Grid Workshop
Bari 25/27.10.2004

Le giornate di Bari sono state molto interessanti per tutti.

Quello che segue rispecchia l'interesse dal punto di vista di un amministratore di sistemi UNIX che segue la farm di INFN Grid di Trieste.

- Test di GPFS a Catania — Rossana Catania INFN-CA
- Esperienza con Torque e Maui — Andrea Chierici INFN-CNAF
- Test di Quattor — Andrea Chierici INFN-CNAF
- La Grid di produzione — Cristina Vistoli INFN-CNAF

Introducing GPFS

- The General Parallel File System (GPFS) for Linux on xSeries® is a high-performance shared-disk file system that can provide data access from all nodes in a Linux cluster environment. Parallel and serial applications can readily access shared files using standard UNIX® file system interfaces, and the same file can be accessed concurrently from multiple nodes. GPFS provides high availability through logging and replication, and can be configured for failover from both disk and server malfunctions.

What does GPFS do?
Why use GPFS?

- **Presents one file system to many nodes – appears to the user as a standard Unix filesystem**
- **Allows nodes concurrent access to the same data**
- **GPFS offers:**
  - scalability,
  - high availability and recoverability
  - high performance
- **GPFS highlights**
  - **Improved system performance!**
  - **Assured file consistency**
  - **High recoverability and increased data availability**
  - **Enhanced system flexibility**
  - **Simplified administration**

Analysis of results

- **Reading from GPFS takes more or less the same time of reading from NFS**
- **Writing on GPFS is faster than on NFS and increases with the number of WNs**

Conclusions and outlook

- Preliminary I/O performance tests in the “NFS” configuration show a worse behaviour w.r.t. to native NFS (about 4:1) ; “Direct attached” is strongly suggest to improve performance
- Network bandwidth of the single servers is VERY important (GPFS sets down to the “slowest” node)
- The proper configuration with GPFS installed both on WNs and servers has been tested:
- Short term (next weeks): tests of reliability
- Medium term (by the end of the year): use GPFS to manage all the disk storage at Catania

Cos’è torque

- Tera-scale Open-source Resource and QUEue manager
- Scheduler basato su OpenPBS 2.3.12
- Dichiarato scalabile su sistemi fino a 2500 CPU
- Più di 100 patch e fix
  - Migliorata fault tolerance (controllo stato dei nodi)
  - Migliorata interfaccia di scheduling (reperibili informazioni addizionali e più accurate sui nodi)
  - Migliorata usabilità (log estesi e più comprensibili)

Usabilità

- Sviluppo continuo con susseguirsi di nuove versioni piuttosto regolare
- Gratuito, come OpenPBS, molto più affidabile
- Suggerita integrazione con prossima versione di INFNGrid (minimo sforzo, T1 pronto ad aiutare)
- Attualmente utilizzato in tutti i principali siti LCG basati su pbs

Maui

- Scheduler avanzato con caratteristiche uniche
  - Supporto per QOS
  - Capacità di fornire priorità a job
  - Supporto per advance reservation
  - Politiche di allocazione dei nodi configurabili
  - Fairshare

Torque + Maui

- E’ sufficiente sostituire il demone dello scheduler di torque/pbs con maui
- E’ importante fare attenzione alla accoppiata di rpm che si usano
- Versione attuale al T1 (aggiornamento imminente):
  - Torque 1.0.1p6
  - Maui 3.2.6p8

Esperienza con Torque e Maui al T1

CONCLUSIONI

- Accoppiata torque/maui molto valida e flessibile, consigliata a tutti i livelli
- torque non risolve comunque tutti i problemi
  - tier2/3, centri di calcolo utilizzino cache wrapper

    (“cache wrapper” scritto da Davide Salomoni di NIKHEF)
- Tier1 ancora al lavoro nell’identificare la soluzione definitiva (PBSPro vs LSF)

QUATTOR - Introduzione

- Sistema in grado di fornire installazione, configurazione e controllo automatizzato di cluster di macchine UNIX (linux, solaris)
- Progetto iniziato all'interno di EDG (WP4), ora coordinato dal CERN in collaborazione con altri istituti (UAM Madrid)
- Contenuto in ELFms (Extremely Large Fabric management system)

Sistemi operativi supportati

- RedHat 7.3 (server e client)
- RedHat 9
- SLC30x
- Fedora Core x
- Solaris 9

- Tutto dipende (o quasi) da ***rpmt*** e ***pkgt***

Differenze rispetto a LCFGng (1)

- Nuovo linguaggio di configurazione (pan)
  - Struttura gerarchica
  - Definizione di tipi e validazione
- Portabilità grazie ad architettura plug-in
- Componenti avanzati, con possibilità di condivisione di configurazione, nuove librerie
- Aderente agli standard dove possibile
  - Il sottosistema di installazione utilizza l'installatore di default (kickstart, jumpstart)
  - I componenti non sostituiscono gli script SysV e init.d

Differenze rispetto a LCFGng (2)

- Modularità
  - Interfacce e protocolli definiti chiaramente
  - Moduli indipendenti
  - Funzionalità “light” (ad esempio per installazione pacchetti personali)
  - Funziona su macchine già installate
- Scalabilità
  - Non più necessario NFS
  - Possibile utilizzo di proxy (http)
- Gestione avanzata dei pacchetti software
  - Possibilità di installare versioni multiple
  - Non necessita di header per gli rpm
  - ACL
- Supporto!!
  - EDG-LCFGng è congelato e obsoleto (non portato su nuove versioni di linux)
  - LCFG → EDG-LCFGng → quattor

Problemi – AI Tier1@infn

- AII (Automated Installation Infrastructure) non ancora perfettamente integrata
- La versione provata non è release ufficiale EGEE/LCG
- rpmt da alcuni problemi su sl302
- Incertezza su OS da usare (SL, SLC, CentOS, ...) e su versione (301, 302, 303)

A chi serve

- Al Tier1@infn, numero elevato di macchine
- Comunque suggerito per tutti i centri di calcolo con un numero di macchine maggiore di 20
- Non è necessario installare tutti i moduli per iniziare (SWrep,...)

Alternative

- Per grossi centri non ci sono attualmente alternative appetibili
- Per piccoli centri, kickstart + config manuale (INGRID ?)

Conclusioni

- Quattor è un software stabile e pronto per la produzione
- Necessario sforzo EGEE/LCG per preparazione template e sviluppo componenti per configurazione
- Nessun altro software da garanzie paragonabili

Grid di produzione
INFN GRID

Cristina Vistoli
INFN-CNAF
Bologna

Workshop di INFN-Grid
25-27 ottobre 2004
Bari

INFN-GRID Release

- **INFN-GRID is a customized release of LCG**

  - All resources are **fully managed** via LCFGng;
  - INFN-GRID does not support the middleware installation without LCFGng;

- INFN-GRID 2.2.0 release is based upon the official LCG-2.2.0 and it is 100% compatible;

INFN-GRID Release

- **Main differences from LCG 2.2.0 to INFN-GRID 2.2.0:**
  - Added support for DAG jobs; (Directed Acyclic Graph)
  - Added support for AFS on the WorkerNodes;
  - Added support for MPI jobs via home syncronisation with ssh;
  - Documented installation of WNs on a private network;
- Added full function VOMS support:
  - INFNGRID, CDF are completely managed via VOMS server.

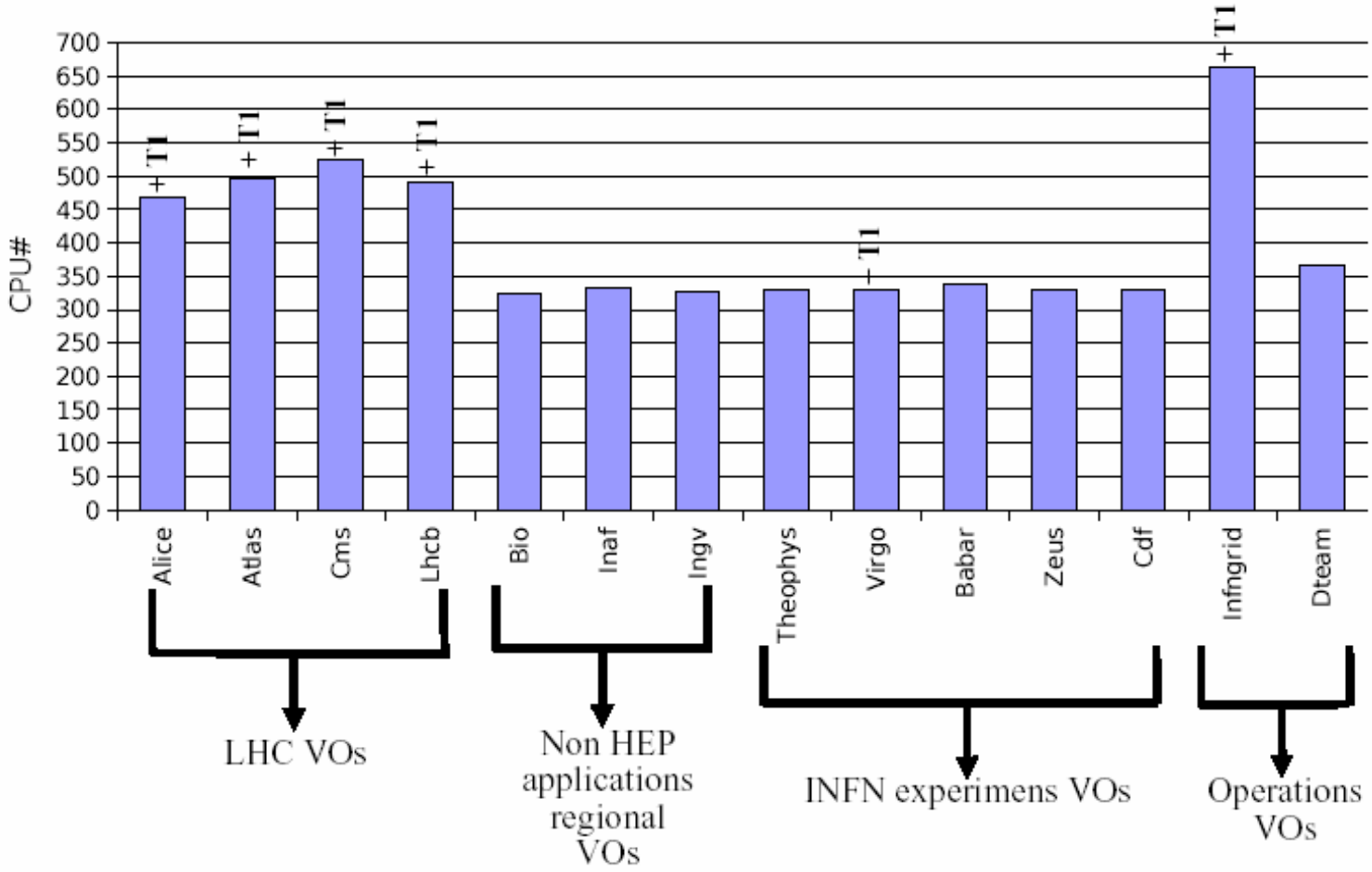
INFN-GRID:
Resources and supported VOs

Site name	CPU#	Storage (GB)	Alice	Atlas	Cms	Lhcb	Bio	Inaf	Infngrid	Ingv	Gridit	Theophys	Virgo	Babar	Zeus	Cdf
INFN-Bari	62	1800	x	x	x	x	x	x	x	x	x	x	x	x	x	x
INFN-Bologna-CMS	22	2700			x				x		x					
INFN-Bologna-CNAF	6	1900	x	x	x	x	x	x	x	x	x	x	x	x	x	x
INFN-Bologna	10	74	x	x	x	x	x	x	x	x	x	x	x	x	x	x
INFN-Cagliari	16	150	x	x	x	x	x	x	x	x	x	x	x	x	x	x
INFN-Catania	60	2100	x	x	x	x	x	x	x	x	x	x	x	x	x	x
INFN-Ferrara	12	25	x	x	x	x	x	x	x	x	x	x	x	x	x	x
INFN-Frascati	6	1100		x					x		x					
INFN-Lecce	2	18		x					x		x					
INFN-Legnaro	142	1300	x	x	x	x			x							
INFN-Milano	64	3200		x												
INFN-Napoli	24	900	x	x	x	x	x	x	x	x	x	x	x	x	x	x
INFN-Napoli-Atlas	12			x					x		x					
INFN-Napoli-Virgo																
INFN-Padova	104	9500	x	x	x	x	x	x	x	x	x	x	x	x	x	x
INFN-Pavia	6			x	x				x		x					
INFN-Perugia	6	225			x				x		x			x		
INFN-Pisa	13	18	x	x	x	x	x	x	x	x	x	x	x	x	x	x
INFN-Roma1-tier2	53	3000		x					x							
INFN-Roma1-Virgo	10	16	x	x	x	x	x	x	x	x	x	x	x	x	x	
INFN-Roma2	6	30	x	x	x	x	x	x	x	x	x	x	x	x	x	x
INFN-Torino	24	2000		x	x	x			x							
INFN-Trieste	2	30	x	x	x	x	x	x	x	x	x	x	x	x	x	x
INAF-Trieste	2	35						x	x		x					
TOTAL	664	30121	467	497	525	491	325	333	600	327	381	331	331	337	331	331
INFN-CNAF-CR	1570(**)	60000	x	x	x	x			x				x			x
TOTAL	2234	90121														

(**) Hyperthreaded

CPU versus VO

Number of CPU where the VO can run (not exclusively - shared)



Tier-1 resources not included. Tier-1 enables LHC, VIRGO and INFNGRID VOs

Upgrade / Installation activity

- Testing if "the grid is working" is not so easy;

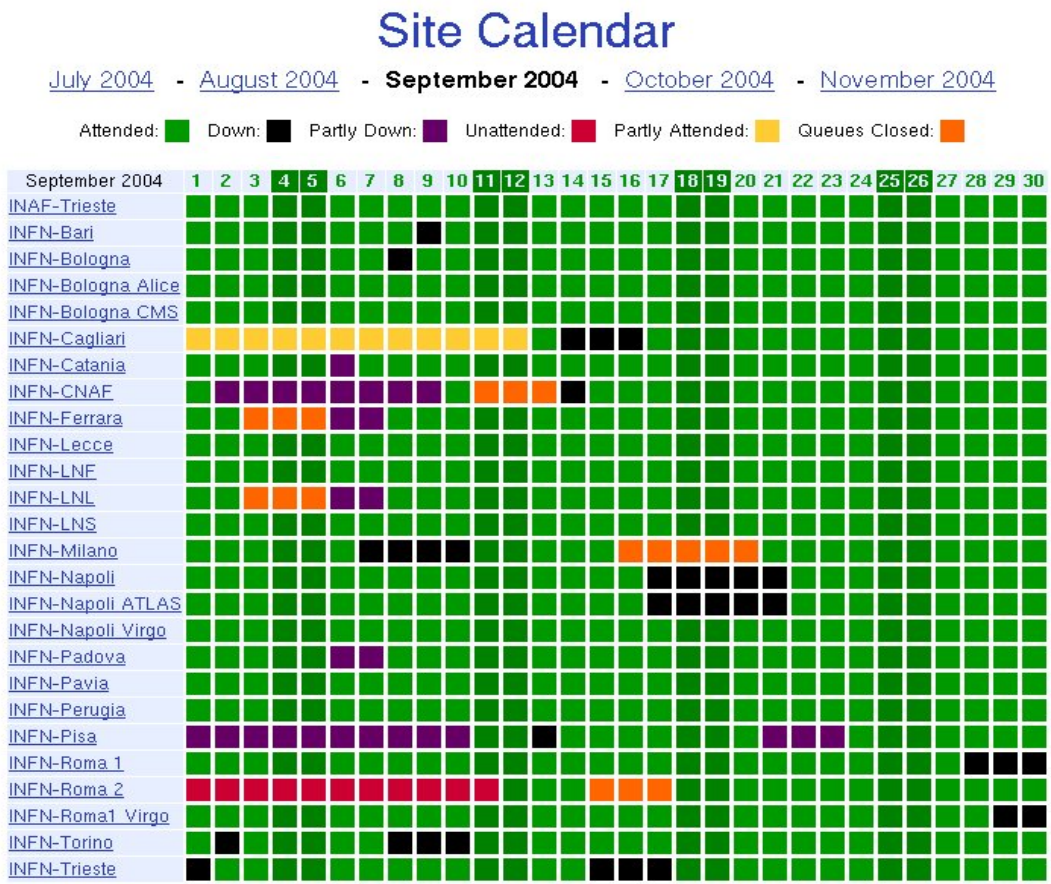
- Certification activity in INFN-GRID can be classified into four levels:
  - Local tests by the local resource center managers;
  - Certification tests by CMT Team;
  - Monitor tests by CMT Team;
  - The fourth level, certification on demand, made both by CMT Team and Application Teams.

Periodic test

- We periodically submit certification jobs to the sites in order to pro-actively find 'troubles' before users find them.

- CMT and system managers, could notify advices about their resources via web inserting a "**Downtime advices**".

- The **Calendar** shows the snapshot of the Production Service Status.



Ticketing system

- INFN-GRID ticketing system is used:
  - from users to ask questions or to communicate troubles;
  - from system manager to communicate about common grid tasks (ex: upgrading to a new grid release)
  - from CMT to system manager to notify a problem
- **Support Groups** are “helper” groups and they exist to resolve the obvious problems arising with the grow of the grid:
  - Support Grid Services (RB, RLS, VOMS, GridICE, etc) Group;
  - Support VO Services Group (each for every VO);
  - Support VOApplications Group (each for every VO);
  - Support Site Group (each for every site)
- **Operative Groups** Operative Central Management Team (CMT);
  - Operative Release & Deployment Team;

Users -> *Create a ticket*
Supporters/**Operatives** -> *Open the ticket*
Users and/or Supporters/**Operatives** -> *Update an open ticket*
Supporters/**Operatives** -> *Close the ticket*

GridAT - Grid Application Test

GridAT has the main goal to provide a general and flexible framework for VO application tests in a grid system.

It permits to test a grid site from the VO viewpoint.

Test report summary

Site	Generic	alice	atlas	babar	bio	cms	gridit	inaf	infngrid	ingv	lhcb	theophys	virgo
ba.infn.it	29/09/2004	06/04/2004	06/04/2004	06/04/2004	06/04/2004	14/04/2004			14/04/2004	07/04/2004			
bo.infn.it								06/04/2004	14/04/2004	14/04/2004			
cr.cnaf.infn.it									14/04/2004				
ct.infn.it		06/04/2004		06/04/2004	06/04/2004		06/04/2004	06/04/2004	14/04/2004	06/04/2004	06/04/2004		
fe.infn.it		06/04/2004		06/04/2004	06/04/2004				14/04/2004	14/04/2004	06/04/2004	06/04/2004	06/04/2004
lnl.infn.it									14/04/2004				
na.infn.it	30/09/2004	06/04/2004		06/04/2004	06/04/2004	06/04/2004	06/04/2004		29/09/2004				06/04/2004
pd.infn.it	06/04/2004	06/04/2004		06/04/2004			06/04/2004		14/04/2004			06/04/2004	
pg.infn.it			06/04/2004						14/04/2004		06/04/2004	06/04/2004	06/04/2004
pi.infn.it		06/04/2004		06/04/2004			06/04/2004	06/04/2004			06/04/2004	06/04/2004	
roma1.infn.it									14/04/2004				
to.infn.it	13/04/2004								14/04/2004				

Results are stored in a central database and browsable on a web page so it will be also used for certification and test activity.

Attivita' in corso

- Sistema di supporto: integrazione in EGEE e copertura supporto distribuito
- Evoluzione di Gridice per job monitoring, application monitoring, SLA monitoring, urgente configurazione notifiche
- Integrazione di DGAS in INFN-GRID → amministrazione sistema di accounting
- Porting di INFN-GRID a SL : nuovo sistema di installazione e configurazione
- Operation support infrastruttura EGEE/LCG a 'rotazione' tra IT/CERN/UK/FR
- Training: corso base e avanzato
- Allargamento infrastruttura a sedi non INFN: Spaci, Enea, etc
- Amministrazione Policy
- Pre-production service per definire il programma di migrazione a Glite
- Middleware certification testbed
- Operational requirements per il middleware

Link

- Fourth INFN Grid Workshop Bari, October 25 - 27 2004
  - http://www.ba.infn.it/~gridwks