Il primo anno di presa dati di LHC

L'esperienza di calcolo nell'esperimento ATLAS

Attività condotte nel 2010 e prospettive future

Lorenzo Rinaldi (INFN-CNAF)

Workshop CCR - Stato e Prospettive del Calcolo Scientifico

Legnaro 14-2-2011

Introduzione

L'esperienza Atlas nel primo anno di presa dati:
il punto di vista di un addetto ai lavori di computing

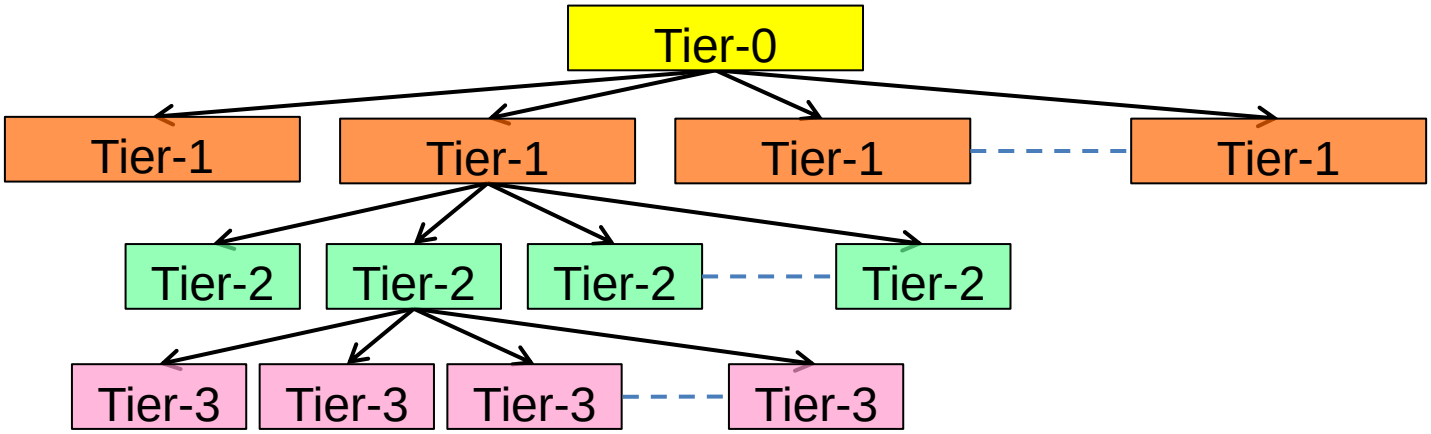
Descrizione del Computing Model

Le attività condotte nel 2010, nel mondo e in Italia

- La distribuzione dei dati su grid
- Le attività di calcolo centrali
- Le attività di analisi degli utenti

L'evoluzione del Computing Model di ATLAS: come il modello si è adattato alle problematiche reali

Il Computing Model di Atlas



Modello a cloud: gerarchico, multi-tier

Link Tier-0 Tier-1s

Cloud Italiana

T1-CNAF

T2
- Napoli
- Milano
- Roma1
- LNF

T3
- Genova
- Trieste
- Roma3
- (Bologna)

Attività nei siti previste dal CM

Tier-0 (CERN)

• Archivio dei RAW data ricevuti dal detector e distribuzione ai Tier1
• Prompt Reconstruction delle calibration e express streams e dei physics streams
• Distribuzione output ricostruzione (ESD, AOD,TAG) ai Tier-1

Tier-1

• Accesso a lungo termine e archivio su tape di un subset di RAW data
• Copia dei RAW data di un altro Tier-1
• Reprocessing della ricostruzione dei propri RAW data con parametri di calibrazioni e allineamenti finali e distribuzione AOD ai Tier-2
• Archivio su disco di tutti gli AOD e TAG data e di una frazione di ESD data
• Archivio dati simulati prodotti nei Tier-2
• Simulazione Monte Carlo
• Analisi di gruppo

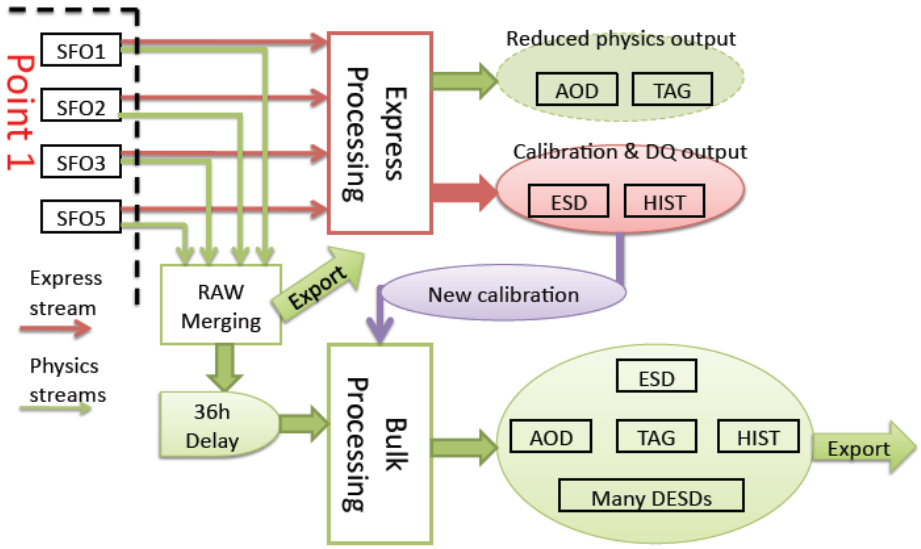
Tier-2

• Simulazione Monte Carlo
• Archivio di una frazione di AOD e TAG data
• Archivio dati utenti
• Analisi utente e di gruppo e sviluppo di codice
• Calibrazione e allineamento per sub-detectors di interesse locale

Tier-3

• Analisi locale e sviluppo di codice
• Archivio dati utenti

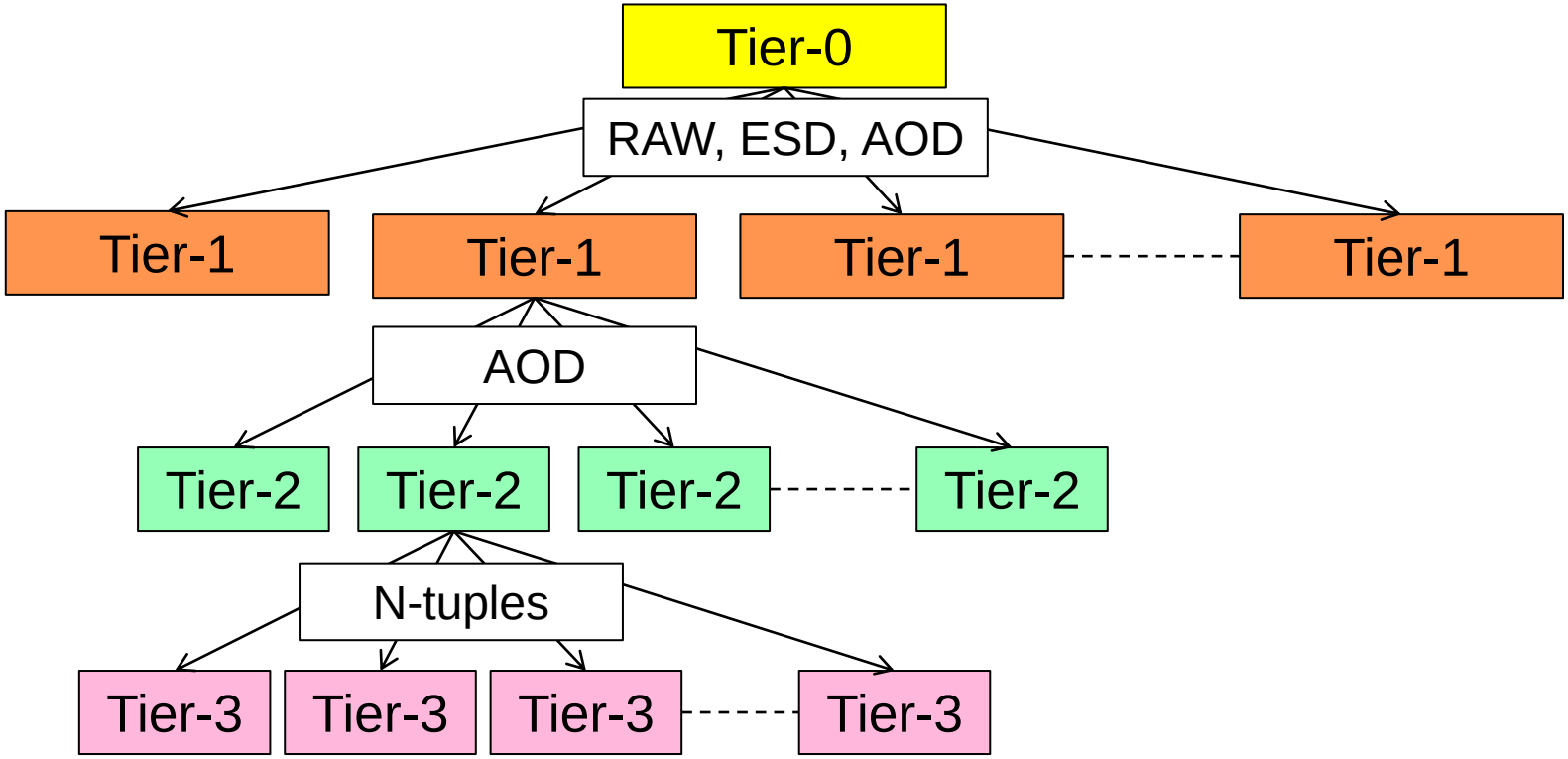
Attività del Tier-0



- Express streams ricostruiti con le calibrazioni best-estimate disponibili
- RAW data esportati subito dopo il merging
- Entro 36h i RAW data processati con le calibrazioni aggiornate
- Export dati derivati

RAW Data: dati in output dal sistema di trigger e acquisizione in formato byte-stream
ESD (Event Summary Data)*:* output della ricostruzione, contiene calibrazione, allineamento, refitting …
Rappresentazione object-oriented POOL/ROOT
AOD (Analysis Object Data)*:* rappresentazione ridotta degli eventi per l'analisi.
Back Navigation agli ESD
Rappresentazione object-oriented POOL/ROOT
TAG: informazioni sintetiche per selezione veloce degli eventi negli AOD e/o ESD.
Formato DB relazionale per ricerche veloci e formato ROOT per istogrammazione veloce
dESD e ***dAOD:*** versione "skimmate" di ESD e AOD

Computing Model: data workflow Tier-1s Tier-2s Tier-3s



- Trasferimenti tra le cloud
  - Tier1 ←→ Tier1: Dati riprocessati (ESD, AOD, dESD per l'analisi)
  - Tier 1/2 → Tier2/3: Output analisi utente su storage area locale
- Trasferimenti nella cloud
  - Tier1 → Tier2: distribuzione nei Tier2 dei dati per l'analisi e cancellazione dal Tier1
  - Tier1 ← Tier2: simulazioni Monte Carlo
  - Tier1/2 → Tier2/3: Output analisi utente su storage area locale

Computing Model: cosa è cambiato subito

Replica dei dati secondo il Computing Model

- ✓ RAW: 1 copia distribuita sull'insieme dei Tier1 (su tape)
- ✓ ESD: 2 copie sull'insieme dei Tier1. Replica nei Tier2 on demand
- ✓ AOD: 2 copie sull'insieme dei Tier1. 10 copie sull'insieme dei Tier2 (~ 1 per cloud)
- ✓ dESD: non replicati ai Tier1. 10 copie sull'insieme dei Tier2 (~ 1 per cloud)
- ✓ Distribuzione nei Tier1 in base al "Tier1 ratio"

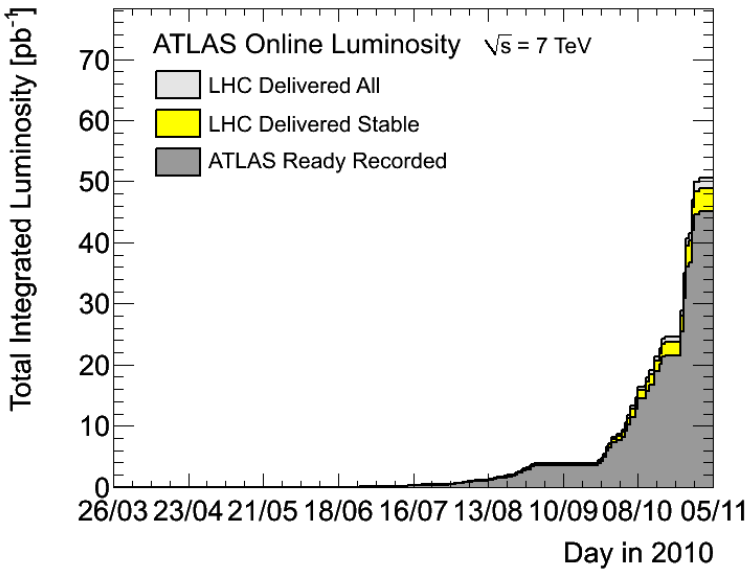
Replica dei dati all'inizio della presa dati

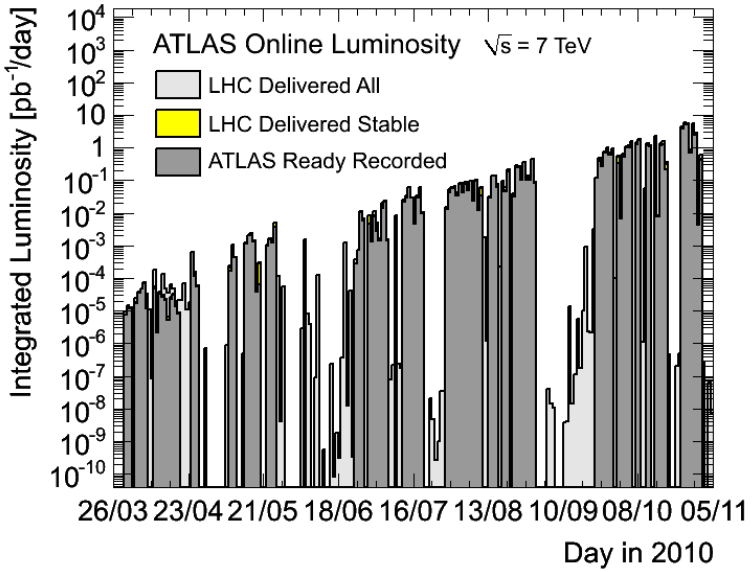
I siti o le cloud grandi copiano più dati dei formati più popolari, rispetto a quanto previsto dal Computing Model, per massimizzare l'analisi

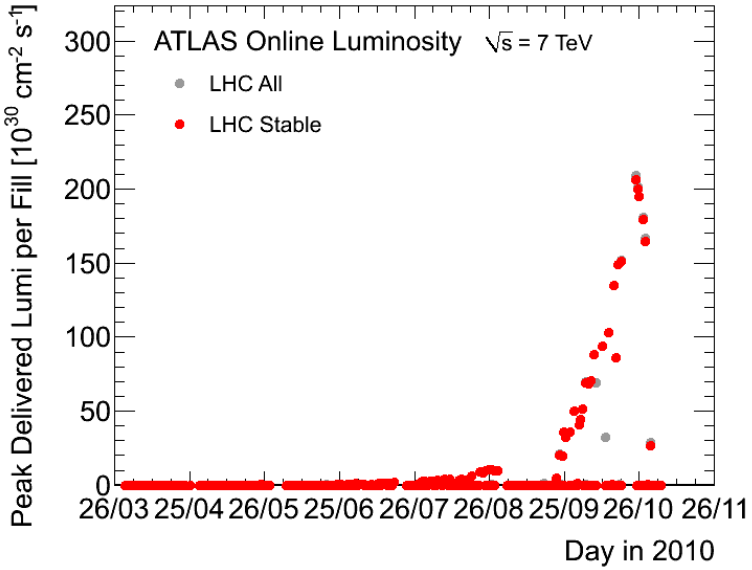
- ESD: 7 copie
  - 3.5 copie in US. Una copia completa a BNL e 2.5 copie nell'insieme dei Tier2
  - 1 copia completa in FR

cloud con piccole percentuali di dati fanno poca analisi e cloud grandi diventano attrattori per i job anche degli stranieri

Summary presa dati 2010 p-p @7 TeV



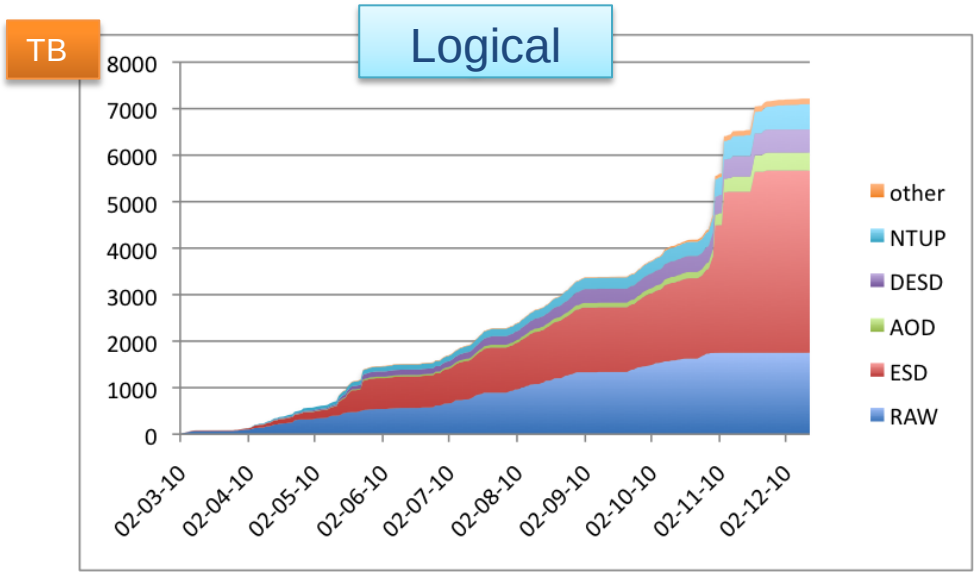




Int Lumi = 45.03 pb $^{-1}$

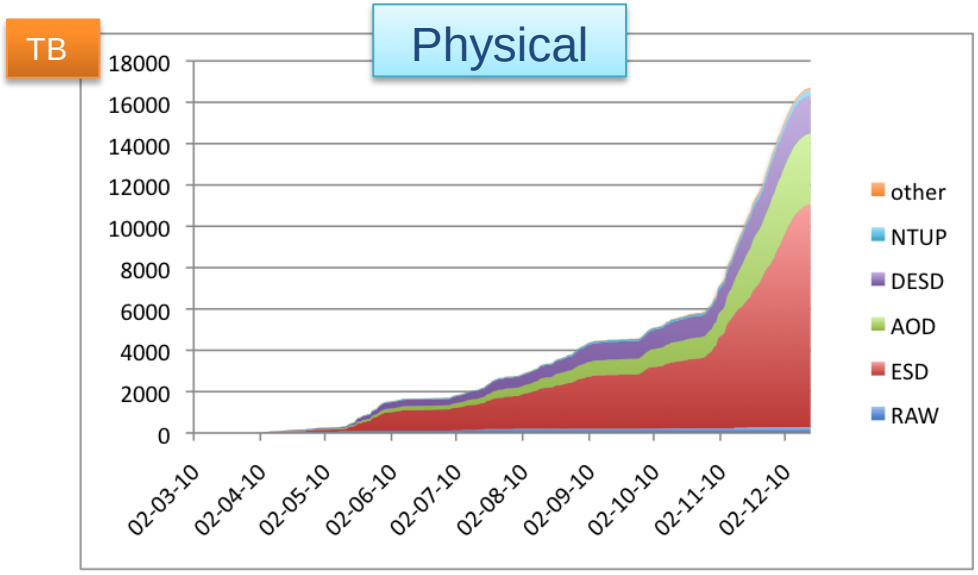
Peak Lumi = 2.07x10^{32} cm $^{-2}$ s $^{-1}$

Volume dati p-p @ 7 TeV



Logical Data
Volume dei dati regiatrati su catalogo (singola copia)

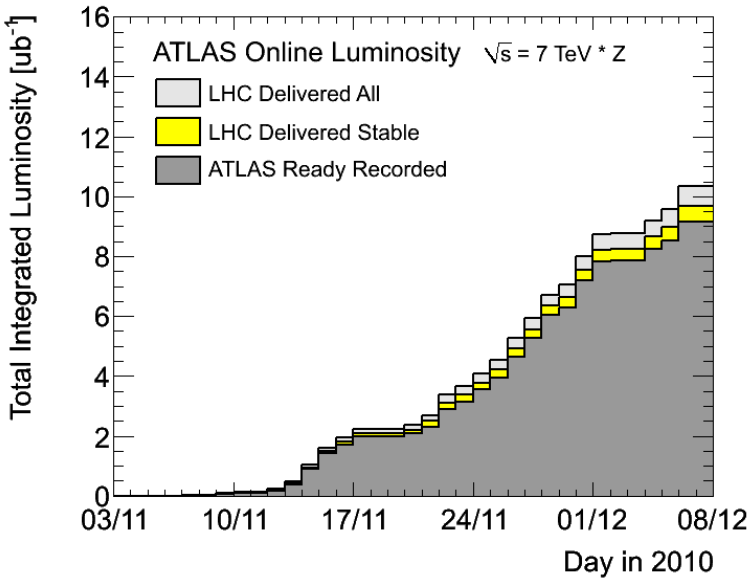
TOTAL	**7216 TB**
RAW	1751 TB
ESD	3920 TB
AOD	384 TB
DESD	498 TB
NTUP	543 TB

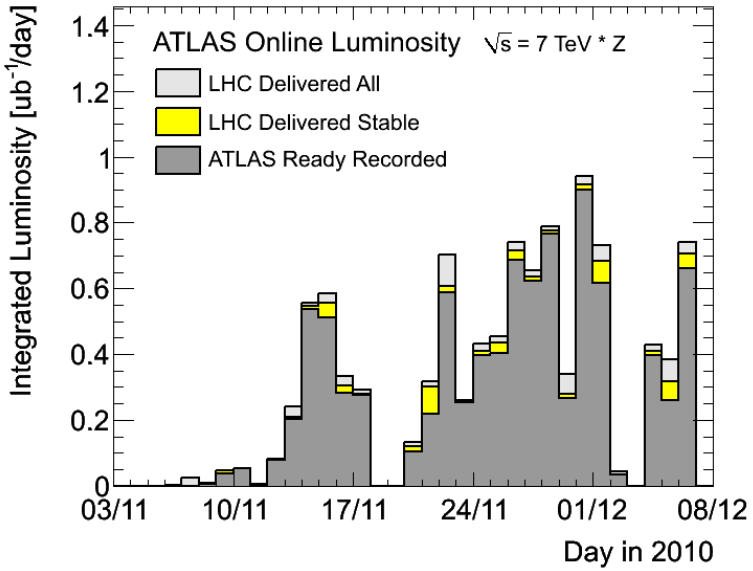


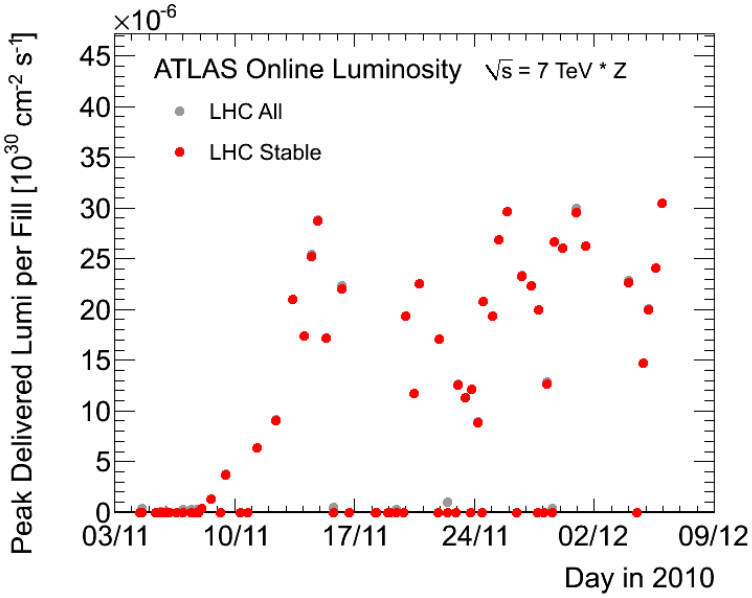
Physical Data
Comprende tutte le repliche distribuite in GRID
Solo dati replicati su disco

TOTAL	**16695 TB**
RAW	279 TB
ESD	10772 TB
AOD	3433 TB
DESD	1835 TB
NTUP	264 TB

Summary presa dati 2010 HI

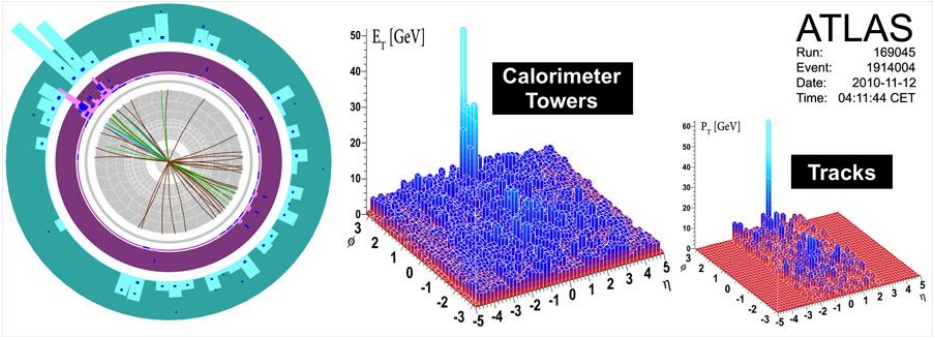




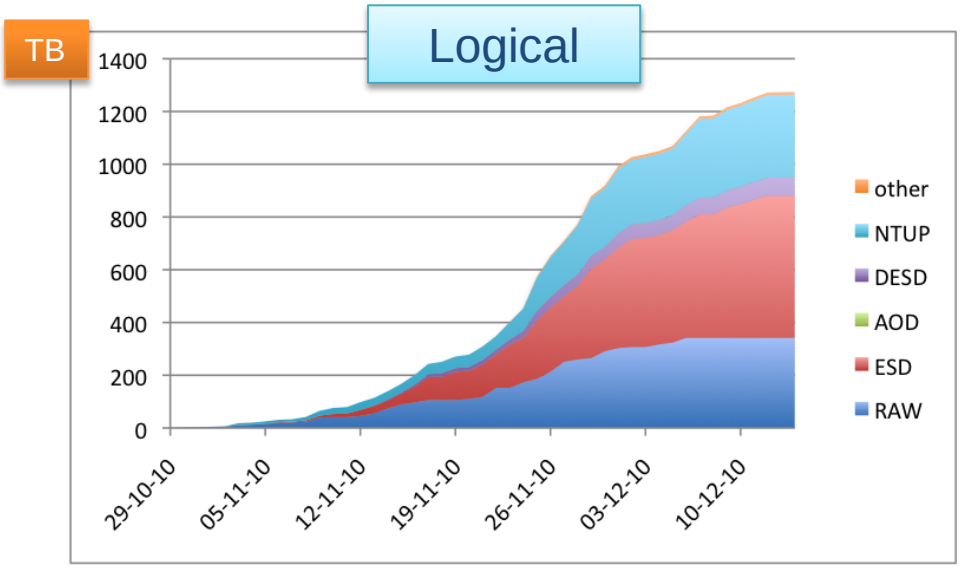


Int Lumi = 9.17 μb^{-1}

Peak Lumi = 3.04×10^{25} cm $^{-2}$ s $^{-1}$

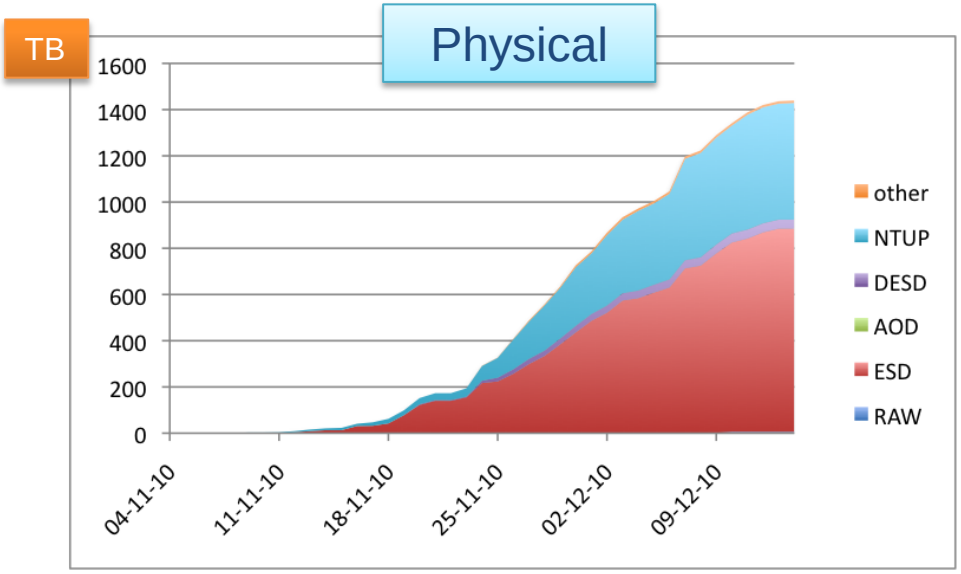


Volume dati Pb-Pb



Logical Data
Volume dei dati regiatrati su catalogo (singola copia)

TOTAL	**1274 TB**
RAW	342 TB
ESD	539 TB
DESD	70 TB
NTUP	313 TB
OTHER	10 TB



Physical Data
Comprende tutte le repliche distribuite in GRID
Solo dati replicati su disco

TOTAL	**1439 TB**
RAW	6 TB
ESD	879 TB
DESD	39 TB
NTUP	505 TB
OTHER	10 TB

Distribuzione totale dei dati 2010

MB/s per day

Reprocessing dati e MC

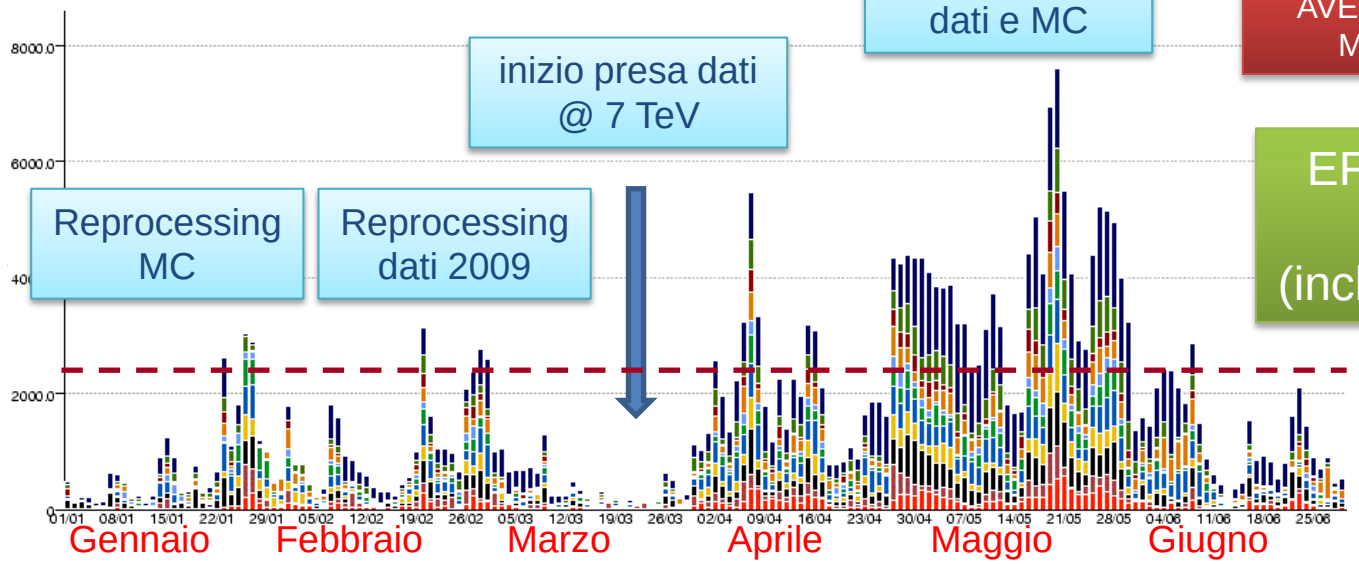
AVERAGE:2.3 GB/s
MAX 7.5 GB/s

inizio presa dati @ 7 TeV

Reprocessing MC

Reprocessing dati 2009

EFFICIENCY: 100% (including retries)



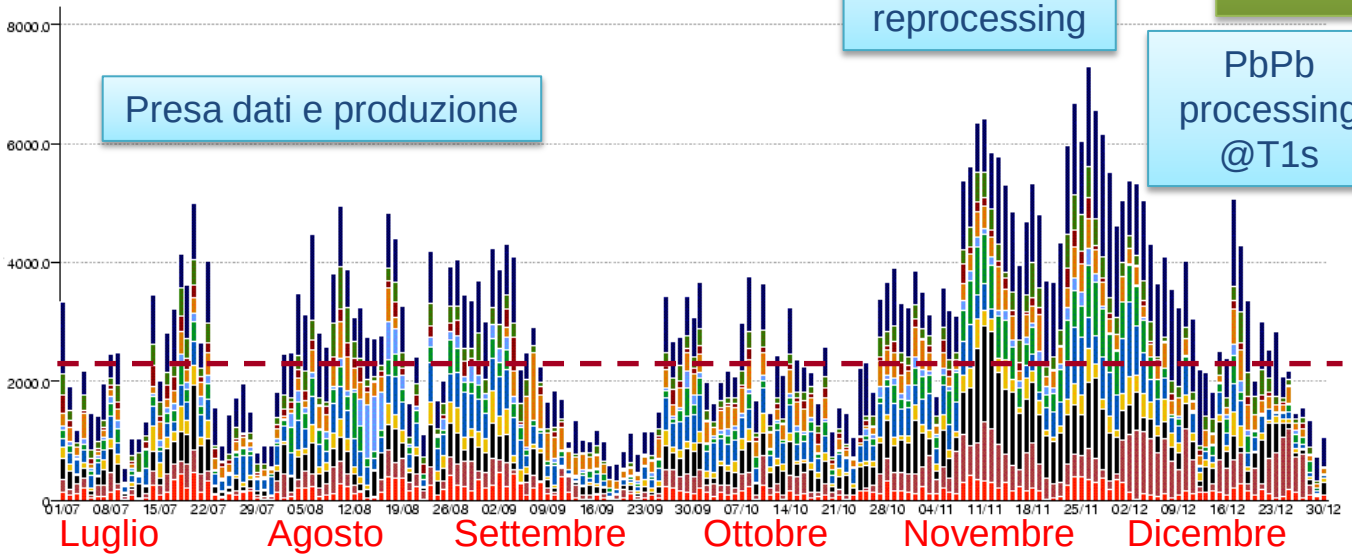
Dati disponibili nei siti dopo poche ore

MB/s per day

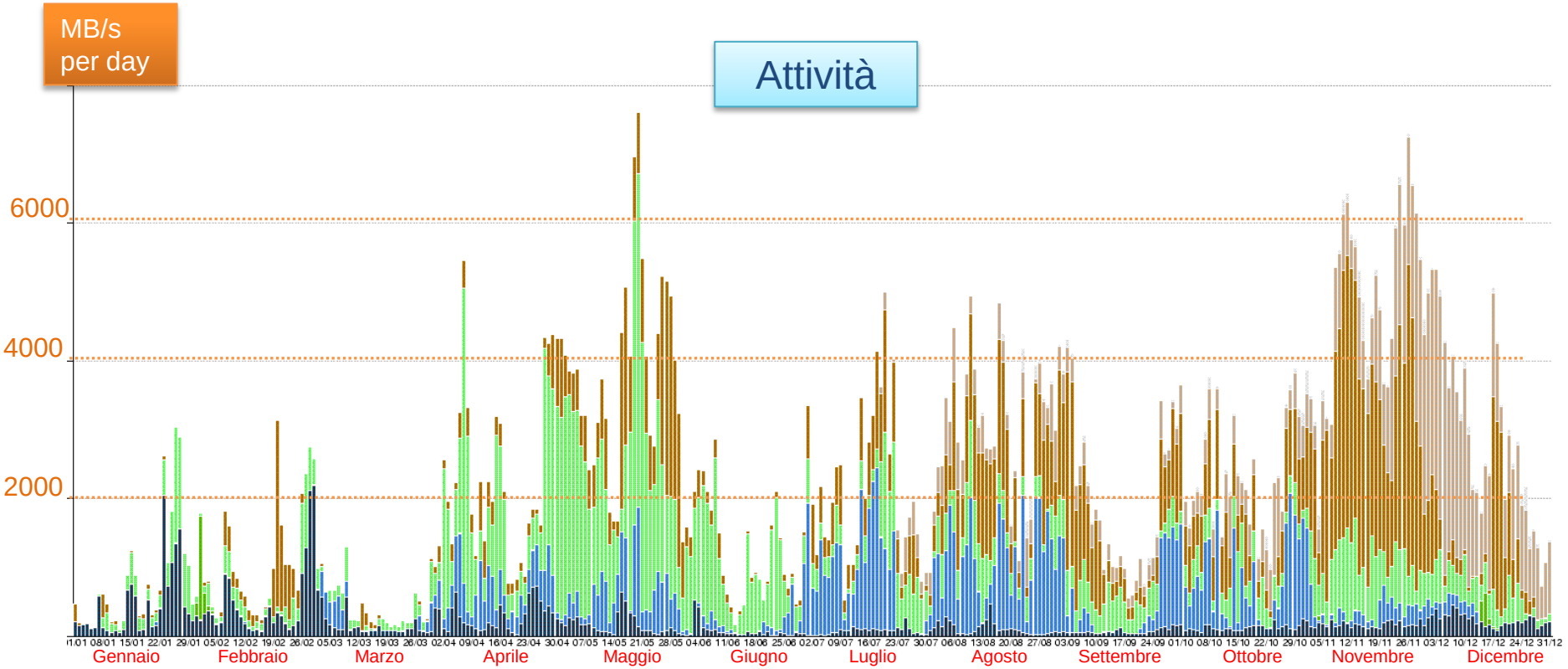
2010 pp reprocessing

PbPb processing @T1s

Presa dati e produzione



Distribuzione totale dei dati 2010



Produzione
- transf. dati da produzione
- interni alle cloud

Tier-0 export
- flusso dati da tier0
- include calibration streams

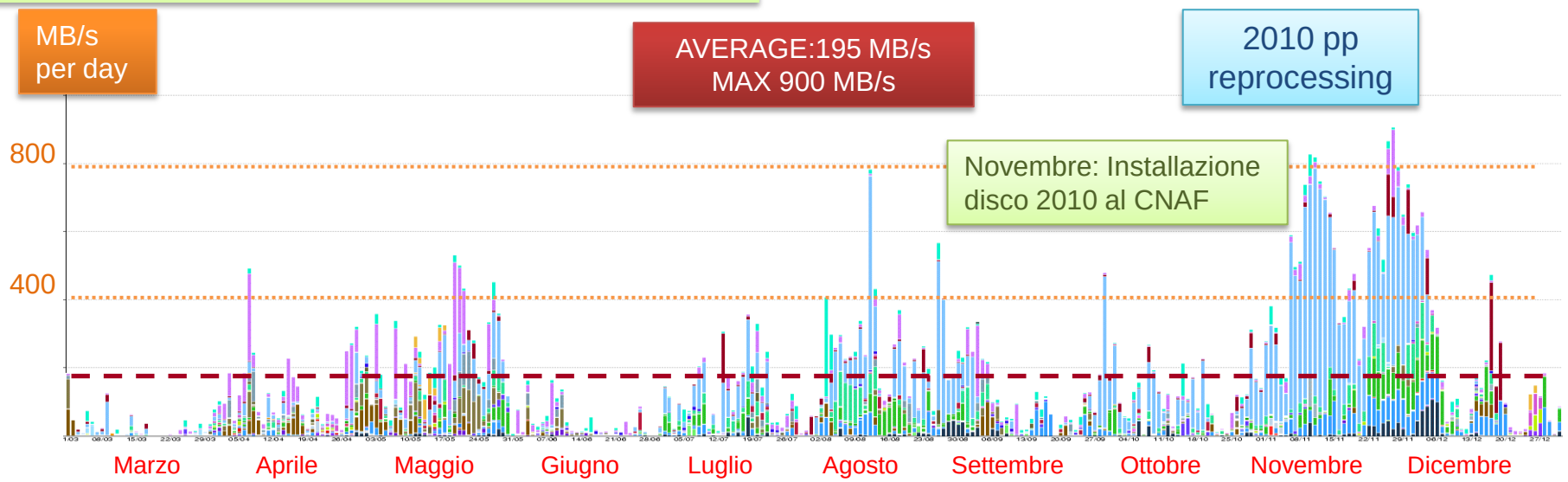
Sottoscrizioni utenti

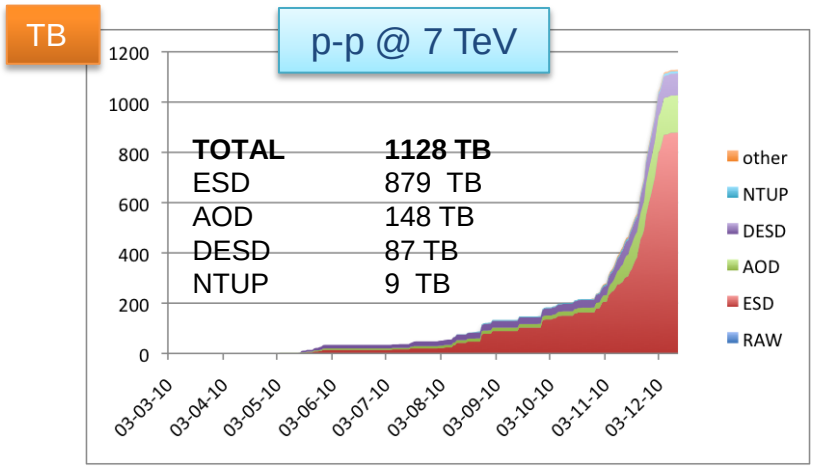
Functional test

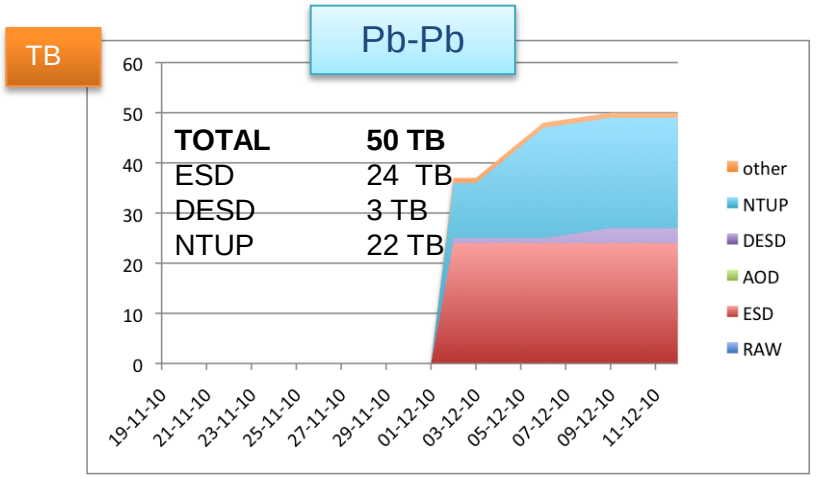
Data consolidation
- Transf. dati da produzione
- extra-cloud

Data brokering
- trasf. dati analisi

Distribuzione dei dati in Italia



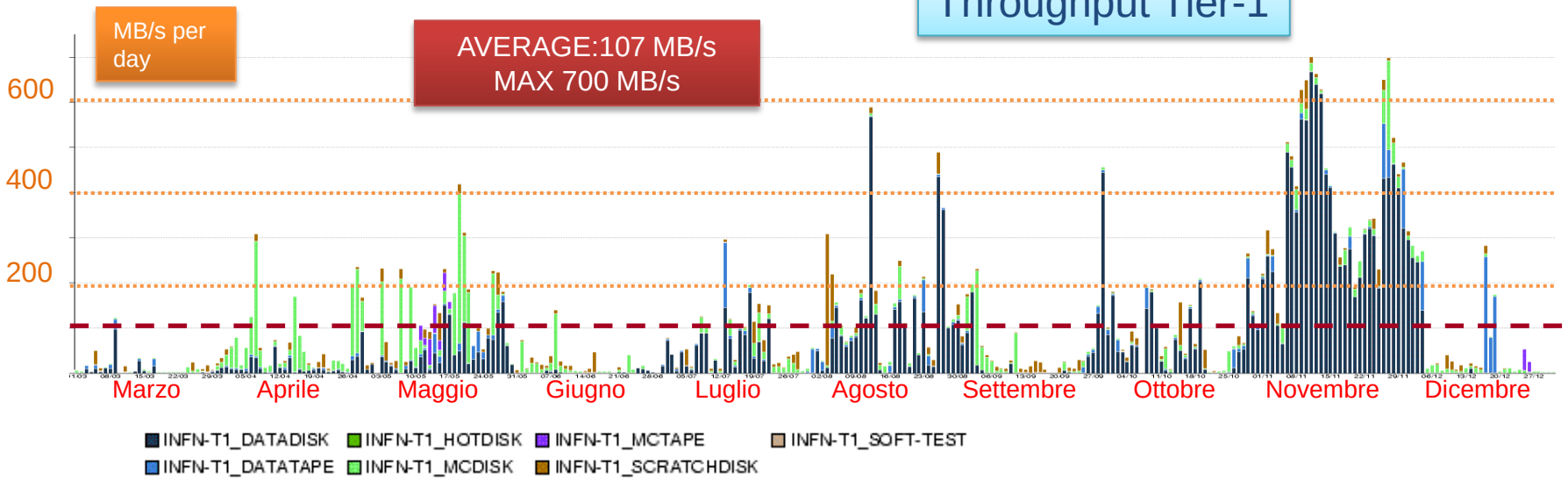




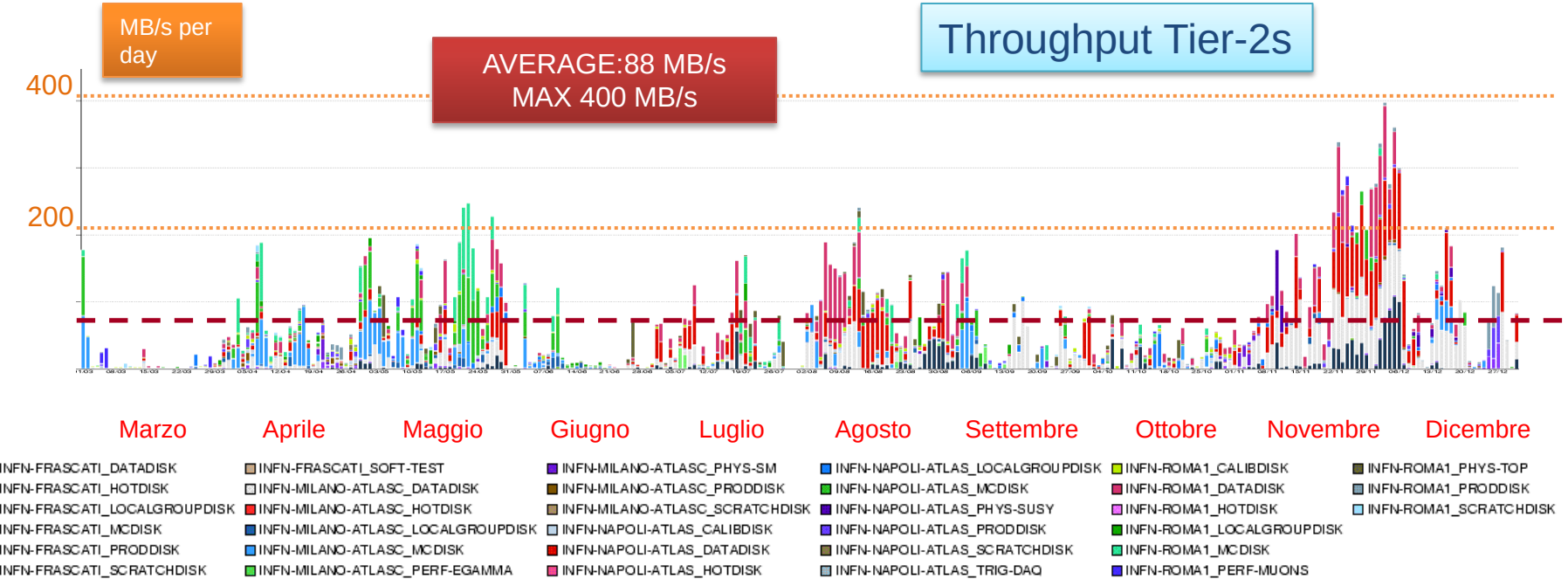
INFN-FRASCATI_DATADISK
INFN-FRASCATI_HOTDISK
INFN-FRASCATI_LOCALGROUPDISK
INFN-FRASCATI_MCDISK
INFN-FRASCATI_PRODDISK
INFN-FRASCATI_SCRATCHDISK
INFN-FRASCATI_SOFT-TEST
INFN-GENOVA_HOTDISK
INFN-GENOVA_LOCALGROUPDISK
INFN-GENOVA_SCRATCHDISK
INFN-MILANO-ATLASC_DATADISK
INFN-MILANO-ATLASC_HOTDISK
INFN-MILANO-ATLASC_LOCALGROUPDISK
INFN-MILANO-ATLASC_MCDISK
INFN-MILANO-ATLASC_PERF-EGAMMA
INFN-MILANO-ATLASC_PHYS-SM
INFN-MILANO-ATLASC_PRODDISK
INFN-MILANO-ATLASC_SCRATCHDISK
INFN-NAPOLI-ATLAS_CALIBDISK
INFN-NAPOLI-ATLAS_DATADISK
INFN-NAPOLI-ATLAS_HOTDISK
INFN-NAPOLI-ATLAS_LOCALGROUPDISK
INFN-NAPOLI-ATLAS_MCDISK
INFN-NAPOLI-ATLAS_PHYS-SUSY
INFN-NAPOLI-ATLAS_PRODDISK
INFN-NAPOLI-ATLAS_SCRATCHDISK
INFN-NAPOLI-ATLAS_TRIG-DAQ
INFN-ROMA1_CALIBDISK
INFN-ROMA1_DATADISK
INFN-ROMA1_HOTDISK
INFN-ROMA1_LOCALGROUPDISK
INFN-ROMA1_MCDISK
INFN-ROMA1_PERF-MUONS
INFN-ROMA1_PHYS-TOP
INFN-ROMA1_PRODDISK
INFN-ROMA1_SCRATCHDISK
INFN-ROMA3_HOTDISK
INFN-ROMA3_LOCALGROUPDISK
INFN-ROMA3_PRODDISK
INFN-ROMA3_SCRATCHDISK
INFN-T1_DATADISK
INFN-T1_DATATAPE
INFN-T1_HOTDISK
INFN-T1_MCDISK
INFN-T1_MCTAPE
INFN-T1_SCRATCHDISK
INFN-T1_SOFT-TEST
INFN-TRIESTE_HOTDISK
INFN-TRIESTE_LOCALGROUPDISK
INFN-TRIESTE_SCRATCHDISK

Distribuzione dei dati in Italia

Throughput Tier-1



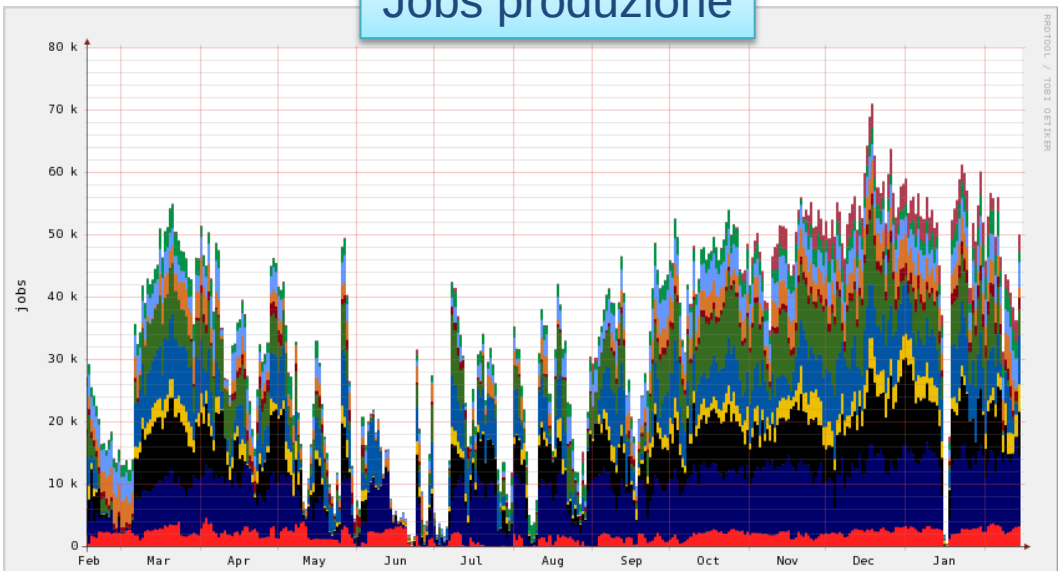
Throughput Tier-2s



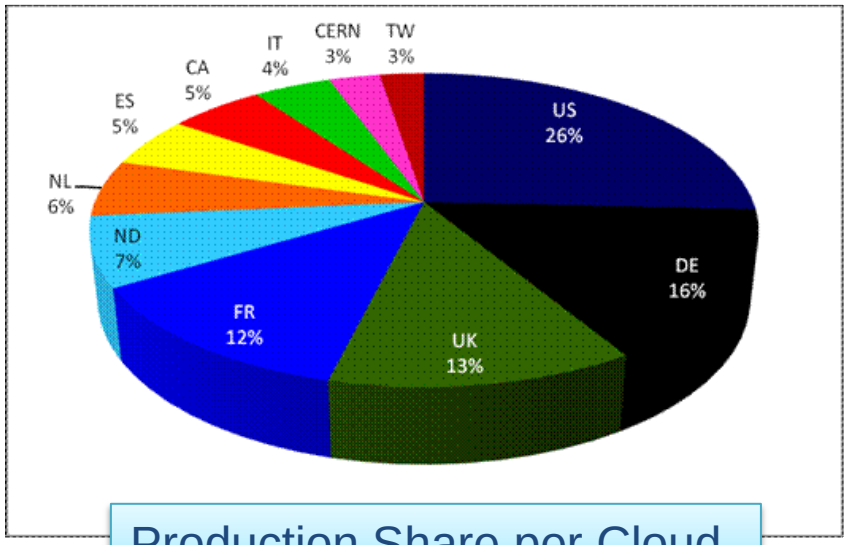
Produzione in ATLAS

Attività di elaborazione dei dati gestite centralmente in Atlas:
- simulazione
- ricostruzione
- reprocessing

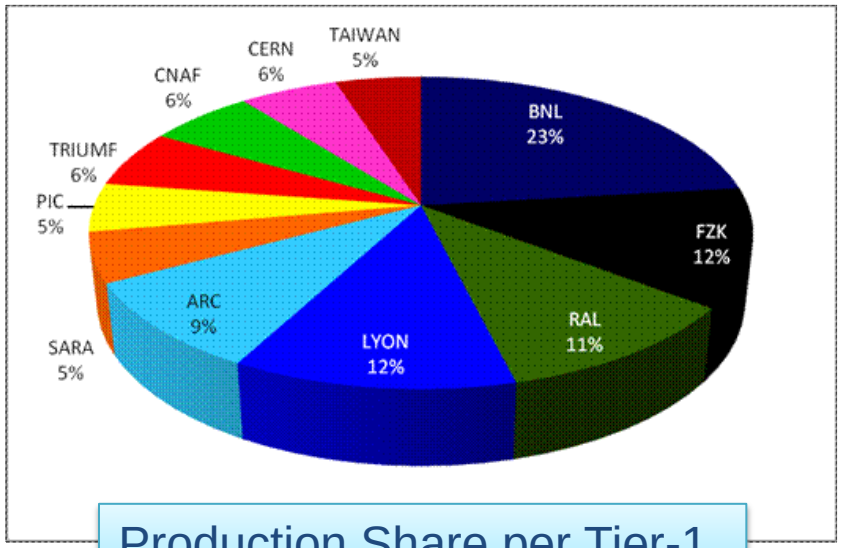
Jobs produzione



Numero medio di jobs running per sito

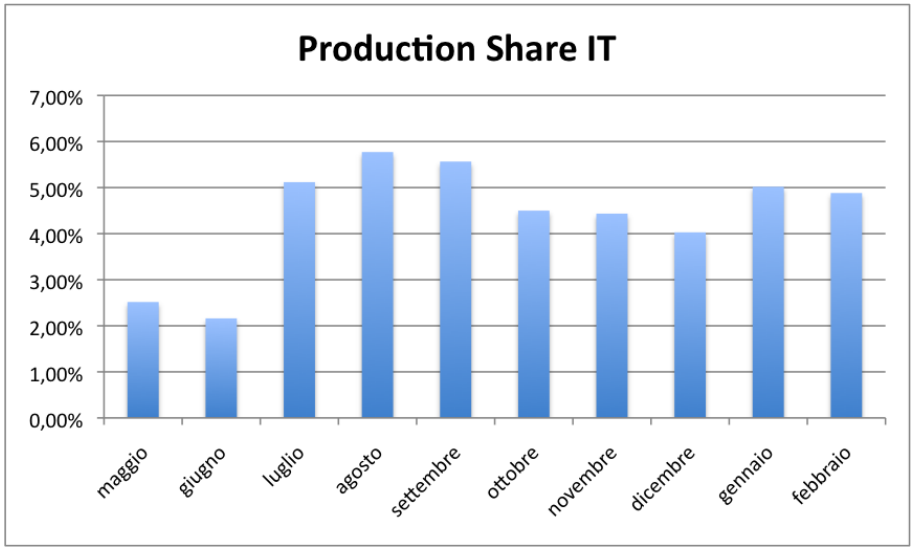


Production Share per Cloud

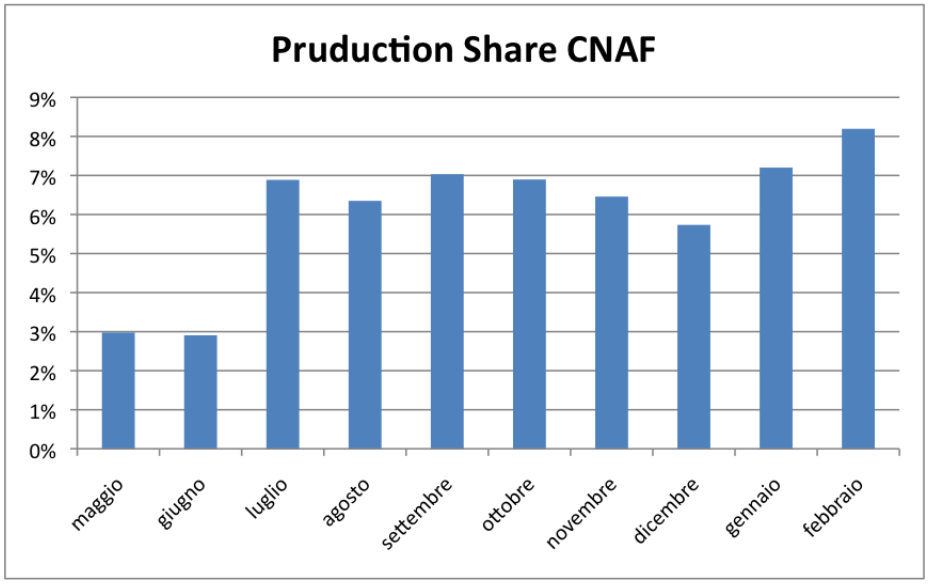


Production Share per Tier-1

Produzione in Italia



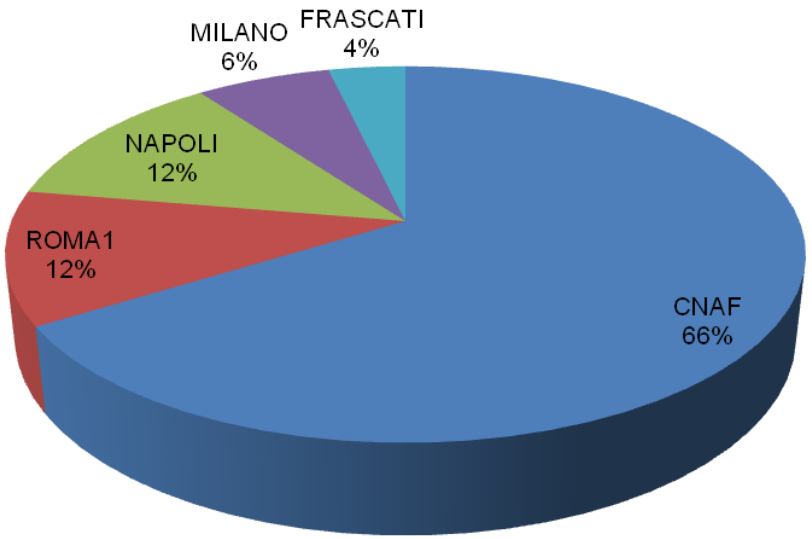
Da Luglio 2010:
- Share dei siti italiani cresciuto dopo l'istallazione delle cpu al CNAF
- variazione dovuta all'aumento delle risorse destinate all'analisi



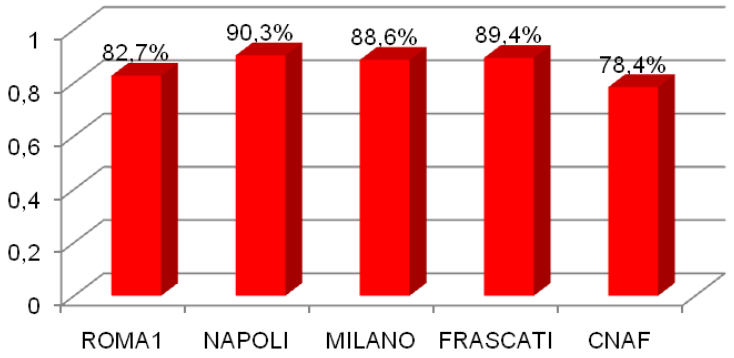
Ultimo mese:
- Share CNAF 8% (=pledge)
- Share IT 5% (pledge non ancora installate nei Tier2)

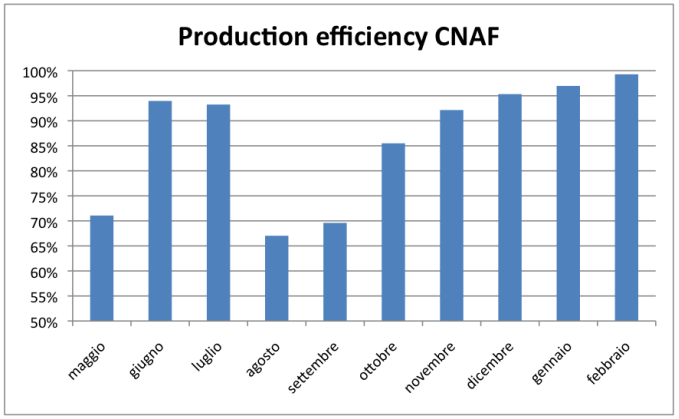
Produzione in Italia

Production share Cloud IT



Production efficiecy





Dati periodo 05/2010 – 02/2011

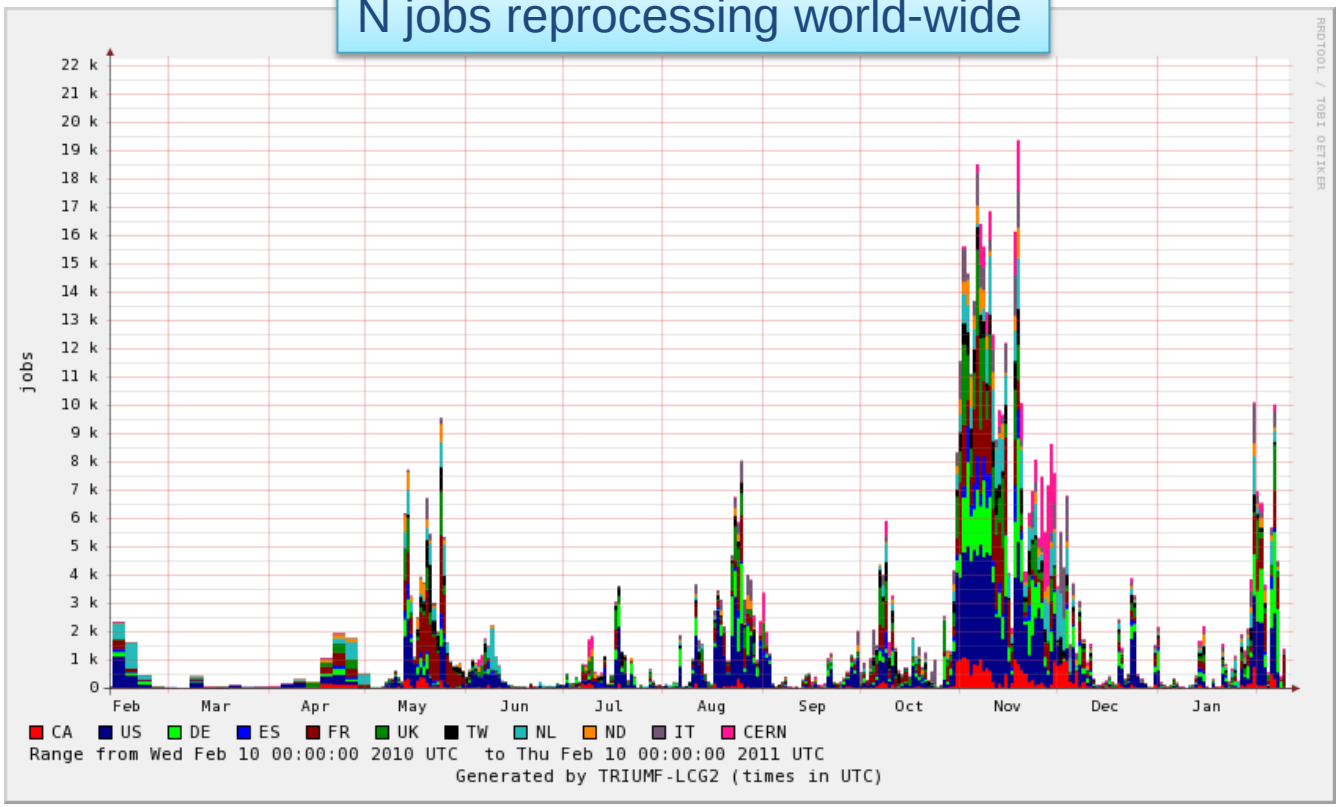
Il CNAF include anche le statistiche di reprocessing (60k jobs)

Efficienza= succ. jobs / totale

- Inefficienze dovute principalmente a problemi nei siti
- Non sempre i siti hanno pieno controllo (power cut)
- Ottime efficienze negli ultimi mesi

Reprocessing

N jobs reprocessing world-wide



4 campagne di reprocessing nel 2010:

- Febbraio: dati 2009 e cosmici
- Aprile: dati 2009 e 2010
- Maggio: dati 2009 e 2010 e MC
- Novembre: dati e MC full 2010 (da tape)

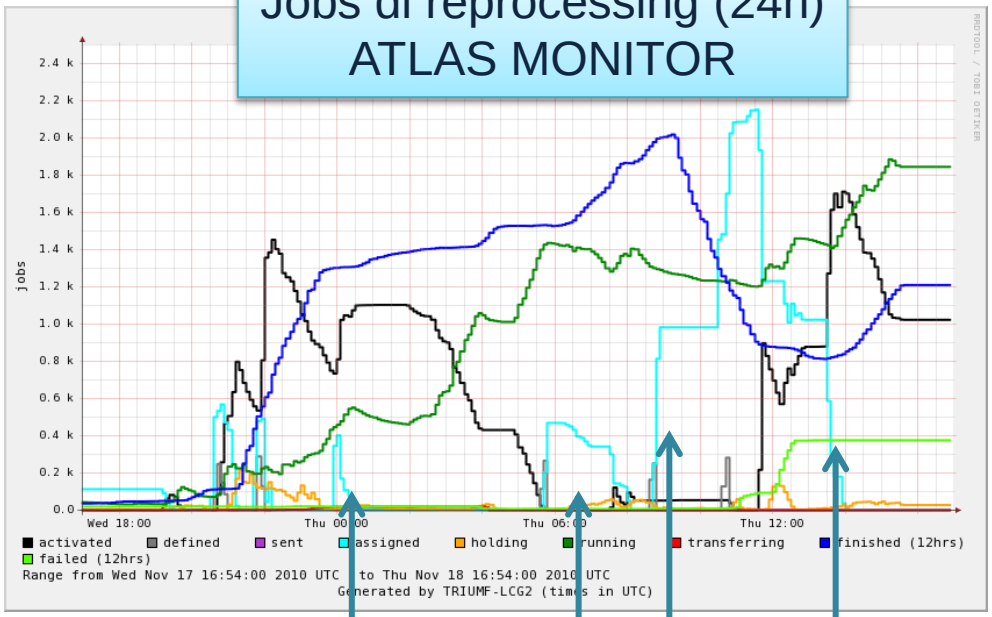
Reprocessing Ioni previsto a Marzo 2010

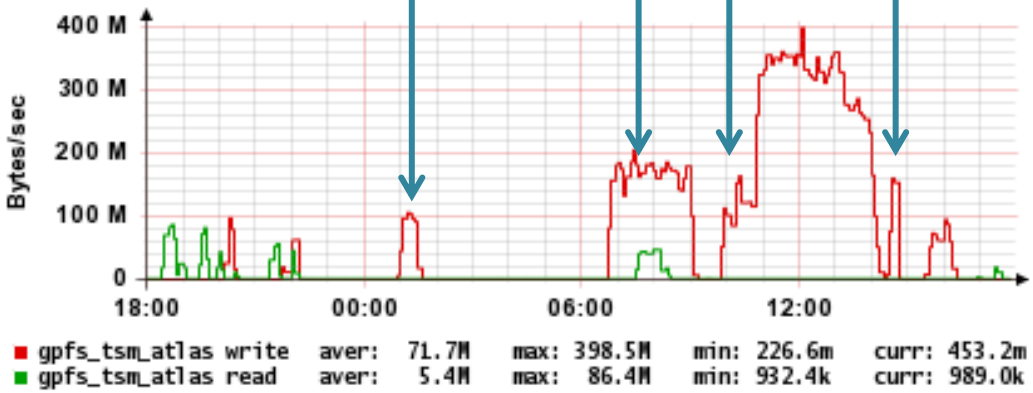
Reprocessamento del 100 % dei dati

- RAW → ESD
- ESD merge
- ESD → AOD, dESD
- Distribuzione nuovi dati nella Grid

Reprocessing FALL 2010 al CNAF

Jobs di reprocessing (24h)
ATLAS MONITOR





Recall da tape (TSM-GPFS)
CNAF MONITOR

Al CNAF riprocessato 5%
Dei dati p-p 2010

54 TB (con recall da tape)
Più di 30k jobs

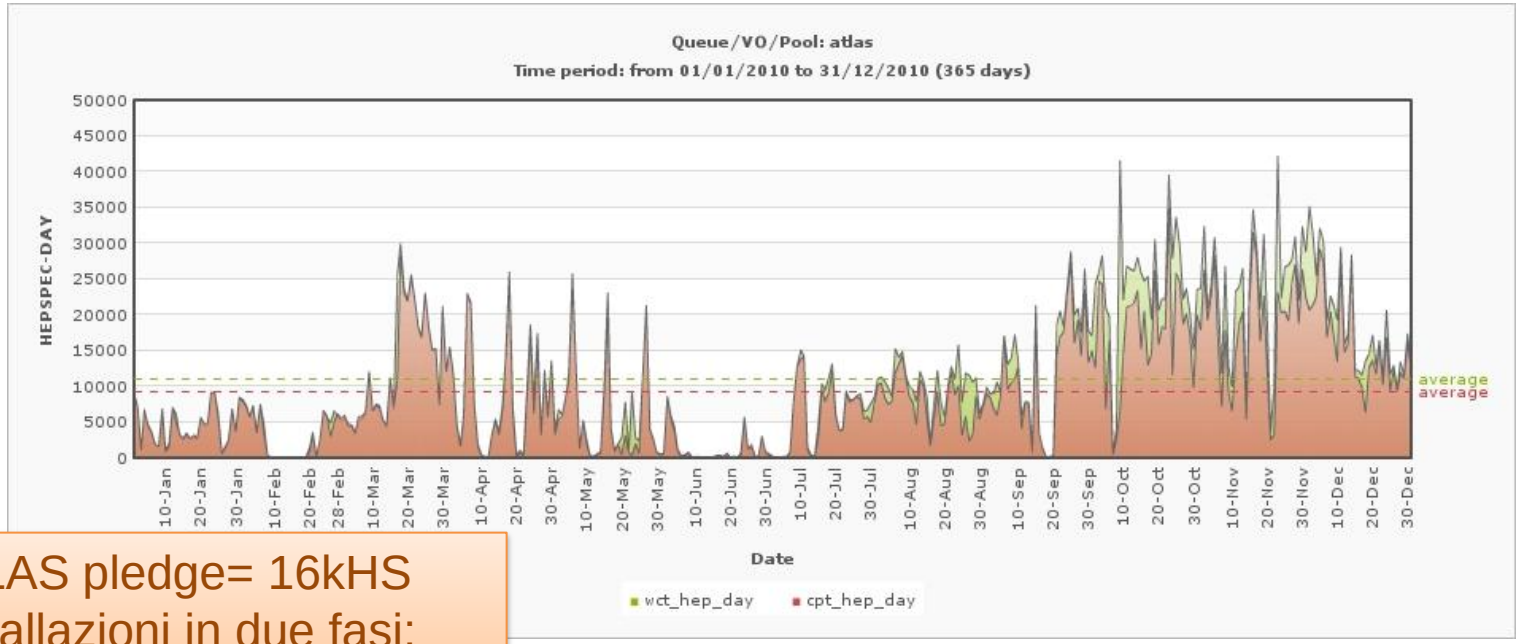
Metrica: 100% in due settimane

Ottime prestazioni del CNAF:
- uso estensivo CPU (jobs terminati in anticipo)
- Storage performante
- Rete al limite della saturazione (concorrenza jobs analisi)

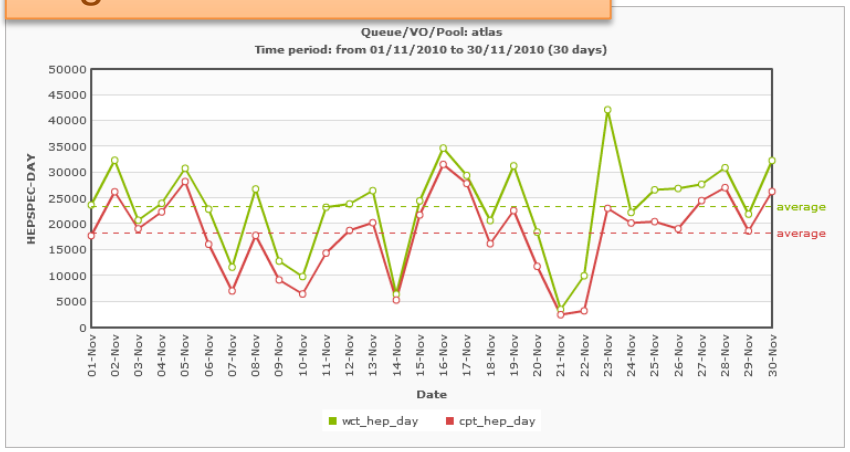
Job workflow:

- activated (pilot)
- defined (tape recall)
- running
- finished

Utiizzo risorse atlas al CNAF 2010



ATLAS pledge= 16kHS
Installazioni in due fasi:
Luglio e Settembre

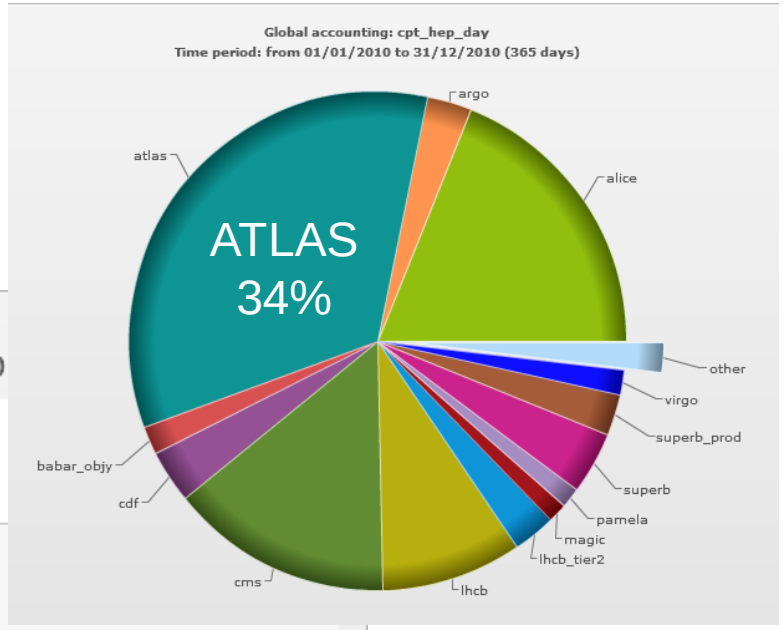


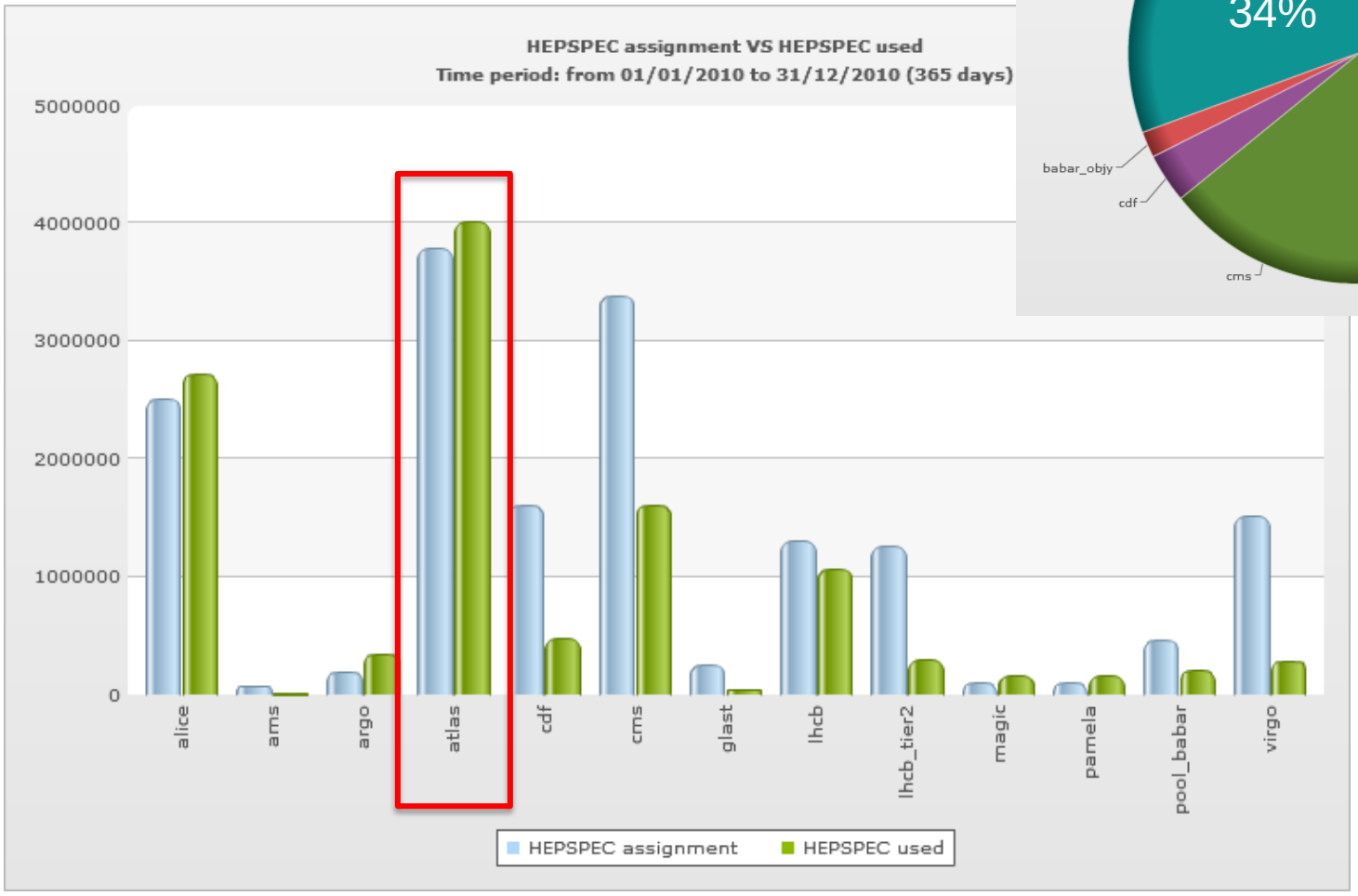
2010	MAX	AVERAGE
WCT (hepspec day)	42158.514	10997.233
CPT (hepspec day)	34883.907	9189.336

Nov. 2010	MAX	AVERAGE
WCT (hepspec day)	42158.514	23316.113
CPT (hepspec day)	31564.501	18228.361

Utiizzo risorse atlas al CNAF 2010

Assegnati: 3.8M HEPSPEC
Utilizzati: 4.0M HEPSPEC



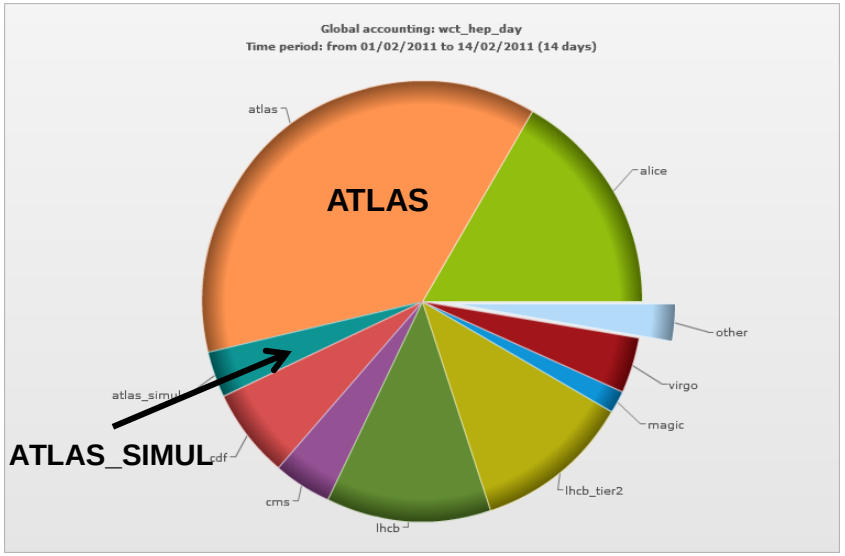


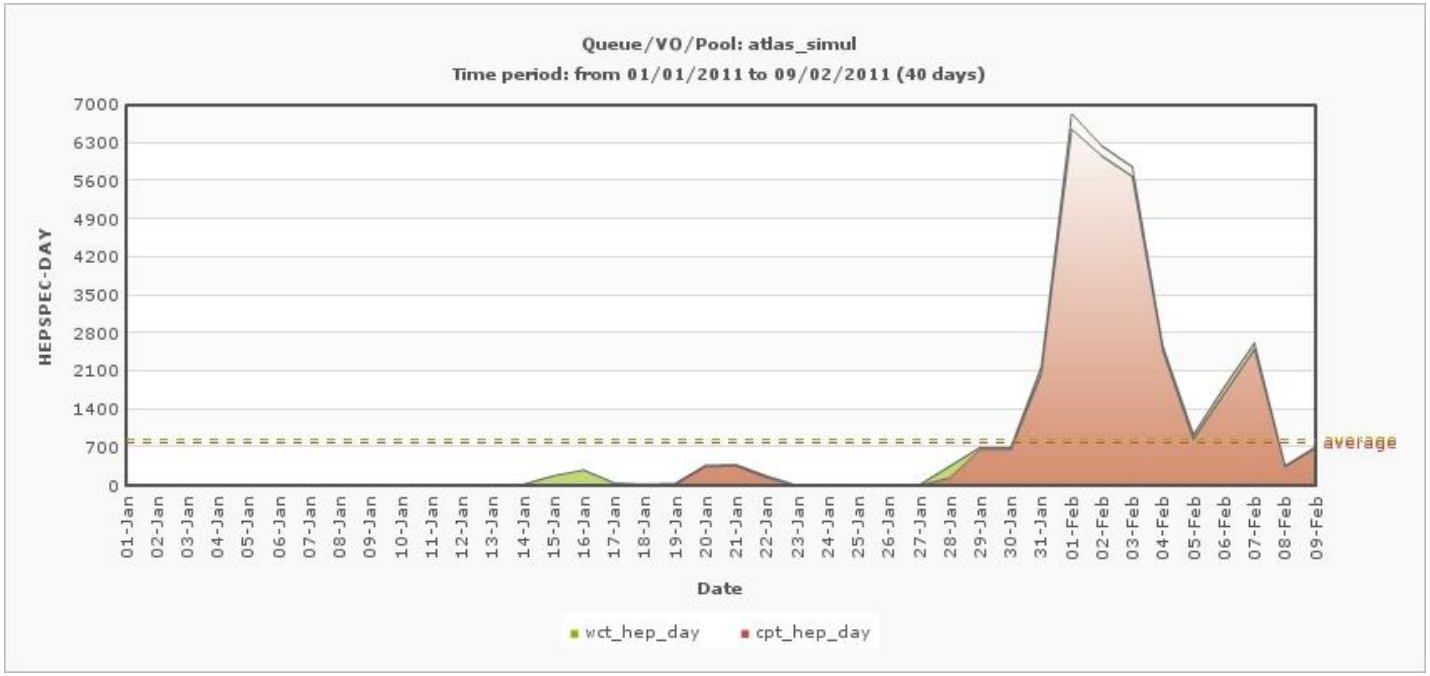
Atlas ha beneficiato bassa attività di altre VO in alcuni periodi

Utiizzo risorse atlas al CNAF 2011

ATLAS_SIMUL:
Abilitata nuova coda per job di simulazione:

- job solo su VWN
- limitato utilizzo di MEM fisica e basso IO
- Share 4%





Workflow analisi utenti

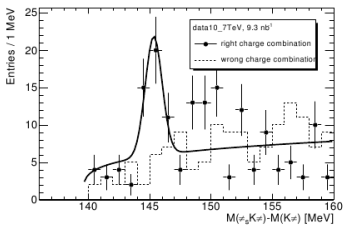
Central processing T0-T1-T2

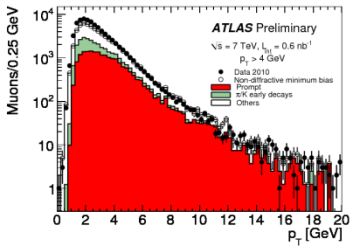
(d)ESD (d)AOD

low-level analysis

N-tuple DP*D

high-level analysis





Formati originali: RAW→(d)ESD→(d)AOD
Processati e distribuiti su grid (Tier0-Tier1-Tier2)

Analisi low-level (Grid/Local):

- SKIMMING: selezione eventi (es. QCD: solo eventi con jets)
- SLIMMING: selezioni data container nell'evento (es. rimozione algoritmi jets)
- THINNING: selezione di parti dei data container (es. rimozione parametri tracce)
- USER INFO: aggiunta di variabili utente

La struttura del dato ridotto resta invariata: stesso codice di analisi su dati originali/ridotti

Analisi high level (istogrammi, fit, …)
Analisi eseguita localmente (possibilià di running su grid, disponibili tool per analisi ROOT-on-GRID)

Analisi distribuita: il ruolo dei Tier-3

ATLAS ha sviluppato un modello di Tier3 definendo le funzionalità e i tool specifici
- Analisi interattiva (Root, Proof) e sviluppo di codice
- Storage locale (ntuple)

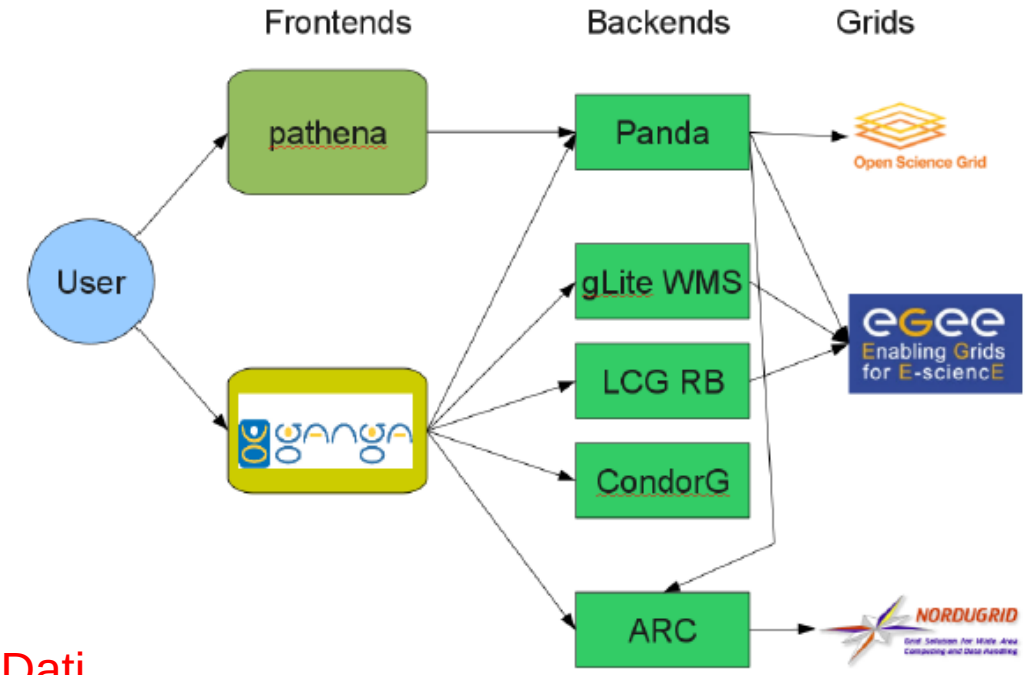
• Molti siti già attivi in tutto il mondo e soprattutto in America, molto importanti per l'attività locale

• Tipologia Tier3:
- sito pienamente "grigliato"
- Tier3 @Tier2 (tipologia più diffusa in ATLAS)
- piccola farm locale non "grigliata" per l'interattivo

I Tier3 "grigliati" hanno duplice funzionalità: GRID e LOCALE
- gli utenti beneficiano di entrambi gli ambienti
- utile per ottimizzare le attività di analisi degli utenti

in Italia
- Tier3 attivi a Genova, Roma3, Trieste/Udine. Bologna in fase di realizzazione
- task force ha portato avanti lo sviluppo del modello in base alle tipologie dei nostri centri
- Tutti i gruppi Atlas-IT hanno delle farm che svolgono funzioni di Tier3

Analisi distribuita su grid



Sistema a 3 griglie:

- OSG → Panda
- EGEE → WMS
- Nordugrid → ARC

Dati

- distribuzione organizza centralmente in base al Computing Model
- il formato dati utilizzato per l'analisi dipende dalle necessità dei gruppi (fisica o locali)

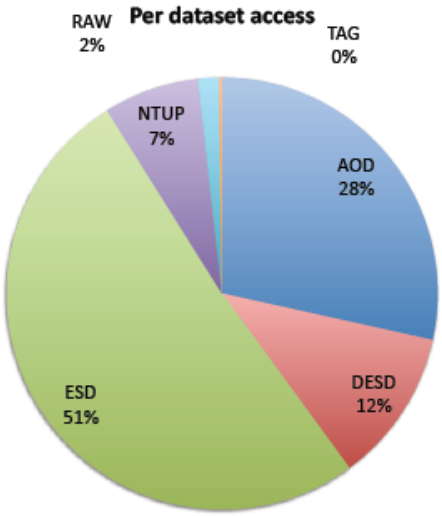
User jobs:

- Modello: ***"i job vanno dove sono i dati"***. I siti devono garantire stabilità e affidabilità
- La banda passante disponibile ha permesso di modificare il modello spostando i dati dove sono disponibili le CPU riducendo il numero di repliche sulla griglia

Scelta del Frontend e del Backend

- Il sistema più usato è pathena-Panda: user friendly, più facile reperire informazioni

Analisi distribuita e accesso ai dati



I job vanno dove ci sono i dati…
… MA ci sono molti dati mai acceduti dai jobs!

Popolarità: definita in base al numero di accessi

Formato più popolare ESD:

- necessario per molti tipi di analisi di performance e detector
- in alcuni casi è un approccio “conservativo” degli utenti (formato dati più completo)
- non può scalare con la luminosità e il numero di utenti

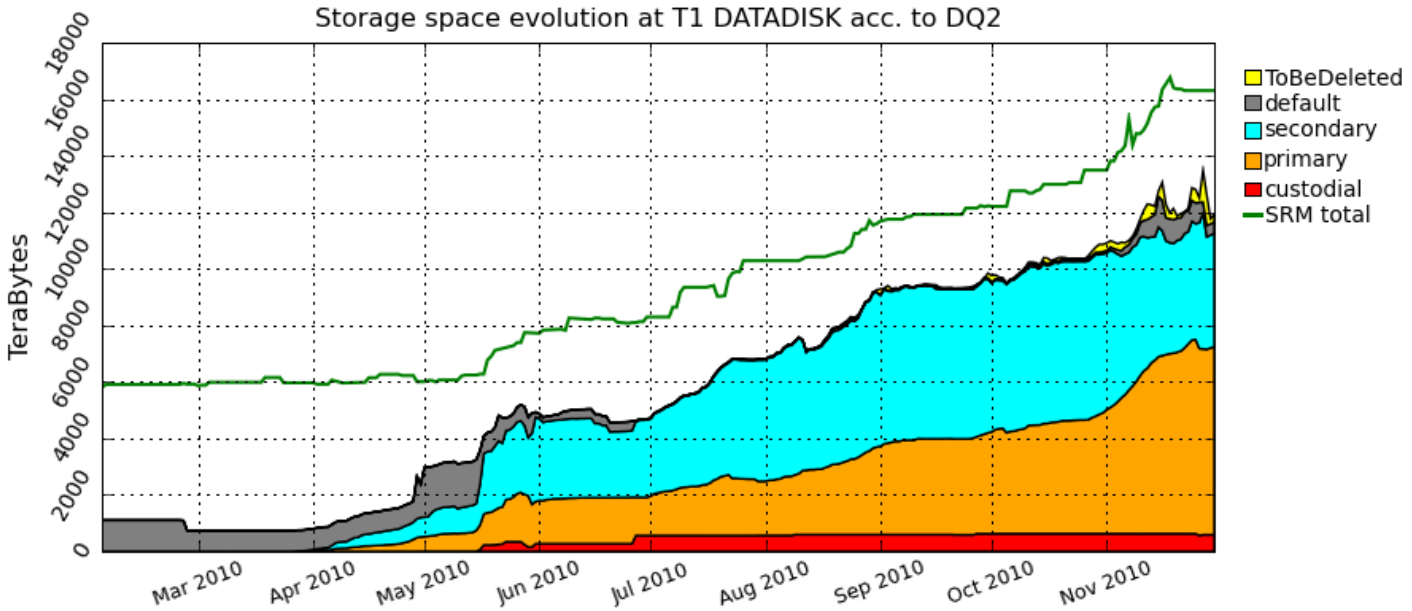
Analisi sulla popolarità ha evidenziato punti deboli del CM

- richiesta eccessiva di spazio disco
  - proliferazione del formato dei dati
  - eccesso di repliche dei dati

Cancellazione dei dati

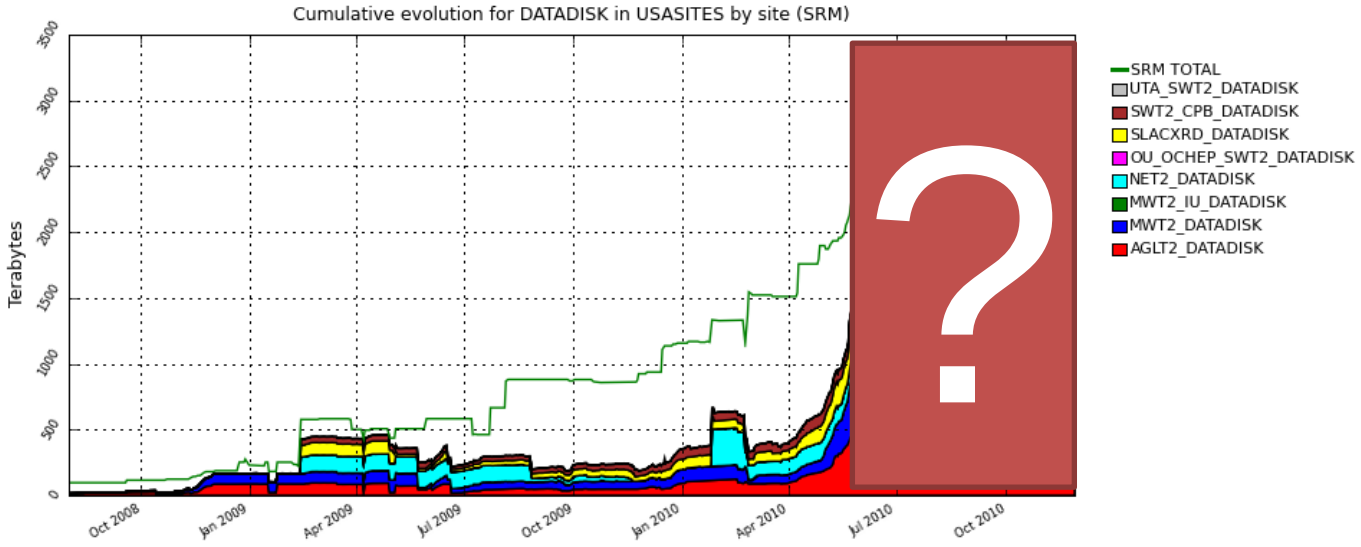
ATLAS ha sviluppato un sistema automatico di cancellazione basato sulla classificazione dei dati e la misura del numero di accessi

- *custodial data:* cancellabili solo se obsoleti (RAW, ESD o AOD prodotti nella cloud)
- *primary data:* cancellabili solo se diventano secondary (dati previsti dal CM)
- *secondary data:* solo questi possono essere cancellati se non popolari in base alla loro anzianità



Crisi del disco

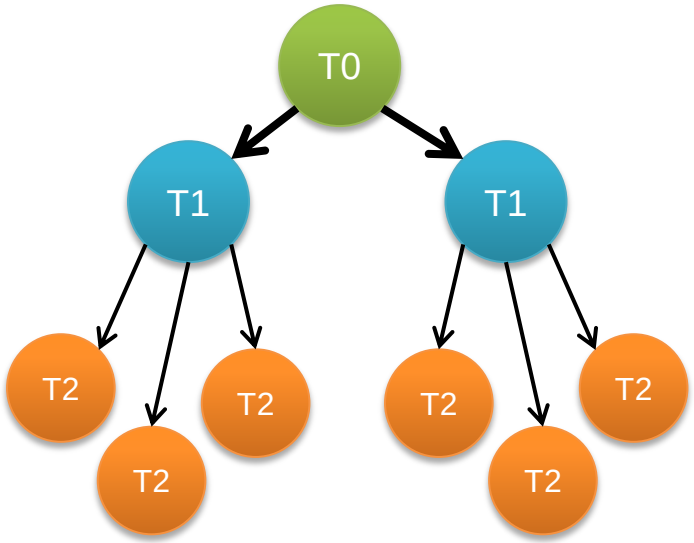
- Maggio 2010: ~ 30% dei siti era *overfull,* con ridotto spazio disco a disposizione
- Cancellare i dati meno popolari dopo essere stati replicati nei siti
  - bisogna assicurare la *custodialità* prevista dal Computing Model
  - permette di replicare sempre tutti i dati nuovi per l'analisi senza penalizzare le cloud più piccole
  - ottimizzazione significativa dello spazio disco



E' tuttavia necessario modificare anche il modello di distribuzione dei dati

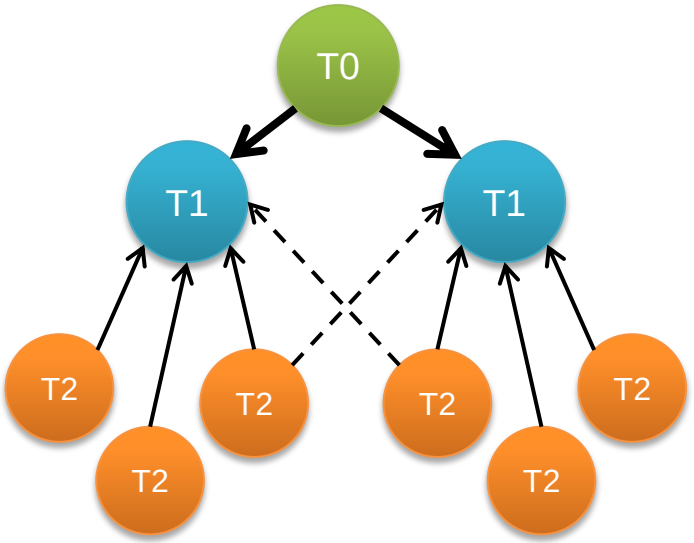
Evoluzione del CM

PUSH MODEL



Maggior parte analisi ai T2
- Inviare velocemente i dati ai T2
- Difficoltà dovuta ai molti formati e molti siti
- Numero delle copie modificato frequentemente
- T2 si riempiono velocemente
- Molti dati replicati nei T2 risultano mai usati

PULL MODEL



- Dati pre-placed solo ai Tier1
- il dato è analizzato per la prima volta al tier 1
- il dato acceduto al tier1 è replicato in un tier2 (che diventa una cache)
- gli utenti possono richiedere repliche personalizzate

Processo adiabatico e trasparente agli utenti
- nessun delay sui jobs running
- nessun cambio user workflow

PD2P: Panda Dynamic Data Placement

Modello di distribuzione dei dati basato sull'idea di considerare gli storage dei Tier2 come cache

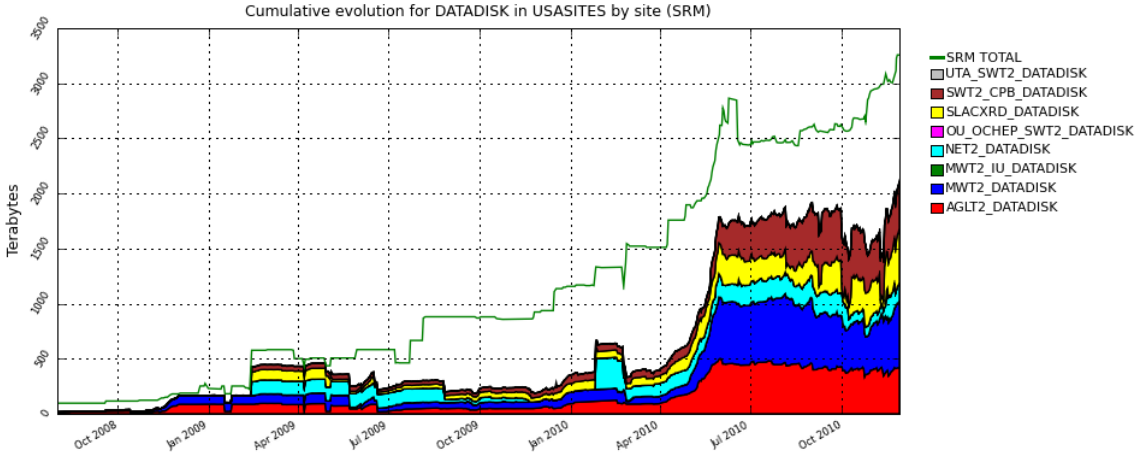
- ✓ Nel modello originario i job vanno verso i dati **pre-placed**
- ✓ nuovo modello è dinamico:
  - immutata la distribuzione dei dati nei Tier1
  - nessun dato (o solo alcuni formati) **pre-placed** nei Tier2, stop alla replica automatica
  - Panda esegue la replica on demand verso i Tier2 (se i dati non sono presenti in altri Tier2)
    - il job gira comunque solo dove sono i dati, per cui **la prima volta al Tier1** (non c'è inefficienza) e successivamente al Tier2 dove è stata eseguita e completata la replica
    - necessità di abilitare l'analisi ai Tier1
  - clean up dei Tier2 quando lo storage è pieno basato sul sistema di popolarità
- ✓ il modello, nella sua fase finale, funzionerà collegando i Tier2 con i Tier1 di ogni cloud
  - anche solo l'applicazione all'interno della singola cloud permette comunque di ottimizzare l'uso dello storage e delle CPU

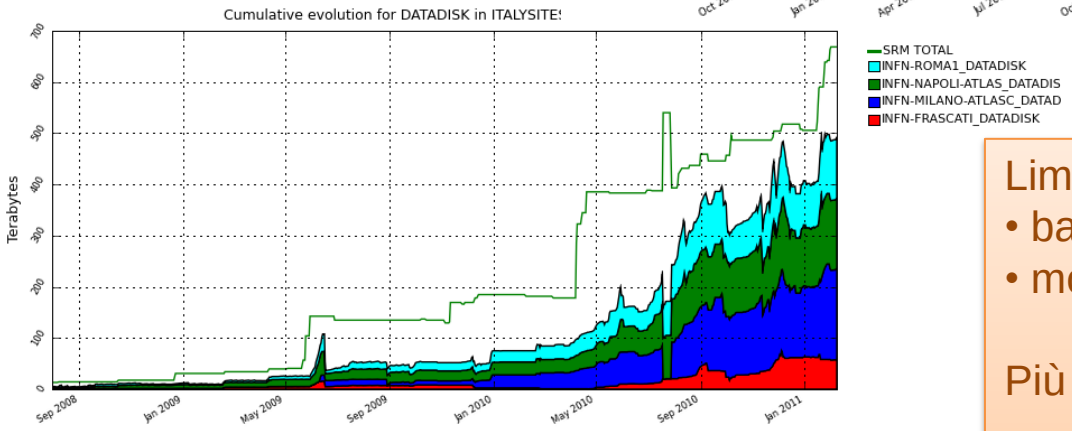
Questo modello utilizzato in tutte le cloud

- l'Italia ha partecipato alla fase di test a fine luglio appena è stata installata la prima parte del disco 2010 ed è stata abilitata l'analisi utente

PD2P: Panda Dynamic Data Placement

Con il sistema di replicazione dinamica si riduce l'occupazione dello spazio disco nei T2 (caching)



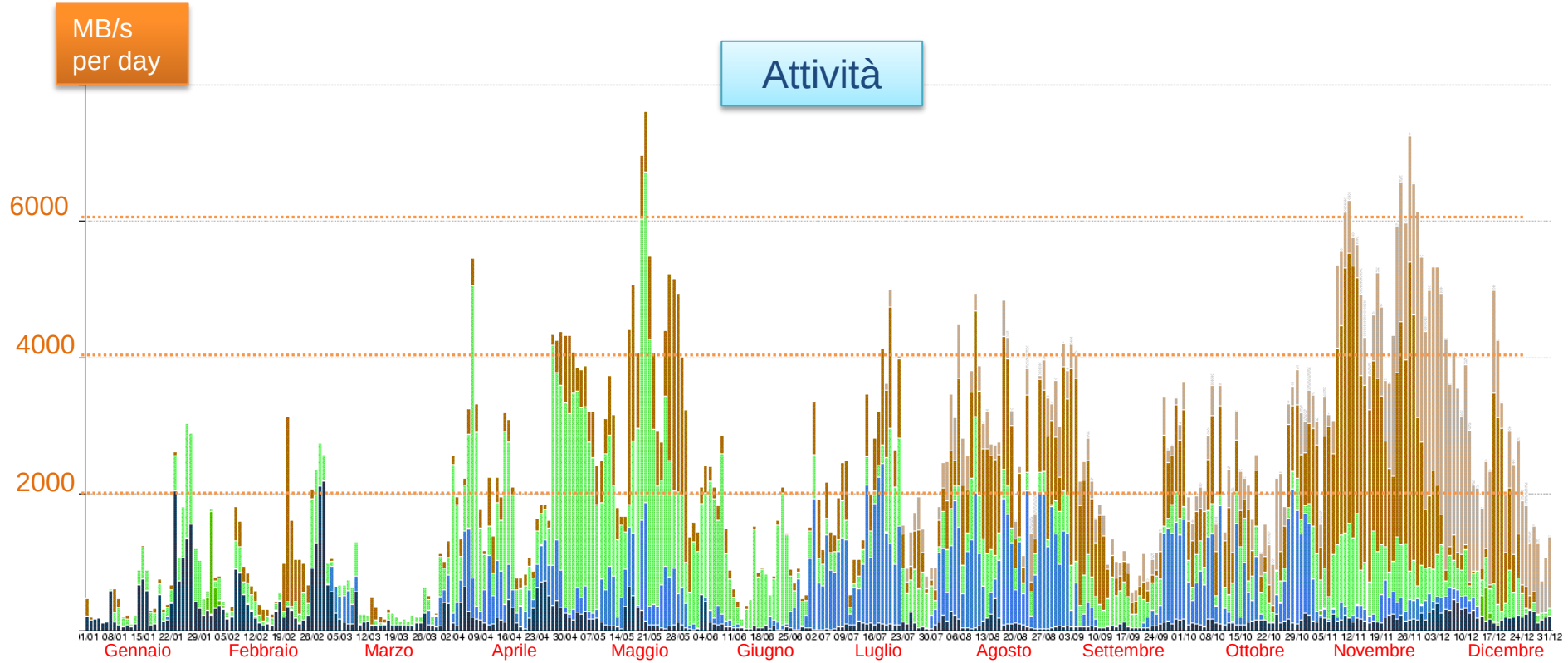


Limitazioni alla replicazione dinamica:
- bassa disponibilità di spazio disco
- molte sottoscrizioni attive

Più repliche se ci sono molti jobs in coda

Possibilità di job Re-brokering:
- un job in coda al Tier-1 può essere inviato al Tier-2 se la replica nel Tier-2 è completata prima che il job vada in esecuzione

Distribuzione totale dei dati 2010



Sottoscrizioni utenti

- grande richiesta nel primo periodo di presa dati

PANDA PD2P brokering

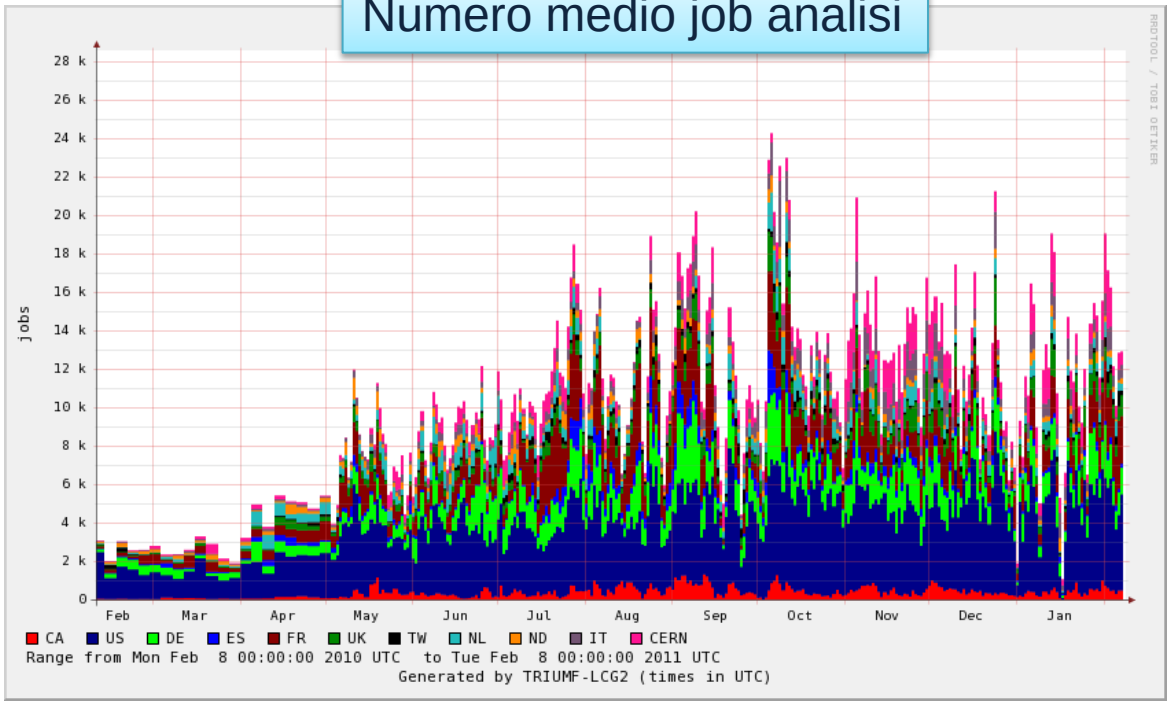
- Fase di test a luglio
- incremento dei trasferimenti
- diminuzione sottoscrizioni utenti

Analisi distribuita su grid world-wide

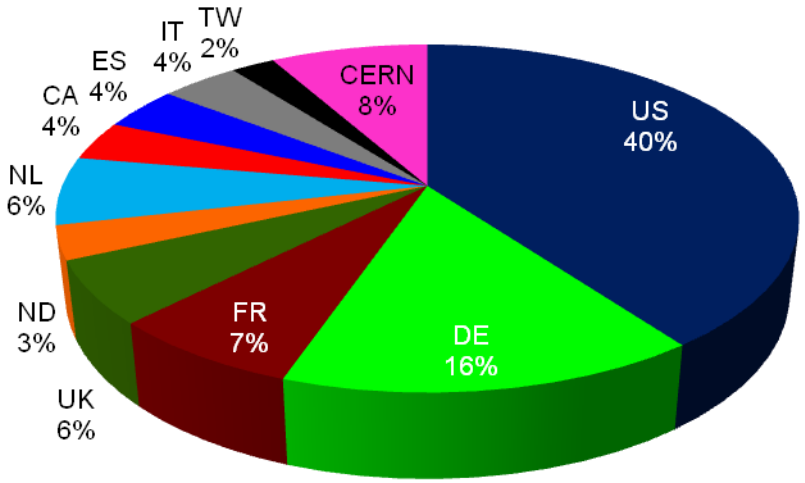
Incremento dell'analisi con l'inizio della presa dati

Poi attività stabile (previsti incrementi in prossimità di conferenze)

Numero medio job analisi



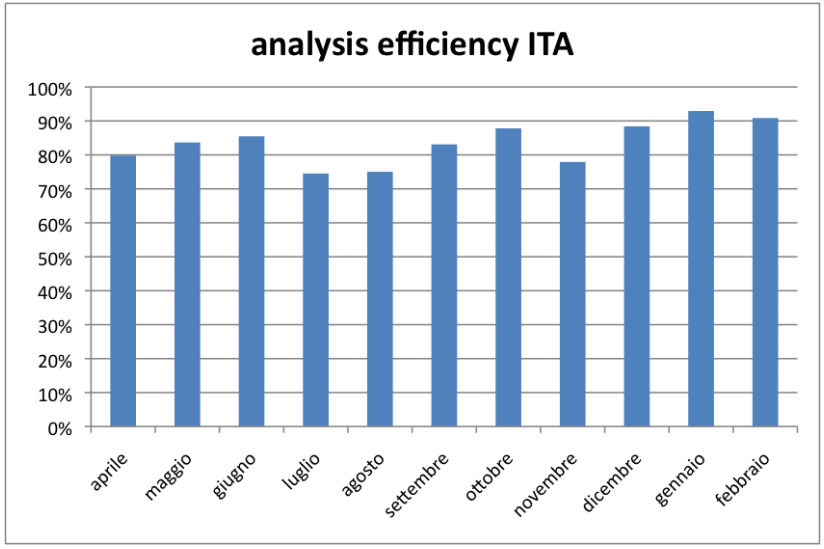
Analysis share per cloud

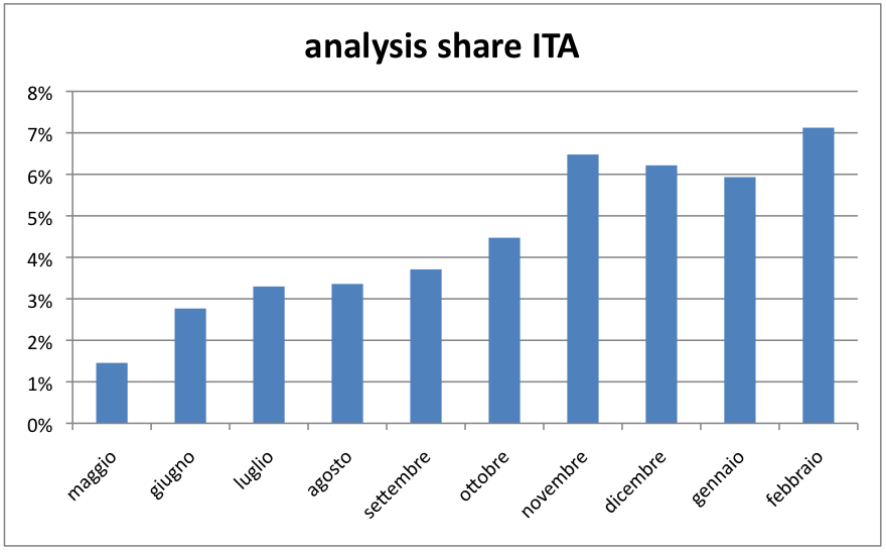


Statistiche relative solo al sistema pathena-Panda per sottomissione jobs

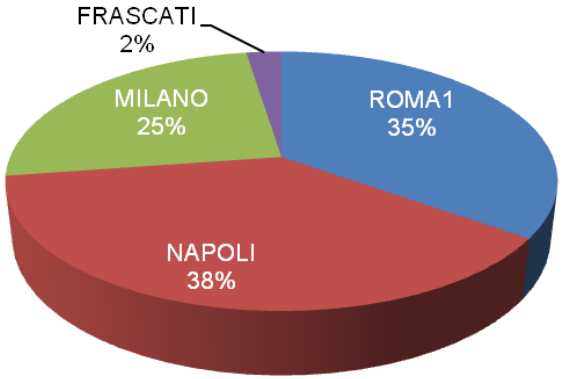
In Italia anche utenti che usano il sistema ganga-WMS (in diminuzione…)

Utiizzo risorse cloud Italiana

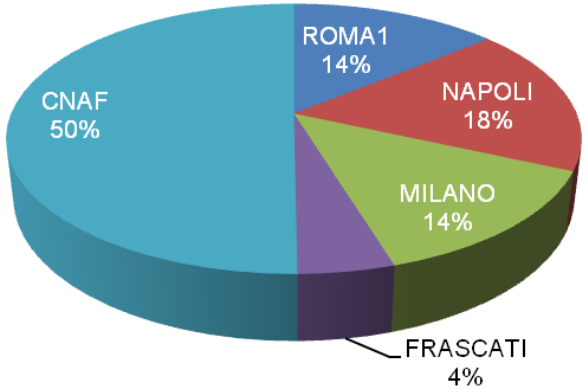




Share analisi Gennaio-giugno



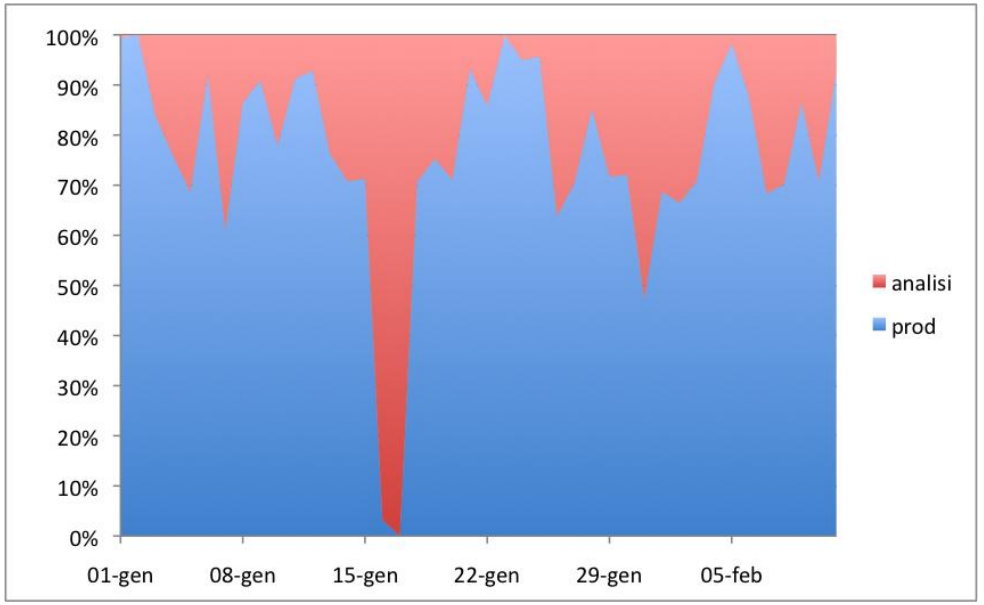
Share analisi Luglio-Dicembre

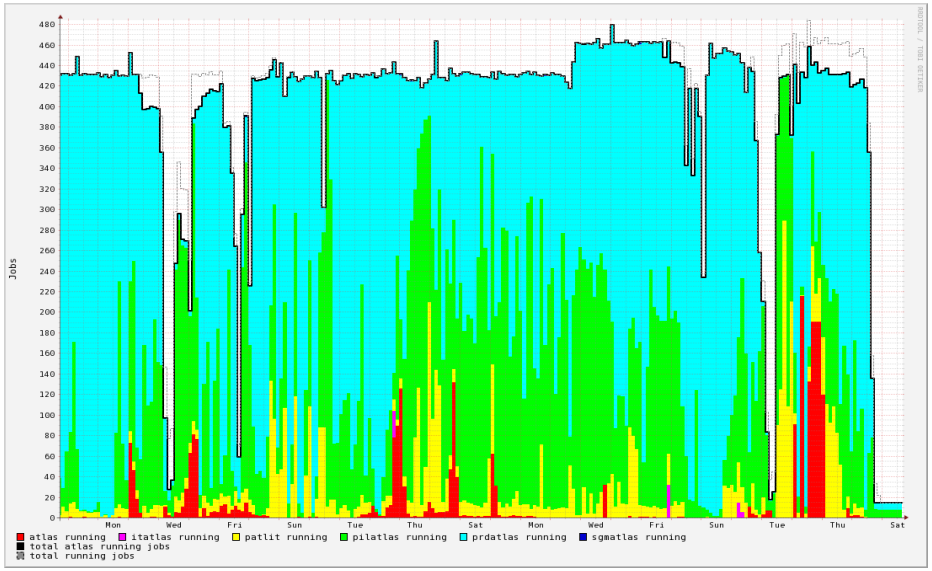


Utiizzo risorse: confronto produzione analisi

Job sharing al Cnaf
Rapporto WCT

78.5% produzione
21.5% analisi





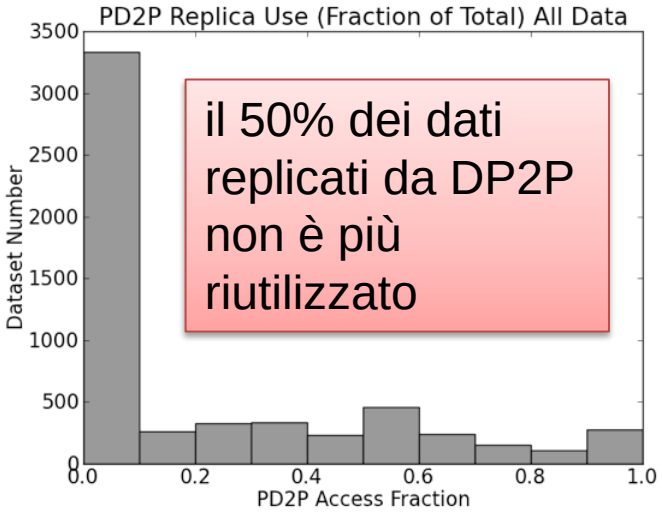
Job running su un Tier2 nell'ultimo mese codice colori:

- Produzione
- Analisi WMS
- Analisi Panda
- Analisi Panda ruolo italiano

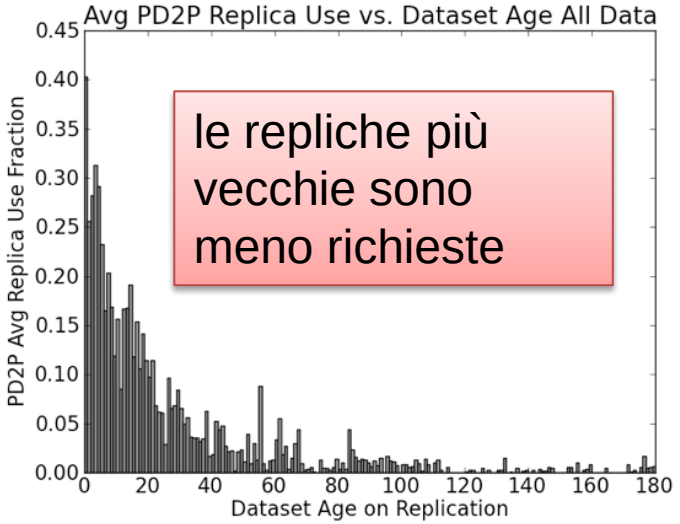
(Gli italiani vengono mappati sia su panda che su panda/it)

Sviluppi futuri

La replicazione dinamica dei dati ha permesso di ottimizzare l'uso dello storage
Nel 2011/2012 previsto incremento del volume dei dati
E' necessario raffinare l'algoritmo di replica
- Al momento replicati dinamicamente solo i dati "popolari"
- I dati replicati dinamicamente sono secondari (caching)



il 50% dei dati replicati da DP2P non è più riutilizzato

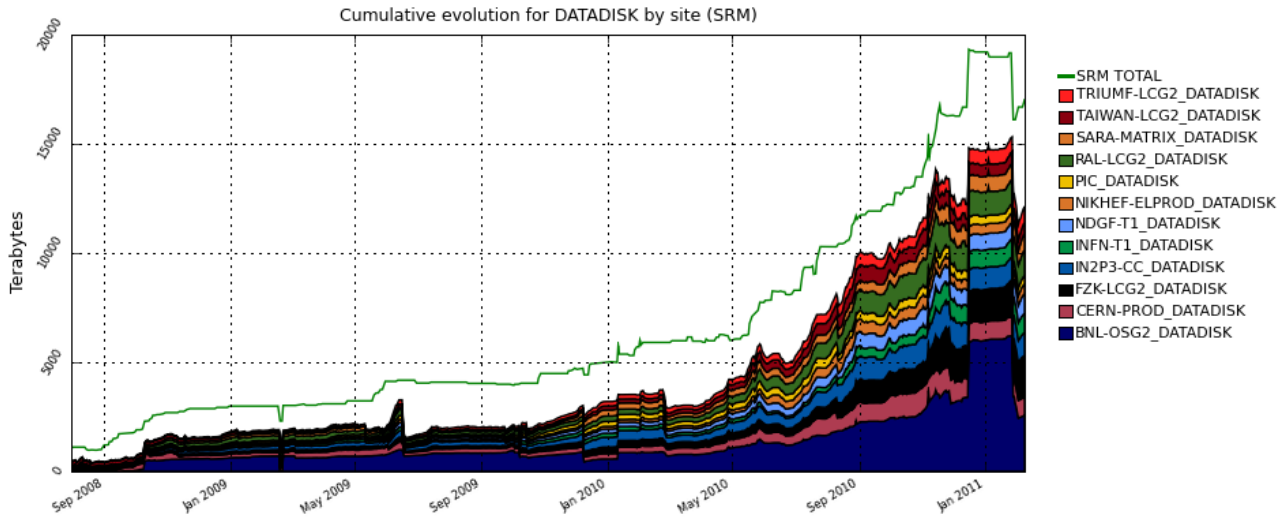


le repliche più vecchie sono meno richieste

Si sta studiando il modo migliore di replicare i dati combinando il numero di accessi e l'anzianità dei dati

Sviluppi futuri

Distribuzione dinamica dei dati nei Tier-1s
- limitare numero di copie ESD primarie
- inviare ulteriori copie ESD on-demand con DP2P



Analisi nei Tier-1
- Ottimizzare le infrastrutture di rete/storage/farming
- Task force al CNAF per studiare la soluzione migliore

Conclusioni

Il Computing Model di Atlas si è rivelato stabile e affidabile

- Distribuzione dei dati solida ed efficiente
- Produzione e analisi sfruttano appieno le risorse

Principale punto debole: eccessiva quantità di dati replicati

Sono state necessarie alcune modifiche per ottimizzare l'uso delle risorse

- distribuzione dei dati e successiva cancellazione effettuate in base alla popolarità dei dati
- la popolarità è determinata dagli utenti
- modifiche sono state adiabatiche e trasparenti

Stretto legame tra attività degli utenti e computing model:

Il modello si adatta alle esigenze degli utenti

Un'ultima considerazione

Opinione personale sull'uso della grid da parte dei fisici

- iniziale diffidenza e difficoltà nell'uso della grid
- con l'aumento della quantità di dati, gli utenti si sono accorti della necessità di usare la grid

- adesso sempre più utenti utilizzano i tool software e middleware, aiutando nel debug e facendo selezione naturale degli strumenti migliori (vedi Panda vs WMS)

- il computing deve offrire stabilità dei siti e disponibilità delle risorse
- necessità di rafforzare il supporto e la collaborazione tra le comunità di computing e di fisica
- I tier3 possono essere degli ambienti ideali per coniugare le attività di fisica e di computing