U-LITE: una proposta per il futuro calcolo scientifico ai LNGS

S. Parlati

LNF 28 novembre 2011

25 anni di calcolo scientifico ai LNGS

- Il Servizio di Calcolo e Reti
  - Nazzareno Taborgna (da 25 anni nel servizio)
  - Sandra Parlati (da 15 anni)
  - Stefano Stalio (da 12 anni)
  - Roberto Giuliani (da 12 anni)
  - Piero Spinnato (da 5 anni)
- Supporto di G. Di Carlo
- Andrea Donati nel servizio fino al 2005
- Borsisti e laureandi

25 anni di calcolo scientifico ai LNGS

- Anni ’80 e ’90: sistema fortemente centralizzato basato su cluster VMS e rete DECNET; Macro, EAS-Top, Gallex, LVD
- Dal 2000: diversificazione dei sistemi operativi (Digital Unix, Linux..); ambienti di calcolo eterogenei tra esperimenti: tendenza degli esperimenti a far ricorso a risorse di calcolo indipendenti (es. Borexino, Icarus, Opera).
- Il Servizio ha continuato a offrire un ambiente di calcolo completo (interattivo, batch, storage e backup) agli esperimenti (es LVD, Luna..). Storage e backup centralizzato utilizzato anche da esperimenti con farm di calcolo propria.
- 2010-2011:modello condiviso, gestito centralmente; nuove tecniche (virtualizzazione) permettono di avere ambienti eterogenei sullo stesso hardware -> U-LITE modello di calcolo per i futuri esperimenti.

Stato del calcolo scientifico ai LNGS

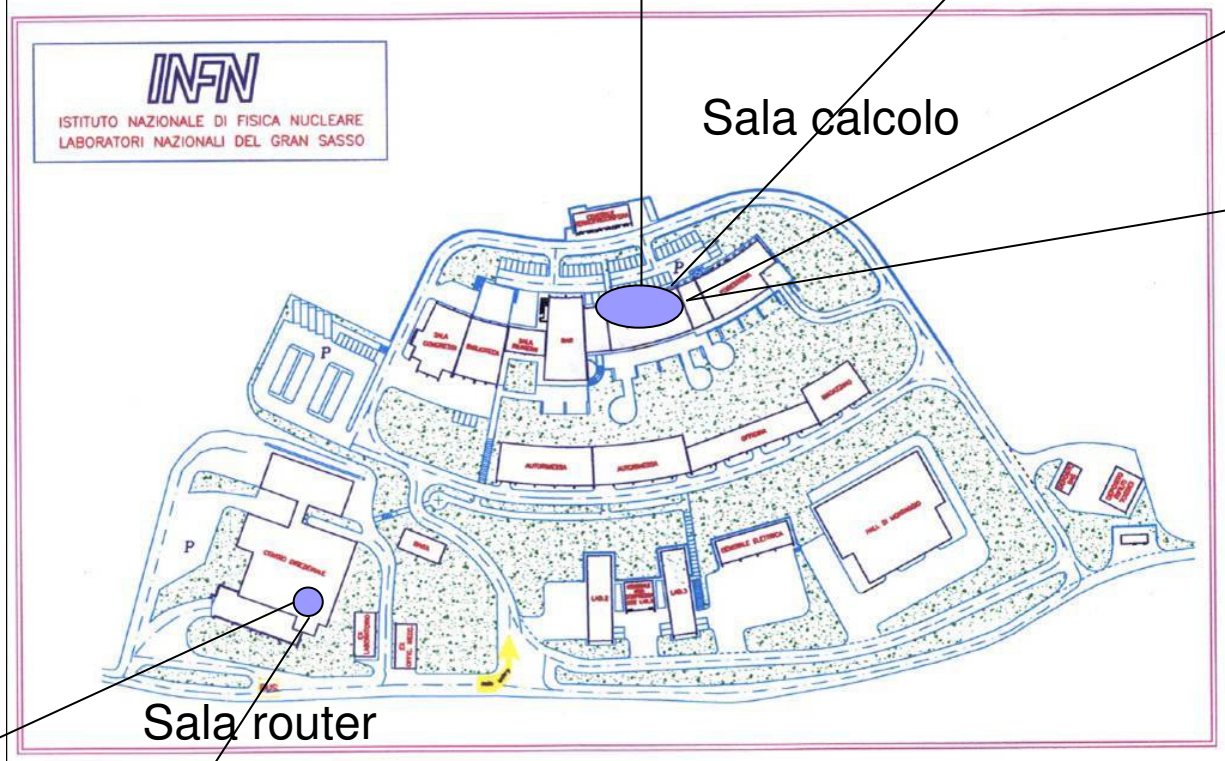
- Il Servizio di calcolo gestisce:
  - Sistemi di storage multipli per un totale di 150TB
  - Due librerie di nastri (120 slot ognuna) per il backup e l'archiviazione situate in due diversi edifici per aumentare l'affidabilità
  - La cella AFS /afs/lngs.infn.it
  - Servizio di public login: circa 600 utenti
  - Cluster LSF
  - Kerberos e LDAP per utenti LNGS e per gli utenti delle farm di esperimento (xenon, warp, darkside,lucifer,gerda ,lvd,teo): circa 1000 utenti registrati.
  - Librerie, software scientifico, etc…

2 sale calcolo distinte, site in edifici differenti, per aumentare l'affidabilità dei sistemi e minimizzare il rischio di perdita di dati

Storage

Cpu

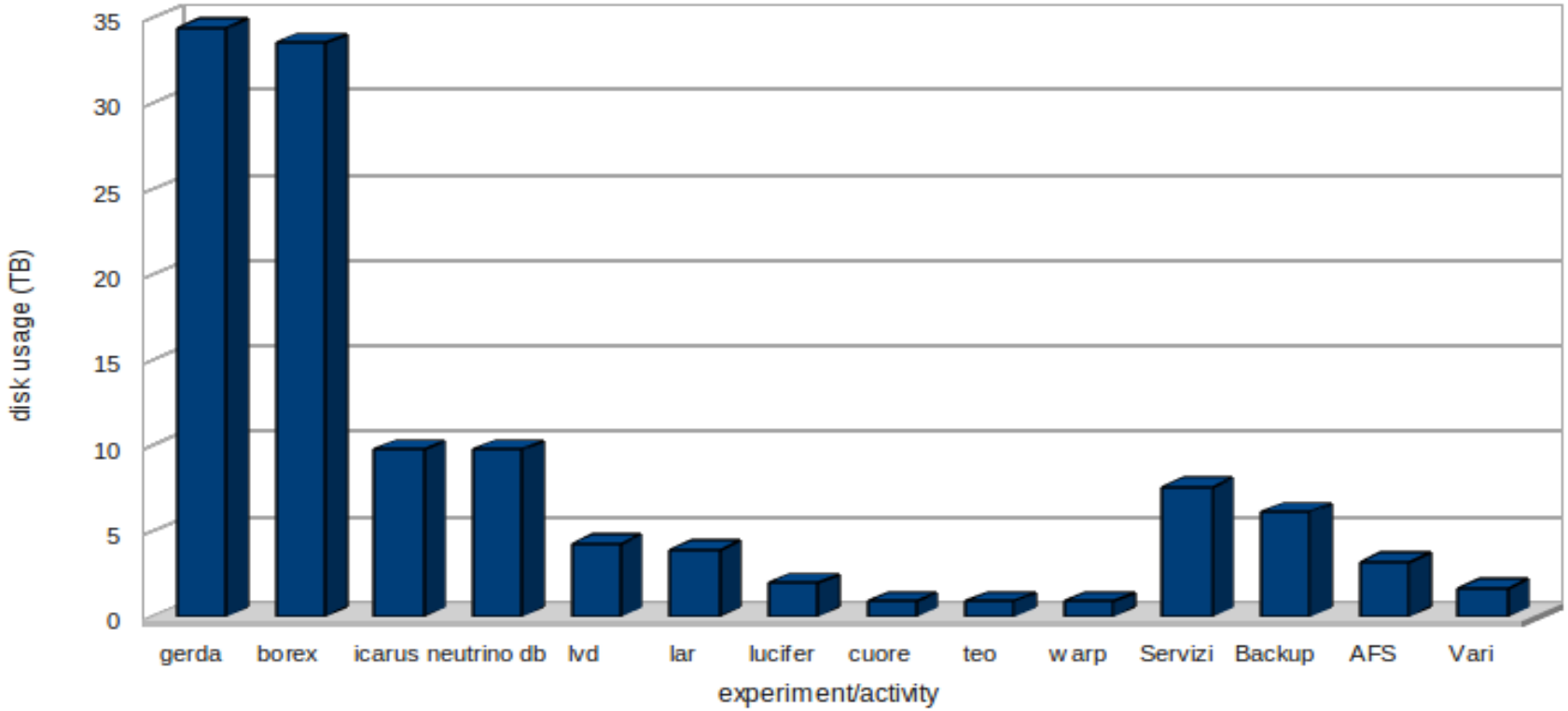
Backup



Servizi

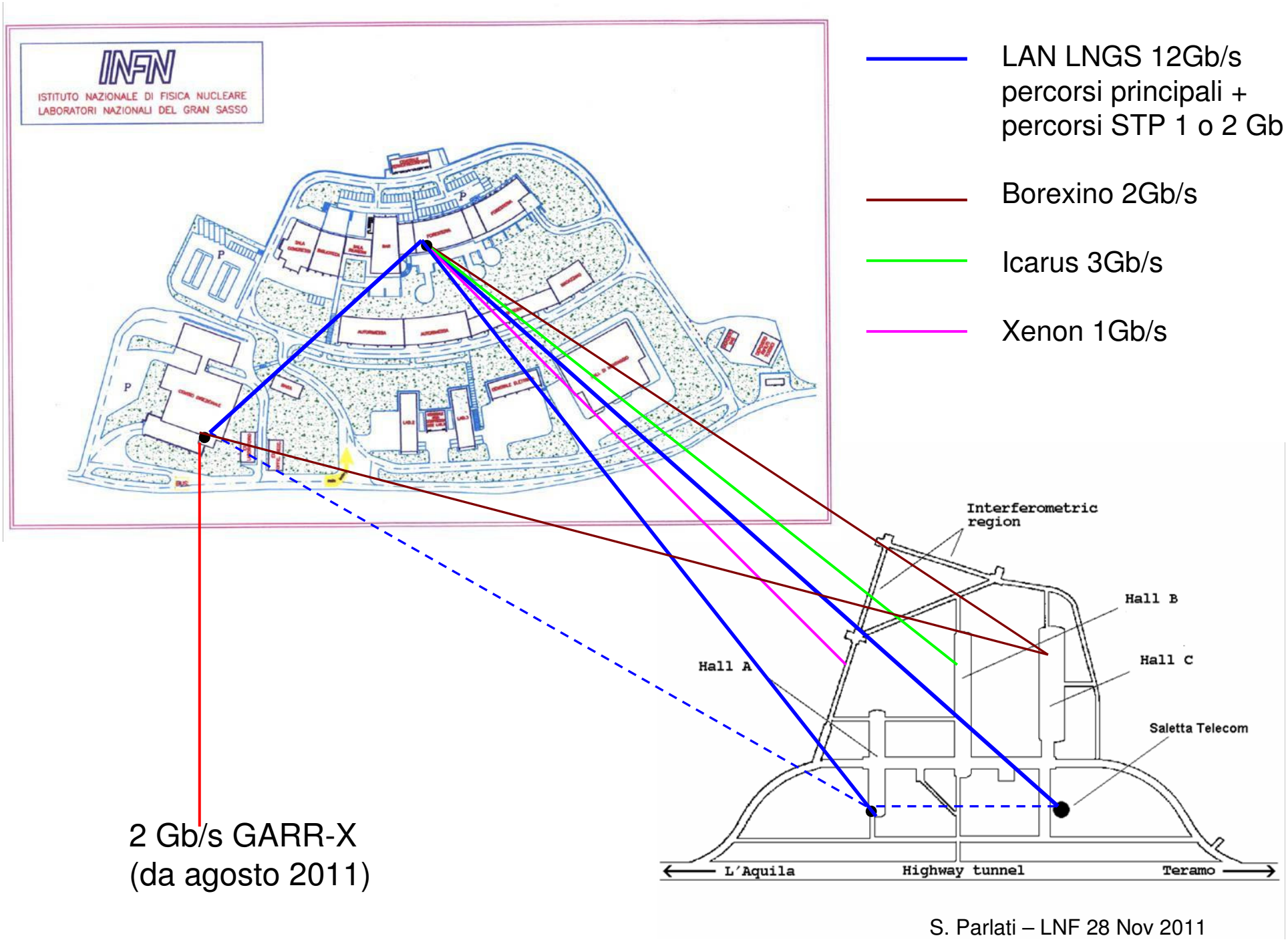
Entrambi i locali sono dotati di condizionamento e sono sotto gruppo di continuità e generatore diesel in caso di back out.

Es: storage gestito centralmente



Stato del calcolo scientifico ai LNGS (2)

- L'affidabilità della rete e' fondamentale quando i dati sono in un tunnel sotto la montagna e distanti 10Km!
- Il Servizio di calcolo gestisce la rete locale in ogni suo aspetto, dando grande importanza all'affidabilità dei collegamenti con il sito sperimentale:
  - Percorsi fisici ridondati e con cammini differenti
  - Link aggregation per aumentare la capacità e l'affidabilità dei link
  - Spanning tree: percorsi di backup pronti ad attivarsi nel caso di failure dei link principali
  - Centro-stella ridondati ai lab esterni e sotterranei e VRRP per ridondare il routing interno
- In aggiunta alla LAN condivisa, alcuni esperimenti utilizzano link punto-punto dedicati per il trasferimento dati dal DAQ allo storage.
- Il Servzio gestisce il link geografico con il GARR



S. Parlati – LNF 28 Nov 2011

Stato del calcolo scientifico ai LNGS (3)

- Tutti i servizi di rete (NFS, AFS, DNS, DHCP, Radius, etc..) lavorano in regime di alta affidabilità.

- Tutti i sistemi (calcolo, rete, servizi) sono controllati da un sistema di monitoring centralizzato (Nagios) che invia alert via mail o via sms in caso di guasti o malfunzionamenti.

Motivazioni per una nuova proposta

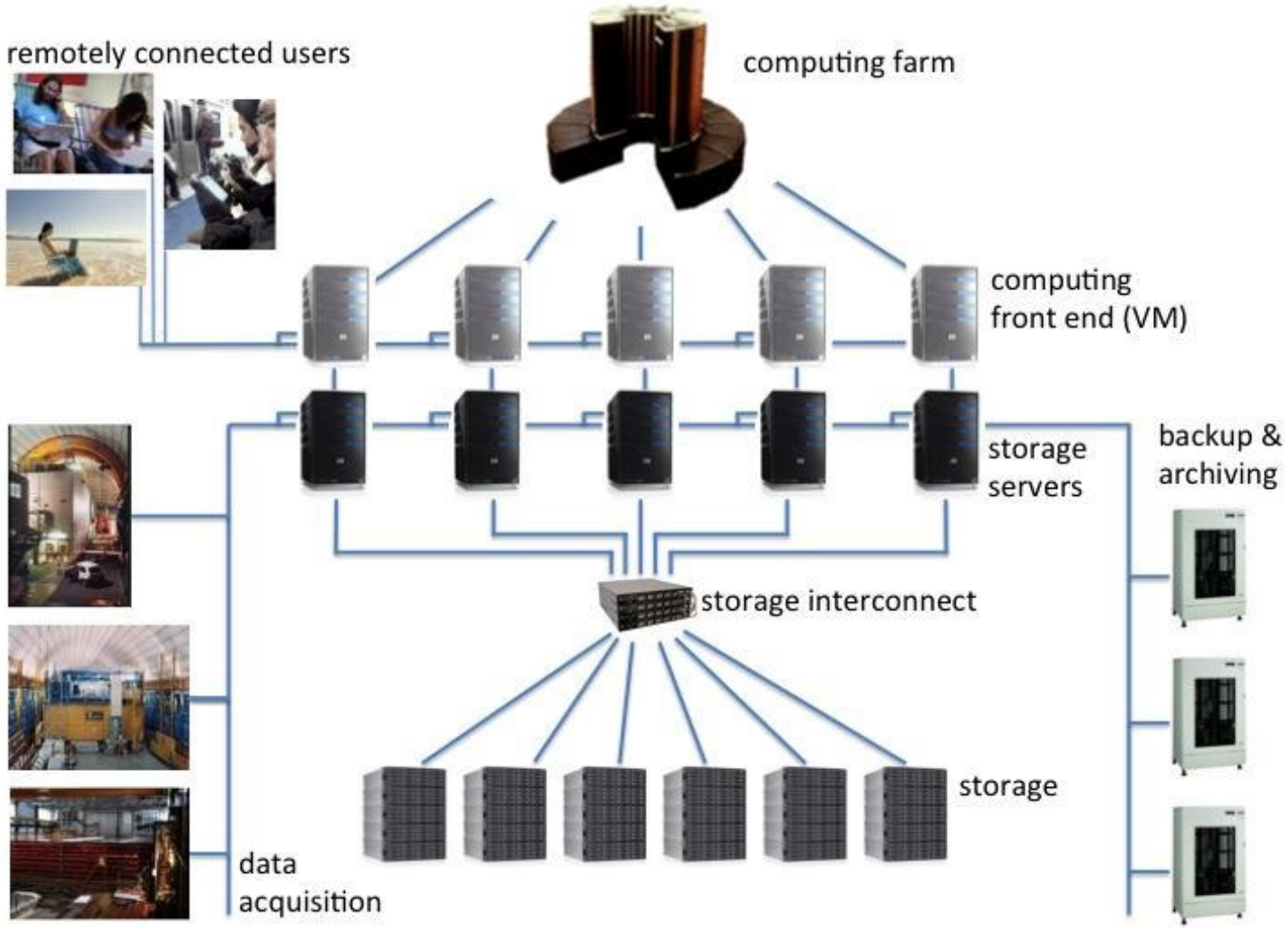
- Offrire un miglior servizio agli esperimenti: creare un ambiente di calcolo completo dove e' più necessario ossia lì dove il dato è acquisito.

- Risorse umane: gestione a carico del servizio di calcolo con personale esperto on-site permette agli esperimenti di risparmiare risorse umane ed previene problemi causati dalla gestione di risorse di calcolo da parte di personale non qualificato o non motivato.

- Aspetto economico: la condivisione delle risorse riduce le spese infrastrutturali e ne ottimizza l'uso.

- La virtualizzazione rende possibile la condivisione dello stesso HW da parte di ambienti sw diversi. L'esperienza acquisita con la virtualizzazione dei servizi ci e' tornata utile nella progettazione di un nuovo modello per il calcolo scientifico.

U-Lite: Unified LNGS IT Infrastructure



U-Lite: gli ingredienti (1)

- Storage servers:
  - 1 o 2 per ciascun esperimento. Copiano i dati dal DAQ sul sistema di storage, eseguono la preanalisi o la riduzione dei dati, salvano i dati sul sistema di backup e presentano i dati alle macchine di front-end e alla computing farm.
  - Per il ruolo delicato che rivestono, a metà tra il DAQ e l'off-line, la gestione degli storage server è a carico degli esperimenti, con possibilità per il Servizio di calcolo di intervenire con diritti amministrativi in caso di necessità.
- Computing front-end: 1 o più per esperimento. Macchine per il login interattivo, per l'analisi dati interattiva, per sviluppo del software e per la sottomissione dei job al computing cluster.

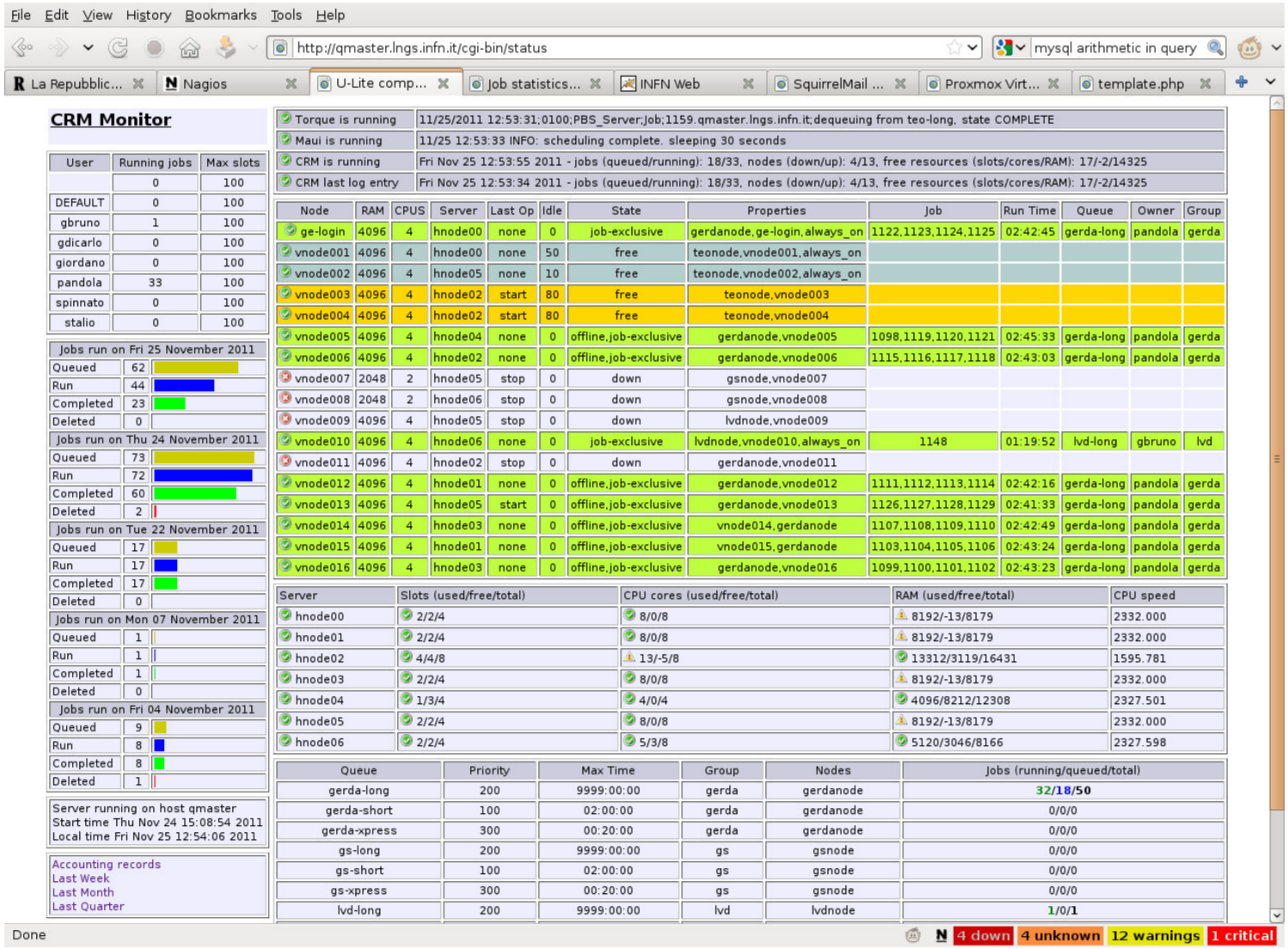
U-Lite: gli ingredienti (2)

- Computing farm:
  - più host virtuali (KVM) che girano su un cluster di server fisici multicore.
  - Ogni gruppo/esperimento ha i propri nodi di calcolo basati su un template di VM, sviluppato dalle collaborazioni sulla base delle proprie esigenze.
  - I job di simulazione o analisi vengono sottomessi ad un sistema di code (Torque/Maui) che li esegue sui nodi di calcolo virtuali.
- Storage: sistemi RAID ridondati (doppio controller, doppio alimentatore).
- Backup e archiviazione: librerie di nastri (attualmente LTO4); software di backup open source (bacula); formato di scrittura dei dati aperto.
- Autenticazione e Autorizzazione: utenti gestiti centralmente attraverso servizi Kerberos e LDAP del laboratorio.

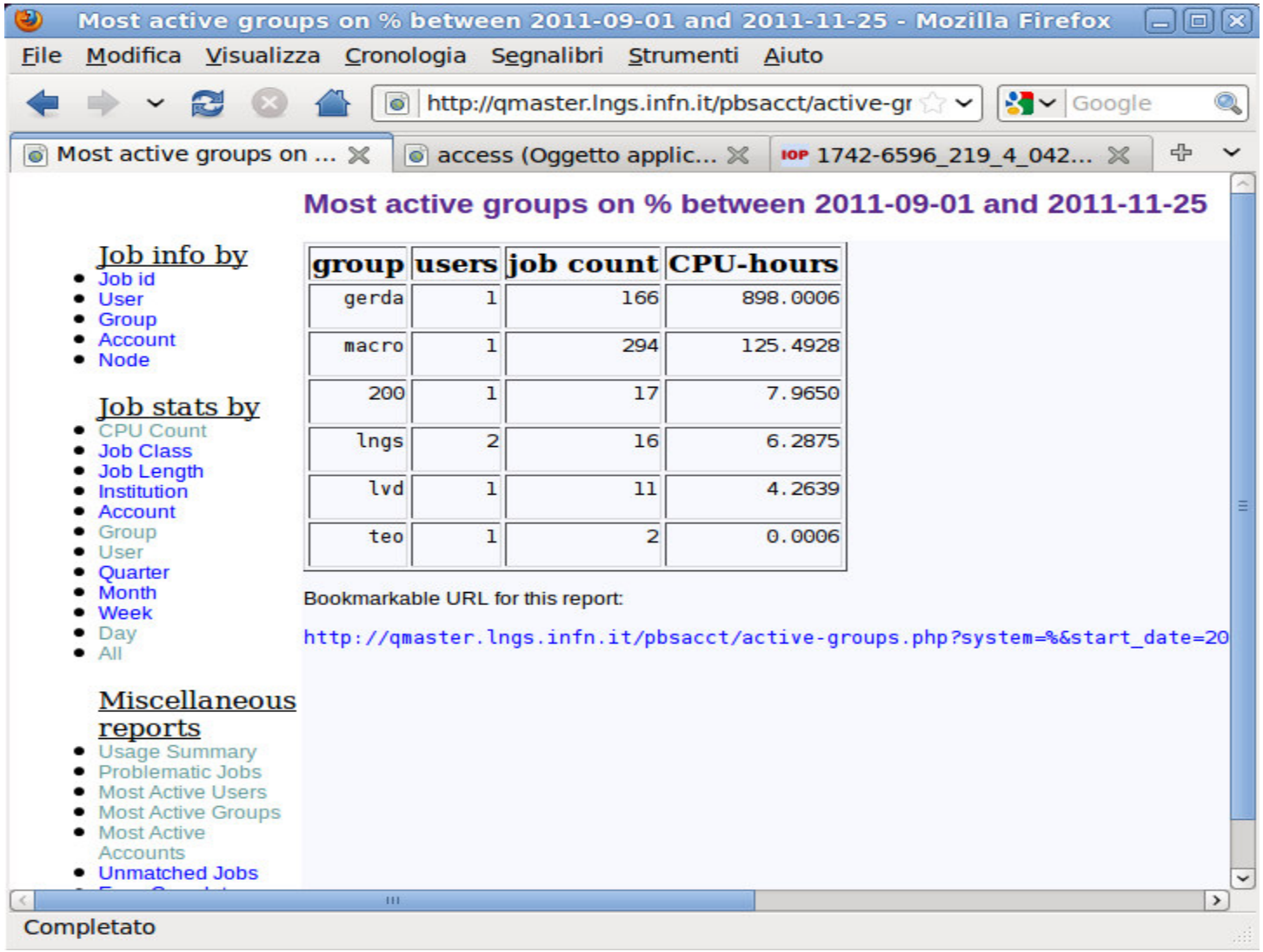
Stato attuale di U-Lite

- Tutte le componenti HW e SW di U-Lite sono testate e pronte all'uso: un primo nucleo di U-Lite e' già operativo da aprile 2011 (6 host fisici,2x4 core ciascuno, 10.25 HepSpec per core, costituiscono il cluster di calcolo)
- Storage e backup sono consolidati; il cluster di calcolo e' operativo, il sw di gestione dei job e' sotto sviluppo.
- Il sistema e' scalabile
- Il pool HW crescerà con l'arrivo di nuovi esperimenti in U-LITE.
- Sono disponibili strumenti per il monitoring on-line dei job e l'accounting.
- Alcuni gruppi cominciano a testare/usare U-Lite (es:Gerda)

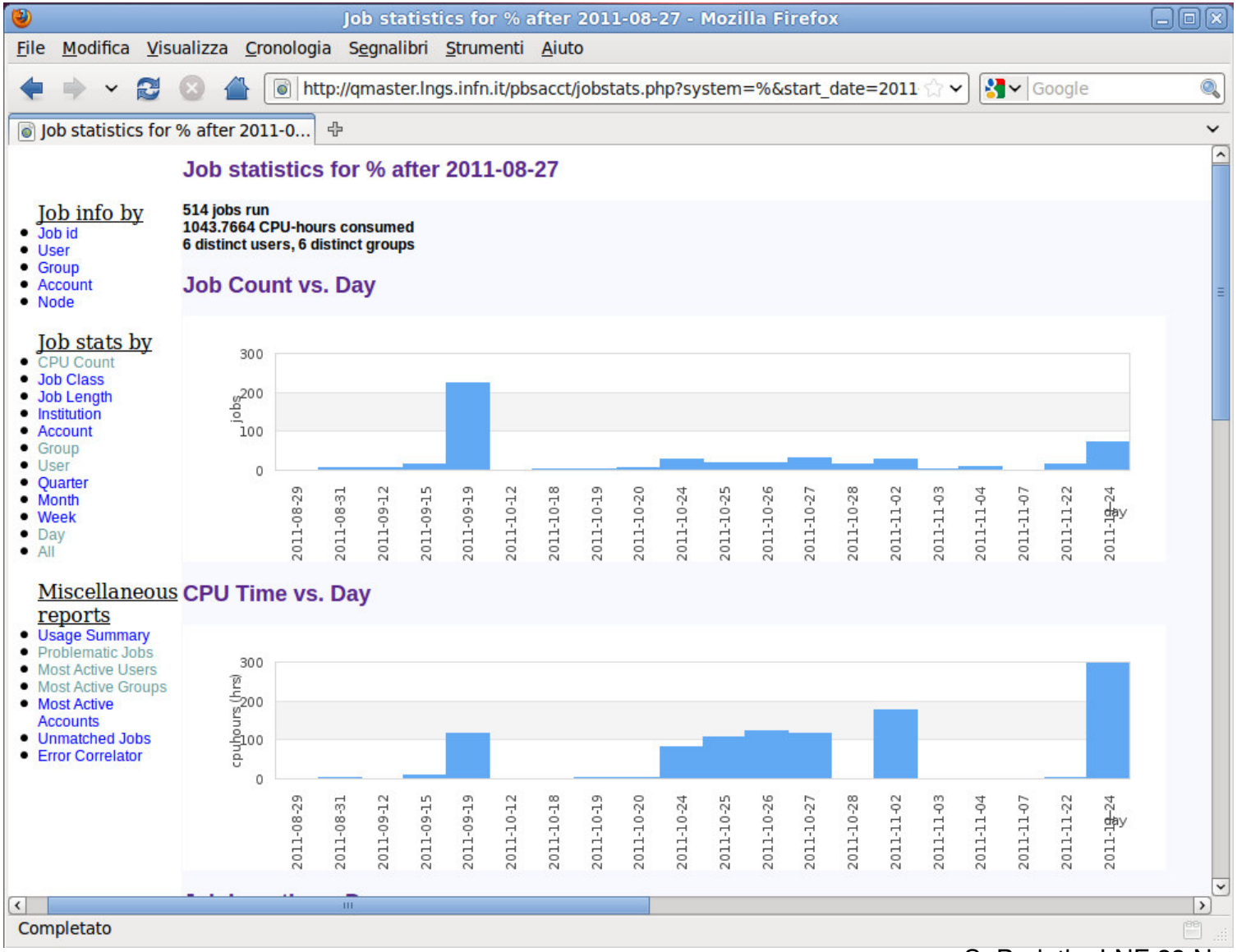
On-line job monitoring di U-Lite



Accounting


Most active groups on % between 2011-09-01 and 2011-11-25 - Mozilla Firefox
File Modifica Visualizza Cronologia Segnalibri Strumenti Aiuto
http://qmaster.lngs.infn.it/pbsacct/active-gr
Google
Most active groups on ...
access (Oggetto applic...
1742-6596_219_4_042...
Most active groups on % between 2011-09-01 and 2011-11-25
Job info by
Job id
User
Group
Account
Node
Job stats by
CPU Count
Job Class
Job Length
Institution
Account
Group
User
Quarter
Month
Week
Day
All
Miscellaneous reports
Usage Summary
Problematic Jobs
Most Active Users
Most Active Groups
Most Active Accounts
Unmatched Jobs
group users job count CPU-hours
gerda 1 166 898.0006
macro 1 294 125.4928
200 1 17 7.9650
lngs 2 16 6.2875
lvd 1 11 4.2639
teo 1 2 0.0006
Bookmarkable URL for this report:
http://qmaster.lngs.infn.it/pbsacct/active-groups.php?system=%&start_date=20
Completato

Accounting


Job statistics for % after 2011-08-27 - Mozilla Firefox
File Modifica Visualizza Cronologia Segnalibri Strumenti Aiuto
http://qmaster.lngs.infn.it/pbsacct/jobstats.php?system=%&start_date=2011
Google
Job statistics for % after 2011-0...
Job statistics for % after 2011-08-27
Job info by
Job id
User
Group
Account
Node
514 jobs run
1043.7664 CPU-hours consumed
6 distinct users, 6 distinct groups
Job Count vs. Day
Job stats by
CPU Count
Job Class
Job Length
Institution
Account
Group
User
Quarter
Month
Week
Day
All
jobs
300
200
100
0
2011-08-29
2011-08-31
2011-09-12
2011-09-15
2011-09-19
2011-10-12
2011-10-18
2011-10-19
2011-10-20
2011-10-24
2011-10-25
2011-10-26
2011-10-27
2011-10-28
2011-11-02
2011-11-03
2011-11-04
2011-11-07
2011-11-22
2011-11-24
day
Miscellaneous reports
CPU Time vs. Day
Usage Summary
Problematic Jobs
Most Active Users
Most Active Groups
Most Active Accounts
Unmatched Jobs
Error Correlator
cpuhours (hrs)
Completato

Chi e' coinvolto in U-Lite?

- Responsabilità
  - Gestione e coordinazione: S. Parlati
  - Responsablità tecniche: P. Spinnato and S. Stalio
  - Supervisione e pianificazione: G. Di Carlo
- Staff tecnico
  - Servizi di base: tutto il Servzio di Calcolo (5 persone)
  - Dedicato staff: P. Spinnato e S. Stalio
  - Necessità stimata di FTE:
    - **3 FTE**, distribuiti tra personale del servizio e personale in formazione (2 borsisti già richiesti per il 2012), per il setup di U-LITE
    - **1.5 FTE** distribuiti tra due o più membri del Servizio quando U-lite sarà pienamente operativo.

Modello di condivisione dei costi

- Il Laboratorio, attraverso il Servizio di Calcolo, ha finanziato l'attuale infrastruttura di U-Lite (rack, rete, SAN, librerie di tape) più un nucleo di macchine di calcolo, con il contributo della CCR.
- Costi vivi di cpu, dischi e tape sono a carico degli esperimenti;
  - dischi e tape sono ad accesso riservato
  - le macchine di calcolo si inseriscono in un ambiente condiviso; ogni collaborazione finanzia solo la cpu necessaria in media mentre i picchi di utilizzo sono coperti dalle risorse di altre collaborazioni o finanziate dal Servizio di Calcolo.
  - Il sistema batch di U-Lite prevede algoritmi di allocazione delle risorse per garantire sempre alle collaborazioni l'accesso alle risorse direttamente finanziate.
- U-Lite e' scalabile e stimiamo che l'infrastruttura attuale rimanga sostanzialmente invariata per una crescita di un fattore 10.
- Gli interventi fisiologici di mantenimento dell'infrastruttura sono a carico del laboratorio.

Vantaggi di U-Lite

- Presenza stabile di personale esperto on-site per la gestione dell'intera infrastruttura.
- Grande attenzione del personale all'affidabilità dei sistemi e consapevolezza dell'importanza dei dati (i LNGS sono il Tier-0 degli esperimenti!).
- Libertà per gli esperimenti di sviluppare l'ambiente di simulazione e analisi su qualunque piattaforma linux.
- Nessuna necessità di adeguarsi all'ambiente GRID
- Risparmio garantito grazie alla condivisione delle risorse

Links

- http://u-lite.lngs.infn.it

  Sito principale di U-LITE
- http://qmaster.lngs.infn.it

  Job monitoring e accounting
- http://computing.lngs.infn.it

  Sito del Servzio di Calcolo e Reti LNGS