Workload Managers
A Flexible Approach

- Blaine Ebeling
- Manager ALPS

- May 8th, 2014

Copyright 2014 Cray Inc. – Confidential and Proprietary

Safe Harbor Statement

This presentation may contain forward-looking statements that are based on our current expectations. Forward looking statements may include statements about our financial guidance and expected operating results, our opportunities and future potential, our product development and new product introduction plans, our ability to expand and penetrate our addressable markets and other statements that are not historical facts. These statements are only predictions and actual results may materially vary from those projected. Please refer to Cray's documents filed with the SEC from time to time concerning factors that could affect the Company and these forward-looking statements.

Copyright 2014 Cray Inc. – Confidential and Proprietary

Agenda

- Application Level Placement Scheduler (ALPS) Background – Traditional Model
- Workload Managers (WLMs)
- Native Workload Managers
- Provided Functionality and Limitations
- CrayPort Site for WLM

Copyright 2014 Cray Inc. – Confidential and Proprietary

Application Level Placement Scheduler (ALPS) The Traditional Model

- **ALPS – resource placement infrastructure for Workload Managers (WLM)**
  - Manages resources, application launch and provides services specific to Cray
  - Uses aprun as the application launch command
  - Supports Cray Process Management Interface (PMI) based applications
- **Cluster Compatibility Mode (CCM) handles the launch of non-Cray PMI applications**

Copyright 2014 Cray Inc. – Confidential and Proprietary

New in ALPS

- Power-state power management options
- Suspend Resume – Job preemption
- Per application level prolog and epilog for site specific actions
- Compute node environment variables for node specific values/actions
- Optional individual files for std in/out/err, one per compute node or processor element
- Inventory Management Size Reduction
  - 26000 homogeneous nodes
  - Before 3.2 million lines, 97.5 megabytes
  - After 22 lines, 5308 bytes.

Copyright 2014 Cray Inc. – Confidential and Proprietary

Workload Manager Functionality

- **WLMs – user interface to run HPC jobs**
  - Provide batch job queuing
  - Currently only supported to work in tandem with ALPS thru the Batch Application Scheduler Interface Layer (BASIL) protocol
  - Interface w/ALPS for node reservations

- **Two widely used Workload Managers on CLE**
  - Moab/TORQUE
  - PBS Professional

Copyright 2014 Cray Inc. – Confidential and Proprietary

Requests for additional Workload Managers

- **Requests to support more Workload Managers**
- **Native WLM support**
  - May allow for earlier availability of WLM functionality

Copyright 2014 Cray Inc. – Confidential and Proprietary

Customer Requests for new Platforms

- Slurm (Simple Linux Utility for Resource Management)
- LSF
- GridEngine

Copyright 2014 Cray Inc. – Confidential and Proprietary

Why More Workload Managers

- Labs/DataCenters
- Administration Staff
- Coverage Across Systems
- Cost Efficiencies

Interest in Slurm

- Open Source WLM
- Developed at Lawrence Livermore National Lab
- Combines both Workload Management and application launch
- SchedMD provides Slurm commercial support and development

Copyright 2014 Cray Inc. – Confidential and Proprietary

“Hybrid” Slurm

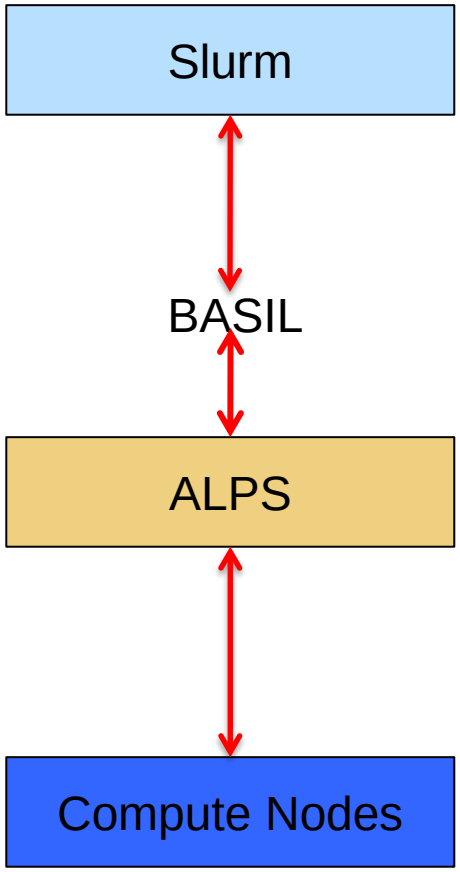
- **Cray’s first integration with Slurm**
- **Hybrid = Slurm + srun wrapper + ALPS/BASIL**
- **Initial release: Slurm 2.6.6**
- **Cluster Compatibility Mode (CCM) support added**
- Developed by 3rd party under contract by Cray
- Limited checkout and exposure
- Beta test (focus on specific use case)
- Integrated in Slurm 14.03.0-pre6
- Requires CLE-5.2.UP00 (or latest)

Copyright 2014 Cray Inc. – Confidential and Proprietary

Hybrid Slurm Architecture for Cray

Slurm

- Prioritizes queue(s) of work
- Decides when and where to start jobs
- Terminates job when appropriate
- Accounting for jobs and job steps
- No daemons on compute nodes



ALPS

- Allocates and releases resources for jobs
- Launches tasks
- Invokes node health
- NHC manages node state
- Has daemons on compute nodes
- Manages Cray network resources

Slurm is a scheduler layer above ALPS, not currently a replacement

Copyright 2014 Cray Inc. – Confidential and Proprietary

ALPS Refactoring: First Steps

- **Phase I – Create C library common interfaces of ALPS functions**
  - Network initialization
- **Phase II - Develop a native Slurm implementation**
  - Cray developed plugins to provide following services:
    - Dynamic node state change information
    - Protection key management
    - Node Health Check support
    - Network performance counter management
    - PMI port assignment management (when more than one application per compute node)
  - Plugins will be open source

Copyright 2014 Cray Inc. – Confidential and Proprietary

WLM Roadmap: Phase I

Slurm | PBS Pro | Moab | GridEngine

BASIL
ALPS

Common
Libs

2013				2014			
Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4

Prep Work (refactoring)

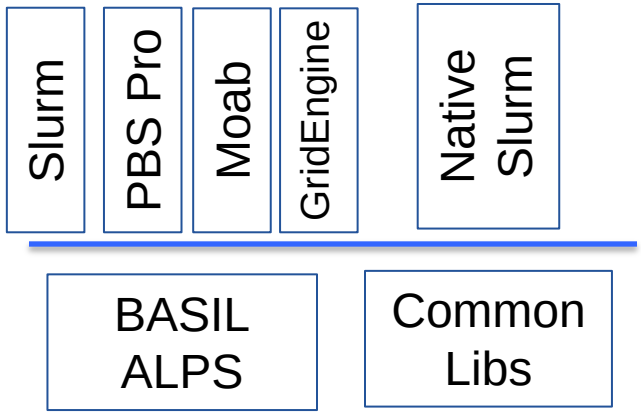
WLM common libraries split apart from ALPS: Functionality required for all WLMs (network initialization, etc…) organized into common libraries with common API

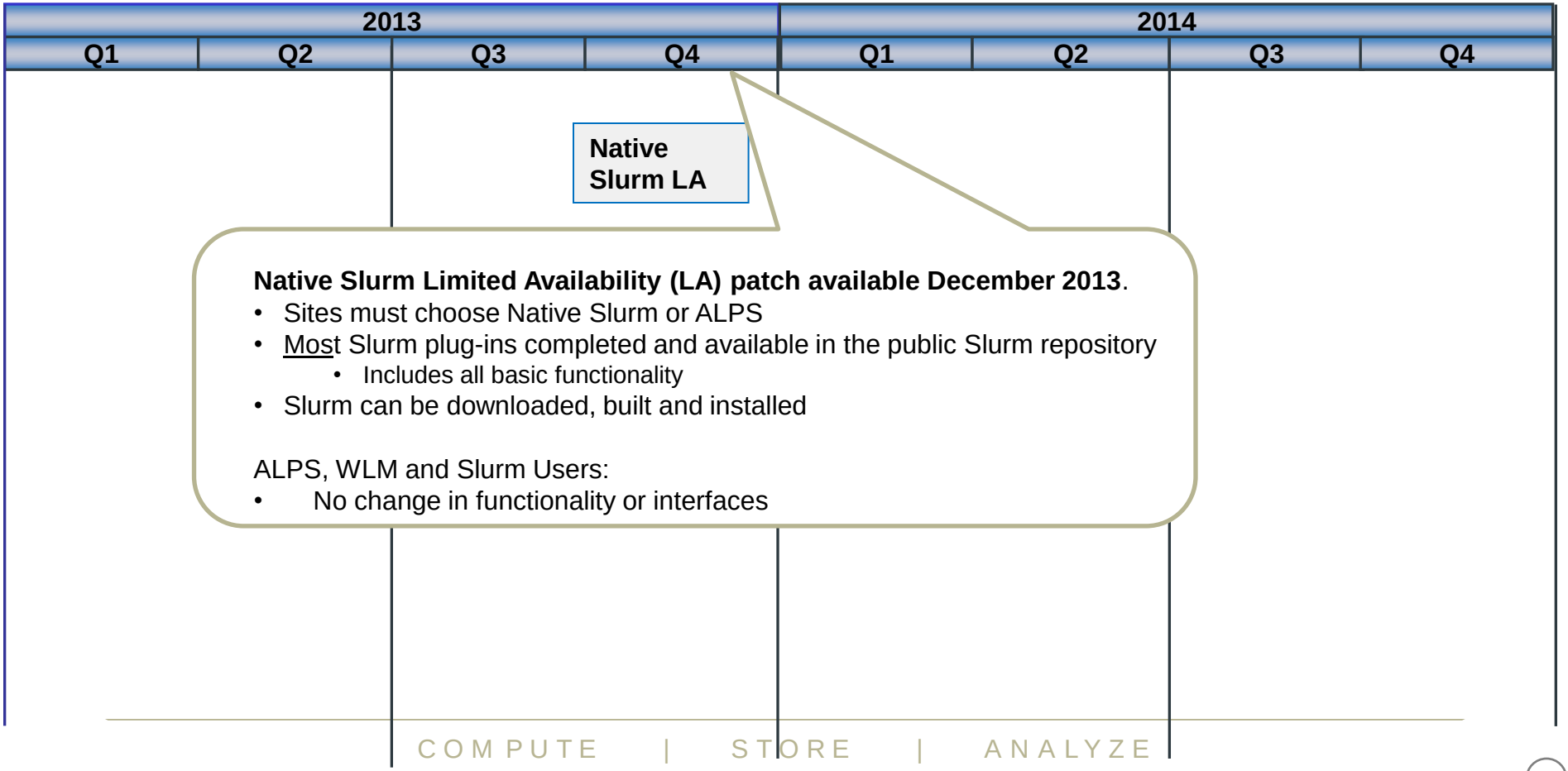
ALPS, WLM and Slurm Users:

- No change in functionality or interfaces

Copyright 2014 Cray Inc. – Confidential and Proprietary

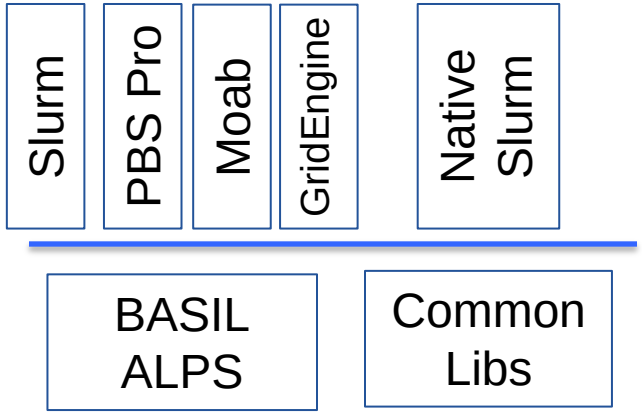
WLM Roadmap: Phase 2

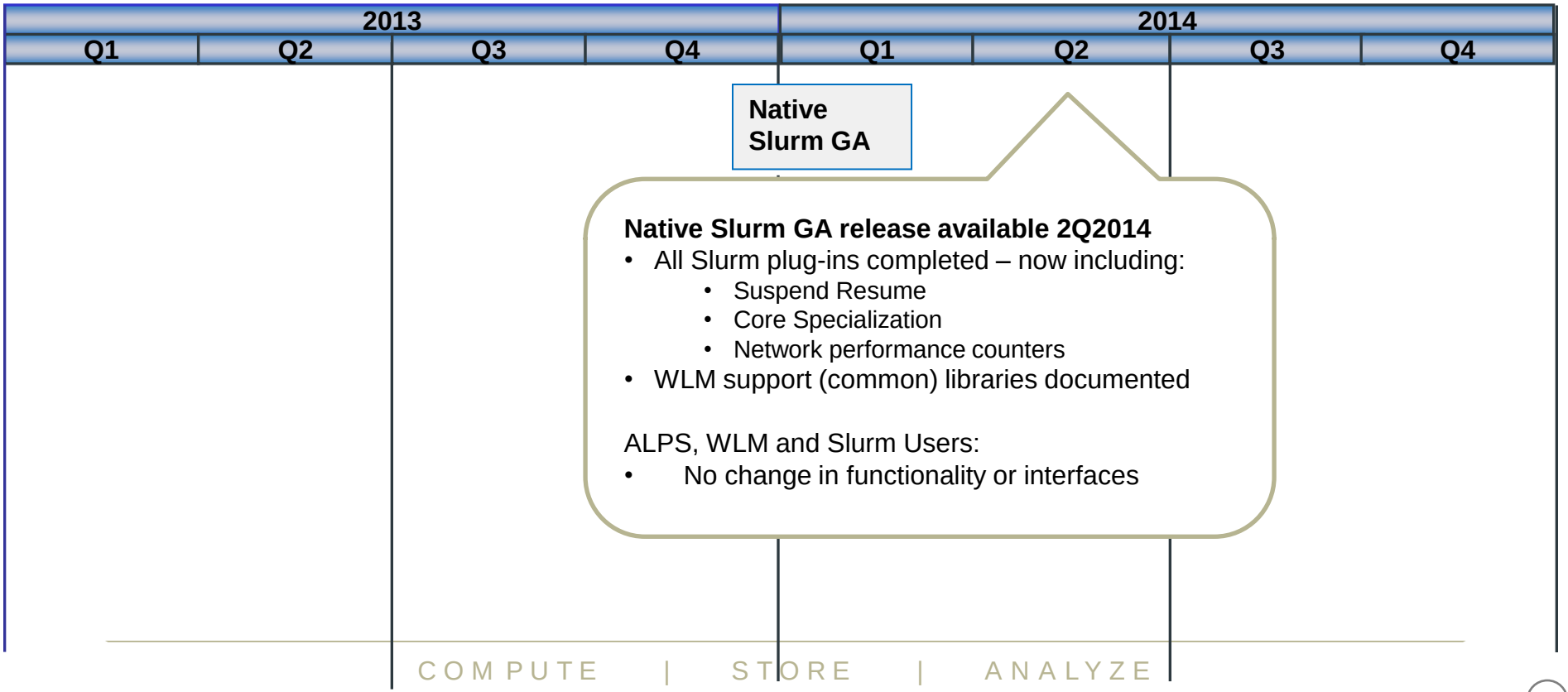




Copyright 2014 Cray Inc. – Confidential and Proprietary

WLM Roadmap: Phase 2a



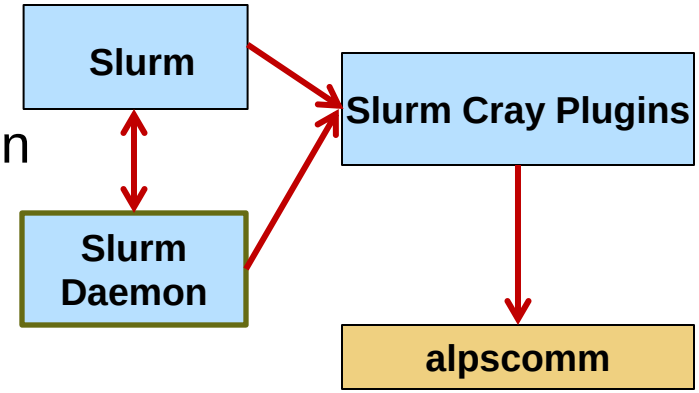


Copyright 2014 Cray Inc. – Confidential and Proprietary

Native Slurm Architecture for Cray

Slurm

- Prioritizes queue(s) of work
- Decides when and where to start jobs
- Terminates job when appropriate
- Accounting for jobs and job steps
- Allocates and releases resources for jobs



Slurm

- Launches tasks
- Monitors node health
- Manages node state
- Has daemons on compute nodes
- Plugin changes to:
  - Select
  - Switch
  - Task
  - Job Container (new)

alpscomm

- Low level interfaces for network management

Copyright 2014 Cray Inc. – Confidential and Proprietary

Any WLM could use srun as a launcher

- **Native use of Workload Managers**
  - Native WLM functionality will provide resource management, scheduling and reporting functions
  - Target srun as a launcher for any WLM without a launch mechanism
    - Majority of ‘mpirun’ implementations already interface to *srun* as launcher

Copyright 2014 Cray Inc. – Confidential and Proprietary

Summary of CLE Infrastructure Changes

- ALPS code extracted to stand alone C library APIs for Cray services
- Service node and compute node libraries provided
- API services provided in WLM plugins
- **No changes in programming model code from PE**
- **Three new daemons**
  - ncmd – network cookie management daemon
    - One per system, runs on boot node
  - aeld – provides application placement info to HSS
    - Used in network congestion management
  - apptermd – application termination daemon
    - Kills apps as directed by network congestion management
- **No ALPS daemons or user commands in native mode**

Copyright 2014 Cray Inc. – Confidential and Proprietary

Cray Specific Services that alpscomm provides

- Cookie/protection Key management
- Configure Aries driver
- Configure reserved access to Network Performance Counters (NPC)
- Provide topology info for NPC
- Memory compaction
- Compute Node Clean Up
- Provide info to ISV Application Acceleration (IAA) for third party application launches
- Suspend/Resume

Copyright 2014 Cray Inc. – Confidential and Proprietary

Functionality Differences in native mode

- New ALPS APIs can be invoked by the native WLM (Slurm) plugins
- srun only (or WLM launcher)
- A `viewcookies` command is provided to display info about assigned protection Keys (pKeys) and cookies
- WLM can use its own daemons on the compute nodes
- Native WLM will provide its own user/launch, status and admin commands
- Application launch will need to know if suspend/resume is enabled
  - Allows for network resources to be scaled for each launch
- Third party MPI launch commands will be supported for non-Cray PMI based apps

Copyright 2014 Cray Inc. – Confidential and Proprietary

Functionality Not Provided in native mode

- Aries only, no Gemini support
- Checkpoint/Restart is not supported on Aries systems
- Following srun options are not supported
  - Checkpoint, reboot and tmp
- No Cray accounting support (WLM accounting is provided)
- Pre-reservation of huge pages
- No RCA or event support on compute node
  - WLM must detect node failures on its own

Copyright 2014 Cray Inc. – Confidential and Proprietary

The Future of ALPS

- ALPS is not being deprecated or removed
- Existing WLMs are still functional with ALPS
  - PBSPro, Moab/Torque, GridEngine, Slurm
- Code within the new APIs will not contain any WLM specific code
- Two models were never intended to be equivalent in functionality

Copyright 2014 Cray Inc. – Confidential and Proprietary

Current Native Slurm Progress

- **LA released Dec 2013 and is being tested**
- **Planned GA Content – Mid 2014**
  - Core Specialization
  - Network Performance Counters
  - GPU Support
  - MIC Support for accelerated mode
  - Suspend/Resume – Job Preemption
  - Third Party Application Launch Support
  - Defect fixes

Copyright 2014 Cray Inc. – Confidential and Proprietary

Workload Manager Support Site

- Certification Process
- Certified WLM versions
- Significant Known Problems
- Links to each WLM homepage

Copyright 2014 Cray Inc. – Confidential and Proprietary

Crayport Workload Manager Portal

- http://crayport.cray.com/3rdPartyBatchSW/Forms/compatibility_info.aspx

Copyright 2014 Cray Inc. – Confidential and Proprietary

CrayPort Page Customer Input

- **How many use this site?**
- **Suggestions of what you would like to see here**
  - Send input to:

bce@cray.com or spswlm@cray.com

Copyright 2014 Cray Inc. – Confidential and Proprietary

Conclusions – Take Aways

- Native infrastructure is available to all WLM partners
- Functionality between models is NOT the same
- Take advantage of the Cray WLM Portal Information
- Customers select their WLM to run
- Cray remains agnostic to the choice of a WLM
- WLMs enhance the Cray experience

Copyright 2014 Cray Inc. – Confidential and Proprietary

Partnerships

- Vendor partnerships help us provide superior products
- Customer partnerships are key to our success

Copyright 2014 Cray Inc. – Confidential and Proprietary

Thank you.

Questions?

Copyright 2014 Cray Inc. – Confidential and Proprietary

Legal Disclaimer

Information in this document is provided in connection with Cray Inc. products. No license, express or implied, to any intellectual property rights is granted by this document.

Cray Inc. may make changes to specifications and product descriptions at any time, without notice.

All products, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.

Cray hardware and software products may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Cray uses codenames internally to identify products that are in development and not yet publically announced for release. Customers and other third parties are not authorized by Cray Inc. to use codenames in advertising, promotion or marketing and any use of Cray Inc. internal codenames is at the sole risk of the user.

Performance tests and ratings are measured using specific systems and/or components and reflect the approximate performance of Cray Inc. products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.

The following are trademarks of Cray Inc. and are registered in the United States and other countries: CRAY and design, SONEXION, URIKA, and YARCDATA. The following are trademarks of Cray Inc.: ACE, APPRENTICE2, CHAPEL, CLUSTER CONNECT, CRAYPAT, CRAYPORT, ECOPHLEX, LIBSCI, NODEKARE, THREADSTORM. The following system family marks, and associated model number marks, are trademarks of Cray Inc.: CS, CX, XC, XE, XK, XMT, and XT. The registered trademark LINUX is used pursuant to a sublicense from LMI, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis. Other trademarks used in this document are the property of their respective owners.

Copyright 2013 Cray Inc.