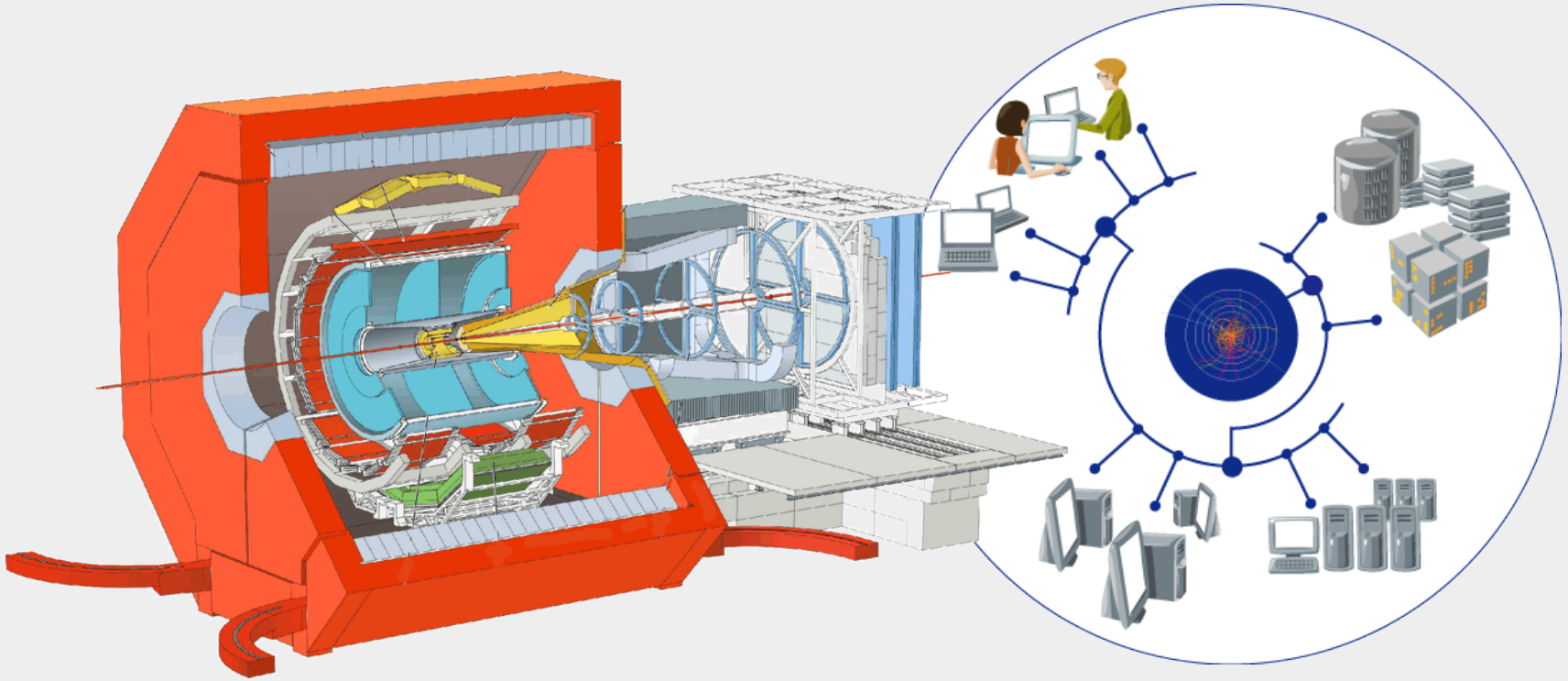


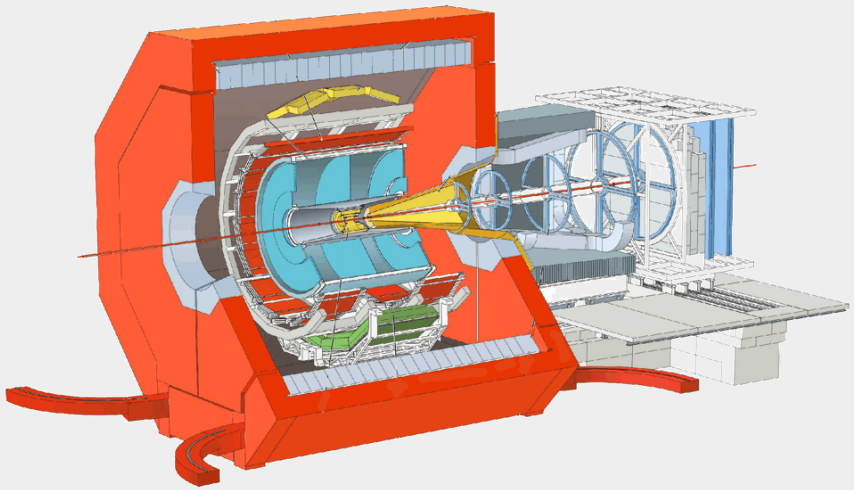
ALICE DATA MANAGEMENT

S. Bagnasco, INFN Torino

Workshop congiunto CCR/INFNGRID
La Biodola, May 16-20, 2011

OUTLINE

- The ALICE computing model
  - Yet again, with a focus on Data Management
- Scheduled activity
  - HI processing
- Chaotic activity
  - Getting ready for QM2011
- Things change
  - Evolution of the CM
- Impact on CPU efficiency
  - [Put a punchline here]



THE ORIGINAL ALICE COMPUTING MODEL

- Tier-0

  Does: first pass reconstruction; calibration and alignment

  Stores: one copy of RAW, calibration data and first-pass ESDs

- Tier-1

  Does: reconstructions and scheduled batch analysis

  Stores: second collective copy of RAW, one copy of all data to be kept, disk replicas of ESDs and AODs, replica of calibration data

- Tier-2

  Does: simulation and end-user analysis

  Stores: disk replicas of AODs and ESDs

Tier role distinction is becoming more shaded: except for reconstruction, everybody does everything if needed or possible

COMPUTING STRATEGY

- AliEn as a common front-end for all distributed resources
  - Using transparent interfaces to different grids where needed
  - Xrootd as a common file access protocol
- Resources are shared
  - No “localization” of data
  - File and job quota enforced in the Central Services
  - Prioritisation of jobs in the central Task Queue
- Data access only through the GRID and AAF
  - No backdoor access to data
  - No “private” processing on shared resources
  - No “private” resources outside of the grid

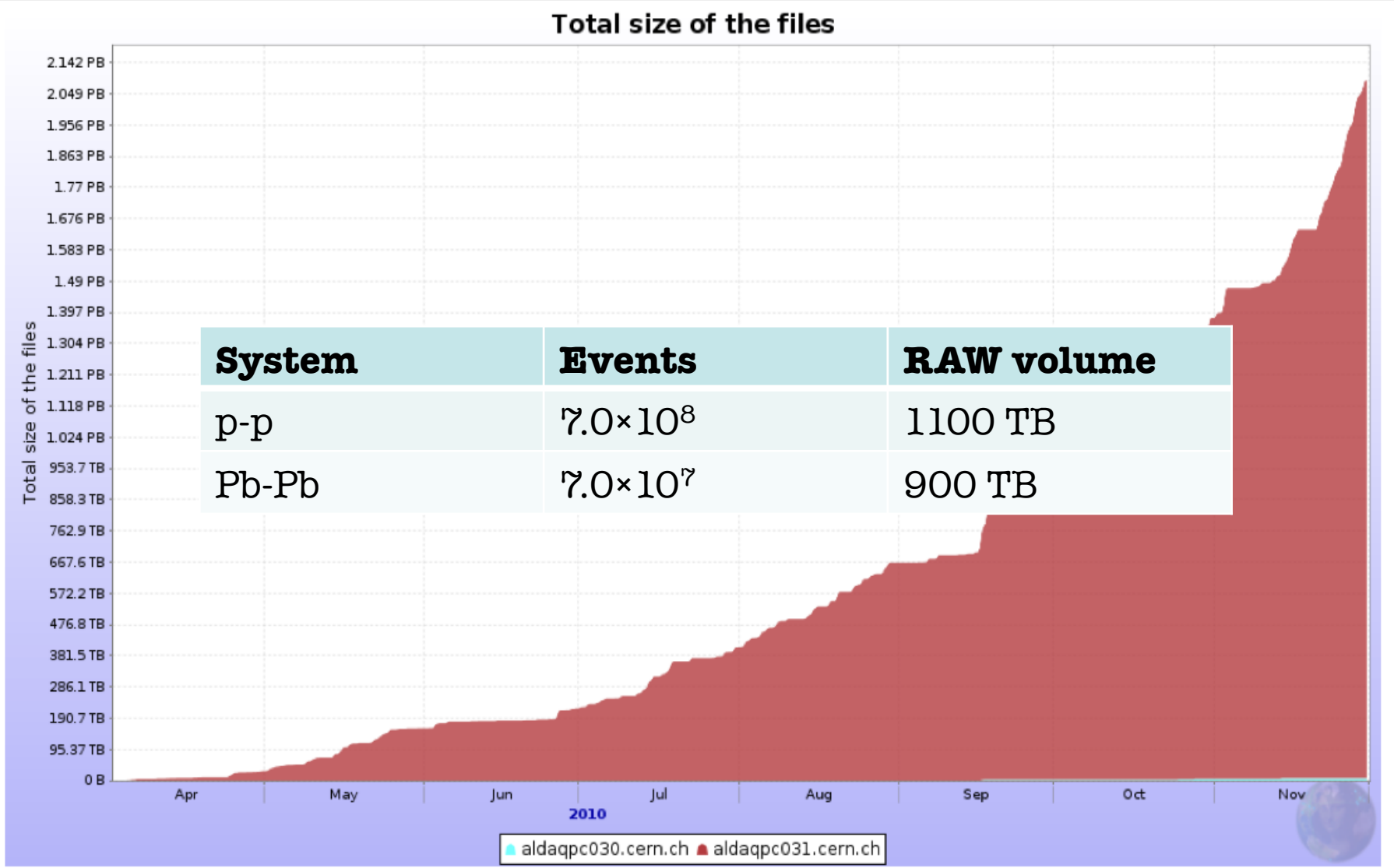
DATA MANAGEMENT KEY CONCEPTS

- Central File Catalogue
  - Central DB of all file produced
  - Enforcement of access rights, quotas, policies etc.
  - FS-like browsable interface for users
- Calibrations and conditions data are no different
  - Root files accessible via catalogue entries
  - Database structure (OCDB) for structured access
  - Replicated in Tier-1s (not yet)
- Xrootd as uniform access protocol
  - Across **sites**, **storage architectures** and **use cases**
  - Run the same analysis macro locally, on PROOF or on the Grid accessing data regardlessly of their physical location
- Central transfers queue
  - Manages data transfers
  - Uses xrd3cp for transfers

THREE JOB CLASSES

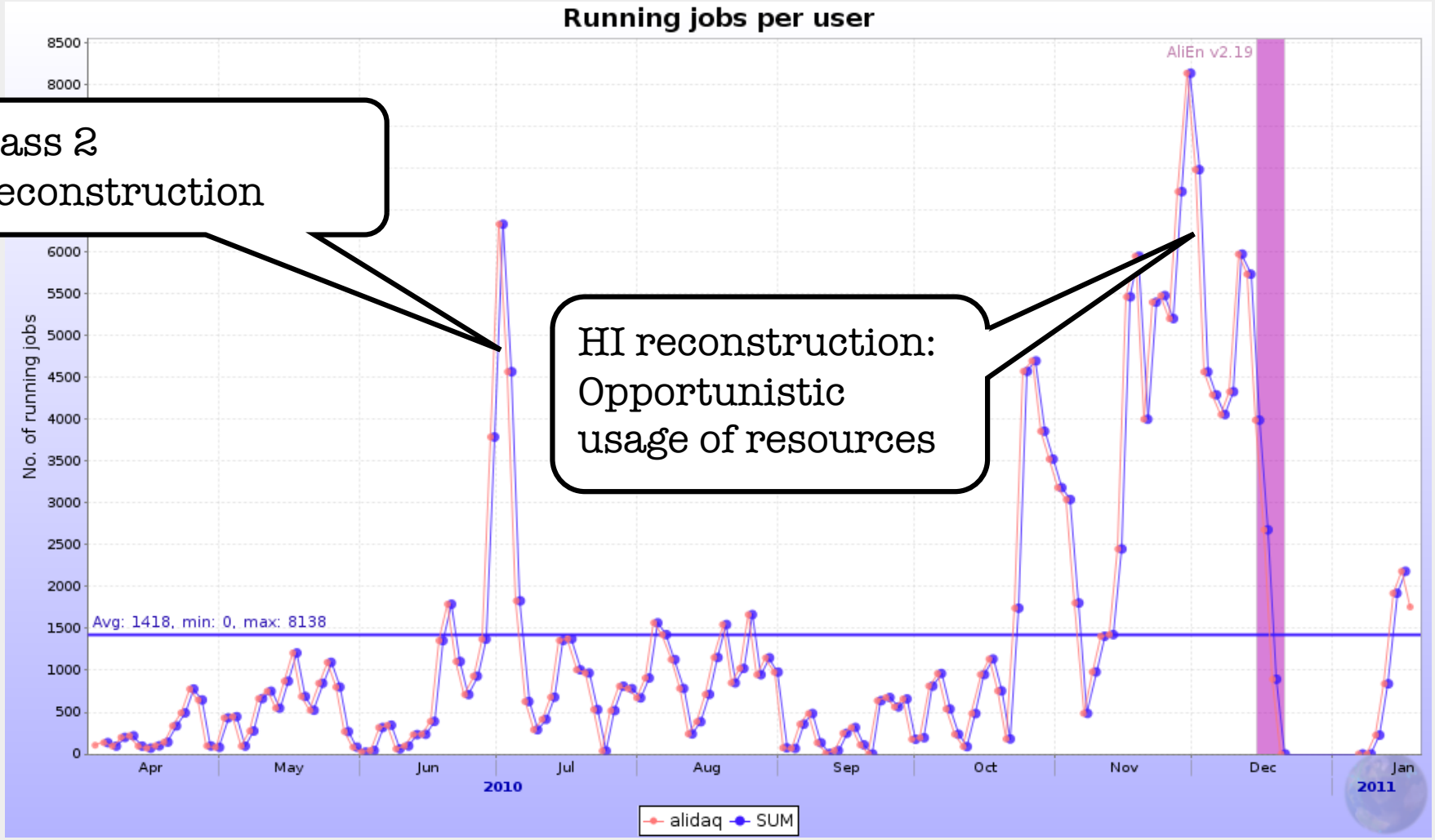
- MC simulation & reco production
  - Low I/O, high CPU efficiency
  - Data export (several copies) after job completion
  - Managed, scheduled
- Analysis Trains
  - Optimized I/O (read once, do many tasks)
  - Streamlined code (as much as possible…)
  - Managed, scheduled
- User jobs
  - Lowest CPU efficiency
  - Variable job duration, many failures, far-from-perfect code
  - Unmanaged, chaotic

2010 DATA SAMPLE

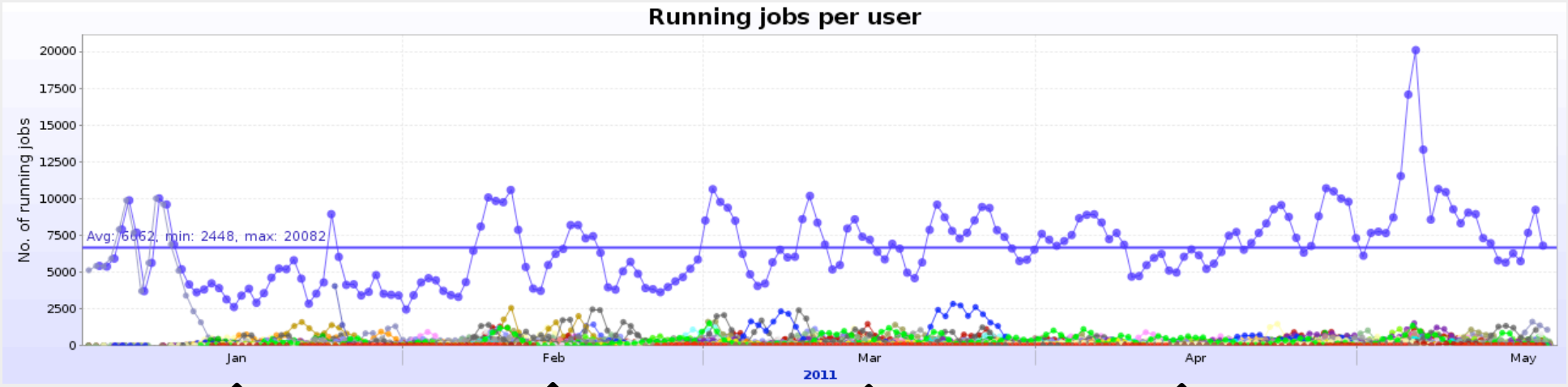


System	Events	RAW volume
p-p	7.0×10^{8}	1100 TB
Pb-Pb	7.0×10^{7}	900 TB

HI DATA PROCESSING



USER ANALYSIS JOBS 2011



January:

- 5000 jobs
- 190 users

(×2.5 increase over 2010 av.)

February:

- 5600 jobs
- 240 users

(+12%, +20%)

March:

- 7100 jobs
- 280 users

(+27%, +16%, ¼ of resources)

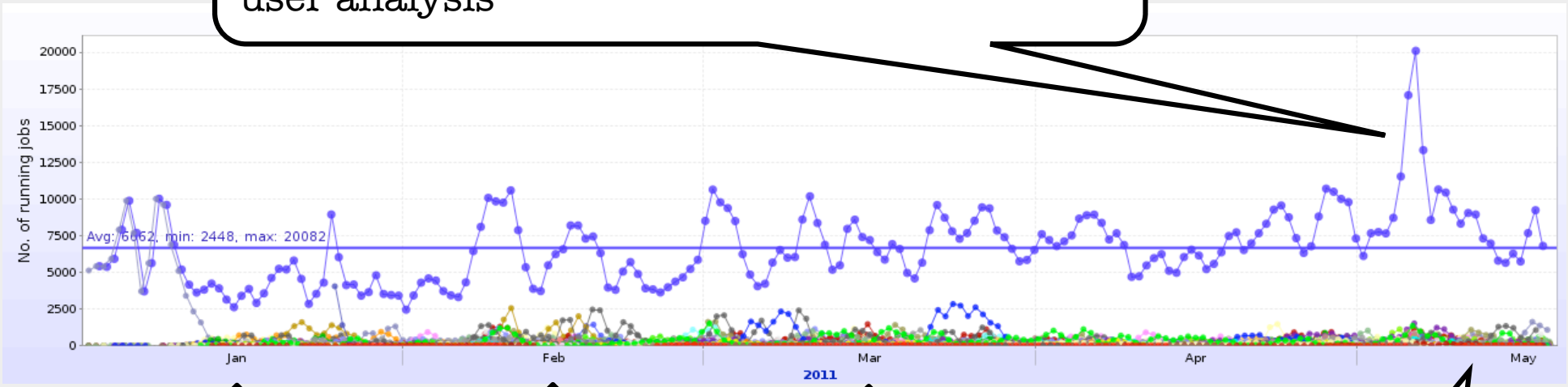
April:

- 7500 jobs
- 290 users

(+27%, +16%)

USER ANALYSIS JOBS 2011

5-6 May: all grid resources freed for users.
80% job slots utilization ONLY from chaotic user analysis



January:

- 5000 jobs
- 190 users

(×2.5 increase over 2010 av.)

February:

- 5600 jobs
- 240 users

(+12%, +20%)

March:

- 7100 jobs
- 280 users

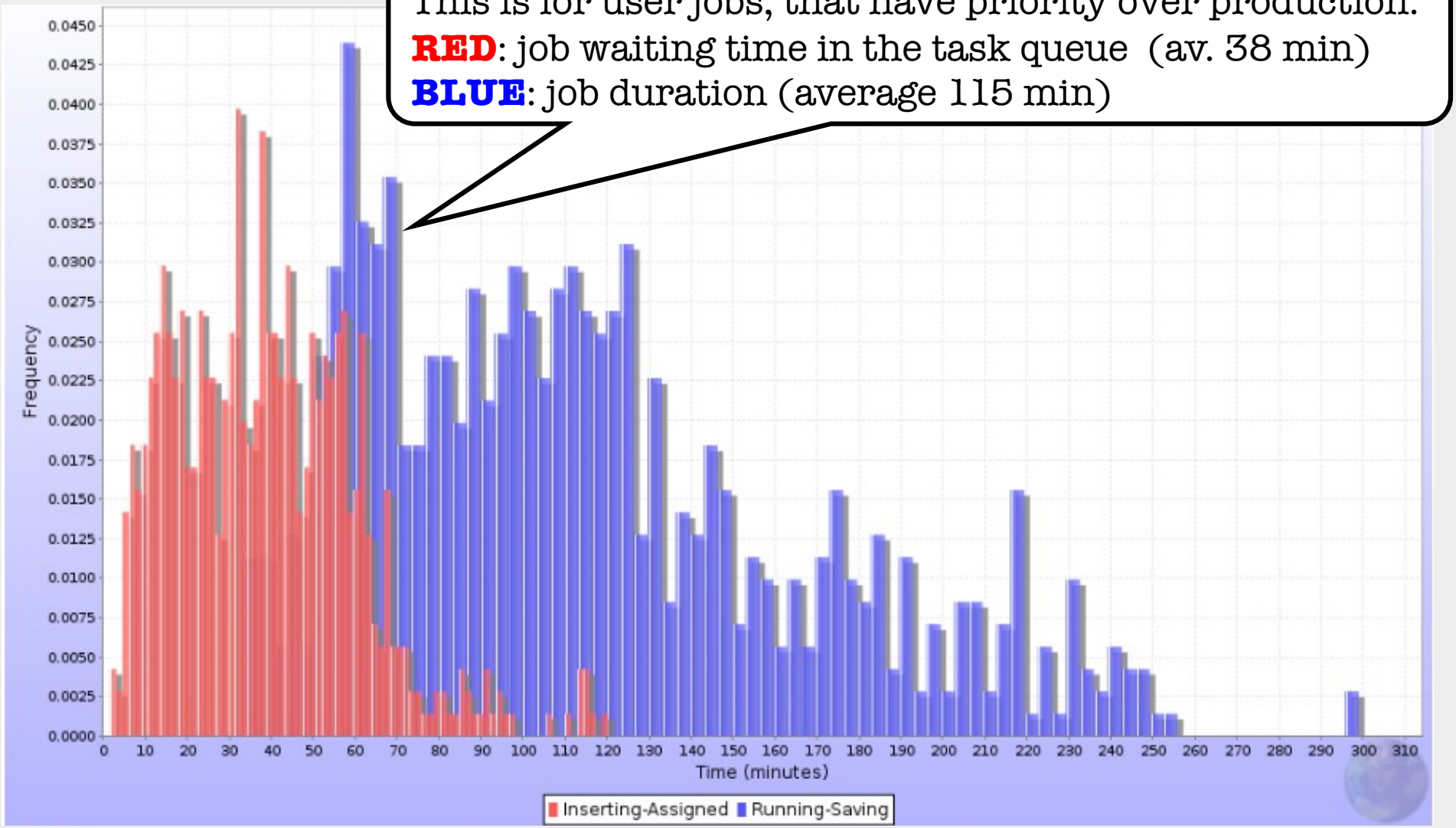
(+27%, +16%, ¼ of resources)

Last two weeks:

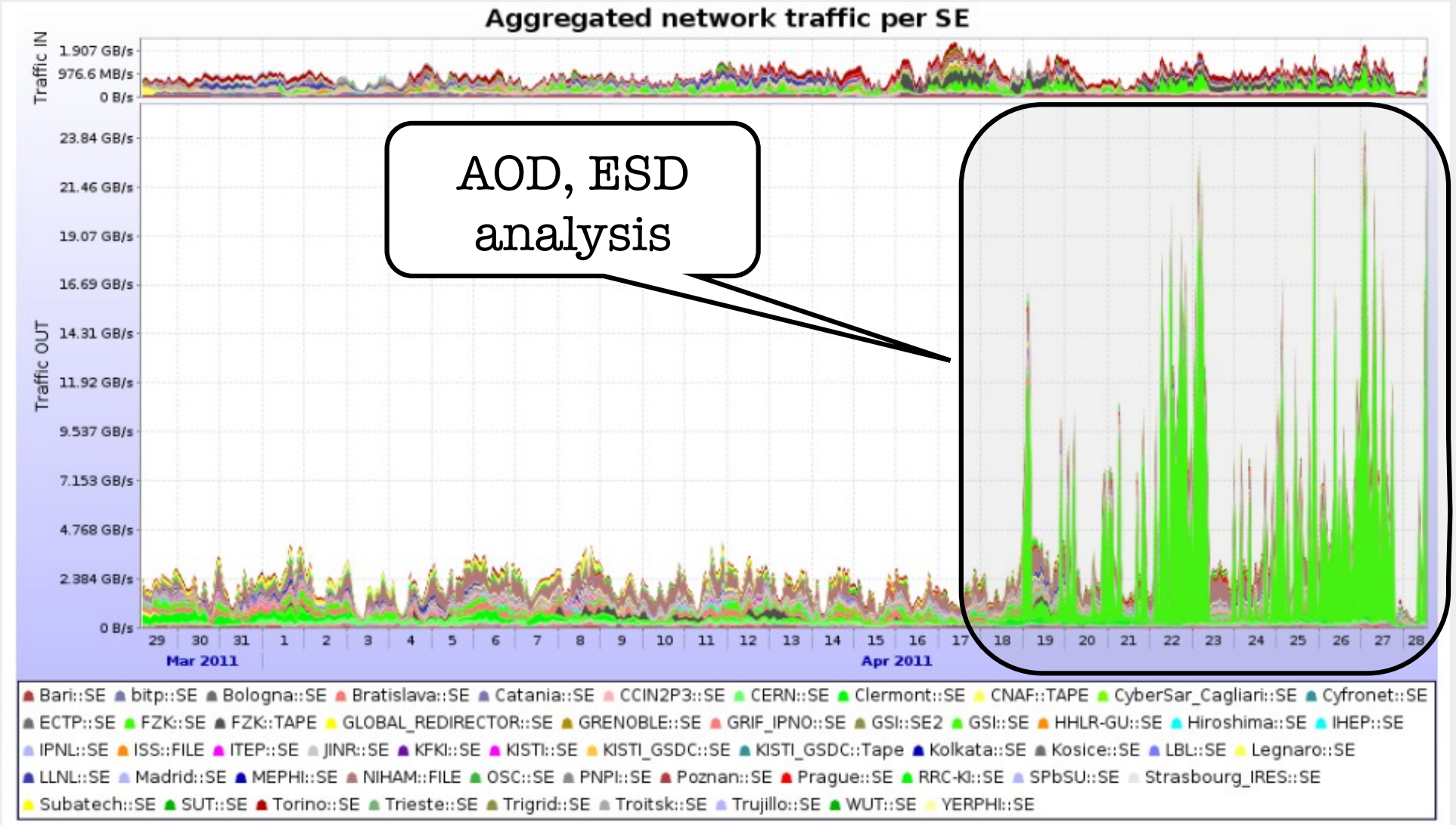
- 10000 av. jobs
- >20000 max jobs

GRID RESPONSIVENESS

This is for user jobs, that have priority over production.
RED: job waiting time in the task queue (av. 38 min)
BLUE: job duration (average 115 min)



USER ANALYSIS ACCESS TO STORAGE



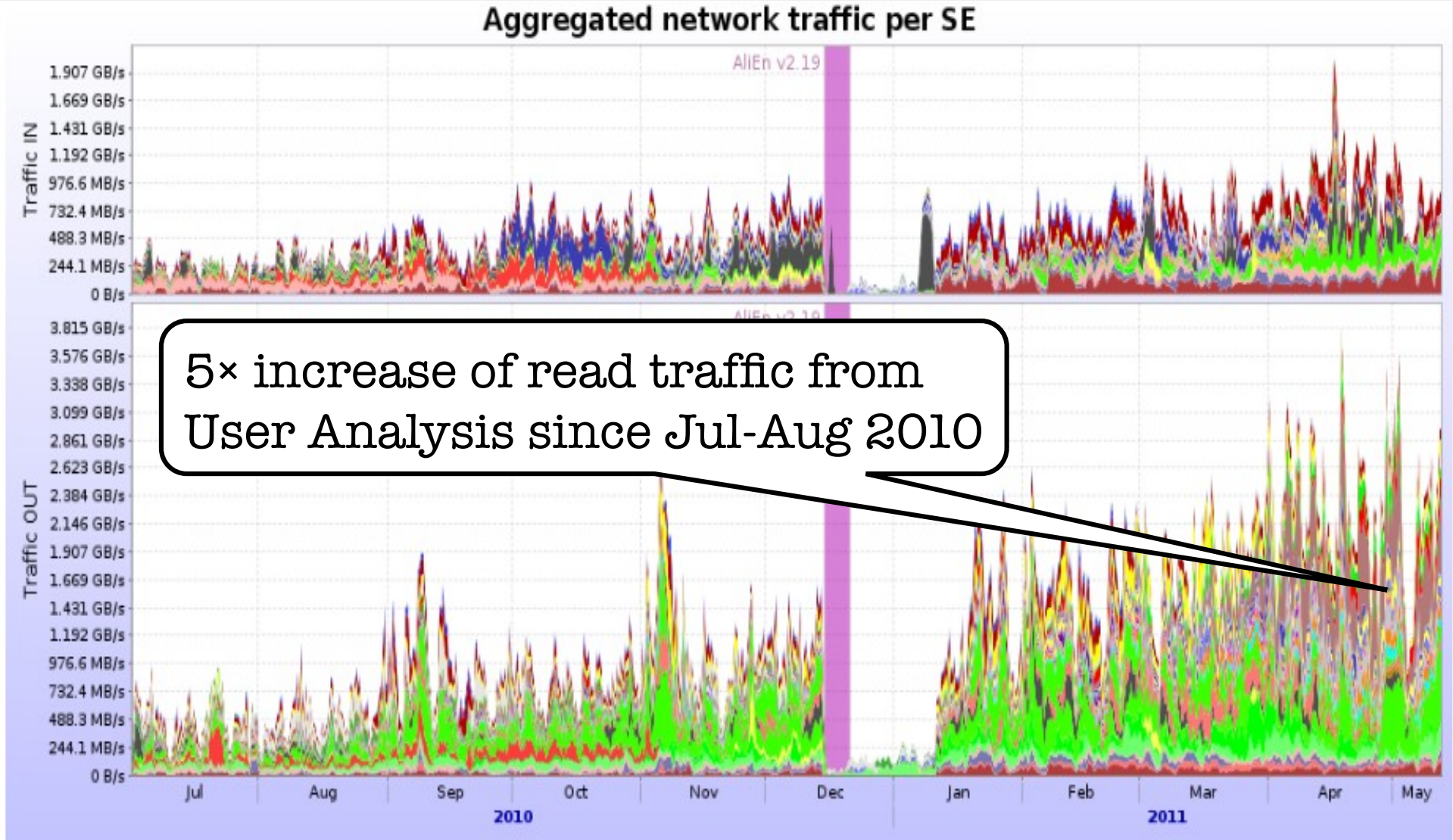
THINGS CHANGE

- Some extra tasks added to original CM
  - Offline calibration
  - QA Analysis trains
  - More to follow
- User Analysis on such a scale was never tried on the Grid before
  - The structure and demands are becoming more clear
  - Again, more to follow
- Everything worked but **computing model is evolving** to take this into account

USER JOBS ARE MESSY

- Diverging memory allocation
  - Killed jobs or even stuck WNs
  - A safety is in place with new AliEn release
- Coding and JDL errors
  - Private code is never tested enough
  - Thousands of jobs can be failing **very** quickly
  - User problem or site problem?
- A user will do anything with an open query
  - E.g. queries with o(10^6) files
  - Whether it makes sense or not
  - Protections need to be in place everywhere

STORAGE RATES HISTORY



UPDATED CM PARAMETERS

	pp/event	PbPb/event
CPU reco (KHEP06×s)	0.07 (+10%)	9.75 (+71%)
CPU MC (KHEP06×s)	1.30 (+40%)	150.00 (+4%)
Raw size (MB)	1.3 (+18%)	12.5 (0%)
ESD size (MB)	0.08 (+37%)	1.20 (-65%)
MC Raw size (MB)	0.4 (0%)	61.5 (0%)
MC ESD size (MB)	0.26 (0%)	50 (0%)

F. Carminati

COMPUTING MODEL EVOLUTION

- More files than ever anticipated
  - Original model:
    1 RAW → 1 ESD → 1 AOD (×3 passes)
  - Current cascade:
    1RAW → 5× ESD-related (×3 passes) →
    → 6× AOD-related (per train) (×*N* passes)
  - MC is more difficult to describe, but also a substantial generator of files
  - Users are prolific generators of files (*.root)
  - In one year we have accumulated **25×10⁶** files in the catalogue (RAW is 1.1 ×10⁶)
  - The physical replication of the above is about **4.2**

COMPUTING MODEL EVOLUTION

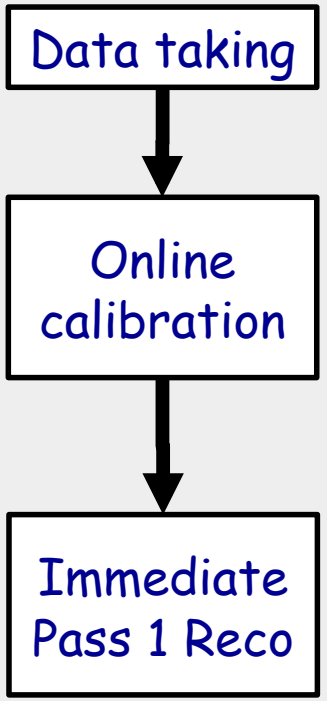
- More complex job structure
  - Added few more reconstruction passes and analysis trains to the original processing model
  - MC is increasing in complexity and is more fragmented (PWG requests,...)
  - User access strongly depends on the file fragmentation from the productions
  - In general, the **jobs are becoming more complex and demanding** on the entire Grid infrastructure
  - "Sending jobs where data is" is becoming more difficult

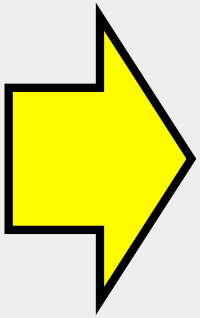
COMPUTING MODEL EVOLUTION

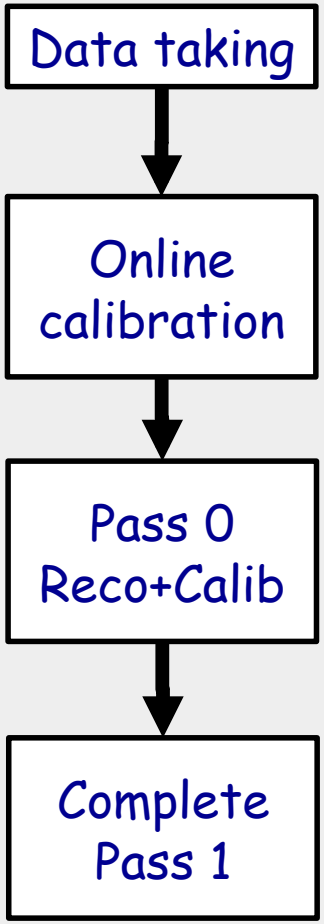
- More access to calibration
  - OCDB is 5× bigger (in number of objects) than originally anticipated
  - Access to OCDB is ~30 more frequent than original projections
  - Will increase substantially with more Pass0, Trains and Tenders – but how much?
  - All of the above has increased the load on the AliEn catalogue and access services dramatically
  - In addition to the massive file access within and outside of the Grid infrastructure

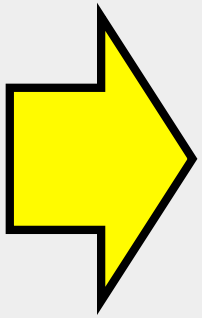
COMPUTING MODEL EVOLUTION

Original CM

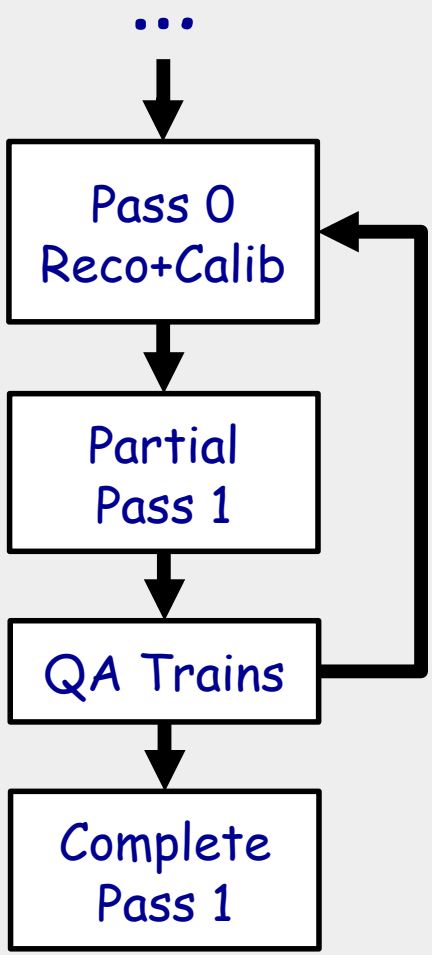








Current implementation



SCHEDULED VS CHAOTIC ANALYSIS

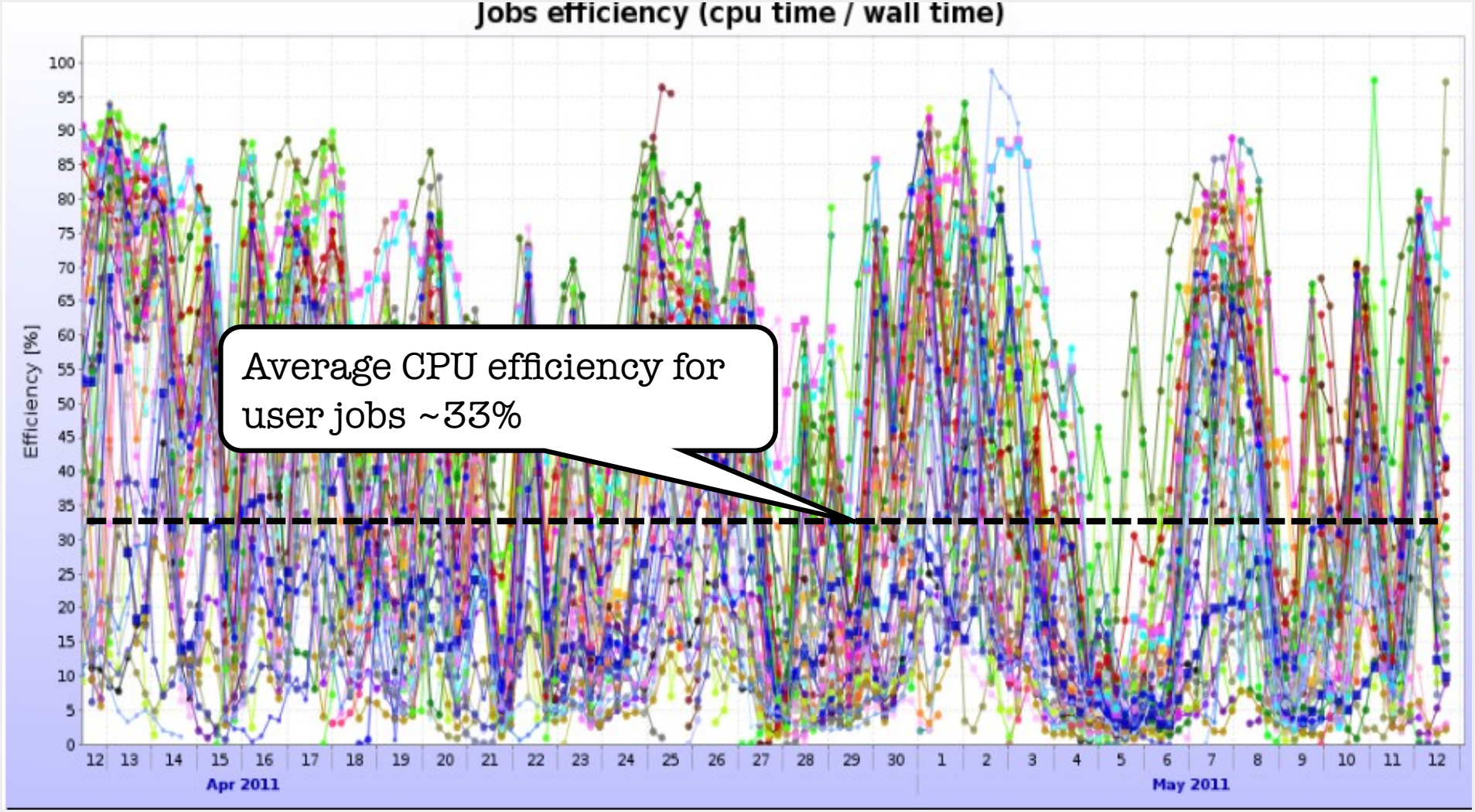
- Scheduled analysis means “analysis trains”
  - “Trains” of several independent “analysis tasks” to reduce number or reads
  - All tasks inherit from a common abstract interface
  - Run on ESD and AOD, may **make heavy use of OCDB** (conditions database).
- User analysis generally mean single tasks
  - A Master Job is generated e.g. by the AliEn Plugin during an interactive ROOT session.
  - The Master Jobs is split in a number o(100) of subjobs
  - Subjob **usually** run where data is located.
  - Subjob results are then **merged** either interactively or by a job.

Boldface red means remote data access.

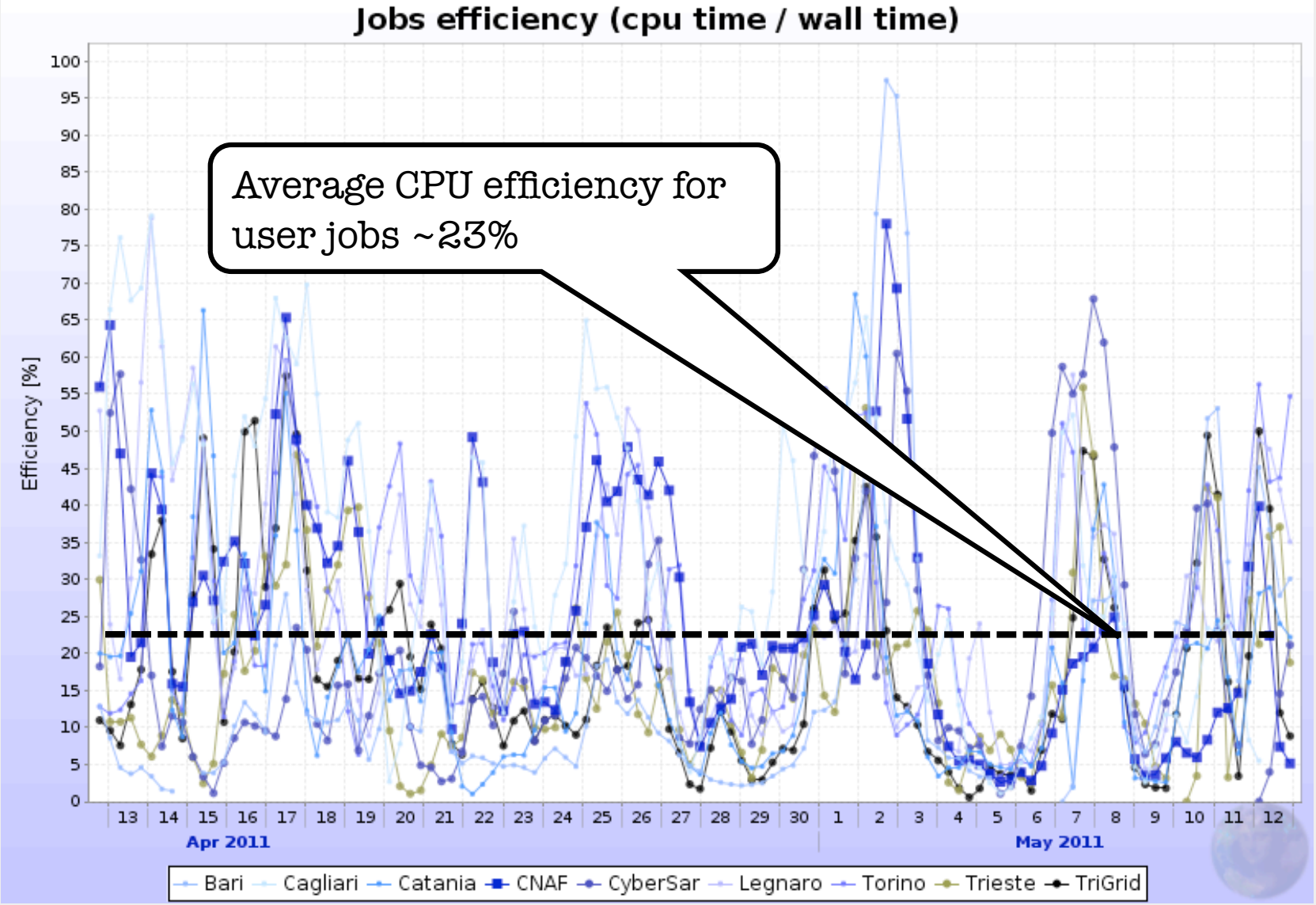
IMPACT ON CPU EFFICIENCY

- Remote data access reduces CPU efficiency
  - How much?
  - Unfortunately, other contributions to CPU efficiency loss appeared more or less at the same time
  - Difficult to decouple, see next slides
  - Also moving data around requires resources ("hidden" inefficiencies)
- Investigation and optimization will be one of the next priorities
  - E.g.: "best" SE discovery

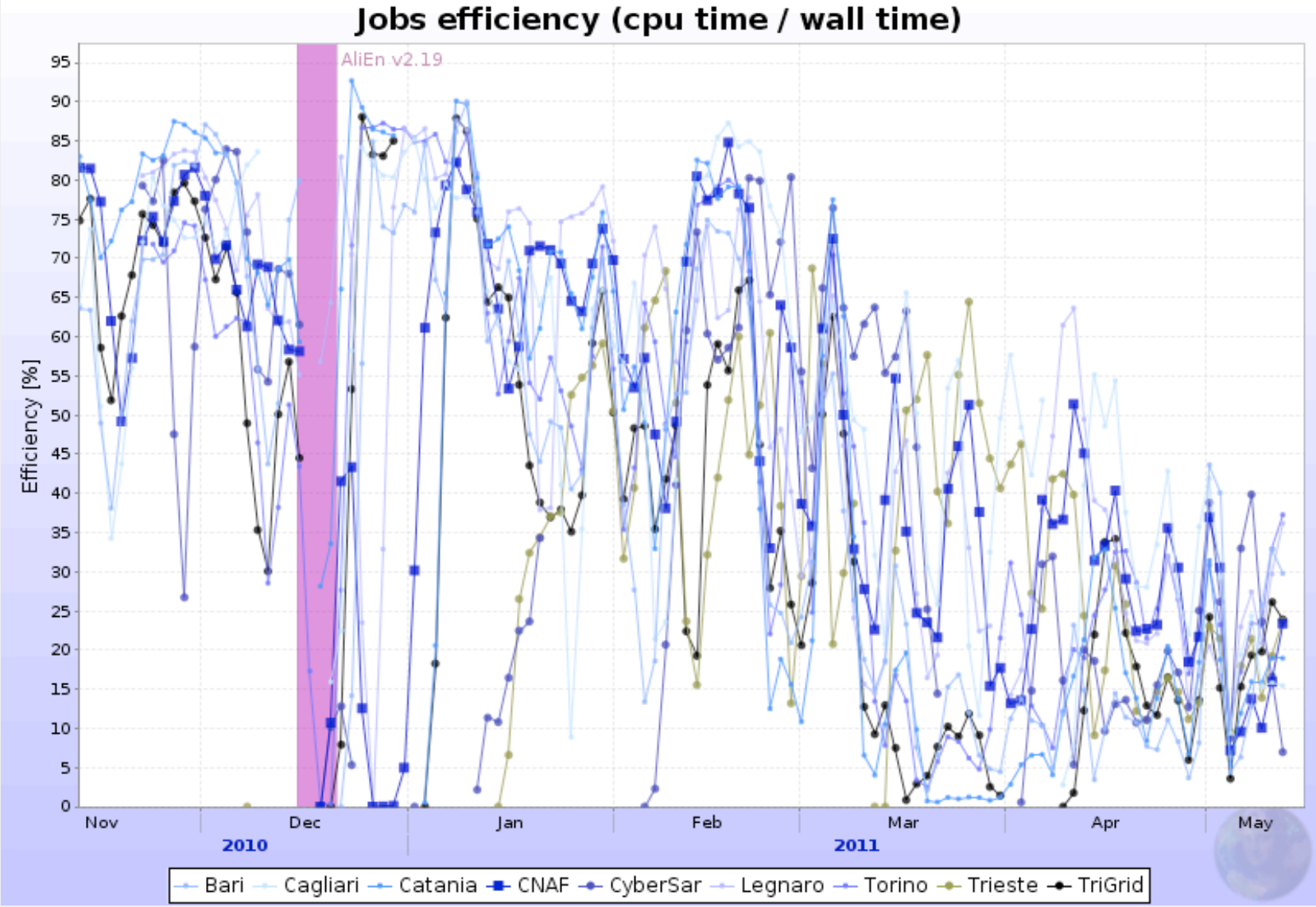
CPU EFFICIENCY HISTORY



CPU EFFICIENCY HISTORY



CPU EFFICIENCY HISTORY



BUGS AND OTHER FEATURES

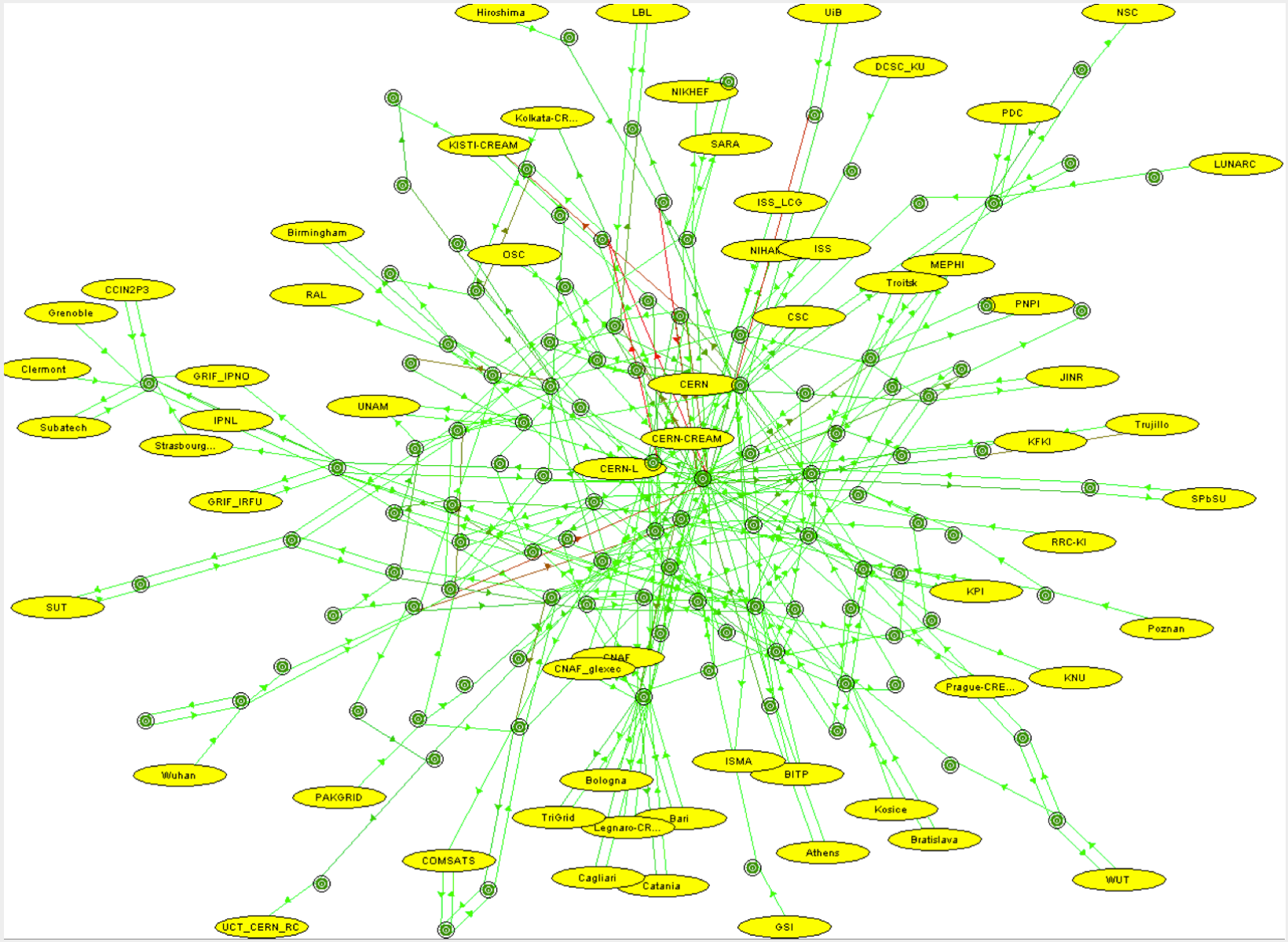
- SE selection mechanism briefly did not work
  - Small, transient effect
- Low-level bug in data access code
  - Read more data than actually needed
  - Bug fixed, will deploy shortly
  - Candidate for being one of the major sources of inefficiency (we'll see)
- "Hanging" jobs
  - Small contribution, and shrinking
- Composition of Analysis Trains can be optimized
  - First priority after QM

Question: which are the best 4 Storage Elements for me to send my files to (or read my data from)?

- Network topology
  - Each SE is associated to a set of Ips (VO-Box, xrootd redirector and servers)
  - Tracepath/traceroute between all sites
  - Measure RTT for the full matrix
- SE "reputability"
  - "Demotion factor" =

$$0.75 \times \frac{\#\,\text{failed tests}}{\#\,\text{tests last day}} + 0.25 \times \frac{\#\,\text{failed tests}}{\#\,\text{tests last week}}$$

NETWORK TOPOLOGY BY AS



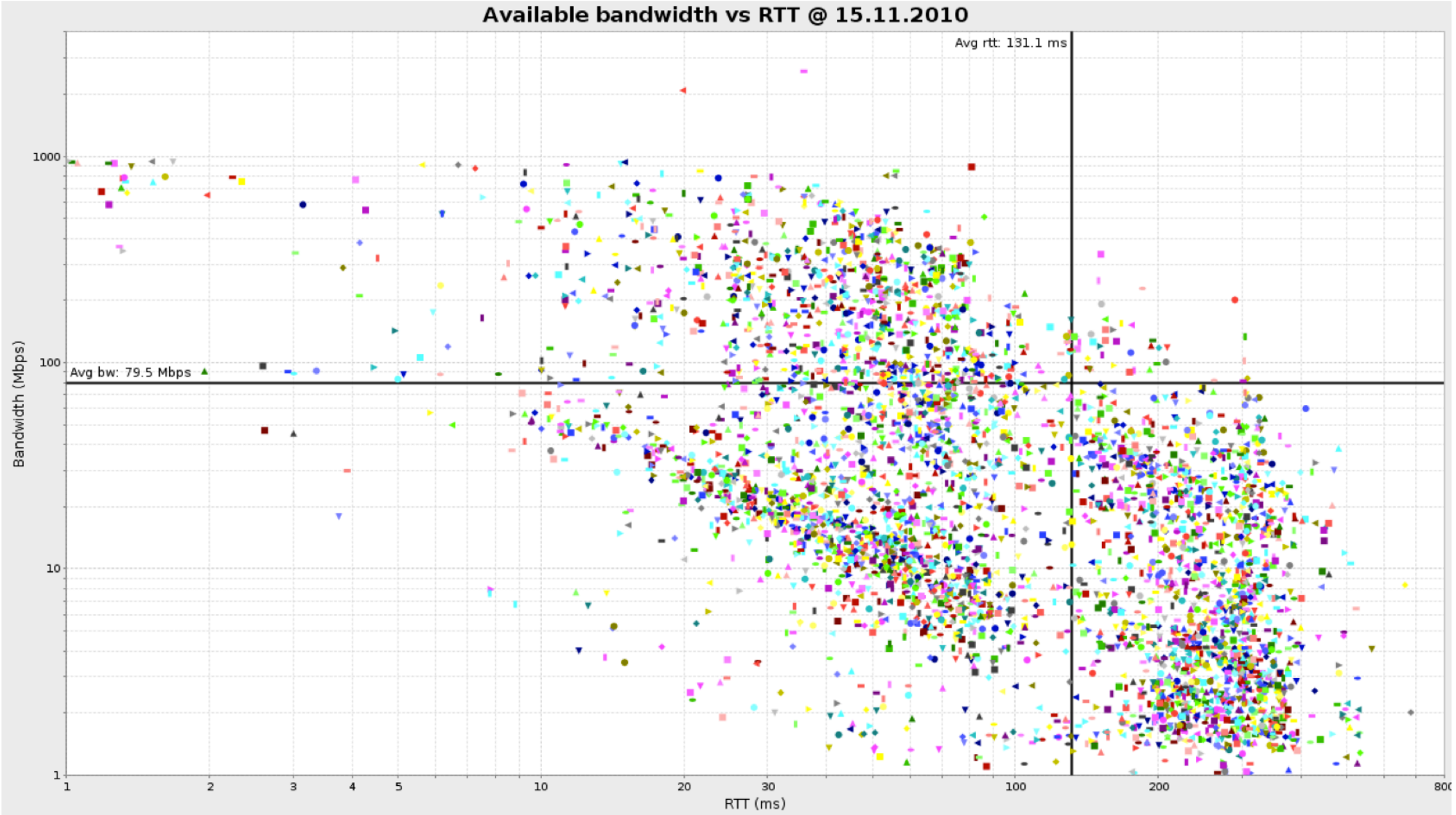
SE METRICS IN ML

Demotion factor

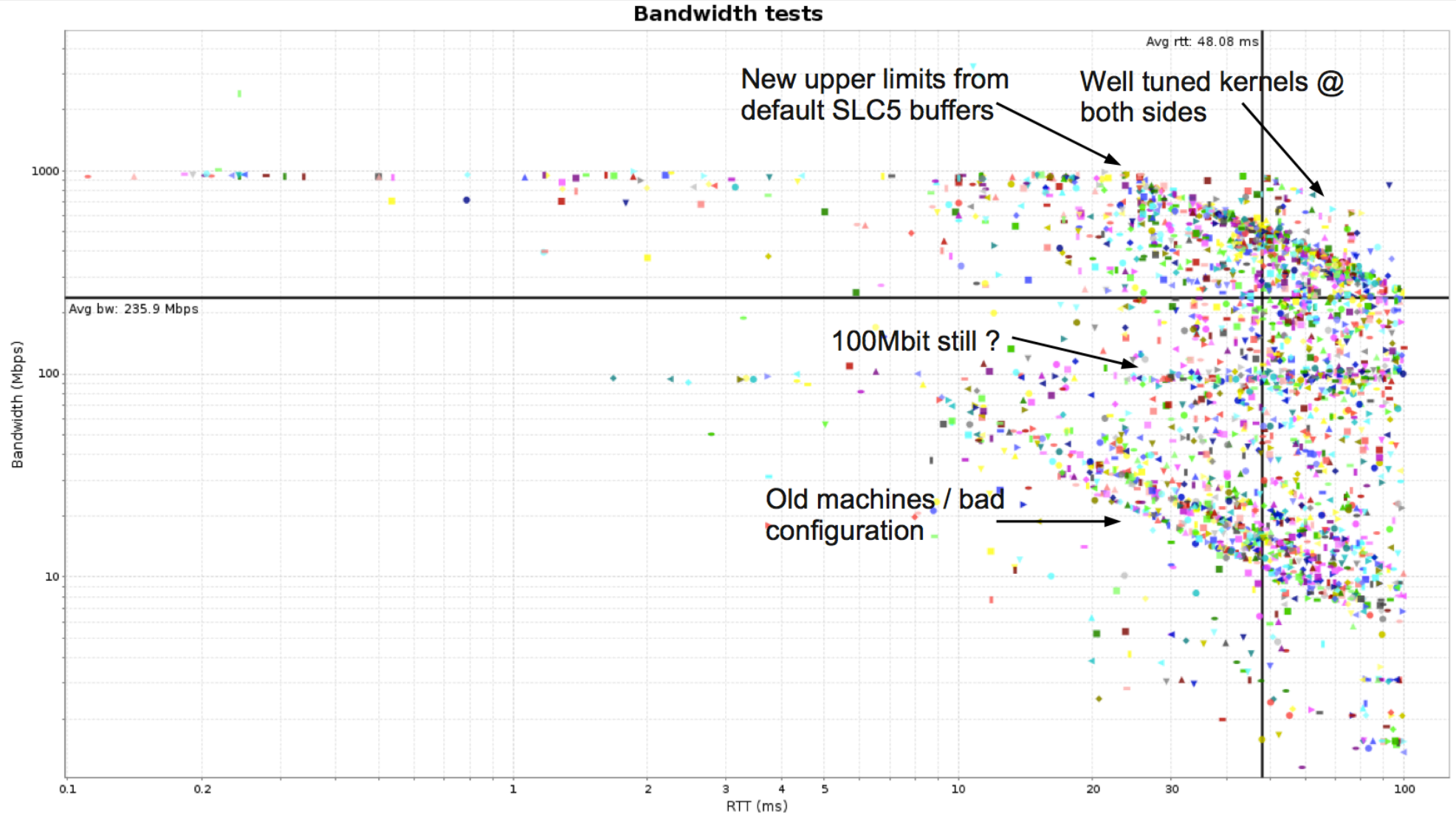
Disk storage elements

	AliEn SE	Statistics						Xrootd info					Functional tests						Last day tests		Demotion
SE Name	AliEn name	Size	Used	Free	Usage	No. of files	Type	Size	Used	Free	Usage	Version	add	ls	get	whereis	rm	Last OK test	Successful	Failed	factor
1. Bari - SE	ALICE::Bari::SE	893.4 TB	128.1 TB	765.3 TB	14.34%	3,019,005	File	1.878 PB	1.775 PB	106.2 TB	94.48%	20100510-1509_dbg						18.05.2011 18:01	12	0	0.298%
2. Bratislava - SE	ALICE::Bratislava::SE	112.8 TB	33.2 TB	79.6 TB	29.44%	985,709	File	112.8 TB	49.08 TB	63.69 TB	43.52%	20100510-1509_dbg						18.05.2011 18:02	12	0	3.869%
3. Catania - SE	ALICE::Catania::SE	100.4 TB	121 TB	-	120.5%	2,459,576	File	158.7 TB	134.4 TB	24.22 TB	84.74%	20100510-1509_dbg						18.05.2011 18:05	12	0	2.679%
4. CCIN2P3 - SE	ALICE::CCIN2P3::SE	112.4 TB	117.5 TB	-	104.5%	2,462,113	File	-	-	-	-							18.05.2011 18:04	12	0	0
5. CCIN2P3 - SE2	ALICE::CCIN2P3::SE2	96 TB	23.53 TB	72.47 TB	24.51%	33,081	File	-	-	-	-							18.05.2011 18:01	12	0	0
6. CERN - ALICEDISK	ALICE::CERN::ALICEDISK	849.6 TB	851.1 TB	-	100.2%	13,510,138	CASTOR	-	-	-	-							18.05.2011 18:04	12	0	0
7. CERN - GLOBAL	ALICE::CERN::GLOBAL	-	0	1.863 TB	-	5,086	root	-	-	-	-							18.05.2011 18:01	12	0	0.298%
8. CERN - SE	ALICE::CERN::SE	20.49 TB	14.04 TB	6.451 TB	68.52%	3,633,414	File	20.46 TB	7.151 TB	13.31 TB	34.95%	20100510-1509_dbg						18.05.2011 18:02	12	0	0
9. Clermont - SE	ALICE::Clermont::SE	179.9 TB	152 TB	27.94 TB	84.47%	3,268,605	File	179.9 TB	175.8 TB	4.024 TB	97.76%	20100510-1509_dbg	Use ...	Last...	Last...	Last...	Last...	29.03.2011 06:01	0	12	100%
10. CNAF - SE	ALICE::CNAF::SE	873.3 TB	493.9 TB	379.4 TB	56.55%	8,430,394	File	873.3 TB	546.3 TB	327 TB	62.55%	20100510-1509_dbg						18.05.2011 18:01	12	0	0
11. CyberSar_Cagliari - SE	ALICE::CyberSar_Cagliari::SE	30.83 TB	33.46 TB	-	108.5%	869,975	File	92.71 TB	88.67 TB	4.035 TB	95.65%	20100510-1509_dbg						18.05.2011 18:03	12	0	0
12. Cyfronet - SE	ALICE::Cyfronet::SE	10 TB	11.69 TB	-	116.9%	519,308	File	9.995 TB	9.556 TB	449.3 GB	95.61%	20100510-1509_dbg						18.05.2011 18:02	12	0	0.298%
13. FZK - SE	ALICE::FZK::SE	1.254 PB	649.1 TB	634.9 TB	50.55%	9,891,737	File	1.252 PB	824.7 TB	457.8 TB	64.3%	20100510-1509_dbg						18.05.2011 18:02	12	0	0.595%
14. Grenoble - DPM	ALICE::Grenoble::DPM	72 TB	6.112 TB	65.89 TB	8.488%	204,323	SRM	-	-	-	-							18.05.2011 18:01	12	0	0
15. Grenoble - SE	ALICE::Grenoble::SE	31 TB	20.09 TB	10.91 TB	64.8%	401,896	File	-	-	-	-		Use ...	Last...	Last...	Last...	Last...	18.04.2011 18:04	0	12	100%
16. GRIF_IPNO - DPM	ALICE::GRIF_IPNO::DPM	85.24 TB	82.96 TB	2.276 TB	97.33%	2,399,219	SRM	-	-	-	-		Use ...	Last...	Last...	Last...	Last...	13.05.2011 00:05	0	12	95.54%
17. GRIF_IPNO - SE	ALICE::GRIF_IPNO::SE	153.1 TB	127.3 TB	25.84 TB	83.13%	3,305,002	File	153.1 TB	150 TB	3.181 TB	97.92%	20100510-1509_dbg	Use ...	Last...	Last...	Last...	Last...	23.03.2011 02:04	0	12	100%
18. GRIF_IRFU - DPM	ALICE::GRIF_IRFU::DPM	171 TB	42.39 TB	128.6 TB	24.79%	782,932	SRM	-	-	-	-							18.05.2011 18:01	12	0	1.19%
19. GSI - SE	ALICE::GSI::SE	312.6 TB	331.5 TB	-	106%	6,196,054	File	297.1 TB	278.9 TB	18.23 TB	93.87%	20100510-1509_dbg	Use ...					18.05.2011 14:01	7	5	36.31%
20. GSI - SE2	ALICE::GSI::SE2	28 TB	459.8 GB	27.55 TB	1.604%	5,144	File	29.08 TB	1.964 TB	27.12 TB	6.751%	20100510-1509_dbg						18.05.2011 18:03	12	0	0
21. HHLR_GU - SE	ALICE::HHLR_GU::SE	200 TB	367.5 GB	199.6 TB	0.179%	5,724	File	-	-	-	-		Use ...	Last...	Last...	Last...	Last...	04.04.2011 14:07	0	12	100%
22. Hiroshima - SE	ALICE::Hiroshima::SE	79 TB	42.03 TB	36.97 TB	53.2%	1,208,761	File	78.78 TB	54.15 TB	24.63 TB	68.73%	20100510-1509_dbg						18.05.2011 18:02	12	0	0
23. IHEP - SE	ALICE::IHEP::SE	35.55 TB	9.147 TB	26.4 TB	25.73%	596,354	File	36.38 TB	9.548 TB	26.83 TB	26.25%	20100510-1509_dbg						18.05.2011 18:06	12	0	0.298%
24. IPNL - SE	ALICE::IPNL::SE	36 TB	50.62 TB	-	140.6%	1,127,971	File	37.3 TB	36.51 TB	801.6 GB	97.9%	20100510-1509_dbg	Use ...	Last...	Last...	Last...	Last...	01.05.2011 00:03	0	12	100%
25. ISS - FILE	ALICE::ISS::FILE	140.5 TB	109.2 TB	31.34 TB	77.69%	3,511,205	File	140.5 TB	134.8 TB	5.722 TB	95.93%	20100510-1509_dbg						18.05.2011 18:01	12	0	0
26. ITEP - SE	ALICE::ITEP::SE	100 TB	41.38 TB	58.62 TB	41.38%	1,099,659	File	99.93 TB	44.75 TB	55.18 TB	44.78%	20100510-1509_dbg						18.05.2011 18:03	12	0	0.298%
27. JINR - SE	ALICE::JINR::SE	112.3 TB	80.44 TB	31.87 TB	71.62%	3,564,624	File	149.1 TB	85.9 TB	63.2 TB	57.61%	20100510-1509_dbg						18.05.2011 18:01	12	0	0.298%
28. KFKI - SE	ALICE::KFKI::SE	39.34 TB	27.02 TB	12.32 TB	68.68%	780,936	File	36.38 TB	35.32 TB	1.055 TB	97.1%	20100510-1509_dbg						18.05.2011 18:03	12	0	0
29. KISTI_GSDC - SE	ALICE::KISTI_GSDC::SE	100 TB	31.88 TB	68.12 TB	31.88%	827,924	File	101.8 TB	47.89 TB	53.88 TB	47.06%	20100510-1509_dbg						18.05.2011 18:03	11	1	10.42%
30. KISTI - SE	ALICE::KISTI::SE	49.95 TB	35.49 TB	14.46 TB	71.05%	1,049,511	File	49.95 TB	35.39 TB	14.55 TB	70.86%	20100510-1509_dbg						18.05.2011 18:05	12	0	0
31. Kolkata - SE	ALICE::Kolkata::SE	73.24 TB	15.62 TB	57.62 TB	21.33%	502,177	File	70.46 TB	33.97 TB	36.48 TB	48.22%	20100510-1509_dbg						18.05.2011 18:03	12	0	0
32. Kosice - SE	ALICE::Kosice::SE	41.84 TB	34.15 TB	7.691 TB	81.62%	1,020,050	File	61.84 TB	46.61 TB	15.23 TB	75.37%	20100115.1117_dbg						18.05.2011 18:02	12	0	0.595%

AVAILABLE BANDWIDTH VS RTT



NETWORK WITHIN CONTINENTS



OTHER INTERESTING ITEMS

- Data staging to AAF
  - Virtual Mass Storage by xrootd
  - Automatic staging or SE access?
- Xrootd advanced features
  - Not all are used by ALICE
- Catalogue optimization
  - More files than foreseen
  - Automatic “crawlers” to ensure consistency
- …

CONCLUSIONS

- The ALICE Computing model is evolving
  - Remote access to data is becoming an unavoidable, and maybe desirable, feature
- The general model works
  - Both for heavy scheduled activity (PbPb processing), massive chaotic analysis (QM preparation) and combinations of the two.
  - Also, sites performed brilliantly
- The storage was able to cope with the load
  - Never tested before in this I/O regime
- However, there is much room for improvement
  - User tasks show markedly low CPU/Wall time efficiency and high memory footprint
  - Will probably need some new tools to diagnose new and exciting large-scale effects

ALICE