



# XtreemOS: an Operating System for Next Generation Grids

**Christine Morin**

**Centre de recherche INRIA Rennes - Bretagne Atlantique**

**XtreemOS Scientific coordinator**

**[xtreemos-info@irisa.fr](mailto:xtreemos-info@irisa.fr)**





## Outline

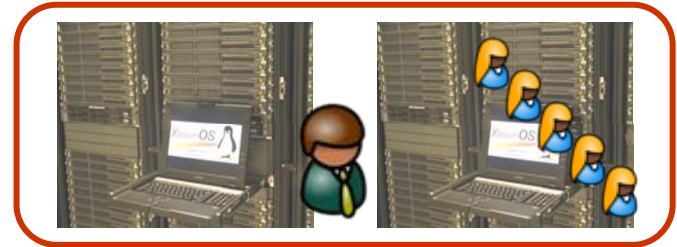
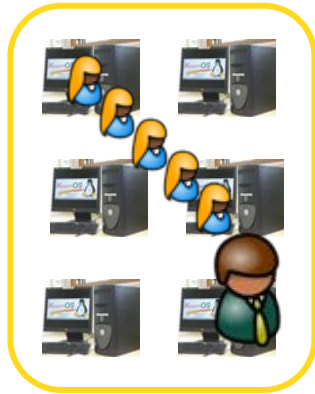
- ❑ Virtual organizations & Grid computing
- ❑ Overview of XtreemOS project
- ❑ XtreemOS services
- ❑ Conclusion



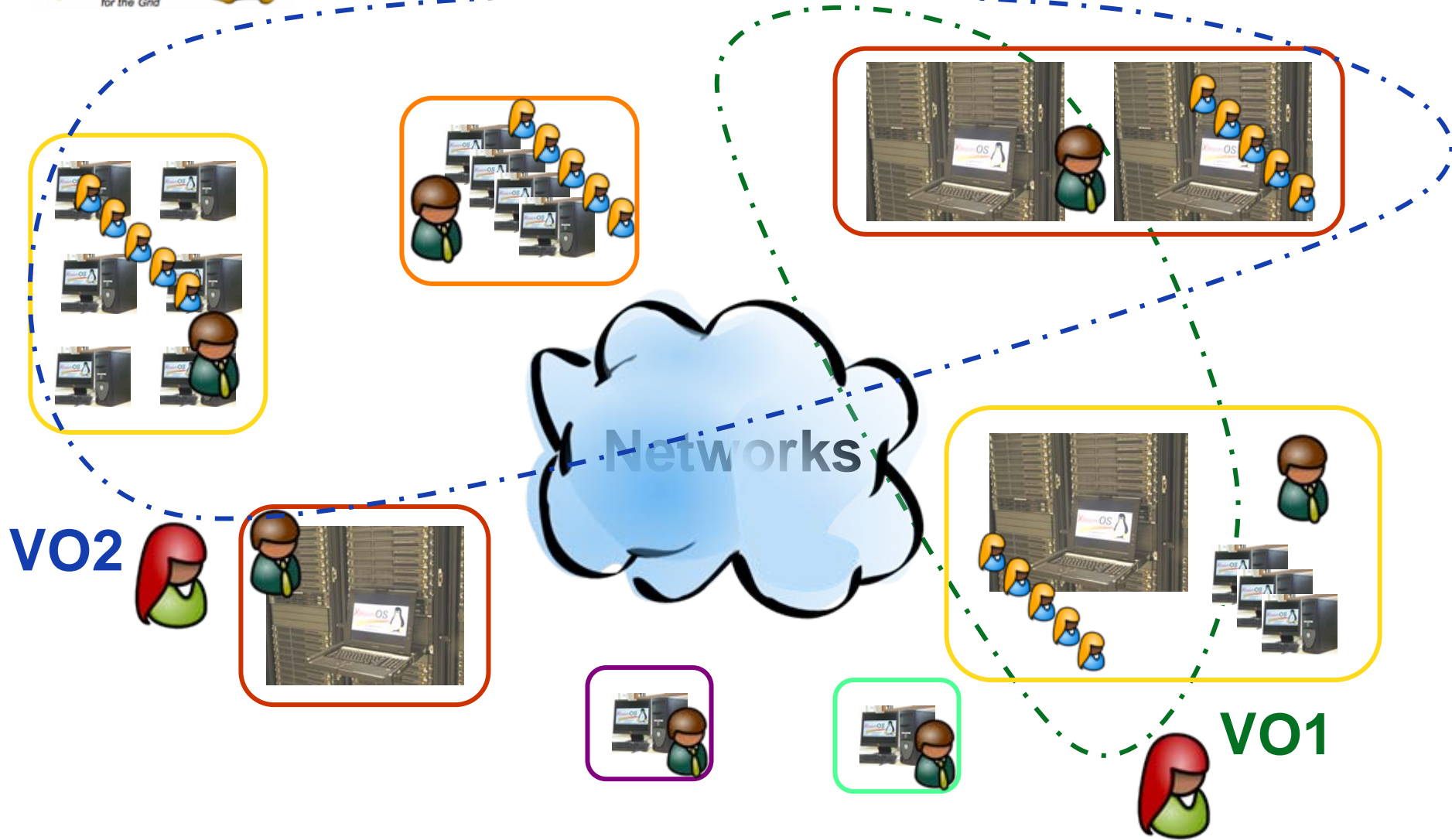
# Virtual Organization Concept

- **Temporary or permanent alliances of enterprises or organizations**
  - sharing resources, skills, core competencies
  - to better respond to business opportunities or large scale application processing requirements
  - whose cooperation is supported by computer networks

# Large Scale Dynamic Grids

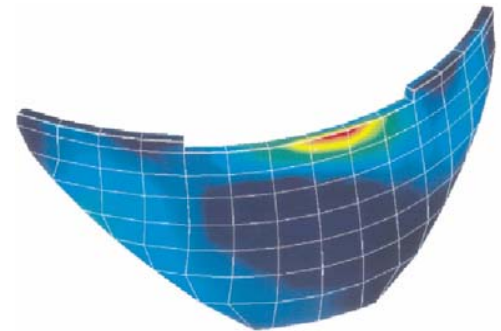


# Virtual Organizations (VO)



# Applications

- ❑ **Computing resources used on demand**
  - Many applications of moderate size
  - Many users
- ❑ **Distributed simulation of physical behaviour**
  - Code coupling
- ❑ **Business services**





# Why it is difficult to use a Grid

- ❑ **Large scale distributed system**
  - Very large number of heterogeneous resources
- ❑ **System used by multiple users simultaneously to run different applications**
  - Very large number of users
- ❑ **Distributed system whose resources belong to multiple institutions**
  - Multiple sites in different autonomous administrative domains
- ❑ **VO Dynamicity**
  - Resources may join or leave the Grid at any time
  - Resource and network failures
  - Changes in VO membership



# Harnessing large scale dynamic Grids

⇒ Easy and scalable VO management

⇒ Efficient, secure, reliable application execution

⇒ Ease of use & programming



# State of the Art

## ❑ Systems offering Grid support

### – Minimal infrastructure (Globus)

- Burden on system administrators, programmers and users

### – Global infrastructure (Xtremweb)

- Lack of flexibility
- Target specific class of applications

## ❑ Implementation level

### – Middleware (Globus, PUNCH, Unicore)

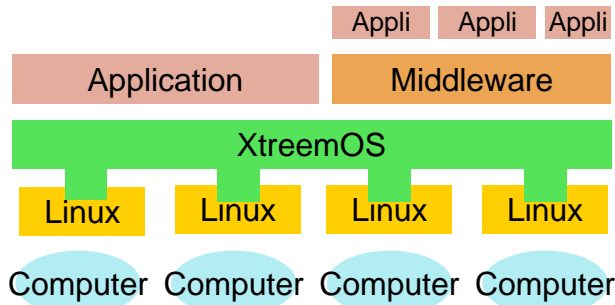
- Performance & security issues due to multiple layers
- Multiple rapidly evolving standards
- Legacy applications need to be modified

### – Grid OS (9Grid, GridOS)

- Implementation of core functionalities to simplify middleware
- No Grid OS currently offers a full set of highly available scalable services



# XtreemOS Project Objectives



□ Design & implement a reference open source **Grid operating system** based on **Linux**

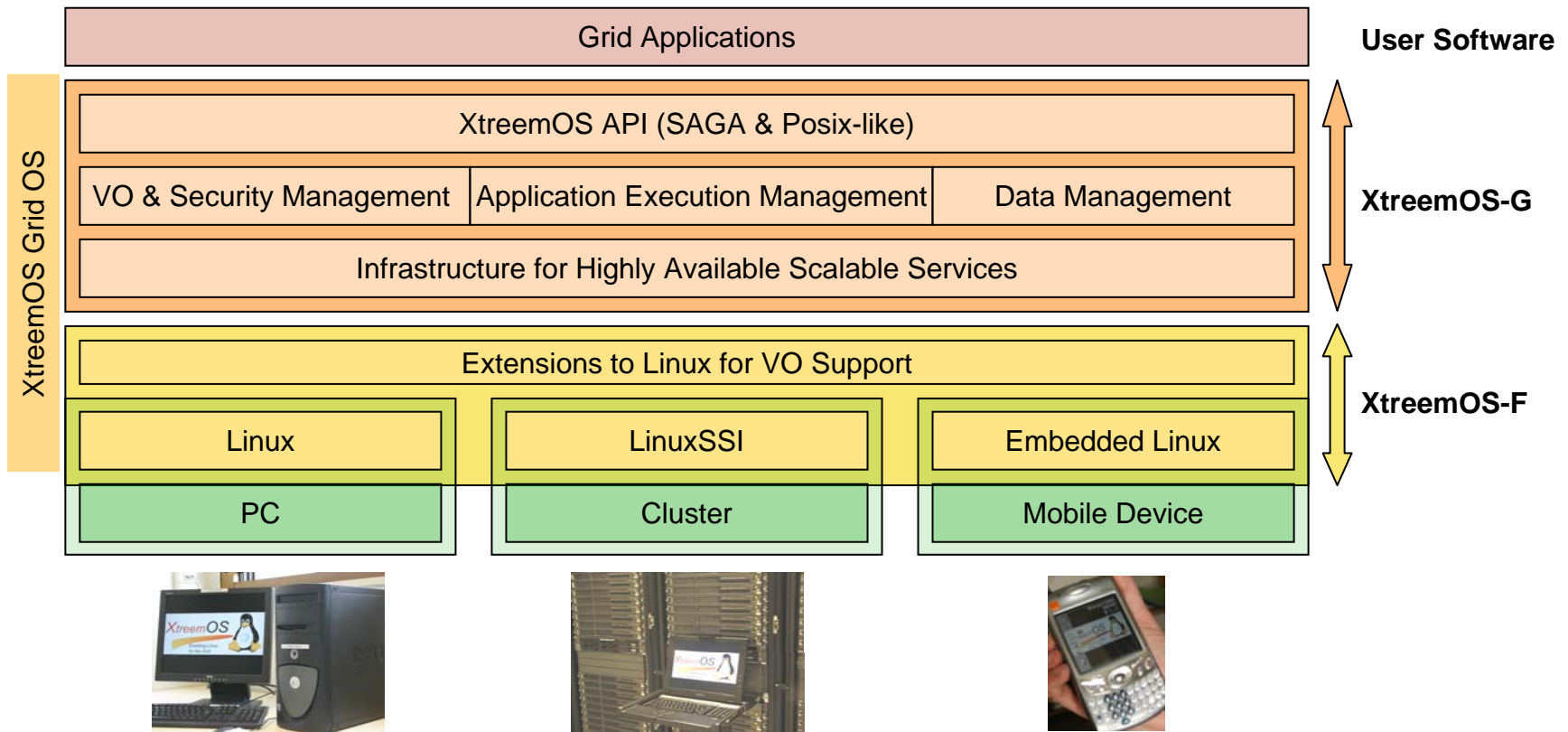
- Get around overheads and security pitfalls brought by layers in existing Grid middleware
- Provide **native VO support**
  - In a **secure** and **scalable** way
  - Without compromising on **flexibility** and **performance**

□ **Validate** the XtreemOS Grid OS with a set of **real use cases** on a **large Grid testbed**

□ **Promote** XtreemOS software in the **Linux community** and **create communities of users and developers**



# Overall XtreemOS Architecture





# XtreemOS API

## ❑ Challenges

- Linux applications should run with little (no) modifications
- Grid applications should run with little (no) modifications
- XtreemOS functionality must be provided to applications

## ❑ SAGA, the Simple API for Grid Applications

- Very close to POSIX
- Compliant to existing OGF standards (DRMAA, JSDL, BES, GridRPC)

## ❑ Implementation of a SAGA engine with Posix Adaptors



# VO & Security Management

## □ Challenges

- **Interoperability with diverse VO framework and security models**
- **Flexibility in policy languages**
- **Scalability of management of dynamic VO**
- **Accurate isolation**
  - Strict access control from service level to system object level
  - Monitoring and logging OS service usage and system object access
    - Audit log must refer to user credentials and be securely provided to the resource owner and the VO manager



# VO and Security

## □ VO level

### – VOM service

- Distributed information management for membership tracking and accounting of users and resources
- Security services

## □ Node level

### – Extended Linux OS

- Mechanisms for recognizing, controlling, and enforcing usage of Grid entities



# VOM Service

## **Identity Service**

Generates and manages globally unique VO IDs and user Ids

## **A Virtual Organization Membership Service**

Checks whether a user is a member of a specified VO. Used by the CDA before issuing an XOS-Cert, and by other subsystems needing to check VO membership of a user

## **A Credential Distribution Agency**

Issues users with VO security credential for accessing grid-wide services and resources

XtreemOS uses X.509 v3 certificates (the 'XOS-Cert')

## **Attribute Service**

Provides users with VO attributes. Used to carry information relating to controlling access to resources, and to allow VO nodes to map global user IDs to local UIDs/GIDs

## **Policy Service**

Provides services such as Policy Information Points and Policy Decision Points



# Node Level VO Support

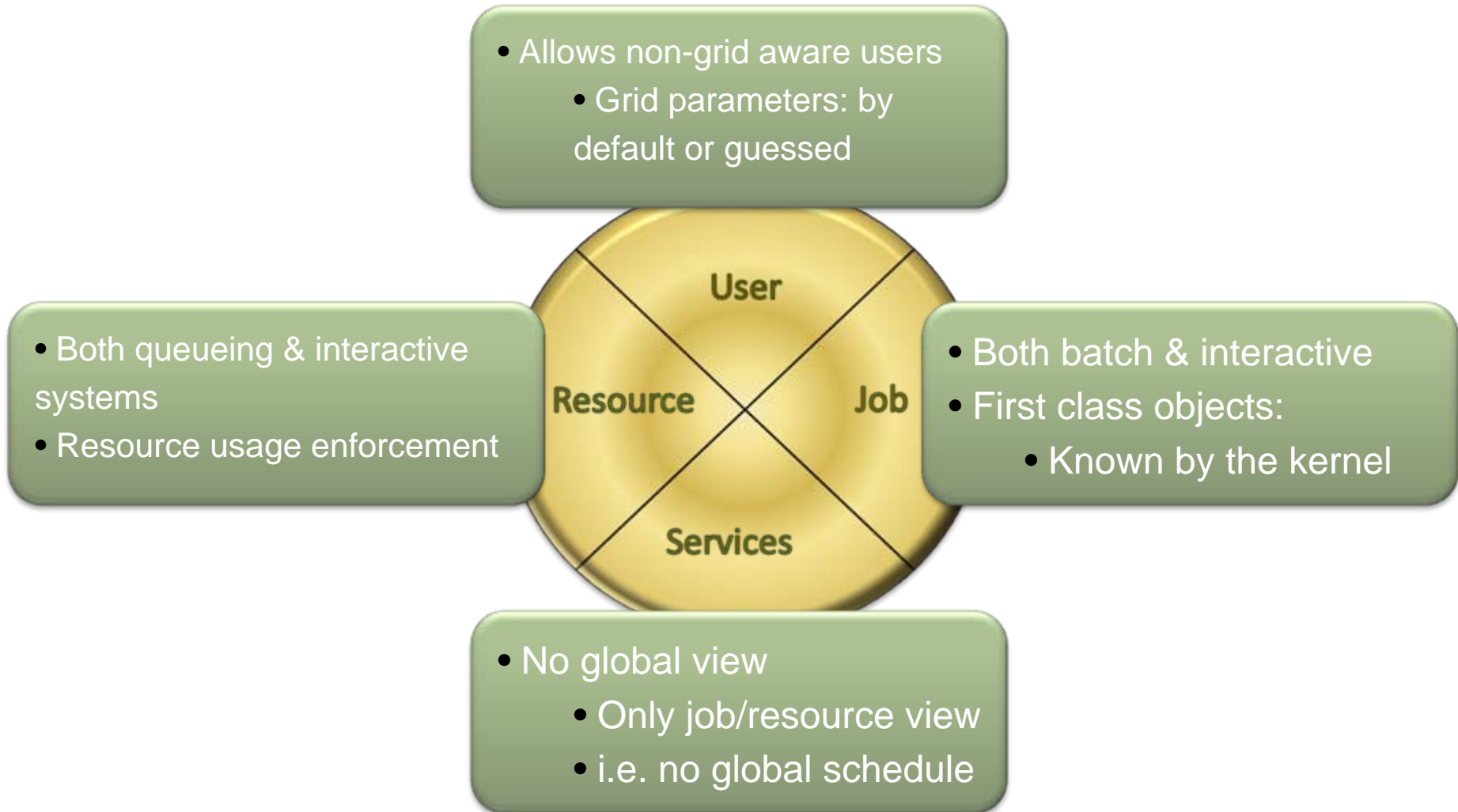
- ❑ Mapping from grid user credentials (User ID, VO id, attributes) to local user credentials (uid, gids)
- ❑ Enforcement of VO/local access control policies and resource usage constraints
- ❑ Isolation of multiple VO accesses on the same node
  - By dynamic creating local accounts for isolation
- ❑ Internal interfaces are exposed via
  - PAM APIs (*libpam*)
  - NSS APIs (*libc*)
  - Kernel Key Retention Service APIs (*libc*)



# Application Execution Management

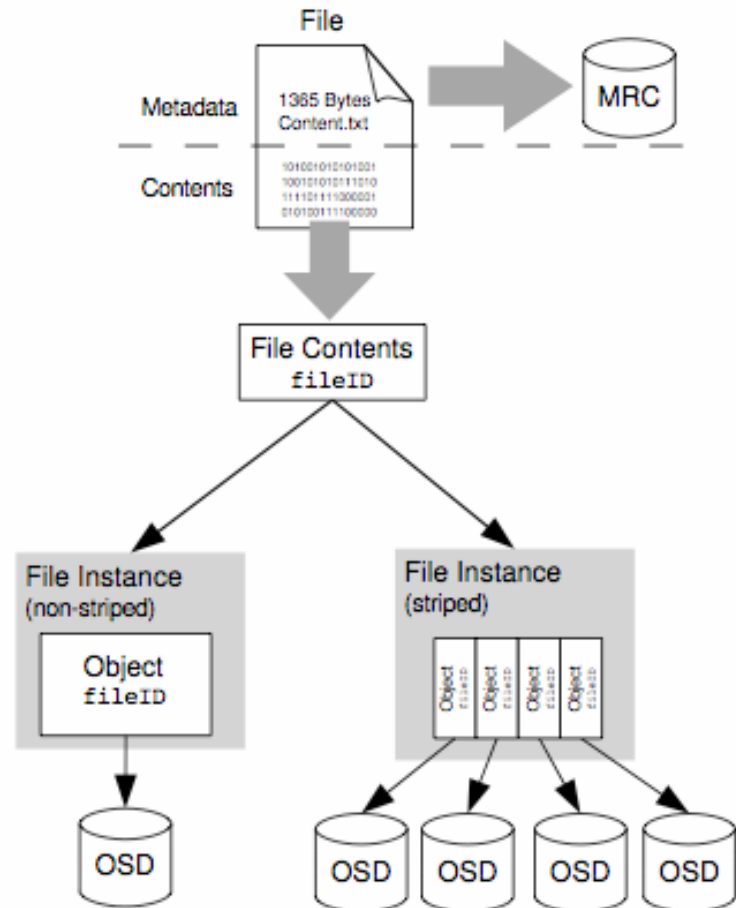
- ❑ **Objective: provide functionality to execute jobs**
  - Services to start, monitor and control applications
  - Services to select and allocate resources
- ❑ **Challenges:**
  - Scaling to  $10^6$  of nodes/users/jobs/...
  - Heterogeneity of resources
  - Benefit from integration with other components
    - Synergies
    - Better accuracy in information

# Comparison with SOA



A distributed file system (POSIX interface):

- Federated installations over multiple VOs
- Designed for cross-org. high-latency WANs
- File replication (control interface for AEM)
- File striping and redundancy
- Metadata replication
- Coordinated client-side data caching with advanced semantics (interface: mmap)





# Infrastructure for highly available scalable Services

- ❑ **Mechanisms for transparent fault tolerant service replication**
  - Based on IPv6
  - AEM: job controller
  - VOM: security services
- ❑ **Publish/subscribe communication**
  - XtreamFS: reliable dissemination of meta data changes
- ❑ **Directory service**
  - AEM: job directory in AEM
  - XtreamFS: global index of file system volumes
- ❑ **Resource discovery**
  - Multi-range queries
  - AEM: find resources matching job requirements

➔ Based on **overlay networks** (P2P technology) for **scalability**



# Cluster Flavour

## ❑ Objectives

- **Efficient** execution of applications requiring a large amount of resources
  - make efficient use of the cluster hardware
- Provide a **simple interface**
  - make resource distribution transparent

## ❑ Single System Image Technology

- A SSI cluster looks like a **single powerful PC** for software executed on top of the OS
  - Legacy applications can be executed on a SSI without modification or recompilation
- Ease of management: a **single distributed OS** managing all cluster nodes

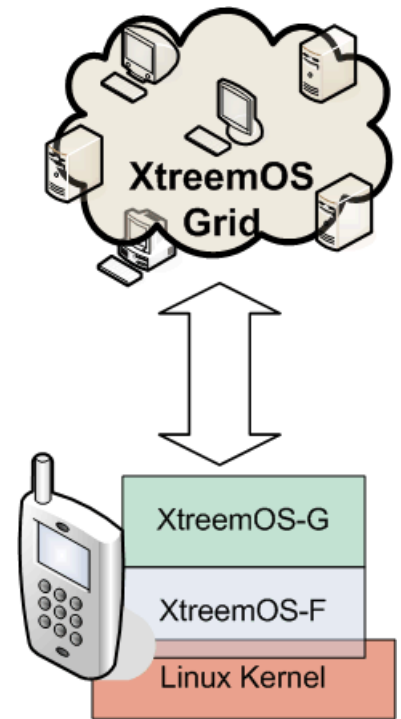


# Cluster Flavour

- ❑ Leverage the **open source Kerrighed** SSI cluster OS originally developed by INRIA in collaboration with EDF R&D
  - Extension to Linux kernel
    - Kernel modules + patch
  - Most recent version Kerrighed 2.1.0 based on Linux 2.6.20
    - <http://www.kerrighed.org>
- ❑ **Linux SSI**
  - KDFS distributed file system exploiting disks attached to compute nodes
  - Customizable scheduler
  - Parallel application checkpointing
  - Scalable SSI

## Mobile Device Flavour

- ❑ Provide **support for VO activities** in a mobile and ubiquitous scenario, by integrating those functionalities in a **Linux** distribution for **mobile devices** (PDAs and Mobile Phones).
- ❑ Composed of two layers:
  - **XtreamOS-F**: Foundation layer, low level, integrated in the OS (kernel, modules...)
  - **XtreamOS-G**: Services layer, a subset of all XtreamOS services
- ❑ Two versions:
  - **Basic** (PDAs): more stable platform, more processing and storage power
  - **Advanced** (Smartphones): more optimizations needed, unstable (but promising) future/market





# Mobile Device Flavour

- ❑ Guide the development of new features in Mobile Grids
- ❑ Help spotting potential “killer apps”
- ❑ Just to name a few:
  - eLearning
  - eHealth
  - Crisis management
  - eBusiness (mobile services integration)
  - In general, services requiring more resources than available on MDs (i.e. voice recognition algorithms, biometric identification databases,...) and access to resources from different organizations.
  - ...



# Conclusion & Perspectives

- ❑ **Initial architecture design of XtreemOS Grid OS**
  - a consistent set of scalable and highly available services based on kernel level mechanisms
  - Native VO support
- ❑ **On-going implementation of a first prototype**
  - First fully integrated XtreemOS release planned by May 2008
    - Some individual components released by the end of 2007
- ❑ **Future work**
  - **Refinement of the initial design**
    - Iterative approach based on feed-back from **experimentation with use cases**
  - **Security analysis of XtreemOS**
- ❑ **More information: <http://www.xtreemos.eu>**



# XtremOS Consortium

