Slurm Native Workload Management on Cray Systems

David Bigagli
david@schedmd.com

SchedMD LLC
http://www.schedmd.com

Cray Architecture

- Many of the most powerful computers built by Cray
- Nodes are diskless
- 2 or 3-dimension torus interconnect
  - Multiple nodes at each coordinate on some systems
- Full Linux on front-end nodes
- Lightweight Linux kernel on compute nodes
- Whole nodes must be allocated to jobs

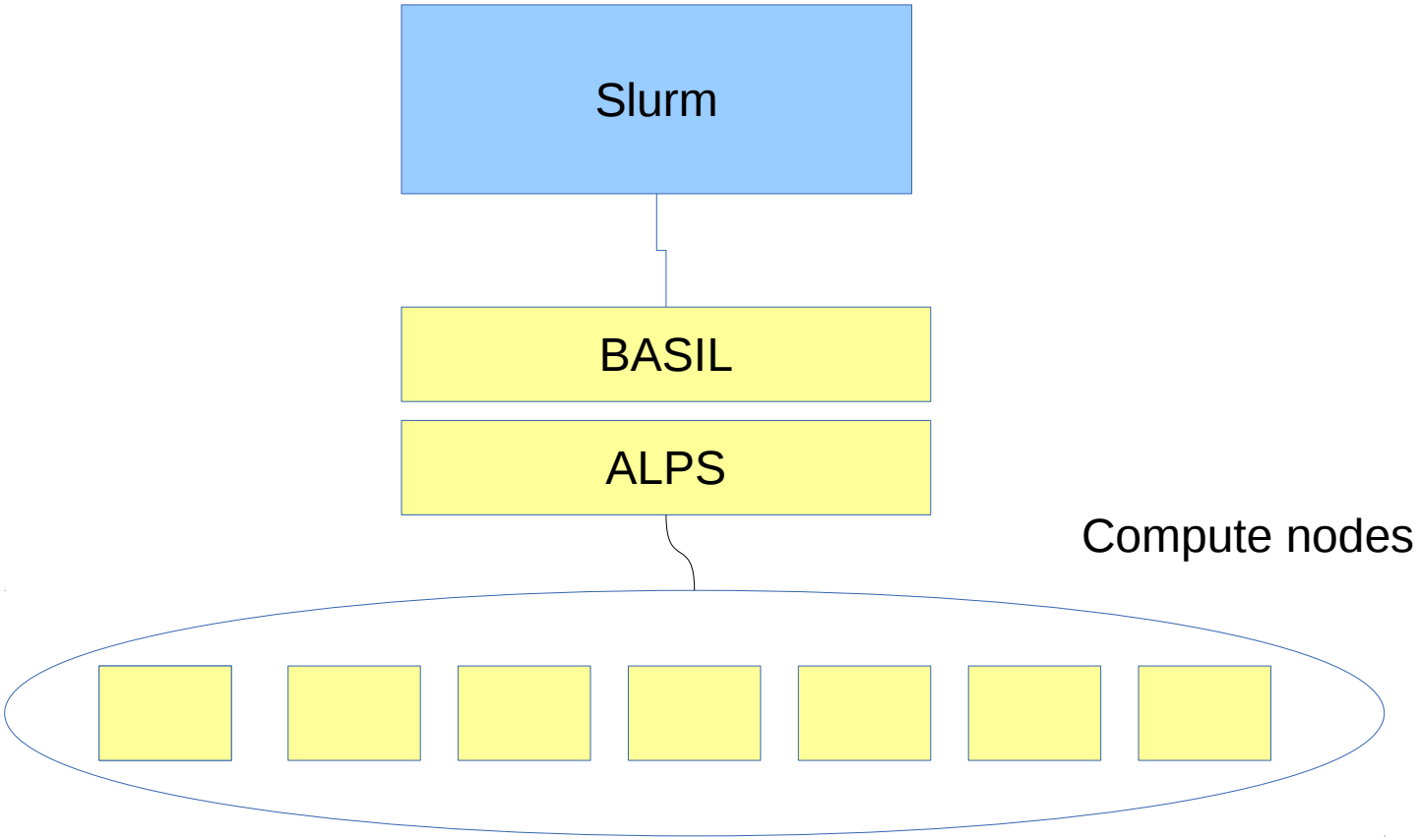
ALPS and BASIL

- ALPS – **A**pplication **L**evel **P**lacement **S**cheduler
  - Cray's resource manager
  - Six daemons plus variety of tools
    - One daemon runs on each compute node to launch user tasks
    - Other daemons run on service nodes
  - Rudimentary scheduling software
    - Dependent upon external scheduler (e.g. SLURM, etc) for workload management
- BASIL – **B**atch **A**pplication **S**cheduler **I**nterface **L**ayer
  - XML interface to ALPS

Slurm and ALPS Functionality

- Slurm
  - Prioritize queues and enforces limits
  - Scheduling and accounting of jobs
  - Manages
- ALPS
  - Allocated and releases resources for jobs
  - Launches tasks
  - Monitors node health

Slurm Architecture for Cray



Job Launch Process

- User submits a job script
- Slurmctld creates an ALPS reservation
- Slurmctld sends the job script to slurmd

- Slurmd claims the reservation for the session ID
- Slurmd launches the user script
- Aprun launches the tasks on compute nodes
- When the job finishes the reservation is released

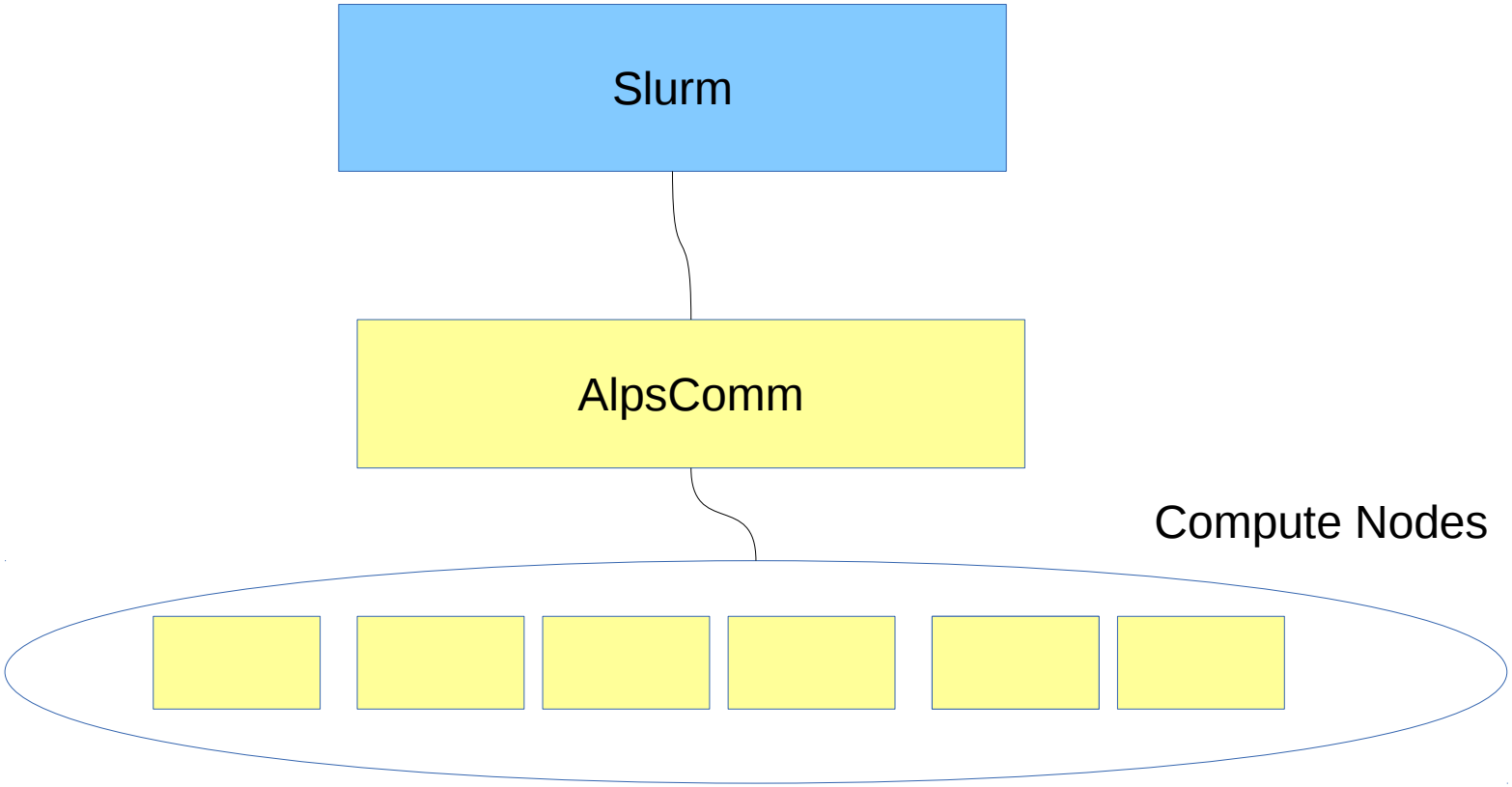
Motivation for Native Slurm Implementation on Cray

- Current architecture has limitations due to the translation from Slurm to ALPS
- Not all features of Slurm can be used, e.g. run multiple jobs per node
- Allow native Slurm functionality scheduling, resource management and reporting
- Majority of mpirun implementations already interface to Slurm as launcher with the srun command

Slurm Native Implementation on Cray System

- Cray and SchedMD developed plugins to provide the following services:
- Dynamic node state change information
- System topology information
- Node Health Check Support
- Network performance counter management
- Congestion management information for Cray Hardware Supervisory System

Slurm Native Architecture



Slurm Cray Specific Feature

- Network Performance Counters
  - To access the Cray's NPC use –network option in sbatch/salloc/srun commands
  - --network=system for the system wide NPC
  - --network=blade for the blade NPC
- Core Specialization
  - Ability to reserve number of cores allocated to the job and not used by the application

Slurm Configuration

- Configure plugins to use Cray without ALPS.
- CoreSpec
  - To use set CoreSpecPlugin=core_spec/cray
- Job Submit
  - To use set JobSubmitPlugin=job_submit/cray
- Process tracking
  - To use set ProctrackType=proctrack/cray
- Select
  - To use set SelectType=select/cray
- Switch
  - To use set SwitchType=switch/cray
- Task
  - To use set TaskPlugin=cray. It could be used with other task plugins as well

Q&A

Thanks!