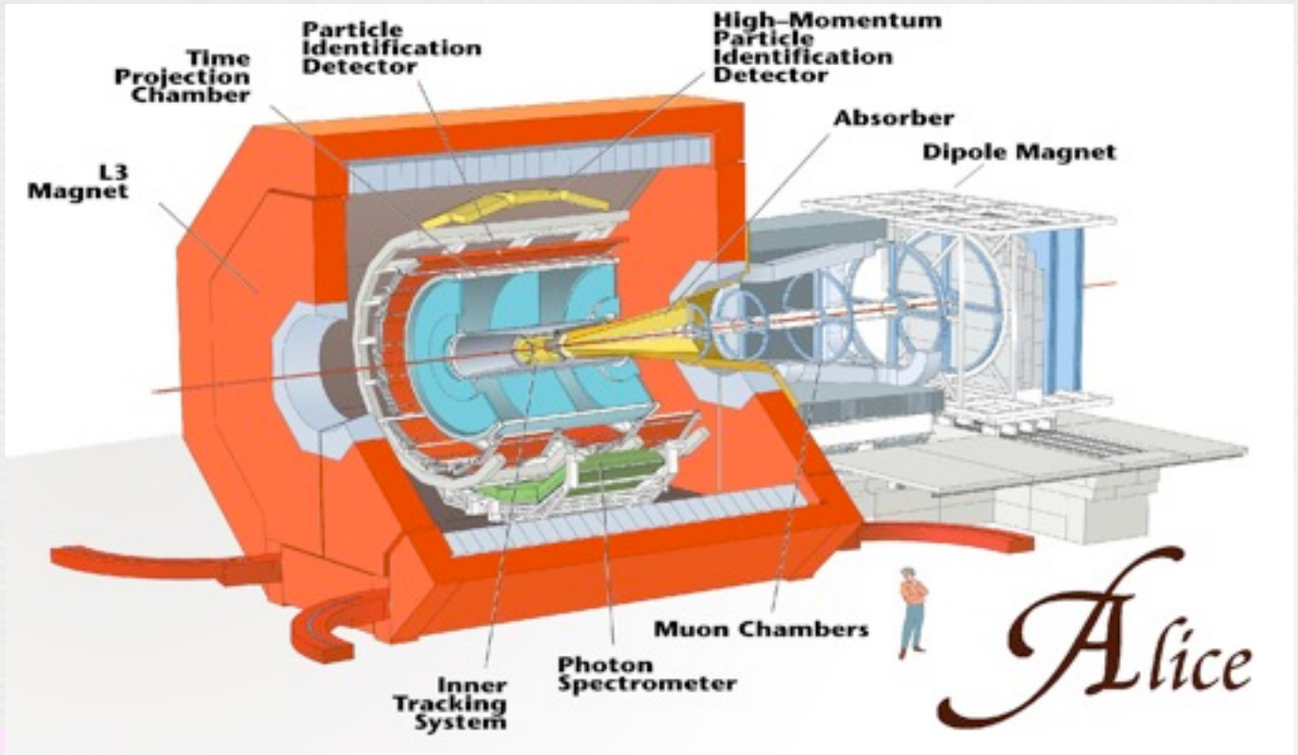
ALICE COMPUTING
AN OVERVIEW

Federico Carminati
T1-T2 workshop
Karlsruhe, January 24, 2012

BEFORE WE START

- Despite many technical difficulties, the LHC computing is a success
  - All experiments have been able to show very quickly results
  - The improvement rate in the quality of the analysis presented is impressive
- This is the first time in the HEP history that interesting results of such quality have been shown so rapidly
  - This is a proof of the maturity of the simulation, reconstruction, calibration and analysis
- We must have been doing something right
  - And we clearly have used well the “extra time” we had

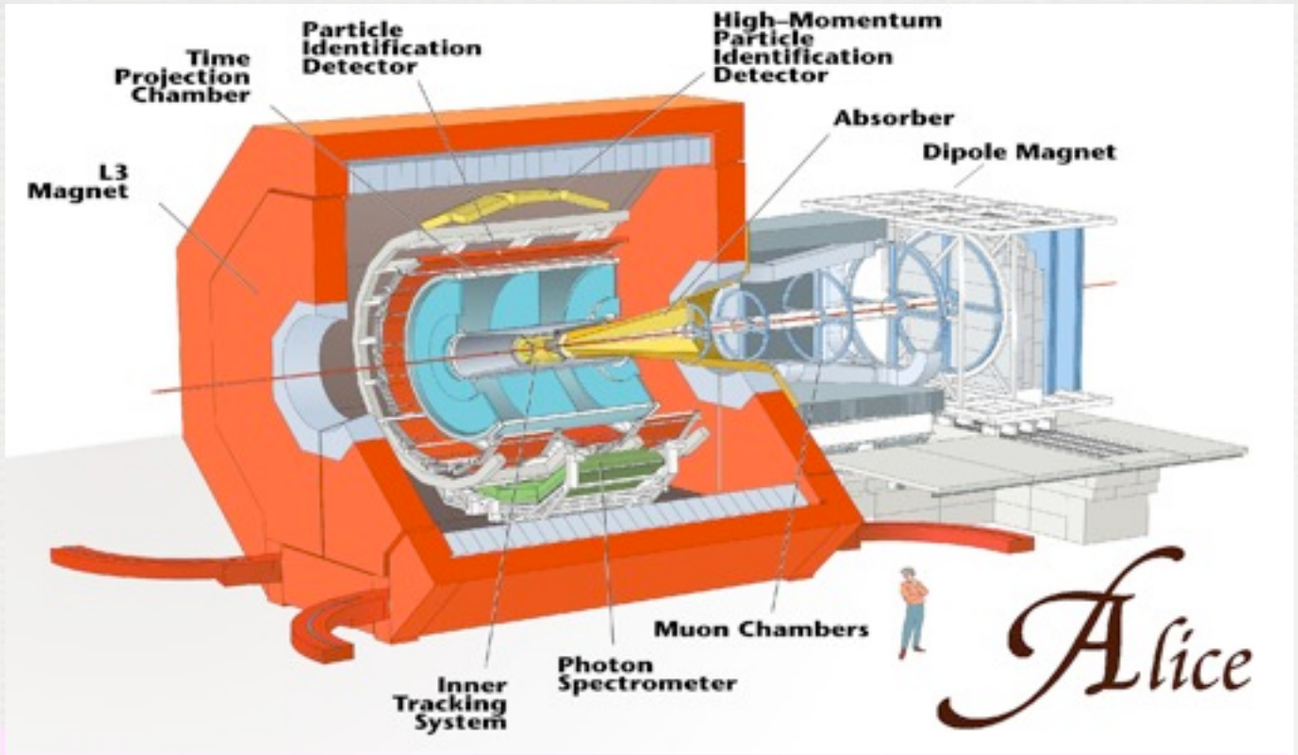


ALICE Collaboration
~ 1/2 ATLAS, CMS, ~ 2x LHCb
~1000 people, 30 countries,
~ 80 Institutes

Total weight	10,000t
Overall diameter	16.00m
Overall length	25m
Magnetic Field	0.5Tesla

8 kHz (160 GB / sec)
level 0 - special hardware
200 Hz (4 GB / sec)
level 1 - embedded processors
30 Hz (2.5 GB / sec)
level 2 - PCs
30 Hz
(1.25 GB / sec)
data recording &
offline analysis

- A full pp programme
- Data rate for pp is 100Hz@1MB



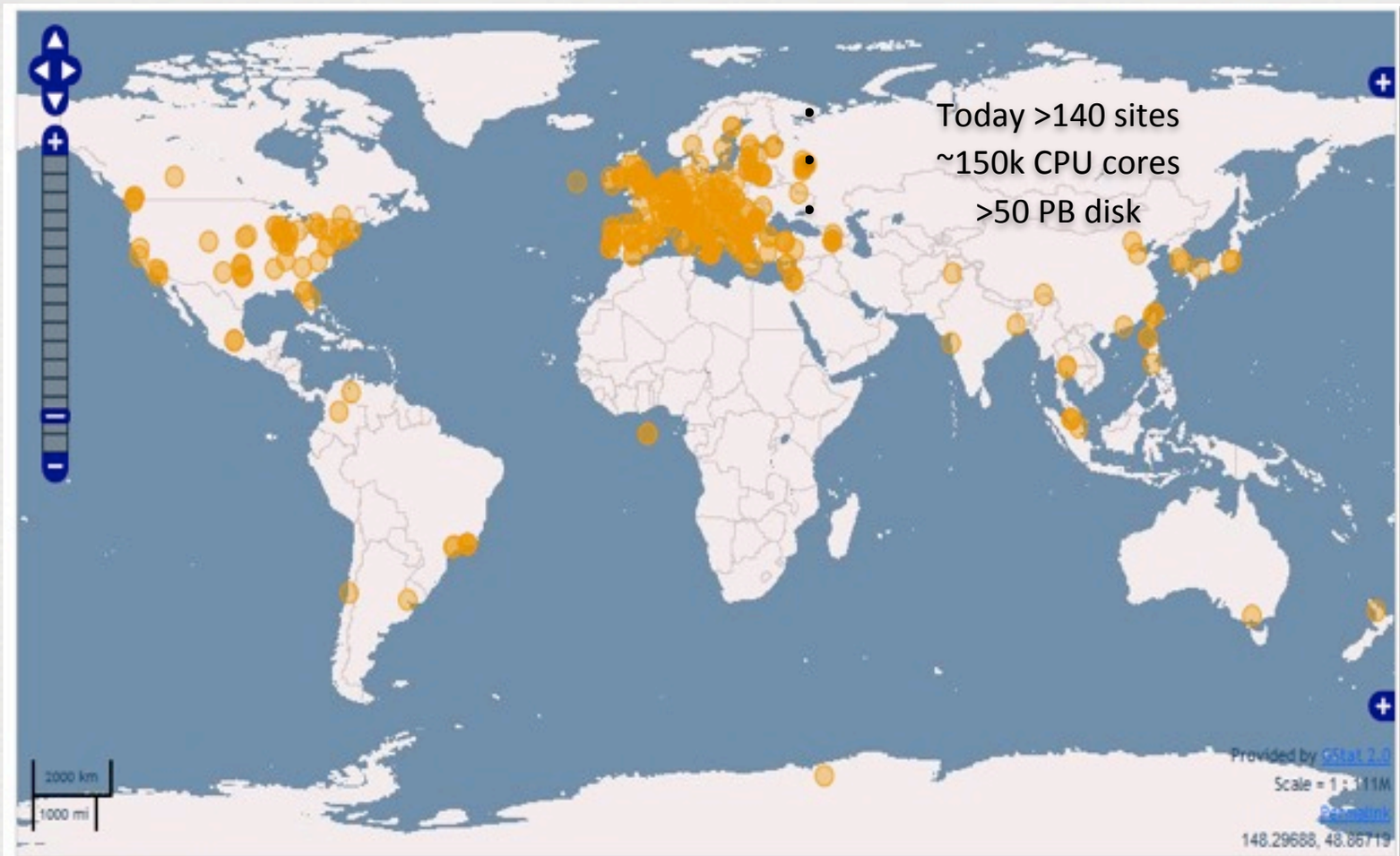
ALICE Collaboration
~ 1/2 ATLAS, CMS, ~ 2x LHCb
~1000 people, 30 countries,
~ 80 Institutes

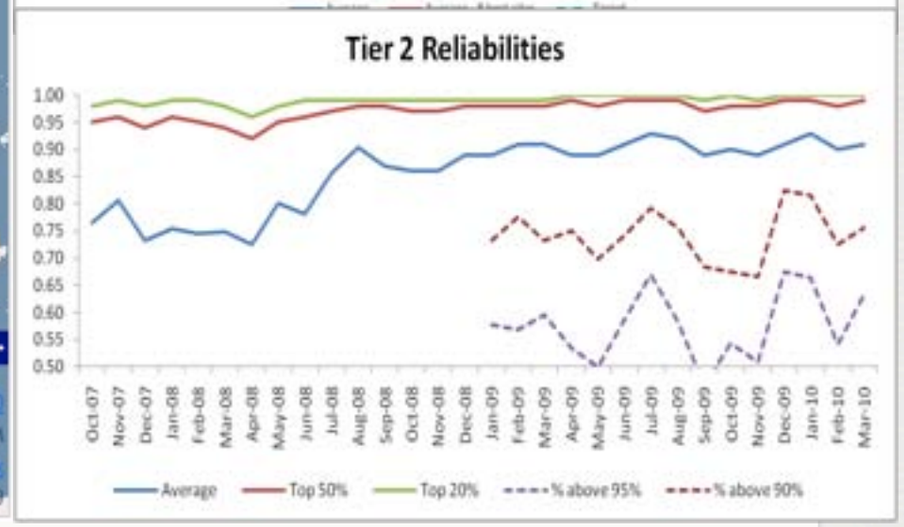
Total weight	10,000t
Overall diameter	16.00m
Overall length	25m
Magnetic Field	0.5Tesla

8 kHz (160 GB / sec)
level 0 - special hardware
200 Hz (4 GB / sec)
level 1 - embedded processors
30 Hz (2.5 GB / sec)
level 2 - PCs
30 Hz
(1.25 GB / sec)
data recording &
offline analysis

Now 4GB/s!!!

- A full pp programme
- Data rate for pp is 100Hz@1MB

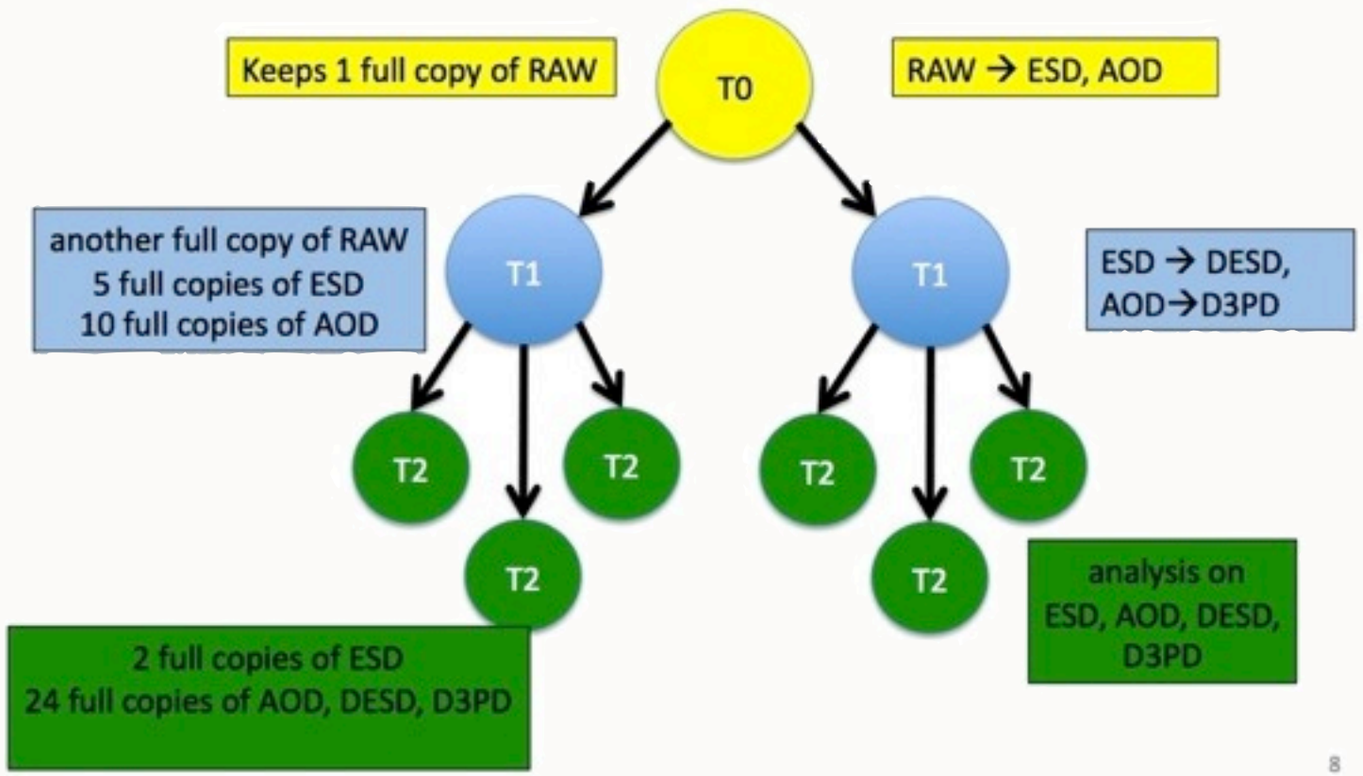




- Site readiness very good
- However the most common reason for job failure is indeed site misconfiguration
  - Apart of course from user errors
- Instabilities coming from BDII are less frequent now
- Good news is that availability is constantly improving
  - Even if the human cost can be very high

THE MONARC MODEL

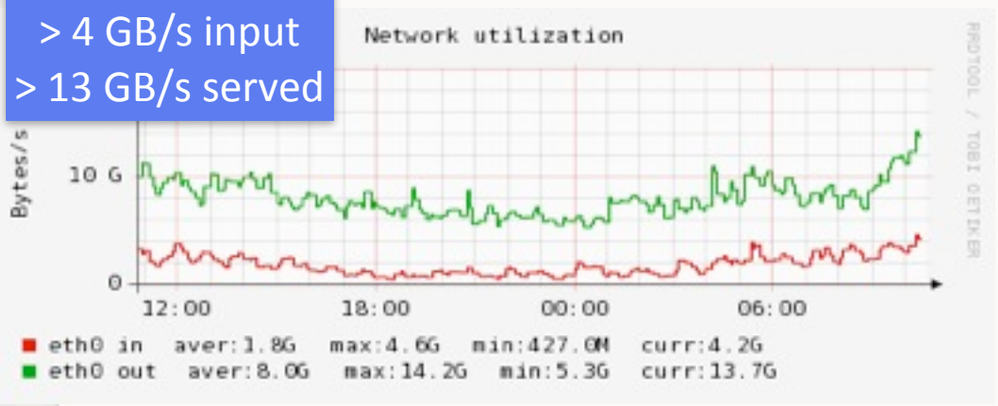
- The Monarc model was designed at the end of the last century based on a "rigid" distribution of tasks between centres of different size and role

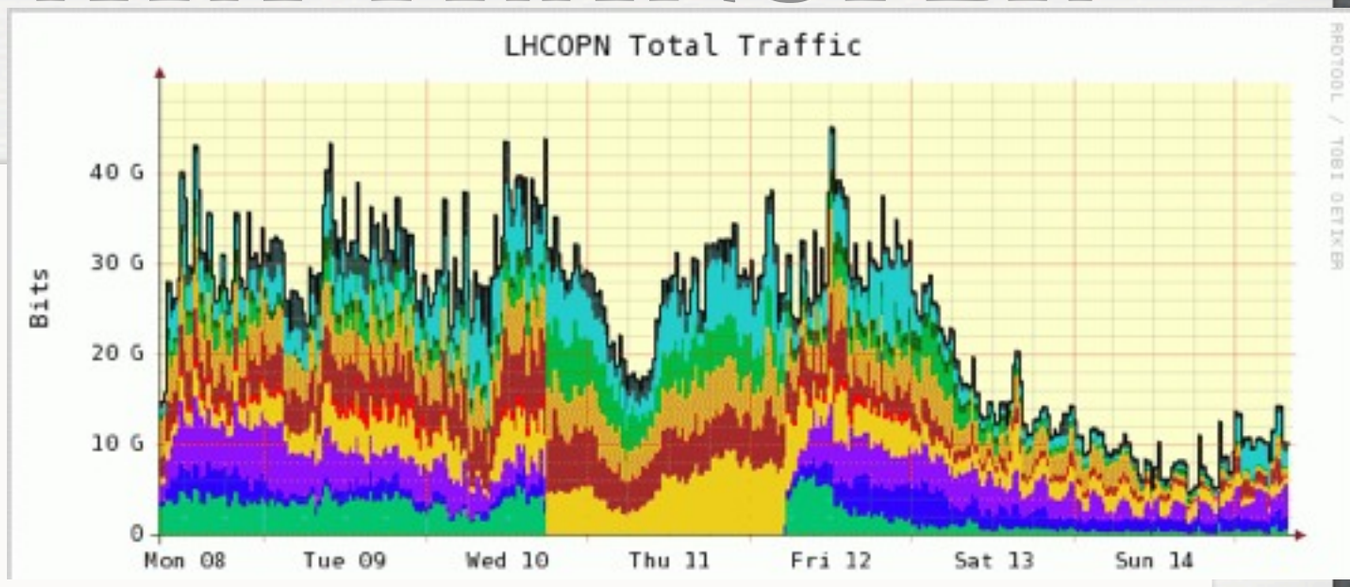


THE GRID – DATA TRANSFER

- Data transfer has been especially successful
  - Out of CERN has peaked above 1GB
  - ...een centres also very good

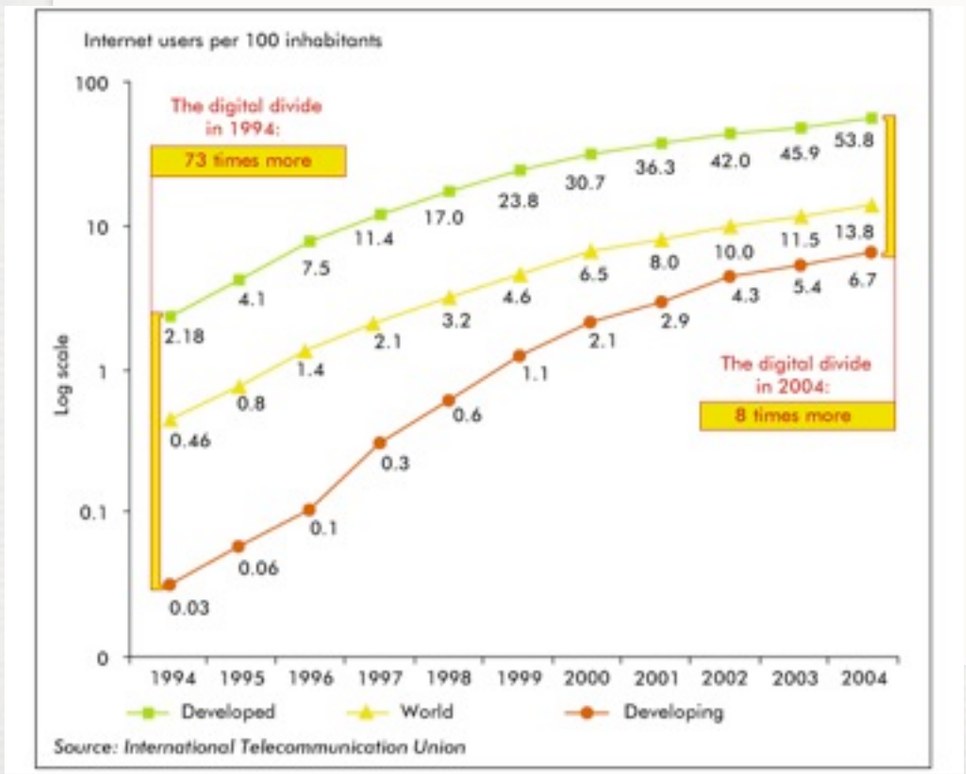
Tier 0 traffic:
> 4 GB/s input
> 13 GB/s served

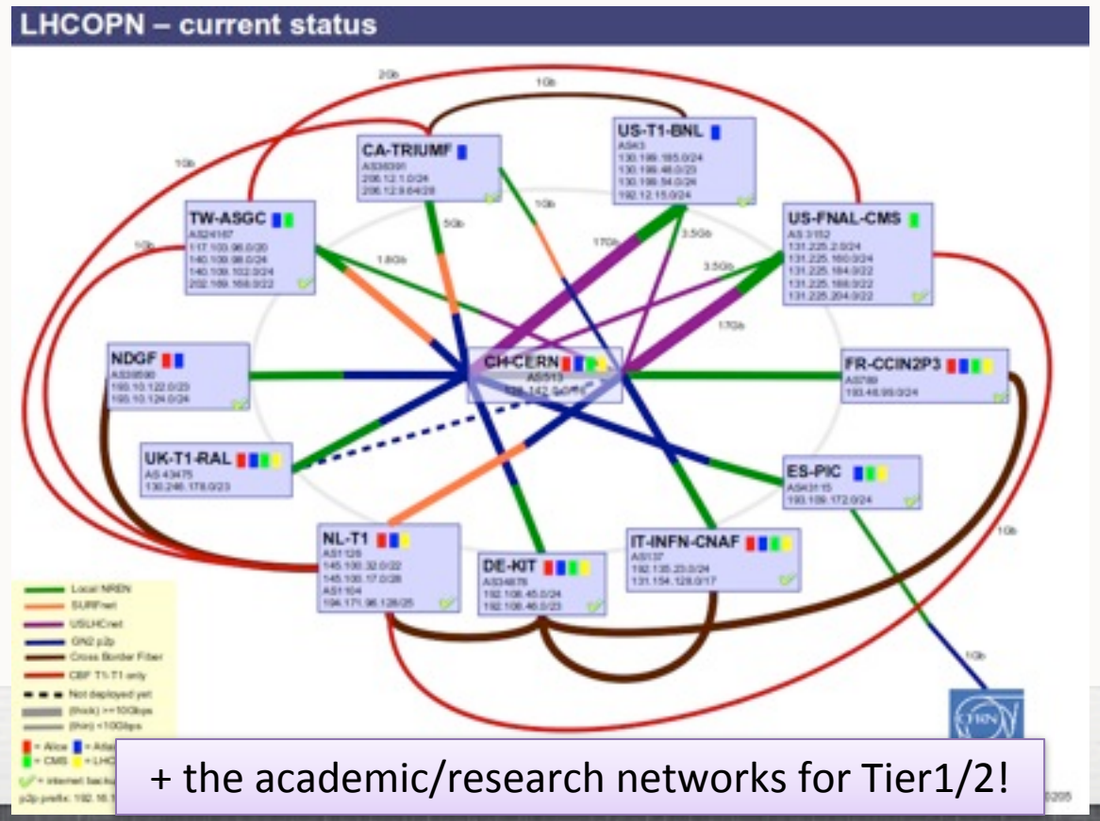




- The network is probably the best surprise here
- Still the least oversubscribed resource we have

Out of CERN

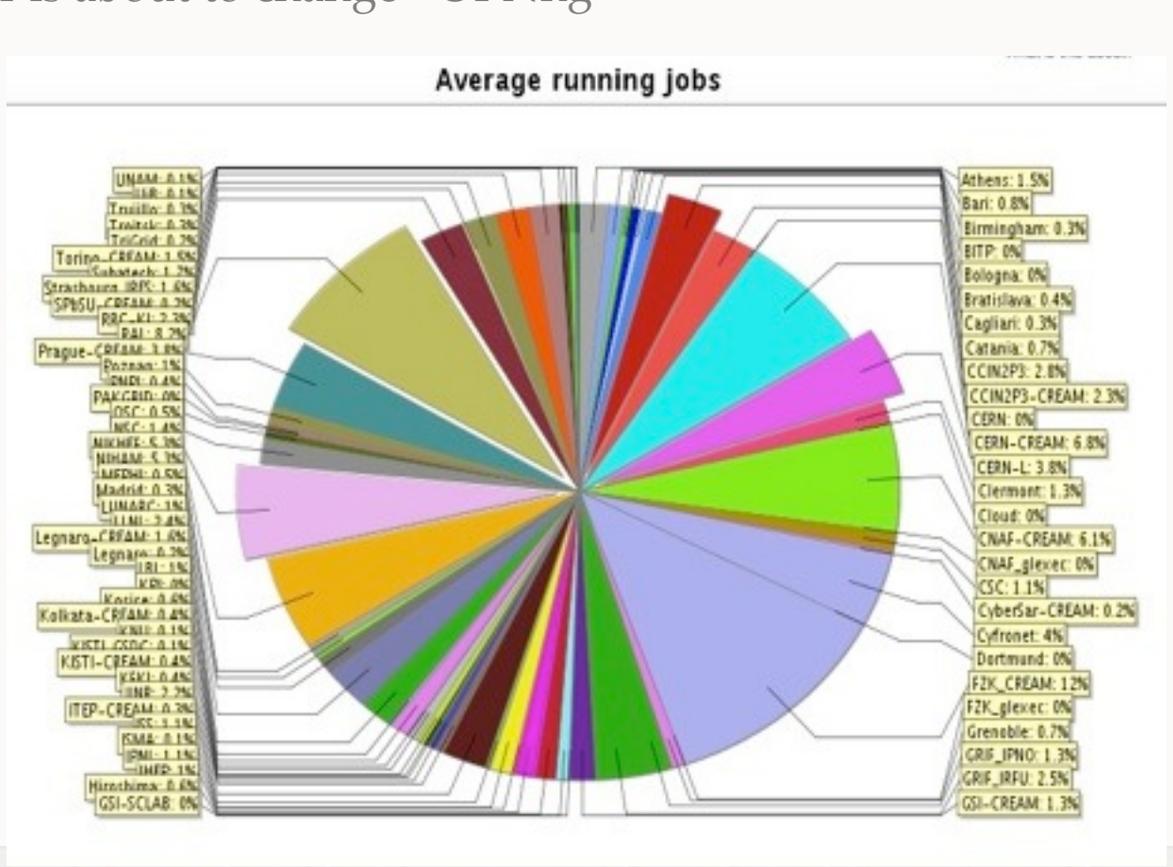




+ the academic/research networks for Tier1/2!

THE GRID – RELATIONS T1-T2-T3

- T2 have been a very good surprise
  - More than 50% of the work in ALICE is done by T2
- The Grid is becoming more and more “cloudy”
  - Not really clear the difference between T1s and T2s apart from data custodial and better network
    - but the latter is about to change - OPNng

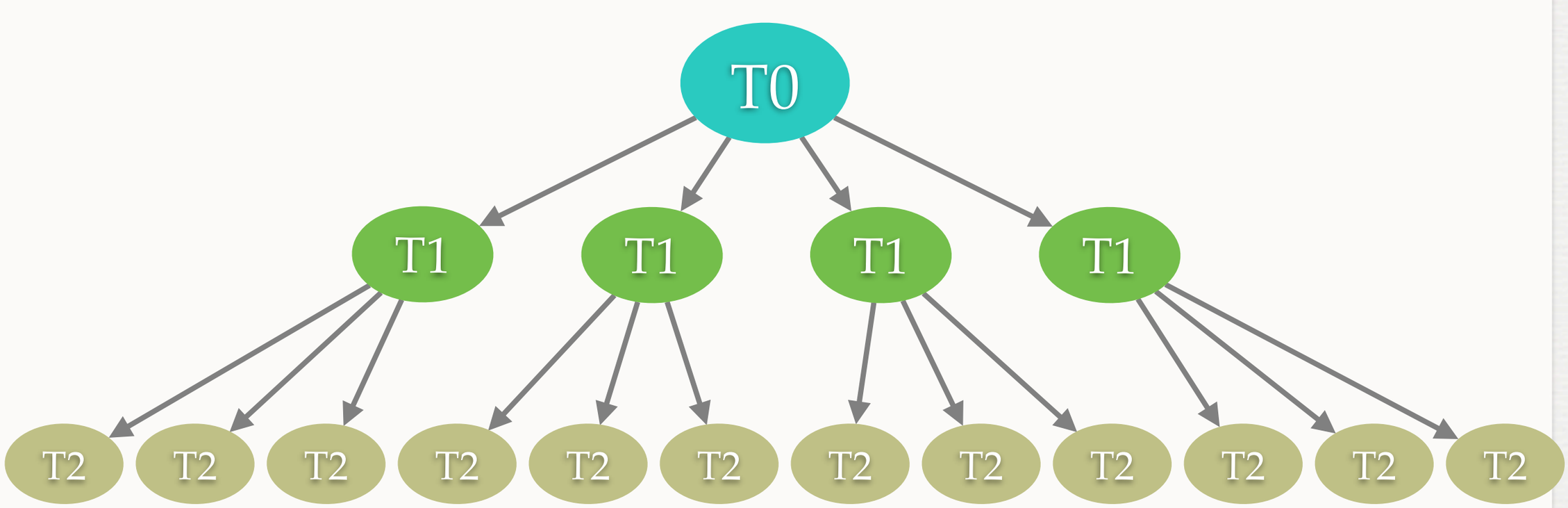


DESTITUTION OF THE MONARC

- Given the good performance of the network and the issues with data placement, the Monarc model is evolving from Grid to Cloud

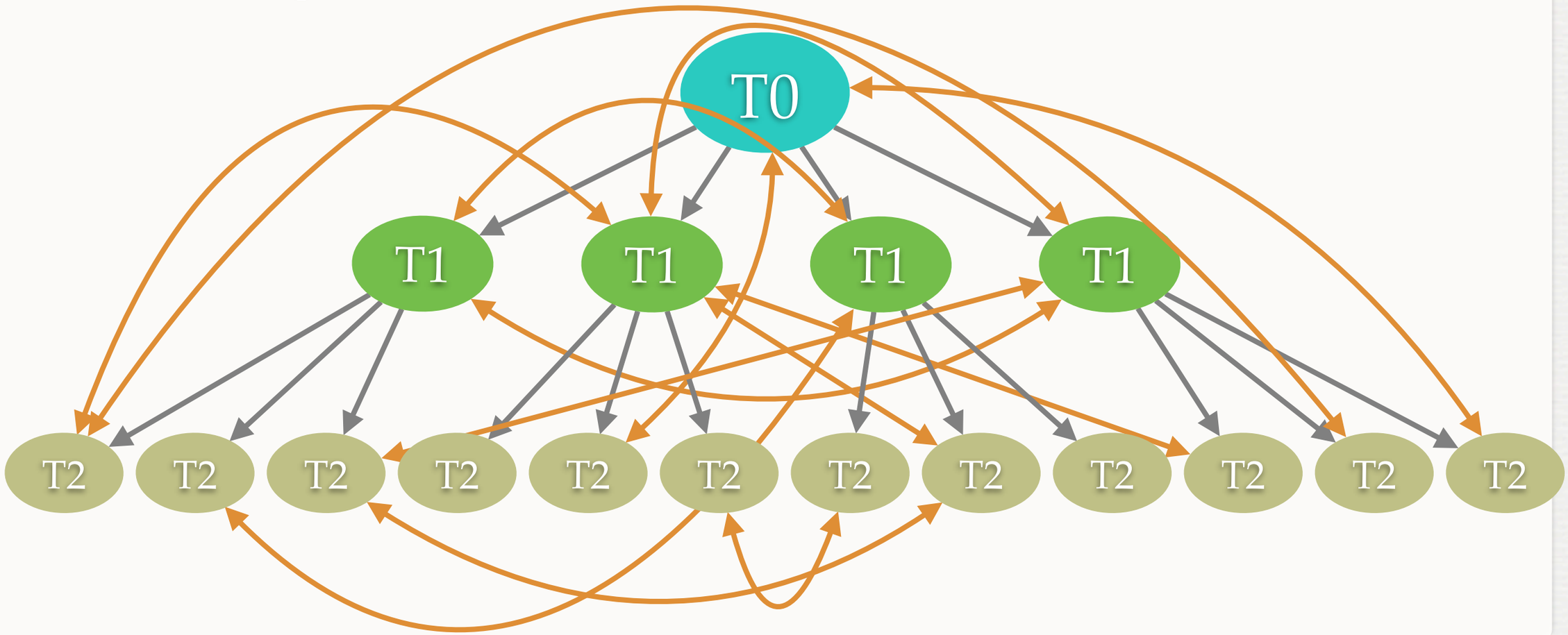
DESTITUTION OF THE MONARC

- Given the good performance of the network and the issues with data placement, the Monarc model is evolving from Grid to Cloud



DESTITUTION OF THE MONARC

- Given the good performance of the network and the issues with data placement, the Monarc model is evolving from Grid to Cloud



The ALICE Grid

T1

The ALICE Grid

T1
NorduGrid

The ALICE Grid

T1
NorduGrid
NIKHEF/SARA

The ALICE Grid

T1
NorduGrid
NIKHEF/SARA
RAL

The ALICE Grid

T1
NorduGrid
NIKHEF/SARA
RAL
FZK

The ALICE Grid

T1
NorduGrid
NIKHEF/SARA
RAL
FZK
CERN

The ALICE Grid

T1
NorduGrid
NIKHEF/SARA
RAL
FZK
CERN
CCIN2P3

The ALICE Grid

T1
NorduGrid
NIKHEF/SARA
RAL
FZK
CERN
CCIN2P3
CNAF

The ALICE Grid

T1
NorduGrid
NIKHEF/SARA
RAL
FZK
CERN
CCIN2P3
CNAF
UNAM

The ALICE Grid

T1
NorduGrid
NIKHEF/SARA
RAL
FZK
CERN
CCIN2P3
CNAF
UNAM
KISTI

The ALICE Grid

T1
NorduGrid
NIKHEF/SARA
RAL
FZK
CERN
CCIN2P3
CNAF
UNAM
KISTI
????

The ALICE Grid

T1
NorduGrid
NIKHEF/SARA
RAL
FZK
CERN
CCIN2P3
CNAF
UNAM
KISTI
????
????

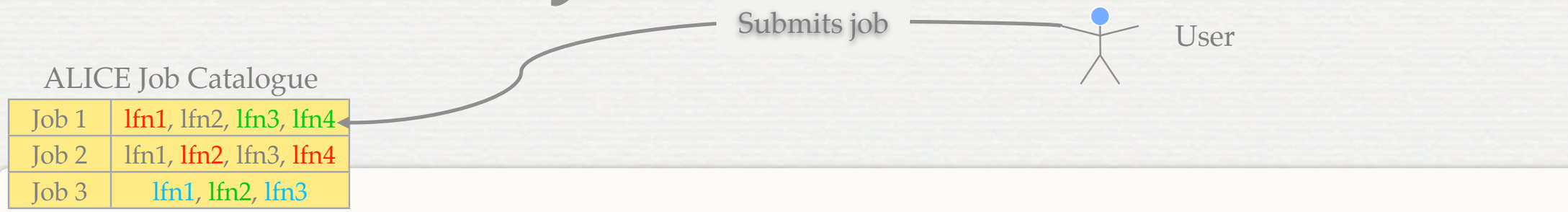
The ALICE Grid

THE GRID – JOB MANAGEMENT

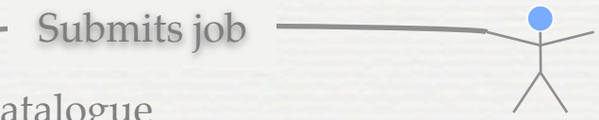
- The priority and quota mechanism is hard to implement Grid-wide
  - Central queues (ATLAS, ALICE) are a single point of failure / bottleneck
  - Distributed queues (CMS) makes it more difficult to manage priorities
- Permissions and quotas on files are also a problem
  - See above for central vs distributed catalogues
- "Upgrading" the Grid is a very long process
  - CREAM
  - SL5
  - glexec
- EMI / EGI may still change the pattern

SENDING JOBS TO DATA

SENDING JOBS TO DATA



SENDING JOBS TO DATA



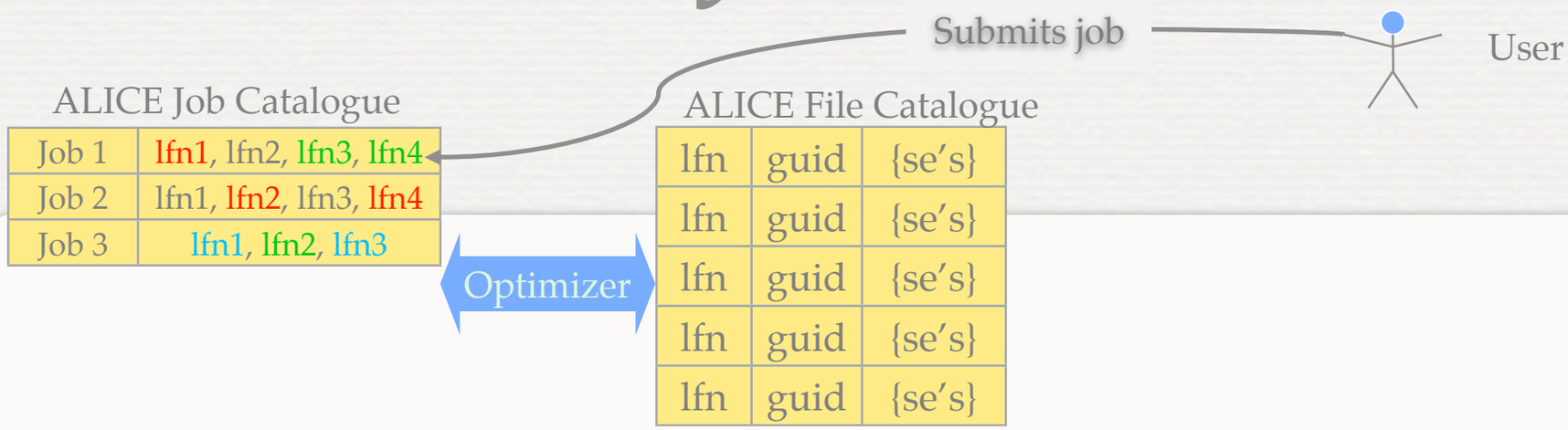
ALICE Job Catalogue

Job 1	lfn1, lfn2, lfn3, lfn4
Job 2	lfn1, lfn2, lfn3, lfn4
Job 3	lfn1, lfn2, lfn3

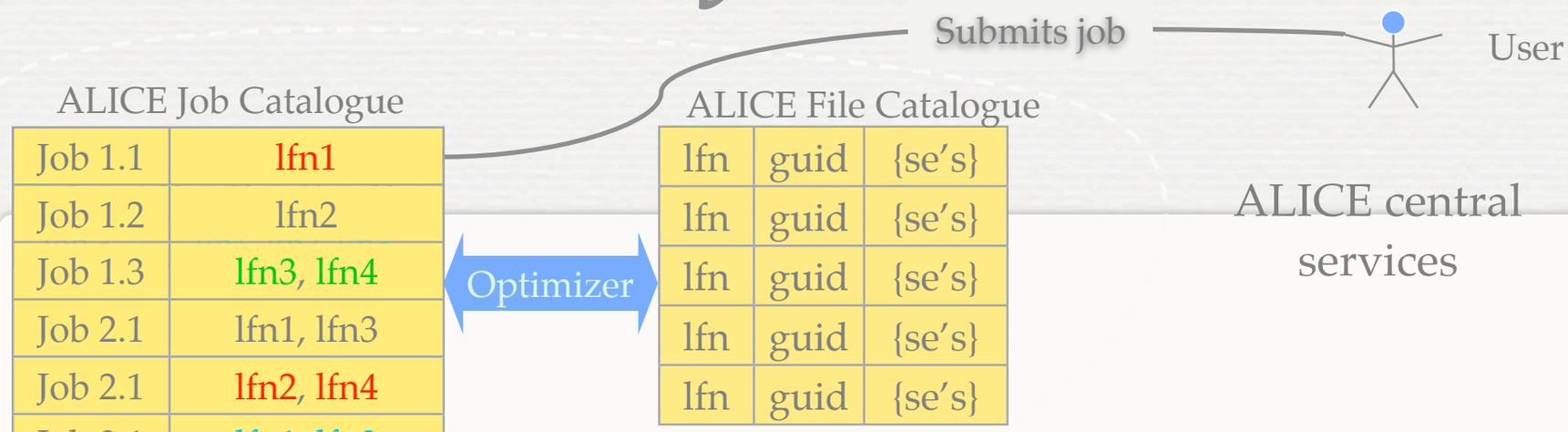
ALICE File Catalogue

lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}

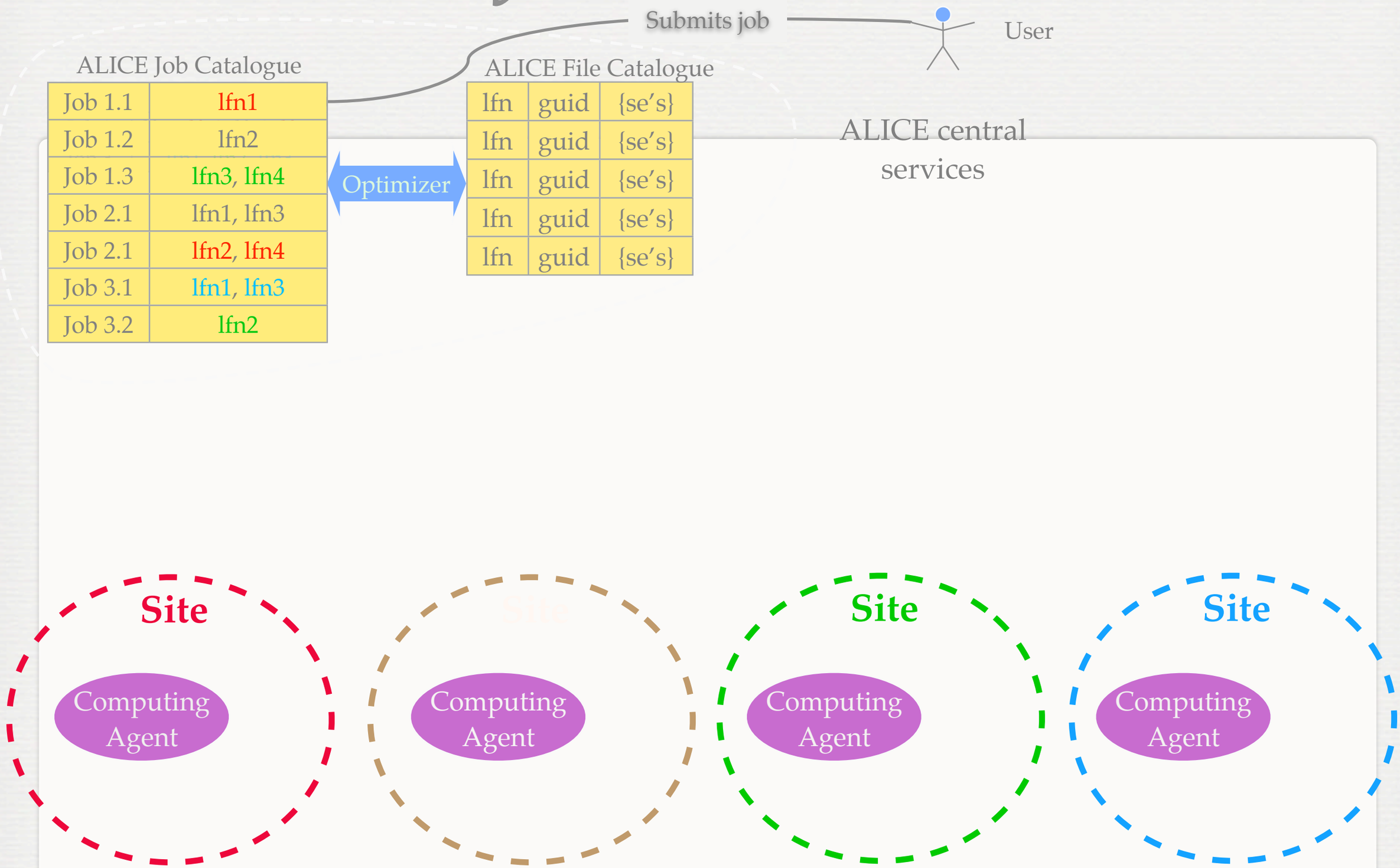
SENDING JOBS TO DATA



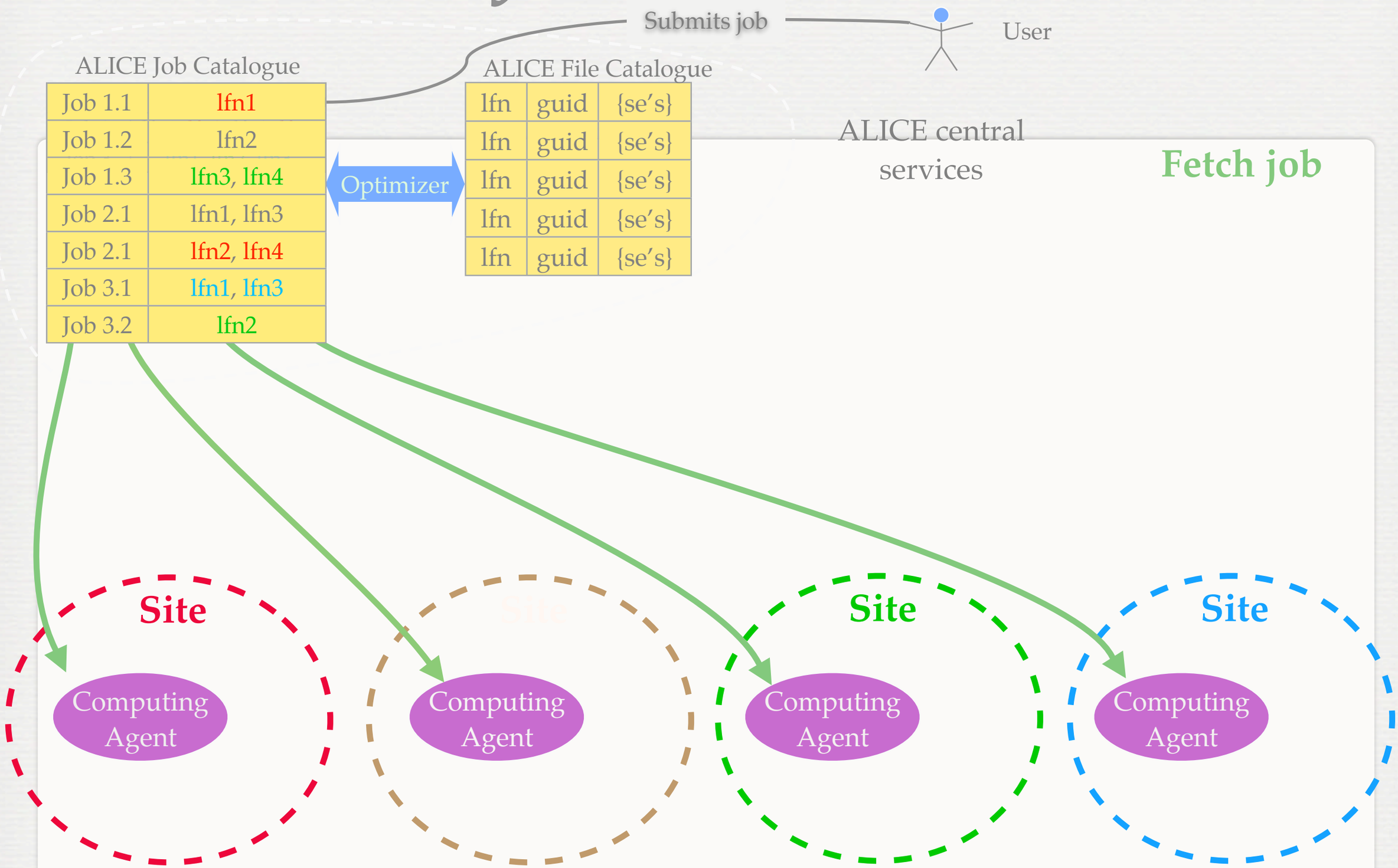
SENDING JOBS TO DATA



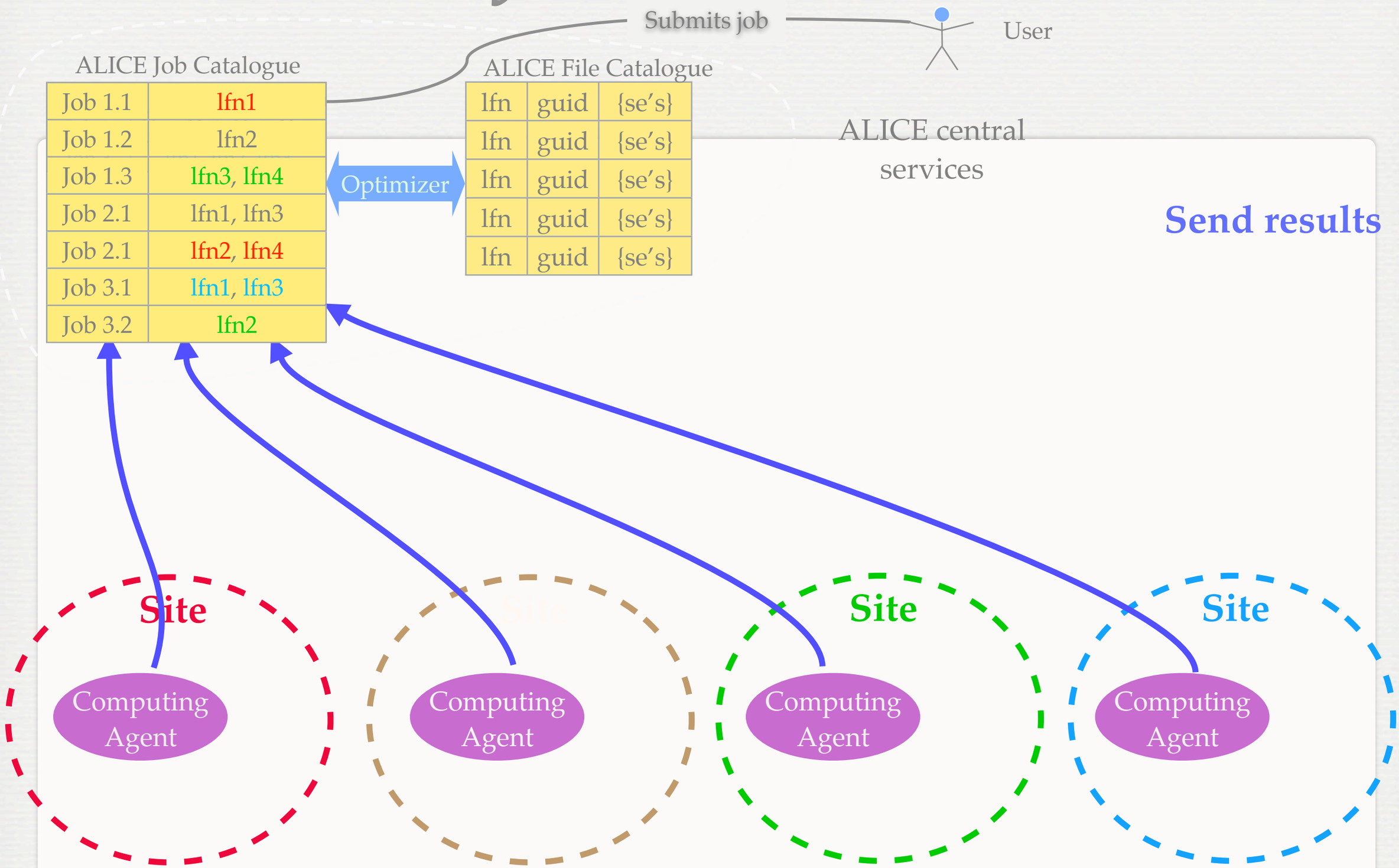
SENDING JOBS TO DATA



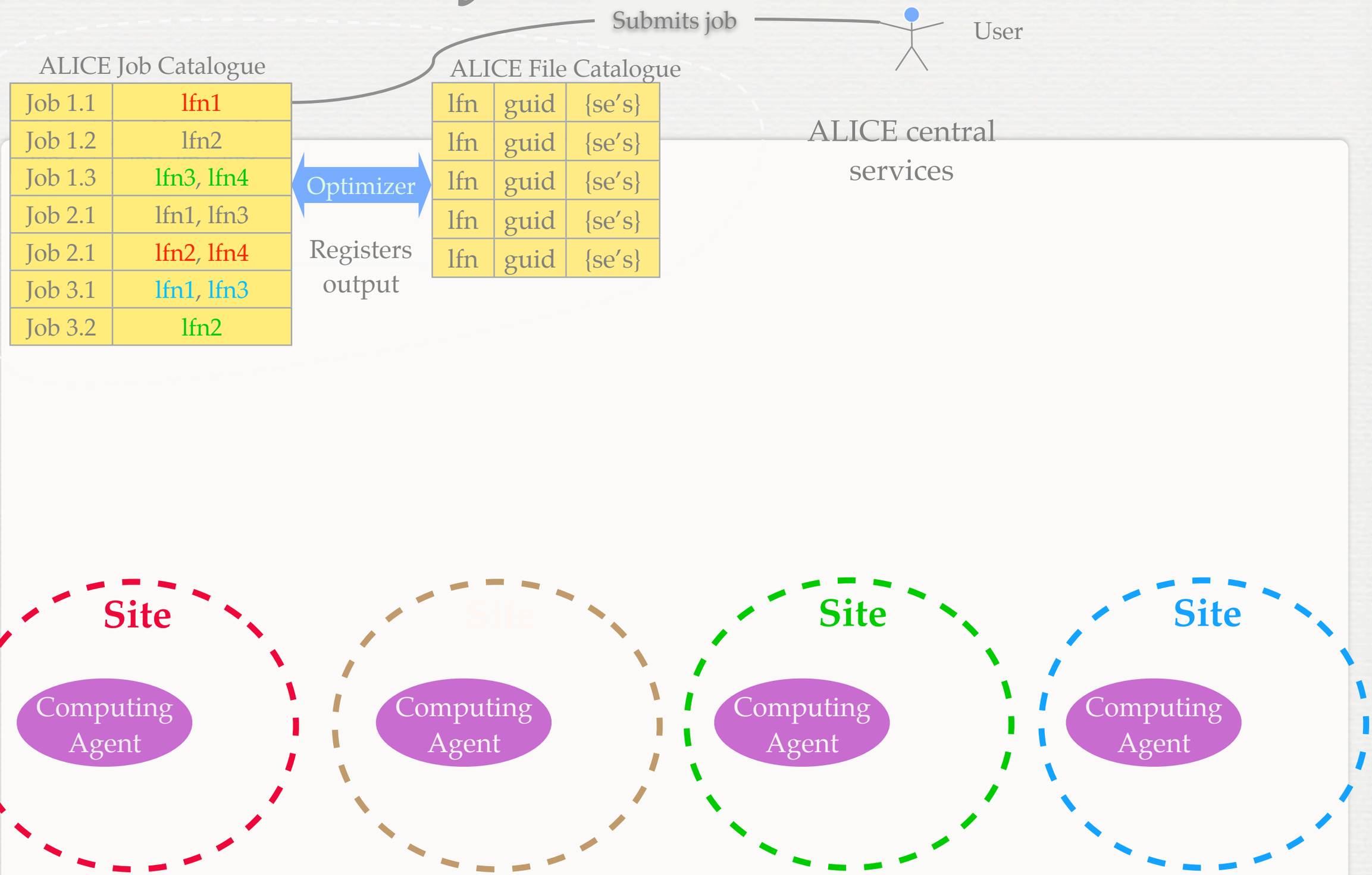
SENDING JOBS TO DATA



SENDING JOBS TO DATA

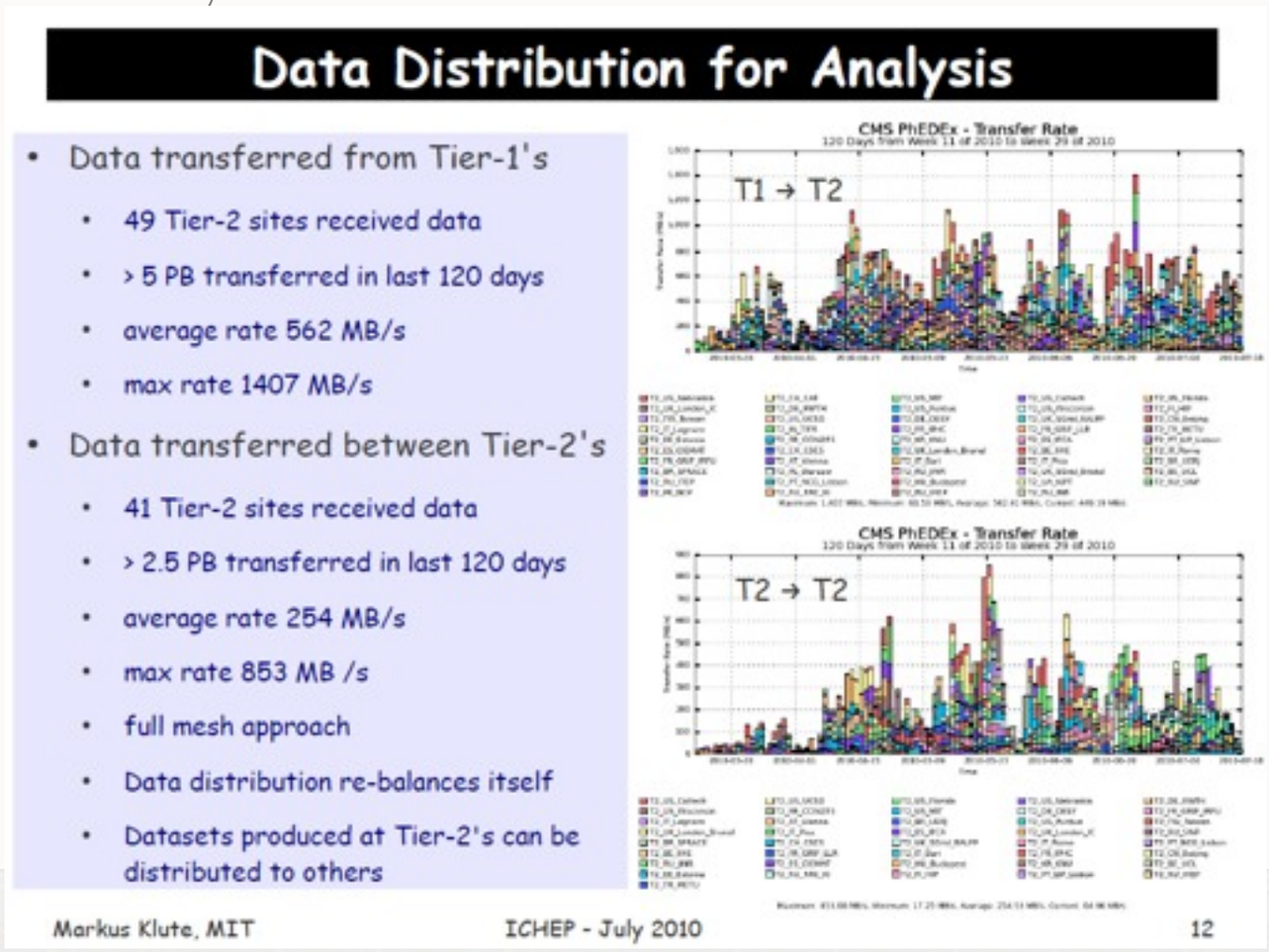


SENDING JOBS TO DATA



DATA IS STILL THE PROBLEM

- Data placement is the main problem, particularly for analysis
  - "predictive" data placement for ATLAS and CMS
  - "opportunistic" data placement for ALICE
- Data distribution "per se" works very well
- With "infinite" disk space the two would be equivalent
- "opportunistic" data distribution depends on a single central catalogue
- It took ALICE 10 years to get there!

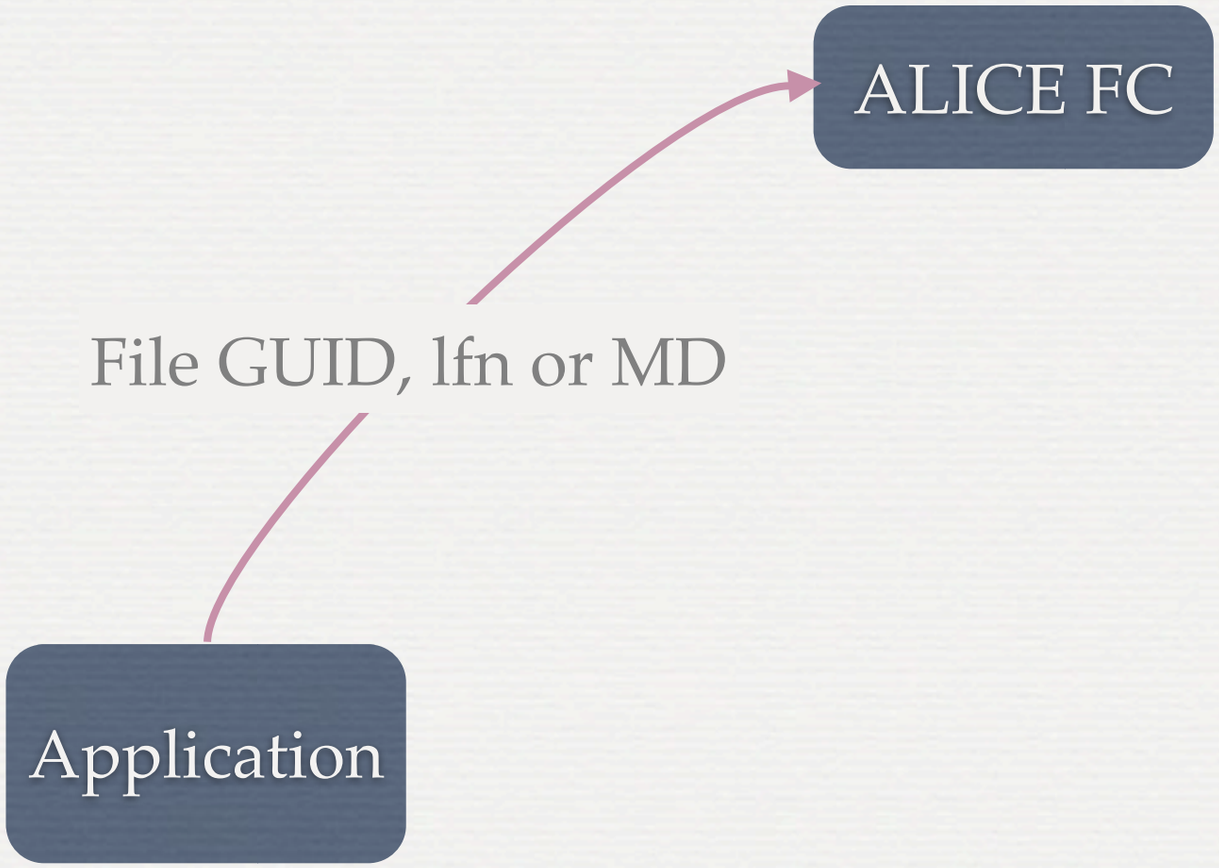


ALICE FILE CATALOGUE

Application

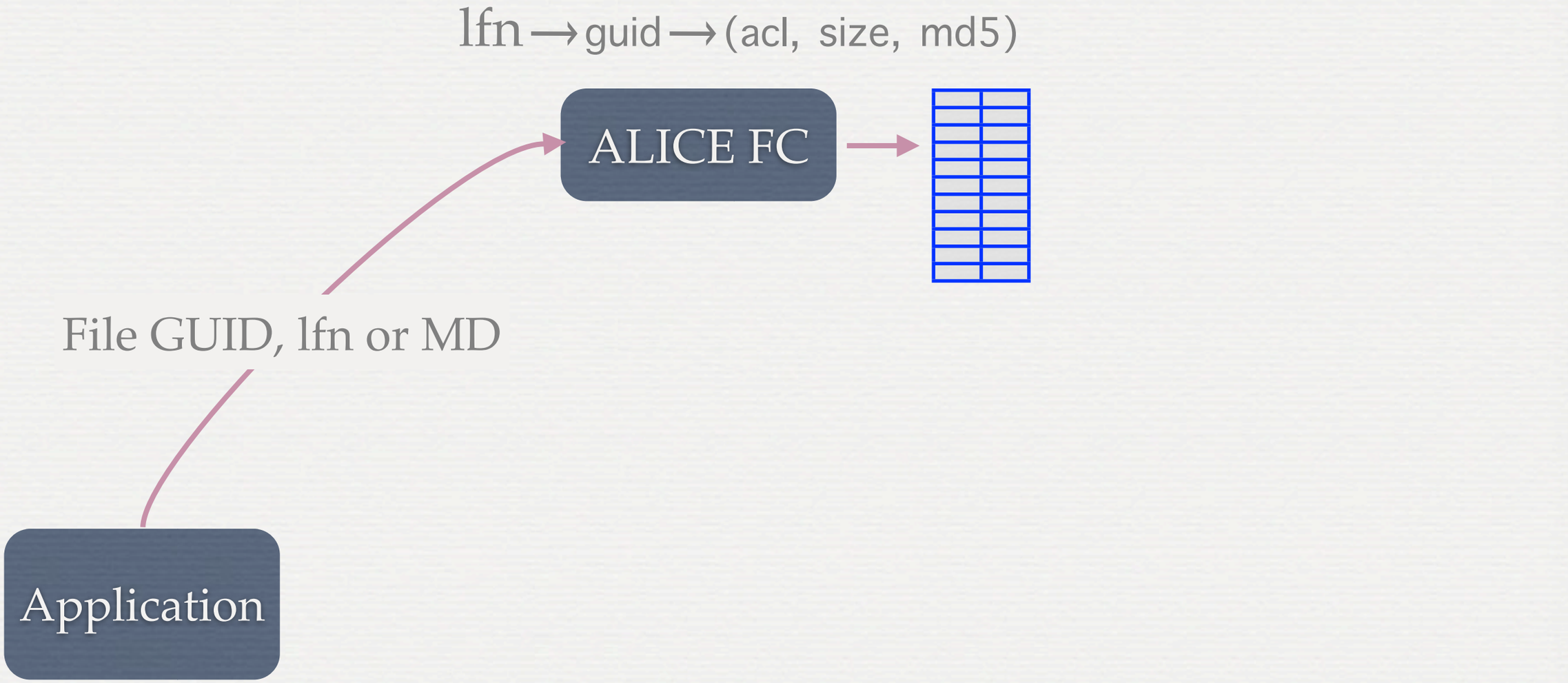
Direct access to data
via TAliEn/TGrid interface

ALICE FILE CATALOGUE



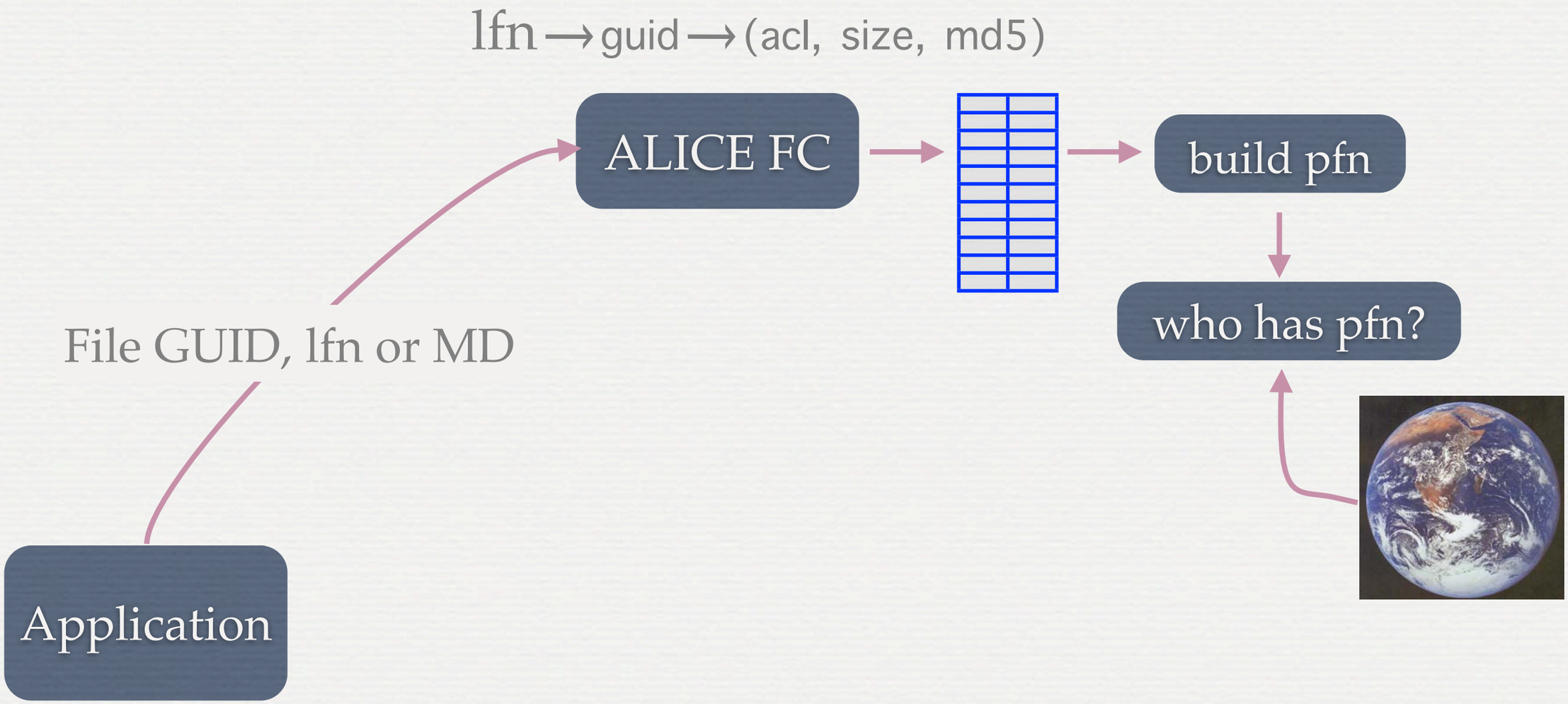
Direct access to data
via TAliEn/TGrid interface

ALICE FILE CATALOGUE



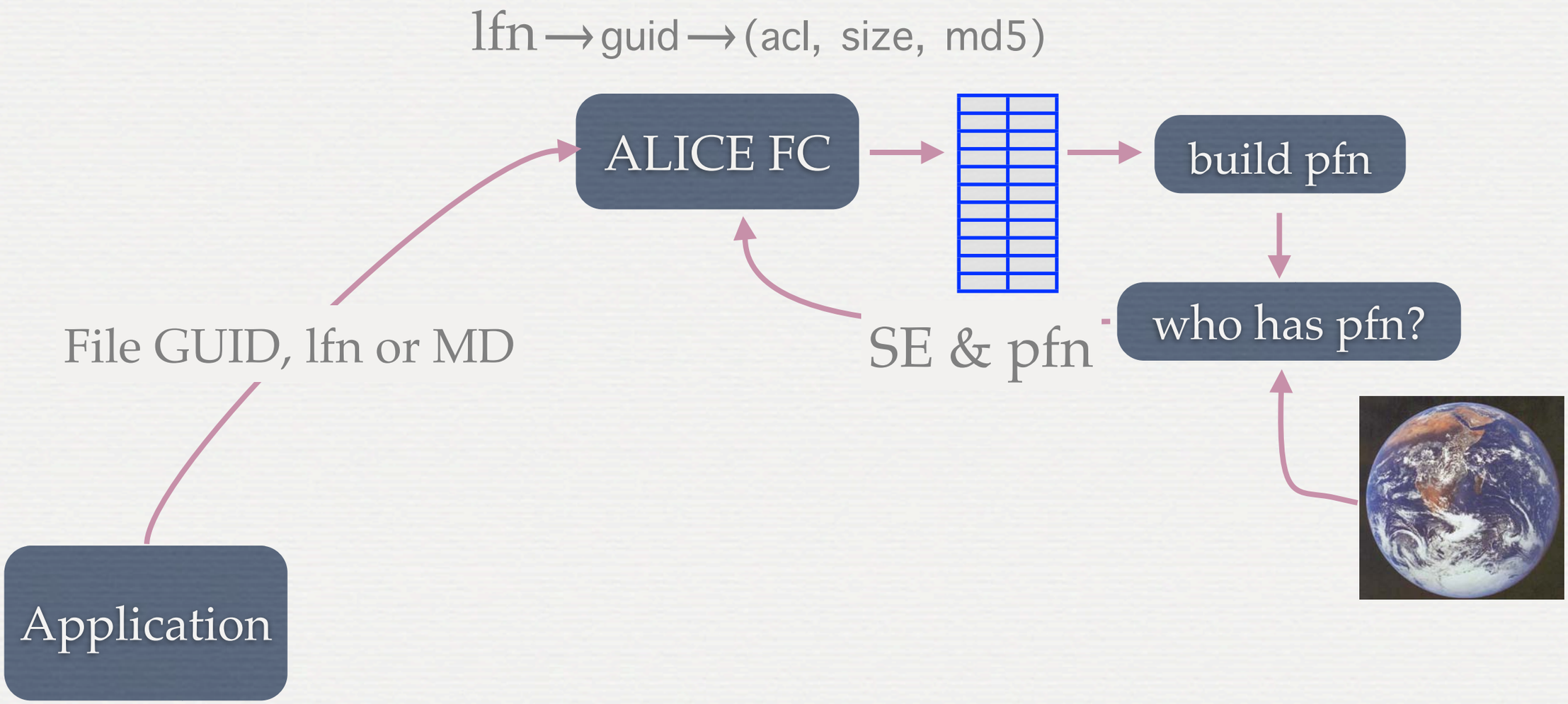
Direct access to data
via TAliEn/TGrid interface

ALICE FILE CATALOGUE



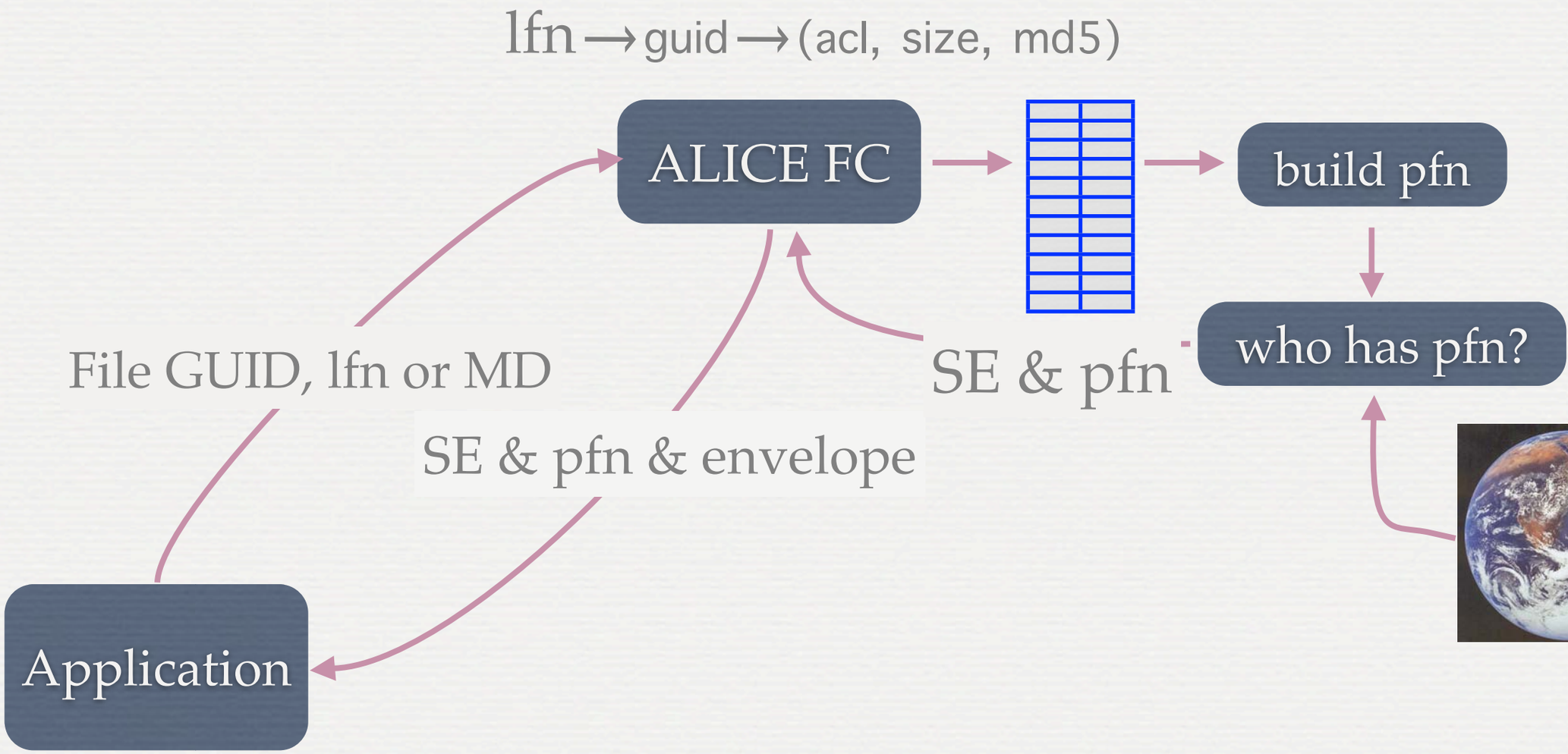
Direct access to data
via TAliEn/TGrid interface

ALICE FILE CATALOGUE



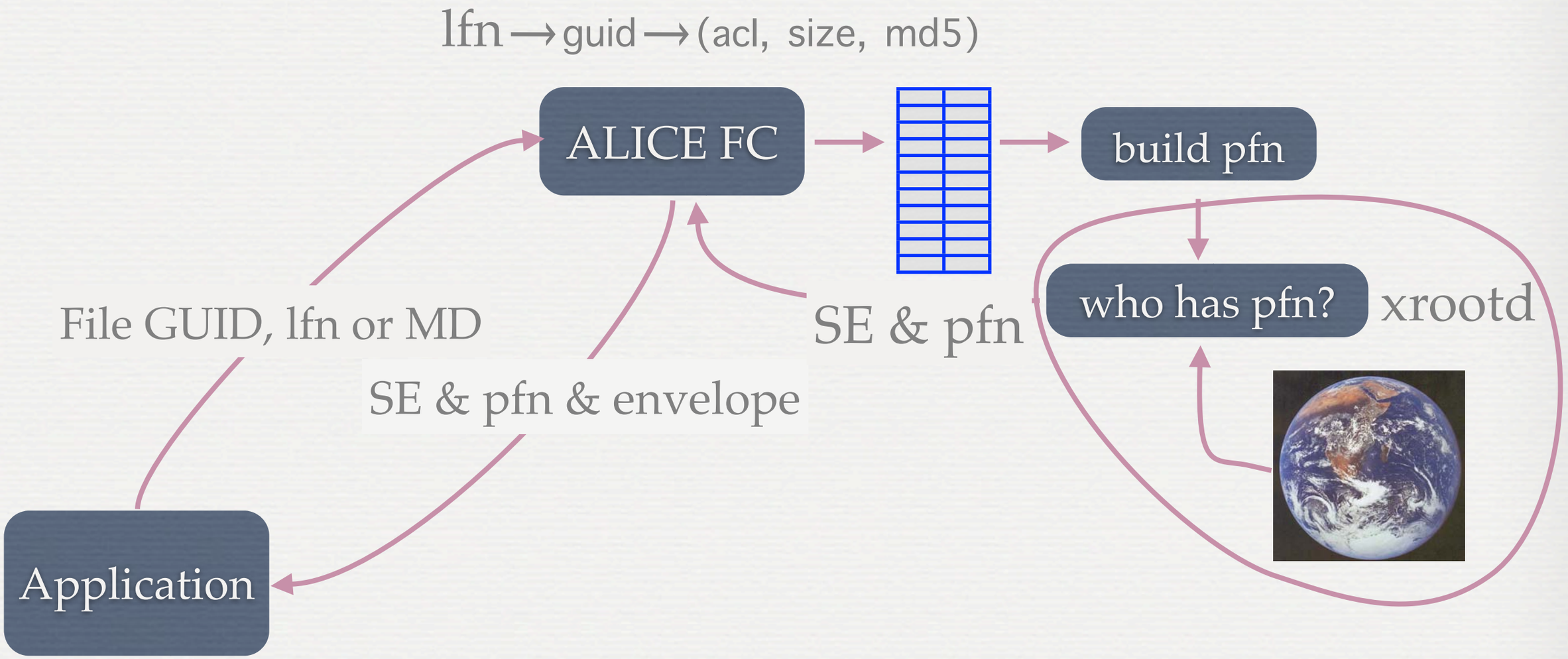
Direct access to data
via TAliEn/TGrid interface

ALICE FILE CATALOGUE



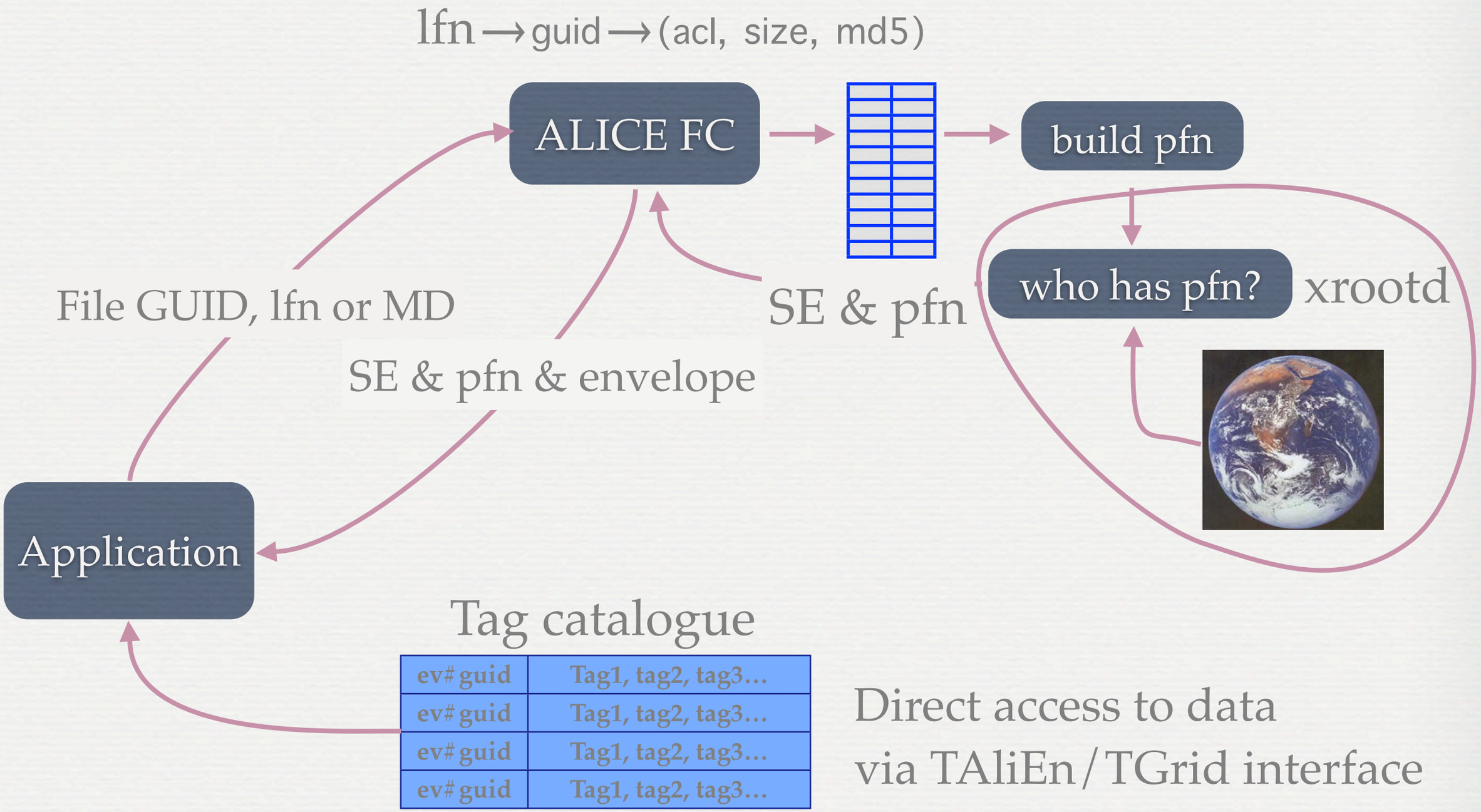
Direct access to data
via TAliEn/TGrid interface

ALICE FILE CATALOGUE



Direct access to data
via TAliEn/TGrid interface

ALICE FILE CATALOGUE

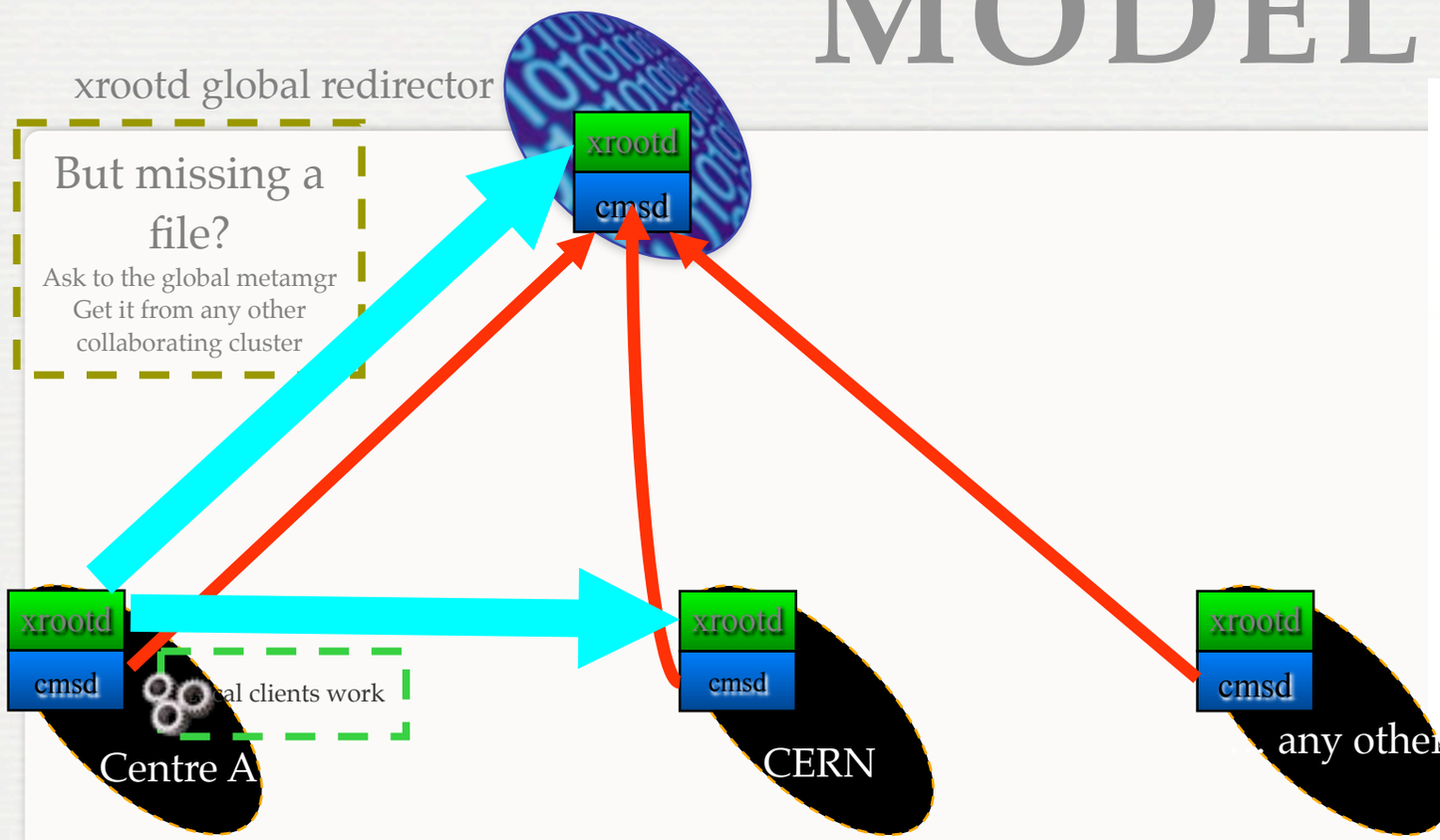


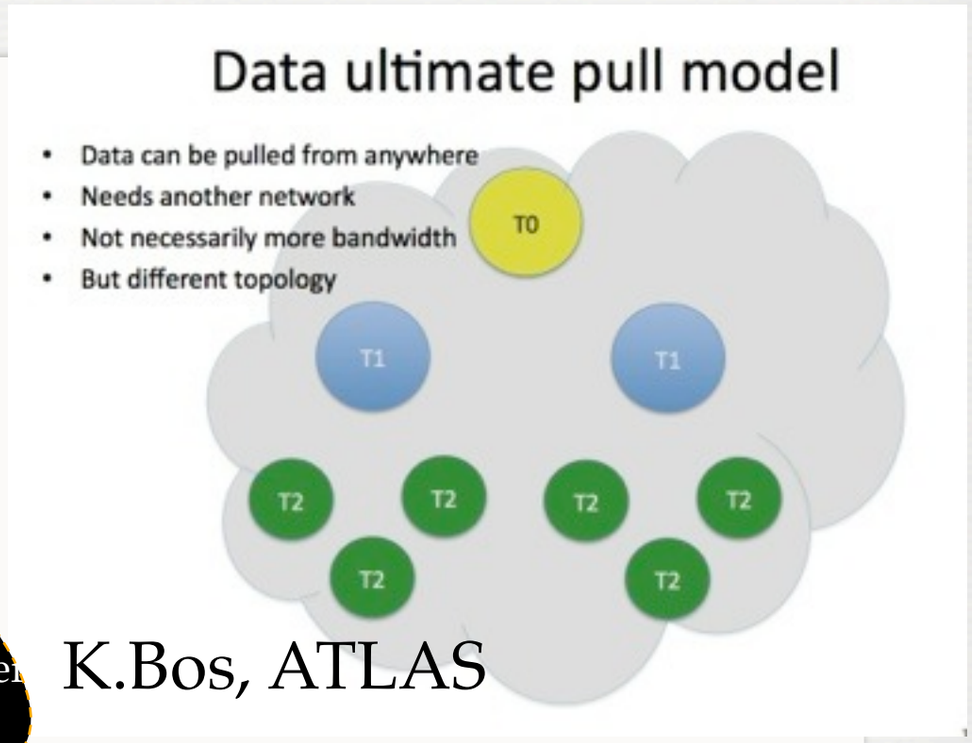
Tag catalogue

ev# guid	Tag1, tag2, tag3…
ev# guid	Tag1, tag2, tag3…
ev# guid	Tag1, tag2, tag3…
ev# guid	Tag1, tag2, tag3…

Direct access to data
via TAliEn/TGrid interface

COMPLETE "PULL" MODEL?





- With careful caching and overlapping access over the network can be slower by a factor 2-3
- xrootd offers this now
  - Will other products go the same way soon?

HOW TO OPTIMISE STORAGE?

- How to efficiently write N replicas of a file ?
- 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)

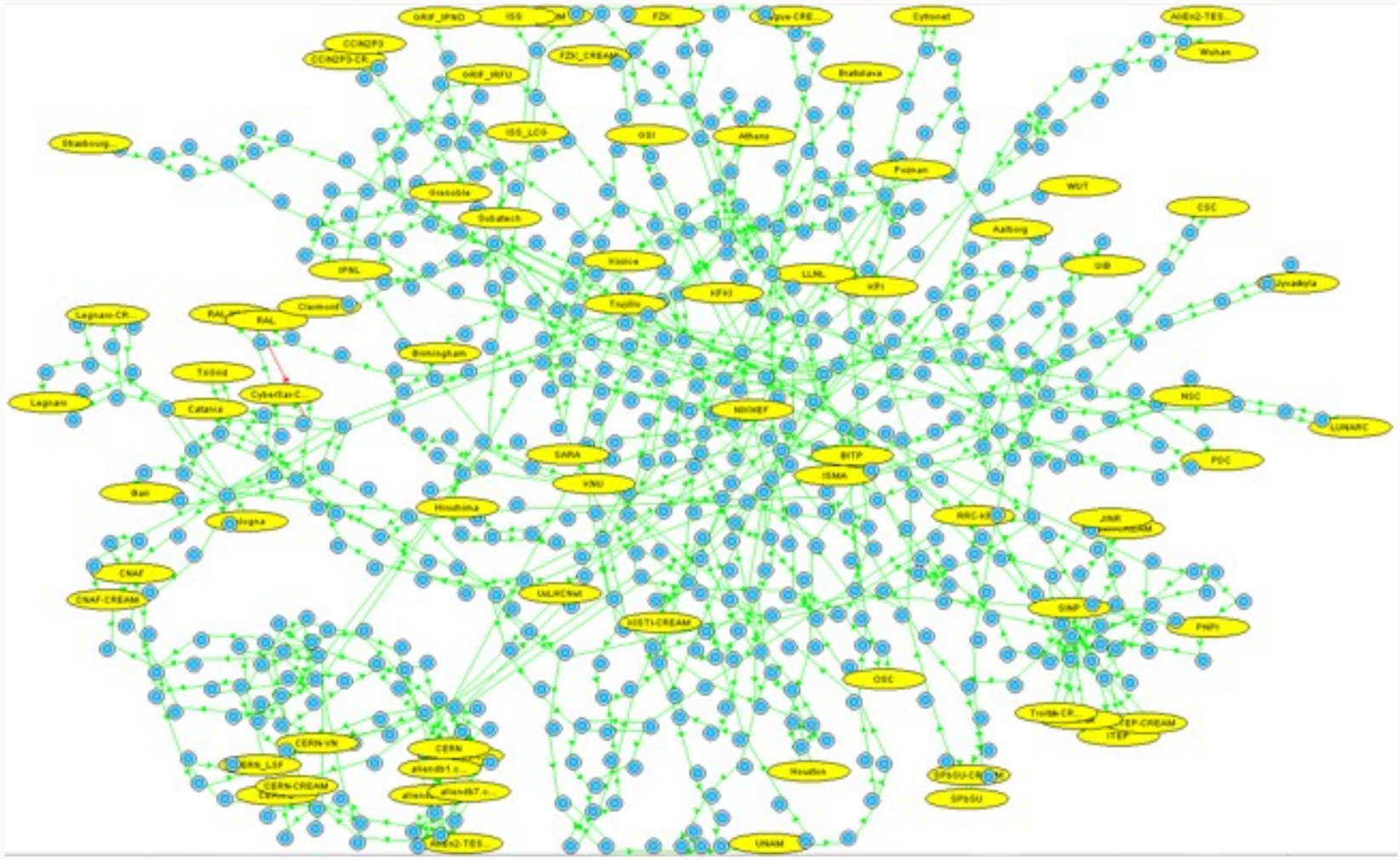
	Statistics					Functional tests						Last day tests	
SE Name	Size	Used	Free	Usage	No. of files	add	ls	get	whereis	rm	Last OK test	Successful	Failed
1. Bari - SE	33.69 TB	1.398 TB	32.29 TB	4.149%	75,820						25.02.2010 06:00	12	0
2. Bologna - SE	500 GB	94.45 GB	405.6 GB	18.89%	28,280	Feb ...	Last...	Last...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

Message

Feb 25 06:00:42 info Getting a security envelope..
Feb 25 06:00:43 info According to the envelope: root://lpnsedpm.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 xrdop!!
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'*)

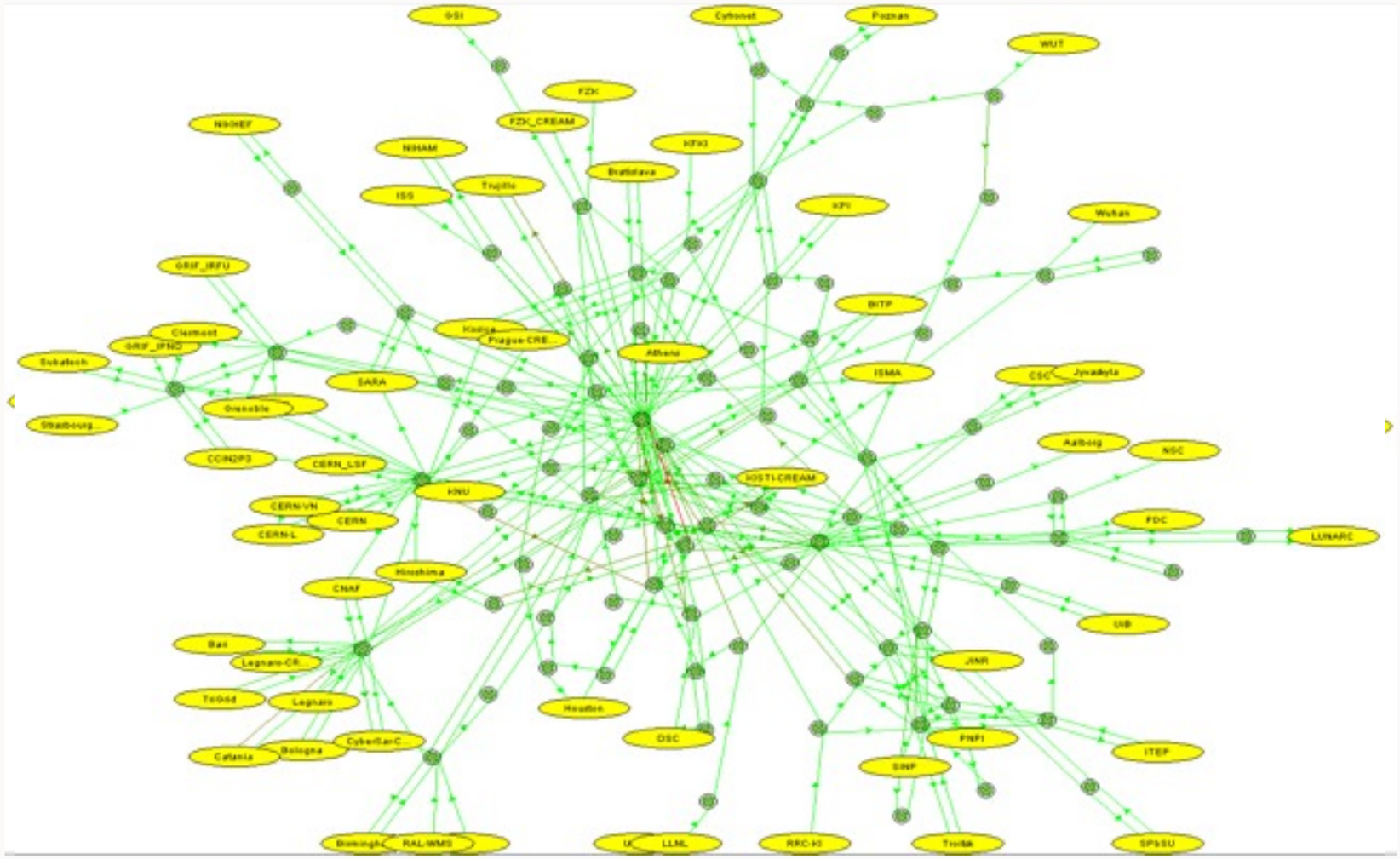
STEP 2 – DISCOVER NETWORK TOPOLOGY

- Each SE is associated a set of IP addresses (VO-Box, xrootd)
- MonALISA records RTT & BW & status between all VO-Boxes



STEP 2 – DISCOVER NETWORK TOPOLOGY

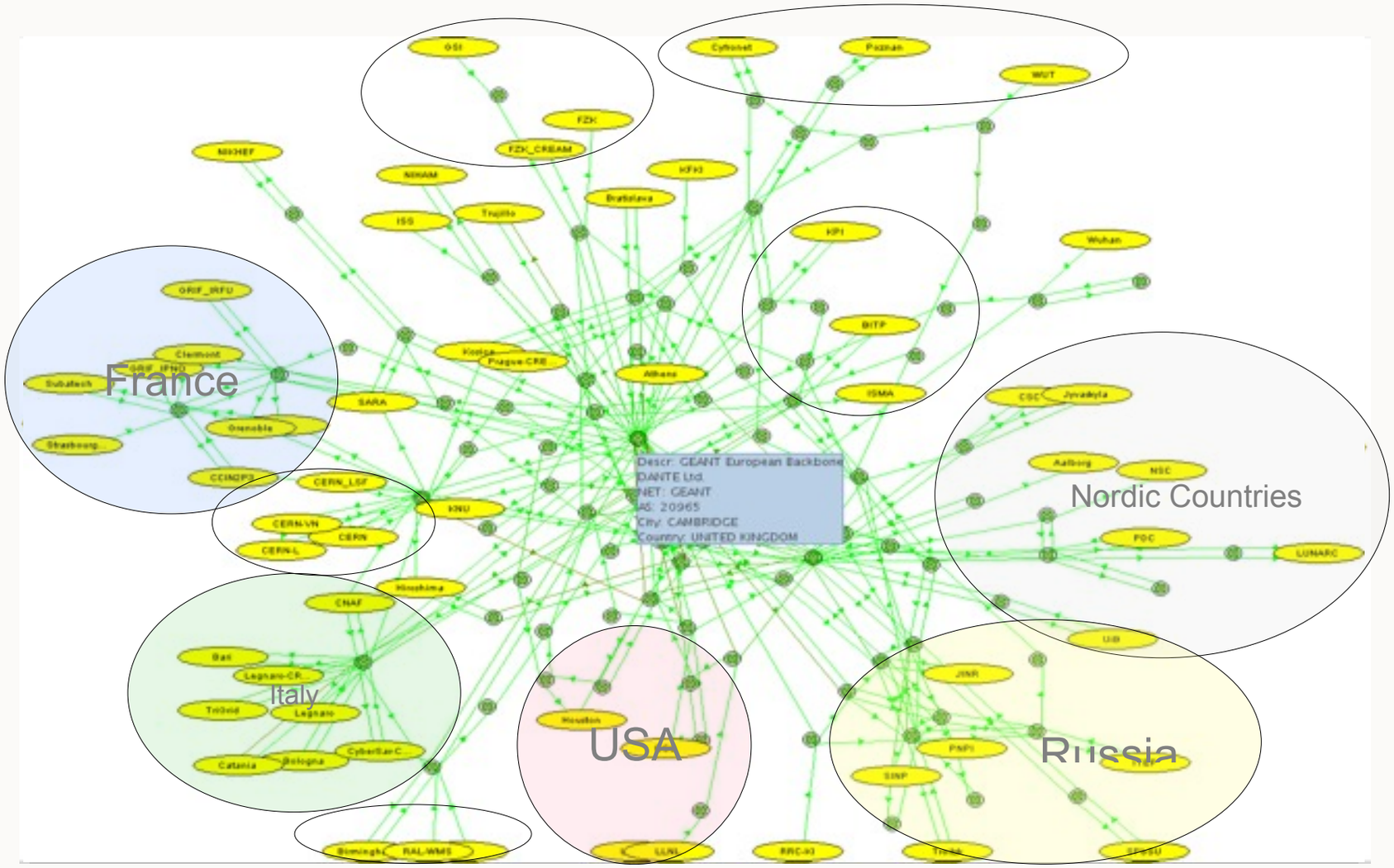
- Each SE is associated a set of IP addresses (VO-Box, xrootd)
- MonALISA records RTT & BW & status between all VO-Boxes



- Group routers in AS
- Measure RTT distance

STEP 2 – DISCOVER NETWORK TOPOLOGY

- Each SE is associated a set of IP addresses (VO-Box, xrootd)
- MonALISA records RTT & BW & status between all VO-Boxes



- Group routers in AS
- Measure RTT distance

STEP 3 – CLIENT TO STORAGE DISTANCE

0

- distance(IP, IP)
  - 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

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

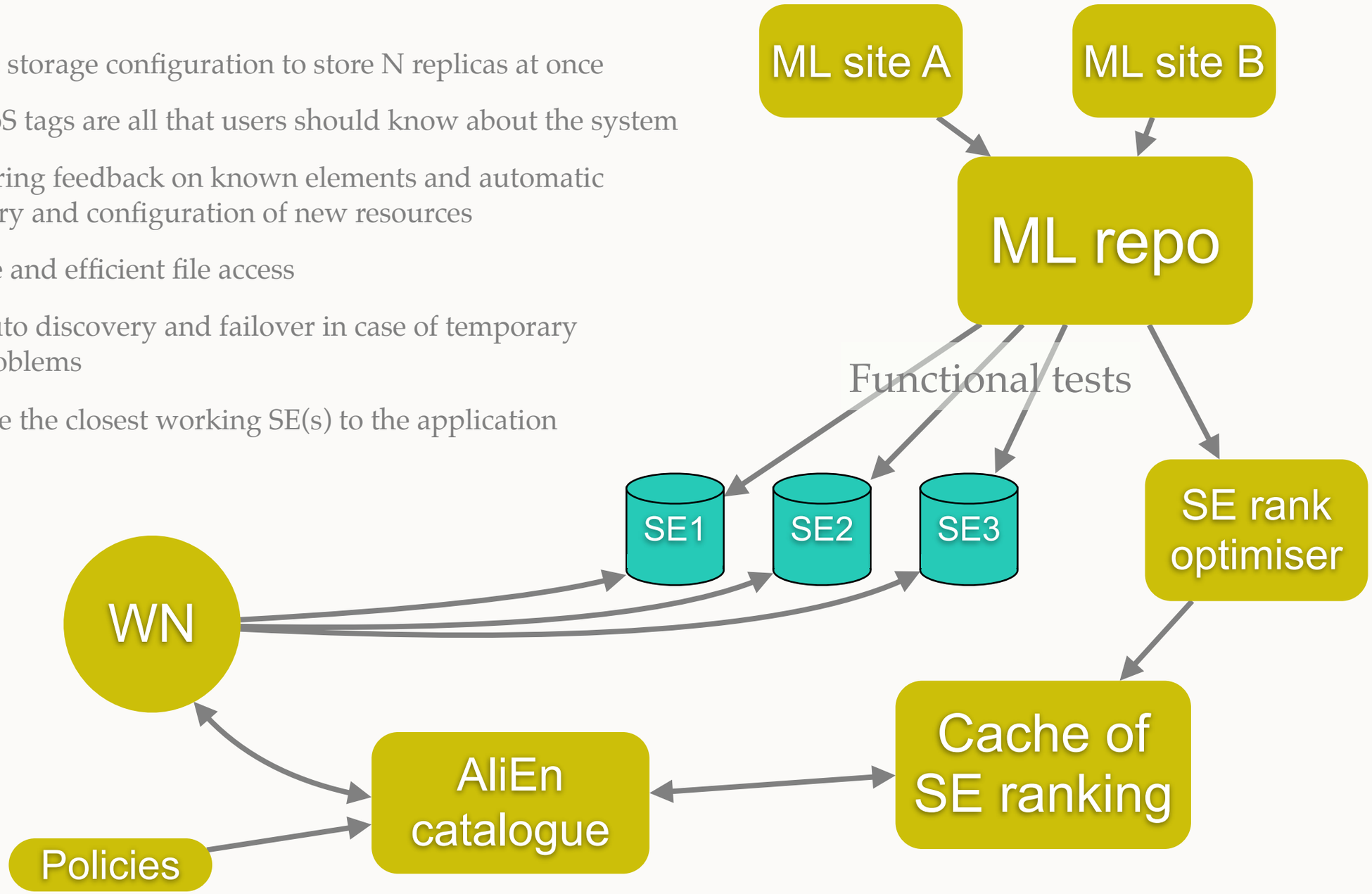
List of SEs

ALICE::**CCIN2P3**::SE
ALICE::**KOLKATA**::SE
ALICE::**CATANIA**::SE
ALICE::**BARI**::SE

Job executed at KOLKATA

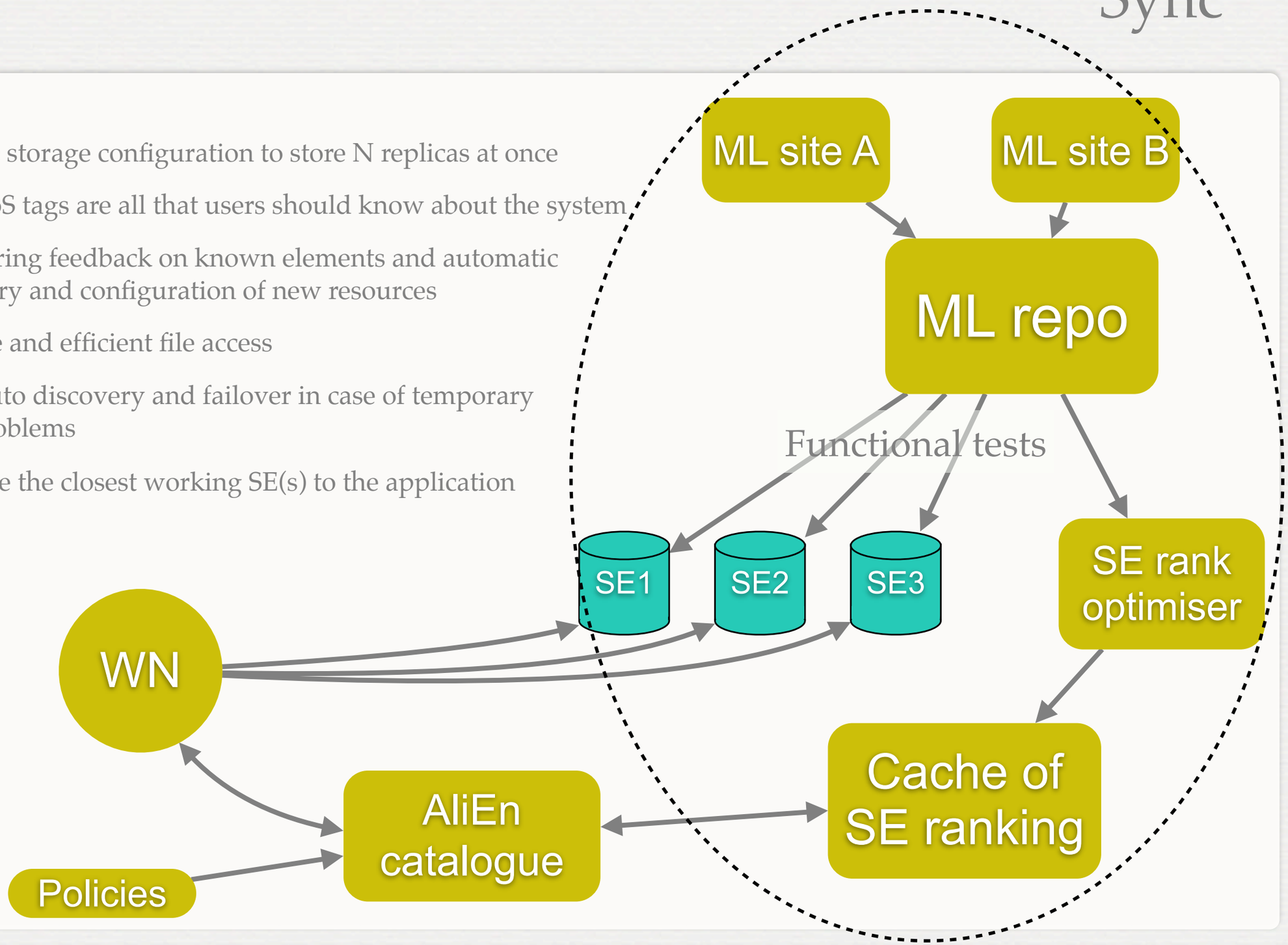
BOTTOM LINE

- Flexible storage configuration to store N replicas at once
  - QoS tags are all that users should know about the system
- Monitoring feedback on known elements and automatic discovery and configuration of new resources
- Reliable and efficient file access
  - Auto discovery and failover in case of temporary problems
  - Use the closest working SE(s) to the application



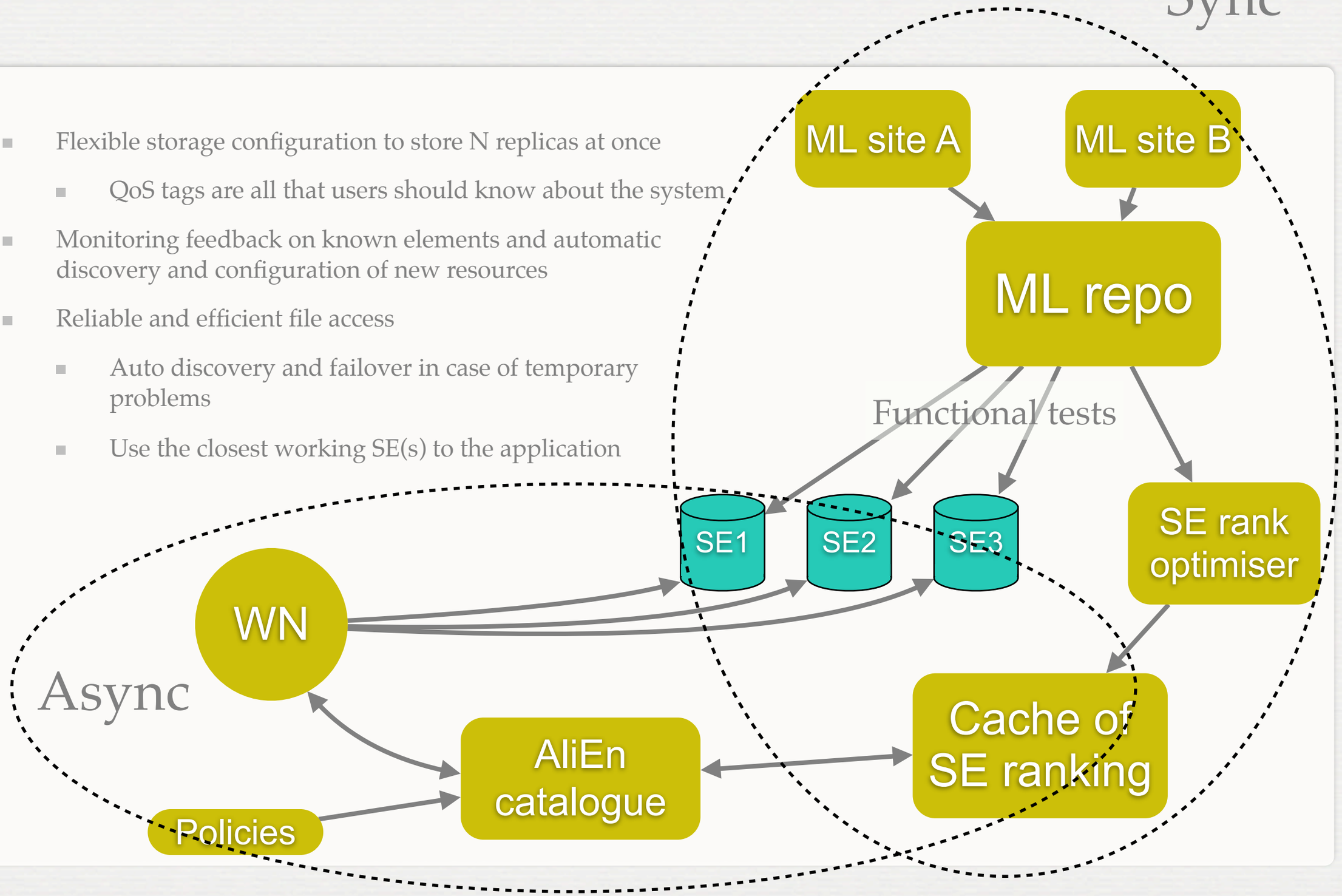
BOTTOM LINE

- Flexible storage configuration to store N replicas at once
  - QoS tags are all that users should know about the system
- Monitoring feedback on known elements and automatic discovery and configuration of new resources
- Reliable and efficient file access
  - Auto discovery and failover in case of temporary problems
  - Use the closest working SE(s) to the application



BOTTOM LINE

- Flexible storage configuration to store N replicas at once
  - QoS tags are all that users should know about the system
- Monitoring feedback on known elements and automatic discovery and configuration of new resources
- Reliable and efficient file access
  - Auto discovery and failover in case of temporary problems
  - Use the closest working SE(s) to the application

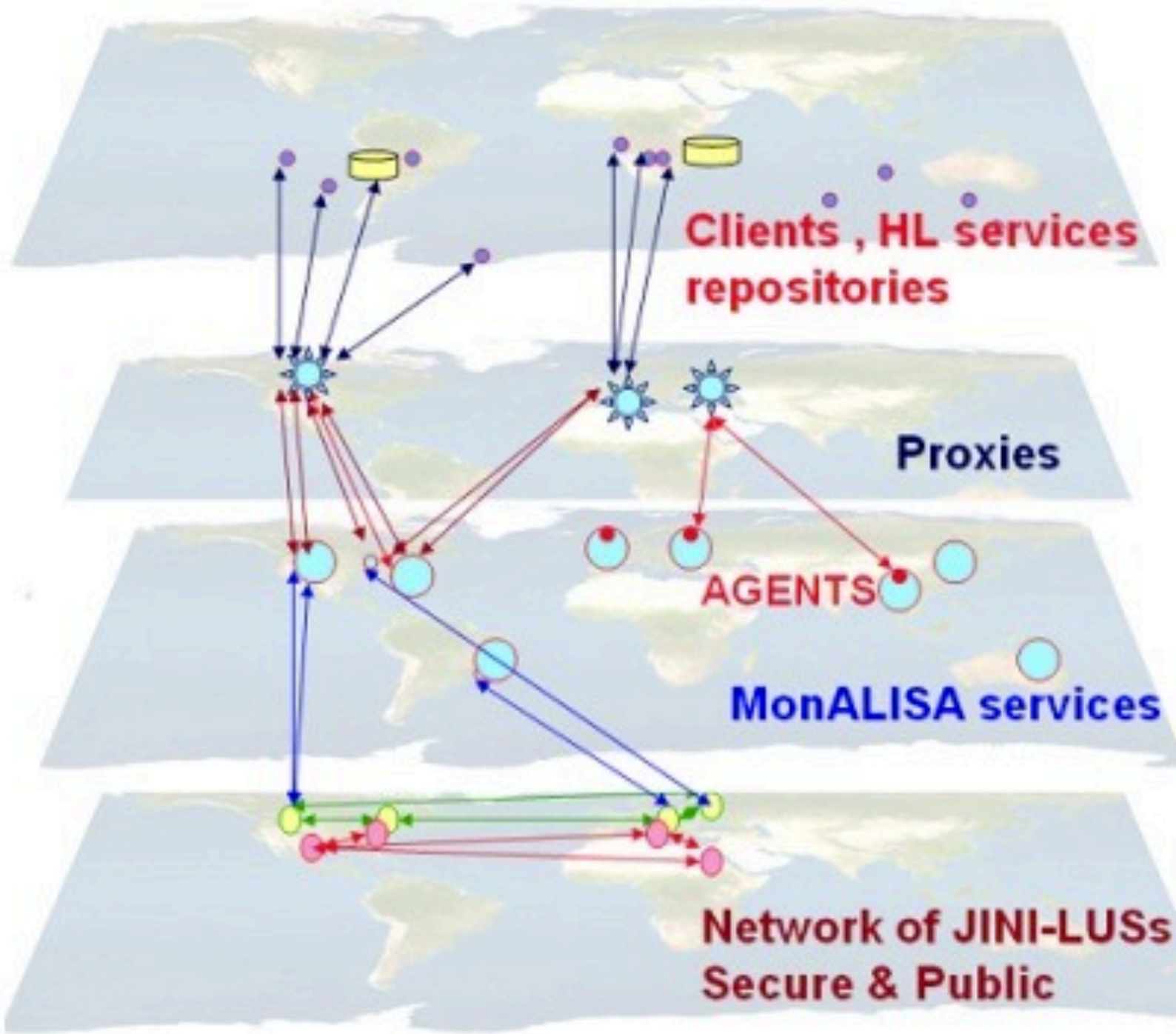


THE ALICE GRID

- AliEn working prototype in 2002
  - Single interface to distributed computing for all ALICE physicists
  - File catalogue, job submission and control, software management, user analysis
- ~80 participating sites now
  - 1 T0 (CERN/Switzerland)
  - 6 T1s (France, Germany, Italy, The Netherlands, Nordic DataGrid Facility, UK)
  - KISTI, UNAM and India coming (!)
  - ~73 T2s spread over 4 continents
- ~30,000 (out of ~150,000 WLCG) cores and 8.5 PB of disk
- Resources are “pooled” together
  - No localisation of roles / functions
  - National resources must integrate seamlessly into the global grid to be accounted for
  - FAs contribute proportionally to the number of PhDs (M&O-A share)
  - T3s have the same role than T2s, even if they do not sign the MoU

http://alien.cern.ch

ALL IS IN MONALISA



Global Services or Clients

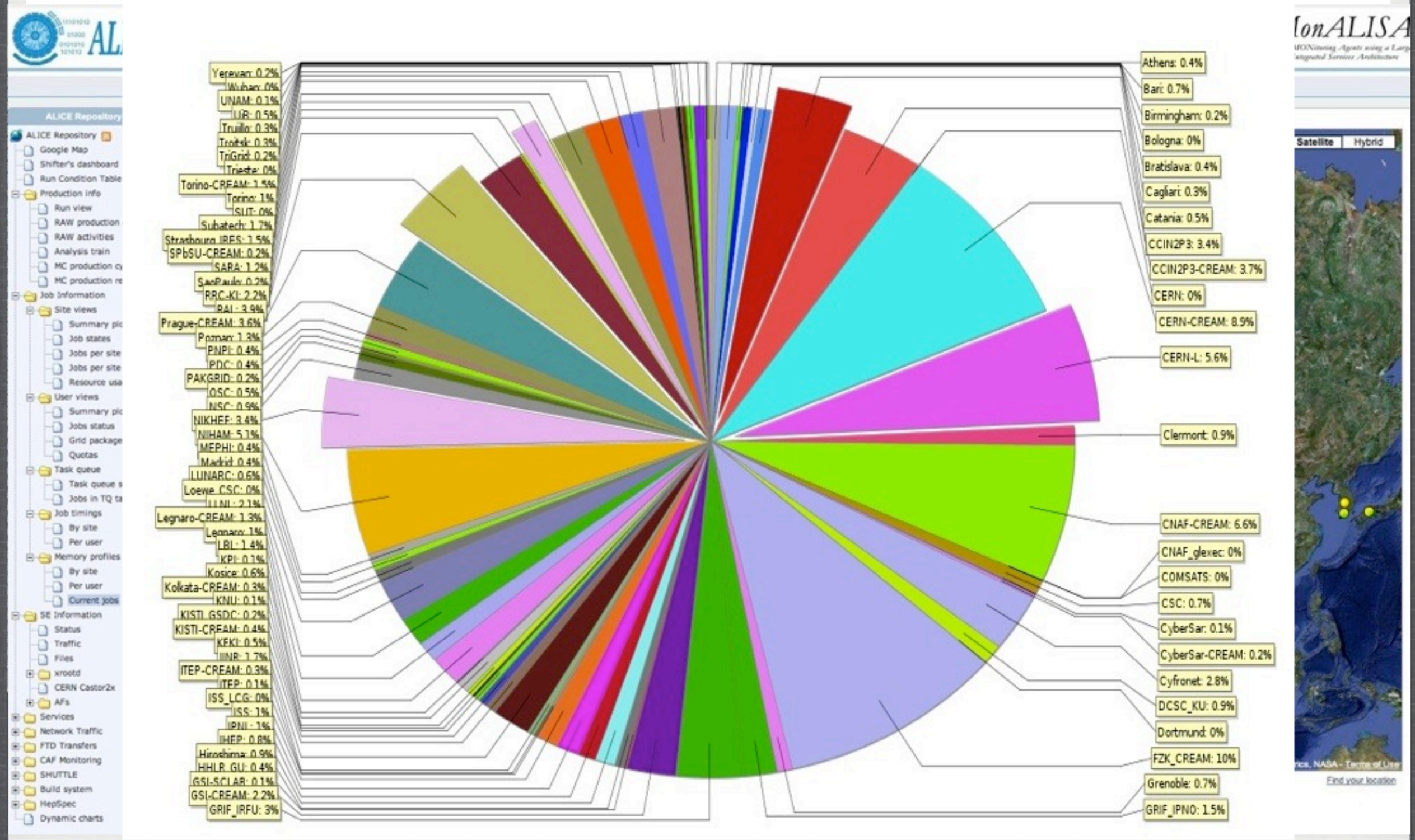
Dynamic load balancing
Scalability & Replication
Security AAA for Clients

Distributed System
for gathering and
Analyzing Information

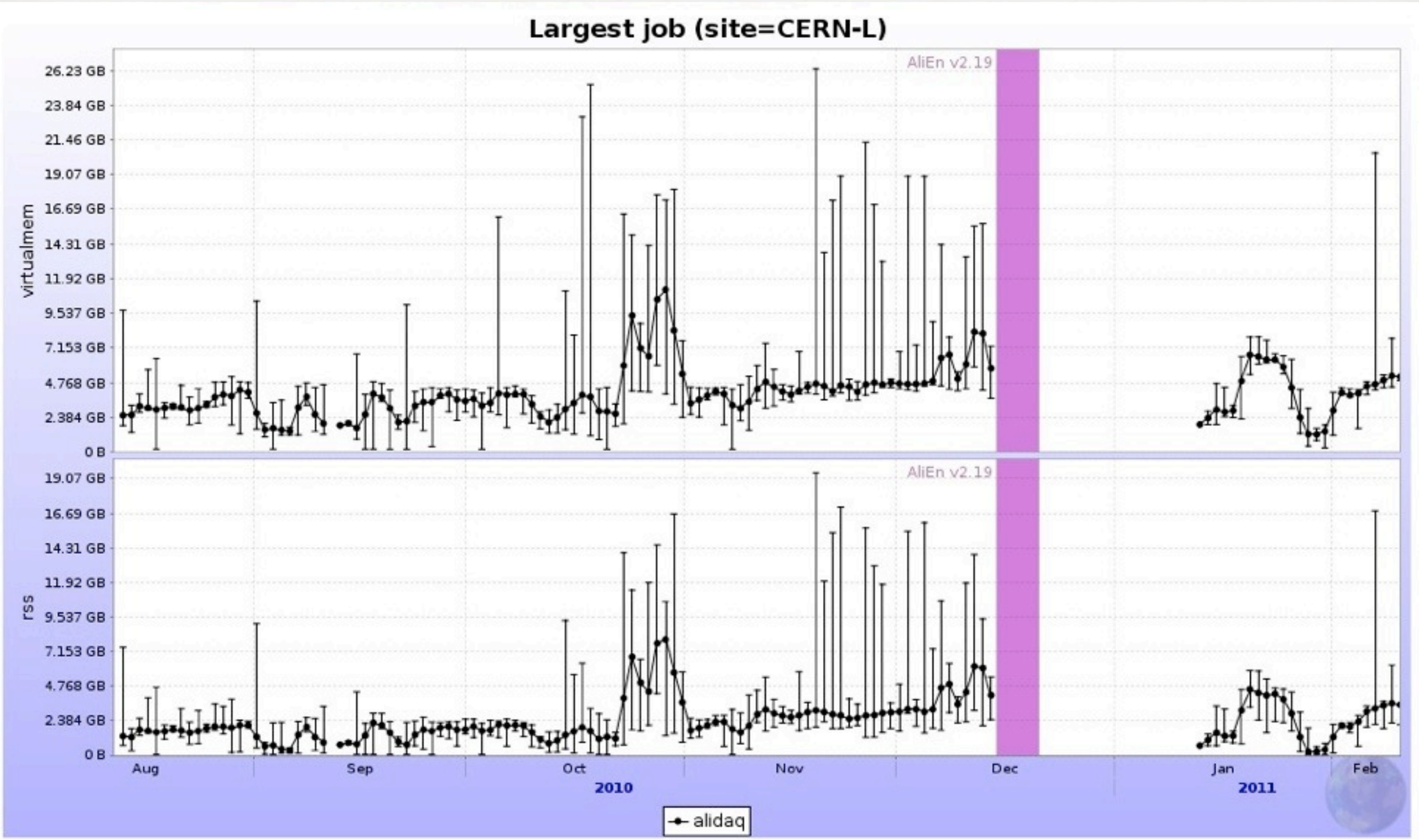
Distributed Dynamic
Discovery- based on
a lease Mechanism
and REN

ALL IS IN MONALISA

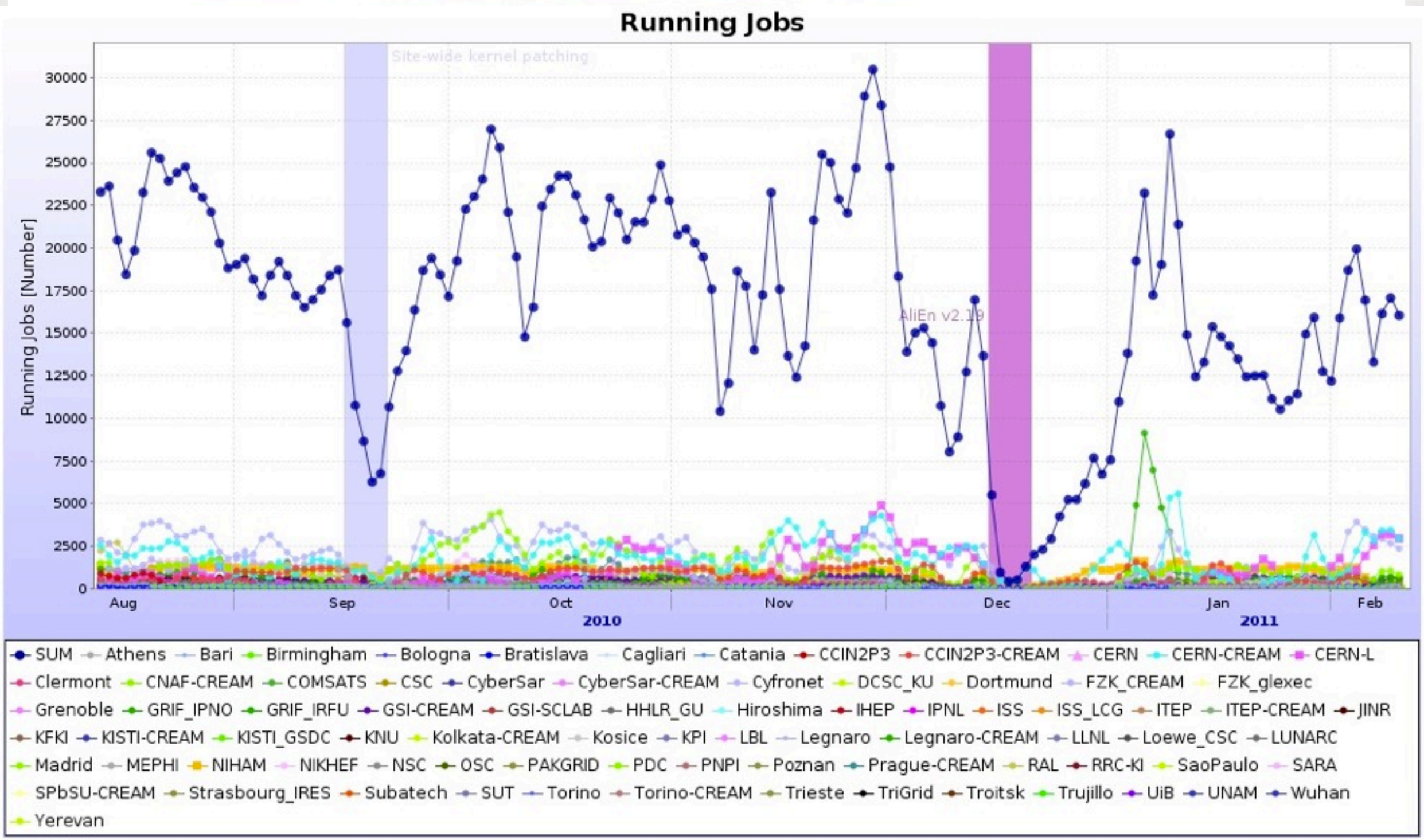
ALL IS IN MONALISA



ALL IS IN MONALISA



ALL IS IN MONALISA



ALL IS IN MONALISA

Production type: AOD

Production info							Jobs status						
		- Any -											
ID	Tag	Status	Done%	Cfg	Out	Links	Total	Done	Active	Waiting	Runs	Output events	Production description
929	FILTER_Pb-Pb_049_LHC10h	Running	91%				46020	41960	463	412	171 (136851 - 139517)	73,015,892	FILTER_Pb-Pb_049_LHC10h: AODs, TOF tender
934	FILTER_Pb-Pb_049_LHC10h_Stage1	Running	77%				3277	2534	2	739	81 (136854 - 139514)		
939	FILTER_Pb-Pb_049_LHC10h_Stage2	Running	56%				248	139		109	15 (138154 - 139513)		
933	FILTER_Pb-Pb_049_LHC10h_Stage3	Completed	100%				23	23			23 (136851 - 139504)	537,938	
923	FILTER_p-p_046_LHC11b2	Running	98%				1339	1315	2	12	36 (127719 - 130848)	1,091,879	FILTER_p-p_046_LHC11b2: stdAOD(+jets_new)/vertexing
925	FILTER_p-p_046_LHC11b2_Stage1	Running	50%				525	267	64	97	35 (127719 - 130848)		
926	FILTER_p-p_046_LHC11b2_Stage2	Running	54%				130	71	9	40	16 (127719 - 130847)		
922	FILTER_Pb-Pb_048_LHC11a10a	Completed	94%				285	269			1 (138653 - 138653)	27,750	FILTER_Pb-Pb_048_LHC11a10a: stdAOD(+jets_new)/vertexing
930	FILTER_Pb-Pb_048_LHC11a10a_Stage1	Running	1%				89	1	4	4	1 (138653 - 138653)		
920	FILTER_Pb-Pb_048_LHC11a10b	Running	96%				24297	23543	3	8	123 (137161 - 139517)	1,951,623	FILTER_Pb-Pb_048_LHC11a10b: stdAOD(+jets_new)/vertexing
924	FILTER_Pb-Pb_048_LHC11a10b_Stage1	Running	7%				5706	433	651	615	108 (137161 - 139514)		
927	FILTER_Pb-Pb_048_LHC11a10b_Stage2	Running	58%				62	36	3	11	10 (137165 - 139042)		
908	FILTER_p-p_047_LHC11a	Completed	95%				12156	11582			216 (141795 - 146860)	396,950,151	FILTER_p-p_047_LHC11a: No tender
913	FILTER_p-p_047_LHC11a_Stage1	Completed	99%				2318	2307			107 (-1 - 146860)		
918	FILTER_p-p_047_LHC11a_Stage2	Completed	99%				472	471			49 (145674 - 146858)		
915	FILTER_p-p_047_LHC11a_Stage3	Completed	100%				210	210			210 (141805 - 146860)	385,761,764	
904	FILTER_p-p_046_LHC10e20	Completed	98%				128	126			2 (130847 - 130848)	302,400	FILTER_p-p_046_LHC10e20: stdAOD(+jets_new)/vertexing
911	FILTER_p-p_046_LHC10e20_Stage1	Completed	80%				76	61			2 (130847 - 130848)		
928	FILTER_p-p_046_LHC10e20_Stage2	Completed	100%				14	14			1 (130847 - 130847)		
897	FILTER_p-p_045_LHC11a	Technical stop	30%				2026	628			23 (146686 - 146860)	27,701,871	FILTER_p-p_045_LHC11a: Vertex tender
907	FILTER_p-p_045_LHC11a_Stage1	Technical stop	73%				19	14			1 (146859 - 146859)		
912	FILTER_p-p_045_LHC11a_Stage3	Completed	100%				3	3			3 (146686 - 146859)	2,780,783	

Yerevan

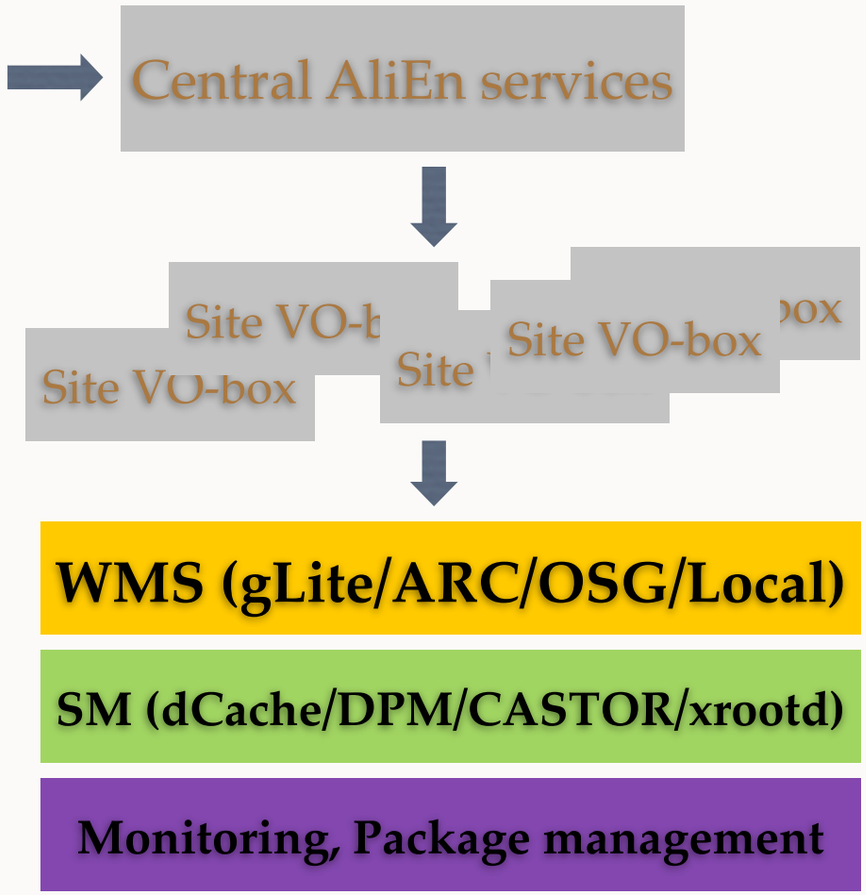
ALL IS IN MONALISA

Production type: AOD

Disk storage elements

	AliEn SE	Statistics						Xrootd info					Functional tests			
SE Name	AliEn name	Size	Used	Free	Usage	No. of files	Type	Size	Used	Free	Usage	Version	add	ls	get	whereis
1. Bari - SE	ALICE::Bari::SE	893.4 TB	114.8 TB	778.6 TB	12.85%	2,573,640	File	1.721 PB	1.539 PB	186.4 TB	89.42%	20100510-1509_dbg				
2. Bratislava - SE	ALICE::Bratislava::SE	112.8 TB	32.86 TB	79.94 TB	29.13%	969,569	File	112.8 TB	48.46 TB	64.31 TB	42.97%	20100510-1509_dbg				
3. Catania - SE	ALICE::Catania::SE	100.4 TB	118.1 TB	-	117.6%	2,314,380	File	158.7 TB	152.4 TB	6.291 TB	96.03%	20100510-1509_dbg				
4. CCIN2P3 - SE	ALICE::CCIN2P3::SE	96 TB	116.7 TB	-	121.5%	2,429,195	File	-	-	-	-					
5. CERN - ALICEDISK	ALICE::CERN::ALICEDISK	849.6 TB	850.3 TB	-	100.1%	13,314,358	CASTOR	-	-	-	-					
6. CERN - GLOBAL	ALICE::CERN::GLOBAL	-	0	1.863 TB	-	4,761	root	-	-	-	-					
7. CERN - SE	ALICE::CERN::SE	20.49 TB	13.94 TB	6.545 TB	68.05%	3,546,153	File	20.46 TB	7.047 TB	13.41 TB	34.44%	20100510-1509_dbg				
8. 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...
9. CNAF - SE	ALICE::CNAF::SE	873.3 TB	461.6 TB	411.7 TB	52.86%	7,585,473	File	873.3 TB	509.5 TB	363.7 TB	58.35%	20100510-1509_dbg				
10. CyberSar_Cagliari - SE	ALICE::CyberSar_Cagliari::SE	30.83 TB	33.44 TB	-	108.5%	869,370	File	92.71 TB	90.69 TB	2.02 TB	97.82%	20100510-1509_dbg				
11. Cyfronet - SE	ALICE::Cyfronet::SE	10 TB	11.69 TB	-	116.9%	518,759	File	9.995 TB	9.547 TB	458.4 GB	95.52%	20100510-1509_dbg				
12. FZK - SE	ALICE::FZK::SE	1.254 PB	614.3 TB	669.7 TB	47.84%	9,176,837	File	1.284 PB	699.9 TB	614.7 TB	53.24%	20100510-1509_dbg				
13. Grenoble - DPM	ALICE::Grenoble::DPM	72 TB	6.083 TB	65.92 TB	8.449%	200,049	SRM	-	-	-	-					
14. Grenoble - SE	ALICE::Grenoble::SE	31 TB	20.09 TB	10.91 TB	64.8%	401,894	File	-	-	-	-		Use ...	Last...	Last...	Last...
15. GRIF_IPNO - DPM	ALICE::GRIF_IPNO::DPM	85.24 TB	81.35 TB	3.89 TB	95.44%	2,240,416	SRM	-	-	-	-		Use ...			
16. 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...
17. GRIF_IRFU - DPM	ALICE::GRIF_IRFU::DPM	171 TB	42.31 TB	128.7 TB	24.74%	782,168	SRM	-	-	-	-					
18. GSI - SE	ALICE::GSI::SE	312.6 TB	330.4 TB	-	105.7%	6,114,838	File	291.7 TB	272.3 TB	19.35 TB	93.37%	20100510-1509_dbg	Use ...	Last...	Last...	Last...
19. GSI - SE2	ALICE::GSI::SE2	28 TB	457.3 GB	27.55 TB	1.595%	3,409	File	0	0	0	-	20100510-1509_dbg	Use ...	Last...	Last...	Last...
20. HHLR_GU - SE	ALICE::HHLR_GU::SE	200 TB	367.5 GB	199.6 TB	0.179%	5,724	File	-	-	-	-		Use ...	Last...	Last...	Last...
21. Hiroshima - SE	ALICE::Hiroshima::SE	79 TB	38.63 TB	40.37 TB	48.89%	941,095	File	78.78 TB	48.76 TB	30.02 TB	61.89%	20100510-1509_dbg				
22. IHEP - SE	ALICE::IHEP::SE	35.55 TB	8.916 TB	26.63 TB	25.08%	569,733	File	36.38 TB	9.274 TB	27.11 TB	25.49%	20100510-1509_dbg				
23. IPNL - SE	ALICE::IPNL::SE	36 TB	50.53 TB	-	140.4%	1,120,881	File	37.3 TB	36.4 TB	916.4 GB	97.6%	20100510-1509_dbg				
24. ISS - FILE	ALICE::ISS::FILE	140.5 TB	104.5 TB	35.99 TB	74.38%	3,049,353	File	140.5 TB	129.1 TB	11.37 TB	91.91%	20100510-1509_dbg				
25. ITEP - SE	ALICE::ITEP::SE	100 TB	41.37 TB	58.63 TB	41.37%	1,097,375	File	99.93 TB	44.74 TB	55.19 TB	44.78%	20100510-1509_dbg				
26. JINR - SE	ALICE::JINR::SE	112.3 TB	75.95 TB	36.35 TB	67.63%	3,144,338	File	149.1 TB	80.47 TB	68.62 TB	53.97%	20100510-1509_dbg				
27. KFKI - SE	ALICE::KFKI::SE	39.34 TB	26.68 TB	12.66 TB	67.83%	731,616	File	36.38 TB	34.73 TB	1.652 TB	95.46%	20100510-1509_dbg				
28. KISTI_GSDC - SE	ALICE::KISTI_GSDC::SE	100 TB	29.55 TB	70.45 TB	29.55%	619,638	File	101.8 TB	43.9 TB	57.88 TB	43.13%	20100510-1509_dbg				
29. KISTI - SE	ALICE::KISTI::SE	49.95 TB	32.18 TB	17.77 TB	64.43%	787,787	File	49.95 TB	30.81 TB	19.14 TB	61.68%	20100510-1509_dbg				
30. Kolkata - SE	ALICE::Kolkata::SE	73.24 TB	14.91 TB	58.33 TB	20.35%	454,585	File	70.46 TB	31.77 TB	38.69 TB	45.09%	20100510-1509_dbg				
31. Kosice - SE	ALICE::Kosice::SE	41.84 TB	29.37 TB	12.47 TB	70.21%	736,513	File	61.84 TB	39.03 TB	22.81 TB	63.11%	20100115.1117_dbg				
32. LBL - SE	ALICE::LBL::SE	143.2 TB	43.56 TB	99.64 TB	30.42%	1,053,500	File	214.8 TB	60.02 TB	154.8 TB	27.94%	20100510-1509_dbg				
33. Legnaro - SE	ALICE::Legnaro::SE	138.3 TB	85.71 TB	52.59 TB	61.97%	2,415,156	File	138.3 TB	103.7 TB	34.58 TB	74.99%	20100510-1509_dbg				
34. LLNL - SE	ALICE::LLNL::SE	688 TB	22.44 TB	665.6 TB	3.262%	793,608	File	687.8 TB	79.85 TB	607.9 TB	11.61%	20100510-1509_dbg				
35. Madrid - SE	ALICE::Madrid::SE	37.5 TB	12.32 TB	25.18 TB	32.85%	403,412	File	36.6 TB	15.47 TB	21.13 TB	42.27%	20100510-1509_dbg				
36. MEPHI - SE	ALICE::MEPHI::SE	18.2 TB	7.981 TB	10.22 TB	43.85%	254,085	File	18.19 TB	12.73 TB	5.457 TB	70%	20100510-1509_dbg				
37. NDGF - DCACHE	ALICE::NDGF::DCACHE	204.6 TB	171.3 TB	33.33 TB	83.71%	2,689,674	srm	-	-	-	-					
38. NIHAM - FILE	ALICE::NIHAM::FILE	855 TB	220.3 TB	634.7 TB	25.77%	11,499,646	File	854.9 TB	221.3 TB	633.6 TB	25.89%	v3.0.2				

GRID OPERATION PRINCIPLE



- The VO-box system (very controversial in the beginning)
  - Has been extensively tested
  - Allows for site services scaling
  - Is a simple isolation layer for the VO in case of troubles

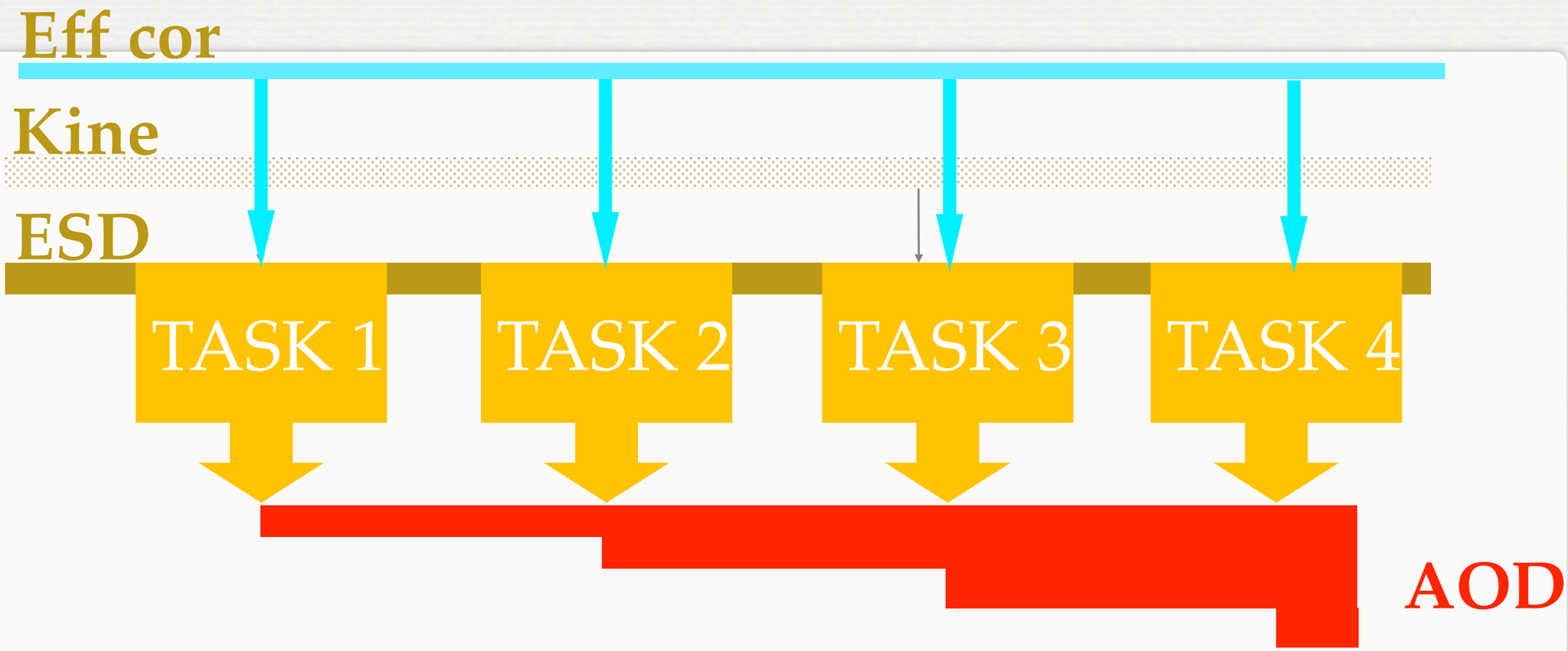
OPERATION – CENTRAL/ SITE SUPPORT

- Central services support (2 FTEs equivalent)
  - There are no experts which do exclusively support – there are 6 highly-qualified experts doing development/support
- Site services support - handled by ‘regional experts’ (one per country) in collaboration with local cluster administrators
  - Extremely important part of the system
  - In normal operation ~0.2FTEs/site
- Regular weekly discussions and active all-activities mailing lists

ANALYSIS

- Much more successful than anticipated
  - At least by ALICE
  - We can really do analysis on the Grid
- In some sense analysis is victim of its own success
  - In ALICE users are “abusing” the “par file” system
    - Local compilation of code fragments
  - The access to the calibration database from analysis jobs is overloading the AliEn catalogue
  - In ATLAS the Data Distribution Model is running way above the design values
- Multiplication of data formats and reduction in the file size is a common curse

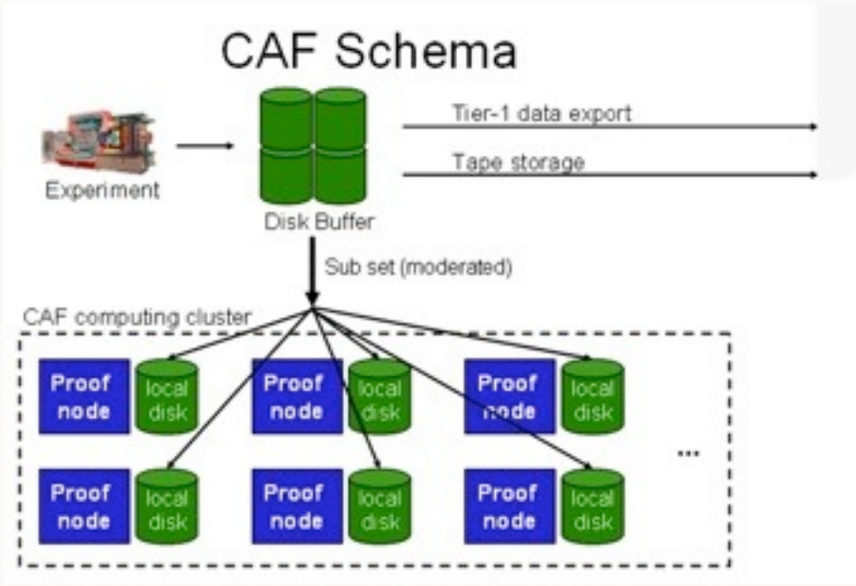
ANALYSIS TRAIN

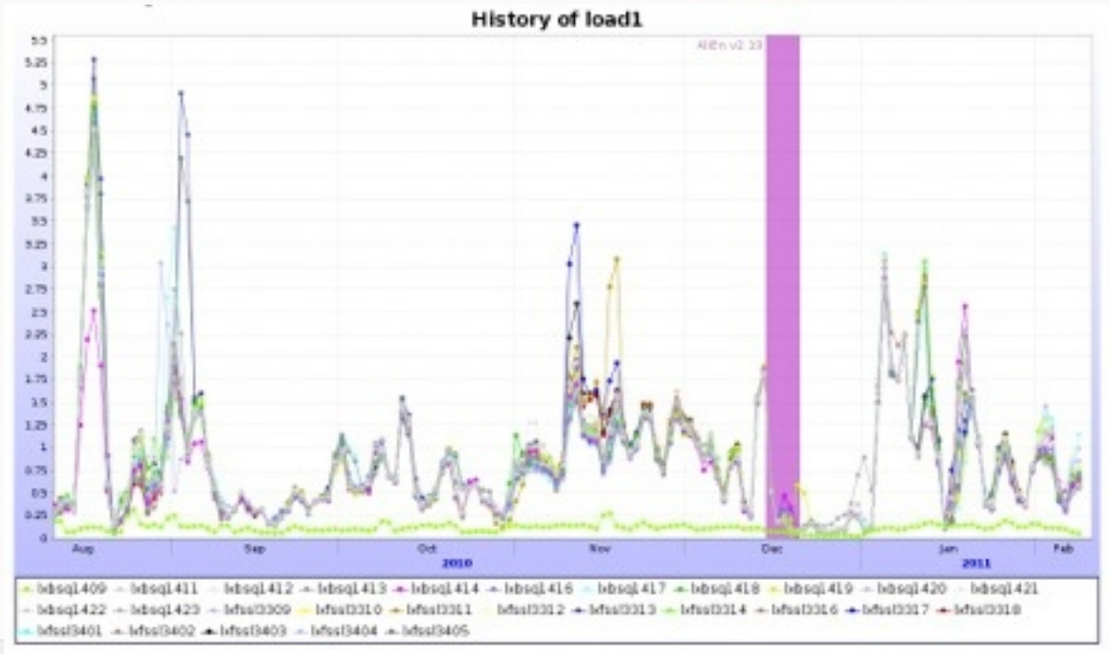


- AOD production will be organized in a ‘train’ of tasks
  - To maximize efficiency of full dataset processing
  - To optimize CPU/IO
  - Using the analysis framework

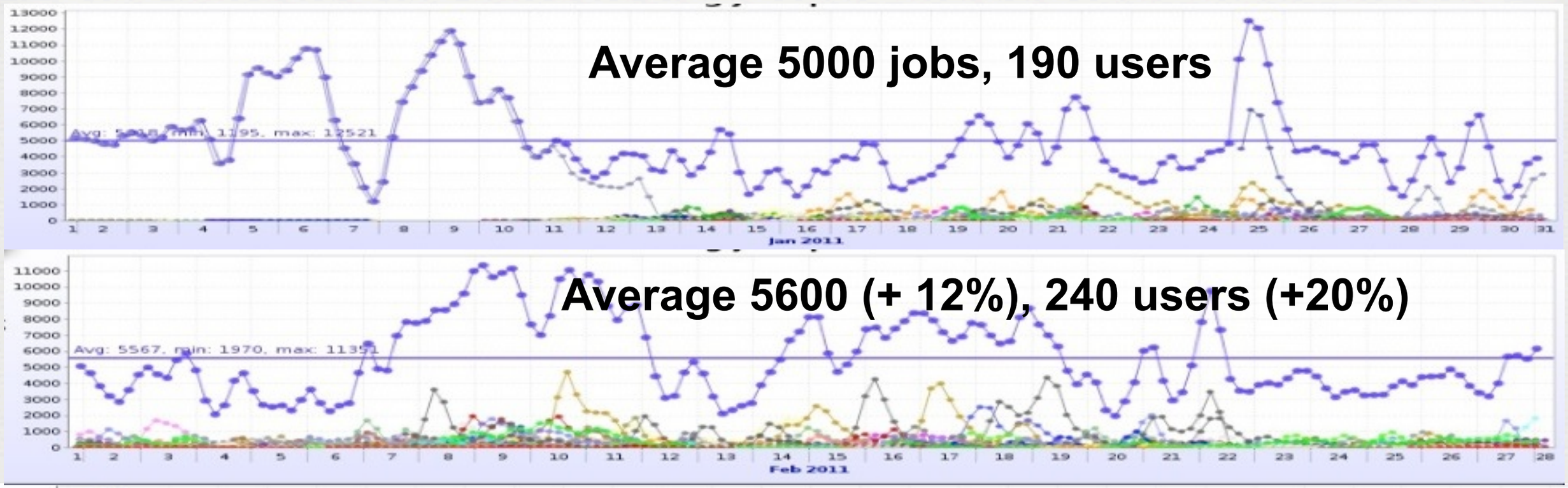
THE ALICE ANALYSIS FACILITIES

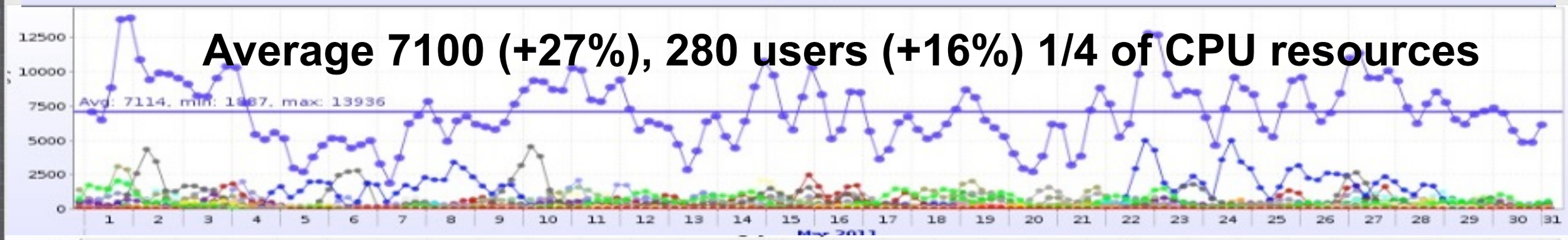
- Proof-enabled, Grid-aware parallel computing platform
- Used for early discovery physics, calibration
- “Victim of its own success” has doubled twice in the last year at CERN, 480 cores in few days

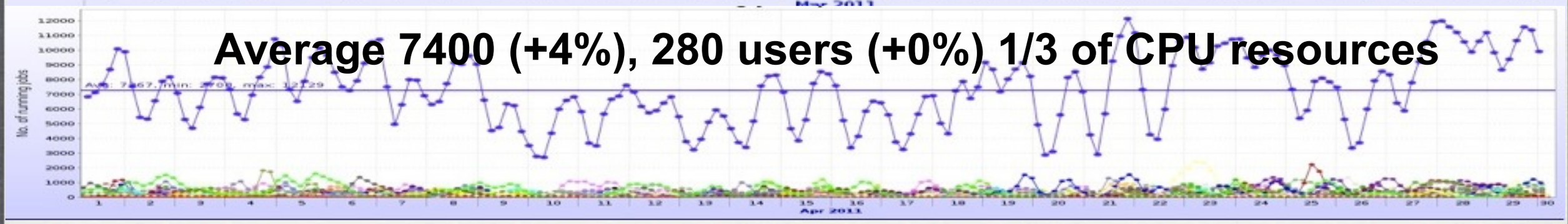




USER ACTIVITY – MONTH ON MONTH INCREASE







ALIROOT

- AliRoot started officially in 1999
- There was never a “reset” of the code, but constant evolution
  - With very heavy refactoring
- One tag release per week
- One full release every ~6 month
- A daunting task

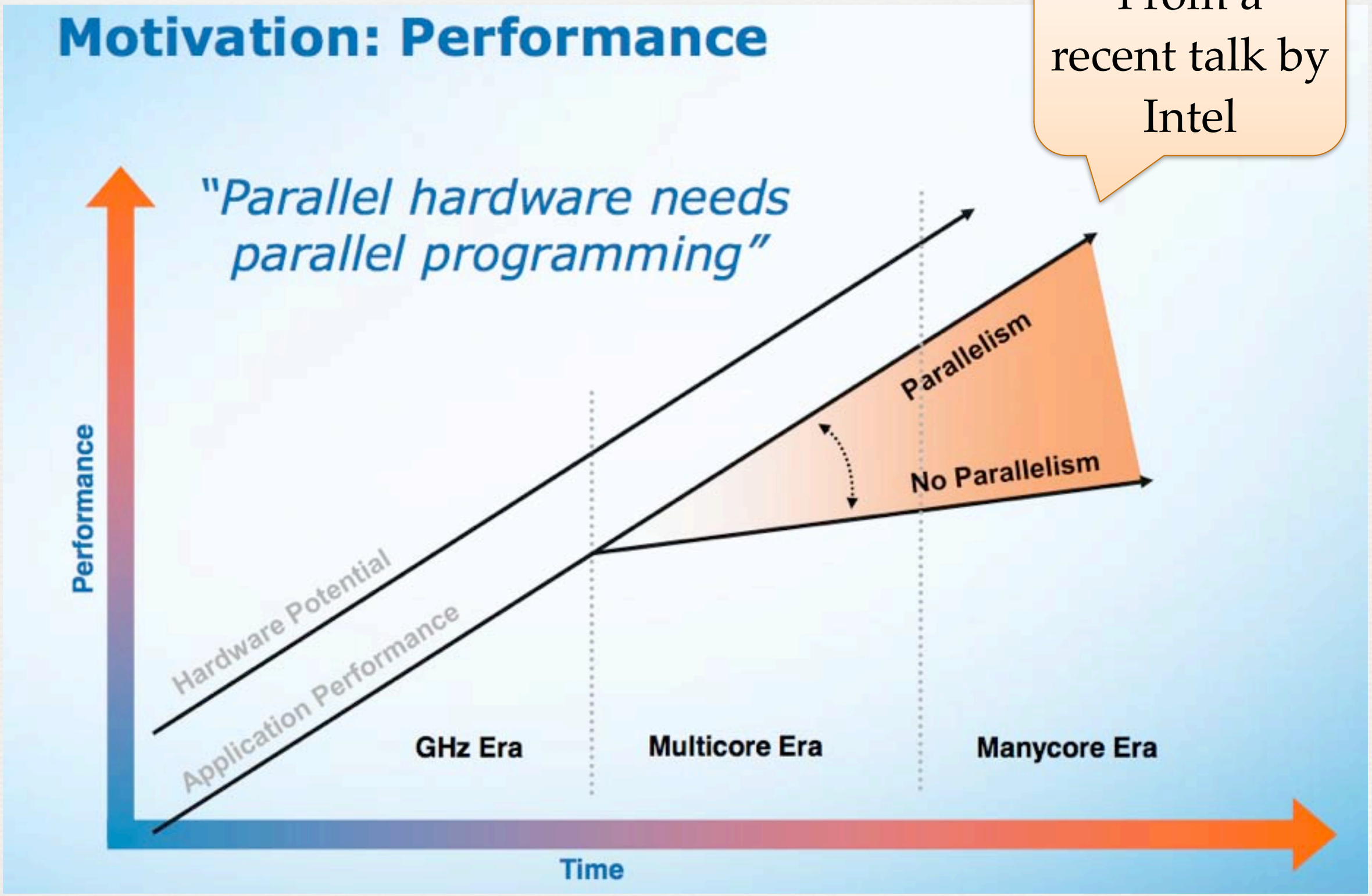
BUT WHAT'S NEXT?

ALIROOT OPTIMISATION

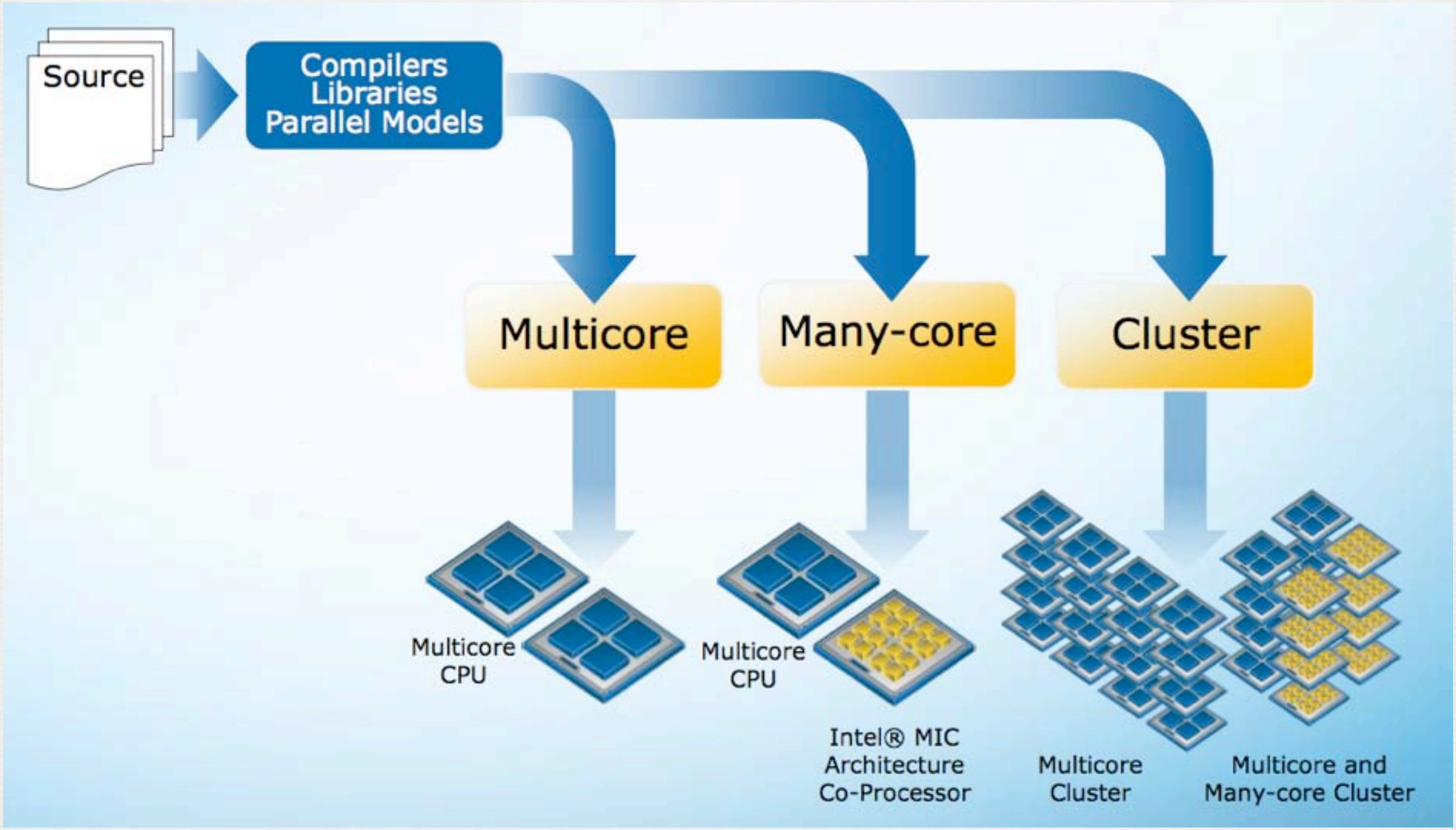
- The HEP code
  - An embarrassing parallelism
  - An inextricable mix of branches / integer / float / double
  - A “flat” timing distribution – no “hot spots”
- We always got away with clock rate, now it is not possible any more
  - Parallelism is there to stay
- We cannot claim that we are resource-hungry and then exploit ~10%-50% of the hardware
  - Just think what it means in terms of money

PARALLELISM

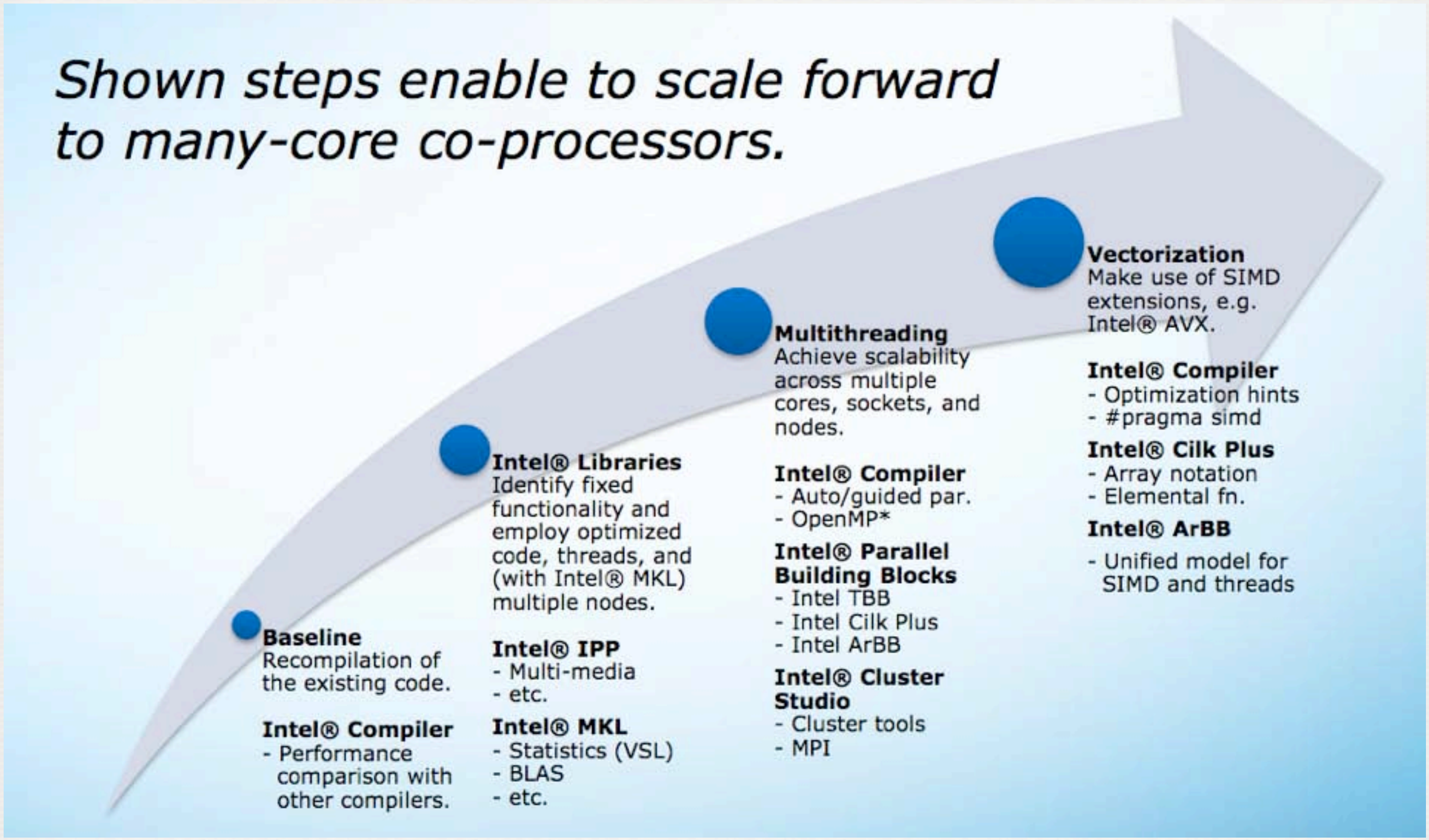
From a recent talk by Intel



IF YOU TRUST INTEL



IF YOU TRUST INTEL 2



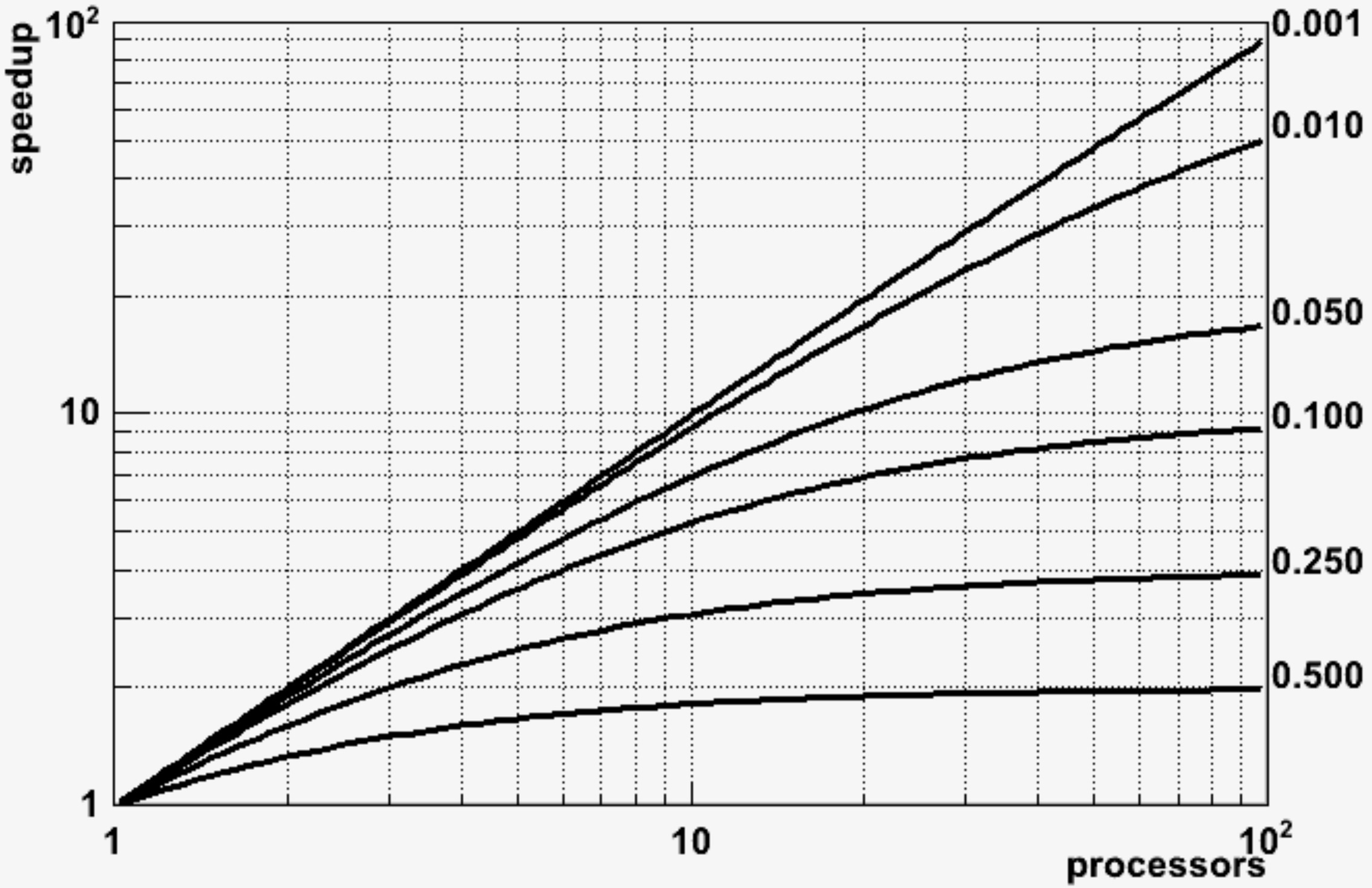
WHY IT IS SO DIFFICULT?

- No clear kernel
- C++ code generation / optimisation not well understood
- Most of the technology is coming out now
  - Lack of standards
  - Technological risk
- Non professional coders
- Fast evolving code
- No control on hardware acquisition

WHY IT IS SO DIFFICULT (CONT)?

- Amdhal law sets stringent limits to the results that can be achieved
  - No “low level” optimisation alone will yield results
- Heterogeneous parallelism forces multi-level parallelisation
- Essentially the code (all of it!) will have to be re-written

Amdahl law

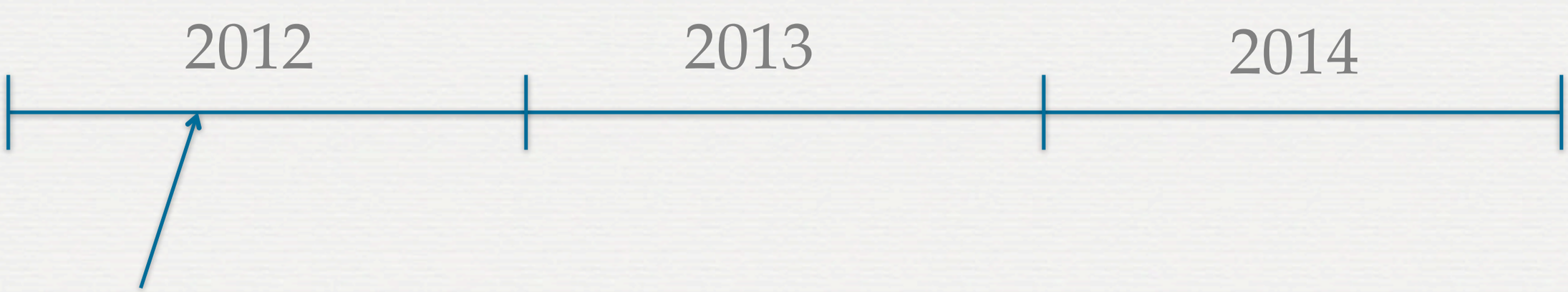

speedup
10^2
10
1
1
10
10^2
processors
0.001
0.010
0.050
0.100
0.250
0.500

ALICE STRATEGY (UNAUTHORISED)

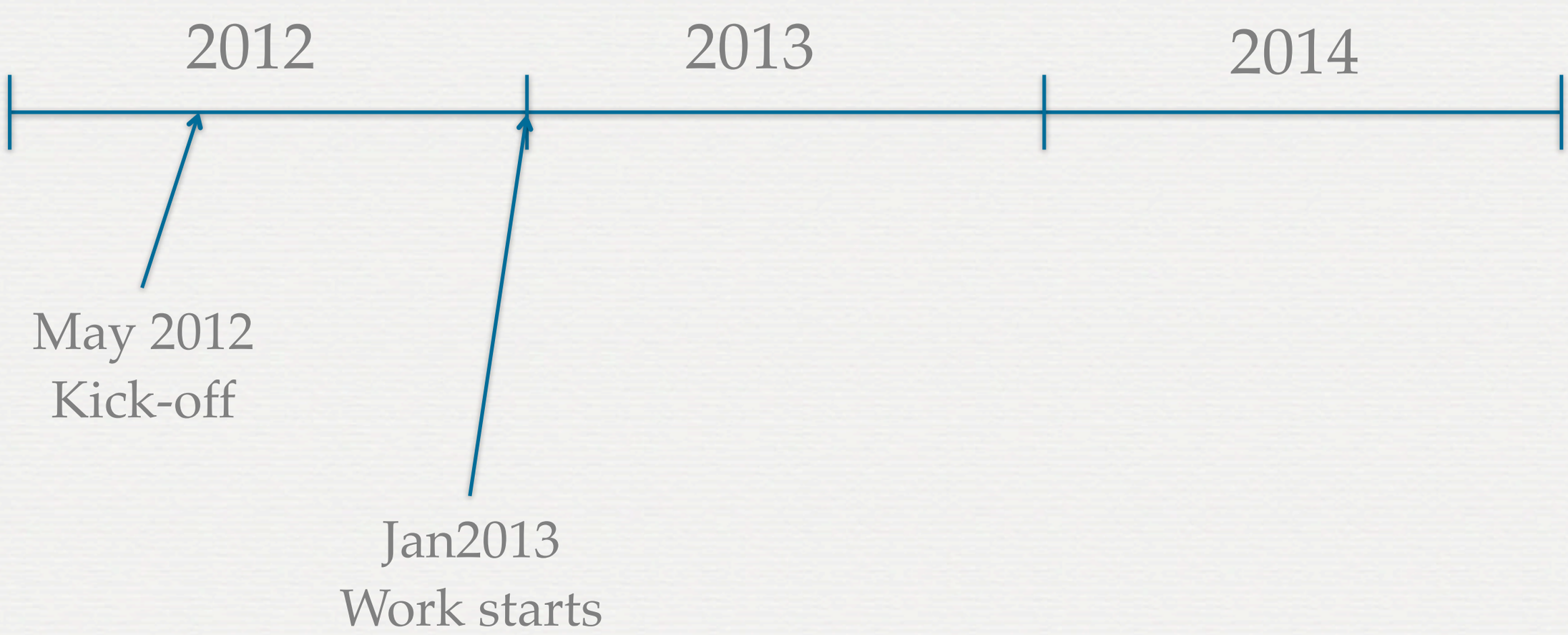
- Use the LSD-1 to essentially re-write AliRoot
- Use the LSD-2 to expand the parallelism to the Grid
  - Hopefully the major thrust will be on MiddleWare
- Refactor the code in order to expose the maximum of parallelism present at each level
- Keep the code in C++ (no CUDA, OpenCL etc.)
- Explore the possible use of #pragma's (OpenMP, OpenACC)
- Experiment on all hardware at hand (OpenLab, but not only)

TIMELINE

TIMELINE



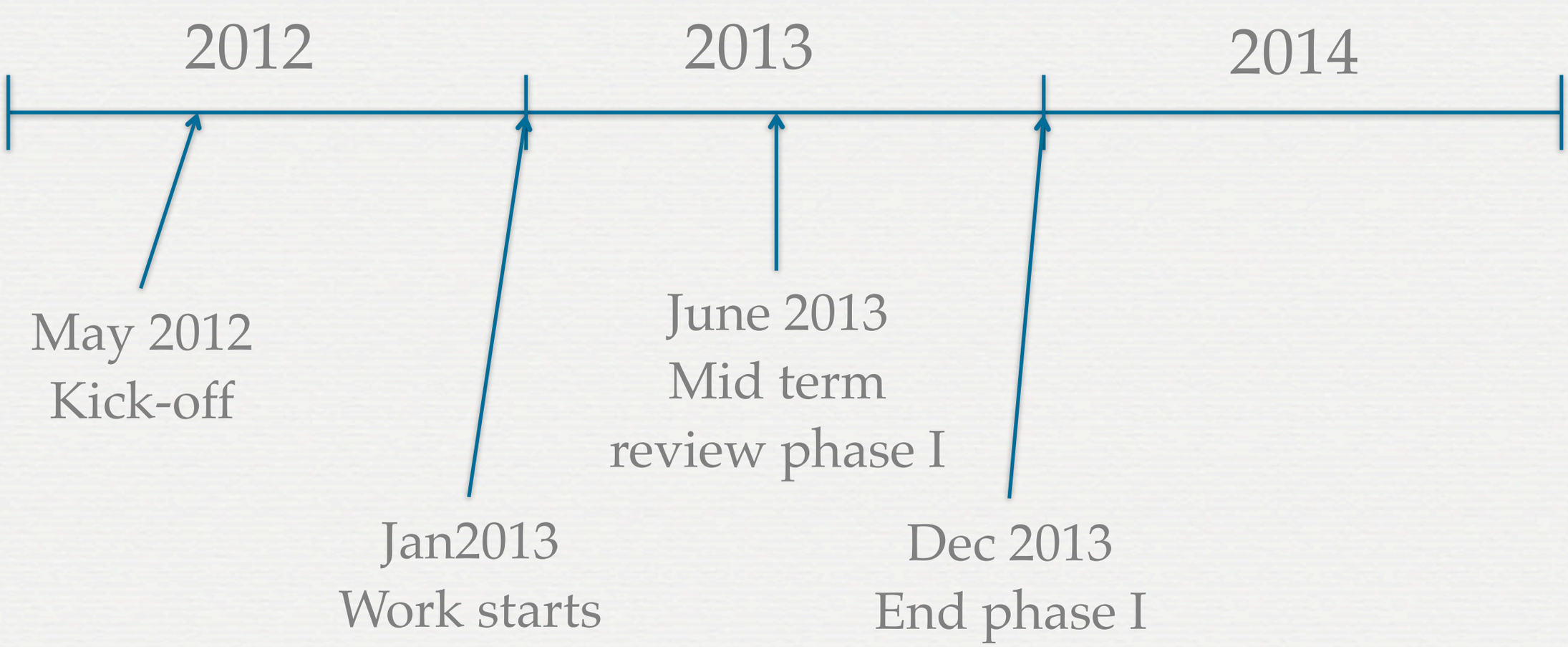
TIMELINE



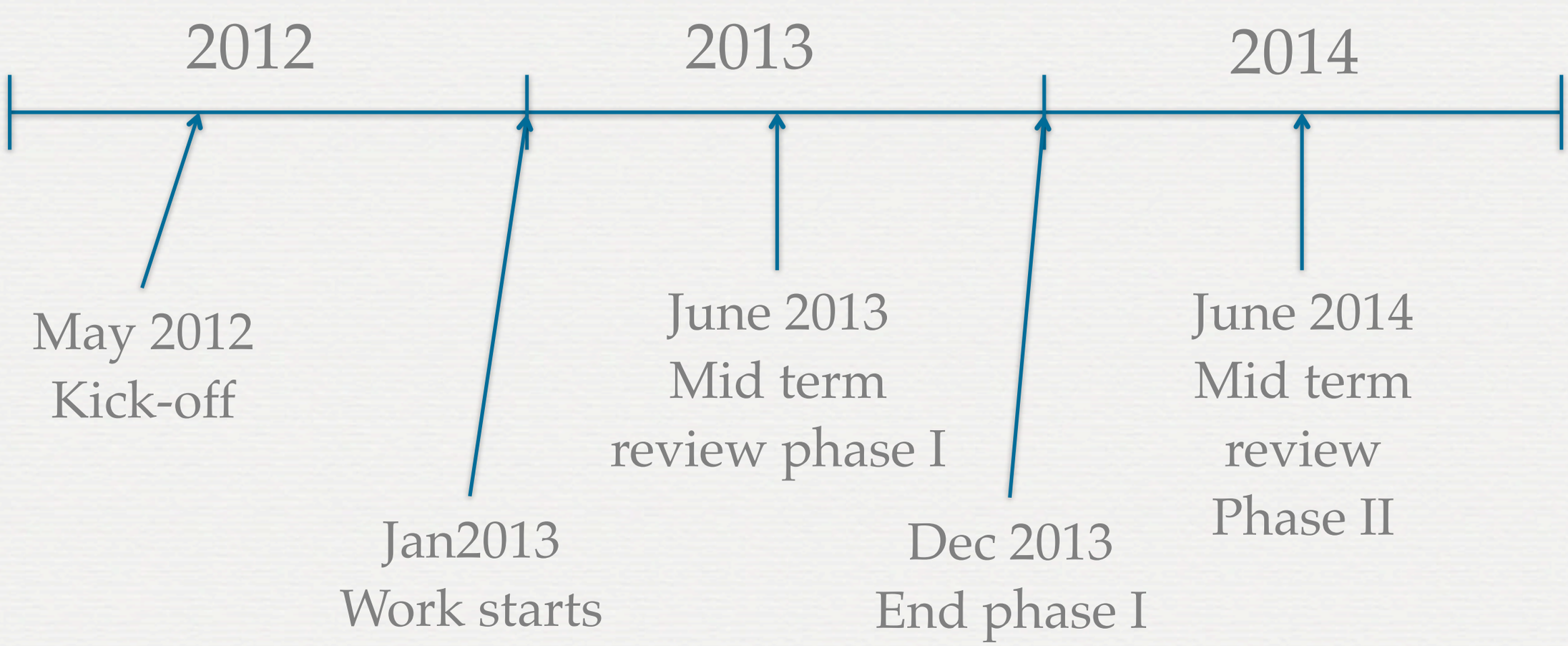
TIMELINE


2012
2013
2014
May 2012
Kick-off
Jan2013
Work starts
June 2013
Mid term
review phase I

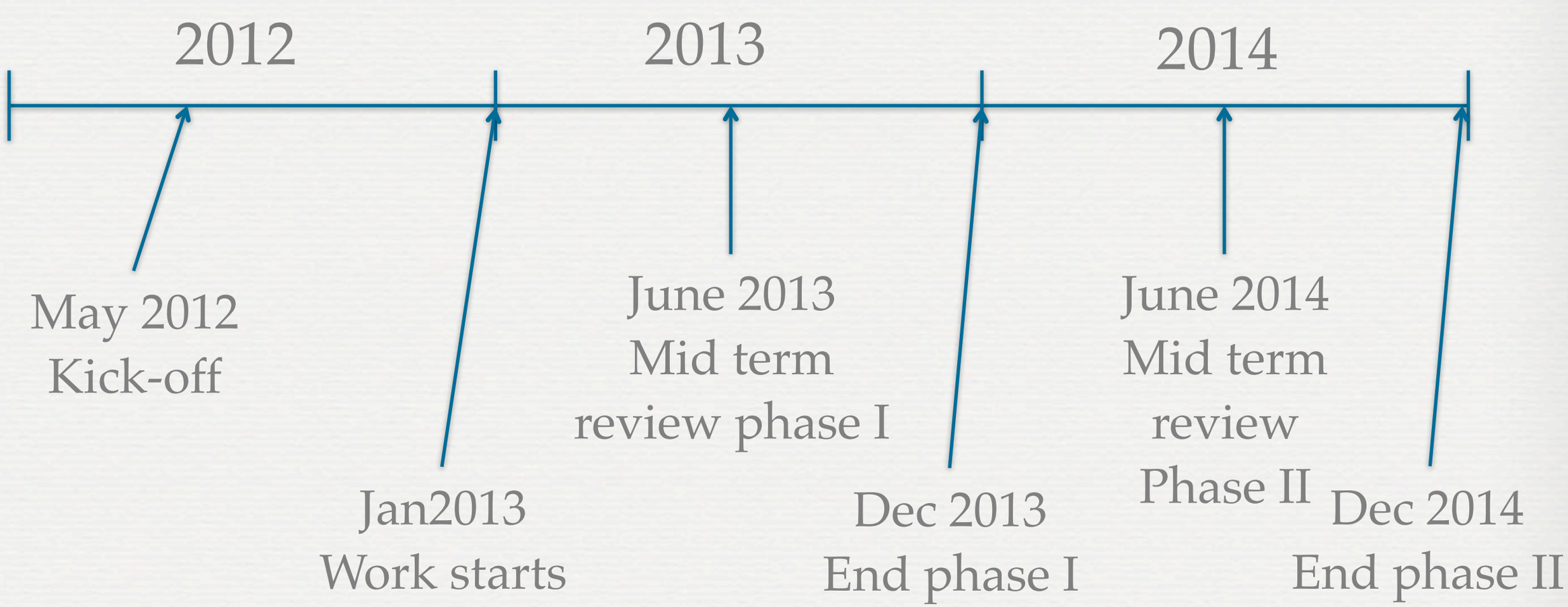
TIMELINE


2012
2013
2014
May 2012
Kick-off
Jan2013
Work starts
June 2013
Mid term
review phase I
Dec 2013
End phase I

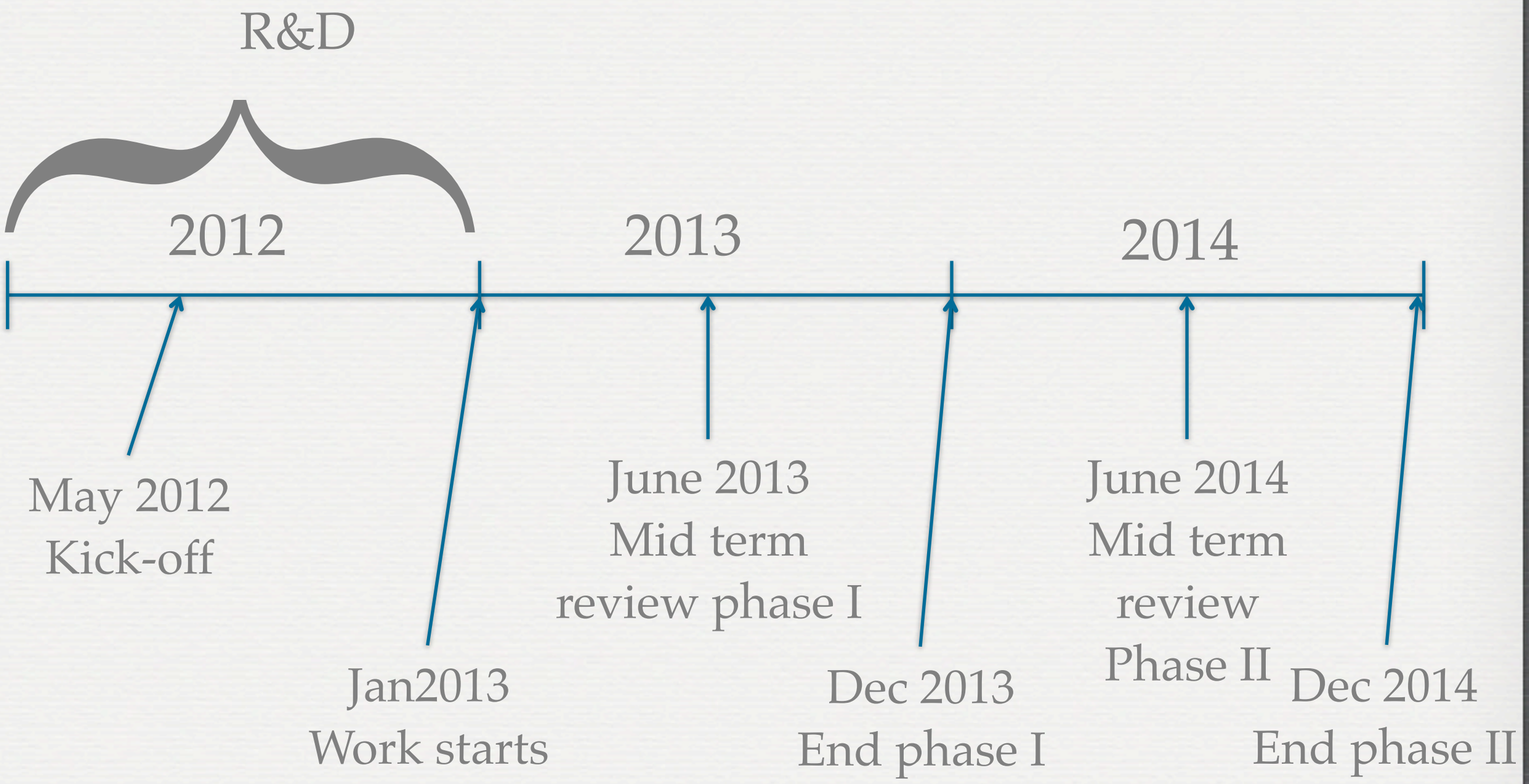
TIMELINE


2012
2013
2014
May 2012
Kick-off
Jan2013
Work starts
June 2013
Mid term
review phase I
Dec 2013
End phase I
June 2014
Mid term
review
Phase II

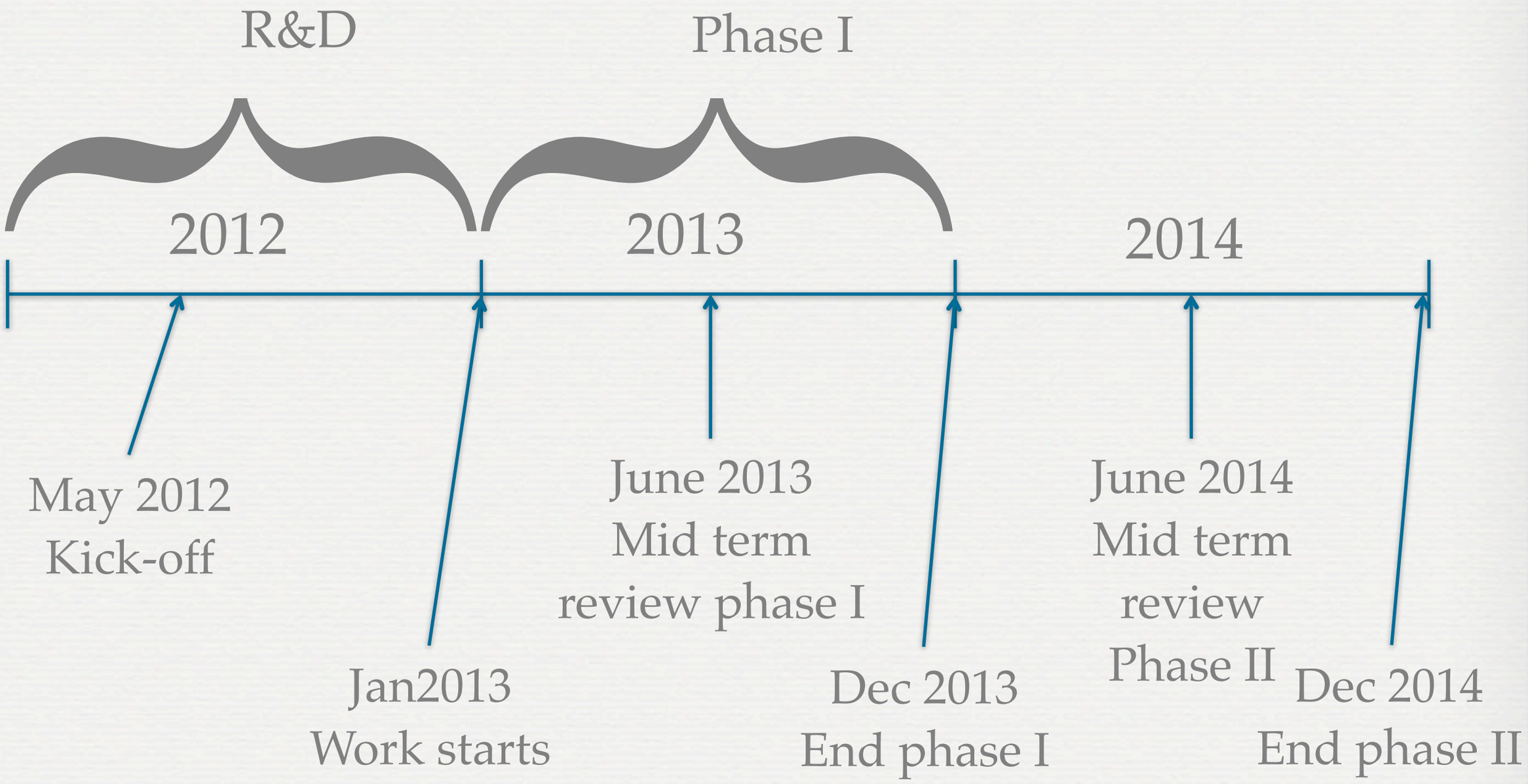
TIMELINE


2012
2013
2014
May 2012
Kick-off
Jan2013
Work starts
June 2013
Mid term
review phase I
Dec 2013
End phase I
June 2014
Mid term
review
Phase II
Dec 2014
End phase II

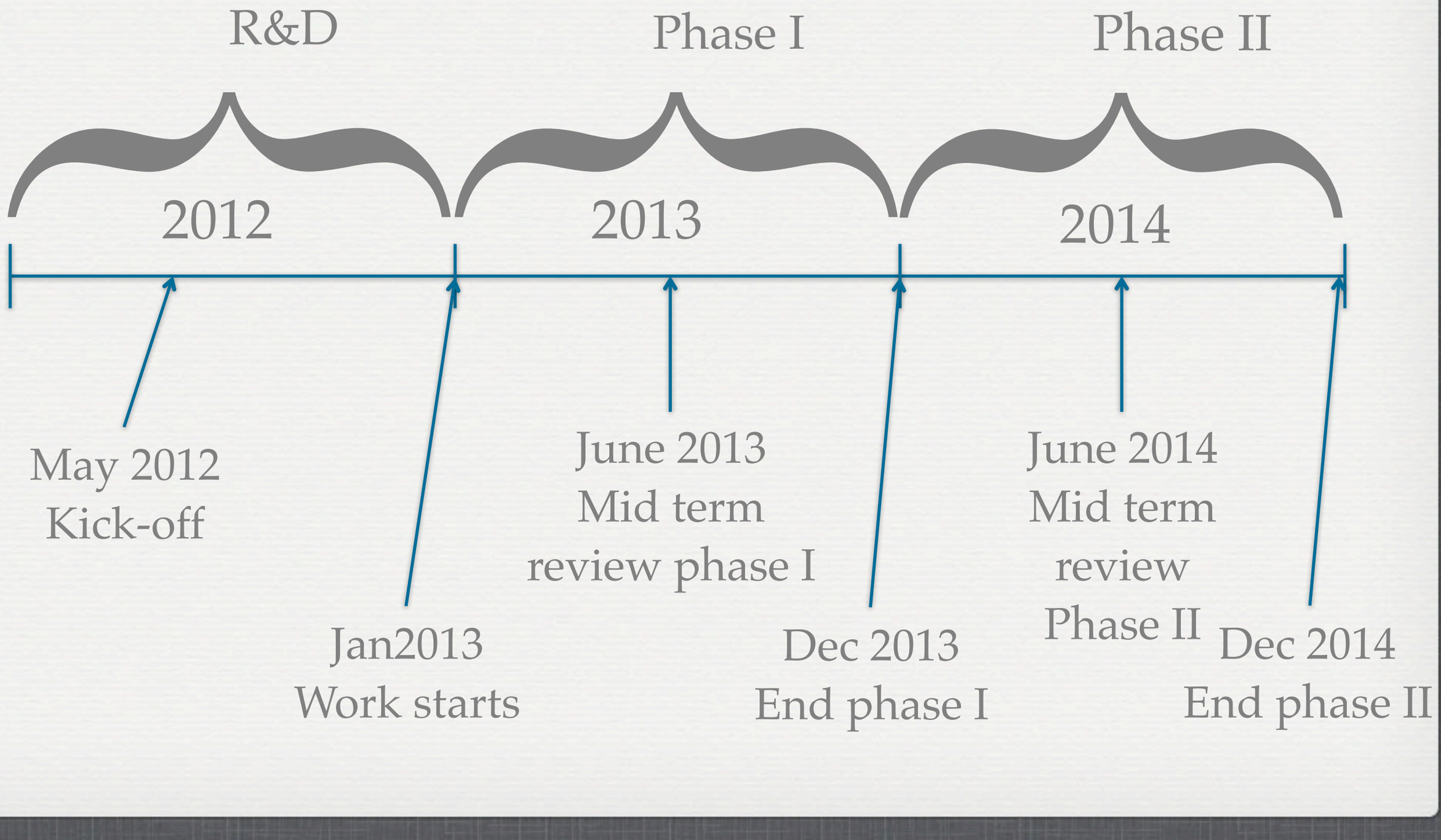
TIMELINE


R&D
2012
2013
2014
May 2012
Kick-off
Jan2013
Work starts
June 2013
Mid term
review phase I
Dec 2013
End phase I
June 2014
Mid term
review
Phase II
Dec 2014
End phase II

TIMELINE


R&D
Phase I
2012
2013
2014
May 2012
Kick-off
Jan2013
Work starts
June 2013
Mid term
review phase I
Dec 2013
End phase I
June 2014
Mid term
review
Phase II
Dec 2014
End phase II

TIMELINE


R&D
Phase I
Phase II
2012
2013
2014
May 2012
Kick-off
Jan2013
Work starts
June 2013
Mid term
review phase I
Dec 2013
End phase I
June 2014
Mid term
review
Phase II
Dec 2014
End phase II

ONE EXAMPLE – SIMULATION

- The LHC experiments use extensively G4 as main simulation engine. They have invested in validation procedures
- One of the reasons why the experiments develop their own fast MC solution is the fact that a full simulation is too slow for several physics analysis
- We would like an architecture where fast and full MC can be run together with the highest performance on parallel systems
- To make it possible one must have a separate particle stack
- However the particle stack depends strongly on the constraints of parallelism. Multiple threads cannot update efficiently a tree data structure.

CONVENTIONAL TRANSPORT

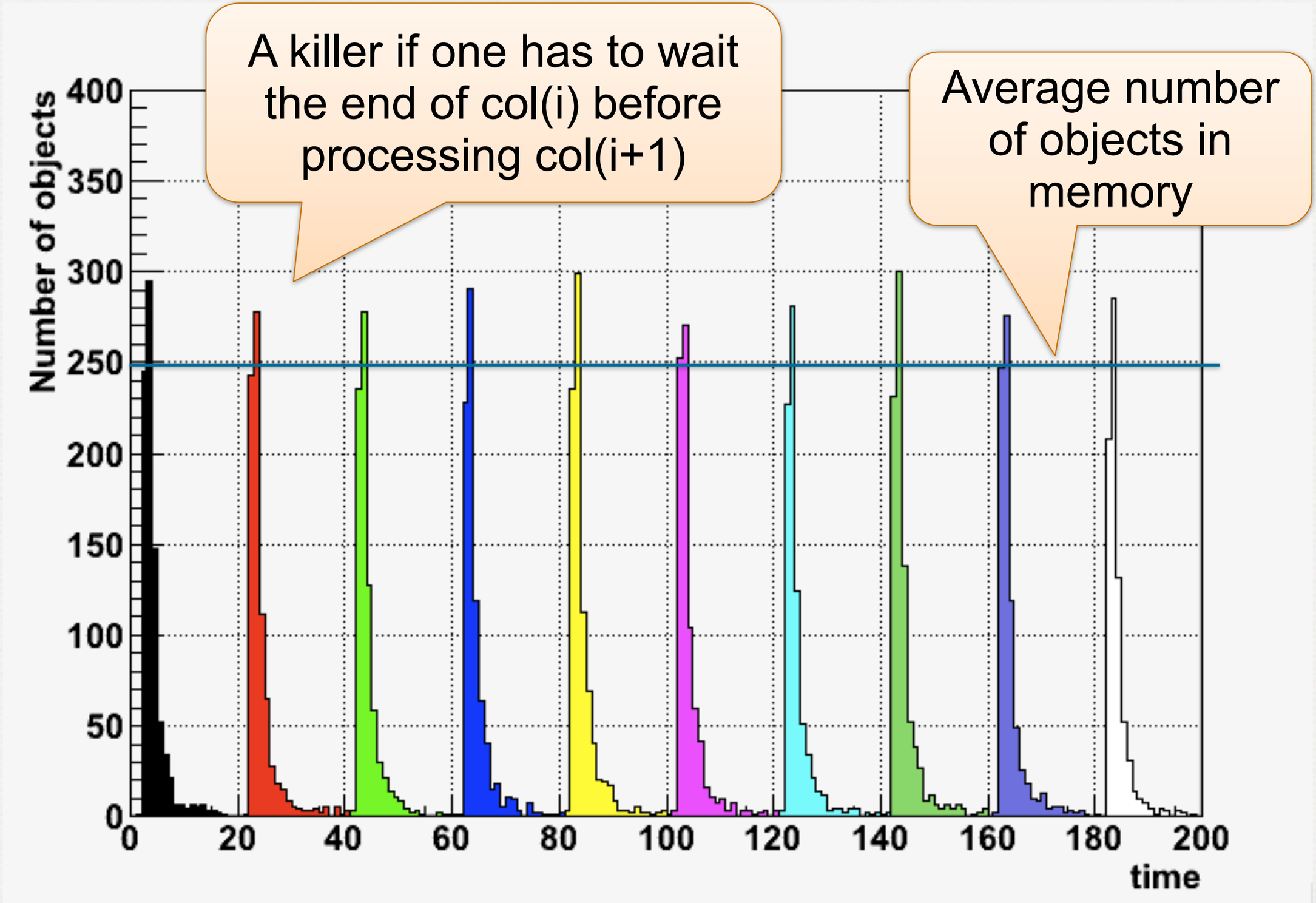
- At each step, the navigator *nav has the state of the particle x,y,z,px,py,pz, the volume instance volume*, etc.
- We compute the distance to the next boundary with something like
  - Dist = nav->DistoOut(volume,x,y,z,px,py,pz)
- Or the distance to one physics process with, eg
  - Distp = nav->DistPhotoEffect(volume,x,y,z,px,py,pz)

Parallelism everywhere again… but how to exploit it?

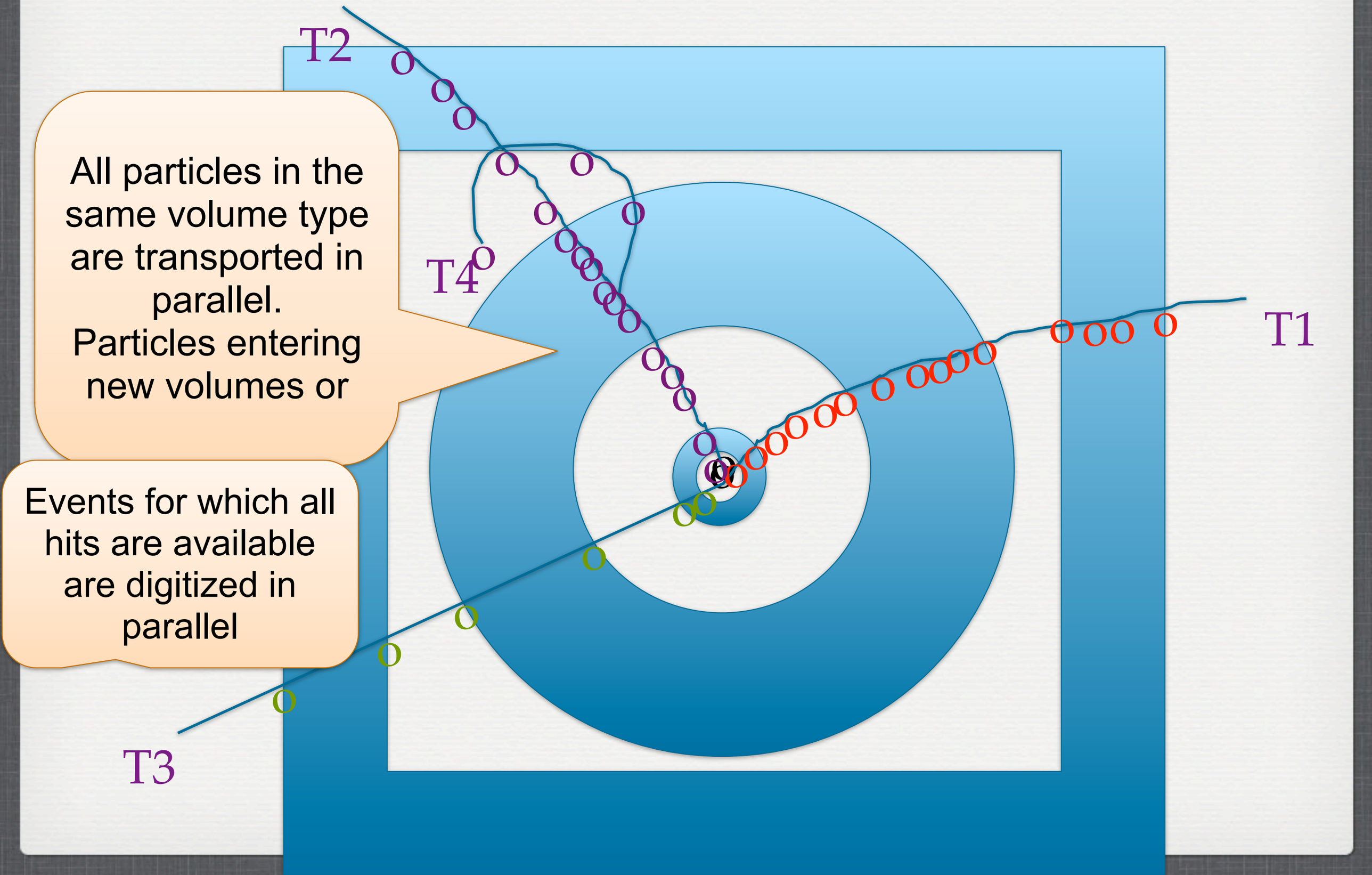
CURRENT SITUATION

- We run jobs in parallel, one per core => it does not scale in case of many cores because it requires too much memory
- A multithreaded version may reduce (say by a factor 2 or 3) the amount of required memory, but it does not fit well with a hierarchy of processors
- We need data structures with internal relations only to allow parallel execution
- When looping on collections, one must avoid the navigation in large memory areas killing the cache
- We must generate vectors well matched to the degree of parallelism and the amount of memory
- We must find a system to avoid the tail effects

TAILS TAILS TAILS...



NEW TRANSPORT SCHEME



All particles in the same volume type are transported in parallel.
Particles entering new volumes or

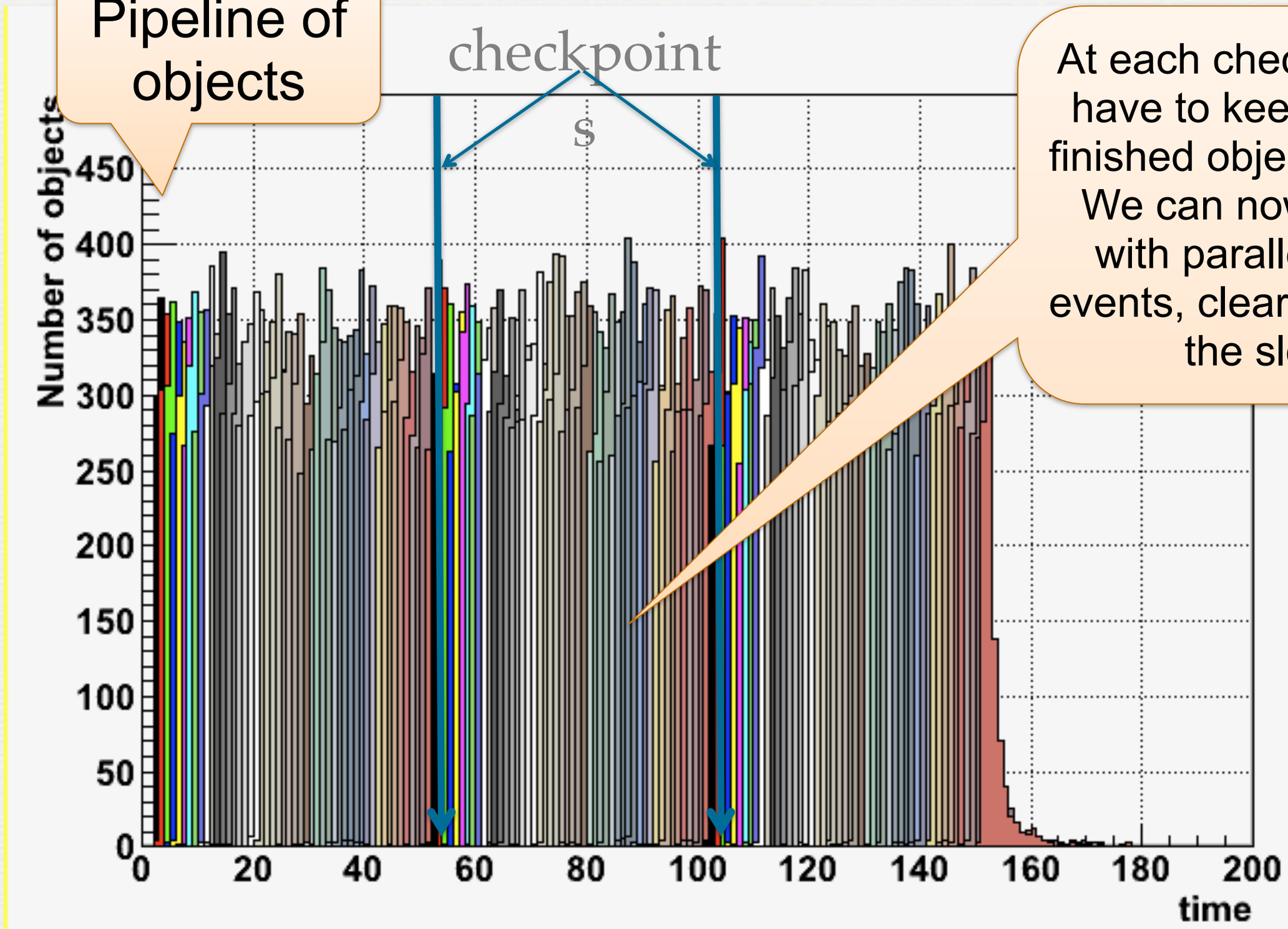
Events for which all hits are available are digitized in parallel

GENERATIONS OF BASKETS

- When a particle enters a volume or is generated, it is added to the basket of particles for the volume type.
- The navigator selects the basket with the highest score (with a high and low water mark algorithm).
- At each step, the navigator *nav has the state of the particles *x,*y,*z,*px,*py,*pz, the volume instances volume** and we compute the distances (array *Dist) to the next boundaries e.g.
  - nav->DistoOut(volume,x,y,z,px,py,pz,Dist)
- Or the distances to one physics process with, eg
  - nav->DistPhotoEffect(volume,x,y,z,px,py,pz,DispP)

A BETTER BETTER SOLUTION

Pipeline of objects



At each checkpoint we have to keep the non finished objects/events. We can now digitize with parallelism on events, clear and reuse the slots.

VECTORIZING THE GEOMETRY

```
Double_t TGeoPara::Safety(Double_t *point, Bool_t in) const
{
   // computes the closest distance from given point to this shape.
   Double_t saf[3];
   // distance from point to higher Z face
   saf[0] = fZ-TMath::Abs(point[2]); // Z

   Double_t yt = point[1]-fTyz*point[2];
   saf[1] = fY-TMath::Abs(yt);       // Y
   // cos of angle YZ
   Double_t cty = 1.0/TMath::Sqrt(1.0+fTyz*fTyz);

   Double_t xt = point[0]-fTxz*point[2]-fTxy*yt;
   saf[2] = fX-TMath::Abs(xt);       // X
   // cos of angle XZ
   Double_t ctx = 1.0/TMath::Sqrt(1.0+fTxy*fTxy+fTxz*fTxz);
   saf[2] *= ctx;
   saf[1] *= cty;
   if (in) return saf[TMath::LocMin(3,saf)];
   for (Int_t i=0; i<3; i++) saf[i]=-saf[i];
   return saf[TMath::LocMax(3,saf)];
}
```

Huge performance gain expected in this type of code where shape constants can be computed outside the loop

PLAN AHEAD (NO TIMING YET)

- Continue exploring all concurrency opportunities
- Develop “virtual transporter” to include a full and fast option
- Introduce embryonic physics processes (em) to simulate shower development
- Evaluate the prototype on parallel architectures
- Evaluate different “parallel” languages (OpenMP, CUDA, OpenCL…)
- Cooperate with experiments
  - For instance with ATLAS ISF (Integrated Simulation Framework) to put together the fast and full MC.

BACK TO ALIROOT

- In the MC example we see how we came to the conclusion that a complete rewrite is necessary
- Possibly a similar conclusion will apply to AliRoot, hence the plan sketched above
- This is why an year of R&D is necessary


"That's
all
folks!"