Open Source Cluster Application Resources (OSCAR) : design, implementation and interest for the [computer] scientific community.

Benoît des Ligneris[a*], Stephen Scott[b], Thomas Naughton[b] Neil Gorsuch[c]

[a]Centre de Calcul Scientifique, Université de Sherbroke, Québec, Canada

[b]Computer Science Group, Oak Ridge National Laboratory, Tennessee, USA

[c]NCSA, University of Illinois at Urbana-Champaign, Illinois, USA

The Open Source Cluster Application Resources (OSCAR) project is the founding working group of the Open Cluster Group (OCG). The OCG is an informal group of people dedicated to making cluster computing practical for high performance computing and more recently, clustering in general (high availability, diskless). OSCAR is a package that makes it easy to build clusters for high performance computing. Everything you need for a cluster – installation and configuration, maintenance, [parallel] programming, queuing system, scheduler – is included in OSCAR.

In this article we will present a brief history of OSCAR, the actual components of OSCAR as well as the design and implementation proposal for OSCAR 2.1 and beyond.

Le projet OSCAR, Open Source Cluster Application Ressource, constitue un groupe de travail du Open Cluster Groupe (OCG). L'OCG est un rassemblement informel d'individus dont le but est de faciliter l'installation, l'utilisation et la maintenance de cluster dédiés au calcul haute performance. Récemment, cet objectif a été étendu au clustering en général : cluster de haute disponibilité et cluster sans disque. OSCAR est donc une suite logicielle qui facilite grandement la création et l'administration de clusters. Tout ce qui est nécessaire à ces objectifs : installation, configuration, maintenance, programmation parallèle, système de gestion de travaux, programmateur est inclu dans la suite OSCAR.

Dans cet article, nous présenterons une courte histoire du projet OSCAR, les composants actuels de OSCAR ainsi que l'évolution préssentie de OSCAR pour les version ultérieures.

1 Introduction

Cluster computing is the "new" way of doing parallel computing. From the modest creation of this field with the Wiglaf project [1,2] to the huge supercomputers that are now built, a wide range of new problems can be addressed and solved. The price and as a consequence, the accessibility of supercomputers is going down as the ratio power/price of commodity hardware (PC components) drops.

There are several cluster distributions available [3–11], of which, OSCAR is a very popular choice for clustering today [12]. There have been slightly less than 85,000[2] downloads of the software, which shows sound indication that the project has good stability and a large user base. OSCAR is not specific to a particular distribution and is not, as such, linked to any Linux vendor. The current offering has support for Red Hat and Mandrake Linux and efforts within the group are investigating support for SuSE Linux.

OSCAR provides a cluster computing solution, including development of a set of tools for cluster installation and management[3]. The focus of the development is "best cluster practices", taking the best of what is currently available and integrating it into one package. This focus offers the user the ability to get a cluster up and running with standard tools in a timely fashion.

*Corresponding author : benoit@des.ligneris.net

[2]Number based on a snapshot taken January 28, 2003.

2 History

A complete history of the project has already be published [13]. OSCAR was created in April 2000 (first OSCAR meeting) because, it was apparent that the assembly of a cluster was challenging for some and often tedious by all. Although reference material was available [14], each cluster was an experiment in itself and cluster builders had to download, compile, buy, and test a large variety of software and hardware components. Moreover, technological evolution (hardware and software) was happening fast and as a consequence cluster computing was a perpetual (re)invention of the wheel.

After this first meeting the group decided to choose "best practices" for High Performance Computing (HPC) clusters. The open-source solutions for each component were identified and work began on the software stack. This work is still in progress as detailed in Table 1. A primary objective of the OSCAR working group is to make the installation, configuration and management of modest sized clusters easier.

OSCAR is the first project set forth by the Open Cluster Group (OCG). The OCG is an informal group dedicated to making cluster-computing practical for high-performance computing research and development While membership to the OCG is open to anyone, it is directed by a steering committee with committee positions up for election every two years. Currently, the steering committee is made up of representatives from IBM, Indiana University, Intel,

Release	Date	Distribution(s)
2.1	2002-12-12	RedHat 7.2, 7.3, IA64 Mandrake 8.2, 9.0(exp)
2.0	2002-11-20	RedHat 7.2, 7.3, IA64 Mandrake 8.2
1.4	2002-09-30	RedHat 7.2, 7.3, IA64 Mandrake 8.2
1.3	2002-07-15	RedHat 7.1, 7.2 Mandrake 7.2(exp)
1.2.1rh72	2002-04-12	RedHat 7.2
1.2	2002-02-11	RedHat 7.1
1.1	2001-08-03	RedHat 7.1
1.0	2001-02	RedHat 6.2

Table 1
OSCAR software release history. Experimentally supported distributions are marked with '(exp)'.

MSC.Software, National Center for Supercomputing Applications (NCSA), and Oak Ridge National Laboratory (ORNL) [3].

3 Overview

The installation of a computing cluster typically involves the installation and configuration of a *headnode*, which in turn provides services to the compute nodes. Once the headnode has been setup the compute nodes must be "built". This latter portion is where a great deal of time and effort is often duplicated unnecessarily. OSCAR assists with the headnode configuration and then builds the compute nodes based upon a user specified description of the desired software packages. This build and configuration greatly reduces the effort and expertise necessary to setup a cluster.

3.1 Installation

In order to use OSCAR one simply does a standard 'Workstation' install with a supported Linux distribution and architecture, see also Table 1. The distribution RPMs should be copied to a default directory for use during the node image build phase. The user is then able to download, extract and start the OSCAR install wizard.

A graphical user interface (GUI) is available to assist with the installation. This GUI based OSCAR wizard consists of several steps that the user walks through in sequence to perform necessary operations to obtain a master-node (node executing wizard) and a set of compute nodes. The necessary headnode services are configured during this process as well as the number of nodes and what software to install on the cluster.

The heart of the OSCAR installation is the construction of a logical node "image"[4] (SIS image see Section 4.1) that contains a base operating system, plus the selected cluster software packages. The cluster nodes are defined (IP address, etc.) and associated with an SIS image. Then each node network boots (via a floppy or PXE), exchanges identification information (MAC address), is remotely formatted and installed based upon the SIS image. This entire cluster installation process is managed by OSCAR and therefore greatly reduces the expertise and time required to build and configure a cluster.

3.2 Design

High performance computing software evolves very rapidly and, as a consequence, the OSCAR framework has to be able to do frequent releases in order to provide, at a given time, a snapshot of the best known methods for building, programming, and using clusters. This strong requirement leads to a distributed architecture based on OSCAR packages. The OSCAR framework includes some core utilities as well as the core packages that have to be installed on every OSCAR cluster. Once this is done, a collection of additional packages can be downloaded and installed that provide functionality to the cluster.

With this modular design, packages can be updated independently of the core utilities and core packages. Additional flexibility to this model is provided by the Oscar Package Downloader (see also: Section 3.2.1), which allows for multiple repositories so that each package maintainer can provide package updates. The OSCAR toolkit is freely redistributable and therefore requires all distributed packages to conform, each organization is free to setup their own OSCAR repository for internally developed (i.e. not free) software.

This distributed system allows the OSCAR community to efficiently manage contributions from every interested developer : each organization has the complete control of the packages they maintain. Functionality can be added and bugs can be fixed every time it is necessary without affecting functionality of other packages. Regular check points are made so that all packages remain compatible : this correspond to a major OSCAR release (as specified in Table 1).

3.2.1 Modular Packaging

The modular packing facility enables OSCAR to install and configure software that has been dropped into a prescribed directory format. The packages are based upon the RedHat Package Manager (RPM) system and allow for additional scripts for cluster configuration. The OSCAR packages are basically the RPM-ized software plus an XML meta file describing the package and various installation criteria, e.g. where to install the software: server, clients, both. The OSCAR packaging API allows authors to provide additional documentation as well as tests to validate the installation.

A selected set of packages are "included" with OSCAR and others are available from remote locations. The OS-

[3] For more information on OCG and its working groups see, http://www.openclustergroup.org/

[4] This is in contrast to a binary "bit" image, e.g. as used by the `dd` utility.

CAR Package Downloader (OPD) is similar to the Perl CPAN utilities for acquisition of packages from remote repositories. This capability enables package authors to provide updates and/or packages that might have restrictions prohibiting direct inclusion in OSCAR. Once the OSCAR packages are obtained they are added to the list of available software for installation on the cluster.

4 Components

The common software packages that are installed on High Performance Computing Clusters (HPCC) are bundled with OSCAR. A small set of packages are "core" or required to support the base OSCAR framework. This core set of software is used for the installation and configuration of the cluster.

This section discusses the four major groups of cluster software found in OSCAR. The *Core Infrastructure and Management* facilities as mentioned above are required and tasked with the installation and configuration of the cluster. The *Administration/Configuration* packages are essential for cluster maintenance [5]. The remaining two sections summarize the HPCC and Security packages.

4.1 Core Infrastructure/Management

Environment Switcher

A standard issue with system administration is management of a user's environment. This includes things like which implementation of MPI is the default for users. These defaults should also be easily modifiable on a per user basis. Typically this environment management is performed by manually modifying ".dot" files, e.g. *.bashrc, .cshrc*.

The Environment Switcher tool, or simply Switcher, provides a clean mechanism to manipulate system- or user-level defaults through a command-line interface [15]. This allows for cluster administrators to set a system default, e.g. MPI = lam-6.5.7. Then a user has the option of accepting system defaults, checking for other available software and selecting an alternative at the per user basis, e.g. MPI = mpich-1.2.4. Again, these environment changes are via the Switcher command-line interface and do not require the user to manually edit files.

Switcher extends the Modules tool by making the environment settings persistent upon subsequent shell invocations. Another feature of switcher is that environment is maintained across non-interactive shells, e.g. rsh/ssh.

C3 - Cluster, Command and Control

The C3 power tools offer a command-line interface for cluster system administration and parallel user tools [16]. They are the product of scalable systems research being performed at ORNL [17,18]. The tool set includes commands to execute (`cexec`) across the entire cluster (or a subset of nodes) in parallel. The scatter/gather (`cpush`/`cget`) operations are also available. The C3 power tools have been developed to span multiple clusters. This multi-cluster capability is not fully harnessed by OSCAR currently but is available for administrators or standard users.

C3 is used internally throughout the OSCAR toolkit to distribute files and perform parallel operations on the cluster. The OPIUM tools (see Section 4.2) use C3 as well to provide clusterized user management tools, e.g. `useradd`. Since C3 enables standard Linux commands to be run in parallel, administrators can use the tools to maintain clusters. A common example is the installation of a Red Hat RPM software package across the cluster using for example `cexec rpm -ivh foo.rpm`.

SIS - System Installation Suite

OSCAR uses the System Installation Suite (SIS) to perform the initial installation of the compute nodes [19]. SIS is based upon the popular *SystemImager* tool. The *System Installer* and *System Configurator* tools extend SystemImager by adding a description based image construction interface and dynamic configuration facility, respectively.

The OSCAR wizard leads the user through the creation of a SIS image. This image is a "virtual node" having all the software contained on the cluster node(s). The nodes boot and copy an install script over the network that partitions and formats the hard disks, and then copies all software to the newly configured disk. The installation of the compute nodes is done automatically and after the nodes build and reboot they are accessible via the network.

SIS can be used beyond the initial installation to maintain a cluster by making modifications to the image and synchronizing the cluster nodes to transfer the changes. This image based cluster management is also useful for maintaining diskless clusters and used by the Thin-OSCAR working group, (see also: Section 5.1).

ODA (OSCAR DAtabase)

The ODA package is used by the OSCAR wizard and OSCAR packages to store and retrieve data. It uses the mySQL database as the underlying database engine. ODA adds user and program friendly interfaces for database access, and the capability to expand database commands in various ways, and for administrators and programs to define new database commands. These definitions are themselves stored in the database, and allow OSCAR and ODA to provide a database API that hides the details of the database. Each OSCAR package can define and modify it's own set of database commands that all OSCAR packages can use, allowing each OSCAR package to provide it's own database API. The package specific definitions can either be statically supplied by OSCAR packages, or created dynamically by the packages. Initially, the ODA package populates the database with various cluster information and static information supplied by each OSCAR package.

[5] Note, most of the "core" tools are as well used for cluster administration and maintenance.

4.2 Administration/Configuration

A number of tools are provided to assist cluster administration, including the previously mentioned SIS and C3 packages. The OSCAR Password Installer and User Management (OPIUM) utility synchronizes user account files across the cluster. The standard user add and delete commands are wrapped to be cluster aware. The Kernel-Picker tool allows one to substitute a given kernel into the SIS image prior to building nodes.

OSCAR also helps with the configuration of standard cluster services, such as Network File System (NFS) to share the user `/home` filesystem across the cluster. This enables users to easily access their files from any of the cluster nodes. The Network Time Protocol (NTP) service is also setup to keep a consistent time throughout the cluster as well as syncing the headnode to one of the atomic clocks on the Internet. Another useful configuration that is performed by the Loghost package is to forward system log (syslog) entries from compute nodes to the headnode. This centralization is helpful when trying to maintain a cluster.

4.3 HPC Services/Tools

Ganglia Monitoring System

Ganglia offers a scalable monitoring system with both command line and GUI based access tools. The base system as configured by OSCAR includes a monitor daemon on each node and a server on the headnode gathering the information. The system communicates over a multicast channel using XML as the data format language. Metrics such as CPU usage and uptime as well as identification information like OS version are added to the multicast stream. These metrics can be extended if desired to include further data.

The monitoring system offers a very nice web based GUI to browse historical data and view usage information for the cluster. The command line interface provides access to machine cpu load, which can be used to for creating host files if working outside a batch scheduling system.

HDF5 (Hierarchical Data Format)

HDF5 is a general purpose library and file format for storing scientific data. The data can be grouped into hierarchical user-defined structures with various datatypes and attributes. There are a variety of datatypes available, including compound and user-defined datatypes. The library is intended for users working with large files, theoretically in the terabyte and greater range.

Parallel Libraries

The message-passing libraries of PVM (Parallel Virtual Machine)[20] and the MPIs [21], LAM/MPI [22] and MPICH [23], provide the requisite capability in the OSCAR environment that enables one to take advantage of this distributed computing environment for the creation of parallel distributed programs. The OSCAR distribution tracks the most recent release of each of these packages and provides the necessary configuration at installation for the above libraries. There is no modification of the standard library package provided by each of these groups; OSCAR is only a repackaging of the standard library distribution with the addition of the cluster configuration/installation automation. OSCAR enjoys an inside track to the LAM/MPI and PVM packages, as core members of the OSCAR team are also members of these two parallel library teams.

OpenPBS/MAUI

The Portable Batch System, OpenPBS [24] is a flexible workload management system. It operates on networked, multi-platform UNIX environments, including heterogeneous clusters of workstations, supercomputers, and massively parallel systems. The OSCAR package includes several useful enhancements (node failure "support", support of more than 500 nodes, ...).

The default scheduler included with OpenPBS is a simple First In First Out (FIFO). Therefore, the more advanced MAUI [25] scheduler is included with OSCAR and setup as default. MAUI is very powerful and offers numerous advanced features, many of which go beyond the default OSCAR configuration. Numerous scheduling policies are supported as well as dynamic priorities, node reservations and fairshare.

4.4 Security

Clusters are under increasing attacks through outside networks. Providing cluster security is especially important for private research and even, in certain case sensitive (military, ...) research. Most of the provided OSCAR tools were not built with security in mind. The preferred networking model for clusters uses a private network for the cluster with limited outside access. The use of private network for securing the cluster from the outside world, strong encryption whenever possible (OpenSSH [26]) for internal and external communication as well as general usage of local firewalls (Pfilter[27,28]) both on the master node and on the regular nodes provide, we believe, a secure clustering solution. OSCAR configures both ssh and the core packages to use ssh within the cluster for communication wherever possible. This also provides secured logins into the cluster from outside. Pfilter is configured to provide packet filtering firewalls on each machine in the cluster. The firewalls are configured to allow unrestricted network communication between machines in a cluster, prevent all but a few types of authorized methods of outside access to the cluster, and allow complete unrestricted access from within the cluster to outside network resources.

Pfilter

Pfilter provides configurable packet filtering firewalls [27,28]. Pfilter is installed on each system as a system service that can be started, stopped, and restarted. Easy to understand, high level configuration files are "compiled"

by Pfilter into packet filtering ruleset command files. Pfilter also handles multiple network interfaces, allowing one machine to be a packet forwarding firewall for machine that are "behind" it on a private network, allowing the protected machines to access the outside network in a safe manner. It can also provide pseudo interfaces for protected hosts, allowing the protected hosts to be visible from the outside network.

OpenSSH

Access to a cluster is typically from remote locations (not the console). The Secure SHell (SSH) enables users to connect to the cluster in a safe manner. The open source implementation, OpenSSH, is configured and used by the default OSCAR configuration. The authentication and encryption schemes incorporated can be scaled for less or more stringent security needs. The usage is similar to that of telnet and rsh/rlogin with the provided security masked from the user.

5 Future

Since, the first OSCAR release in early 2001 there have been approximately eight major releases. Each of these releases has seen OSCAR evolve and improve. The current development path includes an overhaul of the base installation harness. This MetaMenu tool is being designed to provide a state-machine facility that will allow developers to more easily modify and extend installation and maintenance procedures.

Another goal is to enhance the OSCAR maintenance facilities. This includes the ability to both add, delete and update software from a uniform interface. This will include both the OSCAR packages as well as standard system updates. The OSCAR package itself is also going to be rebundled to allow for simpler upgrades between releases. The Linux Standards Base (LSB) is now being supported by most major distributions. Current OSCAR plans are to embrace this standardization for better multi-distribution support.

The scalability of OSCAR and its components is also being investigated. This may be an opportunity for the various OCG working groups to interact and extend current cluster practices.

5.1 Thin-OSCAR

The Thin OSCAR work-group [29] was created in Summer 2001 to specifically analyze and solve problems that will arise while adding diskless support to OSCAR clusters. The current implementation use the SIS image and create an initial RAM disk that contain a minimal Linux system. Users programs (/usr), OSCAR binaries (/opt) and users directories (/home) are exported via NFS.

The long term goals of the Thin OSCAR work-group are to support systemless nodes (i.e. nodes without any system on disk), heterogeneous clusters that are constituted of nodes with disk and system, systemless nodes as well as diskless nodes. Finally, Thin-OSCAR would like to provide several clustering solution by including cutting edge kernel with specialized clustering solutions for instance OpenMOSIX, BProc, LTT as well as distributed filesystems dedicated to clustering (e.g. cluster).

5.2 HA-OSCAR

The High Availability OSCAR working group was established in Summer 2002. This group is tasked with looking into and developing techniques to provide high-availability for both general purpose high-performance cluster computing as well as for special interest groups such as the telecommunication and defense industries. This group is presently studying the problems associated with achieving high-availability to determine the cost of implementing such operations. A reliability-modeling project is presently under way with preliminary findings presented in [30] submitted to this conference. Other efforts are underway including the development of tools for reliability estimation and the study and subsequent improvement of the various system infrastructure and framework to enhance self-healing and failure management. The long-term goal of this group is to develop a complete, cost effective, and open source solution to the high-availability problems associated with high-performance cluster computing.

6 Conclusion

OSCAR is a very powerful package that allows anyone interested by clustering and/or parallel programming to quickly and easily deploy a medium-size cluster. The modularity of OSCAR allows computer scientist, researcher and companies using computer science to easily develop new cluster concepts without being encumbered by installation, configuration and maintenance details. This is a big step toward general use of clustering in all areas of research as present-day commodity hardware is perfectly capable. This is the primary goal of the OSCAR working group.

Acknowledgments

We would like to thank the other OSCAR developers for their time and efforts in making such a sound clustering toolkit. Special thanks to Jeff Squyres who has done a wonderful job as 2002 OSCAR Chair, while managing an extremely militant schedule.

References

1. T. Sterling, D. Savarese, D. J. Becker, J. E. Dorband, U. A. Ranawake, and C. V. Packer. BEOWULF: A parallel workstation for scientific computation. In *Proceedings of the 24th International Conference on Parallel Processing*, volume I, Architecture, pages I:11–14, Boca Raton, FL, August 1995. CRC Press.
2. Phil Merkey. Beowulf, Introduction - History - Overview, May 1998. Revised May 2000. http://beowulf.gsfc.nasa.gov/overview.html.

3. Open Cluster Group: OSCAR Working Group. OSCAR: A packaged cluster software for High Performance Computing. http://www.OpenClusterGroup.org/OSCAR.
4. SCore Cluster System Software, http://www.pccluster.org.
5. Scyld Beowulf Scalable Computing, http//www.scyld.com.
6. MSC.Linux, http://www.msclinux.com.
7. NPACI Rocks Cluster Distribution, http://rocks.npaci.edu.
8. Scalable Cluster Environment (SCE), http://www.opensce.org.
9. ClusterMatic, http://www.clustermatic.org.
10. Cheos – LLNL, http://cheos.llnl.org.
11. KA-Tools, http://ka-tools.sourceforge.net.
12. Poll: *What Cluster system (Distribution) do you use?*, 233 votes (Feb. 01, 2002). Other 24%, OSCAR 23%, Score 15%, Scyld 12%, MSC.Linux 12%, SCE 6%, http://clusters.top500.org.
13. Richard Ferri. The OSCAR revolution. *Linux Journal*, (98), June 2002. http://www.linuxjournal.com/article.php?sid=5559.
14. Thomas Sterling, John Salmon, Donald J. Becker, and Daniel F. Savarese. *How to Build a Beowulf*. MIT Press, May 1999.
15. Env-Switcher, http://env-switcher.sourceforge.net.
16. Cluster Command & Control (C3) power tools, http://www.csm.ornl.gov/torc/C3.
17. M. Brim, R. Flanery, A. Geist, B. Luethke, and S. Scott. Cluster Command & Control (C3) tools suite. In *To be published in, Parallel and Distributed Computing Practices, DAPSYS Special Edition*, 2002.
18. Al Geist et al. Scalable Systems Software Enabling Technology Center, March 7, 2001. http://www.csm.ornl.gov/scidac/ScalableSystems.
19. System Installation Suite (SIS), http://www.sisuite.org.
20. Parallel Virtual Machine (PVM), http://www.csm.ornl.gov/pvm/.
21. Message Passing Interface (MPI) Forum, http://www.mpi-forum.org.
22. LAM/MPI: Implementation of MPI, http://www.lam-mpi.org.
23. MPICH: Implementation of MPI, http://www-unix.mcs.anl.gov/mpi/mpich/.
24. Portable Batch System (OpenPBS), http://www.openpbs.org.
25. MAUI Scheduler, http://supercluster.org/maui/.
26. Secure Shell (OpenSSH), http://www.openssh.org.
27. Network Packet Filter, http://pfilter.sourceforge.net.
28. Neil Gorsuch. PFILTER in OSCAR - Industrial Strength Cluster Firewalls in an Open Source Environment. In *To be published in, OSCAR2003 Conference*, May 11-14, 2003.
29. Benoît des Ligneris and Francis Giraldeau. Thin-OSCAR : Design and future implementation. In *To be published in, OSCAR2003 Conference*, May 11-14, 2003.
30. C. Leangsuksun, L. Shen, H. Song, S.L. Scott, and I. Haddad. The Modeling and Dependability Analysis of High Availability OSCAR Cluster System. In *To be published in, HPCS2003 Conference*, May 11-14, 2003.