OSCAR on Debian: the EDF Experience

Geoffroy Vallée, Jean-Yves Berthou, Hugues Prisker, Daniel Leprince

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

OSCAR on Debian: the EDF Experience

Geoffroy Vallée [1 2] , Jean-Yves Berthou [3] , Hugues Prisker [3] , Daniel Leprince [3]

N°5537

Mars 2005

Systèmes communicants

ISSN 0249-6399 ISRN INRIA/RR--5537--FR+ENG

OSCAR on Debian: the EDF Experience

Geoffroy Vallée [1 2] , Jean-Yves Berthou [3] , Hugues Prisker [3] , Daniel Leprince [3]

Systèmes communicants
Projet PARIS

Rapport de recherche n˚5537 — Mars 2005 — 13 pages

Abstract: Linux clusters are now an interesting solution to execute numerical simulations. But cluster use implies problems of deployment, because of their distributed architecture. In the past, each company developed his own solution based on various tools. Today, software suites like OSCAR are available and can be freely used. But such solutions may not be adapted to the current computing architecture of companies. For example, the current solution deployed inside a company may be based on a Linux distribution not supported by a software suite like OSCAR. Nevertheless, the use of a software suite like OSCAR is very interesting for companies that are not specialized in cluster management.

This paper presents the OSCAR port to the Debian distribution in the context of the numerical simulation platform deployed at Électricité de France (Electricity of France).

Key-words: cluster management, OSCAR, software packaging

(Résumé : tsvp)

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex (France)
Téléphone : 02 99 84 71 00 - International : +33 2 99 84 71 00
Télécopie : 02 99 84 71 71 - International : +33 2 99 84 71 71

OSCAR sur Debian : l'experience d'EDF

Résumé : Les grappes de calculateurs Linux sont maintenant une solution intéressante pour exécuter des simulations numériques. Mais les grappes de calculateurs posent des problèmes de déploiement à cause de leur architecture distribuée. Jusque maintenant, chaque entreprise a développé sa propre solution fondée sur divers outils. Aujourd'hui, des suites logicielles telles que OSCAR sont disponibles et peuvent être librement utilisées. Mais de telles solutions peuvent ne pas être adaptées à l'architecture informatique des entreprises. Par exemple, la solution déployée au sein de l'entreprise peut être fondée sur une distribution Linux qui n'est supportée par une suite logicielle telle que OSCAR. Néanmoins, l'utilisation d'une suite logicielle telle que OSCAR est très intéressante pour les entreprises non spécialisée en administration de grappes.

Ce document présente le port d'OSCAR sur la distribution Linux Debian dans le cadre de la plate-forme de simulation numérique déployée à Électricité de France.

Mots-clé : adiministration de grappes de calculateurs, OSCAR, empaquetage logiciel

1 Introduction

Clusters are widely used to execute numerical simulations. But the distributed architecture of clusters is still an important drawback to deploy clusters in an industrial environment. Without standard solutions for clustering, companies still need to have expertise and man power for cluster management, even if the company is not specialized in computer science. In the past few years, solutions have appeared to simplify cluster installation and management: OSCAR [4] and Rocks [7] are both created for high performance clustering using Linux. These two tools allow a user to easily and quickly install a cluster without specific expertise in clustering.

In companies, numerical applications are developed and executed on "reference" platforms, *i.e.*, platforms with a set of characteristics (programing tools libraries and runtimes environment) compatible with their engineering needs. To have a simple and cheep step between the development phase and the production phase, it is interesting to have the same platform for development and production. For Linux clusters, companies must choose first of all a Linux distribution.

Today, the Debian distribution is an interesting Linux distribution for scientific development. At Electricity of France (EDF), engineers have uses Debian since two years the "reference" Linux distribution. A solution has been developed to install and manage these clusters but a more generic approach must be found, since cluster management is not EDF's objectives but rather a responsability that requires a great deal of effort and is already addressed by tools such as OSCAR. Therefore, EDF is currently porting OSCAR to Debian with the goal to include this port directly into OSCAR.

This paper presents the OSCAR port on Debian at EDF. The remainder of this paper is organized as follows. In Section 2, we present an OSCAR overview. Section 3 presents an overview of the clustering platform at EDF. Section 4 presents the OSCAR port on Debian. Finally, Section 5 provides concluding remarks.

2 OSCAR Overview

OSCAR is a snapshot of the best known methods for building, programming, and using Beowulf clusters. It consists of a fully integrated software bundle designed for high performance cluster computing. Everything needed to install, build, maintain, and use a modest sized Linux cluster is included in the suite, making it unnecessary to download or even install any individual software packages on your cluster.

2.1 OSCAR Architecture

An OSCAR cluster is composed of a head node which is the file server (using NFS), front-end for user connections to the cluster and may be a Domain Name Server (see Figure 1). The other systems in the cluster are compute nodes.

[1]ORNL/INRIA/EDF - Computer Science and Mathematics Division - Oak Ridge National Laboratory - Oak Ridge, TN 37831, USA

[2]INRIA Postdoc co-funded by EDF R&D and ORNL

[3]EDF R&D - 1 avenue de Général de Gaulle, BP408, 92141 Clamart, France

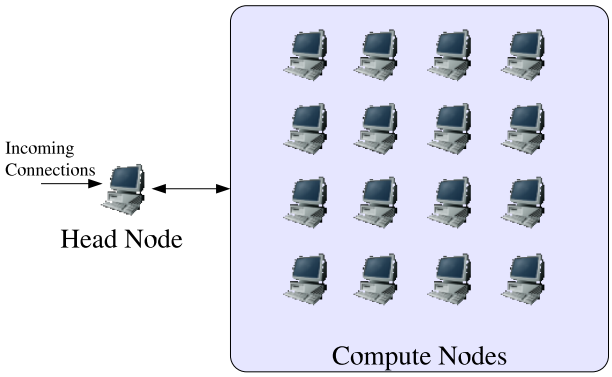

Figure 1: OSCAR Cluster Architecture

OSCAR merges different Linux packages in one *packages set*. To be able to manage these packages at the cluster scale, a new level of package has been created: *OSCAR packages*[5]. OSCAR packages and tools to manage these packages allows for management of software components at the cluster level. Typically, a package's state on nodes (*e.g.* installation state, versioning) are managed through the OSCAR database.

An OSCAR package is composed of an XML file, some configuration scripts and binary packages (see Figure 2). The XML file specifies information about the package version and dependences

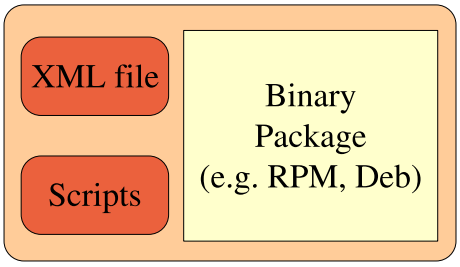

Figure 2: Architecture of OSCAR Packages

with other OSCAR packages. Scripts are used to install, update and remove OSCAR packages at the cluster scale (*versus* scripts included in binary files which install, update and remove packages at the local scale). When an OSCAR package is installed, updated or removed from a cluster, the OSCAR database is used to store and manage this package configuration information. This database records the current state of installed packages, including dependencies between OSCAR packages. Finally, OSCAR manages packages at the cluster scale, by leveraging tools used by Linux distribution for local software management.

The lastest OSCAR version (OSCAR 4.0) supports various Linux distributions, including Red Hat 9, Fedora Core 2 and Red Hat Enterprise Linux 3. Future OSCAR releases will support other Linux distributions. The port on a new Linux distribution is simplified by an abstraction layer, which allows the system to manage various package format in a uniform manner (see Figure 3). This layer is composed of *Packman* and *Depman*, a generic interface for binary package management. Finally this layer virtualizes OSCAR packages regarding the Linux distribution (*i.e.* the binary package format).

Packman is a tool to virtualize the binary package management facilities of supported Linux distributions. With this tool, it is possible to have different binary package formats for different Linux distribution, and manipulate them in a consistent manner.

Depman is a tool to manage dependencies between binary packages, for a given Linux distribution. Currently, Packman and Depman are only available for RPM based packages management systems.

Two kind of packages compose OSCAR: (i) *OSCAR application packages* and (ii) *OSCAR tools packages*. *Application packages* provides well-known applications, libraries to program and use the clusters, *i.e.*, MPI[3, 1], PVM[9], Ganglia[2]. *OSCAR tools packages* provide some internal facilities that are used to configure the cluster. For example, the package *kernel_picker* allows a user to select the kernel installed on the compute nodes.

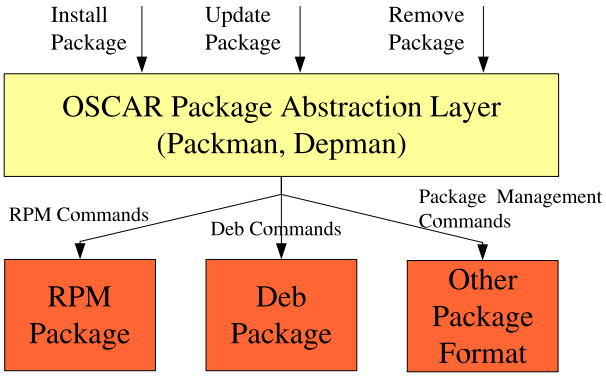

Figure 3: OSCAR Package Abstraction Layer

2.2 OSCAR Cluster Installation

An OSCAR cluster is installed in seven steps (see Figure 4): (i) package selection, (ii) packages configuration, (iii) head node installation (iv) creation of the image for compute node, (v) definition of the compute nodes, (vi) network installation of nodes, (vii) final cluster configurations. The details of each step are beyond the scope of this paper but an overview of an OSCAR cluster installation is given in this section. For a more complete discussion see [8, 6].

First of all, the head node are installed: the Linux distribution and the OSCAR software have to be installed. To be able to install head nodes and compute nodes, OSCAR uses a local repository of all packages included in the supported Linux distribution being used, which is in a local */tftpboot/rpm* folder.

When the head node is completely setup, OSCAR can be launched to install computes nodes. During the OSCAR initialization, some scripts install and setup basic packages. These packages are composed of OSCAR tools to manage a complete cluster, *e.g.*, the OSCAR database, Packman/Depman, .etc...

After the initialization, to install and setup a cluster, OSCAR first setups up the head node for OSCAR needs (package installations and setups), and then an image is created form the compute nodes. With this image, the cluster administrator can define the set of compute nodes that are to be based on this image. Then, a network boot of the compute nodes, start an automatic installation and transparently builds and configures all compute nodes.

Finally, the cluster installation is completed by executing post-install scripts for the applicable packages.

3 EDF Background

Electricité De France (EDF) is the national french electricity company and the first producer of electricity in Europe. EDF transports and sells electricity throughout Europe. The EDF company has a research and development group composed of 2,400 engineers and researchers to anticipate market evolutions and design future solutions to produce, transport and sell electricity.

Over the years, numerical simulation has become an essential and strategic tool for the firm in guaranteeing the safety of, in increasing the life span of, or in optimizing the performance of facilities for generating power from fossil fuels, hydroelectric power, or power from other renewable energies, and in transporting, distributing, and selling electricity.

In recent decades, EDF has developed computational codes in fields as varied as neutronics, molecular dynamics, thermo-hydraulics, structure mechanics, hydrology, hydraulics, financial mathematics, or indeed combinatory optimization. In the late nineteen nineties, it became apparent that simulation needed to move forward in order to offer new benefits. Among the changes deemed necessary were multi-scale and multi-physics 3D code coupling, and making high performance computational resources available.

EDF R&D develops numerical simulations for mostly generation and risk management, material simulation, thermo-hydraulics, structure mechanics and neutronics. Linux clusters, being a computing resource between workstations and HPC computers, are an interesting solution to execute such applications. To simplify the development and the deployment of these applications, a common software platform for both clusters and workstations is used by engineers.

3.1 The Calibre Distribution

The *Calibre* Project was launched at the end of 2000 for defining and implementing a workstation frame of reference that is common to the entire scientific computing community of EDF, and based

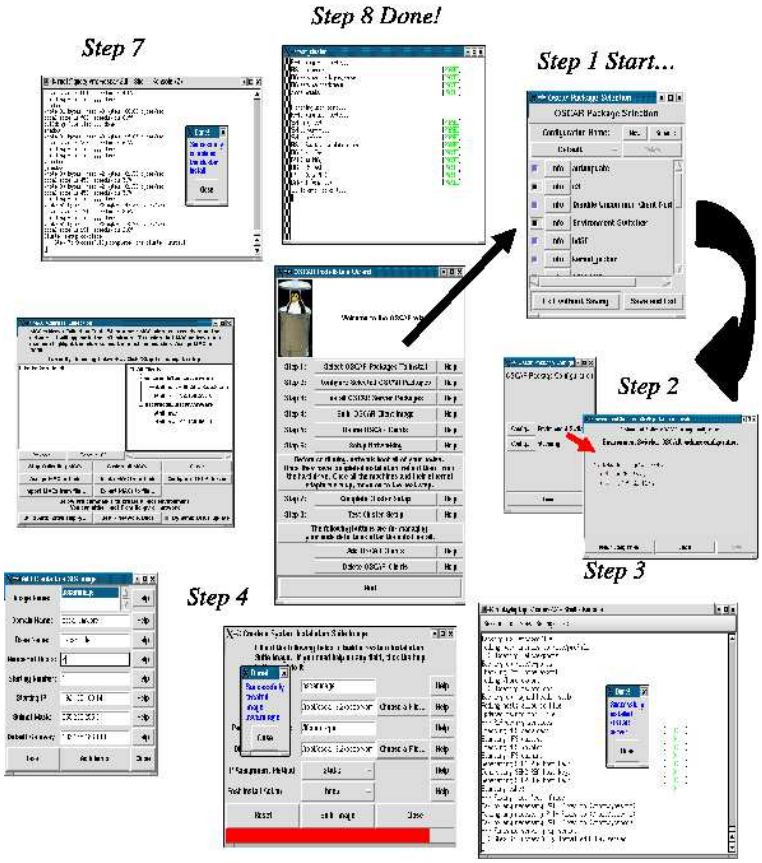

Figure 4: Cluster Installation Using OSCAR

on the PC Intel and Linux tandem. In October 2001, the *Calibre* distribution became the reference Linux system (SR-LINUX) of EDF. Initially, the *Calibre* project was based on the Red Hat Linux distribution, to provide all the tools needed to develop applications on an engineer's workstation and to execute applications on clusters.

The *Calibre* 4.0 version has released in 2004 at EDF R&D and is based on the Debian distribution. The Debian distribution was chosen for the release policy of the Debian project which allows EDF engineers to create only one *Calibre* version each year. Another advantage of Debian is the distribution's architecture, which separates the hardware management and user space software. With a separation of hardware and user space software, the company may buy new hardware without to change the complete *Calibre* distribution (*i.e.* changing only the kernel or adding drivers). The *Calibre* distribution is currently the standard for Linux deployment into the EDF company and therefore will be deployed on 1000 engineers's workstations and on 20 clusters (from 16 to 200 processors) in 2006.

3.2 Calibre Architecture

The calibre architecture looks like the OSCAR architecture. To be deployed on a whole cluster, *Calibre* needs a Debian mirror and a directory on the master node where all the node profiles are stored. In a second step, the master node is setup to create a cluster profile and to answer to network boot requests from compute nodes to initiate node installation. In a Calibre cluster, the master node executes network service (*i.e.* bind, NIS, DHCP) and the batch system. A separate file server acts as NFS server (see Figure 5). All the nodes use the NFS distributed for users home directories file system (the NFS's automounter for the */home* and */local* directories).

An important work of packaging was made to create a complete set of packages for scientific developments. *Calibre* is based on these packages.

3.3 Summary

EDF R&D developed a solution to install and manage clusters and engineer workstations. But EDF is not a company that aims at creating and maintaining a Linux distribution like *Calibre*. Therefore, a new solution based on a "standard software suite", must be deployed. OSCAR is a good suite for clustering, unfortunately OSCAR does not support the Debian distribution. However, the design of the *Calibre* distribution is very close to the OSCAR design. Therefore, a port of OSCAR on Debian was initiated.

4 OSCAR on Debian

There are three different issues to port OSCAR on Debian: (i) port of OSCAR initialization scripts, (ii) port of *OSCAR tools packages*, and (iii) port of OSCAR packages for *OSCAR application packages*.

In software suites like OSCAR, which want to simplify package management at the cluster scale, a major issue to support a new Linus distribution is the support for a new binary packages format.

Linux distributions like RedHat, Fedora Core, Mandrake or SUSE are based on the Red Hat package management (RPM); Linux distributions like Debian, Knoppix, Progeny or Ubuntu are based on Deb packages. These two formats are incompatible and the current tools to automatically convert between binary package formats are not powerful enough to guarantee a valid conversion. Some packages have to be manually converted to the new format (*i.e.* Deb format).

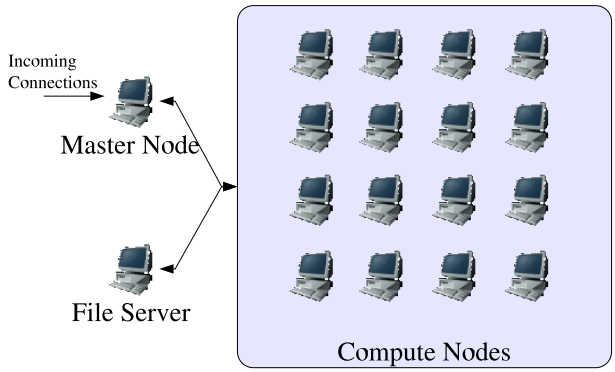

Figure 5: Architecture of a Calibre Cluster

4.1 Port of Initialization Scripts

Initially, OSCAR was developed for RPM based Linux distributions. Therefore some scripts are RPM based. For example, all OSCAR initialization scripts (scripts executed before the installation of Packman and Depman) are more or less dependent on RPM tools. Therefore, these scripts are not fully compliant with a Debian distribution. To simplify the maintenance of the OSCAR source and to simplify the OSCAR port to new Linux distributions, we tried to implement a high-level API for basic package management commands (to install and remove binary packages). The goal is not to create a new Packman/Depman package but to have a very simple generic way to write initialization scripts.

With such a layer, initialization scripts can be implemented both for RPM and Deb based Linux distributions. Moreover by decoupling commands that manage binary packages and initialization scripts, the maintenance of these scripts is simpler.

4.2 Port of OSCAR Tool Packages

The *OSCAR tool packages* are not available in the binary Debian package format. Therefore such packages have to be created from scratch. OSCAR packages are also composed of scripts used to manage packages at the cluster level. These scripts used in the current OSCAR packages may also contain dependencies or items that are specific to RPM based distributions. Therefore an audit must be performed for all existing OSCAR packages scripts to identify and resolve any incompatibility

with the Debian distribution. Once these scripts are compatible with both RPM based and Deb based Linux distributions, Debian packages can be integrated into OSCAR packages. Such OSCAR packages can be used both with Debian and RPM distributions.

Moreover, OSCAR 4.0 introduces the notion of binary package virtualization called Packman/Depman. Packman/Depman provides an API to manage binary packages (*i.e.* to install, update and remove packages). In Section 2 we saw that this abstraction layer allows a developer to more easily port to new Linux distributions that are based on other binary package formats. Therefore, to port OSCAR on Debian, a new Packman/Depman module must be created.

4.3 Port of OSCAR Application Packages

Like for *OSCAR tool packages*, the port of *OSCAR application packages* is composed of: (i) the port of binary packages and (ii) an audit of scripts for compliance with Deb based Linux distributions. The difference with *OSCAR tool packages* is that some existing binary packages may already exist.

The difficulty in creating binary packages for *OSCAR application packages* depends on the official Debian support for these packages. For example, well-known packages (*e.g.* MPICH) are available in both Debian packages and RPM packages. In this case, the port on Debian is very simple, the package allready exists.

In some case, packages are available in Debian packages but package names differ. In this case, the port to Debian just needs to modify dependencies with other packages. The complexity of this step depends on the dependence with other packages. In the worst case, it is possible to consider that the package is not available for Debian and create new packages.

If the package is not available with the standard Debian channel, it needs to be created from sources and in this case, some compiling issues may occur. Each Linux distribution provides a different version of compilers, some Linux distribution also modify compilers and finally, some applications are difficult to port on new distributions.

The last case is when packages are not available but the targeted distribution offers a similar package. In this case, the initial OSCAR package may be replaced by a new tool, directly provided in the targeted Linux distribution.

Finally, the port to Debian is composed of two phases: (i) identification of which packages already exist for the target Linux distribution (*i.e.* the package allready exists and does not have to be ported, the package must be rebuild) and (ii) the port of non-existent packages (see Table 1).

Like for *OSCAR tool packages*, to be sure that all existing scripts of *OSCAR application packages* are compatible with the Debian distribution, an audit must be performed.

4.4 Current State

The OSCAR initialization scripts were ported to work on Debian. A first implementation of an abstraction layer to install basic OSCAR packages during the OSCAR initialization was created. Using the abstraction layer, the *install_cluster* script, the *setup* script for Packman and the *setup* script of *XML-Parser* were modified to be compatible with both RPM based and Deb based distributions. This layer also provides generic functions to audit the underlaying Linux distribution.

	Package Name	In Debian	In Debian w/ other name	Not in Debian
OSCAR tools	C3			X
	disable-services		X	
	kernel-picker			X
	loghost		X	
	ntpconfig	X		
	networking		X	
	oda			X
	opium		X	
	perl-Qt	X		
	pfilter	X		
	sis	X		
	switcher			X
OSCAR applications	hdf5	X		
	lam-mpi	X		
	maui			X
	mpich	X		
	pbs			X
	pvm	X		

Table 1: Status of Binary Packages On Debian

The first port was based on OSCAR 3.0. The OSCAR 3.0 version does not support the abstraction layer of binary packages. Therefore it is difficult to accommodate new non-RPM based Linux distributions. The OSCAR 4.0 version has support for Packman and Depman so the OSCAR to Debian is based on OSCAR 4.0. Using Packman and Depman, a unique distribution for both RPM based and Deb based Linux distributions can be created (see Figure 6).

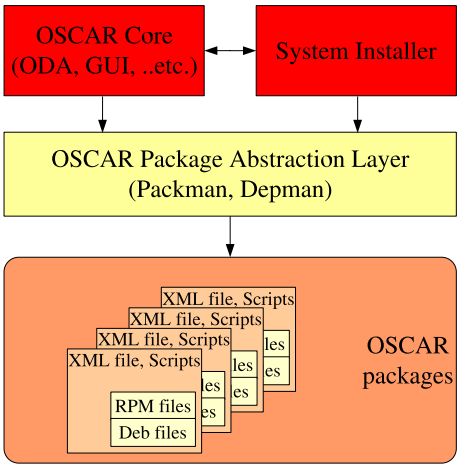

Figure 6: Design of OSCAR on Debian

The current Debian port includes all packages for *OSCAR application packages*. Some of them come directly from the Debian distribution and others were created from scratch. To port OSCAR packages to Debian, OSCAR scripts must be audited, but this audit is not yet done. Moreover, some other packages were specifically created for EDF needs (packages coming from the *Calibre* distribution) but are not integrated into OSCAR packages. For example, the MAUI scheduler has been packaged on Debian, with specific patches in order to be compliant to EDF needs (changes at Debian level instead of a general Debian and then make customizations at the OSCAR package level).

A subset of *OSCAR tool packages* were ported to Debian, specifically Packman and Depman. With initialization scripts, Packman and Depman allow a user to launch OSCAR and take advantage of basic OSCAR mechanisms, *e.g.* package installation at the cluster scale.

5 Conclusion

This paper presents the OSCAR port to the Debian distribution, packaging and porting the OSCAR cluster toolkit. Currently, the packaging effort is done. The port of OSCAR cluster tools is ongoing. Finally, the EDF effort will allow a user to have complete support for the Debian distribution in OSCAR.

This port provides a case study for the versatility of OSCAR and its use on other distributions, highlighting areas that need modifications to support a range of distributions.

The support of a large kind of different Linux distributions is very important to use OSCAR in an industrial environment, because without this capability, OSCAR cannot be integrated into the computer infrastructure of companies. Once integrated into the software framework of companies, OSCAR can become a reference for Linux clustering. Therefore, companies can have access to a complete open source solution for clustering without being dependent on any contractor or particular software company.

The paper presents the first step of the OSCAR integration in an industrial context. The second step is to validate the OSCAR architecture regarding the computing context of the company. This validation comprises a "scientific audit" and a "technical audit". The "scientific audit" must validates that all tools needed by EDF engineers are available in OSCAR. The "technical audit" must validates that the OSCAR architecture is compatible with the cluster architecture currently deployed in the company. Once these audits performed OSCAR will be used as reference for clustering.

References

[1] Al Geist, William Gropp, Steve Huss-Lederman, Andrew Lumsdaine, Ewing Lusk, William Saphir, Tony Skjellum, and Marc Snir. MPI-2: Extending the Message-Passing Interface. In Luc Bouge, Pierre Fraigniaud, Anne Mignotte, and Yves Robert, editors, *Euro-Par '96 Parallel Processing*, number 1123 in Lecture Notes in Computer Science, pages 128–135. Springer Verlag, 1996.

[2] http://ganglia.sourceforge.net. Ganglia: Distributed monitoring and execution system.

[3] Message Passing Interface Forum. MPI: A Message Passing Interface. In *Proc. of Supercomputing '93*, pages 878–883. IEEE Computer Society Press, November 1993.

[4] John Mugler, Thomas Naughton, Stephen L. Scott, Brian Barret, Andrew Lumsdaine, Jeffrey M. Squyres, Benoît des Ligneris, Francis Giraldeau, and Chokchai Leangsuksun. Oscar clusters. In *Linux Symposium*, Ottawa, Ontario, Canada, July 2003.

[5] Open Cluster Group, OSCAR working group. *HOWTO: Create and OSCAR Package*.

[6] Open Cluster Group, OSCAR working group. *OSCAR Installation Guide*.

[7] Philip M. Papadopoulos, Mason J. Katz, and Greg Bruno. Npaci rocks: Tools and techniques for easily deploying manageable linux clusters. In *CLUSTER*, pages 258–, 2001.

[8] Joseph D. Sloan. *High Performance Linux Clusters: with Oscar, Rocks, openMosix, and MPI*. Nutshell Handbooks. O'Reilly & Associates, November 2004.

[9] V. S. Sunderam. PVM: A framework for parallel distributed computing concurrency. In *Practice and Experience*, volume 2, pages 315–339, December 1990.

Unité de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399