筑波大学
University of Tsukuba

ALICE Grid operations: last year and perspectives (+ some general remarks)

ALICE T1/T2 workshop
Tsukuba
5 March 2014
Latchezar Betev
Updated for the ALICE week 20/03/2014

On the T1/T2 workshop

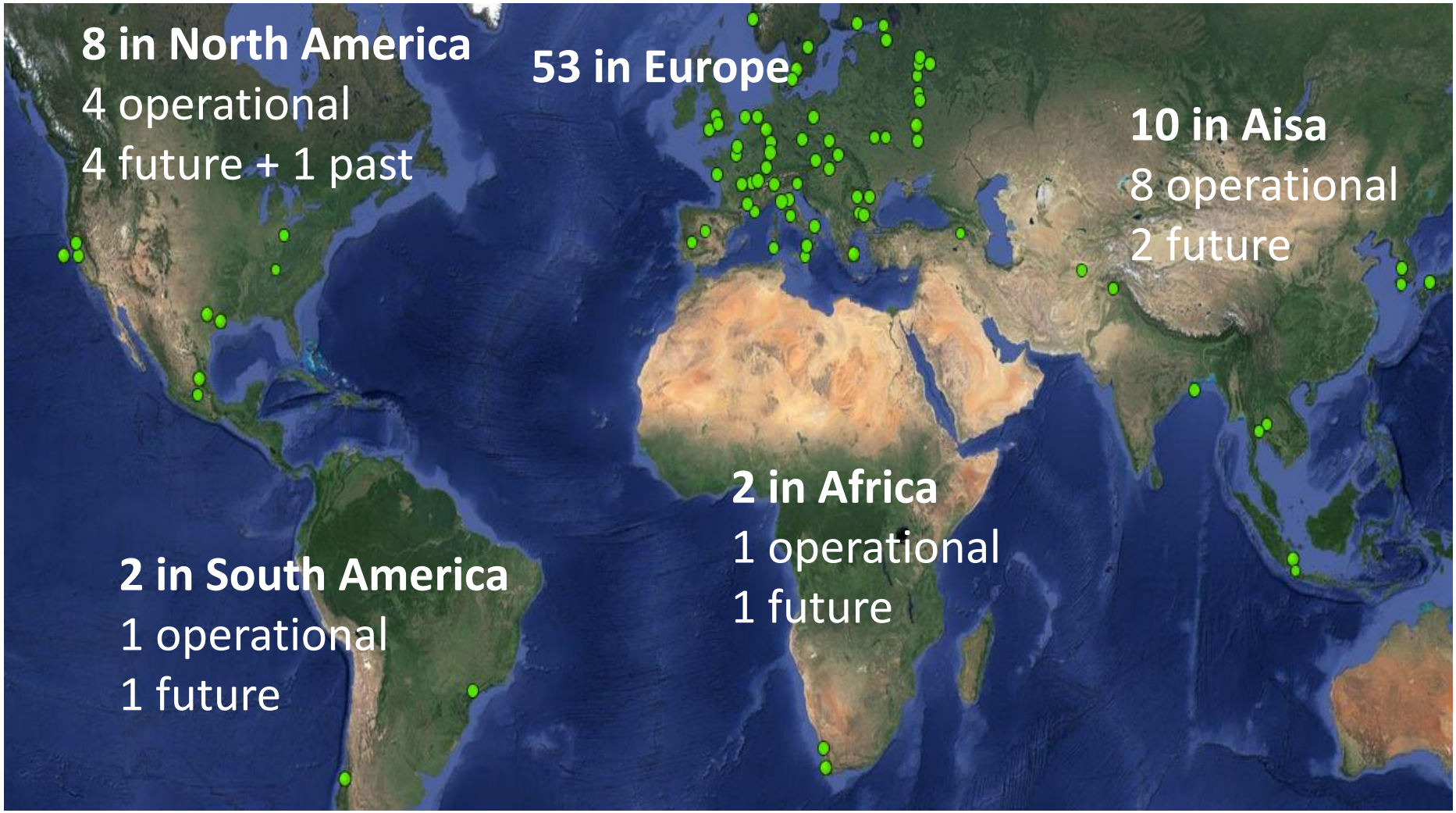
- Fourth workshop in this series
  - CERN – May 2009 (pre-data-taking) - ~45 participants
  - KIT – January 2012 – 47 participants counted
  - CCIN2P3 – June 2013 – 46 registered (45 counted)
  - Tsukuba* - March 2014 – ~45 participants (Grid sites)
- Main venue for discussions on ALICE-specific Grid operations, past and future
  - Site experts+Grid software developers
  - Throughout the year - communication by e-mail
  - …and tickets (the most de-humanizing system)

***-the only city without a computing centre for ALICE**

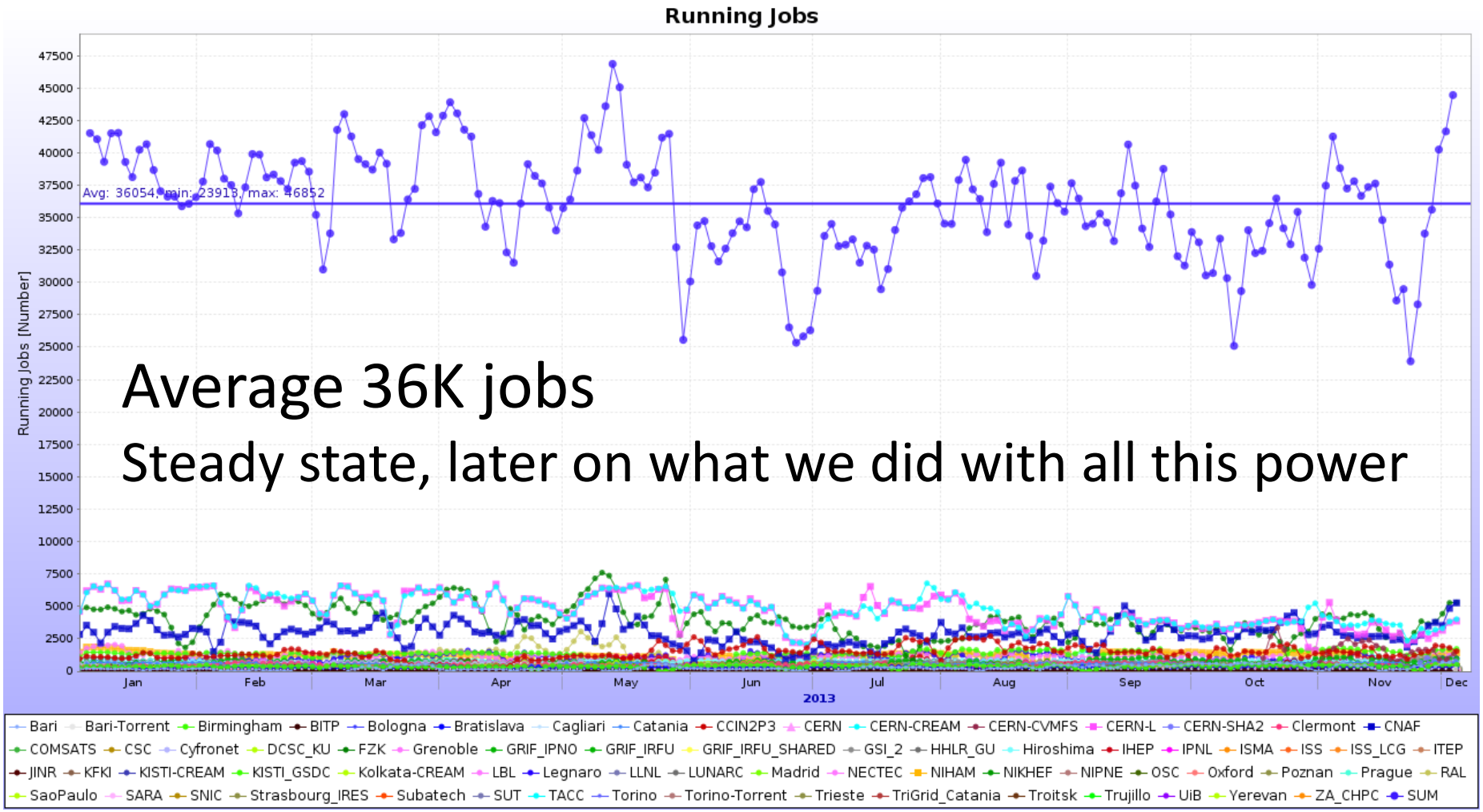
On the T1/T22 workshop (2)

筑波大学
University of Tsukuba

The ALICE Grid



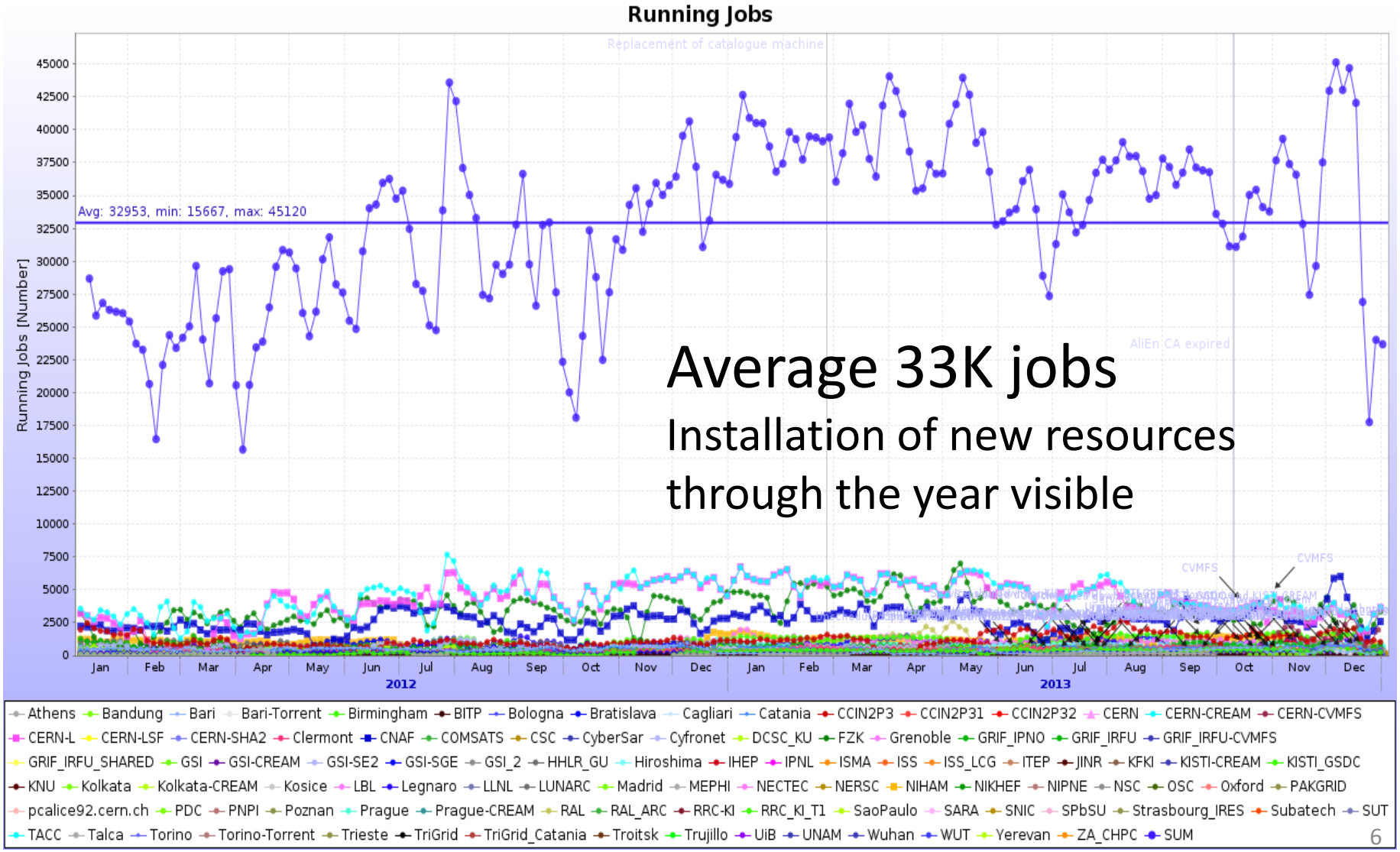
Grid job profile in 2013



Average 36K jobs

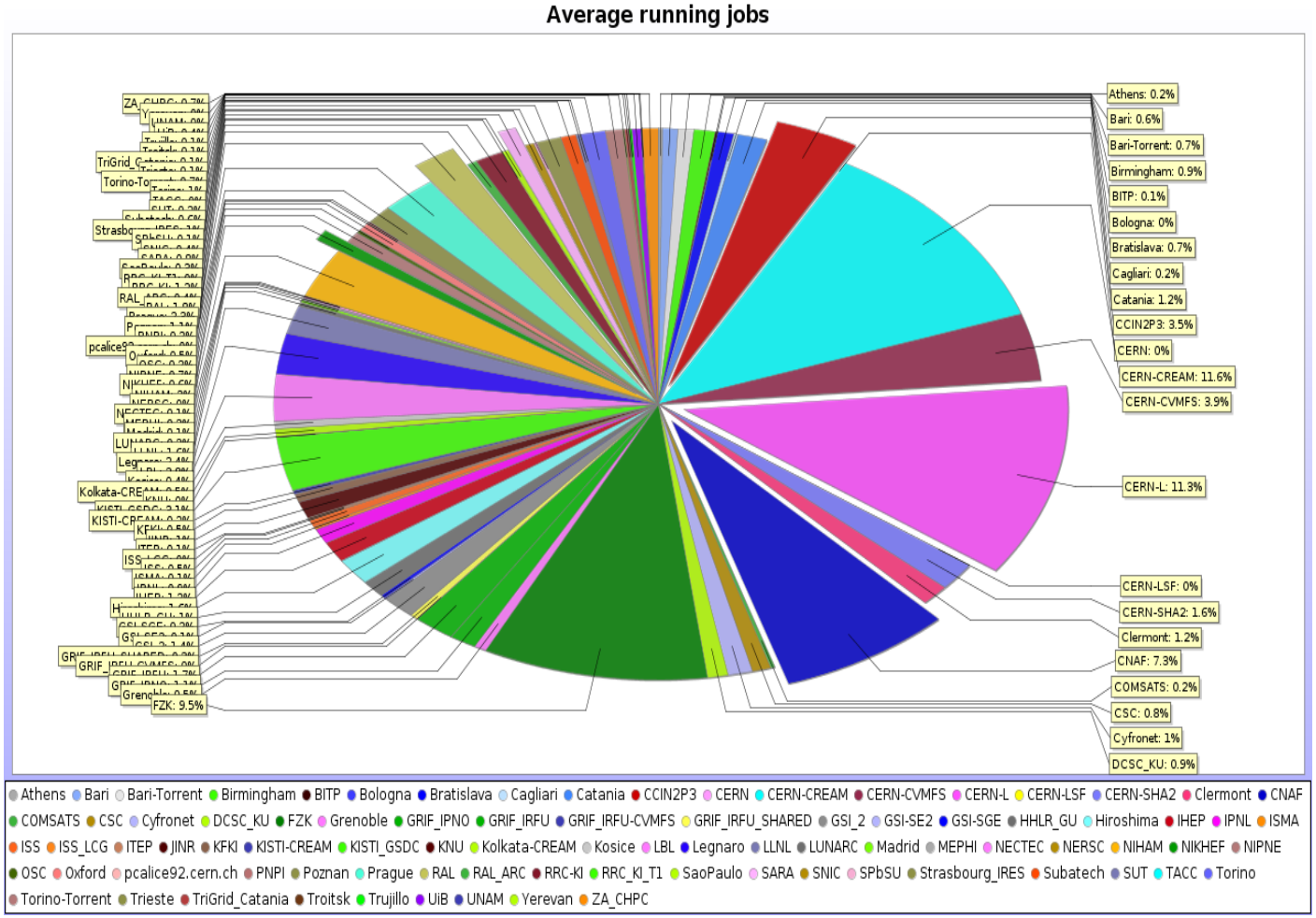
Steady state, later on what we did with all this power

The GRID job profile in 2012

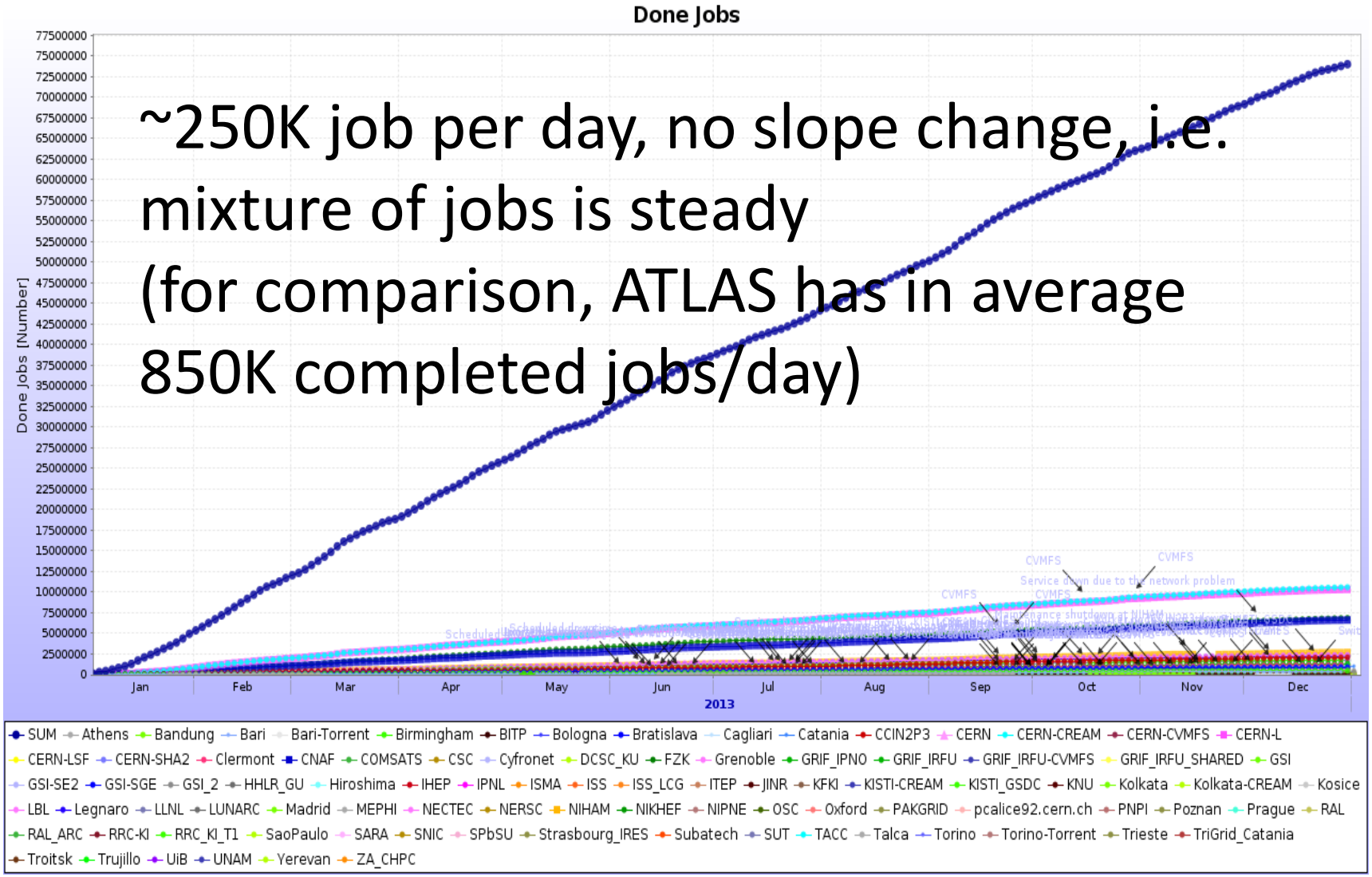


Resources delivery distribution

The remarkable 50/50 share T1/T2 is still alive and well



Done jobs



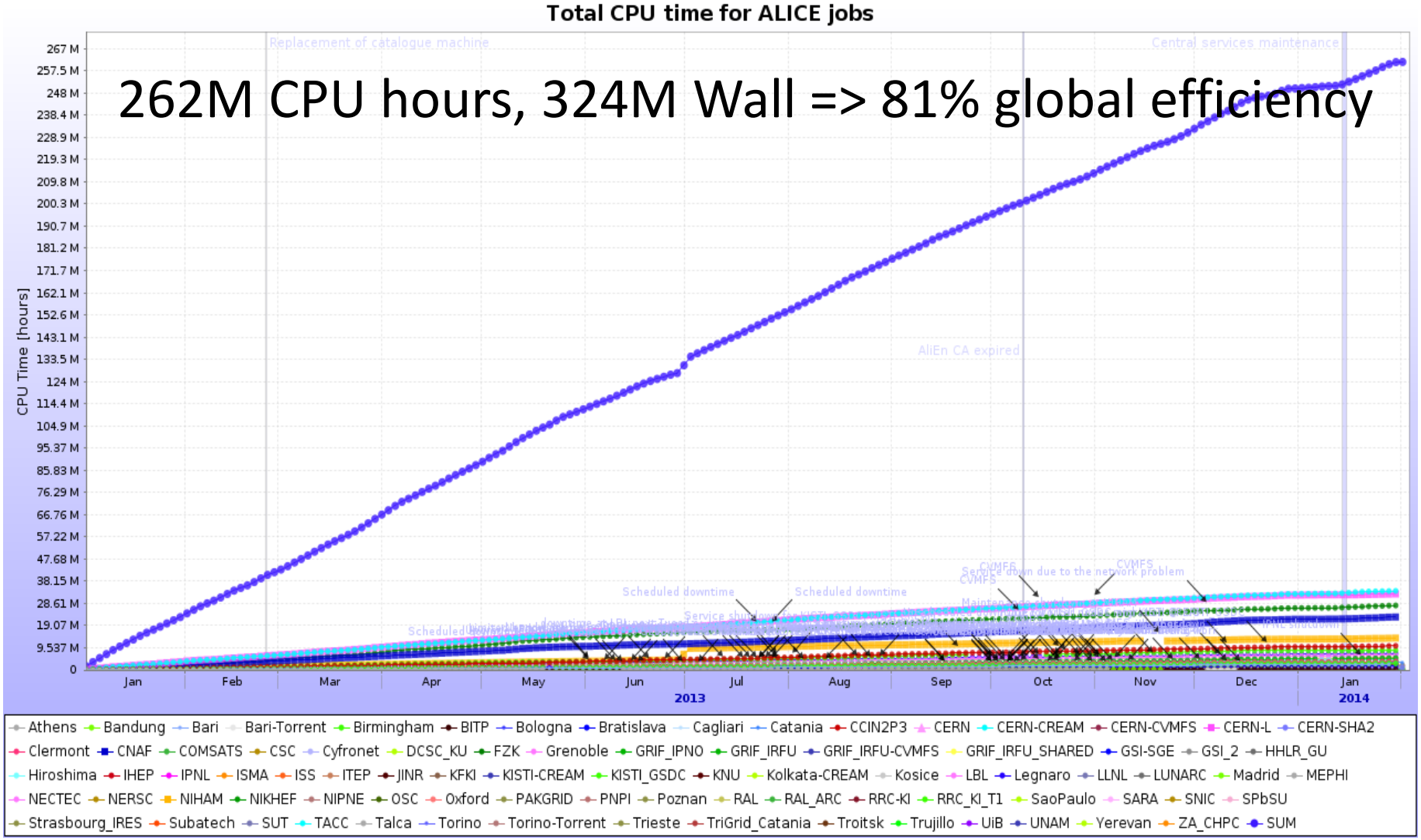
~250K job per day, no slope change, i.e. mixture of jobs is steady

(for comparison, ATLAS has in average 850K completed jobs/day)

Job mixture

69% MC, 8% RAW, 11% LEGO, 12% individual, 447 individual users

CPU and Wall time



Year 2013 in brief

- ‘Flat’ CPU and storage resources
  - However we had 8% more job slots in average in 2013 than in (second half) of 2012
  - Mostly due to Asian (KISTI) sites increasing their CPU capacity, some additional capacity installed at few European sites
  - Storage capacity has increased by 5%
- Stable performance of the Grid in general
  - The productions and analysis unaffected by upgrade stops at many sites

Production cycles MC

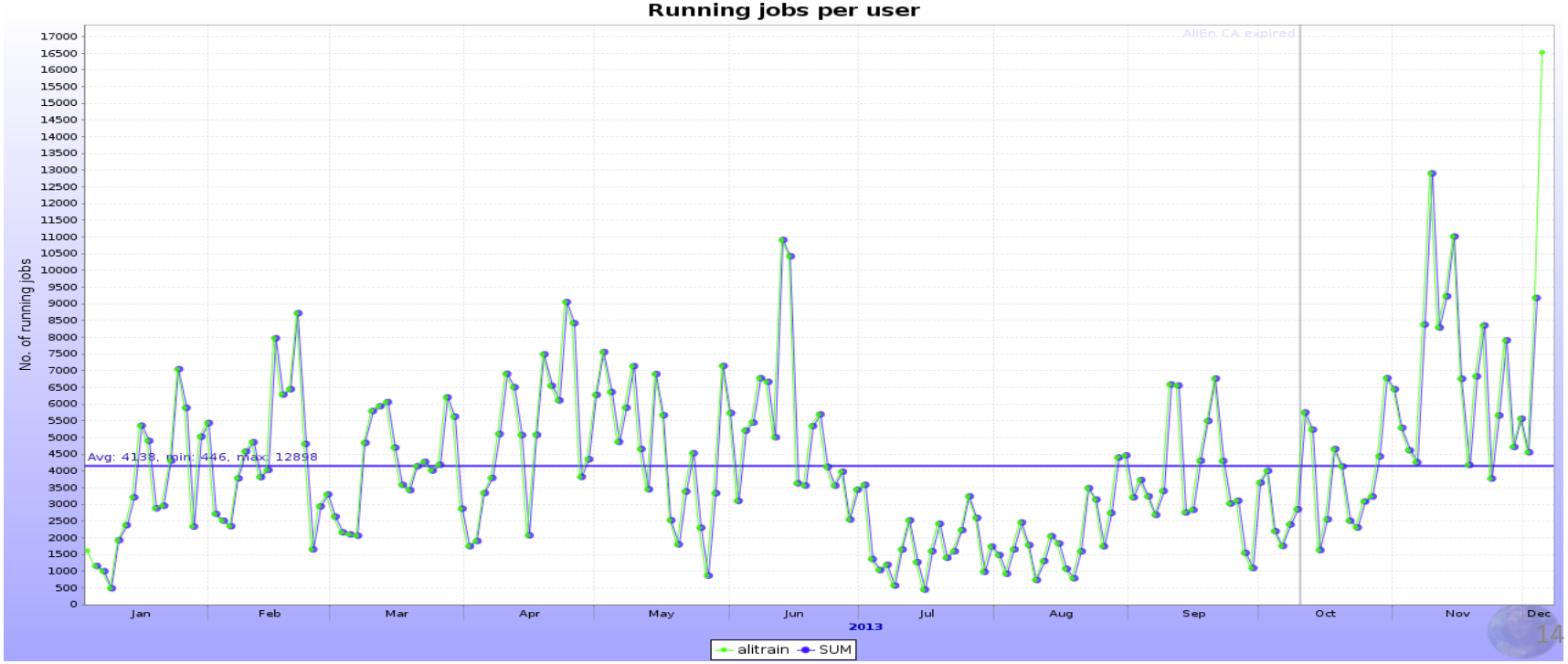
- 93 production cycles from beginning of the calendar year
  - For comparison – 123 cycles in 2012; 639,597,409 events
- 767,433,329 events
  - All types – p+p, p+A, A+A
  - Anchored to all data-taking years – from 2010 to 2013

AOD re-filtering

- 46 cycles
  - From MC and RAW, from 2010 to 2013
- Most of the RAW data cycles have been ‘refiltered’
- Same for the main MC cycles
- This method is fast and reduces the need for RAW data reprocessing

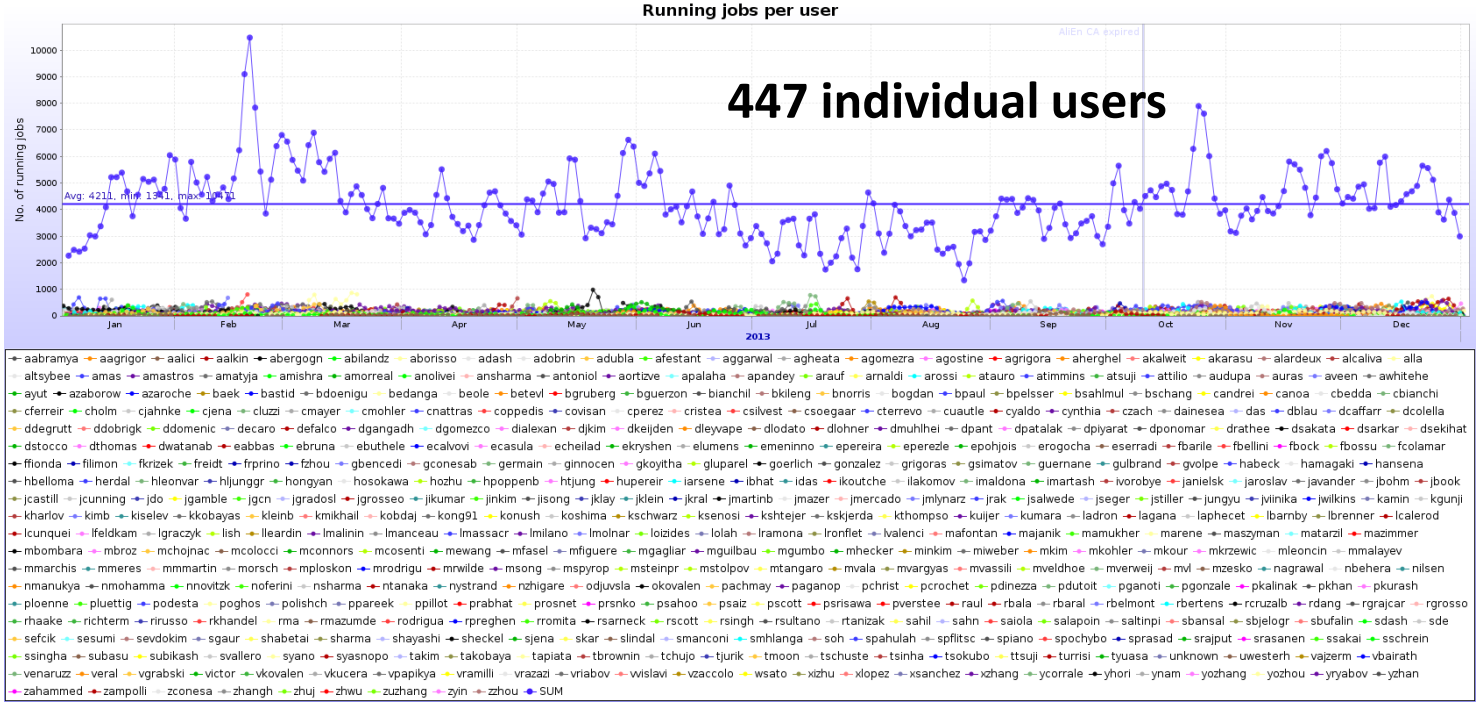
Analysis Train

- More active in specific periods, increase in the past months (QM)
- 4100 jobs, 11% of Grid resources
- 75 train sets for the 8 ALICE PWGs
- 1400 train departure/arrivals in 49 weeks => 28 trains per week…

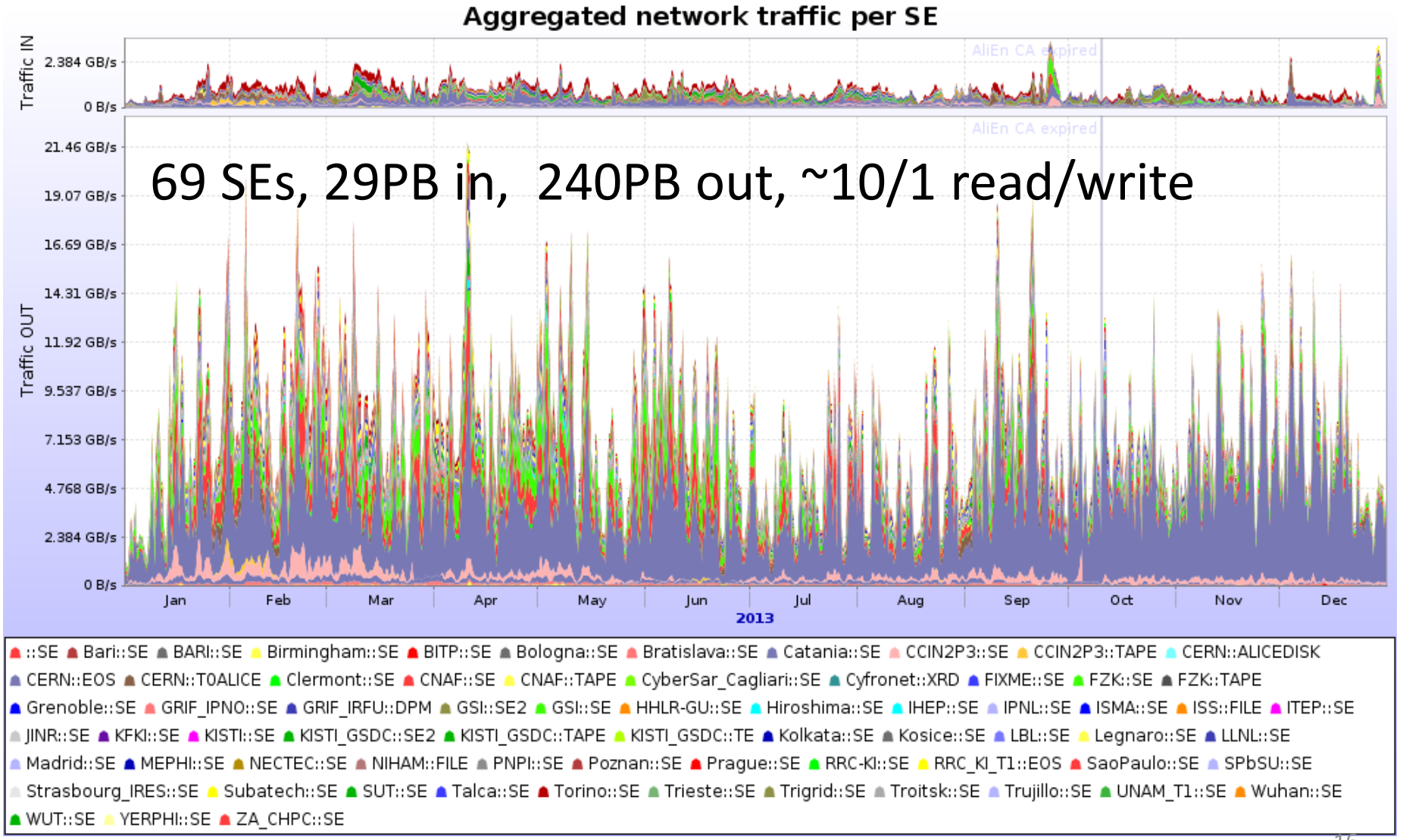


Summary on resources utilization

- The above activities use up to 88% if the total resources made available to ALICE
- The remaining 12% is individual user analysis



Access to data (disk SEs)



Data access 2

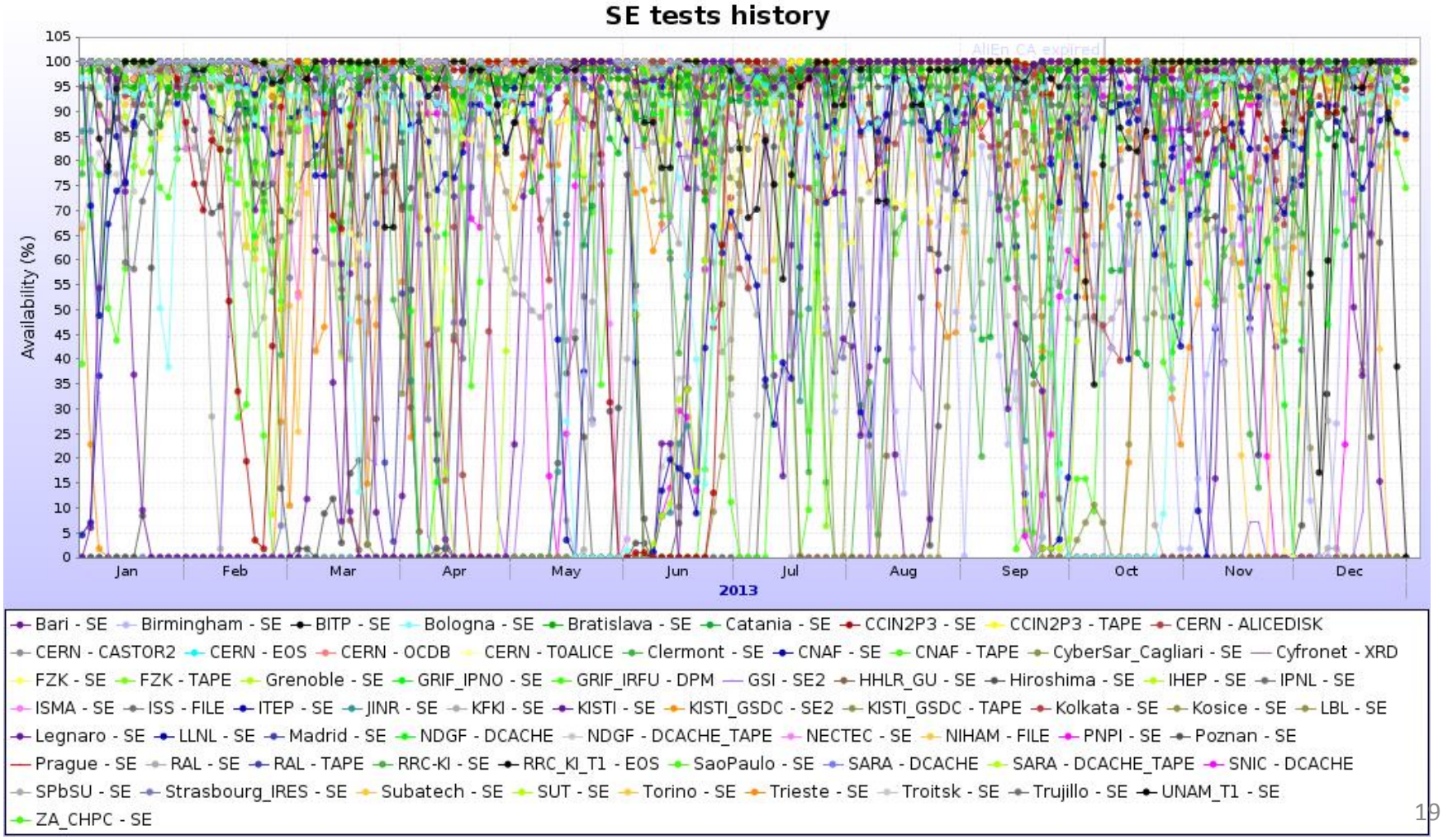
- 99% of the data read are input (ESDs/AODs) to analysis jobs, the remaining 1% are configurations and macros
- From LEGO train statistics, ~93% of the data is read locally
  - The job is sent to the data
- The 7% is file cannot be accessed locally (either server not returning it or file missing)
  - In all such cases, the file is read remotely
  - Or the job has waited for too long and is allowed to run anywhere to complete the train (last train jobs)
- Eliminating some of the remote access (not all possible) will increase the global efficiency by few percent
  - This is not a showstopper at all, especially with better network

Storage availability

- More important question – availability of storage
- ALICE computing model – 2 replicas => if SE is down, we lose efficiency and may overload the remaining SE
  - The CPU resources must access data remotely, otherwise there will be not enough to satisfy the demand
- In the future, we may be forced to go to one replica
  - Cannot be done for popular data

Storage availability (2)

- Average SE availability in the last year: 86%

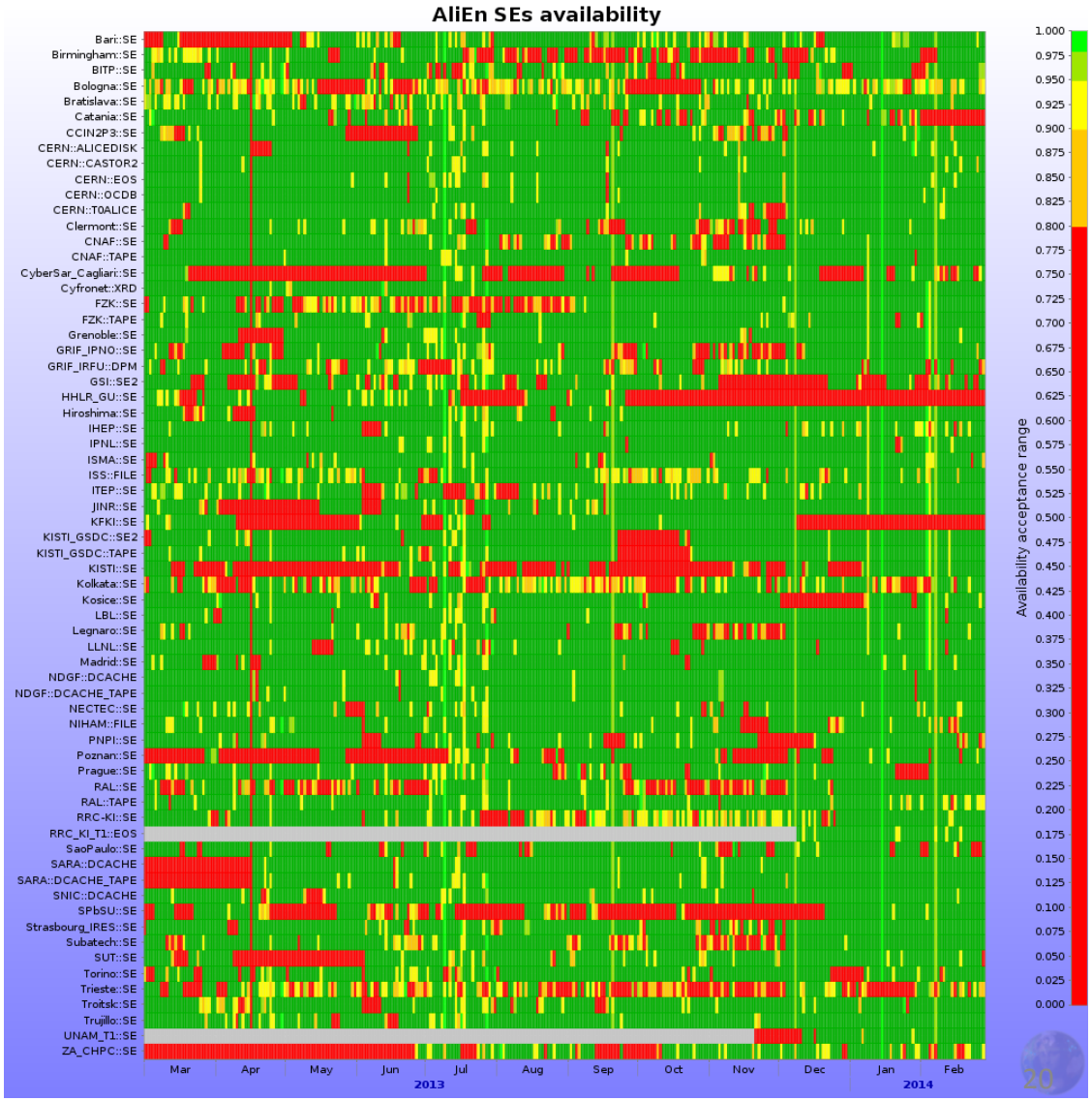


Alternative representation

Green – good
Red – bad
Yellow/orange - bad

Some SEs do have extended ‘repair’ times…

Oscillating ‘availa-bility’ is also well visible



Storage availability

- Extensive ‘repair’, upgrade times, down times
  - Tolerated due to the existing second replica for all files
- Troubles with underlying FS
  - Some SEs – xrootd gateways over GPFS/Lustre/Other
  - Fast file access and multiple open files are is not always supported well
  - Issues with tuning of xrootd parameters
  - Limited number of gateways (traffic routing), can hurt the site performance
  - xrootd works best over a simple Linux FS
- How to solve this – **storage session on Thursday**
- Goal for SE availability >95%

Other services

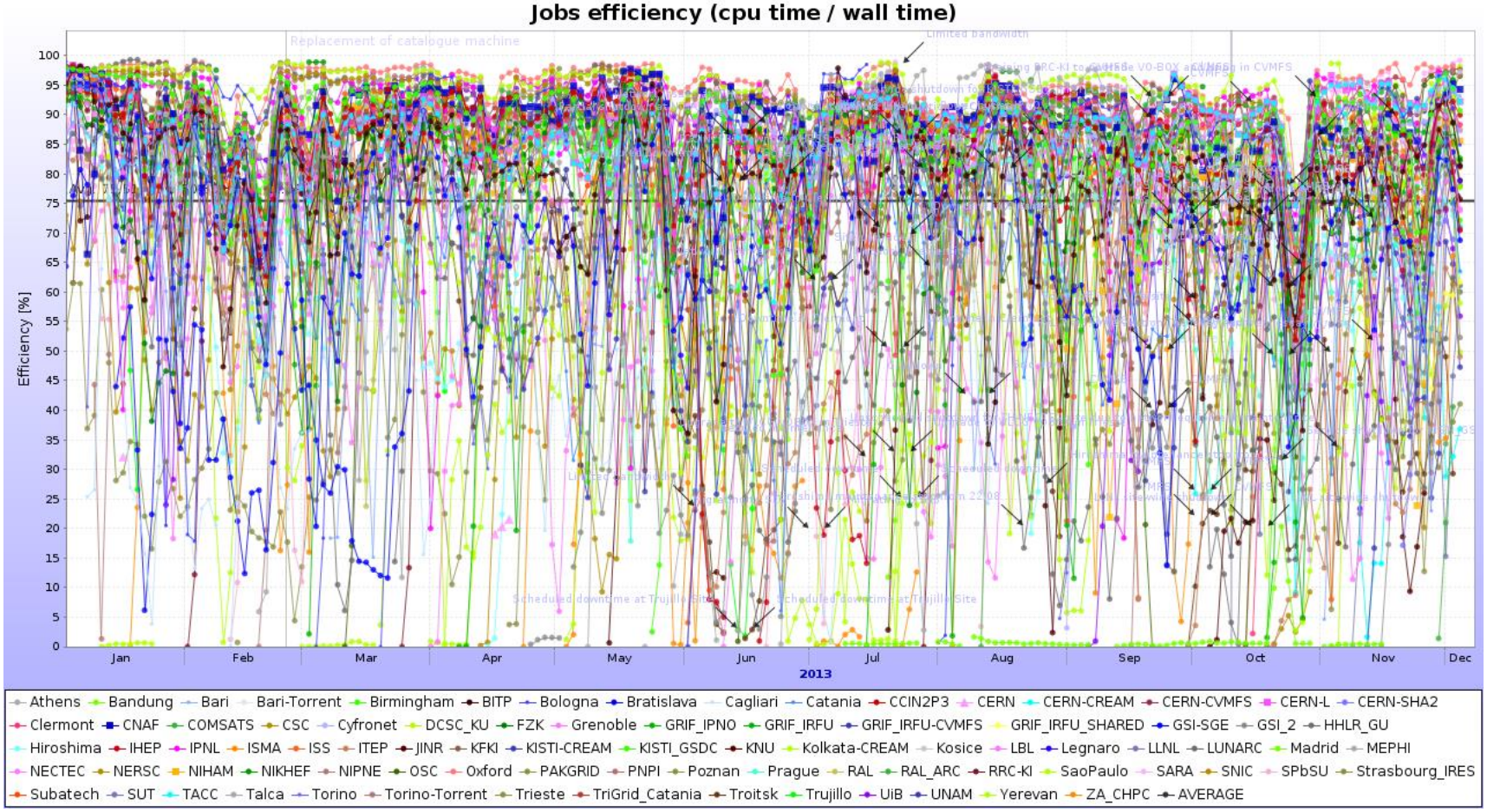
- Nothing special to report
  - Services are mature and stable
  - Operators are well aware of what is to be done and where
  - Ample monitoring is available for every service (more on this will be reported throughout the workshop)
  - Personal reminders needed from time to time
  - Several services updates were done in 2013…

Major upgrade events

- xrootd version – smooth, but not yet done at all sites
  - Purpose – more stable server performance, rehearsal for xrootd v.4 (IPv6-compliant)
- EMI2/3 (including new VO-box) – mostly smooth – more in Maarten's talk
- SL(C)5 (or equivalent) ->SL(C)6 (or equivalent) – smooth, for some reason not yet complete…
- Torrent->CVMFS – quite smooth, two (small) sites remaining

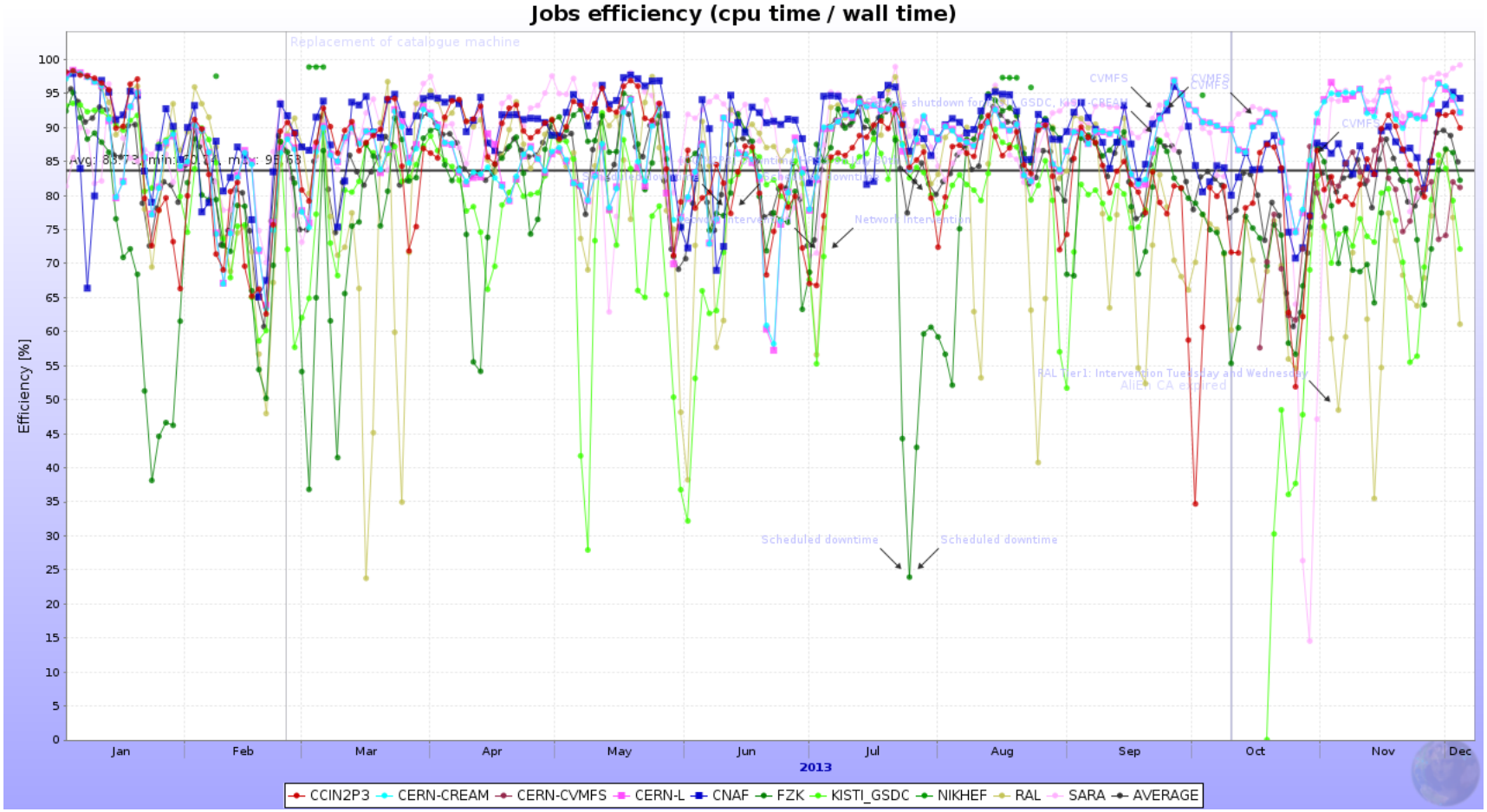
The Efficiency

Average of all sites: 75% (unweighted)



Closer look – T0/T1s

Average – 85% (unweighted)



Summary on efficiency

- Stable throughout the year
- T2s efficiencies are not much below T0/T1s
  - It is possible to equalize all, it is in the storage and networking
- Biggest gains through
  - Inter-sites network improvement (LHCONE); networking session on Friday
  - Storage – keep it simple – xrootd works best directly on a Linux FS and on generic storage boxes

What's in store for 2014

- Production and analysis will not stop – know how to handle these, nothing to worry about
  - Some of the RAW data production is left over from 2013
- Another 'flat' resources year – no increase in requirements
- Year 2015
  - Start of LHC RUN2 - higher luminosity, higher energy
  - Upgraded ALICE detector/DAQ – higher data taking rate; basically 2x the RUN1 rate

What's in store for 2014 - sites

- We should finish with the largest upgrades before March 2015
  - Storage – new xrootd/EOS
  - Services updates
  - Network – IPv6, LHCONE
  - New sites installation – Indonesia, US, Mexico, South Africa
  - Build and validate new T1s – UNAM, RRC-KI (already on the way)

Ramp up to 2015

- Some (cosmics trigger) data taking will start June-October 2014
  - This concerns the Offline team – nothing specific for the sites
- Depending on the ‘intensity’ of this data taking, or how many thing got broken in the past 2 years
  - The central team may be a bit less responsive for site queries

Last trimester of 2014

- ALICE will start standard shifts
- Technical, calibration and cosmics trigger runs
- Test of new DAQ cluster – high throughput data transfers to CERN T0
  - Does not affect T1s… since we do data transfer continuously
- Reconstruction of calibration/cosmics trigger data will be done
- Expected start of data taking – spring 2015

Summary

- Stable and productive Grid operations in 2013
- Resources fully used
- Software updates successfully completed
- MC productions completed according to requests and planning
  - Next year – continue with RAW data reprocessing and associated MC
- Analysis – OK
- 2014 - focus on SE consolidation, resources ramp-up for 2015 (where applicable), networking, new sites installation and validation

A big Thank You to all sites providing resources for ALICE and their ever-vigilant administrators

A big Thank You to the Tsukuba organizing committee for hosting this workshop

Summary of the workshop

- 63 participants (first day – common session)
- 54 participants next days
- Record participation!

General Themes

- Wednesday – Grid operations, computing model, AliEn development, WLCG development, resources
  - Two very interesting external presentations on Tokyo T2 and Belle II experiment – we thank the presenters for sharing their experiences and ideas
- Thursday – Storage and monitoring
- Friday - Networking

Site themes

- 17 regional presentations
- 2 site-specific presentations
- News on Indonesia, US and China

Finally… the group photo