Argonne NATIONAL LABORATORY

INTREPID TO AURORA, THE EVOLUTION OF HPC ARCHITECTURES AT THE ALCF

SCOTT PARKER
Lead, Performance Engineering Team
Argonne Leadership Computing Facility

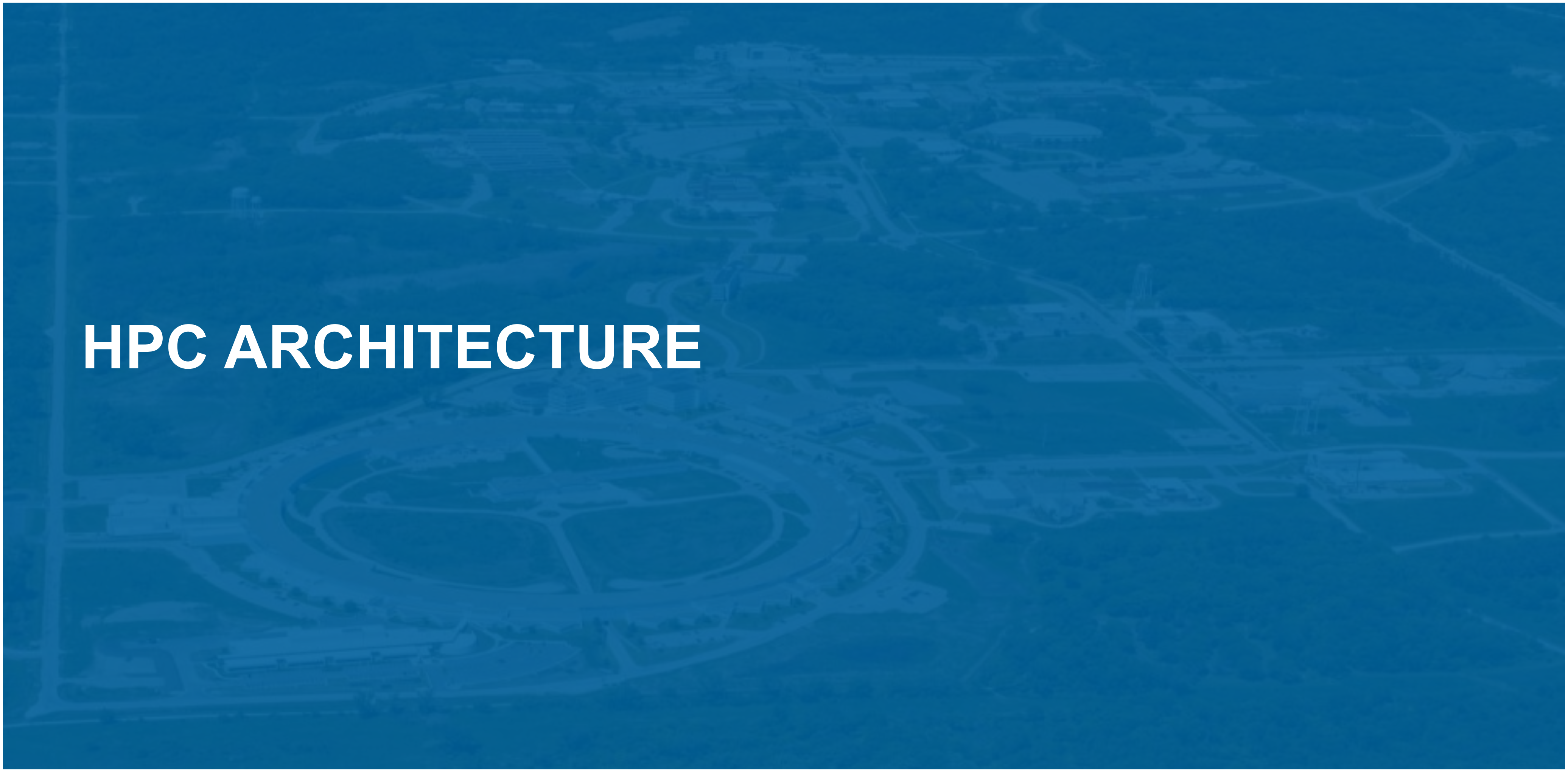
July 27, 2020

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 Denard 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

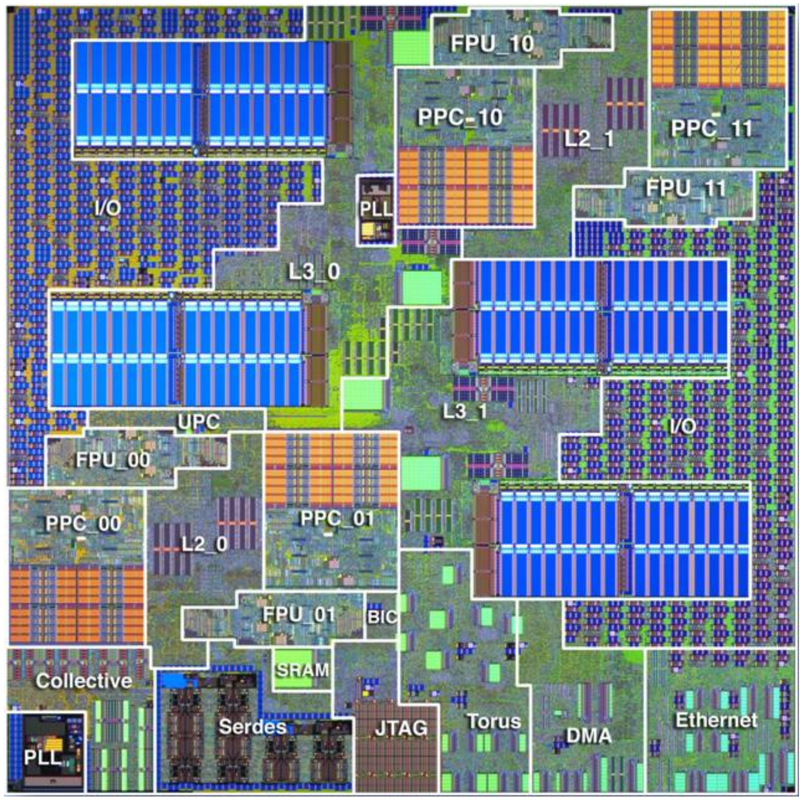
INTREPID: IBM BLUE GENE/P POWERPC 450

INTREPID

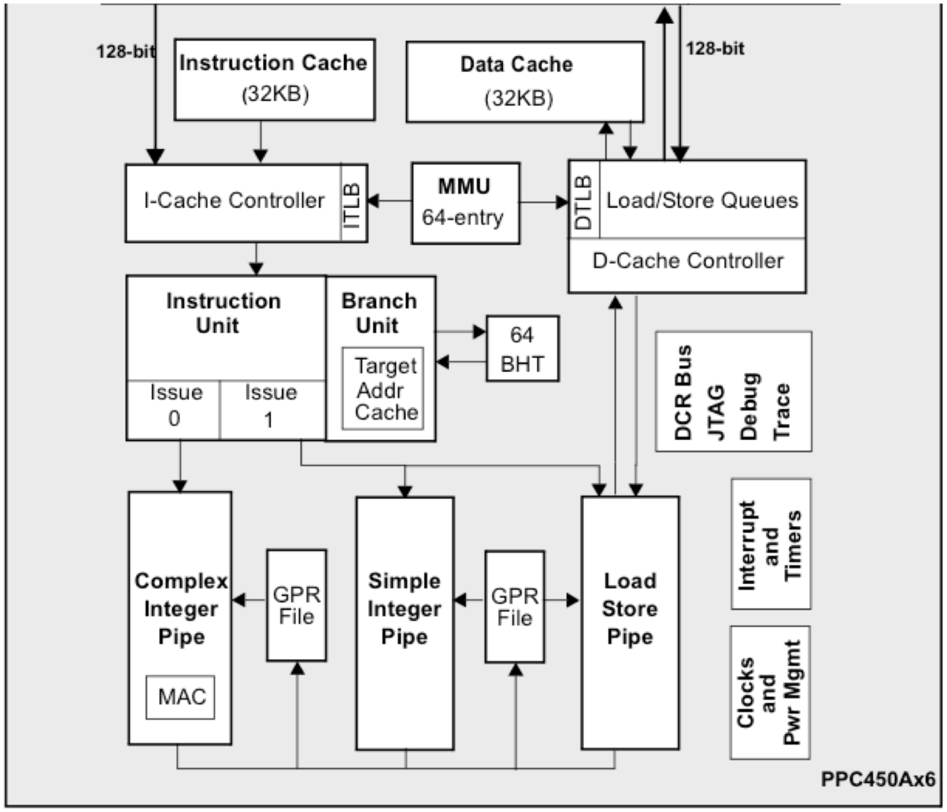
- 2008 ALCF Blue Gene/P System:
  - **40,960 nodes / 163,840 PPC cores**
  - 80 Terabytes of memory
  - Peak flop rate: 557 Teraflops
  - Linpack flop rate: 450.3
  - #6 on the Top500 list
- Storage:
  - 8 Petabytes of disk storage with an I/O rate of 80 GB/s
  - 8 Petabytes of archival storage (10,000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

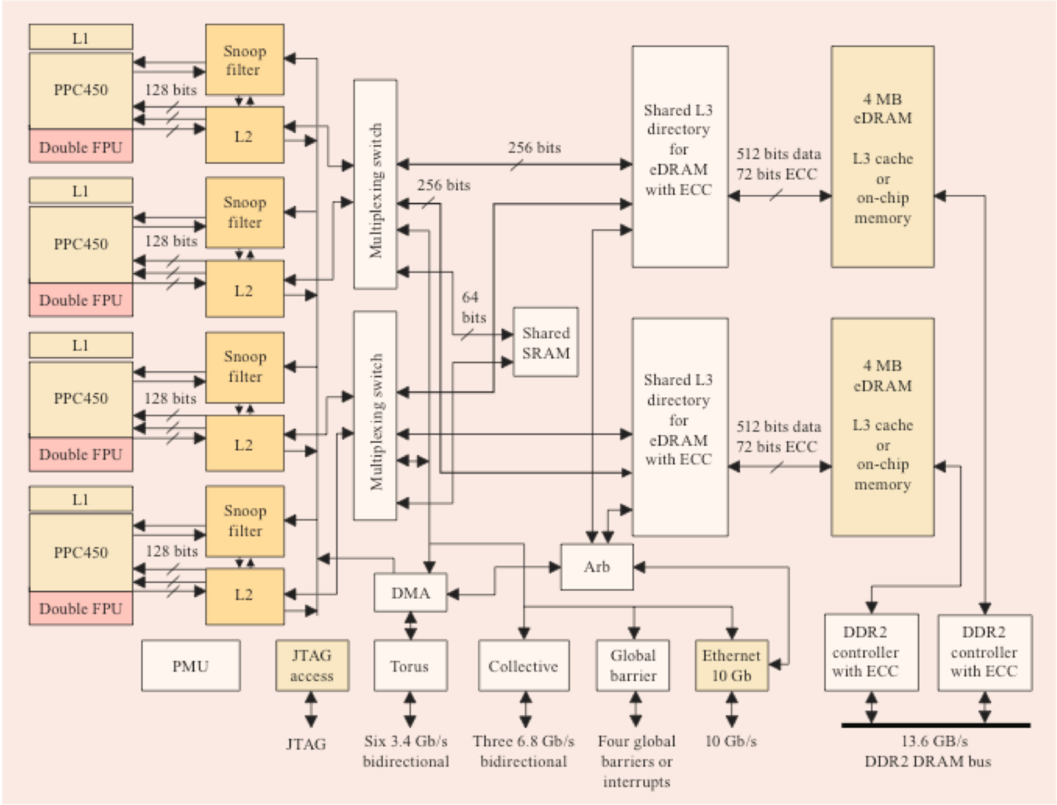


POWERPC 450 CPU



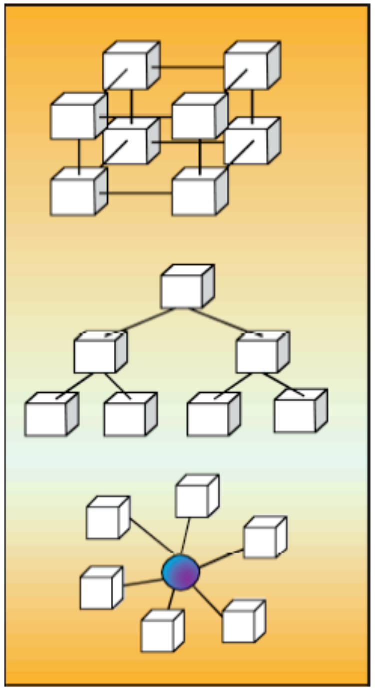
- **In order execution**
- **Dual Issue – can issues two instructions per cycle, must be to different pipelines**
- **Two wide floating point vector instructions**
- **Four Execution Pipelines:**
  - Load/Store (L-Pipe)
  - Simple Integer (J-Pipe)
  - Complex Integer (I-Pipe)
  - Floating Point
    - 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)

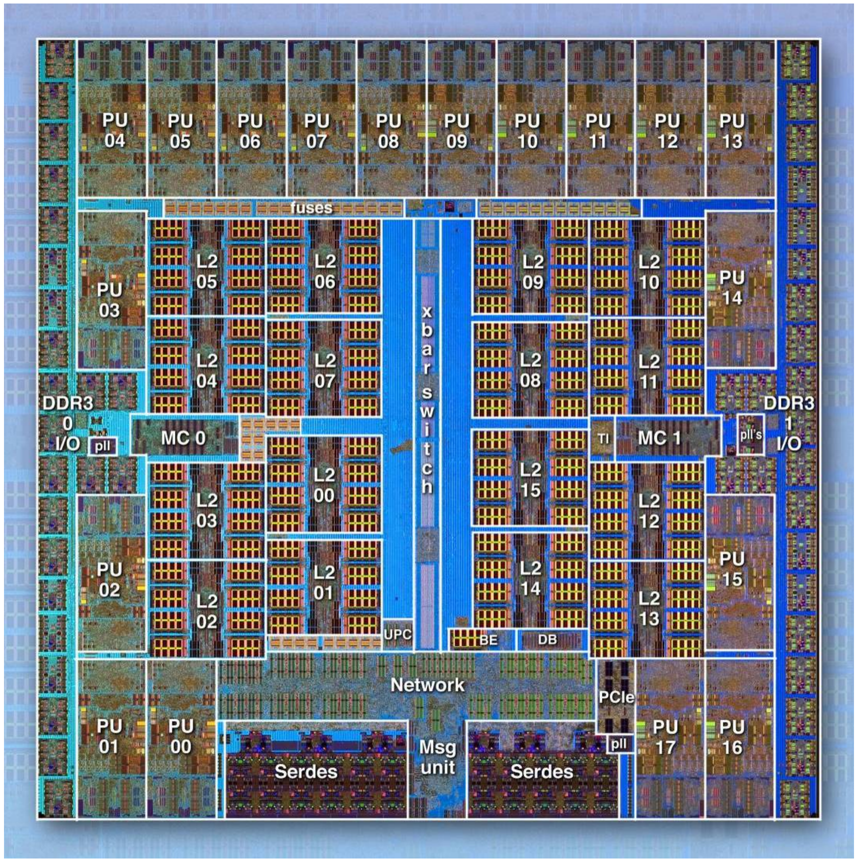
- Service node/system management

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 mm^2 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

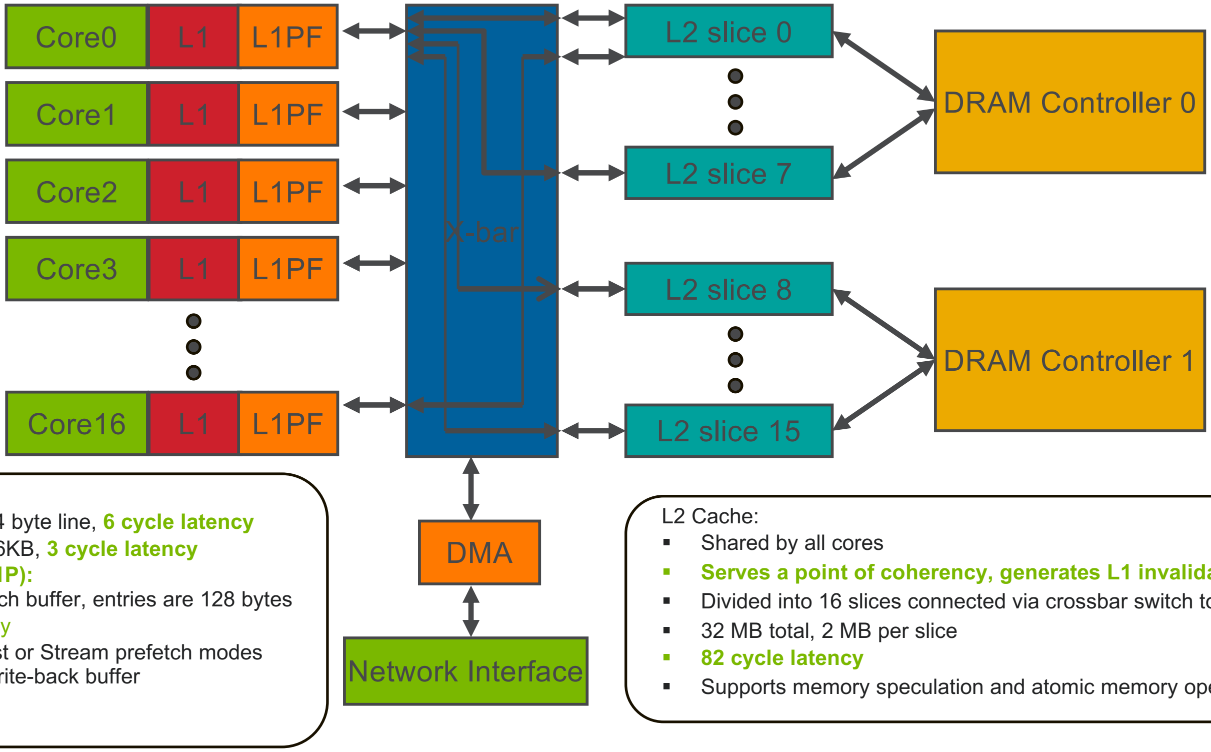
- Router logic integrated into BQC chip
- DMA, remote put/get, collective operations
- 11 network ports

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**



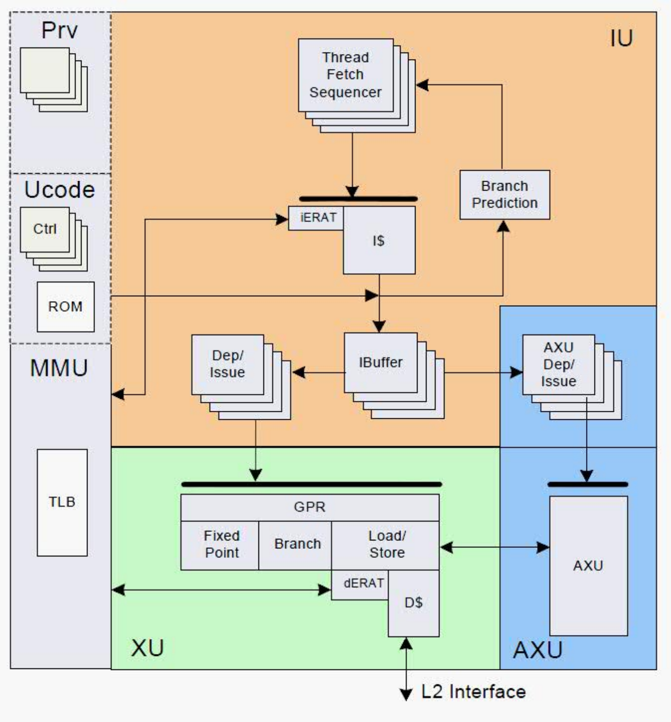
L1 Cache:
- **Data**: 16KB, 64 byte line, **6 cycle latency**
- **Instruction**: 16KB, **3 cycle latency**

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:

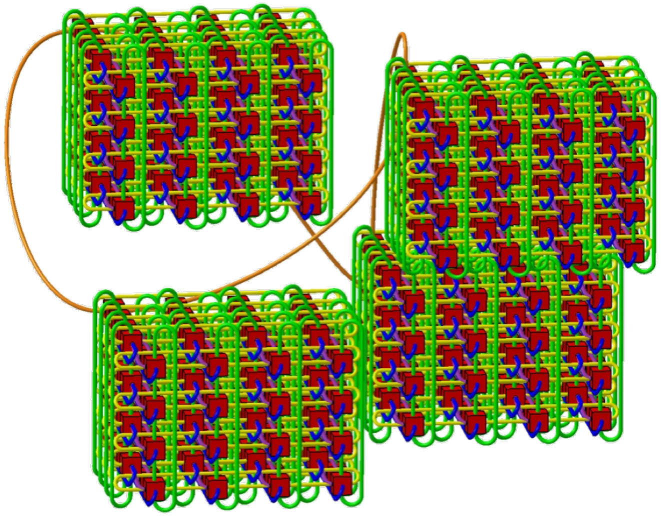
- IU – instructions fetch and decode
- XU – Branch, Integer, Load/Store instructions
- AXU – Floating point instructions
  - Standard PowerPC instructions
  - QPX 4 wide SIMD
- MMU – memory management (TLB)

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

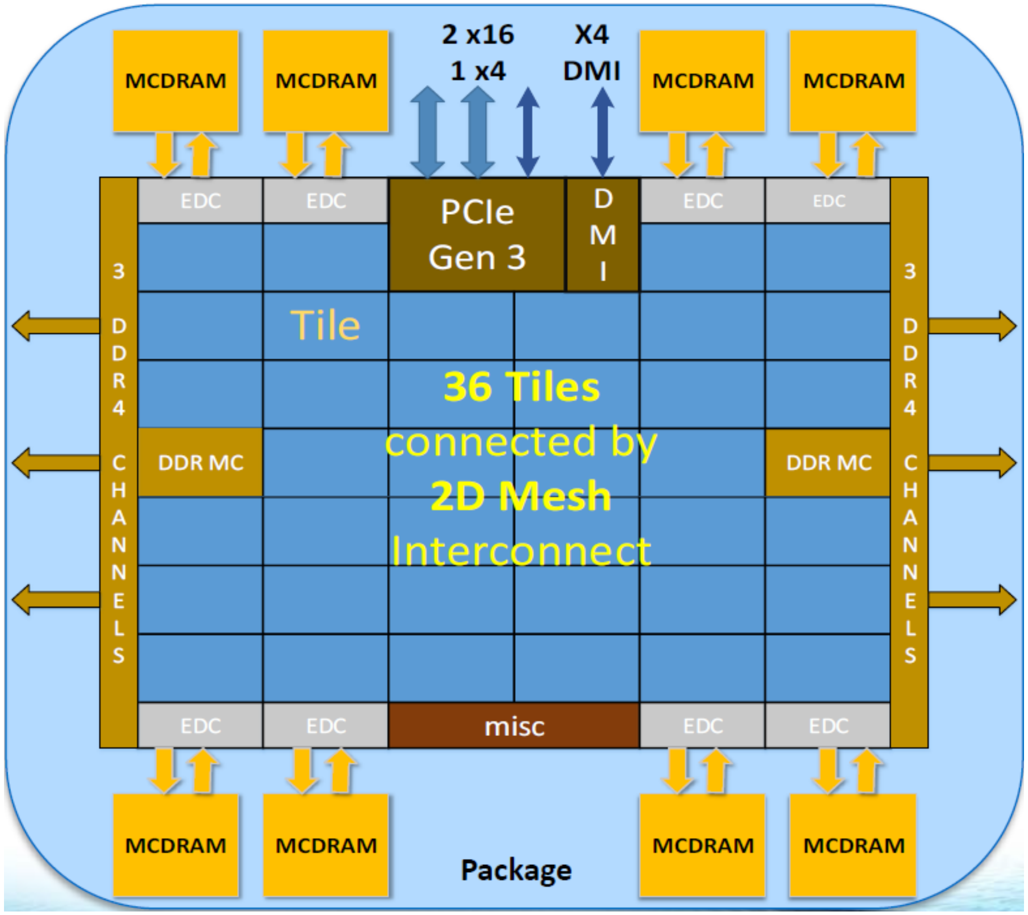


THETA: INTEL XEON PHI KNIGHTS LANDING

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)

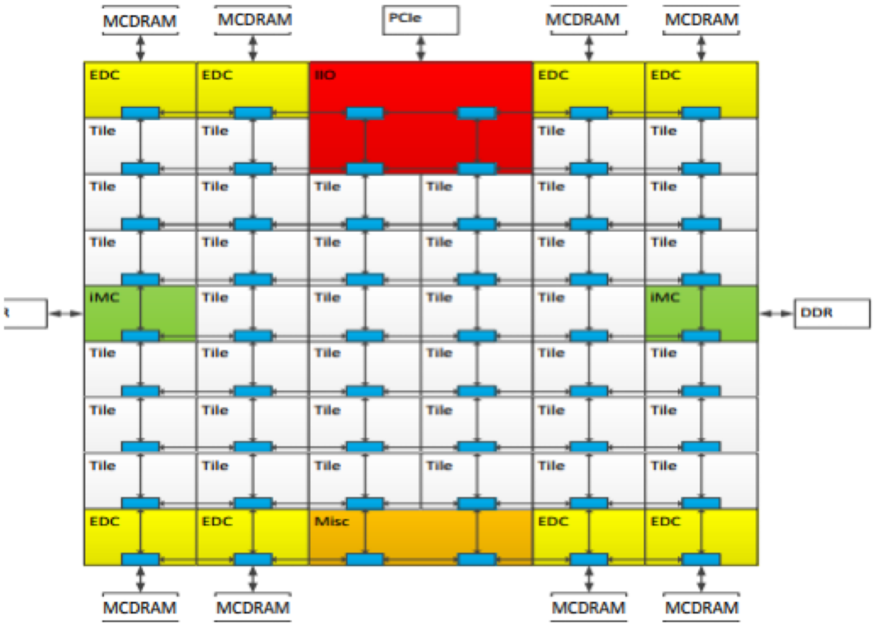
2D Mesh Interconnect
- Tiles connected by 2D mesh

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

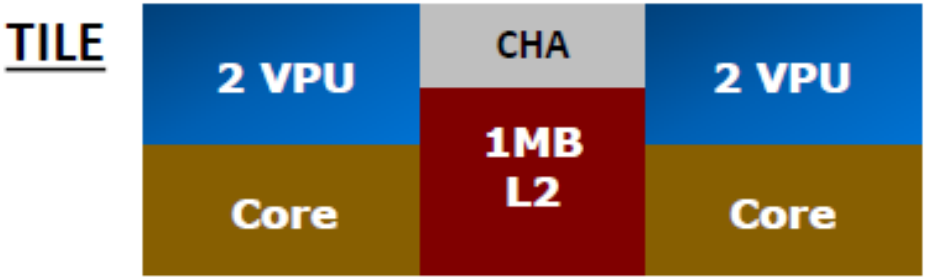
On Socket Networking
- Omni-Path NIC on package
- Connected by PCIe

KNL Mesh Interconnect



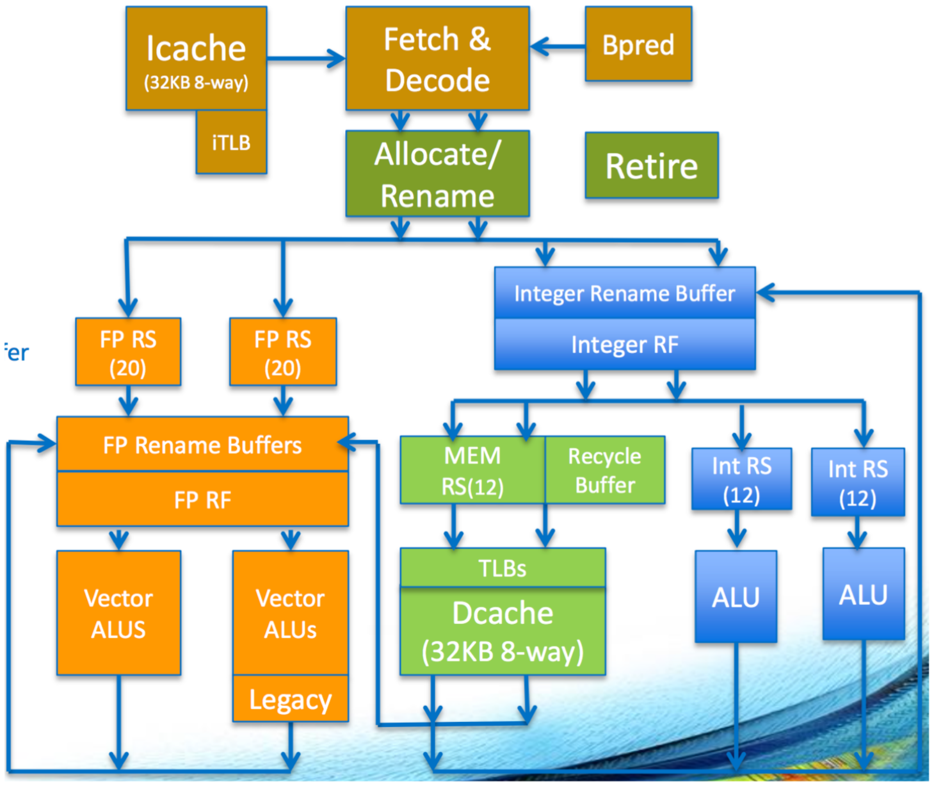
- 2D mesh interconnect connects
  - Tiles (CHA)
  - MCDRAM controllers
  - DDR controllers
  - Off chip I/O (PCIe, DMI)
- YX routing:
  - Go in Y→ turn → Go in X
  - Messages arbitrate on injection and on turn
- Cache coherent
  - Uses MESIF protocol
- Clustering mode allow traffic localization
  - All-to-all,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

KNL TILE



- 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

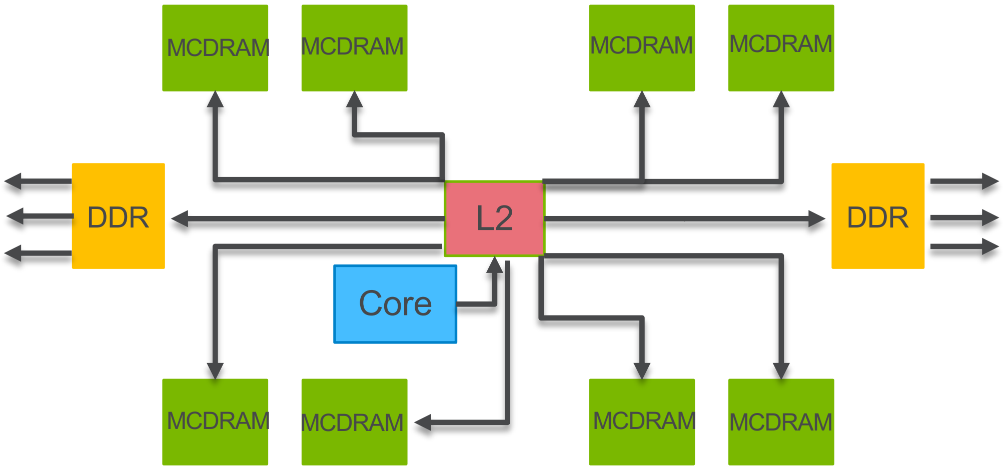
KNL CORE



- 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

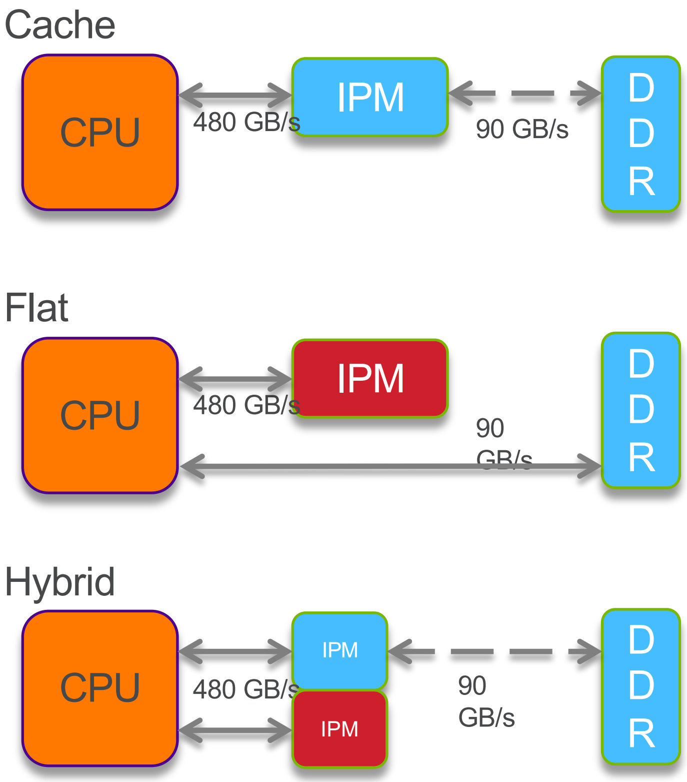
MEMORY

- **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



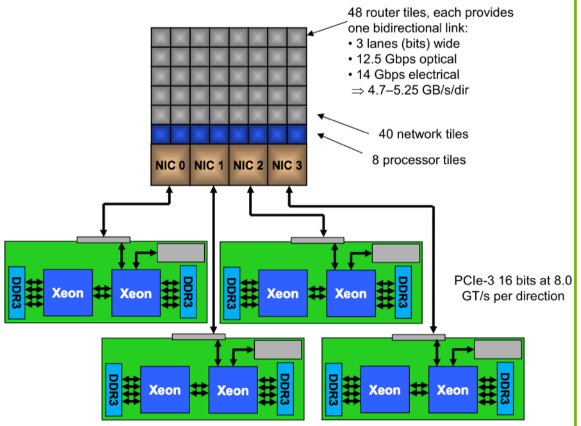
- **Memory configurations**
  - Cached:
    - DDR fully cached by IPM
    - No code modification required
    - Less addressable memory
    - Bandwidth and latency worse than flat mode
  - Flat:
    - Data location completely user managed
    - Better bandwidth and latency
    - More addressable memory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

Argonne NATIONAL LABORATORY

ARIES DRAGONFLY NETWORK

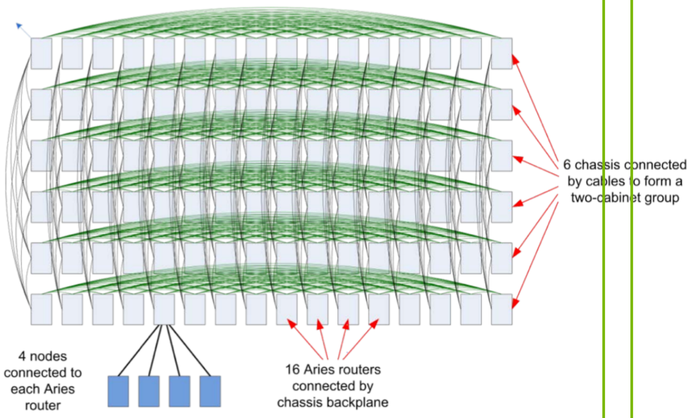
Aries Router:

- 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



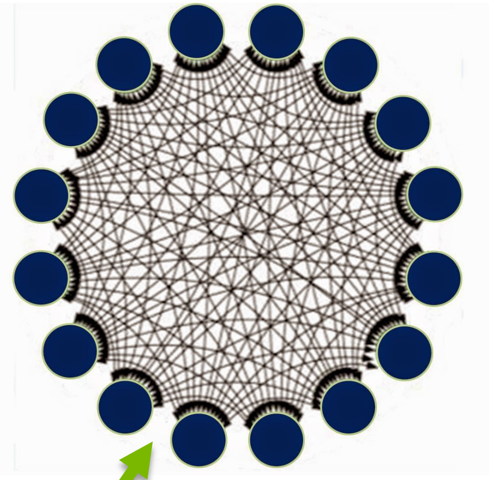
Connections within a group:

- 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

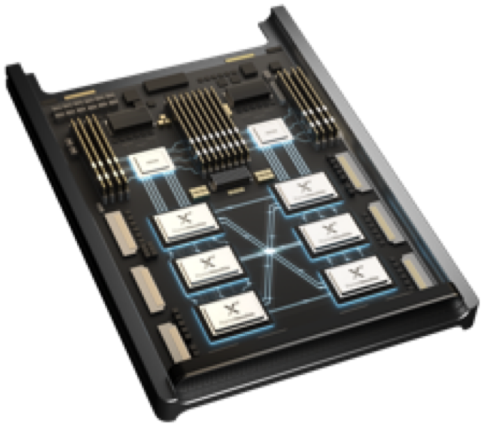


Theta has 12 groups with 12 links between each group

AURORA: INTRODUCING THE INTEL X^{E} GPU

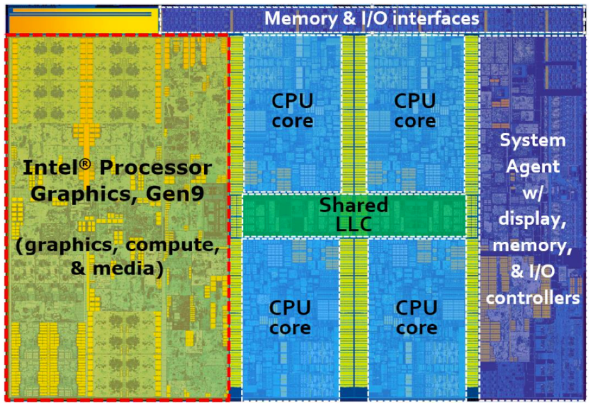
AURORA: A HIGH-LEVEL VIEW

- Intel-Cray machine arriving at Argonne in 2021
  - 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

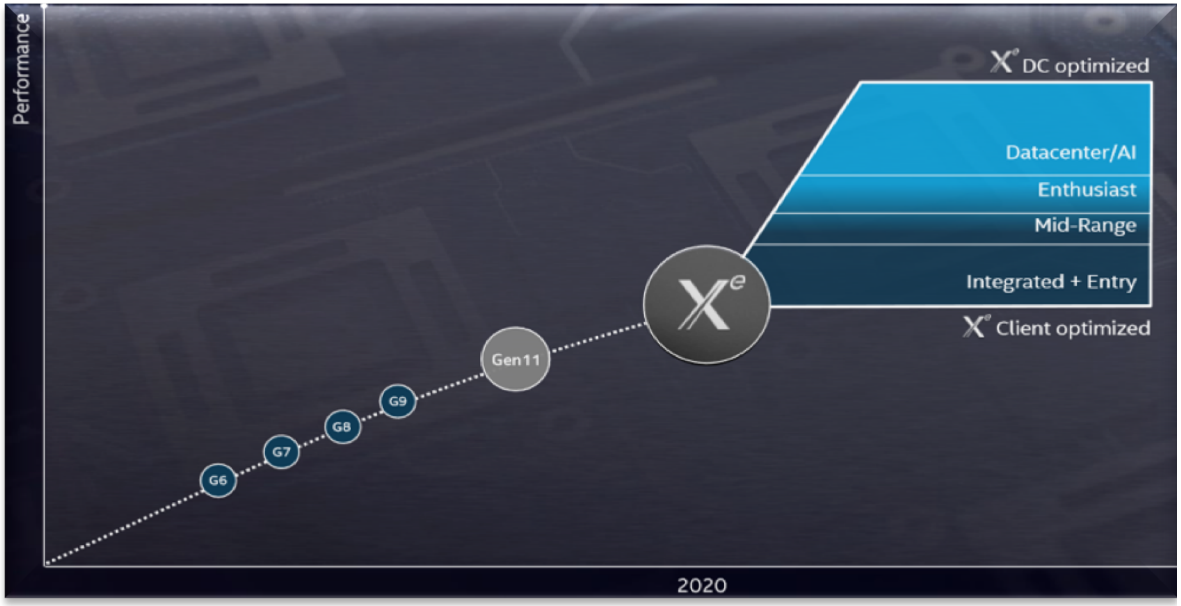


INTEL GPUS

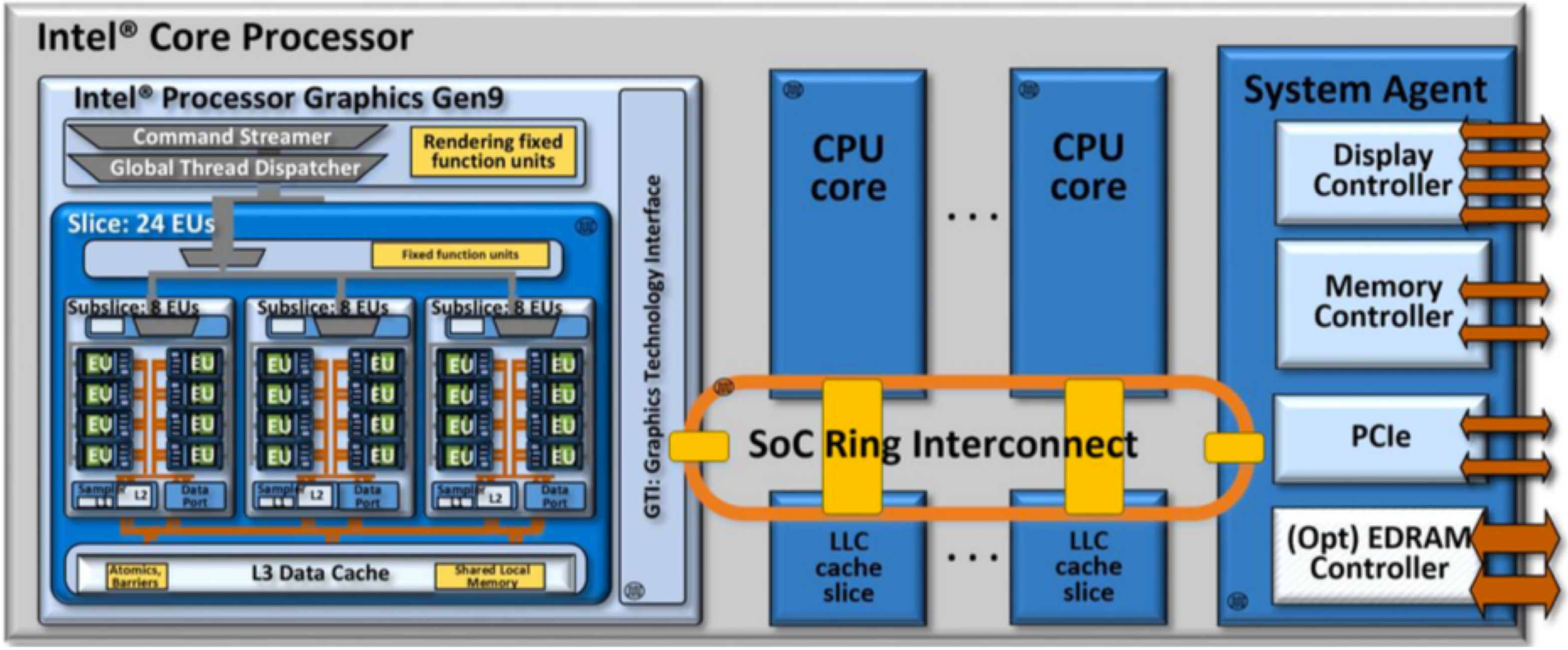
- Intel has been building GPUs integrated with CPUs for over a decade
- Currently released products use the Gen and Gen 11 versions
  - Gen9 – used in Skylake
  - Gen11 – used in Ice Lake
- Low performance by design due to power and space limits
  - Gen9 peak DP flops: 100-300 GF
  - Gen 9 introduce in 2015
- Next is the X^e (Gen 12) line of integrated and discrete GPUs


Memory & I/O interfaces
Intel® Processor Graphics, Gen9
(graphics, compute, & media)
CPU core
CPU core
Shared LLC
CPU core
CPU core
System Agent w/ display, memory, & I/O controllers

Architecture components layout for an Intel Core i7



INTEL INTEGRATED GRAPHICS

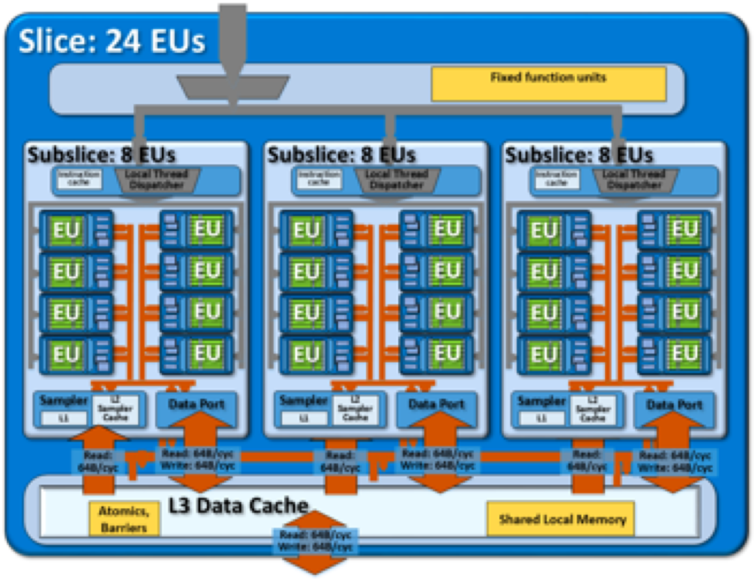


- Cores and GPU on the same chip and connected by a ring interconnect
- Same memory used by CPU and GPU
- Shared Last Level Cache

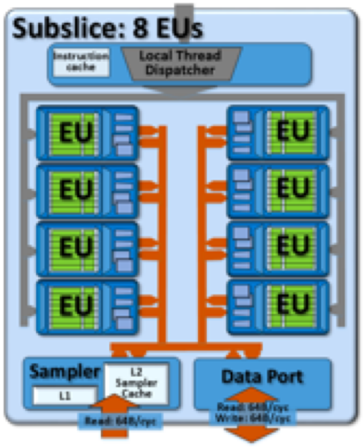
INTEL GEN9 ARCHITECTURE HIERARCHY

- GPU architectures have hardware hierarchies
  - Built out of smaller scalable units, each level shares a set of resources

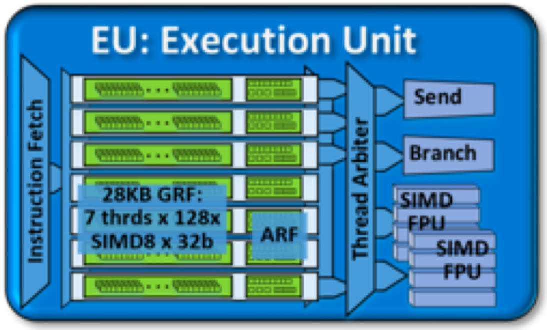
Slice



Sub-Slice

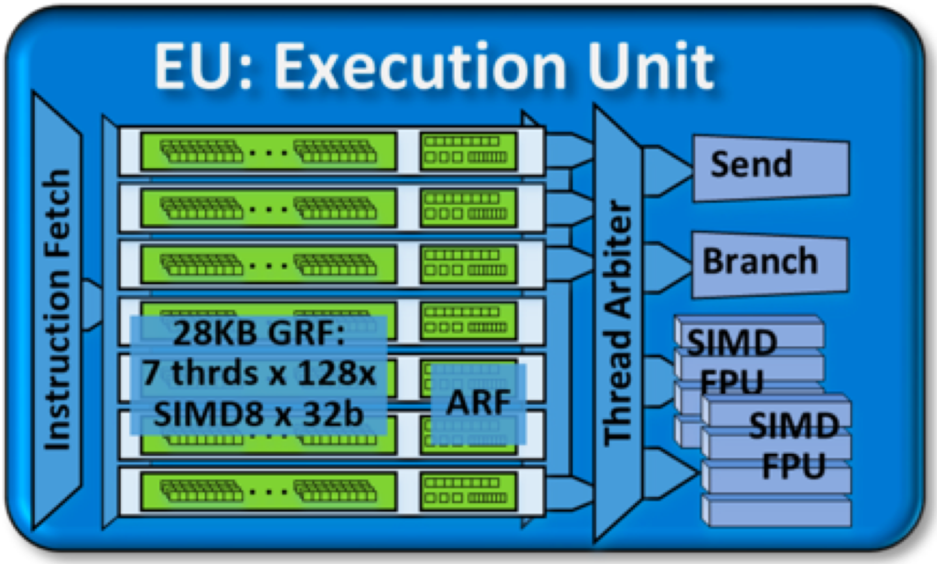


Execution Unit (EU)



INTEL GEN9 BUILDING BLOCKS: EU

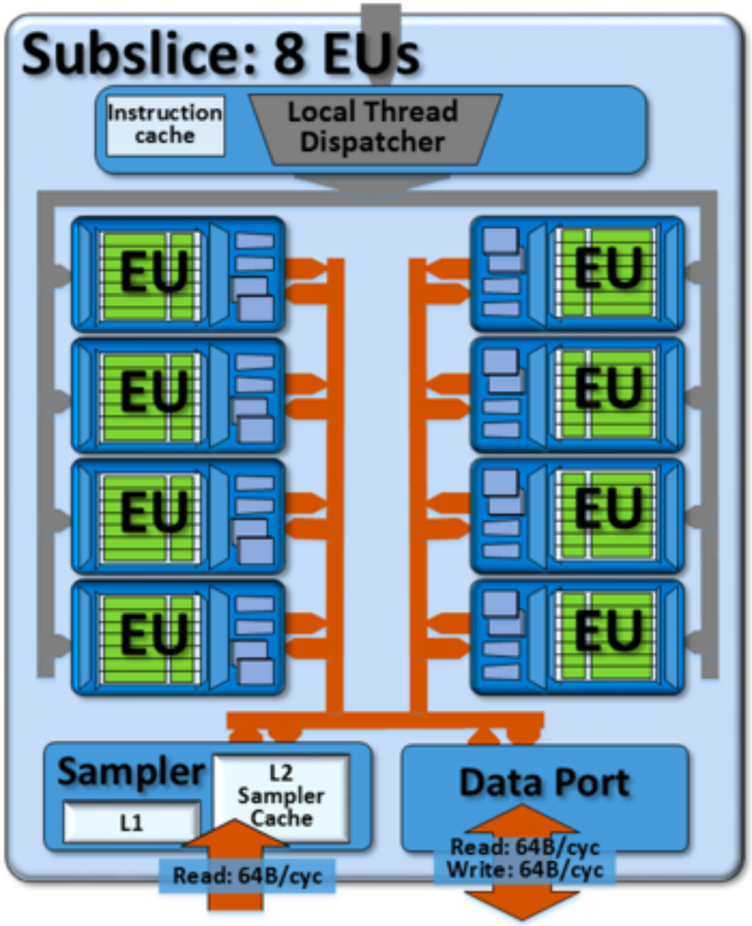
EU: Execution Unit



- Compute processors that executes instructions
- In-order execution
- 7 hardware threads, each with own inst. pointer
- Can issue 4 instructions per cycle (different threads)
- 4 functional units
  - 2 SIMD "FPU"s
    - Floating point instructions
    - Integer instructions
    - One unit performs double precision
    - Fully pipelined
    - Supports FMA
  - Branch
  - Send (memory load/store)
- 28 KB register file
- SIMD instructions
  - SIMD hardware is 128 bits wide
    - 2 DP, 4 SP, 8 HP, …
  - Instruction SIMD width can vary (1-32)

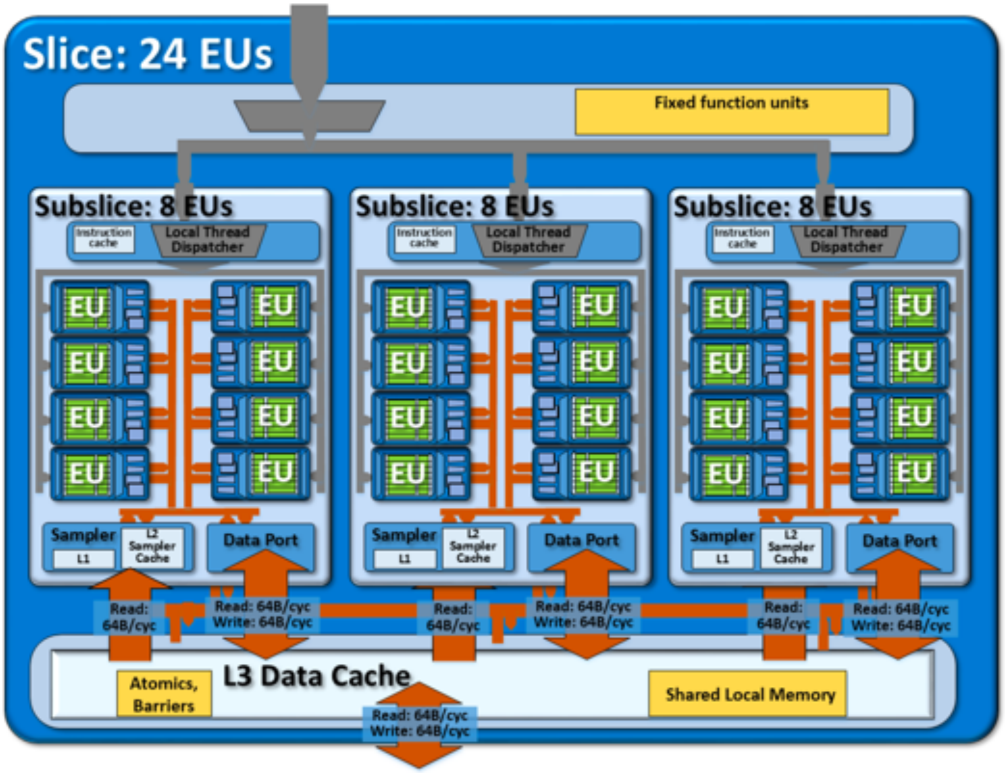
IDF-2015 release of "The Compute Architecture of Intel Processor Graphics Gen9" – Stephen Junkins

INTEL GEN9 BUILDING BLOCKS: SUBSLICE



- A sub-slice contains:
  - 8 EUs
  - Instruction cache
  - Local thread dispatcher
  - L1 & L2 sampler caches
  - Data port (memory load/store)
    - Efficient read/write operations
    - Provides scatter/gather
    - Shared local memory

Intel Gen9 Building Blocks: Slice



A slice contains:

- 3 subslices
- L3 cache
- Shared local memory
- Fixed functional units
- Slices are a scalable architectural unit
  - Products available with 1-3 slices
- Slices are connected at the L3

GEN9 (GT4) GPU CHARACTERISTICS

Characteristics	Value	Notes
Clock Freq.	1.15 GHz	
Slices	3	
EUs	72	3 slice * 3 sub-slices * 8 EUs
Hardware Threads	504	72 EUs * 7 threads
Concurrent Kernel Instances	16,128	504 threads * SIMD-32
L3 Data Cache Size	1.5 MB	3 slices * 0.5 MB/slice
Max Shared Local Memory	576 KB	3 slice * 3 sub-slices * 64 KB/sub-slice
Last Level Cache Size	8 MB	
eDRAM size	128 MB	
32b float FLOPS	1152 FLOPS/cycle	72 EUs * 2 FPUs * SIMD-4 * (MUL + ADD)
64b float FLOPS	288 FLOPS/cycle	72 EUs * 1 FPU * SIMD-2 * (MUL + ADD)
32b integer IOPS	576 IOPS/cycle	72 EUs * 2 FPUs * SIMD-4

331.2 DP GFlops

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

Argonne
NATIONAL LABORATORY