LHCONE: il punto di vista di ATLAS/CMS

Daniele Bonacorsi

[deputy CMS Computing coordinator – University of Bologna, Italy]

Workshop CCR INFN GRID 2011

16-20 May 2011 Hotel Hermitage - Isola d'Elba

DISCLAIMER:

Most discussions/activities in this context involve only ATLAS/CMS - so far

WLCG Tiers

Computing Model(s)

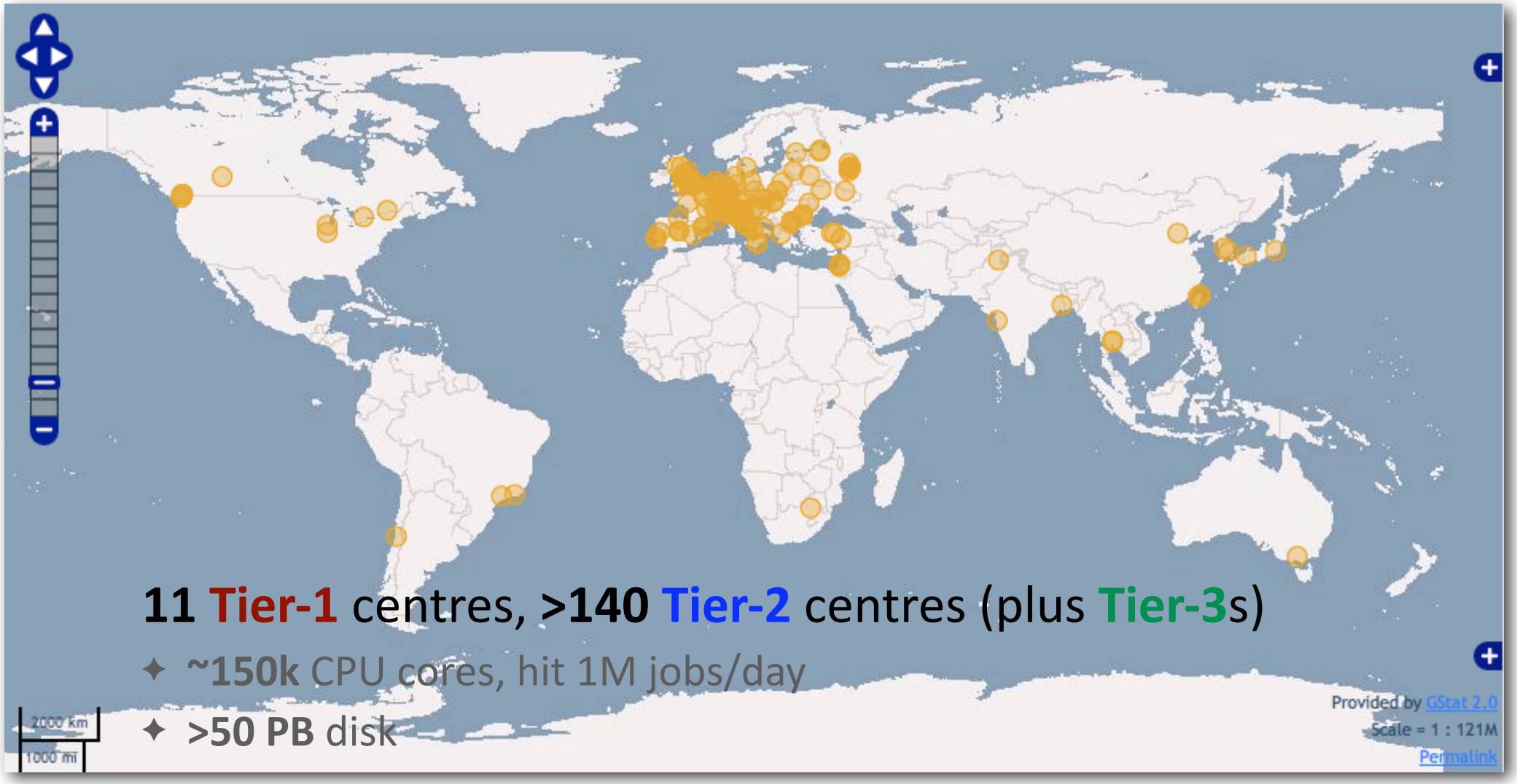
Operations (focus on data transfer and access)

Motivations for LHCONE

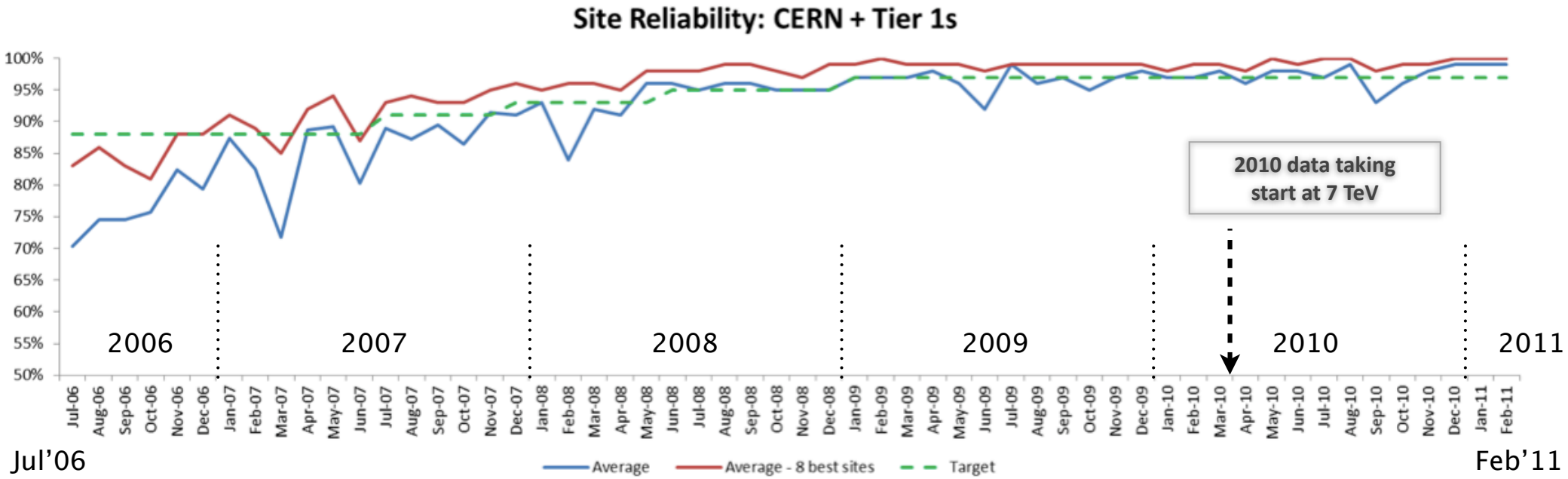
LHCONE and ATLAS/CMS view

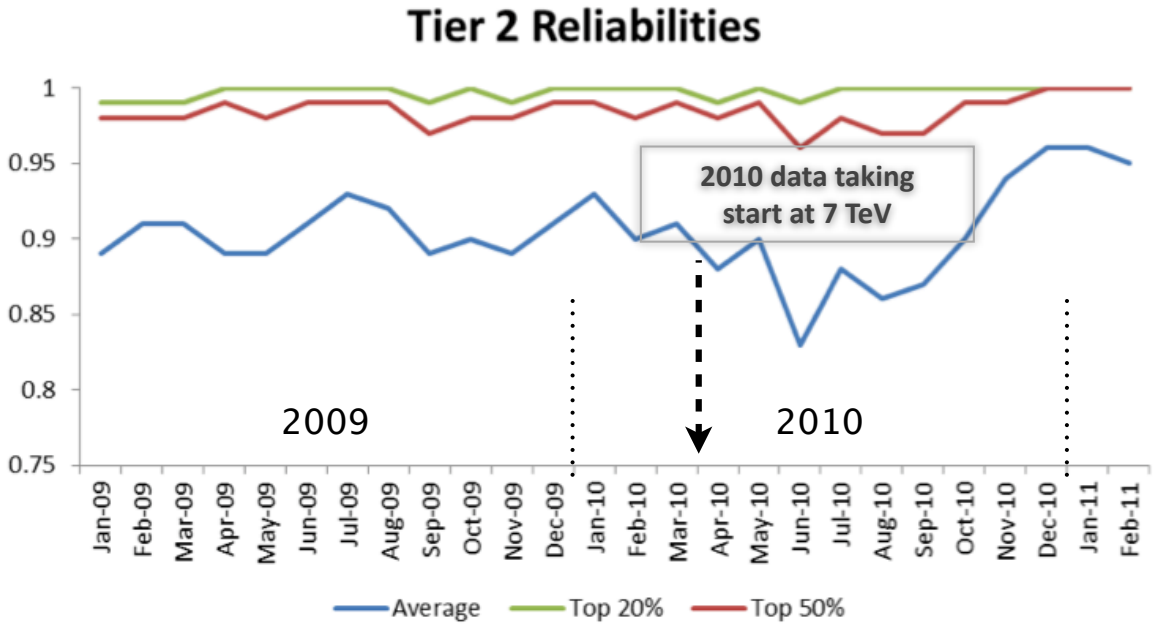
Work in progress

WLCG today for LHC experiments



Site reliability in WLCG





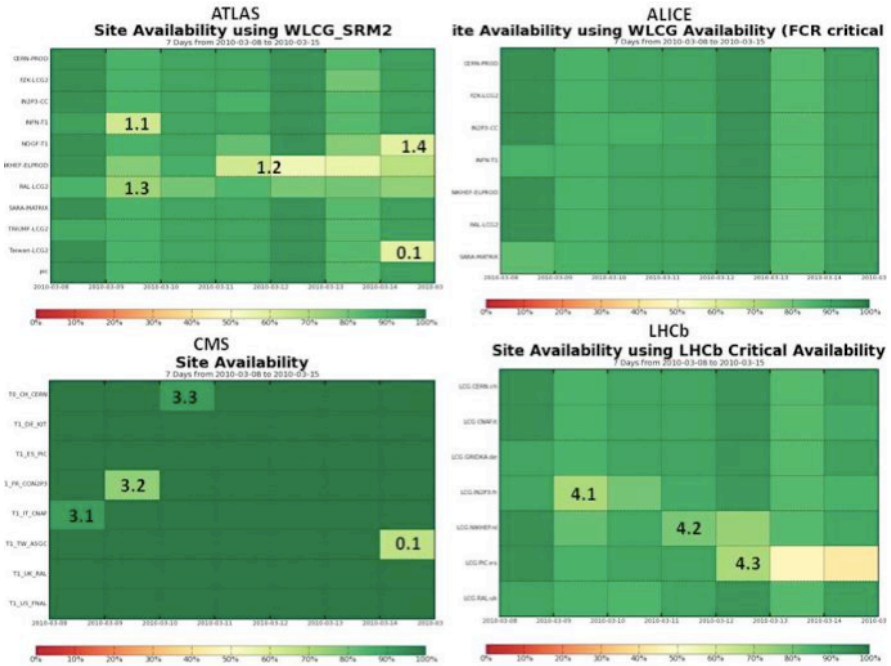
Basic monitoring of WLCG services

- at Tier-0/1/2 levels

Sites reliability is a key ingredient in the success of LHC Computing

- We have reliable T2s as we have reliable T1s
- A variegate community, but it's meaningful to rely on computing activities at T2s, more and more

Readiness of WLCG Tiers

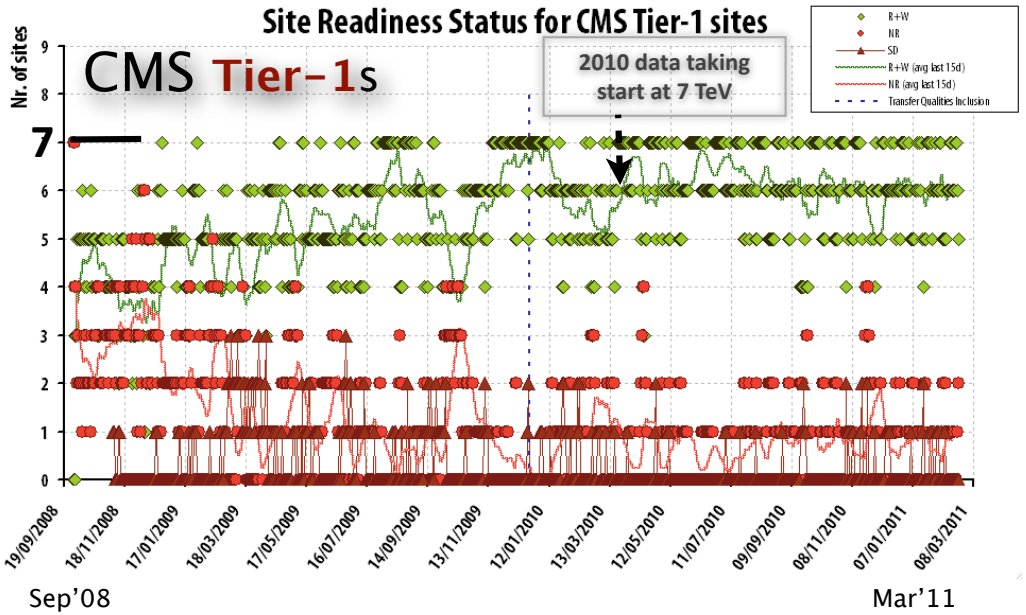


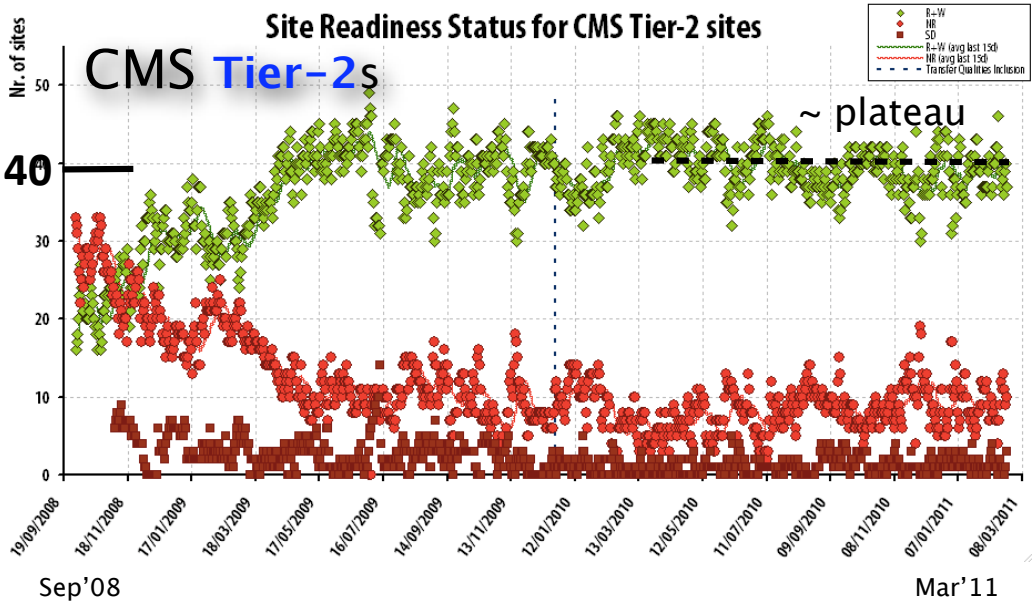
Site Availability Monitoring

- Critical tests, per Tier, per experiment

Some experiments built their own readiness criteria on top of basic ones

- e.g. CMS defines a “site readiness” based on a boolean ‘AND’ of many tests
  - Easy to be OK on some
  - Hard to be OK on all, and in a stable manner...





WLCG Tiers

Computing Model(s)

Operations (focus on data transfer and access)

Motivations for LHCONE

LHCONE and ATLAS/CMS view

Work in progress

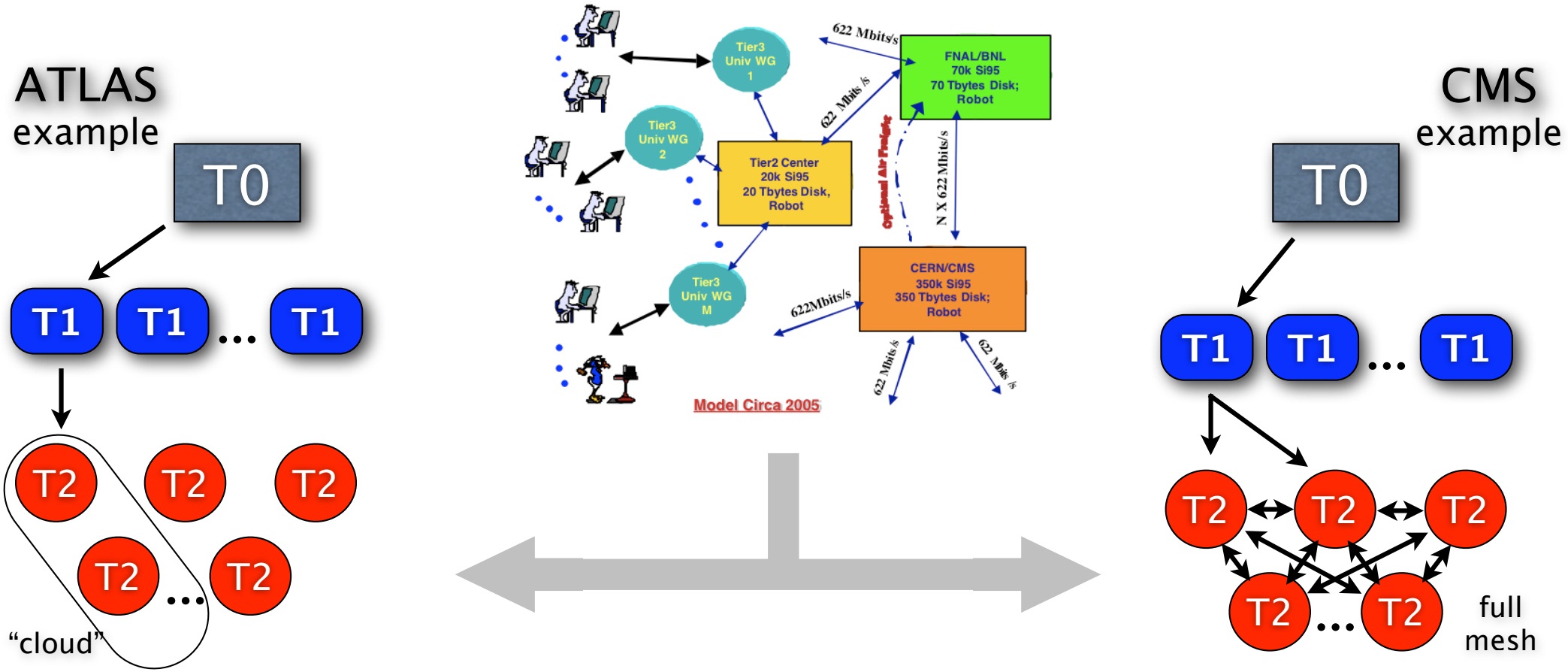
LHC Computing models

LHC Computing models are based on the MONARC model

- Tiered computing facilities to meet the needs of the LHC experiments

MONARC was developed more than a decade ago

- It served the community remarkably well, evolutions in progress



From commissioning to data taking

"**Data Challenges**":
experiment-specific, independent tests
(first full chain of computing models on grids)

"**Service Challenges**":
since 2004, to demonstrate service aspects:
- DM and sustained data transfers
- WM and scaling of job workloads
- Support processes
- Interoperability
- Security incidents ("fire drills")

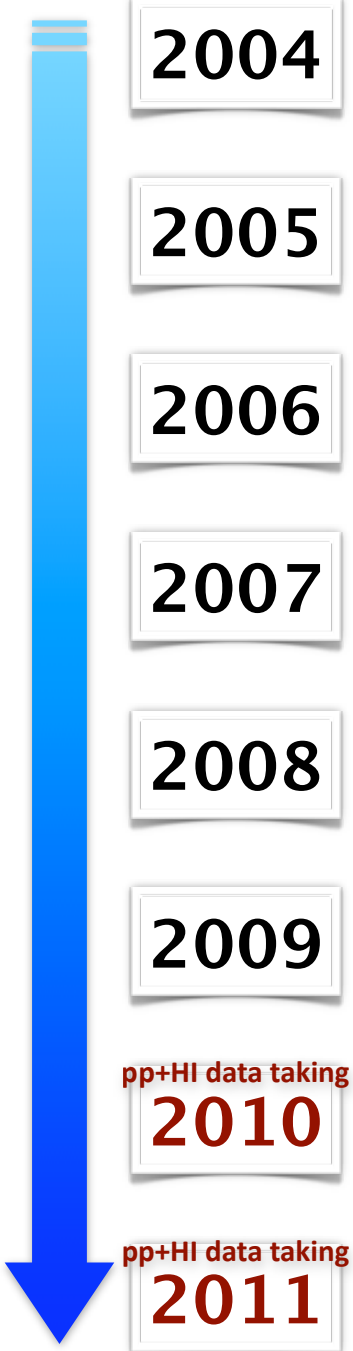
Run the service(s):

Focus on real and continuous production use of the services over several years:
- simulations (since 2003)
- cosmics data taking, …

+

"**Readiness/Scale Challenges**":

Data/Service Challenges to exercise aspects of the overall service at the same time
- if possible with VO overlap



DC04 (ALICE, CMS, LHCb)
DC2 (ATLAS)

SC1 (network transfer tests)

SC2 (network transfer tests)

SC3 (sustained transfer rates, DM, service reliability)

More experiment-specific challenges…

SC4 (nominal LHC rates, disk→tape tests, all T1, some T2s)

More experiment-specific challenges…

CCRC08 (phase I - II)
(readiness challenge, all exps, ~full computing models)

STEP'09
(scale challenges, all exps + multi-VO overlap, FULL computing models)

WLCG Tiers

Computing Model(s)

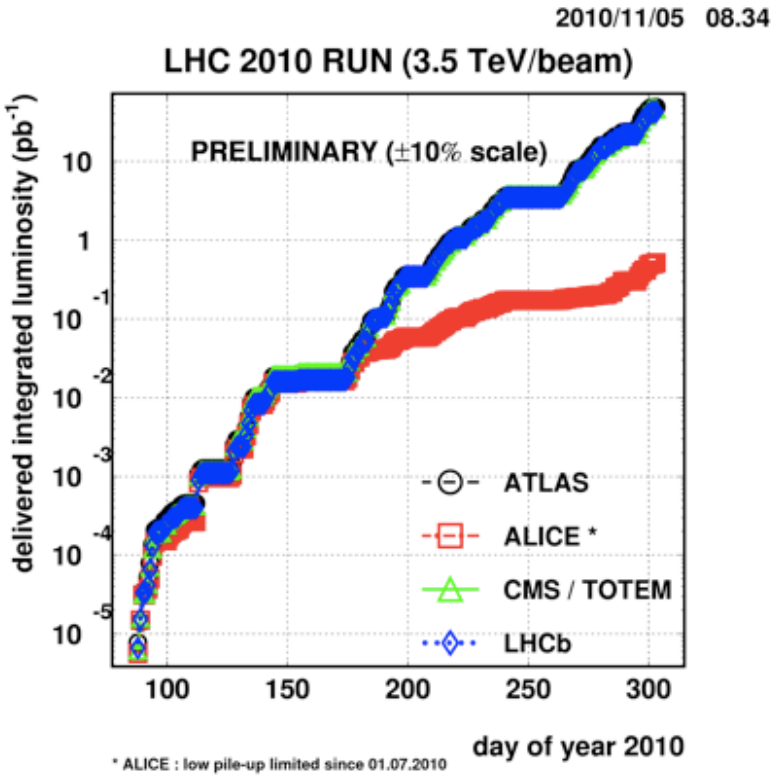
Operations (focus on data transfer and access)

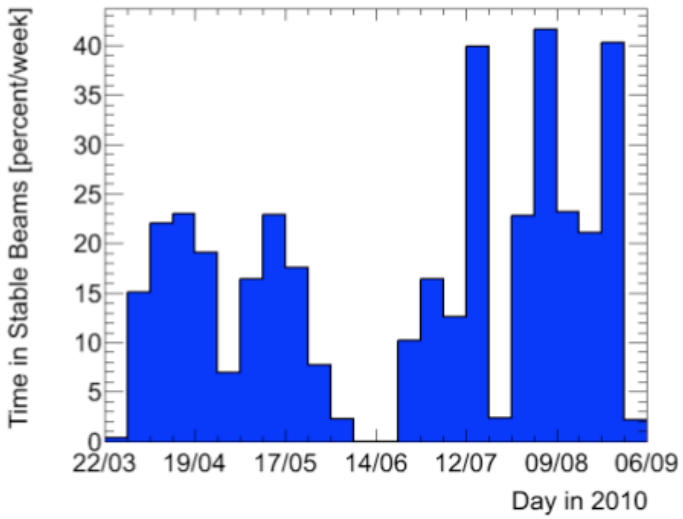
Motivations for LHCONE

LHCONE and ATLAS/CMS view

Work in progress

LHC data taking 2010





Remarkable ramp-up in lumi in 2010

- At the beginning, a "good" weekend could double or triple the dataset
- a significant failure or outage for a fill would be a big fraction of the total data

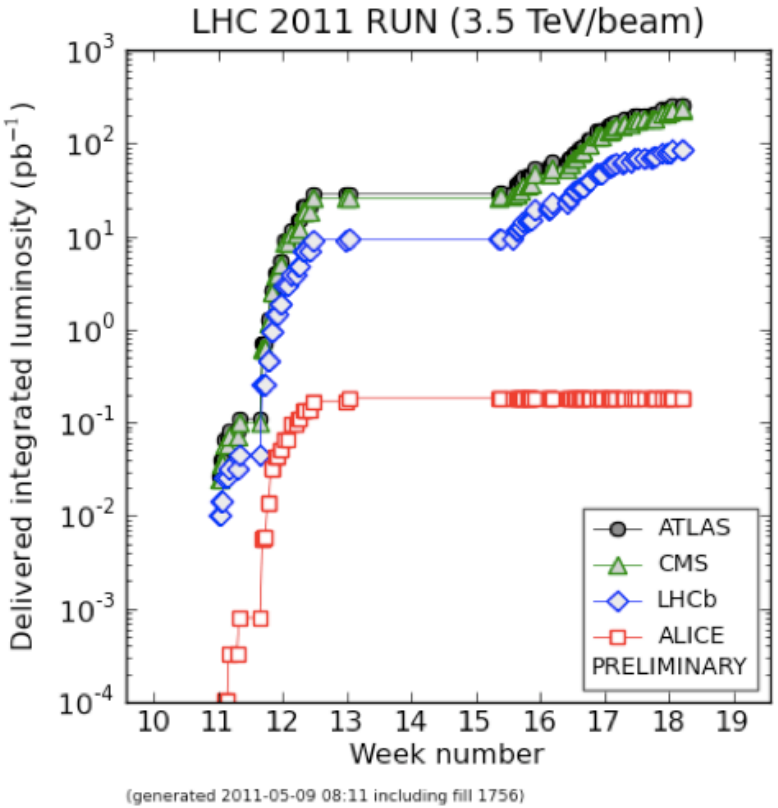
Original planning for Computing in 2010 foresaw higher data volumes

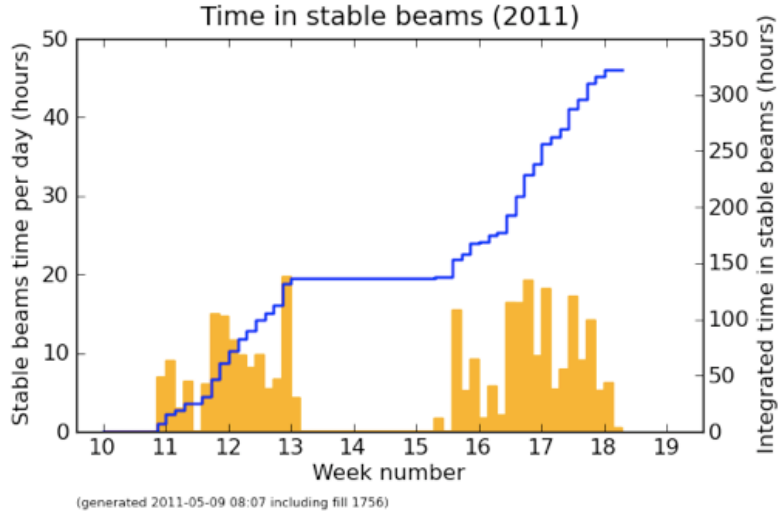
- Time in stable beams per week reached 40% only few times

Load on computing systems lower than expected, no stress on resources

- Slower ramp has allowed predicted activities to be performed more frequently

LHC data taking 2011





Going extremely well

- ~250-300 pb^{-1} so far
- ~1.7 E11 protons/bunch
- Expect ~1400 bunches in June

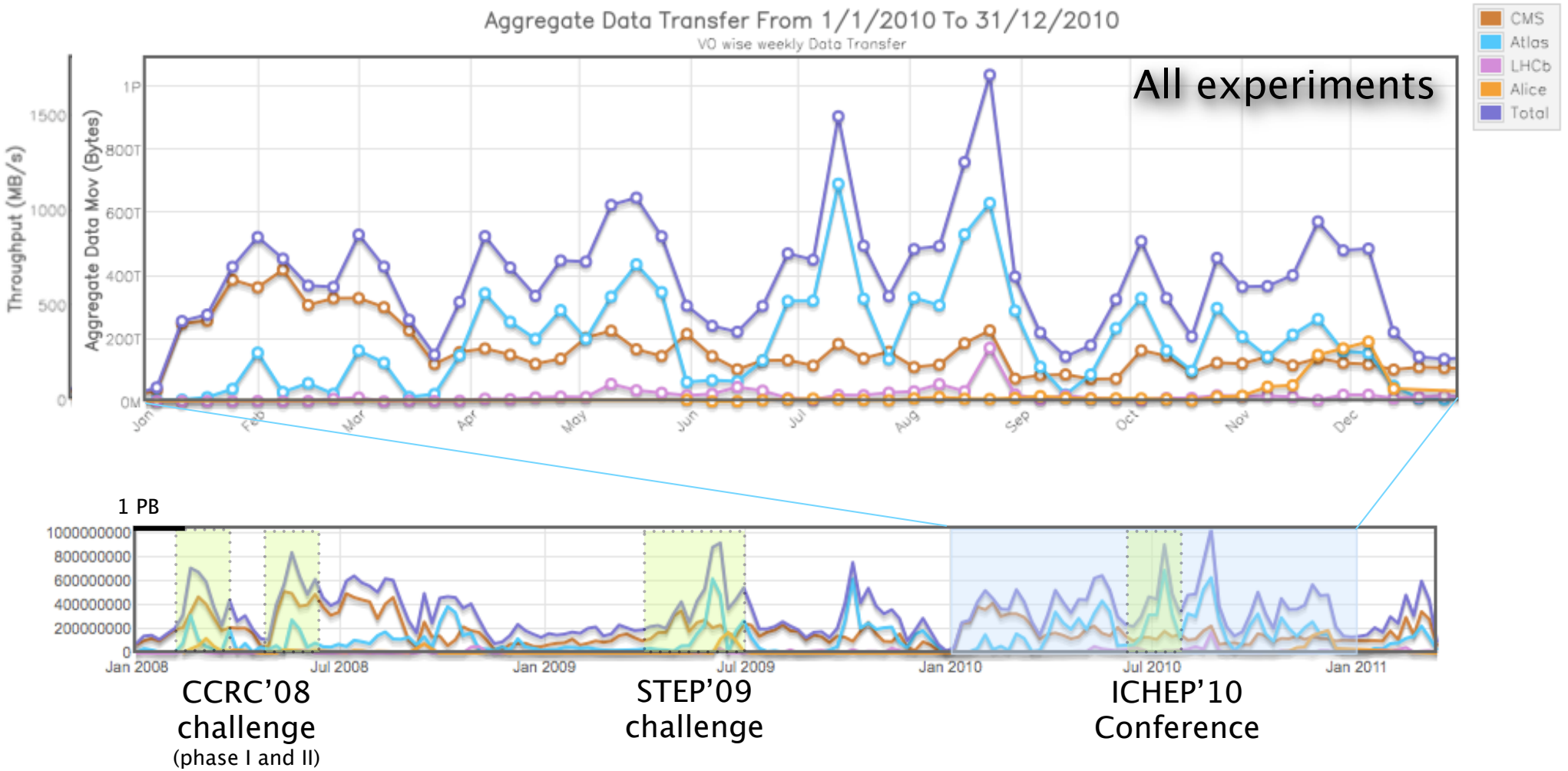
2011: consistent load on resources

- we will be resource constrained

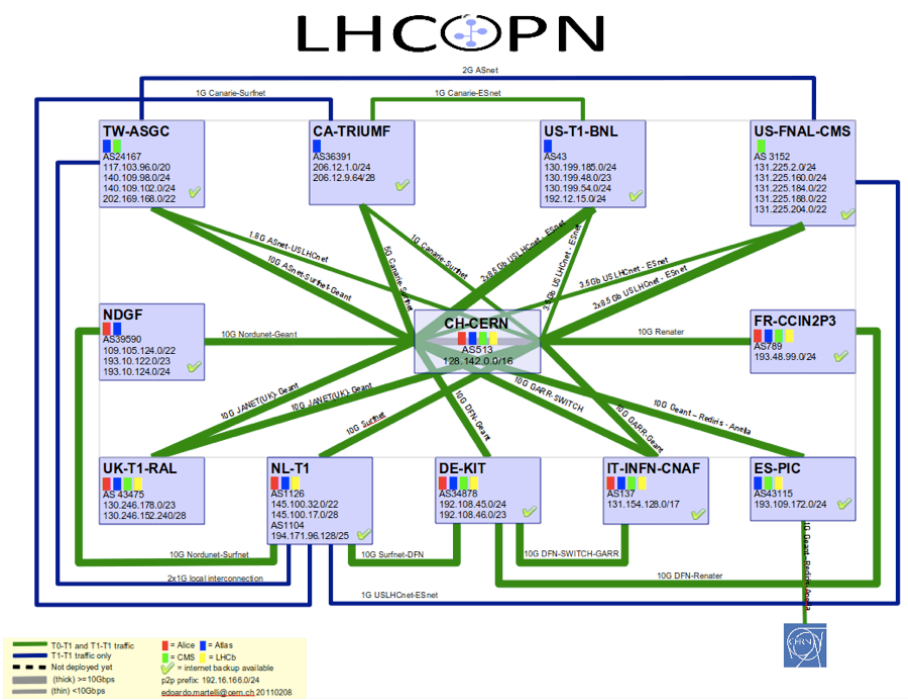
CERN→T1 data transfers

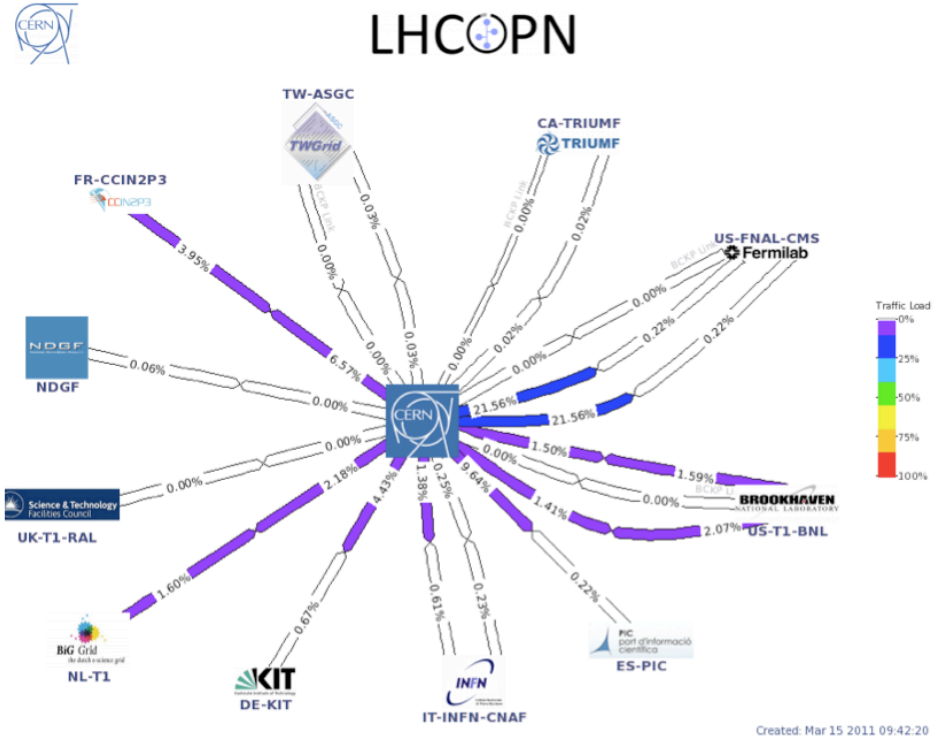
CERN outbound traffic showed high performance and reliability

- ✦ Very well serving the needs of LHC experiments
- ✦ Under control



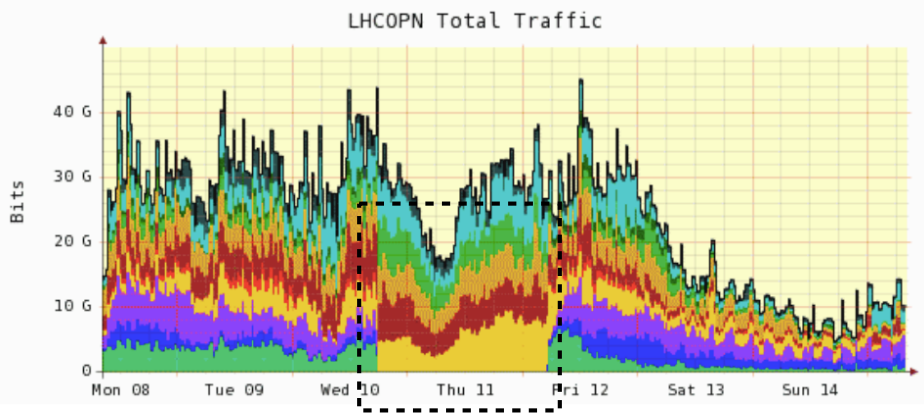
LHCOPN network in operations





OPN links now fully redundant

- Means no service interruptions
  - See the fiber cut during STEP'09



WLCG Tiers

Computing Model(s)

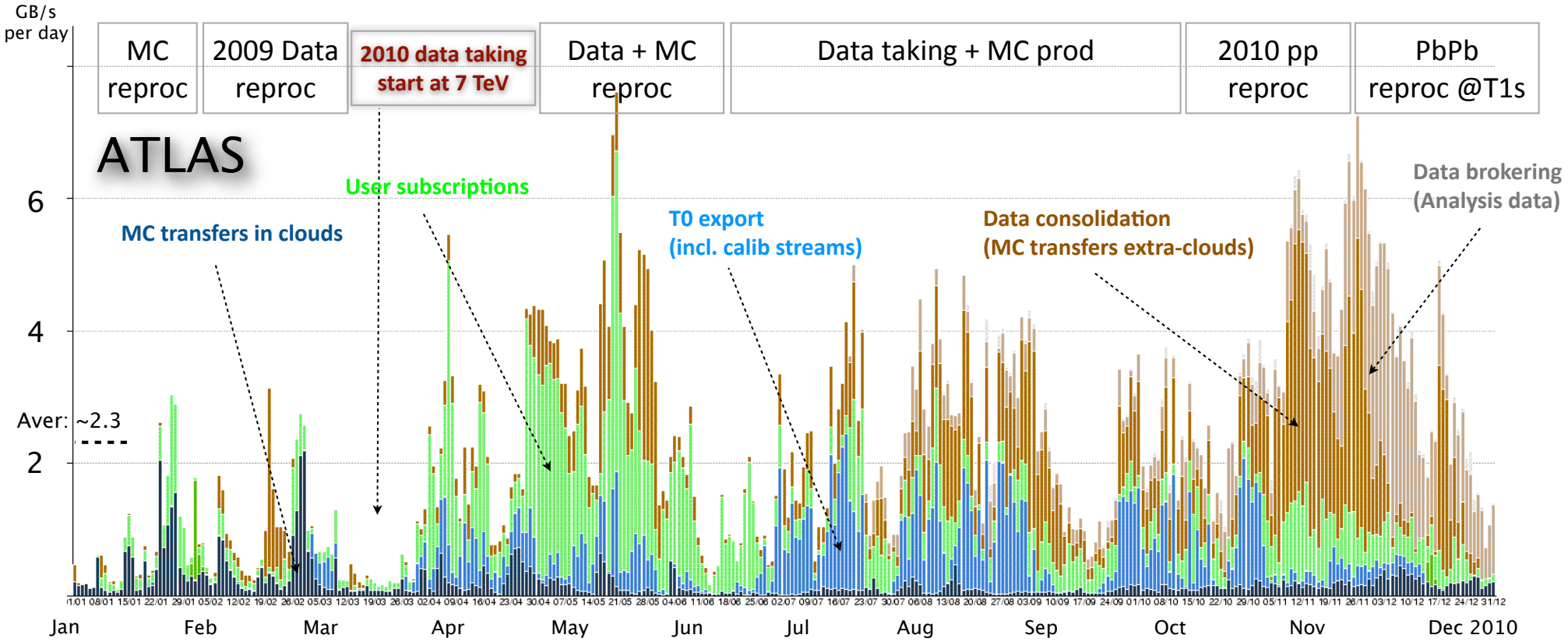
Operations (focus on data transfer and access)

Motivations for LHCONE

LHCONE and ATLAS/CMS view

Work in progress

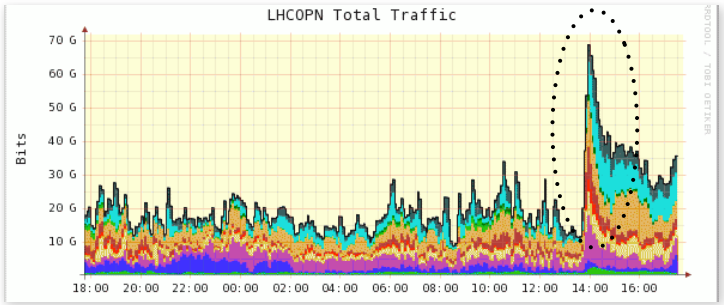
An example: **ATLAS data transfers**



Transfers on all routes (among all Tier levels)

- Average: **~2.3 GB/s** (daily average)
- Peak: **~7 GB/s** (daily average)

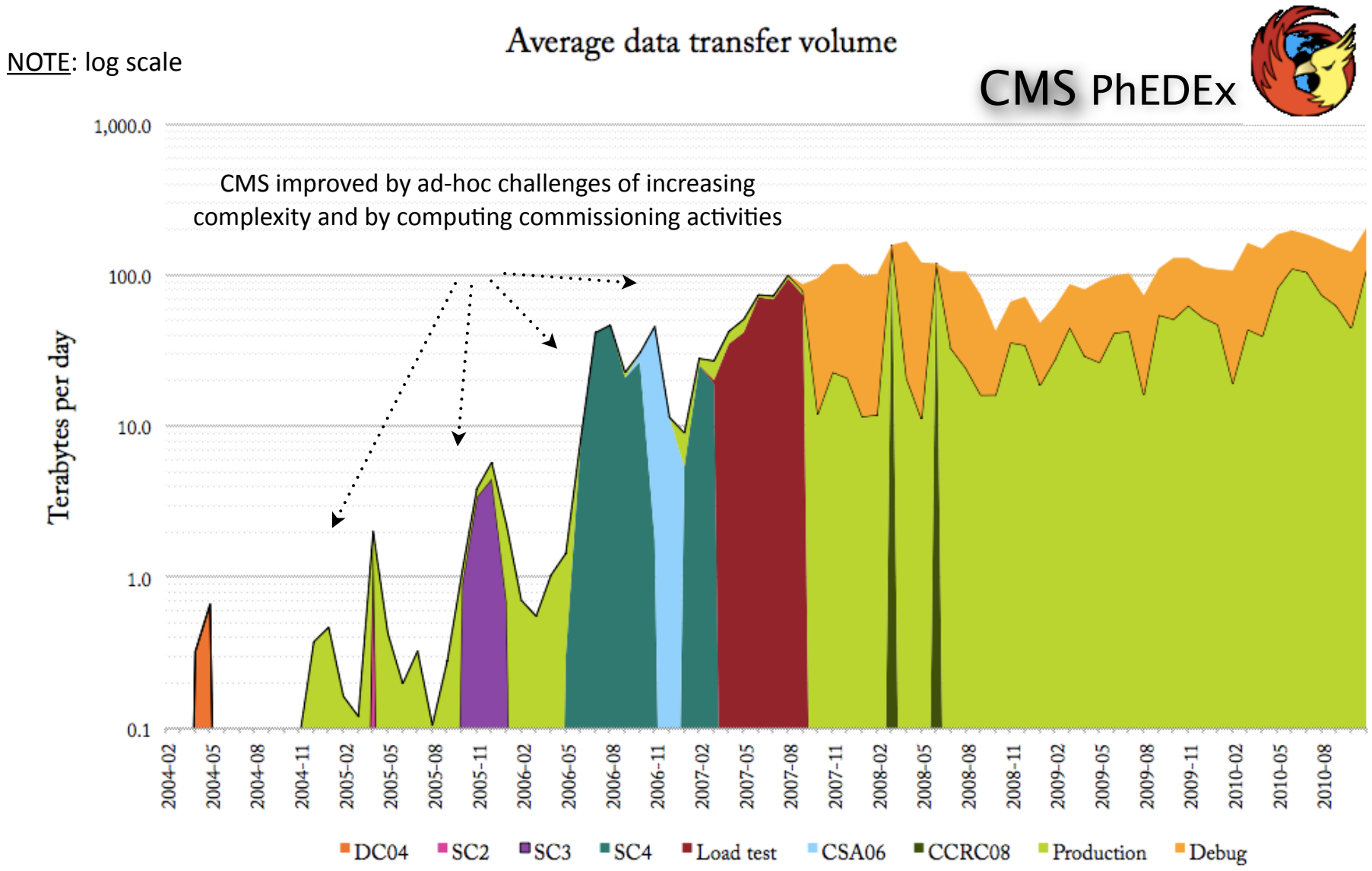
Data available on-site after few hrs.



Traffic on OPN measured up to 70 Gbps

- ATLAS massive reprocessing campaigns

An example: **CMS data transfers**



Massive commissioning, now in continuous production-mode of ops

- Can sustain up to >200 TB/day of production transfers on the overall topology

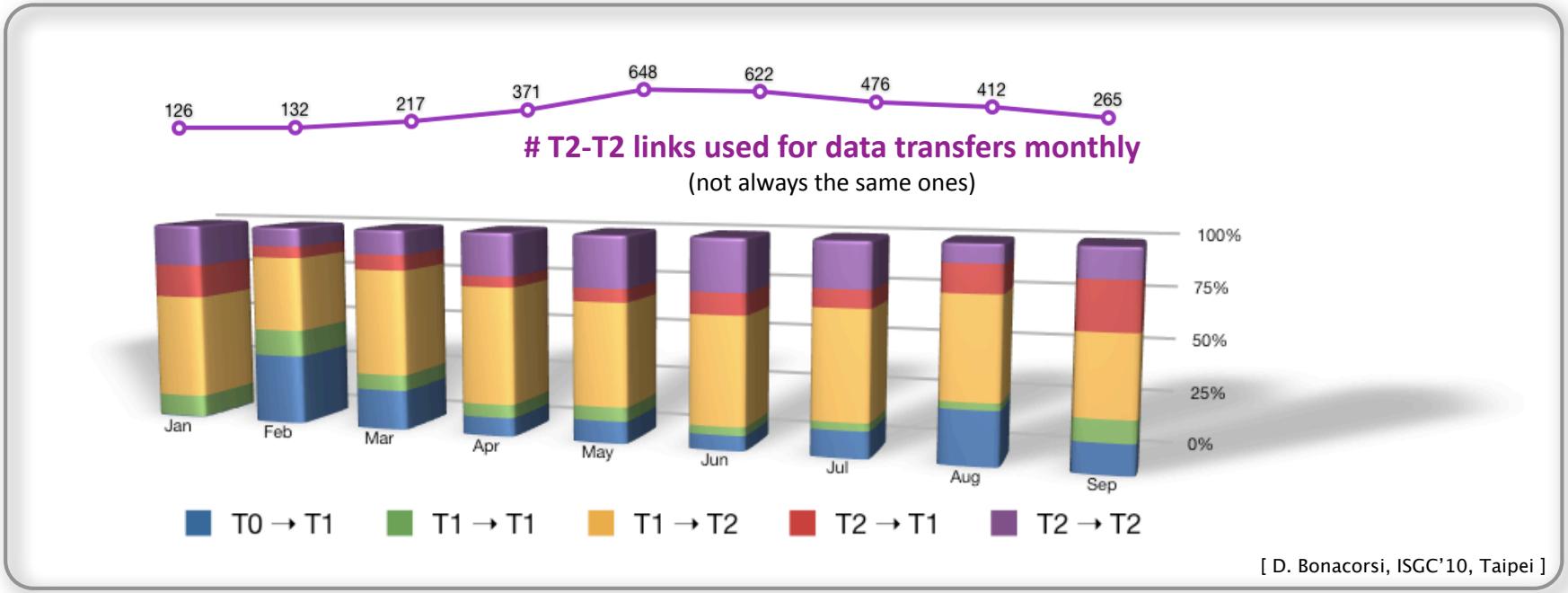
Data placement for analysis: an example

Data population and access by analysis applications at T2 level by CMS

- Largest fraction of analysis computing at LHC is at the T2 level
- Flexibility of the transfer model help to reduce the latency seen by the analysis end-users

This triggered the interest of DANTE

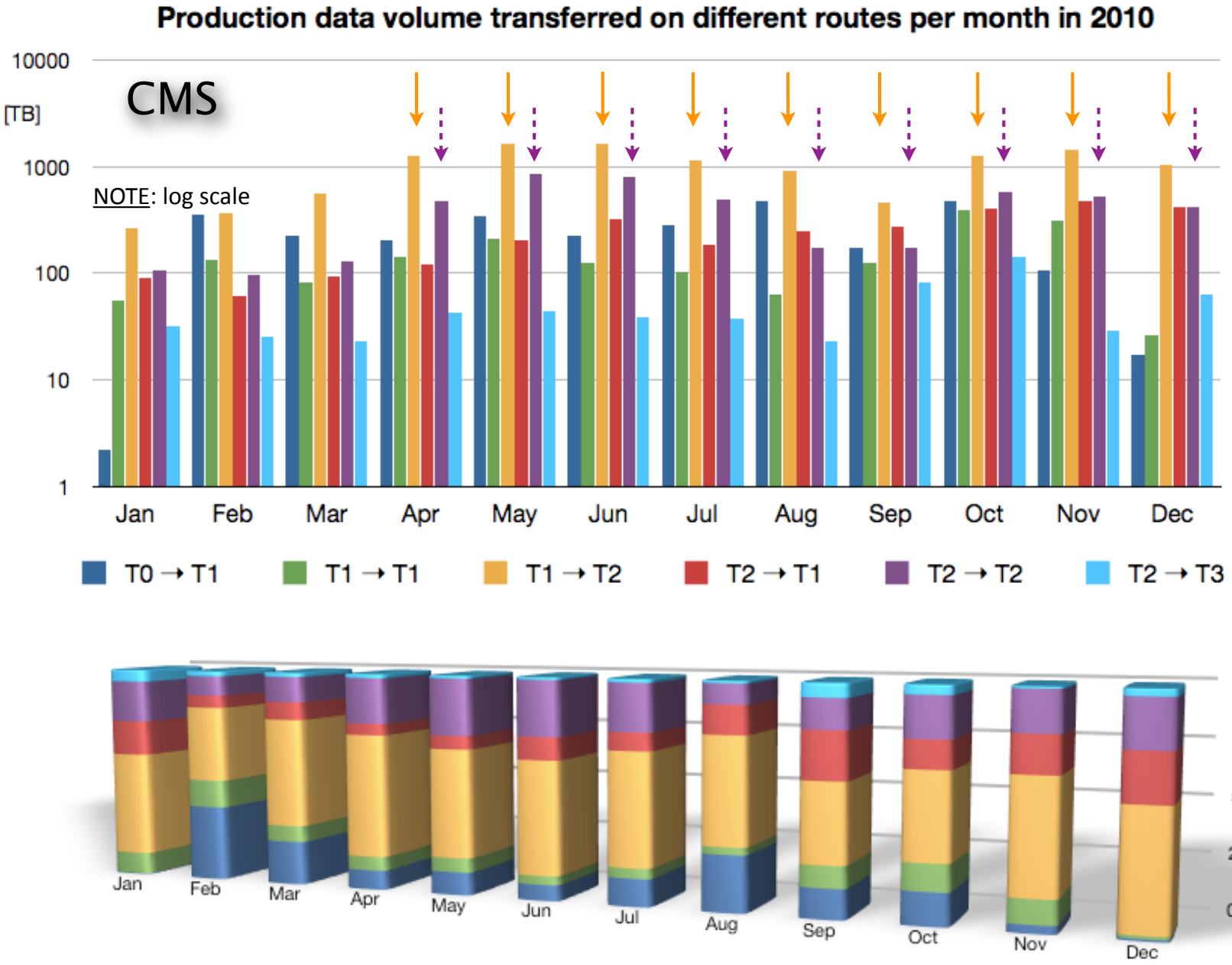
- in ISGC'10 in Taipei and in EGI-UF 2011 in Vilnius



[D. Bonacorsi, ISGC'10, Taipei]

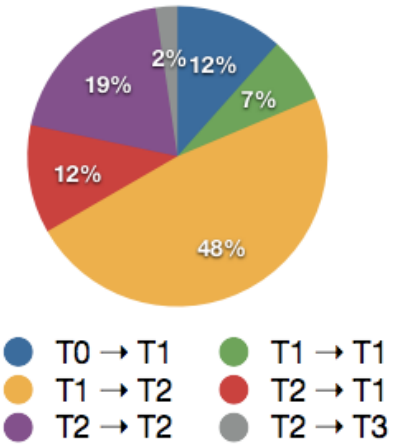
See updated plot

2010 Tx-Ty traffic breakdown in CMS

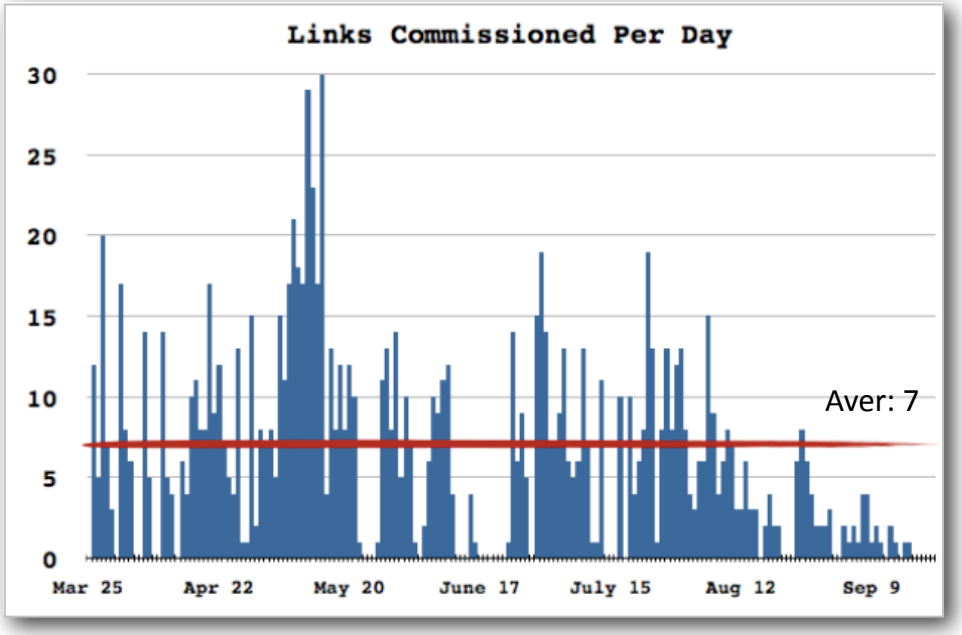


T1-T2 dominates

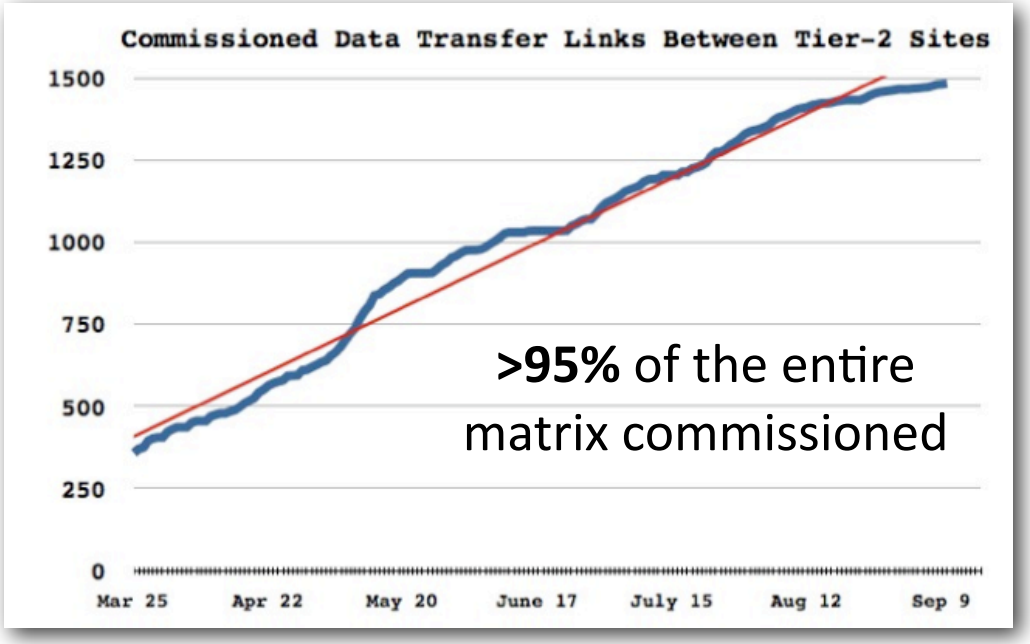
T2-T2 emerges



T2-T2 commissioning



Up to 30 links commissioned per day in CMS in 2010, average is **~7 links/day** over the first 6 months of data taking



ATLAS is doing something similar to CMS now, in testing and starting to use “extra-cloud” links among T2s

WLCG Tiers

Computing Model(s)

Operations (focus on data transfer and access)

Motivations for LHCONE

LHCONE and ATLAS/CMS view

Work in progress

LHCONE for LHC experiments

LHCONE is described elsewhere

- See lhcone.net

Food for thoughts from some LHCONE discussions:

- It addresses a problem we (experiments) do not have yet
  - Avoid things from decaying
- A lot is going on, at different levels (also: not technical)
  - Build the cheapest infrastructure that satisfies the requirements, do not let anyone out
- Allow e-Sciences to grow and operate
  - Avoid congestioning GPNs with “not-MONARC” LHC traffic
- Opportunistic approach, so far
  - “collect the low-hanging fruits first” vs “open env: should not increase the digital divide”
  - Reach a critical mass that pushes the entire process
- It’s a work in progress
  - Both in the network communities and in the experiments communities

ATLAS-specific view

Private communication with S. Jezequel and I Ueda, followed by:
[*] meeting with D.Foster, A.Barczyk, I.Fisk, D.Bonacorsi

ATLAS experienced several "issues" with the network, e.g.:

- more and more often saturating, accidents, low-performance events
  - e.g. CNAF-BNL recently
- quite unpredictable. And seeing an event does not necessarily justify actions.
  - What actually is a problem and should trigger actions needs to be clarified

Expectation from LHCONE:

- Get a list of (ATLAS) sites which will be connected to LHCONE before July (hopefully with a timescale)
  - This was somehow understood at a meeting [*] that the proactive sites will eventually rule the game, and if experiments want some site to go first they should speak up
- Ensure that network team will validate the new path before ATLAS starts transfer tests

Impact on ATLAS

- Check that the transfer rate for single files between these LHCONE sites and some selected T1s is identical or better than before
  - based on ATLAS sonar test
- Define some stress transfer tests (before and after the migration) between these LHCONE sites and some selected T1s to measure any possible improvement in transfer rate
  - maybe need to create some activity on the public network

If LHCONE has any request or better view of interesting tests, this can be discussed.

CMS-specific view

My input, also discussed at:
[*] meeting with S. Jezequel, I Ueda, D.Foster, A.Barczyk, I.Fisk

CMS data on transfers in most routes exist.

But are historically and contextually diverse in richness

- CMS did T2-T2 commissioning and measurements
  - e.g. we have a-la heartbeat (commissioning the links) plus 24-hrs best periods, for most links
  - But we cannot reliably predict how a given is link is working if used tomorrow
- Some links have never been run at high level
  - a little hard for >2.5k links to say if this is because they were not tested, or because a request/subscription was never made.

Important for CMS:

- We need to test **BEFORE** and **AFTER** any change
  - The BEFORE gives you the benchmark, the AFTER gives the feedback to exps and Networks
- To know when site-X will be connected - with some advance
  - Quantified [*] in at least 1 week. This point needs to be reinforced, and sites informed.
- CMS does not expect a major improvements in the network
  - We expect to verify it does not get worse, and it gets eventually more predictable
- Get guidance by network experts about the change at site-X
  - OK on lhcone.net: what changes were done, when, how we can monitor now
- Get prompt feedback in case experiments notify post-change issues
  - We will inform network experts through communication channels we will be suggested to use

WLCG Tiers

Computing Model(s)

Operations (focus on data transfer and access)

Motivations for LHCONE

LHCONE and ATLAS/CMS view

Work in progress

What's going on?

So far, smooth convergence in ATLAS and CMS on the needs and general ideas on how to get prepared.

ATLAS: a meeting on the subject at the end of May.

CMS: decided to adopt a strategy consisting of two complementary approaches

① Use **PhEDEx /Debug** and the **LoadTest** (LT) infrastructure with **Debugging Data Transfers** (DDT) procedures

- LoadTest infrastructure is reliable, fully PhEDEx-integrated, versatile
- a PhEDEx /Debug instance is available to play with (CMS did it for T2-T2 commissioning)

Need to reduce the # links (only the ones in the Bos-Fisk document, not the whole 2.5k matrix!)

② Use the work-in-progress on **FTS/FTM parsing**

Gives complementary, customizable, detailed info for each link (rate per stream, throughput)

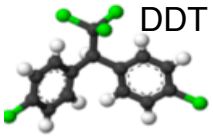
- Once you collect data, the problem is addressed once and scales for N links

NOTE: part ② of the CMS approach is not CMS-specific, and can be adopted by other experiments (if they use FTS/FTM).

1nd approach:

Use PhEDEx and LT based on DDT procedures

①

PhEDEx LoadTest (LT)

- A flexible infrastructure to generate "test" data transfer load among sites
  - "fake" but "real": test files fully integrated in PhEDEx
- Flexible and customizable (e.g. you choose source site, destination site, rate)
- ~24/7 activity since early-2007

Debugging Data Transfers (DDT)

- A program to maintain a high-quality transfer network via commissioning links
- Metric, monitor, troubleshooting, site involvement, doc of success stories
- a Task Force in charge since mid-2007

The idea for LHCONE tests is:

- exploit the LT infrastructure to test links among the list of first-comers
  - as from the Bos/Fisk document
- fully change the DDT procedure for this purpose
  - transition it from a "commissioning" scope to a "performance monitoring" scope

Work is mainly manual. Started already.

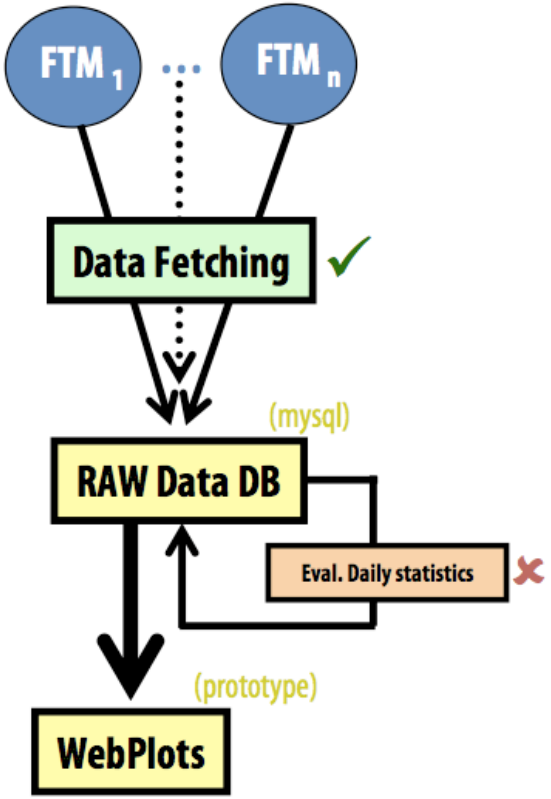
2nd approach:

An accounting tool for FTS transfers

②

The tool is meant to provide an accounting of FTS transfers at the individual file level, which is:

- historical: keeping statistics over a long time period;
- global: attempt to get data from all FTS servers used by CMS;
- low level: collect data as rates, rates/stream, queue time, SRM overheads, etc;



Experiments do not have such tool. E.g. in CMS we have:

- PhEDEx: global and historical but "high level";
- FTS Monitors: "low level" but local to each FTS and no history

This is useful for a number of tasks

- Optimizing of FTS channels settings & identif of congested channels
- Creating of "cloud" FTS channels to improve channel occupancy
- Spotting general problems with endpoints and/or links
- Checking network performances, e.g. feedback on LHC{OPN,ONE}

Also triggered by LHCONE needs, CMS developed and deployed one

- It's not VO-specific. It will eventually converge in a Dashboard object

Work is mainly automatic. Started already.

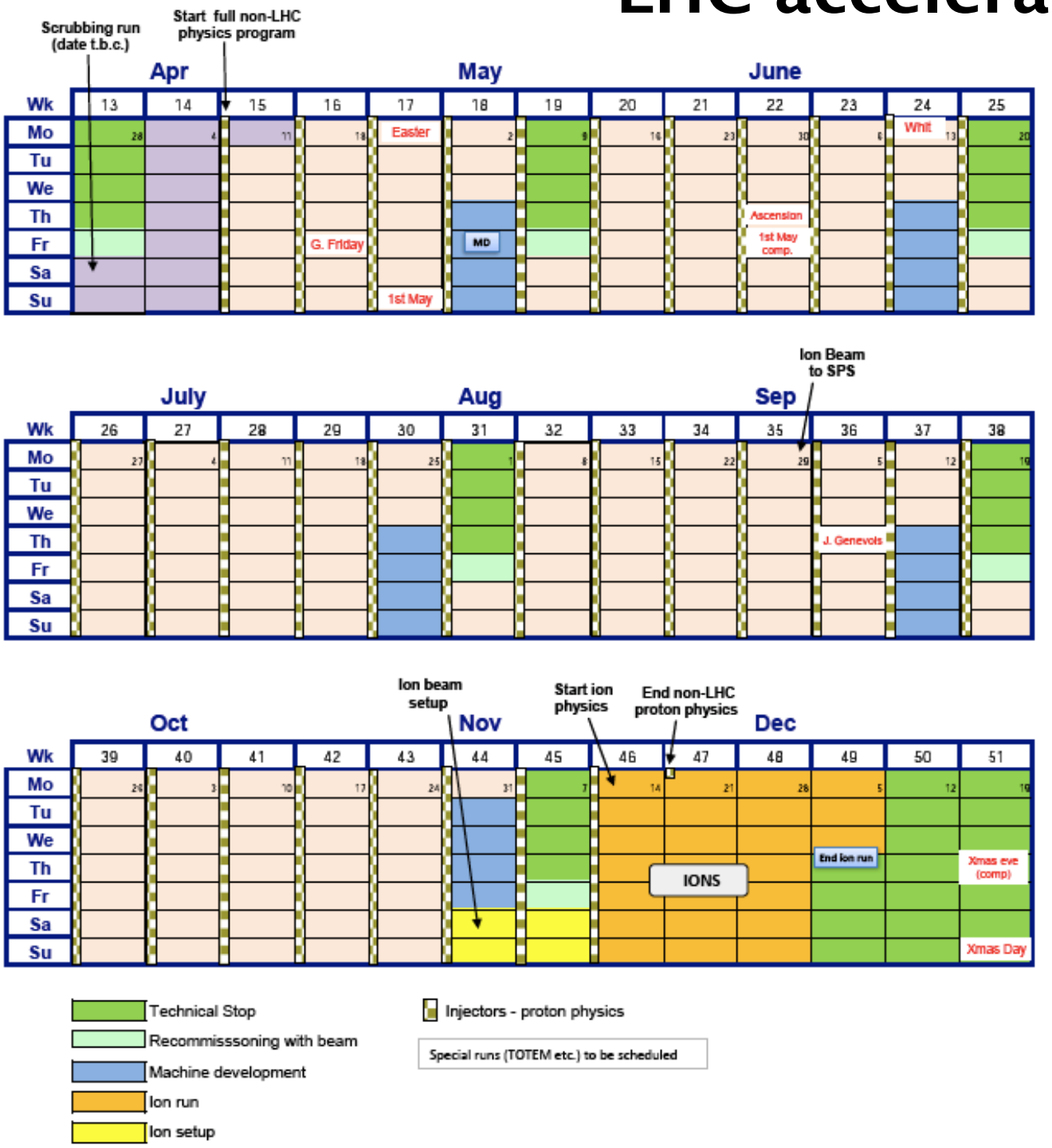
Discussion points

Points for discussions (now, and at the BOF):

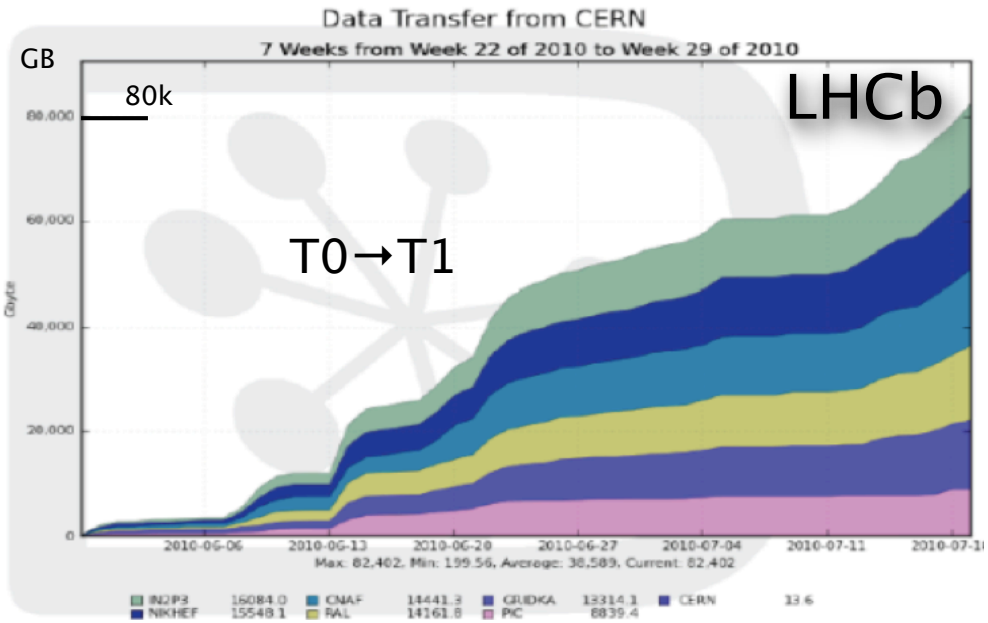
- Experiments↔Networks communication on LHCONE topics
  - Do Networks/Experiments know what Experiments/Networks are doing - planning to do?
  - Are we OK, or more cross-fertilization should be encouraged and enforced?
- Do we need to re-activate some work to describe network utilization in next years by experiments?
  - If so, in which form? Will the Networks community be actors or customers of such work?
- How do we set priorities in picking up sites and plug them in LHCONE ?
  - Prio on best performing sites (heavy/useful load on the system and want to protect it)…
  - … or prio on those that performs poorly (the biggest potential gain - some have plenty of resources)
- Any feedback by network experts on the way experiments plan to do tests?

Back-up

LHC accelerator schedule



More examples: **ALICE and LHCb data transfers**

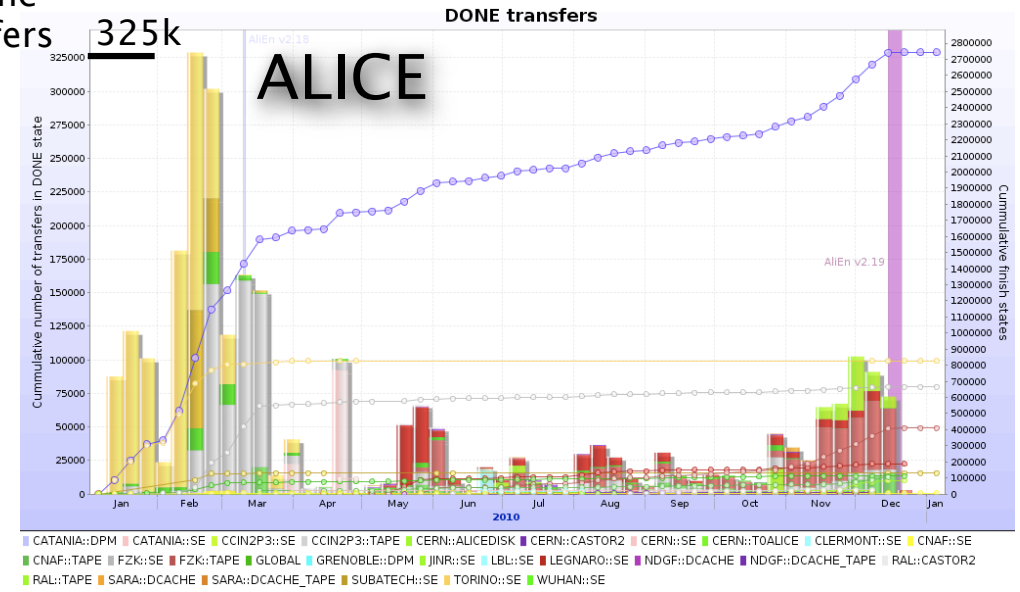


LHCb data is successfully transferred on a regular basis

- RAW data is replicated to one of the T1 sites

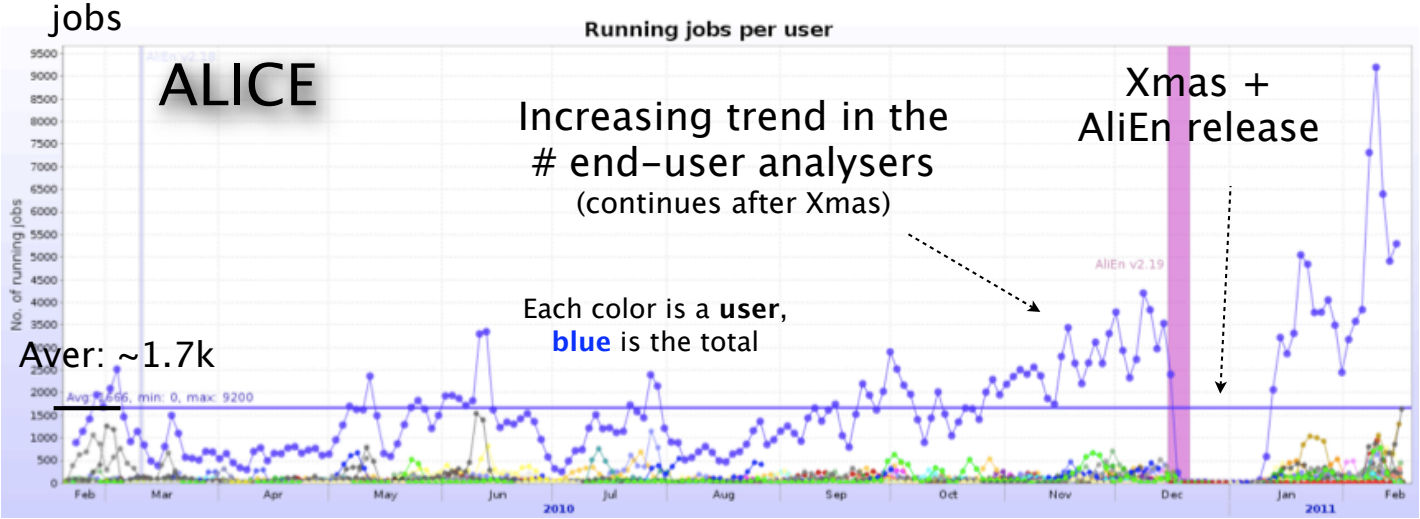
ALICE transfers among all Tiers →

done transfers



Analysis in ALICE

running jobs



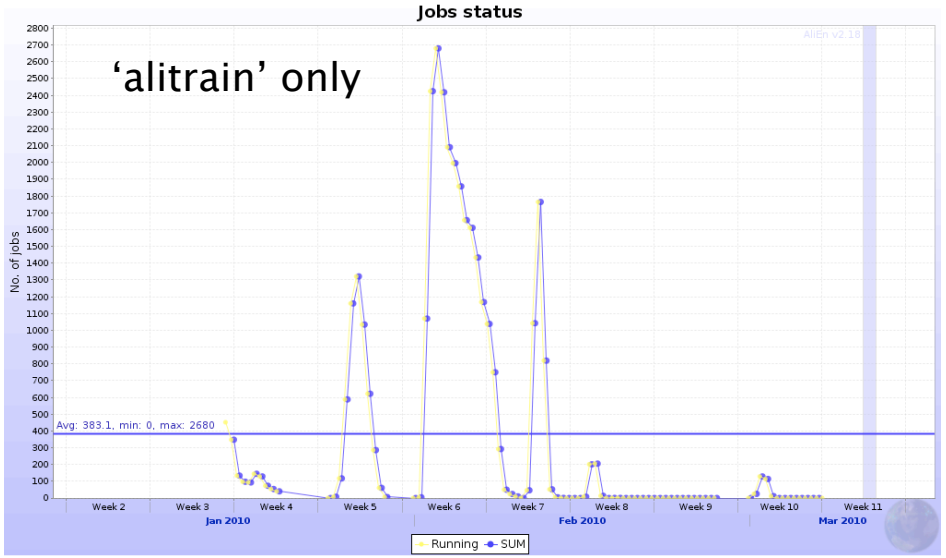
On average, 1.7k concurrent user jobs in 2010

- >9M user jobs completed over last 12 months

~200 distinct users on average, and increasing

Interesting analysis train model

- User code is picked up and executed with other analyses



Analysis Trains:

- Optimized I/O (read once, do many tasks)
- Streamlined code (as much as possible)
- Managed, scheduled (like MC sim or reco)

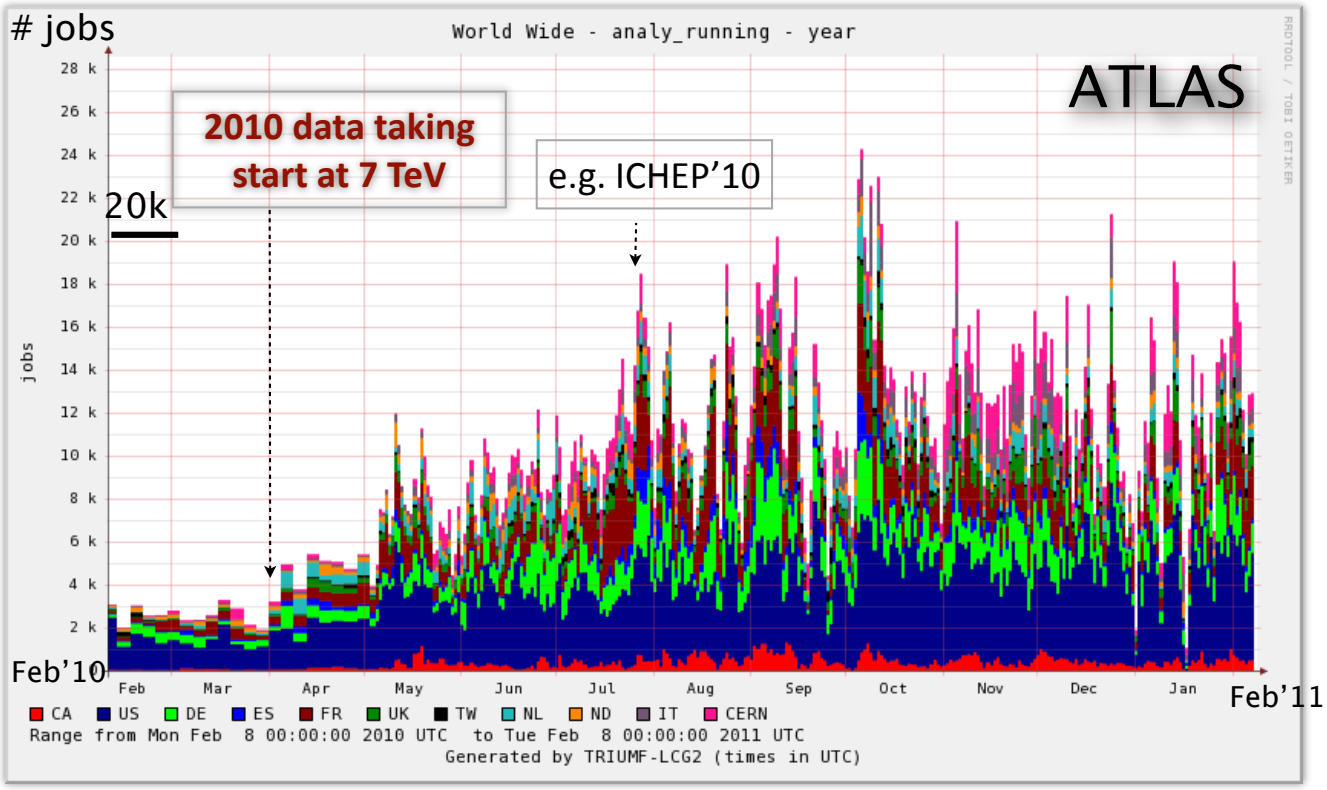
User jobs:

- Low CPU efficiency (wrt MC sim or reco)
- Variable job duration, many failures, far-from-perfect code
- Unmanaged, chaotic

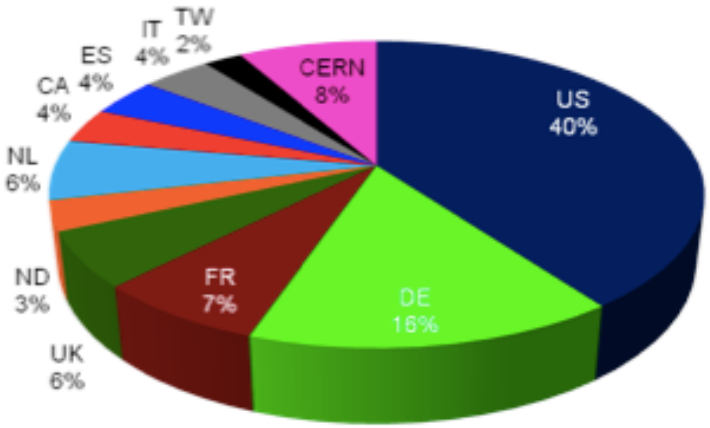
Analysis in ATLAS

Increase in analysis load after the start of 2010 data taking

- After that, roughly stable load
  - Holidays holes, as well as activities peaks before major conferences, are visible

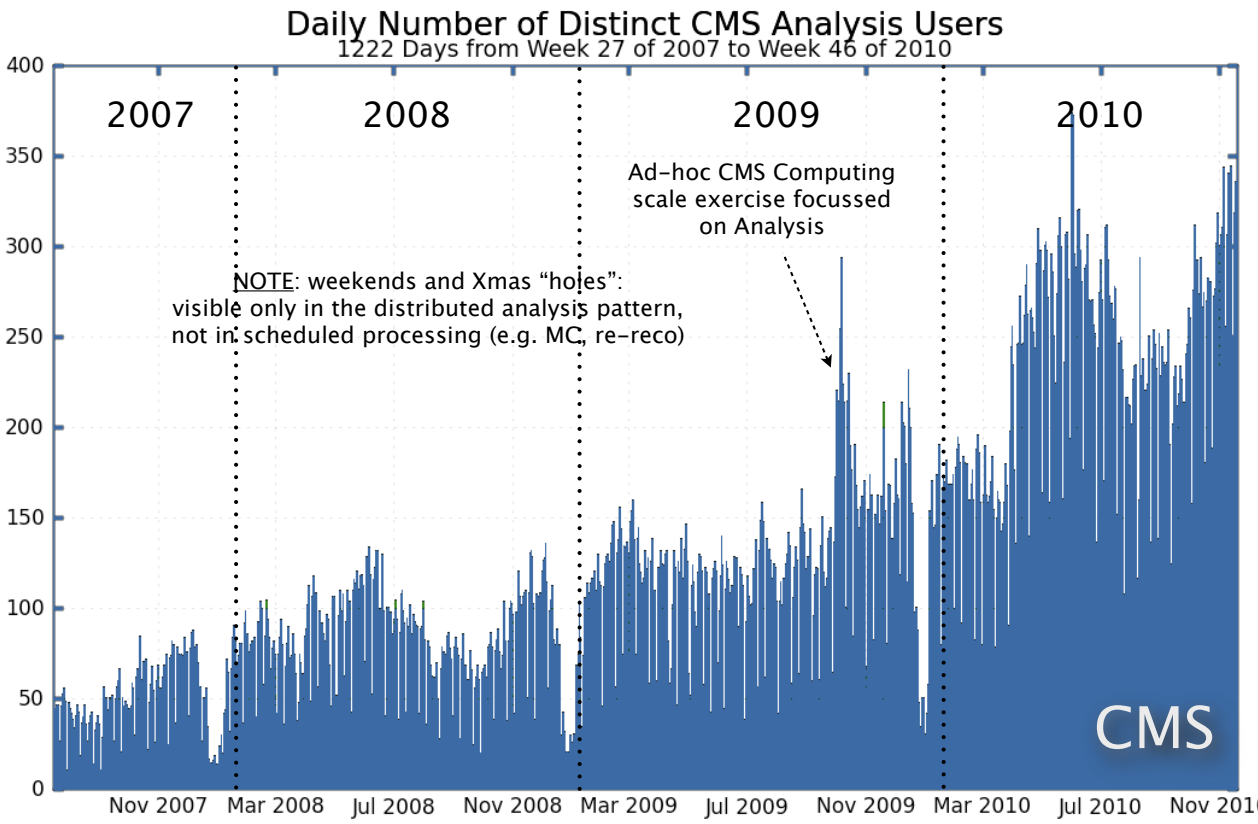


Analysis share per "cloud"



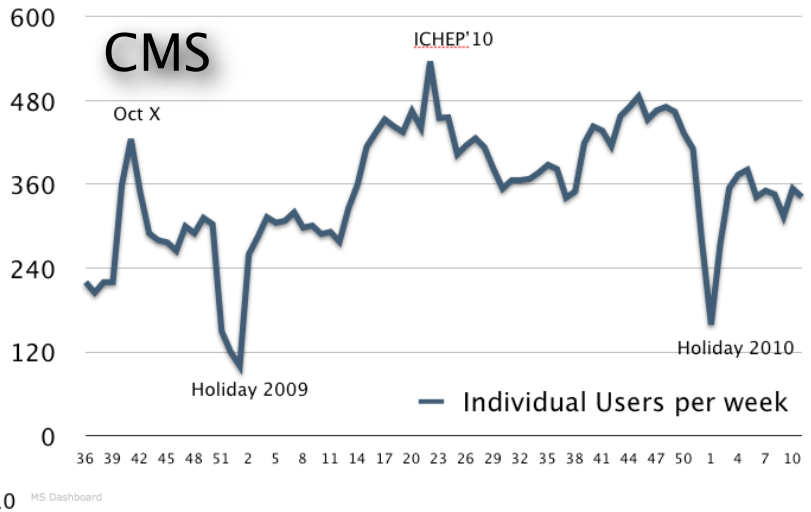
(only for pAthena-Panda system; ganga-WMS not counted)

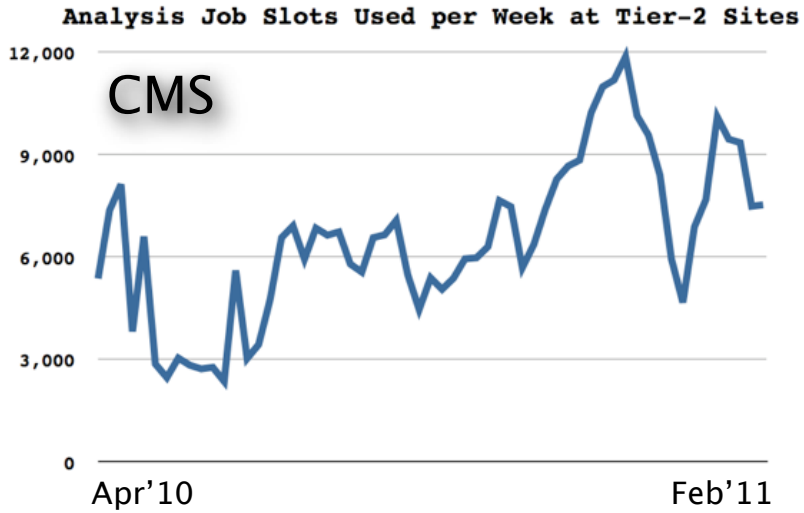
Analysis in CMS



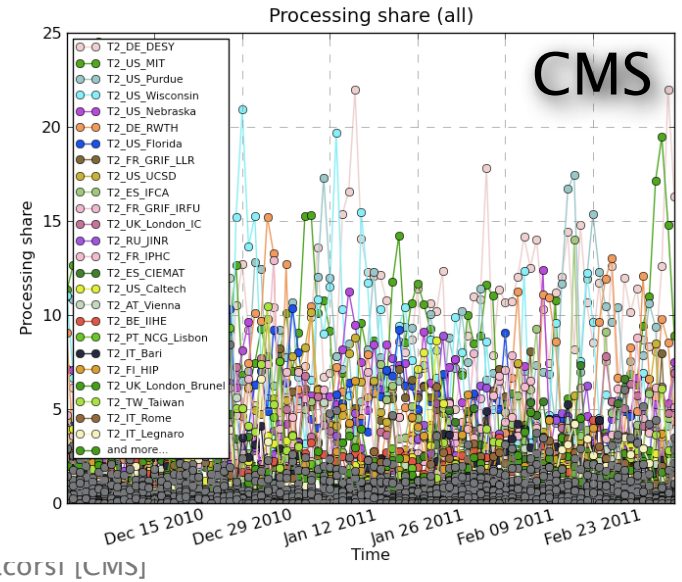
Constant increase in # users

- ~300-350 distinct daily users
- Up to >500 users per week during peaks
- >800 individuals per month





Analysis at the T2 level.



Analysis in LHCb

No a-priori assignment of site

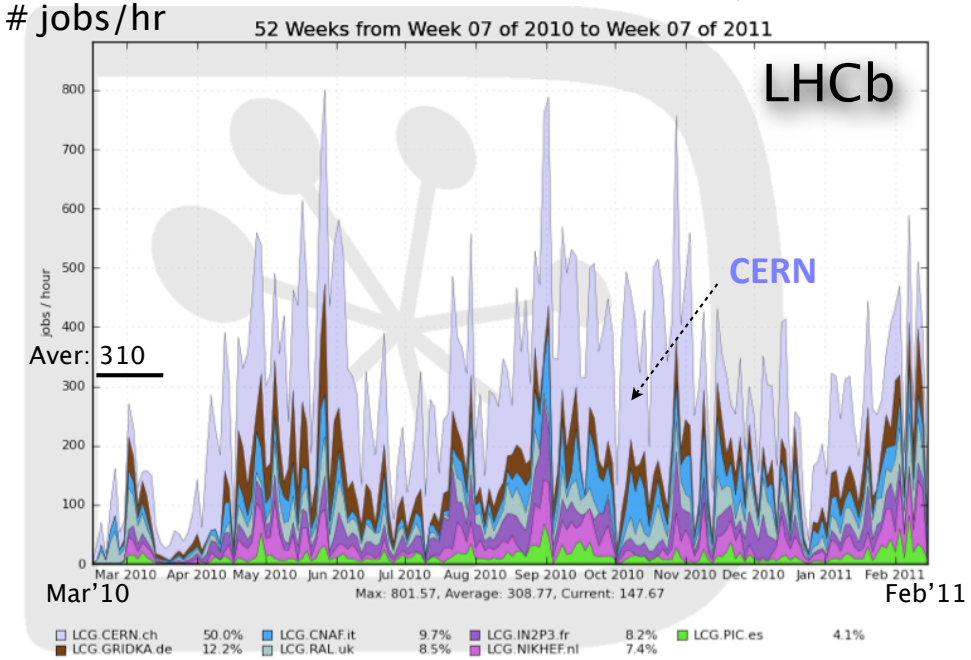
- Share by availability of resources and data

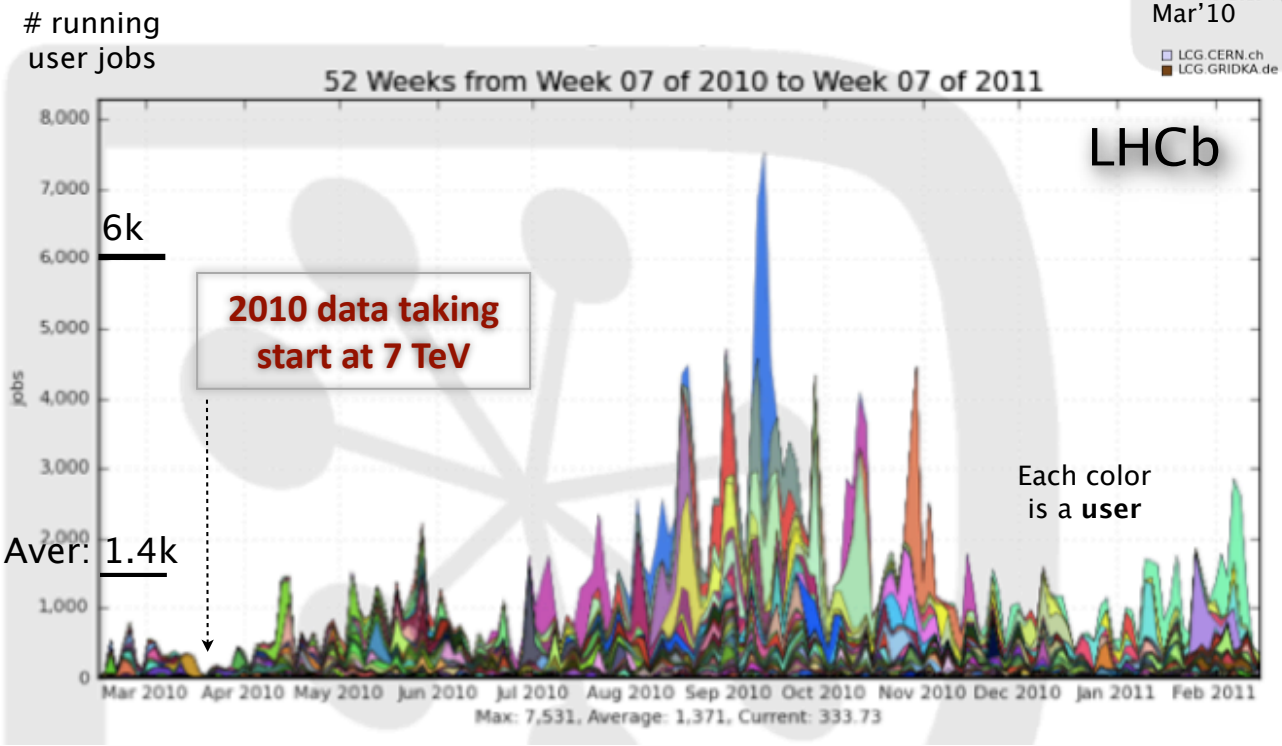
Only ~2% of analysis at T2s

- Toy MC, private small simulations, etc

~320 unique analysis users

Successful user jobs at T1s





Roughly, ~50% of LHCb analysis is performed out of CERN