
Deploying SLURM on XT, XE, and Future Cray Systems

Kris Howard

CSCS

Presenting research by:

Gerrit Renker, Neil Stringfellow, Sadaf Alam, Stephen Trofinoff, and Kris Howard

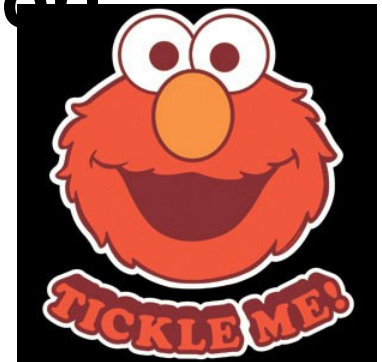
Introduction

CSCS

- **Cray Early Adopter**
 - **1st XT3 in Europe**
 - **1st XE6 & XMT2 in the World**
- **Customized Scheduling and Accounting**

History

- **PBSPPro**
- **TCL Scheduler**
 - **Initially developed by Jason Coverston of Cray**
 - **No more TCL support in version 11**



What we need

Requirements

- **Agressive Backfilling**
- **Bottom feeder Policies**
- **Specific MeteoSwiss Extensions**
- **Robust Extension (Scheduler health Monitor, etc)**

Alternatives

- **LSF, PBSPro 11, Torque w/ MOAB...**

What we found

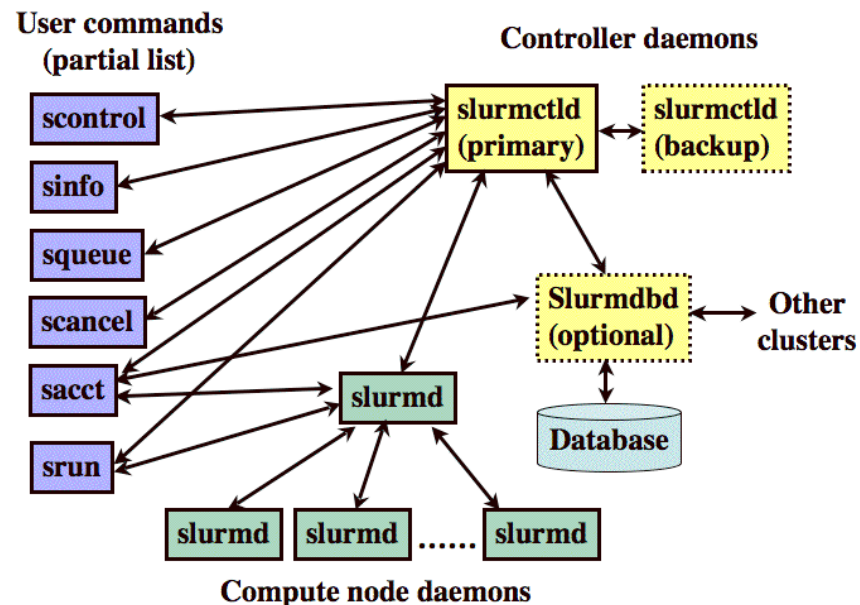
SLURM

- **LLNL**
- **Fulfills Key Requirements**
 - **Extensibility**
 - **Flexibility**
 - **Maintainability**
- **Top Systems**
 - **Tiahne 1-A, CEA, Japan, Bluegene/Q**

SLURM

SLURM Architecture

- **Commands**
- **Accounting**
 - Multicluster
 - Limit Enforcement
- **Plugins**
- **External Schedulers**

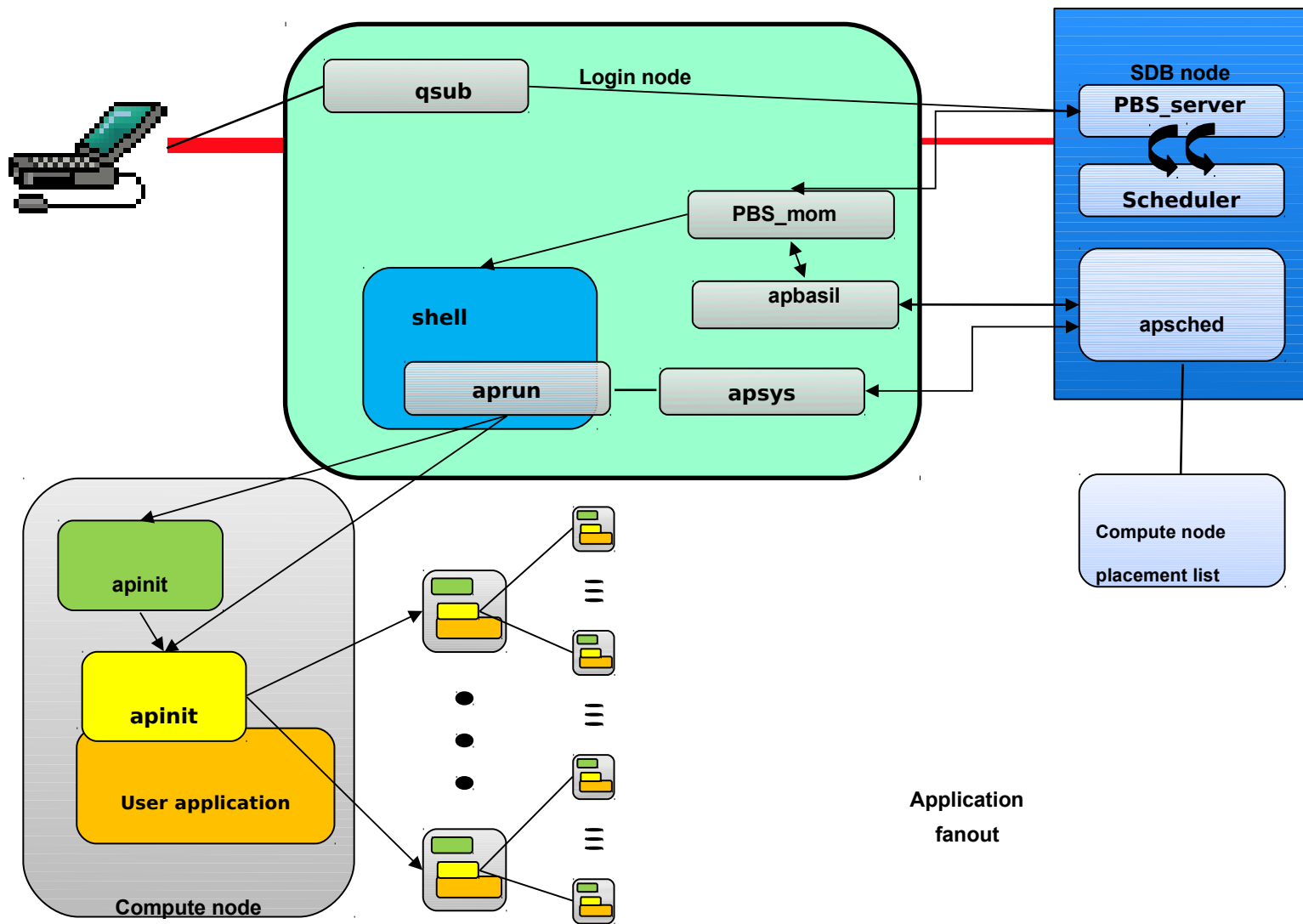


Porting Efforts

Porting

- **XML-RPC interface to ALPS/BASIL**
 - **QUERY**
 - **RESERVE**
 - **CONFIRM**
 - **RELEASE**
- **Documentation**
- **SRUN vs. APRUN**

ALPS



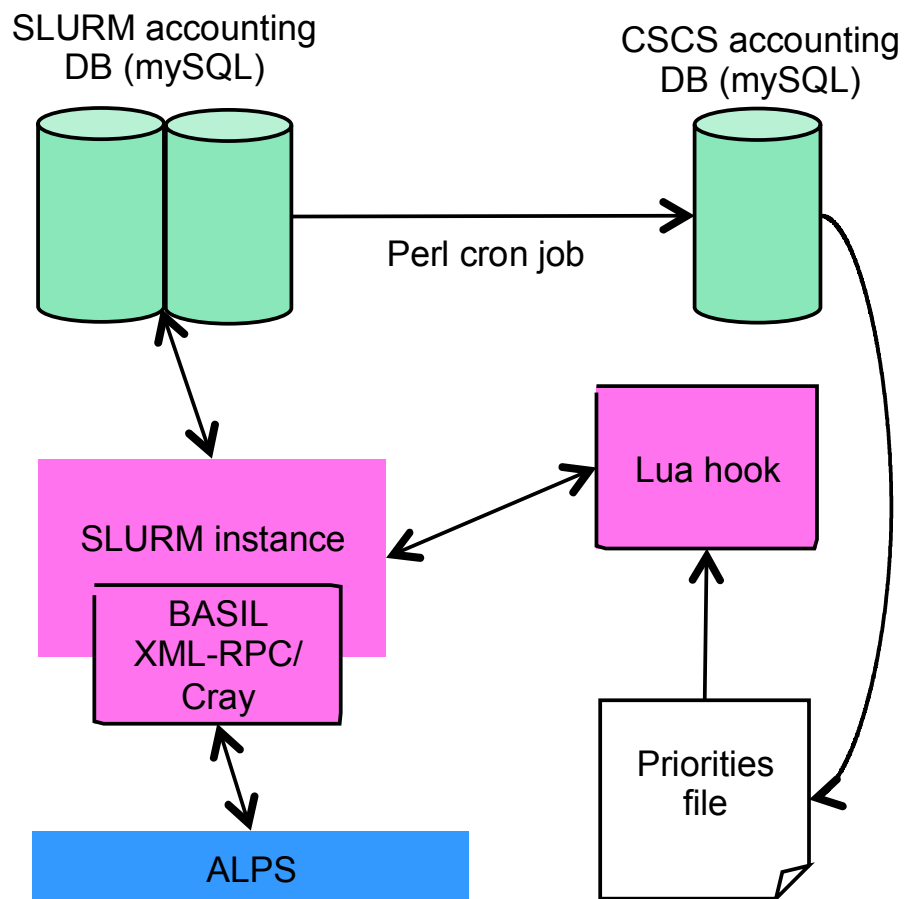
Getting used to the new quirks

- **PBS qsub**
- **SLURM sbatch/salloc**
- **Login node “MoM”**
- **Batch vs. Interactive**
- **Shell Variables in Batch scripts**
- **Job limits**
- **Ulimits**

Biting the Bullet

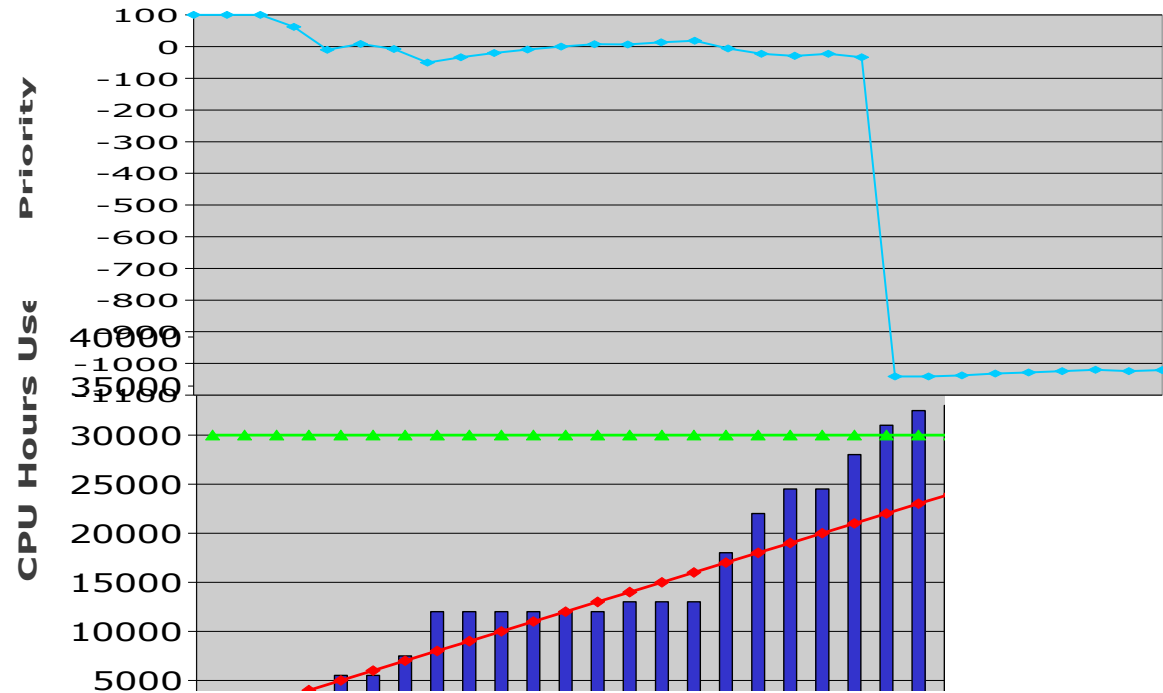
- **GELE XT5 1 cab TDS**
- **PALU XE6 2 cab**
- **ROSA XT5 20 cab**
- **(FUJI, RIGI, ROTHORN, JULIER)**
x86_64 clusters

Accounting



Priorities

**Actual job priority = Aging*age_multiplier +
Group Priority*group_multiplier +
Job_size*num_cores*job_multiplier**



Examples of Top Dog and Bottom Feeders

Job priority

11,124

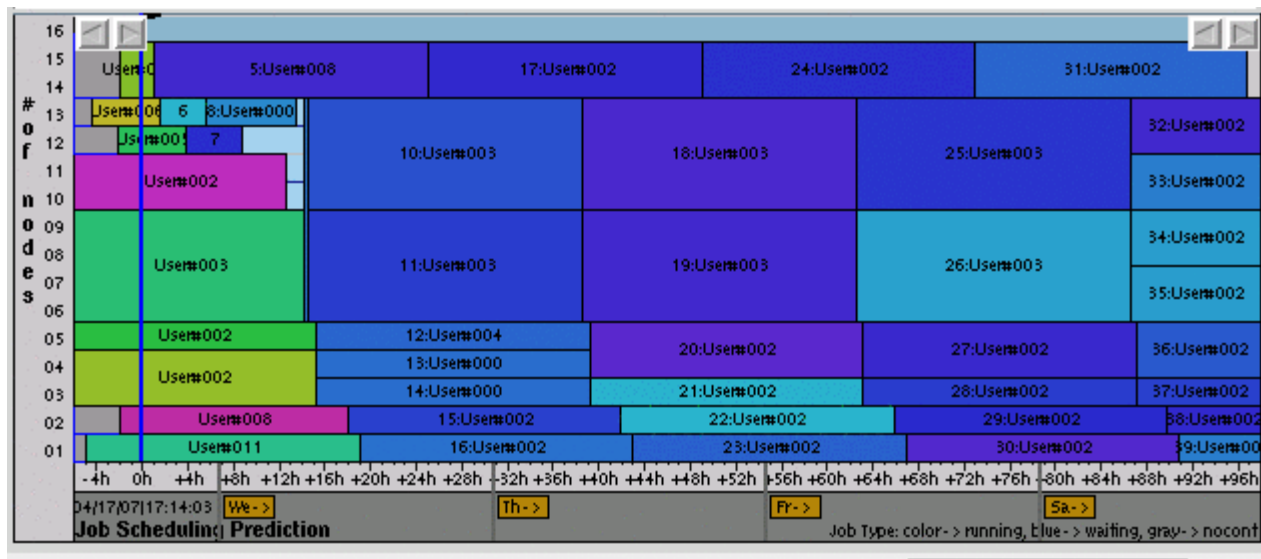
10,240

8,000

...

4

2



Lview image for example purposes

Eating our own SLURM



Tastes Great!

More Filling!

Future

Scheduling – porting the TCL Scheduler

Node sharing

- **Code Auto tuning**
- **Testing for scaling to larger machines**

Accounting

- **Mazama**

GPUs



Questions?