de estas ejecuciones los valores de los parámetros de la política (t_i)

Algoritmo 1 $(\lambda_{max}^{T}, \boldsymbol{t}_{opt})$=capacity$(\boldsymbol{p_{max}}, \boldsymbol{f}, \boldsymbol{\mu_c}, \boldsymbol{\mu_r}, \boldsymbol{b}, C)$

$\varepsilon_1 :=<$ precisión deseada $>$
$L := 0$
$U :=<$ high value $>$
(possible, $\boldsymbol{t}$) := solve_MFGC$(\boldsymbol{p_{max}}, U\boldsymbol{f}, \boldsymbol{\mu_c}, \boldsymbol{\mu_r}, \boldsymbol{b}, C)$
atLeastOnce:=FALSE;

while possible **do**
  $L := U$
  $\boldsymbol{t}_L := \boldsymbol{t}$
  atLeastOnce:=TRUE
  $U := 2U$
  (possible, $\boldsymbol{t}$) := solve_MFGC$(\boldsymbol{p_{max}}, U\boldsymbol{f}, \boldsymbol{\mu_c}, \boldsymbol{\mu_r}, \boldsymbol{b}, C)$
end while{esto asegura que $U > \lambda_{max}^{T}$}

repeat
  $\lambda := (L + U)/2$
  (possible, $\boldsymbol{t}$) := solve_MFGC$(\boldsymbol{p_{max}}, \lambda\boldsymbol{f}, \boldsymbol{\mu_c}, \boldsymbol{\mu_r}, \boldsymbol{b}, C)$
  if possible **then**
    $L := \lambda$
    $\boldsymbol{t}_L := \boldsymbol{t}$
    atLeastOnce:=TRUE;
  else
    $U := \lambda$
  end if
until $(U - L)/L \leq \varepsilon_1$ AND atLeastOnce
$\lambda_{max}^{T} := L$
$\boldsymbol{t} := \boldsymbol{t}_L$

Algoritmo 2 (possible, $\boldsymbol{t}$)=solveMFGC($\boldsymbol{p_{max}}, \boldsymbol{\lambda_n}, \boldsymbol{\mu_c}, \boldsymbol{\mu_r}, \boldsymbol{b}, C$)

INPUTS: $\boldsymbol{p_{max}}, \boldsymbol{\lambda_n}, \boldsymbol{\mu_c}, \boldsymbol{\mu_r}, \boldsymbol{b}, C$
OUTPUTS: possible, $\boldsymbol{t}$

```
 1:
 2: ε2 :=< desired precision >
 3: δ :=< small value >
 4: t := (δ, δ, ..., δ)
 5: p := MFGC(t, λn, μc, μr, b, C)
 6:
 7: repeat
 8:    canConverge:=TRUE;
 9:    i := 1;
10:
11:    repeat
12:       if p(i) > pmax(i) then
13:          t' := t; t'(i) := C
14:          p' := MFGC(t', λn, μc, μr, b, C)
15:
16:          if p'(i) > pmax(i) then
17:             canConvege:=FALSE;
18:          else
19:             L := t(i); U := C
20:             repeat
21:                t(i) := (L + U)/2
22:                p := MFGC(t, λn, μc, μr, b, C)
23:                if p(i) > pmax(i) then
24:                   L := t(i)
25:                else
26:                   U := t(i)
27:                end if
28:             until (1 − ε2)pmax(i) ≤ p(i) ≤ pmax(i)
29:          end if
30:
31:       end if
32:       i := i + 1
33:    until (i > 2N) OR ( NOT(canConverge))
34:
35:    if canConverge then
36:       if p(i) ≤ pmax(i)   ∀i then
37:          possible:=TRUE; exit:=TRUE;
38:       else
39:          exit:=FALSE;
40:       end if
41:    else
42:       possible:=FALSE; exit:=TRUE;
43:    end if
44:
45: until exit
```

se incrementan de forma secuencial ($i = 1, \ldots, 2N$; bucle *repeat*, líneas 11–33) y cíclica (bucle *repeat*, líneas 7–45) buscando cumplir las restricciones de QoS : $(1-\epsilon)\boldsymbol{p_{max}}(i) \leq \boldsymbol{p}(i) \leq \boldsymbol{p_{max}}(i)$. Tras inicializar cada t_i con un valor suficientemente pequeño (δ), la ejecución del bucle secuencial comienza con el flujo 1. Para este flujo se verifica si dando el valor máximo posible a t_1 ($t_1 = C$) es posible cumplir la restricción de calidad de servicio correspondiente a este flujo $\boldsymbol{p}(1) \leq \boldsymbol{p_{max}}(1)$. Si esto es posible se procede a calcular el valor adecuado de t_1, mientras que si no lo es se concluye que para el valor de λ^T que se había tomado no existe a una configuración factible que cumpla los requisitos de QoS, y el algoritmo termina. En el primer caso, es decir, si es posible encontrar un valor apropiado para t_1, se procede de idéntico modo con el flujo 2 excepto que ahora el valor utilizado para t_1 es el que acaba de calcularse. Este procedimiento continúa de forma cíclica con el resto de flujos hasta que el objetivo de QoS se satisface simultáneamente para todos los flujos ($\boldsymbol{p}(i) \leq \boldsymbol{p_{max}}(i) \quad \forall i$, véase la figura 5.9), en cuyo caso el algoritmo devuelve el valor possible=TRUE; o el algoritmo abandona porque se da cuenta de que el objetivo de QoS es inalcanzable (véase la figura 5.10), en cuyo caso el algoritmo devuelve el valor possible=FALSE. Cuando se da la primera de las posibilidades (possible=TRUE) sabemos que existe una configuración de la política de admisión que satisface los requisitos de QoS. Por tanto, $\lambda^T \leq \lambda^T_{max}$ y se puede probar un valor mayor para a λ^T. En la figura 5.11 se muestra una traza en la que el valor de λ^T era mayor que en la traza mostrada en la figura 5.9, de nuevo se encuentra una configuración para la que se cumple la especificación de QoS aunque, como puede observarse aumentar el valor de λ^T ha producido como efecto la disminución de la superficie del espacio de configuraciones que satisfacen las restricciones de QoS. Si `solveMFGC` termina porque el objetivo de QoS no es factible (possible=FALSE), se verifica que $\lambda^T > \lambda^T_{max}$. El algoritmo principal `capacity` consiste básicamente en una búsqueda binaria de λ^T_{max} que en cada iteración llama a `solveMFGC`.

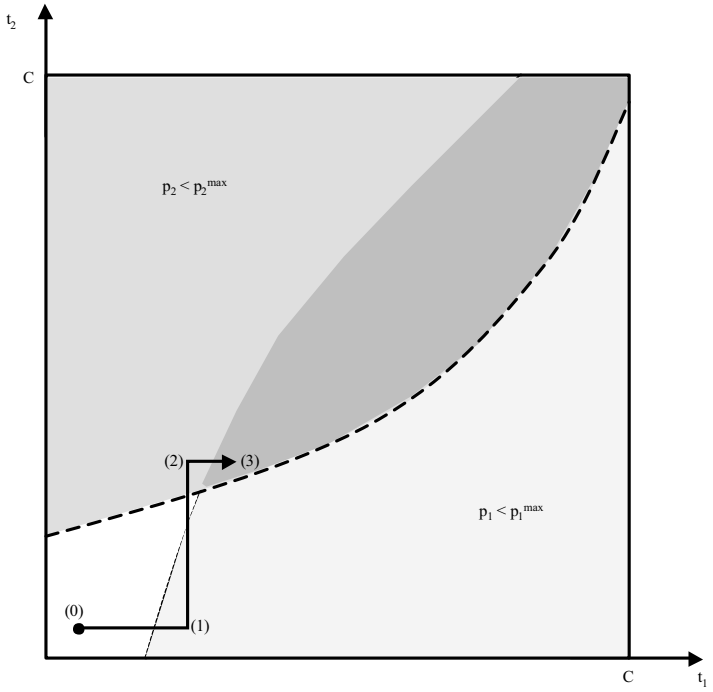


Figura 5.9: Traza gráfica de una ejecución de `solveMFGC`; $\lambda_1^T \leq \lambda_{max}^T$.

Sobre el procedimiento `MFGC`

El procedimiento `MFGC`, el cual es invocado desde el bucle más interno del algoritmo, se utiliza para calcular las probabilidades de bloqueo ($\boldsymbol{p} :=$ MFGC$(\boldsymbol{t}, \boldsymbol{\lambda_n}, \boldsymbol{\mu_c}, \boldsymbol{\mu_r}, \boldsymbol{b}, C)$). En este cálculo se emplea un procedimiento iterativo para calcular las tasas de las peticiones de handover. En cada iteración se tiene que resolver un sistema de nacimiento y muerte multidimensional que, en general, tendrá un elevado número de estados y constituye la principal contribución al coste computacional del algoritmo.

La siguiente observación se utiliza para acelerar el algoritmo pues ésta permite eliminar el mencionado procedimiento iterativo y, así, cada vez que

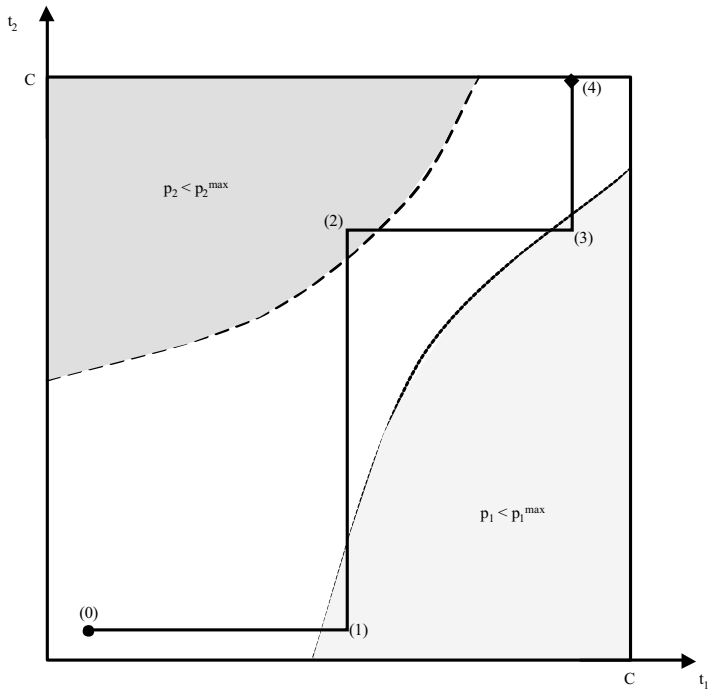


Figura 5.10: Traza gráfica de una ejecución de `solveMFGC`; $\lambda_3^T > \lambda_{max}^T$

se llame al procedimiento `MFGC` se resolverá una sola vez un sistema de ecuaciones lineales en vez de hacerlo varias veces (una por cada iteración). La observación se basa en que cada ejecución del procedimiento se intenta encontrar unos valores de $\boldsymbol{t}$ para los cuales $\boldsymbol{p} = \boldsymbol{p_{max}}$ (dentro del margen de tolerancia especificado). Por tanto, en vez de utilizar la expresión (5.18) para calcular λ_i^h proponemos utilizar la expresión

$$\lambda_i^h = \lambda_i^n \frac{1 - B_i^n}{\mu_i^c/\mu_i^r + B_i^h}. \tag{5.20}$$

Aunque (5.18) y (5.20) son aparentemente muy semejantes existe una diferencia sustancial entre ambas pues en (5.20) el valor de λ_i^h se define de forma explícita mientras que en (5.18) no es así ya que P_i^n and P_i^h dependen de λ_i^h.

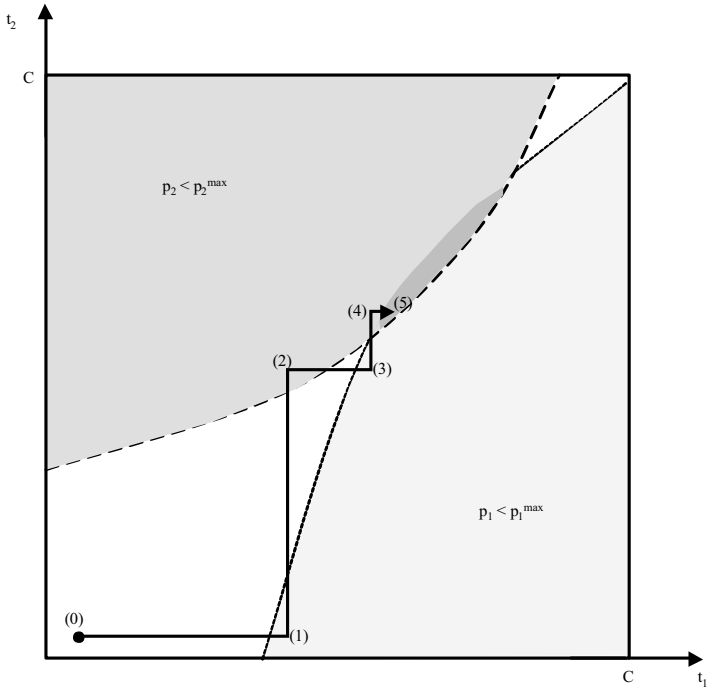


Figura 5.11: Traza gráfica de una ejecución de `solveMFGC`; $\lambda_1^T \leq \lambda_2^T \leq \lambda_{max}^T$

5.2.4. Evaluación numérica de la complejidad computacional

Para evaluar la complejidad computacional de nuestro algoritmo hemos tomado como referencia el algoritmo propuesto por Heredia et al. en [HUCPOG03c, HUCPOG03b, HUCPOG03a]. En adelante nos referiremos a este algoritmo como HCO, nombre tomado de las iniciales de sus autores.

El algoritmo HCO toma como entrada el *orden de prioridad* (*prioritization order*) óptimo, es decir una lista de los tipos de llamada ordenada según la priridad de cada tipo [HUCPOG03a]. Si $\boldsymbol{t}$ es la configuración de la política para la que se alcanza la máxima capacidad el orden de prioridad óptimo es la permutación $\sigma^* \in \boldsymbol{\Sigma}$, $\boldsymbol{\Sigma} := \{(\sigma_i, \ldots, \sigma_{2N}) : \ \sigma_i \in \mathbb{N}, 1 \leq \sigma_i \leq 2N\}$, tal que

$t(\sigma_1^*) \leq t(\sigma_2^*) \leq \ldots \leq t(\sigma_{2N}^*) = C$. Elegir el orden de prioridad óptimo no es una tarea fácil pues depende de la especificación de QoS así como de los parámetros del sistema. En general existirá un total de $(2N)!$ ordenes de prioridad distintos. En [HUCPOG03a] los autores proponen ciertos criterios que permiten construir un lista parcialmente ordenada de ordenes de prioridad según la probabilidad de ser el orden de prioridad óptimo. Posteriormente se sigue un procedimiento de ensayo y error tomando secuencialmente los elementos de esta lista parcialmente ordenada hasta que se encuentra el orden de prioridad óptimo. Para cada elemento de la lista que se ensaya se ejecuta el algoritmo y si éste no converge tras un número elevado de iteraciones se pasa al siguiente elemento.

El algoritmo que proponemos no requiere conocer *a priori* el orden de prioridad lo cual constituye por sí mismo una ventaja importante respecto al algoritmo HCO. Además, en los ejemplos numericos siguientes se demuestra que nuestro algoritmo es más eficiente aun cuando se proporciona como entrada a HCO el orden de prioridad óptimo.

En los ejemplos numericos se ha considerado un sistema con dos servicios ($N = 2$) y mientras no se indique los contrario se utilizarán los siguientes valores $\boldsymbol{b} = (1, 2)$, $\boldsymbol{f} = (0.8, 0.2)$, $\boldsymbol{\mu_c} = (1/180, 1/300)$, $\boldsymbol{\mu_r} = (1/900, 1/1000)$, $\boldsymbol{B}^n = (0.02, 0.02)$, $\boldsymbol{B}^{ft} = (0.002, 0.002)$; todas las tolerancias se han fijado a $\epsilon = 10^{-2}$. A partir de (5.19) tenemos que $\boldsymbol{B}^h \approx (0.01002, 0.00668)$ y $\boldsymbol{p_{max}} \approx (0.02, 0.02, 0.01002, 0.00668)$.

En la tabla 5.4 y en la figura 5.12(a) se muestra una comparación del número de operaciones de coma flotante (*flops*) empleadas en la ejecución de cada algoritmo. Aunque el algoritmo HCO original no incorpora la técnica de acelaración (véase (5.20) y comentario anterior en la página 156) sí es susceptible de hacerlo, por lo que para poder evaluar por separado la ganancia del algoritmo propuesto y de la técnica de acelaración, hemos considerado los cuatro casos que resultan de tomar el algoritmo HCO y nuestro algoritmo ambos con y sin la acelaración. Como puede observarse, en ambos casos la técnica de aceleración reduce el número de *flops* aproximadamente a la

Tabla 5.4: Comparación del algoritmo HCO (con orden de priorización optimo conocido) y nuestro algoritmo, con y sin aceleración (cifras en *Mflops*).

	HCO		nuestro algoritmo	
C	—	aceleración	—	aceleración
5	5.70	2.00	1.17	0.39
10	60.20	20.00	13.80	4.53
20	438.00	156.00	145.00	46.60

tercera parte.

Para evaluar el impacto de la movilidad sobre la complejidad computacional se ha tomado el ejemplo anterior y se han variado los parámetros μ_i^r. De este modo se han configurado las cuatro combinaciones de factores de movilidad (μ_i^r/μ_i^c) siguientes:

a. $\mu_1^r = 0.2\mu_1^c$, $\mu_2^r = 0.2\mu_2^c$

b. $\mu_1^r = 0.2\mu_1^c$, $\mu_2^r = 1\mu_2^c$

c. $\mu_1^r = 1\mu_1^c$, $\mu_2^r = 0.2\mu_2^c$

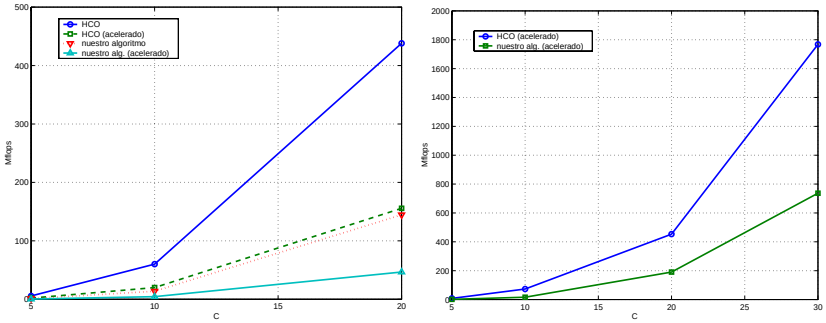
d. $\mu_1^r = 1\mu_1^c$, $\mu_2^r = 1\mu_2^c$

El coste computacional por escenarios se muestra en la tabla 5.5 y en la figura 5.12(b) se representa el agregado de los cuatro. Los resultados comparan nuestro algoritmo con el algoritmo HCO utilizando aceleración en ambos y proporcionando al algoritmo HCO con el orden de prioridad óptimo.

Es importante destacar que, tal y como era previsible, en todos los casos las diferencias entre los valores de la capacidad óptima que se han obtenido mediante distintos métodos estaban dentro de los márgenes de la tolerancia establecida en los algoritmos. Lo mismo puede decirse de los valores de configuración de la política, $\boldsymbol{t}$.

Tabla 5.5: Comparación del algoritmo HCO (con orden de priorización optimo conocido) y nuestro algoritmo para distintos factores de movilidad (cifras en *Mflops*).

		C				
		5	10	20	30	**Total**
HCO (aceleración)	a	2.08	17.54	74.33	407.74	501.69
	b	2.67	14.06	147.13	487.41	651.27
	c	1.12	24.54	110.41	410.93	547.00
	d	2.24	16.86	121.39	462.12	602.62
	Total	8.11	73.00	453.26	1768.20	2302.60
nuestro algoritmo (aceleración)	a	0.35	4.42	53.64	199.69	258.10
	b	0.34	3.87	43.01	172.73	219.95
	c	0.38	3.93	47.95	191.66	243.92
	d	0.31	3.93	45.92	172.58	222.74
	Total	1.39	16.15	190.51	736.66	944.71



(a) con y sin aceleración.

(b) con aceleración;valores agregado para los escenarios A,B,C y D.

Figura 5.12: Comparación del coste computacional del algoritmo HCO y el algoritmo propuesto.

5.3. Control de admisión óptimo empleando predicción de handovers

Tal y como ya hemos comentado anteriormente en varias ocasiones, en la literatura existen un buen número de estudios y propuestas sobre el control de admisión en redes celulares monoservicio y también, aunque en un número mucho menor, multiservicio. La mayoría de estos trabajos se proponen esquemas de CA basados en una reserva de recursos realizada de forma intuitiva o heurística y, en un grupo menor, esta reserva se realiza desde una perspectiva de optimización; en las secciones anteriores de este mismo capítulo se pueden encontrar ejemplos de esto último. En todos los casos referidos hasta ahora (monoservicio o multiservicio, con reserva heurística o basada en un criterio de optimización) el CA adopta la decisión de admitir o rechazar una petición basándose únicamente en el estado actual de la célula sobre la que actúa el CA. En las redes celulares con terminales móviles es posible disponer de un cierto conocimiento anticipado de peticiones futuras y, lo que es más importante, esta previsión afecta a las peticiones más sensibles o de mayor prioridad: las de handover. Siguiendo esta línea se han propuesto diferentes mecanismos de predicción de movilidad (PM) al cual se asocia un método de CA que aprovecha, de una forma heurística, la información proporcionada por la PM (véase por ejemplo [LAN97, CS98, HF01, YL02, SK04, ZK04] y sus referencias).

En este capítulo se estudia el CA desde una perspectiva de optimización y basado en información predictiva de handovers, con el objetivo de evaluar en qué medida mejoran las prestaciones del CA cuando se dispone de un cierto nivel de información predictiva. El único antecedente que conocemos en esta línea es el trabajo de Yener y Rose [YR97] en el que en un escenario monoservicio, se obtiene una política de admisión cuasi-óptima mediante un algoritmo genético que considera el estado de las célula (número de llamadas en curso) y también el de la células colindantes. Sin embargo, los resultados de este estudio revelan que la ganancia de utilizar esta información adicional

no es significativa. Utilizando un método de optimización de *mejoras sucesivas de la política (policy improvement)* [Ros70] que conduce a soluciones óptimas (en vez de cuasi-óptimas), nosotros hemos llegado a la misma conclusión. Estos resultados sugieren que la predicción de potenciales handovers obtenida a partir del nivel de ocupación de las células vecinas no es suficientemente específica, por lo que se ha decidido evaluar la ganancia que podría obtenerse si se dotaba al algoritmo de CA de una información más específica. En nuestro estudio hemos utilizado un modelo de agente predictor (AP) que divide el conjunto de los terminales activos en la inmediaciones de la célula en dos grupos: aquéllos que con una probabilidad alta se prevé que realizarán un handover hacia la célula y aquellos para los cuales la previsión es la contraria —muy probablemente no realizarán un handover—. Aunque estudiar modelos del AP que proporcionan mayor información, como por ejemplo una estimación del momento en que se va a producir el handover, es un aspecto de indudable interés, esto se ha dejado para un trabajo futuro. Por otra parte, aunque cuanto mayor sea la cantidad de información proporcionada al CA mejores serán las prestaciones obtenidas, también es cierto que la complejidad del AP y del proceso de optimización también aumentarán pudiendo llegar a ser irrealizables.

5.3.1. Descripción del modelo

El modelo del tráfico es el mismo que se ha descrito en 5.1.1 y 5.2.1. Al considerarse un único servicio se ha simplificado la notación eliminando el subíndice que se refería al servicio. Dado que en este estudio nuestro interés no es el diseño del AP sino la utilización de la información que éste proporciona para el CA, vamos a utilizar un modelo genérico de este elemento que describimos a continuación. El AP informa al CA del número de terminales activos en las proximidades de la célula de los que se prevé van a realizar un handover hacia ésta. El tiempo transcurrido desde que se realiza la predicción hasta que efectivamente se produce el handover —o se descarta definitivamente esta posibilidad— lo modelamos mediante una variable aleatoria con

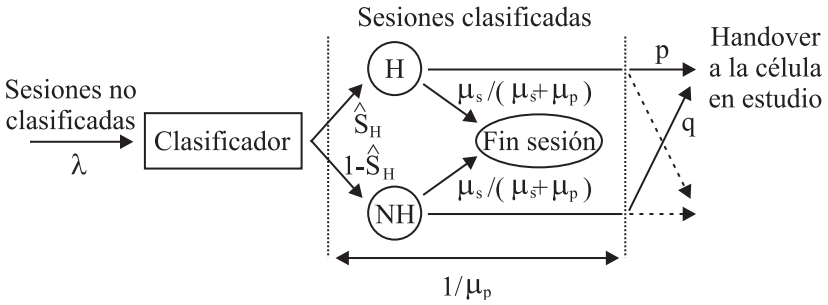


Figura 5.13: Modelo del agente predictor (AP): diagrama de funcionamiento.

una distribución exponencial. En nuestro modelo del AP suponemos que la clasificación de los terminales activos no es totalmente precisa sino que existe cierta probabilidad de error que modelamos mediante las probabilidades de *no detección* q, y *falso positivo* $1 - p$. Esto se muestra de forma esquemática en la figura 5.13.

En el momento en el que un terminal activo entra en la zona próxima a la célula, o uno que ya estaba en esta zona pasa a estar activo, el AP lo clasifica como H (probablemente va a producir un handover) o NH (probablemente no va a producir un handover) atendiendo a las características del terminal (posición, trayectoria, velocidad, perfil histórico, ...) y otra información (mapa de calles y carreteras, hora del día, ...). Tras un tiempo aleatorio, que en nuestro modelo sigue una distribución exponencial, el destino final del terminal se concreta y bien acaba produciéndose el handover o esta posibilidad se descarta definitivamente (por ejemplo porque el terminal se desplaza a otra célula) y, en su caso, el terminal deja de estar en el grupo de los clasificados como H. Un terminal también sale del grupo de los clasificados como H cundo su sesión termina. El sistema de CA conoce en todo momento la cantidad de terminales clasificados como H.

El modelo del AP se caracteriza mediante tres parámetros: el tiempo medio desde que se realiza la clasificación hasta que esta predicción se concreta μ_p^{-1}, la probabilidad p de provocar un handover si el terminal se ha etique-

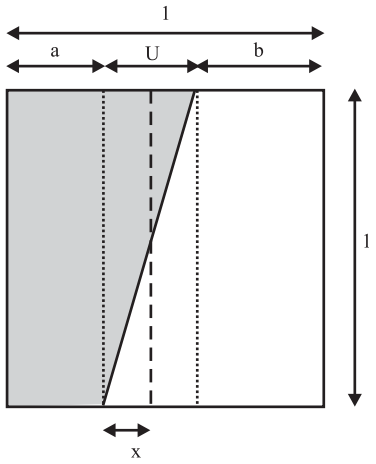


Figura 5.14: Modelo del agente predictor (AP): incertidumbre en la predicción.

tado como H y la probabilidad q de provocar un handover si el terminal se ha etiquetado como NH. Nótese que en general $q \neq 1 - p$. Los valores de p y q están relacionados a través del modelo del AP (ver figura 5.14). En la figura se muestra un cuadrado (de superficie unidad) que representa la población de los terminales que van a ser clasificados por el AP. El área sombreada representa la fracción de los terminales que finalmente realizarán un handover y el área no sombreada el resto de terminales. El clasificador fija un umbral (línea discontinua vertical en la figura) que utiliza para discriminar entre aquellos terminales que probablemente realizarán un handover (a la izquierda de la línea) y los que no (a la derecha de la línea). En la figura puede apreciarse que existe una zona de incertidumbre de anchura U ya que la línea de separación entre la zona sombreada y la zona no sombreada es oblicua, mientras la decisión del clasificador adopta la forma de un corte vertical. Esta incertidumbre es la causante de los errores de clasificación: la zona triangular no sombreada a la izquierda del umbral (falso positivo) y el triángulo sombreado a la derecha del umbral (no detección). El parámetro x representa la posición relativa dentro de la zona de incertidumbre del umbral

del clasificador. Referida a la figura 5.14 introducimos la notación siguiente: S_H denota la superficie sombreada y por tanto es la fracción del terminales que realizarán un handover; $\hat{S}_H$ denota la superficie a la izquierda del umbral y es la fracción de terminales clasificados como H; $\hat{S}^e_H$ denota la superficie no sombreada a la izquierda del umbral y es la fracción de terminales clasificados como H que no realizarán un handover; $\hat{S}^e_{NH}$ denota la superficie sombreada a la derecha del umbral y es la fracción de terminales clasificados como NH que sí producirán un handover. A partir de la figura 5.14 se obtiene fácilmente que:

$$1 - p = \frac{\hat{S}^e_H}{\hat{S}_H} = \frac{x^2}{2U(a+x)} \; ; \qquad q = \frac{\hat{S}^e_{NH}}{1-\hat{S}_H} = \frac{(U-x)^2}{2U(1-a-x)}$$

Los parámetros a y b de la figura pueden expresarse en función de la fracción de termales que realizarán un handover S_H y del grado de incertidumbre en la predicción U,

$$a = S_H - U/2 \; ; \qquad b = 1 - S_H - U/2$$

y, por tanto

$$1 - p = \frac{\hat{S}^e_H}{\hat{S}_H} = \frac{x^2}{U(2S_H - U + 2x)} \; ; \qquad q = \frac{\hat{S}^e_{NH}}{1-\hat{S}_H} = \frac{(U-x)^2}{U(2 - 2S_H + U - 2x)}$$

5.3.2. Optimización de la politica de admisión

La función de coste que se utiliza para comparar las distintas políticas y definir la política óptima es una suma ponderada de la tasa de pérdidas para cada flujo de llegada: sesiones nuevas y handovers. Como modelo del sistema se utiliza un *proceso de decisión markoviano* (MDP) [Ros70] y el proceso de optimización se realiza mediante la técnica de *programación dinámica* conocida como *mejoras sucesivas de la política (policy improvement)* [Ros70].

Dado que aquí sólo consideramos un servicio simplificamos la notación omitiendo el subíndice que se refiere al tipo de servicio, esto es, $\lambda_n = \lambda^n_1$,

$\lambda_h = \lambda_1^h$, $\mu_c = \mu_1^c$, $\mu_r = \mu_1^r$, $\mu = \mu_1$. Además, suponemos, sin pérdida de generalidad, que $b = b_1 = 1$.

Representamos el estado del sistema mediante el par de números naturales (i, j), donde i es el número de sesiones activas en la célula y j es el número de terminales clasificados como H. El conjunto de estados posibles del sistema es

$$S := \left\{ \boldsymbol{x} = (i, j) : 0 \leq i \leq C;\ 0 \leq j \leq C_p \right\}$$

donde C_p representa el número máximo de terminales que puede haber en el conjunto de los clasificados como H en un momento dado. Para este parámetro C_p utilizaremos en todos los casos un valor lo suficientemente alto de manera que en la práctica no tenga ningún impacto en los resultados numéricos. Para cada estado $(i, j), i < C$, el conjunto de acciones posibles es $A := \{a : a = 0, 1\}$, siendo $a = 0$ la acción que rechaza la petición de una nueva sesión y $a = 1$ la acción que la acepta. Las peticiones de handover tienen prioridad sobre las sesiones nuevas por lo que una petición de handover se aceptará siempre que haya suficientes recursos disponibles, es decir, si $i < C$. En los estados (C, j) únicamente es posible la acción $a = 0$.

La figura 5.15 muestra las transiciones desde y hacia el estado (i, j). Nótese que algunas tasas de transición dependen de la decisión $a = 0, 1$. En la figura se ha introducido el parámetro λ_h' que representa la tasa de llegada de handovers no previstos, y cuyo valor puede expresarse como

$$\lambda_h' = (1 - \hat{S}_H) \frac{\mu_p}{\mu_p + \mu_c} q\lambda$$

donde λ es la tasa de entrada al AP.

Para convertir el proceso de Markov de tiempo continuo a una cadena de Markov de tiempo discreto equivalente empleamos la técnica de uniformización [Wol89, Section 4.7]. Es fácil demostrar que $\Gamma = C_p(\mu_p + \mu_c) + C(\mu_r + \mu_c) + \lambda + \lambda_n$ es una cota superior uniforme de la tasa saliente de cualquier estado. Si $r_{\boldsymbol{xy}}(a)$ representa la tasa de transición desde el estado $\boldsymbol{x}$ al estado $\boldsymbol{y}$

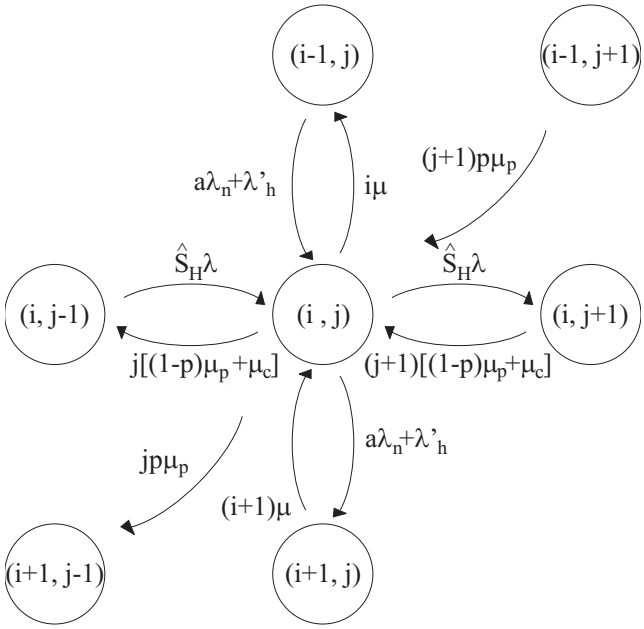


Figura 5.15: Diagrama de transiciones.

cuando se adopta la acción a, entonces las probabilidades de transición de la cadena resultante son

$$p_{xy}(a) = \frac{r_{xy}(a)}{\Gamma} \quad \text{si } \boldsymbol{y} \neq \boldsymbol{x} \qquad \text{y} \qquad p_{xx}(a) = 1 - \sum_{y \in S} p_{xy}(a).$$

El coste incurrido cuando en el estado $\boldsymbol{x}$ se adopta la acción a es

$$C(\boldsymbol{x}, a) = \begin{cases} (1-a)\lambda_n, & i < C,\ a = 0,1 \\ \lambda_n + \beta(\lambda'_h + jp\mu_p), & i = C,\ a = 0 \end{cases}$$

El factor de ponderación β (> 1) representa el hecho de que rechazar una petición de handover es menos deseable que rechazar una sesión nueva. Los costes se han definido de manera que el coste medio esperado $L(\pi)$ sea igual a la suma ponderada de la tasa de pérdidas de sesiones nuevas $L_n(\pi)$ y de

handovers $L_h(\pi)$,

$$L(\pi) = L_n(\pi) + \beta L_h(\pi) = \lim_{n\to\infty} E\left[\frac{1}{n+1}\sum_{t=0}^{n} C\big(\boldsymbol{x}(t), \pi(\boldsymbol{x}(t))\big)\right]$$

donde $\boldsymbol{x}(t)$ representa el estado del sistema en el instante t cuando se está aplicando la política π. Las tasas de pérdidas pueden expresarse como

$$L_n(\pi) = \sum_{\boldsymbol{x}:\ \pi(\boldsymbol{x})=0} \lambda_n p(\boldsymbol{x})\ ; \qquad L_h(\pi) = \sum_{\substack{\boldsymbol{x}=(C,j)\\ 0\le j\le C_p}} \big(\lambda'_h + jp\mu_p\big)p(\boldsymbol{x})$$

donde $p(\boldsymbol{x})$ es la probabilidad estacionaria del estado $\boldsymbol{x}$. Por tanto, el objetivo del problema de optimización es encontrar la política π^* que minimiza $L(\pi)$. Dado que el estado $(0,0)$ es alcanzable desde cualquier otro estado independientemente de la política que se aplique, en virtud del Corolario 6.20 de [Ros70,] y el comentario subsiguiente, podemos asegurar la existencia de una política óptima.

Si $l_{\boldsymbol{x}}(\pi)$ denota el coste relativo del estado $\boldsymbol{x}$ cuando se aplica la política π, podemos escribir que

$$l_{\boldsymbol{x}}(\pi) = C(\boldsymbol{x}, \pi(\boldsymbol{x})) - L(\pi) + \sum_{\boldsymbol{y}} p_{\boldsymbol{xy}}(\pi(\boldsymbol{x}))l_{\boldsymbol{y}}(\pi) \qquad \forall \boldsymbol{x} \qquad (5.21)$$

y de aquí obtener el coste medio $L(\pi)$ y los costes relativos $l_{\boldsymbol{x}}(\pi)$ a falta de una constante. Para deshacernos de dicha constante, arbitrariamente fijamos $l_{(0,0)}(\pi) = 0$ y resolvemos el sistema de ecuaciones lineales (5.21) para obtener $L(\pi)$ y $l_{\boldsymbol{x}}(\pi)$, $\forall \boldsymbol{x}$. Una vez calculados los costes relativos correspondientes a la política π, podemos calcular una política mejorada π' como

$$\pi'(\boldsymbol{x}) = \arg\min_{a=0,1} \Big\{C(\boldsymbol{x}, a) - L(\pi) + \sum_{\boldsymbol{y}} p_{\boldsymbol{xy}}(a)l_{\boldsymbol{y}}(\pi)\Big\}$$

de manera que se cumple que $L(\pi') \le L(\pi)$. Es más, en caso de que se dé la igualdad se tiene que $\pi' = \pi = \pi^*$, donde π^* denota la política optima, es decir, $L(\pi^*) \le L(\pi)\ \forall \pi$.

La repetición iterativa de este ciclo (resolución del sistema (5.21) y mejora de política) hasta que se obtiene una política que no cambia tras la mejora

se conoce también como *iteración de política* [Put94, sección 8.6] y conduce a la política óptima en un numero finito —y generalmente pequeño— de iteraciones.

5.3.3. Resultados numéricos

Para evaluar la mejora de prestaciones que supone incorporar la información predictiva en el proceso de CA, se compara el coste medio esperado $L(\pi)$ para las políticas óptimas con y sin la información predictiva. En el caso en el que no se considera la predicción la optimización se realiza ignorando la segunda componente del estado del sistema (el número de terminales clasificados como H) y utilizando que sabemos [RNT97] que la política óptima es del tipo *guard channel*.

En los resultados numéricos que se presentan a continuación se ha utilizado un escenario básico a partir del cual se va variando el valor de los distintos parámetros de interés para explorar cuál es su influencia en las prestaciones. Los parámetros de este escenario básico toman los valores siguientes: $C = 10$, $C_p = 60$, $N_h = \mu_r/\mu_c = 2$, $\mu_p^{-1}/\mu_r^{-1} = 0.5$, $\beta = 20$, $x = U/2$, $S_H = 0.4$, $\lambda_n = 1$. El valor de λ se ha elegido de forma que en el sistema haya un equilibrio de flujos [Jab96], esto es, para que la tasa de handovers que entran y salen de una célula coincidan,

$$\lambda = \frac{1}{S_H}\frac{\mu_c + \mu_p}{\mu_p}(1 - P_n)(1 - P_{ft})N_h\lambda_n$$

donde P_n es la probabilidad de bloqueo de una nueva sesión y P_{ft} la probabilidad de terminación forzosa. Tomando valores típicos para estas probabilidades ($P_n = 10^{-2}$, $P_{ft} = 10^{-3}$) obtenemos que

$$\lambda \approx 0.989\frac{1}{S_H}\left(N_h + \frac{\mu_r}{\mu_p}\right)\lambda_n.$$

En las figuras de la 5.16 a la 5.19 las curvas representan el cociente de la ponderación de la tasa de pérdidas sin predicción y con predicción. Como era

previsible, la utilización de predicción induce una mejora de las prestaciones en todos los casos y esta mejora disminuye cuando el grado de incertidumbre de la predicción (U) aumenta. En la figura 5.16 se muestran los resultados cuando varía el número medio de handovers por sesión. En la figura 5.17 se ha variado el factor de ponderación β que cuantifica la mayor importancia que se concede a los handovers frente a la sesiones nuevas. En esta figura se observa que cuanto mayor es esta importancia relativa, mayor es el beneficio que se obtiene de utilizar la predicción. La posición relativa del umbral de decisión dentro de la zona de incertidumbre se evalúa en la figura 5.18. La curvas de esta figura muestran que la mejor elección es la posición intermedia. Finalmente, la figura 5.19 muestra el efecto del tiempo medio de antelación con el que se dispone de la predicción (normalizado respecto al tiempo medio de permanencia en una célula). Tanto los valores bajos como los altos tienen un impacto negativo en las prestaciones; mientras que lo primero parece lógico lo segundo podría parecer contradictorio. Para explicar este efecto aparentemente contradictorio se debe tener en cuenta lo siguiente: en el modelo el tiempo de antelación no es determinista y por tanto al aumentar su valor medio también aumenta su variabilidad (cuantificada por ejemplo con la varianza); por otra parte, al aumentar el tiempo que media entre el instante en el que se produce la predicción hasta que el movimiento del terminal concreta esta predicción, aumenta la probabilidad de que la sesión termine antes de que esto último ocurra y, por tanto, la predicción es de algún modo más incierta.

En todos los casos examinados la política óptima presenta una estructura en forma de *guard channel* dinámico, en la que el número de canales reservados aumenta con el número de terminales en el conjunto de los clasificados como H. De un modo más formal, si $p(i,j)$ es la probabilidad de aceptar una nueva sesión cuando el sistema está en el estado (i,j), se cumple que

$$p(i,j) = \begin{cases} 1, & \text{si } i \leq i_{th}(j) \\ 0, & \text{si } i > i_{th}(j) \end{cases} \qquad \text{y} \quad i_{th}(j) \leq i_{th}(j') \quad \text{si} \quad j > j',$$

donde $i_{th}(j)$ es el umbral para un para un valor dado de j.

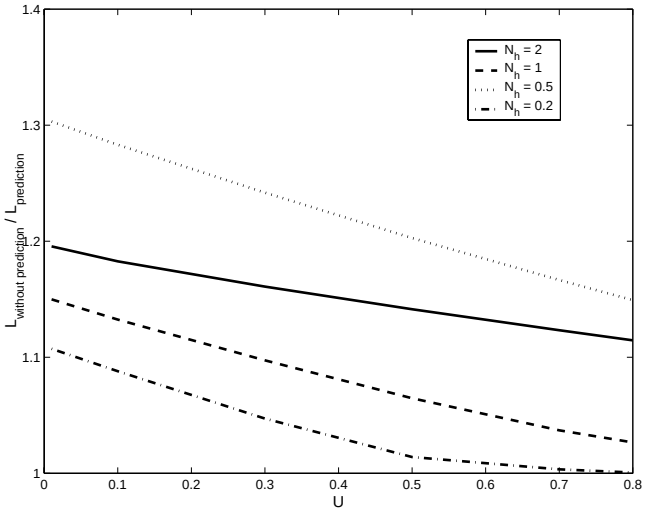


Figura 5.16: Influencia de la movilidad, N_h.

5.4. Conclusiones

En este capítulo se han considerado distintas políticas para el control de admisión en redes celulares multiservicio. Entre estas políticas se incluyen la generalización a un entorno multiservicio de las más populares en el ámbito monoservicio (GC y FGC) y la familia más general de la políticas estacionarias aleatorizadas (RS). En general se concluye que las políticas del tipo RS presentan ciertas ventajas ya que, por una parte permiten conseguir una mayor capacidad para una misma cantidad de recursos — especialmente cuando éstos son reducidos— y, por la otra, el ajuste de los parámetros de este tipo de políticas puede formularse de una forma más adecuada para el diseño mediante la aplicación de la teoría de los procesos de decisión markovianos junto con técnicas de programación lineal. Dentro de este marco proponemos diferentes métodos para el diseño de políticas de admisión del tipo RS que son útiles no sólo para el dimensionado de la red sino también para la fase de operación.

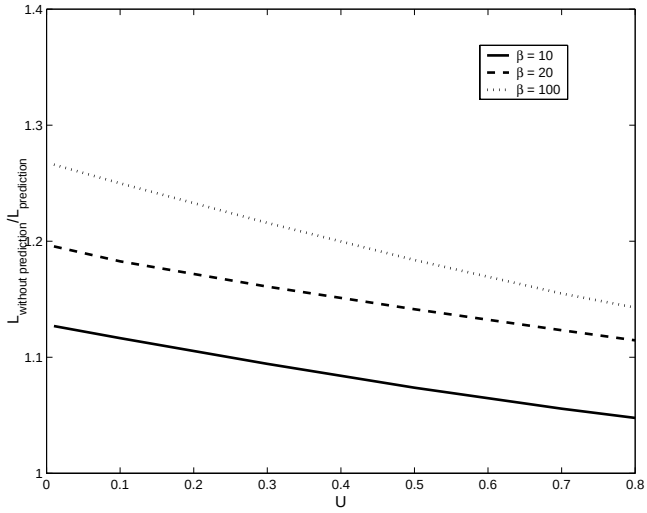


Figura 5.17: Influencia del factor de ponderación, β.

Por otra parte, la mejor política dentro de la familia MFGC consigue una capacidad que sin ser la óptima está muy próxima al valor óptimo, característica ésta que se acentúa todavía más al aumentar los recursos del sistema. Además las políticas del tipo MFGC presentan la ventaja frente a las del tipo RS de necesitar un menor número de parámetros para describir la política. Sin embargo, los métodos de optimación basados en la formulación de un programa lineal no sirven para optimizar la política dentro de la familia MFGC. Estas razones hacen interesante disponer de un método para realizar el ajuste óptimo de los parámetros de una política MFGC. En este capítulo se expone un algoritmo que realiza esta función y que es más simple y computacionalmente más eficiente que el único precedente encontrado en la literatura.

Finalmente se explora el impacto de incorporar información proveniente de una predicción de futuros handovers al proceso de optimización de la política de control de admisión. Utilizando un modelo de agente predictor imperfecto (que comete errores en la predicción) se ha realizado una evaluación

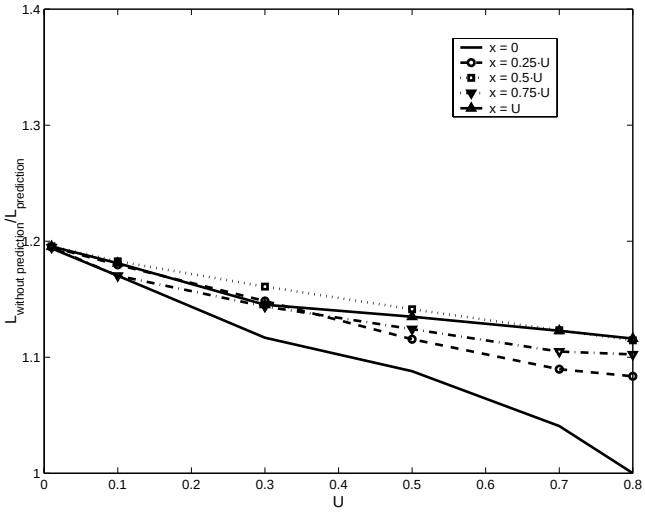


Figura 5.18: Influencia del umbral de decisión, x/U.

numérica que cuantifica la mejora de las prestaciones en unos valores típicos entorno al 10 %, aunque en algunos escenarios llegan a alcanzarse mejoras de hasta el 30 %. Aunque estos resultados son en gran medida dependientes del modelo de predictor utilizado, el modelo general y la metodología empleados son fácilmente adaptables a otros modelos de predictor.

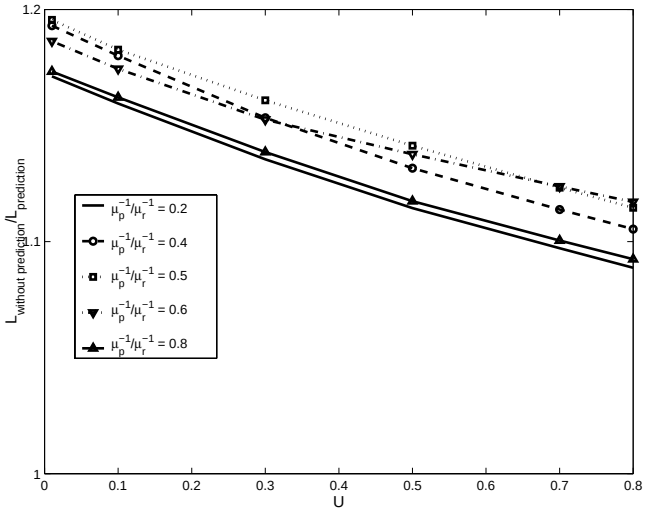


Figura 5.19: Influencia del tiempo de anticipación, μ_p^{-1}/μ_r^{-1}.

Capítulo 6

Conclusiones

Durante los últimos años las comunicaciones móviles han experimentado un enorme y rápido crecimiento, y es previsible que esta tendencia continué al menos durante un tiempo. Este crecimiento, sobre todo el que se prevé, no se refiere únicamente a un aumento del número de usuarios o del volumen de tráfico "tradicional" por usuario, sino que también incluye —y en parte está causado por— un aumento de la diversidad de servicios ofrecidos sobre las redes móviles celulares. Para hacer compatible este crecimiento con las inherentes limitaciones del espectro radioeléctrico es necesario un diseño eficiente de, entre otros, los mecanismos de gestión de recursos en la interfaz radio.

Por otra parte el análisis y diseño de este tipo de mecanismos no es una tarea trivial pues debe enfrentarse a numerosas dificultades que en términos generales incluyen, al menos: la impredecibilidad del tráfico, de la movilidad de los terminales y de las condiciones de propagación; y la interacción de distintos servicios con diferentes características y requisitos de calidad de servicio. Probablemente, es por esta complejidad también creciente que en la literatura especializada se observa un aumento de los estudios basados en simulaciones. Sin embargo, aunque un enfoque basado en simulación por eventos discretos es capaz de abordar el estudio de sistemas cuya com-

plejidad sería difícilmente abordable mediante un modelo analítico, creemos que es igualmente importante disponer de modelos analíticos cuya aplicación complemente a la de los modelos de simulación. En general los modelos analíticos son más adecuados para descubrir tendencias generales y proporcionar resultados cualitativos que facilitan una mejor comprensión del funcionamiento del sistema. Además, este mejor conocimiento es útil como guía para diseñar los experimentos adecuados a realizar mediante simulación o la construcción de un prototipo, experimentos que en general son más costosos en tiempo y recursos. Con este cometido, en esta tesis se ha pretendido contribuir al desarrollo de modelos analíticos para la evaluación de la gestión de recursos radio en redes móviles celulares. En particular, se han abordado los siguientes aspectos. En el capítulo 3 se ha propuesto y analizado de una forma unificada, una familia de algoritmos para asignar recursos en redes celulares otorgando prioridad al tráfico de peticiones de handover. Además de un método de análisis del tipo geométrico-matricial se ha desarrollado también un análisis basado en técnicas espectrales matriciales. En determinadas situaciones esta última metodología ofrece una precisión mayor y un coste computacional inferior. En el capítulo 4 se han estudiado diversos aspectos relacionados con la permanencia del terminal móvil en el área de handover y sus repercusiones en los modelos para la evaluación de la gestión de recursos. Se ha desarrollado un método analítico-numérico para caracterizar estadísticamente el tiempo de residencia en el área de handover y de ocupación de los recursos mientras el móvil está en esta zona. Se evalúa también la sensibilidad de los parámetros de medida de prestaciones frente a la variabilidad del tiempo de residencia en el área de handover. Finalmente, se desarrolla un modelo de colas que permite evaluar la aplicación de distintas disciplinas de servicio (FIFO, LIFO o SIRO), y de gestión del espacio de almacenamiento, a las peticiones de handover que por no poder ser atendidas inmediatamente esperan mientras el terminal permanece en el área de handover. En el capítulo 5 se ha abordado el diseño del control de admisión en redes celulares como un problema de optimización. En este contexto se proponen distintos criterios de optimización para redes multiservicio que dan lugar a la formu-

lación de un problema de programación lineal. Además, hemos propuesto un algoritmo para optimizar la política de control de admisión dentro de un conjunto más restringido, el de las políticas del tipo *trunk reservation*. Por último, se estudia la ganancia que puede obtenerse si se dota al proceso de optimización con una predicción sobre la llegada de peticiones de handover.

Desde la perspectiva en la que nos sitúa el trabajo desarrollado en esta tesis nos aventuramos a apuntar algunas líneas de investigación —dentro del ámbito en el que se inscribe este trabajo— que consideramos relevantes a tenor de las tendencias que se vislumbran o comienzan a observarse en las redes de acceso móviles. El aumento y la diversificación de servios ofrecidos sobre este tipo de redes conlleva una reducción de la proporción del tráfico de voz a favor del tráfico multimedia. A medida que el uso de estos nuevos servicios se extienda se dispondrá de medidas de tráfico cuyo análisis será necesario para validar los modelos actuales de tráfico y movilidad o para desarrollar otros más adecuados. La introducción de tráfico de datos, la introducción de sistemas de modulación y codificación adaptativos para el tráfico de tiempo real y la utilización de la conmutación de paquetes hacen conveniente, si no necesario, realizar la gestión de recursos también a un nivel de paquete, por lo que es importante desarrollar métodos y modelos para diseñar y evaluar las prestaciones de estos mecanismos. Por otra parte, para el nivel de sesión se necesitan nuevos mecanismos de gestión de recursos y de los modelos que se utilizan para su estudio ya que algunas características del nuevo escenario tecnológico lo diferencian sustancialmente del anterior en al menos dos aspectos: la utilización de técnicas de acceso múltiple que están limitadas por interferencia, como CDMA, introducen una fuerte interdependencia entre la gestión de recursos llevada a cabo en células próximas; con la aparición de escenarios multiacceso en los que coexisten y cooperan distintas redes de acceso radio (GSM, GPRS, UMTS, WLAN,...) con terminales capaces de utilizar diferentes tecnologías de acceso, será fundamental realizar un gestión de los recursos radio coordinada entre las distintas redes.

Apéndices

Apéndice A

Notación, variables y parámetros más utilizados

b_i	número de recursos necesarios para cursar una petición
$1/\mu_i$	tiempo medio de ocupación de los recursos por la conexión
$1/\mu_i^r$	tiempo medio de residencia en una célula
$1/\mu_i^c$	tiempo medio de duración de la conexión
λ_i^n	tasa de llegadas de peticiones correspondientes a llamadas nuevas
λ_i^h	tasa de llegadas de peticiones originadas por un *handover*
P_i^n	probabilidad de que una llamada nueva no sea admitida
P_i^h	probabilidad de que una petición de *handover* no sea admitida
P_i^{ft}	probabilidad de que una petición de *handover* no sea admitida

$\lfloor x \rfloor$	mayor de los enteros menores o iguales que x
$\lceil x \rceil$	menor de los enteros mayores o iguales que x
x_i	elemento i-ésimo del vector $\boldsymbol{x}$
M_{ij}	elemento de la fila i, columna j de la la matriz $\boldsymbol{M}$
M^t	matriz traspuesta de la matriz $\boldsymbol{M}$
$\otimes$	producto de Kronecker [LR99, p. 17] de dos matrices (ejemplo: $A \otimes B$)
diag$\{\cdot\}$	operador que devuelve una matriz diagonal con los elementos del

	vector argumento
$\boldsymbol{e}$	vector columna de unos
$\boldsymbol{e}_i$	vector cuyas entradas son todas nulas excepto la i-ésima que es 1
$\mathbf{0}$	vector columna de ceros
$\bar{\bar{\mathbf{0}}}$	matriz de ceros
$\boldsymbol{I}$	matriz identidad

La dimensión de los vectores y matrices: $\boldsymbol{e}$, $\mathbf{0}$, $\bar{\bar{\mathbf{0}}}$, $\boldsymbol{I}$, ..., será la apropiada en cada caso. Cuando sea necesario especificarla para evitar ambigüedad o facilitar la lectura, ésta se indicará mediante un subíndice.

Apéndice B

Abreviaturas y acrónimos

CAC	Call Admission Control
CHT	Channel Holding Time
CRT	Cell Residence Time
CV	Coefficient of Variation $CV_X = \sigma_X / E[X]$
DCA	Dynamic Channel Allocation
DP	Dynamic Programming
EDGE	Enhanced Data rates for Global Evolution
ETSI	European Telecommunications Standars Institute
FGC	Fractional Guard Channel
FIFO	First In First Out
GC	Guard Channel
GPRS	General Packet Radio Service
GSM	Global System for Mobile communications
HART	Handover Area Residence Time
HSCSD	High-Speed Circuit-Swithced Data
LEO	Low Earth Orbit
LIFO	Last In First Out
MAP	Markovian Arrival Process
MDP	Markov Decision Process

MFGC	Multiple Fractional Guard Channel
MMPP	Markov Modulated Poisson Process
PH	PHhase type
QBD	Proceso cuasi de nacimiento y muerte (Quasi-Birth-and-Death)
SIRO	Service In Random Order
TIC	Tecnologías de la Información y las Comunicaciones
W-CDMA	Wideband Code Division Multiple Access
WLAN	Wireless Local Area Network

Apéndice C

Matrices-bloque del generador infinitesimal de la sección 3.1.4

C.1. Algoritmo FGC

C.1.1. Matrices $A_0^{(i)}$ ($i = -1, \ldots, Q_n - 1$)

$i = -1$; dimensión $= m \times (n + 2 + Q_h)$

$$A_0^{(-1)} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \\ \lambda & 0 & \cdots & 0 \end{bmatrix}.$$

$i = 0, \ldots, Q_n - 1$; dimensión $= (n + 2 + Q_h) \times (n + 2 + Q_h)$

$$A_0^{(i)} = \lambda_n \begin{bmatrix} 1 - f & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}$$

C.1.2. Matrices $A_1^{(i)}$ ($i = -1, \ldots, Q_n$)

$\underline{i = -1; \qquad \text{dimensión} = m \times m}$

$$A_1^{(-1)} = \begin{bmatrix} * & \lambda & & & & & \\ \mu & * & \lambda & & & & \\ & 2\mu & * & \lambda & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & (m-2)\mu & * & \lambda \\ & & & & (m-1)\mu & * \end{bmatrix}$$

$\underline{i = 0; \qquad \text{dimensión} = (n+2+Q_h) \times (n+2+Q_h)}$

$$A_1^{(0)} = \begin{bmatrix} * & \lambda_h + f\lambda_n & & & & & & & \\ (m+1)\mu & * & \lambda_h & & & & & & \\ & \ddots & \ddots & \ddots & & & & & \\ & & C\mu & * & \lambda_h & & & & \\ & & & C\mu+\gamma & * & \lambda_h & & & \\ & & & & \ddots & \ddots & \ddots & & \\ & & & & & C\mu+(Q_h-1)\gamma & * & \lambda_h \\ & & & & & & C\mu+Q_h\gamma & * \end{bmatrix}$$

$\underline{i = 1, \ldots, Q_n; \qquad \text{dimensión} = (n+2+Q_h) \times (n+2+Q_h)}$

$$A_1^{(i)} = \begin{bmatrix} * & \lambda_h + f\lambda_n & & & & & & & & \\ (1-f)(m+1)\mu & * & \lambda_h & & & & & & & \\ & (m+2)\mu & * & \lambda_h & & & & & & \\ & & \ddots & \ddots & \ddots & & & & & \\ & & & C\mu & * & \lambda_h & & & & \\ & & & & C\mu+\gamma & * & \lambda_h & & & \\ & & & & & \ddots & \ddots & \ddots & & \\ & & & & & & \ddots & * & \lambda_h \\ & & & & & & & C\mu+Q_h\gamma & * \end{bmatrix}$$

Los elementos de la diagonal de $\boldsymbol{A}_1^{(i)}$, que se han representados mediante asteriscos, toman los valores necesarios para que las filas de $\boldsymbol{Q}$ sumen cero, es decir, $\boldsymbol{A}_1^{(-1)}\boldsymbol{e} + \boldsymbol{A}_0^{(-1)}\boldsymbol{e} = \boldsymbol{0}$ y $\boldsymbol{A}_2^{(i)}\boldsymbol{e} + \boldsymbol{A}_1^{(i)}\boldsymbol{e} + \boldsymbol{A}_0^{(i)}\boldsymbol{e} = \boldsymbol{0}$ $(i = 0, \ldots, Q_n)$.

C.1.3. Matrices $\boldsymbol{A}_2^{(i)}$ $(i = 0, \ldots, Q_n)$

$i = 0$; dimensión $= (n + 2 + Q_h) \times m$

$$\boldsymbol{A}_2^{(0)} = \begin{bmatrix} 0 & 0 & \cdots & m\mu \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

$i = 1, \ldots, Q_n$; dimensión $= (n + 2 + Q_h) \times (n + 2 + Q_h)$

$$\boldsymbol{A}_2^{(i)} = i\eta \boldsymbol{I}_{n+2+Q_h} + \mu \left[\begin{array}{cc} m & \\ & f(m+1) \\ \hline & \\ & \bar{\bar{\boldsymbol{0}}} \\ & \end{array}\right]$$

C.2. Algoritmo F-HOPSWR

C.2.1. Matrices $\boldsymbol{A}_0^{(i)}$ $(i = -1, \ldots, Q_n - 1)$

Estas matrices coinciden con las matrices correspondientes del algoritmo FGC.

C.2.2. Matrices $\boldsymbol{A}_1^{(i)}$ $(i = -1, \ldots, Q_n)$

$i = -1$; dimensión $= m \times m$

Esta matriz coincide con la matriz correspondiente del algoritmo FGC.

$i = 0;$ dimensión $= (n+2+Q_h) \times (n+2+Q_h)$

Esta matriz coincide con la matriz correspondiente del algoritmo FGC.

$i = 1, \ldots, Q_n;$ dimensión $= (n+2+Q_h) \times (n+2+Q_h)$

$$\boldsymbol{A}_1^{(i)} = \begin{bmatrix} * & \lambda_h + f\lambda_n & & & & & & & \\ (m+1-t)\mu & * & \lambda_h & & & & & & \\ & \ddots & \ddots & \ddots & & & & & \\ & & (C-t)\mu & * & \lambda_h & & & & \\ & & & C\mu+\gamma & * & \lambda_h & & & \\ & & & & \ddots & \ddots & \ddots & & \\ & & & & & \ddots & * & \lambda_h & \\ & & & & & & C\mu + Q_h\gamma & * & \end{bmatrix}$$

C.2.3. Matrices $\boldsymbol{A}_2^{(i)}$ ($i = 0, \ldots, Q_n$)

$i = 0;$ dimensión $= (n+2+Q_h) \times m$

Esta matriz coincide con la matriz correspondiente del algoritmo FGC.

$i = 1, \ldots, Q_n;$ dimensión $= (n+2+Q_h) \times (n+2+Q_h)$

$$\boldsymbol{A}_2^{(i)} = i\eta \boldsymbol{I}_{n+2+Q_h} + \mu \left[\begin{array}{c|c|c} m & & \\ \hline & t\boldsymbol{I}_{n+1} & \\ \hline & & \overline{\overline{\mathbf{0}}} \end{array}\right]$$

C.3. Algoritmo F-HOPS

C.3.1. Matrices $A_0^{(i)}$ ($i = -m, \ldots, Q_n - 1$)

$i = -m, \ldots, -2$; dimensión $= 2(n+1) \times 2(n+1)$

$$\boldsymbol{A}_0^{(i)} = \lambda \boldsymbol{I}_{2(n+1)}$$

$i = -1$; dimensión $= 2(n+1) \times [2(n+1) + Q_h]$

$$\boldsymbol{A}_0^{(-1)} = \left[\begin{array}{c|c} \lambda \boldsymbol{I}_{2(n+1)} & \bar{\bar{\boldsymbol{0}}} \end{array}\right]$$

$i = 0, \ldots, Q_n - 1$; dimensión $= [2(n+1) + Q_h] \times [2(n+1) + Q_h]$

$$\boldsymbol{A}_0^{(i)} = \left[\begin{array}{c|c} (1-f)\lambda_n \boldsymbol{I}_{n+1} & \bar{\bar{\boldsymbol{0}}} \\ \hline \bar{\bar{\boldsymbol{0}}} & \lambda_n \boldsymbol{I}_{n+1+Q_h} \end{array}\right]$$

C.3.2. Matrices $A_1^{(i)}$ ($i = -m, \ldots, Q_n$)

Introducimos las matrices auxiliares $\boldsymbol{U}_1$, $\boldsymbol{U}_2$ y $\boldsymbol{U}_3$

$$\boldsymbol{U}_1 = \begin{bmatrix} * & 0 & & & & \\ \mu & * & 0 & & & \\ & 2\mu & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & * & 0 \\ & & & & n\mu & * \end{bmatrix}, \quad \boldsymbol{U}_2 = \begin{bmatrix} * & \lambda_h & & & & \\ \mu & * & \lambda_h & & & \\ & 2\mu & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & * & \lambda_h \\ & & & & n\mu & * \end{bmatrix}$$

$$\boldsymbol{U}_3 = \begin{bmatrix} * & \lambda_h & & & & & \\ C\mu + 2\gamma & * & \lambda_h & & & & \\ & C\mu + 3\gamma & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & * & \lambda_h & \\ & & & & C\mu + Q_h\gamma & * & \end{bmatrix}$$

$i = -m, \ldots, -1;$ dimensión $= 2(n+1) \times 2(n+1)$

$$\boldsymbol{A}_1^{(i)} = \left[\begin{array}{c|c} \boldsymbol{U}_1 & \bar{\bar{\boldsymbol{0}}} \\ \hline \mu \boldsymbol{I}_{n+1} & \boldsymbol{U}_1 \end{array}\right]$$

$i = 0;$ dimensión $= [2(n+1) + Q_h] \times [2(n+1) + Q_h]$

$$\boldsymbol{A}_1^{(0)} = \left[\begin{array}{c|ccc|ccc} \boldsymbol{U}_1 & \multicolumn{3}{c|}{(f\lambda_n + \lambda_h)\boldsymbol{I}_{n+1}} & \multicolumn{3}{c}{\bar{\bar{\boldsymbol{0}}}} \\ \hline \mu \boldsymbol{I}_{n+1} & \multicolumn{3}{c|}{\boldsymbol{U}_2} & 0 & \cdots & 0 \\ & & & & \vdots & & \vdots \\ & & & & \lambda_h & \cdots & 0 \\ \hline & 0 & \cdots & C\mu + \gamma & & & \\ \bar{\bar{\boldsymbol{0}}} & \vdots & & \vdots & \multicolumn{3}{c}{\boldsymbol{U}_3} \\ & 0 & \cdots & 0 & & & \end{array}\right]$$

$i = 1, \ldots, Q_n$; dimensión $= [2(n+1) + Q_h] \times [2(n+1) + Q_h]$

$$A_1^{(i)} = \left[\begin{array}{c|ccc|ccc} \multicolumn{1}{c|}{\boldsymbol{U}_1} & \multicolumn{3}{c|}{(f\lambda_n + \lambda_h)\boldsymbol{I}_{n+1}} & \multicolumn{3}{c}{\overline{\overline{\mathbf{0}}}} \\ \hline \multicolumn{1}{c|}{} & & & & 0 & \cdots & 0 \\ \multicolumn{1}{c|}{(1-f)\mu\boldsymbol{I}_{n+1}} & \multicolumn{3}{c|}{\boldsymbol{U}_2} & \vdots & & \vdots \\ \multicolumn{1}{c|}{} & & & & \lambda_h & \cdots & 0 \\ \hline & 0 & \cdots & C\mu + \gamma & & & \\ \overline{\overline{\mathbf{0}}} & \vdots & & \vdots & \multicolumn{3}{c}{\boldsymbol{U}_3} \\ & 0 & \cdots & 0 & & & \end{array}\right]$$

C.3.3. Matrices $A_2^{(i)}$ ($i = -(m-1), \ldots, Q_n$)

$i = -(m-1), \ldots, -1$; dimensión $= 2(n+1) \times 2(n+1)$

$$A_2^{(i)} = (m+i)\mu\boldsymbol{I}_{2(n+1)}$$

$i = 0$; dimensión $= [2(n+1) + Q_h] \times 2(n+1)$

$$A_2^{(i)} = \left[\begin{array}{c} m\mu\boldsymbol{I}_{2(n+1)} \\ \hline \overline{\overline{\mathbf{0}}} \end{array}\right]$$

$i = 1, \ldots, Q_n$; dimensión $= [2(n+1) + Q_h] \times [2(n+1) + Q_h]$

$$\boldsymbol{A}_2^{(i)} = \left[\begin{array}{c|c|c} (m\mu + i\eta)\boldsymbol{I}_{n+1} & \bar{\bar{\mathbf{0}}} & \\ \hline \bar{\bar{\mathbf{0}}} & (t\mu + i\eta)\boldsymbol{I}_{n+1} & \bar{\bar{\mathbf{0}}} \\ \hline \bar{\bar{\mathbf{0}}} & \bar{\bar{\mathbf{0}}} & i\eta\boldsymbol{I}_{n+1} \end{array}\right]$$

C.4. Algoritmo F-HOSP

C.4.1. Matrices $\boldsymbol{A}_0^{(i)}$ ($i = -m, \ldots, Q_n - 1$)

$i = -m, \ldots, -2$; dimensión $= 2(n+1) \times 2(n+1)$

$$\boldsymbol{A}_0^{(i)} = \begin{bmatrix} \lambda_n & & & \\ & \ddots & & \\ & & \lambda_n & \\ & & & \lambda \end{bmatrix}$$

$i = -1$; dimensión $= 2(n+1) \times [2(n+1) + Q_h]$

$$\boldsymbol{A}_0^{(-1)} = \left[\begin{array}{c|c} \boldsymbol{A}_0^{(-2)} & \bar{\bar{\mathbf{0}}} \end{array}\right]$$

$i = 0, \ldots, Q_n - 1$; dimensión $= [2(n+1) + Q_h] \times [2(n+1) + Q_h]$

Estas matrices coinciden con las matrices correspondientes del algoritmo F-HOPS.

C.4.2. Matrices $A_1^{(i)}$ ($i = -m, \ldots, Q_n$)

$i = -m, \ldots, -1$; dimensión $= 2(n+1) \times 2(n+1)$

$$A_1^{(i)} = \left[\begin{array}{c|ccc} & 0 & \cdots & 0 \\ \boldsymbol{U}_2 & \vdots & & \vdots \\ & 0 & \cdots & \lambda_h \\ \hline \mu \boldsymbol{I}_{n+1} & & \boldsymbol{U}_2 & \end{array}\right]$$

$i = 0$; dimensión $= [2(n+1) + Q_h] \times [2(n+1) + Q_h]$

$$A_1^{(0)} = \left[\begin{array}{c|ccc|ccc} & f\lambda_n & & & & & \\ & & \ddots & & & & \\ \boldsymbol{U}_2 & & & \ddots & & \overline{\overline{\boldsymbol{0}}} & \\ & & & & f\lambda_n & & \\ & & & & f\lambda_n + \lambda_h & & \\ \hline & & & & & 0 \cdots 0 & \\ \mu \boldsymbol{I}_{n+1} & & \boldsymbol{U}_2 & & & \vdots \quad \vdots & \\ & & & & & \lambda_h \cdots 0 & \\ \hline & 0 & \cdots & C\mu + \gamma & & & \\ \overline{\overline{\boldsymbol{0}}} & \vdots & & \vdots & & \boldsymbol{U}_3 & \\ & 0 & \cdots & 0 & & & \end{array}\right]$$

$i = 1, \ldots, Q_n$; dimensión $= [2(n+1) + Q_h] \times [2(n+1) + Q_h]$

$$
\boldsymbol{A}_1^{(i)} = \left[\begin{array}{c|ccc|ccc}
 & f\lambda_n & & & & & \\
 & & \ddots & & & & \\
\boldsymbol{U}_2 & & & \ddots & & \bar{\bar{\mathbf{0}}} & \\
 & & & & f\lambda_n & & \\
 & & & & & f\lambda_n + \lambda_h & \\
\hline
 & & & & & 0 \cdots & 0 \\
(1-f)\mu \boldsymbol{I}_{n+1} & & \boldsymbol{U}_2 & & & \vdots & \vdots \\
 & & & & & \lambda_h \cdots & 0 \\
\hline
 & 0 & \cdots & & C\mu + \gamma & & \\
\bar{\bar{\mathbf{0}}} & \vdots & & & \vdots & \boldsymbol{U}_3 & \\
 & 0 & \cdots & & 0 & &
\end{array}\right]
$$

C.4.3. Matrices $\boldsymbol{A}_2^{(i)}$ ($i = -(m-1), \ldots, Q_n$)

Estas matrices coinciden con las matrices correspondientes del algoritmo F-HOPS.

Apéndice D

Publicaciones

D.1. Relacionas con la tesis

D.1.1. Revista

1. Vicent Pla and Vicente Casares-Giner. Análisis y diseño de políticas de control de admisión en redes celulares multiservicio. *Revista IEEE América Latina*, 2(1):10–19, March 2004. (pdf).

2. Vicent Pla and Vicente Casares-Giner. Analysis of priority channel assignment schemes in mobile cellular communication systems: a spectral theory approach. *Performance Evaluation*, 59(2-3):199–224, February 2005. (pdf).

3. Vicent Pla, Jorge Martínez, and Vicente Casares-Giner. Algorithmic computation of optimal capacity in multiservice mobile wireless networks. *IEICE Transactions on Communications*, E88-B(2):797–799, February 2005. (pdf).

4. David García, Jorge Martínez, and Vicent Pla. Admission control policies in multiservice cellular networks: Optimum configuration and

sensitivity. *Lecture Note in Computer Science, SPRINGER-VERLAG Berlin Heidelberg*, Mobile and Wireless Systems, LNCS 3427:121–135, 2005. (pdf).

5. Vicent Pla and Vicente Casares-Giner. A spectral-based analysis of priority channel assignment schemes in mobile cellular communication systems. *International Journal of Wireless Information Networks*, 2005. Accepted for publication. (pdf).

6. Vicent Pla, José Manuel Giménez-Guzmán, Jorge Martínez, and Vicente Casares-Giner. Optimal bandwidth reservation in multiservice mobile cellular networks with movement prediction. *IEICE Transactions on Communications*, 2005. Accepted for publication. (pdf).

D.1.2. Congreso

Internacional

1. Vicent Pla and Vicente Casares. Delay-loss analysis of channel assignment schemes in mobile cellular with handoff priority and hysteresis control. In *Proceedings of 14th ITC Specialist Seminar on Access Networks and Systems*, pp. 221–230, 2001. (pdf).

2. Vicent Pla and Vicente Casares-Giner. Effect of the handoff area sojourn time distribution on the performance of cellular networks. In *Proceedings of IEEE MWCN*, pp. 401–405, September 2002. (pdf).

3. Vicent Pla and Vicente Casares-Giner. Analytical-numerical study of the handoff area sojourn time. In *Proceedings of IEEE GLOBECOM*, pp. 886–890, November 2002. (pdf).

4. Vicent Pla and Vicente Casares-Giner. Optimal admission control policies in multiservice cellular networks. In *Proceedings of the International Network Optimization Conference (INOC)*, pp. 466–471, October 2003. (pdf).

5. Vicent Pla and Vicente Casares-Giner. Analysis of priority channel assignment schemes in mobile cellular communication systems. In *Proceedings of the 1st International Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks (HET-NETs'03)*, pp. 5/1–10, July 2003. (pdf).

6. David Garcia, Jorge Martínez, and Vicent Pla. Comparative evaluation of admission control policies in cellular multiservice networks. In *Proceedings of the 16th International Conference on Wireless Communications (Wireless 2004)*, pp. 517–531, July 2004. (pdf).

7. Vicent Pla, Jorge Martínez, and Vicente Casares-Giner. Efficient computation of optimal capacity in multiservice mobile wireless networks. In *Proceedings of the 2nd International Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks (HET-NETs'04)*, pp. 35/1–10, July 2004. (pdf).

8. Vicent Pla, José Manuel Giménez-Guzmán, Jorge Martínez, and Vicente Casares-Giner. Optimal admission control using handover prediction in mobile cellular networks. In *Proceedings of the 2nd International Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks (HET-NETs'04)*, pp. 44/1–10, July 2004. (pdf).

9. Vicent Pla, Vicente Casares-Giner, and Jorge Martínez. On a multiserver finite buffer queue with impatient customers. In *Proceedings of the 16th ITC Specialist Seminar on Performance Evaluation of Wireless and Mobile Systems*, pp. 55–62, September 2004. (pdf).

10. Jorge Martínez, David Garcia, and Vicent Pla. Optimality and sensitivity study of admission control policies for multimedia wireless networks. In *Proceedings of the IEEE International Workshop on MultiMedia Signal Processing (MMSP'04)*, September 2004. (pdf).

11. David García-Roger, María-José Domenech-Benlloch, Jorge Martínez-Bauset, and Vicent Pla. Adaptive scheme for admission control policies

in multiservice mobile wireless cellular networks. In *Proceedings of NGI 2005, Traffic Engineering for the Next Generation Internet*, April 2005. (pdf).

12. Vicent Pla, Vicente Casares-Giner, and Jorge Martínez. A matrix analytic solution of a finite buffer queue with PH distribued customers' impatience. In *Proceedings of the 19th International Teletraffic Congress (ITC19)*, 2005. a celebrar. (pdf).

Nacional

1. Vicent Pla Boscà y Vicente Casares Giner. Análisis y estudio comparativo de la capacidad de tráfico de un sistema LMDS multiservicio con diferentes procedimientos de reserva de canales. In *Actas de las III Jornadas de Ingeniería Telemática (JITEL'01)*, pp. 223–228, septiembre 2001. (pdf).

2. Vicent Pla Bsocà y Vicente Casares Giner. Análisis y diseño de políticas de control de admisión en redes celulares multiservicio. In *Actas de las IV Jornadas de Ingeniería Telemática (JITEL'03)*, pp. 487–494, septiembre 2003. (pdf).

3. David García Roger, Jorge Martínez Bauset, y Vicent Pla Boscà. Optimización y sensibilidad frentre a sobrecargas de políticas de control de admisión para redes celulares multiservicio. In *Actas de las XIV Jornadas Telecom I+D*, noviembre 2004. (pdf).

4. Vicent Pla Boscà, Vicente Casares Giner, y Jorge Martínez Bauset. Análisis de algoritmos de asignación de recursos a dos flujos de tráfico. In *Actas de las V Jornadas de Ingeniería Telemática (JITEL'05)*, 2005. (a celebrar).

5. José Manuel Giménez Guzmán, Jorge Martínez Bauset, Vicent Pla Boscà, y Vicente Casares Giner. Control de admisión Óptimo en redes móviles celulares con predicción de movimiento. In *Actas de las V Jornadas de Ingeniería Telemática (JITEL'05)*, 2005. (a celebrar).

6. David García Roger, M.ª José Domenech Benlloch, Jorge Martínez Bauset, y Vicent Pla Boscà. Esquema adaptativo de reserva para redes móviles celulares. In *Actas de las V Jornadas de Ingeniería Telemática (JITEL'05)*, 2005. (a celebrar).

D.2. Otras publicaciones

D.2.1. Revista

1. Roch A. Guérin and Vicent Pla. Aggregation and conformance in differentiated service networks. A case study. *ACM SIGCOMM Computer Communications Review (CCR)*, 31(1):21–32, January 2001. (pdf).

2. Vicente Casares-Giner, Pablo García-Escalle, and Vicent Pla. Evaluation of CELLULAR IP mobility tracking procedures. *Computer Networks: The International Journal of Computer and Telecommunications Networking Computer Networks and ISDN Systems*, 45(3):261–279, June 2004. Special issue in memory of Olga Casals. (pdf).

D.2.2. Congreso

Internacional

1. Luis Guijarro, Vicent Pla, and Jorge Martínez Jose R. Vidal. Multi-layer simulation approach for evaluation of data service support in ATM networks. In *Proceedings of ICATM'98*, pp. 270–277, June 1998. (pdf).

2. José Ramón Vidal, Luis Guijarro, Vicent Pla, and Jorge Martínez. A SDL modelling approach for performance evaluation of ATM networks. In *Proceedings of the 1st International Workshop on Formal Methods and Telecommunications*, pp. 22–39, 1999. (pdf).

3. Roch Guérin and Vicent Pla. Aggregation and conformance in differentiated service networks. a case study. In *Proceedings of the 13 ITC Specialist Seminar on IP Traffic Measurement, Modelling and Management*, pp. 26/1–11, 2000. (pdf).

4. Antonio León and Vicent Pla. Performance evaluation of handover schemes in ip micromobility systems. In *Proceedings of the 2nd International Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks (HET-NETs'04)*, pp. 30/1–10, July 2004. (pdf).

5. Vicente Casares, Pablo García, and Vicent Pla. Fractional hybrid movement-distance-based location update with selective paging. In *Proceedings of the 7th INFORMS Telecommunications Conference*, pp. 22–27, March 2004. (pdf).

6. Vicente Casares-Giner, Pablo García-Escalle, and Vicent Pla. The use of fractional memory in the hybrid movement-distance-based location update schemes with selective paging. In *Proceedings of the 11th European Wireless Conference*, pp. 466–471, June 2005. (pdf).

Nacional

1. José Ramón Vidal, Luis Guijarro, Vicent Pla, y Jorge Martínez. Evaluación de TCP con tráfico esporádico sobre ATM con servicios ABR y UBR. In *Actas de URSI´98*, pp. 559–560, 1998.

Bibliografía

[AL02] A.S. Alfa y W. Li, *PCS networks with correlated arrival process and retrial phenomenon*, IEEE Transactions on Wireless **1** (2002), no. 4, 630–637, (pdf).

[Bar01] Novella Bartolini, *Handoff and optimal channel assignment in wireless networks*, Mobile Networks and Applications (MONET) **6** (2001), no. 6, 511–524, (pdf).

[Bar04] Francisco Barceló, *Performance analysis of handoff resource allocation strategies through the state-dependent rejection scheme*, IEEE Transactions on Wireless **3** (2004), no. 3, 900–909, (pdf).

[BB97] Francisco Barceló y S. Bueno, *Idle and inter-arrival time statistics in public access mobile radio (pamr) systems*, Proceedings of IEEE GLOBECOM, 1997, pp. 126–130, (pdf).

[BBP01] André-Luc Beylot, Selma Boumerdassi, y Guy Pujolle, *NACR: A new adaptive channel reservation in cellular communication systems*, Telecommunication Systems **17** (2001), 233–241, (pdf).

[BC02] Novella Bartolini y Imrich Chlamtac, *Call admission control in wireless multimedia networks*, Proceedings of IEEE PIMRC, 2002, (pdf).

[BdW94] O.J. Boxma y P.R. de Waal, *Multiserver queues with impatient customers*, Proceedings of ITC 14, Elsevier Science, 1994, pp. 743–756, (pdf).

[BF01] Cory C. Beard y Victor S. Frost, *Prioritized resource allocation for stressed networks*, IEEE/ACM Transactions on Networking **9** (2001), no. 5, 618–633, (pdf).

[BH81] François Baccelli y Gérard Hebuterne, *On queues with impatient customers*, Performance '81 (F.J. Kylstra, ed.), North-Hollad Publishing Company, 1981, pp. 159–179, (pdf).

[BJ00] Francisco Barceló y Javier Jordán, *Channel holding time distribution in public telephony systems (PAMR and PCS)*, IEEE Transactions on Vehicular Technology **49** (2000), no. 5, 1615–1625, (pdf).

[BM98] S.C. Borst y D. Mitra, *Virtual partitioning for robust resource sharing: computational techniques for heterogeneous traffic*, IEEE Journal on Selected Areas in Communications **16** (1998), no. 5, 668 – 678, (pdf).

[BS97] S.K. Biswas y B. Sengupta, *Call admissibility for multirate traffic in wireless atm networks*, Proceedings of IEEE INFOCOM, vol. 2, 1997, pp. 649–657, (pdf).

[BS99] F. Barcelo y J.I. Sanchez, *Probability distribution of the inter-arrival time to cellular telephony channels*, Proceedings of IEEE 49th Vehicular Technology Conference, 1999, pp. 762–766, (pdf).

[CB86] R.V. Churchill y J. W. Brown, *Complex variables and applications*, 5th ed., McGraw-Hill, 1986.

[CB00] Ming-Hsing Chiu y M.A. Bassiouni, *Predictive schemes for handoff prioritization in cellular networks based on mobile positioning*, IEEE Journal on Selected Areas in Communications **18** (2000), no. 3, 510–522, (pdf).

[CC97] Chi-chao Chao y Wai Chen, *Connection admission control for mobile multiple-class personal communications networks*, IEEE Journal on Selected Areas in Communications **15** (1997), no. 8, 1618–1626, (pdf).

[CG01] Vicente Casares-Giner, *Integration of dispatch and interconnect traffic in a land mobile trunking system. waiting time distributions*, Telecommunication Systems **16** (2001), no. 3,4, 539–554, (pdf) Previously presented at 4th INFORMS Telecommunications Conference, Boca Ratón (Florida) March 8–11, 1998.

[CH96] V. Casares y J.M. Holtzman, *Dispatch versus interconnect traffic. a comparative analysis in a land mobile trunking system*, Proceedings of the 46th VTC, abril 1996, pp. 242–246, (pdf), see for more details WINLAB TR-118, Rutgers University, NJ, May 1996.

[CKS+98] Mooho Cho, Kwangsik Kim, Ferenc Szidarovszky, Younggap, y Kyoungrok Cho, *Numerical analysis of the dwelll time distribution in mobile cellular communication systems*, IEICE Transactions on Communications **E81-B** (1998), no. 4, 715–721, (pdf).

[CL95] E. Chlebus y W. Ludwin, *Is handoff traffic really poissonian ?*, Proceedings of ICUPC'95, 1995, pp. 348–353, (pdf).

[CPVAOG04] F.A. Cruz-Perez, J.L. Vazquez-Avila, y L. Ortigoza-Guerrero, *Recurrent formulas for the multiple fractional channel reservation strategy in multi-service mobile cellular networks*, IEEE Communications Letters **8** (2004), no. 10, 629–631, (pdf).

[CS98] Sunghyun Choi y Kang G. Shin, *Predictive and adaptive bandwidth reservation for handoffs in QoS-sensitive cellular networks*, Proceedings of ACM SIGCOMM'98, septiembre 1998, pp. 155–166, (pdf).

[CS02] _____, *Adaptive bandwidth reservation and admission control in QoS-sensitive cellular networks*, IEEE Transactions on Parallel and Distributed Systems **13** (2002), no. 9, 882–897, (pdf).

[CSC94] Chung-Ju Chang, Tian-Tsair Su, y Yueh-Yiing Chiang, *Analysis of a cutoff priority cellular radio system with finite queueing and renegingdropping*, IEEE/ACM Transactions on Networking **2** (1994), no. 2, 166–175, (pdf).

[DH86] Bharat T. Doshi y Harry Heffes, *Overload performance of several processor queueing disciplines for the M/M/1 queue*, IEEE Transactions on Communications **COM-34** (1986), no. 6, 538–546, (pdf).

[DJ92] J.N. Daigle y N. Jain, *A queueing system with two arrival streams and reserved servers with application to cellular telephone*, Proceedings of INFOCOM'92, IEEE, mayo 1992, pp. 9C.2.1–9.C.2.7, (pdf).

[DRFG99a] Enrico Del Re, Romano Fantacci, y Giovanni Giambene, *Different queuing policies for handover request in low earth orbit mobile satellite systems*, IEEE Transactions on Vehicular Technology **48** (1999), no. 2, 448–458, (pdf).

[DRFG99b] _____, *Handover queuing strategies with dynamic and fixed channel allocation techniques in low earth orbit mobile satellite systems*, IEEE Transactions on Communications **47** (1999), no. 1, 89–102, (pdf).

[DS02] D. J. Daley y L. D. Servi, *Loss probabilities of hand-in traffic under various protocols. I. models and algebraic results*, Telecommunication Systems **19** (2002), no. 2, 209–226, (pdf).

[DS04] _____, *Loss probabilities of hand-in traffic under various protocols: II. model comparisons*, Performance Evaluation **55** (2004), no. 3-4, 231–249, (pdf).

[DTL03] S. Dharmaraja, K.S. Trivedi, y D. Logothetis, *Performance modeling of wireless networks with generally distributed handoff interarrival times*, Computer Communications **26** (2003), 1747–1755, (pdf).

[EAYH01a] E.-S. El-Alfy, Yu-Dong Yao, y H. Heffes, *Adaptive resource allocation with prioritized handoff in cellular mobile networks under QoS provisioning*, Proceedings of IEEE the 54th Vehicular Technology Conference (VTC Fall), 2001, pp. 2113–2117, (pdf).

[EAYH01b] _____, *Autonomous call admission control with prioritized handoff in cellular networks*, Proceedings of IEEE ICC, 2001, pp. 1386 – 1390, (pdf).

[EAYH01c] _____, *A learning approach for call admission control with prioritized handoff in mobile multimedia networks*, Proceedings of IEEE the 53rd Vehicular Technology Conference (VTC Spring), 2001, pp. 972 – 976, (pdf).

[EAYH01d] _____, *A model-based q-learning scheme for wireless channel allocation with prioritized handoff*, Proceedings of IEEE GLOBECOM, 2001, pp. 3668 – 3672, (pdf).

[EE99] Jamie S. Evans y David Everitt, *Effective bandwidth-based admission control for multiservice CDMA cellular networks*, IEEE

Transactions on Vehicular Technology **48** (1999), no. 1, 36–46, (pdf).

[Esp03] Real Academia Española (ed.), *Diccionario de la lengua española (edición electrónica)*, 22ª ed., Espasa Calpe, 2003.

[ET99] Howard G. Ebersman y Ozan K. Tonguz, *Handoff ordering using signal prediction priority queuing in personal communication systems*, IEEE Transactions on Vehicular Technology **48** (1999), no. 1, 20–35, (pdf).

[Fan00] Romano Fantacci, *Performance evaluation of prioritized handoff schemes in mobile cellular networks*, IEEE Transactions on Vehicular Technology **49** (2000), no. 2, 485–493, (pdf).

[Fan01] Yuguang Fang, *Hyper-erlang distribution and its applications in wireless and mobile networks*, Wireless Networks (WINET) **7** (2001), no. 3, 211–219, (pdf).

[GJ03] Eva Gustafsson y Annika Jonsson, *Always best connected*, IEEE Wireless Communications **10** (2003), 49–55, (pdf).

[GJL84] D.P. Gaver, P.A. Jacobs, y G. Latouche, *Finite birth-and-death models in randomly changing environments*, Advances in Applied Probability **16** (1984), 715–731.

[GMP04] David Garcia, Jorge Martínez, y Vicent Pla, *Comparative evaluation of admission control policies in cellular multiservice networks*, Proceedings of the 16th International Conference on Wireless Communications (Wireless 2004), julio 2004, pp. 517–531, (pdf).

[GMP05] David García, Jorge Martínez, y Vicent Pla, *Admission control policies in multiservice cellular networks: Optimum configuration and sensitivity*, Mobile and Wireless Systems, LNCS 3427. Lecture Note in Computer Science (LCNS), vol. 3427, SPRINGER-VERLAG Berlin Heidelberg, 2005, pp. 121–135, (pdf).

[GSM] *http://www.gsmworld.com/news/statistics*, visitada en abril de 2005.

[Gué87] Roch A. Guérin, *Channel occupancy time distribution in a cellular radio system*, IEEE Transactions on Vehicular Technology **35** (1987), 89–99.

[Gué88] _____, *Queueing-blocking system with two arrival streams and guard channel*, IEEE Transactions on Communications **36** (1988), no. 2, 153–163, (pdf).

[HF01] Jiongkuan Hou y Yuguang Fang, *Mobility-based call admission control schemes for wireless mobile networks*, Wireless Communications and Mobile Computing **1** (2001), no. 3, 269–282, (pdf).

[HR86] Daehyong Hong y Stephen S. Rappaport, *Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures*, IEEE Transactions on Vehicular Technology **VT-35** (1986), no. 3, 77–92, See also: CEAS Technical Report No. 773, June 1, 1999, College of Engineering and Applied Sciences, State University of New York, Stony Brook, NY 11794, USA. (pdf).

[HSSK01] Hirotoshi Hidaka, Kazuyoshi Saitoh, Noriteru Shinagawa, y Takehiko Kobayashi, *Teletraffic characteristics of cellular communication for different types of vehicle motion*, IEICE Transactions on Communications **E84-B** (2001), no. 3, 558–565, (pdf).

[HSSK02] _____, *Self-similarity in cell dwell time caused by terminal motion and its effects on teletraffic of cellular communication networks*, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences **E85-A** (2002), no. 7, 1445–1453, (pdf).

[HUCPOG03a] Heraclio Heredia-Ureta, Felipe A. Cruz-Pérez, y Lauro Ortigoza-Guerrero, *Capacity optimization in multiservice mobile wireless networks with multiple fractional channel reservation*, IEEE Transactions on Vehicular Technology **52** (2003), no. 6, 1519 – 1539, (pdf).

[HUCPOG03b] _____, *Multiple fractional channel reservation for multi-service cellular networks*, Proceedings of IEEE ICC, 2003, pp. 964 – 968, (pdf).

[HUCPOG03c] _____, *Multiple fractional channel reservation for optimum system capcity in multi-service cellular networks*, Electronics Letters **39** (2003), no. 1, 133–134, (pdf).

[Ive02] Villy Bæk Iversen, *Teletraffic engineering handbook*, ITU-D SG 2 and ITC, 2002, http://www.tele.dtu.dk/teletraffic/.

[Jab96] Bijan Jabbari, *Teletraffic aspects of evolving and next-generation wireless communication networks*, IEEE Personal Communications (1996), 4–9, (pdf).

[JGA01] Aruna Jayasuriya, David Gree, y Asenstorfer, *Modelling service time distribution in cellular networks using phase-type service distributions*, Proceedings of ICC'01, 2001, (pdf).

[JL96] C. Jedrzycki y V.C.M. Leung, *Probability distributions of channel holding time in cellular telephony systems*, Proceedings of VTC'96, mayo 1996, pp. 247–251, (pdf).

[JV94] S. Jordan y P.P. Varaiya, *Control of multiple service, multiple resource communication networks*, IEEE Transactions on Communications **42** (1994), no. 11, 2979 – 2988, (pdf).

[KA99] Mohana Dhamayanthi Kulavaratharasah y A. H. Aghvami, *Teletraffic performance evaluation of microcellular personal communication networks (PCN's) with prioritized handoff procedures*, IEEE Transactions on Vehicular Technology **48** (1999), no. 1, 137–152, (pdf).

[Kei79] J. Keilson, *Markov chain models — rarity and exponentiality*, Springer-Verlag, 1979.

[KI95] J. Keilson y O. C. Ibe, *Cutoff priority in mobile cellular communications systems*, IEEE Transactions on Communications **43** (1995), no. 2/3/4, 1038–1045, (pdf).

[Kle75] Leonard Kleinrock, *Queuing systems*, vol. I: Theory, John Wiley & Sons, 1975.

[KN96] I. Katzela y M. Naghshineh, *Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey*, IEEE Personal Communications (1996), 10–31, (pdf).

[KS97] Jae Kyun Kwon y Dan Keun Sung, *Soft handoff modeling in CDMA cellular systems*, Proceeding of VTC '97 (Phoenix, USA), mayo 1997, pp. 1548–1551, (pdf).

[KS99] Duk Kyung Kim y Dan Keung Sung, *Characterization of soft handoff in CDMA systems*, IEEE Transactions on Vehicular Technology **48** (1999), no. 4, 1195–1202, (pdf).

[KZ97] Farooq Khan y Djamal Zeghlache, *Effect of cell residence time distribution on the performance of cellular mobile networks*, Proceedings of VTC'97, IEEE, 1997, pp. 949 – 953, (pdf).

[LA95] Chin-Tau Lea y Anwar Alyatama, *Bandwidth quantization and states reduction in the broadband isdn*, IEEE/ACM Transactions on Networking **3** (1995), no. 3, 352–360, (pdf).

[LAN97] D.A. Levine, I.F. Akyildiz, y M. Naghshineh, *A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept*, IEEE/ACM Transactions on Networking **5** (1997), no. 1, 1–12, (pdf).

[LC97] Yi-Bing Lin y Imrich Chlamtac, *Effects of Erlang call holding times on PCS call completion*, Proceedings of INFOCOM'97, IEEE, 1997, (pdf).

[Lit61] J.D.C. Little, *A proof of the formula: $L = \lambda W$*, Operations Research **9** (1961), 383–387.

[LJCP03] Wei Kuang Lai, Yu-Jyr Jin, Hsin Wei Chen, y Chieh Ying Pan, *Channel assignment for initial and handoff calls to improve the call-completion probability*, IEEE Transactions on Vehicular Technology **52** (2003), no. 4, 876–890, (pdf).

[LLC98] Bo Li, Chuang Lin, y Samuel T. Chanson, *Analysis of a hybrid cutoff priority scheme for multiple classes of traffic in multimedia wireless networks*, Wireless Networks Journal (WINET) **4** (1998), no. 4, 279–290, (pdf).

[LMN94] Yi-Bing Lin, Seshadri Mohan, y Anthony Noerpel, *Queueing priority channel assignment strategies for PCS hand-off and initial access*, IEEE Transactions on Vehicular Technology **43** (1994), no. 3, 704–712, (pdf).

[LR99] G. Latouche y V. Ramaswami, *Introduction to matrix analytic methods in stochastic modeling*, ASA-SIAM, 1999.

[Mac05] Fumiaki Machihara, *Mobile telecommunication systems and generalized Erlang loss formula*, IEICE Transactions on Communications **E88-B** (2005), no. 1, 183–189, (pdf).

[Man04] Avishai Mandelbaum, *Call centers (centres), research bibliography with abstracts*, Tech. report, Faculty of Industrial

Engineering and Management Technion—Israel Institute of Technology, Haifa32000, Israel, diciembre 2004, http://iew3.technion.ac.il/serveng/References/references.html (pdf).

[McM91] David McMillan, *Traffic modelling and analysis for cellular mobile networks*, Proceedigs of ITC-13 on Teletraffic and Datatraffic in a Period of Change (Copenhaguen) (A. Jensen y V.B. Iversen, eds.), IAC, Elsevier Science, junio 1991, pp. 627–632.

[McM95] D. McMillan, *Delay analysis of a cellular mobile priority queuing system*, IEEE/ACM Transactions on Networking **3** (1995), no. 3, 310–319, (pdf).

[MEBC02] I. Martin-Escalona, F. Barceló, y J. Casademont, *Teletraffic simulation of cellular networks: modeling the hand-off arrivals and the hand-off delay*, Proceedings of the 13th IEEE PIMRC, 2002, pp. 2209–2213, (pdf).

[Mit95] Isi Mitrani, *Advances in queueing. Theory, methods and open problems (editor jewgeni h. dshalalow)*, ch. 13: The spectral expansion solution method for Markov processes on lattice strips, pp. 337–352, CRC Press, 1995.

[MK00] Werner Mohr y Walter Konhäuser, *Access network evolution beyond third generation mobile communications*, IEEE Communications Magazine **38** (2000), 122–133, (pdf).

[MM00] Michela Meo y Marco Ajmone Marsan, *Approximate analytical models for dual-band GSM networks design and planning*, Proceedings of IEEE INFOCOM, 2000, (pdf).

[MMDCL+01] Marco Ajmone Marsan, Giovanni Marco De Carolis, Emilio Leonardi, Renato Lo Cigno, y Michela Meo, *Efficient estimation of call blocking probabilities in cellular mobile telephony networks with customer retrials*, IEEE Journal on Selected Areas in Communications **19** (2001), no. 2, 332–346, (pdf).

[Mov98] Ali Movaghar, *On queueing with customer impatience until the beginning of service*, Queuing Systems **29** (1998), no. 2, 337–350, (pdf).

[MS01] K. Mitchell y K. Sohraby, *An analysis of the effects of mobility on bandwidth allocation strategies in multi-class cellular wireless networks*, Procedings of IEEE INFOCOM, vol. 2, 2001, pp. 1075–1084, (pdf).

[Neu81] M. Neuts, *Matrix-geometric solutions in stochastic models: An algorithmic approach*, The Johns Hopkins University Press, 1981.

[NS96] Mahmoud Naghshineh y Mischa Schwartz, *Distributed call admision control in mobile wireless networks*, IEEE Journal on Selected Areas in Communications **14** (1996), no. 4, 711–717, (pdf).

[OR01] Philip V. Orlik y Stephen S. Rappaport, *On the handoff arrival process in cellular communications*, Wireless Networks Journal (WINET) **7** (2001), no. 2, 147–157, (pdf).

[PC01] Vicent Pla y Vicente Casares, *Delay-loss analysis of channel assignment schemes in mobile cellular with handoff priority and hysteresis control*, Proceedings of 14th ITC Specialist Seminar on Access Networks and Systems, 2001, pp. 221–230, (pdf).

[PCG02a] Vicent Pla y Vicente Casares-Giner, *Analytical-numerical study of the handoff area sojourn time*, Proceedings of IEEE GLOBECOM, noviembre 2002, pp. 886–890, (pdf).

[PCG02b] _____, *Effect of the handoff area sojourn time distribution on the performance of cellular networks*, Proceedings of IEEE MWCN, septiembre 2002, pp. 401–405, (pdf).

[PCG04] _____, *Análisis y diseño de políticas de control de admisión en redes celulares multiservicio*, Revista IEEE América Latina **2** (2004), no. 1, 10–19, (pdf).

[PG85] E. C. Posner y R. Guérin, *Traffic policies in cellular radio that minimize blocking of handoff calls*, Proceedings of ITC 11, 1985, (pdf).

[PGGMCG04] Vicent Pla, José Manuel Giménez-Guzmán, Jorge Martínez, y Vicente Casares-Giner, *Optimal admission control using handover prediction in mobile cellular networks*, Proceedings of the 2nd International Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks (HET-NETs'04), julio 2004, pp. 44/1–10, (pdf).

[Pol96] Gregory P. Pollini, *Trends in handover desing*, IEEE Communications Magazine (1996), 82–90, (pdf).

[PSS96] Suwon Park, Ho Shin Sho, y Dan Keun Sung, *Modeling and analysis of CDMA soft handoff*, Proceedings of the VTC'96, 1996, (pdf).

[Put94] Martin L. Puterman, *Markov decision processes : Discrete stochastic dynamic programming*, John Wiley & Sons, 1994.

[Rap94] Stephen S. Rappaport, *Microcellular communications systems with hierarchical macrocel overlays: Traffic performance models and analysis*, Proceedings of the IEEE **82** (1994), no. 9, 1383–1396, (pdf).

[RGS98] Marina Ruggieri, Fabio Graziosi, y Fortunato Santucci, *Modeling of the handover dwell time in cellular mobile communications systems*, IEEE Transactions on Vehicular Technology **47** (1998), no. 2, 489–498, (pdf).

[RNT97] R. Ramjee, R. Nagarajan, y D Towsley, *On optimal call admission control in cellular networks*, Wireless Networks Journal (WINET) **3** (1997), no. 1, 29–41, (pdf).

[Ros70] Sheldon M. Ross, *Applied probability models with optimization applications*, Holden-Day, 1970.

[Ros89] Keith W. Ross, *Randomized and past-dependent policies for markov decision processes with multiple constraints*, Operations Research **37** (1989), no. 3, 474–477, (pdf).

[Ros95] _____, *Multiservice loss models for broadband telecommunication networks*, Springer Verlag, 1995.

[RT00] Myuran Rajaratnam y Fambirai Takawira, *Nonclassical traffic modeling and performance analysis of cellular mobile networks with and without channel reservation*, IEEE Transactions on Vehicular Technology **49** (2000), no. 3, 817–834, (pdf).

[RT01] _____, *Handoff traffic characterization in cellular networks under nonclassical arrivals and service time distributions*, IEEE Transactions on Vehicular Technology **50** (2001), no. 4, 954–970, (pdf).

[Sal04] Apostolis K. Salkintzis, *Interworking techniques and architectures for WLAN/3G integration toward 4G mobile data networks*, IEEE Wireless Communications **11** (2004), 50–61, (pdf).

[SB98] R. Sutton y A. G. Barto, *Reinforcement learning*, The MIT press, Cambridge, Massachusetts, 1998, http://www.cs.ualberta.ca/∼sutton/book/ebook/the-book.html.

[Sch90] R. Schehrer, *On a cut-off priority queueing system with hysteresis and unlimited waiting room*, Computer Networks and ISDN Systems **20** (1990), 45–56, (pdf).

[Sch03] Mathias Schweigel, *The cell residence time in rectangular cells and its exponential approximation*, Proceedings of ITC 18, 2003, (pdf).

[SK04] Wee-Seng Soh y Hyong S. Kim, *Dynamic bandwidth reservation in cellular networks using road topology based mobility prediction*, Proceedings of IEEE INFOCOM, 2004, (pdf).

[Soc02] IEEE Communications Society (ed.), *A brief history of communications*, IEEE, 2002.

[SS97] M. Sidi y D. Starobinski, *New call blocking versus handoff blocking in cellular networks*, Wireless Networks Journal (WINET) **3** (1997), no. 1, 15–27, (pdf).

[SW04] Haw-Yun Shin y Jean-Lien C. Wu, *The study of dynamic multichannel scheme with channel de-allocation in wireless networks*, Computer Networks **45** (2004), 463–482, (pdf).

[TGM97] P. Tran-Gia y M. Mandjes, *Modeling of customer retrial phenomenon*, IEEE Journal on Selected Areas in Communications **15** (1997), no. 8, 1406–1414, (pdf).

[TJ91] Sirin Tekinay y Bijan Jabbari, *Handover and channel assignment in mobile cellular networks*, IEEE Communications Magazine (1991), 42–46, (pdf).

[TJ92] ———, *A measurement-based prioritization scheme for handovers in mobile cellular networks*, IEEE Journal on Selected Areas in Communications **10** (1992), no. 8, 1343–1350, (pdf).

[TP92] D. Towsley y S.S. Panwar, *Optimality of the stochastic earliest deadline policy for the G/M/c queue serving customers with deadlines*, Proceedings of the 2nd ORSA Telecommunications Conference, 1992, (pdf).

[TRV98] Nishith D. Tripathi, Jeffrey H. Reed, y Hugh F. VanLandingham, *Handoff in cellular systems*, IEEE Personal Communications (1998), 26–37, (pdf).

[VC00] A. Valkó y A. Campbell, *An efficiency limit of cellular mobile systems*, Computer Communications **23** (2000), no. 5-6, 441–451, (pdf).

[vDT03] E.A. van Doorn y A.T.K. Ta, *Proofs for some conjectures of Rajaratnam and Takawira on the peakedness of handoff traffic*, IEEE Transactions on Vehicular Technology **52** (2003), no. 4, 953–957, (pdf).

[VLLX02] Johan De Vriendt, Philippe Lainé, Christophe Lerouge, y Xiaofeng Xu, *Mobile network evolution: A revolution on the move*, IEEE Communications Magazine **40** (2002), 104–111, (pdf).

[Wes02] Krzysztof Wesołowski, *Mobile communication systems*, John Wiley & Sons, 2002.

[Wil78] J. H. Wilkinson, *The algebraic eigenvalue problem*, Oxford: Clarendon Press, 1978.

[Wol82] R. W. Wolff, *Poisson arrivals see time averages*, Operation Research **30** (1982), no. 2, 223–231.

[Wol89] _____, *Stochastic modeling and the theory of queues*, Prentice Hall, Englewood Cliffs, NJ, 1989.

[WWL02] Si Wu, K.Y. Michael Wong, y Bo Li, *A dynamic call admission policy with precision QoS guarantee using stochastic control for mobile wireless networks*, IEEE/ACM Transactions on Networking **10** (2002), no. 2, 257–271, (pdf).

[WZA03] Jingao Wang, Qing-An Zeng, y Dharma P. Agrawal, *Performance analysis of a preemptive and priority reservation handoff scheme for integrated service-based wireless mobile networks*, IEEE Transactions on Mobile Computing **2** (2003), no. 1, 65–75, (pdf).

[XCW00] Yang Xiao, Philip Chen, y Yan Wang, *A near optimal call admission control with genetic algorithm for multimedia services in wireless/mobile networks*, Proceedings of the National Aerospace and Electronics Conference (NAECON), 2000, pp. 787 – 792, (pdf).

[XCW01] _____, *Optimal admission control for multi-class of wireless adaptive multimedia services*, IEICE Transactions on Communications **E84-B** (2001), no. 4, 795–804, (pdf).

[XG93] H. Xie y D.J. Goodman, *Mobility models and biasing sample problem*, Proceedings of ICUPC'93 (Ottawa, Canada), IEEE, octubre 1993, pp. 804–807, (pdf).

[XT03] Ariton E. Xhafa y Ozan K. Tonguz, *Does mixed lognormal channel holding time affect the handover performance of guard channel scheme ?*, Proceedings of IEEE GLOBECOM, 2003, pp. 3452–3456, (pdf).

[XT04] _____, *Dynamic priority queueing of handover calls in wireless networks: An analytical framework*, IEEE Journal on Selected Areas in Communications **22** (2004), no. 5, 904 – 916, (pdf).

[YL02] Fei Yu y Victor Leung, *Mobility-based predictive call admission control and bandwidth reservation in wireless cellular networks*, Computer Networks **38** (2002), no. 5, 577–589, (pdf).

[YMW+04] Jianxin Yao, J.W. Mark, Tung Chong Wong, Yong Huat Chew, Kin Mun Lye, y Kee-Chaing Chua, *Virtual partitioning resource allocation for multiclass traffic in cellular systems with QoS constraints*, IEEE Transactions on Vehicular Technology **53** (2004), no. 3, 847 – 864, (pdf).

[YR97] A. Yener y C. Rose, *Genetic algorithms applied to cellular call admission: local policies*, IEEE Transactions on Vehicular Technology **46** (1997), no. 1, 72–79, (pdf).

[ZA95] Yiqiang Quennel Zhao y Attahiru Sule Alfa, *Performance analysis of a telephone system with both patient and impatient customers*, Telecommunication Systems **4** (1995), 201–215, (pdf).

[ZA02] Qing-An Zeng y Dharma P. Agrawal, *Modeling and efficient handling of handoffs in integrated wireless mobile networks*, IEEE

Transactions on Vehicular Technology **51** (2002), no. 6, 1469 – 1478, (pdf).

[ZAB00] M. Zeng, A. Annamalai, y V. K. Bhargava, *A harmonization of global third-generation mobile systems*, IEEE Communications Magazine **38** (2000), 94–104, (pdf).

[ZC99] H . Zeng y I. Chlamtac, *Handoff traffic distribution in cellular networks*, Proceedings of IEEE Wireless Communications and Networking Conference (WCNC), 1999, pp. 413–417, (pdf).

[ZD97] Mahmodd M. Zonoozi y Prem Dassanayake, *User mobility modeling and characterization of mobility patterns*, IEEE Journal on Selected Areas in Communications **15** (1997), no. 7, 1239–1252, (pdf).

[ZK04] Roland Zander y Johan M. Karlsson, *Predictive and adaptive resource reservation (parr) for cellular networks*, International Journal of Wireless Information Networks **11** (2004), no. 3, 161–171, (pdf).

[ZTH+00] George I. Zysman, Joseph A. Tarallo, Richard E. Howard, John Freindenfelds, Reinaldo A. Valenzuela, y Paul M. Mankiewich, *Technology evolution for mobile and personal communications*, Bell Labs Technical Journal (2000), 107–129, (pdf).