The Cielo Petascale Capability Supercomputer

Doug Doerfler[2], Manuel Vigil[1], Sudip Dosanjh[2] and John Morrison[1]

[1] *Los Alamos National Laboratory, Los Alamos, NM*
[2] *Sandia National Laboratories, Albuquerque, NM*

ABSTRACT: *Los Alamos and Sandia National Laboratories have formed a new high performance computing center, the Alliance for Computing at the Extreme Scale (ACES). The two labs will jointly architect, develop, procure and operate capability systems for DOE's Advanced Simulation and Computing Program. This paper discusses a petascale production capability system, Cielo, that is currently being deployed by ACES in partnership with Cray and Panasas.*

KEYWORDS: Cielo, Petascale

1. Introduction

Los Alamos National Laboratory and Sandia National Laboratories have collaborated to create a New Mexico center for high performance computing, the Alliance for Computing at the Extreme Scale (ACES). ACES is funded by the U.S. Department of Energy's Advanced Simulation and Computing (ASC) program and was formed to enable the solution of critical national security problems through the development and deployment of high performance computing technologies. The first ACES project is developing and deploying a production petascale supercomputer, Cielo

Cielo is replacing the Purple supercomputer at Lawrence Livermore National Laboratory as the next ASC capability platform. Many targeted national security problems are extremely large and will require most of the nodes on Cielo for a single simulation. Consequently, hardware and software scalability are critical concerns. Reliability is also a key because the mean time between interrupts for an application executing on Cielo decreases with the number of nodes it uses. Another design consideration was effectively supporting existing ASC computer codes with little or no modification. That is, applications that ran on Purple must execute efficiently on Cielo. The overall goal is to provide an order of magnitude increase in capability over Purple. After a competitive procurement process, Cray was awarded the contract.

ACES is also planning and architecting a 2015 transpetaflops system, Trinity. Other industrial partnerships are being pursued both within the context of reaching exascale and providing production capability computing.

2. Cielo Architecture

The ACES design team was focused on a few key attributes that drove the selection of Cray's XE6 architecture for Cielo: reliability, power, hardware scaling, system software scaling, and application scaling. In this section we provide an overview of the Cielo platform and conclude with application performance demonstrated during the acceptance of the initial deployment.

Cielo Hardware Architecture

Cielo is one of the first instantiations of Cray's XE6 supercomputer architecture. The Cielo configuration and options are described in the following sections.

Computational Partition
Cielo is the latest ASC Tri-Lab capability computing system and is one of the first instantiations of the Cray XE6 architecture [1]. Cielo is being deployed in two phases. Phase 1 is composed of 6,704 compute nodes, each configured with dual Advanced Micro Devices 2.4 GHz, eight-core (model 6136) Magny-Cours processor for a total of 107,264 compute cores and a peak performance of 1.03 PFLOPS. In May 2011, the Phase 2 upgrade grows the system to 8,894 compute nodes, for a total of 142,304 cores and 1.37 PFLOPS peak performance.

Each compute node has two processors, with each processor consisting of two four-core dies for a total of sixteen cores per node, arranged as four separate NUMA regions. HyperTransport™ links connect the dies. The compute nodes are configured with 2 GB of memory per core for a total 32 GB per node.

For the XE6 architecture the Gemini high-speed interconnect replaces the SeaStar interconnect used in the XT. Gemini was designed to better support multicore processors and scales to millions of cores in a single system. The Cielo system is configured as an 18x8x24 3D-torus Gemini network. A pictorial representation of Cielo identifying the principal components is shown in Figure 1.

Visualization and Data Analysis Partition
Four cabinets of the compute section will have double the memory of the rest of the compute partition. This partition is dedicated to support visualization and data analysis applications. It will be configured with 4 GB of memory per core, as opposed to the 2 GB per core for the rest of the compute partition, for a total of 64 GB per node. The four large memory cabinets form a 4x2x24 sub-mesh in the Cielo topology.

Parallel File System and PaScalBB Integration
Cielo is integrated into the LANL Parallel Scalable Backbone Global Parallel File System (PaScalBB). Cielo has been configured to provide greater than 200 GB/s of sustained TCP/IP bandwidth to the PaScalBB. The PaScalBB has been expanded to include an additional 10 PB of user available storage capacity and an additional 160 GB/s of sustained file system performance using Panasas high performance parallel storage [3].

In Cielo's final configuration, 228 I/O nodes, each with dual-port 10 GigE connections, will be connected to the PaScalBB. Each I/O node will provide more than 1.2 GB/s of sustained network bandwidth into the PaScalBB. The parallel file system consists of Cray's Data Virtualization Service (DVS) executing on the compute nodes, forwarding file system calls to Panasas PanFS clients running on the 228 I/O nodes, which in turn transfer data into the PaScallBB where Panasas storage servers reside.

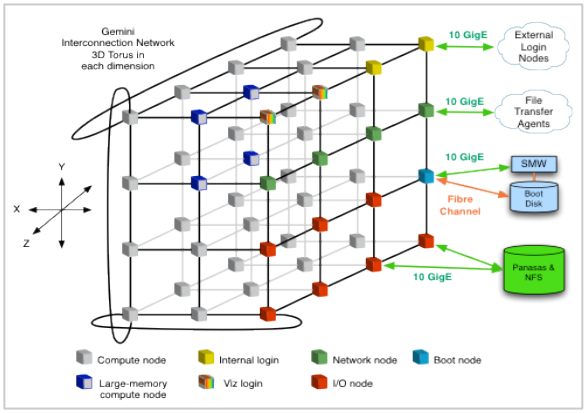


Figure 1: Cielo Partitioning and Configuration

Software Architecture
Cielo utilizes the Cray Linux Environment™ (CLE) [2]. CLE is configured to support MOAB/Torque batch scheduling. In addition, support is provided for PGI, Cray and Intel compiler environments [4, 2, 5]. The visualization partition supports the ParaView, Ensight and VisIt visualization tools [6, 7, 8]. TotalView is provided for debugging [9].

Cielo Phase 1 Application Performance
As a part of acceptance for the Phase 1 deployment, a suite of ASC Tri-lab applications were chosen for scale testing: CTH and Charon from SNL, SAGE and xNobel from LANL, and AMG2006 and UMT2006 from LLNL. The purpose of the scaling study was to demonstrate the increased capability of the Cielo platform relative to its programmatic predecessor, the ASC Purple platform. The requirement was to demonstrate at least a six times improvement (6x) in capability, defined to be the product of increased problem size and runtime performance speedup relative to Purple. For example, if the problem size executed on Cielo is eight times larger then the one executed on Purple (i.e. 8x weak scaling) and the runtime metric of interest demonstrates a speedup of 1.25 relative to Purple, then the capability improvement becomes 8x * 1.25 = 10x. The details of the applications, the acceptance test method and results can be found in Doerfler's CUG 2011 paper [4], but results are summarized in Figure 2. Cielo performed very well for all six of the applications and the demonstrated overall improvement factor of 9.6x at 64K PEs and 10.5x if you include those applications scaling to greater then 64K PEs.

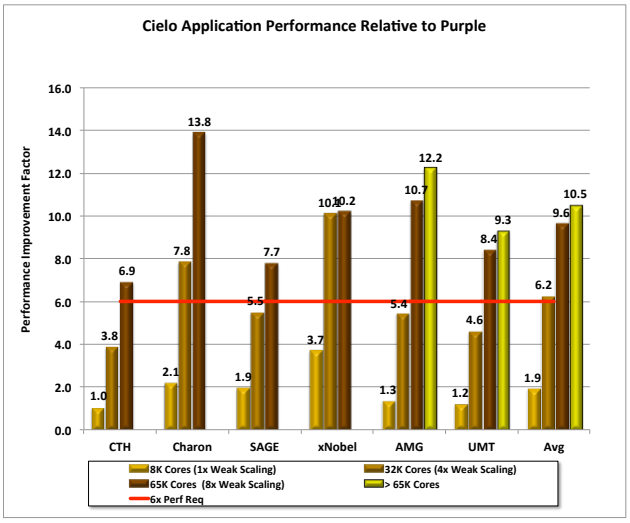


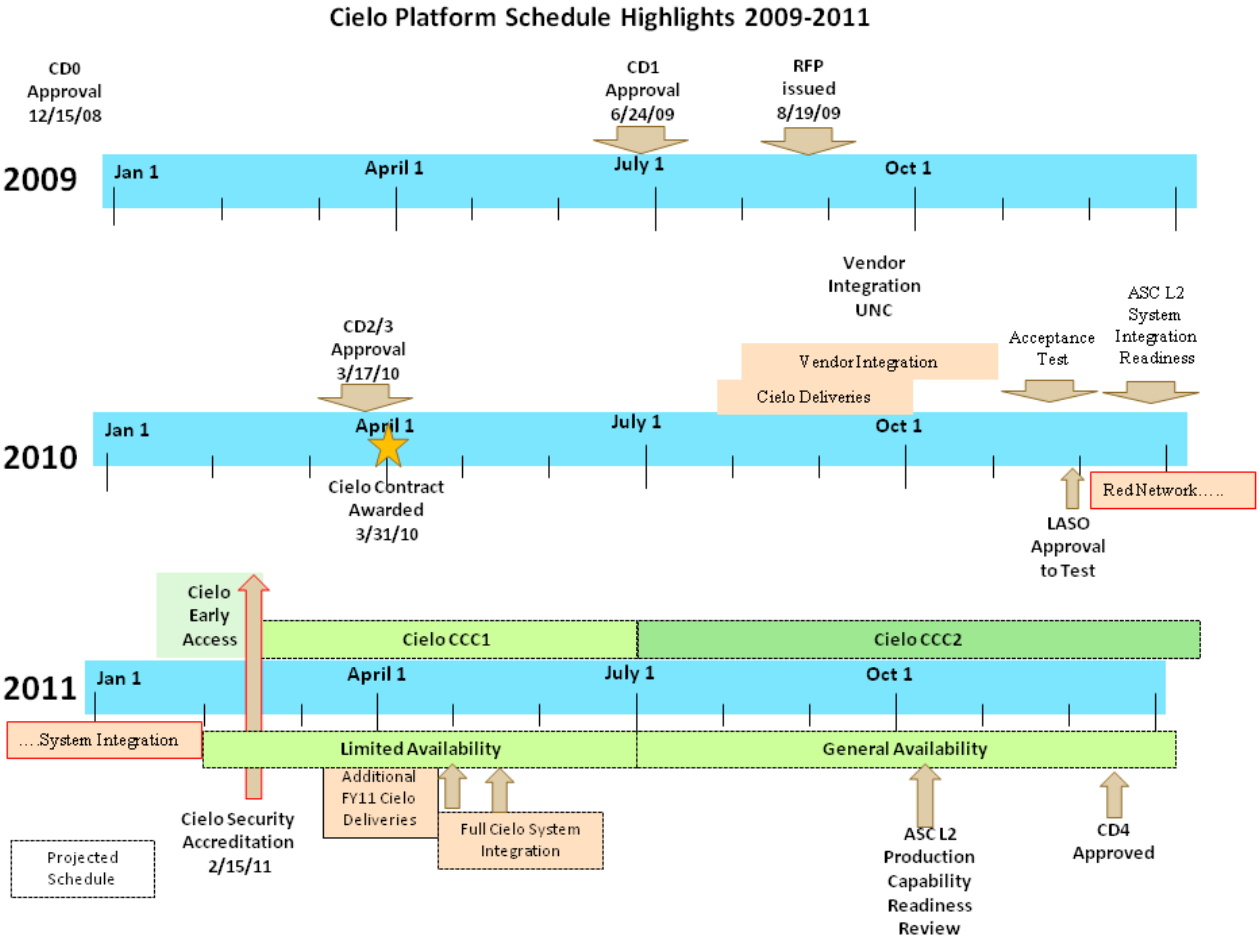
Figure 2: Summary of Cielo Phase 1 application acceptance testing

4. Schedule and System Integration

Following the contract award to Cray in March 2010 the ACES/Cray partnership has achieved significant progress, under a very aggressive schedule (see chart below), in preparation for tri-lab simulations work in 2011. This is the first time a multi-lab partnership has been involved in deploying an Advanced Simulation and Computing (ASC) capability machine. This partnership has worked well in helping meet all contractual, project and program milestones.

The Cielo platform is targeted to be the signature production classified capability platform resource for running integrated weapons simulations for the tri-lab (LLNL, LANL, and SNL) in the 2011-2015 time frame. The Cielo platform provides a replacement computing resource for existing simulation codes as well as provides a larger resource for ever-increasing capability computing requirements. Cielo will be sited at Los Alamos but will operate under a national user facility paradigm by the ASC program and will be available to the tri-lab community.

In March, 2010 Cray, Inc. was selected to deliver the Cielo platform for ACES. The deliveries for Cielo were set up to provide a ~1.03 Petaflop/s system in FY10 and additional deliveries in FY11 for a total system peak capability of ~1.37 Petaflop/s. Cielo will also be the first large Cray system to use the Panasas file system for storage. Visualization nodes were acquired with twice the memory of other compute nodes. The initial FY10 system



consisted of 72 racks, which are already installed at LANL's Strategic Computing Center (SCC) (see picture below).

The successful deployment and stabilization of a large computing system requires a strong partnership between the vendor providing the system and the ACES integration team. Since the contract award in March, 2010 there has been a lot of effort to build, deliver, test, and accept the Cielo system. ACES is responsible for the overall integration and operations of the platform after final system acceptance.

The activities associated with the acquisition consist of contract award, vendor design and system build, vendor project risk plans, pre-ship and post-ship testing, system delivery and installation at the Los Alamos site, and acceptance testing.

Other efforts include project management, system integration, acceptance testing, facilities upgrades, site preparation, file system infrastructure deployment, on-site analyst support, stabilization testing, applications porting and testing, performance tuning and testing, and preparations for production operations of the system, including providing maintenance after the initial warranty period.

Under a very aggressive schedule the ACES/Cray system integration team has made significant progress in 2020 in getting the system ready for applications use. All the original project schedules have been met thus far. A few high level accomplishments include:

- 72 Cray cabinets build, delivered and installed at LANL
- System integration in the classified network, including scaling testing
- Panasas system delivered, installed and tested in the unclassified network
- Completion of official Acceptance Test
- Transitioning the system to LANL's classified network for integration

The early applications work has also demonstrated that most codes tested thus far have been able to port to Cielo with minimal difficulty.

Each Laboratory (LLNL, SNL, LANL) selected an early application for demonstrating the use of Cielo and to help with system stabilization. Allocations for use of Cielo are based on the NNSA ASC Capability Computing Campaigns (CCC) model. Cielo will be operated as the NNSA's National User Facility for capability computing. After security accreditation in early February 2011 the Cielo CCC1 period began. This period is being used to further run tri-lab simulations as well as to help tri-lab application code teams develop or port applications to Cielo and to allow for application scaling work on those applications in preparation for submitting CCC2 proposals.

A system upgrade in early May 2011 is scheduled to add 24 additional cabinets to the Cielo system that will add approximately 33% more computing capability, from 1.03 PF/s to 1.37 PF/s.

Additional smaller systems and applications testbeds were also acquired as part of the acquisition. These systems have been used for both systems testing and applications development and have proved to be extremely useful for the Cielo system integration.

Cielo was acquired under the NNSA Office of the Chief Information Office (OCIO) Project Execution Model (PEM) for IT investments. Under this model Cielo is required to meet several Critical Decision (CD) milestones for acquisition, integration and production operations. Cielo must also meet two ASC programmatic milestones (one has been completed) before it is formally approved for production capability.

4. Conclusion

ACES is partnering with Cray and Panasas to deploy a production petascale capability platform, Cielo. Cielo is being used to solve critical DOE/NNSA national security problems.

Acknowledgments

The authors would like to thank Robert Meisner and Sander Lee for their encouragement and the DOE ASC program for support.

About the Authors

Douglas Doerfler is a Principle Member of Technical Staff at Sandia National Laboratories. Doug is the ACES Cielo Architect and his research interests include high-performance computer architectures and performance analysis.

Manuel Vigil is a Program-Project Director at Los Alamos National Laboratory working on project management, planning, acquisition, and integration of current and future high performance computing systems. Manuel is the Project Manager for the Cielo system as part of the ACES Collaboration.

Sudip Dosanjh is the head of extreme-scale computing at Sandia National Laboratories. He is co-director of ACES and the Science Partnership for Extreme-scale Computing. He has served on DOE's Exascale Initiative Steering Committee.

John Morrison is director of the High Performance Computing Division at Los Alamos National Laboratory. John is co-director of ACES and has served on DOE's Exascale Initiative Steering Committee.

References

[1] Cray XE6, http://www.cray.com/Products/XE/Systems/XE6.aspx/
[2] Cray Linux Environment, http://www.cray.com/Products/XE/Software.aspx/
[3] Panasas High Performance Parallel Storage, http://www.panasas.com/
[4] Portland Group Compiler Suite, http://www.pgroup.com/
[5] Intel Compiler Suite, http://software.intel.com/en-us/articles/intel-compilers/
[6] ParaView, http://www.paraview.org/
[7] Ensight, http://www.ensight.com/
[8] VisIt, https://wci.llnl.gov/codes/visit/
[9] TotalView, http://www.roguewave.com/products/totalview-family/totalview.aspx
[10] D. Doerfler, et al., "Application-Driven Acceptance of Cielo, an XE6 Petascale Capability Platform", Cray User Group (CUG) 2011, Fairbanks, Alaska, May 2011.