Modelado de una Red Fotónica para Computación Exascale

José Duro, Salvador Petit, Julio Sahuquillo and María E. Gómez[1]

Resumen— **La tecnología fotónica se está convirtiendo en una prometedora y viable alternativa para las redes on-chip y off-chip en los futuros sistemas Exascale. Sin embargo, esta tecnología no es lo suficientemente madura todavía, por lo que esfuerzos centrados en las redes fotónicas siguen siendo necesarios para lograr implementaciones apropiadas. Los simuladores de redes fotónicas a nivel de sistema pueden ayudar a guiar a los diseñadores a evaluar las múltiples opciones.**

La mayoría de las investigaciones actuales se realizan en simuladores desarrollados para redes eléctricas, cuyos componentes difieren mucho de los componentes fotónicos. Además, la tecnología fotónica añade nuevos componentes que no están presentes en las redes eléctricas. Este artículo discute cómo se puede construir una herramienta de simulación de redes fotónicas a partir de un simulador de redes eléctricas. Se resume y compara el comportamiento de ambas tecnologías –eléctrica y fotónica– y se discuten las extensiones propuestas. Entre otros aspectos, las extensiones propuestas modelan routers ópticos, wavelength-division multiplexing, conmutación de circuitos y algoritmos de encaminamiento.

Este trabajo tiene como objetivo avanzar en la investigación de redes ópticas off-chip en el contexto del proyecto European Exascale System Interconnect and Storage (ExaNeSt). Los experimentos presentados en este artículo estudian múltiples configuraciones de redes fotónicas realistas y se han ejecutado con fragmentos de trazas reales. Los resultados obtenidos muestran que, comparadas con las redes eléctricas, las redes ópticas pueden reducir el tiempo de ejecución de las cargas en varios órdenes de magnitud. Nuestro estudio revela que las futuras tecnologías ópticas que presentan un ancho de banda agregado de 3.2 Tbps no proporcionan beneficios en rendimiento sobre enlaces ópticos de 1.6 Tbps en las cargas estudiadas, sino que 1.6 Tbps es suficiente para alcanzar el rendimiento máximo. Atendiendo a la configuración de los enlaces, el ancho de banda por cada canal óptico es el parámetro que mayor impacto tiene tanto en los retardos en la red como en el tiempo de ejecución. Por otro lado, para un ancho de banda por canal óptico dado, la mejor estrategia es reducir el tamaño del phit.

Palabras clave— **Fotónica, Simulación, Exascale**

I. Introducción

LOS supercomputadores más potentes del mundo [1] se clasifican por su potencia de cálculo en términos de operaciones de punto flotante ejecutadas por segundo (FLOPS). El Sunway TaihuLight, el superordenador que encabeza la lista en noviembre de 2016, alcanza los 93 PetaFlops (10^{15}) con 10,5 millones de núcleos. La lista Top500 recoge la potencia computacional de los supercomputadores desde 1971 y, según las tendencias actuales, se espera que los supercomputadores rompan la barrera ExaFlop (10^{18}) para 2020. Alcanzar este objetivo, sin embargo, es un desafío y requiere de múltiples soluciones simultáneas que aborden, entre otros, la computación a nivel de chip (nodos del sistema), el movimiento de datos a través del sistema, el almacenamiento distribuido, la gestión de energía, etc.

El desafío del movimiento de datos es probablemente la tarea más crítica a lograr, principalmente debido al creciente número de nodos de computación y, por lo tanto, al aumento de los requerimientos de comunicación. Las redes Exascale contarán con miles de nodos, por lo que la transmisión de datos se convierte en un importante aspecto de diseño ya que las nuevas necesidades no sólo aumentan en términos de rendimiento, sino también en eficiencia energética. En estos sistemas, la tecnología de red subyacente [2], [3] es una elección de diseño crítica.

En este sentido, las interconexiones fotónicas - tanto on-chip como off-chip- han surgido como una alternativa tecnológica que aborda las restricciones clave de las redes eléctricas tradicionales. Esta tecnología proporciona mucho más ancho de banda que la tecnología eléctrica con mucho menos consumo energético [4], [5]. Se espera que las redes ópticas on-chip (ONoCs) [6], [7] se conviertan en una opción viable para la creciente demanda de aplicaciones de computación de alto rendimiento (HPC) que las redes eléctricas no pueden manejar eficientemente. Por otro lado, la tecnología fotónica off-chip puede proporcionar lo necesario para cubrir los crecientes requerimientos en computación Exascale, aportando ventajas adicionales sobre la tecnología eléctrica en ancho de banda de los enlaces (inter e intra-rack) y al ahorro de energía.

El desafío Exascale se ha convertido en una gran preocupación para la industria de los supercomputadores y los países de todo el mundo. En este sentido, algunos proyectos de investigación se han incrementado en Europa para abordar los desafíos involucrados. Este trabajo se ha desarrollado bajo el paraguas del proyecto europeo Exascale System Interconnect and Storage (ExaNeSt) [8], cuyo objetivo principal es construir un sistema capaz de escalar hasta decenas de millones de núcleos ARM de bajo consumo para resolver grandes cálculos científicos y tratar grandes cantidades de datos [9].

El desarrollo de tal sistema requiere de múltiples estudios de evaluación para guiar la construcción del sistema. Con respecto a las interconexiones, los entornos de simulación se utilizan para evaluar el ancho de banda de la red y la latencia de la red. Sin embargo, los entornos de simulación tradicionales modelan redes de interconexión eléctrica y no están preparados para simular redes con tec-

[1]Departamento de Ingeniería de Sistemas y Computadores, Universitat Politècnica de València, e-mail: jodugo1@gap.upv.es

nología fotónica.

El modelado de redes fotónicas sobre un entorno de simulación eléctrica no es un proceso sencillo, y requiere de un sólido conocimiento sobre los fundamentos de la fotónica y las redes eléctricas. Muchos aspectos son muy diferentes ya que la fotónica ofrece nuevas posibilidades a la vez que está limitada en otros aspectos. Por ejemplo, la fotónica proporciona la capacidad de enviar múltiples mensajes simultáneamente a través de un enlace, pero no soporta el almacenamiento de flits en los routers de la red.

Este trabajo discute los pasos que se han llevado a cabo para modelar interconexiones fotónicas mediante la ampliación del simulador de redes utilizado en ExaNeSt, INSEE [10]. El código de INSEE original ha sido ampliamente modificado con el fin de i) modelar los componentes principales de la red fotónica tales como los enlaces de red o los routers; ii) modelar las particularidades de las comunicaciones fotónicas como la conmutación de circuitos y la multiplexación por división de longitud de onda (técnología Wavelength Division-Multiplexing o WDM); iii) modelar el encaminamiento fotónico; y iv) adaptar las topologías de red para imitar adecuadamente el comportamiento de la fotónica.

II. Background sobre la Tecnología Fotónica

Como se ha mencionado anteriormente, se espera que durante la década actual aumente el rendimiento de la red para soportar la computación Exascale. En esta sección, resumimos los avances recientes en la tecnología fotónica y el estado actual de las interconexiones dentro y fuera del chip.

A. Fotónica fuera del Chip

Las interconexiones basadas en tecnología fotónica están siendo ampliamente implementadas en sistemas de comunicaciones debido a su potencial para lograr una integración a bajo coste y consumo energético. Esto se debe a un mayor ancho de banda, un mejor compromiso entre distancia y velocidad y una gestión de cables más fácil. Respecto al ancho de banda, lograr más de 10 Gbps con cables de cobre convencionales sigue siendo un desafío, mientras que una sola fibra óptica puede ofrecer anchos de banda en la escala de los Terahercios. Con respecto al compromiso entre distancia y velocidad, las fibras ópticas son capaces de transmitir datos a lo largo de varios kilómetros sin penalizaciones de ancho de banda. Por último, debido a su ligereza y delgadez inherentes, el uso de cables de fibra óptica en lugar de cables de cobre, reduce considerablemente la densidad del cableado que facilita considerablemente su gestión.

Sin embargo, las actuales tecnologías fotónicas de última generación requieren de transceptores para transformar las señales eléctricas en señales ópticas y viceversa, lo que limita la implantación de un sistema completamente fotónico. Para hacer frente a esta deficiencia, la investigación actual tiene como objetivo integrar dispositivos ópticos en chips. Este objetivo aún no se ha alcanzado y tendría un impacto significativo en la topología de las redes de comunicaciones.

El límite de ancho de banda de los actuales transceptores QSFP (Quad Small Form-factor Pluggable) basados en la tecnología VCSEL (Vertical-Cavity Surface Emitting Laser) es de 40 Gbps. Sin embargo, se espera que las interconexiones fotónicas alcancen los 100 Gbps o más en un futuro próximo. Por ejemplo, Intel Corporation y otras grandes empresas como IBM o Cisco Systems han producido pototipos que funcionan a velocidades de hasta 100 Gbps. Por otra parte, Intel y Corning están desarrollando actualmente el conector MXC, que soporta hasta 64 fibras comunicándose a 25 Gbps, alcanzando una capacidad de transmisión de datos sin precedentes de 1,6 Tbps a una distancia de 300 metros.

B. Fotónica dentro del Chip

La necesidad de bajas latencias y de altos anchos de banda en las transmisiones de datos entre procesadores multinúcleo ha llevado a plantear la fotónica compatible con CMOS como alternativa para el diseño en las redes en chip. Por otra parte, las redes fotónicas dentro del chip permiten la implementación de routers fotónicos en silicio, los cuales son un desarrollo clave para la implementación de redes completamente ópticas inter-rack e intra-rack en sistemas Exascale. En este sentido, los esfuerzos actuales se han concentrado en la fabricación en el silicio de láseres confiables, moduladores electro-ópticos, anillos resonadores y receptores; esto es, los componentes más críticos de los circuitos fotónicos.

Los láseres inyectan luz en las guías de onda del chip. Los láseres son probablemente los dispositivos más difíciles de integrar en el silicio. *Duan et al.* han desarrollado láseres híbridos de silicio / III-V con menor consumo que en trabajos anteriores [11], [12], aunque todavía no alcanzan un consumo lo suficientemente bajo para poder reducir significativamente los costes de encapsulado.

Los moduladores electro-ópticos establecen la capacidad de conmutación, es decir, el ancho de banda de funcionamiento de cualquier circuito integrado fotónico (PIC).

Los anillos resonadores ópticos son el componente clave para aprovechar la tecnología de multiplexación por división de longitud de onda (Wavelength-Division Multiplexing o WDM) [13]. WDM permite dividir la señal óptica en múltiples longitudes de onda independientes. Un anillo resonador captura longitudes de onda específicas; por lo tanto, puede redirigir estas longitudes de onda a otros waveguides y receptores, permitiendo así la implementación de complejas redes ópticas en routers fotónicos y otros tipos de circuitos integrados.

Por último, ya se han integrado receptores ópticos coherentes (también conocidos como fotodetectores) que convierten la amplitud, la fase y la polarización

de una señal óptica en el dominio eléctrico y proporcionan tasas de conversión de datos muy altas (hasta 224 Gbps con señales PDM-16-QAM) [14], [15].

III. Entorno de Simulación INSEE

Esta sección resume las principales características del entorno de simulación y evaluación de redes de interconexión INSEE [10], que se ha extendido para modelar redes fotónicas. El simulador INSEE fue desarrollado originalmente en la Universidad de Manchester con el objetivo de modelar redes eléctricas. Este entorno es de dominio público y es una de las principales plataformas de simulación del proyecto ExaNeSt [9].

INSEE implementa múltiples topologías y permite múltiples métodos de generación de tráfico (e.j. sintéticos, trazas u otros simuladores). Tal flexibilidad no se da en otros simuladores existentes [16]. INSEE consigue esta flexibilidad con un uso eficiente de los recursos en términos de memoria y potencia computacional, permitiendo la simulación de grandes sistemas (por ejemplo, con decenas de miles de nodos) en pocos días.

El entorno de simulación consiste en un simulador funcional (FSIN) y un generador de tráfico (TrGen). Ambos componentes cuentan con un diseño modular que se puede extender con nuevos módulos o ampliar las capacidades de los existentes. Hemos aprovechado este diseño modular para implementar múltiples extensiones, permitiendo a INSEE modelar redes totalmente ópticas.

Como simulador diseñado originalmente para estudiar redes eléctricas, INSEE implementa sus principales componentes, la mayoría de los cuales pueden reutilizarse para modelar redes fotónicas. Sin embargo, hay componentes clave que deben ser modificados considerablemente para imitar el comportamiento de la tecnología y componentes fotónicos. Este es el caso de los routers, las técnicas de conmutación y los enlaces, además de otras técnicas que sólo se aplican en la tecnología fotónica. Por ejemplo, la tecnología fotónica actual permite utilizar un solo enlace de fibra físico con múltiples canales al mismo tiempo.

IV. Extensiones Fotónicas Propuestas

Esta sección presenta las extensiones de INSEE desarrolladas para modelar redes fotónicas. Para este propósito, discutimos brevemente el comportamiento de los componentes principales de una red de altas prestaciones. Este análisis se lleva a cabo basándose en la tecnología subyacente (es decir, eléctrica o fotónica).

A. Routers Ópticos frente a Routers Eléctricos

Los routers eléctricos implementan memorias intermedias internas (buffers) que proporcionan almacenamiento temporal local para paquetes en tránsito (o unidades de datos más pequeñas, dependiendo de la técnica de conmutación). Los paquetes se guardan en estos buffers en caso de que no puedan avanzar debido a restricciones de tráfico o de red.

En comparación, las redes ópticas no proporcionan capacidad de almacenamiento en los routers fotónicos, lo que significa que una vez que los datos se inyectan en la red deben viajar sin experimentar bloqueos. Una solución para hacer frente a este inconveniente es el uso de routers electro-ópticos híbridos. Este enfoque, sin embargo, requiere de conversiones electro-ópticas y opto-eléctricas para escribir y leer en los buffers, limitando las mejoras en ancho de banda y latencias. Como se analiza más adelante, el hecho de que las redes totalmente ópticas no proporcionen soporte de almacenamiento intermedio, tiene consecuencias importantes para las técnicas de conmutación utilizadas en este tipo de redes.

B. Conmutación de Circuitos frente a Conmutación de Paquetes

En las redes eléctricas se pueden utilizar dos técnicas principales de conmutación, conmutación de circuitos y conmutación de paquetes. En la primera, se establece un circuito que se utiliza para transmitir el mensaje; así, el encaminamiento, el arbitraje y la conmutación se realizan una sola vez para cada mensaje. En la segunda, el encaminamiento, el arbitraje y la conmutación se realizan por cada paquete.

Obsérvese que la conmutación de paquetes no puede ser soportada por un diseño puramente óptico, ya que, como se mencionó anteriormente, el almacenamiento en buffers no es compatible con los routers ópticos. Por otro lado, la conmutación de circuitos no se implementa en las redes eléctricas modernas, ya que puede desaprovechar gran parte del ancho de banda de red. Esto significa que la conmutación de circuitos no suele ser modelada en los simuladores de red actuales, por lo que se amplió INSEE con el propósito de soportar esta técnica.

C. Wavelength-Division Multiplexing

Aunque la conmutación de circuitos se considera un método de conmutación prácticamente obsoleto, tiene importantes ventajas en las redes ópticas que soportan multiplexación por división de longitud de onda o wavelengt-division multiplexing (WDM). WDM es una técnica utilizada en comunicaciones ópticas que permite multiplexar en frecuencia varias longitudes de onda en la misma fibra óptica. La cantidad de longitudes de onda multiplexadas depende de la separación entre ellas (por ejemplo, con separaciones estándar de 100, 50 ó 25 GHz se pueden multiplexar hasta 40, 80 ó 160 longitudes de onda, respectivamente). Cuando se utiliza WDM, el ancho de banda total del enlace (es decir, considerando todas las longitudes de onda), conocido como ancho de banda agregado, está dado por la suma de los anchos de banda proporcionados por cada longitud de onda individual.

Para modelar tanto WDM como conmutación de circuitos juntos, se han superado varios problemas de diseño. En primer lugar, ya que hay múltiples longi-

tudes de onda en el mismo enlace, la conmutación clásica de circuitos necesita ser adaptada ya que puede definirse más de un circuito por enlace al mismo tiempo, es decir, cada canal (definido como un conjunto de longitudes de onda) puede ser parte de un posible circuito. Por lo tanto, cuando un mensaje está listo para ser inyectado en la red, el número de rutas posibles que puede reservar es mucho mayor que en las redes eléctricas.

D. Enlaces Fotónicos frente a Enlaces Eléctricos

Como se ha explicado anteriormente, los enlaces eléctricos sólo permiten enviar información de un único mensaje o paquete en un ciclo de red. En comparación, los enlaces ópticos están divididos en canales[1], cada uno de los cuales usa un conjunto diferente de longitudes de onda.

Para permitir múltiples canales, hemos desarrollado en INSEE soporte para configurar el número de longitudes de onda por enlace óptico y el reparto de estas longitudes de onda en canales independientes, cada uno dedicado a la transmisión de un mensaje diferente. En resumen, la primera opción de configuración permite especificar el ancho de banda del enlace y la segunda el ancho de banda por canal.

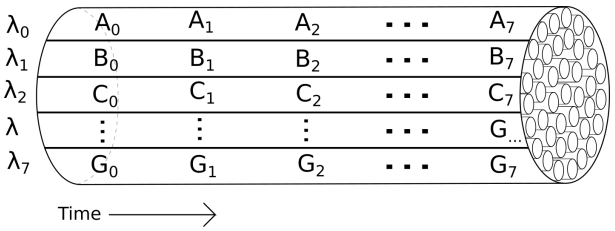
E. Unidades de Transmisión: Phits frente a Bits

Virtual Cut-Through y Wormhole son las técnicas de conmutación más utilizadas en las redes eléctricas. Estas técnicas dividen un paquete en flits, que son las unidades de control de flujo. Los flits son a su vez dividido en phits (unidades físicas). Un phit es la cantidad de bits que se pueden transferir en un solo ciclo de reloj. Por el contrario, las redes ópticas sólo pueden transferir un solo bit por longitud de onda y por ciclo (nótese que los ciclos en las redes ópticas son mucho más cortos).

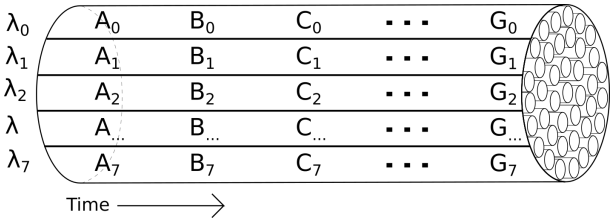
En general, los simuladores de redes eléctricas, e INSEE en particular, definen el tamaño phit como una cantidad entera de bytes (8 bits). Al adaptar este de simulador a las redes fotónicas, se ha mantenido el byte como unidad de transferencia mínima por ciclo. Sin embargo, como se mencionó anteriormente, los enlaces ópticos transfieren un bit por longitud de onda y ciclo. Por lo tanto, se requiere un nuevo enfoque para resolver este desajuste.

Se han ideado dos enfoques principales, ciclos de grano grueso y longitudes de onda empaquetadas, para modelar la transmisión de bits en lugar de phits en INSEE. El primer enfoque define un ciclo de grano grueso que se subdivide en 8 ciclos de grano fino como la unidad de ciclo, lo que permite enviar 8 bits (es decir, 1 byte, el tamaño mínimo de phit en INSEE) por ciclo usando una sola longitud de onda. El segundo enfoque agrupa 8 longitudes de onda, que actúan síncronamente, es decir, se utilizan 8 longitudes de onda para transmitir 8 bits del mismo mensaje en un solo ciclo fotónico, lo que implica que el tamaño mínimo del canal es de 8 longitudes de onda. La Figura 1 presenta un ejemplo en el que se transmiten 8 phits (etiquetados de A a G) usando ambos enfoques. En el enfoque de ciclos de grano grueso, cada phit se transmite en una longitud de onda diferente mientras que en el enfoque de longitudes de onda empaquetadas varias longitudes de onda cooperan para transmitir el mismo phit en paralelo.

[1]El término canal se utiliza en la literatura de tecnología óptica para referirse a una sola longitud de onda. Este artículo utiliza este término desde una perspectiva informática para referirse al conjunto de longitudes de onda utilizadas para transmitir el mismo mensaje.



(a) Enfoque de ciclo de grano grueso.



(b) Enfoque de longitudes de onda empaquetadas.

Fig. 1: Ejemplo de uso de ocho longitudes de onda para transmitir ocho phits, referidos de A a G, con los enfoques de transmisión propuestos.

Como se muestra en la Tabla I, elegir entre ambos enfoques presenta un compromiso entre las características de la red. Para un enlace óptico compuesto de 40 longitudes de onda, el enfoque de ciclo de grano grueso puede proporcionar hasta 40 canales paralelos, pero cada uno de estos canales sólo puede ofrecer el ancho de banda de una sola longitud de onda (40 Gbps). Por el contrario, el segundo enfoque ofrece un número máximo limitado (5) de canales, pero cada uno agrega el ancho de banda de 8 longitudes de onda ($40 \times 8 = 320$ Gbps).

Obsérvese que entre ambos enfoques hay varias posibilidades híbridas. Por ejemplo, la cantidad de canales puede ser reducida a la mitad con respecto a la aproximación de ciclos de grano grueso (segunda línea de la tabla). Entonces, en lugar de transmitir 1 byte en 1 ciclo usando 1 longitud de onda, el byte se puede dividir en 2 fragmentos que son transmitidos por 2 longitudes de onda (duplicando la frecuencia de red).

F. Topología y Encaminamiento

Las topologías de red implementadas en INSEE necesitan ser adaptadas para trabajar bajo la tecnología fotónica. En particular, el algoritmo de encaminamiento debe adaptarse a la conmutación de circuitos. Implementamos el encaminamiento mínimo para establecer la ruta entre cada par origen-destino. Es decir, sólo se consideran los caminos que

	# Canales	Ancho de banda por canal
Ciclos grano grueso	40	40 Gbps
Híbrida	20	80 Gbps
Híbrida	10	160 Gbps
Longitud de onda empaquetada	5	320 Gbps

TABLA I: Posibles compromisos de los enfoques de transmisión estudiados para un enlace óptico con 40 longitudes de onda.

hacen un número mínimo de saltos. Para explorar los beneficios de la fotónica, se ha modelado una topología de toro 3D. Utilizamos el encaminamiento mínimo determinista para el toro, en particular XYZ. Se dejan otras topologías para trabajo futuro.

V. Evaluación de Rendimiento

El objetivo principal de esta sección es ilustrar cómo una red fotónica -consistente de los mecanismos y métodos discutidos en la Sección IV- se comporta en términos de ancho de banda y latencia de red con respecto a una red eléctrica. En otras palabras, comparamos los resultados de las extensiones propuestas con los resultados proporcionados originalmente por INSEE. A continuación, se discuten las opciones de diseño que se han seleccionado para llevar a cabo los experimentos presentados en este documento.

A. Detalles del Sistema

Esta sección especifica el ancho de banda, el número de longitudes de onda por enlace y la topología de red que se han considerado para obtener los resultados.

Para la red eléctrica, se ha modelado una red 10 Gigabit Ethernet. En el caso de la red fotónica se asume que una longitud de onda proporciona 40 Gbps. El ancho de banda de la red eléctrica se ha elegido porque un conjunto importante (un 35.6%) de los supercomputadores ubicados en el Top500 [1] implementan esta tecnología de red, mientras que el de la red óptica se ha seleccionado de acuerdo con la actual tecnología de láseres comentada en la Sección II.

Por otro lado, las redes fotónicas están limitadas por la banda de comunicación óptica [17] y, como se explica en la Sección IV-C, la cantidad de longitudes de onda depende de la distancia entre ellas. En la actualidad, se suele emplear una separación de canales de 100 GHz, lo cual proporciona 40 longitudes de onda por enlace de fibra óptica [18], pero esta distancia podría reducirse para permitir más longitudes de onda por enlace. Por ejemplo, una separación de 50 GHz permite multiplexar 80 longitudes de onda, y recientemente [19], se han alcanzado 160 longitudes de onda con una separación de 25 GHz.

La Tabla II resume las principales opciones de diseño de las configuraciones estudiadas para un enlace fotónico con 40 longitudes de onda. Además de los dos enfoques principales de transmisión, denominados *empaquetamiento de longitud de onda* y *ciclo de grano grueso*, se han estudiado esquemas híbridos que combinan ambos enfoques. Todas las

Técnica	# Canales	Tamaño de Phit (bytes)	Ancho de Banda por Canal (Gbps)
Eléctrica	-	4	10
Empaquetado	5	1	320
Híbrida	10	1	160
Híbrida	10	2	160
Híbrida	10	4	160
Híbrida	20	1	80
Híbrida	20	2	80
Grano grueso	40	1	40

TABLA II: Configuraciones de red considerando un enlace con 40 longitudes de onda.

configuraciones presentan un ancho de banda agregado de 1,6 Tbps con 40 longitudes de onda por enlace. Este ancho de banda agregado está repartido equitativamente entre los canales de cada configuración. En este trabajo, también estudiamos futuros enlaces ópticos con 80 longitudes de onda, lo que da 3.2 Tbps de ancho de banda agregado por cada enlace. Para llevar a cabo este último estudio, hemos definido nuevas configuraciones duplicando la cantidad de canales de las que se muestran en la Tabla II.

Para lanzar los experimentos se han usado dos trazas recogidas de la ejecución de dos aplicaciones ExaNest; Gadget y Lammps. Estas trazas tienen un número diferente de nodos; por lo tanto, para Gadget, que tiene 72 nodos, hemos utilizado un toro 3D con 4x6x3 nodos y para Lammps, que tiene 192 nodos, el toro 3D está configurado con 4x8x6 nodos.

B. Resultados

La Figura 2 y la Figura 3 muestran el tiempo de ejecución obtenido con la red eléctrica y las distintas configuraciones de red fotónicas con ambas trazas. Cada configuración tiene cuatro argumentos asociados, referentes a (de izquierda a derecha) la cantidad de longitudes de onda por enlace, el número de canales posibles, el tamaño del phit y el ancho de banda por canal.

En primer lugar, se puede observar que las configuraciones con 40 longitudes de onda por enlace proporcionan los mismos resultados que aquellas con 80 longitudes de onda independientemente de la aplicación. Esto significa que un ancho de banda por enlace de 1,6 Tbps (alcanzable con la tecnología fotónica actual) es suficiente para lograr el mejor rendimiento. Así, a partir de ahora, nos centraremos en las configuraciones con 40 longitudes de onda por enlace. Entre ellas, la configuración que proporciona los mejores resultados en ambas cargas de trabajo estudiadas es la de empaquetamiento de longitud de onda (ver Tabla II). Esto se debe a que cuenta con el mayor ancho de banda por canal (320 Gbps), lo que significa que este parámetro es el que más impacto tiene en el rendimiento.

Por otro lado, para un ancho de banda dado por canal (p.e. 160 Gbps), la mejor estrategia es reducir el tamaño del phit. Por ejemplo, en ambas apli-

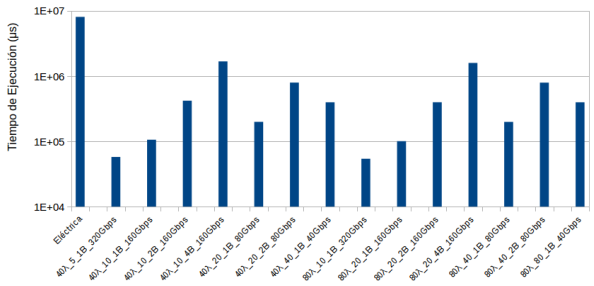


Fig. 2: Tiempo de ejecución (en us) para *Gadget*.

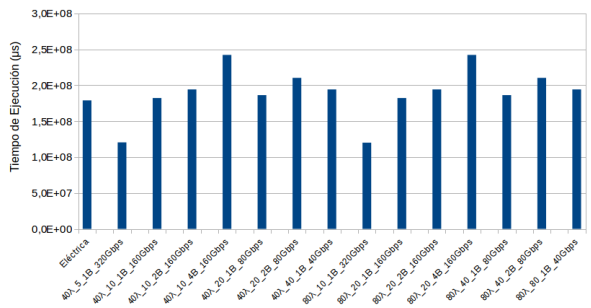


Fig. 3: Tiempo de ejecución (en us) para *Lammps*.

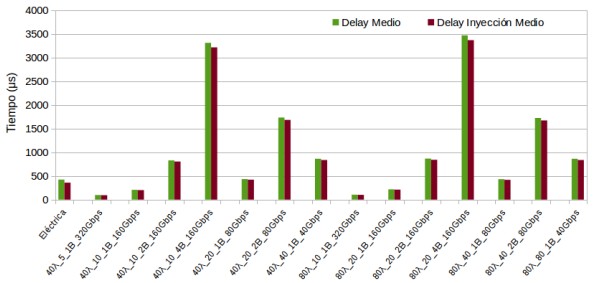


Fig. 4: Latencia media de red (en us) para *Gadget*.

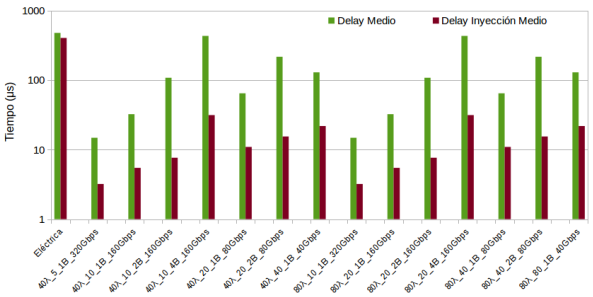


Fig. 5: Latencia media de red (en us) para *Lammps*.

caciones, la mejor configuración para 80 Gbps por canal es la que tiene un tamaño phit de 1 byte (es decir, *40λ_20_1B_80_Gbps*). Esto se debe a que los tamaños de phit pequeños utilizan el ancho de banda que estaría infrautilizado en tamaños de phit mayores. Este efecto es tan importante que a veces es mejor reducir el tamaño phit incluso si el ancho de banda por canal se reduce (por ejemplo, comparar *40λ_10_2B_160Gbps* con *40λ_20_1B_80Gbps* en Gadget).

Se observa que mientras que en Gadget la red fotónica mejora ampliamente el rendimiento de la red eléctrica (variando los beneficios entre 1 y 2 órdenes de magnitud), en Lammps el tiempo de ejecución es similar o mejor en la red eléctrica salvo en las configuraciones óptimas con 320 Gbps por canal. La razón detrás de estos resultados es que en Gadget dominan las transacciones de red con tiempo de cómputo escaso, mientras que en Lammps hay una gran cantidad de tiempo de cómputo entre accesos consecutivos a la red. Por lo tanto, los beneficios de la fotónica en Lammps sólo son significativos con los anchos de banda mas altos.

La Figura 4 presenta la latencia media de red, que incluye la latencia de inyección y la latencia del tránsito de los paquetes, para Gadget. Como puede observarse, casi toda la latencia se debe a la inyección. Aunque en esta figura la red eléctrica tiene una latencia baja, debe tenerse en cuenta que en la red eléctrica la latencia se calcula por cada paquete de 64 bytes mientras que en las configuraciones fotónicas se calcula para la transmisión de un mensaje entero y por lo tanto, no son directamente comparables. Sin embargo, los resultados confirman una fuerte correlación entre la latencia de red y el tiempo de ejecución de las configuraciones ópticas.

En Lammps, como se muestra en la Figura 5, la distribución de las latencias es completamente diferente entre las redes eléctricas y fotónicas. Mientras que en las primeras la mayor parte de la latencia es causada por la inyección, en el segundo la causas se distribuyen uniformemente entre las latencias de inyección y de tránsito. Esto se debe a que el tamaño de los mensajes en Lammps es más grande en promedio. Por lo tanto, un ancho de banda fotónico bajo por canal tiene un alto impacto en la latencia total, afectando en consecuencia el tiempo de ejecución.

VI. Trabajo Relacionado

El interés de la comunidad académica y de la industria en el desarrollo de los sistemas Exascale y en la mejora de las arquitecturas dentro del chip ha fomentado la investigación sobre la tecnología fotónica. Para estudiar estos sistemas, se requieren nuevas herramientas de simulación y estimación.

Los simuladores actuales de red [20], [21], [22] se centran en las redes eléctricas de conmutación de paquetes. Estas herramientas se pueden adaptar fácilmente a modelos de conmutación de paquetes para las redes híbridas electro-ópticas. Sin embargo, con el fin de adaptarlos para modelar las capacidades de conmutación de circuitos de las redes ópticas, se requiere un alto esfuerzo de programación. Debido a esto, algunas herramientas propuestas recientemente han sido diseñadas desde cero para soportar redes ópticas. En este sentido, un simulador bien conocido es PhoenixSim [23], [24]. Este entorno de simulación modela sistemas multiprocesadores que utilizan redes eléctricas, redes ópticas

y redes híbridas. PhoenixSim se basa en el entorno de simulación OMNeT++ [25] y permite el análisis de redes de interconexión desde el nivel físico (e.j., pérdida de inserción óptica, diafonía, disipación de energía) hasta el nivel de sistema (e.j., latencia, rendimiento y tiempo de ejecución).

La herramienta *Design Space Exploration of Networks Tool* (DSENT) [26] mejora el modelo de PhoenixSim de la interfaz de circuitos electro-ópticos tales como moduladores y receptores, modelando intercambios entre dispositivos fotónicos y especificaciones de modulador/receptor que pueden ser explotados para alcanzar configuraciones óptimas en términos de área y potencia. DSENT está diseñado para permitir la evaluación rápida de área y potencia de múltiples configuraciones de redes ópticas y, cuando está acoplado con un simulador que modela una arquitectura, se pueden obtener estimaciones de potencia y área para la red simulada. Sin embargo, DSENT no modela conmutadores fotónicos por lo que no se puede utilizar para simular redes de conmutación de circuitos como se ha realizado en este trabajo. Además, DSENT no admite patrones de tráfico y trazas de aplicaciones, por lo que no puede proporcionar los detalles de una simulación a nivel de sistema.

LioeSim [27] es un simulador de red eléctrica y óptica que utiliza Orion [28] para los modelos de router eléctricos y enlaces. A diferencia de DSENT, modela conmutadores fotónicos y permite analizar resultados a nivel físico y de sistema. Desafortunadamente, LioeSim se centra en redes dentro del chip. Por el contrario, en este trabajo se realizan simulaciones de comunicaciones fuera del chip, para redes ópticas intra-rack e inter-rack.

Por último, también existe la necesidad de ayudar a los diseñadores en tareas como colocar visualmente los dispositivos fotónicos, conectar guías de ondas, etc. Para este fin, en [29], Hendry et al. presentan la herramienta *Visual Automated Nanophotonic Design And Layout* (VANDAL), que también se puede interconectar con otras herramientas estándares de la industria basadas en los procesos de la fabricación de los chips.

VII. Conclusiones

La comunicación eficiente de datos es uno de los desafíos más críticos para alcanzar la computación Exascale en futuros supercomputadores. Las redes Exascale contarán con miles de nodos, y la fotónica ha surgido como una tecnología prometedora para afrontar este desafío. Para orientar a los diseñadores en la toma de decisiones se requieren herramientas de simulación de redes fotónicas.

En este trabajo se ha discutido el proceso requerido para desarrollar un simulador de red fotónica a partir del simulador de red eléctrica INSEE. Hemos identificado los componentes clave que están involucrados, se han descrito las principales diferencias entre estos componentes dependiendo de la tecnología subyacente (eléctrica y óptica), y se han estudiado las características ópticas de algunos componentes (p.e. WDM). En este artículo se han discutido las extensiones principales del simulador en la arquitectura del router y el enlace, la implementación de WDM sobre conmutación del circuitos y los métodos específicos de encaminamiento.

Con fines ilustrativos, se han realizado algunos experimentos para comparar las redes fotónicas con las redes eléctricas en una topología de toro 3D. Se han evaluado varias configuraciones variando cuatro parámetros principales: la cantidad de longitudes de onda, el número de canales posibles, el tamaño de phit y el ancho de banda por canal. Los resultados experimentales obtenidos con trazas de aplicaciones reales utilizadas en el proyecto ExaNeSt muestran que, dependiendo de la configuración de la red óptica, el tiempo de ejecución de la aplicación puede variar ampliamente. En general, un ancho de banda futurista por enlace de 3.2 Tbps no proporciona beneficios de rendimiento adicionales con respecto a 1.6 Tbps. Además, encontramos que el parámetro que más repercute en el rendimiento es el ancho de banda por canal, logrando los mejores resultados con canales de 320 Gbps. Por último, para canales con anchos de banda inferiores (p.e. 160 y 80 Gbps), reducir el tamaño de phit proporciona los mejores resultados.

Agradecimientos

Este trabajo ha sido apoyado por el proyecto ExaNest, financiado por el programa de investigación e innovación Horizon 2020 de la Unión Europea bajo el acuerdo de subvención nº 671553, y los fondos del Ministerio de Economía y Competitividad (MINECO) y el Plan E bajo la subvención TIN2015-66972-C5-1-R.

Referencias

[1] "Top500 website," 2015, Apr.

[2] Avinash Karanth Kodi, Brian Neel, and William C Brantley, "Photonic interconnects for exascale and datacenter architectures," *IEEE Micro*, vol. 34, no. 5, pp. 18–30, 2014.

[3] Sébastien Rumley, Dessislava Nikolova, Robert Hendry, Qi Li, David Calhoun, and Keren Bergman, "Silicon photonics for exascale systems," *Journal of Lightwave Technology*, vol. 33, no. 3, pp. 547–562, 2015.

[4] Assaf Shacham, Keren Bergman, and Luca P Carloni, "Photonic networks-on-chip for future generations of chip multiprocessors," *IEEE Transactions on Computers*, vol. 57, no. 9, pp. 1246–1260, 2008.

[5] Christopher Batten, Ajay Joshi, Jason Orcutt, Anatol Khilo, Benjamin Moss, Charles W Holzwarth, Miloš A Popovic, Hanqing Li, Henry I Smith, Judy L Hoyt, et al., "Building many-core processor-to-dram networks with monolithic cmos silicon photonics," *IEEE Micro*, vol. 29, no. 4, 2009.

[6] Sebastian Werner, Javier Navaridas, and Mikel Lujan, "Designing low-power, low-latency networks-on-chip by optimally combining electrical and optical links," in *The 23rd IEEE Symposium on High Performance Computer Architecture*. IEEE, 2016.

[7] José Puche, Sergio Lechago, Salvador Petit, María E Gómez, and Julio Sahuquillo, "Accurately modeling a photonic noc in a detailed cmp simulation framework," in *High Performance Computing & Simulation (HPCS), 2016 International Conference on*. IEEE, 2016, pp. 387–394.

[8] "ExaNeSt website," 2017, Apr.

[9] M Katevenis, N Chrysos, M Marazakis, I Mavroidis, F Chaix, N Kallimanis, J Navaridas, J Goodacre, P Vicini, A Biagioni, et al., "The exanest project: Interconnects, storage, and packaging for exascale systems," in *Digital System Design (DSD), 2016 Euromicro Conference on*. IEEE, 2016, pp. 60–67.

[10] Fco. Javier Ridruejo Perez and José Miguel-Alonso, "Insee: An interconnection network simulation and evaluation environment," in *Proceedings of the 11th International Euro-Par Conference on Parallel Processing*, Berlin, Heidelberg, 2005, Euro-Par'05, pp. 1014–1023, Springer-Verlag.

[11] G. .. H. Duan, C. Jany, A. Le Liepvre, J. G. Provost, D. Make, F. Lelarge, M. Lamponi, F. Poingt, J. M. Fedeli, S. Messaoudene, D. Bordel, S. Brision, S. Keyvaninia, G. Roelkens, D. Van Thourhout, D. J. Thomson, F. Y. Gardes, and G. T. Reed, "10 gb/s integrated tunable hybrid iii-v/si laser and silicon mach-zehnder modulator," in *2012 38th European Conference and Exhibition on Optical Communications*, Sept 2012, pp. 1–3.

[12] G. H. Duan, C. Jany, A. Le Liepvre, M. Lamponi, A. Accard, F. Poingt, D. Make, F. Lelarge, S. Messaoudene, D. Bordel, J. M. Fedeli, S. Keyvaninia, G. Roelkens, D. Van Thourhout, D. J. Thomson, F. Y. Gardes, and G. T. Reed, "Integrated hybrid iii-v/si laser and transmitter," in *2012 International Conference on Indium Phosphide and Related Materials*, Aug 2012, pp. 16–19.

[13] Keren Bergman, Luca P Carloni, Aleksandr Biberman, Johnnie Chan, and Gilbert Hendry, *Photonic network-on-chip design*, Springer.

[14] Po Dong, Long Chen, Chongjin Xie, Lawrence L. Buhl, and Young-Kai Chen, "50-gb/s silicon quadrature phase-shift keying modulator," *Opt. Express*, vol. 20, no. 19, pp. 21181–21186, Sep 2012.

[15] Po Dong, Xiang Liu, Chandrasekhar Sethumadhavan, Lawrence L. Buhl, Ricardo Aroca, Yves Baeyens, and Young-Kai Chen, "224-gb/s pdm-16-qam modulator and receiver based on silicon photonic integrated circuits," in *Optical Fiber Communication Conference/National Fiber Optic Engineers Conference 2013*. 2013, p. PDP5C.6, Optical Society of America.

[16] Javier Navaridas, Jose Miguel-Alonso, Jose A. Pascual, and Francisco J. Ridruejo, "Simulating and evaluating interconnection networks with {INSEE}," *Simulation Modelling Practice and Theory*, vol. 19, no. 1, pp. 494 – 515, 2011, Modeling and Performance Analysis of Networking and Collaborative Systems.

[17] Vivek Alwayn, *Optical network design and implementation*, Cisco Press, 2004.

[18] René-Jean Essiambre and Robert W Tkach, "Capacity trends and limits of optical communication networks," *Proceedings of the IEEE*, vol. 100, no. 5, pp. 1035–1055, 2012.

[19] E Temprana, E Myslivets, BP-P Kuo, L Liu, V Ataie, N Alic, and S Radic, "Overcoming kerr-induced capacity limit in optical fiber transmission," *Science*, vol. 348, no. 6242, pp. 1445–1448, 2015.

[20] Yaniv Ben-Itzhak, Eitan Zahavi, Israel Cidon, and Avinoam Kolodny, "Hnocs: Modular open-source simulator for heterogeneous nocs," in *Embedded Computer Systems (SAMOS), 2012 International Conference on*. IEEE, 2012, pp. 51–57.

[21] Hemayet Hossain, Mostak Ahmed, Abdullah Al-Nayeem, Tanzima Zerin Islam, and Md Mostofa Akbar, "Gpnocsim-a general purpose simulator for network-on-chip," in *Information and Communication Technology, 2007. ICICT'07. International Conference on*. IEEE, 2007, pp. 254–257.

[22] Lavina Jain, BM Al-Hashimi, MS Gaur, V Laxmi, and A Narayanan, "Nirgam: a simulator for noc interconnect routing and application modeling," in *Design, Automation and Test in Europe Conference*, 2007, pp. 16–20.

[23] Johnnie Chan, Gilbert Hendry, Aleksandr Biberman, Keren Bergman, and Luca P. Carloni, "Phoenixsim: A simulator for physical-layer analysis of chip-scale photonic interconnection networks," in *Proceedings of the Conference on Design, Automation and Test in Europe*, 3001 Leuven, Belgium, Belgium, 2010, DATE '10, pp. 691–696, European Design and Automation Association.

[24] Sébastien Rumley, Meisam Bahadori, Ke Wen, Dessislava Nikolova, and Keren Bergman, "Phoenixsim: Cross-layer design and modeling of silicon photonic interconnects," in *Proceedings of the 1st International Workshop on Advanced Interconnect Solutions and Technologies for Emerging Computing Systems*. ACM, 2016, p. 7.

[25] András Varga and Rudolf Hornig, "An overview of the omnet++ simulation environment," in *Proceedings of the 1st International Conference on Simulation Tools and Techniques for Communications, Networks and Systems & Workshops*, ICST, Brussels, Belgium, Belgium, 2008, Simutools '08, pp. 60:1–60:10, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[26] Chen Sun, Chia-Hsin Owen Chen, George Kurian, Lan Wei, Jason Miller, Anant Agarwal, Li-Shiuan Peh, and Vladimir Stojanovic, "Dsent-a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling," in *Networks on Chip (NoCS), 2012 Sixth IEEE/ACM International Symposium on*. IEEE, 2012, pp. 201–210.

[27] X. Ma, J. Yu, X. Hua, C. Wei, Y. Huang, L. Yang, D. Li, Q. Hao, P. Liu, X. Jiang, and J. Yang, "Lioesim: A network simulator for hybrid opto-electronic networks-on-chip analysis," *Journal of Lightwave Technology*, vol. 32, no. 22, pp. 4301–4310, Nov 2014.

[28] Andrew B Kahng, Bin Li, Li-Shiuan Peh, and Kambiz Samadi, "Orion 2.0: a fast and accurate noc power and area model for early-stage design space exploration," in *Proceedings of the conference on Design, Automation and Test in Europe*. European Design and Automation Association, 2009, pp. 423–428.

[29] J. Chan, G. Hendry, K. Bergman, and L. P. Carloni, "Physical-layer modeling and system-level design of chip-scale photonic interconnection networks," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 30, no. 10, pp. 1507–1520, Oct 2011.