Alliance for Computing at Extreme Scale (ACES)

CUG 2010, Edinburgh

May 27th, 2010

Jim Ang, Douglas Doerfler, Sudip Dosanjh Scott Hemmert
Sandia National Laboratories

Ken Koch, John Morrison, Manuel Vigil
Los Alamos National Laboratory

SAND 2010-3078 C
Unlimited Release
Printed May, 2010

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.

Los Alamos NATIONAL LABORATORY EST.1943

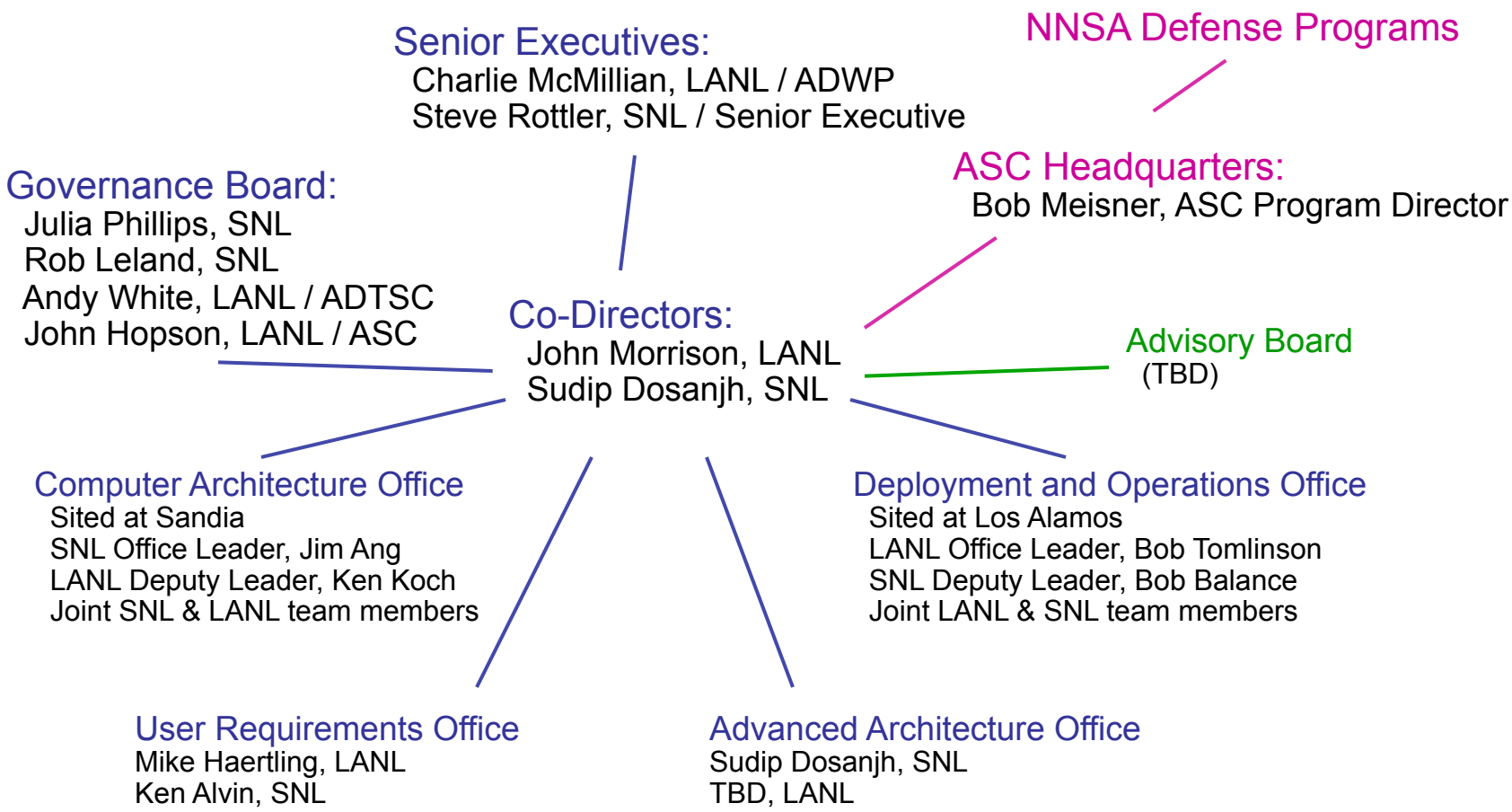
Operated by the Los Alamos National Security, LLC for the DOE/NNSA

Sandia National Laboratories

ACES: LANL & SNL Partnership

- 3/2008: LANL & SNL Memorandum of Understanding
- ACES: The NNSA New Mexico Alliance for Computing at Extreme Scale
  - Joint design, architecture, development, deployment and operation of production capability systems for NNSA
- Driven by mission needs
- Commitment to the development and use of world class computing
- Continued leadership in high performance computing
- Sharing intellectual capabilities of both laboratories

ACES Alliance Organization Chart



ACES Initiatives

- Cielo
  - First ACES project with the responsibility to acquire, integrate, deploy and operate the NNSA ASC Program's next generation *capability computing* platform
  - Support NNSA ASC Program Capability Computing Campaigns, replacing ASC Purple at LLNL
  - Cray Baker supercomputer architecture with an order of magnitude improvement over ASC Purple in application performance
- Interconnection Network Project
  - Collaboration with Cray to develop a future generation of high-speed interconnect, *Pisces*
- Future Platforms
  - Planning for future ASC *capability* and *advanced architecture* platforms

The Cielo Platform

- Cray, Inc. selected to deliver Cielo Platform for ACES
  - Cielo Platform (1.03 Peak PF in FY10)
  - Other Options (FY11):
    - Additional Delivery Option (0.33 additional PFs in FY11)
  - Total of 1.37 PF system (with room for expansion)
- NM Alliance for Computing at Extreme Scales (ACES) partnership developed Cielo RFP requirements
  - Including input received from presentations at LANL, SNL, and LLNL
- Technical Evaluation Team from all three laboratories
  - Including applications staff
- Cielo Project team is ready to execute project deliverables

cielo
NNSA New Mexico Alliance for Computing at Extreme Scale

NNSA ASC

Design Philosophy & Goals

- Petascale production capability to be deployed in Q1FY11
  - Take over the role Purple currently plays
  - Usage Model will follow the Capability Computing Campaign (CCC) process
  - Capability: Capable of running a single application across the entire machine
- Easy migration of existing integrated weapons codes
  - MPI Everywhere is the nominal programming model
  - 2GB memory per core (minimum) to support current application requirements
- Performance goal is to achieve a 6x to 8x improvement over Purple on representative CCC applications
  - Memory subsystem performance will be the major contributor to node performance
  - Interconnect performance will be major contributor to scaling performance
  - Reliability will be major contributor to CCC total time to solution
- Upgrade path to allow increased capability in out years
- Key challenges: Reliability, Power, HW and SW Scalability, Algorithmic Scaling to 80K to 100K MPI ranks

Cielo High-Level Performance Metrics

Performance Metric	RFP Specification	Cielo Phase 1	Cielo Phase 2
Application Performance	> 6x Purple	> 6x Purple	> 1.17 Phase 1
Peak FP	1.0 PF	1.03 PF	1.37 PF
Total Memory Capacity	> 200 TB	227 TB	298 TB
Memory per core	> 2 GB	2 GB*	2 GB*
Peak Memory BW	> 400 TB/s	572 TB/s	763 TB/s
Sustained Bisection BW	> 20 TB/s	12.6 TB/s	15.3 TB/s
Sustained Msg Injection Rate	> 50 GMsgs/s	53.6 GMsgs/s	71.5 GMsgs/s
Sustained Off Platform I/O BW	> 200 GB/s (160 GB/s DVS)	200 GB/s (160 GB/s DVS)	271 GB/s (160 GB/s DVS)
System Power (max)	< 8 MW	3.9 MW (est)	4.4 MW (est)
Full System Job MTBI	> 25 hours	25 hours	25 hours
System MTBI	> 200 hours	200 hours	200 hours

cielo
NNSA New Mexico Alliance for Computing at Extreme Scale

NNSA ASC

Cielo High-Level Software Architecture

Feature	LWOS	FFOS
Pedigree	Linux derivative	Linux derivative
Personality	Streamlined for good application performance; configurable job-by-job	Full featured
Target functionality	Compute	Compute & Service
Language Support	Fortran, C, C++, Python, Perl, Java, Shells	Fortran, C, C++, Python, Perl, Java, Shells
Programming Models	MPI-2 within LWOS, OpenMP, POSIX Threads	SLES support for MPI, OpenMP, POSIX Threads
High-speed Interconnect protocols	Native high-speed for MPI-2	Native high-speed for MPI-2, Sockets, NFS & TCP/IP
Supported Libraries	Cray optimized scientific libraries, libm, libgsl, FFTW, BLAS1-3, LAPACK, dynamic libs and dlopen()	Standard libraries as part of SLES distribution
Application tools	CrayPat and Apprentice2 w/support for hardware counters, memory usage, performance profilers, MPI tracing and profilers	Hardware counters, memory usage, performance profilers, MPI tracing and profilers
Application Debugger	Yes, at least 8192 MPI ranks	
Data Analysis & Geometry Extraction	Support for VisIt, ParaView and EnSight	
Other	Support for MOAB, scalable job launch	Support for MOAB, scalable job launch

cielo
NNSA New Mexico Alliance for Computing at Extreme Scale

NNSA ASC

Cielo: Cray's "Baker" Architecture

	Phase 1	Phase 2
Layout	4 Rows of 18 Cabinets	4 Rows of 24 Cabinets
High Speed Interconnect	Gemini	
Processor	AMD Model 6136 (Magny-Cours)	
# of Cabinets	72	96
# of Service Nodes	208	272
# of Compute Nodes	6,704	8,944
# of Compute Cores	107,264	143,104
Peak Memory BW	572 TB/s	763 TB/s
Compute Memory Capacity	226.6 TB	298.2 TB
Peak Compute FLOPS	1.03 PF	1.37 PF
Sustained PFS BW	160 GB/s	

Cielo User Environment

- External Login Servers – 5 total
  - Quad socket, Quad core workstations
  - 128 GB memory
  - 10 GigE connectivity to service nodes
  - Fully featured Linux OS
    - Including swap
- Platform System Software
  - Linux based OS
    - CNL (LWOS) on compute nodes
    - SLES Linux (FFOS) on service nodes (login, file system, networking, etc.)
    - No Swap
  - Cray MPI-2
    - an alternative MPI as part of Phase 2
  - MOAB scheduler

Cielo User Environment

- System Software cont'd
  - Programming models
    - **MPI Everywhere (Cray, PGI, GNU)**
    - OpenMP (Cray, PGI only)
    - POSIX Threads
    - SHMEM (Cray)
    - Fortran 2008 with coarrays
    - UPC
  - Optimized Libraries
    - Libm, libgsl, FFTW (V2 and V3), BLAS1 through BLAS3, and LAPACK libraries optimized for parallel applications
    - Trilinos & PETSc with Cray optimized sparse kernels (CASK)
    - FFTW with Cray optimized FFT (CRAFFT)
  - ALPS runtime system (as opposed to yod, mpiexec, mpirun, etc.)
    - 30 second job launch
    - Processor and memory affinity support

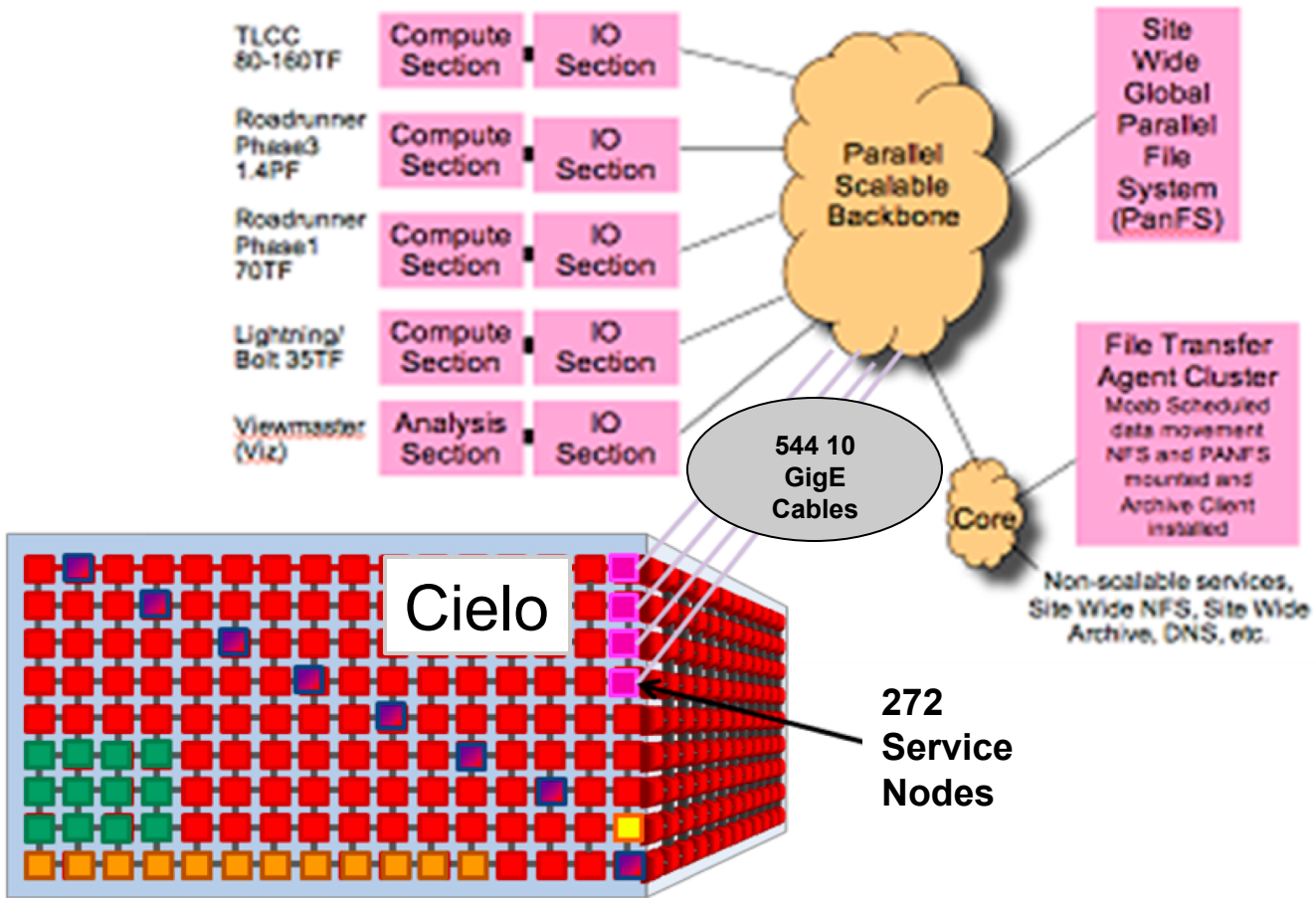
Cielo User Environment

- Tools
  - Compilers: Cray, PGI, GNU
  - CrayPat and Apprentice2
  - TotalView debugger (up to 8192 ranks)
- Visualization Partition
  - Four cabinets will be configured for visualization apps VisIt, ParaView and EnSight
  - 376 compute nodes configured with 4 GB per core (12 TB total)
  - 8 login nodes (service blades) with FFOS
  - Does NOT expand as part of Phase 2

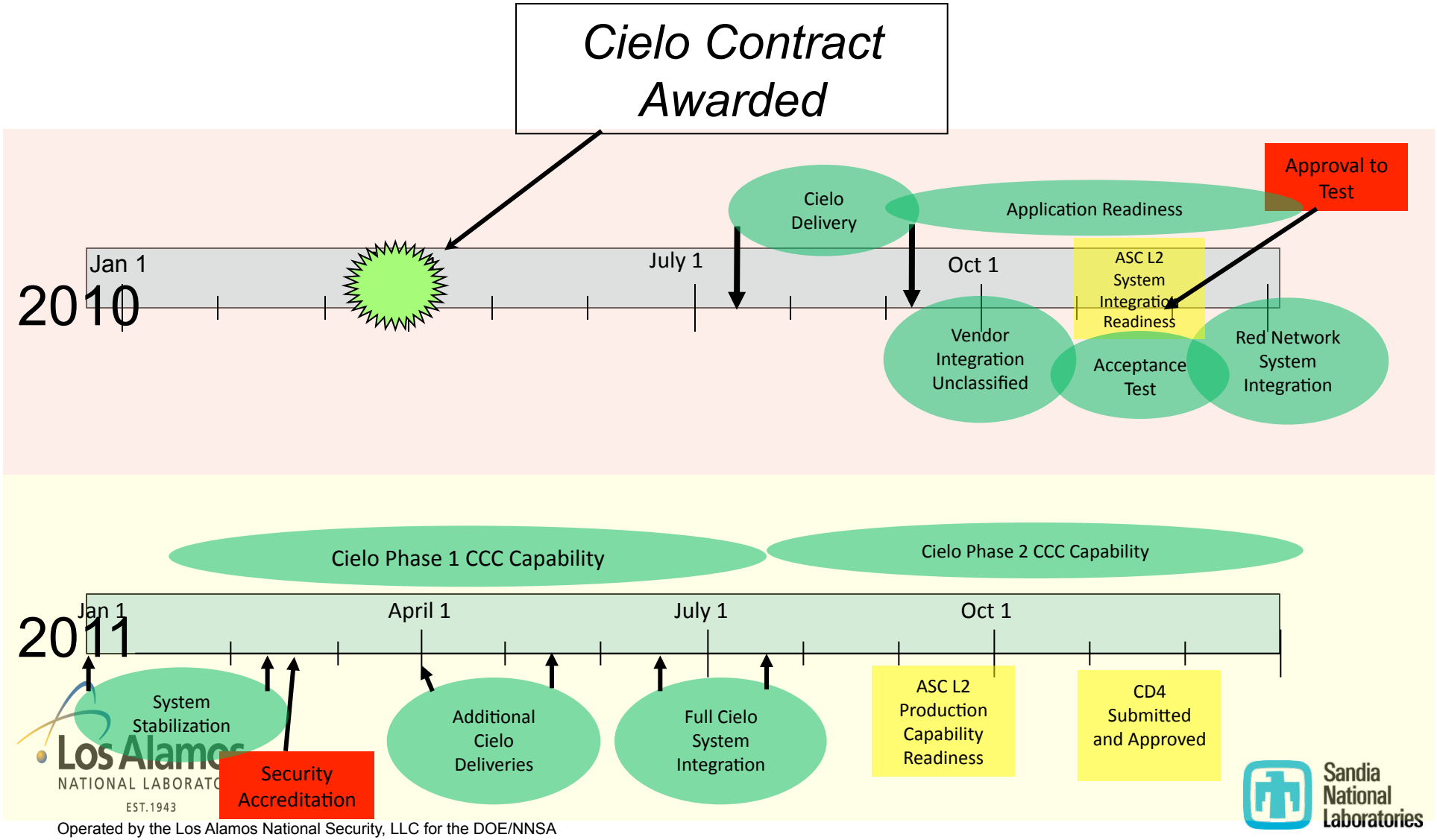
Parallel File System

- Cray DVS file system on compute nodes
  - Gemini high-speed network connectivity to service partition
  - DVS over native transport protocols
- Panasas PanFS on service partition
  - DVS to PanFS translation
  - 10 GigE connectivity to LANL Parallel Scalable Backbone
  - 200 GB/sec sustained network bandwidth
- LANL Parallel Scalable Backbone
  - 10 PB user available storage acquired as a part of the Cielo project
  - 160 GB/sec sustained bandwidth
  - 10 GigE connectivity

Cielo Integration into LANL PaScalBB



CIELO High Level Schedule



Cielo will be operated jointly by ACES: A Los Alamos & Sandia partnership

- The joint management arrangement for Cielo is unusual: management by 2 labs rather than a single laboratory.
  - We are striving to make this as invisible as possible to users, while also getting the most out of the strengths of both organizations.
  - Joint teams have been discussing how to operationally do this for several months.

Some example highlights:

- General planning for deployment:
  - We are working together to plan the deployment and operations of Cielo and it's infrastructure.
- Infrastructure:
  - We are jointly testing the new Panasas infrastructure.
  - The next-generation, 10 Gb/s encrypted WAN with dual-path (northern & southern routes) has been recently deployed.
- User support:
  - We are working to develop a jointly supported consulting (hotline) call center.
- Administration:
  - The security plan will allow local and remote system administrators.
  - We are planning for integration of trouble ticket systems for Cielo issues.

Cielo Usage and Operational Model

- We plan to run Cielo in an environment very similar to the current capability system (Purple).

- A Capability Computing Campaign (CCC) process will be used to allocate resources on the machine.
  - There may be proposed changes to the process, but any changes will be negotiated with NNSA HQ and all stakeholders.

- The Cielo Usage Model is our "contract" with users for how to best use the machine.
  - In developing Cielo's Usage Model we started with Purple's, and applied additional lessons from Red Storm.
  - Both the form & functionality described should be familiar to uses of prior capability systems.
  - The Usage Model will form baseline requirements for the 2011 Production Capability Readiness L2 Milestone.
  - Another "road show" presenting the Cielo Usage Model and gathering feedback will occur in the spring.

Interconnection Network Project

- NNSA/ASC asked ACES to consider definition of, and technical oversight for D&E project with Cray, Inc.
- Cray Interconnect Genealogy
  - Generation 1: SeaStar, current XT architecture interconnect
  - Generation 2: Gemini, to be deployed with “Baker” architecture
  - Generation 3: Aries, to be deployed with “Cascade” architecture
- Final agreement on SOW expected spring 2010
  - Will primarily focus on a potential future, post Aries, interconnection network referred to as *Pisces*, tentatively planned for the CY 2015 timeframe

Interconnect Network Tasks

- NIC Studies and Analysis
  - Analyze Gemini interconnect performance
  - Look for improvements which can be enhanced in Pisces
  - Focus on NIC, with emphasis on occupancy, latency, MPI message throughput, and independent progress.
- Router & Network Studies and Analysis
  - Analyze Aries interconnect performance
  - Look for improvements which can be enhanced in Pisces
  - Focus on network routing
- Pisces
  - Initial architectural specification in collaboration with ACES
  - Perform a comparative study between Pisces and other state of the art interconnects, such as InfiniBand, and possibly one or two others

Comments and Questions

Sudip Dosanjh: sudip@sandia.gov

John Morrison: jfm@lanl.gov

Los Alamos NATIONAL LABORATORY EST.1943

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

Sandia National Laboratories