XTREEMOS, AN OPEN-SOURCE GRID OPERATING SYSTEM TARGETING THE FUTURE INTERNET

Carlini, Emanuele, CNR-ISTI, Pisa and IMT, Lucca emanuele.carlini@isti.cnr.it
Coppola, Massimo, CNR-ISTI, Pisa, massimo.coppola@isti.cnr.it
Laforenza, Domenico, CNR-IIT, Pisa, Italy domenico.laforenza@iit.cnr.it

Abstract

XtreemOS is an EU-funded project of the 6th Framework Programme for developing a Linux-based Operating System to go beyond Grids, integrating as a single computing platform many different kinds of devices, from mobile ones to large clusters. The XtreemOS project tackles the challenge to develop the first planetary-scale, reliable and open computing platform, providing scalable support for Virtual Organizations over secure resource federations. Our ultimate goal is to develop both an open source platform, and to foster a community of developers, that will exploit XtreemOS and maintain it. This motivates many of the project design choices, aimed at easing the interaction with the open-source community and at adopting its best practices.

We describe how the overall computing platform is structured, stacking two main software layers to support the platform locally and at the network level, and including three different version of the Operating System targeting conventional computing resources, Single-System-Image clusters and Linux-powered mobile devices.

The challenge of providing the same ease of use, the same ease of administration, and the reliability of an ordinary Operating System, while exploiting an heterogeneous and ubiquitous platform with high performance and scalability, is the challenge for XtreemOS to represent a reference computing platform for the Future Generation Internet.

We outline the CNR contribution to the project and focus on some of its key components, namely the provision of mechanisms for Virtual Organization management and Security, of Data Management services across federated resources, and of scalable directory services to locate physical and abstract resources.

The XtreemOS Open Source approach leverages on existing technologies that are already part of GNU/Linux systems, reducing the impact on the existing code base and promoting broader acceptance in the open source community. The huge effort in developing such an unprecedented software platform is coordinated through public code repositories, wiki sites and other tools for open collaborative development. The resulting system, whose second public release is scheduled for June 2009, is distributed as a set of additional packages for current Linux distributions.

As the development of the system progresses, the consortium will push its use first within the Grid and SOA communities, focusing on variuos classes of users' needs and interests, ultimately targeting the future Internet of Things and Internet of Services.

- *Easy, flexible execution of large computational simulation.*
- *Set up of large and secure computation-providing platforms for industrial needs (e.g. Service-oriented and Cloud Computing) with no licensing costs and constraints, better suiting public and no-profit organizations.*
- *Pervasive open platforms integrating mobile devices, ubiquitous data storage and high performance computing.*

Keywords: Distributed Operating Systems, Future Internet, Grid Computing, Grid File System, Linux, Open-source Development, Ubiquitous Computing, Virtual Organizations

1 THE XtreemOS PROJECT

XtreemOS [XOS] is a research project of the European Community[1], aiming at producing an Open-Source, Grid-enabled Operating System. The XtreemOS system can federate resources over geographic networks into a large, open platform, in order to solve large-scale computational and data intensive problems in science, engineering and business. The research project, led by the French institution INRIA, spans 4 years and involves 19 research and industrial partners from Europe and China. Among them, the Italian CNR (Consiglio Nazionale delle Ricerche) participates with two of its ITC Research Institutes, ISTI and ICAR.

The XtreemOS consortium aims at producing a new Operating System by leveraging on the huge and ever-growing amount of open source software and tools that converged toward the GNU/Linux Operating System. Of course this cannot happen by merely borrowing code from the network. Productive interaction and exchange with the open-source community is more than essential, it is part of the XtreemOS final goal. The converging of XtreemOS and the Linux open-source community has several implications beside the ability to run existing POSIX and Linux applications onto the XtreemOS platform. We strive to keep compatibility with the open-source software development process, with existing tools and solutions for administration and development, and to attract a community of developers.

Building a Grid-enabled Operating System is an ambitious goal. Modern Operating System provide the user with a robust machine abstraction, that is rich of features and easy to interact with. Isolation among multiple users, security mechanisms, as well as continued support over time and compatibility with existing applications are all essential characteristics of both commercial and open-source Operating Systems. In tackling a composite and pervasive platform, an Operating System must thus go far beyond the features of current Grid middlewares [Foster et al., 2001], providing not only mechanisms to access and control various kinds of dynamical and distributed resources, but also wrapping them under a well-defined and homogeneous abstract machine.

The XtreemOS consortium includes many research institutions, working to advance the state of the art in operating systems, distributed computing and data management, security, networking and Peer-to-peer techniques(P2P). Several European industries contribute with a mix of research and experienced development, and providing immediate usability requirements and significant test-case applications. Last but not least, the contribute of two open-source software distributors Mandriva and Red Flag is essential in making XtreemOS available as a customized Linux distribution.

In the following we discuss the XtreemOS approach toward open source development (§1.1), and the overall XtreemOS System Architecture (§2). We emphasize the crucial aspects of building a Virtual Organization (VO) infrastructure (§2.1), a Grid-aware file system (§2.2), and a scalable and reliable service for resource and service discovery (§2.3). In §3 we survey on the tools and resources used in the XtreemOS project, and report about the current development status and the upcoming 2nd public release of the XtreemOS distribution.

1.1 Open Source Approach

The XtreemOS project basically develops new services and daemons on top of the Linux kernel and the components of modern GNU/Linux distributions. All the software and the source code coming along with XtreemOS is released as open source. The base licensing schemes used for any code developed by the consortium are either BSD or GPL2. However single components of XtreemOS may opt for other open-source licenses, as the GPL3 or Apache licences, possibly to accommodate the use of other third-party software who constrains them to different licensing schemes.

Simply releasing new code under an open license is just the first step for it to be accepted and adopted by a user base. Beside the technological achievement that the project pursues, the consortium follows a

[1]The authors acknowledge the support of Project FP6-033576, Building and Promoting a Linux-based Operating System to Support Virtual Organizations for Next Generation Grids (2006-2010).

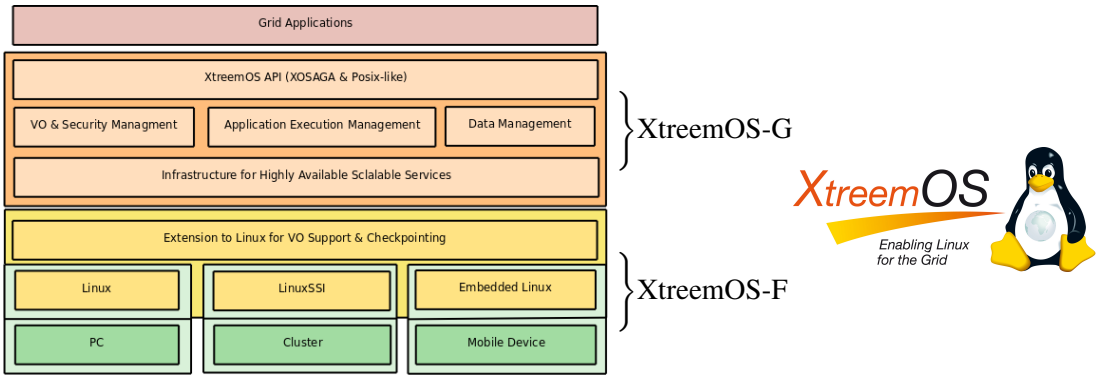

Figure 1. Layering of the XtreemOS architecture.

precise strategy, with respect to kernel and operating system changes, aiming at maximizing GNU/Linux developer acceptance of the XtreemOS contributions to the Linux Operating System.

1. Exploit existing Linux technologies, possibly extend them to achieve XtreemOS peculiar targets.
2. Add daemons or programs, to provide new features complementary to the GNU/Linux standards.
3. Patch existing open source code if unavoidable, in a local and optional way (e.g. by adding kernel modules, or Pluggable Authentication Modules [Samar and Schemers, 1995]).
4. Both ease of use (e.g. graphical frontends) and compatibility with basic standard tools for system use and development (e.g. command line and scripting interfaces) have to be provided.

The rationale is that changes to a complex component, whose development is critical for the system, are unlikely to be welcome by its developer community, especially if changes are not immediately useful to a large fraction of users. Instead, file access control lists help managing users' files across a Grid, IPV6 is exploited in supporting mobility and resiliency of computing, and existing kernel modules provide local VO support (see §2.1). These standard technologies bear an obvious advantage over custom modifications to the Linux sources, as the former are already included in Linux Distributions, and are maintained by active communities.

Being able to exploit separately a subset of the XtreemOS components is another key point of the project strategy. As some system components and solutions can be even adopted separately from the whole, possibly also on non-Linux platforms, and they provide useful services, larger user and developer communities will grow around them. XtreemOS and its adopters will gain support from this increased community size. In the project scenario, the Mandriva and Red Flag organizations have the task of helping the application of open source development practices and tools. They apply their experience about monitoring of the development process, management, integration, testing and deployment of large open-source software collections, up to to the actual packaging as an XtreemOS Linux distribution.

2 SYSTEM ARCHITECTURE

The XtreemOS architecture model comprises two basic layers (Figure 1), following a simplified version of the Open Grid Service Architecture (OGSA) stack [Foster et al., 2001]. XtreemOS provides support for VOs on all software layers involved, starting from the operating system level, passing through the services, and finally to the direct applications support.

The low-level layer, **XtreemOS-F**, comprises all XtreemOS-specific services which need to be performed on the single system. They represent a superset of the Linux OS, extending the kernel behaviour in order to provide grid functionalities to different hardware scenarios such as simple workstations, clusters or mobile devices. Part of the development effort required by this layer is summarized in §2.1. **XtreemOS-G**, the upper layer, ties together all the machines in the platform. Here the focus is on ex-

ploiting the infrastructure connecting the various nodes, and providing a software infrastructure for managing them as a whole. The biggest effort has been put in scalability, combining hierarchical and peer to peer distributed solution in order to allow the practical use of VO with at least ten thousands of machines and users. XtreemOS provides Grid-wide Operating Systems services on top of this two software layers, through the **XOSAGA** implementation of the SAGA API (Simple API for Grid Applications).

Future Internet will see an unprecedented range of devices integrated over the net: large clusters providing service oriented computing, millions of personal workstations of increasing computational power and efficiency, and also millions of personal mobile devices whose capabilities and flexibility of use are steadily growing up and reaching those of full-fledged computers. The emergence of such a composite platform has seen some major Operating System providers competing for the market with mobile versions of their OS. The Linux platform has been initially slowed down in this process by the lack of a strong, explicit commitment from industries to provide porting on widespread hardware. Finally, however, several resource-constrained Linux distributions have started to spread, and Google's commitment to the Android software platform is bringing massive Linux support in the cell-phone world.
XtreemOS early accomodated this trend, by including a wide range of machines in its planned platform, and by targeting three different classes of hardware with specifically optimized XtreemOS "flavours", all of them sharing the same principles and being able to interoperate.
XtreemOS-PC The base version of XtreemOS, that can integrate any number of machines into the platform either as a resource nodes (providing resources to their VO) or as client nodes (belonging to the platform, but solely used to access the real resources).
XtreemOS-SSI XtreemOS-SSI specifically targets homogeneous clusters by employing a Single System Image (SSI) Linux kernel which is an evolution of Kerrighed-Linux. SSI Kernel instances are interconnected and coordinated among themselves, sharing physical kernel resources. An SSI cluster effectively provides the user with the illusion of a single, huge and powerful shared-memory Linux system.
XtreemOS-mobile A lightweight XtreemOS flavour targets mobile phones and PDAs running Linux distributions. These "client" machines do not provide computation or storage to the XtreemOS platform, but allow users to access their own data and steer applications ubiquitously.

2.1 Scalable VO Support

In XtreemOS, credential management integrates grid-level authentication within system-level authentication. The overall design of VO and security support aims at the best combination of local support and properly structured Grid-level support. Dynamic mapping of VO level entities onto local entities exploits currently adopted technologies like Pluggable Authentication Modules (PAM), the Kernel Key Retention Service (KKRS) and the Name Service Switch.
The Grid-level implementation of the security, authentication and access mechanisms consists of a Grid infrastructure (local daemons, communication protocols and certificate and information servers) building up hierarchies of trust and user information management, typically one per VO or group of VOs. Users and resources may belong to multiple VOs, and VOs can be created and destroyed with low overhead. A web interface fo VO creation and administration is also available. The resulting architecture of the VO and Security support of the XtreemOS platform is described in [Coppola et al., 2008].
From the user viewpoint, running applications on top of XtreemOS is as simple as logging into a console and provide his Grid identity certificate, and then runnig ordinary and XtreemOS-specific commands, both being enabled to Grid-level execution.

2.2 XtreemFS

XtreemFS [XFS] is the distributed file system which XtreemOS leans on. The fundamental role of XtreemFS within XtreemOS is to provide an abstraction of the home directory to the generic user, regardless of its current working site, and in fact hiding the complexity of a geo-distributed file system.

XtreemFS has been designed for deployment in wide-area networks and Grids, providing ubiquitous access to user data with high performance and reliability, and supporting data replication strategies and parallel access. XtreemFS provides a fully POSIX-compliant interface to the user, but it is an object-based filesystem. Files are managed as collections of elementary data objects over the network, memorized inside *storage servers*. Complementary information is retained by *metadata servers*, that store object management information and attributes of whole files, like access time and permissions, and perform user authentication. A *central server* maintains the registry of all the servers composing an XtreemFS infrastructure. The CNR team contributed to the first implementation of the storage servers, and is currently involved in testing XtreeemFS POSIX compliance.
Both the metadata and central servers are essentially special cases of a directory service, and thus they can also exploit other similar services provided within XtreemOS such as the one described in 2.3. The XtreemFS architecture is fully described in Hupfeld et al. [2008].

2.3 Service/Resource Discovery System

The Service/Resource Discovery System (SRDS) is a key component of the XtreemOS-G layer (See Figure 1) developed by the CNR team. SRDS offers to other XtreemOS components and applications a scalable and efficient directory service capable of managing entities described by structured, static and dynamically changing attributes [XOS, See documents D3.2.4 and D.3.2.8].
The main challenge is to cope with an highly dynamic environment, where entities (applications, users, computing nodes) can unpredictably join or leave the system. The shared resources change over time, and the system needs to keep the information as updated as possible. Many P2P networks enjoy scalability and reliability, thanks to the absence of centralization points and the spreading of information. The SRDS builds upon P2P key-value storage networks in order to provide more advanced search functionalities. Local SRDS processes join P2P overlays over the Grid.
The heterogeneity of entities and services is also a driving factor in SRDS design. Its modular architecture can dynamically manage different overlays at the same time, allowing to configure and exploit their characteristics (e.g. replication degree, P2P strategy) paying the relative overhead only when actually needed (e.g., enhancing fault tolerance by increasing the DHT replication overhead).
Currently, three different P2P overlays are integrated within SRDS. The Resource Selection Service (RSS) is a hierarchical P2P network built on top of a gossiping-based overlay. It is developed by VUA (Vrije Universiteit Amsterdam) inside the XtreemOS project and is used for resource discovery w.r.t. a static description of the resources. To support general-purpose directory service we exploit two different distributed hash tables (DHTs). Scalaris [Scalaris] is a DHT with transactional capability, that is developed and used inside the XtreemOS project. Overlay Weaver (OW) [Shudo et al., 2008, OW] provides a common DHT API that can be itself extensively configured at run-time, e.g. in the the transport protocol, the database implementation and routing algorithms used.

3 DEVELOPMENT STATUS, TOOLS AND PUBLIC RESOURCES

XtreemOS code has begun its development within an EU-funded project. It was clearly foreseen in advance a need of integration between the initial goal-oriented research and development activity, and the classical forms of open-source software development. To this purpose, proper tools and infrastructures have been adopted since the beginning (see Table 1). Developed code is officially hosted on the INRIA gforge SVN-repository and bug-tracking system. Some of the modules, most notably XtreemFS [XFS], are actually developed on other public repositories like Google's, and are daily synchronized to the main repository. We also use the other typical tools of collaborative development: mailing lists, IRC channels, a WWW portal [XOS] and a Tikiwiki site. Some of these tools are reserved for project developers and for administration, while other ones are intended to outlive the project and support the community.

After a first phase of integration alpha-testing, XtreemOS for non-mobile devices reached a first stable organization and was **first released** in December 2008. The SVN repository has been since then open to the public (read-only), and an open developer mailing lists has been created. The 1.0 and 1.1 releases are distributed as CDs and as special set of packages compatible with the Mandriva 2008 distribution.
The **Mobile flavour** of XtreemOS, first released in April 2009, is compatible with the Maemo 2008 and Angstrom 2007.12 distributions. It provides automatic and manual Grid login, automounts XtreemFS home volumes, launches processes and supports a subset of the SAGA API.
Release 2.0 of the XtreemOS system includes all the essential components of the XtreemOS architecture in Fig. 1. It eases installation and configuration, and is packaged on top of Mandriva 2009. After the *feature freeze* phase in May, the new release is currently in the *code freeze* packaging phase. CD images, as well as ready-to-go disk images for Virtual Machine software, are expected to be released June 2009. After release, a new IRC channel, devoted to user support, will be activated on (`irc.freenode.net`). Further releases of XtreemOS are planned to follow about 6 months apart from each other.

Community tool	**Web reference**
Official WWW and Blog	http://www.xtreeemos.eu https://www.xtreemos.org/blog
TikiWiki site	https://xtreemos.wiki.irisa.fr/
Developer mailing list	http://lists.gforge.inria.fr/cgi-bin/mailman/listinfo/xtreemos-developers
IRC user support	channel to be activated on irc.freenode.net
XtreemOS Development site **Public SVN**	https://gforge.inria.fr/projects/xtreemos/ http://gforge.inria.fr/plugins/scmsvn/viewcvs.php/?root=xtreemos
Mirror for CD Images	ftp://ftp.free.fr/mirrors/ftp.mandriva.com/MandrivaLinux/devel/iso/xtreemos/
XtreemOS Mobile distribution	ftp://ftp.free.fr/mirrors/ftp.mandriva.com/MandrivaLinux/devel/xtreemos/2008.0/mobile/
Architecture documents	http://www.xtreemos.eu/publications/project-deliverables/

Table 1. XtreemOS users' and developers' resources available on the network.

References

M. Coppola, Y. Jegou, B. Matthews, C. Morin, L. P. Prieto, O. D. Sanchez, E. Y. Yang, and H. Yu. Virtual Organization Support within a Grid-wide Operating System. *IEEE Internet Computing, Sp.Issue on Virtual Organization Management*, 12(2), March/April 2008.

I. Foster, C. Kesselman, and S. Tuecke. The anatomy of the grid: Enabling scalable virtual organizations. *Int. J. High Performance Computing Applications*, 15(3):200–222, August 2001.

F. Hupfeld, T. Cortes, B. Kolbeck, J. Stender, E. Focht, M. Hess, J. Malo, J. Marti, and E. Cesario. The XtreemFS architecture—a case for object-based file systems in Grids. *Concurrency and Computation: Practice and Experience*, 20(17):2049–2060, 2008.

OW. Overlay Weaver, An Overlay Construction Toolkit. Public web site, 2006–. URL `http://overlayweaver.sourceforge.net/`.

V. Samar and R. J. Schemers. Unified Login with Pluggable Authentication Modules (PAM). Open Software Foundation RFC 86.0, October 1995.

Scalaris. Scalaris, Distributed Transactional Key-Value Store. Public web site, 2009. URL `http://code.google.com/p/scalaris/`.

K. Shudo, Y. Tanaka, and S. Sekiguchi. Overlay weaver: An overlay construction toolkit. *Computer Communications*, 31(2):402–412, February 2008.

XFS. XtreemFS public web site, 2006–. URL `http://www.xtreemfs.com/`. Development site: `http://code.google.com/p/xtreemfs/`.

XOS. Xtreemos public web site, 2006–. URL `http://www.xtreeemos.eu`.