
TENDENCIAS EN HPC
asac
comunicaciones

HPC

- La *informática de alto rendimiento (HPC)* se utiliza para generar y autentificar nuevos conocimientos.

- Como estas ideas suelen ser novedosas, suelen ser bastante valiosas.

- El objetivo principal de los sistemas HPC es maximizar el ritmo al que producen propiedad intelectual valiosa y al mismo tiempo minimizar los costes.

- La mayor parte del tiempo en HPC, la propiedad intelectual se genera al completar tareas de cargas de trabajo o "jobs".

- Para optimizar el valor del sistema HPC, debemos centrarnos en intentar maximizar el *rendimiento*.

HPC

- *¿Cuáles son los mejores indicadores de rendimiento?*

- La mayoría de las personas que tienen experiencia en HPC piensan inmediatamente en el *ranking TOP500* y en el *benchmarking HP Linpack*, que se ha estado utilizando durante décadas atrás para clasificar los supercomputadores.

- La naturaleza de la informática de alto rendimiento está evolucionando. Las cargas de trabajo intensivas en datos, los algoritmos de precisión mixtos y reducidos, la inteligencia artificial y el aprendizaje automático (ML) están cambiando la forma en que se realiza HPC.

- Es por ello que surgen nuevas herramientas de comparación cuantitativas centradas en métodos emergentes y precisiones no cubiertas por *HP Linpack.*

HPC

- *Algunas de estas nuevas herramientas:*

a) *Green500*: Utiliza la puntuación de *HP Linpack* enviada para TOP500 dividida por el uso máximo de vatios durante la ejecución para proporcionar una métrica combinada de *flops/watt*.

b) *High-Performance Conjugate Gradients (HPCG)*: Su objetivo es modelar los patrones de acceso a datos de aplicaciones del mundo real, como cálculos de matrices dispersas, probando así el efecto de las limitaciones del subsistema de memoria y la interconexión interna del clúster en su rendimiento computacional.

c) *HPL-AI*: Es un *benchmarking* sintético similar a HPL, que realiza cálculos matriciales con precisión de 64 bits. en punto flotante.

… La versión de IA realiza cálculos matriciales con niveles más bajos de precisión, teniendo en cuenta que los modelos de aprendizaje automático a menudo aprovechan cálculos de menor precisión para el entrenamiento (fp32, 16, 8 e incluso int8) y logran un mayor rendimiento sin sacrificar la precisión.

d) *MLPerf HPC - ML Commons*: Es otro enfoque para definir grupos de *benchmarking* centrados en el aprendizaje automático (ML).

… Graph500, HPC Challenge (HPCC), SciML …

Tendencias en HPC

- *Énfasis en la eficiencia energética de los sistemas*: Varios factores están contribuyendo a la conciencia del impacto ambiental que tiene la informática, siendo el principal la tasa de crecimiento.

- Aunque el HPC representa un pequeño porcentaje del total de la informática consumida en el mundo, se utiliza como laboratorio para desarrollar y probar tecnologías que se mercantilizan y, finalmente, se utilizan en dispositivos más modestos.

- Esto llevó en 2005 a la elaboración de la lista *Green500*, *benchmarking* que puntúa el rendimiento y la potencia juntos. Esta lista es ahora la principal forma de reconocimiento para los líderes en HPC que se esfuerzan por maximizar la eficiencia energética de los entornos HPC que administran.

Tendencias en HPC

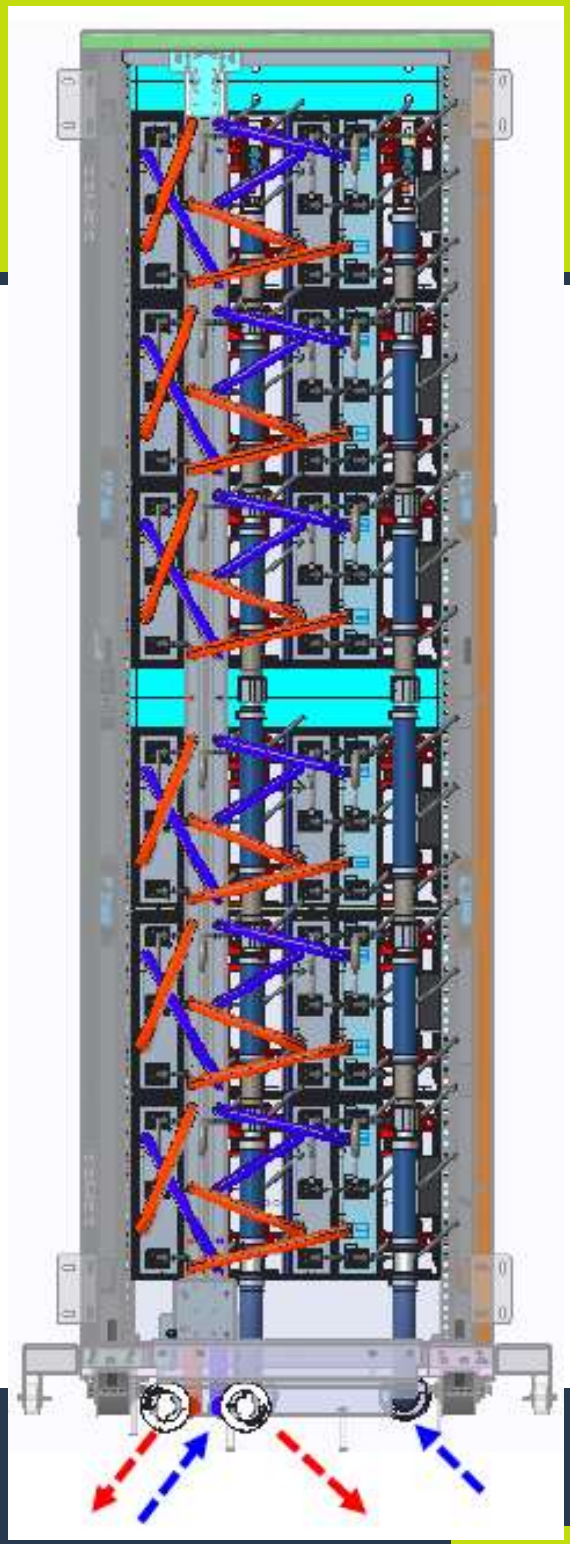
- Una medida de la eficiencia energética del CPD es la eficacia del uso de energía (*PUE = Power Usage Effectiveness*).

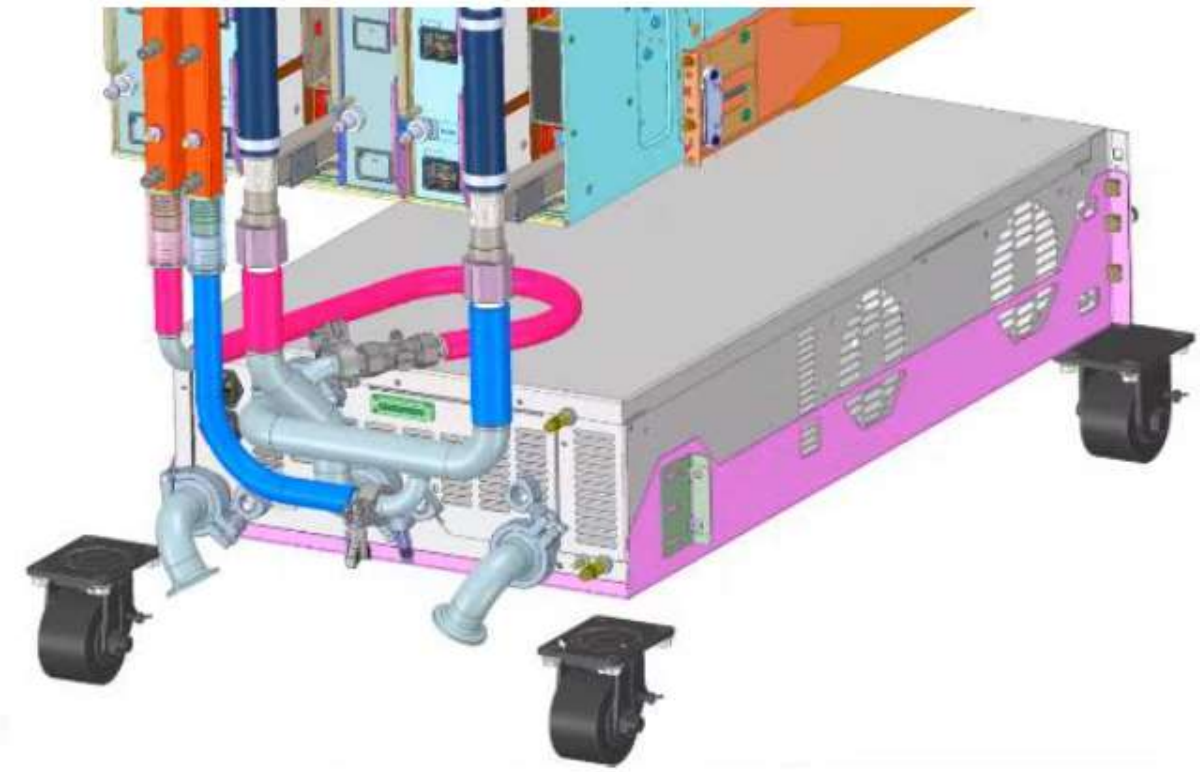
- El objetivo es acercar el *PUE* del centro de datos a 1,00 lo máximo posible, para que toda la energía consumida se utilice para impulsar la informática y no para otras necesidades parásitas como refrigeración, movimiento de aire, bombeo de agua, conversión de CA a CC, etc.

- La gran mayoría de los CPDs actuales utilizan aire enfriado para mantener sus equipos funcionando dentro de las temperaturas de operación. Desafortunadamente, la termodinámica del enfriamiento del aire hace que esta sea una solución ineficiente.

- ALTERNATIVA: refrigeración por agua directa en las partes del servidor que consuman mucha energía y disipen más calor.

Tendencias en HPC

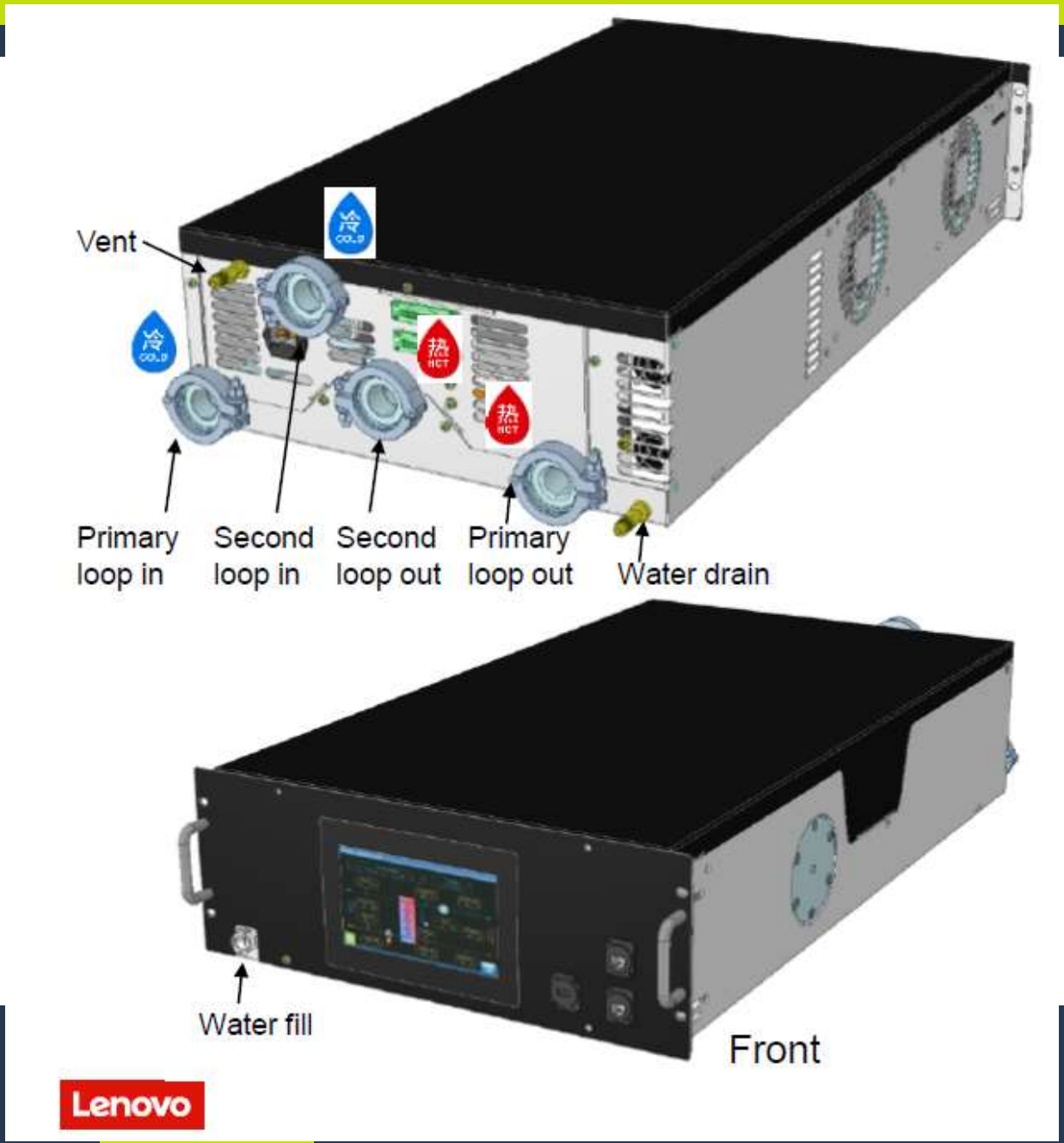




CDU = Coolant Distribution Unit

asac comunicaciones

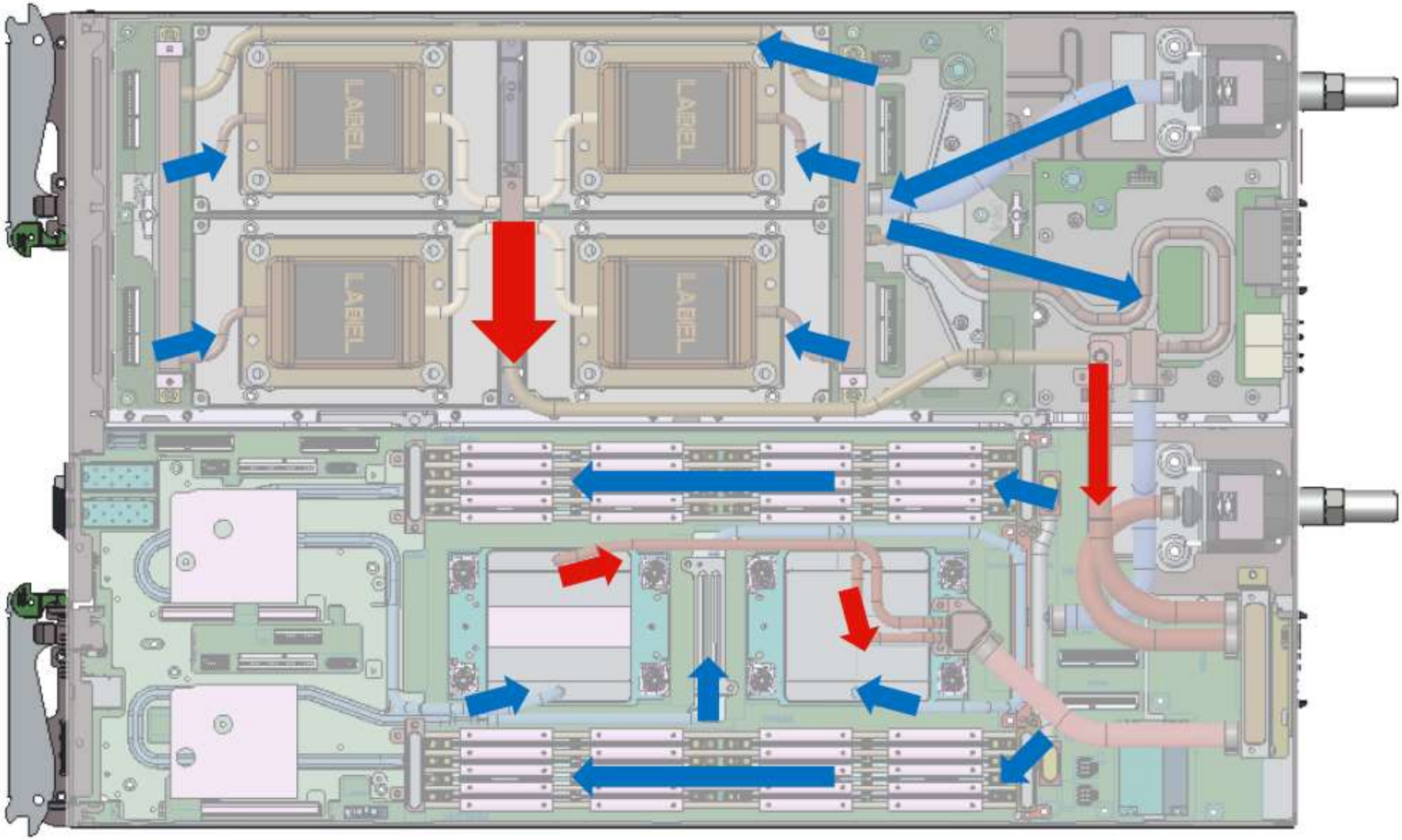
Tendencias en HPC



Lenovo

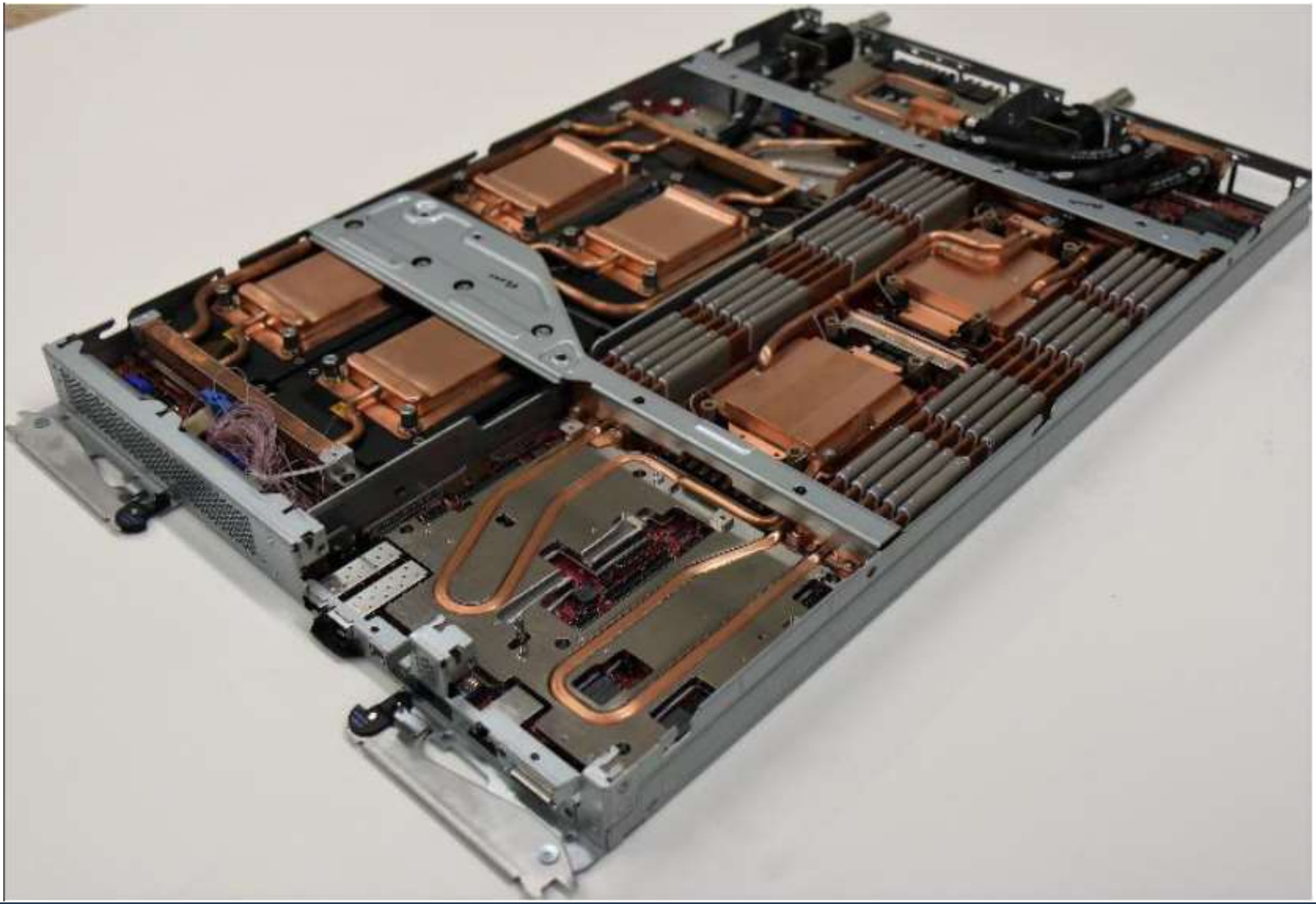
asac comunicaciones

Tendencias en HPC



asac comunicaciones

Tendencias en HPC



Tendencias en HPC

- Esto es mucho más eficiente que el enfriamiento por aire, pero aún requiere bombas, enfriadores y una cantidad significativa de tuberías.

- Los centros de datos más innovadores están explorando la refrigeración por inmersión, en la que todo el servidor se sumerge en algún tipo de líquido conductor de calor que está eléctricamente aislado. Esto es muy ventajoso debido al hecho de que las propiedades térmicas de un líquido son infinitamente mejores en términos de transferencia de calor y capacitancia.

- En la búsqueda de un PUE más alto y *gigaflops/watt* optimizados, replantear la forma en que se distribuye la energía puede ser otra alternativa que vale la pena explorar: CPDs alimentados por corriente continua.

Tendencias en HPC

- *Adopción de nuevas arquitecturas para HPC: Arm64,* con muy bajo *gigaflops/watt*. Con la introducción de la computación acelerada para Arm64, se pueden esperar grandes avances en eficiencia a medida que se vayan portando aplicaciones a esta nueva arquitectura (*GAMESS, GROMACS, LAMMPS, OpenFOAM, WRF, NAMD y MILC* ya están disponibles en Arm64).

- *CPU NVIDIA Grace basada en Arm64*.

asac comunicaciones

Tendencias en HPC

ANNOUNCING NVIDIA GRACE

Breakthrough CPU Designed for Giant-Scale AI and HPC Applications

FASTEST INTERCONNECTS
>900 GB/s Cache Coherent NVLink CPU To GPU (14x)
>600GB/s CPU To CPU (2x)

HIGHEST MEMORY BANDWIDTH
>500GB/s LPDDR5x w/ ECC
>2x Higher B/W
10x Higher Energy Efficiency

NEXT GENERATION ARM NEOVERSE CORES
>300 SPECrate2017_int_base
Availability 2023

asac comunicaciones

Tendencias en HPC

- La convergencia de HPC y la IA también ha generado cierto interés en el uso de aceleradores de IA especializados aplicados a cargas de trabajo de HPC, por ejemplo:

a) *Cerebras System CS-1* / ordenador AI basado en *Wafer-Scale Engine (WSE)*, diseñado para AI y ML (@LLNL).

b) *SambaNova Systems DataScale / Reconfigurable Dataflow Unit (RDU) / SambaNova Systems Cardinal SN10 RDU* / (@LLNL).

c) *Graphcore's Bow Pod systems / Intelligence Processing Unit (IPU).*

(LLNL = Lawrence Livermore National Laboratory)

FPGAs en HPC

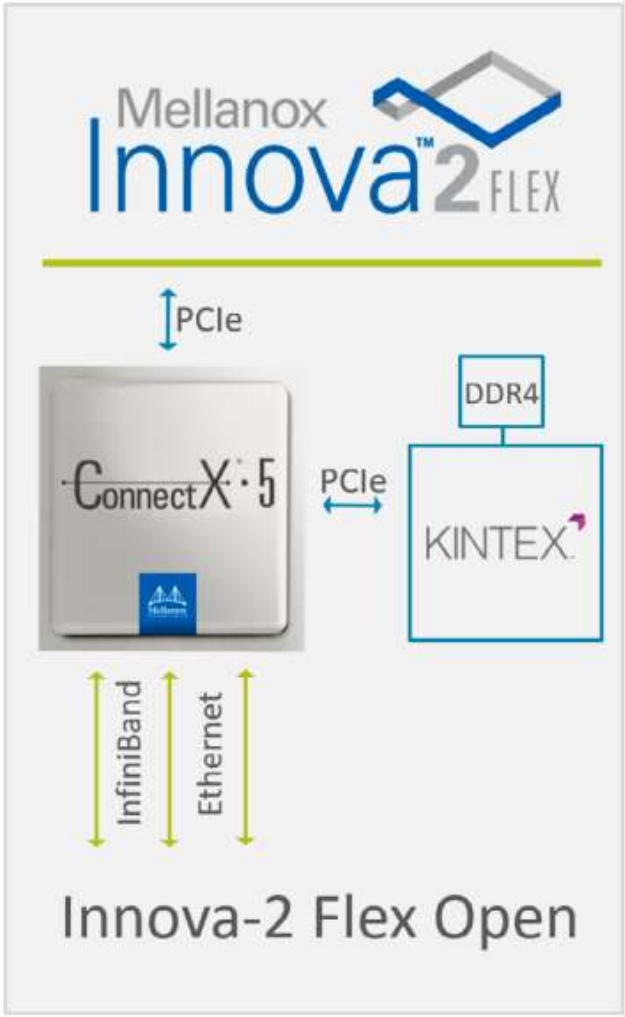
- HPC implica procesar grandes cantidades de datos y realizar cálculos complejos a muy alta velocidad.
- Las CPUs y GPUs tradicionales tienen limitaciones en términos de rendimiento y flexibilidad.
- Posible solución: uso de FPGAs.
- Ofrecen ventajas únicas en HPC:
    - Baja latencia (latencia determinista para aplicaciones de tiempo real).
    - Alta eficiencia energética.
    - Alto paralelismo.
    - Versatilidad (reconfiguración).
    - Procesamiento con gran ancho de banda.

FPGAs en HPC

- Aplicaciones potenciales de uso de FPGAs en sistemas HPC:
  - Análisis de *Big Data (Hadoop / Spark)*.
  - Inteligencia artificial (AI) y aprendizaje automático (ML).
  - Simulaciones científicas.
  - Comercio financiero.
  - Procesamiento de paquetes de red.
  - Procesamiento de imágenes.
- Las FPGAs se programan utilizando un lenguaje de descripción hardware (HDL), como *VHDL* y *Verilog*.
- Las GPUs se configuran utilizando lenguajes de programación de software de propósito general: C, C++, Java, Python …

FPGAs en HPC

- Caso práctico: acelerador de inferencia de aprendizaje profundo (DL) para impulsar cargas de trabajo de redes neuronales convolucionales (CNNs) computacionalmente intensivas para el reconocimiento y clasificación de vídeo e imágenes en tiempo real u *offline* sobre hardware FPGA optimizado.



FPGAs en HPC

- Primitivas CNN implementadas:
  - *Convolution.*
  - *Rectified Linear Unit (ReLU).*
  - *Local Response Normalization (LRN).*
  - *Pooling (max & average).*
  - *Fully-connected.*
  - *Concatenation.*
- Topologías de red neuronal: *AlexNet, GoogLeNet, CaffeNet, LeNet, VGG-16, SqueezeNet*.
- Herramientas: librería *Intel MKL-DNN, framework* de *deep learning Caffe* y el *runtime OpenVINO*, abstrayendo la complejidad de la programación FPGA de bajo nivel.

FPGAs en HPC

asac comunicaciones

DPUs en HPC

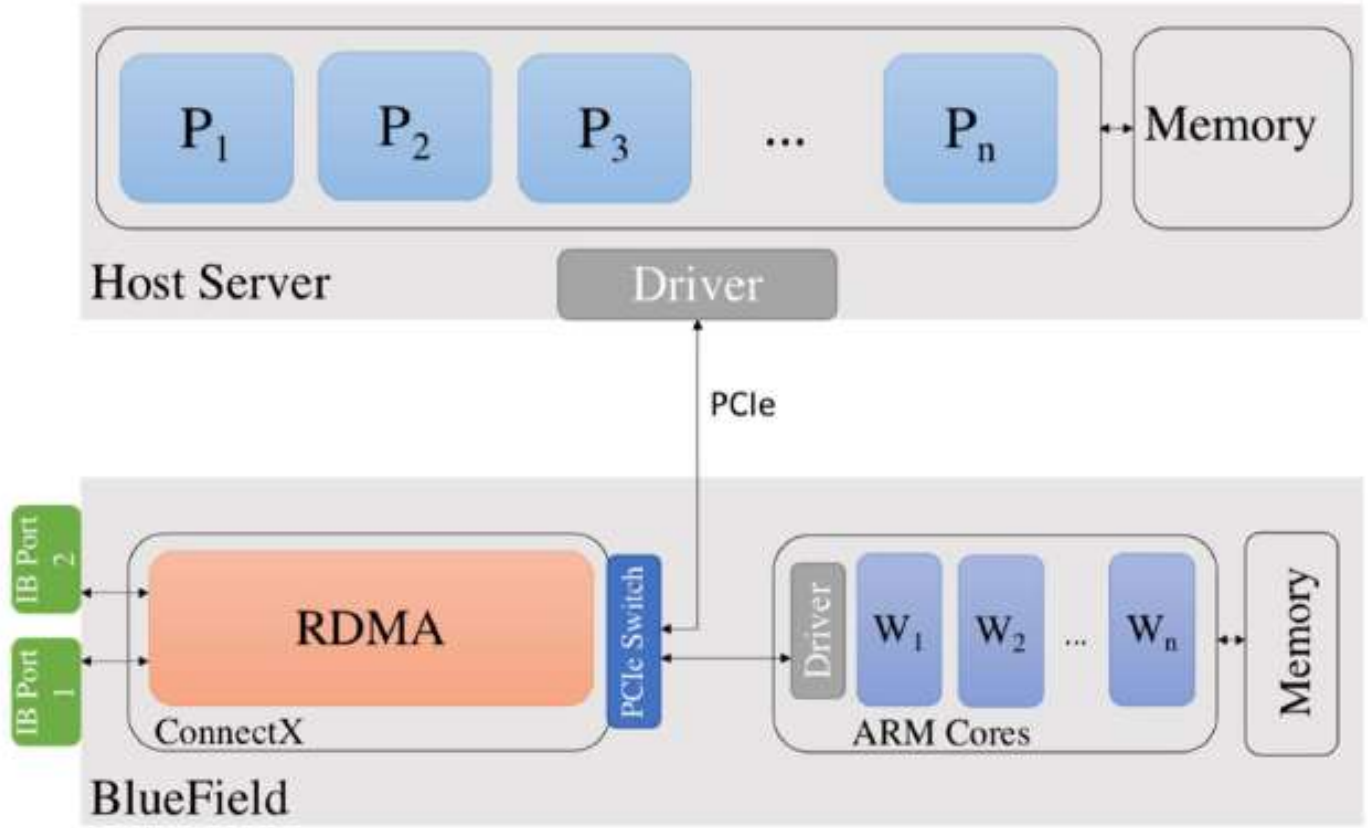
- Las DPUs son una nueva clase de procesador programable especializado en mover datos en CPDs que combinan 3 elementos fundamentales:

1.- CPU multicore de alto rendimiento programable por software , generalmente basada en arquitectura Arm de 64 bits, integrada de forma óptima con el resto de componentes del SoC.

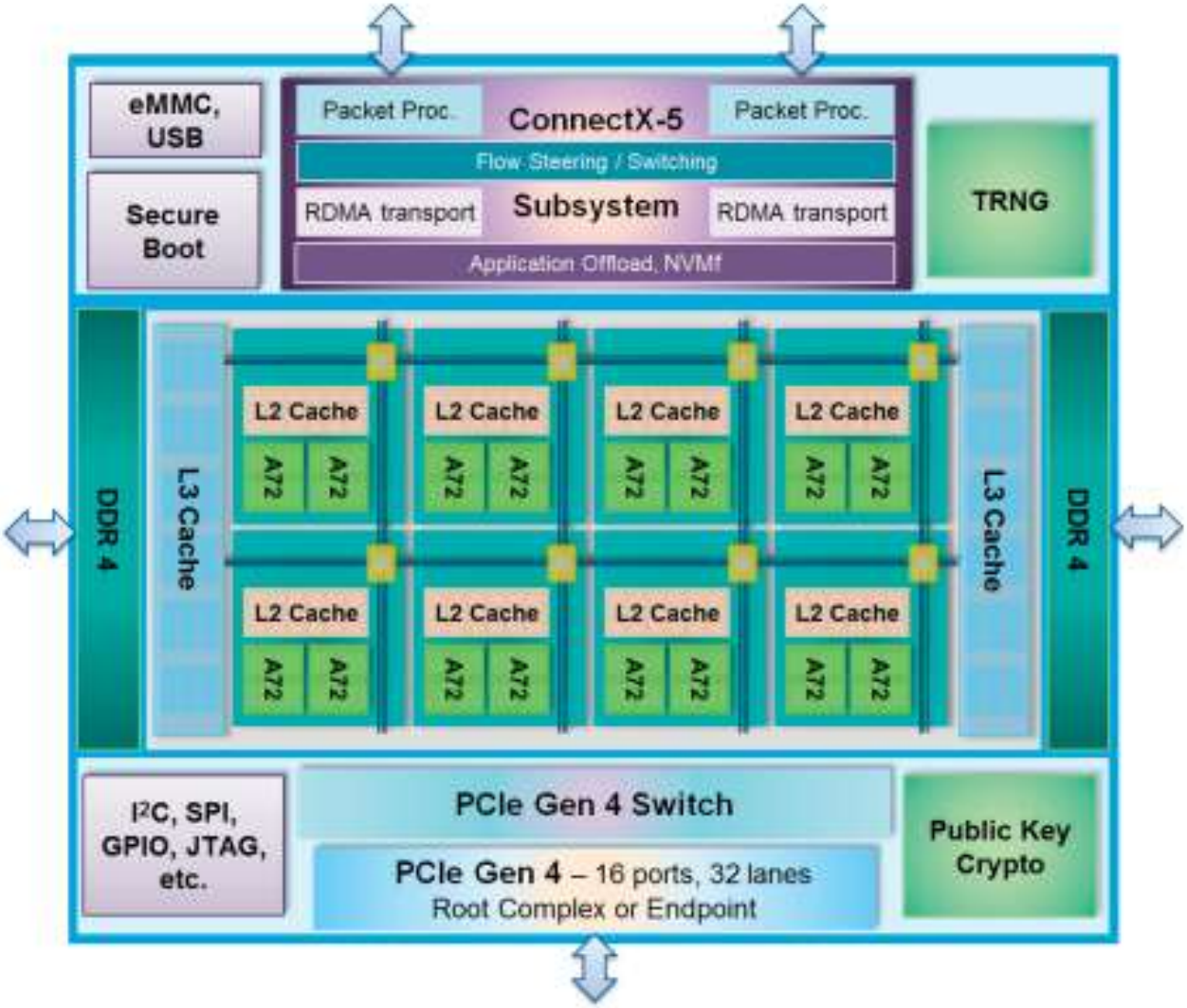
2.- Una interfaz de red de alto rendimiento capaz de analizar, procesar y transferir datos a GPU y a CPU de manera eficiente y a muy alta velocidad.

3.- Un amplio conjunto de motores de aceleración flexibles y programables que descargan y mejoran el rendimiento de las aplicaciones para AI, ML, seguridad, red, almacenamiento, …

DPUs en HPC


P1
P2
P3
...
Pn
Memory
Host Server
Driver
PCIe
IB Port 2
IB Port 1
RDMA
ConnectX
PCIe Switch
Driver
W1
W2
...
Wn
ARM Cores
Memory
BlueField

DPUs en HPC


eMMC, USB
Secure Boot
Packet Proc.
ConnectX-5
Packet Proc.
Flow Steering / Switching
RDMA transport
Subsystem
RDMA transport
Application Offload, NVMf
TRNG
DDR 4
L3 Cache
L2 Cache
A72
L3 Cache
DDR 4
I²C, SPI, GPIO, JTAG, etc.
PCIe Gen 4 Switch
PCIe Gen 4 – 16 ports, 32 lanes
Root Complex or Endpoint
Public Key Crypto

DPUs en HPC

- DPUs para entornos de almacenamiento:
  - Acceso arrays AFA / JBOF / usando NVMe-oF / Ceph / Lustre.
  - iSCSI/TCP offload.
  - Compresión / descompresión / deduplicación de datos.
  - Implementación de RDMA: ofrece rendimiento de acceso remoto al almacenamiento equivalente al del almacenamiento local, con una sobrecarga mínima de CPU.
- DPUs para entornos de red y seguridad:
  - Ofrecen total flexibilidad para implementar el plano de datos y control.
  - *Encryption/SSL offloading*: para criptografía simétrica y asimétrica.

DPUs en HPC

- Aceleración de Open vSwitch (OVS) para ofrecer servicios altamente eficientes de *switching* virtual con capacidad de *routing*.
- Overlay Networks: VxLAN, NVGRE, GENEVE.

- DPUs para ML:
  - Aceleración RDMA: el controlador de red del *data path* utiliza tecnología RDMA y RoCE, ofreciendo baja latencia y alto rendimiento con ciclos de CPU casi nulos.
  - Permite la interconexión de GPUs a través del switch PCIe integrado.
  - *PeerDirect / GPUDirect RDMA*: comunicación *peer-to-peer* entre DPUs y GPUs remotos a través del *fabric* sin comprometer recursos del sistema (memoria, CPU).

DPUs en HPC

- Las arquitecturas de DPUs modernas ofrecen la posibilidad de descargar la computación y comunicación a los núcleos del SoC, lo que permite:

a) Una capacidad computacional adicional.

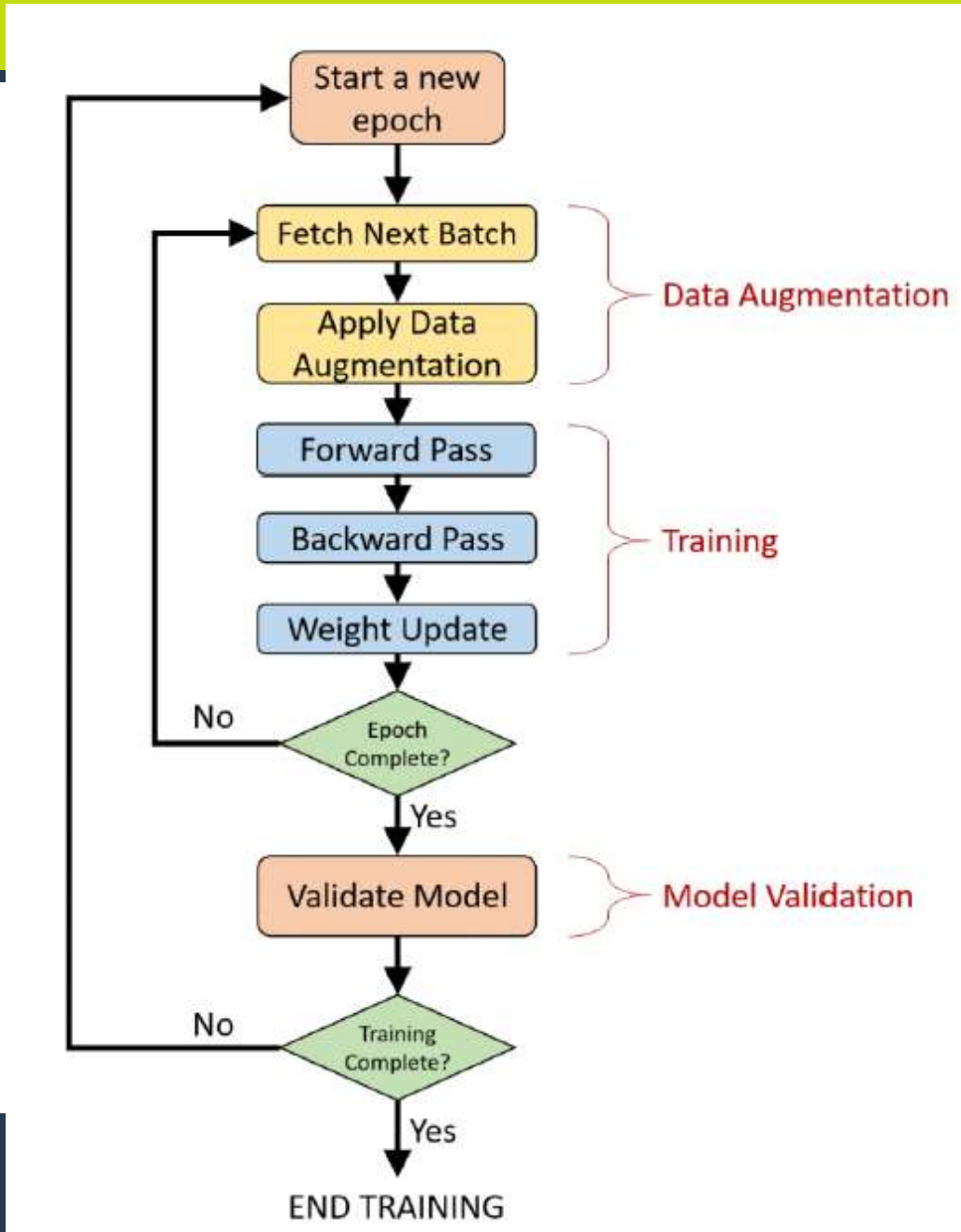
b) El delegar partes del cálculo de un programa sobre los núcleos Arm.

c) Potencial para descargar la comunicación desde la CPU del host a los núcleos Arm.

- "Offloading MPI level Non blocking Collectives Communication"

- "Offloading some parts of computation in DNN Training (Offloads data augmentation and model validation to DPUs)"

DPUs en HPC


Start a new epoch
Fetch Next Batch
Apply Data Augmentation
Data Augmentation
Forward Pass
Backward Pass
Weight Update
Training
No
Epoch Complete?
Yes
Validate Model
Model Validation
No
Training Complete?
Yes
END TRAINING

ICNs en HPC

- Una red de interconexión de alto rendimiento es una parte importante de un sistema HPC y su rendimiento tiene un impacto decisivo en la eficiencia de las aplicaciones paralelas.

- La tendencia de desarrollo de las ICNs se refleja principalmente en el aumento de la escala y el ancho de banda de las redes.

- Las ICNs pueden clasificarse de muchas maneras y caracterizarse en función de varios parámetros, principalmente:
  - Ancho de banda de la red.
  - Latencia.
  - Switch radix.
  - Topología de la red.

- La elección de la ICN está influenciada por el rendimiento del nodo y la tecnología de interconexión.

ICNs en HPC

- *InfiniBand* es un bus de comunicaciones serie de alta velocidad, baja latencia y de baja sobrecarga de CPU, diseñado tanto para conexiones internas como externas.

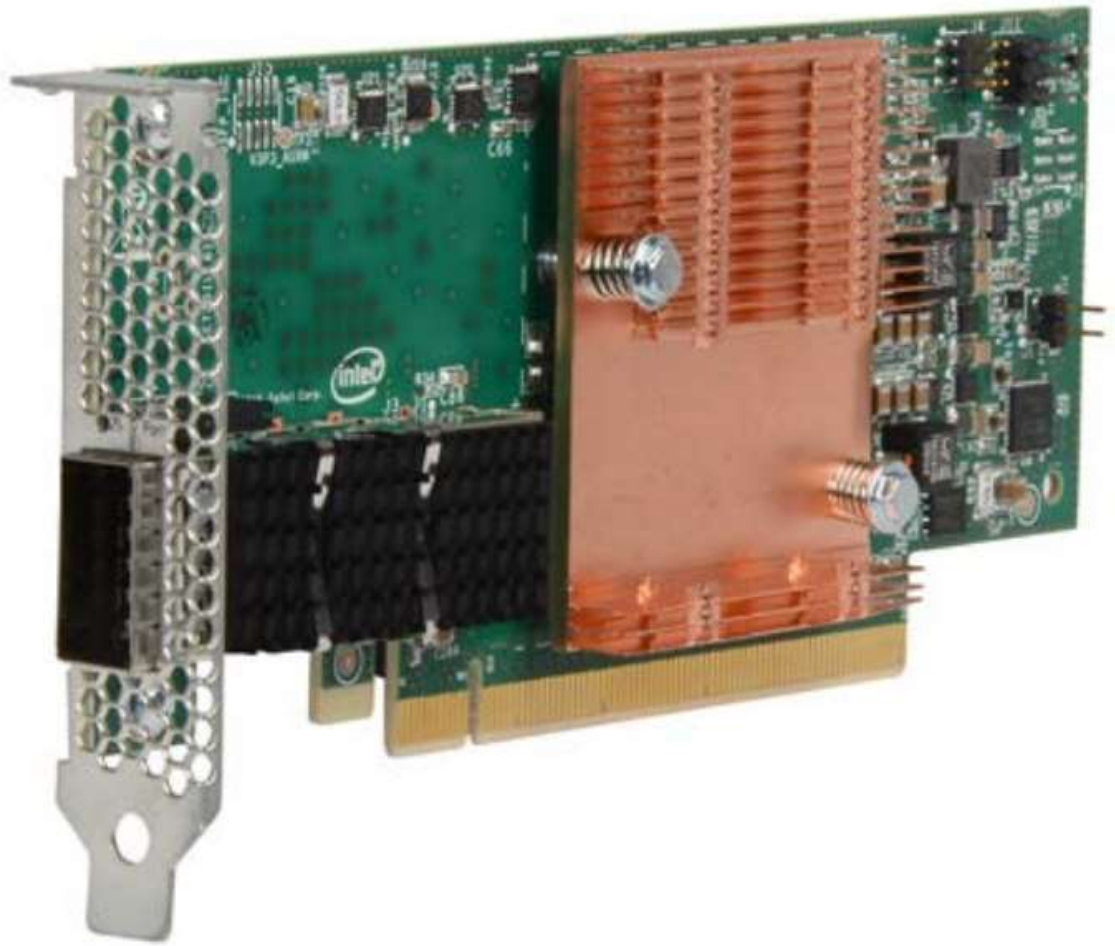
Sus especificaciones son desarrolladas y mantenidas por la InfiniBand Trade Association (IBTA).

asac comunicaciones

ICNs en HPC

- ***Omni-Path (OPA)*** **es una arquitectura de comunicación de alto rendimiento en desuso propiedad de Intel. Su objetivo es lograr una baja latencia de comunicación, un bajo consumo de energía y un alto rendimiento. En 2021** ***Cornelis Networks*** **anunció su continuidad con el lanzamiento de** ***Omni-Path Express HPC Interconnect*****.**

ICNs en HPC



asac comunicaciones

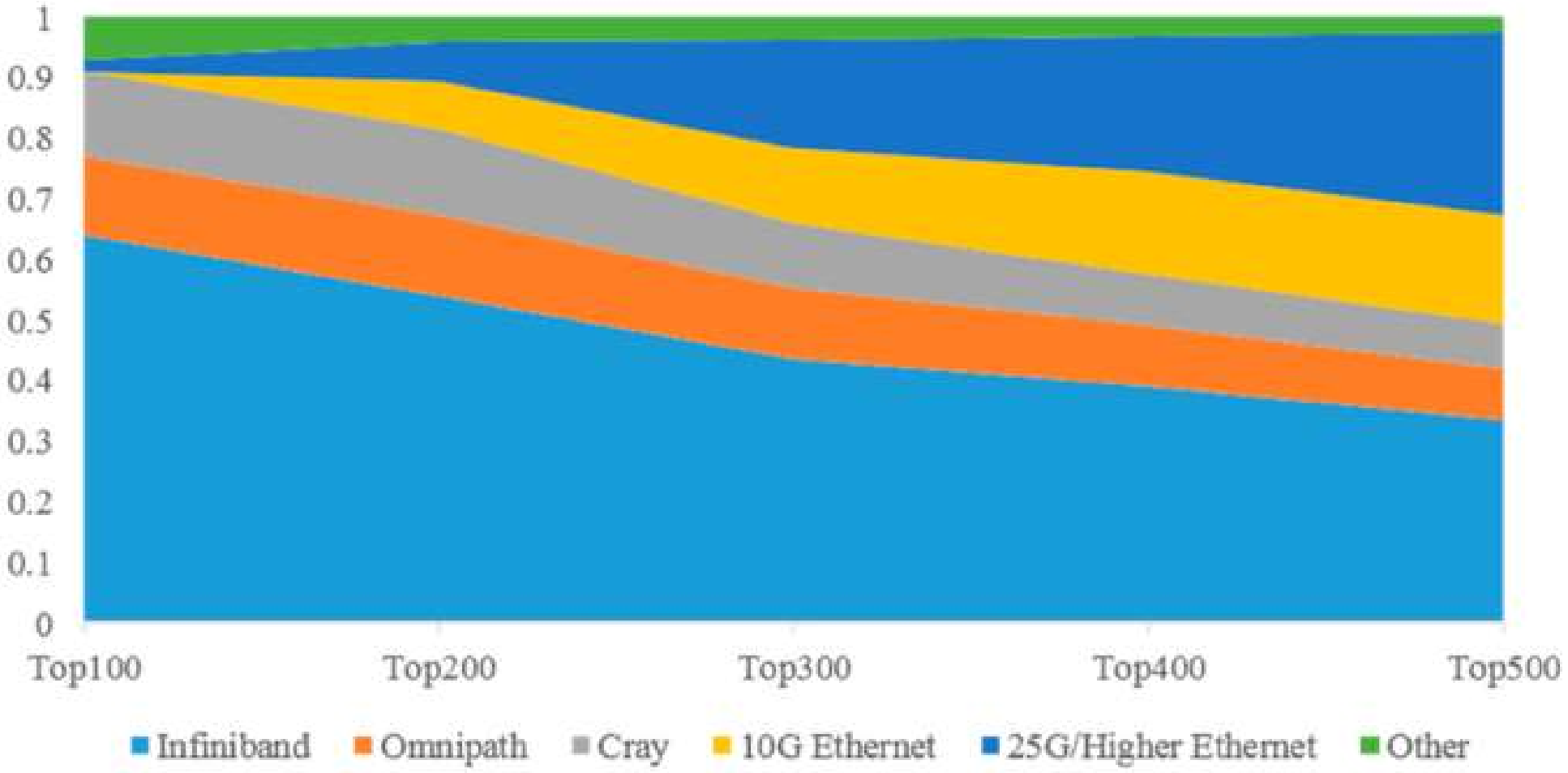
ICNs en HPC

- *Slingshot (HPE/Cray)*: Es una versión personalizada de Ethernet optimizada específicamente para HPC. Agrega optimizaciones de protocolo especiales al tiempo que permite mezclar el tráfico Ethernet estándar.

HPC Ethernet / Hace uso del switch ASIC custom *Rosetta* de Cray.

(*Frontier, 1º TOP500, Slingshot-11*)

ICNs en HPC



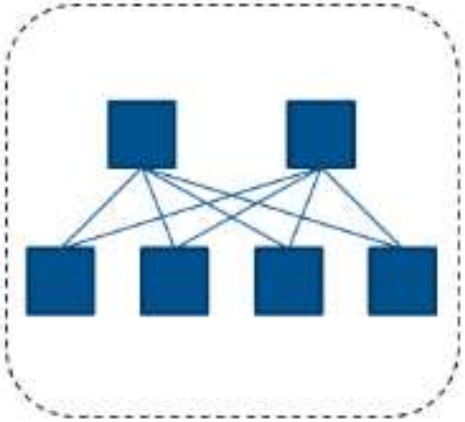
- Otras ICNs: ***Tofu, TH Express/Sunway, Bull BXI***

Conectividad IB en HPC

- Conmutadores *NVIDIA Quatum-2*.
- Implementan *SHARP* - Protocolo Escalable de Agregación y Reducción Jerárquica.
- Las aplicaciones de informática científica e inteligencia artificial hacen un uso extensivo de operaciones colectivas de MPI como *All-Reduce* y *All-to-All*.
- *SHARP*, para computación en la red, permite descargar estas operaciones colectivas desde la CPU del host a la red del conmutador, reduciendo drásticamente el tiempo de las operaciones MPI.

asac comunicaciones

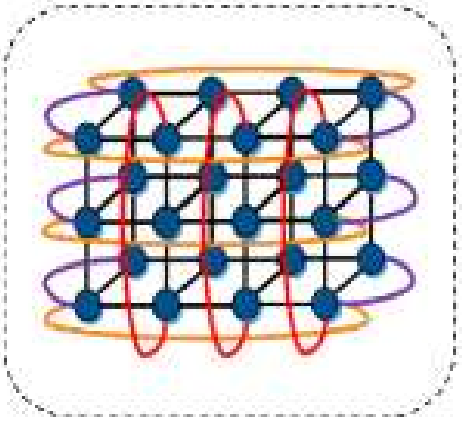
Topologías de red en HPC



- ***Fat Tree***:
  - Es una de las topologías más utilizadas.
  - Es una buena opción para gran variedad de aplicaciones, ya que proporciona baja latencia y permite distintas opciones de rendimiento, desde conectividad sin bloqueo hasta suscripciones excesivas.
  - Maximiza el rendimiento de datos para distintos tipos de patrones de tráfico.
  - INCONVENIENTE: Costosa a gran escala debido a la gran cantidad de conmutadores y enlaces que requiere.

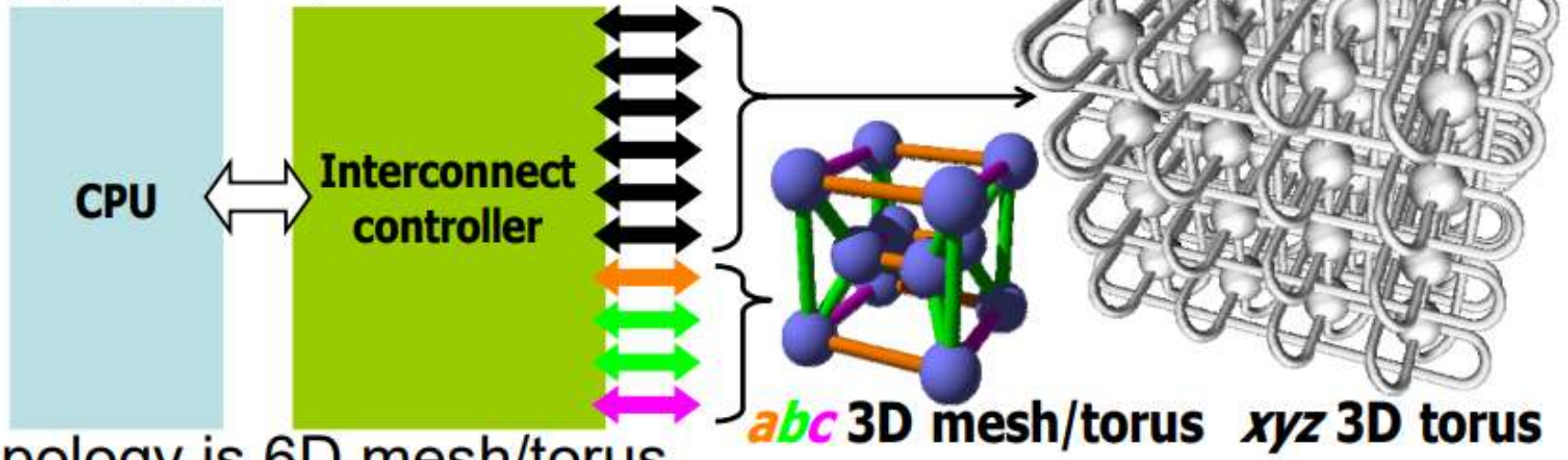
Topologías de red en HPC

- ***Torus***:
  - **Interconectan directamente un host con varios de sus vecinos en una red de *k-dimensional*.**
  - **Son económicas pero proporcionan un bajo rendimiento de red para patrones de tráfico no favorables.**
  - **Debido a su naturaleza de bloqueo y mayor latencia, no es la opción preferida para entornos HPC.**
  - ***Fugaku, 2º TOP500,* utiliza *Tofu (Torus Fusion) Interconnect D*, basada en *topología toroidal 6D-mesh*.**



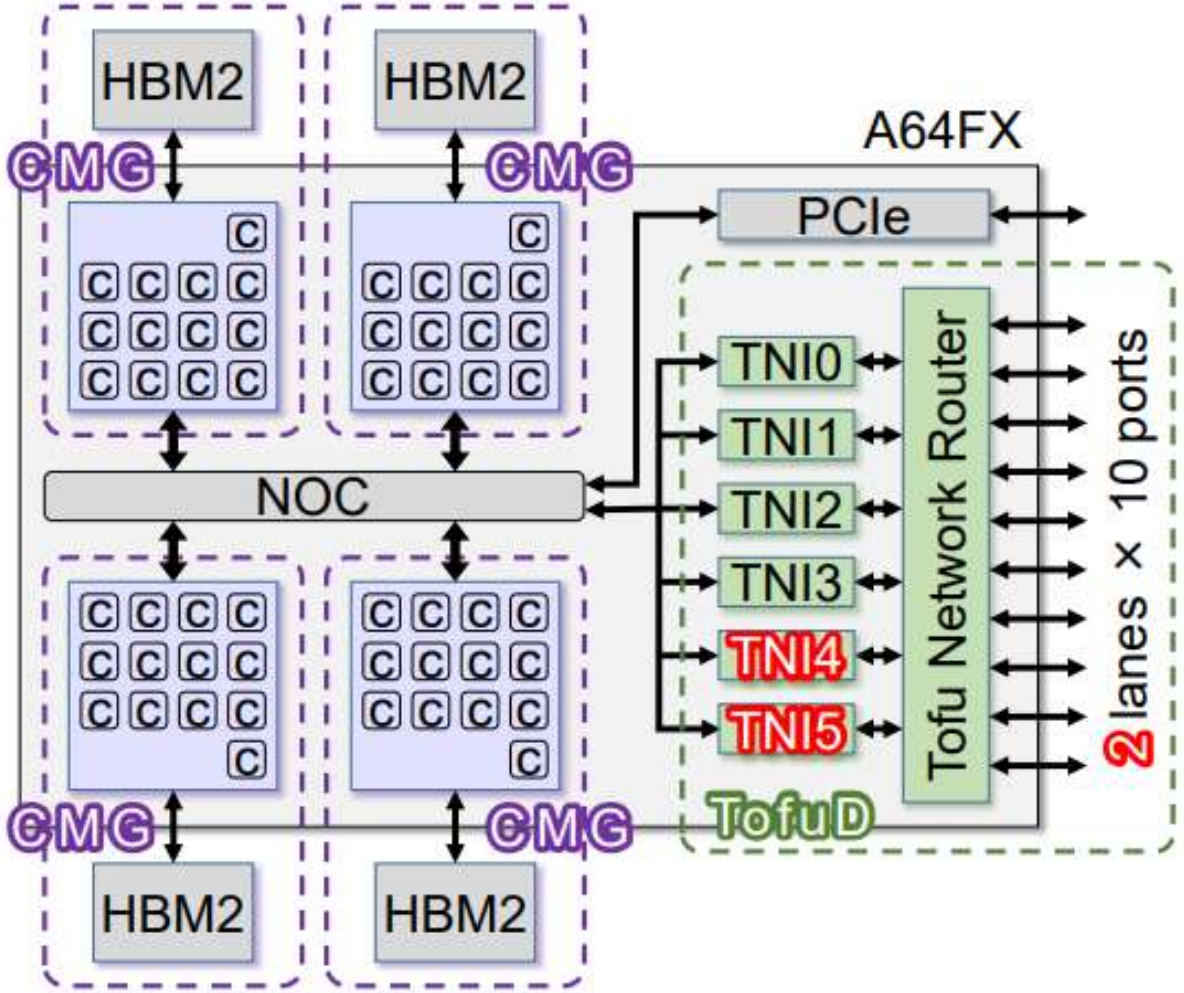
Topologías de red en HPC

- 6 links ⇨ Scalable ***xyz*** 3D torus
- 4 links ⇨ Fixed size ***abc*** 3D mesh/torus
  - |***a***|=2, |***b***|=3, |***c***|=2 ⇨ 12 nodes

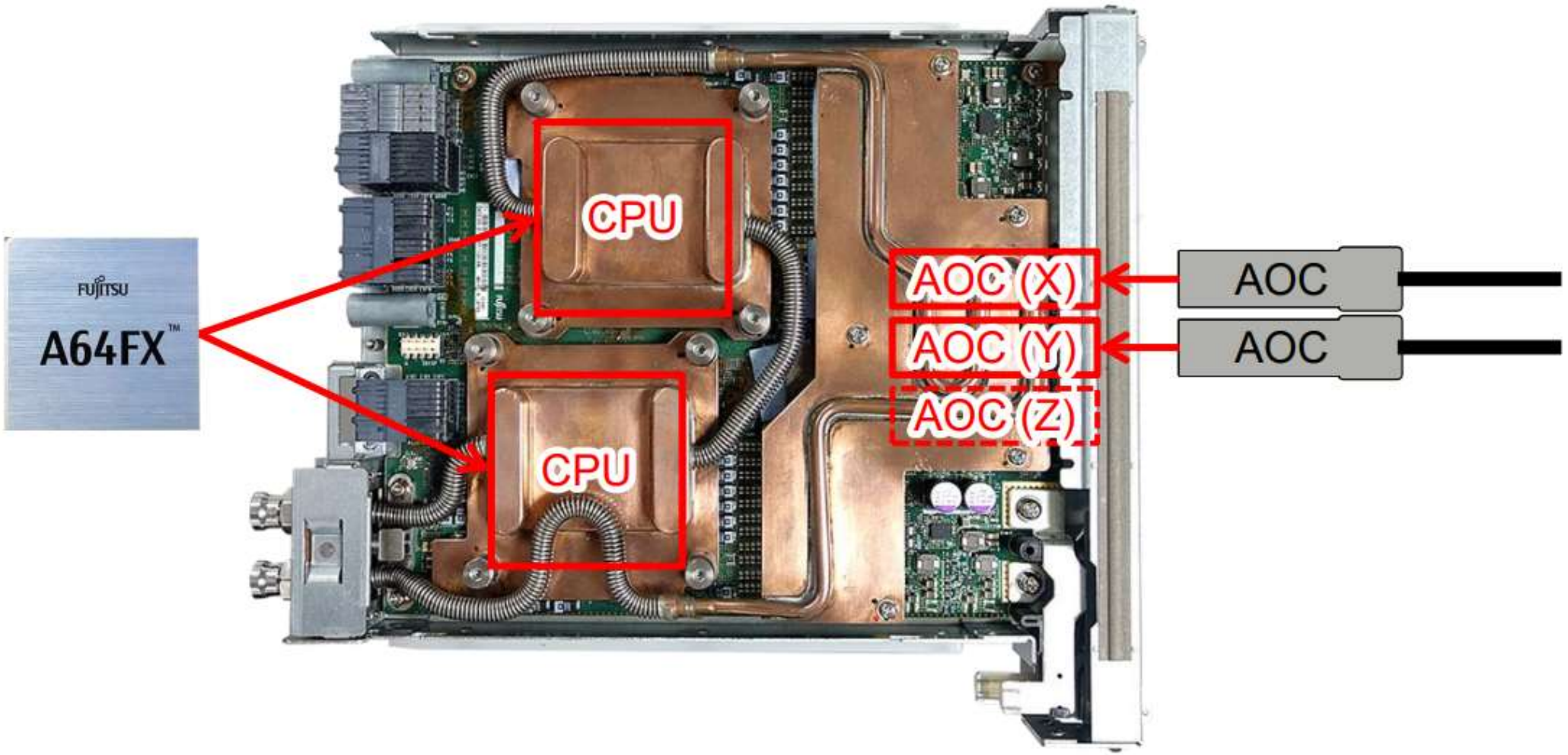


- Total topology is 6D mesh/torus
  - Cartesian product of *xyz* and *abc* mesh/torus

Topologías de red en HPC

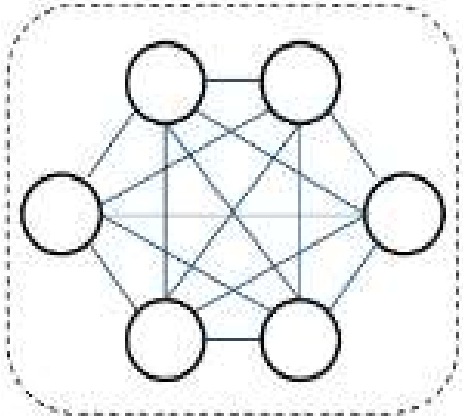


Topologías de red en HPC



Topologías de red en HPC

- ***Dragonfly+***:



- **Proporciona un buen rendimiento para gran variedad de patrones de comunicación.**
- **Reduce los costes de red en comparación con otras topologías, al reducir la cantidad de enlaces largos.**
- **Proporciona un rendimiento conocido en caso peor para el mismo número de enlaces globales entre grupos y permite una mejor utilización del *búfer* de conmutación.**
- **Admite múltiples rutas desde el conmutador de entrada al conmutador de salida y, por lo tanto, ofrece el mayor rendimiento de datos.**
- **El clúster escala sin necesidad de recablear enlaces largos.**