The ORNL Cluster Computing Experience…

Stephen L. Scott

Oak Ridge National Laboratory
Computer Science and Mathematics Division
Network and Cluster Computing Group

Scalable Systems Software

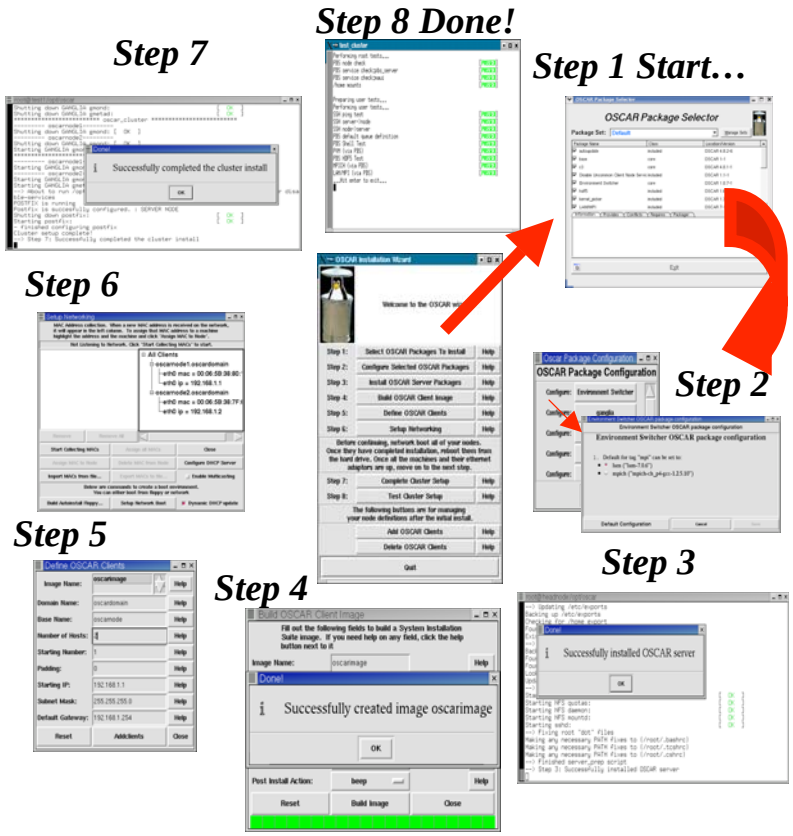
December 12, 2005
RAMS Workshop
Oak Ridge, TN

scottsl@ornl.gov www.csm.ornl.gov/~sscott

Open Source Cluster Application Resources

What is OSCAR?

- OSCAR Framework (cluster installation configuration and management)
  - Remote installation facility
  - Small set of “core” components
  - Modular package & test facility
  - Package repositories
- Use “best known methods”
  - Leverage existing technology where possible
- Wizard based cluster software installation
  - Operating system
  - Cluster environment
    - Administration
    - Operation
- Automatically configures cluster components
- Increases consistency among cluster builds
- Reduces time to build / install a cluster
- Reduces need for expertise



OSCAR Components

- Administration/Configuration
  - SIS, C3, OPIUM, Kernel-Picker, NTPconfig cluster services (dhcp, nfs, ...)
  - Security: Pfilter, OpenSSH

- HPC Services/Tools
  - Parallel Libs: MPICH, LAM/MPI, PVM
  - Torque, Maui, OpenPBS
  - HDF5
  - Ganglia, Clumon, … [monitoring systems]
  - *Other 3rd party OSCAR Packages*

- Core Infrastructure/Management
  - System Installation Suite (SIS), Cluster Command & Control (C3), Env-Switcher,
  - OSCAR DAtabase (ODA), OSCAR Package Downloader (OPD)

OSCAR Background

- Concept first discussed in January 2000
- First organizational meeting in April 2000
  - Cluster assembly is time consuming & repetitive
  - Nice to offer a toolkit to automate
- First public review at SC 2000
- First public release in April 2001
- Use “best practices” for HPC clusters
  - Leverage wealth of open source components
  - Targeted modest size cluster (single network switch)
- Form umbrella organization to oversee cluster efforts
  - Open Cluster Group (OCG)

Open Source Community Development Effort

- 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 (core group)
  - HA-OSCAR (High Availability)
  - SSS-OSCAR (Scalable Systems Software)
  - SSI-OSCAR (Single System Image)

OSCAR Core Participants

- Intel
- Bald Guy Software
- Revolution Linux
- INRIA
- EDF
- Canada's Michael Smith Genome Sciences Center
- Indiana University
- Oak Ridge National Laboratory
- Louisiana Tech Univ.
- NEC HPC Europe

SSI-OSCAR

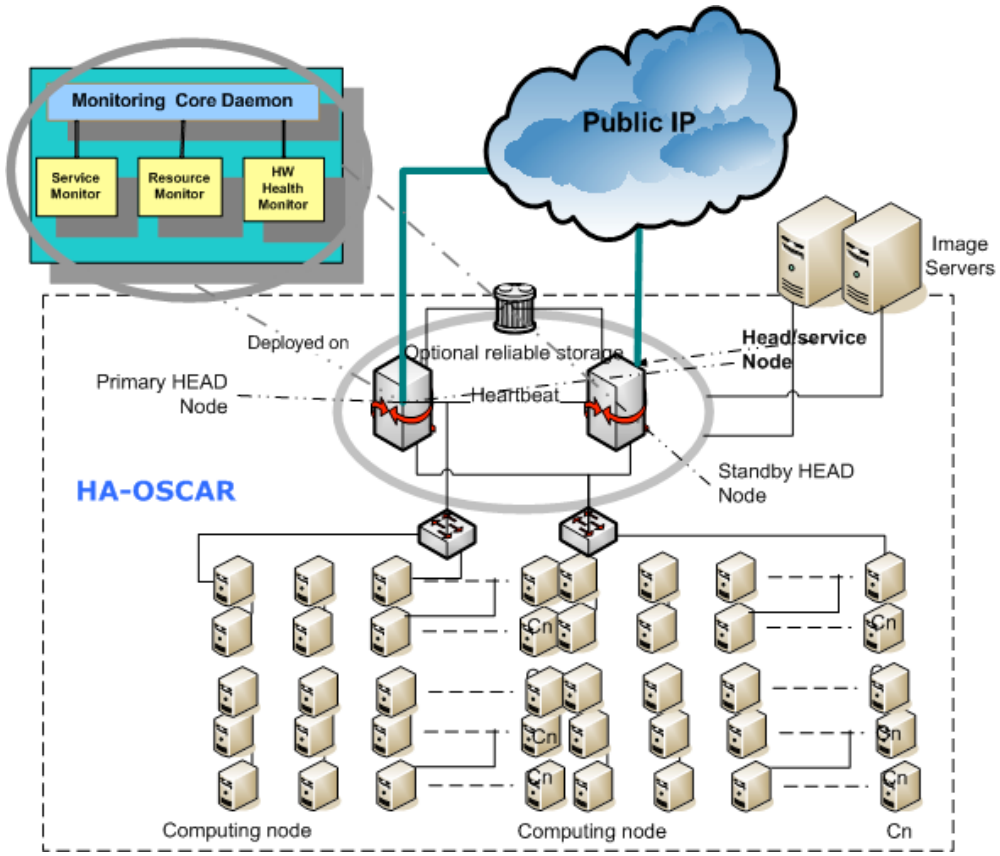
Single System Image Open Source Application Resources

- Easy use thanks to SSI systems
  - SMP illusion
  - High Performance
  - Fault Tolerance
- Easy management thanks to OSCAR
  - Automatic cluster install / update

OAK RIDGE NATIONAL LABORATORY

HA-OSCAR:

RAS Management for HPC cluster



- The first known field-grade open source HA Beowulf cluster release
- Self-configuration Multi-head Beowulf system
- HA and HPC clustering techniques to enable critical HPC infrastructure
- Active/Hot Standby
- Self-healing with 3-5 sec automatic failover time

SSS-OSCAR

Scalable System Software

Leverage OSCAR framework to package and distribute the Scalable System Software (SSS) suite, *sss-oscar*.

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

SSS project developing standard interface for scalable tools

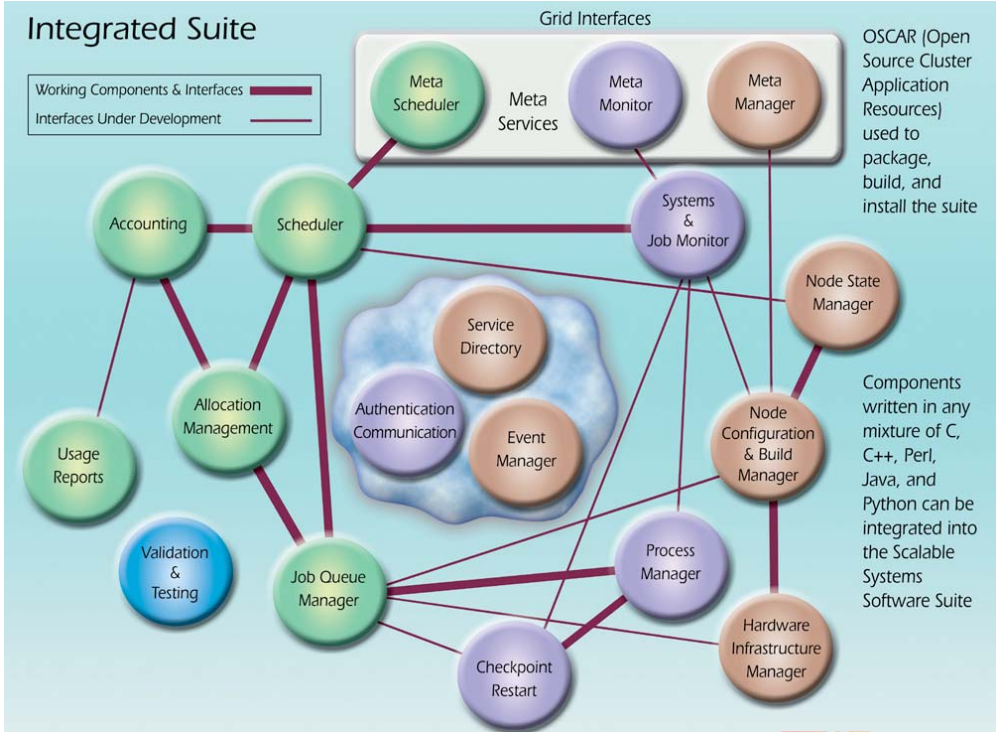
- Improve interoperability
- Improve long-term usability & manageability
- Reduce costs for supercomputing centers

Map out functional areas

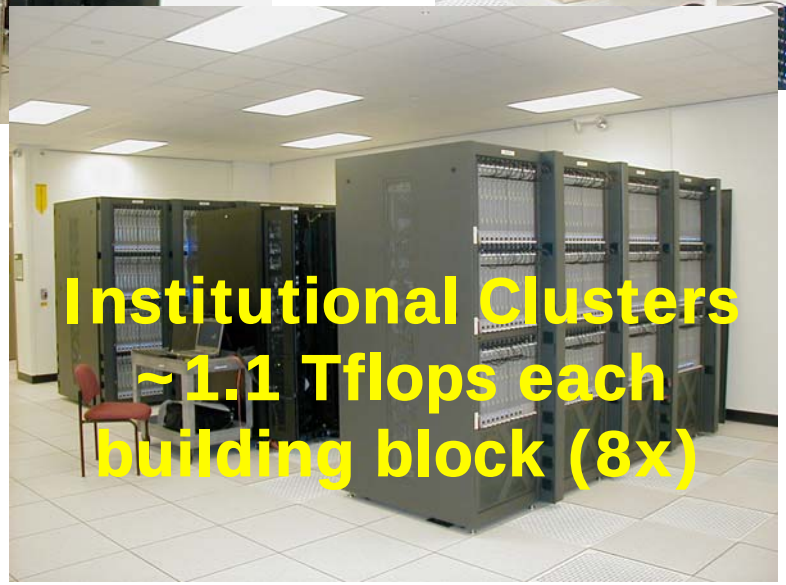
- Schedulers, Job Mangers
- System Monitors
- Accounting & User management
- Checkpoint/Restart
- Build & Configuration systems

Standardize the system interfaces

- Open forum of universities, labs, industry reps
- Define component interfaces in XML
- Develop communication infrastructure



Powered by OSCAR



C3 Power Tools

- Command-line interface for cluster system administration and parallel user tools.
- Parallel execution **`cexec`**
  - Execute across a single cluster or multiple clusters at same time
- Scatter/gather operations **`cpush/cget`**
  - Distribute or fetch files for all node(s)/cluster(s)
- Used throughout OSCAR and as underlying mechanism for tools like OPIUM's *useradd* enhancements.

C3 Building Blocks

- System administration
  - cpushimage - “push” image across cluster
  - cshutdown - Remote shutdown to reboot or halt cluster

- User & system tools
  - cpush - push single file -to- directory
  - crm - delete single file -to- directory
  - cget - retrieve files from each node
  - ckill - kill a process on each node
  - cexec - execute arbitrary command on each node
    - cexecs – serial mode, useful for debugging
  - clist – list each cluster available and it’s type
  - cname – returns a node name from a given node position
  - cnum – returns a node position from a given node name

What do I look for in students?

Attributes Leading to Success!

Personality & Attitude

- Adventurous
- Self starter
- Self learner
- Dedication
- Willing to work long hours
- Able to manage time
- Willing to fail (what!?)
- Work experience
- Responsible
- Mature personal and professional behavior

Academic

- Minimum of Sophomore standing
- CS major
- Above average GPA
- Extremely high faculty recommendations
- Good communication skills
- Two or more programming languages
- Data structures
- Software engineering

The ORNL Cluster Computing Experience…

Stephen L. Scott

Oak Ridge National Laboratory
Computer Science and Mathematics Division
Network and Cluster Computing Group

December 12, 2005
RAM Workshop
Oak Ridge, TN

scottsl@ornl.gov www.csm.ornl.gov/~sscott