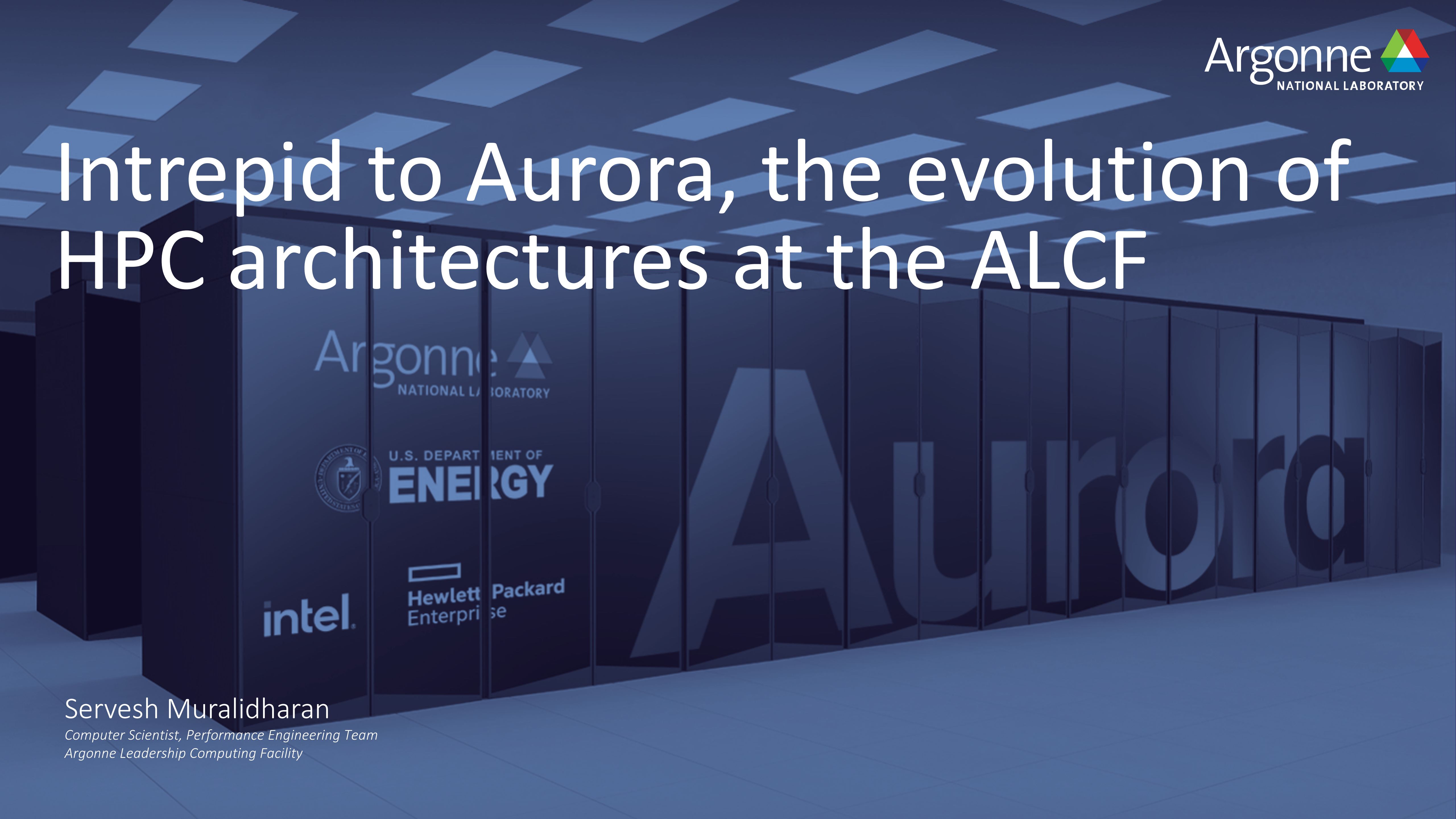

Argonne NATIONAL LABORATORY
U.S. DEPARTMENT OF ENERGY
intel
Hewlett Packard Enterprise
Aurora

Argonne NATIONAL LABORATORY

Intrepid to Aurora, the evolution of HPC architectures at the ALCF

Servesh Muralidharan

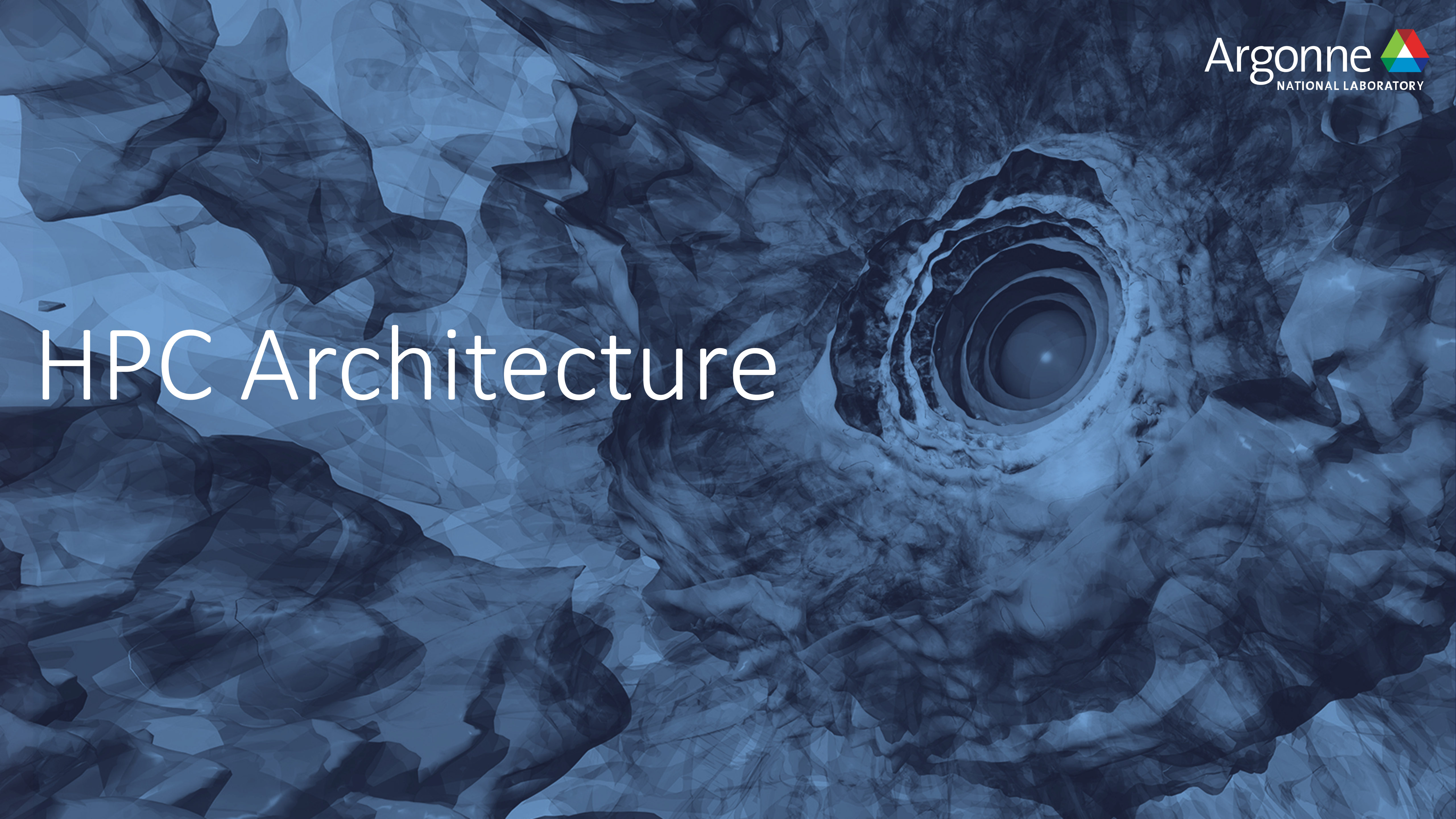
Computer Scientist, Performance Engineering Team
Argonne Leadership Computing Facility

Argonne Leadership Computing Facility

- The Argonne Leadership Computing Facility (ALCF) was established in 2006 as one of two DOE funded leadership computing facilities, along with the Oak Ridge LCF
  - Goal of the LCFs is to provide the computational science community with a leading-edge computing capability dedicated to breakthrough science and engineering
  - Typical have systems at or near the top of the Top 500 list
  - Allocations provided through open INCITE program
- Broader HPC landscape:
  - Other DOE funded facilities:
    - National Energy Research Scientific Computer Center (NERSC)
    - NNSA – Lawrence Livermore, Los Alamos, Sandia
  - Exascale Computing Project
  - National Science Foundation – XSEDE (TACC, PSC, SDSC, NCSA)
  - World wide: Japan, China, Europe

Argonne Leadership computing facility Resources

- 2008: Intrepid
  - — ALCF accepts 40 racks (160k cores) of Blue Gene/P (557 TF)
- 2012: Mira
  - — 48 racks of Blue Gene/Q (10 PF) in production at ALCF
- 2016: Theta
  - — ALCF accepts 12 PF Cray XC40 with Xeon Phi (KNL)
- 2021: Aurora
  - — One Exaflop Intel/Cray GPU machine to be delivered in 2021



HPC Architecture

Elements of a supercomputer

- **Processor** – architecturally optimized to balance complexity, cost, performance, and **power**
- Memory – generally commodity DDR, amount limited by cost
- Node – may contain multiple processors, memory, and network interface
- Network – optimized for latency, bandwidth, and cost
- IO System – complex array of disks, servers, and network
- Software Stack – compilers, libraries, tools, debuggers, …
- Control System – job launcher, system management

Processor performance

Many different approaches to increasing processor performance-

- Increase serial performance:
  - Increase clock speed
    - clock speed increases until around 2006 were enabled by Dennard scaling
  - Lower memory latency:
    - Caches
    - Pre-fetchers
  - Specialized instructions and hardware - multiply-add instructions, tensor operations

- Add Parallelism:
  - Instruction level parallelism
    - Instruction pipe-lining
    - Superscalar execution
    - Out-of-order execution
    - Speculative execution & Branch prediction
  - Vectorization
  - Hardware threads
  - Multiple cores
  - Multiple sockets
  - Multiple nodes

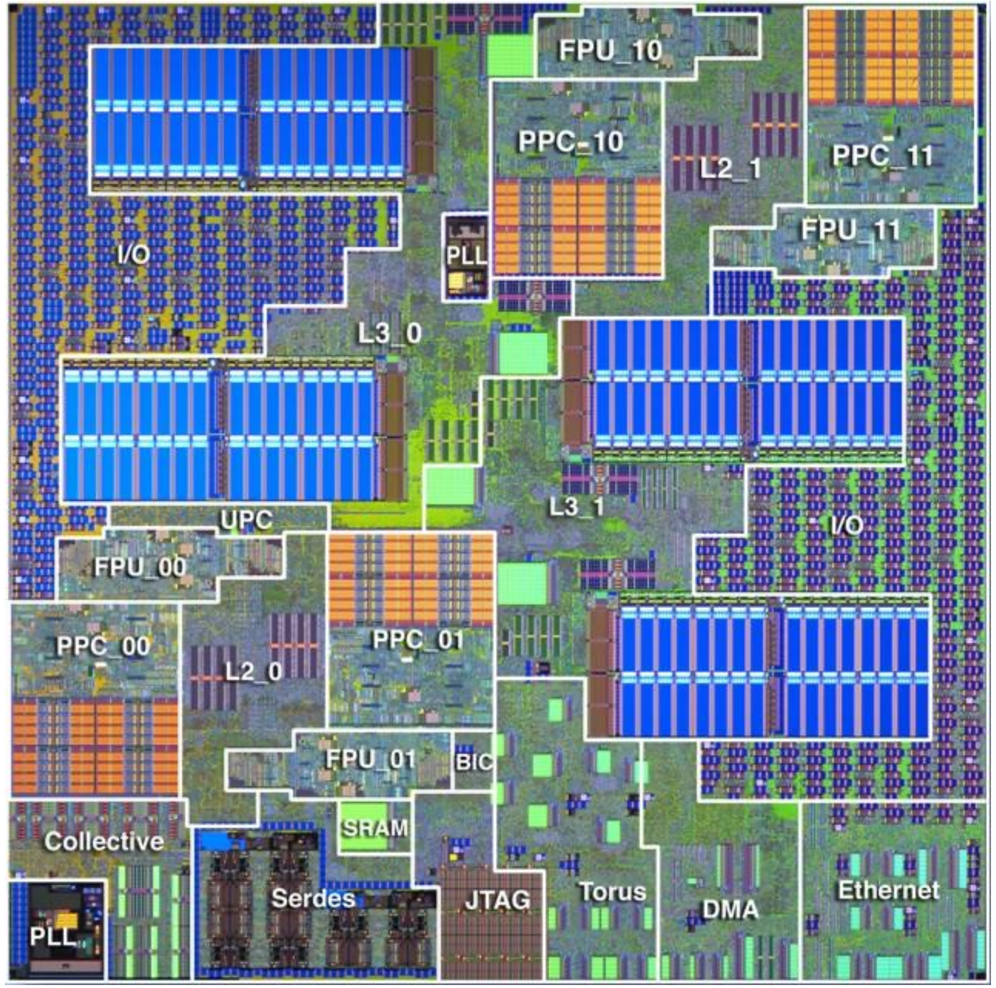
INTREPID: IBM BLUE GENE/P POWERPC 450

Intrepid

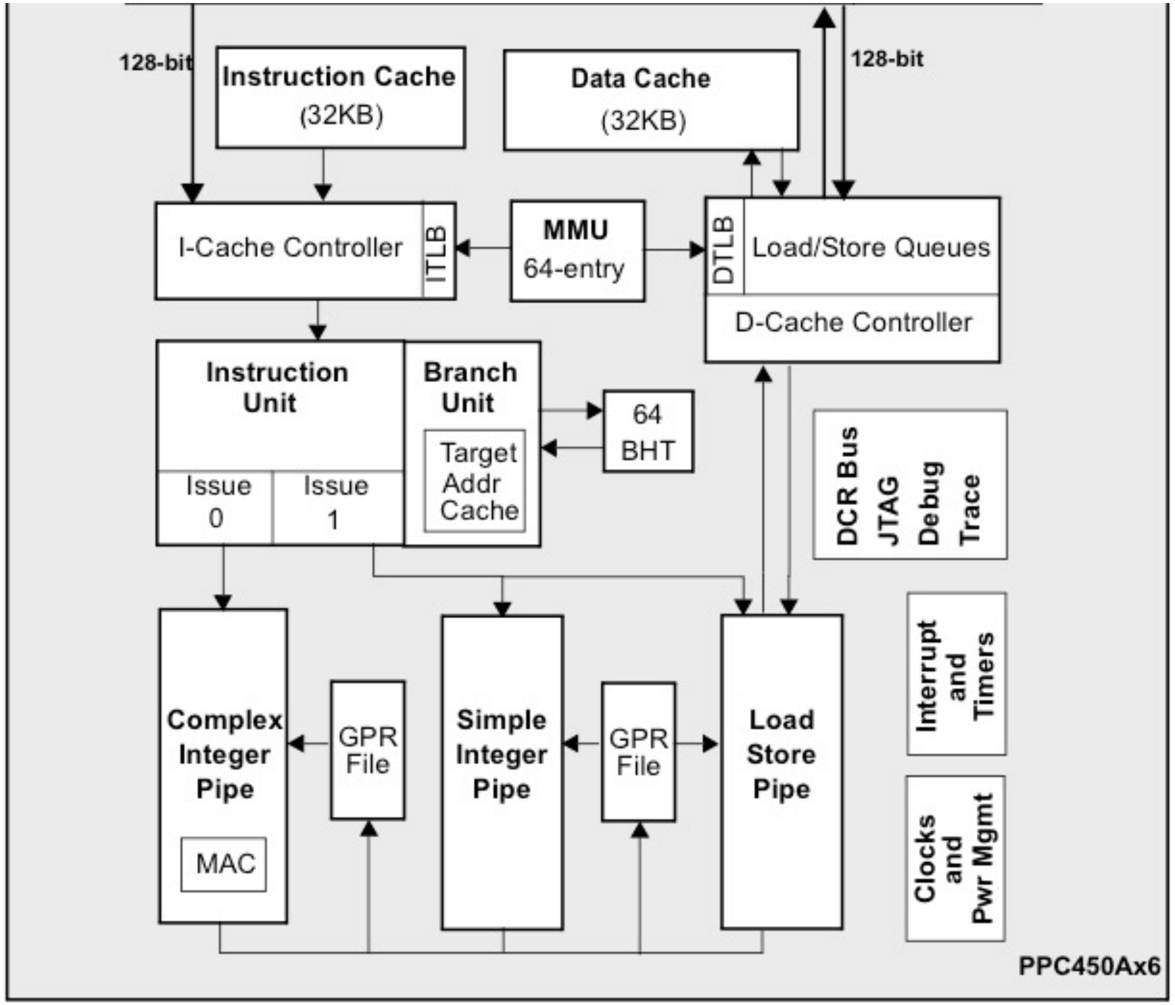
- 2008 ALCF Blue Gene/P System:
  - **40,960 nodes / 163,840 PPC cores**
  - 80 Terabytes of memory
  - Peak flop rate: 557 Teraflops
  - Linpack flop rate: 450.3
  - #6 on the Top500 list
- Storage:
  - 8 Petabytes of disk storage with an I/O rate of 80 GB/s
  - 8 Petabytes of archival storage (10,000 volume tape archive)

Blue Gene/P Compute Chip Die Photo

- Size: 170 mm (13mm x 13 mm)
- Process : 90 nm
- Transistors: 208 M
- 4 CPU core per node
- Clock Speed: 850 MHz
- Peak performance: 3.4 GFlops/core, 13.6 GFlops/node
- 2 GB of DDR 2 memory per node
- 5 network interfaces on chip

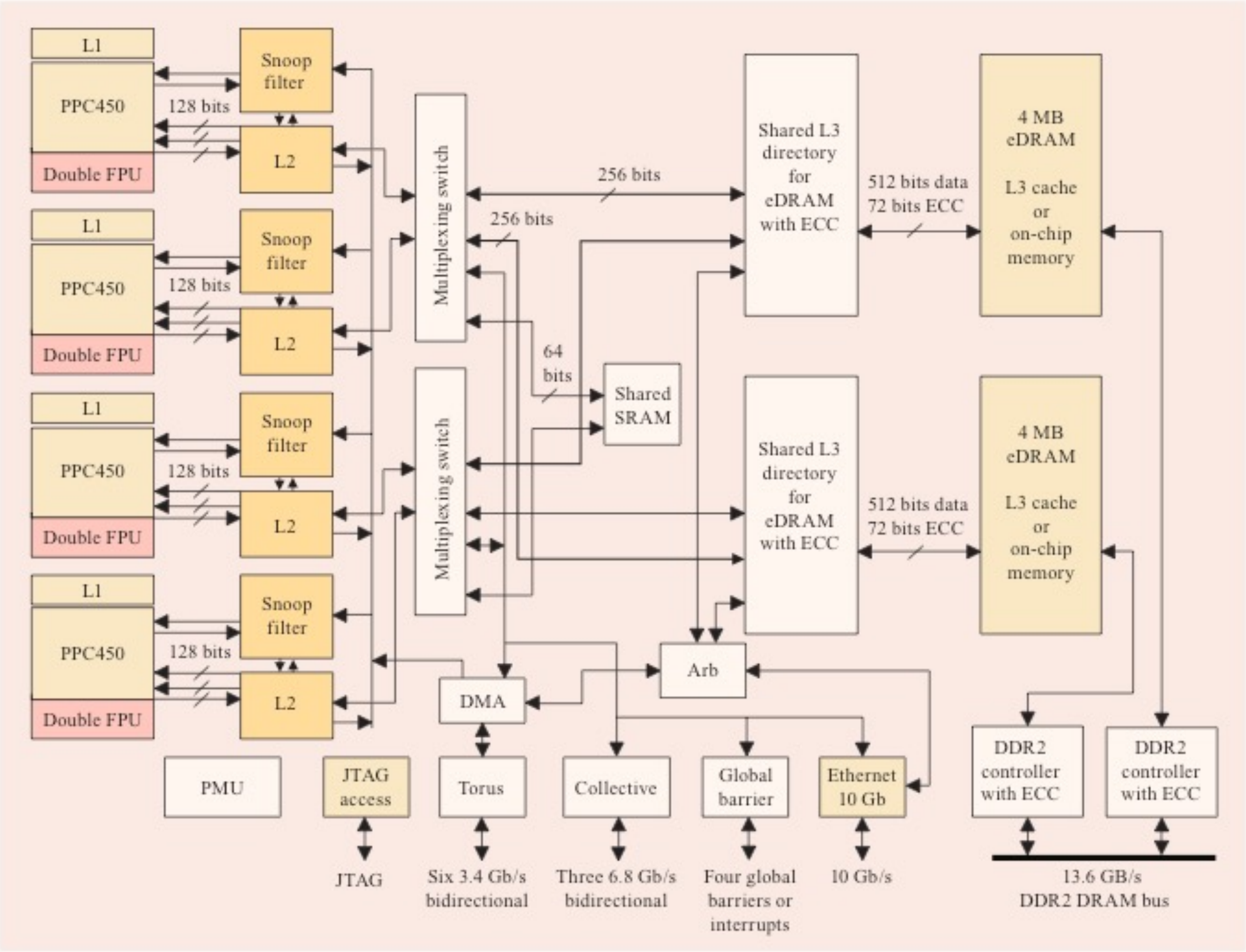


PowerPC 450 CPU



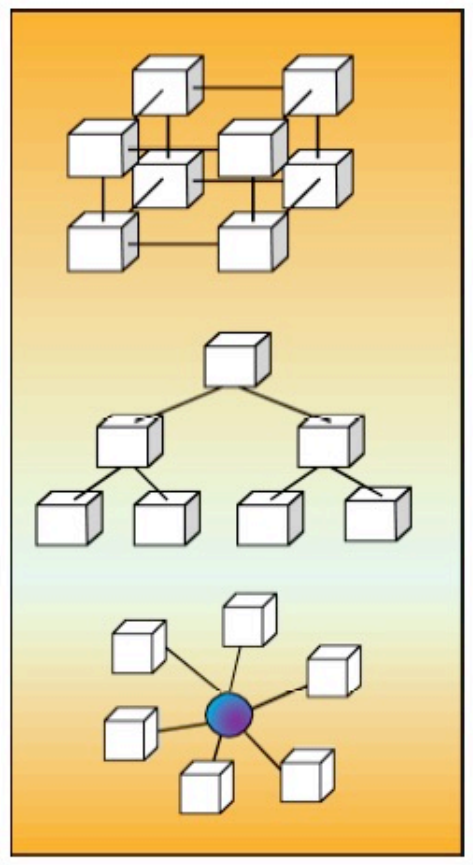
- **In order execution**
- **Dual Issue – can issues two instructions per cycle, must be to different pipelines**
- **Two wide floating point vector instructions**
- **Four Execution Pipelines:**
  - Load/Store (L-Pipe)
  - Simple Integer (J-Pipe)
  - Complex Integer (I-Pipe)
  - Floating Point
    - FMA
    - Vector
- **7 Stage instruction pipeline:**
  - Instruction Fetch
  - Instruction Decode
  - Issue
  - Register Access
  - Pipeline line stage 1
  - Pipeline line stage 2
  - Write Back

BG/P Memory Hierarchy



- L1 Instruction and L1 Data caches:
  - 32 KB total size, **4 cycle latency**, 32-Byte line size
- L2 Data cache:
  - **2KB prefetch buffer**, **12 cycle latency**, 16 lines, 128-byte line size
- L3 Data cache:
  - 8 MB, **50 cycles latency,** 128-byte line size,
- Memory:
  - **Two memory channels**
  - **13.6 GB/s memory bandwidth**
  - 2GB DDR-2 at 425 MHz, **104 cycles**

Blue Gene/P Network



3 Dimensional Torus

- Interconnects all compute nodes
- Communications backbone for point-to-point (send/receive)
- 3.4 Gb/s on all 12 node links (5.1 GB/s per node)
- 0.5 µs latency between nearest neighbors, 5 µs to the farthest
- MPI: 3 µs latency for one hop, 10 µs to the farthest
- *Requires half-rack or larger partition*

Collective Network

- One-to-all broadcast functionality
- Reduction operations for integers and doubles
- 6.8 Gb/s of bandwidth per link per direction
- Latency of one way tree traversal 1.3 µs, MPI 5 µs
- Interconnects all compute nodes and I/O nodes

Low Latency Global Barrier and Interrupt

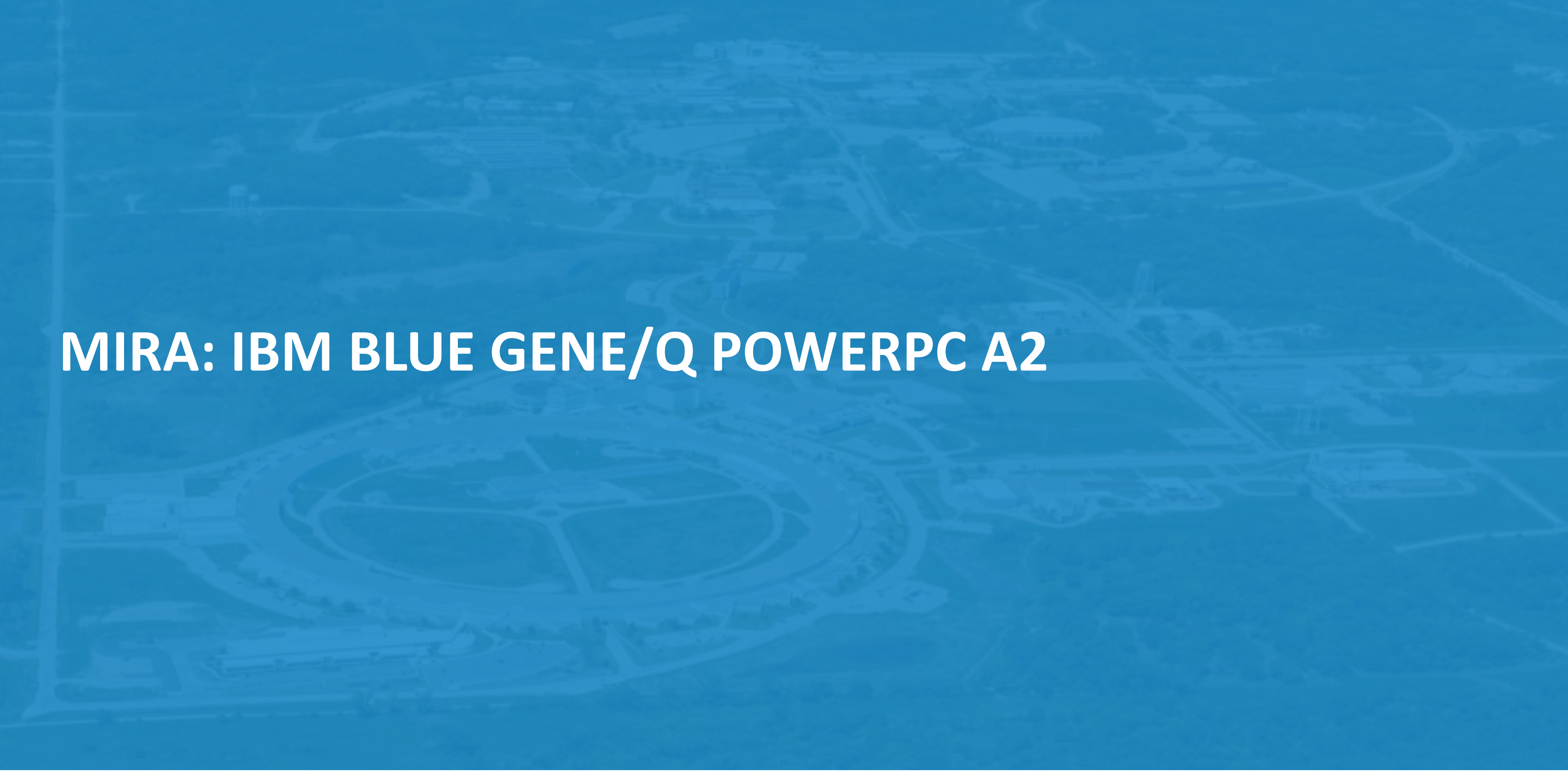
- Latency of one way to reach 72K nodes 0.65 µs, MPI 1.6 µs

10 Gb/s functional Ethernet

- Disk I/O

1Gb private control (JTAG)

- Service node/system management

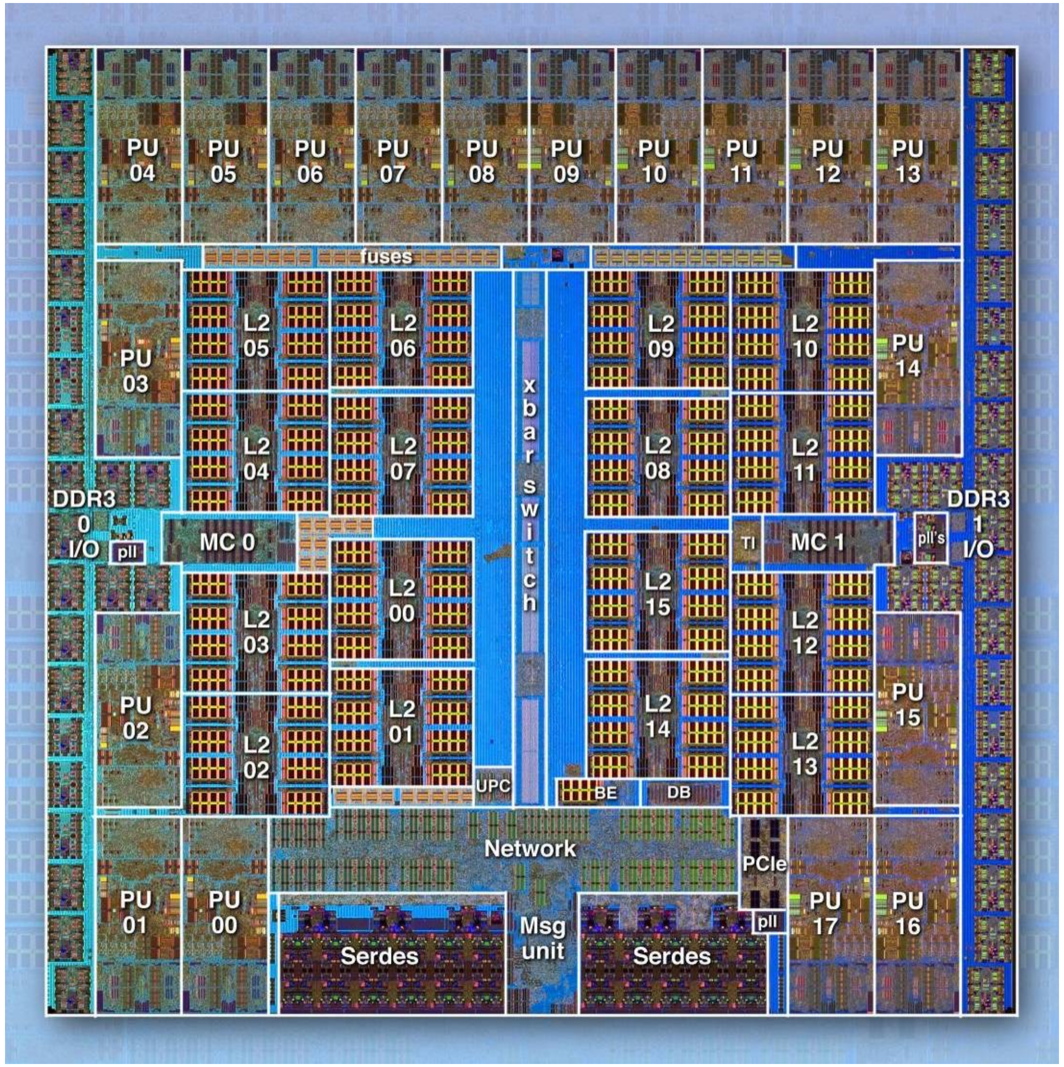


MIRA: IBM BLUE GENE/Q POWERPC A2

ALCF BG/Q Systems

- *2012 Mira* – BG/Q system
  - —**49,152 nodes / 786,432 cores**
  - —768 TB of memory
  - —Peak flop rate: 10 PF
  - —Linpack flop rate: 8.1 PF
  - —#3 on Top 500
- Storage
  - —Scratch: 28.8 PB raw capacity, 240 GB/s bw
  - —Home: 1.8 PB raw capacity, 45 GB/s bw

BlueGene/Q Compute Chip



Chip

- 360 mm² Cu-45 technology (SOI)
- 1.5 B transistors

18 Cores

- **16 compute cores – 205 GF total**
- 17th core for system functions (OS, RAS)
- plus 1 redundant processor
- L1 I/D cache = 16kB/16kB

Crossbar switch

- Each core connected to shared L2
- Aggregate read rate of 409.6 GB/s

Central shared L2 cache

- 32 MB eDRAM
- 16 slices

Dual memory controller

- 16 GB external DDR3 memory
- **42.6 GB/s bandwidth**

On Chip Networking

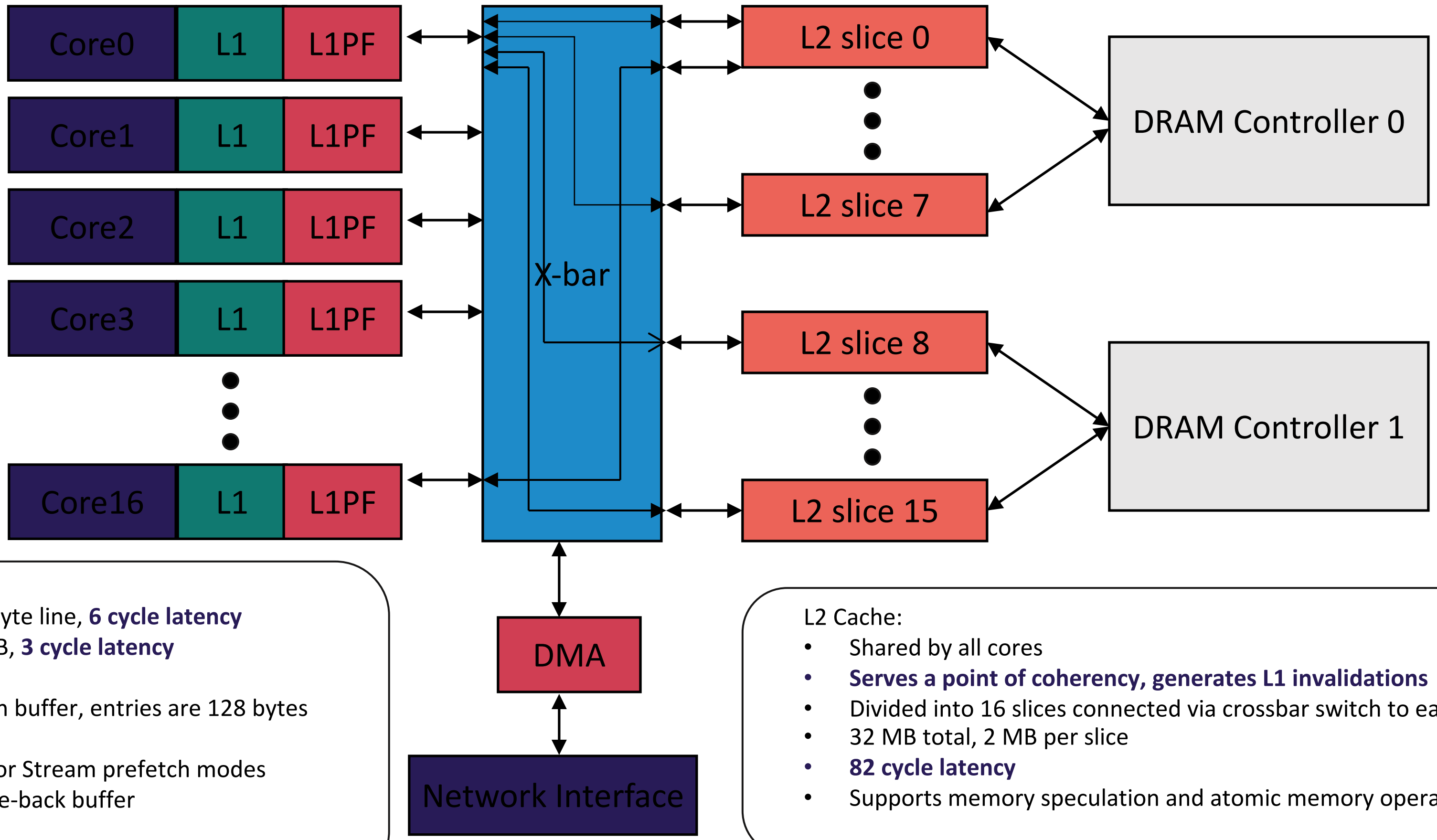
- Router logic integrated into BQC chip
- DMA, remote put/get, collective operations
- 11 network ports

BG/Q Memory Hierarchy

Crossbar switch connects:
- L1P’s, L2 slices, Network, PCIe interface

Aggregate bandwidth across slices:
- Read: 409.6 GB/s, Write: 204.8 GB/s

Memory:
- **Two on chip memory controllers**
- Each connects to 8 L2 slices via 2 ring buses
- Each controller drives a 16+2 byte DDR-3 channel at 1.33 GT/s
- Peak bandwidth is 42.67 GB/s (excluding ECC)
- **Latency > 350 cycles**



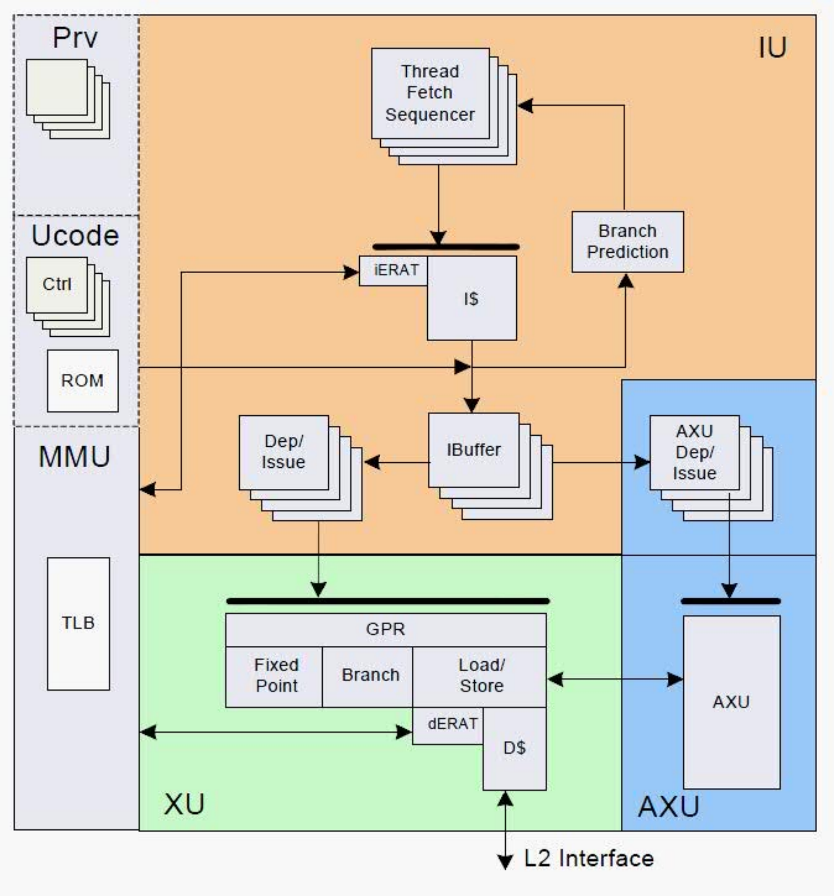
L1 Cache:
- **Data**: 16KB, 64 byte line, **6 cycle latency**
- **Instruction**: 16KB, **3 cycle latency**

L1 Prefetcher (L1P):
- 32 entry prefetch buffer, entries are 128 bytes
- 24 cycle latency
- Operates in List or Stream prefetch modes
- Operates as write-back buffer

L2 Cache:
- Shared by all cores
- **Serves a point of coherency, generates L1 invalidations**
- Divided into 16 slices connected via crossbar switch to each core
- 32 MB total, 2 MB per slice
- **82 cycle latency**
- Supports memory speculation and atomic memory operations

BG/Q Core

- **In-order execution**
- **Runs at 1.6 GHz**
- **4-way Simultaneous Multi-Threading**
- **Four wide floating point vector instructions**



Four Functional Units:

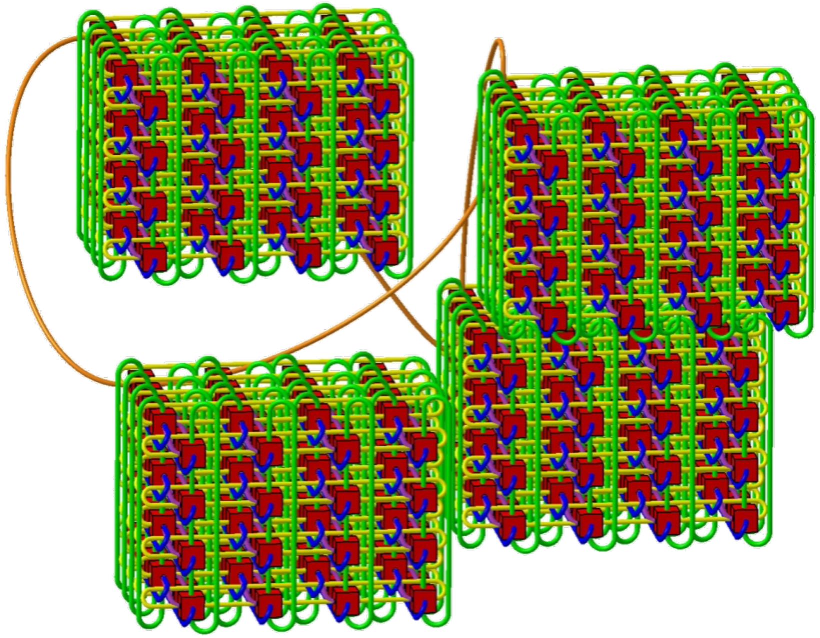
- IU – instructions fetch and decode
- XU – Branch, Integer, Load/Store instructions
- AXU – Floating point instructions
  - Standard PowerPC instructions
  - QPX 4 wide SIMD
- MMU – memory management (TLB)

Instruction Issue:

- **2-way concurrent issue if 1 XU + 1 AXU instruction**
- **A given thread may only issue 1 instruction per cycle**
- **Two threads may each issue 1 instruction each cycle**

The BG/Q Network

- **5D torus network:**
  - –Achieves high nearest neighbor bandwidth while increasing bisectional bandwidth and reducing hops vs 3D torus
  - –Allows machine to be partitioned into independent sub machines
    - – No impact from concurrently running codes.
  - –Hardware assists for collective & barrier functions over COMM_WORLD and rectangular sub communicators
  - –Half rack (midplane) is 4x4x4x4x2 torus (last dim always 2)
- **No separate Collectives or Barrier network:**
  - –Single network used for point-to-point, collectives, and barrier operations
- **Additional 11th link to IO nodes**
- **Two type of network links**
  - Optical links between midplanes
  - Electrical inside midplane

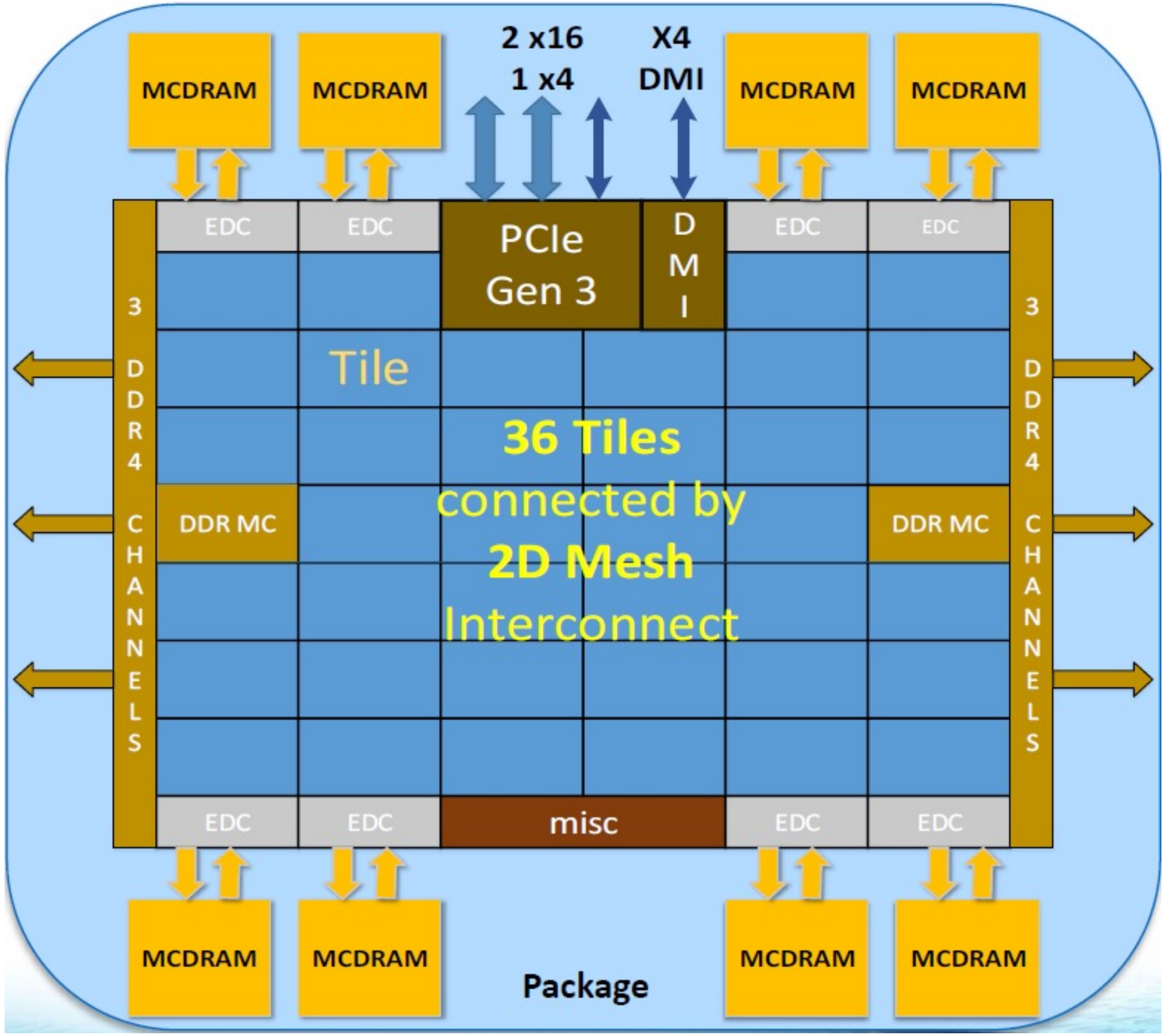


THETA: INTEL XEON PHI KNIGHTS LANDING

Theta

- **2016 Theta:**
  - Cray XC40 system
  - **4,392 compute nodes/ 281,088 cores**
  - 11.7 PetaFlops peak performance
- **Memory:**
  - 892 TB of total system memory
    - **16 GB IPM per node**
    - **192 GB DDR4-2400 per node**
- **Network:**
  - Cray Aries interconnect
  - Dragonfly network topology
- **Filesystems:**
  - Project directories: 10 PB Lustre file system
  - Home directories: GPFS

Theta KNL Processor (KNL 7230)



Chip
- 683 mm²
- 14 nm process
- 8 Billion transistors

64 Cores (up to 72)
- 32 tiles (up to 36)
- 2 cores per tile
- Up to 3 TF per node
- 1.3 GHz, (1.1 – 1.5 GHz Turbo)

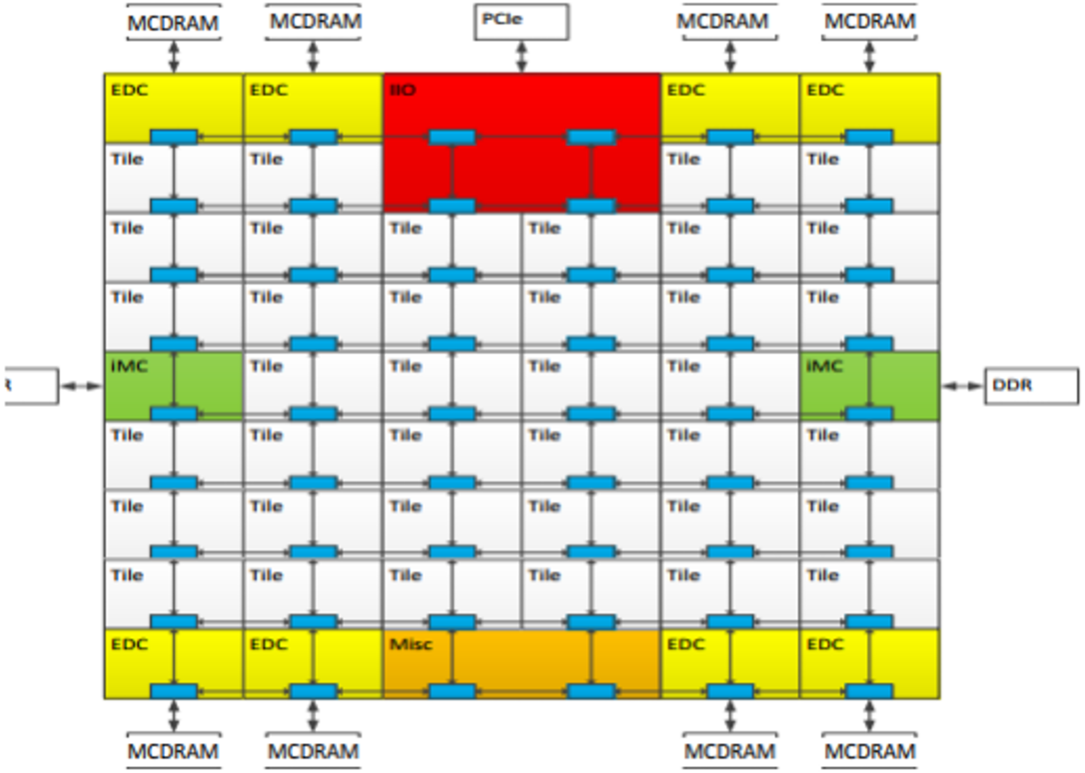
2D Mesh Interconnect
- Tiles connected by 2D mesh

On Package Memory
- 16 GB MCDRAM
- 8 Stacks
- 485 GB/s bandwidth

6 DDR4 memory channels
- 2 controllers
- up to 384 GB external DDR4
- 90 GB/s bandwidth

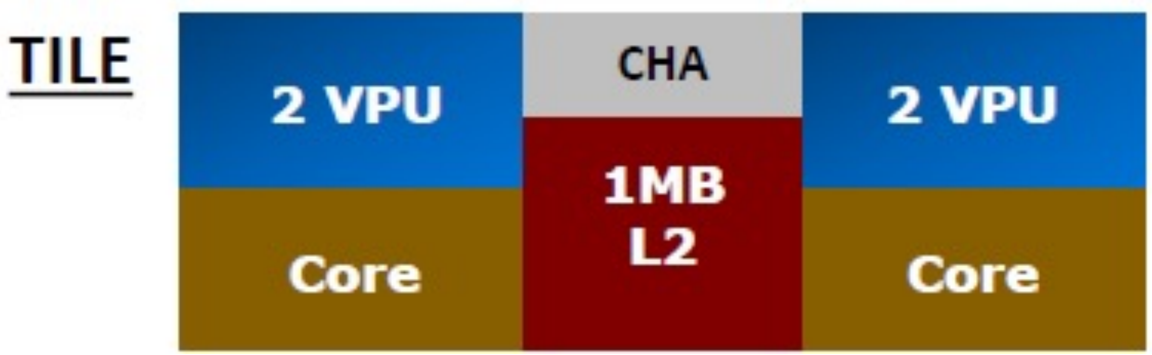
On Socket Networking
- Omni-Path NIC on package
- Connected by PCIe

KNL Mesh Interconnect



- 2D mesh interconnect connects
  - Tiles (CHA)
  - MCDRAM controllers
  - DDR controllers
  - Off chip I/O (PCIe, DMI)
- YX routing:
  - Go in Y→ turn → Go in X
  - Messages arbitrate on injection and on turn
- Cache coherent
  - Uses MESIF protocol
- Clustering mode allow traffic localization
  - All-to-all, Quadrant, Sub-NUMA

KNL Tile



TILE

2 VPU

CHA

1MB L2

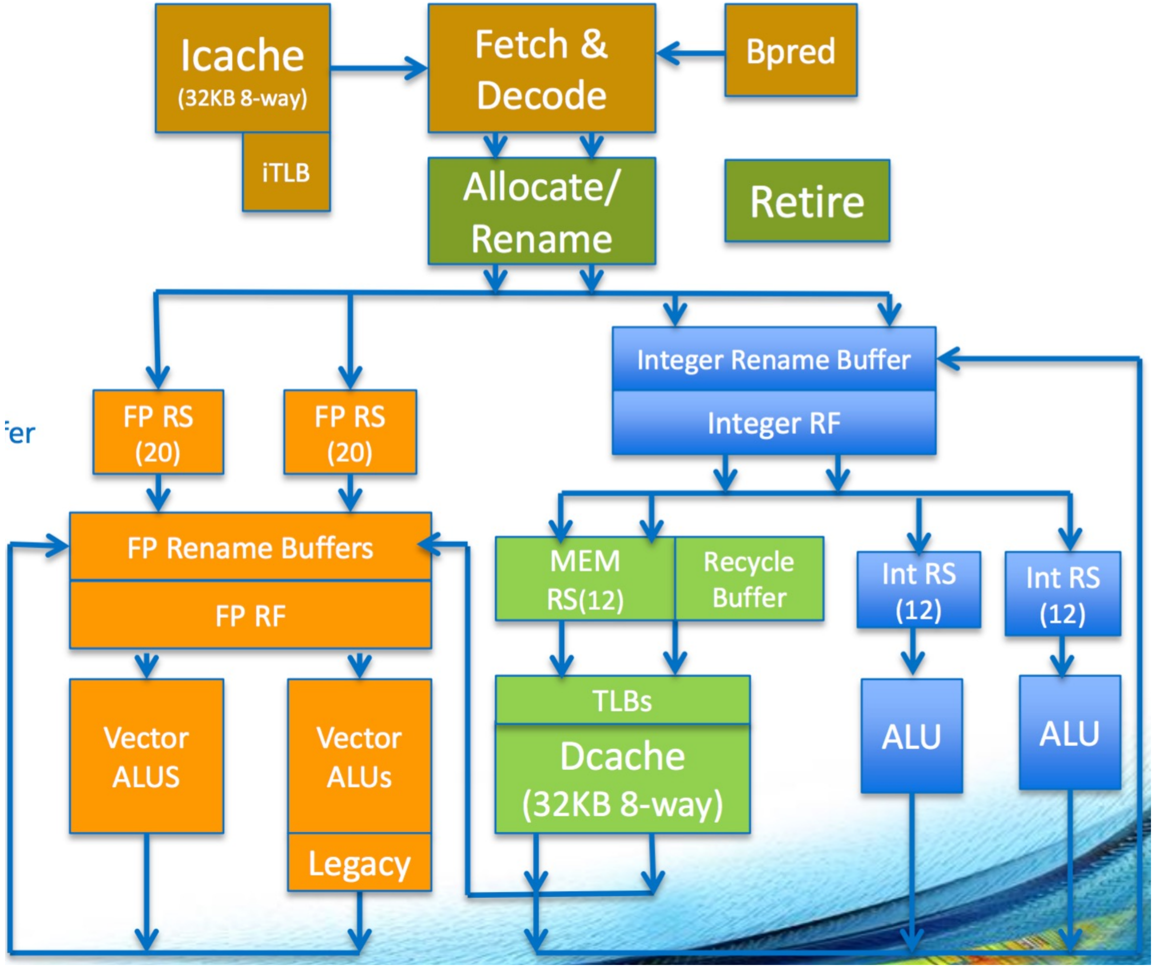
2 VPU

Core

Core

- Two CPUs
- 2 vector units (VPUs) per core
- 1 MB Shared L2 cache
  - Coherent across all tiles (32-36 MB total)
  - 16 Way
  - 1 line read and ½ line write per cycle
- Caching/Home agent
  - Distributed tag directory, keeps L2s coherent
  - Implements MESIF cache coherence protocol
  - Interface to mesh

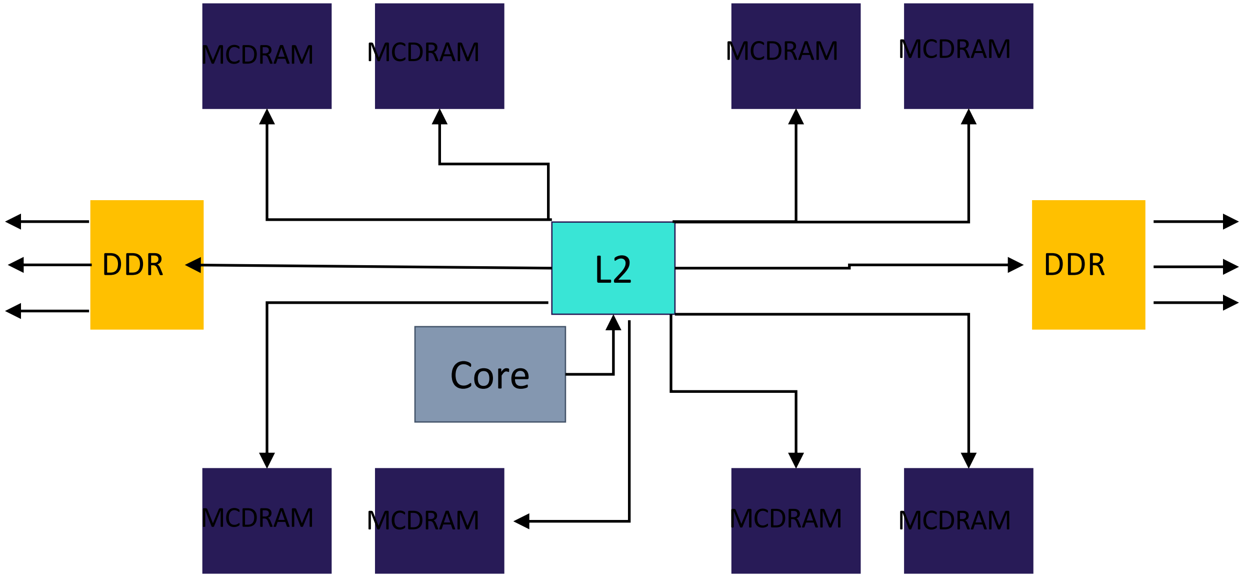
KNL CORE



- Based on Silvermont (Atom)
  - Lower power design
  - Out of order execution
  - Binary compatible with Xeon
  - Introduced AVX-512 vector instructions
  - Includes hardware gather/scatter engine
- Instruction Issue & Execute:
  - 2 wide decode/rename/retire
  - 6 wide execute
- Functional units:
  - 2 Integer ALUs (Out of Order)
  - 2 Memory units (In Order reserve, OoO complete)
  - 2 VPU's with *AVX-512* (Out of Order)
- L1 data cache
  - 32 KB, 8 way associative
  - 2 64B load ports, 1 64B write port
- 4 Hardware threads per core
  - 1 active thread can use full resources of core
  - ROB, Rename buffer, RD dynamically partitioned between threads
  - Caches and TLBs shared

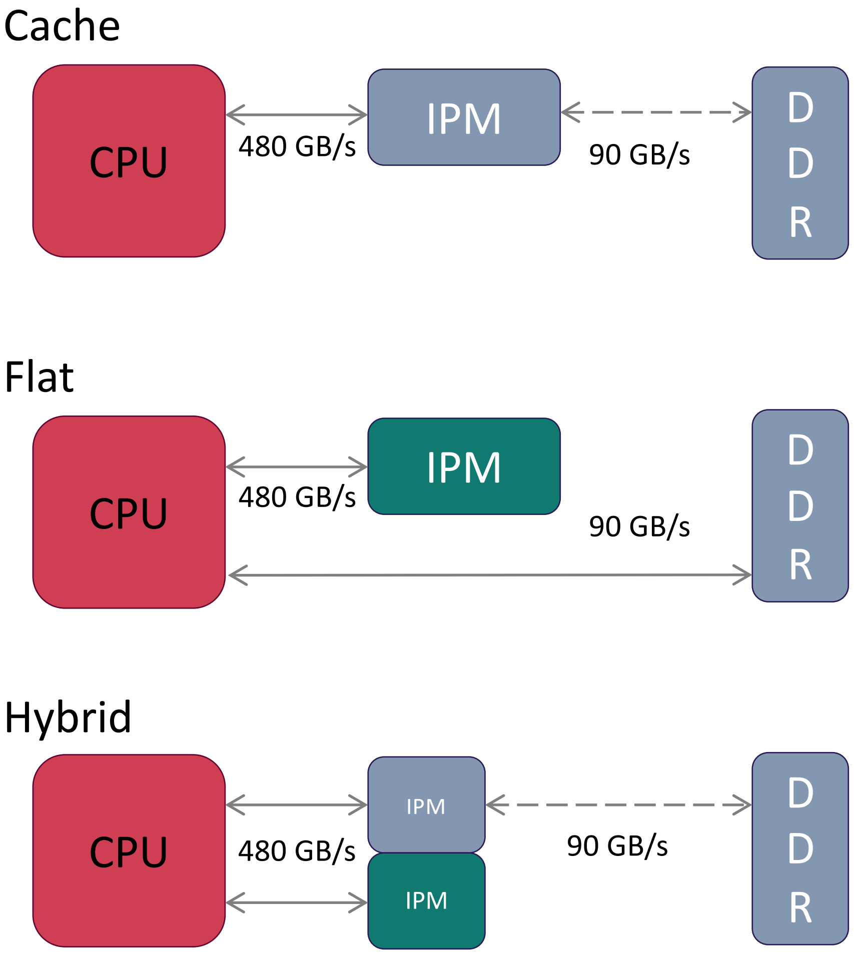
Memory

- **Two memory types**
  - In Package Memory (IPM)
    - 16 GB MCDRAM
    - ~485 GB/s bandwidth
  - Off Package Memory (DDR)
    - Up to 384 GB
    - ~90 GB/s bandwidth
- **One address space**
  - Minor NUMA effects
  - Sub-NUMA clustering mode creates four NUMA domains



Memory Modes - IPM and DDR

SELECTED AT NODE BOOT TIME

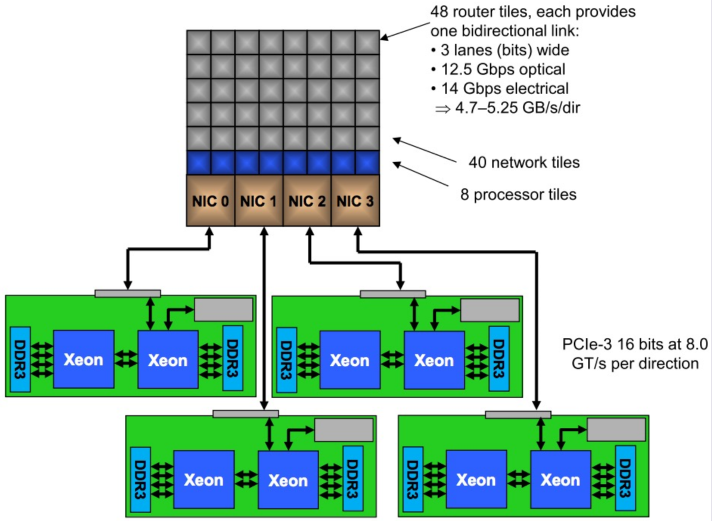


- **Memory configurations**
  - Cached:
    - DDR fully cached by IPM
    - No code modification required
    - Less addressable memory
    - Bandwidth and latency worse than flat mode
  - Flat:
    - Data location completely user managed
    - Better bandwidth and latency
    - More addressable memory
  - Hybrid:
    - ¼, ½ IPM used as cache rest is flat
- **Managing memory:**
  - jemalloc & memkind libraries
  - numctl command
  - Pragmas for static memory allocations

Aries Dragonfly Network

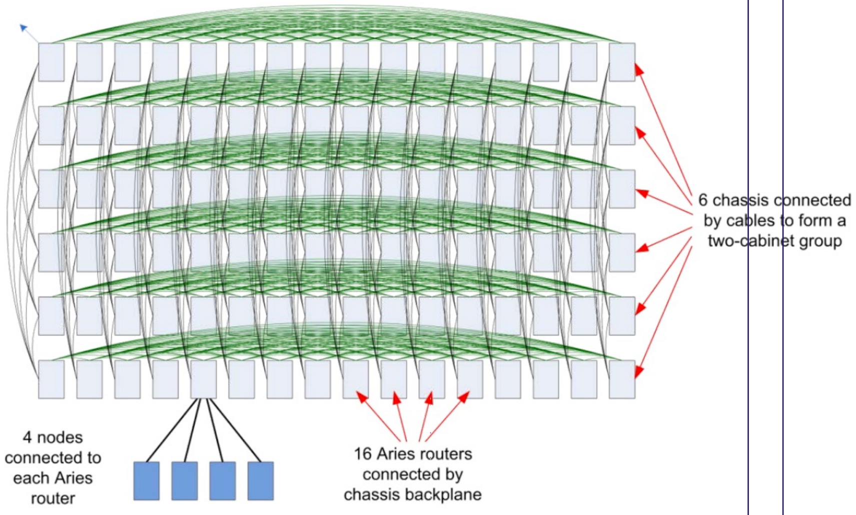
Aries Router:

- 4 Nodes connect to an Aries
- 4 NIC's connected via PCIe
- 40 Network tiles/links
- 4.7-5.25 GB/s/dir per link



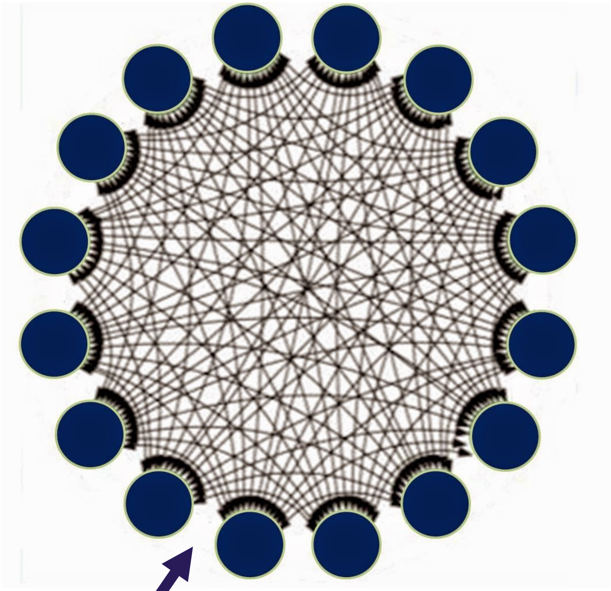
Connections within a group:

- 2 Local all-to-all dimensions
  - 16 all-to-all horizontal
  - 6 all-to-all vertical
- 384 nodes in local group

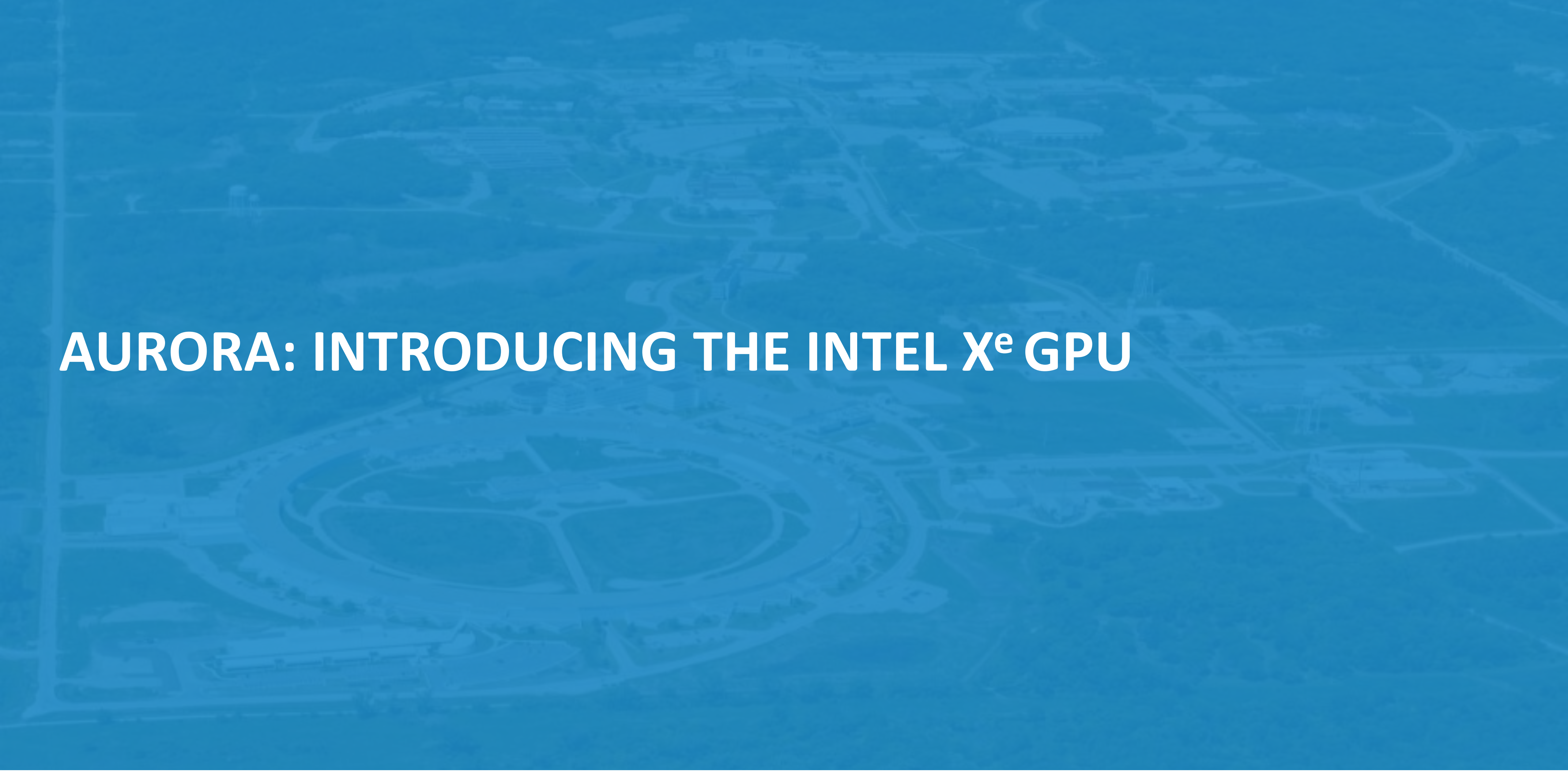

6 chassis connected by cables to form a two-cabinet group
4 nodes connected to each Aries router
16 Aries routers connected by chassis backplane

Connectivity between groups:

- Each group connected to every other group
- Restricted bandwidth between groups



Theta has 12 groups with 12 links between each group

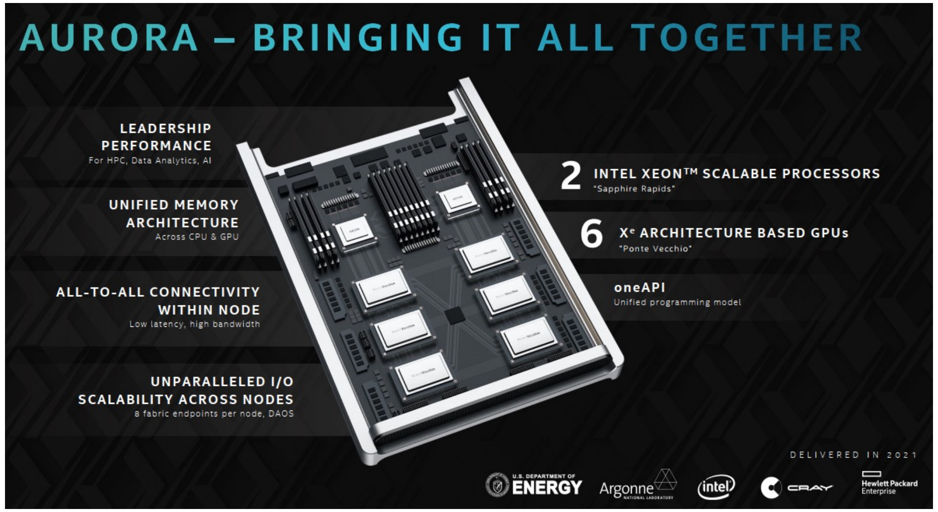


AURORA: INTRODUCING THE INTEL X^e GPU

Aurora: A High-level View

- Intel-Cray machine arriving at Argonne soon
  - Sustained Performance > 1Exaflops
- Intel Xeon processors and Intel X^e GPUs
  - 2 Xeons (Sapphire Rapids)
  - 6 GPUs (Ponte Vecchio [PVC])
    - All to all connection
    - Low latency and high bandwidth
- Greater than 10 PB of total memory
  - Unified memory architecture across CPUs and GPUs
- Cray Slingshot fabric and Shasta platform
  - 8 fabric end points per node
- Filesystem
  - Distributed Asynchronous Object Store (DAOS)
    - ≥ 230 PB of storage capacity
    - Bandwidth of > 25 TB/s
  - Lustre
    - 150 PB of storage capacity
    - Bandwidth of ~1TB/s

Argonne NATIONAL LABORATORY
U.S. DEPARTMENT OF ENERGY
intel
Hewlett Packard Enterprise
Aurora



AURORA – BRINGING IT ALL TOGETHER

LEADERSHIP PERFORMANCE
For HPC, Data Analytics, AI

UNIFIED MEMORY ARCHITECTURE
Across CPU & GPU

ALL-TO-ALL CONNECTIVITY WITHIN NODE
Low latency, high bandwidth

UNPARALLELED I/O SCALABILITY ACROSS NODES
8 fabric endpoints per node, DAOS

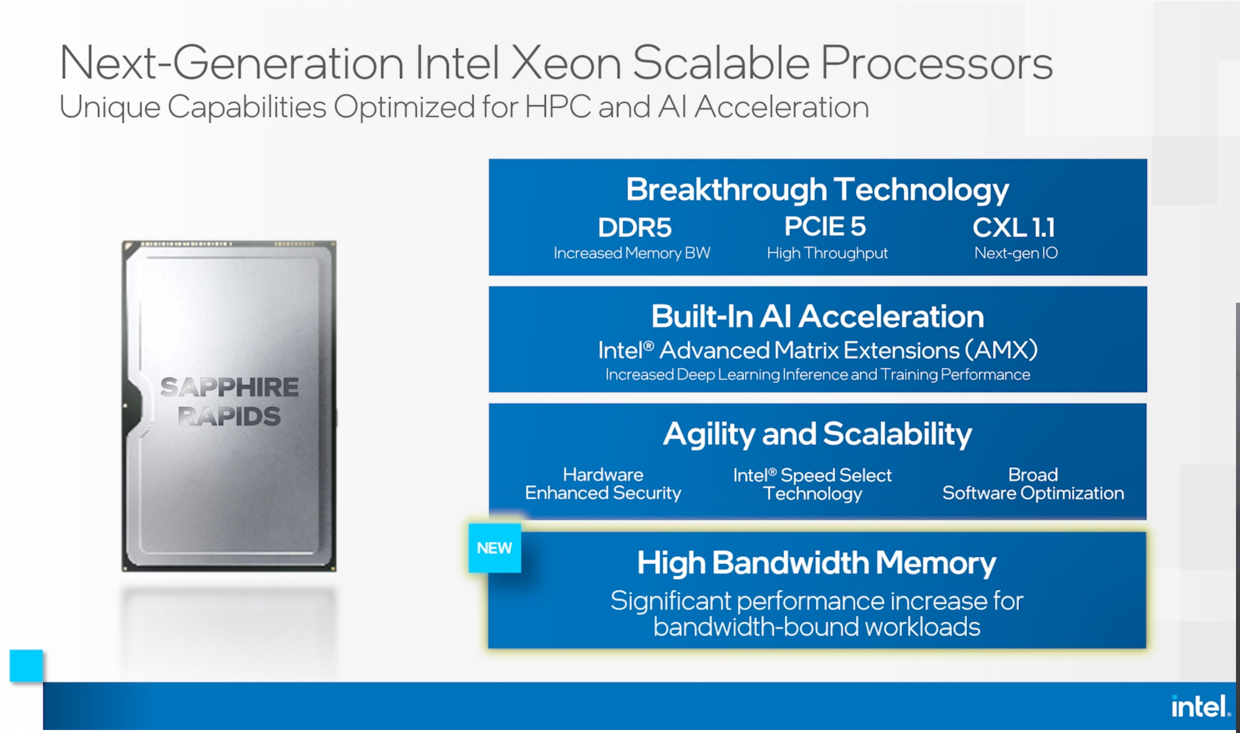
2 INTEL XEON™ SCALABLE PROCESSORS
"Sapphire Rapids"

6 X^e ARCHITECTURE BASED GPUs
"Ponte Vecchio"

oneAPI
Unified programming model

DELIVERED IN 2021

Sapphire Rapids CPU – Ponte Vecchio GPU



Next-Generation Intel Xeon Scalable Processors

Unique Capabilities Optimized for HPC and AI Acceleration

SAPPHIRE RAPIDS

Breakthrough Technology

DDR5	PCIE 5	CXL 1.1
Increased Memory BW	High Throughput	Next-gen IO

Built-In AI Acceleration

Intel® Advanced Matrix Extensions (AMX)

Increased Deep Learning Inference and Training Performance

Agility and Scalability

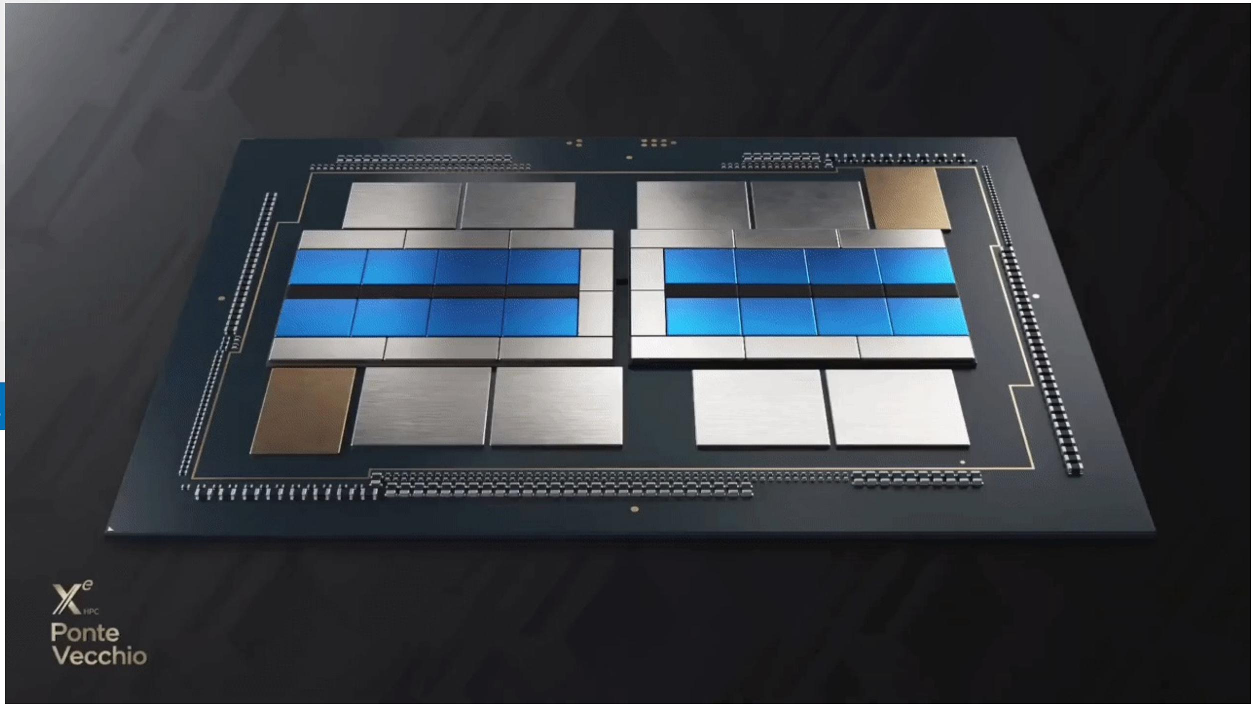
Hardware Enhanced Security	Intel® Speed Select Technology	Broad Software Optimization

NEW

High Bandwidth Memory

Significant performance increase for bandwidth-bound workloads

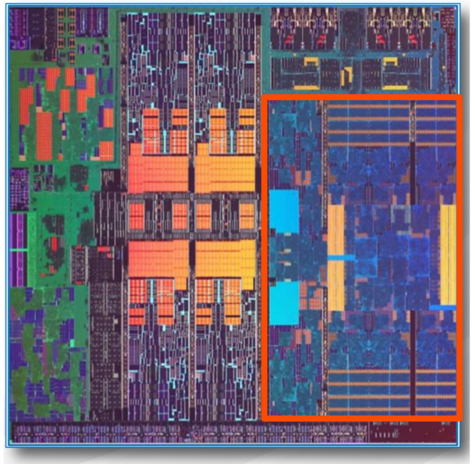
intel.



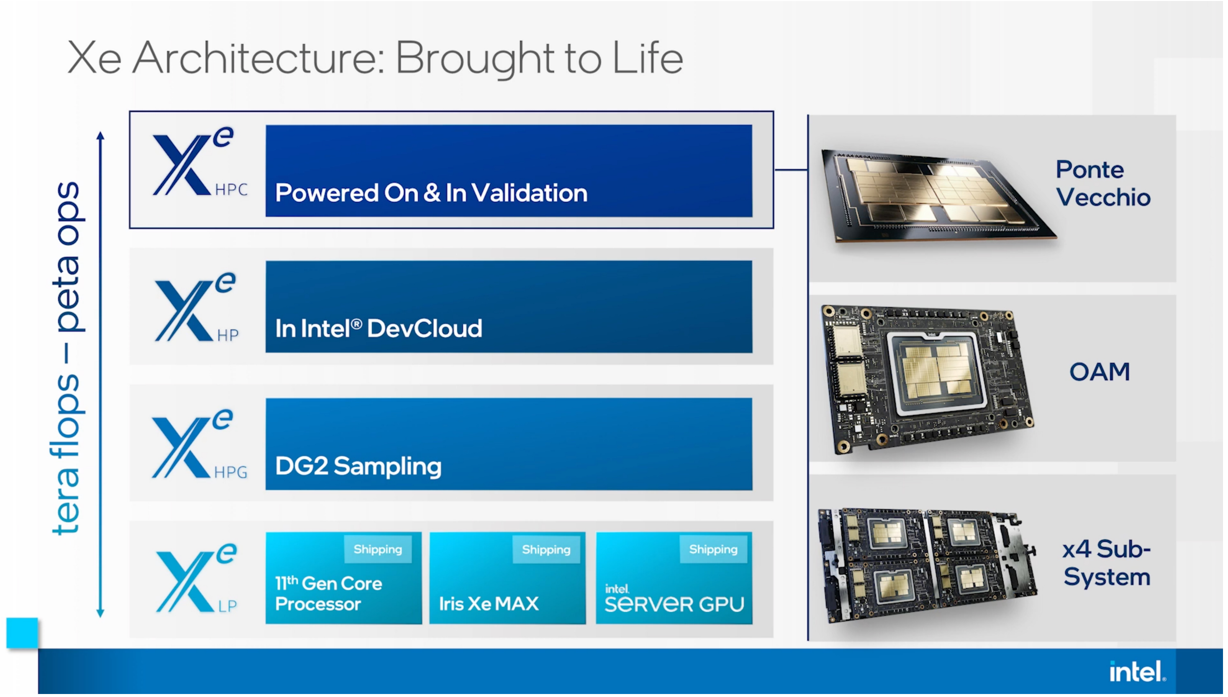
Intel's HPC GM Trish Damkroger Keynote ISC 2021
https://www.youtube.com/watch?v=PuEcCRJLrvs
https://download.intel.com/newsroom/2021/data-center/Intel-ISC2021-keynote-presentation.pdf

Intel GPUs

- Intel has been building GPUs integrated with CPUs for over a decade
- Currently released products use the Gen and Gen 11 versions
  - Gen9 – used in Skylake
  - Gen11 – used in Ice Lake
- Low performance by design due to power and space limits
  - Gen9 peak DP flops: 100-300 GF
  - Gen 9 introduce in 2015
- X^e LP
  - Platforms: Tiger Lake, DG1, SG1
  - Integrated & Discrete



Tiger Lake SoC with X$^e_{LP}$ GPU



Xe Architecture: Brought to Life

tera flops – peta ops

X$^e_{HPC}$ Powered On & In Validation — Ponte Vecchio

X$^e_{HP}$ In Intel® DevCloud — OAM

X$^e_{HPG}$ DG2 Sampling — x4 Sub-System

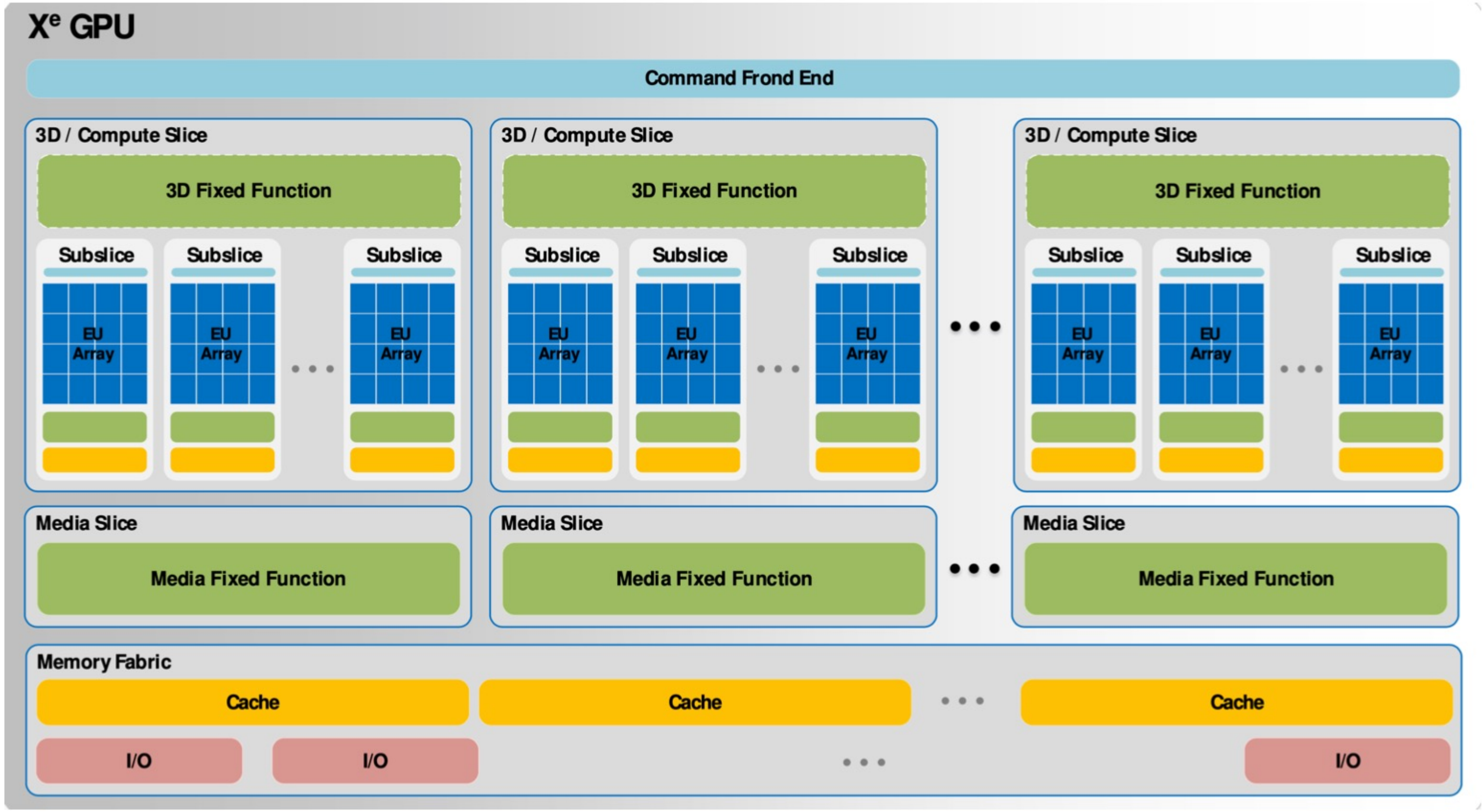
X$^e_{LP}$ Shipping 11th Gen Core Processor | Shipping Iris Xe MAX | Shipping intel SERVER GPU

intel.

Intel's HPC GM Trish Damkroger Keynotes 2021 ISC
https://www.youtube.com/watch?v=PuEcCRJLrvs

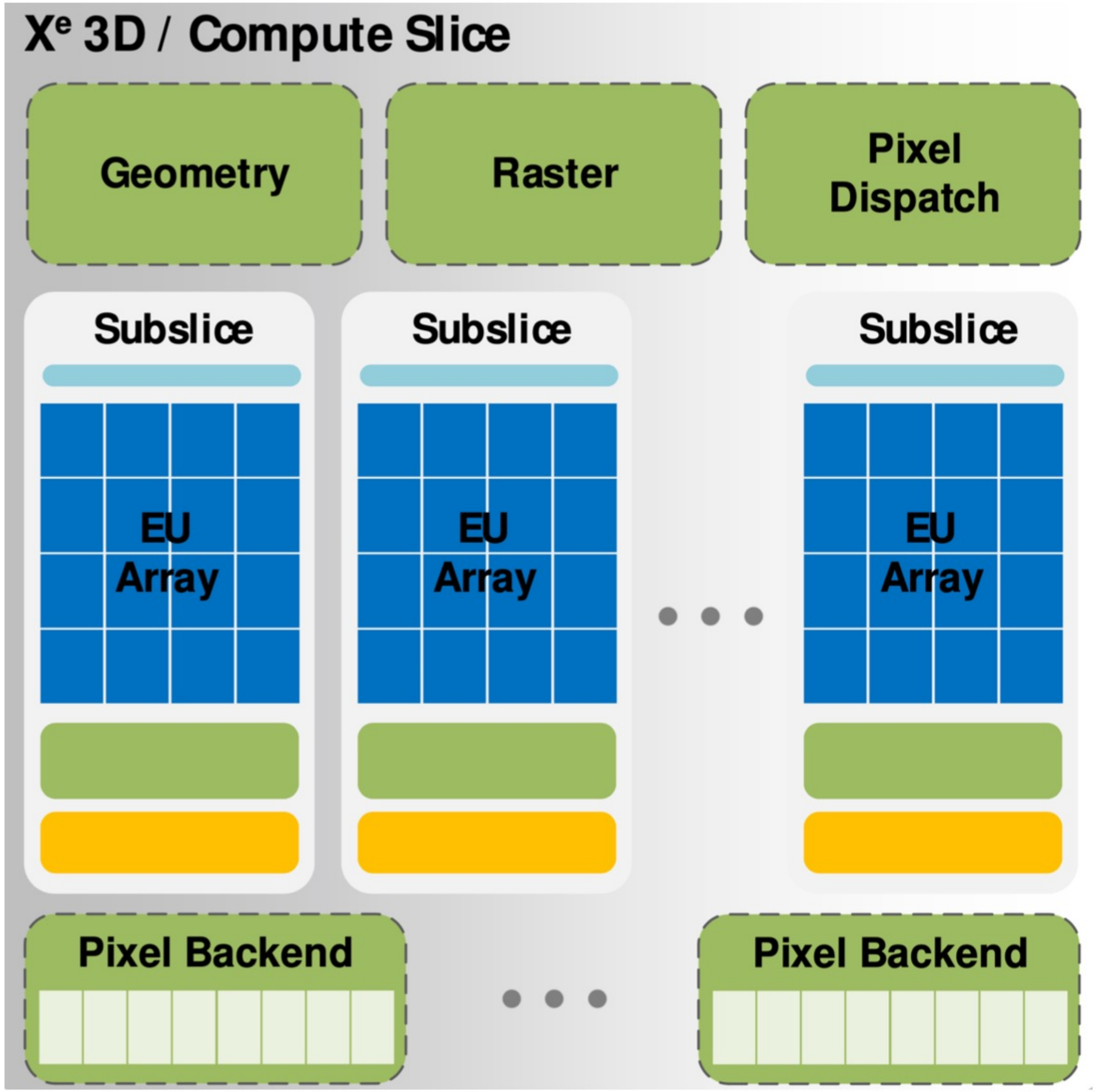
High Level Xe Architecture

- X^e GPU is composed of
  - 3D/Compute Slice
  - Media Slice
  - Memory Fabric / Cache



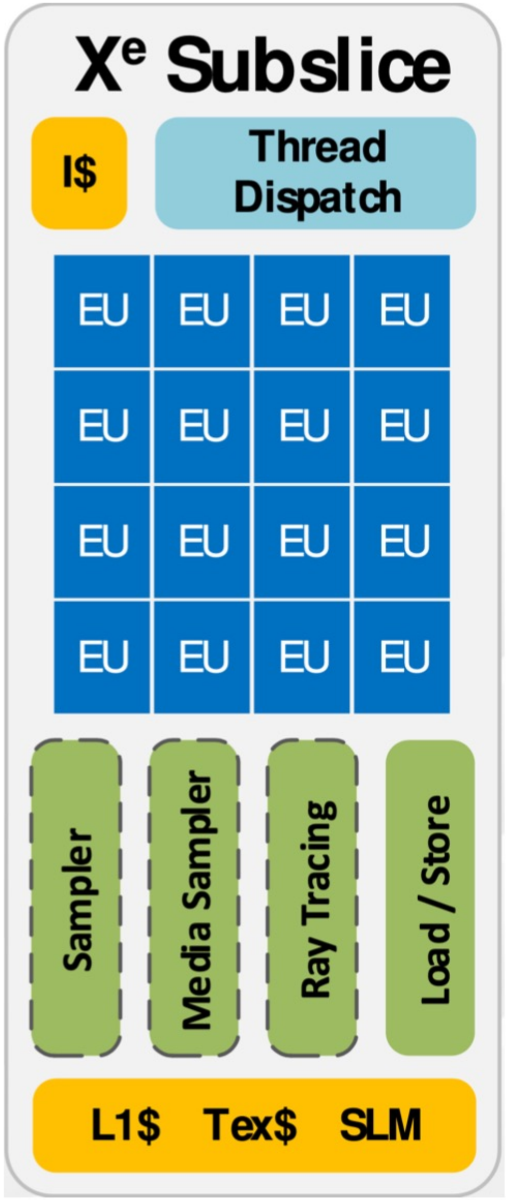
XE 3D/Compute Slice

- A slice contains
  - Variable number of subslices
  - 3D Fixed Function (optional)
    - Geometry
    - Raster



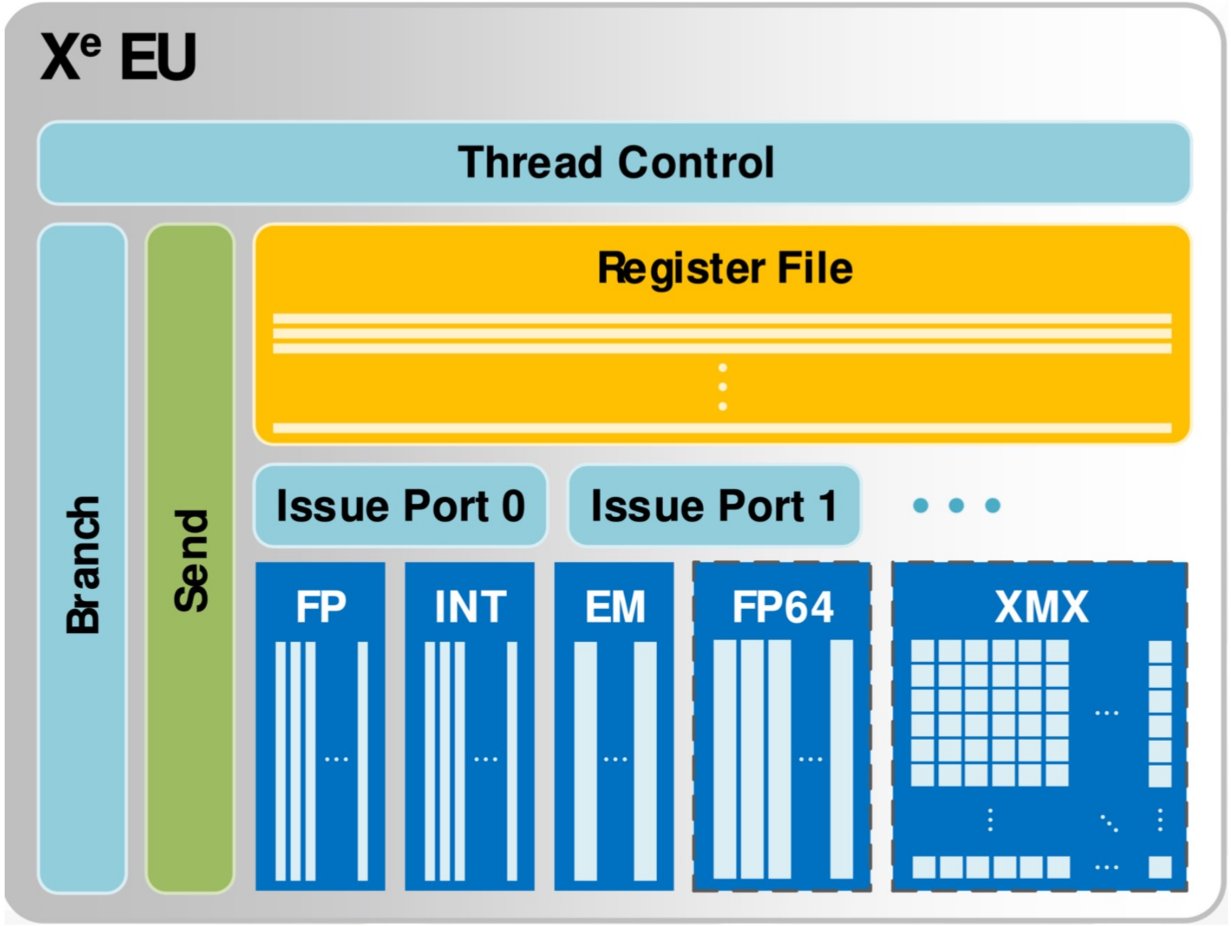
XE Subslice

- A sub-slice contains:
  - 16 EUs
  - Thread dispatch
  - Instruction cache
  - L1, texture cache, and shared local memory
  - Load/Store
  - Fixed Function (optional)
    - 3D Sampler
    - Media Sampler
    - Ray Tracing



XE Execution Unit

- ❑ The EU executes instructions
  - ❑ Register file
  - ❑ Multiple issue ports
  - ❑ Vector pipelines
    - ❑ Float Point
    - ❑ Integer
    - ❑ Extended Math
    - ❑ FP 64 (optional)
    - ❑ Matrix Extension (XMX) (optional)
  - ❑ Thread control
  - ❑ Branch
  - ❑ Send (memory)



Intel Devcloud

- Intel GPUs and oneAPI software are available to try out on the Intel DevCloud
- oneAPI collection of software components:
  - Compilers (C, C++, Fortran)
  - Programming models (DPC++, OpenMP, OpenCL)
  - Libraries (OneMKL, OneDNN, …)
  - Tools (Vtune, Advisor)
- A development sandbox to develop, test and run workloads across a range of Intel CPUs, GPUS, and FPGAs using Intel openAPI Beta software
- Try the oneAPI toolkits, compilers, performance libraries, and tools
- No downloads, no hardware acquisition, no installation
- Free access:
  - https://software.intel.com/content/www/us/en/develop/tools/devcloud.html

QUESTIONS?

www.anl.gov