The Penguin in the Pail – OSCAR Cluster Installation Tool

Thomas Naughton and Stephen L. Scott*
Computer Science and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, TN
{naughtont, scottsl}@ornl.gov

Brian Barrett†, Jeff Squyres‡and Andrew Lumsdaine‡
Department of Computer Science
Indiana University
Bloomington, IN
{brbarret, jsquyres, lums}@lam-mpi.org

Yung-Chin Fang
Scalable Systems Group
Dell Computer Corporation
Austin, TX
yung-chin_fang@dell.com

Abstract

The Open Source Cluster Application Resources (OSCAR) project grew from the need for cluster tools to build, configure and manage the pile-of-PC clusters, which are continuing to grow in popularity. The current "best cluster practices" are leveraged to enable users to install and configure a working cluster in a timely fashion. The current release offers a number of popular tools, as well as improvements to the installation process for a cluster. As OSCAR has evolved, new research and development has been explored to extend the current practices. This paper discusses the status of the OSCAR project and its goals and objectives. The current development is discussed as OSCAR heads toward a newly designed OSCAR 2.0 infrastructure.

1 Introduction

The small scale commodity off-the-shelf cluster computing experiment that began in the 1990's [8] has evolved into a viable option for high-performance computing. The installation time on these early experiments were acceptable, but as the scale increased installation has become a challenging and time consuming element. Even after a traditional cluster is running, configuration and maintenance presents a significant time requirement. The Open Source Cluster Application Resources (OSCAR) project was founded to study this problem and provide a reasonable solution to cluster installation, configuration and administration. The outcome of this project is the OSCAR package [4].

The OSCAR project is the first project set forth by the Open Cluster Group (OCG) [6]. The OCG is an informal group dedicated to making cluster-computing practical for high-performance computing research and development. More information on the OCG and the projects working under the OCG is available at the OCG web page, `http://www.openclustergroup.org/`. While membership to the OCG is open to anyone, it is directed by a steering committee with committe positions up for election every two years. Currently, the steering committee is made up of representitives from IBM, Indiana University, Intel, MSC.Software, National Center for Supercomputing Applications (NCSA), and Oak Ridge National Laboratory (ORNL).

The OSCAR project provides a cluster computing solution, including development of a set of tools for cluster installation and management. 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. Currently, the only hardware configuration supported by OSCAR is a single head node and multiple compute nodes networked to the head node[1]. While sufficient for many clusters, there are limitations to this design. Therefore, as the design of OSCAR progresses, more cluster configurations will be supported. A newly built OSCAR cluster will have all of the following features installed and configured:

- Head node services, e.g. DHCP and NFS
- Internal cluster networking configuration

*This work was supported by the U.S. Department of Energy, under Contract DE-AC05-00OR22725.

†Supported by a Department of Energy High Performance Computer Science Fellowship.

‡Supported by a grant from the Lilly Endowment

[1] *head node* is used in the standard fashion of central cluster server (NFS,DHCP,PBS,Image), plus where the ODR resides.

- System Installation Suite (SIS) bootstrap of the installation as well as image management tools
- OS for clients installed via network (PXE) or floppy boot procedure
- OpenSSH/OpenSSL for secure cluster interactions
- C3 power tools for cluster administration, management, and users
- OpenPBS standard batch queue
- MAUI scheduler
- Message passing libraries: LAM/MPI, MPICH, and PVM

The OSCAR project is comprised of a mixture of industry and research/academic members. The current members include: Bald Guy Software, Dell, IBM, Indiana University (IU), Intel, Lawrence Livermore National Laboratory (LLNL), MSC.Software, National Center for Supercomputing Applications (NCSA), Oak Ridge National Laboratory (ORNL), and Silicon Graphics, Inc. The project is open to new collaborators, and makes all available resources to begin development available on the OSCAR web page, `http://oscar.sourceforge.net/`.

The following sections will provide a brief history of the OSCAR project (Section 2) and then detail the current architecture (Section 3). We will follow this with discussion on plans for OSCAR 2.0 (Section 4), features planned for upcoming releases (Section 5) and our conclusions regarding current OSCAR development (Section 6).

2 Project History

The OSCAR project began life in early 2001, with the first public version, OSCAR 1.0, released in April of the same year. The 1.0 release was layered on top of RedHat 6.2 and used the Linux Utility for cluster Installation (LUI) [5], from IBM, and the C3 power tools [1, 2], from ORNL, for the installation of the cluster. The subsequent release, OSCAR 1.1, upgraded the Linux distribution to RedHat 7.1 and provided improvements to the documentation, as well as improving other support features.

The latest release, in February 2002, is OSCAR 1.2. Again, the release is based on RedHat 7.1. Unlike previous editions of OSCAR, version 1.2 uses the SIS tools [7], also from IBM, instead of LUI. The release updated many of the core packages and provided improved documentation. The number of steps required to install OSCAR (and an entire cluster) has been reduced, shrinking the total time to install a cluster. User feedback suggests that the OSCAR project is meeting its goal to become easier to use with each release.

The project moved to SourceForge.net during development of 1.1. In the six and a half months since the 1.1 release there have been over 19,000 downloads. The series of beta releases for OSCAR 1.2 have seen slightly less than 2,900 downloads over the six week release stabilization period. While hosted at SourceForge, over 22,600 copies of the OSCAR distribution have been downloaded. The download numbers reveal a significant interest in the OSCAR project. Also, a recent poll taken at Clusters@Top500.org [9] had OSCAR with a strong showing as compared to other cluster products. OSCAR collected $\approx 23\%$ of the 269 votes cast, second only to the *Other* category with $\approx 25\%$ of the votes[2]. This anecdotal data shows that OSCAR is active from not only a development standpoint but also by the HPC community.

3 OSCAR Architecture

The current OSCAR design provides for a system to install a set of packages, without differentiation between packages. However, the packages currently shipped with OSCAR can be divided into two groups: a small set of "core" packages, required for installation and configuration and a set of "selected" packages that are not required for installation and configuration, but that greatly enhance the functionality of the cluster. As design progresses, these groups will be more formalized and will be treated differently by the installation process (Section 4.2).

There are currently three items which comprise the "core" packages: SIS, C3, and the OSCAR Wizard. The following sections will highlight these three items and comment on their current status. In addition, there are a number of packages in OSCAR that provide significant benifit to cluster users, including user maintenance and SSH configuration.

3.1 System Installation Suite

The System Installation Suite (SIS) is new to OSCAR as of 1.2, replacing LUI. SIS is based on the well known SystemImager tool. The IBM enhancements come in the form of SystemInstaller and SystemConfigurator. These two components extend the standard SystemImager to allow for a description of the target to be used to build an image[3] on the head node. This image has certain aspects generalized for on-the-fly customization via SystemConfigurator, (e.g. `/etc/modules.conf`). SIS

[2] Based upon Feb 16, 2002 snapshot. Note, both MSC.Linux ($\approx 12\%$) and NCSA's Cluster-in-a-box use OSCAR.

[3] Here image is defined to be a directory tree that comprises an entire filesystem for a machine.

offers improved support for the heterogeneity within the cluster nodes, while leveraging the established SystemImager management model.

SIS is used to "bootstrap" the node installs – kernel boot, disk partitioning, filesystem formatting, and base OS installation. The image used during the installation can also be used to maintain the cluster nodes. Modifying the image is as straight-forward as modifying a local filesystem. Once the image is updated, `rsync`[4] is used to update the local filesystem on the cluster nodes. This method can be used to install and manage an entire cluster, if desired.

3.2 Cluster, Command & Control (C3) Power Tools

The C3 power tools are developed at ORNL and offer a command-line interface for cluster system administration and parallel user tools. They are the product of scalable systems research being performed at ORNL [1, 3]. The version of C3, 2.7.2, that is packaged with OSCAR 1.2 is written in Perl and provides parallel execution as well as scatter/gather operations. The most recent version, C3 3.x, will be packaged with OSCAR 1.3 and has been rewritten in Python. Among other enhancements to C3 3.x, the tools will support multi-cluster environments.

3.3 OSCAR Wizard

The last "core" package is the OSCAR Wizard. This currently consists of a set of screens that guide a user through the process of creating an image, defining the number of nodes, and configuring the network settings. Also, there is a step to run a test to confirm the cluster setup was successful.

Current development is directed toward creating a command-line interface (CLI) which will be used by future Wizards to drive the installation. This CLI will provide improved scriptability to system administrators that do not desire to use the GUI-based Wizard. This CLI will provide the necessary access to the ODR while performing the cluster configuration and installation tasks.

3.4 User Maintenance

OSCAR 1.2 provides a set of tools that significantly reduces the required work for maintaining user data across the entire cluster. Using the C3 tools, the important user files, including `/etc/passwd`, `/etc/shadow`, `/etc/group` are automatically synchronized accross the entire cluster. All maintenance is transparent to the users and system administrators.

[4] `rsync` is a tool to transfer files similar to `rcp/scp` [10].

3.5 Secure Shell Configuration

The Secure Shell, or SSH, provides a secure replacement for RSH. With added security comes greater configuration requirements, which places a burden on system administrators and cluster users wishing to use SSH on their cluster. OSCAR 1.2 installs OpenSSH and automatically sets up all the required configuration files for SSH, allowing for transparent replacement of RSH with SSH.

4 Roadmap to OSCAR 2.0

OSCAR currently offers a solid means for building and configuring a cluster. Development is currently under way toward the next major release of OSCAR – version 2.0. The design goals of OSCAR 2.0 include increasing installation flexibility and extending cluster management capabilities once the cluster has been installed. The following section will highlight some of the features that are slated to help meet these design goals. The proposed changes will be slowly integrated into OSCAR through a series of releases from the current 1.2, leading to OSCAR 2.0.

4.1 Cluster Management

The current release of OSCAR provides a sound tool for cluster construction and configuration, using SIS and C3. While these are powerful tools, there is no exposed high-level management package. This is an acknowledged issue and one that the OSCAR group seeks to remedy in the form of a standard interface to a set of tools for node addition/deletion and package management. The interface will allow the underlying mechanism to be masked. This masking of the underlying implementation (e.g. SIS, SIS+C3, etc.) allows others to extend or replace the management system if desired. This is important within the OSCAR group because system administrators often have strong convictions as to how things **should** be done.

The two major "camps" regarding cluster management, have been the strictly image-based management approach and what has been termed a *hybrid-model*. This latter approach combines the image-based approach with a more distributed mechanism such as one using C3 and maintaining "deltas" from a base image. The key component in the design is the OSCAR Data Repository (ODR), which will house information regarding the state of the cluster. The ODR is discussed in detail in Section 4.3.

The current strategy is to leverage the existing SIS tool to add improved functionality in a timely fashion. In parallel to this development the infrastructure for a more hybrid-model is being developed. The hybrid

model requires the implementation of a storage repository, the ODR, and a stated set of policies regarding package and node management. There may be tools offered to shift management schemes but to speed the development cycle, the features are being restricted for simplicity.

Regardless of the management model selected, one of the key features is the offering of a standard interface to the tools. The interface for cluster maintenence is being developed incrementally during a series of releases leading to the 2.0 release. Users and developers are encouraged to comment on the interface during development, in order to provide an interface satisfying the users' needs.

4.2 Modular Architecture

As OSCAR has evolved one of the clear burdens has been the integration process required for a major release. The current design has yielded a solid tool, but it requires a tight coupling of all packages that are contained in an OSCAR release. Another goal of OSCAR 2.0 is the introduction of a modular architecture for OS installation and package upgrades/changes. The decoupling of the OS install will allow a set of nodes to be installed by other means (e.g. CD-rom, KickStart, etc.) and then use OSCAR to do the remaining installation and configuration.

The introduction of a modular packaging system will remove the tight coupling between OSCAR packages. In addition, it will allow developers outside of the OSCAR team to contribute packages to the OSCAR system, extending the base components of OSCAR. The upcoming release of OSCAR, 1.3, will contain a prototype modular package system, for use with existing OSCAR packages. The interface requires a standard package, which is currently the widely used RPM system from RedHat. In addition to a RPM, the package creator may provide a set of scripts to perform configuration steps not possible within the RPM framework. As work progresses, the API for package maintainers will be available as part of the architecture document on the OSCAR web site.

4.3 OSCAR Data Repository

Current OSCAR designs offer very limited information about the cluster to packages or users. As the flexibility of OSCAR grows, there is a need for reliable data about the state of the cluster. The OSCAR Data Repository (ODR) is a generic interface to data on the cluster. The API will be implementable using a variety of data storage systems, but will likely resemble an SQL interface. Access to the data will be available from any node in the cluster. The API to the repository will be coupled with the improvements to the standard OSCAR Wizard. The modular packaging system will also enable scripts to query for information from the ODR at specified times to obtain information such as number of nodes, head node IP address, etc. This will allow many cluster-aware packages to configure themselves, reducing the load on a cluster system administrator.

4.4 Improved OSCAR Wizard

The OSCAR graphical user interface (GUI) and companion command-line interface (CLI) are targeting better usability. The design goal is an intuitive interface that is extensible. The maintainers of large clusters typically use CLIs to expedite matters and to craft new functionality beyond what the standard cluster administrator might need. The availability of both the GUI and CLI enables the standard administrator to function easily without regard for the underlying implementation and allow sites with special needs to extend the base functionality to fit their requirements.

The addition of an underlying CLI also enables different GUIs to be contributed without having to overhaul the entire system. The current GUIs that have been discussed for OSCAR include: Perl/TK, Webmin, and Python/Tkinter. The future might see someone else contributing an *ncurses* based GUI that as well leverages the CLI. The key being that the underlying CLI would remain the same for accesses into the ODR.

5 Upcoming Features

In addition to the planned architectural changes to OSCAR, there are a number of functionality enhancements planned for future releases. As these features become available, information will be added to the OSCAR web page.

- **IA-64 support** – There are a number of users who have begun to receive IA-64 hardware. The current 1.2 release has some support but has not been extensively tested.
- **External IP address support** – Several users have mentioned the need for external IP addresses. There have been some patches and testing for this and it is likely to be included in the near future.
- **Current RedHat release support** – OSCAR 1.2 supports RedHat 7.1. It is planned that future versions will support the current RedHat releases with little effort from the cluster maintainer.
- **Support for more Operating Systems** – While OSCAR will be RPM based for some time, there are plans to support more RPM-based distributions, noteably Mandrake Linux.

- **Entended interconnect support** – There is also effort into extending the supported interconnection networks beyond Ethernet, (e.g. Myrinet).

6 Conclusions

The OSCAR project has emerged as a useful tool for cluster installation and administration. The most recent release, OSCAR 1.2, is significantly simpler than previous versions. The introduction of SIS into OSCAR is a key factor in this increased ease of use. The project is growing also in functionality, while trying to maintain a balance between flexibility and simplicity. These improvements are the path to the next incarnation of OSCAR. OSCAR 2.0 will offer improved cluster management, a modular architecture, an enhanced OSCAR Data Repository (ODR) and extended GUI/CLI Wizard tools.

As development progresses toward OSCAR 2.0 the project will begin to extend, and even define, "best cluster practices". These extensions will provide improved cluster management, both at the node and package levels. The development of prototypes will begin to be integrated throughout the interim 1.x releases. OSCAR will also serve as a testbed for reference prototypes being developed by core members for the SciDAC Scalable Systems Software project [3].

The OSCAR project continues to be the cluster computing solution stack and provides powerful tools for cluster installation As new research/development begins the group continues to evolve. The coming releases offer significant features as well as potential standards for cluster APIs. The pail is most certainly half full!

References

[1] 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.

[2] Cluster Command & Control (C3) power tools, http://www.csm.ornl.gov/torc/C3.

[3] Al Geist et al. Scalable Systems Software Enabling Technology Center, March 7, 2001. http://www.csm.ornl.gov/scidac/ScalableSystems/.

[4] Open Cluster Group. OSCAR: A packaged cluster software for High Performance Computing. http://oscar.sourceforge.net.

[5] Linux Utility for cluster Installation (LUI), http://oss.software.ibm.com/developerworks /projects/lui/.

[6] The Open Cluster Group, http://www.OpenClusterGroup.org.

[7] System Installation Suite, http://sisuite.sourceforge.net.

[8] 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.

[9] Poll: *What Cluster system (Distribution) do you use?*, http://clusters.top500.org/pollbooth.php.

[10] A. Tridgell and P. Mackerras. The rsync algorithm. Technical Report TR-CS-96-05, Australian National University, Department of Computer Science, June 1996. (see also: http://rsync.samba.org).