UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Departamento de Informática de Sistemas y Computadores

Una estrategia para la reducción del consumo de potencia en redes de interconexión

TESIS DOCTORAL

Marina Alonso Díaz

DIRIGIDA POR:

Dr. Vicente Santonja

Dr. Pedro López

Valencia, Junio 2012

A Joan y Salva, siempre sereis mi prioridad...

Papá, mamá... gracias por estar siempre justo en el sitio en que os necesito...

A Salva, porque nunca has dejado que termine un día sin hacerme más feliz, sin enseñarme de quien realmente puedo aprender...

Una estrategia para la reducción del consumo de potencia en redes de interconexión

Marina Alonso Díaz

Resumen

El alto nivel de potencia de cálculo requerida por algunas aplicaciones sólo puede ser alcanzado por sistemas multiprocesador. Estos sistemas consisten en varios procesadores que se comunican mediante una red de interconexión. El enorme aumento tanto en el tamaño como la complejidad de los sistemas multiprocesador ha disparado su consumo de energía. Las técnicas de reducción de consumo de potencia se están aplicando a todos los niveles en los computadores y la red de interconexión no puede ser una excepción.

En este entorno, las redes de interconexión más ampliamente utilizadas están basadas en topologías regulares: directas, como los toros, e indirectas, como los *fat-tree*. En ambos casos el consumo de potencia de la circuitería de la red de interconexión contribuye significativamente al total del sistema.

En esta tesis, proponemos una estrategia para reducir el consumo de potencia en las redes de interconexión, tanto directas como indirectas. Dicha estrategia se materializa en forma de un mecanismo que combina dos técnicas alternativas: (i) la conexión y desconexión dinámica de los enlaces de la red en función del tráfico (cualquier enlace puede ser desconectado, con tal de que la conectividad de red esté garantizada), (ii) el ajuste dinámico del ancho de banda de los enlaces en función del tráfico. En ambos casos, la topología de la red no se ve modificada. Por lo tanto, el mismo algoritmo de encaminamiento puede ser usado independientemente de las acciones de ahorro en el consumo llevadas a cabo, simplificando así el diseño del router.

Nuestros resultados muestran que el consumo de potencia de la red se puede reducir muy significativamente, a costa de algún incremento en la latencia. Sin embargo, la reducción de potencia alcanzada es siempre mayor que la penalización en la latencia.

A Strategy for Power Saving in Interconnection Networks

Marina Alonso Díaz

Abstract

The high level of computing power required for some applications can only be achieved by multiprocessor systems. These systems consist of several processors that communicate by means of an interconnection network. The huge increase both in size and complexity of high-end multiprocessor systems has triggered up their power consumption. Power consumption reduction techniques are being applied everywhere in computer systems and the interconnection network must not be an exception.

In this scenario, the most widely used interconnection networks are based on regular topologies: direct topologies, as torus, and indirect topologies, as fat-tree. In both cases the power consumed by the interconnect circuitry has a non-negligible contribution to the total system budget.

In this thesis, we propose a strategy to reduce interconnection network, based both on direct and indirect topologies, power consumption. The strategy is implemented by means of a mechanism that combines two alternative techniques: (i) dynamically switching on and off network links as a function of traffic (any link can be switched off, provided that network connectivity is guaranteed), (ii) dynamically setting the available link bandwidth as a function of traffic. In both cases, the topology of the network is not modified. Therefore, the same routing algorithm can be used regardless of the power saving actions taken, thus simplifying router design.

Our results show that the network power consumption can be greatly reduced, at the expense of some increase in latency. However, the achieved power reduction is always higher than the latency penalty.

Una estratègia per a la reducció del consum de potència en xarxes d'interconnexió

Marina Alonso Díaz

Resum

L'alt nivell de potència de càlcul requerit per algunes aplicacions només pot ser aconseguit per sistemes multiprocessador. Estos sistemes consistixen en diversos processadors que es comuniquen per mitjà d'una xarxa d'interconnexió. L'enorme augment tant en la grandària com la complexitat dels sistemes multiprocessador ha disparat el seu consum d'energia. Les tècniques de reducció de consum de potència s'estan aplicant a tots els nivells en els computadors i la xarxa d'interconnexió no pot ser una excepció.

En este entorn, les xarxes d'interconnexió més ampliament utilitzades estan basades en topologies regulars: directes, com els torus, i indirectes com els *fat-tree*. En ambdós casos el consum de potència de la circuiteria de la xarxa d'interconnexió contribuïx significativament al total del sistema.

En esta tesi, proposem una estratègia per a reduir el consum de potència en les xarxes d'interconnexió, tant directes com indirectes. Aquesta estratègia es materialitza en un mecanisme que combina dos tècniques alternatives: (i) la connexió i desconnexió dinàmica dels enllaços de la xarxa en funció del tràfic (qualsevol enllaç pot ser apagat, sempre que la connectivitat de xarxa estiga garantida), (ii) l'ajust dinàmic de l'ample de banda dels enllaços en funció del tràfic. En ambdós casos, la topologia de la xarxa no es veu modificada. Per tant, el mateix algoritme d'encaminament pot ser usat independentment de les accions d'estalvi en el consum dutes a terme, simplificant així el disseny del router. Els nostres resultats mostren que el consum de potència de la xarxa es pot reduir enormement, a costa d'algun increment en la latència. No obstant això, la reducció de potència aconseguida és sempre major que la penalització en la latència.

Agradecimientos

Al finalizar un trabajo tan arduo y lleno de dificultades como el desarrollo de una tesis, al menos de mi tesis, no puedes evitar pensar que este aporte hubiera sido imposible sin la participación de personas que te han facilidado el camino. Es para mí un verdadero placer utilizar este espacio para expresarles a todos y cada uno de ellos mis mayores agradecimientos.

En primer lugar a Salva, por su apoyo, comprensión y amor, que me permite sentir que puedo lograr lo que me proponga;

a mis padres, por ser un ejemplo de generosidad, entrega y superación;

a Vicente Santonja, por sus consejos, enseñanzas, comentarios, sugerencias, colaboración y ayuda, por su predisposición permanente e incondicional durante el largo lapso de mi tesis;

a Pedro López, por su capacidad para originar este trabajo y guiar las ideas que han ido surgiendo en su desarrollo;

a Juan Miguel Martínez, por su estímulo, colaboración, aporte y participación activa en los trabajos relacionados con la tesis;

a José Duato, por permirtirme formar parte del GAP y transmitirme su entusiasmo;

y en general a todos los que de una forma u otra han colaborado en el desarrollo de mi tesis y sin los que su ayuda y conocimiento no estaría donde me encuentro ahora.

Índice general

Índice de figuras

Índice de tablas

1 Introducción

1.1. Contexto

El contexto de este trabajo son las redes de interconexión. Las redes de interconexión son elementos críticos en los sistemas digitales actuales, pues en muchos casos éstas son factores limitadores de su rendimiento, no su lógica o su memoria, por ejemplo.

La importancia de este tipo de redes está en aumento debido a que, además de proporcionar conectividad exterior, las redes se utilizan para conectar componentes dentro de un computador a distintos niveles, incluyendo el microprocesador. Tradicionalmente las redes de interconexión se han usado en grandes computadores paralelos (supercomputadores), pero hoy en día es habitual encontrar estos diseños también con computadores personales debido a la demanda en ancho de banda para la comunicación que se necesita para permitir el incremento en potencia de cálculo y capacidad de almacenamiento. Las redes de interconexión están reemplazando a los buses como medio de comunicación entre procesadores, dispositivos de entrada/salida, tarjetas, chips y elementos dentro del chip [34].

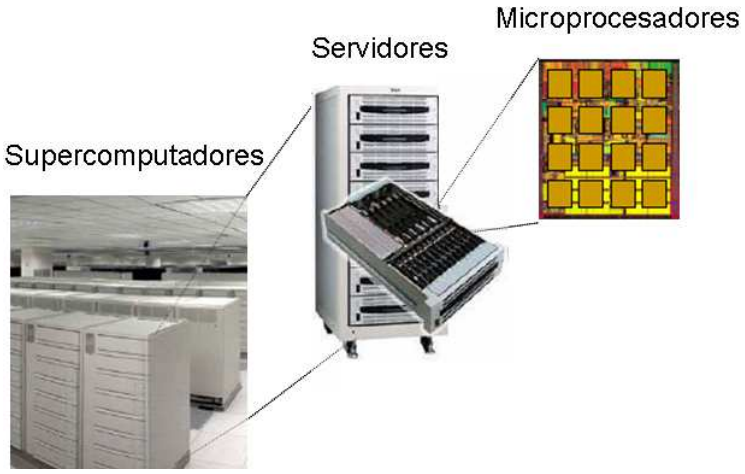


Figura 1.1: Sistemas de propósito general

Las redes de interconexión se utilizan por tanto en sistemas de propósito general, como supercomputadores, servidores y microprocesadores (Figura 1.1) y en sistemas de aplicación específica, como routers para internet o sistemas en chip (Figura 1.2); en ambos casos los sistemas de computación van creciendo a base de introducir cada vez más elementos, de modo que en definitiva los sistemas dependen de la red de interconexión para poder escalar adecuadamente [45].

Tal como se ha señalado, las redes de interconexión se diseñan para ser utilizadas en distintos niveles, dentro y entre sistemas de computación. Dependiendo del número de dispositivos conectados y su proximidad, es posible agrupar las redes de interconexión en cuatro dominios principales [34]:

- Redes en chip (Networks on Chip, NoCs o On Chip Networks, OCNs), que se utilizan para interconectar unidades funcionales a nivel de microarquitectura, bancos de registros, caches, procesadores, etc. dentro de una misma pastilla o chip.

- Redes para sistemas de almacenamiento o redes de sistema (Storage or System Area Networks), que se utilizan para interconectar procesadores o procesador-memoria en sistemas multiprocesadores y multicomputadores, y también para conectar elementos de almacenamiento y de entrada/salida.

- Redes de área local (Local Area Networks, LANs), que se utilizan para interconectar computadores autónomos distribuidos en un ámbito geográfico restrin-

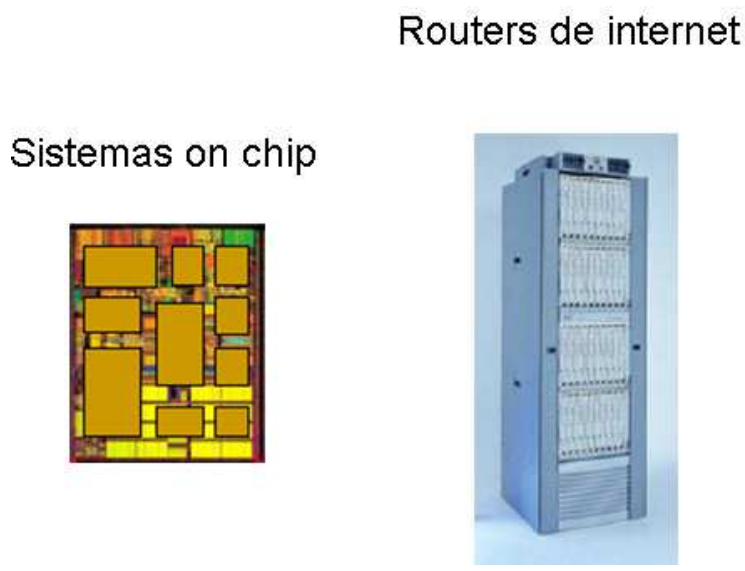


Figura 1.2: Sistemas de propósito específico

gido como una habitación, edificio, campus, etc. El ejemplo típico es la interconexión de PCs en un cluster.

- Redes de área ancha (Wide Area Networks, WANs), que se utilizan para conectar sistemas de computadores distribuidos en un ámbito geográfico más amplio: regional, nacional, intercontinental, ...y que generalmente requiere soporte de Internet.

En estos ámbitos descritos, exiten soluciones que se han convertido en estándares comerciales, y otras que son soluciones propietarias. En cualquier caso, aunque los detalles puedan diferir significativamente entre unos dominios y otros, los problemas y los términos usados para solucionarlos permanecen casi inalterables entre los distintos ámbitos. Con independencia del dominio, las redes se deben diseñar para evitar ser el cuello de botella del sistema en términos de prestaciones (latencia y productividad), y coste (no sólo económico sino también en consumo de área y potencia); por lo tanto, la meta de todo diseño de una red de interconexión es un coste lo más bajo posible que permita transferir la máxima cantidad de información en el menor tiempo. El presente trabajo se centra en la reducción del coste en el ámbito del consumo de potencia de las redes de interconexión. En particular, se estudian las

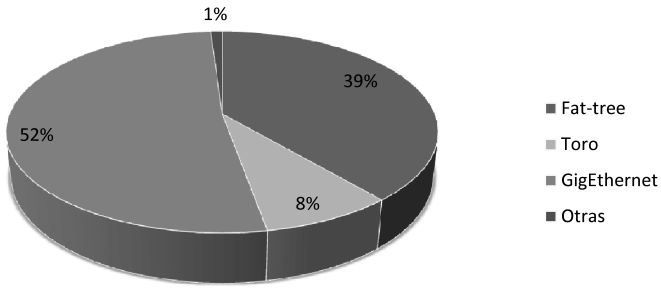


Figura 1.3: Topologías más utilizadas en el TOP500

redes directas con topología toroidal y las redes indirectas con topología de fat-tree. Aunque se han propuesto muchas topologías de redes en los últimos 50 años, sólo unas pocas se han empleado en sistemas reales (han llegado a implementarse en la práctica) [26], [34]. Un análisis de la tendencia seguida actualmente en los diseños que ocupan las posiciones más relevantes entre los supercomputadores mundiales listados en TOP500 indica claramente que éstas son las topologías más utilizadas actualmente excluyendo las redes basadas en tecnología Ethernet [66], donde el 8 % utilizan una red de interconexión con topología toroidal y el 39 % con topología fat-tree.

1.2. Motivación

Durante la última década, se ha consolidado una creciente preocupación por una gestión eficiente de los recursos energéticos. El mundo de los computadores no es ajeno a dicha tendencia. Ello ha provocado que la gestión inteligente del consumo energético se haya convertido en un campo de creciente interés investigador. Lo que se ha dado en llamar "computación verde", ha sido también marcado como objetivo por distintas agencias gubernamentales, definiendo requerimientos de eficiencia energética en los equipos [10].

Iniciativas como la lista Green500 que construye una clasificación con los supercomputadores más eficientes desde el punto de vista energético a nivel mundial [33], o la reciente inclusión de los datos de consumo energético en la lista TOP500 de los supercomputadores más potentes [66], son buena prueba de la creciente importancia que se otorga a la gestión eficiente de la energía. Máxime cuando el mundo de la

supercomputación tradicionalmente no se había preocupado por el consumo energético: el objetivo era simplemente obtener las mejores prestaciones a cualquier coste. Durante décadas la noción de "prestaciones" ha sido sinónimo de "velocidad", medida en *FLOPS, operaciones en coma flotante por segundo*. Desafortunadamente, este punto de vista ha llevado a los supercomputadores a consumir unas cantidades enormes de potencia y producir tanto calor que ha sido necesario construir dispositivos de refrigeración sofisticados para asegurar su funcionamiento correcto. Además, el énfasis en la velocidad como métrica de prestaciones ha hecho que se ignoren otras métricas importantes. Como consecuencia, se ha incrementado significativamente el coste total de un supercomputador.

Además de la relevancia de la gestión energética desde el punto de vista del desarrollo sostenible, el impacto del consumo de energía sobre la usabilidad, la fiabilidad y el coste de mantenimiento de los sistemas informáticos es también significativo. En el ámbito de la computación móvil o en sistemas empotrados, una gestión eficiente del consumo de potencia permite prolongar la vida de las baterías y su autonomía [32]. En el campo de la computación de altas prestaciones, la disipación de potencia incide significativamente en los costes operativos de suministro eléctrico y refrigeración y también en la fiabilidad [24].

1.3. Descripción del problema

Como se ha justificado en la Sección 1.1, las redes de interconexión han adquirido un papel muy significativo en los sistemas computadores actuales, a todas las escalas: desde los grandes supercomputadores a los dispositivos móviles. Las redes de interconexión propuestas y empleadas en los distintos ámbitos descritos han ido mejorando en los últimos años, proporcionando incrementos de productividad y reducciones de latencia. Se han diseñado routers de alta velocidad y enlaces de comunicacion con tasas de envío de datos del orden de varios gigabytes/s y latencias del orden de pocos microsegundos, para adecuarse a los crecientes requerimientos de rendimiento.

Esta tendencia en el diseño ha contribuido a incrementar también el consumo de potencia de las redes de interconexión imponiendo restricciones bastante considerables en la arquitectura de las redes de comunicación y en la escalabilidad de los sistemas. Sirva como ejemplo que varios estudios estiman que la potencia consumida por la infraestructura de las redes de comunicación en Estados Unidos se sitúa entre 5 y 24 TWh/año, con un coste asociado exhorbitante [44, 49]. A nivel de sistema, las

redes de interconexión son un importante consumidor del total de la potencia disponible, debido a que los enlaces consumen una parte muy significativa de la potencia total [54]. Los routers y los enlaces serán, por lo tanto, en el futuro más cercano, los componentes críticos, debido en parte a que los esfuerzos en reducir el consumo se han centrado tradicionalmente en otros ámbitos como los elementos de proceso o las memorias pero no en las redes de interconexión.

A continuación detallamos algunos ejemplos que muestran la significativa contribución de las redes de interconexión al consumo total de los sistemas de los que forman parte:

- En una red de interconexión con tecnología Infiniband (de la firma Mellanox), el router y los enlaces disipan 15W del total de 40W, esto es un 37.5 %, comparable al consumo del procesador que incorpora [6].
- Un switch Infiniband 12X de 8 puertos de IBM, se ha estimado que consume 31W de los cuales los enlaces suponen un 65 % (20W)[3].
- En un router Avici Terabit la potencia consumida por la red de interconexión es aproximadamente un tercio de la disipada por la tarjeta completa [25].
- Los enlaces y el router integrado en el procesador Alpha 21364 consumen 23W, un 18.4 % del total de la potencia del chip estimada en 125W, de los que el router consume 7.6W y la circuitería 13.3W[36].

Otra muestra de la relevancia del consumo de potencia en las redes de interconexión es que las nuevas especificaciones para interconexiones punto a punto en las nuevas arquitecturas de fabricantes como Intel (Quickpath) [12] y AMD (Hypertransport) [11], que pronto serán comunes en la mayoría de sistemas informáticos a nivel mundial, definen mecanismos para una gestión eficiente de la energía. También se ha lanzado recientemente una iniciativa en el ámbito de Ethernet (*IEEE P802.3az Energy Efficient Ethernet Task Force*) que está trabajando para estudiar soluciones en este campo [35].

El escenario presentado muestra la relevante contribución de la red de interconexión al consumo de energía de los sistemas computadores actuales a todos los niveles, desde las redes en chip hasta las WANs. Por esta razón, es muy relevante buscar soluciones que permitan reducir dicho consumo sin degradar significativamente las prestaciones.

2
Objetivos

El escenario presentado en el capítulo 1 muestra, por un lado, la relevancia de las redes de interconexión en los sistemas computadores actuales a todas las escalas, y en particular en las máquinas paralelas de altas prestaciones (tanto clusters como máquinas masivamente paralelas). Por otro lado, es de destacar la creciente importancia de la gestión eficiente de los recursos energéticos a todos los niveles, donde las máquinas paralelas no son una excepción. En estos sistemas, las redes de interconexión son responsables de una fracción muy significativa del consumo total de potencia. Consumo que no tiene perspectivas de reducirse sino de aumentar conforme se va incrementando el ancho de banda.

Esta tesis pretende reducir el consumo de potencia de las redes de interconexión, proporcionando una alternativa a las técnicas existentes. Para alcanzar este objetivo se plantean las siguientes preguntas:

- ¿Es posible definir un mecanismo que permita reducir el consumo de potencia sin requerir complejos elementos hardware?
- ¿Es posible que además no introduzca modificaciones significativas en el funcionamiento habitual de la red, y en concreto, que no requiera cambiar el algo-

ritmo de encaminamiento?

- ¿Es, además, posible reducir el consumo con un impacto poco significativo en las prestaciones?

Nuestra solucion debe proporcionar el maximo ahorro de potencia con el mínimo impacto en las prestaciones y al mismo tiempo no requerir modificaciones significativas de los algoritmos de encaminamiento; al final, esto redundará en simplicidad en la implementación y por lo tanto switches mas sencillos y con un coste más bajo.

En este trabajo se va a desarrollar, caracterizar y evaluar una nueva estrategia para reducir el consumo de potencia de la red de interconexión. La solución propuesta estará basada en la gestión dinámica del estado de los enlaces de la red, en cada uno de los entornos más populares en el escenario actual de la computacion de altas prestaciones:

- en el caso de redes con topologías directas, tipo toro y malla, se va a conseguir el objetivo mediante un mecanismo basado en dos técnicas: la conexión/desconexión de los enlaces y el ajuste de su ancho de banda.

- en el caso de redes con topologías indirectas, basadas en *fat-tree*, la solución propuesta estará basada en el apagado selectivo de enlaces.

Se pretende con ello hacer un uso eficiente de los recursos, adaptando el ancho de banda total de la red a los requerimientos impuestos por la carga de trabajo. Se abordarán algunos problemas: la estimación de manera fiable del tráfico en la red; la definición de una política adecuada de gestión del estado de los enlaces que permita regular el nivel de exigencia del mecanismo en términos de ahorro y también su capacidad de reaccionar ante cambios de tráfico; la gestión adecuada de situaciones de estrés de la red debidas a la reducción de sus prestaciones cuando actua el mecanismo; y, finalmente, la propuesta planteada deberá ser caracterizada y evaluada experimentalmente.

3

Fundamentos y trabajos previos

Para enmarcar el trabajo de investigación presentado en esta tesis se introducen a continuación algunos de los conceptos básicos. En la Sección 3.1 se presentan las topologías de referencia sobre las que se ha realizado el trabajo y en la Sección 3.2 se repasan las propuestas previas para la reducción del consumo de potencia en redes de interconexión.

3.1. Redes de interconexión. Topologías.

3.1.1. Introducción.

En el diseño de una red de interconexión hay tres factores clave: la topología, el algoritmo de encaminamiento y la conmutación [28].

- La topología describe la configuración de una red de interconexión y el modo en que los elementos (enlaces y conmutadores) se conectan unos a otros. La topología de una red de interconexión es análoga a un mapa de carreteras: los paquetes (coches) circulan por los enlaces (carriles o carreteras) desde un nodo

a otro, atravesando conmutadores (ciudades) y encaminadores (intersecciones de carreteras).

- El algoritmo de encaminamiento describe cómo se determina el camino que un paquete debe seguir para alcanzar su nodo destino desde su nodo origen.
- La conmutación describe cómo se realiza la asignación de recursos (canales y buffers) a los paquetes conforme van avanzando en su ruta.

En las topologías regulares, los elementos están enlazados siguiendo un mismo patrón de interconexión. Estas redes pueden por lo tanto escalar mejor a un mayor número de nodos. Dentro de las redes regulares podemos encontrar las *redes directas* o de conmutación distribuida y las *redes indirectas* o multietapa (también se les llama redes de conmutación centralizada) [34].

- Las redes directas son redes estrictamente ortogonales, donde tenemos conectado cada nodo o procesador con un conmutador y donde cada nodo tiene al menos un enlace en cada dirección [28], esto hace que el encaminamiento sea sencillo. Ejemplos de este tipo de redes son los toros y las mallas. Este tipo de topologías viene definido por el número de dimensiones y el número de nodos de cada dimensión. Los nodos se sitúan en un espacio de n dimensiones con k nodos en cada dimensión. De este modo, los toros también son referenciados como k-ary n-cube y las mallas como k-ary n-mesh, siendo k el número de nodos en cada una de las n dimensiones. Las mallas son similares a los toros pero no tienen los enlaces que se usan para conectar los nodos de los extremos de cada dimensión. Por lo tanto, en estas topologías, a diferencia de los toros, hay algunos nodos con un número diferente de nodos vecinos, en concreto los nodos situados en los bordes de la red. Ejemplos de algunos sistemas reales que han estado en las primeras posiciones de la lista de los supercomputadores más potentes del mundo, TOP500 [66], son: Jaguar [58] (Figura 3.1) que usa un toro 3D situado en la posición 1, Kraken [60] en la posición 3 basado también en un toro 3D, JUGENE [59] que es una implementación de BlueGene/P [57] basado en un toro 3D en la posición 4, una implementación de BlueGene/L [65] en la posición 7 y otra implementación de BlueGene/P que emplea la misma topología situado en la posición 8 del TOP500.

- La alternativa son las redes indirectas en las que tenemos los nodos conectados sólo al primer nivel de conmutadores y la forma de comunicarse unos

Figura 3.1: Supercomputador Jaguar

con otros es atravesando varias etapas o niveles de conmutadores. Este tipo de topologías se usan en entornos donde se prefiere conseguir mayores anchos de banda con una latencia previsible, que suele ser constante en ausencia de contención. El ejemplo más popular de este tipo de redes son los fat-trees. Por ejemplo, el computador que ha ocupado la posición más alta con esta topología en los TOP500, posición 2, es Roadrunner en Los Alamos National Laboratory [63] (Figura 3.2), situada además en la posición 6 dentro del ranking Green500, que clasifica a los supercomputadores de acuerdo con su consumo de potencia (relativo a su potencia de cálculo) [33]. También aparecen en la posición 5 Tianhe-1 [64] (posición 5 en el Green500), 9 Ranger [61] y 10 Red Sky [62], todos ellos con topología fat-tree

Este trabajo se ha realizado planteando y evaluando soluciones para los dos tipos de redes, tanto redes directas como redes indirectas. En el apartado 3.1.2 se presentan estas topologías y se describen sus características principales.

Figura 3.2: Supercomputador Roadrunner

3.1.2. Fundamentos básicos.

3.1.2.1. Redes directas.

Las redes directas se han modelado tradicionalmente como un grafo de interconexión $G(N, C)$[28], donde los vértices del grafo N representan el conjunto de nodos de procesamiento, y las aristas del grafo C representan el conjunto de canales de comunicación. A través de este grafo, la topología define las relaciones de interconexión existentes entre los distintos nodos: los nodos vecinos pueden enviarse mensajes entre ellos directamente, mientras que los nodos que no están conectados entre sí deben hacerlo a través de otros nodos que actúan de nodos intermedios. Este es un modelo muy simple que no considera aspectos de implementación. Sin embargo, permite el estudio de muchas propiedades interesantes de las redes:

- Grado: El grado de un nodo es el número de canales que conectan un nodo con sus nodos vecinos. Si el grado es demasiado alto también lo será el número de conexiones y aumentará la complejidad de la red.

- Diámetro: La distancia máxima entre dos nodos en la red. El diámetro está relacionado con la latencia máxima que podría tener una comunicación.

- Regularidad: Esta propiedad hace referencia a que todos los nodos tienen el mismo grado.
- Simetría: Una red se dice simétrica si la visión es la misma desde cualquier nodo. Esta característica facilita la elección de las diversas rutas para la comunicación.

Tres factores fundamentales caracterizan las redes directas: topología, encaminamiento y conmutación.

- La *topología* define cómo los nodos se interconectan con canales y habitualmente se modela con un grafo como hemos indicado. Para redes directas, la topología ideal sería aquella que conecta cualquier nodo con todos los demás nodos; ningún mensaje tendría entonces que pasar a través de nodos vecinos para alcanzar su objetivo. Es evidente que esta topología totalmente conectada tiene un coste prohibitivo para redes de tamaño moderado a grande y unas limitaciones físicas considerables, lo que ha llevado al desarrollo de topologías alternativas como las mallas y los toros.
- Para las topologías en las cuales los paquetes (un paquete es la unidad más pequeña de comunicación, que contiene además de los datos, la dirección del nodo destino e información de secuencia) deben atravesar algunos nodos intermedios, el algoritmo de *encaminamiento* determina el camino seleccionado por un paquete para alcanzar su destino.
- Cuando la cabecera de un paquete alcanza un nodo intermedio, un mecanismo de *conmutación* determina cómo y cuando se le asignan los recursos de red necesarios para seguir su ruta hasta el destino. Por ejemplo, si un paquete debe recibirse por completo en un switch antes de ser enviado al siguiente.

Dentro de esta categoría, muchas de las redes implementadas en la práctica siguen una topología ortogonal. Una topología es ortogonal si y sólo si los nodos se pueden organizar en un espacio n-dimensional, y cada enlace se puede disponer de manera que produzca un desplazamiento en una sola dimensión. En particular en las topologías estrictamente ortogonales cada nodo tiene al menos un enlace cruzando en cada dimensión. Una ventaja de estas topologías es que permiten aplicar sobre ellas algoritmos de encaminamiento muy simples. Las más populares son la malla n-dimensional y el k-ary n-cube o toro.

Formalmente, una malla n-dimensional tiene $k_0 \times k_1 \times \ldots \times k_{n-2} \times k_{n-1}$ nodos, siendo k_i el número de nodos en cada dimensión i, donde $k_i \geq 2$ y $0 \leq i \leq n-1$.

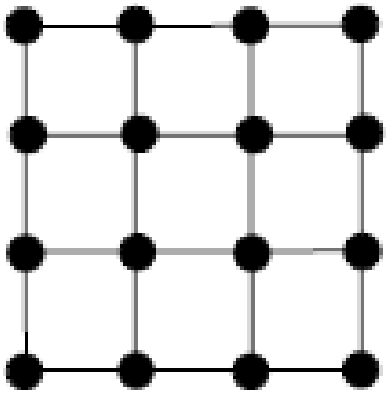


Figura 3.3: Malla bidimensional de 4×4

Cada nodo X está definido por sus n coordenadas, $(x_{n-1}, x_{n-2}, \ldots, x_1, x_0)$, donde $0 \leq x_i \leq k_i - 1$ para $0 \leq i \leq n-1$. Dos nodos X e Y son vecinos si y sólo si $y_i = x_i$ para todo i, $0 \leq i \leq n-1$, excepto uno, j, donde $y_j = x_j \pm 1$. Así, los nodos pueden tener de n a $2n$ vecinos, dependiendo de su localización en la malla [28]. La figura 3.3 muestra una malla de 4×4. La malla no representa una topología regular porque el grado de todos los nodos no es el mismo, en la figura los nodos de los vértices sólo tienen dos enlaces, los de los lados tres y el resto cuatro. Además, todos los nodos no tienen la misma imagen de la red.

Formalmente, en un k-ary n-cube, todos los nodos tienen el mismo número de vecinos. La definición del k-ary n-cube difiere de la de malla n-dimensional en que dos nodos X e Y son vecinos si y sólo si $y_i = x_i$ para todo i, $0 \leq i \leq n-1$, excepto uno, j, donde $y_j = (x_j \pm 1) \, mod \, k$. El cambio a aritmética modular en la dirección añade el canal entre los nodos $k-1$ y 0 en cada dimensión en los n-cubos k-arios, dándole regularidad y simetría. Cada nodo tiene n vecinos si $k = 2$ y $2n$ vecinos si $k > 2$. Cuando $n = 1$ el n-cubo k-ario se colapsa en un anillo bidireccional con k nodos [28] y cuando $k = 2$, el n-cubo k-ario es igual a un hipercubo de dimensión n. A las redes k-ary n-cube también se les llama toros. La figura 3.4 muestra un 4-ary 2-cube.

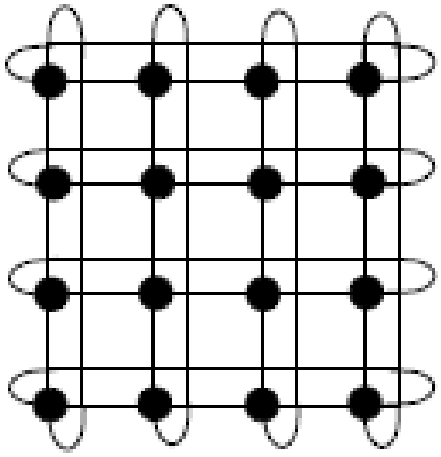


Figura 3.4: Toro bidimensional de 4×4, (4-ary 2-cube)

3.1.2.2. Redes indirectas.

Las redes indirectas también pueden modelarse con un grafo $G(N, C)$, donde N es el conjunto de switches, y C es el conjunto de enlaces unidireccionales o bidireccionales entre switches. Para el análisis de la mayoría de propiedades, no es necesario incluir explícitamente los nodos de procesamiento en el grafo. Aunque las redes indirectas pueden modelarse de forma similar a las directas, existen algunas diferencias entre ellas. Cada switch o conmutador en una red indirecta puede conectarse a cero, uno o más procesadores. Obviamente sólo los switches conectados a algún procesador pueden ser el origen o el destino de un mensaje. Además, transmitir un mensaje de un nodo a otro requiere atravesar el enlace entre el nodo origen y el switch al que se conecta, y el enlace entre el último switch de la ruta y el nodo destino. Por tanto, la distancia entre dos nodos es la distancia entre los switches conectados directamente a ellos más dos. Podría considerarse que no es necesario añadir dos unidades porque las redes directas también tienen enlaces internos entre switches y nodos de procesamiento. Sin embargo, esos enlaces son externos en el caso de las redes indirectas.

De forma similar a las redes directas, las redes indirectas están caracterizadas fundamentalmente por tres factores de diseño: topología, encaminamiento y conmutación. Para redes indirectas con N nodos, la topología ideal conectaría esos nodos a través de un solo conmutador de $N \times N$ puertos. Un conmutador así se conoce con el nombre de *crossbar*. El inconveniente de un crossbar es que su coste es aún prohibitivo para redes grandes. Al igual que con las redes directas, el número de conexiones

físicas de un switch está limitado por restricciones del hardware, como el número de pins disponibles o el área disponible para el cableado. Como consecuencia de estas limitaciones, se han propuesto muchas topologías alternativas. En estas topologías, los mensajes deben atravesar varios conmutadores antes de alcanzar el nuevo destino. Esos conmutadores son habitualmente idénticos y suelen estar organizados como un conjunto de etapas, cada una de las etapas está sólo conectada a la anterior y la siguiente empleando patrones de conexión regulares. De entre todas las topologías de redes indirectas propuestas, en la práctica la topología más ampliamente utilizada es la topología fat-tree (k-ary n-tree).

Un k-ary n-tree está formado por dos tipos de vértices [39, 47]: $N = k^n$ nodos de procesamiento y $S = nk^{n-1}$ switches $k \times k$ (un k-ary n-tree de dimensión $n = 0$ está compuesto solo por un nodo de procesamiento). Cada nodo de procesamiento está identificado por una tupla de longitud n, n-tupla, $(p_0, p_1, ..., p_{n-1})$ donde $p_i \in \{0, 1, ..., k-1\}$ para $0 \leq i \leq n-1$. Cada switch se define como un par ordenado (w, l), donde w es una tupla de longitud $n-1$, $(w_0, w_1, \ldots, w_{n-2})$, donde $w_i \in \{0, 1, \ldots, k-1\}$ y $l \in \{0, 1, ..., n-1\}$ es el nivel del switch (0 es el nivel raiz).

- Dados dos switches,$(w_0, w_1, ..., w_{n-2}, l)$ y $\left(w_0^{'}, w_1^{'}, ..., w_{n-2}^{'}, l\right)$, estos están conectados por un enlace si y sólo si $l^{'} = l + 1$ y $w_i = w_i^{'} \forall i \neq l$. El enlace que conecta ambos switches está etiquetado con $w_l^{'}$ en el switch del nivel l y con $k + w_l$ en el switch $l^{'}$. Este patrón de conexión es conocido como *mariposa* (butterfly).

  De este modo cada switch tiene $2k$ enlaces de salida de los cuales k están conectados a los switches o nodos de procesamiento del nivel $l + 1$ (enlaces descendentes) y los restantes k a los switches del nivel $l - 1$ (enlaces ascendentes).

- Hay un enlace entre el switch del nivel inferior $(w_0, w_1, ..., w_{n-2}, n-1)$ y el nodo de procesamiento $(p_0, p_1, ..., p_{n-1})$ si y sólo si $w_i = p_i, \forall i \in i\,\{0, 1, ..., n-2\}$. El enlace está etiquetado con p_{n-1} en el switch de nivel $n - 1$.

El esquema de etiquetado mostrado en las definiciones previas convierte el k-ary n-tree en una red delta [26, 28]: cualquier camino que empiece en un switch del nivel 0 y dirigido a un nodo $(p_0, p_1, ..., p_{n-1})$ atraviesa la misma secuencia de enlaces etiquetados $p_0, p_1, ..., p_{n-1}$ [28]. Un ejemplo de este etiquetado se muestra en la figura 3.5, para un *fat-tree* cuaternario de dimensión 3 (red de 64 nodos), es decir un 4-ary 3-tree.

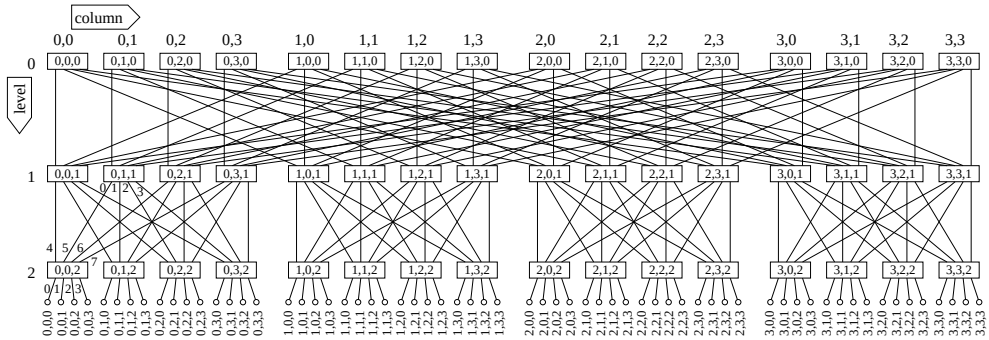


Figure 3.5: Etiquetado para los nodos, switches y enlaces de un 4-ary 3-tree

El camino mínimo entre cualquier par de nodos de procesamiento se puede obtener enviando un mensaje a uno de los switches ancestros comunes más cercanos y desde ahí al destino. Por lo tanto cada mensaje experimenta dos fases de encaminamiento: una fase ascendente, desde el nodo de procesamiento al ancestro común más cercano, seguida de una fase descendente al nodo destino. Mientras que la fase descendente es necesariamente determinista, puesto que hay un único camino desde el switch ancestro común más cercano al destino, es posible que existan rutas alternativas para alcanzar el ancestro común más cercano. La disponibilidad de rutas alternativas hace posible elegir aleatoriamente los enlaces ascendentes o incluso implementar un algoritmo adaptativo que tome las decisiones de acuerdo con el estado local del switch evitando los enlaces congestionados.

3.2. Trabajos previos.

En este apartado se presentan con detalle las soluciones propuestas hasta el momento al problema del consumo de potencia de las redes de interconexión. Los trabajos previos realizados sobre este tema se centran exclusivamente en redes regulares directas. Por lo que nosotros sabemos, no existe ningún estudio previo al nuestro realizado en redes indirectas.

En cuanto a las redes regulares directas, existen dos aproximaciones, que actúan reduciendo el consumo de potencia debido a los enlaces:

1. Escalado (ajuste) dinámico de la tensión y la frecuencia de los enlaces.

2. Gestión dinámica de la potencia usando enlaces on/off.

3.2.1. Escalado dinámico de la tensión y la frecuencia de los enlaces.

El escalado dinámico de la tensión (*DVS*), es una técnica de ahorro de potencia originalmente propuesta para microprocesadores. Esta técnica aprovecha la variabilidad en la carga en los procesadores, donde solamente durante una fracción del tiempo de cómputo se utiliza todo el rendimiento del procesador. Modificando la tensión de alimentación y la frecuencia de reloj bajo demanda, *DVS* proporciona las máximas prestaciones cuando estas son necesarias mientras minimiza el consumo de energía durante los periodos de poca actividad [21]. Esto es posible debido a la relación existente entre la potencia dinámica, la frecuencia y el voltaje. El consumo dinámico se produce debido a la carga y descarga de la capacidad de los transistores y las conexiones, y depende de la actividad del circuito. Es decir, ocurre únicamente durante las transiciones, cuando las puertas lógicas están conmutando. Por lo tanto, es proporcional a la frecuencia de conmutación y cuanto mayor sea el número de conmutaciones mayor será el consumo de potencia dinámico. La siguiente ecuación representa la potencia dinámica consumida [4]:

$$P_{dinámica} = KCV^2f \tag{3.1}$$

donde

- K es la actividad de conmutación, es decir, el número medio de transiciones durante un ciclo de reloj.
- C es la capacidad en cada nodo que conmuta.
- f es la frecuencia de reloj.
- V es el valor de la tensión de alimentación proporcionada.

De dicha ecuación se desprende que, puesto que la potencia dinámica es directamente proporcional a la frecuencia y tiene una relación cuadrática con el voltaje, si se ajusta la frecuencia y la tensión es posible disminuir la potencia minimizando el consumo.

Del mismo modo que la actividad en los microprocesadores, existe una gran variabilidad en la utilización de los enlaces en las redes de interconexión, que depende de los patrones de comunicación de las aplicaciones. Esto permite que teóricamente se puedan alcanzar grandes ahorros en el consumo de potencia si el ancho de banda efectivo de la red se ajusta a la utilización real. La idea es utilizar enlaces con

frecuencia variable [38], en los cuales se ajustan sus niveles de tensión al mínimo valor admisible para una frecuencia dada, proporcionando así un mecanismo que potencialmente puede disminuir el consumo de potencia. Estos enlaces de frecuencia variable pueden ser gestionados con técnicas *DVS*, permitiendo aplicar políticas que mejoren la potencia y las prestaciones sobre toda la red.

La primera contribución de una técnica *DVS* aplicada a la red de interconexión es de Shang, Peh y Jha [51], que proponen una implementación basada en el histórico de utilización de los enlaces. Consiste en usar la utilización de los enlaces en el pasado para predecir el tráfico futuro, ajustando dinámicamente la frecuencia y la tensión de los mismos para reducir el consumo de potencia en la red.

Uno de los aspectos cruciales de las técnicas *DVS e*s la política de transición, que determina cuándo y cuánto modificar la frecuencia del canal. El ancho de banda del canal decrece linealmente con la frecuencia del enlace, lo que queda reflejado en la ecuación:

$$b = wf \tag{3.2}$$

donde

- b es el ancho de banda del canal.
- w es el ancho del canal.
- f es la frecuencia del enlace.

Un canal con un ancho de banda reducido debido a la reducción en la frecuencia de operación del enlace puede causar degradación tanto en latencia de la red como en su productividad. Una buena política de transición ajustará el compromiso entre el ahorro en la potencia y la degradación en las prestaciones. El algoritmo basado en el histórico para minimizar el impacto en las prestaciones lo que hace es, en primer lugar, categorizar el tráfico existente en dos clases: cargas ligeras y congestión. Una red poco cargada es más sensible a la degradación en latencia introducida por la bajada de la frecuencia y la tensión, por lo tanto el escalado de la tensión/frecuencia debería ser conservador para minimizar la penalización en las prestaciones. Por otra parte, cuando la red está congestionada la transmisión del tráfico ya no depende tanto de la velocidad del enlace, en consecuencia y según los autores de esta propuesta, se puede aplicar un escalado de la tensión/frecuencia más agresivo para ahorrar más potencia.

El método que se propone se formaliza en un algoritmo donde cada conmutador predice el tráfico de sus comunicaciones futuras basándose en la utilización previa de los enlaces y de los buffers de entrada, de este modo ajusta dinámicamente la frecuencia del enlace y la correspondiente tensión al uso que se prevé.

La utilización del enlace, que se denomina U_{link}, se define como:

$$U_{link_i} = \frac{\sum_{t=1}^{H} A(t)}{H} \tag{3.3}$$

donde

- $A(t) = 1$ si hay tráfico en el enlace i en el ciclo t.
- $A(t) = 0$ si no hay tráfico en el enlace i en el ciclo t.
- H es el tamaño de la ventana temporal de cálculo medido en ciclos.

La utilización de los buffers de entrada que se denomina U_{buffer} se define como:

$$U_{buffer} = \frac{\sum_{t=1}^{H} (F(t)/B)}{H} \tag{3.4}$$

donde

- $F(t)$ es el número de buffers de entrada ocupados en el instante t.
- B es el número total de buffers de entrada.
- H es el tamaño de la ventana temporal de cálculo medido en ciclos.

Tanto la utilización futura de los enlaces como la de los buffers se predice usando una media ponderada exponencialmente de acuerdo con la siguiente expresión:

$$Par_{predict} = \frac{weight \times Par_{current} + Par_{past}}{weight + 1} \tag{3.5}$$

donde Par indica el parámetro que se predice (U_{link} o U_{buffer})

- $Par_{predict}$ es el valor de la predicción.
- $Par_{current}$ es el valor actual del parámetro.
- Par_{past} es el valor de la predicción en el periodo anterior.
- $weight$ es el valor de ponderación empleado en la media.

La predicción de la utilización de los buffers se emplea para estimar si la red está congestionada y en función de ello la utilización del enlace se compara con unos umbrales fijos seleccionados entre dos posibles parejas. Una pareja se emplea cuando la red está ligeramente cargada y otra cuando la red está congestionada.

Este mecanismo recopila únicamente información local al enlace, de este modo se evita la sobrecarga en comunicación.

Para la evaluación de esta técnica, las métricas consideradas son el ahorro de potencia y la degradación de la latencia con una carga de trabajo auto-similar. Por término medio consume 4 veces menos que un sistema original al que no se le ha aplicado el mecanismo *DVS* a costa de un incremento de la latencia del 15 % y de una reducción en la productividad del 2,5 %.

El principal inconveniente es que requiere componentes de hardware complejos como los reguladores para adaptar la tensión y los sintetizadores para la frecuencia; estos componentes hardware implican mayores costes y limitan la aplicabilidad de la técnica *DVS*. Los enlaces *DVS* en los que se varía la tensión y la frecuencia deben soportar cambios rápidos y necesitan por lo tanto mecanismos hardware sofisticados para asegurar el funcionamiento correcto durante el escalado, provocando sobrecargas importantes y un area CMOS adicional.

Stine y Carter [56] comparan el *DVS* con el uso de encaminamiento adaptativo en enlaces que no son *DVS*. Muestran que, mientras la red proporcione bastante ancho de banda para cubrir las necesidades de las aplicaciones, una red con enlaces de frecuencia fija y encaminamiento adaptativo mejora las prestaciones de una con enlaces *DVS*. Sólo se obtienen mejores latencias en algunos casos en los que la red con DVS alcanza una reducción en el consumo de potencia moderado y justifican este comportamiento porque los enlaces DVS reaccionan de forma muy lenta a las variaciones en la demanda de la red y porque los enlaces DVS operan a combinaciones frecuencia de reloj/tensión que no son óptimos debido a que cambian de una frecuencia de reloj a la siguiente.

Shin y Kim [52] proponen un algoritmo de asignación estática de la velocidad de los enlaces para redes en chip con enlaces de tensión escalable. Este esquema, asigna a priori niveles fijos de frecuencia y tensión a los enlaces , a partir del grafo de tareas de una aplicación. Este esquema es por tanto adecuado para aplicaciones de tiempo real habitualmente ejecutadas en sistemas empotrados, donde los diseñadores pueden predecir los retardos de comunicación durante el diseño.

Finalmente, Chen y Peh [23] muestran distintas alternativas de circuitos para implementar *DVS* en enlaces optoelectrónicos.

3.2.2. Gestión dinámica de la potencia usando enlaces On/Off.

Soteriou y Peh en [54] proponen una técnica que consiste en una reducción dinámica de la potencia basándose en la desconexión total de algunos enlaces dependiendo de la utilización de la red. En adelante nos referiremos a esta técnica como *DPM (Dynamic Power Management)*. En las redes de interconexión hay una gran variabilidad en los patrones de comunicación espacial y temporal dando lugar a utilizaciones de los enlaces irregularmente distribuidas. En realidad este trabajo surge de la idea de llevar al extremo la política anterior, desconectando los enlaces o conectándolos en respuesta a las variaciones del tráfico. En comparación con los enlaces *DVS*, esta técnica requiere un hardware más sencillo y además el ahorro de potencia alcanzable es superior, ya que los enlaces *DVS* consumen potencia aunque estén ociosos.

En esta aproximación, al desconectar enlaces, la disponibilidad de rutas alternativas se reduce drásticamente, pudiendo teóricamente, incluso llegar a desconectarse porciones de la red. Además la red podría bloquearse si el algoritmo de encaminamiento no está adaptado para soportar este mecanismo.

En[54] se ha diseñado un grafo de conectividad de potencia-prestaciones para topologías de mallas 2D sobre el que se construye un algoritmo de encaminamiento totalmente adaptativo libre de bloqueo que garantice la entrega de paquetes.

Los aspectos claves del estudio son:

- Un grafo de conectividad que dicta los enlaces candidatos a poder ser desconectados o conectados, según el caso, (que denominaremos enlaces on/off) garantizando siempre la entrega de los paquetes. En este grafo conviven enlaces candidatos a ser desconectados y otros que deben permanecer siempre conectados. En cada router se permite como máximo desconectar 2 de los 4 enlaces para asegurar la conectividad de la red. De este modo, incluso cuando todos los enlaces posibles están desconectados, la red se mantiene totalmente conectada. Otra restricción estriba en que los enlaces situados en los bordes de la red (a lo largo del perímetro) no pueden desconectarse para asegurar ausencia de bloqueos.

- Un algoritmo de encaminamiento libre de bloqueo que haga llegar los paquetes a su destino independientemente del número total de enlaces desconectados de entre los candidatos del grafo. En primer lugar el algoritmo divide la red en dos redes virtuales separadas o clases de canales virtuales (*VC0* y *VC1*) sin interacción entre ellas. Se asigna un protocolo de encaminamiento distinto para cada una, de modo que la combinación de ambas proporciona un algoritmo de

encaminamiento libre de bloqueo, puesto que cada protocolo forma un ciclo incompleto en el grafo de dependencias[31].

Respecto a los enlaces, hay una serie de características que afectan a la funcionalidad y la eficiencia de un enlace on/off. Un enlace no puede desconectarse inmediatamente debido a las limitaciones de tiempo físicas para reducir el nivel de tensión original del circuito a cero. Además, es necesario un tiempo para resincronizar el reloj cuando un enlace se pone de nuevo en funcionamiento. Hay que considerar, por tanto, el tiempo necesario para conectar y desconectar un enlace, que se expresará en términos de ciclos de reloj y que influye en la implementación de este mecanismo.

En [54] se asume un modelo de consumo sencillo: el consumo de potencia del enlace cuando está activo permanece constante debido a que la potencia nominal y la consumida en el peor caso son similares, y para un enlace se asume potencia cero cuando está desconectado. El enlace no está operativo (no puede transmitir datos) durante las transiciones, pero se asume que el consumo de potencia es el mismo que cuando el enlace está activo y operando normalmente.

Los autores del estudio realizan una implementación sobre un simulador de redes de interconexión dirigido por eventos para topologías de malla bidimensional, empleando una carga de trabajo con distribución uniforme de destinos.

El método que se propone se formaliza en un mecanismo que toma las decisiones de conectar o desconectar un enlace basándose en unos contadores que reflejan el tráfico actual en la red; en concreto la utilización de los buffers de entrada es la métrica que se utiliza en las políticas de encendido y apagado de enlaces. La utilización del buffer de entrada, U_{buffer}, refleja la utilización agregada de los buffers de entrada en un router determinada del siguiente modo:

$$U_{buffer} = \frac{\sum_{p=1}^{P} \sum_{t=1}^{H} \left(F\left(t, p\right) / B \right)}{H \cdot P} \tag{3.6}$$

siendo $0 <= U_{buffer} <= 1$ y donde

- $F(t, p)$ es el número de buffers de entrada que están ocupados en el instante t para el puerto activo p.
- B es el número total de buffers de entrada.
- H es el tamaño de la ventana de muestreo en ciclos de reloj.
- P es el número de puertos de entrada activos en un conmutador.

Cuando U_{buffer} es relativamente bajo, se desconecta un enlace, y cuando U_{buffer} es relativamente alto, entonces un enlace que estaba desconectado se puede reactivar para evitar congestión y por lo tanto degradación en las prestaciones. Para implementar este mecanismo se eligen unos umbrales que deciden la transición del estado del enlace, de modo que cuando U_{buffer} se sitúa por debajo del umbral inferior, un enlace del router se puede desconectar y cuando U_{buffer} sobrepasa el umbral superior el enlace que estaba desconectado se vuelve a conectar. Los enlaces se eligen aleatoriamente de entre los posibles candidatos de acuerdo con el grafo de conectividad.

Los resultados muestran que con un enlace desconectado por router el aumento en latencia es del 29 % y con dos enlaces desconectados por router es del 48.5 %. El ahorro de potencia total en la red es del 21.4 % en el primer caso y del 37.4 % en el segundo. Este mecanismo evita sobrecostes en la comunicación debido a la compartición global de información. Además, los enlaces requieren un hardware más simple para su implementación que en el mecanismo *DVS y* pueden operar a una velocidad mayor. Como contrapartida la necesidad de un algoritmo de encaminamiento especial cuando la red está en modo de ahorro, complica de forma significativa la implementación de este mecanismo en una red real. De hecho, tanto el grafo de conectividad como el encaminamiento libre de bloqueos propuesto en *DPM* son específicos para cada tipo de red, lo que hace compleja la utilización real del mecanismo [8]. En cualquier caso, destacar que las mejoras en el ahorro de potencia son mayores con una política de conexión/desconexión de los enlaces porque los enlaces consumen incluso cuando permanecen ociosos, sin transmisión alguna [54].

En [30] se propone también una técnica basada en desconexión de enlaces para el caso particular de Infiniband basándose en la suposición de que cada destino es alcanzable con dos algoritmos de encaminamiento, *XY* y *SPF*. El algoritmo que toma la decisión de apagar un enlace actúa de la siguiente forma:

- en primer lugar busca el conjunto de enlaces más utilizados, en base a un umbral T_{MAX} para mantenerlos conectados.

- si todos los enlaces están en dicho conjunto, entonces busca apagar enlaces de forma individual utilizando otro umbral denominado T_I.

- en el caso de que no todos los enlaces estén incluidos en el conjunto inicial significa que hay un conjunto de nodos al que se accede con este subconjunto de enlaces menos cargados y sólo con ellos, a partir de ahí de forma recursiva

busca el subconjunto de enlaces mínimo que garantiza conectividad para así desconectar el resto de enlaces.

4

Reducción del Consumo de Potencia en Redes Directas

En este capítulo se justifica la importancia de las redes directas en el ámbito de la supercomputación. Se realiza una descripción del mecanismo propuesto y se definen y analizan sus parámetros característicos. Se incluye una exhaustiva evaluación experimental basada en simulación y se exponen los resultados obtenidos.

4.1. Introducción

En este capítulo se presenta un mecanismo original para reducir el consumo de potencia en redes de interconexión que utilizan topologías regulares construidas con conmutadores de alto grado.

En primer lugar, es significativo destacar que las redes con topologías regulares (en particular los toros 3-D) mantienen una posición muy relevante como las redes de interconexión preferidas en los supercomputadores más potentes. La lista de los TOP500 [66] muestra que, conforme la posición en la lista es más cercana a la cima,

la fracción de computadores basados en topologías regulares para la red de interconexión aumenta: 30 % para los top 100 y 80 % para las top 10. Por otro lado, desde hace ya algún tiempo, los cluster de PCs son una alternativa viable a las máquinas paralelas de gran escala. Actualmente, los clusters utilizan tecnologías de interconexión basadas en el uso de conmutadores y enlaces punto a punto. Myrinet [7], Quadrics [9], InfiniBand [22] o HyperTransport [1] son algunos ejemplos. El progresivo aumento en el tamaño de los clusters ha motivado la utilización de topologías regulares, tanto directas como indirectas (estas últimas serán tratadas en el capítulo 5), en lugar de las irregulares empleadas en un principio.

Por otra parte, el número de puertos por conmutador (grado) ha ido también aumentado progresivamente. Por ejemplo, Mellanox [5] comercializa conmutadores con 36 puertos de 40 Gb/s; y Myricom [7] dispone de conmutadores que llegan hasta 32 puertos de 10Gb/s. Hay varias alternativas para aprovechar la disponibilidad de conmutadores de grado alto. Las redes multietapa son una buena elección[20], [46]. Sin embargo, cuando el tamaño de la red aumenta, la longitud del conexionado y su complejidad crecen notablemente. Por otra parte, las topologías regulares directas (p.e. toro o malla 2-D/3-D) permiten construir redes grandes sin comprometer la longitud del conexionado o su complejidad. De hecho, la disponibilidad de conmutadores de grado alto permite implementar cada dimensión de la red con varios canales físicos en paralelo. Los múltiples canales físicos pueden combinarse para trabajar como un único enlace más ancho (así se incrementa el ancho de banda del canal) o pueden utilizarse para incrementar el número de caminos en la red, aumentando así la flexibilidad de encaminamiento y reduciendo la contención. En ambos casos, se incrementan las prestaciones de la red de interconexión. La técnica consistente en interconectar conmutadores por medio de varios enlaces en paralelo se conoce comúnmente como *link trunking*, *link aggregation* o *port trunking*; en castellano, agregación de enlaces. Incluso una combinación de las dos opciones es posible. Por ejemplo, cuatro puertos InfiniBand 1X se pueden combinar para trabajar como un puerto 4X. Estos, a su vez, se pueden organizar como enlaces agregados compuestos por cuatro enlaces 4X. Por ejemplo, el switch Infiniband de Mellanox InfiniScale IV se puede configurar bien como un switch de 36 puertos 4X o bien como uno de 12 puertos 12X [5]. La figura 4.1 muestra un ejemplo para una malla 2-D donde cada par de conmutadores se conecta mediante un agregado de 4 enlaces, cada uno de ellos formados por canales físicos individuales que trabajan en paralelo componiendo un solo enlace.

Como se muestra en el ejemplo de la figura 4.1, si se emplean enlaces agregados

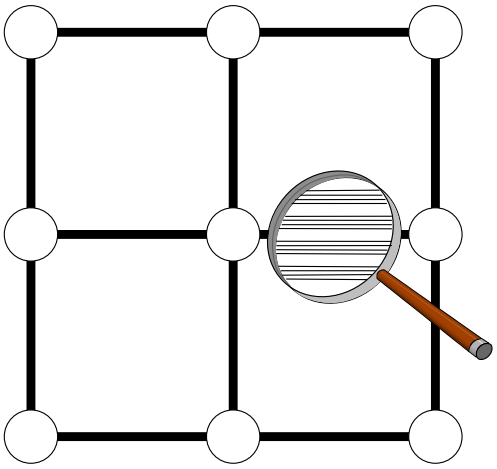


Figura 4.1: Malla 2-D con enlaces múltiples

en cada enlace de la red, se generan varias réplicas físicas de la topología. Esto proporciona flexibilidad en el encaminamiento, puesto que el número de opciones se incrementa significativamente, mejorando también la productividad y reduciendo la congestión. Adicionalmente, si varios enlaces estrechos se combinan en un enlace más ancho, el número de opciones de encaminamiento se reduce pero el ancho de banda del canal se ve incrementado y por lo tanto la latencia de los paquetes disminuye.

Este trabajo se enmarca en este escenario, donde se plantean configuraciones de la red de interconexión en las que se explota la disponibilidad de conmutadores de alto grado, y la agregación de los canales físicos, para conseguir un equilibrio entre la flexibilidad del encaminamiento (proporcionada por los múltiples enlaces disponibles entre cada par de conmutadores) y la disponibilidad del ancho de banda (proporcionada por la combinación de enlaces en un solo enlace más ancho).

El mecanismo de reducción del consumo de potencia, desarrollado originalmente en esta tesis y presentado en este capítulo, se basa en conectar y desconectar canales físicos de la red en función de su utilización, con la restricción de que los enlaces agregados nunca se desconectan completamente. Esta restricción garantiza la conectividad básica entre todos los nodos de la red y, por lo tanto que se pueda utilizar el mismo algoritmo de encaminamiento independientemente del mecanismo de reducción del consumo de potencia, lo que simplifica el diseño del conmutador.

Un aspecto clave, subyacente al mecanismo propuesto, es el hecho de que cuando el tráfico es bajo no se utilizan todos los enlaces, aunque los enlaces ociosos si-

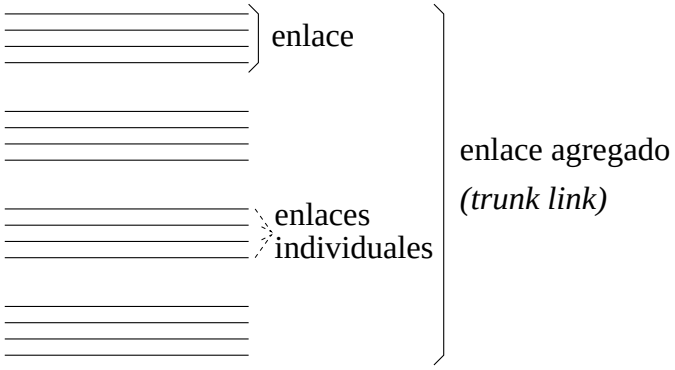


Figura 4.2: Nomenclatura empleada en los enlaces.

guen consumiendo una cantidad significativa de potencia [68]. Diversos experimentos realizados con conmutadores *Gigabit* y *Myrinet* indican que ambos switches que no tienen ningún mecanismo de gestión de la potencia, consumen continuamente 57*W* y 69*W*, respectivamente, independientemente del tráfico de la red. Otro ejemplo es el conmutador $QsNet^{II}$, que consume 36*W* cuando está ocioso y 42*W* cuando hay tráfico. Resulta, por lo tanto, de gran interés aplicar alguna técnica que permita minimizar este consumo de potencia innecesario en los enlaces.

La estrategia planteada en este capítulo se puede aplicar en dos fases o aproximaciones distintas:

- En una primera fase (descrita en la sección 4.2.1), menos agresiva, se actuará únicamente sobre los enlaces que forman un enlace agregado, conectándolos o desconectándolos en función del tráfico. Recuérdese que cada uno de ellos podrá estar compuesto por la agrupación de varios canales físicos (por ejemplo, como en la figura 4.2, cada enlace agregado podría estar compuesto por 4 enlaces 4X, cada uno formado por 4 canales físicos 1X). Esta aproximación permitirá proporcionar una reducción significativa del consumo de potencia en condiciones de poco tráfico, con poco impacto en las prestaciones de la red puesto que los enlaces que se mantienen conectados siguen funcionando en régimen nominal (4X en el ejemplo mencionado).

- En una segunda fase (descrita en la sección 4.2.4), más agresiva, que se aplica cuando solamente queda un enlace conectado. En este caso, se actuará sobre los canales físicos que forman parte de este último enlace, conectándolos o desconectándolos de la misma manera. El efecto es que se varía dinámicamente el

ancho del enlace. Esta aproximación permitirá reducciones muy significativas de potencia para condiciones de poco tráfico pero a costa de una degradación significativa de las prestaciones de la red, dado que los enlaces ven reducido progresivamente su ancho de banda (por ejemplo de 4X a 2X o a 1X). Esta aproximación también sería aplicable en el caso de enlaces paralelos en los que se iría adaptando la anchura física de los enlaces en función del tráfico, con el límite de anchura de 1 bit para garantizar la conectividad o incluso a enlaces serie a los que se pudiera reducir la frecuencia de funcionamiento.

Es importante destacar que la capacidad de los conmutadores de una red determinada para habilitar o deshabilitar canales físicos para ahorrar en el consumo de potencia o de dividir los puertos de un conmutador en puertos más pequeños (proporcionando por consiguiente diferentes anchos de banda) ya está soportado por algunos estándares actuales como *HyperTransport I/O Link Protocol Specification [11]* o *Intel QuickPath Interconnect* [48]. Las propuestas presentadas en esta tesis parten de la base de la disponibilidad de estos mecanismos a nivel hardware.

4.2. Reducción del consumo de potencia en redes directas

4.2.1. Descripción del mecanismo

El mecanismo propuesto se basa en incrementar o reducir dinámicamente el número de enlaces operativos en un enlace agregado o *trunk link*, en función del tráfico presente en la red [19]. Cuando el tráfico es alto, todos los enlaces de un enlace agregado se encontrarán ocupados, y deberán mantenerse en funcionamiento para no provocar una significativa degradación de las prestaciones. Por otra parte, cuando el tráfico es bajo, hasta un máximo de $n - 1$ de los n enlaces que componen el enlace agregado se podrán desconectar para reducir el consumo de potencia. Nótese que, al mantener en todo momento al menos uno de los enlaces siempre operativo, siempre se puede utilizar en la red de interconexión el mismo algoritmo de encaminamiento. Obviamente, cuando los enlaces son desconectados no pueden ser utilizados por el algoritmo de encaminamiento como posibles opciones de comunicación, hasta que no se vuelven a conectar. Además, debe tenerse en cuenta que la conexión y desconexión de los enlaces requiere un tiempo, durante el cual el enlace no puede ser utilizado. En esta tesis supondremos, como hipótesis pesimista, que durante este

tiempo el enlace consume como si estuviera conectado.

Es importante señalar que la restricción de mantener siempre la topología original de la red, a pesar de la conexión y desconexión dinámica de enlaces, es un aspecto clave de la técnica propuesta. Esta limitación evita el uso de complejos algoritmos de encaminamiento adaptativos o tolerantes a fallos [54] o la aplicación de mecanismos de reconfiguración dinámica de la red [43].

Otro de los aspectos fundamentales que requiere el mecanismo propuesto es que cada conmutador debe tener la capacidad de calcular de forma local una estimación del tráfico de la red. Se han realizado distintas propuestas para realizar dicha estimación: basadas en el número de canales virtuales ocupados [41, 54], basadas en la longitud de los buffers asociados a los enlaces [54] o basadas en la utilización del enlace [54]. Ésta última, la utilización del enlace tiene la ventaja de representar fielmente el nivel de tráfico del mismo. Sin embargo tiene el inconveniente de que cuando la red presenta saturación puede llevar a conclusiones erróneas debido a la degradación de prestaciones que sufre [41]: en este caso, el tráfico aceptado (y la utilización del enlace) decrece drásticamente y la latencia del mensaje crece de forma brusca. Como consecuencia, si se toma como métrica la utilización del enlace sin ninguna medida adicional, es posible que se desconecte incorrectamente un enlace cuando la red está congestionada, empeorando por lo tanto una situación ya de por sí mala. Aunque Soteriou y Peh en [54] utilizan como comprobación la utilización de los buffers de entrada, nuestros experimentos concluyen que cuando la red está congestionada, los patrones reales de tráfico pasan a ser atípicos debido a las limitaciones del encaminamiento empleado (como por ejemplo la disponibilidad de los canales de escape, o los requerimientos de asignación atómica de canales). Para superar estas limitaciones, en este trabajo se utiliza una alternativa eficaz pero sencilla: cuando la red está congestionada, suponiendo que todos los nodos están inyectando mensajes a la red (lo que debería ser lo habitual en este caso), los mensajes nuevos generados no se podrán encaminar porque los canales están ocupados durante un periodo de tiempo largo. Como consecuencia, estos mensajes quedan encolados localmente. Se propone el tamaño de esta cola como test de verificación de situaciones de congestión en la red. De esta manera, el mecanismo de ahorro de potencia debería actuar sobre un enlace sólo si su utilización es baja y además la cola de inyección de mensajes del procesador local al conmutador está vacía. Así los enlaces no se desconectarán cuando su utilización sea baja debido a la congestión de la red.

Por otra parte, partiendo de una situación en la que se han desconectado varios enlaces, cuando la utilización de la red crece, el número de enlaces activos en el enlace

agregado se debería ir incrementando progresivamente hasta su valor nominal. En este caso, la congestión de la red puede también disminuir la utilización de los enlaces. Por ejemplo, supongamos que, el tráfico ha sido bajo en el periodo de muestreo anterior; como consecuencia de este tráfico bajo se desconectaron algunos enlaces del enlace agregado y la mayoría de enlaces del enlace agregado están trabajando sólo a una fracción de su ancho de banda total. En ese momento, aparece en la red una ráfaga de tráfico que incrementa significativamente la carga. Aunque la carga generada por la ráfaga de tráfico no sea suficiente para saturar la red en régimen nominal (con todos sus enlaces operativos), puede ser suficiente para congestionar la red cuando trabaja en bajo consumo, llevándole al punto de saturación en el que se degradan notablemente sus prestaciones. Esta situación transitoria de congestión provocaría una reducción de la utilización de los enlaces (del modo en que se ha detallado anteriormente), provocando que los enlaces no se conectaran cuando sí deberían hacerlo y agravando la situación. De nuevo, la ocupación de la cola de inyección debería usarse como un método de comprobación para detectar la congestión en la red. Cuando una situación como la descrita es detectada, los enlaces deberían ser conectados de la manera más rápida posible para que trabajen de nuevo al máximo ancho de banda disponible con el objeto de minimizar el impacto de la congestión.

El tráfico a través de los enlaces agregados se mide a partir de la utilización de cada uno de sus enlaces, las cuales se obtienen mediante contadores que se incrementan cada vez que se transmite un phit por cualquiera de ellos. La utilización en un periodo determinado de tiempo se puede obtener dividiendo el valor del contador por el número de ciclos de reloj transcurridos, suponiendo que los enlaces tienen un ancho de banda de 1 phit por ciclo de reloj. Es importante señalar que, a efectos de implementación en hardware del mecanismo, no sería necesario realizar la división ya que, al tratarse de periodos de tiempo fijos, se emplearía como métrica de utilización el número de phits transmitidos. Este contador se comprueba periódicamente a los efectos de aplicar, si procede, alguna de las acciones de ahorro de potencia.

Tras obtener la utilización de un enlace agregado, sumando las utilizaciones de los enlaces que lo componen y dividiendo por el número de enlaces activos, se actúa en función de las siguientes reglas generales (ver figura 4.3):

- Se definen dos umbrales:
  - U_{on}, que define la utilización para la cual debe intentarse conectar un enlace adicional; y
  - U_{off} , que define la utilización para la cual se debe desconectar un enlace.

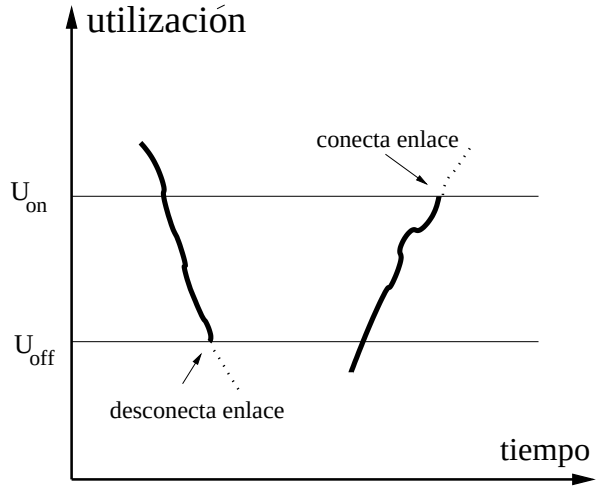


Figura 4.3: Conexión y desconexión de los enlaces en función de los umbrales U_{on} y U_{off} .

- Si la utilización del enlace agregado decrece por debajo del umbral de desconexión U_{off} , uno de sus enlaces se desconecta sólo si éste no está siendo utilizado por algún paquete. Tal y como se ha comentado anteriormente, los enlaces sólo se pueden desconectar si la cola de inyección del procesador local está vacía. Además, se debe prestar especial atención para evitar el apagado del último enlace operativo en el enlace agregado, es decir, esta decisión se puede tomar sólo si el número de enlaces activos es mayor que la unidad. Los enlaces agregados nunca se pueden desconectar por completo para garantizar que se puede seguir utilizando el mismo algoritmo de encaminamiento.

- Si la utilización del enlace agregado aumenta por encima del umbral de conexión U_{on}, y existe un enlace disponible para su conexión, este enlace se conecta. Como se ha indicado anteriormente, es necesario tener en cuenta una consideración adicional para evitar congestión en la red. Una red con enlaces desconectados, es decir con baja capacidad máxima respecto a la situación nominal, es muy sensible a la congestión ante incrementos rápidos del tráfico. Podría darse el caso de que la utilización de los canales nunca aumentara (debido a la propia saturación) y los enlaces ni siquiera se reconectarían. Es más, cuando el tráfico crece en un conmutador, en el que uno o más de uno de sus enlaces están desconectados, el retardo necesario para secuencialmente reconectar los enlaces puede ser demasiado largo, llevando potencialmente a la saturación

de la red. Para detectar esta situación, el mecanismo también comprueba si hay mensajes pendientes generados localmente esperando a ser inyectados en la red (en la cola de inyección). Si los hay, es una señal que alerta de congestión en la red. Si se detecta esta situación, todos los enlaces que están desconectados en el conmutador se conectan. Este test de detección de situaciones de congestión trabaja asíncronamente con el mecanismo de ahorro de potencia para aumentar la sensibilidad o capacidad de respuesta del sistema.

4.2.2. Parámetros de control del mecanismo. Agresividad y sensibilidad.

El comportamiento del sistema vendrá condicionado por la pareja de umbrales U_{on} y U_{off}. El efecto de estos dos parámetros se puede analizar en términos de dos aspectos complementarios:

- Agresividad del mecanismo. Viene determinada por el valor medio de los dos umbrales, u_{avg} siendo

$$u_{avg} = \frac{(U_{on} + U_{\text{off}})}{2} \tag{4.1}$$

Un valor alto de u_{avg} proporciona una política de reducción del consumo agresiva, ya que los enlaces se desconectan con cargas relativamente altas, y se mantienen desconectados incluso con cargas muy altas. Por otra parte, si u_{avg} es bajo, se aplicará una política conservadora ya que la desconexión de enlaces (y el consiguiente ahorro de consumo de potencia) sólo se activará con cargas muy bajas y la reconexión también se hace para cargas relativamente bajas.

- Capacidad de respuesta (*responsiveness*) o sensibilidad del mecanismo. El ancho de la banda de histéresis, definido como la diferencia entre $U_{on} - U_{\text{off}}$, controla la sensibilidad del mecanismo ante variaciones en el tráfico. Con valores elevados de anchura de la banda de histéresis, el mecanismo es poco sensible: serán necesarias variaciones significativas en el tráfico para que se conecten o desconecten enlaces, mientras que pequeñas variaciones de tráfico no afectarán al estado del sistema. Un valor demasiado alto haría inútil al mecanismo. En el caso contrario, con una banda de histéresis estrecha, pequeñas variaciones en la utilización de los enlaces provocarán conexiones y desconexiones de enlaces. Un valor demasiado bajo provocaría constantes cambios en el modo de funcionamiento de los enlaces y la sobrecarga debida al propio proceso de co-

nexiones/desconexiones pudiera resultar en una degradación de prestaciones sin ahorro de potencia.

Teniendo en cuenta las consideraciones anteriores, nuestra propuesta puede sintonizarse para conseguir distintos comportamientos en base a ajustar la agresividad y la sensibilidad. Sin embargo, aparecen ciertas limitaciones en relación a los valores posibles de los umbrales:

- Ambos umbrales deben ser positivos y distintos de cero, $U_{on} > 0$ y $U_{\text{off}} > 0$.
- U_{on} debe ser mayor que U_{off}, $U_{on} > U_{\text{off}}$.
- U_{on} debe ser menor que la utilización máxima alcanzada por los enlaces con la carga más alta aceptada por la red (U_{MAX}). En caso contrario, la red entraría en saturación antes de intentar conectar los enlaces desconectados si los hubiere. Por tanto, $U_{on} < U_{MAX}$
- Finalmente, la diferencia entre los umbrales debe ser suficiente para evitar la presencia de ciclos de conexión/desconexión. Seguidamente se explica este aspecto utilizando un sencillo ejemplo.

Consideremos una red basada en enlaces agregados con cuatro enlaces (figura 4.4). En este caso, habrá cuatro posibles estados del enlace agregado, y, consecuentemente, cuatro anchos de banda posibles. El ancho de banda B se alcanzará cuando estén conectados los cuatro enlaces (*estado S4*); B' se alcanzará con tres conectados y uno desconectado (*estado S3*); B'' con dos enlaces conectados y dos desconectados (*estado S2*); y B''' cuando se encuentre un enlace conectado y tres desconectados (*estado S1*). Como se indicó anteriormente, al menos un enlace del enlace agregado debe estar conectado.

Consideremos la situación en la que el enlace agregado de la figura 4.4 está en el estado *S4*. Si la utilización del enlace agregado disminuye por debajo del umbral de desconexión U_{off}, el mecanismo propuesto desconectará un enlace. El ancho de banda disponible se reducirá a $3/4$ del nominal, y suponiendo que no ha habido variación en el tráfico, la utilización del enlace agregado relativa a la nueva capacidad del enlace aumentará en la proporción inversa, $4/3$ (para la transición entre *S3* y *S2* la utilización escalaría en un factor $3/2$ y para la transición entre *S2* y *S1* el factor aplicado sería $2/1$). Este aumento en la utilización se detectará en el siguiente muestreo, y, en función de los umbrales escogidos, puede ocurrir que la utilización medida, u, sea mayor que U_{on} $(u > U_{on})$, volviendo a conectar el enlace desconectado previamente. Este fenómeno podría producir oscilaciones en el estado del enlace agregado.

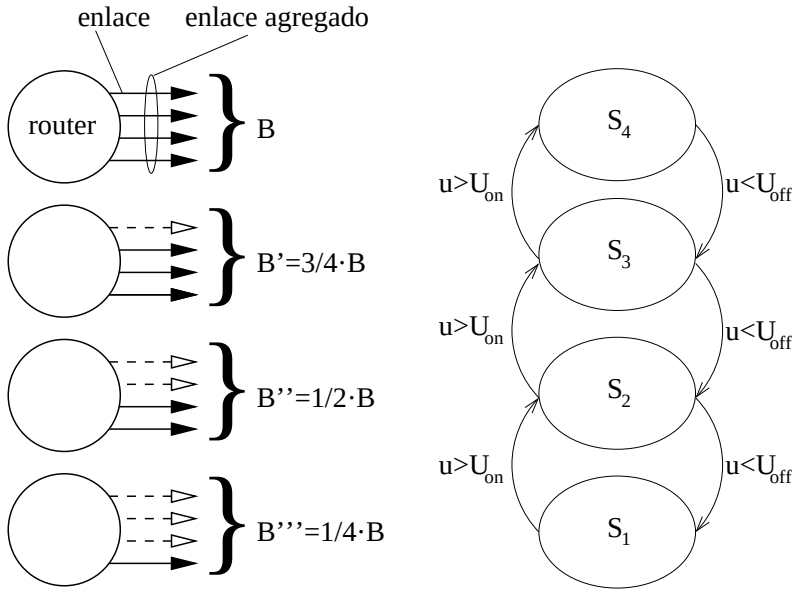


Figura 4.4: Ancho de banda del enlace agregado en función del estado del enlace agregado, S_i donde i indica el número de enlaces activos.

Como el encendido y apagado de los enlaces requiere un tiempo no despreciable y durante este tiempo el enlace no puede ser utilizado pero consume como si estuviera conectado, estas oscilaciones deben evitarse.

La figura 4.5 muestra un ejemplo de estas oscilaciones en el comportamiento del mecanismo para $U_{on} = 0{,}6$, $U_{\text{off}} = 0{,}4$ y una utilización en el enlace agregado de $u = 0{,}18$. En este caso, el mecanismo desconecta, durante tres iteraciones consecutivas, 3 enlaces de los 4 (en el estado S_1 sólo queda un enlace conectado). Analizando paso a paso la secuencia de operaciones, en la primera iteración se comprueba que la utilización del enlace agregado es inferior al umbral de desconexión, $0{,}18 < 0{,}4$, y se desconecta un enlace. Con la premisa de que la tasa de tráfico se mantiene constante durante el ejemplo, la utilización del enlace agregado tras desconectar el primer enlace se escalará por $4/3$, de forma que la utilización medida será $u' = 4/3 \times 0{,}18 = 0{,}24$. En la segunda iteración, de nuevo se producirá la desconexión de un enlace, pues $u' = 0{,}24 < 0{,}4$. Como consecuencia, la utilización relativa del enlace agregado aumentará, en esta caso en un factor $3/2$, de forma que $u'' = 3/2 \times 0{,}24 = 0{,}36$. Nuevamente se producirá una desconexión más pues $u'' = 0{,}36 < 0{,}4$ y la utilización en el único enlace operativo será $u''' = 0{,}72$. La carga efectiva en este enlace (y en el enlace agregado reducido ahora a un solo enlace) se ha multiplicado por 4 al pasar

de 4 enlaces a 1 enlace operativo tras las desconexiones:

$$u''' = 0{,}18 \times \frac{4}{3} \times \frac{3}{2} \times \frac{2}{1} = 0{,}72 \tag{4.2}$$

Puesto que esta utilización es mayor que U_{on} , $u''' > 0{,}6$, se dispara una transición a S_2, reduciendo la utilización medida en el enlace agregado de nuevo al valor $u'' = 0{,}36$ y entrando así en un ciclo entre S_2 y S_1, como se ha descrito en el párrafo anterior. En particular, las oscilaciones o ciclos entre estados podrían ser especialmente perjudiciales si se producen entre los estados S_1 y S_4. Esto podría ocurrir como consecuencia del test de detección de la saturación mencionado anteriormente. Si, en el ejemplo que se muestra en la figura 4.5 se detecta congestión en el estado S_1, el mecanismo conectará todos los enlaces, alcanzando el estado S_4. Esto hará que el enlace agregado empiece otro ciclo pasando a través de los mismos estados. Afortunadamente, esto no ocurrirá si el mecanismo se configura de manera que nos aseguramos de que el tráfico de la red sea lo suficientemente bajo antes de realizar desconexiones de enlaces (esto es, se utilizan valores de U_{off} lo suficientemente bajos). Por consiguiente, la selección de los umbrales viene condicionada por la posibilidad de transiciones de estado cíclicas que hagan que el sistema se transforme en inestable.

Para asegurar ausencia total de ciclos en el diagrama de estados de un enlace agregado, el valor del umbral U_{on} debe ser mayor que la utilización media resultante del enlace agregado tras la desconexión de un enlace cuando se pasa del estado S_i al S_{i-1}. El incremento en la utilización durante la desconexión de enlaces depende del ratio entre los anchos de banda de los estados origen y destino en cada transición. En nuestro ejemplo con cuatro enlaces por cada enlace agregado, estos ratios son $4/3$, $3/2$, y $2/1$ cuando hay respectivamente 4, 3, 2 enlaces conectados y el mecanismo desconecta uno de ellos. El peor caso ocurre cuando se pasa del estado S_2 al estado S_1, pues la utilización se multiplica por un factor $2/1$. Así pues, para evitar transiciones de estado cíclicas, se debe cumplir que $U_{on} \geq 2U_{\text{off}}$ con lo que en el peor caso, cuando se cambia del estado S_2 al S_1, el incremento en la utilización media del enlace agregado no hará que se vuelva al estado anterior.

Por lo tanto, teniendo en cuenta todas las consideraciones anteriores, las restricciones para elegir los umbrales U_{off} y U_{on} son:

- $U_{\text{off}} > 0$
- $U_{on} < U_{MAX}$

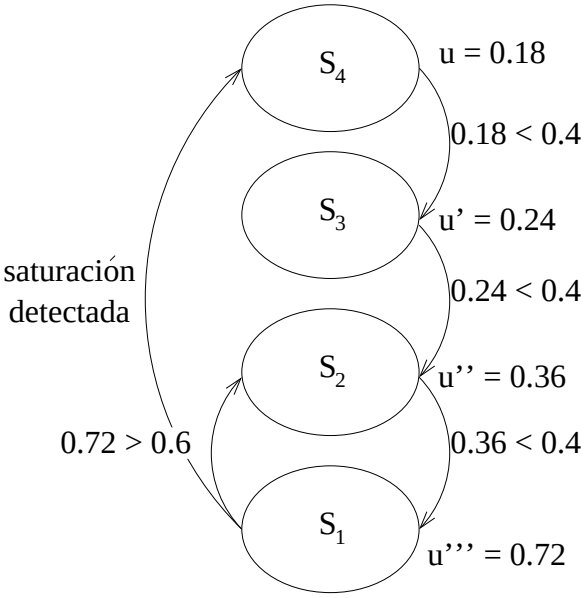


Figura 4.5: Ejemplo de las transiciones entre estados en un enlace agregado con dos posibles ciclos: S_2, S_1,S_2,S_1... y S_4,S_3,S_2,S_1,S_4,S_3,S_2,S_1.... En este caso $U_{on=}$0,6 y $U_{off} = 0{,}4$

- $U_{on} \geq 2U_{\text{off}}$

Respecto a la tercera restricción, se puede reescribir de la siguiente forma: $U_{\text{off}} \leq U_{on}/2$ y aplicando entonces la segunda restricción tendremos: $U_{\text{off}} < U_{MAX}/2$. Esta condición ayuda a evitar el ciclo entre el estado donde aparece el ancho de banda al completo (S_4 en la figura 4.5) y el estado con el ancho de banda menor (S_1 en la figura 4.5).

La figura 4.6 muestra la región de los umbrales posibles teniendo en cuenta las restricciones anteriores. Cualquier punto dentro de la zona sombreada proporciona una configuración válida de la política de ahorro de potencia, con distintas combinaciones de sensibilidad y agresividad. Este diagrama puede aplicarse a sistemas basados en conmutadores con cualquier número de enlaces por enlace agregado, donde $n-1$ de los n enlaces pueden desconectarse. Ello se debe a que, en cualquier caso, el sistema tendrá una transición entre un estado con dos enlaces conectados a un estado con sólo un enlace conectado. Y ése es el caso peor de incremento de utilización por desconexión de enlaces (es decir, el más propenso a provocar ciclos), debiendo así mantenerse la condición indicada anteriormente, $U_{on} \geq 2U_{\text{off}}$, que garantiza ausen-

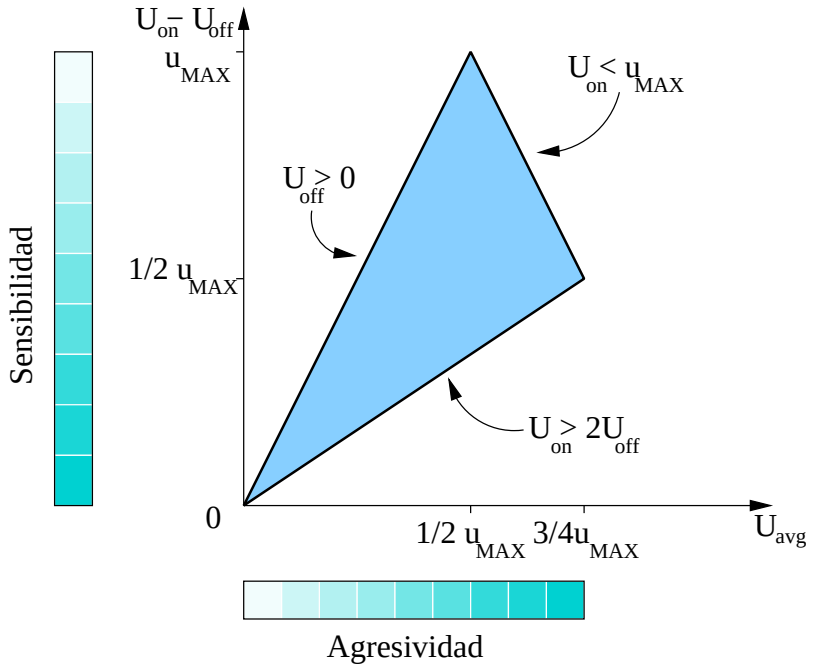


Figura 4.6: Mapa de posibles umbrales. El área sombreada corresponde con el espacio de valores válidos para U_{on} y U_{off}.

cia de ciclos en ese caso.

4.2.3. Comportamiento dinámico

A pesar de que los umbrales que definen una configuración determinada pueden parecer estáticos, por estar definidos con valores constantes, el mecanismo de ahorro de potencia intrínsecamente proporciona una adaptación dinámica de los umbrales de acuerdo con el estado del enlace agregado.

El mecanismo de ahorro de potencia usa como métrica para estimar el tráfico la utilización del enlace agregado medida con respecto al ancho de banda disponible, no sobre el ancho de banda total del enlace agregado. Por esta razón, y tal como se ha descrito en la sección anterior, cuando, por ejemplo, la carga de un enlace agregado hace que éste cambie del estado S_4 al estado S_3, la utilización se ve incrementada en un factor $4/3$ (el mismo tráfico circula por un enlace que ha reducido su ancho de banda). Otro punto de vista alternativo corresponde al caso de que la utilización se mida respecto al ancho de banda total del enlace agregado. En este caso, el me-

	Umbrales estáticos			Umbrales efectivos	
Enlaces activos	On	Off	Factor	On	Off
4		U_{off}	1		U_{off}
3	U_{on}	U_{off}	$3/4$	$3/4U_{on}$	$3/4U_{\text{off}}$
2	U_{on}	U_{off}	$2/4$	$2/4U_{on}$	$2/4U_{\text{off}}$
1	U_{on}		$1/4$	$1/4U_{on}$	

Tabla 4.1: Valores de los umbrales estáticos en función de los enlaces disponibles para un switch con cuatro enlaces por enlace agregado.

canismo se comporta como si los umbrales se ajustaran al estado del enlace. Para el caso de la transición del estado S_4 al estado S_3 es como si los umbrales hubieran decrecido exactamente según la inversa de este factor, es decir, $3/4$ de su valor original. Con este enfoque, para cualquier estado dado S_i, los umbrales efectivos relativos a la carga absoluta del enlace agregado se escalan a $i/i+1$. La tabla 4.1 recoge los umbrales originales y su valor efectivo para una red con una agregación de cuatro enlaces por cada enlace agregado. Es importante destacar que en el caso extremo de cuatro enlaces activos, el umbral U_{on} no se aplicará nunca dado que no se pueden conectar enlaces adicionales; de forma equivalente, en el caso de un solo enlace activo, el umbral U_{off} no actúa ya que no se pueden desconectar más enlaces. Teniendo en cuenta la información recogida en la tabla, la utilización mínima para tener todos los enlaces activos es $3/4U_{on}$, mientras que $2/4U_{\text{off}}$ es la utilización máxima que garantiza el ahorro de potencia máximo para un switch en particular.

La figura 4.7 muestra un diagrama representando el estado del enlace frente a la carga del enlace agregado, cuando los umbrales de conexión y desconexión (*on/off*) están en $U_{on} = 0{,}4$ y $U_{\text{off}} = 0{,}16$. En esta figura, la carga corresponde a la utilización medida sobre el ancho de banda nominal (correspondiente al estado S_4 en la figura 4.5). Tal y como se indica en las flechas del gráfico, las transiciones en la dirección descendente pueden producirse sólo siguiendo las lineas verticales de la izquierda, mientras que las transiciones ascendentes ocurren siguiendo las lineas verticales de la derecha. El efecto que se consigue es que, a medida que el número de enlaces conectados decrece, la agresividad del mecanismo decrece y viceversa. Por otra parte, cuando el número de enlaces conectados disminuye, la sensibilidad aumenta y a la inversa. Este comportamiento es ventajoso para el mecanismo: al reducirse la disponibilidad de enlaces con la desconexión, las probabilidades de que se produzcan situaciones de congestión por incrementos de tráfico menores aumentan. Que el mecanismo sea menos agresivo reduce la tendencia a la desconexión a la vez que se

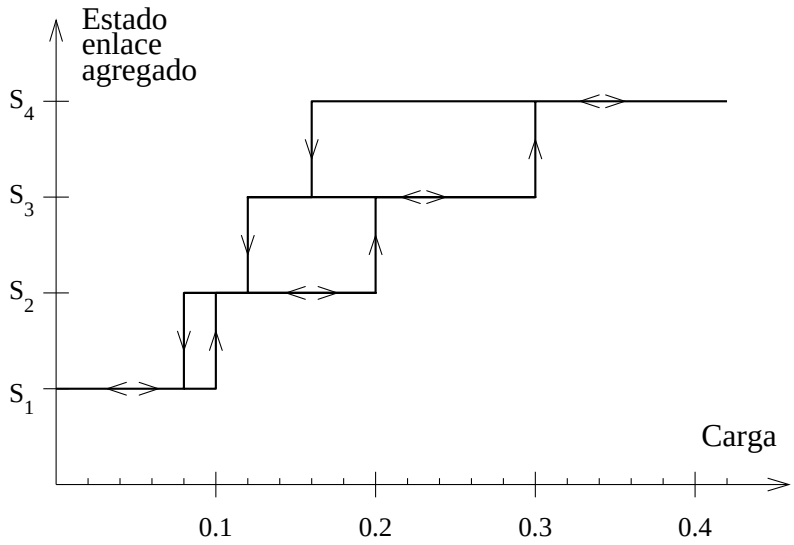


Figura 4.7: Diagrama de histéresis para $U_{on} = 0{,}4$ y $U_{\text{off}} = 0{,}16$.

vuelve más sensible a los cambios de utilización y por tanto aumenta su agilidad para responder a cambios de tráfico de menor magnitud.

4.2.4. Reducción del consumo con un solo enlace conectado.

Cuando la red de interconexión no contempla el uso de enlaces agregados, o teniéndolos sólo queda un enlace del enlace agregado conectado (recuérdese la nomenclatura fijada en la figura 4.2), todavía es posible conseguir reducciones de potencia adicionales [14]. Con ese objeto, hemos diseñado una extensión del mecanismo presentado en las secciones anteriores que permite conseguir ese ahorro extra. Se basa en medir la utilización del único enlace disponible y dinámicamente ajustar el ancho de banda del enlace a las necesidades exigidas por el tráfico en cada momento. El ajuste del ancho de banda del enlace se puede implementar de diferentes maneras dependiendo de la configuración de la red. En el caso empleado como referencia (mostrado en la figura 4.2), se puede ajustar este parámetro por medio de la conexión/desconexión dinámica de los enlaces individuales de un enlace. Alternativamente se podría ajustar el ancho de un enlace (en enlaces paralelos) o su frecuencia de funcionamiento. De la misma manera que con la desconexión de enlaces descrita en las secciones anteriores, siempre se mantendrá operativo al menos un enlace individual (o parte de un enlace en el caso de una implementación basada en enlaces

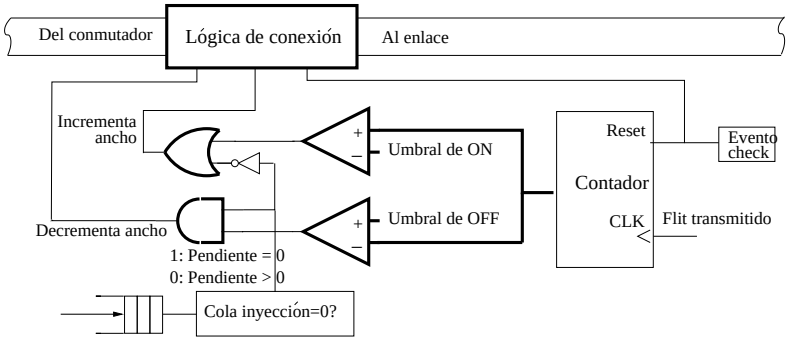


Figura 4.8: Hardware necesario para implementar el esquema propuesto de ahorro de potencia

paralelos), y por ello el algoritmo de encaminamiento de la red no se ve afectado ni requiere funcionalidades especiales.

Utilizando este enfoque, la potencia se puede reducir dinámicamente a estados de menor consumo. El límite inferior del consumo de potencia puede ser tan bajo como el ancho de banda mínimo del enlace que garantiza la conectividad: en el caso de enlaces agregados sería suficiente mantener un enlace individual; en un enlace paralelo, bastaría con mantener una conexión mínima de un bit. Con el objeto de simplificar la implementación de este mecanismo hemos limitado los ajustes válidos del ancho de banda a algunos tamaños predefinidos. En particular, como el ancho del enlace (en número de bits o canales físicos combinados) normalmente es potencia de dos, cada uno de las reducciones de anchura se obtendrán dividiendo por dos la anchura anterior. Por ejemplo, un enlace $4X$ resultado de combinar 4 canales físicos ($1X$) podría reducirse a $2X$ y $1X$. En el caso de una conexión paralela, un link de 8 bits de ancho se podría reducir a 4, 2 y 1 bit. En este caso, un flit de 8 bits se transmitirá por medio de 1, 2, 4 o 8 phits, respectivamente. De esta manera, el diseño de la lógica de serialización/deserialización es relativamente simple. La figura 4.8 muestra de forma simplificada el hardware que se necesita para implementar el mecanismo [17]. Como se ha indicado anteriormente, la funcionalidad de red necesaria para ajustar dinámicamente el ancho del enlace ya está definido en los actuales estándares de redes de interconexión de alta velocidad *[11, 48]*.

El precio a pagar con este mecanismo es que la latencia del mensaje se verá incrementada con cargas bajas, cuando los mensajes están atravesando enlaces estrechos, puesto que uno de los componentes de la latencia de los paquetes depende del ancho de banda del enlace. A pesar de que esto puede ser una preocupación para las aplica-

ciones sensibles a la latencia, los beneficios en ahorro de potencia pueden compensar el incremento de la latencia. La forma de intentar mitigar este problema es utilizando una función de selección de los enlaces de salida en los conmutadores que dé prioridad a los enlaces más anchos [27]. En realidad, tanto la anchura del enlace como el grado de multiplexación de los canales virtuales, si lo hay, deberían ser considerados en la función de selección. Se selecciona, en primer lugar, el enlace más ancho; si hay varios enlaces con el mismo ancho de banda entonces, en segundo lugar, se selecciona aquél con el menor número de canales virtuales ocupados. Si hubiera varios en igualdad de condiciones, entonces se selecciona el primero disponible. También es factible deshabilitar el mecanismo de ahorro de potencia cuando se está ejecutando una aplicación sensible a la latencia.

La forma en que se calcula el tráfico cuando se aplica este mecanismo extra de ahorro de potencia podría ser muy similar a la que se utiliza cuando se desconectan enlaces en un enlace agregado: un contador que se incrementa cada vez que se transfiere un phit por el enlace. Este contador se lee periódicamente para calcular la utilización de cada enlace, dividiendo el valor del contador por el número de ciclos de reloj transcurridos. Este contador se resetea tras cada periodo de comprobación. Teniendo en cuenta que el mecanismo incrementará o decrementará el ancho efectivo de cada enlace es importante destacar que es una acción dirigida a cada uno de los enlaces individualmente y no basada en la utilización global del conmutador [54].

En función de que la utilización u del enlace sea menor que el valor del umbral U_{off} o mayor que U_{on} (ver 4.3), la anchura del enlace se aumentará o se reducirá, respectivamente.

Los parámetros de diseño que influyen en el comportamiento del mecanismo son:

- los valores concretos de los umbrales,
- factor de reducción/aumento del ancho del enlace,
- y cuántas veces se puede reducir progresivamente el ancho del enlace

Respecto al primer aspecto, las consideraciones para elegir los umbrales son idénticas a las propuestas en el apartado 4.2.2.

En relación con los dos últimos parámetros, el número de reducciones que se permiten es una especificación del diseño que afecta al límite inferior del consumo de potencia del enlace. En este trabajo se proponen dos reducciones del ancho inicial del enlace, imponiendo por lo tanto un límite inferior del 25 % del consumo nominal de

potencia, trabajando el enlace al 100 %, 50 % o 25 % del ancho nominal. Es importante destacar en este momento que son posibles más reducciones del ancho del enlace, pero imponen una severa penalización en la latencia para cargas bajas, como hemos comprobado a nivel experimental.

4.3. Evaluación de prestaciones del mecanismo propuesto

En esta sección, se evalúa el comportamiento del mecanismo de ahorro de potencia propuesto en la sección 4.2. Mediante simulación, se calcula la reducción de potencia que se obtiene al utilizar los mecanismos propuestos y se cuantifica el impacto sobre las prestaciones de la red. Los resultados obtenidos con estas técnicas se comparan con las prestaciones que proporciona el sistema original, sin el mecanismo de ahorro de potencia.

Las métricas de prestaciones empleadas son dos:

- la latencia media de un mensaje medida desde el momento de su generación en el nodo fuente hasta el momento de recepción por su destino.
- el consumo de potencia de los enlaces relativo al sistema original, es decir, con la red en régimen de funcionamiento normal, sin aplicación de mecanismo alguno. Por tanto, la potencia se expresa de forma relativa como un porcentaje del consumo total de la red cuando están todos los enlaces activos. Sólo se tiene en cuenta la potencia disipada en los enlaces puesto que este trabajo se centra exclusivamente en dicho componente de la red. Es por ello que se evalúa el impacto de esta técnica de forma aislada de otras propuestas que se podrían aplicar en diferentes partes de la red (como por ejemplo en los buffers de los conmutadores).

4.3.1. Modelo de red

Nuestro simulador modela una red de interconexión basada en wormhole a nivel de flit [27]. Cada nodo de la red está compuesto por un procesador, su memoria local y un conmutador o router. El router[1] contiene una unidad de encaminamiento y arbitraje, un conmutador o switch interno, y varios enlaces físicos. Todos estos

[1]Este router se denomina así por tradición en este campo, redes de interconexión regulares [28, 26], pero corresponde a lo que, hasta ahora, hemos llamado conmutador o switch.

componentes realizan su tarea en 1 ciclo de reloj. Hay cuatro canales de memoria independientes conectando el conmutador con la memoria local. Se utiliza un algoritmo de encaminamiento totalmente adaptativo y libre de bloqueos [27].

Los nodos se interconectan empleando la técnica de agregación de enlaces descrita en la sección 4.1 y detallada en la figura 4.2. Los enlaces agregados dispuestos entre cada pareja de nodos están compuestos por 4 enlaces, siendo cada uno de ellos el resultado de agrupar 4 canales físicos. Los canales físicos se descomponen en tres canales virtuales (uno adaptativo y dos de escape). Cada canal virtual tiene asociado un buffer con capacidad para cuatro flits. En cada nodo se incluye la implementación del mecanismo de ahorro de potencia propuesto en este trabajo. Se han evaluado, por representar las arquitecturas más populares actualmente, redes con topología de toro 2D de 16×16 nodos y 3D de $8 \times 8 \times 8$ nodos.

4.3.2. Modelo de tráfico

El patrón de tráfico en la red se define en base a los siguientes parámetros: la distribución espacial de los destinos, la tasa media de inyección de los mensajes, junto con su distribución temporal, y la longitud de los mensajes.

Los experimentos han sido diseñados con el objetivo de presentar situaciones, por un lado representativas de cargas reales y empleadas en el análisis de prestaciones de redes de interconexión, y por otro lado poco favorables a que el mecanismo propuesto presentara un funcionamiento óptimo. El peor caso lo constituyen cargas que a largo o medio término no presentan variaciones temporales ni espaciales como el tráfico uniforme aleatorio (distribución espacial y temporal uniforme) [54]. Ello es debido a que los flujos de tráfico son distribuidos uniformemente por la red a lo largo del tiempo. De esa manera es virtualmente imposible detectar cambios en la utilización de la red que puedan disparar sistemáticamente el mecanismo de reducción del consumo de potencia y maximicen el ahorro con un mínimo impacto negativo en las prestaciones. Con tráfico intenso es difícil desconectar un enlace y para tráfico ligero prácticamente todos los candidatos a la desconexión serán desconectados, llevando al sistema al nivel más bajo de consumo, con mayor impacto en las prestaciones. El extremo opuesto sería un patrón de tráfico con una distribución irregular que permitiera tener desconectadas de forma permanente secciones de la red por las que nunca circula tráfico.

Como se ha justificado, con el objeto de evaluar el comportamiento de nuestro mecanismo en las peores condiciones posibles, el modelo de tráfico se basa en una

distribución uniforme de destinos. Con respecto a los tiempos de inyección de cada mensaje la mayoría de los experimentos se han realizado con una distribución uniforme (aleatoria) para evaluar el peor caso. Alternativamente, se presentan algunos resultados en los que se emplea una distribución auto-similar debido a la naturaleza auto-similar del tráfico Ethernet [40], tan popular en sistemas tanto domésticos como de altas prestaciones en el momento de redactar esta memoria (Gigabit Ethernet representa el 52 % de las tecnologías de red presentes en el TOP500).

Puesto que se ha evaluado nuestro mecanismo tanto en condiciones estáticas como dinámicas, el valor medio de la tasa de inyección es constante para la evaluación estática y variable para el análisis de carga dinámica. Primero se analiza el comportamiento estático de la red ejecutando simulaciones independientes con tasas de inyección de mensajes de media constante, evaluando el rango completo de tráfico, partiendo de una carga baja hasta llegar a la saturación (ver sección 4.3.5. El segundo grupo de experimentos tiene como objetivo estudiar el comportamiento dinámico de la red, utilizando tasas de inyección variable durante cada simulación (ver sección 4.3.6).

El tamaño del mensaje se ha fijado a 16 flits, excepto para el análisis del efecto de la longitud de los mensajes donde se han probado mensajes largos de 256 flits. Como se comprobará la sección 4.3.4.3, de nuevo se ha optado por evaluar el mecanismo con una configuración conservadora. Es decir, los mensajes cortos no favorecen el ahorro de potencia que se consigue con el mecanismo propuesto.

4.3.3. Parámetros del mecanismo propuesto

La dinámica del modelo está gobernada por los umbrales de conexión U_{on} y de desconexión U_{off}. La figura 4.6 muestra el área de valores posibles para estos umbrales. En los experimentos mostrados, se explora dicha área seleccionando diferentes valores para los umbrales U_{on} y U_{off} con el objetivo de alcanzar distintas metas de agresividad y sensibilidad para el mecanismo de ahorro de potencia.

Por otra parte, un enlace no se puede conectar o desconectar de forma instantánea. El tiempo necesario para incrementar y reducir el ancho de banda de un enlace, T_{on} y T_{off}depende del retardo necesario para reactivar algunas de sus lineas. Si nos basamos en los valores indicados en [30, 54] podemos considerar unos valores $T_{on} = T_{off} = 1000\,ciclos$.

Cuanto mayores sean estos tiempos, menos adaptable será el mecanismo a cambios en la carga de la red. Por otra parte, durante estas transiciones se asumirá el caso

peor:

- cuando se está realizando la desconexión de un enlace se asume que el ancho de banda disminuye de forma instantánea pero que continua consumiendo la potencia del estado inicial hasta que transcurren T_{off} instantes de tiempo.
- cuando se está realizando la conexión de un enlace se asume que el nuevo ancho no está disponible hasta transcurridos T_{on} ciclos, pero que el consumo de potencia se incrementa desde el instante inicial.

Por otra parte, las decisiones sobre el ajuste de estado del enlace se deben hacer con un periodo mayor que T_{on} y T_{off} para permitir que los enlaces se estabilicen después de los cambios. En concreto y después de realizar diversas pruebas, se ha fijado el periodo de chequeo en 2000 ciclos de reloj.

Mientras no se indique lo contrario los enlaces agregados están inicialmente al 100 % de su ancho nominal, es decir todos los enlaces están conectados. Finalmente, se asume que el consumo de potencia de cada enlace es proporcional a su ancho real (o a la fracción del mismo que está conectada).

4.3.4. Evaluación de las prestaciones básicas de la red

Se presenta en esta sección una evaluación de las prestaciones de la red, como referencia para el posterior análisis del comportamiento de la red cuando el mecanismo de ahorro de potencia está activado.

4.3.4.1. Efecto de la función de selección

La función de encaminamiento y la función de selección constituyen el algoritmo de encaminamiento [28]. La función de encaminamiento suministra un conjunto de canales de salida que permiten llegar a un nodo destino desde el nodo actual. La función de selección escoge un canal libre (si es posible) de entre los suministrados por la función de encaminamiento de acuerdo a una determinada estrategia. Si bien la función de encaminamiento determina si un algoritmo de encaminamiento es libre de bloqueos o no, la función de selección solamente tiene efecto sobre las prestaciones.

Hemos considerado relevante incluir este apartado ya que la función de selección determina cómo se distribuye el tráfico entre los canales disponibles y ello puede tener impacto en las posibilidades para desconectar enlaces cuando la carga sea baja. Por ejemplo, una función de selección que concentre todo el tráfico en unos pocos

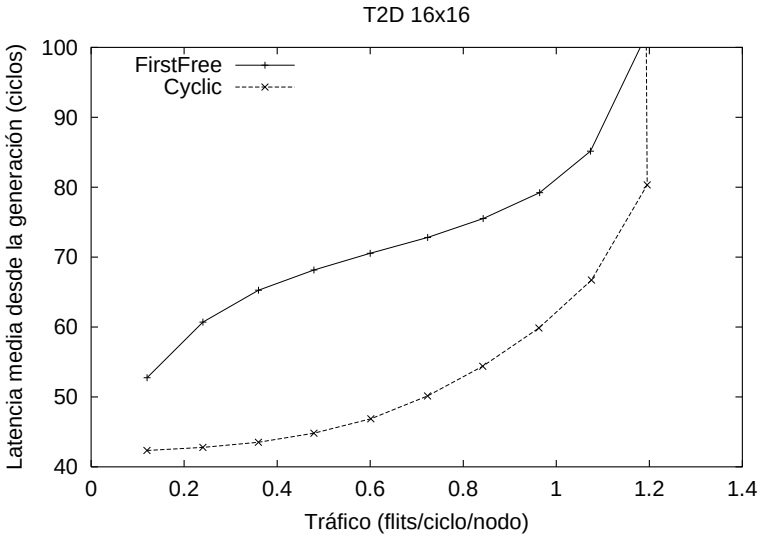
canales físicos permitirá desconectar los otros y por lo tanto reducir el consumo de potencia con más facilidad que en el caso de que el tráfico se distribuya por todos lo canales disponibles.

Las funciones de selección evaluadas han sido dos y los resultados obtenidos para la red en régimen nominal (sin desconectar ningún enlace) se muestran en la figura 4.9. La primera función de selección, etiquetada *FirstFree* en las gráficas, escoge siempre el primer canal virtual libre de entre los proporcionados por la función de encaminamiento; tiende pues a concentrar el tráfico en los canales virtuales de menor índice (y por tanto en las dimensiones de la topología que se comprueban en primer lugar). La segunda función, denominada *Cyclic*, elige los canales asignando prioridades repartidas de forma cíclica entre las dimensiones, consiguiendo con ello una mejor distribución del tráfico en la red. Esta mejor distribución del tráfico en la red es la que, en términos de prestaciones, penaliza a la función de selección First-Free frente a Cyclic. Como se observa en la figura, para baja carga, la latencia con la función FirstFree es significativamente más elevada porque todo el tráfico tiende a concentrarse en unos pocos canales. Solamente cuando el tráfico aumenta, las diferencias tienden a disminuir ligeramente porque se van ocupando todos los canales disponibles en ambos casos. Sin embargo la red alcanza antes la saturación con la función FirstFree (esto se observa más claramente en la topología 3D). Ello se debe a que cuando el tráfico es elevado la sobrecarga que sufren los primeros canales provoca un estado de sobreocupación de los recursos de red que no ya no se puede resolver aunque se ocupen los canales libres, que están siendo usados con menos frecuencia.

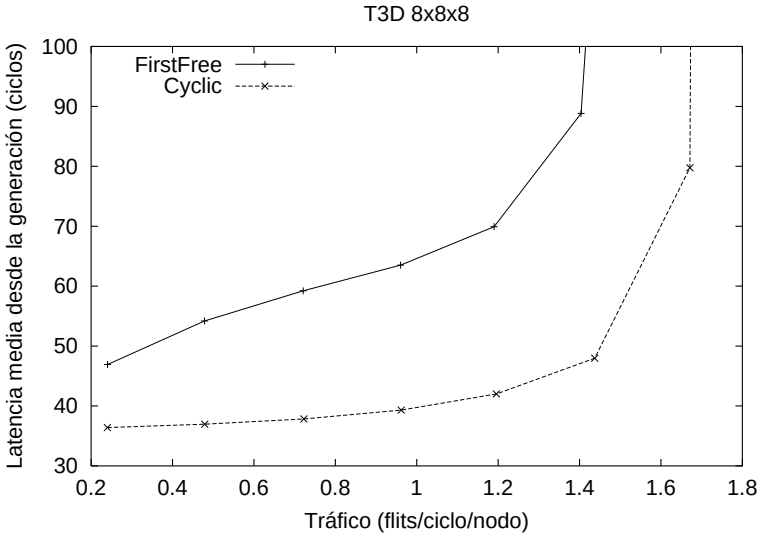
En las siguientes secciones, donde se analiza el efecto de la agregación de enlaces y el efecto de la longitud de los mensajes. Sólo se muestran resultados basados en la función Cyclic, que es la que proporciona mejores prestaciones.

4.3.4.2. Efecto de la agregación de enlaces

Con el objeto de proporcionar un punto de referencia para el posterior análisis de resultados, se ha realizado una evaluación del impacto de la agregación de enlaces sobre las prestaciones de la redes evaluadas. Se han realizado experimentos en los que se mantienen fijos todos los parámetros de la red y se evalúan la agregación de enlaces de factor 2X y factor 4X, donde cada enlace agregado está constituido por 2 o 4 enlaces, respectivamente. Los resultados presentados en la figura 4.10 muestran que la agregación de enlaces produce una mejora muy significativa en la productividad de la red. Para el caso del toro 2D, la productividad pasa de 0,24 *flits/ciclo/nodo*



(a) Latencia media frente a tráfico entregado en un toro 2D para tráfico uniforme.



(b) Latencia media frente a tráfico entregado en un toro 3D para tráfico uniforme.

Figura 4.9: Efecto de la función de selección.

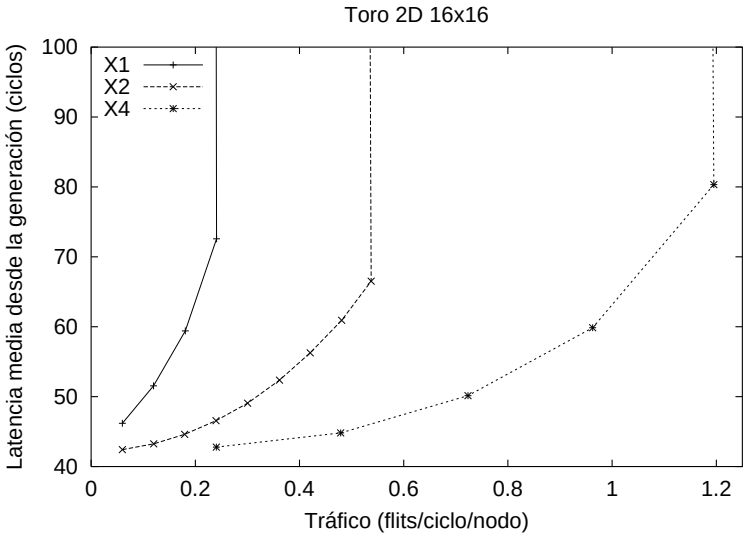
a 0,58 *flits/ciclo/nodo* y 1,2 *flits/ciclo/nodo*, con agregaciones de factor 2 y 4 respectivamente. Es importante señalar que la disponibilidad de múltiples enlaces entre cada par de nodos hace posible alcanzar productividades superiores a 1 flit/ciclo/switch, recuérdese que hay 4 canales de inyección/eyección en la red. Para baja carga no hay diferencias en la latencia de los mensajes puesto que en esas situaciones la disponibilidad de múltiples enlaces no supone una ventaja: existen enlaces disponibles pero no están en uso. Como se muestra en la figura 4.10b, los resultados para un toro 3D son equivalentes pero con unos valores de productividad mayores. La disponibilidad de una dimensión adicional, unida al empleo de un algoritmo de encaminamiento adaptativo, provoca que aumente la cantidad de tráfico que la red es capaz de entregar.

4.3.4.3. Efecto de la longitud de los mensajes

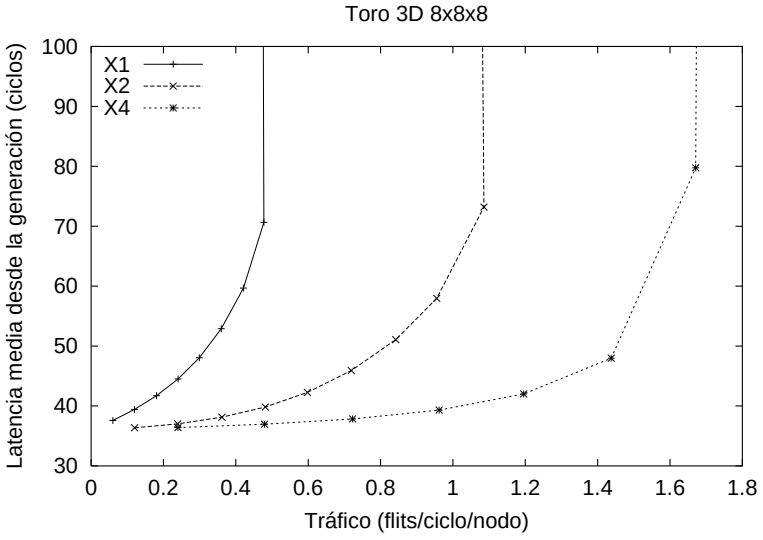
Se han realizado experimentos con mensajes cortos, de 16 flits, y mensajes largos, de 256 flits, para las dos configuraciones de red planteadas. La figura 4.11 recoge los resultados obtenidos. La latencia media por flit es inferior para los mensajes largos debido a que la transmisión de los mensajes se realiza de forma segmentada (wormhole), de manera que el establecimiento de la ruta se amortiza entre más flits cuando los mensajes son largos. Además, los flits de datos avanzan más rápido que los de cabecera porque estos últimos han de ser encaminados, esperando que la unidad de encaminamiento de cada nodo calcule el canal de salida, y posiblemente esperando a que dicho canal esté libre. Así, cuando la cabecera alcanza el destino, los flits de datos avanzan más rápido, favoreciendo a los mensajes más largos. En cuanto a la productividad, ésta es ligeramente superior para mensajes cortos, ya que en este caso los enlaces se matienen ocupados durante menos tiempo, contribuyendo así a que el tráfico fluya mejor. [28].

4.3.5. Evaluación estática del mecanismo de reducción del consumo de potencia

En esta sección se analizan las prestaciones de la red con el mecanismo de reducción del consumo de potencia en funcionamiento. Se realiza una evaluación estática del mismo, es decir, en condiciones de tráfico constantes. Los resultados presentados han sido obtenidos para configuraciones en las que todos los enlaces se encuentran conectados en el instante inicial y seguidamente se aplica un escalón de tráfico inyectado. Los experimentos han sido diseñados para que se intercambien 500.000

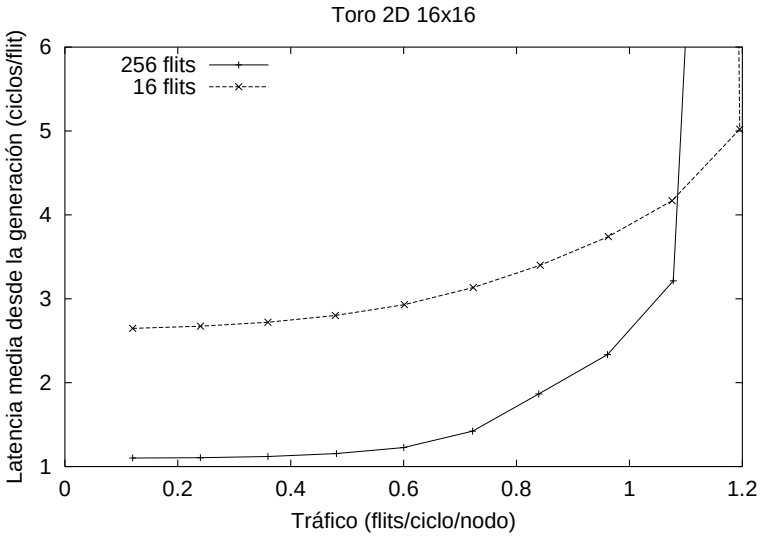


(a) Latencia media desde la generación frente a tráfico entregado en un toro 2D para tráfico uniforme y mensajes de 16 flits.

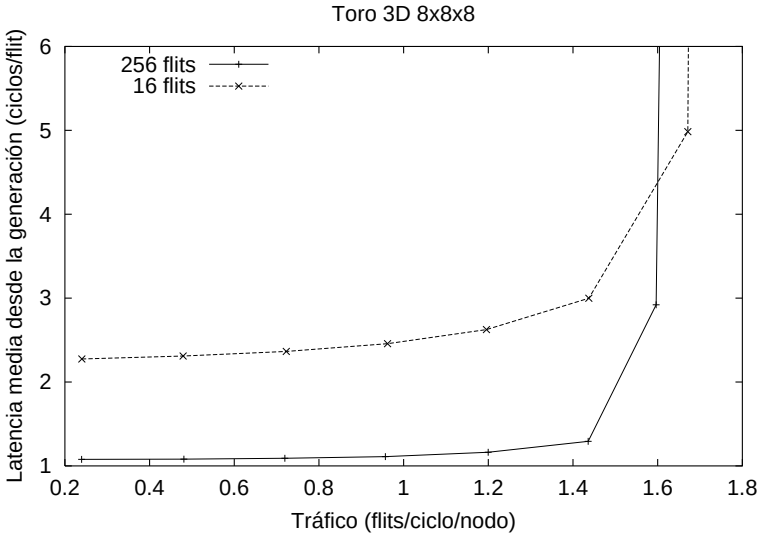


(b) Latencia media desde la generación frente a tráfico entregado en un toro 3D para tráfico uniforme y mensajes de 16 flits.

Figura 4.10: Efecto de la agregación de enlaces.



(a) Latencia media por flit frente a tráfico entregado en un toro 2D para tráfico uniforme.



(b) Latencia media por flit frente a tráfico entregado en un toro 3D para tráfico uniforme.

Figura 4.11: Efecto de la longitud de los mensajes.

mensajes para una tasa de inyección media constante. Dicha tasa de inyección se ha modificado para explorar situaciones desde baja carga hasta la saturación. Como se ha descrito, el comportamiento del mecanismo de ahorro de potencia depende de los umbrales de conexión y desconexión de enlaces, U_{on} y U_{off} respectivamente. Es por ello que se han explorado diferentes configuraciones compatibles con el mapa de umbrales posibles definido en la sección 4.2.2.

El cálculo de los umbrales objeto de esta evaluación se ha realizado de acuerdo con el mapa de posibles umbrales. Se ha buscado una distribución regular de puntos que proporcione un abanico representativo de configuraciones. De esta forma, se mostrarán diferentes ajustes de sensibilidad y agresividad del mecanismo que tendrán impacto en el comportamiento del mismo. Los puntos seleccionados aparecen resaltados en la figura 4.12. Tal como se muestra en la figura, la distribución de puntos, y por tanto los valores concretos de los umbrales, dependen del parámetro U_{MAX} o máxima utilización de la red. De acuerdo con los resultados presentados en la figura 4.10, el valor de utilización máxima alcanzada es de $U_{MAX} = 0{,}30$ para el toro 2D y de $U_{MAX} = 0{,}42$ para el toro 3D. Con respecto a dichos valores máximos se han calculado los valores de los umbrales de referencia evaluados, que se muestran en la tabla 4.2. El conjunto de valores se ha calculado distribuyendo regularmente los puntos de test, referenciados respecto de su histéresis o sensibilidad ($U_{on} - U_{\text{off}}$) y su media o agresividad ($(U_{on}+U_{\text{off}})/2$). La tabla 4.2(a) representa en dos secciones, para el toro 2D, los valores correspondientes al umbral de conexión (U_{on}), en la parte superior, y al umbral de desconexión (U_{off}), en la parte inferior. Idéntico criterio se utiliza en la tabla 4.2(b) para el toro 3D. Se han sombreado las trece parejas de umbrales que cumplen con las restricciones. También se han incluido puntos fuera del área recomendada con el objeto de evaluar el comportamiento del mecanismo cuando no se cumple la restricción de ausencia de ciclos de conexión/desconexión.

En las figuras siguientes se presentan las curvas de latencia media desde la generación de los mensajes y las curva de potencia relativa consumida por los enlaces frente al tráfico entregado, tanto para el toro 2D (figura 4.13 y figura 4.14) como para el toro 3D (figura 4.15 y figura 4.16). Las curvas están identificadas por etiquetas con el formato "$U_{\text{off}} - U_{on}$", excepto la curva etiquetada "*Nominal*", que representa la latencia de los mensajes cuando el mecanismo de ahorro de potencia no actúa. Con el fin de presentar los resultados de una manera ordenada, éstos se han agrupado para que cada gráfica muestre curvas correspondientes a configuraciones con sensibilidad (histéresis) constante. Es decir, se agrupan en la misma gráfica resultados para una fila de la matriz de puntos de test representada sobre la figura 4.12, que a su vez se

Histéresis	u_{on}	Media					
		0,0375	0,075	0,1125	0,15	0,1875	0,225
	0,05	**0,0625**	**0,1**	0,1375	0,175	0,2125	0,25
	0,1		**0,125**	**0,1625**	**0,2**	0,2375	0,275
	0,15			**0,1875**	**0,225**	**0,2625**	**0,3**
	0,2			**0,2125**	**0,25**	**0,2875**	
	0,25				**0,275**		

Histéresis	u_{off}	Media					
		0,0375	0,075	0,1125	0,15	0,1875	0,225
	0,05	**0,0125**	**0,05**	0,0875	0,125	0,1625	0,2
	0,1		**0,025**	**0,0625**	**0,1**	0,1375	0,175
	0,15			**0,0375**	**0,075**	**0,1125**	**0,15**
	0,2			**0,0125**	**0,05**	**0,0875**	
	0,25				**0,025**		

(a) Umbrales para el toro 2D con $u_{MAX} = 0{,}30$.

Histéresis	u_{on}	Media					
		0,0525	0,105	0,1575	0,21	0,2625	0,315
	0,07	**0,0875**	**0,14**	0,1925	0,245	0,2975	0,35
	0,14		**0,175**	**0,2275**	**0,28**	0,3325	0,385
	0,21			**0,2625**	**0,315**	**0,3675**	**0,42**
	0,28			**0,2975**	**0,35**	**0,4025**	
	0,35				**0,385**		

Histéresis	u_{off}	Media					
		0,0525	0,105	0,1575	0,21	0,2625	0,315
	0,07	**0,0175**	**0,07**	0,1225	0,175	0,2275	0,28
	0,14		**0,035**	**0,0875**	**0,14**	0,1925	0,245
	0,21			**0,0525**	**0,105**	**0,1575**	**0,21**
	0,28			**0,0175**	**0,07**	**0,1225**	
	0,35				**0,035**		

(b) Umbrales para el toro 3D con $u_{MAX} = 0{,}42$.

Tabla 4.2: Umbrales de test empleados en la evaluación.

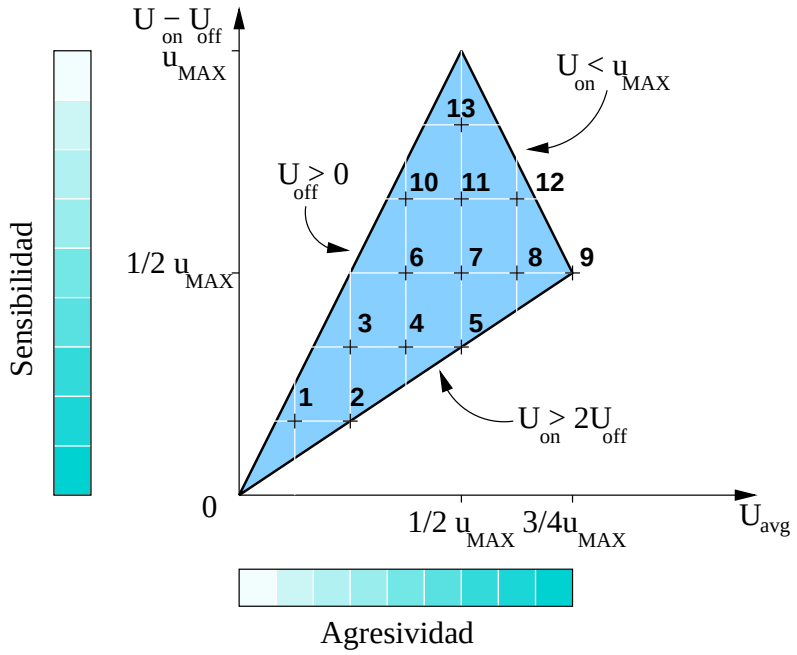


Figura 4.12: Mapa de umbrales posibles con indicación de los puntos evaluados.

corresponden a una fila de la tabla 4.2(a) (toro 2D) o de la tabla 4.2(b). Adicionalmente, las etiquetas se ordenan de arriba a abajo en orden creciente de agresividad. Los resultados muestran, tanto para el toro 2D como para el toro 3D, significativos descensos del consumo de potencia para baja carga con un incremento muy moderado en la latencia de los mensajes.

En general, para baja carga, la latencia media sufre un ligero incremento debido a que el mecanismo de ahorro de potencia ha desconectado enlaces. Es en la zona correspondiente a la baja carga donde se produce una reducción en el consumo de potencia, como se puede observar en las gráficas. Para cargas mayores, la latencia se reduce aproximándose a la nominal, al tiempo que la potencia aumenta. Ello se debe a menor desconexión de enlaces que realiza el mecanismo al estar su utilización por debajo del umbral de conexión (U_{on}).

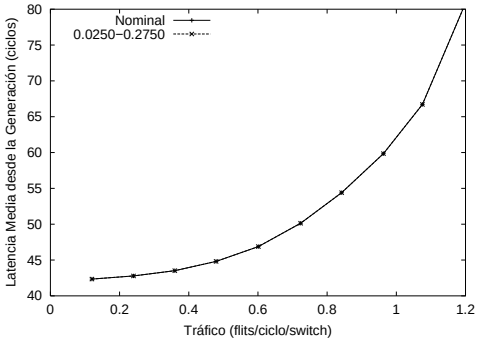
Para el toro 2D los mejores resultados se obtienen con la configuración más agresiva posible, que se corresponde con los umbrales $U_{\text{off}} = 0{,}15$ y $U_{on} = 0{,}30$ (figura 4.13(e) y figura 4.13(f)). La potencia relativa oscila entre el 30 % para baja carga y el 100 % para cargas superiores a $0{,}48\ flits/ciclo/nodo$. Es importante señalar que la cota mínima de potencia es del 25 % (lo que significa un ahorro del 75 %) que se ob-

tendría al conectar un solo enlace de los cuatro disponibles en cada enlace agregado. En este caso no se alcanza esa situación porque no se han realizado experimentos con cargas tan bajas. Para la configuración evaluada, la latencia sufre una penalización máxima del 19, 6 % en el punto de máximo ahorro de potencia situado en el 70 %. Para el toro 3D los mejores resultados se obtienen de nuevo para la configuración más agresiva, $U_{\text{off}} = 0{,}21$ y $U_{on} = 0{,}42$ (figura 4.15(e) y figura 4.15(f)). Para esta situación, la potencia oscila entre el 27 % y el 100 % para las cargas experimentadas, mientras la penalización en latencia se mantiene por debajo del 22 % en todos los casos.

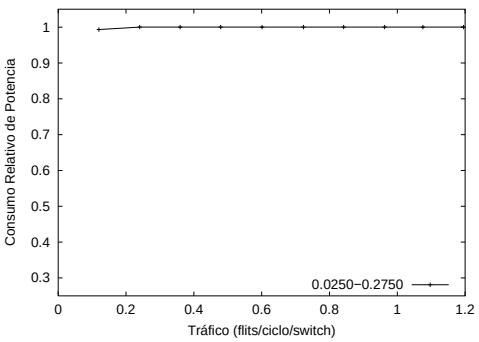
Si se comparan los resultados obtenidos para el toro 2D con los del toro 3D, se observa que el mecanismo proporciona mejores resultados con la configuración de toro 3D. El intervalo de tráfico, entre baja carga y saturación, para la cual se consigue que la red opere por debajo del 100 % de la potencia es mayor. Por ejemplo, la máxima carga para el cual el mecanismo proporciona ahorro en el toro 2D es aproximadamente de 0,5 $flits/ciclo/nodo$, o un 42 % de la carga máxima; en cambio, en el toro 3D ese valor es de 1,4 $flits/ciclo/nodo$, que representa un 83 % de la carga máxima. La mayor diversidad de rutas proporcionada por la topología 3D resulta en este caso favorable al permitir la desconexión de una fracción mayor de enlaces con un bajo impacto en la latencia.

Con el objetivo de proporcionar un punto de vista complementario que confirme la efectividad de la estrategia de ahorro de potencia, se ha evaluado el producto $L_{rel} \times P_{rel}$, donde L_{rel} es la latencia relativa con respecto a la latencia obtenida cuando no se aplica ninguna política de ahorro; y P_{rel} es el consumo relativo de potencia. Si no se utiliza ninguna estrategia de ahorro de potencia $L_{rel} \times P_{rel} = 1$, puesto que los dos factores son iguales a uno. Al aplicar el mecanismo de ahorro de potencia, la latencia relativa se ve incrementada y el consumo relativo de potencia decrece (tal y como se aprecia en las figuras previas). Si se consigue mantener este factor por debajo de la unidad, $L_{rel} \times P_{rel} \leq 1$, es una indicación de que el ahorro en el consumo de potencia tiende a compensar el aumento en la latencia media de los mensajes. Este factor, resultado del producto de tiempo por potencia, ofrece una indicación del ahorro medio de energía consumida para las condiciones testeadas. El indicador $L_{rel} \times P_{rel}$ se trata de una contribución original de esta tesis [18] que ha sido posteriormente empleado por otros equipos de investigación [53].

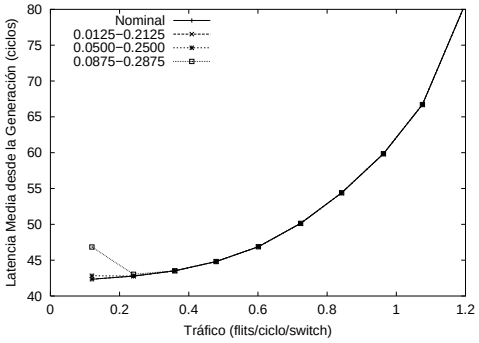
La figuras 4.17 y 4.18 muestran (para la topología 2D y 3D, respectivamente) los resultados obtenidos para el indicador $L_{rel} \times P_{rel}$ a partir de los resultados de latencia y potencia relativa presentados en las figuras 4.13 a 4.16. De los resultados mostrados, se verifica que el ahorro de potencia obtenido compensa ampliamente el incremento



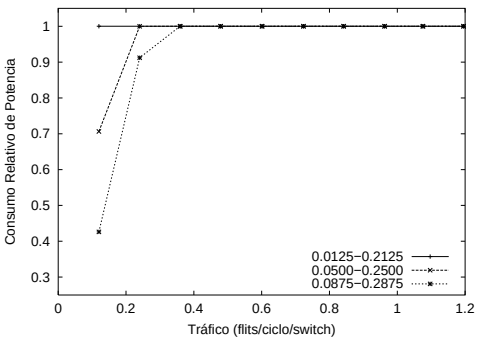
(a) Latencia con histéresis de 0,25.



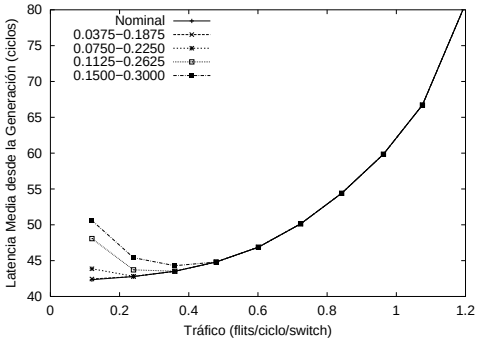
(b) Potencia con histéresis de 0,25.



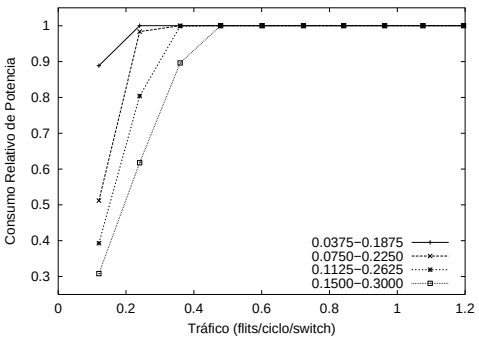
(c) Latencia con histéresis de 0,2.



(d) Potencia con histéresis de 0,2.

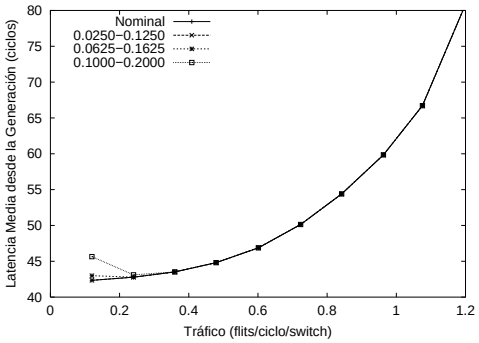


(e) Latencia con histéresis de 0,15.

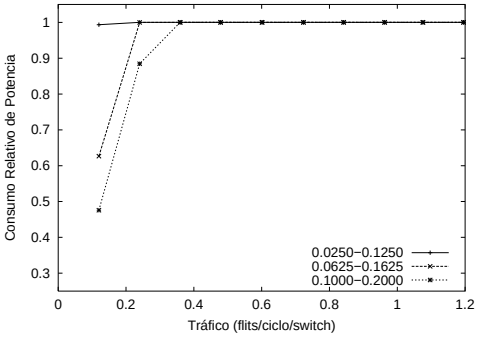


(f) Potencia con histéresis de 0,15.

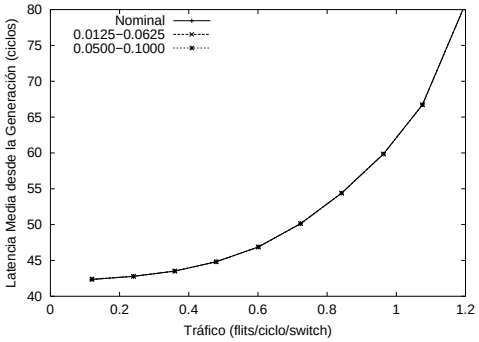
Figura 4.13: Resultados para el toro 2D (primera parte).



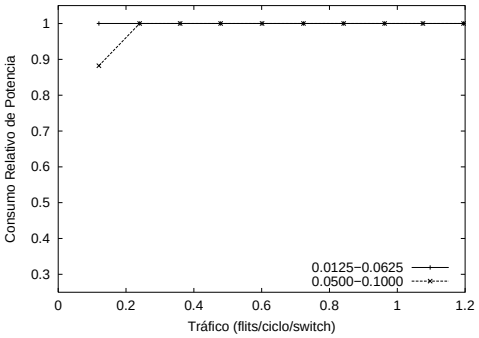
(a) Latencia con histéresis de 0, 1.



(b) Potencia con histéresis de 0, 1.



(c) Latencia con histéresis de 0, 05.



(d) Potencia con histéresis de 0, 05.

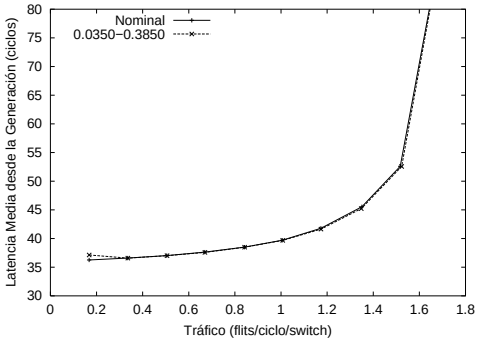
Figura 4.14: Resultados para el toro 2D (segunda parte).

de latencia para todas las configuraciones probadas. Incluso las configuraciones más agresivas, que consiguen las mayores reducciones de consumo, obtienen un valor del indicador muy favorable.

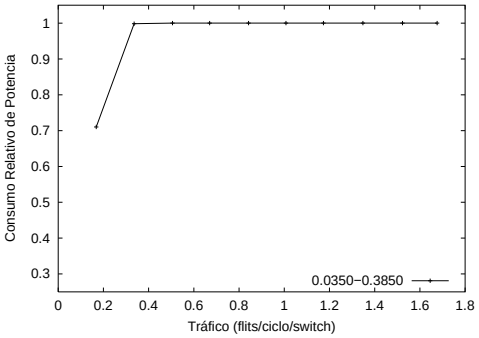
4.3.5.1. Efecto de la longitud de los mensajes

Al objeto de analizar la influencia de la longitud de los mensajes sobre la efectividad del mecanismo, se ha realizado también una evaluación del mismo con mensajes largos de 256 flits. Se han repetido los experimentos presentados en la sección anterior empleando el tamaño de mensaje más largo. Los resultados se presentan en las figuras 4.19 a 4.22.

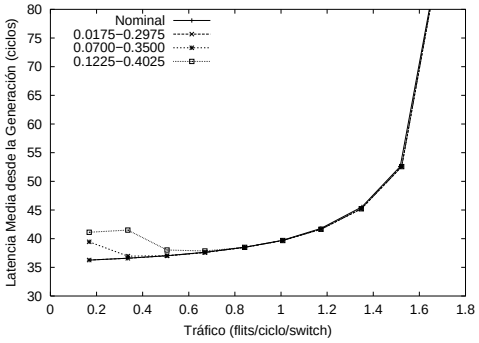
De nuevo, los mejores resultados se obtienen para el toro 2D con la configuración más agresiva, umbrales $U_{\text{off}} = 0{,}15$ y $U_{on} = 0{,}30$ (figura 4.19(e) y figura 4.19(f)). En



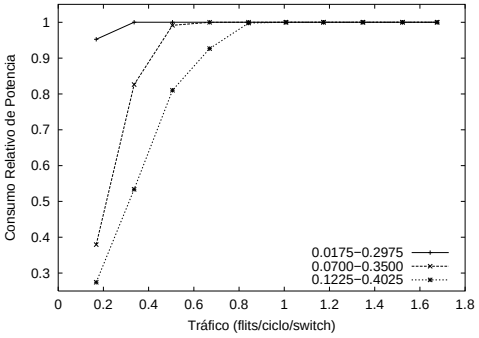
(a) Latencia con histéresis de 0,35.



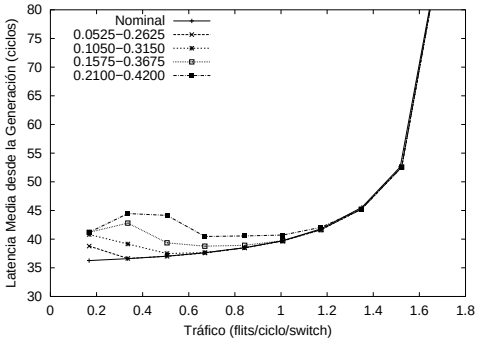
(b) Potencia con histéresis de 0,35.



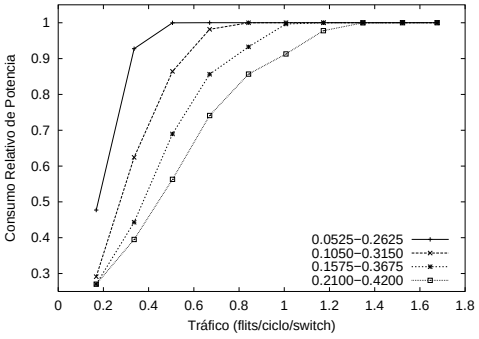
(c) Latencia con histéresis de 0,28.



(d) Potencia con histéresis de 0,28.

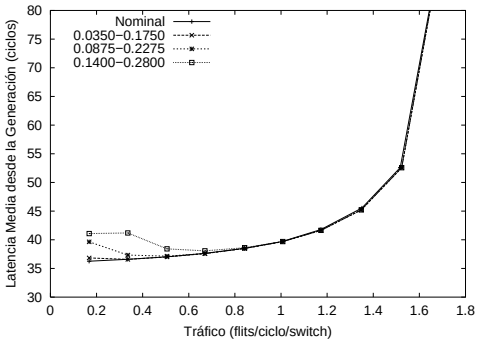


(e) Latencia con histéresis de 0,21.

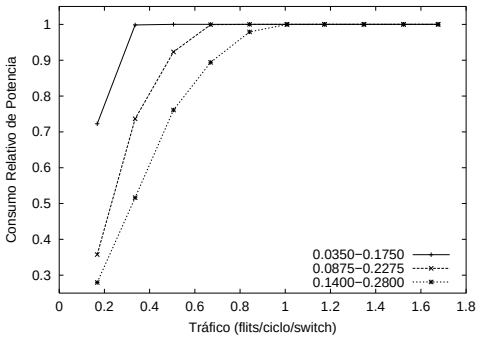


(f) Potencia con histéresis de 0,21.

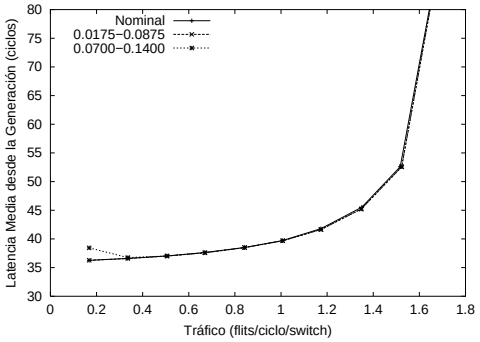
Figura 4.15: Resultados para el toro 3D (primera parte).



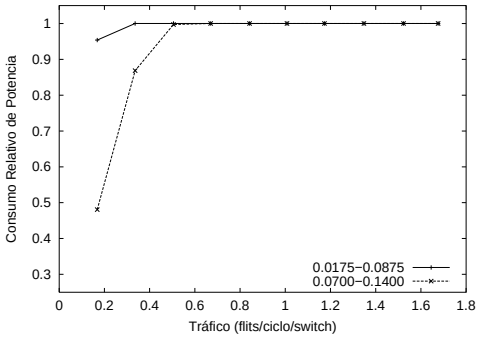
(a) Latencia con histéresis de 0, 14.



(b) Potencia con histéresis de 0, 14.



(c) Latencia con histéresis de 0, 07.

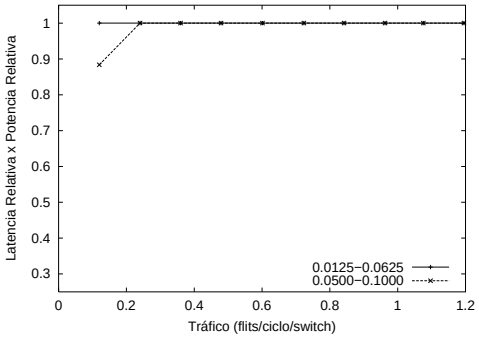


(d) Potencia con histéresis de 0, 07.

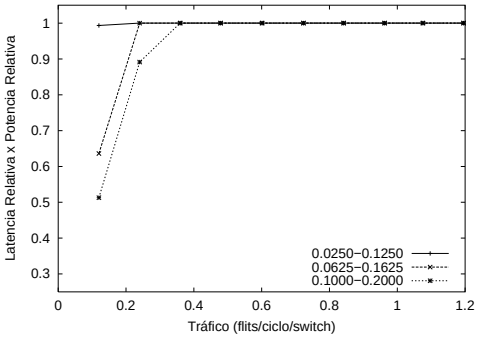
Figura 4.16: Resultados para el toro 3D (segunda parte).

los experimentos realizados, la potencia relativa oscila entre el 38 % para baja carga y el 100 % para cargas superiores a 0, 6 $flits/ciclo/switch$. Para esta configuración, la latencia sufre una penalización máxima del 51 % en el punto de máximo ahorro de potencia. Para el toro 3D, los mejores resultados también se obtienen con la configuración más agresiva, $U_{off} = 0{,}21$ y $U_{on} = 0{,}42$ (figura 4.21(e) y figura 4.21(f)). Para esta situación, la potencia oscila entre el 25 % y el 100 % para las cargas experimentadas, mientras la penalización en latencia se mantiene por debajo del 62 % en todos los casos.

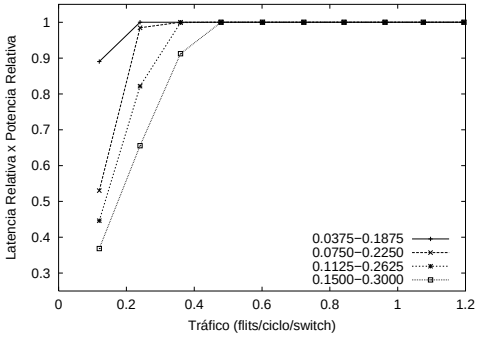
Se puede observar que el mecanismo de ahorro de potencia presenta resultados ligeramente más favorables para mensajes más largos. En el mejor caso del toro 2D, con mensajes de 16 flits se ahorra potencia para cargas por debajo del 42 % de la carga de saturación, mientras que con 256 flits este valor es del 50 %. En el caso del toro 3D estas cifras son del 83 % y del 95 %, respectivamente. El rango de tráfico para el cual



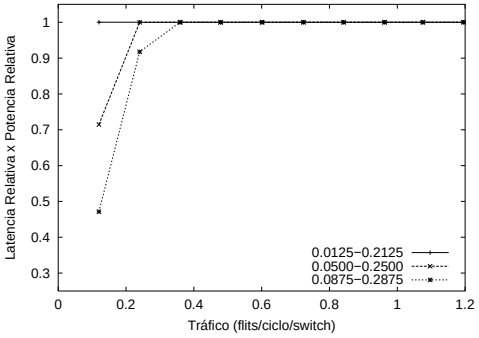
(a) $L_{rel} \times P_{rel}$ con histéresis de 0,25.



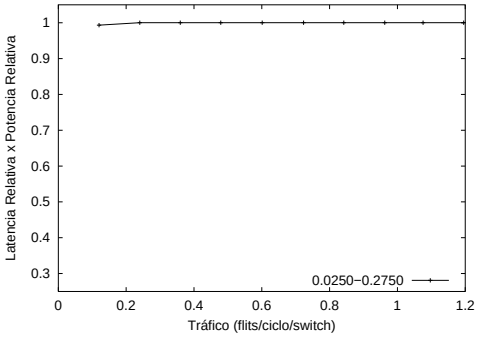
(b) $L_{rel} \times P_{rel}$ con histéresis de 0,2.



(c) $L_{rel} \times P_{rel}$ con histéresis de 0,15.

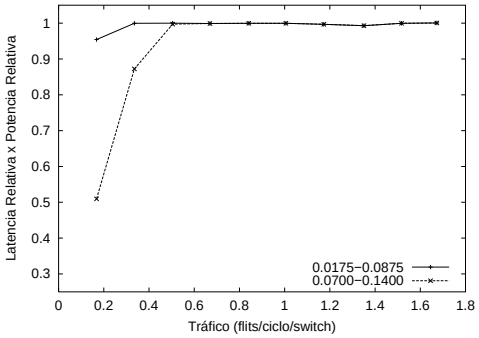


(d) $L_{rel} \times P_{rel}$ con histéresis de 0,1.

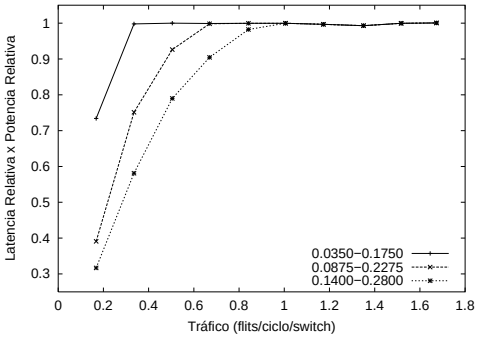


(e) $L_{rel} \times P_{rel}$ con histéresis de 0,05.

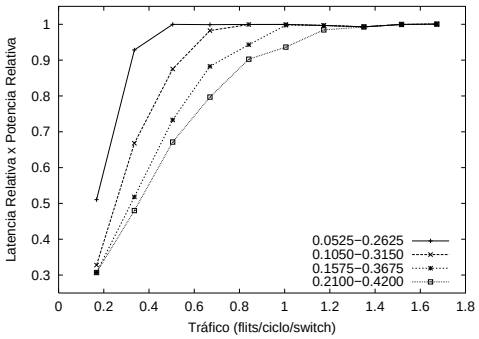
Figura 4.17: Resultados para el toro 2D.



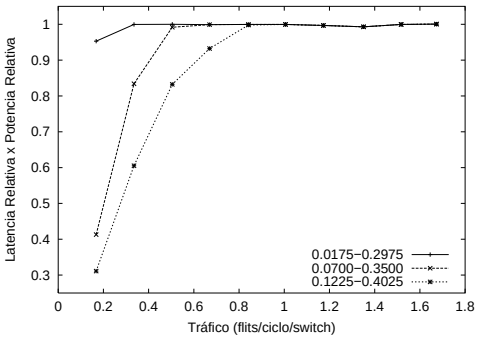
(a) $L_{rel} \times P_{rel}$ con histéresis de 0,35.



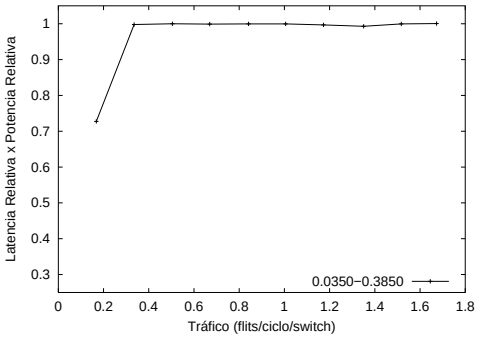
(b) $L_{rel} \times P_{rel}$ con histéresis de 0,28.



(c) $L_{rel} \times P_{rel}$ con histéresis de 0,21.



(d) $L_{rel} \times P_{rel}$ con histéresis de 0,14.



(e) $L_{rel} \times P_{rel}$ con histéresis de 0,07.

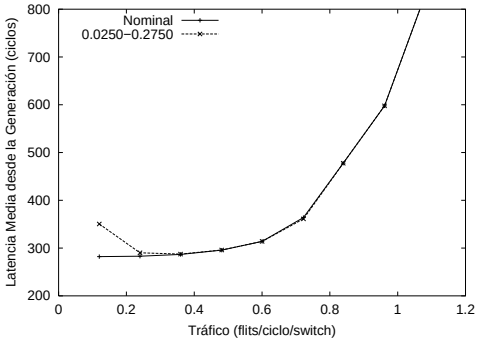
Figura 4.18: Resultados para el toro 3D.

se consigue reducir el consumo de potencia aumenta con mensajes más largos, aún cuando se emplean los mismo valores para los umbrales de conexión y desconexión de enlaces. Este efecto se debe a que la distribución del trafico en mensajes de mayor tamaño reduce la sobrecarga debida a las cabeceras y permite reducir la utilización efectiva de los enlaces y por tanto conseguir una ahorro de potencia adicional.

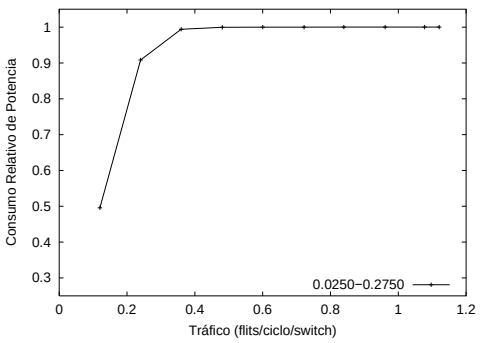
Los resultados obtenidos para la métrica $L_{rel} \times P_{rel}$ (figuras 4.23 y 4.24) confirman que la mayor penalización observada en la latencia es ampliamente compensada por la reducción del consumo de potencia, al igual que sucede con mensajes de 16 flits. Solamente una de las configuraciones del mecanismo sobre el toro 3D, la más agresiva, proporciona valores del producto $L_{rel} \times P_{rel}$ superiores a la unidad (el valor máximo medido es de 1,11). Esta configuración presenta un valor de $U_{on} = 0{,}42$ que es la máxima utilización posible de la red y $U_{\text{off}} = 0{,}21$, con lo que $U_{\text{on}} = 2 \times U_{off}$, por lo que pueden aparecer ciclos de conexión-desconexión indeseables. Es decir, está en el límite de lo que hemos definido como aceptable. Cualquiera de las demás configuraciones probadas proporciona resultados favorables para este indicador.

4.3.5.2. Efecto de la función de selección

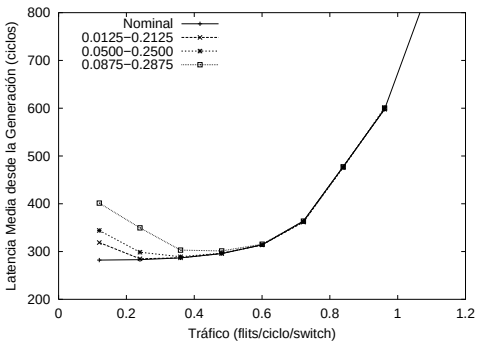
Como se ha constatado en la sección 4.3.4.1, la función de selección que proporciona las mejores oportunidades para ahorrar potencia, FirstFree, es la que ofrece peores prestaciones de la red. Todos los resultados mostrados hasta el momento, se han obtenido con la función de selección Cyclic. No obstante, se ha realizado un análisis experimental del mecanismo de ahorro de potencia en combinación con esta función de selección. El objetivo ha sido constatar si el ahorro de potencia que se consigue compensa los peores resultados de latencia de partida. Por esta razón, en los resultados de $L_{rel} \times P_{rel}$ la latencia relativa se ha calculado tomando como referencia los resultados obtenidos con la función de selección Cyclic cuando el mecanismo de ahorro de potencia no está operativo. Se han evaluado todas las configuraciones de umbrales presentadas en los experimentos anteriores y se resumen los resultados por medio de las representación gráfica del indicador $L_{rel} \times P_{rel}$ en función del tráfico. Los resultados, recogidos en las figuras 4.25 y 4.26, indican que solamente en algunas configuraciones de umbrales y con muy baja carga, inferior a 0,2 $flits/ciclo/switch$, se obtienen valores favorables del indicador. A la vista de estos resultados, concluimos que la función de selección Cyclic, que es la que proporciona las mejores prestaciones para la topología de toro, con encaminamiento totalmente adaptativo, es también la más adecuada cuando se aplica nuestra propuesta de reducción del consumo de potencia. Idénticos resultados se han obtenido al evaluar la función de selección



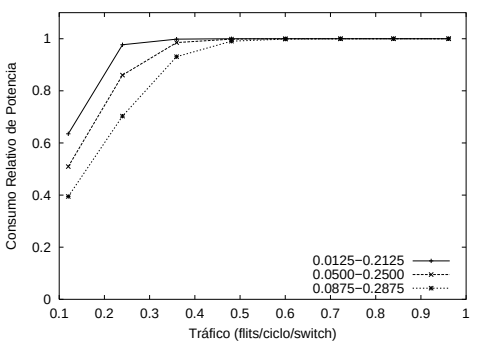
(a) Latencia con histéresis de 0,25.



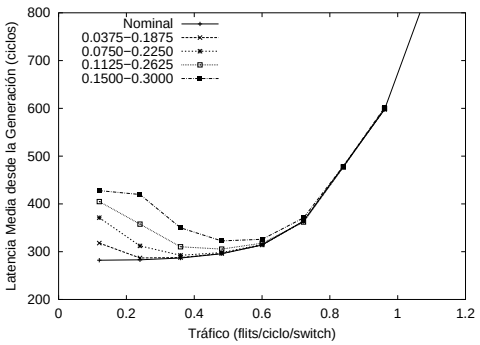
(b) Potencia con histéresis de 0,25.



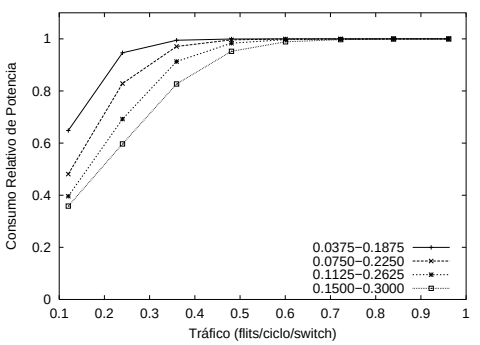
(c) Latencia con histéresis de 0,2.



(d) Potencia con histéresis de 0,2.

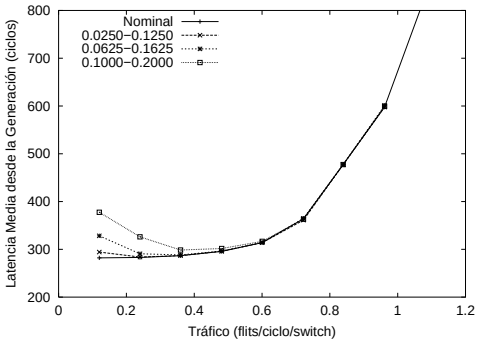


(e) Latencia con histéresis de 0,15.

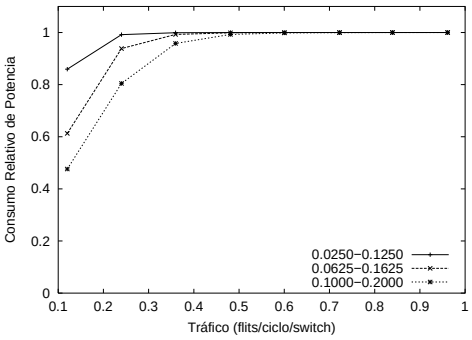


(f) Potencia con histéresis de 0,15.

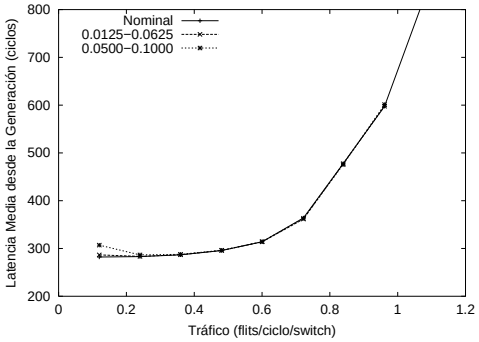
Figura 4.19: Latencia media y consumo relativo para el toro 2D con mensajes de 256 flits (primera parte).



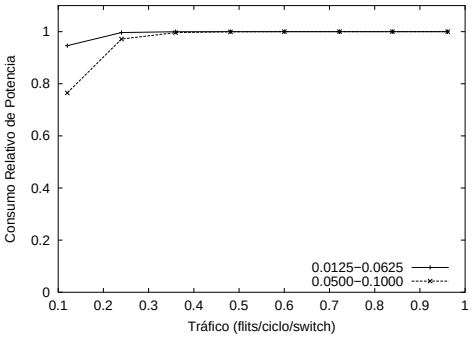
(a) Latencia con histéresis de 0, 1.



(b) Potencia con histéresis de 0, 1.



(c) Latencia con histéresis de 0, 05.



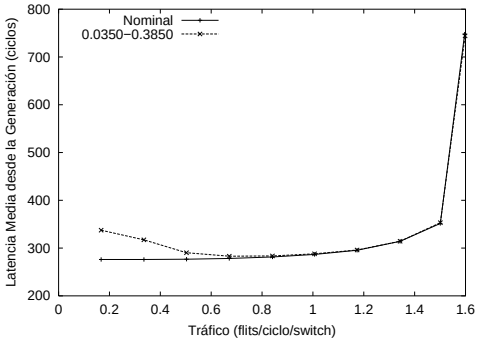
(d) Potencia con histéresis de 0, 05.

Figura 4.20: Latencia media y consumo relativo para el toro 2D con mensajes de 256 flits (segunda parte).

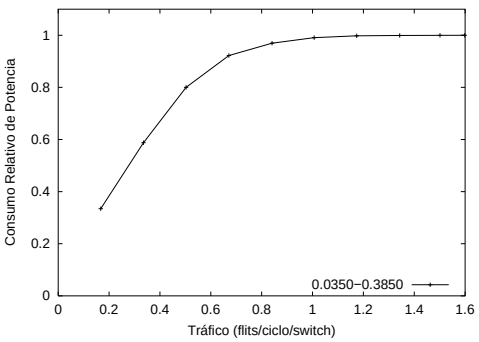
sobre mensajes de 256 flits.

4.3.5.3. Energía consumida

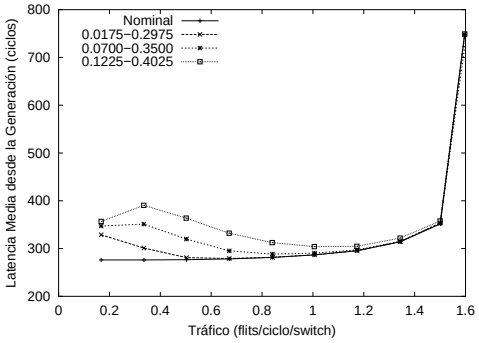
Los resultados anteriores se completan con medidas de energía que proporcionan evidencias adicionales sobre los beneficios potenciales de la estrategia propuesta. El consumo de energía en la red se ha evaluado usando un modelo de carga cerrado. En este modelo, la simulación está caracterizada por el intercambio de un número fijo de mensajes a través de la red. Para cada simulación los nodos activos inyectan otros tantos mensajes a destinos aleatorios. Por otro lado, cada vez que se recibe un mensaje, se responde con otro. De esta manera la latencia de los mensajes tiene influencia sobre el tiempo de generación de los mismos, a diferencia del modelo



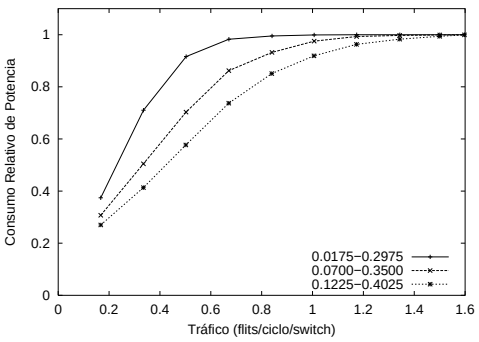
(a) Latencia con histéresis de 0,35.



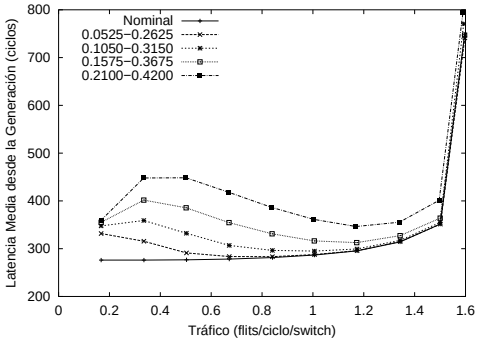
(b) Potencia con histéresis de 0,35.



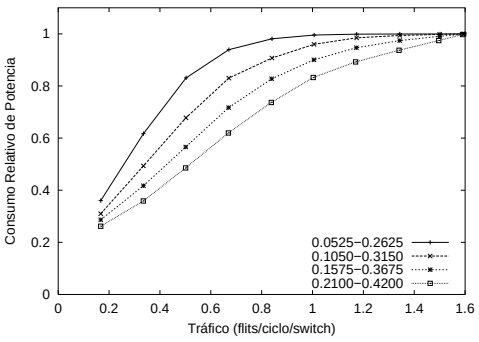
(c) Latencia con histéresis de 0,28.



(d) Potencia con histéresis de 0,28.

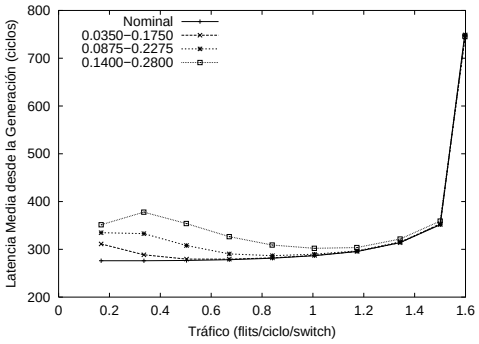


(e) Latencia con histéresis de 0,21.

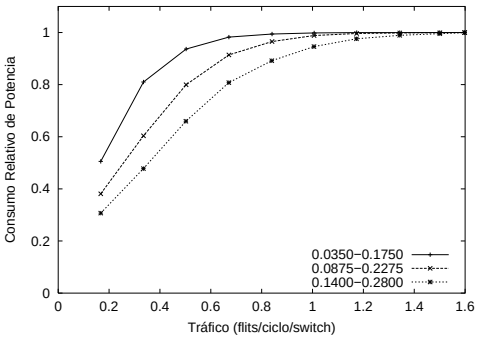


(f) Potencia con histéresis de 0,21.

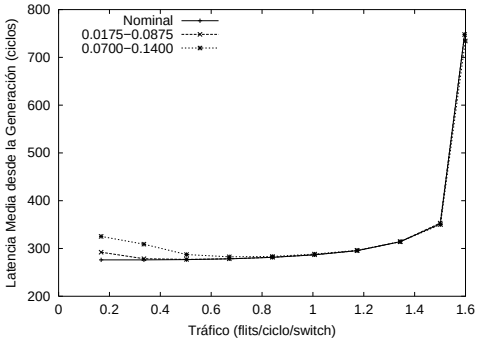
Figura 4.21: Latencia media y consumo relativo para el toro 3D con mensajes de 256 flits (primera parte).



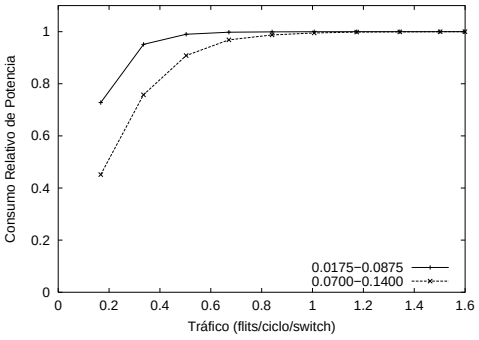
(a) Latencia con histéresis de 0, 14.



(b) Potencia con histéresis de 0, 14.



(c) Latencia con histéresis de 0, 07.



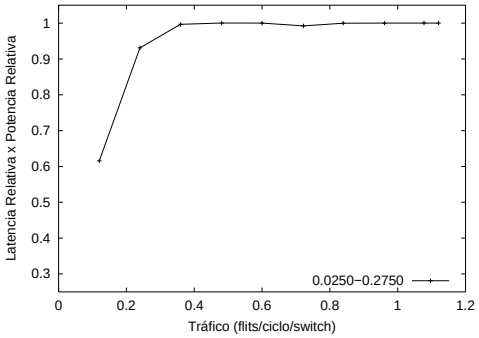
(d) Potencia con histéresis de 0, 07.

Figura 4.22: Latencia media y consumo relativo para el toro 3D con mensajes de 256 flits (segunda parte).

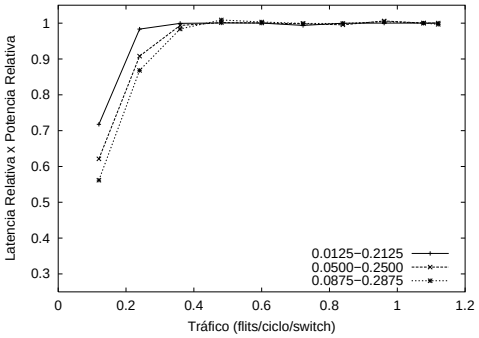
abierto de carga empleado en los experimentos anteriores. Es el tipo de tráfico que se denomina de petición-respuesta utilizado por otros autores [26, 2].

Durante cada punto de simulación los nodos inyectan mensajes empleando la máxima carga para la topología analizada. La carga global de la red se ajusta haciendo participar solo a una parte de los nodos. Con ello, para cada punto de simulación, hay un porcentaje de nodos de la red inyectando mensajes a destinos aleatorios, este porcentaje indica que intervienen el 100 %, 90 %,..,10 % de los nodos respectivamente. Por ejemplo, para simular una carga de red del 10 % un 10 % de los nodos de la red inyectan tráfico a su máxima carga. Se procede de manera similar para escalones de carga espaciados al 10 % para obtener un barrido significativo hasta el 100 %.

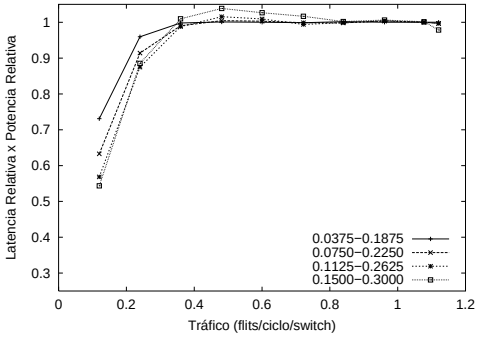
Las simulaciones están realizadas teniendo en cuenta que para la configuración utilizada, toro 3-D $8 \times 8 \times 8$, la carga máxima es de 0,45. La simulación se prolon-



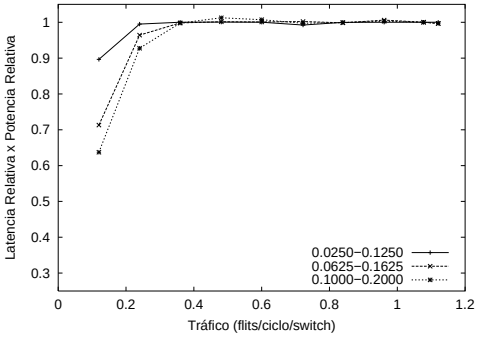
(a) $L_{rel} \times P_{rel}$ con histéresis de 0,25.



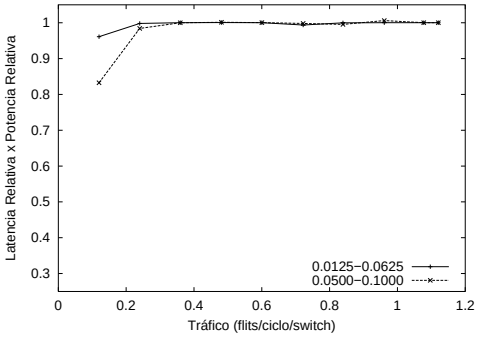
(b) $L_{rel} \times P_{rel}$ con histéresis de 0,2.



(c) $L_{rel} \times P_{rel}$ con histéresis de 0,15.

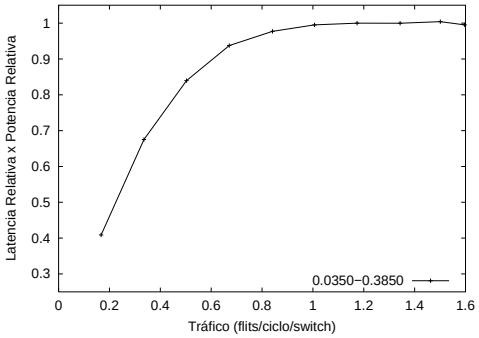


(d) $L_{rel} \times P_{rel}$ con histéresis de 0,1.

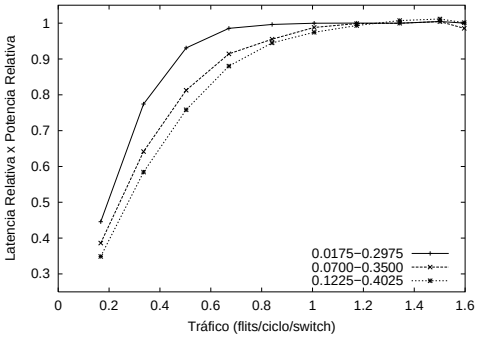


(e) $L_{rel} \times P_{rel}$ con histéresis de 0,05.

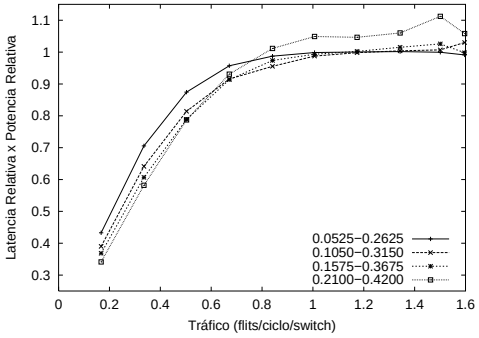
Figura 4.23: Resultados para el toro 2D con mensajes de 256 flits.



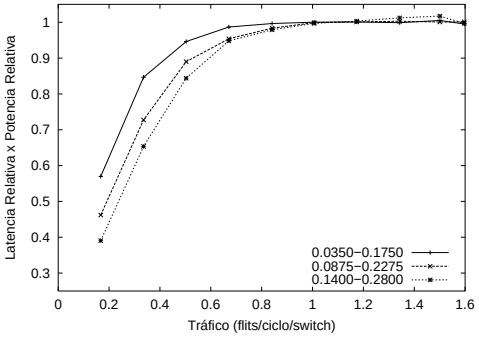
(a) $L_{rel} \times P_{rel}$ con histéresis de 0,35.



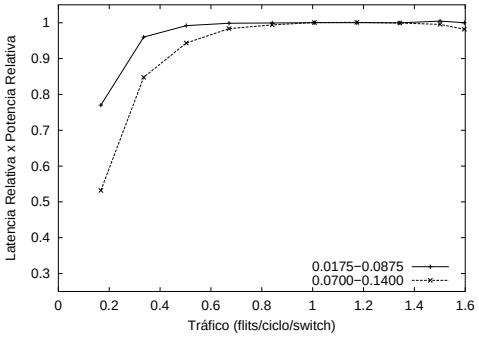
(b) $L_{rel} \times P_{rel}$ con histéresis de 0,28.



(c) $L_{rel} \times P_{rel}$ con histéresis de 0,21.

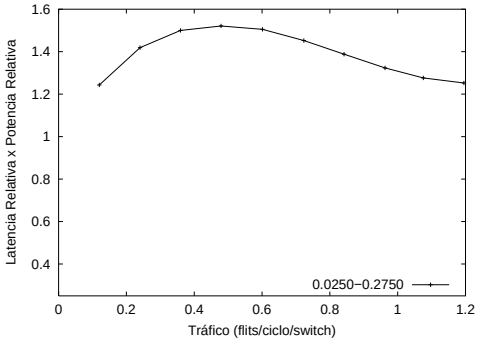


(d) $L_{rel} \times P_{rel}$ con histéresis de 0,14.

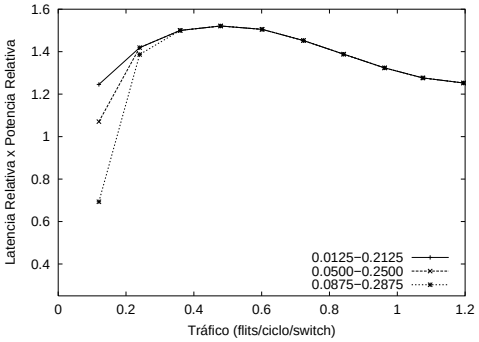


(e) $L_{rel} \times P_{rel}$ con histéresis de 0,07.

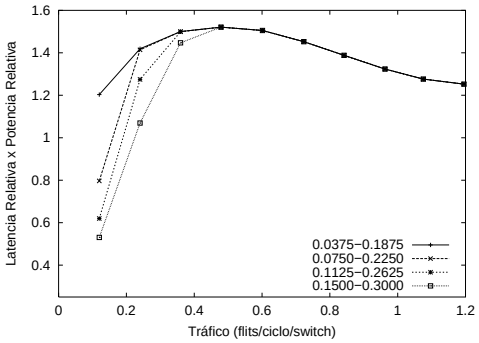
Figura 4.24: Resultados para el toro 3D con mensajes de 256 flits.



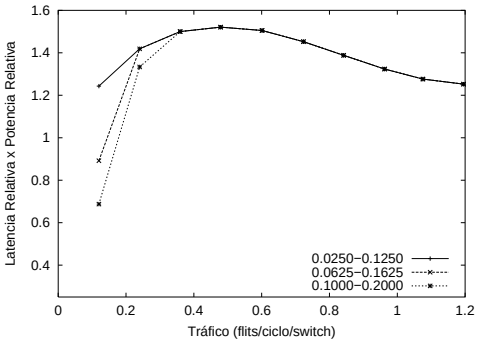
(a) $L_{rel} \times P_{rel}$ con histéresis de 0,25.



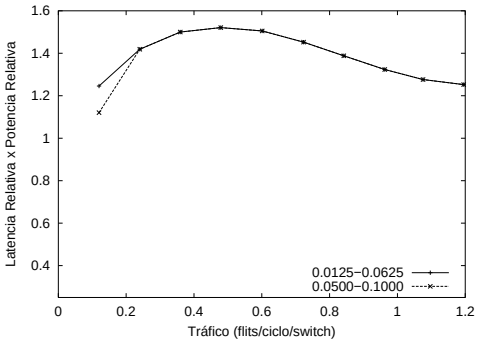
(b) $L_{rel} \times P_{rel}$ con histéresis de 0,2.



(c) $L_{rel} \times P_{rel}$ con histéresis de 0,15.

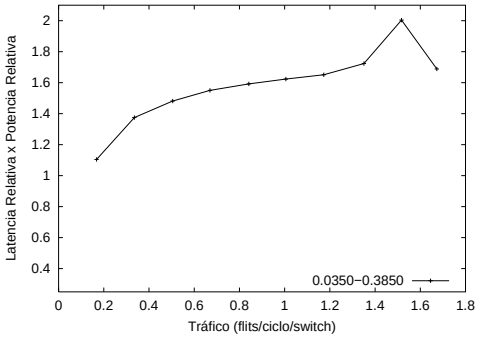


(d) $L_{rel} \times P_{rel}$ con histéresis de 0,1.

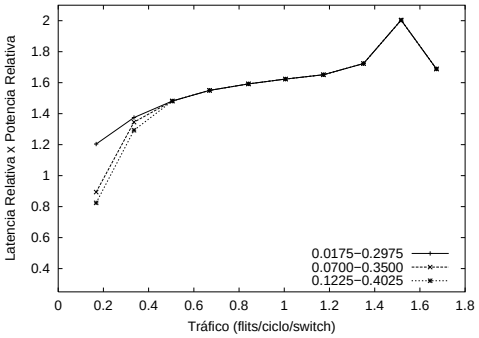


(e) $L_{rel} \times P_{rel}$ con histéresis de 0,05.

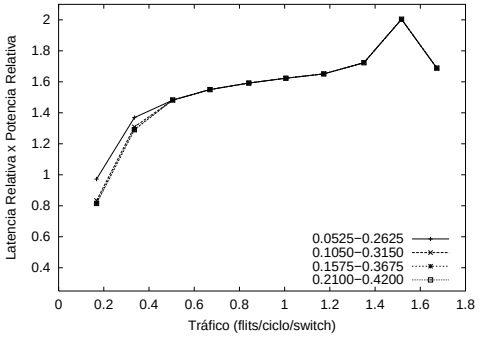
Figura 4.25: Resultados para el toro 2D con función de selección FirstFree.



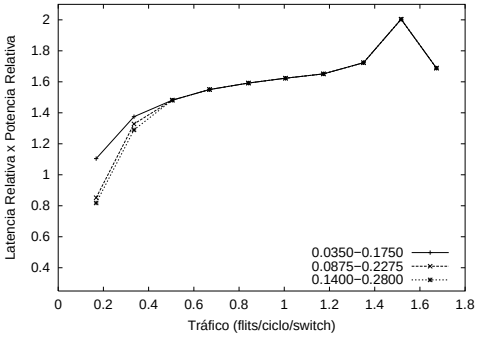
(a) $L_{rel} \times P_{rel}$ con histéresis de 0,35.



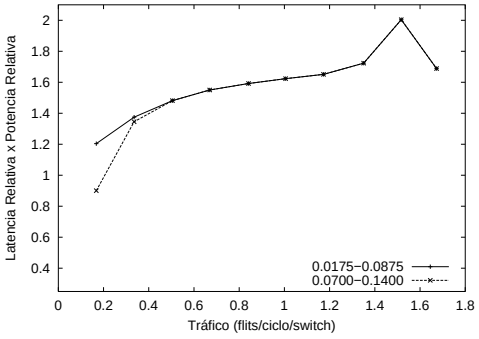
(b) $L_{rel} \times P_{rel}$ con histéresis de 0,28.



(c) $L_{rel} \times P_{rel}$ con histéresis de 0,21.



(d) $L_{rel} \times P_{rel}$ con histéresis de 0,14.



(e) $L_{rel} \times P_{rel}$ con histéresis de 0,07.

Figura 4.26: Resultados para el toro 3D con función de selección FirstFree.

Nodos activos	T Nominal (ciclos)		Power Aware		
		$T(ciclos)$	$\triangle T$	*Potencia*	$\triangle$*Energía*
10 %	*91429*	*101587*	*11.11 %*	*28.83 %*	*-54.05 %*
20 %	*46113*	*52459*	*13.76 %*	*43.98 %*	*-38.75 %*
30 %	*31115*	*32821*	*5.48 %*	*69.21 %*	*-21.28 %*
40 %	*23733*	*24316*	*2.46 %*	*85.32 %*	*-9.94 %*
50 %	*19405*	*19573*	*0.87 %*	*92.75 %*	*-5.16 %*
60 %	*16563*	*16594*	*0.19 %*	*98.96 %*	*-0.67 %*
70 %	*14682*	*14688*	*0.04 %*	*99.98 %*	*0.03 %*
80 %	*13367*	*13367*	*0 %*	*100 %*	*0 %*
90 %	*12419*	*12419*	*0 %*	*100 %*	*0 %*
100 %	*11718*	*11718*	*0 %*	*100 %*	*0 %*

Tabla 4.3: Impacto del consumo de energía para un toro 3-D.

ga hasta que 500.000 mensajes llegan a sus nodos destinos. Como se ha descrito, la carga de trabajo se controla cambiando el número de nodos activos en la red. Se ha trabajado con 10 puntos de carga equiespaciados respecto a la máxima. Para cada simulación, se mide el tiempo necesario para entregar los mensajes y la potencia consumida por los enlaces. La tabla 4.3 muestra los resultados para una configuración con un toro 3-D con cargas en la red situadas en un rango entre el 10 % al 100 % de la saturación y unos umbrales $U_{\text{off}} = 0{,}15$ y $U_{on} = 0{,}30$.

Analizando los resultados presentados se observa que cuando se aplica el mecanismo de ahorro de potencia, el tiempo de simulación (tercera columna de la tabla 4.3) crece comparado con el sistema original (segunda columna de la tabla). El incremento en el tiempo de simulación se muestra en la cuarta columna de la tabla. La desconexión de enlaces provocada por el mecanismo de ahorro de potencia hace que la potencia consumida por los mismos (quinta columna) se reduzca a una fracción del valor nominal.

Con el objetivo de realizar estimaciones del consumo de energía total de la red también se ha considerado la energía que consumen los conmutadores o switches. Se ha incluido el consumo de los conmutadores porque durante el tiempo de ejecución adicional provocado por el mecanismo de ahorro de potencia todos los componentes de la red (no sólo los enlaces) están activos y contribuyen al consumo de energía. El cálculo de la energía se ha realizado teniendo en cuenta el consumo de los enlaces y conmutadores de acuerdo con la siguiente expresión:

$$E = (P_{links} \times T_{sim}) \times a + (P_{routers} \times T_{sim}) \times (1 - a)$$

donde E es energía, P_{links} es potencia relativa consumida por los links, $P_{routers}$ es potencia relativa consumida por los routers, T_{sim} es la duración de la simulación (tiempo requerido para entregar todos los mensajes) y a es la fracción de potencia consumida por los enlaces. $P_{routers} = 1$ puesto que no se evaluan técnicas de ahorro de potencia a nivel de router.

En los experimentos realizados, cuando no se aplican técnicas de reducción del consumo de potencia $P_{links} = 1$ y T_{sim} es el nominal, por lo tanto $E = T_{sim}$. En cambio, cuando se aplican dichas técnicas, se espera, y así lo demuestran los resultados, que T_{sim}aumente pero reduciendo E. La columna situada más a la derecha de la tabla 4.3 presenta el incremento de energía respecto al nomimal que se obtiene al ejecutar las simulaciones con el mecanismo de ahorro de potencia activado. Se muestra que la ponderación en la energía es siempre favorable a la red con el mecanismo de ahorro de potencia a pesar del incremento en el tiempo de simulación .

4.3.6. Evaluación dinámica del mecanismo de reducción del consumo de potencia

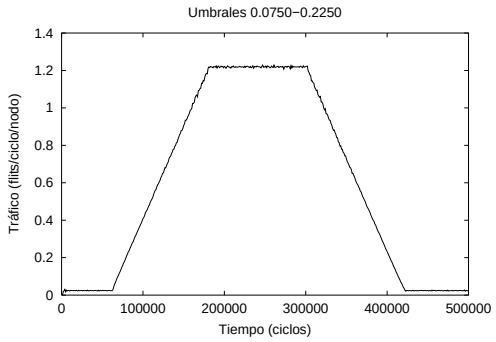
Para analizar el comportamiento dinámico del mecanismo propuesto, se han ejecutado simulaciones utilizando niveles de carga variables. Se inicia la simulación con un nivel de tráfico equivalente al 2 % del máximo soportado por la red, tras un período de tiempo, la carga crece de forma constante durante 120000 ciclos hasta un valor del 100 % de la carga de saturación (U_{MAX}). Esta carga elevada se mantiene durante 120000 ciclos. Entonces, el tráfico de entrada se decrementa nuevamente de forma constante hasta alcanzar el valor inicial. Este experimento ilustra el comportamiento del sistema bajo cargas que tengan pendientes crecientes y decrecientes. El estado de la red en el instante inicial es de desconexión de todos los enlaces posibles, de acuerdo con el mecanismo propuesto (el consumo relativo de potencia es del 25 %).

La figura 4.27 muestra los resultados obtenidos para un toro 2D, empleando mensajes de 16 flits, para una configuración de los umbrales aproximadamente en el punto central del mapa de posibles umbrales (figura 5.3), en concreto $U_{\text{off}} = 0{,}0750$ y $U_{on} = 0{,}2250$. En todos los casos, el eje de abscisas indica el tiempo transcurrido en ciclos. La figura 4.27(a) muestra el tráfico entregado, donde se aprecia la evolución en forma de doble rampa descrita. La figura 4.27(b) presenta la latencia media de los mensajes, calculada para un periodo de medida de 2000 ciclos y la figura 4.27(c) proporciona los resultados de consumo relativo de potencia. Por último, la figura 4.27(d) combina en una misma gráfica la representación de la latencia y de la potencia rela-

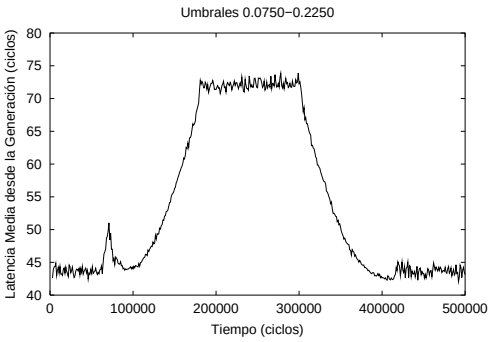
tiva. El comportamiento del mecanismo de ahorro de potencia se observa muy bien en esta última. Mientras el tráfico es bajo, el 75 % de los enlaces de la red están desconectados y la latencia se mantiene constante entorno a 44 ciclos. Cuando el tráfico empieza a incrementarse, la latencia de los mensaje tiende a subir porque sólo hay un enlace conectado. Se incrementa la utilización de los enlaces y ello provoca que se inicie el proceso de conexión de los mismos. Tras la conexión de enlaces, la latencia experimenta un descenso debido a la disponibilidad de ancho de banda adicional. Sin embargo, al mantenerse constante la tasa de incremento de tráfico inyectado, la utilización sigue aumentando y se continúan conectando enlaces adicionales. Una vez todos los enlaces han sido conectados (en este caso eso sucede alrededor de los 100000 ciclos), el aumento adicional de tráfico provoca un aumento de latencia hasta alcanzar un valor aproximado de 73 ciclos en el que se estabiliza, una vez el tráfico queda fijado en el valor máximo (en este caso 1,2 $flits/ciclo/switch$). Por otra parte, cuando el tráfico inicia la rampa de bajada, se produce una progresiva reducción de la latencia. Cuando la utilización de los enlaces es inferior a U_{off} actúa el mecanismo de desconexión de enlaces y se produce un ligero incremento transitorio de la latencia (aproximadamente en los 400000 ciclos), previo a su estabilización en un valor de 44 ciclos.

La figuras 4.28 y 4.29 muestran los resultados obtenidos para las configuraciones más significativas del mapa de posibles umbrales para el toro 2D (se ha incluido el mapa para facilitar el análisis de los resultados). Se observa que el comportamiento del mecanismo es similar al del caso anterior pero con un impacto mayor en las curvas de latencia a medida que los requerimiento de ahorro de potencia son más exigentes por el uso de configuraciones más agresivas.

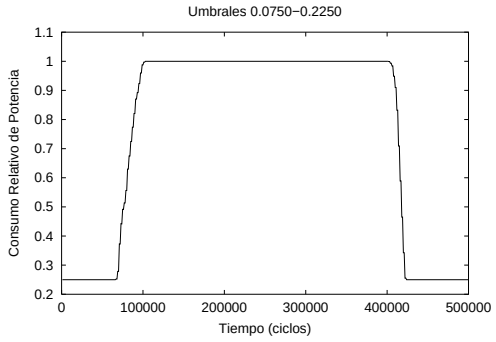
Para el caso del toro 3D se han resumido todos los resultados en las figuras 4.30 y 4.31. Se verifica el comportamiento esperado del mecanismo con un impacto más significativo en el ahorro de potencia, como se ha constatado en la evaluación estática (sección 4.3.5). Ello va ligado a un mayor incremento en la latencia de los mensajes. Se observa en algunos casos muy claramente (por ejemplo con los escalones en la figura 4.31(b)) que la red pasa por distintos estados que se corresponden a niveles de consumo de potencia que se mantienen estables durante un intervalo de tiempo. Estas pequeñas mesetas en las curvas de potencia se corresponden a periodos de crecimiento en la latencia. La latencia aumenta hasta que el incremento en el tráfico es suficiente para conectar enlaces adicionales y evolucionar hacia otro estado de la red al tiempo que la latencia vuelve a decrecer. Hasta este punto se constata que el comportamiento del mecanismo es el adecuado a los objetivos de diseño.



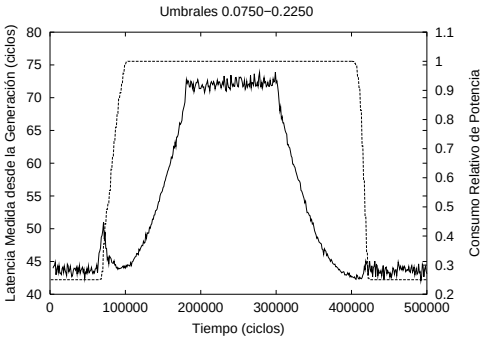
(a) Tráfico entregado en función del tiempo.



(b) Latencia en función del tiempo.



(c) Potencia consumida en función del tiempo.

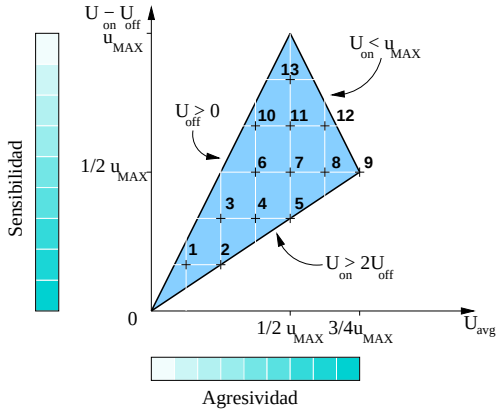


(d) Latencia y potencia superpuestas.

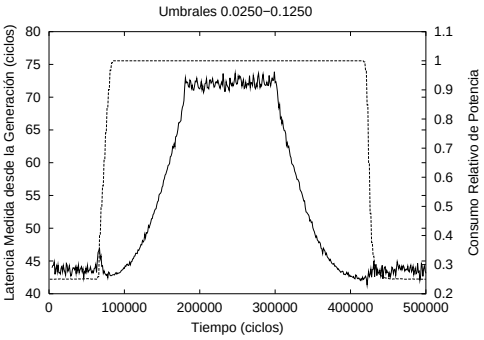
Figura 4.27: Evaluación dinámica para el toro 2D con $U_{off} = 0{,}075$ y $U_{on} = 0{,}225$.

Un análisis más detallado del comportamiento de la red puede realizarse estudiando la evolución de la latencia de los mensajes cuando el tráfico aumenta o disminuye. En la figura 4.32 se representa la latencia media frente al tráfico entregado, tanto para los tramos de carga ascendente como descendente, en las dos topologías analizadas. La línea etiquetada "Ascendente" representa la latencia cuando el tráfico aumenta desde el 2 % al 100 % de la carga máxima, mientras que la línea "Descendente" representa la situación contraria (debe ser leída, por tanto, de derecha a izquierda). Junto a estas dos curvas se incluyen también, como referencia, las curvas estáticas de latencia para redes con un nivel de agregación de 1, 2 y 4 (X1, X2, X4). Por brevedad se ha seleccionado una sola de las configuraciones probadas, en particular la correspondiente al punto 9, que es la que presenta la mayor agresividad teórica (resultados similares se han obtenido con las restantes configuraciones).

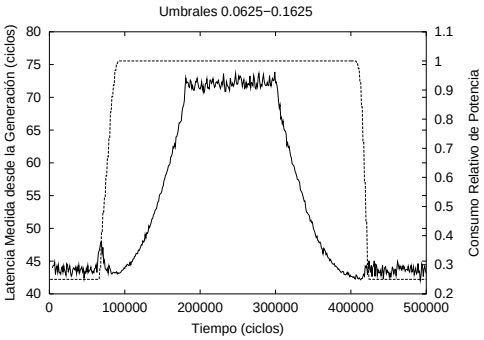
Es importante señalar que las curvas estáticas superpuestas en la figura son resul-



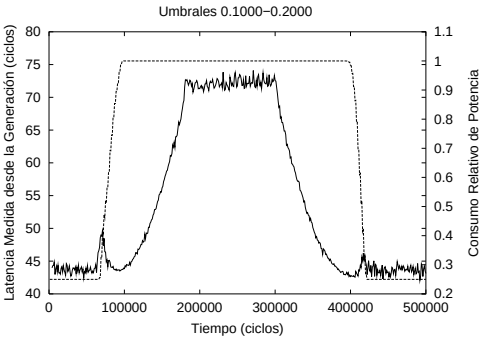
(a) Mapa de posibles umbrales.



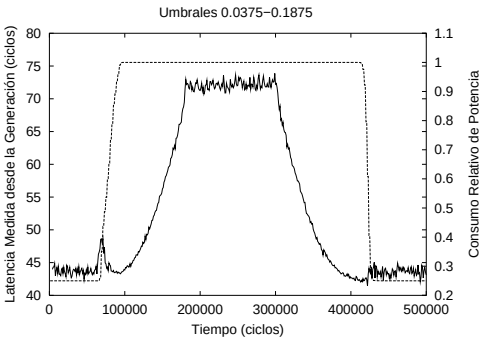
(b) Latencia y potencia en la configuración 3.



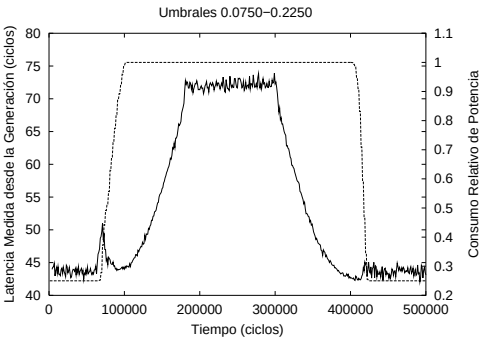
(c) Latencia y potencia en la configuración 4.



(d) Latencia y potencia en la configuración 5.

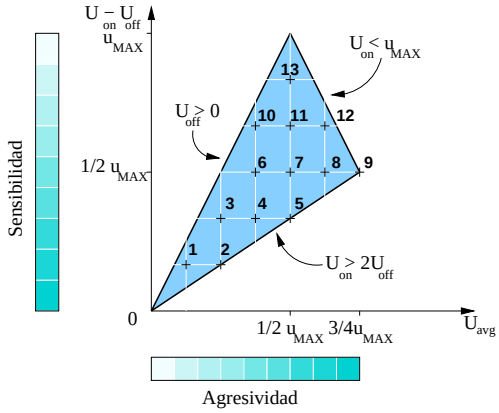


(e) Latencia y potencia en la configuración 6.

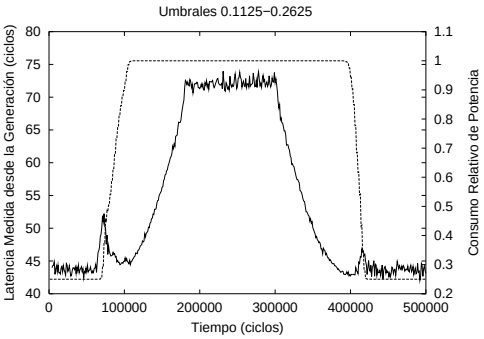


(f) Latencia y potencia en la configuración 7.

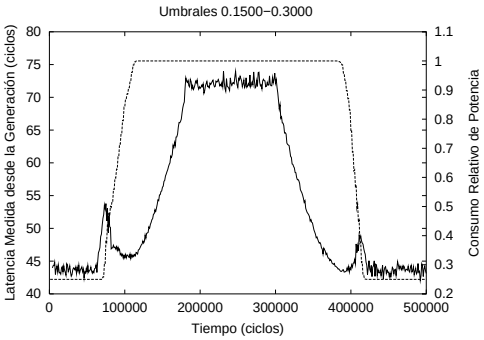
Figura 4.28: Evaluación dinámica para el toro 2D para distintos puntos del mapa de posibles umbrales.



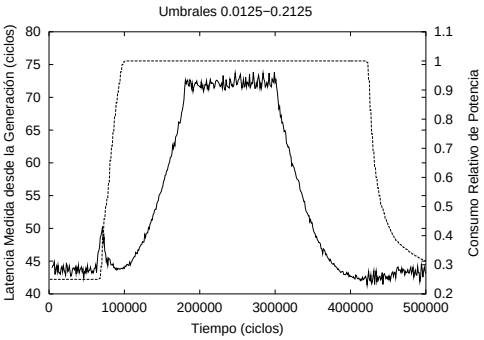
(a) Mapa de posibles umbrales.



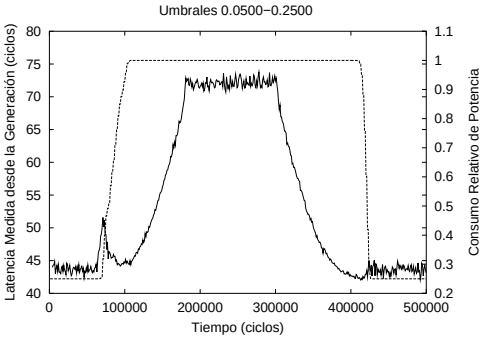
(b) Latencia y potencia en la configuración 8.



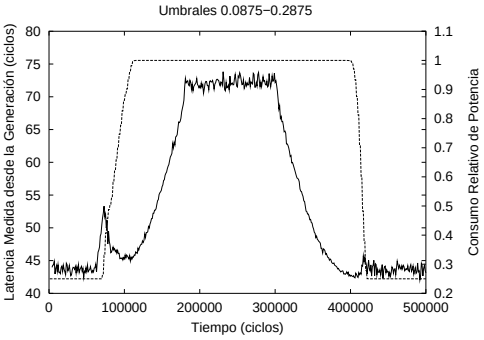
(c) Latencia y potencia en la configuración 9.



(d) Latencia y potencia en la configuración 10.

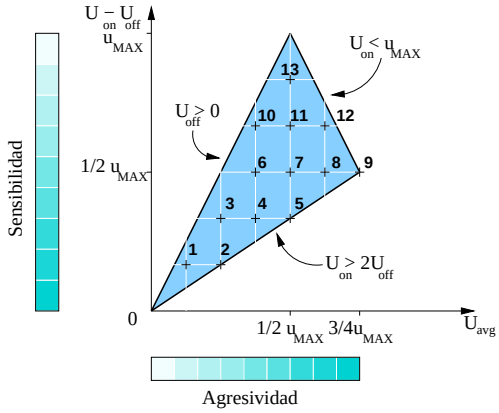


(e) Latencia y potencia en la configuración 11.

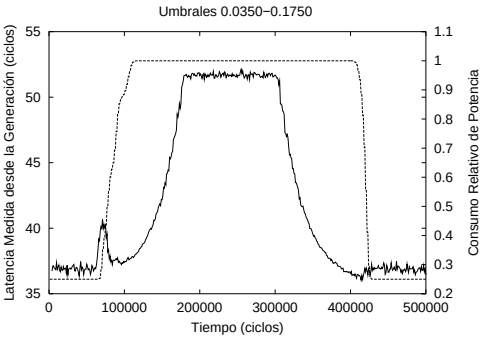


(f) Latencia y potencia en la configuración 12.

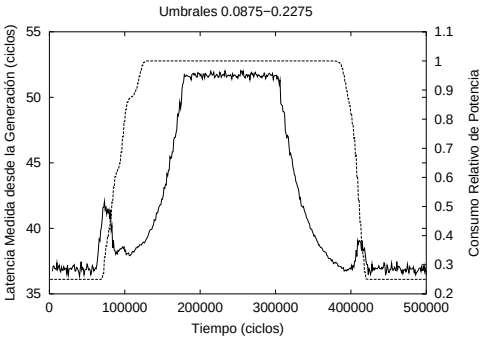
Figura 4.29: Evaluación dinámica para el toro 2D para distintos puntos del mapa de posibles umbrales.



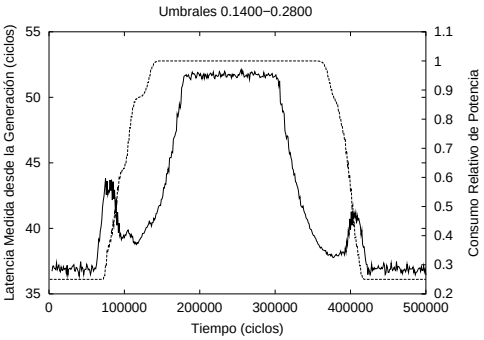
(a) Mapa de posibles umbrales.



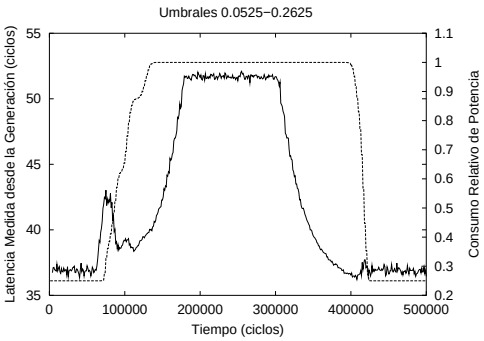
(b) Latencia y potencia en la configuración 3.



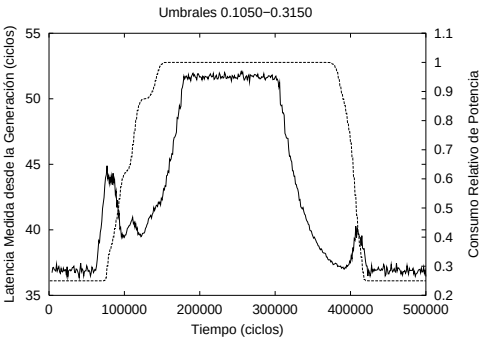
(c) Latencia y potencia en la configuración 4.



(d) Latencia y potencia en la configuración 5.

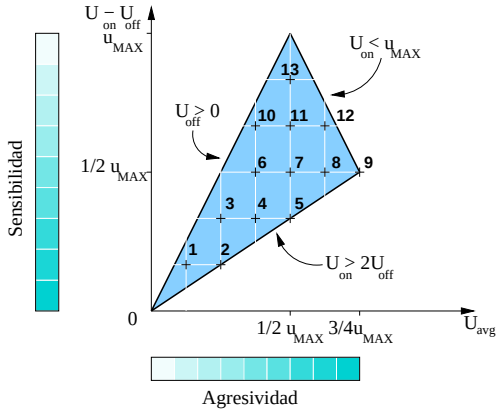


(e) Latencia y potencia en la configuración 6.

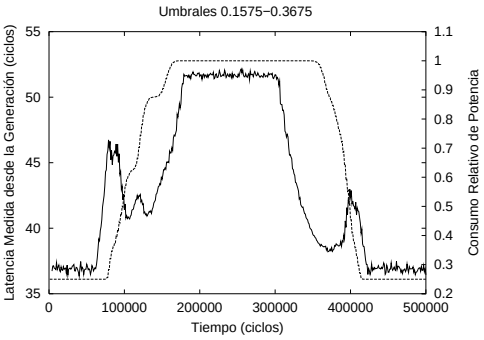


(f) Latencia y potencia en la configuración 7.

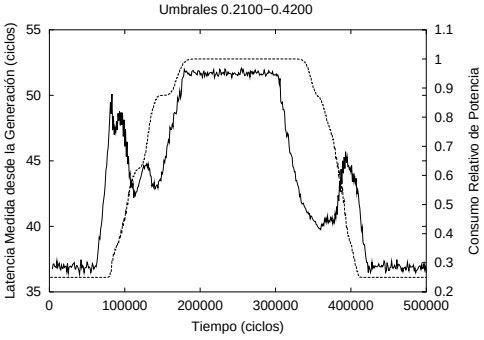
Figura 4.30: Evaluación dinámica para el toro 3D para distintos puntos del mapa de posibles umbrales.



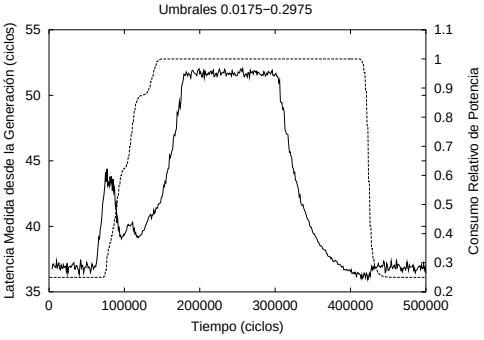
(a) Mapa de posibles umbrales.



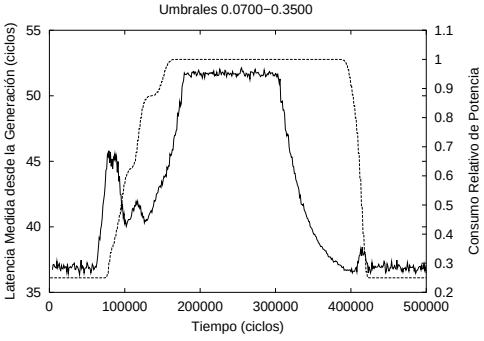
(b) Latencia y potencia en la configuración 8.



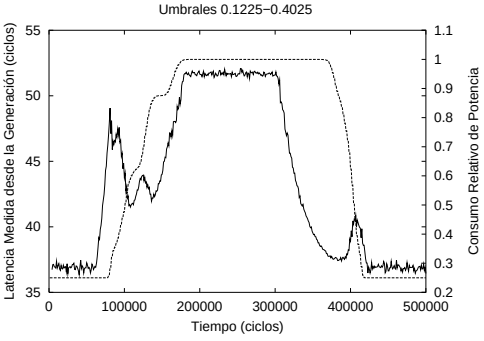
(c) Latencia y potencia en la configuración 9.



(d) Latencia y potencia en la configuración 10.

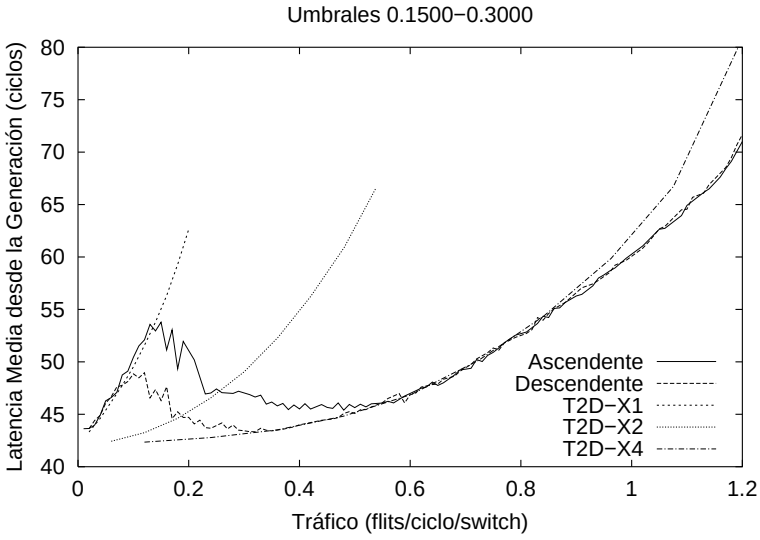


(e) Latencia y potencia en la configuración 11.

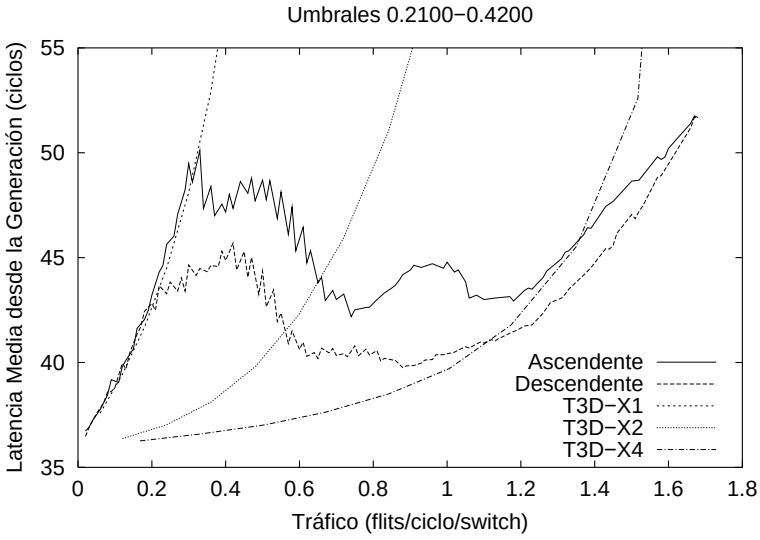


(f) Latencia y potencia en la configuración 12.

Figura 4.31: Evaluación dinámica para el toro 3D para distintos puntos del mapa de posibles umbrales.



(a) Toro 2D.



(b) Toro 3D.

Figura 4.32: Latencia frente a tráfico para carga ascendente y descendente para el punto 9 del mapa de umbrales.

tado de unir entre sí puntos obtenidos para diferentes escalones de tráfico en simulaciones independientes. En cambio, las curvas dinámicas, se han generado a partir de una sola simulación en la que el tráfico sube o baja en forma de rampa.

En la figura se aprecia que, para cargas bajas, donde la mayoría de enlaces están desconectados, la latencia sigue aproximadamente el comportamiento de la red X1. Cuando el tráfico es mayor la latencia se sitúa en un valor intermedio a la de las configuraciones X1 y X2. Una parte de los enlaces agregados tienen todos los enlaces desconectados (equivalente a X1) y otra parte ha iniciado ya las conexiones (X2 o superior). Para tráfico aún más alto la curva de latencia tiende a aproximarse a la curva X4 cuando la mayoría de los enlaces trabajan al 100 %. La diferencia observada para tráfico alto se debe a que las curvas no son estrictamente comparables puesto que han sido generadas de forma distinta.

4.3.6.1. Evaluación con tráfico autosimilar

Además de la evaluación experimental con carga basada en un tiempo entre generación de mensajes distribuido uniformemente, se ha completado el estudio utilizando tráfico auto-similar, más representativo de las cargas de trabajo que se pueden encontrar en los *clusters [29].* La carga auto-similar se ha generado por la agregación de fuentes de ON/OFF. Las fuentes de ON/OFF concatenan períodos de inyección de tráfico con periodos de inactividad, ambos con distribución Pareto. En nuestro simulador, cada nodo actúa como una fuente ON/OFF. Para cada enlace de la red, el tráfico generado por cada nodo se combina de acuerdo con la distribución de destinos y el algoritmo de encaminamiento, produciendo así un tráfico auto-similar. Hemos verificado la autosimilitud del tráfico mediante la herramienta *Selfis* [37], obteniendo, en todos los casos, valores para el parámetro *Hurst* mayores a *0.7*[2].

Los experimentos se han realizado empleando rampas de tráfico similares a las empleadas con tráfico uniforme. Las simulaciones se inician con niveles de tráfico del 2 % del máximo soportado por la red; tras un período de tiempo, la carga crece durante 120000 ciclos de forma constante hasta un valor promedio del 85 % de la carga de saturación medida con tráfico uniforme (U_{MAX}). Se ha usado una carga inferior al 100 % porque la naturaleza de mayor variabilidad de la carga autosimilar puede provocar picos rápidos de carga que pueden saturar momentaneamente la red. Esta carga elevada se mantiene durante 120000 ciclos. Entonces, el tráfico de

[2]El grado de autosimilitud se puede expresar usando únicamente un parámetro que expresa la rapidez del decrecimiento de la función de autocorrelación. A este parámetro se le denomina parámetro de *Hurst*. El tráfico generado es estadísticamente autosimilar si $0{,}5 \leq H \leq 1$. A medida que $H \rightarrow 1$ el grado de autosimilitud se incrementa [40, 42].

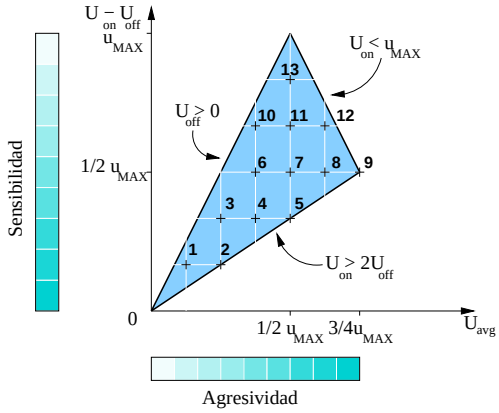
entrada se decrementa nuevamente de forma constante hasta alcanzar el valor inicial. El estado de la red en el instante inicial es de desconexión de todos los enlaces posibles, de acuerdo con el mecanismo propuesto (el consumo relativo de potencia es del 25 %).

Los resultados se muestran en la figura 4.33, para el toro 2D, y en la figura 4.34, para el toro 3D. Se presentan resultados para las configuraciones más agresivas de entre las posibles de acuerdo con el mapa de umbrales. En ambos casos, y con propósitos de comparación, la subfigura (b) muestra la latencia nominal de la red (sin mecanismo de gestión dinámica de los enlaces) junto con el tráfico recibido (que es el mismo para todas las configuraciones probadas).

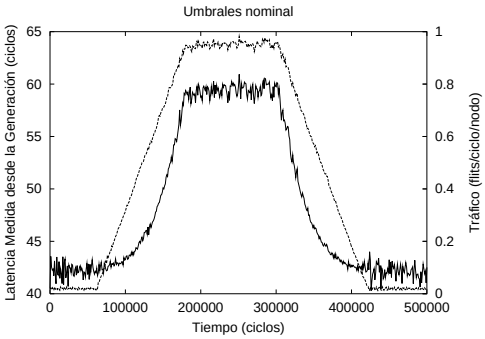
El comportamiento de la red bajo la carga autosimilar es comparable al obtenido para carga uniforme. Inicialmente, todos los enlaces de la red están desconectados y el consumo es del 25 % del valor nominal. Como el tráfico inicial es bajo, el mecanismo de ahorro de energía mantiene los enlaces desconectados. Conforme el tráfico aumenta, la latencia inicia un ligero aumento. El incremento consiguiente de la utilización de los enlaces provoca a su vez que se inicie la reconexión de enlaces desconectados. Ello provoca una disminución de la latencia (alrededor de los 100000 ciclos) que vuelve a aumentar una vez todos los enlaces han sido conectados y el tráfico sigue aumentando. En la primera parte del incremento de tráfico, muchos de los enlaces todavía continúan desconectados, por lo que la latencia de los mensajes sufre un pequeño pico que desaparece cuando el mecanismo de ahorro de potencia reacciona conectando enlaces adicionales. Cuando el tráfico alcanza su valor máximo y todos los enlaces están totalmente operativos, la latencia se estabiliza en el mismo valor alcanzado por la red cuando no se utiliza mecanismo de reducción de consumo de potencia. Cuando el tráfico baja, se produce una disminución de la latencia y el consumo de potencia. Cuando los enlaces comienzan a desconectarse (hasta que quede únicamente uno activo en cada enlace agregado) se produce un pequeño incremento en la latencia, simétrico al producido frente a la rampa de tráfico ascendente.

4.3.6.2. Diagramas de histéresis

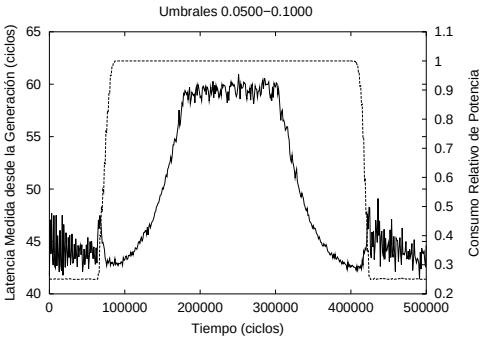
Con el fin de caracterizar totalmente el comportamiento dinámico del mecanismo de ahorro de potencia, se han construido los diagramas de histéresis de potencia consumida frente a tráfico entregado. En estos diagramas se representa la evolución de la potencia para tráfico creciente y decreciente. Se ilustra así el proceso de conexión de enlaces a medida que la utilización de los mismos se incrementa y la desconexión



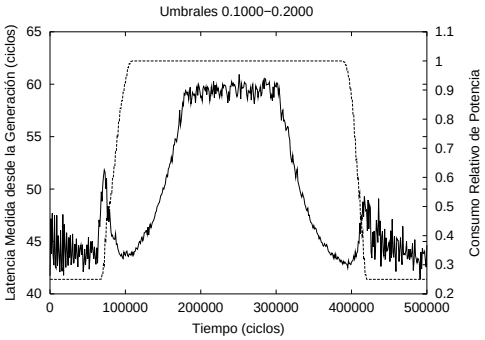
(a) Mapa de posibles umbrales.



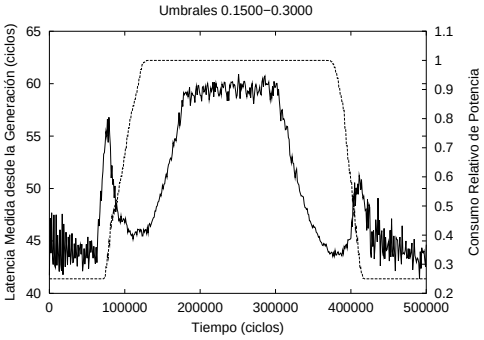
(b) Latencia y tráfico en la configuración nominal.



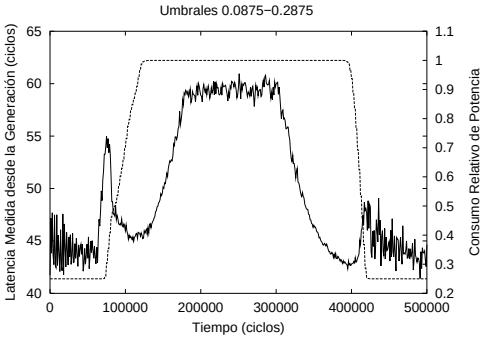
(c) Latencia y potencia en la configuración 2.



(d) Latencia y potencia en la configuración 5.

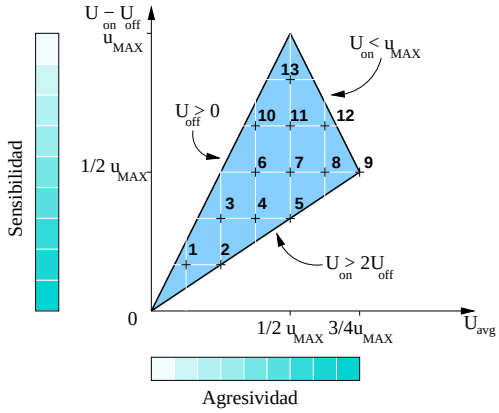


(e) Latencia y potencia en la configuración 9.

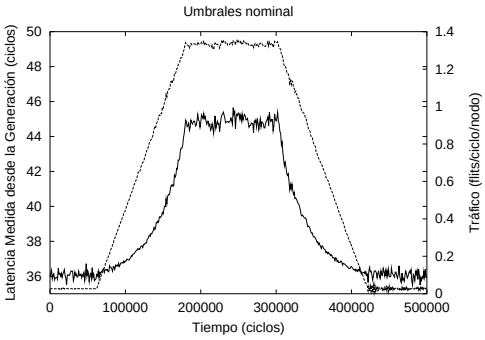


(f) Latencia y potencia en la configuración 12.

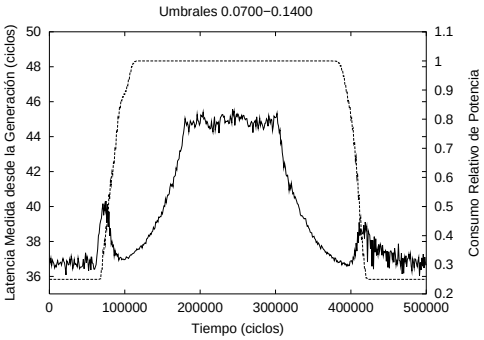
Figura 4.33: Evaluación dinámica con tráfico autosimilar para el toro 2D para distintos puntos del mapa de posibles umbrales.



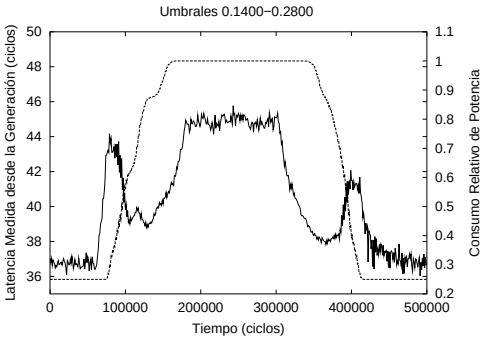
(a) Mapa de posibles umbrales.



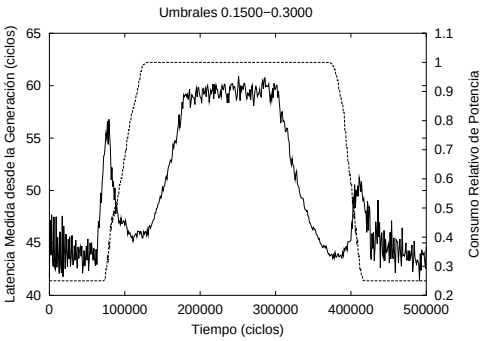
(b) Latencia y tráfico en la configuración nominal.



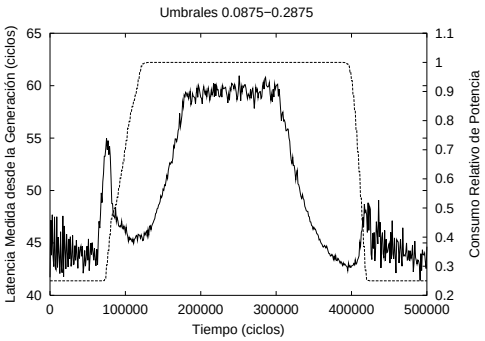
(c) Latencia y potencia en la configuración 2.



(d) Latencia y potencia en la configuración 5.



(e) Latencia y potencia en la configuración 9.



(f) Latencia y potencia en la configuración 12.

Figura 4.34: Evaluación dinámica con tráfico autosimilar para el toro 3D para distintos puntos del mapa de posibles umbrales.

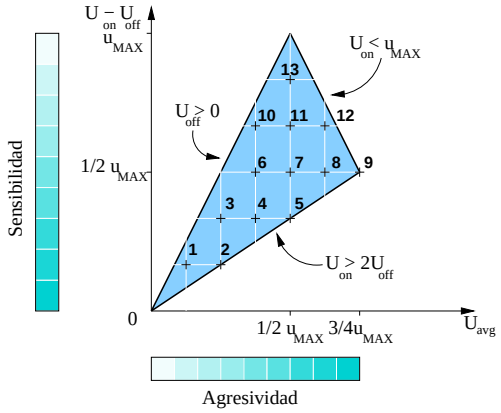
cuando su utilización desciende. Las figuras 4.35 a 4.38 recogen las configuraciones 3 a 12 del mapa de umbrales posibles para las topologías analizadas. Se observa que el comportamiento medido de la red se ajusta a lo predicho por el análisis teórico realizado en la sección 4.2.3.

Como se ha constatado en los resultados presentados hasta este punto, se observa un comportamiento similar del mecanismo para las dos topologías. También los diagramas de histéresis muestran que el rango de tráfico para el cual se produce un ahorro de potencia es mayor en la topología 3D. Como el diseño del mecanismo predice, configuraciones con una menor diferencia entre los umbrales ($U_{on} - U_{\text{off}}$) proporcionan mayor sensibilidad. Esto se verifica con las estrechas bandas de histéresis obtenidas con las configuraciones 3, 4 o 5. Por otro lado, también se verifica que configuraciones con un valor más alto del promedio entre los umbrales son más exigentes y reducen el consumo de potencia con valores de tráfico más altos.

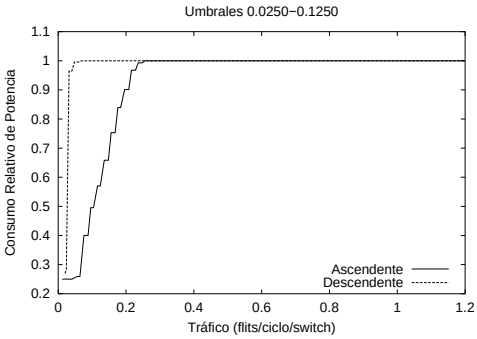
Configuraciones como la 5 o la 9 proporcionan simultáneamente una alta sensibilidad (la banda de histéresis obtenida es muy estrecha y constante) y una alta agresividad (las curvas presentan una baja pendiente). Son estas configuraciones las que reunen las mejores características de rapidez de respuesta ante cambios de la carga y buen ahorro de potencia.

4.3.7. Cómo obtener ahorro adicional en la potencia

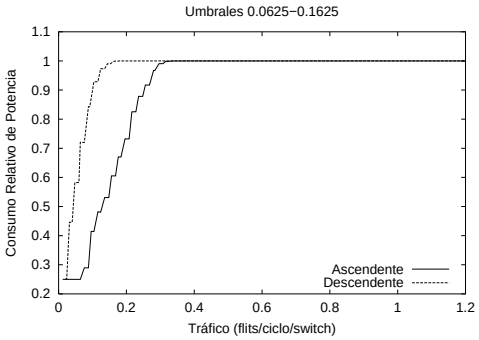
Tal y como se ha explicado en la sección 4.2.4, cuando la técnica anterior ha reducido el número de enlaces físicos activos en un enlace agregado o *trunk link* a uno solo, o cuando el diseño de la red de interconexión no incluye el uso de enlaces agregados, aun es posible obtener ahorros de potencia mayores reduciendo dinámicamente el ancho del único enlace activo que queda mientras el tráfico sea cada vez más bajo. En este apartado se valora y evalúa la reducción de potencia adicional que se puede obtener con esta técnica y se cuantifica su impacto en las prestaciones de la red. Se han realizado los experimentos considerando una red en la que ya han sido desconectados todos los enlaces posibles, de manera que solo existe un enlace disponible entre cada par de nodos. El consumo de potencia relativo máximo es, por tanto, del 25 %. De este modo, se evalúa el efecto de la reducción del ancho del enlace de forma independiente del mecanismo de conexión/desconexión de enlaces. En los experimentos presentados, los anchos de enlace posibles son del 100 %, 50 % y 25 % . Las figuras 4.39 y 4.40 presentan los resultados obtenidos sobre el toro 2D para los distintos umbrales. Las figuras 4.41 y 4.42 recogen los resultados para el



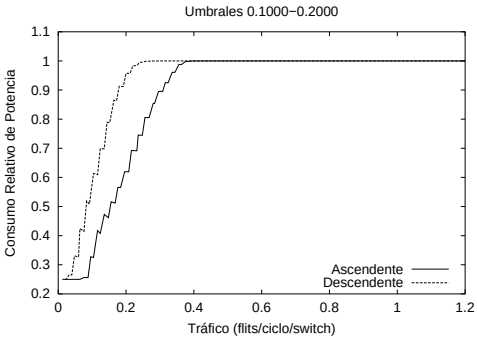
(a) Mapa de posibles umbrales.



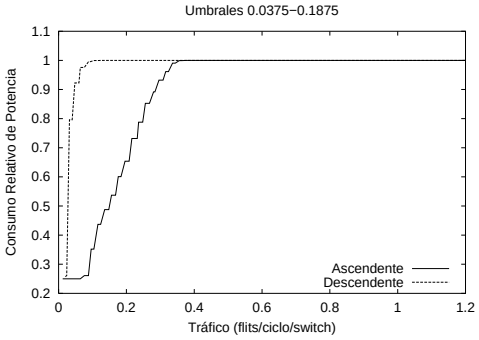
(b) Latencia y potencia en la configuración 3.



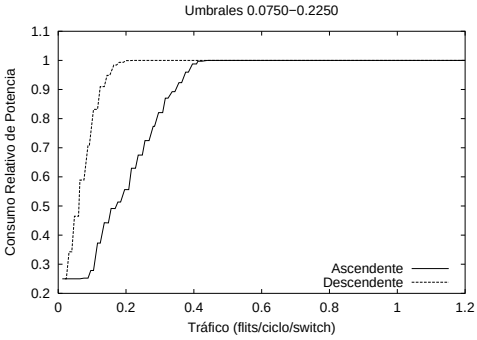
(c) Latencia y potencia en la configuración 4.



(d) Latencia y potencia en la configuración 5.

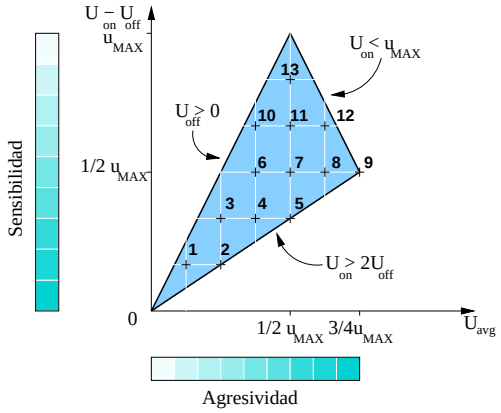


(e) Latencia y potencia en la configuración 6.

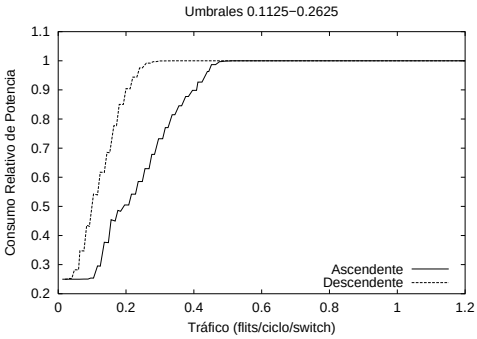


(f) Latencia y potencia en la configuración 7.

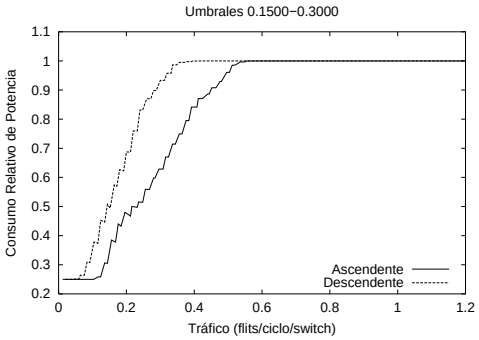
Figura 4.35: Diagramas de histéresis para el toro 2D para distintos puntos del mapa de posibles umbrales.



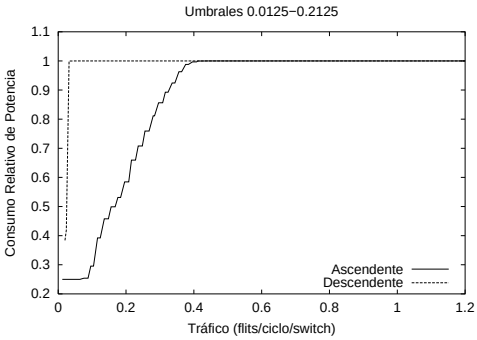
(a) Mapa de posibles umbrales.



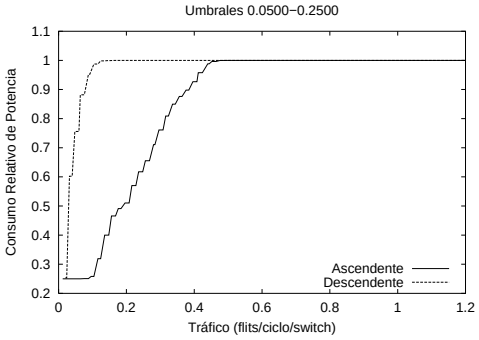
(b) Latencia y potencia en la configuración 8.



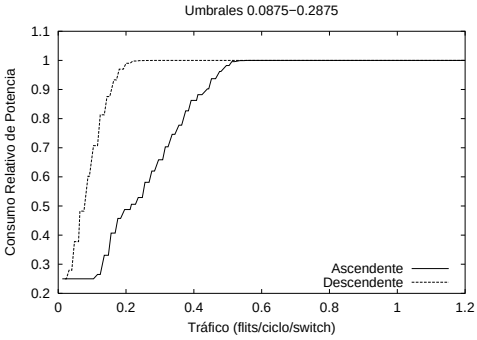
(c) Latencia y potencia en la configuración 9.



(d) Latencia y potencia en la configuración 10.

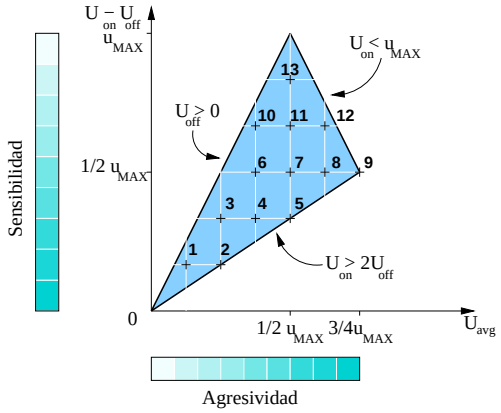


(e) Latencia y potencia en la configuración 11.

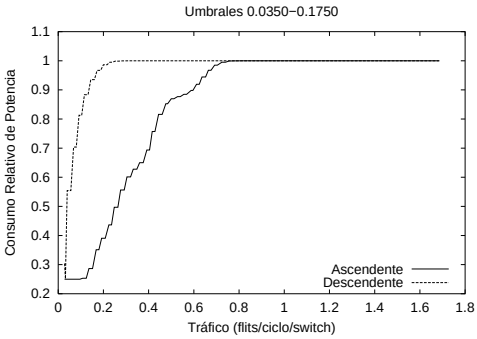


(f) Latencia y potencia en la configuración 7.

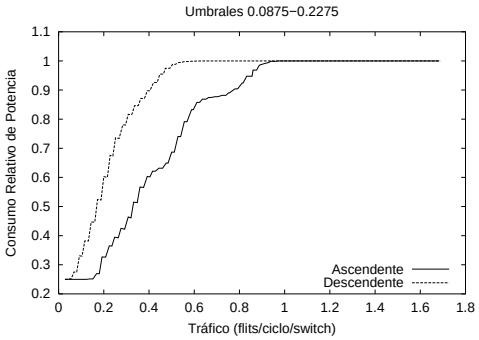
Figura 4.36: Diagramas de histéresis para el toro 2D para distintos puntos del mapa de posibles umbrales.



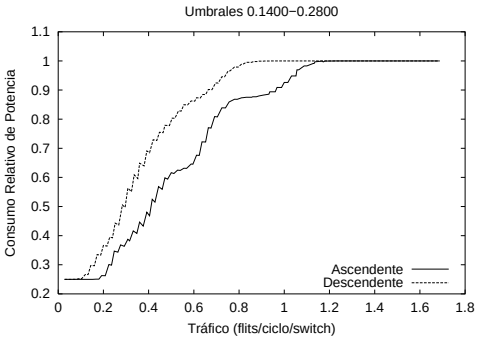
(a) Mapa de posibles umbrales.



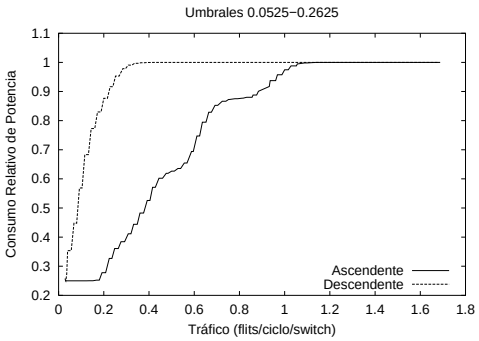
(b) Latencia y potencia en la configuración 3.



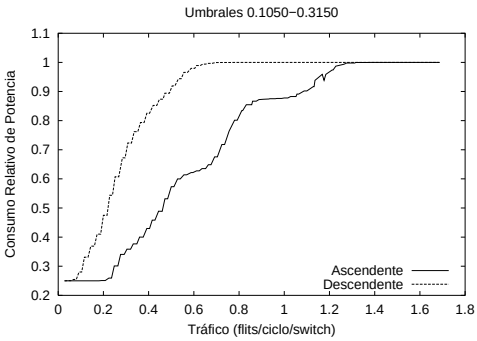
(c) Latencia y potencia en la configuración 4.



(d) Latencia y potencia en la configuración 5.

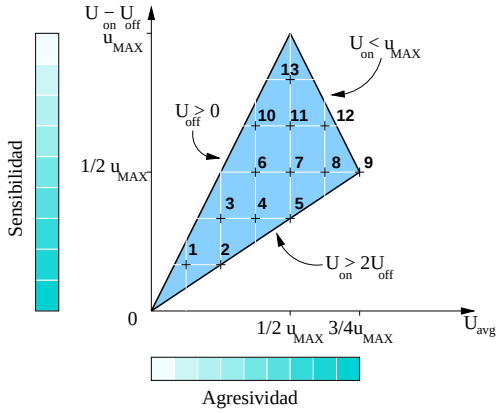


(e) Latencia y potencia en la configuración 6.

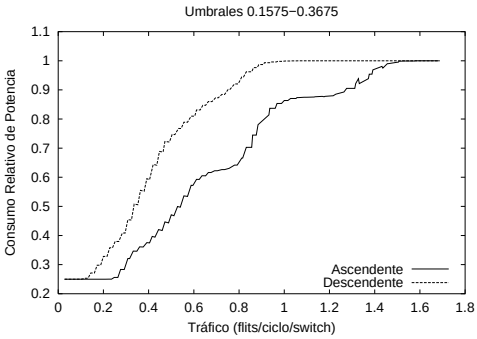


(f) Latencia y potencia en la configuración 7.

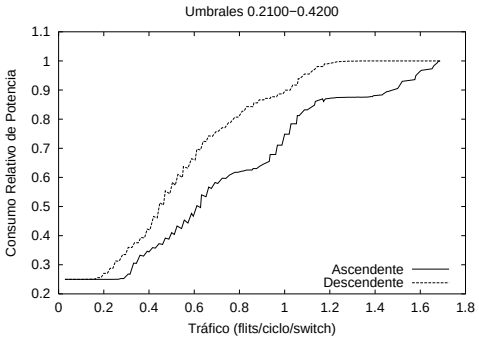
Figura 4.37: Diagramas de histéresis para el toro 3D para distintos puntos del mapa de posibles umbrales.



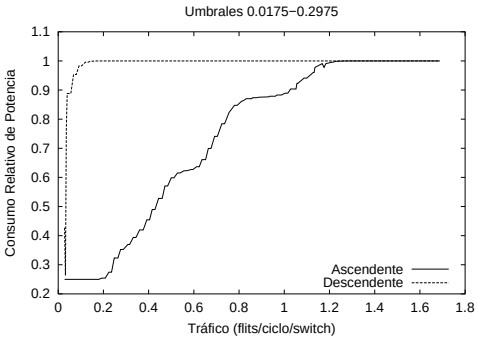
(a) Mapa de posibles umbrales.



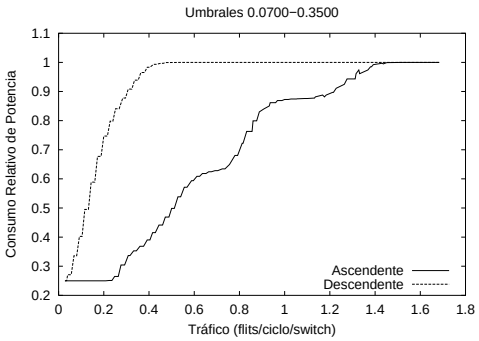
(b) Latencia y potencia en la configuración 8.



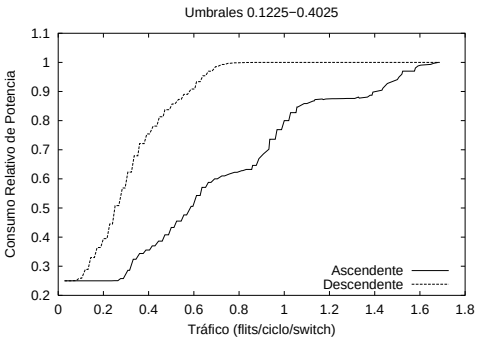
(c) Latencia y potencia en la configuración 9.



(d) Latencia y potencia en la configuración 10.



(e) Latencia y potencia en la configuración 11.



(f) Latencia y potencia en la configuración 7.

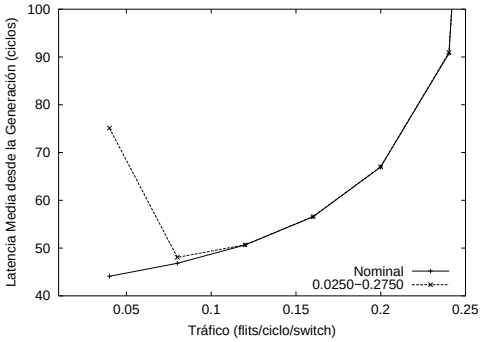
Figura 4.38: Diagramas de histéresis para el toro 3D para distintos puntos del mapa de posibles umbrales.

toro 3D. Estas figuras presentan resultados obtenidos para una evaluación estática siguiendo el modelo empleado en la sección 4.3.5. Se muestran resultados para tráficos de hasta 0,24 $flits/ciclo/nodo$, en el toro 2D, y de hasta 0,40 $flits/ciclo/nodo$, en el toro 3D, puesto que es por debajo de estos umbrales cuando actúa el mecanismo de reducción del ancho de banda. En este punto, la red con enlaces agregados ($4X$) se comporta como una red formada por enlaces simples ($1X$), cuyas prestaciones básicas se recogen en la figura 4.10. Se observa en los resultados que, cuando el tráfico se acerca a los límites indicados, la curvas de latencia tienden hacia las curvas nominales (conmutadores con un enlace por dimensión funcionando al 100 %). Es a partir de este punto cuando se activa el mecanismo de conexión dinámica de enlaces cuya evaluación ha sido ya presentada. Ello es debido a que la utilización de los enlaces es lo suficientemente alta para provocar la conexión de enlaces adicionales (véanse los resultados de la evaluación estática del mecanismo, sección 4.3.5).

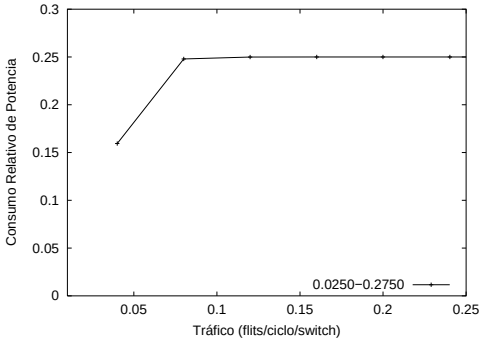
Tal y como se puede ver, si se reduce el ancho de los enlaces de acuerdo con su utilización, se obtienen ahorros en la potencia adicionales a expensas de un incremento en la latencia de los mensajes. La penalización en latencia es significativamente superior al caso del mecanismo operando con varios enlaces en paralelo. No obstante, se ha verificado que en todos los casos, para todos los umbrales probados, el producto $L_{rel} \times P_{rel}$ es todavía inferior a la unidad, lo que indica que el beneficio por la reducción en potencia supera al coste por el incremento en la latencia. La activación de este mecanismo extra de ahorro de potencia depende de si la aplicación que se está ejecutando en el sistema es sensible a la latencia y por tanto se puede ver muy afectada por ella o no. El administrador del sistema debería decidir para estas aplicaciones en particular si resulta apropiado o interesante activar el mecanismo con el objetivo de obtener un equilibrio aceptable entre el consumo de potencia y el incremento en la latencia. Estos resultados muestran que hay una oportunidad para ahorros en la potencia adicionales y que simplemente se debe considerar la penalización en prestaciones a la hora de aplicar las reducciones de potencia extras para ver su compensación.

El comportamiento del mecanismo en términos de ahorro frente a agresividad de los umbrales es equivalente a lo observado para la red operando con enlaces agregados. Umbrales más agresivos proporcionan una mayor reducción del consumo de potencia, y una mayor penalización en la latencia. Como en el caso anterior, de nuevo la topología 3D es la que proporciona mayores márgenes de ahorro.

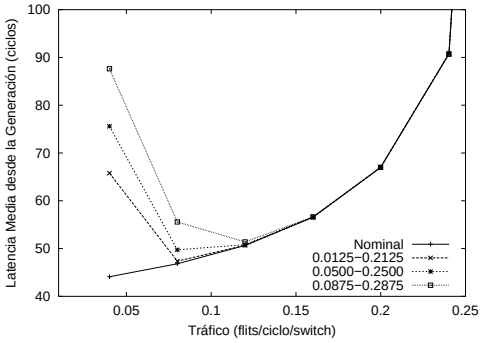
Nótese que para valores de tráfico correspondientes a la transición entre uno y varios enlaces conectados, tendremos a los dos mecanismos trabajando juntos en pa-



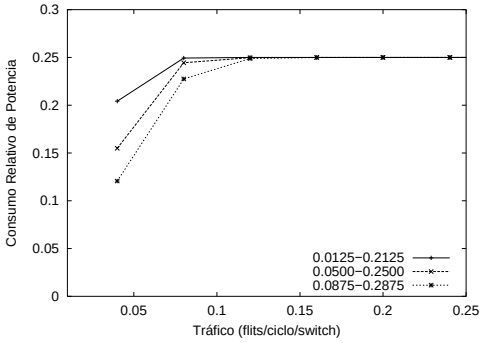
(a) Latencia con histéresis de 0,25.



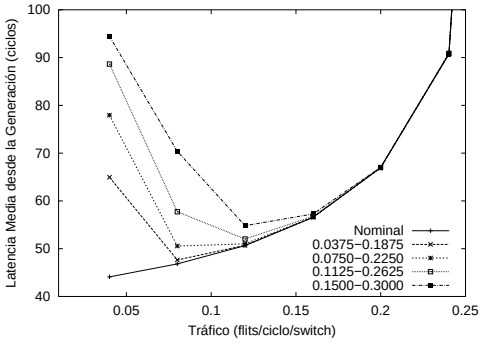
(b) Potencia con histéresis de 0,25.



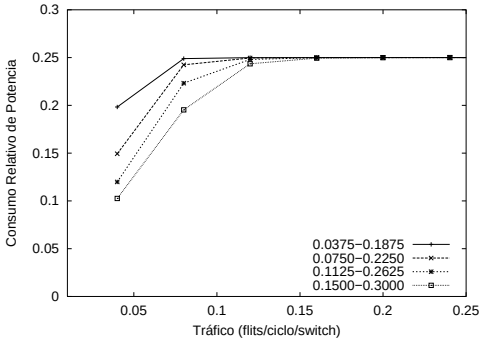
(c) Latencia con histéresis de 0,2.



(d) Potencia con histéresis de 0,2.

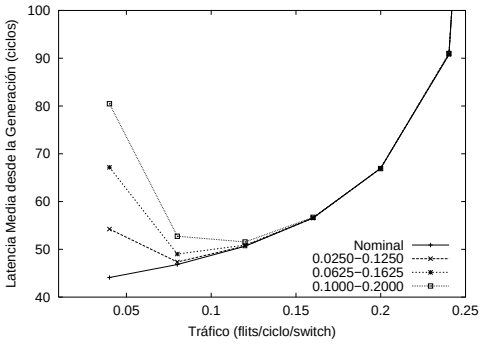


(e) Latencia con histéresis de 0,15.

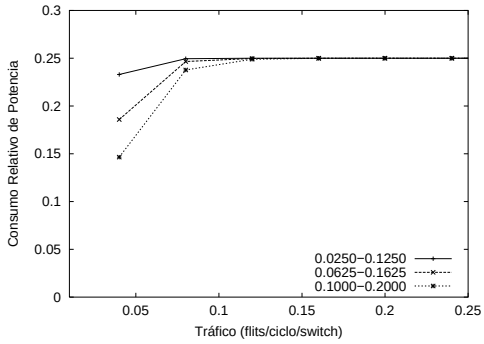


(f) Potencia con histéresis de 0,15.

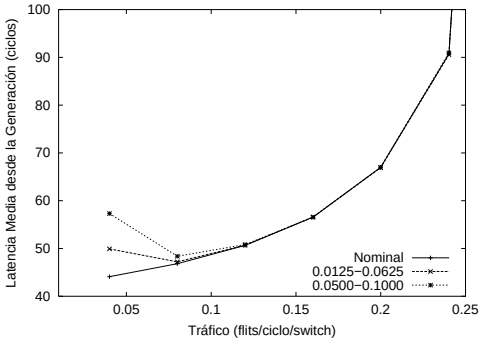
Figura 4.39: Latencia media y consumo relativo para el toro 2D actuando sobre un solo enlace (primera parte).



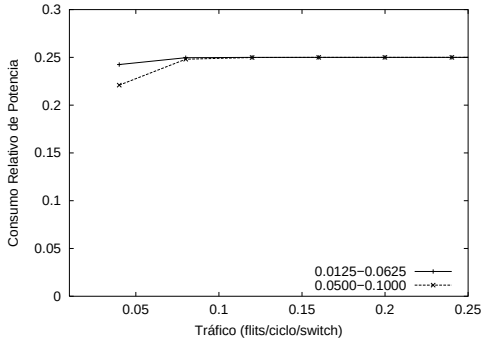
(a) Latencia con histéresis de 0, 1.



(b) Potencia con histéresis de 0, 1.



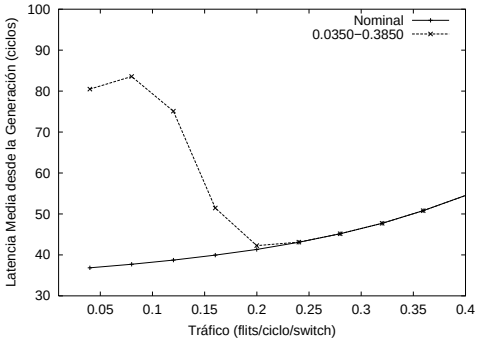
(c) Latencia con histéresis de 0, 05.



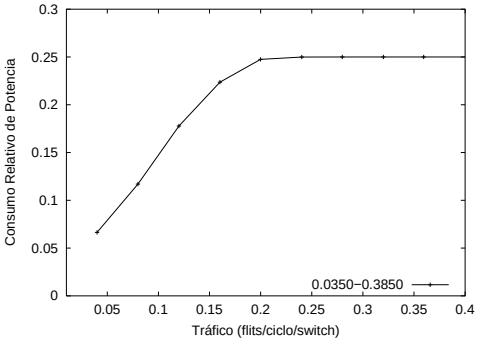
(d) Potencia con histéresis de 0, 05.

Figura 4.40: Latencia media y consumo relativo para el toro 2D actuando sobre un solo enlace (segunda parte).

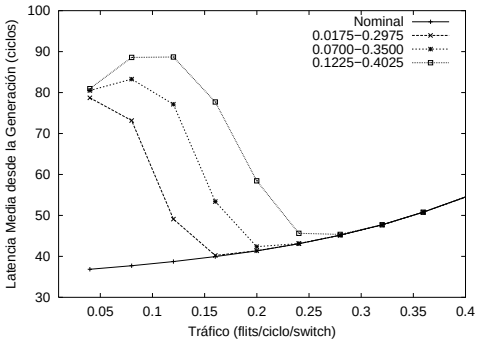
ralelo. De este modo habrá unos pocos enlaces con varios enlaces activos por enlace agregado, varios enlaces con sólo un enlace activo por enlace agregado, y finalmente muchos enlaces con su tamaño reducido a la mitad o incluso a un cuarto de su tamaño original. Para un enlace en particular, sólo se puede aplicar la técnica de reducción del ancho del enlace cuando el primer mecanismo ya ha desconectado todos los enlaces físicos del enlace agregado menos uno. Los resultados de latencia y de consumo de potencia se sitúan en ese caso a caballo de los dos casos presentados en este capítulo.



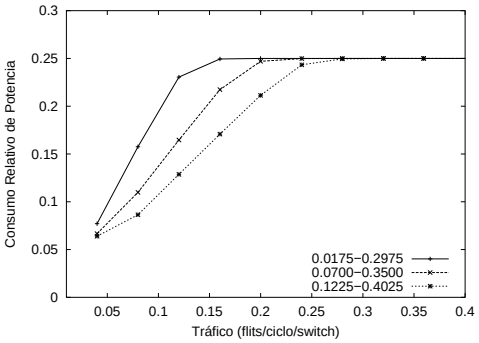
(a) Latencia con histéresis de 0,35.



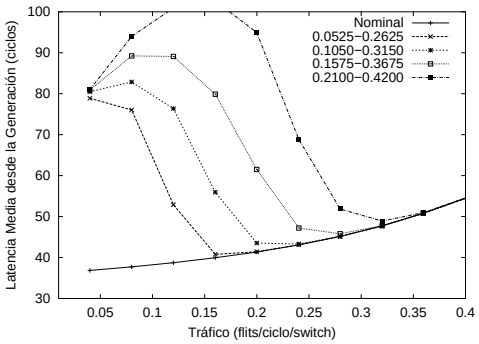
(b) Potencia con histéresis de 0,35.



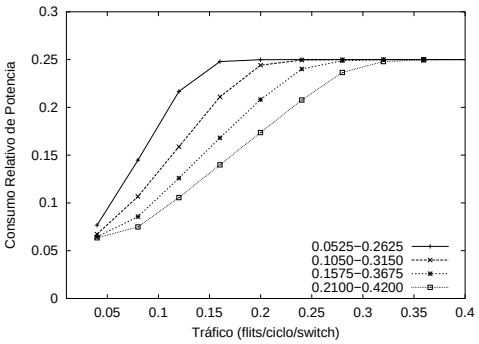
(c) Latencia con histéresis de 0,28.



(d) Potencia con histéresis de 0,28.

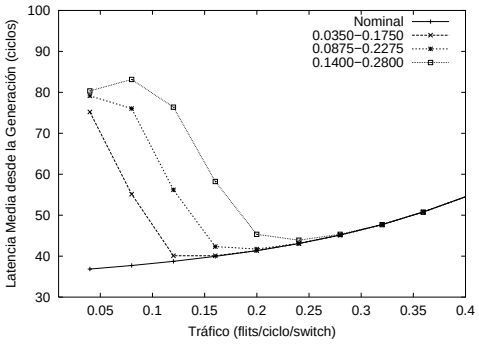


(e) Latencia con histéresis de 0,21.

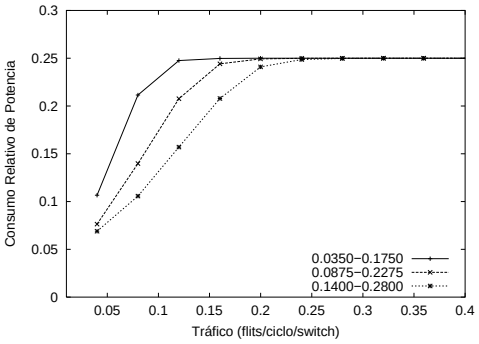


(f) Potencia con histéresis de 0,21.

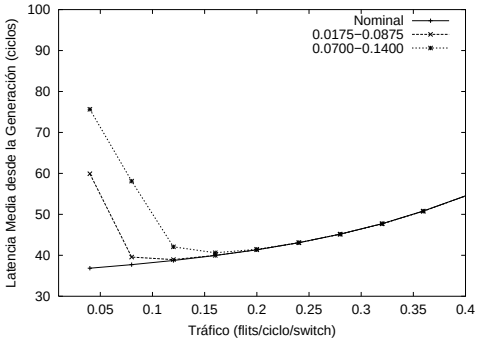
Figura 4.41: Latencia media y consumo relativo para el toro 3D actuando sobre un solo enlace (primera parte).



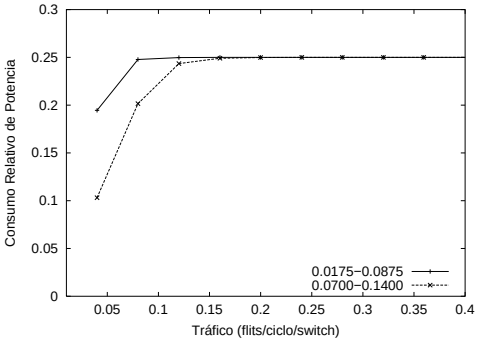
(a) Latencia con histéresis de 0, 14.



(b) Potencia con histéresis de 0, 14.



(c) Latencia con histéresis de 0, 07.



(d) Potencia con histéresis de 0, 07.

Figura 4.42: Latencia media y consumo relativo para el toro 3D actuando sobre un solo enlace (segunda parte).

4.4. Conclusiones

En este capítulo se ha presentado un mecanismo de ahorro del consumo de potencia para redes de interconexión directas construidas con conmutadores de alto grado, donde cada dimensión de la red está implementada con varios enlaces en paralelo (enlaces agregados o *trunk link*). El mecanismo propuesto se basa en conectar o desconectar de forma dinámica los enlaces que componen el enlace agregado en función del tráfico presente en la red. Se ha realizado una caracterización del mecanismo en función de los parámetros que lo gobiernan, los cuales permiten ajustar la política de gestión energética de la red. La exhaustiva evaluación experimental presentada revela el potencial del mecanismo para reducir significativamente el consumo de potencia, a expensas de moderados incrementos de las latencia de los mensajes. En

todos los experimentos realizados, se ha verificado que el producto de potencia relativa por latencia relativa proporciona resultados favorables. Ello es una muestra de la bondad del mecanismo para admitir configuraciones agresivas y aun así mantener siempre un comportamiento favorable.

La principal conclusión que se puede extraer de las pruebas realizadas sobre el efecto de la longitud de los mensajes (sección 4.3.5.1) es que el mecanismo de ahorro de potencia presenta resultados ligeramente más favorables para mensajes de mayor longitud. Esto es debido a que la distribución del trafico en mensajes más largos, reduce la sobrecarga debida a las cabeceras y por tanto la utilización efectiva de los enlaces, incrementando así el ahorro de potencia.

En el caso de la función de selección, se ha verificado (sección 4.3.5.2) que una distribución más equilibrada del tráfico entre las dimensiones de la red (función de seleccion Cyclic) proporciona mejores resultados que una concentración de tráfico en unos pocos enlaces que busca descargar otros para facilitar su desconexión (función de selección FirstFree), debido a las mayores prestaciones que ofrece la red.

Respecto a la selección de umbrales, se ha probado todo el mapa de umbrales posibles, y verificado que en todos los casos el balance incremento de latencia-consumo de potencia resulta favorable. Es por esto que el mecanismo ofrece la máxima flexibilidad para ser configurado de acuerdo con los parámetros de agresividad, sensibilidad e impacto en la latencia más adecuados en cada caso. Este aspecto se refuerza con el análisis del consumo de energía presentado en la sección 4.3.5.3, donde se verifica que el balance energético es favorable cuando se emplea el mecanismo de ahorro propuesto.

5

Reducción del Consumo de Potencia en Redes Indirectas

En este capítulo se presenta nuestro mecanismo de reducción del consumo de potencia para redes indirectas. Se incluye una exhaustiva evaluación experimental basada en simulación y se exponen los resultados obtenidos.

5.1. Introducción

Las tendencias actuales en computadores paralelos de altas prestaciones muestran que las redes indirectas, en particular la red de interconexión fat-tree, son una de las topologías más populares. Actualmente, un gran número de sistemas de computación de altas prestaciones son clusters. Esta es la arquitectura elegida por 416 de los 500 sistemas listados en la edición de Noviembre de 2011 de los Top500[66]. Esto representa un 83 % de los computadores de la lista (cinco años antes, sólo el 59 % de los sistemas que estaban en la lista eran clusters). Una parte muy significativa de estos clusters, el 47 %, utilizan redes de interconexión basadas en la topología indirecta

fat-tree.

La popularidad de esta topología es en parte debida a su gran ancho de banda de la bisección y la facilidad para mapear aplicaciones con topologías de comunicación arbitrarias[47]. No obstante, la mayoría de aplicaciones tienen requerimientos de topologías de comunicación que están bastante lejos de la conectividad total que proporcionan los fat-trees. Vetter y Mueller muestran que las aplicaciones que escalan más eficientemente a un gran número de procesadores utilizan patrones de comunicación punto a punto donde el número medio de destinos distintos es relativamente pequeño [67]. Esto proporciona una fuerte evidencia de que las topologías de comunicación de muchas aplicaciones utilizan una pequeña fracción de los recursos que proporcionan los fat-trees [50]. Además, el tráfico en una red de interconexión presenta grandes variaciones espaciales y temporales, dando lugar a periodos de inactividad en varios enlaces de la red [55].

Las particulares características de esta topología, que proporciona múltiples caminos alternativos para cada par origen/destino la hace una candidata excelente para aplicar en ella técnicas de reducción de consumo de potencia: es posible conseguir una reducción en el consumo de potencia apagando o encendiendo enlaces dinámicamente, mientras se ejecutan un conjunto de aplicaciones, basándose en el tráfico presente en la red. Como ya se ha tratado en esta tesis, estas técnicas de reducción de potencia son relevantes al considerar que las redes de interconexión consumen una parte significativa de la potencia del sistema. Por ejemplo, los routers y los enlaces de un servidor tipo *blade* Mellanox consumen casi la misma potencia que un procesador (15*W*) y alrededor del 37 % de la cantidad de potencia total consumida[30]. Se han propuesto varias técnicas de reducción del consumo de potencia en redes de interconexión 3.2, pero se centran exclusivamente en redes regulares directas, por lo que nosotros sabemos, no existe ningún estudio previo al nuestro realizado en redes indirectas.

5.2. Fat-trees (k-ary n-tree)

Un k-ary n-tree está formado por dos tipos de vértices: $N = k^n$ nodos de procesamiento y $S = nk^{n-1}$ switches $k \times k$ (un k-ary n-tree de dimensión $n = 0$ estaría compuesto solo por un nodo de procesamiento). Cada nodo de procesamiento está identificado por una tupla de longitud n, n-tupla, $(p_0, p_1, ..., p_{n-1})$ donde $p_i \in \{0, 1, ..., k-1\}$ para $0 \leq i \leq n-1$. Cada switch se define como un par ordenado (w, l), donde w es una tupla de longitud $n-1$, $(w_0, w_1, \ldots, w_{n-2})$, donde

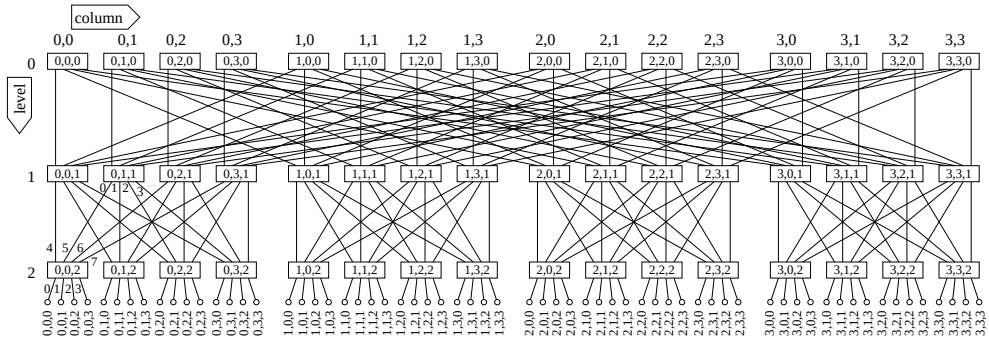


Figure 5.1: Etiquetado para los nodos, switches y enlaces de un 4-ary 3-tree .

$w_i \in \{0, 1, \ldots, k-1\}$ y $l \in \{0, 1, ..., n-1\}$ es el nivel del switch (0 es el nivel raíz).

- Dados dos switches$(w_0, w_1, ..., w_{n-2}, l)$ y $\left(w_0', w_1', ..., w_{n-2}', l\right)$, estos están conectados por un enlace si y sólo si $l' = l + 1$ y $w_i = w_i' \, \forall i \neq l$. El enlace que conecta ambos switches está etiquetado con w_l' en el switch del nivel l y con $k + w_l$ en el switch l'. Este patrón de conexión definido es conocido como *mariposa* (butterfly).

  De este modo, cada switch tiene $2k$ enlaces de salida de los cuales k están conectados a los switches o nodos de procesamiento del nivel $l + 1$ (enlaces descendentes) y los restantes k a los switches del nivel $l - 1$(enlaces ascendentes).

- Hay un enlace entre el switch del nivel inferior $(w_0, w_1, ..., w_{n-2}, n-1)$ y el nodo de procesamiento $(p_0, p_1, ..., p_{n-1})$ si y sólo si $w_i = p_i, \forall i \in i \, \{0, 1, ..., n-2\}$. El enlace está etiquetado con p_{n-1} en el switch de nivel $n - 1$.

El esquema de etiquetado mostrado en las definiciones previas convierte el k-ary n-tree en una red delta [26, 28]: cualquier camino que empiece en un switch del nivel 0 y dirigido a un nodo $(p_0, p_1, ..., p_{n-1})$ atraviesa la misma secuencia de enlaces etiquetados $p_0, p_1, ..., p_{n-1}$ [28]. Un ejemplo de este etiquetado se muestra en la figura 5.1, para un *fat-tree* cuaternario de dimensión 3 (red de 64 nodos), es decir un 4-ary 3-tree.

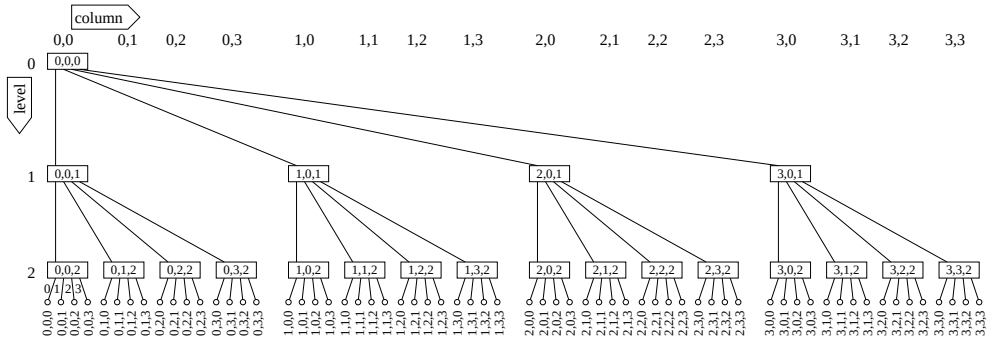


Figura 5.2: Árbol Mínimo para un *4-ary 3-tree*.

5.3. Reducción del consumo de potencia

La idea básica que hay detrás del mecanismo propuesto es idéntica a la empleada en las redes directas: conmutar dinámicamente enlaces a encendido y apagado (*on/off*), en función del tráfico requerido en la red. En nuestro modelo, se consideran enlaces bidireccionales que se pueden conectar o desconectar en una determinada dirección, bien ascendente o bien descendente. Para hacer eso, cada switch de la red mide el tráfico saliente y controla el estado de los enlaces de salida dependiendo de las variaciones de tráfico. Hay un subconjunto de los enlaces de la red, que se definen como Árbol Mínimo (*Minimal Tree*) no se puede apagar para mantener la conectividad de la red. Esto limita el nivel de ahorro de potencia que se puede conseguir, pero al mismo tiempo evita realizar sofisticadas modificaciones en el algorimo de encaminamiento [13].

Dado un *k*-ary *n*-tree, el *Minimal Tree* (MT) es el subconjunto del *k*-ary *n*-tree compuesto por todos los nodos de procesamiento, un subconjunto de los switches de comunicación y los enlaces que comunican ambos. Considerando la definición formal presentada en la Sección 5.2, un switch $(w_0, w_1, ..., w_{n-2}, l)$ pertenece al MT si cumple una de las siguientes propiedades:

1. $l < n-1$ y $w_i = 0 \, \forall i \in \{l, ..., n-2\}$
2. $l = n-1$ (todos los switches del nivel $n-1$ pertenecen al MT)

En la figura 5.2 se muestra el Árbol Mínimo para un *fat tree* cuaternario de dimensión 3 (4-ary 3-tree).

El MT está formado por los switches y enlaces que no se pueden apagar porque proporcionan los caminos mínimos necesarios para mantener todos los nodos de

procesamiento conectados. Dentro de estos switches, todos los enlaces descendentes y el enlace ascendente con el índice k también pertenecen al MT. Así hay $k+1$ enlaces en el MT por switch en el MT, excepto en el switch raíz: sólo sus k enlaces descendentes están en el MT. Los enlaces que conectan los $N = k^n$ nodos de procesamiento con los $S = nk^{n-1}$ switches también pertenecen al MT.

El número de switches en el MT es:

$$|MT_S| = k^{n-1} + k^{n-2} + ... + k + 1 = \frac{1-k^n}{1-k}$$

El número de enlaces unidireccionales en el MT es:

$$|MT_L| = (k+1)\,|MT_S| - 1 + k^n = 2k\,|MT_S|$$

Considerando sólo el consumo de potencia de los enlaces, el consumo de potencia relativo mínimo viene dado por el ratio entre el número de enlaces en el árbol mínimo y el número de enlaces total o enlaces en el *fat-tree*.

$$p_{min} = \frac{|MT_L|}{2k \times S} = \frac{|MT_S|}{S}$$

A modo de ejemplo:

- Para la topología 4-ary 3-tree los valores son:
  - $|MT_S| = 21$
  - $|MT_L| = 168$
  - $p_{min} = 0{,}4375$
- Para 4-ary 4-tree los valores son:
  - $|MT_S| = 85$
  - $|MT_L| = 680$
  - $p_{min} = 0{,}3320$

En la tabla 5.1 se muestran cuáles serían los valores para la potencia relativa mínima obtenidos para distintas configuraciones de dimensiones del *fat-tree (n)* y aridad de los switches *(k)*. Los resultados indican que el potencial para la reducción del consumo de potencia aumenta tanto cuando aumenta el número de dimensiones como cuando aumenta el tamaño de los switches.

		n				
		2	3	4	5	6
k	2	0,75000	0,58333	0,46875	0,38750	0,32813
	4	0,62500	0,43750	0,33203	0,26641	0,22217
	8	0,56250	0,38021	0,28564	0,22856	0,19048
	16	0,53125	0,35547	0,26666	0,21333	0,17778

Tabla 5.1: Potencia relativa mínima para distintas configuraciones de *fat-tree.*

5.3.1. Descripción del mecanismo

El mecanismo propuesto de ahorro de potencia controla el estado de los enlaces de la red de acuerdo las reglas descritas a continuación. El término enlace ascendente se aplica a aquellos que permiten llegar a un switch situado gráficamente más arriba en la red, de acuerdo con las representaciones incluidas en las figuras 5.1 y 5.2. Con el mismo criterio, enlace descendente se refiere a los que conectan con nodos situados gráficamente más abajo en el árbol.

- Enlaces ascendentes: la utilización de los enlaces ascendentes de un determinado switch se usa para tomar la decisión de conectar o desconectar dichos enlaces ascendentes. Esta decisión se propaga de forma ascendente para garantizar al menos un camino a los switches del nivel 0 (esto es necesario para proporcionar un camino a cada uno de los destinos posibles), y se propaga también de forma descendente para garantizar rutas descendentes a los procesadores, por el mismo motivo.

- Enlaces descendentes: los enlaces de un switch dado se conectan o desconectan todos al mismo tiempo. Se desconectan cuando ese switch no puede recibir tráfico descendente, esto es, cuando todos sus enlaces de entrada (desde niveles superiores o inferiores) han sido ya desconectados. Estos enlaces se conectarán de nuevo cuando se conecte al menos uno cualquiera de los enlaces de entrada, puesto que los mensajes que lleguen a través de ese enlace pueden requerir utilizarlos.

- Un switch se pone en estado de bajo consumo (*standby*) en el caso de que todos sus enlaces de entrada, tanto los que llegan de niveles superiores como inferiores, estén inactivos. En este estado, la única funcionalidad necesaria es poder detectar que se activa un enlace de entrada y que por tanto se debe iniciar la reconexión del switch al modo de funcionamiento normal. Si bien esto redunda

en un ahorro adicional de potencia, en este trabajo no se considera la reducción de potencia adicional que se obtiene si el switch se sitúa en estado de *standby*.

Teniendo en cuenta las consideraciones generales que se acaban de presentar, el mecanismo de ahorro de potencia se implementa a nivel de switch. De forma equivalente al mecanismo propuesto para redes directas (capítulo 4), una pareja de umbrales de utilización permite ajustar las características de la estrategia de ahorro de potencia: U_{off} es el umbral de desconexión y U_{on} es el umbral de conexión. Un nivel de tráfico que implique una baja utilización de los enlaces de un switch (por debajo de U_{off}) provocará la desconexión de enlaces, mientras que un tráfico elevado (por encima de U_{on}) indicará la conveniencia de conectar más enlaces. El comportamiento concreto de un switch de la red vendrá condicionado por el tráfico que soporte con relación a los umbrales de control y de su posición física en la red (en particular si pertenece o no al árbol mínimo y su nivel en la red) de acuerdo con la siguiente estrategia:

- Switches pertenecientes al árbol mínimo (MT). Para el switch raiz del árbol, $(0, 0, \ldots, 0)$, no se aplica ningún mecanismo de ahorro de potencia porque sus enlaces de salida (que son solo descendentes) deben estar siempre conectados. Para el resto de switches del árbol mínimo, cada uno ellos mide la utilización de los enlaces de salida ascendentes y periódicamente, calcula la utilización media de dichos enlaces, u_{up}. Los enlaces susceptibles de ser desconectados son exclusivamente aquellos que no pertenecen al árbol mínimo. Es decir, todos los enlaces ascendentes menos uno; en particular, los enlaces numerados entre $k+1$ y $2k-1$. Los enlaces descendentes no se pueden desconectar porque pertenecen al MT y proporcionan la conectividad necesaria hacia los nodos de procesamiento.

  - Cuando $u_{up} < U_{off}$ entonces se desconecta uno de los enlaces de salida ascendentes. El primer enlace que se debe desconectar es el enlace $2k-1$ después $2k-2$ y así sucesivamente. El enlace k no se puede apagar porque pertenece al MT y proporciona por lo tanto la conectividad mínima hacia el nivel superior de switches. Cuando la carga es baja, la secuencia de desconexión concentra el tráfico en unos pocos switches en el nivel superior, favoreciendo la estrategia de ahorro de consumo de potencia.
  - Cuando $u_{up} > U_{on}$, y existen enlaces de salida ascendentes desconectados, se activa uno de estos enlaces. El enlace seleccionado en este caso es el que tiene el índice más bajo de entre los desconectados.

- Switches no pertenecientes al árbol mínimo (MT).
  - Enlaces en dirección ascendente (nótese que los switches de nivel 0 no disponen de estos enlaces): para estos enlaces se aplica, para todos ellos sin excepción, la misma técnica de conexión/desconexión basada en utilización que para los switches del árbol mínimo. Como funcionalidad adicional, se anticipan acciones de conexión/desconexión en función de la actividad de los switches vecinos:
    - Cuando un switch detecta que uno de sus enlaces de entrada en dirección ascendente ha sido desconectado (inicia la transición desde el estado *on* al estado *off*) entonces desconecta un enlace de subida. Se trata de proporcionar un ancho de banda de salida del switch equivalente al de entrada. En concreto, cuando detecta que el enlace conectado al puerto i ($0 \leq i < k$) se desconecta, entonces se desconecta el enlace de subida $i + k$. Como se recibe menos tráfico del nivel inferior, necesita menos ancho de banda hacia el nivel superior. Esta política consigue un buen nivel de equilibrio entre el tráfico de entrada de los niveles inferiores y el tráfico de salida disponible hacia los niveles superiores. Si todos los enlaces de entrada desde el nivel inferior se desconectan, todos los enlaces de salida hacia el nivel superior se pueden desconectar, puesto que el switch no va a recibir tráfico en dirección ascendente.
    - Cuando un switch detecta que un enlace de entrada desde el nivel inferior (por tanto en dirección ascendente) está siendo conectado (el enlace comienza una transición desde el estado apagado, *off*, al estado encendido, *on*) el switch inmediatamente inicia la conexión del correspondiente enlace de subida. Es importante tener en cuenta que ambos procesos pueden ocurrir simultáneamente; además, la conexión del enlace de subida puede disparar más reconexiones en niveles superiores.
  - Enlaces en dirección descendente: el comportamiento de los enlaces descendentes es diferente puesto que no pueden desconectarse de forma independiente. Un mensaje en su fase de encaminamiento descendente dispone de una ruta determinista para alcanzar al nodo destino. Como se ha descrito en el Capítulo 3, un mensaje que se dirige a un nodo de procesamiento $(p_0, p_1, ..., p_{n-1})$ debe usar el puerto p_l en el nivel l. Mientras

que un switch tenga enlaces de entrada activos desde los niveles superior o inferior, todos los enlaces descendentes tienen que estar activos (en el estado *on*), puesto que proporcionan la conectividad necesaria para los mensajes descendentes hasta los nodos destino.

- Cuando todos los enlaces de entrada del switch se desconectan, se puede ahorrar potencia adicional conmutando el switch a un estado de reposo o *standby*, desconectando elementos adicionales del switch (por ejemplo los buffers). Tan pronto como uno de los enlaces de entrada inicie su proceso de activación, el switch vuelve a un estado activo y todos sus enlaces descendentes se conectan para proporcionar caminos descendentes hacia los nodos de procesamiento. En el caso particular de que el switch se active por un enlace ascendente, además de todos sus enlaces descendentes se debe activar también uno de sus enlaces ascendentes. Esto es necesario para garantizar que los mensajes ascendentes pueden alcanzar todos los destinos posibles.

5.3.2. Parámetros de control del mecanismo. Agresividad y sensibilidad.

El comportamiento del sistema de ahorro de potencia está gobernado por la pareja de umbrales U_{off} y U_{on} que se utilizan para gestionar el mecanismo de conexión dinámica de los enlaces. El efecto de estos dos parámetros se puede analizar en términos de dos aspectos complementarios: agresividad y sensibilidad del mecanismo. Una descripción detallada de estos aspectos ha sido incluida en la sección 4.2.2, dedicada a las redes directas.

Al igual que en el caso analizado para redes directas, existe un número de limitaciones que se deben aplicar al conjunto de posibles valores de los umbrales.

- Ambos umbrales deben ser positivos y distintos de cero, $U_{on} > 0$ y $U_{off} > 0$.
- U_{on} debe ser mayor que U_{off}, $U_{on} > U_{off}$.
- U_{on} debe ser menor que la utilización máxima alcanzada por los enlaces con la carga más alta aceptada por la red (U_{MAX}). En caso contrario, la red entraría en saturación antes de intentar conectar los enlaces desconectados, si los hubiere. Por tanto, $U_{on} < U_{MAX}$.
- La diferencia entre los umbrales debe ser suficiente para evitar la presencia de ciclos de conexión/desconexión.

Para la última restricción se ha presentado un análisis en detalle del efecto que se produce sobre switches de alto grado (sección 4.2.2) y las conclusiones que se han obtenido se pueden trasladar facilmente a las redes *fat-tree.* En un *fat-tree,* esta restricción viene condicionada por la fracción de ancho de banda en dirección ascendente disponible en un switch. Cada vez que se desconecta un enlace ascendente la carga de los enlaces que permanecen activos se ve incrementada. El mayor incremento en la carga se produce cuando se pasa de dos enlaces activos ascendentes a uno, entonces la carga se multiplica por dos. Para evitar transiciones de estado cíclicas, U_{on} debe ser mayor o igual a $2 \cdot U_{off}$. La razón es que, en el peor caso, el aumento en la utilización de los enlaces de subida cuando éstos se desconectan no será suficientemente alto para hacer que los enlaces vuelvan a reconectarse inmediatamente.

El mapa de posibles umbrales de acuerdo con las anteriores restricciones aparece representado en la figura 5.3. Al igual que en el capítulo anterior, este diagrama viene representado como una función de la diferencia de los umbrales y la media. Cualquier punto dentro de la región sombreada proporciona una configuración válida, con un nivel de sensibilidad y agresividad.

5.3.3. Umbrales estáticos

El mecanismo de gestión dinámica de los enlaces se basa en la medida de la utilización de los conmutadores en dirección ascendente, considerando el ancho de banda disponible (el proporcionado por los enlaces activos en cada momento). Por comparación de dicha utilización con los umbrales prefijados se inician las acciones de conexión/desconexión cuando se verifican las condiciones presentadas en la sección 5.3.1. Tal como se ha descrito, el mecanismo emplea un par de umbrales fijos o *estáticos* que establecen su configuración. Es importante destacar que el uso de umbrales constantes simplifica ligeramente la implementación del mecanismo, puesto que sólo es necesario comparar con un único par de umbrales [16].

No obstante lo anterior, cuando el mecanismo está en funcionamiento se producen cambios del ancho de banda disponible en función del número de enlaces operativos en cada momento. Esto hace que la conexión/desconexión de enlaces provoque cambios en la utilización de los conmutadores (considerando este efecto independientemente de eventuales cambios en la carga). Se trata del mismo fenómeno descrito en capítulo 4, sección 4.2.3. Por ejemplo, desconectar 1 enlace de 4 disponibles multiplica por 4/3 la utilización de los 3 enlaces que permanecen conectados. Ese valor de utilización es el que se volverá a comparar con los umbrales en una nueva iteración.

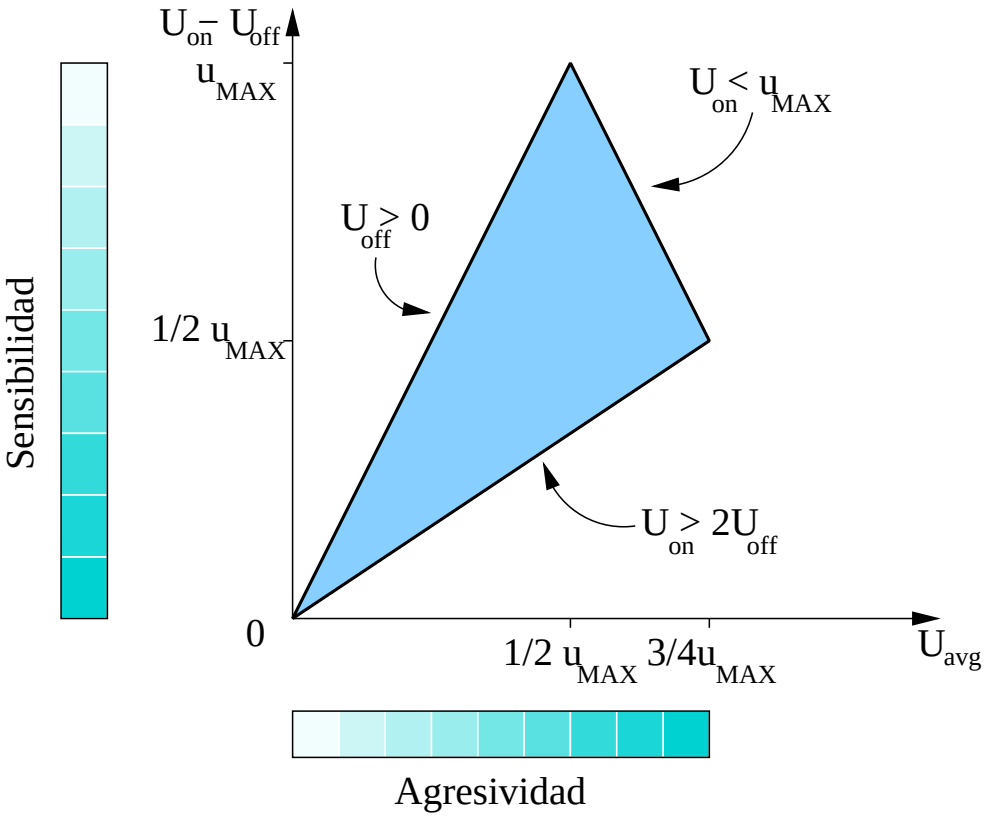


Figura 5.3: Mapa de posibles umbrales

Este comportamiento sería equivalente a reducir el valor de los umbrales en un factor 3/4 y calcular la utilización respecto al ancho de banda total y no solo respecto al disponible. A estos nuevos umbrales los denominaremos umbrales efectivos.

Los valores de los umbrales efectivos como una función del número de enlaces de salida en dirección ascendente para un *4-ary n-tree* se indican en la tabla 5.2. La columna *factor* muestra la fracción de ancho de banda de salida disponible usada para calcular los umbrales efectivos. Teniendo en cuenta los resultados presentados en la tabla, $\frac{3}{4}U_{on}$ se puede considerar como la utilización mínima para tener todos los enlaces activos, mientras que $\frac{2}{4}U_{off}$ es la utilización máxima que garantiza el ahorro de potencia máximo (3 enlaces desconectados de 4 posibles) para un switch en particular.

La figura 5.4 muestra el estado del switch (dado por el número de enlaces ascendentes activos) frente a la carga que atraviesa el switch en dirección ascendente para

	Umbrales estáticos			Umbrales efectivos	
Enlaces ascendentes activos	On	Off	Factor	On	Off
4	U_{on}	U_{off}	1	U_{on}	U_{off}
3	U_{on}	U_{off}	$3/4$	$3/4U_{on}$	$3/4U_{off}$
2	U_{on}	U_{off}	$2/4$	$2/4U_{on}$	$2/4U_{off}$
1	U_{on}	U_{off}	$1/4$	$1/4U_{on}$	$1/4U_{off}$

Tabla 5.2: Valores de los umbrales estáticos conforme a los enlaces disponibles para un switch *4-ary*.

un *4-ary n-tree.* El estado con todos los enlaces en *off* no se muestra, debido a que el último enlace ascendente se desconecta (excepto si pertenece al *MT*) cuando todos los enlaces de entrada en dirección ascendente ya han sido desconectados (la carga sería nula). La forma de interpretar la figura consiste en seguir las curvas mostradas de acuerdo con las flechas: de derecha a izquierda para carga descendente o de izquierda a derecha para carga ascendente. Por ejemplo, partiendo de una carga que genere una utilización de los enlaces superior a U_{on}, los 4 enlaces ascendentes del switch estarán conectados. Si el tráfico disminuye por debajo de U_{off} el número de enlaces ascendentes conectados pasa sucesivamente de 4 a 3, a 2, y a 1 siguiendo la curva escalonada de la parte superior. Si a continuación el tráfico aumenta hasta el valor original se producirá la reconexión de los enlaces siguiendo la curva de la parte inferior. La distancia horizontal (en utilización) entre las dos curvas representa la sensibilidad (histéresis) instantánea del mecanismo para una situación dada. El área disponible entre las curvas mostradas y una linea horizontal situada en 4 enlaces conectados representa las zonas de funcionamiento de la red donde se produce ahorro de potencia.

El análisis de la figura indica que la reducción en el ancho de banda disponible cuando disminuye el numero de enlaces de salida activos genera un efecto que puede ser visto como una reducción en la banda de histéresis (y consiguiente incremento de sensibilidad) del mecanismo. Este es un efecto positivo, puesto que la red es más sensible a la congestión cuando se reduce la fracción de enlaces activos respecto a las especificaciones nominales. En estas situaciones, interesa tener una sensibilidad mayor porque se incrementará la agilidad del mecanismo para reaccionar contra cambios de menor intensidad en el tráfico, proporcionando enlaces activos adicionales si es necesario. Por el contrario, una sensibilidad pequeña podría provocar congestión en la red durante periodos limitados de tiempo. Considerando que la red habrá sido diseñada en función de la carga máxima, en situaciones con pocos enlaces ac-

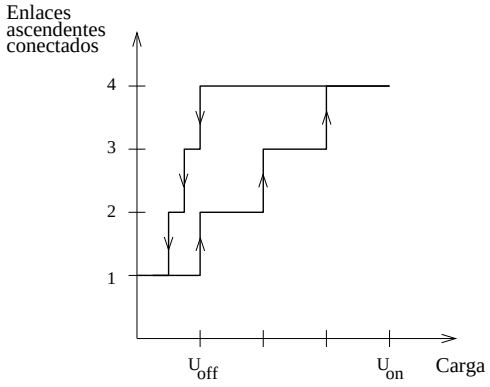


Figura 5.4: Estado del switch en función del tráfico con umbrales estáticos.

tivos y aumentos bruscos de tráfico, es probable que se produjese congestión si la reconexión no se produce con rapidez.

Una limitación importante de los umbrales estáticos es que U_{on} debe ser mayor que $2 \cdot U_{off}$. Esto hace que el promedio de los umbrales sea bajo en situaciones donde hay varios enlaces activos, y por lo tanto hace el mecanismo menos agresivo, lo que implica reducir el margen del ahorro de potencia. Además, esta condición impide el uso de una configuración que sea al mismo tiempo muy agresiva y muy sensible (figura 5.3).

5.3.4. Umbrales dinámicos

Se ha desarrollado una versión alternativa del mecanismo de ahorro de potencia que está basada en umbrales dinámicos. El objetivo es aumentar las zonas de funcionamiento donde se reduce el consumo de potencia. Gráficamente puede visualizarse como conseguir una banda de histéresis estrecha y constante para los umbrales efectivos, intentando acercar la curva seguida al desconectar enlaces con la que se sigue en el proceso de reconexión. Ello incrementaría el margen para reducir el consumo de potencia.

La estrategia consiste en tratar de dividir el rango total de utilización de la red en tantos intervalos como indica la aridad de la red. La intuición que hay debajo de esta propuesta es que cada switch distribuirá el tráfico ascendente entre k enlaces cuando la red esté totalmente activa. Si el tráfico decrece en un factor $\frac{1}{k}$ por debajo del valor nominal de tráfico para un switch totalmente conectado, parece razonable reducir la productividad disponible en exactamente la misma cantidad, es decir, apagando un

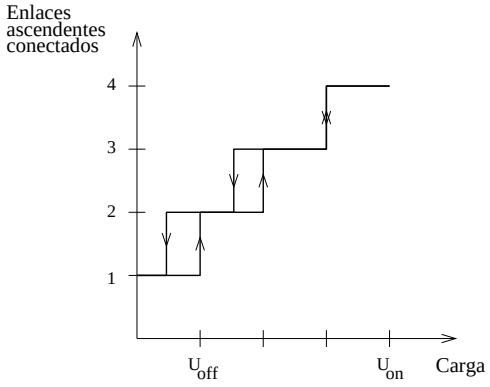


Figura 5.5: Estado del switch en función del tráfico con umbrales dinámicos.

enlace [15].

La implementación de los umbrales dinámicos se basa en un umbral de "conexión" fijo, U_{on} , y una versión dinámica del umbral de "desconexión", U_{off} , que depende del número de enlaces de salida activos de acuerdo a la siguiente expresión:

$$U_{off_i} = \frac{U_{on}(i-1)}{k}$$

siendo i el número de enlaces de salida activos en dirección ascendente. Cada switch de la red mantiene actualizado su umbral de desconexión de acuerdo con la expresión anterior.

El diagrama de transición de estados para un *4-ay n-tree* se muestra en la figura 5.5.

Como conclusión directa, el análisis teórico del comportamiento del mecanismo indica que la variante dinámica de los umbrales proporcionará mejoras significativas en términos de ahorro de potencia, puesto que la porción de utilización de switch donde hay algunos enlaces desconectados aumenta con respecto a la versión estática. Los resultados se validan en la sección siguiente (5.4.5). La superposición de las transiciones entre 3 y 4 enlaces a *on* no es un problema (posible transiciones alternativas múltiples debido a oscilaciones de tráfico) puesto que la operación del mecanismo se basa en la utilización media del switch durante periodos de longitud fija que tienen el efecto de filtrar los cambios de tráfico rápidos. Por último, si consideramos un *4-ary n-tree*, el conjunto de umbrales dinámicos junto con los umbrales efectivos se muestran en la tabla 5.3.

	Umbrales dinámicos			Umbrales efectivos	
Enlaces ascendentes activos	On	Off	Factor	On	Off
4	U_{on}	$3/4U_{on}$	1	U_{on}	$3/4U_{on}$
3	U_{on}	$2/4U_{on}$	$3/4$	$3/4U_{on}$	$6/16U_{on}$
2	U_{on}	$1/4U_{on}$	$2/4$	$2/4U_{on}$	$2/16U_{on}$
1	U_{on}	0	$1/4$	$1/4U_{on}$	0

Tabla 5.3: Valores de los umbrales dinámicos conforme a los enlaces disponibles para un switch *4-ary*.

5.4. Evaluación de prestaciones del mecanismo propuesto

5.4.1. Modelo de red

Nuestro simulador modela una red de interconexión basada en wormhole a nivel de flit [27]. La red se compone de dos tipos de nodos, nodos procesador y nodos *switch* o conmutador. Los switches contienen una unidad de control de encaminamiento, un crossbar y tantos enlaces físicos como indica la aridad de la red. Los canales físicos se dividen en tres canales virtuales, cada canal virtual tiene asociado un buffer con capacidad para cuatro flits. Se utiliza un algoritmo de encaminamiento adaptativo libre de bloqueos que emplea rutas mínimas [26]. Para encaminar un paquete de un nodo fuente a un nodo destino, el paquete es encaminado hacia uno de los ancestros comunes más cercanos, y desde allí al destino. El encaminamiento consta pues de dos fases. En la primera se sigue una ruta adaptativa en dirección ascendente (hacia niveles de valor numérico inferior, véase la figura 5.1) hasta uno de los switches ancestro común más cercanos. En la segunda fase se sigue una ruta determinista (solo existe una) en dirección descendente hasta el destino.

En cada nodo se incluye la implementación del mecanismo de ahorro de potencia propuesto en este trabajo. Los resultados presentados se han obtenido para *fat-trees* cuaternarios de dimensión 4 (256 nodos).

5.4.2. Modelo de tráfico

El patrón de tráfico en la red se define en base a los siguientes parámetros: la distribución espacial de los destinos, la tasa media de inyección de los mensajes, junto con su distribución temporal, y la longitud de los mensajes.

Los experimentos han sido diseñados con el objetivo de presentar situaciones, por

un lado representativas de cargas reales y empleadas en el análisis de prestaciones de redes de interconexión, y por otro lado poco favorables a que el mecanismo propuesto presentara un funcionamiento óptimo. El peor caso lo constituyen cargas que no presentan variaciones temporales ni espaciales como el tráfico uniforme aleatorio (distribución espacial y temporal uniforme) [54]. Ello es debido a que los flujos de tráfico son distribuidos uniformemente por la red a lo largo del tiempo. De esa manera es virtualmente imposible detectar variaciones en la utilización de la red que puedan disparar sistemáticamente el mecanismo de reducción del consumo de potencia y maximicen el ahorro con un mínimo impacto negativo en las prestaciones. Con tráfico intenso es difícil desconectar un enlace y para tráfico ligero prácticamente todos los candidatos a la desconexión serán desconectados, provocando una degradación de las prestaciones de la red. El extremo opuesto sería un patrón de tráfico con una distribución irregular que permitiera tener desconectadas de forma permanente secciones de la red por las que nunca circula tráfico.

Como se ha justificado, con el objeto de evaluar el comportamiento de nuestro mecanismo en las peores condiciones posibles, el modelo de tráfico se basa en una distribución uniforme de destinos. Con respecto a los tiempos de inyección de cada mensaje la mayoría de los experimentos se han realizado con una distribución uniforme (aleatoria) para evaluar el peor caso. Alternativamente, se presentan algunos resultados en los que se emplea una distribución auto-similar debido a la naturaleza auto-similar del tráfico Ethernet [40] tan popular en sistemas tanto domésticos como de altas prestaciones (Gigabit Ethernet representa el 52 % de las tecnologías de red presentes en el TOP500).

Puesto que se ha evaluado nuestro mecanismo tanto en condiciones estáticas como dinámicas, el valor medio de la tasa de inyección es constante para la evaluación estática y variable para el análisis de carga dinámica. Primero se analiza el comportamiento estático de la red ejecutando simulaciones independientes con tasas de inyección de mensajes de media constante, evaluando el rango completo de tráfico, partiendo de una carga baja hasta llegar a la saturación 4.3.5. El segundo grupo de experimentos tiene como objetivo estudiar el comportamiento dinámico de la red, utilizando tasas de inyección variable durante cada simulación 4.3.6.

El tamaño del mensaje se ha fijado a 16 flits, excepto para el análisis del efecto de la longitud de los mensajes, donde se han probado mensajes más largos (256 flits).

5.4.3. Evaluación de las prestaciones básicas de la red

Se presenta en esta sección una evaluación de las prestaciones de la red con topología fat-tree, como referencia para el posterior análisis del comportamiento de la red cuando el mecanismo de ahorro de potencia está activado.

5.4.3.1. Efecto de la función de selección

Como se ha destacado en el capítulo dedicado a la redes directas, un algoritmo de encaminamiento es modelado mediante una función de encaminamiento y una función de selección [28]. La primera suministra un conjunto de canales de salida que permiten llegar a un nodo destino desde el nodo actual. La segunda escoge un canal libre (si es posible) de entre los suministrados por la función de encaminamiento de acuerdo a una determinada estrategia. Si bien la función de encaminamiento determina si un algoritmo de encaminamiento es libre de bloqueos o no, la función de selección solamente tiene efecto sobre las prestaciones.

Hemos considerado relevante incluir este apartado ya que la función de selección determina cómo se distribuye el tráfico entre los canales disponibles y ello puede tener impacto en las posibilidades para desconectar enlaces cuando la carga sea baja. Por ejemplo, una función de selección que concentre todo el tráfico en unos pocos canales físicos permitirá desconectar los otros y por lo tanto reducir el consumo de potencia con más facilidad que en el caso de que el tráfico se distribuya por todos lo canales disponibles.

Las funciones de selección evaluadas han sido dos y los resultados obtenidos para la red en régimen nominal se muestran en la figura 5.6. La primera función de selección, etiquetada *FirstFree* en las gráficas, es determinista: escoge siempre el primer canal virtual libre de entre los proporcionados por la función de encaminamiento; tiende pues a concentrar el tráfico en los canales virtuales de menor índice (y por tanto pertenecientes al árbol mínimo 5.2). Se ha evaluado una variante adaptativa, *FirstFreeAdaptive*, que selecciona el primer canal virtual libre perteneciente al primer canal físico con mayor número de canales virtuales libres. El canal seleccionado cuando todos están ocupados se escoge de forma cíclica entre los ocupados. Tiende, por tanto, a distribuir el tráfico de una manera más uniforme en la red.

Esta diferente distribución del tráfico en la red es la que, en términos de prestaciones, penaliza a la función de selección *FirstFree* frente a la variante adaptativa. Como se observa en la figura 5.6, para baja carga la latencia con la función *FirstFree* es significativamente más elevada porque todo el tráfico tiende a concentrarse

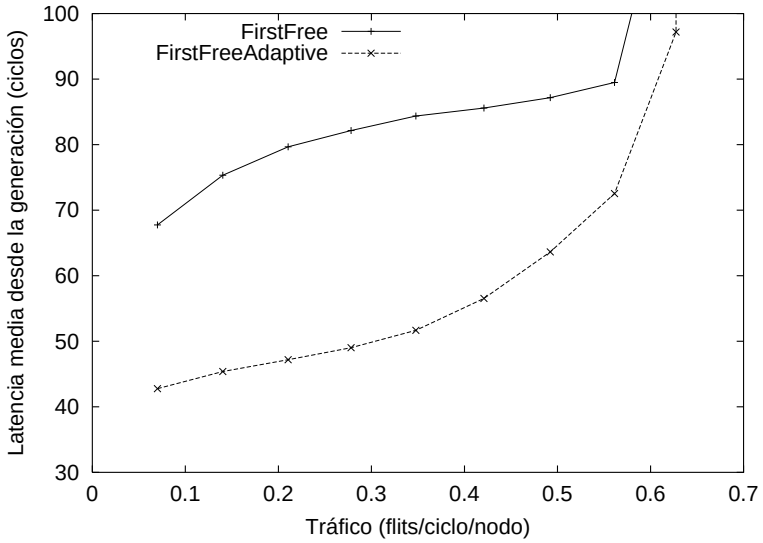


Figura 5.6: Efecto de la función de selección. Latencia media desde la generación frente a tráfico entregado en un 4-ary 4-tree para tráfico uniforme y mensajes de 16 flits.

en unos pocos canales. Solamente cuando el tráfico aumenta, las diferencias tienden a disminuir ligeramente porque se van ocupando todos los canales disponibles en ambos casos. Sin embargo la red alcanza antes la congestión con la función *FirstFree*. Ello se debe a que cuando el tráfico es elevado, la sobrecarga que sufren los primeros canales provoca un estado de saturación de los recursos de red que no ya no se puede resolver aunque se ocupen los canales libres, que están siendo usados con menos frecuencia.

5.4.3.2. Efecto de la longitud de los mensajes

Se han realizado experimentos con mensajes de dos longitudes diferentes, 16 flits y 256 flits. A falta de presentar el impacto de la función de selección en el comportamiento del mecanismo de ahorro de potencia (sección 5.4.4.2), se recogen en esta sección experimentos para las funciones de selección presentadas. La figura 5.7 recoge los resultados obtenidos. Al igual que en los resultados para redes directas, la latencia media por flit es inferior para los mensajes largos debido a que la transmisión de los mensajes se realiza de forma segmentada y el establecimiento de la ruta se amortiza entre más flits cuando los mensajes son largos. Además, los flits de datos

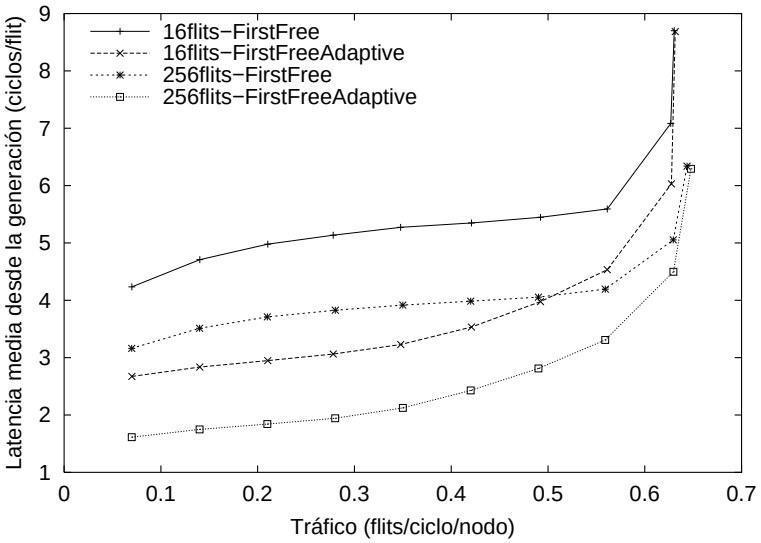


Figura 5.7: Efecto de la longitud de los mensajes. Latencia media por flit frente a tráfico entregado para tráfico uniforme.

avanzan más rápido que los de cabecera porque estos últimos han de ser encaminados, esperando que la unidad de encaminamiento de cada nodo calcule el canal de salida, y posiblemente esperando a que dicho canal esté libre. Así, cuando la cabecera alcanza el destino los flits de datos avanzan más rápido, favoreciendo a los mensajes más largos.

5.4.3.3. Prestaciones del Árbol Mínimo

El Árbol Mínimo proporciona la mínima infraestructura de red que garantiza la conectividad entre todos los nodos. Define, por lo tanto, el limite inferior del consumo de potencia. Para baja carga, representa el estado al que tiende la red cuando se encuentra operativo el mecanismo propuesto de desconexión de enlaces. Con el objeto de ofrecer información sobre la red en dichas condiciones, se han evaluado las prestaciones del Árbol Mínimo en condiciones nominales, para tráfico uniforme y mensajes de 16 bits (figura 5.8). La carga máxima admisible para la configuración evaluada (4-ary 4-tree) se sitúa en un valor extremadamente bajo (alrededor de 0,015flits/ciclo/nodo) cuando la comparamos con el fat-tree completo. Ello es debido a la reducción del ancho de banda disponible que presenta un árbol a medida que se debe ascender más niveles en la red para alcanzar un destino. Como consecuencia,

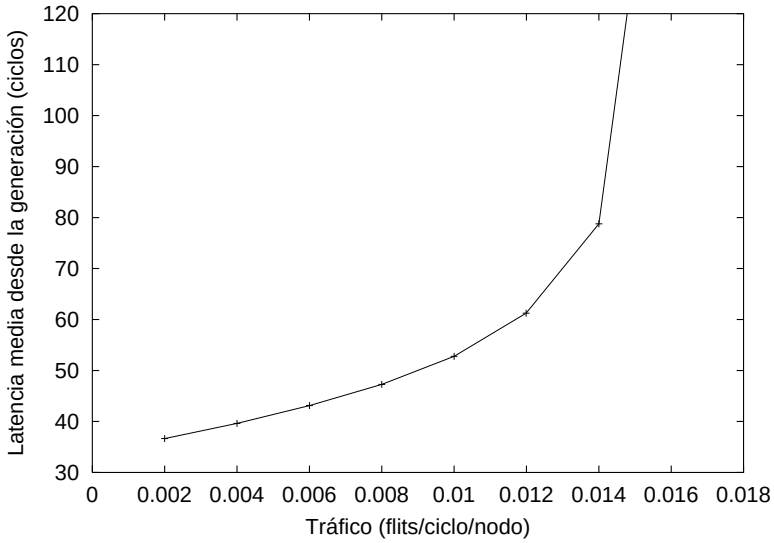


Figura 5.8: Latencia media desde la generación frente a tráfico entregado en el Árbol Mínimo de un 4-ary 4-tree para tráfico uniforme y mensajes de 16 flits.

cuando el tráfico aumenta, rápidamente se congestiona la raíz del árbol y se colapsa la capacidad de la red de absorber más tráfico.

5.4.4. Evaluación estática

En esta sección se analizan las prestaciones de la red con el mecanismo de reducción del consumo de potencia en funcionamiento. Se realiza una evaluación estática del mismo, es decir, en condiciones de tráfico constantes. Los resultados presentados han sido obtenidos para configuraciones en las que todos los enlaces se encuentran conectados en el instante inicial. Los experimentos han sido diseñados para que se intercambien 500.000 mensajes para una tasa de inyección media constante. Dicha tasa de inyección se ha modificado para explorar situaciones desde baja carga hasta la saturación. Como se ha descrito, el comportamiento del mecanismo de ahorro de potencia depende de los umbrales de conexión y desconexión de enlaces, U_{on} y U_{off}, respectivamente. Es por ello que se han explorado diferentes configuraciones compatibles con el mapa de umbrales posibles definido en la sección 5.3.2.

El cálculo de los umbrales objeto de esta evaluación se ha realizado buscando una distribución regular de puntos que proporcione un abanico representativo de configuraciones. De esta forma, se mostrarán diferentes ajustes de sensibilidad y agre-

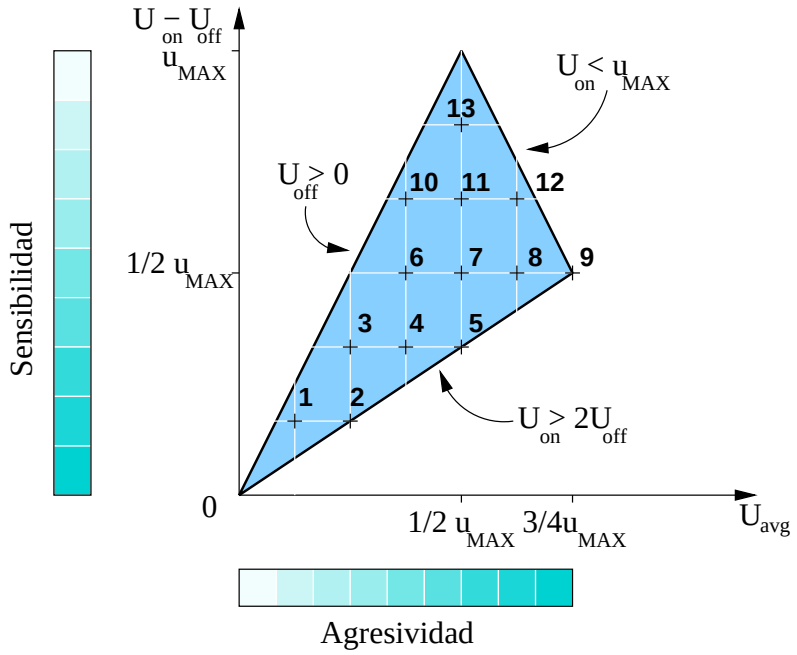


Figura 5.9: Mapa de umbrales posibles con indicación de los puntos evaluados.

sividad del mecanismo que tendrán impacto en el comportamiento del mismo. Los puntos seleccionados aparecen resaltados en la figura 5.9. Tal como se muestra en la figura, la distribución de puntos, y por tanto los valores concretos de los umbrales, dependen del parámetro u_{MAX} o máxima utilización de la red. De acuerdo con los resultados presentados en la figura 5.6, el valor de productividad máxima para la topología evaluada es aproximadamente de $u_{MAX} = 0{,}63\ flits/ciclo/nodo$. A partir de este resultado, y de acuerdo con el mapa de posibles umbrales de la figura 5.9, se han calculado los valores de los umbrales de referencia evaluados, que se muestran en la tabla 5.4. El conjunto de valores se ha calculado distribuyéndolos uniformemente, referenciados respecto de su histéresis o sensibilidad ($U_{on} - U_{off}$) y su media o agresividad ($(U_{on}+U_{off})/2$). La tabla 5.4 muestra los valores correspondientes al umbral de conexión (U_{on}) en la parte superior, y al umbral de desconexión (U_{off}) en la parte inferior. Se han sombreado las trece parejas de umbrales que cumplen con las restricciones. También se han incluido puntos fuera del área recomendada con el objeto de evaluar el comportamiento del mecanismo cuando no se cumple la restricción de ausencia de ciclos de conexión/desconexión.

En las figuras siguientes se presentan las curvas de latencia media desde la gene-

Histéresis	U_{on}	Media					
		0,07875	0,1575	0,23625	0,315	0,39375	0,4725
	0,105	**0,13125**	**0,21**	0,28875	0,3675	0,44625	0,525
	0,21		**0,2625**	**0,34125**	**0,42**	0,49875	0,5775
	0,315			**0,39375**	**0,4725**	**0,55125**	**0,63**
	0,42			**0,44625**	**0,525**	**0,60375**	
	0,525				**0,5775**		

Histéresis	U_{off}	Media					
		0,07875	0,1575	0,23625	0,315	0,39375	0,4725
	0,105	**0,02625**	**0,105**	0,18375	0,2625	0,34125	0,42
	0,21		**0,0525**	**0,13125**	**0,21**	0,28875	0,3675
	0,315			**0,07875**	**0,1575**	**0,23625**	**0,315**
	0,42			**0,02625**	**0,105**	**0,18375**	
	0,525				**0,0525**		

Tabla 5.4: Umbrales de test empleados en la evaluación con $u_{MAX} = 0{,}63$.

ración de los mensajes y las curva de potencia relativa consumida por los enlaces frente al tráfico entregado (figuras 5.10 y 5.11). Las curvas corresponden a la variante estática de los umbrales y están identificadas por etiquetas con el formato "$U_{off} - U_{on}$", excepto la curva etiquetada "*Nominal*", que representa la latencia de los mensajes cuando el mecanismo de ahorro de potencia no actúa. Con el fin de presentar los resultados de una manera ordenada, éstos se han agrupado para que cada gráfica muestre curvas correspondientes a configuraciones con sensibilidad (histéresis) constante. Es decir, se agrupan en la misma gráfica resultados para una fila de la matriz de puntos de test representada sobre la figura 5.9, que a su vez se corresponden a una fila de la tabla 5.4. Adicionalmente, las etiquetas se ordenan de arriba a abajo en orden creciente de agresividad. Los resultados mostrados corresponden a la función de selección *FirstFreeAdaptive*, la cual proporciona los mejores resultados de prestaciones (en la sección 5.4.4.2 se hace un análisis comparativo). Los resultados muestran, en todos los casos, significativos descensos del consumo de potencia para baja carga con un incremento muy moderado en la latencia de los mensajes.

Se observa que el valor de los umbrales tiene un impacto significativo en el consumo de potencia. Los umbrales más agresivos consiguen mejores ahorros de potencia. El precio a pagar es un incremento de la latencia para cargas bajas debido a la reducción del número de enlaces disponibles en la red por la desconexión de los mismos. El impacto más significativo en latencia (y en ahorro de potencia) se produce para la configuración más agresiva ($U_{off} = 0{,}315$, $U_{on} = 0{,}630$), que se corresponde con

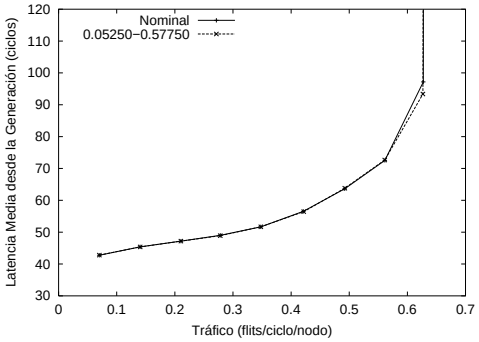
el punto 9 del mapa de umbrales posibles (figura 5.9). En este caso, se ahorra potencia para tráfico por debajo de aproximadamente 0,34 $flits/ciclo/switch$. Cuando el tráfico en la red aumenta, la progresiva reconexión de enlaces hace que la latencia tienda hacia su valor nominal.

Si se evalúa el producto $L_{rel} \times P_{rel}$ (figura 5.12) se constata que en todas las configuraciones probadas se obtienen resultados favorables del mecanismo. Se observa que las configuraciones más agresivas proporcionan un amplio margen para el ahorro de potencia, representado por los valores del producto $L_{rel} \times P_{rel}$ situados por debajo de la unidad.

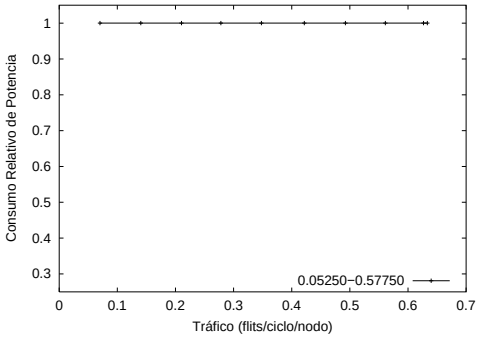
Umbrales dinámicos A continuación se presentan los resultados obtenidos con idéntica configuración pero empleando los umbrales dinámicos descritos en la sección 5.3.4. Los experimentos se han realizado empleando como umbrales de conexión iniciales (U_{on}) los mapeados en la figura 5.9. Las curvas están identificadas en este caso por etiquetas con el formato "U_{on}", excepto la curva etiquetada "*Nominal*", que representa la latencia de los mensajes cuando el mecanismo de ahorro de potencia no actúa. En este caso no tiene sentido hacer referencia a una etiqueta U_{off}, dado que los umbrales para cada conmutador se ajustan en función de su estado. Se ha mantenido el formato de presentación de los resultados para permitir una comparación sencilla respecto a los presentados para los umbrales estáticos con los mismos valores para U_{on}.

Los resultados muestran un mayor potencial para el ahorro de potencia que el mecanismo basado en umbrales estáticos. Ello se constata por una ampliación de los niveles de tráfico para los cuales la potencia consumida se mantiene por debajo de la nominal. Esta mejora en potencia se consigue a cambio de ligeros incrementos de latencia para niveles de tráfico por debajo de 0,45 $flits/ciclo/switch$. Además, el aumento de latencia se manifiesta como un desplazamiento hacia arriba de la curva con bajas cargas, sin picos significativos. El resultado final es que la curva de latencia para baja potencia muestra un comportamiento de la red similar al que se obtiene con potencia nominal. Por otro lado, con los umbrales estáticos el impacto del mecanismo en el consumo de potencia es mucho menor y las curvas obtenidas muestran un comportamiento más variable (manifestado en forma de picos más pronunciados) en la evolución de la latencia.

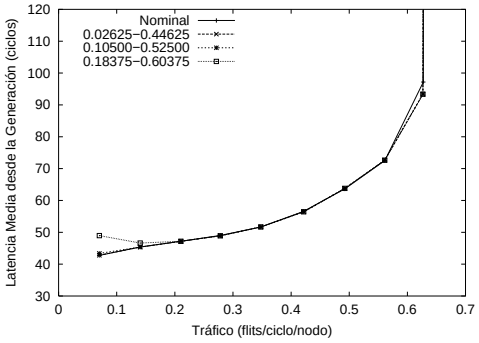
El análisis del balance latencia/potencia a través del factor $L_{rel} \times P_{rel}$ (figura 5.15) indica que para las configuraciones más agresivas, con umbrales dinámicos, existe una penalización. Para los tests realizados cuando el umbral U_{on} es superior a 0,47



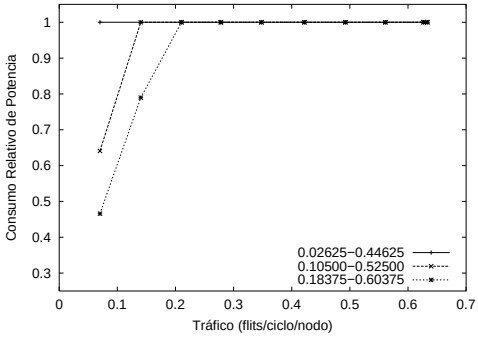
(a) Latencia con histéresis de 0,525.



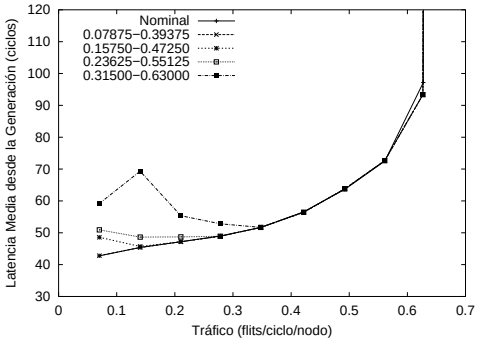
(b) Potencia con histéresis de 0,525.



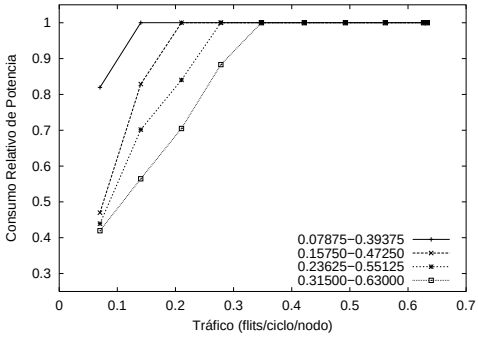
(c) Latencia con histéresis de 0,42.



(d) Potencia con histéresis de 0,42.

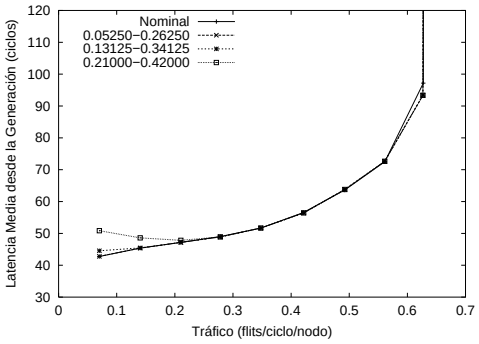


(e) Latencia con histéresis de 0,315.

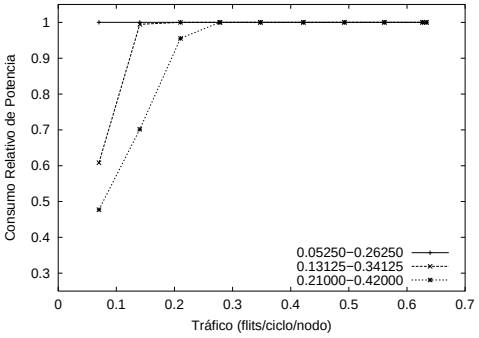


(f) Potencia con histéresis de 0,315.

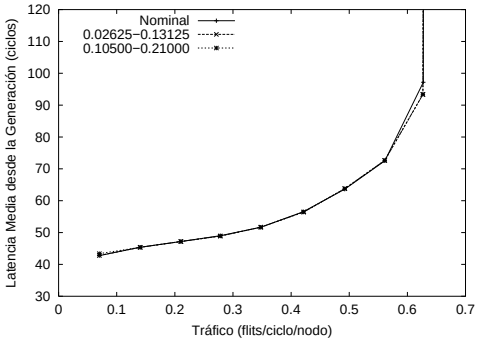
Figura 5.10: Resultados con umbrales estáticos (primera parte).



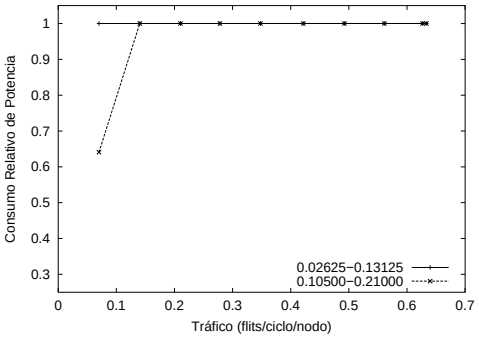
(a) Latencia con histéresis de 0,21.



(b) Potencia con histéresis de 0,21.



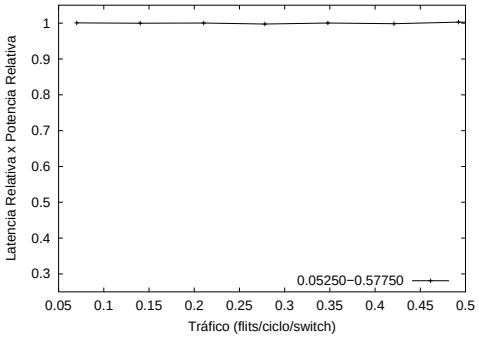
(c) Latencia con histéresis de 0,105.



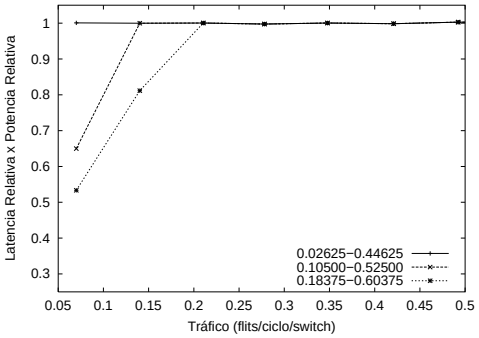
(d) Potencia con histéresis de 0,105.

Figura 5.11: Resultados con umbrales estáticos (segunda parte).

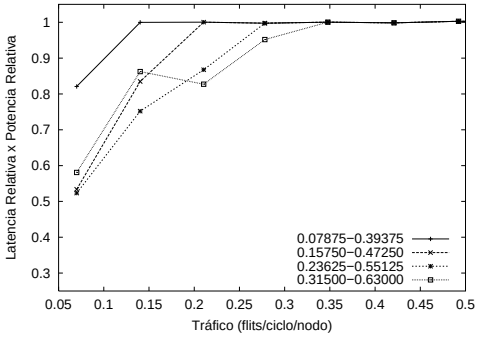
se obtienen resultados de $L_{rel} \times P_{rel}$ superiores a la unidad (el incremento en la latencia de los mensajes supera el ahorro de potencia relativa). Para valores inferiores del umbral de conexión, se obtienen resultados favorables en todos los casos y existe un amplio margen para definir políticas de ahorro más o menos agresivas. La figura 5.16 reúne en una sola gráfica algunas configuraciones seleccionadas que ilustran el comportamiento del mecanismo para valores crecientes del umbral de conexión. La configuración que proporciona resultados favorables para un mayor margen de tráficos es $U_{on} = 0{,}4725$ pero con resultados muy cercanos a $U_{on} = 0{,}42$. Políticas más agresivas (figura 5.15) no proporcionan mejoras y en cambio generan penalizaciones para tráfico superior a 0,35 $flits/ciclo/switch$. De acuerdo con los resultados obtenidos, el umbral de conexión más agresivo recomendado se situaría en valores alrededor de 0,45.



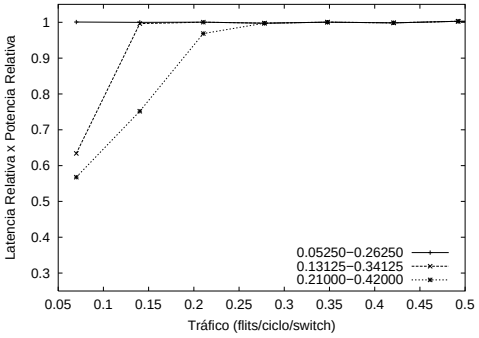
(a) $L_{rel} \times P_{rel}$ con histéresis de 0,525.



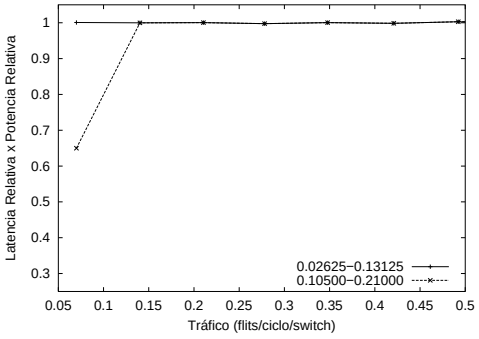
(b) $L_{rel} \times P_{rel}$ con histéresis de 0,42.



(c) $L_{rel} \times P_{rel}$ con histéresis de 0,315.



(d) $L_{rel} \times P_{rel}$ con histéresis de 0,21.



(e) $L_{rel} \times P_{rel}$ con histéresis de 0,105.

Figura 5.12: Resultados umbrales estáticos.

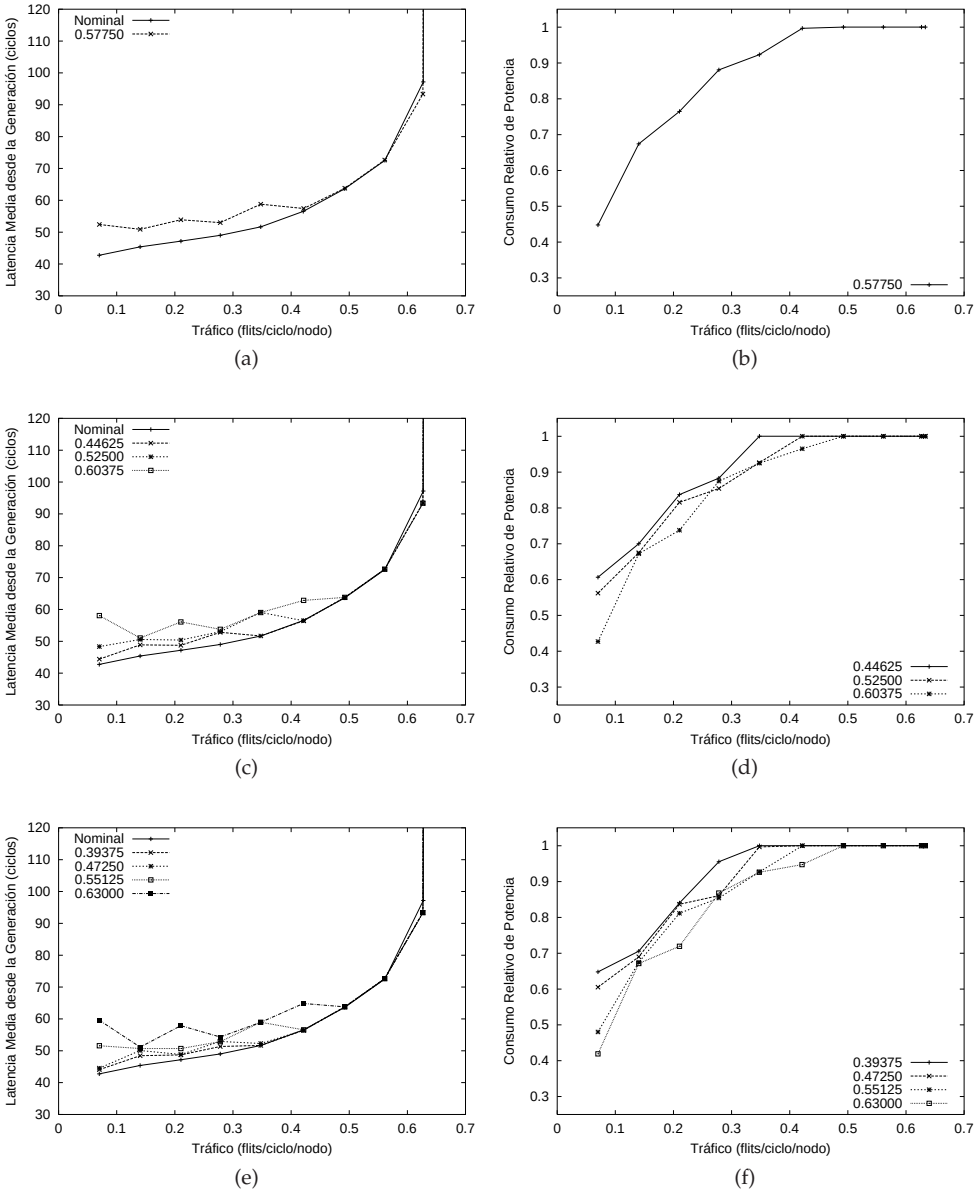


Figura 5.13: Resultados con umbrales dinámicos (primera parte).

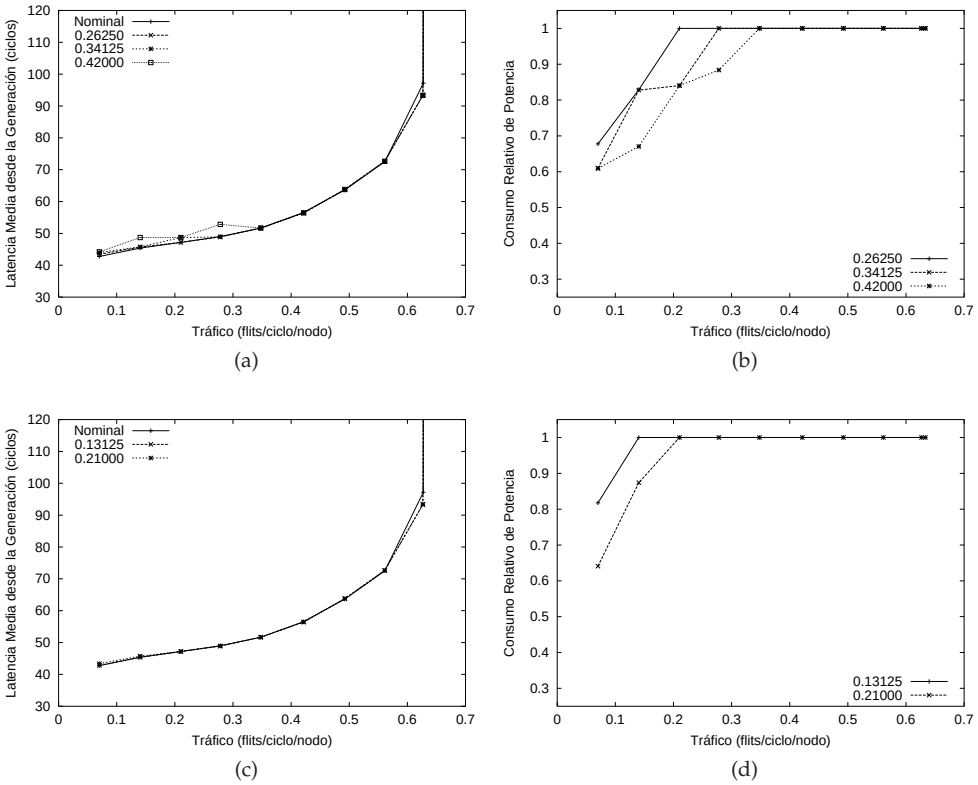


Figura 5.14: Resultados con umbrales dinámicos (segunda parte).

5.4.4.1. Efecto de la longitud de los mensajes

Con el objetivo de estudiar la influencia de la longitud de los mensajes sobre el comportamiento del mecanismo, se ha realizado una evaluación del mismo con mensajes largos de 256 flits. Se han repetido los experimentos presentados en la sección anterior empleando dicho tamaño de mensaje. Los resultados se presentan en las figuras 5.17 y 5.18.

Al igual que en la sección anterior, las curvas corresponden a la variante estática de los umbrales y están identificadas por etiquetas con el formato "$U_{off} - U_{on}$", excepto la curva etiquetada "*Nominal*", que representa la latencia de los mensajes cuando el mecanismo de ahorro de potencia no actúa. Los resultados se han agrupado para que cada gráfica muestre curvas correspondientes a configuraciones con sensibilidad (histéresis) constante. Las etiquetas se ordenan de arriba a abajo en orden creciente

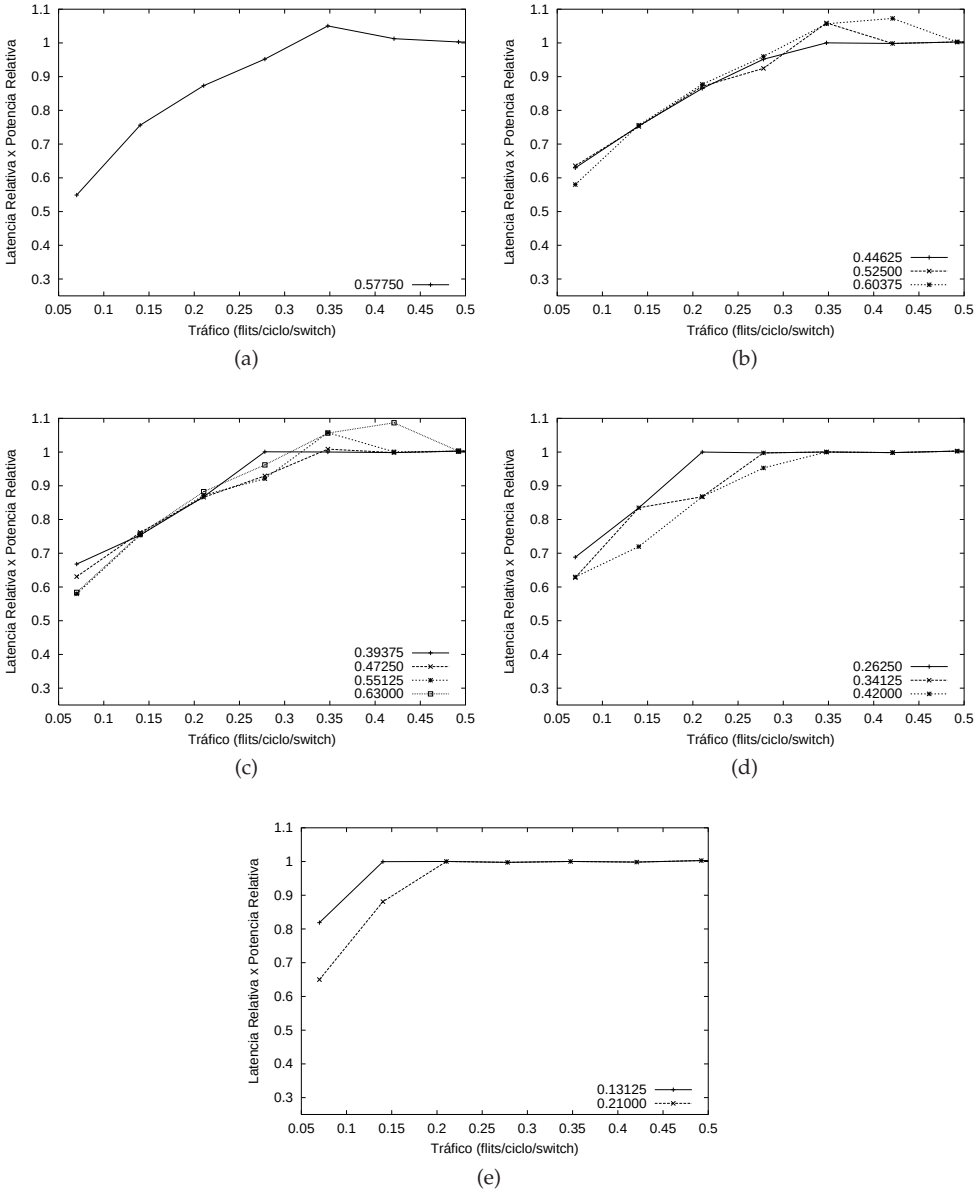


Figura 5.15: Resultados con umbrales dinámicos.

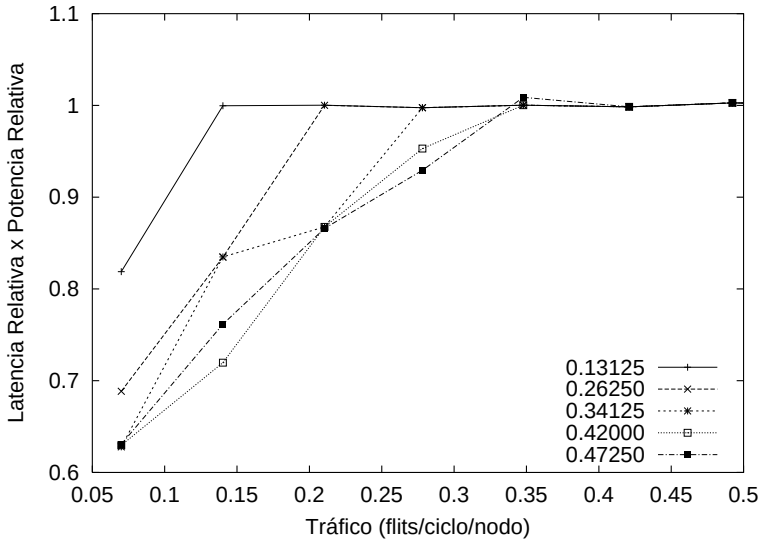
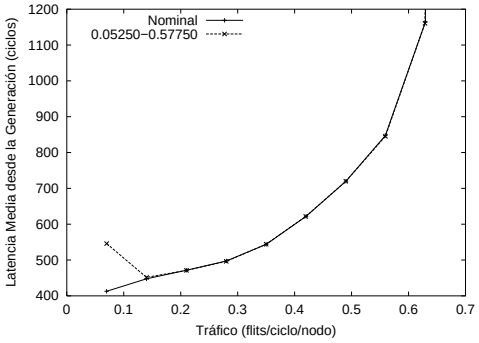


Figura 5.16: Producto $L_{rel} \times P_{rel}$ para configuraciones favorables con umbrales dinámicos.

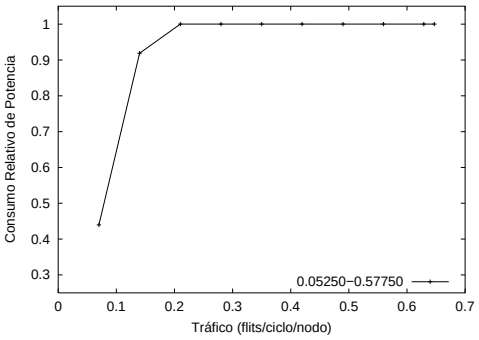
de agresividad. Se muestran resultados para la función de selección que permite obtener mayores prestaciones *(FirstFreeAdaptive*).

Los resultados muestran, en todos los casos, resultados equivalentes a los obtenidos para mensajes cortos pero con mejores tasas de ahorro de potencia para baja carga, con un incremento muy moderado en la latencia de los mensajes. En el mejor caso para mensajes de 16 flits se genera ahorro de potencia cuando la carga es inferior al 33 % de la carga de saturación, mientras que con 256 flits este valor es del 55 %. Al igual que en el caso de las redes directas, la distribución del trafico en mensajes de mayor tamaño reduce la sobrecarga debida a las cabeceras y permite reducir la utilización efectiva de los enlaces y por tanto conseguir un ahorro de potencia adicional; es decir, con mensajes más largos hay menos encaminamientos, por lo que hay menos oportunidades de escoger otros canales y al estar el tráfico más agrupado existen mayores posibilidades para el ahorro.

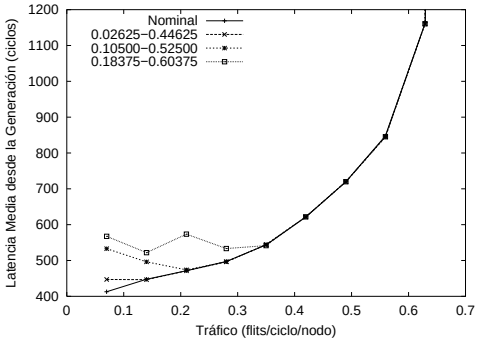
Los resultados obtenidos para la métrica $L_{rel} \times P_{rel}$ (figura 5.19) confirman que la mayor penalización observada en la latencia es ampliamente compensada por la reducción del consumo de potencia, al igual que sucede con mensajes de 16 flits. Solamente una de las configuraciones del mecanismo, la más agresiva, proporciona valores del producto $L_{rel} \times P_{rel}$ superiores a la unidad (el valor máximo medido es de 1,04). Esta configuración presenta un valor de $U_{on} = 0{,}630$ que es la máxima uti-



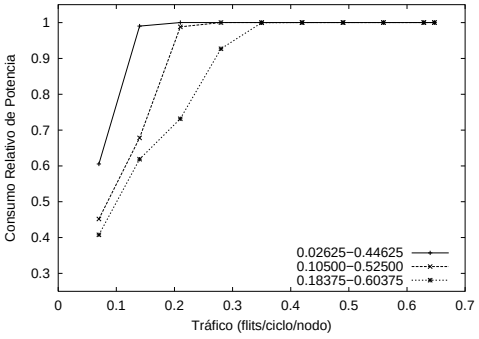
(a) Latencia con histéresis de 0,525.



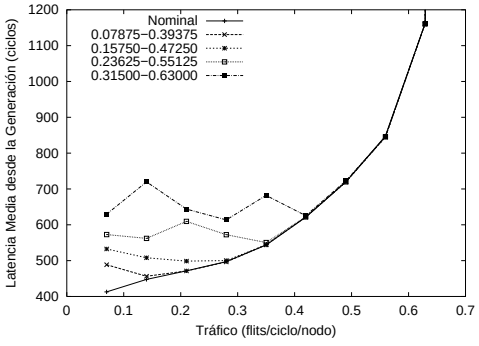
(b) Potencia con histéresis de 0,525.



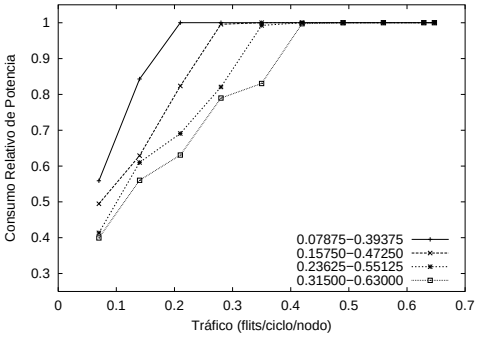
(c) Latencia con histéresis de 0,42.



(d) Potencia con histéresis de 0,42.

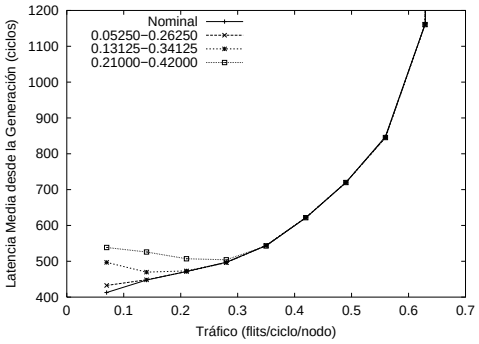


(e) Latencia con histéresis de 0,315.

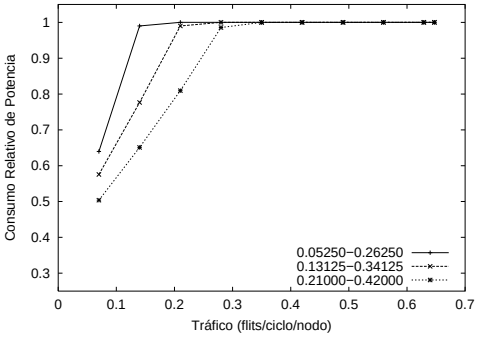


(f) Potencia con histéresis de 0,315.

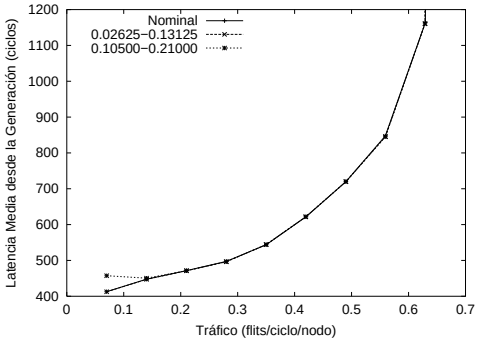
Figura 5.17: Resultados con umbrales estáticos y mensajes de 256 flits (primera parte).



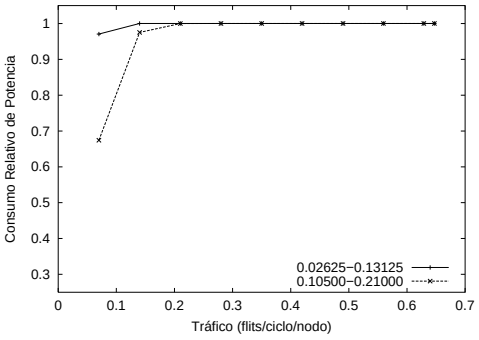
(a) Latencia con histéresis de 0,21.



(b) Potencia con histéresis de 0,21.



(c) Latencia con histéresis de 0,105.

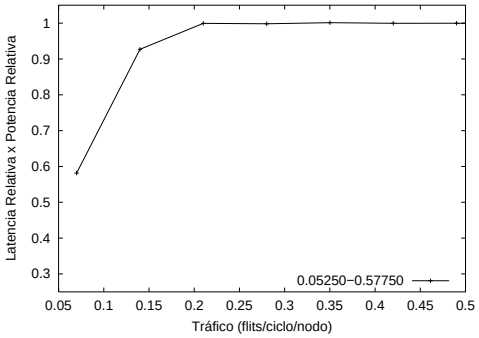


(d) Potencia con histéresis de 0,105.

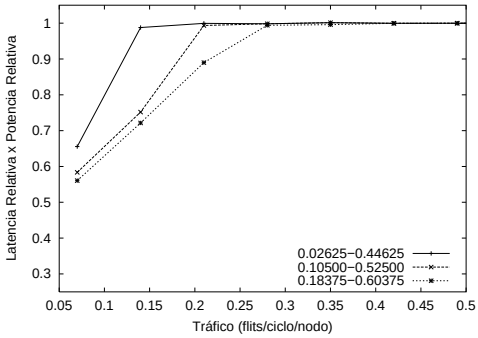
Figura 5.18: Resultados con umbrales estáticos y mensajes de 256 flits (segunda parte).

lización posible de la red y $U_{off} = 0{,}315$, con lo que $U_{on} = 2 \times U_{off}$ y son posibles los ciclos de conexión desconexión. Es decir, está en el límite de lo que hemos definido como aceptable. Cualquiera de las demás configuraciones probadas proporciona resultados favorables para este indicador.

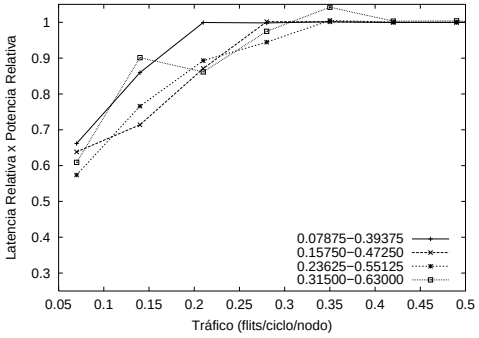
Umbrales dinámicos En este apartado se presentan los resultados para las mismas configuraciones pero con el mecanismo basado en umbrales dinámicos. Las figuras 5.20 y 5.21 contienen los resultas de latencia y potencia relativa para todos los umbrales probados. Al igual que con mensajes cortos, se constata un mayor potencial para el ahorro de potencia que con el mecanismo basado en umbrales estáticos. Alguna configuración permite conseguir ahorro de potencia para niveles de tráfico de hasta 0,50 $flits/ciclo/nodo$.



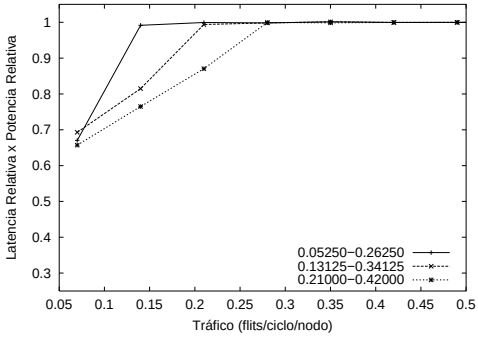
(a) $L_{rel} \times P_{rel}$ con histéresis de 0,525.



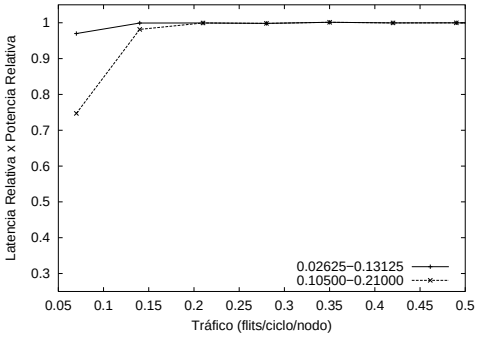
(b) $L_{rel} \times P_{rel}$ con histéresis de 0,42.



(c) $L_{rel} \times P_{rel}$ con histéresis de 0,315.



(d) $L_{rel} \times P_{rel}$ con histéresis de 0,21.



(e) $L_{rel} \times P_{rel}$ con histéresis de 0,105.

Figura 5.19: Resultados con umbrales estáticos y 256 flits.

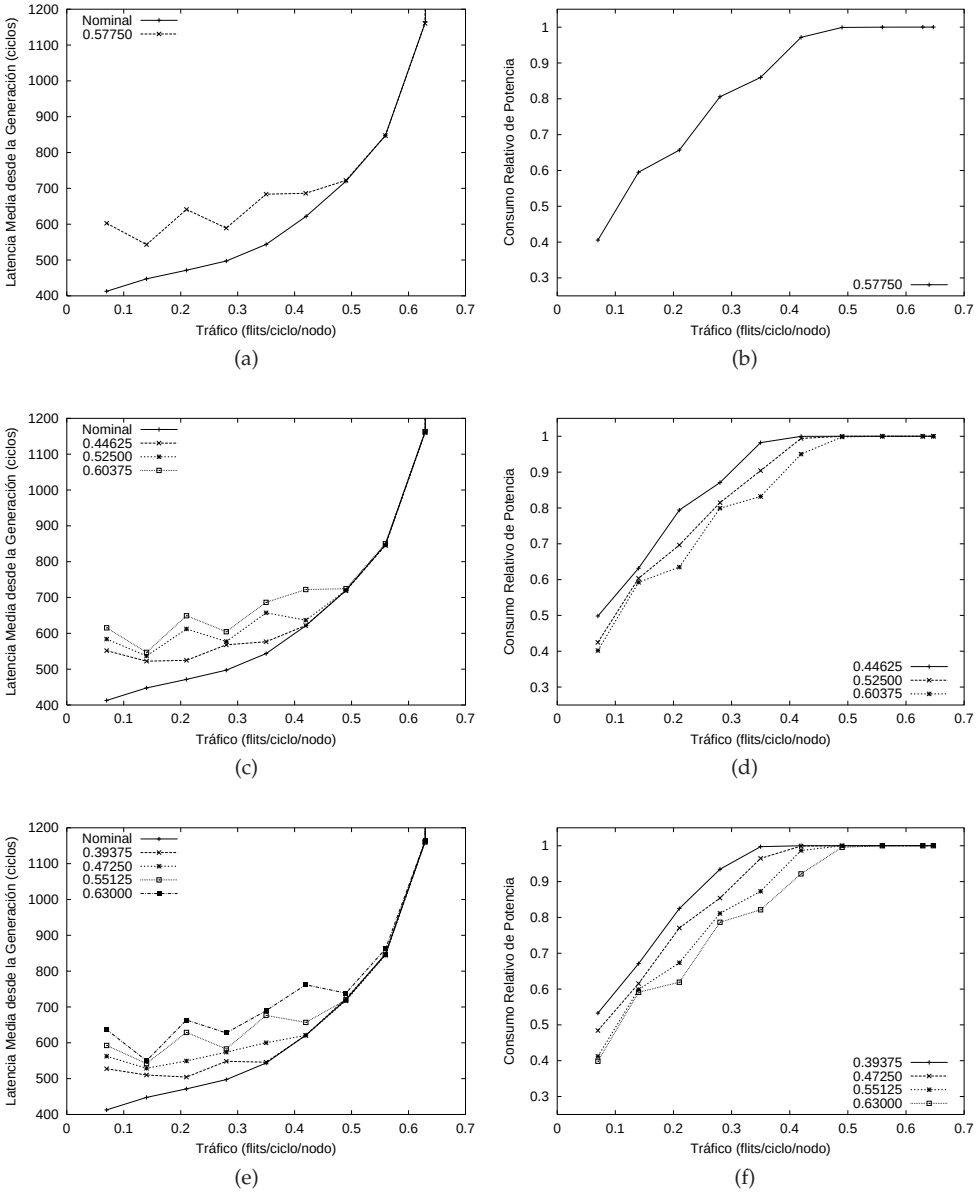


Figura 5.20: Resultados con umbrales dinámicos (primera parte).

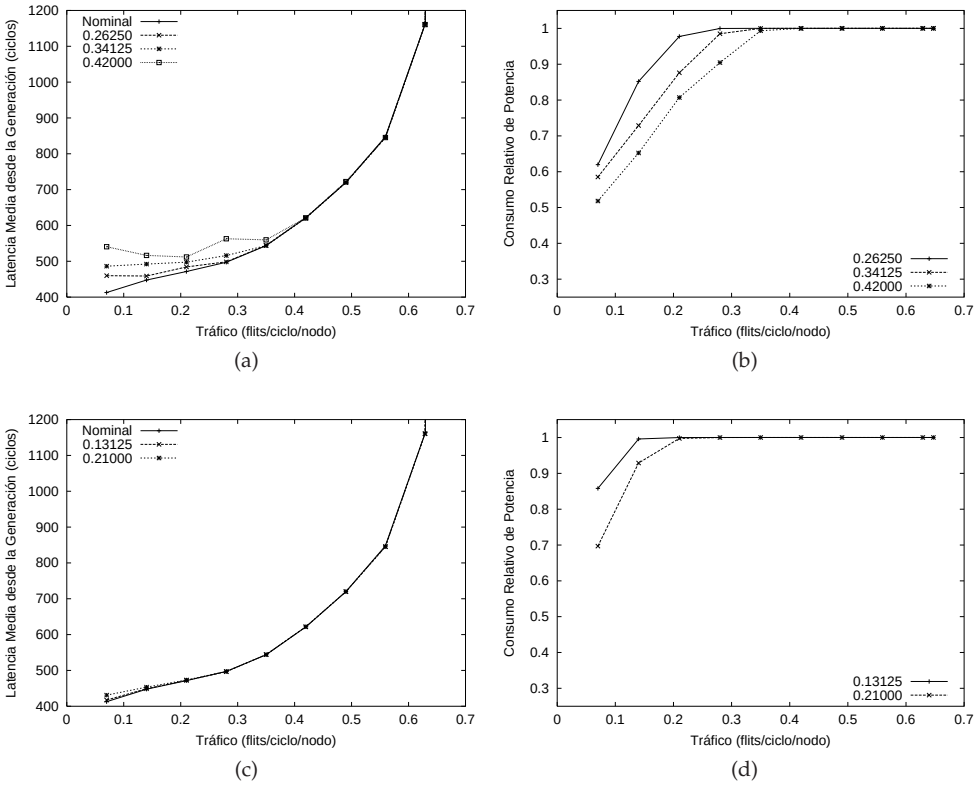


Figura 5.21: Resultados con umbrales dinámicos (segunda parte).

No obstante, el balance $L_{rel} \times P_{rel}$ (figura 5.22) indica que, para las configuraciones más agresivas, con umbrales dinámicos, existe una penalización. Para los tests realizados cuando el umbral U_{on} es superior a 0,34 se obtienen resultados de $L_{rel} \times P_{rel}$ superiores a la unidad (el incremento en la latencia de los mensajes supera el ahorro de potencia relativa). Para valores inferiores del umbral de conexión, se obtienen resultados favorables en todos los casos y existe un amplio margen para definir políticas de ahorro más o menos agresivas. La figura 5.23 reúne en una sola gráfica algunas configuraciones seleccionadas que ilustran el comportamiento del mecanismo para valores crecientes del umbral de conexión. La configuración que proporciona resultados razonablemente favorables para un mayor margen de tráficos es $U_{on} = 0{,}42$, pero con resultados muy cercanos a $U_{on} = 0{,}34125$ (en estos casos hay una ligera penalización para tráfico entre 0,3 $flits/ciclo/switch$ y $0,35$ $flits/ciclo/switch$. Políticas más agresivas (figura 5.22) no proporcionan mejoras y en cambio generan penaliza-

ciones para tráfico superior a 0,30 $flits/ciclo/switch$. De acuerdo con los resultados obtenidos, el umbral de conexión más agresivo recomendado se situaría en valores alrededor de 0,35 en este caso, ligeramente por debajo del máximo admisible con mensajes más cortos.

5.4.4.2. Efecto de la función de selección

De acuerdo con los resultados presentados en la sección 5.4.3.1, la función de selección que proporciona las mejores oportunidades para ahorrar potencia, *FirstFree*, es la que ofrece peores prestaciones de la red. No obstante, se ha realizado un análisis experimental del mecanismo de ahorro de potencia cuando se utiliza esta función de selección. El objetivo ha sido constatar si el ahorro de potencia que se consigue compensa los peores resultados de latencia de partida. Por esta razón, la latencia relativa se ha calculado tomando como referencia los resultados obtenidos con la función de selección *FirstFreeAdaptive,* que es la proporciona menores latencias. Se han evaluado todas las configuraciones de umbrales presentadas en los experimentos anteriores y se resumen los resultados por medio de las representación gráfica del indicador $L_{rel} \times P_{rel}$ en función del tráfico para algunas configuraciones seleccionadas. Los resultados, recogidos en las figuras 5.24 y 5.25, indican que solamente para algunas configuraciones de umbrales y con muy baja carga, inferior a 0,15 $flits/ciclo/nodo$, se obtienen valores favorables del indicador con la función de selección *FirstFree*. Esto sucede tanto para umbrales estáticos como para umbrales dinámicos. A la vista de estos resultados concluimos que, en condiciones estáticas de carga, la función de selección *FirstFreeAdaptive* proporciona las mejores prestaciones cuando se aplica nuestra propuesta de reducción del consumo de potencia. Idénticos resultados se han obtenido al evaluar la función de selección sobre mensajes de 256 flits.

5.4.4.3. Energía consumida

La evaluación estática presentada se ha completado con medidas de energía para proporcionar evidencias adicionales sobre el impacto de la estrategia propuesta. El consumo de energía en la red se ha evaluado usando un modelo de carga cerrado. La simulación se configura para que se intercambie un número fijo de mensajes, en particular 500.000. Existen dependencias entre los mensajes, de modo que, cuando un nodo recibe un mensaje, responde con otro. De este modo la latencia de los mensajes tiene influencia sobre el tiempo de generación de los mismos, a diferencia del modelo abierto de carga anterior. Es posible controlar la carga de trabajo cambiando el

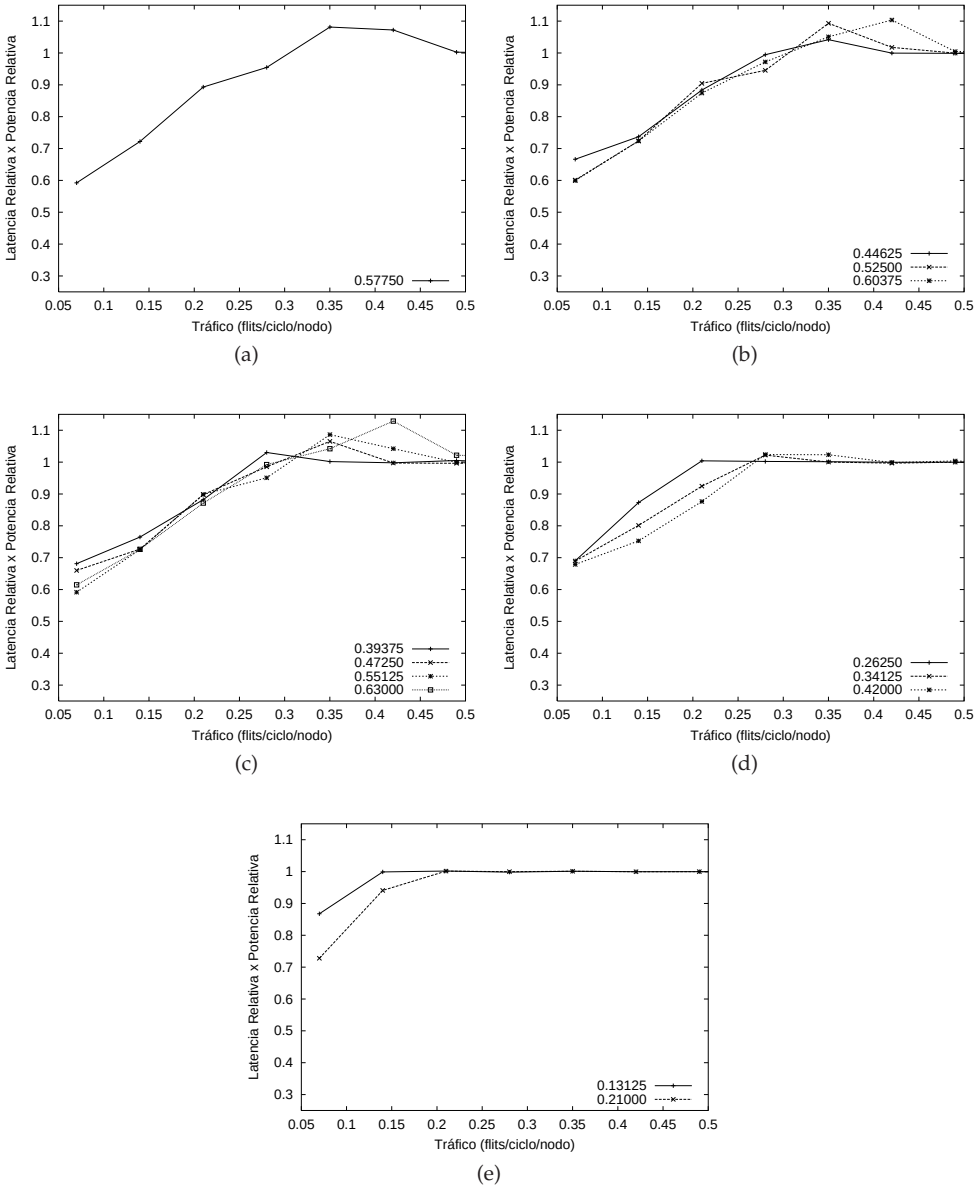


Figura 5.22: Resultados con umbrales dinámicos y mensajes de 256 flits.

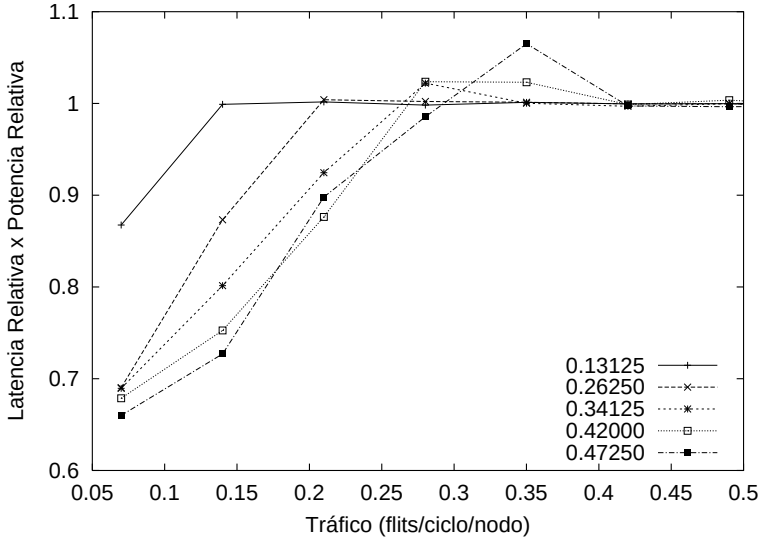
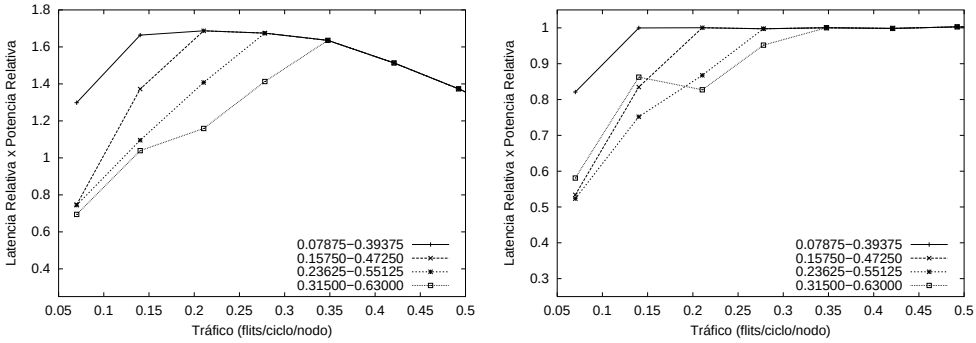


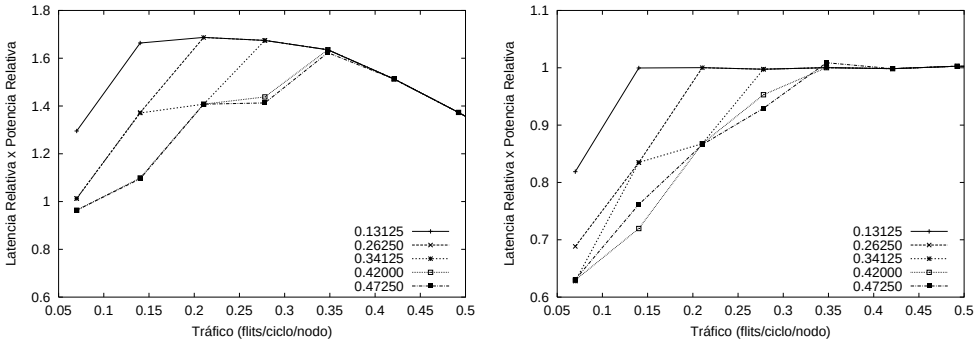
Figura 5.23: Producto $L_{rel} \times P_{rel}$ para configuraciones favorables con umbrales dinámicos y mensajes de 256 flits.



(a) $L_{rel} \times P_{rel}$ con histéresis de 0,315. Función de selección *FirstFree* y umbrales estáticos.

(b) $L_{rel} \times P_{rel}$ con histéresis de 0,315. Función de selección *FirsFreeAdaptive* y umbrales estáticos.

Figura 5.24: Efecto de la función de selección para umbrales estáticos.



(a) $L_{rel} \times P_{rel}$ con función de selección *FirstFree* y umbrales dinámicos.

(b) $L_{rel} \times P_{rel}$ con función de selección *FirsFreeAdaptive* y umbrales dinámicos.

Figura 5.25: Efecto de la función de selección para umbrales dinámicos.

Nodos activos	T Nominal (ciclos)	$T(ciclos)$	Power Aware $\triangle T$	$Potencia$	$\triangle Energía$
10 %	*829275*	*917576*	*10.65 %*	*37.78 %*	*-46.08 %*
20 %	*426181*	*484425*	*13.67 %*	*44.51 %*	*-38.31 %*
30 %	*293784*	*317378*	*8.03 %*	*61.35 %*	*-26.37 %*
40 %	*224031*	*238939*	*6.65 %*	*68.14 %*	*-21.35 %*
50 %	*183735*	*188027*	*2.34 %*	*85.29 %*	*-10.07 %*
60 %	*155652*	*155930*	*0.18 %*	*99.36 %*	*-0.35 %*
70 %	*135697*	*135697*	*0 %*	*100 %*	*0 %*
80 %	*121729*	*121729*	*0 %*	*100 %*	*0 %*
90 %	*110223*	*110223*	*0 %*	*100 %*	*0 %*
100 %	*101183*	*101183*	*0 %*	*100 %*	*0 %*

Tabla 5.5: Impacto del consumo de energía para umbrales estáticos.

número de nodos activos en la red. Para cada simulación, se mide el tiempo necesario para entregar los mensajes y la potencia consumida por los enlaces. Se muestran resultados obtenidos para umbrales estáticos, $U_{\text{off}} = 0{,}1575$ y $U_{on} = 0{,}4725$ (tabla 5.5), y para umbrales dinámicos, $U_{on} = 0{,}4725$ (tabla 5.6), con cargas de red (número de nodos generando mensajes) entre el 10 % y el 100 %. la configuración presentada corresponde al punto 7 del mapa de posibles umbrales (5.9).

La tabla muestra en la segunda columna (TNominal) el tiempo de ejecución para una red sin mecanismo de ahorro de potencia. El tiempo de ejecución, que corresponde al tiempo que necesita la red para entregar todos los mensajes, se alarga cuando este mecanismo está activado (tercera y cuarta columnas de la tabla 4.3, T, $\triangle T$), ya

Nodos activos	T Nominal (ciclos)	$T(ciclos)$	Power $\triangle T$	Aware *Potencia*	$\triangle$*Energía*
10 %	*829275*	*927982*	*11.90 %*	*42.54 %*	*-41.08 %*
20 %	*426181*	*499542*	*17.21 %*	*41.91 %*	*-38.89 %*
30 %	*293784*	*314330*	*6.99 %*	*61.74 %*	*-26.74 %*
40 %	*224031*	*247215*	*10.34 %*	*69.86 %*	*-17.06 %*
50 %	*183735*	*202670*	*10.30 %*	*68.83 %*	*-18.03 %*
60 %	*155652*	*173772*	*11.64 %*	*74.03 %*	*-12.25 %*
70 %	*135697*	*142056*	*4.69 %*	*87.85 %*	*-0.06 %*
80 %	*121729*	*128125*	*5.25 %*	*86.18 %*	*-0.07 %*
90 %	*110223*	*117142*	*6.28 %*	*88.31 %*	*-0.04 %*
100 %	*101183*	*108239*	*6.97 %*	*88.43 %*	*-0.03 %*

Tabla 5.6: Impacto del consumo de energía para umbrales dinámicos.

que se realizan desconexiones de enlaces de acuerdo con la configuración de umbrales seleccionada. Este efecto se pone de manifiesto porque la potencia consumida por los enlaces de la red es una fracción del valor nominal (columna quinta, *Potencia*). En el análisis del consumo de energía también se ha considerado la energía que consumen los conmutadores o switches, asumiendo que disipan el 17,6 % del total de la potencia de la red, siendo el resto para los enlaces [51]. Se ha incluido el consumo de los conmutadores porque durante el tiempo de ejecución adicional provocado por el mecanismo de ahorro de potencia todos los componentes de la red (no sólo los enlaces) están activos y contribuyen al consumo de energía. La columna situada más a la derecha de la tabla 4.3 ($\triangle$*Energía*) muestra que el consumo de energía resulta siempre favorable cuando actúa el mecanismo de ahorro de potencia, a pesar del incremento en el tiempo de simulación. Resultados similares se han obtenido con el resto de configuraciones posibles de los umbrales.

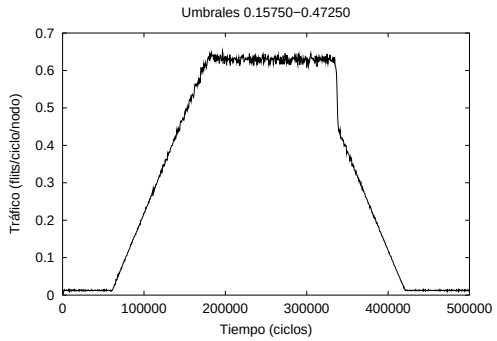
5.4.5. Evaluación dinámica

Para analizar el comportamiento dinámico del mecanismo propuesto, se han ejecutado simulaciones utilizando niveles de tráfico variables en función del tiempo. Se ha empleado la misma carga de trabajo que en el caso de las redes directas. Se inicia la simulación con un nivel de tráfico equivalente al 2 % del máximo soportado por la red, tras un período de tiempo, la carga crece lentamente y de forma constante hasta un valor del 100 % de la carga de saturación (U_{MAX}). Esta carga elevada se mantiene durante *60000* ciclos. Entonces, el tráfico de entrada disminuye nuevamente a una tasa constante hasta alcanzar el valor inicial. Este experimento ilustra el comporta-

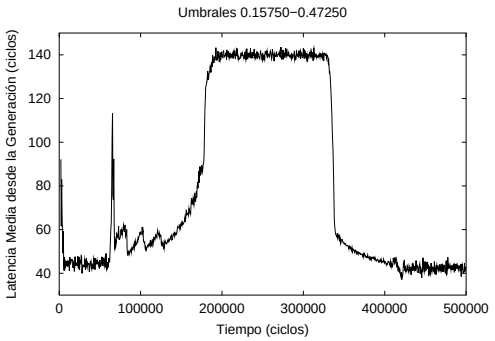
miento del sistema bajo cargas que tengan pendientes crecientes y decrecientes. El estado de la red en el instante inicial es de desconexión de todos los enlaces posibles, de acuerdo con el mecanismo propuesto, es decir solo está conectado el árbol mínimo (el consumo relativo de potencia es del 33,2 % para la red evaluada).

La figura 5.26 muestra los resultados obtenidos, para mensajes de 16 flits y una configuración de los umbrales aproximadamente en el punto central del mapa de posibles umbrales (figura 5.9), en concreto $U_{off} = 0{,}1575$ y $U_{on} = 0{,}4725$. En todos los casos, el eje de abscisas indica el tiempo transcurrido en ciclos. La figura 5.26(a) muestra el tráfico entregado, donde se aprecia la evolución en forma de doble rampa descrita. La figura 5.26(b) presenta la latencia media de los mensajes y la figura 5.26(c) proporciona los resultados de consumo relativo de potencia. Por último, la figura 5.26(d) combina en una misma gráfica la representación de la latencia y de la potencia relativa. El comportamiento del mecanismo de ahorro de potencia se observa en en esta última. En la primera fase de la simulación con carga constante, el consumo relativo de potencia se sitúa en $0,336$ frente a $0,332$ del árbol mínimo (este transitorio inicial es el responsable del pico de latencia que se produce al inicio del experimento). Ello se debe a que ni aun con el 2 % de la carga máxima se puede absorber todo el tráfico empleando solamente el árbol mínimo. Para esta configuración, mientras el tráfico es bajo solamente los enlaces del árbol mínimo más unos pocos enlaces adicionales están conectados y la latencia se mantiene constante entre 42 y 44 ciclos. Cuando la carga inicia la rampa ascendente, la latencia de los mensajes tiende a subir porque se incrementa la utilización de los enlaces. Estos incrementos de latencia son más bruscos para cargas más bajas porque la red está más cerca del árbol mínimo y es mucho más sensible a la congestión. Como respuesta, el mecanismo de control de consumo activa el proceso de conexión de enlaces. Inicialmente la conexión de enlaces hace que la latencia experimente un descenso, aun cuando la carga sigue creciendo, debido a la disponibilidad de ancho de banda adicional. Sin embargo, al mantenerse constante la tasa de incremento de tráfico inyectado, la utilización sigue aumentando y se continúan conectando más enlaces. Una vez todos los enlaces han sido conectados (en este caso eso sucede alrededor de los 125000 ciclos), el aumento adicional de tráfico provoca un aumento de latencia hasta alcanzar un valor aproximado de 140 ciclos en el que se estabiliza una vez el tráfico ha alcanzado el valor máximo (en este caso 0,63 $flits/ciclo/nodo$).

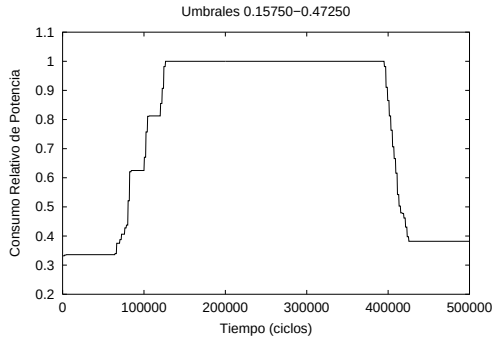
El descenso de tráfico inyectado no se percibe de forma instantánea en el tráfico entregado porque al haber llevado la red a su carga máxima, al límite de la saturación, transcurren algunos ciclos hasta que el tráfico entregado sigue al inyectado.



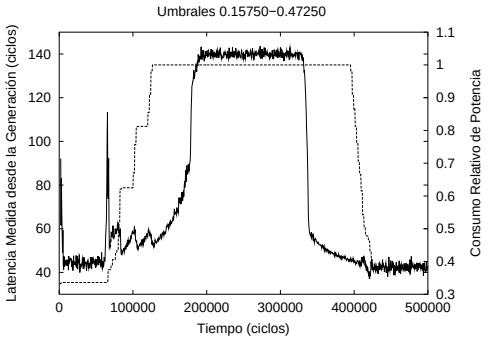
(a) Tráfico entregado en función del tiempo.



(b) Latencia en función del tiempo.



(c) Potencia consumida en función del tiempo.



(d) Latencia y potencia superpuestas.

Figura 5.26: Evaluación dinámica para un fat-tree con umbrales estáticos $U_{off} = 0{,}1575$ y $U_{on} = 0{,}4725$.

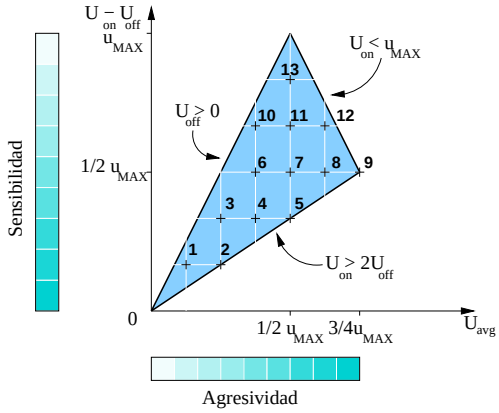
Este fenómeno se identifica por el codo en ángulo recto de descenso de carga, previamente a la rampa. El primer resultado de la reducción de tráfico es un descenso en la latencia. Cuando la reducción de tráfico hace que la utilización de los enlaces sea suficientemente baja (aproximadamente en 400000 ciclos), actúa el mecanismo de desconexión de enlaces y se produce, en este caso, un ligero incremento transitorio de la latencia previo a su estabilización en un valor mínimo de alrededor de 43 ciclos.

Las figuras 5.27 y 5.28 muestran los resultados obtenidos con umbrales estáticos para las configuraciones más significativas del mapa de posibles umbrales (se ha incluido el mapa para facilitar el análisis de los resultados). Se observa que el comportamiento del mecanismo se ajusta al del ejemplo analizado en detalle en la figura 5.26. Se verifica un impacto mayor en las curvas de latencia a medida que los requerimientos de ahorro de potencia son más exigentes por el uso de configuraciones

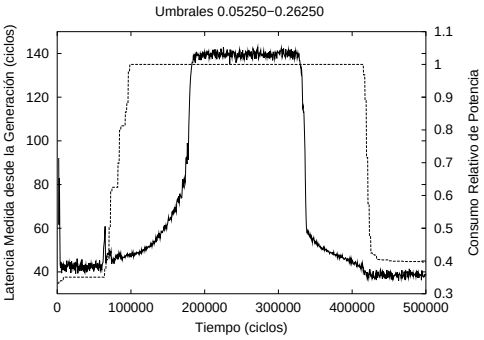
más agresivas. El caso más extremo es el mostrado en la figura 5.28(c), que se corresponde a la configuración más agresiva de todas las posibles. Cuando la red parte de un estado de conexiones cercano al del árbol mínimo y se emplean umbrales agresivos, se espera a tener utilizaciones muy altas antes de conectar enlaces disponibles y, debido a que la conexión no es instantánea, se producen transitorios en los que la latencia presenta picos positivos importantes que son absorbidos poco después.

Para los experimentos presentados, en algunos casos también se observa que el consumo de potencia al final del experimento es ligeramente superior que el inicial, para el mismo nivel de tráfico. Ello se debe a la banda de histéresis de los umbrales y el estado de partida de la red. Inicialmente se parte de una red con todos los enlaces posibles desconectados y el tráfico inyectado no genera bastante utilización para que se reconecte un número significativo de enlaces. El consumo de potencia es, por tanto, el mínimo. En cambio, al final, se parte de la red con todos los enlaces conectados y el descenso de tráfico (y de utilización de enlaces) es el que provoca la desconexión. En el experimento cuyos resultados se presentan, no se alcanza la potencia mínima ya que el tráfico mínimo en este caso particular, que es del 2 % del máximo, no es lo suficientemente bajo para que los enlaces se desconecten según los umbrales correspondientes. Este comportamiento está de acuerdo con los resultados obtenidos en la sección 5.4.3.3. Sin embargo, experimentos con niveles de tráfico mínimo inferiores han mostrado que el proceso de desconexión es correcto y se llega hasta el árbol mínimo.

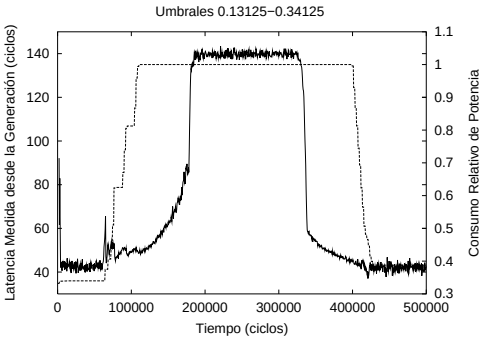
También resulta interesante destacar los tramos horizontales que presentan las curvas de potencia en las zonas correspondientes a transiciones de tráfico. Estas zonas de consumo de potencia constante, en forma de meseta, se aprecian mejor en los tramos de tráfico ascendente. En general se pueden distinguir cuatro niveles equiespaciados que se corresponden con cuatro estados promedio de la red. Dichos estados son: aquél en que los switches conectados a procesadores tienen un solo enlace ascendente conectado (árbol mínimo); el estado con switches con dos enlaces ascendentes conectados, el estado con tres o con los cuatro enlaces ascendentes conectados. Como las conexiones y desconexiones se transmiten a switches vecinos según criterios de conectividad, el estado de los switches directamente conectados a los nodos procesadores condiciona y define el estado de conexión (y de consumo) global de la red. Ello hace que la red, en promedio, presente cuatro estados de consumo estables que se reparten de manera aproximadamente equidistante entre el consumo mínimo del 33, 2 % y el máximo del 100 %.



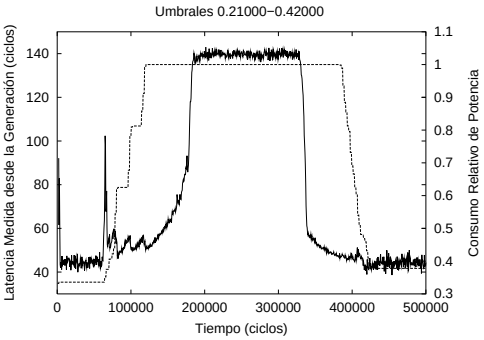
(a) Mapa de posibles umbrales.



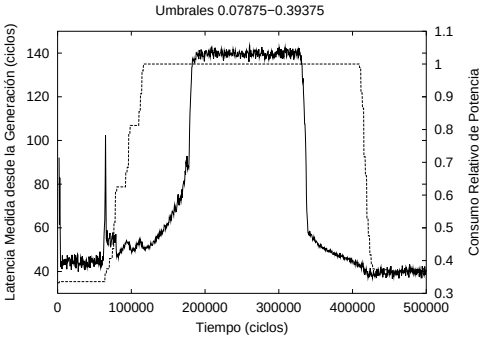
(b) Latencia y potencia en la configuración 3.



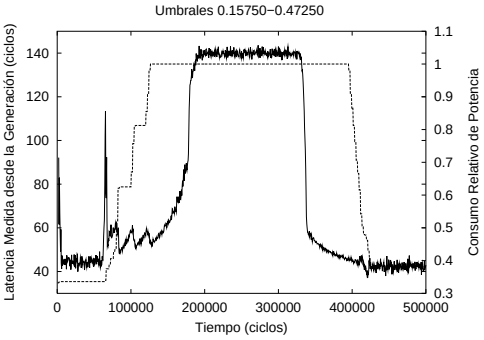
(c) Latencia y potencia en la configuración 4.



(d) Latencia y potencia en la configuración 5.

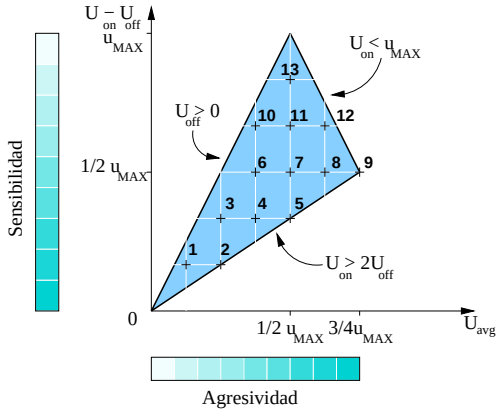


(e) Latencia y potencia en la configuración 6.

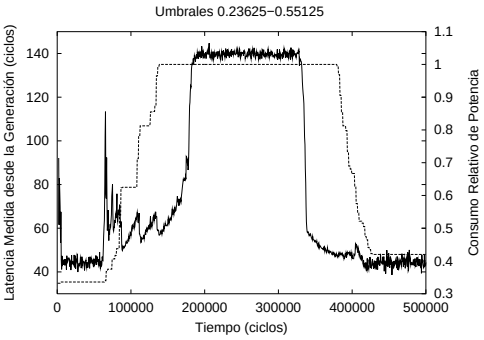


(f) Latencia y potencia en la configuración 7.

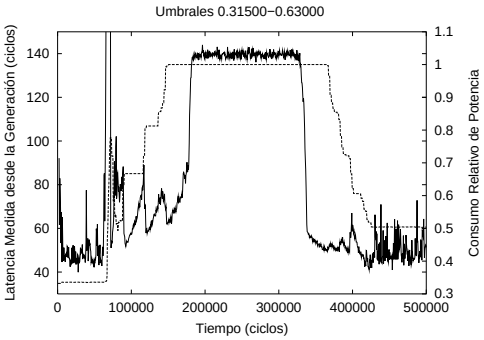
Figura 5.27: Evaluación dinámica para distintos puntos del mapa de posibles umbrales. Umbrales estáticos.



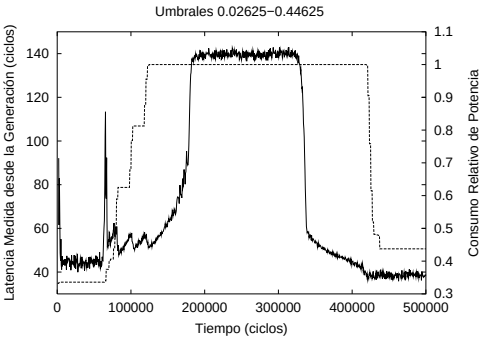
(a) Mapa de posibles umbrales.



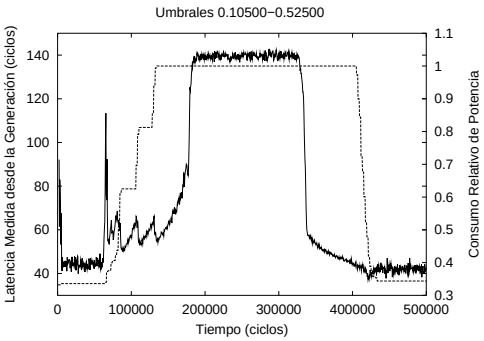
(b) Latencia y potencia en la configuración 8.



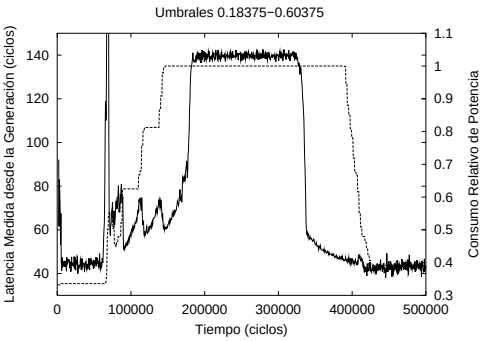
(c) Latencia y potencia en la configuración 9.



(d) Latencia y potencia en la configuración 10.



(e) Latencia y potencia en la configuración 11.



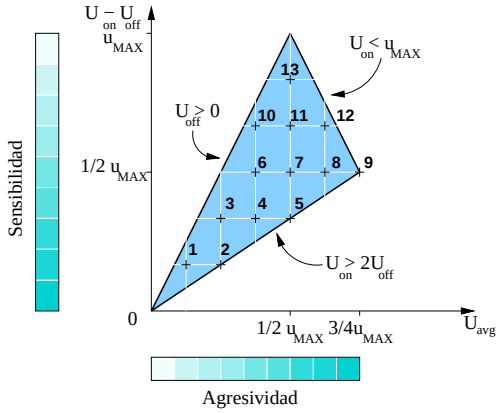
(f) Latencia y potencia en la configuración 12.

Figura 5.28: Evaluación dinámica para distintos puntos del mapa de posibles umbrales. Umbrales estáticos.

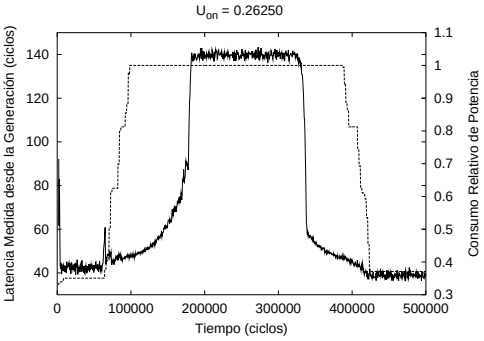
Umbrales dinámicos Se ha repetido el mismo experimento empleando la versión dinámica de los umbrales. La figuras 5.29 y 5.30 muestran los resultados obtenidos con umbrales dinámicos. En este caso solamente existe un umbral característico de cada una de las configuraciones (U_{on}) y su valor se ha empleado en los títulos de las gráficas. Se ha mantenido el formato en la presentación de los resultados para facilitar su comparación con los obtenidos con umbrales estáticos. La diferencia más significativa respecto a la variante estática de los umbrales es la reducción del consumo de potencia para cargas decrecientes. Se observa que la curva potencia presenta un comportamiento escalonado en su tramo de bajada similar al que aparece en el tramo de subida. La razón es que la variante dinámica de los umbrales hace que el umbral de desconexión efectivo (que es el que condiciona la rampa de bajada de potencia) se mantenga en valores más altos (figura 5.5). También se aprecia un ligero impacto en la latencia de los mensajes en forma de picos de latencia provocados por la desconexión de enlaces a medida que el tráfico decrece.

La evolución de la latencia de los mensajes junto con la potencia consumida cuando el tráfico aumenta o disminuye permite completar el análisis del comportamiento del mecanismo. En la figura 5.31 se representan la latencia media y la potencia frente al tráfico entregado, tanto para carga ascendente como descendente (entre el 2 % y el 100 % y viceversa), en las dos variantes del mecanismo. Las curvas para tráfico descendente deben ser leídas de derecha a izquierda. Por brevedad, se ha seleccionado una sola de las configuraciones probadas, en particular la correspondiente al punto 9, que es la que presenta la mayor agresividad teórica (con las restantes configuraciones se han obtenido resultados comparables). La escala de la gráfica recorta el pico máximo de latencia (alrededor de 400 ciclos) para mostrar con mejor resolución su comportamiento en todo el rango de tráfico.

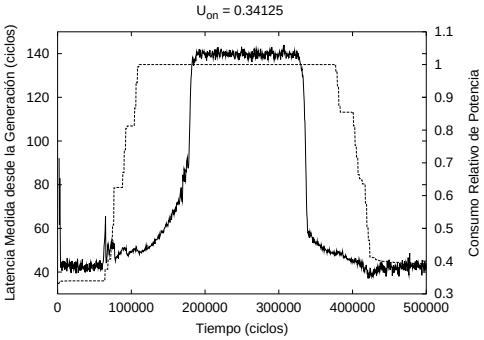
En la figura se aprecia que las curvas para tráfico ascendente son muy similares para los dos tipos de umbrales. El incremento de tráfico provoca un aumento de la utilización de los enlaces, que se manifiesta en un aumento de latencia, que a su vez dispara el mecanismo de conexión de enlaces cuando se supera el umbral U_{on}. Para cargas bajas, donde la mayoría de enlaces están desconectados, la red es muy sensible a los incrementos bruscos de tráfico (el fat-tree se reduce a un árbol simple) y esa es la causa del primer pico de latencia. Este pico queda controlado cuando los enlaces necesarios son efectivamente conectados. Como ya se ha observado, existen intervalos de tráfico para los que la red permanece en un estado de consumo de potencia estable. Para aumentos adicionales de tráfico, el comportamiento se repite con picos de latencia más suaves ya que la red se va aproximando a un fat-tree y es



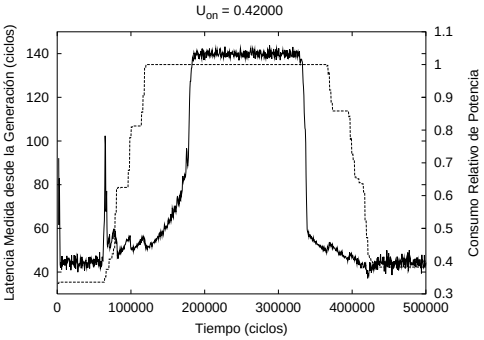
(a) Mapa de posibles umbrales.



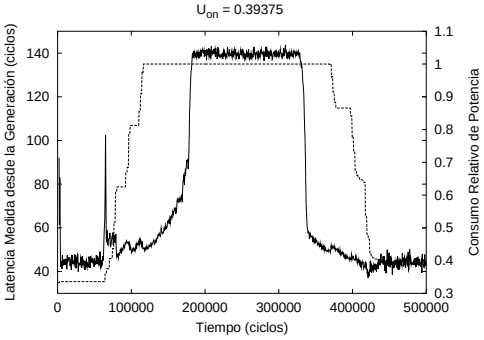
(b) Latencia y potencia en la configuración 3.



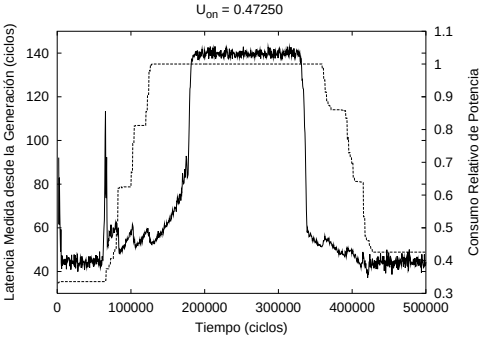
(c) Latencia y potencia en la configuración 4.



(d) Latencia y potencia en la configuración 5.

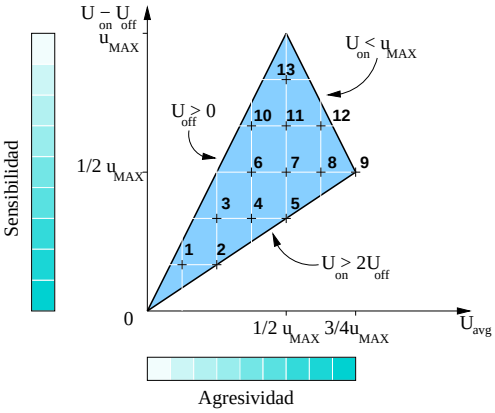


(e) Latencia y potencia en la configuración 6.

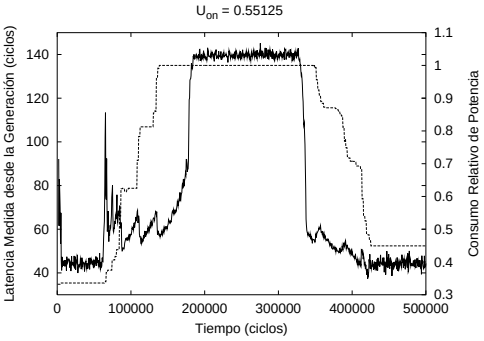


(f) Latencia y potencia en la configuración 7.

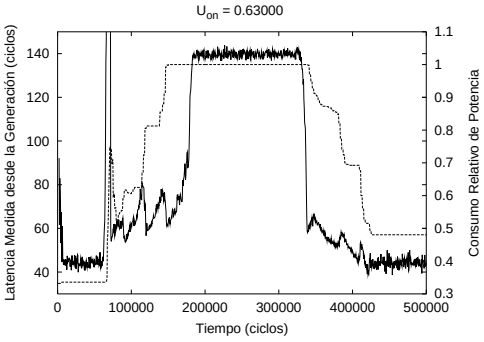
Figura 5.29: Evaluación dinámica para distintos puntos del mapa de posibles umbrales. Umbrales dinámicos.



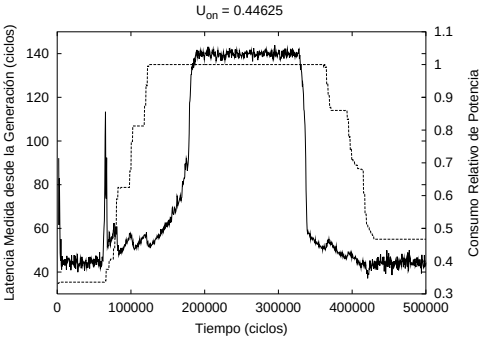
(a) Mapa de posibles umbrales.



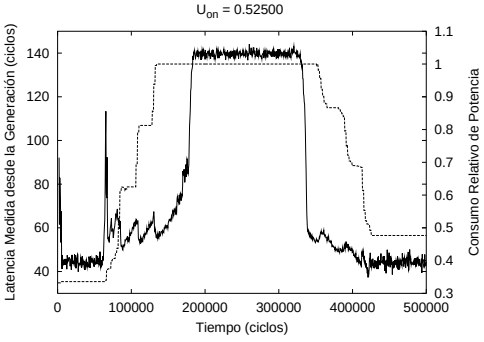
(b) Latencia y potencia en la configuración 8.



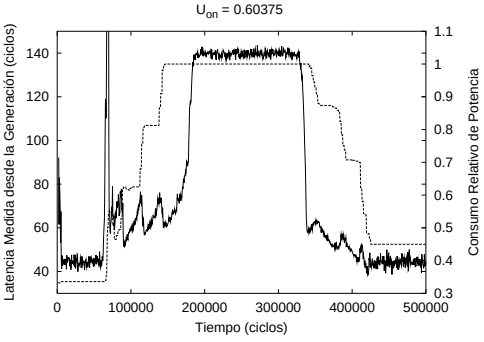
(c) Latencia y potencia en la configuración 9.



(d) Latencia y potencia en la configuración 10.

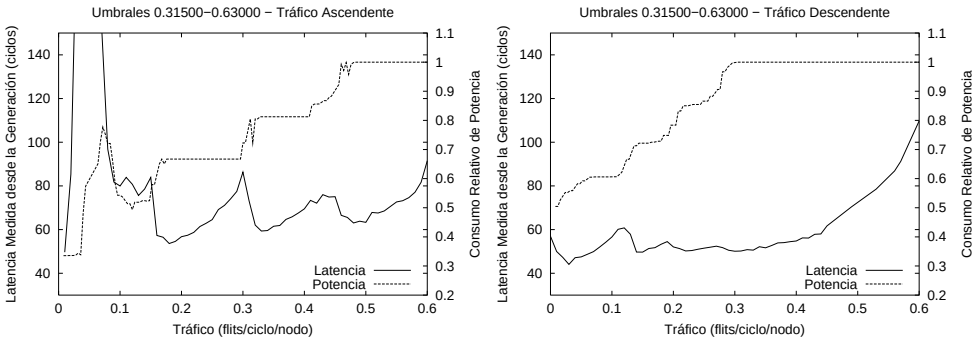


(e) Latencia y potencia en la configuración 11.



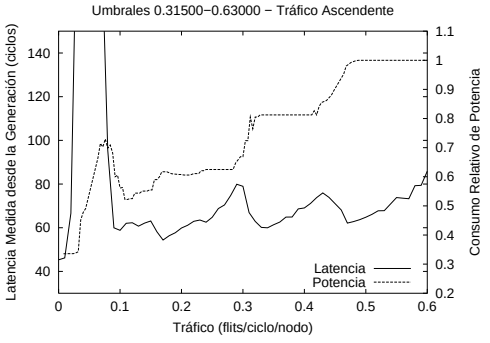
(f) Latencia y potencia en la configuración 12.

Figura 5.30: Evaluación dinámica para distintos puntos del mapa de posibles umbrales. Umbrales dinámicos.

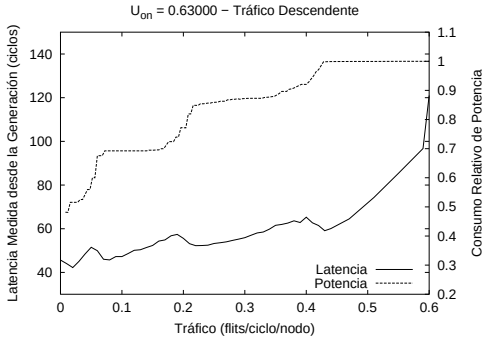


(a) Umbrales estáticos. Tráfico ascendente.

(b) Umbrales estáticos. Tráfico descendente.



(c) Umbrales dinámicos. Tráfico ascendente.



(d) Umbrales dinámicos. Tráfico descendente.

Figura 5.31: Latencia y potencia frente a tráfico con carga ascendente y descendente para el punto 9 del mapa de umbrales.

capaz de absorber mejor incrementos de tráfico.

En las curvas con tráfico descendente se manifiestan las diferencias constatadas anteriormente. Se aumenta el margen de tráfico en el que se ahorra potencia (desde un valor de 0,42 $flits/ciclo/nodo$ en umbrales dinámicos frente a 0,30 $flits/ciclo/nodo$ en umbrales estáticos) y por tanto mejora la prestaciones del mecanismo. El impacto en latencia se limita a un ligero incremento de la misma entre 0,42 y 0,30 $flits/ciclo/nodo$.

5.4.5.1. Evaluación con tráfico autosimilar

Al igual que en el capítulo anterior, tras la evaluación experimental con carga distribuida uniformemente en el tiempo, se ha completado el estudio utilizando tráfico auto-similar, más representativo de las cargas de trabajo que se pueden encontrar

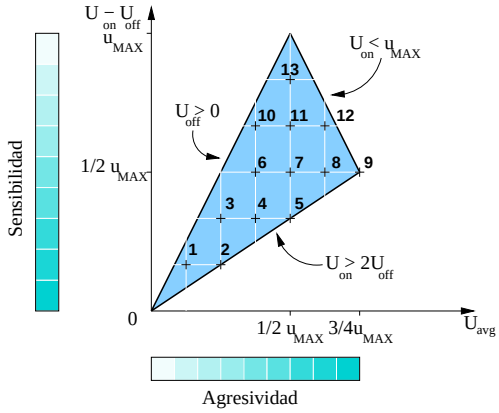
en los *clusters.* La carga auto-similar se ha generado por la agregación de fuentes de ON/OFF. Las fuentes de ON/OFF concatenan períodos de inyección de tráfico con periodos de inactividad, ambos con distribución Pareto. En nuestro simulador, cada nodo actúa como una fuente ON/OFF. Para cada enlace de la red, el tráfico generado por cada nodo se combina de acuerdo con la distribución de destinos y el algoritmo de encaminamiento, produciendo un tráfico auto-similar. Hemos verificado la autosimilitud del tráfico mediante la herramienta *Selfis* [37], obteniendo en todos los casos valores para el parámetro *Hurst* mayores a 0,7.

Los experimentos se han realizado empleando rampas de tráfico similares a las empleadas con tráfico uniforme, pero empleando diferentes niveles de tráfico. En este experimento las simulaciones se inician con niveles de tráfico del 2 % del máximo soportado por la red. Tras un período de tiempo, la carga crece durante 120000 ciclos de forma constante hasta un valor promedio del 85 % de la carga de saturación medida con tráfico uniforme (U_{MAX}). Se ha usado una carga inferior al 100 % porque la naturaleza de mayor variabilidad de la carga autosimilar hace que picos rápidos de carga puedan saturar momentáneamente la red. Esta carga elevada se mantiene durante 60000 ciclos. Entonces, el tráfico de entrada se decrementa nuevamente de forma constante hasta alcanzar el valor inicial. El estado de la red en el instante inicial es de desconexión de todos los enlaces posibles de acuerdo con el mecanismo propuesto (el consumo relativo de potencia es del 33,2 % para la red evaluada).

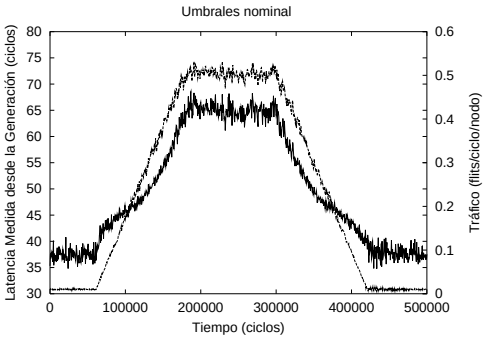
Los resultados se muestran en la figura 5.32, para umbrales estáticos, y la figura 5.33, para umbrales dinámicos. Se presentan resultados para las configuraciones más agresivas de entre las posibles de acuerdo con el mapa de umbrales. En ambos casos, la subfigura (b) muestra la latencia nominal de la red (sin mecanismo de gestión dinámica de los enlaces) junto con el tráfico recibido (que es el mismo para todas las configuraciones probadas).

Si se comparan los resultados con los obtenidos para tráfico uniforme, la diferencia más destacable es la mayor variabilidad de la latencia alrededor de su valor medio. Ello se debe al tráfico autosimilar, que se caracteriza por ráfagas de corta duración. El tráfico es también responsable de que las curvas de potencia no presenten el aspecto escalonado que se observa con tráfico uniforme. Si con tráfico uniforme la red transita por cuatro estados de consumo de potencia claramente diferenciados, con tráfico autosimilar no aparecen estados estables sino una transición progresiva entre los niveles de mínima y máxima potencia.

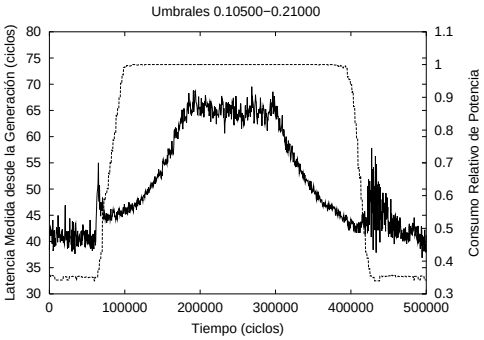
Las diferencias para umbrales estáticos y dinámicos confirman lo observado con tráfico uniforme. El comportamiento con carga ascendente, que provoca conexiones



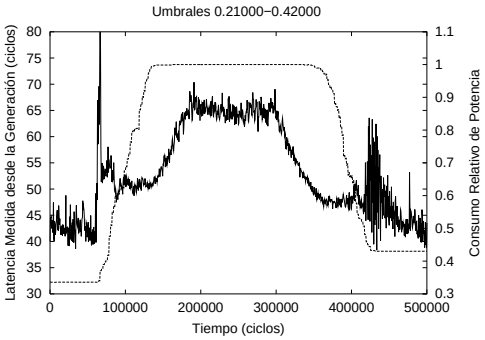
(a) Mapa de posibles umbrales.



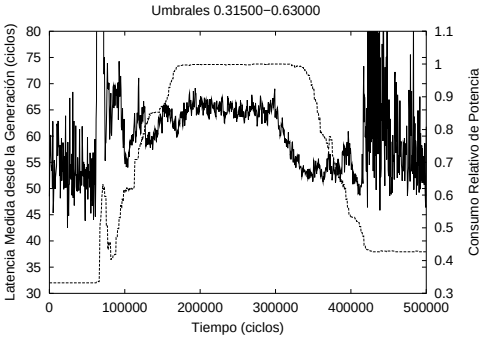
(b) Latencia y tráfico en la configuración nominal.



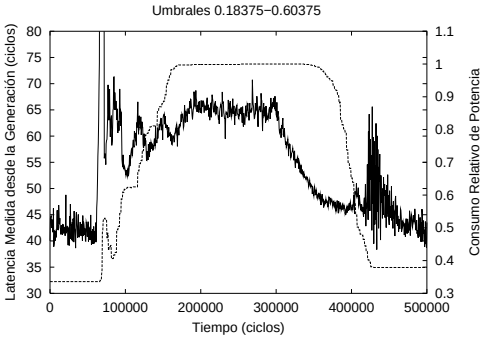
(c) Latencia y potencia en la configuración 2.



(d) Latencia y potencia en la configuración 5.

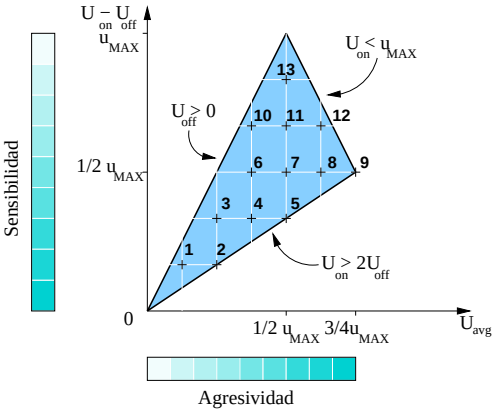


(e) Latencia y potencia en la configuración 9.

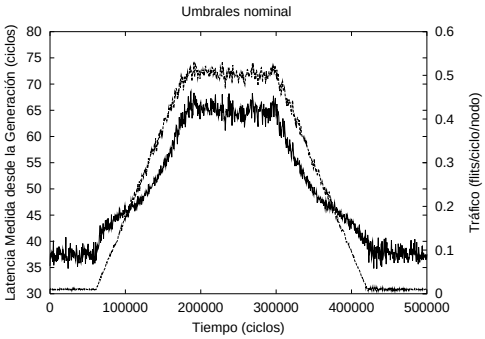


(f) Latencia y potencia en la configuración 12.

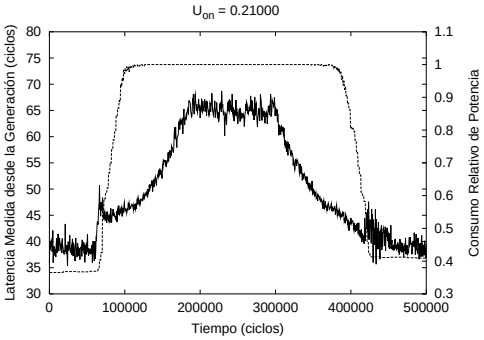
Figura 5.32: Evaluación dinámica con tráfico autosimilar usando umbrales estáticos para distintos puntos del mapa de posibles umbrales.



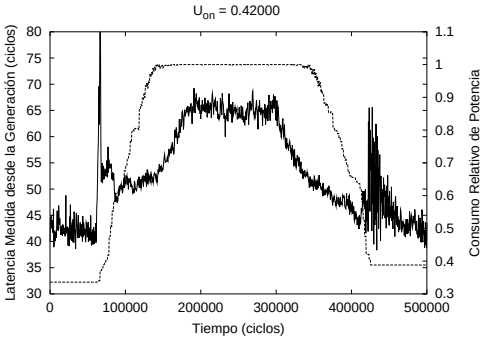
(a) Mapa de posibles umbrales.



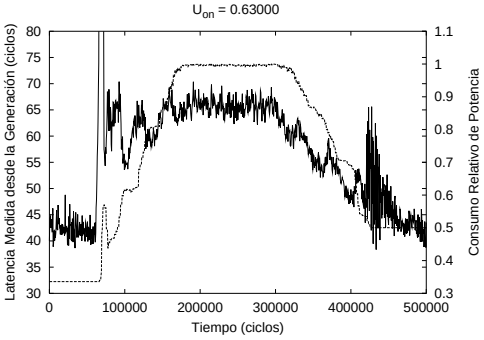
(b) Latencia y tráfico en la configuración nominal.



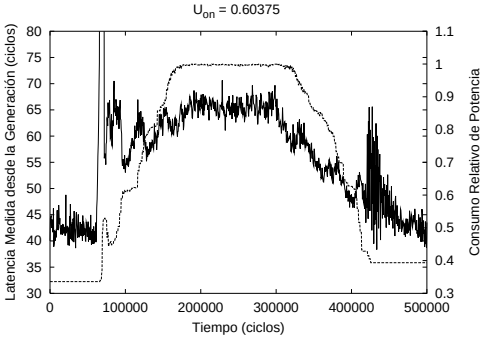
(c) Latencia y potencia en la configuración 2.



(d) Latencia y potencia en la configuración 5.



(e) Latencia y potencia en la configuración 9.



(f) Latencia y potencia en la configuración 12.

Figura 5.33: Evaluación dinámica con tráfico autosimilar usando umbrales dinámicos para distintos puntos del mapa de posibles umbrales.

de enlaces, es muy similar para ambas modalidades. En cambio, con tráfico descendente, la variante con umbrales dinámicos permite desconectar antes (para cargas más altas) los enlaces, con un impacto significativo en las curvas de potencia. Por ejemplo, en la figura 5.33(e) (que se corresponde con la configuración más agresiva de los umbrales dinámicos) se inicia el ahorro de potencia alrededor del ciclo 300000, momento significativamente anterior al equivalente con umbrales estáticos (alrededor del ciclo 360.000, figura 5.32(e)).

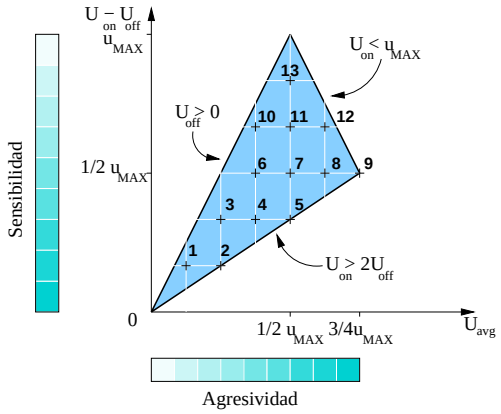
5.4.5.2. Diagramas de histéresis

Con el fin de caracterizar totalmente el comportamiento dinámico del mecanismo de ahorro de potencia se han construido los diagramas de histéresis de potencia consumida frente a tráfico entregado. En estos diagramas se representa la evolución de la potencia ante aumentos y decrementos de tráfico. Se ilustra así el proceso de conexión de enlaces a medida que la utilización de los mismos se incrementa y la desconexión cuando su utilización desciende. Las figuras 5.34 a 5.37 recogen las configuraciones 3 a 12 del mapa de umbrales posibles para las configuraciones con umbrales estáticos y dinámicos.

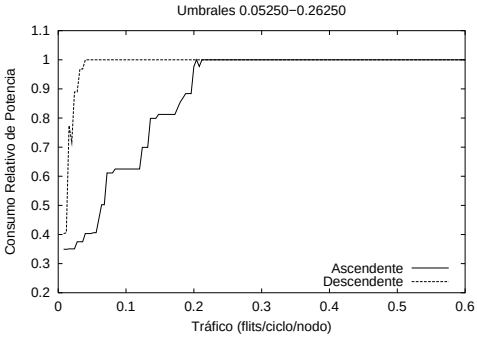
Se observa que el comportamiento medido de la red se ajusta a lo predicho por el análisis teórico realizado en la sección 5.3. La morfología de los diagramas de histéresis se aproxima a las representaciones de la figura 5.4, para umbrales estáticos, y la figura 5.5, para umbrales dinámicos. Un análisis comparativo pone en evidencia que la diferencia en prestaciones de las dos variantes de umbrales se manifiesta a través de la curva de potencia para tráfico descendente. Sin embargo, la curva con tráfico ascendente no cambia significativamente al pasar de umbrales estáticos a dinámicos, tal y como predice el análisis teórico. Se constata experimentalmente que se consigue el efecto deseado de mantener constante la banda de histéresis efectiva mediante los umbrales dinámicos. Como resultado, se incrementa el conjunto de niveles de tráfico para los cuales se consigue ahorrar potencia.

5.5. Conclusiones

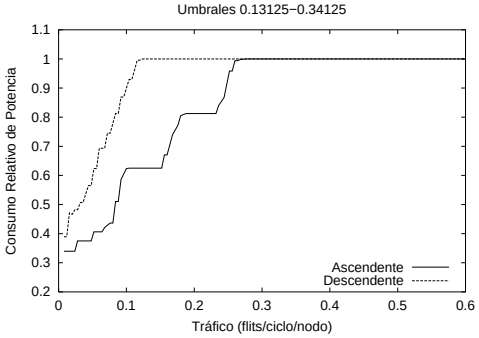
En este capítulo se ha presentado un mecanismo de ahorro de potencia para redes de interconexión indirectas o multietapa, de entre las que destaca el fat-tree. El mecanismo propuesto se basa en conectar o desconectar los enlaces redundantes de la red en función del tráfico, manteniendo siempre conectividad total entre los



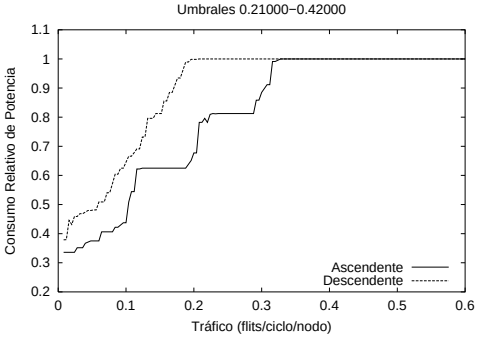
(a) Mapa de posibles umbrales.



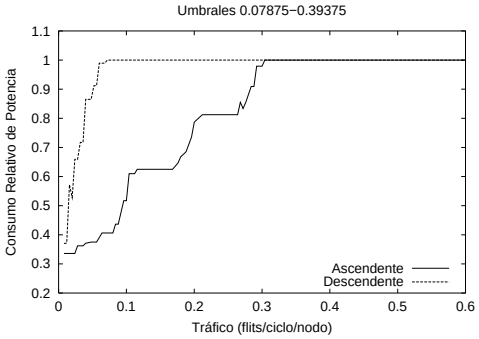
(b) Potencia consumida para la configuración 3.



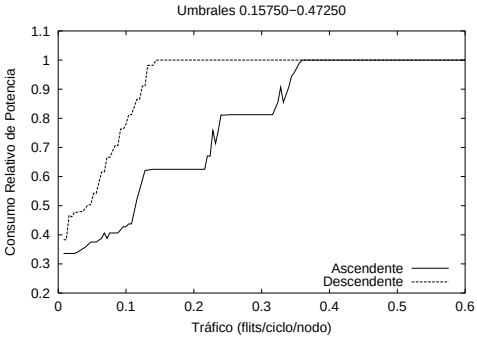
(c) Potencia consumida para la configuración 4.



(d) Potencia consumida para la configuración 5.

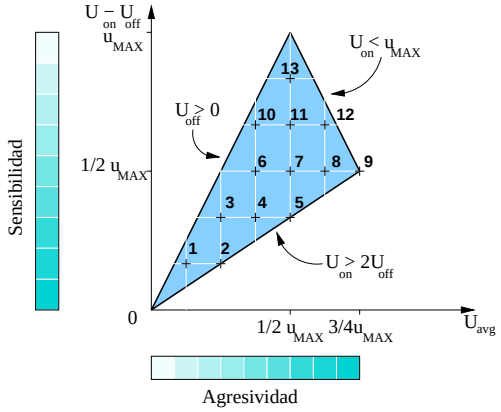


(e) Potencia consumida para la configuración 6.

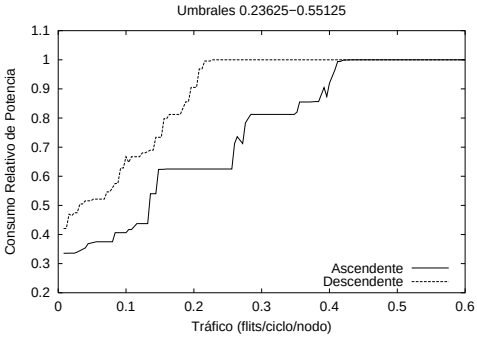


(f) Potencia consumida para la configuración 7.

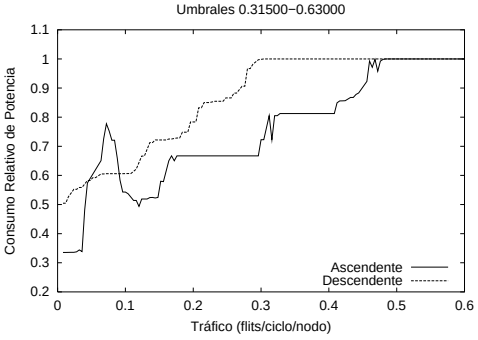
Figura 5.34: Diagramas de histéresis con umbrales estáticos.



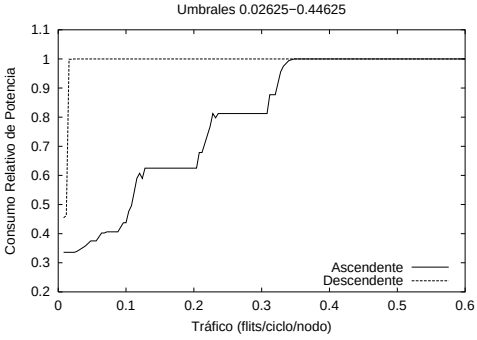
(a) Mapa de posibles umbrales.



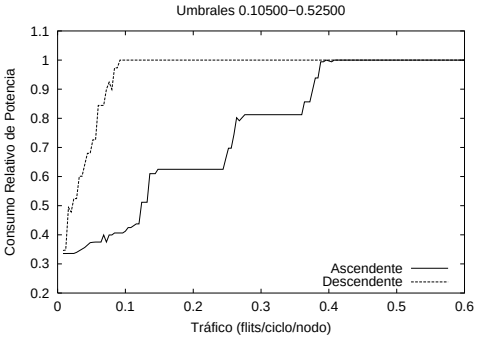
(b) Potencia consumida para la configuración 8.



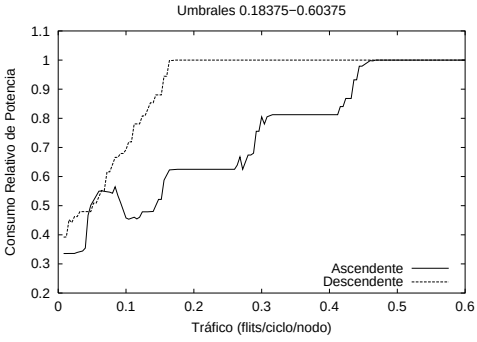
(c) Potencia consumida para la configuración 9.



(d) Potencia consumida para la configuración 10.

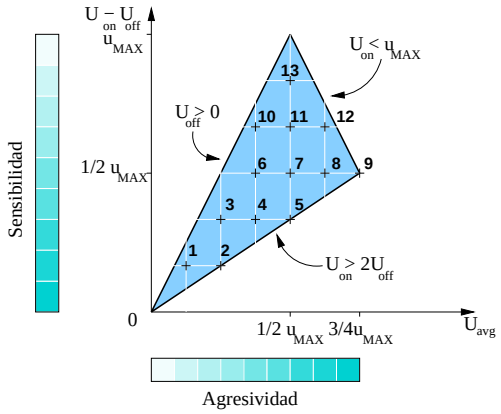


(e) Potencia consumida para la configuración 11.

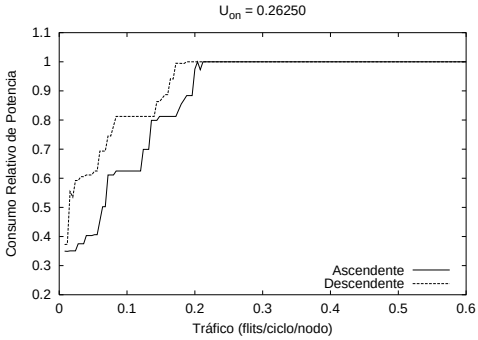


(f) Potencia consumida para la configuración 12.

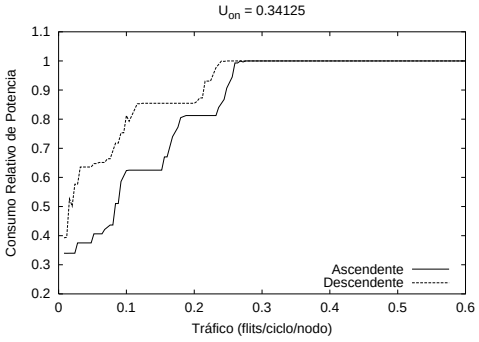
Figura 5.35: Diagramas de histéresis con umbrales estáticos.



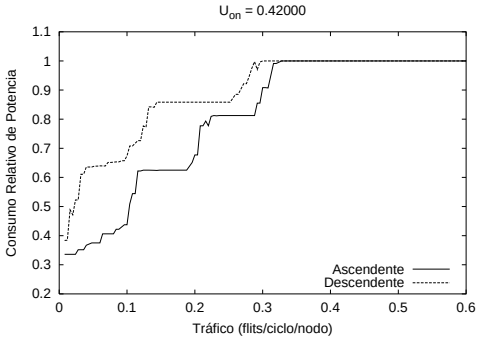
(a) Mapa de posibles umbrales.



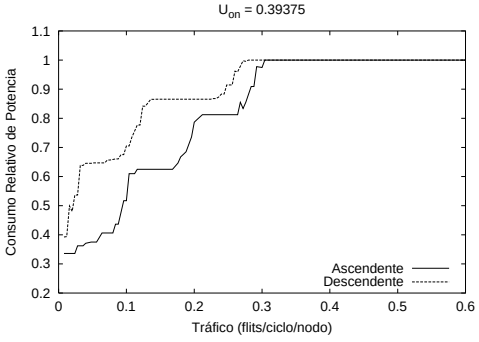
(b) Potencia consumida para la configuración 3.



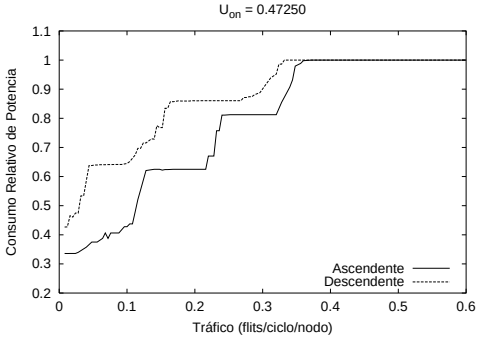
(c) Potencia consumida para la configuración 4.



(d) Potencia consumida para la configuración 5.

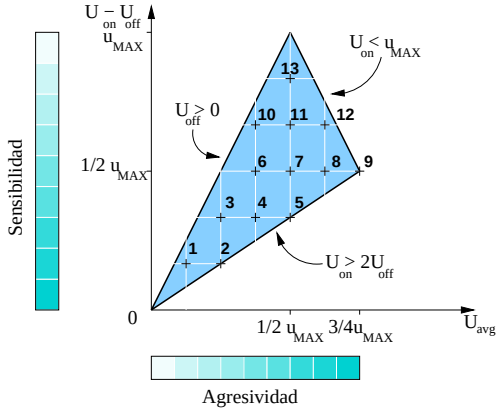


(e) Potencia consumida para la configuración 6.

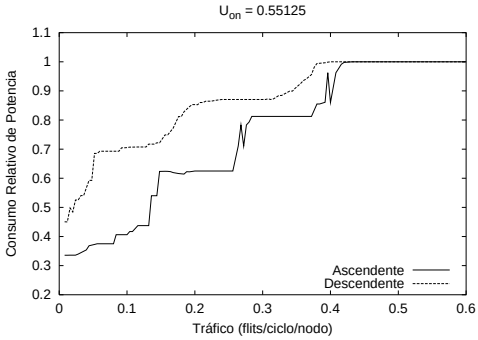


(f) Potencia consumida para la configuración 7.

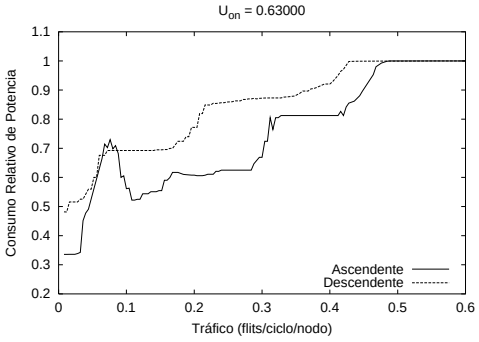
Figura 5.36: Diagramas de histéresis con umbrales dinámicos.



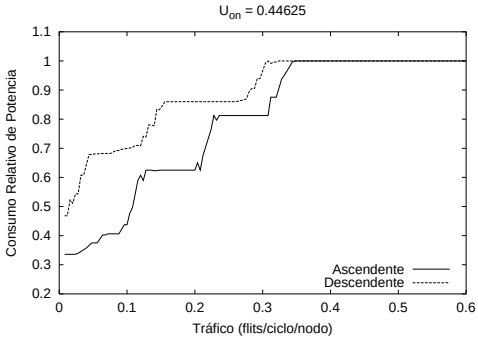
(a) Mapa de posibles umbrales.



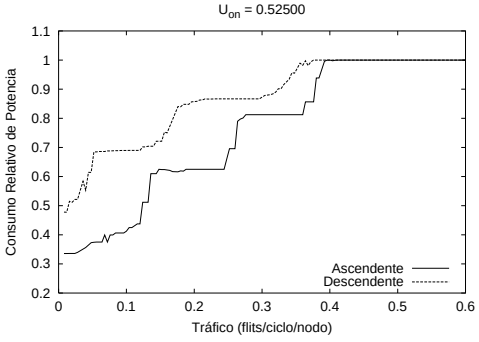
(b) Potencia consumida para la configuración 8.



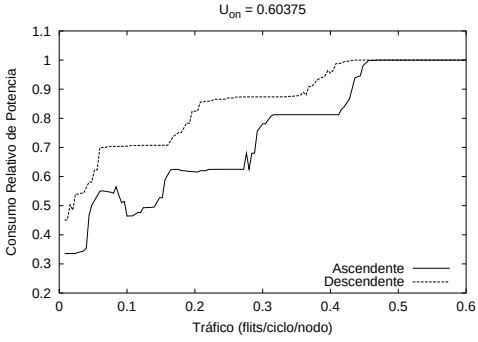
(c) Potencia consumida para la configuración 9.



(d) Potencia consumida para la configuración 10.



(e) Potencia consumida para la configuración 11.



(f) Potencia consumida para la configuración 12.

Figura 5.37: Diagramas de histéresis con umbrales dinámicos.

nodos. Se ha caracterizado el mecanismo en función de los parámetros que lo gobiernan, los cuales permiten ajustar la política de gestión energética de la red. La exhaustiva evaluación experimental presentada demuestra la capacidad del mecanismo para propocionar un significativo ahorro de potencia, con un impacto muy ligero en latencia. Se ha verificado que el producto de potencia relativa por latencia relativa proporciona resultados muy favorables. Solamente en determinadas configuraciones extremadamente agresivas con mensajes largos y umbrales dinámicos se han observado resultados ligeramente desfavorables. Aún así, el mecanismo admite un amplio abanico de configuraciones que combinan agresividad y sensibilidad con un balance potencia-latencia favorable.

Analizando el impacto de la longitud de los mensajes (sección 5.4.4.1) sobre las prestaciones del mecanismo de ahorro de potencia, la principal conclusión es que el mecanismo presenta resultados ligeramente más favorables para mensajes de mayor longitud. Al igual que en el caso de las redes directas, esto se debe a que la distribución del trafico en mensajes más largos, reduce la sobrecarga debida a las cabeceras y por tanto la utilización efectiva de los enlaces, incrementando el ahorro de potencia.

La evaluación del efecto de la función de selección (sección 5.4.4.2) indica que una concentración del tráfico en unos pocos enlaces con el objeto de generar oportunidades de desconexión de los restantes, por baja utilización, no resulta favorable. El impacto en la latencia es tan significativo que la reducción de potencia que se consigue, en ninguno de las configuraciones probadas, proporciona resultados favorables. Los mejores resultados se obtienen con una función de selección adaptativa que proporciona una más uniforme distribución de carga en la red. Este resultado coincide plenamente con las conclusiones obtenidas para redes directas.

Respecto a la selección de umbrales, para umbrales estáticos se ha probado todo el mapa de umbrales posibles, y verificado que en todos los casos el balance incremento de latencia-consumo de potencia resulta favorable. En el caso de los umbrales dinámicos, se obtienen resultados favorables con un umbral de conexión U_{on} por debajo de 0,35. Si se comparan las dos variantes, umbrales estáticos y umbrales dinámicos, los resultados que se consiguen son significativamente mejores para el caso de umbrales dinámicos. El rango de tráfico para el cual se reduce el consumo de potencia es más amplio en el caso de los umbrales dinámicos. La evaluación con cargas dinámicas confirma este extremo al mostrar unas bandas de histéresis más estrechas cuando se emplean umbrales dinámicos. El análisis de la energía consumida en experimentos con carga cerrada (sección 5.4.4.3) concluye que para baja carga el ahorro en el consumo de potencia es equiparable para umbrales estáticos y dinámicos, pe-

ro cuando la carga es más elevada, los umbrales dinámicos proporcionan mejores ahorros en el consumo.

6

Conclusiones y trabajo futuro

La principal contribución de esta tesis es el desarrollo de un nuevo mecanismo para la reducción del consumo de potencia en redes de interconexión. La técnica propuesta se basa en la gestión dinámica del estado de los enlaces de la red para adaptarlo a los requerimientos cambiantes de la carga del sistema. Debido a la variabilidad característica del tráfico de red, en muchas ocasiones, existen enlaces ociosos o poco utilizados. Con un mecanismo distribuido de gestión del estado de los enlaces, se determina, en función del tráfico, cuál es la mejor configuración de los mismos. El mecanismo planteado es flexible y puede ser totalmente sintonizado de acuerdo a parámetros predefinidos de agresividad y sensibilidad. Se consigue con ello hacer una gestión eficiente de los recursos, evitando mantener activos recursos que no se están utilizando o que no van a proporcionar una mejora significativa de las prestaciones.

El diseño de la estrategia para la reducción del consumo de potencia ha requerido abordar y resolver varios problemas:

- Selección de una métrica que permite cuantificar el tráfico en la red de manera distribuida y que no requiere de sofisticados elementos hardware. Hemos

seleccionado la utilización de los enlaces como la métrica más adecuada para representar el tráfico y hemos verificado que se ajusta a los requerimientos planteados. Se ha propuesto una métrica alternativa a la utilización de los enlaces para detectar situaciones de congestión, ya que la utilización no garantiza la detectabilidad de estas situaciones, más probables debido a la alta sensibilidad de la red a picos de carga cuando está operando en modo de bajo consumo. Se ha definido también un mecanismo para eliminarla.

- Una estrategia para gestionar dinámicamente los enlaces. Hemos desarrollado un mecanismo que determina el comportamiento de los enlaces en función del tráfico estimado y de unos parámetros de control definidos por unos umbrales de conexión y desconexión.

- La caracterización de la estrategia de ahorro de potencia en función de unos parámetros de control que permitan definir políticas de ahorro más o menos agresivas. Se ha realizado un análisis detallado de dichos parámetros, su impacto sobre el comportamiento del sistema y sus limitaciones.

- La adaptación de la estrategia propuesta a las topologías de red de mayor interés en la actualidad. Se han definido los mecanismos de ahorro de potencia tanto para toros, como ejemplo más representativo de las topologías directas, como para fat-tree, topología indirecta más popular.

- La caracterización y evaluación experimental del mecanismo y su impacto tanto sobre la potencia consumida como su efecto sobre las prestaciones. Se ha implementado la política de ahorro de potencia sobre nuestro simulador de redes de interconexión. Hemos realizado una extensiva evaluación experimental que ha permitido valorar su impacto con múltiples configuraciones y bajo distintas condiciones de carga.

  Un aspecto significativo de nuestra propuesta es que se implementa por medio de un mecanismo totalmente distribuido. Ello la dota de una escalabilidad óptima frente a propuestas basadas en soluciones de gestión de la red centralizadas. Por otro lado, nuestra estrategia es completamente configurable en base a únicamente dos parámetros: un umbral de conexión y un umbral de desconexión de los enlaces en función de su utilización. Hemos demostrado que la política de ahorro puede ser totalmente sintonizada en términos del objetivo deseado de ahorro (agresividad) y de su velocidad de respuesta (sensibilidad).

Nuestros resultados demuestran que la estrategia de ahorro de potencia propuesta proporciona reducciones muy significativas del consumo con un impacto muy ligero sobre las prestaciones de la red de interconexión. Los resultados para las topologías evaluadas, representativas del escenario actual de las redes de interconexión, son comparables. En todos los casos, para todas las configuraciones y bajo todas las condiciones probadas, se ha verificado que el pequeño aumento en la latencia sufrida por los mensajes, provocado por la desconexión de enlaces con baja carga, es ampliamente compensado por la reducción de la potencia. La conclusión global de mi trabajo de tesis es que nuestra estrategia para la reducción del consumo de potencia constituye una solución eficaz para el ahorro energético en redes de interconexión. Se trata de una solución de sencilla implementación, pues no requiere de complicados mecanismos en hardware, que no requiere modificaciones en el algoritmo de encaminamiento subyacente y que permite una sencilla configuración en función de los objetivos de ahorro deseados. Existe además la garantía de que incluso con los niveles de ahorro más exigentes el balance energético será favorable.

El trabajo presentado en esta tesis ha planteado algunos retos que merecen mayor atención y que orientan alguna líneas de trabajo futuro. Nos gustaría explorar nuevas metodologías para la detección y el control de la congestión. Esta situación se presenta con más probabilidad en redes en las que se aplican estrategias de ahorro basadas en la desconexión de enlaces, por ejemplo cuando se producen picos de carga tras periodos de baja carga. También merece una atención especial la adaptación de los mecanismos planteados a las particularidades de las redes de interconexión en chip. Adicionalmente, tenemos previsto estudiar las implicaciones de los parámetros de diseño de la red en las oportunidades de reducción de consumo de potencia mediante la gestión dinámica de los enlaces y su relación con el coste de implementación.

Bibliografía

[1] Advanced Switching Interconnect SIG home page. Disponible en http://www.asi-sig.org/home.

[2] Booksim 2.0 user's guide. Version 2.0. Disponible en http://nocs.stanford.edu/cgi-bin/trac.cgi/wiki/Resources/BookSim.

[3] *IBM InfiniBand 8-port 12x switch.* Disponible en http://www-3.ibm.com/chips/products/infiniband.

[4] Low power design guide. Disponible en http://www.lowpower.de/index.php.

[5] Mellanox home page. Disponible en http://www.mellanox.com.

[6] *Mellanox technologies performance, price, power, volume (PPPV).* Disponible en http://www.mellanox.com/pdf/whitepapers/Blade_WP_120.pdf.

[7] Myricom Inc. home page. http://www.myricom.com.

[8] Power utilization techniques with links of interconnection networks home page. http://cva.stanford.edu/classes/ee382c/research.

[9] Quadrics home page. http://www.quadrics.com.

[10] EPA Announces New Computer Efficiency Requirements. Disponible en http://www.epa.gov.

[11] *HyperTransport I/O Link Specification. Revision 3.10*, 2008. Disponible en http://www.hypertransport.org/docs/twgdocs/HTC20051222-00046-0028.pdf.

[12] *Intel QuickPath Architecture. A new system architecture for unleashing the performance of future generations of Intel multi-core microprocessors*, 2008. Disponible en http://www.intel.com/technology/quickpath/whitepaper.pdf.

[13] ALONSO, M., COLL, S., MARTÍNEZ, J.-M., SANTONJA, V., LÓPEZ, P., AND DUATO, J. Dynamic power saving in fat-tree interconnections networks using on/off links. In *Proceedings of the 20th IEEE International Parallel and Distributed Processing Symposium (IPDPS'05)* (Los Alamitos, CA, USA, 2006), IPDPS '06, IEEE Computer Society.

[14] ALONSO, M., COLL, S., MARTÍNEZ, J.-M., SANTONJA, V., LÓPEZ, P., AND DUATO, J. Power saving in regular interconnection networks. *PARALLEL COMPUTING 36* (December 2010), 696–712.

[15] ALONSO, M., COLL, S., SANTONJA, V., MARTÍNEZ, J.-M., LÓPEZ, P., AND DUATO, J. Power aware fat-tree networks using on/off links. In *Proceedings of the 3rd International Conference on High Performance Computing and Communications (HPCC 2007)* (2007), HPCC '07, LECTURE NOTES IN COMPUTER SCIENCE.

[16] ALONSO, M., COLL, S., SANTONJA, V., MARTÍNEZ, J.-M., LÓPEZ, P., AND DUATO, J. Reducing power consumption in fat-tree networks using on/off links. In *Congreso Español de Informática (CEDI), XVIII Jornadas de Paralelismo (JP'2007)* (2007), JP'07, Thomson Paraninfo.

[17] ALONSO, M., MARTÍNEZ, J.-M., SANTONJA, V., AND LÓPEZ, P. Reducing Power Consumption in Interconnection Networks by Dynamically Adjusting Link Width. In *Lecture Notes in Computer Science* (Springer-Verlag, 2004), vol. 3149, pp. 882–890.

[18] ALONSO, M., MARTÍNEZ, J.-M., SANTONJA, V., LÓPEZ, P., AND DUATO, J. Power saving in regular interconnection networks built with high-degree switches. In *Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium (IPDPS'05)* (Los Alamitos, CA, USA, 2005), IPDPS '05, IEEE Computer Society.

[19] ALONSO, M., MARTÍNEZ, J.-M., SANTONJA, V., LÓPEZ, P., AND DUATO, J. Reducing power consumption in fat-tree networks using on/off links. In *Congreso Español de Informática (CEDI), XVI Jornadas de Paralelismo (JP'2005)* (2005), JP'05, Thomson Paraninfo.

[20] BODEN, J. J., COHEN, D., FELDERMAN, R. E., KULAWICK, A. E., SEITZ, C. L., SEIZOVIC, J. N., AND SU, W.-K. Myrinet: A Gigabit-per-Second Local Area Network. *IEEE Micro 15*, 1 (January 1995), 29–36.

[21] BURD, T. D., AND BRODERSEN, R. W. Design issues for dynamic voltage scaling. In *ISPLED'00: Proceedings of the 2000 International Symposium on Low Power Electronics and Design* (2000), pp. 99–14.

[22] CASSIDAY, D. Infiniband architecture tutorial. Hot Chips 12 Tutorial, August 2000.

[23] CHEN, X., PEH, L.-S., WEI, G.-Y., HUANG, Y.-K., AND PRUCNAL, P. Exploring the Design Space of Power-Aware Opto-Electronic Network Systems. In *Proceedings of the 11th International Symposium on High-Performance Computer Architecture* (San Francisco, CA, February 2005).

[24] CHUN FENG, W., AND CAMERON, K. W. The green500 list: Encouraging sustainable supercomputing. *Computer 40*, 12 (December 2007), 50–55. Disponible en `http://doi.ieeecomputersociety.org/10.1109/MC.2007.445`.

[25] DALLY, W. J., CARVEY, P. P., AND DENNISON, L. R. The avici terabit switch/router. In *Hot Interconnects VI Symposium* (1998).

[26] DALLY, W. J., AND TOWLES, B. *Principles and Practices of Interconnection Networks*. Morgan Kaufmann, March 2004.

[27] DUATO, J. A New Theory of Deadlock-Free Adaptive Multicast Routing in Wormhole Networks. In *Proceedings of the 5th IEEE Symposium on Parallel and Distributed Processing (SPDP)* (December 1993), pp. 64–71.

[28] DUATO, J., YALAMANCHILI, S., AND NI, L. *Interconnection Networks: an Engineering Approach*. Morgan Kaufmann, August 2002.

[29] ERSOZ, D., YOUSIF, M., AND DAS, C. Characterizing network traffic in a cluster-based, multi-tier data center. In *Distributed Computing Systems, 2007. ICDCS '07. 27th International Conference on* (june 2007), p. 59.

[30] ET. AL, E. J. K. Energy optimization techniques in cluster interconnects. In *Proc. of the International Symposium on Low Power Electronics and Design (ISLPED'03)* (Aug 2003), pp. 459–464.

[31] GLASS, C. J., AND NI, L. M. The Turn Model for Adaptive Routing. In *Proceedings of the 19th International Symposium on Computer Architecture* (May 1992).

[32] GOVIL, K., CHAN, E., AND WASSERMAN, H. Comparing algorithms for dynamic speed-setting of a low-power CPU. In *MobiCom '95: Proceedings of the 1st annual international conference on Mobile computing and networking* (New York, NY, USA, 1995), ACM, pp. 13–25. Disponible en `http://doi.acm.org/10.1145/215530.215546`.

[33] THE GREEN500 LIST. 2009.
Disponible en `http://www.green500.org`.

[34] HENNESSY, J. L., AND PATTERSON, D. A. *Computer Architecture: Quantitative Approach*, fourth ed. Morgan Kaufmann Publishers, 2006.

[35] IEEE P802.3AZ ENERGY EFFICIENT ETHERNET TASK FORCE. 2009.
Disponible en `http://www.ieee802.org/3/az/index.html`.

[36] JAIN, A., ANDERSON, W., BENNINGHOFF, T., BERUCCI, D., BRAGANZA, M., BURNETIE, J., CHANG, T., EBLE, J., FABER, R., GOWDA, O., GRODSTEIN, J., HESS, G., KOWALESKI, J., KUMAR, A., MILLER, B., MUELLER, R., PAUL, P., PICKHOLTZ, J., RUSSELL, S., SHEN, M., TRUEX, T., VARDHARAJAN, A., XANTHOPOULOS, D., AND ZOU, T. A 1.2 ghz alpha microprocessor with 44.8 gb/s chip pin bandwidth. In *Solid-State Circuits Conference, 2001. Digest of Technical Papers. ISSCC. 2001 IEEE International* (2001), pp. 240–241.

[37] KARAGIANNIS, T., AND FALOUTSOS, M. SELFIS: A Tool For Self-Similarity and Long-Range Dependence Analysis. In *1st Workshop on Fractals and Self-Similarity in Data Mining: Issues and Approaches* (Edmonton, Canada, July 2002).

[38] KIM, J., AND HOROWITZ, M. A. Adaptive Supply Serial Links with sub-1V Operation and Per-pin Clock Recovery. *IEEE Journal of Solid State Circuits* (November 2002), 1403–1413.

[39] LEISERSON, C. E. Fat-Trees: Universal Networks for Hardware Efficient Supercomputing. *IEEE Transactions on Computers C-34*, 10 (October 1985), 892–901.

[40] LELAND, W. E., TAQQU, M. S., WILLINGER, W., AND WILSON, D. V. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Trans. Netw. 2*, 1 (1994), 1–15.

[41] LOPEZ, P., AND DUATO, J. Deadlock-free adaptive routing algorithms for 3D-torus: limitations and solutions. In *PARLE'93: Proceedings of the 5th international PARLE conference on parallel architectures and languages europe* (June 1993).

[42] MANDELBROT, B., AND TAQQU, M. Robust r/s analysis of long run serial correlation. In *Proc. 42nd Session ISI* (1979), vol. XLVIII of *2*, pp. 69–99.

[43] MONTANANA, J. M., KOIBUCHI, M., MATSUTANI, H., AND AMANO, H. Stabilizing path modification of power-aware on/off interconnection networks. *Networking, Architecture, and Storage, International Conference on 0* (2010), 218–227.

[44] NORDMAN, B. Energy Efficient Ethernet: Outstanding Questions. Tech. rep., Lawrence Berkeley National Laboratory, January 2007. Disponible en `http://www.ieee802.org/3/eee_study/public/jan07/nordman_01_0107.pdf`.

[45] PEH, L.-S. Low-power Interconnection Networks. In *Workshop on On- and Off-Chip Interconnection Networks for Multicore Systems* (Stanford University, July 2006).
Disponible en `http://www.ee.princeton.edu/~peh/talks/Lowpowernws.pdf`.

[46] PETRINI, F., CHUN FENG, W., HOISIE, A., COLL, S., AND FRACHTENBERG, E. The Quadrics Network: High Performance Clustering Technology. *IEEE Micro 22*, 1 (January-February 2002), 46–57.

[47] PETRINI, F., AND VANNESCHI, M. Performance Analysis of Wormhole Routed *k*-ary *n*-trees. *International Journal on Foundations of Computer Science 9*, 2 (June 1998), 157–177. Available from `http://www.c3.lanl.gov/~fabrizio/papers/ijfcs98.ps.gz`.

[48] ROBERT A., M., SINGH, G., AND SAFRANEK, R. J. The architecture of the intel (r) quickpath interconnect.

[49] ROTH, K. W., GOLDSTEIN, F., AND KLEINMAN, J. Energy Consumption by Office and Telecommunications Equipment in Commercial Buildings. Volume I: Energy Consumption Baseline. Tech. rep., Arthur D. Little Inc. for Office of Building Technology State and Community Programs (DOE), January 2002.

[50] SHALF, J., KAMIL, S., OLIKER, L., AND SKINNER, D. Analyzing ultrascale application communication requirements for a reconfigurable hybrid interconnect. In *Super Computing* (2005).

[51] SHANG, L., PEH, L.-S., AND JHA, N. K. Dynamic Voltage Scaling with Links for Power Optimization of Interconnection Networks. In *Proceedings of the 9th International Symposium on High-Performance Computer Architecture (HPCA)* (Anaheim, CA, January 2002).

[52] SHIN, D. K., AND KIM, J. Power-aware comunication optimization for networks-on-chips with voltage scalable links. In *CODESMASISSS'04: Proceedings of the 2nd IEEE/ACM/IFIP international conference on hardware/software codesign and system synthesis* (2004), pp. 170–175.

[53] SOTERIOU, V., EISLEY, N., AND PEH, L.-S. Software-directed power-aware interconnection networks. In *Proceedings of the International Conference on Compilers, Architecture and Synthesis for Embedded Systems (CASES)* (San Francisco, September 2005).

[54] SOTERIOU, V., AND PEH, L.-S. Dynamic Power Management for Power Optimization of Interconnection Networks Using On/Off Links. In *Hot Interconnects 11* (Stanford University, Palo Alto CA, August 2003).

[55] SOTERIOU, V., AND PEH, L.-S. Design-space exploration of power-aware on/off interconnection networks. In *Proceedings of the 22nd International Conference on Computer Design (ICCD'04)* (San Jose, October 2004), pp. 510–517.

[56] STINE, J. M., AND CARTER, N. P. Comparing adaptive routing and dynamic voltage scaling for link power reduction. *Computer Architecture Letters* (June 2004).

[57] SUPERCOMPUTADOR BLUEGENE/P EN ARGONNE NATIONAL LABORATORY. Disponible en `http://www.alcf.anl.gov/`.

[58] SUPERCOMPUTADOR JAGUAR. Disponible en `http://www.nccs.gov/computing-resources/jaguar`.

[59] SUPERCOMPUTADOR JUGENE. Disponible en `http://www.fz-juelich.de/jsc/jugene`.

[60] SUPERCOMPUTADOR KRAKEN. Disponible en `http://www.nics.tennessee.edu/computing-resources/kraken`.

[61] SUPERCOMPUTADOR RANGER. Disponible en `http://www.tacc.utexas.edu/resources/hpc`.

[62] SUPERCOMPUTADOR REDSKY. Disponible en `http://www.sandia.gov`.

[63] SUPERCOMPUTADOR ROADRUNNER. Disponible en `http://www.lanl.gov/roadrunner`.

[64] SUPERCOMPUTADOR TIANHE.
Disponible en http://www.top500.org/system/10186.

[65] THE BLUEGENE/L TEAM. An Overview of the BlueGene/L Supercomputer. In *IEEE/ACM SC2002* (Baltimore, MD, November 2002). Available from http://sc-2002.org/paperpdfs/pap.pap207.pdf.

[66] TOP500 SUPERCOMPUTER SITES. 2010.
Disponible en http://www.top500.org.

[67] VETTER, J., AND MUELLER, F. Communication characteristics of large-scale scientific applications for contemporary cluster architectures. In *Proc. International Parallel and Distributed Processing Symposium (IPDPS)* (2002).

[68] ZAMANI, R., AFSAHI, A., QIAN, Y., AND HAMACHER, C. A feasibility analysis of power-awareness and energy minimization in modern interconnects for high-performance computing. In *CLUSTER '07: Proceedings of the 2007 IEEE International Conference on Cluster Computing* (Washington, DC, USA, 2007), IEEE Computer Society, pp. 118–128.