



# Scalable Systems Software

## Developing Systems Management Tools for TeraScale Computer Centers

Ames, ANL, Cray, IBM, Intel, LANL, LBNL, ORNL, NCSA, PNNL, PSC, SGI, SNL

### Problem

System administrators and managers of terascale computer centers are facing a crisis:

- Computer centers use incompatible, ad hoc set of systems tools
- Present tools are not designed to scale to multi-Teraflop systems
- Commercial HPC solutions not happening as business forces drive industry towards servers rather than HPC

### Integrated Suite

Leverage OSCAR (Open Source Cluster Application Resources)

- Benefits
  - Modular cluster package API
  - OSCAR framework
  - Installation/distribution process
- Community
  - Has been adopted by many cluster vendors
  - Ten's of thousands of downloads
  - Raises SSS software suite's profile and availability



### SSS-OSCAR

- Version 1.0 release at SC2004
- Version 1.1 release June 2005

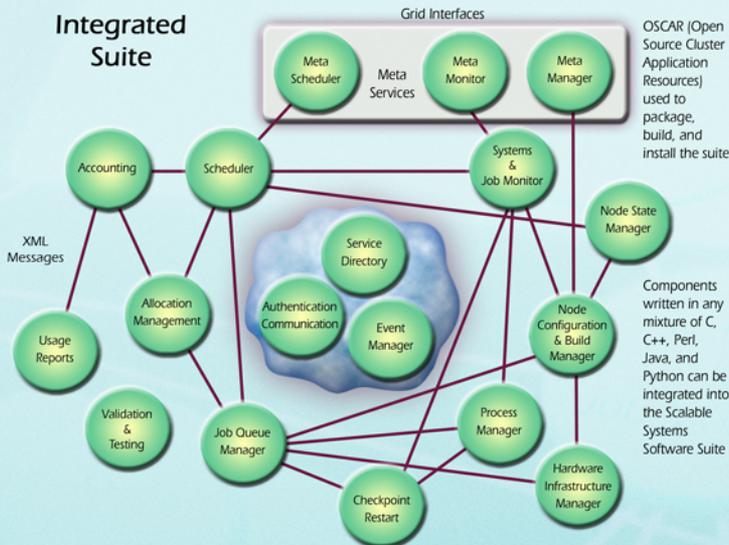
### Cobalt

- SSS suite for IBM BlueGene

### Impact

Fundamentally change the way future high-end systems software is developed and distributed

- Reduced facility management costs
  - reduce duplication of effort rewriting components
  - reduce need to support ad hoc software
  - better systems tools available
  - able to get machines up and running faster and keep running
- More effective use of machines by scientific applications
  - scalable launch of jobs and checkpoint/restart
  - job monitoring and management tools
  - allocation management interface



### Goals

Design a modular system software architecture

- Portable across diverse hardware, make easy to adopt - allows plug and play components, and is language and wire protocol independent.

Collectively (with industry) agree on and specify standardized interfaces between system components

- MPI-like process to promote interoperability, portability, and long-term usability.

Produce a fully integrated suite of systems software and tools

- Reference Implementation for the management and utilization of terascale computational resources.

### Modular Architecture Design

- Make it easy for sites to Adopt
  - Easily replace a component that doesn't meet their needs
  - Use only parts of the suite that they need
  - Components can be shared across facilities
  - Open Source to allow sites to modify at will
- Components have well defined roles
  - Independent of language and wire protocol
  - Communicate through XML messages
- Service Directory, Event Manager and Communication Lib
  - Form core and interact with all other components
  - Provide plug and play registration and notification
- Multiple communication protocols are supported.
  - Components can use one or more of the wire protocols supplied in the communication library
  - http(s), ssl, tcp, zlib, challenge authentication, more...
  - The set of wire protocols is extensible

### Production Users

- Running a Full Suite in Production for over a year
  - Argonne National Lab – 200 node Chiba City, BG/L
  - Ames Lab
- Running one or more components in Production
  - Pacific Northwest National Lab – 11.4 TF cluster + others
  - NCSA
- Running full suite on development systems
  - Most participants

### Adoption of API

- Maui Scheduler now uses our API in client and server
  - 3,000 downloads/month
  - 75 of the top 100 supercomputers in TOP 500
- Commercial Moab Scheduler uses SSS API
  - Users: Amazon.com, Boeing, Ford, Dow Chemical, Lockheed-Martin, more...
- New Capabilities added to Schedulers due to API
  - fairness, higher system utilization, improved response time