ATLAS Computing Report

Dario Barberis

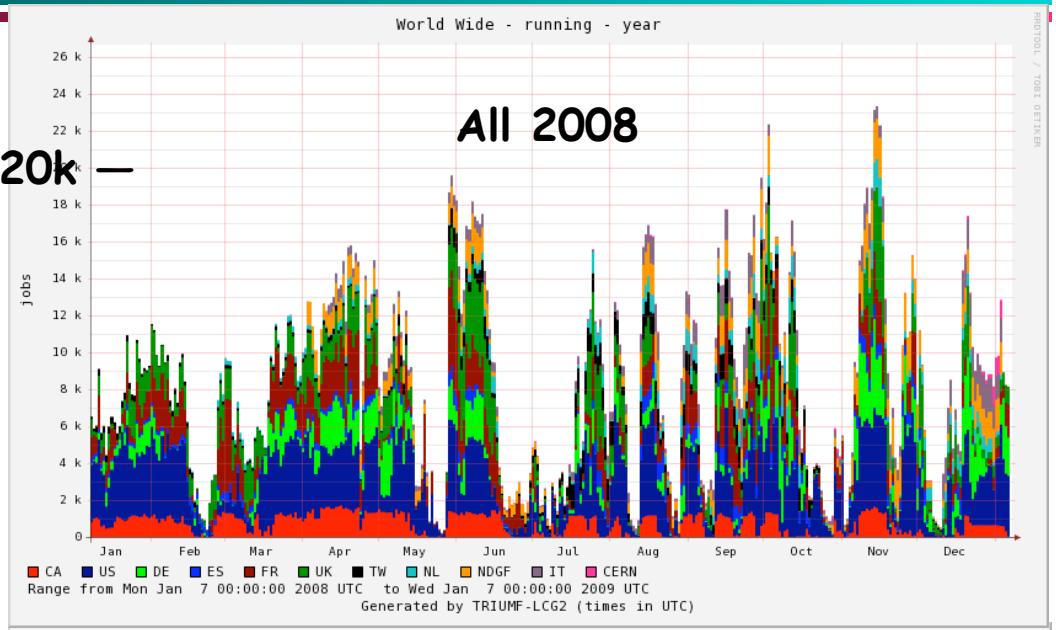
CERN & Genoa University/INFN

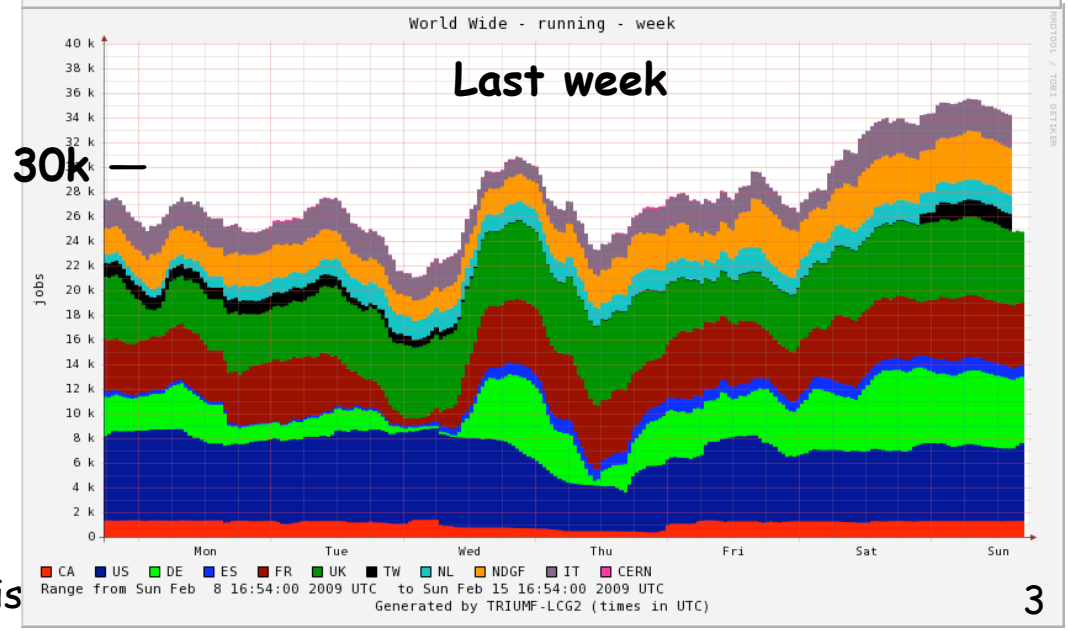
2008 Activity Summary

- Real data:
  - Collected 500 TB of good cosmics data in August-November
    - >200M events with the full detector operational
    - Nominal ~350 MB/s average data-taking rate
  - … and a few runs of single-beam events the week of Sept 10th
  - All RAW data were processed at Tier-0 and distributed according to the Computing Model
- Simulation production:
  - Large production campaigns interleaved by validation of new releases
- Reprocessing:
  - Campaign in December 2008 – January 2009
    - 500 TB of cosmics and single-beam data reprocessed at Tier-1s (and CERN)
  - All outputs distributed to Tier-1/2s according to the Computing Model
- Monitoring and testing:
  - Continuous Functional Tests (FT) of data distribution at 5% level
  - WLCG monitoring tools (SAM etc) extensively used for site status monitoring
  - Stress tests of Distributed Analysis performed periodically on groups of sites

Simulation production

- Simulation production in 2008 used most of the available CPU capacity
  - Apart from the validation phases
- Production will continue, using now Geant4.9
  - The Geant4 step now takes ~1000 kSI2k-s/event
- Fast simulation samples are produced in addition, for large statistics studies
  - Atlfast-II takes on average ~100 kSI2k-s/event





Reprocessing campaigns

- First ever reprocessing campaign of real (cosmic) data took place during the end-of-year 2008 period
  - Software was finally validated on 22nd December
  - Site validation could not be completed on all Tier-1s so we used 8 Tier-1s plus CERN as back-up
    - FZK had problems with glibc version (different numerical results for validation samples)
    - ASGC had mass storage problems throughout December
    - Both problems are now understood and fixed
  - A number of Tier-1 sites were working on "best effort" during the holiday period
  - We decided to reduce risks to the minimum:
    - Decrease the dependencies on Oracle servers
      - More on that in next slide
    - Read input data from disk instead of tape (80% were on disk already)
- Next campaign will be mid-March
  - Same 500 TB RAW data
  - Bug-fixed software
  - Better calibrations

Use of Oracle databases for reprocessing

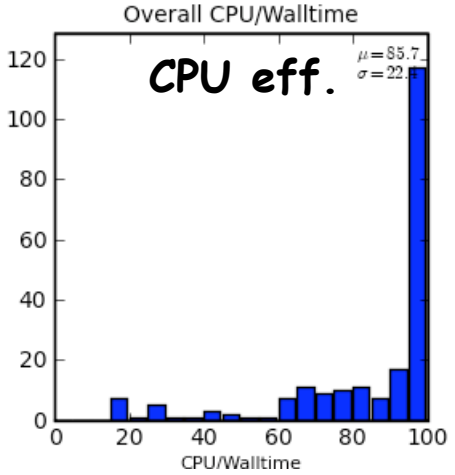
- The original computing model assumed that
  - All conditions data are in the COOL/Oracle databases
  - All calib/align, reprocessing and analysis jobs access conditions data directly from Oracle
- It turns out that:
  - Many detector calibration and alignment constants (LAr and InDet) are too large to store directly in Oracle
    - They are stored instead in POOL files referenced by COOL/Oracle
  - Tests done in Summer 2008 showed that the global load on Oracle servers from O(1000) concurrently running reprocessing jobs at Tier-1s would need a substantial upgrade of database servers at most sites
  - Access by O(1000) concurrent jobs to the POOL calibration files on a local SE can easily block the local storage system
- The preferred (and proven) solution is now the Conditions Database Release
  - A single tar file containing a SQLite database with all data for a given run (or run period) and all POOL calibration file
  - The tar file (1-2 GB) is copied to the worker node and accessed locally
- This solution relieves the Oracle servers from the load due to organised reprocessing tasks and leaves them free for database-intensive tasks, such as calib/align and detector performance analysis

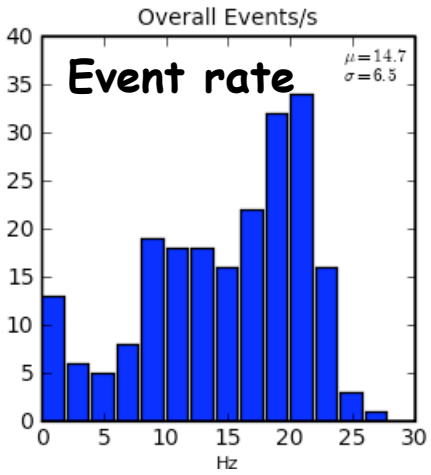
Stress-tests of distributed analysis (1)

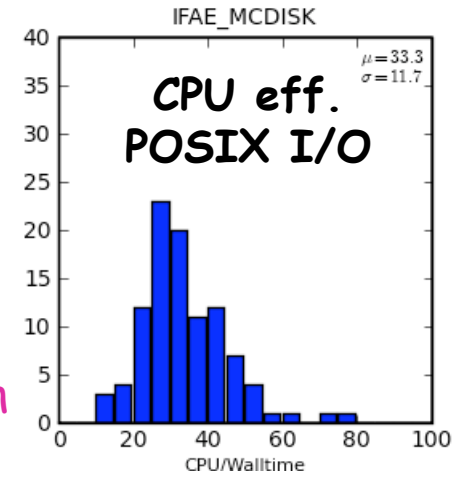
- HammerCloud (HC) is a Distributed Analysis stress testing system
  - It's a database with command-line and web front-ends used for scheduling, running, and reporting on Ganga-based DA stress tests.
  - One specifies where, what, when, and how many jobs to run; HC runs the test and reports the results.
    - At present it can run tests on EGEE/WMS and NDGF/ARC sites.
    - Panda support needs work (in progress).
- Presently we submit large numbers of input data intensive jobs.
  - Focusing on finding site limitations, e.g. heavy loads on storage and network resources.
  - But looking out for grid limitations, e.g. broker, WMS, client.
- Specifically, HC runs an AOD muon analysis everywhere right now
  - Each HC run processes an entire dataset (200 files)
- Soon analyses of real cosmics datasets will be added (with database access)

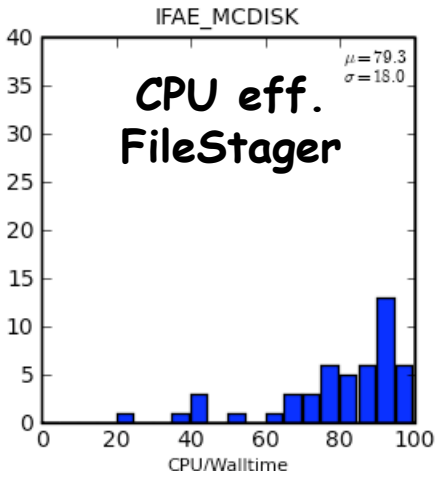
Stress-tests of distributed analysis (2)

- The initial target for these tests was for sites to achieve >80% success efficiency and >15 Hz event rate
  - Example of a good site:
- In several cases we found out that default access methods could be changed or the local hardware setup had limitations
  - Example of improvements:
- Every site has a different hardware setup
  - Every site needs a separate investigation and optimisation
  - Adding CPUs and disks is not enough if the internal network is not correspondingly upgraded at the same time









CERN Computing Facility

- Tier-0: automatic processing tasks
  - Prompt reconstruction
  - Offline data quality monitoring
  - Automatic part of prompt calibration and alignment tasks
    - Tier-0 processing has been tested in advance for several years and exercised throughout 2008 with cosmics and single beam data
- CAF: high-priority central activities
  - Non-automatic prompt calibration and alignment tasks that need iterations and (for the time being) human checking
  - Detector performance investigations
  - Trigger performance studies
    - CAF processing tasks have been developed over the last year and are still under active development
    - The increase of resource needs for the CAF we requested last year is fully justified by the need to have a readily available amount of RAW and processed data for the above tasks, and being able to run them
- User facility: general user access to CERN computing facilities
- Castor pools right now:

Tier-0:	t0atlas+t0merge	CAPACITY 283 TB
CAF:	atlcal+atldata+atlprod+atlspecial	CAPACITY 625 TB
User/Group:	atlasgroupdisk+atlasuserdisk	CAPACITY 215 TB
Tape access:	default	CAPACITY 124 TB
Total:		1247 TB

Evolution of the computing model (1)

- The computing model is fundamentally unchanged since the Computing TDR (2005)
- Some model parameters have been adjusted all along, based on our experience with real data (cosmics) and intensive simulation campaigns during the last few years, and to reflect reality:
  - ESD and AOD sizes doubled to 1 MB and 0.2 MB/event to include more information needed to study detector performance
    - We try to avoid to go back to RAW (from ESD) or ESD (from AOD) when not necessary
    - Sizes will most probably decrease again as we gain experience with real data
  - Full simulation processing times increased from 100 to 1000 kSI2k-s/ev to allow for a very detailed detector geometry and much better physics models in Geant4
    - This level of detail is needed for some physics processes
    - We have developed a complete suite of full, medium and fast simulations that we will use and mix according to the physics needs and the availability of CPU capacity
    - Tuning the Geant4 parameters on real data will help us to optimise CPU usage
  - For reconstruction, CPU time is in the correct ballpark (15 kSI2k-s/event) at low luminosity, but more work is needed for higher luminosities ($>2\times10^{33}$ $cm^{-2}s^{-1}$)
    - Targeted effort for memory reduction is now starting to pay off (improvements in the forthcoming s/w release)

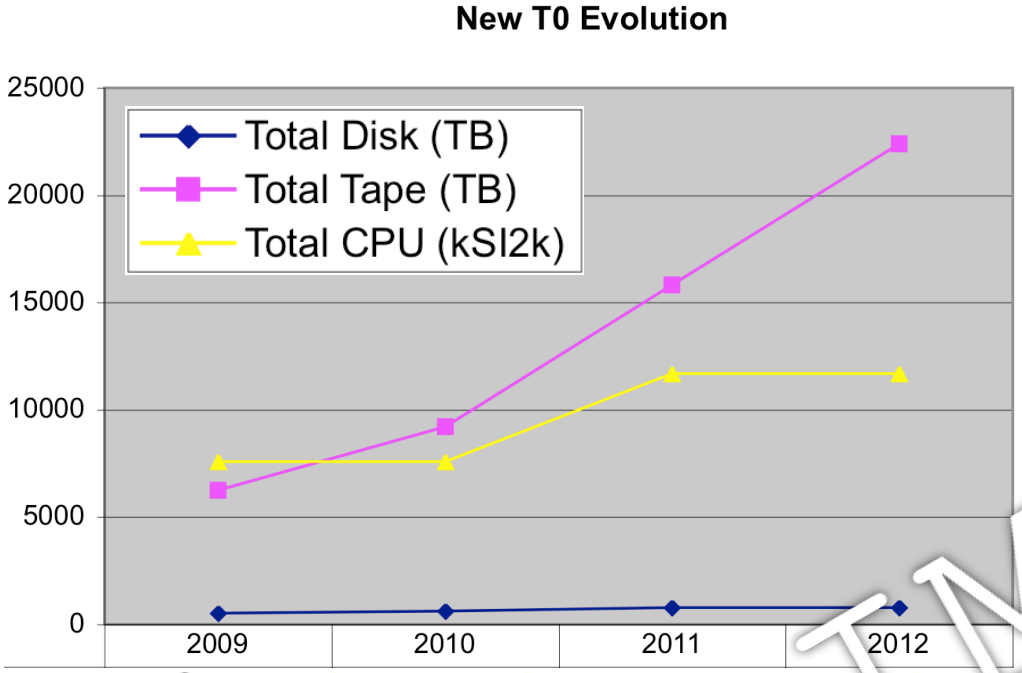
Evolution of the computing model (2)

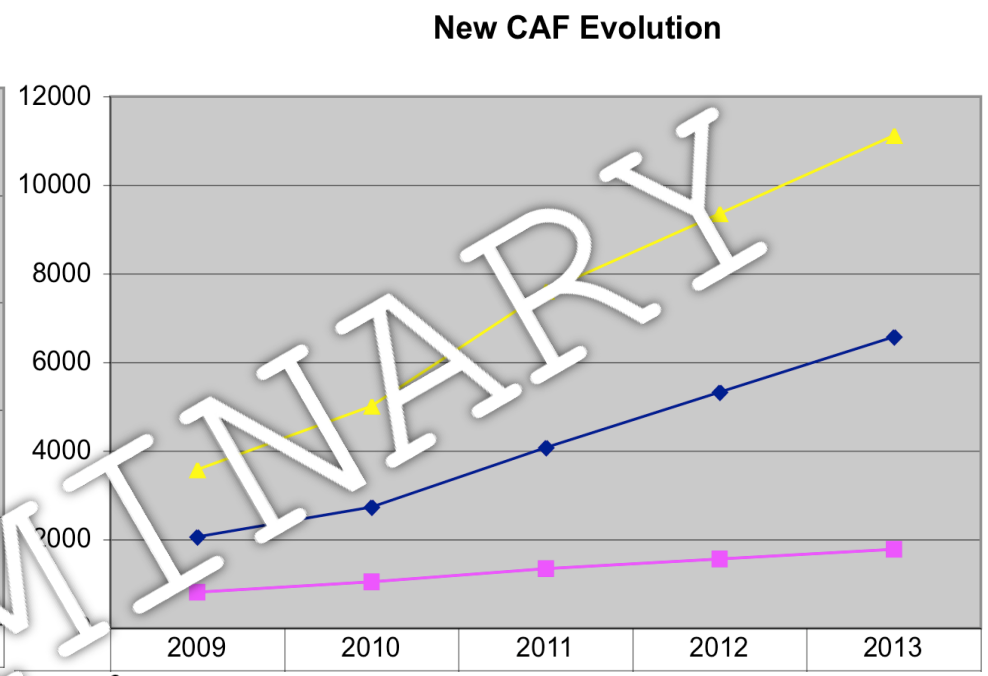
- The DPDs (Derived Physics Data), mentioned in the C-TDR, have been in the meantime defined and implemented in the same data formats as ESD and AOD:
  - Commissioning DPDs are subsets (skims) of ESDs for particular detector studies
  - Physics DPDs are subsets (skim/slim/thin) of AODs for physics studies
  - DPDs are in principle distributed like AODs (a full set per Tier-1 and per Tier-2 cloud)
- We'll have to take operational decisions if the available disk is << than the pledge
  - We may have to delete AOD/DPD datasets from (some) Tier-1s and leave them only in Tier-2s for user analysis
  - Or recognise that not all Tier-1s (and their Tier-2 clouds) are equal and adjust AOD/DPD replicas accordingly
- In the meantime, we have developed tools to collect statistics on dataset usage and we will use it to evaluate how many replicas are needed for different dataset families

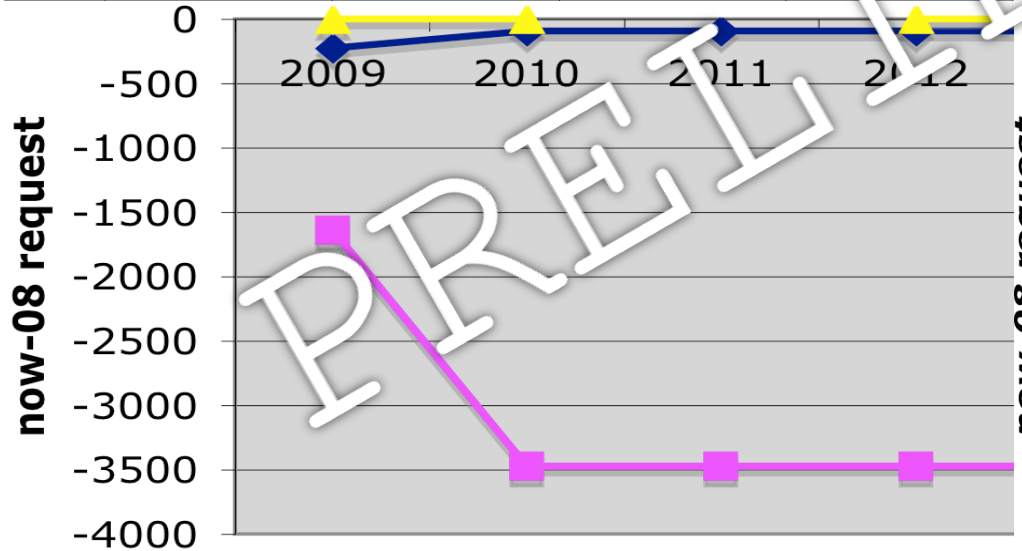
Re-evaluation of resource needs

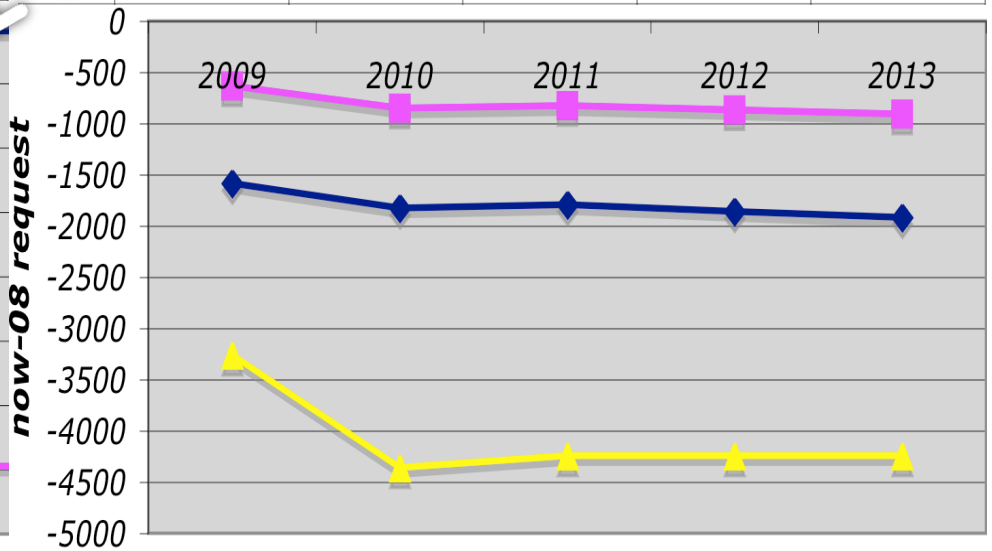
- After hearing last week of the new LHC schedule, we have made a quick calculation based on:
  - The established computing model
  - 3 10^6 seconds of live time in Q4-2009 plus Q1-2010
  - 6 10^6 seconds of live time in Q2-2010 to the end of 2010
- The result has NOT been circulated, discussed and approved by the ATLAS Collaboration yet
  - A revised and formally approved resource request (detailed by quarter instead of year) will be presented to the C-RRB in advance of its April meeting
- We use the "WLCG year" (from April to March) as reference
  - The resources for 2010 are assumed to be available on 1 April 2010
  - The resources for real data in 2009 are mostly needed only on 1 October 2009
  - Part of the 2009 resources (about 1/3) are needed earlier for cosmic ray runs (starting late Spring 2009) and simulation production (continuously)
- Reminder:
  - Some of the Tier-1s have just about reached their 2007 pledges!

Preliminary resource needs at CERN

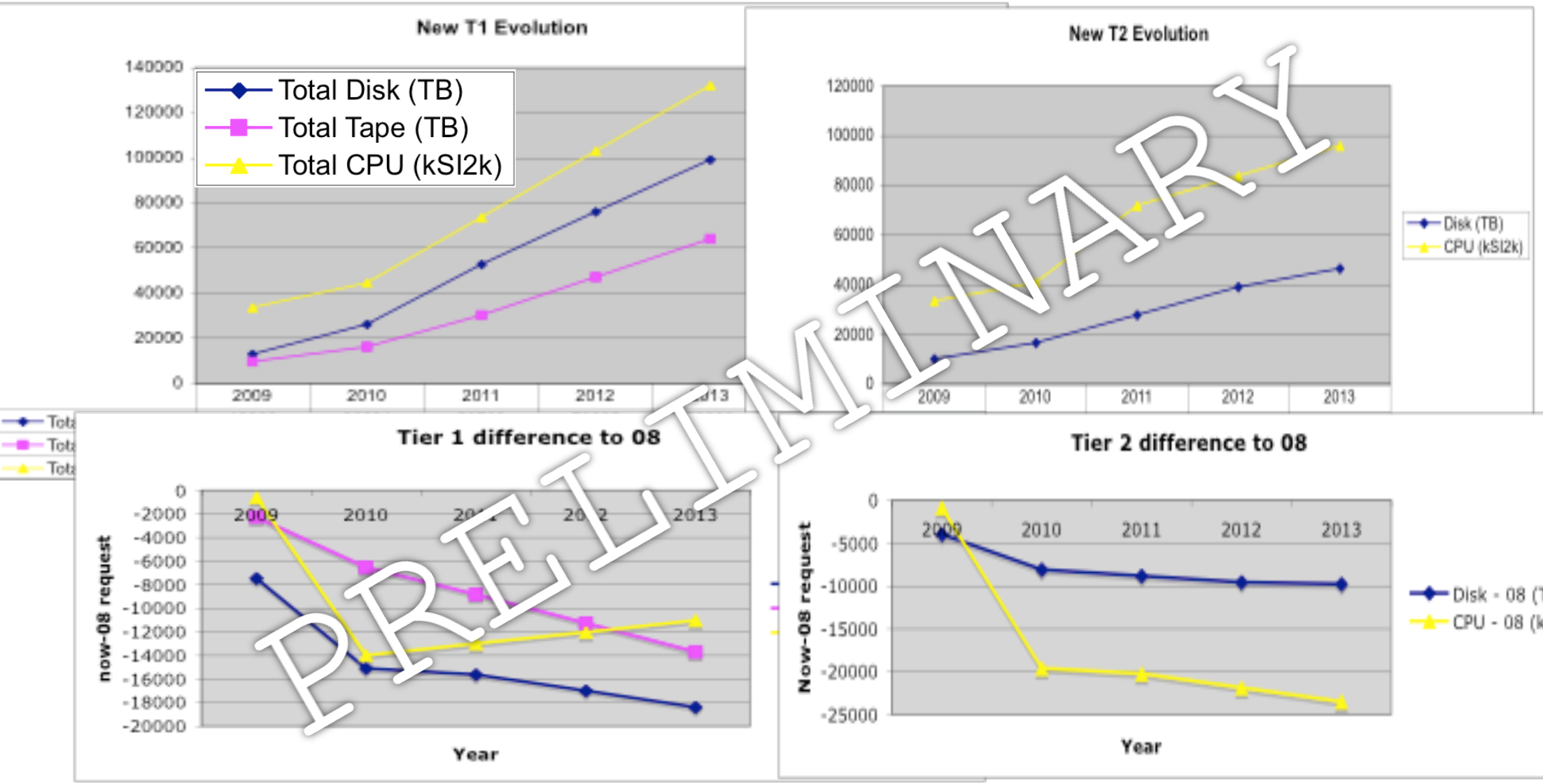








Preliminary resource needs at Tier-1/2s



Work plan for 2009

- In progress: tests of pre-staging from tape at Tier-1s
  - Needed for bulk reprocessing of RAW data in future operations
- End February: software release 15.0.0
  - Baseline release for 2009 operations
    - Several performance improvements, new ROOT etc.
- Mid-March: 2nd reprocessing campaign
- Now to end April: further tests of database access methods for different task types
- Continuously: HammerCloud tests of data access for distributed analysis
- May onwards: start of cosmics data-taking at increasing rates
- October: ready for LHC data