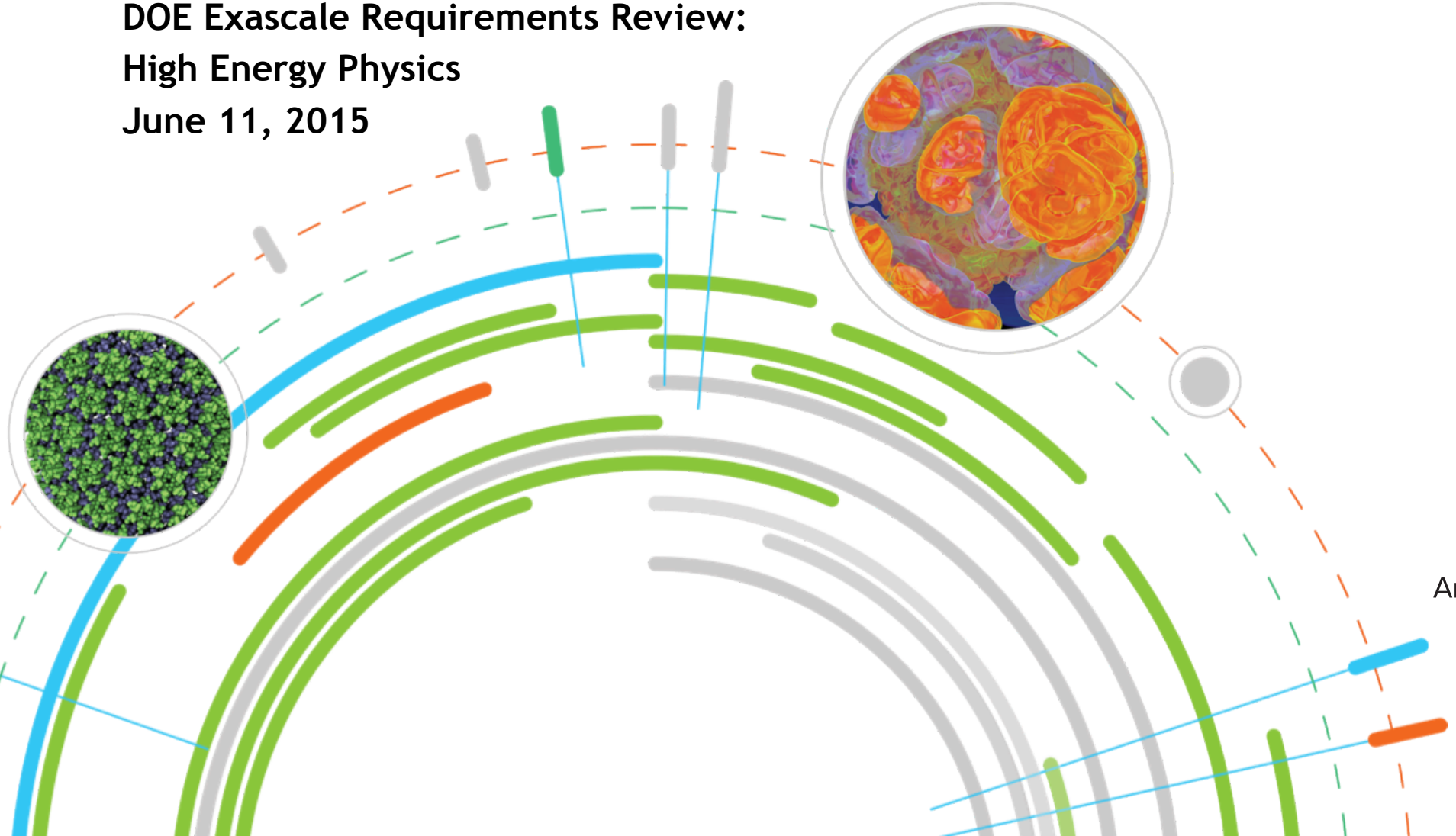
ALCF Future Systems

Tim Williams, Argonne Leadership Computing Facility

DOE Exascale Requirements Review:
High Energy Physics
June 11, 2015



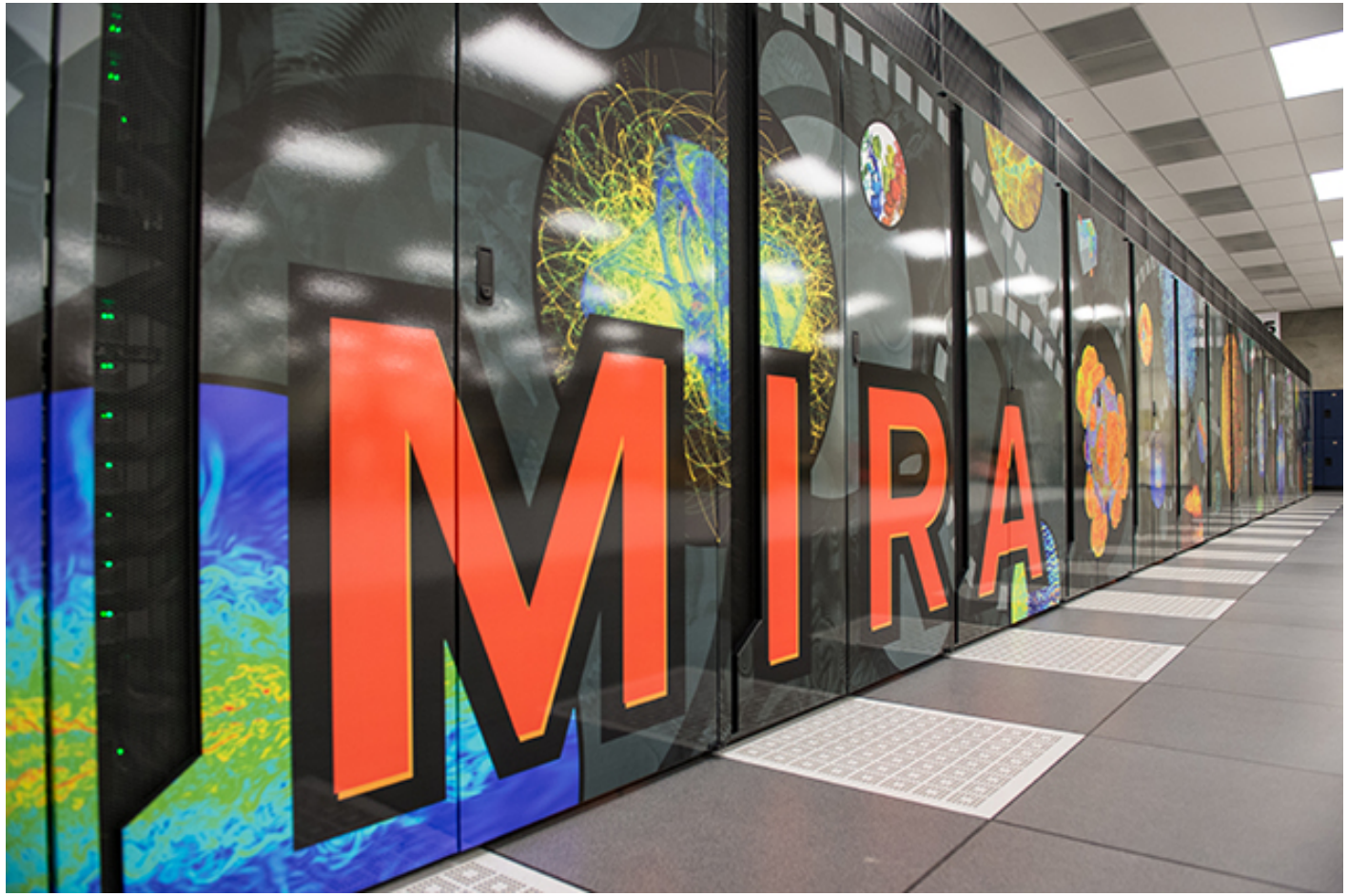
Argonne **Leadership Computing** Facility

Argonne NATIONAL LABORATORY

Production Systems (ALCF-2)

Mira - IBM Blue Gene/Q

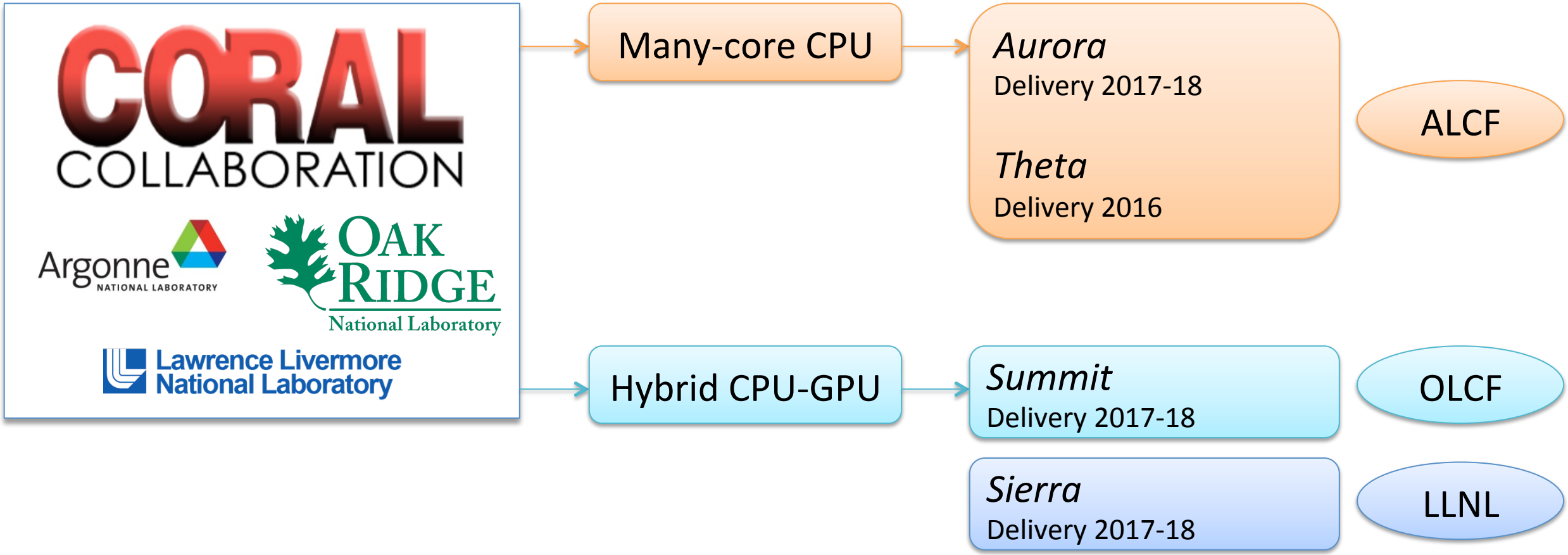
- 49,152 nodes
  - PowerPC A2 cpu
    - 16 cores, 4 HW threads/core
  - 16 GB RAM
- Aggregate
  - 768 TB RAM, 768K cores
  - Peak 10 PetaFLOPS
- 5D torus interconnect



Cooley - Viz/Analysis cluster

- 126 nodes:
  - Two 2.4 GHz Intel Haswell 6-core
    - 384 GB RAM
  - NVIDIA Tesla K80 (two GPUs)
    - 24 GB GPU RAM
- Mounts *Mira* file system (GPFS)

Next-Generation ALCF-3 System



Aurora, Summit are LCF's **pre-exascale** systems.

Aurora

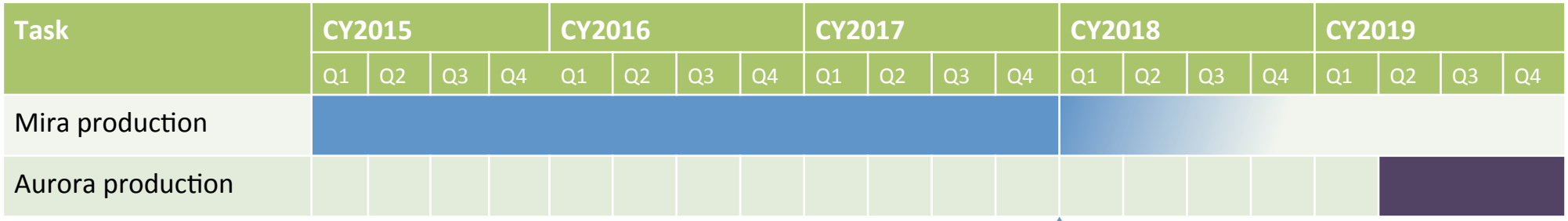
- Homogeneous
- Many-core
- Self-hosted
- Water cooled

- 18x *Mira* speed
- 2.7x *Mira* peak power consumption
- Similar node count to *Mira*
- Intel Architecture (x86-64) Compatibility

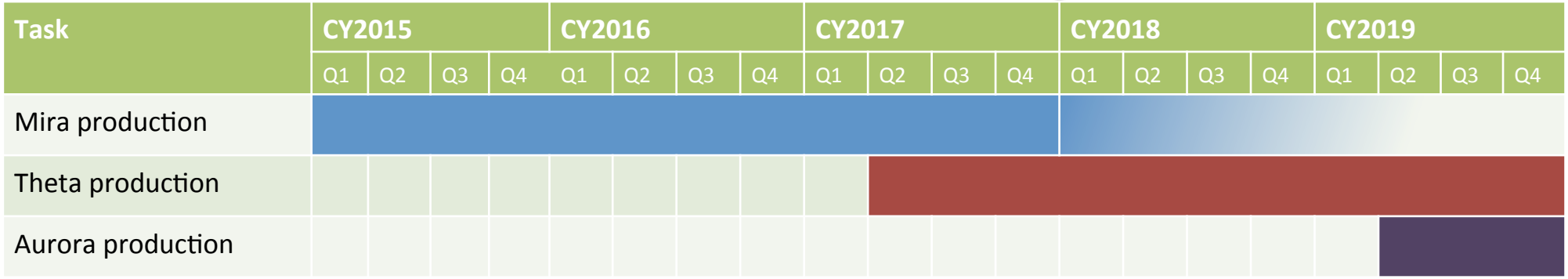
Aurora Details

System Feature	*Aurora*
Peak System performance (FLOPs)	180 - 450 PetaFLOPS
Processor	3rd Generation Intel® Xeon Phi™ processor (code name Knights Hill)
Number of Nodes	>50,000
Compute Platform	Cray Shasta next generation supercomputing platform
High Bandwidth On-Package Memory, Local Memory, and Persistent Memory	>7 Petabytes
System Interconnect	2nd Generation Intel® Omni-Path Architecture with silicon photonics
Interconnect interface	Integrated
Burst Storage Buffer	Intel® SSDs, 2nd Generation Intel® Omni-Path Architecture
File System	Intel Lustre* File System
File System Capacity	>150 Petabytes
File System Throughput	>1 Terabyte/s
Peak Power Consumption	13 Megawatts
FLOPS/watt	>13 GFLOPS/watt
Delivery Timeline	2018
Facility Area	~3,000 sq. ft.

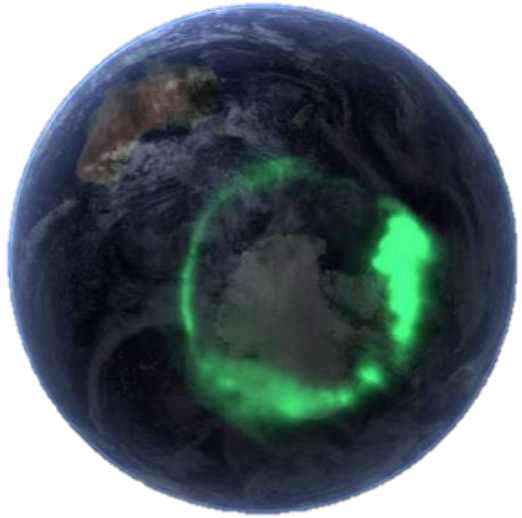
Transition to Aurora



Mira planned end-of-life:



Theta



- Homogeneous
- Many-core
- Self-hosted
- Water cooled

- 0.85x *Mira* speed
- 0.35x *Mira* peak power consumption
- >2500 nodes
- Intel Architecture (x86-64) Compatibility

Theta Details

System Feature	*Theta*
Peak System performance (FLOPs)	>8.5 PetaFLOP/s
Processor	2nd Generation Intel® Xeon Phi™ processors (Code name: Knights Landing)
Number of Nodes	>2,500 single socket nodes
Compute Platform	Cray* XC* supercomputing platform
Compute Node Peak Performance	>3 TeraFLOP/s per compute node
Cores Per Node	>60 cores
High Bandwidth On-Package Memory	Up to 16 Gigabytes per compute node
DDR4 Memory	192 Gigabytes
On-node storage	128 GB SSD
File System	Intel Lustre* File System
System Interconnect	Cray Aries* high speed Dragonfly* topology interconnect
File System Capacity (Initial)	10 Petabytes
File System Throughput (Initial)	210 Gigabytes/s
Peak Power Consumption	1.7 Megawatts
Delivery Timeline	Mid-2016

Theta Details (cont'd)

System Feature	*Theta*
Number of Nodes	>2,500 single socket nodes
Cores Per Node	>60 cores with four hardware threads per core
High Bandwidth On-Package Memory BW	Projected to be 5X the bandwidth of DDR4 DRAM memory, >400 Gigabytes/sec
DDR4 Memory	192 Gigabytes using 6 channels per compute node
Microarchitecture	Intel® "Silvermont" enhanced for HPC, with 2X the out-of-order buffer depth of current Silvermont, gather/scatter in hardware, advanced branch prediction, 32KB Icache and dcache, 2 x 64B load ports in dcache, and 46/48 physical/virtual address bits to match Xeon
Vector functionality	AVX512 vector pipelines with a hardware vector length of 512 bits (eight double-precision elements)
Interconnect details	Processor cores connected in a 2D mesh network with 2 cores per tile, with a 1-MB cache-coherent L2 cache shared between 2 cores in a tile, with two vector processing units per core, and with multiple NUMA domain support per socket

Programming the *Mira* → *Theta* → *Aurora* Systems

- Most common: MPI + OpenMP
- *Theta* & *Aurora*: OpenMP 4.x
  - `simd` pragma: portable control of vectorization
  - `target` pragma: directives-based GPU parallelism

Theta vs. *Mira* Run Configurations

Run entirely in IPM on *Theta*

Mira Node

Ranks	Threads	RAM per rank
16	4	1 GB
8	8	2 GB

Theta Node (assume 60 cores)

Ranks	Threads	IPM per rank	BW
15	16	1.06 GB	>9.3x *Mira*
10	24	1.60 GB	>7.4x *Mira*

BW = memory bandwidth per rank

Use DDR4 on *Theta*

Mira Node

Ranks	Threads	RAM per rank
16	4	1 GB
8	8	2 GB

Theta Node (assume 60 cores)

Ranks	Threads	DDR4 per rank	BW
15	16	13.86 GB	>1.99x *Mira*
10	24	20.8 GB	>1.49x *Mira*

Applications Readiness: ALCF-3 Early Science Program

- Prepare applications for architecture & scale of next-generation ALCF systems
- Two phases:

Theta

- 6 projects
  - 3 pre-selected
  - 3 chosen via CFP
- 4 funded postdocs

Aurora

- 10 projects
  - Chosen via CFP
- 10 funded postdocs

ESP Project Support

SUPPORT

- Funded postdoctoral appointee
- Catalyst staff member support
- Planned ALCF-vendor center of excellence

TRAINING

- Training on HW and programming
  - Virtual Kick-Off workshop
  - Hands-on workshop using early hardware
- Community workshop to share lessons learned

COMPUTE RESOURCES

- Time on current systems (*Mira*) for development
- Advanced simulator access
- Precursor-hardware tiny test systems (JLSE)
- Earliest possible hardware access
- 3 months dedicated Early Science access
  - Pre-production (post-acceptance)
  - Large time allocation
  - Continued access for rest of year