Proyecto RED-SEA: Resultados Intermedios

José Duro[1], Adrián Castelló[1], María E. Gómez[1], Julio Sahuquillo[1], Enrique Quintana[1], Gabriel Gómez[2], Miguel Sánchez[2], Jesús Escudero-Sahuquillo[2], Pedro J. García[2], Francisco J. Alfaro[2], José L. Sánchez[2], Francisco J. Quiles[2]

Resumen— **El objetivo general de RED-SEA es diseñar una nueva generación de red de interconexión europea, que posibilite la computación exascala en Europa, mediante una interconexión económicamente viable y tecnológicamente eficiente, aprovechando tecnología de interconexión europea (BXI) junto a la tecnología estándar y madura (Ethernet), iniciativas anteriores financiadas por la UE, como ExaNeSt, EuroEXA, ECOSCALE, Mont-Blanc, los proyectos DEEP y el proyecto de procesador europeo (EPI), así como estándares abiertos y API compatibles.**

Para alcanzar este objetivo global, el proyecto RED-SEA se desarrolla en torno a cuatro pilares fundamentales: i) arquitectura y codiseño - con el objetivo de optimizar el ajuste con los otros proyectos EuroHPC y con los procesadores EPI; ii) desarrollo de un bridge de altas prestaciones, baja latencia y sin fisuras con Ethernet iii) gestión de recursos de red, incluyendo congestión y calidad de servicio; y iv) funciones de extremo a extremo implementadas en la red.

Este artículo presenta los principales logros alcanzados a mitad del proyecto por los 2 socios españoles que participan en el proyecto, es decir, la Universitat Politècnica de València (UPV) y la Universidad de Castilla La-Mancha (UCLM), contribuyentes a los pilares 1 y 3. En este sentido, cabe destacar i) la definición de los requisitos de la red y la arquitectura de la red, una lista inicial de aplicaciones y el modelado de la arquitectura BXI3 para poder evaluar las prestaciones de las propuestas del proyecto; ii) la caracterización de la congestión de las aplicaciones y las propuestas para reducir esta congestión mediante la optimización de las primitivas de comunicación colectiva.

Palabras clave— **Redes de interconexión, HPC, congestión, datacenter, primitivas de comunicación colectiva, baja latencia.**

I. Introducción

Las redes de interconexión de próxima generación deben escalar para soportar sistemas de procesamiento masivamente paralelo (cientos de miles de nodos y millones de núcleos) y proporcionar un conjunto de características que permitan a las aplicaciones HPC, HPDA e IA alcanzar la computación exascala, beneficiándose al mismo tiempo de las nuevas tendencias de hardware y software. En este sentido, el sistema debe permitir que las aplicaciones escalen eficientemente, estar preparado para aceleradores y unidades de computación de bajo consumo, y soportar aplicaciones emergentes y generalizadas centradas en datos y relacionadas con la IA.

El consorcio RED-SEA reúne a los mejores centros académicos con las principales fuerzas industriales europeas en este ámbito. El consorcio RED-SEA persigue el objetivo mencionado, aprovechando las competencias y los antecedentes europeos clave, incluida BXI, que es una tecnología de red de interconexión europea en producción, así como los resultados de una serie de proyectos financiados por la UE sobre interconexiones y sistemas HPC.

RED-SEA está apoyando e impulsando la computación exascala y las tecnologías basadas en datos dentro de Europa mediante la ampliación y optimización de la tecnología de red BXI Exascale para anticiparse a los requisitos de los sistemas en el horizonte temporal 2022-2025. BXI versión 2 está actualmente en producción y aparece en los sistemas Top 500. El proyecto RED-SEA está poniendo en marcha la tercera generación de la interconexión BXI (conocidad como BXI3), contribuyendo a su hoja de ruta mediante: (i) la definición del proyecto de arquitectura y los modelos de simulación correspondientes; (ii) el diseño de los nuevos bloques de construcción (IP) necesarios para abordar los nuevos retos de los superordenadores modulares; (iii) la realización de una prueba de concepto inicial de sus componentes críticos en aplicaciones de la vida real; y (iv) el desarrollo del ecosistema y la creación de una comunidad más amplia de usuarios y desarrolladores que combine equipos de investigación e industriales.

El resto del artículo se organiza de la siguiente manera. La sección II resume el trabajo previo relacionado con el proyecto. En la sección III se presentan los socios que participan en RED-SEA. En la sección IV se discuten los principales retos que debemos afrontar para llevar a cabo el proyecto. En la sección V se describe el enfoque de RED-SEA para abordar estos retos. En la sección VI se describen las metodologías empleadas. En la Sección VII se presentan los paquetes de trabajo implementados en el proyecto para alcanzar su objetivo final, y en la Sección VIII los logros alcanzados hasta el momento en el proyecto por parte de los dos socios españoles. Finalmente, en el apartado IX se exponen las principales conclusiones del trabajo.

II. Estado del arte

En los últimos años se han hecho muchos intentos de implementar redes de alta velocidad, orientadas a la computación HPC, en sistemas basados en FPGA. En el marco del proyecto ExaNeSt [1] se diseñó, verificó y desplegó una interconexión de altas prestaciones y baja latencia basada en FPGA en un banco de pruebas de 128 SoCs Zynq Ultrascale+. ExaSNet es una interconexión de red jerárquica caracterizada por una topología todo a todo a nivel de nodo (formada por 4 FPGA interconectadas) y una red Torus 3D escalable para la interconexión entre nodos a nivel de bastidor. ExaNeSt integra una interfaz de red

[1]Universitat Politècnica de València, e-mail: e-mail: megomez@disca.upv.es

[2]Universidad de Castilla La-Mancha, e-mail: Jesus.Escudero@uclm.es

optimizada y embebida orientada a ARM, que ofrece múltiples canales de hardware para transferencias RDMA y packetizer-mailbox fiables a nivel de usuario [2], [**?**], y un bloque switch/router de alto rendimiento procedente de APEnet con 6 enlaces bidireccionales 3D Torus de hasta 32 Gb/s, en una única FPGA. En el mismo proyecto se ha implementado un rumtime MPI especial, y se han portado y ejecutado aplicaciones HPC en un banco de pruebas de 8 palas con 128 FPGAs (en 128 SoCs) o 512 núcleos ARMv8. EuroEXA [3] aprovecha ExaNeSt para impulsar el concepto de ïnterconexión escalable de topología híbrida"(*TriFeCta*) a escala extrema. EuroEXA diseñó una versión mejorada de la arquitectura ExaNet que proporciona diferentes topologías y características en los distintos niveles de la jerarquía de red. EuroEXA diseñó un innovador çonmutador personalizado"basado en una única FPGA Virtex Ultrascale+ que implementa una topología todo a todo de 2 saltos a nivel de placa, una red Torus 3D a nivel de bastidor y un puente ExaNet-Ethernet 100/200G para la conectividad entre bastidores.

III. Consorcio

El consorcio reúne a grupos de investigación bien establecidos en toda Europa, con una larga experiencia en redes de interconexión, incluido el diseño, despliegue y evaluación de redes. La UPV y la UCLM (España) han desarrollado técnicas punteras en casi todos los aspectos de las interconexiones, incluyendo topologías eficientes, estrategias de encaminamiento adaptativo sin bloqueo, control de flujo, provisión de QoS, técnicas de gestión de la congestión y estrategias de encaminamiento tolerantes a fallos. Además de estos socios españoles firmantes de este trabajo, participan otros socios que describimos a continuación. FORTH (Grecia) ha sido pionera en la gestión de la congestión y la regulación por flujo. ETH Zúrich (Suiza) aporta investigación puntera sobre redes escalables de altas prestaciones, programación paralela (destacando en MPI) y computación programable en red. INFN (Italia), el desarrollador de APEnet [4], tiene una larga experiencia en la creación de prototipos de sistemas y en códigos HPC eficientes. EXTOLL (Alemania), una empresa con sede en la UE, spin-off de la Universidad de Heidelberg, que ofrece IPs de baja latencia, de última generación y altas prestaciones para interconexiones HPC; ParTec (Alemania) desarrolla runtimes MPI para interconexiones, ExactLab optimiza códigos HPC y ExaPSYS es una joven startup con buenos conocimientos sobre simulaciones y sistemas heterogéneos. CEA (Francia), con actividades que abarcan desde el funcionamiento de grandes despliegues de HPC hasta nuevas tecnologías de hardware y software para sistemas e interconexiones. FZJ (Alemania) y CEA albergan dos de los mayores centros de supercomputación de Europa, y llevan a cabo investigación científica sobre códigos de computación de alto rendimiento. Atos (Francia) y CEA colaboran estrechamente en el programa nacional francés de exaescala. Por último, pero no por ello menos importante, el coordinador del proyecto, Atos/Bull, único fabricante europeo de ordenadores, cuenta con equipos de investigación y desarrollo experimentados que han producido una serie de interconexiones y sistemas comerciales on-chip y off-chip.

IV. Retos

A continuación se enumeran los principales retos para lograr las prestaciones y la compatibilidad deseados en RED-SEA:

- *Escalabilidad, fiabilidad*: Demostrar formas de escalar las interconexiones más allá de 100 K nodos, cumpliendo al mismo tiempo los objetivos clave de prestaciones y fiabilidad, y satisfaciendo diversos requisitos, desde bibliotecas de comunicación (MPI) e IA hasta aplicaciones centradas en datos.
- *Convergencia HPC/datacenter*: desarrollar y demostrar métodos a nivel de producto que integren de forma óptima IP y el tráfico Ethernet y RoCE (RDMA sobre Ethernet convergente) en una interconexión HPC, consiguiendo una baja latencia y altas velocidades de mensajes.
- *Throughput, ancho de banda*: Multiplicar por 4 el ancho de banda y la tasa de mensajes disponibles para cada punto final de la red duplicando la frecuencia del enlace (hasta 200 Gbps) y duplicando el número de interfaz de red para cada proceso (multi-rail).
- *Calidad del servicio*: Desarrollar nuevos algoritmos de control de la congestión y mecanismos de provisión de QoS adecuados para entornos HPC ágiles centrados en datos, evaluarlos en plataformas y modelos de simulación escalables, y esbozar su camino hacia el producto de interconexión.
- *Programabilidad, latencia*: Desarrollar formas de configurar mediante programación el motor de descarga de red, permitiendo también el cómputo en red y consiguiendo una mejor latencia/eficiencia energética.
- *Nuevos procesadores, relación con EPI*: Demostrar la interoperabilidad de la interconexión diseñada con componentes de la Iniciativa Europea de Procesadores, como procesadores y aceleradores Arm y RISC-V, y definir arquitecturas de red alternativas para los sistemas europeos Exascale.
- *Protección, compartición*: Demostrar métodos para la partición de un sistema HPC existente, clúster en múltiples nubes (privadas), manteniendo la protección, la seguridad y el aislamiento.
- *Ir al mercado / impacto*: Definir una trayectoria de salida al mercado y optimizar nuestras posibilidades de tener una parte importante de estas IP y resultados europeos utilizados en los principales sistemas europeos en el horizonte 2022-23, al tiempo que reforzamos nuestras posiciones actuales en el segmento de mercado de la interconexión.

V. La aproximación RED-SEA

Los sistemas HPC y los orientados a datos de próxima generación serán heterogéneos en cuanto a los dispositivos que utilizarán, incluidos procesadores ARM y RISC-V de bajo consumo, CPU de gama alta, unidades de aceleración vectorial y GPU adecuadas para cargas de trabajo masivas de una sola instrucción y múltiples datos (SIMD), así como diseños FPGA y ASIC adaptados para códigos personalizados de consumo extremadamente eficiente [5]. Las modernas unidades de procesamiento paralelo de datos, como las GPU y los aceleradores vectoriales, pueden procesar datos a velocidades asombrosas (decenas de TFLOPS). En este panorama, la red se está convirtiendo en el próximo gran cuello de botella. El consorcio RED-SEA está abordando estos retos mediante:

- Ethernet de altas prestaciones como red de federación con semántica de comunicación RDMA de baja latencia de última generación.
- BXI como estructura HPC, que consta de dos componentes discretos: una NIC BXI, un conmutador BXI y el gestor de estructura BXI. La tercera generación de BXI añade nuevas funciones y aumentará sus prestaciones para alcanzar los objetivos enumerados.

A. RED-SEA Capa Física

El continuo aumento del ancho de banda conduce a enlaces serie cada vez más sofisticados. El proyecto se centra en enlaces de 200 Gb/s por dirección, formados por cuatro carriles diferenciales independientes que funcionan a 56 Gb/s. El proyecto se centra en el desarrollo de IP modulares MAC y PCS que puedan reutilizarse tanto para enlaces Ethernet como para futuros enlaces BXI.

B. RED-SEA Capa de Transporte

Instalar sistemas de producción que cuenten con más de 100 K nodos ya es un reto, pero escalar las prestaciones es el verdadero requisito. Las prestaciones globales dependen directamente del comportamiento de la red. Los requisitos de fiabilidad siguen a la explosión del número nodos y el mecanismo de fiabilidad extremo a extremo debe diseñarse para soportar simultáneamente hasta 100 K nodos y mantener las prestaciones de 200 Gb/s para cada enlace. El proyecto diseña un IP de fiabilidad E2E que proporcione un mecanismo de recuperación para fallos transitorios y permanentes que garantice la integridad de los mensajes, el orden de los mensajes y la entrega de los mensajes.

C. Interfaz de host RED-SEA

El proyecto tiene como objetivo una reducción muy agresiva de la latencia entre los procesadores de host y la red. Este objetivo se lleva a cabo de dos formas principales. El cambio más disruptivo consiste en eliminar la interfaz PCIe estándar y disponer de un acceso directo a los núcleos de procesador de bajo consumo a través de una interfaz coherente para reducir la latencia y simplificar la interfaz de software.

D. Entorno de software de RED-SEA

El proyecto tiene como objetivo desarrollar la pila de software y las bibliotecas para aprovechar las capacidades de descarga de BXI, como las operaciones colectivas de altas prestaciones. La red BXI se basa en la API Portals 4. Portals 4 [6] es un estándar desarrollado en colaboración por Sandia National Labs y por la Universidad de Nuevo México. Se eligió porque es la única interfaz disponible que soporta tanto MPI como PGAS, a la vez que proporciona bloques de construcción apropiados para las comunicaciones de software del sistema (E/S paralela, lanzamiento de trabajos).

E. RED-SEA Congestión y QoS

Abordar la congestión de la red es clave para proporcionar una gestión eficiente de los recursos de red. RED-SEA propone, implementa y evalúa nuevos mecanismos de gestión de la congestión para la tecnología de red BXI3, como se muestra en la Figura 1. En este punto es donde la UPV y la UCLM están haciendo las mayores aportaciones.

En primer lugar, en el contexto de este proyecto con componentes de computación y datos extremos, las aplicaciones paralelas generán una gran congestión debido a las operaciones colectivas como presentaremos en este trabajo posteriormente. En el proyecto abordamos la optimización de primitivas de comunicación colectiva clave.

Otras técnicas previstas para hacer frente a la congestión son: encaminamiento adaptativo de grano fino y medio, gestión de la congestión inteligente y con capacidad de respuesta. Además, el proyecto aborda una QoS altamente flexible. Las decisiones clave deben tomarse lo antes posible en el hardware para aumentar la eficacia de las técnicas propuestas, diseñando e implementando mecanismos inmediatamente en la tarjeta de interfaz de red o en los conmutadores de red. Además, los conmutadores deben ofrecer una visión global de la red. La solución de gestión global de la congestión incluye innovaciones desarrolladas por los socios a varios niveles: i) la definición del protocolo y la especificación de las sondas de hardware para supervisar el estado de la red; ii) los algoritmos para tomar las mejores decisiones de encaminamiento adaptativo y regulación de la inyección; y iii) el soporte para la gestión de la congestión adaptado a las operaciones colectivas.

Las IP desarrolladas abarcan desde los módulos de hardware en los puertos BXI hasta los módulos de firmware que se ejecutan en los componentes y en el software de gestión global del red.

F. Puente RED-SEA Ethernet

Los sistemas de exaescala ya no son sistemas monolíticos y cerrados. Son híbridos, compuestos de particiones especializadas y deben estar abiertos al

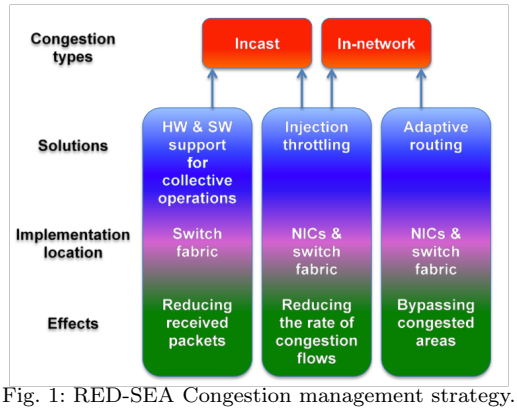


Fig. 1: RED-SEA Congestion management strategy.

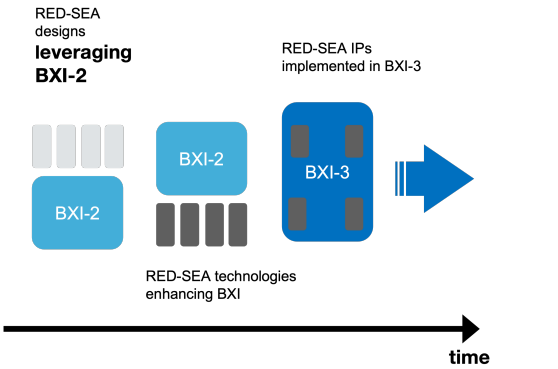


Fig. 2: In the RED-SEA project, we will leverage the existing BXI interconnect to develop, test and evaluate new technologies, planning to integrate many of them to advance BXI..

mundo exterior. Se comunican con otros superordenadores, nubes híbridas y servidores Edge para participar en un flujo de trabajo global. Esto constituye un enfoque de computación continua. Se trata de un cambio disruptivo en la naturaleza de los superordenadores. Como consecuencia, ya no es posible comprometerse en cuestiones de seguridad, y una segunda consecuencia es que las pasarelas a las redes externas están bajo presión. Uno de los demostradores del proyecto es un prototipo de pasarela Ethernet que conecta el la red HPC a la red de almacenamiento Ethernet. La pasarela se implementa como una FPGA de gama alta ajustada a los puertos de los conmutadores BXI que interactúan con los conmutadores Ethernet.

VI. Metodologías

Este proyecto hace uso de varias plataformas de evaluación, incluidos modelos y cálculos analíticos, modelos de simulación por ordenador existentes o nuevos, así como plataformas de emulación de hardware que combinan FPGA, servidores y equipos comerciales (incluido BXI).

En las siguientes subsecciones resumimos las metodologías del proyecto. Se puede encontrar una descripción más extensa en [7].

A. Conexión entre la investigación y BXI

Como se muestra en la Figura 2, el proyecto RED-SEA está creando un bucle de retroalimentación positiva entre la investigación y la industria. El proyecto aprovecha la tecnología BXI2, que se utiliza en sistemas HPC operativos y ofrece altas velocidades de enlace (100-200 Gb/s), conmutadores de alta radix (48 puertos), control de flujo por VC, retransmisiones de extremo a extremo y salto a salto, y descarga de red avanzada. Muchas de las actividades de RED-SEA utilizan BXI2 y sus actuales capacidades de alto rendimiento para probar y validar nuevas IP. Esto proporciona a los equipos de investigación de RED-SEA una ventaja competitiva para avanzar en el estado del arte de las interconexiones industriales, que pueden servir como punto de partida para el desarrollo de ASICs BXI3.

B. Codiseño

En RED-SEA, el codiseño se entiende como un proceso colectivo e iterativo en el que (i) las aplicaciones, cargas de trabajo y ecosistemas relevantes (requisitos); (ii) los diseñadores de interconexión (nuevas soluciones/IPs); y (iii) la empresa Atos (consideraciones de mercado) se reúnen para dar forma, evaluar y promover la IP de interconexión desarrollada en RED-SEA. La interconexión BXI y su evolución son fundamentales para el éxito general de este esfuerzo. Los propietarios de aplicaciones, los proveedores de plataformas HPC y los diseñadores de interconexiones experimentados han trabajado en la confección de una lista ampliada de aplicaciones y cargas de trabajo propias o ampliamente conocidas. Como resultado de este proceso, el consorcio ha confeccionado una serie de cargas de trabajo previstas, asociadas a objetivos de rendimiento relacionados. Todas las soluciones propuestas por los diseñadores de redes dentro de RED-SEA se están evaluando en función de las cargas de trabajo y los objetivos de rendimiento seleccionados por el codiseño de RED-SEA. En RED-SEA, utilizaremos una combinación de plataformas para desarrollar, probar y verificar las soluciones propuestas. Las plataformas de evaluación previstas incluyen un conjunto de modelos de simulación y bancos de pruebas de hardware para ajustarse a los requisitos particulares de cada tarea y potenciar la productividad.

C. Plataformas y modelos de simulación

En RED-SEA, utilizamos modelos de simulación de red con diferentes niveles de abstracción, para evaluar las soluciones previstas. En general, se aceptan ciertas simplificaciones para mejorar el tiempo hasta la solución, en lo que respecta a las evaluaciones de simulación. Cuando se estudia la gestión de la congestión, por ejemplo, las simulaciones ayudan a evaluar diferentes soluciones en diferentes escenarios de congestión dentro de la red, y por lo tanto el modelo no tiene que capturar los detalles del procesador. Cuando se estudia la interfaz de red, pueden ser necesarios modelos detallados de la ruta que conecta el procesador con la interfaz de red para capturar con precisión la latencia y la tasa de mensajes.

Los modelos SystemC y RTL, herramientas que utilizan los equipos de diseño de hardware, también se emplean con frecuencia para examinar en detalle el comportamiento y la corrección de las IP bajo diversas cargas de trabajo. Entre las herramientas de simulación que utilizará el consorcio figuran: OMNeT++ [8], NS3 [9], SystemC, Verilog/SystemVerilog, Gem5 [10], Custom frameworks [11].

En concreto, como se describe más adelante, la UPV y la UCLM están utilizando el simulador de red SAURON [12], para modelar la arquitectura de la tecnología de red BXI, y simular cientos de miles de nodos. Así mismo, para generar tráfico realista que alimente las redes modeladas, se hace uso del framework VEF Traces [13], que también se ha utilizado para capturar los patrones de comunicación de las aplicaciones HPC que se han definido como casos de uso del proyecto: NEST, LAMMPS, LinkTest, etc. Finalment, en cuanto al modelado de estas aplicaciones, la UCLM y la UPV, han impulsado en RED-SEA la colaboración con otros proyectos financiados con fondos europeos, como el DEEP-SEA, IO-SEA y MAELSTROM, con el objetivo de capturar trazas de tráfico en la red adicionales en el formato VEF, conforme a las aplicaciones y casos de uso de estos proyectos. Se han obtenido trazas de aplicaciones como GROMACS, PATMOS o LinkPack Todas estas trazas están disponibles en un repositorio público[1].

D. Estrategia de emulación de hardware

Uno de los objetivos del proyecto RED-SEA es desarrollar nuevos IPs hardware que puedan ser utilizados para mejorar la interconexión BXI. Se utilizan simulaciones a nivel funcional para evaluar el impacto en el rendimiento de estas IPs bajo patrones de tráfico sintéticos o trazas de mini-aplicaciones. RED- SEA está probando además el rendimiento de estas IPs en despliegues reales utilizando una plataforma de emulación de hardware, que mezcla placas comerciales/verificadas (por ejemplo, conmutador BXI) programables FPGAs programables con modelos modificables de componentes BXI (por ejemplo, conmutadores BXI y NICs), servidores informáticos y placas FPGA.

Testbeds de pequeña escala. El consorcio está utilizando bancos de pruebas a pequeña escala para fines de emulación que incluyan placas FPGA acopladas a componentes existentes, como conmutadores BXI, tarjetas de interfaz de red (NIC) BXI, conmutadores y adaptadores Ethernet, con el fin de validar la funcionalidad de las IP y medir su rendimiento.

Testbeds de alta escala. El consorcio aprovecha la plataforma multi-FPGA del proyecto ExaNeSt [1] para desarrollar y probar protocolos y puntos finales heterogéneos, como procesadores RISC-V y aceleradores FPGA.

Dibona: Arm-based HPC Cluster. El cluster Dibona es una máquina diseñada en el proyecto Mont-Blanc 3. Dibona se está reutilizando para análisis y optimizaciones de prestaciones de BXI. Dibona cuenta con procesadores ARMv8 ThunderX2 conectados mediante interconexión InfiniBand o BXI en una topología fat-tree. Dibona se compone de 3 placas (nodos de cálculo), cada uno de ellas optimizado para integrar 2 zócalos (CPU) y 16 canales de memoria. En el contexto de este proyecto, las placas base se actualizarán con soporte BXI en el mezzanine de interconexión.

E. Cargas HPC y Datacentre Relevantes

La lista de aplicaciones y puntos de referencia de red específicos seleccionados en el proyecto es:

NEST. NEST es una aplicación consolidada que abarca una gama mucho más amplia de modelos neuronales y de conectividad sináptica, que admite la ejecución paralela mediante un híbrido MPI y OpenMP, y que proporciona una interfaz Python para facilitar la configuración y la interoperabilidad con códigos para la manipulación algebraica y la investigación estadística de la red simulada y su dinámica.

LAMMPS. Large-scale Atomic/Molecular Massively Parallel Simulator es un motor clásico de dinámica molecular centrado en el modelado de materiales. Se utiliza ampliamente en varias ramas de la ciencia: física del estado sólido, química computacional, biofísica y muchas otras.

SOM. Los mapas autoorganizados (SOM) son redes neuronales artificiales que se utilizan en el contexto del aprendizaje automático no supervisado. En el contexto de RED SEA, se ha desarrollado una implementación paralela del algoritmo SOM. El paquete, llamado DIAPASOM, explota los enfoques de paralelización MPI y PGAS (a través de OPENSHMEM). El código se ha publicado bajo licencia BSD-4-Clause y está a disposición del público en https://github.com/exactlab/diapasom.

DAW. DAW (Datacentre-inspired adversarial workloads) es un conjunto de generadores de escenarios de banco de pruebas para reproducir cargas de trabajo interesantes de la plataforma a gran escala ExaNeSt que ponen a prueba las capacidades de interfaz de red a escala y las capacidades de calidad de servicio de la interconexión.

LinkTest. LinkTest es una herramienta para la evaluación comparativa escalable de API de comunicación. Las API y el hardware de comunicación asociado se evalúan mediante el envío de mensajes entre tareas alojadas en la misma o en diferentes CPU/GPU. Los mensajes pueden enviarse entre dos tareas en paralelo: una tarea envía su mensaje a la otra mientras ésta le devuelve el suyo. Alternativamente, los mensajes pueden enviarse uno tras otro. Además, se puede controlar la ubicación en la que se almacenan estos mensajes. Pueden residir en la RAM de la

[1] https://gitraap.i3a.info/jesus.escudero/vef-traces-repository

CPU o en la RAM de la GPU.

PCVS. PCVS (Parallel Computing Validation Suite) es un motor de validación diseñado para evaluar las capacidades de descarga de la red de alta velocidad ejecutando grandes bases de pruebas de forma escalable, aprovechando entornos altamente paralelos para reducir su tiempo de obtención de resultados, mejorando posteriormente la eficiencia del proyecto gracias a un proceso de validación más regular.

VII. Implementación

El proyecto RED-SEA se está llevando a cabo en torno a cuatro pilares clave y en torno a estos pilares, RED-SEA define 4 paquetes de trabajo técnicos que se muestran en la Figura 3.

El paquete de trabajo WP1, "Co-design & performance", tiene como objetivo recopilar los requisitos de las aplicaciones y ecosistemas objetivo. Además, el WP1 cubre el trabajo necesario para construir o adaptar las plataformas de evaluación existentes (bancos de pruebas de emulación y simulación) y evaluar las soluciones nuevas o existentes, cubriendo aspectos como la calidad del servicio y la fiabilidad a escala, la portabilidad de los códigos existentes, frente a las aplicaciones y las cargas de trabajo establecidas anteriormente.

El paquete de trabajo WP2, "High performance Ethernet"(Ethernet de alto rendimiento), persigue desarrollar IPs que permitan puentes de altas prestaciones y baja latencia entre BXI y Ethernet, para eliminar la necesidad de sockets Ethernet en los servidores y de switches/routers Ethernet consolidando el tráfico Ethernet.

El paquete de trabajo 3, "Gestión eficiente de los recursos de red", aprovecha los recursos disponibles en la arquitectura BXI para mejorar el rendimiento de la red, centrándose principalmente en la gestión de la congestión y la calidad del servicio, pero también en el enrutamiento adaptativo y la gestión de la energía. En cuanto a la gestión de la congestión, se estudian distintos enfoques (optimización colectiva de las comunicaciones primitivas, estrangulamiento de la inyección, esquemas de colas, etc.) para tratar de forma más eficiente los distintos tipos de congestión.

Por último, el paquete de trabajo 4, "Funciones de punto final y fiabilidad", persigue desarrollar protocolos de extremo a extremo mejorados para avanzar en la fiabilidad a la escala de sistema prevista, trabajar en optimizaciones de MPI y en modelos de programación para aceleradores de red, así como en la interoperabilidad de la interconexión con tecnologías emergentes de procesadores y aceleradores de bajo consumo, como las diseñadas en EPI.

VIII. Hitos Intermedios

Durante la primera mitad del proyecto RED-SEA, se han llevado a cabo acciones fructíferas en todos los paquetes de trabajo técnicos, que son los pilares clave para guiar y alcanzar los retos globales del proyecto. A continuación, resumimos los principales logros para los paquetes de trabajo 1 y 3, en los que UPV y UCLM están más activos.

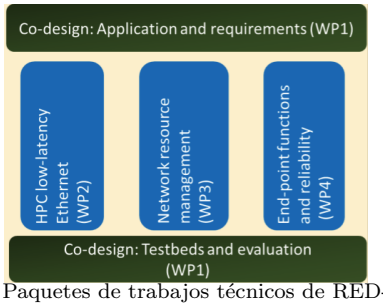


Fig. 3: Paquetes de trabajos técnicos de RED-SEA.

A. WP1: Architecture, Co-design, and Performance

El diseño de hardware de la red de interconexión y el diseño de software de middleware necesitan del conocimiento de los patrones de comunicación de la red, en un entorno de codiseño de hardware y software.

Partiendo de una descripción detallada del ámbito y campo de aplicación de cada una de las aplicaciones, hemos esbozado los requisitos de hardware y software de cada uno de ellos. VEF Traces framework [13], desarrollada por la UCLM, se ha establecido como el marco de referencia en el proyecto para recopilar trazas y utilizarlas con fines de simulación. Además, el proyecto ha proporcionado apoyo a otros proyectos SEA para instalar y ejecutar el entorno. De hecho, el proyecto DEEP-SEA ha compartido con RED-SEA algunas trazas VEF recogidas de aplicaciones objetivo de DEEP-SEA, como GROMACS, PATMOS y LinPack. A continuación, mediante el análisis de las trazas de red, hemos proporcionado recomendaciones para el diseño del sistema de red en relación con la latencia, el ancho de banda en rangos específicos de tamaño de mensaje, el número de mensajes MPI y la velocidad de transmisión.

Con el fin de proporcionar recomendaciones para la arquitectura de red, hemos realizado un análisis ampliado de las dos principales aplicaciones disponibles en RED-SEA, LAMMPS y NEST, utilizando el conjunto de herramientas VEFtraces.

Por último mencionar que la UPV y la UCLM han invertido un esfuerzo importante en el simualdor Sauron **??** con el objetivo de modelar la arquitectura BXI3, y validarla contra el diseño de bajo nivel de ATOS (en SystemC). Este modelo se está ampliando para incluir y evaluar las propuestas realizadas en el WP3.

B. WP3: Efficient Network Resources Management

En el paquete de trabajo 3, un punto clave a atacar para gestionar eficientemente los recursos de red es la congestión. Por ello, uno de los primeros pasos de este paquete de trabajo ha sido realizar la caracterización de los escenarios de congestión de red derivados del tráfico generado por las aplicaciones objetivo del proyecto RED-SEA (LAMMPS y NEST) identificadas en el paquete de trabajo 1. Realizamos esta caracterización con el objetivo de adquirir un cono-

cimiento completo del comportamiento de la congestión causada por las aplicaciones objetivo, con el fin de diseñar los mecanismos de control de la congestión de este paquete de trabajo. Para realizar esta caracterización, se han utilizado trazas obtenidas mediante el framework VEF Trace [13] al ejecutar las dos aplicaciones mencionadas en el clúster Dibona.

Hemos observado que las prestaciones de la red están muy determinadas por las primitivas colectivas (tipo y frecuencia) que realiza cada carga de trabajo. Más concretamente, las prestaciones de la red vienen determinadas por los patrones de comunicación definidos por estas primitivas colectivas de las cargas de trabajo en ejecución. Por este motivo, como primer paso, analizamos el tipo, la frecuencia y los patrones de comunicación de las primitivas de comunicación colectiva de las aplicaciones en ejecución. Los resultados experimentales han mostrado que LAMMPS es una aplicación con una variedad de primitivas de comunicación colectiva (ver Figura 4a), mientras que NEST está dominada por la primitiva All2All (ver Figura 4b) y esto afecta a la distribución de mensajes y tráfico en la red. En el caso de NEST (ver Figura 5b), la distribución global de mensajes, considerando toda la traza, es uniforme entre pares de tareas, pero en el caso de LAMMPS la distribución no es uniforme (ver Figura 5a), ya que en LAMMPS existen primitivas colectivas donde los mensajes son enviados o recibidos por la tarea número cero (una sola tarea).

Otra observación interesante es el comportamiento a ráfagas del tráfico de red generado por ambas aplicaciones, que puede impedir que los paquetes avancen, generando una alta ocupación en las colas, aumentando así la latencia de los paquetes y reduciendo la productividad de la red.

Con el objetivo de reducir la congestión provocada por las primitivas de comunicación colectiva, estamos llevando a cabo una optimización software-hardware de algunas primitivas de comunicación colectiva cruciales teniendo en cuenta la topología de la red. En comparación con nuestro objetivo, las instancias actuales de MPI para la primitiva colectiva son agnósticas a la topología de la red. Al no tener en cuenta la topología, estas implementaciones malgastan el ancho de banda de la red, creando congestión. La optimización del hardware se basa en el enrutamiento multidifusión asistido por hardware. El objetivo es reducir el número de mensajes generados, así como el volumen de tráfico.

Las optimizaciones hardware se basan en tener soporte de los switches de la red para reducir el número de mensajes enviados por la red en la implementación de la primitivas de comunicación colectivas y de esta forma reducir la congestión.

Las optimizaciones software se basan en optimizar los algoritmos que implementan las primitivas de comunicación colectiva (CCP). En esta línea, distinguimos dos técnicas de optimización de software. En primer lugar, la aplicación de pipelining de comunicación mediante la segmentación del mensaje, con el fin de lograr un solapamiento más eficiente con el cómputo. En segundo lugar, la introducción de técnicas que tienen en cuenta la topología para lograr una comunicación más eficiente.

La figura 6 muestra el efecto sobre las prestaciones de la segmentación de mensajes (solapamiento de cálculo y comunicación) para dos CCP asíncronos de OpenMPI: MPI-Ibcast (gráfico de la izquierda), MPI-Ireduce-scatter, y (gráfico de la derecha), mostrando que es posible acelerar el rendimiento aplicando pipelining.

Además se han implementado dos algoritmos conscientes de la topología dragonfly y la primitiva broadcast: LLF (Local Level First) and GLF (Global Level First). En la Figura 7 puede verse una comparativa de los algoritmos tradicionales (N y log N) con las optimizaciones hardware y los dos algoritmos software optimizados (LLF and GLF). En la figura se muestra el speedup para los algoritmos con respecto al algoritmo tradicional N. Como se puede ver el algoritmo HW es el que mejora más las prestaciones seguido del LLF y además son capaces de mantener la mejora con el aumento del tamaño del mensaje.

Además, con el objetivo de abordar la congestión, en este paquete de trabajo se está también trabajando en propuestas de control de la congestión basadas en colas y encaminamiento adaptativo. Respecto al encaminamiento adaptativo se ha acordado con Atos el diseño de un algoritmo que requiere bajo coste e impedirá la propagación de la congestión.

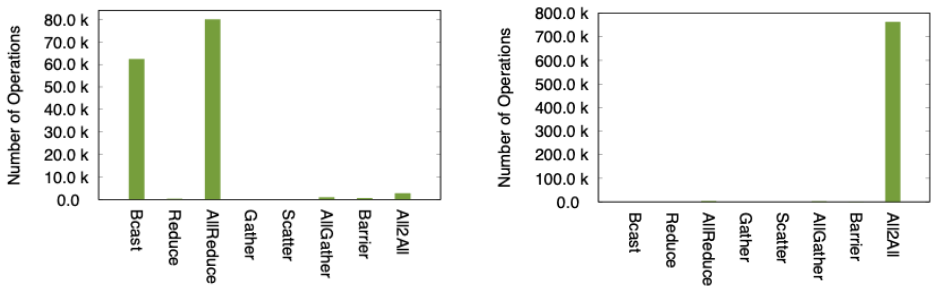
En este paquete de la trabajo la UPV y la UCLM también están colaborando junto con Atos en el diseño de técnicas de QoS y aislamiento de tráfico que sean compatibles con BXI y de técnicas para la reducción del consumo de energía en la red.

IX. Conclusiones

Este trabajo ha resumido los retos impuestos para desarrollar la próxima red europea para un sistema de escala extrema. Hemos presentado el enfoque de RED-SEA para abordar estos retos durante un plazo de 3 años. Estos ambiciosos objetivos son alcanzables, como demuestran las acciones llevadas a cabo con éxito a mitad del proyecto.

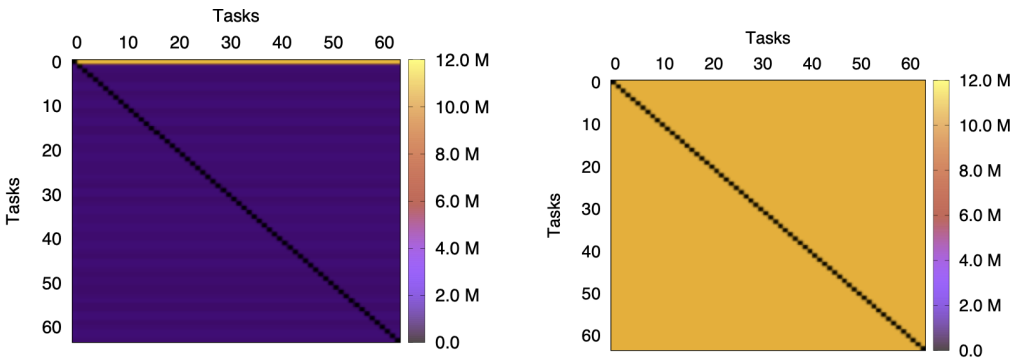
Entre los hitos más relevantes nos gustaría citar los siguientes:

- Se han definido los requisitos de la red, la arquitectura y una lista inicial de puntos de referencia y aplicaciones. En cuanto al tipo típico de comunicaciones MPI, las dos aplicaciones consideradas hacen uso de MPIAllToAll, MPIBroadcast y MPIAllReduce principalmente, por lo que el correspondiente tipo de patrón de comunicaciones requerirá especial atención durante el diseño de la red.
- Se ha caracterizado la congestión en el sistema objetivo y se ha dado cuenta del alto impacto de algunas primitivas de comunicación colectiva en las prestaciones y se han propuesto mecanismos para hacer frente a la congestión mostrando los beneficios de prestaciones que se pueden lograr.



(a) LAMMPS (b) NEST

Fig. 4: Número de llamadas a colectivas por tipo para Lammps y Nest.



(a) LAMMPS (b) NEST

Fig. 5: Bytes transferidos entre los pares de tareas.

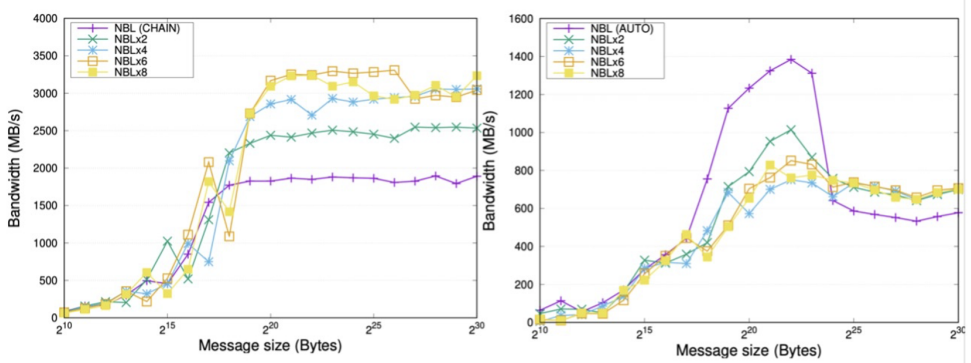


Fig. 6: Prestaciones de MPI-Ibcast y MPI-Ireduce-scatter OpenMPI con número fijo de segmentos. 8 nodos conectador por red EDR. NBL corresponde con la implementación inicial.

Fig. 7: Prestaciones de MPI-Ibcast y MPI-Ireduce-scatter OpenMPI con número fijo de segmentos. 8 nodos conectador por red EDR. NBL corresponde con la implementación inicial.

Agradecimientos

El presente trabajo ha sido financiado por la Comisión Europea (programa Horizon 2020, call H2020-JTI-EuroHPC-2019-1 - grant agreement ID: 955776) y Ministerio de Ciencia e Innovación de España (proyectos PCI2021-121934 y PCI2021-121976).

Referencias

[1] Manolis Katevenis et al., "Next generation of exascale-class systems: Exanest project and the status of its interconnect and storage development," *Microprocessors and Microsystems*, vol. 61, pp. 58–71, 2018.

[2] Manolis Ploumidis, Nikolaos D. Kallimanis, Marios Asiminakis, Nikos Chrysos, Pantelis Xirouchakis, Michalis Gianoudis, Leandros Tzanakis, Nikolaos Dimou, Antonis Psistakis, Panagiotis Peristerakis, Giorgos Kalokairinos, Vassilis Papaefstathiou, and Manolis Katevenis, "Software and hardware co-design for low-power hpc platforms," Berlin, Heidelberg, 2019, p. 88–100, Springer-Verlag.

[3] Biagioni, Andrea et al., "Euroexa custom switch: an innovative fpga-based system for extreme scale computing in europe," *EPJ Web Conf.*, vol. 245, pp. 09004, 2020.

[4] R Ammendola, A Biagioni, O Frezza, F Lo Cicero, A Lonardo, P S Paolucci, D Rossetti, F Simula, L Tosoratto, and P Vicini, "APEnet+: a 3D torus network optimized for GPU-based HPC systems," *Journal of Physics: Conference Series*, vol. 396, no. 4, pp. 042059, dec 2012.

[5] Jeffrey S Vetter, Ron Brightwell, Maya Gokhale, Pat McCormick, Rob Ross, John Shalf, Katie Antypas, David Donofrio, Travis Humble, Catherine Schuman, et al., "Extreme heterogeneity 2018-productive computational science in the era of extreme heterogeneity: Report for doe ascr workshop on extreme heterogeneity," 2022.

[6] Ken Raffenetti, Antonio J Pena, and Pavan Balaji, "Toward implementing robust support for portals 4 networks in mpich," in *2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*. IEEE, 2015, pp. 1173–1176.

[7] Andrea Biagioni et al., "RED-SEA: network solution for exascale architectures," in *25th Euromicro Conference on Digital System Design, DSD 2022, Maspalomas, Spain, August 31 - Sept. 2, 2022*. 2022, pp. 712–719, IEEE.

[8] András Varga, "Omnet++," in *Modeling and Tools for Network Simulation*, Klaus Wehrle, Mesut Günes, and James Gross, Eds., pp. 35–59. Springer, 2010.

[9] George F. Riley and Thomas R. Henderson, *The ns-3 Network Simulator*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[10] Nathan Binkert et al., "The gem5 simulator," *SIGARCH Comput. Archit. News*, vol. 39, no. 2, pp. 1–7, aug 2011.

[11] Nikolaos Tampouratzis, Ioannis Papaefstathiou, Antonios Nikitakis, Andreas Brokalakis, Stamatis Andrianakis, Apostolos Dollas, Marco Marcon, and Emanuele Plebani, "A novel, highly integrated simulator for parallel and distributed systems," *ACM Trans. Archit. Code Optim.*, vol. 17, no. 1, mar 2020.

[12] P. Yebenes, J. Escudero-Sahuquillo, P. J. Garcia, and F. J. Quiles, "Networks of exascale systems with omnet++.," in *Euromicro International Conference on Parallel, Distributed, and Network-Based Processing*, 2013, pp. 203–207.

[13] F. J. Andújar, J. A. Villar, F.J. Alfaro, and et al., "An open-source family of tools to reproduce mpi-based workloads in interconnection network simulators.," *Journal of Supercomputing*, , no. 72, pp. 042059, 2016.

[14] Daniele De Sensi, Salvatore Di Girolamo, Kim H. McMahon, Duncan Roweth, and Torsten Hoefler, "An in-depth analysis of the slingshot interconnect," in *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, 2020, pp. 1–14.

[15] Mark S. Birrittella, Mark Debbage, Ram Huggahalli, James Kunz, Tom Lovett, Todd Rimmer, Keith D. Underwood, and Robert C. Zak, "Intel® omni-path architecture: Enabling scalable, high performance fabrics," in *2015 IEEE 23rd Annual Symposium on High-Performance Interconnects*, 2015, pp. 1–9.

[16] Saïd Derradji, Thibaut Palfer-Sollier, Jean-Pierre Panziera, Axel Poudes, and François Wellenreiter Atos, "The bxi interconnect architecture," in *2015 IEEE 23rd Annual Symposium on High-Performance Interconnects*. IEEE, 2015, pp. 18–25.

[17] Roberto Ammendola, Massimo Bernaschi, Andrea Biagioni, Mauro Bisson, Massimiliano Fatica, Ottorino Frezza, Francesca Lo Cicero, Alessandro Lonardo, Enrico Mastrostefano, Pier Stanislao Paolucci, Davide Rossetti, Francesco Simula, Laura Tosoratto, and Piero Vicini, "Gpu peer-to-peer techniques applied to a cluster interconnect," in *2013 IEEE International Symposium on Parallel Distributed Processing, Workshops and Phd Forum*, 2013, pp. 806–815.

[18] Adrià Armejach, Bine Brank, Jordi Cortina, François Dolique, Timothy Hayes, Nam Ho, Pierre-Axel Lagadec, Romain Lemaire, Guillem López-Paradís, Laurent Marliac, Miquel Moretó, Pedro Marcuello, Dirk Pleiter, Xubin Tan, and Said Derradji, "Mont-blanc 2020: Towards scalable and power efficient european hpc processors," in *2021 Design, Automation Test in Europe Conference Exhibition (DATE)*, 2021, pp. 136–141.

[19] Norbert Eicker, Thomas Lippert, Thomas Moschny, Estela Suarez, and for the DEEP project, "The deep project an alternative approach to heterogeneous cluster-computing in the many-core era," *Concurrency and Computation: Practice and Experience*, vol. 28, no. 8, pp. 2394–2411, 2016.

[20] Theodoropoulos et al., "The AXIOM project (Agile, eXtensible, fast I/O Module)," in *2015 International Conference on Embedded Computer Systems: Architectures, Modeling, and Simulation (SAMOS)*, 2015, pp. 262–269.

[21] "Epi: European processor iniciative," .

[22] Salvatore Di Girolamo, Andreas Kurth, Alexandru Calotoiu, Thomas Benz, Timo Schneider, Jakub Beránek, Luca Benini, and Torsten Hoefler, "A risc-v in-network accelerator for flexible high-performance low-power packet processing," in *Proceedings of the 48th Annual International Symposium on Computer Architecture*, 2021, ISCA '21.

[23] Daniele De Sensi, Salvatore Di Girolamo, Saleh Ashkboos, Shigang Li, and Torsten Hoefler, "Flare: Flexible in-network allreduce," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, New York, NY, USA, 2021, SC '21, Association for Computing Machinery.

[24] Nikolaos Chrysos and Manolis Katevenis, "Scheduling in non-blocking buffered three-stage switching fabrics.," in *INFOCOM*, 2006, vol. 6, pp. 1–13.

[25] Antonis Psistakis, Nikos Chrysos, Fabien Chaix, Marios Asiminakis, Michalis Gianioudis, Pantelis Xirouchakis, Vassilis Papaefstathiou, and Manolis Katevenis, "Optimized page fault handling during rdma," *IEEE Transactions on Parallel and Distributed Systems*, pp. 1–1, 2022.