

# OSCAR and the Beowulf Arms Race for the "Cluster Standard"

Stephen L. Scott

Cluster 2001 - October 9, 2001





# What is OSCAR?

---

- Open Source Cluster Application Resources
  - Originally known as: Community Cluster Development Kit (CCDK)
- Cluster on a CD...
- “best practices” for cluster computing
- Built by consortium
  - aka – by committee – yikes!



# Cluster “industry” Problems

---

- Cluster Standard vs Standard Cluster
  - One size does not fit all
  - One type does not work for all
  - There should be a unified view for all
- Stumbling block to commercialization
  - I want the software for free
    - Oh, and I want it to be of commercial quality
    - Did I mention support – I’d like that free too...
  - Open source does not mean free
    - What are you willing to pay?



- Extreme Linux
- May 13, 1998
- \$29.95 CD

redhat.com | Announcing Extreme Linux

redhat

SEARCH RED HAT:  Go

Products and Services | Red Hat Marketplace | Partner Programs | Support and Docs | Training | About Red Hat

### About Red Hat

[Corporate Information](#) | [Serving Our Community](#) | [Careers](#) | [Press Center](#) | [Investor Relations](#) | [Events & Presentations](#) | [Contact Red Hat](#) | [Success Stories](#)

#### Red Hat Press Releases

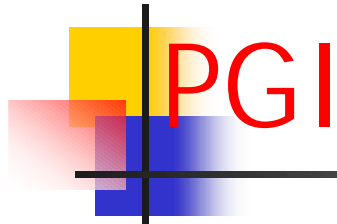
## Announcing Extreme Linux

Called Extreme Linux, and also known as The Beowulf Project, this collaboration between Red Hat, Inc. NASA Goddard Space Flight Center, and over two dozen leading research centers will bring the speed and power of multiple computers—parallel processing as one computer—to students, researchers, and end-users worldwide. Extreme Linux is perhaps the most effective example of how the cooperative software development model that has produced the award-winning Linux OS results in better technology at a revolutionarily low-cost.

Building a computer cluster with the OS and tools that are included in this \$29.95 CD-ROM product will provide researchers with radical improvements in the amount of processing power available to them for a given dollar of investment. Having access to complete source code of these tools will allow the students, researchers, and technical end users to understand this technology at a level never before possible, resulting in a more effective, higher performance computing platform.

For more information, check [here](#).

For More Information:  
Bryan Scanlon or Dan Ring  
Schwartz Communications for Red Hat  
Phone: (781) 684-0770  
[redhat@schwartz-pr.com](mailto:redhat@schwartz-pr.com)



- Cluster Development Kit (CDK)
- Includes “floating” seat compiler licenses

**The Portland Group**

Home | About PGI | Support | Documentation | News | Jobs | Y2K

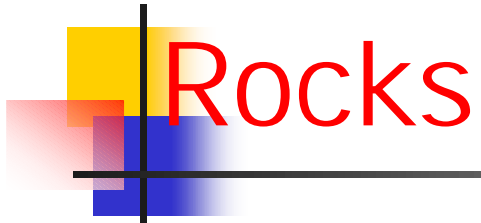
**PGI® CDK™**  
**Cluster Development Kit™**  
**Software for Linux**

If you're building a Linux cluster, we've put everything but the hardware in one convenient package called the *PGI CDK Cluster Development Kit*. Linux clusters are now easily affordable by any organization interested in serious scientific computing. Getting your Linux hardware installed and your network up and running is just the first step in building a cluster. The *PGI CDK* contains the software development and cluster management tools you need to get a Linux cluster running production applications:

- Floating multi-user seats for PGI's parallel Fortran, C, and C++ compilers for Linux -- industry-leading single-processor performance and integrated native support for all 3 popular parallel programming models: HPF, OpenMP, and MPI.
- Graphical debugging and parallel performance profiling tools
- Pre-compiled/pre-configured MPI-CH message-passing libraries and utilities
- Pre-compiled/pre-configured PBS batch queueing system to manage the workload and throughput on your cluster
- Pre-compiled PVM Message-passing library
- Pre-compiled ScaLAPACK parallel math library
- Optimized BLAS and LAPACK serial math libraries
- Tutorial examples and programs to help you get your codes up and running quickly using HPF, OpenMP, and MPI messaging
- Installation utilities to simplify the setup and management of your Linux cluster

Whether you have legacy FORTRAN 77 that relies on Cray, DEC, or IBM extensions, or are writing modern parallel codes in Fortran 90, High Performance Fortran (HPF), or C/C++ the *PGI CDK* has all the features you need. PGI's compilers are highly optimized for Pentium III/III workstations and servers running Linux. You'll have the option of parallelizing your applications using MPI, or using explicit [HPF](#) or [OpenMP](#) directives. PGI's industry-leading performance - including a [30% performance improvement](#) over g77/egcs on SPECfp\_base85 - allows you to build high performance applications for single, dual, or quad-processor workstations which can then run unchanged on workstation clusters, shared-memory servers, or high-end distributed-memory or NUMA supercomputers.

In fact, PGI's *PGHPF*® data parallel Fortran compiler for high performance computing systems has been adopted as the



- Cluster distribution
- Cluster tools
- RedHat / Kickstart  
/ RPM based

## Rocks Clustering Toolkit

[Home](#)  
[News](#)  
[Faq](#)  
[Cluster Overview](#)  
[Getting Started](#)  
[System Administration](#)  
[Real Rocks Clusters](#)  
[Cluster Software](#)  
[Downloads](#)  
[How To Contribute](#)  
[Collaborators](#)  
[Talks / Papers / Press](#)  
[Benchmarks](#)

**Current Release: [2.1](#)**

### Welcome to the NPACI Rocks Clusters Website!

The NPACI Rocks clustering activity is a collection of open-source software tools, software, management techniques, monitoring infrastructure and other good things for building commodity off-the-shelf clusters. We're working to make "cookie-cutter" clusters that groups can easily replicate, manage, and update.

This site is a gathering point of information and software that pertain directly to clusters and includes tools and ideas from people across [NPACI](#) (and beyond). If you've got cluster-specific tools that should be included in this toolkit then [contact us](#). Rocks is a community site.

Rocks builds with a standard [RedHat Linux](#) distribution so the software can stay current without managing a complete distribution. All software is packaged as an [RPM](#), to greatly simplify installation and automation. We use RedHat's kickstart installation method to automate builds of front-ends and cluster nodes.

Rocks is a "complete" clustering distribution as well as a toolset collection. You can pick and choose the components that you want for your cluster. We espouse a particular [management strategy](#) for clusters, but you don't have to lock into "our way" if you prefer something else.

Take a look at our [cluster software](#) for a more detailed description of the current component set or go directly to [software downloads](#) to get the packages that you want. You will see things like reinstalling your cluster in 10 minutes, remote installation *without* KVM (keyboard/video/mouse) switches or serial concentrators, scalable and secure remote execution, cluster monitoring, batch schedulers, MPI, PVM, PBS, Globus and other great tools.

Rock On!

# SCore



## Parallel and Distributed System Software Laboratory, **RWCP**

### ■ Claims NOT a beowulf "type" cluster

- PMv2 communication library
- Not TCP/IP stack based

#### Japanese

What's New  
updated on  
07/06/2001

Clustering Technologies

SCore Users Group

Software Distribution

SCore Cluster System Software Documents

SCore Tips

SCore Users

Benchmark results of RWC Clusters

Collaborators' Only

Published Papers

Presentation Slides

Technical Reports

RWC SCore is NOT a beowulf type cluster in the sense that the SCore cluster system software is designed for the high performance cluster environment without using the TCP/IP protocol stack. The following are the distinguishing SCore features:

#### 1. High Performance Communication

The PMv2 high performance communication library is dedicated to cluster computing. PMv2 supports 10/100/1000 Ethernet links and Myrinet. It is also available on SMP using the shared memory facility. MPI on top of PMv2, called MPICH-SCore, has been developed based on MPICH. The MPICH-SCore performance is shown below. It was measured using RWC SCore II cluster which is 64 NEC Express Servers (Dual Pentium-III's whose clock is 800 MHz and 512MByte Main Memory).

Network	Bandwidth (MByte/sec)	RTT Latency (microsecond)
Myrinet Lanal9 in 33MHz PCI64	146.9	20
Gigabit Ethernet(Sysconnect)	73.4	61
Fast Ethernet(EEPRO100)	11.9	100
Fast Ethernet x 2(EEPRO100)	23.9	105
Fast Ethernet x 3(EEPRO100)	30.7	111

#### Legend:

RTT Latency means that round trip latency.

Fast Ethernet x 2 means that two Fast Ethernet Links are used. Fast Ethernet x 3 means that three Fast Ethernet Links are used. PMv2 for Ethernet utilizes Ethernet links, whose functionality is called *Network Trunking*.



# Scyld

- Unique and Cool
- Built to cluster
- Single point administration
- Load and run


**SCYLD COMPUTING CORPORATION™**

[HOME](#)
[PRODUCTS](#)
[SUPPORT](#)
[VENDORS](#)
[ABOUT SCYLD](#)
[SEARCH](#)

**Products**  
 Professional Edition  
 Enterprise Solutions  
 Basic Edition  
 Training  
 Overview

**Support**  
 Documentation  
 FAQs  
 Network Drivers

**Vendors**

**About Scyld**  
 Press Releases  
 Employment  
 Contact

**Search**

**Scyld Beowulf Cluster Operating System**  
  
**Latest Professional Release Now Available!**

- Second Generation Beowulf Clustering
- The original Beowulf team
- A stable platform for turn-key commercial applications
- Setting the standard for cluster computing



**News and Announcements**

- Important News! Latest Professional Product Release includes advanced software with new features, full documentation, and full support. Read more [here](#).
- Updated Basic Edition also available for \$249 [Click here](#) for more information.
- Penguin Computing and Scyld Computing Partner to Offer High Performance Computing Solutions [Read more](#).
- Clemson University's newly procured 520 processor cluster to run Scyld Beowulf.
- Full professional training programs at all levels of expertise now available for Beowulf clustering. [Click here](#) for more information.
- Compaq announces new linux initiatives including Linux Beowulf clusters on industry-standard ProLiant servers using commercial Scyld Beowulf. [Click here](#) for more information.



© 2001 Scyld Computing Corporation



# EnFuzion

- Same job everywhere model

PRODUCTS

SERVICES

SOLUTIONS

PARTNERS

NEWS

ABOUT US

**turbolinux** Powerful Thinking.



SUPPORT  
DOWNLOAD  
SECURITY CENTER

DEVELOPER ZONE  
PARTNER ACCESS  
CLUB TURBO

[VISIT TURBOLINUX WORLDWIDE](#)

[PowerCockpit Workstation 7.0](#)  
[Server Cluster Server](#)  
[EnFuzion 7.0](#)  
[EnFuzion DataServer with IBM DB2](#)  
[DataServer for Oracle® Series](#)  
[pSeries S/390](#)

[Turbolinux in the News](#)  
[Product Registration](#)  
[Product Info](#)  
[Why to Buy](#)  
[White papers](#)  
[Product Reviews](#)

[License Agreement](#)  
[Hardware Compatibility](#)



**Turbolinux EnFuzion 7.0**  

EnFuzion clusters all available computing resources on a corporate network to create a powerful "virtual supercomputer" and, as a result, allows companies to reduce time and costs associated with computationally demanding data processing jobs. Traditionally, th jobs - such as complex financial calculations - have been handled expensive high-end servers. With the growing need to process increasing volumes of complex jobs in a shorter time period, the of traditional solutions becomes prohibitive.

[Data Sheet](#) <sup>(PDF)</sup> | [Buy Offline](#)

Copyright © 2001 Turbolinux Inc. All Rights Reserved

[CONTACT](#) [PRIVACY STATEMENT](#)

# PowerCockpit

- Cool GUI
- Image based system
- Global commands

PRODUCTS SERVICES SOLUTIONS PARTNERS NEWS ABOUT US

**turbolinux.** Powerful Thinking

POWERCOCKPIT DEMO - REALPLAYER  
POWERCOCKPIT DEMO - WINDOWS MINI

**TurboLinux PowerCockpit 1.0**

PowerCockpit combines server management, rapid deployment and customization of racks of Linux servers in an easy to use, comprehensive tool. With PowerCockpit you can convert an entire rack of raw hardware to fully configured Linux servers in minutes. You can quickly collect complete images from existing Linux servers and organize a repository of images that can be rapidly deployed onto new or existing servers.

PowerCockpit reduces costs by simultaneously deploying multiple servers without hours of repetitive work and without an expensive KVM switch. PowerCockpit can standardize your enterprise by deploying preconfigured file/print, e-mail, caching proxy, intranet web, or firewall servers worldwide. You can pre-load new servers with an image from the repository, quickly delivering build-to-order solutions. With PowerCockpit you can also recover existing servers painlessly, without complicated tape archives.

**PowerCockpit 1.0 Features:**

**Interface**

- Uniquely powerful GUI for interactive use.
- Perl script interfaces.
- Scalable control: operate on one server or racks simultaneously.
- Online context sensitive documentation.

**Deployment**

- Collect and deploy images.
- Manage disk layouts.
- Store images in a repository.

**Management**

- Organize servers in overlapping or hierarchical sets.
- Perform remote collective commands.
- Send and collect file sets.
- Install and manage RPM sets.
- Monitor server health with heartbeats.



# CPlant

- Emphasis on scalability
  - Of running codes
  - Not of building...
- Help me!!!

The Computational Plant project at [Sandia National Laboratories](#) is developing a large-scale, massively parallel computing resource from a cluster of commodity computing and networking components. We are combining the knowledge and research of previous and ongoing commodity cluster projects with our expertise in designing, developing, using, and maintaining large-scale MPP machines. Our goal is to provide a commodity-based, large-scale computing resource that meets the level of compute performance needed by Sandia's critical applications.





# Unlimited Scale

[home](#) [products](#) [news](#) [careers](#) [about](#) [contact](#)

## products

### Unlimited Linux - The Unlimited Scale Product

Clustering is revolutionizing high-performance computing by providing unprecedented performance at commodity prices. Linux-based clustering offers the potential of a large platform-independent application catalog as well as a significant core technology base that can be leveraged to reduce the cost of providing the sophisticated environment required by demanding user communities. Building on the strengths of Linux, Unlimited Linux with its resource management, scheduling and system administration capabilities will enable customers to effectively manage clusters of tens to thousands of commodity servers as a single unified system.

The initial product release, based on Sandia National Laboratories' Cplant™ clustering software environment, will be available early next year. Unlimited Linux will be available initially supporting Compaq Alphaserver Platforms connected via Myrinet. Versions supporting Intel-based servers and other types of

- We will help you!
- When?
- Will it support a range of system sizes?
  - aka – small user vs big user problem



- Migration is cool
  - Performance hit
- Longevity award

## MOSIX Scalable Cluster Computing for Linux

MOSIX home page

[http://www.mosix.org/cgi\\_main.html](http://www.mosix.org/cgi_main.html)

MOSIX is a software that enhances the Linux kernel with cluster computing capabilities. It allows any size cluster of X86/Pentium/AMD based workstations and servers to work cooperatively as if part of a single system.

To run in a MOSIX cluster, there is no need to modify applications or to link with any library, or even to assign processes to different nodes. MOSIX does it automatically - just "[fork and forget](#)", like in an SMP. For example, you can create many processes in your (login) node and let MOSIX assign these processes to other nodes. If you type "ps", then you will see all your processes, as if they run in your node.

The core of MOSIX are adaptive management algorithms that monitor and respond to uneven resource distribution among the nodes. These algorithms use preemptive process migration to assign and reassign processes among the nodes, to continuously take advantage of the best available resources. The MOSIX algorithms are geared for maximal overall performance, overhead-free scalability and ease-of-use.

Because MOSIX is implemented in the Linux kernel, its operations are completely transparent to the applications. It can be used to define different [cluster types](#), even a cluster with different CPU or LAN speeds, like our 72 processors cluster:



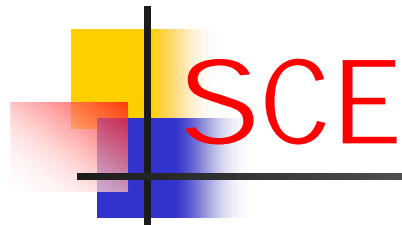
(click on image for more details).

---

To operate your Linux cluster like a single system [download and install the latest release of MOSIX](#). The installation is straight forward for [RedHat](#), [SuSE](#), [Debian](#), [Mandrake](#) and [Slackware](#).

For further information, please contact: [mosix@cs.huji.ac.il](mailto:mosix@cs.huji.ac.il)

Copyright © 2001 Amnon Bank (amnon@cs.huji.ac.il). All rights reserved.



- Scalable Cluster Environment
- Diskless clusters
- Tools included

SCE Home

<http://pg.cpe.ku.ac.th/tesach/scr>



## Welcome to SCE Project

Scalable Cluster Environment (SCE) is a set of interoperable opensource tools that enable users to build and use Beowulf cluster effectively to solve their problems. This project is part of our research at [Parallel Research Group](#), Computer and Network System Research Laboratory, Department of Computer Engineering, The Faculty of Engineering, Kasetsart University



[SCE on SourceForge](#)

### Mailing-List

- [SCE User mailing-list](#)

### What 's new ?

October 9, 2001    [www.opensce.org is up!](#)  
September 6, 2001    [Release SCE 1.2 \(SCE with its document 115MB\)](#)  
July 8, 2001    [SCE got 4 penguins on mcows](#)  
June 21, 2001    [Release SCE 1.0](#)

SCE project is supported in Part by



KU Research and Development Institute



Advanced Microdevice Far East Inc.

### Partners

SCE and related software has been used on cluster product delivered by these partners.

Oak Ridge National Laboratory



# OSCAR

- v1.x – repackage current best practices
  - One button – it was pressed for you at the factory
- v2.x – repackage AND more...
  - Install
  - Maintenance
  - Daily operation

OSCAR Homepage

<http://oscar.sourceforge.net/>



[www.openclustergroup.org](http://www.openclustergroup.org)

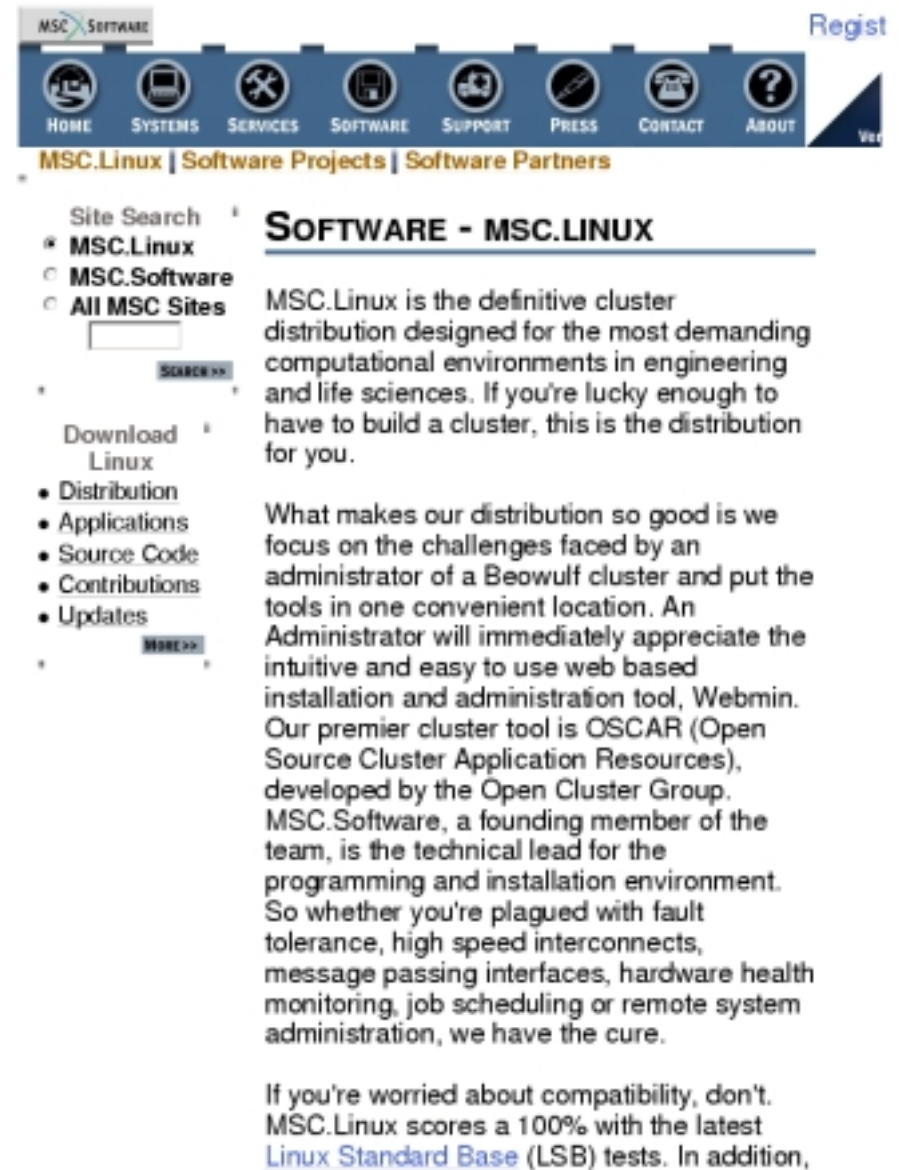
Oak Ridge National Laboratory



# MSC.Linux

- OSCAR based
- Adds
  - Webmin tool
  - Commercial grade integration and testing

Oak Ridge National Laboratory



The screenshot shows the 'Software' section of the MSC.Linux website. At the top, there is a navigation bar with icons for Home, Systems, Services, Software, Support, Press, Contact, and About. Below this is a search bar and a list of links: MSC.Linux, MSC.Software, and All MSC Sites. The main content area is titled 'SOFTWARE - MSC.LINUX' and contains a paragraph describing MSC.Linux as a definitive cluster distribution. It mentions that MSC.Linux is designed for demanding computational environments and highlights the inclusion of the Webmin tool for administration. The text also mentions OSCAR (Open Source Cluster Application Resources) as a premier cluster tool. At the bottom of the page, there is a link to the Linux Standard Base (LSB) tests.

MSC.Linux | Software Projects | Software Partners

Site Search  
\* MSC.Linux  
○ MSC.Software  
○ All MSC Sites  
SEARCH >>

Download Linux  
• Distribution  
• Applications  
• Source Code  
• Contributions  
• Updates  
MORE >>

## SOFTWARE - MSC.LINUX

MSC.Linux is the definitive cluster distribution designed for the most demanding computational environments in engineering and life sciences. If you're lucky enough to have to build a cluster, this is the distribution for you.

What makes our distribution so good is we focus on the challenges faced by an administrator of a Beowulf cluster and put the tools in one convenient location. An Administrator will immediately appreciate the intuitive and easy to use web based installation and administration tool, Webmin. Our premier cluster tool is OSCAR (Open Source Cluster Application Resources), developed by the Open Cluster Group. MSC.Software, a founding member of the team, is the technical lead for the programming and installation environment. So whether you're plagued with fault tolerance, high speed interconnects, message passing interfaces, hardware health monitoring, job scheduling or remote system administration, we have the cure.

If you're worried about compatibility, don't. MSC.Linux scores a 100% with the latest [Linux Standard Base \(LSB\)](#) tests. In addition,

# Cluster-in-a-Box

- OSCAR based
- Adds
  - Myrinet
  - Additional Alliance software
  - IA-64

Cluster-in-a-Box

visit: <http://www.ncsa.uiuc.edu/TechFocus/Deployment/CIB>

At the heart of the Alliance strategy is a Cluster-in-a-Box (CIB) effort that acknowledges the growing interest in commodity-based cluster computing and open source software in both academia and the private sector. The CIB effort has two goals: to develop software that greatly simplifies the task of installing and running a parallel Linux cluster that is compatible with the Alliance's large-scale production clusters; and to provide a software foundation on which other software packages—including grid toolkits and scalable display wall software—can be built.

The initial work on developing CIB software was done by a collaboration among academic and private industry researchers called the Open Cluster Group, which resulted in the software package Open Source Cluster Applications Resources (OSCAR). OSCAR targets clusters up to about 64 nodes that support Ethernet-based messaging. Using OSCAR, a user can get a small to mid-sized cluster up and running in a matter of hours, instead of days. The larger Alliance CIB effort expands on the work of the Open Cluster Group by allowing users to create clusters of more than 64 nodes, by supporting Myricom's Myrinet as the interconnect among cluster processors, and by integrating Alliance software into the overall package, including Condor, PVHS, and tools for monitoring jobs, measuring performance, and accounting.

Read more about OSCAR at <http://access.ncsa.uiuc.edu/Releases/010125.OSCAR.html>.

Contact [technotes@ncsa.uiuc.edu](mailto:technotes@ncsa.uiuc.edu) with questions regarding this page.  
Last updated Jun 13, 2001. All rights reserved.  
©2001 Board of Trustees of the University of Illinois.



# OSCAR Components

---

- Functional Areas
  - Cluster Installation
  - Programming Environment
  - Workload Management
  - Security
  - General Administration & Maintenance
- Other
  - Packaging
  - Documentation



# OSCAR statistics

---

- Version 1.1 released August 2, 2001
  - August
    - Page views: 9,013    Downloads: 1,744
  - September
    - Page views: 31,698    Downloads: 5,006
  - October (close of 10-7-2001)
    - Page views: 13,046    Downloads: 1,422
  - Does not include systems using OSCAR
    - MSC.linux
    - NCSA's Cluster-in-a-Box



# OSCAR systems in the news

---

- LLNL (~Sept 1 time frame)
  - People – Jim Garlick, Chris Dunlap, Mark Seager
  - Vendors: Linux NetworkX & SGI
  - 3 clusters (236-nodes, 472-processors)
    - Parallel Computing Resource (PCR)
      - 126 dual P4 nodes
    - PCR production cluster
      - 86 dual P4 nodes
    - PCR development cluster
      - 24 dual P4 nodes



# Failures of OSCAR v1.x

---

- Static environment
  - No options during installation - all-or-nothing
  - No cluster configuration changes after initial installation
- Install scripts too task specific
  - Each written for their package with no reuse plan
- LUI was too new and both architecture and distribution specific
  - x86 only
  - RedHat 7.1 only - (v1.0 was RH 6.2)
- Integration tasks
  - We worked hard, not smart!

The logo for OSCAR 2 features a stylized graphic on the left consisting of overlapping yellow, red, and blue squares with a black crosshair. To the right of this graphic, the text "OSCAR 2" is written in a large, red, sans-serif font.

# OSCAR 2

- Decouple OS and cluster environment (OSCAR) install
  - Advantages
    - Allow OSCAR install over existing OS and user environment
    - Users may select alternative OS installation scheme
      - Example: RedHat KickStart – or – from distribution CD
    - Better support for site differences
      - Example: non-private cluster, special purpose nodes or networks
  - Disadvantage
    - OSCAR can't rely on SI S to load cluster software by default with OS
  - Default OS installer is System Installer Suite (SI S) – (next generation LUI)
    - Combination of: LUI & System Imager
    - Initial setup
      - Architecture: x86, alpha
      - Distribution: RedHat, MSC Linux, TurboLinux, Mandrake, Debian



# OSCAR 2 (continued)

- New OSCAR Wizard
  - Installation modes
    - Simple
      - Minimum set of questions
      - Default OSCAR packages
      - OS install gets default package
    - Standard
      - Same minimum set of questions as "Simple"
      - Can't decide yet...
    - Expert
      - Complete control over installation options







# OSCAR 2 (continued)

---

## OSCAR maintenance modes

- Node
  - Add
  - Delete
  - Update
  - Reinstall
  - Status – up/down
- Package
  - Add
  - Delete
- Interface
  - Status – up/down
- Boot time initialization script
  - Run as part of init process for “off-line” cluster build
- Configuration report



# OSCAR 2 (continued)

---

- Standard API requirement for OSCAR packages
  - Package provides single machine instantiation of:
    - Install
    - Uninstall
    - Configure
  - Package provides list of required files (e.g. tarball, RPM, etc)
  - OSCAR provides “cluster aware” backend to distribute to all nodes via two technologies
    - C3
    - SIS

# OSCAR partners

- MSC Software
- ORNL
- Indiana University
- IBM
- NCSA
- Intel
- LLNL
- Dell
- Ericsson, Canada
- Your name here...

