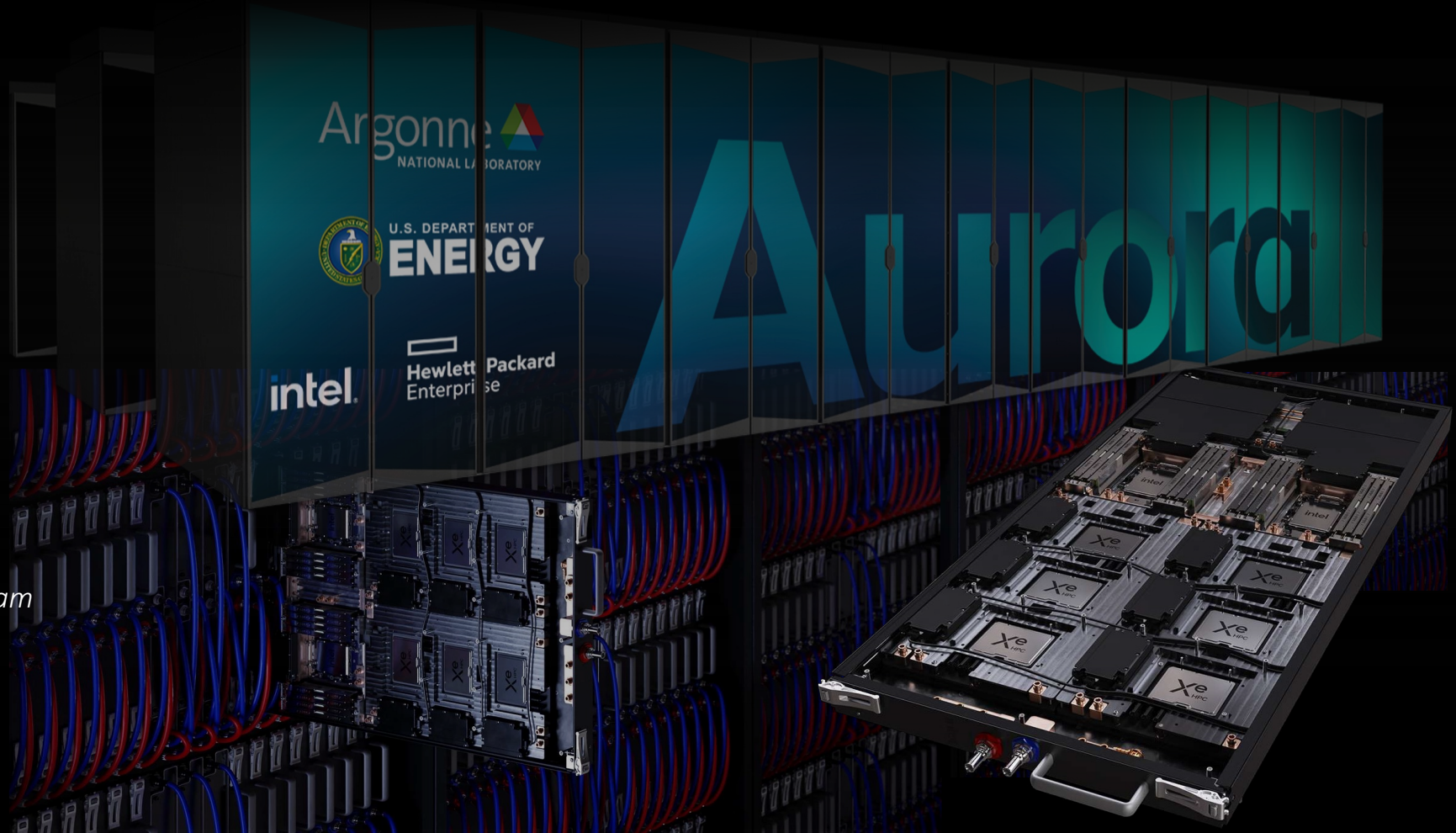
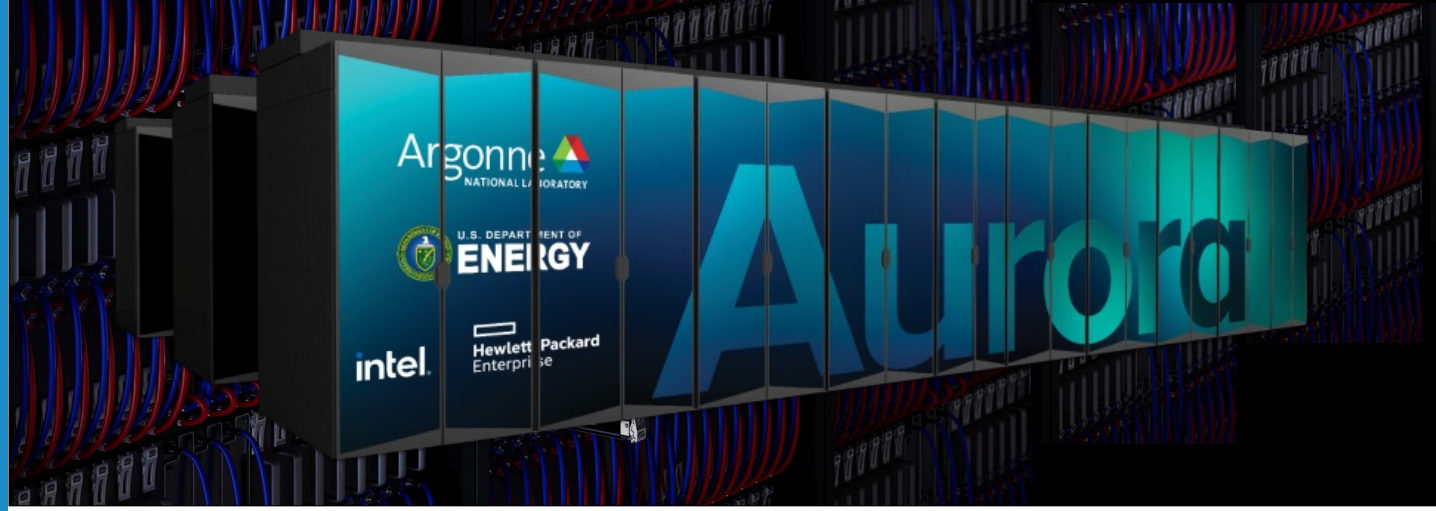
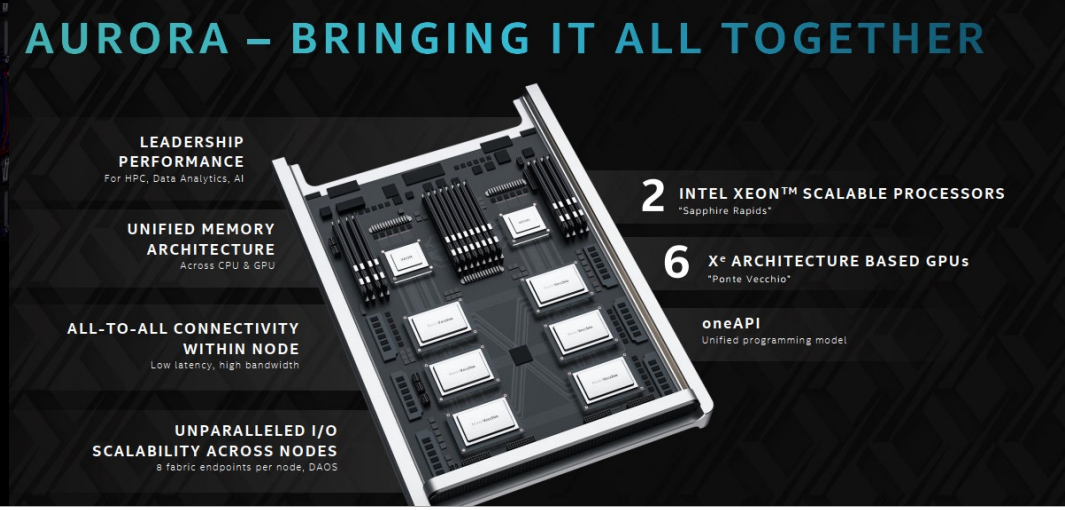
Overview on Aurora Exascale Compute Blade



Servesh Muralidharan
servesh@anl.gov
Computer Scientist, Performance Engineering Team
Argonne Leadership Computing Facility





Peak Performance
≧ 2 Exaflops DP

Intel GPU
Intel® Data Center GPU Max Series 1550

Intel Xeon Processor
Intel® Xeon Max Series 9470C CPU with High Bandwidth Memory

Platform
HPE Cray-Ex

Compute Node
2x Intel® Xeon Max Series processors
6x Intel® Data Center GPU Max Series
8x Slingshot11 fabric endpoints

GPU Architecture
Intel XeHPC architecture
High Bandwidth Memory

Node Performance
>130 TF

System Size
166 Cabinets
10,624 Nodes
21,248 CPUs
63,744 GPUs

System Memory
1.36PB HBM CPU Capacity
10.9PB DDR5 Capacity
8.16PB HBM GPU Capacity

System Memory Bandwidth
30.58PB/s Peak HBM BW CPU
5.95PB/s Peak DDR5 BW
208.9PB/s Peak HBM BW GPU

High-Performance Storage
230PB
31TB/s DAOS bandwidth
1024 DAOS Nodes

System Interconnect
HPE Slingshot 11
Dragonfly topology with adaptive routing

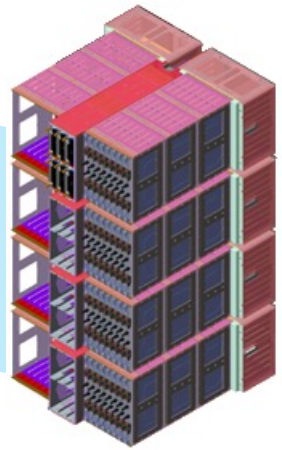
System Interconnect BW
Peak Injection BW 2.12PB/s
Peak Bisection BW 0.69PB/s

Network Switch
25.6 Tb/s per switch (64x 200 Gb/s ports)
Links with 25 GB/s per direction

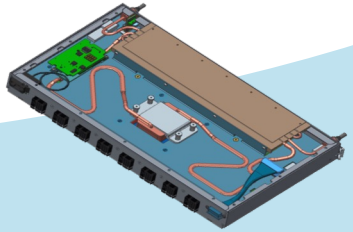
Programming Environment
- C/C++, Fortran
- SYCL/DPC++
- OpenMP 5.0
- Kokkos, RAJA

Aurora High-level System Overview

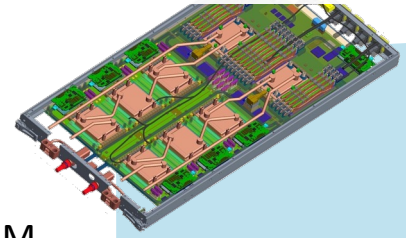
- Aurora compute cluster is organized into groups of blades, chassis and racks
- 6x GPUs, 2x CPUs, 16x DDR DIMMs and 8x NICs are densely packed into an Exascale Compute Blade (ECB)
- 8 compute blades and 4 network switches form a chassis
- 8 chassis composed of 64 compute blades and 32 switch blades form a rack
- 166 racks form the entire compute cluster



COMPUTE RACK
64 Compute blades
32 Switch blades
GPU: 49.1 TB HBM
CPU: 8.2 TB HBM, 64 TB DDR5

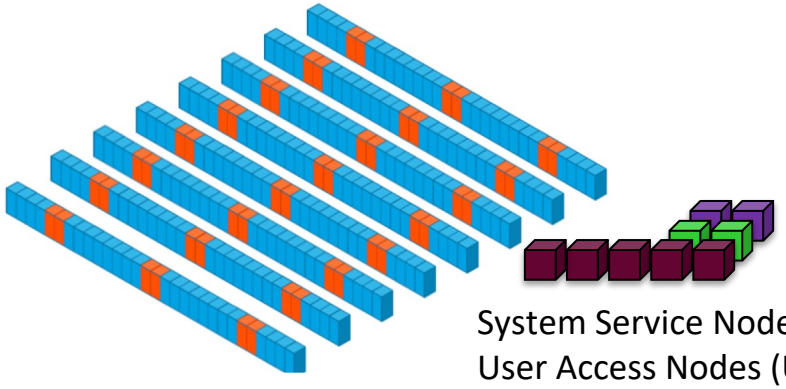


SWITCH BLADE
1 Slingshot switch
64 ports
Dragonfly topology



COMPUTE BLADE
2x Intel Xeon Max Series w HBM
6x Intel Data Center GPU Max Series
GPU: 768 GB HBM
CPU: 128 GB HBM, 1024 GB DDR5

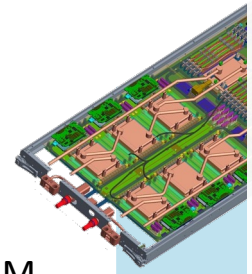
Aurora High-level System Overview



System Service Nodes (SSNs)
User Access Nodes (UANs)
DAOS Nodes (DNs)
Gateway Nodes (GNs)
IOF service, scalable library loading
DAOS <-> Lustre data mover

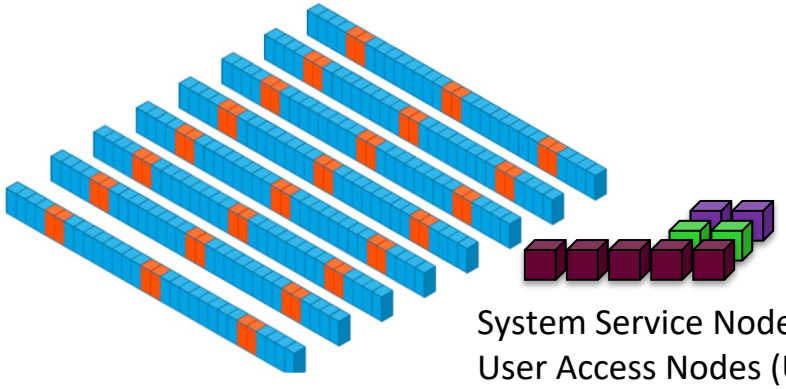
AURORA SYSTEM
166 Compute racks
10,624 Nodes
GPU: 8.16 PB HBM
CPU: 1.36 PB HBM, 10.9 PB DDR5
DAOS: 64 racks, 1024 nodes
230 PB (usable), 31 TB/s

- Set of user facing login nodes referred to as User Access Nodes (UANs) act as the interface to the compute cluster
- Compute applications are launched through the PBS job scheduler from the User Access Nodes
- DAOS nodes provide persistent storage
- System Service Nodes and Gateway Nodes are backend nodes that provide system services and connectivity for the operation of the cluster



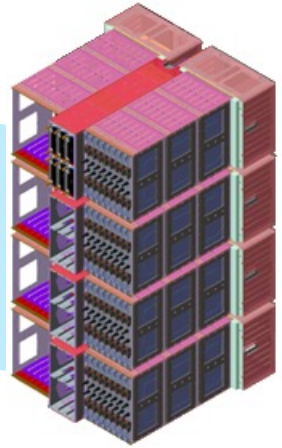
COMPUTE BLADE
2x Intel Xeon Max Series w HBM
6x Intel Data Center GPU Max Series
GPU: 768 GB HBM
CPU: 128 GB HBM, 1024 GB DDR5

Aurora High-level System Overview

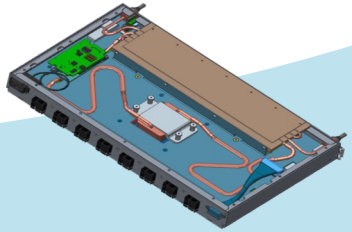


System Service Nodes (SSNs)
User Access Nodes (UANs)
DAOS Nodes (DNs)
Gateway Nodes (GNs)
IOF service, scalable library loading
DAOS <-> Lustre data mover

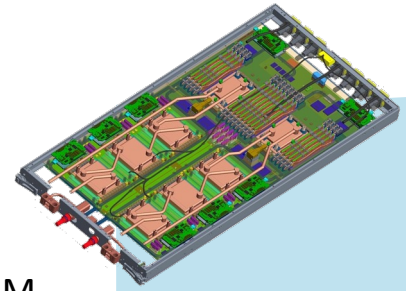
AURORA SYSTEM
166 Compute racks
10,624 Nodes
GPU: 8.16 PB HBM
CPU: 1.36 PB HBM, 10.9 PB DDR5
DAOS: 64 racks, 1024 nodes
230 PB (usable), 31 TB/s



COMPUTE RACK
64 Compute blades
32 Switch blades
GPU: 49.1 TB HBM
CPU: 8.2 TB HBM, 64 TB DDR5



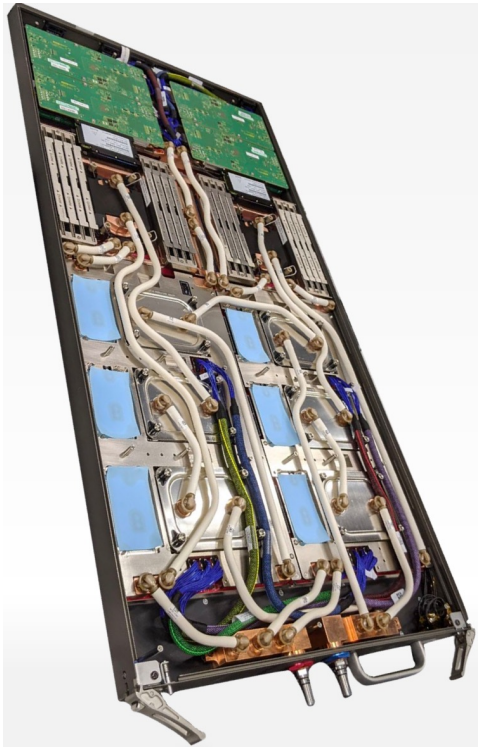
SWITCH BLADE
1 Slingshot switch
64 ports
Dragonfly topology

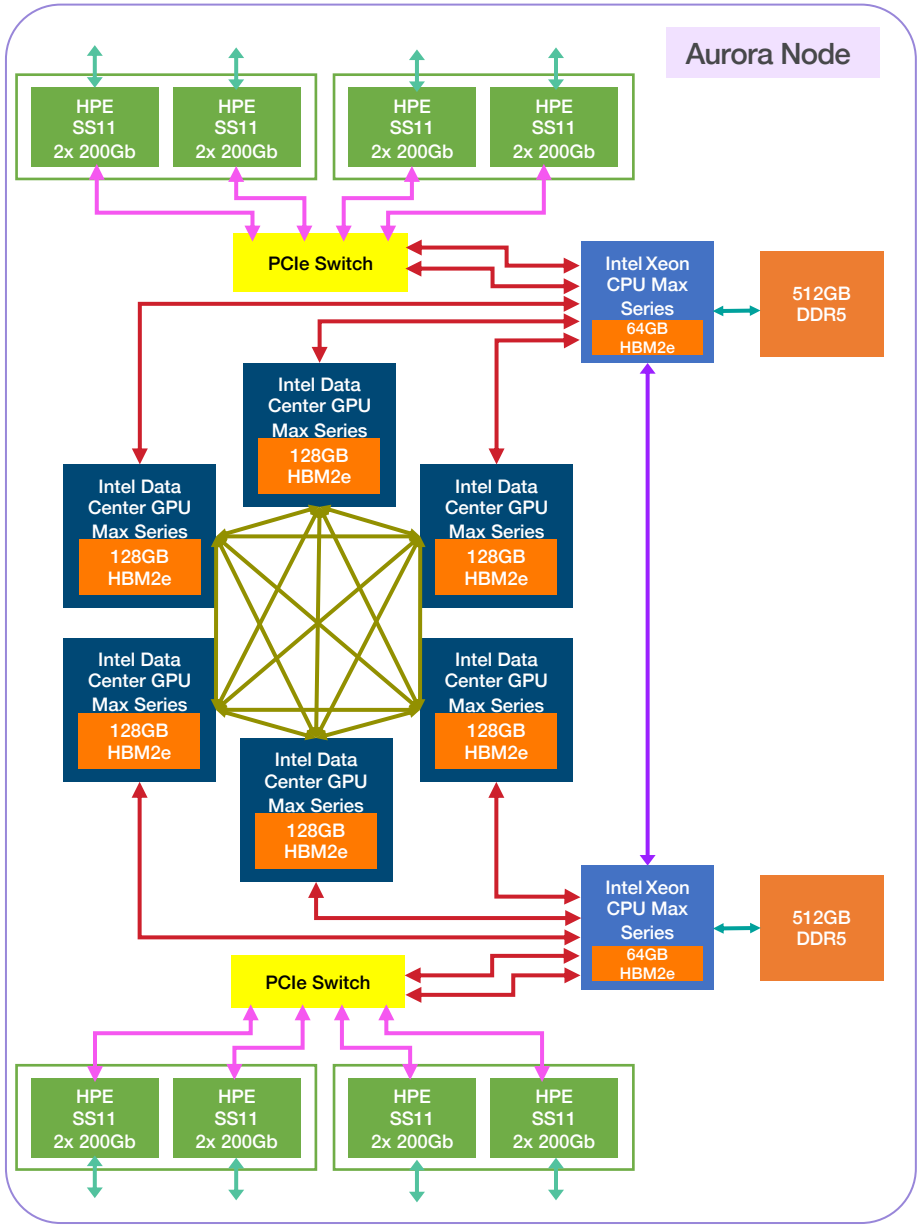


COMPUTE BLADE
2x Intel Xeon Max Series w HBM
6x Intel Data Center GPU Max Series
GPU: 768 GB HBM
CPU: 128 GB HBM, 1024 GB DDR5

Aurora Exascale Compute Blade

NODE CHARACTERISTICS	
6	GPU - Intel Data Center GPU Max Series (#)
2	CPU - Intel Xeon CPU Max Series (#)
768	GPU HBM Memory (GB)
19.66	Peak GPU HBM BW (TB/s)
128	CPU HBM Memory (GB)
2.87	Peak CPU HBM BW (TB/s)
1024	CPU DDR5 Memory (GB)
0.56	Peak CPU DDR5 BW (TB/s)
≧ 130	Peak Node DP FLOPS (TF)
200	Max Fabric Injection (GB/s)
8	NICs (#)





Network Switch

Consistent, Repeatable Application Performance

- Advanced congestion control
- Fine grained adaptive routing
- Very low average and tail latency

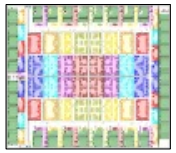
Extremely Scalable RDMA Performance

- Connectionless protocol
- Fine grained flow control
- MPI HW tag matching & progress engine
- Dragonfly topology – 3 switch hops (typical)

Native Ethernet

- Native IP – no encapsulation
- High-scale bandwidth integration to campus

HPE Slingshot Switches - 64 ports @ 200 Gbps



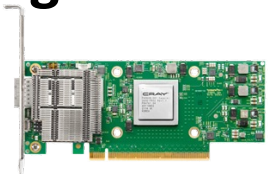
HPE Switch ASIC

Rack switches

100% DLC Switches

HPE Slingshot NICs - 200 Gbps

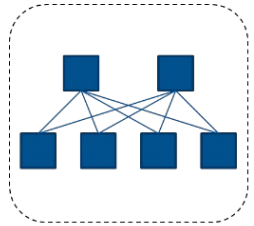
HPE NIC ASIC



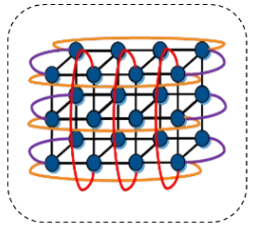
PCIe Adapters

100% DLC NIC Mezz

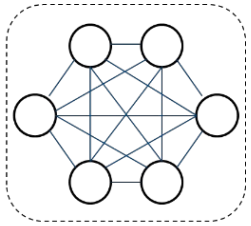
Interconnect Topology



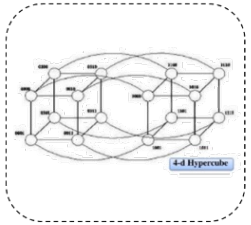
Fat Tree



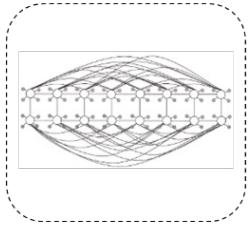
Torus



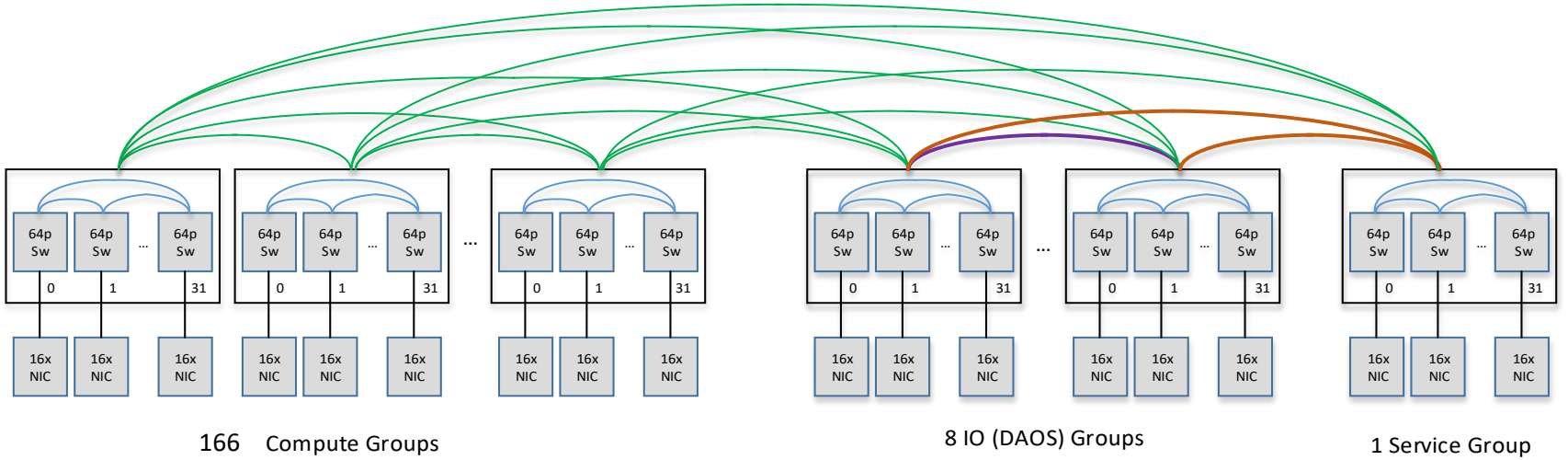
Dragonfly



Hypercube



HyperX



- 1-D Dragonfly Topology - 175 total groups (166 compute + 8 IO + 1 Service),
- All the global links are optical, all the local links in compute groups are electrical
- 2 global links between any two compute groups
- 24 links between any two IO groups, 8 links between the Service group and each IO group
- Total injection bandwidth: 2.12PB/s
- Total bisection bandwidth: 0.69PB/s

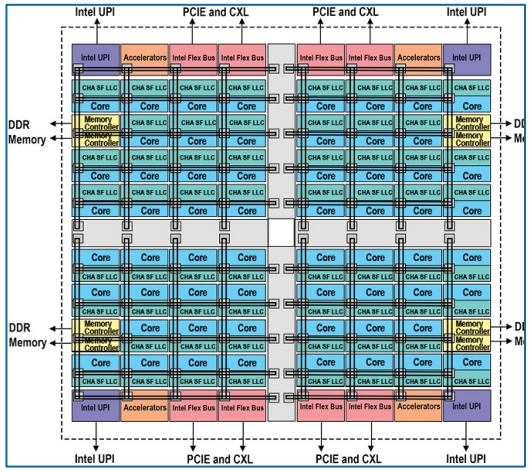
Aurora Exascale Compute Blade

Compute and Data Flow

Intel Xeon Max Series CPU w HBM

- Dual socket
- 52 cores
- First Level Cache: 32 KB Instruction Cache
  48 KB Data Cache
- Mid-Level Cache: 2 MB private per core
- Last Level Cache: 1.875 MB per core
- 8 channels DDR5 @ 4400MT/s
- 1TB DDR5 Memory
- 64GB HBM2e per socket
- 80 PCIe lanes with PCIe Gen 5.0 support
  - PCIe bifurcation support: x16, x8, x4, x2(Gen4)

https://www.hc33.hotchips.org/assets/program/conference/day1/HC2021.C1.4%20Intel%20Arijit.pdf

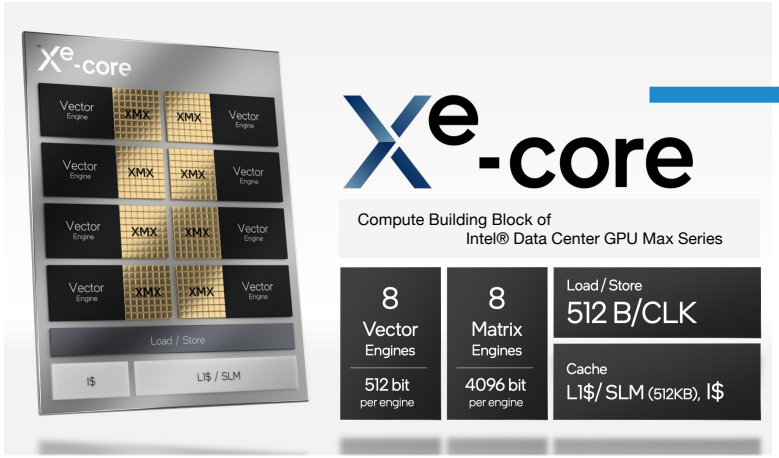


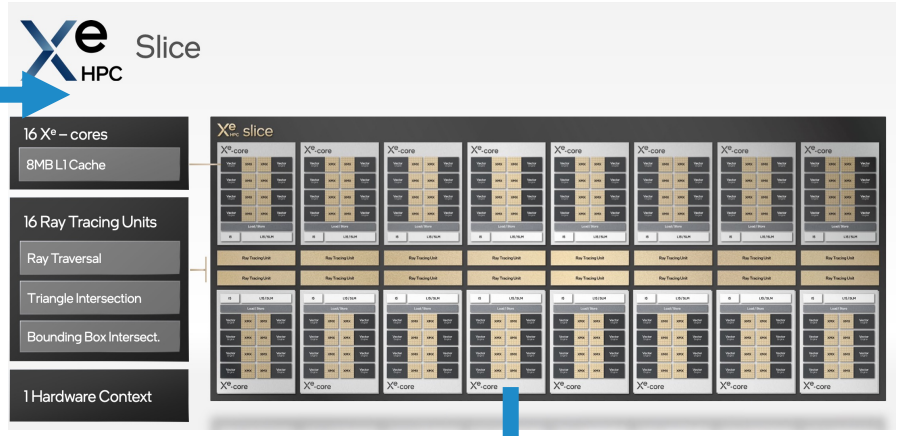
https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html

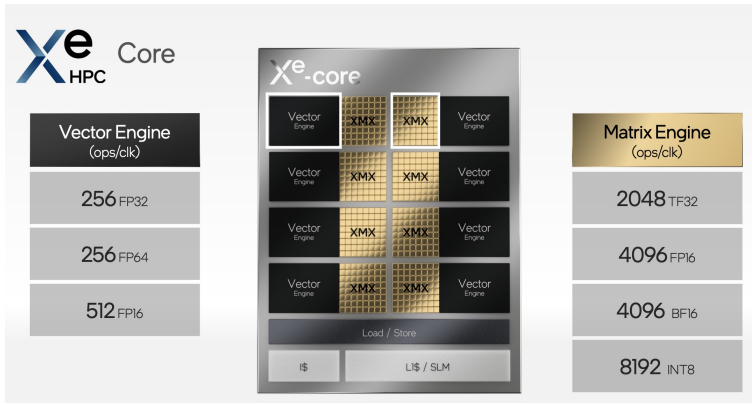
Argonne NATIONAL LABORATORY

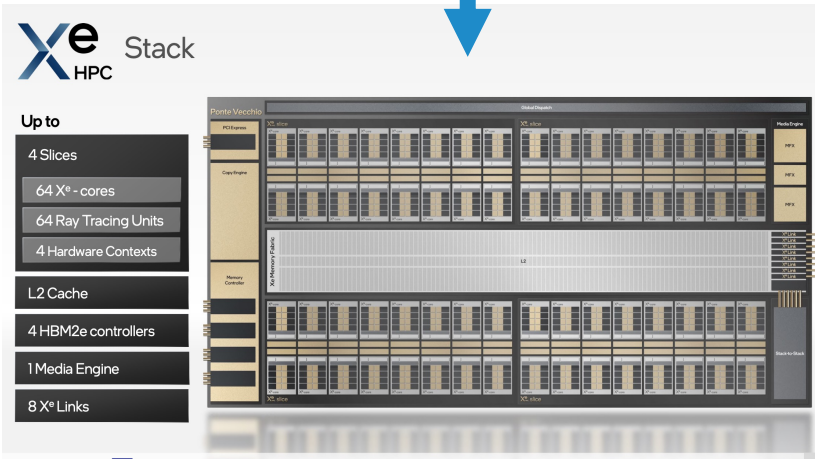
Intel Data Center GPU Max Series Architectural Components

- Xe Cores
  - Vector Engine
    - Traditional compute pipeline
  - Matrix Engine
    - Low precision systolic pipeline
  - L1 Data Cache
    - Shared Local Memory
  - Instruction Cache
- Xe Slice
  - Hardware Context
  - Offload Units
- Xe Stack
  - LLC
  - HBM2e controllers
  - Xe link
  - Cache Memory Fabric
  - PCIe Endpoint
  - Hardware specific engines
  - Stack to Stack Interconnect
  - Xe links
    - Multi GPU Interconnect







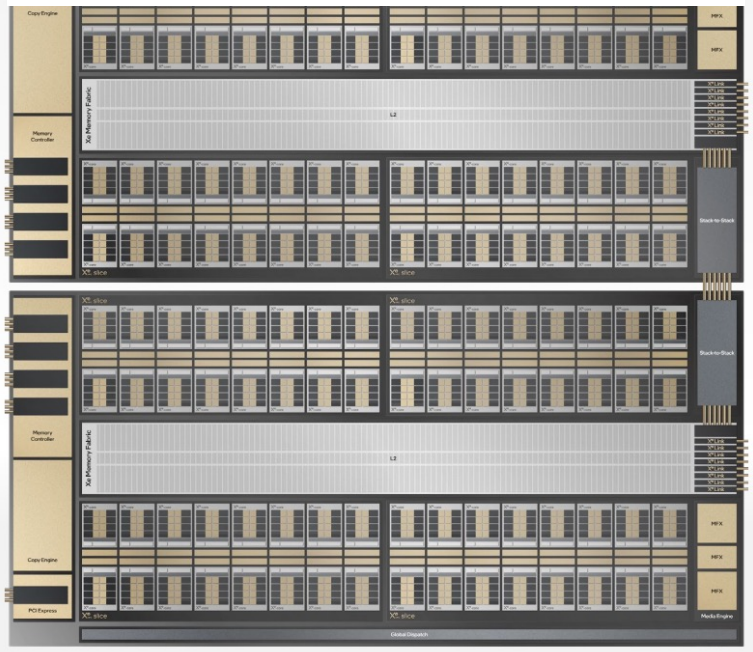


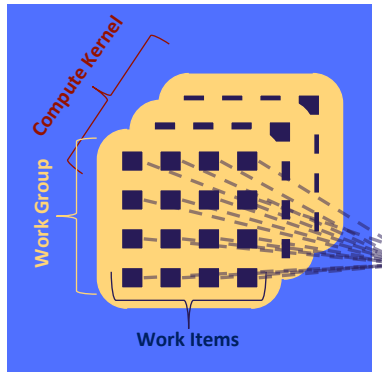
https://hc33.hotchips.org/assets/program/conference/day2/hc2021_pvc_final.pdf

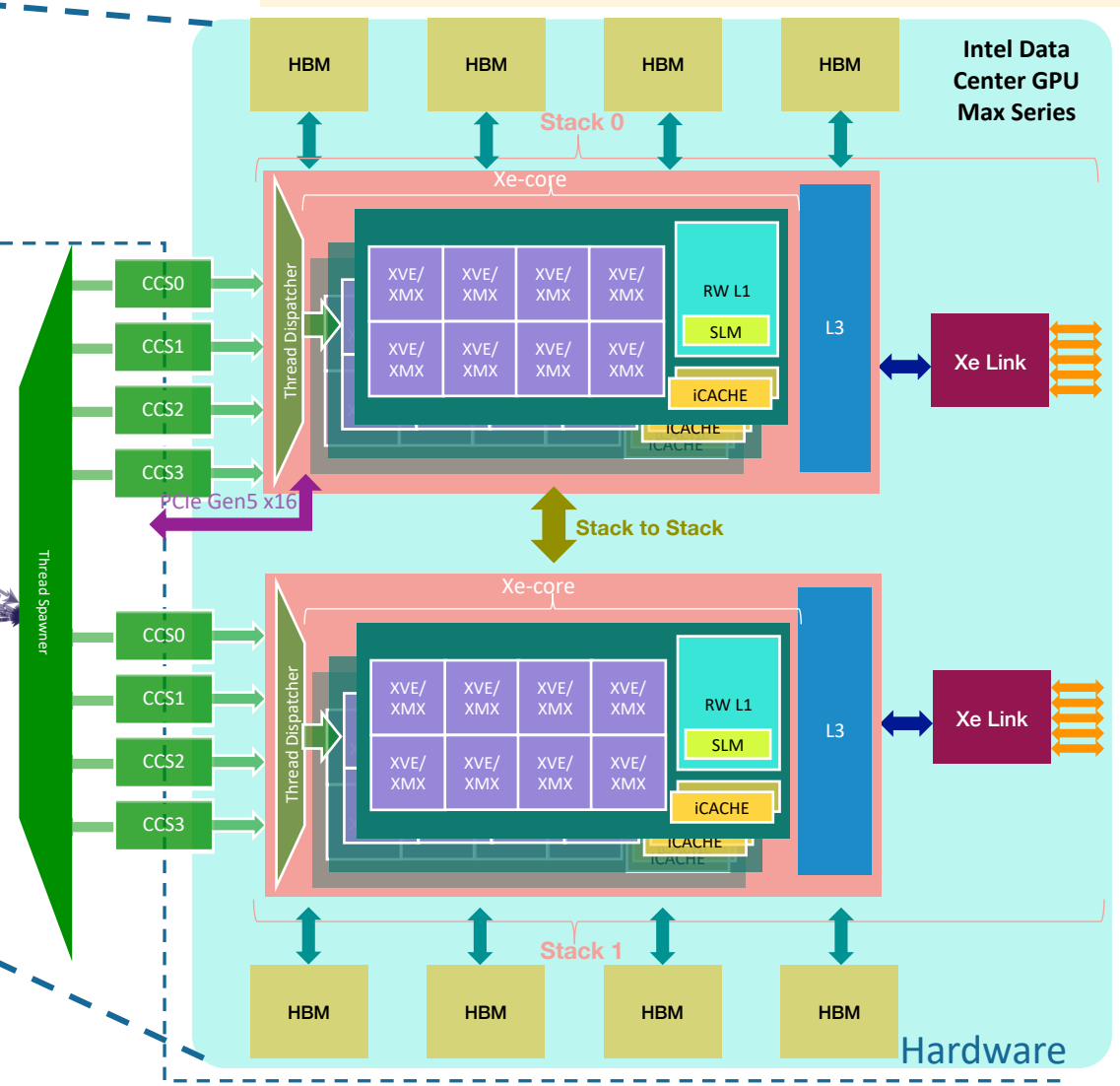
GPU Compute Execution

XVE – Xe Vector Engine
GRF – General Register File
SLM – Shared Local Memory
RW L1 – Read/Write L1
HBM – High Bandwidth Memory
iCACHE – Instruction Cache
CCS – Compute Command Streamer
SIMD – Single Instruction Multiple Data

https://www.intel.com/content/www/us/en/docs/oneapi/optimization-guide-gpu/2024-0/intel-xe-gpu-architecture.html



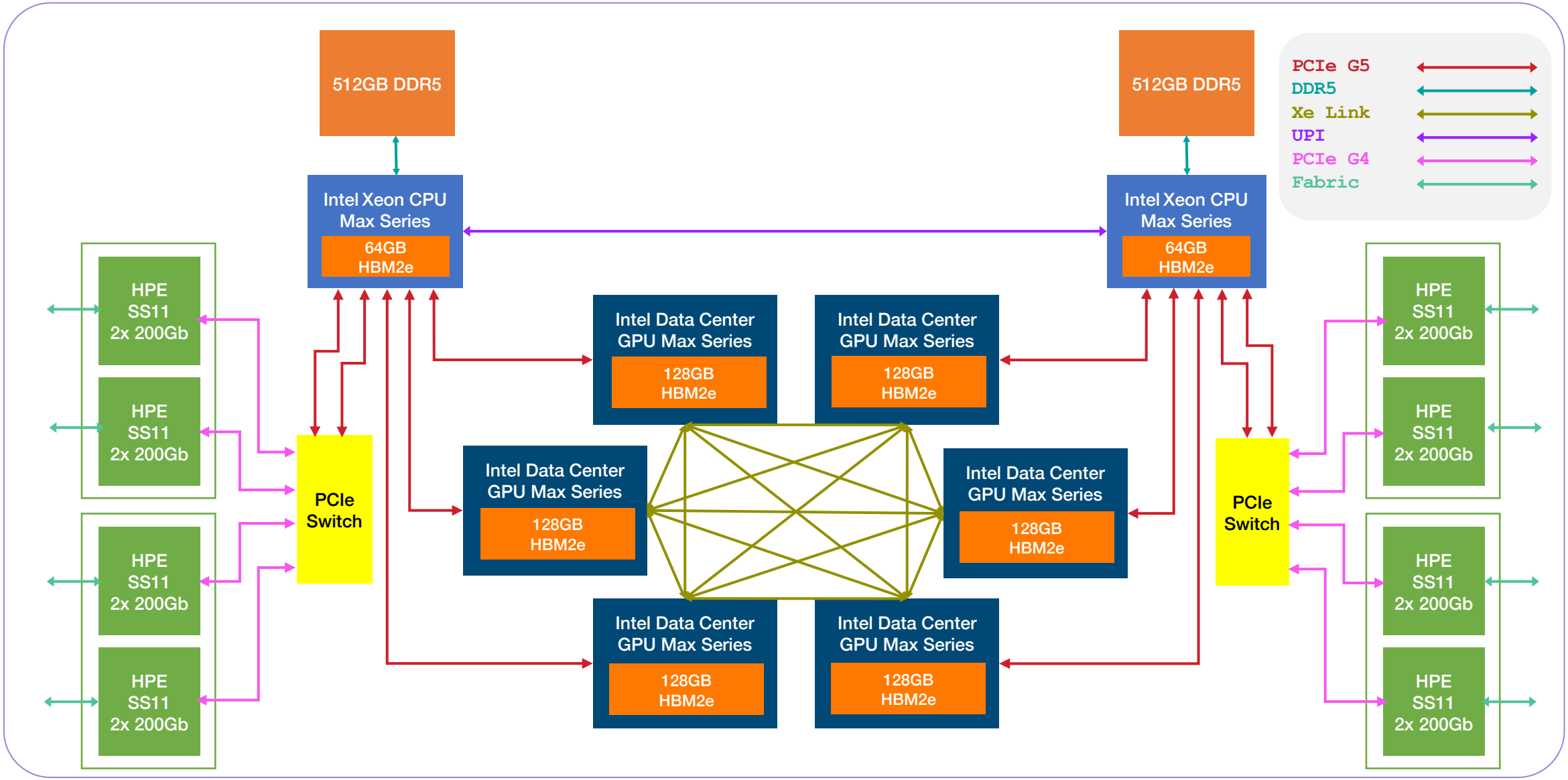




- Execution on the GPU starts with the allocation of memory and the compute kernel scheduled on the GPU
- The GPU threads are spawned and scheduled through the CCS
- Execution stops when the kernel hits the “end of thread” instruction
- Shared vs Device allocation implies different latencies for accessing the data
- GPU threads can switch when any of the stall condition occurs
  - However during execution threads cannot be interrupted

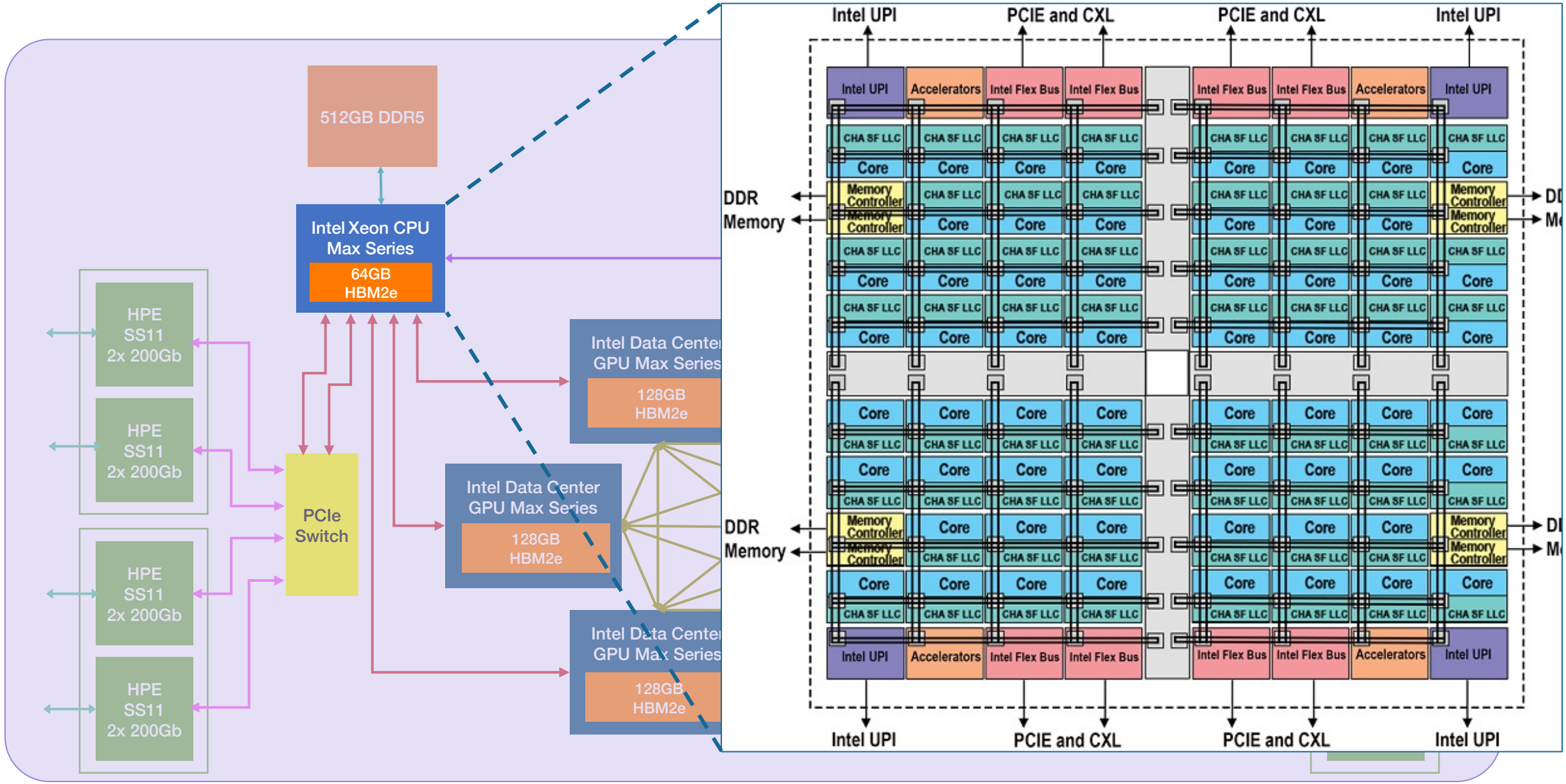
https://www.intel.com/content/www/us/en/docs/oneapi/optimization-guide-gpu/2024-0/execution-model-overview.html

Aurora Exascale Compute Blade – Data Flow



Intel Xeon Max Series CPU w HBM

https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html

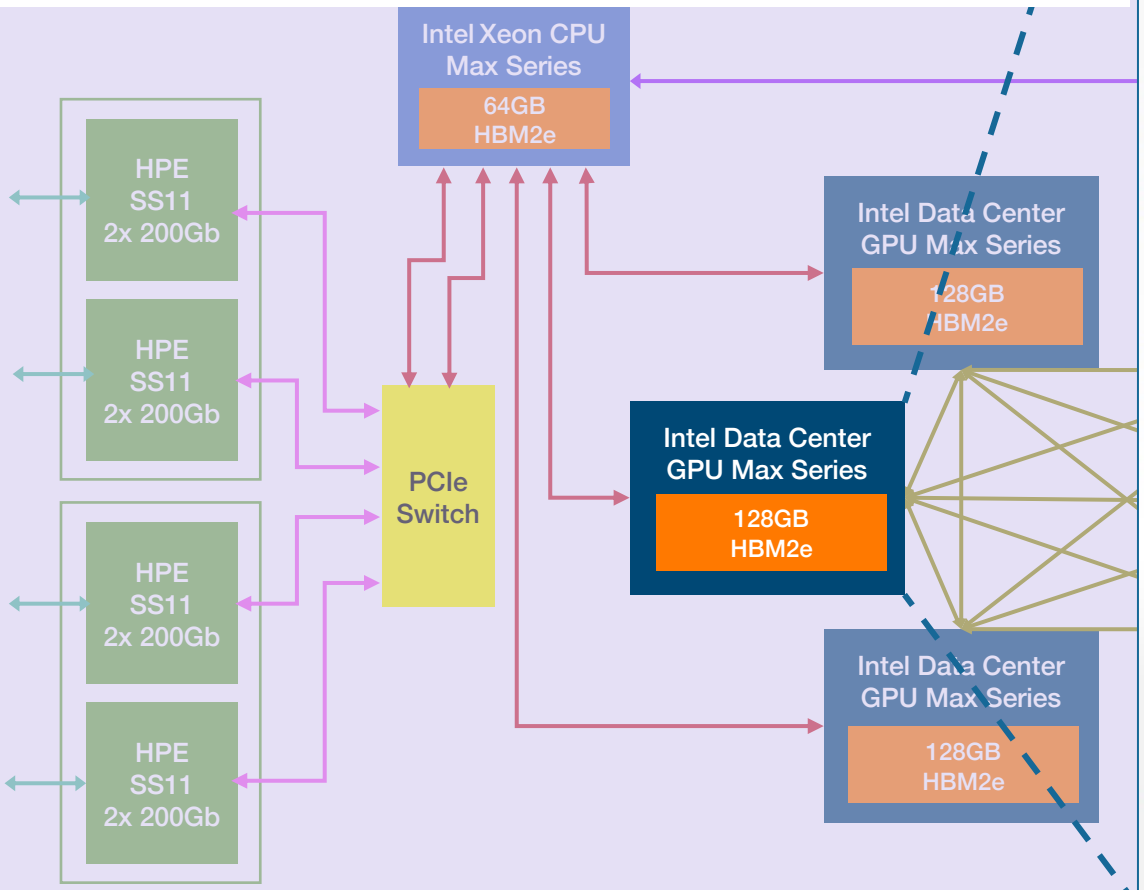


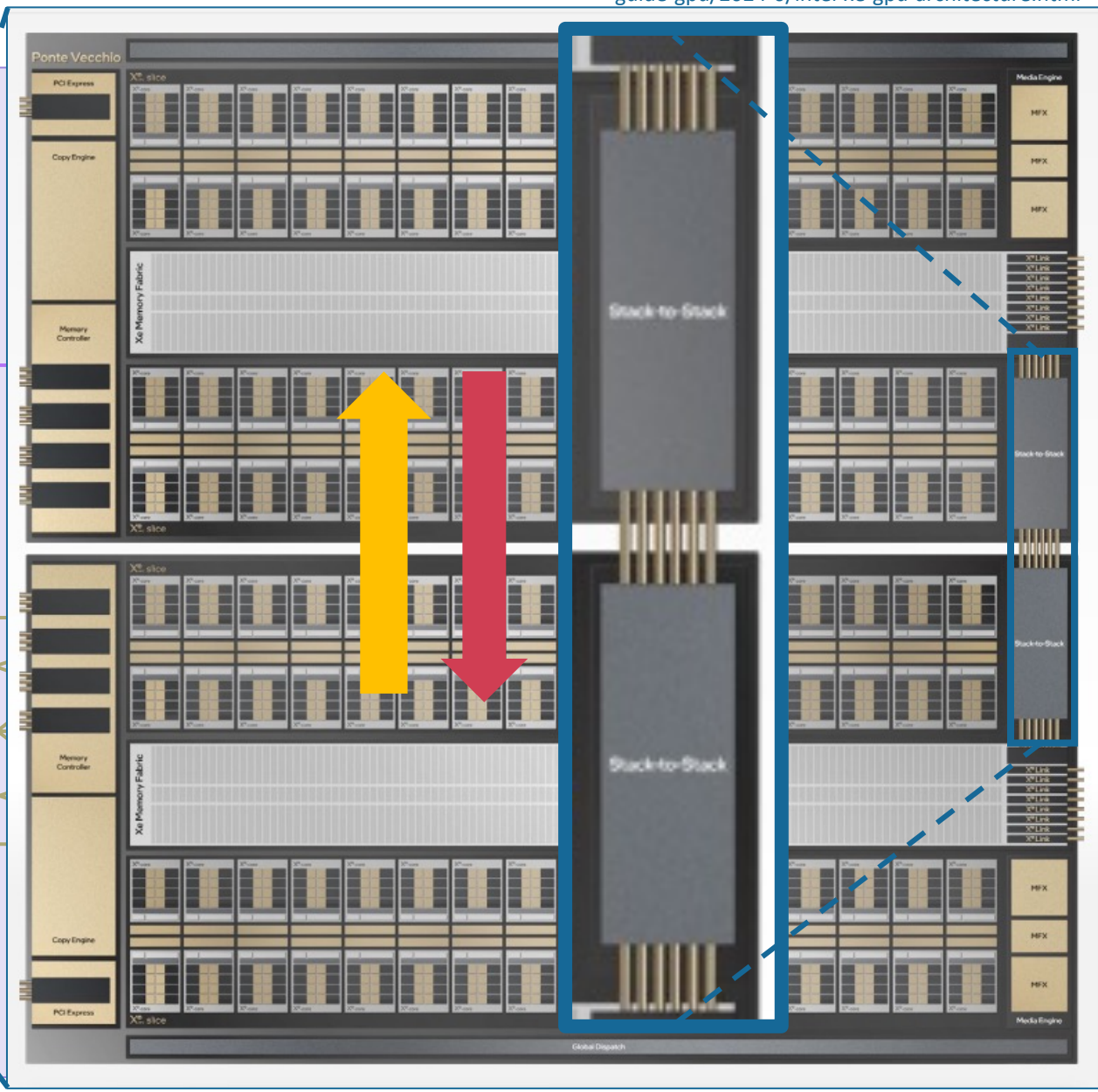
Argonne NATIONAL LABORATORY

Intel Data Center GPU

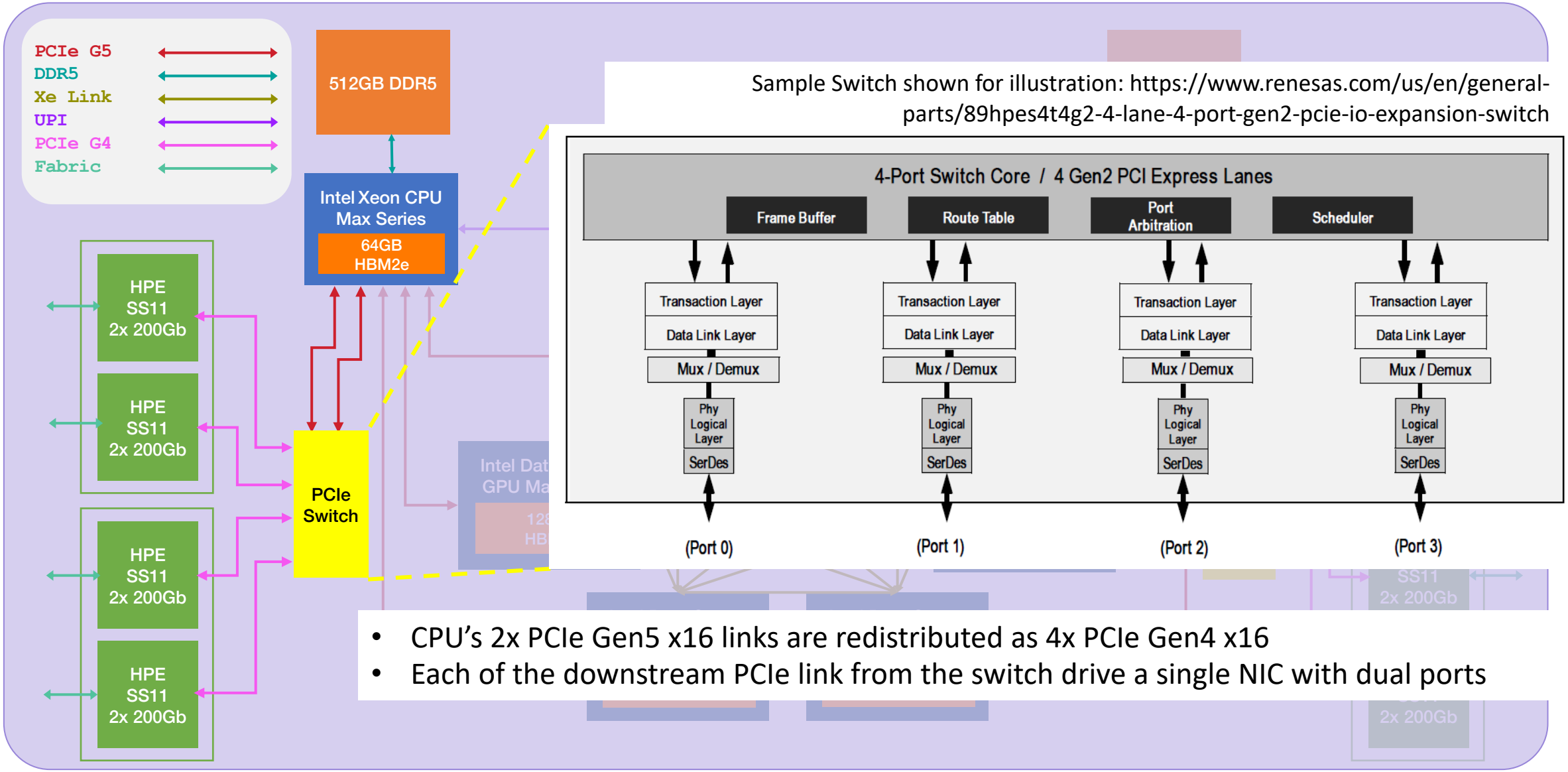
https://www.intel.com/content/www/us/en/docs/oneapi/optimization-guide-gpu/2024-0/intel-xe-gpu-architecture.html

- Each GPU is actually composed of dual stacks
- The PCIe endpoint is present in only one of the stack
- Data movement between stacks happens through stack to stack interconnect



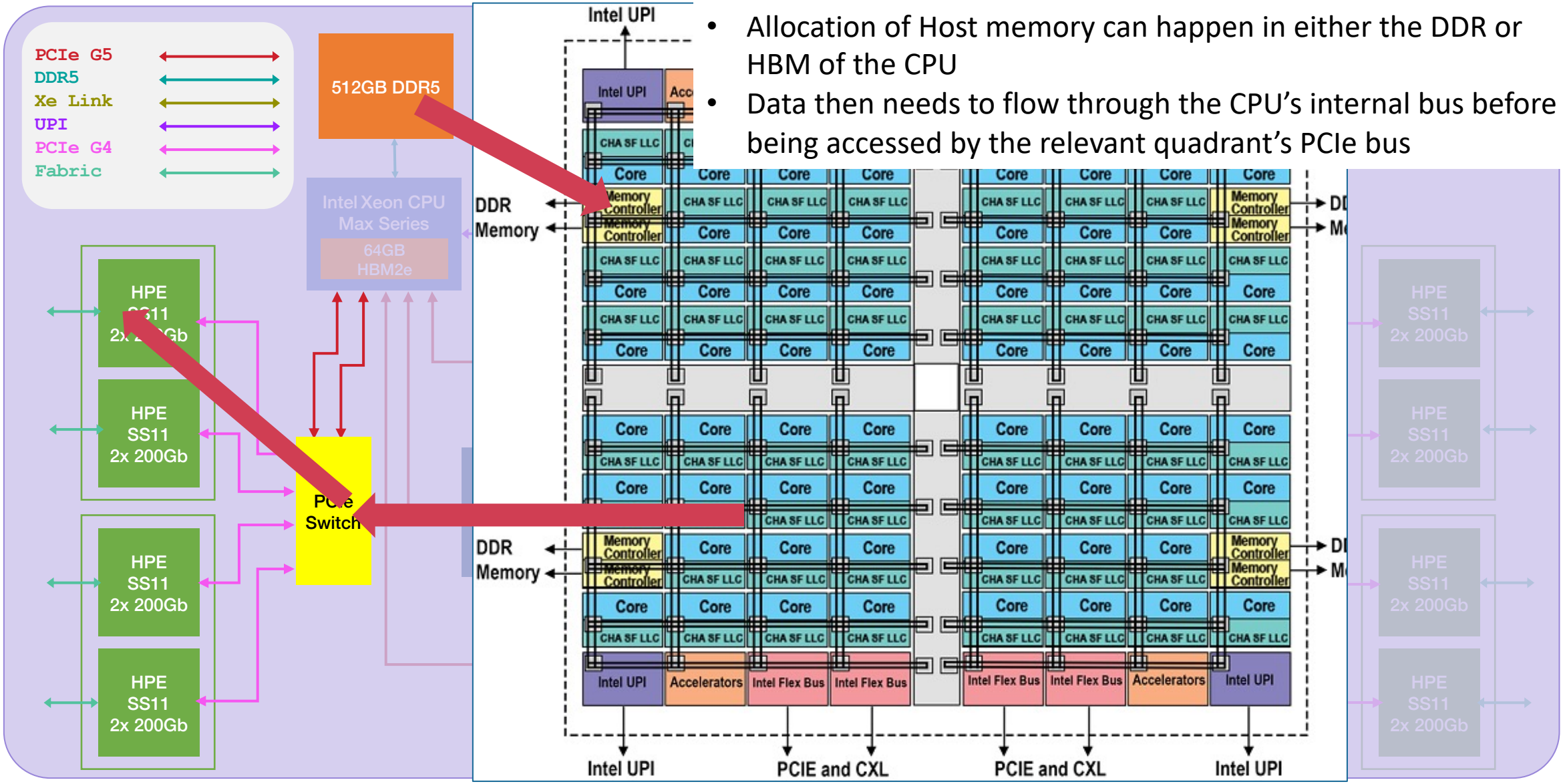


CPU – NIC PCIe Switch Interface



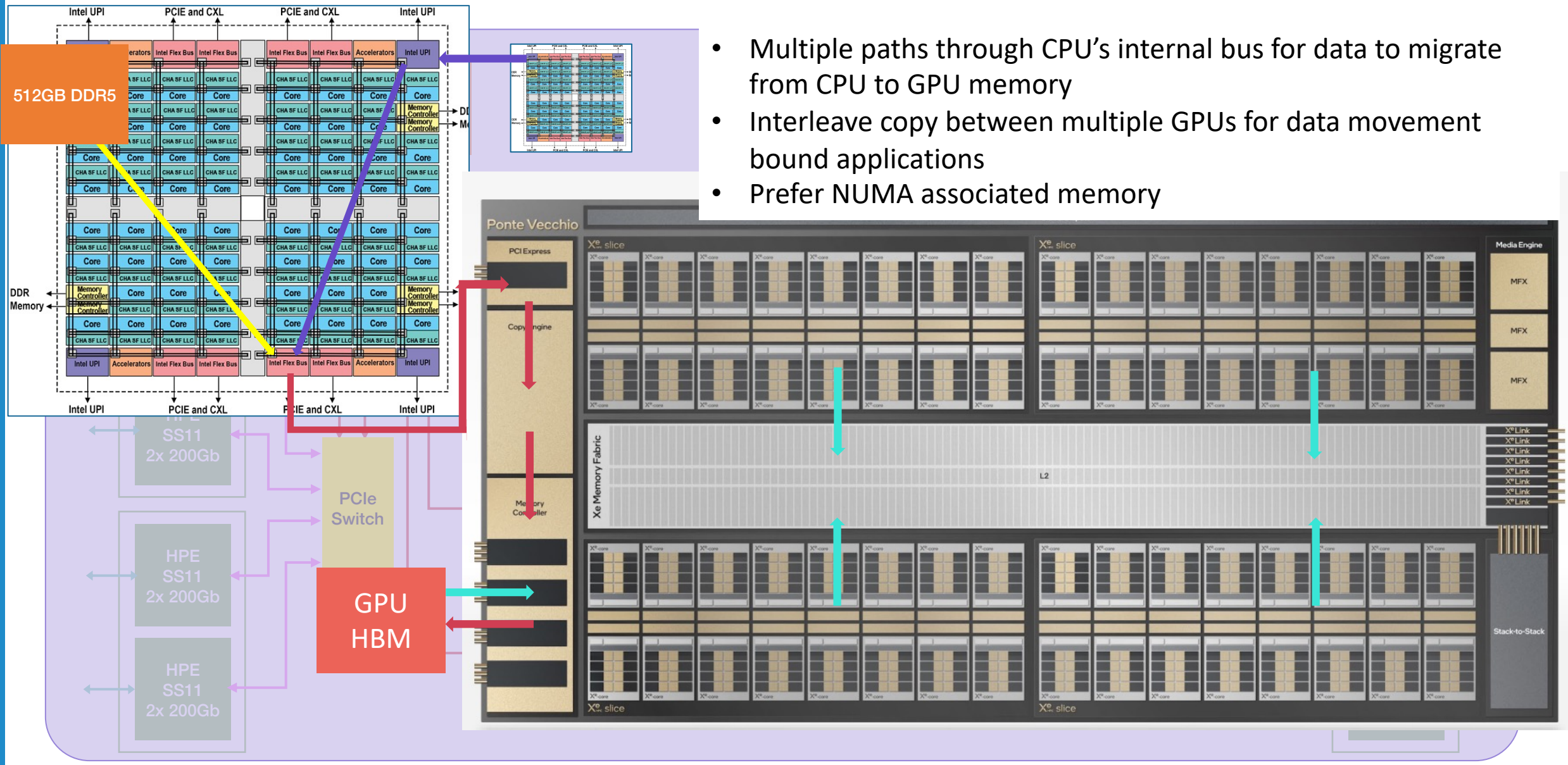
- CPU's 2x PCIe Gen5 x16 links are redistributed as 4x PCIe Gen4 x16
- Each of the downstream PCIe link from the switch drive a single NIC with dual ports

CPU – NIC Data Flow



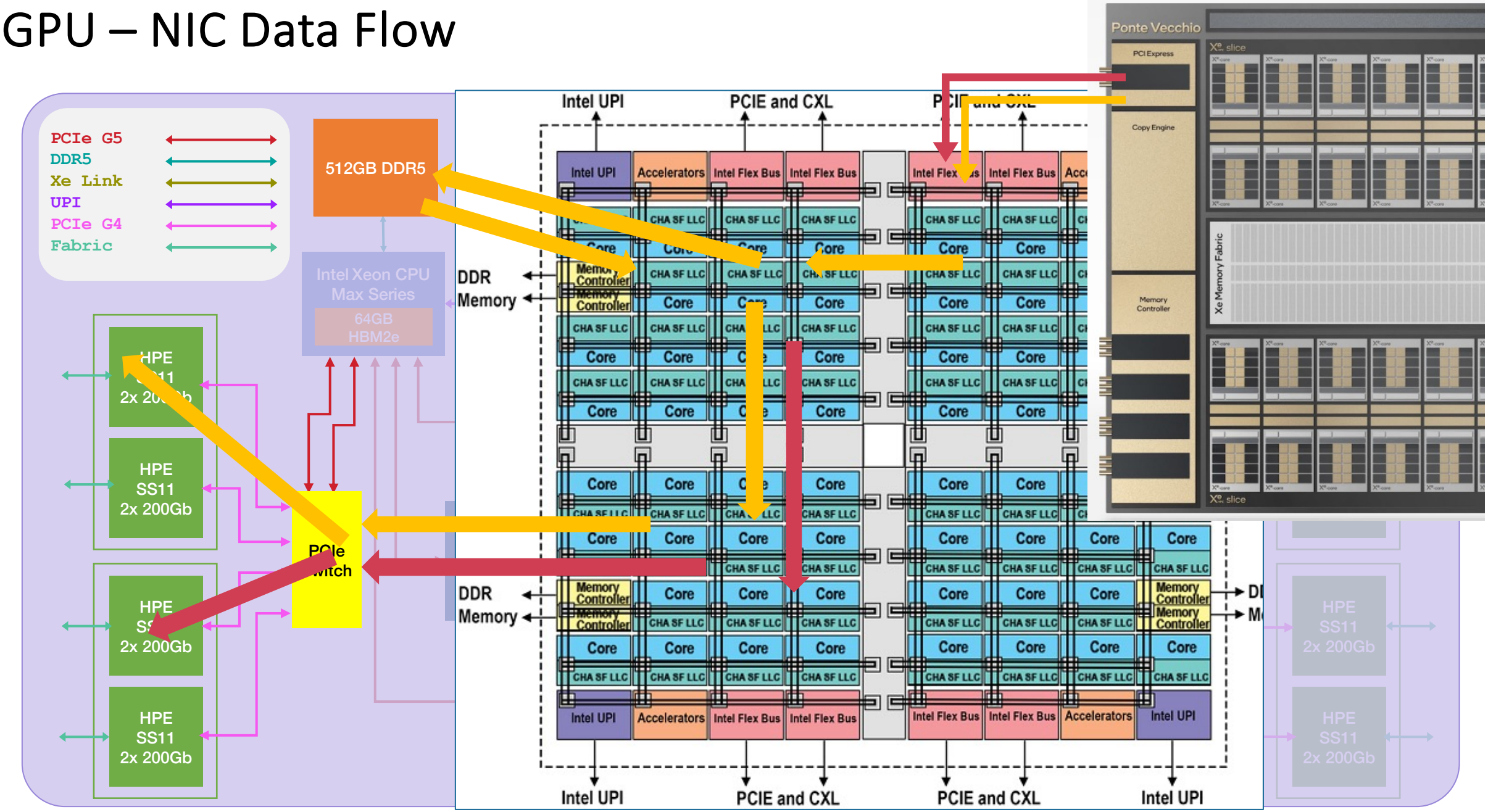
- Allocation of Host memory can happen in either the DDR or HBM of the CPU
- Data then needs to flow through the CPU's internal bus before being accessed by the relevant quadrant's PCIe bus

CPU to GPU Data Flow

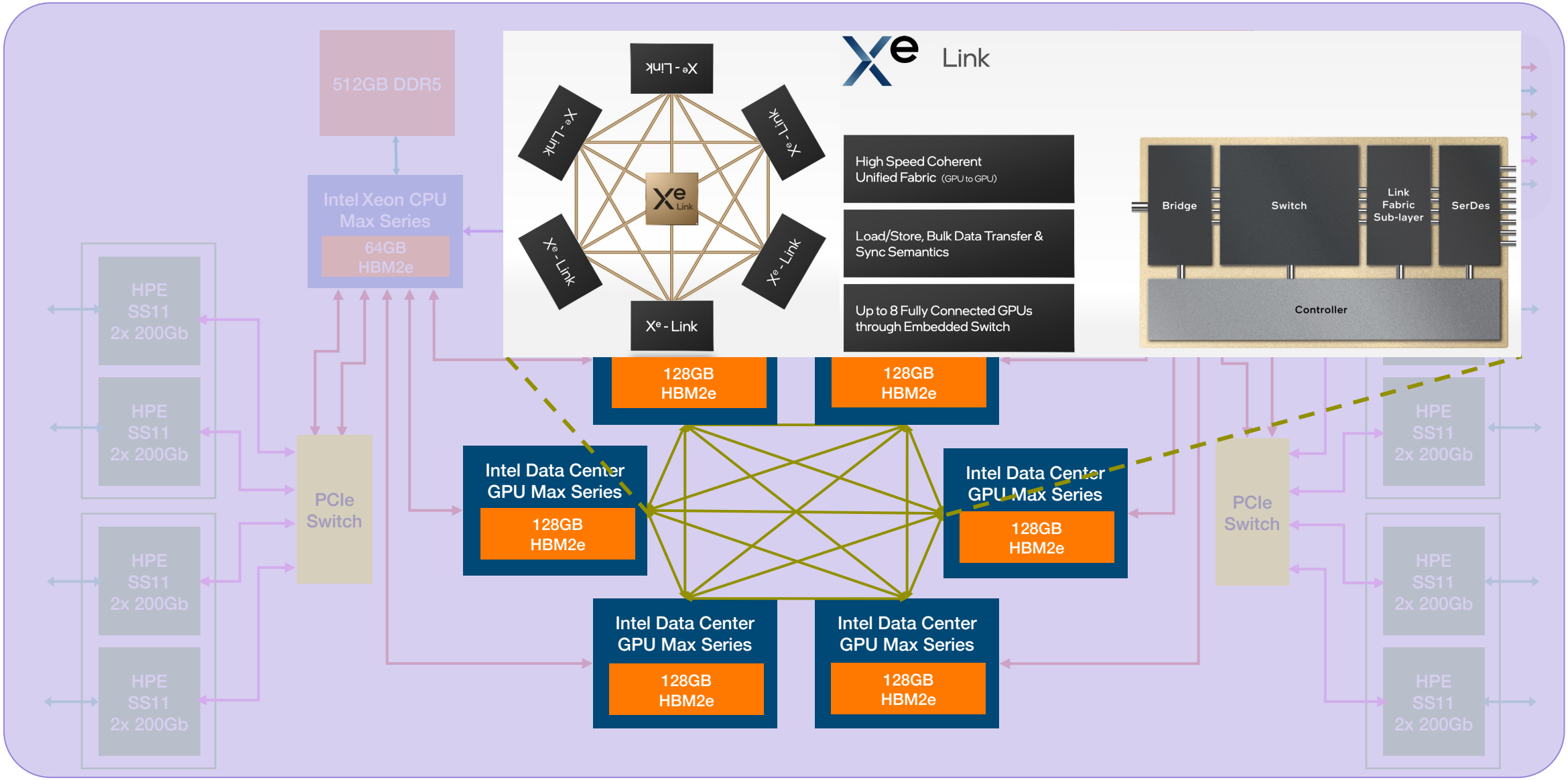


- Multiple paths through CPU's internal bus for data to migrate from CPU to GPU memory
- Interleave copy between multiple GPUs for data movement bound applications
- Prefer NUMA associated memory

GPU – NIC Data Flow

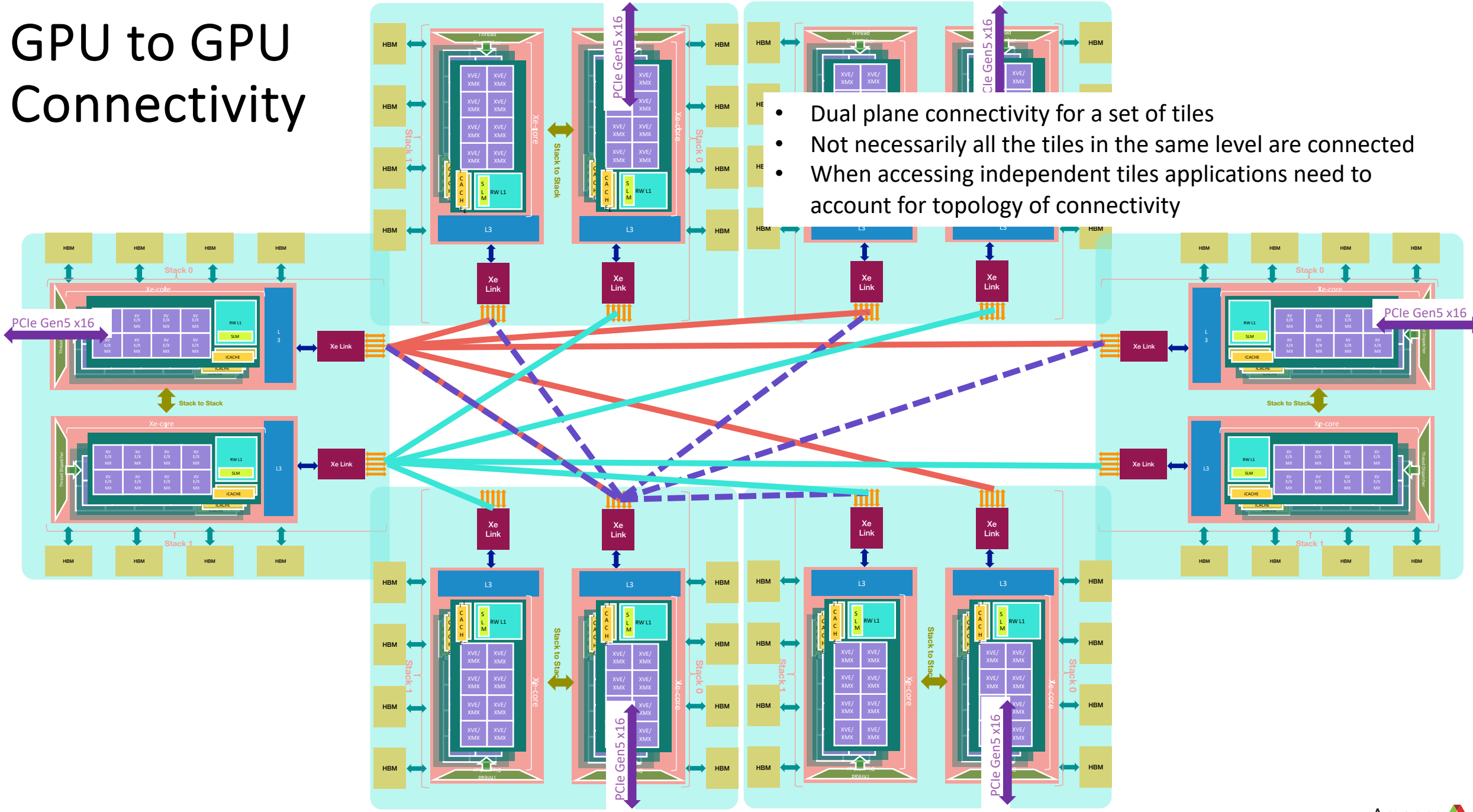


GPU to GPU Connectivity



GPU to GPU Connectivity

- Dual plane connectivity for a set of tiles
- Not necessarily all the tiles in the same level are connected
- When accessing independent tiles applications need to account for topology of connectivity



Conclusions

- Aurora's Exascale Compute Blade has an intricate data flow design
  - Multiple paths exist to move data between NIC<->CPU<->GPU
  - Optimize data flow to reduce bottlenecks
- Compute dominated by the complexity of scheduling work
- Overlap asynchronous communication across multiple devices

QUESTIONS?

SERVESH@ANL.GOV