

# The Integration of Scalable Systems Software with the OSCAR Clustering Toolkit

John Mugler, Thomas Naughton and Stephen L. Scott  
Oak Ridge National Laboratory  
Oak Ridge, TN, USA

# Introduction

- OSCAR Overview
- SSS Project Description
- SSS + OSCAR Integration/Deployment
- Challenges and Observations
- Release Status

# OSCAR

## Open Source Cluster Application Resources

Snapshot of best known methods for building, programming and using clusters.

Consortium of academic/research & industry members.



# OSCAR Project Organization

- Open Cluster Group (OCG)
  - Informal group formed to make cluster computing more practical for HPC research and development
  - Membership is open, direct by steering committee
- OCG working groups
  - OSCAR
  - Thin-OSCAR (diskless)
  - HA-OSCAR (high availability)

# OSCAR 2004 Core Members

- Intel
- RevolutionLinux
- Bald Guy Software
- Indiana University
- NCSA
- Oak Ridge National Lab

*The project direction is determined through consensus of core organization voting.*

# What does OSCAR do?

- Wizard based cluster software installation
  - Operating system
  - Cluster environment
- Automatically configures cluster components
- Increases consistency among cluster builds
- Reduces time to build / install a cluster
- Reduces need for expertise

# Basic Design

- Use “best known methods”
  - Leverage existing technology where possible
- OSCAR framework
  - Remote installation facility
  - Small set of “core” components
  - Modular package & test facility
  - Package repositories

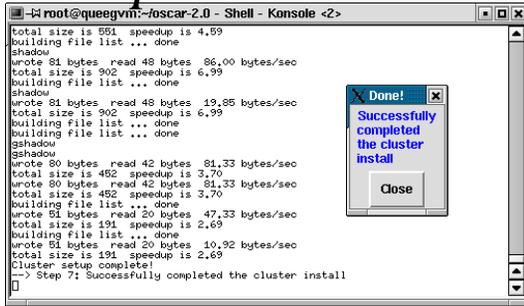
# Component Areas

- Core Infrastructure/Management
  - System Installation Suite (SIS), C3, Env-Switcher,
  - Database (ODA),
  - Package Downloader (OPD)
- Administration/Configuration
  - SIS, C3, OPIUM, cluster services (dhcp, nfs, ntp...)
  - Security
- HPC Services/Tools
  - Parallel & Scientific Libraries
  - Batch scheduler & queuing system
  - Monitoring systems

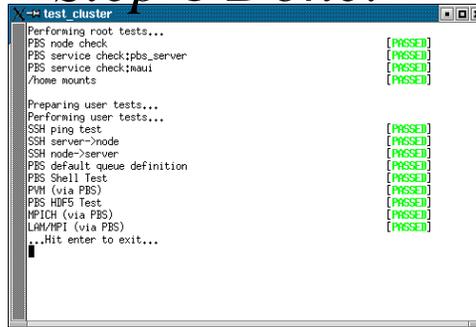
# Package-based Framework

- Content:
  - Software + Configuration, Tests, Docs
  - RPM's for Software
- Types:
  - Core: SIS, C3, Switcher, ODA, OPD, Support Libs
  - Non-core: selected & third-party
- Access:
  - Repositories accessible via OPD/OPDer

# Step 7



# Step 8 Done!



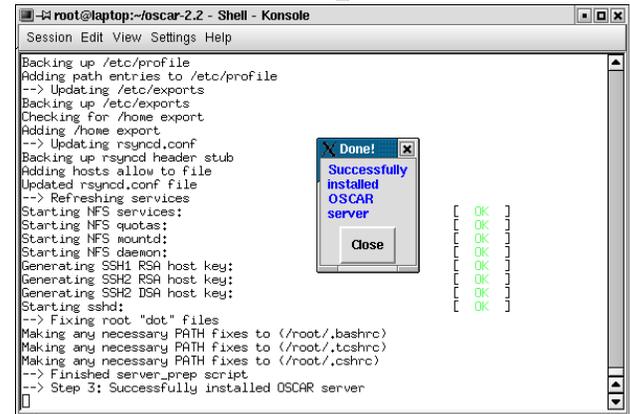
# Step 1 Start...



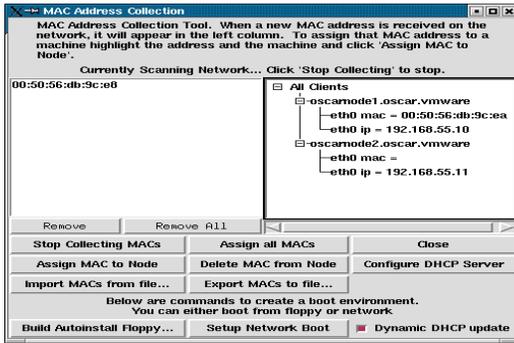
# Step 2



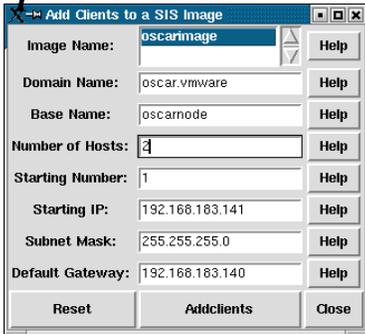
# Step 3



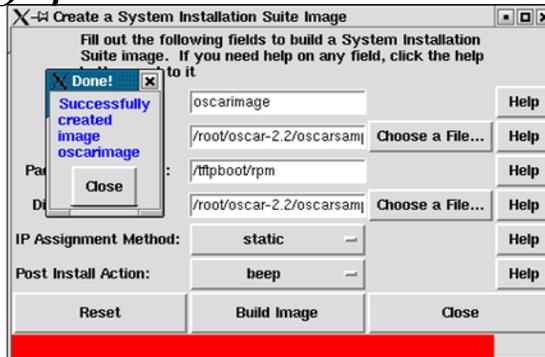
# Step 6



# Step 5



# Step 4



# Scalable System Software

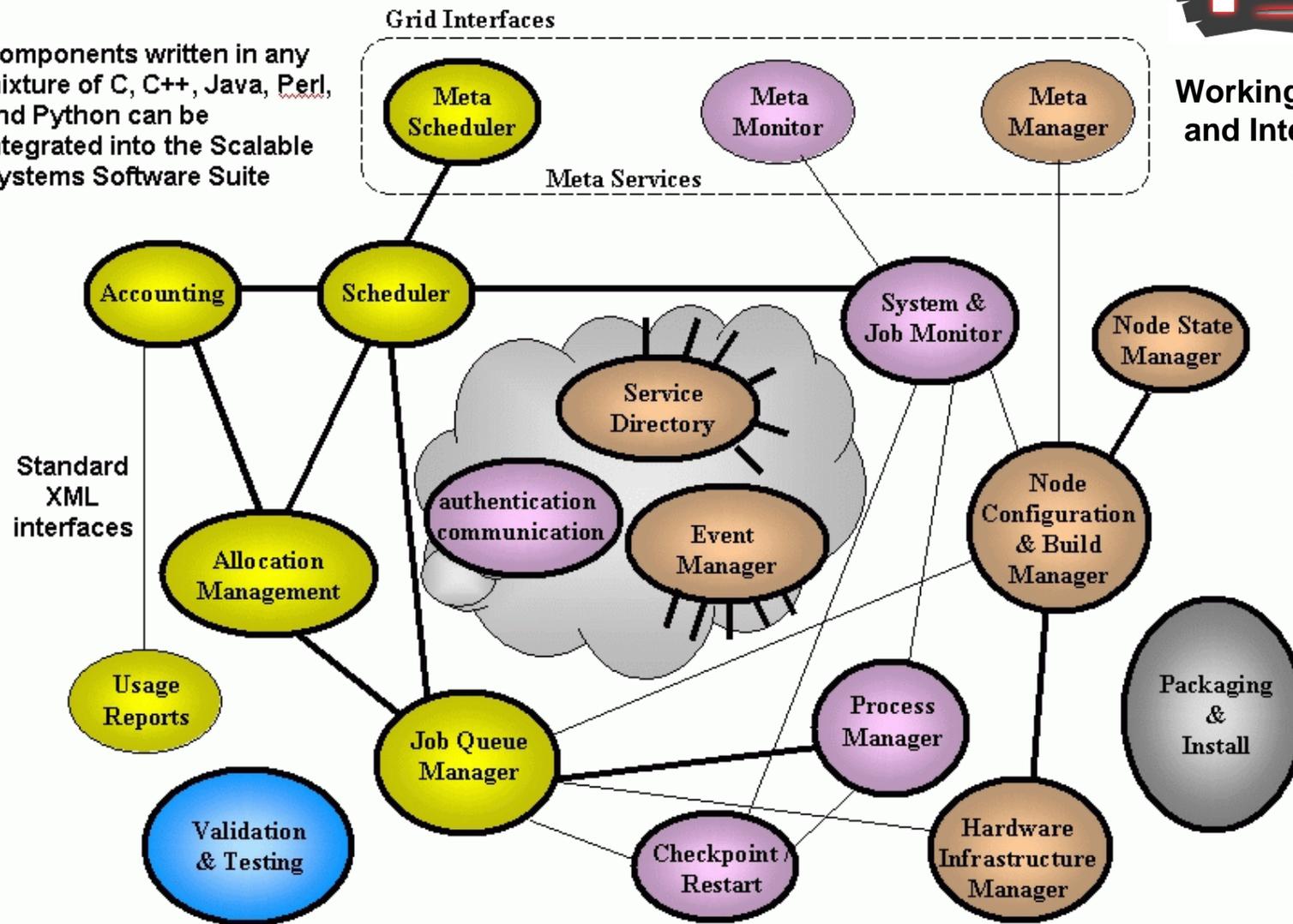


- Problems
  - Computer centers use incompatible, ad hoc set of systems tools
  - Tools are not designed to scale to multi-Teraflop systems
  - Duplication of work to try and scale tools
  - System growth vs. Administrator growth
- Goals
  - Define standard interfaces for system components
  - Create scalable, standardized management tools
  - (Subsequently) reduce costs & improve efficiency at centers
- Participants
  - Nat'l Labs: ORNL, ANL, LBNL, PNNL, SNL, LANL, Ames
  - Academics Inst.: NCSA, PSC, SDSC
  - Industry: IBM, Cray, Intel, Unlimited Scale

# SSS Project Outline

- Map out functional areas
  - Schedulers, Job Mangers
  - System Monitors
  - Accounting & User management
  - Checkpoint/Restart
  - Build & Configuration systems
- Standardize the system interfaces
  - Open forum of universities, labs, industry reps
  - Define component interfaces in XML
  - Develop communication infrastructure

Components written in any mixture of C, C++, Java, Perl, and Python can be integrated into the Scalable Systems Software Suite



# Using OSCAR for SSS

***Problem:*** Helping users obtain and install SSS software.

***Solution:*** Leverage OSCAR framework to package and distribute the SSS suite, *sss-oscar*.

*sss-oscar* → A release of OSCAR containing all SSS software in single downloadable bundle.

# OSCAR-ized SSS Components

- Bamboo – *Queue/Job Manager*
- BLCR – *Berkeley Checkpoint/Restart*
- Gold – *Accounting & Allocation Management System*
- LAM/MPI (w/ BLCR) – *Checkpoint/Restart enabled MPI*
- MAUI-SSS – *Job Scheduler*
- SSSLib – *SSS Communication library*
  - *Includes: SD, EM, PM, BCM, NSM, NWI*
- Warehouse – *Distributed System Monitor*
- MPD2 – *MPI Process Manager*

*\* As of April 2004*

# Common Ground

- Distributed developer group
  - Setup central repository, tracker, mailing lists
  - Settle upon RedHat 9.0 (x86) and oscar-3.0
  - Settle on OSD compliant licenses
- Setup testbed for devel, integration & testing
  - Add 2<sup>nd</sup> headnode to 64 node xtorc cluster
- Diverse group
  - Range of experience in cluster build & configuration
  - Range of opinions on cluster build & configuration ☺

# The Learning Curve

- Developers are new to OSCAR
- Update of OSCAR Package HOWTO (docs)
- Generate RPMs for software components
- OSCAR packaging API
  - Script selection: *post\_configure*, *post\_server\_rpm\_install*  
*post\_clients*, *post\_install*
  - Where to put “hooks”?

# Rules of Thumb

- Dividing line between RPM & OSCAR scripts?
- When to use Env-Switcher instead of profile.d?
- Installation directories, /opt vs. /usr/local/bin
- How configurable should a package be?

# Packaging Issues

- OSCAR sets up a “reasonable default”
- Which OSCAR script to use?
- Ordering within a script phase
- Shared package data, e.g., “shared key”

# Miscellaneous Issues

- Integrating lots of software pieces is tough
- This lends credence to the “package set” idea
- Removing key HPC services disturbs OSCAR test framework (currently), e.g., PBS

# Suggestions

- Improve package author tools
  - XML DTD or Schema, `xmlint` is your friend!
  - Mechanism to isolate package install/testing
  - OASIS tool?
- Script ordering within a phase
  - Can work as-is but helpful for new pkg authors
- Improve Test Framework
  - SSS project's APItest looks promising
- Higher level abstraction for dependence
  - Per package, package sets, testing, script ordering

# SSS Summary

- OSCAR serving as a SSS deployment vehicle
  - SSS integration feeding back improvements
- SSS project developing standard interface for scalable tools
  - Improve interoperability
  - Improve long-term usability & manageability
  - Reduce costs for supercomputing centers
- Currently doing testing on 2<sup>nd</sup> sss-oscar pre-release \*
  - Builds full working cluster with current SSS pkgs
  - sss-oscar-0.2a6-v3.0
  - <http://www.csm.ornl.gov/oscar/sss/>

*\* Release information as of 5/15/04*

# Online Resources

- OSCAR

<http://www.OpenClusterGroup.org/OSCAR/>

- Scalable System Software (SSS)

<http://www.scidac.org/ScalableSystems>

<http://www.csm.ornl.gov/oscar/sss/>

- “OSCAR Package HOWTO”

<http://www.csm.ornl.gov/~naughton/sss-oscar/>

Or <http://sss-oscar.sourceforge.net>