# ATLAS Operation report

# GDB

S. Jézéquel

ANNECY lapp

9 March 2011

# Outline

- **Operation report (Dec 2010- Mar 2011)**
- **Short term implementation : Gradual breaking of cloud model**
- **ATLAS Data Distribution model : 2011**

# Operation activities(Dec 2010-Mar 2011)

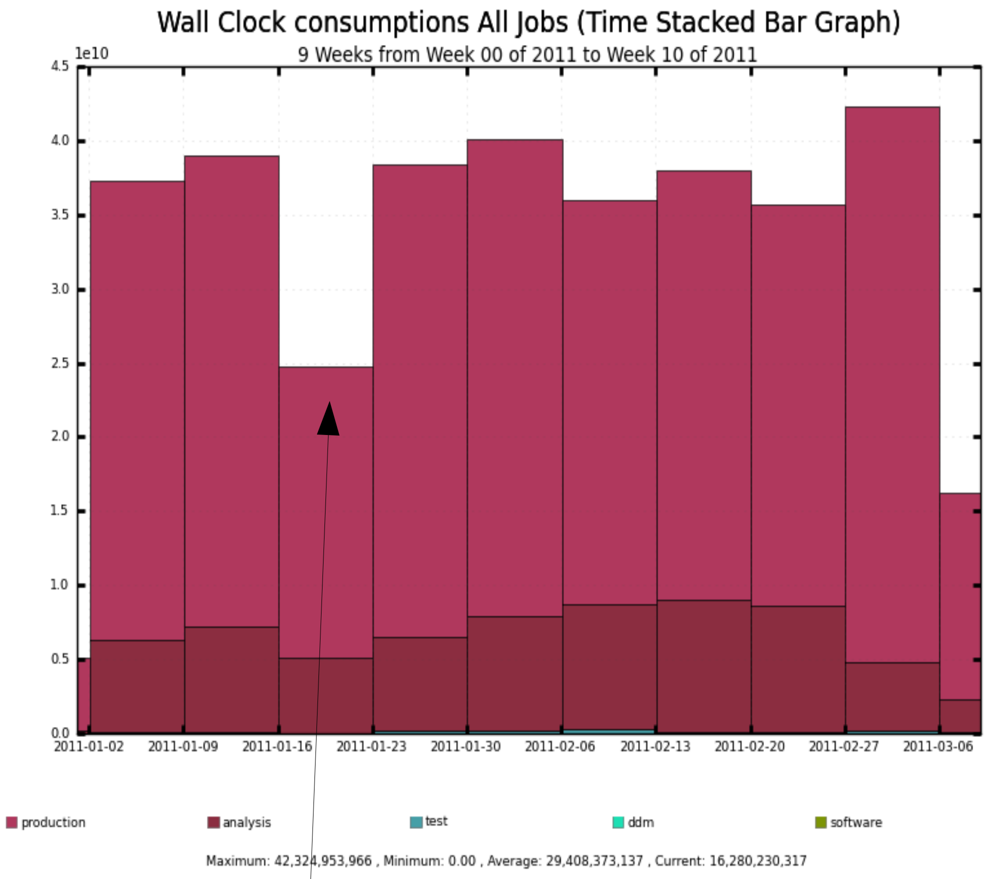
- **No specific activity during this period**

- **Consolidation (ATLAS+sites) and preparation for 2011 activities :**
  - **Service certificate replace Kors certificate for DDM activities**
  - **Oracle upgrade at CERN + ATLAS database splitting**
  - **Validation of EOS technology at CERN**
  - **Taiwan DISK : Castor → DPM (to be finished)**
  - **Setup automatic export of ID calibration data from T0 to TAIWAN/Valencia**

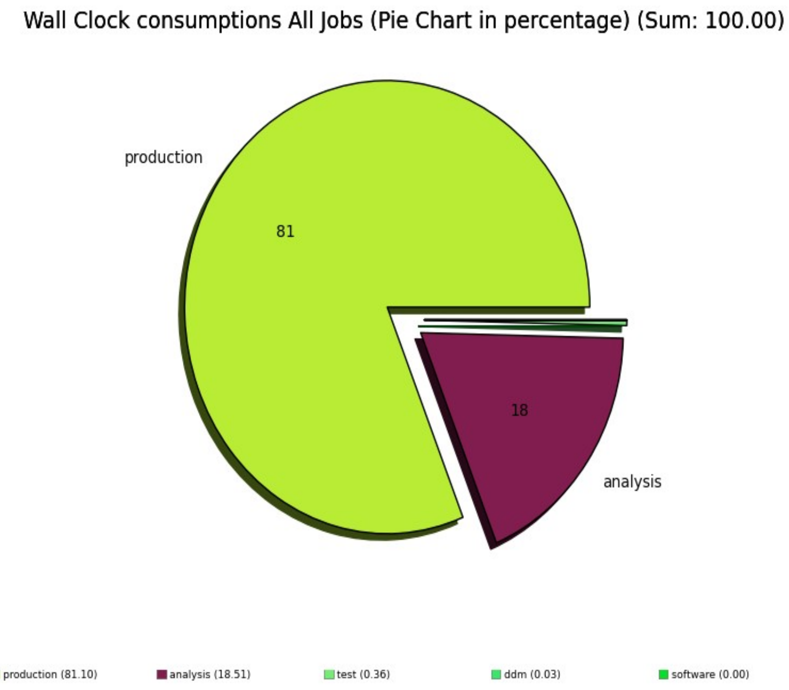
- **2010 Heavy-Ion reprocessing starting this week**

# Processing activity: Jan-March 2011

All ATLAS Grid activity (January-March 2011)



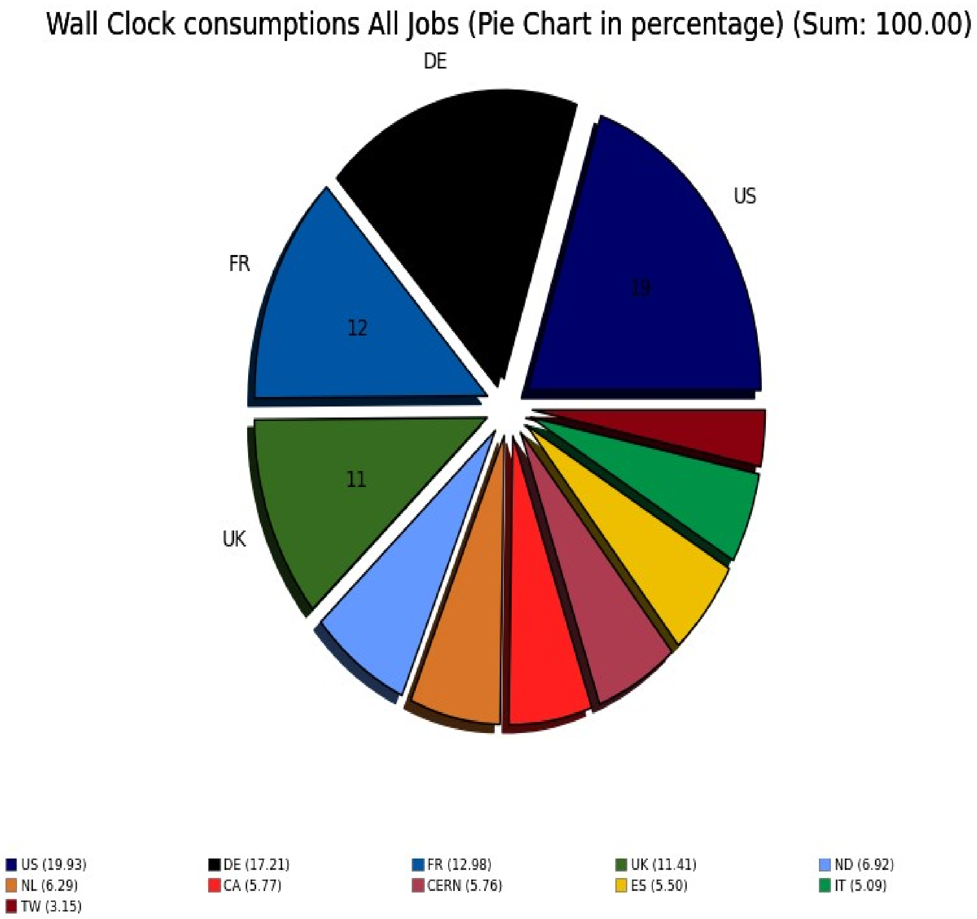
~66 k full CPUs



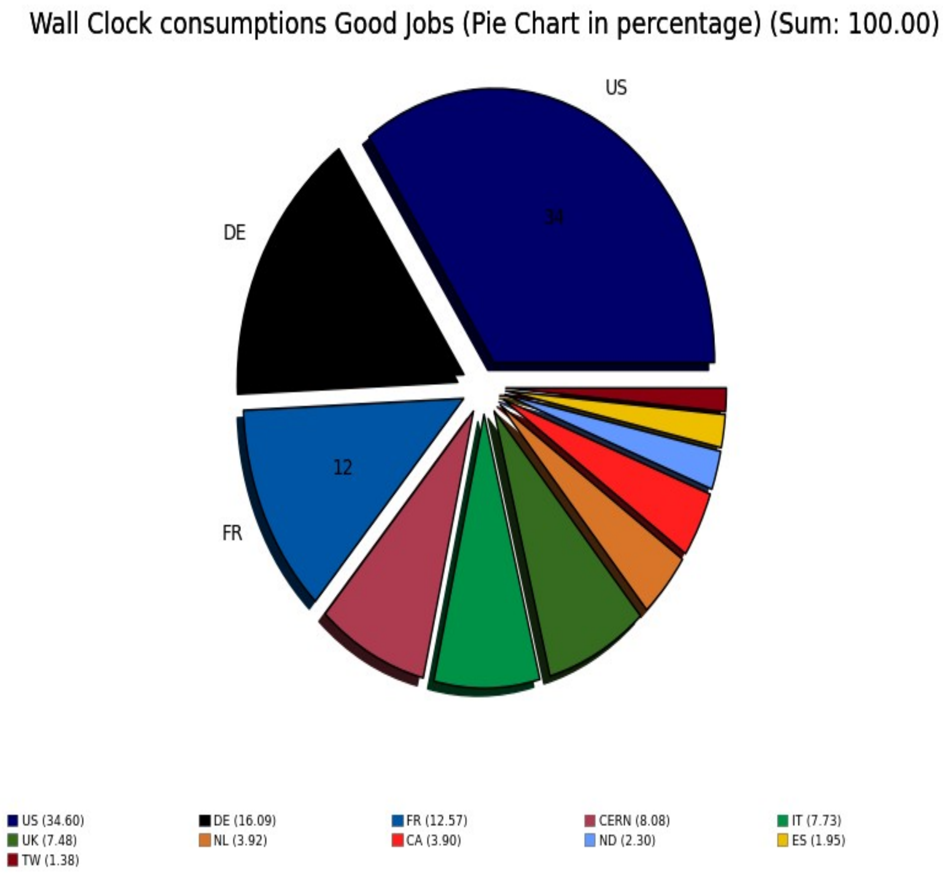
CERN Oracle upgrade : 16 Jan (LCGR) -17 Jan (ADCR+ATLR)

# Processing activity (2)

## Production (MC + group production)



## Analysis (user+phys. groups)



# Storage news

## Merging of space tokens DATADISK/MCDISK

- Data migrated with DDM (FTS transfer + central deletion)
- Only primary replicas were transfered (minimize activity)
- Exception : RAL : Internal Castor migration

- Timescale
  - January : T2/T3 (no other major transfer activity)
  - February-March : T1 (no other major transfer activity)
  - CERN : Will be done when migration Castor → EOS

Site responsability : When migration is finished, clean remaining dark data

## Space token shares in 2011

- Reference : https://twiki.cern.ch/twiki/bin/view/Atlas/StorageSetUp
- 2011 shares similar to 2010 (validated on 7 March 2011)
- Main changes :
  - Shares ATLASMCDISK and ATLASDATADISK merged

    → 1 space token host 75-80 % of storage ressources
  - ATLASPRODDISK : 25 TB in T1s (MC production at T1 + ATLAS managed stagin buffer)

# Operation issues

- RAL :
  - Problem with Castor during migration from bad servers (Christmas period)
  - MCDISK was full because of bug in ATLAS central deletion

    → No MC production in UK cloud during a week

- PIC : Temporary lost 800k files/250 TB after file system corruption (GGUS : 66409)
  - Took a week to get list of unaccessible files
  - PIC announced that most of files could be accessible after internal migration
  - Too many files to recover from outside (1/3 was available outside)

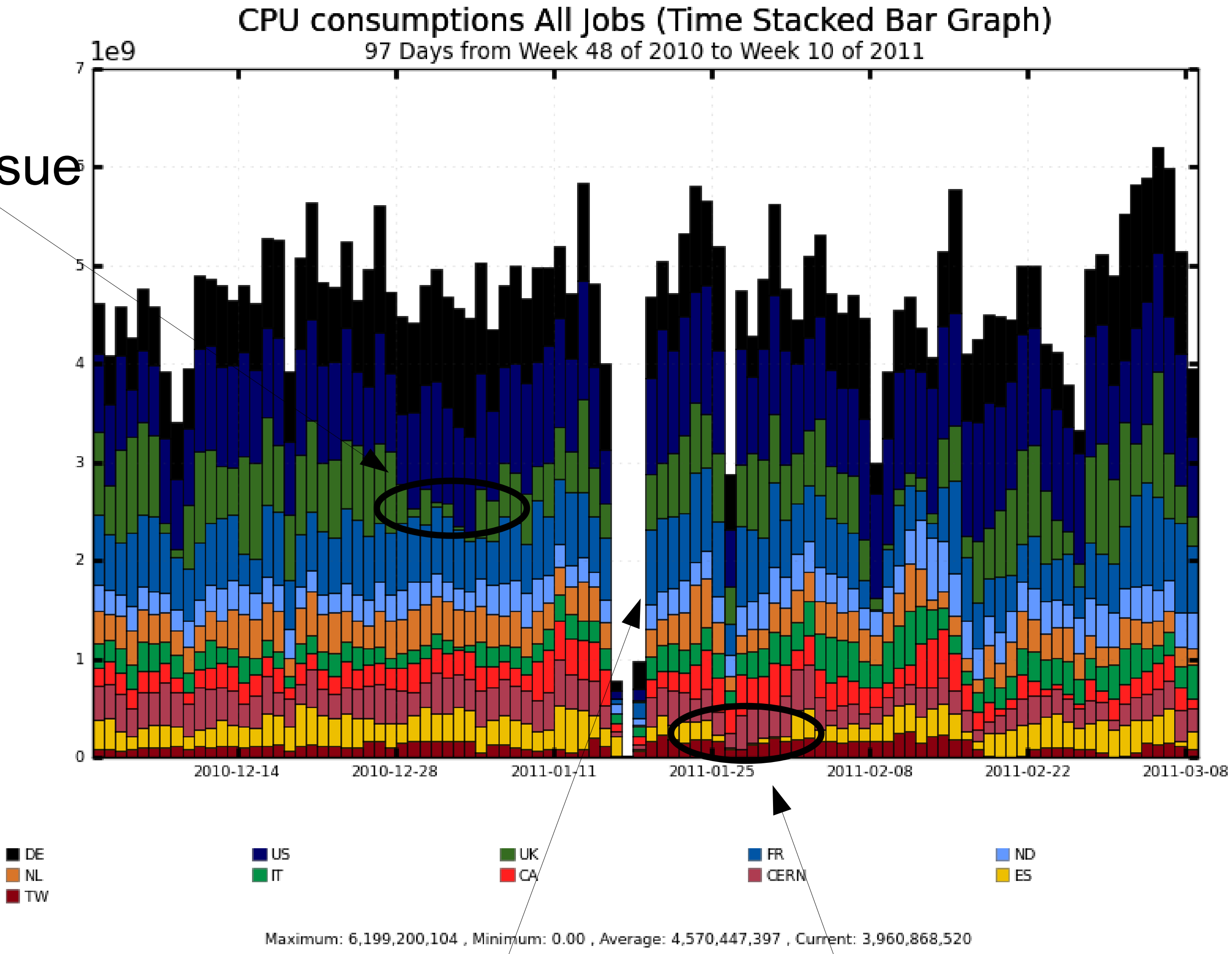
    → Data recovered/consolidated within PIC (~1 week)

    → No production in ES cloud during 2 weeks + some unaccessible data

- Castor/srm upgrade to 2.10 (CERN) (GGUS :):
  - Was not able to rollback after experiencing instabilities
  - Solved within a day by patching 2.10

- Thousands of jobs from same user using > 4 GB memory : Solved by contacting user

# Operation issues



RAL issue

CERN Oracle upgrade

PIC issue

# Breaking cloud model

- ATLAS wants to break the cloud model to get more flexibility
- Obvious constraint : Should match the network connectivity between sites

- Done step-by-step :
  - To ensure that scalibility issues can now be overcome
  - Adapt the monitoring tools
  - Train ATLAS shifters : who is responsible in case of problem

- Current actions :
  - Prepare LFC consolidation at CERN
  - Some T2s running G4 simulation with input/output files transfered from/to different T1s
  - Direct transfers between some T2s and all T1s

- Future actions :
  - Promote 'good' T2s to host primary replicas (only in T1s today)

# LFC consolidation at CERN

- Goal :
  - All LFCs agregated in a single LFC at CERN
  - Read-only replica in another site (probably BNL)

- Reason :
  - ATLAS experienced LFC downtime over few weeks (Summer 2010)
  - Current LFC model: single point of failure

    → all stored data within cloud can be unaccessible

- Status :
  - Discussion between ATLAS and WLCG/CERN to validate the merging procedure
  - Identifiy possible inconsistencies between catalogs before merging

- Timescale :
  - One LFC migration at a time (should be done within days)
  - Expected to be done during spring/summer 2011

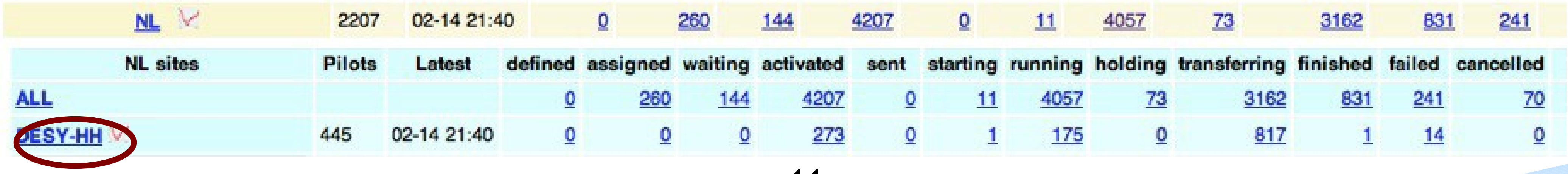
# Cross-cloud production

- Reason :
  - Allocate more CPU ressources for urgent/big simulation tasks

    → Gather CPUs from sites outside the cloud
  - Avoid inbalance between T1 storage ressource and CPU ressource within cloud
  - Continue to use T2s CPUs when T1 SE is down

- Necessary connectivity :G4 simulation jobs in T2s with output transfered to T1

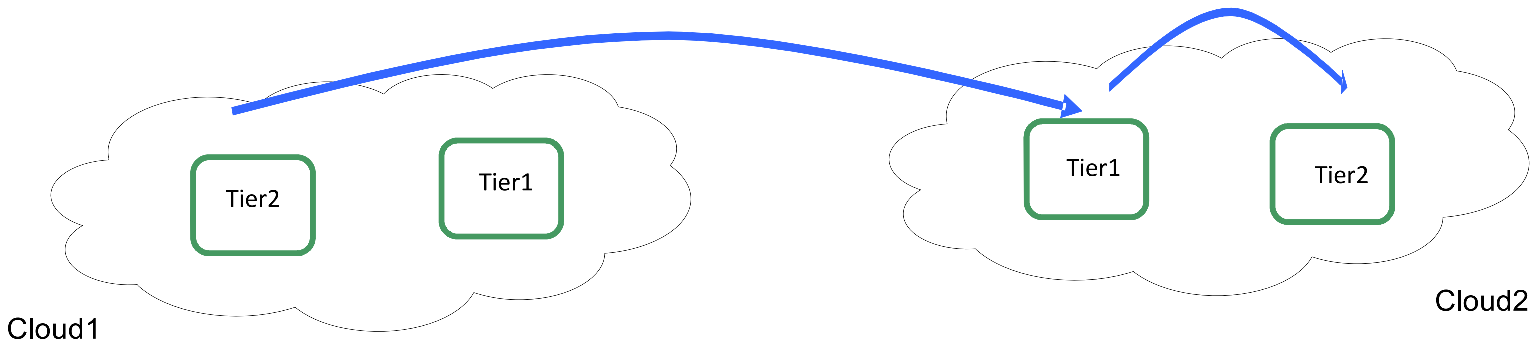
  1000 cores : Transfer rate ~ O(1) MB/s : Easy
- Main issue :
  - Setup FTS channels to avoid to go through STAR-T1/T2 channels (T2D topic)
  - Adapt monitoring
- Current situation
  - Some big T2 sites already associated to many Tier1s



| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NL | 2207 | 02-14 21:40 | 0 | 260 | 144 | 4207 | 0 | 11 | 4057 | 73 | 3162 | 831 | 241 | |
| NL sites | Pilots | Latest | defined | assigned | waiting | activated | sent | starting | running | holding | transferring | finished | failed | cancelled |
| ALL | | | 0 | 260 | 144 | 4207 | 0 | 11 | 4057 | 73 | 3162 | 831 | 241 | 70 |
| DESY-HH | 445 | 02-14 21:40 | 0 | 0 | 0 | 273 | 0 | 1 | 175 | 0 | 817 | 1 | 14 | 0 |

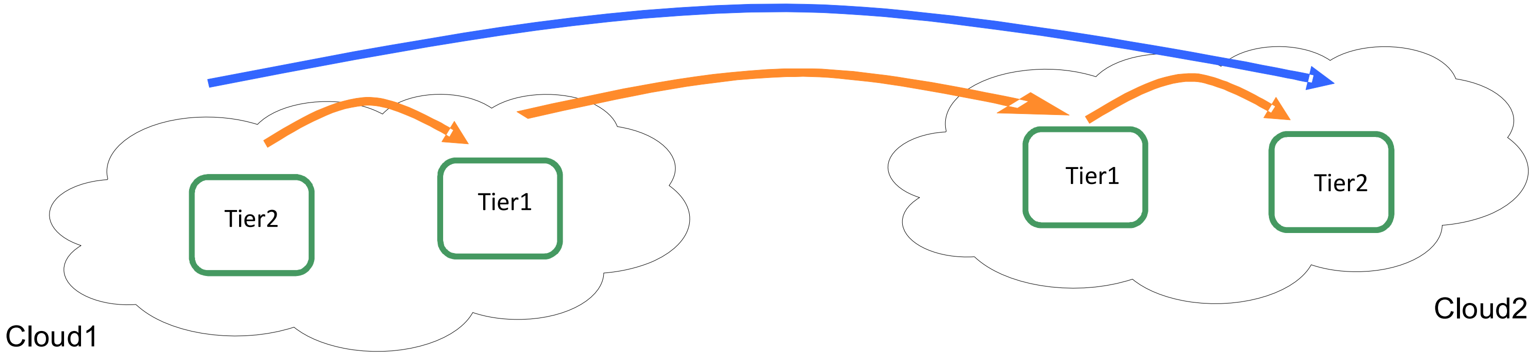
# Data collection into T2s

## Current DDM model



ISSUES T2→ T1 for T2S WITH BAD CONNECTIVITY

## DDM model : Version +1 (Under validation in Italy cloud)



2 possible paths → Need monitoring to optimise path (Depend on file size)

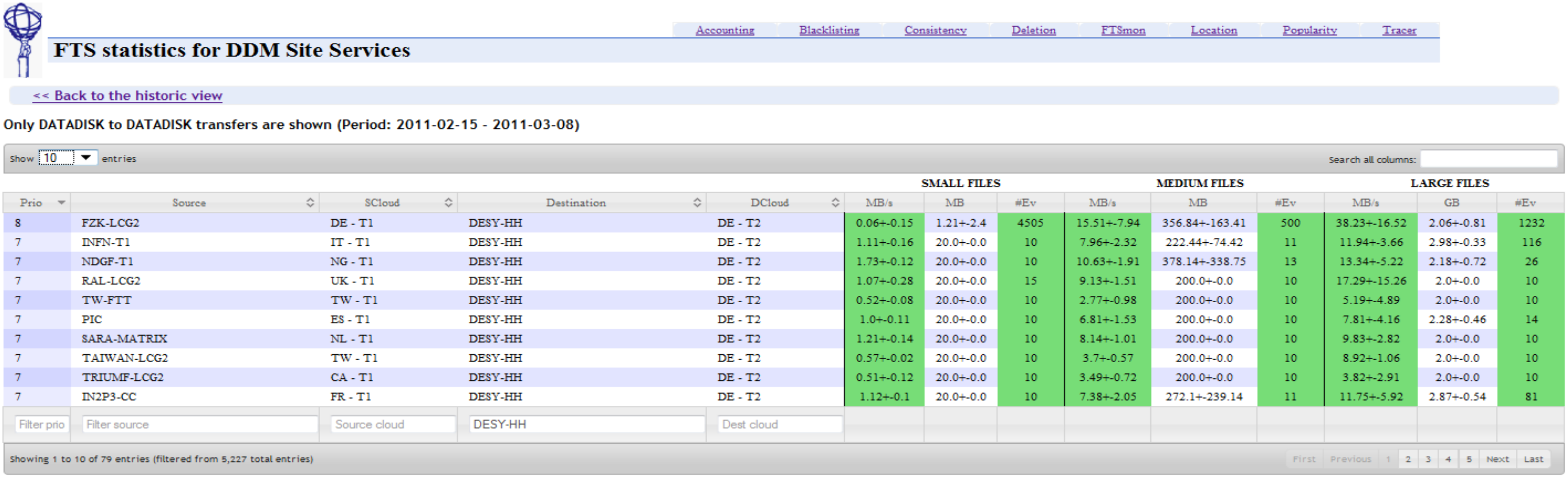
Expected to use direct transfers for small files

# T2 connectivity : Monitoring

- **Extend current channel validation (T0→T1, T1a→T1b, T1a→T2a)**

  **to T1a→T2b and T2a→T2b**

**http://bourricot.cern.ch/dq2/ftsmon/sonar  view/cached/**



Accounting | Blacklisting | Consistency | Deletion | FTSmon | Location | Popularity | Tracer

**FTS statistics for DDM Site Services**

<< Back to the historic view

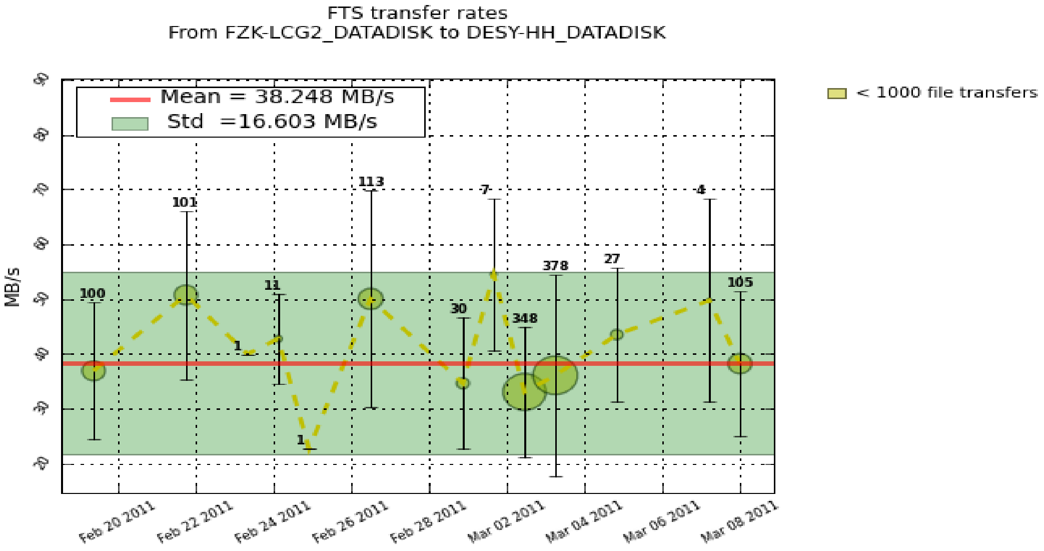
Only DATADISK to DATADISK transfers are shown (Period: 2011-02-15 - 2011-03-08)

Show 10 entries | Search all columns:

| | | | | | SMALL FILES | | | MEDIUM FILES | | | LARGE FILES | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Prio | Source | SCloud | Destination | DCloud | MB/s | MB | #Ev | MB/s | MB | #Ev | MB/s | GB | #Ev |
| 8 | FZK-LCG2 | DE - T1 | DESY-HH | DE - T2 | 0.06+-0.15 | 1.21+-2.4 | 4505 | 15.51+-7.94 | 356.84+-163.41 | 500 | 38.23+-16.52 | 2.06+-0.81 | 1232 |
| 7 | INFN-T1 | IT - T1 | DESY-HH | DE - T2 | 1.11+-0.16 | 20.0+-0.0 | 10 | 7.96+-2.32 | 222.44+-74.42 | 11 | 11.94+-3.66 | 2.98+-0.33 | 116 |
| 7 | NDGF-T1 | NG - T1 | DESY-HH | DE - T2 | 1.73+-0.12 | 20.0+-0.0 | 10 | 10.63+-1.91 | 378.14+-338.75 | 13 | 13.34+-5.22 | 2.18+-0.72 | 26 |
| 7 | RAL-LCG2 | UK - T1 | DESY-HH | DE - T2 | 1.07+-0.28 | 20.0+-0.0 | 15 | 9.13+-1.51 | 200.0+-0.0 | 10 | 17.29+-15.26 | 2.0+-0.0 | 10 |
| 7 | TW-FTT | TW - T1 | DESY-HH | DE - T2 | 0.52+-0.08 | 20.0+-0.0 | 10 | 2.77+-0.98 | 200.0+-0.0 | 10 | 5.19+-4.89 | 2.0+-0.0 | 10 |
| 7 | PIC | ES - T1 | DESY-HH | DE - T2 | 1.0+-0.11 | 20.0+-0.0 | 10 | 6.81+-1.53 | 200.0+-0.0 | 10 | 7.81+-4.16 | 2.28+-0.46 | 14 |
| 7 | SARA-MATRIX | NL - T1 | DESY-HH | DE - T2 | 1.21+-0.14 | 20.0+-0.0 | 10 | 8.14+-1.01 | 200.0+-0.0 | 10 | 9.83+-2.82 | 2.0+-0.0 | 10 |
| 7 | TAIWAN-LCG2 | TW - T1 | DESY-HH | DE - T2 | 0.57+-0.02 | 20.0+-0.0 | 10 | 3.7+-0.57 | 200.0+-0.0 | 10 | 8.92+-1.06 | 2.0+-0.0 | 10 |
| 7 | TRIUMF-LCG2 | CA - T1 | DESY-HH | DE - T2 | 0.51+-0.12 | 20.0+-0.0 | 10 | 3.49+-0.72 | 200.0+-0.0 | 10 | 3.82+-2.91 | 2.0+-0.0 | 10 |
| 7 | IN2P3-CC | FR - T1 | DESY-HH | DE - T2 | 1.12+-0.1 | 20.0+-0.0 | 10 | 7.38+-2.05 | 272.1+-239.14 | 11 | 11.75+-5.92 | 2.87+-0.54 | 81 |
| Filter prio | Filter source | Source cloud | DESY-HH | Dest cloud | | | | | | | | | |

Showing 1 to 10 of 79 entries (filtered from 5,227 total entries)

First | Previous | 1 | 2 | 3 | 4 | 5 | Next | Last

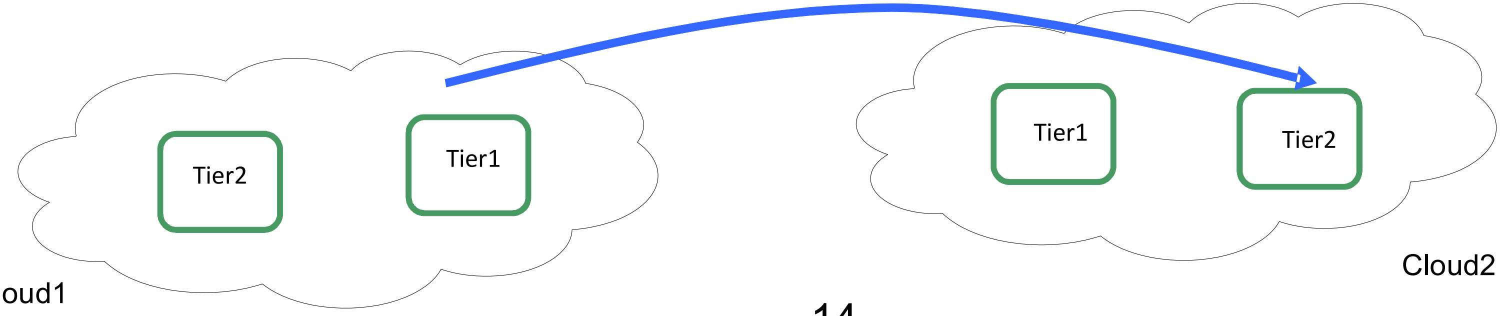


FTS transfer rates
From FZK-LCG2_DATADISK to DESY-HH_DATADISK

# Data collection into T2s (2)

- Issue :
  - T2s with big storage capacity collect lots of data
  - Current model : cross-cloud transfers go through T1s
    - Reason : Good connectivity between clouds through T1s/LHCOPN
  - But :
    - Triggers additional activity in T1s SE (copy+delete)
    - Transfers to T2s are stuck if T1 SE is down/full
    - Transfers can be delayed if huge activity for T1 transfer
- Target :
  - Minimize useless load on T1 SE
  - Minimize intermediate steps → Less sensititive to intermediate site availability
  - First use case : Collect group production at T1s into group storage at T2

DDM model : Version +1' (Implemented)



# Direct cross-cloud T2 connectivity

- **Select good T2 sites which will always transfer from/to all T1s**
  - **T2Ds**
- **T2D current list :**
  - **All US T2s**
  - **DESY-HH,DESY-ZN, GRIF-LAL,GRIF-LPNHE, INFN-NAPOLI, IFIC**
- **A long list of sites under probation**
- **ATLAS would like to have as many sites as possible**
- **Triggered many network studies (UK for example)**

**→ More sites will be added soon**

**LHCONE will be a key component for this policy**

- **Request additional FTS channels :**
  - **T1s → T2D**
  - **T2Ds → T1 (currently go through STAR-T1 channel)**
  - **Implementation/optimisation discussed in T1 Coordination meeting**
  - **Triggered discussions with FTS dev.**

# T2 connectivity : Summary

- Full multi-hop model under validation
  - Hopefully generalised in March 2011

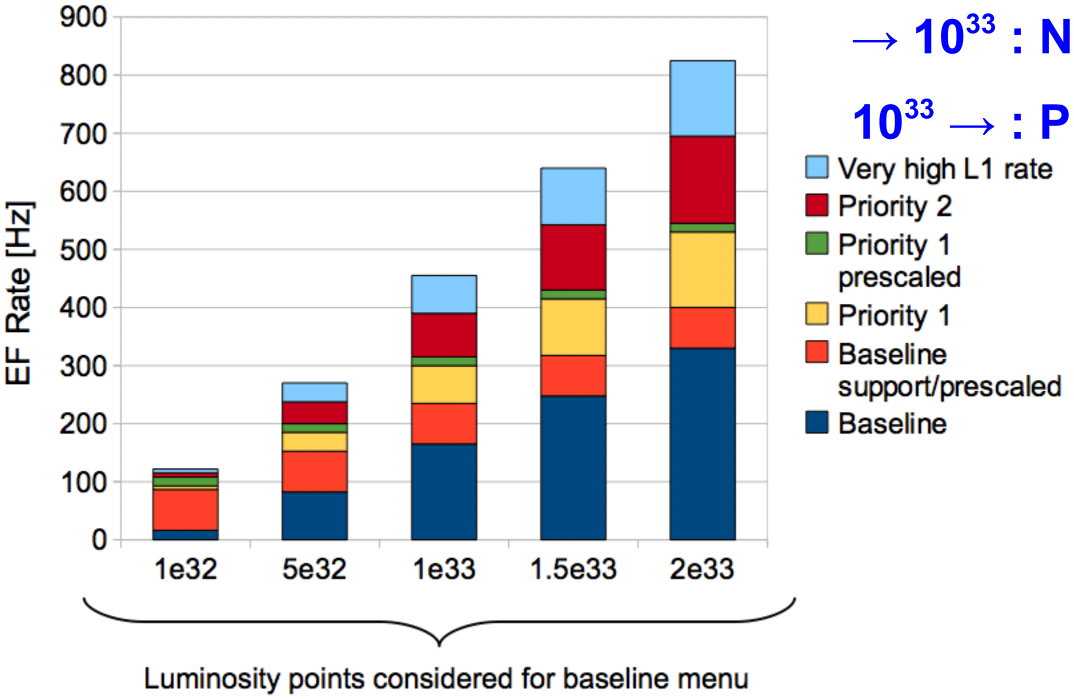
- Monitoring cross-cloud transfers : Done
  - Migration to well supported monitoring framework under way

- DDM connectsT2Ds to all T1s : Done
  - First list of T2Ds defined
  - FTS channel setup being optimised

# 2011 Data Distribution Model

- New requirements from ATLAS physic/trigger community to collect more data

→ Increase mean Event Filter mean rate to 400 Hz (limitation is T0 capacity)



→ $10^{33}$ : No additional cut

$10^{33}$ → : Priorities defined to limit to 400 Hz

**Computing issues**

- Transfer rate of fresh data from CERN
- Storage capacity for primary replicas produced over year

→ Review of the ATLAS Data Distribution Model

# 2011 Data Distribution Model

- Reduce RAW size : zip files
  - Gain a factor 2 (many empty calorimeter cells)

    → Compression factor close to 1 when writing on TAPE
  - Zipping is done at T0 level
  - Unzipping files is done during file reading (< 0.1 s in addition)
  - Will be implemented in coming days

- If needed, transfer CERN CAF CPU resources to T0 from prompt data reco.

  → CERN cannot replace a stuck T1 for reprocessing campain

  → All Tier1s should reach promised availability

- In addition to TAPE copies,
  - 1 RAW copy on DISK to allow prompt access for 'discovery' studies

    (few to few 100k events accessible within 24 hours)

# 2011 Data Distribution Model (2)

- 'Life without ESD' :
  - Restrict number of ESD replicas (2 copies)

    → Promote analysis from AOD/DESD
  - Lifetime of 6-8 weeks for bulk ESD streams

- AOD/DESD:
  - Number of replicas adapted to available DISK ressources
  - Promote big T2Ds to host primary replicas (+10 TB)

    → Include them in schedulded downtime coordination ?

LHC one : key component to deliver datasets from these T2s

# 2011 Data Distribution Model: Rate

- Transfer rate T0 → T1s for 400 Hz
  - 2 RAW copies : 2*320 MB/s (0.8 MB/evt)
  - 1 ESD copy : 640 MB/s (1.6 MB/evt)
  - 2 AOD +2 DESD copies : 4 * 110 MB/s (0.275 MB/evt)

→ Total : 1720 MB/s (does not include internal read/write within T0)

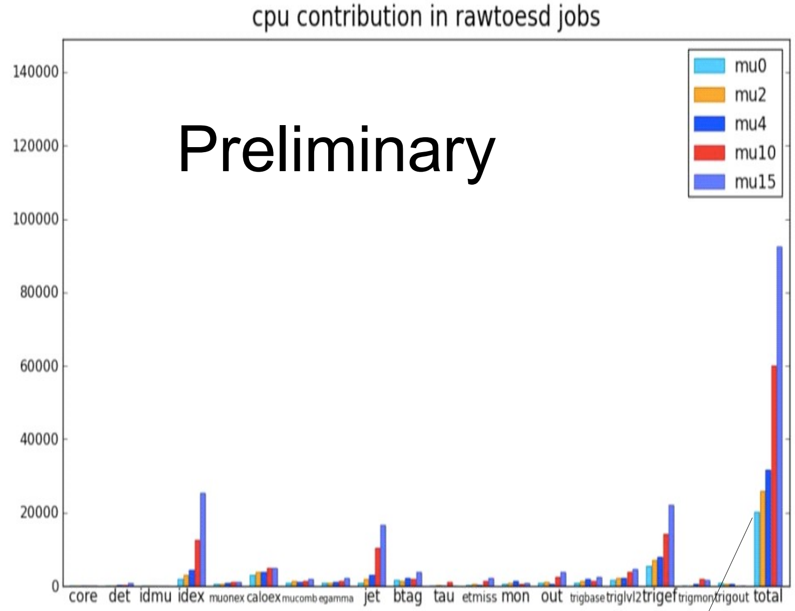
- Transfer activity will be smoother : No run validation before exporting data

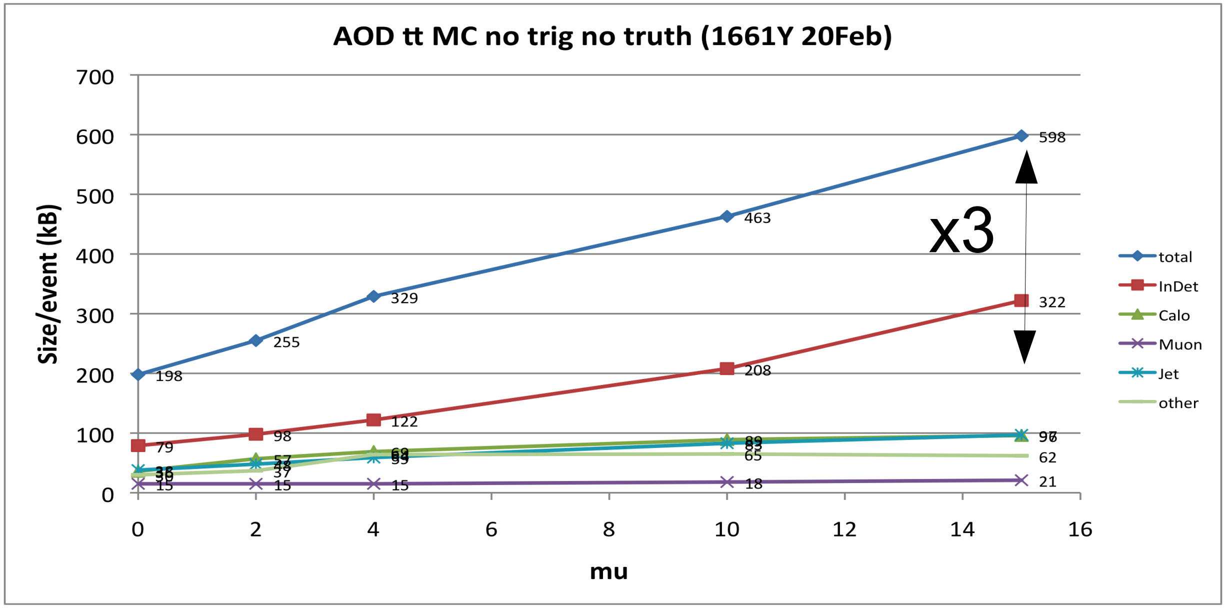
# Pileup effect

- LHC luminosity increase ← including more protons per bunch

  → Number pile-up events increases rapidly with luminosity

- Pile-up beginning/end 2011 (L ~O($10^{32}$)/O($10^{33}$)) ~0/15





Less CPU ressources for CERN CAF activities → CERN not backup for T1s reprocessing

•Pressure on software developpers to reduce the size increase → Improvments each day

Another potential source of tension for computing ressources
→ Monitoring and reactivity will be necessary

# Conclusion

- Smooth ATLAS computing activity during the last 3 months
  - Includes usual rate of site/tools issues

- Next months will be much more challenging than 2010
  - All consolidation activities should proceed quickly
  - Sites (especillay T1s) should keep up with expected availability

- Many medium/long term developments also coming hopefully

in collaboration with WLCG and other LHC experiments