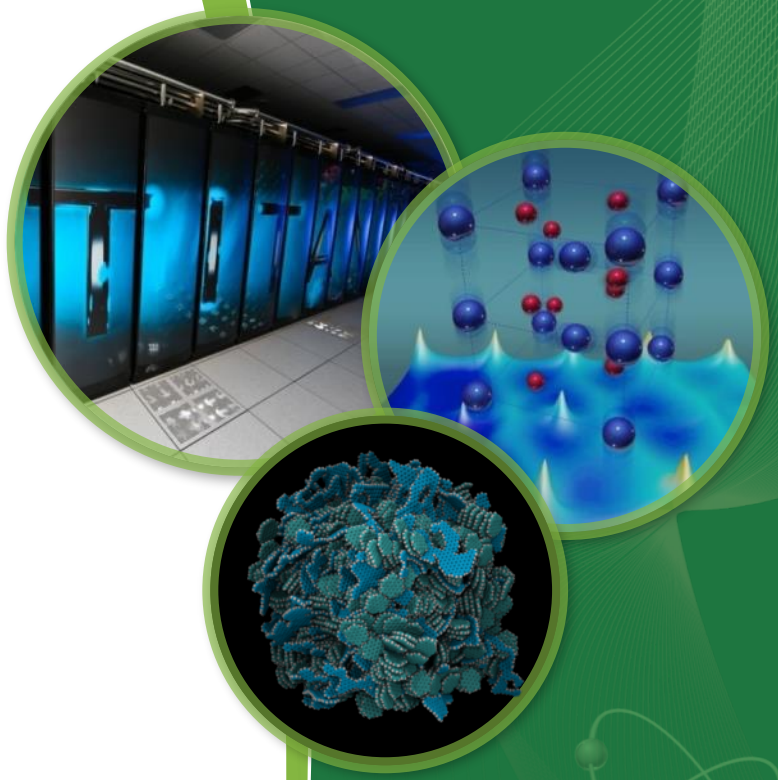
Preparing GPU-Accelerated Applications for the Summit Supercomputer



Fernanda Foertter
HPC User Assistance Group
Training Lead

foertterfs@ornl.gov

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

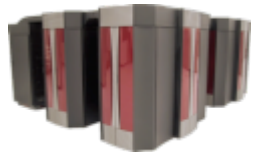
What is the Leadership Computing Facility

- Collaborative DOE Office of Science program at ORNL and ANL
- Mission: Provide the computational and data resources required to solve the most challenging problems.
- 2-centers/2-architectures to address diverse and growing computational needs of the scientific community
- Highly competitive user allocation programs (INCITE, ALCC).
- Projects receive 10x to 100x more resource than at other generally available centers.
- LCF centers partner with users to enable science & engineering breakthroughs (Liaisons, Catalysts).

OAK RIDGE National Laboratory
Argonne NATIONAL LABORATORY

The OLCF has delivered five systems and six upgrades to our users since 2004

- Increased our system capability by 10,000x
- Strong partnerships with computer designers and architects
- Worked with users to scale codes by 10,000x
- Science delivered through strong user partnerships to scale codes and algorithms



Phoenix X1
- Doubled size
- X1e

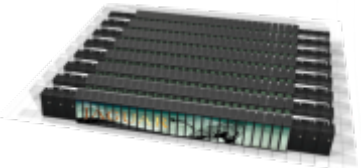
2004

Jaguar XT3
- Dual core upgrade

2005

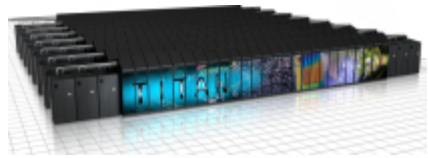
Jaguar XT4
- Quad core upgrade

2007



Jaguar XT5
- 6 core upgrade

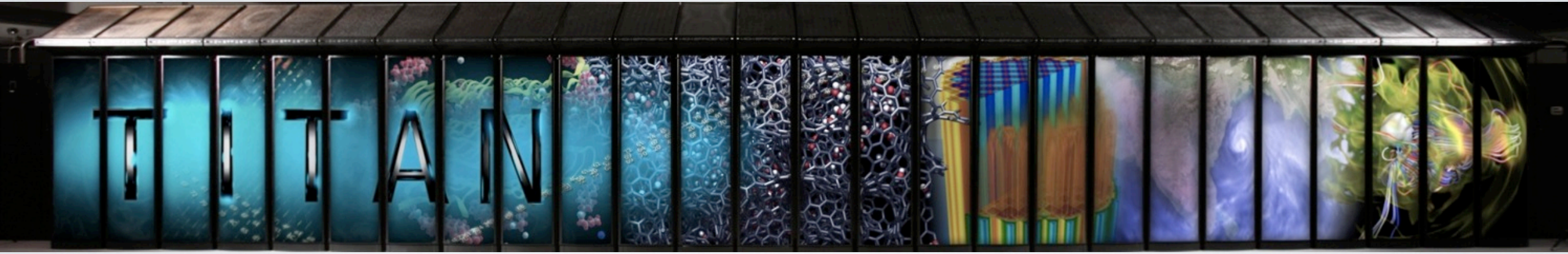
2008

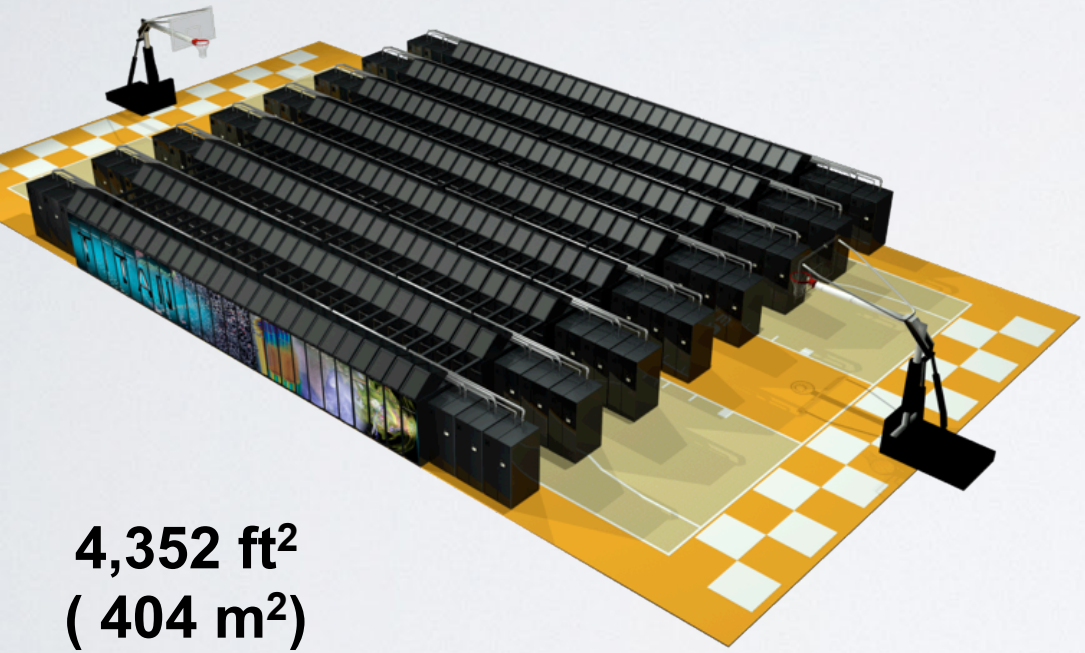


Titan XK7
- GPU upgrade

2012

ORNL'S TITAN HYBRID SYSTEM



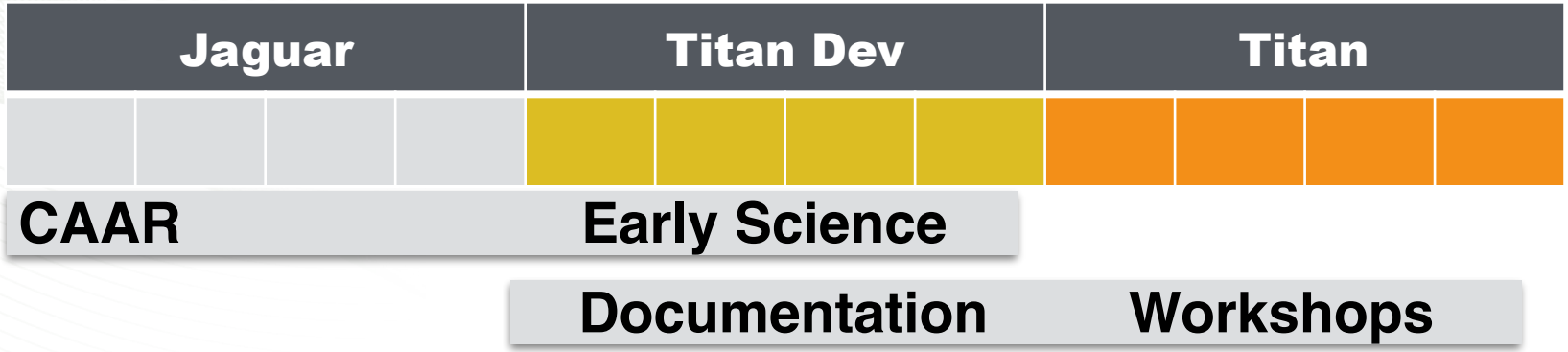


4,352 ft²
(404 m²)

SYSTEM SPECIFICATIONS:

- Peak performance of 27.1 PF
  - 24.5 GPU + 2.6 CPU
- 18,688 Compute Nodes each with:
  - 16-Core AMD Opteron CPU
  - NVIDIA Tesla "K20x" GPU
  - 32 + 6 GB memory
- 512 Service and I/O nodes
- 200 Cabinets
- 710 TB total system memory
- Cray Gemini 3D Torus Interconnect
- 8.9 MW peak power

Migrating to Titan



ELSEVIER

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Accelerated application development: The ORNL Titan experience ☆,☆☆,★

Wayne Joubert [a,*], Rick Archibald [a], Mark Berrill [a], W. Michael Brown [a], Markus Eisenbach [a], Ray Grout [c], Jeff Larkin [d], John Levesque [b], Bronson Messer [a], Matt Norman [a], Bobby Philip [a], Ramanan Sankaran [a], Arnold Tharrington [a], John Turner [a]

[a] *Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37831, United States*
[b] *Cray, Inc., Knoxville, TN, United States*
[c] *National Renewable Energy Laboratory, 15013 Denver West Parkway, Golden, CO 80401, United States*
[d] *NVIDIA Corp., P.O. Box 2008, MS-6008, Oak Ridge, TN 37831, United States*

Post-Moore's Law Era: Two Architecture Paths for Future Systems

Power concerns for large supercomputers are driving the largest systems to either Hybrid or Many-core architectures

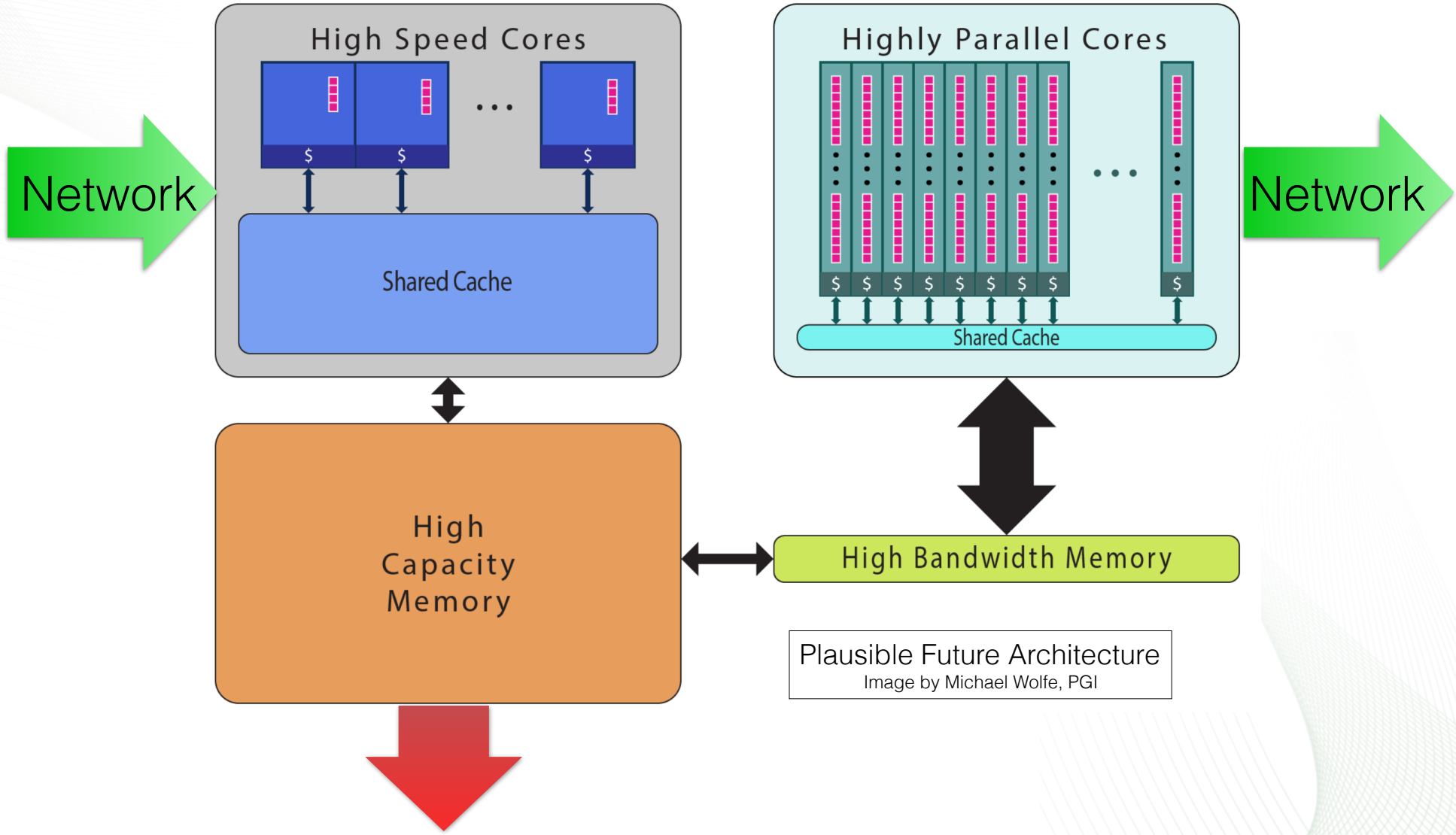
Hybrid Multi-Core (like Titan)

- CPU / GPU hybrid systems
- Likely to have multiple CPUs and GPUs per node
- Small number of very powerful nodes
- Expect data movement issues to be much easier than previous systems – coherent shared memory within a node
- Multiple levels of memory – on package, DDR, and non-volatile

Many Core (like Sequoia/Mira)

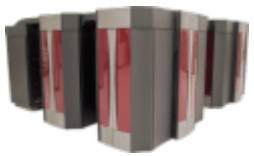
- 10's of thousands of nodes with millions of cores
- Homogeneous cores
- Multiple levels of memory – on package, DDR, and non-volatile
- Unlike prior generations, future products are likely to be self hosted

Data movement and layout matters most



Path of future architectures shows increasing parallelism

- Hierarchical parallelism
- Hierarchical data spaces



Phoenix X1
- Doubled size
- X1e

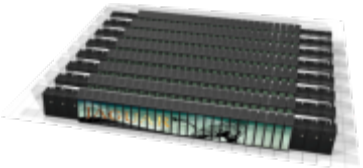
2004

Jaguar XT3
- Dual core upgrade

2005

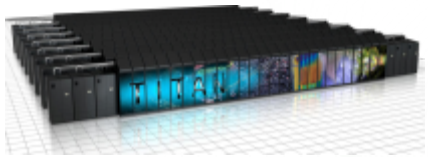
Jaguar XT4
- Quad core upgrade

2007



Jaguar XT5
- 6 core upgrade

2008



Titan XK7
- GPU upgrade

2012

Summit
- Hybrid Accelerated

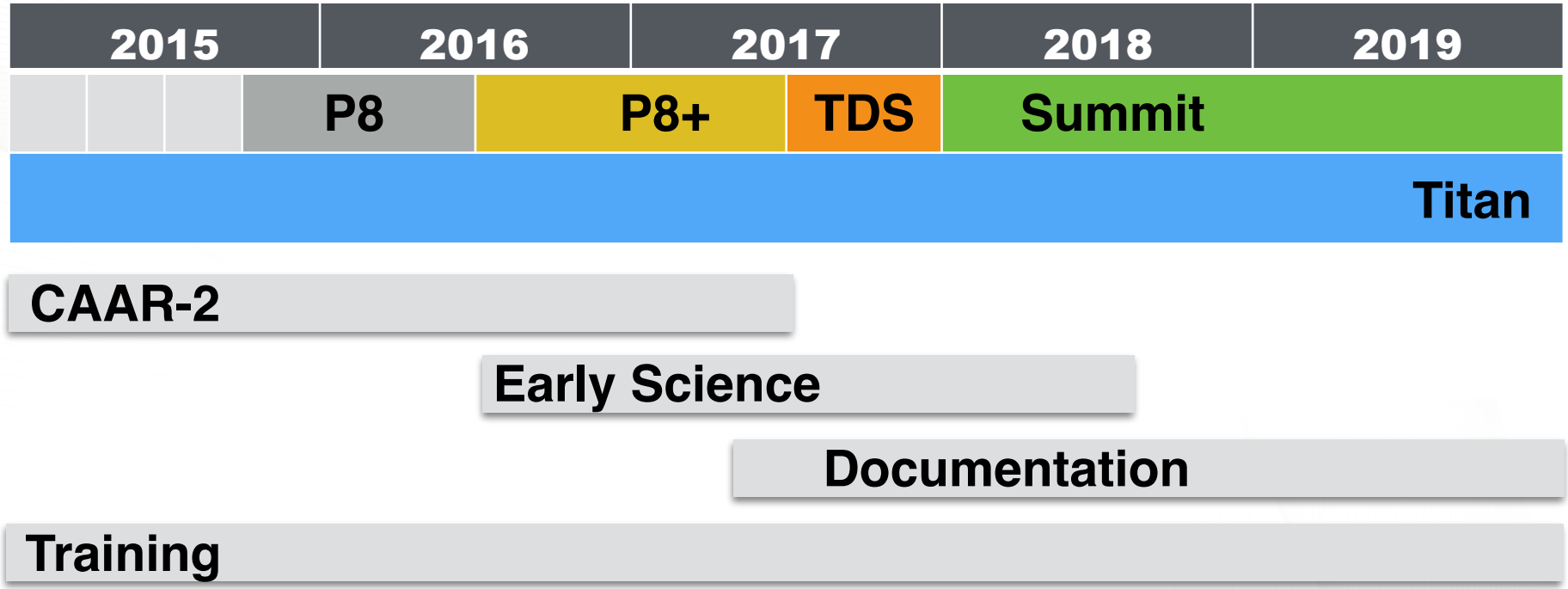
2017

Summit will replace Titan as OLCF's leadership supercomputer

- 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) Or Dual Rail HDR-IB (48 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

Migrating to Summit



- Continuous training program
  - Focus on porting
  - Portability
  - Performance

OLCF Hackathons

Moving domain applications forward

2014	2015	2016	2017
OAK RIDGE National Laboratory	NCSA	TECHNISCHE UNIVERSITÄT DRESDEN	JÜLICH FORSCHUNGSZENTRUM
	CSCS Centro Svizzero di Calcolo Scientifico Swiss National Supercomputing Centre	UNIVERSITY OF DELAWARE	BROOKHAVEN NATIONAL LABORATORY
	OAK RIDGE National Laboratory	CSCS Centro Svizzero di Calcolo Scientifico Swiss National Supercomputing Centre	NASA Langley Research Center
		OAK RIDGE National Laboratory	CSCS Centro Svizzero di Calcolo Scientifico Swiss National Supercomputing Centre
			OAK RIDGE National Laboratory

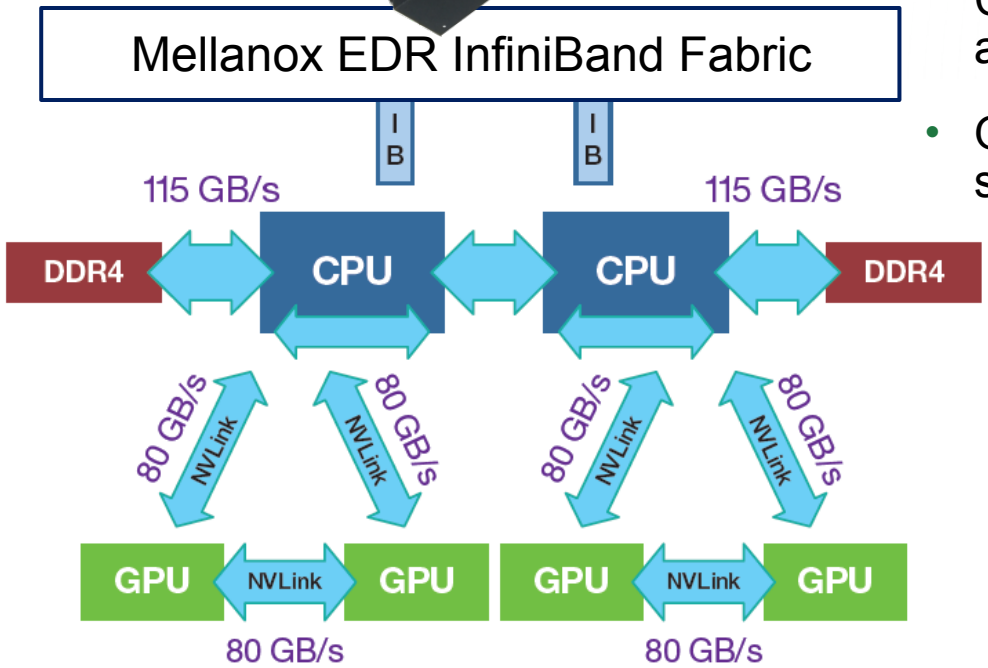
75 teams + apps

+230 attendees

8 partner sites

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



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

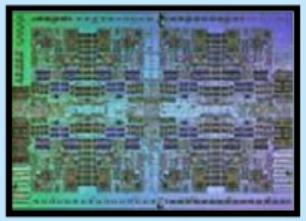
Summit EA System:

- Two 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

Summit Programming Environment

- **System**
  - Linux®
  - IBM Elastic Storage (GPFS™)
  - IBM Platform Computing™ (LSF)
  - IBM Platform Cluster Manager™ (xCAT)
- **Programming Environment**
  - Compilers supporting OpenMP, OpenACC, CUDA
    - IBM XL, PGI, LLVM, GNU, NVIDIA
  - Libraries
    - IBM Engineering and Scientific Subroutine Library (ESSL)
    - FFTW, ScaLAPACK, PETSc, Trilinos, BLAS-1,-2,-3, NVBLAS
    - cuFFT, cuSPARSE, cuRAND, NPP, Thrust
  - Debugging
    - Allinea DDT, IBM Parallel Environment Runtime Edition (pdb)
    - Cuda-gdb, Cuda-memcheck, valgrind, memcheck, helgrind, stacktrace
  - Profiling
    - IBM Parallel Environment Developer Edition (HPC Toolkit)
    - VAMPIR, Tau, Open|Speedshop, nvprof, gprof, Rice HPCToolkit

POWER Processor Technology Roadmap

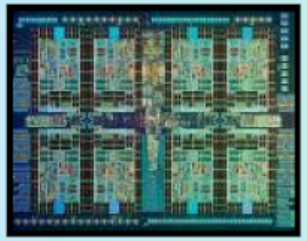


POWER7
45 nm

Enterprise

- 8 Cores
- SMT4
- eDRAM L3 Cache

1H10

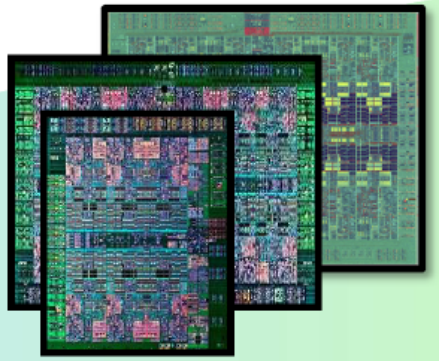


POWER7+
32 nm

Enterprise

- 2.5x Larger L3 cache
- On-die acceleration
- Zero-power core idle state

2H12

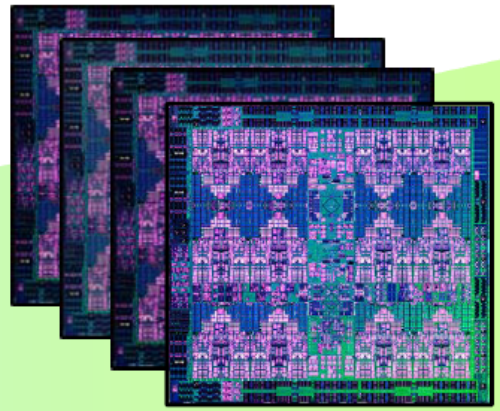


POWER8 Family
22nm

Enterprise & Big Data Optimized

- Up to 12 Cores
- SMT8
- CAPI Acceleration
- High Bandwidth GPU Attach

1H14 – 2H16



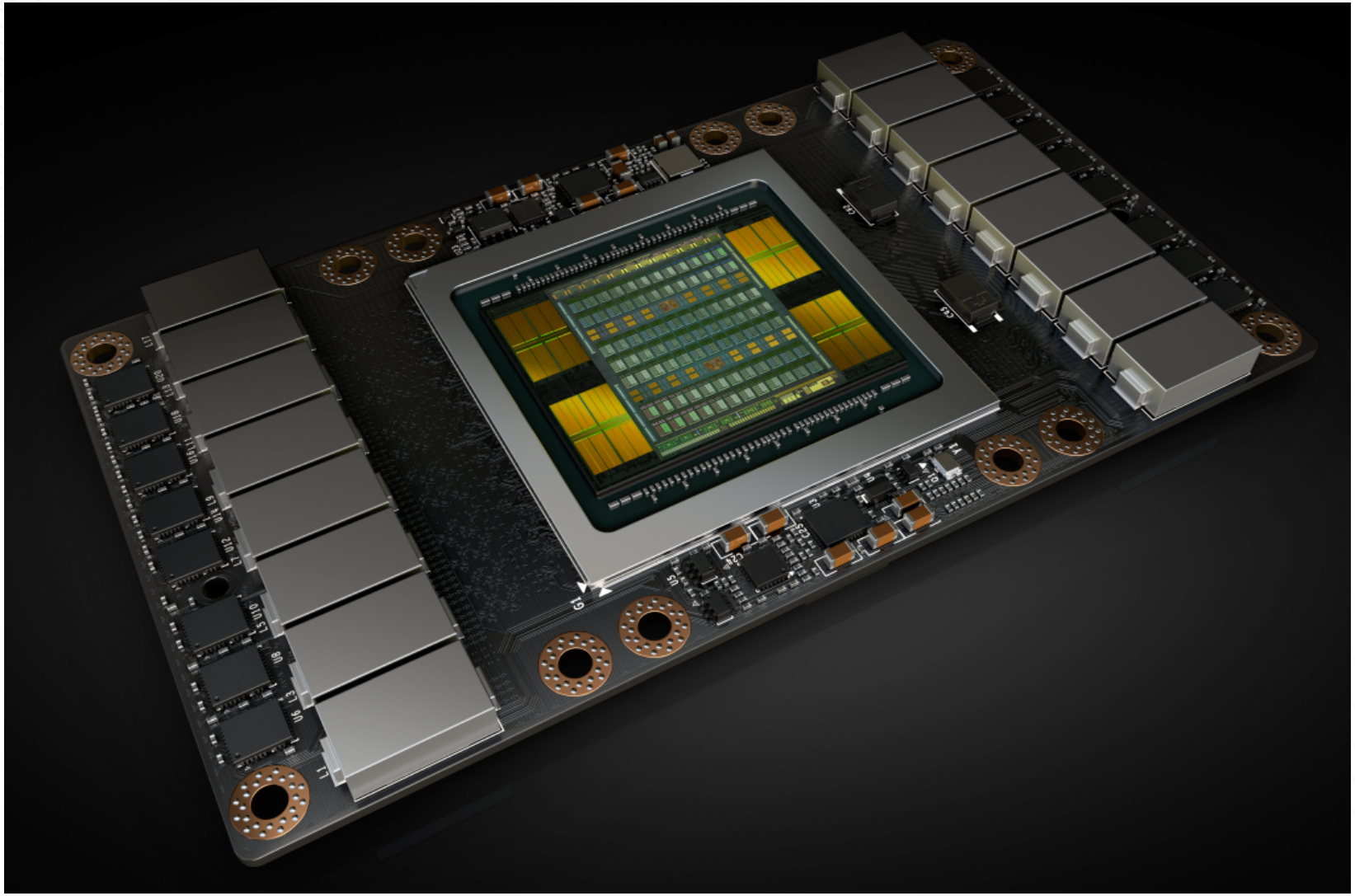
POWER9 Family
14nm

Built for the Cognitive Era

- Enhanced Core and Chip Architecture Optimized for Emerging Workloads
- Processor Family with Scale-Up and Scale-Out Optimized Silicon
- Premier Platform for Accelerated Computing

2H17 – 2H18+

Volta



- Volta
  - NVLink 2nd Gen
  - 16GB HBM2
  - 7.5 TFLOP/s DP (FP64)
  - 15 TFLOP SP (FP32)
  - 120 Tensor TFLOP/s
  - 80 SMs
  - 5120 cores

Options for accelerating apps

APPLICATIONS

LIBRARIES trillinos, petsc, blas	**DIRECTIVES** OpenMP, OpenACC	**PROGRAMMING LAGUAGES** F/C/C++
✳ Drop-in ✳ HW Vendor ✳ Reuse	✳ Simple ✳ May be limited ✳ Performance close	✳ Flexible ✳ Hand coding ✳ Adv features issues

Conclusions

- More parallelism is the new way of life
- Future applications need diverse teams
- Code will have to restructure expose parallelism
- Portability across platforms is a must
- We are here to help you!
- We would like to partner with your organization
  - If you're unsure where to begin, come talk to me

More thoughts on future…

- Memory addressing may be unified, but good performance will still require good placement
  - don’t depend on page fault hw fetch
- Memory hierarchies are increasing with no answer in sight
- Communication layer evolving slowest
- Data structure must be looked at again
- CPU clock going down, parallelism going up
  - vector lengths may increase or even be variable!
- Communities must adopt good SW Eng practices

Partnering with us

Open Tools

- Join us in developing open tools

Open Standards

- Join us in developing open API standards

Open Training

- Join us in improving scientific applications

Open Science

- We do Open Science! Become a user.

POWER9 Processor – Common Features

New Core Microarchitecture

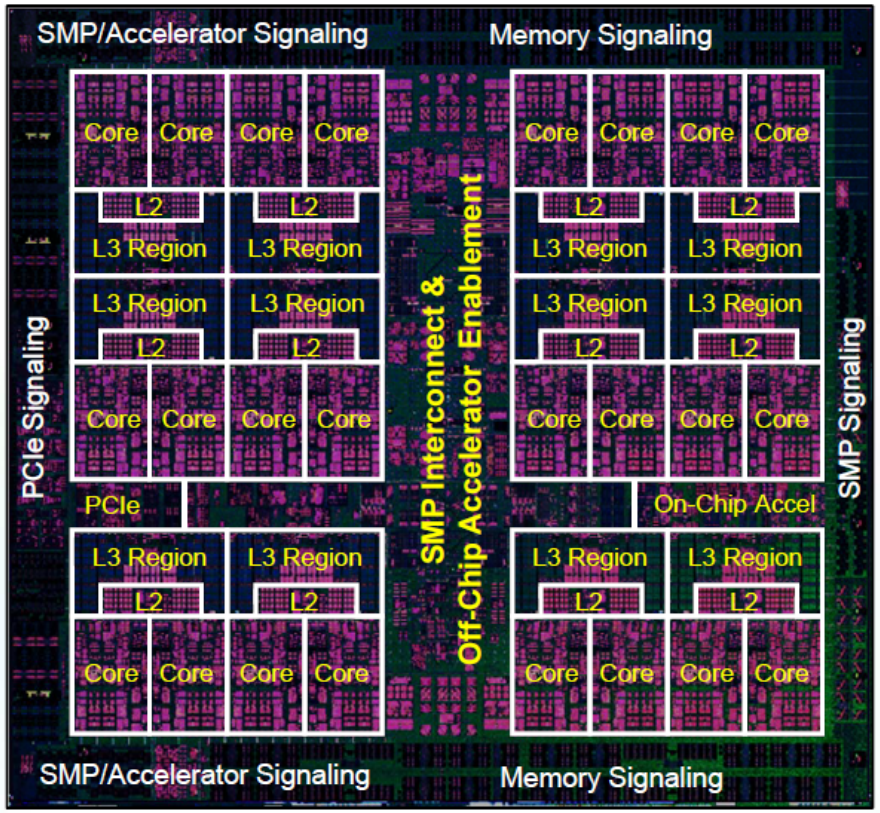
- Stronger thread performance
- Efficient agile pipeline
- POWER ISA v3.0

Enhanced Cache Hierarchy

- 120MB NUCA L3 architecture
- 12 x 20-way associative regions
- Advanced replacement policies
- Fed by 7 TB/s on-chip bandwidth

Cloud + Virtualization Innovation

- Quality of service assists
- New interrupt architecture
- Workload optimized frequency
- Hardware enforced trusted execution



14nm finFET Semiconductor Process

- Improved device performance and reduced energy
- 17 layer metal stack and eDRAM
- 8.0 billion transistors

Leadership Hardware Acceleration Platform

- Enhanced on-chip acceleration
- Nvidia NVLink 2.0: High bandwidth and advanced new features (BlueLink)
- CAPI 2.0: Coherent accelerator and storage attach (PCIe G4)
- New CAPI: Improved latency and bandwidth, open interface (BlueLink)

State of the Art I/O Subsystem

- PCIe Gen4 – 48 lanes

High Bandwidth Signaling Technology

- 16 Gb/s interface
  - Local SMP
- 25 Gb/s IBM BlueLink interface
  - Accelerator, remote SMP