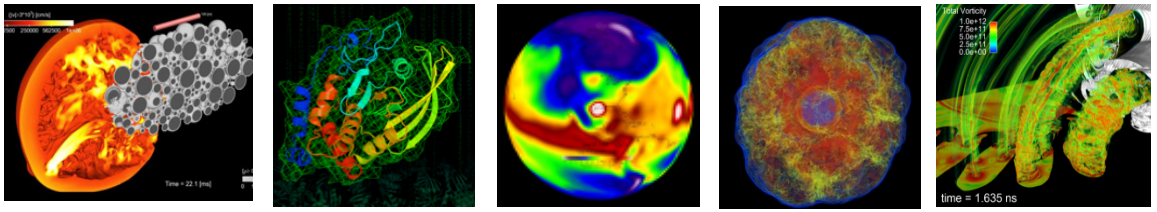
Trinity: Advanced Technology System for the ASC Program

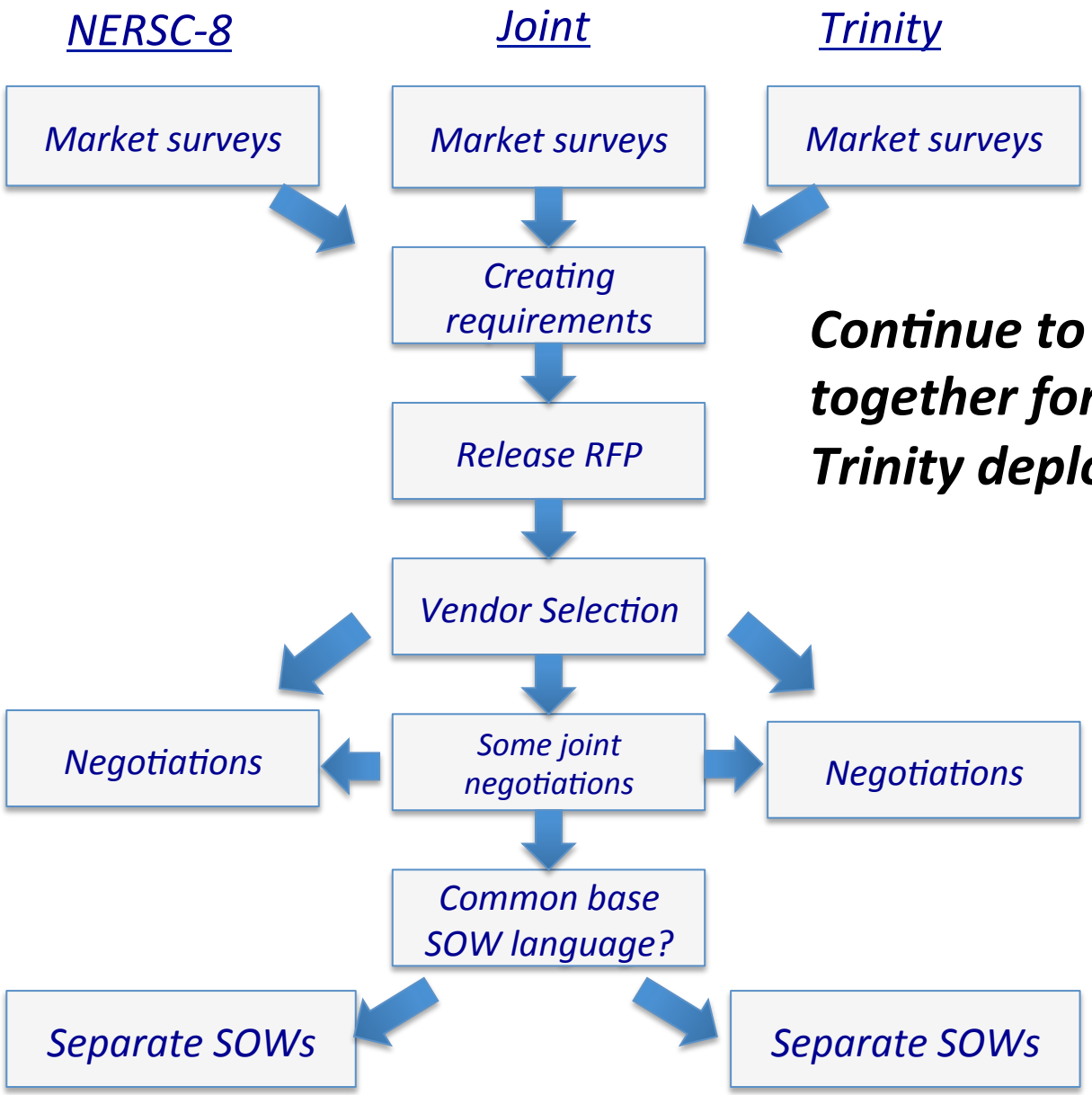


Manuel Vigil
Trinity Project Director
High Performance Computing Division
Los Alamos National Laboratory

HPC User Forum
September 17, 2014

LA-UR-14-27024

NERSC-8 and Trinity team activities



Continue to work together for CORI and Trinity deployments

Topics

- Trinity Status
- ASC Computing Strategy
- Trinity Project Drivers & Mission Need
- The Trinity System
  - High-level architecture overview
  - High-level capabilities
  - Schedule
- Summary

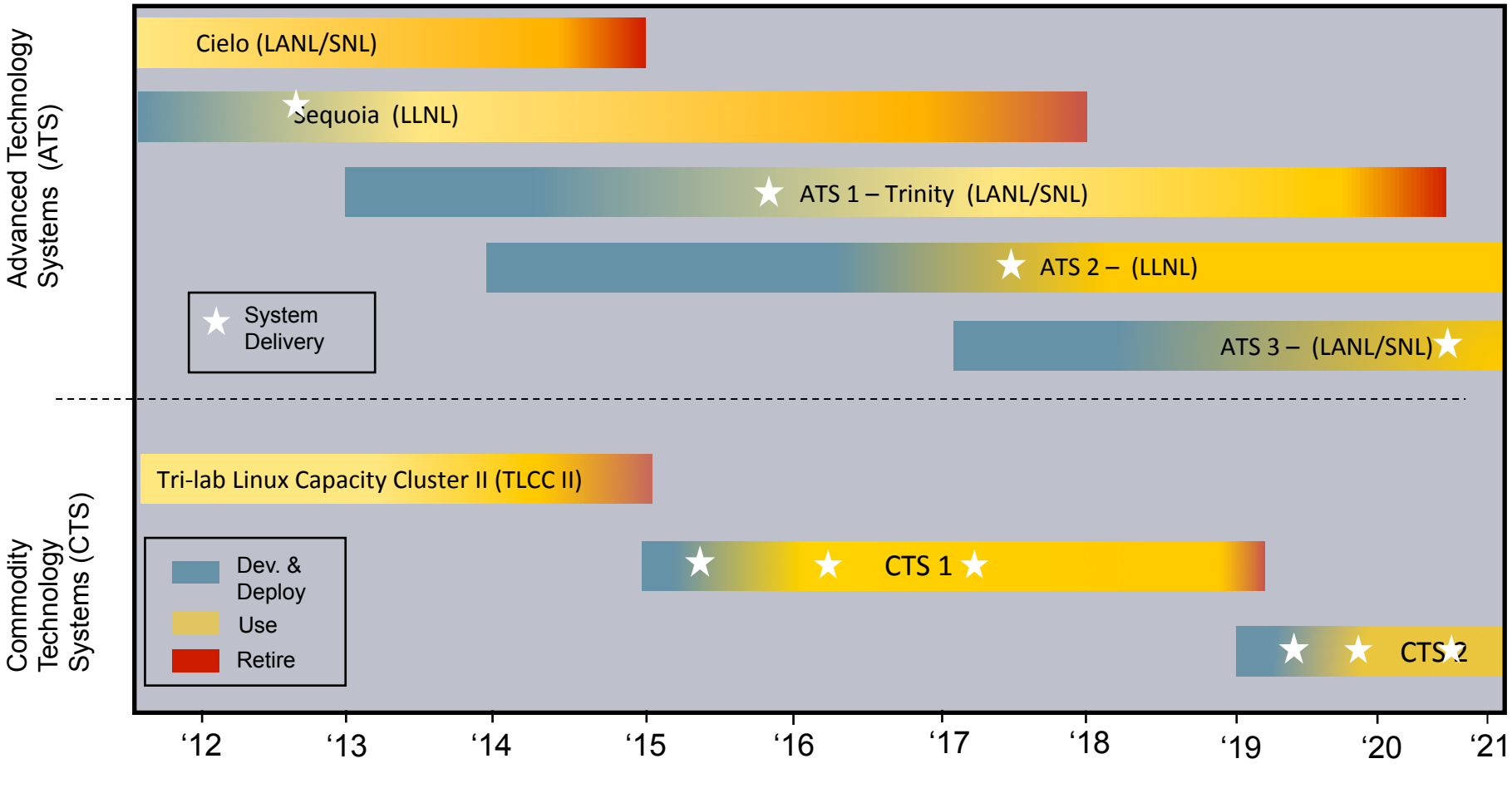
Trinity Status

- Formal Design Review and Independent Project Review completed in 2013
- Trinity/NERSC8 RFP released August 2013
- Technical Evaluation of the proposals completed September 2013
- Initial negotiations for both systems completed November 2013
- Initial Trinity system procurement completed in late 2013/early 2014
  - Before final approval Trinity went back to the proposing vendors for a Best and Final Offer (BAFO)
  - Target delivery date of September 2015 is unchanged
- Trinity Best and Final Offer RFP released to vendors March 2014
- Trinity proposal evaluations and negotiations completed April 2014
- Trinity Procurement Approval – Notice of Consent from NNSA received May 16, 2014
- Trinity Independent Cost Review completed May 2014
- Trinity CD-2/3b approved July 3, 2014
- Trinity Contract Awarded to Cray, Inc. on July 9, 2014

ASC Computing Strategy

- Approach: Two classes of systems
  - Advanced Technology: First-of-a-kind systems that identify and foster technical capabilities and features that are beneficial to ASC applications
  - Commodity Technology: Robust, cost-effective systems to meet the day-to-day simulation workload needs of the program
- Advanced Technology Systems
  - Leadership-class platforms
  - Pursue promising new technology paths
  - These systems are to meet unique mission needs *and* to help prepare the program for future system designs
  - Includes Non-Recurring Engineering (NRE) funding to enable delivery of leading-edge platforms
  - Acquire right-sized platforms to meet the mission needs
  - Trinity will be deployed by ACES (New Mexico Alliance for Computing at Extreme Scale, i.e., Los Alamos & Sandia)

ASC Platform Timeline



ASC Computing Strategy includes application code transition for all platforms

Trinity Project Drivers and Mission Need

- Satisfy the mission need for more capable platforms
  - Trinity is designed to support the largest, most demanding ASC applications
  - Increases in geometric and physics fidelities while satisfying analysts' time-to- solution expectations
- Mission Need developed with tri-lab input
- Trinity will support the tri-lab applications community at LLNL, SNL, and LANL
- Mission Need Requirements are primarily driving memory capacity
  - Over 2 PB of aggregate main memory
  - Trinity is sized to run several jobs using about 750 TBytes of memory

Overview of Trinity Award

- Subcontractor
  - Cray, Inc.
- Firm Fixed Price Subcontract:
  - Trinity Platform (including File System)
  - Burst Buffer
  - 2 Application Regression Test Systems
  - 1 System Development Test System
  - On-site System and Application Analysts
  - Center of Excellence for Application Transition Support
  - Advanced Power Management
  - Trinity System Maintenance

Trinity Platform

- Trinity is a single system that contains both Intel Haswell and Knights Landing processors
  - Haswell partition satisfies FY15 mission needs (well suited to existing codes).
  - KNL partition delivered in FY16 results in a system significantly more capable than current platforms and provides the application developers with an attractive next-generation target (and significant challenges)
  - Aries interconnect with the Dragonfly network topology

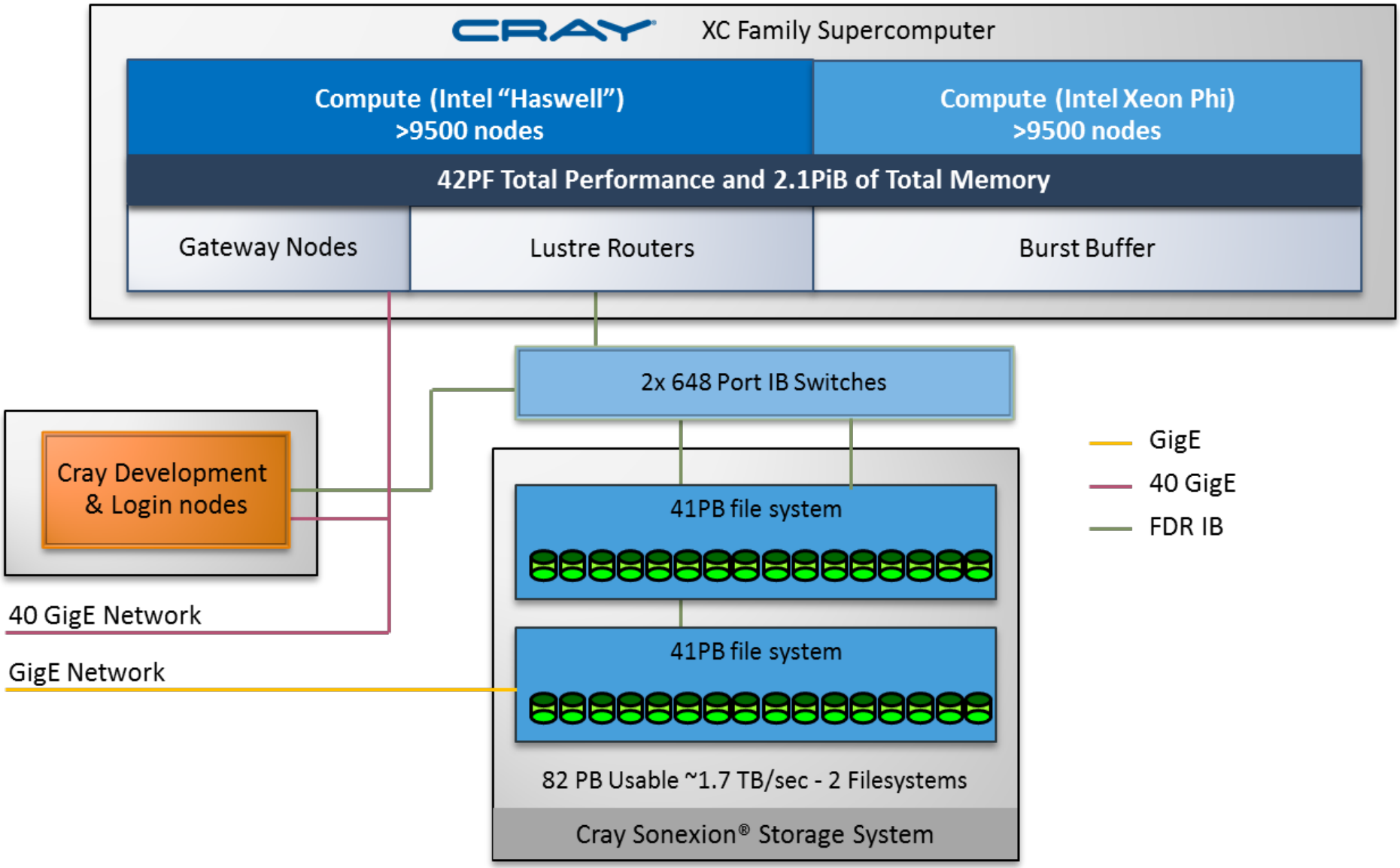
- Based on mature Cray XC30 architecture with Trinity introducing new architectural features
  - Intel Knights Landing (KNL) processors
  - Burst Buffer storage nodes
  - Advanced power management system software enhancements

Trinity Platform

- Trinity is enabling new architecture features in a production computing environment
  - Trinity's architecture will introduce new challenges for code teams: transition from multi-core to many-core, high-speed on-chip memory subsystem, wider SIMD/vector units
  - Tightly coupled solid state storage serves as a "burst buffer" for checkpoint/restart file I/O & data analytics, enabling improved time-to- solution efficiencies
  - Advanced power management features enable measurement and control at the system, node, and component levels, allowing exploration of application performance/watt and reducing total cost of ownership

- Managed Risk
  - Cray XC30 architecture minimizes system software risk and provides a mature high-speed interconnect
  - Haswell partition is low risk as technology; available in Fall CY14
  - KNL is higher risk due to new technology, but provides a reasonable path, and resource, for code teams to transition to the many-core architecture

Cray Compute and Storage Infrastructure for "Trinity"



Trinity Architecture Overview

Metric	Trinity		
Node Architecture	KNL + Haswell	Haswell Partition	KNL Partition
Memory Capacity	2.11 PB	>1 PB	>1 PB
Memory BW	>7PB/sec	>1 PB/s	>1PB/s +>4PB/s
Peak FLOPS	42.2 PF	11.5 PF	30.7 PF
Number of Nodes	19,000+	>9,500	>9500
Number of Cores	>760,000	>190,000	>570000
Number of Cabs (incl I/O & BB)	112		
PFS Capacity (usable)	82 PB usable		
PFS Bandwidth (sustained)	1.45 TB/s		
BB Capacity (usable)	3.7 PB		
BB Bandwidth (sustained)	3.3 TB/s		

Compute Node specifications

	Haswell	Knights Landing
Memory Capacity (DDR)	2x64=128 GB	Comparable to Intel® Xeon® processor
Memory Bandwidth (DDR)	136.5 GB/s	Comparable to Intel® Xeon® processor
# of sockets per node	2	N/A
# of cores	2x16=32	60+ cores
Core frequency (GHz)	2.3	N/A
# of memory channels	2x4=8	N/A
Memory Technology	2133 MHz DDR4	MCDRAM & DDR4
Threads per core	2	4
Vector units & width (per core)	1x256 AVX2	AVX-512
On-chip MCDRAM	N/A	Up to 16GB at launch, over 5x STREAM vs. DDR4

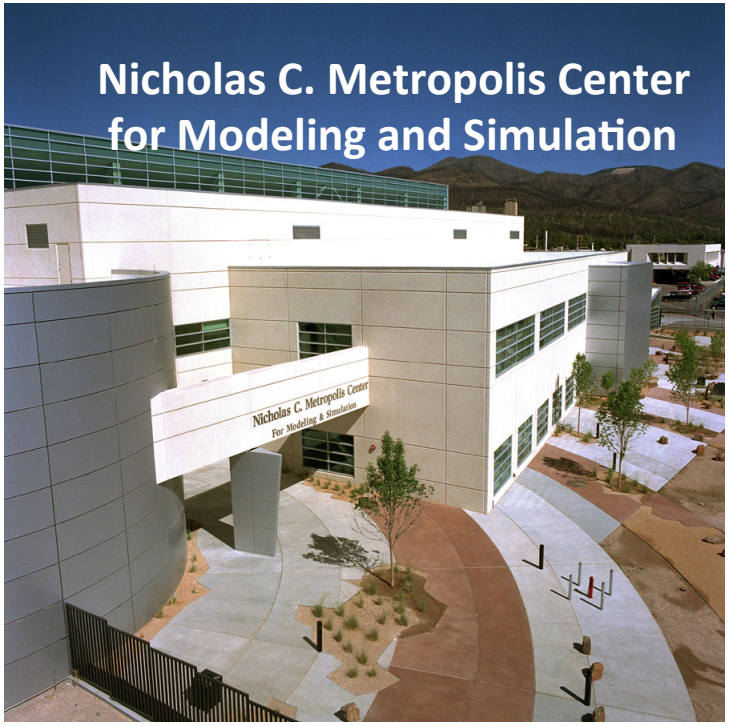
Trinity Capabilities

- Each partition will accommodate 1 to 2 large mission problems
- Capability relative to Cielo
  - 8x to 12x improvement in fidelity, physics and performance
  - > 30x increase in peak FLOPS
  - > 2x increase in node-level parallelism
  - > 6x increase in cores
  - > 20x increase in threads
- Capability relative to Sequoia
  - 2x increase in peak FLOPS
  - Similar complexity relative to core and thread level parallelism

The Trinity Center of Excellence & Application Transition Challenges

- Center of Excellence
  - Work with select NW application code teams to ensure KNL Partition is used effectively upon initial deployment
  - Nominally one application per laboratory (SNL, LANL, LLNL)
  - Chosen such that they impact the NW program in FY17
  - Facilitate the transition to next-generation ATS code migration issues
  - This is NOT a benchmarking effort
- Intel Knights Landing processor
  - From multi-core to many-core
  - > 10x increase in thread level parallelism
  - A reduction in per core throughput (1/4 to 1/3 the performance of a Xeon core)
  - MCDRAM: Fast but limited capacity (~5x the BW, ~1/5 the capacity of DDR4 memory)
  - Dual AVX-512 SIMD units: Does your code vectorize?
- Burst Buffer
  - Data analytics use cases need to be developed and/or deployed into production codes
  - Checkpoint/Restart should “just work”, although advanced features may require code changes

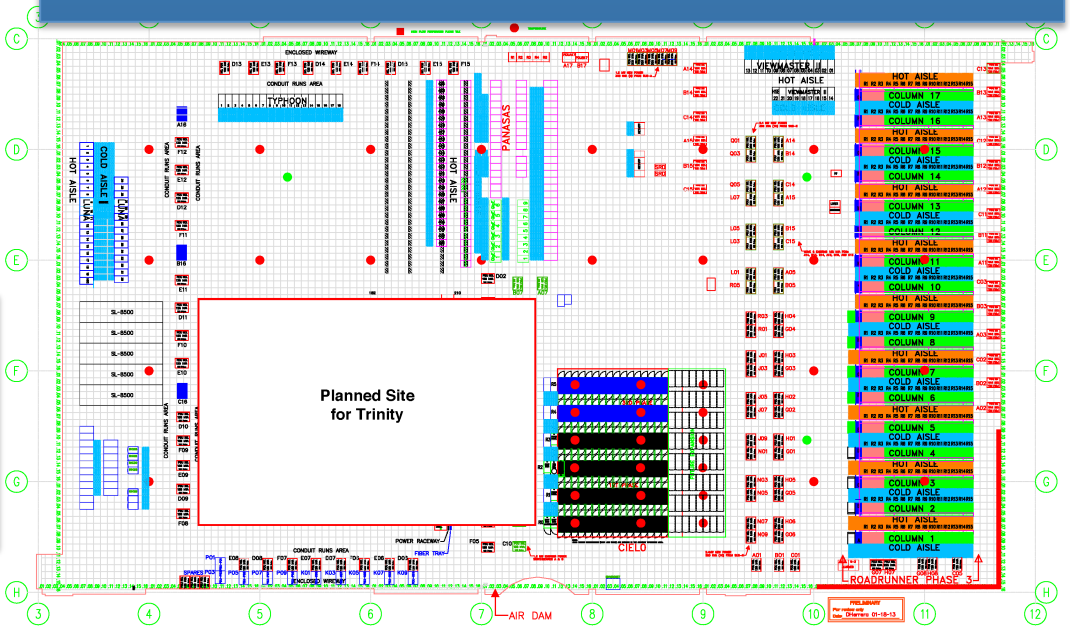
Trinity will be located at the Los Alamos Nicholas C. Metropolis Center for Modeling and Simulation



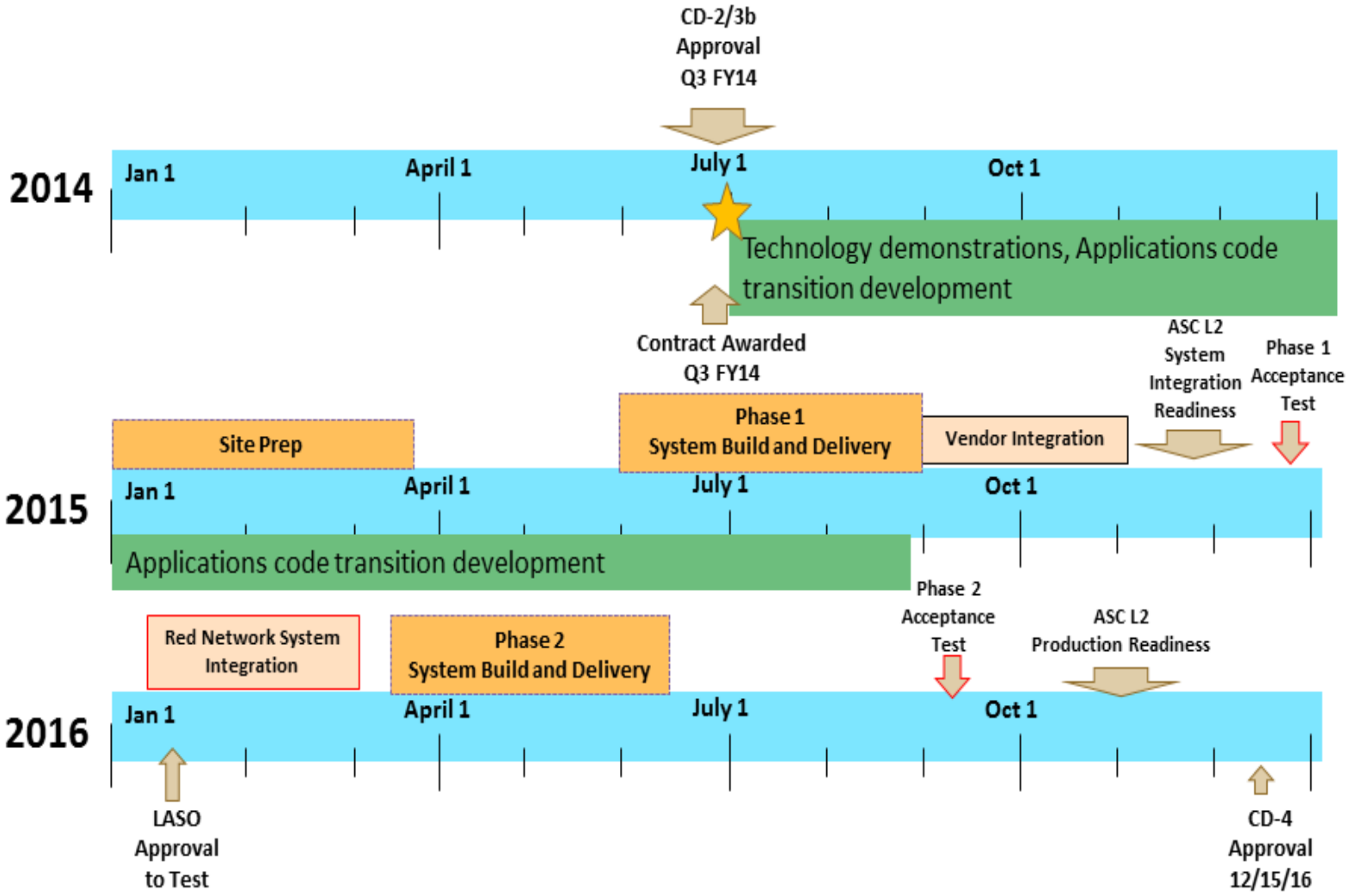
- Classified computing
- 15MW / 12MW water, 3MW air
- 42” subfloor, 300 lbs/sqf
- 80’x100’ (8,000 sqf)

Trinity Power and Cooling

- At least 80% of the platform will be water cooled
- First large water cooled platform at Los Alamos
- Concerns
  - Idle power efficiency
  - Rapid ramp up / ramp down load on power grid over 2MW



Trinity Platform Schedule Highlights 2014-2016



Challenges and Opportunities

- Application transition work using next generation technologies (for Trinity, ATS-2, ATS-3, …)
  - Many-core, hierarchical memory, burst buffer
- Operating a large supercomputer using liquid cooling technology
- Operating Trinity using a mix of Haswell and KNL nodes
- The Burst Buffer concepts and technology for improving application efficiency and exploring other user cases
- On the road to Exascale…

- ….

Trinity System Summary

- Trinity is the first instantiation of the ASC's ATS platform
- Trinity meets or exceeds the goals set by the ACES Design Team
- Relative to Cielo, Trinity will require applications to transition to an MPI+X programming environment and requires increases in thread and vector level parallelism to be exposed
- Trinity introduces Active Power Management, Burst Buffer storage acceleration, and the concept of a Center of Excellence to ASC production platforms

Special thanks to the following for contribution of slides used in this talk

- Doug Doerfler - SNL
- Thuc Hoang – NNSA ASC
- And a cast of others…….

Questions?

- For more info visit the Trinity Web Site:

  trinity.lanl.gov