ATLAS Quarterly Report and Plans

Dario Barberis

CERN & Genoa University/INFN

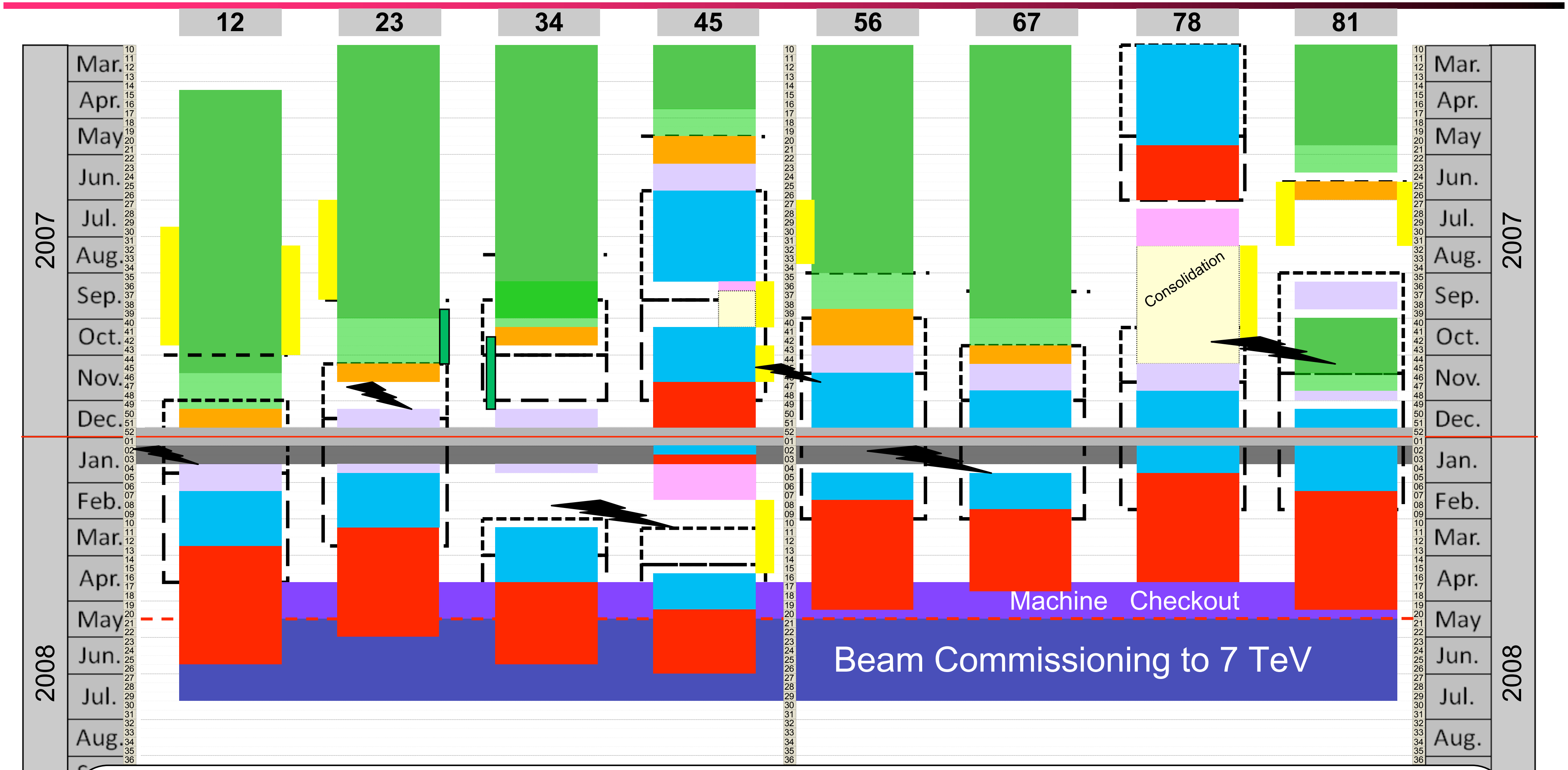
Outline

- Recent news from ATLAS
- Evolution of the computing and analysis model
- Data streaming decisions
- Event size and performance issues
- Database distribution and tests of TAGs
- Simulation production
- Progress with Distributed Analysis tools
- Data throughput tests
- Distributed Computing organisation
- Evolution of the distributed production system
- Plan of activities until LHC turn-on

ATLAS Construction

- Construction and assembly at the surface come to an end
- Installation in the cavern is also nearing completion
- The cosmics Milestones Weeks give a taste of the excitement we will enjoy in less than one year from now…

Schedule by LHC sectors (October 9th)



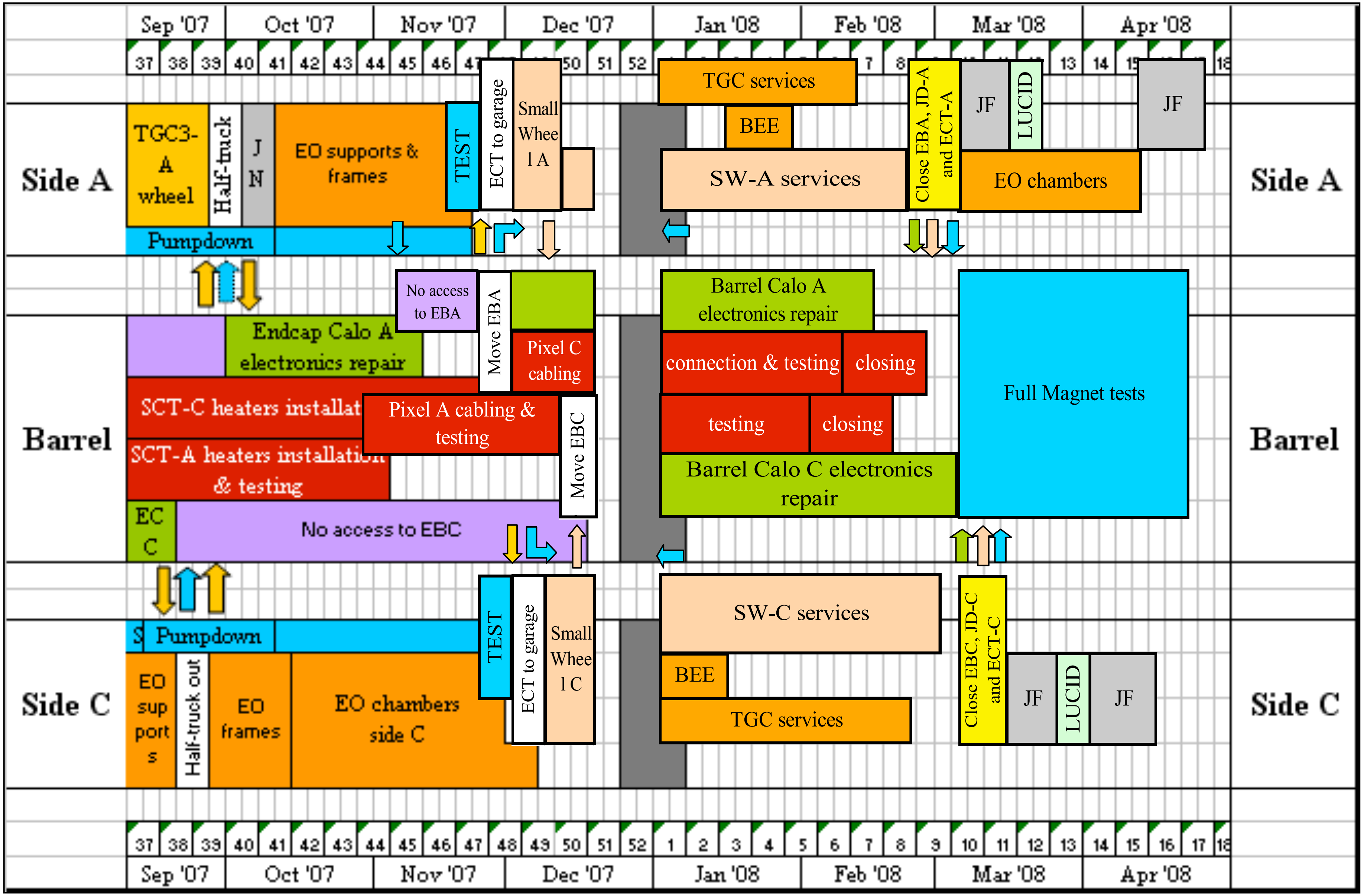
General schedule Baseline rev. 4.0

- ···· Global pressure test &Consolidation
- □ Cool-down
- ⌜⌟ Powering Tests

- 🟩 Interconnection of the continuous cryostat
- 🟩 Leak tests of the last sub-sectors
- 🟨 Inner Triplets repairs & interconnections
- 🟧 Global pressure test &Consolidation

- 🟪 Flushing
- 🟦 Cool-down
- 🟪 Warm up
- 🟥 Powering Tests

ATLAS completion schedule



Sep '07 | Oct '07 | Nov '07 | Dec '07 | Jan '08 | Feb '08 | Mar '08 | Apr '08

37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Side A

TGC3-A wheel

Half-truck

J N

EO supports & frames

TEST

ECT to garage

Small Wheel A

Pumpdown

TGC services

BEE

SW-A services

Close EBA, JD-A and ECT-A

JF

LUCID

JF

EO chambers

Barrel

No access to EBA

Move EBA

Endcap Calo A electronics repair

Pixel C cabling

SCT-C heaters installat

Pixel A cabling & testing

SCT-A heaters installation & testing

Move EBC

EC C

No access to EBC

Barrel Calo A electronics repair

connection & testing

closing

testing

closing

Barrel Calo C electronics repair

Full Magnet tests

Side C

S Pumpdown

TEST

ECT to garage

Small Wheel C

EO supports

Half-truck out

EO frames

EO chambers side C

SW-C services

BEE

TGC services

Close EBC, JD-C and ECT-C

JF

LUCID

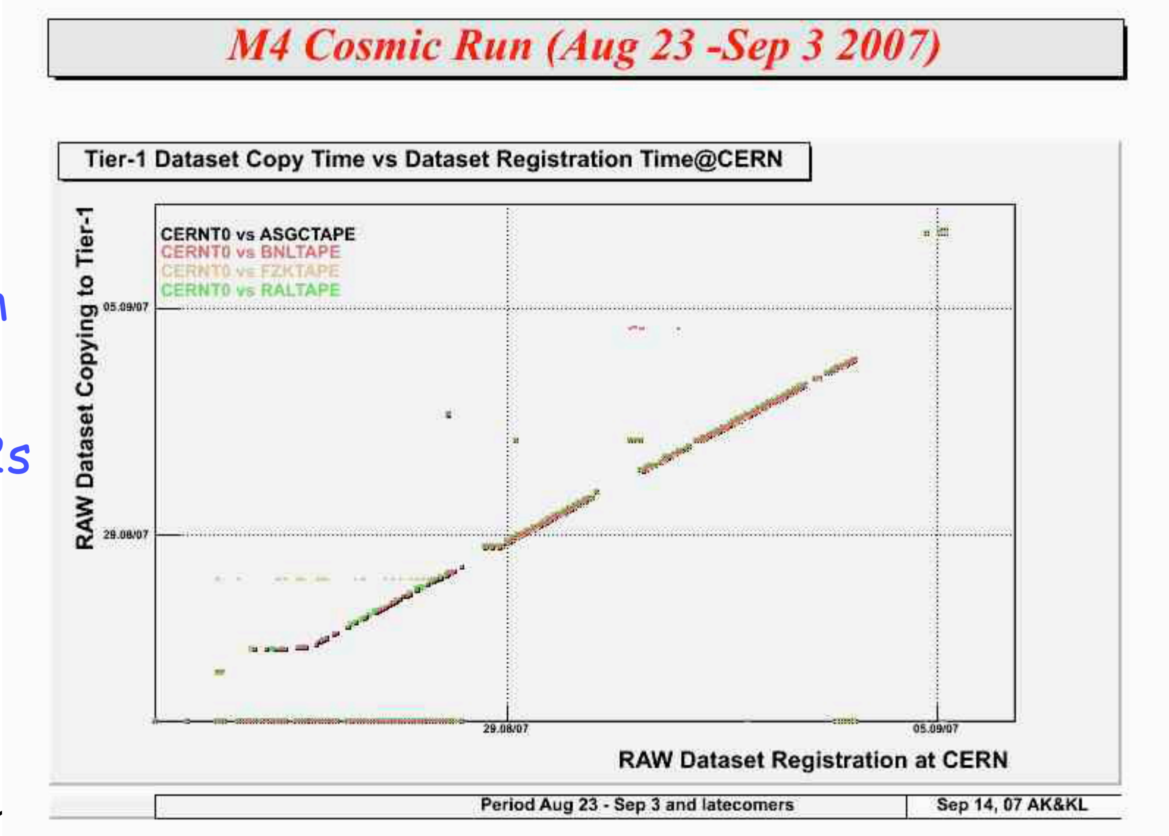
JF

M4 Objectives

- *August 23 – September 3*
- Using 4 SFOs: rate < ~250 MB/s
- Data written into Castor 2
- Full Tier-0 operation
- RAW data subscribed to Tier-1 tape
- ESD data subscribed to Tier-1 disk
- ESD data subscribed from Tier-1s to Tier-2s
- *Analyse M4 data at Tier-2s*

M4 Objectives

- ✓ August 23 – September 3
- ✓ Using 4 SFOs: rate < ~250 MB/s
- ✓ Data written into Castor 2 (~40 TB)
- ✓ Full Tier-0 operation
- ✓ RAW data subscribed to Tier-1 tape
- ✓ ESD data subscribed to Tier-1 disk
- ✓ ESD data subscribed from Tier-1s to Tier-2s
- ✓ Analyse M4 data at Tier-2s



M4 Cosmic Run (Aug 23 -Sep 3 2007)

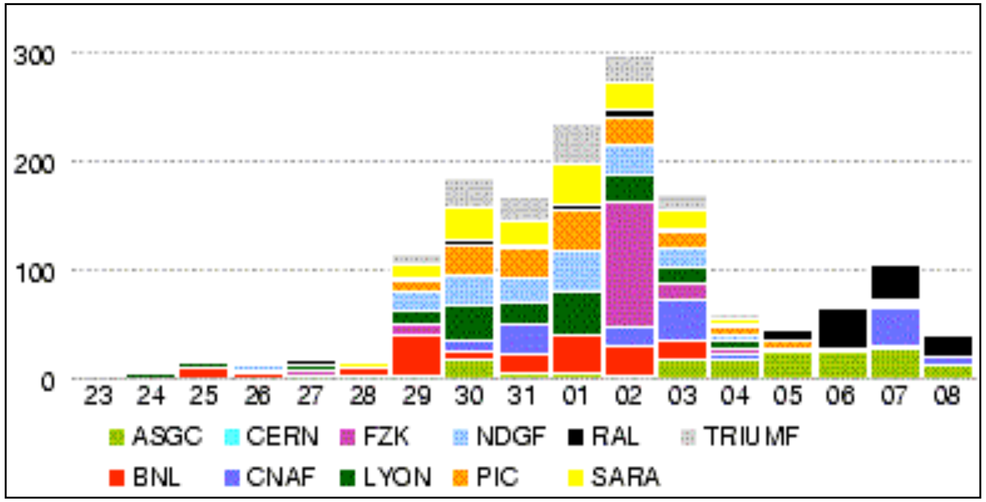
Tier-1 Dataset Copy Time vs Dataset Registration Time@CERN

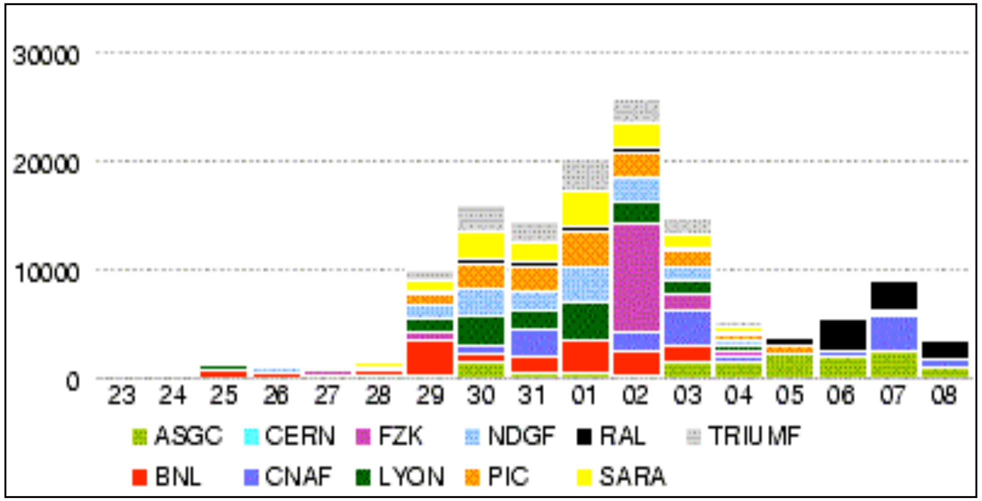
Period Aug 23 - Sep 3 and latecomers

Sep 14, 07 AK&KL

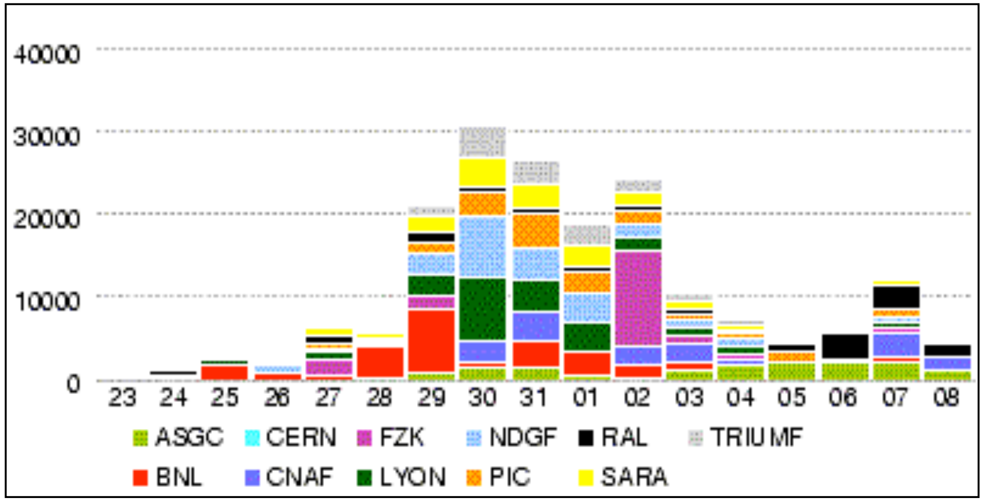
The whole M4 run

Total throughput (MB/s)
Aug 23 – Sep 8



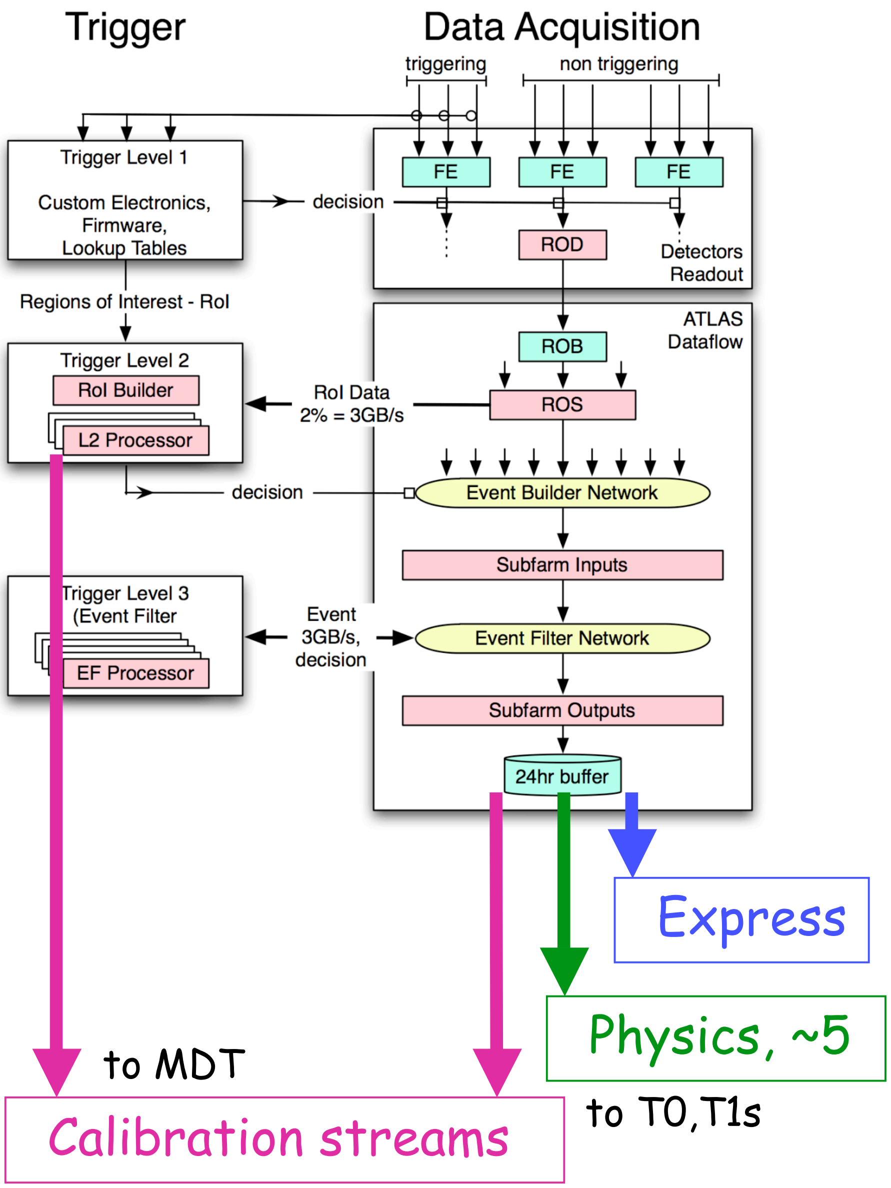


Data transferred (GB/day)
Aug 23 – Sep 8



Completed file transfers
Aug 23 – Sep 8

Data Streaming



Express

Physics, ~5

to MDT

Calibration streams

to T0,T1s

- There is no 'obvious' right way to stream
  - Flexibility is vital
  - Overlaps vary with luminosity
- Streaming (RAW, ESD, AOD) is based on trigger decisions
- Raw data streams baseline
  - ~5 physics streams, Express stream, Calibration streams
  - Physics streams are inclusive: one event may be in >1 streams (e+γ, μ+Bphys, jets, τ+E_t^{miss} , minbias; overlap ~10%)
- ESD streams = Raw streams
- AOD streams from central production and reprocessing
  - Respect Raw/ESD boundaries, may split streams

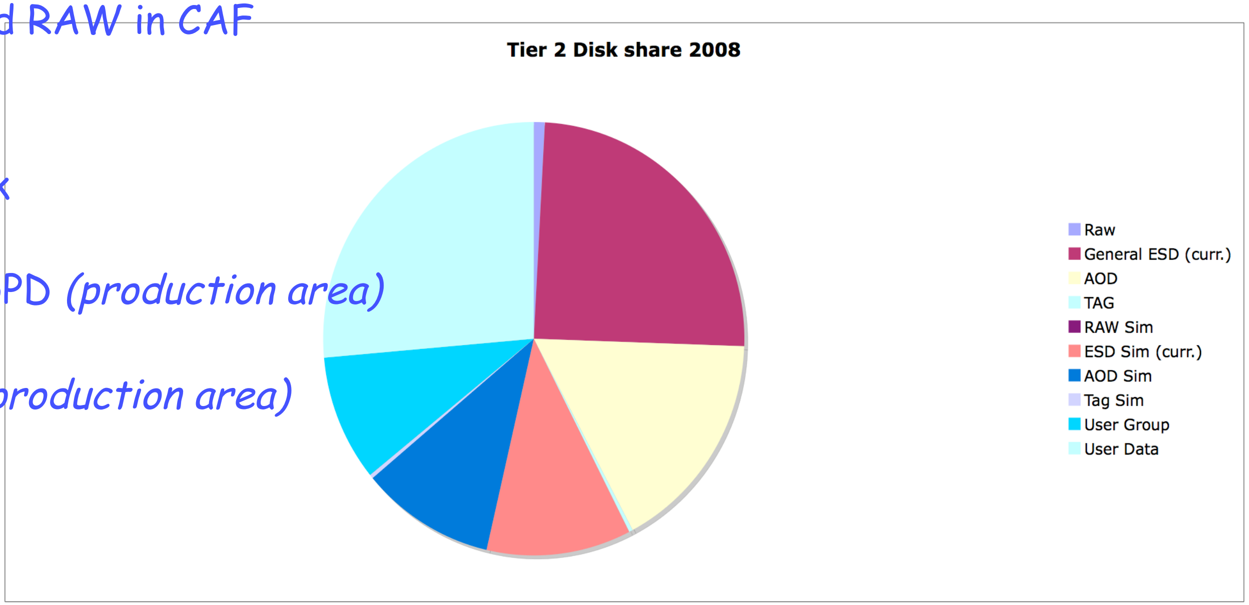
Analysis Formats

- Evolving view of what 'Derived Physics Datasets' (DPD) are
  - In the C-TDR, used to represent many derivations
    - Skimmed AOD, data collections, augmented AOD, other formats (Athena-aware Ntuples, root-tuples)
  - Much effort to see if one format can cover most needs
    - Saves resources
    - But diversity will remain
    - 'Everyone ends-up with a flat tuple'
  - In each case, aim is to be faster, smaller and more portable
  - May span data streams
- Group-level DPD to be produced in scheduled activity at Tier 1s
  - 'Big trains' becomes small meetings!
  - Overall co-ordinator, production people in each group

Tier-2 Data on Disk

- ~30 Tier-2 sites of very, very different size contain:
- Some of ESD and RAW
  - In 2007: 10% of RAW and 30% of ESD in Tier-2 cloud
  - In 2008: 30% of RAW and 150% of ESD in Tier-2 cloud
  - In 2009 and after: 10% of RAW and 30% of ESD in Tier-2 cloud
  - *This will largely be 'pre-placed' in early running*
  - *recall of small samples through the group production at T1*
- Additional access to ESD and RAW in CAF
  - 1/18 RAW and 10% ESD
- 10 copies of full AOD on disk
- A full set of official group DPD *(production area)*
- Lots of small group DPD *(in production area)*
- User data
  - Access is 'on demand'

How does all this map to space tokens if at all?



On-demand Analysis

- Restricted Tier 2s and CERN Analysis Facility (CAF)
  - *Note: CAF is 'on demand', group analysis at T1 is scheduled*
  - Can specialise some Tier 2s for some groups
  - All Tier 2s are for ATLAS-wide usage
- Most ATLAS Tier 2 data should be 'placed' with lifetime ~ months
  - *Lifetime matches ~4 group DPDs a year*
  - Tier 2 bandwidth is vastly lower, job efficiency higher
  - *Group DPD in 'production' area and 'pinned' to disk*
- *Role and group based quotas are essential (but are not emerging quickly!)*
  - *CPU fair-shares are quite easily done*
  - *We can at least easily split 'production' from user space*
  - *Quotas to be determined per group not per user*
    - *User files can be garbage collected - effectively ~SCR$MONTH unless 'adopted' by a physics/detector group*
    - *'Adoption' implies the group 'production' role moves (or reallocates) the files into the group's production quota and 'pins' it*
  - *The details of this migration need to be fleshed-out*
  - *The details of the garbage collection also need to be fleshed-out*
    - *By ATLAS, as deleted files need to be removed from the catalogues*

Tier-3s?

- These will have many forms
- Basically represent resources not for general ATLAS usage
  - Some fraction of T1/T2 resources
  - Local University clusters
  - Desktop/laptop machines
  - T3 task force will provide recommended solutions (plural!)
- Concern over the apparent belief that T3s can host large samples
  - Required storage and effort, network and server loads at T2s
- Network access
  - ATLAS policy in outline:
    - O(10GB/day/user) who cares?
    - O(50GB/day/user) rate throttled
    - O(10TB/day/user) user throttled!
    - Planned large movements possible if negotiated

Early Data

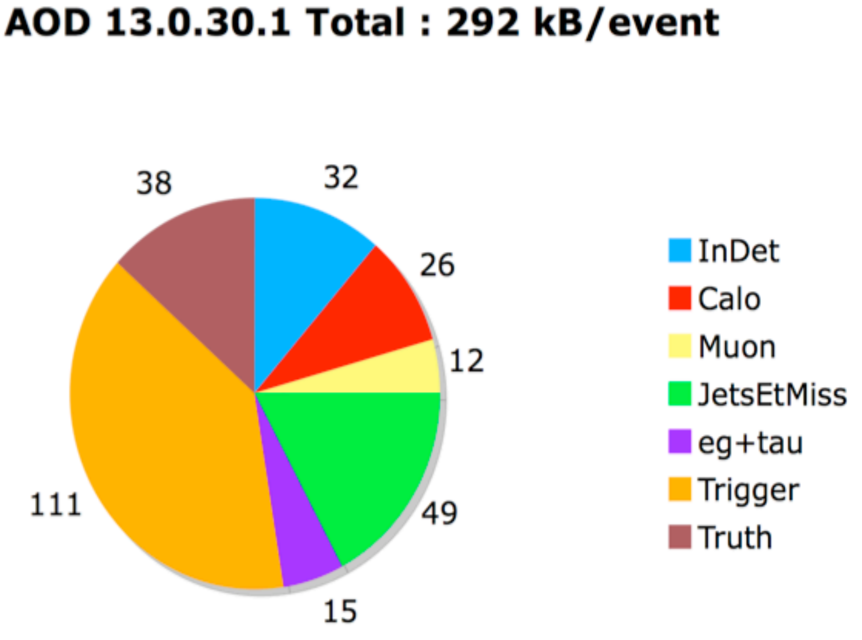
- Storage: has to be in place before usage
  - Possible to use in the short term
    - More ESD - so long as you clear the extra events for new data
    - Bigger AOD - so long as you reduce it later
  - Hard to remove AOD features from users
  - Note: augmented AOD for well defined subsets or tasks is not problem
    - This is a use case for group DPD!
- CPU:
  - At the Tier 1, the CPU is going to be busy much of the time
    - The full reprocessing will obviously wait for calibration and algorithmic development, so the capacity is available until then
    - *The group analysis/big trains have a large resource allocation: the balance between DPD production and reprocessing is adjustable*
  - We anticipate some samples being reprocessed often
    - Output short-lived
    - 'Proper' processing for physics results
    - Must beware inconsistent processing with inclusive streams

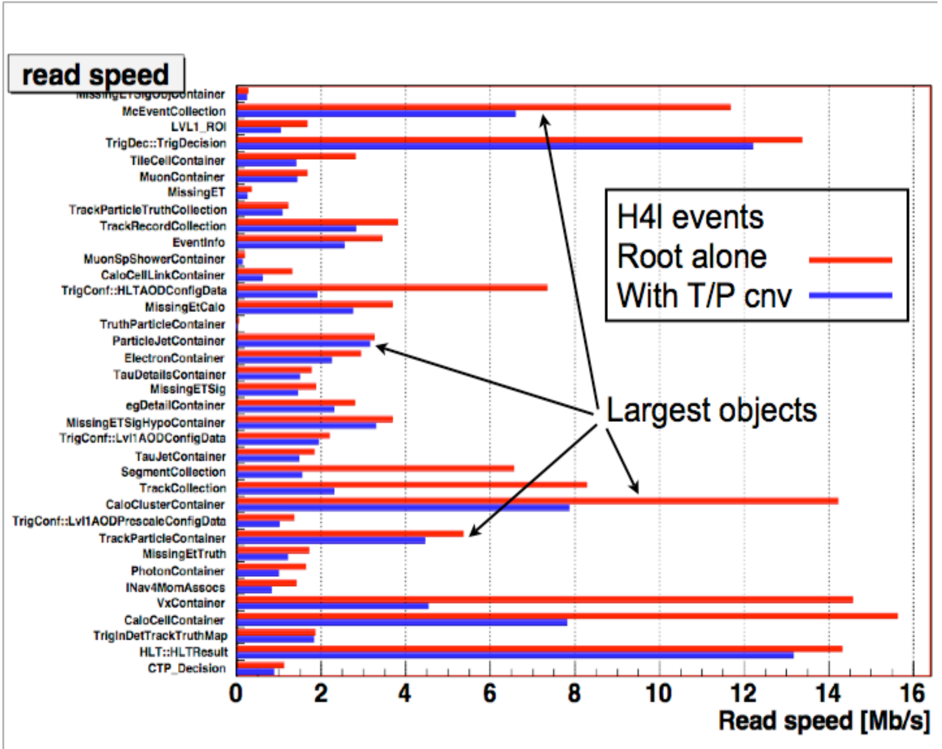
Other Issues

- Event size matters!
  - Lots of good work on the event sizes etc
  - Hopeful we will have 'good' event sizes
- Simulation times
  - We seem to be running to stand still!
    - Again lots of hard work, but from my uninformed position each reduction in time seems to be matched by a slow down for better modelling
    - Ultimately, this could limit our simulation numbers and our physics, so the trade-off must be kept in mind
    - If there are samples that can be done with faster options, we should use them
- Memory
  - Memory is similar cost to CPU
  - Memory costs are volatile
    - Tier 1s purchase often and can benefit from drops
    - Tier 2s purchase less often and have to pay the price
      - Much of the Tier 2 capacity for day 1 is already bought
    - Simulation nodes need 2GB/core (memory more expensive than the CPU)
    - Tier 2 analysis nodes often have lower memory
      - Especially reconstruction must try to keep profile down to allow use on most Tier 2 nodes
  - If we want high memory nodes, we need to reduce our CPU expectations (money is largely fixed!)

Event Data Model: Size and Performance

- ESD Size: Rel 12 (~1700kB) => Rel 13 (~800kB)
- AOD Size: Rel 12 (~200kB) => Rel 13 (~250kB)
  - Truth reduced, added egamma/muon track + calo cells
  - Trigger EDM size x0.5 but larger exploratory menu and lower thresholds
  - Still duplications (muon tracking, jet collections, etc.)
- Early work on investigating and improving the I/O performance

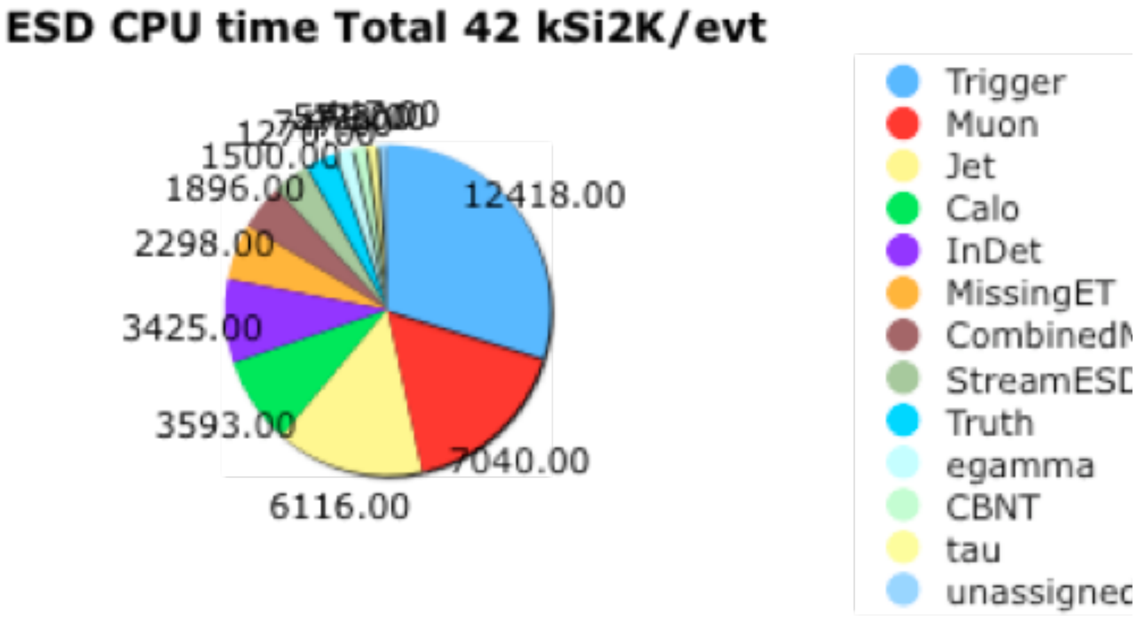


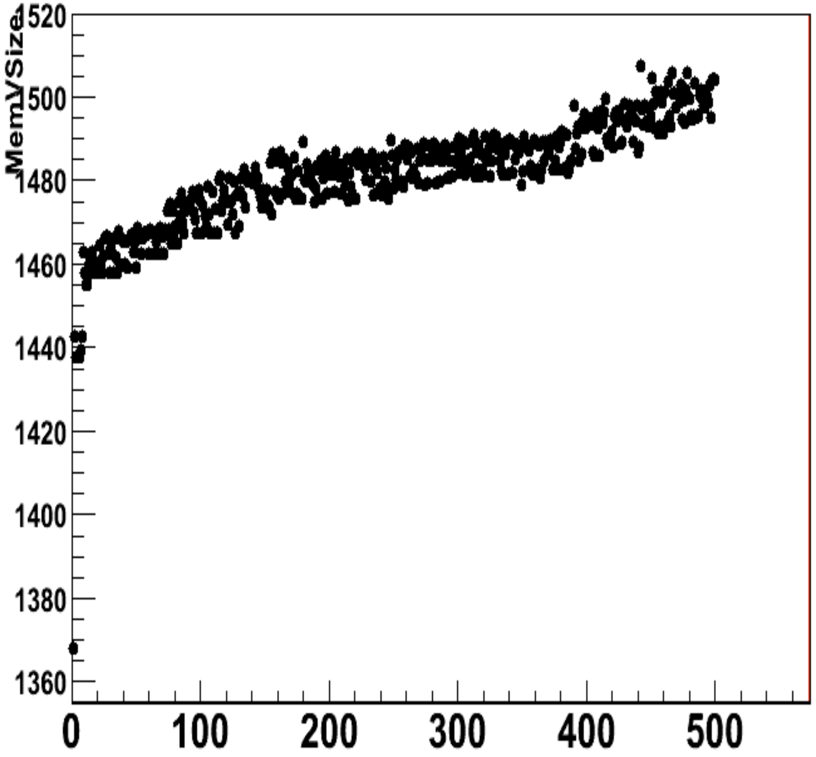


Reconstruction: Need Optimization Work

- Configuration, documentation
- Preparation for data (add DQ monitoring; express stream handling, etc.)
- CPU - improved tools (PTF); remove duplications; focussed optimizations
- Memory size and Leaks
  - Size <1.9MB
  - Leak <100kB (was ~500kB for 13.0.20)

ESD CPU time Total 42 kSi2K/evt



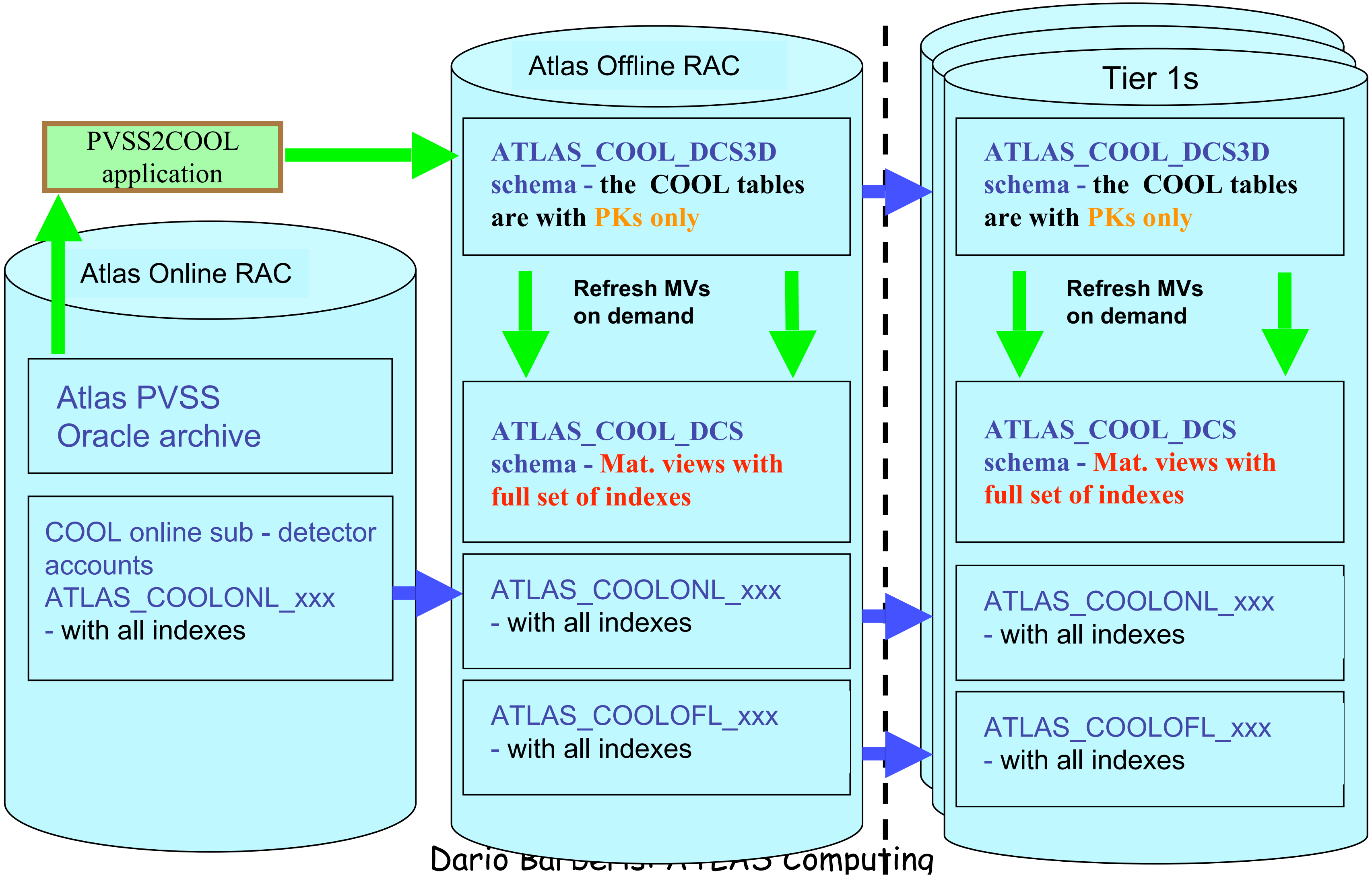


Physics Analysis Tools

- People start using TAGs
  - Support direct navigation to events (RAW, ESD, AOD)
  - A selection of e.g. 5% of events via TAG query is really x20 faster than reading all events and rejecting 95%
- DPD (Derived Physics Data) topics
  - More support for thinning (dropping of containers from event)
  - Improved Athena/ROOT access to user data
- Possible role of PROOF parallel analysis facilities
  - Tier-3 task force
- Interactive Athena and/or ROOT; C++ and/or Python
  - Support for different working styles improving (essentially symmetrical)
    - Apart from lack of database access from within pure ROOT
  - Support CINT, Python or compiled C++
    - CINT is slowest (x10), then Python (x2-3), then C++
- Extensions to EventView toolkit underway

Database Replication

- Oracle Streams in production use: ATONR -> ATLR -> all Tier1s
  - Used for almost everything



PVSS2COOL application

Atlas Online RAC

Atlas PVSS Oracle archive

COOL online sub - detector accounts ATLAS_COOLONL_xxx - with all indexes

Atlas Offline RAC

ATLAS_COOL_DCS3D schema - the COOL tables are with PKs only

Refresh MVs on demand

ATLAS_COOL_DCS schema - Mat. views with full set of indexes

ATLAS_COOLONL_xxx - with all indexes

ATLAS_COOLOFL_xxx - with all indexes

Tier 1s

ATLAS_COOL_DCS3D schema - the COOL tables are with PKs only

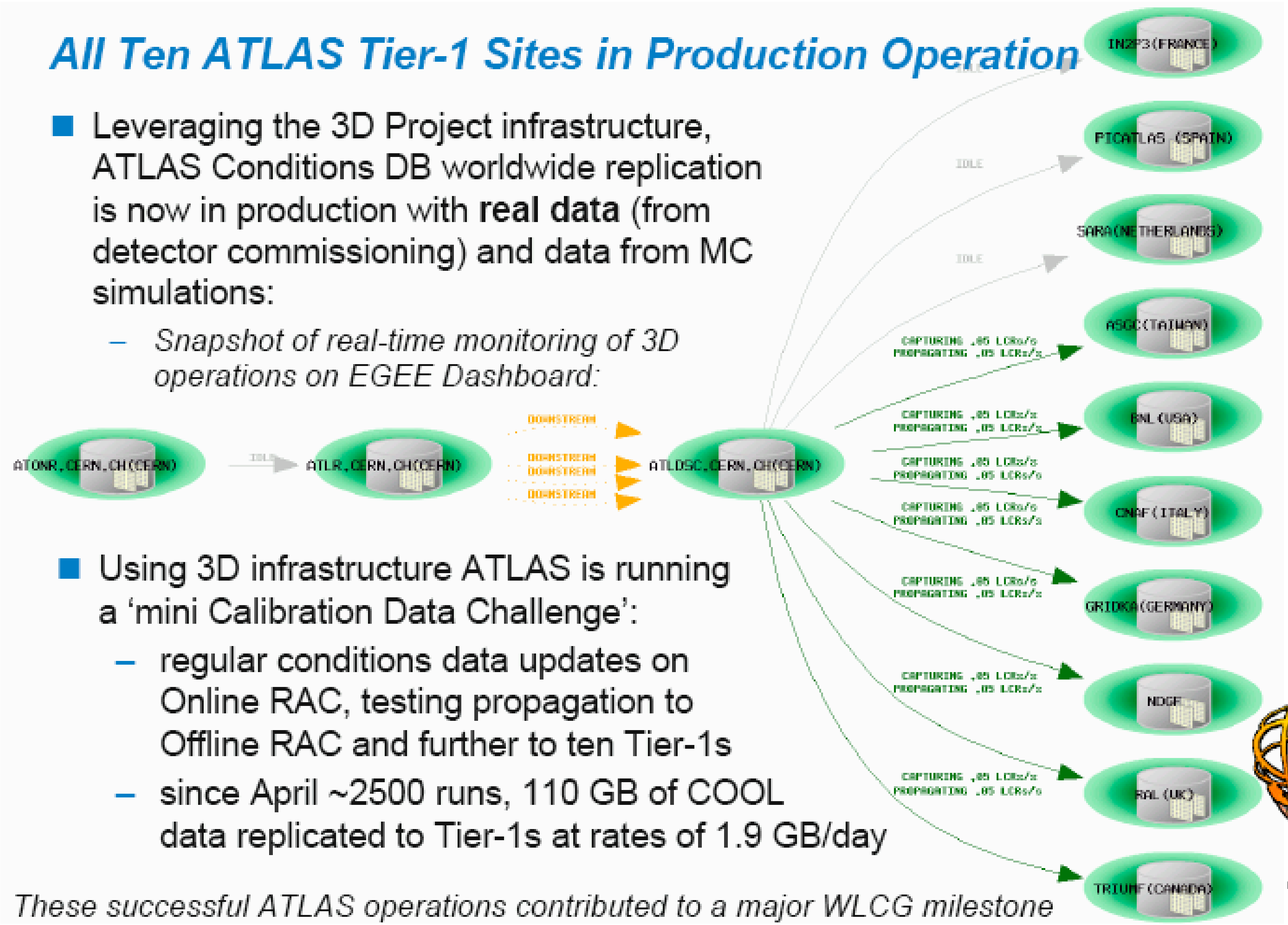
Refresh MVs on demand

ATLAS_COOL_DCS schema - Mat. views with full set of indexes

ATLAS_COOLONL_xxx - with all indexes

ATLAS_COOLOFL_xxx - with all indexes

Database Replication - Status



All Ten ATLAS Tier-1 Sites in Production Operation

- Leveraging the 3D Project infrastructure, ATLAS Conditions DB worldwide replication is now in production with **real data** (from detector commissioning) and data from MC simulations:
  - *Snapshot of real-time monitoring of 3D operations on EGEE Dashboard:*

- Using 3D infrastructure ATLAS is running a ‘mini Calibration Data Challenge’:
  - regular conditions data updates on Online RAC, testing propagation to Offline RAC and further to ten Tier-1s
  - since April ~2500 runs, 110 GB of COOL data replicated to Tier-1s at rates of 1.9 GB/day

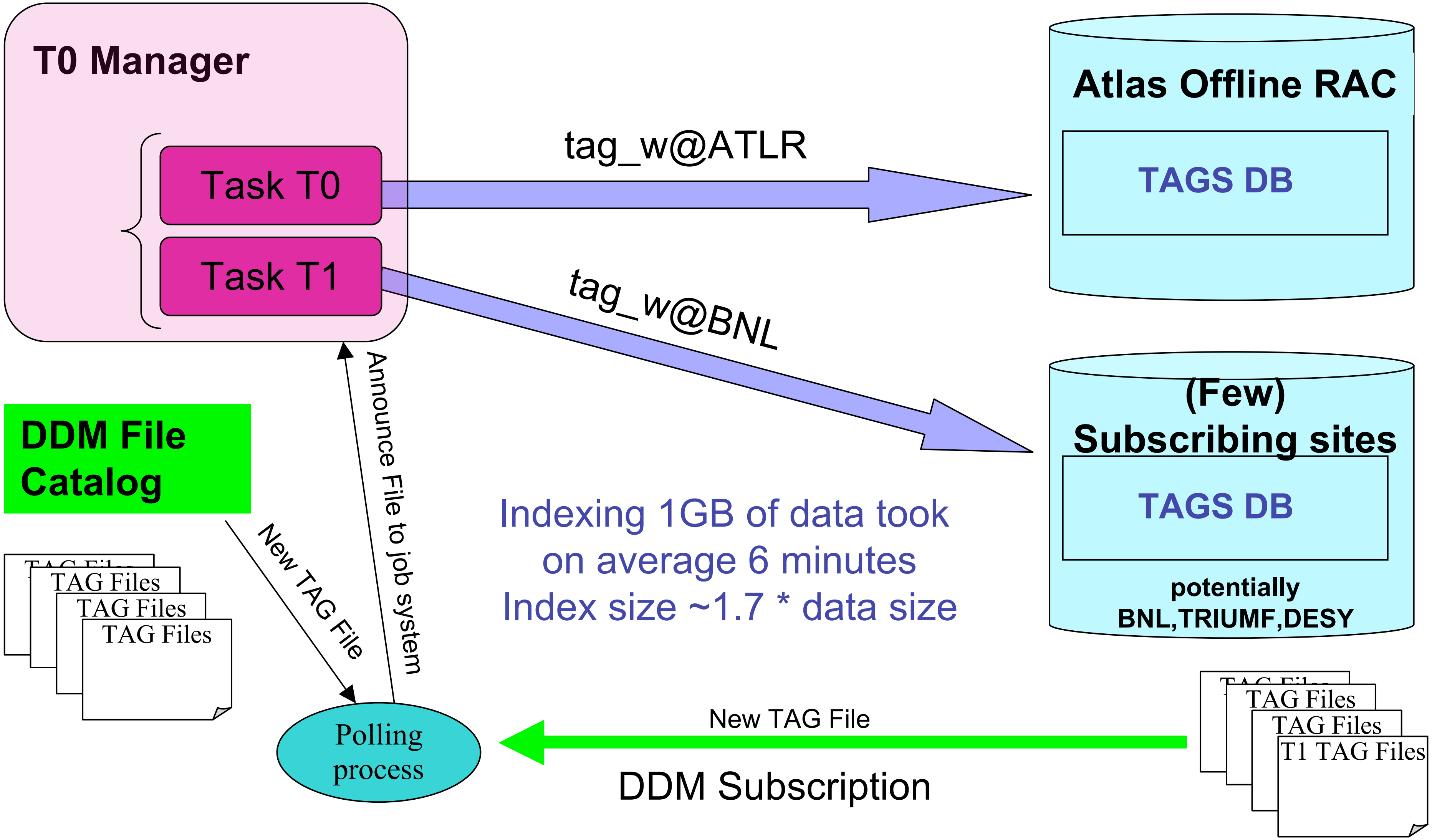
These successful ATLAS operations contributed to a major WLCG milestone

Tests of TAGs

Two types of tests done:

- Small data volume (~3M events, 5h running, 4GB) "real" physics content
  - In the framework of the data streaming tests
  - "real" means mixed Monte-Carlo events without truth information
  - Event selection done with TAGs for "real" analyses
- Large data volume (~0.6G events, 1/4 yr running, 1TB) of fake data
  - To measure timing performance for event counting and for event sample writing
  - Queries got optimized - interactive speed obtained for counting-type queries, and ~1 job/min for a sequence of several counting then one writing query (mimics user workstyle)
  - Also tested replication of TAG database

Loading of TAGs centrally



T0 Manager

Task T0

Task T1

tag_w@ATLR

Atlas Offline RAC

TAGS DB

tag_w@BNL

(Few) Subscribing sites

TAGS DB

potentially BNL,TRIUMF,DESY

DDM File Catalog

New TAG File

Announce File to job system

Polling process

TAG Files

TAG Files

TAG Files

TAG Files

T1 TAG Files

New TAG File

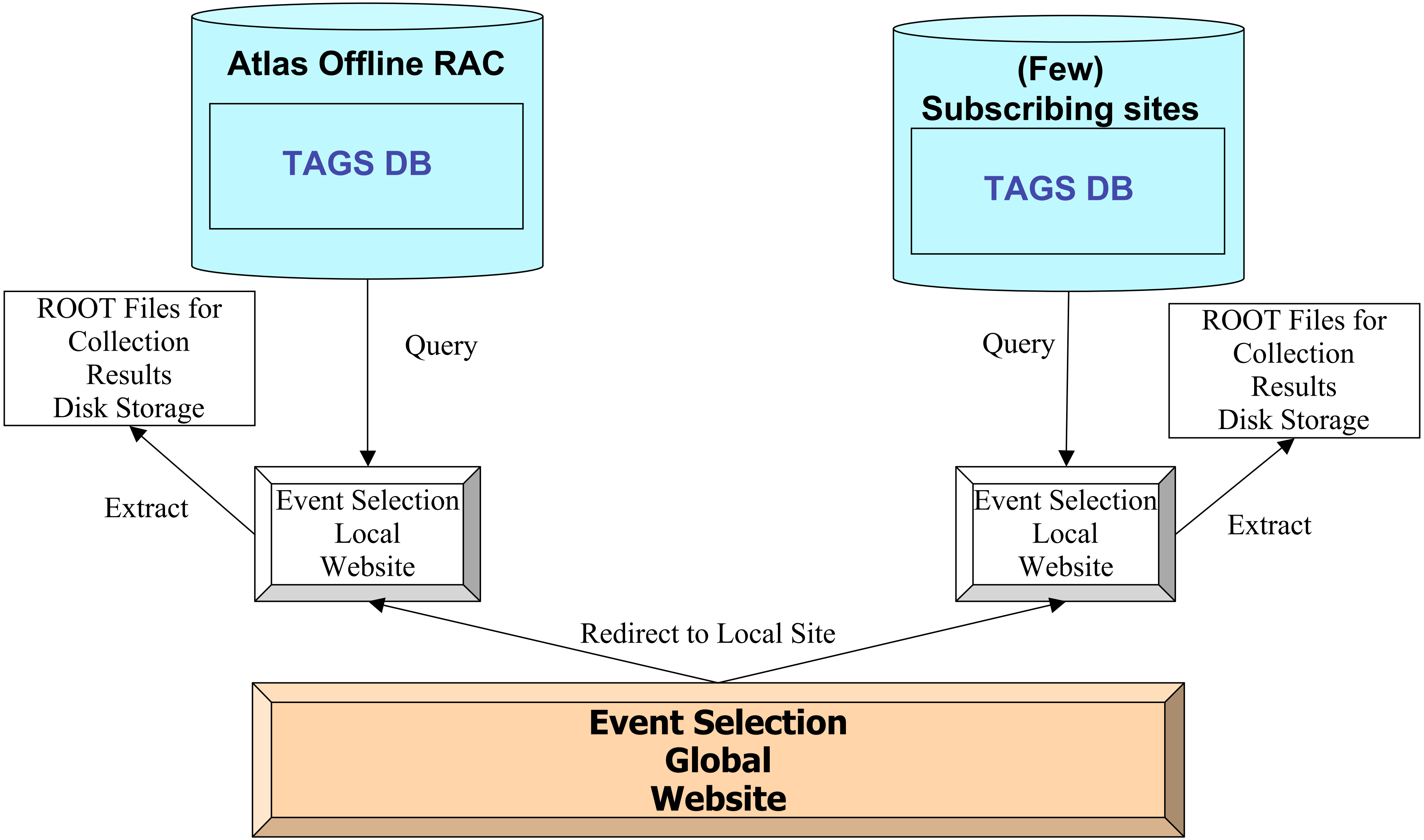
DDM Subscription

Indexing 1GB of data took on average 6 minutes

Index size ~1.7 * data size

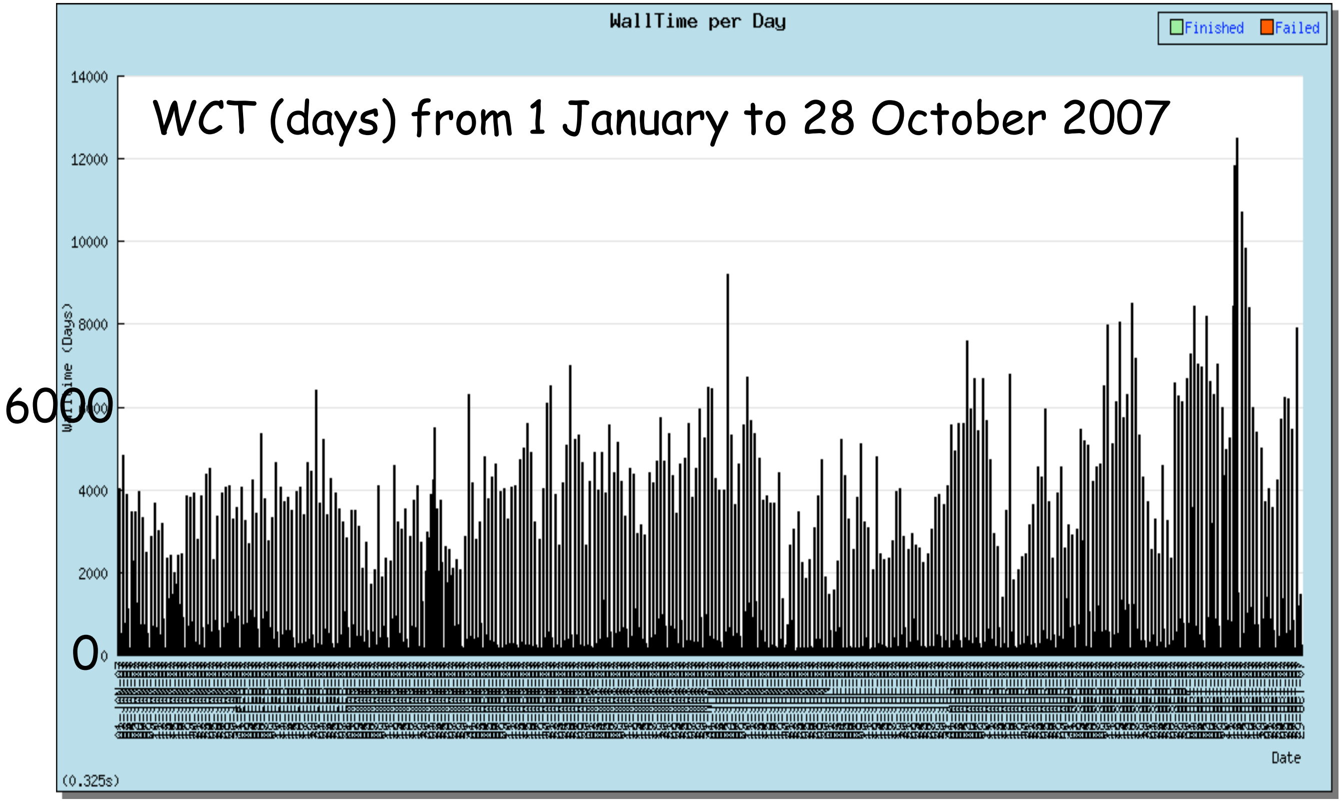
→ POOL Collection utilities

Usage of local TAGs via Web Interface


Atlas Offline RAC
TAGS DB
(Few) Subscribing sites
TAGS DB
Query
Query
ROOT Files for Collection Results Disk Storage
ROOT Files for Collection Results Disk Storage
Extract
Extract
Event Selection Local Website
Event Selection Local Website
Redirect to Local Site
Event Selection Global Website

Distributed Simulation Production

- Simulation production continues all the time on the 3 Grids (EGEE, OSG and NorduGrid) and reached 1M events/day recently
  - The rate is limited by the needs and by the availability of data storage more than by resources
- Validation of simulation and reconstruction with release 13 is in progress
  - Large-scale reconstruction will start soon for the detector paper and the FDR


WCT (days) from 1 January to 28 October 2007

Production at Tier-0/1/2/3…

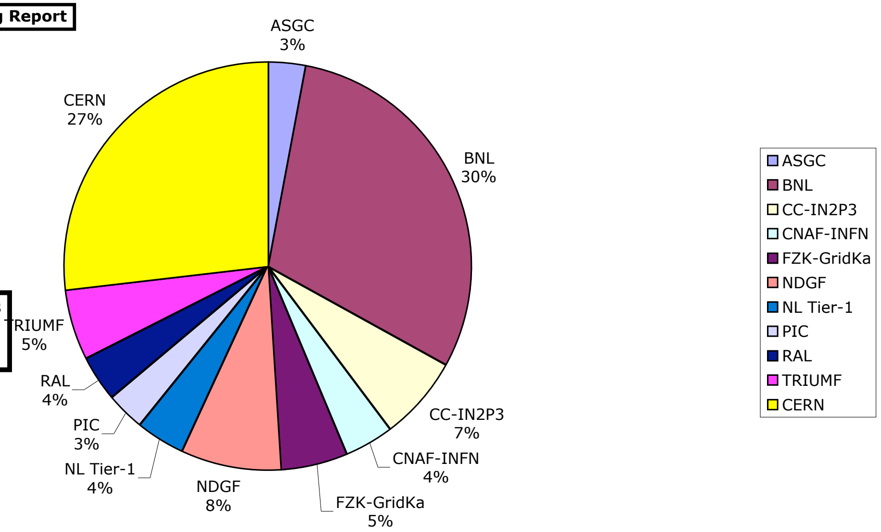
ATLAS CPU at Tier-1s & Tier-0 in 2007 (Jan-Sep)

Source: WLCG Accounting Report

~ 952000 kSi2k-days

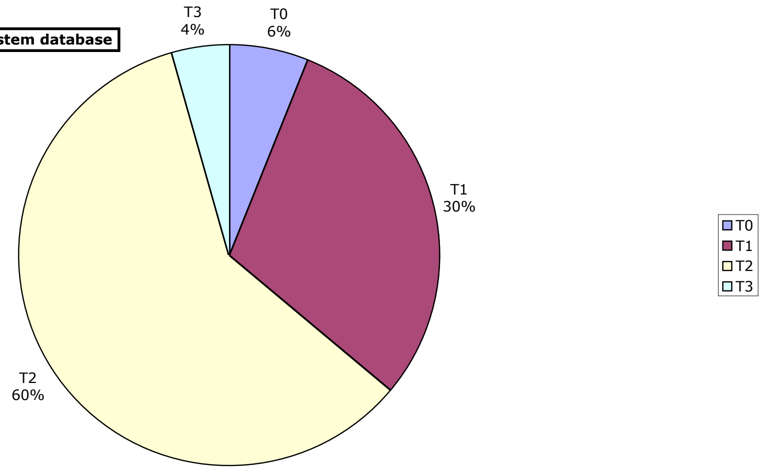
~ 3500 kSI2k

Grid: ~ 72 %
Non-Grid: ~ 48 %



Production at ATLAS Tiers - Number of jobs (Jan-Sep 2007)

Source: ATLAS Production System database

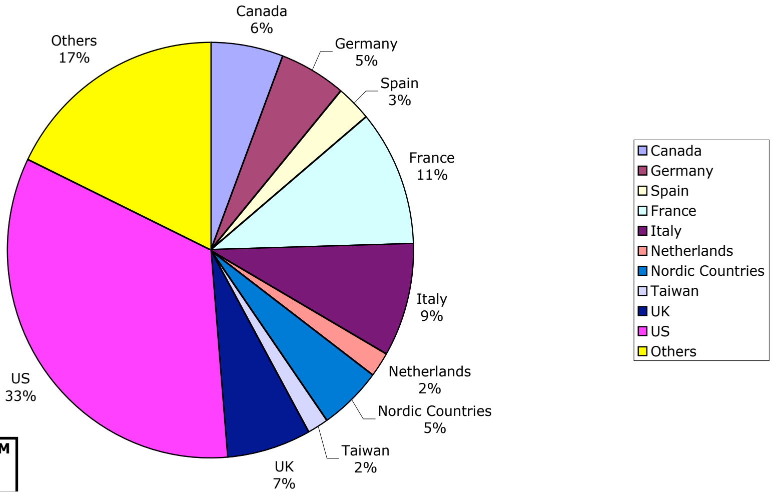


Total number of jobs: ~ 5M

Efficiency: ~70%

ATLAS Production (per country) - Number of Jobs (Jan-Sep 2007)

Source: ATLAS Production System database

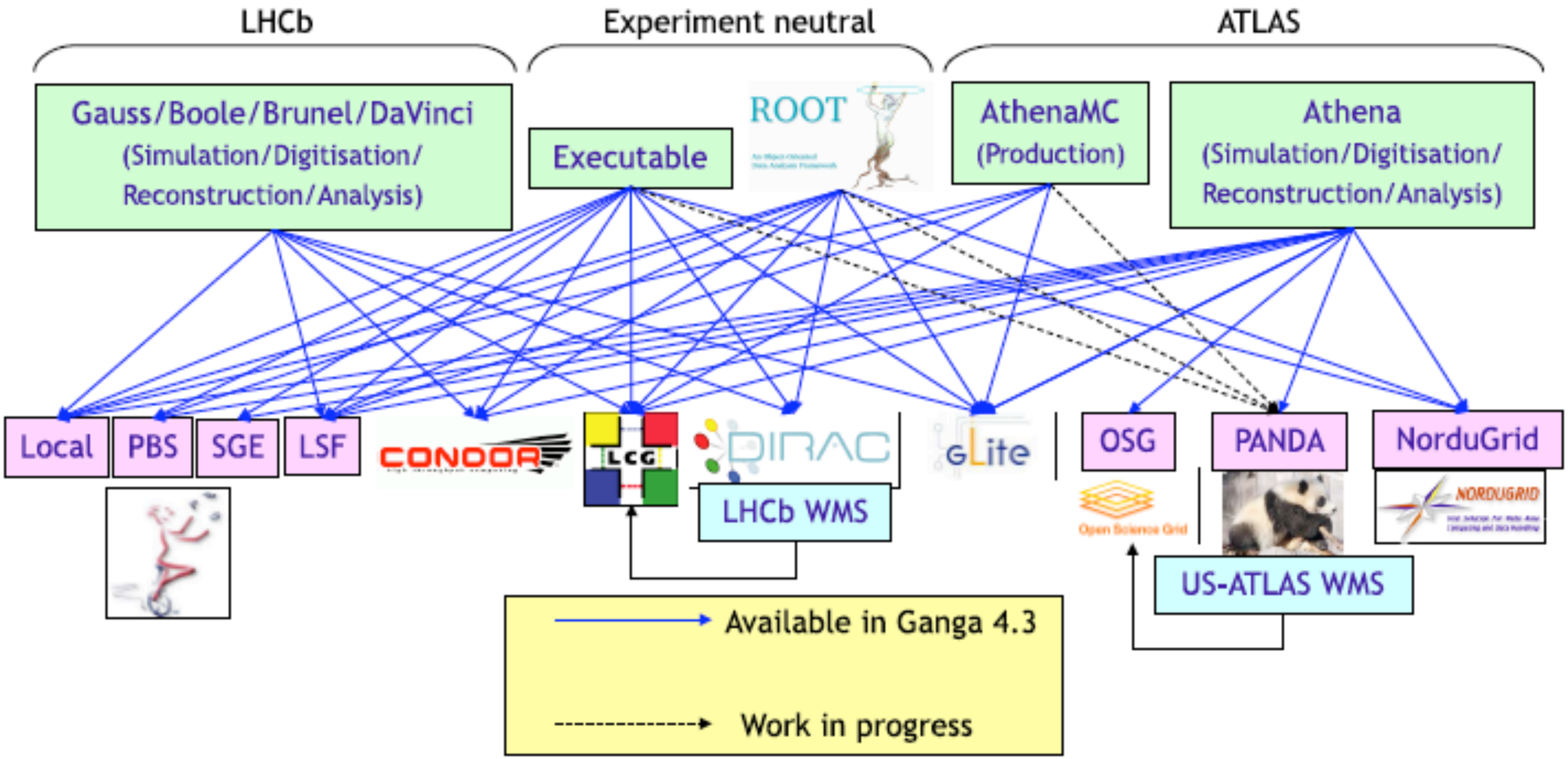


Total number of jobs: ~ 5M

Efficiency: ~70%

Distributed Analysis with Ganga (1)

- GANGA simplifies running of ATLAS (and LHCb) applications on a variety of Grid and non-Grid back-ends



LHCb

Experiment neutral

ATLAS

Gauss/Boole/Brunel/DaVinci (Simulation/Digitisation/Reconstruction/Analysis)

Executable

ROOT

AthenaMC (Production)

Athena (Simulation/Digitisation/Reconstruction/Analysis)

Local

PBS

SGE

LSF

CONDOR

LCG

DIRAC

gLite

OSG

PANDA

NorduGrid

LHCb WMS

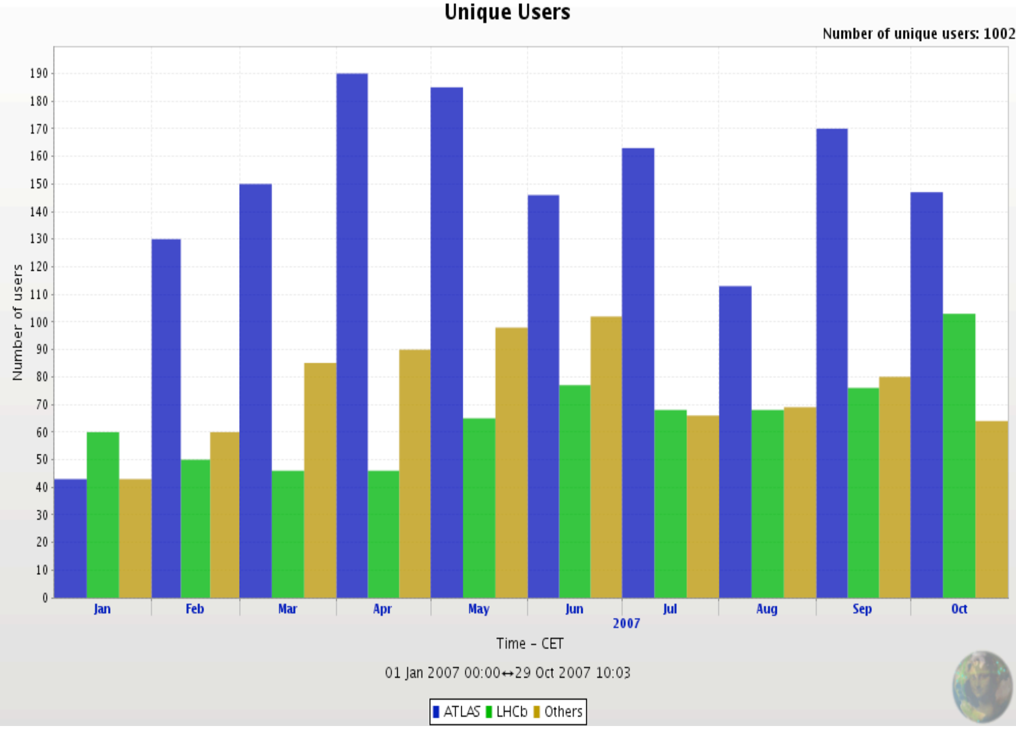
US-ATLAS WMS

Available in Ganga 4.3

Work in progress

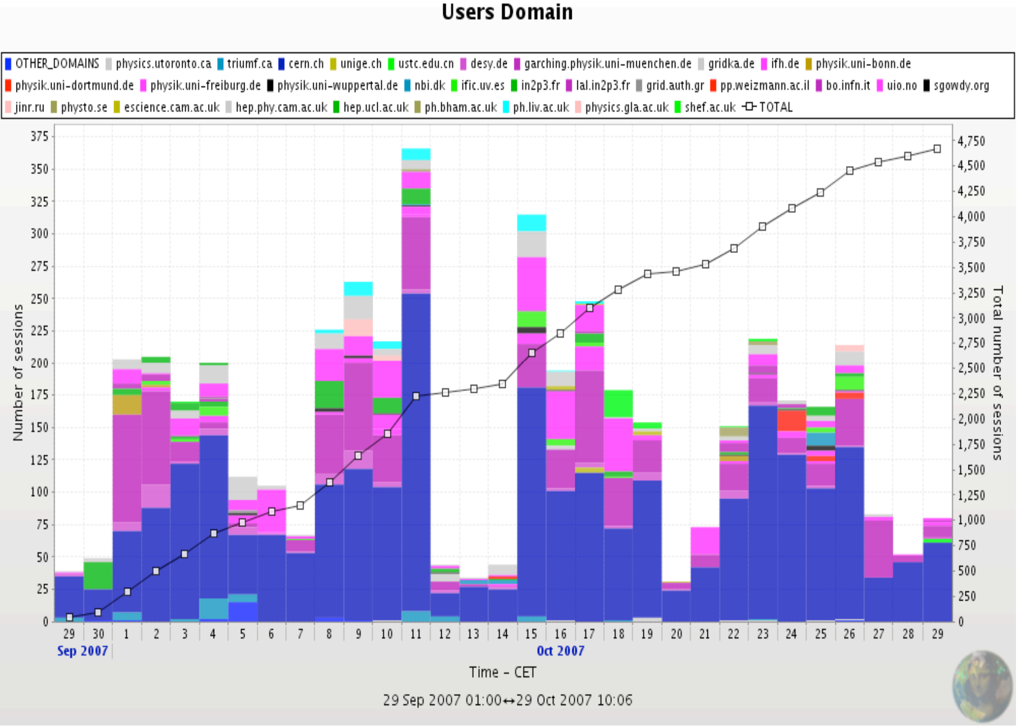
Distributed Analysis with Ganga (2)

- ATLAS end users are finally learning to use the appropriate tools (such as Ganga) to send jobs to their input data
  - Rather than copying files to their local computing clusters and running locally, which turns out to be close to impossible
- A major improvement in job splitting and brokering is forthcoming:
  - Use the information from DDM catalogues to split jobs along the boundaries of (parts of) datasets present at different sites, then direct each subjob to the location of all its input files



Unique Users

Number of unique users: 1002

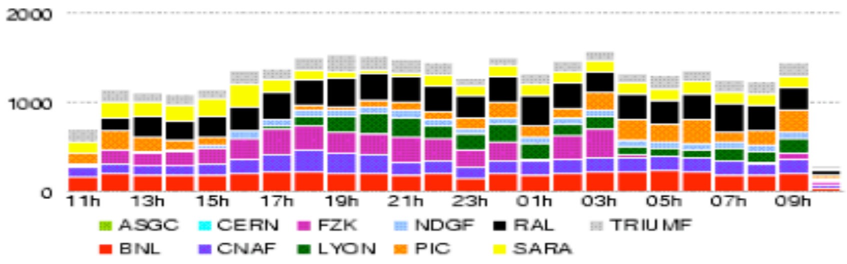


Users Domain

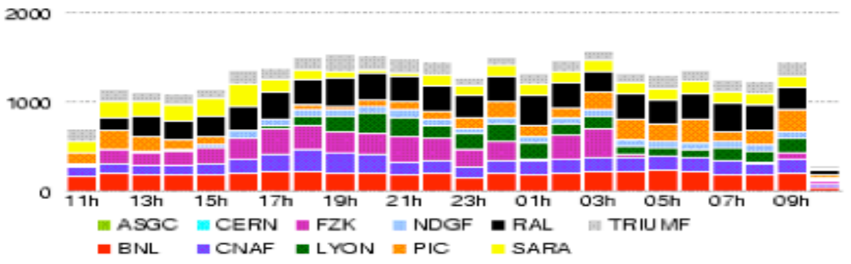
Tier-0 → Tier-1 export works

19-20 October 2007

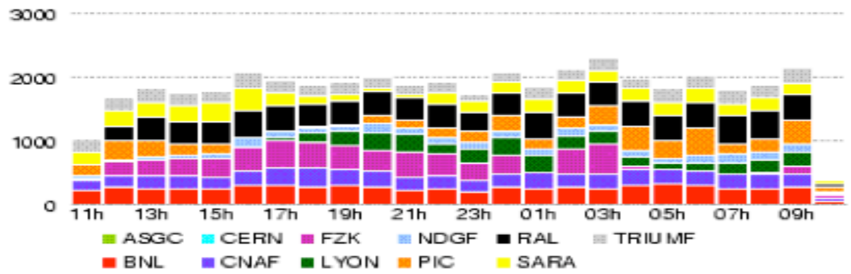
Throughput MB/s



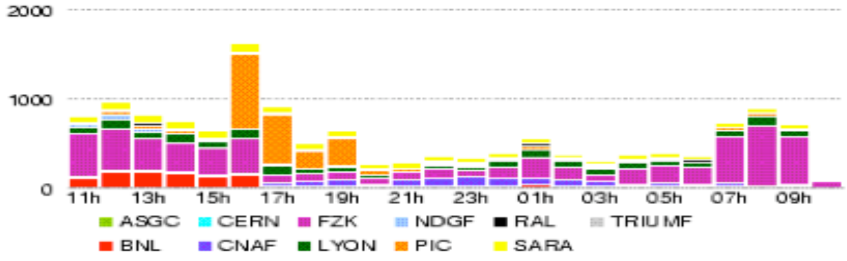
Data transferred GB



Completed filetransfers



Total number of errors



- Last data throughput tests showed that all obstacles to data export from CERN have been identified and removed
  - An export rate ~1.2 GB/s could be sustained for prolonged periods using an incomplete set of Tier-1s
  - BNL took less than their nominal rate (but we know they can take a lot more)
  - ASGC was not included but will join in next time (November) as problems were since fixed

Data Distribution Tests

- The throughput tests will continue (a few days/month) until all data paths are shown to perform at nominal rates
  - This includes:
    - a) Tier-0 → Tier-1s → Tier-2s for real data distribution
    - b) Tier-2 → Tier-1 → Tier-1s → Tier-2s for simulation production
    - c) Tier-1 ⇔ Tier-1 for reprocessing
- Test a) is now OK almost everywhere; next rounds will concentrate on b) and c)
- The Functional Test will also be run in the background approximately once/month in an automatic way
  - The FT consists in low rate tests of all data flows, including performance measurements of the completion of dataset subscriptions
  - The FT is run in the background, without requiring any special attention from site managers
  - It checks the response of the ATLAS DDM and Grid m/w components as experienced by most end users

Distributed Computing (1)

- Until now we had two separate areas within Software & Computing, covering respectively the development and operation of Grid Tools & Services
  - This structure turned out to be less than optimal to ensure good communication between developers and operators, and also cross-communication between activity areas
  - We are also limited by available manpower; separating activities too much meant that people focussed their attention on their narrow area of responsibility, creating many potential single points of failure
  - An example is the situation with small files generated by simulation production; while it was evident to DDM developers that small files are not a good idea, production operators optimised their setup to improve job efficiency (i.e. having shorter jobs that produce small output files)

Distributed Computing (2)

- To overcome this situation, we decided to create a "Distributed Computing" project, that includes both development and operations activities, within which people can be assigned to tasks in a more flexible way
  - As in the near future the needs of operations have to set the priorities for everybody, KORS BOS, currently Computing Operations Coordinator, will lead the Distributed Computing Project
    - Jim Shank will be Deputy Distributed Computing PL
    - Massimo Lamanna will be responsible for all development activities
    - Alexei Klimentov will be responsible for all operation activities
- The first task of the 4 people named above (plus myself) is to write down, in close consultation with all people currently involved in these activities:
  - A description of scope and organisation of the Distributed Computing project
  - The global system architecture we realistically think we can have in mid-2008
  - The work plan to get there
  - The list of deliverables and milestones, taking external constraints into account
    - M* runs, SRM2.2 readiness, FDR, CCRC, see later slides
  - The manpower needed and available for each task
- As soon as this is completed (in 2-3 weeks), the new organisation will be effective

Evolution of the Production System

- During the ATLAS Computing Operations Meeting in the Software & Computing Week it was discussed and decided that the ATLAS production system will evolve towards having just one way of submitting and running production jobs on the OSG and EGEE Grid resources.
  - A suite of ATLAS and Middleware tools and services (the new names of Pallas and Palette were proposed) will be selected to make this happen.
  - Two important choices of input to the baseline system were made already during the meeting: the Panda pilot job technology and the Local File Catalog LFC will be used.
- While this may have longer term implications for distributed analysis, the decision does not imply that the same tool will be used for that purpose; both the problems to be addressed and the scale are rather different.
- In the short term, while the developers work together to turn the currently available set of tools into a coherent and modular system suitable for the longer-term production needs of ATLAS, the production will continue at full speed with the system used till now, with the usual bug fixes and with the minimal evolutions needed for good operation.
- It was realised that for NorduGrid this evolution would not be straightforward, as pilot jobs do not really fit that architecture, which is already performing very well.
  - NorduGrid and NDGF support the idea of having just one way of submitting jobs to all the grids.
  - A complete and concise technical documentation and a proof of concept of the new system must however be provided before any decision can be made. This concerns both the "pilot job" option of submitting jobs and the choice of the file catalogue.

Global schedule: M*, FDR & CCRC'08

- FDR must test the full ATLAS data flow system, end to end
  - SFO → Tier-0 → calib/align/recon → Tier-1s → Tier-2s → analyse
  - Stage-in (Tier-1s) → reprocess → Tier-2s → analyse
  - Simulate (Tier-2s) → Tier-1s → Tier-2s → analyse
- The SFO→Tier-0 tests interfere with cosmic data-taking
- We must decouple these tests from the global data distribution and distributed operation tests as much as possible
- CCRC'08 must test the full distributed operations at the same time for all LHC experiments
  - As requested by Tier-1 centres to check their own infrastructure
- Proposal:
  - Decouple CCRC'08 from M* and FDR
    - CCRC'08 has to have fixed timescales as many people are involved
    - CCRC'08 can use any datasets prepared for the FDR, starting from Tier-0 disks
    - CCRC'08 can then run in parallel with cosmic data-taking
      - Tier-0 possible interference and total load has to be checked
      - Cosmic data distribution can be done in parallel as data flow is irregular and on average much lower than nominal rates

Plans

- Software releases:
  - 13.0.30.3
    - Week of 05-09 Nov
  - 13.0.40
    - Week of 19-23 Nov
  - 13.1.0
    - Week of 5-9 Nov (note clash with 13.0.30.3 - should be manageable)
  - 13.2.0
    - Week of 3-7 Dec
  - 14.0.X
    - Staged release build starts week of 17-21 Dec; base release 14.0.0 available Mid-end Feb 2008
  - 15.0.X (tentative)
    - Mid 2008
- Cosmic runs:
  - M6
    - (Not earlier than) second half of February 2008
  - Continuous mode
    - Start late April 2008 (depends on LHC schedule)
- FDR:
  - Phase I
    - February 2008 (before M6)
  - Phase II
    - April 2008 (before start of continuous data-taking mode)
- CCRC'08
  - Phase I
    - February 2008 (coincides with FDR/I)
  - Phase II
    - May 2008 (in parallel with cosmic data-taking activities)

Conclusions

- Much progress has been done in the last year
  - And much remains to be done!
- Distributed Data Management (DDM) dominates our worries
  - If it doesn't work, nothing else can work
- Scheduled DDM operations appear to work much better than user access to the data
  - This gives us confidence that the direction is about right
  - But it does not satisfy the average ATLAS physicist who often cannot get to the data as fast as (s)he is used to!
- The new organisation of Distributed Computing will give us more flexibility in assigning people to tasks and focus on the absolute priorities
- The extensive tests planned for the next 6-9 months will help us to finally commission the whole system:
  - Grid middleware tools
  - ATLAS layer and user interfaces