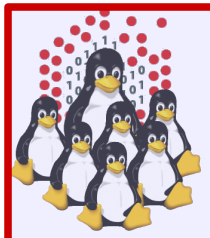
Moreno Baricevic

CNR-INFM DEMOCRITOS
Trieste, ITALY

Installation Procedures for Clusters

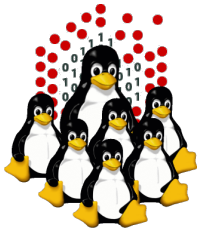
PART 3



Agenda

- Cluster Services
- Overview on Installation Procedures
- Configuration and Setup of a NETBOOT Environment
- Troubleshooting
- Cluster Management Tools
- Notes on Security
- Hands-on Laboratory Session

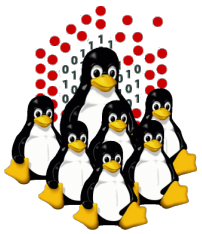
Cluster management tools



CLUSTER MANAGEMENT
Administration Tools

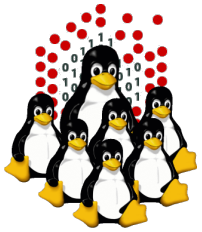
Requirements:

- ✔ cluster-wide command execution
- ✔ cluster-wide file distribution and gathering
- ✔ password-less environment
- ✔ must be simple, efficient, easy to use for CLI addicted



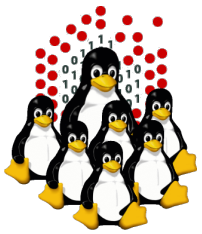
CLUSTER MANAGEMENT
Administration Tools

- C3 tools – The Cluster Command and Control tool suite
  - allows configurable clusters and subsets of machines
  - concurrently execution of commands
  - supplies many utilities
    - cexec (parallel execution of standard commands on all cluster nodes)
    - cexecs (as the above but serial execution, useful for troubleshooting and debugging)
    - cpush (distribute files or directories to all cluster nodes)
    - cget (retrieves files or directory from all cluster nodes)
    - crm (cluster-wide remove)
    - ... and many more
- PDSH – Parallel Distributed SHell
  - same features as C3 tools, few utilities
    - pdsh, pdcp, rpdcp, dshbak
- Cluster-Fork – NPACI Rocks
  - serial execution only
- ClusterSSH
  - multiple xterm windows handled through one input grabber
  - Spawn an xterm for each node! DO NOT EVEN TRY IT ON A LARGE CLUSTER!



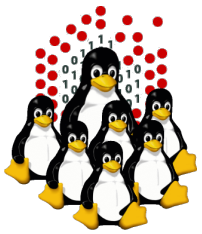
CLUSTER MANAGEMENT
Administration Tools – C4 Tools

- C4 tools – under development, inspired by c3:
  - provides all the c3 features and wrappers (exec, push, get, ...)
  - written in Perl instead of Python
  - better threads handling
  - configurable timeouts
  - configurable default commands (ssh, ping or any other command-line utility or script)
  - allows configurable clusters and subsets of machines, REGEXP are handled as well
  - can use Torque/PBS “nodes” definition file (nodes' “features” define subset of nodes)
  - more command-line options:
    - ssh/rsh client options (or valid options for the *command/script*)
    - variable number of threads
    - selectable features and nodes using REGEXP
    - ...



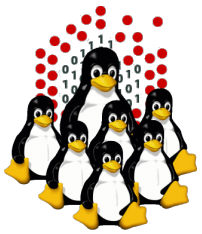
CLUSTER MANAGEMENT
Monitoring Tools

- Ad-hoc scripts (BASH, PERL, ...) + cron
- Ganglia
  - excellent graphic tool
  - XML data representation
  - web-based interface for visualization
  - http://ganglia.sourceforge.net/
- Nagios®
  - complex but can interact with other software
  - configurable alarms, SNMP, E-mail, SMS, ...
  - optional web interface
  - http://www.nagios.org/

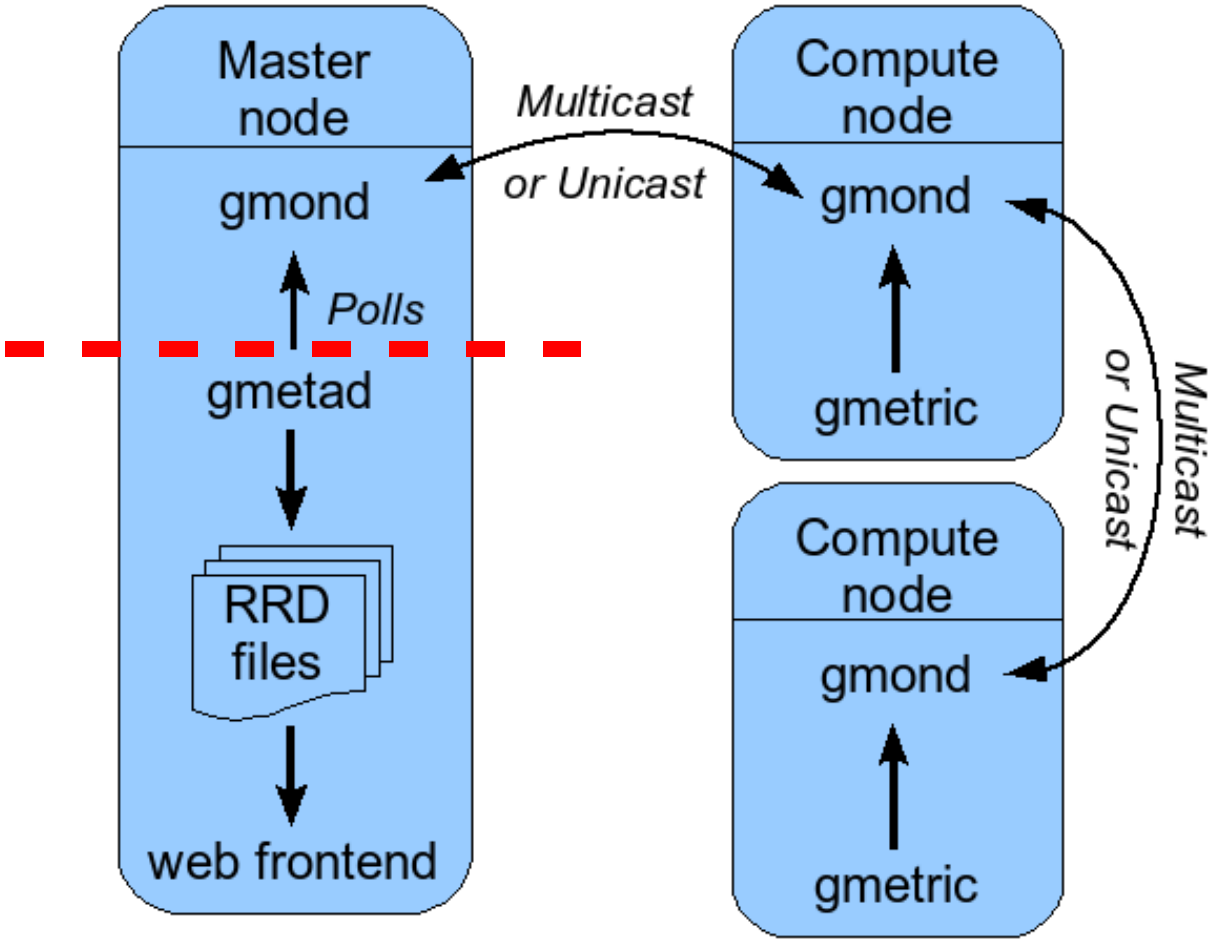


CLUSTER MONITORING
About Ganglia

- is a cluster-monitoring program
- a web-based front-end displays real-time data (aggregate cluster and each single system)
- collects and communicates the host state in real time (a multithreaded daemon process runs on each cluster node)
- monitors a collection of metrics (CPU load, memory usage, network traffic, ...)
- *gmetric* allows to extend the set of metrics to monitor

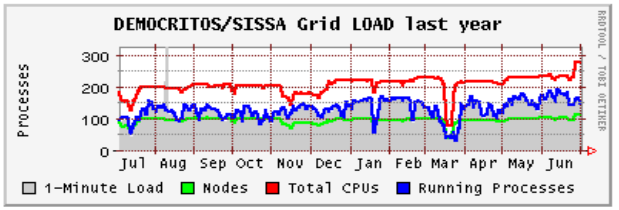
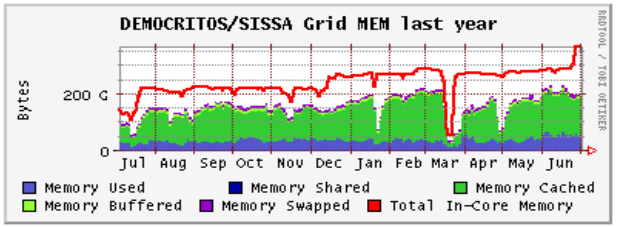
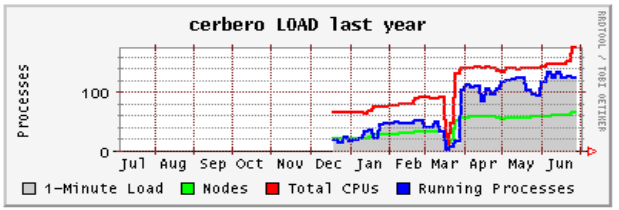
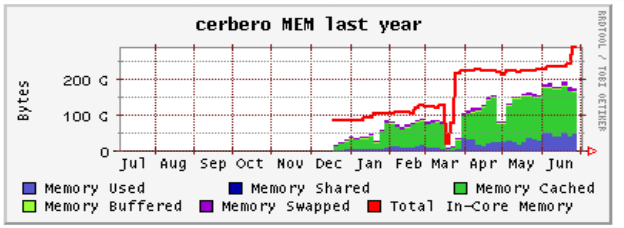
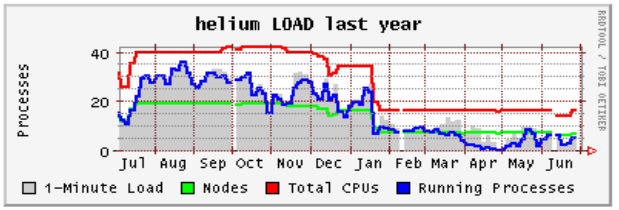
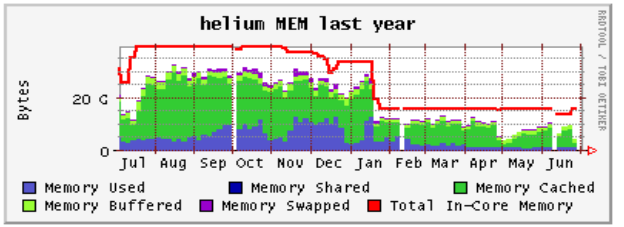
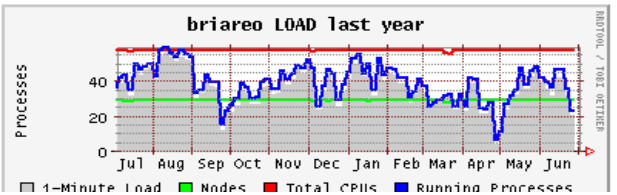
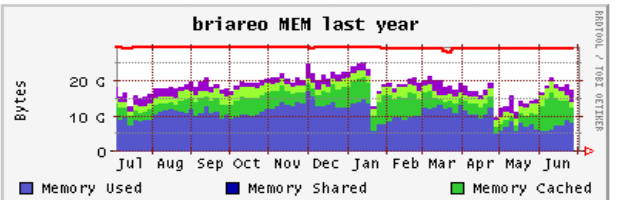


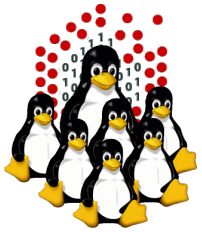
CLUSTER MONITORING
About Ganglia



CLUSTER MONITORING
Ganglia at work /1

DEMOCRITOS/SISSA Grid > --Choose a Source

Name / Info	Load Averages			%CPU User, Nice, System, Idle			
DEMOCRITOS/SISSA Grid (4 sources) (tree view)	124.76	124.33	124.26	45.5	1.3	1.0	52.6
Hosts up: **113** (**276** CPUs Total) Hosts down: **1**						
cerbero (physical view)	111.72	111.80	112.15	65.4	2.1	1.5	29.7
Cluster Localtime: **July 2, 2006, 9:19 pm** Hosts up: **70** (**188** CPUs Total) Hosts down: **0**						
helium (physical view)	4.00	4.00	3.75	28.6	0.0	0.0	71.4
Cluster Localtime: **July 2, 2006, 9:19 pm** Hosts up: **7** (**16** CPUs Total) Hosts down: **0**						
briareo (physical view)	8.73	8.49	8.35	12.4	0.0	0.4	92.1
Cluster Localtime: **July 2, 2006, 9:19 pm** Hosts up: **29** (**58** CPUs Total) Hosts down: **0**						



CLUSTER MONITORING
Ganglia at work /2

Ganglia

DEMOCRITOS/SISSA Grid > cerbero > a103.hpc

a103.hpc Info

a103.hpc
10.1.2.3
Location: Unknown

Last heartbeat received 4 seconds ago.
Uptime 66 days, 12:33

Load: 3.84 4.00 3.99
1m 5m 15m

CPU Utilization: 94.2 4.0 1.6
user sys idle

Hardware
CPUs: 4 x 2192 Mhz
Memory (RAM): 3964 MB
Local Disk: Using 17.074 of 68.024 GB
Most Full Disk Partition: 25.2% used.

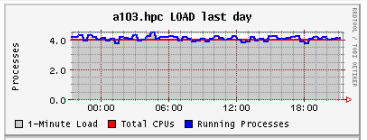
Software
OS: Linux 2.6.13.3 (x86_64)
Booted: April 27, 2006, 8:50 am
Uptime: 66 days, 12:33
Swap: Using 8.7 of 4094.7 MB swap.

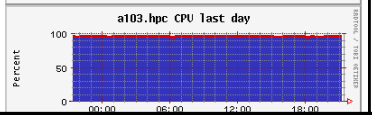
Physical View | Reload

DEMOCRITOS/SISSA Grid > cerbero > a103.hpc

a103.hpc Overview

This node is up and running





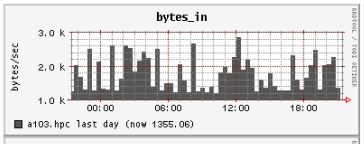
Time and String Metrics

Name	Value
boottime	Thu, 27 Apr 2006 08:50:03 +0200
gexec	OFF
machine_type	x86_64
os_name	Linux
os_release	2.6.13.3
sys_clock	Thu, 27 Apr 2006 08:51:14 +0200
uptime	66 days, 12:33

Constant Metrics

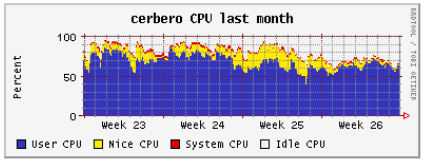
Name	Value
cpu_aidle	17.5 %
cpu_num	4
cpu_speed	2192 MHz
mem_total	4059676 KB
mtu	1500 B
swap_total	4192956 KB

Graphs of Volatile Metrics. Rang

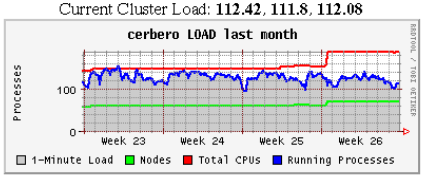


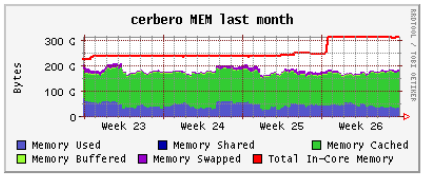
DEMOCRITOS/SISSA Grid > cerbero > --Choose a Node

There are **70** nodes (**188** CPUs) up and running.
There are no nodes down.



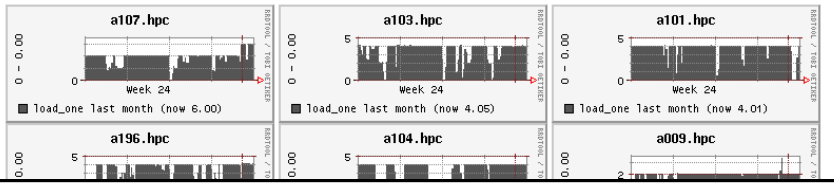
Current Cluster Load: **112.42, 111.8, 112.08**

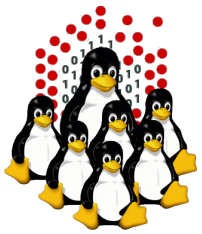




Snapshot of cerbero | Legend

cerbero load_one



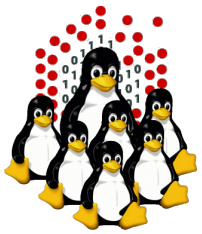


CLUSTER MONITORING
What does Nagios provide?

Nagios®

- ✔ Comprehensive Network Monitoring
- ✔ Problem Remediation
- ✔ Proactive Planning
- ✔ Immediate Awareness and Insight
- ✔ Reporting Options
- ✔ Multi-Tenant/Multi-User Capabilites
- ✔ Integration With Your Existing Applications
- ✔ Customizable Code
- ✔ Easily Extendable Architecture
- ✔ Stable, Reliable, and Respected Platform
- ✔ Huge Community

from http://www.nagios.org/about/



CLUSTER MONITORING
Nagios at work /1 – Tactical Overview

Nagios®

Hosts

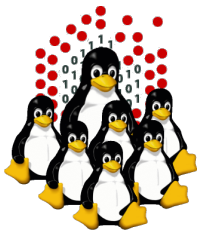
4 Down	0 Unreachable	164 Up	0 Pending
1 Scheduled			
4 Acknowledged			

Services

37 Critical	3 Warning	0 Unknown	1392 Ok	91 Pending
2 Unhandled Problems	3 Unhandled Problems		64 Disabled	91 Disabled
35 on Problem Hosts				

Monitoring Features

Flap Detection	Notifications	Event Handlers	Active Checks	Passive Checks
Enabled	Enabled	Enabled	Enabled	Enabled
All Services Enabled	472 Services Disabled	All Services Enabled	155 Services Disabled	All Services Enabled
No Services Flapping	2 Hosts Disabled	All Hosts Enabled	All Hosts Enabled	All Hosts Enabled
All Hosts Enabled				
No Hosts Flapping				



CLUSTER MONITORING
Nagios at work /2 – Host Status

Nagios®

Host Information
Last Updated: Fri Mar 20 12:51:53 CET 2009
Updated every 90 seconds
Nagios® 3.0.6 - www.nagios.org
Logged in as *nagiosadmin*

View Status Detail For This Host
View Alert History For This Host
View Trends For This Host
View Alert Histogram For This Host
View Availability Report For This Host
View Notifications This Host

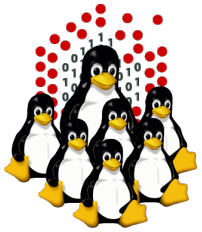
Host
c007
(c007)

Member of
c-nodes

10.2.10.7

Host State Information

Host Status:	**UP** (for 1d 0h 48m 9s)
Status Information:	PING OK - Packet loss = 0%, RTA = 0.21 ms
Performance Data:	rta=0.207000ms;3000.000000;5000.000000;0.000000 pl=0%;80;100;0
Current Attempt:	1/15 (HARD state)
Last Check Time:	03-20-2009 12:51:34
Check Type:	ACTIVE
Check Latency / Duration:	0.590 / 4.276 seconds
Next Scheduled Active Check:	03-20-2009 12:56:44
Last State Change:	03-19-2009 12:03:44
Last Notification:	N/A (notification 0)
Is This Host Flapping?	**NO** (0.00% state change)
In Scheduled Downtime?	**NO**
Last Update:	03-20-2009 12:51:44 (0d 0h 0m 9s ago)



CLUSTER MONITORING
Nagios at work /3 – Service Status Detail

Nagios®

Current Network Status
Last Updated: Fri Mar 20 12:51:28 CET 2009
Updated every 90 seconds
Nagios® 3.0.6 - www.nagios.org
Logged in as *nagiosadmin*

View History For This Host
View Notifications For This Host
View Service Status Detail For All Hosts

Host Status Totals

Up	Down	Unreachable	Pending
1	0	0	0

All Problems	All Types
0	1

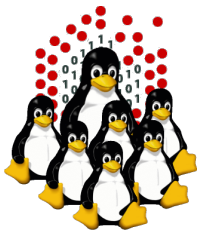
Service Status Totals

Ok	Warning	Unknown	Critical	Pending
9	1	0	0	0

All Problems	All Types
1	10

Service Status Details For Host 'c007'

Host	Service	Status	Last Check	Duration	Attempt	Status Information
c007	EDAC memory errors	WARNING	03-20-2009 12:42:12	0d 14h 9m 16s	3/3	WARNING - several correctable memory errors
	NFS mounts from local fstab	OK	03-20-2009 12:44:38	1d 0h 46m 53s	1/3	NFS mounts OK
	NTP server	OK	03-20-2009 12:48:58	1d 0h 22m 31s	1/3	NTP OK: Offset 0.001392602921 secs
	PING	OK	03-20-2009 12:48:31	1d 0h 42m 57s	1/3	PING OK - Packet loss = 0%, RTA = 0.24 ms
	SSH	OK	03-20-2009 12:42:48	1d 0h 28m 41s	1/3	SSH OK - OpenSSH_4.3 (protocol 2.0)
	job events	OK	03-19-2009 11:15:51	5d 19h 16m 0s	1/3	job 22372 by smogunov/tosatti
	load average	OK	03-20-2009 12:47:14	1d 0h 4m 14s	1/3	OK - load average: 0.00, 0.00, 0.00
	lustre client	OK	03-20-2009 12:47:08	1d 0h 4m 21s	1/3	lustre client OK
	pbs_mom	OK	03-20-2009 12:47:21	1d 0h 44m 7s	1/4	TCP OK - 0.000 second response time on port 15002
	reverse_ping_IB	OK	03-20-2009 12:46:02	1d 0h 45m 26s	1/3	PING OK - Packet loss = 0%, RTA = 0.39 ms



CLUSTER MONITORING
Nagios at work /4 – Service Problems

Nagios®

Current Network Status
Last Updated: Fri Mar 20 12:50:50 CET 2009
Updated every 90 seconds
Nagios® 3.0.6 - www.nagios.org
Logged in as *nagiosadmin*

View History For all hosts
View Notifications For All Hosts
View Host Status Detail For All Hosts

Host Status Totals

Up	Down	Unreachable	Pending
164	4	0	0

All Problems	All Types
4	168

Service Status Totals

Ok	Warning	Unknown	Critical	Pending
1474	3	0	36	10

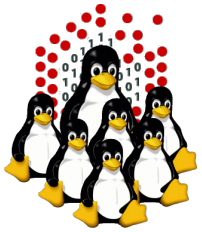
All Problems	All Types
39	1523

Display Filters:

Host Status Types:	Pending | Up
Host Properties:	Any
Service Status Types:	All Problems
Service Properties:	Not In Scheduled Downtime & Has Not Been Acknowledged & Active Checks Enabled

Service Status Details For All Hosts

Host	Service	Status	Last Check	Duration	Attempt	Status Information
a199	EDAC memory errors	CRITICAL	03-20-2009 12:42:09	10d 1h 28m 46s	3/3	CRITICAL - many correctable memory errors
c007	EDAC memory errors	WARNING	03-20-2009 12:42:12	0d 14h 8m 38s	3/3	WARNING - several correctable memory errors
m038	EDAC memory errors	WARNING	03-20-2009 12:44:44	10d 1h 26m 10s	3/3	WARNING - several correctable memory errors
m045	EDAC memory errors	WARNING	03-20-2009 12:43:54	10d 1h 27m 25s	3/3	WARNING - several correctable memory errors



LOCAL AND REMOTE ACCESS

LOCAL ACCESS

- LOCAL CONSOLE (max ~10m for PS2, ~30m VGA) (*)
- KVM (max ~30m) (*)
- SERIAL CONSOLE (RS232, max ~15m@19200baud / ~150m@9600baud) (*)

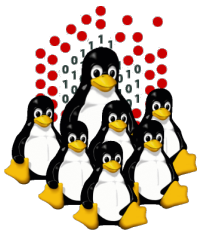
* repeaters and transceivers increase the max length

REMOTE ACCESS (OS dependent, **in-band**)

- SSH
- VNC, remote desktop, ...

REMOTE ACCESS (OS in-dependent, **out-of-band**)

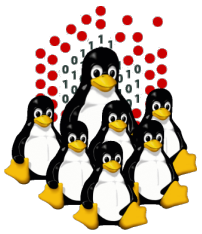
- KVM over IP (hardware)
- SERIAL over IP (hardware; serial hubs, IBM RSA and other LOM systems)
- SERIAL over LAN (hardware; IPMI)
- JAVA CONSOLE, web appliances (hardware+sw; SUN and other vendors)



REMOTE MANAGEMENT

SysAdmins are lazy, and IT-button-pusher-slaves cost too much. We want remote management NOW!

What does the market offer?
- in-band and out-of-band controllers
- either built-in or pluggable
- proprietary controllers and protocols (SUN, IBM, HP, ...)
- well-known standards based SPs (IPMI/SNMP) (good)
- some provides ssh access (good)
- some allows only web-based management (bad)
- some requires java (bad)
- some requires weird tools, often closed-source (bad)
- some implements more of the above (VERY GOOD)
- some don't work... (REALLY BAD)

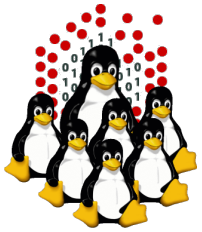


REMOTE MANAGEMENT
IPMI - Intelligent Platform Management Interface

IPMI (Intelligent Platform Management Interface)

- sensor monitoring
- system event monitoring
- power control
- serial-over-LAN (SOL)
- independent of the operating system, but works locally as well

- **OpenIPMI**
  - http://openipmi.sourceforge.net/
  - *ipmicmd, ipmilan, ipmish, ...*
- **GNU FreeIPMI**
  - http://www.gnu.org/software/freeipmi/
  - *bmc-config, ipmi-chassis, ipmi-fru, ipmiping, ipmipower, ...*
- **ipmitool**
  - http://ipmitool.sourceforge.net/
  - *ipmitool*
- **ipmiutil**
  - http://ipmiutil.sourceforge.net/
  - *ipmiutil*



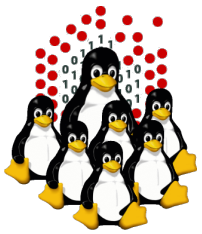
REMOTE MANAGEMENT
SNMP - Simple Network Management Protocol

SNMP (Simple Network Management Protocol)

- monitor network-attached devices (switches, routers, UPSs, PDUs, hosts, ...)
- retrieve and manipulate configuration information (*get/set/trap* actions)
- v1: clear text, no auth (community string)
- v2: clear text, auth (but v2c uses comm. str.)
- v3: privacy, auth, access control
- depends on the NOS/FW, hosts need a local agent

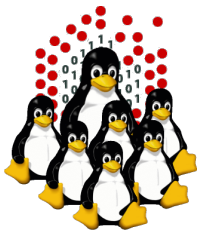
- **Net-SNMP**
  - http://www.net-snmp.org
  - *snmpset*
  - *snmpget*
  - *snmpwalk*
  - many more...

SECURITY



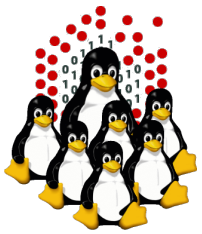
SECURITY NOTES
What you should care of

- physical access / boot security
- active services
- software updates
- filesystem permissions
- user access
- intrusion detection
- system hardening
- virtualization



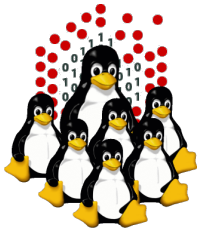
SECURITY NOTES
Hints /1

- PAM: `/etc/pam.d/*, /etc/security/*`
  - `limits.conf`: per-user resources limits (cputime, memory, number of processes, ...)
  - `access.conf`: which user from where
- SSH: `/etc/ssh/sshd_config`
- *TCPwrapper*: `/etc/hosts.{allow,deny}`, only for services handled by *(x)inetd* or compiled against *libwrap*
- firewall: OK on external network; overkill on the cluster network
- services: the least possible



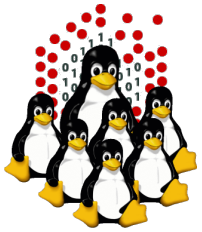
SECURITY NOTES
Hints /2

- ownerships/permissions: local users+exported services, NFS *root_squash* for rw dirs
- *chroot* jails: for some (untrusted) services
- avoid automatic updates, manually patch as far as possible
- beware of test-accounts and passwordless environment outside the cluster
- *grsec*: if you are really paranoid... like we are and you should be ;)
- network devices: default passwords, SNMP, SP/IPMI, CDP and the like, ...



SECURITY NOTES
Security Policy

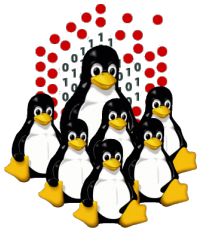
- **HARDWARE**
  - physical access
  - redundancy
- **SOFTWARE**
  - hardening
  - configuration
  - update
  - backup
- **USERS' EDUCATION**
  - “strong” passwords
  - no account sharing
  - prevent social engineering / phishing



That's All Folks!


```
( questions ; comments ) | mail -s uheilaaa baro@democritos.it

( complaints ; insults ) &>/dev/null
```



REFERENCES AND USEFUL LINKS

Cluster Toolkits:

- OSCAR – Open Source Cluster Application Resources
  http://oscar.openclustergroup.org/
- NPACI Rocks
  http://www.rocksclusters.org/
- Scyld Beowulf
  http://www.beowulf.org/
- CSM – IBM Cluster Systems Management
  http://www.ibm.com/servers/eserver/clusters/software/
- xCAT – eXtreme Cluster Administration Toolkit
  http://www.xcat.org/
- Warewulf/PERCEUS
  http://www.warewulf-cluster.org/ http://www.perceus.org/

Installation Software:

- SystemImager http://www.systemimager.org/
- FAI http://www.informatik.uni-koeln.de/fai/
- Anaconda/Kickstart http://fedoraproject.org/wiki/Anaconda/Kickstart

Management Tools:

- openssh/openssl
  http://www.openssh.com
  http://www.openssl.org
- C3 tools – The Cluster Command and Control tool suite
  http://www.csm.ornl.gov/torc/C3/
- PDSH – Parallel Distributed SHell
  https://computing.llnl.gov/linux/pdsh.html
- DSH – Distributed SHell
  http://www.netfort.gr.jp/~dancer/software/dsh.html.en
- ClusterSSH
  http://clusterssh.sourceforge.net/
- C4 tools – Cluster Command & Control Console
  http://gforge.escience-lab.org/projects/c-4/

Monitoring Tools:

- Ganglia http://ganglia.sourceforge.net/
- Nagios http://www.nagios.org/
- Zabbix http://www.zabbix.org/

Network traffic analyzer:

- tcpdump http://www.tcpdump.org
- wireshark http://www.wireshark.org

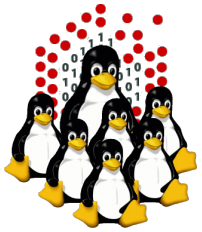
UnionFS:

- Hopeless, a system for building disk-less clusters
  http://www.evolware.org/chri/hopeless.html
- UnionFS – A Stackable Unification File System
  http://www.unionfs.org
  http://www.fsl.cs.sunysb.edu/project-unionfs.html

RFC: (http://www.rfc.net)

- RFC 1350 – The TFTP Protocol (Revision 2)
  http://www.rfc.net/rfc1350.html
- RFC 2131 – Dynamic Host Configuration Protocol
  http://www.rfc.net/rfc2131.html
- RFC 2132 – DHCP Options and BOOTP Vendor Extensions
  http://www.rfc.net/rfc2132.html
- RFC 4578 – DHCP PXE Options
  http://www.rfc.net/rfc4578.html
- RFC 4390 – DHCP over Infiniband
  http://www.rfc.net/rfc4390.html

- PXE specification
  http://www.pix.net/software/pxeboot/archive/pxespec.pdf
- SYSLINUX http://syslinux.zytor.com/



Some acronyms...

ICTP – the Abdus Salam International Centre for Theoretical Physics
DEMOCRITOS – Democritos Modeling Center for Research In aTOmistic Simulations
INFM – Istituto Nazionale per la Fisica della Materia (Italian National Institute for the Physics of Matter)
CNR – Consiglio Nazionale delle Ricerche (Italian National Research Council)

HPC – High Performance Computing

OS – Operating System
LINUX – LINUX is not UNIX
GNU – GNU is not UNIX
RPM – RPM Package Manager

CLI – Command Line Interface
BASH – Bourne Again SHell
PERL – Practical Extraction and Report Language

PXE – Preboot Execution Environment
INITRD – INITial RamDisk

NFS – Network File System
SSH – Secure SHell
LDAP – Lightweight Directory Access Protocol
NIS – Network Information Service
DNS – Domain Name System

PAM – Pluggable Authentication Modules

LAN – Local Area Network

IP – Internet Protocol
TCP – Transmission Control Protocol
UDP – User Datagram Protocol
DHCP – Dynamic Host Configuration Protocol
TFTP – Trivial File Transfer Protocol
FTP – File Transfer Protocol
HTTP – Hyper Text Transfer Protocol
NTP – Network Time Protocol

NIC – Network Interface Card/Controller
MAC – Media Access Control
OUI – Organizationally Unique Identifier

API – Application Program Interface
UNDI – Universal Network Driver Interface
PROM – Programmable Read-Only Memory
BIOS – Basic Input/Output System

SNMP – Simple Network Management Protocol
IPMI – Intelligent Platform Management Interface

LOM – Lights-Out Management
RSA – IBM Remote Supervisor Adapter