REAL ACADEMIA DE CIENCIAS
EXACTAS, FÍSICAS Y NATURALES

APORTACIONES AL CONTROL DE CONGESTIÓN EN REDES DE INTERCONEXIÓN DE ALTA VELOCIDAD

DISCURSO LEÍDO EN EL ACTO DE SU RECEPCIÓN
COMO ACADÉMICO DE NÚMERO POR EL

EXCMO. SR. D. JOSÉ FRANCISCO DUATO MARÍN

Y CONTESTACIÓN DEL

EXCMO. SR. D. JAVIER JIMÉNEZ SENDÍN

EL DÍA 13 DE NOVIEMBRE DE 2019

MADRID
Domicilio de la Academia
Valverde, 22

ISSN: 0214-9540
ISBN: 978-84-87125-69-0
Depósito legal: M-33587-2019

Índice general

DISCURSO DE INGRESO

DEL

EXCMO. SR. D. JOSÉ FRANCISCO DUATO MARÍN

Capítulo 1

Introducción y agradecimientos

Excmo. Señor Presidente de la Real Academia de Ciencias, Excmas. Señoras y Señores Académicos, señoras y señores.

Mis primeras palabras son para agradecer a todos los miembros de la Real Academia de Ciencias que me hayan aceptado en esta prestigiosa institución. Quiero agradecer muy especialmente a los Excmos. Sres. D. Manuel López Pellicer, D. Jesús María Sanz Serna y D. Javier Jiménez Sendín que propusieran mi candidatura para la Medalla número 5. Un agradecimiento muy sentido, pues los tres han sido clave para mi ingreso en la Real Academia de Ciencias. A D. Jesús María Sanz Serna debo agradecerle que iniciara el procedimiento para proponer mi candidatura, así como sus continuos ánimos a lo largo de todo el proceso. A D. Javier Jiménez Sendín debo agradecerle el enorme interés mostrado por mis investigaciones relacionadas con el enrutamiento adaptativo en el supercomputador IBM BlueGene y, sobre todo, que haya aceptado la tarea de contestar a mi discurso. Y a D. Manuel López Pellicer quiero expresarle mi más profundo agradecimiento por su incondicional apoyo en todo momento. Desde que propuso mi candidatura para cubrir una plaza de académico correspondiente nacional en la Sección de Exactas, ha sido mi mentor y ha guiado todos mis pasos en la Real Academia de Ciencias. Es un extraordinario ejemplo a seguir, como persona y como científico. Como persona, su dedicación incondicional a los demás es una referencia. No en vano fue elegido defensor de la comunidad universitaria en la Universidad Politécnica de Valencia. Y como científico, me maravilla su afán por descubrir nuevos resultados y por completar algunos aspectos de la obra de su maestro D. Manuel Valdivia Ureña. Finalmente, quiero agradecer a D. Jesús María Sanz Serna y a D. José Bonet Solves la infinita paciencia que han tenido conmigo, mientras modificaba una u otra vez este discurso de ingreso para incluir los últimos

avances que estamos desarrollando en colaboración con la industria, llegando incluso a reescribir buena parte del mismo.

A lo largo de mi carrera como investigador he colaborado con muchas personas. Creo en el trabajo en equipo y estoy convencido de que los equipos de investigación de gran tamaño tienen la capacidad de abordar proyectos de mayor envergadura e interdisciplinaridad. Desde hace décadas, el Grupo de Arquitecturas Paralelas de la Universitat Politècnica de València ha venido colaborando con grupos de investigación de otras universidades, que en la actualidad incluye la Universidad de Castilla - La Mancha, la Universidad de Murcia, la Universitat de València, la Universidad Miguel Hernández y la Universitat Jaume I. La lista de personas es demasiado larga para nombrarlos a todos. A todos ellos, mi más sincero agradecimiento. También han sido muchas las colaboraciones internacionales, algunas tan intensas que he llegado a hacer verdaderos amigos en otros países. De entre ellos quiero estacar a Timothy Pinkston, Olav Lysne, Sudhakar Yalamanchili, DK Panda y Mitch Gusat. Finalmente, quiero expresar mi más sentido agradecimiento a mi familia, que me ha ayudado y apoyado en todo momento.

Al provenir de campos científicos tan distintos, mi interacción con mi antecesor en la Medalla número 5 de la Real Academia de Ciencias, el Excmo. Sr. D. Darío Maravall Casesnoves, fue muy reducida, limitándose prácticamente al acto de investidura como Doctor Honoris Causa por la Universidad Politécnica de Valencia el 3 de Junio de 1997. Pero sí que pude sentir, a través de las intervenciones de sus antiguos discípulos en el homenaje que se celebró el día 23 de Mayo de 2018 en el Instituto de la Ingeniería de España, la gran humanidad que desprendía en todas sus actuaciones. Y también pude aprender sobre la enorme amplitud y profundidad de sus resultados científicos. Sus numerosos trabajos sobre teoría de oscilaciones me resultaron fascinantes. Me recordaron las oscilaciones que se producen en un sistema de control de congestión de extremo a extremo. La intuición me dice que si D. Darío hubiese conocido este complejo problema, habría sido capaz de encontrar una solución que permitiera eliminar dichas oscilaciones mediante un adecuado ajuste de los parámetros del control de congestión.

Según comentaron algunos de sus discípulos durante el homenaje, D. Darío era admirado, querido y respetado por todos sus compañeros y alumnos. D. Darío disfrutaba de una buena charla ya fuera de ingeniería, matemáticas, política o filosofía con una naturalidad y fluidez al alcance de solo unos pocos elegidos. Su ímpetu y su curiosidad hicieron que fuera un estudioso de casi todas las ciencias. "¿De qué no sabrá Don Darío?" se preguntaban sus alum-

nos y amigos, mostrando la fascinación que era capaz de despertar entre sus allegados y homólogos.

Una faceta que me resulta sorprendente de D. Darío, por contraposición a mi propia forma de actuar, es su habilidad para encontrar inspiración en la resolución de problemas prácticos para desarrollar complejas teorías. Mientras que mi actuación se limita a desarrollar investigaciones motivadas por problemas reales con el objetivo de resolver dichos problemas, habiendo sido capaz en el mejor de los casos de generalizar la solución encontrada a un número más amplio de sistemas, D. Darío iba mucho más allá y desarrollaba teorías que trascendían el problema concreto que las inspiró. Tal es el caso de unos trabajos prácticos encargados por la Cámara Agrícola de Madrid, relativos al catastro, en los cuales encontró inspiración para obtener analogías matemáticas entre las oscilaciones de relajación en series estadísticas temporales y la variación del número de "huecos" en la pantalla fotoeléctrica del iconoscopio de Zworykin.

Sin duda, un elemento clave para muchos de sus logros fue su interés por adoptar una perspectiva filosófica en el estudio de todos los problemas científicos y técnicos que abordó en su dilatada vida, por muy especializados y abstractos que fueran, tales como las que denominó oscilaciones teleológicas, que descubrió y desarrolló en sus estudios sobre poblaciones biológicas y en el análisis de sistemas económicos complejos.

Mi discurso versa sobre un problema cuya solución es crucial para el desarrollo de servicios inteligentes con respuesta en tiempo real, los cuales empiezan a estar disponibles en la nube y accesibles a través de Internet. En la primera parte, se describe el contexto y las tendencias actuales, tanto en el mundo de la supercomputación como de los servidores de Internet. En la segunda parte, se describe y analiza el problema de la congestión en las redes de comunicaciones. También se presentan soluciones eficientes, a varias de las cuales he contribuido, con un nivel comprensible para los científicos de otras ramas, así como una lista de problemas abiertos con algunas pinceladas sobre cómo atacarlos. Y en la tercera parte hablo sobre ciencia, transferencia de conocimiento y sociedad, presentando mi visión personal sobre una de las vías que los científicos podemos escoger para servir a la sociedad a la que nos debemos.

Capítulo 2

Descripción del contexto

Hace unos pocos siglos, una persona no percibía a lo largo de su vida ninguna evolución en la sociedad ni en la tecnología empleada. Hoy en día, la evolución tecnológica es tan rápida que podemos percibir cambios bastante significativos de cada década a la siguiente. Y esos avances tecnológicos están introduciendo cambios apreciables en el comportamiento de las personas, sus hábitos y su forma de relacionarse.

Varios son los frentes en los que la tecnología nos ha abierto nuevas oportunidades y ha cambiado nuestra forma de actuar. Tal es el caso de la generación de energía, los medios de transporte, las tecnologías para el procesado de los alimentos, la atención sanitaria, las comunicaciones y la educación, entre otros. La introducción de muchas de estas tecnologías ha marcado hitos en la historia de la Humanidad, tales como el tren, el automóvil, el avión, la energía eléctrica, el teléfono y los ordenadores personales de cualquier tipo y tamaño. Algunas de estas tecnologías, tras décadas o siglos de progreso, han experimentado un cierto estancamiento. Tal es el caso del avión, limitado en su versión comercial a velocidades subsónicas por motivos de coste y contaminación, o del automóvil, limitado en su velocidad máxima de circulación por vías públicas por motivos de seguridad, contaminación y coste. Otras tecnologías han desaparecido, al haber sido superadas por otras tecnologías más potentes, flexibles y/o económicas, como es el caso de la película fotográfica.

Sin embargo, la evolución de las tecnologías de la información y de las comunicaciones durante las últimas décadas ha sido absolutamente espectacular. Desde un punto de vista tecnológico, la clave de ese progreso tan espectacular ha sido la fabricación de transistores cada vez más pequeños, con diseños que han permitido la codificación de información en señales que requerían cada vez menos potencia para su almacenamiento y su procesamiento. Pero ese

progreso tan espectacular no hubiese sido posible sin la enorme aceptación que han tenido los dispositivos de cálculo y de comunicaciones. Dicha aceptación ha dado lugar a una realimentación positiva que ha acelerado el progreso hasta límites impensables en las etapas iniciales de la evolución de estas tecnologías. Es famosa la frase de 1943 «I think there is a world market for maybe five computers», atribuida a Thomas J. Watson, entonces presidente de IBM, si bien no hay evidencias directas de que la pronunciara. Un mayor número de ventas ha permitido mayores beneficios brutos y mayor inversión por parte de las empresas en investigación y desarrollo de nuevos o mejores productos, lo que ha permitido precios más reducidos y un nuevo incremento en las ventas, cerrando así el ciclo. Es más, cada vez que la tecnología parecía estancarse o los consumidores se sentían satisfechos con los dispositivos existentes en el mercado, los fabricantes han recurrido a la miniaturización y a incorporar nueva funcionalidad para mantener, e incluso incrementar, las ventas. Tal ha sido el caso de la sucesiva introducción en el mercado de los ordenadores personales, los portátiles, las tabletas y los teléfonos móviles actuales.

Pero las ventas antes mencionadas no se han limitado al consumo directo de productos por parte del usuario final. De hecho, durante décadas, la mayor parte de las ventas han procedido del uso de dispositivos de cálculo, de almacenamiento o de comunicaciones como parte integrante de sistemas más complejos. Como ejemplo cabe citar el uso de microcontroladores en los vehículos a motor. Un automóvil de fabricación reciente contiene decenas de microcontroladores, conectados entre sí, para gobernar los diferentes componentes que lo constituyen.

Por otra parte, las tecnologías de la información han sido un instrumento esencial para el progreso acelerado en muchas ramas de la ciencia. La disponibilidad de supercomputadores cada vez más potentes y más accesibles ha permitido simular sistemas complejos o de grandes dimensiones en tiempos razonables. Como ejemplos, podemos enumerar la reconstrucción del genoma de diferentes especies, la simulación de colisiones entre galaxias, la simulación del plegado de proteínas al introducirse en un medio acuoso o la simulación de la evolución global del clima en la Tierra. En muchos de estos casos, la demanda de mayores potencias de cálculo no deja de crecer, con objeto de poder simular sistemas de mayor tamaño, analizar la evolución durante un periodo de tiempo más largo o conseguir una mayor resolución o precisión de los resultados.

También la industria demanda cada vez mayores potencias de cálculo. La exploración del subsuelo para buscar nuevas bolsas de petróleo y gas se hace mediante la simulación de complejos modelos a partir de las ondas sísmicas

medidas tras producir múltiples explosiones controladas. Los ensayos de deformación de prototipos de carrocerías de automóviles tras una colisión han sido sustituidos por simulaciones, permitiendo ensayar muchos más prototipos por unidad de tiempo y con un menor coste. Los cálculos de la aerodinámica y de la resistencia estructural del fuselaje de los aviones se realizan mediante complejas simulaciones, permitiendo niveles de optimización antes impensables. Incluso se han diseñado simuladores tan detallados que replican a la perfección el funcionamiento individual de cada motor de avión. A partir de los datos medidos por un gran número de sensores ubicados en un motor real, su "gemelo virtual" reproduce fielmente su evolución y es capaz de predecir cuándo va a fallar y qué componente va a fallar. Las diversas aplicaciones bioinformáticas, la mayoría de las cuales comparan fragmentos de cadenas de aminoácidos con un genoma de referencia para obtener algún tipo de conclusión, también requieren enormes potencias de cálculo debido al tamaño del genoma de las diferentes especies. Estos son solo algunos ejemplos, pero son muchas las aplicaciones industriales de los supercomputadores.

Pero a pesar de la gran relevancia que tienen los supercomputadores y las aplicaciones científicas, comerciales y de defensa que en ellos se ejecutan, la mayor parte de las ventas de dispositivos de cálculo y comunicaciones provienen del mercado de consumo. Por ello, durante las últimas décadas las grandes empresas del sector han centrado sus diseños y productos en dicho mercado, y en él tienen puestas sus mayores expectativas. De hecho, tanto las líneas de investigación como los temas de interés han estado muy influidos por la evolución tecnológica y por la demanda del mercado ante la sucesiva introducción de nuevos productos. Hace unas décadas, cuando se introdujeron en el mercado los dispositivos de cálculo personales, la evolución de estos dispositivos y la de los grandes computadores de aquella época caminaban por separado. Hoy en día, no se concibe un dispositivo personal que no esté conectado a una variedad de redes de datos. Esta conexión ha abierto un sinfín de oportunidades y continuamente están apareciendo nuevas aplicaciones que han llegado a cambiar los hábitos sociales, especialmente entre las generaciones más jóvenes. Pero para almacenar y procesar los datos solicitados hacen falta computadores muy potentes, lo que ha unido de forma muy estrecha la evolución de los pequeños dispositivos de cálculo y la de los grandes computadores. No obstante, siguen existiendo diferencias significativas entre los computadores diseñados para conseguir muy altas prestaciones y los diseñados para funcionar como servidores para aplicaciones que acceden a los mismos a través de algún tipo de red, por lo que los analizaremos por separado.

2.1. Evolución de los supercomputadores

En la década de los noventa se produjo un punto de inflexión en la evolución de los sistemas de computación de altas prestaciones. Los microprocesadores, que en aquella década experimentaron el mayor incremento relativo de prestaciones que se ha podido observar en su historia [Danowitz12], permitieron un cambio radical en el diseño de los supercomputadores. En lugar de interconectar un número reducido de procesadores muy potentes, los diseñadores empezaron a considerar arquitecturas de sistema basadas en interconectar un número elevado de microprocesadores [Becker95, Kessler93]. Estos dispositivos tenían una potencia de cálculo reducida, pero conectando un gran número de ellos se podían alcanzar potencias de cálculo superiores a las de los supercomputadores más potentes de la época. Pero lo que realmente motivó el cambio fue la reducción en el coste de las máquinas. Un supercomputador basado en microprocesadores tenía un coste de fabricación al menos diez veces inferior al de un supercomputador tradicional de prestaciones similares. Esto fue posible gracias al reducido precio de los microprocesadores, consecuencia directa del elevado volumen de ventas de estos dispositivos, ya que se trataba del componente principal de los ordenadores de sobremesa y portátiles.

La interconexión de un número tan elevado de microprocesadores planteó inicialmente algunos problemas, pues las tecnologías de las redes de área local de aquella época, tales como Fast Ethernet, no estaban diseñadas para esta finalidad. Pero esos problemas fueron fácilmente salvables, sin más que diseñar redes de interconexión especialmente adaptadas a las necesidades de los supercomputadores [Scott94, Scott96]. Muchas de las soluciones que se aplicaron en aquella época, tales como la conmutación segmentada o el control de flujo [Dally87, Dally92], ya existían antes de que surgiera esta nueva tendencia en el diseño de los supercomputadores (ver, por ejemplo, [Kermani79]). Otras tendencias, tales como la utilización de topologías de red con un reducido número de dimensiones [Dally90], como por ejemplo el toro o la malla de dos y tres dimensiones (ver Figura 3.1), fueron abandonadas pocos años después y reemplazadas por topologías de red con una conectividad mucho más elevada. Tal es el caso de la red de Beneš plegada, mostrada en la Figura 2.1, también denominada en algunos ámbitos como red de Clos, y más conocida popularmente como "árbol gordo" (fat tree en inglés) [Leiserson85].

Tradicionalmente, la tecnología desarrollada para las redes de interconexión ha ido por delante de las necesidades de comunicación de los microprocesadores, gracias en buena medida al esfuerzo de los programadores por desarrollar programas paralelos que minimizaran la comunicación entre procesos que se ejecutan en diferentes procesadores de un supercomputador. Sin embar-

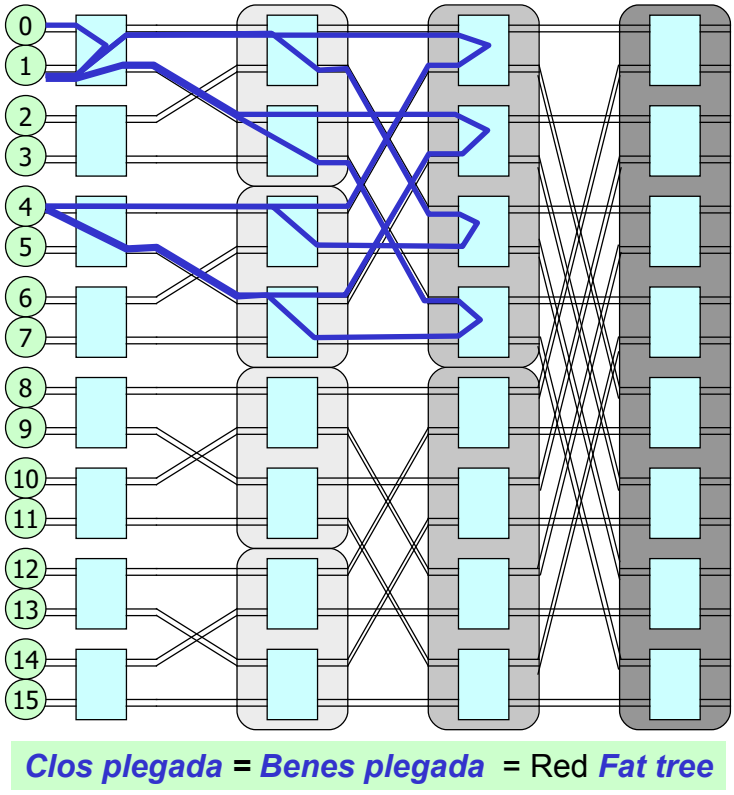


Figura 2.1: Topología “fat tree”, también denominada red de Clos plegada o red de Beneš plegada.

go, la continua demanda de mayores potencias de cálculo, acrecentada durante la última década por la carrera hacia el Exaflop (un trillón de operaciones en coma flotante por segundo) [Kogge08], ha llevado a diseños de supercomputadores con un número de procesadores cada vez mayor, existiendo ya diseños con cientos de miles de nodos de procesamiento. La interconexión de un número tan elevado de procesadores está creando nuevos problemas en las redes de interconexión, hasta ahora inexistentes, tales como la congestión de la red.

La mayor dificultad encontrada hasta el momento para construir supercomputadores tan potentes radica en el consumo de energía eléctrica. Se podría haber construido ya un supercomputador con más de un Exaflop, pero hubiera consumido una potencia cercana a la producida por una central nuclear. Con objeto de diseñar supercomputadores más eficientes energéticamente, muchos de los diseños desarrollados durante la última década incluyen aceleradores es-

pecializados, que sólo pueden usarse para determinados cálculos, pero son más eficientes energéticamente [Green500]. El tipo de acelerador más popular es el derivado de las tarjetas gráficas de gama alta [Subramaniam13], comúnmente designados con las siglas GPU. Sin embargo, la utilización media de estos caros dispositivos aceleradores es bastante baja [Peña14], lo que está cuestionando su uso. Esta baja utilización proviene, en buena medida, de que solo pueden ser utilizados desde el procesador al que están directamente conectados. Una solución a este problema consiste en virtualizar los aceleradores y compartirlos, de modo que cualquier procesador pueda utilizar cualquier acelerador, independientemente de donde esté ubicado. Esta solución, propuesta originalmente en [Duato10] y desarrollada por mi equipo de investigación para las GPUs con interfaz de programación CUDA, permite acceder a cualquier GPU desde cualquier procesador a través de la red de interconexión. Esto permite asignar tantas GPUs como haga falta a una aplicación determinada, así como aumentar la productividad global del sistema y reducir el consumo global de energía eléctrica para una carga dada. Pero esta flexibilidad tiene un precio, ya que el acceso a aceleradores remotos requiere una red de interconexión con un tiempo de comunicación de extremo a extremo, también denominado latencia, del orden de un microsegundo o inferior. Además, el acceso a aceleradores remotos incrementa notablemente el caudal de tráfico por la red, lo que puede agravar los problemas de congestión.

2.2. Evolución de la sociedad digital

Como ya se ha indicado, los supercomputadores son máquinas especializadas que representan una pequeña fracción de los dispositivos de computación que existen en el mundo. Lo que realmente ha dinamizado y seguirá impulsando la investigación y la innovación en el sector de las tecnologías de la información es la digitalización de la sociedad.

Para bien o para mal, nuestras vidas han cambiado para siempre gracias a la tecnología digital. El número de servicios y aplicaciones disponibles a través de Internet, en especial para dispositivos móviles, crece incesantemente. Muchas de estas aplicaciones son accedidas como servicio desde la nube [Zhang10, Regalado11]. Entre los más jóvenes, los servicios en la nube (en particular, las redes sociales y la mensajería multimedia) se están convirtiendo en una parte importante, si no imprescindible, de su vida natural. La interacción con los servicios en la nube se realiza cada vez más de forma humana y natural, a través de órdenes de voz y reconocimiento visual. Algún día, en un futuro no muy lejano, como predijo el futurista Ray Kurzweil [Kurzweil05], la

forma en que pensamos se verá aumentada por la nube. Hoy en día, los servicios en la nube se personalizan según nuestros gustos individuales a través de servicios de datos en línea intensivos. Nos estamos acostumbrando a tener acceso casi instantáneo a cantidades masivas de contenido digital como respuesta a nuestras órdenes de voz.

Pero el concepto de la nube requiere, para poder funcionar, que existan un enorme número de servidores de Internet, también denominados centros de datos, para poder almacenar de forma redundante y procesar en tiempo real las ingentes cantidades de datos a los que habitualmente accedemos, y que siguen creciendo cada día a un ritmo vertiginoso. Para 2019, se prevé que un solo usuario generará 1,6 Gigabytes de almacenamiento en la nube por mes, en comparación con los 992 Megabytes por mes en 2014. Y los datos creados por los dispositivos que conforman la Internet de las Cosas se prevé que alcanzarán los 507,5 Zettabytes por año para 2019, lo que supone un gran aumento respecto a los 134,5 Zettabytes por año en 2014 [Kleyman16]. Es más, para poder conseguir que cantidades masivas de datos pueden convertirse en información útil dentro del escaso margen de tiempo que permite la interacción con una inteligencia humana en tiempo real, los centros de datos utilizan tecnologías similares a las utilizadas en los supercomputadores. Aprovechando la enorme potencia de cálculo de los procesadores actuales y la capacidad de almacenamiento de los discos actuales, se construyen granjas de procesadores que pueden alcanzar decenas e incluso centenares de miles de dispositivos interconectados.

Sin embargo, no basta con incrementar el número de dispositivos de cálculo y el de unidades de almacenamiento de información, ya que al aumentar el número de dispositivos también aumentan tanto la probabilidad de que falle algún componente como los retardos en la propagación de la información, lo que dificulta enormemente la consecución de resultados en tiempo real. La respuesta de las empresas líderes en el mercado no se ha hecho esperar. El requisito de integrar la tecnología digital en nuestras vidas naturales está impulsando la innovación en los centros de datos. Esta innovación responde a la necesidad de nuevos niveles de rendimiento, escalabilidad y fiabilidad de la infraestructura. El requisito de respuesta en tiempo real se traduce en que la entrega de información por parte del centro de datos en la nube debe ser muy rápida. Y esos requisitos tan estrictos implican, a su vez, que la red de interconexión que interconecta los dispositivos en un centro de datos debe ser muy rápida, consiguiendo una latencia de las comunicaciones muy baja, incluso cuando tiene que transmitir enormes cantidades de datos por unidad de tiempo.

2.3. Tendencias en el centro de datos

Con objeto de poder procesar la enorme cantidad de complejas peticiones que llegan a un centro de datos cada segundo, respondiendo a cada una de ellas en tan solo unas decenas de milisegundo, los diseñadores están introduciendo cuatro estrategias innovadoras en los centros de datos. En primer lugar, se están rediseñando las aplicaciones para que puedan utilizar de forma eficiente múltiples recursos de cálculo y almacenamiento concurrentemente, al tiempo que se garantiza un tiempo de respuesta acotado. En segundo lugar, se están diseñando e implantando aceleradores de cálculo específicos para ejecutar determinadas tareas, de modo que se incrementa notablemente la potencia de cálculo al ejecutar dichas tareas al tiempo que se mejora la eficiencia energética del centro de datos. En tercer lugar, se están rediseñando los subsistemas de almacenamiento para aprovechar eficientemente las nuevas tecnologías de almacenamiento de estado sólido sin incrementar excesivamente el coste total del sistema. Finalmente, se están rediseñando las redes de interconexión que posibilitan la comunicación y sincronización entre los diferentes componentes, pasando de redes con pérdida de paquetes a redes sin pérdidas para poder garantizar una latencia reducida y acotada.

De estas cuatro líneas de innovación, nos vamos a centrar en los cambios introducidos en la red de interconexión. Este subsistema es un elemento clave de los centros de datos, debiendo ser capaz de interconectar un elevado número de dispositivos, proporcionando elevados anchos de banda para poder transmitir grandes caudales de tráfico por la red y consiguiendo tiempos de comunicación muy reducidos para soportar eficientemente las aplicaciones interactivas. Al igual que en el caso de los supercomputadores, la red de interconexión puede sufrir situaciones de elevada congestión, que se produce cuando se transmiten grandes caudales de tráfico por la red. Cabría pensar que la congestión se produce porque la red de interconexión no está correctamente dimensionada, pero no es así. Por una parte, el tráfico generado por las aplicaciones no se distribuye uniformemente a lo largo del tiempo, por lo que para atender adecuadamente los picos de tráfico habría que sobredimensionar excesivamente la red, con el consiguiente coste y consumo de energía. Y por otra, la aplicaciones pueden generar patrones de tráfico que requieren que múltiples fuentes envíen simultáneamente información a un mismo destino, haciendo imposible absorber dicho caudal de tráfico independientemente del diseño de la red.

Seguidamente vamos a analizar estas situaciones mediante dos casos de uso. Por su mayor relevancia económica, se han escogido casos de uso relacionados con los centros de datos. Estos casos de uso [Congdon18] versan sobre

tecnologías que han empezado muy recientemente a introducirse en el mercado, y muestran claramente la necesidad de diseñar redes de interconexión de muy altas prestaciones. En particular, muestran situaciones que dan lugar a congestión muy severa en la red, por lo que se hace imprescindible atacar este problema, aplicando soluciones eficientes. El primer caso de uso abarca los diferentes servicios en línea intensivos en datos (en inglés, On-Line Data Intensive, o OLDI) a gran escala, tales como los sistemas de recomendación automatizados para compras en línea, redes sociales y búsqueda en la web. El segundo incluye las numerosas aplicaciones basadas en una evolución de las tradicionales redes neuronales, las denominadas redes de aprendizaje profundo de alto rendimiento.

2.3.1. Servicios en línea intensivos en datos (OLDI)

La diferencia fundamental entre los servicios en línea intensivos en datos y sus equivalentes fuera de línea es que requieren respuestas casi inmediatas a las solicitudes que llegan con una cadencia muy elevada. El control de la latencia de respuesta es fundamental en estos servicios. La experiencia del usuario final depende en gran medida de la capacidad de respuesta del sistema, e incluso las demoras moderadas de menos de un segundo pueden tener un impacto medible en las consultas individuales y sus ingresos publicitarios asociados. Una gran parte de la demora, inevitable debido a los límites impuestos por la velocidad de la luz, se debe a las comunicaciones a través de Internet para acceder a un sistema remoto, integrado en la nube, que va a ser la fuente de decisión e información. Esto impone aún más presión sobre las demoras máximas admisibles dentro del centro de datos en sí.

Para abordar estos problemas de latencia, los servicios OLDI despliegan cada solicitud individual sobre miles de procesadores simultáneamente. Las respuestas de estos procesadores se coordinan y agregan para formar las mejores recomendaciones o respuestas. Pero los tiempos de obtención de estas respuestas son variables y se ven agravados por los flujos de comunicación rezagados a causa de la congestión en la red. Los estudios han demostrado que la red de interconexión se ha convertido en un componente significativo de la latencia total del centro de datos cuando se produce congestión en la red [Kapoor12]. Esto crea una distribución estadística de cola larga para la latencia, tanto peor cuanto más se paraleliza el cálculo. Para acotar la latencia de las respuestas, los procesadores a menudo se organizan en una jerarquía, como se muestra en la Figura 2.2, con plazos estrictos en cada nivel para producir una respuesta. Si los datos llegan tarde debido a la latencia en la red, dichos datos simplemente se descartan y se devuelve una respuesta parcial o subóptima.

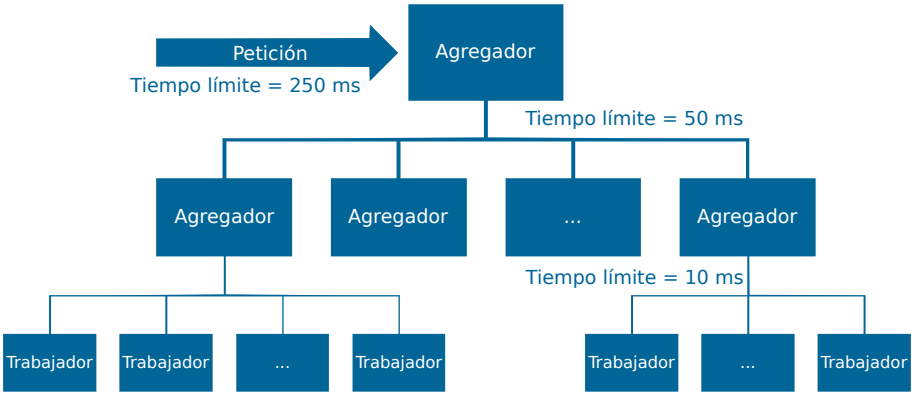


Figura 2.2: Jerarquía de procesadores para conseguir respuesta en tiempo real ante una petición compleja

La larga cola de la distribución de la latencia en los centros de datos suele ser causada por un par de factores [Jalaparti13]. El primero está relacionado con la mezcla de tráfico entre los mensajes de control (cortos) y los de datos (largos). Si bien la mayoría de los mensajes en el centro de datos son cortos, la mayor parte de la información transferida a través de la red se debe a los mensajes largos. Por lo tanto, un pequeño número de flujos de mensajes largos puede retrasar los mensajes de control, demorando el funcionamiento de todo el sistema.

Debido a que los centros de datos OLDI lanzan el procesamiento de cada solicitud a miles de procesadores simultáneamente, la segunda causa de las largas latencias se debe a que los nodos de procesamiento devuelven sus respuestas a un padre común en la jerarquía casi al mismo tiempo. Esto puede causar situaciones de elevada congestión en la red. Hay que tener en cuenta que la organización jerárquica no sólo permite paralelizar la agregación de resultados sino que también distribuye mejor el tráfico por la red, tanto en el espacio como en el tiempo. Esta mejor distribución del tráfico disminuye la severidad de las situaciones de congestión, pero no es capaz de eliminarlas completamente, ya que el tiempo de respuesta acotado impone un límite en el número de niveles de la jerarquía.

2.3.2. Aprendizaje profundo

El aprendizaje profundo (en inglés, deep learning) [Deng14] es una rama del aprendizaje automático (en inglés, machine learning) que está teniendo un enorme éxito, ya que permite resolver problemas muy complejos sin tener que

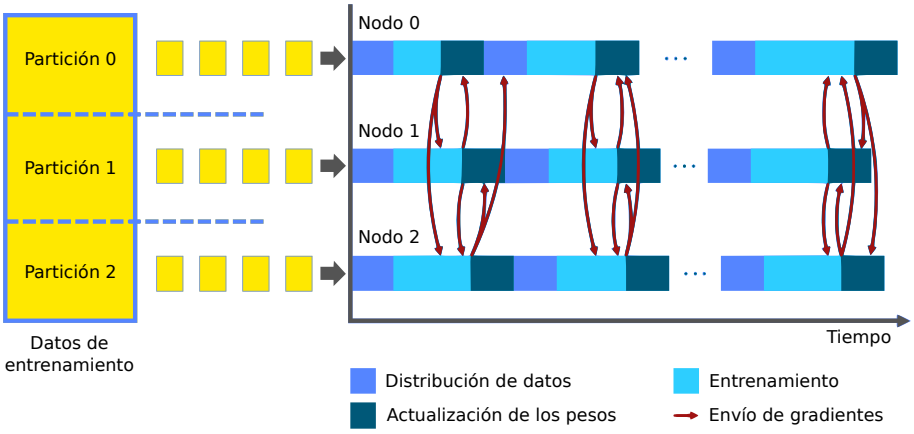


Figura 2.3: Entrenamiento de una red neuronal usando paralelismo de datos

desarrollar soluciones específicas para ello. Los sistemas de aprendizaje actuales utilizan redes neuronales muy grandes y con muchas capas, requiriendo el entrenamiento de millones, e incluso miles de millones de parámetros en algunos casos. Estas redes neuronales ejecutan tareas humanas cotidianas, como el reconocimiento de voz y de imágenes, y pueden integrarse en un servicio en línea. De hecho, complementan muy bien los servicios OLDI, al permitir que las aplicaciones y servicios que se ejecutan en la nube puedan ver y escuchar. Algunas tareas complejas, como el filtrado en las redes sociales y la detección de fraude y anomalías, se realizan sin esfuerzo una vez que se entrenan estas redes neuronales, de modo análogo a como lo haría un cerebro con sus millones de interconexiones neuronales. En general, cuanto más grande sea la red neuronal, mejor podrá desarrollar su trabajo.

Las redes neuronales actuales para aprendizaje profundo pueden tener miles de millones de parámetros y millones de interconexiones [Dean12]. El ajuste de dichos parámetros para que la red realice una determinada función es un proceso iterativo llamado entrenamiento o aprendizaje. Para conseguir un aprendizaje permanente en tiempo real, el aprendizaje se implanta como una aplicación altamente paralela que requiere baja latencia y alto rendimiento. Dedicar más recursos de cálculo al problema puede mejorar el tiempo que conlleva crear y ajustar un modelo. Sin embargo, la sobrecarga de comunicación involucrada en la aplicación paralela puede contrarrestar las ganancias obtenidas al utilizar más procesadores o más aceleradores. Como se ve en la Figura 2.3, los enormes conjuntos de datos de entrenamiento se dividen en fragmentos y se distribuyen a través de una serie de grupos de trabajo. Cada

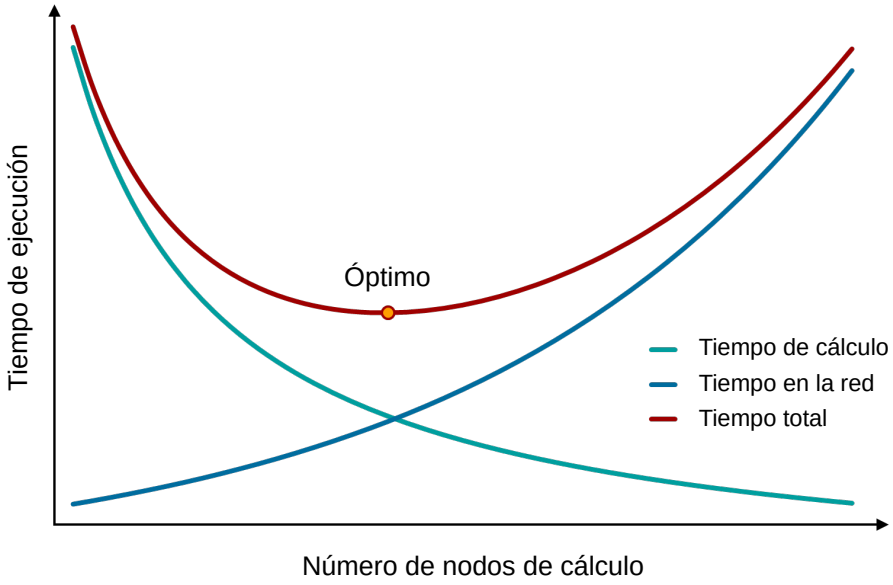


Figura 2.4: Configuración óptima del número de nodos

procesador procesa fragmentos de datos separados y devuelve los resultados del gradiente para que un servidor de parámetros común los actualice de forma coordinada. El proceso se repite, afinando y redistribuyendo los parámetros del modelo hasta que éste pueda reconocer una entrada conocida con un nivel aceptable de precisión. Una vez que se han creado y ajustado los modelos, se pueden distribuir y usar como parte de un nuevo tipo de servicio OLDI que requiera una entrada compleja, como voz, escritura a mano, imágenes de alta resolución o vídeo.

Como se ha indicado, cuando las aplicaciones requieren aprendizaje permanente en tiempo real, los modelos de aprendizaje profundo se entrenan constantemente. Cuando se utiliza un servidor de parámetros en el proceso de aprendizaje, existe un problema inherente de confluencia en la red. Los nodos de trabajo devuelven resultados de cálculo de gradientes al servidor de parámetros casi al mismo tiempo. Este escenario de confluencia crea congestión en el conmutador que conecta el servidor de parámetros al resto de la red y puede dar como resultado grandes retrasos de sincronización y que no se consiga un aprendizaje en tiempo real. Una mayor paralelización del problema solo complica la situación, ya que se requieren comunicaciones entre un mayor número de nodos de procesamiento, lo que multiplica el impacto de la congestión en la

red. La Figura 2.4 muestra que existe una configuración óptima del número de nodos que trabajan en paralelo, que consigue minimizar el tiempo que conlleva entrenar un modelo. Un mejor diseño de la red puede permitir un mayor número de nodos trabajando en paralelo para entrenar el modelo, lo que reduciría el tiempo total de ejecución.

Capítulo 3

El problema de la congestión

Una vez analizadas las tendencias en los centros de supercomputación y en los centros de datos, así como las necesidades de comunicaciones de las aplicaciones y las situaciones de congestión en la red que de ellas se derivan, voy a definir el problema de la congestión.

3.1. Modelo de red de interconexión

Una red de interconexión [Duato03] está constituida por conmutadores, enlaces e interfaces de red, en número arbitrario e interconectados según un determinado patrón que se denomina topología de la red. La Figura 3.1 muestra algunos ejemplos de topología de red. Los enlaces permiten la transmisión de información. Los conmutadores permiten seleccionar la ruta que debe seguir cada fragmento de información transmitida. La unidad de información a transmitir se denomina mensaje e incluye una cabecera que proporciona la información necesaria para establecer la ruta a seguir. Puede descomponerse en unidades más pequeñas, denominadas paquetes, que también incluyen una cabecera que permite establecer la ruta a seguir, así como un número de secuencia para reconstruir el mensaje original. Finalmente, las interfaces de red permiten conectar a la red los diferentes dispositivos que necesitan comunicarse y sincronizarse.

En la Figura 3.2 se representa un diagrama de bloques de un conmutador básico [Duato03]. Cuando se recibe un paquete por un enlace conectado a un puerto de entrada de un conmutador, se empieza a almacenar en una memoria asociada a dicho puerto. Dicha memoria suele estar organizada en múltiples colas. La cabecera del paquete entrante se procesa, determinando el puerto de salida del conmutador por el que debe retransmitirse dicho paquete y esta-

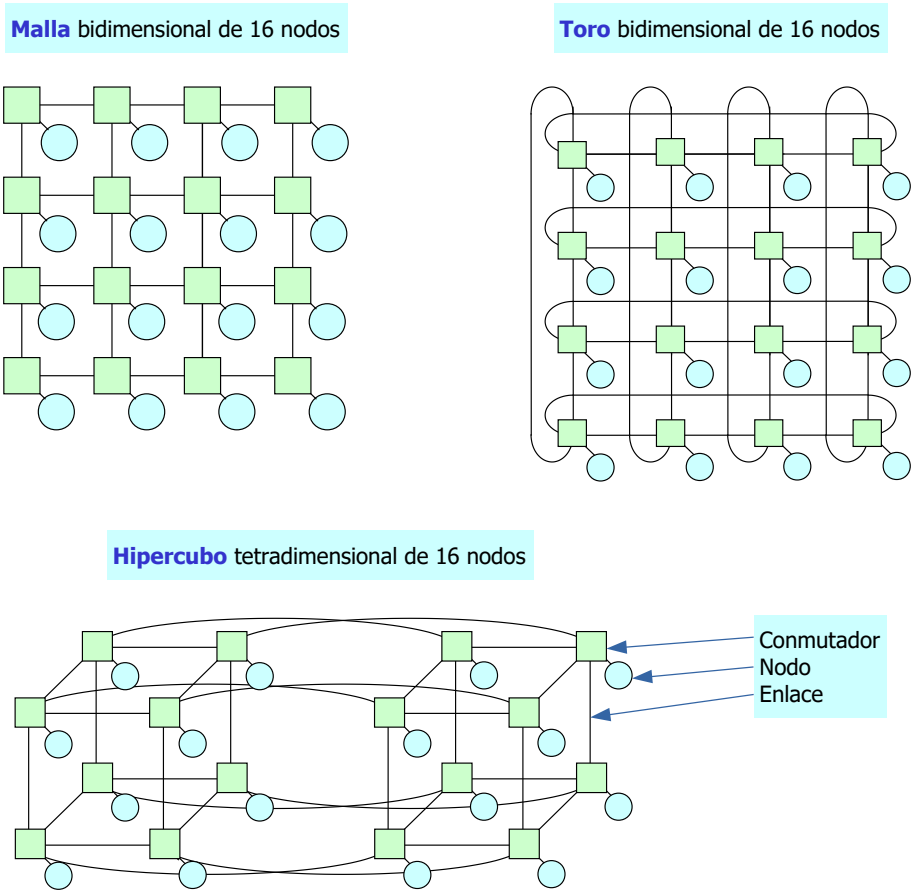


Figura 3.1: Ejemplos de topologías de red

bleciendo una conexión a través del conmutador interno entre los puertos de entrada y de salida correspondientes. Seguidamente, el paquete se transmite a la memoria asociada al puerto de salida y se planifica su transmisión por el siguiente enlace.

3.2. Contienda y congestión

Los recursos disponibles en la red de interconexión se comparten entre los diferentes paquetes que se transmiten concurrentemente. Esta compartición se realiza en el espacio y en el tiempo. En el espacio, diversos paquetes usan simultáneamente diferentes enlaces de comunicaciones de la red y/o se alma-

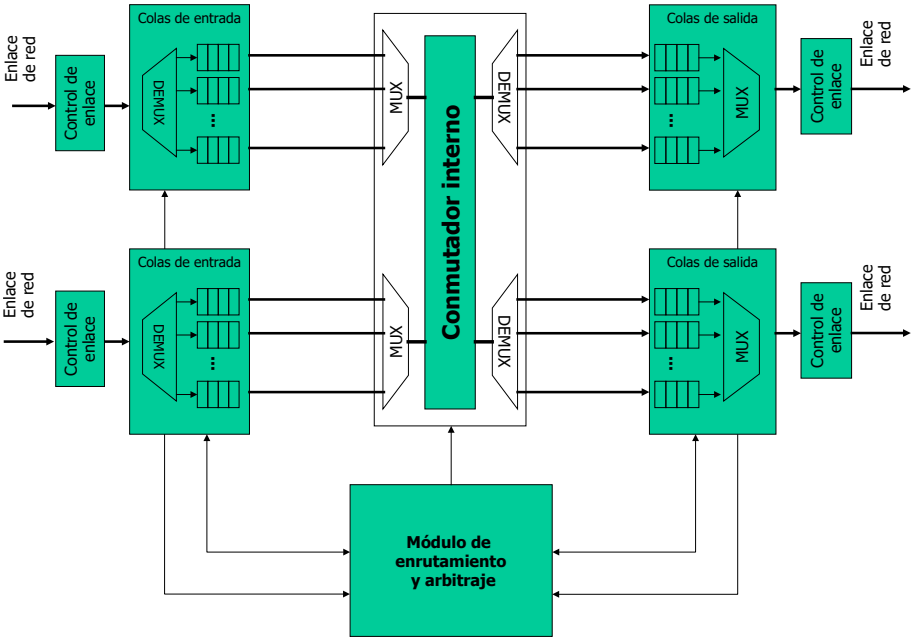


Figura 3.2: Diagrama de bloques de un conmutador

cenan en colas diferentes en los conmutadores. En el tiempo, un enlace de comunicaciones se utiliza sucesivamente por diferentes paquetes o fragmentos de paquetes.

Esta compartición de recursos da lugar a contiendas entre diferentes paquetes por el uso de los recursos. Dichas contiendas se resuelven mediante arbitraje [Duato03]. Cada vez que dos o más paquetes solicitan simultáneamente un mismo recurso, el árbitro correspondiente selecciona uno de ellos. El paquete ganador utiliza el recurso concedido mientras el resto de solicitantes esperan en la cola en la que están almacenados hasta que se libere el recurso y haya un nuevo arbitraje.

Cuando el tráfico de paquetes supera la capacidad de la red en alguna zona de la misma, la tasa de llegada de paquetes que solicitan un determinado recurso supera la tasa de salida de paquetes que ya han utilizado dicho recurso para avanzar hacia su destino. En este caso se produce congestión en esa zona de la red [García06a, García05b]. La congestión puede ser transitoria, debida a un desequilibrio de la carga en la red, o persistir durante bastante tiempo porque el tráfico inyectado en una zona de la red supera la capacidad de la misma. Si persiste, las colas dispuestas en los conmutadores para el almacenamiento de los paquetes en tránsito se llenan.

Tradicionalmente han existido dos estrategias para tratar el problema del llenado de las colas, que dan lugar a dos tendencias de diseño con comportamientos completamente distintos. La estrategia más sencilla consiste en descartar los paquetes que lleguen cuando la cola ya está llena. En este caso, la información contenida en dichos paquetes se pierde. Sin embargo, esta sencilla estrategia se ha combinado tradicionalmente con protocolos de comunicación de más alto nivel que notifican la correcta recepción de los paquetes a la fuente de la información. Dicha fuente mantiene una copia de todos los paquetes transmitidos hasta que recibe la correspondiente notificación. Si dicha notificación no se produce dentro de un intervalo de tiempo prefijado, la información se vuelve a transmitir.

Cabe pensar que si las colas de paquetes de un determinado conmutador ya estaban llenas, cuando lleguen los paquetes retransmitidos pueden seguir llenas, volviendo a descartar dichos paquetes. De hecho, así suele ocurrir. Por ello, la estrategia basada en el descarte de paquetes suele combinarse con un mecanismo de control de congestión que bruscamente reduce la tasa de inyección en las fuentes que empiezan a retransmitir paquetes, aumentando después dicha tasa de inyección de forma progresiva a lo largo del tiempo. Esta estrategia de descarte de paquetes, junto con los mecanismos de retransmisión de paquetes y de control de congestión asociados, es adecuada para comunicación a largas distancias, y se ha usado desde hace décadas en Internet y su famosa pila de protocolos de comunicación, habitualmente designados mediante las siglas TCP/IP (del inglés, Transmission Control Protocol / Internet Protocol) [Brakmo95].

El descarte de paquetes simplifica muchos aspectos del diseño de una red, tales como el dimensionado de las colas de los conmutadores, el diseño de algoritmos distribuidos para decidir la ruta que deben seguir los paquetes, la incorporación dinámica de nuevos dispositivos a la red y la supresión dinámica de aquellos componentes que se desconectan o fallan. A pesar de su sencillez, el descarte de paquetes tiene consecuencias nefastas para las prestaciones. Por una parte, la retransmisión de paquetes consume ancho de banda adicional en los enlaces de comunicaciones por los que circulan, precisamente en un escenario en el que el ancho de banda ya escaseaba y por eso se había llegado a producir congestión. Y por otra, los paquetes retransmitidos van a experimentar una latencia mucho mayor, ya que dicha latencia hay que contabilizarla desde la transmisión de la primera copia. Este incremento de latencia puede traducirse en una notable pérdida de prestaciones cuando los paquetes descartados forman parte de la comunicación entre dos procesos de una aplicación paralela. Por este motivo, la estrategia basada en el descarte de paquetes resul-

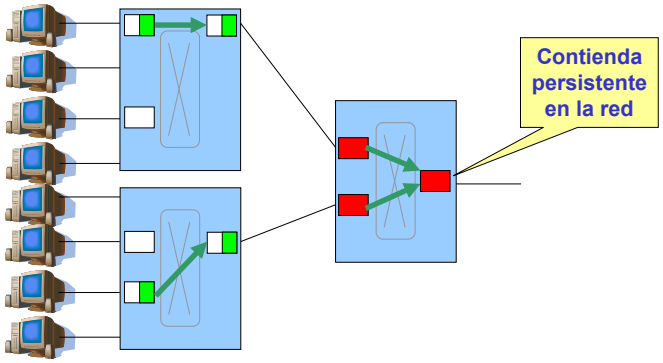
ta inaceptable en el diseño de la red de interconexión de un supercomputador. Los centros de datos sí que han usado tradicionalmente tecnologías basadas en el descarte de paquetes. Sin embargo, la reciente implantación de las aplicaciones paralelas antes descritas, en las que cada petición de usuario es procesada por miles de procesadores en paralelo, ha hecho que los centros de datos estén evolucionando hacia tecnologías sin descarte de paquetes [Congdon18]. Por todo ello, en adelante nos referiremos exclusivamente a sistemas sin descarte de paquetes.

Para evitar el descarte de paquetes se requiere un protocolo de control de flujo a nivel de enlace [Duato03], es decir, entre los conmutadores situados en ambos extremos de cada enlace de comunicaciones. Este protocolo permite al conmutador que recibe los paquetes notificar su disponibilidad de espacio al conmutador que los transmite. Hay varios modos de comunicar esta disponibilidad, desde la notificación cada vez que se libera espacio en una cola hasta la desactivación y reactivación del tráfico en función de la capacidad restante de la cola, como si de un semáforo se tratase. En cualquier caso, el diseño del protocolo de control de flujo debe ser tal que nunca se transmita un paquete si no hay espacio suficiente para poderlo almacenar en el conmutador que debe recibirlo.

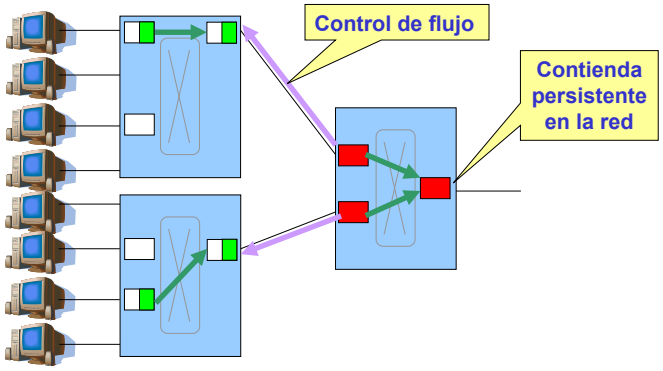
Cuando se utiliza un protocolo de control de flujo, no se descartan paquetes por falta de capacidad en las colas. No se requieren retransmisiones por descarte de paquetes y, como consecuencia de ello, mejoran mucho las prestaciones del sistema. Sin embargo, el control de flujo complica mucho el funcionamiento de la red cuando se produce congestión. En concreto, cuando se llena una cola, el protocolo de control de flujo de la cola que se ha llenado notifica al conmutador que hay al otro extremo del enlace que no envíe más paquetes. Como dicho conmutador no puede transmitir el tráfico de paquetes que le está llegando, sus colas también acabarán llenándose. Así pues, si la congestión perdura, ésta se propaga a los conmutadores vecinos, y de éstos a sus vecinos, llegando eventualmente hasta los procesadores que están inyectando el tráfico en la red e impidiendo o limitando dicha inyección [García05b].

3.3. Árboles de congestión y su evolución

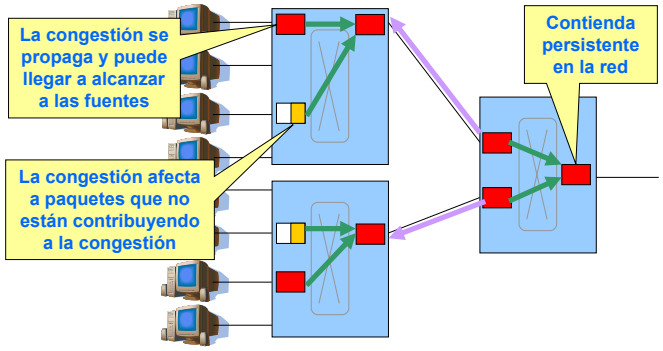
Cuando hay contienda, dos o más paquetes que han llegado por diferentes entradas de un conmutador solicitan un mismo recurso. Por tanto, cuando se produce congestión en un punto de la red, son varias las colas que pueden llenarse y propagar dicha situación. Como dichas colas de entrada se llenan con tráfico proveniente de diferentes conmutadores, la propagación de la con-



(a)



(b)



(c)

Figura 3.3: Evolución de los árboles de congestión: (a) Un canal de salida no dispone de ancho de banda suficiente para atender la demanda; (b) la situación se propaga a los conmutadores vecinos por la actuación del control de flujo; (c) la congestión se extiende y afecta a otros paquetes que no están contribuyendo a la congestión.

gestión no se producirá a lo largo de una única ruta lineal. Por el contrario, la congestión se ramificará y extenderá por las diferentes rutas que confluyen en la zona congestionada y que aportan tráfico al punto donde se originó la congestión. Esta evolución puede verse en la Figura 3.3. Por este motivo, las zonas congestionadas se denominan árboles de congestión [García05b]. Las diferentes rutas que convergen en el punto de máxima congestión son las ramas del árbol de congestión, y el punto donde convergen es la raíz del árbol de congestión. Asimismo, se denominan hojas del árbol de congestión a los extremos de las diferentes ramas que forman el árbol. Dichas hojas pueden ser conmutadores de la red, si la congestión no se ha extendido demasiado, o procesadores que están inyectando tráfico, si las ramas del árbol han crecido lo suficiente. Por supuesto, caben situaciones en las que algunas hojas son conmutadores y otras hojas son procesadores, en función de las rutas que permite la red y, sobre todo, del caudal de tráfico inyectado por cada procesador, que puede hacer que diferentes ramas crezcan a diferente velocidad. En la Figura 3.4 se representan gráficamente estos conceptos.

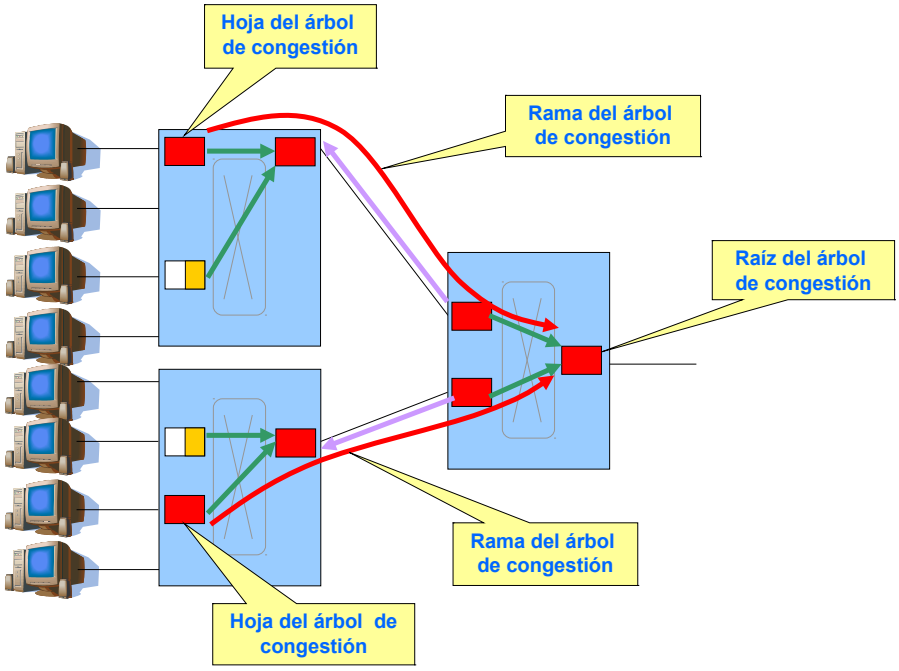


Figura 3.4: Estructura de un árbol de congestión

La evolución de los árboles de congestión es un proceso con una dinámica generalmente muy rápida [García05b]. Es muy fácil exceder temporalmente o de forma permanente la capacidad de la red. Basta con que muchos procesadores envíen a la vez grandes volúmenes de datos al mismo destino. De forma análoga, los árboles de congestión pueden desaparecer por sí mismos. Basta con que los procesadores que estaban inyectando tráfico hacia el mismo destino dejen de hacerlo.

Pero la dinámica de un árbol de congestión puede ser bastante más compleja que su crecimiento y posterior desaparición [García05b]. Puede haber varios árboles de congestión que coexistan en el tiempo, aunque no se hayan originado en el mismo instante de tiempo. Dado que los árboles crecen extendiendo sus ramas, es probable que dos o más árboles acaben solapándose en parte y compartiendo algunos fragmentos de rama. Es más, los árboles no siempre crecen desde la raíz hacia las hojas y se desvanecen en orden inverso. Un árbol de congestión ya formado, cuya raíz está dentro de la red, puede crecer hacia el destino de los paquetes, combinándose con tráfico intenso por otros enlaces o con otro árbol de congestión para formar un nuevo árbol de mayor tamaño con una nueva raíz. Es decir, se extienden dos o más árboles de congestión ya existentes, uniendo sus raíces a una nueva raíz recién formada y combinándose en un único árbol de mayor tamaño, o bien se extiende un solo árbol ya existente hacia abajo, desplazando su raíz, de la cual surgen ahora nuevas ramas. En casos extremos de crecimiento muy rápido y con muchas fuentes inyectando tráfico hacia el mismo destino, es posible que el árbol de congestión crezca desde las hojas hacia la raíz. En este caso, en cada conmutador que contendrá hojas del árbol final se crea un primer árbol con una raíz en dicho conmutador y con ramas muy cortas, de un solo enlace. Varios grupos de dichos árboles se combinan a continuación, formando varios árboles de mayor tamaño, cuyas raíces están más cerca del destino de los paquetes. Este proceso de fusión de árboles se repite varias veces hasta que queda un único árbol, cuya raíz suele estar ubicada en algún nodo destino, es decir, que la raíz del árbol final estará en un extremo de la red y no dentro de la misma.

A diferencia del crecimiento, la dinámica de la desaparición de un árbol de congestión es casi siempre la misma. Cuando la suma de los caudales de los tráficos que circulan por dos o más ramas del árbol pasa a ser inferior al ancho de banda disponible en el punto de confluencia de dichas ramas, las colas empiezan a vaciarse. Cuando las colas de una rama progresivamente se vacían, dicha rama deja de formar parte del árbol de congestión. Eventualmente, la desaparición de las ramas alcanza a la raíz del árbol, que también desaparece. Sin embargo, pueden darse situaciones complejas en las que dos o más árboles

solapan parcialmente algunas de sus ramas. Si el tráfico en una de esas ramas debido a uno de los árboles es mucho más intenso que el debido al otro árbol, pueden llegar a vaciarse las colas de uno de los árboles en dicha rama. Eso no quiere decir que esa rama del árbol con tráfico menos intenso haya desaparecido, pero todos los parámetros locales que se pueden medir apuntan a que sí lo ha hecho. Sin embargo, cuando el otro árbol deje de enviar tráfico tan intenso, reaparecerá la rama del árbol que parecía haber desaparecido.

3.4. Bloqueo en la cabeza de las colas

La aparición de un árbol de congestión en la red no se debe necesariamente a un mal diseño de la misma. En particular, cuando las aplicaciones se han diseñado de tal modo que un nodo de la red distribuye trabajo a otros nodos y posteriormente recoge los resultados, puede ocurrir que muchos de esos resultados se devuelvan casi al mismo tiempo. Como el destino de todos los paquetes que transportan dichos resultados es el mismo, el nodo destino no va a ser capaz de absorber tal caudal de tráfico, acumulándose los paquetes en la red y produciendo congestión. Pero a pesar de la congestión, la red está entregando los paquetes al nodo destino en el menor tiempo posible, ya que el enlace que conecta la red con el nodo destino está funcionando al cien por cien de su capacidad. Y gracias al control de flujo, no se descarta ningún paquete. Eventualmente, las notificaciones de control de flujo pueden llegar hasta los nodos que están inyectando paquetes en la red, ajustando de forma automática las tasas de inyección de cada uno de los nodos al tráfico que es capaz de aceptar la red.

Pero cabe plantearse si la situación recién descrita corresponde al comportamiento general, o al menos habitual, de la red. Más concretamente, es importante determinar si una red congestionada está consiguiendo las máximas prestaciones que permiten los enlaces que la componen y está transmitiendo los paquetes con la mínima latencia posible.

Los estudios empíricos sobre el comportamiento de las redes de interconexión arrojan unos resultados que difieren mucho del caso particular antes descrito [García06a]. Como puede verse en la Figura 3.5, cuando una red se congestiona, su productividad, o caudal agregado de tráfico que puede transmitir, suele degradarse notablemente. Esta reducción puede superar el 50 % de su productividad máxima. Pero las consecuencias de la congestión son aún más graves. En muchos casos, tras desaparecer las causas de la congestión y volver a las tasas de inyección de tráfico existentes antes de que la red saturase, la red

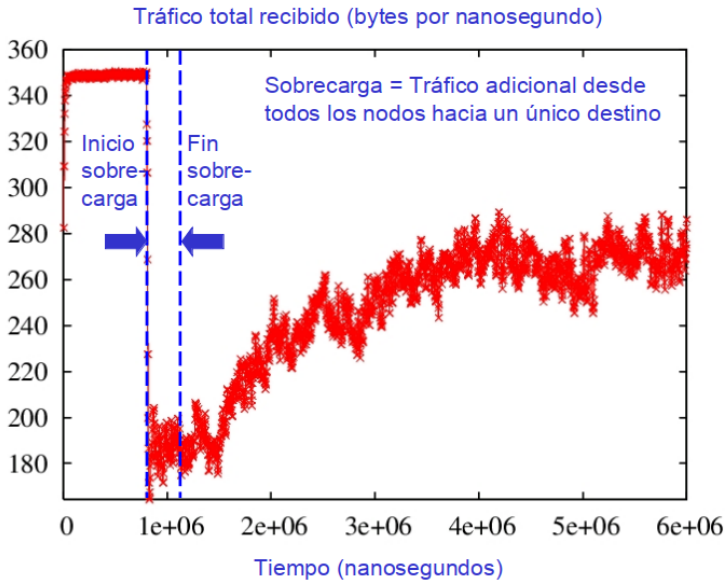


Figura 3.5: Degradación de la productividad de una red al sobrepasar su capacidad máxima

es incapaz de recuperar sus niveles de productividad anteriores a la aparición de la congestión.

Por otra parte, un conmutador con una sola cola por puerto de entrada y que recibe paquetes de forma ininterrumpida por sus enlaces de entrada es incapaz de reenviar todos esos paquetes por los enlaces de salida, a pesar de que el ancho de banda agregado de los enlaces de entrada y de salida es el mismo. Esta situación se da incluso cuando la distribución de destinos de los paquetes entrantes es uniforme, es decir, que solicitan cualquiera de los enlaces de salida del conmutador con la misma probabilidad, no consiguiéndose en este caso más que un 58 % de utilización del ancho de banda disponible [Karol87].

La causa de ambos comportamientos deficientes, según han demostrado diversas investigaciones, es esencialmente la misma. Si el paquete que está en la cabeza de una cola solicita un recurso y pierde el arbitraje, no sólo tendrá que esperar el paquete que ha perdido el arbitraje sino también todos los que le siguen en la misma cola. Esta situación se conoce como bloqueo en la cabeza de la cola (en inglés, head-of-line blocking) [Tamir92, García05b]. Como se muestra en la Figura 3.6, el motivo por el que dicho bloqueo en la cabeza de la cola afecta tanto a las prestaciones es porque, en muchas ocasiones, alguno de los paquetes que están esperando en la cola va a solicitar otro recurso dife-

rente, que no está congestionado. Si dicho paquete estuviese en la cabeza de la cola, ganaría el arbitraje, al no competir con ningún otro paquete, y podría transmitirse. El bloqueo en la cabeza de la cola hace que se pierdan oportunidades de transmitir paquetes, disminuyendo la productividad global de la red e incrementando la latencia de los paquetes que tienen que esperar a pesar de existir enlaces libres por los que podrían avanzar.

Tal como se ha definido, el bloqueo en la cabeza de la cola implica que tanto los paquetes bloqueados como el paquete que impide el avance de los demás están almacenados en la misma cola. Esta situación se denomina bloqueo de primer orden en cabeza de cola, para distinguirla del bloqueo de segundo orden (ver Figura 3.6). En el bloqueo de segundo orden en cabeza de cola [Duato05], los paquetes almacenados en una cola van a solicitar el mismo recurso en el conmutador actual, pero en cambio van a solicitar recursos diferentes en los sucesivos conmutadores de sus respectivas rutas. Y es posible que el paquete que está en la cabeza de la cola no pueda avanzar una vez llegue al siguiente conmutador, mientras que alguno de los paquetes que le siguen en la misma cola sí que podrán avanzar cuando lleguen al siguiente conmutador. En este caso no serviría de nada intercambiar las posiciones de los paquetes en una misma cola. De hecho, haría falta intercambiar las posiciones de paquetes que están en colas de diferentes conmutadores, lo cual es inviable.

El problema del bloqueo de primer orden en la cabeza de la cola es relativamente fácil de solucionar, ya que diferentes paquetes de la misma cola van a solicitar recursos diferentes en el conmutador en el que están almacenados. Basta para ello modificar la estructura del circuito que almacena los paquetes, pasando de usar una cola a otra estructura que permita varios puntos de lectura de paquetes. Sin embargo, la solución más conocida consiste en utilizar, en cada entrada de un conmutador, tantas colas como enlaces de salida tiene el conmutador, en lugar de una sola cola [Anderson93]. De este modo, los paquetes entrantes se almacenan en colas diferentes, en función de cual sea el enlace de salida que van a solicitar. Dos paquetes que vayan a solicitar recursos diferentes se almacenarán en colas diferentes, evitando así que uno de los paquetes tenga que hacer esperar al otro innecesariamente. Con esta solución se elimina completamente el bloqueo de primer orden en cabeza de cola, pero el coste es elevado.

En teoría es posible extender la solución antes mencionada para resolver también el bloqueo de segundo orden. Consiste en incorporar a cada entrada de cada conmutador tantas colas separadas como destinos ofrezca la red de interconexión. Esta solución es tan sencilla como inviable. No hay más que pensar en algunos diseños actuales, cuyo objetivo es dar soporte para la interconexión

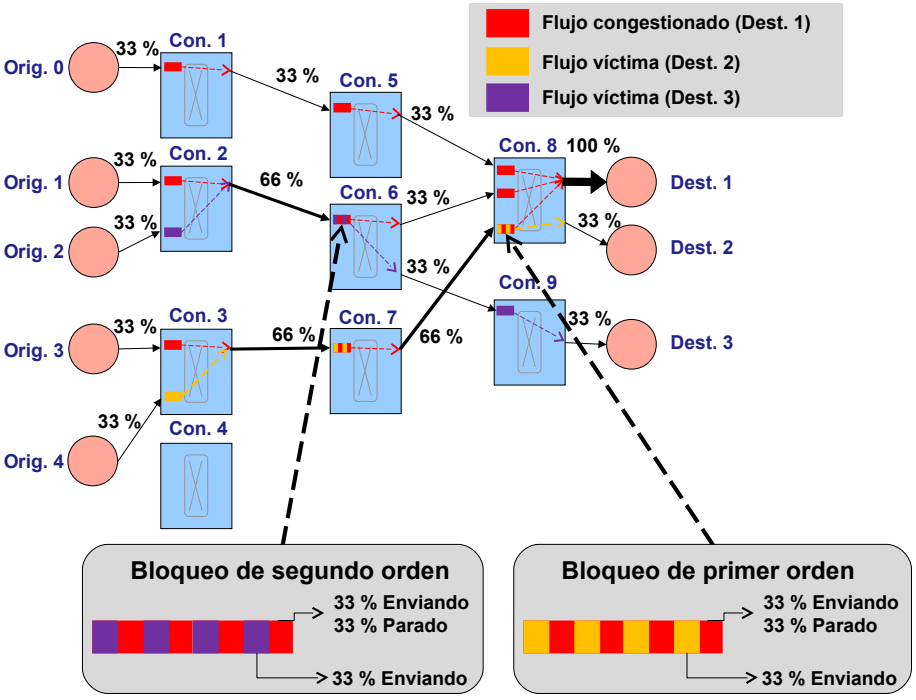


Figura 3.6: Bloqueo en cabeza de cola de primer y segundo orden. Los flujos que son víctimas de la congestión se ven frenados como consecuencia de dicho bloqueo, deteniéndose la transmisión durante parte del tiempo. Si no hubiese bloqueo en cabeza de cola, los enlaces representados con trazo más grueso transmitirían al 100 % de su capacidad, en lugar del 66 %.

de cientos de miles de dispositivos, sean procesadores o dispositivos de almacenamiento. El coste de incluir cientos de miles de colas por cada enlace de un conmutador es totalmente prohibitivo.

Así pues, la degradación de prestaciones asociada a la congestión se debe, fundamentalmente, a la aparición masiva de situaciones de bloqueo en cabeza de cola, la mayoría de ellas de segundo orden, que hacen que muchos enlaces no se utilicen aunque existan paquetes en las colas que potencialmente podrían usarlos. Para evitar esta degradación de prestaciones hace falta diseñar soluciones específicas para atacar el problema de la congestión.

3.5. Análisis global de las causas de la congestión

Antes de proceder a analizar las soluciones al problema de la congestión, es importante conocer las causas de la misma, ya que podría requerirse una solución especializada en función de la causa. Las redes de interconexión se diseñan, en general, de tal modo que los enlaces de comunicaciones tengan el mismo ancho de banda en ambos sentidos. Como cada nodo de procesamiento o de almacenamiento del sistema se conecta a la red mediante uno o unos pocos enlaces, el resultado es que cada nodo puede recibir el mismo caudal de tráfico que es capaz de inyectar. Si agregamos estos valores para todos los nodos de la red, tendremos que el caudal máximo de tráfico que los nodos de la red pueden recibir es igual al caudal máximo que pueden inyectar. Dicho en otros términos, el ancho de banda agregado de inyección es igual al ancho de banda agregado de recepción de la red.

Cabe ahora plantearse si cualquier sección de la red va a tener el mismo ancho de banda o si, por el contrario, puede haber secciones con mayor o menor ancho de banda que la inyección o la recepción. Como existen muchas secciones diferentes de una red dada, los análisis suelen limitarse a la bisección de la red [Dally90], la cual se obtiene al seccionar la red en dos mitades iguales por su sección más estrecha. Una de las decisiones de diseño más importantes en una red de interconexión es la selección de la topología de la misma. Existen topologías en las cuales el ancho de banda de la bisección es menor que el ancho de banda de inyección. Este es el caso de la malla y el toro representados en la Figura 3.1. En estas topologías resulta evidente que si todos los nodos inyectan tráfico a la vez con la máxima tasa de inyección posible, de tal modo que el destino de cada paquete esté ubicado al otro lado de la bisección de la red, ésta no va a poder atender tal demanda de tráfico y se va a producir congestión. Estas topologías tienen sentido cuando se conocen a priori los patrones de comunicaciones habituales de las aplicaciones y se diseña la topología para que coincida con dichos patrones. De este modo, puede llegar a conseguirse que la mayoría de paquetes solamente atraviesen un enlace y lleguen directamente a su destino. Tal es el caso de los supercomputadores con una topología de toro tridimensional, diseñados para ejecutar simulaciones de diversos sistemas físicos en un mundo tridimensional.

Cuando no se conocen con detalle los patrones de tráfico generados por las aplicaciones, dichos patrones cambian a lo largo del tiempo, o se ejecutan concurrentemente múltiples aplicaciones con diferentes patrones de tráfico, es habitual elegir topologías en las cuales el ancho de banda de la bisección no sea inferior al ancho de banda agregado de inyección, con objeto de evitar que la red de interconexión constituya un cuello de botella. Este es el motivo por el

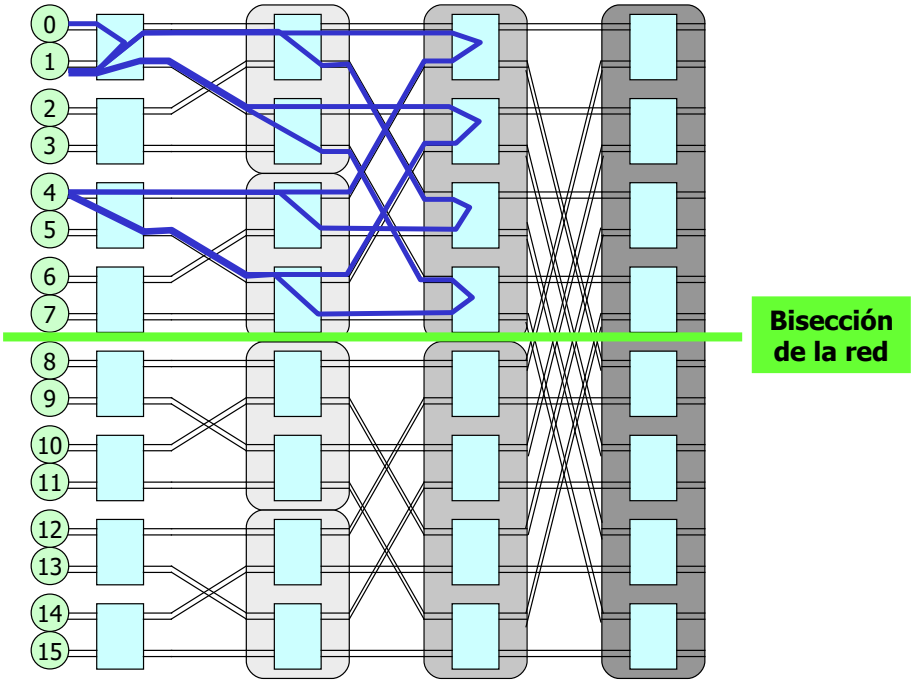


Figura 3.7: Red “fat tree” de 16 nodos, mostrando la ubicación de la bisección de la red.

que la topología “fat tree” se ha hecho tan popular tanto en supercomputadores como en centros de datos, ya que el ancho de banda de la bisección es igual al ancho de banda agregado de inyección, como puede apreciarse en la Figura 3.7, al tiempo que consigue muy baja latencia. De hecho, esta topología es una red multietapa que mantiene un ancho de banda constante en todas las etapas.

Sin embargo, no basta con tener en todas las secciones de la red un ancho de banda igual al ancho de banda agregado de inyección para evitar la congestión en la red. Ésta puede producirse por dos causas bien diferentes. En primer lugar, es posible que los destinos de los paquetes que se están inyectando en un instante dado sean tales que determinados destinos son solicitados por múltiples paquetes. Cuando dichos paquetes lleguen concurrentemente a su destino, éste no tendrá suficiente ancho de banda para absorberlos todos a la vez y se producirá congestión en los enlaces que conectan con los destinos más solicitados. En este caso, la congestión se origina en los nodos conectados a la red, si bien posteriormente puede propagarse por la misma. Es importante resaltar que no hay ninguna topología que pueda resolver este problema. La causa de la congestión es el patrón de tráfico de las aplicaciones que se están ejecutando.

En segundo lugar, incluso cuando el patrón de tráfico consiste en una permutación, con lo cual cada paquete inyectado tiene un destino diferente, puede producirse congestión en la red. Esta congestión se debe a que los algoritmos de enrutamiento pueden generar rutas tales que dos paquetes con orígenes y destinos diferentes compartan uno o más enlaces a lo largo de sus respectivas rutas. Este problema se ha estudiado exhaustivamente en el contexto de las redes multietapa con enlaces unidireccionales, habiéndose propuesto diferentes patrones de conexión entre etapas [Duato03]. Tras descubrirse que los diferentes patrones de interconexión eran isomórficos [Kruskal86] y que bastaba una renumeración de los nodos para convertir un patrón de conexión en otro, se abandonó esta línea de investigación. Pero el problema no se había resuelto, ya que para cualquier patrón de conexión existían permutaciones que daban lugar a congestión. Esta congestión se produce dentro de la red, no originándose a partir de los nodos destino.

Se han encontrado soluciones teóricas al problema de la congestión en la red para tráfico consistente en permutaciones. Una vía de solución consiste en sobredimensionar adecuadamente la red, de modo que las secciones intermedias de la misma tengan más ancho de banda que la inyección o la recepción, dando lugar a las denominadas redes no bloqueantes [Duato03]. Esta vía de solución requiere muchos conmutadores y enlaces adicionales y encarece mucho la red. Otra vía de solución consiste en añadir etapas a la red para poder definir rutas alternativas, de modo que siempre sea posible elegir el conjunto de rutas para el tráfico con una determinada permutación de modo que no se produzca congestión, dando lugar a las denominadas redes reconfigurables [Duato03]. Dichos estudios teóricos se han aprovechado a nivel comercial, ya que la topología “fat tree” es equivalente a plegar una red reconfigurable simétrica por su etapa intermedia, convirtiendo así los enlaces unidireccionales de la red reconfigurable en los enlaces bidireccionales del “fat tree” tras el plegado. Así pues, el “fat tree” hereda las rutas alternativas y las propiedades de reconfiguración de las redes reconfigurables. No obstante, ni el tráfico generado por las aplicaciones consiste en permutaciones ni es viable implantar un control centralizado para realizar la reconfiguración en las redes de interconexión actuales.

En resumen, la congestión puede aparecer dentro de la red o en los nodos destino. Dentro de la red se produce bien porque el ancho de banda de la bisección es insuficiente o porque los algoritmos de enrutamiento son incapaces de evitar las colisiones entre paquetes a lo largo de sus respectivas rutas. En los nodos destino se produce debido a que los patrones de tráfico de las aplicaciones son tales que determinados destinos reciben más tráfico del que pueden absorber mientras otros destinos no están recibiendo tráfico.

Capítulo 4

Soluciones tradicionales al problema de la congestión

El problema de la congestión se ha estado intentando resolver desde hace más de 30 años [Pfister85] y ha sido muy estudiado en redes sin control de flujo [Yang95], en las cuales se descartan paquetes cuando se llenan las colas. En dichas redes, cuando un paquete se descarta debe ser retransmitido, con los consiguientes incrementos en la latencia y en el consumo de recursos y de energía. Sin embargo, a pesar de esta pérdida de prestaciones, muchos aspectos de diseño se simplifican notablemente. En particular, al descartar paquetes, los árboles de congestión no crecen más allá de la raíz. Las soluciones adoptadas para tratar la congestión se limitan, fundamentalmente, a ajustar las tasas de inyección de paquetes para minimizar el número de paquetes descartados. Este es el caso del protocolo TCP, uno de los pilares de Internet [Brakmo95].

Lamentablemente, las soluciones adoptadas en redes con descarte de paquetes no funcionan adecuadamente en redes con control de flujo, por lo que ha habido que desarrollar nuevas soluciones. Estas soluciones son muy variadas [Yang95, Dandamudi99] y persiguen diferentes objetivos, relacionados en mayor o menor medida con la reducción o eliminación de la congestión o de sus efectos nocivos. Entre dichos objetivos cabe destacar los siguientes:

1. Garantizar la calidad de servicio. O dicho de otro modo, garantizar el comportamiento de ciertos flujos de información. Para ello se utilizan técnicas proactivas, que reservan recursos antes de iniciar la transmisión de información [Yew87]. La transmisión de determinados flujos de información puede requerir la entrega de los paquetes que los constituyen con una latencia acotada. En algunos casos se requiere incluso acotar las variaciones de la latencia respecto a su valor medio. Tal es el caso de la

transmisión de audio y de vídeo digitalizado. En estas transmisiones, la llegada tardía de un paquete provocaría la escucha de voces o sonido a destiempo o la presentación de imágenes fuera de secuencia. El establecimiento de garantías de comportamiento tiene como principal enemigo a la congestión en la red, ya que la latencia de los paquetes transmitidos puede verse notablemente incrementada debido al bloqueo en la cabeza de las colas. Cuando en una red se incluye este objetivo de diseño no se pretende eliminar la congestión que pueda producirse. Sin embargo, los mecanismos diseñados para establecer garantías de comportamiento deberán funcionar correctamente incluso en presencia de congestión. Para conseguirlo, se reservan recursos o ancho de banda para los flujos que requieren calidad de servicio, en detrimento del resto de flujos.

2. Equilibrar la carga en la red. La práctica totalidad de las topologías de red utilizadas en la actualidad proporcionan varias rutas alternativas entre todos los nodos fuente y destino, o al menos entre buena parte de ellos. La existencia de dichas rutas permite diseñar mecanismos que distribuyan el tráfico entre las diferentes rutas. Cuando se incluye este objetivo en el diseño de una red, se pretende equilibrar el tráfico por los diferentes enlaces de la red y permitir cargas de tráfico más elevadas sin que la red se congestione.

3. Eliminar las situaciones de congestión en la red. Las redes que incluyen este objetivo de diseño deben incorporar mecanismos de control de congestión para planificar o reducir la carga en la red de modo que la congestión que pueda producirse sea transitoria y desaparezca al cabo de un corto periodo de tiempo.

4. Eliminar el bloqueo en la cabeza de las colas. Como ya se ha indicado anteriormente, el efecto más pernicioso de la congestión es la introducción de bloqueo en las cabezas de las colas, siendo ésta la verdadera causa de la notable degradación de prestaciones que sufren las redes congestionadas. Por ello, en algunos diseños no se plantea como objetivo eliminar la congestión que pueda producirse, sino tan solo reducir o eliminar el bloqueo en la cabeza de las colas.

Seguidamente se describen algunas técnicas utilizadas tradicionalmente para conseguir los objetivos arriba indicados.

4.1. Calidad de servicio

Algunos de los servicios de comunicaciones que utilizamos a diario empezaron utilizando tecnología analógica y posteriormente se digitalizaron. Tal es el caso del teléfono y de la televisión. Cuando se empezaron a diseñar sistemas de transmisión digital de audio y vídeo se vio la necesidad de garantizar la entrega de la información con unos márgenes temporales muy estrictos. Más concretamente, los paquetes debían entregarse con una latencia acotada, requiriendo en algunos casos acotar incluso las variaciones de la latencia respecto a su valor medio. A las técnicas desarrolladas para ofrecer estas garantías se las ha denominado tradicionalmente calidad de servicio y es frecuente el uso de las siglas QoS (de su denominación en inglés, Quality of Service) [QoS].

Hoy en día, casi todos los estándares de comunicaciones incorporan algún tipo de soporte para garantizar la calidad de servicio [InfiniBand, Ethernet]. Aunque los detalles pueden variar mucho de un estándar a otro, básicamente se utilizan dos mecanismos para conseguir este tipo de garantías. En primer lugar, se incorpora un mecanismo de control de admisión, que lleva una contabilidad de los recursos ya asignados y sólo acepta el establecimiento de un nuevo flujo de información si existen recursos suficientes para poder atenderlo con las garantías solicitadas. Entre las verificaciones que realiza el control de admisión siempre se incluye la comprobación de que existe suficiente ancho de banda disponible en todos los enlaces de la ruta que van a seguir los paquetes. De este modo, se garantiza que el conjunto de flujos de información con requisitos de calidad de servicio no va a congestionar la red en ningún punto. El segundo mecanismo requerido es un planificador de transmisión en cada enlace de la red. Dicho planificador cumple una doble función. Por una parte, garantiza que los diferentes flujos con requisitos de calidad de servicio reciben los recursos previamente asignados y que ninguno de estos flujos consume más recursos de los asignados. Por otra parte, como es habitual que la red sea compartida con otro tráfico que no requiere garantías de calidad de servicio, el planificador asegura que los flujos de información con garantías de calidad de servicio reciben los recursos previamente acordados, incluso si la red llegara a congestionarse.

Los mecanismos incorporados en los diferentes estándares de comunicación para realizar la planificación suelen incluir algunos de los siguientes componentes: niveles de servicio especificados en la cabecera de los paquetes, colas de paquetes con niveles de prioridad, tablas de planificación que permiten descomponer un intervalo de tiempo en ranuras e indicar qué ranuras se asignan a cada nivel de servicio, planificadores que observan las priorida-

des de forma estricta, planificadores basados en tablas de planificación, etc. [InfiniBand, Ethernet]

Es importante resaltar que los mecanismos de calidad de servicio permiten ofrecer unas garantías mediante la reserva a priori de los recursos necesarios y mediante el establecimiento de una serie de prioridades en la utilización de los recursos. Cuando una nueva solicitud no puede atenderse con las garantías solicitadas, simplemente se rechaza. Por tanto, esta estrategia puede no resultar aceptable para determinadas aplicaciones. Es más, los mecanismos de calidad de servicio no son capaces de evitar la congestión de la red en el escenario más habitual, en el cual coexisten flujos de información con y sin garantías de calidad de servicio. Así pues, los mecanismos de calidad de servicio cumplen su función pero no son un mecanismo de control de congestión. No obstante, como veremos más adelante, algunos diseños recientes de mecanismos de control de congestión aprovechan los recursos para calidad de servicio existentes en la red, cuando éstos no se utilizan para realizar su función original.

4.2. Equilibrado de la carga en la red

Como se ha indicado anteriormente, puede aparecer congestión en la red incluso cuando el ancho de banda en la bisección de la red sea igual o mayor que el ancho de banda agregado de inyección. Ello se debe a que los algoritmos de enrutamiento determinan la ruta a seguir por los diferentes paquetes y, para determinados patrones de tráfico, dichas rutas pueden compartir varios enlaces mientras otros enlaces se quedan sin utilizar. El diseño de nuevos algoritmos de enrutamiento no resuelve el problema de forma general, pues si bien puede aliviar la situación para determinados patrones de tráfico, puede empeorarla para otros.

Tradicionalmente han sido dos las vías de solución adoptadas para resolver este problema. Ambas vías de solución son incompatibles entre sí. La primera vía de solución consiste en definir múltiples rutas alternativas desde cada fuente a cada destino y repartir el tráfico entre las mismas [ECMP, Franco99, Singh04]. Como ejemplo, en la Figura 3.7 pueden verse varias rutas alternativas entre los nodos 1 y 4. Dichas rutas alternativas deben ser todas ellas rutas válidas. En concreto, ninguna de las rutas debe utilizar recursos en un orden que pueda conducir a situaciones de interbloqueo. Según la topología de la red, es frecuente que el número de rutas alternativas varíe en función de la distancia entre los nodos fuente y destino. Cuanto más alejados estén, más rutas alternativas existen entre dichos nodos, como puede verse en la Figura 3.7 comparando las rutas entre los nodos 1 y 0 por un lado, y 1 y 4 por otro. Una

vez definidas las rutas alternativas entre un par de nodos dados, el tráfico entre los mismos se distribuye de forma equitativa entre las diferentes rutas. De este modo se consigue una utilización más equilibrada de los enlaces de la red, evitando situaciones en que unos enlaces están muy cargados mientras otros se quedan sin utilizar. Incluso si el tráfico no queda totalmente equilibrado y alguna de las rutas alternativas sufre más congestión que otras, dicha congestión será mucho menos severa que si todo el tráfico se transmitiese por la misma ruta.

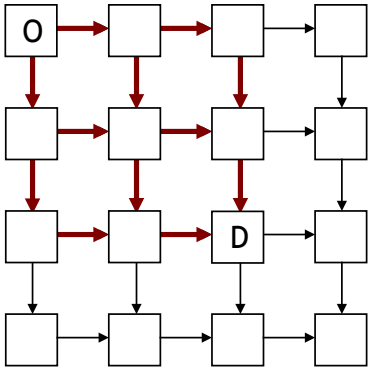


Figura 4.1: Enrutamiento adaptativo. Para ir del origen O al destino D se puede usar cualquiera de las rutas señaladas con trazo grueso. En cada conmutador de la ruta se decide dinámicamente el siguiente enlace a utilizar.

Existen topologías en las cuales el ancho de banda de la bisección es bastante menor que el ancho de banda agregado de inyección. En dichas topologías las técnicas de equilibrado de carga no suelen funcionar bien. También existen topologías en las cuales no se pueden definir apenas rutas alternativas sin que algunas de ellas puedan dar lugar a situaciones de interbloqueo. Es obvio que sin rutas alternativas o sin un número suficiente de las mismas, esta técnica de equilibrado de carga no va a funcionar bien. En estos casos puede recurrirse a otra vía diferente para equilibrar la carga, consistente en utilizar enrutamiento adaptativo [Duato93], como se observa en la Figura 4.1. Con esta técnica, la ruta seguida por cada paquete individual puede variar dinámicamente en cada conmutador de la misma, en función del tráfico existente en esa zona de la red. El enrutamiento adaptativo tiene dos partes: la función de enrutamiento y la función de selección [Duato93]. La primera calcula las opciones válidas de enrutamiento para un paquete con un destino determinado, entendiendo como tales aquellas que no pueden dar lugar a situaciones de interbloqueo, y la segunda elige entre las opciones válidas en función del tráfico

local. Puede tenerse en cuenta para ello la ocupación de las colas para cada una de las opciones. Si varias colas están vacías, también pueden utilizarse otros criterios para decidir entre ellas.

Las dos técnicas descritas para equilibrar la carga en la red consiguen, en general, una mejor distribución del tráfico y una mejor utilización de los recursos, con lo que aumenta el caudal máximo de tráfico que puede soportar la red antes de que se empiecen a producir situaciones de congestión. Sin embargo, si a pesar de distribuir mejor el tráfico la red empieza a congestionarse en alguna zona, las técnicas descritas no pueden evitar la aparición de árboles de congestión. De hecho, debido a que la carga está mejor distribuida, cuando una zona de la red se congestiona, las zonas colindantes se congestionarán con mayor rapidez que si no se emplean técnicas de equilibrado de la carga.

4.3. Control de flujo de extremo a extremo

Al igual que puede definirse un protocolo de control de flujo a nivel de enlace para evitar el descarte de paquetes cuando se llenan las colas en el conmutador receptor, también puede definirse un protocolo de control de flujo de extremo a extremo. A diferencia del control de flujo a nivel de enlace, el control de flujo de extremo a extremo se establece entre los nodos transmisor y receptor de un determinado flujo de información. El objetivo habitual de dicho control de flujo es evitar la pérdida de información en el nodo receptor si no es capaz de procesarla al ritmo que le está llegando. La forma más frecuente de implantar un control de flujo de extremo a extremo consiste en definir una ventana deslizante [Peterson00], que se irá desplazando sobre los paquetes de información a enviar. El tamaño máximo de la ventana lo fija el nodo receptor. En un momento dado, la ventana abarca los paquetes ya enviados cuya recepción no ha sido confirmada por el receptor. Cada vez que el transmisor recibe una confirmación o reconocimiento de la correcta recepción de uno o varios paquetes, elimina de la ventana los paquetes confirmados, desplazando así la ventana y pudiendo enviar nueva información. Este protocolo de comunicaciones permite que el nodo transmisor pueda enviar múltiples paquetes sin esperar a recibir las correspondientes confirmaciones de recepción. La versión más popular del protocolo de ventana deslizante está implantada en TCP, uno de los dos pilares principales de Internet, y puede verse en la Figura 4.2.

Además de su utilidad como control de flujo de extremo a extremo, un protocolo de ventana deslizante también permite reducir la probabilidad de que se congestione la red. Efectivamente, el tamaño máximo de la ventana deslizante define la cantidad de información que el nodo transmisor puede enviar sin ha-

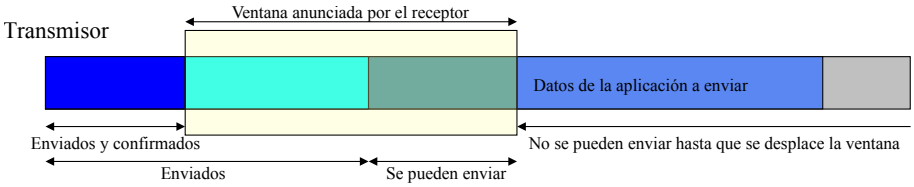


Figura 4.2: Protocolo de ventana deslizante.

ber recibido las correspondientes confirmaciones de recepción. El efecto global es una limitación en la cantidad de información que circula por la red, lo que se traduce en una menor probabilidad de que se produzca congestión, especialmente cuando se combina con técnicas de equilibrado de carga en la red. Pero el efecto beneficioso del control de flujo de extremo a extremo sobre la congestión es aún mayor. En primer lugar, si se produce congestión, los paquetes afectados tardarán más en llegar a su destino, por lo que los correspondientes reconocimientos también tardarán más en ser devueltos al nodo transmisor. El resultado es que éste agotará su ventana de transmisión y dejará de transmitir durante un cierto tiempo, con lo que de forma automática disminuirá su contribución a la congestión. En segundo lugar, cuando se dan situaciones de tráfico de varios a uno, el nodo receptor tendrá que atender tráfico de múltiples fuentes a la vez, cada una con su ventana deslizante, por lo que las confirmaciones de recepción las repartirá entre las diferentes fuentes. De este modo, cada nodo transmisor recibirá confirmaciones con menor frecuencia que en el caso de tráfico de una sola fuente a un destino, por lo que los nodos que transmiten reducirán automáticamente el caudal de tráfico que envían. Recordemos que una de las situaciones de congestión más problemáticas se produce precisamente cuando las aplicaciones transmiten tráfico de muchos nodos a uno o de muchos nodos a unos pocos.

4.4. Control de congestión de extremo a extremo

También se han desarrollado varias técnicas de control de congestión de extremo a extremo que tienen como objetivo directo detectar y eliminar la congestión. El mecanismo más frecuentemente utilizado en propuestas académicas y en diversos estándares de comunicaciones consiste en detectar las situaciones de congestión, notificar dichas situaciones a las fuentes que están contribuyendo a la congestión y limitar la inyección de paquetes en dichas fuentes [InfiniBand, ECN, Smai98, Thottethodi01, Vogels00], como se muestra en la Figura 4.3. Se trata, por tanto, de un sistema de control en bucle cerra-

do. Una de las características de dicho sistema de control es que incorpora un retardo puro en la cadena de realimentación, ya que las notificaciones tardan un cierto tiempo en llegar a las fuentes donde se va a limitar la inyección. Como es sabido por la teoría de control, los sistemas con un retardo puro en la cadena de realimentación son especialmente difíciles de estabilizar [Isermann81], ya que es frecuente que las actuaciones se sucedan sin esperar a ver el efecto de la actuación anterior, por lo que pueden producirse sobreactuaciones, con las correspondientes oscilaciones, que conducen a inestabilidad y pérdida de prestaciones.

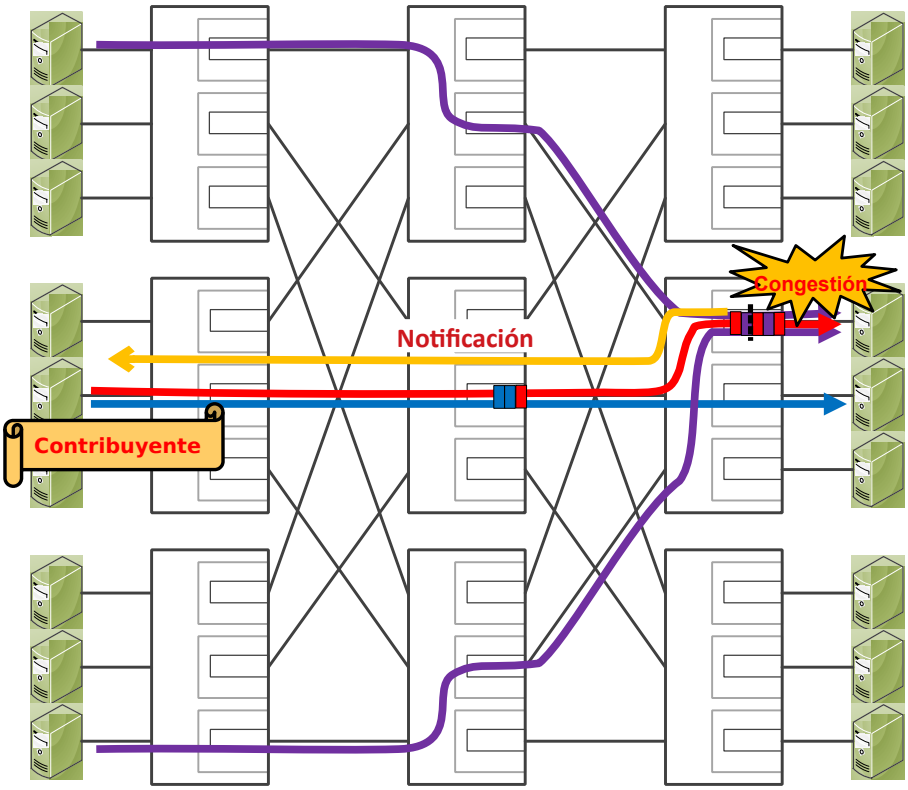


Figura 4.3: Control de congestión de extremo a extremo

Curiosamente, este aspecto no ha sido tenido en cuenta por los diseñadores de la mayoría de mecanismos de control de congestión propuestos hasta la fecha, seguramente porque no contaban con expertos en teoría de control en los equipos de diseño. El resultado es que las notificaciones de congestión, en lugar de ser enviadas directamente desde el conmutador donde se detecta la

congestión a la fuente que ha sido detectada como contribuyente a la misma, se anotan en el propio paquete de datos [InfiniBand, Gran10]. Por tanto, como puede observarse en la Figura 4.3, habrá que esperar a que dicho paquete llegue a su destino en una red congestionada, sea procesado y se devuelva el correspondiente reconocimiento, para que el nodo fuente reciba la notificación de congestión y proceda en consecuencia. Esta forma de proceder simplifica el protocolo de comunicaciones al evitar que los conmutadores se conviertan en fuentes de mensajes, pero da lugar a un retardo en las notificaciones varias veces mayor que en el caso de notificación directa.

Los mecanismos de control de congestión son, en general, bastante complejos. Tanto la detección como la notificación y la limitación de inyección presentan múltiples opciones. Para la detección, en principio basta con establecer un umbral en las colas de paquetes, detectando que hay congestión si la cola se llena por encima de dicho umbral. La dificultad aparece a la hora de identificar tanto las fuentes que están contribuyendo a la congestión como la severidad de la congestión. Si se marcan todos los paquetes que están en la cola como contribuyentes a la congestión, muy probablemente estaremos marcando como contribuyentes a ciertos paquetes que simplemente son víctimas y que están sufriendo las consecuencias del bloqueo en la cabeza de la cola. Resulta más razonable marcar sólo el paquete que está en la cabeza de la cola o el último que ha entrado en la cola. El razonamiento en el primer caso es que la cola se ha llenado y ha alcanzado el umbral de detección precisamente porque el paquete que está en la cabeza de la cola no puede avanzar, siendo el principal responsable de la congestión. El razonamiento en el segundo caso es que la ocupación de la cola puede haber crecido rápidamente debido a la llegada súbita de un grupo de paquetes, y éstos son los responsables de la congestión. Por otra parte, en cuanto a la severidad de la congestión, se puede notificar a las fuentes mediante el marcado permanente de paquetes mientras se esté excediendo el umbral de detección. La duda en este caso surge sobre la frecuencia con la que deben marcarse los paquetes. Si se marcan todos los paquetes que lleguen a la cabeza de la cola, se enviarán notificaciones más frecuentes a las fuentes. Pero esto puede producir que las fuentes sobreactúen. Asimismo, por el razonamiento anterior, también se marcarán todos los paquetes que, sin estar contribuyendo a la congestión, sean víctimas de la misma. Si, por el contrario, se marca cada cierto intervalo de tiempo el paquete que se encuentra en la cabeza de la cola, es más probable que la mayoría de los paquetes marcados estén contribuyendo a la congestión. Como ha habido bastante debate respecto a qué solución es la mejor, los estándares de comunicaciones que incorporan

mecanismos para detectar la congestión suelen dejar estos parámetros como configurables [InfiniBand].

Respecto a la notificación, ya se ha indicado que la solución habitual consiste en marcar paquetes que pasan por zonas congestionadas y dejar a cargo de los correspondientes reconocimientos la notificación de la congestión a las fuentes de esos paquetes. Esto simplifica el protocolo de comunicaciones respecto a las notificaciones directas del conmutador que ha detectado la congestión a las fuentes de la misma, pero incrementa notablemente el retraso en las notificaciones y hace mucho más difícil conseguir un funcionamiento estable del mecanismo de control de congestión.

Finalmente, los algoritmos para calcular cuánto se debe limitar la inyección son diversos y, en general, difíciles de ajustar correctamente [Gran10]. Para limitar la inyección, un enfoque sencillo consiste en esperar como mínimo un tiempo definido, previamente calculado en función de la severidad de la congestión, antes de inyectar el siguiente paquete [Jain89]. Dos dificultades se presentan en este aparentemente sencillo mecanismo. La primera es determinar de forma distribuida la severidad de la congestión en la raíz del árbol de congestión, y la segunda es que el resultado va a ser diferente si los paquetes que se inyectan pueden ser de diferente tamaño. De hecho, los diferentes estándares de comunicaciones permiten una amplia variedad de tamaños de paquetes [InfiniBand, Ethernet]. Por ello se suele recurrir a mecanismos más complejos para fijar la tasa de inyección de cada nodo. Como se ha indicado anteriormente, el protocolo de ventana deslizante tiene unas propiedades intrínsecas muy interesantes y útiles en este contexto. En particular, limita el tráfico inyectado por cada nodo y reacciona de forma automática ante la congestión limitando aún más el tráfico inyectado. Además, como la cantidad de información inyectada por un nodo y de la cual no se ha recibido un reconocimiento de correcta recepción está limitado por el tamaño de la ventana deslizante, una forma directa de reducir la tasa de inyección de un nodo consiste en reducir el tamaño de su ventana deslizante. Así pues, la variación dinámica del tamaño de la ventana deslizante se ha utilizado como mecanismo para limitar la tasa de inyección de los nodos que están contribuyendo a la congestión. Se han desarrollado múltiples algoritmos para establecer la secuencia de tamaños de la ventana deslizante en función de la frecuencia con que se reciben o se dejan de recibir notificaciones de congestión [Brakmo95]. Estas estrategias se han depurado mucho en el contexto de redes con descarte de paquetes como Internet, pero no han tenido demasiado éxito en redes con control de flujo y con una dinámica mucho más rápida, como son las redes de interconexión para supercomputadores y centros de datos. Se han dejado muchos parámetros

como ajustables. Como los administradores del sistema no suelen entender la compleja dinámica de la red durante la congestión, el resultado es que no saben ajustar los parámetros y los mecanismos de control de congestión de extremo a extremo suelen estar desactivados.

4.5. Reducción del bloqueo en la cabeza de la cola

Como se ha indicado anteriormente, la degradación de prestaciones que se produce cuando se congestiona la red proviene mayoritariamente de las situaciones de bloqueo en la cabeza de la cola. Es por ello que se han elaborado e implantado en diseños comerciales diversas soluciones para atacar directamente este problema. Las diferentes soluciones que se han propuesto pueden clasificarse en dos grandes grupos: las que extienden la funcionalidad de una cola mediante la introducción de múltiples puntos de lectura y las que utilizan múltiples colas para separar el tráfico en función de su destino.

Dado que el bloqueo se produce porque el paquete que hay almacenado en la cabeza de la cola no puede avanzar, las soluciones del primer grupo sustituyen las colas por otras estructuras de datos con múltiples puntos de lectura. De este modo, un paquete puede leerse y transmitirse aunque no esté almacenado en la cabeza de la cola, evitando así el bloqueo. La solución más conocida es la multicola dinámicamente asignada (DAMQ, siglas de su nombre en inglés, Dynamically Allocated Multi-Queues) [Tamir92], que consiste en sustituir cada cola individual por una estructura similar pero con múltiples puntos de lectura asignables dinámicamente a cualquier posición de la cola. El resultado es como si tuviéramos varias colas encadenadas de tamaño variable y con posibilidad de leer cada una de ellas por separado, pero con un único punto de entrada para almacenar paquetes.

El segundo grupo de soluciones utiliza múltiples colas para almacenar los paquetes en función de su destino, de modo que un paquete que no puede avanzar no bloquee el avance de otros paquetes que van a un destino diferente, ya que éstos estarán almacenados en otra cola diferente. La versión más popular de esta vía de solución son las denominadas colas de salida virtuales (VOQ, siglas de su nombre en inglés, Virtual Output Queues) [Anderson93], que consiste en disponer en cada puerto de entrada de un conmutador de tantas colas como puertos de salida, como se muestra en la Figura 4.4. De este modo, cada paquete entrante se almacena en una cola diferente en función del puerto de salida por el que tendrá que ser transmitido. Esta solución elimina el bloqueo de primer orden en la cabeza de la cola, pero no es capaz de eliminar el bloqueo de segundo orden. Para eliminar el bloqueo de segundo orden com-

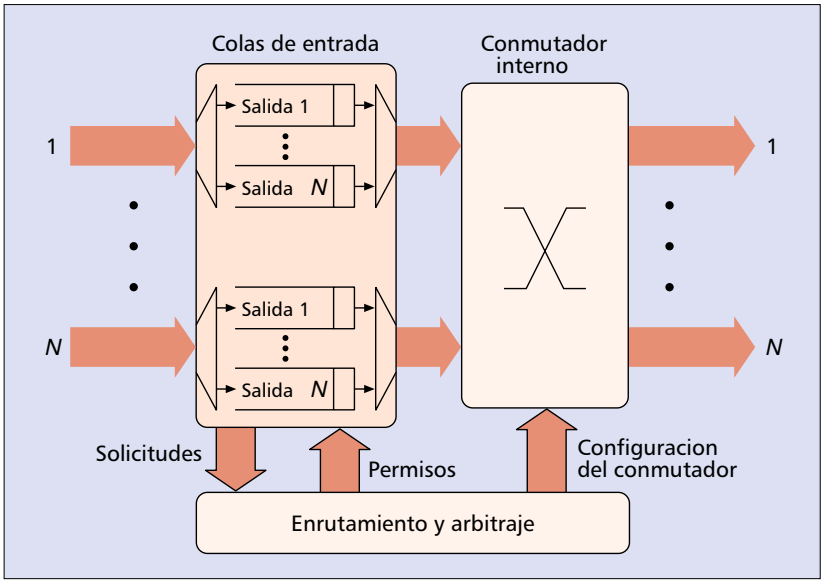


Figura 4.4: Conmutador con colas de salida virtuales

pletamente habría que disponer de tantas colas en cada puerto de entrada de un conmutador como nodos destino tiene la red completa, lo cual es inviable por su elevadísimo coste salvo para redes muy pequeñas [Dally98].

Los dispositivos que incorporan DAMQ o VOQ suelen implantarla mediante una memoria de alta velocidad en la cual se gestionan internamente varias colas virtuales mediante el uso de punteros. De este modo resulta muy fácil insertar y extraer paquetes de cada una de las colas, variando dinámicamente su tamaño en función de la demanda. En este caso, la única diferencia conceptual entre DAMQ y VOQ es que en DAMQ hay un único punto de entrada, por estar las colas encadenadas, y en VOQ hay tantos puntos de entrada como colas. Por tanto, DAMQ cayó en desuso. VOQ también está cayendo en desuso a medida que aumenta el número de puertos de los conmutadores, ya que su coste total crece con el cuadrado del número de puertos.

Capítulo 5

Soluciones eficientes al problema de la congestión

Tal como se ha indicado, las soluciones tradicionales para reducir el bloqueo en la cabeza de las colas no escalan bien, requiriendo un número de recursos que crece cuadráticamente con el número de puertos del conmutador o de la red, lo que las hace inviables en los sistemas actuales. Por otra parte, los mecanismos de control de congestión basados en detección de la congestión y limitación de la inyección son difíciles de ajustar. Con frecuencia dan lugar a oscilaciones en las tasas máximas de inyección permitidas, haciendo que los enlaces afectados pasen periódicamente de estar infrautilizados a estar congestionados. Por este motivo, en la mayoría de redes de interconexión se desactiva el mecanismo de control de congestión de extremo a extremo, consiguiendo en general mejores prestaciones que cuando se activa.

Sin embargo, el problema de la congestión es cada vez más acuciante, especialmente en los centros de datos utilizados como servidores de Internet, por lo que diversas empresas y grupos de investigación están atacando el problema con nuevos planteamientos.

5.1. Asignación eficiente de colas a paquetes

Hoy en día, casi todos los conmutadores comerciales incorporan múltiples colas en cada puerto. Su finalidad es bastante diversa, en función del estándar de comunicaciones en el que está basado. En la mayoría de los casos, estas colas fueron originalmente diseñadas para poder implantar mecanismos de calidad de servicio. Así, por ejemplo, los conmutadores para Ethernet incorporan varias colas con diferentes niveles de prioridad [Ethernet] y los conmu-

tadores de InfiniBand incorporan varias colas, denominadas enlaces virtuales (en inglés, virtual lanes), asignables dinámicamente a 16 niveles de servicio [InfiniBand]. Dado que en la mayoría de aplicaciones de estos conmutadores no se requiere soporte para calidad de servicio, dichas colas pueden ser reutilizadas para planificar el tráfico adecuadamente y reducir los efectos perniciosos de la congestión.

Algunas estrategias de planificación, tales como Homa [Montazeri18], utilizan de forma eficiente las colas de cada puerto de salida de los conmutadores de la última etapa, que están directamente conectados a algún nodo destino, para permitir que los paquetes de mensajes cortos adelanten a los paquetes de mensajes largos, consiguiendo así mejores prestaciones globales. Para ello se aprovecha la circunstancia de que dichas colas sólo pueden ser usadas por paquetes que van al nodo destino que está directamente conectado. Sin embargo, esta estrategia no es extensible a las colas de otros conmutadores, ya que dichas colas están compartidas entre múltiples flujos con diferentes destinos.

Existen diversas estrategias que consiguen aprovechar de forma eficiente las colas disponibles en los conmutadores de las distintas etapas de la red [Guay11, Escudero10a, Escudero11a, Escudero11b, Yébenes13]. En ellas se asignan grupos de destinos a cada cola, de forma equilibrada. El principio de funcionamiento se basa en realizar un particionado del tráfico, de modo que si se produce bloqueo en la cabeza de una cola determinada, este bloqueo sólo afecte a los flujos de información que comparten dicha cola, pero no al resto. La mayor dificultad de esta asignación de destinos a colas estriba en definir una función de asignación que consiga buenos resultados en todas las etapas de la red. Esta es precisamente la mayor diferencia entre las estrategias propuestas. Todas ellas consiguen la máxima efectividad en redes con una sola etapa, pero al aumentar el número de etapas unas estrategias mantienen mejor su efectividad que otras.

La estrategia más efectiva propuesta hasta la fecha, denominada Flow2SL [Escudero10a, Escudero11a, Escudero14], se basa en una sencilla teoría, consistente en definir tres parámetros que miden la calidad de la función de asignación. Estos tres parámetros miden el número de rutas asignadas a cada cola, los desequilibrios en la asignación de rutas a colas y el nivel de solape en la utilización de colas por parte de las diferentes rutas. Este último parámetro es el menos obvio y el que presenta una mayor variabilidad de unas propuestas a otras. Este parámetro se minimiza cuando cada grupo de rutas utiliza un conjunto de colas tal que los conjuntos de colas resultantes son disjuntos entre sí. Es decir, un grupo de rutas no comparte ninguna cola con otro grupo de rutas diferente en ninguna de las etapas de la red. Obviamente, los valores de estos

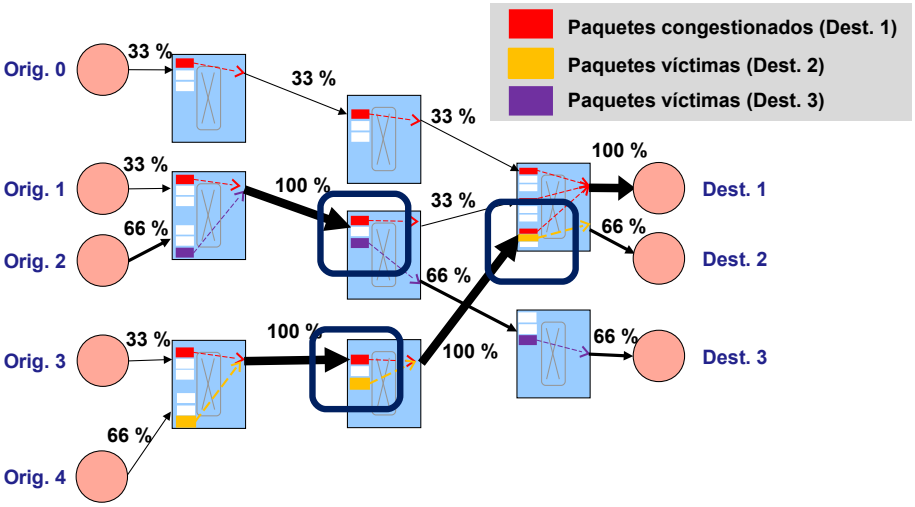


Figura 5.1: Ejemplo de tráfico para la asignación de colas Flow2SL. Los flujos de paquetes víctimas no sufren bloqueo en la cabeza de las colas porque se asignan a colas diferentes de las utilizadas por los paquetes congestionados en las diferentes etapas de la red. Como consecuencia de ello, en este ejemplo se consiguen utilizaciones del 100 % en los enlaces con tráfico no congestionado.

parámetros dependen tanto de la función de asignación como de la topología de la red y del algoritmo de enrutamiento. Por ello, Flow2SL se ha definido para la topología de red más popular en los centros de datos, el "fat tree", y para uno de los algoritmos de enrutamiento más efectivos propuestos para dicha topología, denominado DESTRO [Gómez15]. Para otras topologías y/o algoritmos de enrutamiento se pueden usar los mismos parámetros para medir la calidad de la función de asignación, pero la definición de dicha función deberá ser distinta, con objeto de optimizar dichos parámetros. La definición de funciones de asignación de destinos a colas para otras topologías y algoritmos de enrutamiento es un problema abierto.

5.2. Equilibrado dinámico de la carga en la red

El equilibrado de la carga en la red distribuye el tráfico de forma organizada de modo que se eviten acumulaciones excesivas de tráfico en ninguna zona, evitando o demorando la aparición de congestión dentro de la red. La técnica más frecuentemente utilizada consiste en aprovechar las múltiples rutas que ofrece el algoritmo de enrutamiento desde cada fuente a cada destino,

eligiendo una de ellas en el nodo fuente para cada transmisión. El espacio de diseño para esta técnica de equilibrado de carga ofrece opciones en múltiples dimensiones ortogonales. En una dimensión tenemos la opción de realizar un control centralizado de todas las transmisiones o bien realizar dicho control de forma distribuida en cada nodo fuente. En una segunda dimensión podemos definir la granularidad con la que se aplica el equilibrado de carga, realizando la asignación de rutas a nivel de paquete, a nivel de mensaje o a nivel de flujos de mensajes. En una tercera dimensión la asignación de rutas puede basarse en una política de distribución equitativa sin tener en cuenta el tráfico que circula en ese momento por la red o bien puede tener en cuenta el tráfico actual para evitar las zonas más congestionadas.

Los enfoques centralizados requieren coordinar con un nodo que actúa como planificador antes de realizar cada transmisión. En teoría disponen de información global que permite hacer una mejor planificación que cuando las decisiones se toman a nivel local. Sin embargo, como la distribución del tráfico en la red cambia muy rápidamente, es prácticamente imposible disponer de información actualizada del estado del tráfico en toda la red. Además, los enfoques centralizados tienen dificultad para escalar a grandes tamaños de red, así como para cumplir con los requisitos de latencia en tiempo real.

La selección de la granularidad del equilibrado de carga es un compromiso entre las ventajas que supone hacer una planificación para pequeños volúmenes de datos de cara a conseguir una mayor uniformidad en la distribución de la carga y la complejidad asociada con asegurar que los datos se entreguen en destino en su orden original. En general, sólo se consigue un buen equilibrado de carga si la distribución de rutas alternativas se hace a nivel de paquete, que es la unidad de datos más pequeña con capacidad para ser enrutada de forma independiente. El precio a pagar consiste en tener que reordenar en el nodo destino los paquetes de un mismo mensaje que puedan llegar fuera de orden, al ser transmitidos por rutas distintas. La forma tradicional de resolver la entrega fuera de orden ha consistido en incorporar una memoria en el nodo destino, en la cual se almacenan los paquetes que van llegando hasta que se completa un mensaje. Una vez recibido el mensaje completo, se reordenan los paquetes y se entregan en orden. Otra técnica más reciente y más eficiente consiste en utilizar el número de secuencia de los paquetes para almacenarlos directamente en la posición de memoria que le correspondería dentro del mensaje completo, con lo cual se evita tener que realizar la reordenación de paquetes.

El uso de una política de distribución equitativa funciona muy bien cuando todos los paquetes son del mismo tamaño y no hay interferencia entre el tráfico generado por los diferentes nodos. Sin embargo, en la práctica, cuando la pla-

nificación se realiza de forma distribuida en cada nodo fuente, prácticamente siempre hay interferencia entre los flujos de información generados por los diferentes nodos. Por ello, una política que tenga en cuenta la carga en cada una de las rutas alternativas que un nodo fuente puede elegir suele generar mejores resultados que una política que no tenga en cuenta dicha carga. Es importante resaltar que cada nodo fuente no necesita tener información de la carga en toda la red, sino sólo en las rutas alternativas que está utilizando en un momento dado.

Teniendo en cuenta las consideraciones anteriores, de este espacio de diseño, la configuración que permite obtener mejores resultados es la denominada Load-Aware Packet Spraying (LPS) [Congdon18]. Utiliza un planificador distribuido y escalable en los nodos fuente, funciona a nivel de paquetes y tiene en cuenta la congestión en las rutas alternativas para lograr un equilibrio de carga de grano fino sin causar que los paquetes se entreguen fuera de orden.

Con LPS, los paquetes transmitidos de un nodo a otro se distribuyen a través de las múltiples rutas de acuerdo con el grado de congestión medido en esas rutas. LPS incluye un número de secuencia en cada paquete para permitir que el nodo destino vuelva a ordenar los paquetes en su secuencia original. Como un nodo destino puede estar recibiendo flujos de muchos nodos fuente al mismo tiempo, es necesario disponer de zonas de almacenamiento separadas para cada nodo de la red que esté transmitiendo al nodo destino. Cada nodo fuente de LPS mantiene un indicador de congestión a lo largo de cada ruta a otros nodos. Este indicador puede determinarse mediante cualquier técnica de medición de la congestión. El nodo fuente utiliza el indicador de congestión para determinar cómo distribuir los paquetes a través de las múltiples rutas. Las rutas con carga más ligera transportarán más paquetes que las rutas congestionadas, que pueden omitirse por completo.

Las ventajas de LPS sobre métodos más tradicionales de equilibrado de carga, tales como Equal-cost multi-path routing (ECMP) [ECMP], son tres. LPS evita colisiones entre flujos que transmiten mensajes de gran tamaño porque distribuye el tráfico con granularidad fina a nivel de paquete. LPS puede adaptarse rápidamente a los cambios de estado de la red porque tiene en cuenta la congestión en las diferentes rutas. Finalmente, LPS es más paralelo que ECMP y puede reducir los tiempos de transmisión de un determinado flujo en redes con poca carga al distribuir un solo flujo a través de múltiples rutas paralelas al mismo tiempo.

5.3. Transmisión por iniciativa y bajo demanda

Como se ha visto al analizar el control de flujo de extremo a extremo, a medida que aumenta la congestión, este mecanismo reduce de forma automática el caudal global de tráfico inyectado en la red. También se ha visto que este mecanismo reacciona muy bien ante situaciones de congestión en escenarios de tráfico de muchos a uno o de muchos a unos pocos. El motivo de este buen comportamiento es que cada nodo destino tiene que notificar explícitamente las disponibilidades de espacio a cada nodo del cual está recibiendo información. Por tanto, los retrasos debidos a la congestión se traducen en retrasos en las notificaciones y conducen de forma natural a una reducción en las tasas de inyección. Asimismo, cuando un nodo destino tiene que enviar notificaciones a muchos nodos fuente, el número de notificaciones por unidad de tiempo devueltas a cada nodo fuente se reduce mucho, limitando de forma automática la tasa de inyección en dichos nodos fuente.

Esta característica implícita puede aprovecharse de forma explícita mediante la inclusión de un planificador en cada nodo destino. Con dicho planificador, cada nodo destino puede decidir a qué nodo fuente y con qué frecuencia devuelve notificaciones, que ahora se convierten en auténticos permisos para inyectar tráfico. De este modo, la transmisión de información pasa de realizarse por iniciativa de los nodos fuente a realizarse bajo demanda de los nodos destino. De hecho, con esta técnica un nodo destino puede determinar con precisión las tasas de inyección de los nodos que le están enviando tráfico. Para tomar una decisión, el planificador suele utilizar como información la disponibilidad de espacio de almacenamiento para cada nodo fuente, pero puede además utilizar otras informaciones, tales como el nivel de carga del enlace de entrada a dicho nodo destino. En particular, al menos en teoría, cada nodo destino puede ajustar las tasas de inyección de los nodos que le envían tráfico de modo que no haya congestión en los nodos destino, resolviendo así el caso más complejo de congestión.

A pesar de las enormes ventajas que tiene la transmisión bajo demanda, también tiene limitaciones e introduce algunos problemas, lo que hace que su aplicación no sea trivial ni sea capaz de resolver totalmente el problema de la congestión. El mayor problema que introduce es la necesidad de enviar una solicitud de conexión desde el nodo fuente al nodo destino, previo a cualquier nueva transmisión de información, con objeto de que el nodo destino asigne un espacio de almacenamiento adecuado para dicho nodo fuente e incorpore dicha solicitud a su planificador. Esta solicitud y las acciones asociadas introducen una sobrecarga y una demora que suelen resultar inaceptables tanto en supercomputadores como en centros de datos. En cuanto a las limitaciones, cabe

mencionar que no puede operar hasta que se haya procesado la solicitud correspondiente, que sólo puede ajustar la carga del enlace final de las rutas que conducen al correspondiente nodo destino y que no suele conseguir ajustar con elevada precisión el nivel de carga de dicho enlace. La primera limitación es obvia. La segunda se deriva de que sólo el enlace final de una ruta conduce exclusivamente a un determinado nodo destino, mientras que el resto de la ruta puede estar compartido con rutas que alcanzan otros destinos diferentes. La tercera limitación es más sutil, y proviene de que los permisos tardan un cierto tiempo en llegar a los nodos fuente y, sobre todo, que dichos nodos fuente pueden demorar el procesamiento de dichos permisos si concurrentemente están enviando tráfico a otros destinos de los cuales también están recibiendo permisos.

De cara a evitar las limitaciones y mitigar los problemas que introduce la planificación en los nodos destino y poder aprovechar sus enormes ventajas, se han propuesto varios protocolos de comunicaciones. El protocolo más sofisticado se denomina Homa [Montazeri18] y consiste en una combinación de múltiples estrategias, describiéndose a continuación las más relevantes. Para cada flujo de información, independientemente de su duración, Homa utiliza dos etapas en la transmisión de información: una primera etapa no planificada de corta duración y, si la duración del flujo lo permite, una segunda etapa planificada que abarca el resto de la transmisión. El objetivo principal de la primera etapa no planificada es evitar la demora inicial que supone el establecimiento de la planificación. Esto es especialmente importante para los flujos de muy corta duración, los cuales suelen ser muy frecuentes. Pero dicha etapa no planificada también es útil para los flujos de larga duración, que pueden arrancar de forma inmediata la etapa no planificada, al tiempo que se realiza la comunicación necesaria entre los nodos fuente y destino para configurar el planificador en el nodo destino. Una vez configurado el planificador, se pasa a la etapa planificada de la transmisión, evitando así que se pueda producir saturación en el caso de tráfico desde muchas fuentes a un destino.

Para reducir la probabilidad de que se produzca congestión debido a múltiples flujos simultáneos que estén en la etapa no planificada, Homa utiliza control de flujo de extremo a extremo con una ventana de transmisión de pequeño tamaño. Asimismo, Homa utiliza colas separadas para el tráfico planificado y el no planificado, minimizando así la interacción entre los mismos. Por otra parte, para evitar la tercera limitación antes mencionada, Homa utiliza una técnica de planificación en exceso o sobreplanificación. La idea es planificar más tráfico del que puede aceptar el enlace de entrada a cada nodo destino, en un porcentaje fijado de antemano, enviando los permisos correspondientes. De

esta forma se compensa el porcentaje de permisos concedidos y no utilizados debido a sobrecarga en los nodos fuente.

Otro protocolo, alternativo a Homa, para mitigar los problemas que introduce la planificación en destino, consiste en transmitir información sin usar ningún planificador en los nodos destino mientras no se produzca congestión en los mismos. Pero en cuanto se detecta congestión en algún nodo destino, se pone en marcha un planificador en el mismo para gestionar el tráfico destinado a dicho nodo. Esta estrategia híbrida se denomina Push and Pull Hybrid (PPH) [Congdon18], y funciona muy bien combinada con técnicas de equilibrado dinámico de la carga tales como LPS. Mientras se transmite información sin planificación hacia un determinado nodo destino, la técnica de equilibrado de carga LPS recopila información sobre la carga de las diferentes rutas alternativas y planifica el tráfico desde el nodo fuente. Cuando todas las rutas alternativas se congestionan, se transfiere la planificación al nodo destino, que concede permisos según la disponibilidad de ancho de banda en el enlace de entrada al mismo. A partir de ese momento, el nodo fuente limita el tráfico inyectado en función de los permisos de transmisión que recibe. No obstante, el nodo fuente puede seguir utilizando LPS para distribuir dinámicamente la carga entre las rutas alternativas, evitando así además la congestión dentro de la red. No hay que olvidar que la planificación desde los nodos destino es adecuada para evitar la congestión que se produce en los nodos destino, pero no es adecuada para evitar la congestión dentro de la red.

Tanto Homa como PPH consiguen resolver de forma efectiva el caso más difícil de congestión, que es el que se produce cuando múltiples nodos fuente transmiten simultáneamente grandes cantidades de información a un mismo destino. La activación de la transmisión bajo demanda gestionada mediante un planificador en los nodos destino afectados por la congestión permite limitar el tráfico inyectado en la red a los caudales que pueden ser absorbidos, sin introducir las oscilaciones e inestabilidades habituales en otros mecanismos tales como el control de congestión de extremo a extremo. La combinación de transmisión bajo demanda con equilibrado dinámico de la carga es especialmente interesante, ya que conjuntamente reducen e incluso eliminan los dos tipos de congestión que pueden producirse. Es más, gracias a que esta combinación de mecanismos reduce o elimina la congestión producida, también se reduce la frecuencia con que se activa el control de congestión de extremo a extremo, pudiendo este último llegar a resultar totalmente innecesario.

5.4. Asignación dinámica de flujos congestionados a colas

A pesar de todos los mecanismos mencionados en las secciones anteriores, el hecho de que los diferentes nodos de cálculo y de almacenamiento conectados a una red funcionen asíncronamente y utilicen funciones de enrutamiento de forma autónoma se traduce en que se pueden generar situaciones transitorias de congestión en la red. Cuando esto ocurre, tanto los mecanismos de equilibrado dinámico de la carga como los de planificación en los nodos destino son mecanismos que funcionan de extremo a extremo y tienen unos tiempos de respuesta excesivamente largos. Por tanto, no pueden evitar que se produzcan situaciones de bloqueo en las cabezas de las colas implicadas, con la consiguiente degradación de prestaciones. La asignación eficiente de colas a paquetes restringe notablemente el alcance del problema, pero no es capaz de eliminarlo por completo. Dependiendo del diseño del conmutador, puede que exista o no bloqueo de primer orden, pero sí que se acabará produciendo bloqueo de segundo orden. La única forma de evitar la degradación de prestaciones que se produce consiste en actuar localmente de forma inmediata.

Dicha actuación consiste en separar el flujo de paquetes que está produciendo la congestión del resto de flujos no congestionados. Los detalles de dicha separación varían de un diseño a otro, pero en todos los casos se requiere el uso de colas adicionales para poder realizar dicha separación. En unos diseños se utilizan colas adicionales reservadas explícitamente para este propósito mientras que en otros se reutilizan las colas originalmente diseñadas para soportar tráfico con diferentes niveles de prioridad y que no están siendo utilizadas para ese propósito.

Los primeros diseños basados en esta idea [Katevenis98, Krishnan04] resolvían el problema localmente pero no eran escalables. La primera solución escalable que se propuso para la separación de los flujos congestionados (denominada Regional Explicit Congestion Notification, RECN) [Duato05] utilizó colas dedicadas para cada uno de los flujos congestionados, con el doble objetivo de minimizar la interacción entre múltiples árboles de congestión parcialmente solapados y de permitir la desasignación y recuperación de una determinada cola cuando el árbol de congestión correspondiente se desvaneciera. La necesidad de diseñar soluciones con un coste más reducido ha llevado a la concepción de soluciones que comparten una o varias colas entre diversos flujos congestionados. En estos diseños ha habido que desarrollar soluciones ingeniosas para poder desasignar dinámicamente los flujos que ya no están congestionados, ya que pueden coexistir varios flujos congestionados y

no congestionados en la misma cola. En las siguientes secciones se describen con más detalle cada una de las alternativas para realizar la asignación dinámica de colas a flujos congestionados.

Pero antes se van a analizar varios problemas comunes a todas las soluciones, junto con los mecanismos necesarios para resolverlos. El primer problema que se presenta es detectar cuando se va a producir congestión. Cuando empieza a producirse congestión, los paquetes empiezan a acumularse en una o varias colas. Tal como se ha indicado en la sección 4.4, la forma más sencilla de detectar la aparición de congestión consiste en establecer un umbral en las colas, superado el cual el conmutador interpreta que se está produciendo congestión.

El segundo problema consiste en determinar cuál es el flujo de paquetes que está produciendo la congestión. Este problema ya se analizó en la sección 4.4. La solución más habitual consiste en marcar como responsable de la congestión el paquete que ha hecho que se supere el umbral en una determinada cola. Pero a diferencia de lo indicado en la sección 4.4, no se marcan todos los paquetes que hay en esa cola ni se siguen marcando paquetes con cierta periodicidad. En cambio, se registra la información sobre el flujo que ha producido la congestión en una tabla de flujos congestionados. La información registrada debe bastar para identificar de forma inequívoca todos los paquetes del mismo flujo que puedan llegar en el futuro. Entre otras opciones, esta información puede consistir en un identificador de flujo, la dirección del nodo destino, la dirección relativa de la raíz del árbol de congestión desde el conmutador actual, etc. Además de esta información de identificación, cada entrada en la tabla suele requerir información adicional para gestionar el flujo correspondiente, tal como el número de paquetes de dicho flujo almacenados en la cola correspondiente de flujos congestionados, el identificador de la cola en la que se detectó congestión o un indicador de si está pausado el envío de paquetes a la cola correspondiente de flujos congestionados desde el conmutador anterior.

El tercer problema consiste en clasificar todos los paquetes que llegan a un conmutador por cualquiera de sus puertos de entrada. Debe determinarse con la máxima celeridad si cada paquete entrante pertenece a un flujo congestionado o no. Para ello se consulta la tabla de flujos congestionados asociada al puerto por el que ha llegado el paquete. Si el flujo al que pertenece el paquete está registrado en la tabla, el paquete se almacena en la cola correspondiente de flujos congestionados. En caso contrario, se almacena en la cola que le correspondería normalmente.

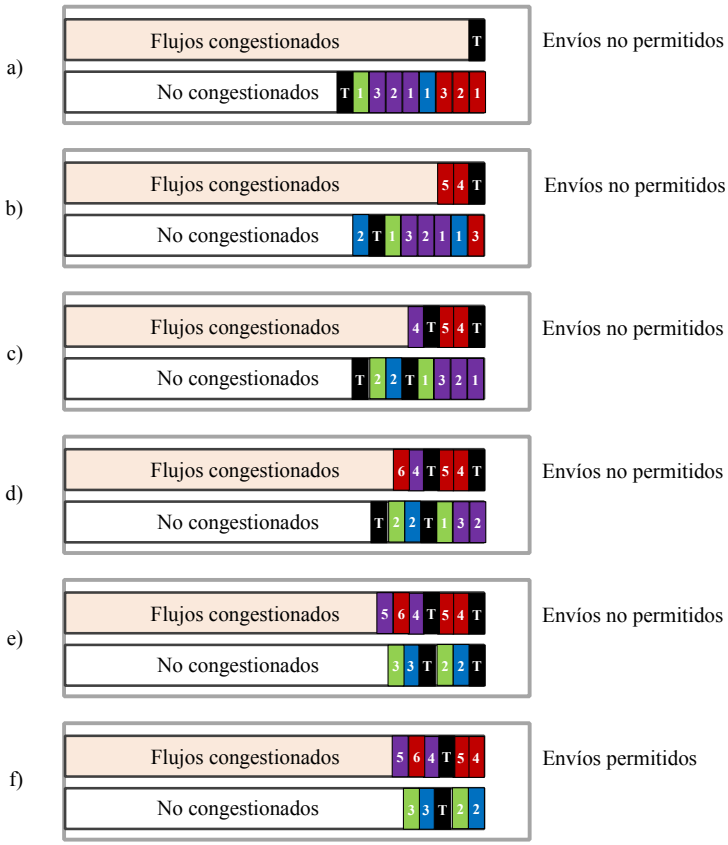


Figura 5.2: Uso de marcas para garantizar la entrega en orden de los paquetes. Se ha detectado congestión y se ha identificado el flujo de paquetes rojos como responsable de la misma, por lo que estos paquetes pasarán a almacenarse en la cola de flujos congestionados. Para evitar la entrega de paquetes fuera de orden: a) Se inserta una marca en cada cola y se desactiva el envío de paquetes desde la cola de flujos congestionados; b) los paquetes congestionados que llegan se almacenan en la cola de flujos congestionados, pero también pueden llegar paquetes de otros flujos no congestionados que se almacenan en la cola de flujos no congestionados; c) si se detecta algún flujo más como congestionado, se inserta otra marca en cada cola y se procede como en el caso anterior; d) mientras tanto, la cola de flujos no congestionados sigue enviando paquetes; e) una vez enviados todos los paquetes que había antes de la marca, dicha marca llega a la cabeza de la cola de flujos no congestionados; f) cuando ambas colas tienen una marca en la cabeza de la cola, se eliminan ambas marcas y se reactiva el envío de paquetes desde la cola de flujos congestionados.

El cuarto problema que se presenta es la entrega de paquetes fuera de orden. En el momento en que se detecta congestión, hay un cierto número de paquetes del flujo que está produciendo la congestión en una cola de flujos no congestionados. Como se acaba de indicar, el siguiente paquete que llegue de ese mismo flujo se almacenará en otra cola, dedicada a flujos congestionados. Si no se hace nada al respecto, es posible que el paquete recién llegado se transmita antes que algunos de los paquetes del mismo flujo que ya estaban en la cola de flujos no congestionados, con lo que se entregaría fuera de orden. En algunas tecnologías de red, esta entrega fuera de orden no está permitida, por lo que debe implantarse un mecanismo para resolver el problema. La solución trivial, consistente en copiar los paquetes del flujo que ha generado congestión de la cola de flujos no congestionados a la cola de flujos congestionados, no es viable por su elevado coste y la demora que supone. Por ello, se han propuesto varias soluciones para este problema.

Seguidamente se describe la solución más efectiva. Consiste en almacenar sendas marcas en ambas colas en el momento en que se asigna un flujo congestionado a una cola, como puede verse en la Figura 5.2. A partir de ese momento, la cola de flujos congestionados no podrá transmitir ningún paquete que se almacene después de su marca hasta que la cola de flujos no congestionados haya transmitido todos los paquetes que hay antes de su marca. Dicho de otro modo, las marcas en las respectivas colas establecen un punto de sincronización temporal. Todos los paquetes de un determinado flujo que hay antes de las respectivas marcas se transmitirán antes que cualquiera de los paquetes de ese flujo que estén almacenados después de las marcas. De ese modo, se garantiza la entrega en orden de los paquetes de cada flujo. Este sencillo mecanismo puede extenderse fácilmente para soportar la detección encadenada de varios flujos congestionados, sin más que insertar una nueva marca en cada una de las colas cada vez que se produce una nueva asignación de un flujo congestionado a una cola de flujos congestionados. Si una marca llega a la cabeza de una de las colas, ésta dejará de transmitir hasta que la marca en la otra cola también llegue a la cabeza de la misma. Los diseños que incorporan este mecanismo suelen incluir contadores del número de marcas almacenado en cada cola para simplificar la gestión de las mismas.

La separación de los flujos congestionados en una o varias colas separadas permite evitar que la cola (o las colas) para flujos no congestionados se llenen. Pero en cambio, si persiste la congestión, es muy probable que las colas de flujos congestionados se llenen, por lo que tendrán que activar el control de flujo a nivel de enlace. Si esto ocurre, se presenta un quinto problema, ya que el conmutador que va a recibir la notificación de control de flujo para pausar

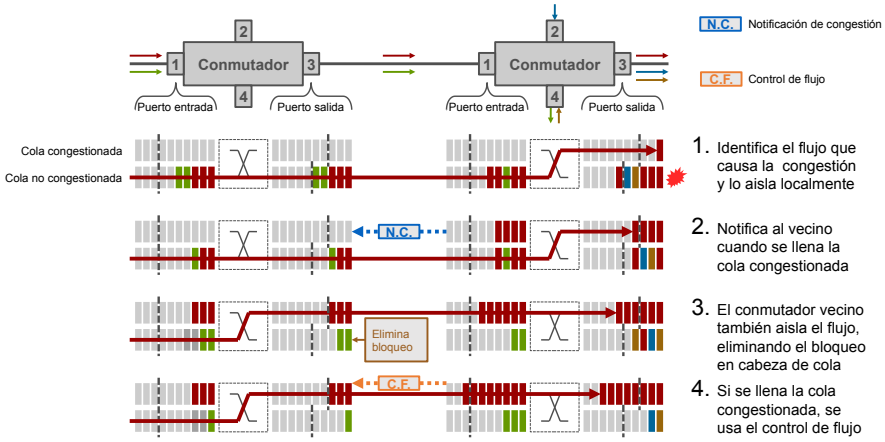


Figura 5.3: Secuencia de detección y notificación de la congestión al conmutador vecino

la transmisión va a tener que detener todo el tráfico, por lo que introducirá bloqueo en la cabeza de la cola, que es precisamente lo que se quiere evitar. Para resolverlo, hay que enviar al conmutador anterior una notificación de que se ha producido congestión, mucho antes de enviar la notificación para pausar el tráfico, como puede verse en la Figura 5.3. De este modo, el conmutador anterior tendrá tiempo de asignar dinámicamente el flujo congestionado a una cola de flujos congestionados, pausando sólo esta cola cuando llegue la notificación de control de flujo. Para poder enviar la notificación de congestión con suficiente antelación, es necesario definir un nuevo umbral en las colas, el cual debe alcanzarse bastante antes que el umbral para pausar el tráfico. Asimismo, cada notificación de congestión debe identificar el flujo responsable de la congestión de forma inequívoca, obteniendo dicha información de la entrada correspondiente de la tabla de flujos congestionados.

5.4.1. Solución con colas dedicadas

En este caso se asigna dinámicamente una nueva cola cada vez que se detecta un nuevo árbol de congestión o se recibe una notificación de congestión que corresponda a un árbol que no está ya registrado en la tabla de flujos congestionados. Dicha cola albergará los paquetes que lleguen en el futuro al conmutador correspondiente y que sean clasificados como contribuyentes al árbol de congestión asociado. Cada vez que se asigna una nueva cola, se regis-

tra una nueva entrada en la tabla de flujos congestionados, especificando toda la información relevante.

Dado que cada vez que se detecta un nuevo árbol de congestión se asigna una nueva cola y estos recursos son muy escasos, es necesario determinar, cada vez que se recibe una notificación de congestión, si corresponde a un nuevo árbol o a una expansión de un árbol ya existente. En este último caso, no se asigna una nueva cola, sino que se actualiza la información en la tabla concerniente al árbol de congestión asociado, indicando dónde se ha situado la nueva raíz del mismo. Este es el caso, por ejemplo, cuando dos árboles de congestión disjuntos se unen para formar otro árbol de mayor tamaño, con una nueva raíz ubicada en el conmutador donde convergen los dos árboles que se han unido. Nótese que cada conmutador de la red sólo puede estar en uno de los árboles que se unen, pero no en ambos.

Dada la escasa disponibilidad de colas para flujos congestionados, es importante desasignar dichas colas y la entrada asociada de la tabla de flujos congestionados en cuanto el árbol correspondiente se desvanece. La desasignación de colas se simplifica mucho por el hecho de que cada cola de flujos congestionados almacena paquetes de un único flujo congestionado. Cuando la congestión generada por dicho flujo se desvanece y las colas de flujos congestionados asociadas empiezan a vaciarse, se van desasignando de forma progresiva. Las condiciones para la desasignación son sencillas. Basta con que una cola sea una hoja del árbol de congestión, no haya pausado el tráfico proveniente del conmutador conectado al puerto de entrada correspondiente y esté vacía para poder proceder a desasignarla. La condición de que la cola sea una hoja del árbol de congestión se verifica mediante el envío de notificaciones de desasignación, similares a las notificaciones de congestión pero enviadas en sentido contrario. Cuando una cola recibe una notificación de que la cola anterior ha sido desasignada, tiene la certeza de que se ha convertido en una hoja del árbol de congestión. Por otra parte, la condición de que la cola esté vacía simplifica mucho la desasignación, ya que no va a introducir problemas de entrega fuera de orden. Sin embargo, en la práctica esta condición obliga a combinar este mecanismo con control de congestión de extremo a extremo para garantizar que los árboles de congestión se eliminan al cabo de poco tiempo desde su formación, pudiendo así desasignar las colas utilizadas, dejándolas disponibles para el siguiente árbol de congestión que pueda formarse.

La solución descrita constituye el mecanismo denominado Regional Explicit Congestion Notification (RECN) [Duato05], con la salvedad de que la garantía de entrega en orden de los paquetes se consigue mediante un mecanismo de postprocesado de las colas, que en lugar de transmitir un paquete

individual lo cambia a la cola congestionada correspondiente si la transmisión hubiera dado lugar a entrega fuera de orden. El mecanismo RECN lo patentamos conjuntamente con la empresa Xyratex, para ser utilizado en redes con tecnología Advanced Switching Interconnect (ASI), que fue una extensión del popular PCI Express usado en todos los computadores personales. También desarrollamos mecanismos similares a RECN, pero para redes con enrutamiento distribuido [Escudero08, Escudero13]. Asimismo, combinamos RECN y mecanismos similares con técnicas de control de congestión de extremo a extremo, con objeto de eliminar de forma paulatina la congestión cuando ésta no es de carácter transitorio, al tiempo que se liberan las colas dinámicamente asignadas por RECN [Escudero11c, Escudero15].

5.4.2. Solución con una cola compartida

La solución basada en colas dedicadas para cada árbol de congestión no resulta aceptable en tecnologías como Ethernet ya que, por su implantación generalizada y su elevada cota de mercado, los fabricantes deben ajustar mucho los precios de los productos. Además, Ethernet tiene que respetar una serie de estándares, manteniendo la compatibilidad con los dispositivos existentes cada vez que se estandariza un nuevo mecanismo. Esta necesidad de diseñar soluciones con un coste más reducido ha llevado a la concepción de soluciones de asignación dinámica que comparten una o varias colas entre los diversos flujos congestionados. De hecho, no se incluyen nuevas colas en el diseño de los conmutadores, sino que se reutilizan las colas existentes para separar los flujos congestionados. Dichas colas fueron originalmente concebidas para poder establecer diferentes niveles de prioridad entre los flujos de información. En general, basta con una sola cola para almacenar los paquetes de los flujos congestionados.

El funcionamiento es el siguiente. Cuando se detecta un nuevo árbol de congestión o se recibe una notificación de congestión que corresponda a un árbol que no está ya registrado en la tabla de flujos congestionados, simplemente se añade una nueva entrada en dicha tabla con la información correspondiente. A partir de ese momento, los paquetes entrantes que pertenezcan al nuevo flujo congestionado se almacenarán en la cola de flujos congestionados. Puede utilizarse el mecanismo de marcas antes descrito para garantizar la entrega de paquetes en orden.

A pesar de que la cola para almacenar paquetes de flujos congestionados se comparte entre todos los árboles de congestión que se hayan detectado, existe un límite en el número de árboles que una cola puede gestionar. Ello se debe a que cada nuevo árbol consume una entrada en la tabla de flujos congestio-

nados, la cual tiene una capacidad limitada. Como el coste de una entrada en dicha tabla es mucho menor que el coste de una cola, el límite en el número de árboles de congestión que se pueden gestionar es mucho más elevado cuando se comparte la cola de flujos congestionados que cuando se utilizan colas dedicadas. No obstante, dado que hay un límite, es necesario establecer mecanismos para liberar entradas de la tabla de flujos congestionados cuando los árboles de congestión correspondientes desaparezcan.

La desasignación de flujos congestionados es más compleja cuando la cola está compartida porque no cabe esperar que todos los árboles de congestión desaparezcan a la vez. Por tanto, a partir del momento en que se detecta congestión por primera vez, rara vez se dará la circunstancia de que la cola de flujos congestionados se vacíe. Y hay que tener en cuenta que no es viable realizar un barrido de la cola para determinar cuántos paquetes hay almacenados y a qué flujos pertenecen. Por tanto, hace falta establecer condiciones más complejas e implantar los mecanismos necesarios para gestionarlas. Una condición fácil de gestionar es liberar una entrada de la tabla de flujos congestionados cuando el número de paquetes de dicho flujo almacenados en la cola sea cero. Para ello basta añadir un campo en cada entrada de la tabla, que permita almacenar el número de paquetes de ese flujo que hay en la cola en cada momento y actúe como contador. Cuando llega un nuevo paquete, tal como se ha indicado anteriormente, hay que consultar la tabla para determinar si hay que almacenar dicho paquete en la cola de flujos congestionados o en otra cola. En ese momento, si el paquete corresponde a un flujo congestionado, se incrementa el contador correspondiente. Cuando se transmite un paquete de la cola de flujos congestionados, se consulta la tabla y se decrementa el contador correspondiente. Si dicho contador se hace cero, simplemente se elimina la entrada correspondiente de la tabla. Como consecuencia de ello, el siguiente paquete del mismo flujo que llegue al conmutador ya se almacenará en una cola para flujos no congestionados. No hay problema de entrega fuera de orden porque no quedan paquetes de dicho flujo en la cola de flujos congestionados. Esta condición de desasignación es muy adecuada tanto para situaciones de congestión súbita pero de corta duración como para resolver los casos de falsa detección, es decir, que se ha identificado como responsable de la congestión a un flujo que no lo es.

Pero esta sencilla condición no suele cumplirse en el caso de situaciones de congestión de larga duración, ni siquiera cuando ya están actuando otros mecanismos de control de congestión que funcionan de extremo a extremo. Estos mecanismos limitan la tasa de inyección de los nodos involucrados para eliminar la congestión, pero no pausan por completo el tráfico, por lo que se

mantiene el envío de paquetes con una cierta cadencia. Otra condición que sí que resulta efectiva en este tipo de situaciones consiste en establecer un umbral para la desasignación. Cuando la ocupación total de la cola de flujos congestionados más el resto de colas baja por debajo del umbral, simplemente se interpreta que se han eliminado todos los árboles de congestión y se borran todas las entradas de la tabla de flujos congestionados. Un mecanismo de marcas como el descrito anteriormente permite garantizar la entrega de paquetes en orden. La condición mencionada no es fácil de comprobar, por lo que se han propuesto diferentes variantes de esta condición. Una variante fácil de comprobar consiste en que todos los paquetes de la cola de flujos congestionados deben pertenecer al mismo flujo, la ocupación de dicha cola debe estar por debajo de un umbral, y la cola a la que se transferirá dicho flujo está vacía.

El mecanismo descrito se ha denominado Dynamic Virtual Lanes (DVL) [Congdon18] y está actualmente en desarrollo a través de una colaboración entre Huawei y nuestro equipo de investigación. Se ha lanzado y aprobado una petición para estandarizar este mecanismo como parte del estándar de Ethernet. Se están tramitando varias patentes conjuntas con la empresa Huawei cuyo contenido no se ha descrito por motivos de confidencialidad. Este mecanismo puede extenderse para utilizar más de una cola compartida para almacenar los paquetes de los flujos congestionados sin más que definir algún criterio sencillo para repartir los árboles de congestión entre dichas colas.

5.5. Funcionamiento combinado de las diferentes técnicas

El enfoque tradicional en las redes con control de flujo, es decir, sin descarte de paquetes, ha consistido bien en no hacer nada y dejar que el control de flujo propague la congestión y acabe ajustando automáticamente la tasa de inyección de los diferentes nodos de la red, o bien en implantar un control de congestión de extremo a extremo para intentar eliminar la congestión y el bloqueo en la cabeza de las colas. Para eliminar dicho bloqueo, al menos el de primer orden, también se han utilizado frecuentemente técnicas como Virtual Output Queuing. Conviene matizar que las redes en la mayoría de diseños han estado sobredimensionadas y que las situaciones de congestión no han sido frecuentes.

Hoy en día, las redes tienen cada vez tamaños más grandes, los conmutadores tienen más puertos y los enlaces son más rápidos. Por ello, las soluciones tales como Virtual Output Queuing han caído en desuso por su elevado coste. Asimismo, debido a las condiciones mencionadas, las técnicas de control de

congestión de extremo a extremo son cada vez más difíciles de ajustar y no parecen ganar adeptos. En cambio, la necesidad de disponer de mecanismos efectivos y eficientes de control de congestión es más imperiosa que nunca. Por este motivo, recientemente se han desarrollado o mejorado varios mecanismos que atacan el problema de la congestión, y otros están actualmente en desarrollo. Sin embargo, cabe preguntarse sobre el grado de compatibilidad de los diferentes mecanismos propuestos y sobre el funcionamiento cuando se combinan varios de ellos. Este es un problema abierto, pero ya se han probado ciertas combinaciones y se han debatido otras. Seguidamente se presenta un análisis sucinto del funcionamiento combinado de varias de las técnicas antes descritas.

En primer lugar, la utilización de técnicas de asignación eficiente de colas a paquetes o a flujos de información permite fragmentar el problema de la congestión y limitar notablemente el alcance de los efectos perniciosos de la congestión. No tiene efectos negativos ni requiere el ajuste de parámetros. No obstante, puede consumir todas las colas disponibles. Dado que otros mecanismos, tales como Homa, RECN o DVL también requieren el uso de parte o de todas las colas disponibles, la implantación de técnicas de asignación eficiente de colas a paquetes deberá supeditarse a la disponibilidad de colas no utilizadas para otros propósitos y restringirse a las colas que queden disponibles.

En segundo lugar, el uso de técnicas de equilibrado dinámico de la carga tales como LPS permite distribuir el tráfico de la forma más uniforme posible, evitando la congestión dentro de la red siempre que ésta disponga de ancho de banda suficiente en la bisección de la misma y, en cualquier caso, demorando en lo posible la aparición de congestión.

En tercer lugar, cuando los patrones de tráfico de las aplicaciones hacen que la aparición de congestión en los nodos destino sea inevitable, la transición a un modo de funcionamiento con transmisión bajo demanda tal como Homa o PPH puede evitar la propagación de la congestión. Tanto Homa como PPH realizan dicha transición de forma automática. Homa hace la transición de manera forzada para cada flujo, una vez ha transcurrido un cierto periodo inicial. PPH hace la transición en función del nivel de congestión en la red y resulta, por tanto, más flexible. Por otra parte, Homa consume todas las colas disponibles, lo que no deja opción a poder combinarlo con otros mecanismos. En cambio, PPH no necesita consumir colas y puede combinarse con LPS, consiguiendo conjuntamente eliminar la congestión dentro de la red y en los nodos destino, siempre que el ancho de banda de la bisección de la red sea suficiente.

A pesar de ello, ni LPS ni PPH pueden reaccionar suficientemente rápido como para evitar la congestión que pueda aparecer de forma súbita en una zona de la red, generalmente debida a la aleatoriedad del tráfico. En este caso, los mecanismos de asignación dinámica de colas tales como RECN o DVL son capaces de reaccionar de forma local y muy rápida. Si bien no eliminan la congestión, sí que son capaces de eliminar el bloqueo en la cabeza de las colas y la degradación súbita de las prestaciones. RECN consume muchas colas. DVL tiene un diseño más reciente y es más eficiente en el uso de recursos, por lo que es una mejor opción.

Una vez ha actuado el mecanismo de asignación dinámica de colas, LPS o PPH reaccionarán, aunque más lentamente, eliminando la congestión, lo que permitirá liberar los recursos consumidos por RECN o DVL. En el caso extremo en que LPS no pueda evitar la saturación dentro de la red por no disponer ésta de ancho de banda suficiente, queda como último recurso la utilización del tradicional control de congestión de extremo a extremo para eliminar la congestión, con todos los problemas de ajuste de parámetros que ello conlleva.

En resumen, en una red correctamente dimensionada, la combinación de equilibrado dinámico de carga y transición dinámica a transmisión bajo demanda con asignación dinámica de flujos congestionados a una cola compartida, es decir, LPS más PPH más DVL, permite dar una respuesta rápida, efectiva y eficiente al problema de la congestión. Si quedan colas no utilizadas en los conmutadores, estos mecanismos pueden complementarse con una asignación eficiente de las colas restantes a paquetes para reducir aún más el alcance de cualquier situación de congestión que pueda aparecer.

Capítulo 6

Aportaciones personales al control de congestión

El problema de la congestión es uno de los más difíciles de resolver en el campo de las redes de interconexión y dicha dificultad crece a medida que aumenta el tamaño de la red y la velocidad de transferencia por los enlaces. Además, ha despertado un gran interés por parte de la industria, ya que las prestaciones pueden reducirse notablemente si la red se congestiona. Por ello, nuestro equipo de investigación viene trabajando desde hace décadas en diferentes soluciones para atacar el problema de la congestión. También ha sido el problema al que más esfuerzo de investigación he dedicado a nivel personal. Entre mis aportaciones personales cabe citar:

1. Una teoría para garantizar la ausencia de interbloqueo cuando se utiliza enrutamiento adaptativo [Duato93].

2. Un diseño de conmutador con soporte para garantizar calidad de servicio [Caminero02b, Caminero03, Caminero06, Caminero02c, Caminero05].

3. Una función eficiente de asignación de destinos a colas [Nachiondo05a, Nachiondo05b, Nachiondo06, Nachiondo10].

4. Una teoría para medir la calidad de las funciones de asignación de destinos a colas [Escudero10a, Escudero11a].

5. Un mecanismo para ajustar automáticamente las tasas de inyección de los nodos que contribuyen a la congestión

6. La concepción del primer mecanismo de asignación dinámica de colas a flujos congestionados que funciona a nivel de la red entera (RECN) [Duato05, García06a].

7. Varios mecanismos necesarios para el correcto funcionamiento de la asignación dinámica de colas con una o varias colas compartidas (patentes solicitadas).

8. La primera propuesta de combinar asignación dinámica de colas a flujos congestionados con un mecanismo de control de congestión de extremo a extremo [Escudero11c, Escudero15].

9. El establecimiento de la relación entre control de congestión y control del consumo de energía.

Asimismo, he dirigido o codirigido varias tesis doctorales en temas relacionados con la congestión en redes de interconexión, cuyos resultados más relevantes se resumen a continuación:

1. El diseño y evaluación de un conmutador con soporte para garantizar calidad de servicio

2. La evaluación comparativa de varias funciones de asignación de destinos a colas

3. La propuesta de la función de asignación de destinos a colas denominada Flow2SL

4. Un análisis detallado del comportamiento dinámico de los árboles de congestión

5. La evaluación comparativa de múltiples opciones de diseño de la técnica denominada RECN

6. La evaluación conjunta y ajuste de parámetros cuando se combina la asignación dinámica de colas a flujos congestionados con control de congestión extremo a extremo

Capítulo 7

Problemas abiertos e ideas a explorar

A continuación se enumeran los principales problemas que siguen abiertos en el tema del control de congestión en redes de interconexión sin descarte de paquetes:

1. La minimización de los recursos necesarios para conseguir separar los flujos congestionados de los que no lo están.

2. El ajuste automático de los parámetros del control de congestión de extremo a extremo en los sistemas que incorporan este mecanismo.

3. La adaptación de las técnicas desarrolladas para redes conmutadas multietapa a las redes directas.

4. La combinación de técnicas de control de congestión con técnicas de gestión de la energía consumida por la red.

5. La combinación de técnicas de control de congestión con algoritmos de enrutamiento adaptativo.

6. La adaptación de las técnicas existentes o el desarrollo de nuevas técnicas para redes con conmutación "wormhole", las cuales disponen de muy poca capacidad de almacenamiento de paquetes en las colas de los conmutadores.

7. La adaptación de las técnicas desarrolladas o el desarrollo de nuevas técnicas de control de congestión para redes con tecnología totalmente óptica.

Respecto a la minimización de los recursos necesarios, recientemente he concebido un nuevo enfoque para eliminar el bloqueo en la cabeza de la cola que, en lugar de asignar dinámicamente los paquetes pertenecientes a flujos congestionados a colas separadas, directamente impide que dichos paquetes congestionados lleguen a la cola en cuestión. Concretamente, reduce la tasa de llegada de dichos paquetes a unos niveles que no lleguen a producir congestión, con lo cual se eliminaría el bloqueo en la cabeza de la cola sin necesidad de utilizar colas adicionales.

Dicho nuevo enfoque consiste en detectar cuándo se empieza a producir congestión con unos umbrales de detección muy estrictos, notificar dicha situación desde el conmutador donde se ha detectado la congestión a los conmutadores que le están enviando tráfico y modificar el comportamiento de los árbitros en dichos conmutadores para que acepten menos paquetes pertenecientes a flujos congestionados de los que aceptarían habitualmente. De este modo se reduciría el caudal de dichos paquetes a unos niveles que permitieran su procesamiento sin esperas en los conmutadores posteriores, eliminando de este modo no solo el bloqueo en la cabeza de la cola sino también la congestión.

Aunque este enfoque es muy prometedor, está todo por desarrollar. Surgen muchas incógnitas, tales como los umbrales de detección más adecuados, la información exacta que debe notificarse a los árbitros de conmutadores anteriores, las políticas de arbitraje más adecuadas para maximizar la utilización de recursos sin llegar a saturarlos, y la conveniencia o no de hacer llegar las notificaciones hasta las fuentes de los paquetes para combinar este mecanismo con uno de limitación de la inyección. Incluso más importante que estas incógnitas, queda por demostrar que este enfoque actuará suficientemente rápido como para eliminar el bloqueo en la cabeza de la cola antes de que sus perniciosos efectos se manifiesten de forma masiva. Es de esperar que su respuesta sea más lenta que la de los mecanismos de asignación dinámica de flujos congestionados a colas, ya que en lugar de actuar localmente de forma inmediata, requiere una notificación a otros conmutadores y su actuación sólo se notará en el tráfico futuro, pero no en el que ya está produciendo la congestión.

Respecto al ajuste automático de los parámetros del mecanismo de control de congestión de extremo a extremo en sistemas actuales, la dificultad del ajuste es tal que la mayoría de los usuarios consiguen peores prestaciones cuando activan dicho mecanismo que cuando no lo activan. Ello se debe a la inherente inestabilidad de dichos mecanismos, que se traduce en oscilaciones en el caudal de tráfico, con el consiguiente desaprovechamiento del ancho de banda de

los enlaces en los intervalos de tiempo en que la limitación de la inyección es excesiva.

Para resolver este problema, he propuesto un nuevo enfoque consistente en utilizar la tasa de llegada de reconocimientos de paquetes previamente transmitidos como referencia para la inyección de nuevos paquetes. El funcionamiento es análogo al de los mecanismos de transmisión bajo demanda salvo que los planificadores de tráfico no se ubican en los nodos destino sino en los conmutadores donde se ubica la raíz de un árbol de congestión. Más concretamente, este enfoque utilizaría un mecanismo de notificación idéntico al utilizado en el estándar InfiniBand [InfiniBand]. En dicho mecanismo, todo paquete que atraviesa una zona congestionada es marcado con un bit denominado Forward Explicit Congestion Notification (FECN). Cuando dicho paquete se recibe en su destino, dicha marca también se incluye en el reconocimiento devuelto al transmisor del paquete, en un bit denominado Backward Explicit Congestion Notification (BECN). En el caso más sencillo, en el que todos los paquetes son del mismo tamaño, la tasa de recepción de reconocimientos marcados con BECN es idéntica a la tasa de paso de los paquetes correspondientes por la zona congestionada. Por tanto, la tasa de recepción de reconocimientos indica con precisión qué fracción del ancho de banda de la zona congestionada es consumida por los paquetes cuyos reconocimientos se están recibiendo.

Esta estrategia es muy potente, pues tiene el potencial de adaptarse dinámicamente a los cambios en el caudal de tráfico inyectado por los diferentes nodos de la red. Efectivamente, si un nodo reduce su tasa de inyección, el resto de nodos que contribuyen a la congestión incrementarán la fracción de ancho de banda de la zona congestionada que consumen, con lo que también se incrementará la tasa de recepción de los reconocimientos correspondientes.

De nuevo, aunque este enfoque es muy prometedor, también está todo por hacer. El principal problema a resolver es determinar si la demora introducida desde el marcado de los paquetes al atravesar la zona congestionada hasta la llegada de los reconocimientos correspondientes a los respectivos nodos fuente va a introducir notables imprecisiones debido a la falta de actualidad de la información recibida. Estas imprecisiones pueden dar lugar incluso a problemas de estabilidad. De forma análoga, las variaciones dinámicas que habitualmente ocurren en las tasas de inyección de los diferentes nodos sufrirán una notable demora hasta que dichas variaciones sean notificadas al resto de nodos a través de los respectivos reconocimientos.

También deben desarrollarse unas políticas de actuación adecuadas. Si la tasa de inyección de nuevos paquetes se hace idéntica a la de recepción de reconocimientos, el árbol de congestión que ya se haya formado persistirá en el

tiempo, con los inconvenientes y la degradación de prestaciones que conlleva asociados. Por tanto, la tasa de inyección de nuevos paquetes deberá ser ligeramente inferior a la tasa de recepción de reconocimientos, con objeto de que el árbol de congestión vaya desapareciendo de forma progresiva. Pero tampoco la tasa de inyección debe ser muy inferior, ya que si bien el árbol de congestión desaparecería más rápidamente, a partir de ese momento podría producirse una degradación de prestaciones debido a una infrautilización de los enlaces de comunicaciones.

Respecto a la adaptación de los mecanismos desarrollados a las redes directas, cabe señalar que la inmensa mayoría de los sistemas de computación de altas prestaciones utilizan redes de interconexión con topologías indirectas. Más precisamente, utilizan redes conmutadas multietapa. Por tanto, hay menos interés en desarrollar soluciones para redes con topología directa que para redes indirectas. No obstante, hay algunas notables excepciones, tales como el IBM BlueGene en sus diferentes versiones, que han utilizado una topología directa, y más concretamente, un toro tridimensional.

La adaptación de las técnicas de control de congestión a las redes directas puede resultar muy compleja. Uno de los motivos es que mientras que el tráfico en las redes multietapa atraviesa las etapas de forma ordenada y está perfectamente definido qué etapas alimentan a otras etapas, en una red directa existen un gran número de ciclos. Esto tiene grandes implicaciones en varias de las técnicas de control de congestión propuestas hasta la fecha. Por ejemplo, si las notificaciones de congestión transmitidas hacia las etapas anteriores en técnicas como RECN circularan varias veces alrededor de alguno de los ciclos, agotarían rápidamente todos los recursos existentes para separar los flujos congestionados, ya que harían una nueva asignación cada vez que circularan por el ciclo.

Para atacar este problema, recientemente he concebido un nuevo enfoque que consiste en considerar no sólo la topología de la red, sino el conjunto de la topología y el algoritmo de enrutamiento. Un algoritmo de enrutamiento cuyo grafo de dependencia de canales sea acíclico, como es el caso de la inmensa mayoría de algoritmos de enrutamiento utilizados en redes directas, permitiría establecer un orden parcial entre todos los enlaces de la red. Según este nuevo enfoque, dicho orden podría utilizarse para determinar el sentido de avance de las notificaciones de congestión, de modo que dichas notificaciones nunca recorrerían ninguno de los ciclos existentes en la red.

En el caso de que el grafo de dependencia de canales no sea acíclico, como es el caso de los algoritmos de enrutamiento basados en el teorema de Duato, dicho teorema establece que debe existir un subgrafo, denominado grafo de

dependencia de canales extendido, que sí sea acíclico. En este caso, podría utilizarse el grafo de dependencia de canales extendido para determinar el orden de propagación de las notificaciones de congestión.

No obstante, quedan otros problemas importantes por resolver antes de conseguir una adaptación efectiva de las técnicas de control de congestión a las redes directas. En general, el diámetro de topologías tales como el toro tridimensional es mucho mayor que el de las redes multietapa para el mismo tamaño del sistema. Por tanto, todos los problemas relacionados con el retardo de propagación de las notificaciones de congestión se agravan de forma muy notable. En particular, los problemas de estabilidad de los sistemas que incorporan control de congestión de extremo a extremo, como sistemas de control en bucle cerrado que son, se agravan muchísimo al aumentar el retardo en la notificación de la congestión.

Respecto a la combinación con técnicas de gestión de la energía consumida, no se prevé que sea necesario desarrollar nuevos enfoques para el control de congestión. Sin embargo, es de esperar que la actuación de los mecanismos de gestión del consumo de energía intensifiquen la necesidad de utilizar control de congestión. Efectivamente, la reducción del consumo de energía se consigue mediante el apagado total o parcial de algunos enlaces de comunicaciones o mediante el uso de otras técnicas que reducen su ancho de banda. Esta reducción conlleva que la utilización de la red se acerque a su punto de saturación, tanto más cuanto más agresivas sean las estrategias de ahorro de energía. Como consecuencia, cuando las técnicas de gestión de energía se implanten de forma masiva en sistemas comerciales, aumentará de forma radical la probabilidad y frecuencia de aparición de árboles de congestión. Pero al mismo tiempo, la reactivación de recursos previamente apagados va a hacer posible una nueva forma de eliminar los árboles de congestión, aumentando el ancho de banda en las zonas de la red donde sea necesario. Por tanto, se hace necesaria una evaluación exhaustiva del comportamiento conjunto de los mecanismos de control de congestión y de gestión de energía. Esta evaluación determinará qué técnicas de control de congestión son las más adecuadas para responder rápida y frecuentemente a las sucesivas situaciones transitorias de congestión.

Respecto a la combinación con enrutamiento adaptativo, cabe destacar que dicho enrutamiento se ha propuesto en numerosas ocasiones como una técnica efectiva para reducir la congestión o retardar la aparición de árboles de congestión. Sin embargo, también existen trabajos en los que se pone de manifiesto que, para determinados patrones de tráfico, el uso de enrutamiento adaptativo reduce el caudal máximo de tráfico que puede transmitir la red. Estos resultados aparentemente contradictorios requieren un análisis más detallado.

En el marco de una colaboración con el laboratorio de IBM en Zurich, identifiqué las causas de dicho comportamiento aparentemente contradictorio. Si la raíz del árbol de congestión está ubicada dentro de la red, la utilización de rutas alternativas permite evitar las zonas más congestionadas, reduciendo o evitando así la formación de árboles de congestión. En cambio, si la raíz del árbol de congestión está ubicada en un nodo destino, no hay forma de evitar que dicho tráfico atraviese la raíz del árbol de congestión. En dicho caso, la utilización de enrutamiento adaptativo, no sólo no reducirá el problema, sino que dispersará el tráfico a través de una zona más amplia de la red. Ello hará que crezca el número de ramas del árbol de congestión, agravando así la situación.

También propuse un primer enfoque para poder determinar en una red con control de flujo basado en créditos, de forma sencilla, si la raíz de un determinado árbol de congestión está dentro de la red o no. Para ello propuse un método sencillo para determinar si un determinado enlace de la red es la raíz del árbol de congestión. Si un enlace de salida tiene varios paquetes almacenados para ser transmitidos pero no dispone de créditos, entonces pertenece a una de las ramas del árbol de congestión. Si dicho enlace tiene varios paquetes almacenados para ser transmitidos y sí que dispone de créditos, entonces se trata de la raíz del árbol de congestión.

Si bien este sencillo método permite distinguir entre la raíz y las ramas del árbol de congestión, un conmutador concreto no contiene información suficiente para determinar dónde está ubicada la raíz del árbol de congestión. Por tanto, sigue abierto el problema de cómo notificar de la forma más eficiente posible la ubicación de la raíz de los diferentes árboles de congestión a los conmutadores, con objeto de que puedan decidir si deben emplear enrutamiento adaptativo o no. Asimismo, se hace necesario desarrollar estrategias adecuadas para el caso en que existan múltiples árboles de congestión simultáneamente. Especialmente complejo es el caso en que un mismo conmutador esté afectado por varios árboles de congestión, algunos de los cuales tengan la raíz dentro de la red mientras otros tienen su raíz en algún nodo destino.

Respecto al desarrollo de técnicas para conmutación "wormhole", cabe destacar que esta técnica de conmutación no sólo es la opción preferida para las futuras redes dentro del chip, sino que está resurgiendo como la opción preferida en diseños recientes de redes de gran tamaño. Concretamente, la empresa Atos Bull, líder europeo en sistemas de supercomputación, ha preferido esta opción en sus diseños actuales y futuros. La gran ventaja de la conmutación "wormhole" es que requiere colas de pequeño tamaño, pudiendo funcionar correctamente incluso cuando en una cola no cabe ni siquiera un solo

paquete de datos. El truco para conseguir un correcto funcionamiento consiste en descomponer un paquete en unidades de control de flujo de pequeño tamaño, permitiendo así que un paquete pueda quedar distribuido entre varias colas cuando no puede avanzar por la red. El resultado global es que se requiere mucha menos área de silicio en los conmutadores para almacenar paquetes y, por tanto, pueden fabricarse conmutadores de mayor tamaño, haciendo así las redes más económicas y compactas.

Esta técnica de conmutación hace prácticamente imposible la aplicación eficiente de muchas de las técnicas de control de congestión conocidas hasta el momento. Efectivamente, la escasa capacidad de las colas hace que éstas se llenen incluso antes de que una notificación específica de control de congestión haya podido propagarse. Es más, la escasa capacidad de las colas hace que los árboles de congestión crezcan muy rápidamente. Dicho crecimiento es tan rápido que las ramas del árbol de congestión pueden alcanzar los nodos fuente del tráfico incluso antes de que lleguen las notificaciones de congestión. Por tanto, las técnicas de control de congestión de extremo a extremo siempre van a reaccionar demasiado tarde. Pero además, las variaciones en la tasa de inyección en los diferentes nodos de la red que están contribuyendo a la congestión van a tener un efecto mucho más rápido y acusado en la aparición y desaparición de los árboles de congestión, por lo que va a resultar mucho más difícil conseguir una respuesta estable de estos mecanismos.

Tampoco resultan adecuados los mecanismos basados en asignar dinámicamente los flujos congestionados a otras colas. No tiene sentido incluir en un diseño colas para flujos congestionados con capacidad para varios paquetes cuando las colas para flujos no congestionados no tienen capacidad ni para un solo paquete. Por otra parte, los beneficios de apartar los paquetes congestionados, almacenándolos en colas separadas, no son tan notables como en el caso de otras técnicas de conmutación. De hecho, en los casos en que en una cola no cabe ni un solo paquete, no se elimina el bloqueo en la cabeza de la cola al apartar un paquete, porque dicho bloqueo no existe al no haber más paquetes en la cola. Ni siquiera la detección de la congestión resulta sencilla, ya que con un fragmento de paquete puede llenarse una cola, y ello no implica necesariamente que haya congestión. Por tanto, no basta con definir un umbral de llenado de una cola para detectar así la congestión.

En cambio, sí que resultan eficaces los mecanismos de asignación eficiente de colas a destinos, pero dichos mecanismos no bastan para eliminar completamente los problemas originados por la congestión. También pueden funcionar correctamente los mecanismos de equilibrado dinámico de la carga y de transmisión bajo demanda. Sin embargo, ambos mecanismos tienen un funcio-

namiento de extremo a extremo y pueden tener una respuesta excesivamente lenta. Por tanto, resulta necesario desarrollar soluciones específicas de control de congestión para sistemas con conmutación "wormhole", siendo este un problema abierto de gran dificultad.

Respecto al desarrollo de técnicas de control de congestión para redes totalmente ópticas, cabe señalar que dichas redes se han investigado durante décadas por parte de varias empresas líderes y unos pocos grupos de investigación académicos. La relación coste/prestaciones de dichas redes ha hecho que se hayan descartado hasta el momento como solución comercial. Sin embargo, algunos análisis recientes apuntan a que dichas redes totalmente ópticas se convertirán en la mejor opción en una o dos décadas. También el ITRS (Integration Technology Roadmap) prevé que las redes totalmente ópticas se implantarán en el año 2025.

Las investigaciones realizadas hasta el momento apuntan a que las redes totalmente ópticas utilizarán técnicas de conmutación diferentes a las empleadas en redes con conmutadores electrónicos. Más concretamente, se ha propuesto utilizar conmutación de circuitos, estableciendo cada ruta completamente antes de iniciar la transmisión de datos, ya que no se pueden almacenar de forma flexible los paquetes en los conmutadores ópticos. Seguramente, el uso de conmutación de circuitos simplificará mucho el control de congestión, pues será posible evitar la congestión a base de reservar rutas sólo cuando hay suficiente ancho de banda disponible. En cualquier caso, se trata de un problema abierto en este campo.

Capítulo 8

Ciencia, transferencia de conocimiento y sociedad

Los científicos desarrollan su labor para el beneficio y progreso de la Humanidad en su conjunto. En la mayoría de los casos, los resultados de investigación se publican, poniéndolos a disposición de la sociedad sin restricciones. Esto permite tanto la utilización por parte de otros científicos, para seguir construyendo sobre la base del conocimiento ya existente, como su posterior utilización en beneficio de la sociedad, dando lugar a nuevos o mejores productos o servicios. Esta vocación de servicio está, en general, muy arraigada en la comunidad científica.

Los científicos tenemos en general la convicción de que toda la ciencia que se genera es útil, y los resultados que aún no han encontrado aplicación la encontrarán más pronto o más tarde. Ya dijo Louis Pasteur (Revue Scientifique, 1871), que «no existe una categoría de ciencia a la que se le pueda dar el nombre de ciencia aplicada. Existen las ciencias y las aplicaciones de la ciencia, unidas como el fruto al árbol que lo lleva». También existe una correlación directa demostrada entre la inversión en I+D y el volumen económico de un país [Klowden12, Domingo18]. Y en base a estos argumentos y evidencias, los científicos constantemente reclamamos mayores porcentajes de inversión en I+D. Lamentablemente, la inversión en I+D en España en el último año analizado (2016) es del 1,19 % del PIB [OECD], muy por debajo de la media de la Unión Europea (1,94 %) y de otros países más desarrollados como Estados Unidos (2,74 %) y Japón (3,14 %). En 2016 España ha invertido un 9,1 % menos en I+D que en 2009, mientras que la Unión Europea en su conjunto ha invertido un 27,4 % más. Además, en la distribución de las fuentes de financiación pesa bastante más la inversión pública que en otros países más

desarrollados. Si en el conjunto de la Unión Europea las inversiones de las empresas en I+D ascienden al 1,07 % del PIB, en España la cifra se queda en el 0,57 % [Cotec]. Por otra parte, da la impresión de que la mayor parte de la sociedad no conoce ni valora adecuadamente las aportaciones de los científicos en España, a excepción de las áreas relacionadas con la salud.

Para mejorar esta situación se requieren actuaciones coordinadas en diversos frentes, que involucran a la comunidad científica, a la administración y al tejido productivo. No obstante, en tanto se consigue esta coordinación, los científicos podemos aportar nuestro granito de arena para mitigar el problema y mejorar la situación, tanto a nivel global como a nivel nacional. Un primer aspecto en el que podemos intervenir es preocuparnos, no ya de si nuestros resultados de investigación se van a utilizar, sino incluso de cuándo y dónde se van a utilizar. El cuándo da lugar, entre otras actuaciones, a que nos planteemos realizar actividades de transferencia de conocimiento. Más concretamente, que dediquemos parte de nuestro esfuerzo a transferir nuestros propios resultados de investigación al tejido productivo y a la sociedad en general. Y el dónde da lugar a que nos planteemos a qué empresas transferimos el conocimiento, priorizando las empresas regionales, nacionales y europeas respecto a las de otros países cuando la financiación de la I+D haya provenido de fuentes regionales, nacionales y europeas, respectivamente. No se trata de que todos los científicos nos planteemos realizar actuaciones de transferencia del conocimiento como parte de nuestra actividad. Considero que tal exigencia sería excesiva, dada la enorme dificultad que ya supone realizar investigación de alto nivel, conseguir la financiación para desarrollarla y formar doctores. Pero sí que es importante que tomemos consciencia del problema y que, de forma voluntaria, emprendamos las acciones que estén a nuestro alcance.

Pero si de verdad consideramos que fomentar la transferencia de conocimiento es importante, no podemos dejar que esta actividad esté basada exclusivamente en el voluntarismo de los investigadores. Un segundo aspecto en el que podemos incidir consiste en fomentar la creación de un sistema de incentivos para que las nuevas generaciones de investigadores consideren la transferencia de conocimiento como una opción válida dentro de su carrera profesional. Dicho sistema de incentivos puede incluir desde reconocimientos y complementos salariales como el recién creado sexenio de transferencia del conocimiento e innovación hasta una valoración adecuada de los méritos de transferencia del conocimiento en los procesos de acreditación y promoción de los profesores e investigadores. Es importante destacar que, para que dichos reconocimientos y complementos supongan un incentivo, han de otorgarse de forma complementaria a los sexenios de investigación. Si no supone

un reconocimiento adicional, no va a suponer un incentivo, como ya lo ha demostrado la escasa aceptación que ha tenido el campo 0 sobre transferencia del conocimiento de la Comisión Nacional Evaluadora de la Actividad Investigadora (CNEAI) [CNEAI, MECD]. También es importante matizar que conviene distinguir entre actividades de transferencia de resultados de investigación y desarrollo tecnológico. No hay que olvidar que este último podría llegar a suponer una competencia desleal con las empresas del sector correspondiente, lo cual no resulta nada deseable. La forma en que el subcomité del campo 0 de la CNEAI, el cual presidí en su etapa inicial, acordó distinguir entre transferencia y desarrollo tecnológico fue mediante la trazabilidad. Los investigadores debían indicar qué resultados de investigación previos, desarrollados por ellos mismos, se habían utilizado en cada actividad de transferencia del conocimiento.

Con la transferencia de conocimiento devolvemos a la sociedad una parte de lo que invierte en nosotros, haciendo que nuestra investigación se convierta de forma más rápida en un beneficio para la sociedad o en un progreso más rápido de la misma. Esta agilización del retorno a la sociedad es importante para que la sociedad y los políticos que la representan adquieran una conciencia más clara del beneficio que supone invertir en investigación. Como consecuencia de ello, cabe esperar que se incrementen los presupuestos destinados a subvencionar la investigación.

Pero además, dentro de las oportunidades de transferencia de conocimiento que se presenten, los científicos podemos escoger en qué entorno realizamos la transferencia. Si analizamos la situación actual, vemos que la inversión en I+D realizada por países como España se traduce, en un elevado porcentaje, en publicaciones y doctores formados. Como muy pocos grupos de investigación realizan la transferencia de sus propios resultados, esta transferencia es realizada por los ingenieros y equipos de investigación de las empresas, casi siempre de grandes empresas. Pero en España tenemos pocas grandes empresas que desarrollen su propia tecnología. De hecho, las empresas que más se benefician de los resultados de investigación publicados están ubicadas casi siempre en los países más desarrollados. Por tanto, de los resultados de investigación publicados por nuestros científicos, la parte que es aprovechada y transferida al tejido productivo redunda en beneficio de empresas de otros países.

Pero esto es sólo una parte de la tragedia. En España las empresas rara vez ofertan puestos de trabajo que requieran el grado de doctor. Así pues, los doctores que formamos, si quieren quedarse en España, buscan empleo en otras universidades y centros de investigación o tienen que conformarse con puestos de trabajo en los que se requiere un nivel de cualificación mucho más bajo y

donde no pueden desarrollar todo su potencial. Una consecuencia directa de esta situación es que muchos de nuestros jóvenes más brillantes están emigrando a países más desarrollados.

La forma más directa de revertir esta situación consiste en que un porcentaje cada vez mayor de nuestros investigadores transfieran parte de los resultados que ellos o sus grupos de investigación obtienen. De este modo podrían elegir a qué empresas pueden beneficiar. Lo ideal es que dicha transferencia se produzca a través de institutos tecnológicos, que existen en muchas regiones de España, para que se beneficie todo un sector productivo y no una sola empresa. Dicha transferencia de conocimiento, cuando se produce de forma organizada y a gran escala, puede traducirse en un enorme beneficio para la región o país donde se desarrolla. Es un hecho contrastado que la incorporación de más tecnología en los productos produce un mayor valor añadido. Este incremento en el valor añadido se traduce en un incremento de la productividad por trabajador, lo que a su vez permite salarios más elevados. Asimismo, el incremento en el nivel tecnológico de una empresa lleva asociada la necesidad de contratación de un mayor número de titulados universitarios e incluso de doctores en el caso de empresas que incorporen sus propios centros de investigación. De este modo, no sólo incrementaríamos la riqueza de nuestro país sino que también reduciríamos la emigración de nuestros jóvenes más capacitados. Sobre la base de estos razonamientos y con el objetivo de apoyar, sistematizar e implantar de forma generalizada la transferencia de los resultados de nuestros investigadores, la Generalitat Valenciana ha creado recientemente la Agencia Valenciana de Innovación (AVI) [AVI], de cuyo Comité Estratégico de Innovación soy miembro.

Pero las oportunidades de transferencia de conocimiento hay que buscarlas. No podemos esperar que una empresa esté interesada precisamente en nuestros resultados de investigación. Eso rara vez ocurre, y menos cuando queremos focalizarnos en empresas ubicadas en una determinada zona geográfica, como es el caso de España. Hay que adaptarse a las necesidades de las empresas. Pero para que nos den a conocer dichas necesidades hace falta establecer previamente una relación de confianza con las empresas, tanto si se trata de empresas nacionales como extranjeras. En muchas ocasiones, la investigación que hemos desarrollado es un buen punto de partida, a partir de la cual podremos desarrollar soluciones especialmente adaptadas a las necesidades particulares de una empresa o sector productivo. Este esfuerzo adicional suele consumir bastante tiempo y recursos humanos, lo que casi siempre se traduce en una notable reducción de la productividad científica, si la medimos estrictamente en número de publicaciones. Esta es una de las razones por las

que muchos investigadores son reacios a iniciar el camino de la transferencia. Pero al mismo tiempo, en mi experiencia personal de colaboración con los laboratorios de investigación de diversas empresas, dicha colaboración puede ser una fuente muy interesante de nuevos problemas a resolver y sobre los que poder investigar. También, a través de dicha colaboración, he aprendido mucho sobre limitaciones de la tecnología, lo que me ha permitido abordar investigaciones más realistas, y sobre tendencias de mercado, lo que me ha permitido adelantarme a algunas de las necesidades futuras de las empresas y elegir mejor los temas de investigación en los que trabajar.

No quiero terminar esta sección sin manifestar que si ha habido algo a nivel profesional que me ha producido una mayor satisfacción que ser testigo del éxito profesional de mis discípulos o haber generado resultados de investigación de gran impacto, ha sido precisamente el haber sido capaz de desarrollar nuevas soluciones para resolver complejos problemas de diseño bajo las estrictas restricciones impuestas por la compatibilidad con estándares previos, el coste de los dispositivos y las limitaciones de la tecnología. Es una combinación de ingeniería y de investigación al más alto nivel, aliñada con la oportunidad de ver el resultado del esfuerzo implantado en un dispositivo que funciona.

Capítulo 9

Comentarios finales

La vida del científico es apasionante. Es fuertemente vocacional. Hay algo dentro de nosotros que nos empuja a dedicar nuestra vida a obtener resultados científicos en beneficio de la sociedad. Cada nuevo descubrimiento o logro científico suele producir una gran satisfacción. Sin embargo, no todos los científicos percibimos esta vocación ni la satisfacción asociada a los logros científicos del mismo modo. Las diferentes percepciones provienen no solo de la diversidad de nuestras preferencias sino también de la madurez de un determinado campo de investigación y de las herramientas disponibles. Así, por ejemplo, cuando un determinado proceso o mecanismo es desconocido y no se dispone de los conocimientos y los medios para descubrirlo, se pueden aplicar métodos estadísticos a muestras poblacionales significativas para obtener conclusiones, sin necesidad de conocer el funcionamiento interno. Tal es el caso de los ensayos clínicos consistentes en la aplicación de fármacos o tratamientos para curar determinadas enfermedades.

Este enfoque no satisface a todos los científicos y hace que muchos de ellos investiguen para descubrir los pormenores del funcionamiento de dichos procesos. Ante un nuevo fenómeno físico, ¿cuándo se siente mayor satisfacción, cuando se descubre dicho fenómeno, cuando se describe con precisión mediante un modelo matemático o cuando se entiende porqué se produce? Creo que no hay una respuesta universal y va a depender de cómo perciba cada investigador su vocación investigadora.

Algunos científicos hemos tenido la suerte de trabajar en un campo científico en el que o bien el nivel de madurez es elevado o bien el nivel de complejidad es suficientemente bajo como para ir más allá del descubrimiento científico y ser capaces de diseñar nuevas soluciones y sistemas que mejoran respecto a

lo conocido hasta el momento. Esta es la forma en que se manifiesta mi vocación científica.

Sé que esta capacidad no es exclusiva de mi campo de investigación. A través de debates y presentaciones de colegas de otros campos, he podido observar con gran satisfacción cómo los investigadores de muchos otros campos también diseñan o inventan nuevas soluciones. Tal es el caso del diseño de moléculas para conseguir una determinada funcionalidad en un proceso químico. Este es también el caso cuando se desarrolla una nueva teoría matemática. En otros campos, lo que se diseña son experimentos para descubrir fenómenos o mecanismos físicos, químicos o biológicos.

Quizá en un futuro no muy lejano, la capacidad de diseño de nuevas soluciones domine la producción científica en muchos campos del saber. Hoy en día ya es posible editar el ADN de diversas especies. Eso abre la puerta a que en el futuro podamos diseñar nuevas funcionalidades e incluirlas en el ADN. Entendemos ya el funcionamiento de las redes neuronales y somos capaces de diseñar complejos sistemas que reproducen partes importantes de la funcionalidad del cerebro, tales como la visión y el reconocimiento del lenguaje hablado y escrito. En algunas tareas de alcance restringido, tales como jugar a un juego de mesa, las redes neuronales han superado no solo a los mejores jugadores humanos sino también a los mejores programas de ordenador existentes hasta el momento. Tal es el caso de Alpha Zero, la inteligencia artificial que ha arrasado jugando al ajedrez. Esa capacidad de la inteligencia artificial está abriendo la puerta a un sinfín de aplicaciones beneficiosas para la Humanidad. Ojalá seamos capaces de utilizar esta capacidad de diseño para salvar a la raza humana y al planeta de los peligros cada vez más acuciantes a los que nos enfrentamos.

Quisiera terminar como he empezado, reiterando mi agradecimiento por haber sido aceptado por la Real Academia de Ciencias. Estoy convencido de que esta percepción que he manifestado, siempre desde el más profundo respeto hacia las opiniones y percepciones de los demás, resultará enriquecedora para los futuros debates en el seno de esta Academia.

Bibliografía

[Alfaro02] Francisco José Alfaro, José L. Sánchez, José Duato, Chita R. Das: A Strategy to Compute the InfiniBand Arbitration Tables. IPDPS 2002.

[Alfaro03] Francisco José Alfaro, José L. Sánchez, José Duato: A New Proposal to Fill in the InfiniBand Arbitration Tables. ICPP 2003: 133-.

[Alfaro04] Francisco José Alfaro, José L. Sánchez, José Duato: QoS in InfiniBand Subnetworks. IEEE Trans. Parallel Distrib. Syst. 15(9): 810-823 (2004).

[Alfaro07] Francisco José Alfaro, José L. Sánchez, M. Menduiña, José Duato: A Formal Model to Manage the InfiniBand Arbitration Tables Providing QoS. IEEE Trans. Computers 56(8): 1024-1039 (2007).

[Alfaro09] Francisco José Alfaro, José L. Sánchez, José Duato: A new strategy to manage the InfiniBand arbitration tables. J. Parallel Distrib. Comput. 69(6): 508-520 (2009).

[Anderson93] T. Anderson, S. Owicki, J. Saxe, and C. Thacker, "High-Speed Switch Scheduling for Local-Area Networks", ACM Trans. on Computer Systems, vol. 11, no. 4, pp. 319–352, Nov. 1993.

[AVI] http://innoavi.es/

[Baydal02] Elvira Baydal, Pedro López, José Duato: Congestion Control Based on Transmission Times. Euro-Par 2002: 781-790.

[Baydal05] Elvira Baydal, Pedro López, José Duato: A Family of Mechanisms for Congestion Control in Wormhole Networks. IEEE Trans. Parallel Distrib. Syst. 16(9): 772-784 (2005).

[Becker95] Becker, D., Sterling, T., et al. Beowulf: A Parallel Workstation for Scientific Computation, Proceedings, International Conference on Parallel Processing, Oconomowoc, Wisconsin, Aug. 1995, p. 11-14.

[Brakmo95] L. S. Brakmo and L. L. Peterson, "TCP Vegas: End To End Congestion Avoidance on a Global Internet", IEEE Journal on Selected Areas in Communication, vol.13, no. 8, pp. 1465–1480, Oct. 1995.

[Caminero02a] María Blanca Caminero, Carmen Carrión, Francisco J. Quiles, José Duato, Sudhakar Yalamanchili: A new switch scheduling algorithm to improve QoS in the multimedia router. IEEE Workshop on Multimedia Signal Processing 2002: 376-379.

[Caminero02b] María Blanca Caminero, Carmen Carrión, Francisco J. Quiles, José Duato, Sudhakar Yalamanchili: A multimedia router architecture to provide high performance and QoS guarantees to mixed traffic. ICME (1) 2002: 313-316.

[Caminero02c] María Blanca Caminero, Carmen Carrión, Francisco J. Quiles, José Duato, Sudhakar Yalamanchili: Investigating Switch Scheduling Algorithms to Support QoS in the Multimedia Router. IPDPS 2002.

[Caminero03] María Blanca Caminero, Carmen Carrión, Francisco J. Quiles, José Duato, Sudhakar Yalamanchili: A Solution for Handling Hybrid Traffic in Clustered Environments: The MultiMedia Router MMR. IPDPS 2003: 197.

[Caminero05] María Blanca Caminero, Carmen Carrión, Francisco J. Quiles, José Duato, Sudhakar Yalamanchili: Traffic Scheduling Solutions with QoS Support for an Input-Buffered MultiMedia Router. IEEE Trans. Parallel Distrib. Syst. 16(11): 1009-1021 (2005).

[Caminero06] María Blanca Caminero, Carmen Carrión, Francisco J. Quiles, José Duato, Sudhakar Yalamanchili: MMR: A MultiMedia Router architecture to support hybrid workloads. J. Parallel Distrib. Comput. 66(2): 307-321 (2006).

[CNEAI] http://www.aneca.es/Programas-de-evaluacion/Evaluacion-de-profesorado/CNEAI

[Congdon18] Paul Congdon (Ed.), Roger Marks, Jose Duato, Barak Gafni, Feng Gao, Liang Guo, Jie Li, Gu Rong, Richard Scheffenegger, Mehmet Toy, Sowmini Varadhan, Jianglong Wang, Ilan Yerushalmi, Yolanda Yu, IEEE 802 Nendica Report: The Lossless Network for Data Centers, Nendica Report 802.1 1-18-0042-00-ICne, The Institute of Electrical and Electronics Engineers, 2018. https://mentor.ieee.org/802.1/dcn/18/1-18-0042-00-ICne-ieee-802-nendica-report-the-lossless-network-for-data-centers.pdf

[Cotec] http://informecotec.es/media/Informe-Cotec_2018_versiónweb.pdf

[Dally87] W.J. Dally and C.L. Seitz, "Deadlock-free message routing in multiprocessor interconnection networks," IEEE Transactions on Computers, vol. C-36, no. 5, pp. 547-553, May 1987.

[Dally90] W.J. Dally, "Performance analysis of k-ary n-cube interconnection networks," IEEE Transactions on Computers, vol. C-39, no. 6, pp. 775-785, June 1990.

[Dally92] W.J. Dally, "Virtual-channel flow control," IEEE Transactions on Parallel and Distributed Systems, vol. 3, no. 2, pp. 194-205, March 1992.

[Dally98] W. J. Dally, P. Carvey, and L. Dennison, "The Avici Terabit Switch/Router", in Proc. Hot Interconnects 6, Aug. 1998.

[Dandamudi99] S.P. Dandamudi, "Reducing Hot-Spot Contention in Shared-Memory Multi- processor Systems," IEEE Concurrency, vol. 7, no 1, Jan. 1999, pp. 48-59.

[Danowitz12] Andrew Danowitz, Kyle Kelley, James Mao, John P. Stevenson, Mark Horowitz: CPU DB: Recording Microprocessor History. acmqueue 10 (4): 1-18 (2012).

[Dean12] J. Dean, G. S. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M. Z. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang and A. Y. Ng, "Large scale distributed deep networks," in Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1, Lake Tahoe, Nevada, 2012.

[Deng14] Deng, L.; Yu, D. (2014). "Deep Learning: Methods and Applications". Foundations and Trends in Signal Processing. 7 (3–4): 1–199.

[Domingo18] Esteban Domingo, Declaración sobre la financiación y gestión de la investigación científica en España, Real Academia de Ciencias Exactas, Físicas y Naturales, Noviembre 2018. http://www.rac.es/ficheros/doc/01171.pdf

[Duato93] J. Duato, A new theory of deadlock-free adaptive routing in wormhole networks, IEEE Transactions on Parallel and Distributed Systems, vol. 4, no. 12, pp. 1320-1331, Dec. 1993.

[Duato03] J. Duato, S. Yalamanchili, and L. M. Ni, Interconnection Networks: An Engineering Approach (Revised printing), Morgan Kaufmann Publishers, 2003.

[Duato05] José Duato, Ian Johnson, Jose Flich, Finbar Naven, Pedro Javier García, Teresa Nachiondo: A New Scalable and Cost-Effective Congestion Management Strategy for Lossless Multistage Interconnection Networks. HPCA 2005: 108-119.

[Duato10] José Duato, Antonio J. Peña, Federico Silla, Rafael Mayo, Enrique S. Quintana-Ortí: rCUDA: Reducing the number of GPU-based accelerators in high performance clusters. HPCS 2010: 224-231.

[ECMP] RFC 2991 - Multipath Issues in Unicast and Multicast Next-Hop Selection. IETF. November 2000. https://tools.ietf.org/html/rfc2991

[ECN] RFC 3168 - The Addition of Explicit Congestion Notification (ECN) to IP. IETF. September 2001. http://tools.ietf.org/html/rfc3168

[Escudero08] Jesús Escudero-Sahuquillo, Pedro Javier García, Francisco J. Quiles, Jose Flich, José Duato: FBICM: Efficient Congestion Management for High-Performance Networks Using Distributed Deterministic Routing. HiPC 2008: 503-517.

[Escudero10a] Jesús Escudero-Sahuquillo, Pedro Javier García, Francisco J. Quiles, José Duato: An Efficient Strategy for Reducing Head-of-Line Blocking in Fat-Trees. Euro-Par (2) 2010: 413-427.

[Escudero10b] Jesús Escudero-Sahuquillo, Pedro Javier García, Francisco J. Quiles, José Flich, José Duato: Cost-Effective Congestion Management for Interconnection Networks Using Distributed Deterministic Routing. ICPADS 2010: 355-364.

[Escudero11a] Jesús Escudero-Sahuquillo, Pedro Javier García, Francisco J. Quiles, José Flich, José Duato: Cost-effective queue schemes for reducing head-of-line blocking in fat-trees. Concurrency and Computation: Practice and Experience 23(17): 2235-2248 (2011).

[Escudero11b] Jesús Escudero-Sahuquillo, Pedro Javier García, Francisco J. Quiles, José Flich, José Duato: OBQA: Smart and cost-efficient queue scheme for Head-of-Line blocking elimination in fat-trees. J. Parallel Distrib. Comput. 71(11): 1460-1472 (2011).

[Escudero11c] Jesús Escudero-Sahuquillo, Ernst Gunnar Gran, Pedro Javier García, Jose Flich, Tor Skeie, Olav Lysne, Francisco J. Quiles, José Duato: Combining Congested-Flow Isolation and Injection Throttling in HPC Interconnection Networks. ICPP 2011: 662-672.

[Escudero13] Jesús Escudero-Sahuquillo, Pedro Javier García, Francisco J. Quiles, Jose Flich, José Duato: An Effective and Feasible Congestion Management Technique for High-Performance MINs with Tag-Based Distributed Routing. IEEE Trans. Parallel Distrib. Syst. 24(10): 1918-1929 (2013).

[Escudero14] Jesús Escudero-Sahuquillo, Pedro Javier García, Francisco J. Quiles, Sven-Arne Reinemo, Tor Skeie, Olav Lysne, José Duato: A new proposal to deal with congestion in InfiniBand-based fat-trees. J. Parallel Distrib. Comput. 74(1): 1802-1819 (2014).

[Escudero15] Jesús Escudero-Sahuquillo, Ernst Gunnar Gran, Pedro Javier García, Jose Flich, Tor Skeie, Olav Lysne, Francisco J. Quiles, José Duato: Efficient and Cost-Effective Hybrid Congestion Control for HPC Interconnection Networks. IEEE Trans. Parallel Distrib. Syst. 26(1): 107-119 (2015).

[Escudero18] Jesús Escudero-Sahuquillo, Pedro Javier García, Francisco J. Quiles, German Maglione Mathey, José Duato Marín: Feasible enhancements to congestion control in InfiniBand-based networks. J. Parallel Distrib. Comput. 112: 35-52 (2018).

[Ethernet] 802.3cc-2017 - IEEE Standard for Ethernet - Amendment 11: Physical Layer and Management Parameters for Serial 25 Gb/s Ethernet Operation Over Single-Mode Fiber. https://ieeexplore.ieee.org/document/8259171

[Ferrer07] Joan-Lluís Ferrer, Elvira Baydal, Antonio Robles, Pedro López, José Duato: Congestion Management in MINs through Marked and Validated Packets. PDP 2007: 254-261.

[Ferrer08] Joan-Lluís Ferrer, Elvira Baydal, Antonio Robles, Pedro López, José Duato: On the Influence of the Packet Marking and Injection Control Schemes in Congestion Management for MINs. Euro-Par 2008: 930-939.

[Ferrer10] Joan-Lluís Ferrer, Elvira Baydal, Antonio Robles, Pedro López, José Duato: A Scalable and Early Congestion Management Mechanism for MINs. PDP 2010: 43-50.

[Ferrer12] Joan-Lluís Ferrer, Elvira Baydal, Antonio Robles, Pedro López, José Duato: Progressive Congestion Management Based on Packet Marking and Validation Techniques. IEEE Trans. Computers 61(9): 1296-1310 (2012).

[Franco99] D. Franco, I. Garces, and E. Luque, “A New Method to Make Communication Latency Uniform: Distributed Routing

Balancing", in Proc. ACM International Conference on Supercomputing (ICS99), pp. 210–219, May 1999.

[García05a] Pedro Javier García, Jose Flich, José Duato, Francisco J. Quiles, Ian Johnson, Finbar Naven: On the Correct Sizing on Meshes Through an Effective Congestion Management Strategy. Euro-Par 2005: 1035-1045.

[García05b] Pedro Javier García, Jose Flich, José Duato, Ian Johnson, Francisco J. Quiles, Finbar Naven: Dynamic Evolution of Congestion Trees: Analysis and Impact on Switch Architecture. HiPEAC 2005: 266-285.

[García06a] Pedro Javier García, Francisco J. Quiles, Jose Flich, José Duato, Ian Johnson, Finbar Naven: Efficient, Scalable Congestion Management for Interconnection Networks. IEEE Micro 26(5): 52-66 (2006).

[García06b] Pedro Javier García, Francisco J. Quiles, Jose Flich, José Duato, Ian Johnson, Finbar Naven: RECN-DD: A Memory-Efficient Congestion Management Technique for Advanced Switching. ICPP 2006: 23-32.

[Gómez03] María Engracia Gómez, Jose Flich, Antonio Robles, Pedro López, José Duato: VOQSW: A Methodology to Reduce HOL Blocking in InfiniBand Networks. IPDPS 2003: 46.

[Gómez15] Crispín Gómez Requena, Francisco Gilabert Villamón, María Engracia Gómez, Pedro López, José Duato: A HoL-blocking aware mechanism for selecting the upward path in fat-tree topologies. The Journal of Supercomputing 71(7): 2339-2364 (2015).

[Gran10] E.G. Gran, M. Eimot, S.A. Reinemo, T. Skeie, O. Lysne, L. Huse, G. Shainer, "First experiences with congestion control in InfiniBand hardware", in Proceedings of IPDPS 2010, pp. 1–12.

[Green500] https://www.top500.org/green500/

[Guay11] Wei Lin Guay, Bartosz Bogdanski, Sven-Arne Reinemo, Olav Lysne, Tor Skeie: vFtree - A Fat-Tree Routing Algorithm Using Virtual Lanes to Alleviate Congestion. IPDPS 2011: 197-208.

[Gusat05] Mitchell Gusat, D. Craddock, Wolfgang E. Denzel, Antonius P. J. Engbersen, Nan Ni, G. Pfister, W. Rooney, José Duato: Congestion Control in InfiniBand Networks. Hot Interconnects 2005: 158-159.

[InfiniBand] InfiniBand Trade Association, “InfiniBand Architecture. Specification Volume 1. Release 1.0”. Available at http://www.infinibandta.com/.

[Isermann81] Rolf Isermann, Digital control systems, Springer-Verlag, 1981.

[Jain89] R. Jain, “A Delay-Based Approach for Congestion Avoidance in Interconnected Heterogeneus Computer Networks”, ACM Computer Communication Review, vol. 19, no. 5, pp. 56–71, Oct. 1989.

[Jalaparti13] V. Jalaparti, P. Bodik, S. Kandula, I. Menache, M. Rybalkin and C. Yan, "Speeding up distributed request-response workflows," in Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM, Hong Kong, China, 2013.

[Kapoor12] R. Kapoor, G. Porter, M. Tewari, G. M. Voelker and A. Vahdat, "Chronos: predictable low latency for data center applications," in Proceedings of the Third ACM Symposium on Cloud Computing, San Jose, California, 2012.

[Karol87] M. Karol, M. Hluchyj, and S. Morgen, Input versus output queueing on a space division switch, IEEE Transactions on Communications, vol. 35, no. 12, pp. 1347-1356, 1987.

[Katevenis98] M. Katevenis, D. Serpanos, E. Spyridakis, “Credit-Flow-Controlled ATM for MP Interconnection: the ATLAS I Single-Chip ATM Switch”, in Proc. 4th Int. Symp. on High-Performance Computer Architecture, pp. 47–56, Feb. 1998.

[Kermani79] P. Kermani and L. Kleinrock, “Virtual cut-through: A new computer communication switching technique,” Computer Networks, vol. 3, pp. 267-286, 1979.

[Kessler93] R.E. Kessler and J.L. Schwarzmeier, “CRAY T3D: A new dimension for Cray Research,” Proceedings of Compcon, pp. 176-182, 1993.

[Kleyman16] Bill Kleyman, Why Google Wants to Rethink Data Center Storage, DataCenter Knowledge, Mayo 2016. https://www.datacenterknowledge.com/archives/2016/05/02/google-wants-rethink-data-center-storage

[Klowden12] Klowden, K, Wolfe, M. 2012. State Technology and Science Index. Milken Institute, Washington DC.

[Kogge08] Peter Kogge, Keren Bergman, Shekhar Borkar, Dan Campbell, William Carlson, William Dally, Monty Denneau, Paul Franzon, William Harrod, Kerry Hill, Jon Hiller, Sherman Karp, Stephen Keckler, Dean Klein, Robert Lucas, Mark Richards, Al Scarpelli, Steven Scott, Allan Snavely, Thomas Sterling, R. Stanley Williams, Katherine Yelick, ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems, ResearchGate, January 2008, https://www.researchgate.net/publication/242366160_ExaScale_Computing_Study_Technology_Challenges_in_Achieving_Exascale_Systems

[Krishnan04] V. Krishnan and D. Mayhew, "A Localized Congestion Control Mechanism for PCI Express Advanced Switching Fabrics", in Proc. 12th IEEE Symp. on Hot Interconnects, Aug. 2004.

[Kruskal86] Clyde P. Kruskal, Marc Snir, A Unified Theory Of Interconnection Network Structure. Theoretical Computer Science 48 (1986) 75-94.

[Kurzweil05] R. Kurzweil, The Singularity Is Near: When Humans Transcend Biology, Penguin Publishing Group, 2005.

[Leiserson85] C.E. Leiserson, "Fat-trees: Universal networks for hardware-efficient supercomputing," IEEE Transactions on Computers, vol. C-34, pp. 892-901, Oct. 1985.

[Martínez05] Alejandro Martínez, Francisco José Alfaro, José L. Sánchez, José Duato: Providing Full QoS Support in Clusters Using Only Two VCs at the Switches. HiPC 2005: 158-169.

[Martínez06a] Alejandro Martínez, Pedro Javier García, Francisco José Alfaro, José L. Sánchez, Jose Flich, Francisco J. Quiles,

José Duato: Towards a Cost-Effective Interconnection Network Architecture with QoS and Congestion Management Support. Euro-Par 2006: 884-895.

[Martínez06b] Alejandro Martínez, Francisco José Alfaro, José L. Sánchez, José Duato: Full QoS Support with 2 VCs for Single-chip Switches. NCA 2006: 239-242.

[Martínez07a] Alejandro Martínez-Vicente, Pedro Javier García, Francisco José Alfaro, José L. Sánchez, Jose Flich, Francisco J. Quiles, José Duato: Integrated QoS Provision and Congestion Management for Interconnection Networks. Euro-Par 2007: 837-847

[Martínez07b] Alejandro Martínez, Francisco José Alfaro, José L. Sánchez, José Duato: Deadline-based QoS Algorithms for High-performance Networks. IPDPS 2007: 1-9.

[Martínez07c] Alejandro Martínez, Francisco José Alfaro, José L. Sánchez, José Duato: Efficient Switches with QoS Support for Clusters. IPDPS 2007: 1-6.

[Martínez08] Alejandro Martínez-Vicente, George Apostolopoulos, Francisco José Alfaro, José L. Sánchez, José Duato: Efficient Deadline-Based QoS Algorithms for High-Performance Networks. IEEE Trans. Computers 57(7): 928-939 (2008).

[Martínez09] Alejandro Martínez, Pedro Javier García, Francisco José Alfaro, José L. Sánchez, Jose Flich, Francisco J. Quiles, José Duato: A Switch Architecture Guaranteeing QoS Provision and HOL Blocking Elimination. IEEE Trans. Parallel Distrib. Syst. 20(1): 13-24 (2009).

[MECD] https://www.mecd.gob.es/servicios-al-ciudadano-mecd/catalogo/general/educacion/050920/ficha.html

[Montazeri18] Behnam Montazeri, Yilong Li, Mohammad Alizadeh, and John Ousterhout, Homa: A Receiver-Driven Low-Latency Transport Protocol Using Network Priorities, Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication

(ACM SIGCOMM 2018), pp. 221-235, August 2018. https://arxiv.org/pdf/1803.09615.pdf

[Mora07] Gaspar Mora, Pedro Javier García, Jose Flich, José Duato: RECN-IQ: A Cost-Effective Input-Queued Switch Architecture with Congestion Management. ICPP 2007: 74.

[Nachiondo05a] Teresa Nachiondo, Jose Flich, José Duato, Mitchell Gusat: Cost / Performance Trade-Offs and Fairness Evaluation of Queue Mapping Policies. Euro-Par 2005: 1024-1034.

[Nachiondo05b] Teresa Nachiondo, Jose Flich, José Duato: Efficient Reduction of HOL Blocking in Multistage Networks. IPDPS 2005.

[Nachiondo06] Teresa Nachiondo, Jose Flich, José Duato: Destination-Based HoL Blocking Elimination. ICPADS (1) 2006: 213-222.

[Nachiondo10] Teresa Nachiondo, Jose Flich, José Duato: Buffer Management Strategies to Reduce HoL Blocking. IEEE Trans. Parallel Distrib. Syst. 21(6): 739-753 (2010).

[OECD] https://data.oecd.org/rd/gross-domestic-spending-on-r-d.htm

[Peña14] Antonio J. Peña, Carlos Reaño, Federico Silla, Rafael Mayo, Enrique S. Quintana-Ortí, José Duato: A complete and efficient CUDA-sharing solution for HPC clusters. Parallel Computing 40(10): 574-588 (2014).

[Peñaranda13] Roberto Peñaranda, Crispín Gómez Requena, María Engracia Gómez, Pedro López, José Duato: Deterministic Routing with HoL-Blocking-Awareness for Direct Topologies. ICCS 2013: 2521-2524.

[Peñaranda14] Roberto Peñaranda Cebrian, Crispín Gómez Requena, María Engracia Gómez Requena, Pedro Juan López Rodríguez, José Duato Marín: HoL-Blocking Avoidance Routing Algorithms in Direct Topologies. HPCC/CSS/ICESS 2014: 11-18.

[Peterson00] Peterson, Larry L. and Davie, Bruce S. "Computer Networks: A Systems Approach", Morgan Kaufmann, 2000.

[Pfister85] G. Pfister and A. Norton, “Hot Spot Contention and Combining in Multistage Interconnect Networks”, IEEE Trans. on Computers, vol. C-34, pp. 943–948, Oct. 1985.

[QoS] E.800: Definitions of terms related to quality of service. ITU-T Recommendation. August 1994. Updated September 2008. https://www.itu.int/rec/T-REC-E.800-200809-I/en

[Regalado11] Antonio Regalado, Who Coined ’Cloud Computing’?, MIT Technology Review, October 2011. https://www.technologyreview.com/s/425970/who-coined-cloud-computing/

[Scott94] S.L. Scott and G. Thorson, “Optimized routing in the Cray T3D,” Proceedings of the Workshop on Parallel Computer Routing and Communication, pp. 281-294, May 1994.

[Scott96] S.L. Scott and G. Thorson, “The Cray T3E network: Adaptive routing in a high performance 3D torus,” Proceedings of Hot Interconnects Symposium IV, Aug. 1996.

[Singh04] A. Singh, W. J. Dally, B. Towles, A. K. Gupta, “Globally Adaptive Load-Balanced Routing on Tori”, Computer Architecture Letters, vol. 3, no. 1, pp. 6–9, July 2004.

[Smai98] A. Smai and L. Thorelli, “Global Reactive Congestion Control in Multicomputer Networks”, in Proc. 5th Int. Conf. on High Performance Computing, 1998.

[Subramaniam13] Balaji Subramaniam, Winston Saunders, Tom Scogland, Wu-chun Feng. Trends in Energy-Efficient Computing: A Perspective from the Green500. In Proceedings of the 4th International Green Computing Conference, Arlington, VA, June 2013.

[Tamir92] Y. Tamir and G.L. Frazier, “Dynamically-Allocated Multi-Queue Buffers for VLSI Communication Switches,” IEEE Trans. Computers, vol. 41, no. 6, June 1992, pp. 725-737.

[Thottethodi01] M. Thottethodi, A. R. Lebeck, S. S. Mukherjee, “Self-Tuned Congestion Control for Multiprocessor Networks”, in Proc. Int. Symp. High-Performance Computer Architecture, Feb. 2001.

[Vogels00] W. Vogels et al, “Tree-Saturation Control in the AC3 Velocity Cluster Interconnect”, in Proc. 8th Conference on Hot Interconnects, Aug. 2000.

[Yang95] C. Q. Yang and A. V. S. Reddy, “A Taxonomy for Congestion Control Algorithms in Packet Switching Networks”, IEEE Network, pp. 34–45, July/Aug. 1995.

[Yébenes13] Pedro Yébenes Segura, Jesús Escudero-Sahuquillo, Crispín Gómez Requena, Pedro Javier García, Francisco J. Quiles, José Duato: BBQ: A Straightforward Queuing Scheme to Reduce HoL-Blocking in High-Performance Hybrid Networks. Euro-Par 2013: 699-712.

[Yébenes14] Pedro Yébenes, Jesús Escudero-Sahuquillo, Crispín Gómez Requena, Pedro Javier García, Francisco J. Alfaro, Francisco J. Quiles, José Duato: Combining HoL-blocking avoidance and differentiated services in high-speed interconnects. HiPC 2014: 1-10.

[Yew87] P. Yew, N. Tzeng, and D.H. Lawrie, “Distributing Hot-Spot Addressing in Large-Scale Multiprocessors,” IEEE Trans. Computers, vol. 36, no. 4, Apr. 1987, pp. 388-395.

[Zhang10] Qi Zhang, Lu Cheng, Raouf Boutaba, Cloud computing: state-of-the-art and research challenges, J. Internet Serv. Appl. (2010) 1: 7–18.

DISCURSO DE CONTESTACIÓN

DEL

EXCMO. SR. D. JAVIER JIMÉNEZ SENDÍN

Excmo. Sr. Presidente,
Excmas. Sras. y Srs. Académicos,
Queridos amigos,

Me corresponde hoy el honor de dar la bienvenida a José Duato Marín en nombre de la Academia, y tengo que reconocer que me hace ilusión. En primer lugar porque me brinda la oportunidad de presentar a una persona excepcionalmente brillante, que no puede por menos que enriquecer nuestra casa, pero también porque el nuevo académico pertenece a una disciplina que estaba poco representada entre nosotros, la Informática, por la que tengo un cierto cariño personal como usuario. Como todos sabemos, las Academias aparecieron en los siglos XVII y XVIII para compensar en parte la osificación de las Universidades, a las que, por aquel entonces, les costaba liberarse de la obligación de transmitir la sabiduría clásica. Era una época en la que surgían con frecuencia nuevos descubrimientos, algunos de los cuales contradecían ideas con mucha tradición intelectual, y hacían falta instituciones que les diesen cobijo y difusión. El invento tuvo éxito, y las Academias se convirtieron durante bastante tiempo en el centro de la innovación científica. Los premios concedidos a científicos jóvenes por la Academie Française y por la Royal Society de Londres fueron en gran parte responsables de que las matemáticas y las ciencias experimentales sean hoy lo que son. Desde entonces, el ámbito de la Ciencia ha crecido mucho y a las academias relativamente pequeñas, como la nuestra, nos vuelve a costar incorporar temas nuevos, entre otras cosas porque parece que, cuando lo hacemos, va en detrimento de los temas de toda la vida, lo que no tendría sentido. Pero también porque no siempre está claro dónde colocar las nuevas disciplinas. Algunos informáticos me dicen que José Duato se dedica a lo que ellos llaman 'hardware', aunque a mí algunos de sus artículos me parecen más bien Matemáticas discretas bastante abstractas. Y, siendo así, ¿dónde debería encuadrarse la Informática? ¿En la sección de Física y Química, por lo del hardware? ¿O en la de Exactas, por lo de las matemáticas? En la duda, me alegra personalmente que la sección de ciencias Exactas se haya adelantado a otras secciones para 'robarles' al nuevo académico, y para acoger un área de conocimiento con un innegable futuro, pero estoy seguro que a todos se nos ocurren ejemplos de disciplinas que aún no están bien representadas en la Academia, y que deberían estarlo, y quizá sea el momento de empezar a pensar en mecanismos específicos para atraerlas a nuestra casa.

Pero ya es hora de dejar de hablar de la Academia, y de ocuparse de nuestro nuevo miembro. José Duato nació en Alberic, en la Ribera Alta de Valencia. Hizo el bachillerato en Carcaixent y el COU en Alzira, y pasó a más tarde a la Universidad Politécnica de Valencia para estudiar Ingeniería Industrial. Cul-

minó la carrera con el premio nacional de finalización de estudios, se doctoró en la misma Universidad, y ha permanecido allí desde entonces, con algunas cortas interrupciones en el extranjero. En la actualidad es catedrático de Arquitectura y Tecnología de Computadores en el departamento de Informática de Sistemas y Computadores de la Escuela Técnica Superior de Ingeniería Informática, donde dirige un grupo numeroso de investigación sobre arquitecturas paralelas y redes de interconexión. Ha ejercido también labores de responsabilidad en el ámbito universitario y en política científica, habiendo llegado a ser Decano de la Facultad de Informática y Vicerrector de Investigación y Desarrollo Tecnológico. En la actualidad preside el Comité Estratégico de Innovación Especializado en Tecnologías Habilitadoras de la Agencia Valenciana de Innovación. Ha recibido numerosos premios, entre los que destacan el premio Rey Jaime I de Nuevas Tecnologías en 2006 y el premio Nacional de Investigación Julio Rey Pastor en 2009.

Su actividad no se ha restringido a España. Ha sido profesor adjunto de la universidad del estado de Ohio, visitante en el laboratorio nacional de Los Álamos, en EEUU, e investigador a tiempo parcial en el laboratorio Simula en Oslo. Ha sido invitado a organizar y a contribuir a conferencias internacionales de prestigio, y ha colaborado con empresas informáticas de todos conocidas, como IBM, Huawei, Sun Microsystems y Mellanox. Ha sido editor de varias revistas internacionales sobre arquitectura de ordenadores, y sigue contribuyendo hasta hoy al diseño de los estándares internacionales para redes de interconexión.

Según nos ha dicho en la primera parte de su discurso, su investigación se ha centrado en el estudio de estas redes. Nos ha explicado su importancia en la informática y en las comunicaciones, y ha hablado algo, aunque menos, sobre las redes que podríamos llamar naturales, desde la ecología a la sociología. Estas últimas son temas que aparecen cada vez más en la vida diaria, y sin duda podemos esperar del nuevo académico que nos explique algún día, por lo menos, si lo que nos dice la prensa sobre ellas tiene sentido o no. Personalmente, poco puedo añadir a un asunto del que no sé casi nada, pero hay dos cosas que querría resaltar.

La primera es el carácter esencialmente aplicado de la investigación de José Duato. Como dije más arriba, una gran parte de sus trabajos los ha llevado a cabo en colaboración con empresas, y en su currículo aparecen más de 400 artículos científicos y un libro de texto importante, pero también siete patentes, cinco de ellas en Estados Unidos, en las que el primer autor es José Duato. Este sesgo es importante en España, donde falta muchas veces el transvase entre la investigación pura y la aplicada, y donde se espera a menudo que el investiga-

dor científico lo haga todo, desde el desarrollo fundamental al industrial. Creo que nuestro país debería estar especialmente agradecido a los investigadores dispuestos a suplir una parte del trabajo que correspondería a los laboratorios de desarrollo y a las empresas, y debemos alegrarnos de que nuestra Academia dé hoy la bienvenida a uno de ellos.

Lo segundo que querría resaltar es más personal. Ya he dicho al principio de esta presentación que mi relación con la informática es la de un usuario agradecido por lo que los ordenadores me han permitido hacer durante mi carrera. He tenido la ocasión y la necesidad de usar grandes ordenadores, y, a partir de los años ochenta del siglo pasado, eso ha significado lidiar con ordenadores paralelos. Los he programado, los he explotado, y, en algunas ocasiones los he tenido que comprar. Siempre han tenido dos partes. Primero estaban los procesadores, de los que en seguida hubo varios miles, pero que, al menos hasta hace unos años, daban pocos quebraderos de cabeza y eran cada vez más rápidos. Luego estaba la red de interconexión entre esos procesadores, que era la otra parte fundamental de la máquina, y que claramente estaba mal resuelta. Era siempre la parte más cara y más lenta, y no hacía más que bloquearse y dar problemas. Un día, hace unos diez años, mi grupo consiguió unas cuantas horas en una máquina nueva, un BlueGene/P, y al poco tiempo de usarla mis estudiantes llegaron a contarme que aquella máquina era distinta a las que habíamos usado hasta entonces. Había que ocuparse otra vez de los procesadores, que eran lentos y más bien pequeños, pero la red de interconexión funcionaba de maravilla. Era rápida y no presentaba ningún síntoma de atascarse. De golpe, nos olvidamos de la pesadilla de llevar los datos de un procesador a otro, y todos interpretamos que por fin había habido alguien, en aquel caso IBM, que había entendido cómo hacer una red, y que era una pena que unos señores tan listos no pudieran encontrarse más que en las grandes universidades y empresas de Estados Unidos. Unos años más tarde, con ocasión de una conferencia en esta casa, descubrí que entre los responsables de aquella red estaba José Duato. José, en nombre mío y de mi equipo de entonces, no puedo por menos que agradecerte lo que hiciste con aquel diseño por los sufridos usuarios.

La segunda parte del discurso del nuevo académico se ha centrado en cómo organizar la política científica en España. No voy a repetir lo que nos ha dicho, ni mucho menos a juzgarlo, pero si querría añadir a sus consideraciones mi desconsuelo por la tragedia de la emigración de nuestros mejores jóvenes científicos, que no es una tragedia para ellos, como se dice a menudo, sino para los que nos quedamos aquí. Y querría sumar mi voz a la de los que claman por la falta de inversión de nuestro país en las generaciones futuras de investigadores, que nos empobrece y nos condena a medio plazo a la categoría de país

menor. No se puede pedir a nuestros estudiantes que renuncien a carreras en el extranjero porque la patria les llama, ni se les puede tratar económicamente como maestros no especializados cuando se quedan aquí, o cuando vuelven. Todos sabemos que, cada vez más, la patria de la ciencia es el mundo, y que el talento acabará yéndose a donde estén las oportunidades.

Pero ha habido en esta segunda parte del discurso otro tema que me ha interesado especialmente, que es la dicotomía entre la 'ciencia por la ciencia' y las ciencias aplicadas. No tanto en el aspecto práctico de esta distinción, del que ya he hablado, sino en la discusión sobre si se rebaja la calidad de la empresa investigadora cuando nos distraemos prestando atención a la posible aplicación de lo que hacemos. Es un tema antiguo, desgraciadamente especialmente extendido en Matemáticas, donde hay una larga tradición que considera a las matemáticas puras como radicalmente distintas (y superiores) a las matemáticas aplicadas. Pero que se da también en otras ciencias, donde siempre ha habido corrientes que sostienen que la investigación básica es la única investigación real, y que las aplicaciones llegarán eventualmente, aunque no nos ocupemos de ellas. Todo eso es verdad, pero Duato nos ha recordado que el asunto es más complicado, y que, más allá de las consideraciones 'morales' de cómo ser más útil a la sociedad que nos sostiene, las aplicaciones son fuentes de problemas muy interesantes, y que hay una satisfacción interior en ver cómo vuela 'tu avión' o cómo una ambulancia cruza 'tu puente', que puede ser comparable a la satisfacción estética que se deriva de haber demostrado el último teorema de Fermat. Al final, la elección es personal, y es importante que la reconozcamos como tal, y que ninguna de las dos partes desprecie a la otra por su elección.

Intentar prohibir, o dejar de apoyar a, la ciencia por la ciencia sería como prohibir a los poetas, y, como alguien dijo en una comisión que discutía la financiación de un instrumento científico especialmente caro y de utilidad poco evidente, "no se trata de cómo la poesía contribuye a la mejora de nuestro país, sino de si merecería la pena mejorar un país en el que no hubiera poesía"[1]. Dicho lo cual, tampoco podemos olvidar que un país que únicamente mantuviese poetas quizá resultase un país bastante pobre.

José Duato nos ha recordado que se puede hacer investigación de altura con resultados prácticos inmediatos y, aunque sólo fuera por continuar esa discusión con él durante más tiempo, sería importante acogerle en nuestra Academia. Pero, tal como nos ha demostrado en su discurso, y como seguro que

[1] R.R. Wilson, director de Fermilab, contestando en 1969 frente al congreso de EEUU a la pregunta de si la financiación de un nuevo acelerador sería útil para la defensa nacional: "No tiene nada que ver con la defensa del país,... sino con hacer el país más digno de ser defendido".

nos seguirá demostrando en el futuro, no hay duda de que nos aportará mucho más.

Por ello, en nombre de la Academia y en el mío propio, sé bienvenido entre nosotros.