CRAY®

“Native” SLURM on Cray XC30

SLURM ‘Birds of a Feather’ SC’13

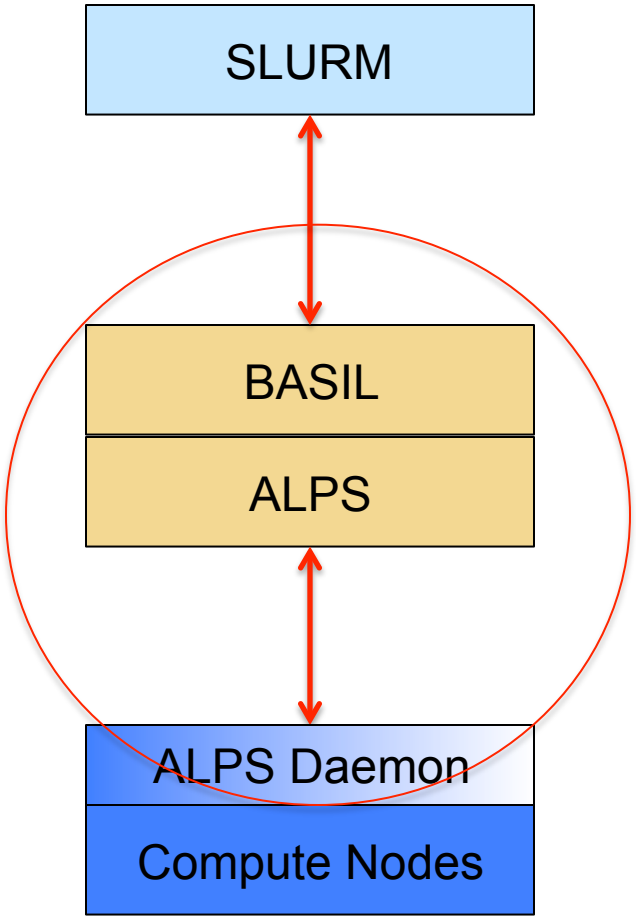
What's being offered?

- **SLURM / ALPS**
  - The current open source version available on the SchedMD/SLURM web page
    - SLURM 2.6 validated for Cray systems
  - Basic WLM functions
  - This version supports most Cray features, but is a subset
    - Cray doesn't support all of the SLURM capabilities, conversely SLURM doesn't support all of the ALPS capabilities
  - Cray Cluster Compatibility Mode **<NEW>**
  - Cray has contract(s) to add enhancements to SLURM for Cray systems
    - These and existing enhancements will be pushed upstream to SchedMD to be included in open source SLURM repository

"Hybrid" SLURM Architecture for Cray

SLURM

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



ALPS

- Allocates and releases resources for jobs (as directed by SLURM)
- Launches tasks
- Monitors node health
- Manages node state
- Has daemons on compute nodes
- Manages Cray network resources

SLURM is a scheduler layer above ALPS and BASIL, not currently a replacement

ALPS Refactoring: Motivation

Evolving system requirements driving changes

- Workload manager role not “just launching jobs”
- The role of a resource manager is managing job’s resource requirements *throughout* job
- The resource manager’s work only starts when a job begins processing
- The information a resource manager needs is constantly changing
- Resources a job needs are constantly changing
- Resiliency is an application’s responsibility with system’s assistance

From SLURM User Group Meeting Keynote Address, 2011

Motivation

Make Cray Systems More Accessible with Additional WLMs

- Increase potential customer base – some RFPs require WLMs that we do not support
  - LSF
  - GridEngine
- Customers have expressed desire to run the same WLM across entire enterprise and/or data center

“Native” SLURM

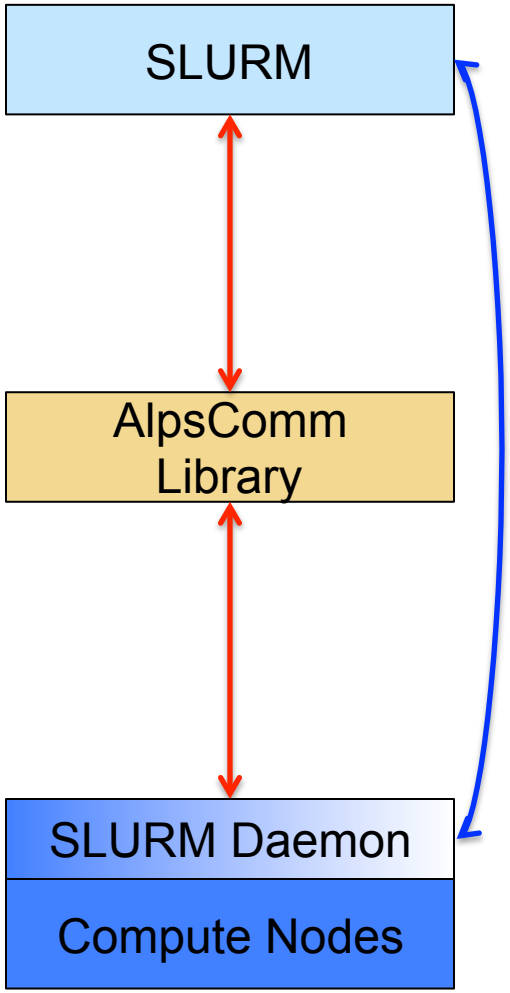
Native SLURM

- **Create library of low-level ALPS functions and h/w APIs**
- **Develop a "native" SLURM implementation**
  - "Native" means no interaction with ALPS/BASIL
  - Cray developed plugins to provide following services (correspond to ALPS common APIs):
    - Dynamic node state change information
    - System topology information
    - Congestion management information for HSS
    - Protection key and protection domain management
    - Node Health Check support
    - Network performance counter management
    - PMI port assignment management (when more than one application per compute node)
  - Working with SchedMD on implementation

“Native” SLURM Architecture for Cray

SLURM

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



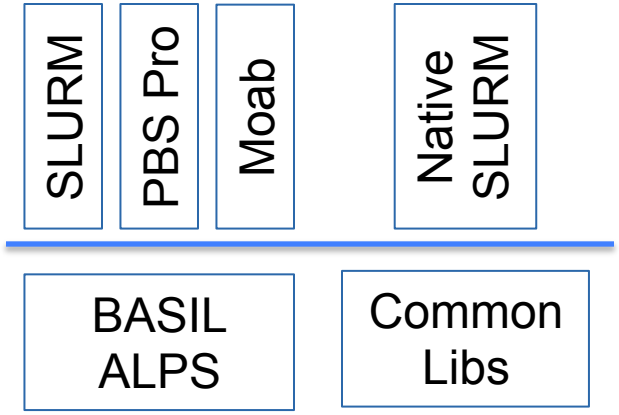
SLURM

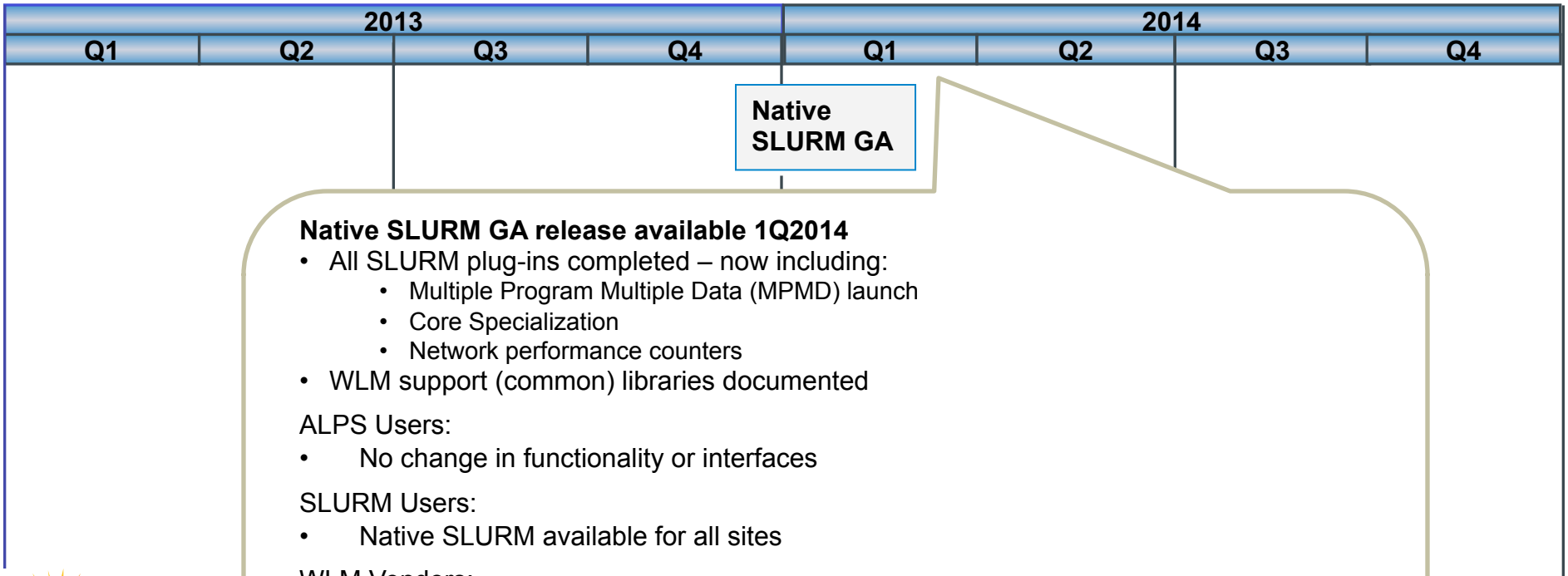
- Launches tasks
- Monitors node health
- Manages node state
- Has daemon on compute nodes
- Plugin changes to:
  - Select
  - Switch
  - Task
  - ProcTrack
  - Cgroup

AlpsComm

- Low level interfaces for network management

WLM Roadmap:





Native SLURM GA release available 1Q2014

- All SLURM plug-ins completed – now including:
  - Multiple Program Multiple Data (MPMD) launch
  - Core Specialization
  - Network performance counters
- WLM support (common) libraries documented

ALPS Users:

- No change in functionality or interfaces

SLURM Users:

- Native SLURM available for all sites

WLM Vendors:

- WLM that are currently integrated with ALPS continue to work
- WLM (common) libraries documented and available for vendors to access/use directly

CRAY®

Questions?