ALICE DATA ACCESS MODEL

Costin.Grigoras@cern.ch

Outline

- ALICE data model
- Some figures
- Infrastructure monitoring
- Replica discovery mechanism

ALICE data model

- Central catalogue of logical file names (LFN)
  - With owner:group and unix-style permissions
  - Size, MD5 of files
  - Metadata on subtrees
- Each LFN is associated a GUID
- Any number of physical file names (PFN) can be associated to a LFN
  - Like root://<redirector>//<HH>/<hhhhh>/<GUID>
    - HH and hhhhh are hashes of the GUID

ALICE data model (2)

- Data files are accessed directly
  - Jobs go to where a copy of the data is
  - Reading from the closest working replica to the job
- Exclusive use of xrootd protocol
  - while 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, 3xAODs, etc...)
- Scheduled data transfers for raw data only with xrd3cp
  - T0 -> one T1 / run, selected at data taking time

Some figures

- 60 disk storage elements + 8 tape-backed (T0 and T1s)
  - 28PB in 307M files (replicas included)
- 2012 averages:
  - 31PB written (1.2GB/s)
    - 2.4PB is the raw data (1.4PB to T0, 1PB replicated to the T1 set – only the good runs)
      - So some 67MB/s are needed for the raw data replication
    - In average ~half is written locally, ~half on the uplink
  - 216PB read back (8.6GB/s) - 7x the amount written
    - Mostly from the local storage
- Sustained periods of 3-4x the above

Analysis train efficiency

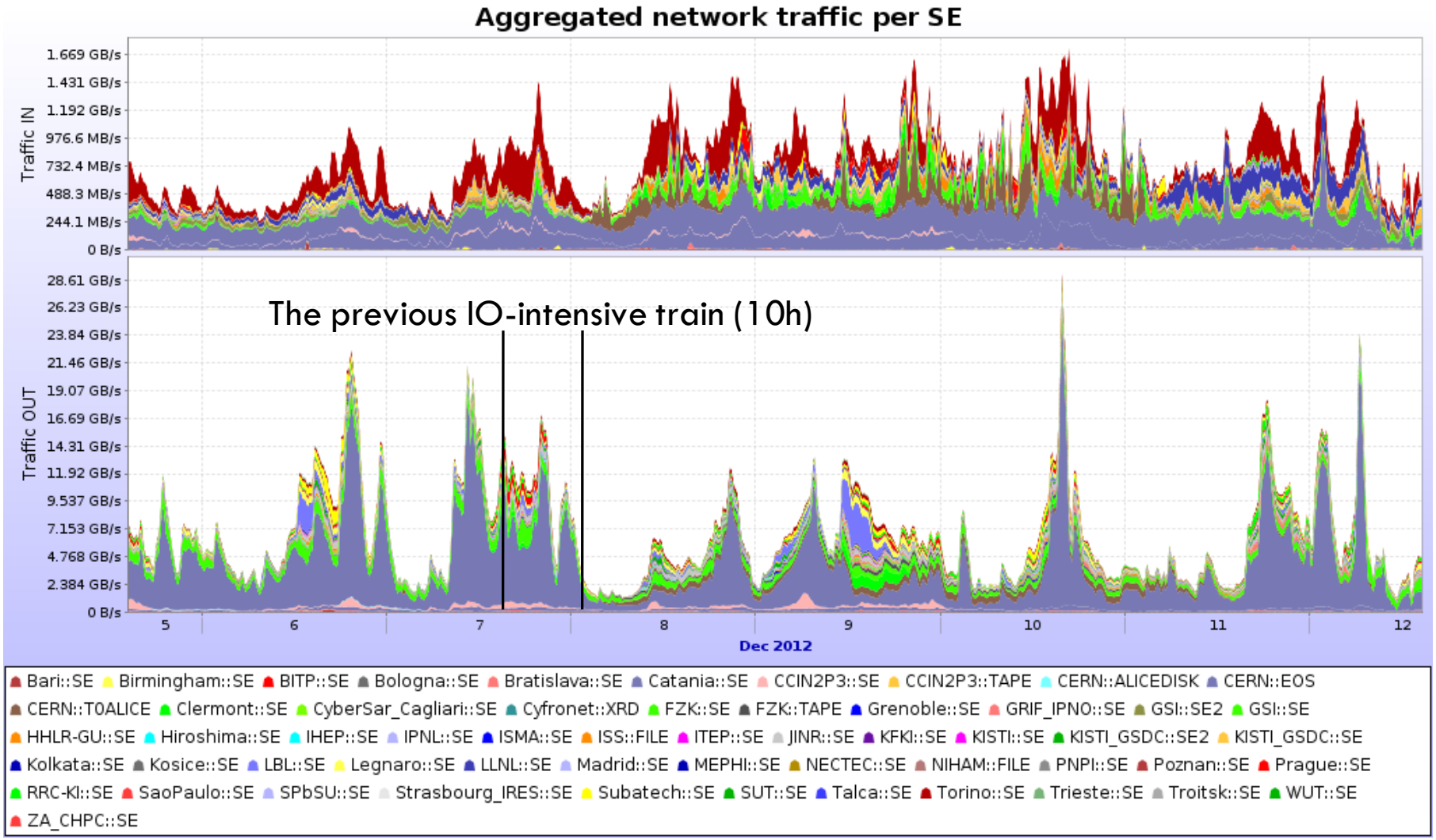
- The most IO-demanding processing
- Last 1M analysis jobs (mix of all types of analysis with aggressive scheduling)
  - 14.2M input files
  - 87.5% accessed from the site local SE at 3.1MB/s
  - 12.5% read from remote at 0.97MB/s
    - Bringing the average processing speed down to 2.8MB/s
- Average job efficiency was 70% for an avg CPU power of 10.14 HepSpec06
- So we would need at least **0.4MB/s/HepSpec06** for the *average* analysis to be efficient
  - And the infrastructure to scale linearly with this …

A particular analysis task …

Site activity												
Site	**Job eff.**	**HepSpec06**	**All files**	**Local files**	**Remote files**	**CERN ALICEDISK**	**CNAF SE**	**NIHAM FILE**	**RRC-KI SE**	**JINR SE**	**PRAGUE SE**	**LBL SE**
CERN 4646 jobs (30.63%)	85.77%	10.25	77602 files 27.6 MB/s	77452 (99.81%) 27.68 MB/s	150 (0.193%) 10.96 MB/s	**77452 (99.81%)** **27.68 MB/s**	107 (0.138%) 10.4 MB/s	1 (0.001%) 3.186 MB/s				
CNAF 3744 jobs (24.68%)	32.27%	10.81	65943 files 12.14 MB/s	65865 (99.88%) 12.14 MB/s	78 (0.118%) 13.83 MB/s		**65865 (99.88%)** **12.14 MB/s**	4 (0.006%) 16.54 MB/s		8296 (12.58%) 17.77 MB/s		
NIHAM 3013 jobs (19.86%)	66.08%	9.176	52974 files 24.21 MB/s	51857 (97.89%) 27.74 MB/s	1117 (2.109%) 3.738 MB/s	1 (0.002%) 14.86 MB/s	164 (0.31%) 5.236 MB/s	**51857 (97.89%)** **27.74 MB/s**	34 (0.064%) 2.246 MB/s		20 (0.038%) 2.944 MB/s	1 (0.002%) 8.849 MB/s
RRC-KI 759 jobs (5.004%)	29.74%	12.06	11244 files 11.24 MB/s	10676 (94.95%) 15.34 MB/s	568 (5.052%) 1.62 MB/s		55 (0.489%) 1.283 MB/s	110 (0.978%) 1.613 MB/s	**10676 (94.95%)** **15.34 MB/s**	1 (0.009%) 20.88 MB/s		
JINR 591 jobs (3.896%)	61.42%	10.86	8337 files 21.68 MB/s	8270 (99.2%) 22.48 MB/s	67 (0.804%) 2.603 MB/s		2 (0.024%) 2.712 MB/s			**8270 (99.2%)** **22.48 MB/s**		
PRAGUE 403 jobs (2.657%)	44.04%	9.463	7174 files 15.69 MB/s	7124 (99.3%) 17.61 MB/s	50 (0.697%) 1.266 MB/s		1 (0.014%) 1.931 MB/s	16 (0.223%) 1.695 MB/s			**7124 (99.3%)** **17.61 MB/s**	
LBL 378 jobs (2.492%)	14.43%	9.279	5315 files 7.022 MB/s	5139 (96.69%) 7.761 MB/s	176 (3.311%) 1.756 MB/s		55 (1.035%) 3.898 MB/s	4 (0.075%) 3.505 MB/s				**5139 (96.69%)** **7.761 MB/s**
TOTAL **15168 jobs**	34.77%	10.14	**249118 files** **12.39 MB/s** **80.81 TB**	**239865 (96.29%)** **18.96 MB/s** **77.8 TB**		**77452 (32.29%)** **27.68 MB/s** **33.81 TB**	**65865 (27.46%)** **12.14 MB/s** **16.91 TB**	**51857 (21.62%)** **27.74 MB/s** **14.38 TB**	**10676 (4.451%)** **15.34 MB/s** **3.182 TB**	**8270 (3.448%)** **22.48 MB/s** **2.185 TB**	**7124 (2.97%)** **17.61 MB/s** **1.416 TB**	**5139 (2.142%)** **7.761 MB/s** **1.383 TB**
					9253 (3.714%) **1.239 MB/s** **3 TB**	**21 (0.227%)** **0.763 MB/s** **10.36 GB**	**4044 (43.7%)** **1.172 MB/s** **1.322 TB**	**186 (2.01%)** **1.755 MB/s** **43.99 GB**	**42 (0.454%)** **0.809 MB/s** **11.83 GB**	**26 (0.281%)** **0.409 MB/s** **11.05 GB**	**66 (0.713%)** **2.232 MB/s** **16.94 GB**	**2 (0.022%)** **1.17 MB/s** **538.2 MB**

- IO-intensive analysis train run
- Small fraction of files accessed remotely
  - With the expected penalty
- However the external connection is the lesser issue …

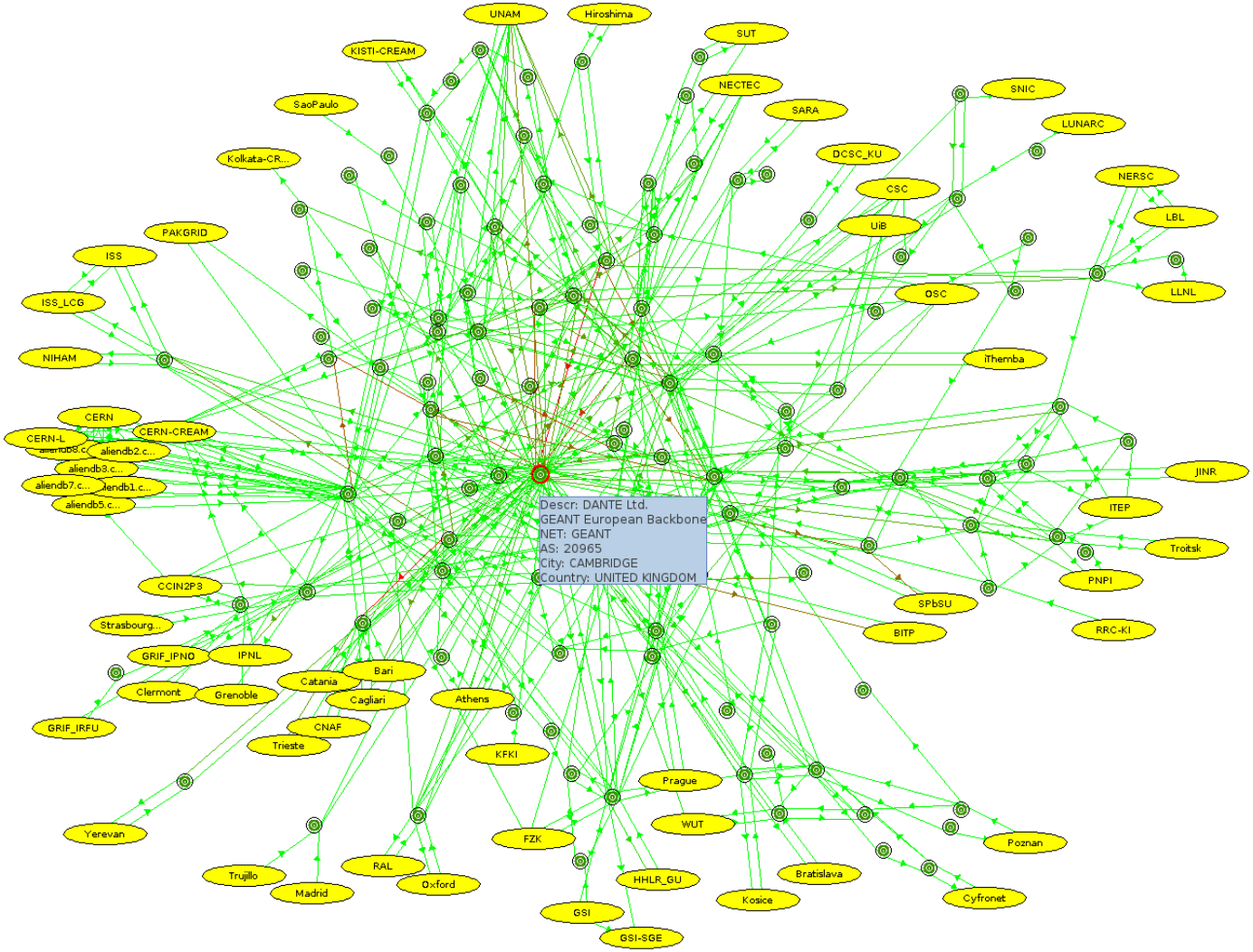
Aggregated SE traffic



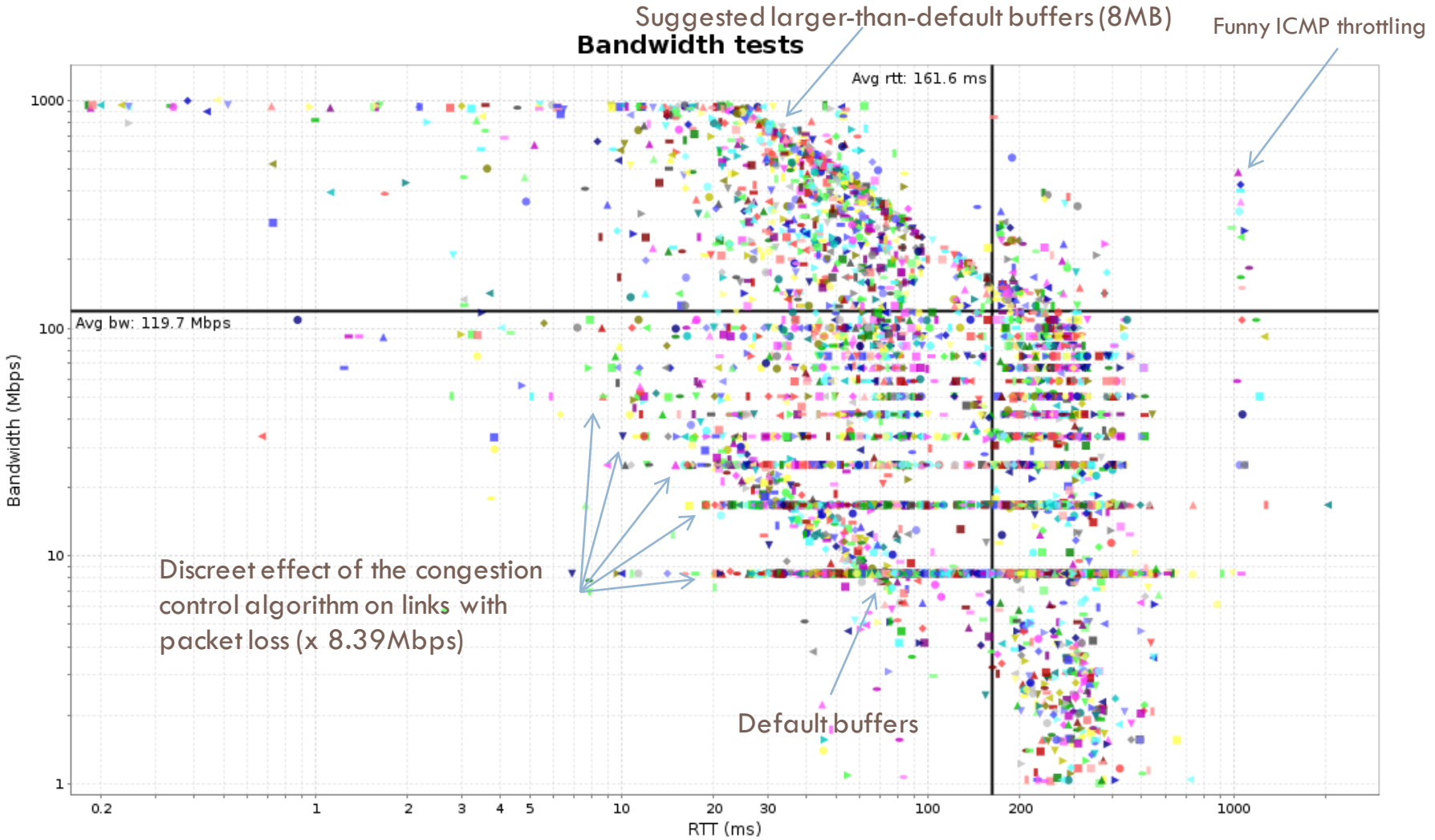
Infrastructure monitoring

- On each site VoBox a MonALISA service collects
  - Job resource consumption, WN host monitoring …
  - Local SEs host monitoring data (network traffic, load, sockets etc)
- And performs VoBox to VoBox measurements
  - traceroute / tracepath
  - 1 stream available bandwidth measurements (FDT)
    - This is what impacts the job efficiency
- All results are archived and we also infer the network topology from them

Network topology view in MonALISA



Available bandwidth per stream



Bandwidth test matrix

- 4 years of archived results for 80x80 sites matrix
- http://alimonitor.cern.ch/speed/

‹NIHAM›

Chart view »

IN from

No.	ID	Site	Speed (Mbps)	Hops	RTT (ms)	Streams
7.	1384109	Athens	679.51	12	33.99	1
50.	1385165	BITP	16.78	14	63.74	1
37.	1388596	Bari	41.90	16	43.91	1
30.	1270037	Birmingham	100.47			1
67.	1383738	Bologna		15	36.16	1
27.	1384012	Bratislava	100.57	13	24.90	1
4.	1384018	CCIN2P3	870.72	17	51.23	1
28.	1253827	CCIN2P31	100.57	17	53.91	1
45.	1383974	CERN-CREAM	25.14	19	46.63	1
46.	1388049	CERN-L	25.14	19	38.42	1
55.	1384206	CNAF	16.61	16	36.49	1

OUT to

No.	ID	Site ▾	Speed (Mbps)	Hops	RTT (ms)	Streams
6.	1384220	Athens	343.95	11	34.74	1
27.	1384064	BITP	41.95			1
9.	1388692	Bari	310.40	15	53.28	1
36.	1271409	Birmingham	25.17	21	77.33	1
47.	1384194	Bologna				1
42.	1388520	Bratislava	8.39	21	987.53	1
48.	1388590	CCIN2P3		16	52.60	1
49.	1253751	CCIN2P31		17	54.07	1
25.	1383846	CERN-CREAM	50.33			1
50.	1384440	CERN-L				1
7.	1384130	CNAF	318.78	15	44.25	1

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
  - Leaving as last resort only the SEs that fail the respective functional test
  - Weighted with their recent reliability
- Writing is slightly randomized for more 'democratic' data distribution

Plans

- Encourage sites to improve their infrastructure
  - This has the largest impact on the efficiency
  - Eg. few Xrootd gateways for large GPFS clusters, insufficient backbone capacity …
- Then provide as much information as possible for them to tackle the uplink problems
  - Deploy a similar test suite on the data servers
  - (Re)enable icmp where it is missing
  - (Re)apply TCP buffer settings
  - Accelerate the migration to SLC6
- But we only see the end-to-end results
  - How can we find out what happens in between?

Conclusions

- ALICE uses all resources uniformly
  - No dedicated SEs / sites / links for particular tasks
- This was an ambitious model that worked only because the network capacity exceeded the initial expectations
- Now the problems are more at an operational level
  - And we need more meaningful / realistic monitoring data to identify them
    - From site fabric to the backbones
    - This is not a one-shot process