Esperienza di un sito INFNGrid che offre risorse di calcolo e supporto T3-like ad utenti locali

Mirko Corosu
Atlas Italia – Tier-3 task force
Workshop CCR-INFNGRID 2010

Sommario

- Breve panoramica della situazione dei T3 integrati in GRID in Italia (ATLAS,CMS,LHCb,ALICE)
- Modello di lavoro e requirements per un T3/GRID in ATLAS
- Esperienza di due T3 integrati in GRID per Atlas (INFN-GENOVA,INFN-ROMA3)
- Esperienza di un T3 integrato in GRID per CMS

Breve panoramica della situazione dei T3 integrati in GRID in Italia

Breve panoramica su T3/GRID

- ATLAS:
  - Modello di lavoro e requirements di un T3 integrato in GRID sono in definizione. E' attiva una task force sull'argomento che ha composto un documento preliminare
- CMS:
  - Vari T3 integrati in GRID sono gia' operanti in diverse sedi. L'attivita' del gruppo di lavoro sui T3 si sta concentrando sul coordinamento
- LHCB e ALICE non sembra abbiano ancora preso in considerazione la tipologia T3/GRID

Modello di lavoro e requirements per un T3/GRID in ATLAS

Funzione di un T3 integrato in Grid e HW requirements per ATLAS

- Simulazioni di piccoli campioni di eventi di segnale e di fondo per il processo di interesse
- Analisi finale degli eventi reali selezionati da preprocessing su T0,T1 e T2
- Strumenti da supportare:
  - Athena
  - Root/Proof
- Sviluppo software e/o simulazione:
  - 2 TB di disco + 8 core per utente
- Analisi interattiva (Root/Proof)
  - 2 TB di disco + 8 core per utente

Modello di lavoro per un T3-GRID per ATLAS (1)

- Sviluppo e test rapido di codice in maniera interattiva
- Test intensivo del codice via batch system locale
- Test preliminare sul nodo GRID locale
- Invio di produzioni private o di gruppo sulla GRID

Modello di lavoro per un T3-GRID per ATLAS (2)

- Di norma l'output dei job locali (interattivi e batch) viene salvato su file system locale condiviso con tutti i nodi
- L'output dei job Grid locali viene salvato sugli space token serviti dall'SRM locale e registrati nei cataloghi LFC e DDM
- Gli utenti che eseguono in interattivo o via batch system locale hanno accesso readonly ai file presenti sugli space token

Configurazione dello Storage

- Gli space token sono gestiti attraverso StoRM + filesystem parallelo (GPFS)
  - Acceduti readonly via posix da tutti i nodi comprese le UI
  - Acceduti read/write via SRM tramite DQ2
- Space token che devono essere definiti:
  - SCRATCHDISK: file di output che vengono cancellati periodicamente
  - HOTDISK: file di calibrazione in formato POOL/ROOT
  - LOCALGROUPDISK: file di output che si desidera mantenere per periodi prolungati
- Viene inoltre resa disponibile un'area dati invisibile alla GRID per le produzioni locali (interattivo o batch)

Installazione software

- Installazione ed aggiornamento automatici via GRID (ATLAS Installation System – Alessandro DeSalvo)
  - Athena
  - DQ2
- Il software viene installato da un job GRID su un'area esportata a tutti i WN e le UI (di norma via NFS)
- L'utente T3 puo' richiedere l'installazione di versioni particolari via interfaccia web (se dotato dei permessi necessari)

Trasferimento dati

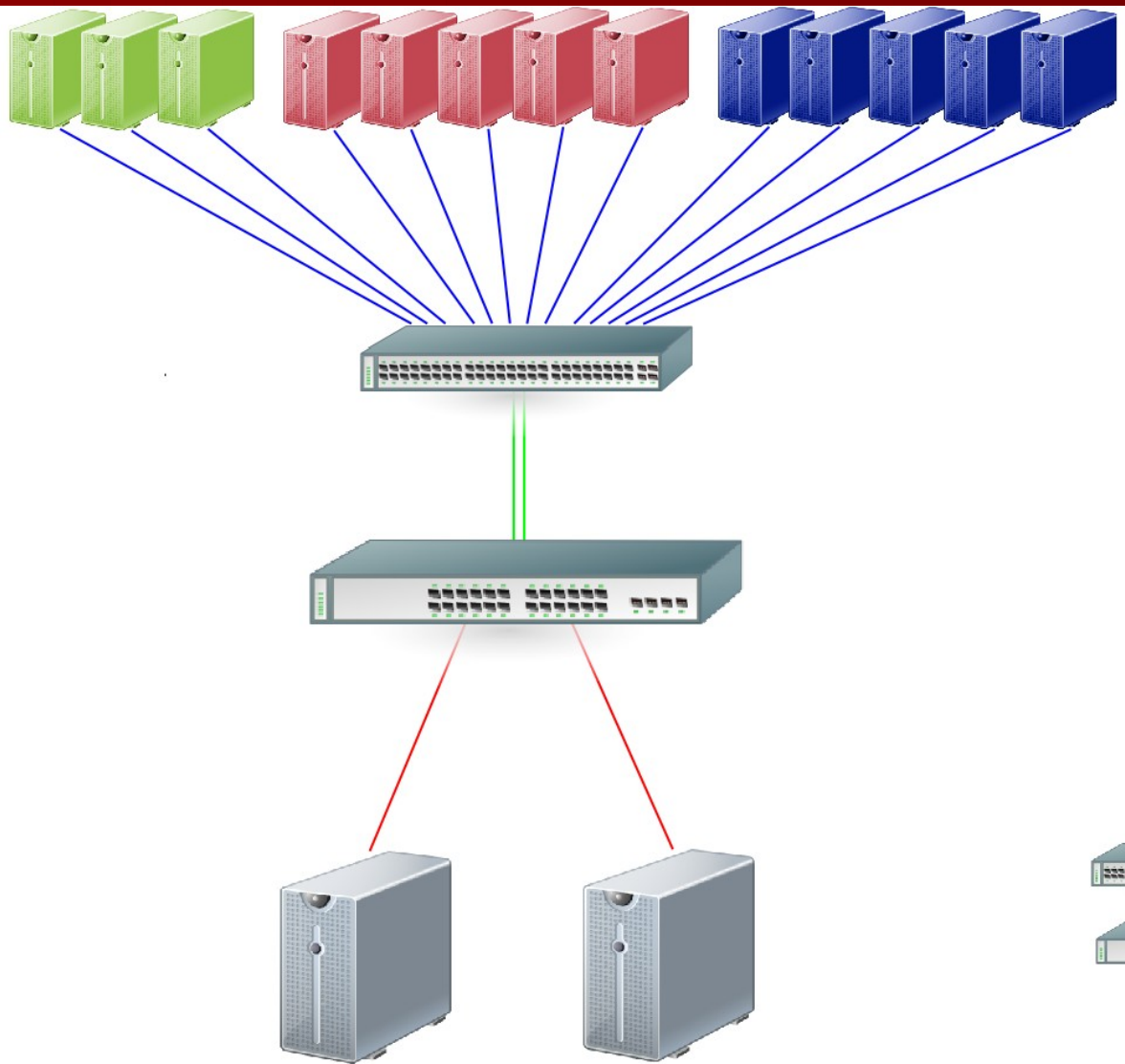
- Il trasferimento dati puo' avvenire:
  - Via DQ2 client o SRM client dalle User Interface
  - Via canale FTS attraverso subscription su DDM
- I files di calibrazione utilizzati da Athena vengono replicati sul token HOTDISK e sincronizzati con quelli della cloud italiana

Esperienza T3 integrati in GRID per Atlas (INFN-GENOVA,INFN-ROMA3)

Struttura T3 INFN-GENOVA

- Tre tipologie di nodo:
  - User Interface:
    - Esecuzione di job interattivi: montano l'area software via NFS, gli space token ed il file system locale via GPFS
    - Sottomissione sul batch system locale
    - Sottomissione job GRID
  - WN Grid:
    - Risorse INFN-GRID accessibili via GRID dagli appartenenti a tutte le VO supportate e dagli utenti T3 locali.
  - WN T3:
    - Risorse T3 accessibili via GRID e via batch system dai soli utenti T3 locali

Configurazione LAN INFN-GENOVA



1 GE

BONDING 2 x 10 GE

LAN 10 GE

DISK SERVER 10 GE CARD
8 Gb/s Dual head
Fiber Channel HBA

WORKER NODE T3

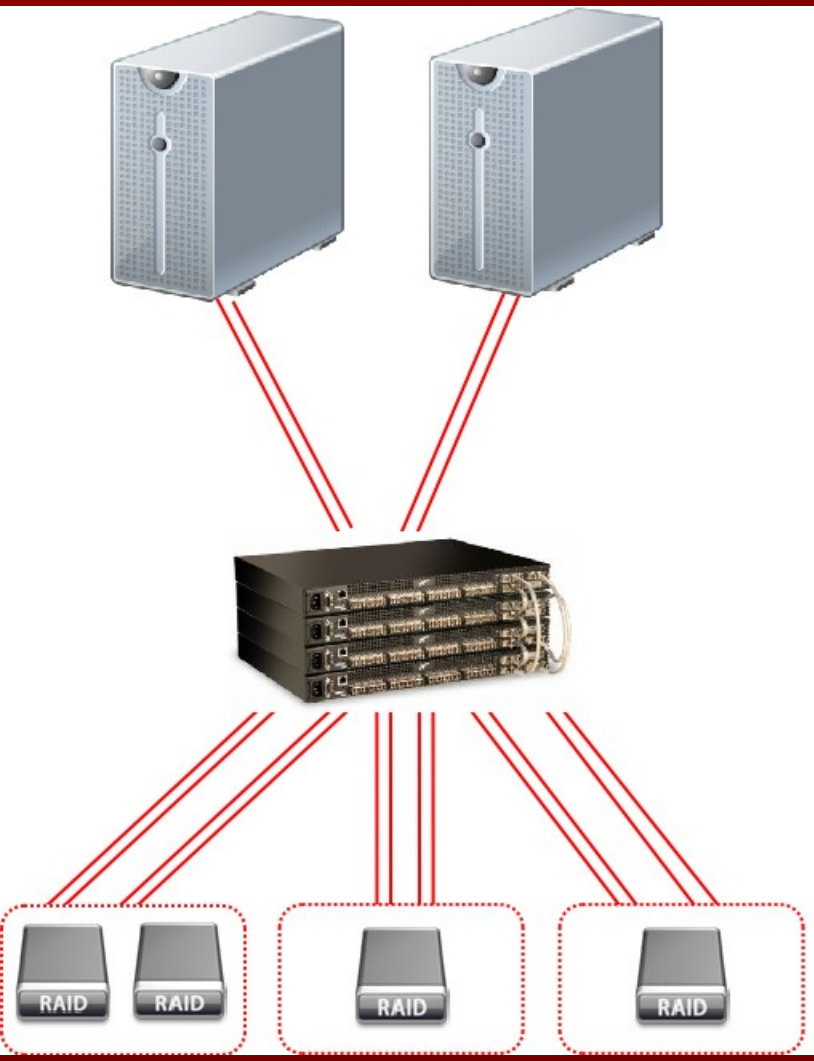
UI GRID
CLIENT LSF
ROOT/PROOF

WORKER NODE GRID

EXTREME NETWORK X350
48 port GE + 2 x 10GE uplink

EXTREME NETWORK X650
24 port 10 GE

Configurazione Storage INFN-GENOVA (1)



SAN FC 4 Gb/s

DISK SERVER 10 GE CARD
8 Gb/s Dual head
Fiber Channel HBA

2 + 2 (redundant) FC SWITCH
16 port 4 Gb/s

RAID6 VOLUME

INFORTREND EONSTOR S16F-R1430
4Gb/s redundant dual controller

Configurazione Storage INFN-GENOVA (2)

- I volumi raid fanno parte di un unico filesystem, sul quale sono configurati 3 fileset (con quota) per i token di Atlas:
  - ATLASHOTDISK (1 TB)
  - ATLASLOCALGROUPDISK (4 TB)
  - ATLASSCRATCHDISK (4 TB)
- ed un fileset ATLAS_GEN (9 TB), destinato all'utilizzo esclusivamente locale
- La quota viene gestita da GPFS ed e' modificabile dinamicamente
- Tutti i nodi (UI e WN) accedono via Posix a tutti i fileset
- E' configurata sull'SRM (StoRM) una ACL posix di default che permette al gruppo degli utenti del Tier-3 di avere accesso readonly ai dati contenuti nei token
- L'area software e' in un filesystem GPFS separato ed esportata via NFS da due CNFS server
- I gridftp services girano sugli NSD servers del filesystem contenente I dati

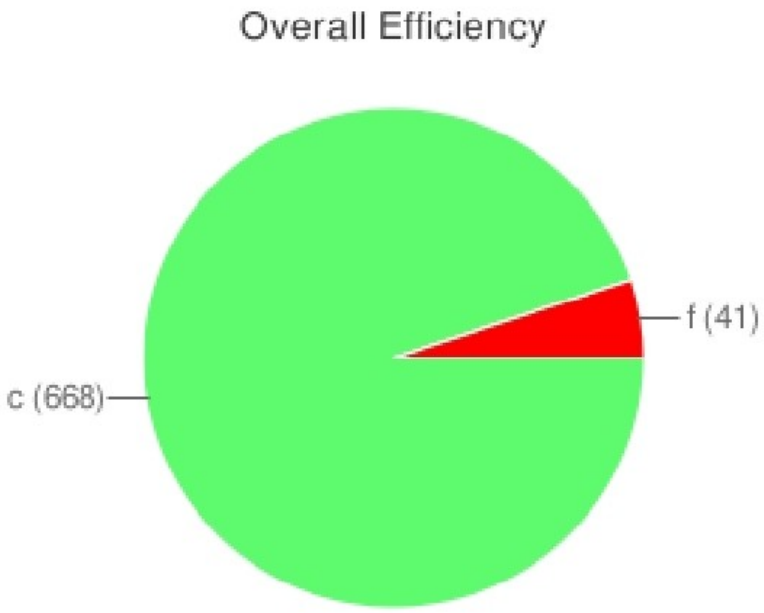
Integrazione User Database

- Le UI condividono lo udb centrale e montano le home directory centrali
- I WN ed il CE utilizzano uno user db locale, con home directory non condivise
  - su queste macchine vengono creati e mantenuti aggiornati account per gli utenti del T3 con stesso username/UID automaticamente via cron job
- Al fine di poter differenziare l'utilizzo delle risorse per gli utenti T3 rispetto all'utenza Grid generica, i certificati di tali utenti sono mappati sul CE a username particolari

Configurazione T3 INFN-ROMA3

- Configurazione simile a quella di genova (tranne il 10GE sui disk server)
  - 16 lame HP dual 5535 + 5 lame SuperMicro dual 5520
  - 2 disk server + 2 CNFS server per il software
  - LAN 1GE dual channel bonding per i WN
  - LAN 1GE quad channel bonding per i disk server
  - SAN FC con doppio switch e HBA dual head sui disk server
  - Storage:
    - 1 SUN/stk6140 (48 dischi da 0.5 TB)
    - 1 E4 6570 (6 dischi da 450GB + 50 dischi da 1TB + 32 dischi da 2TB)

HC test per INFN-ROMA3



- Discreta efficienza
- Eventrate valore poco significativo: dipende dal dataset. Comunque in linea con risultati di altri siti
- Software setup time troppo alto. Da verificare

Job eff	CPU eff	Events/Athena(s)	Eventrate	Software setup time	Prepare inputs time	Athena running time	Output storage time
0.94	40.20	7.66	7.69	236.73	19.62	26350.96	131.36

Futuro per T3-Grid di Atlas

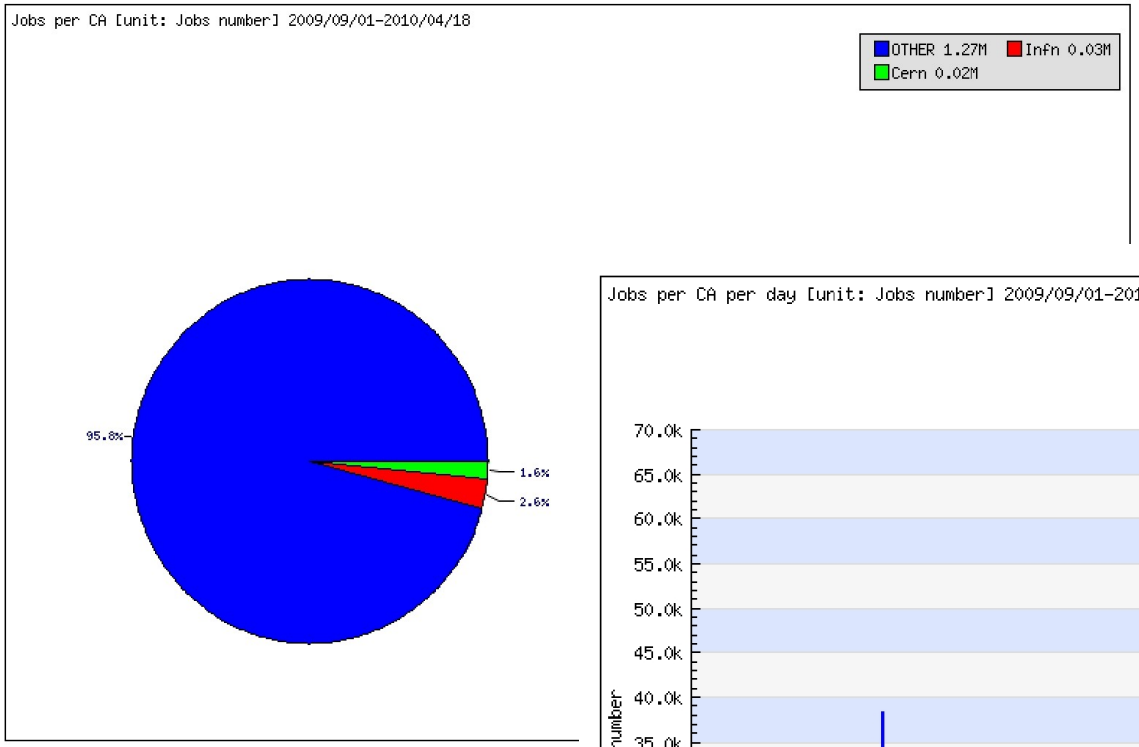
- Test di funzionalita'
  - per l'interattivo
    - interattivo: OK
    - Proof: da fare
  - Batch locale: OK
  - via GRID: da fare quando i trasferimenti di HOTDISK saranno completati
- Test di prestazioni da eseguire:
  - interattivo/proof
  - identificazione colli di bottiglia in funzione della tipologia di data set
  - Grid: da attivare i test Hammer Cloud per genova
- Da approfondire meccanismi di sharing delle risorse tra le tipologie di utilizzo (proof vs. batch/grid)
- Capire come il sistema interagisce con l'accounting di INFN-GRID

Esperienza di un T3 integrato in GRID per CMS

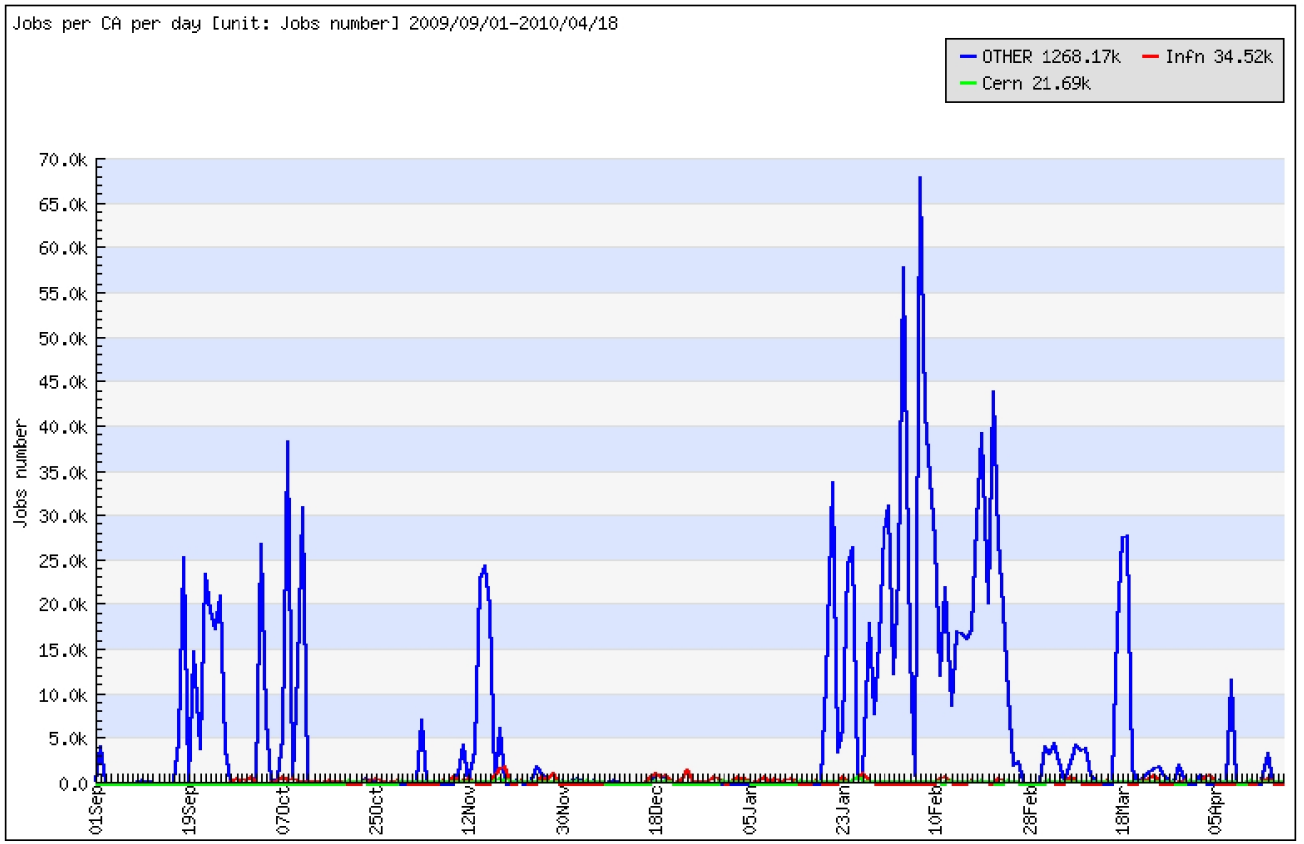
INFN PERUGIA/Farm status

- 35 TB spazio disco netto
- PhedEx up and running
- Storage elements DCache
- ~180 cores SLC4 + 80 SLC5 dedicati a CMS
- UI SLC5 per CMS + UI SLC4

INFN-PERUGIA/Stats



- La grande maggioranza dei jobs e' sottomessa localmente



- 1.3 Mjobs da sett 2009 ad aprile 2010

INFN-PERUGIA/Futuro

- Introduzione del CREAM CE
- Monitoring NAGIOS di sito
- Passaggio definitivo a SLC5 (quando lo consente la release INFN-GRID)
- Realizzazione di un testbed per lo studio e la valutazione di soluzioni per sistemi di storage (file server/file system)
- Consolidamento di soluzione Worker Node on demand
- Ottimizzazione di politiche di scheduling attraverso lo sviluppo di tool per la simulazione e lo studio sullo scheduler MAUI

Contributi

- Fulvio Galeazzi (ex. INFN-ROMA3)
- Leonello Servoli (INFN-PERUGIA)
- Mattia Cinquilli (INFN-PERUGIA/CMS)

Grazie !!