CORAL
COLLABORATION
OAK RIDGE • ARGONNE • LIVERMORE

SIERRA

Matt Leininger
Deputy Advanced Technology Projects
Livermore Computing

OpenFabrics Workshop

March 18, 2015

Lawrence Livermore National Laboratory

LLNL-PRES-666680

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

CORAL is a DOE NNSA & Office of Science project to procure 3 leadership computers for ANL, ORNL, & LLNL with delivery in CY17-18

Modeled on successful LLNL/ANL/IBM Blue Gene partnership (Sequoia/Mira)

LLNL's IBM Blue Gene Systems

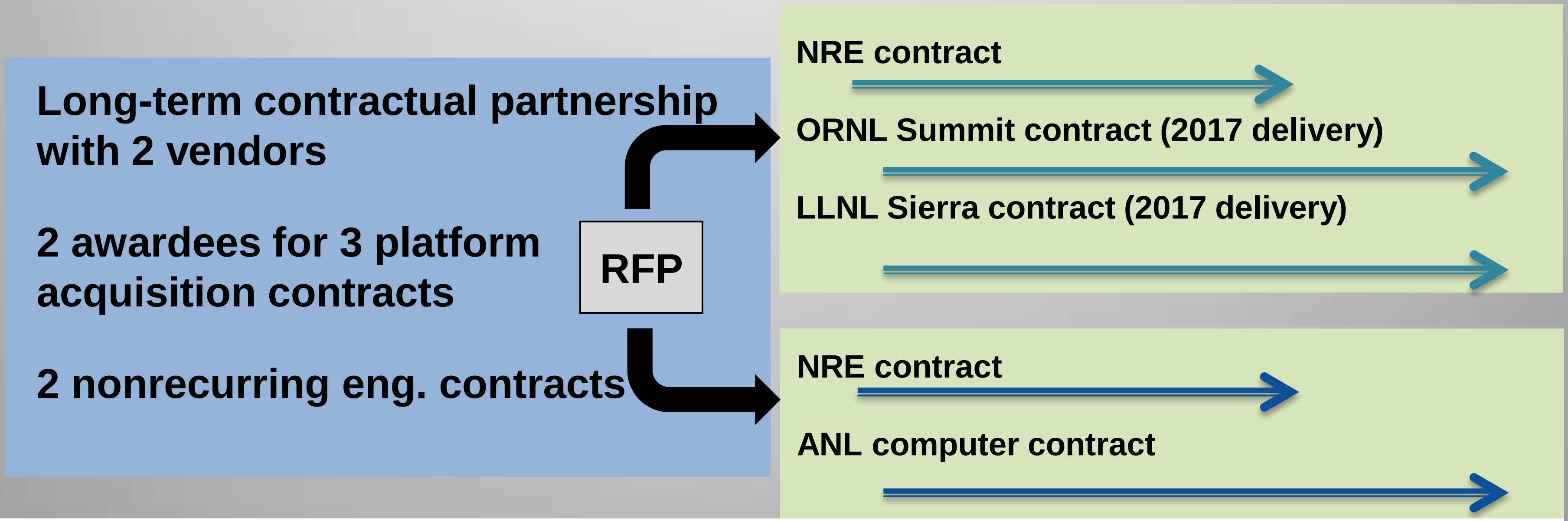
BG/L

BG/P

Dawn

BG/Q

Sequoia



Long-term contractual partnership with 2 vendors

2 awardees for 3 platform acquisition contracts

2 nonrecurring eng. contracts

RFP

NRE contract

ORNL Summit contract (2017 delivery)

LLNL Sierra contract (2017 delivery)

NRE contract

ANL computer contract

CORAL is the next major phase in the U.S. Department of Energy's scientific computing roadmap and path to exascale computing

High Level System Requirements

- Target speedup over current systems of 4x on Scalable benchmarks and 6x on Throughput benchmarks
- Peak Performance ≥ 100 PF
- Aggregate memory of 4 PB and ≥ 1 GB per MPI task (2 GB preferred)
- Maximum power consumption of system and peripherals ≤ 20MW
- Mean Time Between Application Failure that requires human intervention ≥ 6 days
- Architectural Diversity
- Delivery in 2017 with acceptance in 2018

Application Performance Requirements are the Highest Priority to CORAL

An average “figure of merit” (FOM) improvement of 4-8X for scalable science apps and 6-12X for throughput apps over today’s DOE systems.

- The Offerors provided actual, predicted and/or extrapolated performance results for the proposed system for the following:

- CORAL system performance (TR-1)
  - Average FOM over four TR-1 scalable science apps >= 4.0
  - Average FOM over four TR-1 throughput apps >= 6.0
  - Raw results for three TR-1 Data Centric apps and five TR-1 skeleton apps

Example “figures of merit” are number of years simulated per day, and number of particles pushed per second

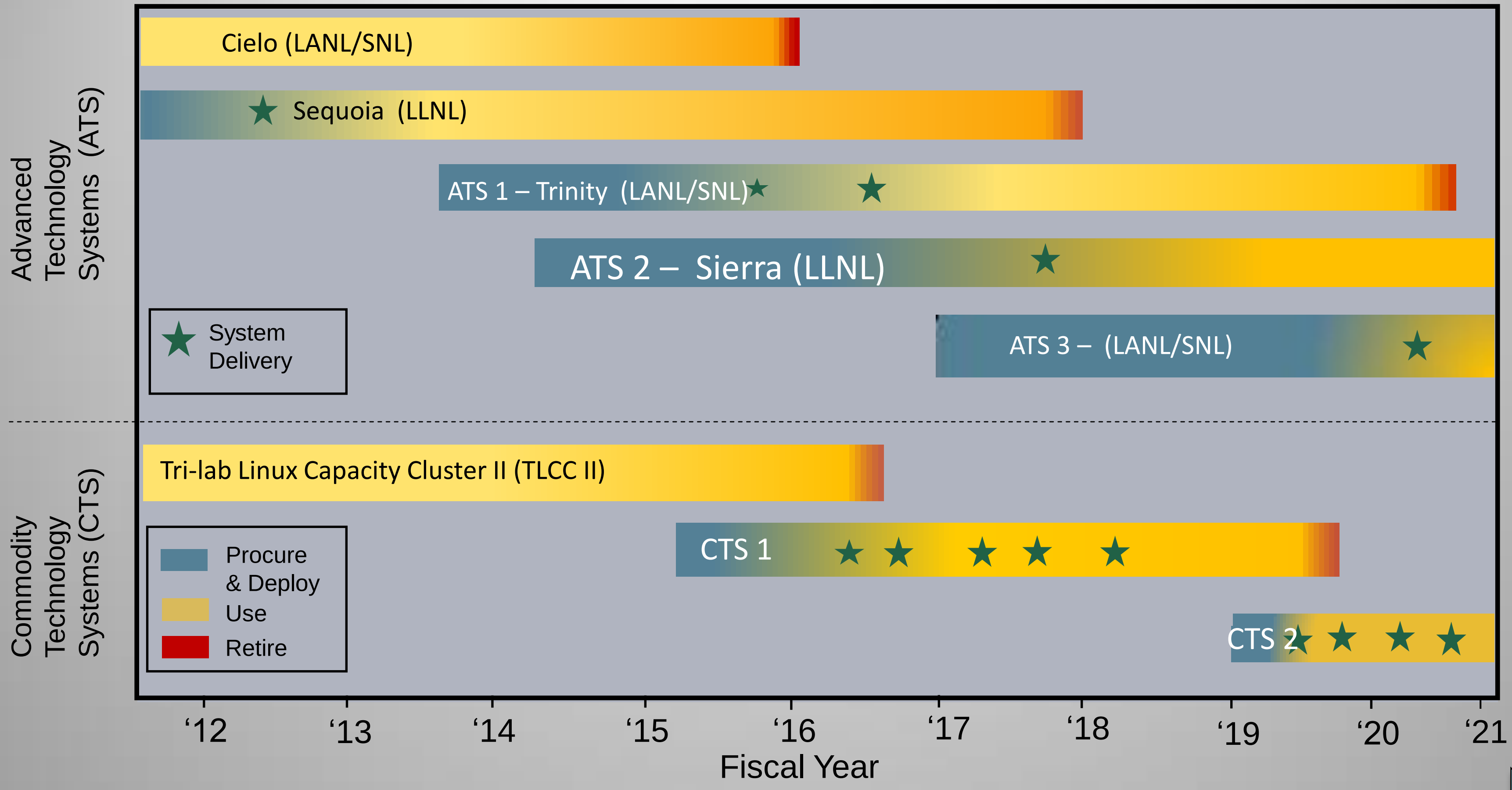
Sierra workloads were derived directly from the needs to fulfill NNSA's Advanced Simulation and Computing (ASC) mission

Sierra will provide computational resources that are essential for nuclear weapon scientists to fulfill the stockpile stewardship mission through simulation in lieu of underground testing.

Two broad simulation classes constitute Sierra's workload

#1 Assess the performance of integrated nuclear weapon systems
#2 Perform weapon's science and engineering calculations

NNSA's Advanced Simulation and Computing (ASC) Platform Timeline

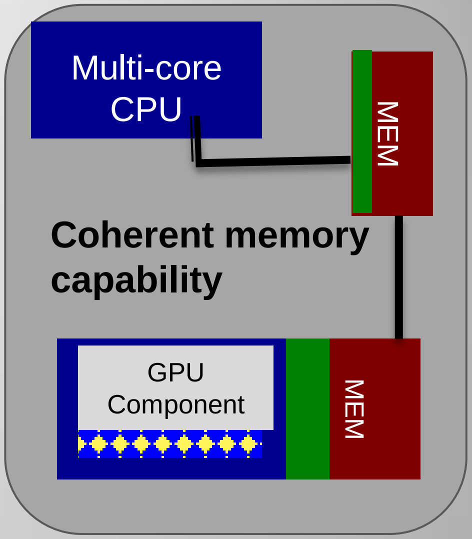


ASC Platform Strategy includes application code transition for all platforms

LLNL selected the most compelling system for NNSA

- Unmodified codes will run on Power® Architecture processor
- Memory rich nodes; high node memory bandwidth
- Volta™ GPUs provide substantial performance potential
- Outstanding benchmark analysis by IBM + NVIDIA
- Cost competitive; low risk solution; outstanding hardware reliability

Notional Sierra node



NRE contract provides significant benefit

- Center of Excellence - expert help with porting and optimizing actual applications
- Motherboard design and novel cooling concept
- GPU reliability; file system performance; open source compiler infrastructure
- Advanced system diagnostics and scheduling; advanced networking capabilities

Sierra System

Compute Node

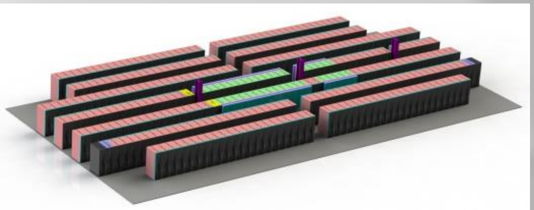
POWER® Architecture Processor
NVIDIA®Volta™
NVMe-compatible PCIe 800GB SSD
> 512 GB DDR4 + HBM
Coherent Shared Memory

Compute Rack

Standard 19”
Warm water cooling

Compute System

2.1 – 2.7 PB Memory
120 -150 PFLOPS
10 MW



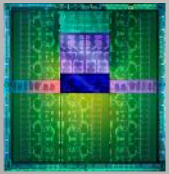
Components

IBM POWER

- NVLink™

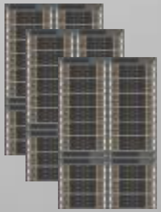
NVIDIA Volta

- HBM
- NVLink



Mellanox® Interconnect

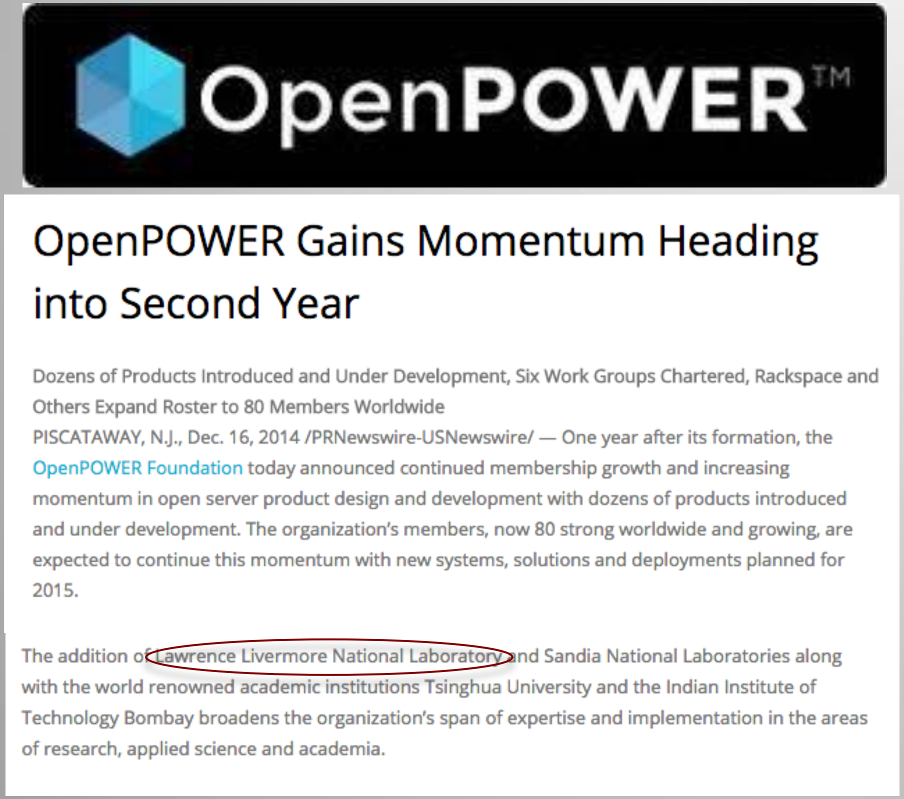
Dual-rail EDR Infiniband®



GPFS™ File System

120 PB usable storage
1.2/1.0 TB/s R/W
bandwidth

OpenPower Foundation

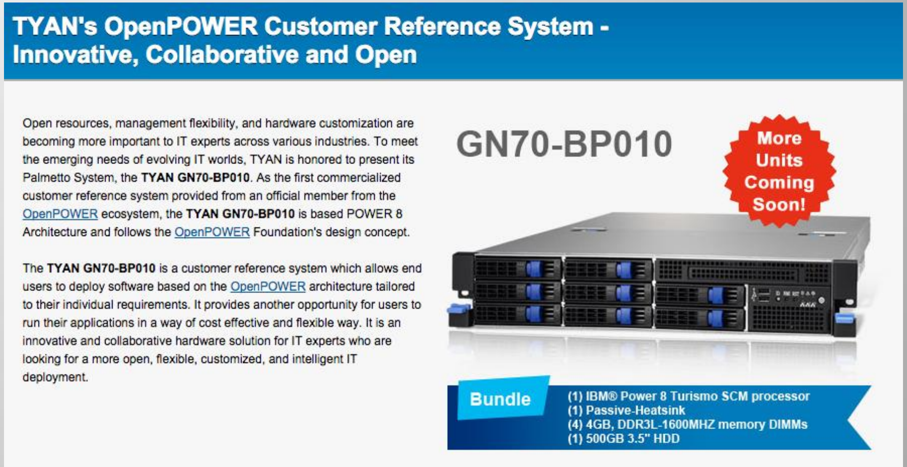


OpenPOWER Gains Momentum Heading into Second Year

Dozens of Products Introduced and Under Development, Six Work Groups Chartered, Rackspace and Others Expand Roster to 80 Members Worldwide

PISCATAWAY, N.J., Dec. 16, 2014 /PRNewswire-USNewswire/ — One year after its formation, the OpenPOWER Foundation today announced continued membership growth and increasing momentum in open server product design and development with dozens of products introduced and under development. The organization's members, now 80 strong worldwide and growing, are expected to continue this momentum with new systems, solutions and deployments planned for 2015.

The addition of Lawrence Livermore National Laboratory and Sandia National Laboratories along with the world renowned academic institutions Tsinghua University and the Indian Institute of Technology Bombay broadens the organization's span of expertise and implementation in the areas of research, applied science and academia.



TYAN's OpenPOWER Customer Reference System - Innovative, Collaborative and Open

Open resources, management flexibility, and hardware customization are becoming more important to IT experts across various industries. To meet the emerging needs of evolving IT worlds, TYAN is honored to present its Palmetto System, the **TYAN GN70-BP010**. As the first commercialized customer reference system provided from an official member from the OpenPOWER ecosystem, the **TYAN GN70-BP010** is based POWER 8 Architecture and follows the OpenPOWER Foundation's design concept.

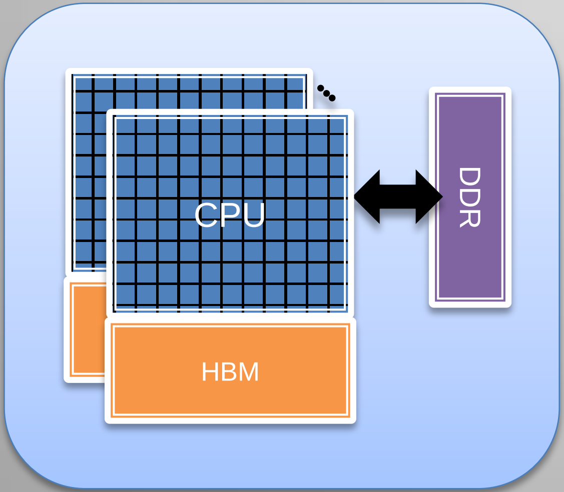
The **TYAN GN70-BP010** is a customer reference system which allows end users to deploy software based on the OpenPOWER architecture tailored to their individual requirements. It provides another opportunity for users to run their applications in a way of cost effective and flexible way. It is an innovative and collaborative hardware solution for IT experts who are looking for a more open, flexible, customized, and intelligent IT deployment.

(1) IBM® Power 8 Turismo SCM processor
(1) Passive-Heatsink
(4) 4GB, DDR3L-1600MHZ memory DIMMs
(1) 500GB 3.5" HDD

Architectural Paths to Exascale

Many-Core (like Trinity)

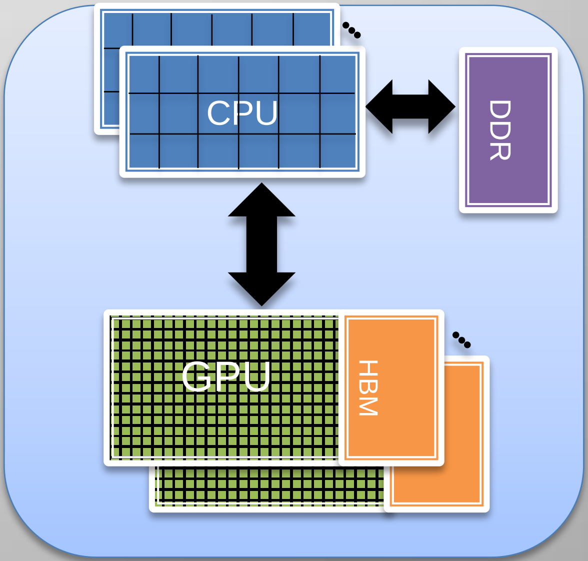
- Many relatively small homogeneous compute cores
- 10's of thousands of nodes with millions of cores
- Multiple levels of memory
- Single/Dual rail high performance network



Notional Many-Core Node

Hybrid Multi-Core (like Sierra)

- Multiple CPUs and accelerators per node
- Small(ish) number of very powerful nodes
- Multiple levels of memory
- Multi-rail high performance network



Notional Hybrid Multi-Core Node

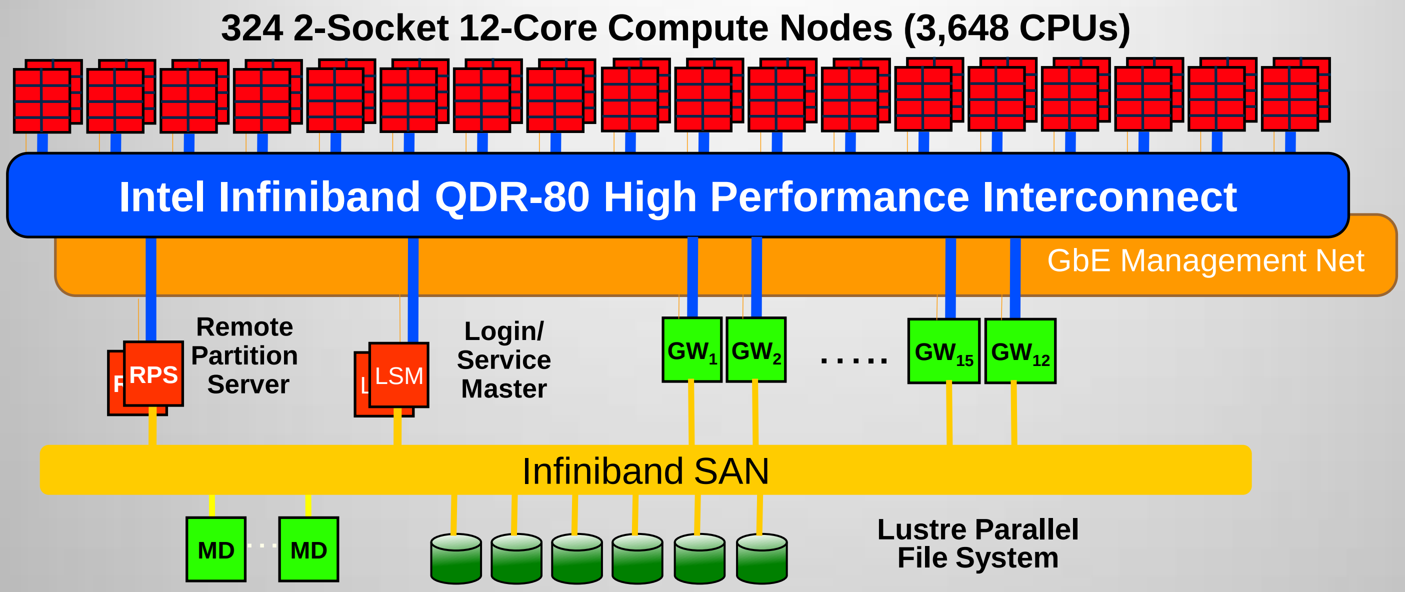
CORAL Working Groups are the Focal Point of Interactions with IBM Partnership

- Eight working groups coordinate technical interactions across all CORAL partners
  - NNSA Tri-laboratories
  - Office of Science CORAL Laboratories (ORNL and Argonne)
  - Vendor partners: IBM, NVIDIA and Mellanox
- Centers of Excellence are effectively a 9th working group focused on applications
- Ensure final systems meet DOE requirements
- Provide co-design directly related to NRE milestones and other system aspects

- Burst Buffer (LLNL Lead: Mark Gary)
- Compilers (LLNL Lead: John Gyllenhaal)
- GPFS (LLNL Lead: Mark Gary)
- Hardware (LLNL Lead: Bronis R. de Supinski)
- Messaging (LLNL Lead: Matt Leininger)
- SRM & LSF (LLNL Lead: Greg Tomaschke)
- System Administration and Management (LLNL Lead: Robin Goldstone)
- Tools (LLNL Lead: Dong Ahn)

DATA INTENSIVE

Catalyst Data Intensive HPC Cluster



324 2-Socket 12-Core Compute Nodes (3,648 CPUs)

Intel Infiniband QDR-80 High Performance Interconnect

GbE Management Net

RPS

Remote Partition Server

LSM

Login/ Service Master

GW_1 GW_2 ..... GW_{15} GW_{12}

Infiniband SAN

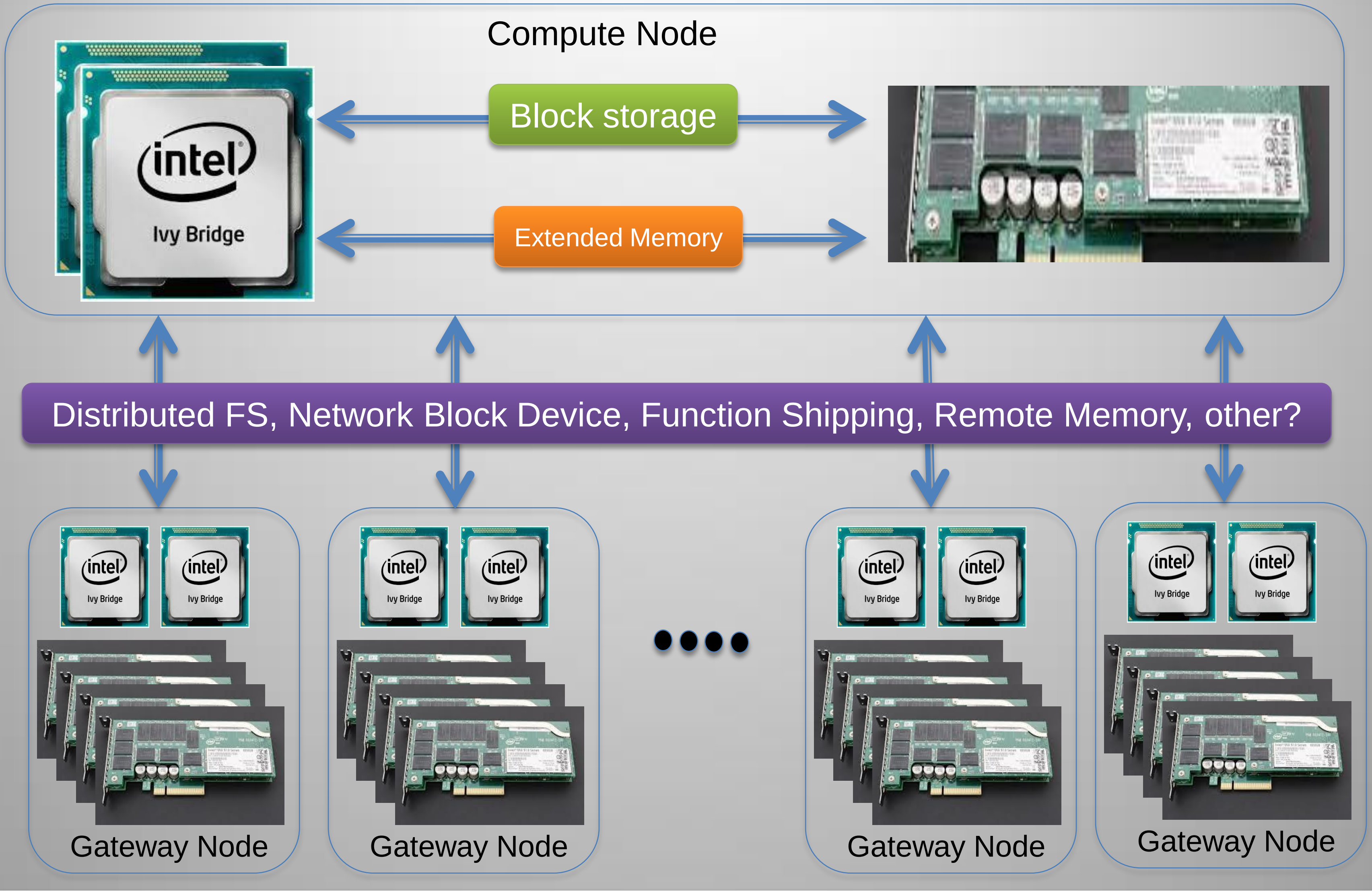
MD ... MD

Lustre Parallel File System

Catalyst Node Configuration

- Dual socket 12-core Intel Ivy Bridge processors @ 2.4Ghz
- 8x16GB DDR3-1866 DIMM, 128 GB memory (5.3 GB/core)
- 60 GB/s memory BW x 2 = 120 GB/s
- Intel QDR-80 dual rail Infiniband – one QDR link per CPU socket
- Intel 910 PCIe NVRAM (1.8 GB/s R; 1.3 GB/s W)
  - 800GB per compute node
  - 3.2TB per gateway node

Catalyst allows exploration of usage models for on-node and network attached NVRAM



Compute Node

Block storage

Extended Memory

Distributed FS, Network Block Device, Function Shipping, Remote Memory, other?

Gateway Node

Gateway Node

••••

Gateway Node

Gateway Node

Traverse huge graphs using advanced architectures

Facilitate the processing of Massive Real-World Graphs

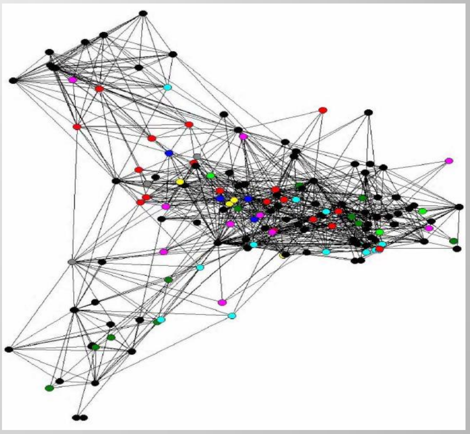
- Billions of vertices and trillions of edges (e.g. Social Networks, Web Graphs)
- Too large to fit in main memory of a workstation

Utilize emerging Non-volatile memory storage technologies (e.g., NAND Flash)

- Develop techniques that tolerate data latencies

We have developed:

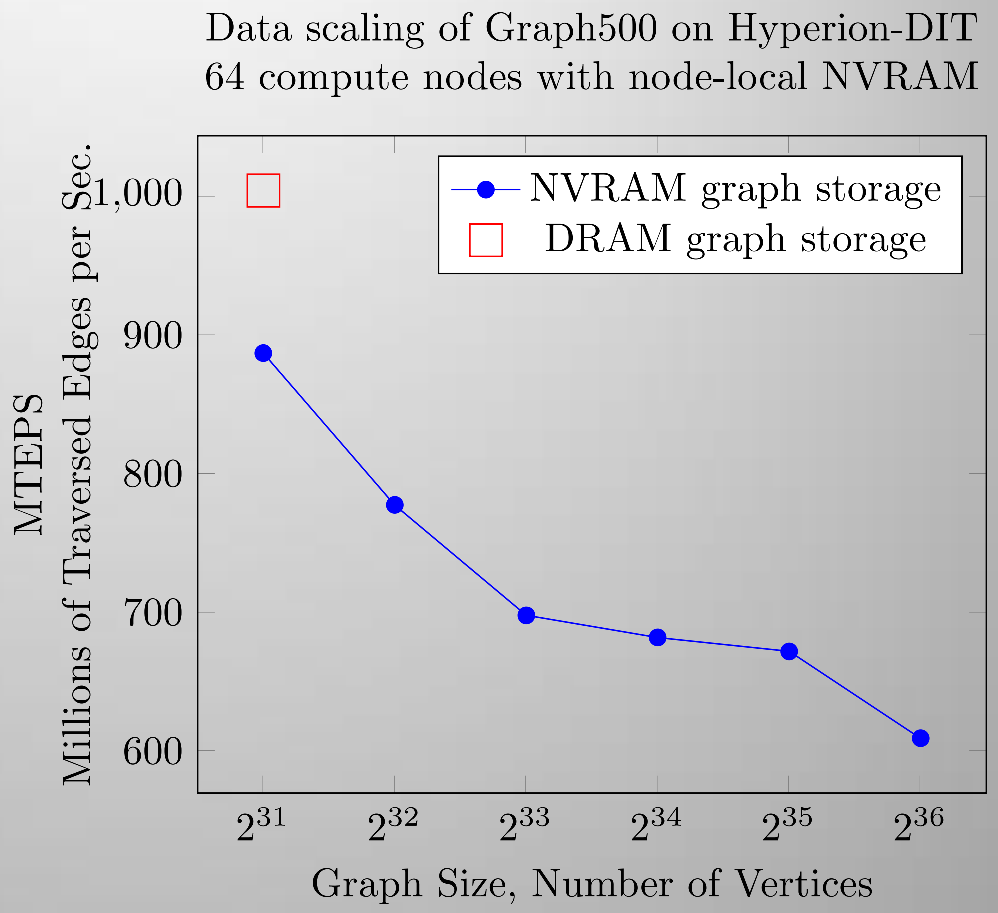
- **Parallel asynchronous technique** that outperforms parallel competitors in shared-memory
- Scale to trillion-edge graphs, both shared and distributed memory
- Scale to large HPC clusters containing NVRAM: LLNL's Catalyst



NVRAM augments DRAM to process trillion edge graphs [IPDPS 2013]

Our approach can process 32x larger datasets with only a 39% performance degradation in TEPS by leveraging node-local NVRAM for expanded graph data storage.

By exploiting both distributed memory processing and node-local NVRAM, significantly larger datasets can be processed than with either approach in isolation.



LLNL algorithm submitted to the Graph500 challenge

Using NVRAM [Nov 2014]

- single-node 2TB DRAM with IB connected Flash from Saratoga Speed
  - Scale 37, ~2 trillion edges, 62.7 MTEPS
- 300 nodes of Catalyst
  - Scale 40, ~17.6 trillion edges, 4.17 GTEPS

Using NVRAM -- Scale 36 (~1 trillion edges) [2011]

- Single compute node with PCI-e NAND Flash -- 52 MTEPS
- 144 compute nodes with commodity SATA Flash – 242 MTEPS
  - Trestles @ SDSC
- 64 compute nodes with PCI-e NAND Flash – 609 MTEPS
  - Hyperion DIT (precursor to Catalyst)

Using distributed memory supercomputers (no NVRAM)

- 15% faster than best Graph500 result on the IBM Blue Gene/P “Intrepid”

Summary

- NNSA Sierra represents the next step towards exascale computing
- Exascale solutions are critical for both HPC and Data Analytics
- OpenFabrics will have a strong role in these systems
  - OFA software is a requirement on all LLNL/NNSA Linux clusters (CTS)
- NNSA labs helped form OFA/OpenIB with two goals:
  - Develop first open source & openly developed high performance network stack
  - Provide a mechanism to allow vendors to innovate new networking features

Lawrence Livermore
National Laboratory