Project no. IST-033576

XtreemOS

Integrated Project
BUILDING AND PROMOTING A LINUX-BASED OPERATING SYSTEM TO SUPPORT VIRTUAL ORGANIZATIONS FOR NEXT GENERATION GRIDS

XtreemOS Test Beds: Activity Report
D4.3.3

Due date of deliverable: August 31^{st}2010
Actual submission date: September 14^{th}2010

Start date of project: June 1^{st} 2006

Type: Deliverable
WP number: WP4.3

Responsible institution: INRIA
Editor & and editor's address: Yvon Jégou
INRIA
Campus de Beaulieu
35042 Rennes Cedex
FRANCE

Version 1.0 / Last edited by Yvon Jégou / Sep, 14, 2010

Project co-funded by the European Commission within the Sixth Framework Programme		
Dissemination Level		
PU	Public	√
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Revision history:

Version	Date	Authors	Institution	Section affected, comments
0.1	14/09/10	Yvon Jégou	INRIA	Initial document
0.2	14/09/10	Peter Linnell	INRIA	Additions about Development Test Bed

Reviewers:
Corina Stratan (VUA) and André Lage (INRIA)

Tasks related to this deliverable:

Task No.	Task description	Partners involved°
T4.3.3	Managing the XtreemOS large scale testbed	INRIA
T4.3.4	Managing the XtreemOS permanent testbed	INRIA, VUA, ICT, TID
T4.3.5	Creating an open testbed from the XtreemOS permanent testbed (opening/extension)	INRIA, VUA

°This task list may not be equivalent to the list of partners contributing as authors to the deliverable
*Task leader

Executive Summary

Operating system developments must be validated on real hardware. The behavior of a grid operating system depends on many parameters such as the number of nodes, their heterogeneity (memory, CPU, devices), the structure of the Grid (small or large clusters), the interconnection network (structure, latency, bottlenecks), the dynamism of the grid, the stability of the grid (node failures), the efficiency of grid services and so on. It is a simple conclusion therefore of the impossibility to evaluate it through simulation alone. Grid operating systems such as XtreemOS must be validated on a realistic test bed with the same characteristics as large grid installations. A vaild grid test bed must provide a significant number of computation nodes for scalability evaluation. The test bed nodes must allow full reconfiguration of the software stack for operating system experimentation from the low level communication layers all the way up to the top level grid services.

XtreemOS is a grid operating system targeting thousands of nodes.

In order to allow large scale experimentation of XtreemOS, the consortium provided access to the Grid'5000 platform to its members. However, as such a platform (more than 1000 servers) must be shared with many other projects, it is not possible to permanently dedicate part of the platform to a single project. On the other hand, during the course of the XtreemOS project, the necessity to have a permanent platform running the last stable release of the operating system appeared clearly for many reasons: to provide a reference installation to all members of the consortium, to provide a base for bug tracking. But, in order to not "disturb" the activities of the consortium members (bug tracking, performance evaluation), this platform cannot be used to experiment and validate new versions of the operating system component. A third test bed, the *testing* or *development* test bed was introduced in the course of the project. This test bed, also open to all consortium members, is used to validate new versions of XtreemOS components before they can be integrated to the official releases and installed on the permanent test bed.

The large scale test bed was used during the first phase of the project to evaluate single components: mainly RSS (resource discovery overlay), DIXI (the operating system bus) and AEM (service in charge of application execution management). A first Grid'5000 image was built upon release 2.0 of XtreemOS and has been used by various consortium members for the evaluation of the operating system on a large platform.

The permanent test bed was set up for the first release of the operating system. The number of nodes dedicated to this tested remained low (less than 5) during the first phase and then grew to 14 for release 2.1 in June 2010.

The number of nodes dedicated to the development test bed has been very variable during the project, as this test bed can be partitioned depending on the testing needs. Finally, this test bed has been the most active, as it is permanently exploited by the major developers of the project. The minimum number of nodes has been 5, but it has scaled up to 12 nodes for certain tests. Currently it runs 9-10 nodes depending on hardware availability. In addition it has hosted and tested the Kerrighed cluster versions of XtreemOS for long periods of time.

Contents

1 Introduction **5**

2 The permanent testbed **5**

2.1 Permanent testbed description 7

2.2 Problems in managing the permanent testbed during the project 7

2.3 The Open Testbed at the end of the project 8

3 The development testbed **8**

4 The large-scale testbed **8**

4.1 Large Scale Test Bed Users 9

5 Major Challenges **10**

6 XtreemOS Computing Challenge **10**

7 Conclusion **11**

1 Introduction

Large scale distributed systems like grids are too complex to be evaluated using theoretical models and simulators. Such systems must be tested on real size, real life experimental platforms. In order to prove the effectiveness of the results, the experiments must be reproducible. The purpose of WP4.3 is to setup an experimentation platform for XtreemOS. Because XtreemOS is an operating system, this platform must allow any type of experiment on the whole software stack. In order evaluate the scalability of XtreemOS, this platform must offer thousands of computation nodes.

The XtreemOS test bed serves multiple purposes in the consortium:

1. Provide permanent access to the last stable release of the operating system for all consortium members;
2. Provide a common base for bug tracking;
3. Allow the validation of new versions of XtreemOS components;
4. Provide a stable base for functionality evaluation by WP4.2 members;
5. Provide a base for scalability evaluation.

Some of these objectives appear to be difficult to support on the same hardware platform. For instance, it is not economically possible to setup a permanent test bed made of thousands of nodes: XtreemOS can be deployed and tested on a large-scale platform (Grid'5000) only during dedicated sessions. In order to support all platform needs, the consortium decided to maintain 3 different hardware platforms:

- The permanent test bed with about 10 nodes distributed among a few partners and running the current stable release of XtreemOS,
- The testing/development test bed dedicated to the validation of new component releases, as well as a reference for testing bugs discovered during development.
- The large-scale test bed dedicated to large-scale experiments.

Section 2 is dedicated to the permanent testbed, section 3 to the development testbed and section 4 to the large-scale testbed. Section 5 lists the major problems which had to be solved during the project and section 7.

2 The permanent testbed

The permanent testbed connects physical and virtual nodes running the current stable release of XtreemOS. These nodes are provided by different partners of the consortium.

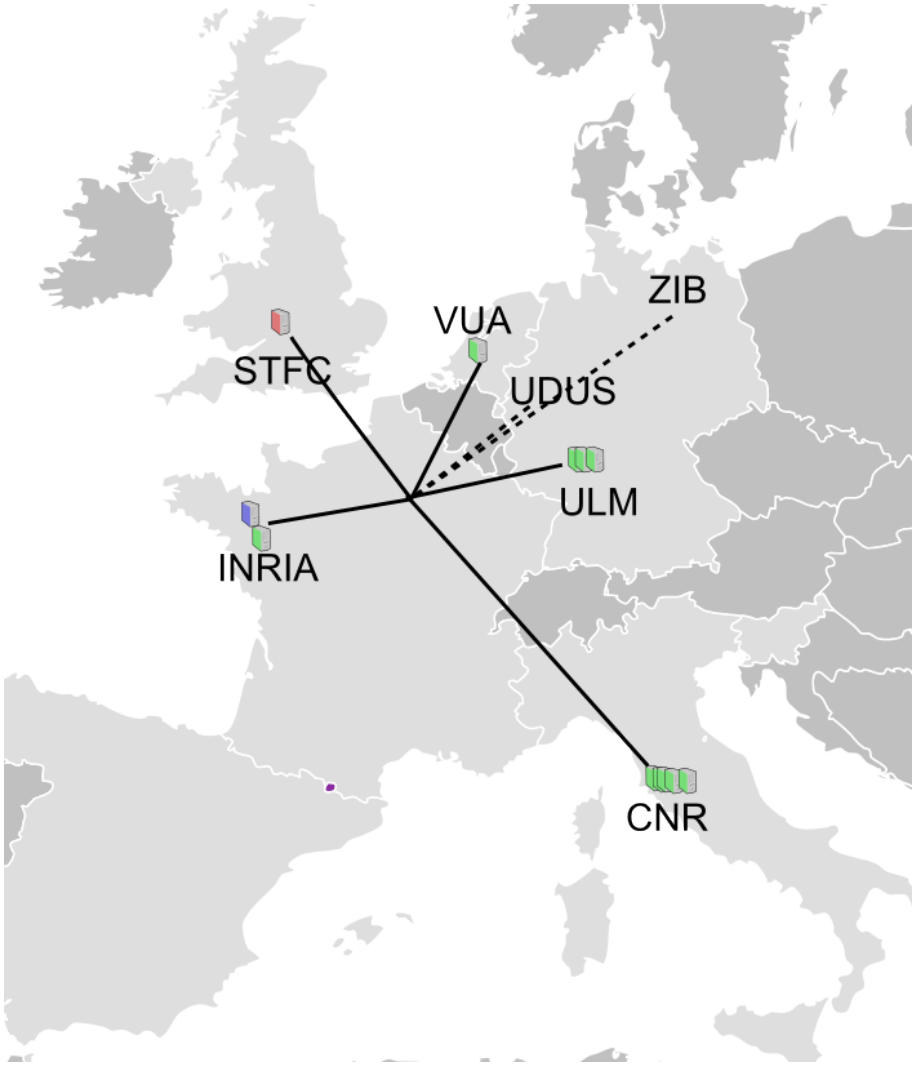

Figure 1: Overview of the permanent testbed in July 2010, during the XtreemOS summer school in Günzburg

2.1 Permanent testbed description

In order to facilitate the setup of the testbed, the entire configuration is described in the consortium's wiki[1]. For each node, this wiki-page provides the IP address, the list of services running on the node, any restriction related to Internet routing (firewall, open ports) and the e-mail address of the person in charge of the node. This page also provides a copy of the certificates necessary to connect to the testbed.

The initial permanent testbed setup consisted in VO management services (XVOMS data-base, certificate distribution authority (CDA) and VOlife web front-end) running at STFC, XtreemFS core services running at ZIB, AEM (Application Execution Management) services running at INRIA and resource nodes provided by INRIA, ICT, ULM, VUA, XLAB, CNR and UDUS.

In July 2010, during the XtreemOS summer school in Günzburg, Germany (see Fig. 1), the permanent testbed was made of

- 1 node running VO management components (XVOMS, CDA and VOlife) at STFC;
- 1 node running XtreemFS services, VOPS and all core AEM services at INRIA;
- resource nodes in other sites: 1 at INRIA, 6 at CNR, 3 at ULM and 1 at VUA.

Extra resource nodes provided by ZIB, UDUS and CNR will be added in September 2010.

2.2 Problems in managing the permanent testbed during the project

This initial plan to setup a permanent testbed appeared to be unrealistic during the first phases of the project. The major problems were due to

- Difficulties to remotely manage an unstable distributed operating system. In the case were some node misbehaves on the testbed, some means must be provided to shut it down in the case where the site admin is not present.
- The need to update all nodes in sync after each new release. Some releases of XtreemOS introduced modifications to the API of some components, making upgraded nodes incompatible with nodes running a previous release. Requesting all node admins to upgrade their nodes simultaneously became unworkable. The solution was to provide root access on all testbed nodes to some admins.
- The need to guarantee a coherent configuration of all nodes. XtreemOS provides some flexibility for node configuration: path names of configuration files, etc. However,this freedom makes bug tracking more complex. The provision of a new configuration tool for XtreemOS after release 2.0 provided a means to configure all nodes of the testbed from the same configuration file.

[1] http://xtreemos.wiki.irisa.fr/tiki-index.php?page=XtreemOS+Testbed&highlight=testbed

- The difficulty to configure the testbed after major releases. As new versions of component packages are validated on the development testbed, each new major release (integrating simultaneous modifications to multiple packages) resulted in major difficulties to re-configure the whole testbed. The problem here is probably due to unavoidable changes to configuration files and the lack of support currently for updating configuration files when a package upgrade is done. As the software matures, this issue will lessen in importance. In order to bootstrap the new release, we first needed to shutdown all testbed nodes. Then, configure a first node running all services (at INRIA), to debug and to reconfigure this single node until the correct behavior is observed. Finally, configure one or more resource nodes. Once the new release behaves correctly with this setup (one node running all services and one or more resource nodes), moving to the default setup is a matter of applying the new reconfiguration to all nodes.

2.3 The Open Testbed at the end of the project

During the last year of the XtreemOS project, it was decided to open access to the permanent testbed to users external to the consortium. Resource nodes located at ZIB and CNR serve as entry points to these external users.

3 The development testbed

During the first phase of the XtreemOS project, each developer group used his own local environment for testing, debugging and validation purposes. This process resulted in difficulty in the integration phases, as different developer groups were using different environments: different versions of compilers, of libraries, etc. The solution adopted by the consortium to face this problem was to setup a development testbed were all components can be validated in the same XtreemOS execution environment. As this testbed needs to support all forms of XtreemOS services (overlays, distributed, etc.), this testbed is similar to the permanent testbed: the nodes are distributed among the consortium partners with the major core services running at INRIA. The major difference is that, as the components being tested are less stable, their developers need to be easily accessible in case of problem. Instant communication between all XtreemOS developers is provided through the `#xtreemos-dev` IRC. More than 20 users have been permanently connected to this IRC during the last phase of the project. Use of IRC has been a major contributor to speed communication and allow real-time interaction amongst partners.

4 The large-scale testbed

The Grid'5000 testbed supports the deployment of the whole operating system stack (kernel, drivers, as well as root file system). Users must provide the operating system image which

is deployed on all nodes allocated to the user session. The Grid'5000 support environment provides a means to automatically install the image on hardware node disks, run user-provided scripts on these nodes and then boot the new operating system.

Building a classical new operating system image requires advanced skills in operating system configuration. In general, Grid'5000 users derive their own images from images made available by the Grid'5000 community. The process is more complex in the case of XtreemOS: it is necessary to configure each node in order to form a coherent set of services. As the nodes are allocated dynamically to the user at the beginning of the session, the whole XtreemOS configuration must be generated in the fly once the image is installed and before booting the nodes.

Few partners in the XtreemOS consortium have enough operating system skills to generate a Grid'5000 image for XtreemOS. INRIA generated such an image for release 2.0 of XtreemOS (updated with 2.1 components) along with the corresponding configuration scripts. This image can be deployed reliable at least 2 Grid'5000 sites (Rennes and Nancy) but fails to deploys on some nodes on other site. The major reason for these failures is due to the unusual size of the XtreemOS system image which results in time-outs during the deployment process. Two solutions to this problem are being evaluated:

- Reconfigure the Grid'5000 deployment system to support larger image sizes;
- Reduce the XtreemOS image size.

The default XtreemOS operating system image built from the `iso` provided by the consortium contains many packages. For instance a full graphical desktop, is utilized for large scale experiments where the user does not have access to each node's desktop.

Some components of XtreemOS, mainly in the Java environment are built using different versions of the same libraries and thus increase the size of the final image. Reconfiguring the development tools to use the same library release should also reduce the image size.

The initial plans when the XtreemOS project was to extend the Grid'5000 large-scale testbed with other clusters provided by XtreemOS partners, for instance DAS-3 in Amsterdam or CNGrid in China. The major difficulty with these clusters is that they do not provide any means to users to deploy their own operating systems. It is possible that advances in cloud technology will provide means to run XtreemOS on these platforms inside virtual machines in the future. However, experiments using nodes provided by VUA at Amsterdam and ICT in China have been run through the permanent and the development testbeds.

4.1 Large Scale Test Bed Users

About 25 users from the XtreemOS consortium have requested an account on the Grid'5000 platform during the project . The platform has been used mainly for testing the scalability of some major XtreemOS components such as the RSS overlay, the DIXI system bus or the AEM

application management system. A few users, mainly related to the WP4.2 evaluation work package have experimented with the XtreemOS system image.

5 Major Challenges

Installing and configuring a distributed operating system is a complex process and requires expertize in many facets of operating systems. The experience of the XtreemOS project shows that very few of the developers were capable of installing the whole system in order to test and validate their developments. A major part of the software developed in the course of the project was tested only outside the XtreemOS environment. This lack of early validation of new components makes the integration phase (when all components are assembled) and the configuration phase (when the components are configured to cooperate) error prone and rather lengthy.

The configuration of XtreemOS remains a complex process and the majority of the developers need some support in order to run experiments on XtreemOS. The permanent and the development test beds mostly resolved this need. The provision of fully configured virtual machines was also very helpful, as it allowed users to run their own configuration independently of the evolution of the other testbeds.

6 XtreemOS Computing Challenge

The consortium launched the *XtreemOS computing challenge* in April 2010[2]. All competitors were given access to the permanent testbed and to Grid'5000. All of them selected the Grid'5000 platform. Here is the result of this challenge:

1. *Monte Carlo Simulation for Single-Photon Emission Computed Tomography with XtreemOS*, Emanuele Carlini, Sebnem Erturk & Giacomo Righetti from University of Pisa, Italy

2. *Grid Security Operation Center*, Syed Raheel Hasan, Jasmina Pazardzievska, Maxime Syrame (supervised by Julien Bourgeois) Laboratoire Informatique, Université de Franche-Comté, France,

3. *Parallel Kriging* Alvaro Parra, Exequiel Sepulveda & Felipe Lema from ALGES lab at Universidad de Chile.

[2]http://www.xtreemos.eu/hotspot_news/xtreemos-computing-challenge

7 Conclusion

In conclusion, a substantial amount of engineering effort and hardware was devoted to establishing, operating and supporting the XtreemOS test beds. There is no doubt this was practical and worthwhile investment. Many improvements to the quality of the software, as well as unanticpated enhancements to XtreemOS only came about via the active use of the test beds. That the remaining partners will continue to put material and engineering efforts validates the ongoing work within this realm of the project.

As the end of the project is close, the consortium has plans to offer easy access to XtreemOS for external users: access to the permanent testbed has been opened, virtual machine images are made available and images for the grid'5000 platform are maintained.