L'esperienza degli esperimenti non-LHC di CSN2 al Tier-1del CNAF-INFN Bologna

A.Karen Calabrese M.
INFN-CNAF

SOMMARIO

- Il CNAF e il Tier -1
- Le risorse attuali
- Le soluzioni di accesso e gestione delle risorse
- Utilizzo delle risorse da parte degli esperimenti di CSN2

MAPSES, Lecce 23-25 Novembre 2011

Il Tier-1 del CNAF-INFN Bologna

- Il CNAF situato a Bologna è il centro INFN per la ricerca e lo sviluppo nel campo delle tecnologie informatiche applicate agli esperimenti di fisica nucleare e delle alte energie.
- Il Tier-1 del CNAF è stato inaugurato nel 2005 con ampliamento finale dei nuovi impianti nel 2009
- Centro italiano di riferimento per il calcolo ad alte prestazioni dei fisici coinvolti in LHC
- Unica sede italiana di primo livello per gli esperimenti LHC.
- Ad oggi capace di fornire supporto a diverse comunità scientifiche con la necessità di gestire, elaborare e archiviare una grande mole di dati

I Tiers

Il grande sistema di calcolo distribuito per LHC si sviluppa su tre livelli detti "TIERS".

Il Tier-0 è il livello base

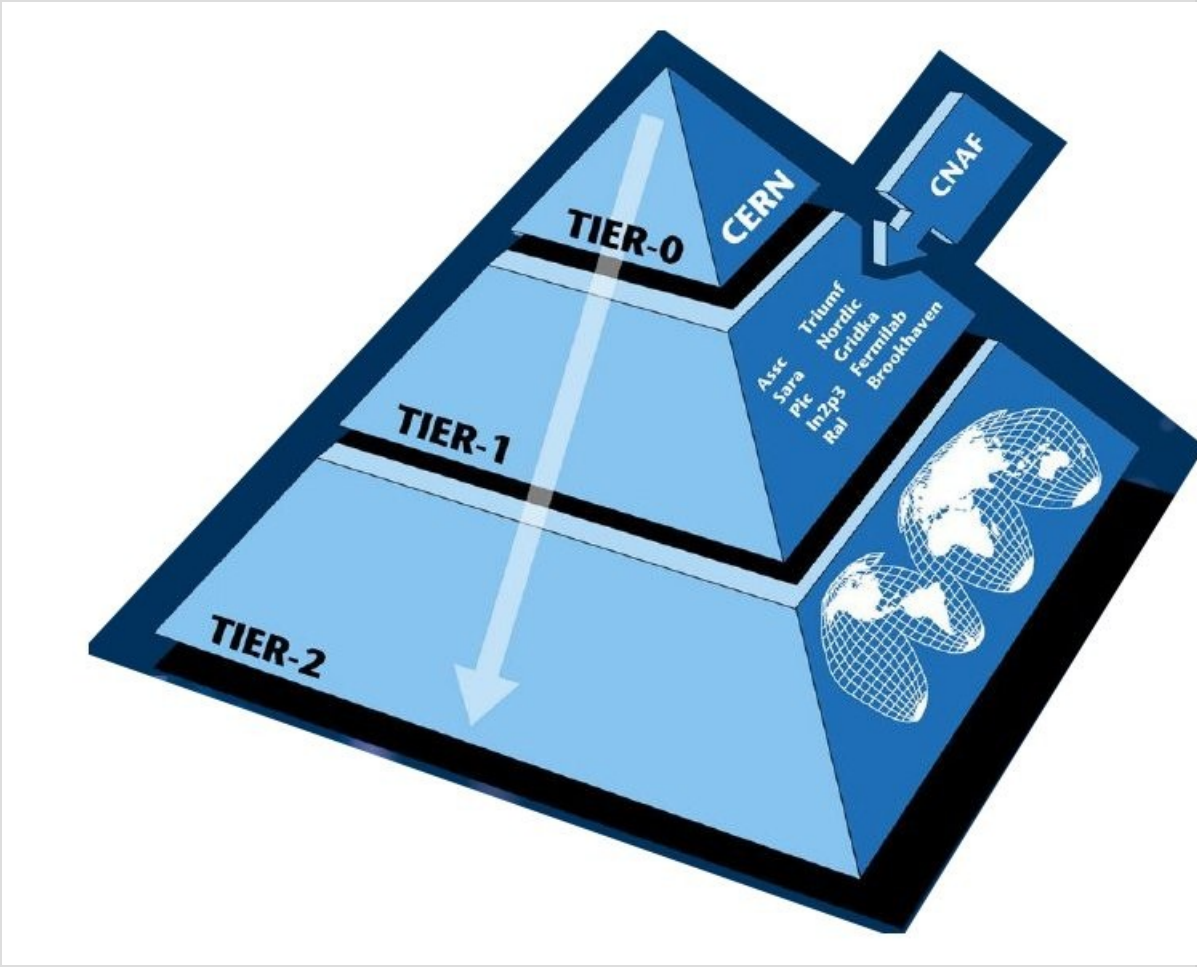
(centro di calcolo del CERN):

- tutti i dati passano da

questo nodo

- fornisce circa il 20%

della capacità di calcolo totale



I Tiers

Il grande sistema di calcolo distribuito per LHC si sviluppa su tre livelli detti "TIERS".

I Tier-1 sono 11 in tutto il mondo.

Si caratterizzano per:

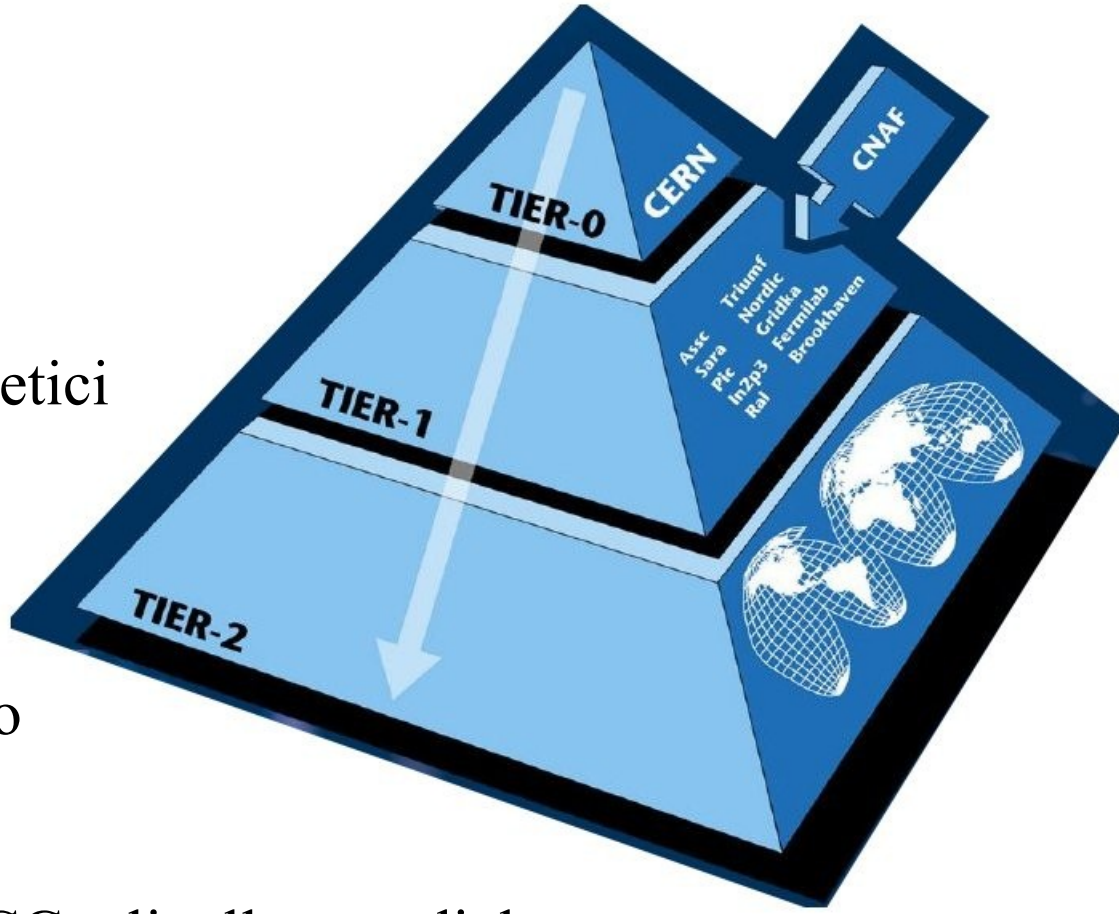
- elevata capacità di calcolo e

di memoria su dischi e nastri magnetici

- il livello del servizio
- la connettività a larga banda
- la responsabilità di primo piano

per garantire il funzionamento

di tutta l'infrastruttura Grid di WLCG a livello mondiale



I Tiers

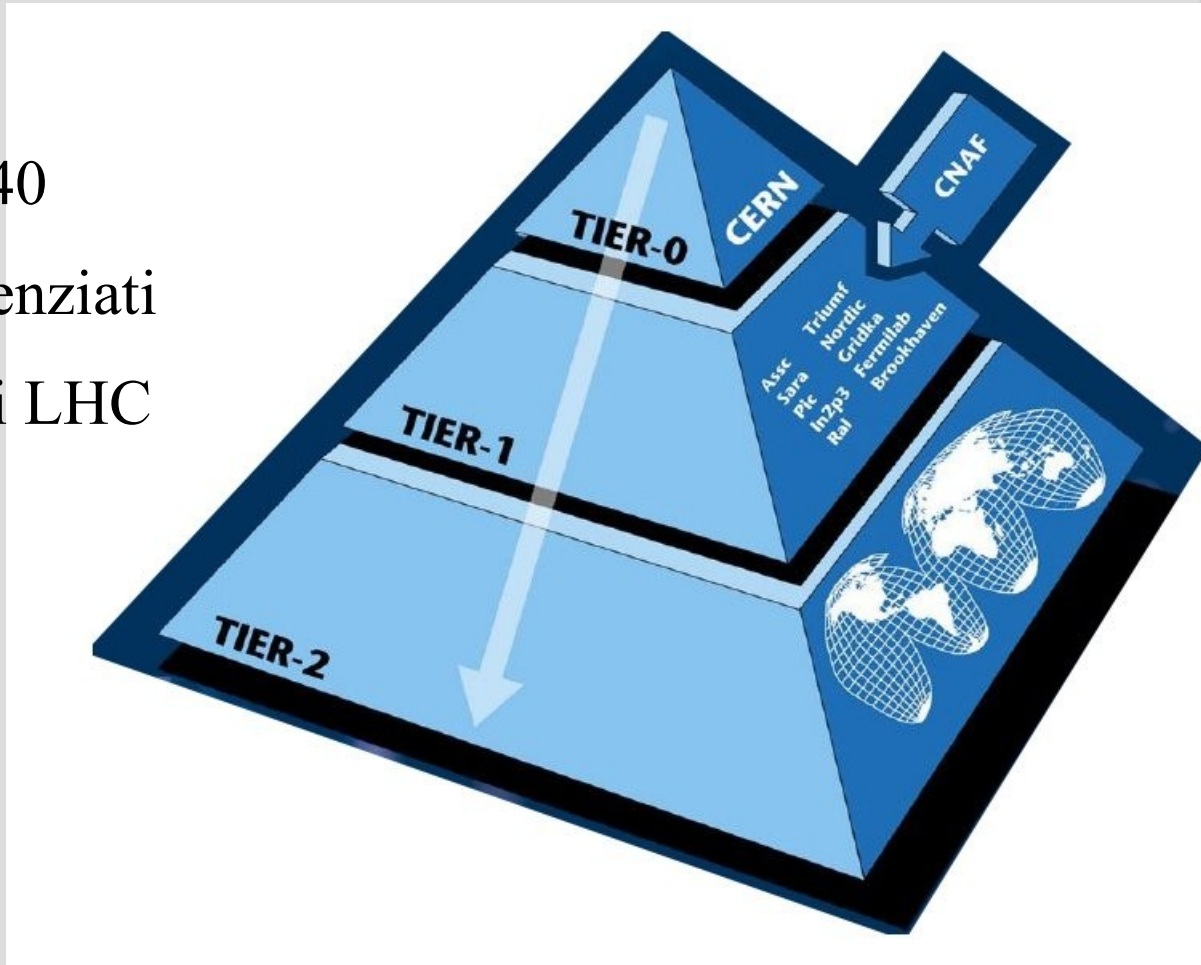
Il grande sistema di calcolo distribuito per LHC si sviluppa su tre livelli detti “TIERS”.

I Tier-2 si localizzano in altri 140

siti permettono ai gruppi di scienziati

di accedere e analizzare i dati di LHC

direttamente nei propri istituti



Il Tier-1 del CNAF in numeri

Infrastruttura

- Sala calcolo 1200 m^2
- 120 armadi per apparecchiature informatiche di calcolo
- Potenza massima dissipabile 1.4 MW
  - consumo massimo previsto 2.6 MW (come una cittadina di 25000 persone)
  - 2+1 trasformatori elettrici da 2.5 MVA ciascuno
  - impianto di continuità elettrica con 2 gruppi rotanti dotati di generatore diesel coassiale da 1.7 MW ciascuno
  - 5+2 chiller per il raffreddamento (capacità di raffreddamento circa 2 MW)

Il Tier-1 del CNAF in numeri

Storage

- 9 PB capacità disco (SAN)
- 110 disk servers (50% 10 Gbit)
- 10 Tape TSM-HSM clients
- 1 Fibre channel centrali
- 20 Fibre chanel periferici
- 10 PB On-Line capacità su nastro

Il Tier-1 del CNAF in numeri

Farming

- 100.000 HEP-SPEC06 ->125.000 HEP-SPEC06 nel 2012
- 9216 slots attuali di calcolo (“CPU” core)
- 20 diverse collaborazioni scientifiche utilizzano le risorse del Tier-1 accessibili in modo distribuito attraverso l'infrastruttura Grid.
- Più di 50.000 job di calcolo eseguiti in media ogni giorno
- macchine virtuali vengono messe a disposizione in maniera trasparente e dinamica attraverso la creazione di nodi di calcolo virtuali utilizzando il servizio WNoDeS(Worker Nodes on Demand Service)

Il Tier-1 del CNAF in numeri

Rete

- 4 switch router di core
- 200x10Gb/s porte
- 468x1Gb/s porte
- 3 link geografici
  - 2x10Gb/s (doppio link condiviso) per le direttrici di traffico T0 (CERN)-T1, T1-T1
  - 1x10Gb/s T1-T2, General Purpose
  - 1x10 Gb/s T1-T1 (Karlsruhe, IN2P3, SARA)

Il Tier-1 del CNAF in numeri

Rete

- Switch di aggregazione
  - 78 switch (1 unità rack)
  - 21 blade switches
  - 4300 x 1Gb/s porte
  - 100 x 100Mb/s porte
  - 36 x 10 Gb/s
- Sistemi di monitoring, allarmistica e logging
  - MRTG: raccoglie tutte le statistiche di banda della rete

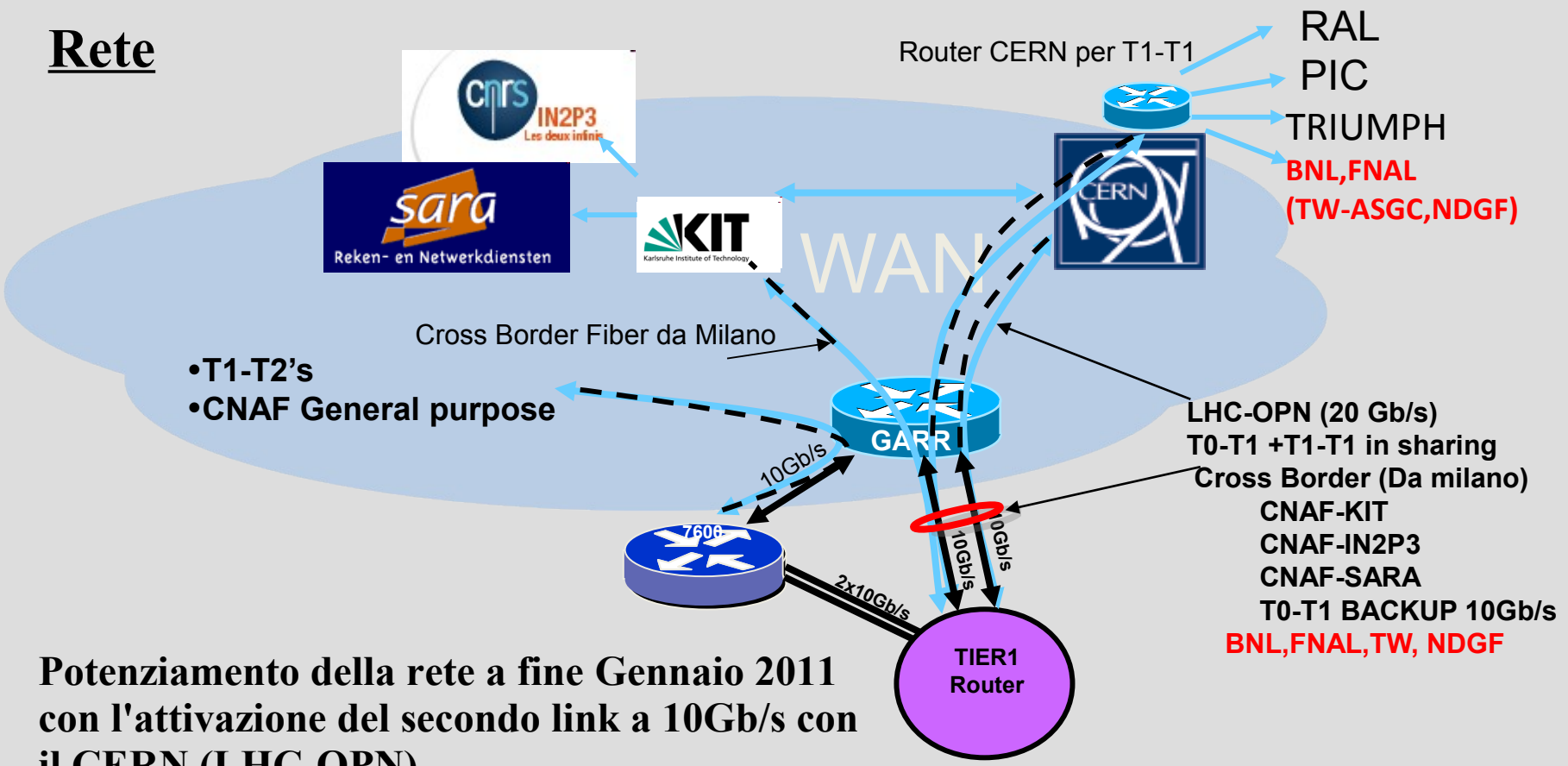
Il Tier-1 del CNAF in numeri

Rete

- Sistemi di monitoring, allarmistica e logging
  - NAGIOS: controlla lo stato di ogni dispositivo di rete (connettività, trunck status, CPU load)
  - Lemon: sistema di monitoring usato dai server dello storage e del farming
  - Syslog-NG: tutti i messaggi di log (genrati dai dispositivi di rete e dai principali server del Tier-1) vengono raccolti da diversi server Syslog-NG

Il Tier-1 del CNAF in numeri

Rete



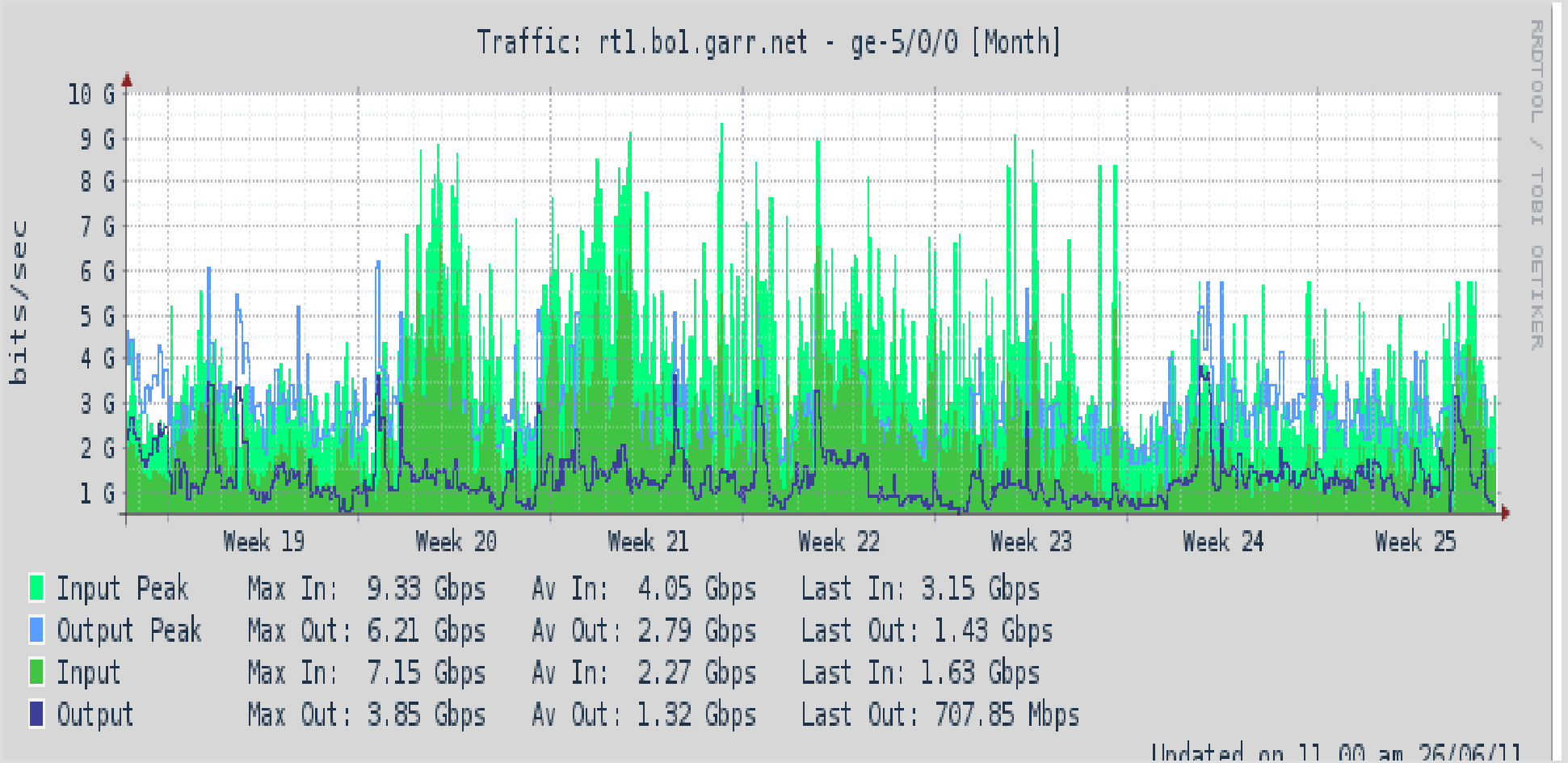
- **Potenziamento della rete a fine Gennaio 2011 con l'attivazione del secondo link a 10Gb/s con il CERN (LHC-OPN)**
- **Link da 10 Gb/s per traffico T1-T2s + general purpose (include traffico per exp di GR2)**

Rete

- La connettività geografica è garantita dalla Rete della Ricerca Italiana (GARR) e dalla rete della ricerca europea (GEANT)
- La banda disponibile per il traffico di rete sul link general purpose è aumentata a seguito del ri-direzionamento del traffico verso i Tier-1 americani sui link verso il CERN (FNAL,BNL)
- Il link general purpose è in grado di far fronte alle esigenze degli esp. di CSN2
- Da qualche settimana introduzione di un nuovo link (per ora in shared con il trunk a 20 Gb/s di LHCOPN) dedicato ad LHCONE (per raggiungere i Tier-2 ed eventualmente anche i Tier-3). A livello italiano questa rete verrà utilizzata per raccogliere tutti i Tier-2.

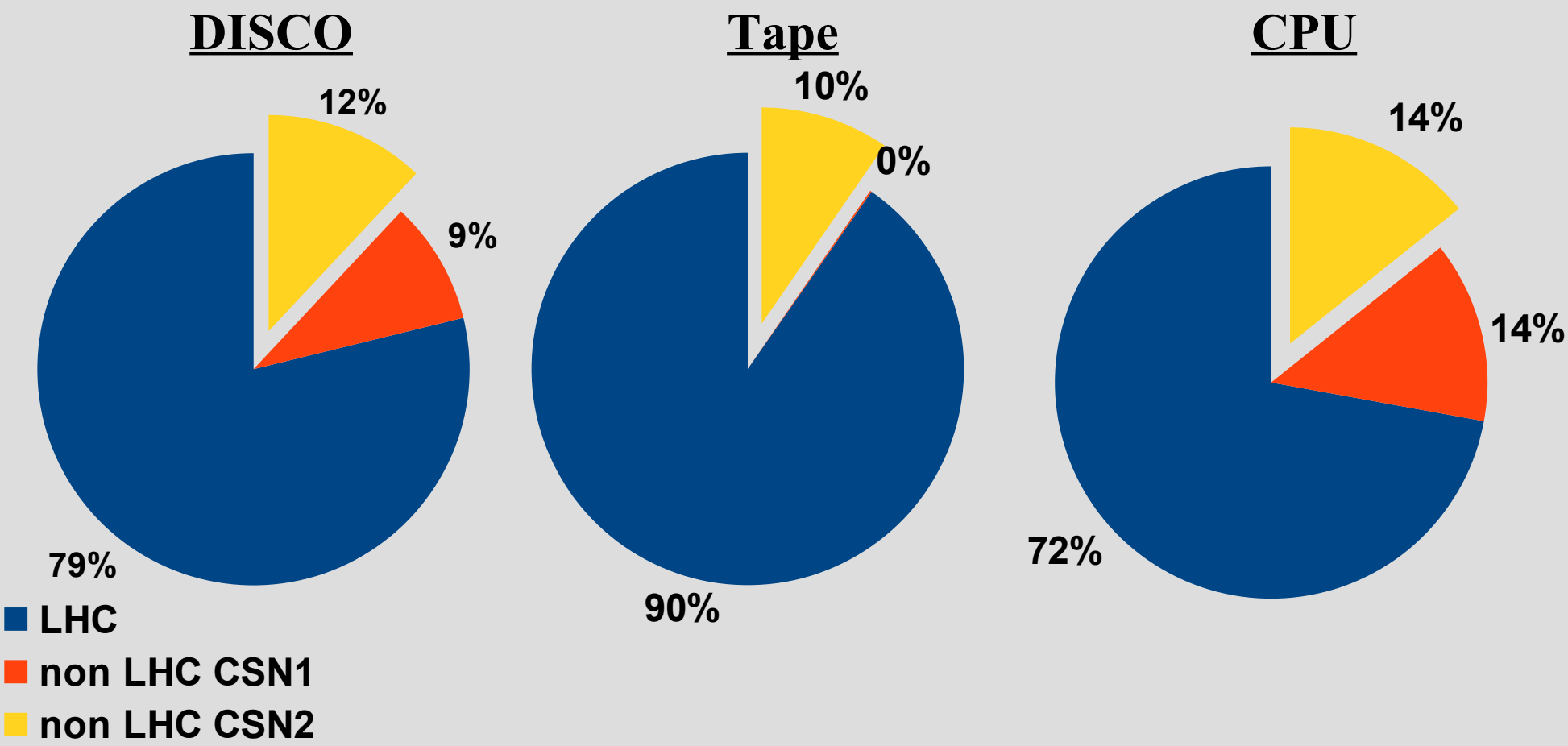
Il Tier-1 del CNAF in numeri

Rete



General purpose

La distribuzione attuale delle risorse al Tier-1



Le risorse degli esperimenti non-LHC di CSN2 al Tier1 rappresentano oggi una parte non trascurabile del totale delle risorse attuali del Tier-1

Gli esperimenti non-LHC di CSN2 al Tier-1

Esperimento	Attività al Tier1
ARGO	• trasferimento dati da YBJ a CNAF • archiviazione su tape dati raw • calibrazione e ricostruzione • archiviazione su nastro root-ple prodotte • distribuzione root-ple ai siti nazionali • analisi ufficiali e dedicate • Simulazioni
MAGIC	• Simulazioni

Gli esperimenti non-LHC di CSN2 al Tier-1

Esperimento	Attività al Tier-1
VIRGO	• trasferimento dati da Cascina a CNAF • archiviazione su tape dati raw • analisi utenti
PAMELA	• trasferimento dei dati da MEPHI a CNAF • simulazioni • analisi
FERMI	• simulazioni

Gli esperimenti non-LHC di CSN2 al Tier-1

Esperimento	Attività al Tier-1
ICARUS	• trasferimento dati da LNGS a CNAF (backup su tape)
BOREXINO 	• trasferimento dati da LNGS a CNAF (backup su disco) • Analisi dati e simulazioni
GERDA 	• trasferimento dati da LNGS a CNAF (backup su disco • simulazioni
XENON 	• Analisi e simulazioni

Gli esperimenti non-LHC di CSN2 al Tier-1

Nel modello a multi-Tier gli exp. di CSN2 utilizzano il Tier-1 del CNAF come:

- **Tier0 :** Archiviazione primaria dei dati raw e reco con copia singola o multipla (es. ARGO, PAMELA, VIRGO, ICARUS). Processamento primario dei dati (es. ARGO)
- **Tier1:** Campagne di simulazioni (es. MAGIC, FERMI, AUGER). Processamento sistematico dei dati (es. ARGO)
- **Tier2 :** Supporto analisi dati (es. PAMELA, ARGO, AUGER)
- **Tier3 :**Allocazione dinamica delle risorse per l'analisi interattiva degli utenti

Accesso locale alle risorse del Tier1

- Sistemi di accesso alle risorse del Tier1: LDAP(autorizzazione) + Kerberos(autenticazione)
- Per un account locale alle risorse: http://www.cnaf.infn.it/it/staff/AUP
- Con un account locale:
  - verrà assegnata una user interface di un esperimento X a cui accedere localmente alle risorse del Tier-1
  - si potrà accedere all'area storage di esperimento (/storage/gpfs_X mounted via gpfs sulla ui dedicata), all'area software di esperimento(/opt/exp-software/X mounted via nfs)
  - si potrà sottomettere job sulla farm del Tier-1

Accesso Grid alle risorse del Tier-1

- Occorre appartenere ad una V.O. ufficiale a scope GLOBAL

  (tutte le info relative per la creazione di una nuova V.O. su http://www.italiangrid.it/grid_operations/users/VO_creation

  e info V.O. esistenti sul CIC portal http://operations-portal.egi.eu/vooperations-portal.egi.eu/vo)
- Richiedere la registrazione ad una V.O.
- con certificato infn e con un proxy “firmare” il permesso di accesso alle risorse tramite il middleware gLite
- sottomettere jobs decidendo le caratteristiche hw e software che il nodo di calcolo di destinazione deve soddisfare specificandoli attraverso il Job Description Language

Soluzioni standard di utilizzo e accesso delle risorse

- **IBM GPFS** file system parallelo per lo storage dei dati
- **LSF** per lo scheduling dei job **(algoritmo:meccanismo di fairshare dinamico)**
- **StorM** implementazione srm sviluppata al CNAF che permette l'accesso diretto alle risorse dello storage sia attraverso protocollo file che protocolli standard Grid.
- **GEMSS** una nuova soluzione di gestione dell'accesso alle librerie di nastri, che ha rimpiazzato Castor (software sviluppato al CERN che ha evidenziato negli anni forti limiti di performance). In produzione dal 2010 per tutti gli esperimenti.
- **WnoDeS** servizio sviluppato al CNAF per un utilizzo ottimizzato e dedicato delle risorse attraverso la virtualizzazione

IBM GPFS (General parallel File System)

- File system parallelo dedicato ai sistemi di HPC dalle prestazioni elevate con dischi condivisi:
  - insieme di dischi che contengono i dati del fs
  - insieme di nodi che usano e gestiscono il fs
  - in insieme di interconnessioni di rete che consentono la comunicazione tra i nodi e lo storage

- Consente a tutti i nodi di calcolo che montano il fs di avere una visione coerente di tutto lo storage ed un accesso concorrente ad esso

IBM GPFS (General parallel File System)

- Caratteristiche:
  - Standard posix, supporto alle quote e alle Access Control List (ACL)
  - ridimensionamento dinamico dei volumi
  - Prestazioni elevate: parallelizzazioni degli accessi e bilanciamento del carico
  - Alta affidabilità per la conservazione dei dati tramite repliche di dati e metadati
  - Esportabilità dei volumi via nfs
- Al Tier1 tutti i disk-server sono connessi allo storage via SAN e i worker nodes raggiungono via rete i disk-server

LSF Batch Manager (Load sharing facility)

- Batch manager fornito da Platform
- Caratteristiche principali:
  - ogni nodo viene identificato dalla sua configurazione hardware e dal sistema operativo, dagli slot ovvero il numero di job che possono essere eseguiti su di esso;
  - per ogni nodo da gestire è possibile definire se i suoi slot sono dedicati in via esclusiva ad un determinato gruppo di utenti, oppure se fanno parte dello shared cluster utilizzabile da chiunque, oppure una combinazione o nessuna delle due possibilità

LSF Batch Manager (Load sharing facility)

- Caratteristiche:
  - l'accesso agli slot, di norma, prevede la sottomissione dei job ad una coda
  - per ogni job che viene eseguito si collezionano le statistiche di utilizzo delle risorse (cputime ed occupazione RAM); queste statistiche servono per l'accounting

    (http://tier1.cnaf.infn.it/monitor/accounting/)

LSF: V.O e meccanismo di fairshare

- Ogni V.O. ha una propria coda, i suoi utenti sono autorizzati a sottoettere i job solamente su questa coda,
- si possono associare priorità differenti in base a singolo gruppo, utente o progetto interno alla V.O., con una politica di prioritizzazione gerarchica.
- ogni V.O. può avere a disposizione in via esclusiva un pool di nodi, ma anche accedere alle risorse condivise fra tutte le V.O.
- Il meccanismo di fair-share dinamico adottato
  - permette di evitare la monopolizzazione delle risorse nell'ambiente condiviso
  - Modifica la priorità in base all'uso effettivo delle risorse

LSF: V.O e meccanismo di fairshare

- ad ogni V.O. viene assegnata una percentuale di utilizzo dello shared cluster, che il batch manager si impegnera a far rispettare nell'arco di una finestra di ore, giorni oppure mesi
- il fattore percentuale unito all'effettivo utilizzo contribuiscono a modificare in modo dinamico la priorità dei jobs in attesa nelle code
- le percentuali di fairshare possono essere a loro volta personalizzate all'interno di una V.O. (fairshare gerarchico)

Fairshare: algoritmo di scheduling

Meccanismo di **fairshare gerarchico.**

• In caso di non contesa delle risorse una VO può usare tutte le risorse disponibili.

• In caso di contesa l'accesso alle risorse di calcolo è regolato dalla priorità dell'utente che è dinamica.

• Questa priorità dinamica è funzione dello share dell'esperimento e delle risorse consumate dal singolo utente definita all'interno di una finestra temporale (sliding window).

• La formula utilizzata per calcolare la priorità dinamica è la seguente:

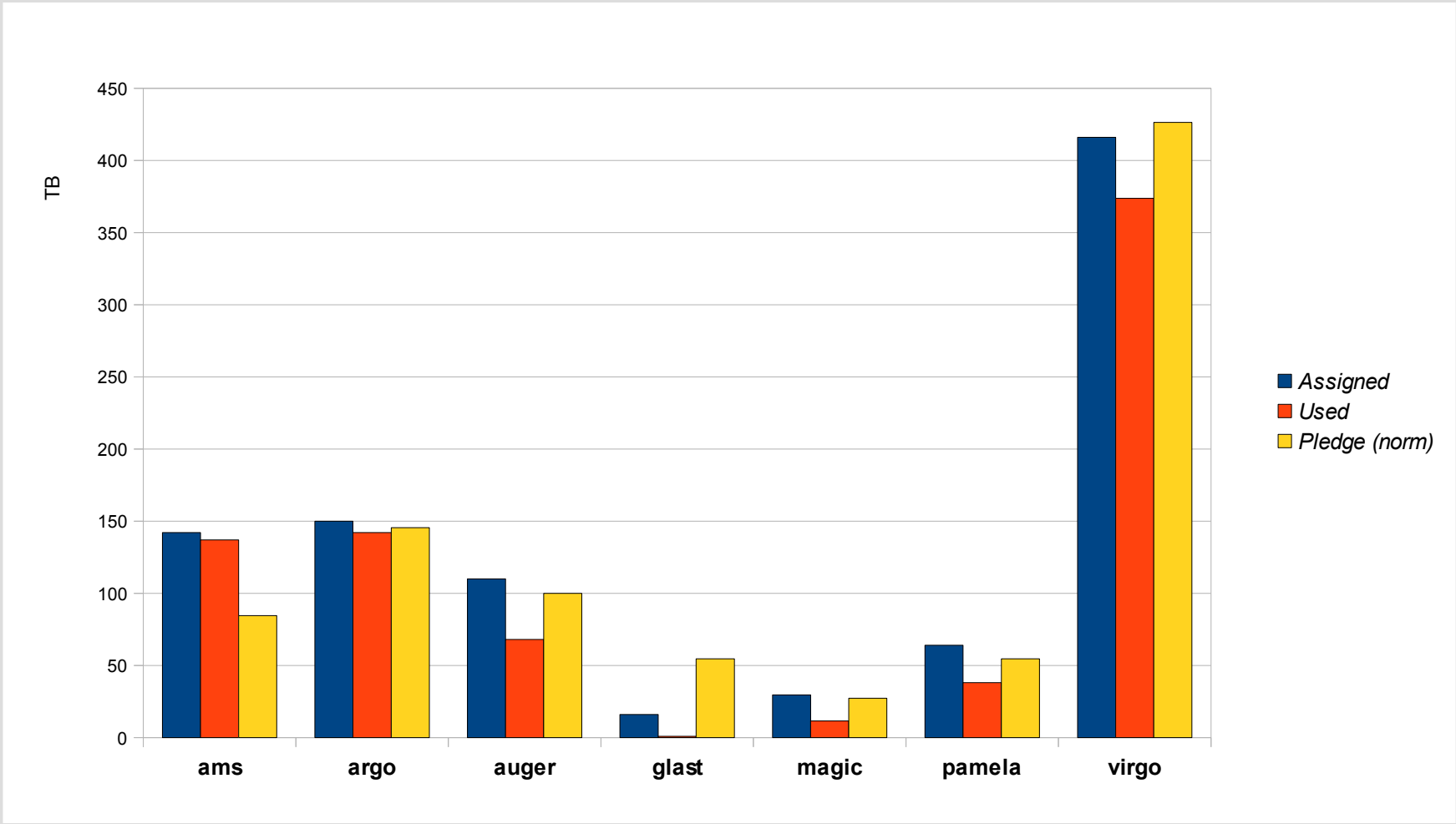
$$priority \propto \frac{share}{(K * WCT + J * CPT + H * RUNJOB + 1)}$$

Considerazioni sull'algoritmo di scheduling utilizzato

- Al Tier-1 l'attuale sliding window è di 48ore.
- I jobs di un Exp. Che ha esaurito la quota all'interno della sliding window,rimangono incoda se c’è contesa di risorse perché avrà una priorità più bassa.
- D’altro canto se non c'è contesa di risorse un Exp. Può usare tutte le risorse a disposizione.Questa situazione si sfrutta al meglio tenendo sempre un certo numero di job incoda.
- **In caso di contesa la VO può usare al massimo quanto garantito dallo share assegnato**

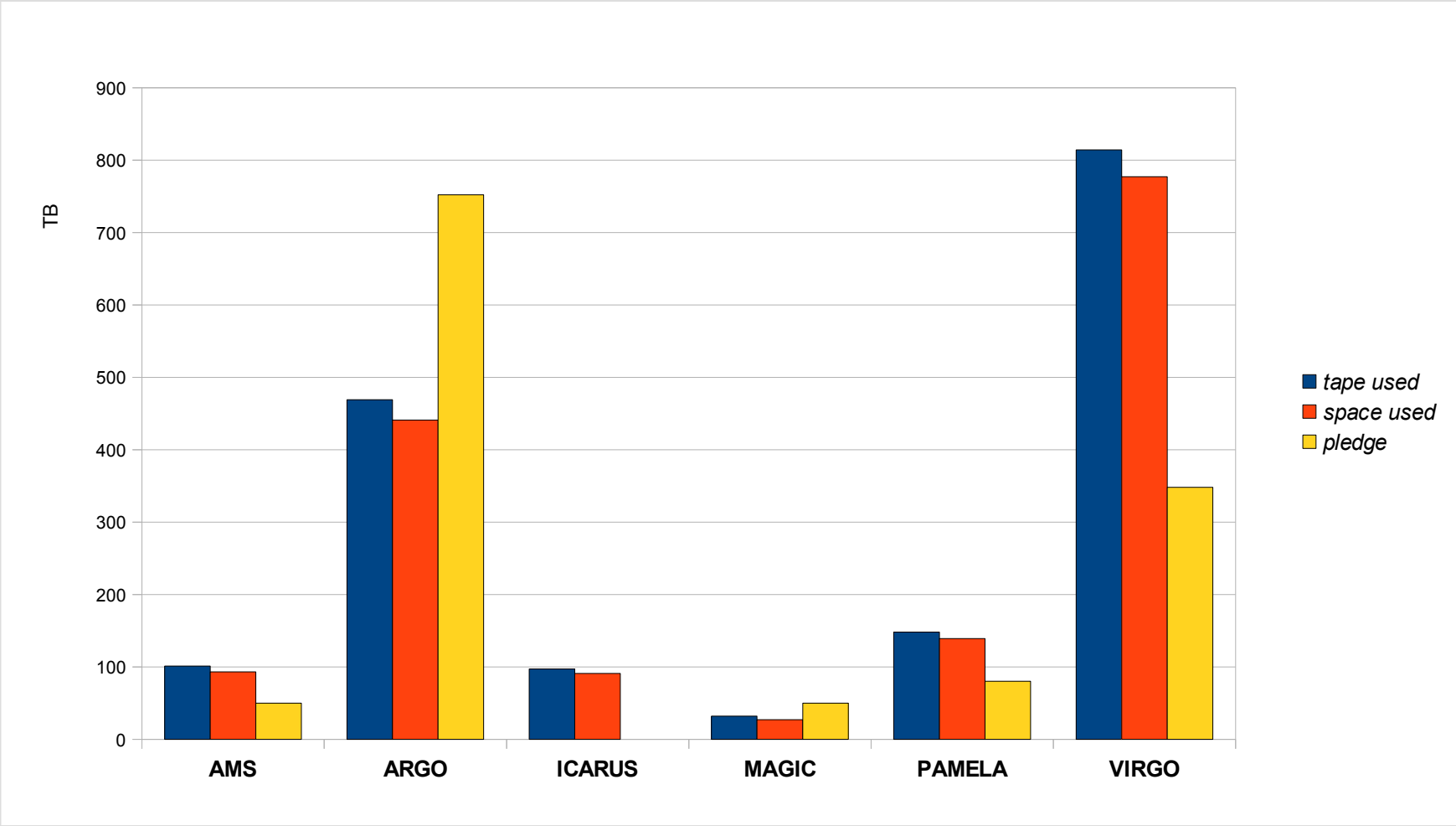
Utilizzo delle risorse: Storage

DISCO: dati aggiornati al 14 Novembre



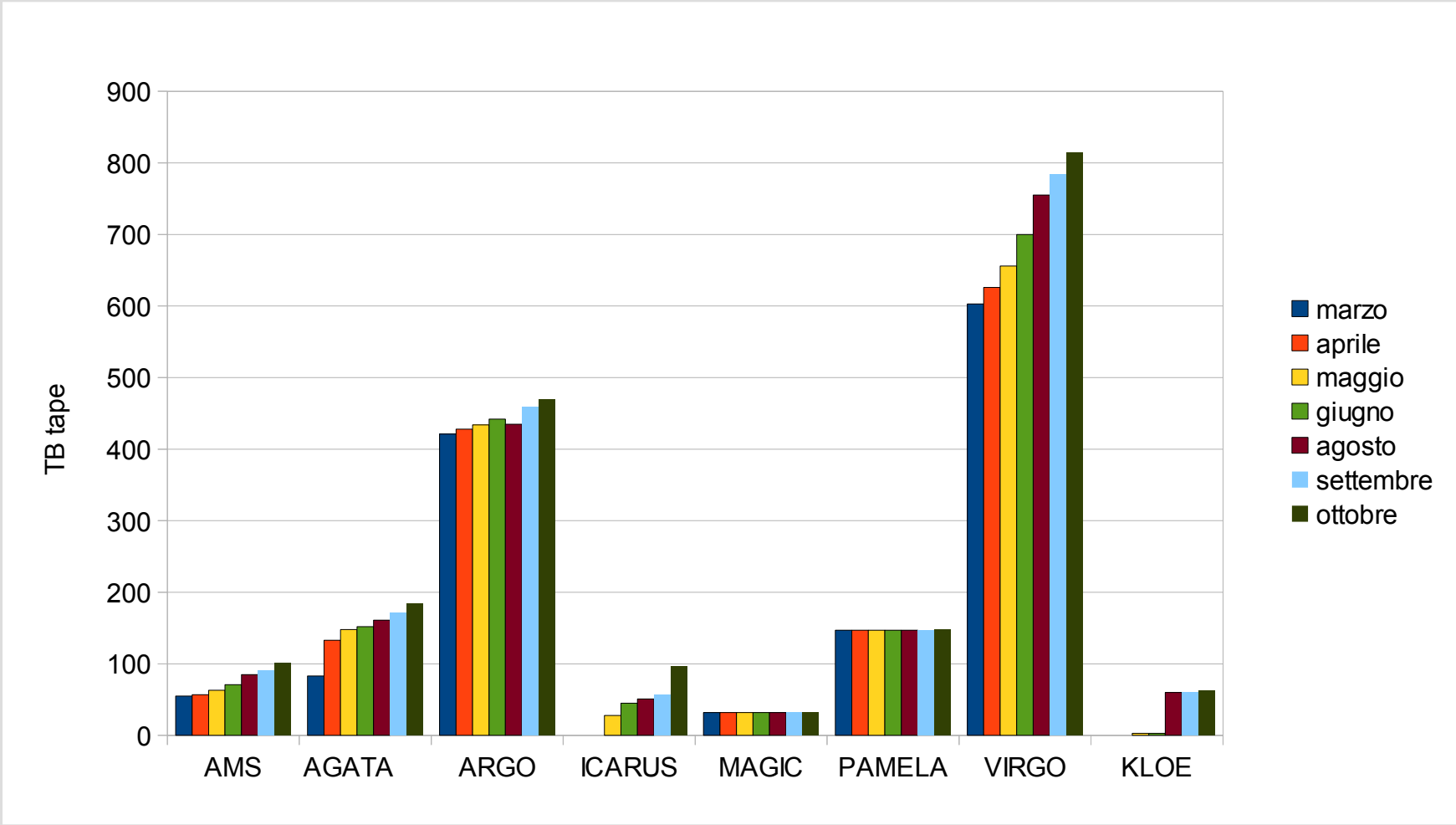
Distribuzione delle risorse: Storage

Tape: dati aggiornati al 14 Novembre



Utilizzo delle risorse:Storage

Tape: storico



Modalità di accesso allo Storage

Accesso locale

- es. ARGO, MAGIC, AUGER, VIRGO,AMS02, FERMI

Accesso tramite SRM/SToRM su disco e GEMSS su tape

- es. AMS02, VIRGO, PAMELA, ICARUS

Considerazioni:

- Il modello di storage utilizzato dagli esperimenti differisce per il tipo differente di attività svolte al Tier1
- Le modalità di accesso sono eterogenee e ha reso necessario un dimensionamento ottimizzato del sistema di storage in termini di numero di server oltre che di capacità disco (il sistema di I/O rappresenta l'effettivo collo di bottiglia).
- Soluzioni adottate sul campo. Fondamentale l'interazione con l'esperimento

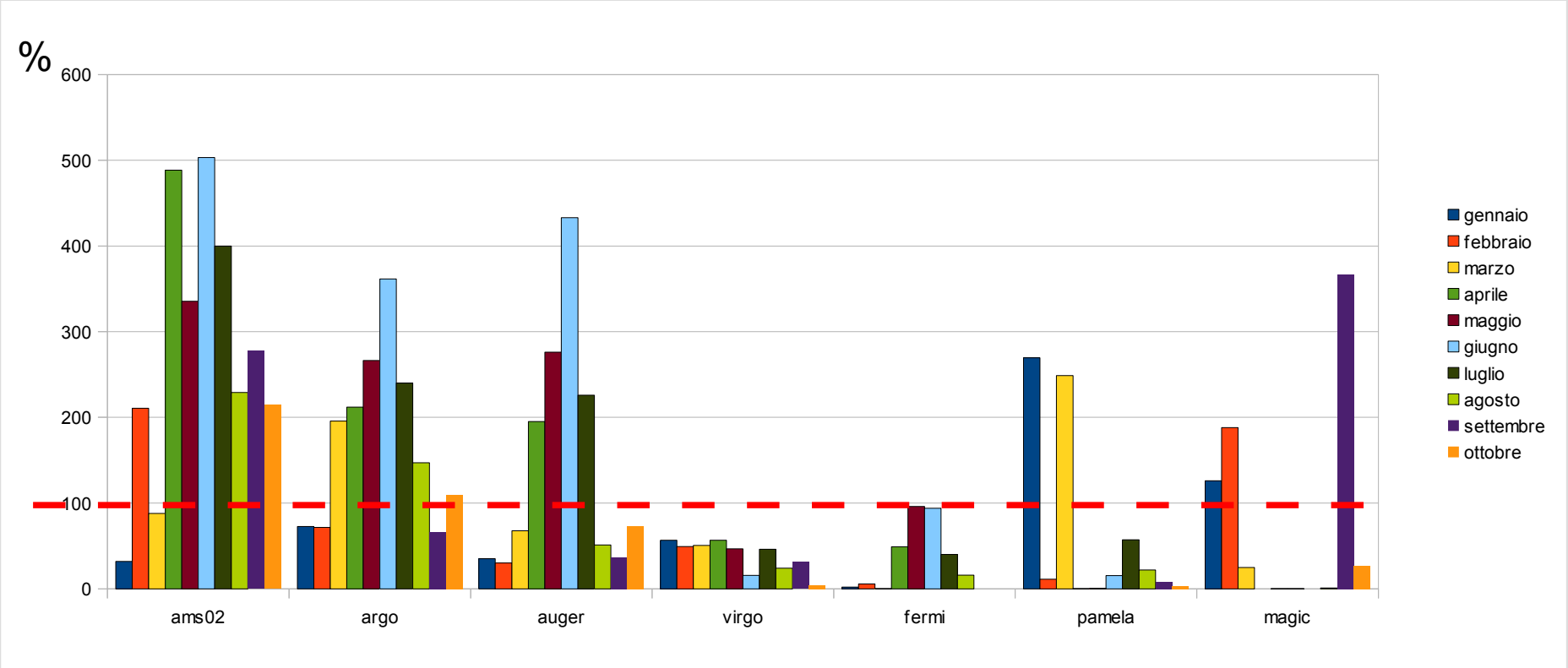
Utilizzo delle risorse: CPU

Parametri di indagine:

- wct (wall clock time) è il tempo in cui la risorsa di calcolo è occupata dal job di esperimento
- cpt (cpu time) è il tempo di cpu utilizzato dal job di esperimento
- **efficienza** di un job (cpt/wct)

Utilizzo delle risorse: CPU

Wall Clock Time medio in HEP-SPEC06 day / Pledged 2011 (%)



Il meccanismo di fairshare dinamico permette anche agli esperimenti non-LHC di utilizzare più delle risorse loro assegnate

Analisi sull'uso delle risorse ad oggi: CPU

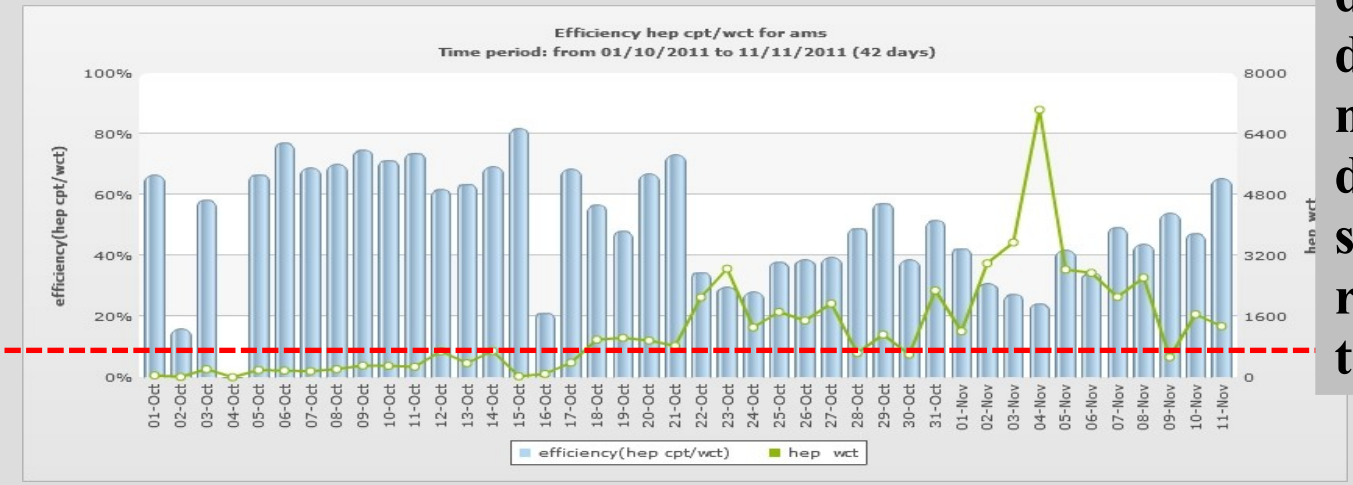
- La maggior parte degli esperimenti fa un utilizzo mediamente buono ed efficiente delle risorse
- Dall'analisi sull'uso delle risorse emergono essenzialmente tre modalità di utilizzo:
  1. uso continuo ed efficiente
  2. uso a “burst” ed efficiente
  5. uso continuo con inefficienze temporanee (risoluzione sul campo)

Efficienza (valor medio cpt/wct)
dal 1/01/2011 al 23/06/2011

Exp.	Eff.(cpt/wct)
ARGO	88%
AMS02	90%
PAMELA	97%
MAGIC	86%
FERMI	98%
AUGER	82%
VIRGO	61%

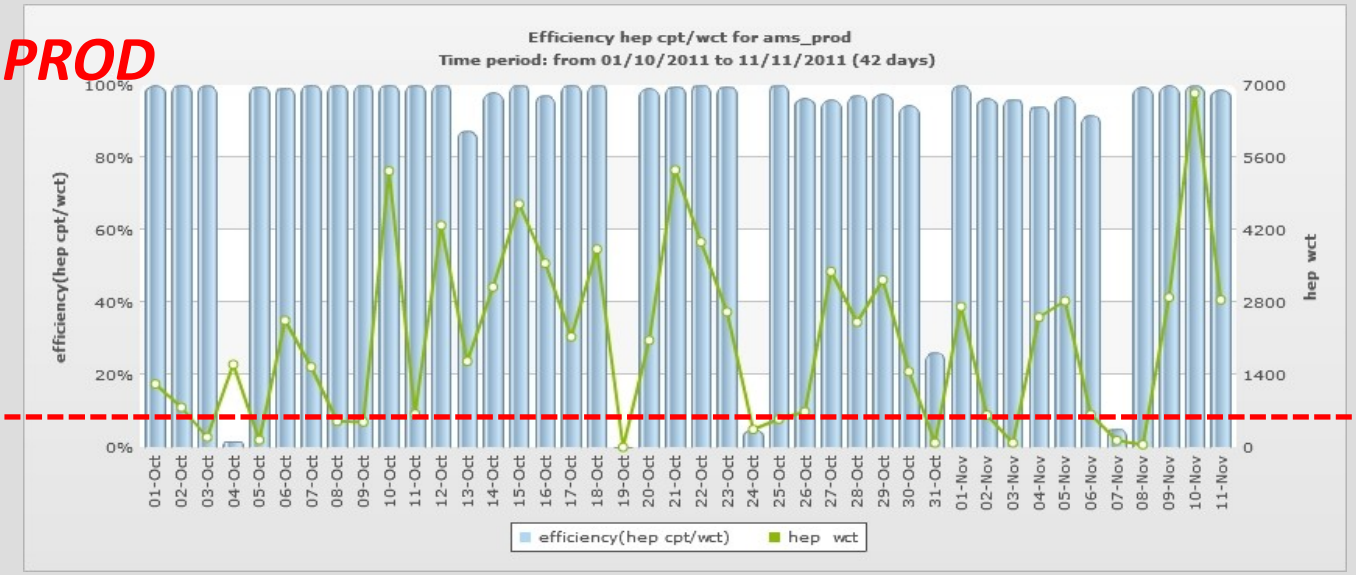
Analisi sull'utilizzo delle risorse: un esempio

AMS



Problema di dimensionamento dello storage: numero diskserver non sufficiente a reggere il throughput

AMS_PROD



Pledge:
432 HepSpec

Conclusioni

- Il Tier-1 da più di un anno sta usando a pieno regime le risorse per gli esperimenti LHC
- E' tuttavia in grado di ospitare agevolmente le risorse degli esperimenti non LHC e di agire come centro di calcolo primario per essi (Tier-0)
- Molti nuovi esperimenti hanno infatti manifestato l'esigenza di avere un backup di dati al Tier-1 del CNAF nell'ultimo periodo
- Molti passi avanti sono stati fatti per utilizzare al meglio le risorse assegnate agli esperimenti definendo sul campo il modello di storage e di computing meglio corrispondente alle esigenze di analisi
- Nuove strategie di virtualizzazione e l'allocazione dinamica delle risorse dovranno far fronte alle esigenze sempre più eterogenee degli esperimenti che si apprestano ad eseguire al Tier-1 analisi dedicate e simulazioni