Executive Summary

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

Executive Summary

The XtreemOS project is building an operating system to support virtual organizations (VOs) in next-generation grids. Unlike the traditional, middleware-based approaches, it is a major goal to provide seamless support for VOs at all the software layers involved, ranging from the operating system of a node, via the VO-global services, up to direct application support.
XtreemOS implementation is based on Linux. XtreemOS integrates operating systems for the various computer architectures used in VOs. For stand-alone PCs (single CPU, or SMP, or multi-core), XtreemOS provides its Linux-XOS flavour with full VO support. For clusters of Linux machines, the LinuxSSI flavour combines VO support with a single system image (SSI) functionality. For mobile devices, XtreemOS provides the XtreemOS-MD flavour with VO support and specially-tailored, lightweight services for application execution, data access, and user management.

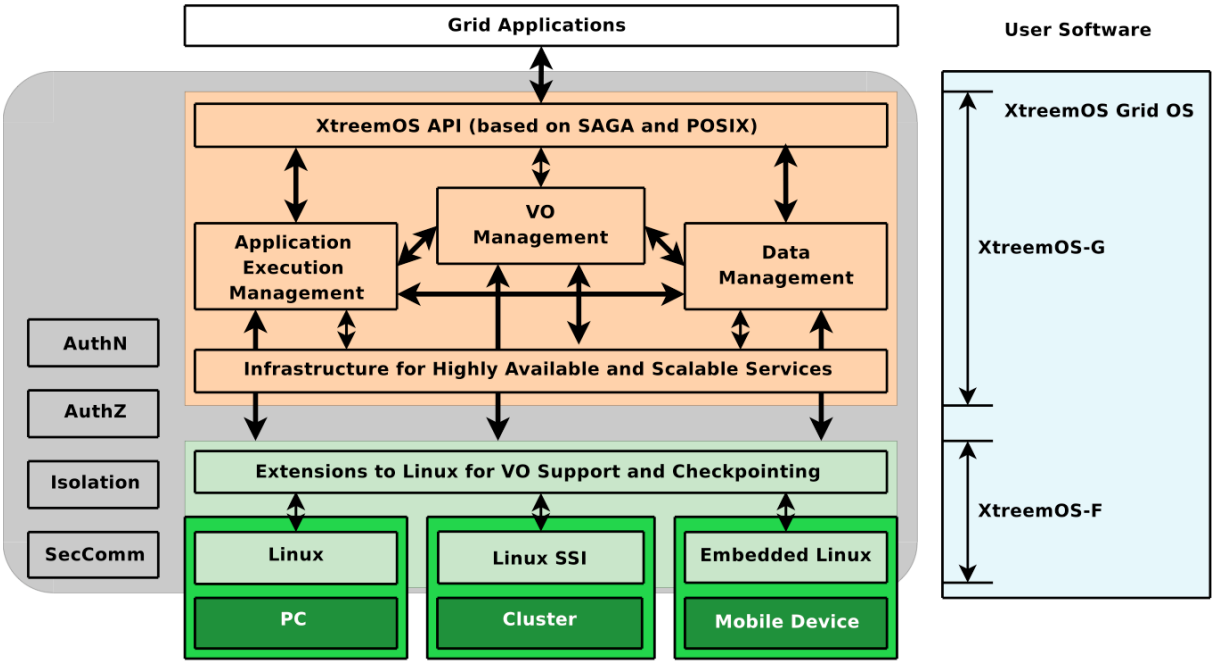

Figure 1: XtreemOS Architecture

During the second year of the project, we produced a first basic version of XtreemOS software components, ranging from Linux kernel modules to application-support libraries. The overall layering of these components, grouped within software packages, is shown in Figure 1. Each layer abstracts further from the underlying physical structure of a Grid, and consists of one or more software packages. A software package provides one or more of the services of XtreemOS. Each service implements its functionality by interacting with other services in the same layer, and the layer below. Here, services can be either ``classical'' Grid services within the XtreemOS-G layer, or Linux extensions (kernel modules etc.) within the XtreemOS-F layer. All

XtreemOS services being designed within a single project in a cooperative way, which ensures better integration and better coordination between these services than in traditional middleware stacks. During the second year of the project, we refined XtreemOS architecture studying the interactions between all XtreemOS services. A particular emphasis has been put on the definition of the overall security architecture and its implementation. We designed and implemented a first basic version of XtreemOS system for PC and clusters. The XtreemOS services have been integrated and packaged to produce a first release of XtreemOS Grid OS. We intend to make this release public after an internal evaluation by XtreemOS partners not involved in XtreemOS software development. Extensive tests have been performed and are still in progress. We also worked on the specification and design of advanced functionalities for the XtreemOS PC and cluster flavours. The resulting new features will be progressively integrated in new releases during the second half of the project. We also made progress on the implementation of a first basic version of the mobile device (MD) flavour of XtreemOS for PDAs while specifying in parallel the advanced version of the XtreemOS MD flavour for mobile phones. During the first period, we had identified a set of 14 reference applications to be used to validate XtreemOS technology. During the second period, we carried out our work to make these applications ready for the experimentation of the first XtreemOS release. We set up a permanent test bed for debugging and demonstration purposes. This test bed covers multiple administrative domains and is composed of a set of PCs provided by different partners in different locations.
Scalability and transparency have been identified as two fundamental properties guiding XtreemOS design as described in the first XtreemOS white paper on our vision of a Grid OS.

VO and Security
XtreemOS supports various VO models, used in scientific as well as business scenarios. Within these models, a user can belong to different VOs, and a resource can provide computation power and storage to multiple VOs. User management and resource management are independent in XtreemOS: there is no need to configure resources when new users are registered in VOs. XtreemOS provides Single-Sign-On (SSO): when a user performs a ``login'' within a VO, he receives 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 basic version of XtreemOS is validated for single VO scenarios.

Application Execution Management
The resource discovery mechanism within XtreemOS is based on a distributed information service using P2P technology. Furthermore, services that take decisions never work with a global view of the whole system, but rather use a local viewpoint. For instance, the scheduling will not try to perform a perfect global schedule, but rather generate a job-oriented scheduling within the subset of resources obtained by the resource discovery mechanism.
To ease the use of the Grid services, it is very important to mimic the well-known Linux functionality as opposed to offer different abstractions and functionality, which are more oriented to the Grid. In this same spirit, reliable monitoring which can be reported to the user using familiar tools is vital to provide assurance to users and administrators. This is a feature, which is normally not found in Grid environments.

Data Management
The data management capabilities of XtreemOS are provided by the XtreemFS file system. With XtreemFS we have chosen to implement a full file system and design it for features that are expected from a grid data management system, such as federation of data storage from different administrative domains, data replication and parallel access. XtreemFS fully integrates with the VO concept and allows applications to transparently access files across the whole Grid without any further mediation by middleware layers. Being a real file system, it has full control over any access to the file data and provides real file semantics even in presence of concurrent accesses.
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 include support for speculative transactions, which alleviate network latency and avoid complicated lock management.

Infrastructure for Highly Available and Scalable Services
The infrastructure for highly available and scalable services provides generic services that can be used by XtreemOS-G services and applications running on top of XtreemOS, which underpin the resource management within XtreemOS in a scalable and transparent manner.

- **Distributed Server.** A distributed server is an abstraction that presents a collection of server processes to its clients as a single entity. The address of a distributed server remains stable, even in the case of nodes joining or leaving the application. This technology is exploited in the project both as a support for highly available services (e.g., the job manager or the VO manager) and by those applications willing to make their internal distribution transparent to their clients.
- **Virtual Nodes.** A group of nodes taking part in an application can request to be organized as a virtual node. A virtual node is a fault-tolerant group where each member can take over the task of the others in case of failure. Several types of virtual nodes may be provided, based on active replication, passive replication, and checkpoint/restart mechanisms provided by the XtreemOS operating system. This technology will be integrated with distributed servers to provide a single platform to support fault-tolerant, highly available services and applications.
- **Publish-Subscribe.** A common form of communication between a large number of nodes taking part in a given service or application is publish-subscribe. We will provide a fully decentralized pub/sub communication system that applications can use for their own purpose. The current implementation is based on a hierarchical topic-based mode while later in the project we will evaluate if a content-based approach is also needed.
- **Resource Selection Service.** The Resource Selection Service (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. This allows to scale up to hundred thousands, if not billions, of nodes and to be extremely resilient to churn and catastrophic failures. This service is invoked by the AEM service.
- **Application Directory Service.** The Application Directory Service (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 NodeIDs received by the RSS, related to the resources involved in the application execution. To provide scalability and reliability, DHT techniques and their extensions to dynamic and complex queries will be used.
- **Application Bootstrapping.** Many applications need to have nodes arranged in specific overlay networks (e.g., a torus, a ring) to operate correctly. Application Bootstrapping is a set of libraries, leveraging off epidemic protocols, to make application nodes self-organize to meet the requirements.

XtreemOS API

In general, the XtreemOS API has to serve three classes of applications: existing Linux applications, using POSIX-standardized interfaces, existing Grid applications, using OGF-standardized interfaces, new applications, using functionality uniquely provided by XtreemOS.

We have selected the emerging OGF standard Simple API for Grid Applications (SAGA) as the first draft API for 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.

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 nodes resources to give the illusion that a Linux cluster is a single Linux node. The Posix interface is offered to users allowing the execution of unmodified legacy sequential or parallel applications and system administration tools. Hence, LinuxSSI makes a cluster appear as a single powerful (SMP-like) Grid node. Based on Kerrighed Single System Image (SSI) technology, LinuxSSI provides additional features such as a global customizable scheduler, the checkpoint/restart of process trees, additional reconfiguration mechanisms, and a distributed file system.

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. This scalability factor can be especially relevant, given the enormous number of mobile devices that exist these days. Moreover, mobile Grid applications will be 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.

Due to the current state of mobile Linux market, another key principle of the mobile flavour is portability. XtreemOS will provide 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
Council for the Central Laboratory of the Research Councils	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
T6	IT
Institute of Computing Technology of Chinese Academy of Sciences	CN
Red Flag Software	CN
Telefónica I+D	ES
Universitaet Düseeldorf	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