XtreemOS European Project: Achievements & Perspectives

Christine Morin
XtreemOS scientific coordinator
Head of Myriads research team
INRIA Rennes - Bretagne Atlantique

CCGSC 2010 – Flat Rock, NC
XtreemOS IP project
is funded by the European Commission under contract IST-FP6-033576
Distributed operating system for large scale dynamic Grids

- “Operating system” approach
  - Comprehensive set of cooperating system services
- Ease of use
  - “Bring the Grid to standard users”
    - Unix system interface
    - SAGA programming interface

- Scalability
  - Dependable system
XtreemOS Flavours
- Open source development
- Release 2.1.1 **packaged** for Mandriva and Asianux Linux distributions
  - Packaging in progress for Debian, Ubuntu, Open Suse
- Ready to use VM images for KVM & Virtual Box
- Open testbed for the community
  - Test your applications without installing XtreemOS
- Tool for **automatic configuration** of the system
  - Deployment on Grid’5000
Overview of Applications

19 applications demonstrating and evaluating XtreemOS from the perspective of industrial and academic end-users

- Electromagnetics
- CAE
- Virtual Reality
- Mobile applications
- Fluid Dynamics
- Particle Physics
- Cloud Computing
- Enterprise solutions
- Optimization
- Cloud Computing
- Optimization
Some Contributions

- **XtreemOS system services**
  - VO & security management
  - XtreemFS Grid file system
  - Job & resource management
  - OSS object sharing system

- **XOSAGA**
  - SAGA programming interface

- **Virtual Node approach**
  - Highly available applications & system services
VO & Security Management

- Scalable VO management
  - Independent user & resource management
  - On-the-fly mapping of Grid credentials to Linux user accounts
  - Customizable isolation, access control and auditing

- Secure and reliable application execution
  - Fine-grained control of resource usage
- **Improved usability**
  - Local resource administrator: autonomous management of local resources
  - VO administrator: flexible management of credential and VO policies
  - End user: login as a Grid user into a VO
    - On-line certificate distribution
    - Single sign-on & delegation
      - System services services trust each other (“operating system approach”)
      - A trusted credential store service associated to each user session
      - There is not need of proxy certificates
Federating storage in different administrative domains
XtreemFS Features

- Posix compatible file system (API, behaviour)
- Provide users a global view of their files in a Grid
  - Each XtreemOS user has a home volume in XtreemFS
  - Transparent location-independent access to data
- Consistent data sharing
- Access control based on VO member credentials
- Autonomous data management with self-organized replication and distribution
- Advanced metadata management
Job & Resource Management

- Job self-scheduling
- Decentralized resource discovery based on overlays
- Resource reservation
- Unix-like job management
- Support for interactive jobs
- Accurate & adaptable monitoring
- Job checkpoint/restart & migration
- Automatic management of the user specified fault tolerance strategy
  - Handling checkpoint/restart for Grid applications
XtreemGCP Service

- **Generic service**
  - Different levels to implement fault tolerance
    - In the application code
    - In a programming environment (MPI …)
    - At system level transparently to the application
    - VM Suspend/restart
  - Different backward error recovery protocols
    - Checkpoint based (coordinated, independent, message induced, …), message logging based (pessimistic, optimistic, causal, …),…
  - Different technologies for process group checkpointing
    - Some do not handle all resources
Process Group

Checkpointers

- Condor
- DMTCP & MTCP
- BLCR
- Epckpt
- KMU
- TICK
- MCR
- CHPOX
- DCR
- zap
- CRAK
- CLIP
- libckpt
- tmPVM
- Dynamite
- VMWare
- player
- LAM/MPI & BLCR
- CoCheck
- VMADump
- Ckpt
- LinuxSSI
- OpenVZ
- SCore
- Linux-native
- CP/R
- UCLiK
User Perspective

- **User/application commands**
  - `$xjobcheckpoint JobID`
  - `$xjobrestart JobID CPversion`

- **JSDL file extensions**
  - Extended by checkpointing tags
  - Checkpointer requirements
  - Protocols and parameters
  - ...
<JobCheckpointing>
  <Initiator>System</Initiator>
  <ProtocolManagement>
    <Name>CoordinatedCheckpointing</Name>
    <Parameter>1 hour</Parameter>
  </ProtocolManagement>
  <FileManagement>
    <ReplicationLevel>5</ReplicationLevel>
  </FileManagement>
  <JobCheckpointerMatching>
    <MultiThread>Yes</MultiThread>
    <Sockets>Yes</Sockets>
  </JobCheckpointerMatching>
</JobCheckpointing>
XtreemOS-GCP Architecture

Grid level

Job Checkpointer
(Job Manager extension)

Node Level

Job-unit Checkpointer
(Execution Manager extension)

Job-unit Checkpointer
(Execution Manager extension)

Common Checkpointer API

SSI-Translib

BLCR-Translib

LinuxSSI Kernel Checkp.

BLCR Checkpointer

XtreemOS-SSI cluster

XtreemOS PC
• Provide a uniform access to different checkpoints
  • translib library
• Translate jobs in Linux process groups
• Translate user credential in Linux user account
• Provide callbacks to applications
  • Processed during checkpoint and restart operations
  • Allow applications to optimize checkpointing
  • Used to drain communication channels
To which extent must existing checkpointers be adapted to support various checkpointing protocols?

We need the following sequences

- Stop
- Checkpoint
- resume_cp

- Rebuild
- resume_rst

Common Checkpointer API
Callback Management

- Implemented in the generic part of translib
- Called before and after a checkpoint and after restart
- Common API for application callback registration

**Usage**
- Application optimizations
- Complement checkpointer incapabilities
- Checkpointing communication channels
Fault tolerance information stored in XtreemFS Grid file system
  - checkpoint replication
  - checkpoint can be accessed from any Grid node

Resource conflict avoidance at restart

Management of security issues regarding the use of fault tolerance information
Current Status

- **XtreemGCP fully integrated in XtreemOS**
  - PC and cluster nodes
  - Sequential, parallel and distributed applications
  - System level checkpointing
- **Kernel checkpointer supported**
  - BLCR, OpenVZ based checkpointer, native Linux checkpointer, Kerrighed checkpointer
  - Call back mechanisms
- **Protocols supported**
  - Coordinated checkpointing (for job migration)
  - Independent checkpointing
What’s coming next?
What’s coming next?

• **Sustainability** of the XtreemOS Grid technology
• **Cloud computing** - Contrail EC funded R&D project
Feasibility studies (2008 - …)
- Extending an XtreemOS Grid with resources gathered from Clouds
- Hbase on top of XtreemFS
- Picture sharing application over XtreemOS in a cloud
- XtreemOS as a system to manage IaaS Clouds
Contrail European Project

• Objectives
  ▪ Design, implement, evaluate and promote an open source system to federate computing resources from different providers in a single cloud easy to access for users

• Approach
  • Vertical integration of
    ▪ Infrastructure-as-a-Service services
    ▪ Runtimes and high level services providing the foundations for Platform-as-a-Service services
Contrail in a Nutshell

High Performance Real-Time Scientific Data Analysis
Distributed Provision of Georeferentiated Data
Large-scale Code Analysis for Open Source software
Multimedia Processing Service Marketplace

Use cases and Exploitation
SP5

Map/Reduce
Bag-of-Tasks
Web-servers
key-value store
structured storage
SQL

Platform-as-a-Service
SP3

Cloud User Interface + API
well known abstractions
(POSIX API, x86 ISA, IP Network)

Core Virtual Infrastructure Layer
SP2

Resource Selection
Monitoring
Accounting
SLAs

Network
Compute
Storage

Security

Cloud Federation Management
SP1

IaaS Federation
(distr. registry, VO management, identity mgmt)

Providers

Integration, Validation and Testing
SP4
Contrail European Integrated Project

- **Coordinator**
  - INRIA, France

- **Academic partners**
  - CNR, Italy
  - STFC, UK
  - Vrije Universiteit Amsterdam, The Netherlands
  - ZIB, Germany

- **Industrial partners**
  - CONSTELLATION, UK
  - GENIAS, The Netherlands
  - HP, Italy
  - TISCALI, Italy
  - XLAB, Slovenia

- **Starting date:** October 2010
- **Duration:** 3 years
- **Budget:** 11,4 M€
- **EC funding:** 8,3 M€
Acknowledgements
More Information

- **XtreemOS**
  - Web site: http://www.xtreemos.eu
  - Software: http://gforge.inria.fr/projects/xtreemos/
    - GPL/BSD licence
  - INRIA/XtreemOS booths at SC 2010

- **Contrail**
  - http://www.contrail-project.eu