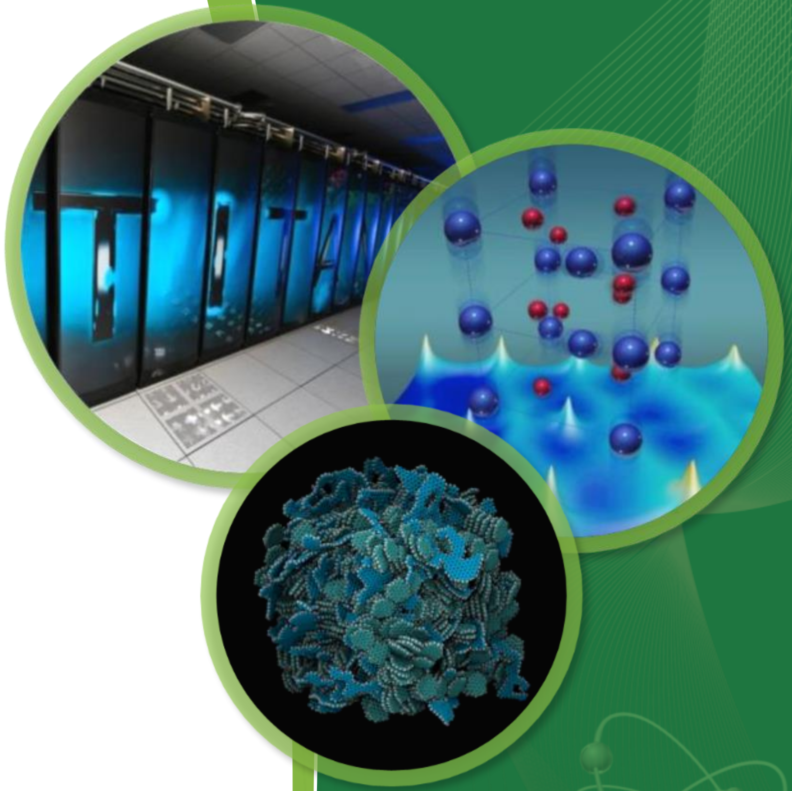
Oak Ridge Leadership Computing Facility: Summit and Beyond

Justin L. Whitt

OLCF-4 Deputy Project Director,
Oak Ridge Leadership Computing Facility
Oak Ridge National Laboratory

March 2017



ORNL is managed by UT-Battelle
for the US Department of Energy

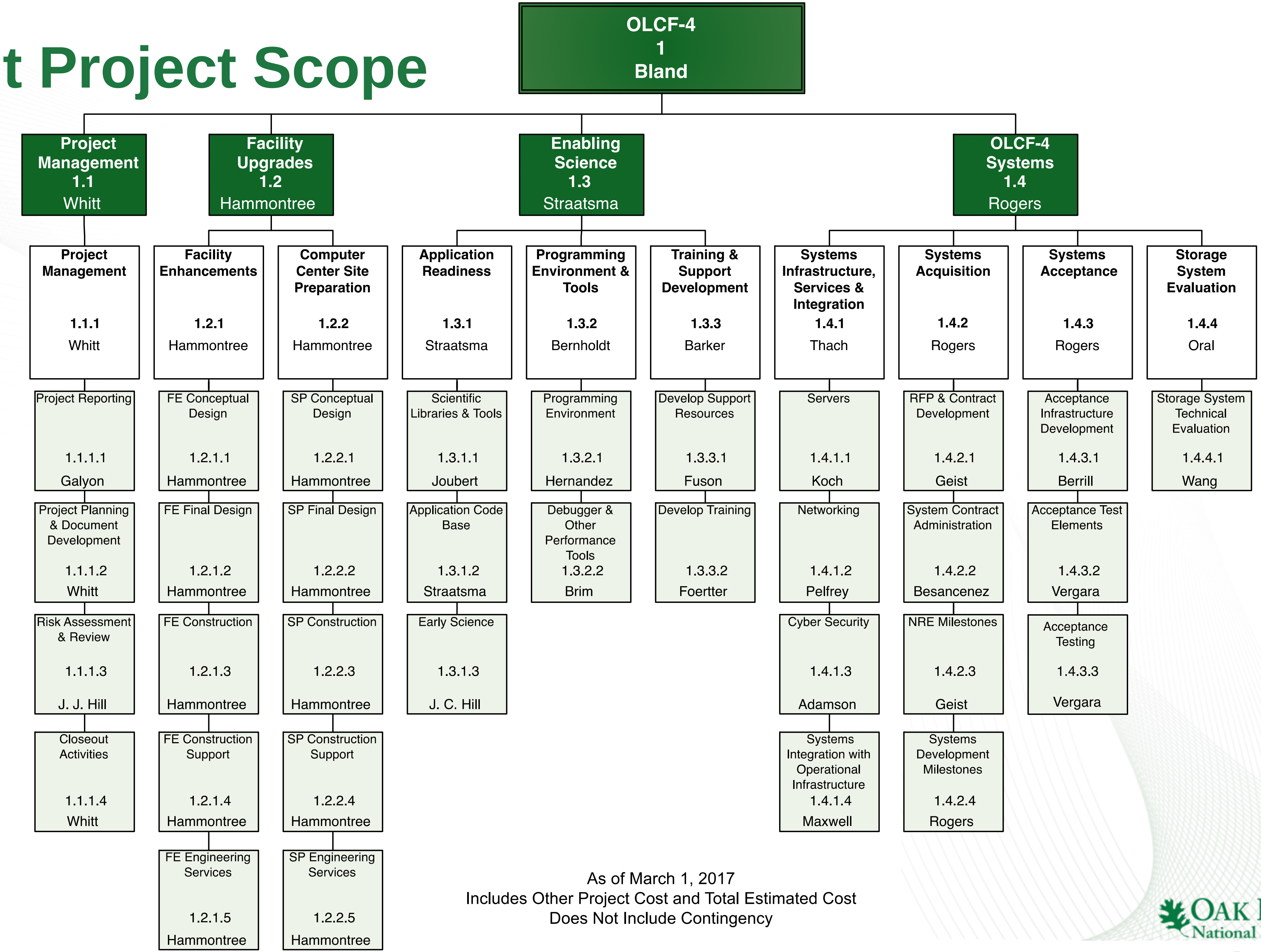
OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

Summit will replace Titan as the OLCF’s leadership supercomputer in 2019

- Many fewer nodes
- Much more powerful nodes
- Much more memory per node and total system memory
- Faster interconnect
- Much higher bandwidth between CPUs and GPUs
- Much larger and faster file system

Feature	Titan	Summit
Application Performance	Baseline	5-10x Titan
Number of Nodes	18,688	~4,600
Node performance	1.4 TF	> 40 TF
Memory per Node	38GB DDR3 + 6GB GDDR5	512 GB DDR4 + HBM
NV memory per Node	0	800 GB
Total System Memory	710 TB	>6 PB DDR4 + HBM + Non-volatile
System Interconnect (node injection bandwidth)	Gemini (6.4 GB/s)	Dual Rail EDR-IB (23 GB/s)
Interconnect Topology	3D Torus	Non-blocking Fat Tree
Processors	1 AMD Opteron™ 1 NVIDIA Kepler™	2 IBM POWER9™ 6 NVIDIA Volta™
File System	32 PB, 1 TB/s, Lustre®	250 PB, 2.5 TB/s, GPFS™
Peak power consumption	9 MW	13 MW

Summit Project Scope



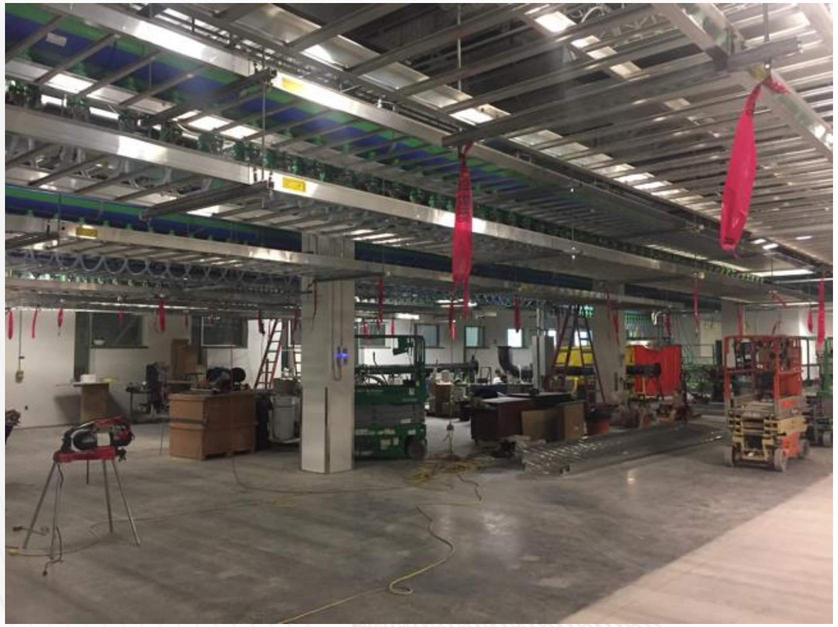
- OLCF-4 1 Bland
  - Project Management 1.1 Whitt
    - Project Management 1.1.1 Whitt
      - Project Reporting 1.1.1.1 Galyon
      - Project Planning & Document Development 1.1.1.2 Whitt
      - Risk Assessment & Review 1.1.1.3 J. J. Hill
      - Closeout Activities 1.1.1.4 Whitt
  - Facility Upgrades 1.2 Hammontree
    - Facility Enhancements 1.2.1 Hammontree
      - FE Conceptual Design 1.2.1.1 Hammontree
      - FE Final Design 1.2.1.2 Hammontree
      - FE Construction 1.2.1.3 Hammontree
      - FE Construction Support 1.2.1.4 Hammontree
      - FE Engineering Services 1.2.1.5 Hammontree
    - Computer Center Site Preparation 1.2.2 Hammontree
      - SP Conceptual Design 1.2.2.1 Hammontree
      - SP Final Design 1.2.2.2 Hammontree
      - SP Construction 1.2.2.3 Hammontree
      - SP Construction Support 1.2.2.4 Hammontree
      - SP Engineering Services 1.2.2.5 Hammontree
  - Enabling Science 1.3 Straatsma
    - Application Readiness 1.3.1 Straatsma
      - Scientific Libraries & Tools 1.3.1.1 Joubert
      - Application Code Base 1.3.1.2 Straatsma
      - Early Science 1.3.1.3 J. C. Hill
    - Programming Environment & Tools 1.3.2 Bernholdt
      - Programming Environment 1.3.2.1 Hernandez
      - Debugger & Other Performance Tools 1.3.2.2 Brim
    - Training & Support Development 1.3.3 Barker
      - Develop Support Resources 1.3.3.1 Fuson
      - Develop Training 1.3.3.2 Foertter
  - OLCF-4 Systems 1.4 Rogers
    - Systems Infrastructure, Services & Integration 1.4.1 Thach
      - Servers 1.4.1.1 Koch
      - Networking 1.4.1.2 Pelfrey
      - Cyber Security 1.4.1.3 Adamson
      - Systems Integration with Operational Infrastructure 1.4.1.4 Maxwell
    - Systems Acquisition 1.4.2 Rogers
      - RFP & Contract Development 1.4.2.1 Geist
      - System Contract Administration 1.4.2.2 Besancenez
      - NRE Milestones 1.4.2.3 Geist
      - Systems Development Milestones 1.4.2.4 Rogers
    - Systems Acceptance 1.4.3 Rogers
      - Acceptance Infrastructure Development 1.4.3.1 Berrill
      - Acceptance Test Elements 1.4.3.2 Vergara
      - Acceptance Testing 1.4.3.3 Vergara
    - Storage System Evaluation 1.4.4 Oral
      - Storage System Technical Evaluation 1.4.4.1 Wang

As of March 1, 2017
Includes Other Project Cost and Total Estimated Cost
Does Not Include Contingency

Facility Enhancements

- 13 MW power
- 20 MW of cooling capacity
- Preparation of the bare room for the computers

- Electrical distribution
- Cooling Water distribution
- Fire protection
- Controls systems



Application Readiness for Summit

The **Center of Accelerated Application Readiness (CAAR)** remains the OLCF's forward facing program to facilitate application readiness on evolving architectures

- ***Build on the experience of a successful application readiness program for OLCF-3 (Titan)***
- ***Thirteen CAAR projects were selected after a call for proposals***
  - Partnership: Application Developers,
    OLCF Scientific Computing staff
    Vendor Center of Excellence

- ***Resources available to CAAR projects***
  - Dedicated collaboration with OLCF Scientific Computing staff
  - Support and consultation from other OLCF staff and vendor Center of Excellence
  - Access to early test systems
  - Eight associated postdoctoral fellow in CSEEN program associated with CAAR projects
  - Allocations to available compute resources at OLCF, ALCF and NERSC in ALCC program
  - Early Science allocations

CAAR Applications

Domain	Application	Methods	PI	Institution	Related to INCITE/ALCC	Related to SciDAC
Astrophysics	FLASH	Grid, AMR	Bronson Messer	ORNL	Zingale	SciDAC II
Chemistry	DIRAC	Particle, LA	Lucas Visscher	VUA	Dixon	
Climate Science	ACME (N)	Unstr Mesh	David Bader	LLNL	Taylor	SciDAC III
Engineering	RAPTOR	Kokkos	Joseph Oefelein	SNL	Oefelein	SciDAC II
Materials Science	QMCPACK	MC	Paul Kent	ORNL	Kent, Ceperley	SciDAC III
Nuclear Physics	NUCCOR	Particle	Gaute Hagen	ORNL	Vary	SciDAC III
Plasma Physics	XGC (N)	PIC, PETSc	CS Chang	PPPL	Chang	SciDAC III
Seismic Science	SPECFEM	Unstr Mesh	Jeroen Tromp	Princeton	Tromp	
Astrophysics	HACC(N,A)	Grid	Salman Habib	ANL	Habib	SciDAC III
Biophysics	NAMD (N)	Particle	Klaus Schulten	UIUC	Klein, Schulten, Tajkhorshid	SciDAC II
Chemistry	NWCHEM (N)	Particle, LA	Karol Kowalski	PNNL	Dixon, Sumpter	SciDAC III
Chemistry	LSDALTON	Particle, LA	Poul Jørgensen	Aarhus	Jørgensen	
Plasma Physics	GTC (N)	PIC	Zhihong Lin	UCI	Lin	SciDAC III

N: NERSC application; A: ALCF application

CAAR: Architecture and Performance Portability

ALCF, NERSC and OLCF Joint Activities and Resources

SC15 Austin, TX | hpc transforms.

- ALCF, NERSC and OLCF participated in each other's proposal reviews
- ALCC Award to support NESAP, CAAR and ESP
- Common applications teams in NESAP, CAAR and ESP will collaborate
- Leveraging training activities at NERSC, OLCF and ALCF
- SC15 workshop "Portability Among HPC Architectures for Scientific Applications" on Sunday, November 15 was chaired by Tim Williams (ALCF), Katie Antypas (NERSC) and Tjerk Straatsma (OLCF)
- All three ASCR facilities have representation on the standards bodies for programming models that facilitate portability (OpenACC and OpenMP)
- We are working with vendors to provide programming environments and tools that enable portability, as part of CORAL and Trinity procurements
- Organized portability workshops
- Portability Research Project shared between OLCF, ALCF, NERSC and their CoE's

Oakland, September 24-25, 2014

Oak Ridge, January 27-29, 2015

Synergy between Application Readiness Programs

NESAP at NERSC - *NERSC Exascale Science Application Program*

- Call for Proposals – June 2014
- 20+26 Projects selected
- Partner with Application Readiness Team and Intel/Cray
- 8 Postdoctoral Fellows

CAAR at OLCF - *Center for Accelerated Application Readiness*

- Call for Proposals – November 2014
- 13 Projects selected
- Partner with Scientific Computing group and IBM/NVIDIA Center of Excellence
- 8 Postdoctoral Associates

ESP at ALCF - *Early Science Program*

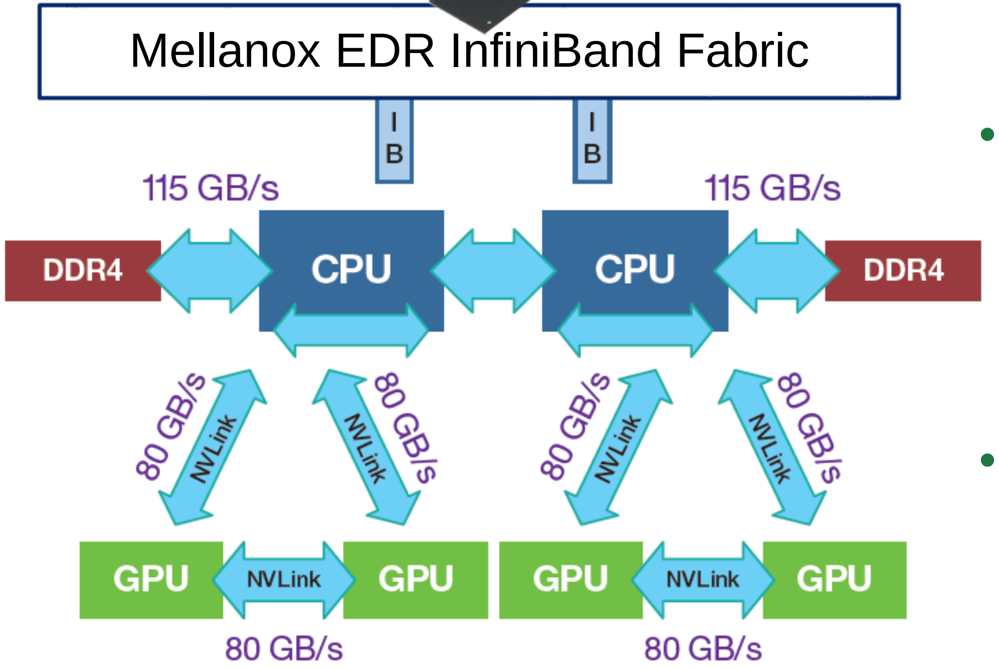
- Call for Proposals – May 2015
- 6 Projects selected in first round
- Partner with Catalyst group and Intel/Cray Center of Excellence
- Postdoctoral Appointee per project

OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

Summit Early Evaluation System

Each IBM S822LC node has:

- 2x IBM POWER8 CPUs
  - 32x 8GB DDR4 memory (256 GB)
  - 10 cores per POWER8,
    each core with 8 HW threads
- 4x NVIDIA Tesla P100 GPUs
  - NVLink 1.0 connects GPUs at 80 GB/s
  - 16 GB HBM2 memory per GPU
- 2x Mellanox EDR InfiniBand
- 800 GB NVMe storage



Summit EA System:

- Three racks, each with 18 nodes
- One rack of login and support servers
- Nodes connected in a full fat-tree via EDR InfiniBand
- Liquid cooled w/ heat exchanger rack
- We will get an additional rack to add to Summit EA for Exascale Computing Project testing, giving us a 54 node system
- One additional 18-node rack is for system software testing

Information and drawing from IBM Power System S822LC for High Performance Computing Data Sheet

Spider 3 @ OLCF

Spider 3 is a center-wide single namespace POSIX file system to serve all OLCF resources, eliminating data islands and enabling seamless data sharing between resources

- Built on IBM’s Elastic Storage Server and uses Spectrum Scale (formerly known as GPFS) parallel filesystem technology utilizing GPFS Native RAID with 8+2 redundancy
- Provides a usable capacity of 250 PB
- Performs at an aggregate sequential peak read/write bandwidth of 2.5 TB/s
- Performs at an aggregate random peak read/write bandwidth of 2.2 TB/s
- Provides rich metadata performance; single directory parallel create rate of 50,000/s
- Provides rich interactive performance; @32 KiB I/O 2.6 million IOPs
- Disk-based, with tens of thousands of disks
- Connected to OLCF’s SION 3 SAN with IB EDR
- Will also serve as the Summit Burst Buffer sink and source on the end-to-end I/O path

OLCF Programming Environment and Tools Focus Areas

Programming Environment

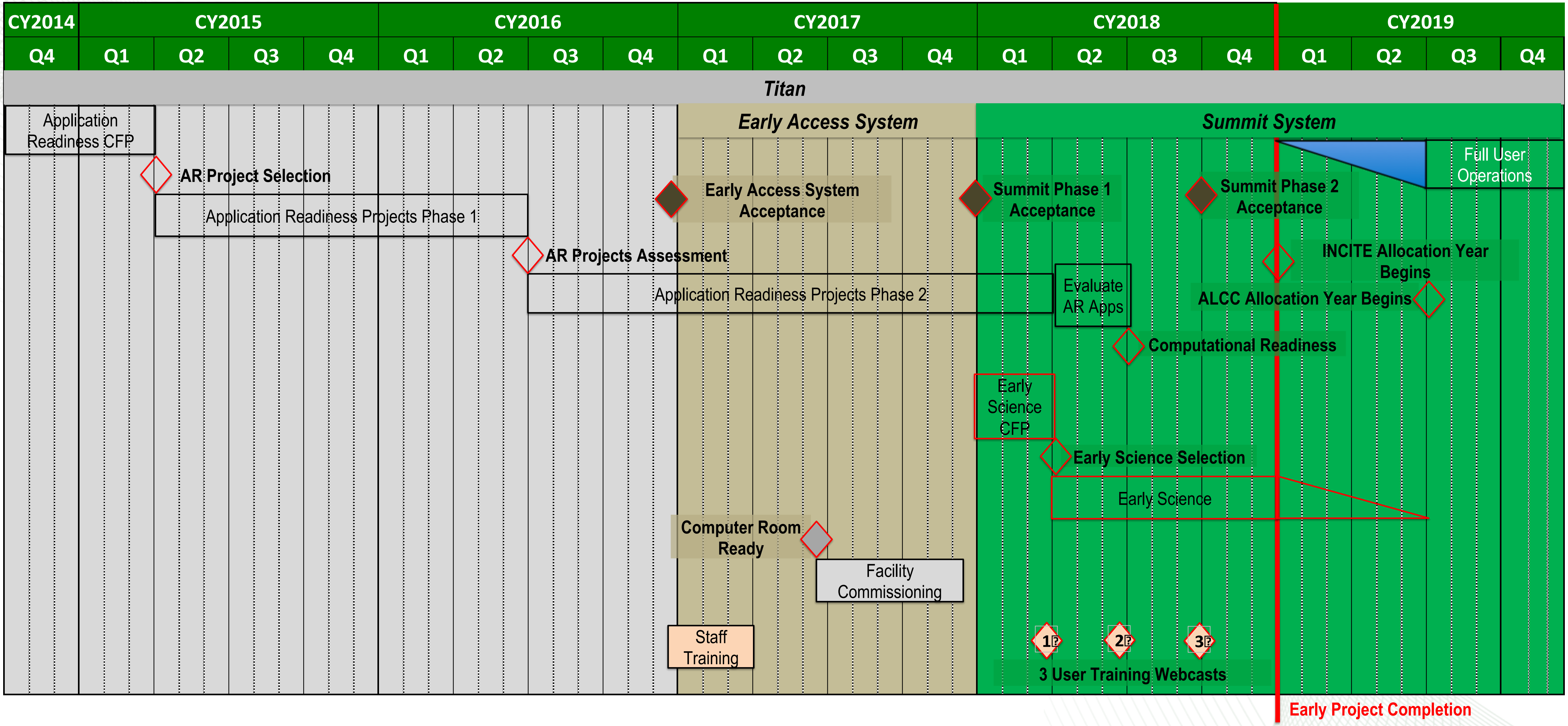
- Directive-based programming
  - **OpenACC** – accelerator offload, unified memory support, error handling
  - **OpenMP** – threading and tasks with accelerator offload under intensive development, memory hierarchy support, tools API, task reductions.
  - **SPEC High-Performance Group** – benchmark suites to drive performance and correctness
- Runtime
  - **MPI** – resilience, collectives, scalability

Tools

- Co-design of hardware and software to ensure maximum capabilities for tools on Summit system
  - **CPU, GPU, memory system, network**
  - **CORAL vendors, labs, tool developers**
- Target tools
  - HPCToolkit, Open|Speedshop, TAU, Valgrind, PAPI, and DynInst
  - Allinea DDT, MAP; Score-P/VAMPIR

- Evaluation of CORAL NRE products, including compilers, tools, and infrastructure
- Support for the Center for Accelerated Applications Readiness (Summit) applications
- Support for current OLCF users (Titan)

Timeline for Summit



CY2014 | CY2015 | CY2016 | CY2017 | CY2018 | CY2019

Q4 | Q1 | Q2 | Q3 | Q4 | Q1 | Q2 | Q3 | Q4 | Q1 | Q2 | Q3 | Q4 | Q1 | Q2 | Q3 | Q4 | Q1 | Q2 | Q3 | Q4

Titan

Early Access System

Summit System

Application Readiness CFP

Full User Operations

AR Project Selection

Early Access System Acceptance

Summit Phase 1 Acceptance

Summit Phase 2 Acceptance

Application Readiness Projects Phase 1

AR Projects Assessment

INCITE Allocation Year Begins

Application Readiness Projects Phase 2

Evaluate AR Apps

ALCC Allocation Year Begins

Computational Readiness

Early Science CFP

Early Science Selection

Early Science

Computer Room Ready

Facility Commissioning

Staff Training

1 2 3

3 User Training Webcasts

Early Project Completion

NVIDIA DGX-1

Specifications

- 8 Tesla P100 GPUs
- GPU Mem: 16 GB per GPU
- System Mem: 512 GB
- Storage: 4x 2 TB SSDs
- Out of the box libraries (theano, caffe, cuDNN, cuBLAS, etc)
- Integrated into the CADES environment

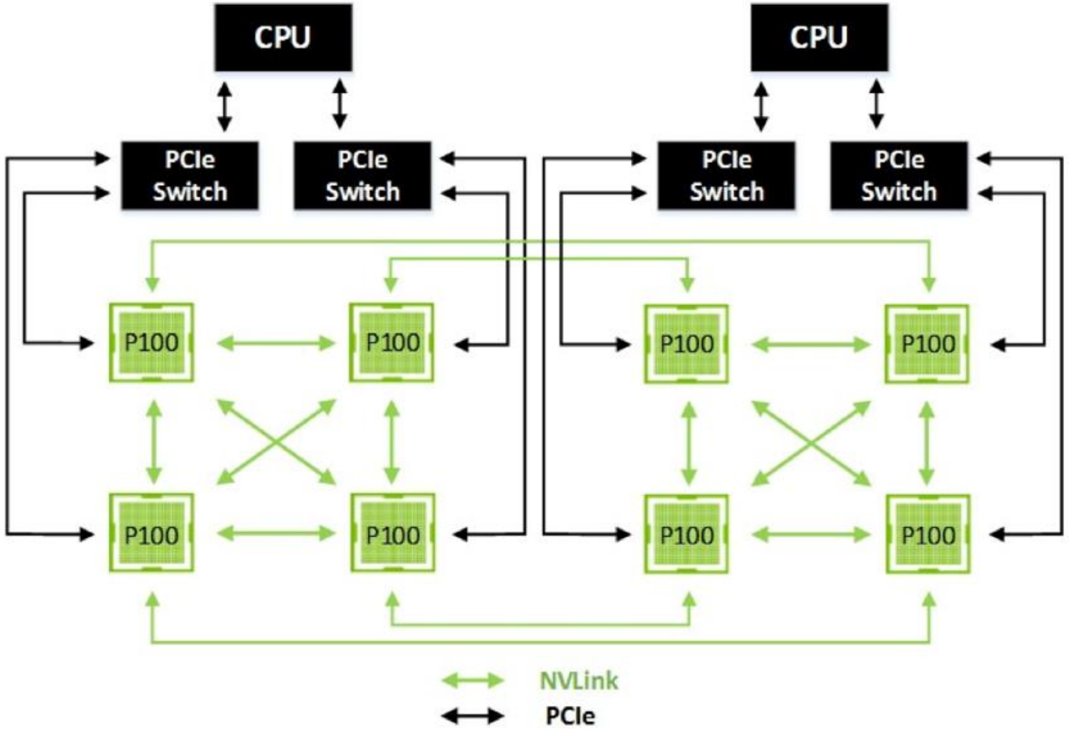
Preparation for Summit

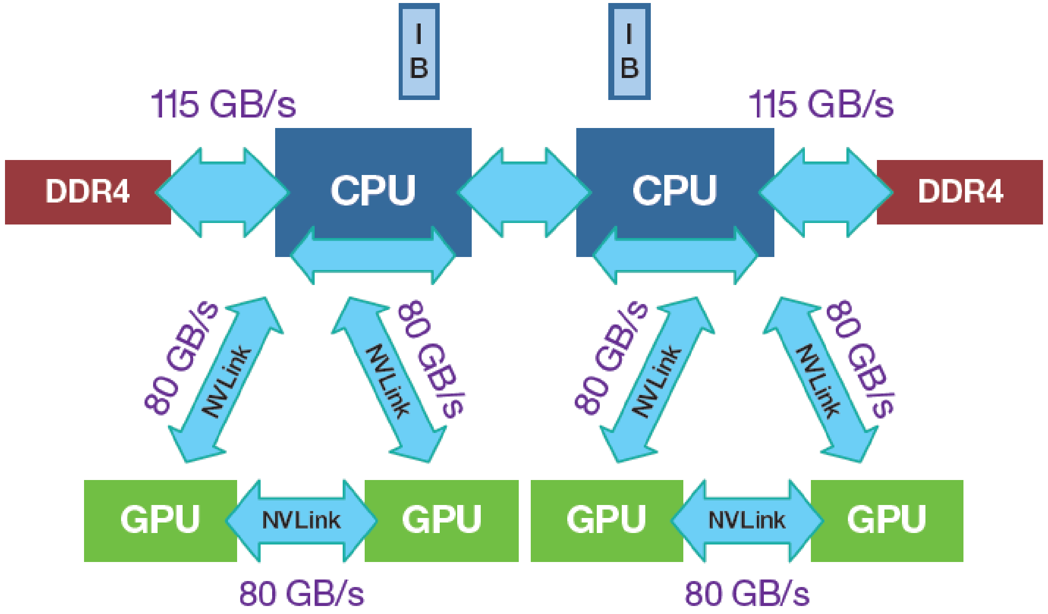
- **DGX-1 architecture and use of NVLink similar to Summit-dev**
  - Ease the preparation for DNN training at larger scales

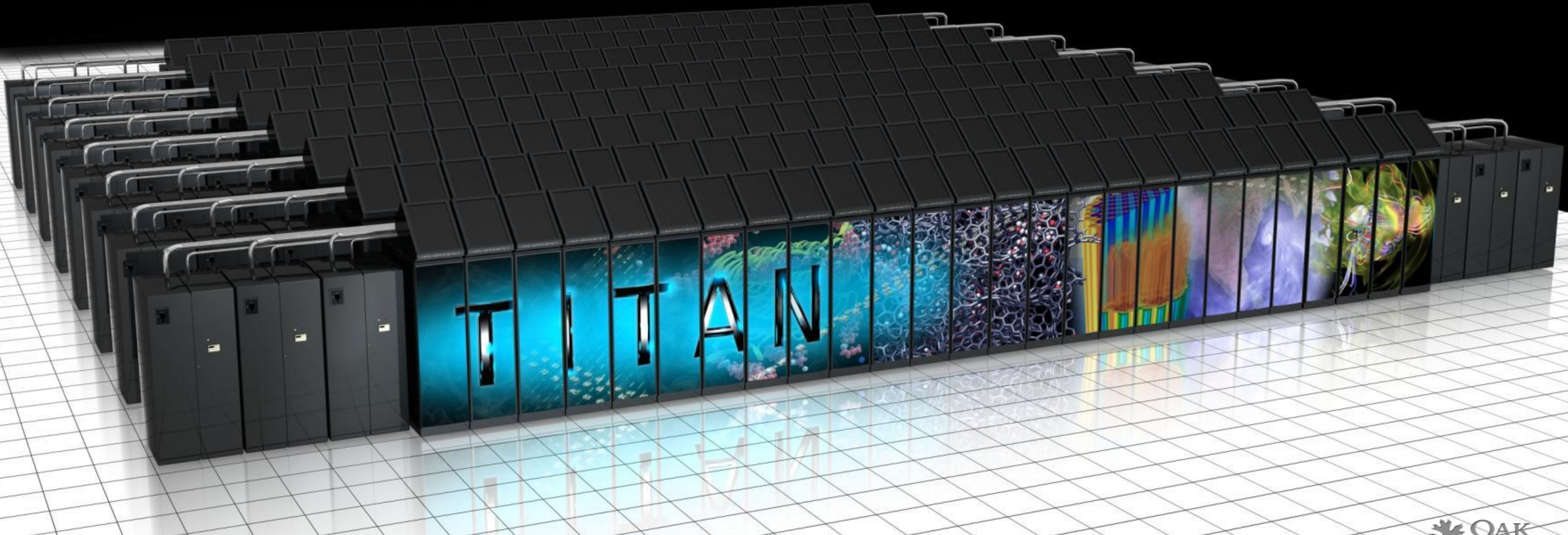
DGX-1

Summit-dev

(IBM's S822LC)






TITAN