LA-UR-15-20221
Approved for public release; distribution is unlimited.

Title: The Trinity System

Author(s): Archer, Billy Joe
Vigil, Benny Manuel

Intended for: Proceeding of NECDC 2014

Issued: 2015-01-13

# The Trinity System

Billy J. Archer and Manuel Vigil
*Los Alamos National Laboratory*

**LA-UR-15-**

***This paper describes the Trinity system, the first ASC Advanced Technology System (ATS-1). We describe the Trinity procurement timeline, the ASC computing strategy, the Trinity specific mission needs, and the Trinity system specifications.***

## Trinity Procurement Timeline

Trinity is the Advanced Simulation and Computing (ASC) major computing system that is being delivered in fiscal year 2015. The system is being procured by New Mexico Alliance for Computing at Extreme Scale (ACES), a joint Los Alamos National Laboratory (LANL) and Sandia National Laboratories (SNL) partnership, and will be installed at LANL. The early phase of the procurement was a joint effort with National Energy Research Scientific Computing (NERSC).

Procurement of a major system is a complex and time consuming process. The mission need was started in September 2011 with the first request for information from the vendors in February 2012. The Critical Decision (CD) process started in December 2012 with the approval of CD-0, the decision that the mission needs justified procuring the system. After a large number of reviews CD-1/3a was approved in July 2013 giving permission to put out a Request For Proposals (RFP) and to select a vendor. The joint RFP by ACES and NERSC was released in August 2013. Due to a problem with meeting the desired delivery schedule, a request for Best and Final Offers (BAFO) was issued in March 2014 for the Trinity system. CD-2/3b was approved on July 3, 2014 and the Trinity contract was awarded to Cray Inc., on July 9, 2014.

## ASC National Computing Strategy

The ASC national computing strategy defines two types of systems [1]. The Commodity Technology Systems (CTS) are robust, cost-effective systems that are designed to meet the day-to-day simulation needs of the Stockpile Stewardship Program (SSP).

The Advanced Technology Systems (ATS) are first-of-a-kind systems that identify and foster technical capabilities / features that are beneficial to ASC applications. These systems have a dual purpose, to meet unique mission needs of the SSP, and to help prepare the ASC Program for future system designs. These are leadership-class systems, among the largest in the world. When procuring an ATS there is a tension between acquiring the right-sized platform to meet the mission needs and pursuing promising new technologies. ATS procurements include Non-Recurring Engineering (NRE) funding to enable delivery of new technologies for leading-edge platforms.

The ASC notational computing platform procurement timeline is shown in Figure 1. The strategy includes deliberate efforts to transition the application codes to each ATS platform. **Trinity is the first ATS procured by ASC.**

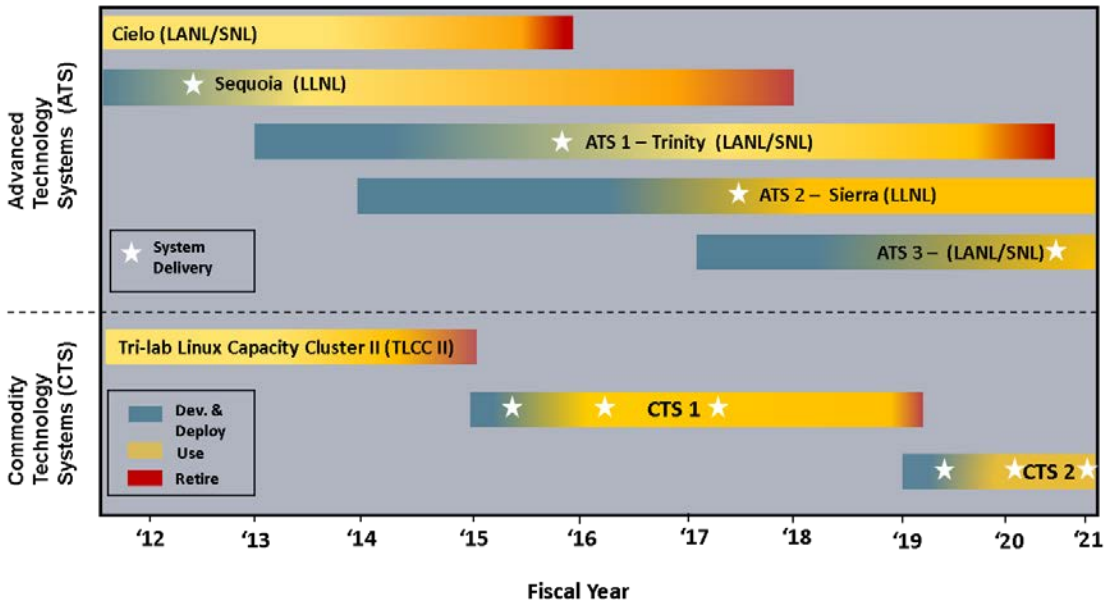


**Figure 1 The notational ASC computing platform timeline, Trinity is ATS-1.**

# Trinity Mission Needs

Trinity is designed to support the largest, most demanding Directed Stockpile Work (DSW) applications that support the SSP. ATS platforms are used by applications from all three nuclear weapons laboratories, and the mission need was developed with tri-lab input. The mission need concentrates on increases in geometric and physics fidelities in 3D while satisfying analysts' time-to- solution expectations. The 3D weapon applications are mainly constrained by available memory. The main driver is the desire to run multiple jobs each using about 750 TB of memory. This means that Trinity needs a minimum of 2 PB of aggregate main memory. **Trinity was the first DOE system specified by memory, not by floating operations per second (FLOPS).**

# The Trinity System

A firm fixed price contract was awarded to Cray, Inc. to provide Trinity. Trinity will be deployed at LANL by ACES and Cray staff.

The platform includes the compute nodes, the Aries interconnect, the Lustre parallel file system, burst buffers, two applications regression test systems, one system development test system, and

on-site system and application analysts, as well as maintenance for the life-time of the system.

NRE covers improved burst buffer software, advanced power management, and the Trinity Center of Excellence (COE). The COE directly supports modifying select applications for Trinity, and is an essential element in making effective use of Trinity.

Trinity is a single system that contains both Intel Haswell and Knights Landing (KNL) processors, shown in Figure 2. It is based on the Cray XC architecture. The Aries interconnect with the Dragonfly network topology is a mature and well understood technology. The Haswell partition, delivered in FY15, is well suited to existing codes and provides the immediate ability to partially satisfy the SSP mission needs while the application codes are modified for the KNL partition. The 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. The mix of processor types results in a platform that meets both of the ATS requirements, support of the SSP while advancing the deployed technology.

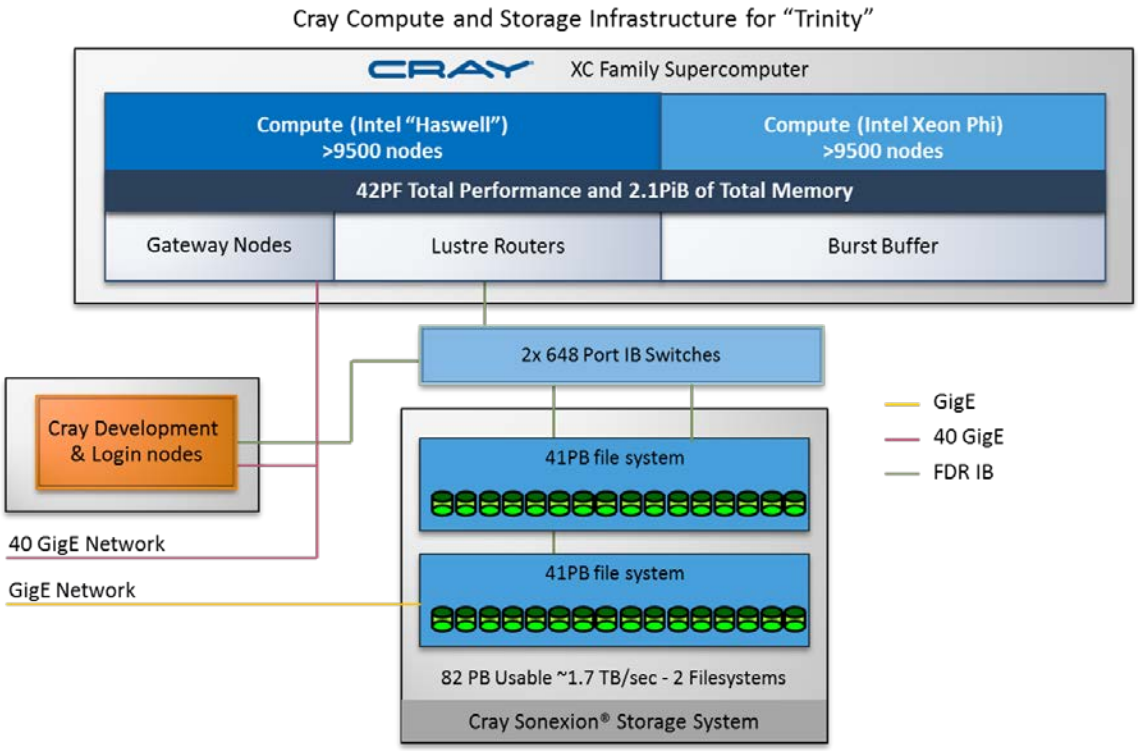


**Figure 2 Trinity Architecture.**

Each partition of Trinity has about 1 PB of memory, enough to accommodate one or two large mission simulations, for two to four total. An overview of the Trinity architecture is given in Table 1. The Trinity platform will provide an improvement in fidelity, physics and performance, 8X to 12X relative to Cielo and 2X relative to Sequoia.

Relative to Cielo there will be a >30X increase in peak FLOPS, but at the cost of a greater than 6X increase in cores per node and a 20X increase in threads per node. The parallel complexity is similar to that of Sequoia. An overview of the nodes is provided in Table 2.

**Table 1 Overview of the Trinity architecture.**

| 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 | | |

**Table 2 Overview of the Trinity nodes.**

| | 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 |

## Advanced Architectural Features

Trinity introduces several significant new architectural features. The KNL are self-hosted many-core processors. That is, the KNL are the processors, they are not accelerators, and there are 60+ cores on each chip, a 6X increase over current multi-core Xeon chips. Each KNL core has lower performance than a Xeon core. Performance is recovered by a greater than 10X increase in thread parallelism. The KNL also doubles the width of vector instructions with the dual AVX-512 SIMD units.

The KNL also introduces hierarchical memory with both high-bandwidth, low capacity, MCDRAM memory and normal DDR memory with high capacity, but low-bandwidth. Compared to DDR4 memory the MCDRAM has about 5X higher bandwidth, but only 1/5 the capacity. Determining the best way to use the hierarchical memory is a research effort by the application code teams.

The burst buffer storage system uses non-volatile memory (NVRAM) to provide a fast storage system between the compute nodes and the parallel file system. The usage model for the burst buffers is being developed. Some possibilities are staging of input/output (I/O) to disk, temporary checkpoint files, and in-situ visualization. The checkpoint use case should "just work", but other use cases will require changes to the application codes.

Continuing the theme of reducing power requirements are modifications to the system software to provide advanced power management. This is the beginning of an effort to actively manage power usage of the advanced systems as they grow in size.

The many-core architecture with hierarchal memory and burst buffers poses significant challenges to application's ability to effectively use the KNL partition. The payoff is significantly more performance at the same power cost as the Haswell partition. The Center of Excellence (COE) is being established to ensure that the KNL partition is effectively used upon initial deployment. The COE will provide support from both Cray and Intel to work with select applications from each laboratory to facilitate the code transition to Trinity. There will nominally be one application from each laboratory.

# Summary

The schedule for installation of Trinity is shown in Figure 3. FY2014 activities concentrated on NRE, including standing up the COE. The Haswell partition will be installed starting in mid-2015. After an integration and open science period, Trinity will be moved to the secure in early 2016. Shortly thereafter the KNL partition will be installed. The entire system will be in production in early FY17.

Trinity will require applications to transition to an MPI+X programming environment and requires increases in thread and vector level parallelism to be exposed. This transition is the beginning of a longer effort required by the next-generation systems that will follow Trinity, such as ATS-2 and ATS-3. Trinity introduces Active Power Management, Burst Buffer storage acceleration, and the concept of a Center of Excellence to ASC production platforms

There will be operational challenges associated with Trinity. A major one is that Trinity is the first liquid cooled system at Los Alamos since the 1990's, and it is much larger than those earlier systems. Operating a system with a mix of two very different processors will bring system management challenges. The sheer volume of data Trinity will be capable of generating will overwhelm existing archiving systems, and calls for a rethinking of user work patterns.

Trinity is the first of the ASC ATS platforms. It meets or exceeds the design goals that were laid out in the CD process, and will meet the mission needs of the SSP.

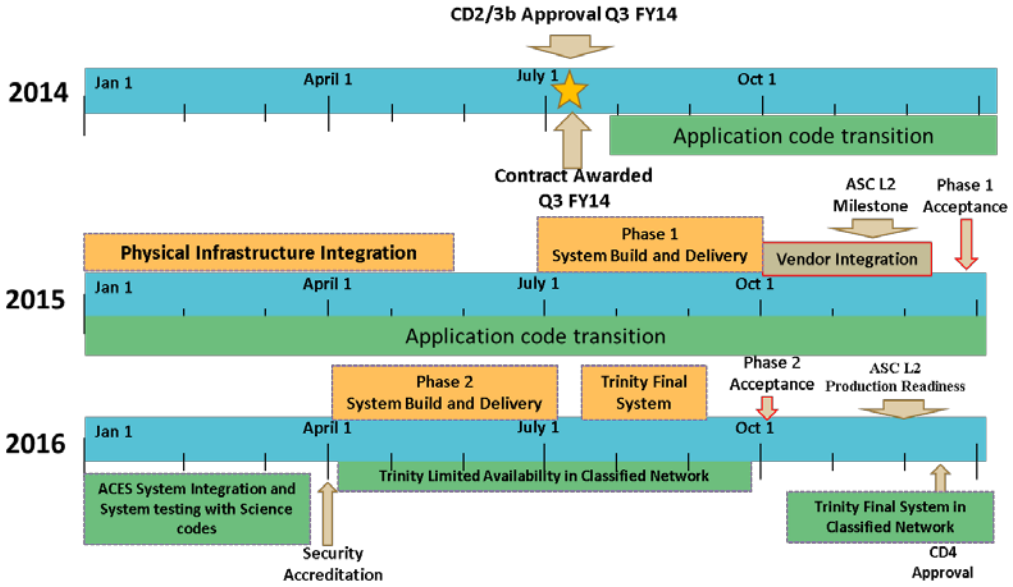


**Figure 3 Trinity platform schedule highlights, 2014-2016.**

# References


1. Ang, J.A., Henning, P.J., Hoang, T.T., and Neely, R. 2013. Advanced Simulation and Computing, Computing Strategy. SAND 2013-3951P