



UNCLASSIFIED

# Cielo: Next Generation Capability Computing Platform for NNSA/ASC

Sudip Dosanjh and John Morrison

ACES Co-Directors

February 9<sup>th</sup>, 2010

LA-UR 10-00633



UNCLASSIFIED



## 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  $> 6x$  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

# High-Level Design Targets

Design Metric	Specification
Application Performance	> 6x Purple
Peak Floating Point (double precision)	> 1 PF
Total Memory Capacity	200 TB
Memory Capacity per Core	> 2 GB
Aggregate Memory BW	> 400 TB/s
Aggregate File System BW	> 160 GB/s
Total Disk Capacity	> 10 PB
System Power	< 8 MW
Full System Job MTBI	> 25 hours
System MTBI	> 200 hours

# High-Level Software Architecture

Feature	LWOS Specification	FFOS Specification
<b>Pedigree</b>	Unspecified	Unix derivative
<b>Personality</b>	Minimal noise & footprint, minimal features via configuration	Full featured
<b>Target functionality</b>	Compute & Service	Compute & Service
<b>Language Support</b>	Fortran, C, C++ & Python	Fortran, C, C++, Python, Perl, Java, Shells
<b>Programming Models</b>	MPI-2 within LWOS, OpenMP, POSIX Threads	MPI-2 within FFOS, OpenMP, POSIX Threads
<b>Binary Type</b>	Static & Dynamic	Static & Dynamic
<b>High-speed Interconnect protocols</b>	Native high-speed for MPI-2	Native high-speed for MPI-2, Sockets, NFS & TCP/IP
<b>Supported Libraries</b>	libm, libgsl, FFTW, BLAS1-3, LAPACK	libm, libgsl, FFTW, BLAS1-3, LAPACK & Mesa
<b>Application tools</b>	Hardware counters, memory usage, performance profilers, MPI tracing and profilers	Hardware counters, memory usage, performance profilers, MPI tracing and profilers
<b>Application Debugger</b>	Yes, at least 8192 MPI ranks	Yes, at least 8192 MPI ranks
<b>Data Analysis &amp; Geometry Extraction</b>		On-platform
<b>Other</b>	Support for MOAB, scalable job launch	Support for MOAB, scalable job launch

# 6x Acceptance Applications

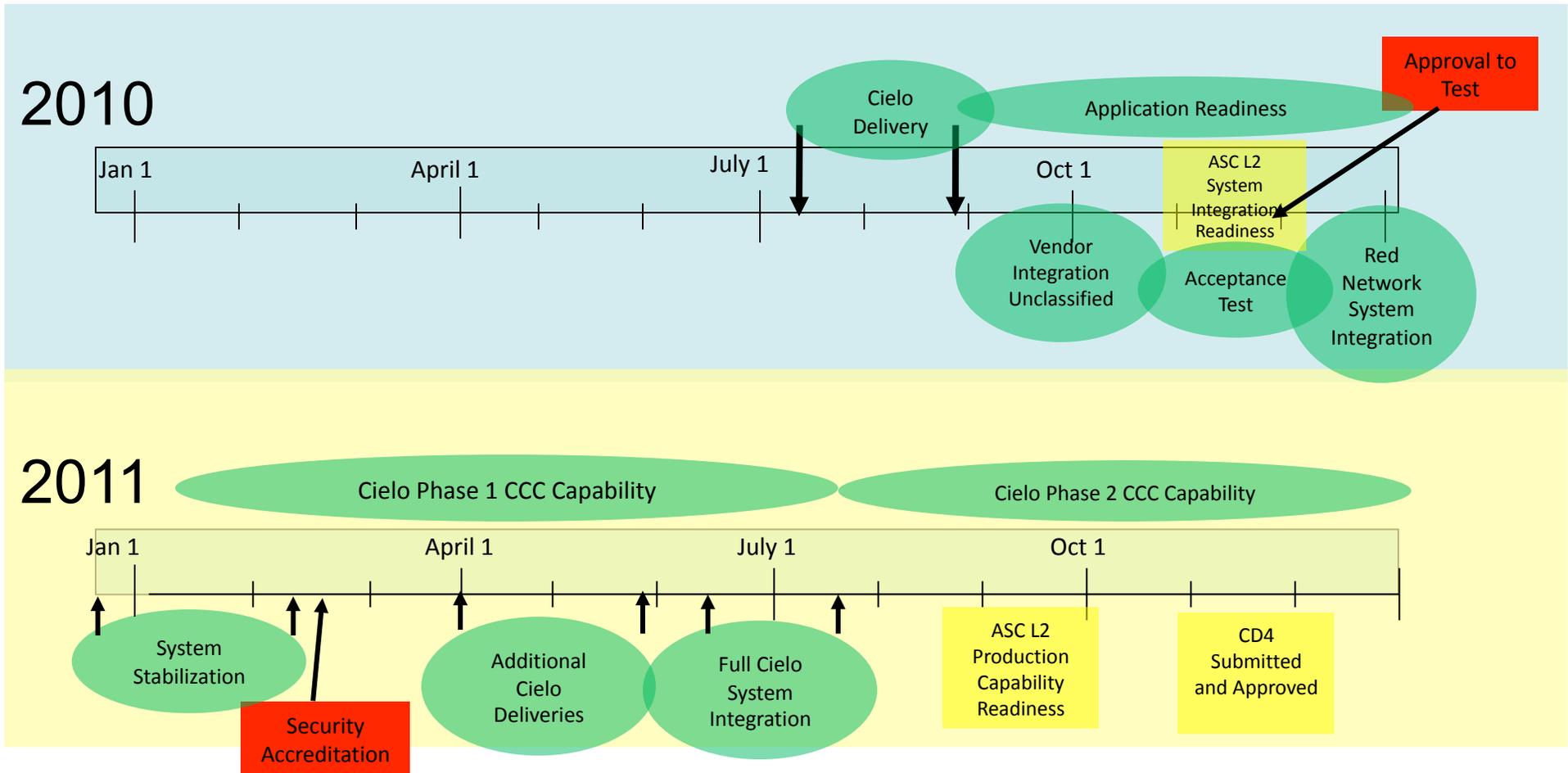
Lab	Code	Fortran	Python	C	C++	MPI	OpenMP	Description
SNL	RAMSES/ Charon			X	X	X		A transport reaction code to simulate the performance of semiconductor devices under irradiation
SNL	CTH	X		X		X		Explicit, multi-material shock hydrodynamics code
LANL	xNOBEL	X		X		X		Continuous Adaptive Mesh Refinement (CAMR) code: Hydrodynamics with adaption and high-explosive burn modeling
LANL	SAGE	X		X		X		Multi-dimensional multi-material Eulerian hydrodynamics code with adaptive mesh refinement.
LLNL	AMG2006			X		X	X	Algebraic Multi-Grid linear system solver for unstructured mesh physics packages
LLNL	UMT2006	X	X	X	X	X	X	Single physics package code. Unstructured-Mesh deterministic radiation Transport.

## Status

---

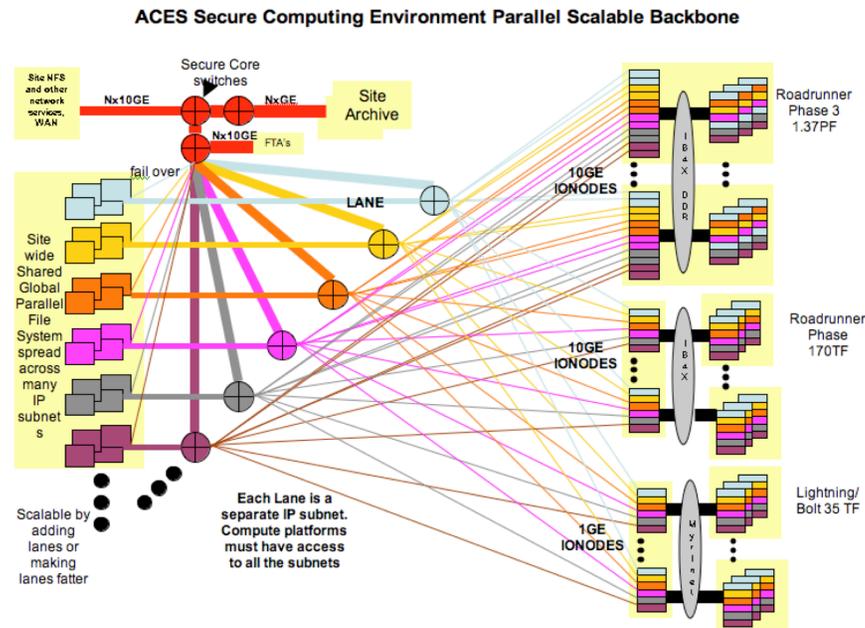
- March 2008: MOU signed between SNL and LANL creating the NNSA New Mexico Alliance for Computing at Extreme Scale (ACES)
  - First task is to manage the Cielo procurement and deployment
- Dec 2008: CD0 approved by NNSA, allowing draft Request for Proposal (RFP) to be released to selected vendors for comment
- March 2009: RFP feedback received from vendors
- June 2009: CD1 approved by NNSA
- August 2009: Final RFP to released
- Oct 2009: Proposals reviewed and selection recommendation
- Nov 2009: Contract negotiations with recommended vendor
- Contract award upon approval of CD2/3
- System delivery → projected for July-August, 2010
- Application readiness and acceptance tests
- Transition to the secure environment and usage for CCC campaigns

# Projected Cielo Delivery and Availability Schedule



# Storage Solution will Expand LANL's PaScaBB

- Contract has been awarded with Panasas for Cielo's storage subsystem
- Will be integrated into LANL's Parallel Scalable BackBone (PaScaBB)
- An additional 10 petaBytes of user available storage
- An additional 160 GB/s of sustained parallel file system bandwidth
- PanFS will be integrated into Cielo cooperatively between ACES, Panasas and the Cielo vendor.



## 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 which can be discussed are:

- 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).
- 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 users of prior capability systems.
  - Another “road show” presenting the Cielo Usage Model and gathering feedback will occur in the spring (April?).
- 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.
- After contract award a Cielo “road show” will brief users on:
  - the details of the system
  - the rest of the infrastructure and environment for the machine, and
  - generally how we plan to operate the system.

On computing remotely:

- The speed of light is not infinite; and this does cause latency delays:
- We are working with each site to validate and test work methodologies (mostly data movement and data analysis) in this new arrangement.