The ATLAS Computing (and gLite)

Alessandro Di Girolamo

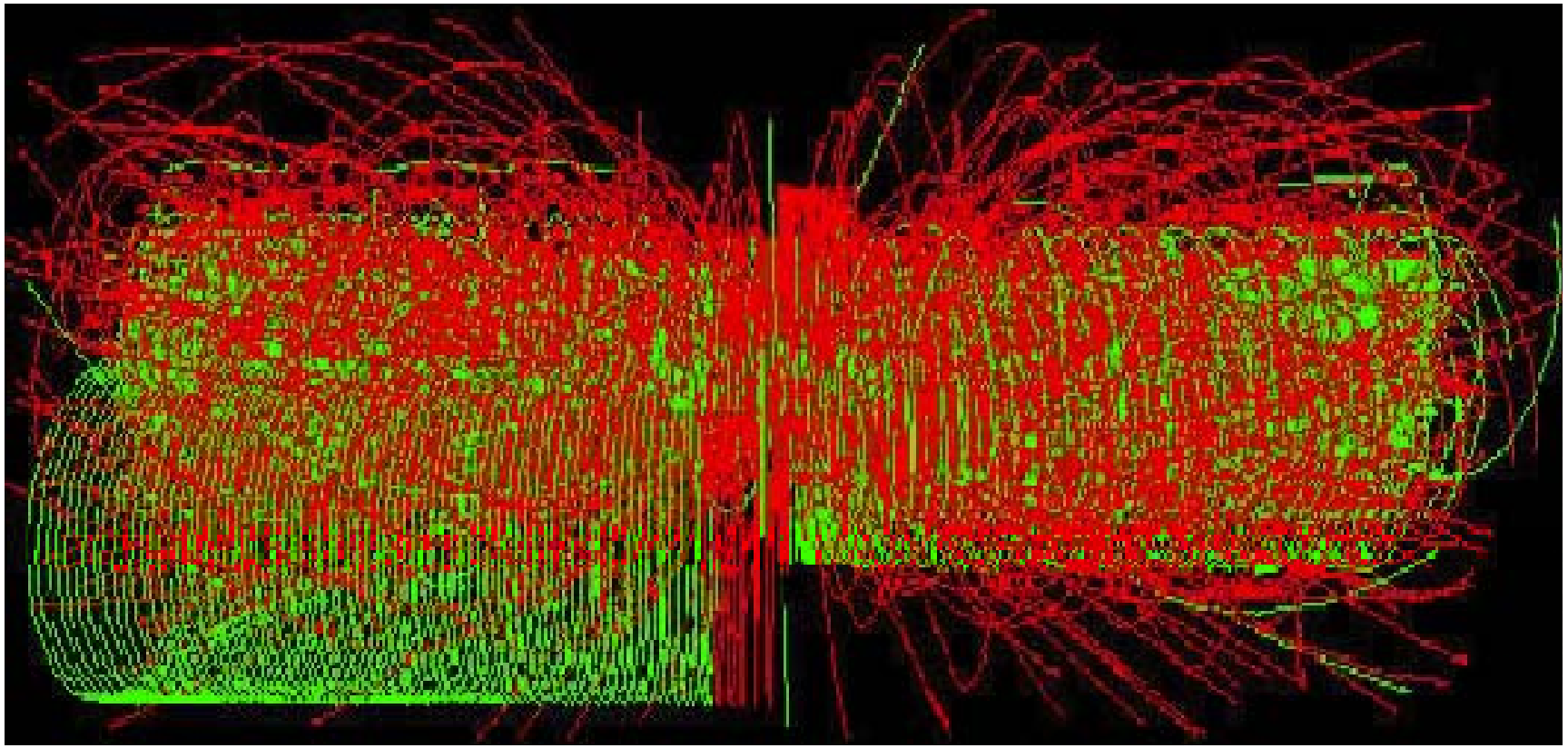
CERN IT/GS

Disclaimer:

For any incorrect or missing ☹ information,

it's only me that should be blamed!

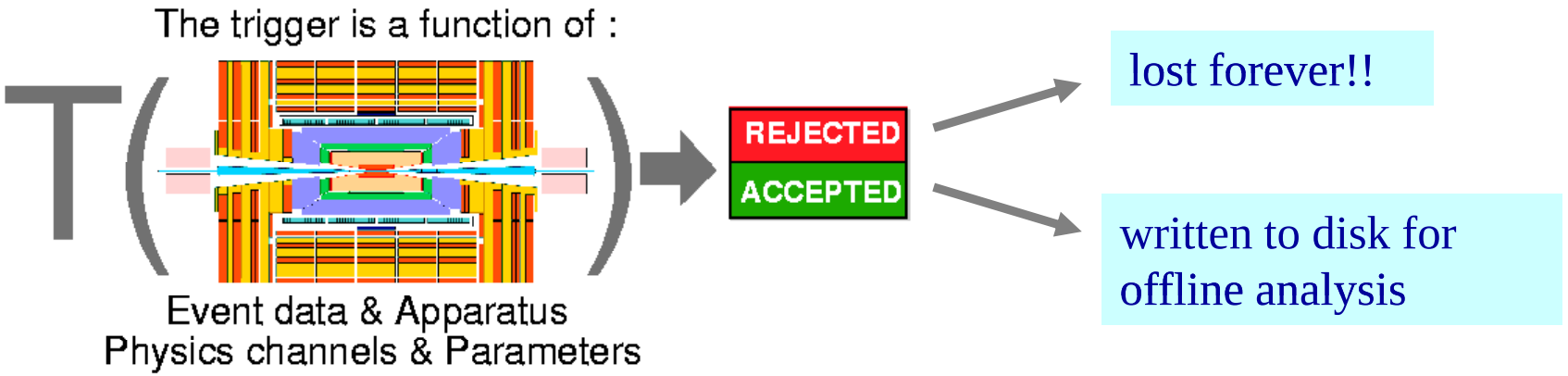
“THE” Challenge



Higgs → 4 μ

The Data selection

25 ns bunch spacing → collision rate 40 MHz
write ~ @200 Hz events to storage (technological and budget limits)
→ need a factor > 10^5 online rejection



Multi-level system:

- Level1 uses (fast) signals from calorimeters and Muon Spectrometers, hardware-based: 40 MHz → ~ 40 kHz
- Higher-Level-Triggers use fast software algorithms: ~ 40 kHz → ~ 200 Hz
  Selection based on high p_T signatures: jets, leptons, photons, …

Remind you:

- Expected Higgs event rate 1 out of 10^{13} interactions

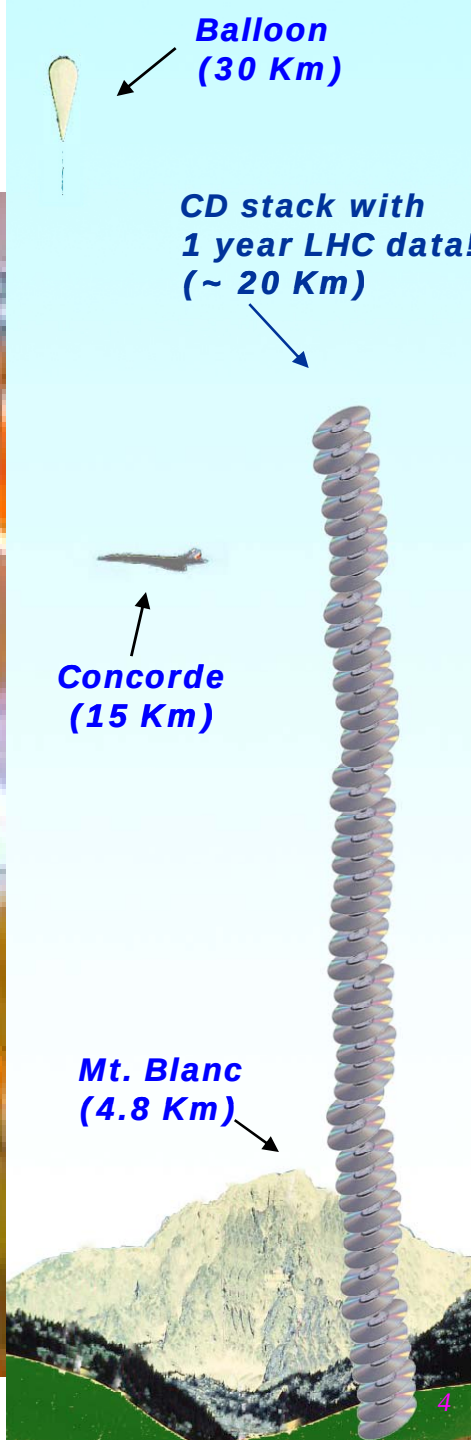
The Data Volume

- ATLAS raw data:

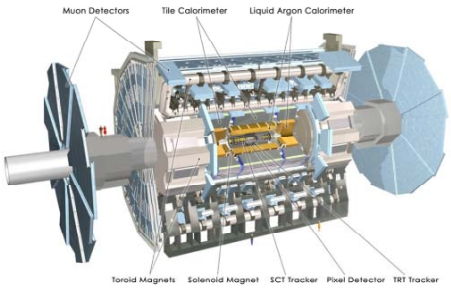
  ~ 3.2 PB / year

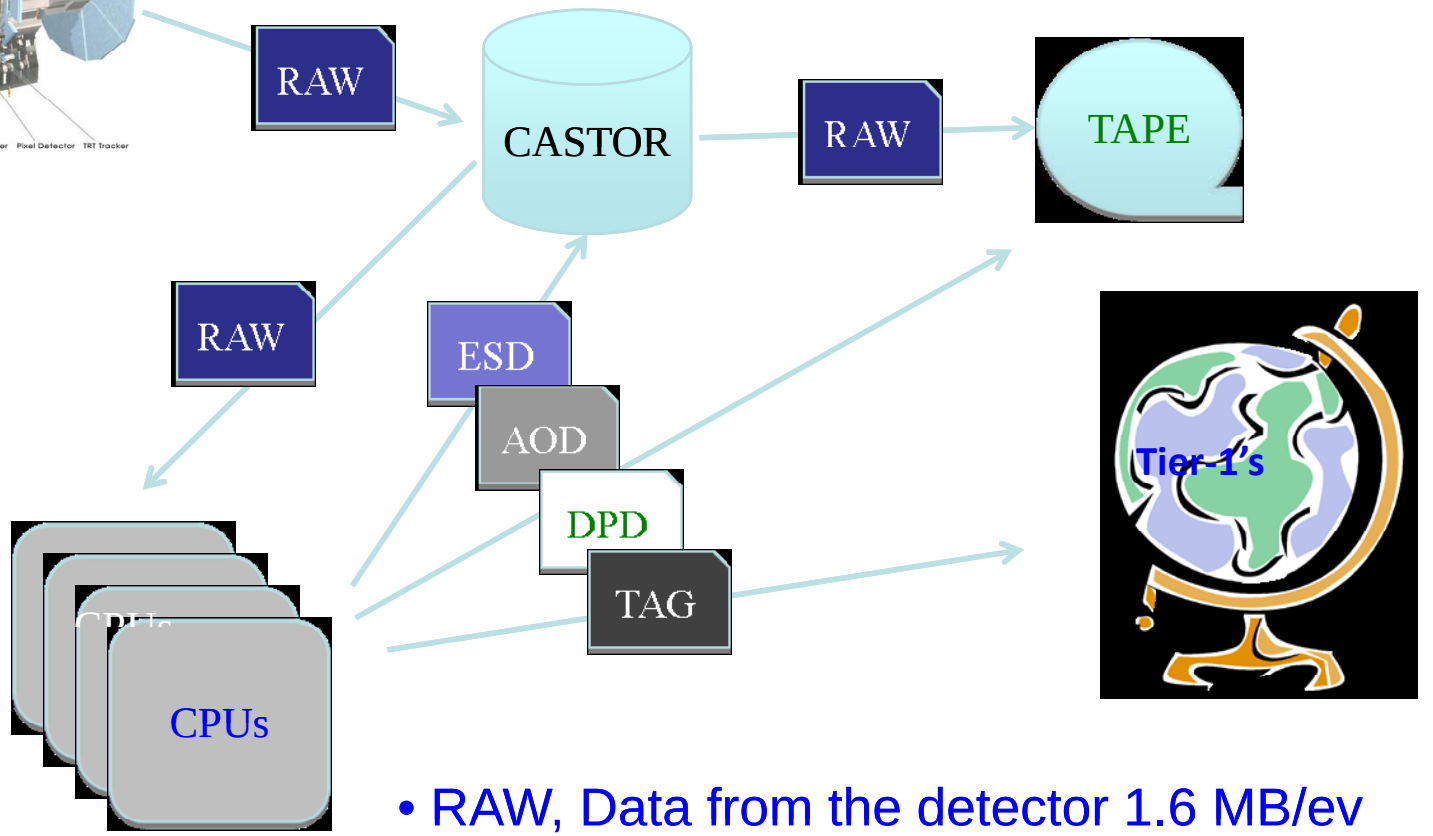
  - 200 day (day = 50k sec)
  - 200Hz,
  - RAW event = 1.6MB

  – 320 MB/s of RAW



The Data Types at the Tier-0





- RAW, Data from the detector 1.6 MB/ev
- ESD, Event Summary Data 1.0 MB/ev
- AOD, Analysis Object Data 0.2 MB/ev
- DPD, Derived Physics Data 0.2 MB/ev
- TAG, Data tag 0.01 MB/ev

The ATLAS Computing

- Data distribution

- Data simulation and reprocessing

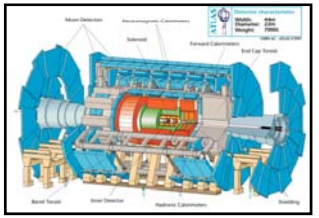
- Data analysis

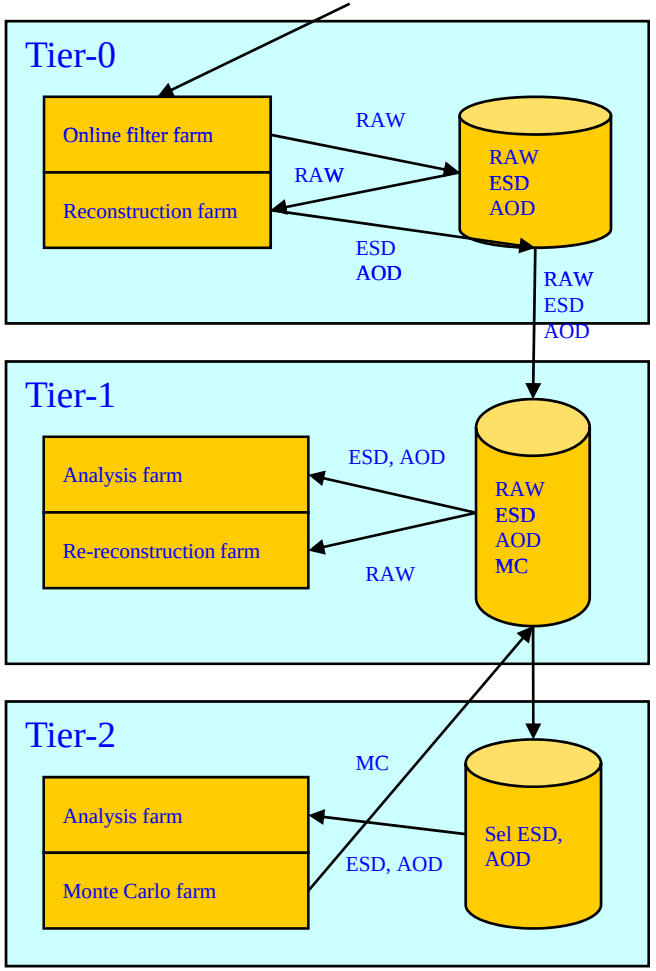
….. and of course the Monitoring

The ATLAS Computing Model

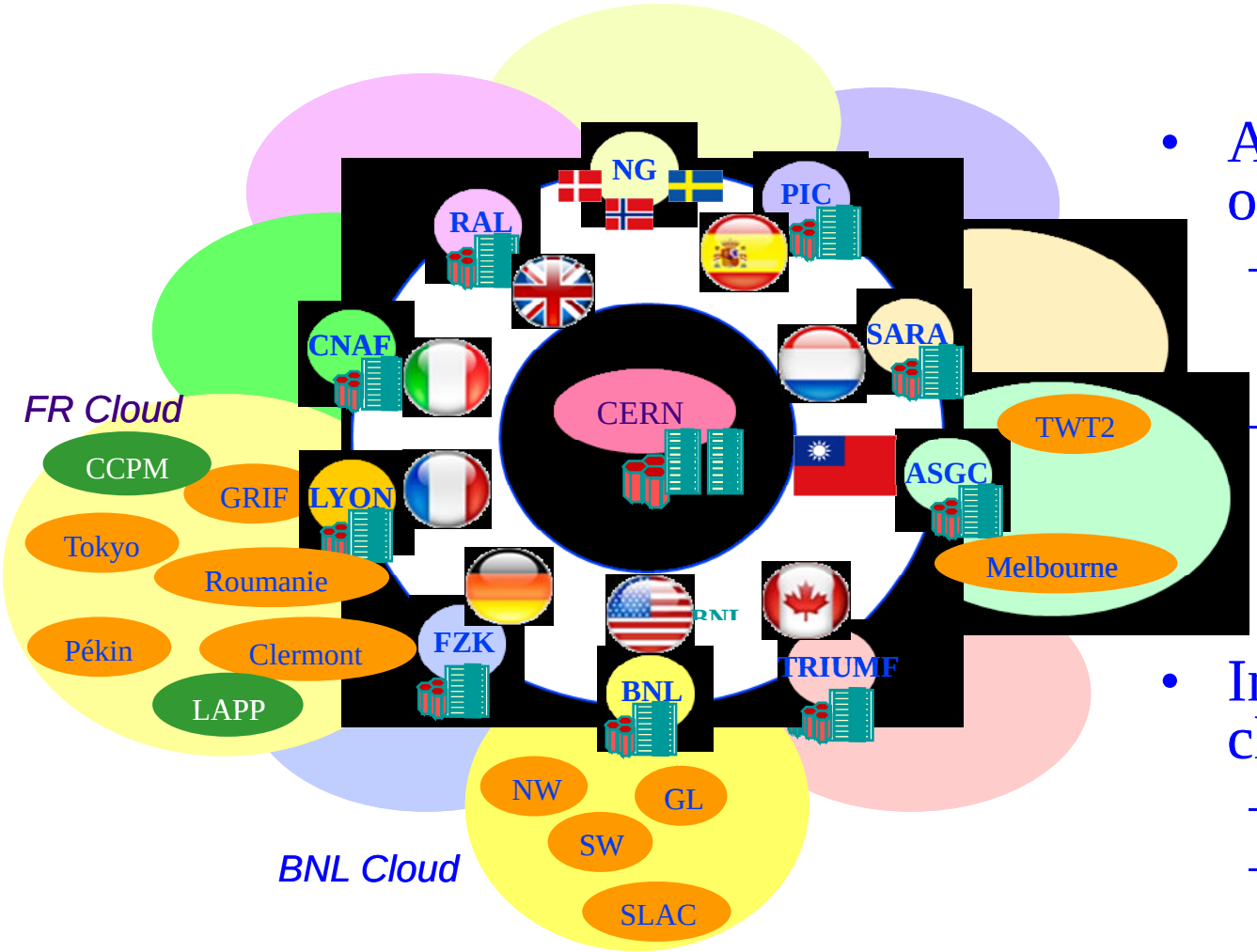
- ATLAS computing model is GRID oriented
  - High level of decentralization

- Sites are organized in a multi-tier structure
  - Hierarchical model
  - Tiers are defined by **ROLE** in the experiment computing model
    - **Tier-0 at CERN**
      - Record RAW data
      - Distribute second copy to Tier-1s
      - Calibrate and do first-pass reconstruction
    - **Tier-1 centers**
      - Manage permanent storage – RAW, simulated, processed
      - Capacity for reprocessing, bulk analysis
    - **Tier-2 centers**
      - Monte Carlo event simulation
      - End-user analysis





ATLAS Tiers: the CLOUD model

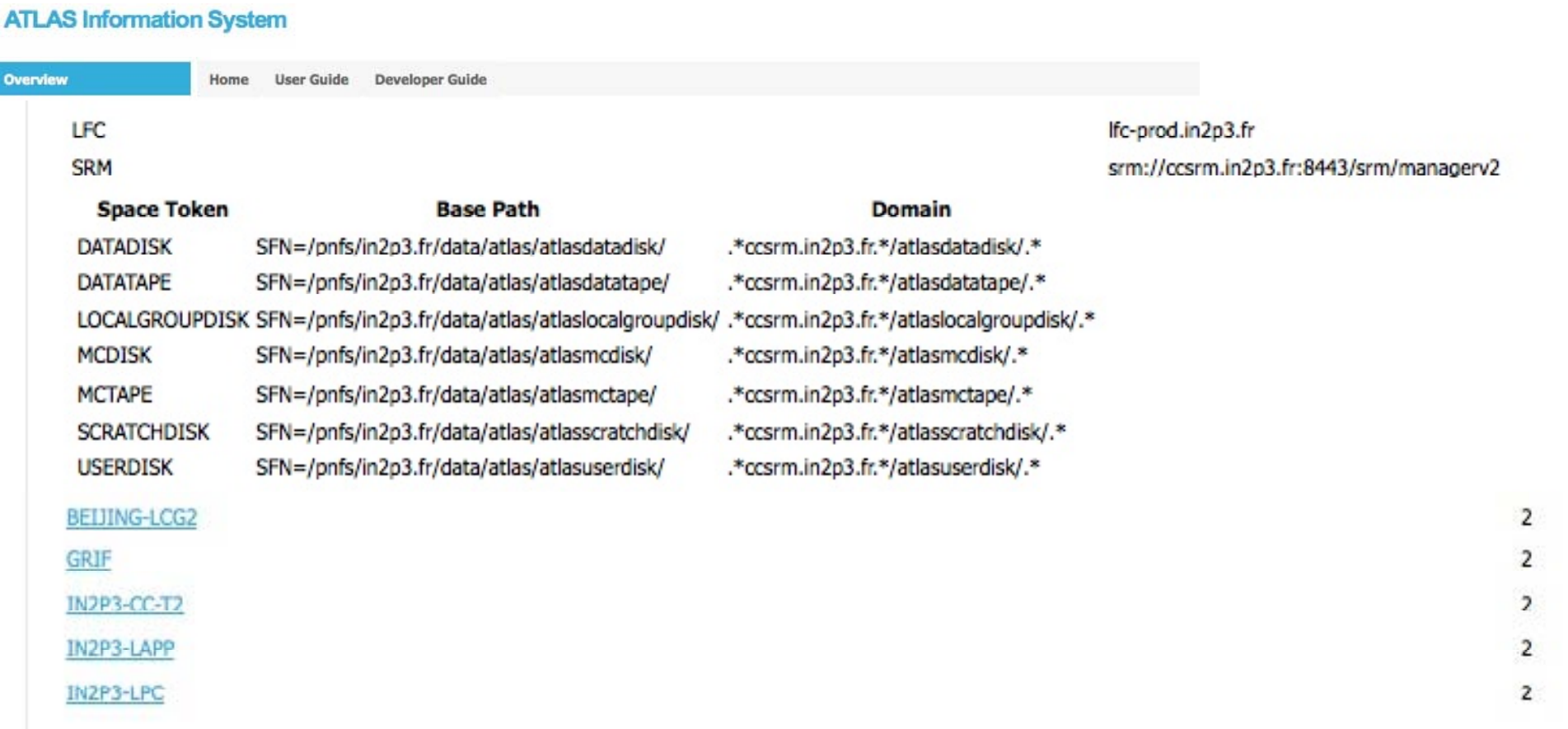


- ATLAS sites are organized in clouds
  - Every cloud consists of a T1 and several T2s
  - Most clouds are defined from geography or founding
    - Not really a rule
- Implications of the cloud model
  - Services deployment
  - Support

AGIS: The ATLAS Grid Information System

The ATLAS Information System provides static and semi-static information about resources, services and topology of the ATLAS grid.

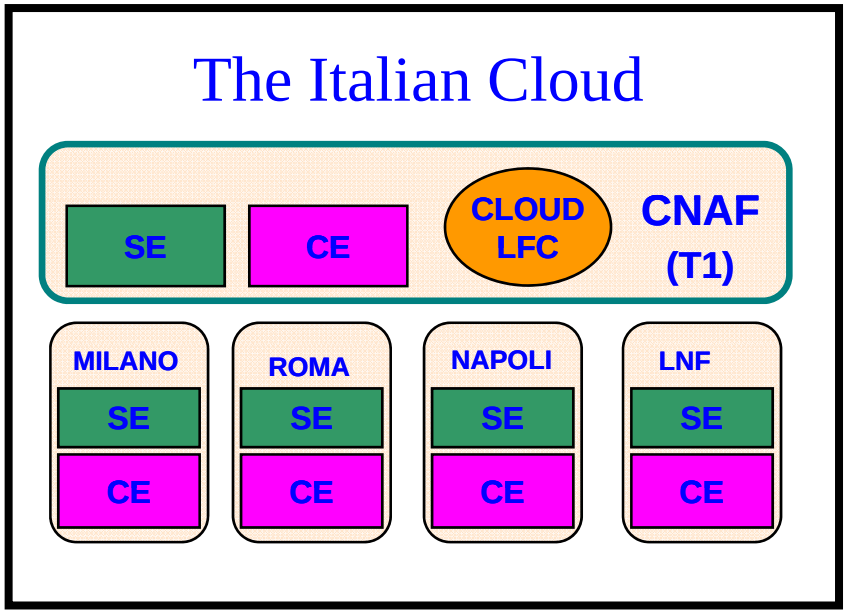
- Complementary to the BDII



Catalogs and Storages in the Cloud Model

- .. i.e how are services really deployed?
- Every site (T1 and T2) hosts a Storage Element

- The Local File Catalog:
  - Relies on the LCG File Catalog (LFC) middleware
  - One LFC per cloud (at T1)
    - Contains info about all files in the T1 and all T2s of the cloud
    - Purely a deployment strategy.

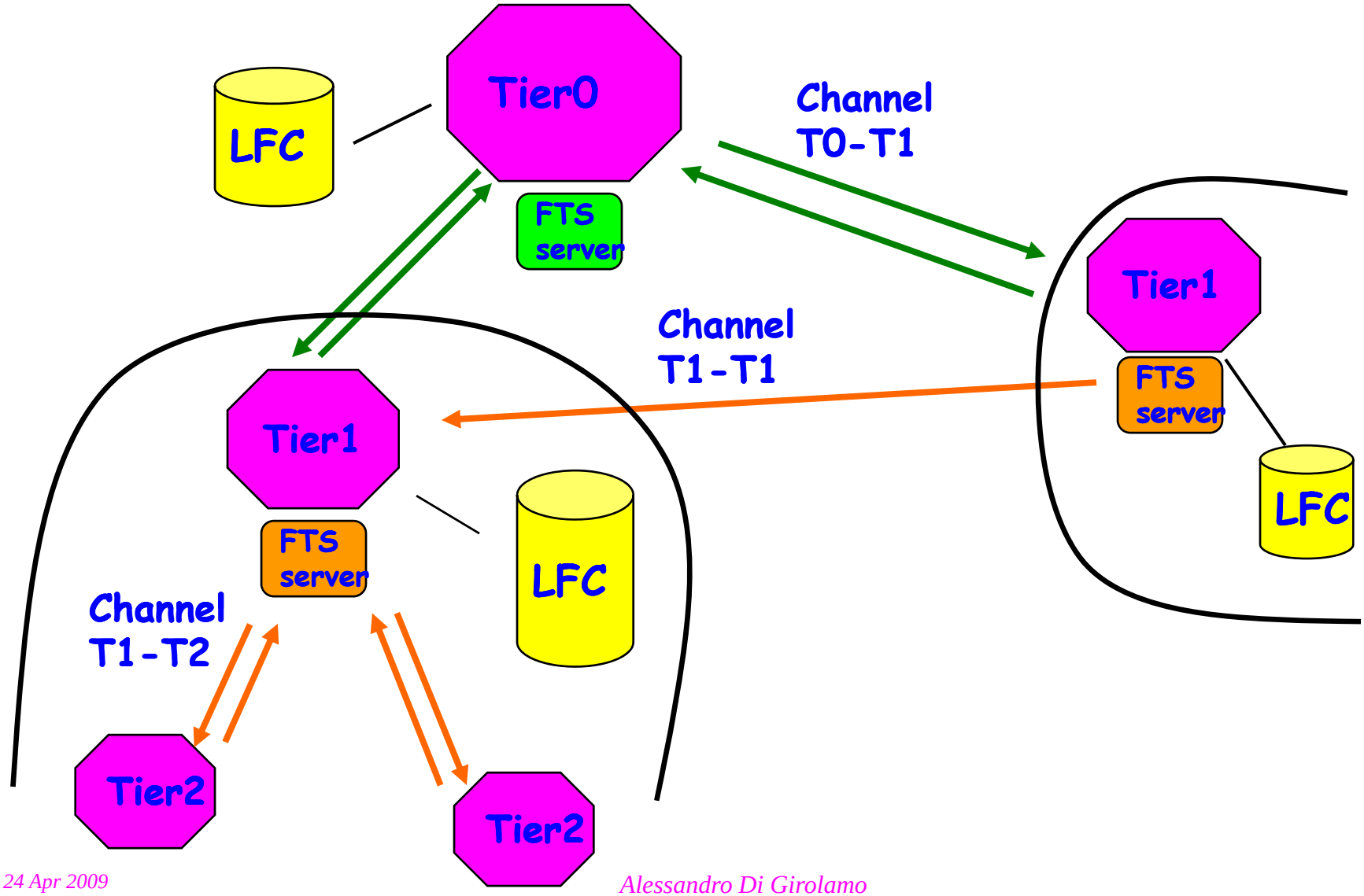


SRM and Space Tokens

- Many Storage Elements implementation
  - Some offer disk-only storage, other offer a gateway to mass storage systems
- The Storage Resource Manager (SRM) offers a common interface to all storages
  - GridFTP is the common transfer protocol
  - Storage specific access protocols
- SRM comes with Space Tokens
  - Partitioning of storage resources for different activities
- A DDM "site" is identified by the Grid Site name + the storage space token

'CERN-PROD_DATADISK':
{[…],'srm': 'token:ATLASDATADISK:srm://srm-atlas.cern.ch:8443/srm/managerv2?SFN=/castor/cern.ch/grid/atlas/atlasdatadisk/',

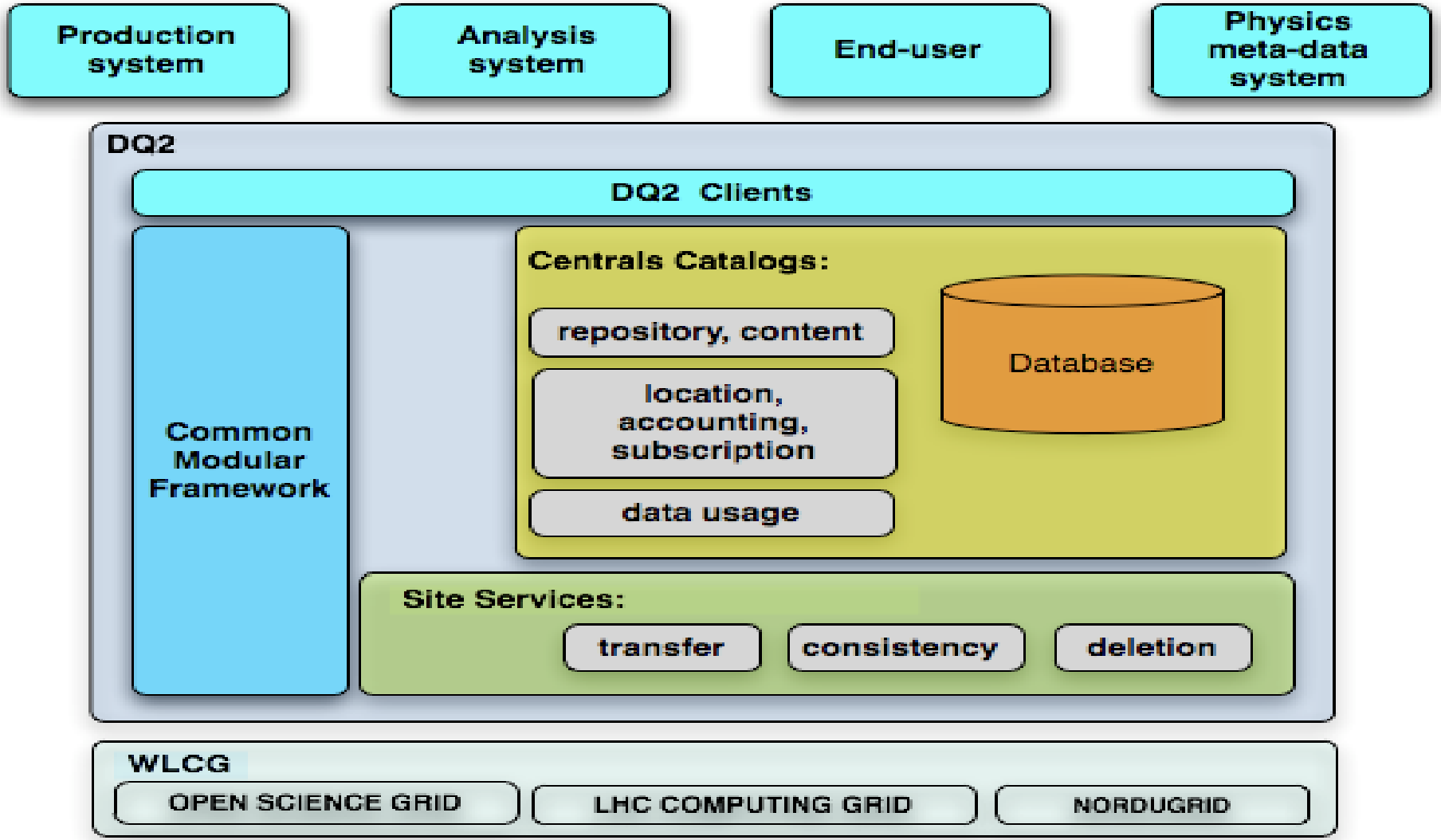
FTS and Data Movement schema



FTS and Data Movement

- FTS is a point to point File transfer service
  - One FTS server per cloud
- FTS channels are defined for “privileged paths”
  - Could be associated with privileged physical networks
  - Other transfers happen via normal network routes
  - No site multi-hops
- The FTS channel at T1 of cloud X defines channels for
  - T1(X)-T2(X) and T2(X)-T1(X)
  - T1s-T1(X)
  - *-T1(X) and *-T2s(X)
  - CERN-T1s are served from CERN FTS

The DDM stack



The DDM in a nutshell

The Distributed Data Management …

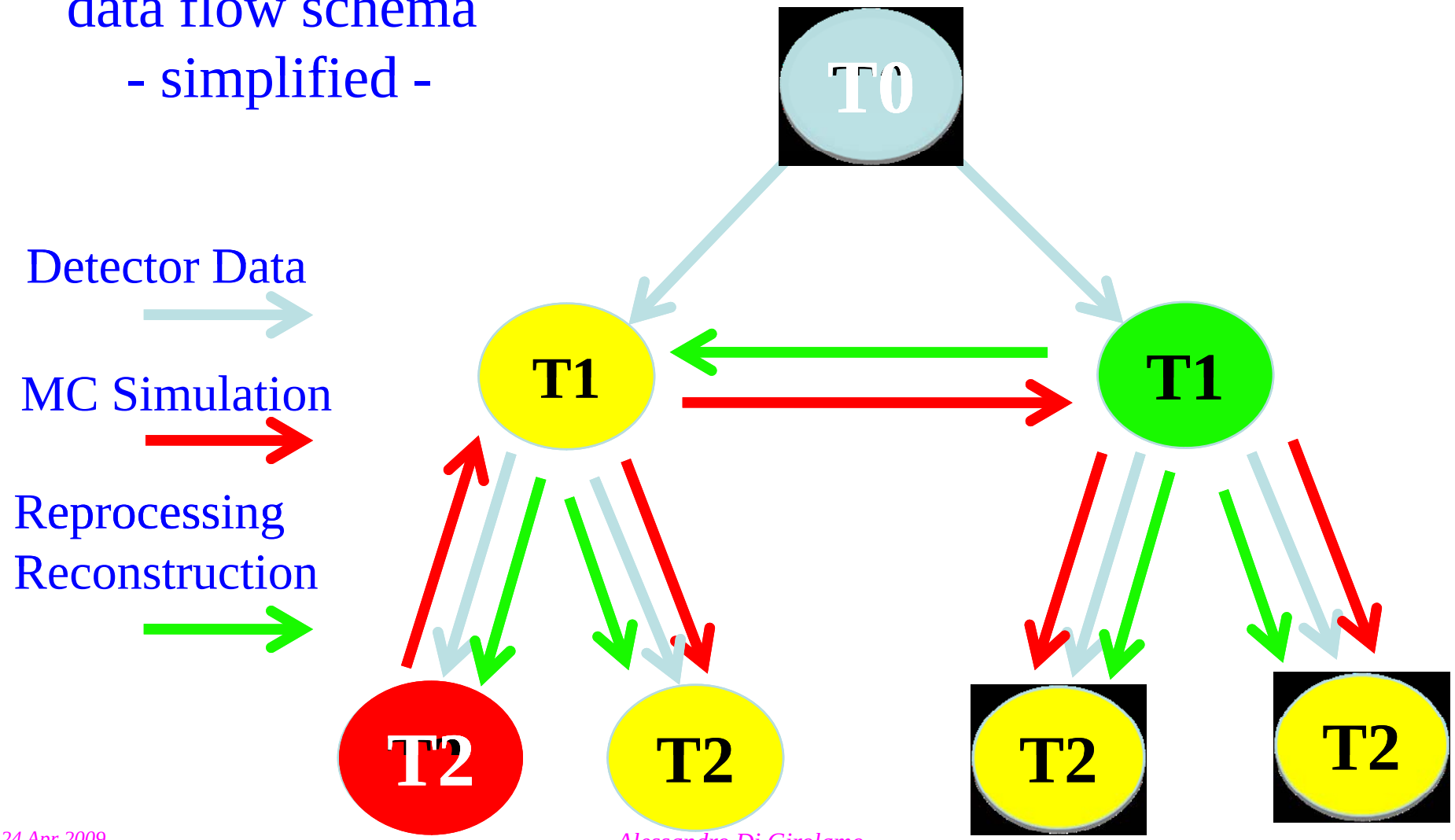
- … enforces the concept of dataset
  - Units of data placement and replication
- … based on a subscription model
  - Datasets are subscribed to sites
  - A series of services enforce the subscription
    - Lookup data location in LFC
    - Trigger data movement via FTS
    - Validate data transfer

Clients and APIs

- Command Line and Python APIS exist for all basic operations
  - Created datasets, register subscriptions, delete datasets etc …
- High level tools allow users to
  - Upload a dataset in DDM (dq2-put)
  - Download a dataset from DDM (dq2-get)
  - List content of a DDM dataset (dq2-ls)

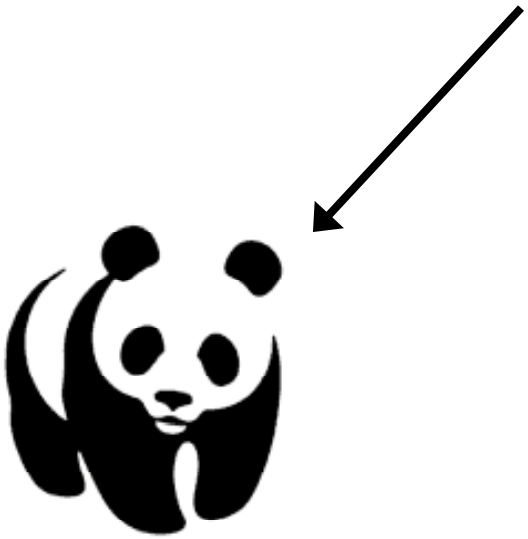
DDM Exports and Consolidation

data flow schema
- simplified -



ATLAS production system

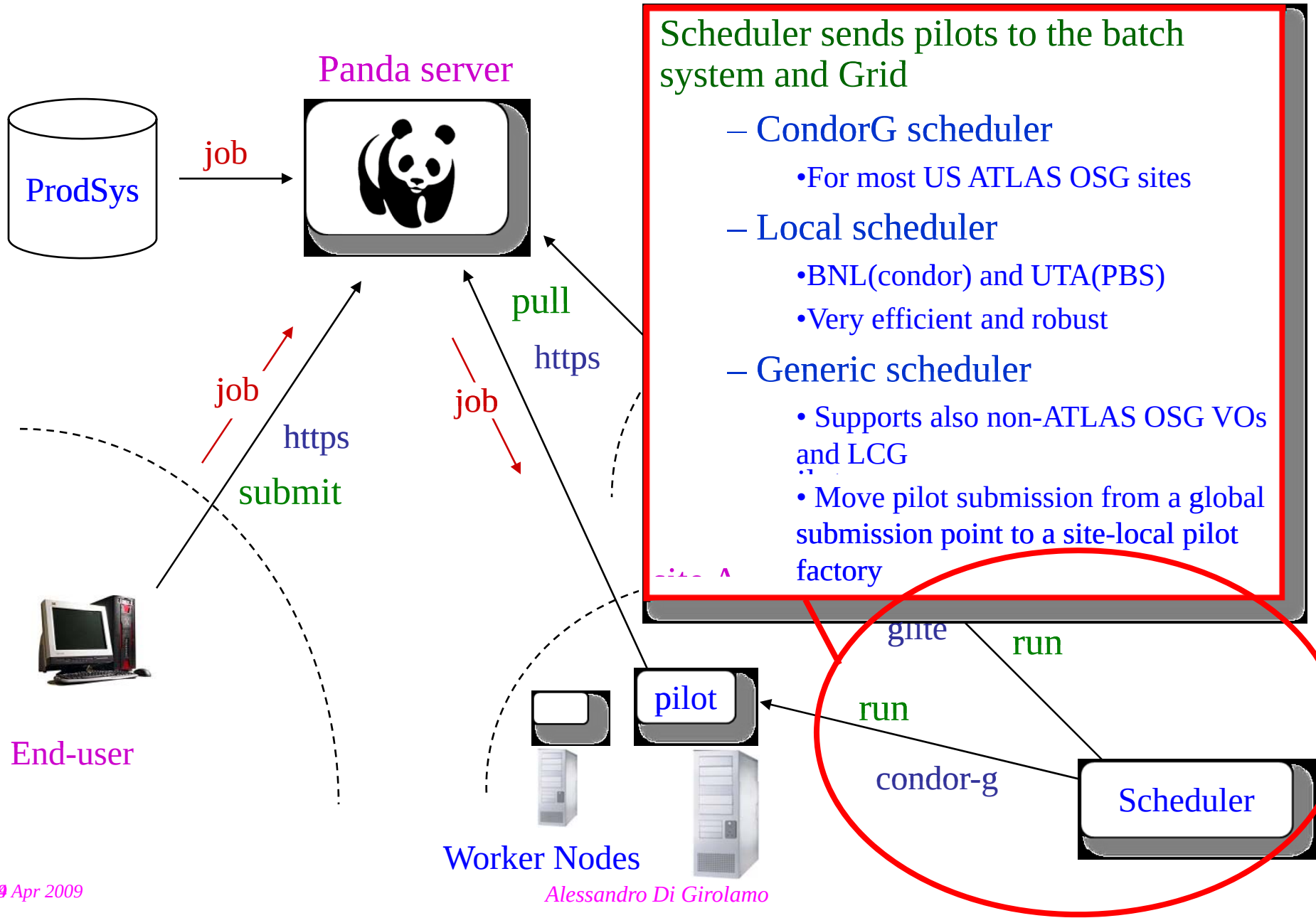
Requirements



- throughput
- scalability
- robustness
- efficient resource utilization
- minimal operations manpower
- tight integration of data management with processing workflow

- Works both with OSG and EGEE middleware
- pilot based system
  - Apache-based central server
  - Pilots retrieve jobs from the server as soon as CPU available, hence low latency
  - Tightly integrated with the DDM system (in/out datasets)

PANDA


Panda server
ProdSys
job
Scheduler sends pilots to the batch system and Grid
– CondorG scheduler
•For most US ATLAS OSG sites
– Local scheduler
•BNL(condor) and UTA(PBS)
•Very efficient and robust
– Generic scheduler
• Supports also non-ATLAS OSG VOs and LCG
• Move pilot submission from a global submission point to a site-local pilot factory
pull
https
job
job
https
submit
End-user
pilot
gLite
run
run
condor-g
Scheduler
Worker Nodes

How the pilot works

- Sends the several parameters to Panda server for job matching (HTTP request)
  - CPU speed
  - Available memory size on the WN
  - List of available ATLAS releases at the site
- Retrieves an “activated” job (HTTP response of the above request)
  - activated → running
- Runs the job immediately (all input files should be already available at the site)
- Sends heartbeat every 30min
- Copy output files to local SE and register them to Catalog

PANDA security

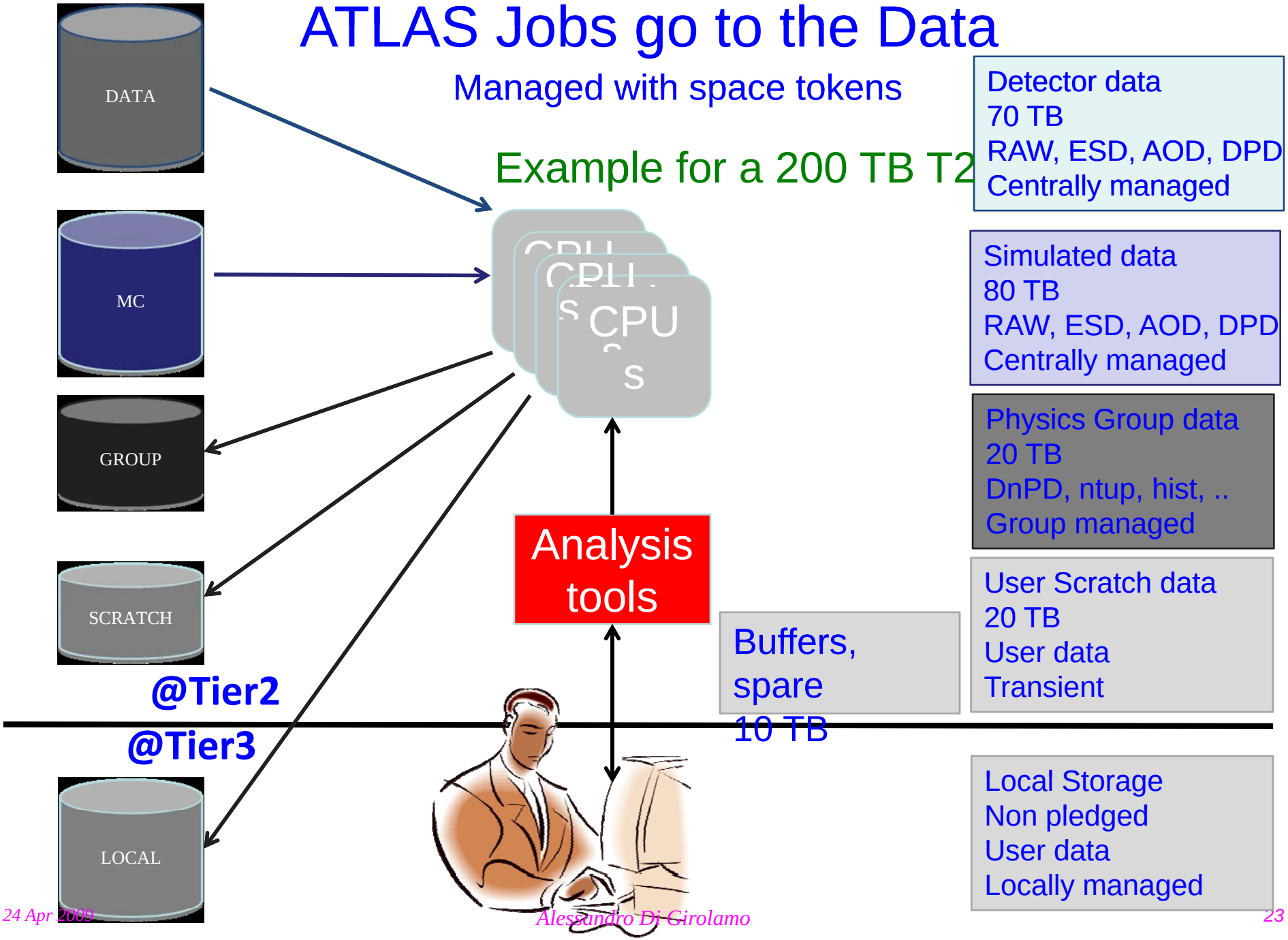
- Panda services use std GSI grid security model of authentication and authorization based on X509 grid certificates
- Proxy's VOMS attributes are checked
- Production job execution and file management relies on Role=production certificates
- Analysis jobs run under production proxy unless gLExec is implemented in identity switching mode
  - gLExec based identity change on WN to submitter identity for user jobs under testing (proxy management done by MyProxy)
  - Security issues have been investigated and clarified for ATLAS
    - gLExec is considered mature

User analysis in Tier2's

- Mostly done on DPD's
  - But also on AOD's with/out TAGs
  - On ESD's for performance (need to be downloaded from Tier1 first)
  - On RAW for special studies (need to be downloaded from Tier1 first)
- Output can be D2PD's or D3PD's or ntuples or histograms
- User can only temporarily store data on the grid (in SCRATCH)
- Permanent storage possible for physics and performance groups (in GROUP)
- User may have permanent storage in (its) Tier3 (in LOCALGROUP)
- Analysis jobs automatically sent to
  - where data resides, and
  - where free CPU's available
- 50% of CPU capacity for analysis and 50% for Simulation Production
  - Analysis shares up to the site to be set

ATLAS Jobs go to the Data

Managed with space tokens



Distributed Analysis tools

- Ganga
  - Runs everywhere and on everything
  - Developed on LCG
  - Can submit jobs to WMS and to PANDA
  - Only for Analysis
- pAthena
  - Requires the panda pilot job framework: does not run at all sites yet
  - Developed on OSG
  - Pull mode (Pilot jobs)
  - Fully integrated with DQ2
    - Outputs (libraries, logs, ntup, etc..) are DQ2 datasets
  - For Production and Analysis
  - At the present moment does NOT support gLExex

ATLAS distributed Analysis

submit

pAthena

job

job

job

pull

job

pull

site A

WN

submit

job

Ganga

job

job push

Site B

WMS

job

WN

The ATLAS DDM Dashboard

ATLAS dashboard | Data: All Activities | Jobs: Production | Jobs: Analysis | Panda: Production | SLS: Central Services

Overview | Dataset Info | Page Help | User Guide | Feedback

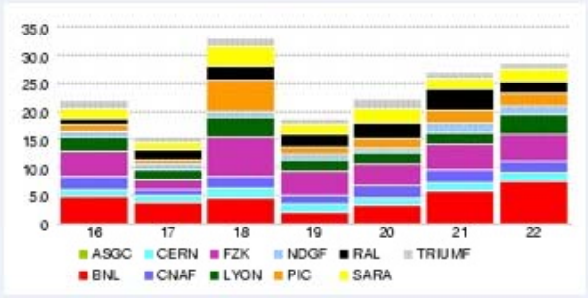
OVERVIEW

Activity Period
- Activity in Last Hour
- Activity in Last 4 Hours
- Activity in Last 24 Hours
- Activity in Last 7 Days
- Activity in Last 30 Days
- Activity in ...

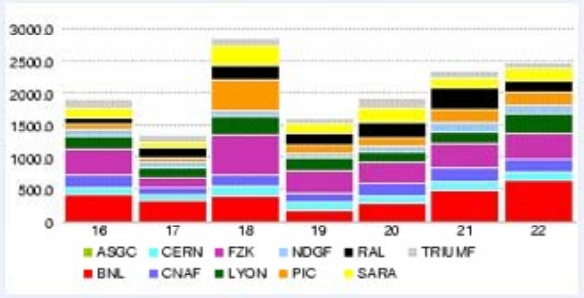
Selected Activities
- MC Production
- Data Export
- Functional Test
- User Subscriptions
- Staging

Selected Cloud
- ASGC Cloud
- BNL Cloud
- CERN Cloud
- CNAF Cloud
- FZK Cloud
- LYON Cloud
- NDGF Cloud
- PIC Cloud
- RAL Cloud
- SARA Cloud
- TRIUMF Cloud

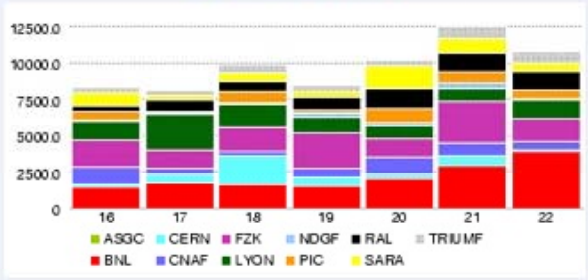
Throughput (MB/s)



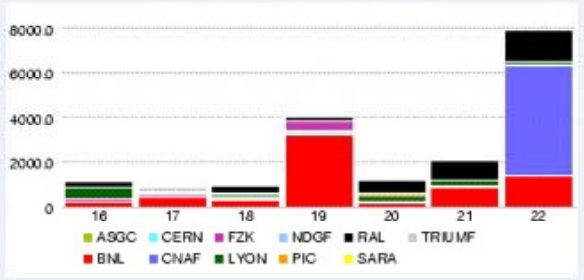
Data Transferred (GBytes)



Completed File Transfers



Total Number Transfer Errors

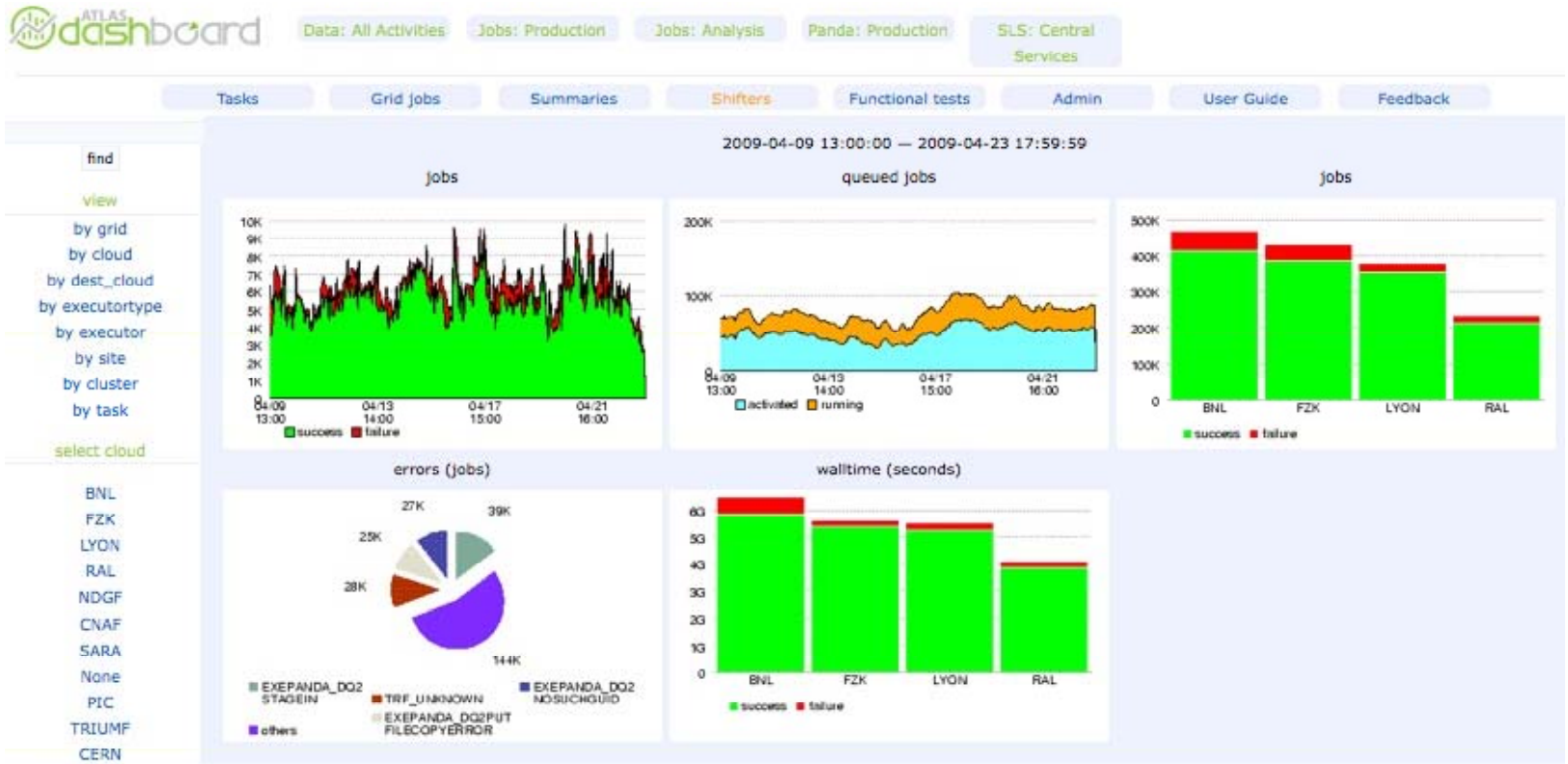


Activity Summary ('*2009-04-16 00:00*' to '*2009-04-23 00:00*')

Click on the cloud name to view list of sites

	Transfers			*Registrations*		*Errors*			*Services*
Cloud	*Efficiency*	*Throughput*	*Successes*	*Datasets*	*Files*	*Transfer*	*Registration*	*Services*	*Grid*
ASGC	100%	0 MB/s	1	1	1	0	0	0	
BNL	71%	6 MB/s	19385	29073	603662	8038	52	0	
CERN	98%	2 MB/s	7471	1853	137884	137	0	0	
CNAF	67%	3 MB/s	10528	11971	215490	5112	0	0	
FZK	94%	6 MB/s	15328	22152	356256	987	0	0	
LYON	89%	3 MB/s	11426	15434	241734	1420	0	0	
NDGF	98%	2 MB/s	1575	3017	63781	24	0	0	
PIC	88%	3 MB/s	4886	10017	144363	681	0	0	
RAL	70%	3 MB/s	9031	17155	322257	3876	0	0	
SARA	99%	3 MB/s	6012	8747	174300	58	0	0	

The ATLAS Production Dashboard



cloud	defined	assigned	waiting	activated	running	holding	transferring	success	failure	efficiency
BNL	0	5599	0	11361	6774	1529	6237	411465	56628	87.9%
FZK	0	56	0	6542	4331	389	2345	386765	43152	90%
LYON	0	16	0	7413	5474	1068	2656	354437	25421	93.3%
RAL	0	7	0	10516	5016	403	2058	212225	21052	91%
NDGF	0	0	0	0	3399	13	6	138778	6846	95.3%
SARA	0	0	0	265	641	433	370	93010	34284	73.1%
CNAF	0	298	0	2329	716	116	947	109773	16157	87.2%
None	1030	124	439	15260	1595	397	868	60086	36092	62.5%
PIC	0	63	0	768	678	276	719	61235	5991	91.1%
TRIUMF	0	0	0	2449	902	442	361	36045	17266	67.6%
CERN	0	10	0	51	14	1	0	2661	308	89.6%
ASGC	0	0	0	0	20	403	0	0	0	-
total	*1030*	*6173*	*439*	*56954*	*29560*	*5470*	*16567*	*1866480*	*263197*	*87.6%*

CRITICAL WARNING NORMAL GOOD NO_ACTIVITY

The Panda Monitoring

CERN monitor | Production | Clouds | DDM | PandaMover | AutoPilot | Sites | Analysis | Physics data | Usage | Plots | ProdDash | DDMDash

1 min old Update

Not logged in. List users

Panda monitor
Times are in UTC

Panda info and help

Panda Production Operations Dashboard

Panda shift guide calendar mailing list
ADCoS twiki calendar mailing list
Production task support mailing list

Click for help

Jobs - search
Recent running, activated, waiting, assigned, defined, finished, failed jobs
Select analysis, prod, install, test jobs

Quick search
Job
Dataset
Task request
Task status
File

Summaries
Blocks: days
Errors: days
Nodes: days
Daily usage

Tasks - search
Generic Task Req
EvGen Task Req
CTBsim Task Req
Task list
New Tag
Bug Report

Datasets - search
Dataset browser
Aborted MC datasets
Panda subscriptions
Dataset Popularity

Datasets Distribution
DDM Req
Req list
AODs
EVNTs
Conditions DS
DB Releases
SIT pacballs
Validation Samples
Functional Tests
ATLAS Data
FDR_Datasets
Reprocessed_Datasets

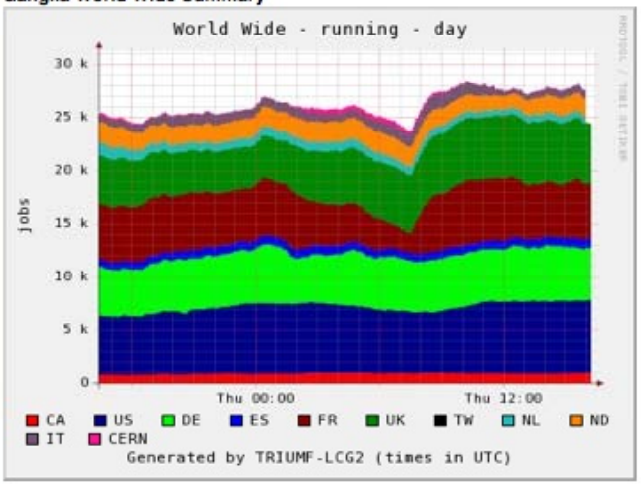
Servers: CERN:OK BNL:Failed Logger:OK Bamboo:OK

Show space available at sites

Active tasks: CA:31 CERN:2 DE:21 ES:11 FR:33 IT:28 ND:19 NL:11 UK:18 US:203
Bamboo task brokerage, job submissions, status over last 12 hours

Jobs updated >12 hrs ago: activated:25912 running:none
Jobs updated >36 hrs ago: transferring:561

Ganglia World Wide Summary



Cloud efficiency history

Production job summary, last 12 hours (Details: errors, nodes)

Cloud Information	Nodes	Jobs	Latest	Pilots (3hrs)	defined	assigned	waiting	activated	sent	running	holding	transferring	finished	failed	tot	trf	other
Overall Production	4643	3762	04-23 16:20	35012	1 / 0	5132 / 0	31 / 0	23729 / 0	3 / 0	11909 / 0	1198 / 0	8789 / 561	30229 / 0	1881 / 0	6%	0%	6%
CA	23	268	04-23 14:36	0	0	0	0	1425	0	0	0	17 / 17	0	134	100%	0%	100%

The Service Level Status

for the ADC Central Services

Service Level Status overview

Home | Search | KPIs | Tags | Admin | Documentation | Help

ATLAS Distributed Computing Central Services 5 Nov 2008 Wed 12:48:53

ADC Central Services

availability:
(more)

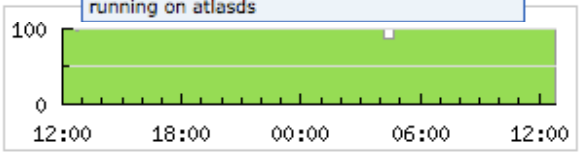
percentage: 100%

status: **available**

this service consists of:

- ATLAS_DDM_VOBOXES
- ATLAS_CC
- ATLAS_DDM_Tracker
- ATLAS_DDM_Deletion

ATLAS_DDM_Deletion
Availability: **100%**, available
NumError=33, Storage|Catalog Cleaner=1914,
NumDone=231,
running on atlasds



Additional information

full name: **ATLAS Distributed Computing Central Services**

short name: ADC Central Services

group: IT-GS

email: **atlas-project-adc-operations@cern.ch**

web site: https://twiki.cern.c...

alarms page: https://prod-grid-lo...

service managers: **Birger Koblitz**, Alessandro Di Girolamo

Availability update

last update: 12:42:52, 5 Nov 2008 (6 minutes ago)

expires after: 77 minutes

rss feed with status changes

http://sls.cern.ch/sls/service.php?id=ADC_CS

Admin

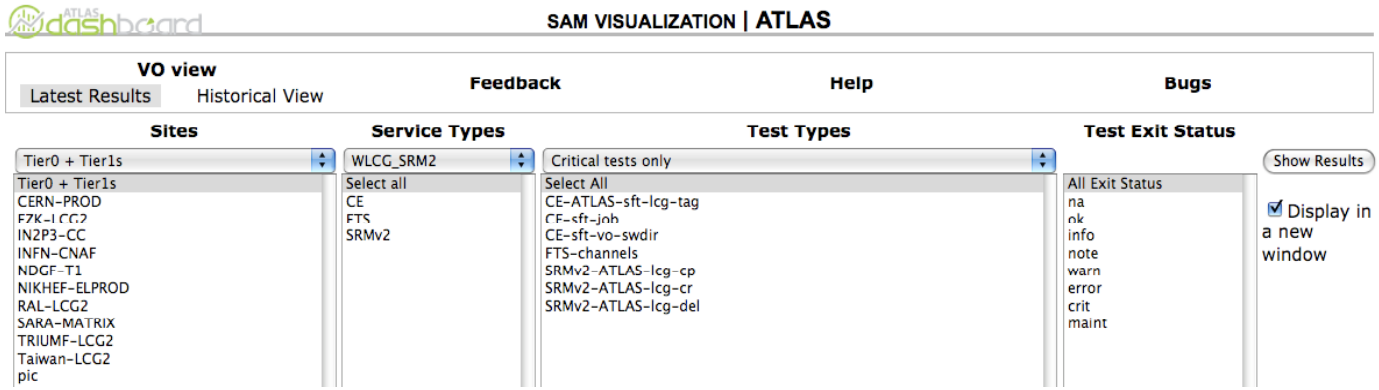
admin tools

Service Availability Monitor

The ATLAS Critical Test

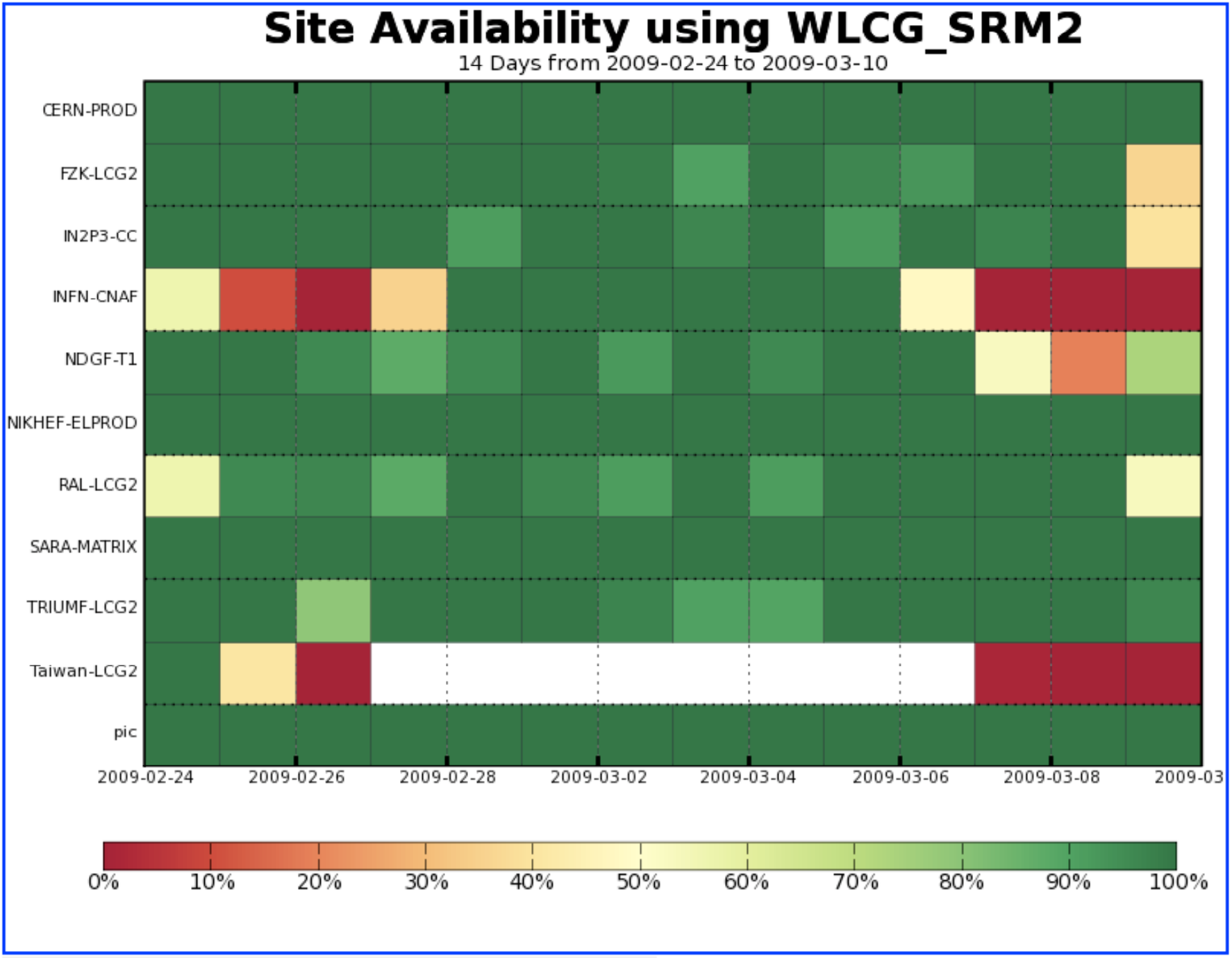
- SRM
  - For each spacetoken
    - Copy and register, copy back, delete (lcg-cr/cp/del)
- CE
  - Job submission, software area, lcgtag
- LFC
  - ping, ls
- FTS
  - check channels list

- Other NOT critical tests launched

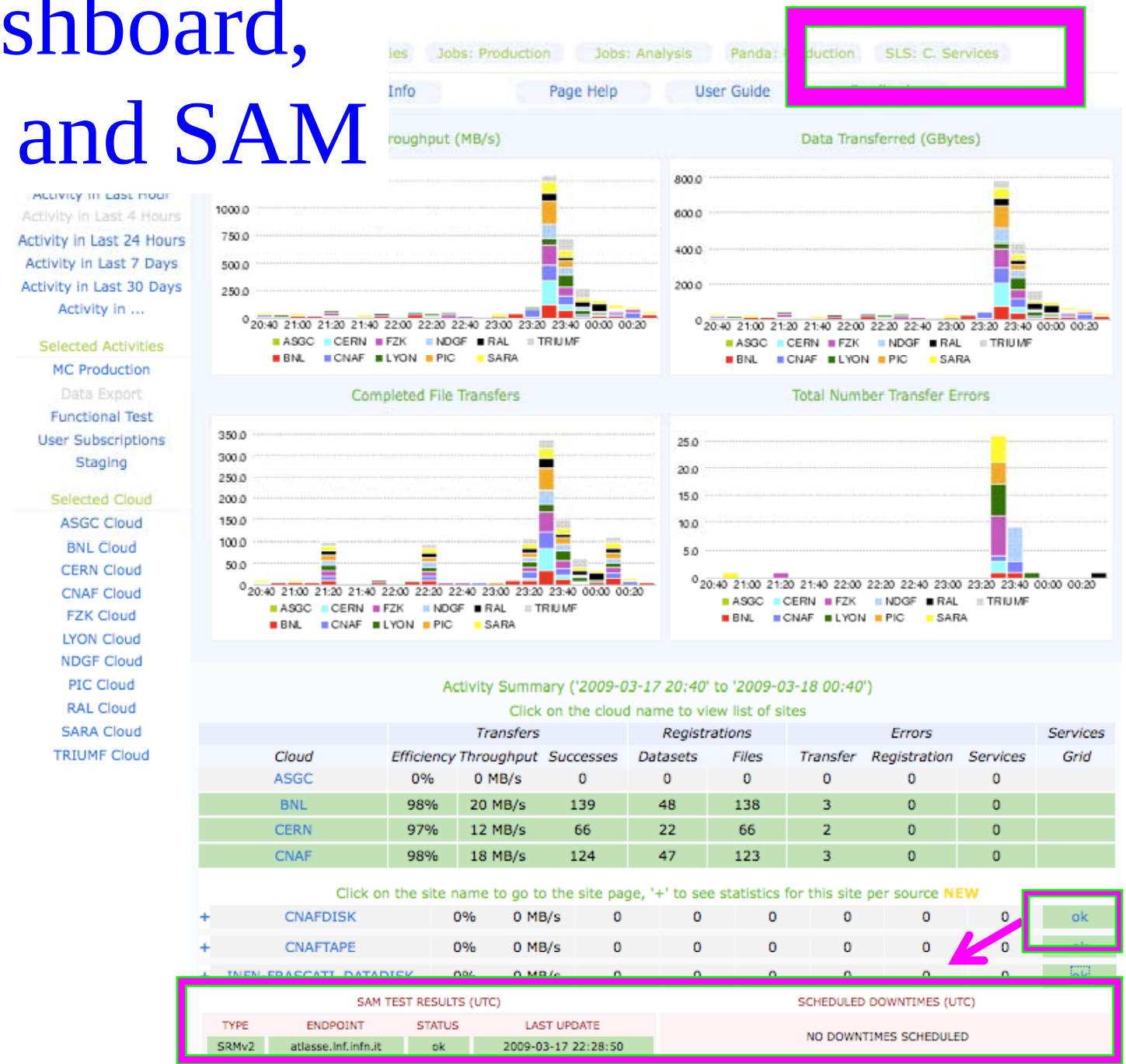


SAM latest results

Sitename	Service Type	Service Name	atlas-lcgtag	js	swdir	ftschn	atlas_srm2_cp	atlas_srm2_cr	atlas__srm2_del
		ce127.cern.ch	ok	ok	ok				
		ce128.cern.ch	ok	ok	ok				
		ce129.cern.ch	ok	ok	ok				
	FTS	fts-t0-export.cern.ch				ok			
		fts-t2-service.cern.ch				ok			
	SRMv2	srm-atlas.cern.ch					ok	ok	ok
FZK-LCG2	CE	ce-1-fzk.gridka.de	ok	ok	ok				
		ce-2-fzk.gridka.de	ok	ok	ok				
		ce-3-fzk.gridka.de	ok	ok	ok				
		ce-4-fzk.gridka.de	ok	ok	ok				
		ce-5-fzk.gridka.de	ok	ok	ok				
	FTS	fts-fzk.gridka.de				ok			
		fts2-fzk.gridka.de				ok			
IN2P3-CC	CE	cclcgceli01.in2p3.fr	ok	maint	ok				
		cclcgceli02.in2p3.fr	ok	maint	ok				
	FTS	cclcgftsprod.in2p3.fr				ok			
	SRMv2	ccsrm.in2p3.fr					error	error	error

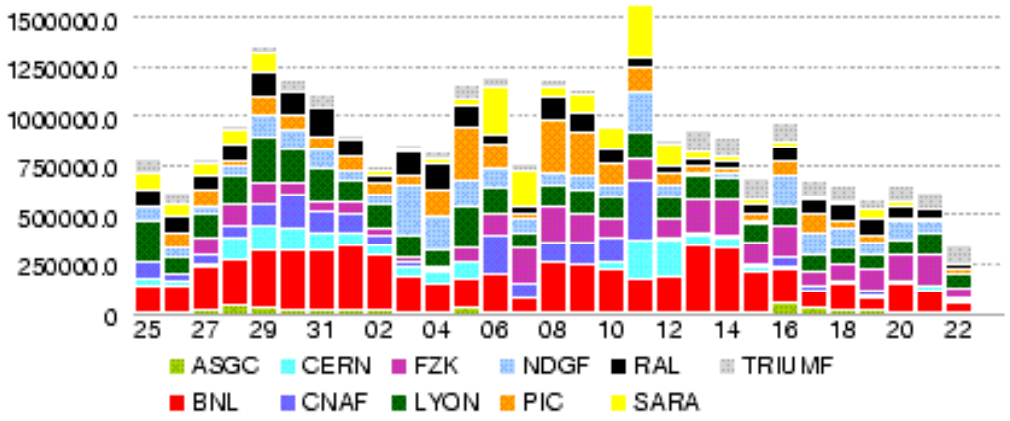

Site Availability using WLCG_SRM2
14 Days from 2009-02-24 to 2009-03-10
CERN-PROD
FZK-LCG2
IN2P3-CC
INFN-CNAF
NDGF-T1
NIKHEF-ELPROD
RAL-LCG2
SARA-MATRIX
TRIUMF-LCG2
Taiwan-LCG2
pic
2009-02-24
2009-02-26
2009-02-28
2009-03-02
2009-03-04
2009-03-06
2009-03-08
2009-03
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%

Dashboard, SLS and SAM

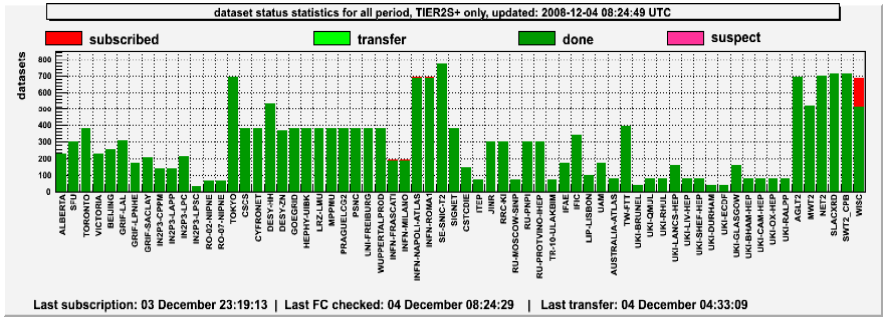


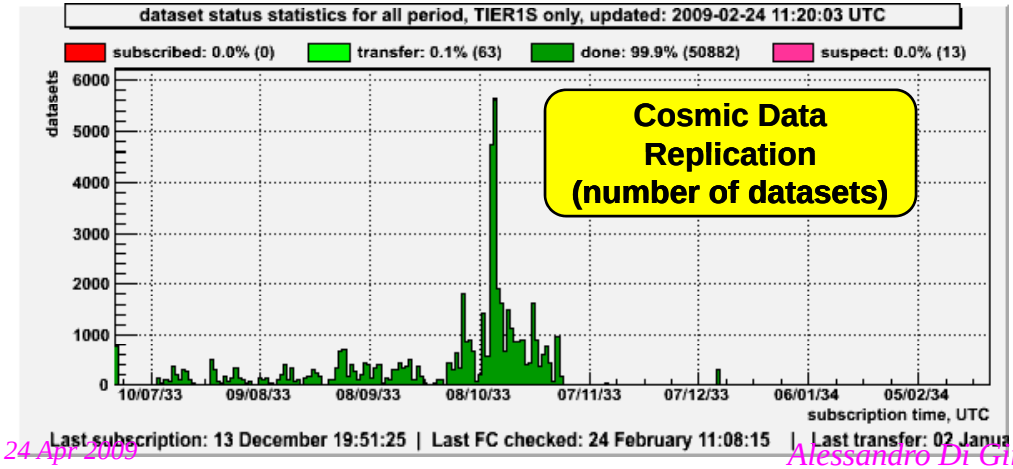
Data Exports and Consolidation

Number of replicated files over a one month period
850K files/day average, paek at 1.5M files/day



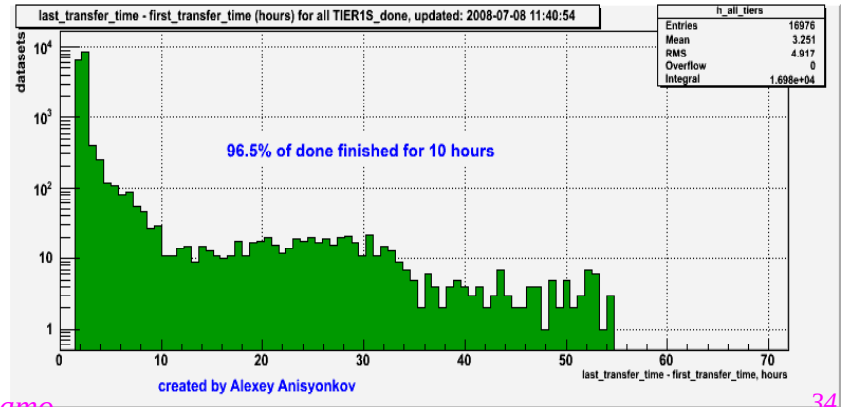
ATLAS Beam and Cosmics data replication to ~70 Tier2s





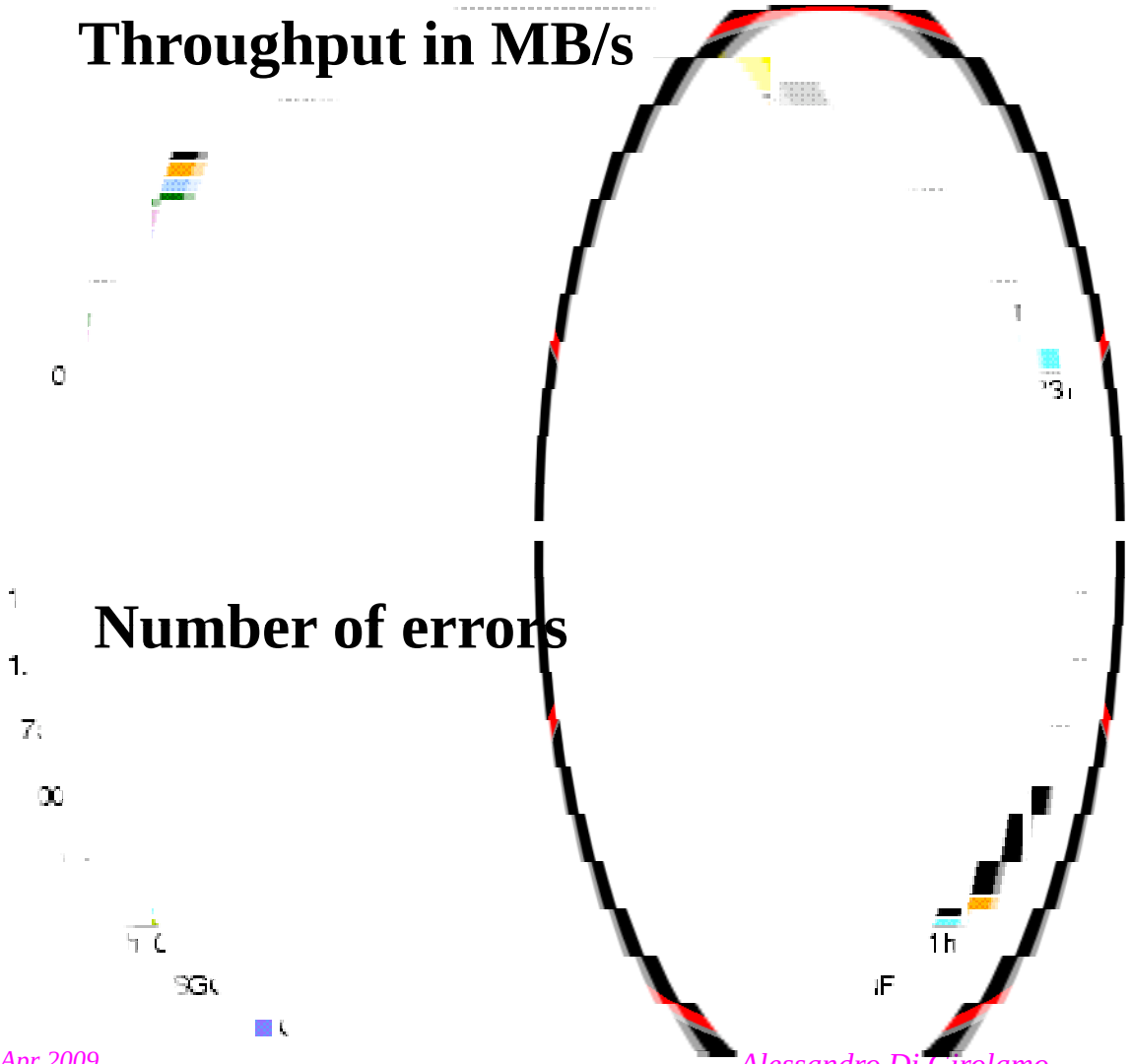
Cosmic Data Replication (number of datasets)

Cosmic Dataset replication time T0-T1s (completed transfers)
96.4% data replication is done (~92% within 3h)



Are we ready for data taking?

… ready? Yes, BUT…



- Interference between user activity and centralized data export operation
  - Overload of disk server
- But the user was not reading the detector data via the GRID…

The Challenge

- Support of user activities

difficult to simulate: real life will provide new challenges and opportunities

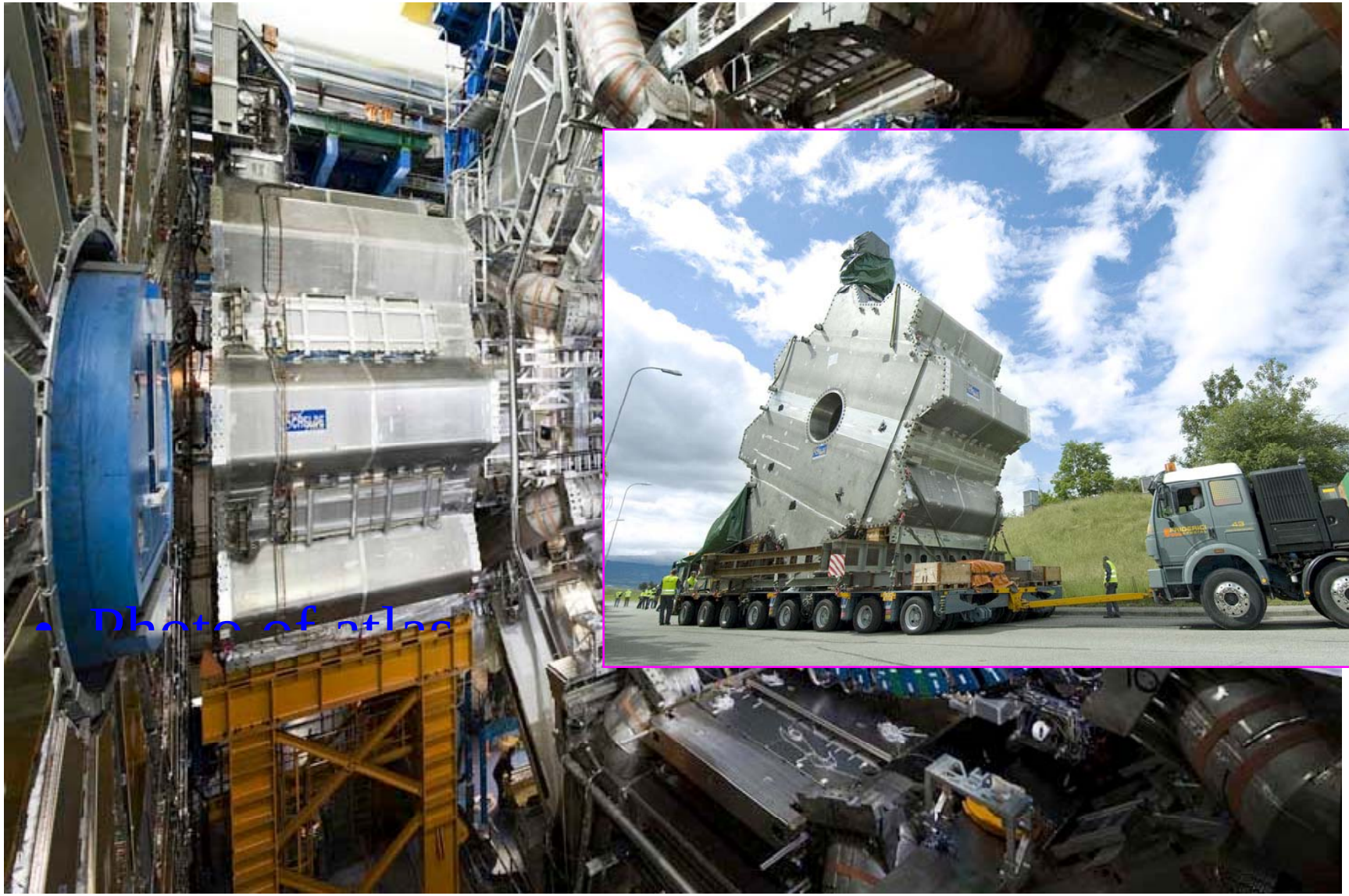
Questions?

Thanks to all (…those who I took these slides from) !

BackUp slides

- …

The ATLAS experiment

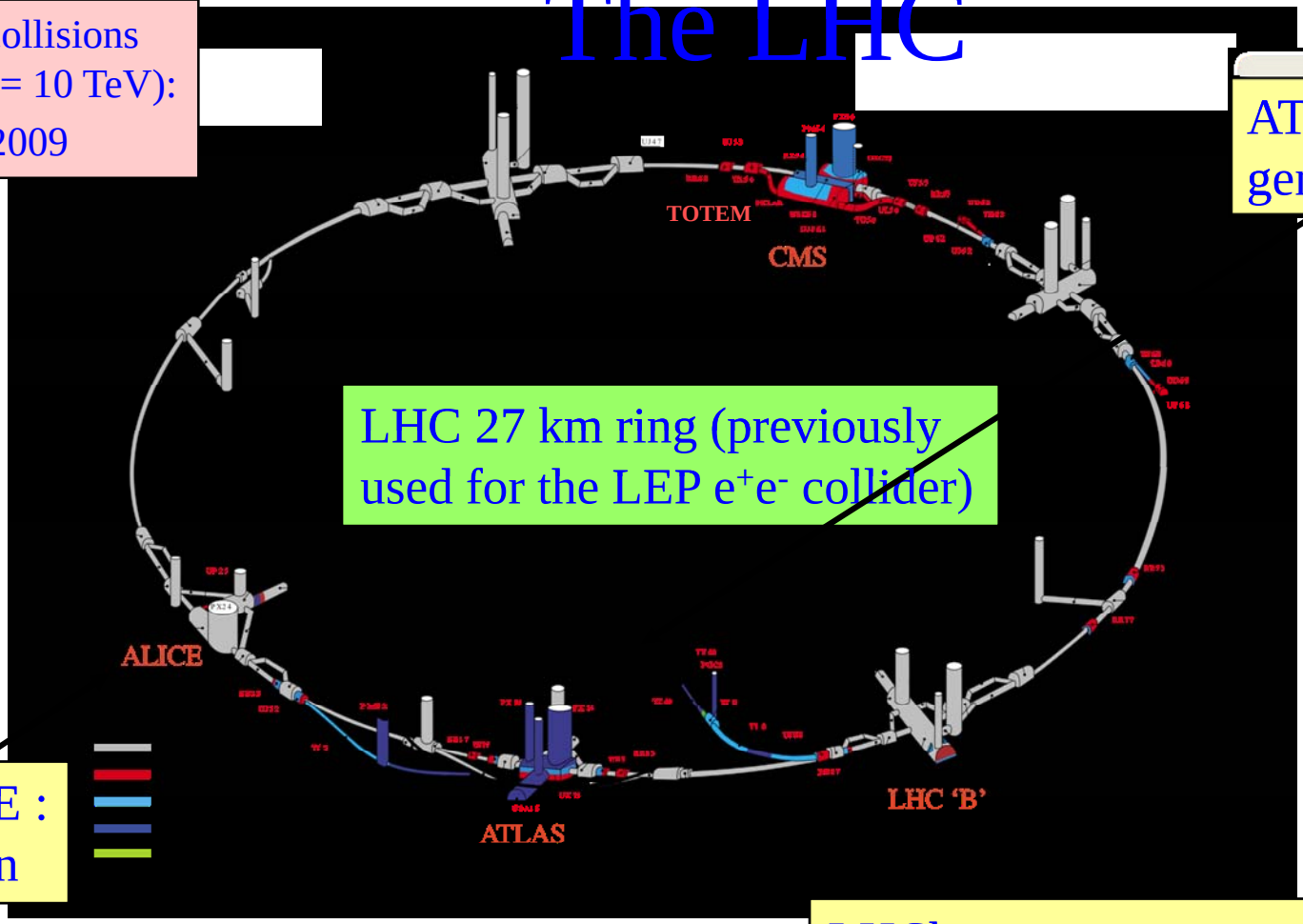


The Large Hadron Collider

The LHC

First collisions (at $\sqrt{s}$ = 10 TeV): Fall 2009

ATLAS and CMS : general purpose



LHC 27 km ring (previously used for the LEP e^+e^- collider)

ALICE : ion-ion

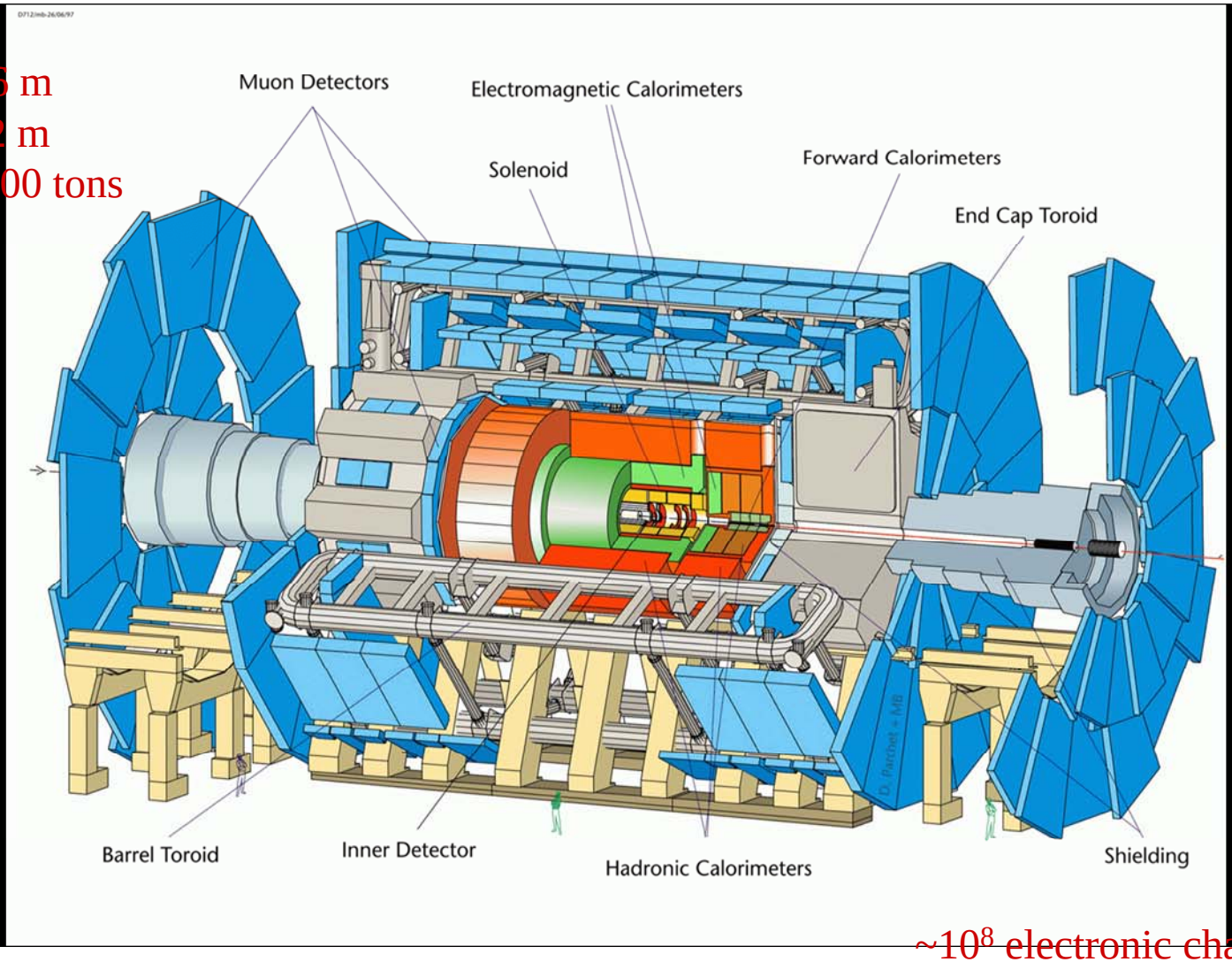
LHCb : pp, B-physics, CP-violation

- **pp: $\sqrt{s}$ = 14 TeV** $L_{design} = 10^{34}$ cm^{-2} s^{-1} (after 2011)
  $L_{initial}$ < few x 10^{33} cm^{-2} s^{-1} (before 2011)
  Note: $\sqrt{s}$ is x7 Tevatron, L_{design} is x30 Tevatron
- Heavy ions: (e.g. Pb-Pb at $\sqrt{s}$ ~ 1000 TeV)

The Cavern

The ATLAS detector

Length : ~ 46 m
Radius : ~ 12 m
Weight : ~ 7000 tons



~10^8 electronic channels
~ 3000 km of cables

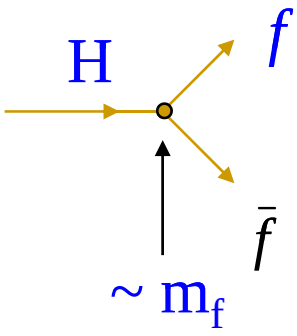
Why? (veryvery short!)

The Standard Model, although very successful, is an incomplete theory:
it's unable to answer many questions, it starts to fail at the TeV scale.

1) What is the origin of the particle masses ?

E.g. why $m_\gamma = 0$
$m_{W,Z} \approx 100$ GeV ?

SM : Higgs mechanism gives mass to particles



$m_H < 1$ TeV from theory

However:
- **Higgs not found yet**: only missing (and essential !) piece of SM
- **present limit : $m_H > 114.4$ GeV** (from LEP)

→ we need a machine to discover/exclude the Higgs particle over 115-1000 GeV

The only example of observed Higgs as of today …

Why we also don't like the Standard Model…

- The Standard Model needs to be "fixed" at the TeV scale
  - if Higgs -> some new physics is needed to stabilize its mass;
  - if !Higgs -> something else must prevent divergences at high E

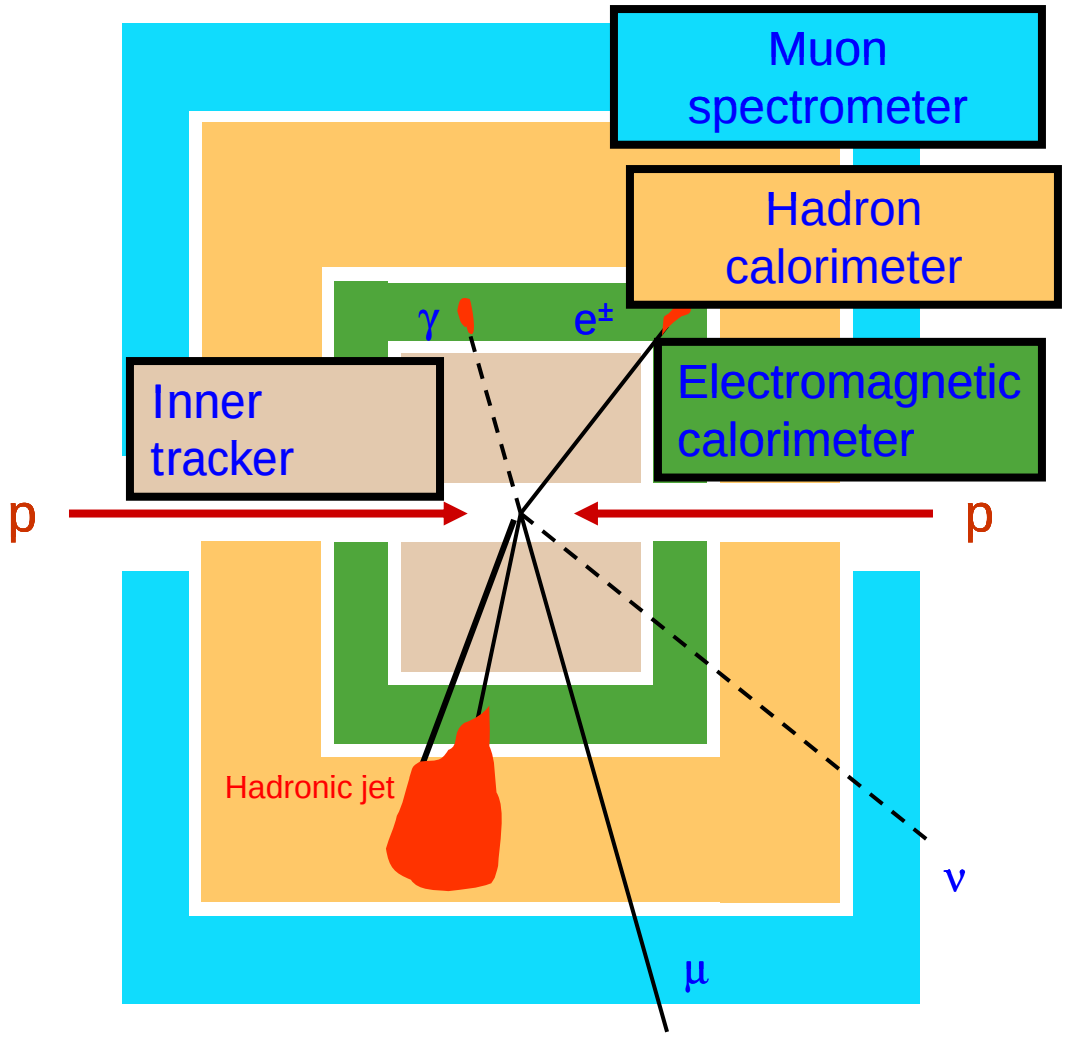
- Many other open questions
  - Why is the first family special ?
  - Are there additional (heavy) leptons and bosons ?
  - Are quarks and leptons really elementary ?
  - "Hierarchy" problem : why $M_{EW}/M_{Planck} \sim 10^{-17}$? Is there anything in between ?
  - What is the origin of the matter / anti-matter asymmetry in the Universe ?
  - Unification of coupling constants ?
  - What is the origin of ν masses ?
  - What is the composition of the Universe dark matter and the origin of dark energy ?
  - Why 3 fermion families?

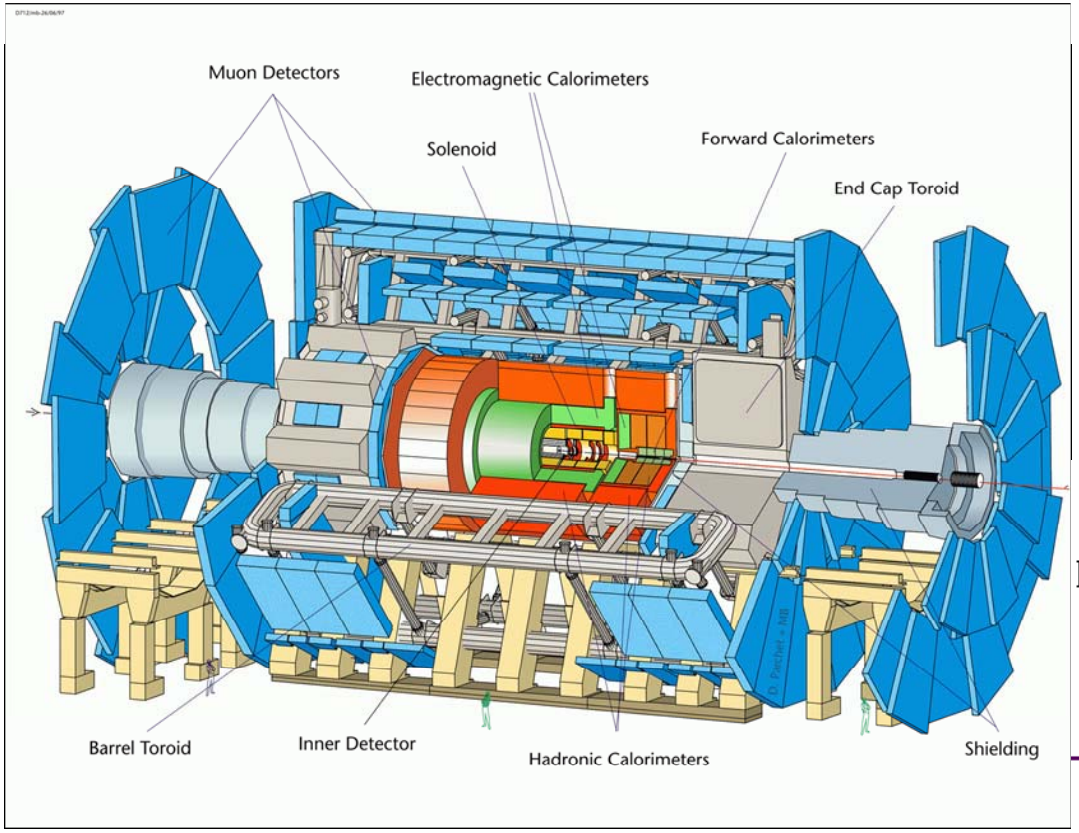
a general purpose detector

Don' t know how New Physics will manifest itself, so …

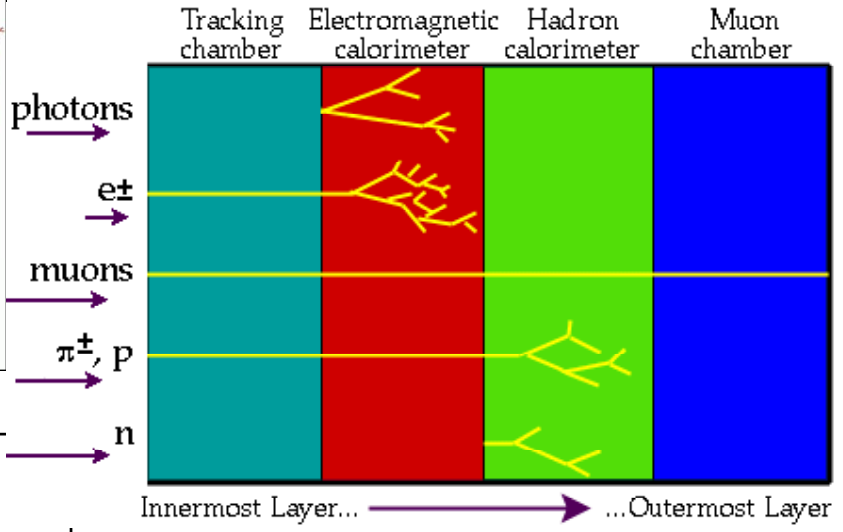
detectors must be able to detect as many particles and signatures as possible: e, μ, τ, ν, γ, jets, b-quarks, ….

With excellent performance over unprecedented energy range :
few GeV → few TeV





Length : ~ 46 m
Radius : ~ 12 m
Weight : ~ 7000 tons
~10^8 electronic channels
~ 3000 km of cables



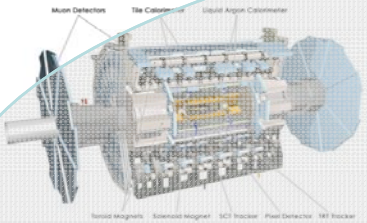
- Inner Detector ($|\eta|<2.5$, B=2T) :
  - -- Si pixels and strips
  - -- Transition Radiation Detector (e/π separation)

- Calorimetry ($|\eta|<5$) :
  - -- EM : Pb-LAr
  - -- HAD: Fe/scintillator (central), Cu/W-LAr (fwd)

- Muon Spectrometer ($|\eta|<2.7$) :
  air-core toroids with muon chambers

And ~2800 physicists from
169 Institutions, 37 countries,
5 continents

ATLAS Workflows



Prompt Reconstruction

Calibration & Alignment
Express Stream Analysis

Tier-0

CAF

CASTOR

650 MB/sec

RAW Re-processing
HITS Reconstruction

50-500 MB/sec

Tier-1

Tier-1

Tier-1

50-500 MB/sec

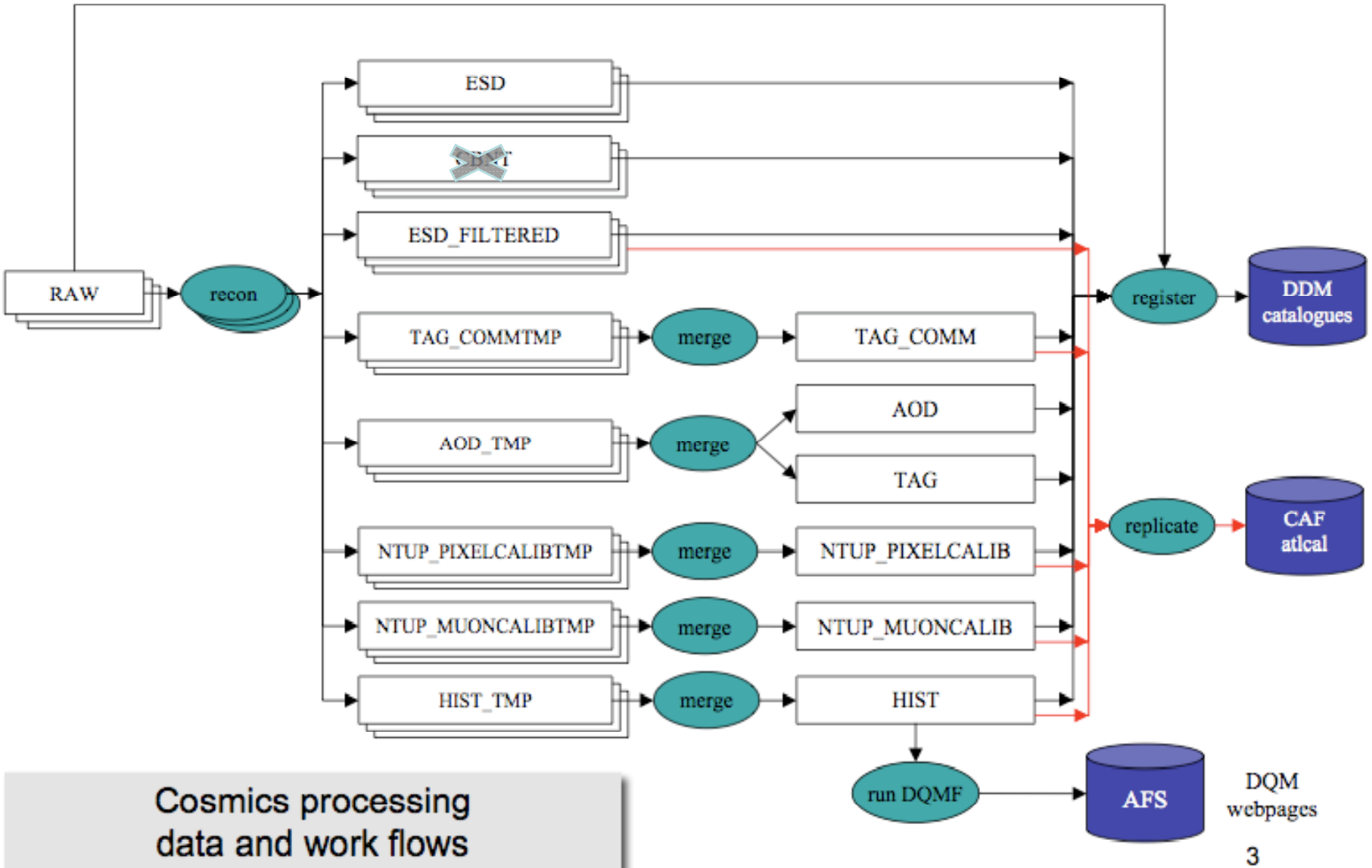
Simulation
Analysis

Tier-2

Tier-2

Tier-2

The Data Format at the Tier0



Cosmics processing
data and work flows
(as of beginning of October; simplified)

At the Tier0

Data Streams

Physics streams

- egamma
- muon
- Jet
- Etmiss
- tau
- Bphys
- minBias

Calibration streams

- Inner Detector Calibration Stream
  - Contains only partial events
- Muon Calibration Stream
  - Contains only partial events
  - Analyzed outside CERN
- Express line
  - Full events, 10% of data

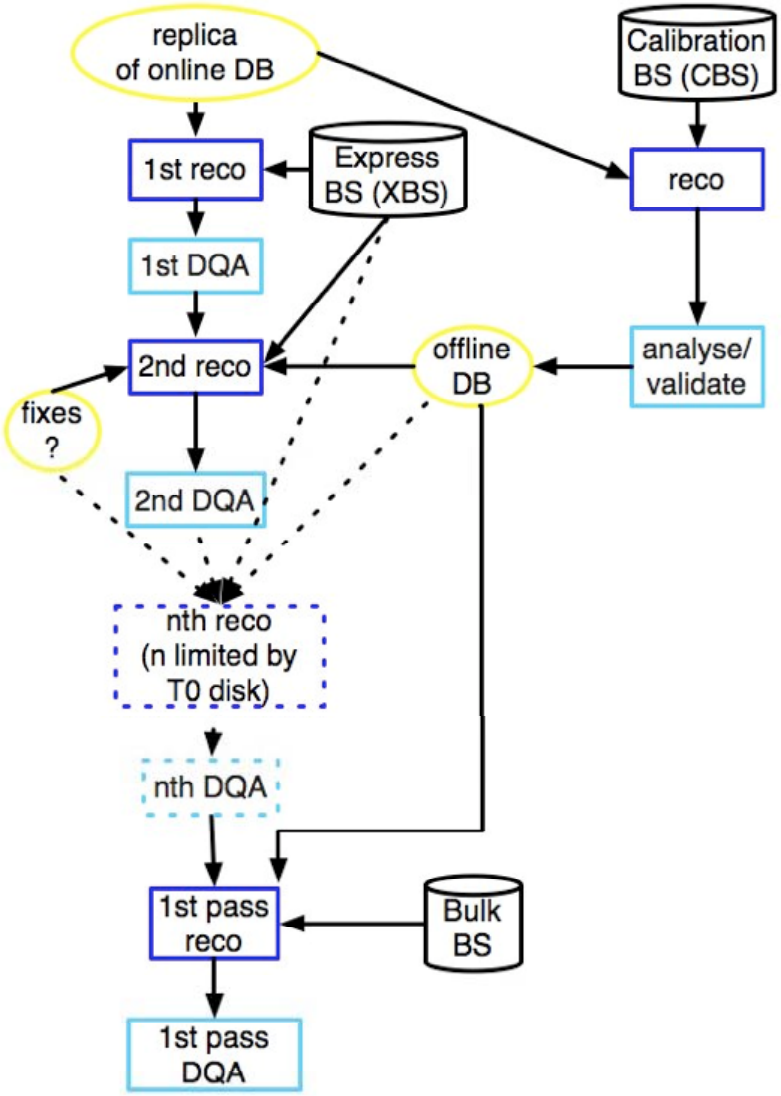
Runs and RAW Merging

- A run is between 2 Luminosity Blocks ~30 seconds
- 200 Hz for 30' is 6000 events but split between ~10 streams
- Streams are unequal and some create too small files
- Small RAW files are merged into >2 GB files
- Only merged files are written to tape and exported

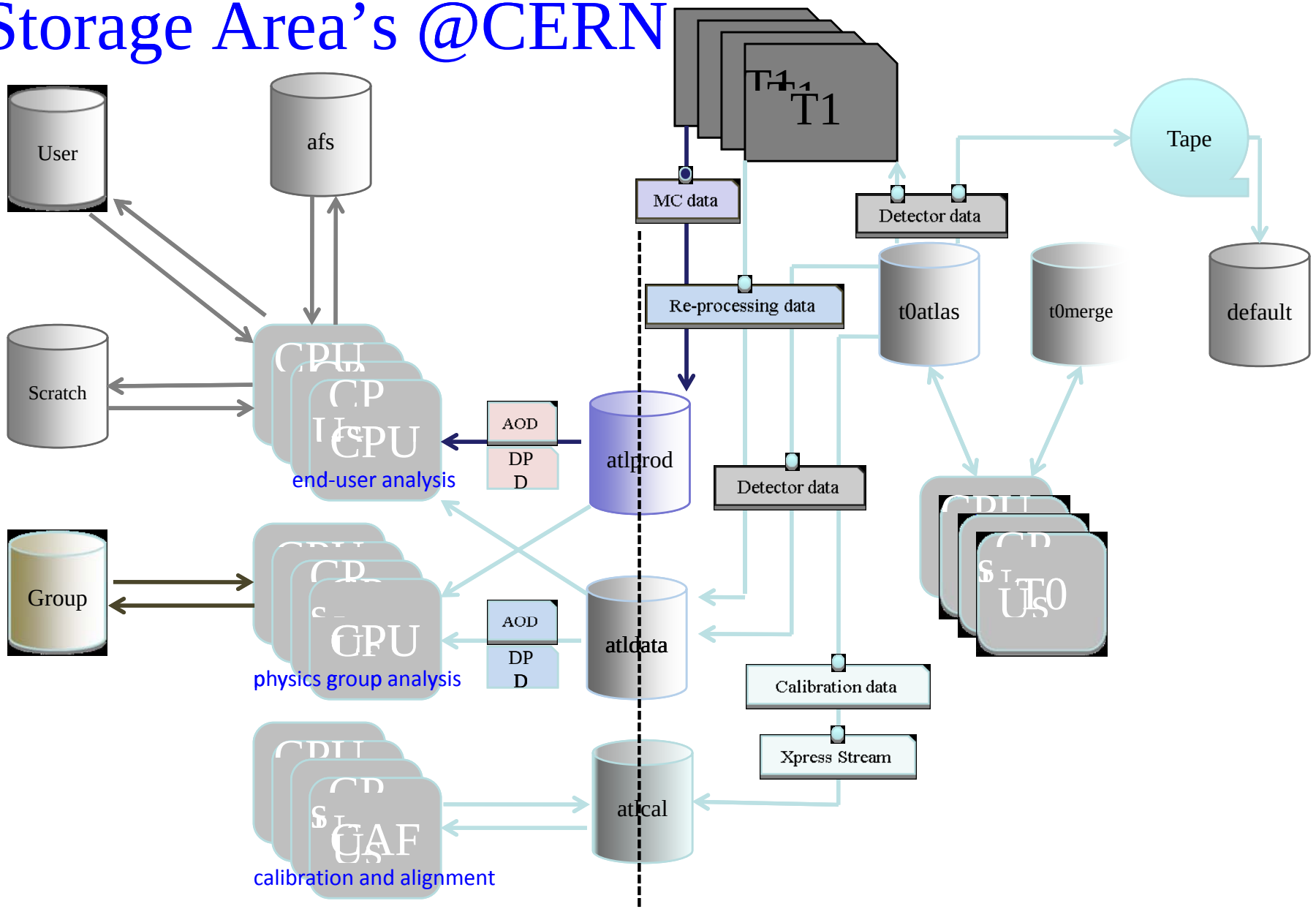
Calibration and Alignment Facility CAF

Per run ..

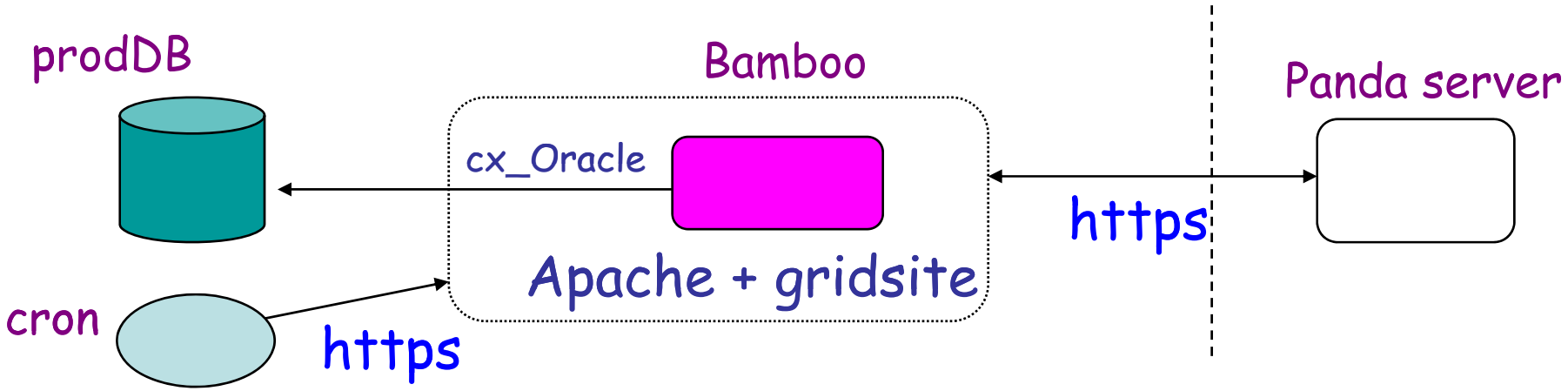
- Express line used for real-time processing
  - Initial calibration used
  - verified by DQ shifters
- Calibration data processed in CAF
  - Initial calibrations used
  - New calibrations into offline db
- Express line processed again
  - New calibrations used
  - Verified by DQ shifters
  - If necessary fixes applied
- Express line processed again if necessary
  - Buffer for several days of data
- Reconstruction of all data triggered
  - Results archived on tape. and
  - Made available at CERN, and
  - Replicated to other clouds



Storage Area's @CERN



Bamboo



- Get jobs from prodDB to submit them to Panda
- Update job status in prodDB
- Assign tasks to clouds dynamically
- Kill TOBEABORTED jobs

- A cron triggers the above procedures every 10 min

Summary of
ATLAS interactions with the middleware

- File Transfer Services (FTS)
  - One per Cloud (plus Tier0)
  - Triggers the third party transfer by contacting the SRM or Gridftp servers
- LCG File Catalog (LFC)
  - One per Cloud (plus Tier0)
  - Keeps track of local file replicas at a site
  - Main source of replica information by the site services
- Storage Resource Manager (SRM)
  - Extra level of abstraction on top of file transfers (e.g. gridftp)
  - Allows operations like pinning and space reservation
- Workload Management System (WMS)
  - User analysis
- Computing Element (CE)
  - gLite CE, OSG CE and CREAM CE