Computing model and data handling ATLAS-CMS

Simone CAMPANA
(*ATLAS CERN*) ,

Daniele SPIGA
(*CMS CERN - Perugia*)

Quinto Workshop italiano sulla fisica p-p ad LHC

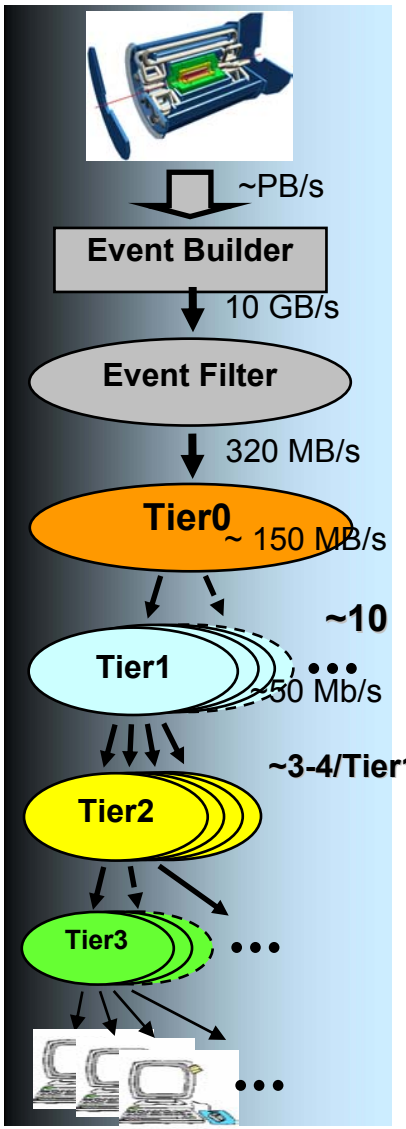
Perugia, 31 Gennaio 2008

Il Computing Model di ATLAS e CMS

Modello di calcolo gerarchico multi-Tier.

Architettura basata su una infrastruttura distribuita di risorse, servizi e strumenti Grid.

ATLAS



Tier-0 (CERN)

- Archiviazione dei RAW data e distribuzione ai Tier1
- Prompt Reconstruction e 1st pass calibration
- Distribuzione output ricostruzione ai Tier-1: ESD, AOD e TAG

Tier-1 (10)

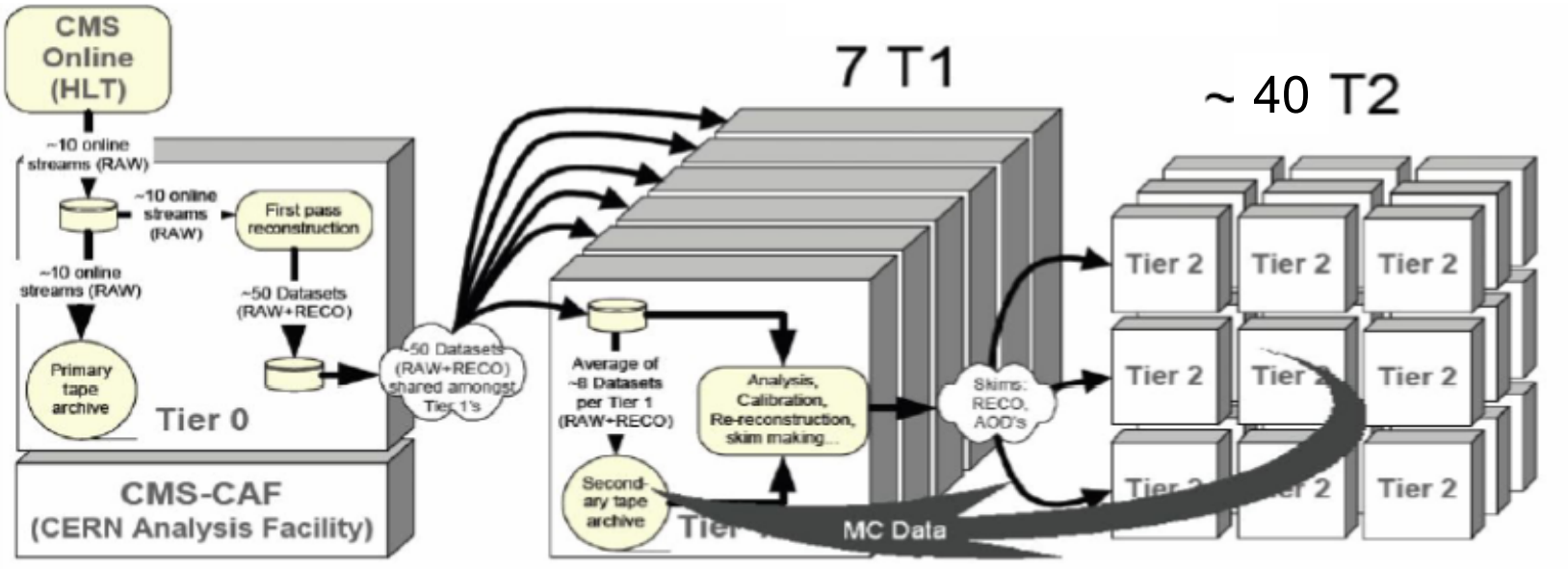
- Archiviazione di un subset di RAW data
- Copia degli ESD di un altro (altri) Tier-1
- Reprocessing dei propri RAW data
- Distribuzione AOD ai Tier-2
- Archivio dati MC prodotti nei Tier-2

Tier-2

- Simulazione Monte Carlo
- Analisi (Single User e Group Analysis)

Modello a cloud: ad ogni Tier-1 sono associati alcuni Tier-2 (spesso in base a considerazioni geografiche).

CMS



- **T0**
  - Ricezione dati dal DAQ
  - Prompt reconstruction
  - Archiviazione RAW data e distribuzione ai T1's

- **CAF** (CERN Anlalysis Facility for CMS)
  - Accesso ai RAW dataset
  - Finalizzata alle attività "latency-critical"
    (diagnostica del rivelatore, trigger performances, definizione costanti Al/Ca, Hot analysis)

CMS

- **T1**

Riporocessamento dati (later pass reco,AOD, Skimming)

Archiviazione Dati (Reali+MC custodial)

Distribuzione dati (serve dati ai T2 per l'analisi)

- **T2**

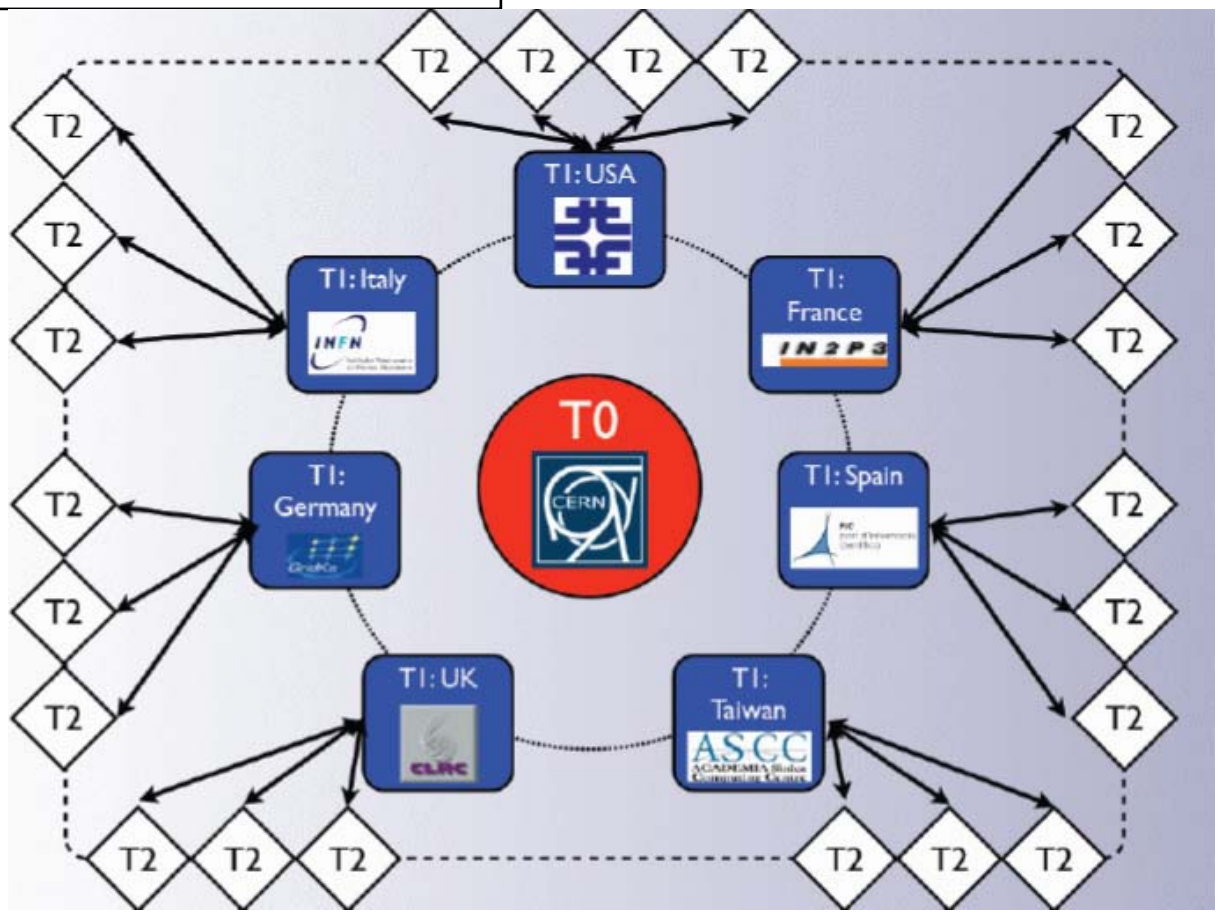
User Analysis

Produzione eventi MC

~20% delle risorse

Sono allocate al CERN, 40% ai T1 e 40% ai T2.

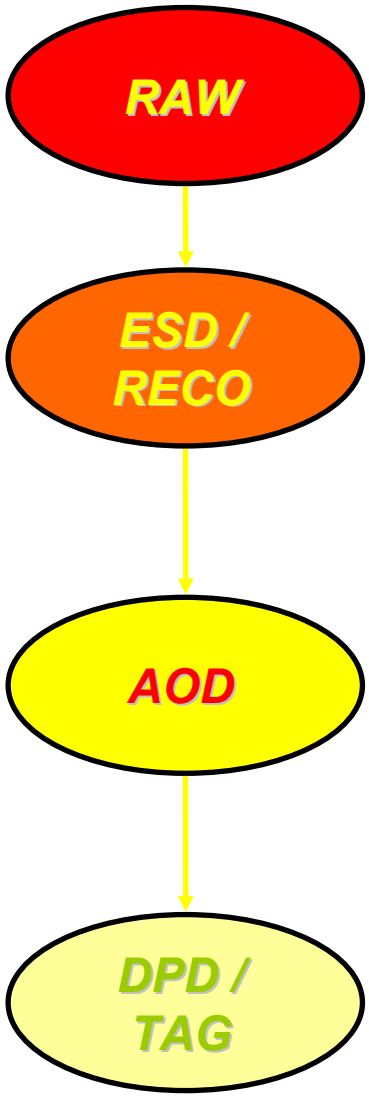
Modello basato sullo sviluppo di strumenti per rendere transparente l'accesso alle risorse.



Data Model ATLAS-CMS

Nelle varie fasi di ricostruzione e analisi ATLAS e CMS utilizzano diversi formati di dati:

ATLAS 1.6 MB
CMS 1.5 MB



target 500 KB

ATLAS 1,3/1,5 MB
CMS ~700 kB

Target 100 kB

ATLAS 110/210 kB
CMS ~200 kB

ATLAS/CMS 10kB

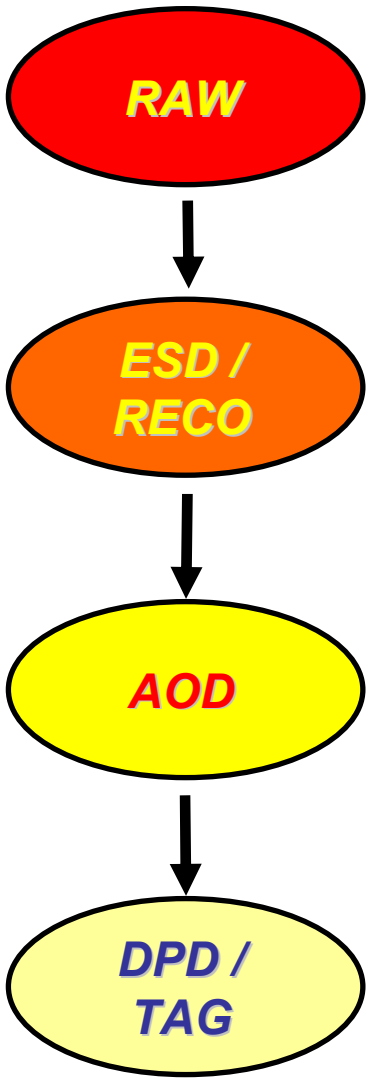
Raw Data dati in output dal sistema di trigger in formato byte-stream

Event Summary Data output della ricostruzione (tracce e hit, celle e cluster nei calorimetri, combined reconstruction objects etc...). Per calibrazione, allineamento, refitting ...

Analysis Object Data: rappresentazione ridotta degli eventi per l'analisi: oggetti "fisici" ricostruiti (elettroni, muoni, jet, missing Et ...)

Derived Physics Data: informazioni ridotte per analisi specifiche in ROOT.

Distribuzione Dei Dati ATLAS-CMS



- Dati originali al Tier0
- Replica completa nell'insieme dei Tier-1

- Gli ESD vengono esportati associati ai RAW in 2 copie nell'insieme dei T1
Frazioni a richiesta nei Tier-2

- I RECO vengono esportati associati ai RAW in 1 copia nell'insieme dei T1

 Una copia completa in ogni Tier-1
 Replicati parzialmente nei Tier-2 (~1/3 – 1/4 in ciascun Tier-2):

Una copia completa in ogni T1 e replicati parzialemte ai T2

- DPD dei gruppi di analisi prodotti nei Tier1
- repliche nei Tier-2 e Tier-3
- In fase di definizione le procedure di produzione

Organizzazione e trasferimento Dati in ATLAS

- Il sistema di Distributed Data Management (DDM) implementa la funzionalita' richieste dal computing model per il trasferimento dati
  - Disctibuzione di raw e reconstructed data (reali e simulati) tra i vari Tier
- I dati di ATLAS sono organizzati in datasets
  - Cataloghi di dataset centrali, suddivisi in vari DB
    - Dataset Repository, Dataset Content Catalog, Dataset Location Catalog, Dataset Subscription Catalog
    - Cataloghi di file distribuiti (locali)
      - mapping nome logico nome fisico: LFC (LCG File Catalog) al Tier1
- Trasferimento dei file attraverso il Sistema di Sottoscrizione:
  - T0 T1 e T1 T1 (inter clouds)
  - T1 T2 e T2 T1 (intra cloud)

Attività trasferimento dati

- **Functional Test:** simulazione del data flow previsto dal CM a basso rate
  - Trasferimenti T0 → T1 e di seguito T1 → T2 della cloud secondo lo share previsto dal Computing Model
  - Trasferimenti T1 → T1 dei dati riprocessati
  - Studio dell'efficienza dei trasferimenti in termini di numero di dataset replicati correttamente e velocità di arrivo dei file, numero di retry
    - Test sporadici. A breve test regolari e risultati riportati nel site Reliabiliry e Availability di WLCG
- **T0 Throughput exercise:** mantenere con stabilità i rate di trasfermento tra il CERN e le clouds previsti dal Computing Model:
  - T0 → ©T1 = 1 GB/s
    - T0 → CNAF = 100 MB/s, 2/3 su disco e 1/3 su nastro
- Run di cosmici M5
  - Trasferimento dei dati (RAW e ESD) al CNAF e nei Tier2 secondo le percentuali previste dal Computing Model

Functional Test Ottobre 2007

Trasferimenti Tier-1 □ Tier-1

	ASGC	BNL	CNAF	FZK	LYON	NGDF	PIC	RAL	SARA	TRIUMF
ASGC										
BNL										
CNAF										
FZK										
LYON										
NDGF										
PIC										
RAL										
SARA										
TRIUMF										

Trasferimenti CERN □ Tier-1s

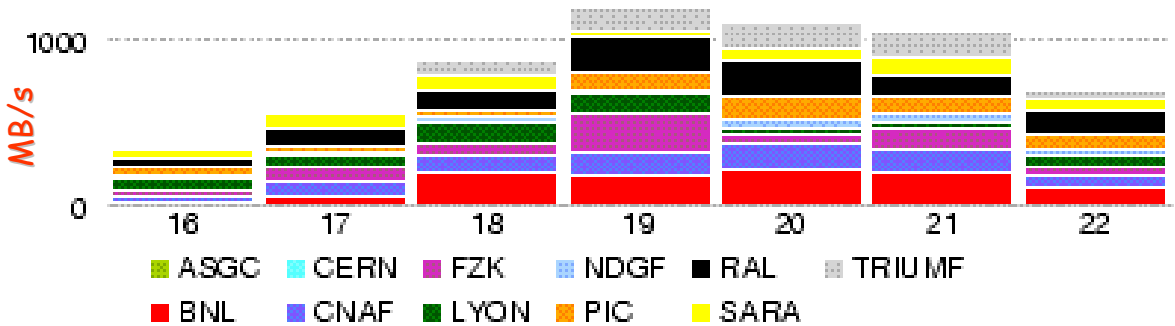
	ASGC	BNL	CNAF	FZK	LYON	NDGF	PIC	RAL	SARA	TRIUMF
CERN										

100% , 90+% , 50%, less than 50%, of data transferred within 24h

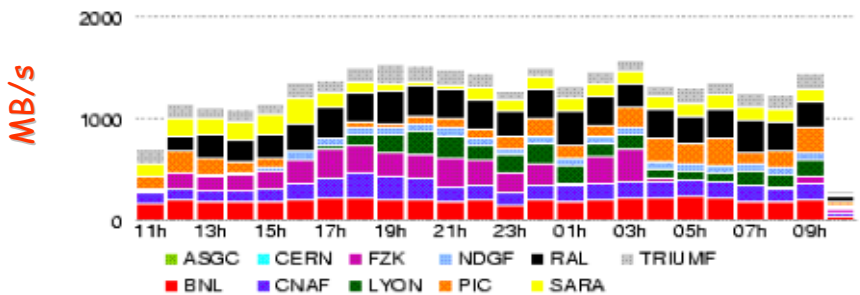
Throughput exercise Ottobre 2007

- Throughput al 100% MoU
  - Come se la macchina operasse 24h/day ~ 1 GB/sec
- Operazioni completamente automatizzate senza intervento
- Corretto share tra dati da inviare su tape (tipo RAW) e su disco (tipo AOD e ESD)

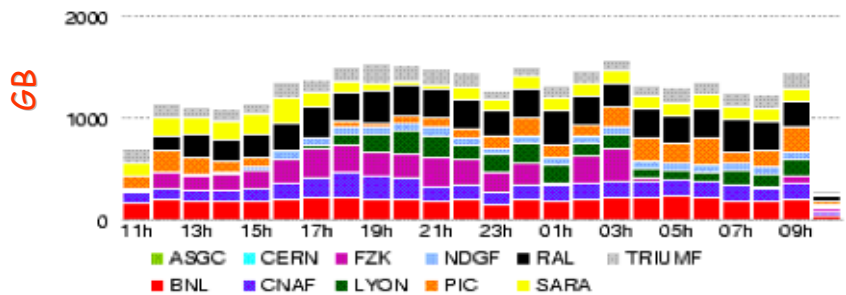
Obiettivo raggiunto !
Rate di ~ 1.2 GB/sec per un periodo prolungato con un set incompleto di Tier-1



Throughput MB/s



Data transferred GB



Organizzazione e trasferimento Dati in CMS

- **I Dati sono organizzati in files** ($O(10^6)$files/year)
- **I files sono raccolti in blocks per ragioni di data management** (10^3 Filebocks/year)
- **I blocks sono raggruppati in Datasets 0.1->1TB** (physics driven definition)

La catalogazione dei dati avviene attraverso il Dataset Bookkeeping System (DBS):
fornisce le seguenti funzionalità:

- Data definition: dataset definition, runs, provenance, etc..
- Data discovery: quali dati esistono e come sono organizzati (files/fileblocks)
- Data location: dove sono i dati

- (per dati privati e pubblici)

Physics Experiment Data Export (PhEDEx)

- Gestisce in maniera affidabile il trasferimento dei dati tra i siti, usando i servizi Grid di file and storage management.
- Tecnicamente basato su agenti software che runnano autonomamente nei siti scambiando informazioni attraverso un databse centrale (Tranfer Management DB TMDB).
- I dati ai T2 vengono richiesti dai gruppi di fisica o singoli (chiunque può sottomettere una richiesta)

https://cmsdoc.cern.ch:8443/cms/aprom/phedex/prod/Request::Create?view=global

CMS LoadTest

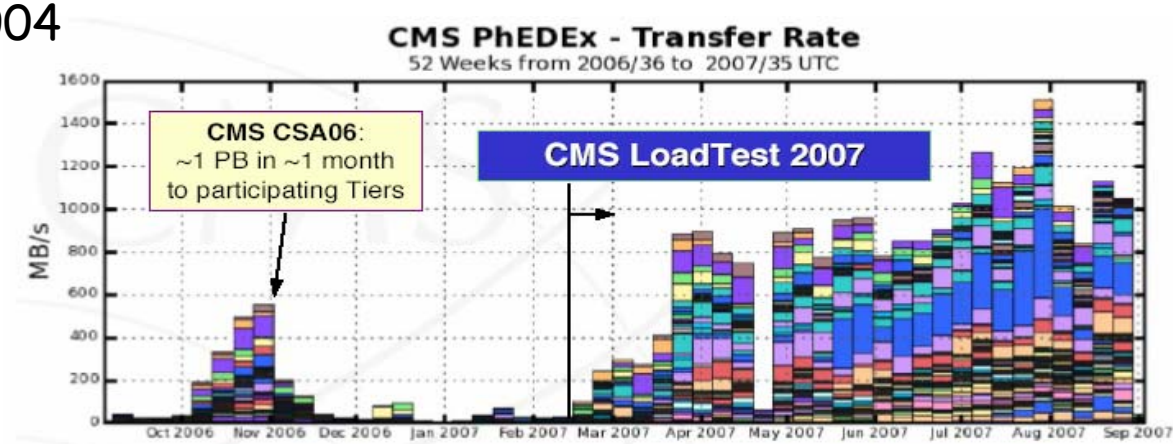
- I siti sono pronti per il trasferimento dei dati?
- Il sistema di trasferimento dati di CMS è adeguato?

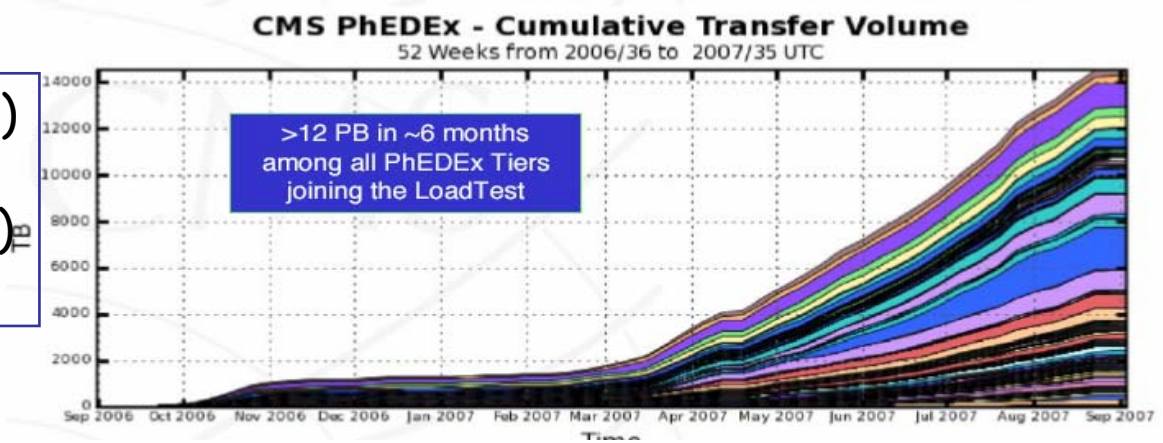
✓ PhEDEx è in produzione dal 2004

✓ Molta esperienza fin dal 2004 (CMS-specific e WLCG-Service challenges)

✓ PhEDEx in continuo test di trasferimento..

Dal 2007 è iniziato LoadTest…

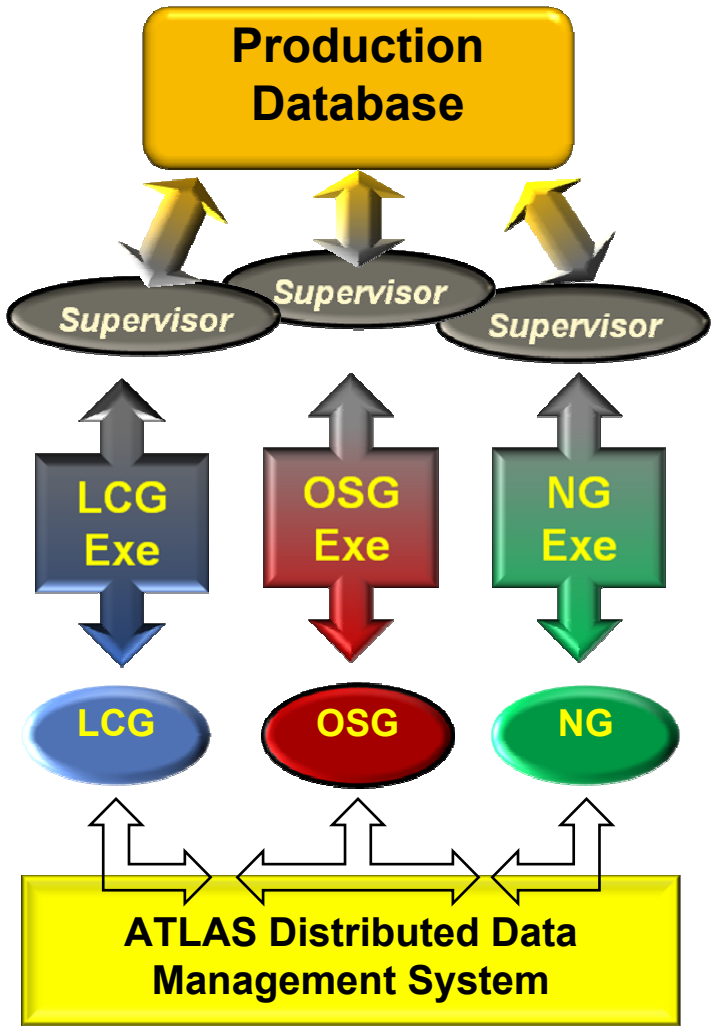




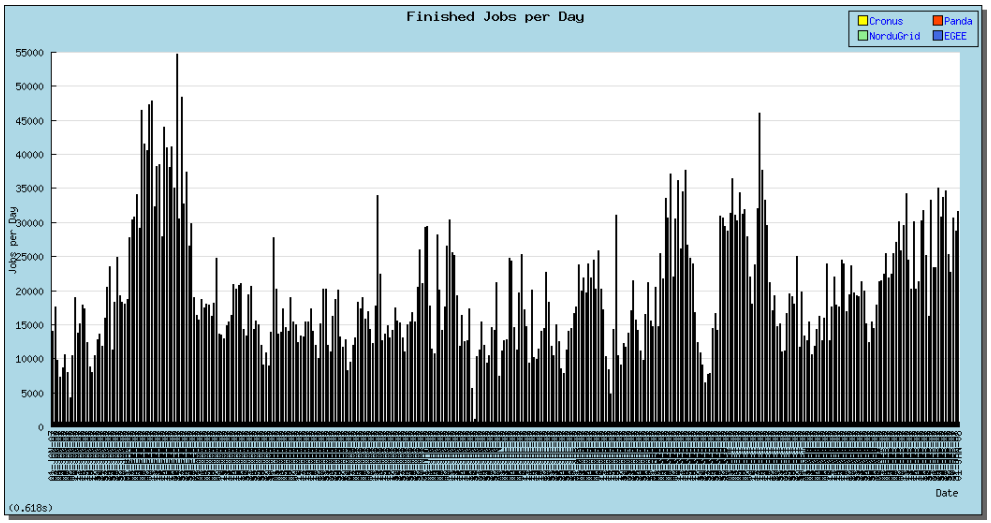
- CERN -T1, T1-CERN (14 links)
- T1-T1 crosslinks (42 links)
- T1 to T2 downlinks (360 links)
- T2 to T1 uplinks (360 links)

Sistema di Produzione Monte Carlo in ATLAS

- Risorse di ATLAS distribuite su 3 Grid
  - EGEE ~3000 CPU, ~360 TB disk
  - NDGF ~500 CPU, ~60 TB disk
  - OSG ~2000 CPU, ~390 TB disk
- Il sistema di Produzione (Prodsys) di ATLAS e' in grado di gestire job di produzione sulle 3 griglie
  - Jobs definiti in un database centrale (ProdDB).
  - L' agente "Supervisor" la sottomissione, monitoring e validazione dei job sulle 3 griglie
    - Supervisor e' Grid-Neutral
    - L' interfaccia con il grid middleware e' implementata dagli "Executor"
  - Il movimento dei dati avviene tramite DDM.
- Evoluzione del Sistema di Produzione verso un unico executor per OSG e EGEE
  - Due scelte di base: la tecnologia dei job pilota di Panda e il Local File Catalog di LCG

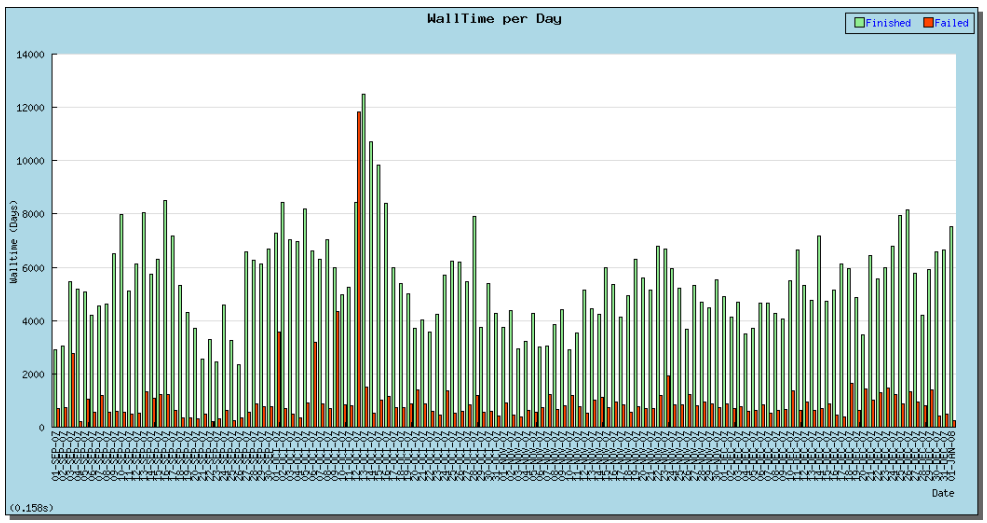


Produzione Monte Carlo 2007 in ATLAS

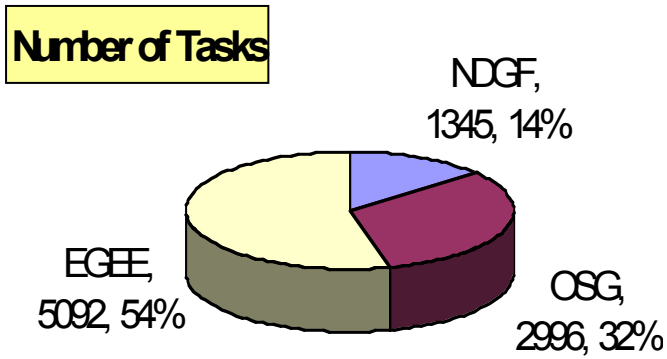


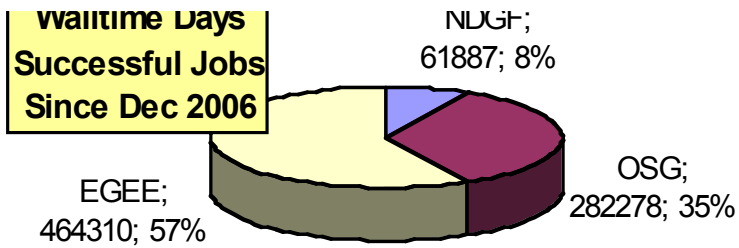
- 50M+ eventi simulati per CSC12
  - >300 TB di MC data su disco
  - Job scaling come previsto nel 2005
- Target: 120M nuovi eventi entro aprile 2008 (FDR2)
- Attualmente 1 M ev/giorno, rate limitato dalle necessità e disponibilità di storage

- Ragionevole efficienza in termini di WallClockTime (>90% in media)
- Problemi residui:
  - Data Management (~40%)
  - ATLAS application (~40%)
  - Configurazione siti (~20%)
- Efficienza in termini di # jobs ancora non soddisfacente (<70%)
  - Sistema "Pilot Jobs" dovrebbe aiutare in questo



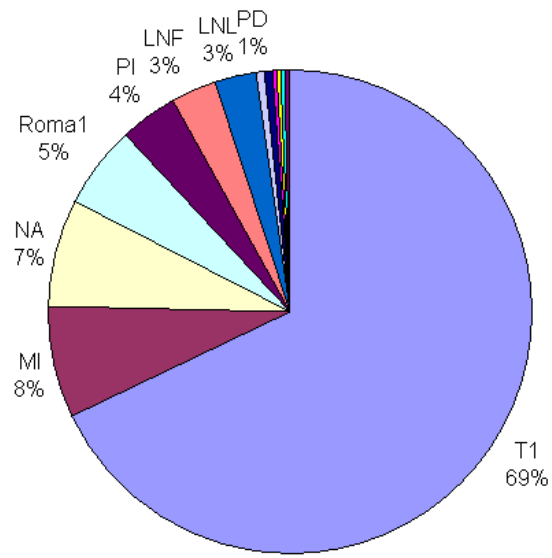
Overview della Produzione MC 2007





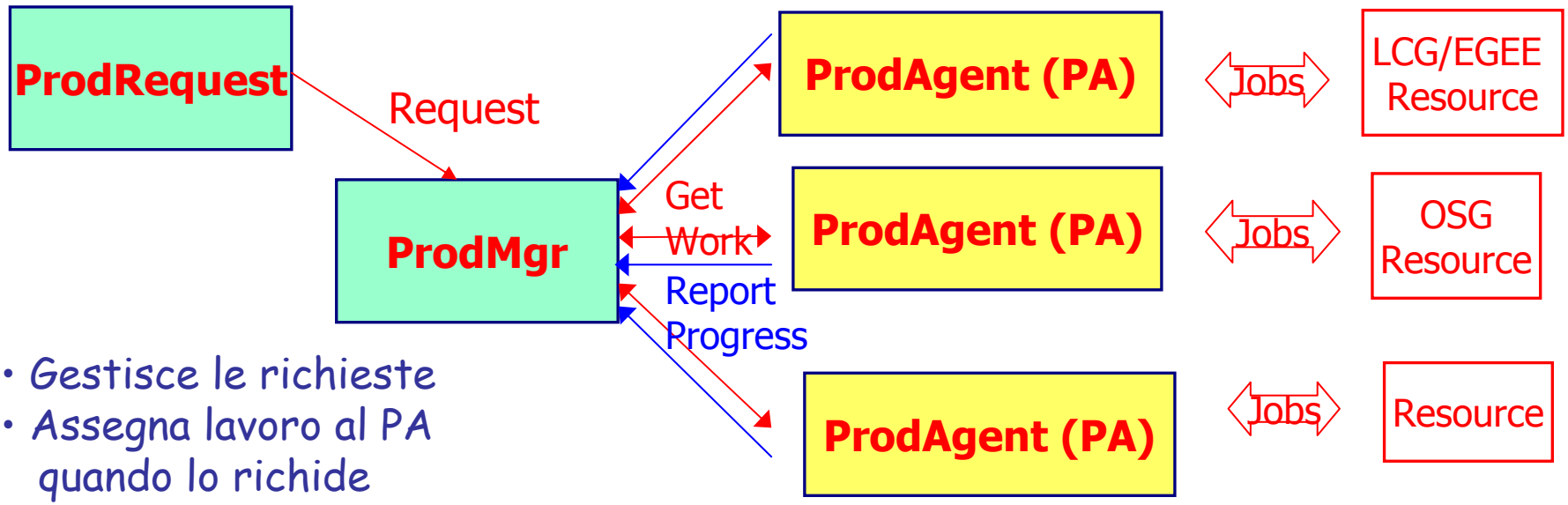
Produzione CSC nei siti INFN

- ATLAS sfrutta risorse tutte le risorse di CPU accessibili
  - CNAF (T1)
  - T2 (Milano, Roma1 e Napoli) e Proto-T2 (Frascati) di ATLAS
  - T2 di altri esperimenti
  - Risorse condivise di INFNGRID



Sistema di Produzione Montecarlo in CMS

Strumento unico per la produzione dati MC, Skimming, riprocessamento e ricostruzione al Tier 0

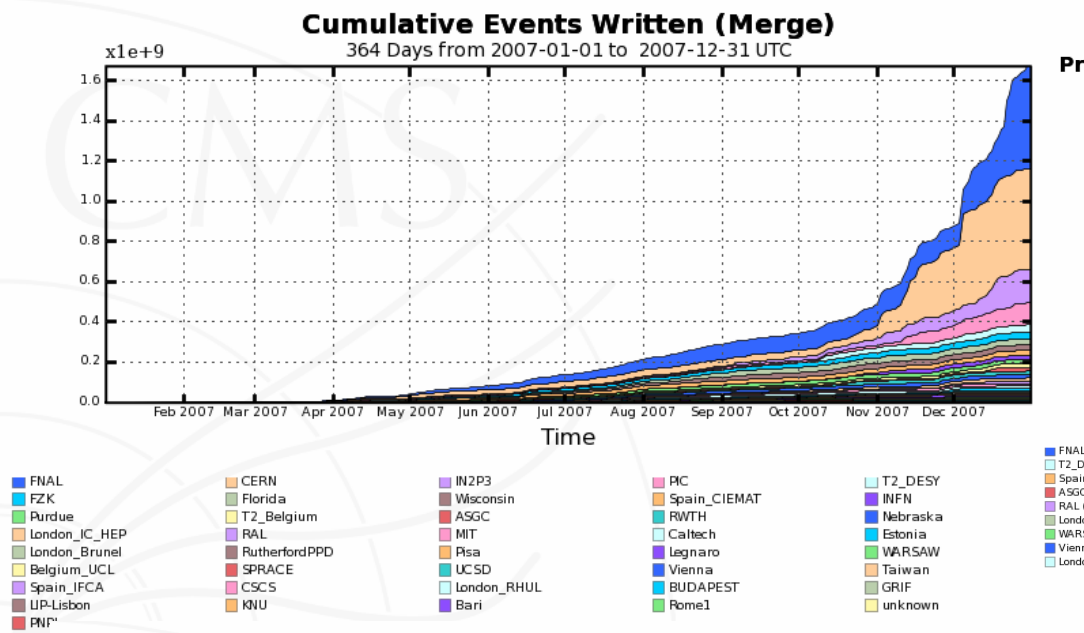


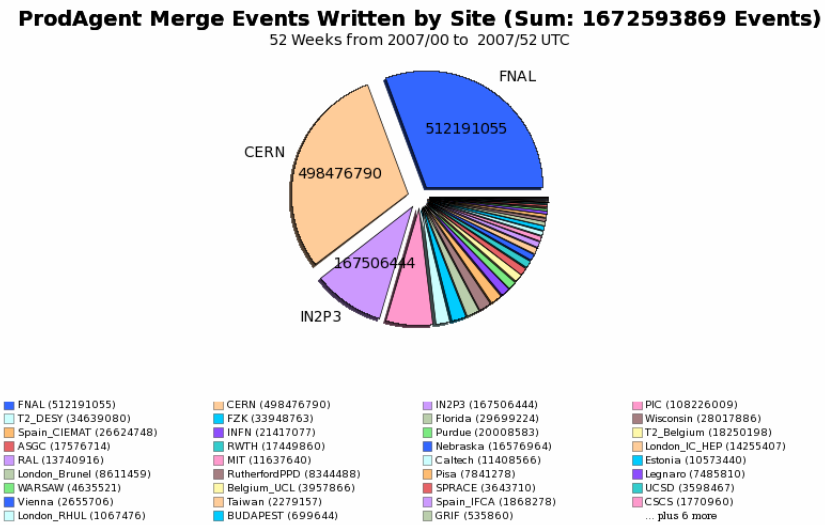
- Gestisce le richieste
- Assegna lavoro al PA quando lo richide
- Traccia il completamento del task

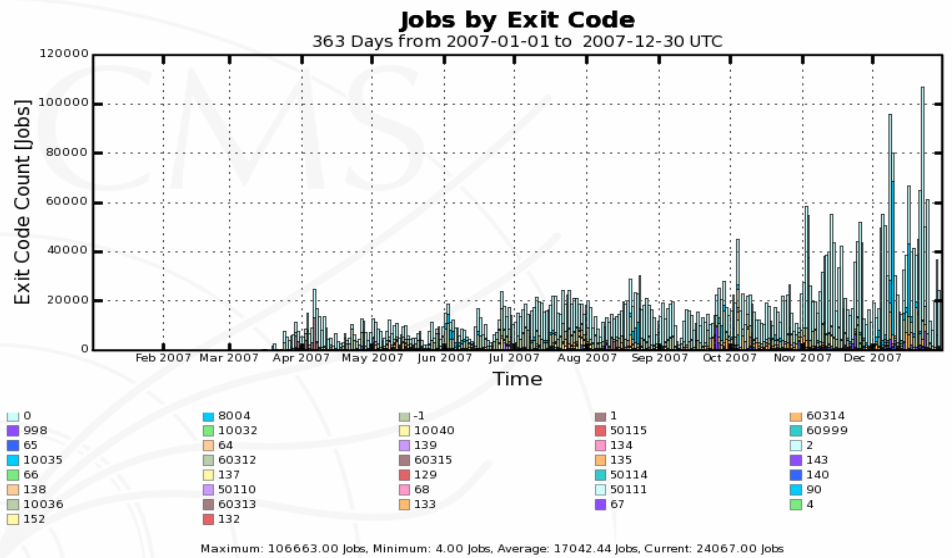
- Richiede lavoro
- Converte lavoro in jobs
- Crea, sottomette, traccia jobs
- Gestisce il merge, errori, risottomette, etc..

- Supporto per differenti Grid/batch Middleware

Produzione Monte Carlo 2007 a CMS







- Raggiunto un rate > 20Kjobs/day
- ~50% della produzione su OSG
- ~75% efficienza comprese Grid failures

Analisi Distribuita in ATLAS

- L' Analysis Model si basa sul principio che i job devono girare dove risiedono i dati
  - Full set di AOD e DPD presenti su disco in tutti i T1 e 1/3 dei T2
    - 2 copie del full set di ESD su disco, suddivise tra i T1
      - Frazione minima anche ad alcuni T2
    - RAW data solo su nastro, 10% anche su disco suddivisa tra vari T1
  - Selezione degli eventi da TAG
    - Supporta backward navigation agli eventi (RAW, ESD, AOD)
    - Procedura di selezione molto veloce
  - Determinazione dei siti dove i dati sono memorizzati
    - Invio in questi siti (tramite i tool di Analisi Distribuita) dei jobs e recupero degli output: DPD e logfiles
- Tool di Analisi Distribuita (job submission):
  - GANGA in EGEE
  - PAthena in OSG

GANGA

- Framework per configurazione, sottomissione, Monitoring dei Job
- Le applicazioni possono essere sottomesse, **in modo assolutamente trasparente**, attraverso vari tipi di risorse
  - **Macchine locali (test, debugging dell' applicazione)**
  - **Sistemi Batch (LSF, PBS, CONDOR)**
  - **Grid**
- Il giudizio obiettivo sulle performance di GANGA è complicata dalla forte dipendenza dal funzionamento di altri sistemi:
  - DDM: **i dataset non sono completamente replicati nei siti**
    - **Alta efficienza se i dataset sono completi**
    - **La replica non completa è il problema più serio**
  - Configurazione dei siti: **i job falliscono per problemi locali**
    - **comprende problemi (anche temporanei) di hardware**

Analisi Distribuita: Pathena

- **Package integrato in Athena**
  - si usa in modo del tutto simile, molto facile da usare !
- **Stesso workload usato per le produzioni ATLAS: PANDA**
  - Output registrato su catalogo in DQ2
  - bookeeping, monitoring, status, facile risottomissione dei job , notifica via email
  - Interfaccia Web molto user friendly.
- **Giudizio degli utenti:**
  - "Not much comments here. PAthena is the closest thing to the user dream"
    - Bisogna solo scegliere il nome del dataset di input e il dataset di output
    - Job, in media, hanno successo con un' efficienza molto alta
- **Pero'...**
  - PAthena funziona perfettamente a BNL dove sono replicati (quasi) totalemente tutti gli AOD (BNL e' un super Tier-1)
    - E' veramente un sistema di analisi distribuita?
    - E' altrettanto possibile runnare PAthena in siti con distribuzioni incomplete dei dati?
    - E' un modello scalabile con il numero degli utenti?
  - Issue di sicurezza da chiarire

- Fino a luglio (tabella) problemi di replicazione dati in IT
  - DDM e problemi allo storage
- Ora, "disk space crisis"
  - Meno del 20% di AOD al CNAF, ma ragionevolmente selezionati

	ASGC	BNL	CERN	CNAF	FZK	LYON	NG	PIC	RAL	SARA	TRIUMF	%
ASGC												80
BNL												92
CERN												45
CNAF												21
FZK												84
LYON												85
NG												82
PIC												X
RAL												25
NIKHEF												36
TRIUMF												36

Analisi Dati Distribuita in CMS

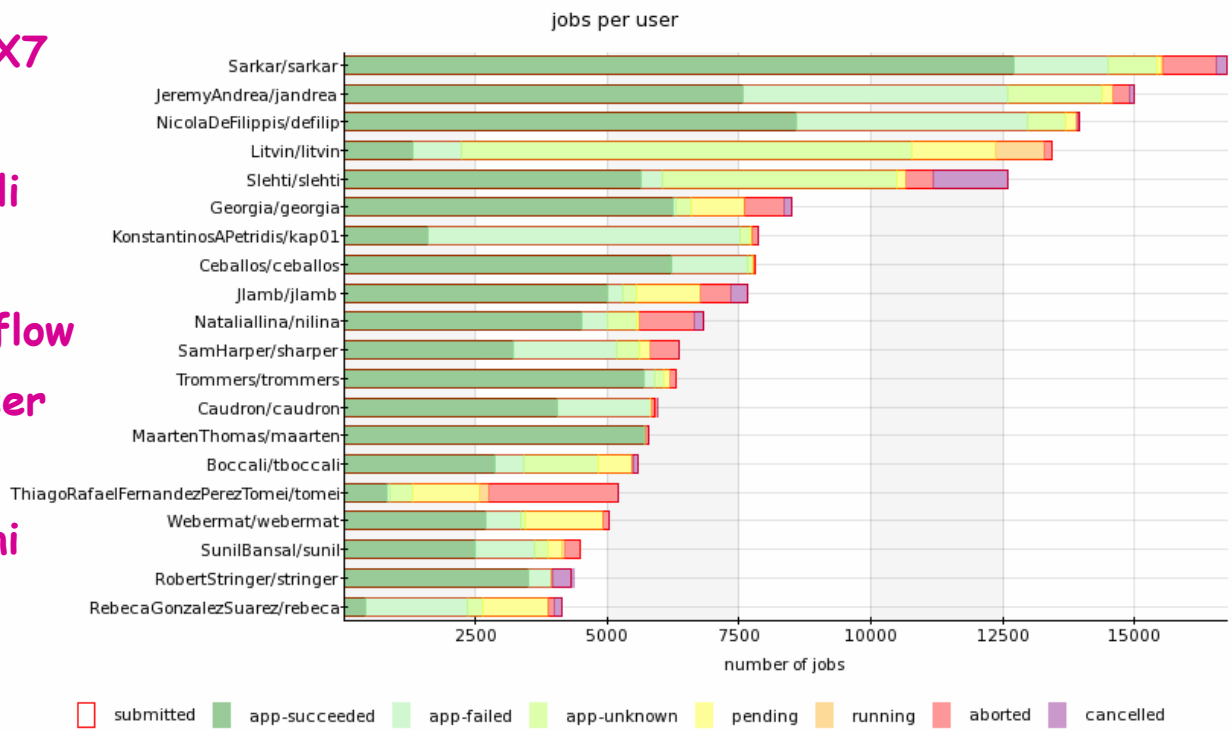
- Data driven Analysis Model
  - Dati distribuiti nei siti
  - Sw di CMS precedentemente installato
- Dal punto di vista dell'utente:
  - Sviluppo e test del codice localmente su piccoli campioni di eventi
  - Selezione/ricerca del dataset e Sottomissione dei jobs alla Grid
  - Ritiro degli outputs prodotti…
- CMS ha sviluppato un unico strumento per l'utilizzo dell'infrastruttura in fase di analisi:
  - CRAB: CMS Remote Analysis Builder

CRAB

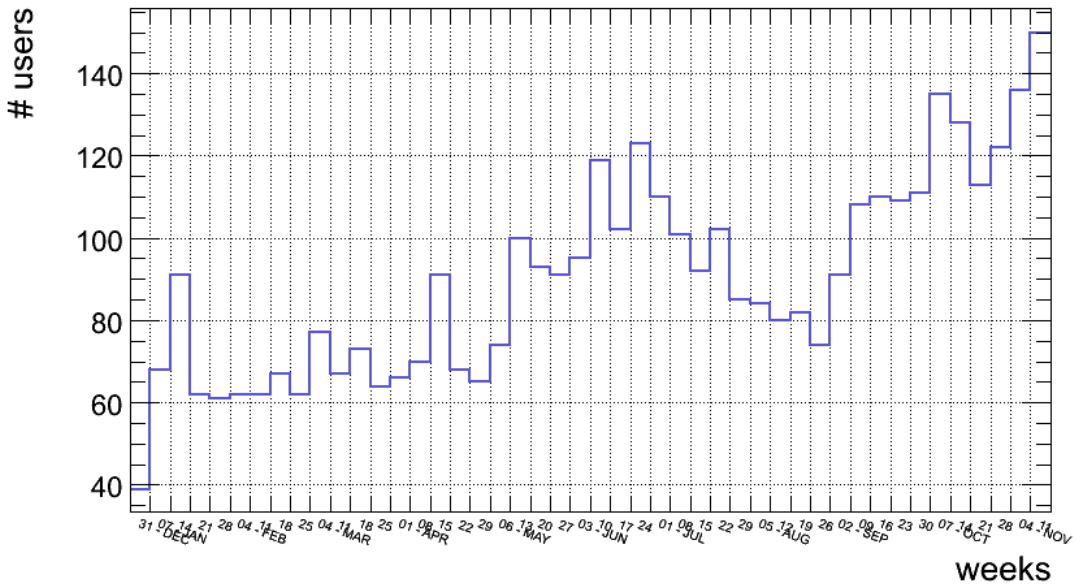
- Rende transparente l'utilizzo delle risorse di CMS per le attività "private" dell'utente:
  - Automatizza la data discovery, la preparazione, configurazione dei jobs, sottomisione…
  - Completamente interfacciato a OSG/LCG/risorse Locali (LSF,CAF)
  - Supporta la sottomissone attraverso un servizio 24X7

Gestione automatica degli errori, del ritiro degli Output…
(automatizza tutto il workflow riducendo il carico dello user aumentando l'efficienza)

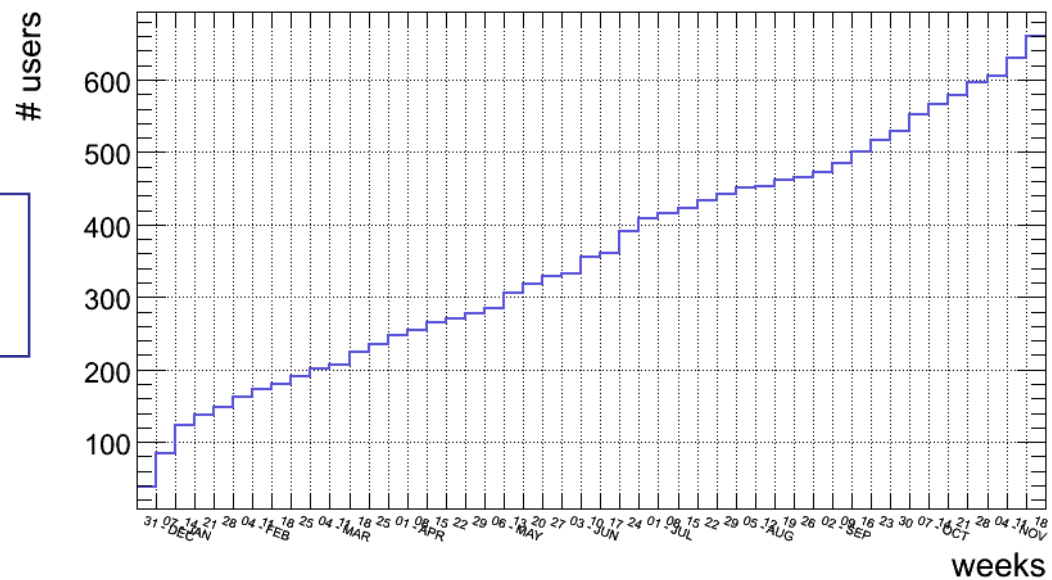
  - In produzione da ~3 anni



User Community nel 2007

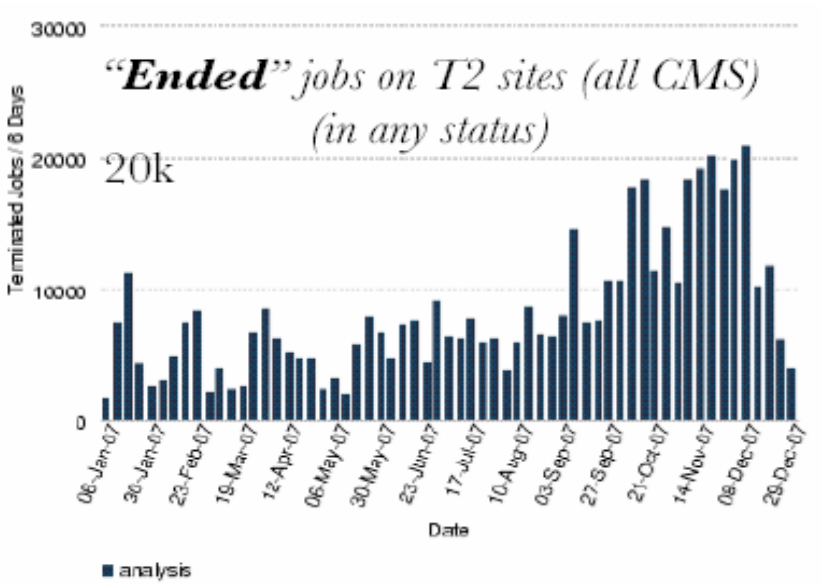


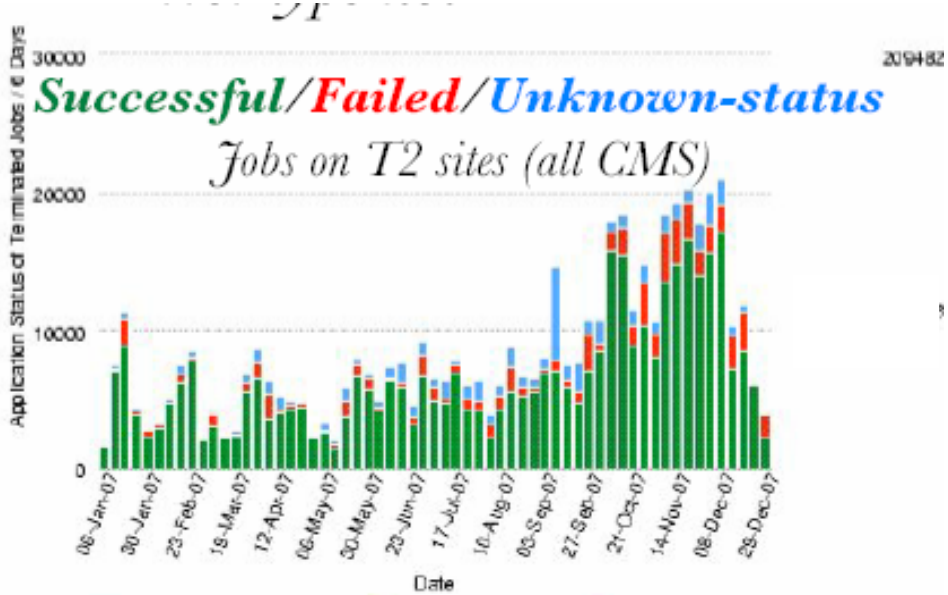
- Numero di utenti per settimana

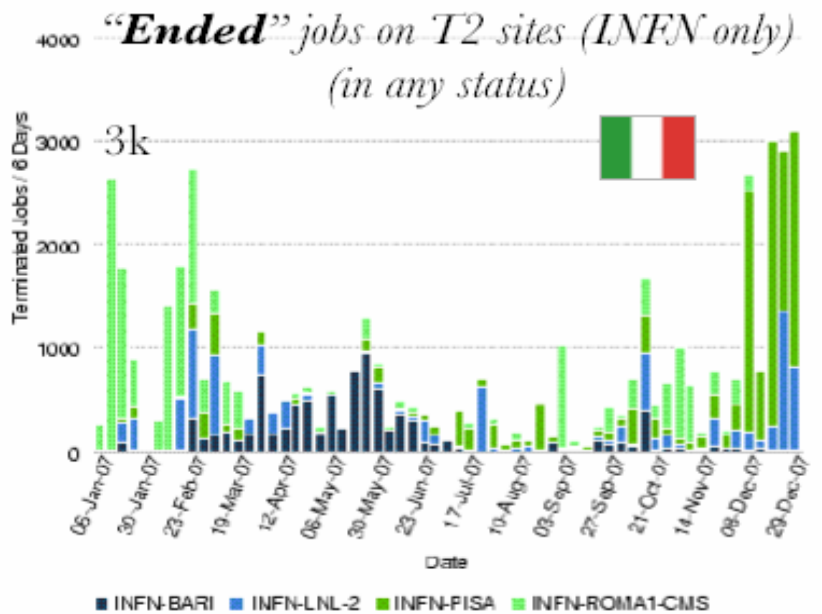


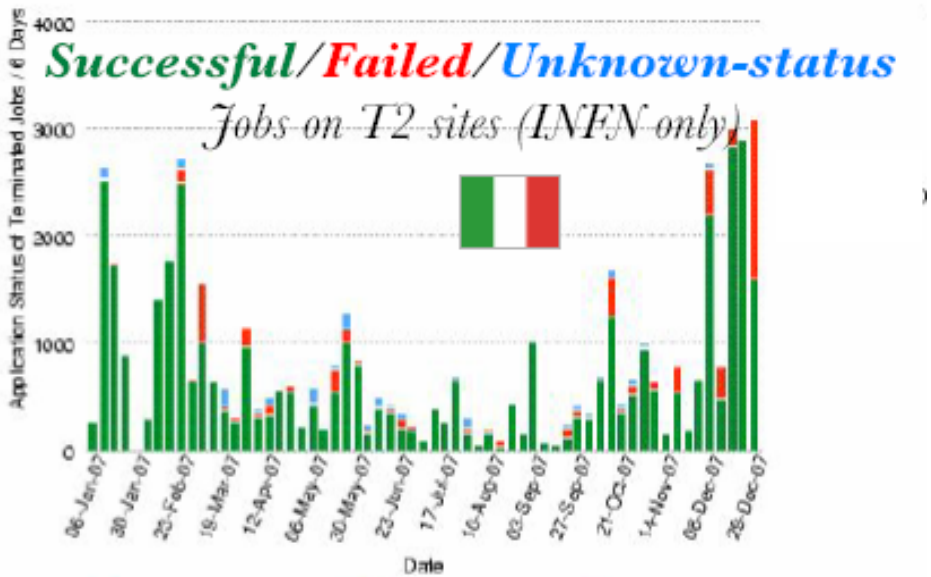
- Numero totale di utenti

CMS Analysis jobs 2007



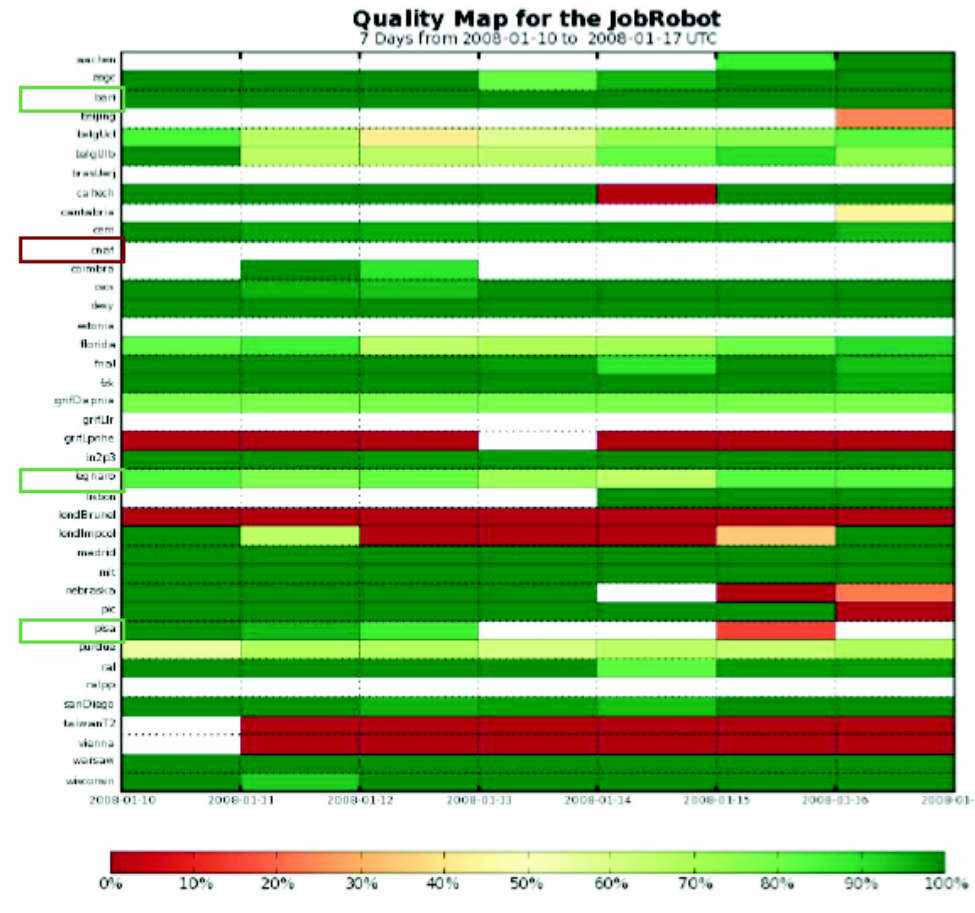


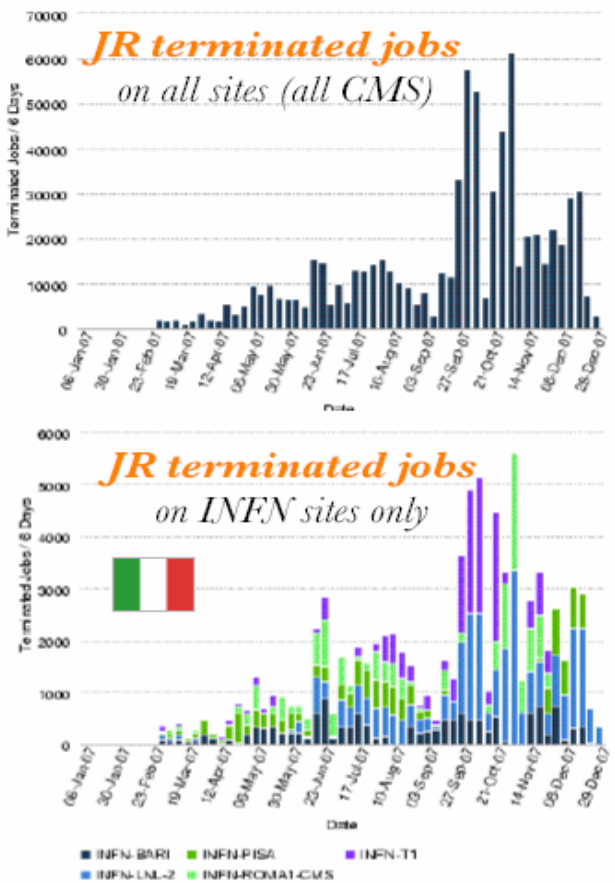




Commissionig dei siti per l'analisi

- JobRobot: generatore di fake-analysis-jobs
  - Implementato nel 2005/2006 come un "semplice wrapper" attorno a CRAB
  - Sottomette jobs a tutti I siti di CMS
  - Accsso via web alle statistiche ed ai log files.
- GangaRobot: generatore di fake-analysis-jobs





Summary

- ATLAS e CMS hanno adottato un modello di computing distribuito basato sulla tecnologia di Grid
- Stanno migliorando la qualità dei tools facendo commissioning di risorse, servizi e strumenti a larga scala con complessi esercizi di computing
- Next steps: see Simone slides….