SAND2014-17093PE

Exceptional service in the national interest

Sandia National Laboratories

Trinity Advanced Technology System Overview

Douglas Doerfler
Distinguished Member of Technical Staff
Sandia National Laboratories
Scalable Computer Architectures Department

SAND 2014-xxxxP
Unlimited Release

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Outline

- ASC ATS Computing Strategy
- Project Drivers and Procurement Process
- Platform Architecture Overview
- Schedule and Status
- Questions, and maybe some answers

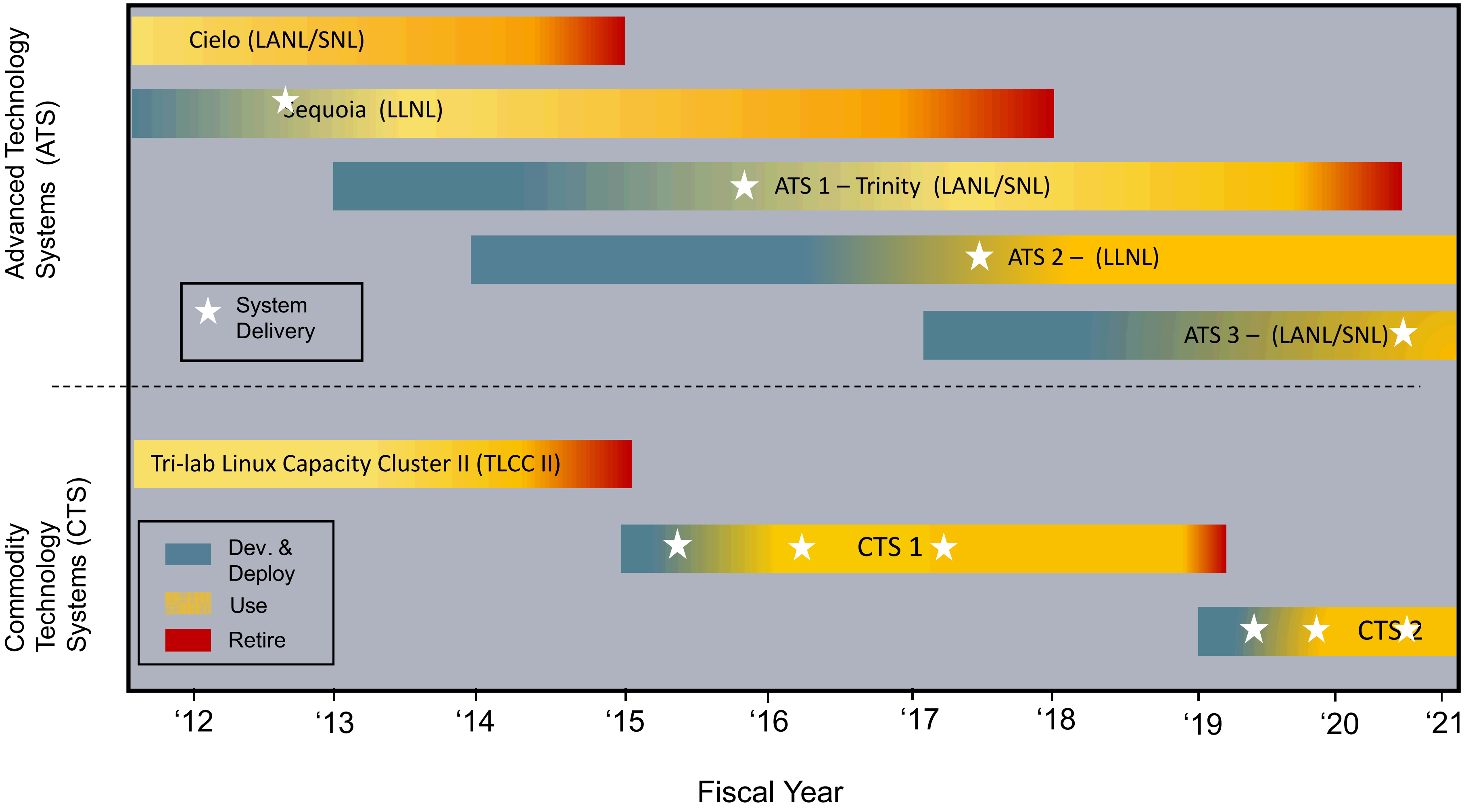
ASC computing strategy

- 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
- Investment Principles
  - Maintain continuity of production
  - Ensure that the needs of the current and future stockpile are met
  - Balance investments in system cost-performance types with computational requirements
  - Partner with industry to introduce new high-end technology constrained by life-cycle costs
  - Acquire right-sized platforms to meet the mission needs

Advanced Technology Systems

- Leadership-class platforms
- Pursue promising new technology paths with industry partners
- 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
- Trinity (ATS-1) will be deployed by ACES (New Mexico Alliance for Computing at Extreme Scale, i.e. Los Alamos & Sandia)
- ATS-2 will be deployed by LLNL

ASC Platform Timeline



Trinity Project Drivers

- 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
  - Foster a competitive environment and influence next generation architectures in the HPC industry
- Trinity is enabling new architecture features in a production computing environment
  - 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
  - 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

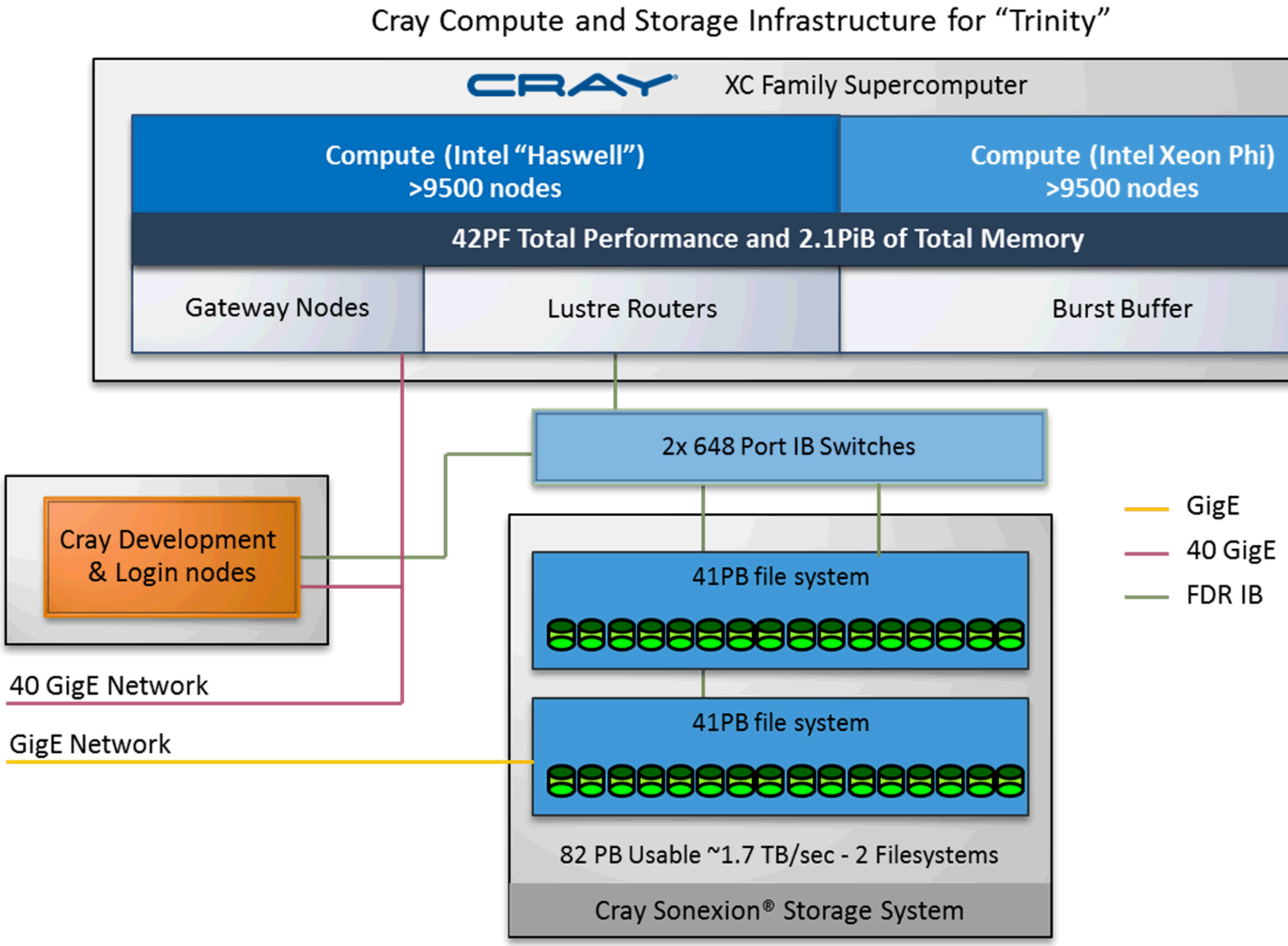
Trinity/NERSC8 Procurement Process Timeline

- ACES (LANL/SNL) Project started late 2011
- Market Survey started January 2012
- Partnered with LBL/NERSC for RFP (NERSC 8)
- Draft Technical Requirements and RFI completed December 2012
- Formal Design Review completed April 2013
- Independent Project Review (Lehman) completed May 2013
- Trinity/NERSC8 RFP released August 2013
- Technical Evaluation of the proposals completed September 2013
- Initial negotiations for both systems completed November 2013
- NNSA Independent Cost Review completed Jan 2014
- NERSC8 awarded April 2014
- Trinity awarded July 2014 after Best and Final Offer (BAFO)

Trinity Platform Solution

- Cray was awarded the contract, July 2014
  - Based on mature Cray XC30 architecture
- with Trinity introducing new architectural features
  - Intel Knights Landing processor
  - Burst Buffer storage nodes
  - Advanced power management system software enhancements
- A single system that contains both Intel Haswell and Knights Landing (KNL) processors
  - Haswell partition satisfies FY15 mission needs (well suited to existing codes) and fits the FY15 budget profile.
  - KNL partition delivered in FY16 results in a system significantly more capable than current platforms, provides Intel with incentive to continue competing in HPC, provides the application developers with an attractive next generation target, and fits the FY16 budget profile.
- Managed Risk
  - Cray XC30 architecture minimizes system software risk and provides a mature high-speed interconnect
  - Haswell partition is low risk as technology is available Fall CY14
  - KNL is higher risk due to new technology, but provides a reasonable path for codes teams to transition to many-core architecture

Trinity High-Level Architecture



Cray Compute and Storage Infrastructure for "Trinity"
CRAY
XC Family Supercomputer
Compute (Intel "Haswell")
>9500 nodes
Compute (Intel Xeon Phi)
>9500 nodes
42PF Total Performance and 2.1PiB of Total Memory
Gateway Nodes
Lustre Routers
Burst Buffer
2x 648 Port IB Switches
Cray Development & Login nodes
40 GigE Network
GigE Network
41PB file system
41PB file system
82 PB Usable ~1.7 TB/sec - 2 Filesystems
Cray Sonexion® Storage System
GigE
40 GigE
FDR IB

Trinity Architecture Details (UUR/public specs)

Metric	Trinity			
Node Architecture	KNL + Haswell		Haswell Partition	KNL Partition
Memory Capacity	2.11 PB		Not disclosed	>1 PB
Memory BW	>7PB/sec		Not disclosed	>1PB/s+>4PB/s
Peak FLOPS	42.2 PF		Not disclosed	30.7 PF
Number of Nodes	19000+		Not disclosed	~10000
Number of Cores	>760,000		Not disclosed	>570,000
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			

Intel KNL Specifications (UUR/public specs)

	Knights Landing
Memory Capacity (DDR)	Comparable to Intel® Xeon® processor
Memory Bandwidth (DDR)	Comparable to Intel® Xeon® processor
# of sockets per node	N/A
# of cores	60+ cores
Core frequency (GHz)	N/A
# of memory channels	N/A
Memory Technology	MCDRAM & DDR4
Threads per core	4
Vector units & width (per core)	AVX-512
On-chip MCDRAM	Up to 16GB at launch, over 5x STREAM vs. DDR4

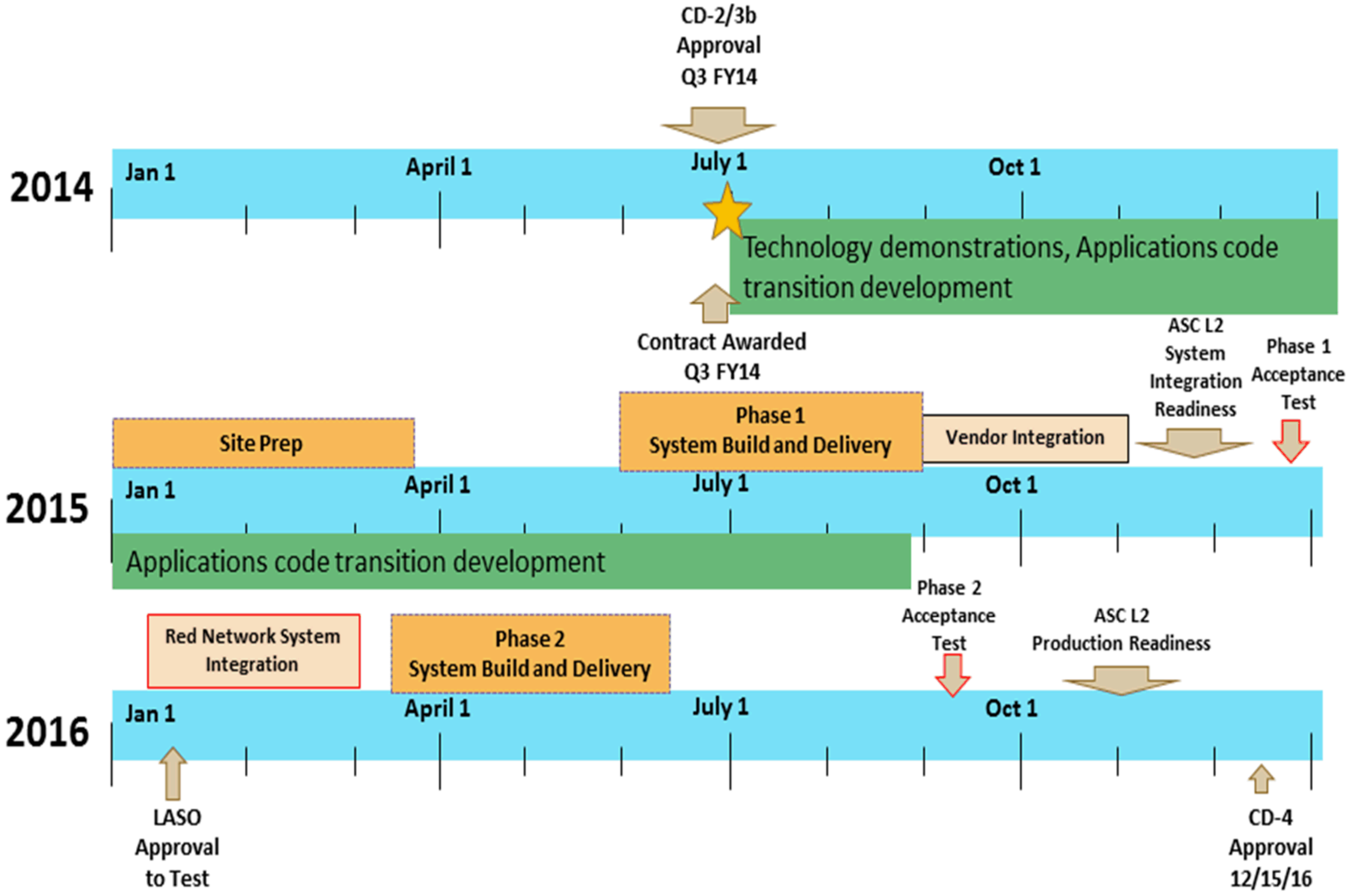
Trinity Capabilities

- Each partition will accommodate 1 to 2 large mission problems (2 to 4 total)
- 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

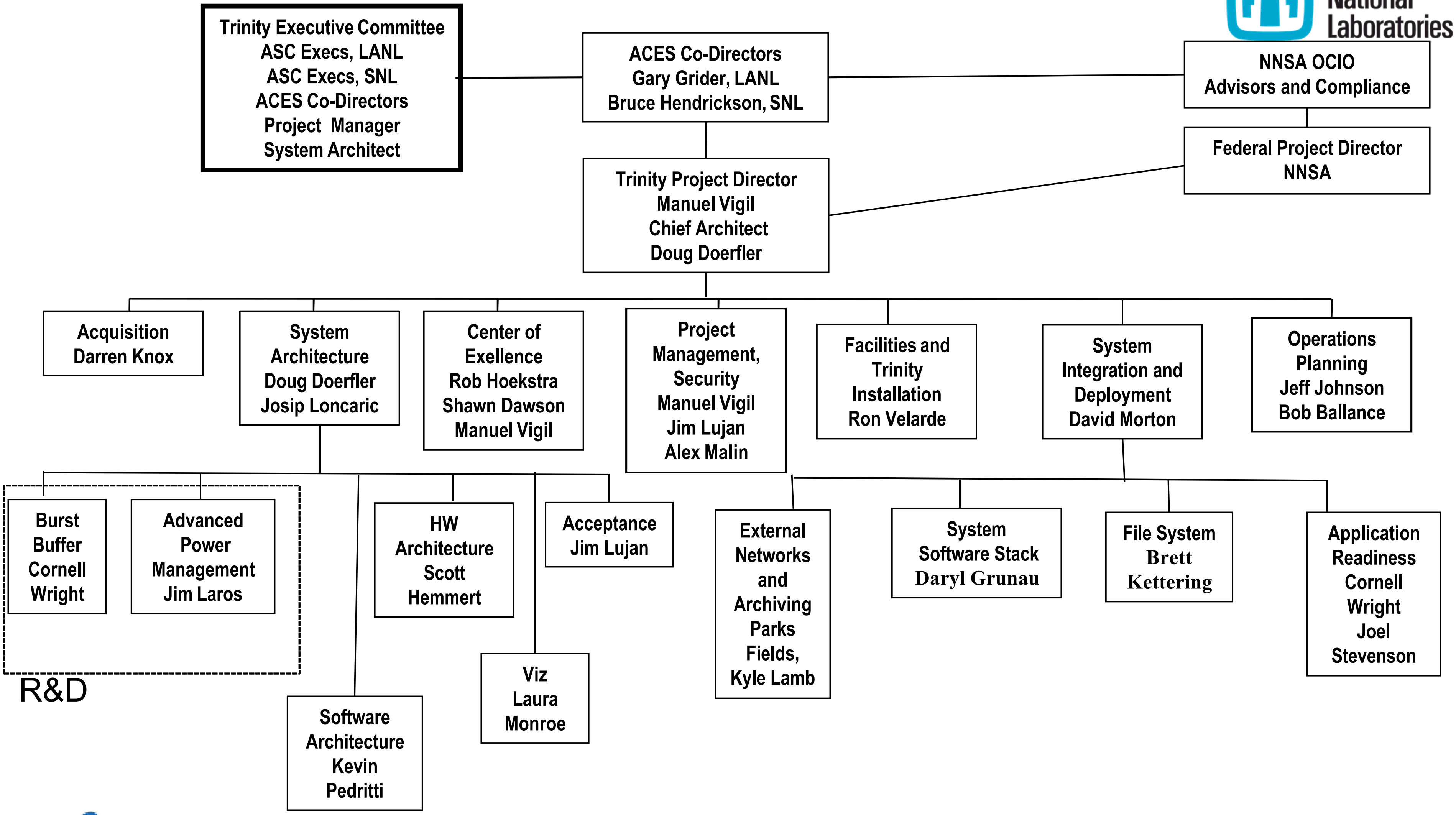
The Trinity Center of Excellence & Application Transition Challenges

- 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”

Trinity Platform Schedule Highlights 2014-2016


CD-2/3b Approval Q3 FY14
2014
Jan 1
April 1
July 1
Oct 1
Technology demonstrations, Applications code transition development
Contract Awarded Q3 FY14
ASC L2 System Integration Readiness
Phase 1 Acceptance Test
Site Prep
Phase 1 System Build and Delivery
Vendor Integration
2015
Jan 1
April 1
July 1
Oct 1
Applications code transition development
Phase 2 Acceptance Test
ASC L2 Production Readiness
Red Network System Integration
Phase 2 System Build and Delivery
2016
Jan 1
April 1
July 1
Oct 1
LASO Approval to Test
CD-4 Approval 12/15/16

Trinity Project Team



Trinity Executive Committee
ASC Execs, LANL
ASC Execs, SNL
ACES Co-Directors
Project Manager
System Architect

ACES Co-Directors
Gary Grider, LANL
Bruce Hendrickson, SNL

NNSA OCIO
Advisors and Compliance

Federal Project Director
NNSA

Trinity Project Director
Manuel Vigil
Chief Architect
Doug Doerfler

Acquisition
Darren Knox

System Architecture
Doug Doerfler
Josip Loncaric

Center of Exellence
Rob Hoekstra
Shawn Dawson
Manuel Vigil

Project Management, Security
Manuel Vigil
Jim Lujan
Alex Malin

Facilities and Trinity Installation
Ron Velarde

System Integration and Deployment
David Morton

Operations Planning
Jeff Johnson
Bob Ballance

Burst Buffer
Cornell Wright

Advanced Power Management
Jim Laros

R&D

HW Architecture
Scott Hemmert

Acceptance
Jim Lujan

Viz
Laura Monroe

Software Architecture
Kevin Pedritti

External Networks and Archiving
Parks Fields,
Kyle Lamb

System Software Stack
Daryl Grunau

File System
Brett Kettering

Application Readiness
Cornell Wright
Joel Stevenson

Questions?

Backup Slides

The ACES partnership since 2008

- SNL/LANL MOU signed March 2008 to integrate and leverage capabilities
- Commitment to the shared development and use of HPC to meet NW mission needs
- Major efforts are executed by project teams chartered by and accountable to the ACES co-directors
  - Cielo delivered ca. 2011
  - Trinity delivery in 2015 is now our dominant focus
- Both Laboratories are fully committed to delivering a successful platform as it’s essential to the Laboratories

Trinity is Sized for High Fidelity Workloads

Table 1. ASC 3D Simulation Memory Requirements over Time

Timeframe/ Fidelity	Physics fidelity	Geometric fidelity	Numerical fidelity	Restart file memory size
1980	Not applicable Could not run the 3D simulations due to memory size limitations	Not applicable	Not applicable	Not applicable
1990	Low	Low Not all items included	Low Cell size was large	3x10^{9} or 3 Gbytes
2000	Low	Medium Not all items included	Medium	3x10^{11} or 300 Gbytes
2010	More physics added	Medium	Medium	8x10^{13} or 80 Tbytes
2015 Projected	Increased physics	Higher	Higher	7.5x10^{14} or 750 Tbytes

Point Design:
A current LANL 3D problem runs on ¼ of Cielo today using about 80 TBytes

Projected capacity for a higher fidelity problem is about 750 TBytes in the Trinity timeframe

Trinity is sized to support 2 to 4 jobs of this class
-> 2 to 4 PBytes
-> ¼ to ½ of the system

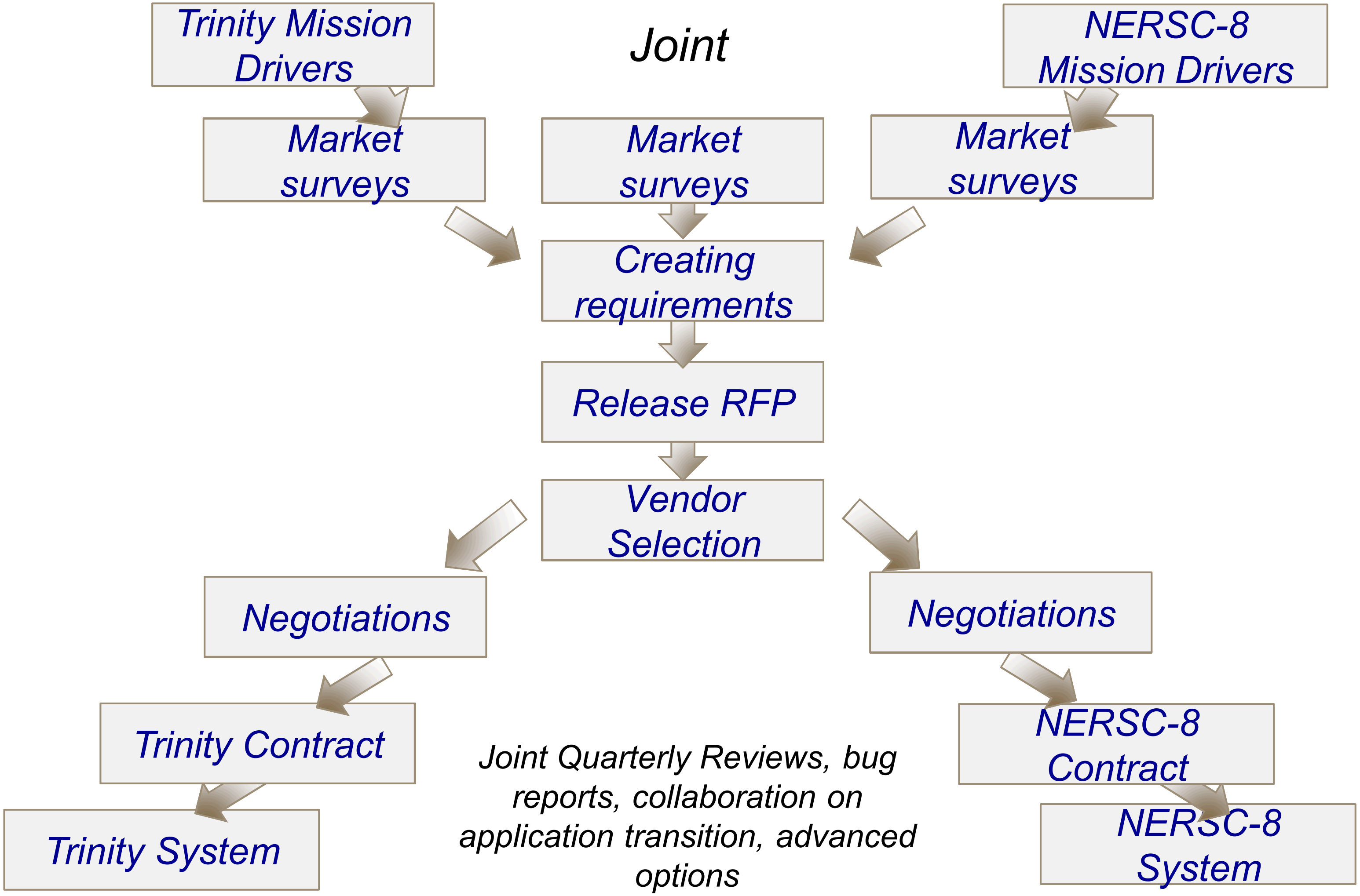
Trinity Facility, Power & Cooling

- Trinity will be located in the Nicholas C. Metropolis center (SCC) at Los Alamos National Lab
- Facility power is one of the primary constraints in the design of Trinity
  - 12 MW water cooling + 2-3 MW (maybe 4 MW) air cooling available
    - Inclusive of storage and any other externally attached equipment
  - 300 lbs per square foot floor loading
  - 10,000 to 12,000 square feet of floor space
- At least 80% of the platform will be water cooled
  - Direct (direct to chip or cold plate) is preferred
  - Indirect (e.g. radiator) method is acceptable
  - Tower water (directly from cooling tower) at up to 32° C is preferred
  - Chilled water at 8.5° C is available but less desirable due to additional $
  - Under floor air at 12.5° C is available to supplement the water cooling method
- Concerns
  - Idle power efficiency
  - Rapid ramp up / ramp down load on power grid over 2 MW

Why is NNSA/ASC and SC/ASCR collaborating?

- The April 2011 MOU between SC & NNSA for coordinating Exascale activities was the impetus for ASC and ASCR to work together on the proposed Exascale Computing Initiative (ECI).
- While ECI is yet to be realized, ASC & ASCR program directors made strategic decisions to co-fund and collaborate on:
  - Technology R&D Investments: FastForward and DesignForward
  - System Acquisitions: Trinity/NERSC-8 and CORAL
  - Great leveraging opportunities to share precious resources (budget & technical expertise) to achieve each program's mission goals, while working out some cultural/bureaucratic differences.
- The Trinity/NERSC-8 collaboration will proceed with joint RFP and selection, separate system awards, attendance at other system's project reviews and collective problem solving.

Trinity & NERSC-8 are two separate projects resulting in two distinct contracts and systems



Trinity Mission Drivers

Joint

NERSC-8 Mission Drivers

Market surveys

Market surveys

Market surveys

Creating requirements

Release RFP

Vendor Selection

Negotiations

Negotiations

Trinity Contract

NERSC-8 Contract

Trinity System

NERSC-8 System

Joint Quarterly Reviews, bug reports, collaboration on application transition, advanced options

Trinity BAFO Drivers and Constraints

1. Satisfy the mission need for increased capability
   - Increase in geometric and physics fidelities, decrease time to solution
2. Deliver a significant system in FY15
3. Keep to ATS-1 budget profile
   - Minimize carry over to FY16
4. Deliver ASC a significant increase over Sequoia
5. Foster competition for next generation architectures in the HPC industry
6. Keep ASC application developers engaged in making the transition to next generation architectures
   - Application developers and users stated at least 50% of compute partition should be KNL in support of code transition

Trinity BAFO Strategy

- A single system that contains both Intel Haswell and Intel Knights Landing (KNL) processors
- Haswell partition delivered in FY15 satisfies the mission needs (well suited to existing codes), the FY15 delivery requirement, and fits the FY15 budget profile.
- KNL partition delivered in FY16 results in a system significantly larger than Sequoia, provides Intel with incentive to continue competing in HPC, provides the application developers with an attractive next generation target, and fits the FY16 budget profile.
- Haswell partition is low risk because it is existing technology
- KNL is higher risk due to new technology, but mitigation is to go to a Haswell only system