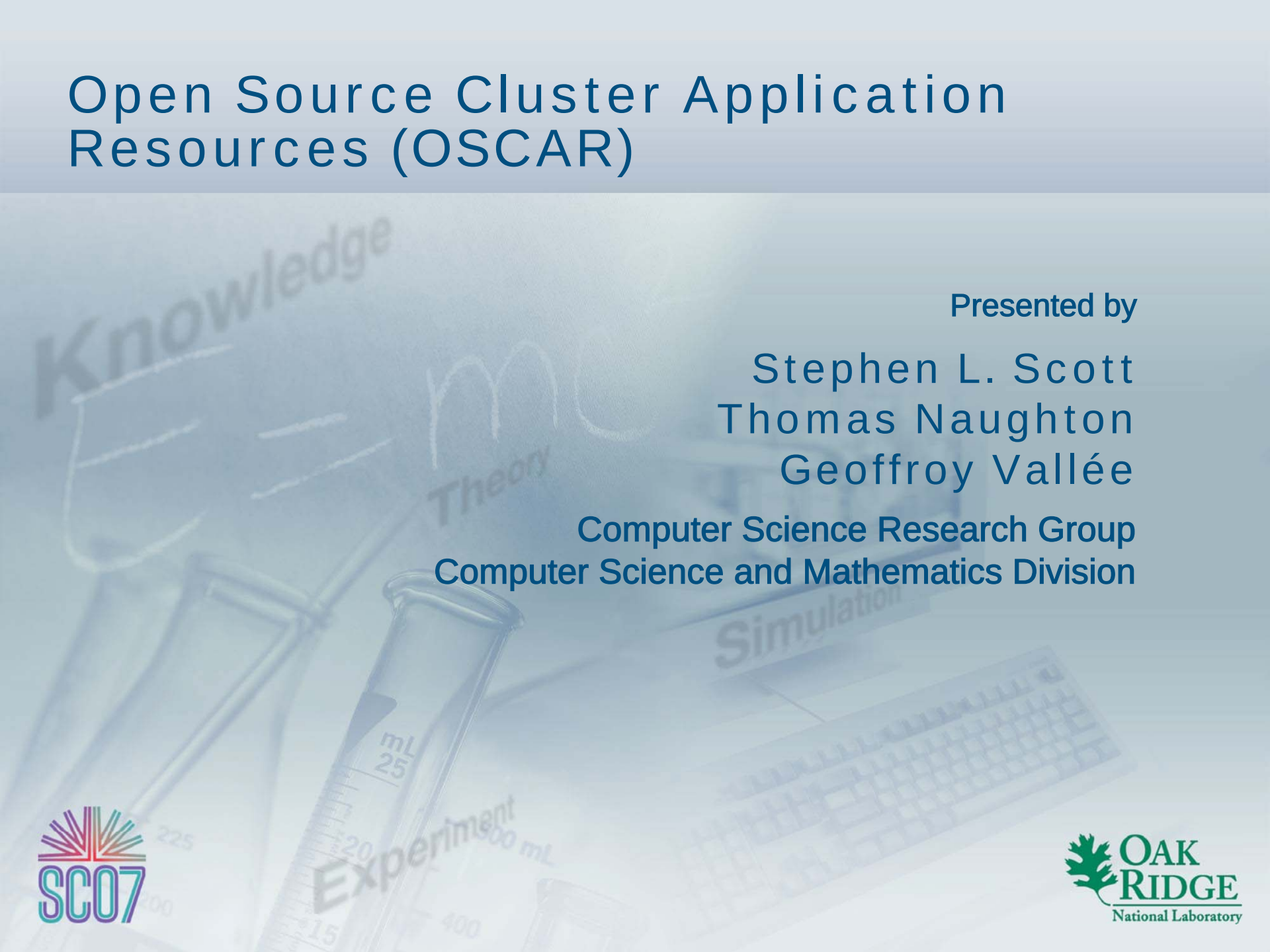
Open Source Cluster Application Resources (OSCAR)

Presented by

Stephen L. Scott
Thomas Naughton
Geoffroy Vallée

Computer Science Research Group
Computer Science and Mathematics Division

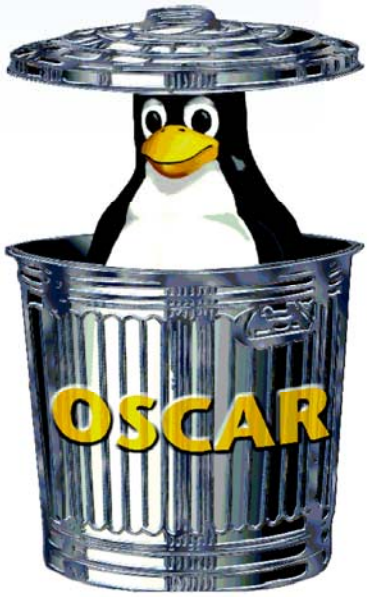


SC07

OAK RIDGE National Laboratory

Open Source Cluster Application Resources

- Snapshot of best known methods for building, programming, and using clusters
- International consortium of academic, research, and industry members



INRIA

OAK RIDGE National Laboratory

Bald Guy SOFTWARE

intel

LOUISIANA TECH UNIVERSITY

pervasive technology labs AT INDIANA UNIVERSITY

NEC

OAK RIDGE National Laboratory

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	Leverage "best practices" whenever possible	Extensibility for new software and projects
• Keep the interface simple • Provide basic operations of cluster software and node administration • Enable others to reuse and extend system—deployment tool	• Native package systems • Existing distributions • Management, system, and applications	• 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

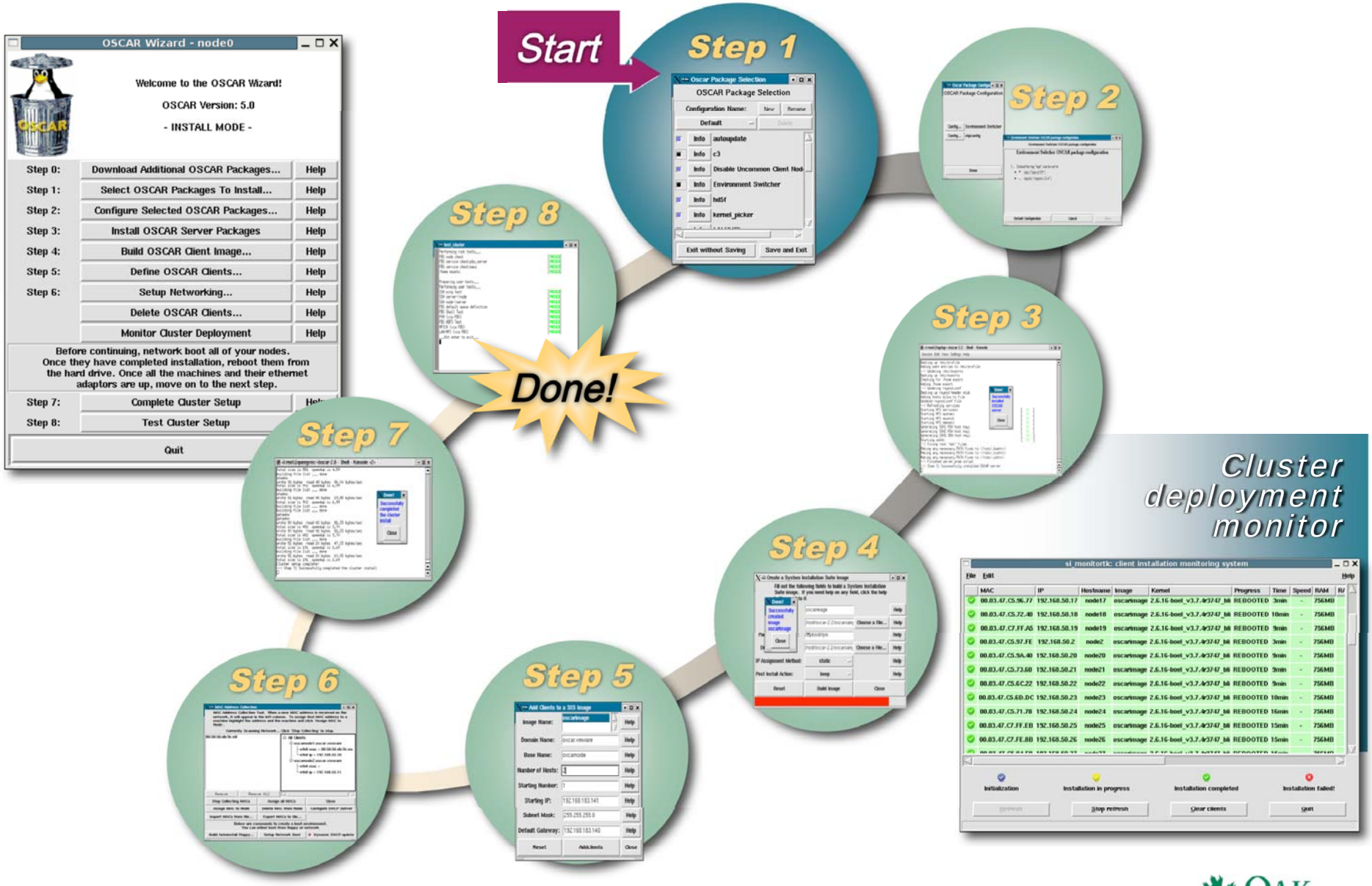
OSCAR packages

- 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 versatilty 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



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)

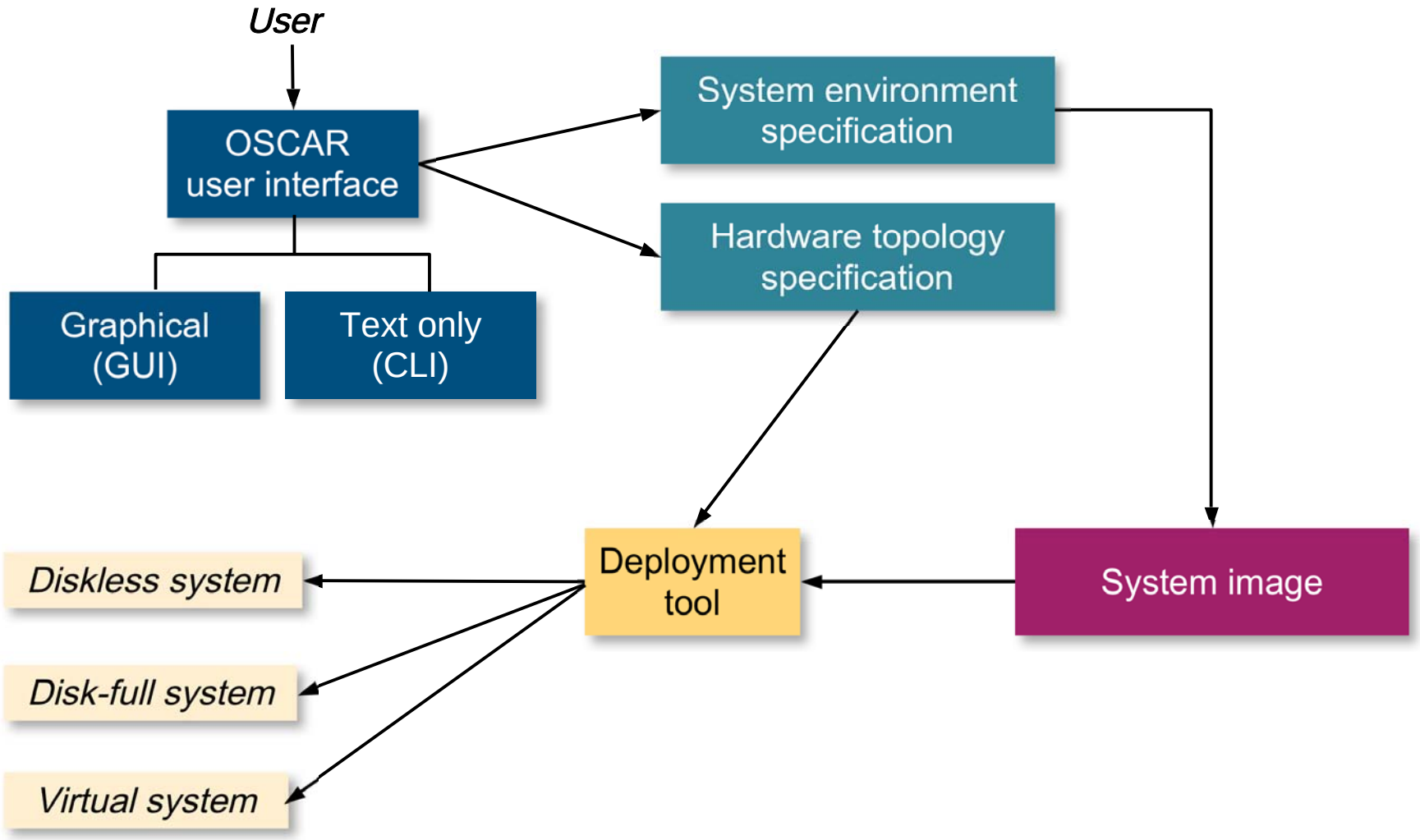
OSCAR: C3 power tools

- 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

OAK RIDGE National Laboratory

OAK RIDGE National Laboratory

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

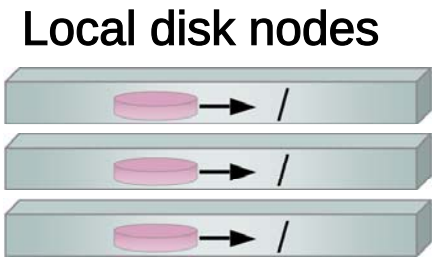
OSCAR

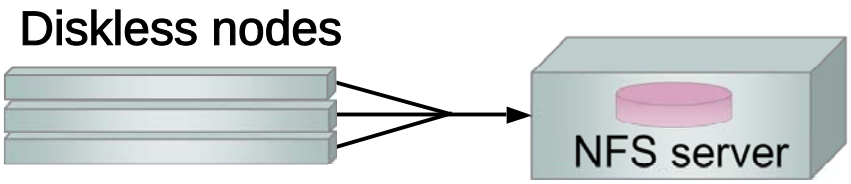
Normal init (disk-full)

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





OSCAR highlights

- 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

Top eight clusters by CPU count from registered list at OSCAR Web site	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.

More OSCAR information…

Home page	oscar.OpenClusterGroup.org
Development page	svn.oscar.openclustergroup.org/trac/oscar
Mailing lists	oscar-users@lists.sourceforge.net oscar-devel@lists.sourceforge.net
Open cluster group	www.OpenClusterGroup.org
OSCAR symposium	www.csm.ornl.gov/srt/oscar08

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 "flavors"

SSS-OSCAR

SSI-OSCAR

OSCAR-V

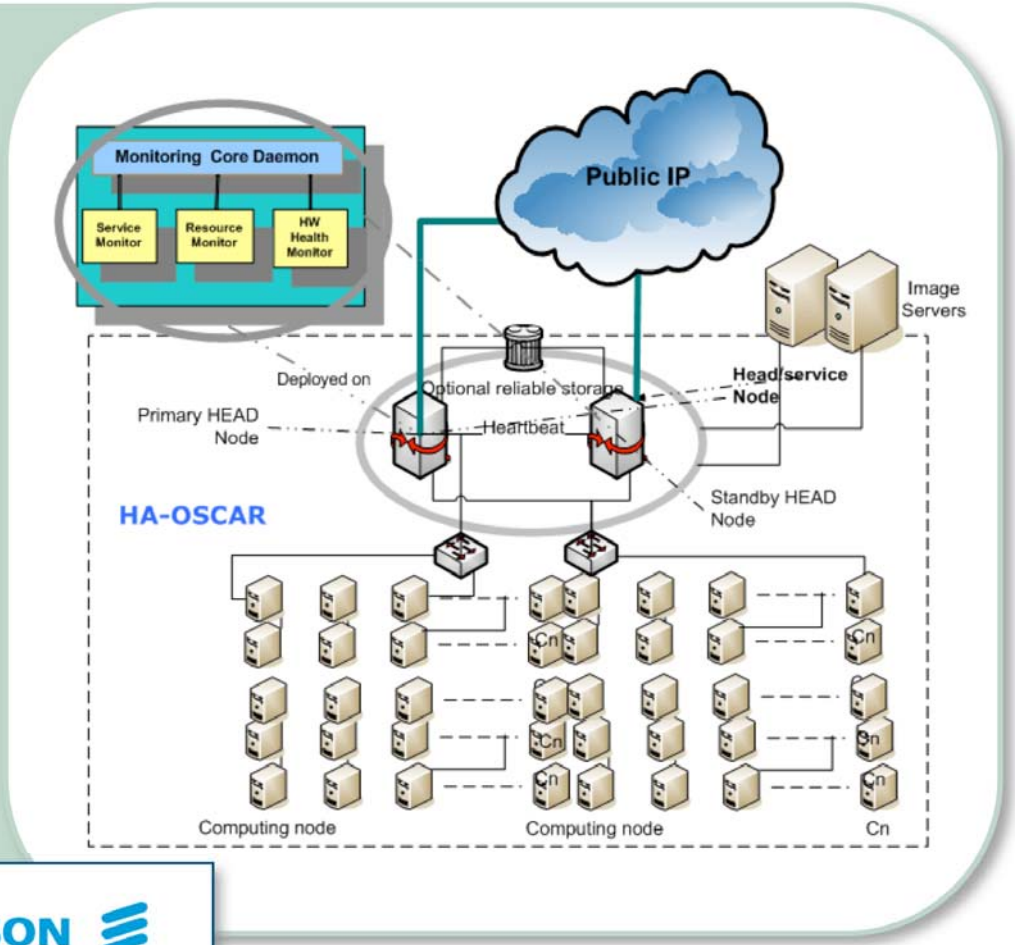
NEC's *OSCAR-Pro*

HA-OSCAR

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



LOUISIANA TECH UNIVERSITY intel ERICSSON

OAK RIDGE National Laboratory

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

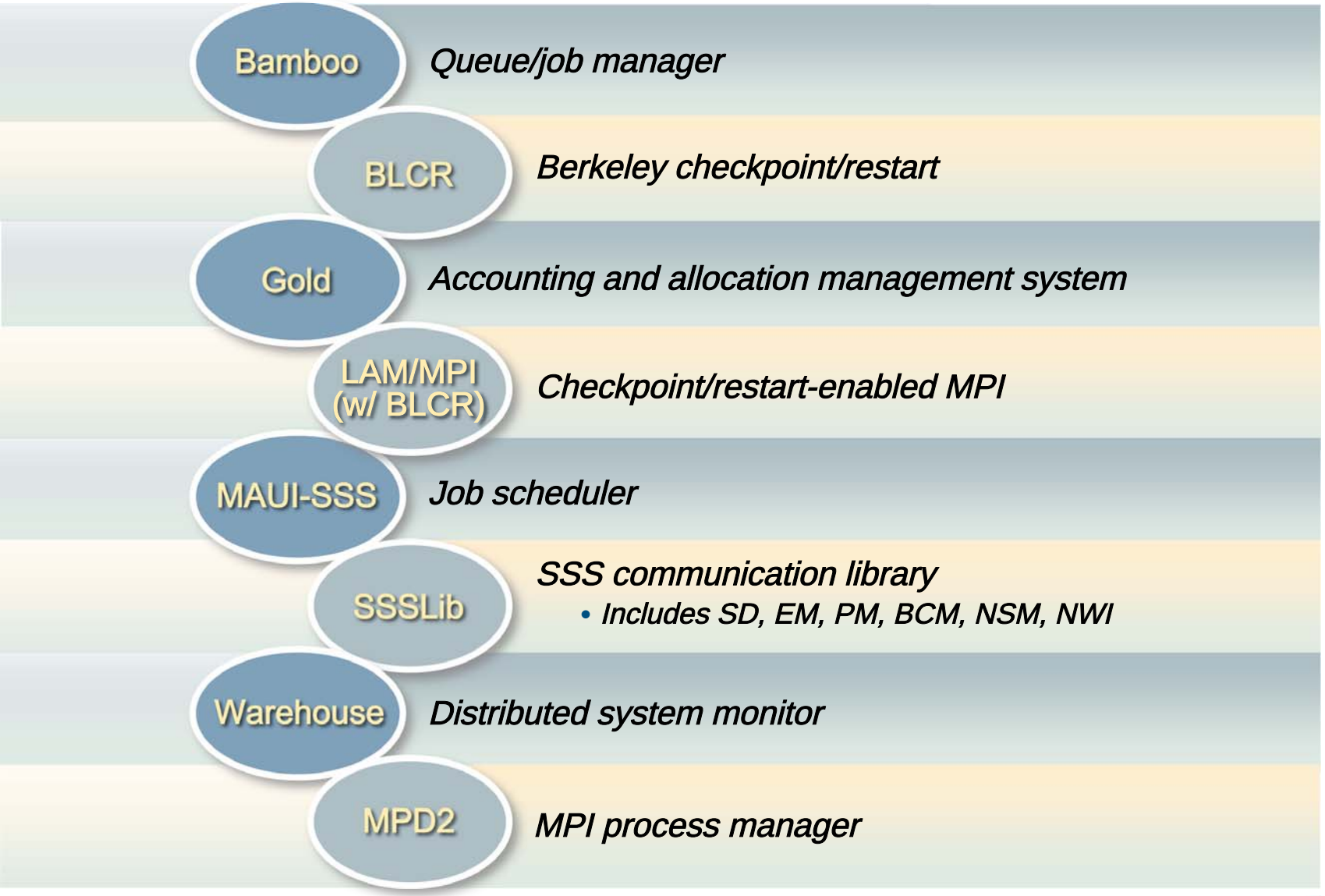
Participants

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

Office of Science
U.S. DEPARTMENT OF ENERGY

OAK RIDGE National Laboratory

SSS-OSCAR components



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

OAK RIDGE National Laboratory

OSCAR-V

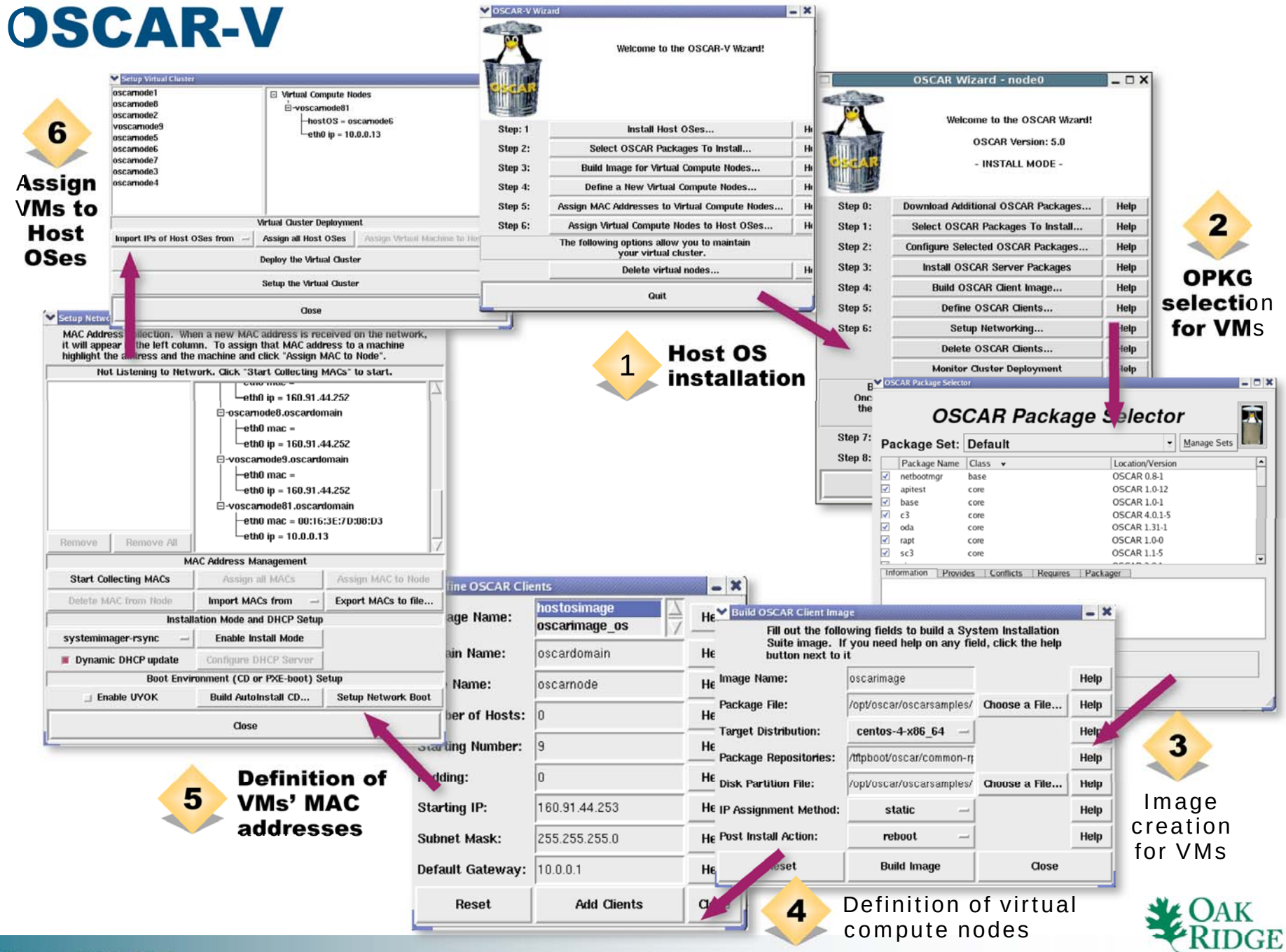
Enhancements to support virtual clusters

- 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



OAK RIDGE National Laboratory

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