Building and Promoting a Linux-based Operating System to Support Virtual Organizations for Next Generation Grids

Executive Summary

The XtreemOS operating system provides for Grids what a traditional operating system offers for a single computer: abstraction from the hardware and secure resource sharing between different users. It thus simplifies the work of users belonging to virtual organizations by giving them the illusion of using a traditional computer while removing the burden of complex resource management issues of a typical Grid environment. When a user runs an application on XtreemOS, the operating system automatically finds all resources necessary for the execution, configures user's credentials on the selected resources and starts the application. The XtreemOS operating system provides three major distributed services to users: application execution management (providing scalable resource discovery and job scheduling for distributed interactive applications), data management (accessing and storing data in XtreemFS, a POSIX-like file system spanning the Grid) and virtual organization management (building and operating dynamic virtual organizations). The implementation of this new Grid operating system introduces new challenges:

- *Scalability*: supporting hundreds of thousands of nodes and millions of users dynamically joining and leaving the Grid,
- *Transparency*: hiding the complexity of the Grid by distributed operating system services allowing to run new and legacy applications seamlessly,
- *Interoperability*: complying with all major standards such as POSIX and SAGA,
- *Dependability*: providing reliability and high availability through checkpointing and replication,
- *Security*: ensuring trust and integrity according to customizable policies.

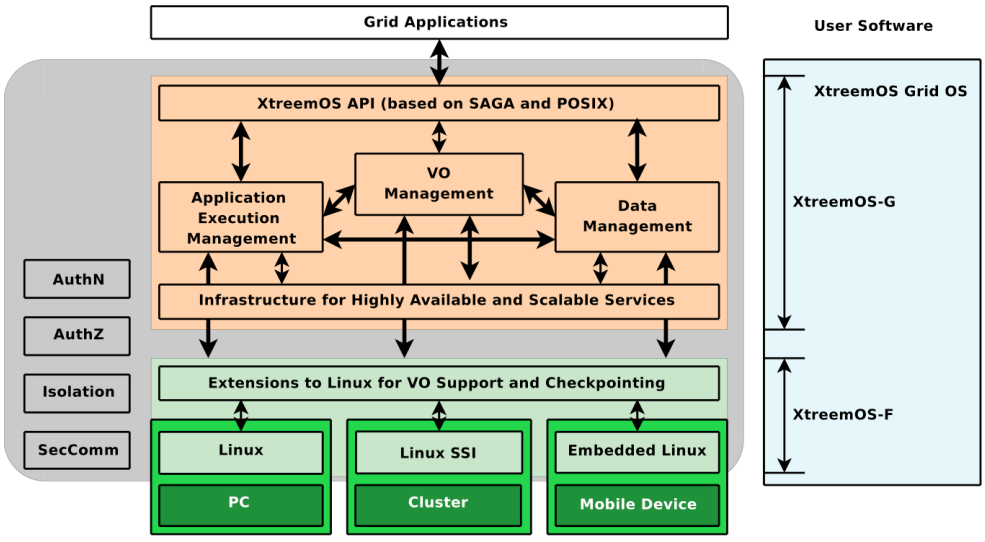

Figure 1: XtreemOS Architecture

During the third year of the project, the main achievements are the first basic versions of the three flavours of XtreemOS system for PCs, clusters and PDAs (XtreemOS 1.0). XtreemOS system integrates a comprehensive set of cooperating system services. XtreemOS software components range from Linux kernel modules to application-support libraries. The overall layering of these components,

grouped within software packages, is shown in Figure 1. XtreemOS 1.0 has been packaged for both Mandriva and RedFlag Linux distributions and has been available as open source software since the end of 2008 under GPL/BSD license. Since then, XtreemOS is openly developed on INRIA's Gforge. XtreemOS 1.0 has been extensively tested with a set of applications from different sectors. This provides valuable feedback taken into account by XtreemOS system developers when implementing XtreemOS 2.0. Two of these applications are already being distributed with XtreemOS 1.0 (Galeb, a financial application and Wissenheim, a virtual world application). Two additional applications will be packaged with XtreemOS 2.0 (jCAE, a computer aided engineering application and OpenTurns for uncertainty treatment by probabilistic methods). XtreemOS 1.0 experiments have been executed on a permanent test bed covering multiple administrative domains. It is composed of a set of PCs provided by different partners in different locations. We have also worked during the last twelve months on the design and implementation of an advanced version of XtreemOS for PCs and clusters while improving the performance, scalability, ease of use, and robustness of the system. This work resulted in the second major version of XtreemOS system, XtreemOS 2.0, packaged for Mandriva Linux distribution. An advanced version of the mobile device flavour is under development for smart phones.

During the last year of the XtreemOS European project, we will further improve the system and prepare a third major release for the three flavours of the system. A number of demonstrations will be presented, for example at EuroPar 2009 in Europe and SC'09 in the US. Virtual machine images of XtreemOS system will be made available to further disseminate the results. Support will be provided to the user community. Ensuring the sustainability of XtreemOS technology will be one of our major goals.

VO Management and Security
XtreemOS supports various VO models, used in scientific as well as business scenarios and multi VO scenarios. User management and resource management are independent: there is no need to configure resources when new users are registered in VOs. XtreemOS provides Single-Sign-On (SSO): when users perform a ``login'' within a VO, they receive credentials recognized by all resources of the VO without any need to re-authenticate. Resource access security in XtreemOS is policy-driven: access rights to a resource are evaluated from policies provided by users, VOs and resource providers; the latter remaining always in control. In XtreemOS, the isolation degree (for QoS, performance and security) can be customized using native Linux mechanisms to execute applications: cgroups, containers, name spaces and virtual machines.

Resource Discovery and Selection
The resource discovery mechanism within XtreemOS is based on the SRDS distributed information service comprising of the Resource Selection Service (RSS) and the Application Directory Service (ADS). RSS takes care of performing a preliminary selection of nodes to allocate to an application, according to range queries upon static attributes. It exploits a fully decentralized approach, based on an overlay network, which is built and maintained through epidemic protocols. As each node represents its own attributes in the overlay, failure management does not require any specific repair operation. ADS handles the second level of resource discovery, answering queries expressed as predicates over the dynamic attributes of the resources. ADS creates an application-specific ``directory service'' using the resources received by the RSS. Dynamic node attributes are periodically updated into the system. To provide scalability and reliability, DHT techniques and their extensions to range and multi-attribute queries are used. ADS supports Scalaris transactional DHT and OverlayWeaver.

Application Execution Management
XtreemOS implements a job-oriented scheduling within the subset of resources obtained by the resource discovery mechanism. To ease the use of the Grid services, XtreemOS mimics the well-known Linux functionality as opposed to offering different abstractions and functionality, which are more oriented to the Grid. For instance, AEM implements job control through signals. The reservation service implemented by AEM provides resource co-allocation allowing the execution of distributed/parallel applications on multiple Grid nodes. Reservations can be dynamically modified. AEM also provides flexible and accurate job monitoring. It manages job dependencies, providing an

interface to external workflow engines. Lastly, XtreemOS fully supports the execution of interactive applications. The XtreemGCP Grid checkpointing service takes care of reliable execution of distributed applications that can take advantage of rollback recovery protocols. XtreemGCP can checkpoint/restart and migrate applications running on multiple heterogeneous Grid nodes. Various kernel checkpointers are supported through a library implementing a common checkpointer interface: BLCR, a checkpointer based on OpenVZ containers on PCs and LinuxSSI checkpointer on clusters.

XtreemFS Grid file system
XtreemFS is a distributed file system designed for deployment in wide-area environments spanning multiple locations in different administrative domains. It allows mounting an XtreemFS volume from any location, given the right permissions. XtreemFS is transparently integrated with the XtreemOS VO infrastructure in the form of dynamic user mappings and automatic mounting of home volumes. XtreemFS stores file data on several storage servers. It implements an object-based file system architecture. The metadata of a file is stored separate from the file content on a metadata server. This metadata server organizes file system metadata as a set of volumes, each of which implements a separate file system namespace in form of a directory tree. XtreemFS is a full-featured file system that supports the full POSIX file interface, including extended attributes. In case of concurrent access by several distributed programs, it provides currently the NFS close-to-open consistency. XtreemFS provides efficient access to file data through file striping on multiple servers. It also manages data replication transparently to users.

Volatile Data Management
The Object Sharing Service (OSS) aims to ease the sharing of volatile data objects by transparently managing replicas and keeping them consistent. Grid applications can share objects through standard file system operations or by using customized functions. The latter includes support for speculative transactions, which alleviates network latency and avoids complicated lock management.

Distributed Server & Virtual nodes.
A distributed server is an abstraction that allows a group of server processes to appear as a single entity, with a single IP address, to its clients. Distributed Servers aim at allowing high-performance client-to-server communication, while being totally transparent to the clients. The only requirement is that the clients support the Mobile IPv6 protocol. The goal of virtual nodes is to provide fault-tolerant functionality to service-oriented applications with minimum effort for the service developers. Virtual nodes allow a programmer to choose a replication policy among several available ones, which include several flavours of passive and active replication. Virtual nodes and distributed servers are highly complementary to each other: virtual nodes provide fault-tolerant replication that is mostly transparent to the service developer, but lack an access method that makes fault-tolerance transparent to the clients. Distributed servers provide a solution for making the service replication transparent to the client. Although each service has its own utility when used in isolation, we see that merging both systems would in principle allow one to build fault-tolerant replicated services where the complexity of replication would be transparent to both the service developer and to the client-side application.

XtreemOS API
The XtreemOS API has to serve three classes of applications: existing Linux applications, using POSIX interface, existing Grid applications, using the OGF SAGA interface, new applications, using functionality uniquely provided by XtreemOS. We have defined an API name space called XOSAGA (XtreemOS extensions to SAGA) that mirrors the SAGA API name space. XOSAGA contains only those packages, classes, and interfaces that require XtreemOS-specific extensions to SAGA. Together, SAGA and XOSAGA form the XtreemOS API, implemented in C/C++, Java, and Python.

XtreemOS Cluster Flavour
The XtreemOS cluster variant is based on LinuxSSI, which implements a full Single System Image (SSI) operating system for computing clusters. A full SSI operating system globally manages all cluster resources to give the illusion that a Linux cluster is a single Linux SMP-like node, allowing the execution of unmodified legacy sequential or parallel applications and system administration tools.

LinuxSSI leverages Kerrighed SSI technology. Additionally, it implements kDFS a distributed / parallel file system exploiting the disks attached to compute nodes and providing support for handling persistent states during checkpoint/restart operations.

XtreemOS Mobile Device Flavour
XtreemOS also provides a mobile device flavour (XtreemOS-MD), which fully integrates most of XtreemOS functionalities, giving users on the move full access to the XtreemOS Grid. This kind of approach is much more scalable than gateway or Grid portal solutions for mobile access, as it eliminates the potential bottlenecks and single-points of failure of these gateways. Moreover, mobile Grid applications are able to run transparently with little or no modifications in mobile devices, due to the inclusion in XtreemOS-MD of OGF's standard SAGA API. Support for context-aware applications is studied for integration in the advanced version of XtreemOS-MD for smart phones. XtreemOS provides not only a full Grid operating system for mobile devices, but also a set of open source software modules that can be easily integrated into any modern mobile Linux distribution, by avoiding excessive reliance on any specific mobile platform.

XtreemOS Fact-sheet
XtreemOS is a four-year European Integrated Project funded by the European Commission that started in June 2006.

Contractors

Organisation name	Country
Caisse des dépôts et consignations	FR
Institut National de Recherche en Informatique et Automatique	FR
Science and Technology Facilities Council	UK
Consiglio Nazionale delle Ricerche	IT
European Aeronautic Defence and Space Company	FR
Electricité de France	FR
Edge-IT	FR
NEC Deutschland GmbH	DE
SAP	DE
Barcelona Supercomputing Center - Centro Nacional de Supercomputación	ES
Universitaet Ulm	DE
Vrije Universiteit Amsterdam	NL
Xlab	SL
Konrad-Zuse-Zentrum für Informationstechnik Berlin	DE
Institute of Computing Technology of Chinese Academy of Sciences	CN
Red Flag Software	CN
Telefónica I+D	ES
Universitaet Düsseldorf	DE

Administrative and Financial Coordinator
Jean-François Forté, Caisse des dépôts et consignations
15 Quai Anatole France
75007 Paris, France

Scientific Coordinator
Christine Morin, INRIA
Campus universitaire de Beaulieu
35042 Rennes cedex, France

Contact: contact@xtreemos.eu
Public web site: http://www.xtreemos.eu