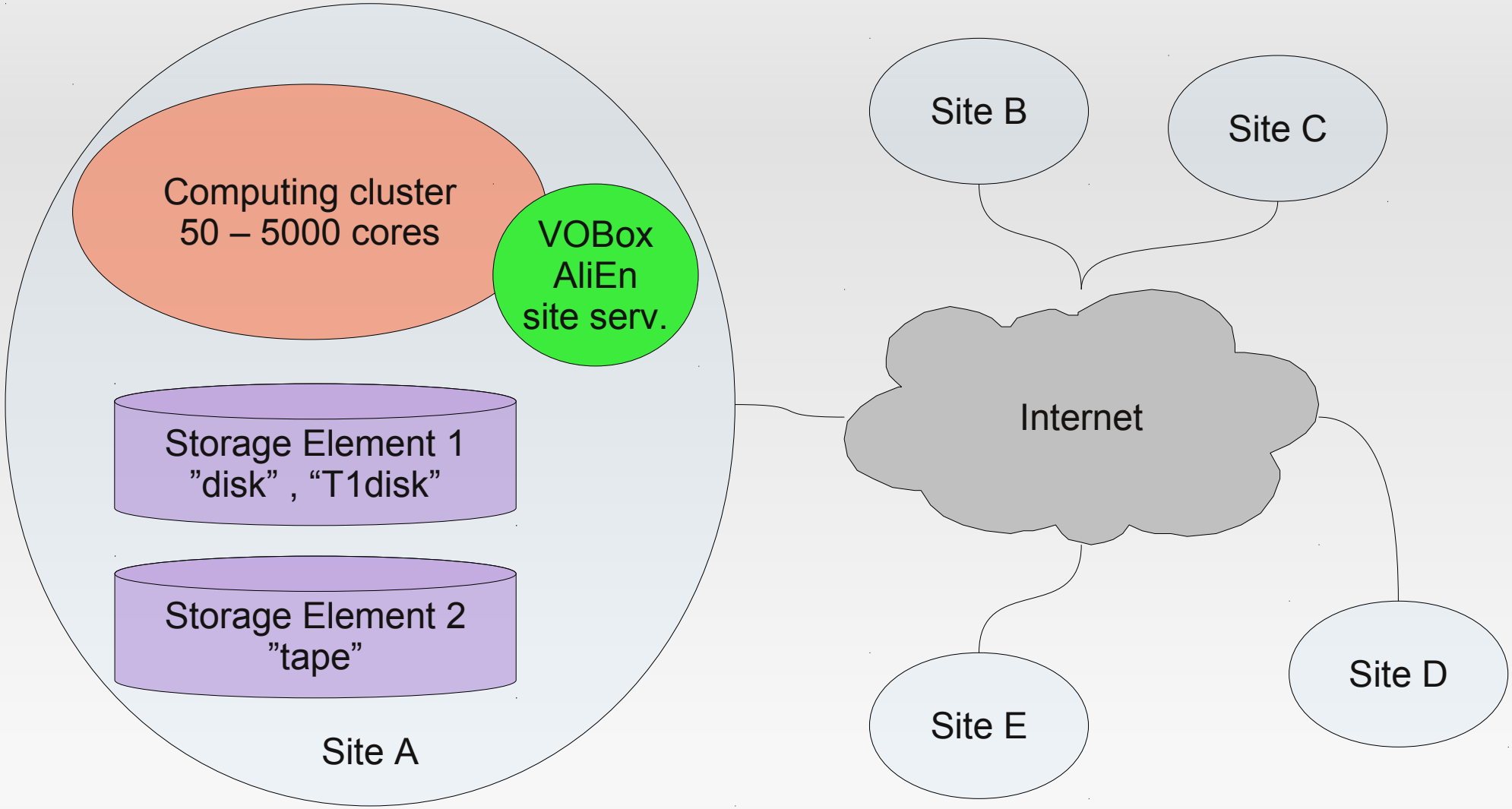
Storage discovery in AliEn

Costin.Grigoras@cern.ch

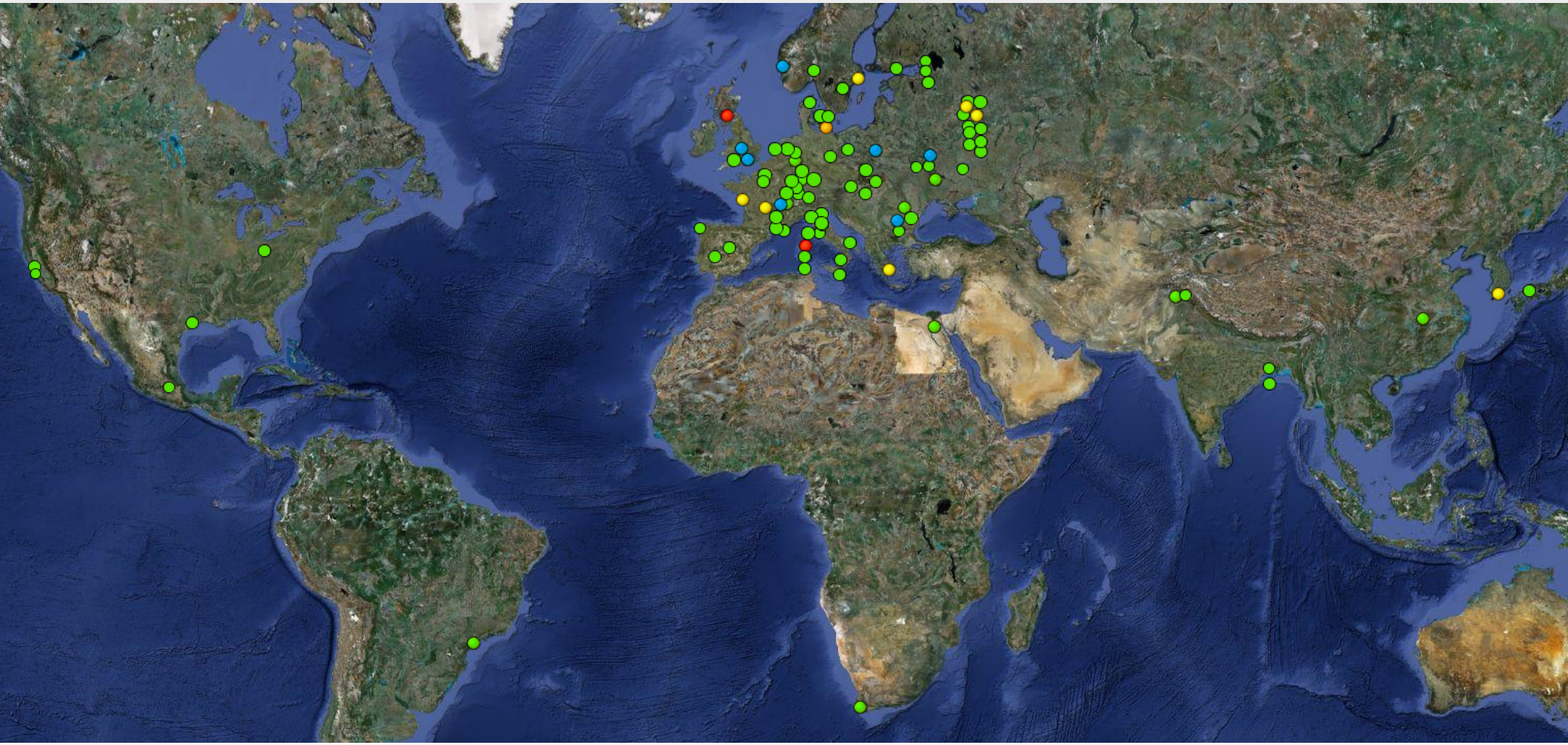
Background – AliEn

- ALIce ENvironment – a lightweight Grid environment, users' door to the Grid
- Central Services
  - File catalogue, AAA services, Job & Transfer queues, various Optimizers
- Site Services
  - CE, ClusterMonitor, MonALISA
- More information on AliEn in Poster 64

Site structure



Sites overview



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

http://alimonitor.cern.ch/

The problem

- We only write to one user-specified target storage
- How to efficiently write N replicas of a file ?
  - from jobs running inside the sites
  - by a user running its software on a laptop while at home
    - same case as for a worker running somewhere in the clouds
- Then, how to efficiently read the data when N replicas are available?
- In the end this is just a variation of the data locality problem

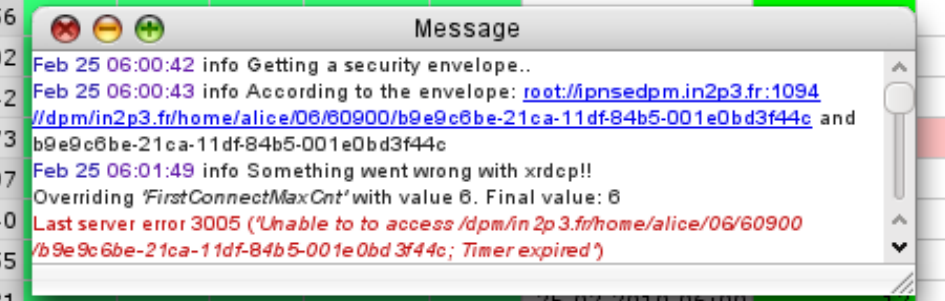
Step 1 – Storage status

- 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

Step 1 – Storage status

- And this is how the monitoring of the storage elements looks like

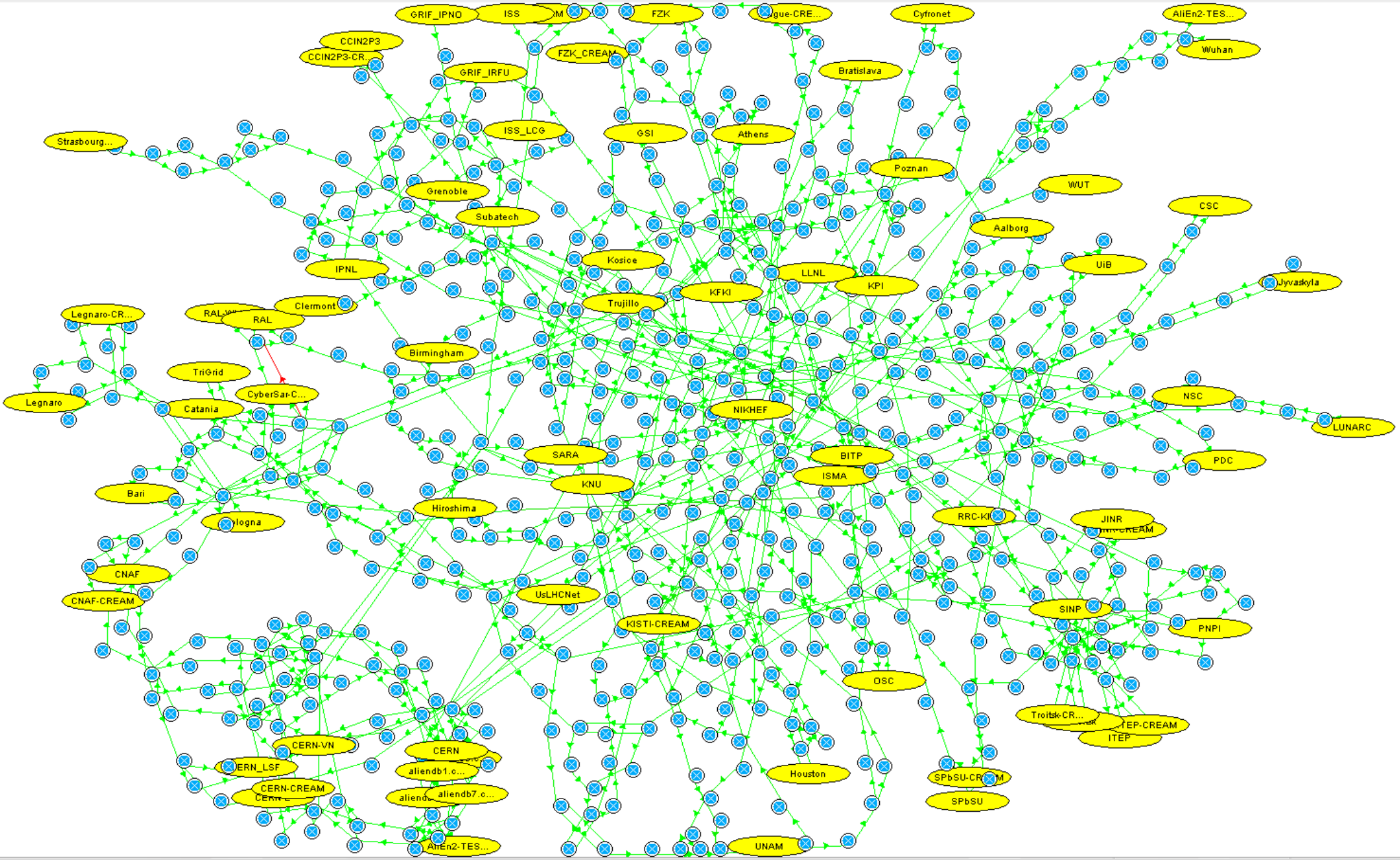
	Statistics					Functional tests						Last day tests	
SE Name	Size	Used	Free	Usage	No. of files	add	ls	get	whereis	rm	Last OK test	Successful	Failed
1. Bari - SE	33.69 TB	1.398 TB	32.29 TB	4.149%	75,820						25.02.2010 06:00	12	0
2. Bologna - SE	500 GB	94.45 GB	405.6 GB	18.89%	28,280	Feb ...	Last...	Last...	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								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



Step 2 – Discover network topology

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

Step 2 – Discover network topology

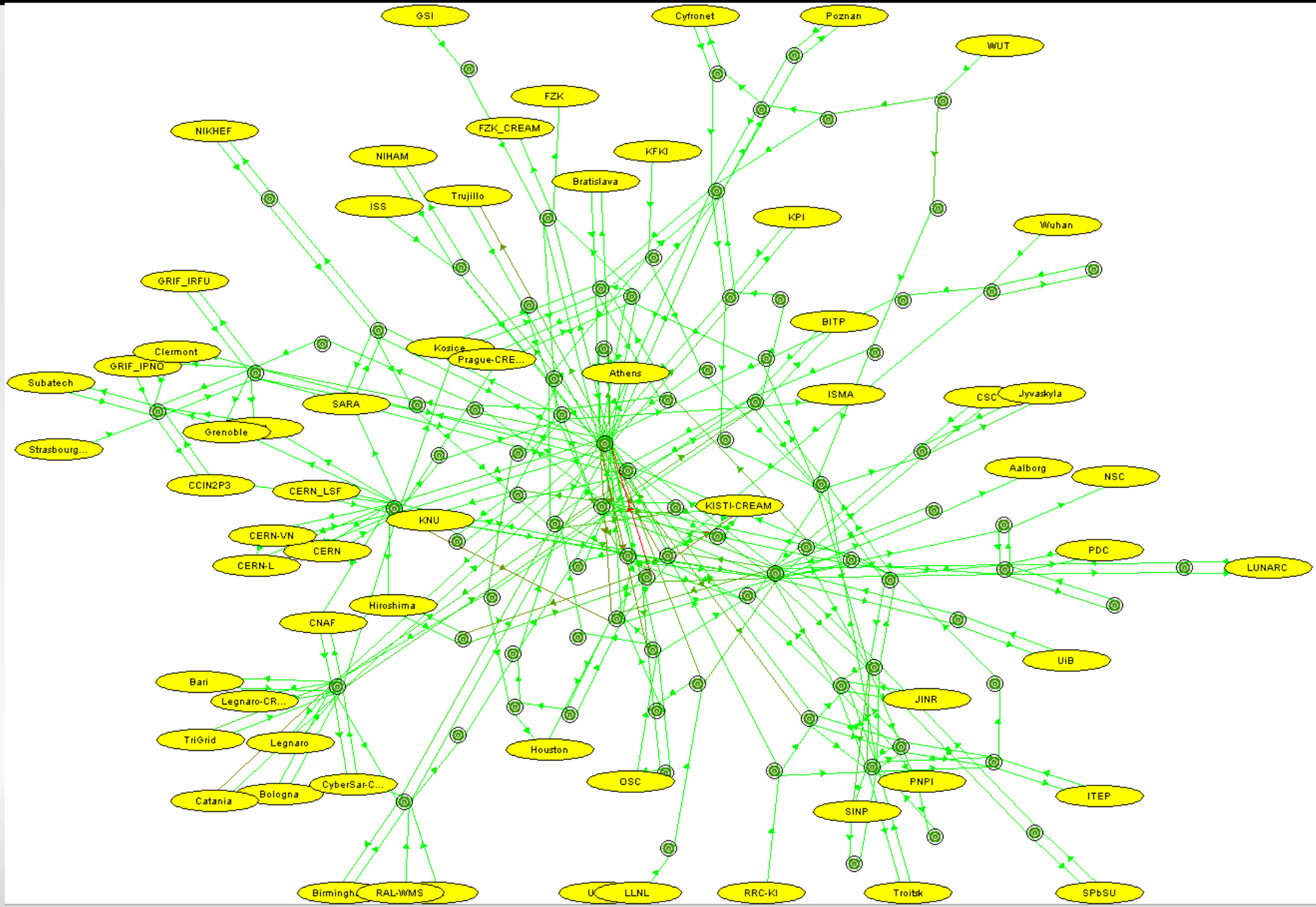


Step 3 – Derived network topology

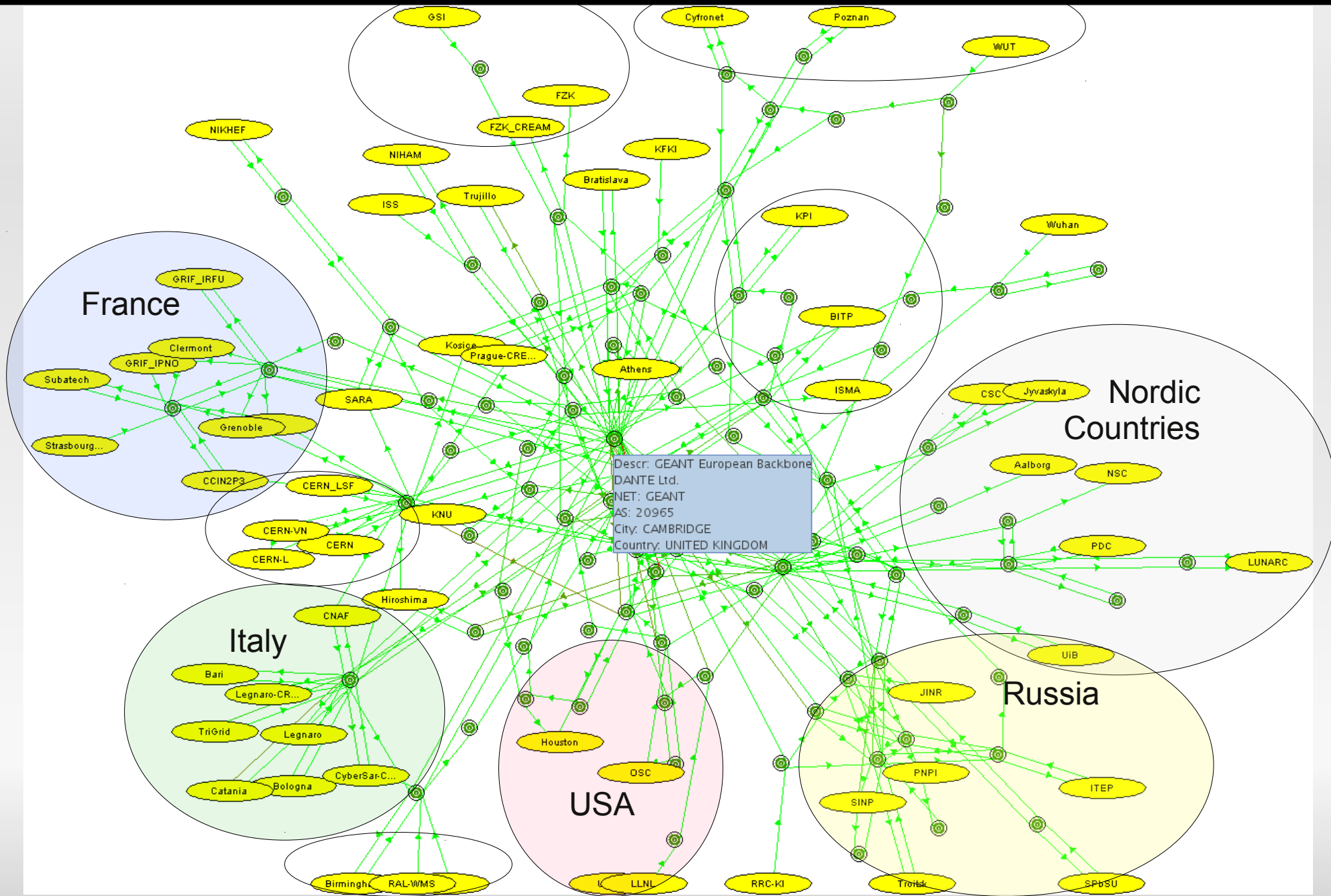
But if we ...

- group the routers in the respective Autonomous Systems (AS)
- compute the distance (RTT) between them
- then we have a better understanding of the relation between sites

Step 3 – Derived network topology



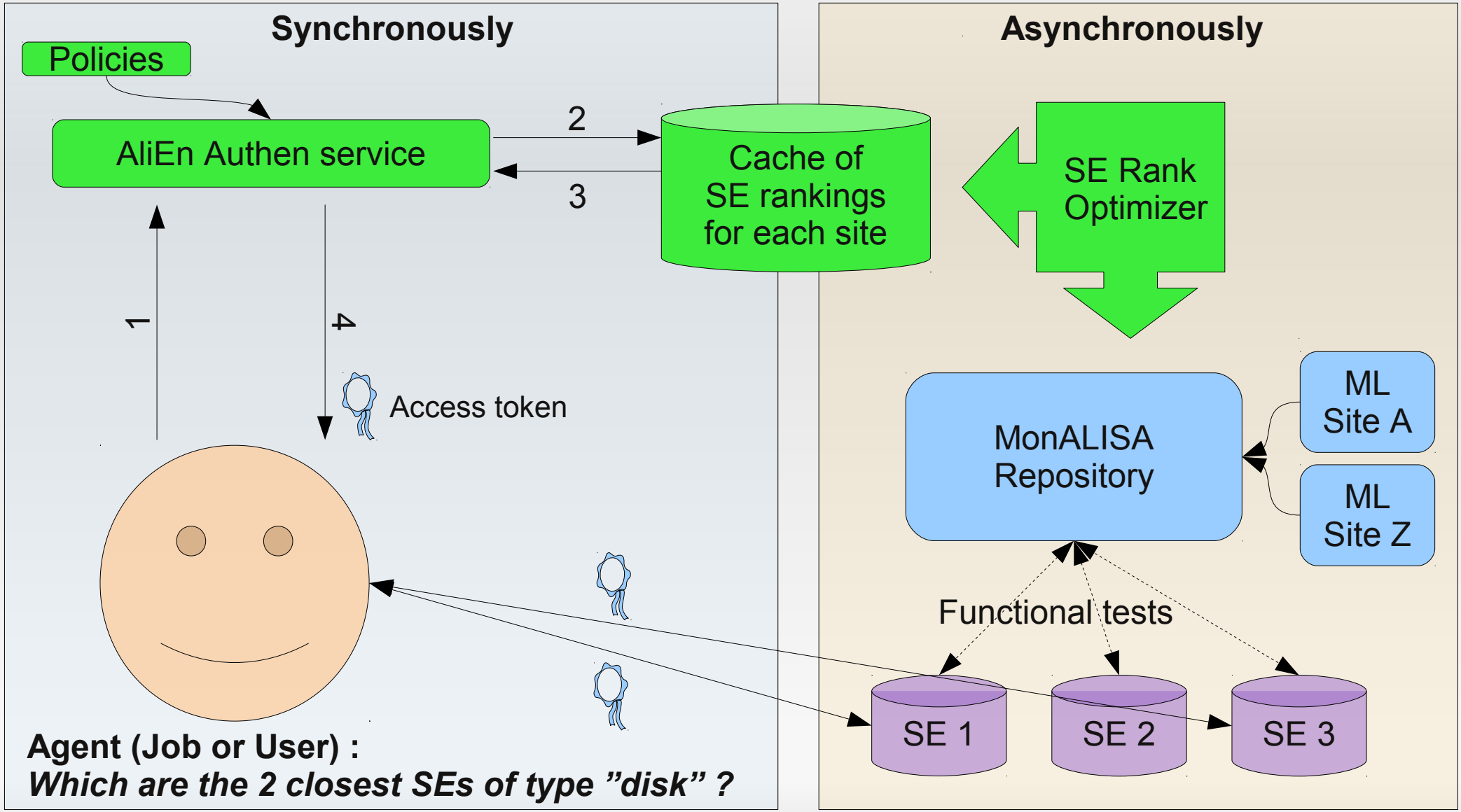
Step 3 – Derived network topology



Step 4 – Client to Storage distance

- distance(IP, IP)
  - 0 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
  - 1 Far far away
- distance(IP, Set<IP>): Client's public IP to all known IPs for the storage

Solution


Synchronously
Asynchronously
Policies
AliEn Authen service
2
3
Cache of
SE rankings
for each site
SE Rank
Optimizer
1
4
Access token
MonALISA
Repository
ML
Site A
ML
Site Z
Functional tests
SE 1
SE 2
SE 3
Agent (Job or User) :
Which are the 2 closest SEs of type ”disk” ?

Samples

/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

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 resources
- 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