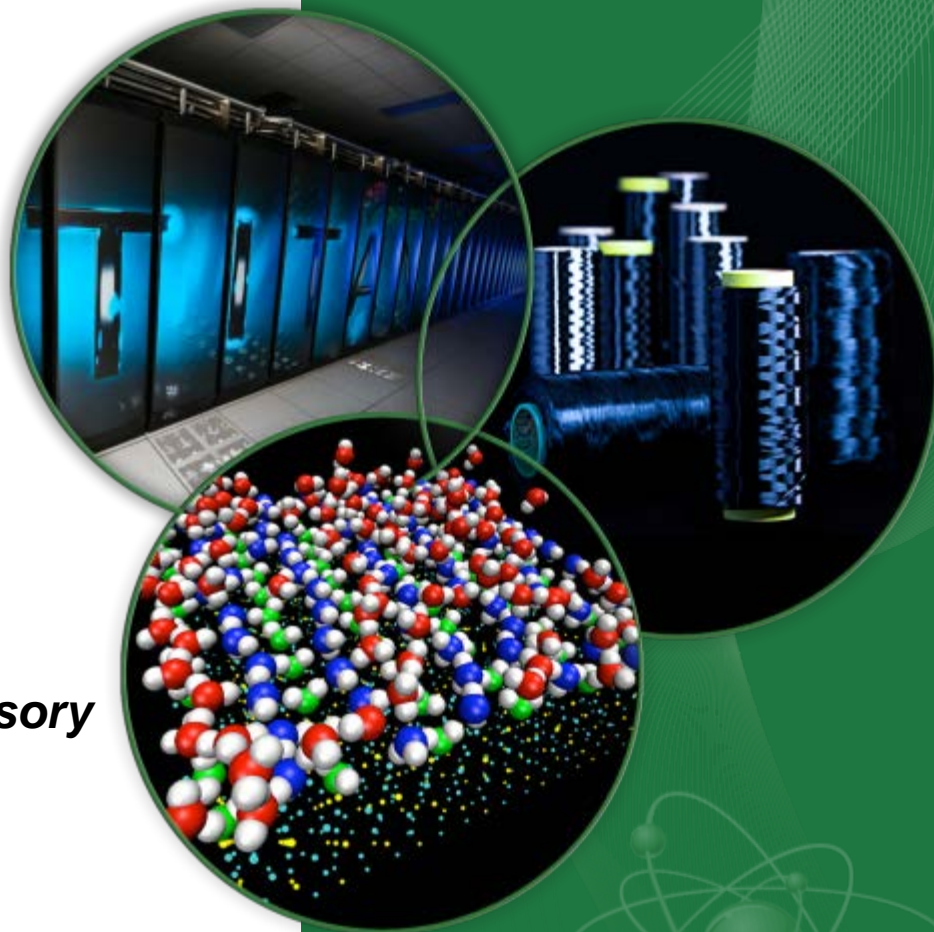


CORAL Acquisition Update: OLCF-4 Project

Buddy Bland

OLCF Project Director

***Presented to:
Advanced Scientific Computing Advisory
Committee
November 21, 2014
New Orleans***



Oak Ridge Leadership Computing Facility (OLCF)

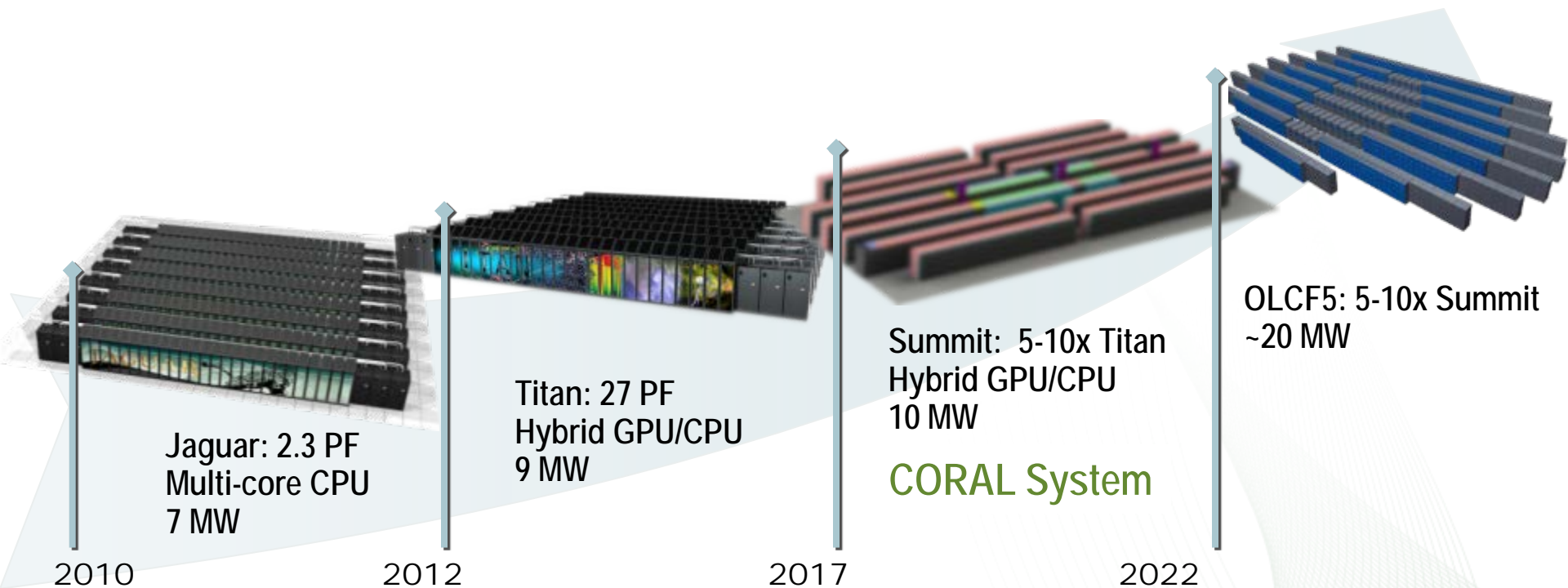
Mission: Deploy and operate the computational resources required to tackle global challenges

- ☐ **Providing world-leading computational and data resources and specialized services for the most computationally intensive problems**
- ☐ **Providing stable hardware/software path of increasing scale to maximize productive applications development**
- ☐ **Providing the resources to investigate otherwise inaccessible systems at every scale: from galaxy formation to supernovae to earth systems to automobiles to nanomaterials**
- ☐ **With our partners, deliver transforming discoveries in materials, biology, climate, energy technologies, and basic science**

Our Science requires that we continue to advance OLCF's computational capability over the next decade on the roadmap to Exascale.

Since clock-rate scaling ended in 2003, HPC performance has been achieved through increased parallelism. Jaguar scaled to 300,000 cores.

Titan and beyond deliver hierarchical parallelism with very powerful nodes. MPI plus thread level parallelism through OpenACC or OpenMP plus vectors

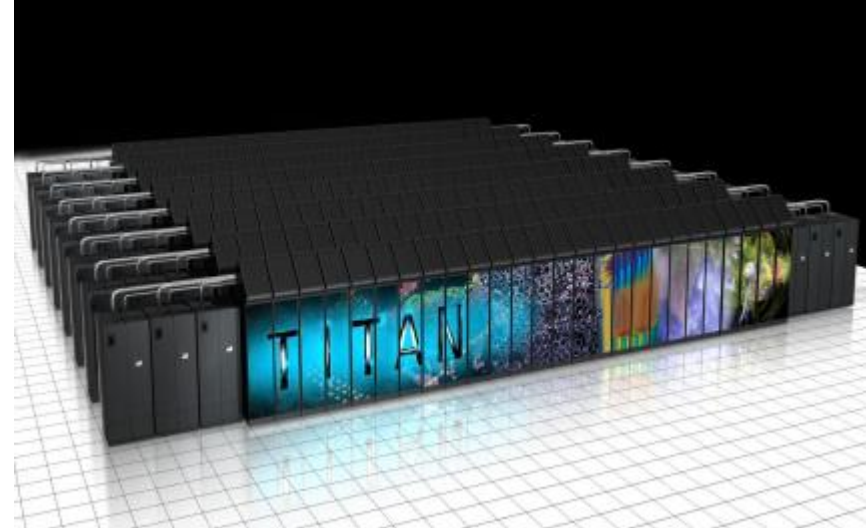


Today's Leadership System - Titan

Hybrid CPU/GPU architecture, Hierarchical Parallelism

Vendors: Cray™ / NVIDIA™

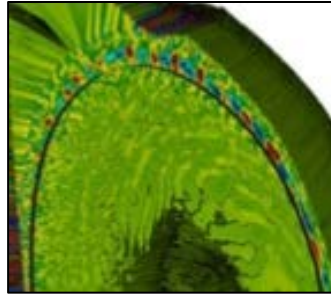
- 27 PF peak
- 18,688 Compute nodes, each with
 - 1.45 TF peak
 - 6 GB GDDR5 memory
 - NVIDIA Kepler™ GPU - 1,311 GF
 - 32 GB DDR3 memory
 - PCIe2 link between GPU and CPU
- Cray Gemini 3-D Torus Interconnect
- 32 PB / 1 TB/s Lustre® file system



Scientific Progress at all Scales

Fusion Energy

A Princeton Plasma Physics Laboratory team led by C.S. Chang was able to simulate and explain nonlinear coherent turbulence structures on the plasma edge in a fusion reactor by exploiting increased performance of the Titan architecture.

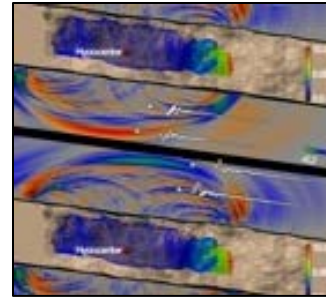
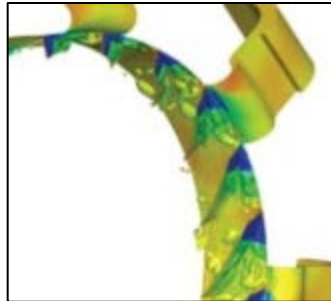


Liquid Crystal Film Stability

ORNL Postdoctoral fellow Trung Nguyen ran unprecedented large-scale molecular dynamics simulations on Titan to model the beginnings of ruptures in thin films wetting a solid substrate.

Turbomachinery

Ramgen Power Systems is using the Titan system to significantly reduce time to solution in the optimization of novel designs based on aerospace shock wave compression technology for gas compression systems, such as carbon dioxide compressors.

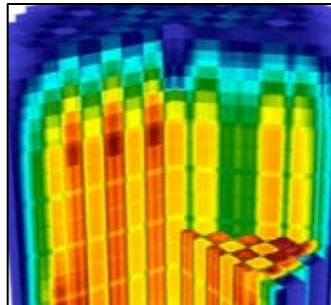


Earthquake Hazard Assessment

To better prepare California for large seismic events SCEC joint researchers have simulated earthquakes at high frequencies allowing structural engineers to conduct realistic physics-based probabilistic seismic hazard analyses.

Nuclear Reactors

Center for the Advanced Simulation of Lightwater Reactors (CASL) investigators successfully performed full core physics power-up simulations of the Westinghouse AP1000 pressurized water reactor core using their Virtual Environment for Reactor Application code that has been designed on Titan architecture.



Climate Science

Simulations on OLCF resources recreated the evolution of the climate state during the first half of the last deglaciation period allowing scientists to explain the mechanisms that triggered the last period of Southern Hemisphere warming and deglaciation.

Two Architecture Paths for Today and Future Leadership 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

Where do we go from here?

- Provide the Leadership computing capabilities needed for the DOE Office of Science mission from 2018 through 2022
 - Capabilities for INCITE and ALCC science projects
- CORAL was formed by grouping the three Labs who would be acquiring Leadership computers in the same timeframe (2017).
 - Benefits include:
 - Shared technical expertise
 - Decreases risks due to the broader experiences, and broader range of expertise of the collaboration
 - Lower collective cost for developing and responding to RFP

CORAL Collaboration ORNL, ANL, LLNL)

Objective - Procure 3 leadership computers to be sited at Argonne, Oak Ridge and Lawrence Livermore in 2017. Two of the contracts have been awarded with the Argonne contract in process.

Current DOE Leadership Computers

Titan (ORNL)
2012 - 2017



Sequoia (LLNL)
2012 - 2017



Mira (ANL)
2012 - 2017



Leadership Computers RFP requests >100 PF, 2 GB/core main memory, local NVRAM, and science performance 4x-8x Titan or Sequoia

Approach

- Competitive process - one RFP (issued by LLNL) leading to 2 R&D contracts and 3 computer procurement contracts
- For risk reduction and to meet a broad set of requirements, 2 architectural paths will be selected and Oak Ridge and Argonne must choose different architectures
- Once Selected, Multi-year Lab-Awardee relationship to co-design computers
- Both R&D contracts jointly managed by the 3 Labs
- Each lab manages and negotiates its own computer procurement contract, and may exercise options to meet their specific needs
- Understanding that long procurement lead-time may impact architectural characteristics and designs of procured computers

CORAL Evaluation Process

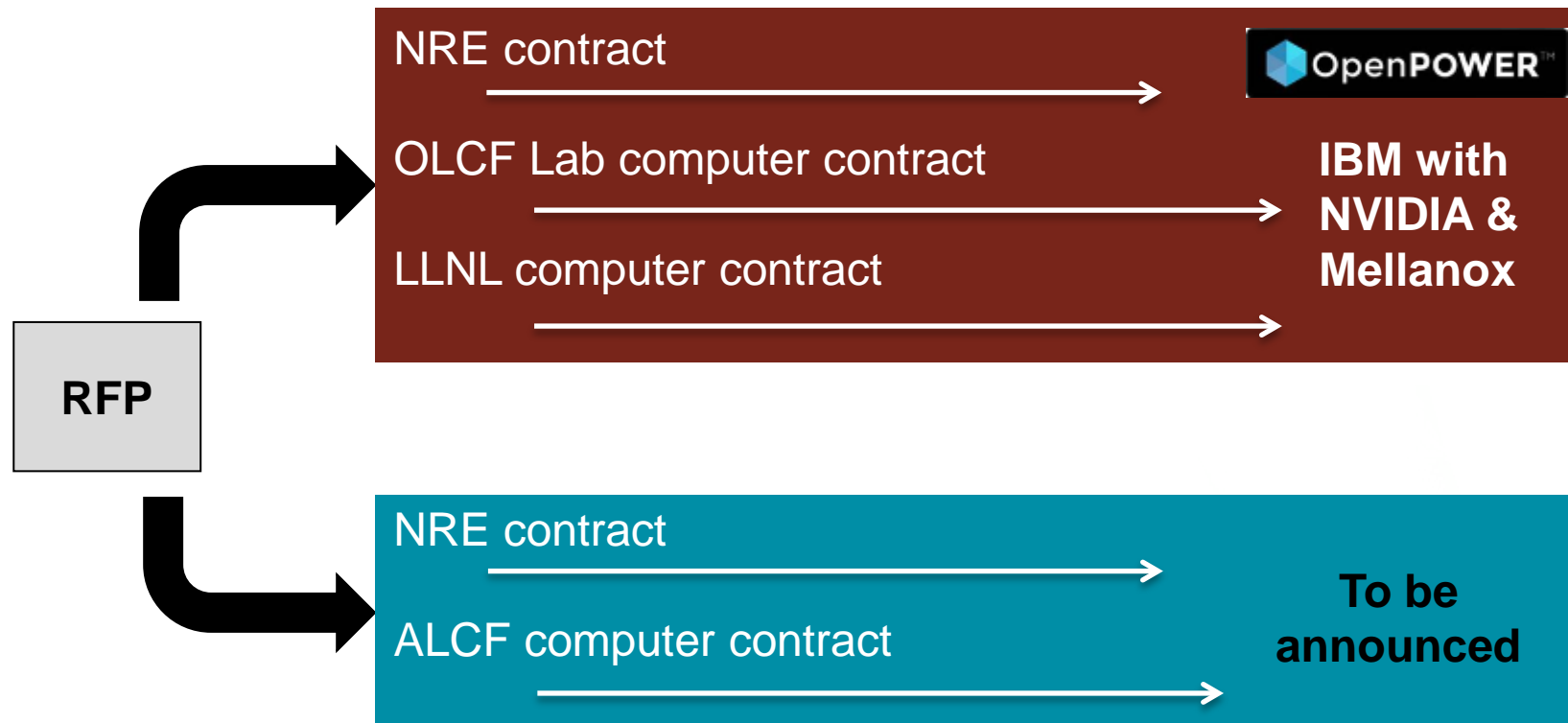
A buying team consisting of the management, technical, and procurement leadership of the three CORAL labs met to select the **set of two proposals** that provided the best value to the government

Evaluation Criteria:

- DOE mission requirements - the best overall **combination** of solutions
- Technical proposal excellence; projected performance on the applications is the single most important criterion
- Feasibility of schedule and performance
- Diversity
- Overall Price
- Supplier attributes
- Risk evaluation

Results of CORAL Procurement

Two Diverse Architecture Paths



2017 OLCF Leadership System

Hybrid CPU/GPU architecture



Vendor: IBM (Prime) / NVIDIA™ / Mellanox Technologies®



At least 5X Titan's Application Performance

Approximately 3,400 nodes, each with:

- Multiple IBM POWER9 CPUs and multiple NVIDIA Tesla® GPUs using the NVIDIA Volta™ architecture
- CPUs and GPUs completely connected with high speed NVLink™
- Large coherent memory: over 512 GB (HBM + DDR4)
 - all directly addressable from the CPUs and GPUs
- An additional 800 GB of NVRAM, which can be configured as either a burst buffer or as extended memory
- over 40 TF peak performance

Dual-rail Mellanox® EDR-IB full, non-blocking fat-tree interconnect

IBM Elastic Storage (GPFS™) - 1TB/s I/O and 120 PB disk capacity.

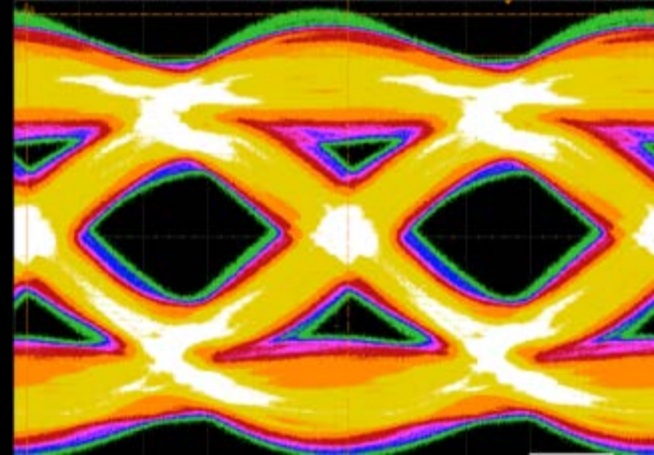


INTRODUCING NVLINK AND STACKED MEMORY

TRANSFORMATIVE TECHNOLOGY FOR 2016 WITH POWER 8+[®], AND BEYOND

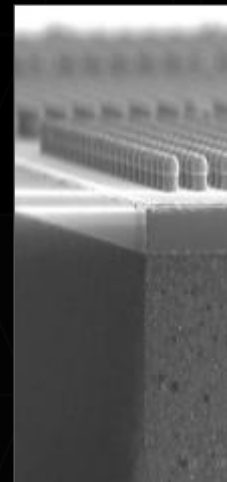
NVLINK

- GPU high speed interconnect
- 5X-12X PCI-E Gen3 Bandwidth
- Planned support for POWER[®] CPUs



Stacked (HBM) Memory

- 4x Higher Bandwidth (~1 TB/s)
- 3x Larger Capacity
- 4x More Energy Efficient per bit

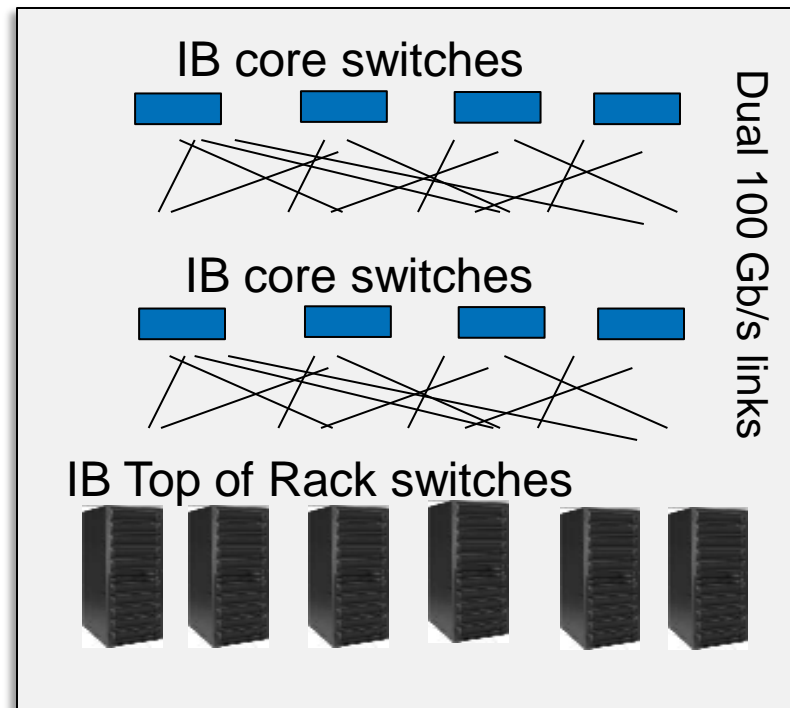


Summit's High-Speed Interconnect

Mellanox Technologies® Dual-Rail EDR Infiniband

InfiniBand Interconnect Three-level Fat Tree Interconnect

- 3-Level Fat Tree
- 23 GB/s (dual plane 100Gb/s)
- 5 hops max
- Adaptive routing



Summit Key Software Components

- **System**

- Linux®
- IBM Elastic Storage (GPFS™)
- IBM Platform Computing™ (LSF)
- IBM Platform Cluster Manager™ (xCAT)

- **Programming Environment**

- Compilers supporting OpenMP and OpenACC
 - 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

How does Summit compare to Titan

| Feature | Summit | Titan |
|---|-----------------------------|--------------------------------|
| Application Performance | 5-10x Titan | Baseline |
| Number of Nodes | ~3,400 | 18,688 |
| Node performance | > 40 TF | 1.4 TF |
| Memory per Node | >512 GB (HBM + DDR4) | 38GB (GDDR5+DDR3) |
| NVRAM per Node | 800 GB | 0 |
| Node Interconnect | NVLink (5-12x PCIe 3) | PCIe 2 |
| System Interconnect (node injection bandwidth) | Dual Rail EDR-IB (23 GB/s) | Gemini (6.4 GB/s) |
| Interconnect Topology | Non-blocking Fat Tree | 3D Torus |
| Processors | IBM POWER9 NVIDIA Volta™ | AMD Opteron™ NVIDIA Kepler™ |
| File System | 120 PB, 1 TB/s, GPFS™ | 32 PB, 1 TB/s, Lustre® |
| Peak power consumption | 10 MW | 9 MW |

Questions?



U.S. DEPARTMENT OF
ENERGY

Office of
Science



GPU Hackathon October 2014

This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725

