ALICE STORAGE MODEL: HOW IT WORKS AND HOW IT IS WORKING DURING DATA TAKING

Stefano Bagnasco

INFN Torino

OUTLINE

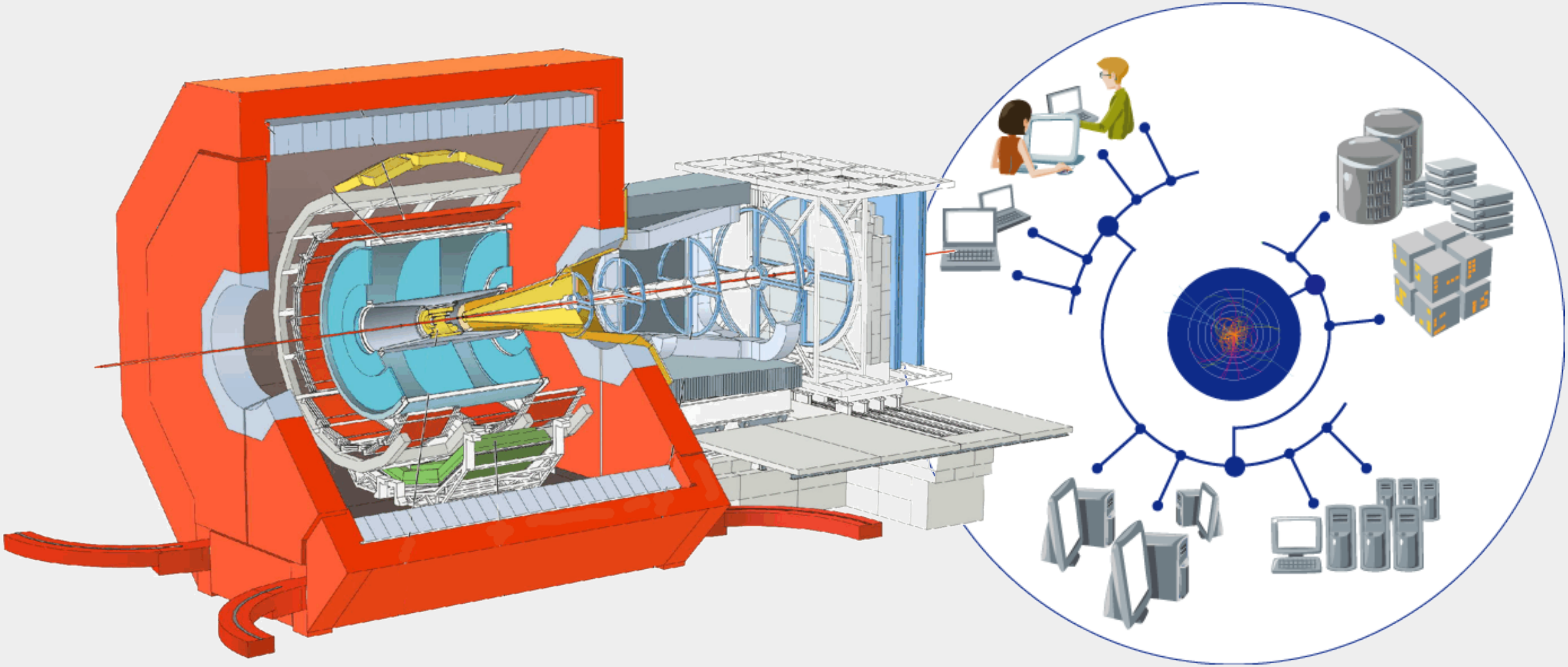
- The ALICE Computing Model
  - In a nuthell

- The ALICE Storage Model
  - In practice
  - Plus some recent activity

- Some Plots

ACKNOWLEDGEMENTS

- **Slides etc. stolen from:**
  - Fabrizio Furano
  - Patricia Mendez Lorenzo
  - Costin Grigoras
  - Jean-Michel Barbet
  - Latchezar Betev
  - Galina Shabratova
  - MonALISA monitoring



THE ALICE COMPUTING MODEL

Stefano Bagnasco – INFN Torino

THE ALICE COMPUTING MODEL

- For **pp** similar to the other experiments
  - Quasi-online data distribution and first reconstruction at T0
  - Further reconstructions at T1's

- For **AA** different model
  - Calibration, alignment, pilot reconstructions and partial data export during data taking
  - Data distribution and first reconstruction at T0 in the four months after AA run (shutdown)
  - Further reconstructions at T1's

THE ALICE COMPUTING MODEL

- Three kinds of data analysis
  - **Fast pilot analysis** of the data “just collected” to tune the first reconstruction at CERN Analysis Facility (CAF)
  - **Scheduled batch analysis** on the Grid (ESDs and AODs)
  - **End-user interactive or batch analysis** using PROOF and GRID (AODs and ESDs)

- **T0 (CERN)**
  - Does: first pass reconstruction; calibration and alignment
  - Stores: one copy of RAW, calibration data and first-pass ESDs
- **T1s**
  - 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
- **T2s**
  - Does: simulation and end-user analysis
  - Stores: disk replicas of AODs and ESDs

THE COMPONENTS

- AliRoot
    - ROOT + Geant3 + ...
- AliEn
    - Data catalogue
    - Job management
- Xrootd
    - Data access
- MonALISA
    - Monitoring
- Underlying infrastructure
    - LCG/INFNGrid
        - But also OSG, NorduGrid,... that use different middleware

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
- Jobs are assigned where data is located
  - All policies (data & CPU) enforced on central servers
  - WMS efficiency not a big issue thanks to JAs strategy
- Resources are shared
  - No “localization” of groups
  - Fairshare Group/Site Contribution and Consumption *will* be regulated by accounting system
  - Prioritisation of jobs in the central ALICE queue
- Data access only through the GRID
  - No backdoor access to data
  - No “private” processing on shared resources
  - No “private” resources outside of the grid

ALIEN COMPONENTS

Job management

- Task Queue
  - Database of all submitted jobs
  - Keeps track of status, etc.
- Job optimizers
  - Run on the TQ
  - Enforce policies, split jobs, etc.
- Job Agents
  - Run jobs on sites
- Cluster Monitor
  - Site service working as a proxy for Job Agents

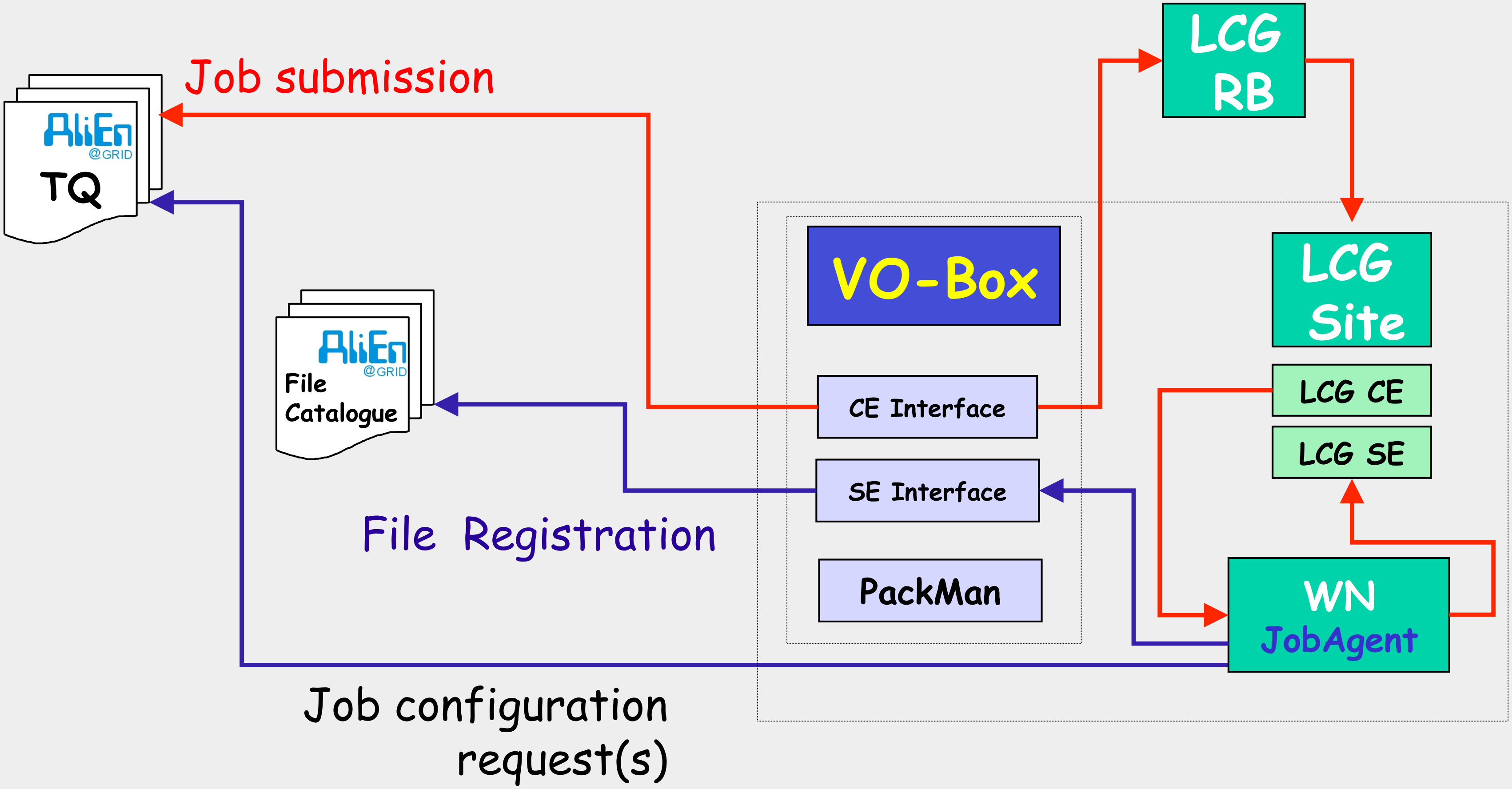
Data management

- File Catalogue
  - With metadata
- File Transfer Service
  - Similar to the Task Queue
  - Uses FTS or xrootd
- Storage Element
  - Not really a piece of AliEn
  - Several “flavours” exist
- Package Manager
  - Did not know where to put this

JOB SUBMISSION KEY CONCEPTS

- Task Queue and Optimizers
  - Central DB of jobs to be executed
  - Optimizers split and arrange jobs according to input data, priority policies and/or user defined criteria
- Site VO-Box
  - Thin interfaces to underlying Grid site services
  - Submits JobAgents to site
  - Takes care of proxy management
- JobAgent
  - Runs on WNs, downloads payload from the TQ and executes it
  - JAs create a “virtual grid” on top of existing Grid infrastructures

THE VO BOX



SB's talk @ almost everywhere

STORAGE KEY POINTS

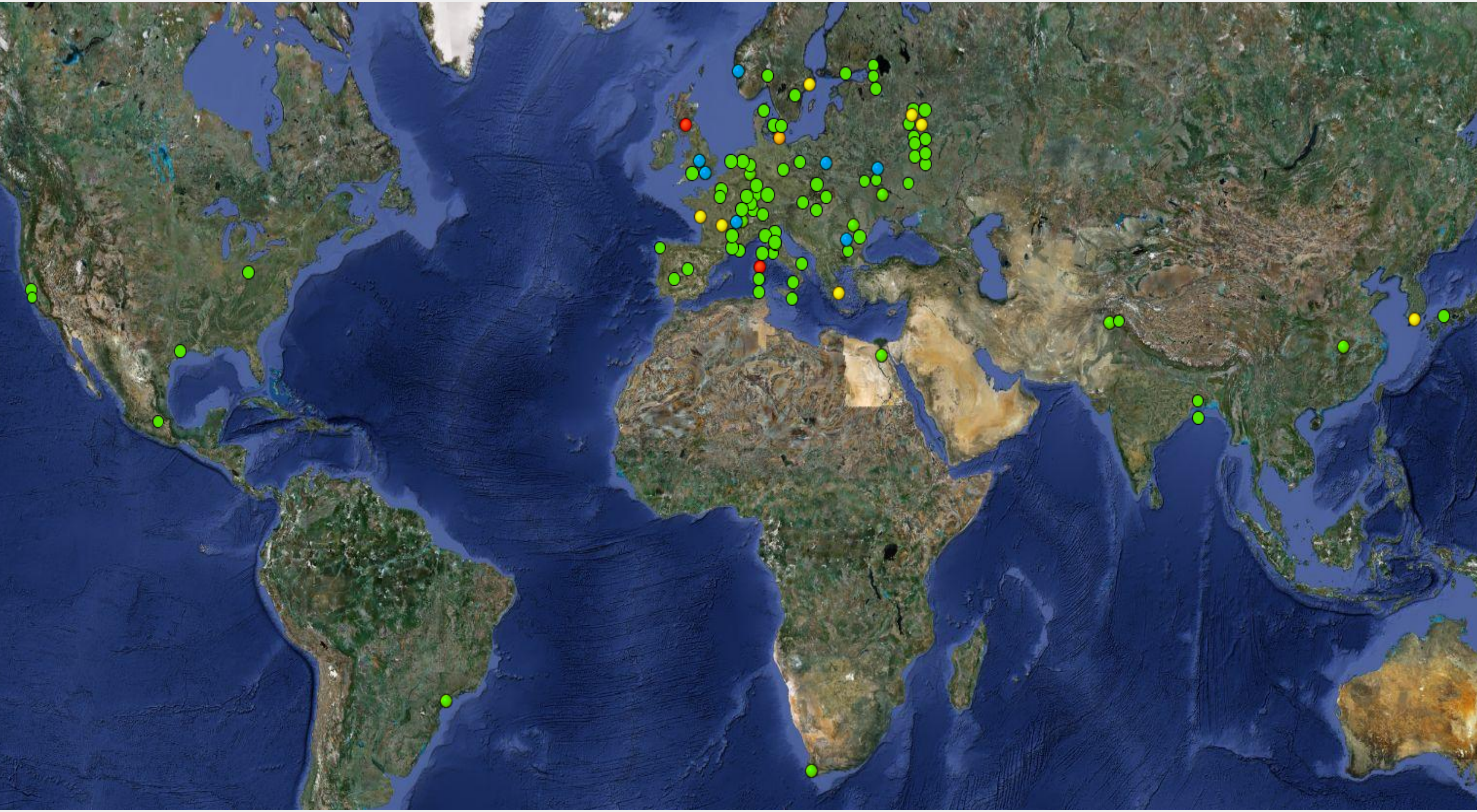
- Move jobs, not data
  - Well, mostly

- Central file and data catalogue

- All permissions, quotas and policies enforced centrally
  - No distributed policies

- All data access through xrootd
  - See next section

SITES OVERVIEW

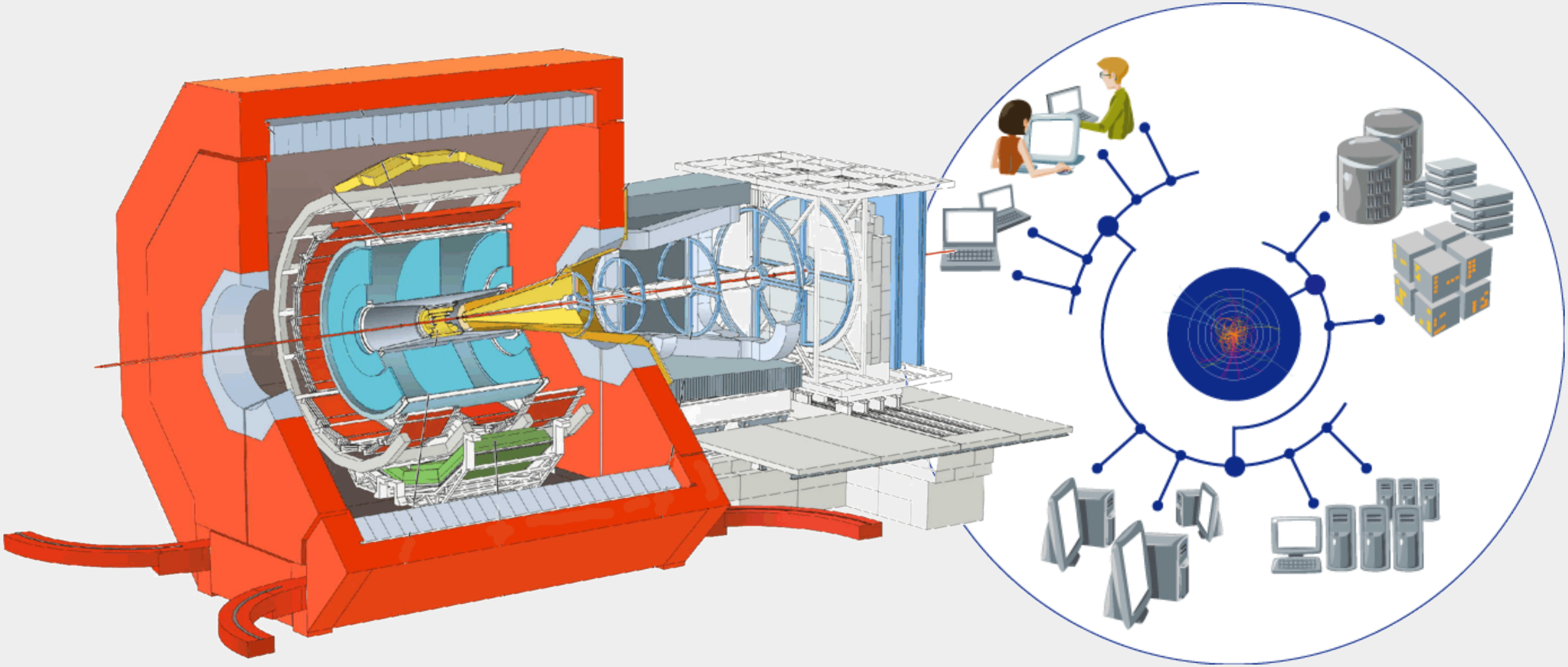


105 VOBoxes in 83 centers, >22000 CPU cores
55 Storage Elements

http://pcalimonitor.cern.ch/

ALICE

Stefano Bagnasco – INFN Torino



ALICE AND XROOTD

XROOTD RATIONALE

- 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
    - ALWAYS use LFN to reference file
- Proven performance, stability and scalability
  - ALICE uses xrootd native servers for some of the most critical data management tasks:
    - Conditions data on the Grid
    - Configuration macros for production and analysis
- “Global redirector” + xrd3cp
  - Xrootd has a highly optimized “WAN mode”
  - Torrent-like “extreme copy”
  - See next slides

- Having an unique WAN+LAN compliant protocol allows to do the right thing
  - Exploit locality whenever possible (=most of the times)
  - Do not worry too much if a job accesses some data files which is not in the same site. This has to be possible and foreseen.
    - Explicitly creating 100s of replicas takes much more time and risk.
  - Access condition data ONLY via WAN

F. Furano, July 2010

XROOTD

- Each server manages a portion of the storage
  - many servers with small disks, or
  - fewer servers with huge disks

- Low overhead DB-free aggregation of servers
  - Gives the functionalities of an unique thing
  - A non-transactional file system

- Efficient LAN/WAN byte-level data access
- Protocol/architecture built on the tough HEP requirements

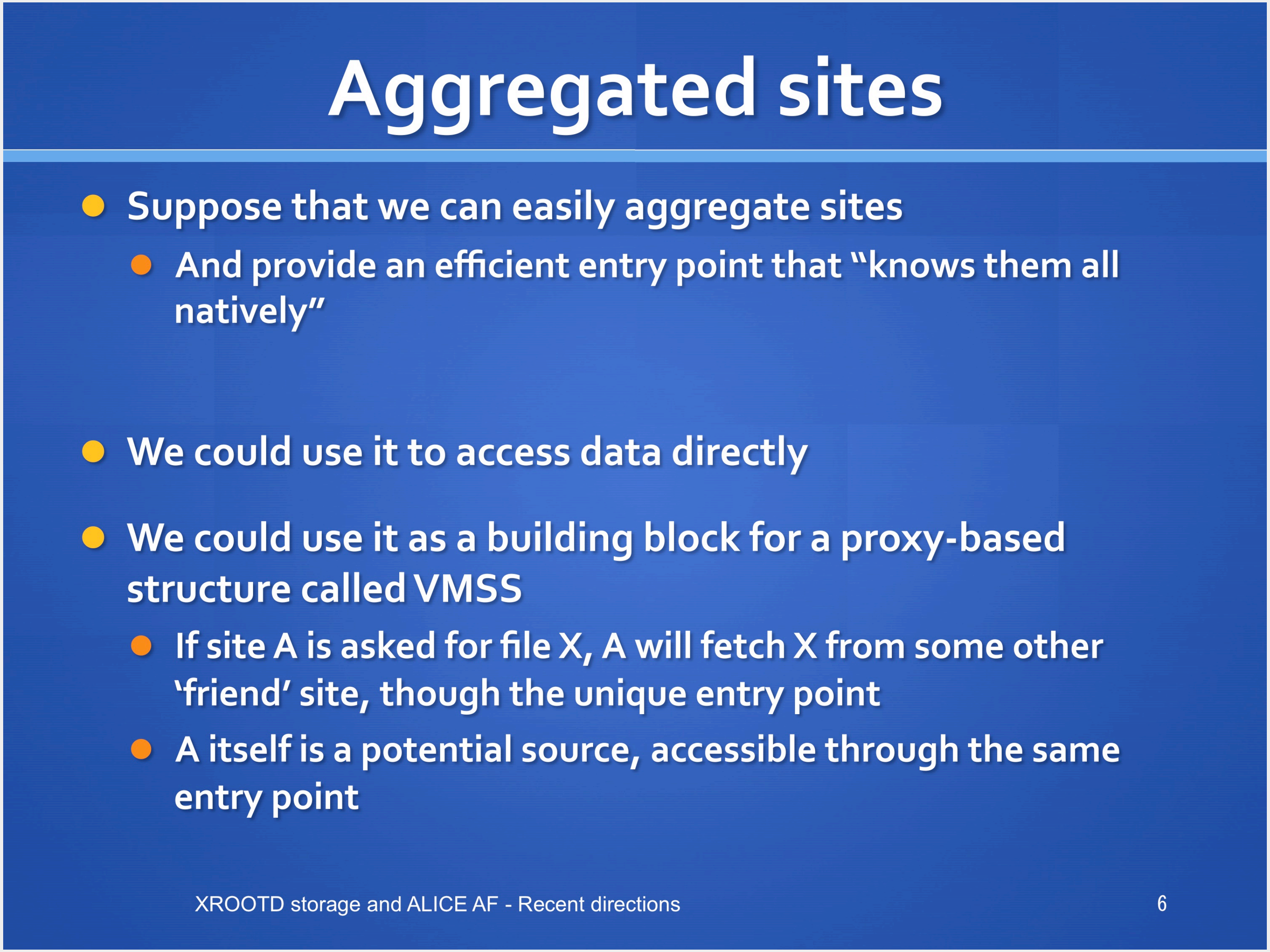
F. Furano, July 2010

DEPLOYMENT

- At CERN
  - xCastor2
  - https://twiki.cern.ch/twiki/bin/view/DataManagement/X2CASTOR
- At Tier-1s
  - Usually over parallel FS + hierarchical SM
  - CNAF Example: GPFS+TSM+Xrootd
  - dCache + xrootd emulation in Germany
- At Tier-2s
  - Sometimes over parallel FS
  - Lustre, GPFS
  - StorM to provide SRM access

- A mix of xrootd-only storages and DPM enabled xrootd (half a dozen of sites)
  - DPM/xrootd was the solution for small sites
  - Later Alice pushed xrootd-only
  - Sites supporting several VO would prefer to have only one solution for all Vos
    - To have less different services to maintain
    - To spread traffic on more servers
  - Many sites moved nevertheless from DPM/xrootd to xrootd-only
- On dCache, xrootd is emulated
  - Protocol re-written in Java

GLOBAL REDIRECTOR



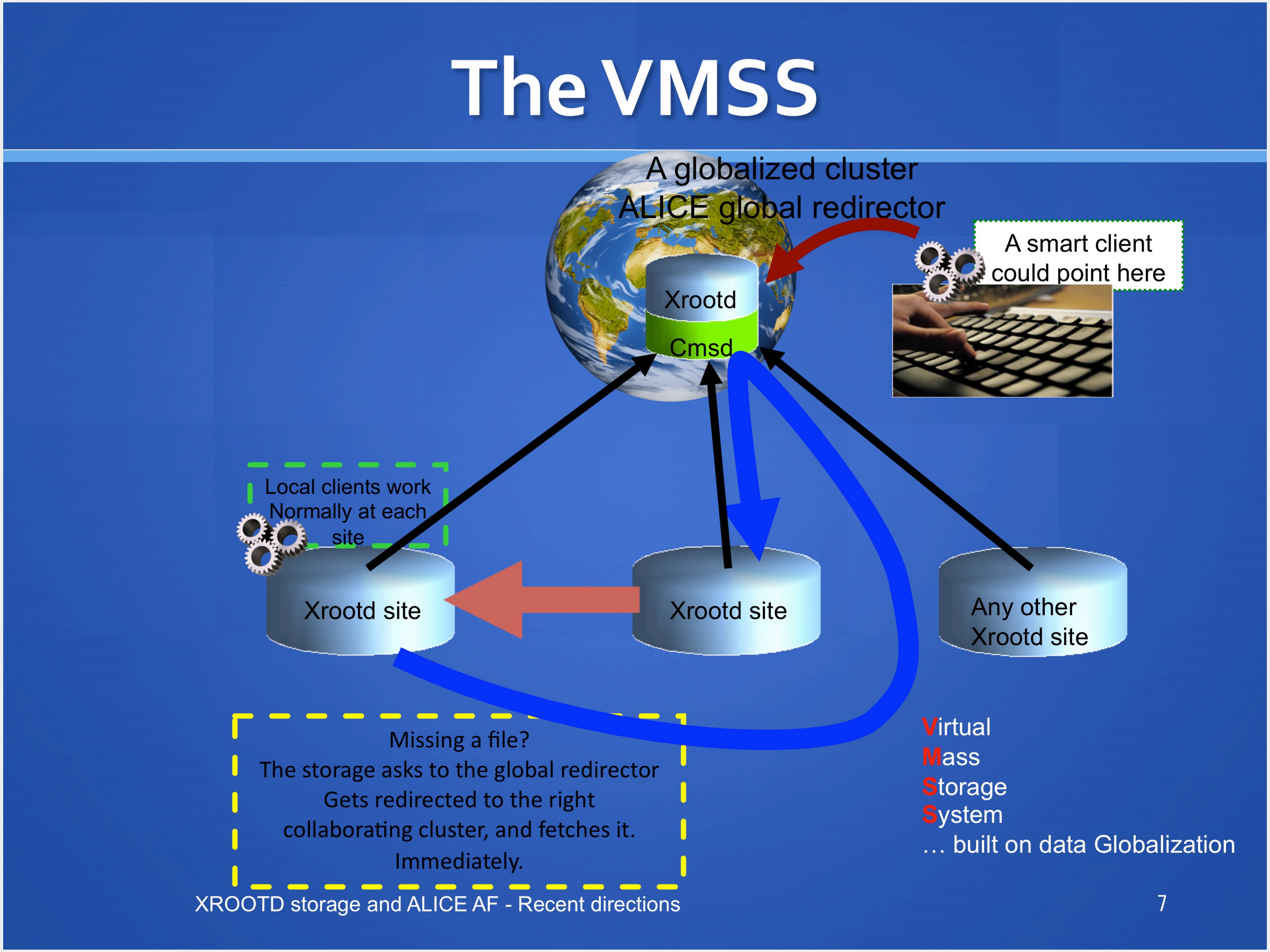
Aggregated sites

- Suppose that we can easily aggregate sites
  - And provide an efficient entry point that "knows them all natively"
- We could use it to access data directly
- We could use it as a building block for a proxy-based structure called VMSS
  - If site A is asked for file X, A will fetch X from some other 'friend' site, though the unique entry point
  - A itself is a potential source, accessible through the same entry point

XROOTD storage and ALICE AF - Recent directions 6

F. Furano, July 2010

GLOBAL REDIRECTOR



The VMSS

Missing a file?
The storage asks to the global redirector
Gets redirected to the right collaborating cluster, and fetches it.
Immediately.

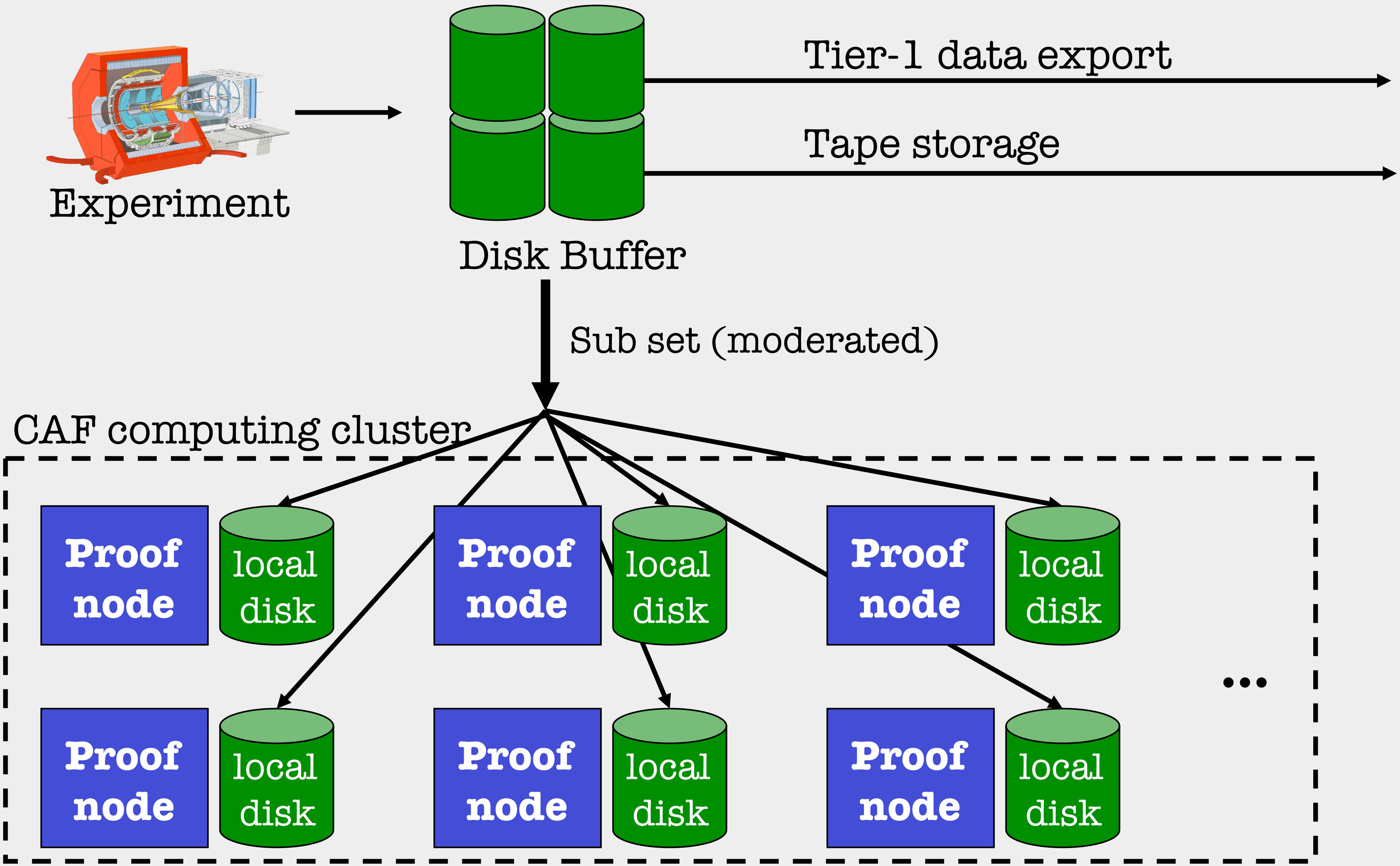
Virtual
Mass
Storage
System
… built on data Globalization

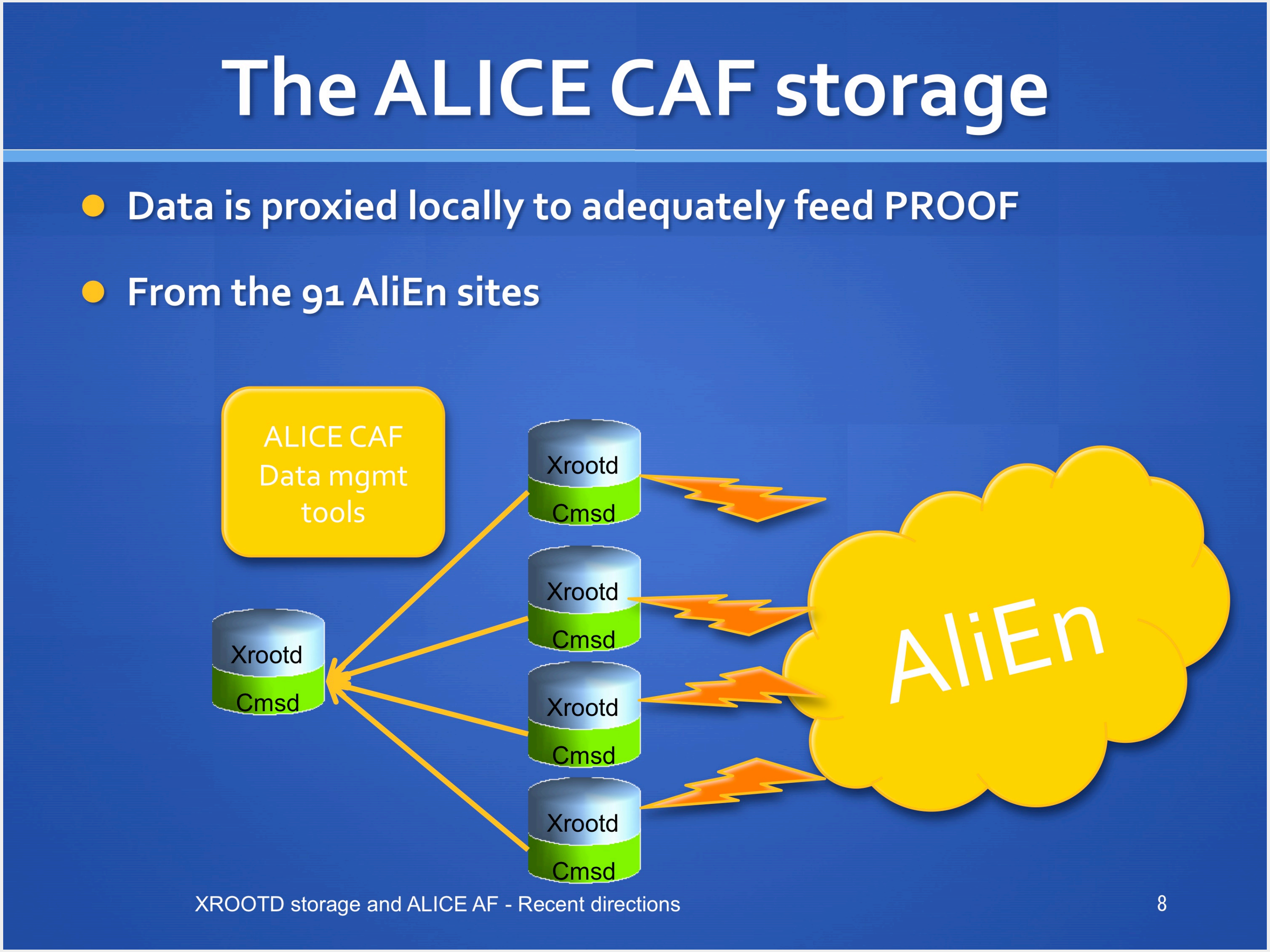
XROOTD storage and ALICE AF - Recent directions

7

F. Furano, July 2010

THE ORIGINAL CAF





The ALICE CAF storage

- Data is proxied locally to adequately feed PROOF
- From the 91 AliEn sites

ALICE CAF
Data mgmt
tools

Xrootd
Cmsd

Xrootd
Cmsd

Xrootd
Cmsd

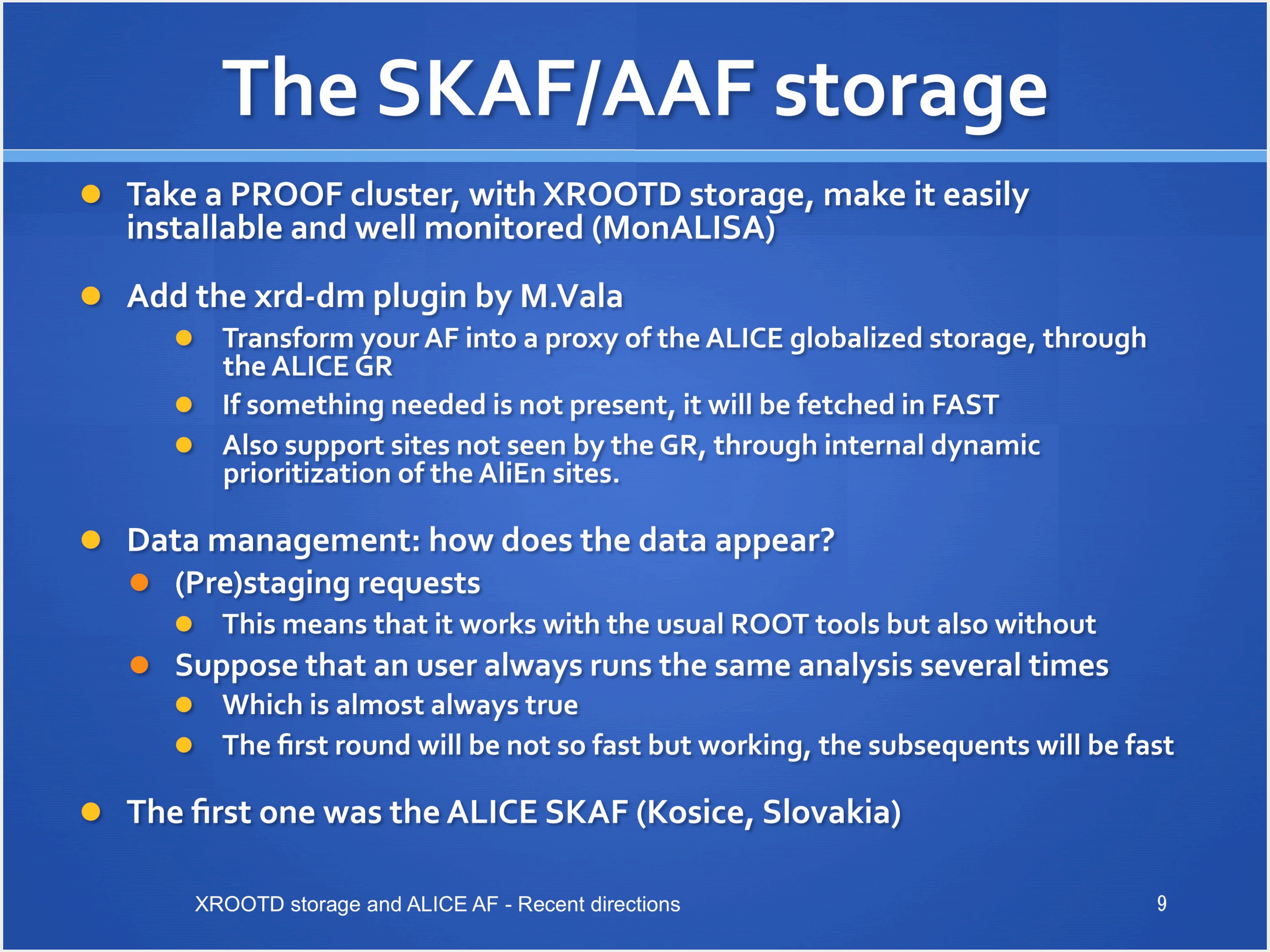
Xrootd
Cmsd

Xrootd
Cmsd

AliEn

F. Furano, July 2010

INTERACTIVE ANALYSIS FACILITIES

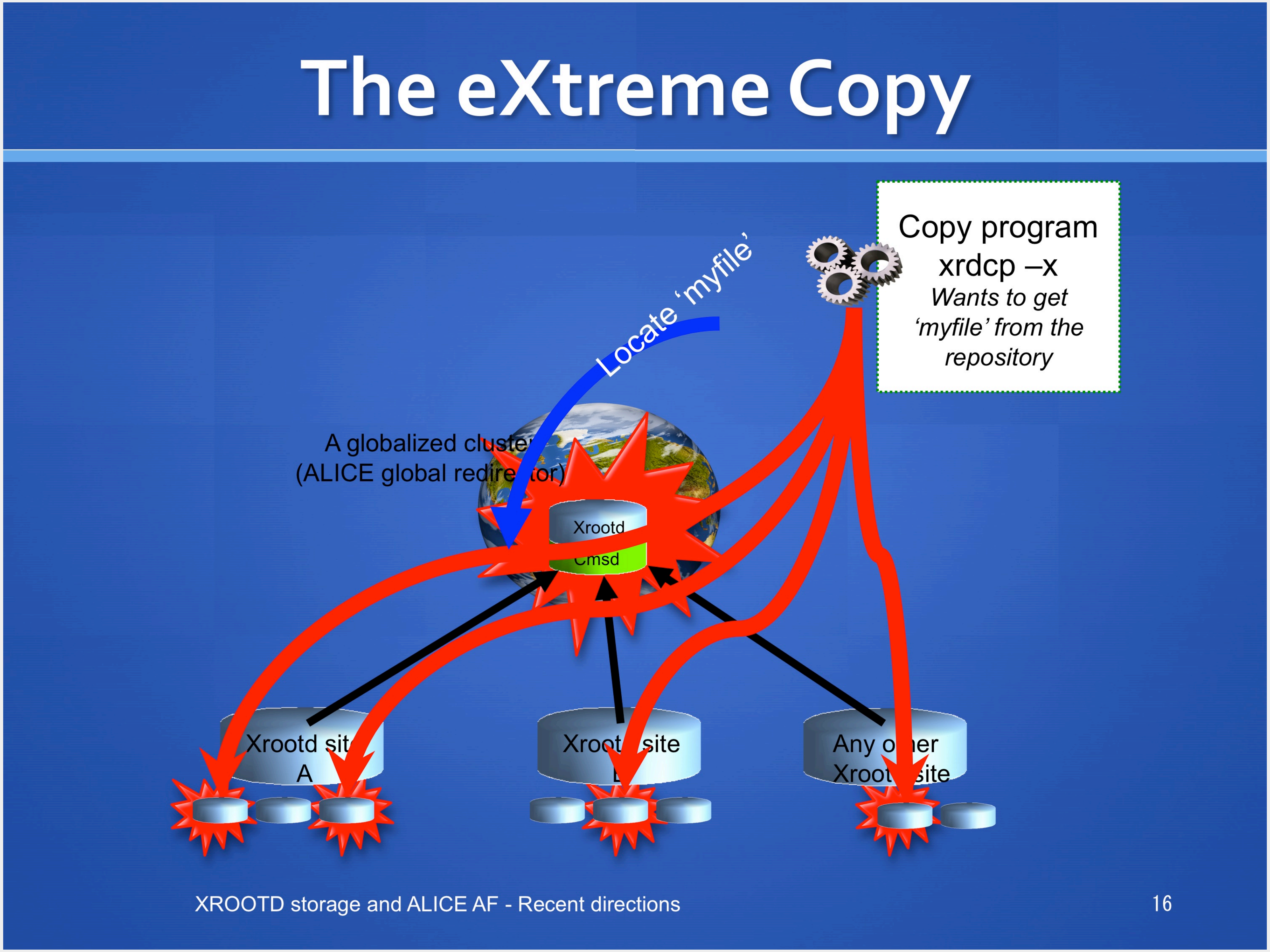


The SKAF/AAF storage

- Take a PROOF cluster, with XROOTD storage, make it easily installable and well monitored (MonALISA)
- Add the xrd-dm plugin by M.Vala
    - Transform your AF into a proxy of the ALICE globalized storage, through the ALICE GR
    - If something needed is not present, it will be fetched in FAST
    - Also support sites not seen by the GR, through internal dynamic prioritization of the AliEn sites.
- Data management: how does the data appear?
    - (Pre)staging requests
        - This means that it works with the usual ROOT tools but also without
    - Suppose that an user always runs the same analysis several times
        - Which is almost always true
        - The first round will be not so fast but working, the subsequents will be fast
- The first one was the ALICE SKAF (Kosice, Slovakia)

F. Furano, July 2010

XTREME COPY



The eXtreme Copy

Copy program
xrdcp –x
Wants to get 'myfile' from the repository

Locate 'myfile'

A globalized cluste
(ALICE global redire tor)

Xrootd
Cmsd

Xrootd sit
A

Xroot site

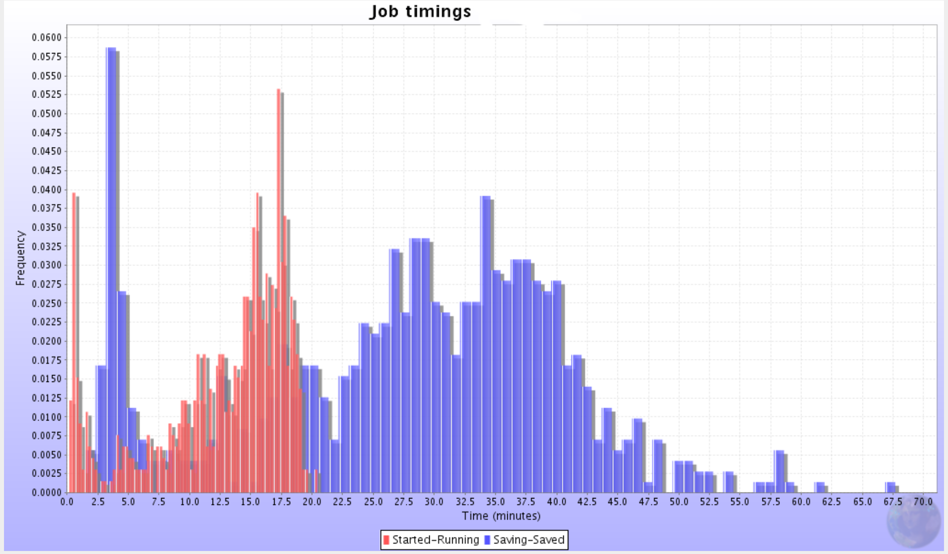
Any o er
Xroot site

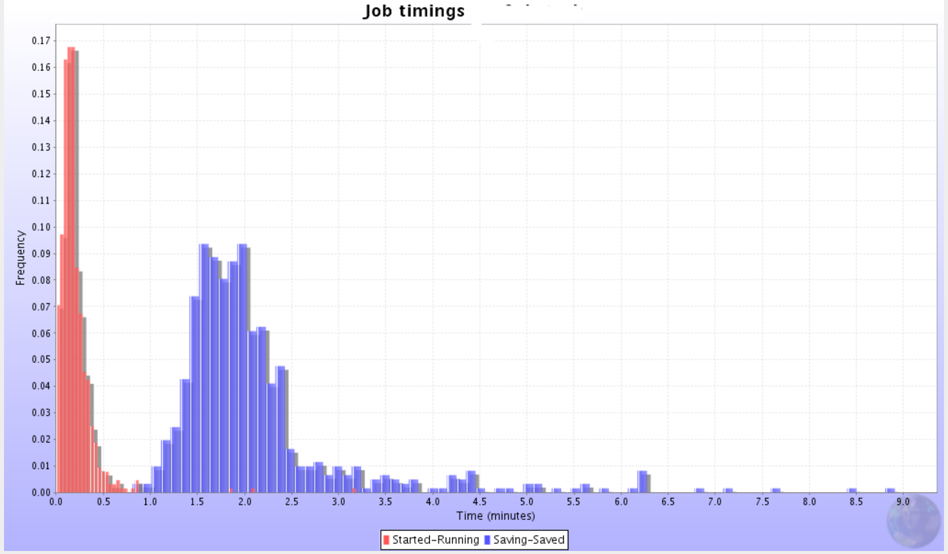
XROOTD storage and ALICE AF - Recent directions

16

F. Furano, July 2010

BAD SITES AND GOOD SITES





J.-M. Barbet, Septemper 2010

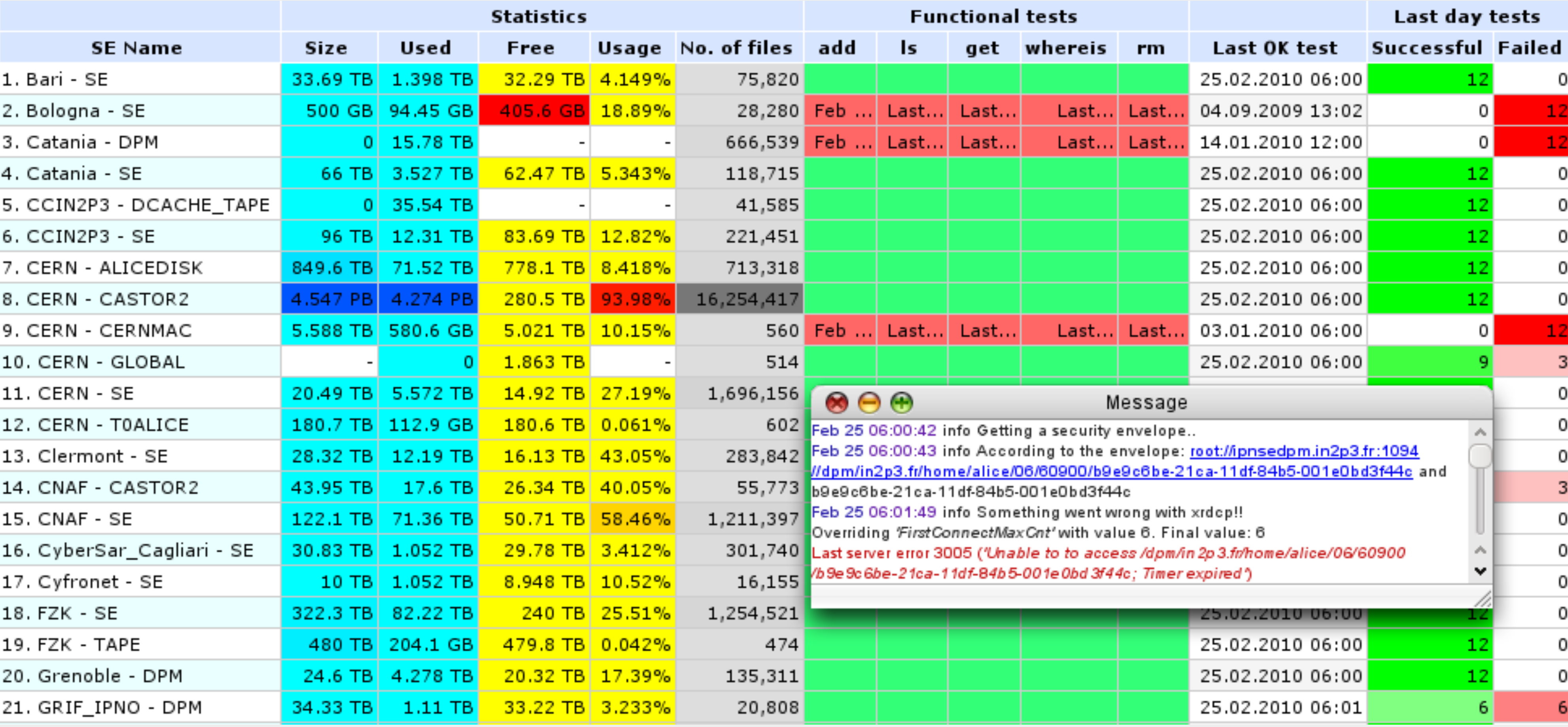
FIRST, AVAILABILITY

- To simplify the decision we first remove the problematic storages from the options

- Periodic functional tests of all known SEs
  - currently every 2h
  - add, get, remove of a test file from a remote location

- The status of an SE can be also set by the administrators

STORAGE MONITORING



	Statistics					Functional tests						Last day tests	
SE Name	Size	Used	Free	Usage	No. of files	add	ls	get	whereis	rm	Last OK test	Successful	Failed
1. Bari - SE	33.69 TB	1.398 TB	32.29 TB	4.149%	75,820						25.02.2010 06:00	12	0
2. Bologna - SE	500 GB	94.45 GB	405.6 GB	18.89%	28,280	Feb ...	Last...	Last...	Last...	Last...	04.09.2009 13:02	0	12
3. Catania - DPM	0	15.78 TB	-	-	666,539	Feb ...	Last...	Last...	Last...	Last...	14.01.2010 12:00	0	12
4. Catania - SE	66 TB	3.527 TB	62.47 TB	5.343%	118,715						25.02.2010 06:00	12	0
5. CCIN2P3 - DCACHE_TAPE	0	35.54 TB	-	-	41,585						25.02.2010 06:00	12	0
6. CCIN2P3 - SE	96 TB	12.31 TB	83.69 TB	12.82%	221,451						25.02.2010 06:00	12	0
7. CERN - ALICEDISK	849.6 TB	71.52 TB	778.1 TB	8.418%	713,318						25.02.2010 06:00	12	0
8. CERN - CASTOR2	4.547 PB	4.274 PB	280.5 TB	93.98%	16,254,417						25.02.2010 06:00	12	0
9. CERN - CERNMAC	5.588 TB	580.6 GB	5.021 TB	10.15%	560	Feb ...	Last...	Last...	Last...	Last...	03.01.2010 06:00	0	12
10. CERN - GLOBAL	-	0	1.863 TB	-	514						25.02.2010 06:00	9	3
11. CERN - SE	20.49 TB	5.572 TB	14.92 TB	27.19%	1,696,156								0
12. CERN - T0ALICE	180.7 TB	112.9 GB	180.6 TB	0.061%	602								0
13. Clermont - SE	28.32 TB	12.19 TB	16.13 TB	43.05%	283,842								0
14. CNAF - CASTOR2	43.95 TB	17.6 TB	26.34 TB	40.05%	55,773								3
15. CNAF - SE	122.1 TB	71.36 TB	50.71 TB	58.46%	1,211,397								0
16. CyberSar_Cagliari - SE	30.83 TB	1.052 TB	29.78 TB	3.412%	301,740								0
17. Cyfronet - SE	10 TB	1.052 TB	8.948 TB	10.52%	16,155								0
18. FZK - SE	322.3 TB	82.22 TB	240 TB	25.51%	1,254,521						25.02.2010 06:00	12	0
19. FZK - TAPE	480 TB	204.1 GB	479.8 TB	0.042%	474						25.02.2010 06:00	12	0
20. Grenoble - DPM	24.6 TB	4.278 TB	20.32 TB	17.39%	135,311						25.02.2010 06:00	12	0
21. GRIF_IPNO - DPM	34.33 TB	1.11 TB	33.22 TB	3.233%	20,808						25.02.2010 06:01	6	6

```
Feb 25 06:00:42 info Getting a security envelope..
Feb 25 06:00:43 info According to the envelope: root://ipnsedpm.in2p3.fr:1094
//dpm/in2p3.fr/home/alice/06/60900/b9e9c6be-21ca-11df-84b5-001e0bd3f44c and
b9e9c6be-21ca-11df-84b5-001e0bd3f44c
Feb 25 06:01:49 info Something went wrong with xrdcp!!
Overriding 'FirstConnectMaxCnt' with value 6. Final value: 6
Last server error 3005 ('Unable to to access /dpm/in2p3.fr/home/alice/06/60900
/b9e9c6be-21ca-11df-84b5-001e0bd3f44c; Timer expired')
```

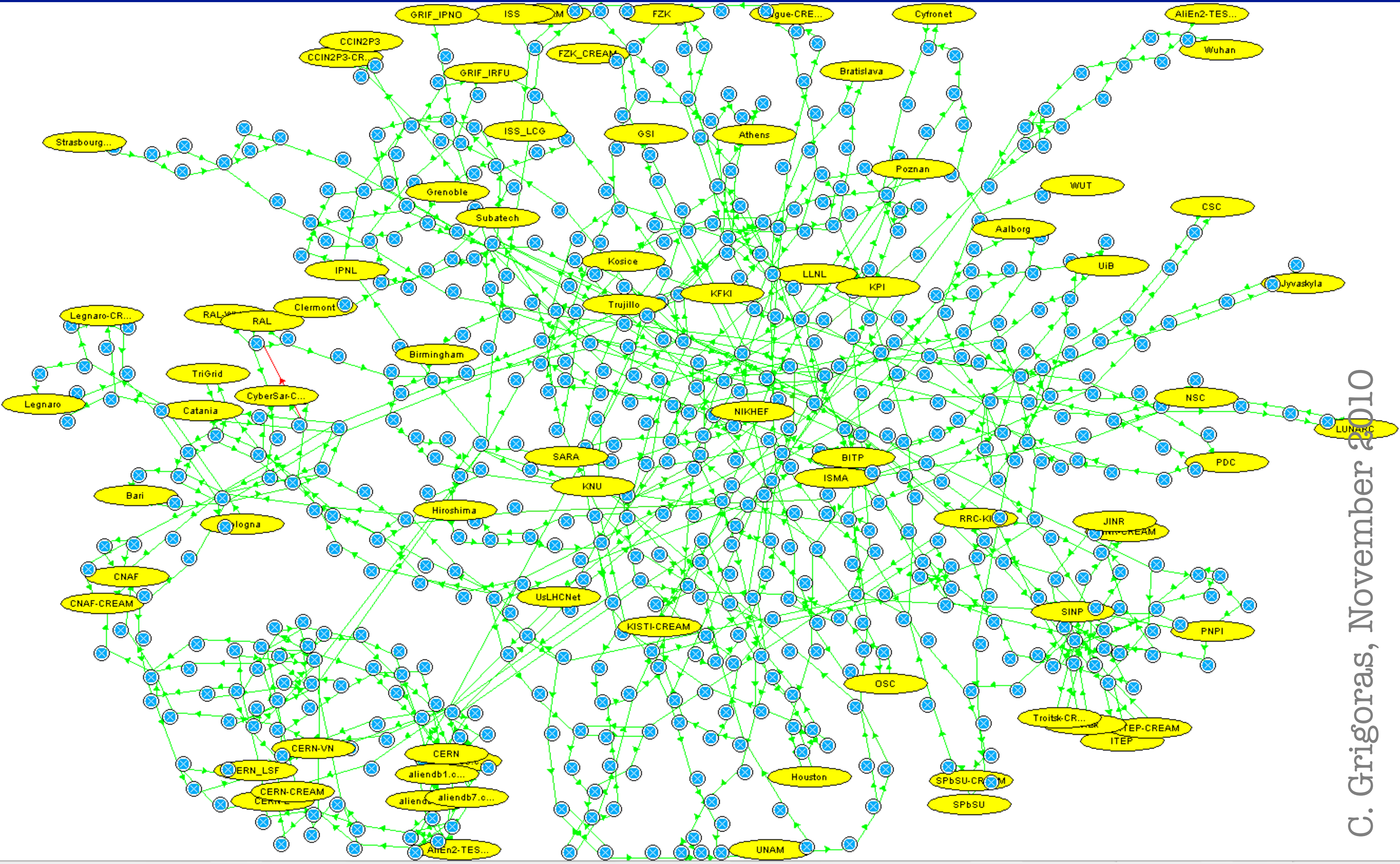
C. Grigoras, Nov 2010

SECOND, TOPOLOGY

- Each SE is associated a set of IP addresses
  - The IP of the VOBox
  - IPs of xrootd redirector & nodes
- tracepath/traceroute tests between all VOBoxes
  - By MonALISA
- Recording all routers and the RTT of each link
  - + status of storage nodes
  - + bandwidth tests between sites

C. Grigoras, November 2010

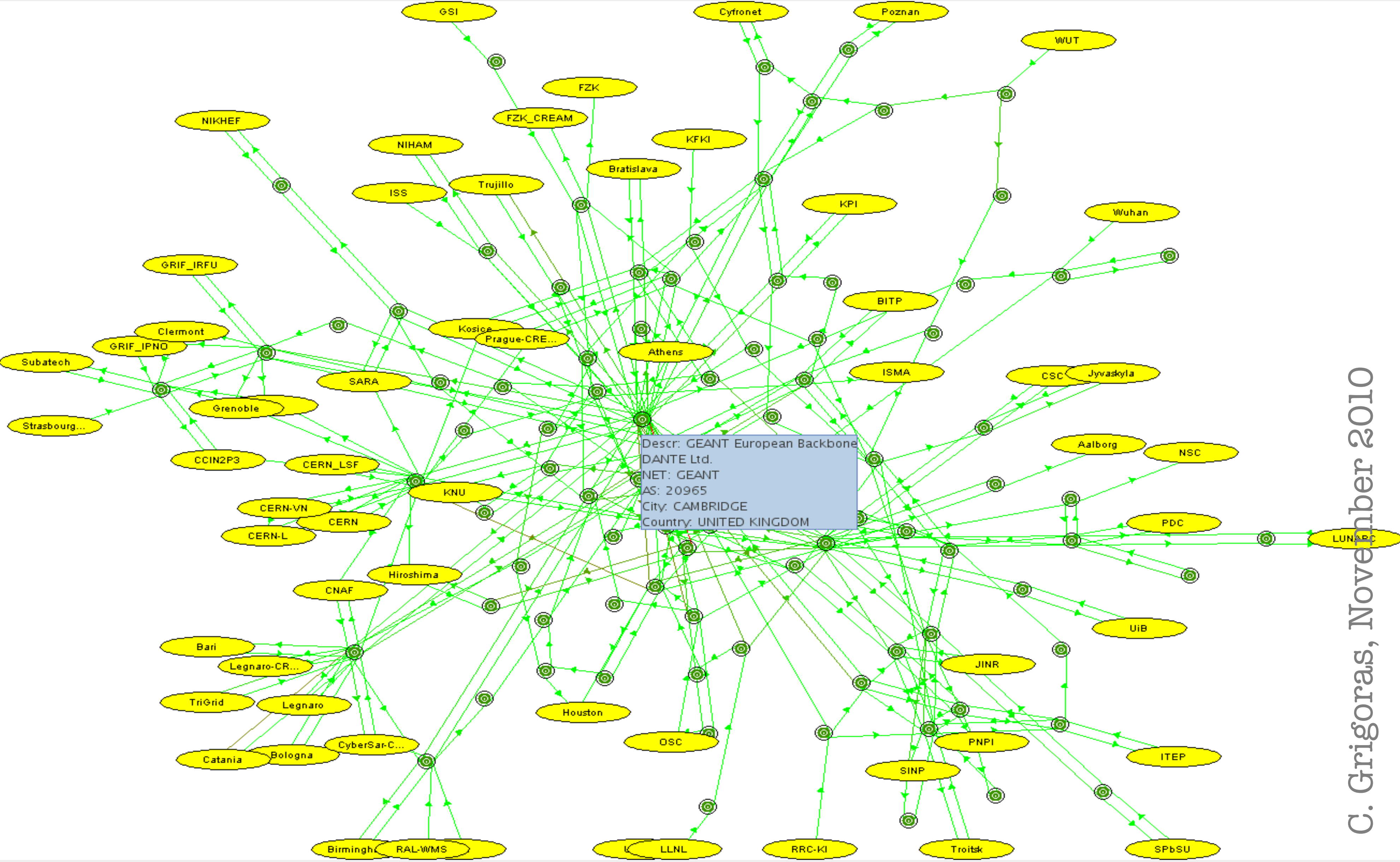
ALICE RAW NETWORK TOPOLOGY



C. Grigoras, November 2010

Stefano Bagnasco – INFN Torino

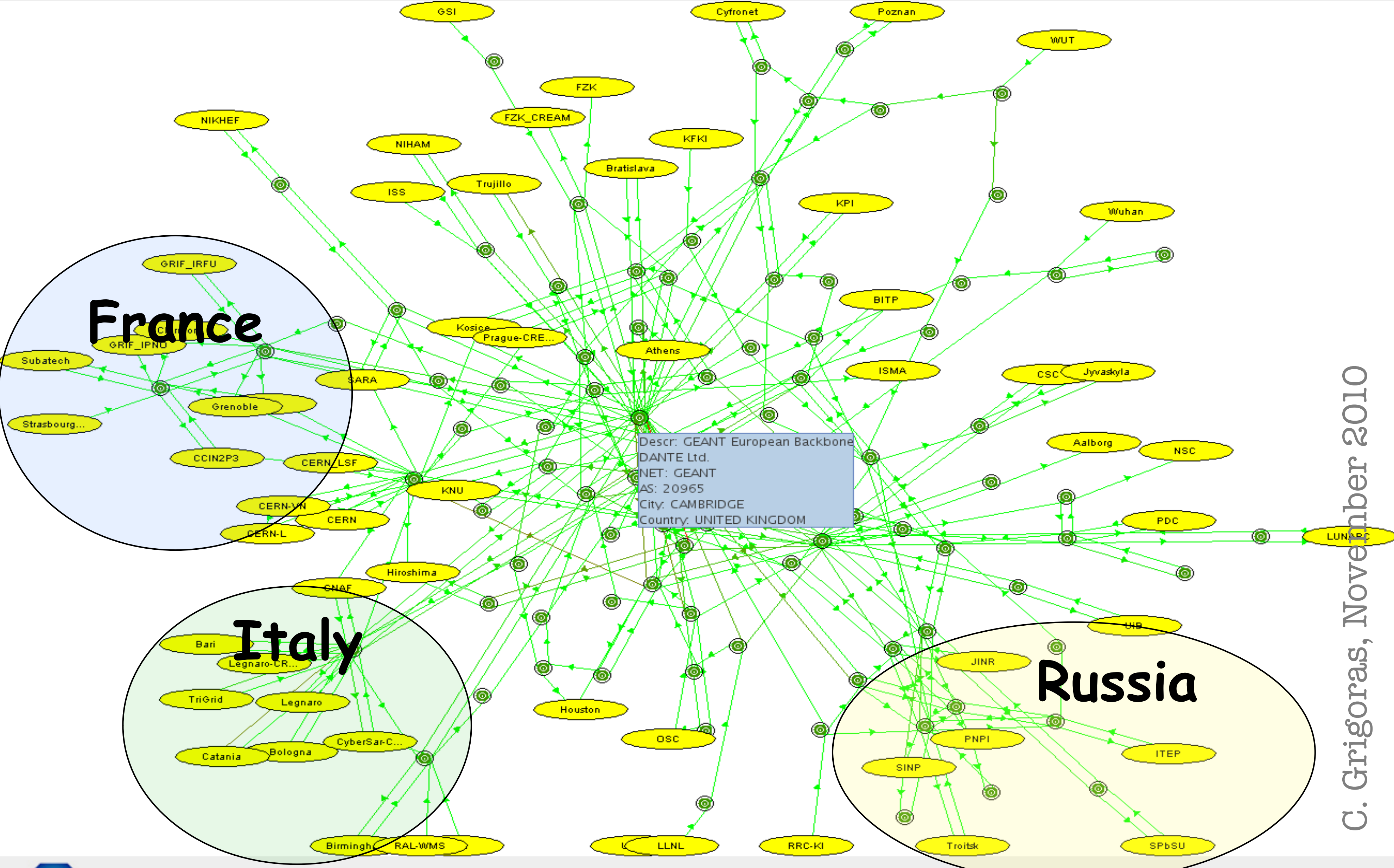
DERIVED NETWORK TOPOLOGY



Descr: GEANT European Backbone
DANTE Ltd.
NET: GEANT
AS: 20965
City: CAMBRIDGE
Country: UNITED KINGDOM

C. Grigoras, November 2010

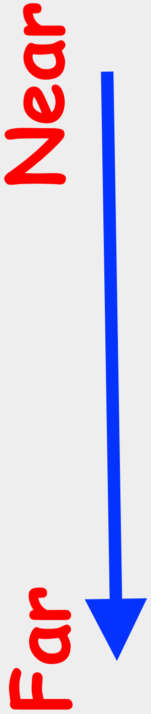
DERIVED NETWORK TOPOLOGY



Descr: GEANT European Backbone
DANTE Ltd.
NET: GEANT
AS: 20965
City: CAMBRIDGE
Country: UNITED KINGDOM

C. Grigoras, November 2010

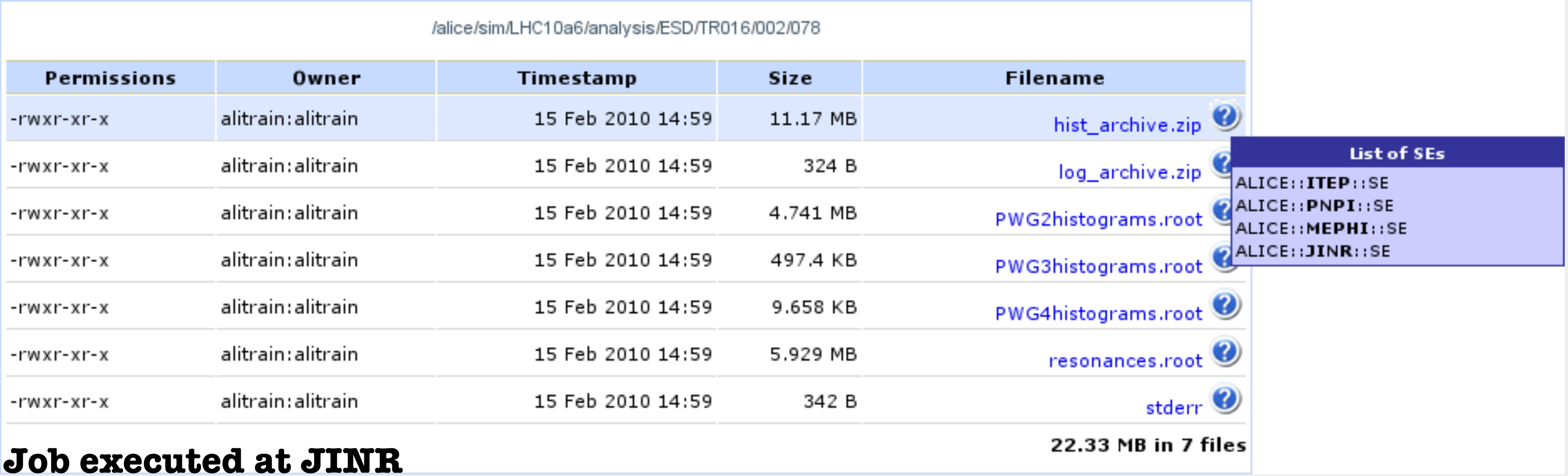
DEFINE A DISTANCE



- Same C-class network
- Common domain name
- Same AS
- Same country (+ function of RTT between the respective AS-es if known)
- If distance between the AS-es is known, use it
- Same continent

- Far far away

C. Grigoras, November 2010



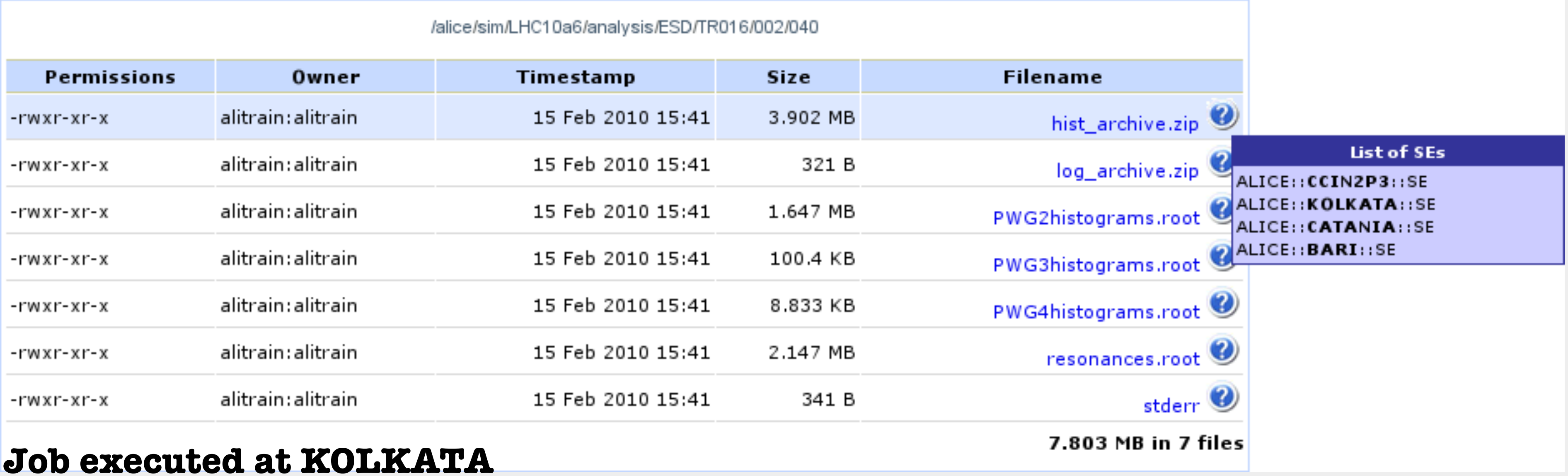
/alice/sim/LHC10a6/analysis/ESD/TR016/002/078

Permissions	Owner	Timestamp	Size	Filename
-rwxr-xr-x	alitrain:alitrain	15 Feb 2010 14:59	11.17 MB	hist_archive.zip
-rwxr-xr-x	alitrain:alitrain	15 Feb 2010 14:59	324 B	log_archive.zip
-rwxr-xr-x	alitrain:alitrain	15 Feb 2010 14:59	4.741 MB	PWG2histograms.root
-rwxr-xr-x	alitrain:alitrain	15 Feb 2010 14:59	497.4 KB	PWG3histograms.root
-rwxr-xr-x	alitrain:alitrain	15 Feb 2010 14:59	9.658 KB	PWG4histograms.root
-rwxr-xr-x	alitrain:alitrain	15 Feb 2010 14:59	5.929 MB	resonances.root
-rwxr-xr-x	alitrain:alitrain	15 Feb 2010 14:59	342 B	stderr

22.33 MB in 7 files

List of SEs
ALICE::**ITEP**::SE
ALICE::**PNPI**::SE
ALICE::**MEPHI**::SE
ALICE::**JINR**::SE

Job executed at JINR



/alice/sim/LHC10a6/analysis/ESD/TR016/002/040

Permissions	Owner	Timestamp	Size	Filename
-rwxr-xr-x	alitrain:alitrain	15 Feb 2010 15:41	3.902 MB	hist_archive.zip
-rwxr-xr-x	alitrain:alitrain	15 Feb 2010 15:41	321 B	log_archive.zip
-rwxr-xr-x	alitrain:alitrain	15 Feb 2010 15:41	1.647 MB	PWG2histograms.root
-rwxr-xr-x	alitrain:alitrain	15 Feb 2010 15:41	100.4 KB	PWG3histograms.root
-rwxr-xr-x	alitrain:alitrain	15 Feb 2010 15:41	8.833 KB	PWG4histograms.root
-rwxr-xr-x	alitrain:alitrain	15 Feb 2010 15:41	2.147 MB	resonances.root
-rwxr-xr-x	alitrain:alitrain	15 Feb 2010 15:41	341 B	stderr

7.803 MB in 7 files

List of SEs
ALICE::**CCIN2P3**::SE
ALICE::**KOLKATA**::SE
ALICE::**CATANIA**::SE
ALICE::**BARI**::SE

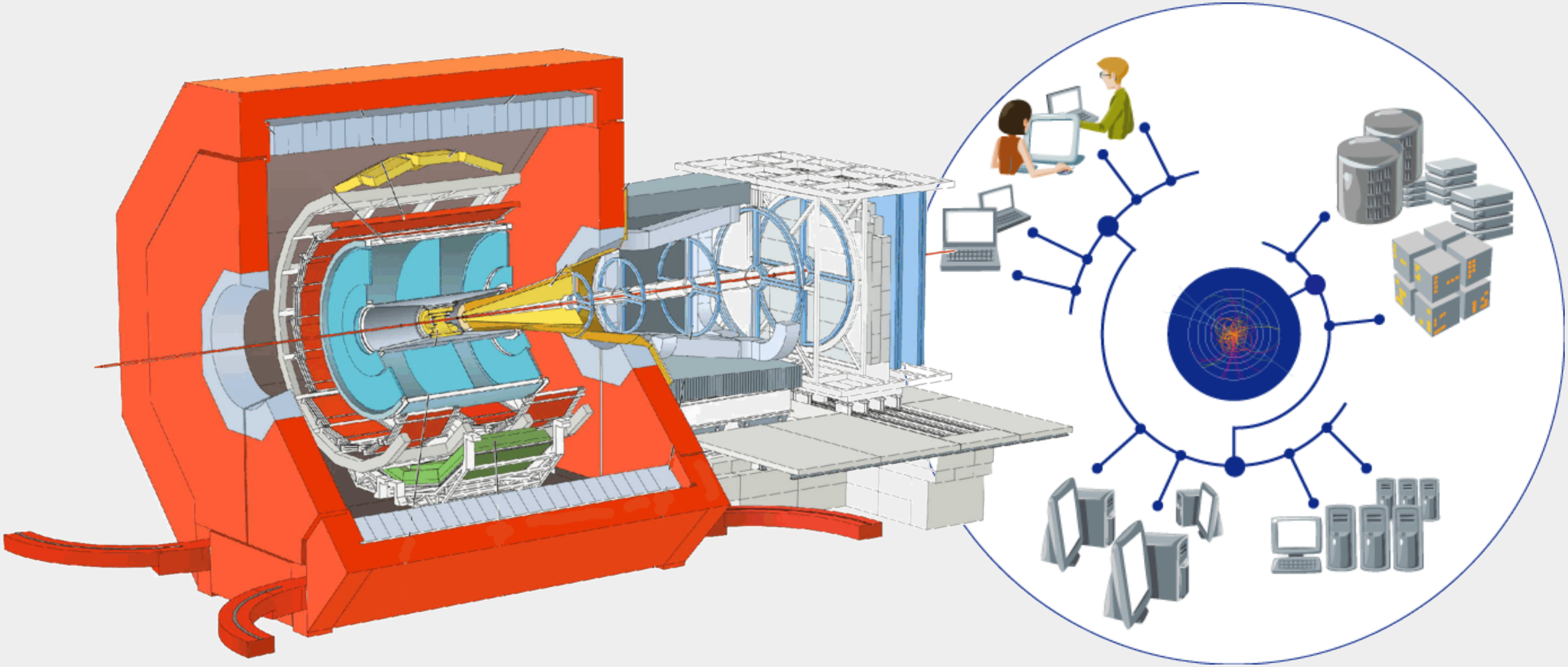
Job executed at KOLKATA

C. Grigoras, November 2010

BOTTOM LINE

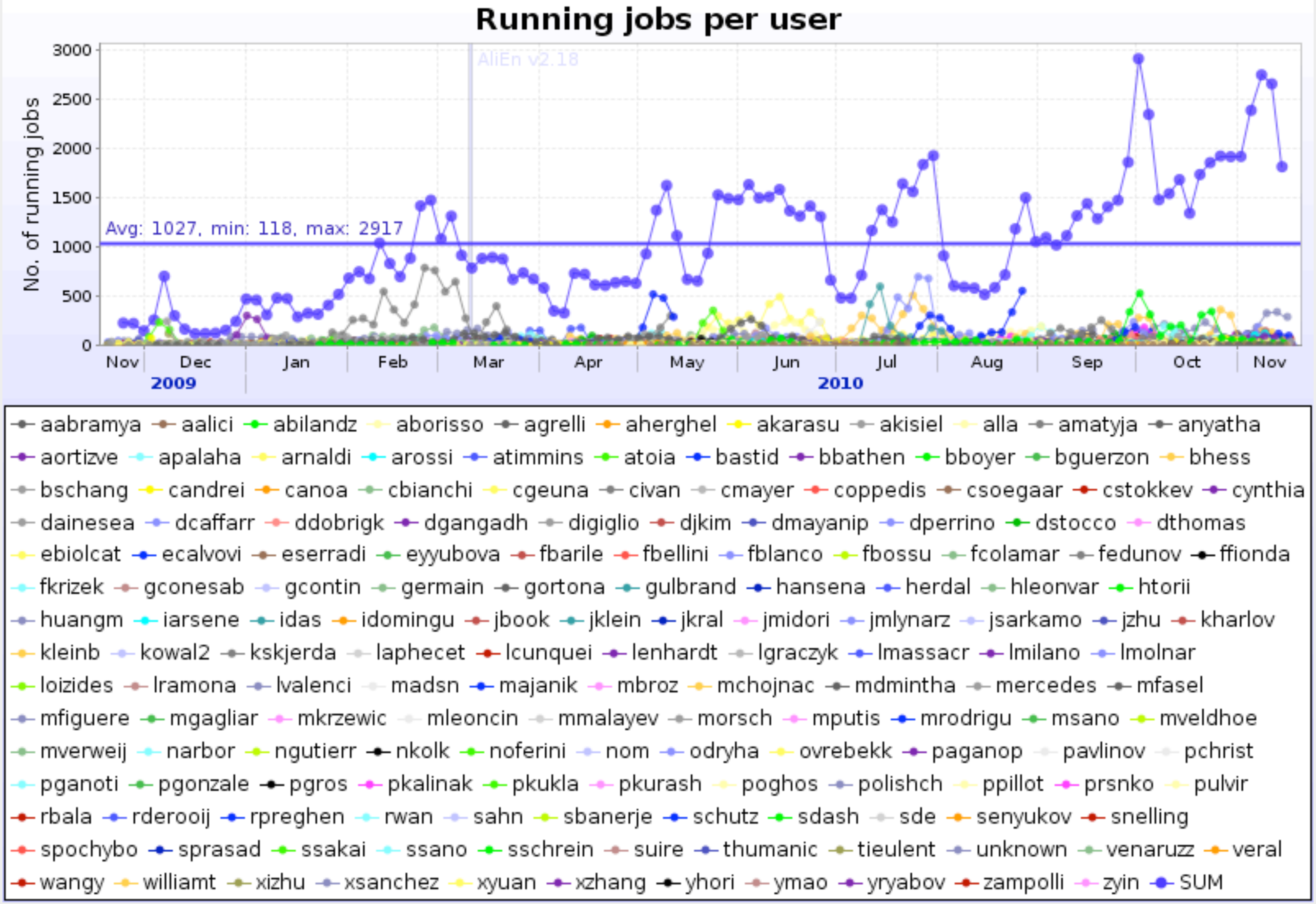
- Flexible storage configuration
  - QoS tags are all that users should know about the system
  - We can store *N* replicas at once
- Maintenance-free system
  - Monitoring feedback on known elements and automatic discovery and configuration of new ones
- Reliable and efficient file access
  - No more failed jobs due to auto discovery and failover in case of temporary problems
  - Use the closest working storage element(s) to where the application runs

C. Grigoras, November 2010

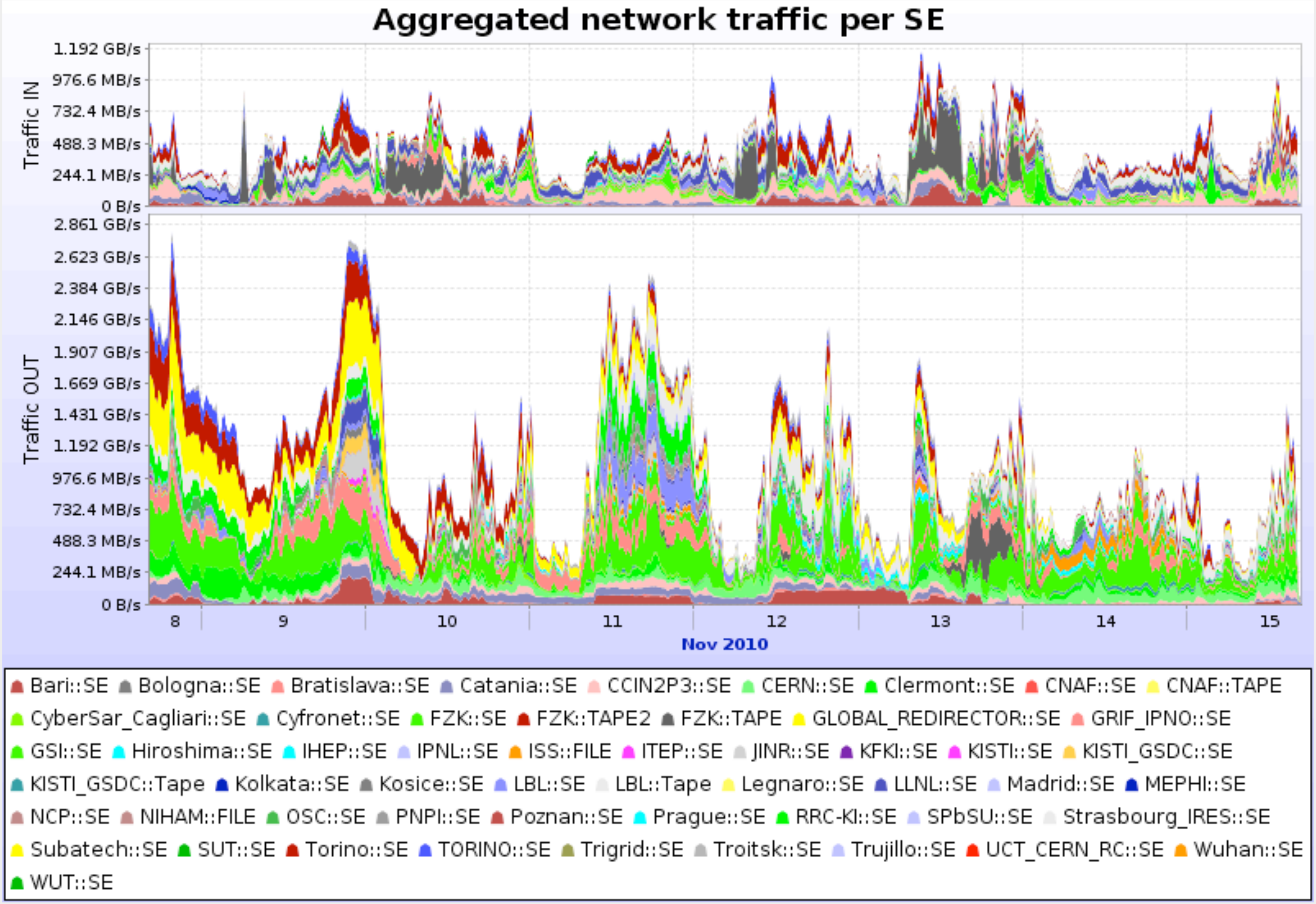


SOME PLOTS

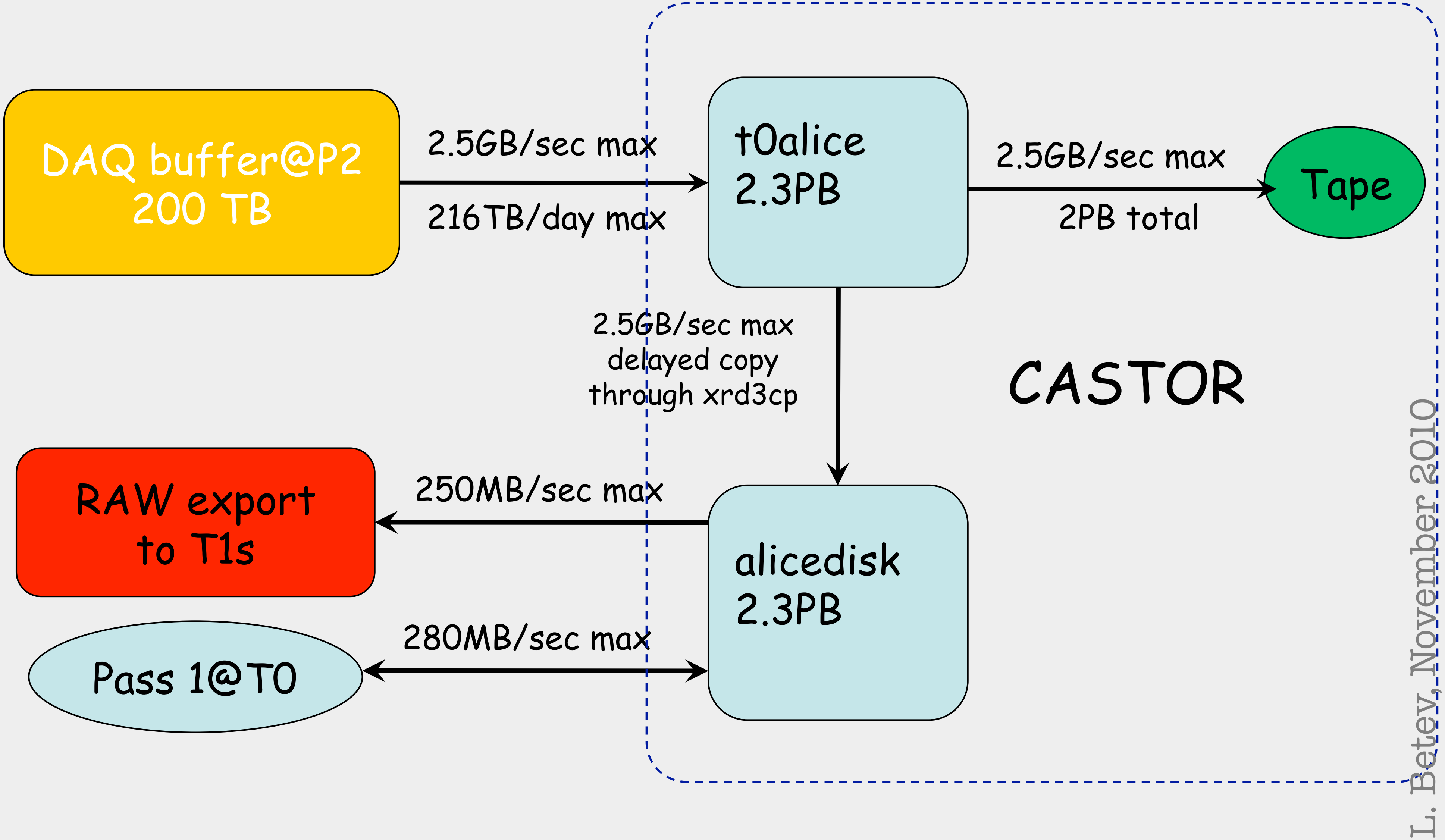
DISTRIBUTED ANALYSIS ON THE GRID



TRAFFIC TO AND FROM SE



ALICE HI DATA FLOWCHART

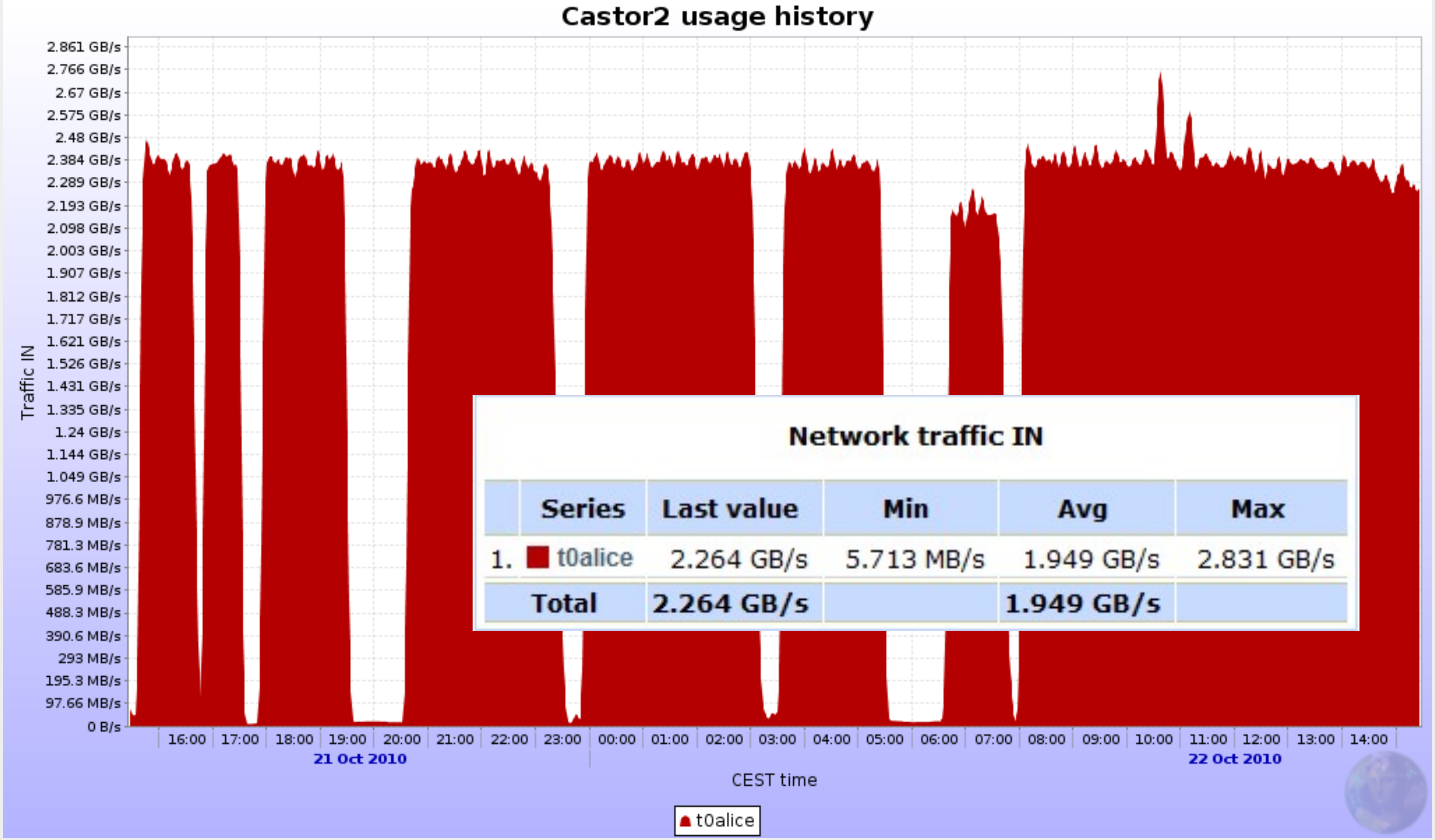


FULL RATE EXERCISE

- Sustained 24 hours data taking and transfer to tape at maximum HI rate
- Simultaneous copy of RAW to disk pool (third copy of data)
- Reconstruct ~10% of data
- CMS transferring data at their max speed at the same time
- Check of IT infrastructure (network and tapes) ability to cope with the combined rate

L. Betev, November 2010

TRANSFER P2 CASTOR

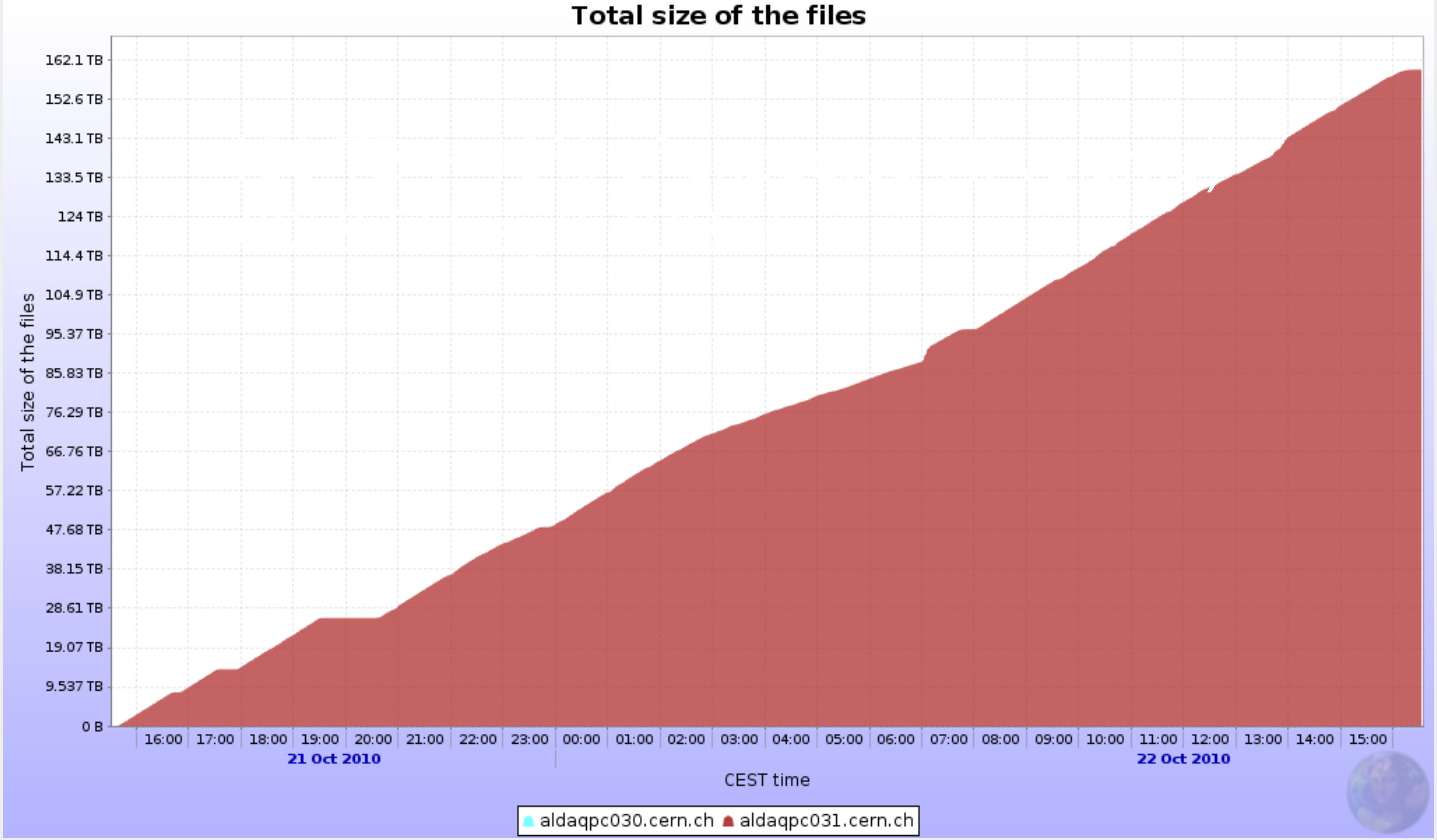


Network traffic IN

	Series	Last value	Min	Avg	Max
1.	🟥 t0alice	2.264 GB/s	5.713 MB/s	1.949 GB/s	2.831 GB/s
	Total	**2.264 GB/s**		**1.949 GB/s**	

L. Betev, November 2010

DATA VOLUME

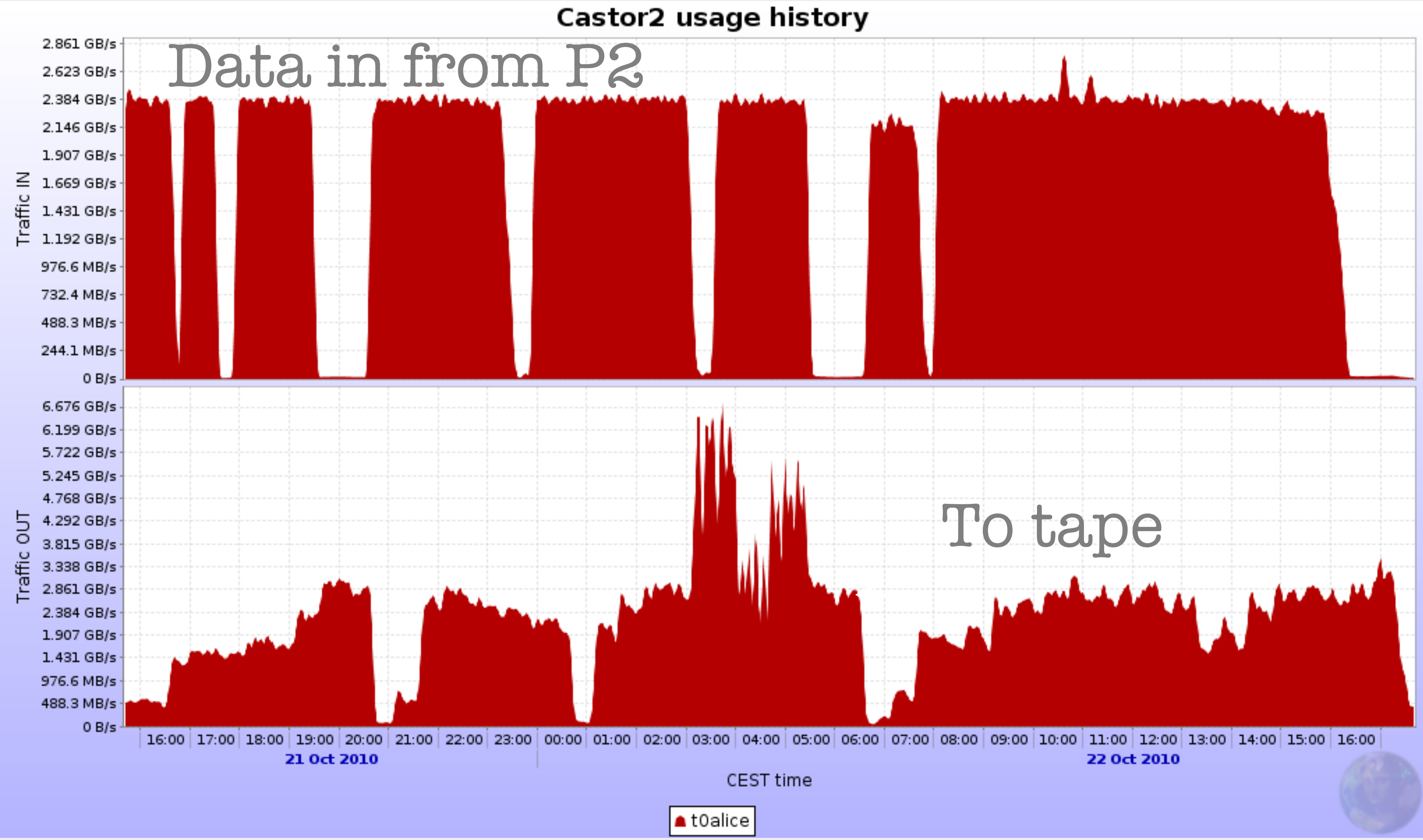

Total size of the files
Total size of the files
162.1 TB
152.6 TB
143.1 TB
133.5 TB
124 TB
114.4 TB
104.9 TB
95.37 TB
85.83 TB
76.29 TB
66.76 TB
57.22 TB
47.68 TB
38.15 TB
28.61 TB
19.07 TB
9.537 TB
0 B
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
00:00
01:00
02:00
03:00
04:00
05:00
06:00
07:00
08:00
09:00
10:00
11:00
12:00
13:00
14:00
15:00
21 Oct 2010
22 Oct 2010
CEST time
aldaqpc030.cern.ch
aldaqpc031.cern.ch

P2 TO CASTOR

- Average rate – 2GB/sec
- Max rate 2.5GB/sec
- ~10% of the expected HI data volume

- Realistic data taking scenario
  - 3 interruptions in data taking for detector reconfiguration
  - 2 interruptions to follow up on data transfer to tapes

DISK BUFFERS



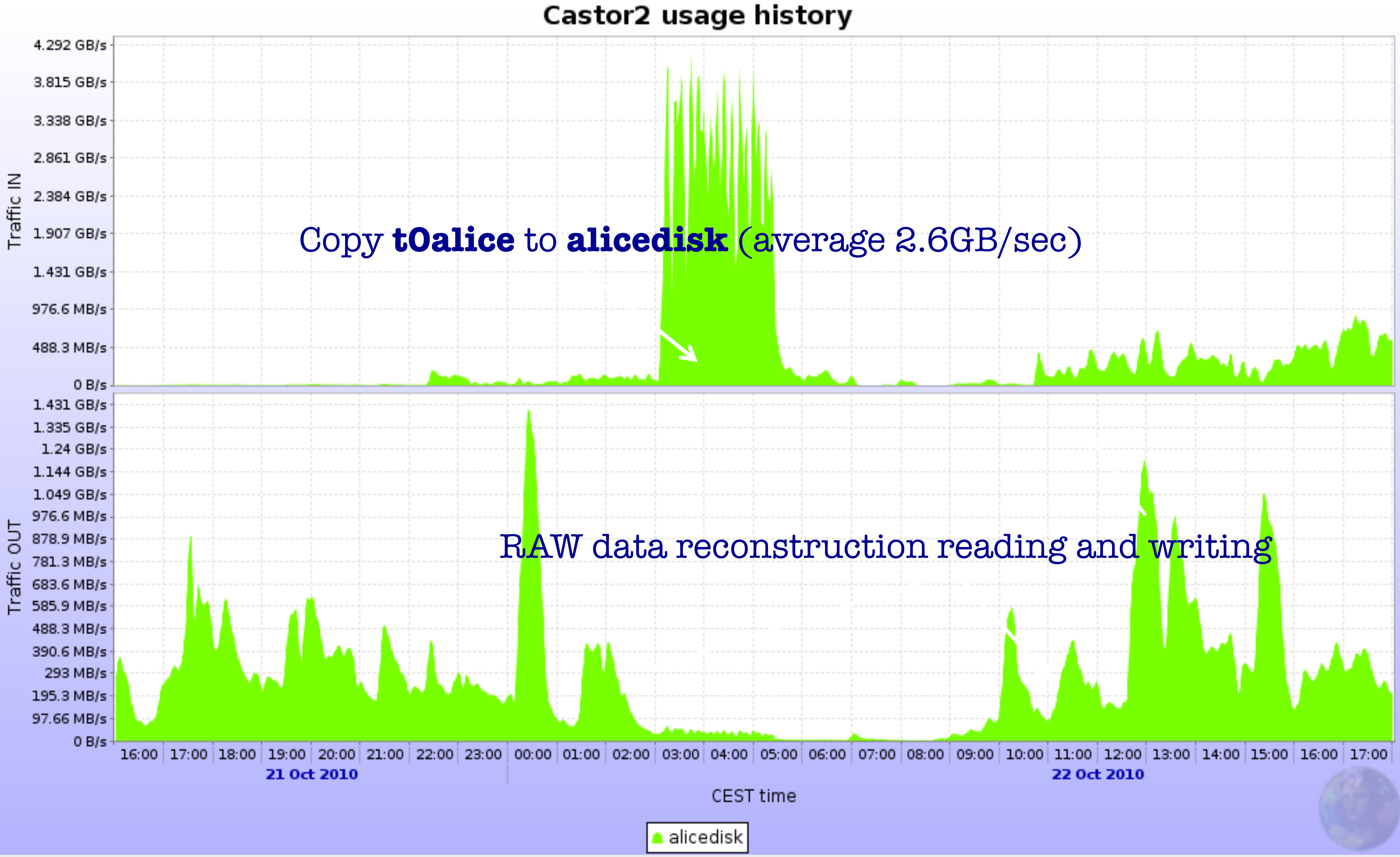
Data in from P2

To tape

CASTOR DISK BUFFER OUT

- Average rate 2.4GB/sec, max 8.7GB/sec
  -
- The data is on tape no later than 1 hour after being written on the disk buffer (**t0alice**)
- The **t0alice** buffer can sustain the combined rate from P2, to tape and the third party copy to **alicedisk** SE

TAPE BUFFERS TO ALICEDISK



Copy **t0alice** to **alicedisk** (average 2.6GB/sec)

RAW data reconstruction reading and writing

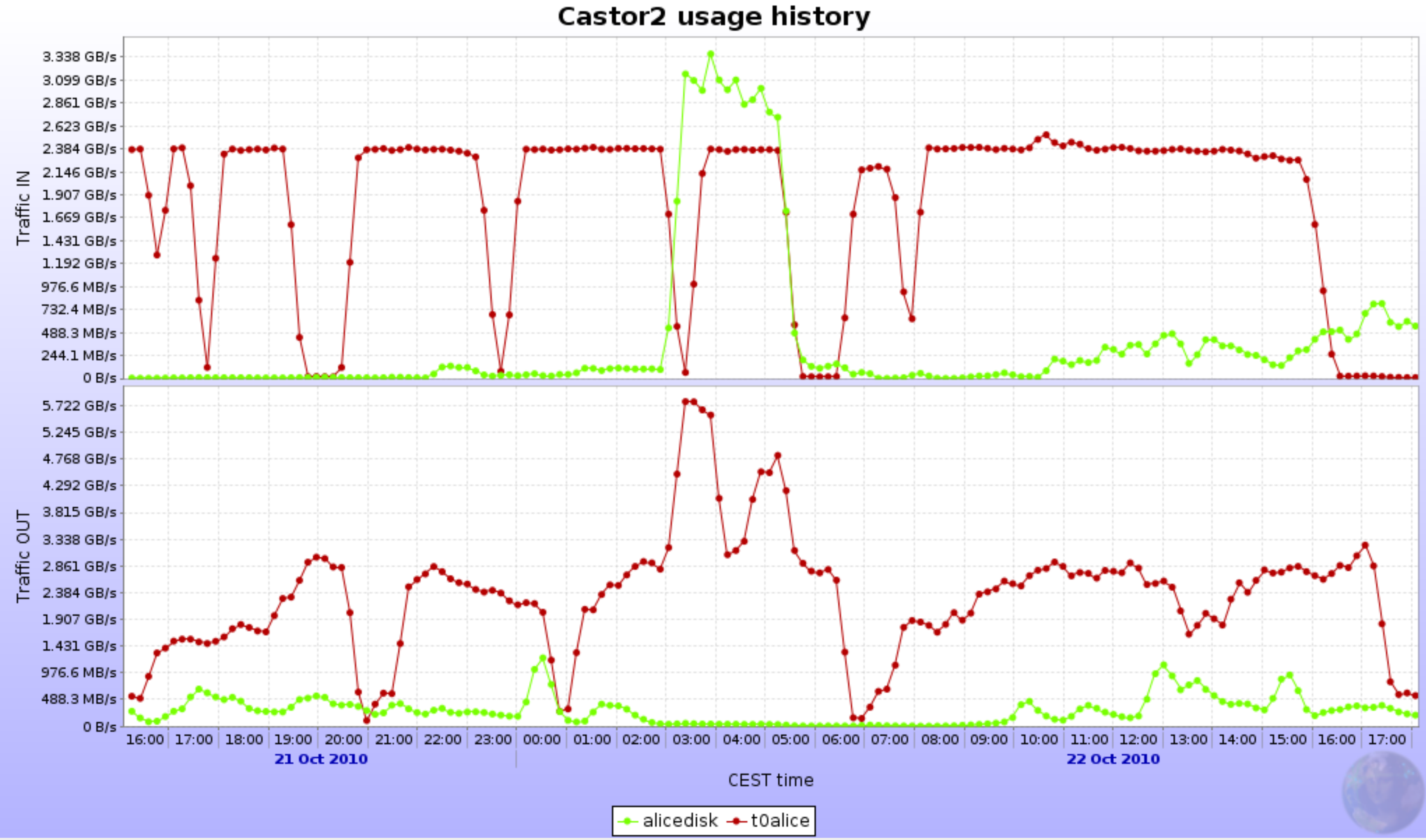
THIRD PARTY COPY TO ALICEDISK AND RECO

- Average copy rate – 2.5GB/sec

- Average reco ‘in’ rate – 200 MB/sec
  - ~10% from RAW

- Average reco ‘out’ rate – 20MB/sec
  - ~10% from reco ‘in’ rate

COMBINED RATE PLOT



THANKS