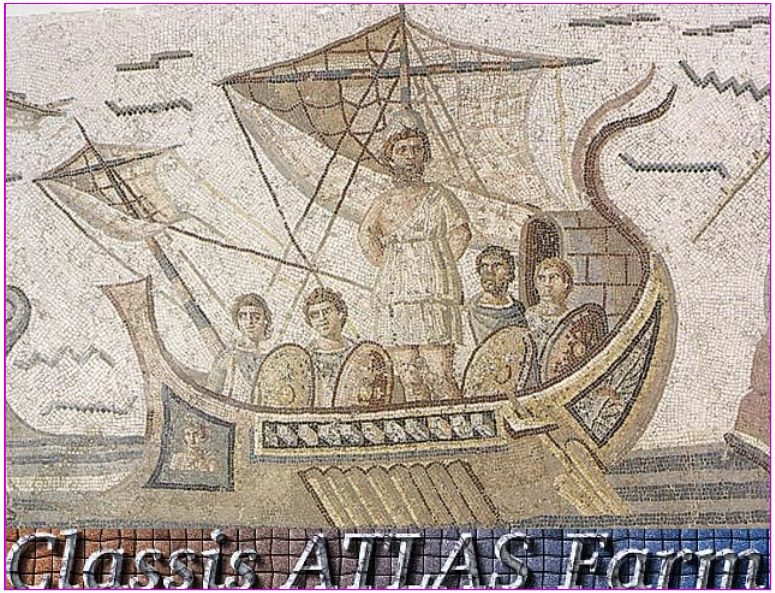


La Farm di Atlas a Roma 1

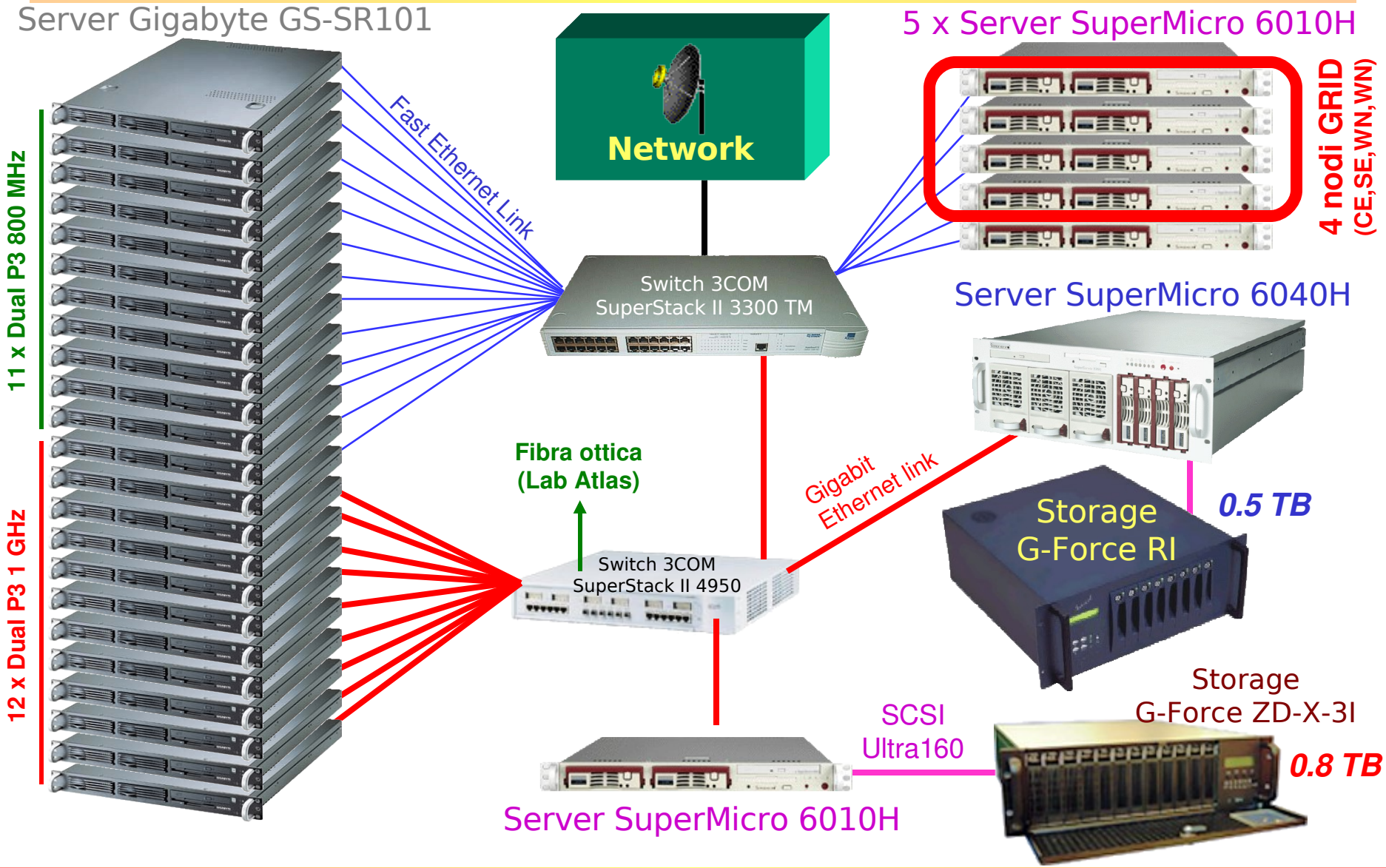
A. De Salvo, A. Di Mattia, L. Luminari, F. Marzano, A. Spanu

Workshop CCR – La Biodola 6-9 maggio 2002

Outline

- Architettura della farm
- Installazione
- Monitoring
- Conclusioni

L'architettura della farm (2002)



Installazione

- Nodi non GRID
  - Installazione via RedHat kickstart
  - IP dinamico attraverso l'utilizzo del server DHCP di sezione
  - Kickstart via rete (RedHat 6.2/7.X)
    - Le macchine client dotate di GE effettuano il kickstart direttamente dall'interfaccia Gigabit Ethernet
  - Il kickstart server è connesso allo switch tramite Gigabit Ethernet
    - Tempo medio di installazione della Farm (in parallelo): 20 min.

- Nodi GRID
  - Installazione via LCFG
  - IP dinamico attraverso il server LCFG
  - EDG 1.1.4
  - 1 CE, 1 SE, 2 WN

Installazione (2)

- **Problemi con RAID Promise FastTrak100 LT**
  - **RedHat Linux 6.2/7.0**
    - **Driver Promise (binary only) per kernel 2.2.x**
  - **RedHat Linux 7.1/7.2**
    - **Driver Promise sperimentale (binary only) per kernel 2.4.2/2.4.7**
    - **Per l'installazione del kernel 2.4.9-31 (ultima versione redhat, supporto ext3 e bugfixes) non è disponibile ancora il driver (e non si sa neanche se lo sarà mai)!**
    - **Il kernel 2.4.9-31 della RedHat supporta parzialmente le devices FastTrak100 LT (alcune opzioni di compilazione non sono attivate)**
      - ***Rebuild del kernel*** **(→ kernel-2.4.9-31a)** ***e update degli RPM***
    - **Il disco di boot via rete della RedHat è creato con il kernel-BOOT-2.4.7, che non supporta il FastTrak100 LT**
      - **Update del dischetto di boot via rete (→ kernel-BOOT-2.4.9-31a) e customizzazione di anaconda per Atlas (logo, rpmlist, opzioni di installazione)**
      - **Creazione di una serie di script per una più facile customizzazione di anaconda**

Installazione (3)

- **Problemi con l'interfaccia Gbit Ethernet Intel e1000**
  - **Il driver incluso nella distribuzione della RedHat non è utilizzabile per questo tipo di hardware**
  - **Compilazione del codice fornito da Intel e generazione degli RPMs per RedHat 6.x e 7.x**
  - **Inclusione del driver corretto (modulo di BOOT) nel dischetto di kickstart**
    - **Le macchine dotate di interfaccia e1000 possono direttamente effettuare l'installazione via rete tramite interfaccia Gbit.**

Software environment

- **Sistemi Operativi:**
  - **Gigabyte GS-SR101: RedHat Linux 7.0 (CASPUR)** ***[RedHat Linux 7.2 (custom, CASPUR based)]***
  - **Supermicro 6010H/6040H: RedHat Linux 6.2 (CASPUR)**
  - **Supermicro utilizzate per GRID: RedHat Linux 6.2 (CERN) via LCFG**

- **Scheduler (batch system)**
  - **DQS v3.3.2 (Standalone)**
  - **OpenPBS 2.3 (GRID)**

- **AFS**
  - **Il client AFS è installato su tutti i nodi non GRID**

- **Software specifico Atlas**
  - **Il software e le librerie runtime di Atlas sono disponibili attraverso AFS e/o installazione locale**
  - **I nodi GRID hanno gli RPMs del kit di Atlas v1.3.0 installati (a breve upgrade alla versione 3.0.1/3.1.0)**
- **Grid software**
  - **INFN globus toolkit v1.2 (+ librerie di bypass – Silvia Resconi/Francesco Prelz)**
  - **EDG (European DataGrid) software (Globus 2) sui nodi GRID.**

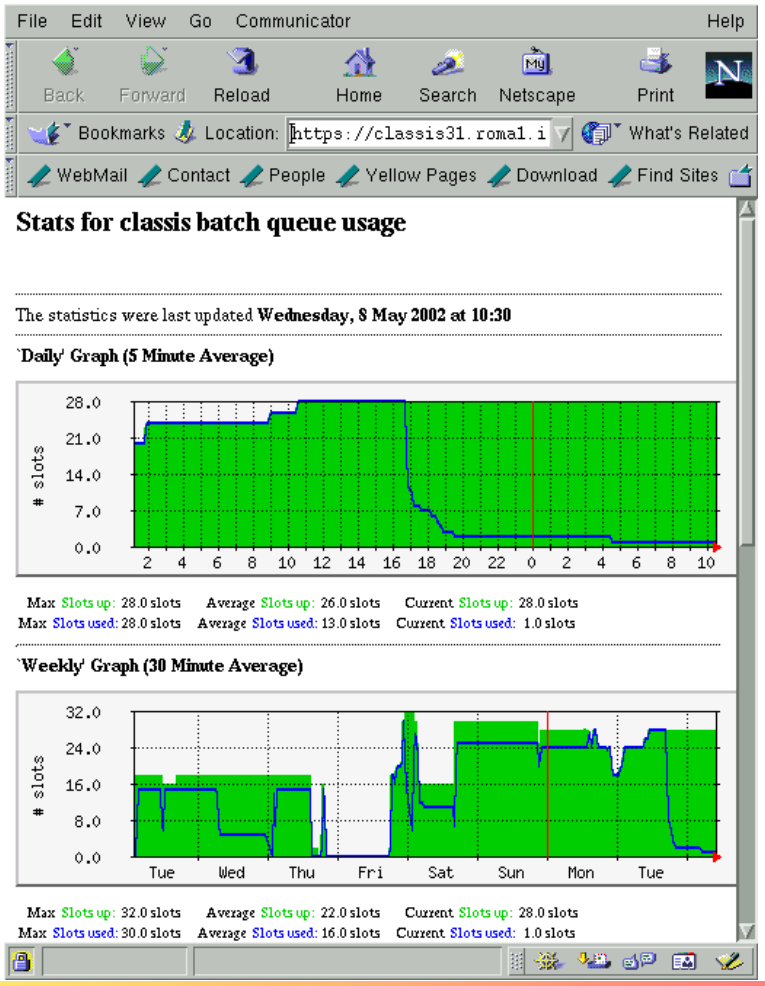
- **Nodi di accesso**
  - **Storicamente classis01 è la macchina di front-end (public access via ssh)**
  - **Sistema multi-server tramite l'alias classis.roma1.infn.it**

Farm management

- Ogni due ore ogni nodo (cron) provvede all'esecuzione dello script centrale di configurazione e update presente sul server di kickstart
  - File passwd/shadow
  - Abilitazione dei servizi
  - Configurazione del firewall
  - Update degli scripts di configurazione
  - Check dei mounts via NFS
  - Configurazione dei demoni del sistema di code
  - Gestione degli upgrades/installazione degli RPM
- Il server di kickstart può in ogni momento forzare l'update delle configurazioni delle macchine o eseguire comandi contemporaneamente su tutto il cluster

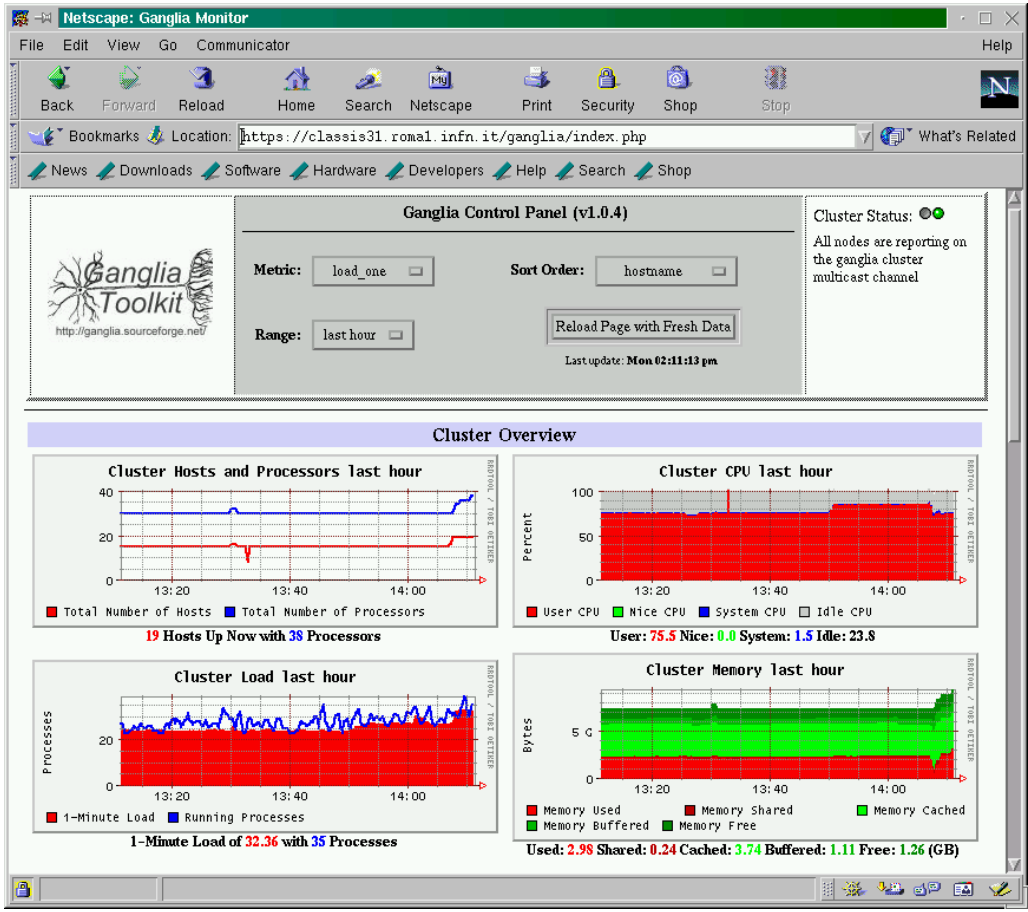
Monitoring (1)

- **Monitoring via *MRTG***
  - I valori delle grandezze da misurare sono ottenuti tramite SNMP
    - Estensione dell'albero di SNMP base
      - Temperatura delle CPUs via lmsensors
      - %CPU
      - % uso del sistema di code
  - I grafici prodotti sono pubblicati su web
    - Esempio:
      https://classis31.roma1.infn.it/mrtg/queue.html
      https://classis31.roma1.infn.it/mrtg/load.html



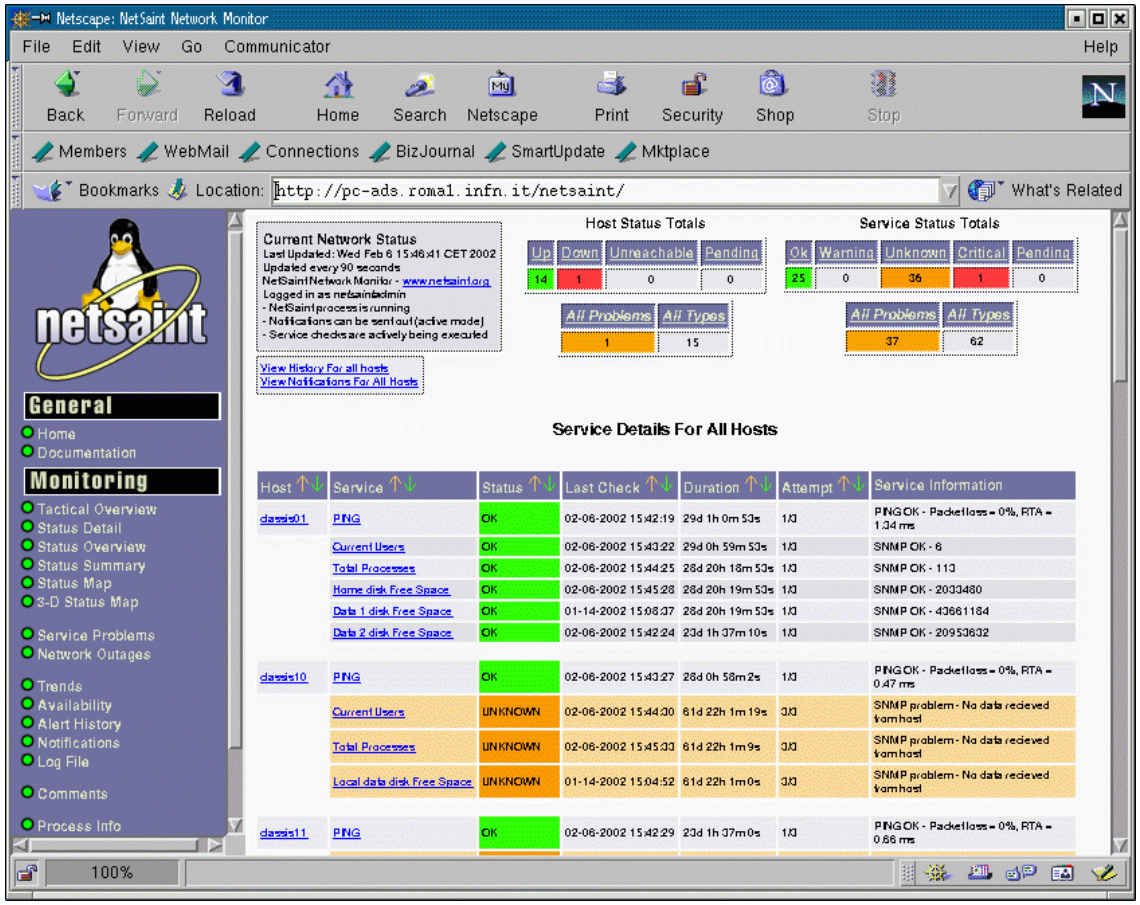
Monitoring (2)

- Cluster monitoring via *Ganglia*
  - Overall monitoring del cluster
  - Ogni nodo esegue il demone gmond che comunica con il server di monitoring via multicast
  - Possibilità di espansione delle metriche da misurare (gmetric)



Monitoring (3)

- **Monitoring via *Netsaint***
  - Accessibile solo ai gestori della farm
  - Alert via email su eventuali problemi



Conclusioni

- **28 Nodi di calcolo Dual Processor + 2 Servers (~2.5 kSPECint95)**

- **Installazione via kickstart**
  - **RedHat 6.2/7.0 CASPUR**
  - **RedHat 7.2 custom (CASPUR based)**

- **Utilizzo di 4 macchine della farm per EDG**
  - **Installazione via LCFG**

- **Monitoring via MRTG/RRDtool, Ganglia, Netsaint**

Statistiche

Users (Roma1/2/3, Frascati, MI, PV, CS)	41
Users “attivi”	12
Utilizzo medio della farm via batch (da 15-02-2001)	~25 %
Spazio disco usato	332 GB (81%)
Hackers	3
Recovery time (hacker 1/2/3)	72 h / 24 h / 45 min
Problemi al raid	3 [t_{perso} = ~15g]
Problemi ai nodi di calcolo	1 [t_{perso} = ???]
Ventole di raffreddamento	> 20 [t_{perso} = 0]

Spazio Disco

- **Spazio disco (dati + utenti)**

Modello	Mount Point	Dimensione [GB]	Utilizzo [%]
G-Force RI	classis01:/home	30	93
G-Force RI	classis01:/storage/data1	192	80
G-Force RI	classis01:/storage/data2	189	89
G-Force ZD-X-3I	*classis02:/storage/data3*	*~350 GB*	*0*
G-Force ZD-X-3I	*classis02:/storage/data4*	*~350 GB*	*0*

- **Aree di backup**
  - **2 dischi SCSI da 36 GB su classis01**
    (+ altri 3 dischi da 36 GB su altre macchine)
  - **Backup delle home directories via Arkeia**
  - **Crash recovery con tar**
    (Crash recovery con mkCDrec in fase di test)