ALICE Experiment Status, Run2 Plans, & Federated Requirements

XRootD Workshop @ UCSD

27 January 2015

Latchezar Betev

xrootd in ALICE - Happy 10 years anniversary!

- February 2005

… an ordinary meeting

Hi All,

As has been discussed by email, Andy Hanushevsky is at CERN next week so we'd like to have a mini-workshop:

place: 40-SS-C01
date: Tuesday 15 February 2005
time: 14:00-18:00

I have put together a rough agenda to guide things:

14:00 xrootd latest developments - Andy
14:40 ROOT authentication/security scheme - Gerri
15:00 LDPM - Jean-Philippe
+ discussion
16:00 Castor
+ discussion

Andy and Gerri will have presentations, but hopefully we can make some progress via informal discussions in understanding how xrootd can be useful for everyone.

The room is a bit large, but it was difficult to find a normal size conference room.

thanks,
Pete

… enter xrootd for the Grid

- Out of this meeting came the initial ideas to adopt xrood for ALICE Grid operations
  - Already in ROOT – a natural extension + WAN + other
- AliEn tools by Andreas Peters (of EOS fame), Derek Feictinger (ARDA), Fabrizio Furano (lead xrootd developer)
- Monitoring by Catalin Cirstoiu (ARDA) and Costin Grigoras
- By end of 2005 the ALICE Grid system was fully xrootd-ready
- … and the rest is history

Back to today

RUN 2 physics programme and rates

- Target - integrated luminosity of $1nb^{-1}$ of Pb-Pb collisions (combined RUN 1+RUN 2)
  - Consistent with the ALICE approved programme
  - 4-fold increase in instant luminosity for Pb-Pb
- Double event rate of TPC/TRD
- Increased capacity of HLT and DAQ systems
  - Rate up to 8GB/sec to T0

Heavy Ion data taking

RUN 2 detector upgrades

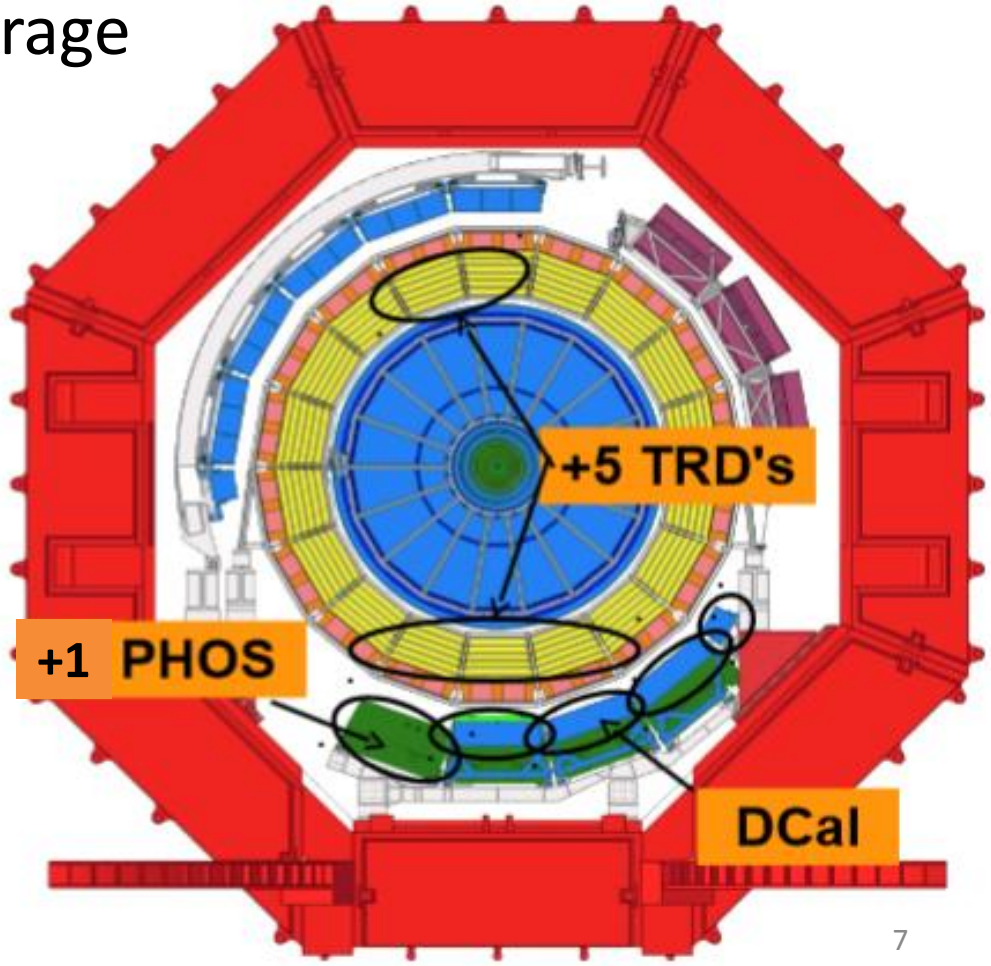
- TPC, TRD readout electronics consolidation
- TRD full azimuthal coverage

(+5 modules)

- +1 PHOS calorimeter

module

- New DCAL calorimeter



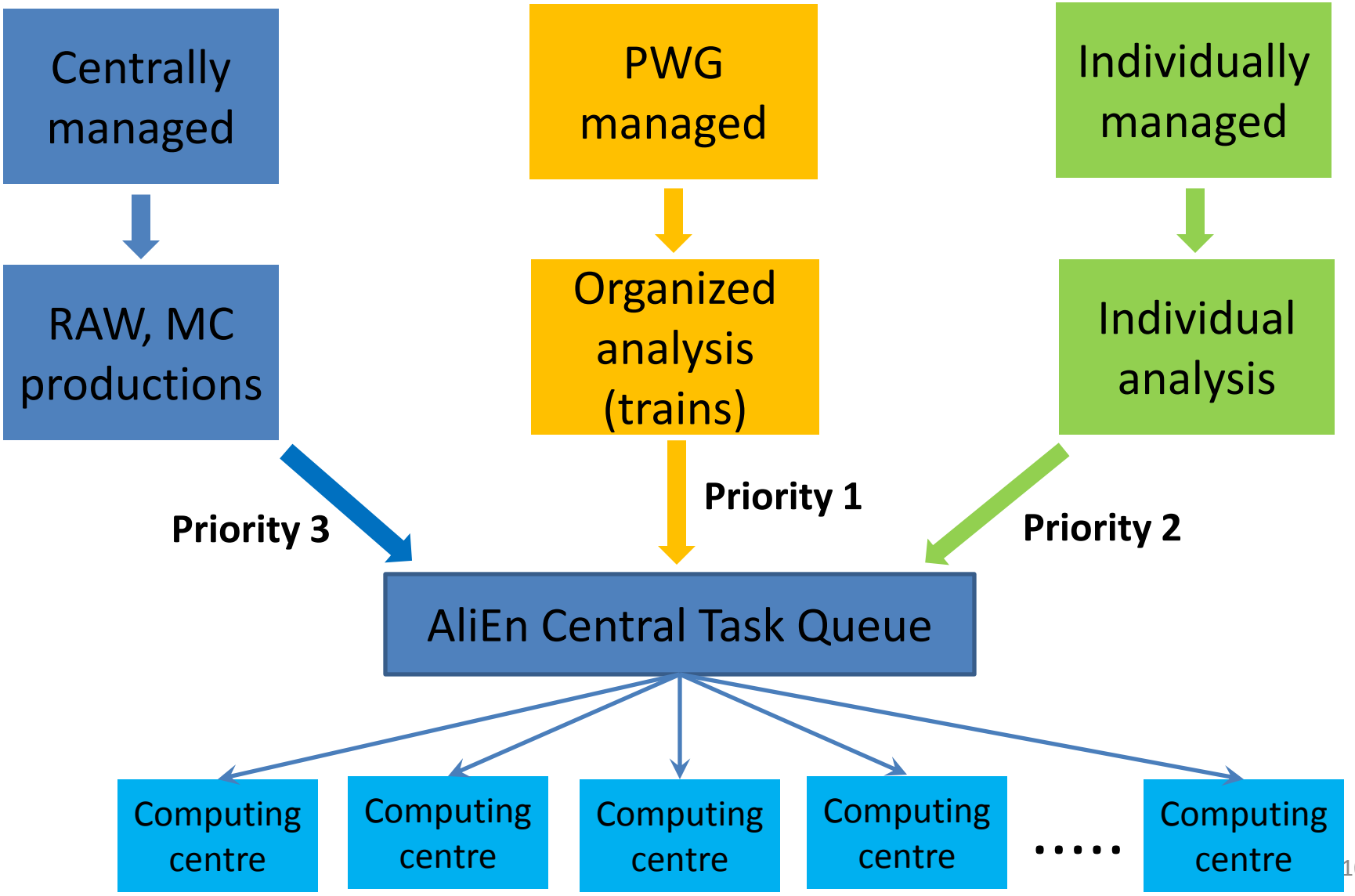
RUN 2 resources considerations

- Same CPU power needed for reconstruction
- 25% larger raw event size
  - Additional detectors
  - Higher track multiplicity with increased beam energy and event pileup
- ALICE requirements for RUN2 were approved by CRSG in April 2014
- The CPU request growth is compatible with 'flat' budget, i.e. depends purely on technology development
- Major demand on resources towards the end of 2015 (Pb-Pb data taking)

Basics for 2015-2018 operation

- ALICE Grid model remains largely unchanged in RUN2
  - Integration of every new computing centre into the Grid
  - Average 2 replicas of analysis objects => dependency on resources stability
  - Low differentiation of tasks – T0/T1s are still RAW data keepers and producers, all other tasks are performed everywhere
  - Tasks are generally send to data, but data can go to the tasks if needed

Computing tasks and workflow

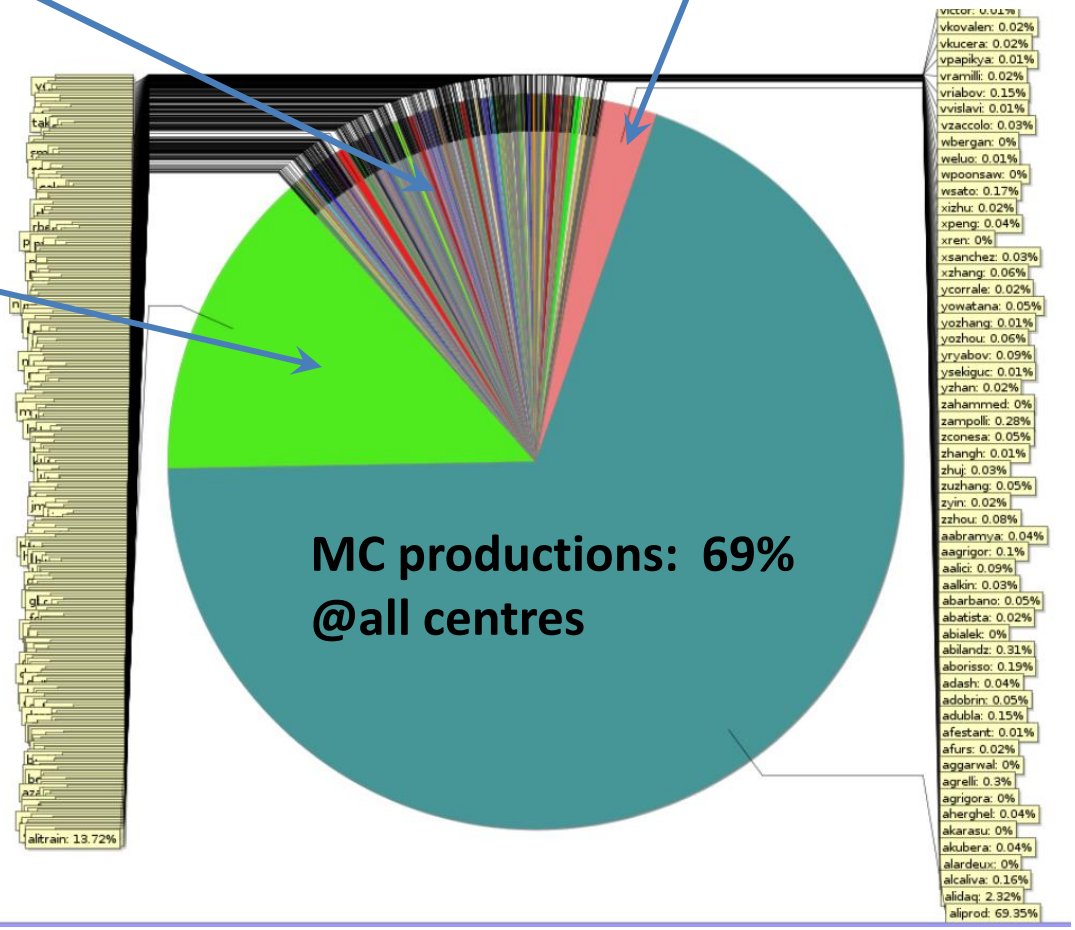


Wall time resources share 2014

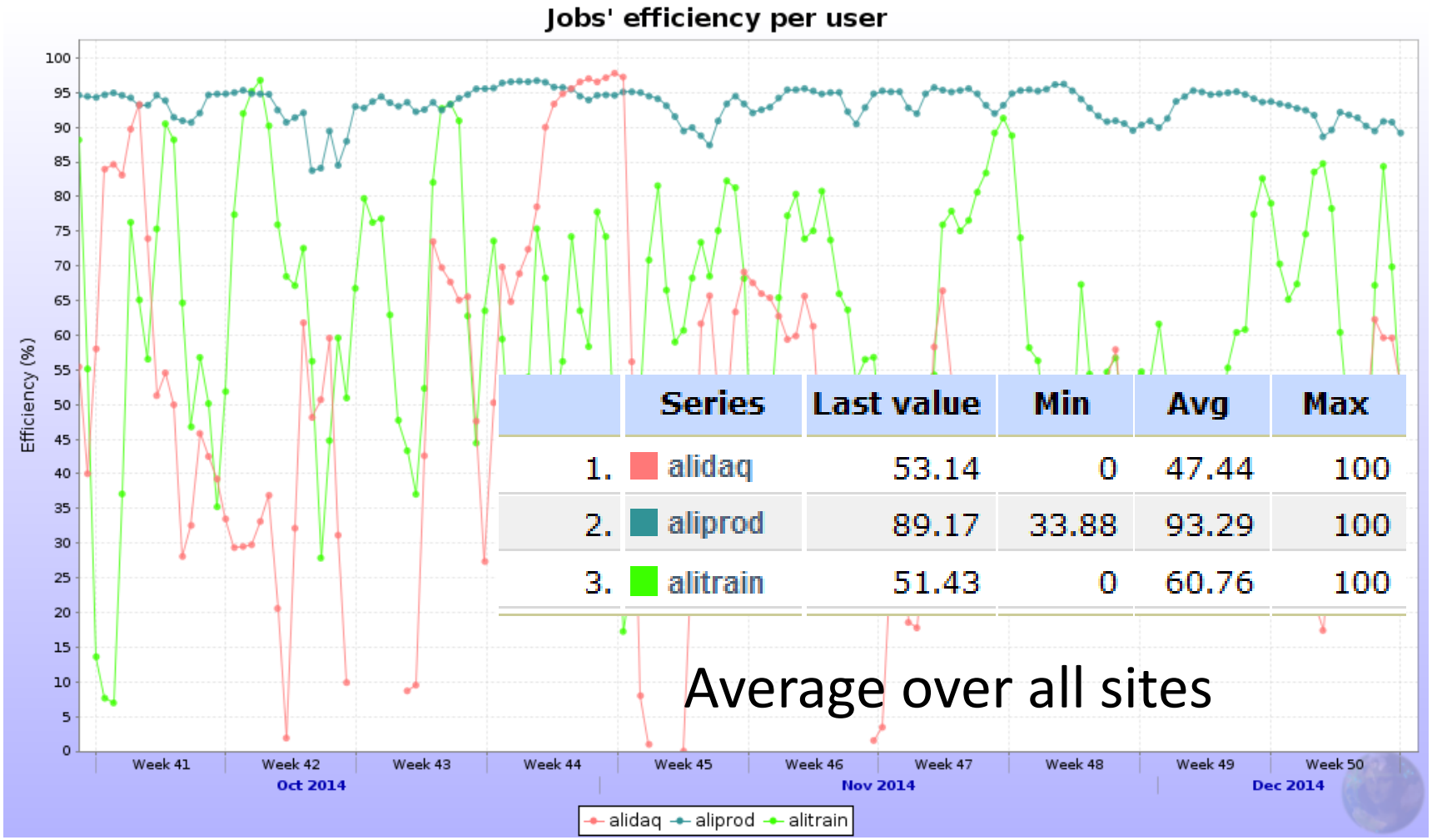
Individual analysis: 14%
@all centres
425 users

RAW data processing: 3%
@ T0/T1s only

Organized analysis: 14%
@all centres



Efficiency per workflow

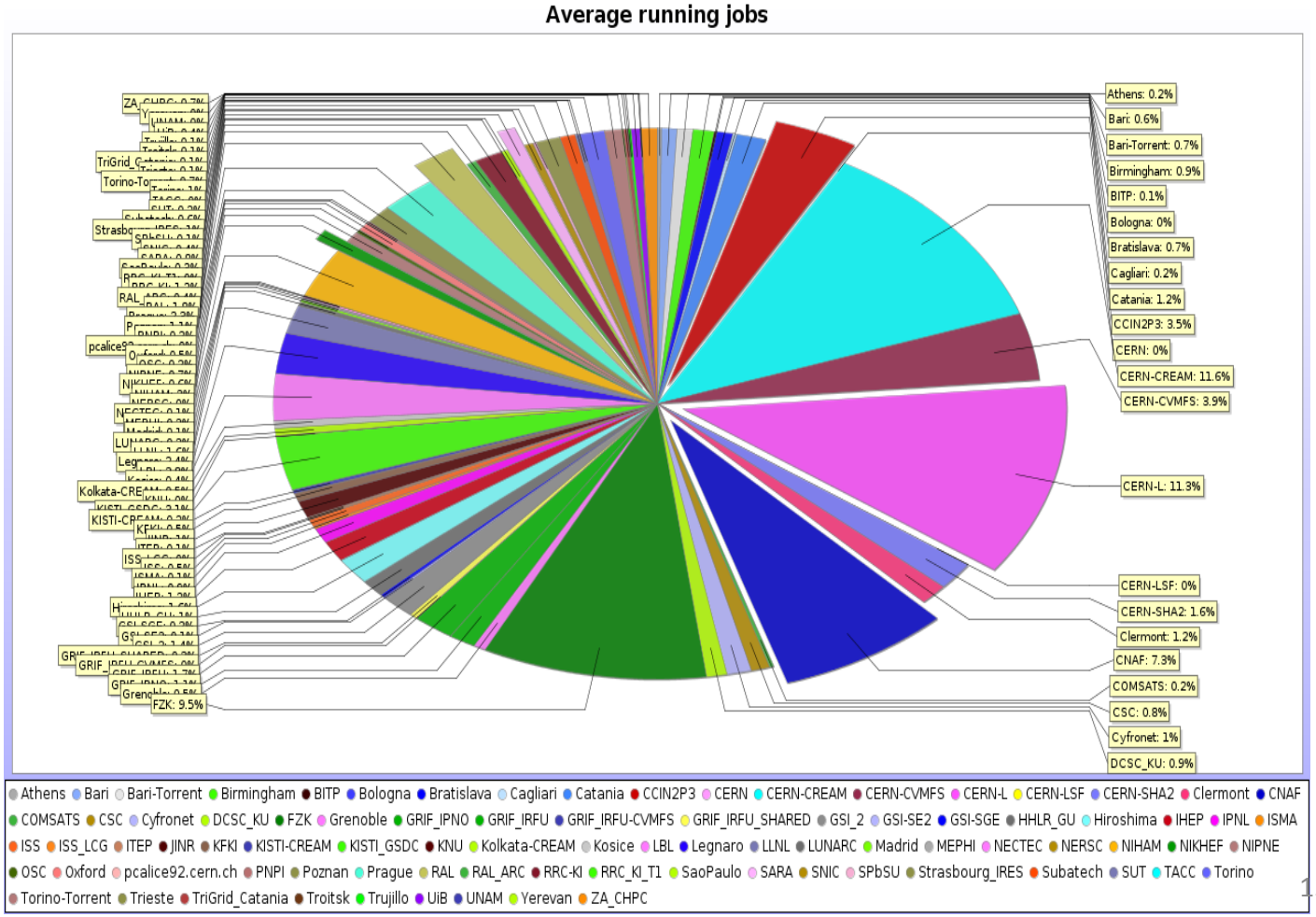


	Series	Last value	Min	Avg	Max
1.	alidaq	53.14	0	47.44	100
2.	aliprod	89.17	33.88	93.29	100
3.	alitrain	51.43	0	60.76	100

Average over all sites

Resources distribution

Remarkable 50/50 share between large (T0/T1) and smaller computing centres



ALICE data model

- All ALICE data are annotated in the AliEn catalogue
  - Including the location on site SEs
- Data files are accessed directly
  - Jobs go to the data, in case of local failure reads from closest replica
  - User access to data is managed through a shell, which connects to the catalogue and downloads/uploads data to the site SEs
- Exclusive use of xrootd protocol
  - Also supporting http, ftp, torrent for downloading other input files
  - At the end of the job N replicas are uploaded from the job itself (2x ESDs, 2xAODs, 1x logs and other service files)

Replica discovery mechanism

- Closest working replicas are used for both reading and writing
  - Sorting the SEs by the network distance to the client making the request
    - Combining network topology data with the geographical location
  - SEs are weighted with reliability and occupancy
- Writing is slightly randomized for more ‘democratic’ data distribution
  - Important for isolated sites
- Assures data security and sufficient number of replicas for efficient analysis

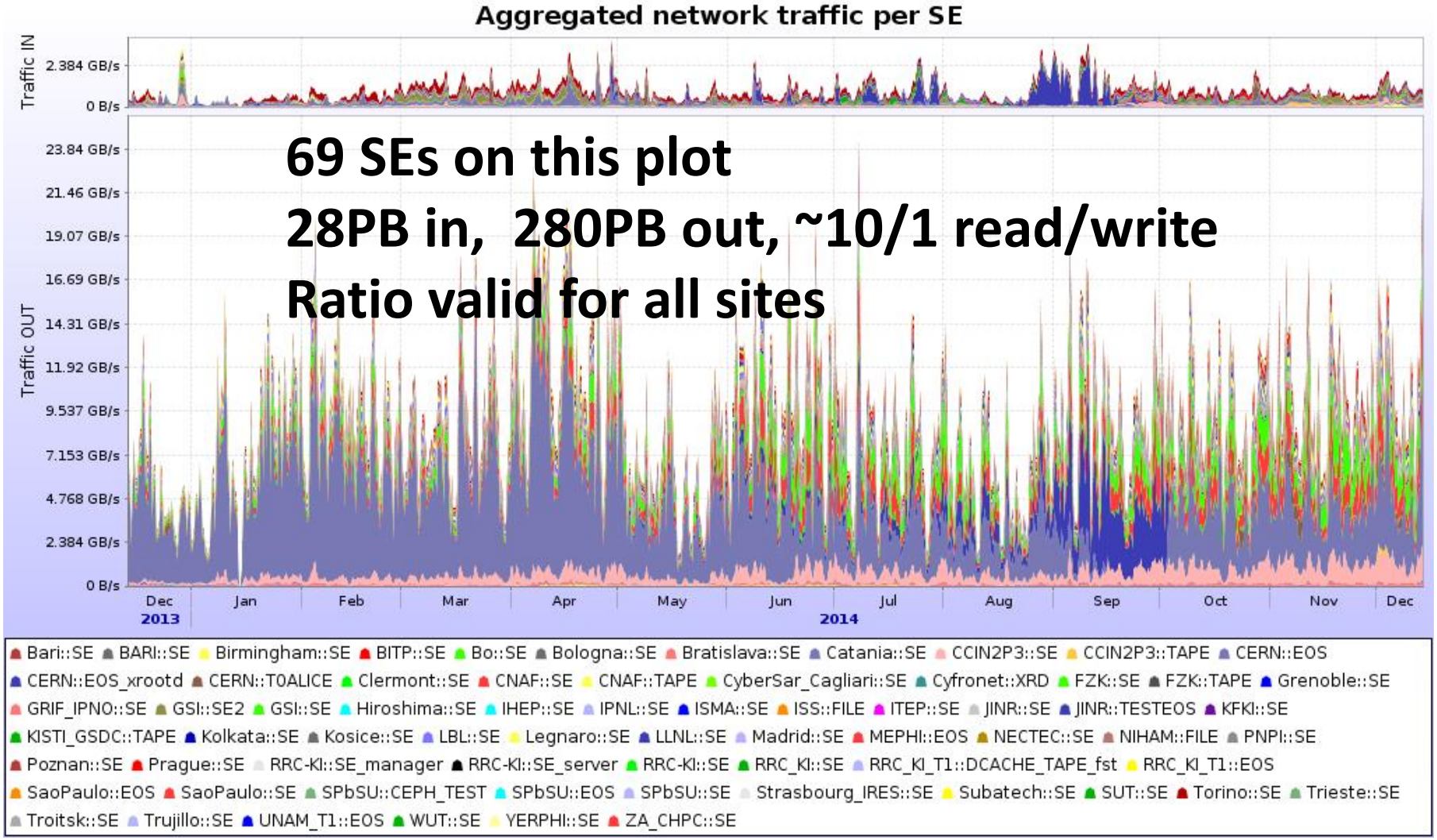
Federation mechanism

- Users and jobs are isolated from the underlying storage complexity
  - Interaction only with the catalogue (Logical File Names)
- All SE access passes through services
  - LFN-PFN (GUIDs) translation
  - Authorization and authentication
  - Quotas
- Jobs and users cannot read/write to any SE directly

Federation mechanism (2)

- The ALICE storage is fully federated
  - Through the AliEn catalogue and the services
  - ‘isolated’ and ‘manage yourself’ SEs don’t exist
  - Direct access to the files from anywhere possible
  - Data replication (RAW/lost files recovery/SE evacuation for repair/updates) through xrd3cp, also integrated in the system
- Bonus
  - Full control of the SE filling, deletion, data migration
  - Low risk of catastrophic data loss in case of full SE failure
  - No need for the end user to know file location, SE occupancy, and plan around these

Typical data access rates



Data access in analysis tasks

- 1M analysis tasks (mix of all types)
  - 14.2M input files
  - 90% accessed from the site local SE at **3.1MB/s**
  - 10% read from remote at 0.97MB/s
  - Average processing speed 2.76MB/s
- Job efficiency 70% for an average CPU power of 10.14 HepSpec06
- => need **0.4MB/s/HepSpec06** for analysis on any site (T0/T1s/T2s)

Summary

- In the period 2015-2018 (LHC RUN2) ALICE will collect data volume ~3x larger than during RUN1
- The computing model remains largely unchanged, storage access exclusively through xrootd
- The planned computing resources increase is expected to meet the demands
- The focus of Grid development will be on improving the analysis efficiency and decreasing the turnaround time of the organized trains