Esperienza di calcolo
nel primo anno di presa dati in CMS

Daniele Bonacorsi

[deputy CMS Computing Coordinator – Università di Bologna, Italy]
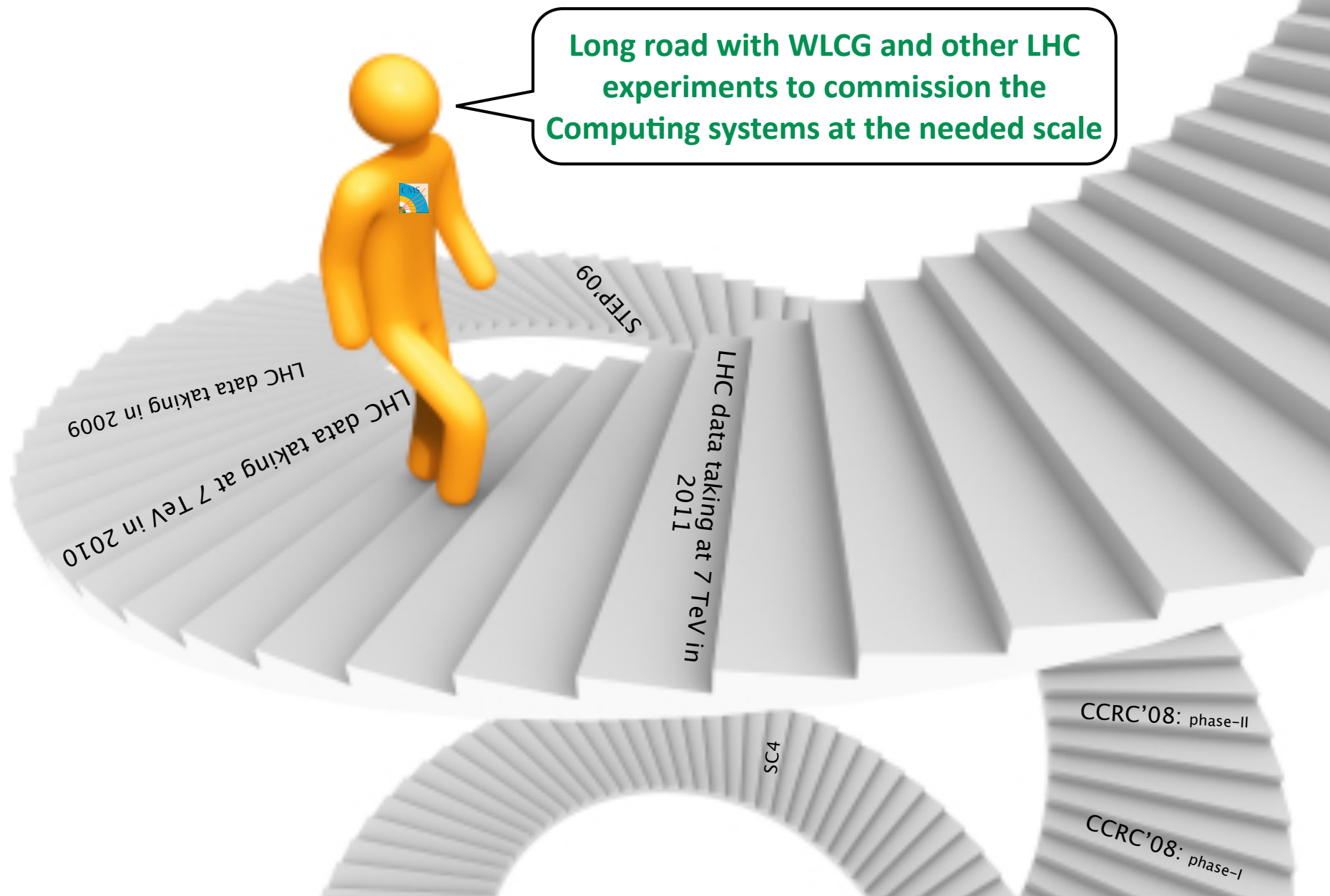
Stato e Prospettive del Calcolo Scientifico

16-18 February 2011   Laboratori Nazionali di Legnaro

CMS Computing and "challenges"

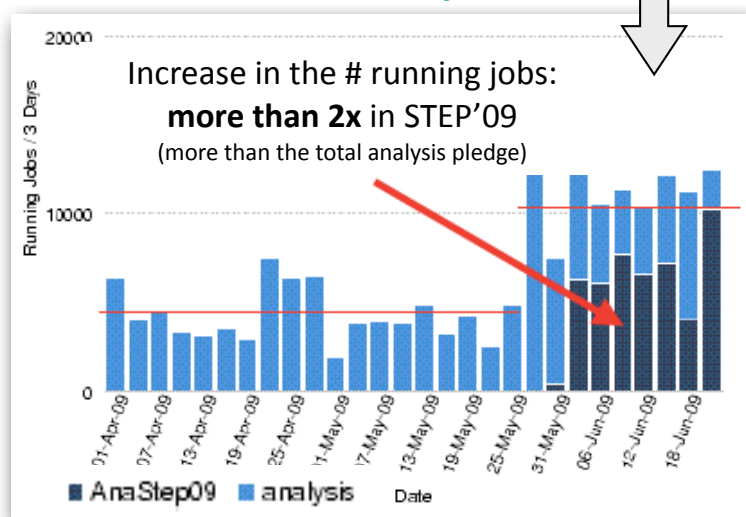# STEP'09: last challenge before LHC collisions data

A WLCG multi-VO exercise involving LHC exps + Tiers-0/1/2

- ✦ Before? CCRC'08 for CMS was a successful and fully integrated scale-test already
- ✦ STEP'09 focussed on several computing aspects while overlapping with other exps
  - T0: data recording to tape
  - T1: pre-staging & rolling processing
  - Distributed transfers: focus on latencies
  - Analysis: focus on ramping up the # used slots to meet the pledged resources at T2 level

Useful to get prepared for LHC p-p computing operations?

- ✦ As other challenges: <u>YES</u>! And quite advanced already. A couple of highlights:
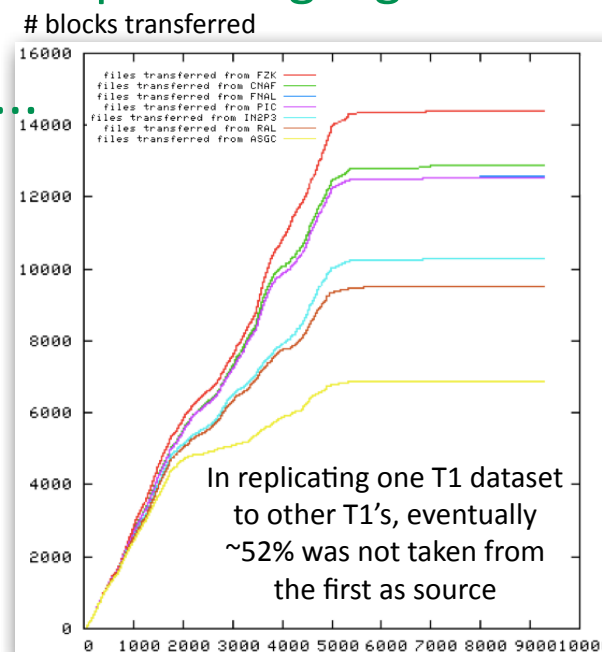
… one on Analysis…

… and one on Transfers…

Increase in the # running jobs:
more than 2x in STEP'09
(more than the total analysis pledge)

Load sharing in data replication patterns

- ✦ evidence of WAN transfers pattern optimization via files being routed from several already existing replicas instead of all from the original source

blocks transferred

In replicating one T1 dataset to other T1's, eventually ~52% was not taken from the first as source

Computing activities in LHC data taking

And now the <u>real</u> "challenge"… ⟹

During collision running CMS performed workflows and activities that were predicted in the Computing Model

- ✦ From a scale-tests playground…
  … to Computing operations on LHC p-p data first, and heavy-ions later on

The CMS Computing Tiers in WLCG performed the specified workflows

- ✦ Prompt processing at the **Tier-0**
- ✦ Output to the **CAF** for Prompt Feedback and Alignment/Calibration activities
- ✦ Transfer to **Tier-1**'s for storage, and distribution to **Tier-2**'s
- ✦ Prompt Skimming and Reprocessing at **Tier-1**'s
- ✦ Monte Carlo production and Analysis activities at **Tier-2**'s

Already now, modifications to the above were minor but not null

- ✦ Not disruptive on computing operations, and fully transparent to Physics

The load so far

Load on computing systems lower than expected in the planning (so far), so:

✦ No stress on resources

 - Total data volume smaller than foreseen

✦ Predicted activities performed more frequently

 - Reprocessing and analysis regularly exercised
 - Data subscribed to all T1s and more T2s

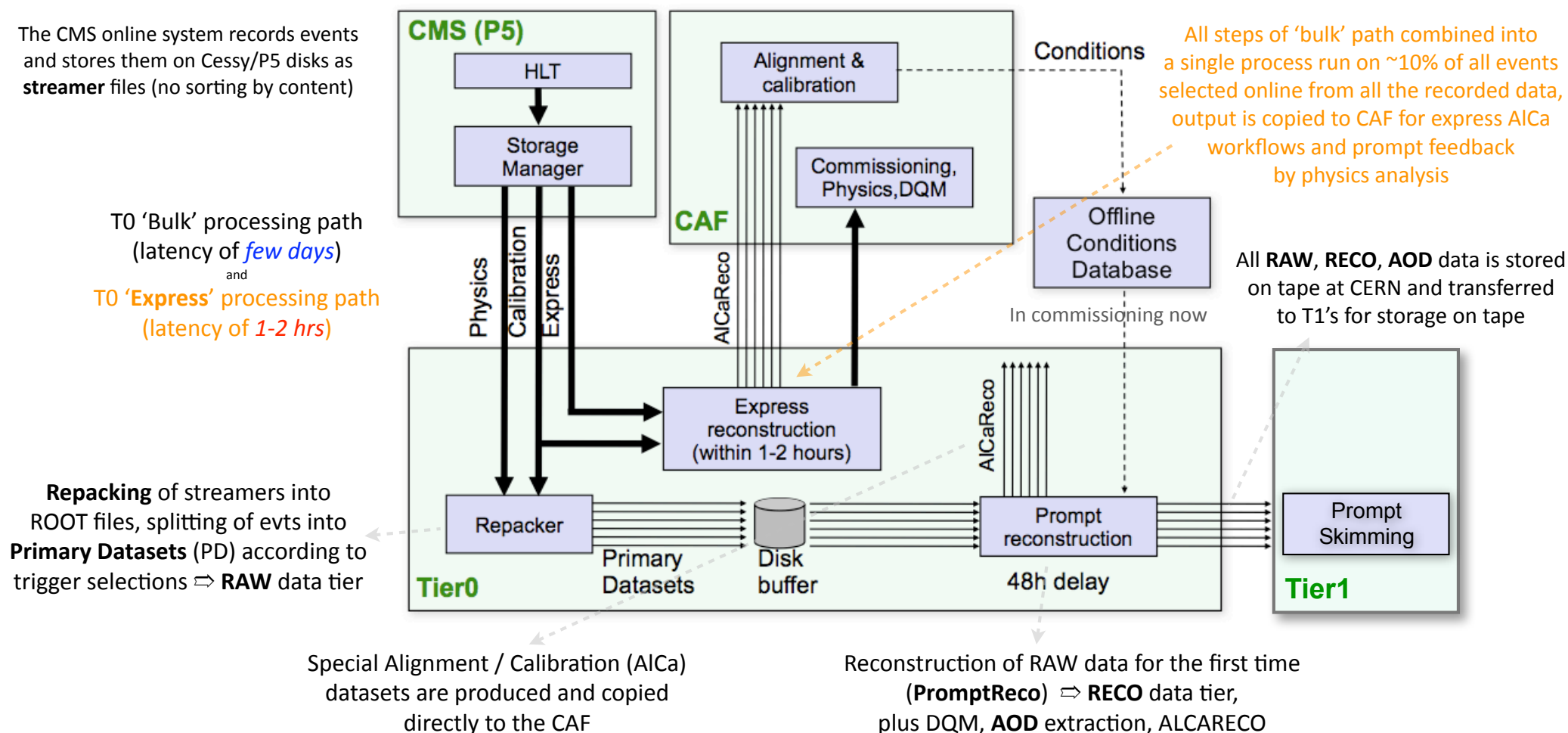✦ Also, room for changes in order to better use available resources
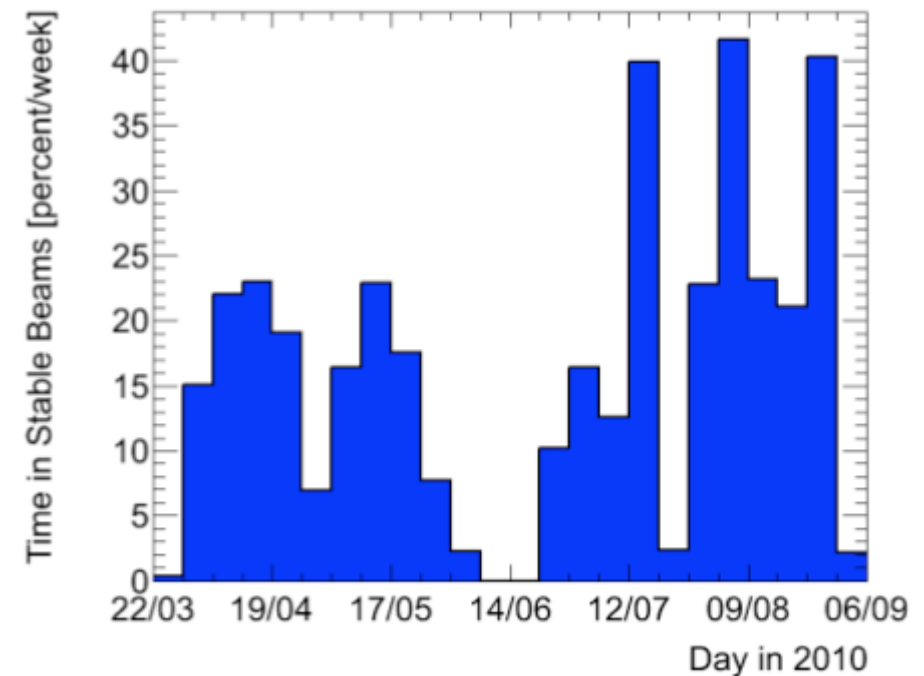
The extension of the 7 TeV run in 2011, the luminosity ramp and the increase in total data will be keeping Computing more and more busy in next months...

T0 data flows

Rolling workflows (fully automated)

✦ Express processing (at Tier-0 level)

✦ Prompt reconstruction (at Tier-0 level)

✦ Prompt skimming (at Tier-1 level - but scheduled by Tier-0 system)

The CMS online system records events and stores them on Cessy/P5 disks as **streamer** files (no sorting by content)

T0 'Bulk' processing path
(latency of *few days*)
and
T0 '**Express**' processing path
(latency of *1-2 hrs*)

Repacking of streamers into ROOT files, splitting of evts into **Primary Datasets** (PD) according to trigger selections ⇨ **RAW** data tier

All steps of 'bulk' path combined into a single process run on ~10% of all events selected online from all the recorded data, output is copied to CAF for express AlCa workflows and prompt feedback by physics analysis

All **RAW**, **RECO**, **AOD** data is stored on tape at CERN and transferred to T1's for storage on tape

CMS (P5)

HLT

Storage Manager

CAF

Alignment & calibration

Commissioning, Physics, DQM

Conditions

Offline Conditions Database

In commissioning now

Physics | Calibration | Express | AlCaReco

Express reconstruction (within 1-2 hours)

AlCaReco

Tier0

Repacker

Primary Datasets

Disk buffer

Prompt reconstruction

48h delay

Tier1

Prompt Skimming

Special Alignment / Calibration (AlCa) datasets are produced and copied directly to the CAF

Reconstruction of RAW data for the first time (**PromptReco**) ⇨ **RECO** data tier, plus DQM, **AOD** extraction, ALCARECO

T0 operations

Generally good operations experience at the Tier-0 so far

✦ Software and system reliability are good, level of automation is high

✦ Performances are remarkably good, the system works well beyond design specs
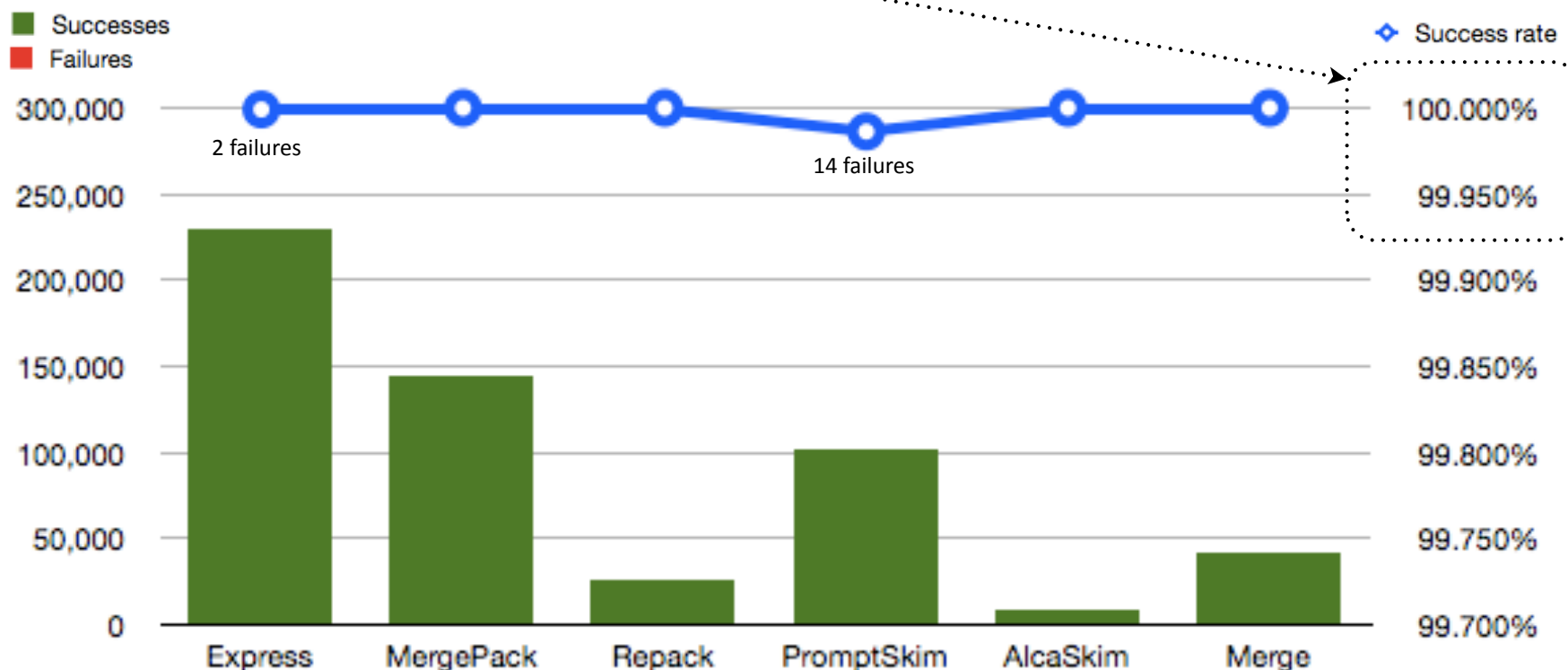
See above T0 job statistics since the end of September technical stop

✦ Repacker >4G evts (RAW), PromptReco >500M evts (RECO), and almost no failures

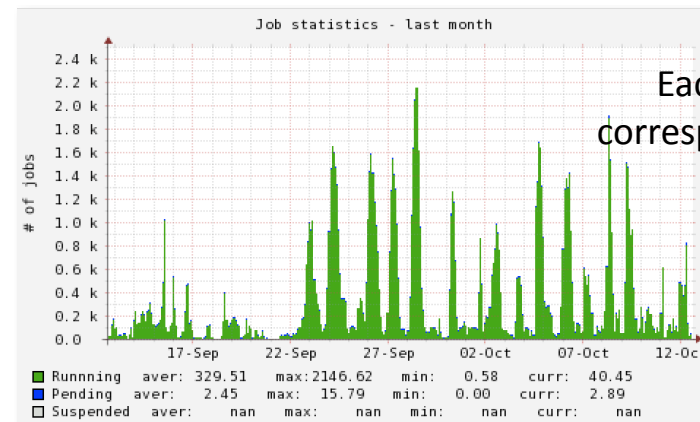# T0 and CAF performances

Processing latency for time-critical applications (e.g. express) are well within the design goals

✦ See e.g. latency of express FEVT to the CAF in Run2010B (until CHEP'10 times)
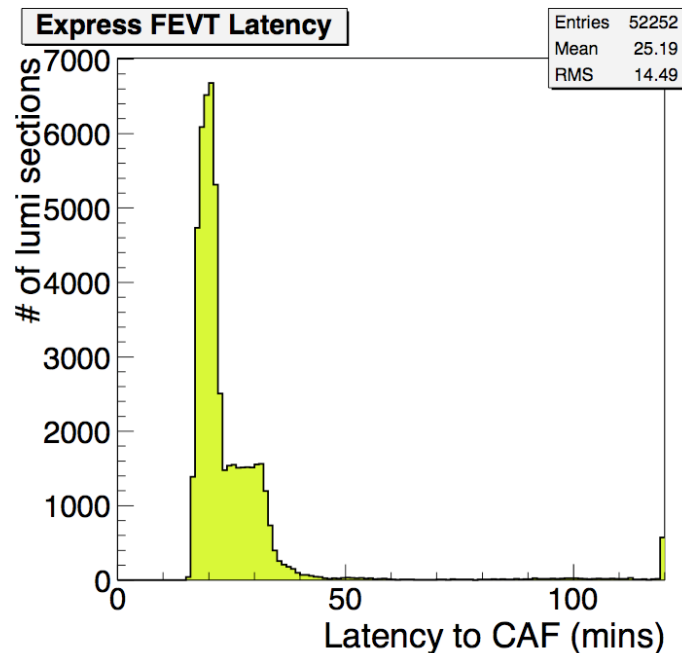
CERN LSF cmst0 slot usage over last ~1 month

Each spike basically corresponds to one LHC fill

Resource utilization is not stressing the system (yet)

✦ LHC live time and event complexity lower than we expect in the future

CAF also heavily used in burst

✦ Jobs start quickly in low latency queues

✦ >250 active users over last 6 months

Tiers readiness for operations

Site Readiness Status for CMS Tier-1 sites

T1

Site Readiness Status for CMS Tier-2 sites

T2
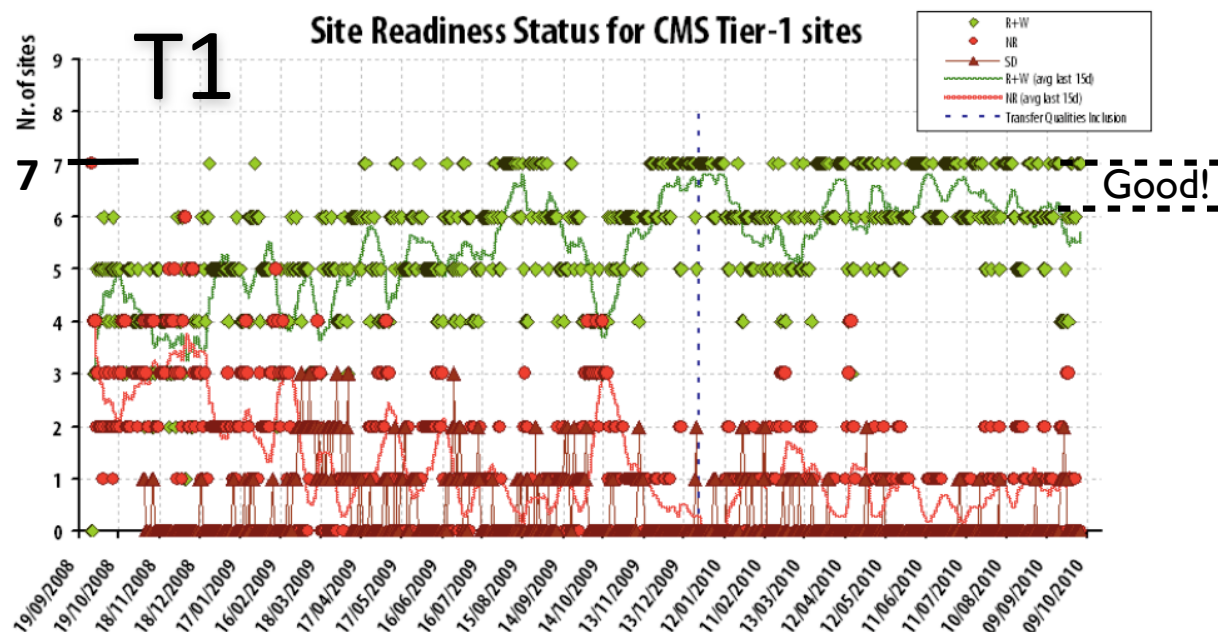
CMS defines a "site readiness" based on a boolean 'AND' of many tests

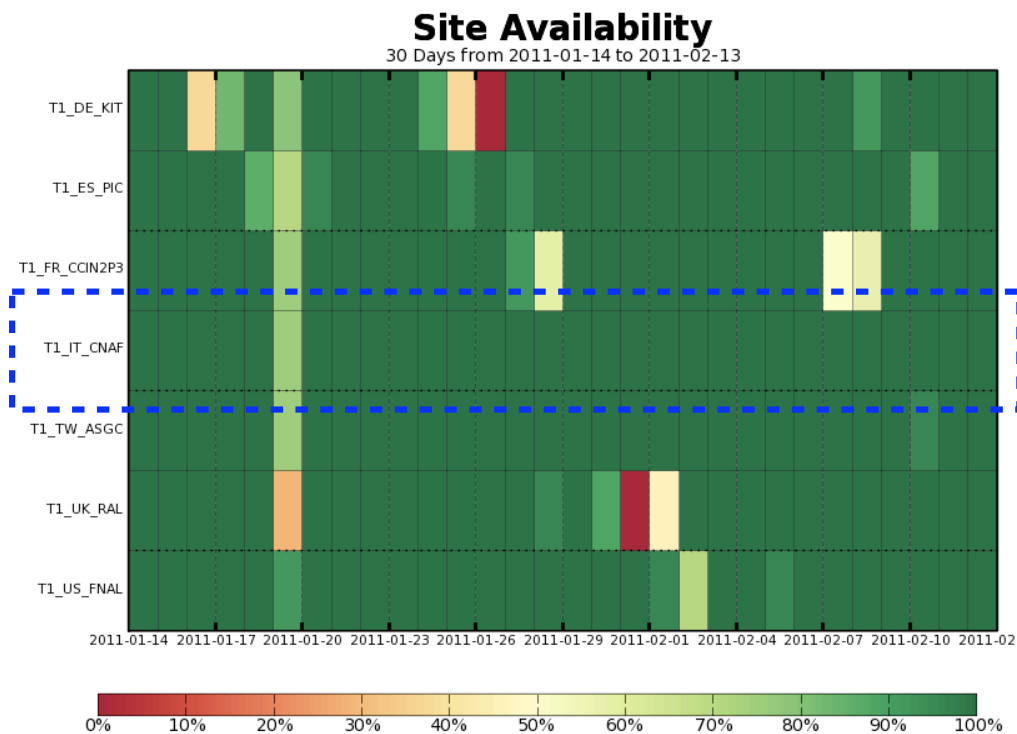✦ Easy to be OK on some

✦ Hard to be OK on all, and stably...

Readiness regularly overviewed and problems followed up

Good collaboration among CMS Computing Operations teams and WLCG Tiers
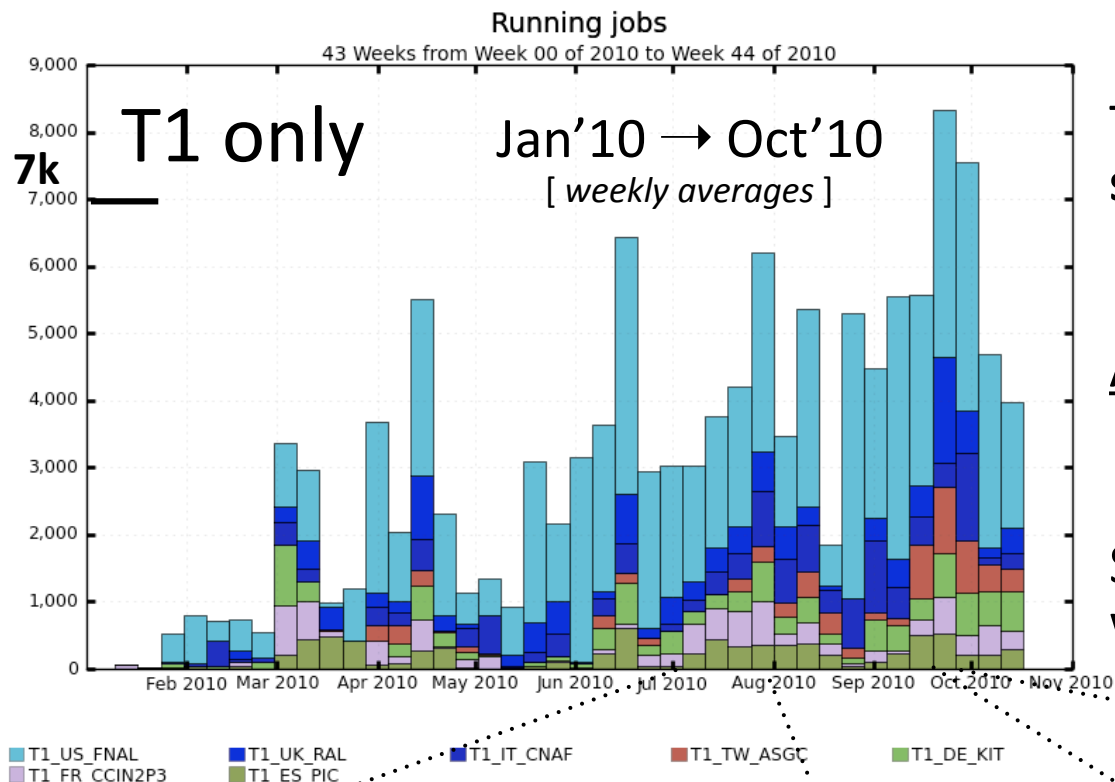
✦ Computing shifters play an important role

One example: CNAF T1

Site Availability
30 Days from 2011-01-14 to 2011-02-13
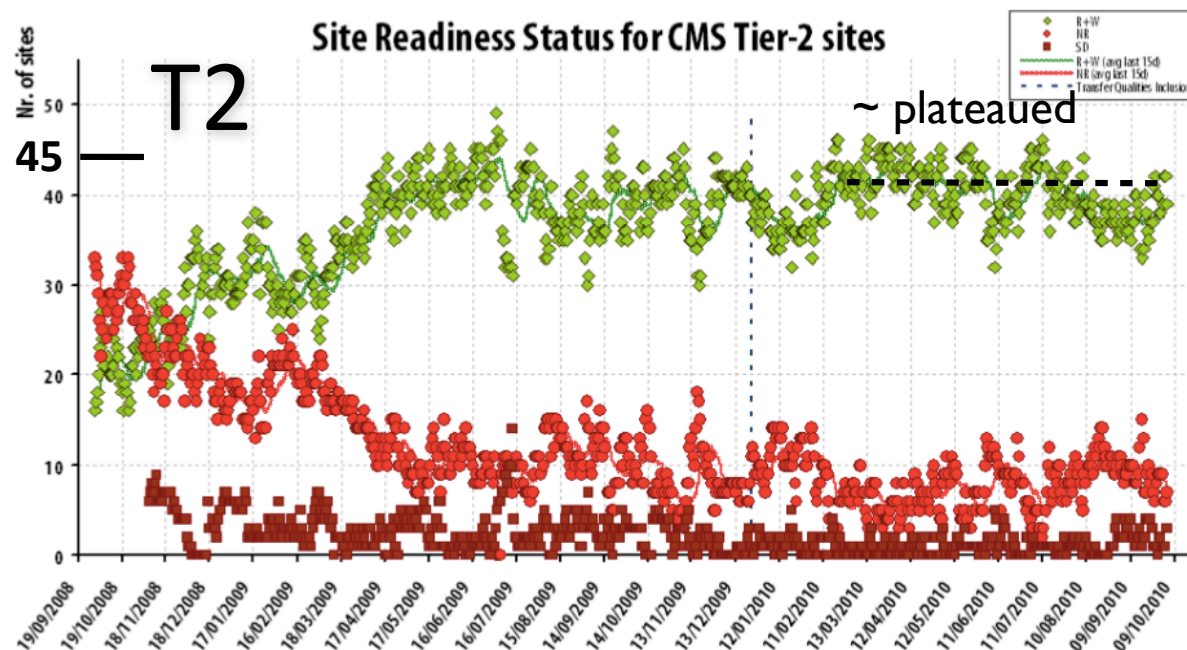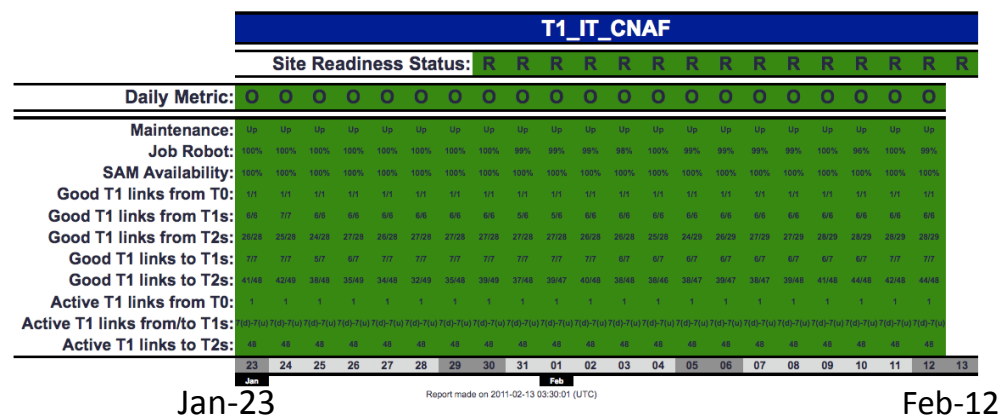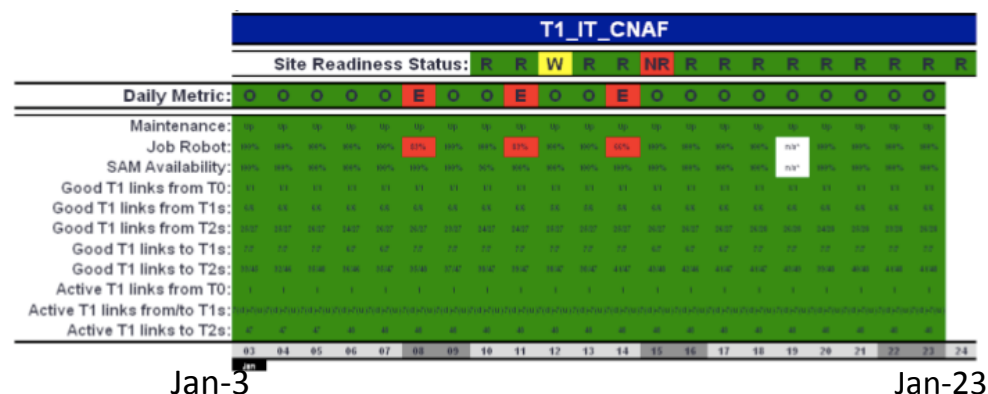
Hard to be green in CMS-SAM, much harded to be green in the CMS SiteReadiness

✦ All is hourly/daily overviewed for all WLCG Tiers supporting the CMS VO
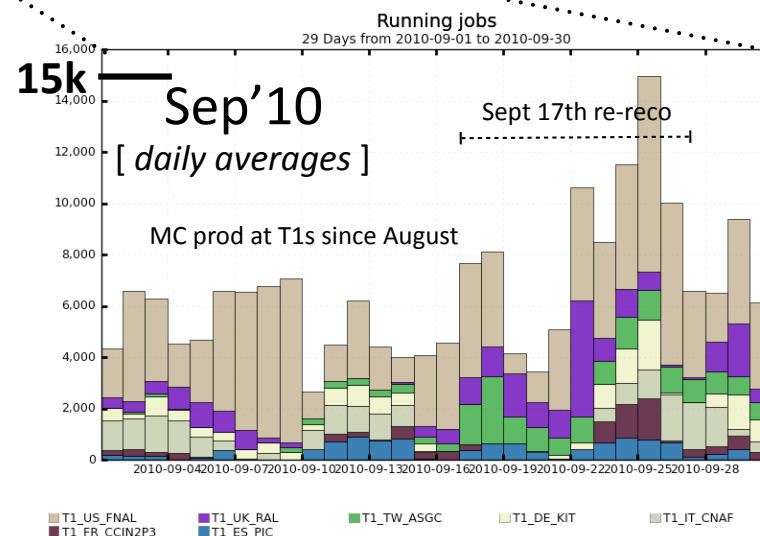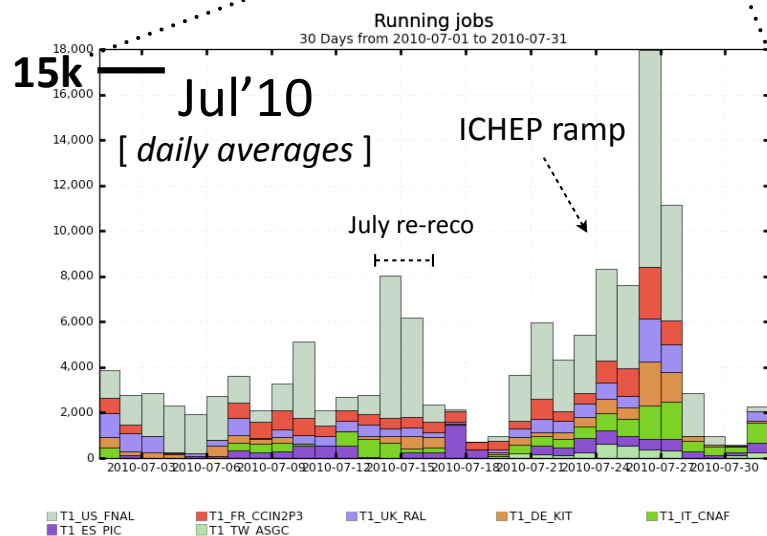
Processing at T1 sites

Tier-1 sites functioning well for **prompt skimming** and **reprocessing**

✦ High level activity driven by the 7 TeV re-reconstruction passes

<u>All</u> 7 T1s participating

✦ # jobs depending on custodial data
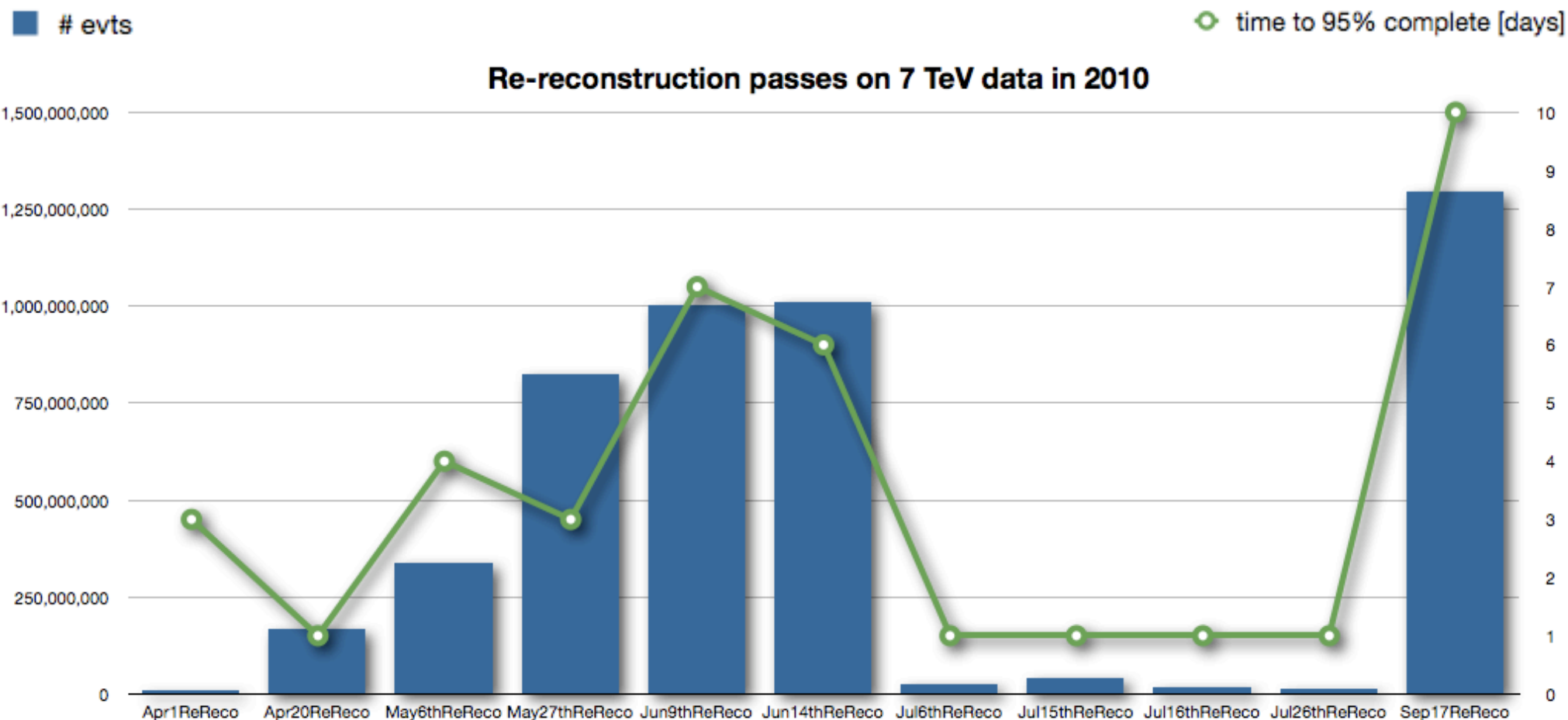
- We are not resources constrained (yet)

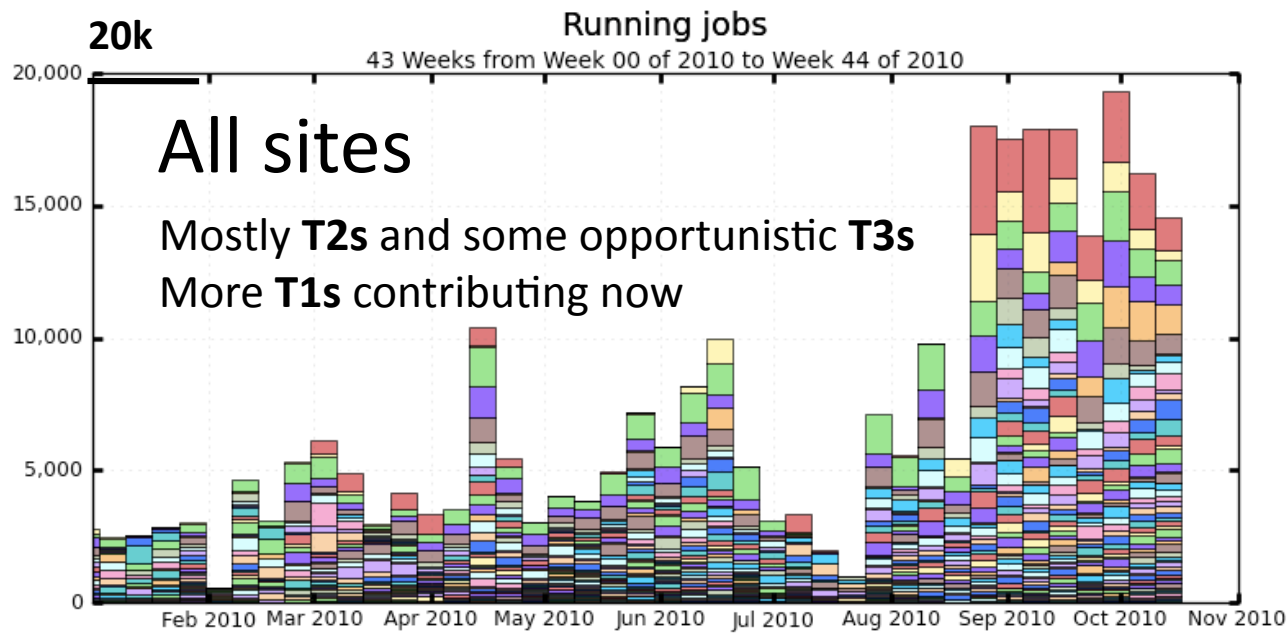Since August 2010, T1s are used together with T2s also for **MC production**

Re-reconstruction passes at T1s

Demonstrated ability to turn over the data passes needed for Summer conferences (e.g. ICHEP'10, HCP'10) in few days

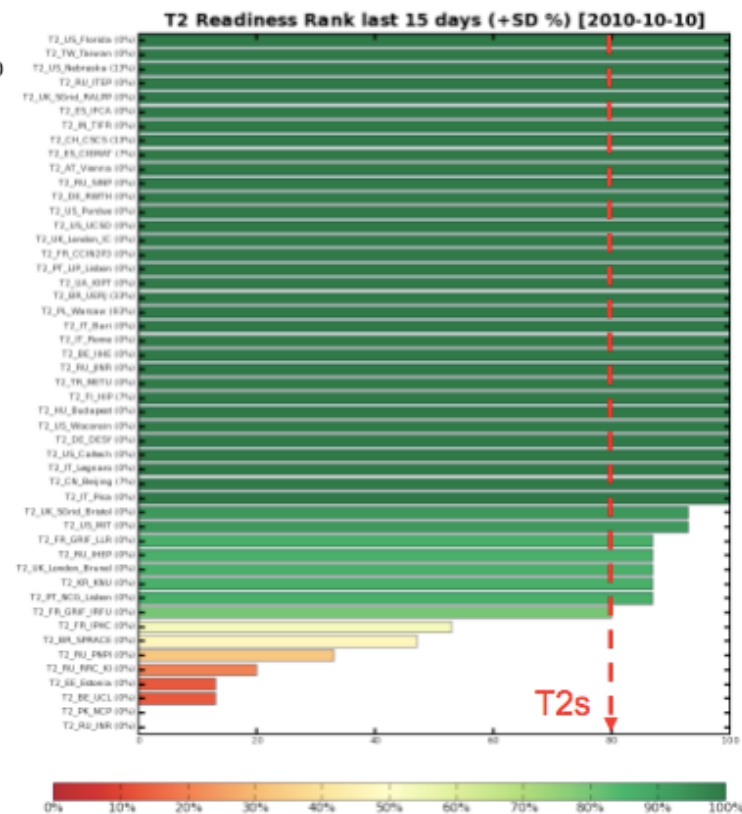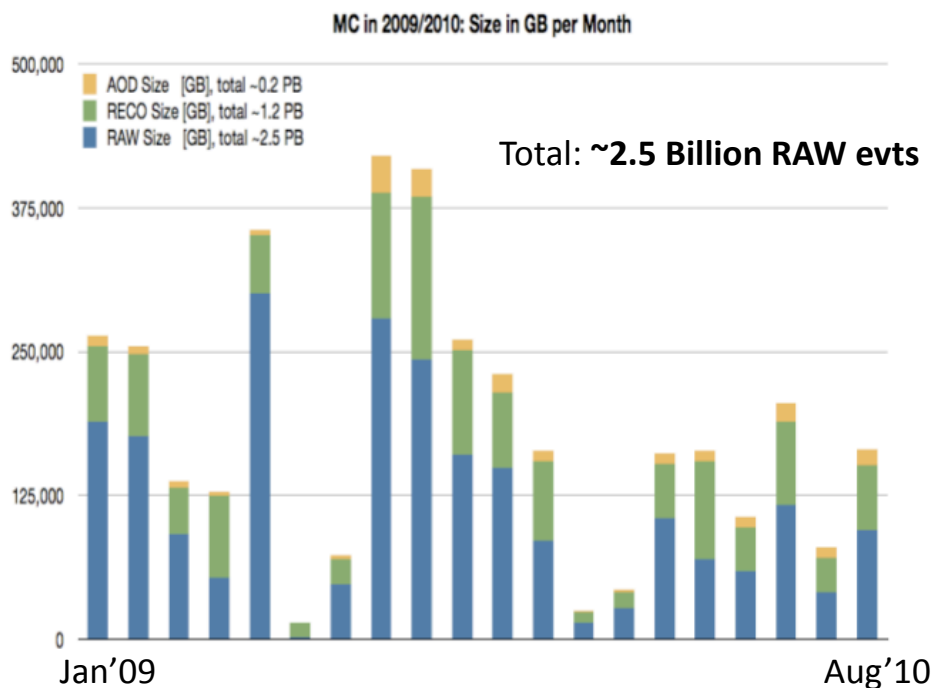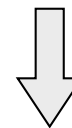(here shown for all pp running in 2010 before HI started)

MC production at CMS Tiers

Running jobs
43 Weeks from Week 00 of 2010 to Week 44 of 2010

20k

All sites

Mostly **T2s** and some opportunistic **T3s**
More **T1s** contributing now

MC in 2009/2010: Size in GB per Month

- AOD Size [GB], total ~0.2 PB
- RECO Size [GB], total ~1.2 PB
- RAW Size [GB], total ~2.5 PB

Total: **~2.5 Billion RAW evts**

Tiers usage for MC is driven by requests

✦ CMS uses Tiers passing the SiteReadiness metrics

T2 Readiness Rank last 15 days (+SD %) [2010-10-10]

T2s

CMS data transfers

NOTE: log scale

Average data transfer volume

PhEDEx

We improved by ad-hoc challenges of increasing
complexity and by computing commissioning activities

Legend: DC04, SC2, SC3, SC4, Load test, CSA06, CCRC08, Production, Debug
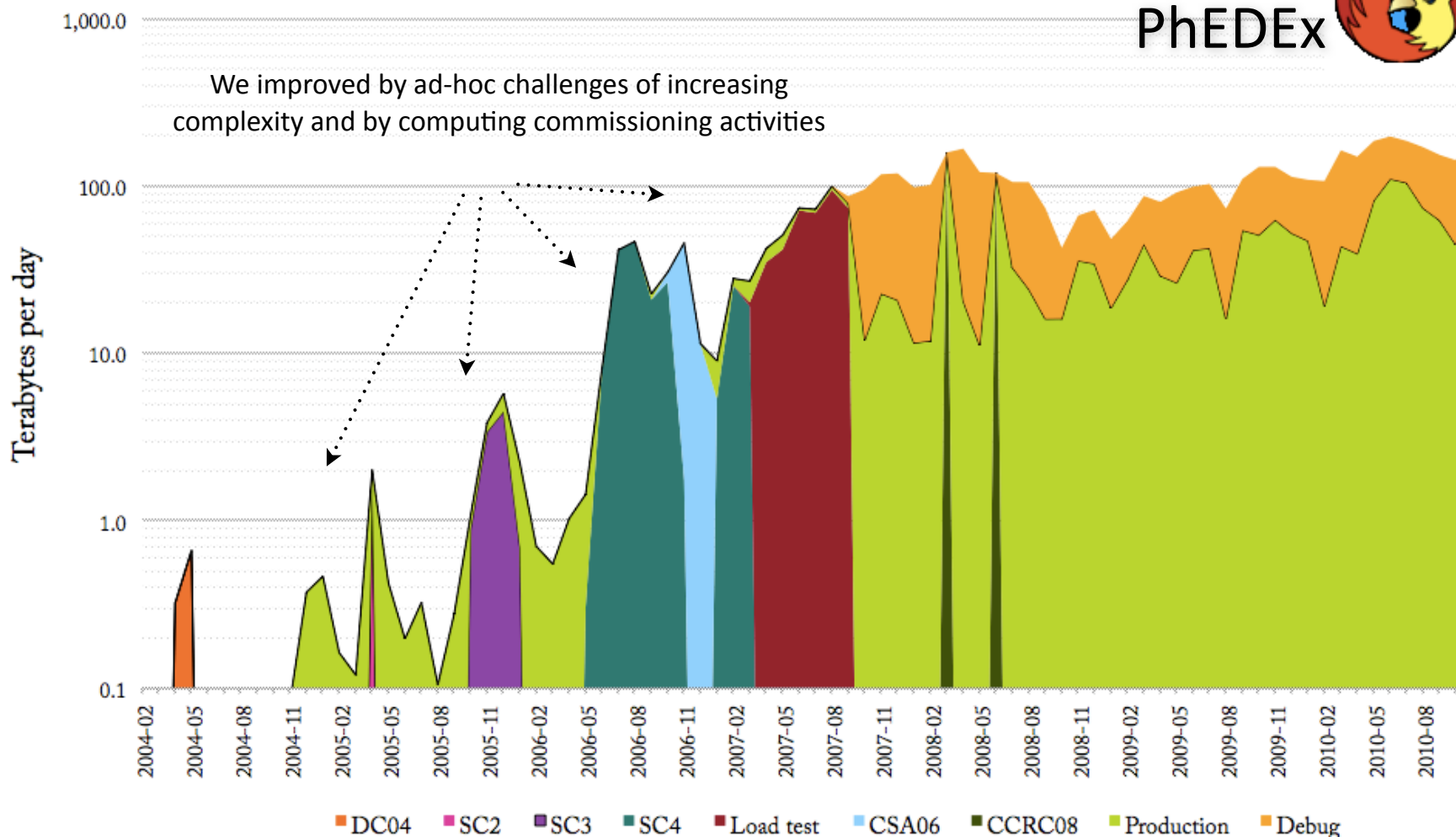
Massive commissioning, now in continuous production-mode of ops

✦ In October 2010, sustained >200 TB/day for the first time

Transferring 7 TeV CMS data among Tiers

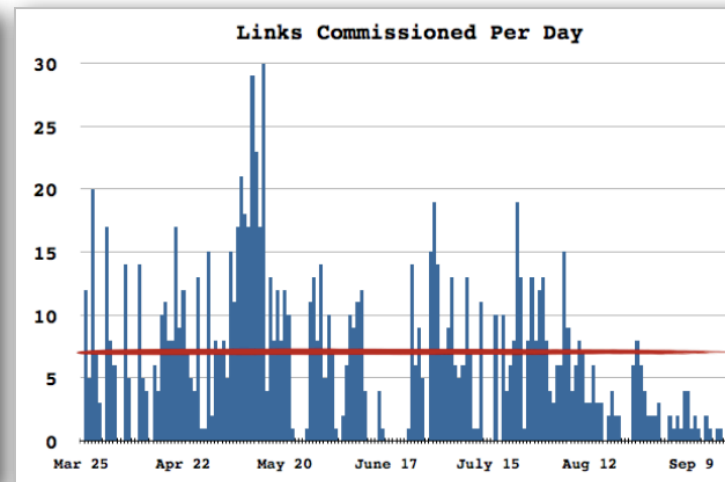## Data moved in transfer operations in 2010 flew in various routes:
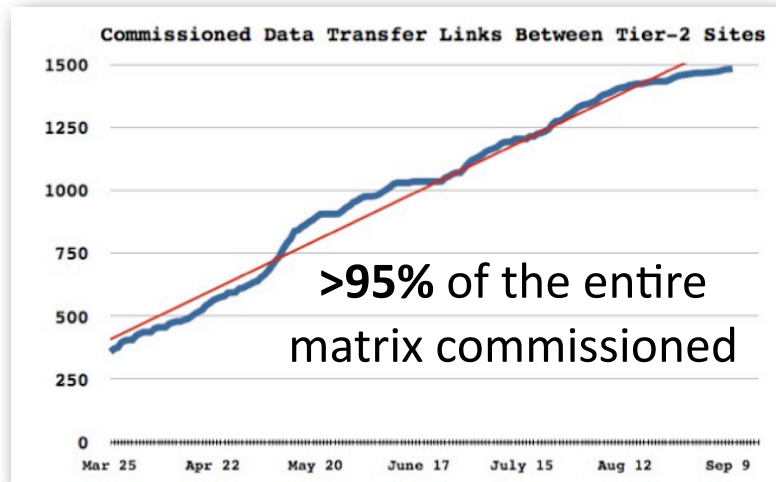
- ✦ **T0→T1** correlated with LHC fills. Also helped by PhEDEx rerouting from T1s
- ✦ **T1→T2** has come up, serving data to the CMS analysis layer (i.e. T2 level)
- ✦ **T2→T2** serious production traffic, after dedicated computing efforts in 2010
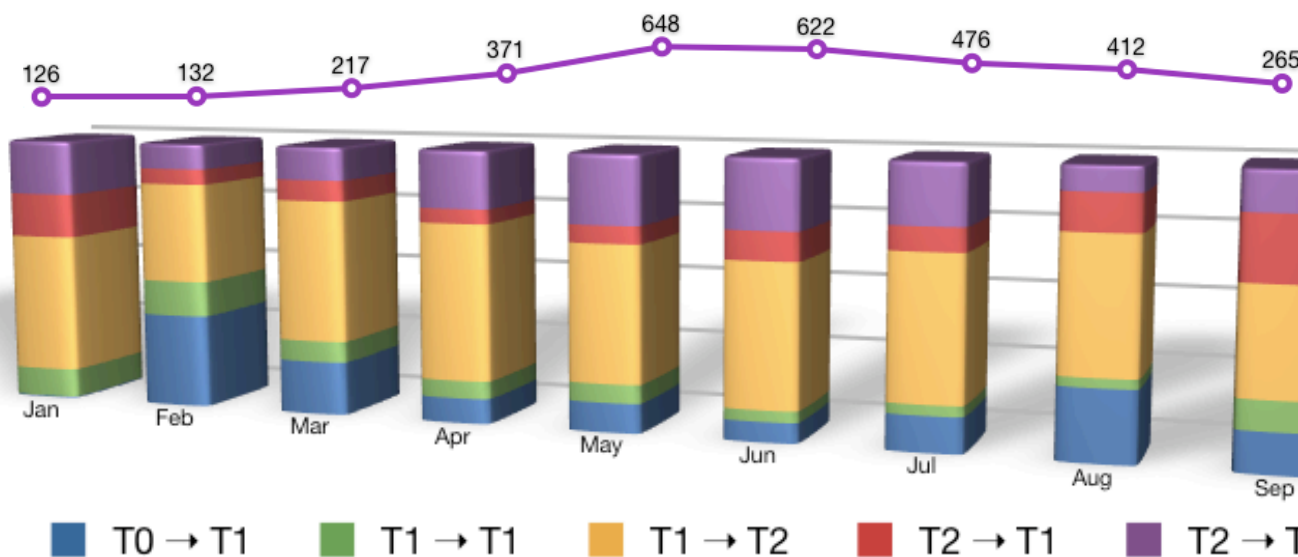

Production data volume transferred on different routes per month in 2010 so far
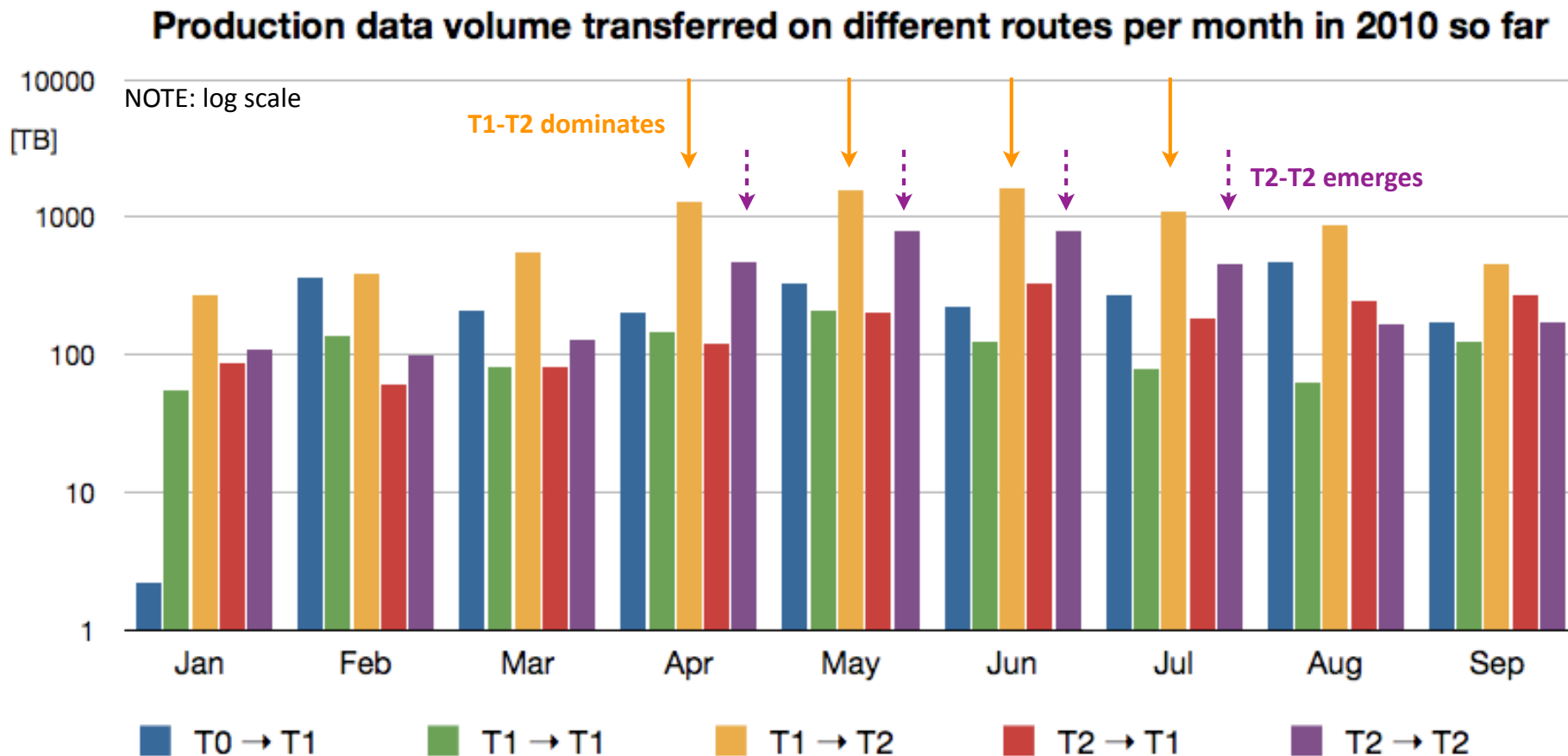
Commissioning the T2–T2 full mesh

Add more flexibility beyond that originally foreseen in the original Model

✦ Smooth transfer of physics results among all T2 associated to same analysis groups

✦ Fully support new fall-back stage-out tactics in CRAB, the CMS Remote Analysis Builder

>95% of the entire matrix commissioned

Up to 30 links commissioned per day, average is **~7 links/day** over last 6 months

126   132   217   371   648   622   476   412   265

 # T2-T2 links used for data transfers monthly
(not always the same ones)

T2-T2 traffic now a reality in CMS transfers ops and addresses Physics needs

Jan   Feb   Mar   Apr   May   Jun   Jul   Aug   Sep
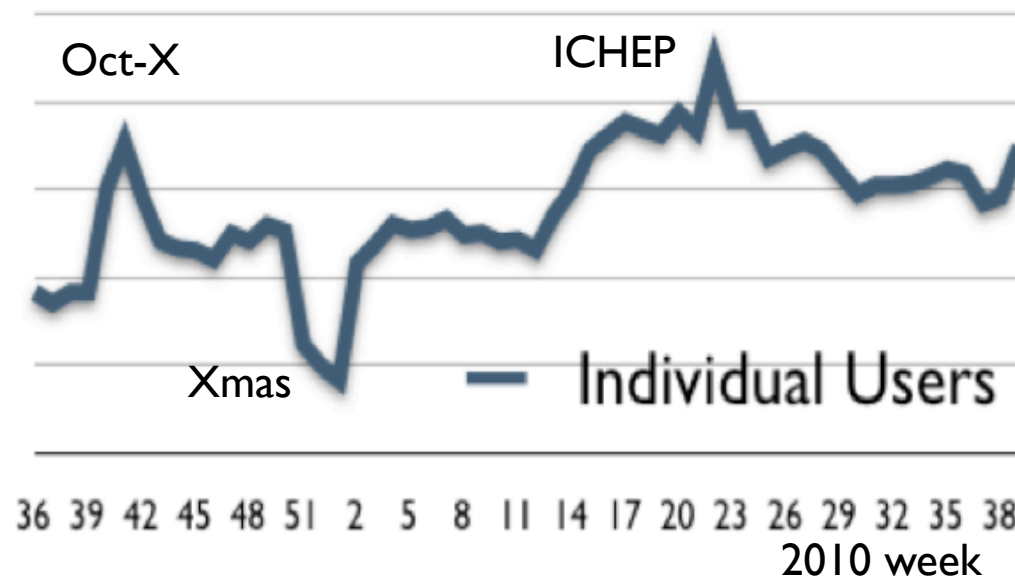
■ T0 → T1   ■ T1 → T1   ■ T1 → T2   ■ T2 → T1   ■ T2 → T2

Analysis at T2s

Number of people participating in CMS analysis is increasing
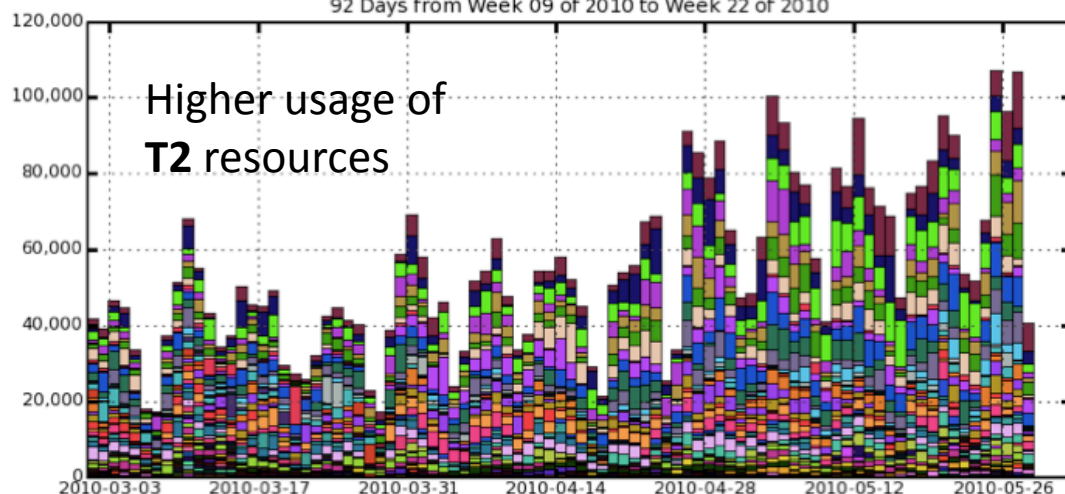
- ✦ ~**800** unique users/months
- ✦ Up to **100k jobs/day**

In total, >1 PB "central space" used at ~50 CMS T2 sites

- ✦ Used: ~50% of the total at T2 level
  - ~20% of sites still not extensively used

Largest access to collision data and the corresponding simulation

Summary

CMS Computing is working smoothly for LHC physics at 7 TeV so far
- ✦ Able to cope with the load in all sectors
  - (Rare) backlogs or (rare) service losses showed no impact whatsoever on physics

Not running in a resource constrained environment right now
- ✦ But soon we will, for sure.
  - The planning calls for interesting times soon (and for quite a long time..)

Stay tuned.