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 attivo dal 2000 (3 classi C), link a 32 Mb/s
- Servizi di rete estesi a tutto il Dipartimento (~400 utenti)

Attivita' di calcolo

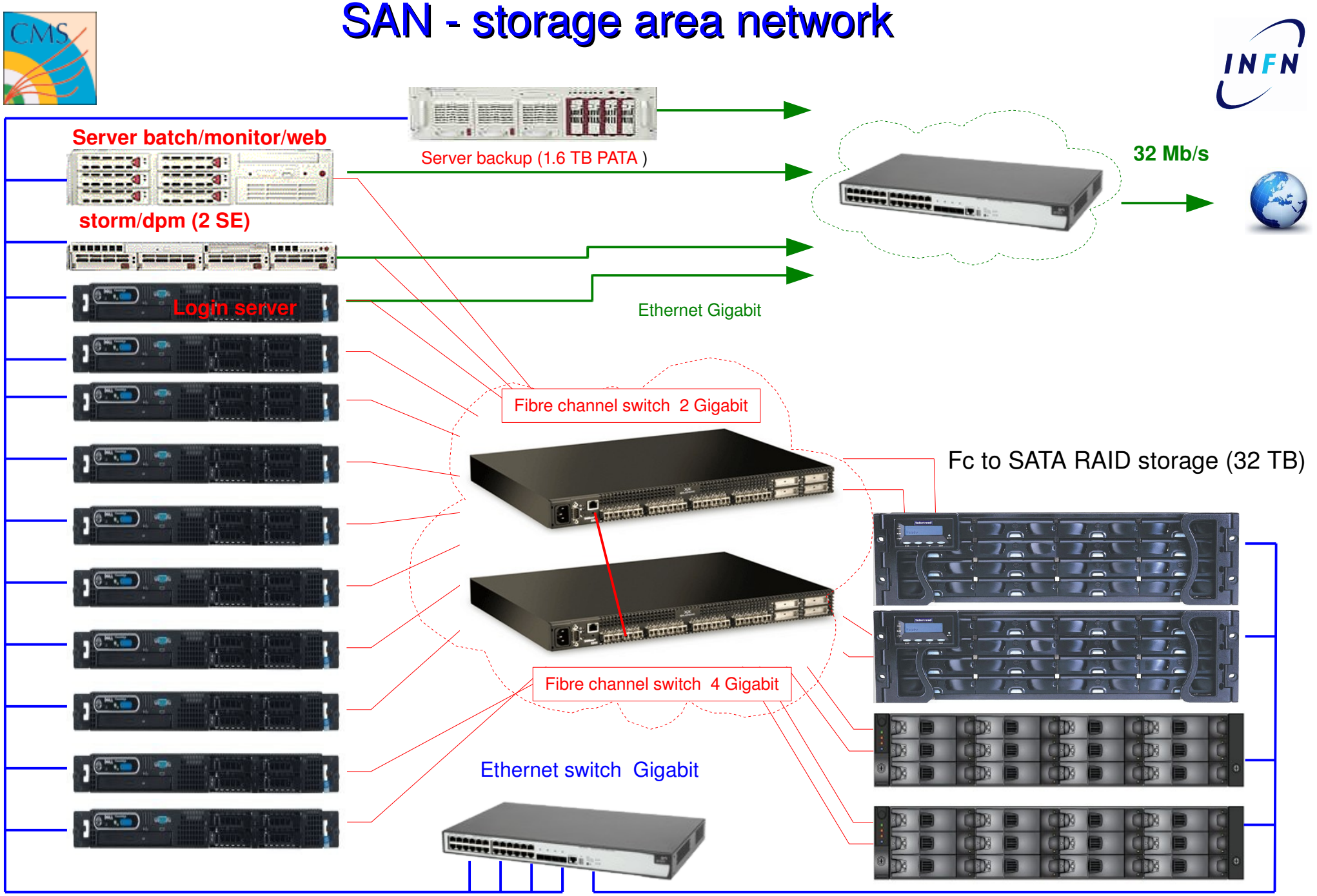
- Gruppo I (CMS), gruppo II (AMS,CUORE,HARP/MICE), gruppo IV
- nuova sala macchine dal giugno 2009

CMS a Milano Bicocca

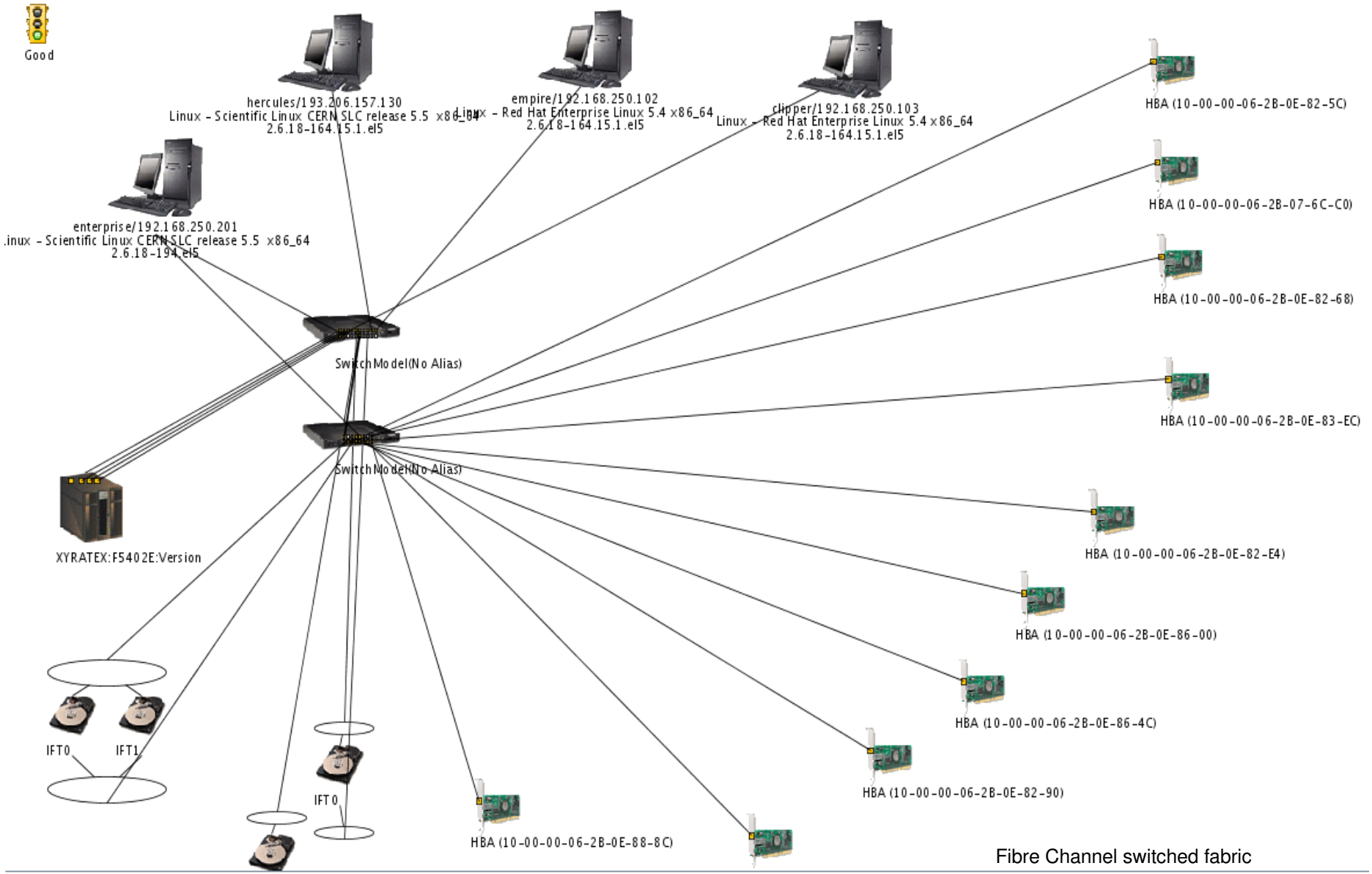
- gruppo ECAL
- gruppo Pixel Milano (Forward Tracker)
- Tipologia delle analisi: Higgs/electro-weak/B-Physics → (N.Skim > 2)

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

SAN - storage area network



Topologia del cluster



Fibre Channel switched fabric

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)

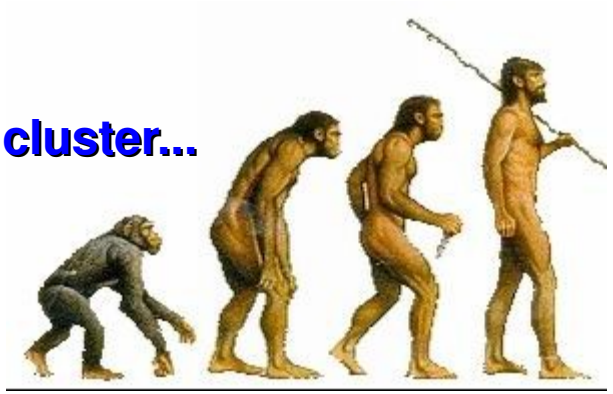
•3 sistemi di storage ridondati per un totale di 56 slot dischi:

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

Throughpout (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

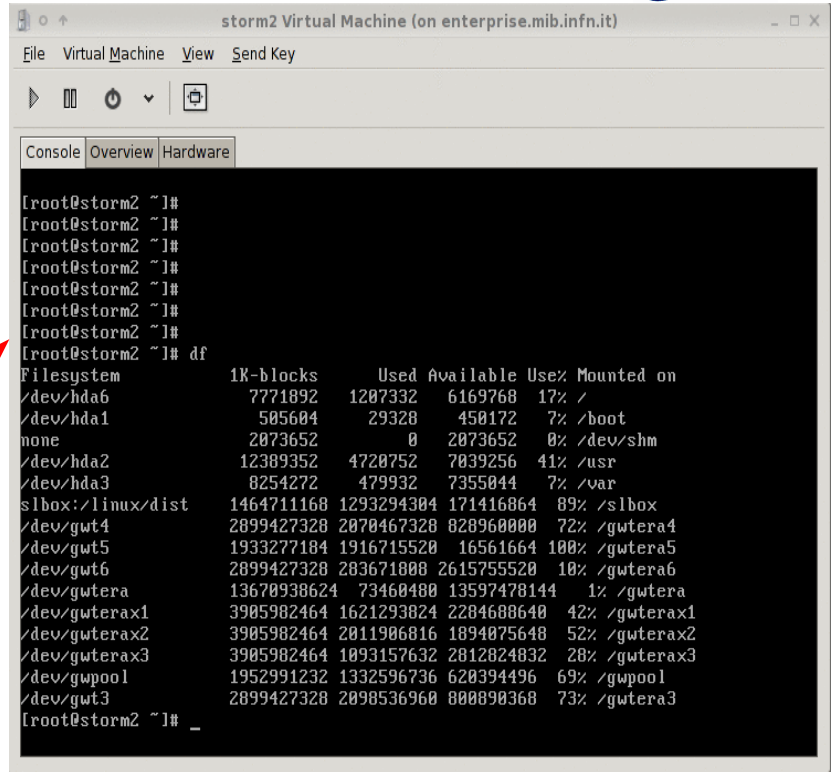
Harp+CMS 2003

CMSCLUSTER (2010)

Nuova sala macchine di Milano-Bicocca (giugno 2009)

- ~ 40 m²
- 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...

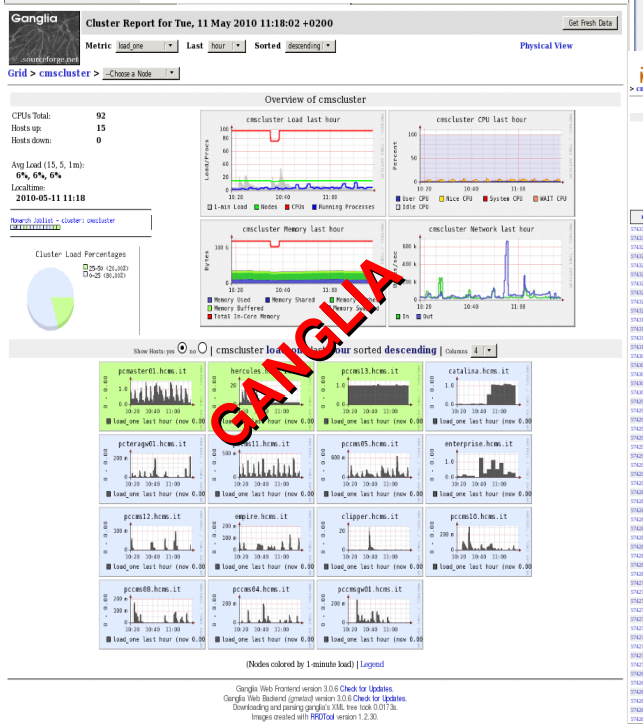
- User Interface: glite_3.2 per slc5 e glite_3.1 per slc4
- CRAB client
- Testati 2 Storage Element : DPM e STORM

⇒ **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!

MONITORING TOOLS



WEBSITE

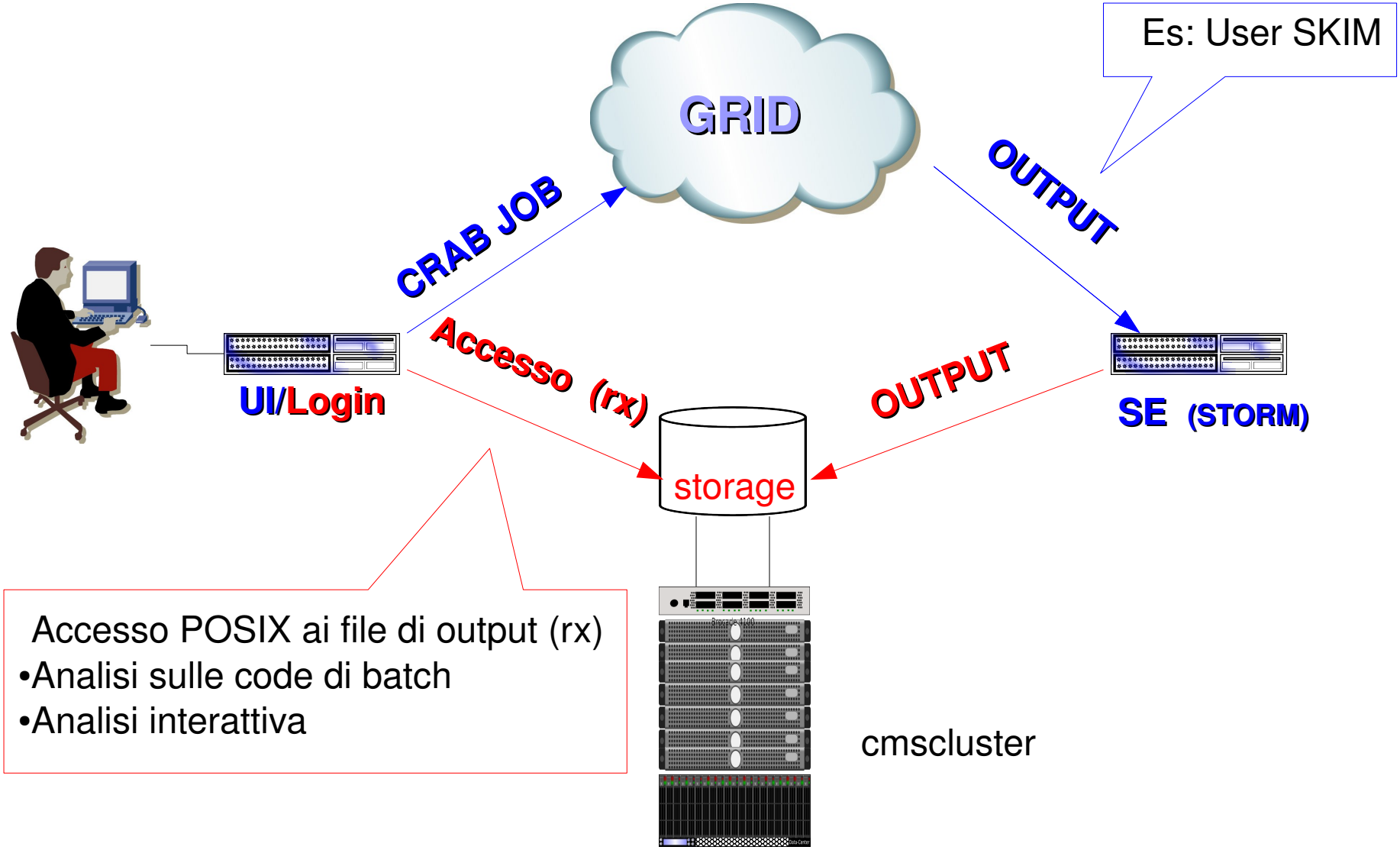
MUNIN

GANGLIA

JOBMONARCH

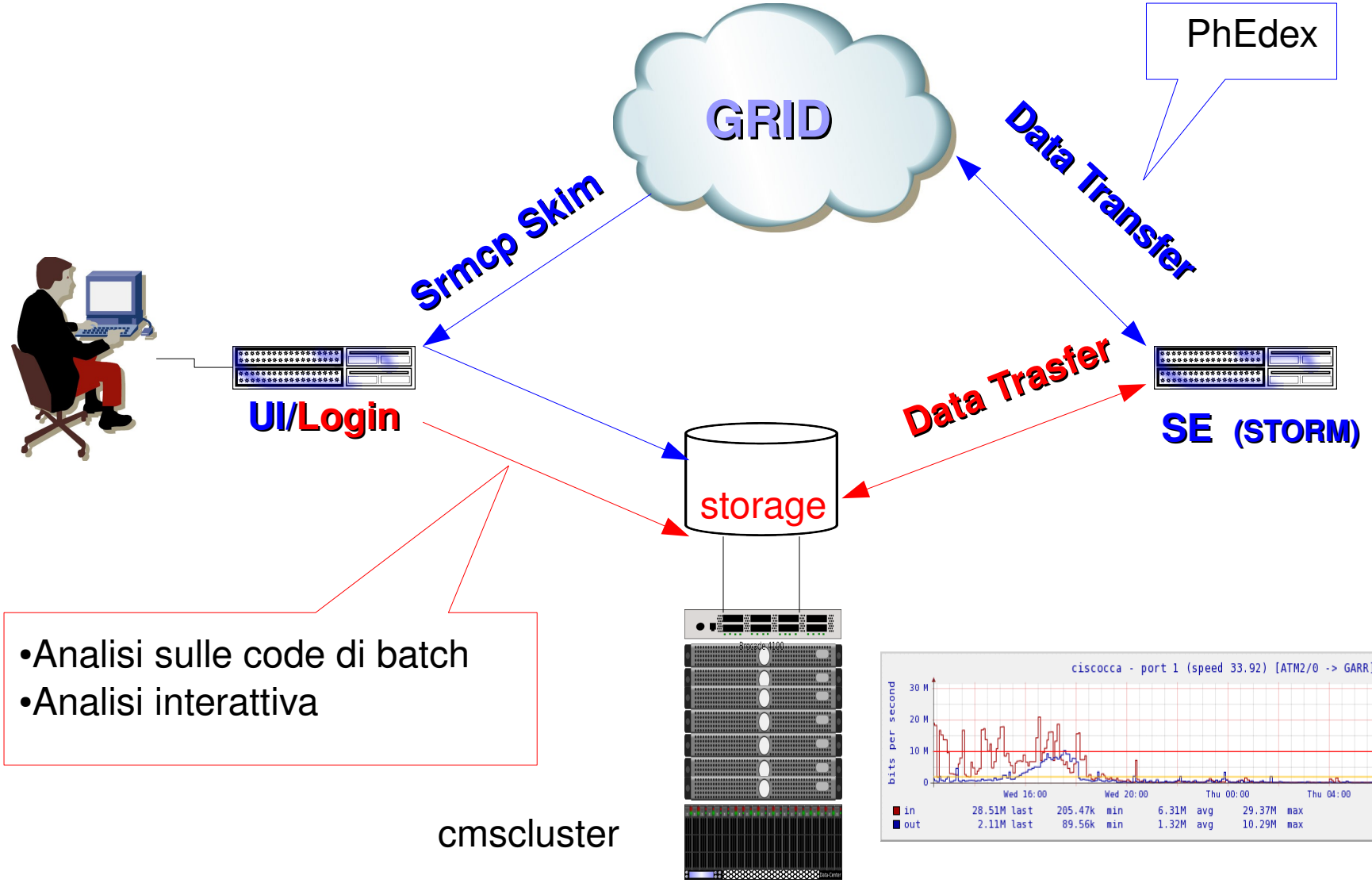
NAGIOS

USE CASE



Accesso POSIX ai file di output (rx)
- Analisi sulle code di batch
- Analisi interattiva

USE CASE



- Analisi sulle code di batch
- Analisi interattiva

cmscluster

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)

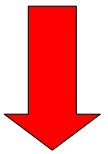
Futuri sviluppi:

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

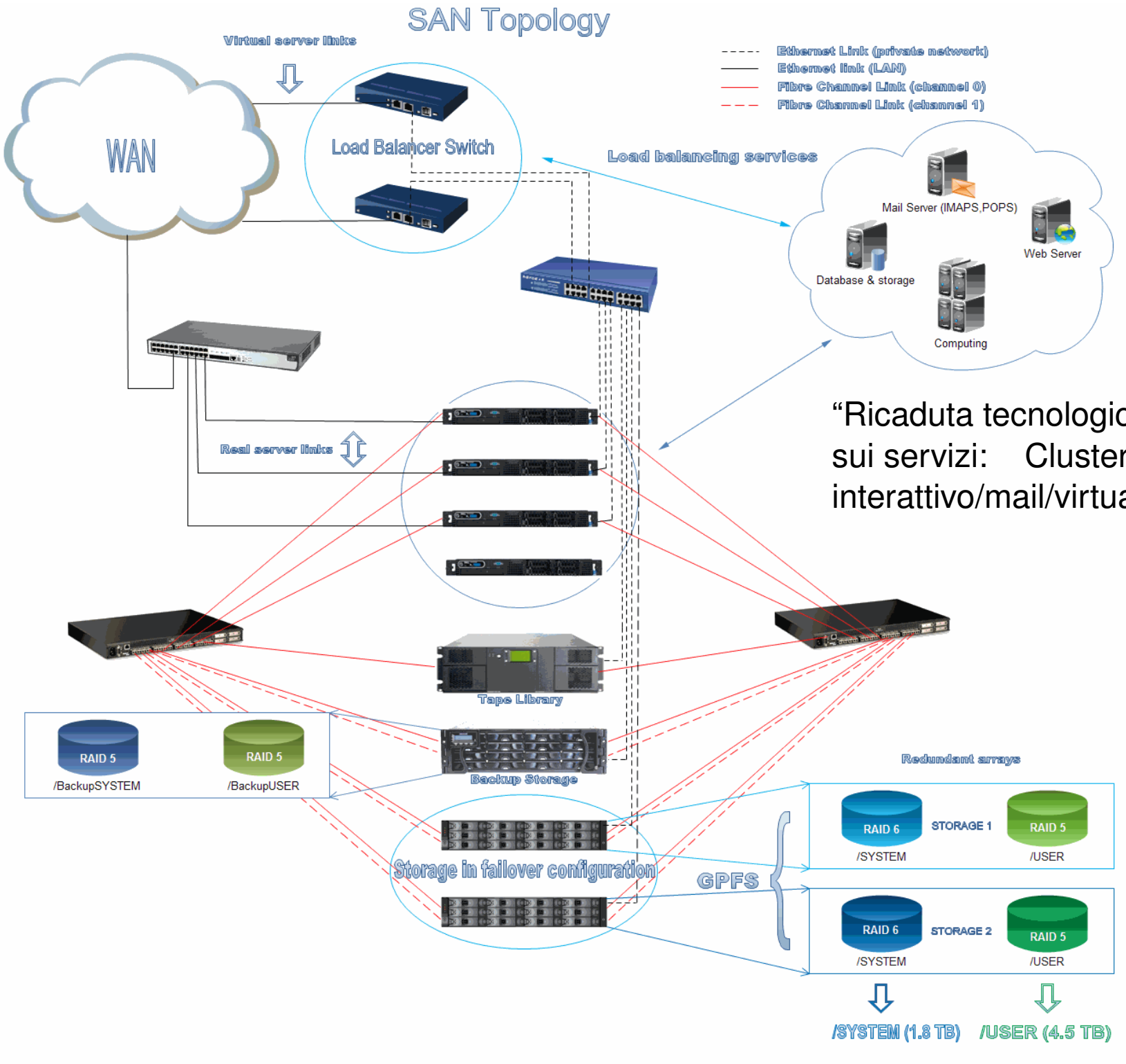
- Produzioni MC locali (CPU+storage)
- User Skim o subSkim (storage, rete!)
- Data replica (storage, rete!)
- Condivisione di dati attraverso PhEdex (...)

… e dalla tipologia degli utenti

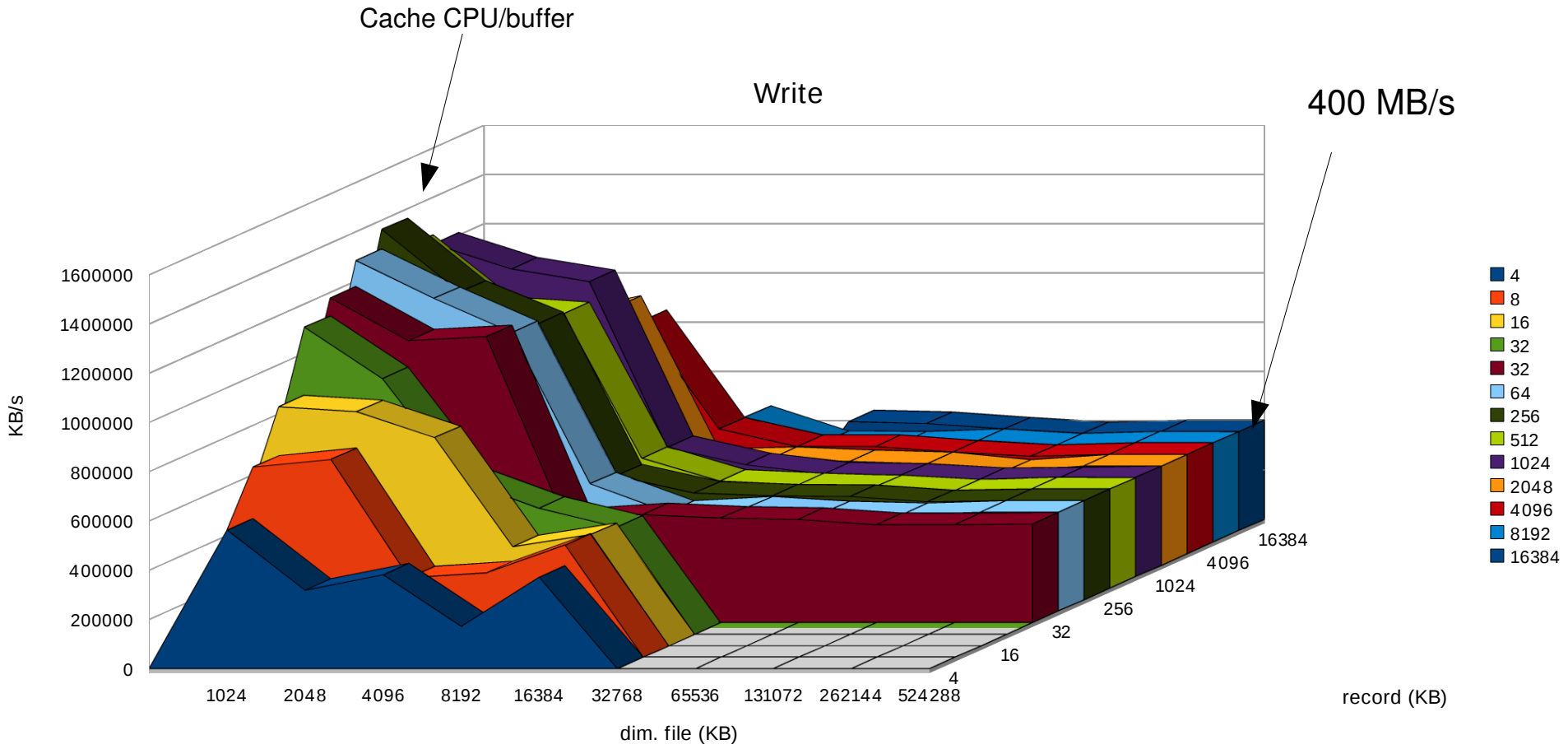
- Utenti locali
- Utenti locali+utenti esterni

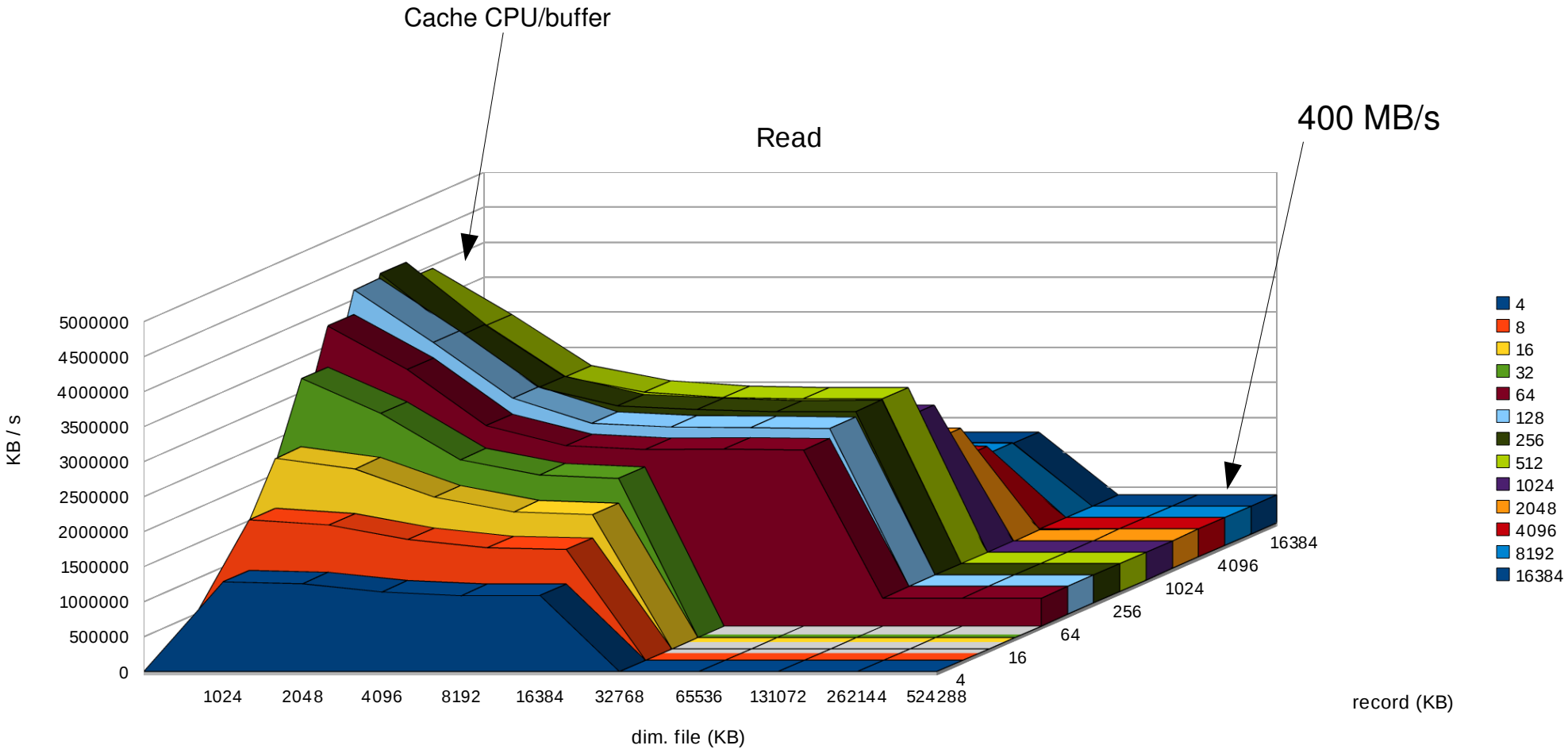


- Connessione a 100 Mb/s (Giugno?)
- Acquisto di 1 JBOD (16 slot) ⇒ + 32 TB
- Acquisto di un nuovo file server x backup on line
- Nodi?
- **PhEdex: da venerdi' T3_IT_MIB**



"Ricaduta tecnologica"
sui servizi: Cluster per
interattivo/mail/virtualizzazione


Cache CPU/buffer
Write
400 MB/s
KB/s
1600000
1400000
1200000
1000000
800000
600000
400000
200000
0
1024
2048
4096
8192
16384
32768
65536
131072
262144
524288
dim. file (KB)
4
16
32
256
1024
4096
16384
record (KB)
4
8
16
32
32
64
256
512
1024
2048
4096
8192
16384


Cache CPU/buffer
Read
400 MB/s
KB / s
5000000
4500000
4000000
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
1024
2048
4096
8192
16384
32768
65536
131072
262144
524288
dim. file (KB)
4
16
64
256
1024
4096
16384
record (KB)
4
8
16
32
64
128
256
512
1024
2048
4096
8192
16384