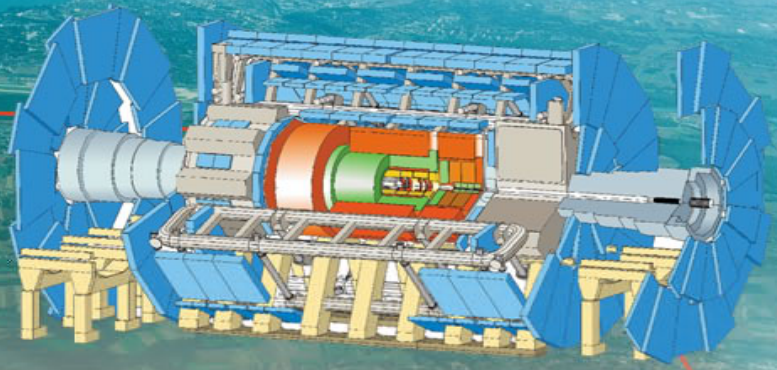
2009 LHC Run Computing

Gianpaolo Carlino
INFN Napoli
Atlas Italia, Roma, 02/02/10



Highlights from:

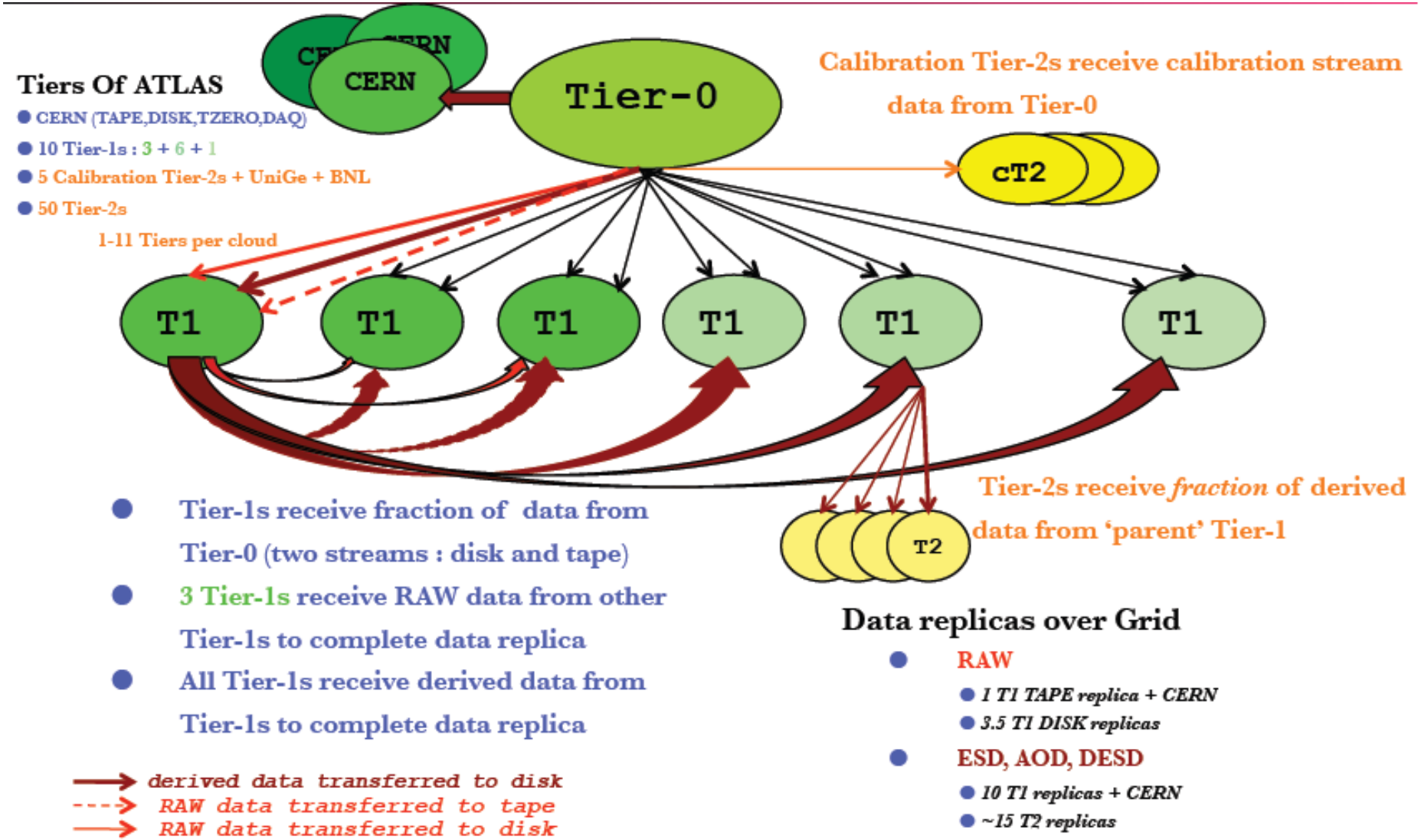
- Computing Post Mortem
- Tier3 Workshop

2009 LHC Run – Tier0 Ops

- Fast and reliable processing of data at Tier0 (and CAF)
  - Collaboration got access to data quickly
- Setup highly automated by now, continuously being improved during running
  - Many new features were added during running

- Tier-0 worked basically fine and reliably
  - Full processing chain in place and working well
  - New processes and features introduced and tested in time, e.g.
    - RAW merging, lumi-block aware AOD and DPD merging
    - Tier-0 data quality in COOL ("RAW file counter")
    - Express-stream processing chain
    - TAG and TAG_COMM uploading to several sites
    - Proof-of-principle for automated calibration processing (beam spot)
  - Enough flexibility for "panic" operations mode and for fast serving of special requests, e.g.
    - Parallel processing of same data in different configurations
    - Reprocessing at the Tier-0
    - Data replication to CAF/atlcal
    - Fast registration with DDM and AMI, selection filters
- Good, efficient collaboration with PROC and DP, DQM, DDM, TAG, s/w validation groups, etc.
  - Working procedure and tools for configuration changes by PROC

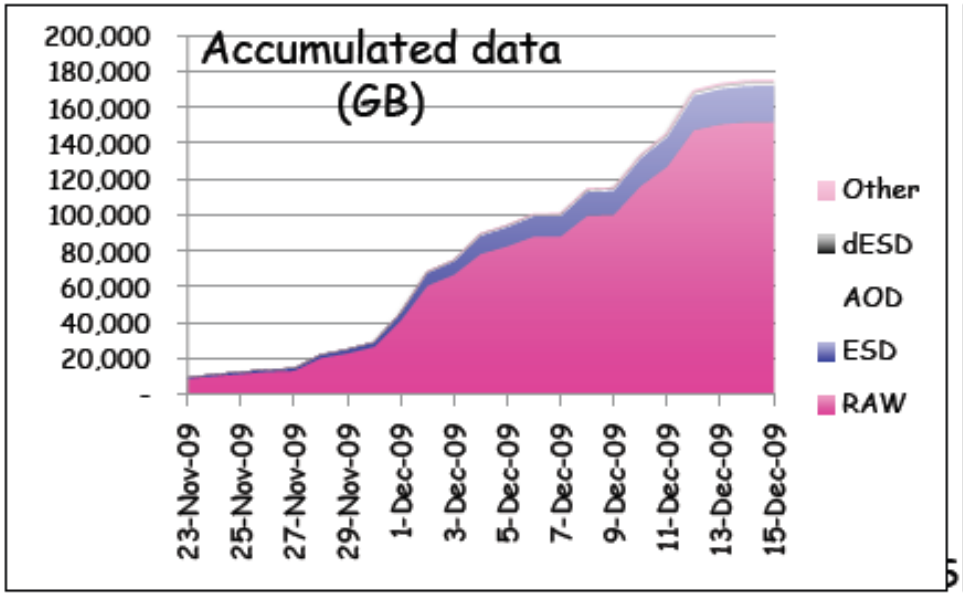
2009 LHC Run – Data Distribution

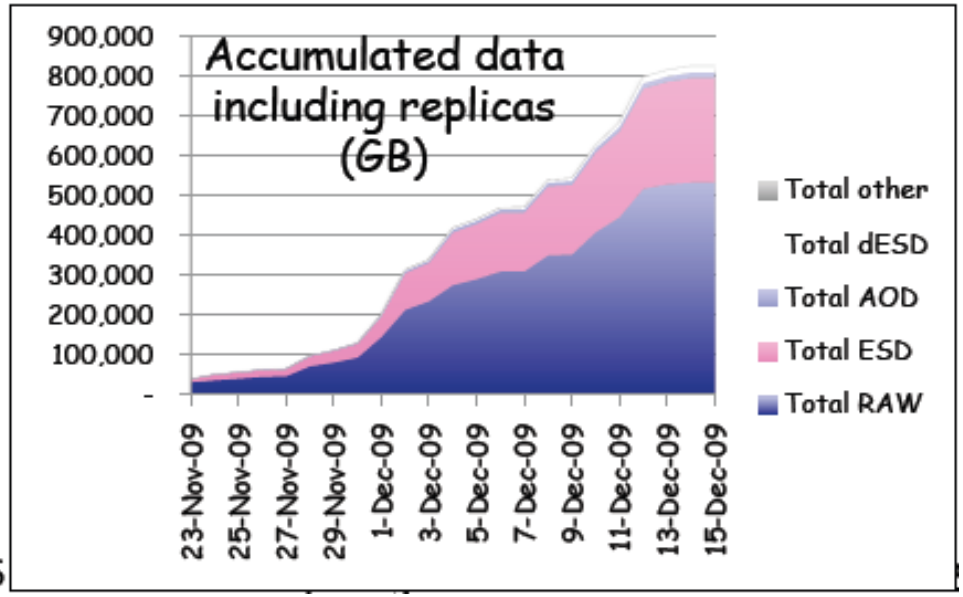


2009 LHC Run – Data Distribution

project	RAW (TB*)	ESD (TB)	AOD (TB)	DESD (TB)	NTUP (TB)	TAG (TB)
1beam	3.8	0.2	0.01	0.01	0.003	0.0002
900GeV	138	21.0	1.1	2.4	0.4	0.005
2TeV	7.4	1.0	0.03	0.04	0.01	0.0002
Total (TB)	149.2	22.2	1.14	2.45	0.41	0.005

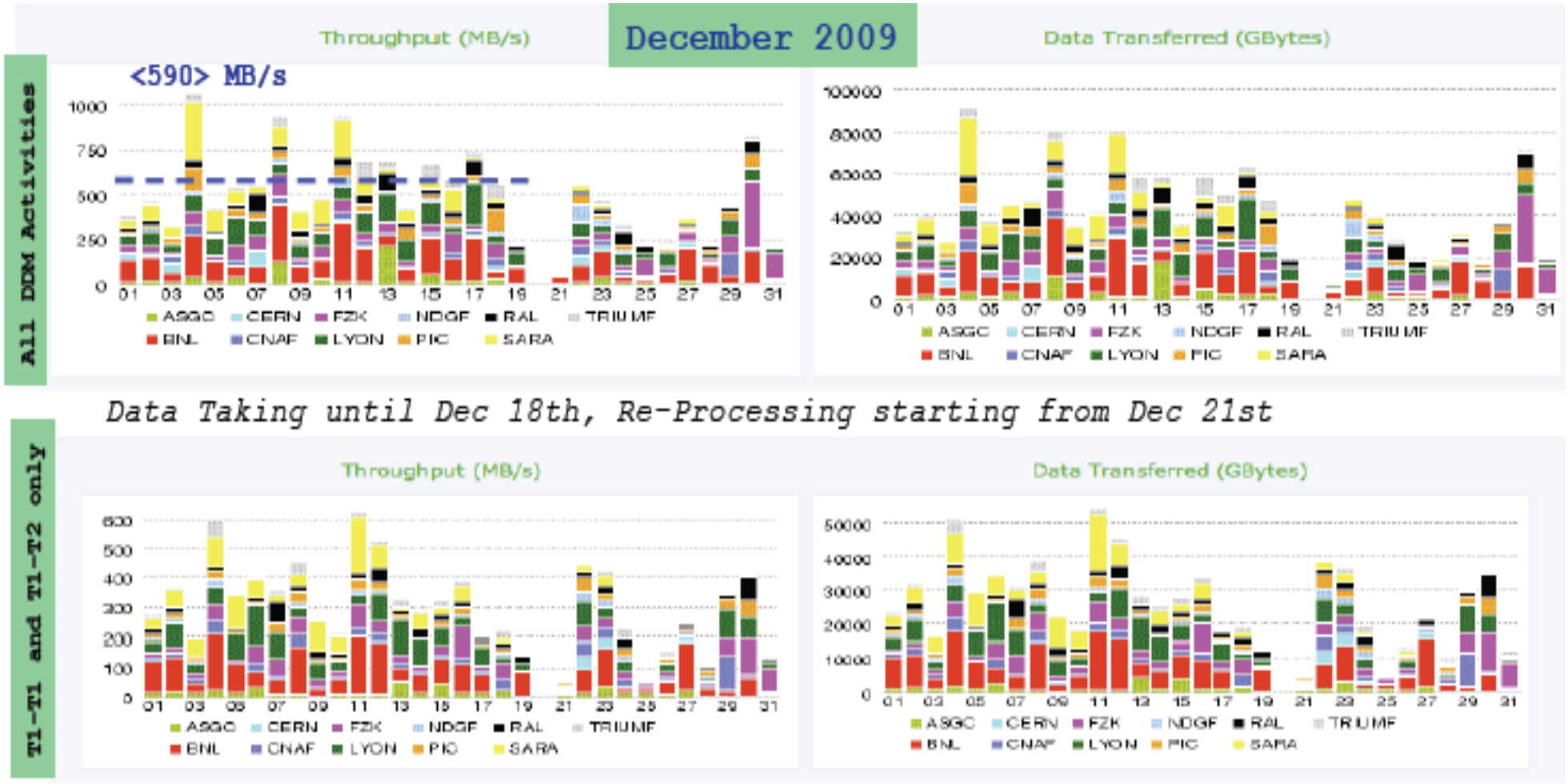
*) express, calibration and debug streams are not counted



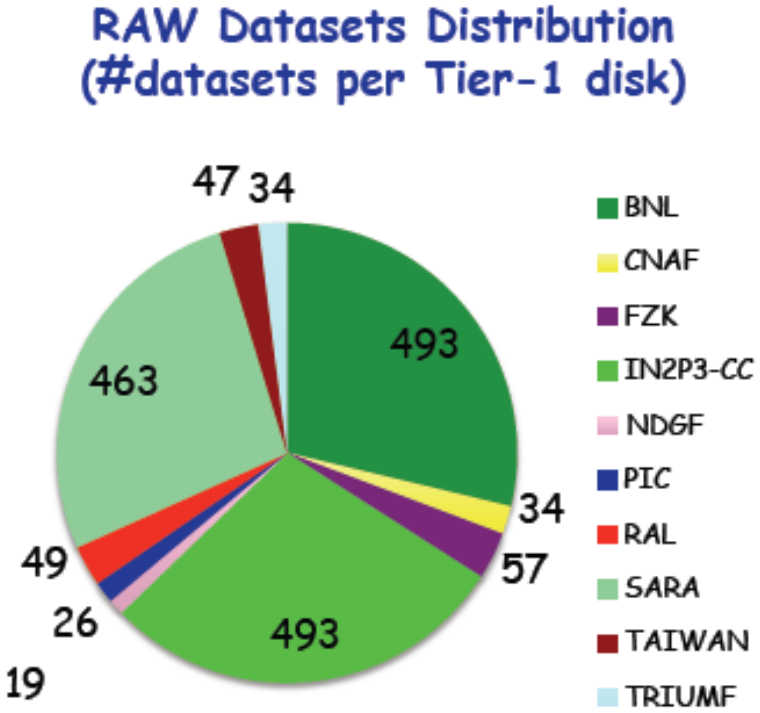


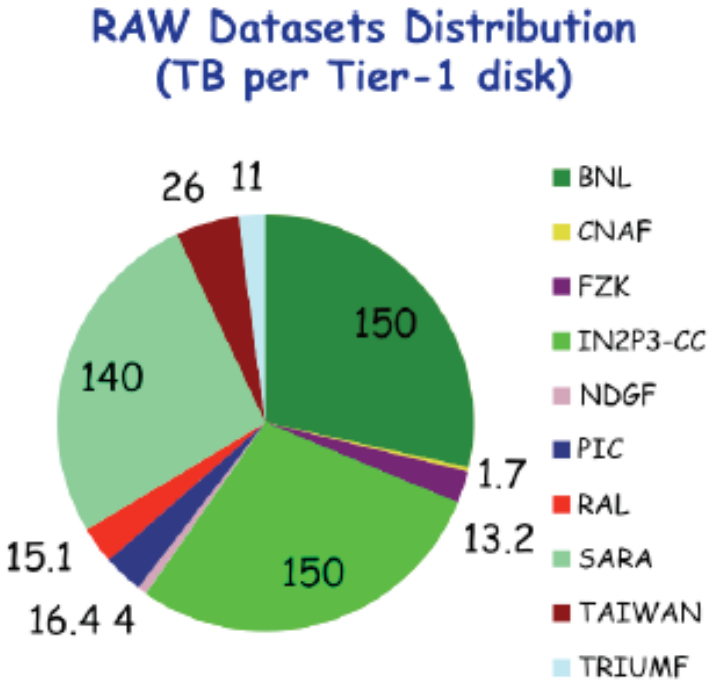
2009 LHC Run – Data Distribution

Efficienza di trasferimento = 100%



2009 LHC Run – Data Distribution





Data distribution between Tiers is 'unequal' :

PIC : 16 TB of RAW data

RAL : 15

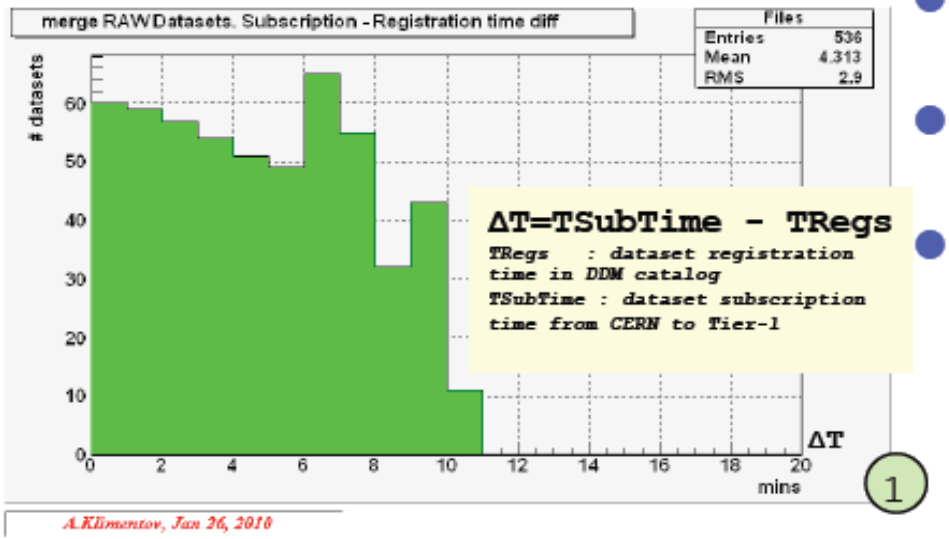
FZK : 13

CNAF and TRIUMF have the same number of datasets and factor 6 difference in data volume

Distribuzione sbilanciata perche' basata sul # di dataset e non il reale volume di dati. Alla lunga si bilancera'

2009 LHC Run – Data Distribution

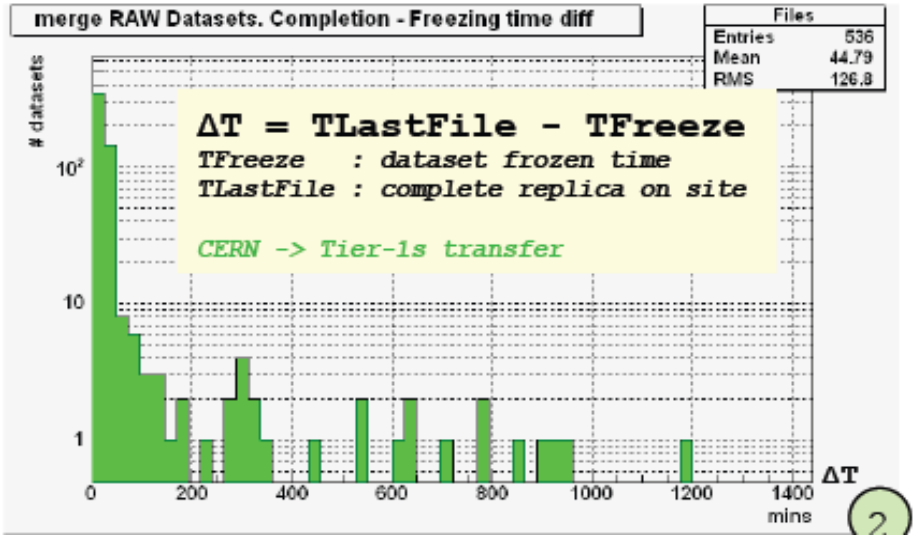
2009 Run. data09_900GeV. Registration time



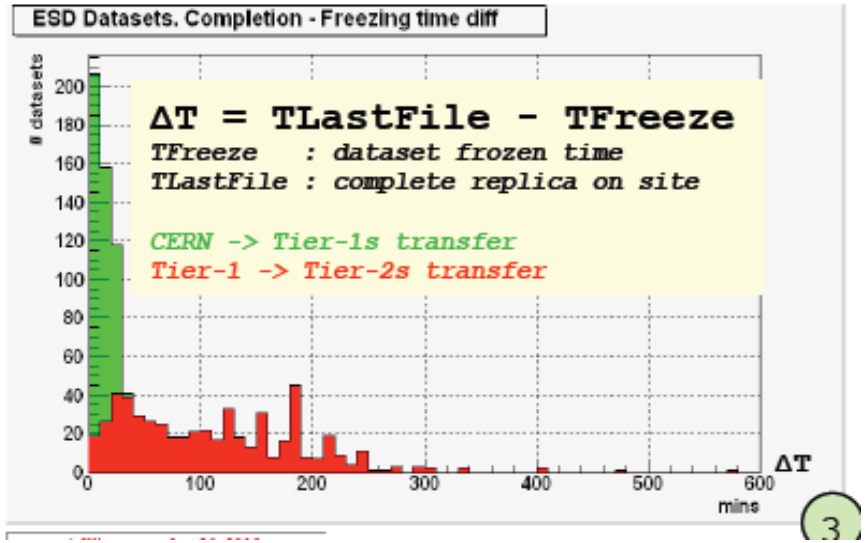
A.Klimentov, Jan 26, 2010

- Datasets are subscribed to Tier-1s in <4'> after registration [1]
- RAW dataset replica is available at Tier-1 in <45'> after dataset is frozen [2]
- ESD dataset replica is available at Tier-1 and 'CERN' in <20'> after dataset is frozen and at Tier-2 in <1h50'> after dataset is frozen [3]

2009 Run. data09_900GeV. Completion time



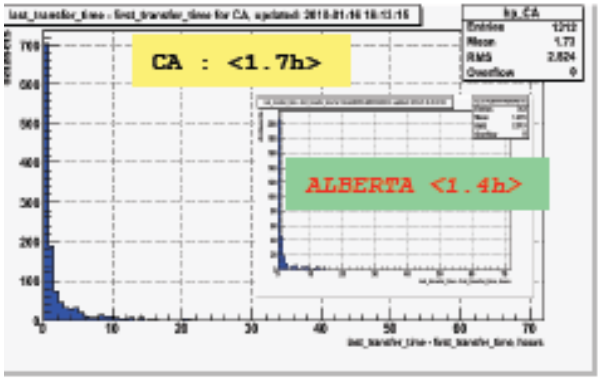
2009 Run. data09_900GeV. Completion time

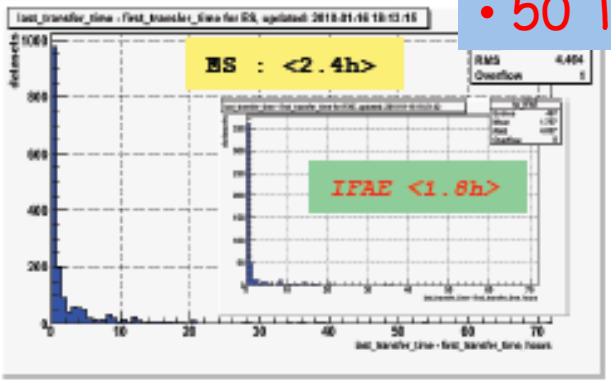


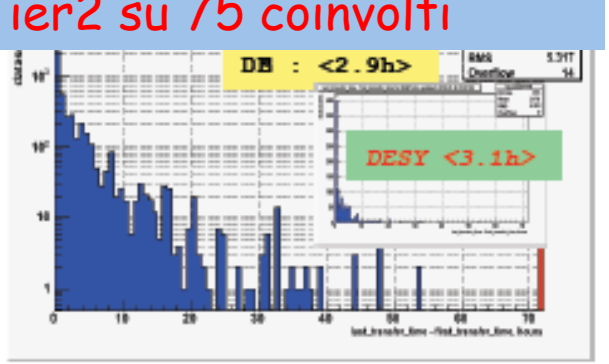
2009 LHC Run – Data Distribution

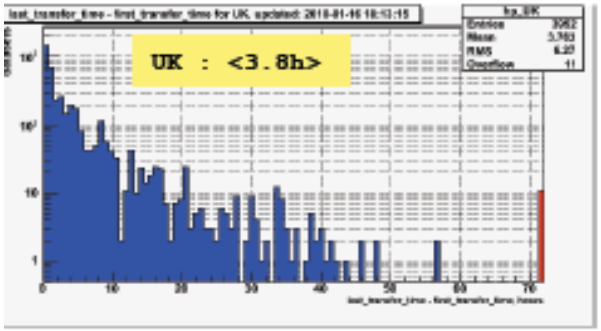
$\Delta T[h]$ = TcompleteReplica – Tsubscription

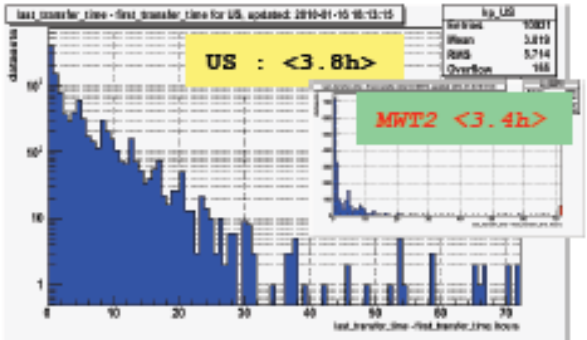
- Maggior parte dei siti stabili
- 50 Tier2 su 75 coinvolti

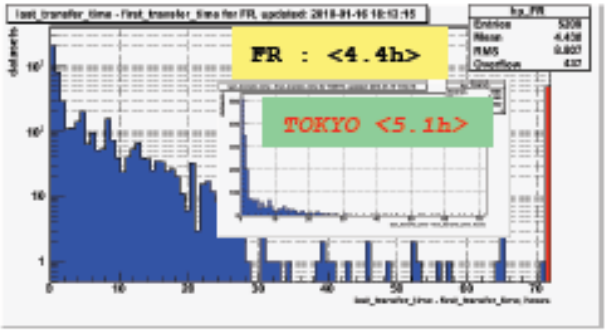


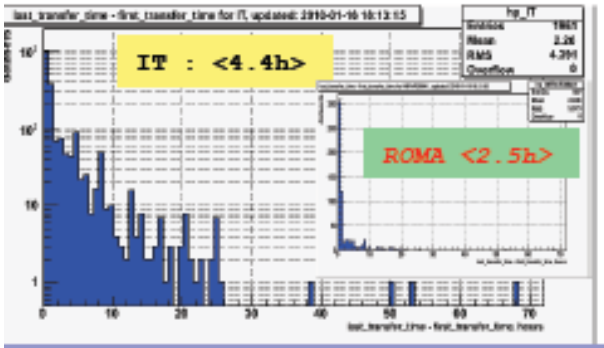


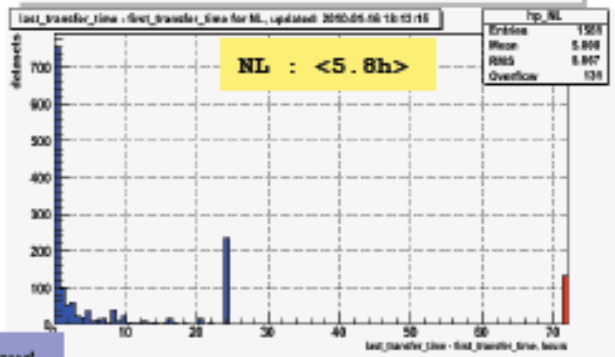


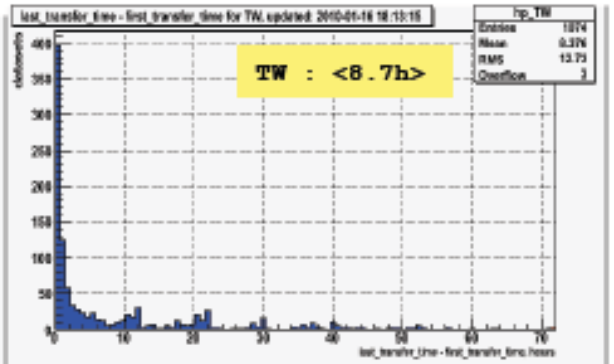








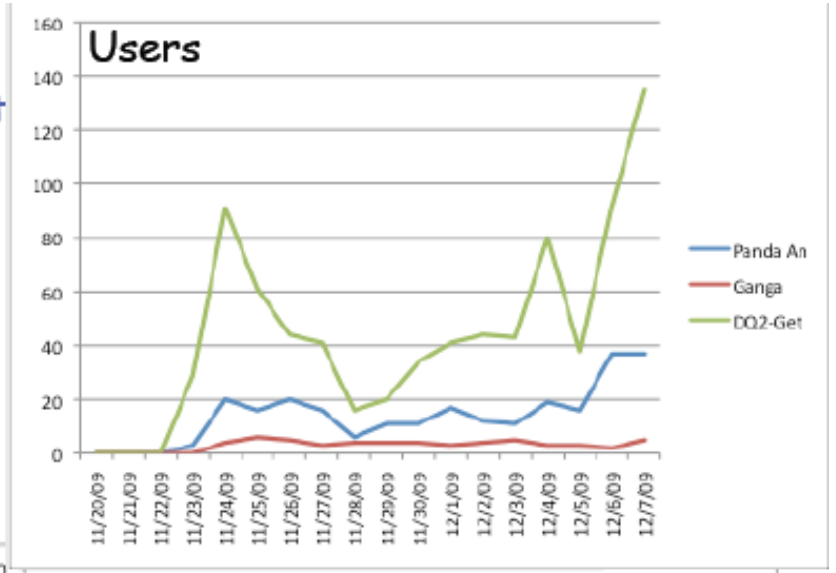


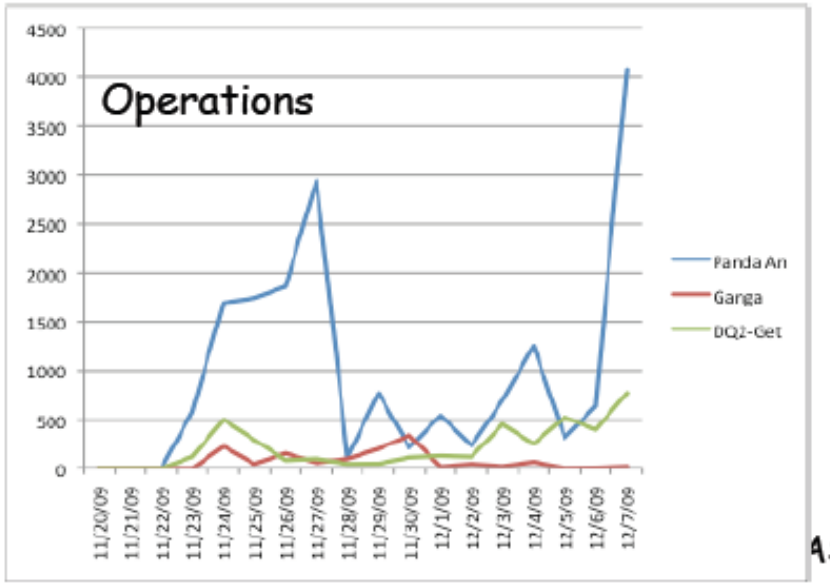


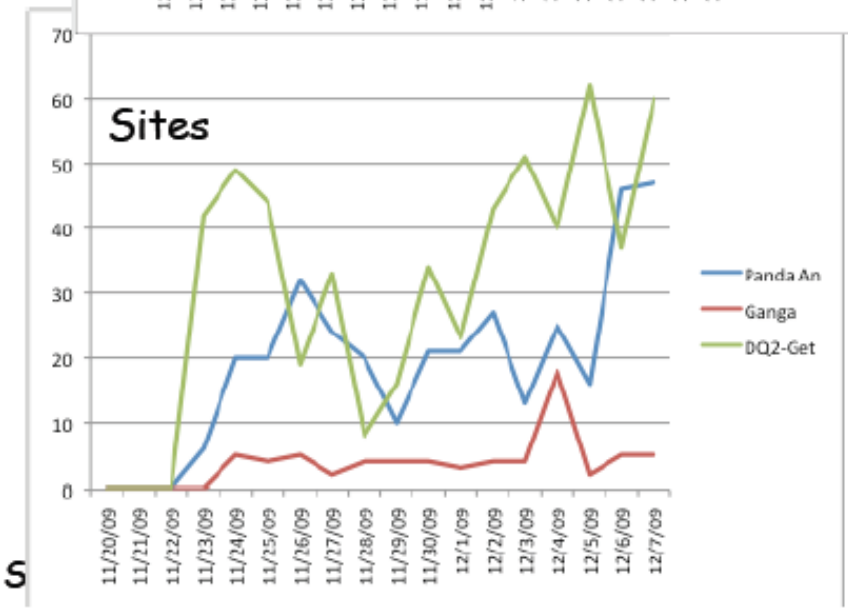
NDGF has unique data distribution model within cloud

2009 LHC Run – Distributed Analysis

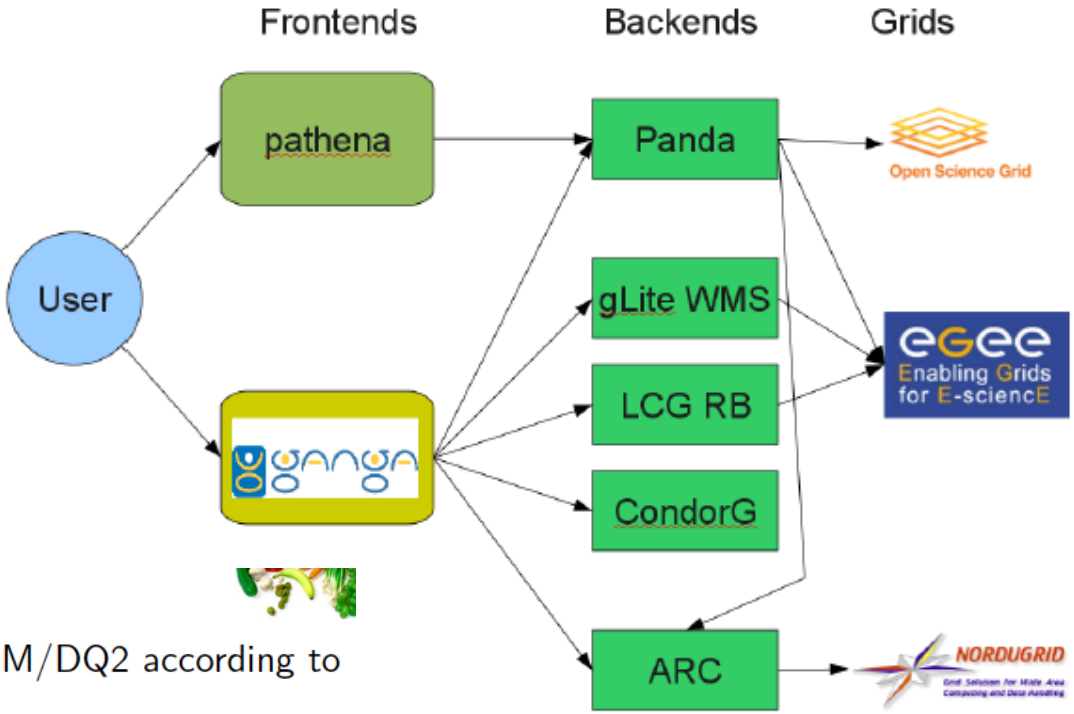
- Most people copied the (relatively) little amount of data to their facilities with dq2-get and analysed them locally
- Popularity fractions:
  - ESD ~90%
  - DESD_COLLCAND ~5%
  - AOD ~1%
  - RAW ~4%







2009 LHC Run – Distributed Analysis



Data

- Centrally organized data distribution by DDM/DQ2 according to computing model

Athena distribution kits

- Centrally organized installation on EGEE, OSG and NG

User jobs

- Model: ,,Job to Data"
- Package user code and ship it to data location

User Output:

- Store on site *SCRATCHDISK* or transfer to remote *LOCALGROUPDISK*
- Retrieval with dq2 command line tools

2009 LHC Run – Distributed Analysis

Panda

- Users in the last 3 days: 208, 7: 280, 30:402, 90:699, 180:839
- On average now 30000 jobs per day (finer grained job splitting !)

glite WMS

- Users in the last 3 days: 47, 7: 101, 30:147, 90:323
- On average 10000 jobs per day
- Some loss in monitoring statistics

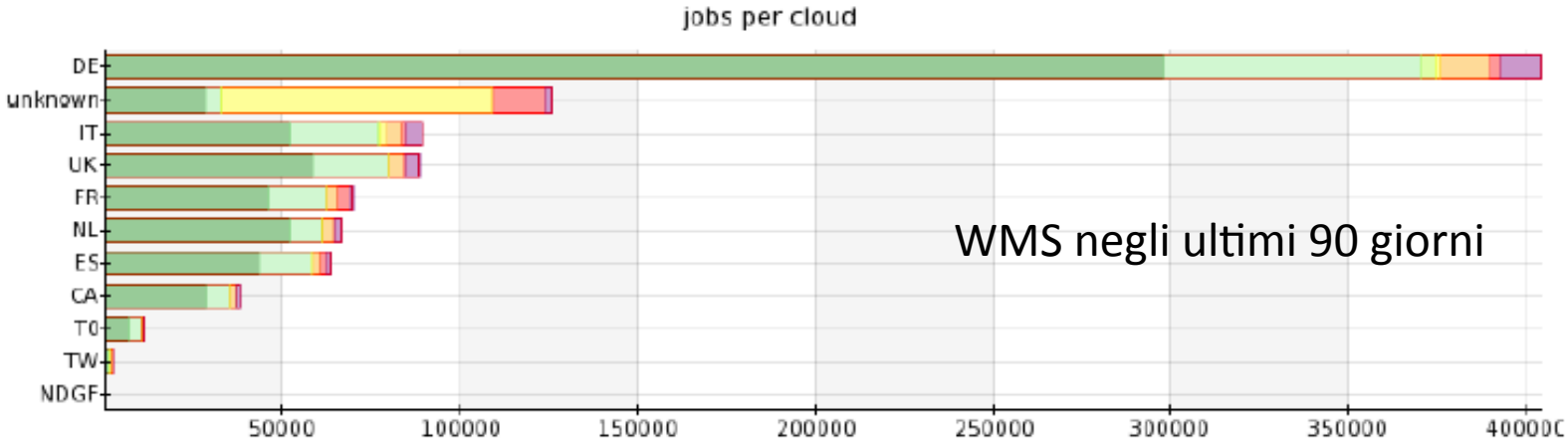
Other

- Some more users on NG and Batch back-ends at local resources

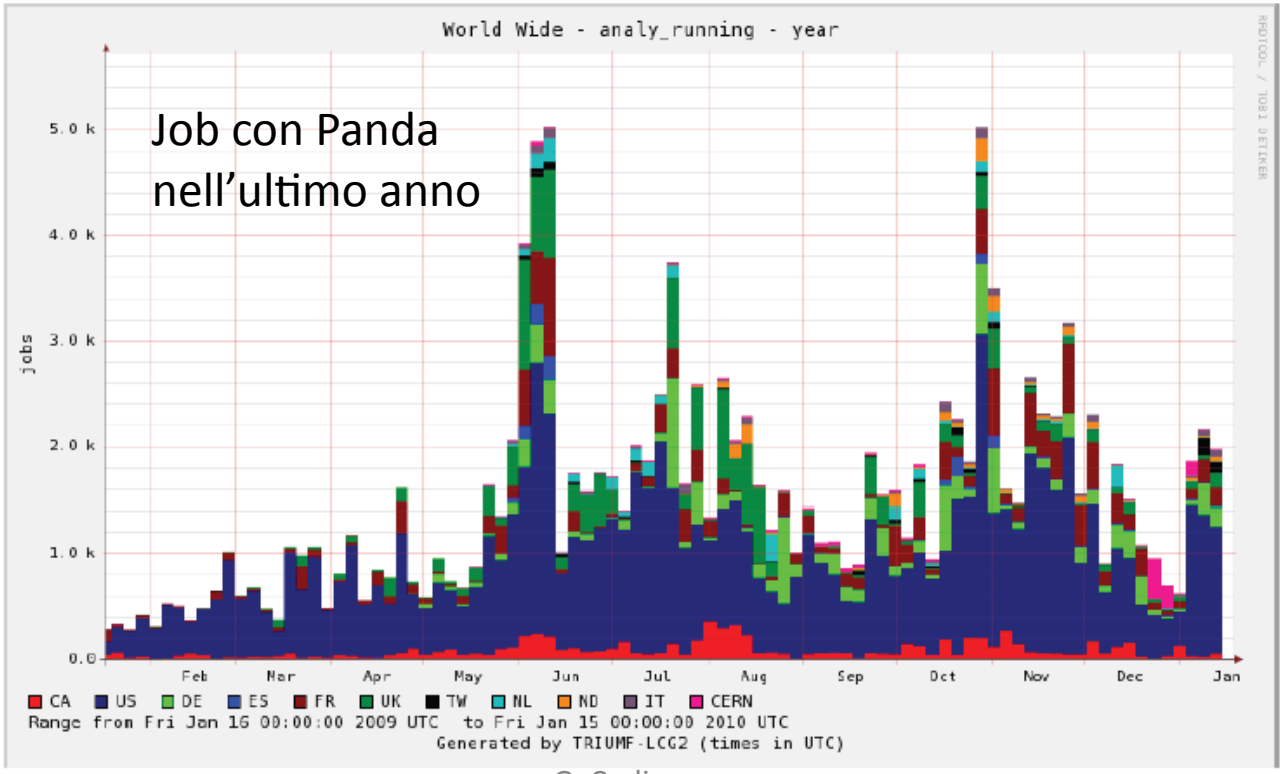
Overall

- Increased number of users esp. after collision start, but still not at full possible capacity
- Feeling is still that many people are running offline

2009 LHC Run – Distributed Analysis



WMS negli ultimi 90 giorni



Job con Panda nell'ultimo anno

2009 LHC Run – Distributed Analysis

Steadily increasing number of users

What is working well:

- Analysis at most of the sites
- Most of standard work-flows on MC and real data

What works, but needs improvement:

- 'Blind' job submission
- Exotic use cases
- Stabilize DB access work-flow
- Overall site availability
- Coherent job monitoring

For the distributed analysis it is vital to have:

- Easy interface that does not scare off physicists
- A reliable and robust service of many components

Atlas Tier3

Workshop Tier3 (25 e 26/01)

http://indico.cern.ch/conferenceTimeTable.py?confId=77057&showDate=all&showSession=all&detailLevel=contribution&viewMode=parallel

Discussione sulle funzionalita' di un Tier3 e le possibili implementazioni tecniche e report dello stato attuale

Phyisics Analysis Workflow (talk Wouter Verkerke):

1. Individuazione dei vari step di cui si compone un'analisi
   - simulazione, produzione di ntuple, analisi di ntuple
2. Analysis workflow
   - Iterazione e tempi di ciascuno step
3. Determinazione delle risorse di computing necessarie
4. Individuazione del sito (Tier n) piu' adatto per ogni fase di analisi
5. Suggerimenti sulla struttura di un Tier3 per l'analisi

Atlas Tier3

4 Possibili Scenari (talk Simone Campana):

1. Service Based Tier3 (small Tier2)
   - Simile ad ogni altro sito in DDM (Tier1/2), associato ad una cloud in ToA. Importa dati con le sottoscrizioni DDM, utilizza i servizi (e il supporto) centrali
   - Richiede un notevole impegno di gestione locale
   - Differenza: risorse non pledged
2. Co-hosted Tier3 (componente Tier3 di un Tier2).
   - risorse dedicate per la (le) comunita' locale (locali)
   - CPU dedicate nel batch system, accesso allo storage del Tier2 e LOCALGROUPDISK per utenti locali
   - >90% dei Tier3 attuali
3. Client Based Tier3
   - Non e' un sito Grid e DDM (usa i DQ2 client per importare i dati)
   - Adatto a piccoli siti per uso interattivo. Facile da gestire ma non scalabile
4. Hybrid Tier3 (simile allo scenario 1)
   - Storage senza SRM
   - Solo destinazione per i trasferimenti DDM e non sorgente
   - Da valutare pros e cons e l'impatto sulle operazioni di ATLAS

Options for T3 sites

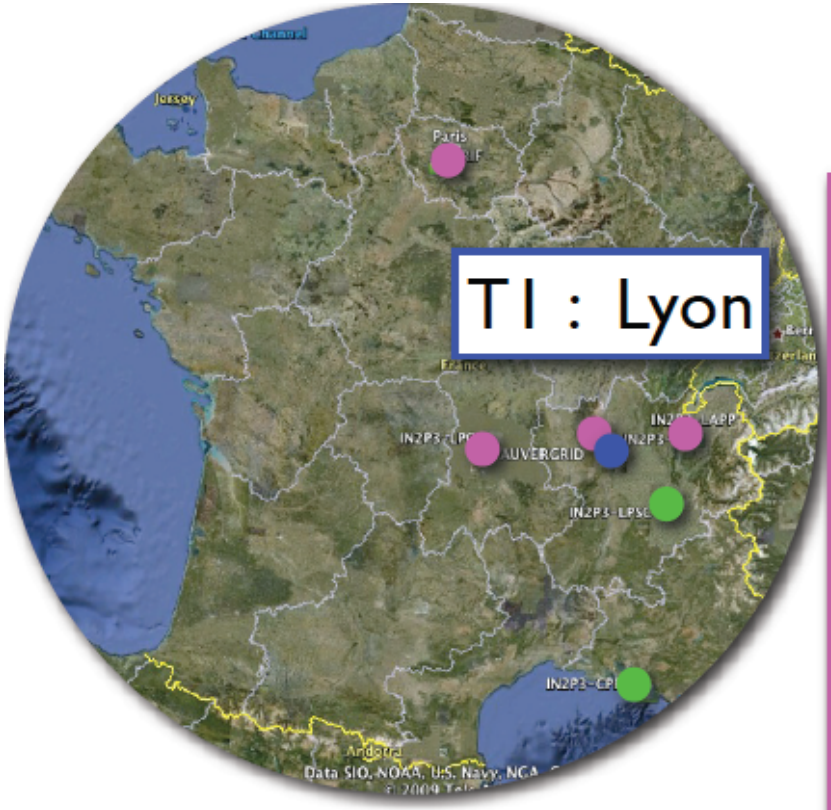
(talk Alessandro De Salvo)

- **Sites integrated in the Grids**
  - **Can take advantage of the automatic installation system**
    - Software releases (+ compilers, where needed)
    - DDM clients
    - PFC/Frontie
    - Other needed software
  - **Condition files / Frontier**
    - For sites not deploying an HOTDISK space token, the installer tool may download and incrementally update only the condition files using a predefined disk location (at least 500 GB needed)
    - Any allowed remote Frontier server can be used
    - See Rod's talk for other options
- **Sites not integrated in the Grids**
  - **Can manually download the installer tool and deploy the needed releases, condition files and Frontier setup, in the same way they are deployed in the grid sites**
    - http://kv.roma1.infn.it/KV/sw-mgr
    - A page will all the needed instructions will be available soon
  - **Experimental Firefox plugin for automatic installation of the software releases**
    - https://atlas-install.roma1.infn.it/atlas_install/firefox

A. De Salvo – ATLAS Software Installation and T3 sites – ATLAS T3 Workshop, CERN 25-1-2010

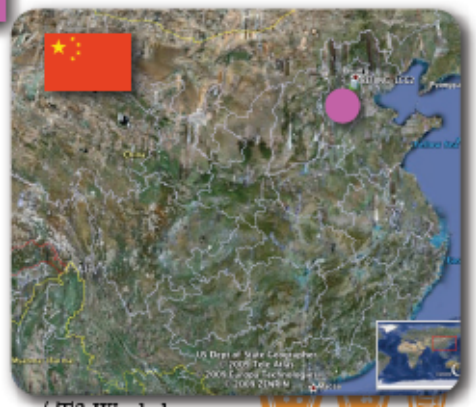
Atlas Tier3 – es.1: Francia

FR-Cloud



T2 : 8
- Annecy
- Bucarest x2
- Clermont
- GRIF
- Lyon
- Beijing
- Tokyo

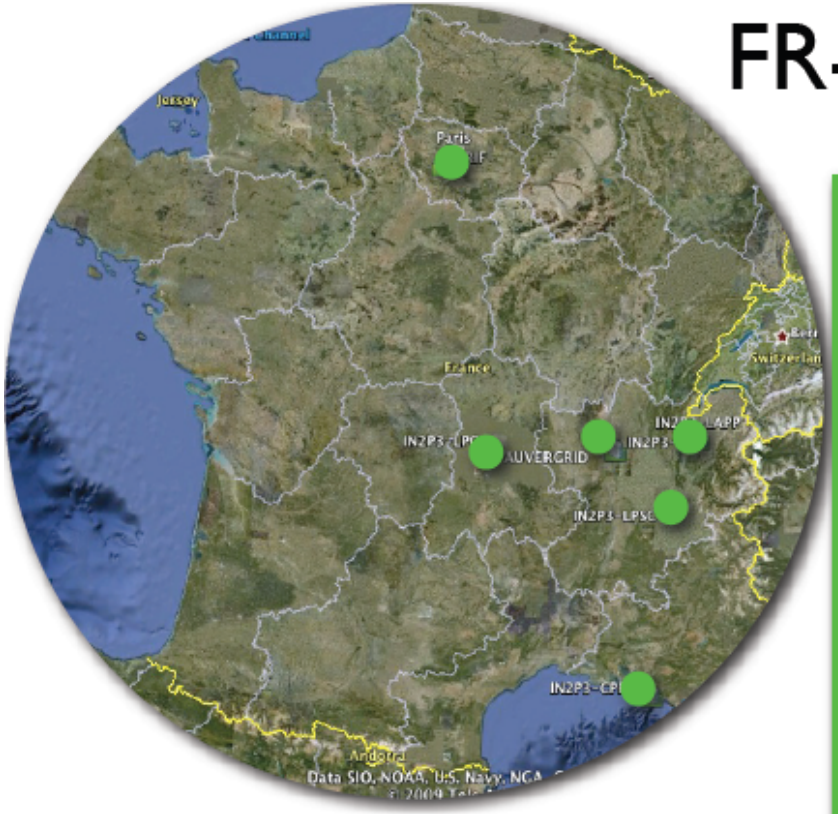
T3 : 2
- Grenoble
- Marseille



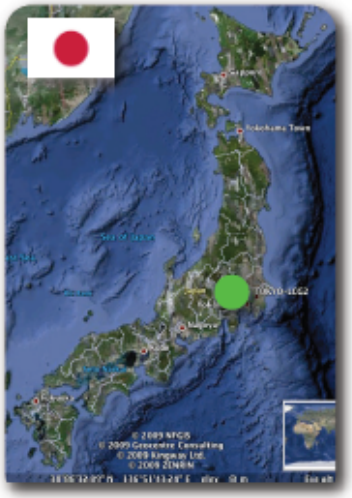
irfu
cea
saclay

Atlas Tier3 – es. 1: Francia

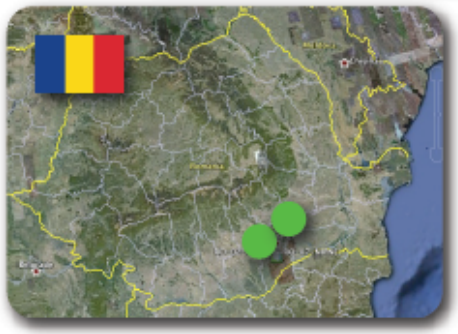
FR-Cloud another view

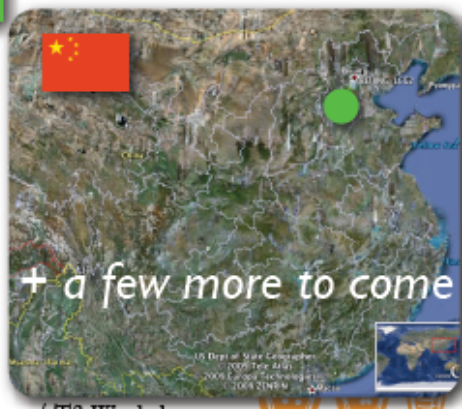


T3 : 8
- Annecy
- Bucarest x2
- Clermont
- GRIF
- Lyon
- Beijing
- Tokyo



T3 : 2
- Grenoble
- Marseille





irfu
cea
saclay

Atlas Tier3 – es. 1: Francia

- All T2s have a T3 component
  - 25-30+% of total resources are T3
  - For Grid and non-Grid usage

T3-only

- Are Grid enabled with DDM end-points
- They are integrated in
  - Software distribution system
  - MC Production
  - Functional Tests (DDM, SAM, etc...)
  - pAthena
- Actively participating in Squad-FR activities

Atlas Tier3 – es. 2: USA

- Tier3 computing resources in US ATLAS belong to the institute.
  - The institute has control over how the T3 resources are used.
  - In this sense, T3 is more an extension of the physicist desktop rather than the extension of the Grid (for most US ATLAS T3 sites)

Some classifications: (defined by Chip Brock's report on T3s for US ATLAS

- T3gs (Tier3 with Grid Services): This type of T3 has the ranges of Grid site services supported at T2. Capable of accepting jobs from outside.
- T3g: This type of T3 gets data from the Grid by does not accept jobs from the outside. It also does not serve the locally stored data to the outside.
- T3w: Single workstations. (not concerned with these in this talk)

Atlas Tier3 – es. 2: USA

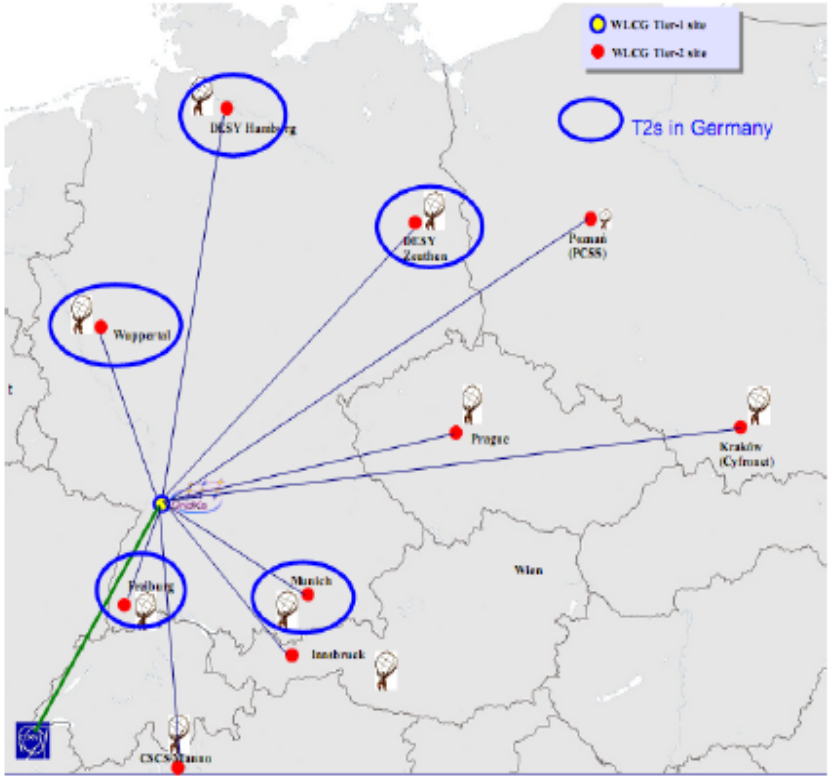
- US ATLAS has 4 T3gs sites.
- US ATLAS has 20 T3g sites.
  - 8 of these are associated with Tier2 or Tier1
  - The remaining 12 are of variety of designs and capacities.
    - Four are coupled to departmental or university infrastructure.
    - Others are largely standalone.
- Most US ATLAS T3's plan an expansion in 2010. Funding approved.

- 16 Institutes plan to build a Tier3 from scratch in 2010. Funding approved and waiting for arrival.
- Expect total of 40 US ATLAS institutes to operate T3g in 2010. All of these will either be completely new, or will have significant upgrades in 2010.

Atlas Tier3 – es. 3: Germania

T3 Resources at the T2s and the T1

- T2 sites include: Freiburg, Goettingen, the Munich sites (MPPU and LMU), Wuppertal and Desy-HH and ZN with the NAF as the T3 part.



- T3 resources funded locally and/or Nationally (the DGrid project)
- About ≈30% of the T2 resources, both CPU and storage, are for T3 usage.
- All the named sites run dCache.

Atlas Tier3 – es. 3: Germania

University Groups T3s

- Six sites in this category:
  - Bonn, Dresden, Dortmund, Maiz, Siegen and Wuerzburg.
  - Dresden and Dortmund have a settings similar to the T2s. Participates in atlas panda-prod.
  - Current status for the group T3s
    - Bonn (dCache and Storm-Lustre): ⇒ More next slide
    - Dresden: 128/512 for LHC VOs. Runs dCache (20TB and 100TB Tape). Exploring directory exporting via NFS4 and also investigating on xroot-access for the T3 users.
    - Maiz : 128 CPU Cores, on dCaChe 26TB. NFS for direct local access.
    - Siegen: 12 cores and DPM 4TB. On default settings.
    - Wuerzburg: Is planning to set a T3. Will appreciate recommendations.

Tier3 in Italia

In Italia la discussione sui Tier3 e' appena cominciata.

Motivazioni principali:
- attivita' di analisi limitata nei Tier2
- "prudenza" dell'INFN ad affrontare questo argomento.

- Fino ad oggi l'analisi extra-Tier2 e' stata effettuata (soprattutto) sui propri desk-laptop o su piccole farm di gruppo off-Grid

- Quasi tutte le sezioni hanno a disposizione farm in Grid (spesso sottoutilizzate) condivise con altri esperimenti o con INFN-GRID. Dimensioni assolute e share per Atlas variabili.

- Informazioni sulle disponibilita' locali fornite dalla maggior parte dei gruppi. Queste farm possono essere considerate come "Tier3 seeds".

I tempi sono maturi!

Tier3 in Italia

E' necessario arrivare ad un Modello Italiano di Tier3:

1. Analizzare le strategie di analisi e identificare gli aspetti che possono caratterizzare il Tier3
   - Analisi interattiva (Proof), sviluppo, simulazione di piccoli sample di prova
   - Evitare che sia solo un piccolo Tier2. Non e' spendibile con l'INFN
2. Dimensionare il Tier3 in base alla dimensione dei gruppi afferenti e al numero di attivita' svolte
   - Quanti TB e CPU mi servono per utente/gruppo/attivita'???
3. Individuare lo scenario o gli scenari piu' adatti ai nostri scopi tra i 4 di Atlas
   - Non tutti partono dalle stesse condizioni: gruppi con grandi farm di sezione, dipartimento o INFN-GRID (scenario I) altri con piccole farm non in grid (scenario III) magari con risorse obsolete
   - Fornire i Tier2 di una componente Tier3
   - Particolare attenzione alle scelte sullo storage
4. Definire un timing credibile
   - Non tutte le farm possono diventare subito Tier3 ufficiali
   - E' necessario individuare siti pilota che hanno gia' un'infrastruttura adeguata e personale competente su cui effettuare tutti gli studi e i test necessari

Tier3 in Italia

Per arrivare ad un Modello di Tier3 italiano il piu' velocemente ed efficientemente possibile e' necessario:

- individuare dei siti con infrastrutture GRID funzionanti su cui effettuare dei test.
- Alcuni siti (GE e UD) hanno gia' avviato discussioni con i direttori INFN e i centri di calcolo per le necessarie autorizzazioni e il supporto locale

Attivita' gia' in corso:

- Test di Proof a Milano e della configurazione della componente Tier3 di un Tier2 (Pi e Mi)
- Test delle funzionalita' di una farm (proto-Tier3) come end-point DDM e di Ganga per sottomissione al batch system locale (UD nella farm di Trieste)
- Studio dello soluzioni di storage per un Tier3 (GE)
- Conoscenza e test dei tool di analisi distribuita (RM3)

- Inserimento nei sistemi di Atlas (Tiers of Atlas) e partecipazione agli Hammer Cloud, Functional Test e eventualmente produzione.
- Collaborazione con le attivita' della cloud. Sinergia comunita' Tier1/2 e Tier3

Altri siti possono inserirsi (BO presto avra' una grande farm di dipartimento condivisa con CMS) ma all'inizio e' bene partire con quelli gia' avviati.

Tier3 in Italia

- formare una task force di persone gia' attive in Atlas nell'ambito del computing per coordinare i test e valutare tutti gli aspetti tecnici.

  - Dario Barberis (chair dal 1/3/10)
  - Marina Cobal
  - Alessandro Brunengo (Ge)
  - Umberto De Sanctis (Ud)
  - Fulvio Galeazzi (Rm3)
  - Massimo Pistolese (Mi)
  - Roberto Vitillo (Pi)
  - + G.C. e Alessandro De Salvo (in veste solo di consulente)

- Scopo: valutare la reale situazione dei siti italiani ed arrivare in pochi mesi ad un modello per la discussione con i referee

- Prematura al momento ogni discussione su eventuali richieste finanziarie.