SAND2012-9804P

Introducing Trinity, the ASC Program's Next-Generation Advanced Technology System

Doug Doerfler,
Sandia National Laboratories
Scalable Computer Architectures Department
Trinity Architecture co-Lead
ACES deputy PM for Sandia

SAND 2012-xxxxP
Unlimited Release

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04- 94AL85000.

Topics Covered

- ASC Platform Strategy
- Partnerships
- High-Level Design Requirements
- Notional Technologies
- Status
- Schedule

ASC platform acquisition strategy
Objectives and Approach

- Objectives
  - Acquire right-sized platforms to meet the mission needs for ASC codes to support stockpile stewardship
  - Invest in prioritized R&D technologies to explore and exploit new and incoming technologies
- Approach
  - Previous classes of systems: Advanced Architectures, Capability, Capacity
  - Reduced to two classes of systems: Commodity Technology Systems and Advanced Technology Systems
  - New Advanced Technology system start every 2 years

ASC platform acquisition strategy

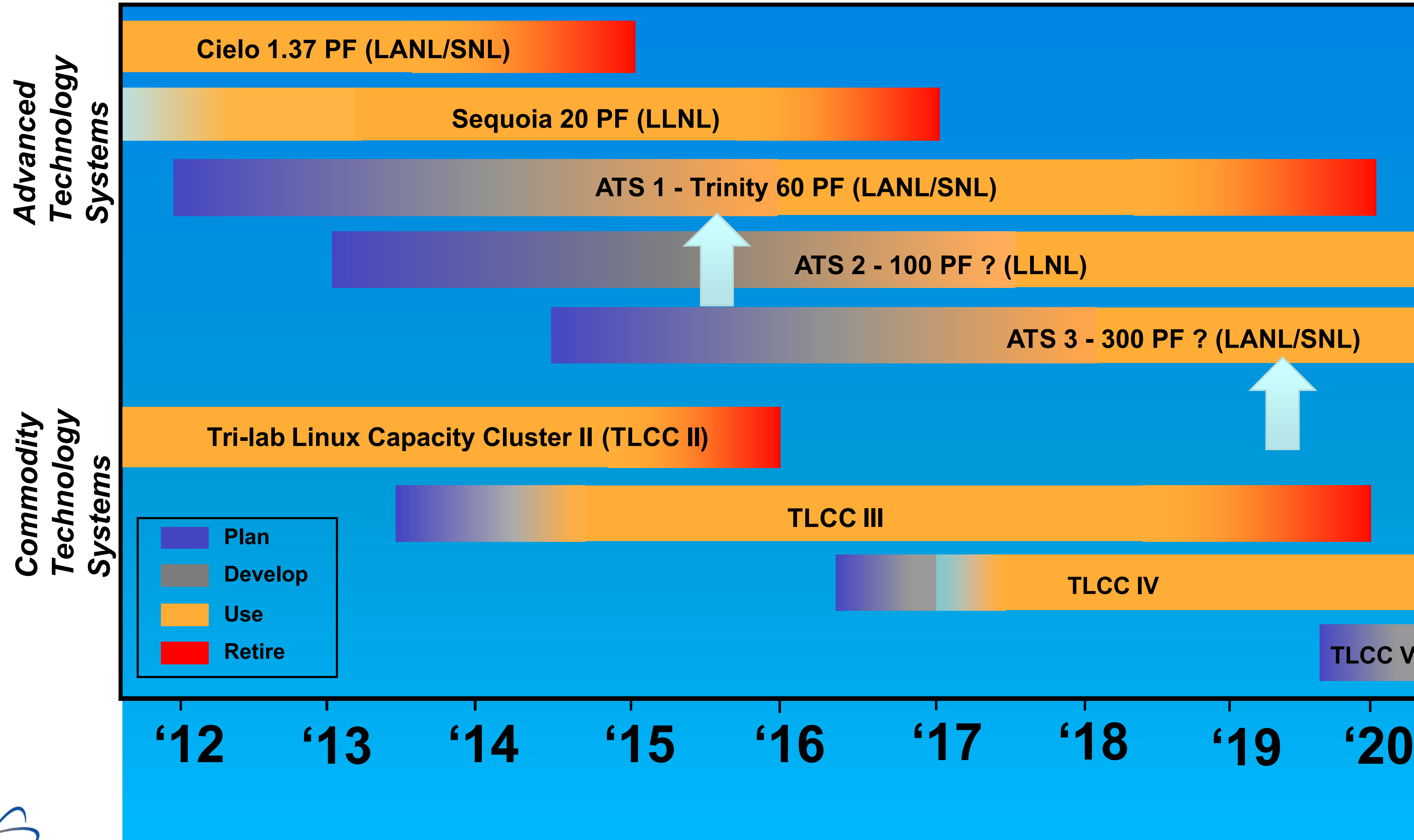
Advanced Technology Systems

- 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
- Important that all Labs use all of the systems to help inform decision-making process
- Includes Non-Recurring Engineering (NRE) funding to enable delivery of leading-edge platforms
- Trinity is the first of the Advanced Technology Systems
- Trinity will be deployed by the ACES partners (Sandia and Los Alamos)

ASC platform acquisition strategy
Commodity Technology Systems

- Provide a stable and reliable computing environment
- Least disruption to applications
- Leverage market advances in technology
- Common tri-lab procurements
- Continue standardization efforts to reduce costs and enhance cross-site utilization
- Need to push petascale application development

NNSA ASC Baseline Platform Acquisition Plan



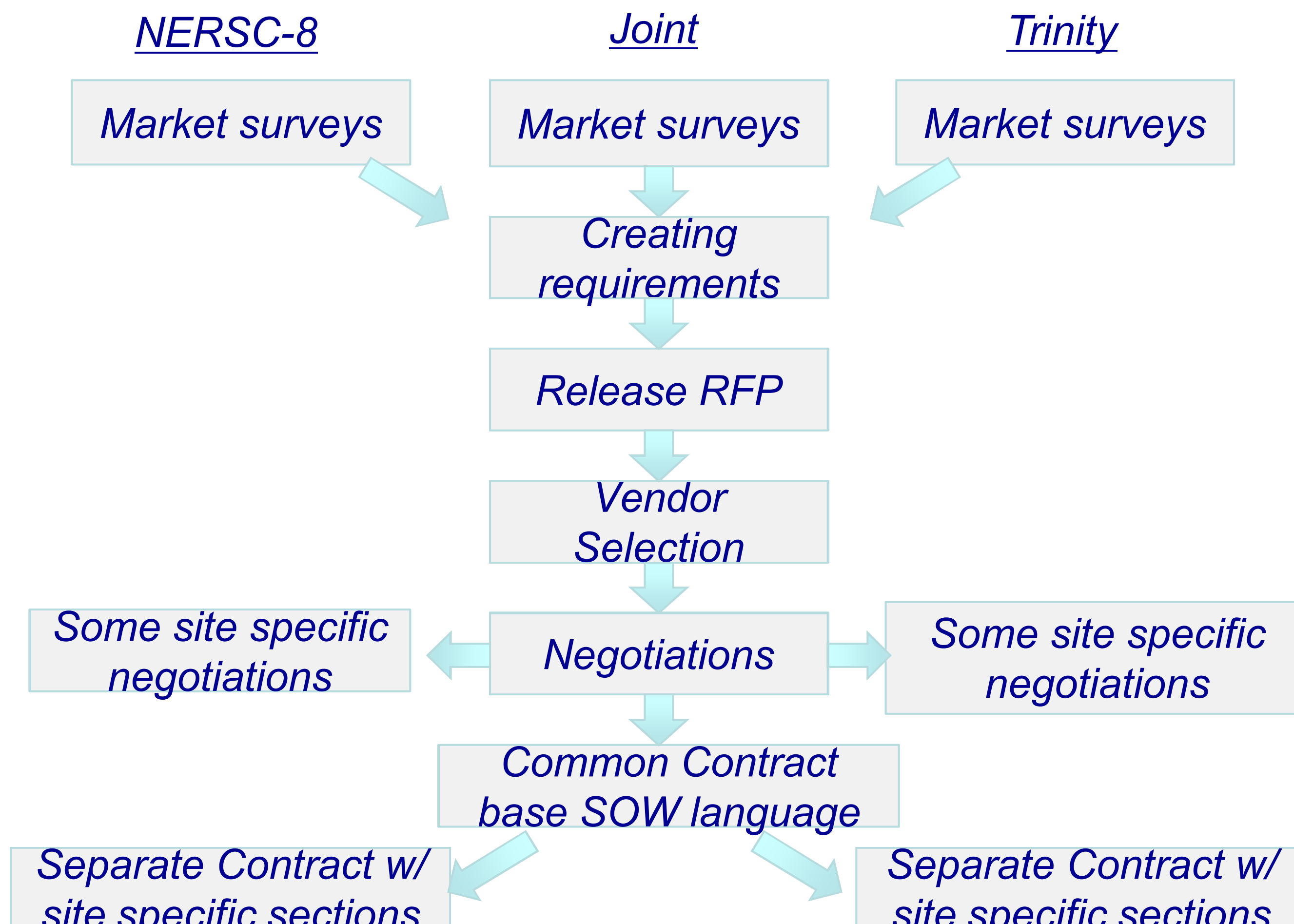
Trinity will Enable an Increase in Predictive Capability for the ASC Program

- An increase in predictive capability requires increases in the fidelity of both geometric and physics models.
- Trinity needs to demonstrate a significant capability improvement over current platforms (>> Cielo, > Sequoia) in key areas of physics
  - Improvement is a function of performance (total time to solution), increased geometries and increased physics capabilities
- Increased capabilities drive improvements in computational resources
  - Higher fidelity models -> increases in aggregate memory capacity
  - While sustaining time to solution -> increases in computational capabilities, memory bandwidth & scaling characteristics
- Advanced resilience techniques will play a major role in improving application efficiency (time to solution) (a key strategy is an I/O burst buffer)
- Active power management techniques within the platform may be required to meet the facility and total cost of ownership constraints

ACES (NNSA) is partnering with NERSC (Office of Science) on the Procurement

- Strengthen alliance between SC/NNSA on road to exascale
- Show vendors a more united path on road to exascale
- Shared technical expertise between labs
- *Should* gain cost benefit
- Saves vendors money/time responding to a single RFP, single set of technical requirements
- Outside perspective reduces risk -- avoids tunnel vision by one lab
- More leverage with vendors by sharing information between labs
- Benefits in production, shared bug reports, quarterly meetings
- Less likely to be a one-off system with multiple sites participating

Proposed joint activities between NERSC-8 and Trinity teams



Negotiations could start together, but different file systems, configurations, $, payment schedules, integration specs will require separate negotiations

Common base SOW language, but site specific sections needed

High-level Design Requirements

- Trinity is to be in production operation early to mid FY2016
- Mission requirements are primarily driving memory capacity
  - 2 to 4 PB of aggregate main memory
- And increasing fidelities and complexity while maintaining the time to solution provided by today's platforms, Cielo and Sequoia
  - Capability Improvement > 8x Cielo on key ASC applications
- NW codes should be able to port in a "reasonable" timeframe
  - Need to support legacy MPI-everywhere model out of the box …
  - but it is expected to take time to become computationally efficient.
- Nominal programming model will be MPI+X
  - MPI for coarse grain parallelism, X for fine grain parallelism
  - X programming model needs to be agnostic and portable to a variety of highly threaded architectures, e.g. Multicore, GPGPU, MIC, APU, etc.
    - OpenMP, OpenACC, CUDA, OpenCL, …
    - There may be requirements for multiple choices of X within the same code
    - Programming models other than MPI+X are also of value, e.g. PGAS

Notional Technology & Design

- 2 PB to 4 PB of aggregate memory
  - Likely to be 20,000 to 35,000 nodes
  - 2x to 3x Cielo in node level parallelism
  - Increased thread level parallelism
  - Increasing vector lengths (SIMD parallelism, AVX, etc)
- Node processor architecture options
  - Many-core, e.g. Intel Phi (MIC) (>50 cores)
  - GPGPU, e.g. Nvidia Fermi or AMD APU
  - IBM Power X
  - x86-64?
- Node memory architecture options
  - 192 GB to 512 GB main memory per node
    - Transition from DDR3 to DDR4 is low risk, however, DDR4 pricing is a concern
  - Fast (near) & Not-so-Fast (far) memory is possible
    - Horizontal memory hierarchy exposed by some accelerators is an optimization that may be necessary to meet application performance goals
    - A single address space is desirable, but data movement may be necessary
    - Fast memory may be application managed

Notional Technology & Design (cont'd)

- External parallel file system
  - Vendor provided and supported
  - Disk-based scratch file system
    - Hold files for a few days
    - Backed up by site provided Archive system
  - 25x to 30x of the aggregate memory size (100 PB to 120 PB)
  - Accessible to external resources via File Transfer Agents (FTAs)
    - Tri-lab WAN, external clusters and site archive
- Visualization & Data Analysis
  - Support CEI's Ensight, LLNL's VisIt and KitWare's ParaView
    - Dedicated, on-platform support for geometry extraction and SW rendering
    - Traditionally this is ~ 5% of total compute partition
  - Burst Buffer will allow for in-transit data analysis
  - Mechanisms to support in-situ analysis will also be included

Focus Areas for Enhanced Capabilities & Advanced Development

- On-system burst buffers to provide high-BW intermediate storage
  - Solid-state-based burst-buffer resource, primarily for defensive checkpointing
    - Alternating A/B checkpoints that only exist during a job's duration but can be used for application restart within a single job session (e.g. restart after a failed node)
    - Staging area for persistent I/O to/from an external disk-based parallel file system
  - External persistent I/O can be dribbled in/out for prefetch & write out
  - 3x aggregate memory size (A+B checkpoints + staging space)
  - Also a goal to support other use cases, e.g. in-transit data analysis
- Power management
  - Platform and facility level monitoring and management tools for application and operational use cases
- Application Readiness
  - System and processor vendor support porting "key" applications after initial acceptance of the system and production availability

Facility, Power & Cooling

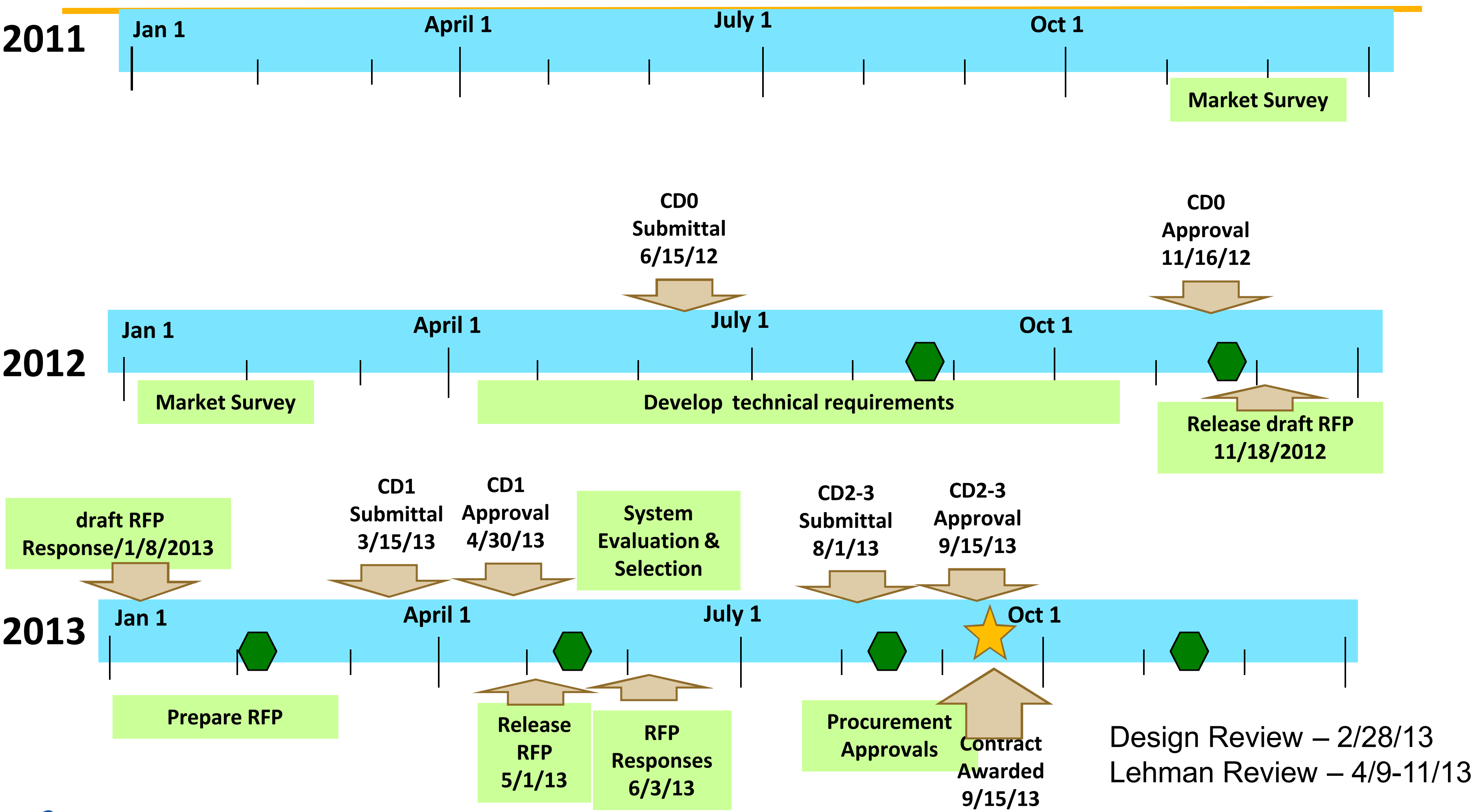
- Trinity will be located in the Nicholas C. Metropolis center (SCC) at Los Alamos National Lab
- Facility power is one of the primary constraints in the design of Trinity
  - 12 MW water cooling + 2-3 MW (maybe 4 MW) air cooling available
    - Inclusive of storage and any other externally attached 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^{o} C is preferred
  - Chilled water at 8.5^{o} C is available but less desirable due to additional $
  - Under floor air at 12.5^{o} 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

Trinity: Current status and activities

- ASC Platforms discussions with HQ & tri-lab
- Mission Need and CD0 activities
- Trinity CD0 is at NNSA ASC HQ
- Schedule issues and discussions
- Ongoing market survey meetings with vendors to identify technology paths for the Trinity system timeframe
- Technical and Project teams being developed to complete the acquisition and deployment of the Trinity system.
- Ongoing technical requirements interaction between ACES and NERSC

Trinity Platform Schedule Highlights 2011-2013

Draft for discussion



Trinity OCIO Project interactions

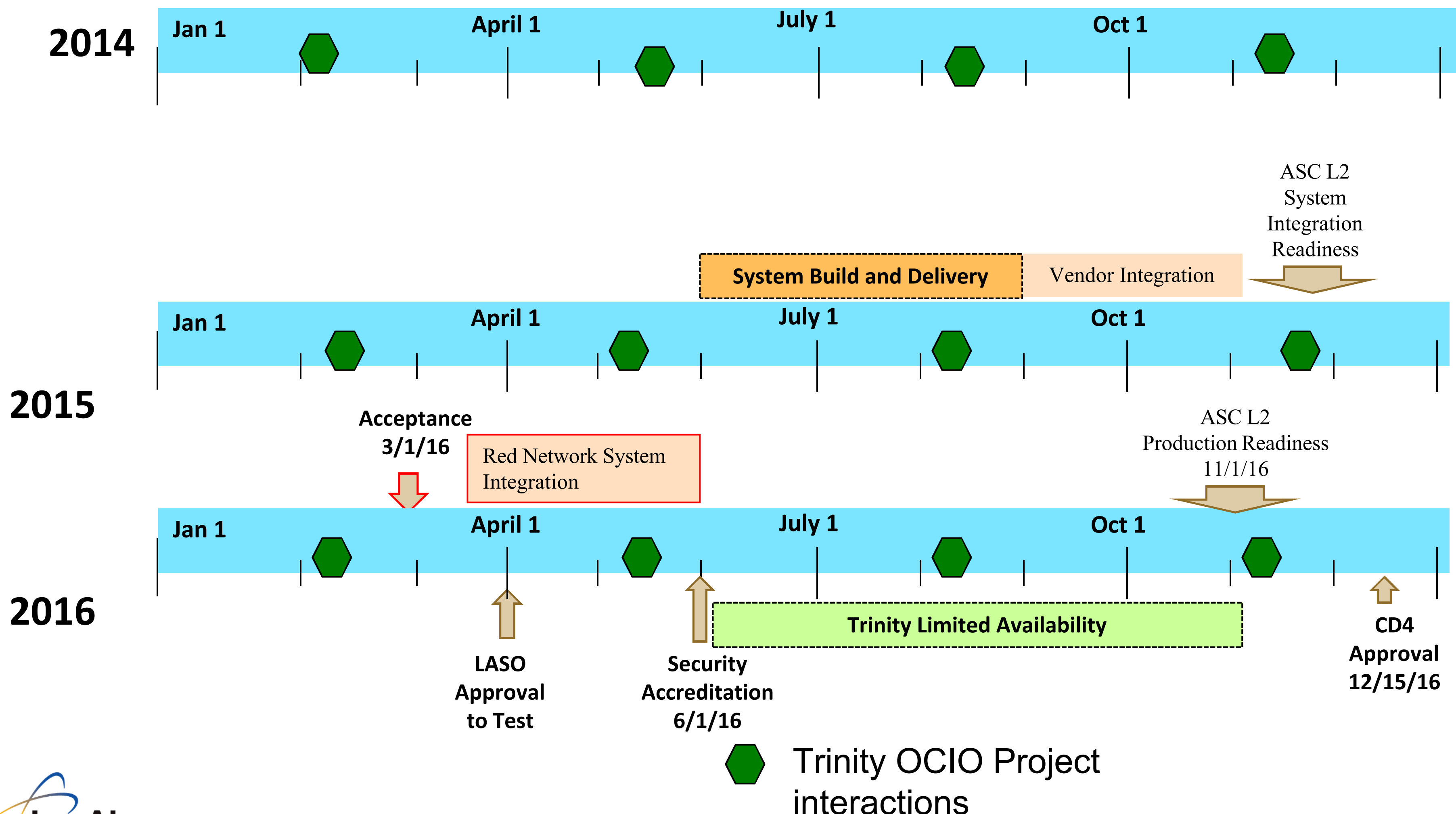
Los Alamos NATIONAL LABORATORY EST.1943

Operated by the Los Alamos National Security, LLC for the DOE/NNSA

Sandia National Laboratories

Trinity Platform Schedule Highlights 2014-2016

Draft for discussion



Questions

dwdoerf@sandia.gov