

# Open Source Cluster Application Resources (OSCAR)

Presented by

Stephen L. Scott  
Thomas Naughton  
Geoffroy Vallée

Computer Science Research Group  
Computer Science and Mathematics Division



# Open Source Cluster Application Resources

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

- International consortium of academic, research, and industry members

# OSCAR



# OSCAR background

- **Concept first discussed in January 2000**
  - First organizational meeting in April 2000
  - Cluster assembly is time consuming and repetitive
  - Nice to offer a toolkit to automate
- **Leverage wealth of open source components**
- **First public release in April 2001**
- **Over six years of project development and six specialized versions**
- **Current Stable: oscar-5.0 ; Development: oscar-5.1/6.0**



# 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**



# OSCAR design goals

## Reduce overhead for cluster management

- Keep the interface simple
- Provide basic operations of cluster software and node administration
- Enable others to reuse and extend system—deployment tool

## Leverage “best practices” whenever possible

- Native package systems
- Existing distributions
- Management, system, and applications

## Extensibility for new software and projects

- Modular metapackage system/API—“OSCAR Packages”
- Keep it simple for package authors
- Open source to foster reuse and community participation
- Fosters “spin-offs” to reuse OSCAR framework

# OSCAR overview

## Framework for cluster management

- Simplifies installation, configuration, and operation
- Reduces time/learning curve for cluster build
  - Requires preinstalled head node with supported Linux distribution
  - Thereafter, wizard guides user through setup/install of entire cluster



## Package-based framework

- Content: Software + configuration, tests, docs
- Types:
  - Core: SIS, C3, Switcher, ODA, OPD, APItest, Support Libs
  - Non-core: Selected and third party (PVM, LAM/MPI, Toque/Maui, etc.)
- Access: Repositories accessible via OPD/OPDer

# OSCAR packages

- Simple way to wrap software & configuration
  - “Do you offer package Foo version X?”
- Basic design goals
  - Keep simple for package authors
  - Modular packaging (each self-contained)
  - Timely release/updates
- Leverage RPM + meta file + scripts, tests, docs, etc.
  - Recently extended to better support RPM, Debs, etc.
- Repositories for downloading via OPD/OPDer
- Leverage native package format via *opkgc*
  - OSCAR Packages compiled into native binary format

# OSCAR Packages (latest enhancements)

- **Maintain versatility and improve manageability**
  - High-level opkg description
  - Use 'opkgc' to convert to lower-level native binary pkg(s)
  - Manage binary opkgs via standard tools (rpm/yum, dpkg/apt)
- **Package repositories**
  - Local repos for restricted access (all via tarball)
  - Online repos for simplified access (opkgs via yum/apt)
- **Basis for future work**
  - Easier upgrades
  - Specialized OSCAR releases (reuse oscar-core with custom opkgs)



# OSCAR – cluster installation wizard

Welcome to the OSCAR Wizard!  
OSCAR Version: 5.0  
- INSTALL MODE -

Step 0:	Download Additional OSCAR Packages...	Help
Step 1:	Select OSCAR Packages To Install...	Help
Step 2:	Configure Selected OSCAR Packages...	Help
Step 3:	Install OSCAR Server Packages	Help
Step 4:	Build OSCAR Client Image...	Help
Step 5:	Define OSCAR Clients...	Help
Step 6:	Setup Networking...	Help
	Delete OSCAR Clients...	Help
	Monitor Cluster Deployment	Help

Before continuing, network boot all of your nodes. Once they have completed installation, reboot them from the hard drive. Once all the machines and their ethernet adapters are up, move on to the next step.

Step 7:	Complete Cluster Setup	Help
Step 8:	Test Cluster Setup	Help

Quit

Start

**Step 1**

**Step 2**

**Step 3**

**Step 4**

**Step 5**

**Step 6**

**Step 7**

**Step 8**

Done!

Cluster deployment monitor

MAC	IP	Hostname	Image	Kernel	Progress	Time	Speed	RAM	RV
00.03.47.C5.96.77	192.168.50.17	node17	oscartimage	2.6.16-boot_v3.7.4937.08	REBOOTED	3min	-	756MB	
00.03.47.C5.72.40	192.168.50.18	node18	oscartimage	2.6.16-boot_v3.7.4937.08	REBOOTED	3min	-	756MB	
00.03.47.C7.FF.A0	192.168.50.19	node19	oscartimage	2.6.16-boot_v3.7.4937.08	REBOOTED	3min	-	756MB	
00.03.47.C5.97.FE	192.168.50.2	node2	oscartimage	2.6.16-boot_v3.7.4937.08	REBOOTED	3min	-	756MB	
00.03.47.C5.9A.40	192.168.50.20	node20	oscartimage	2.6.16-boot_v3.7.4937.08	REBOOTED	3min	-	756MB	
00.03.47.C5.73.50	192.168.50.21	node21	oscartimage	2.6.16-boot_v3.7.4937.08	REBOOTED	3min	-	756MB	
00.03.47.C5.6C.22	192.168.50.22	node22	oscartimage	2.6.16-boot_v3.7.4937.08	REBOOTED	3min	-	756MB	
00.03.47.C5.6D.DC	192.168.50.23	node23	oscartimage	2.6.16-boot_v3.7.4937.08	REBOOTED	3min	-	756MB	
00.03.47.C5.71.78	192.168.50.24	node24	oscartimage	2.6.16-boot_v3.7.4937.08	REBOOTED	15min	-	756MB	
00.03.47.C7.FE.B0	192.168.50.25	node25	oscartimage	2.6.16-boot_v3.7.4937.08	REBOOTED	15min	-	756MB	
00.03.47.C7.FE.B8	192.168.50.26	node26	oscartimage	2.6.16-boot_v3.7.4937.08	REBOOTED	15min	-	756MB	

# OSCAR components

## Administration/ configuration

- System Installation Suite (SIS), Cluster Command & Control (C3), OPIUM, KernelPicker, and cluster services (dhcp, nfs, ntp, etc.)
- Security: Pfilter, OpenSSH

## HPC services/ tools

- Parallel libs: MPICH, LAM/MPI, PVM, Open MPI
- OpenPBS/MAUI, Torque, SGE
- HDF5
- Ganglia, Clumon
- Other third-party OSCAR Packages

## Core infrastructure/ management

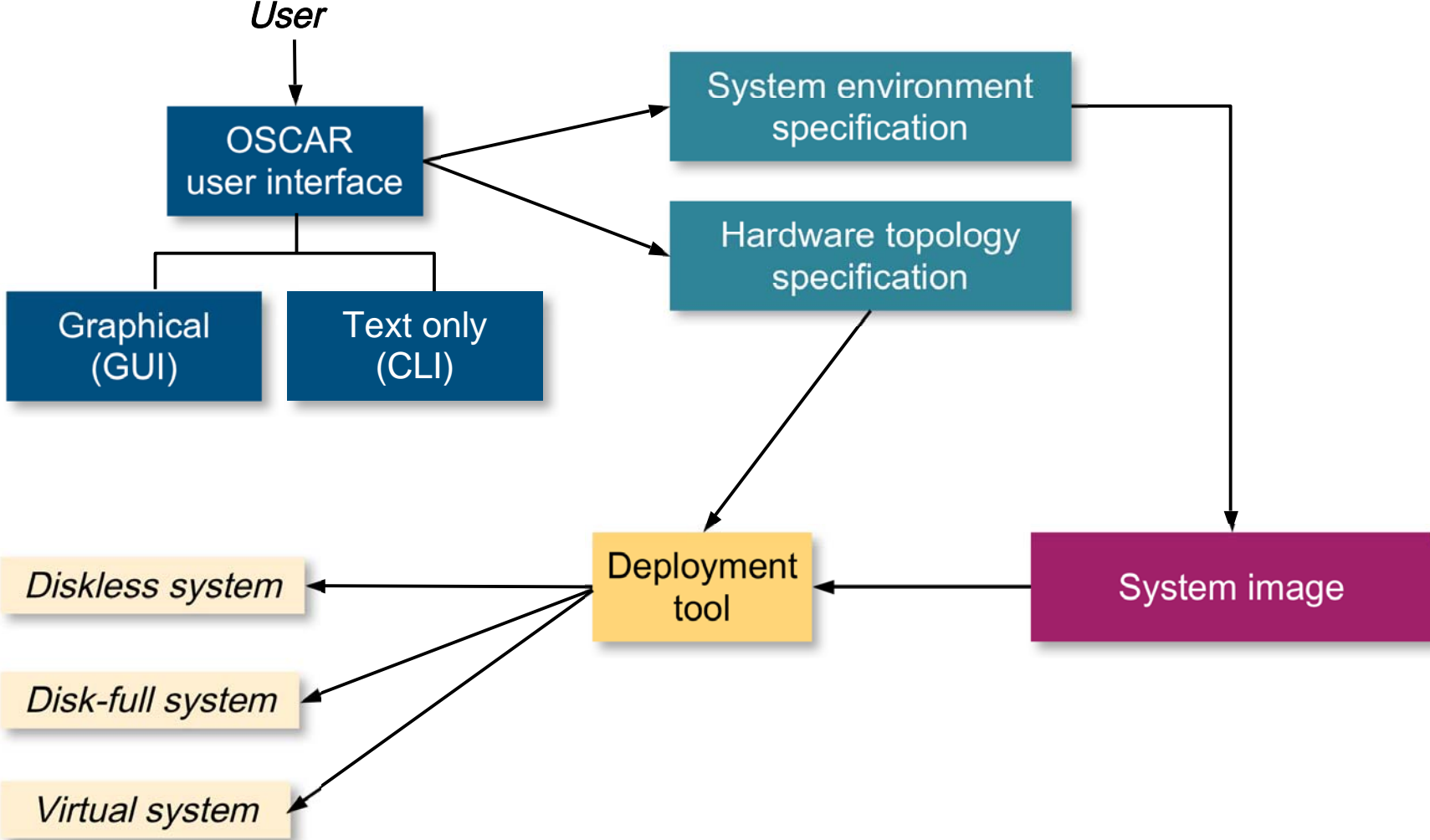
- SIS, C3, Env-Switcher
- OSCAR Database (ODA), OSCAR Package Downloader (OPD)
- OSCAR Package Compiler (OPKGC)

# 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
  - Mechanism for clusterwide operations



# OSCAR architecture



# Diskless OSCAR

- Extension of OSCAR to support diskless and diskfull nodes
- Ensures separation of node specific and shared data
- Current (2007) diskless OSCAR approach
  - Based on NFS-Root for node boot without local disk
  - Changes primarily isolated to System Installation Suite
  - In future will consider parallel filesystems (e.g., PVFS, Lustre)
- Modifies the initialization, *init*, of the compute nodes

# OSCAR

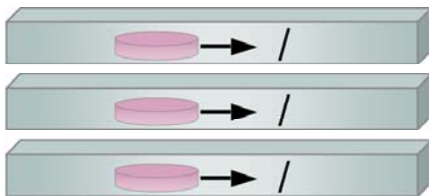
## Normal init (disk-full)

1. Mount/proc
2. Initialize the system
3. Run scripts at run level

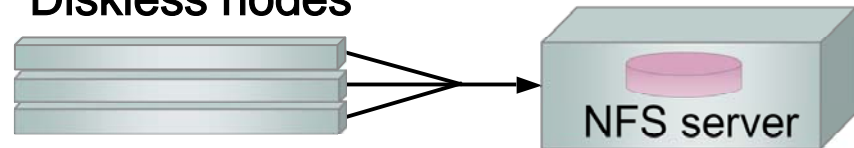
## Modified init (diskless)

1. Mount/proc
2. Start `rpc.lockd`, `portmap`
3. Mount NFS shares
4. Initialize the system
5. Run scripts at run level
6. `Rc.local` mounts hard disks and sends message back to head node

Local disk nodes



Diskless nodes



# OSCAR highlights

OSCAR  
Devel  
(v5.1/6.0)



In  
progress

- Local/remote repository installs
  - Command-line interface
  - Enhancements to OPD/OPDer
  - New OSCAR Package format
  - Targeted platforms:
    - Fedora, Red Hat EL, Debian, SuSE
    - x86, x86\_64
- 
- Diskless OSCAR
  - OPKG/node sets
  - New GUI
  - Google SoC'07: Benchmark, etc.
  - Enhanced native package installation
  - New KernelPicker2 (boot management tool)

# OSCAR: Proven scalability

<b>Top eight clusters by CPU count from registered list at OSCAR Web site</b>	OIC (ORNL)	526 nodes with 1052 CPUs
	Endeavor	232 nodes with 928 CPUs
	McKenzie	264 nodes with 528 CPUs
	SUN-CLUSTER	128 nodes with 512 CPUs
	Cacau	205 nodes with 410 CPUs
	Barossa	184 nodes with 368 CPUs
	Smalley	66 nodes with 264 CPUs
	PS9200-1-auguste	32 nodes with 256 CPUs

Based on data taken on 08/14/2007 from OSCAR Cluster Registration Page, [http://oscar.openclustergroup.org/cluster-register?sort=cpu\\_count](http://oscar.openclustergroup.org/cluster-register?sort=cpu_count).



# More OSCAR information...

Home page	<a href="http://oscar.OpenClusterGroup.org">oscar.OpenClusterGroup.org</a>
Development page	<a href="http://svn.oscar.openclustergroup.org/trac/oscar">svn.oscar.openclustergroup.org/trac/oscar</a>
Mailing lists	<a href="mailto:oscar-users@lists.sourceforge.net">oscar-users@lists.sourceforge.net</a> <a href="mailto:oscar-devel@lists.sourceforge.net">oscar-devel@lists.sourceforge.net</a>
Open cluster group	<a href="http://www.OpenClusterGroup.org">www.OpenClusterGroup.org</a>
OSCAR symposium	<a href="http://www.csm.ornl.gov/srt/oscar08">www.csm.ornl.gov/srt/oscar08</a>

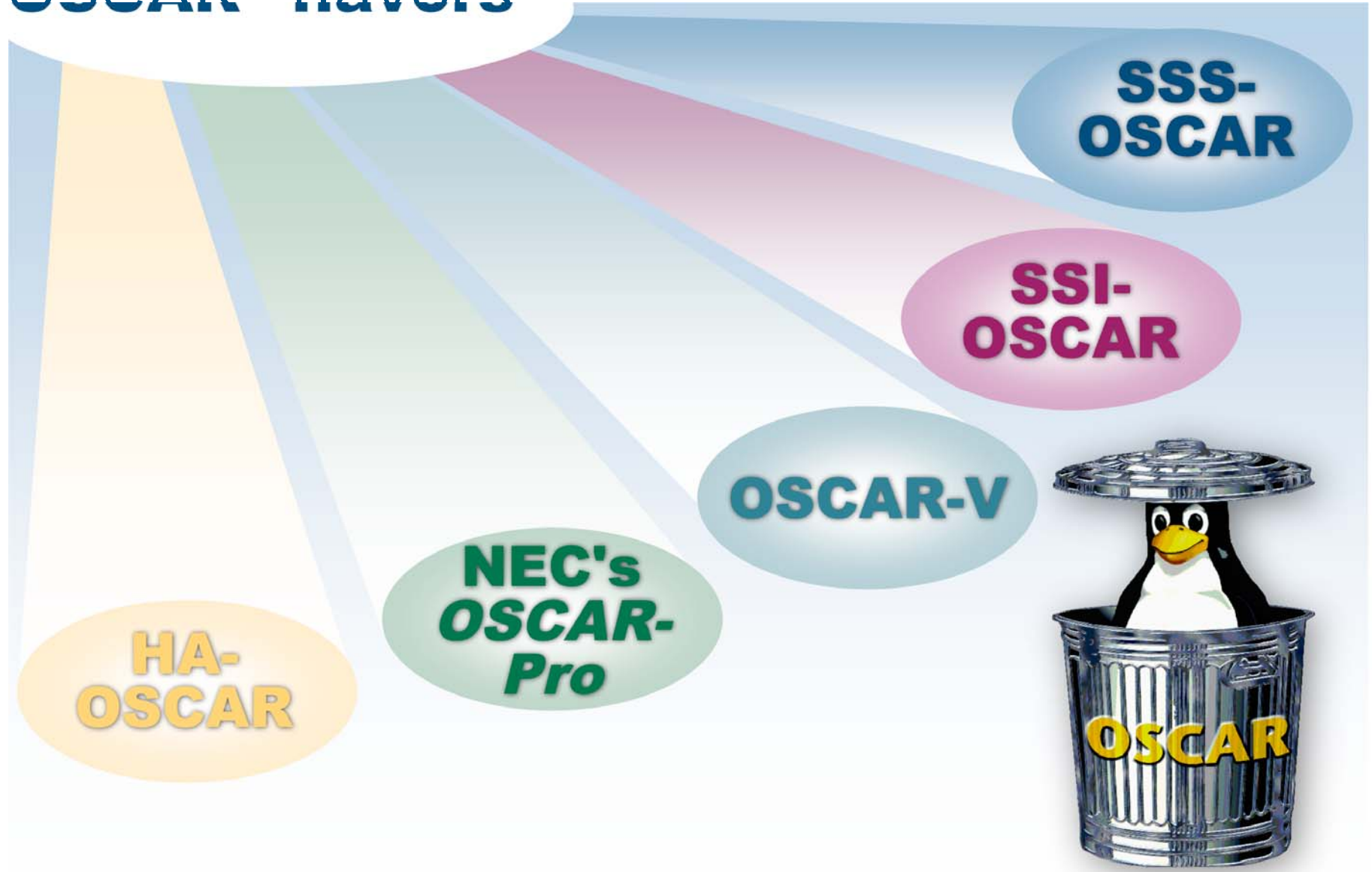
*OSCAR research supported by the*

*Mathematics, Information, and Computational Sciences Office,  
Office of Advanced Scientific Computing Research, Office of Science,  
U. S. Department of Energy, under contract no. DE-AC05-00OR22725 with UT-Battelle, LLC.*

# OSCAR



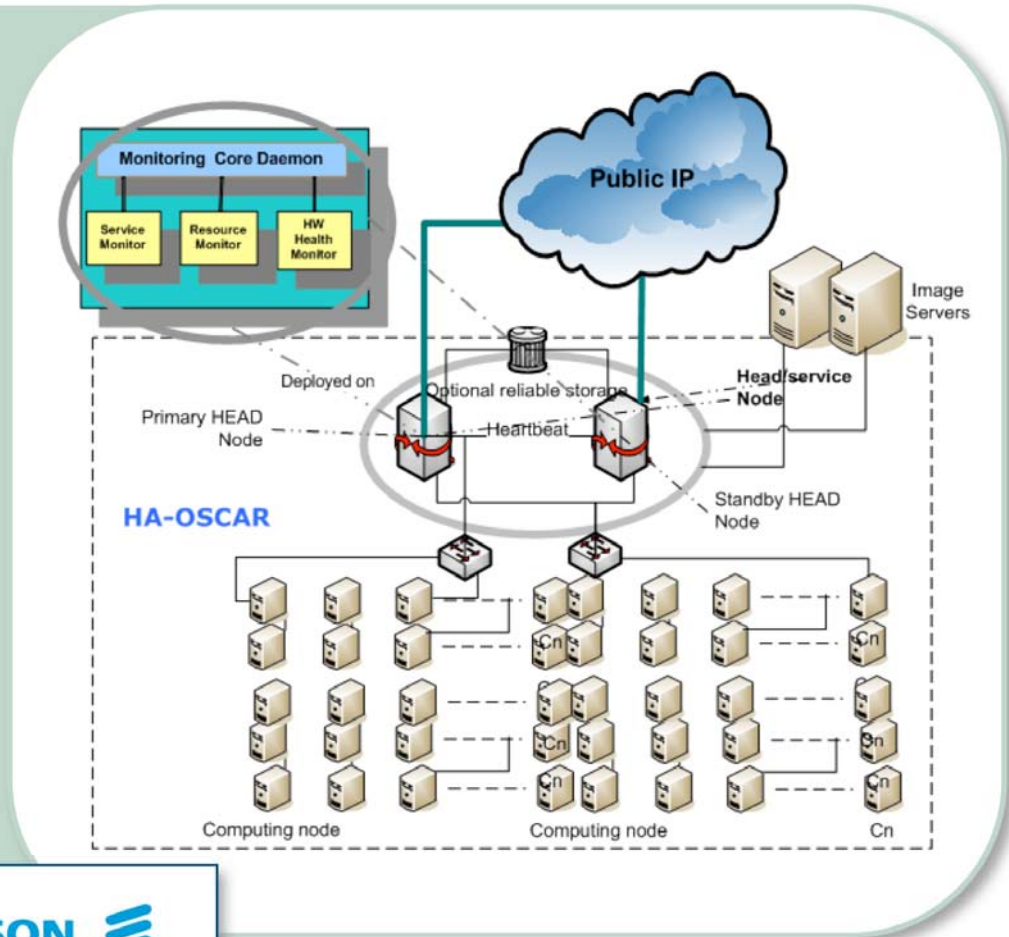
# OSCAR "flavors"



# HA-OSCAR

## RAS management for HPC cluster: Self-awareness

- The first known field-grade, open source HA Beowulf cluster release
- Self-configuration multihead Beowulf system
- HA and HPC clustering techniques to enable critical HPC infrastructure
- Services: Active/hot standby
- Self-healing with 3–5 s automatic failover time



# NEC's OSCAR-Pro

## Presented at OSCAR'06

OSCAR'06 keynote by  
Erich Focht (NEC)

- Leverage open source tool
- Joined project/contributions to OSCAR core

## Commercial enhancements

- Integrate additions when applicable
- Feedback and direction based on user needs

# OSCAR: Scalable systems 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
- Reduce costs and improve efficiency

## Participants

- DOE labs: ORNL, ANL, LBNL, PNNL, SNL, LANL, Ames
- Academics: NCSA, PSC, SDSC
- Industry: IBM, Cray, Intel, SGI



# SSS-OSCAR components

**Bamboo**

*Queue/job manager*

**BLCR**

*Berkeley checkpoint/restart*

**Gold**

*Accounting and 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*

# Single System Image – OSCAR (SSI-OSCAR)

- **Easy use thanks to SSI systems**
  - SMP illusion
  - High performance
  - Fault tolerance
- **Easy management thanks to OCSAR**
  - Automatic cluster install/update



# OSCAR-V

## Enhancements to support virtual clusters

- OSCAR-core modifications
- Create OSCAR Packages for virtualization solutions
- Integrate scripts for automatic installation and configuration

## Abstracts differences in virtualization solutions

- Must provide abstraction layer and tools—*libv3m/v2m*
- Enable easy switch between virtualization solutions
- High-level definition and management of VMs: Mem/cpu/etc., start/stop/pause



# OSCAR-V

6

Assign VMs to Host OSes

Welcome to the OSCAR-V Wizard!

Step 1: Install Host OSes...  
Step 2: Select OSCAR Packages To Install...  
Step 3: Build Image for Virtual Compute Nodes...  
Step 4: Define a New Virtual Compute Nodes...  
Step 5: Assign MAC Addresses to Virtual Compute Nodes...  
Step 6: Assign Virtual Compute Nodes to Host OSes...

Virtual Cluster Deployment

- Import IPs of Host OSes from
- Assign all Host OSes
- Assign Virtual Machine to Host OSes

Deploy the Virtual Cluster

Setup the Virtual Cluster

Close

2

OPKG selection for VMs

Welcome to the OSCAR Wizard!

OSCAR Version: 5.0  
- INSTALL MODE -

Step 0: Download Additional OSCAR Packages...  
Step 1: Select OSCAR Packages To Install...  
Step 2: Configure Selected OSCAR Packages...  
Step 3: Install OSCAR Server Packages...  
Step 4: Build OSCAR Client Image...  
Step 5: Define OSCAR Clients...  
Step 6: Setup Networking...  
Step 7: Delete OSCAR Clients...  
Step 8: Monitor Cluster Deployment

1

Host OS installation

MAC Address Management

Not Listening to Network. Click "Start Collecting MACs" to start.

MAC Address	Machine
eth0 mac = 160.91.44.252	-oscardomain
eth0 ip = 160.91.44.252	-oscardomain
eth0 mac = 160.91.44.252	-voscardome9.oscardomain
eth0 ip = 160.91.44.252	-voscardome9.oscardomain
eth0 mac = 00:16:3E:7D:08:D3	-voscardome81.oscardomain
eth0 ip = 10.0.0.13	-voscardome81.oscardomain

Remove Remove All

OSCAR Package Selector

Package Set: Default

Package Name	Class	Location/Version
<input checked="" type="checkbox"/> netbootmgr	base	OSCAR 0.8-1
<input checked="" type="checkbox"/> apitest	core	OSCAR 1.0-12
<input checked="" type="checkbox"/> base	core	OSCAR 1.0-1
<input checked="" type="checkbox"/> c3	core	OSCAR 4.0.1-5
<input checked="" type="checkbox"/> oda	core	OSCAR 1.31-1
<input checked="" type="checkbox"/> rapt	core	OSCAR 1.0-0
<input checked="" type="checkbox"/> sc3	core	OSCAR 1.1-5

5

Definition of VMs' MAC addresses

MAC Address Management

Start Collecting MACs Assign all MACs Assign MAC to Node

Delete MAC from Node Import MACs from Export MACs to file...

Installation Mode and DHCP Setup

systemimager-rsync Enable Install Mode

Dynamic DHCP update Configure DHCP Server

Boot Environment (CD or PXE-boot) Setup

Enable UYOK Build AutoInstall CD... Setup Network Boot

Close

3

Image creation for VMs

Define OSCAR Clients

Image Name: hostosimage  
oscardomain  
oscardomain  
oscardomain

Number of Hosts: 0

Starting IP: 160.91.44.253

Subnet Mask: 255.255.255.0

Default Gateway: 10.0.0.1

Build OSCAR Client Image

Fill out the following fields to build a System Installation Suite image. If you need help on any field, click the help button next to it

Image Name:	oscardomain	Help
Package File:	/opt/oscar/oscarsamples/	Choose a File... Help
Target Distribution:	centos-4-x86_64	Help
Package Repositories:	/ftpboot/oscar/common-ftp	Help
Disk Partition File:	/opt/oscar/oscarsamples/	Choose a File... Help
IP Assignment Method:	static	Help
Post Install Action:	reboot	Help

Reset Build Image Close

4

Definition of virtual compute nodes



# OSCAR-V: Description of steps

## Initial setup

1. Install supported distro head node (host)
2. Download/set up OSCAR and OSCAR-V
  - OSCAR: untar oscar-common, oscar-base, etc., and copy distro RPMs
  - OSCAR: untar; run “make install”
3. Start Install Wizard
  - run “./oscarv \$network\_interface” and follow setups

# OSCAR-V: Summary

- **Capability to create image for Host OSes**
  - Minimal image
  - Take benefit of OSCAR features for the deployment
  - Automatic configuration of system-level virtualization solutions
  - Complete networking tools for virtualization solutions
- **Capability to create images for VMs**
  - May be based on any OSCAR-supported distribution: Mandriva, SuSE, Debian, Fedora, Red Hat EL, etc.
  - Leverage the default OSCAR configuration for compute nodes
- **Resources**
  - V2M/libv3m: <http://www.csm.ornl.gov/srt/v2m.html>
  - OSCAR-V: <http://www.csm.ornl.gov/srt/oscarv.html>
  - OSCAR: <http://oscar.openclustergroup.org>

# Contacts regarding OSCAR

Stephen L. Scott

Computer Science Research Group  
Computer Science and Mathematics Division  
(865) 574-3144  
scottsl@ornl.gov

Thomas Naughton

Computer Science Research Group  
Computer Science and Mathematics Division  
(865) 576-4184  
naughtont@ornl.gov

Geoffroy Vallée

Computer Science Research Group  
Computer Science and Mathematics Division  
(865) 574-3152  
valleegr@ornl.gov