



# **Cluster di calcolo per CMS a Milano Bicocca**



## INFN a Milano Bicocca (MiB):

### La Sezione

- sede distaccata della Sezione di Milano presso il Dip. Di Fisica dal 2000
- sezione autonoma dal settembre 2007

### Servizio Calcolo

- dominio [mib.infn.it](http://mib.infn.it) attivo dal 2000 (3 classi C), link a 32 Mb/s
- Servizi di rete estesi a tutto il Dipartimento (~400 utenti)

### Attivita' di calcolo

- Gruppo I (CMS), gruppo II (AMS,CUORE,HARP/MICE), gruppo IV
- nuova sala macchine dal giugno 2009

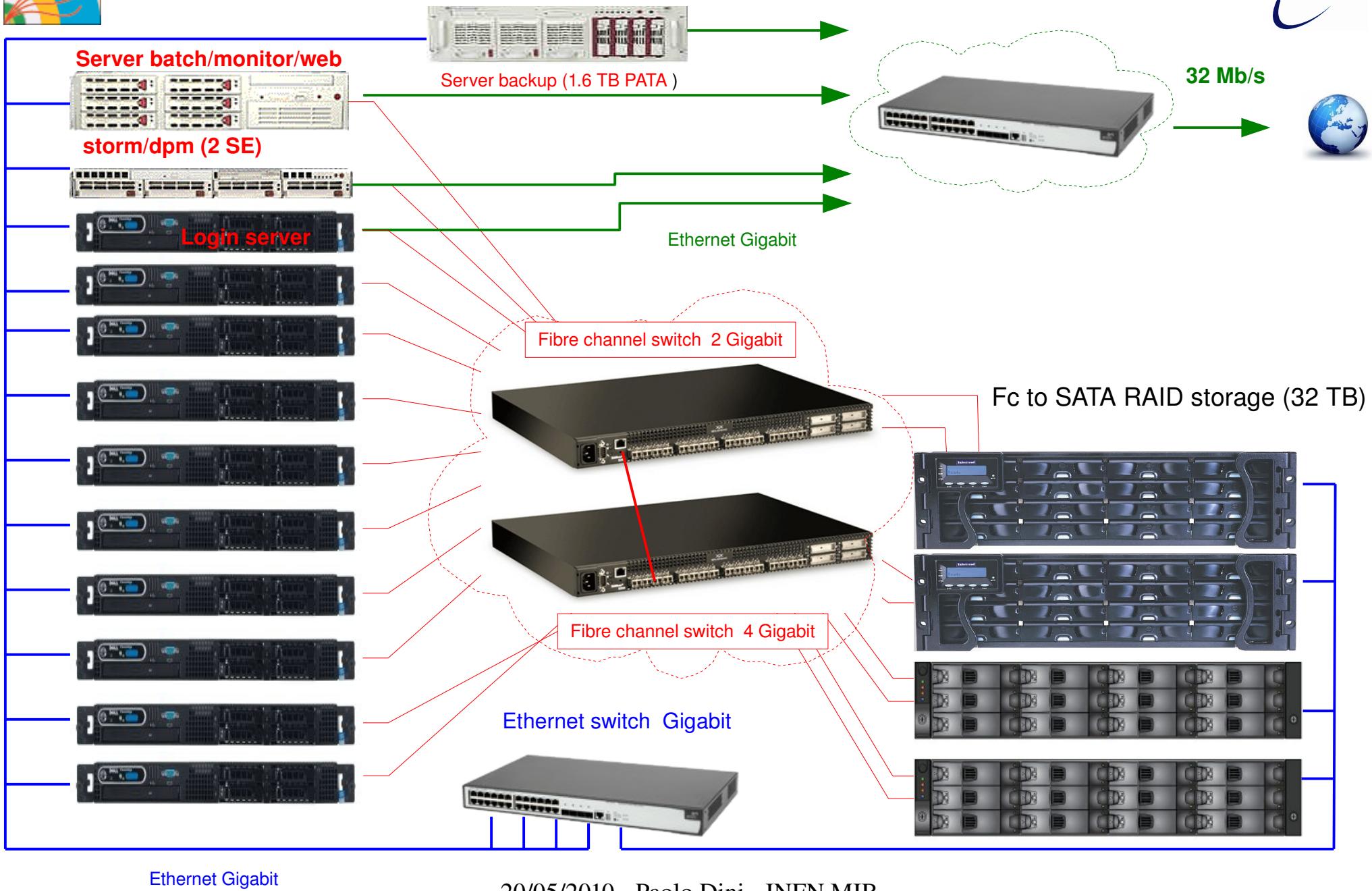
### CMS a Milano Bicocca

- gruppo ECAL
- gruppo Pixel Milano (Forward Tracker)
- Tipologia delle analisi: Higgs/electro-weak/B-Physics → ( N.Skim > 2 )

In totale ~20 tra ricercatori/tecnologi/assegnisti/dottorandi ecc.



# SAN - storage area network

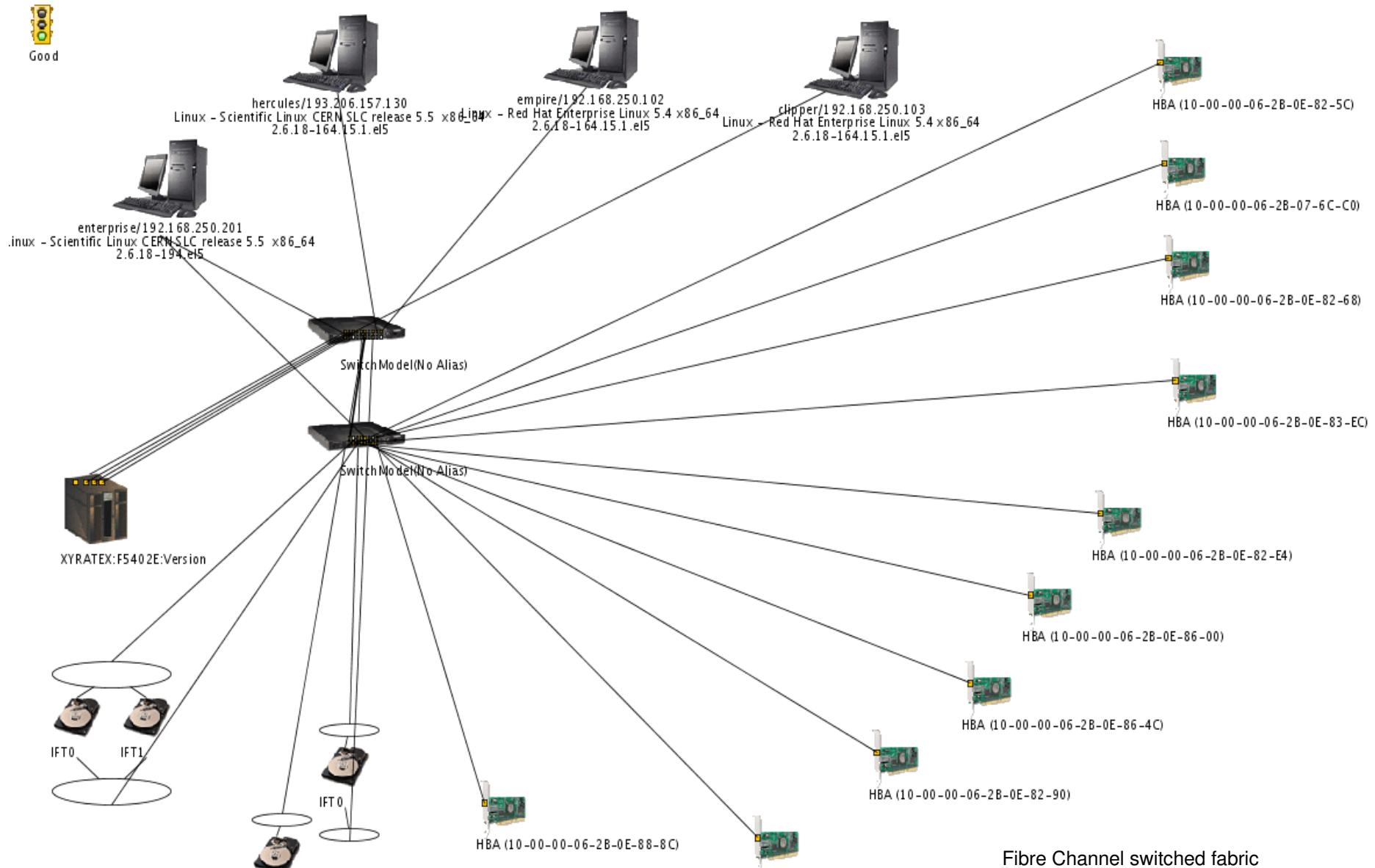


Ethernet Gigabit

20/05/2010 - Paolo Dini - INFN MIB



## Topologia del cluster





## Il cluster in pillole...

Nell'attuale configurazione:

- 15 server biprocessori multicore:

- 1 server di login/interattivo (8 core/16 GB RAM, dischi e alimentatori ridondati) **UI** (glite-3.2) slc5 x86\_64
  - 10 nodi di calcolo (4/8 core) ⇒ **60 slot di calcolo** slc5 x86\_64
  - 1 SE (dpm.mib.infn.it) con **DPM** (dpm 4.0) slc4
  - 1 SE (storm.mib.infn.it) con **STORM** (storm 1.4) slc4
  - 1 file server per backup on line (**1.4 TB**) slc5
  - 1 server per batch system, proxy, web monitor slc4
- (IP in classe nascosta, switch ethernet dedicato, connessione diretta al centro stella di rete)



direttamente connessi via Fibre Channel  
al pool di storage - file system GPFS

(traffico dati separato da quello di rete)

} (path ridondati - 2 switch qlogic a 2/4 Gb/s x 24 porte)

- 3 sistemi di storage ridondati per un totale di 56 slot dischi:

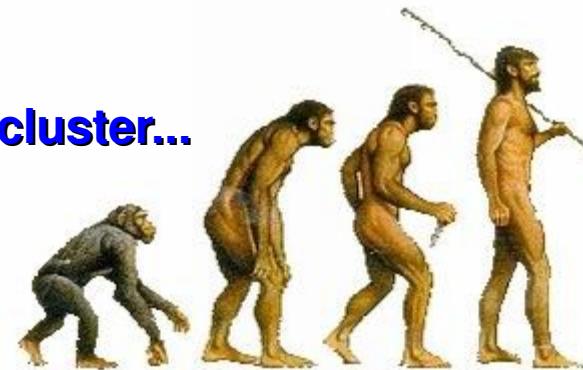
> **32 TB di spazio su disco** (RAID0 e RAID5, dischi da 500 GB/1TB)

Throughput (rw):  
200 MB/s per nodi a 2Gb  
400 MB/s per nodi a 4Gb

} X File da 1GB



## Evoluzione del cluster...



Harp+CMS 2001



Harp+CMS 2003



CMSCLUSTER (2010)

## Nuova sala macchine di Milano-Bicocca (giugno 2009)



- $\sim 40 \text{ m}^2$
- UPS 30 KVA (+15 KVA)
- Potenza frigo 96 KW
- 8 rack (4 cluster piu' svariati server)



## Il cluster in pillole...

- Batch system: [PBS/torque](#) – long/short queues
- Software CMS: [slc4\\_ia32\\_345](#) e [slc5\\_ia32\\_gcc434](#)
- Monitor: ganglia (+ Jobarchive), munin, nagios
- Frontier squid (registrato su <http://frontier.cern.ch>)
- Repository CVS locale
- Web server (<https://cmscluster.mib.infn.it>)
- **Virtualizzazione:** server virtuale (KVM) [GPFS client](#)

GRID...

- User Interface: [glite\\_3.2](#) per slc5 e [glite\\_3.1](#) per slc4
- [CRAB](#) client
- Testati 2 Storage Element : [DPM](#) e [STORM](#)

⇒ vantaggio di **storm**: accesso POSIX ai file su GPFS .

- Usando **CRAB** sul cluster ⇒ job su grid ⇒ scrittura su SE locale ⇒ **accesso diretto ai file su filesystem GPFS!**

[TESTATO CON SUCCESSO!](#)

```
[root@storm2 ~]# df
Filesystem      1K-blocks      Used   Available Use% Mounted on
/dev/hda6        7771892  1207332   6169768  17% /
/dev/hda1       585604    29328   458172   7% /boot
none            2073652      0   2073652  0% /dev/shm
/dev/hda2       12389352  4720752   7039256  41% /usr
/dev/hda3       8254272  479932   7355044   7% /var
s1box:/linux/dist 1464711168 1293294384 171416864  89% /s1box
/dev/gwter4     2899427328 2070467328 828960000  72% /gwter4
/dev/gwter5     1933277184 1916715520 16561664 100% /gwter4
/dev/gwter6     2899427328 283671888 2615755520 10% /gwter6
/dev/gwtera     13678938624 73460480 13597478144   1% /gwtera
/dev/gwterax1   3905982464 1621293824 2284688640  42% /gwterax1
/dev/gwterax2   3905982464 2011906816 1894075648  52% /gwterax2
/dev/gwterax3   3905982464 1093157632 2812824832  28% /gwterax3
/dev/gwpool     1952991232 1332596736 620394496  69% /gwpool
/dev/gwt3       2899427328 2098536960 800890368  73% /gwter3
[root@storm2 ~]# _
```



# MONITORING TOOLS



CMS@MIB



Welcome on CMS@MIB cluster Home page



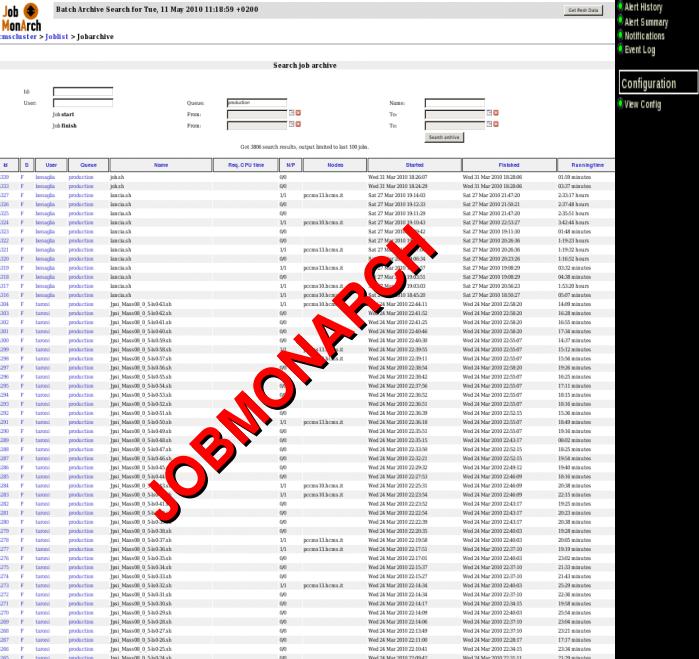
These are the web pages for the [computing](#) of CMS collaboration in Milano-Bicocca<sup>1</sup>. The main project is to offer a local computing resource to the CMS Milano-Bicocca collaboration for analysis data simulation. The cluster is composed by several computing nodes (multiprocessor, on rack-mounted servers) connected together via a Gigabit Switch in a private network. All the compute nodes are members of a [Fibre Channel Area Network](#), connected by a Fibre Channel Storage system. All the nodes have the same disk space via [GPT3.2](#) file system. The storage sys is controlled by an FC SAN, with 32 x 32 RAID0 disk array. The storage pool is mapped to all the compute nodes via a set of links to the storage, managed by the [TOMA](#) system. A gateway server allows users to login and to run interactively jobs. Two cluster nodes are designated as [GRID Storage Element](#), respectively with DPM and S installed. Users of either can run GRID jobs and redirect the output on the SE. The are also dedicated file server services for the management, monitoring and administration of the cluster, the dedicated for the on-line daily backup of the user area.

A photograph showing a row of glass bottles on a shelf, likely beer taps or kegs.



(Nodes colored by 1-minute load) | [Legend](#)

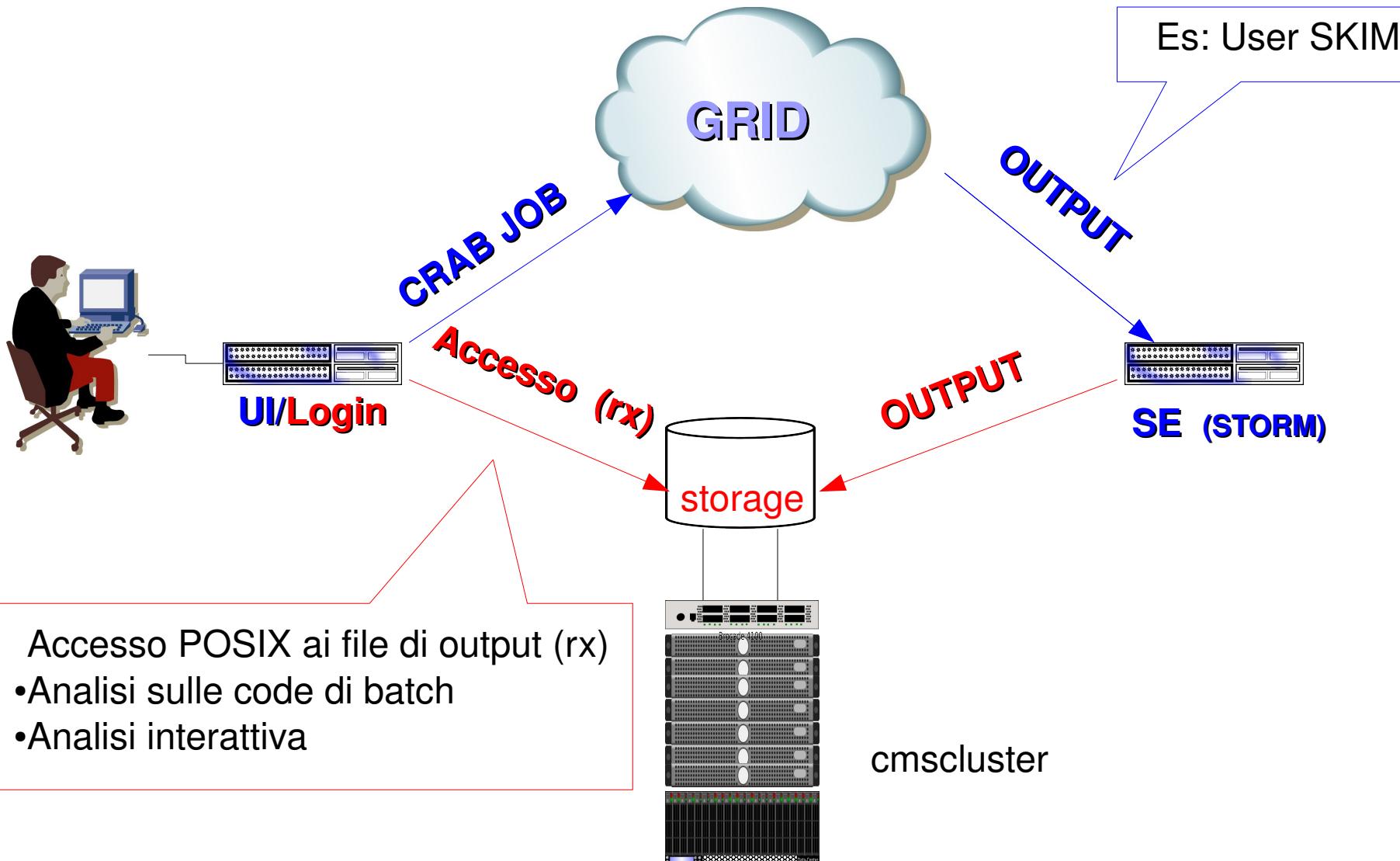
Ganglia Web Frontend version 3.0.6 [Check for Updates](#).  
Ganglia Web Backend (gmetad) version 3.0.6 [Check for Updates](#).  
Downloading and parsing ganglia's XML tree took 0.0173s.  
Images created with [RRDTool](#) version 1.2.30.



20/05/2010 - Paolo Dini - INFN MIB

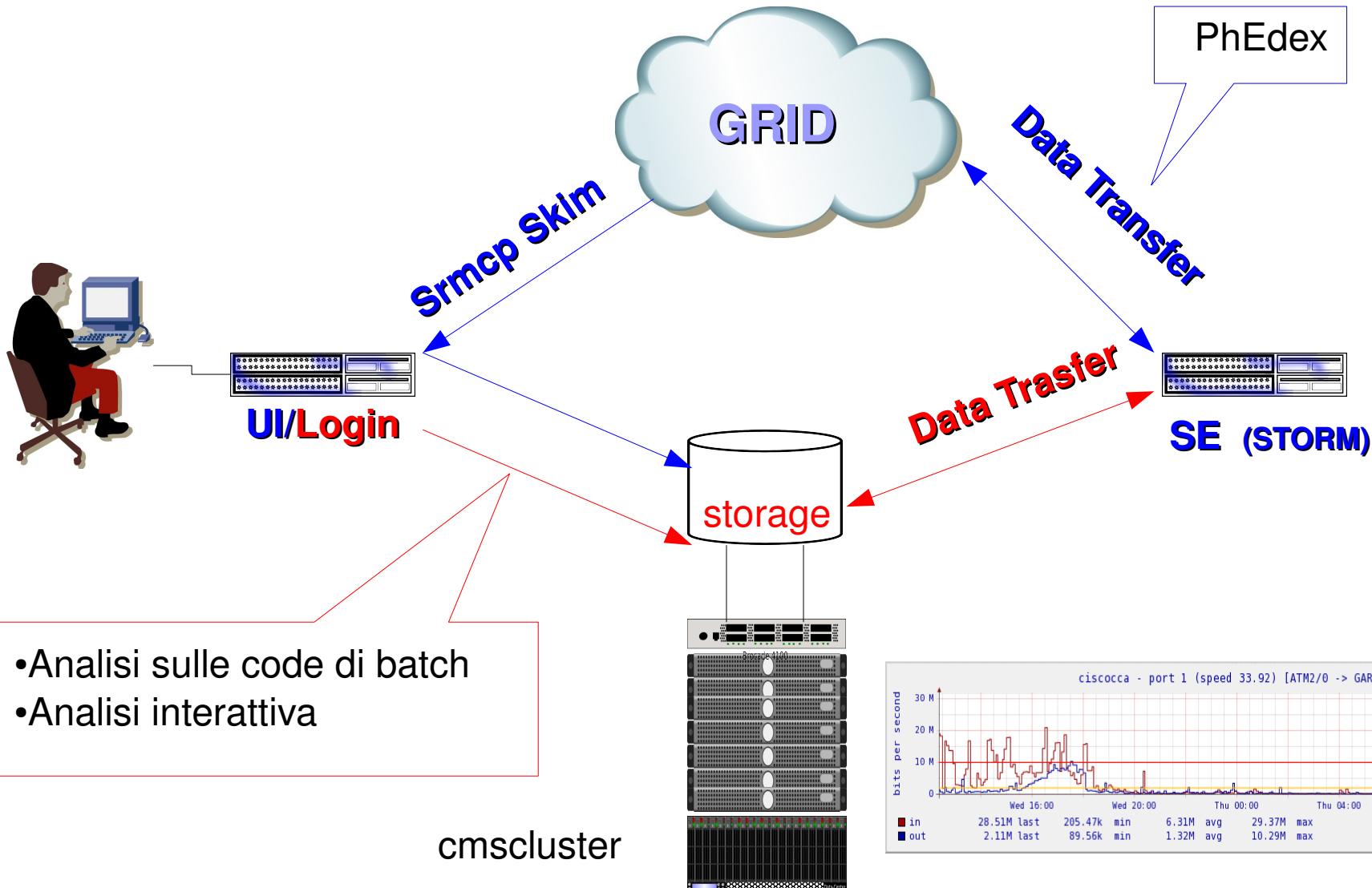


# USE CASE





# USE CASE





## Il cluster in pillole...

Osservazioni su cmscluster:



- servizio “best effort” (“one man show”)
- Accesso diretto ai dischi via FC - un solo tipo di filesystem (GPFS).
- Flessibile e scalabile
- ~ 20 TB in RAID0 (suddivisi in 6 dischi logici per limitare l'eventuale perdita di dati)
- **Soluzioni semplificate**
  - Autenticazione unix passwd (no LDAP) **solo sul server di login**
  - **Nessun nameserver** per la classe privata (risoluzione tramite /etc/hosts)
  - Gestione delle configurazione attraverso chiavi ssh
  - Installazione via rete di Linux (Kickstart/PXE-BOOT)
- Numero di utenti limitato: ~20 utenti
- **Modello di gestione “collaborativo”** per GRID/CMSSW (L. Sala co-admin)
- **Interazione diretta con gli utenti del cluster** (sviluppo collaborativo delle risorse di calcolo)



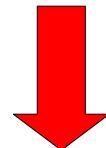
## Futuri sviluppi:

Dipendono dal modello di analisi che si vuole seguire, es:

- Produzioni MC locali (CPU+storage)
- User Skim o subSkim (storage, rete!)
- Data replica (storage, rete!)
- Condivisione di dati attraverso PhEdex (...)

... e dalla tipologia degli utenti

- Utenti locali
- Utenti locali+utenti esterni



- Connessione a 100 Mb/s (Giugno?)
- Acquisto di 1 JBOD (16 slot) ⇒ + 32 TB
- Acquisto di un nuovo file server x backup on line
- Nodi?
- **PhEdex: da venerdi' T3\_IT\_MIB**

# SAN Topology

