

# Clustering with openMosix

Maurizio Davini

Department of Physics

INFN Pisa

*([maurizio.davini@df.unipi.it](mailto:maurizio.davini@df.unipi.it))*

# Introduction

- Linux Clusters in Pisa
- Why openMosix ?
- What is openMosix?
  - Single-System Image
  - Preemptive Process Migration
  - The openMosix File System (MFS)
- The future

# Linux Clusters in Pisa

# Linux Clusters in Pisa(1)

- **Anubis** cluster
- 13 SuperMicro 6010H  
Dual PIII 1Ghz,1GB  
RAM,18 SCSI disk
- RedHat 7.2



# Linux Clusters in Pisa(2)

- **Seth Cluster**
- 27 Appro 1124 Dual AMD Athlon MP 1800+, 1GB RAM, 18GB SCSI disk
- RedHat 7.2
- Ganglia monitor



# The cluster applications

- **Anubis** cluster :QCD simulations full time
- **Seth Cluster**: QCD simulations, Nuclear Physics Simulations, Quantum Chemistry Applications (Gaussian...), Plasma Physics Simulation, Virgo Data Analysis

# The openMosix Project history

- Born early 80s on PDP-11/70. One full PDP and disk-less PDP, therefore process migration idea.
- First implementation on BSD/pdp as MS.c thesis.
- VAX 11/780 implementation (different word size, different memory architecture)
- Motorola / VME bus implementation as Ph.D. thesis in 1993 for under contract from IDF (Israeli Defence Forces)
- 1994 BSDi version
- GNU and Linux since 1997
- Contributed dozens of patches to the standard Linux kernel
- Split Mosix / openMosix November 2001

# What is openMOSIX (today version 1.5.4)

- **Linux kernel extension (2.4.17) for clustering**
- **Single System Image - like an SMP, for:**
  - **No need to modify applications**
  - **Adaptive resource management to dynamic load characteristics (CPU intensive, RAM intensive, I/O etc.)**
  - **Linear scalability (unlike SMP)**



# Single System Image Cluster

- **Users can start from any node in the cluster, or sysadmin setups a few nodes as "login" nodes**
- **use round-robin DNS: “hpc.qclusters” with many IPs assigned to same name**
- **Each process has a Home-Node**
  - **Migrated processes always seem to run at the home node,  
e.g., “ps” show all your processes, even if they run elsewhere**

# A two level technology

## **1. Information gathering and dissemination**

- Support scalable configurations by probabilistic dissemination algorithms**
- Same overhead for 16 nodes or 2056 nodes**

## **2. Pre-emptive process migration that can migrate any process, anywhere, anytime - transparently**

- Supervised by adaptive algorithms that respond to global resource availability**
- Transparent to applications, no change to user interface**

# Level 1: Information gathering and dissemination

- **Each unit of time (1 second) each node gathers and disseminates information about:**
  - CPU(s) speed, load and utilization
  - Free memory
  - Free proc-table/file-table slots
- **Info sent to a randomly selected node**
  - Scalable - more nodes better scattering

# Level 2: Process migration by adaptive resource management algorithms

- **Load balancing:** reduce variance between pairs of nodes to improve the overall performance
- **Memory ushering:** migrate processes from a node that nearly exhausted its free memory, to prevent paging
- **Parallel File I/O:** bring the process to the file-server, direct file I/O from migrated processes

# Performance of process migration

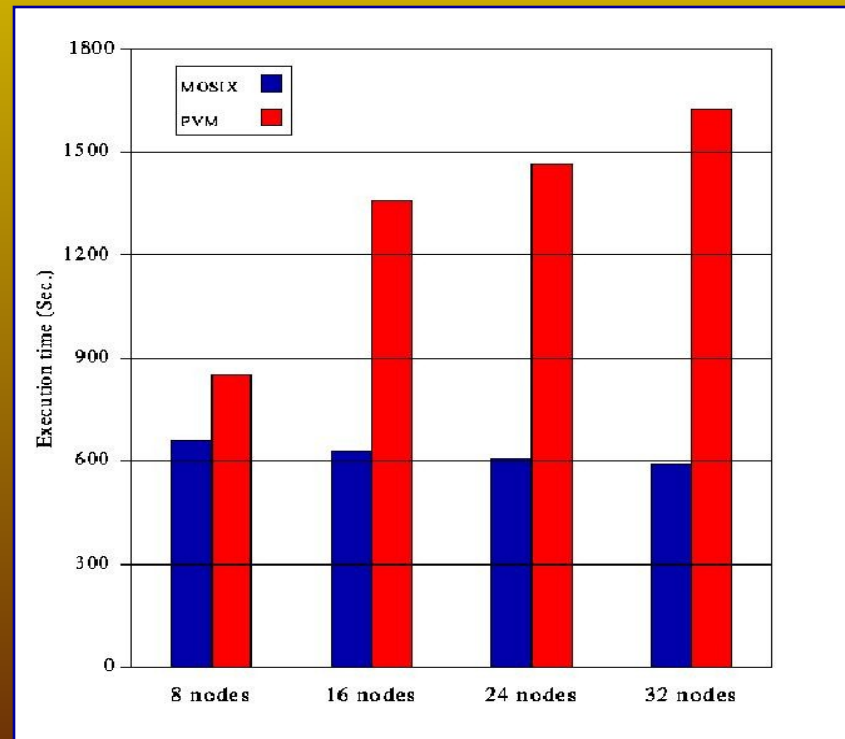
- **CPU: Pentium III 400 MHz**
- **LAN: Fast-Ethernet**
- **For reference: remote system call = 300microsec**
- **Times:**
  - **Initiation time = 1740microsec (less than 6 system calls)**
  - **Migration time = 351microsec per 4KB page**
- **Migration speed: 10.1 MB/Sec = 88.8 Mb/Sec**

# Process migration (MOSIX) vs. static allocation (PVM/MPI)

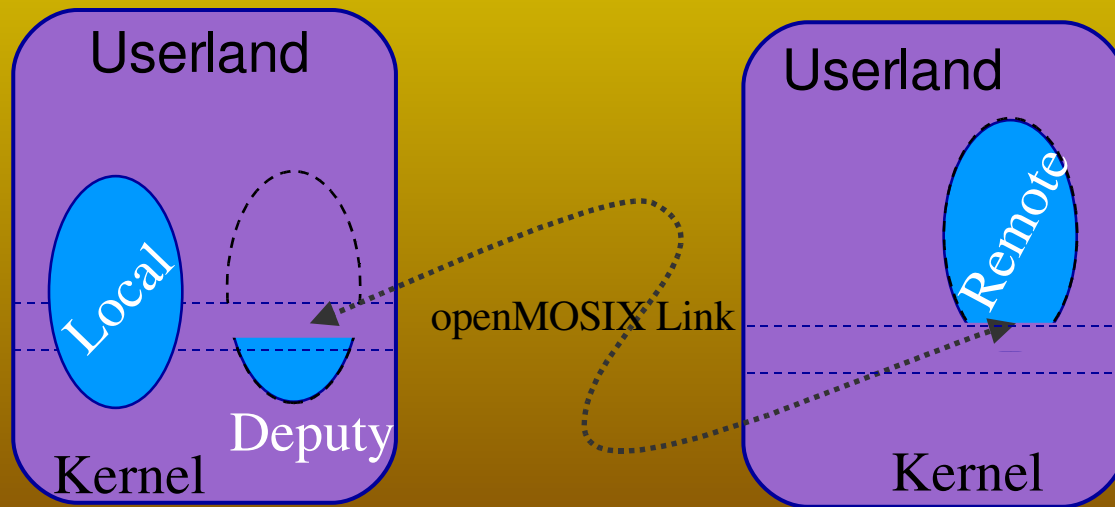
Fixed number of processes  
per node

Random process size with  
average 8MB

Note the performance  
**(un)**scalability !



# Migration - Splitting the Linux process



- **System context (environment) - site dependent- "home" confined**
- **Connected by an exclusive link for both synchronous (system calls) and asynchronous (signals, MOSIX events)**
- **Process context (code, stack, data) - site independent - may migrate**

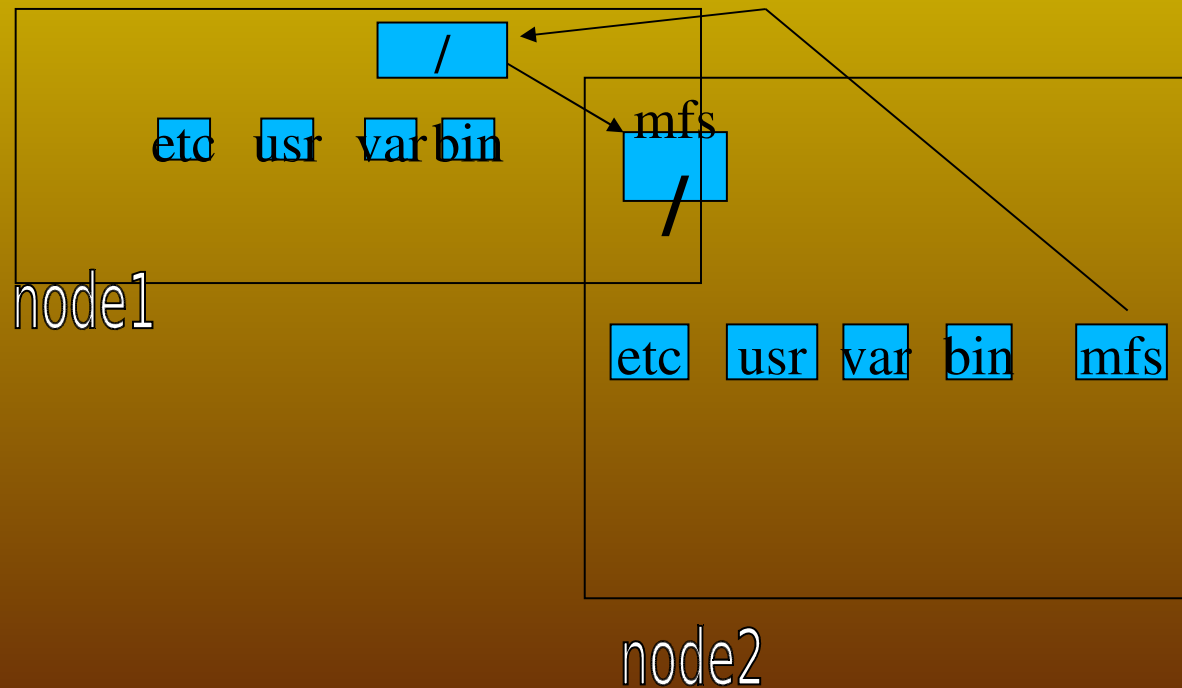
# The Mosix FileSystem



# The MOSIX File System (MFS)

- Not a 'Real Filesystem' but a /proc like filesystem
- Provides a unified view of all files and all mounted FSs on all the nodes of a MOSIX cluster as if they were within a single file system
- Makes all directories and regular files throughout an openMOSIX cluster available from all the nodes
- Provides cache consistency as files viewed from different nodes by maintaining one cache at the server node
- Allows parallel file access by proper distribution of files (each process migrate to the node which has its files)

# The MFS File System Namespace



# Direct File System Access (DFSA)

- **I/O access through the home node incurs high overhead**
- **Direct File System Access (DFSA) compliant file systems allow processes to perform file operations (directly) in the current node - not via the home node**
- **Available operations: all common file-system and I/O system-calls on conforming file systems**
- **Conforming FS: GFS, openMOSIX File System (MFS), Lustre, GPFS and PVFS in the future**

# DFSA Requirements

- **The FS (and symbolic-links) are identically mounted on the same-named mount-points**
- **File consistency: when an operation is completed in one node, any subsequent operation on any other node see the results of that operation**
  - **Required because an openMOSIX process may perform consecutive syscalls from different nodes**
  - **Time-stamp consistency: if file A is modified after B, A must have a timestamp  $\geq$  B's timestamp**

# Global File System (GFS) with DFSA

- **Provides local caching and cache consistency over the cluster using a unique locking mechanism**
- **Provides direct access from any node to any storage entity (via Fiber-channel)**
- **Latest: GFS now includes support for DFSA**
- **GFS + process migration combine the advantages of load-balancing with direct disk access from any node - for parallel file operations**
- **Problem with License (SPL)**

# Postmark (heavy FS load) client-server performance

<i>Access Method</i>	<i>Data Transfer Block Size</i>						
	<b>64B</b>	<b>512B</b>	<b>1KB</b>	<b>2KB</b>	<b>4KB</b>	<b>8KB</b>	<b>16KB</b>
<b>Local (in the server)</b>	102.6	102.1	100.0	102.2	100.2	100.2	101.0
<b>MFS with DFSA</b>	104.8	104.0	103.9	104.1	104.9	105.5	104.4
<b>NFSv3</b>	184.3	169.1	158.0	161.3	156.0	159.5	157.5
<b>MFS without DFSA</b>	<b>1711.0</b>	382.1	277.2	202.9	153.3	136.1	124.5

# The openMosix API

# Kernel 2.4. API and Implementation

- **No new system-calls**
- **Everything done through /proc**

/proc/hpc

/proc/hpc/admin

Administration

/proc/hpc/info

Cluster-wide information

/proc/hpc/nodes/**nnnn**/

Per-node information

/proc/hpc/remote/**pppp**/

Remote proc. information



# Impact on the kernel

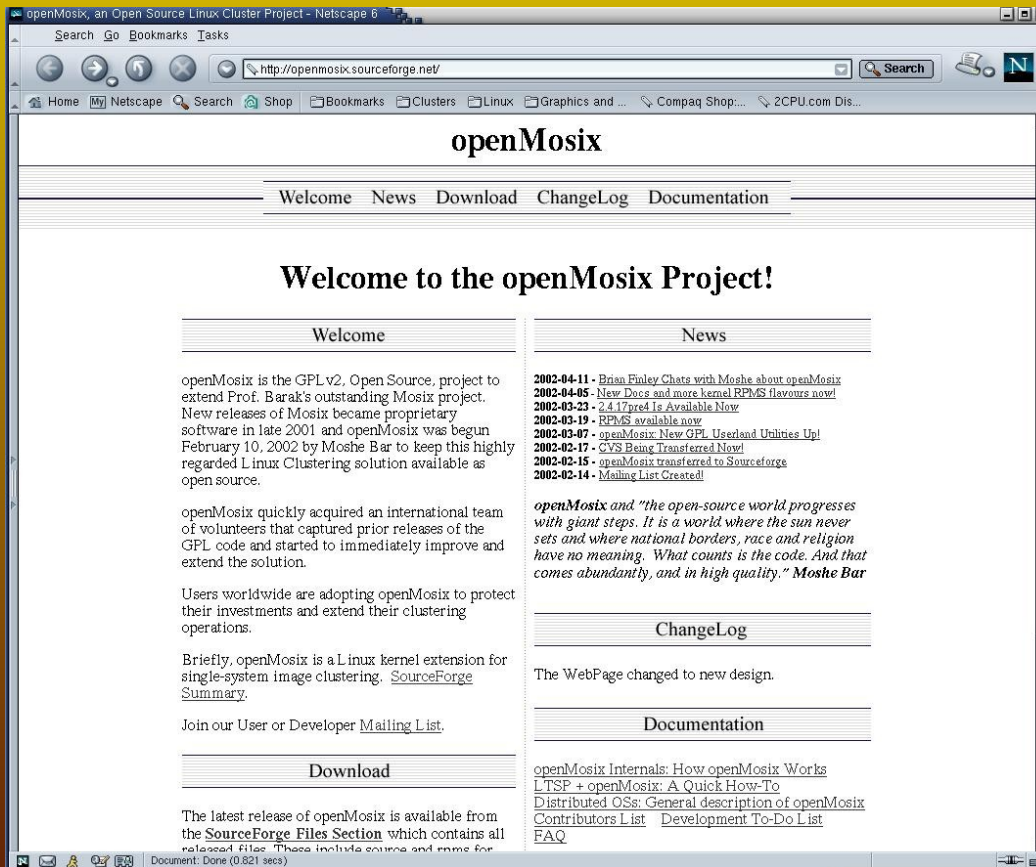
- **MOSIX for the 2.2.19 kernel:**
  - 80 new files (40,000 lines)
  - 109 modified files (7,000 lines changed/added)
  - About 3,000 lines are load-balancing algorithms
- **openMOSIX for Linux 2.4.17**
  - 47 new files (38,500 lines)
  - 126 kernel files modified (5,200 lines changed/added)
  - 48 user-level files (12,000 lines)

# Some Tools

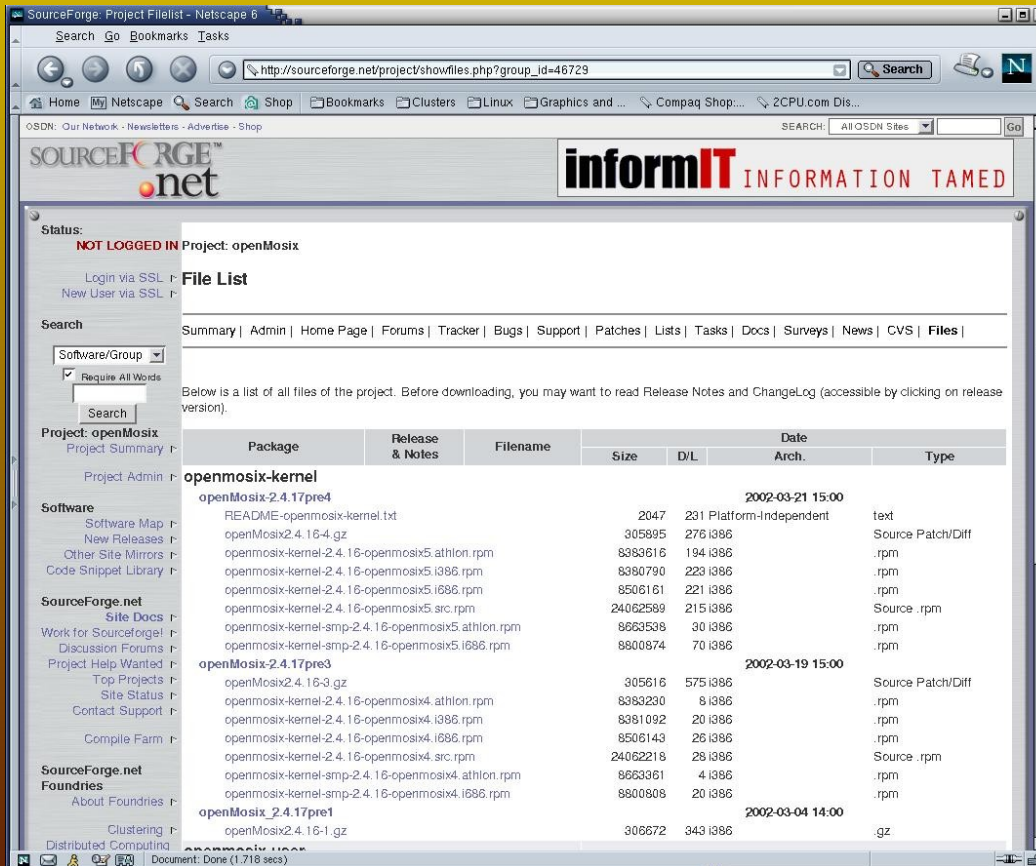
- Some ancillary tools
  - Kernel debugger for 2.2. and 2.4
  - Kernel profiler
  - Parallel make (all `exec()` become `mexec()`)
  - openMosix pvm
  - openMosix mm5
  - openMosix HMMER
  - openMosix Mathematica

# Cluster Installation with OpenMosix

# The openMosix Web Site



# SourceForge download page



The screenshot shows a Netscape browser window displaying the SourceForge project page for 'openMosix'. The browser's address bar shows the URL: `http://sourceforge.net/project/showfiles.php?group_id=46729`. The page features the SourceForge logo and an 'informIT INFORMATION TAMED' banner. A navigation menu includes links for Summary, Admin, Home Page, Forums, Tracker, Bugs, Support, Patches, Lists, Tasks, Docs, Surveys, News, CVS, and Files. A search box is present with a 'Require All Words' checkbox. The main content area displays a 'File List' table with columns for Package, Release & Notes, Filename, Size, D/L, Date, Arch., and Type. The table lists various files for 'openMosix-2.4.17pre4', 'openMosix-2.4.17pre3', and 'openMosix\_2.4.17pre1', including README files, RPM packages, and source code archives.

Package	Release & Notes	Filename	Date		Type	
			Size	D/L		
<b>openMosix-2.4.17pre4</b>			<b>2002-03-21 15:00</b>			
		README-openMosix-kernel.txt	2047	231	Platform-Independent	text
		openMosix2.4.16-4.gz	305595	276	i386	Source Patch/Diff
		openMosix-kernel-2.4.16-openMosix5.athlon.rpm	8383616	194	i386	.rpm
		openMosix-kernel-2.4.16-openMosix5.i386.rpm	8380790	223	i386	.rpm
		openMosix-kernel-2.4.16-openMosix5.i686.rpm	8506161	221	i386	.rpm
		openMosix-kernel-2.4.16-openMosix5.src.rpm	24062589	215	i386	Source .rpm
		openMosix-kernel-smp-2.4.16-openMosix5.athlon.rpm	8663538	30	i386	.rpm
		openMosix-kernel-smp-2.4.16-openMosix5.i686.rpm	8800874	70	i386	.rpm
<b>openMosix-2.4.17pre3</b>			<b>2002-03-19 15:00</b>			
		openMosix2.4.16-3.gz	305616	575	i386	Source Patch/Diff
		openMosix-kernel-2.4.16-openMosix4.athlon.rpm	8383230	8	i386	.rpm
		openMosix-kernel-2.4.16-openMosix4.i386.rpm	8381092	20	i386	.rpm
		openMosix-kernel-2.4.16-openMosix4.i686.rpm	8506143	26	i386	.rpm
		openMosix-kernel-2.4.16-openMosix4.src.rpm	24062219	29	i386	Source .rpm
		openMosix-kernel-smp-2.4.16-openMosix4.athlon.rpm	8663361	4	i386	.rpm
		openMosix-kernel-smp-2.4.16-openMosix4.i686.rpm	8800908	20	i386	.rpm
<b>openMosix_2.4.17pre1</b>			<b>2002-03-04 14:00</b>			
		openMosix2.4.16-1.gz	306672	343	i386	.gz

# Cluster Installation

- Various installation options:
  1. K12LTSP ([www.k12ltsp.org](http://www.k12ltsp.org))
  2. ClumpOs
  3. Debian distribution already includes openMosix
  4. Install RedHat 7.2 and download openMosix RPMS from [sourceforge.net](http://sourceforge.net)
    1. Edit `/etc/mosix.map`
    2. Reboot and ...that's all

# Cluster Administration (1)

- UserLand tools for openMosix
  - Mosctl for node administration
  - Mosrun
  - Migrate
  - .....
- Use 'mps' & 'mtop' for more complete process status information

# openMosix tuning

- 14 parameters to modify openMosix behaviour (`/proc/openMosix/admin/overheads`)
- openMosix provides automated configuration and tuning tools
- Run *prep\_tune* on a node and *tune\_kernel* on another and cat the result in `/proc/openMosix/admin/overheads`

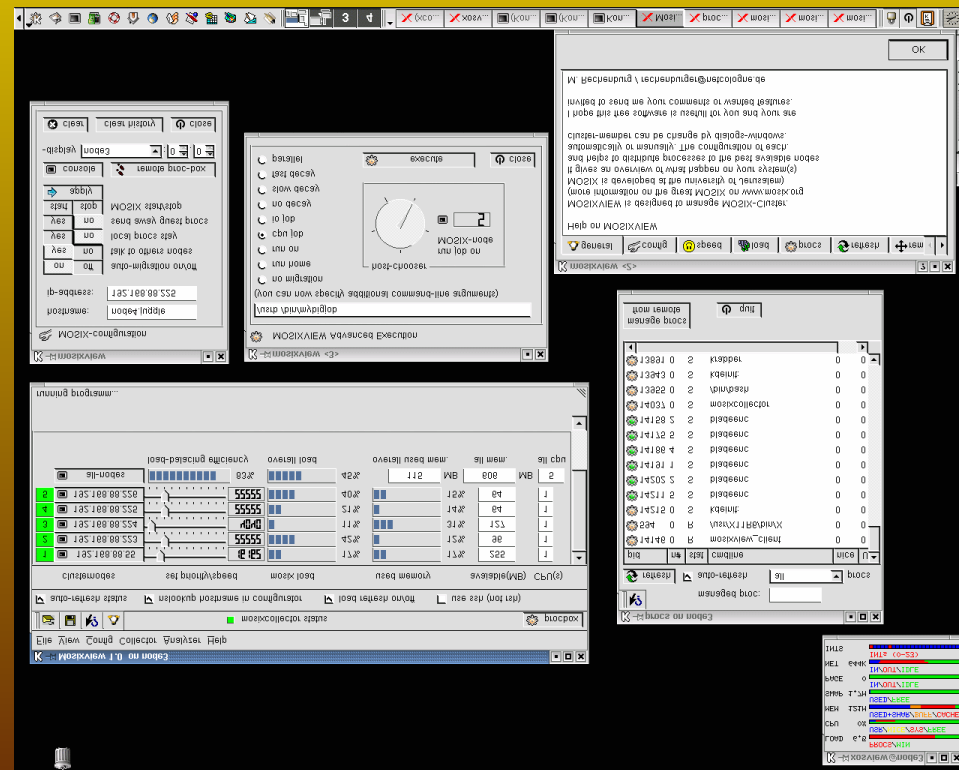


# Cluster Monitoring

- **Cluster monitor - 'mosmon'(or 'qtop')**
  - Displays load, speed, utilization and memory information across the cluster.
  - Uses the /proc/hpc/info interface for the retrieving information
- **Mosixview with X GUI**

# Mosixview

- Developed by Mathias Rechemburg
- [www.mosixview.cc](http://www.mosixview.cc) (and its mirror)



# Applications

# Application Fields

- **Scalable storage area cluster (SAN + Cluster) for parallel file access**
  - Scalable transaction processing systems
- **Scalable web servers: assign new incoming requests to the least loaded node**
  - Scalable to any number of nodes by IP rotation
  - Higher availability
- **Misc. applications - parallel make**

# HPC Applications

Demanding applications:

- Protein classification
- Molecular dynamics
- Weather forecasting (MM5)
- Computational fluid dynamics
- Car crash numerical simulations (parallel Autodyn)
- Military applications

# Example: Parallel Make

- **Assign the next file to the least loaded node**
- **A cluster of 52 4-way 550MHz Xeon nodes**
  - **Runs over a 40 builds of entire code of SAP R/3 (4.7 million lines of code) *concurrently***
  - **Got much better performance vs. LSF cluster for less cost in computing nodes**

# People behind openMosix

- Copyright for openMosix, Moshe Bar
- Barak and Moshe Bar were co-project managers of Mosix until Nov 2001
- Team Members
  - Danny Getz (migration)
  - Avraham Ben Yehudah (MFS and 2.5.x)
  - David Santo Orcero (user-space utilities)
  - Michael Farnbach (extern. Patch matching, ie XFS, JFS etc.)
  - Many others, including help from Ingo Molnar, Alan Cox, Andrea Arcangeli and Rik van Riel

# Present and Future of openMosix



# Current Projects (1)

- Migrating sockets
- Network RAM
- Distributed Shared Memory
- Checkpoint / Restart
- Queue Manager / Scheduler

# Future Plans

- **Inclusion in Linux 2.6**
- **Re-writing MFS**
- **Increase developers to 20-30**

# So....

- openMosix is today still the most advanced HPC clustering option
- A file system like NFS is not really an option in a cluster, MFS, pvfs, GPFS(perhaps) and GFS (...) are.
- openMosix is much more open than the predecessor
- Over 300 installations already switched to openMosix (some classified)
  - University of Pisa
  - STM
  - Intel
  - INFN Napoli
  - SISSA
  - Installation on 1400 nodes (multiprocessor) in Japan

The future

# Clusters in Pisa

- **Amon** cluster
- 5 dual AMD Athlon 1900+, 1GB RAM, 18 GB scsi disk Evolocity Cluster
- RedHat 7.2
- **Donated by AMD**



# Cluster Application

- **Amon Cluster** : target to ‘industry world’
  - Automotive (StarCD,Nastran,Fluent..)
  - Databases

# New machines to test...

- SuperMicro 6022P (2 2.2Xeon 8Gb RAM 1Gb eth+ 1 100Mb eth)
- New Appro Chassis for AMD Athlon 2000/2100+ MP
- Myrinet and Dolphin networks

# Clusters OS

The new frontier



# The new QlusterOS

- Commercial Product
- Release 1.0 announcement of Friday 04-19 at Futurshow in Bologna
- First Installation in Pisa (this weekend)
- First sales to Italy
- Partnership with IBM,RedHat,Compaq,Intel.

# Qclusters OS features (1)

- Based in part on openMosix technology
- Migrating sockets
- Network RAM already implemented
- Cluster parallel Installer,
- Cluster Configurator,
- Qsense ( automatic detection of nodes no-more /etc/mosix.map)
- Monitor (written in Flash),
- Queue Manager ,Launcher, Scheduler
- Job Description Language in XML

# Clusters OS features (2)

- New Load Balancer
- Threaded applications migration
- Linux kernel 2.4.18 with (VM by A.Arcangeli integrated with Reverse Mapping by R.V.Ryel)
- Over 100 patches ( RedHat Quality)
- Kernel latency reduced by 65% due to Robert Love latest pre-emption patch

# Clusters OS features (3)

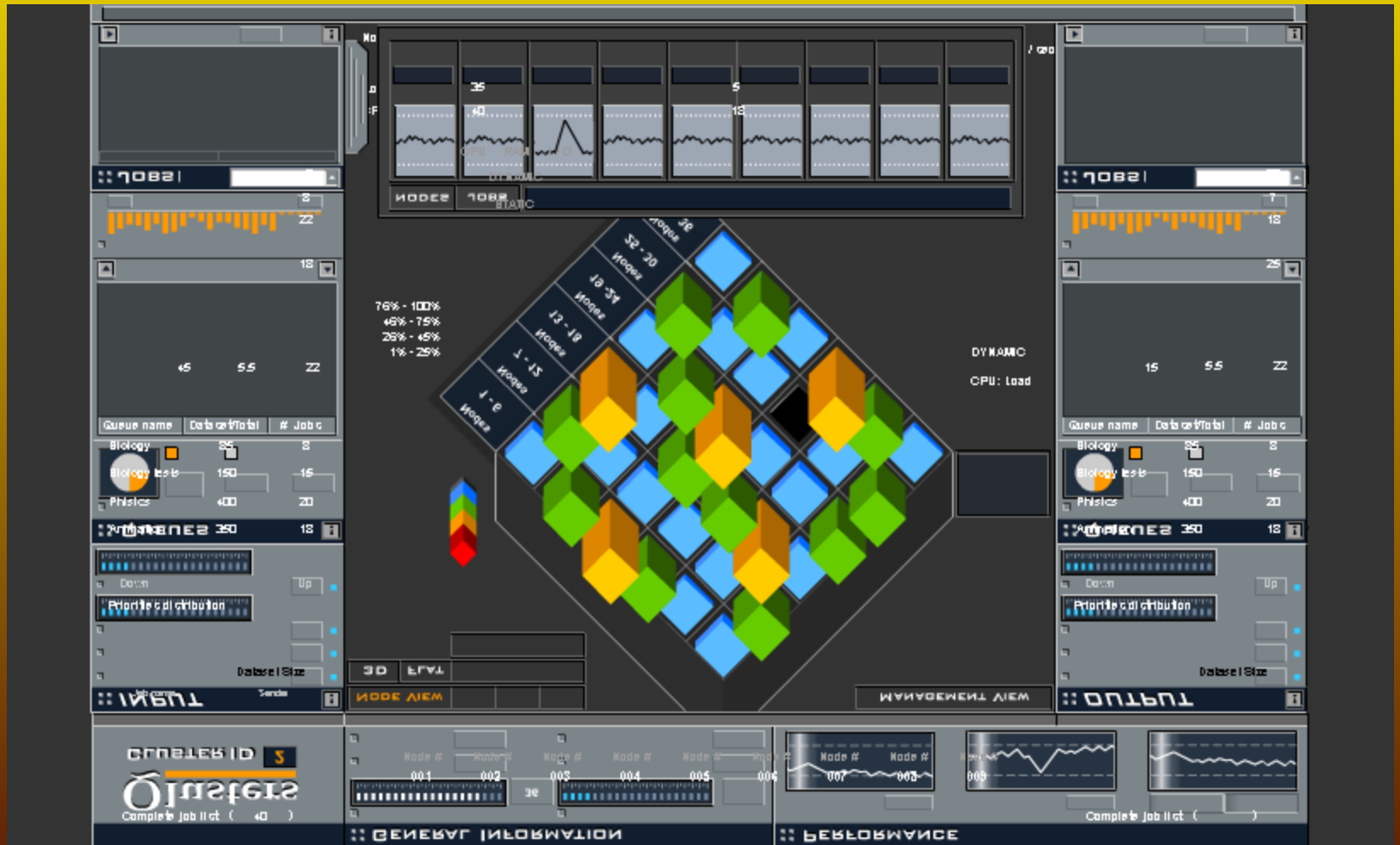
- Support for migration on Myrinet and Dolphin networks
- Integration with GFS completed
- Integration with AFS planned
- IBM xSeries NUMA support
- DSM in a few months

# Qluster Os features (3)

- grid with multiplatform consideration  
(recompiles when transferring on a cluster  
of different architecture )

# The Monitor

# QlusterOS Monitor



# Info on Qclusters OS

- Visit the Web site [www.qclusters.com](http://www.qclusters.com)
- Ask Moshe Bar ( [moshe@moelabs.com](mailto:moshe@moelabs.com) )