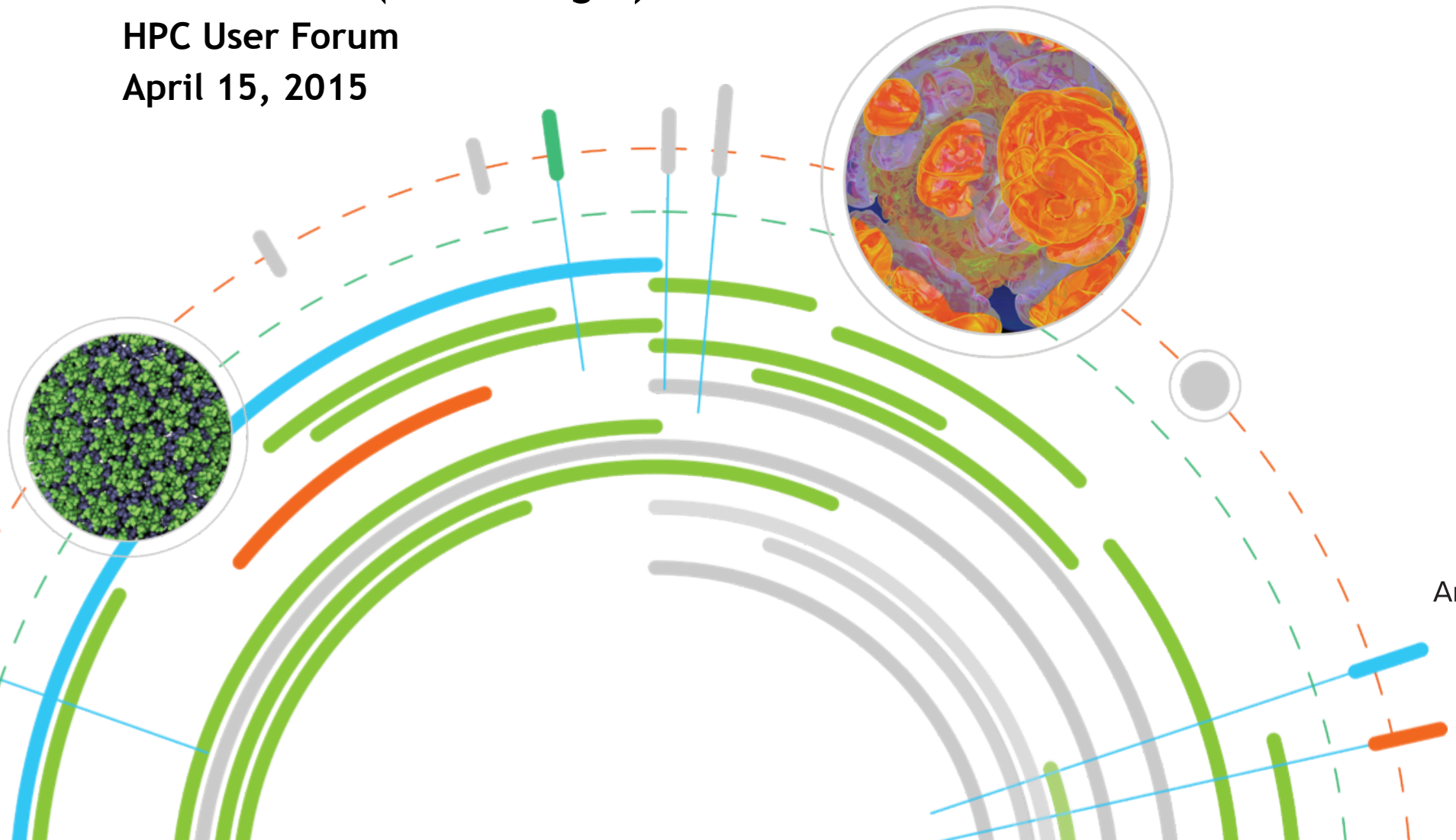
ALCF Early Science Program

Tim Williams *(ESP Manager)*

HPC User Forum

April 15, 2015

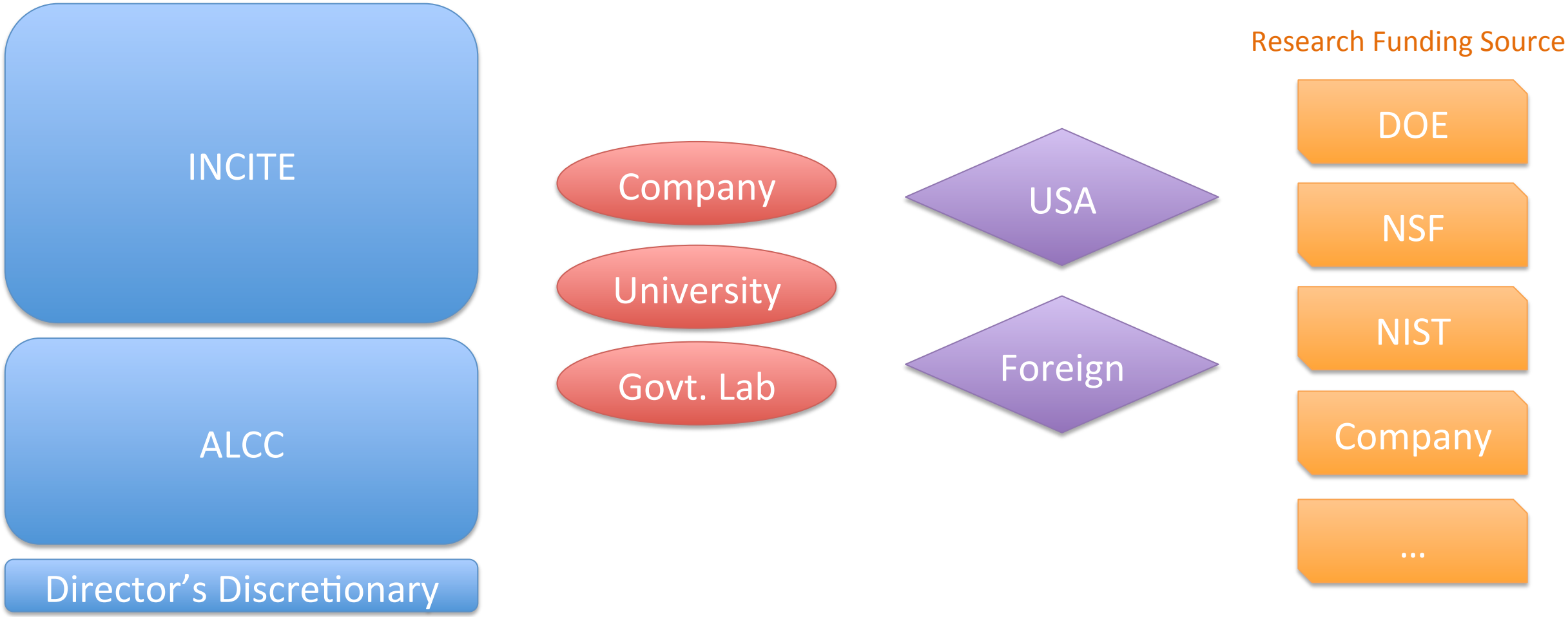


Argonne **Leadership Computing** Facility

Argonne NATIONAL LABORATORY

Argonne Leadership Computing Facility

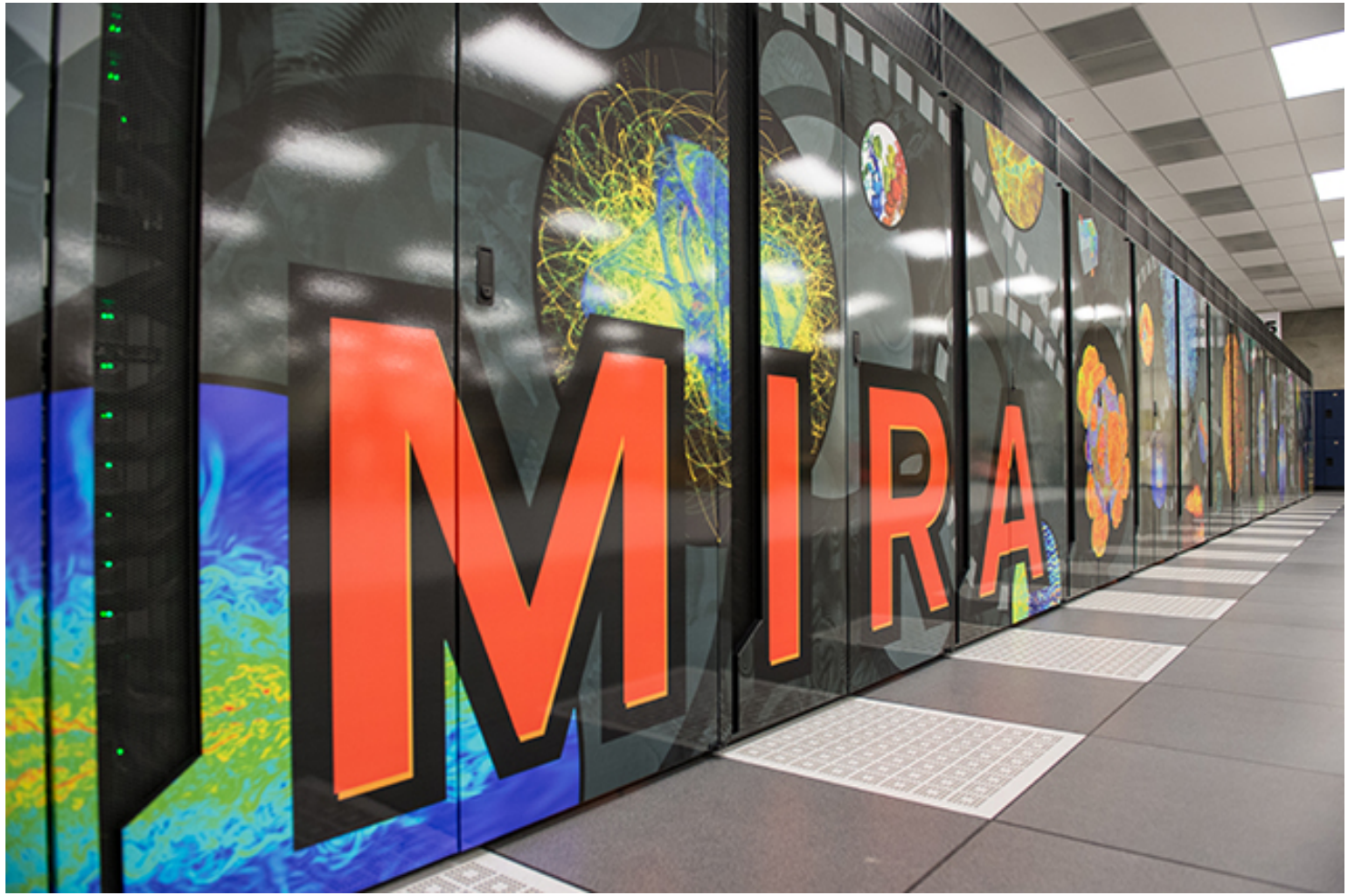
- Mission: *capability* computing (*leadership-class*)



Production Systems (ALCF-2)

Mira – IBM Blue Gene/Q system

- 49,152 nodes / 786,432 cores
- PowerPC A2 cpu – 16 cores, 4 HW threads/core
- 786 TB of memory
- Peak flop rate: 10 PF



Cetus – BG/Q system

- 4096 nodes / 65,536 cores
- 16 TB of memory, 836 TF peak
- Mounts *Mira* file system

Vesta - BG/Q system

- 2048 nodes / 32,768 cores
- 32 TB of memory, 419 TF peak

Tukey – NVIDIA system

- 96 nodes / 1536 AMD CPU cores
- 192 NVIDIA Tesla M2070 GPUs
- 6 TB memory / 1.1TB GPU memory
- GPU Peak flop rate: 99 TF
- Mounts *Mira* file system

Storage - Projects: 28.8 PB raw capacity, 240 GB/s bw (GPFS); Home: 1.8 PB; Tape: 16 PB of archival storage, 15,906 volume tape archive (HPSS)

Next-Generation ALCF-3 System

CORAL partner (ANL, LLNL, ORNL)

- RFP for two architecturally distinct systems
- One architecture: OLCF-4 system *Summit*
- Other architecture: ALCF-3 system *Aurora*
- LLNL picks one: *Sierra*
- Significant speed increase from today's systems

Delivery of *Aurora* system: 2017-18

Delivery of *Theta* system: 2016

Aurora

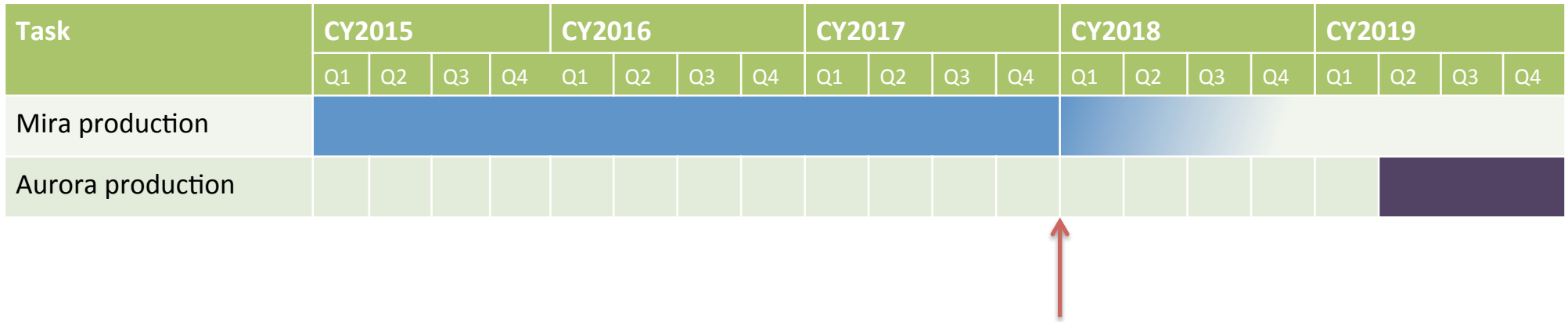
- Homogeneous
- Many-core
- Self-hosted
- Water cooled

- 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

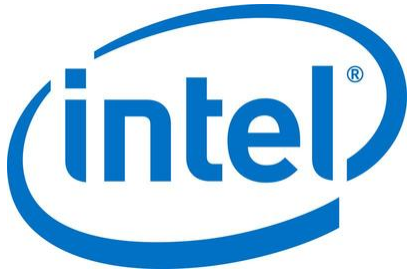


Transition to Aurora

Task	CY2015				CY2016				CY2017				CY2018				CY2019			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Mira production																				
Aurora production																				

Task	CY2015				CY2016				CY2017				CY2018				CY2019			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Mira production																				
Theta production																				
Aurora production																				

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

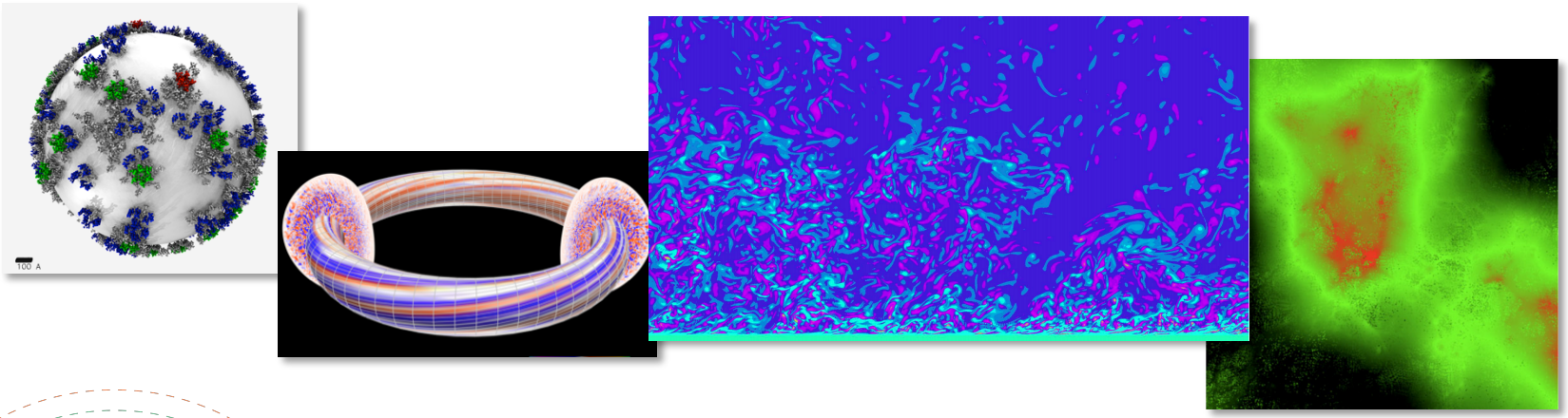
System Feature	*Theta*
Peak System performance (FLOPs)	>8.5 PetaFLOP/s
Processor	2nd Generation Intel® Xeon Phi™ processors (Code name: Knights Landing)
Number of Nodes	>2,500 single socket nodes
Compute Platform	Cray* XC* supercomputing platform
Compute Node Peak Performance	>3 TeraFLOP/s per compute node
Cores Per Node	>60 cores
High Bandwidth On-Package Memory	Up to 16 Gigabytes per compute node
DDR4 Memory	192 Gigabytes
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

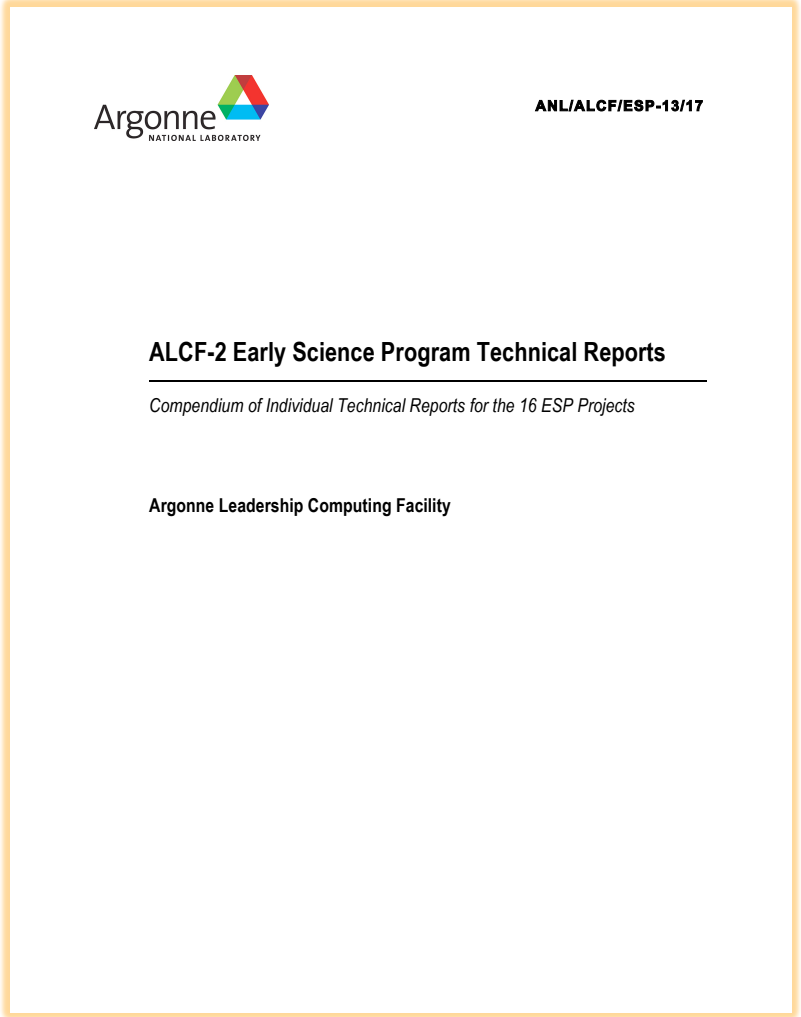
Theta Details (cont'd)

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

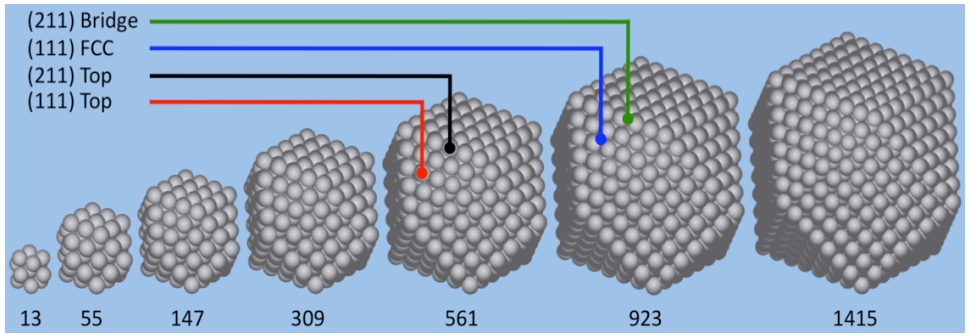
ALCF Applications-Readiness: Early Science Program

- Patterned after successful ALCF-2 (*Mira*) Early Science Program
- *Mira* ESP generated
  - Breakthrough science
  - Technical reports on code porting & tuning
  - Two open community workshops (science & technology)
  - Synergy with Tools & Libraries project
  - Stable production platform (problems shaken out)
  - Persisting culture of apps readiness for next generation
  - Success stories for postdocs





ALCF-2 ESP (Mira)

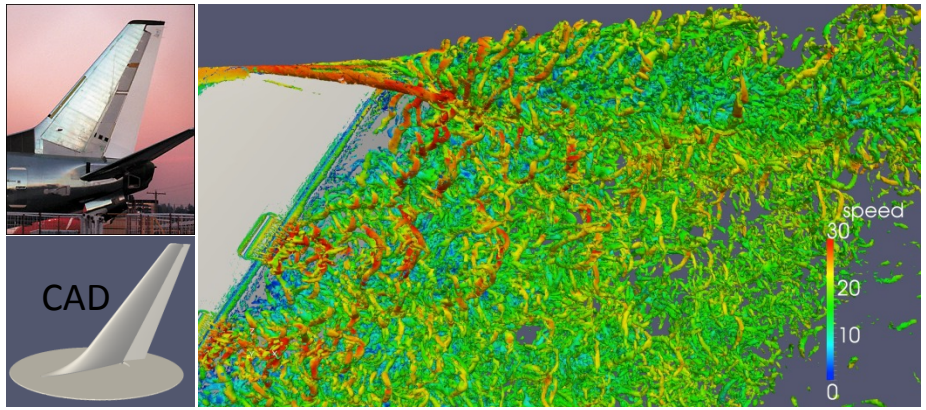


Energy Storage Materials and Catalysis

Quantum Monte Carlo simulations of platinum metal nanoparticles as catalysts for key reactions (QMCPACK).

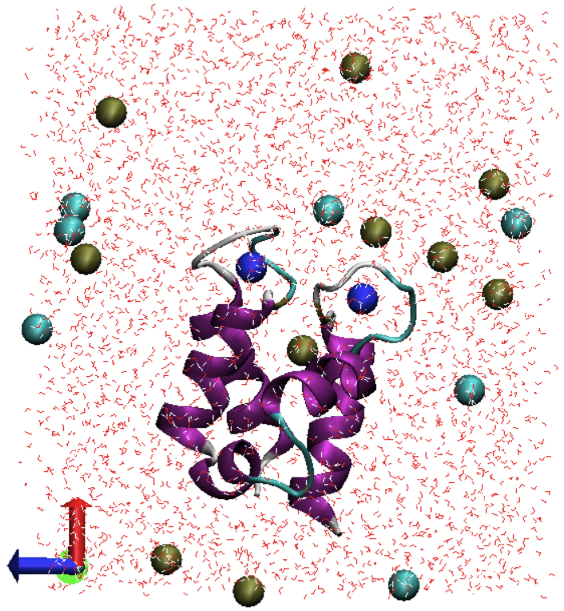
Active Aerodynamic Flow Control

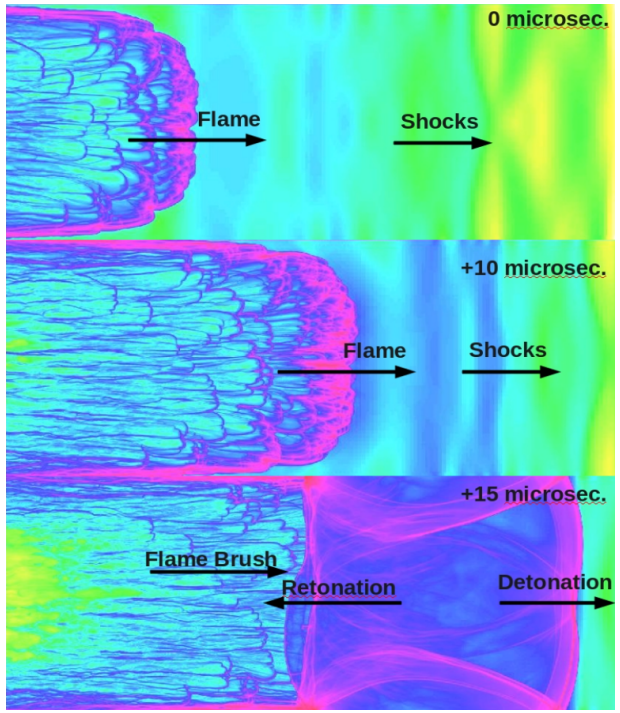
Tiny synthetic jet actuators dramatically improve effectiveness of aerodynamic control surfaces such as rudders. (PHASTA).



Biomolecular Science

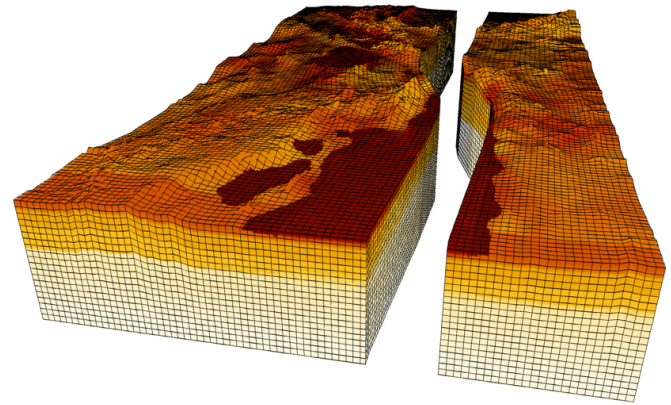
Highly accurate microscopic model of proteins and complexes—polarizable force field (NAMD).





High Speed Combustion and Detonation

Direct numerical simulation of shock tube experiments (ALLA/FTT).



Earthquake Genesis

Realistic 3D fault rupture simulation (SORD).

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 Projects

Proposal

- INCITE-lite
- Ambitious targeted science calculation
- Parallel performance
- Development needed
  - Porting and tuning for *Theta*/*Aurora*
  - New algorithms/physics

Selection

- Computational readiness review
- Science impact review
- Appropriate team (developers)
- Projects span wide range of
  - Scientific domains
  - Algorithms/numerical methods

ESP Projects

Proposal

- INCITE-lite
- Ambitious targeted science calculation
- Parallel performance
- Development needed
  - Porting and tuning for *Theta*/*Aurora*
  - New algorithms/physics

Selection

- Computational readiness review
- Science impact review
- Appropriate team (developers)
- Projects span wide range of
  - Scientific domains
  - Algorithms/numerical methods

Particiption in review by other stakeholders (NERSC, OLCF, Trinity/CORAL partners,….)

ESP Projects

Proposal

- INCITE-lite
- Ambitious targeted science calculation
- Parallel perform
- Development ne
  - Porting and tun
  - New algorithms/physics

Selection

- Computational readiness review
- Science impact review
- ers)
- f
  - Algorithms/numerical methods

http://esp.alcf.anl.gov

ESP Project Support

SUPPORT

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

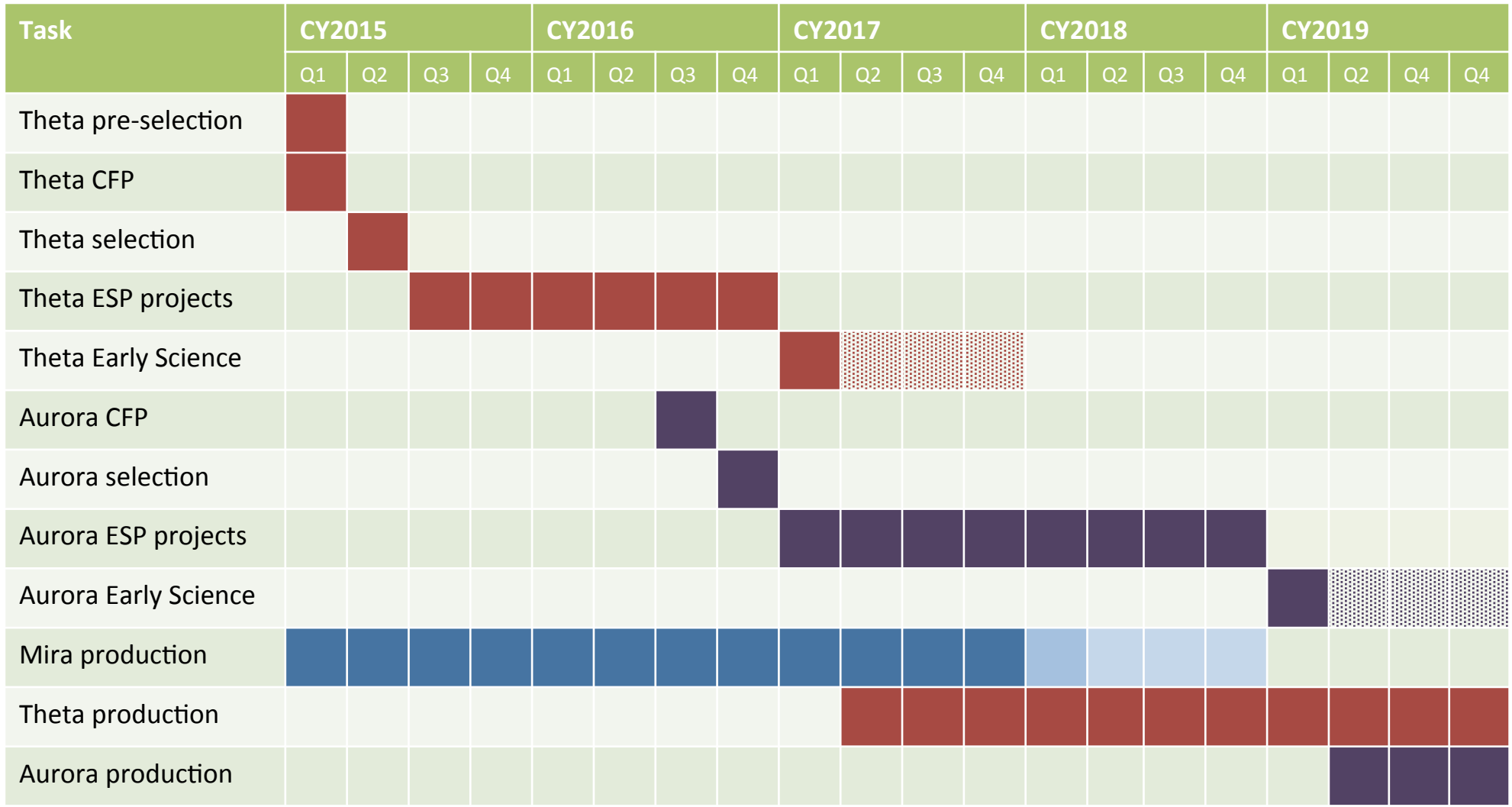
TRAINING

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

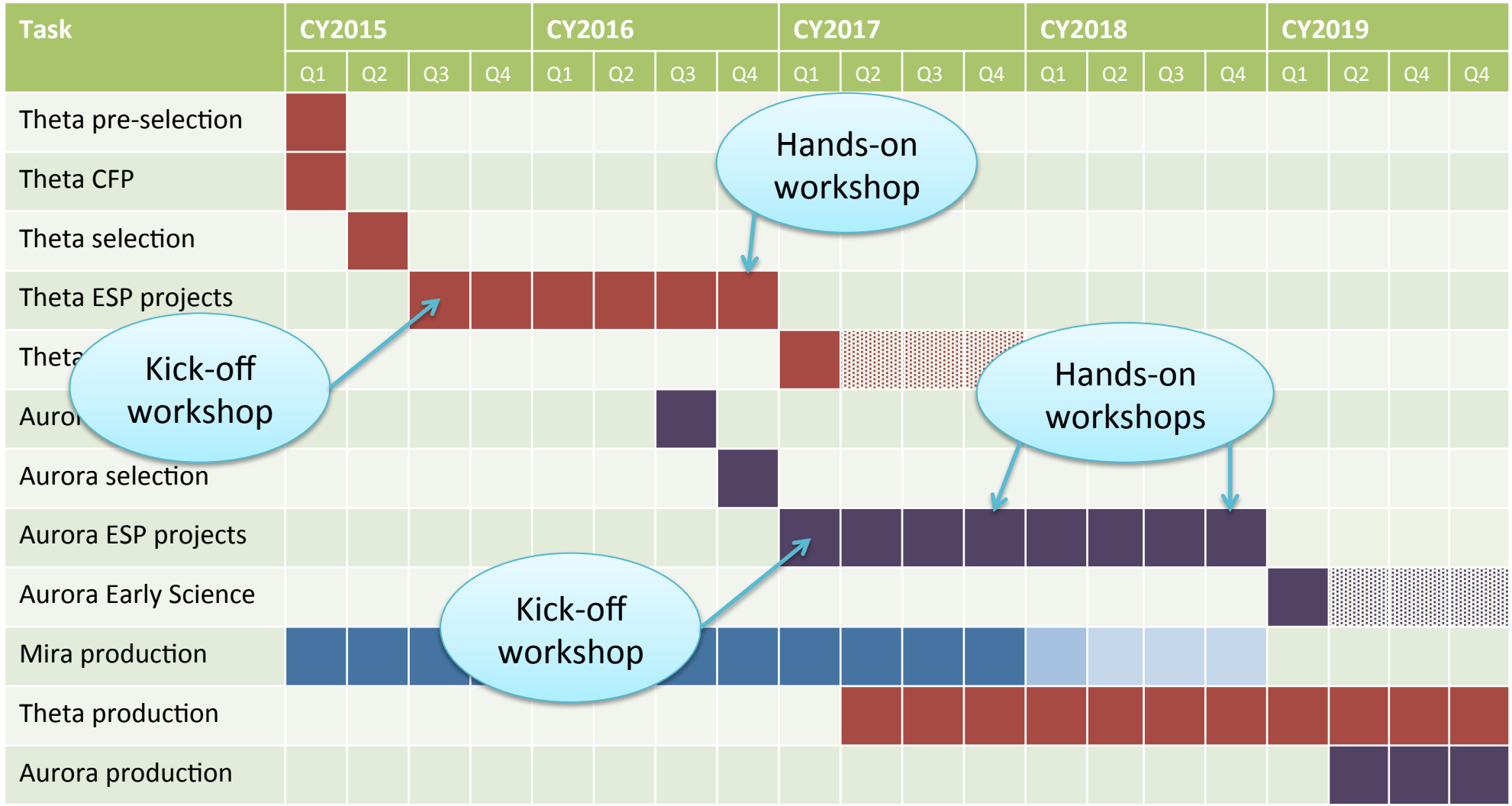
COMPUTE RESOURCES

- 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

ESP Timeline



ESP Timeline



Coordinating SC Centers Efforts

- Semiannual meetings of cross-labs applications readiness staff
- Tools and libraries working group (W. Joubert)
- Cross-lab training committee (F. Foerttner)
  - Shared calendar
  - Shared training events
- Manage nondisclosure, export control challenges
  - CORAL partners, APEX partners (NERSC & Trinity)
- Portability

March 2014 Meeting
- Apps readiness coordination
- ~15 representatives

September 2014 Meeting
- Apps Portability
- Apps readiness coordination
- ~25 representatives

January 2015 Meeting
- Next-gen hardware & software
- Portable programming
- ~40 center staff
- ~10 vendor reps

OpenMP 4.0 – Source Code Portability

```
double x[128], y[128];
#pragma omp target data map(to:x[0:64]) map(tofrom:y[0;64])
{
#pragma omp target
  {
     // y computed on device
  }
}
```

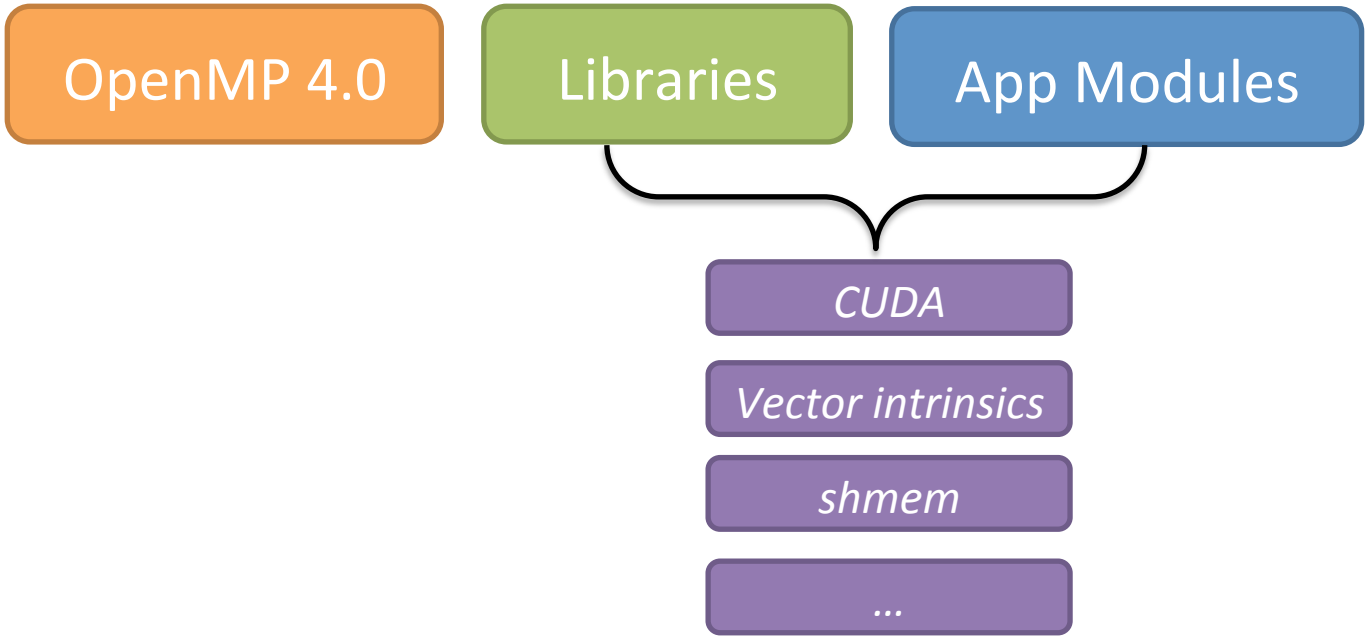
```
double x[128], y[128];
#pragma omp for simd aligned(x, y: 32)
for (int i=0; i<128; i++ ) {
  // thread's iterates → SIMD lanes
}
```

```
#pragma omp target data map(to:x) {
  #pragma omp target map(tofrom:y)
  #pragma omp teams
  #pragma omp distribute
  #pragma omp parallel for
  for (int i = 0; i < n; ++i) {
    y[i] = a*x[i] + y[i];
  }
}
```

- Host (CPU)
- Device (accelerator)
  - GPGPU
  - Intel Mic
  - `omp_get_num_devices()`
- `map` maybe shared location
- `simd` standard vectorization
- gcc 4.9 *getting close*

Performance Portability

- When directives-based approach performs: OpenMP 4.x
- *Use libraries or frameworks to encapsulate node level parallelism*



Charm++ | Trilinos
ADLB | MADNESS
PETSc | Kokkos
Chombo | RAJA

End