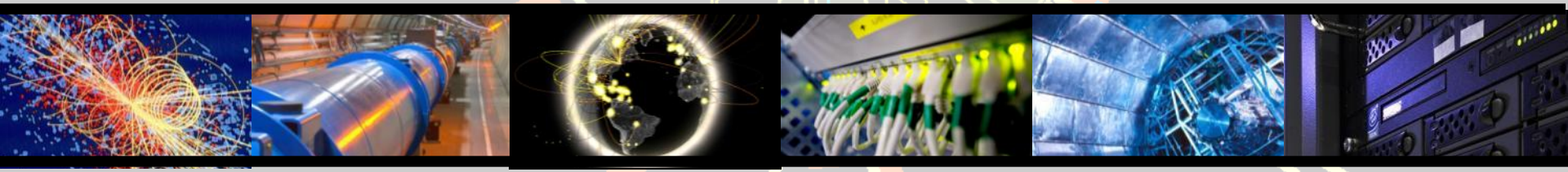
Progress in Computing

Ian Bird

ICHEP 2010
28th July 2010, Paris



Outline

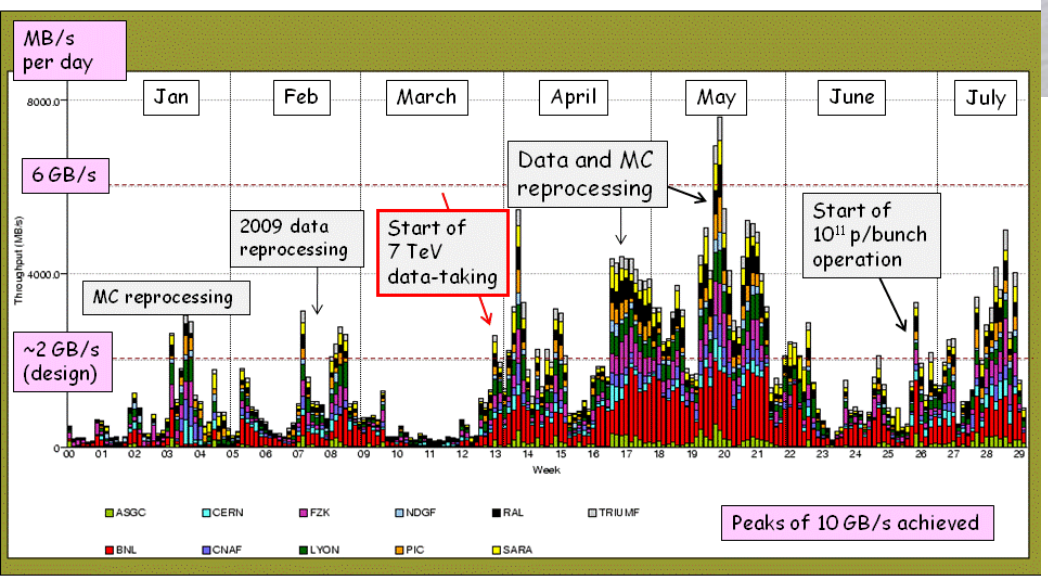
- Progress in Computing (for LHC)
  - Where we are we today & how we got here.
  - Achievements in computing of the experiments
  - Some representative statistics and plots from parallel sessions
- … and the outlook?
  - Grids → clouds? Sustainability?
- Thanks to:
  - Contributors to Track on "Advances in Instrumentation and Computing for HEP" – various slides used to illustrate points

Progress …

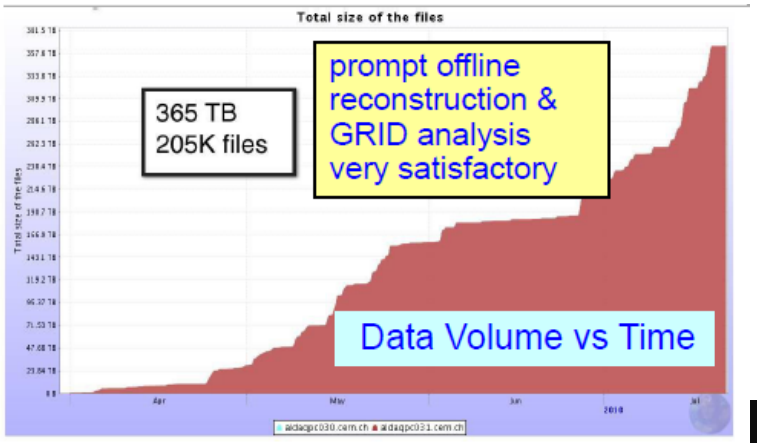
- Overall is clear – physics output in very short time
- Huge effort: Combination of experiment sw & computing and grid infrastructures
- And a lot of testing !

Worldwide data distribution and analysis

Total throughput of ATLAS data through the Grid: from 1st January until yesterday

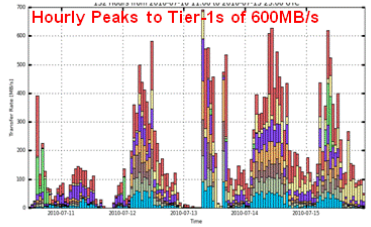


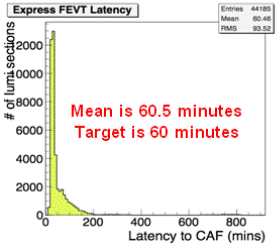
GRID-based analysis in June-July 2010:
>1000 different users, ~ 11 million analysis jobs processed

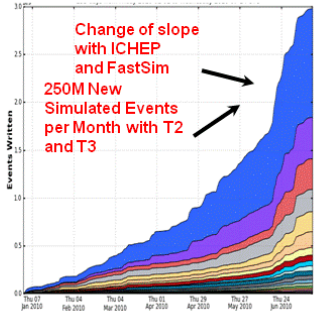


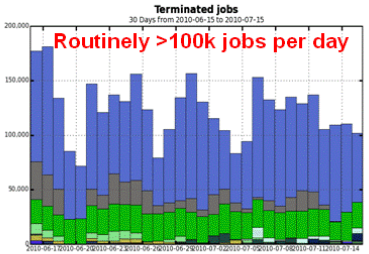
Data Processing, Transfer and Analysis Activities

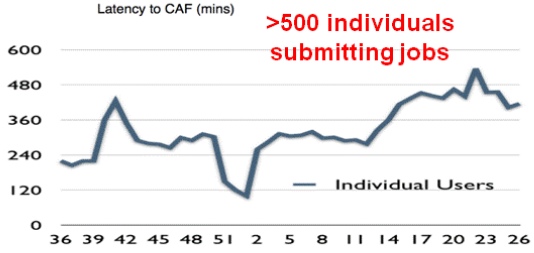
Excellent experience so far: the whole offline and computing organization + GRID infrastructure performing very well.



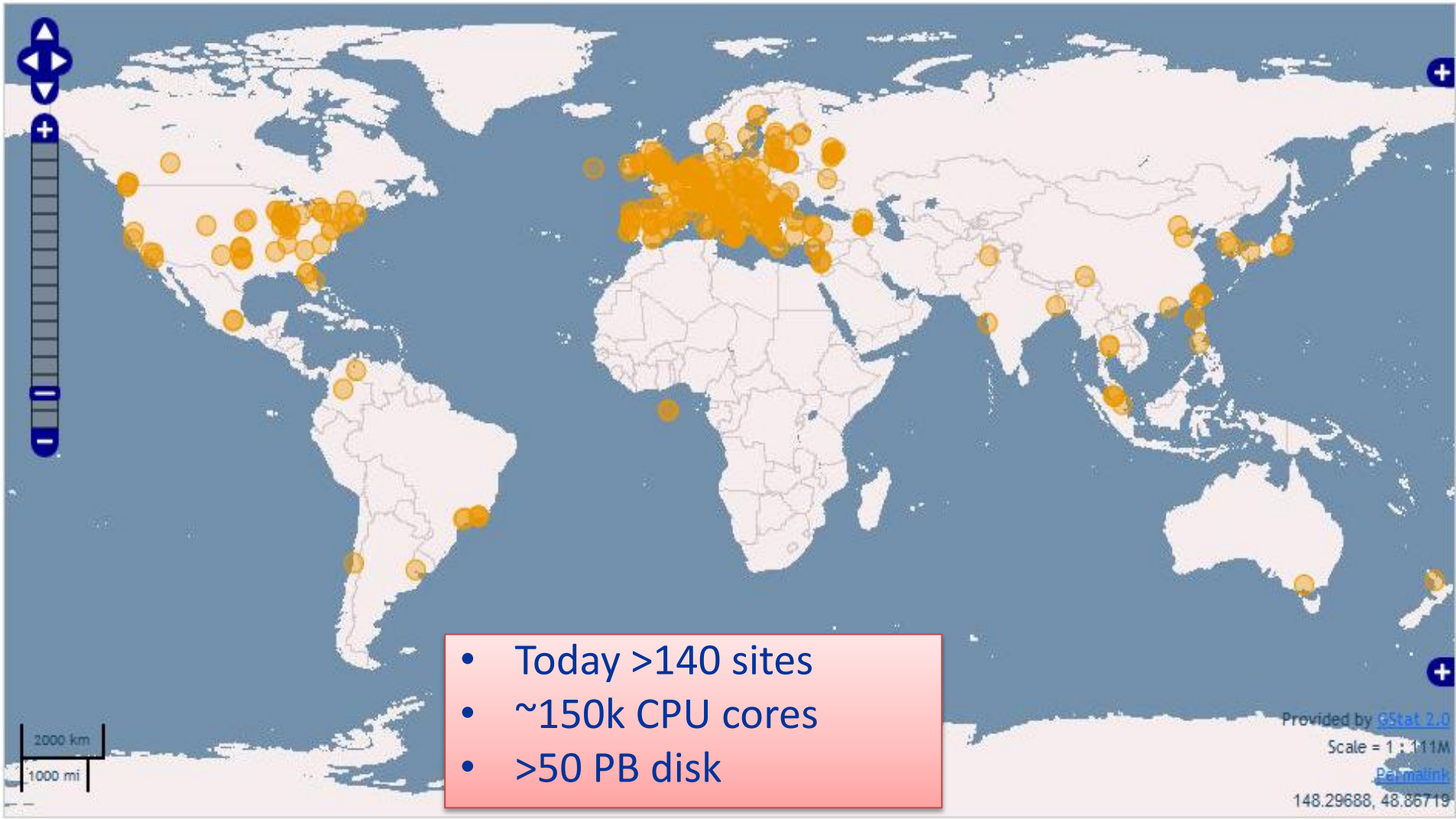








Worldwide resources



- Today >140 sites
- ~150k CPU cores
- >50 PB disk

WLCG Collaboration Status
Tier 0; 11 Tier 1s; 64 Tier 2 federations

Today we have 49 MoU signatories, representing 34 countries:

Australia, Austria, Belgium, Brazil, Canada, China, Czech Rep, Denmark, Estonia, Finland, France, Germany, Hungary, Italy, India, Israel, Japan, Rep. Korea, Netherlands, Norway, Pakistan, Poland, Portugal, Romania, Russia, Slovenia, Spain, Sweden, Switzerland, Taipei, Turkey, UK, Ukraine, USA.

From testing to data:

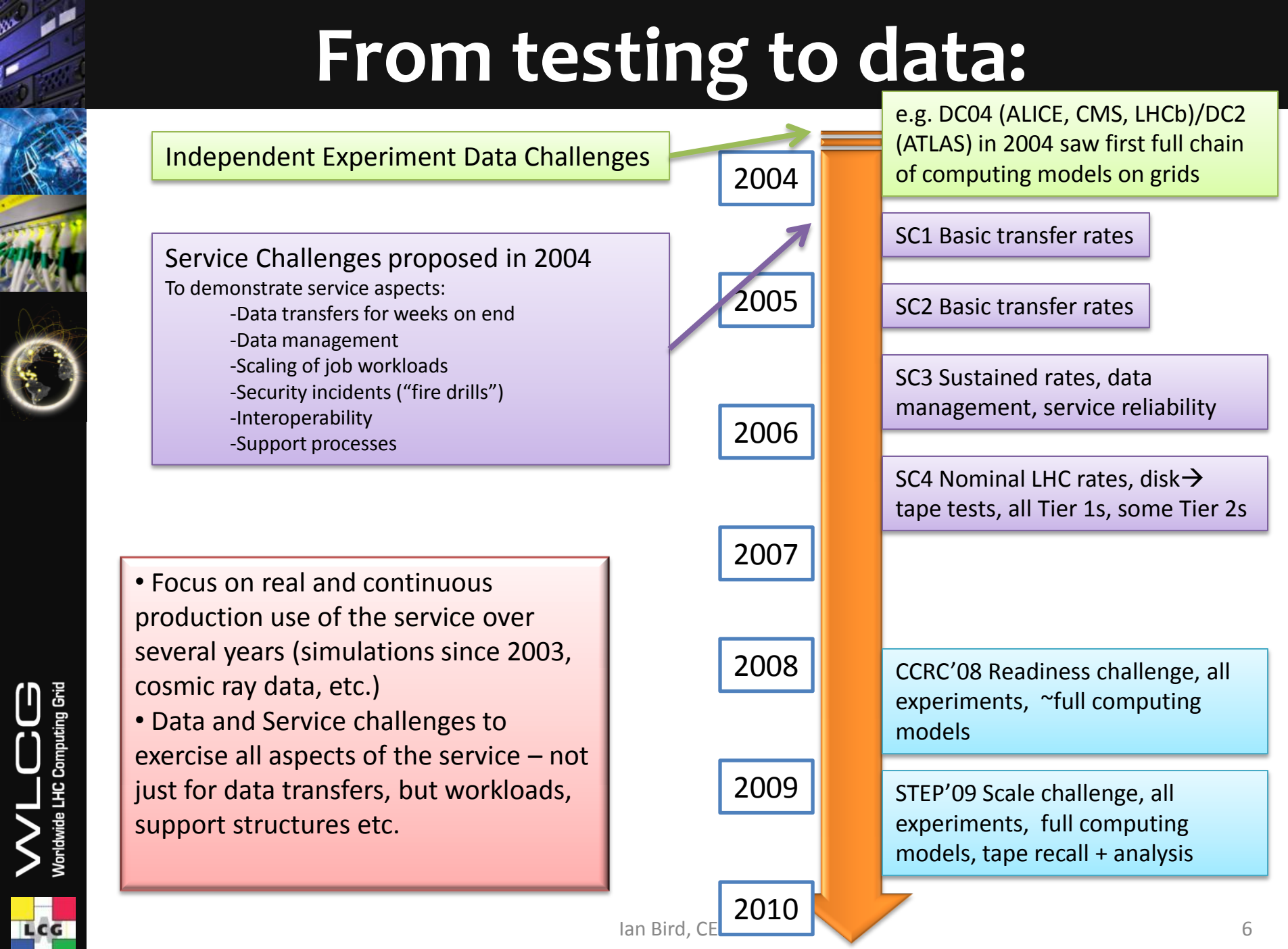
Independent Experiment Data Challenges

Service Challenges proposed in 2004

To demonstrate service aspects:

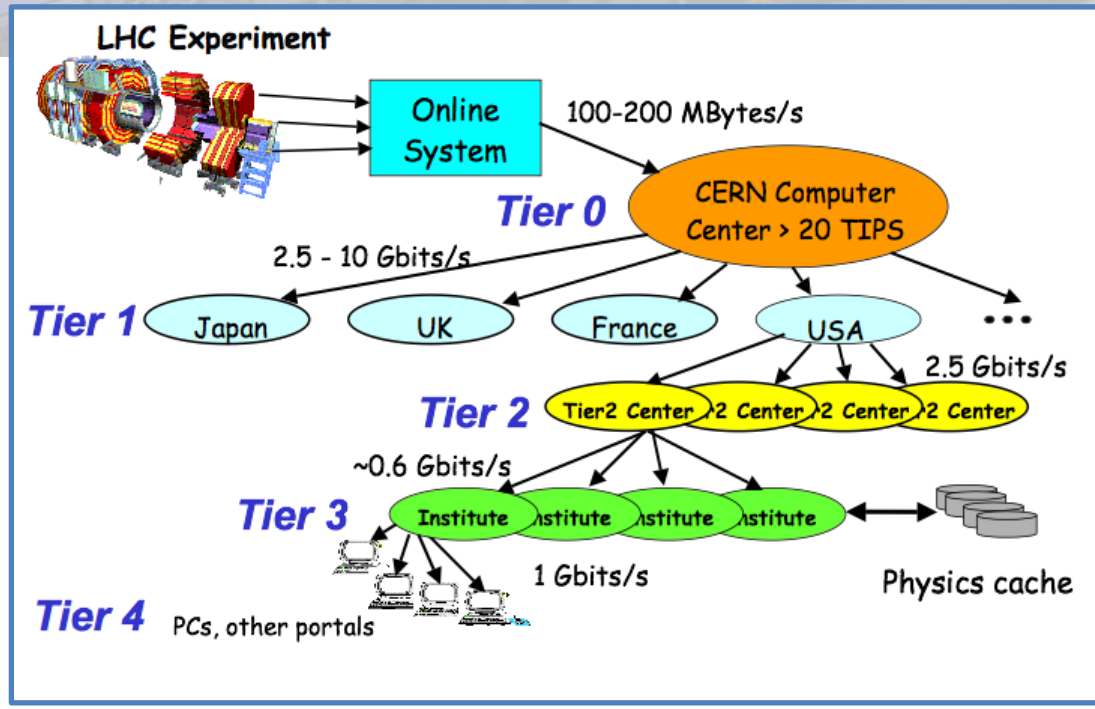
- -Data transfers for weeks on end
- -Data management
- -Scaling of job workloads
- -Security incidents ("fire drills")
- -Interoperability
- -Support processes

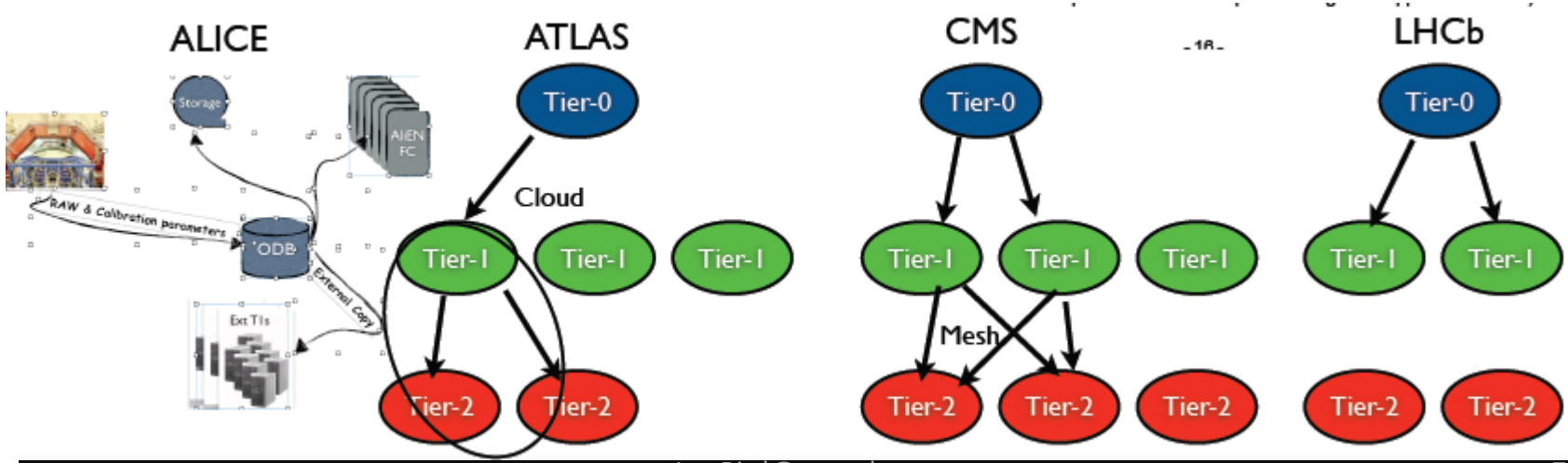
- Focus on real and continuous production use of the service over several years (simulations since 2003, cosmic ray data, etc.)
- Data and Service challenges to exercise all aspects of the service – not just for data transfers, but workloads, support structures etc.



Experiment models have evolved

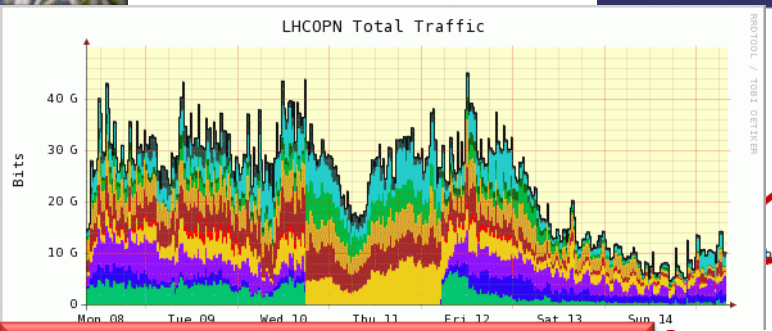
- Models all ~based on the MONARC tiered model of 10 years ago
- Several significant variations, however



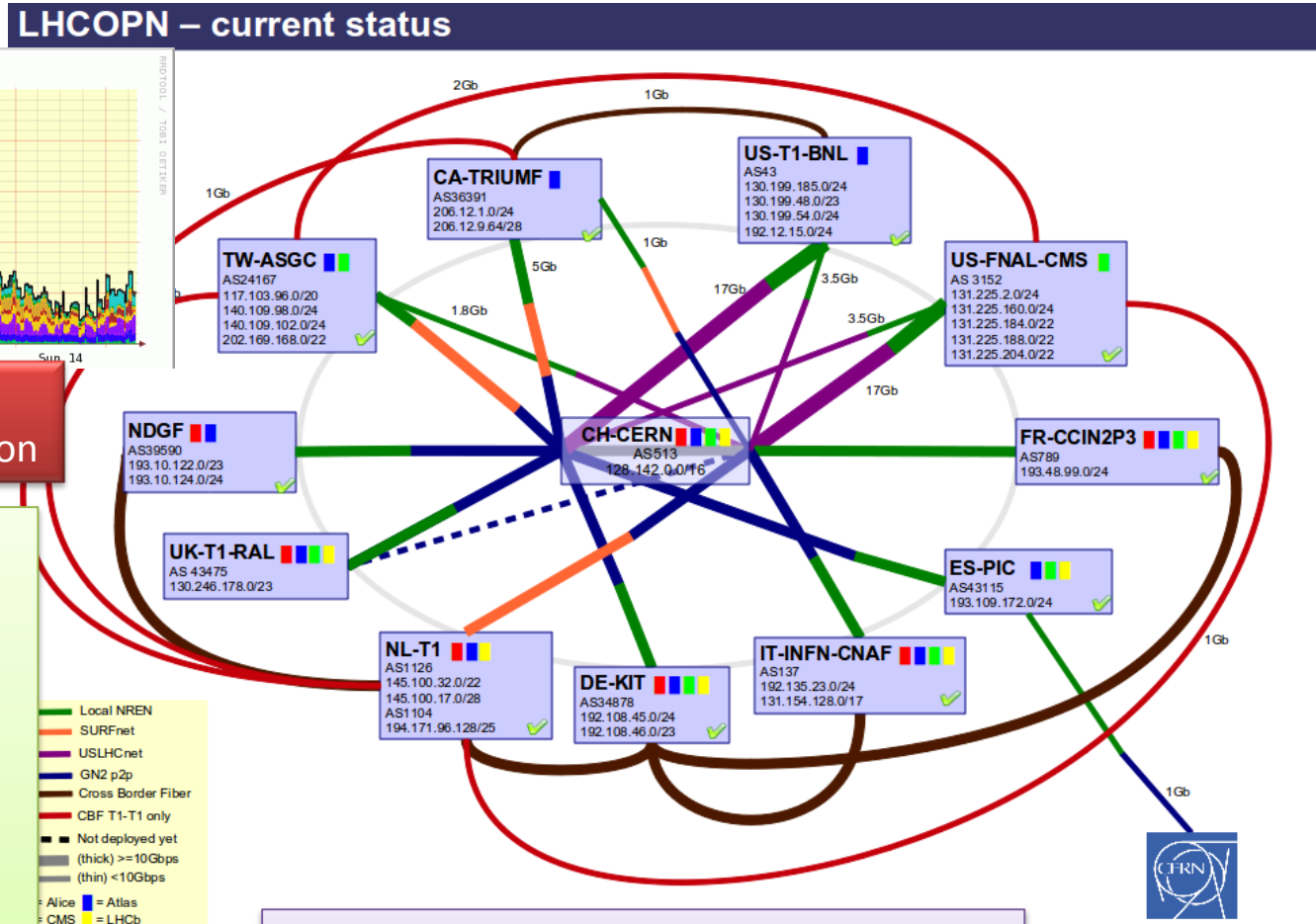


Data transfer

- Data transfer capability today able to manage much higher bandwidths than expected/feared/planned



Fibre cut during STEP'09:
Redundancy meant no interruption



Data transfer:

•SW: gridftp, FTS (interacts with endpoints, recovery), experiment layer

•HW: light paths, routing, coupling to storage

•Operational: monitoring

+ the academic/research networks for Tier1/2!

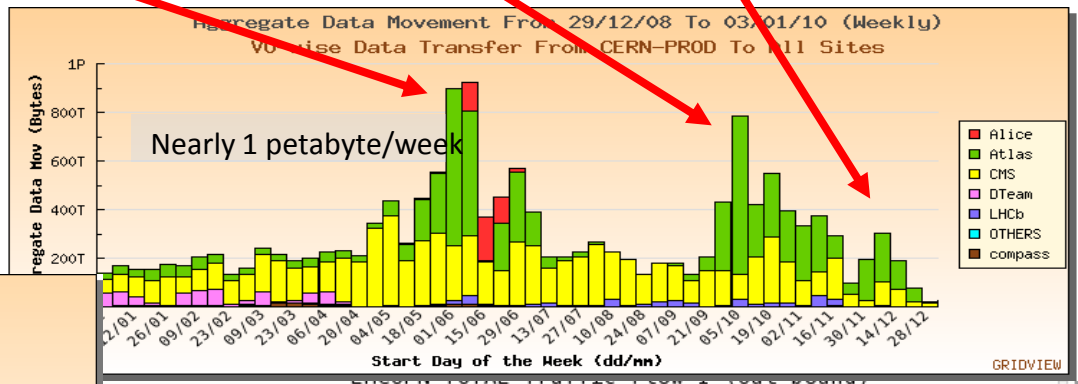
In terms of data transfers …

Final readiness test (STEP'09)

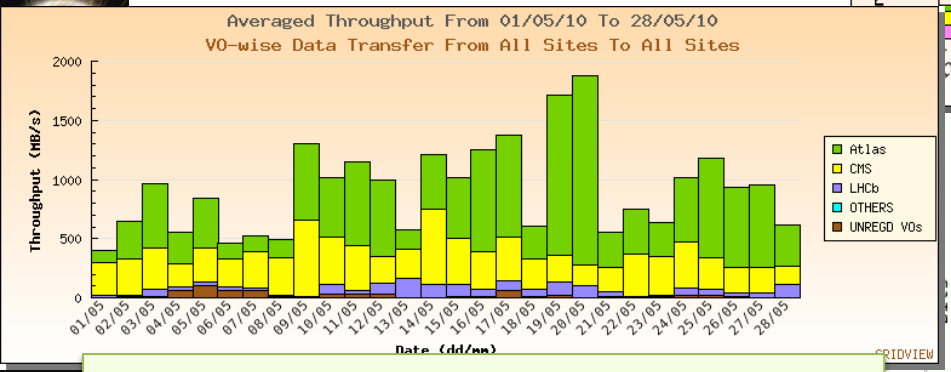
Preparation for LHC startup

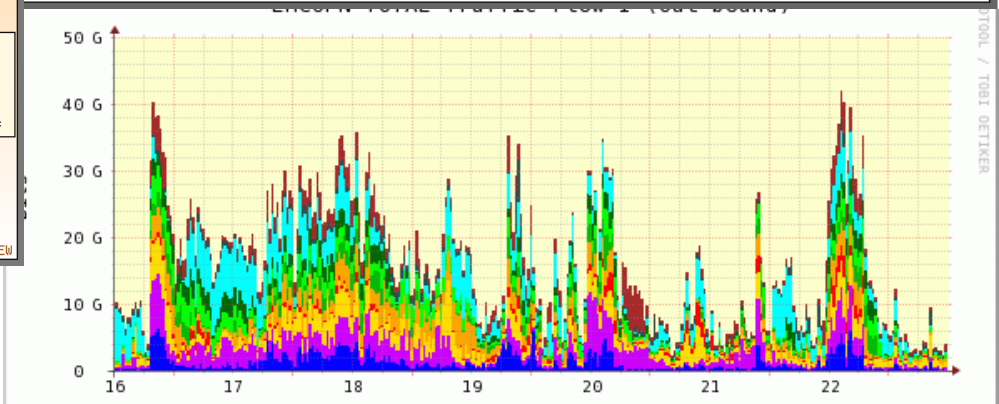
LHC physics data

2009: STEP09 + preparation for data



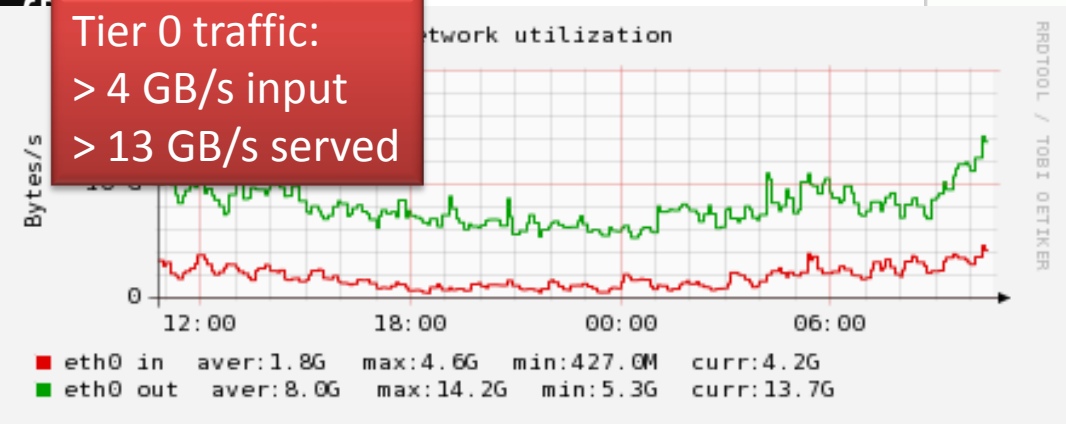
Nearly 1 petabyte/week

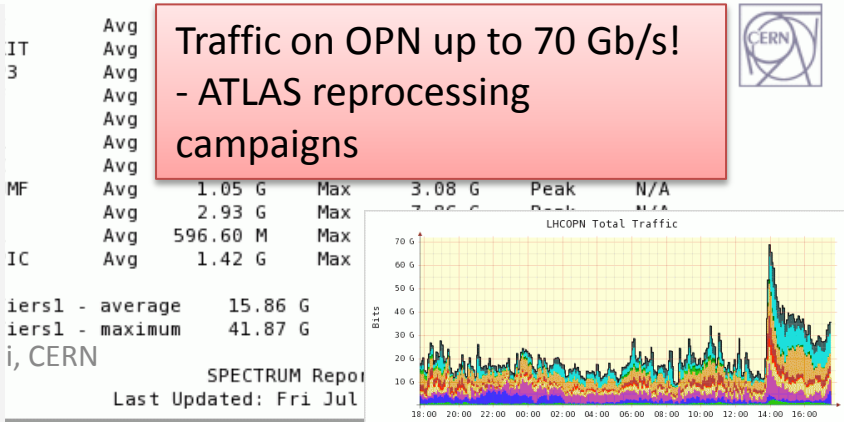




Data export during data taking:
- According to expectations on average

Tier 0 traffic:
> 4 GB/s input
> 13 GB/s served

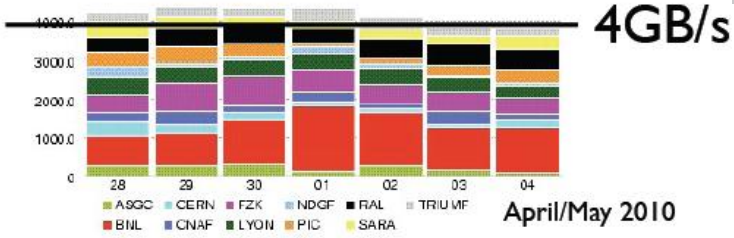




Traffic on OPN up to 70 Gb/s!
- ATLAS reprocessing campaigns

Data distribution

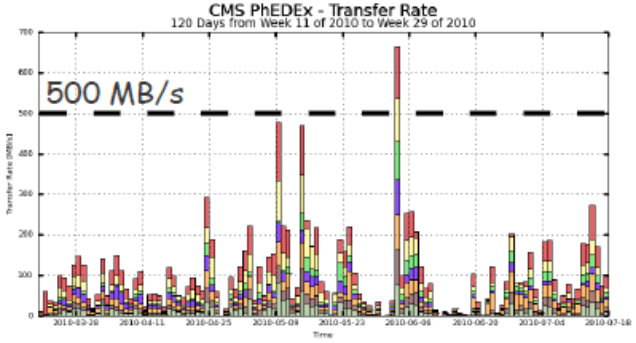
- In concert with data reprocessing we reprocess MC to assure consistency
- This leads to large volumes of data which need to be distributed after reprocessing campaigns
  - This takes a long time!
- Can lead to delays in 'interesting' data arriving



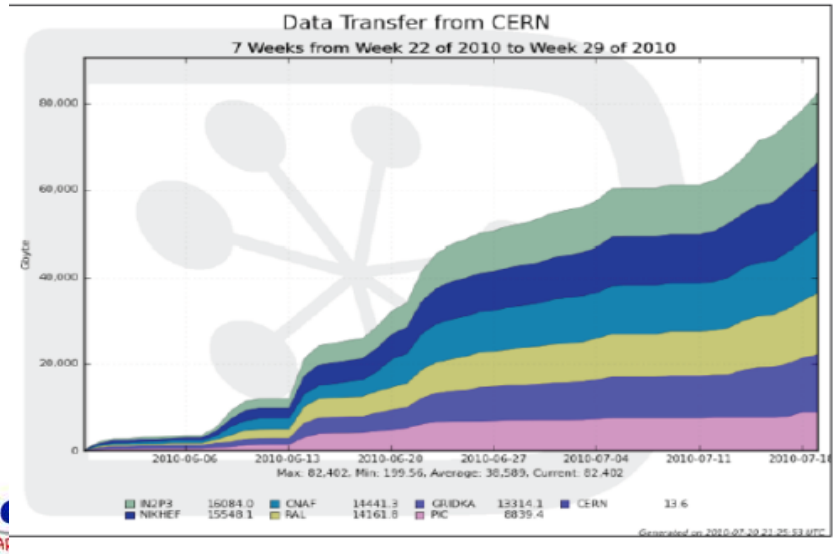
Disk Usage Ramp up on T1s

ATLAS: Total throughput
T0-T1; T1-T1; T1-T2
G. Stewart, 1225

- Resources provisioned for steady data stream from Tier-0 to Tier-1's
- Current reality looks different
- Total volume of 1 PB since April
- Very good transfer quality



CMS: T0 – T1
M. Klute, 1223



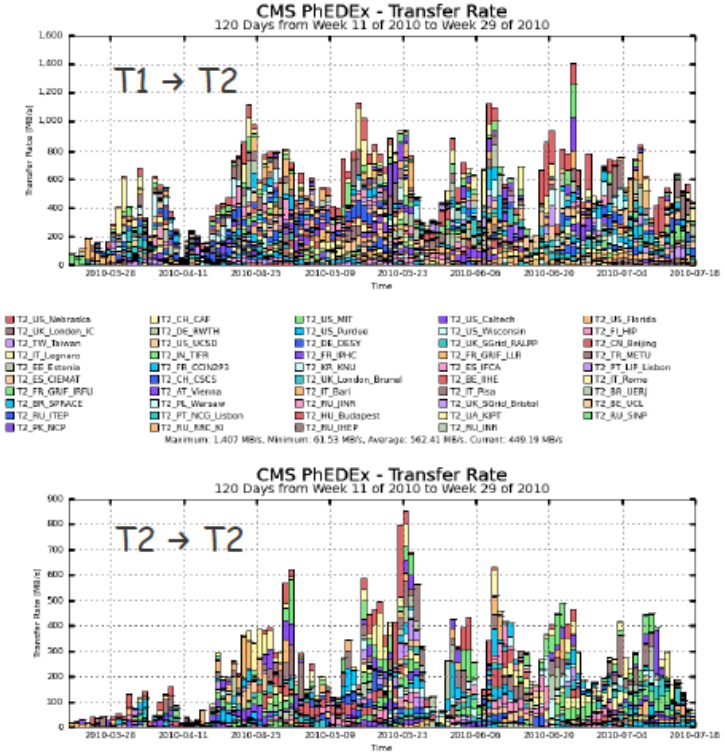
- RAW Data is replicated to one of the Tier-1
- Albeit some initial problem, data is now successfully transferred on regular basis.

LHCb: T0 – T1
M. Adinolfi, 1221

Data distribution for analysis

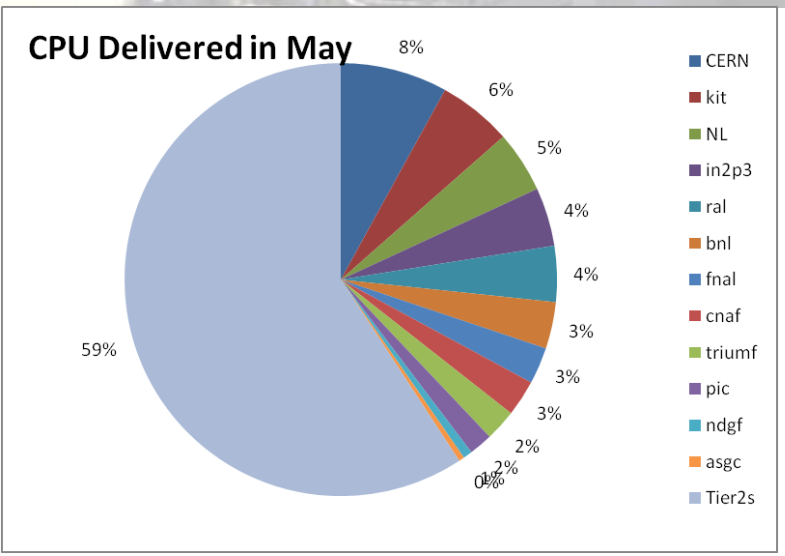
Data Distribution for Analysis

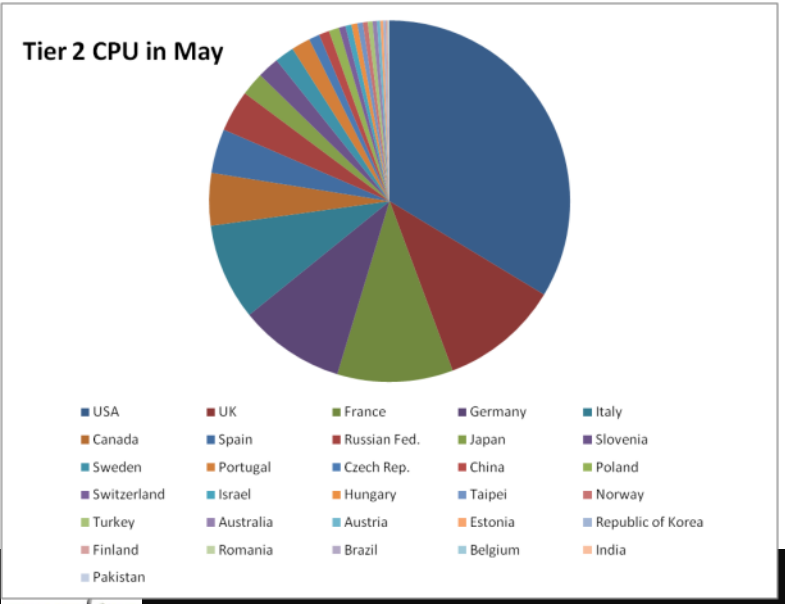
- Data transferred from Tier-1's
  - 49 Tier-2 sites received data
  - > 5 PB transferred in last 120 days
  - average rate 562 MB/s
  - max rate 1407 MB/s
- Data transferred between Tier-2's
  - 41 Tier-2 sites received data
  - > 2.5 PB transferred in last 120 days
  - average rate 254 MB/s
  - max rate 853 MB /s
  - full mesh approach
  - Data distribution re-balances itself
  - Datasets produced at Tier-2's can be distributed to others

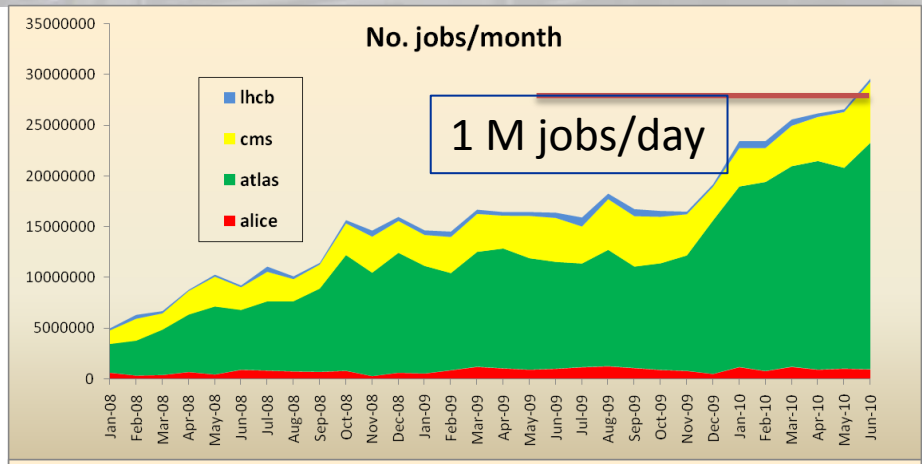


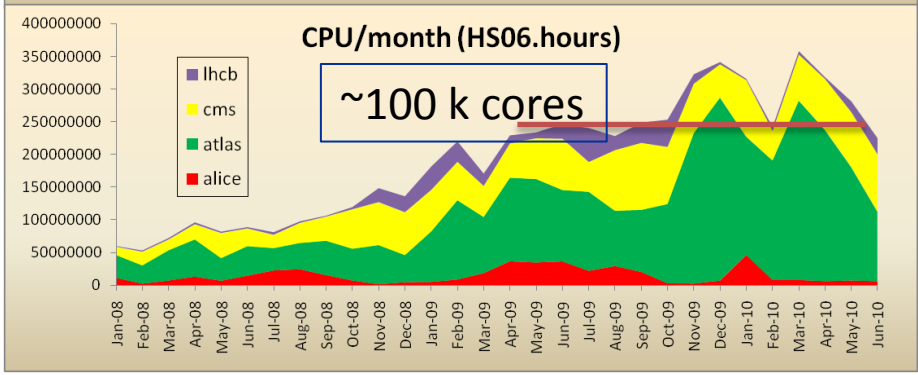
- For all experiments: early data has been available for analysis within hours of data taking

Use of CPU …



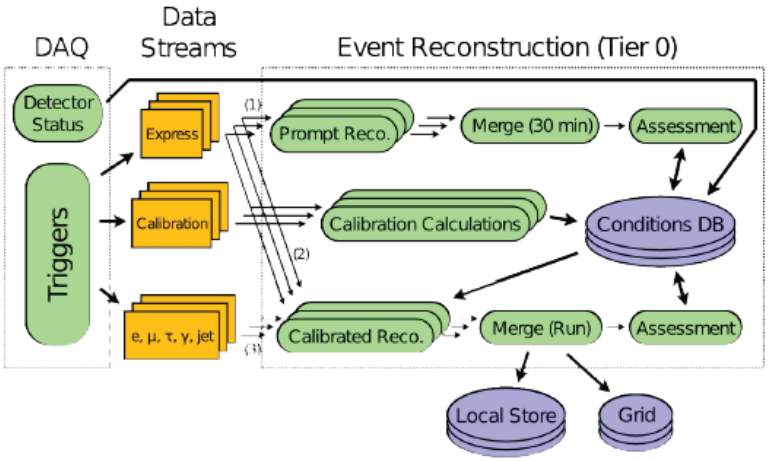






- Peaks of 1M jobs/day now
- Use ~100k cores equivalent
- Tier 2s heavily used wrt Tier 1s

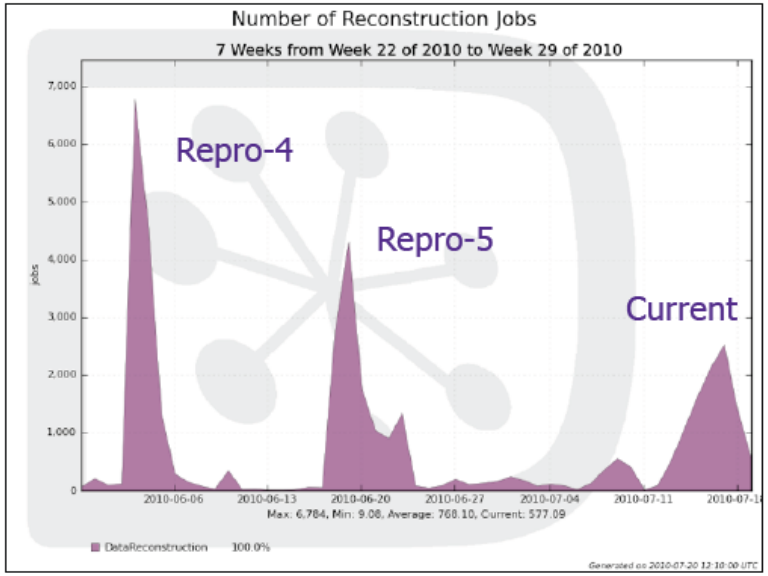
Processing & re-processing



ATLAS processing: P. Onyisi, 1197

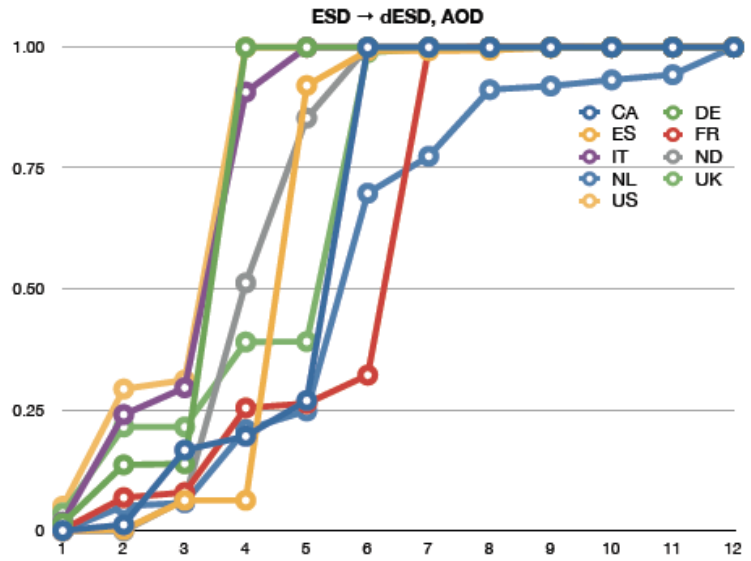
- First data has been reprocessed many times
- Reaching the point where this is slowing down now

Full cycle (first reconstruction pass, calibration, second reconstruction pass, data quality assessment) generally complete in 3–4 days



- Data collected up to early June (~14nb-1) processed several times as new alignment and improved reconstruction are made available.
- 90% of the datasets is reprocessed in abut 3 days.
- Now that nominal conditions have been reached such frequent reprocessing are no longer possible.

Sometimes takes operational effort …



- With ten Tier-1s involved there's lots of scope for problems
- Operationally heavy
  - But sites do respond
- ATLAS Distributed Computing team successful in achieving 100% of events processed in April and May

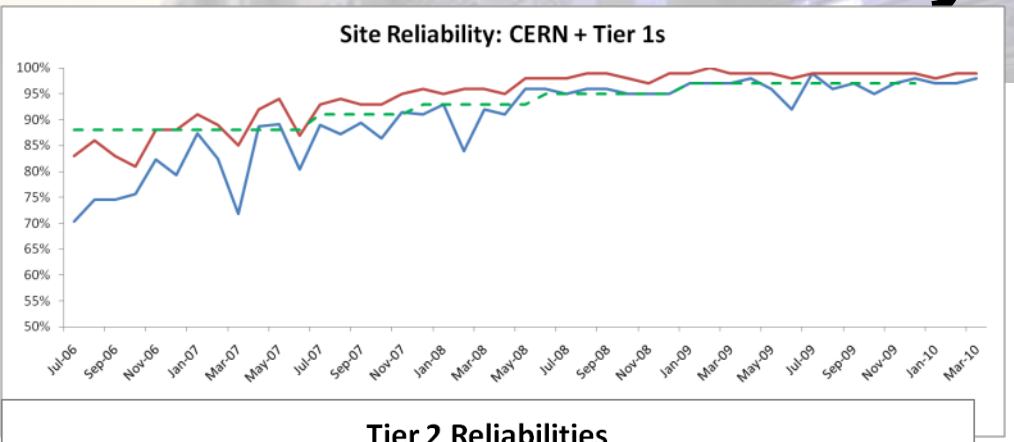
total jobs	9577	9540	7233	13375	1964	6886	26676	19252	25197	119700
total done	9577	9540	7233	13375	1964	6886	26676	19252	25197	119700
% %	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

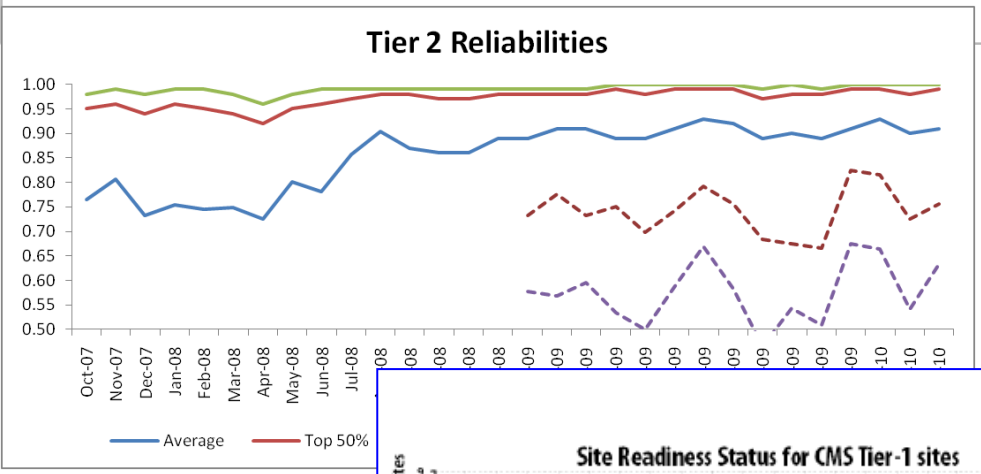
ICHEP 2010, Paris

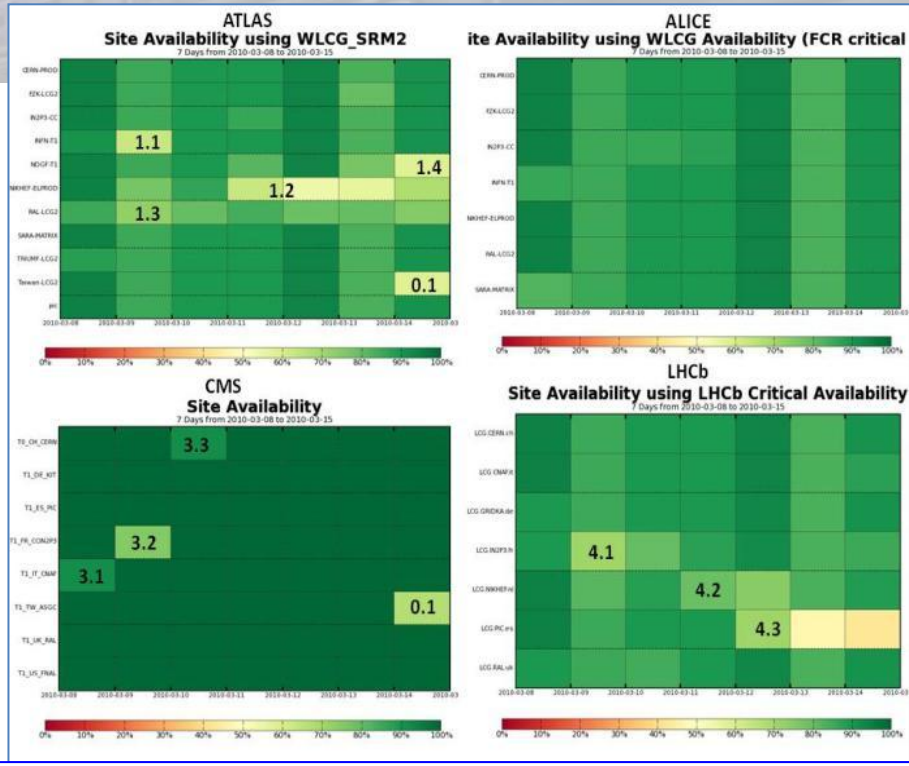
20

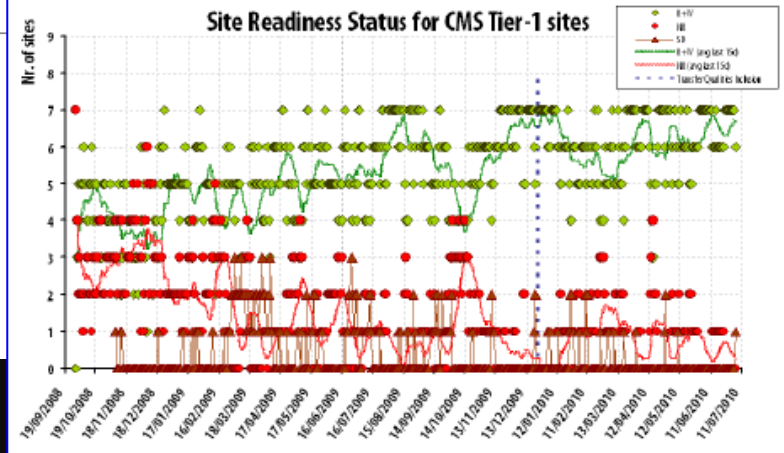
- Daily operations meetings – all experiments – many sites – address exactly these kind of problems
- Still a significant level of manual intervention and coordination required
- But this is now at a level that is sustainable – in CCRC'08 and even STEP09 this was not clear

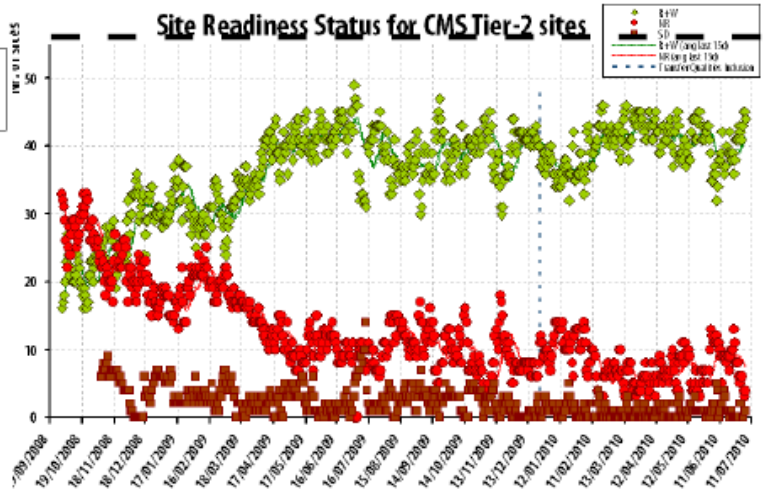
Site availability and readiness





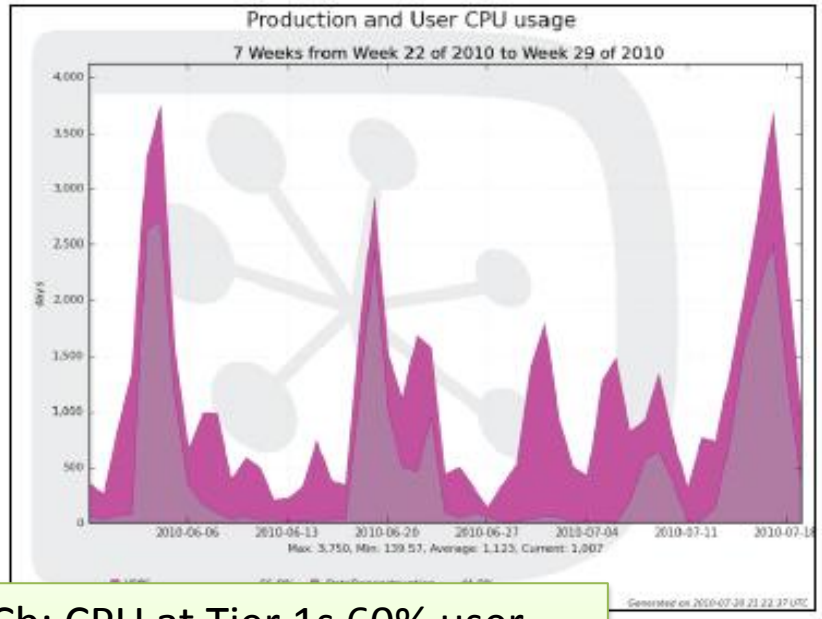


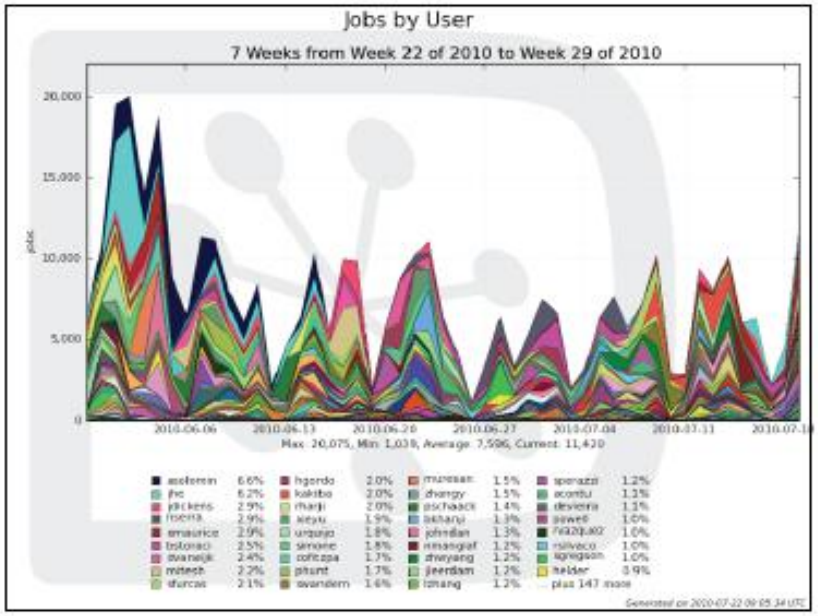




ICHEP Paris 2010

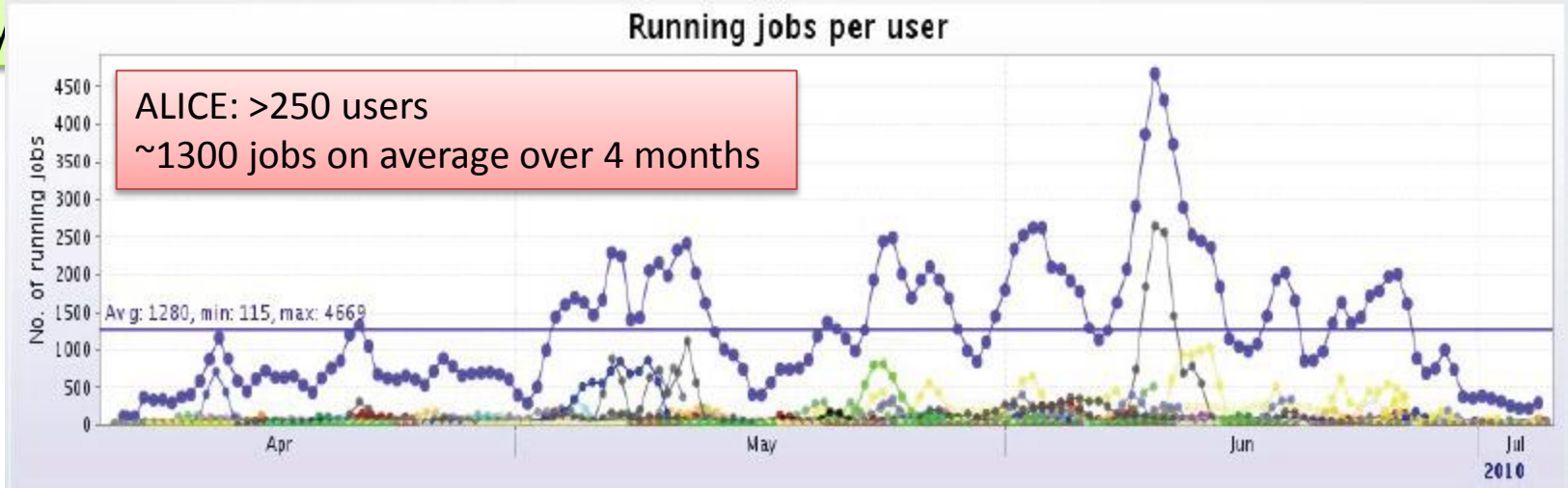
Analysis & users





LHCb: CPU at Tier 1s 60% user and 40% reconstruction;

- 200 users
- 30k jobs/day

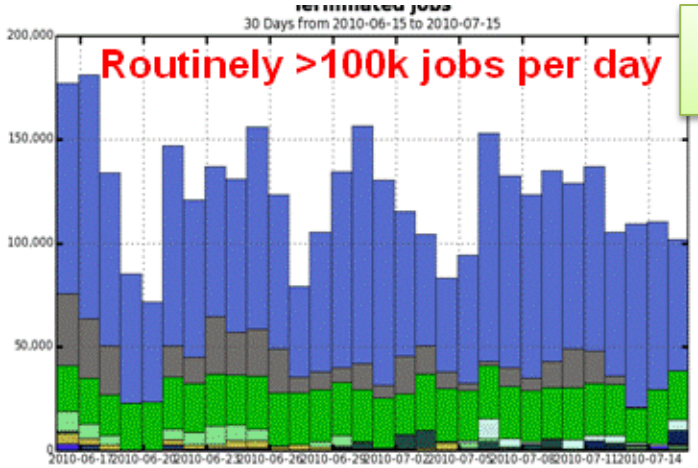


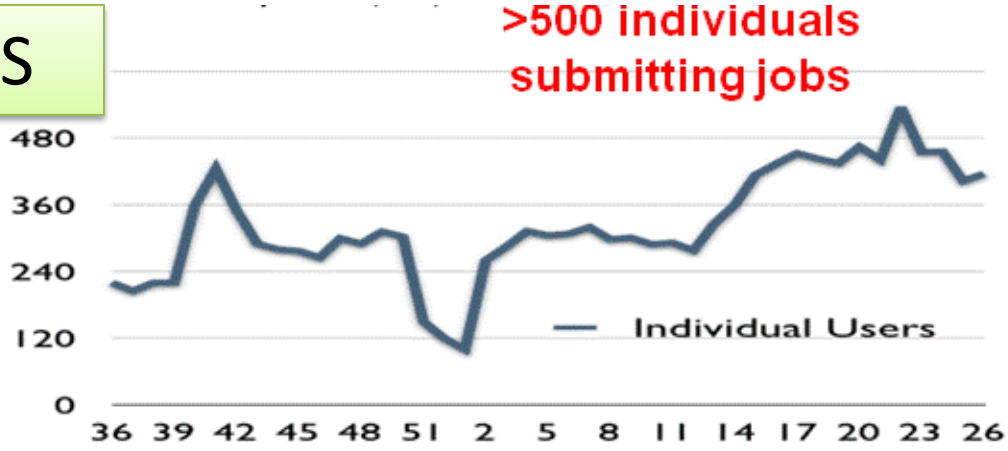
ALICE: >250 users
~1300 jobs on average over 4 months

ICHEP PARIS 2010

Analysis & users – 2

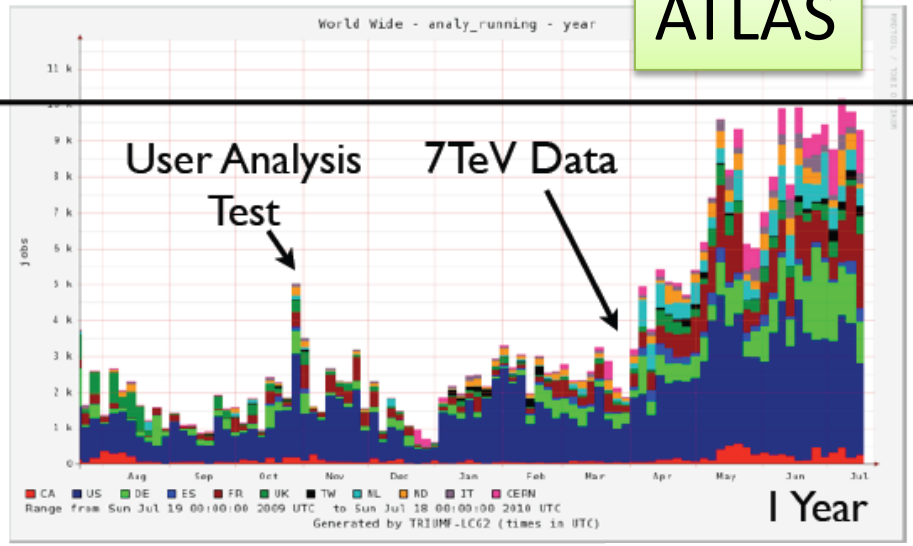
CMS





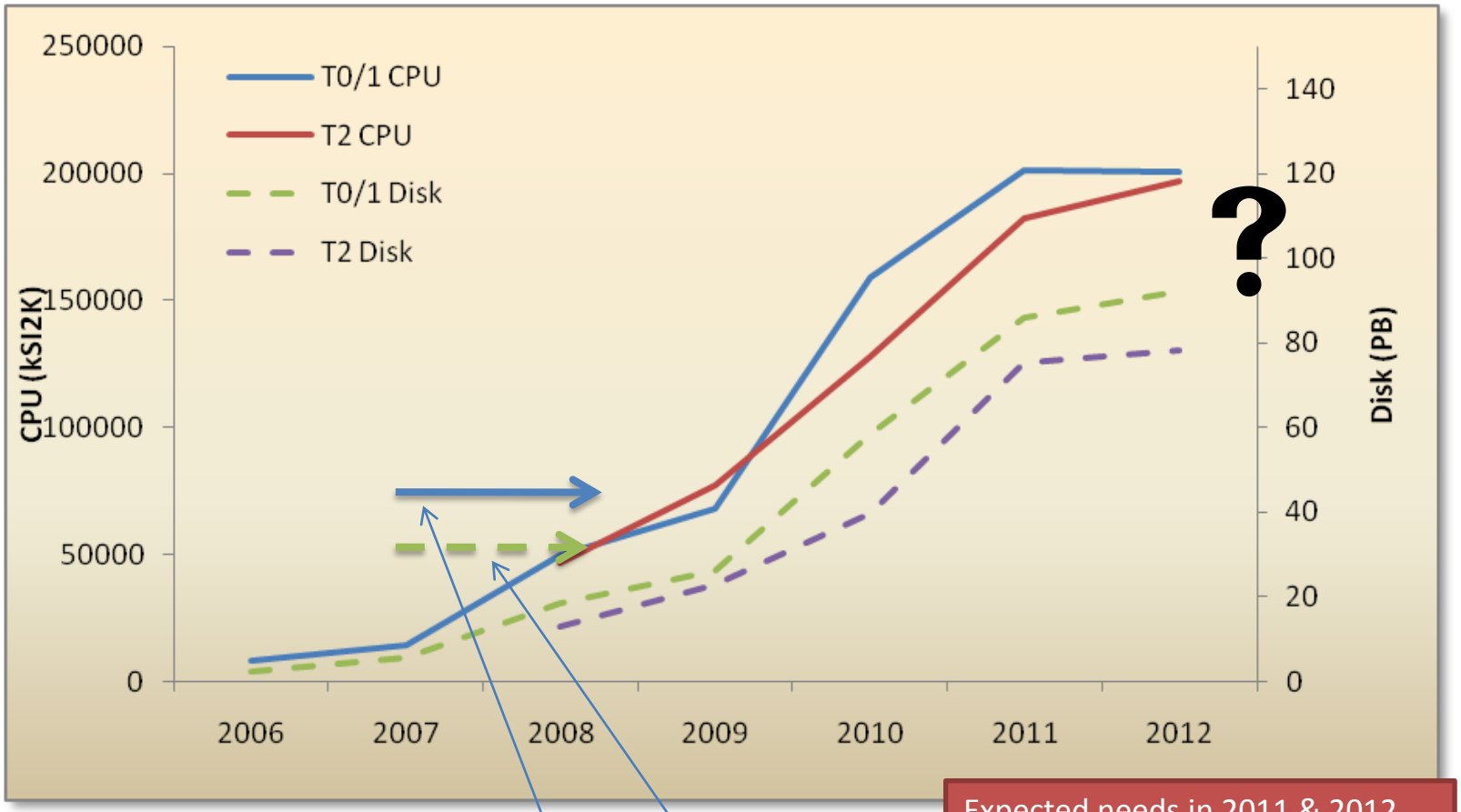
ATLAS

10k running jobs



GRID-based analysis in June-July 2010:
>1000 different users, ~ 11 million analysis jobs processed

Resource Evolution



Expected needs in 2011 & 2012

Need foreseen @ TDR for T0+1 CPU and Disk for 1st nominal year

NB. In 2005 only 10% of 2008 requirement was available. The ramp-up has been enormous!

Prospects for next few years

- We have an infrastructure demonstrated to be able to support LHC data processing and analysis
- Significant science grid (e-science/ cyberscience) infrastructures spun off and used to provide support
  - These are now evolving: EGI in Europe, OSG phase 2, etc
- This is not just software – there are significant operational infrastructures behind it
  - World wide trust → single authentication/authorization
  - Coordinated security policies and operational response
  - Operational processes, monitoring, alarms, reporting, ...
- Must be able to evolve the technical implementation (i.e. Grid middleware) without breaking the overall infrastructure

Evolution and sustainability

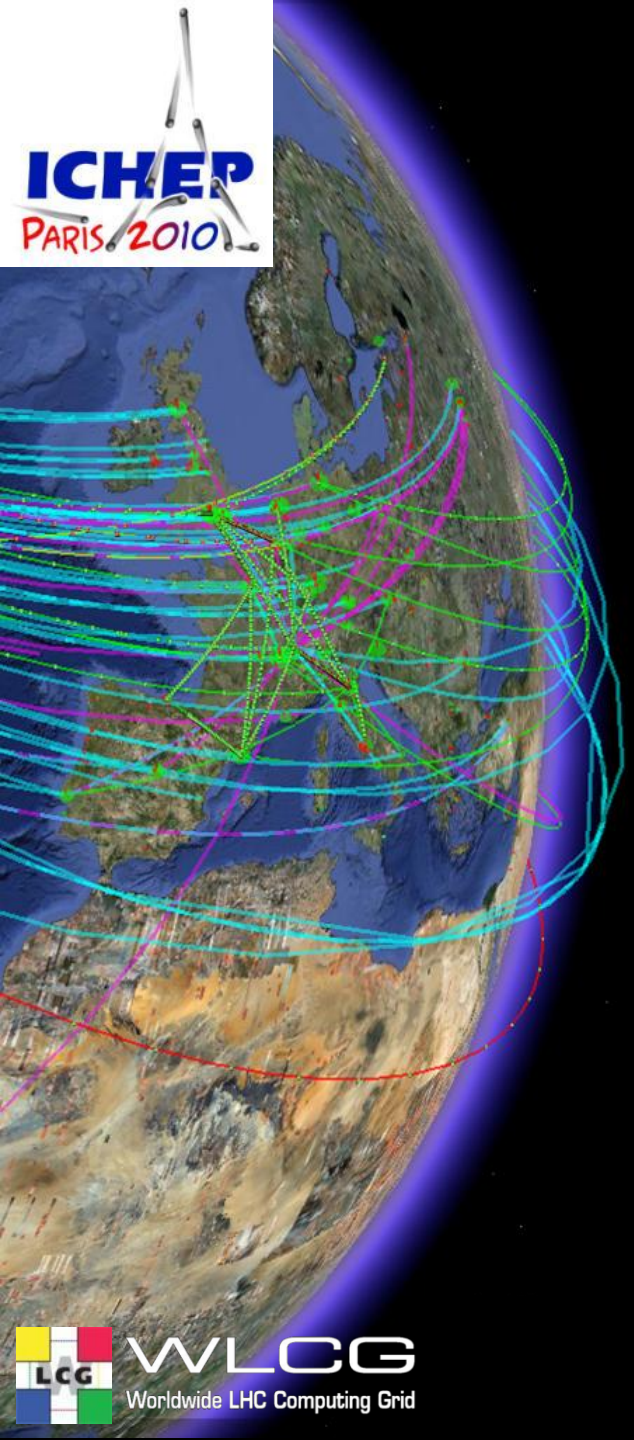
- Need to adapt to changing technologies
  - Major re-think of storage and data access
  - Use of many-core CPUs (and other processor types?)
  - Virtualisation as a solution for job management
    - Brings us in line with industrial technology
    - Integration with public and commercial clouds
- Network infrastructure
  - This is the most reliable service we have
  - Invest in networks and make full use of the distributed system
- Grid Middleware
  - Complexity of today's middleware compared to the actual use cases
  - Evolve by using more "standard" technologies: e.g. Message Brokers, Monitoring systems are first steps
- But: retain the WLCG infrastructure
  - Global collaboration, service management, operational procedures, support processes, etc.
  - Security infrastructure – this is a significant achievement
    - both the global A&A service and trust network (X509) and
    - the operational security & policy frameworks

Evolution of Data Management

- 1st workshop held in June
  - Recognition that network as a very reliable resource can optimize the use of the storage and CPU resources
    - The strict hierarchical MONARC model is no longer necessary
  - Simplification of use of tape and the interfaces
  - Use disk resources more as a cache
  - Recognize that not all data has to be local at a site for a job to run – allow remote access (or fetch to a local cache)
    - Often faster to fetch a file from a remote site than from local tape
- Data management software will evolve
  - A number of short term prototypes have been proposed
  - Simplify the interfaces where possible; hide details from end-users
- Experiment models will evolve
  - To accept that information in a distributed system cannot be fully up-to-date; use remote access to data and caching mechanisms to improve overall robustness
- Timescale: 2013 LHC run

Some observations

- Experiments have truly distributed models
- Needs a lot of support and interactions with sites – heavy but supportable
- Network traffic far in excess of what was anticipated, but it is supportable at the moment
  - Must plan for the future
- Limited amount of data has allowed many reprocessings
  - LHCb already on their nominal model ...
- Today resources are plentiful, and not yet full. This will surely change ...
- Significant numbers of people successfully doing analysis



Conclusions

- Distributed computing for LHC is a reality and enables physics output in a very short time
- Experience with real data and real users suggests areas for improvement –
  - The infrastructure of WLCG can support evolution of the technology