CERN

1st ASTERICS-OBELICS Workshop

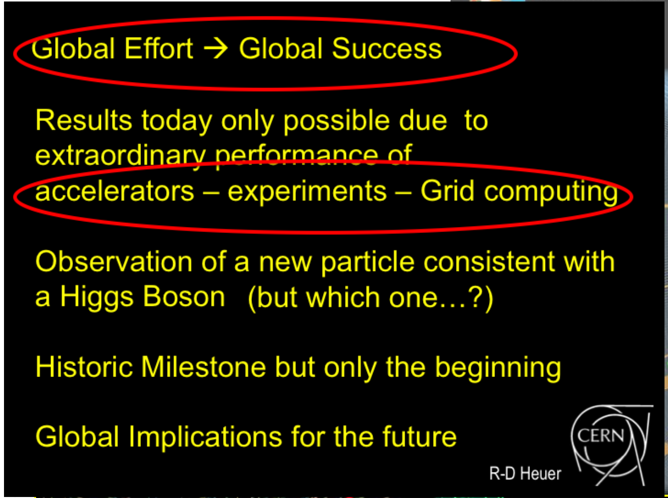
12-14 December 2016, Rome, Italy.

A globally distributed data management solution

WLCG – the LHC's offline computing platform

Oliver Keeble
CERN Storage Group

H2020-Astronomy ESFRI and Research Infrastructure Cluster (Grant Agreement number: 653477).



Global Effort → Global Success

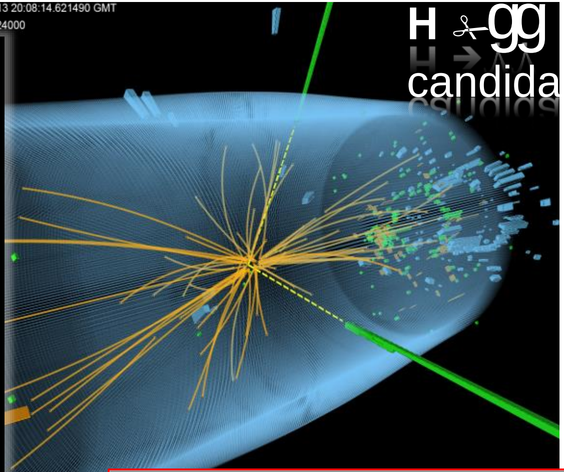
Results today only possible due to extraordinary performance of accelerators – experiments – Grid computing

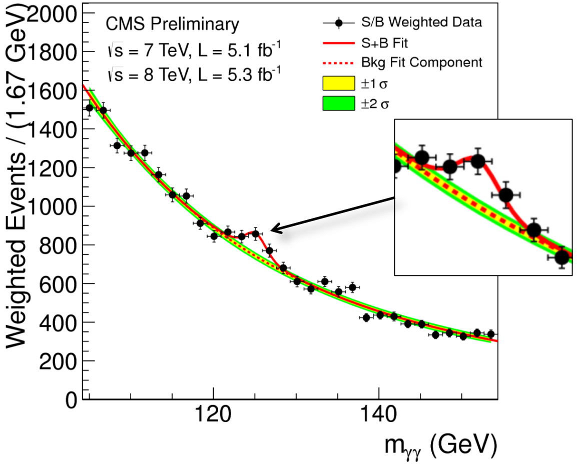
Observation of a new particle consistent with a Higgs Boson (but which one…?)

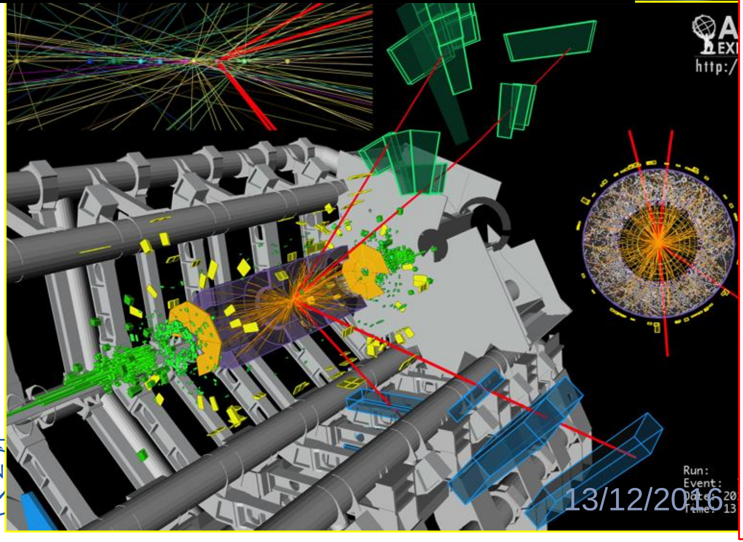
Historic Milestone but only the beginning

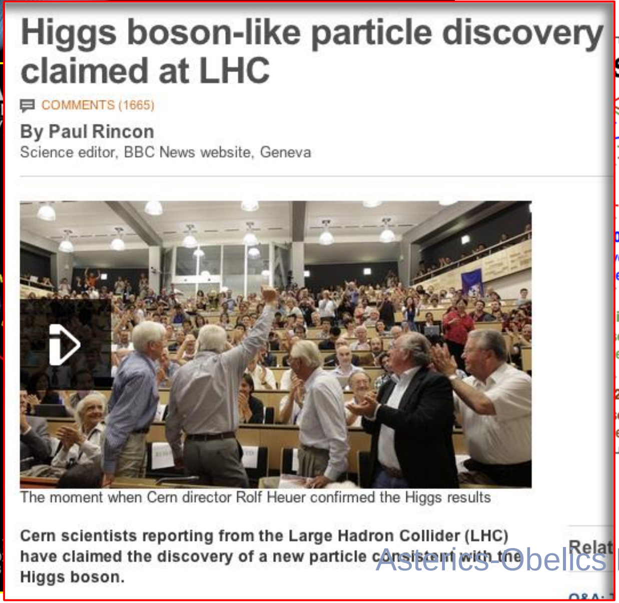
Global Implications for the future

R-D Heuer


Data recorded: 2012-May-13 20:08:14.621490 GMT
Run/Event: 194108 / 564224000
H → gg
candida







Higgs boson-like particle discovery claimed at LHC

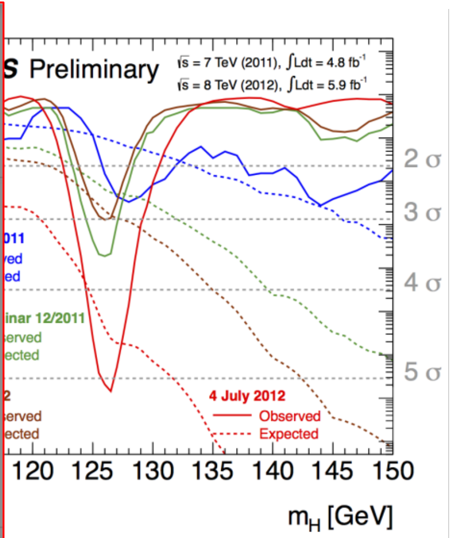
COMMENTS (1665)

By Paul Rincon
Science editor, BBC News website, Geneva

The moment when Cern director Rolf Heuer confirmed the Higgs results

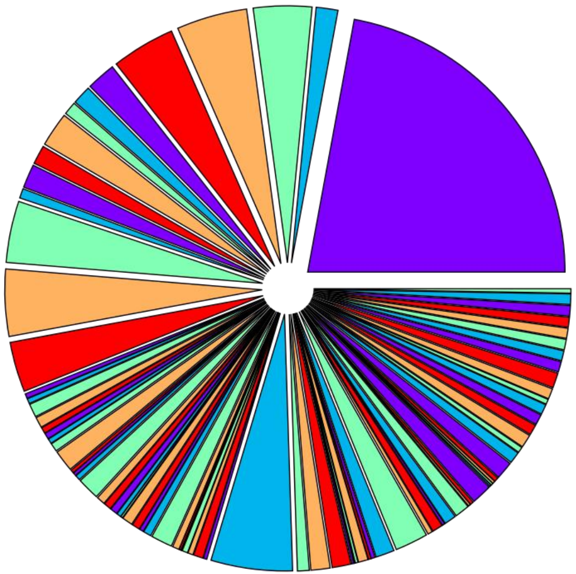
Cern scientists reporting from the Large Hadron Collider (LHC) have claimed the discovery of a new particle consistent with the Higgs boson.

Relat

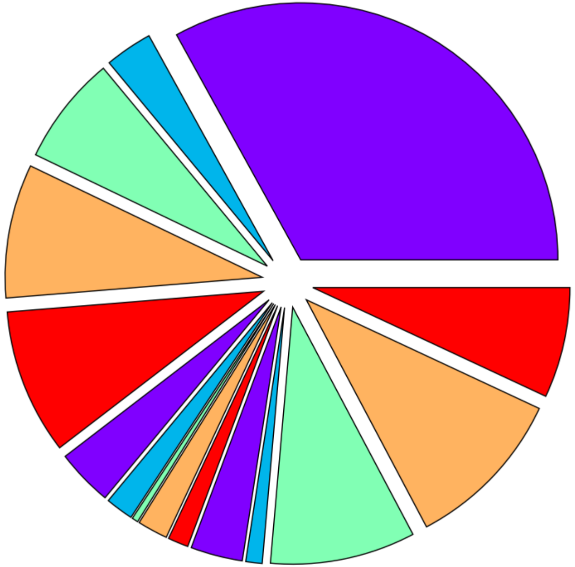


Pledged Resources

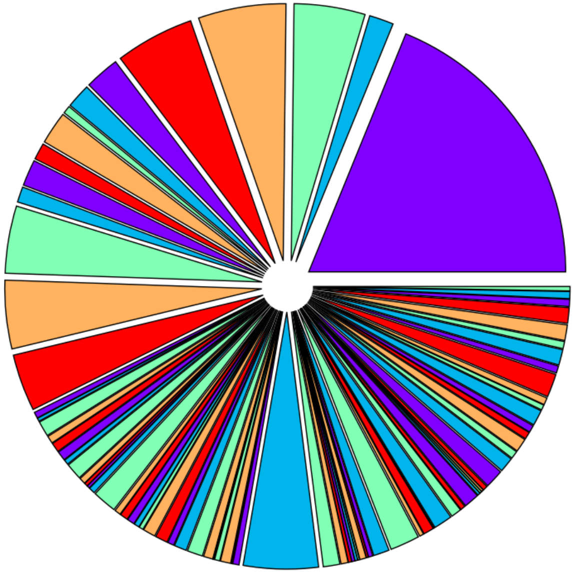
CPU by computer centre
Total ~400k cores

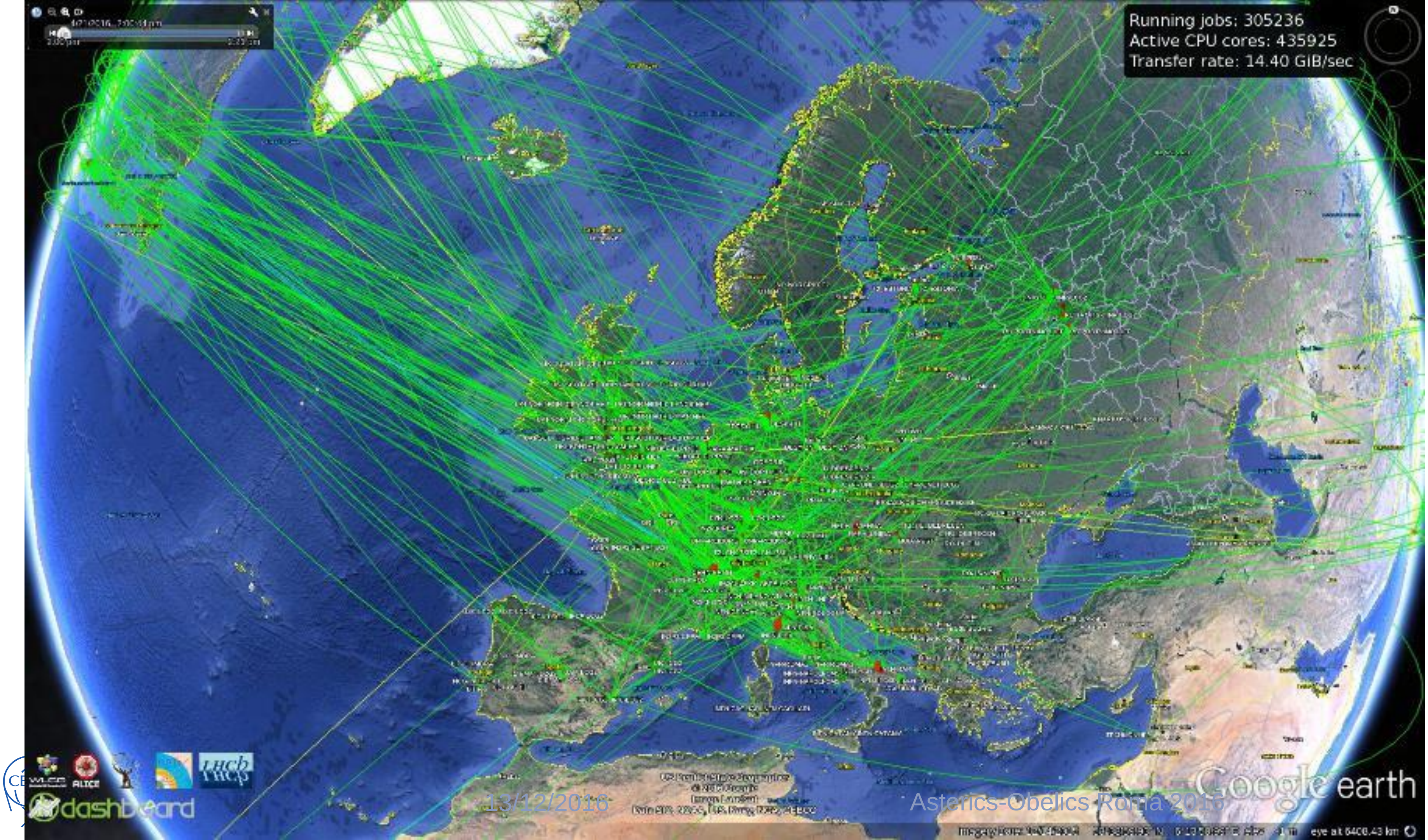


Tape by computer centre
Total ~400 PB



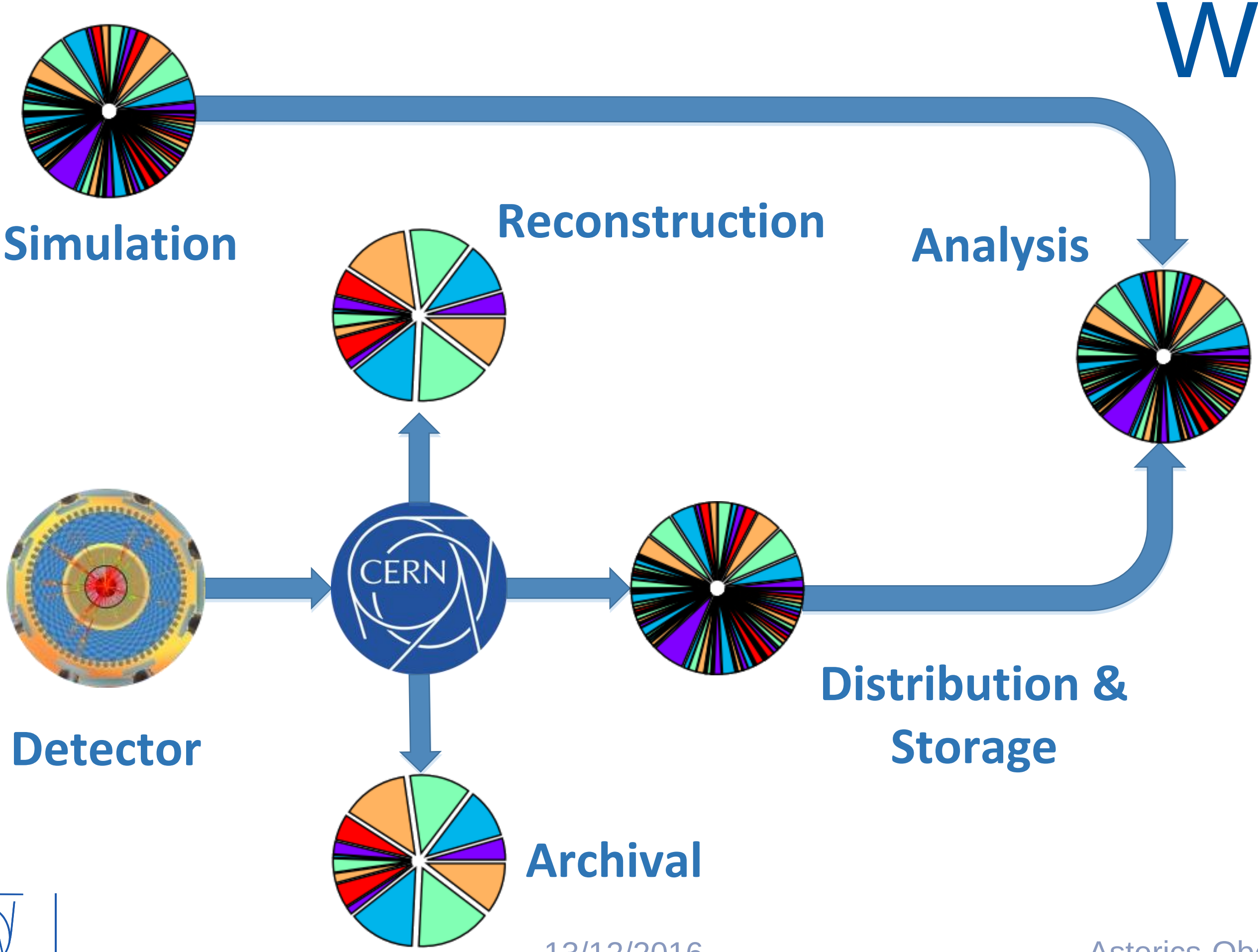
Disk by computer centre
Total ~300 PB





Running jobs: 305236
Active CPU cores: 435925
Transfer rate: 14.40 GiB/sec

Workflows



Simulation

Reconstruction

Analysis

Detector

Distribution & Storage

Archival

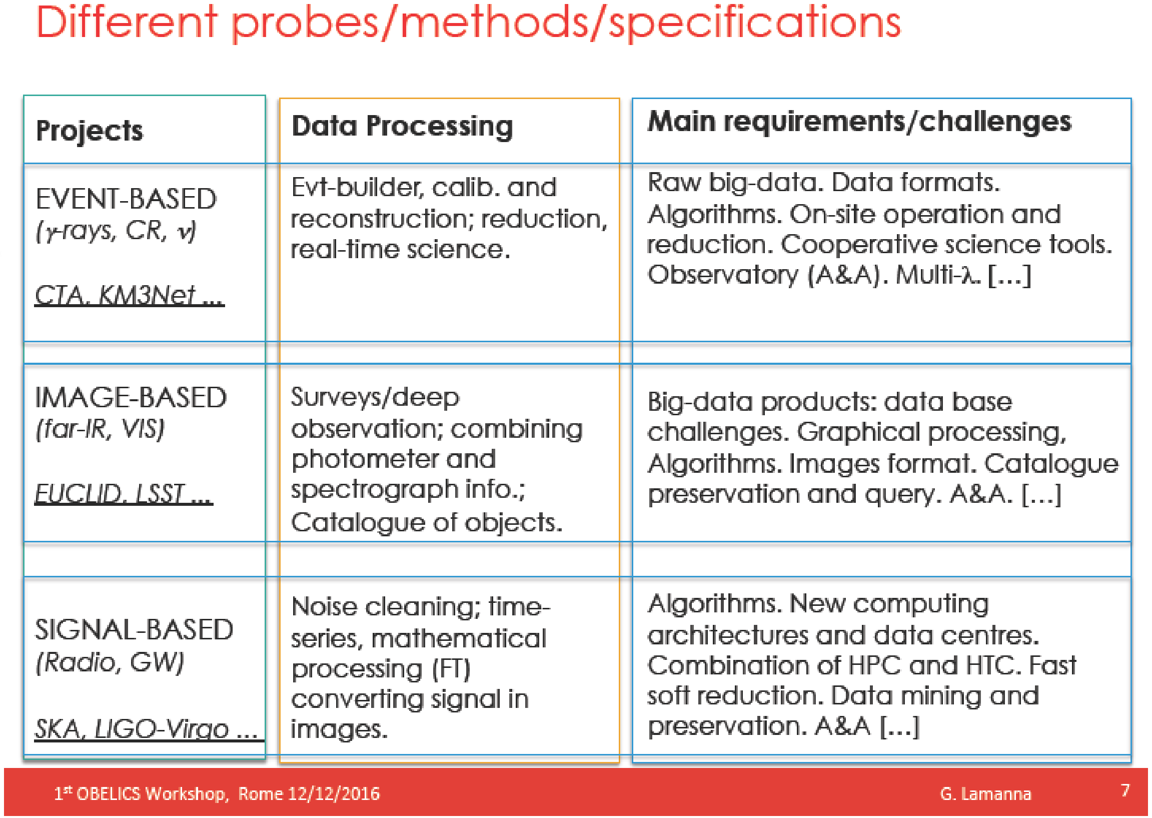
More than half the CPU goes on simulation.

Most of the rest is reconstruction.

The remainder is analysis.

From yesterday…

We are here!



Different probes/methods/specifications

Projects	Data Processing	Main requirements/challenges
EVENT-BASED *(γ-rays, CR, ν)* *CTA, KM3Net …*	Evt-builder, calib. and reconstruction; reduction, real-time science.	Raw big-data. Data formats. Algorithms. On-site operation and reduction. Cooperative science tools. Observatory (A&A). Multi-λ. […]
IMAGE-BASED *(far-IR, VIS)* *EUCLID, LSST …*	Surveys/deep observation; combining photometer and spectrograph info.; Catalogue of objects.	Big-data products: data base challenges. Graphical processing, Algorithms. Images format. Catalogue preservation and query. A&A. […]
SIGNAL-BASED *(Radio, GW)* *SKA, LIGO-Virgo …*	Noise cleaning; time-series, mathematical processing (FT) converting signal in images.	Algorithms. New computing architectures and data centres. Combination of HPC and HTC. Fast soft reduction. Data mining and preservation. A&A […]

1st OBELICS Workshop, Rome 12/12/2016 G. Lamanna 7

What advantages do we have?

- Event independence
- “Read-only” data
- File-based data
- Scientific Linux
- Coarse grained (VO) authorisation
- X509 acceptance
- Large proportion of CPU intensive work

WLCG Stack

Applications	Analysis, reco	ROOT
Community Services	Workflow, Data Management	Atlas, Alice, CMS, LHCb
Infrastructure Services	Transfer, Identity, Mon	FTS, X509, monit
Site services	Storage, Compute	DPM, CREAM
Site, Cloud, HPC, …	Storage, CPU, Network	Ceph, Torque

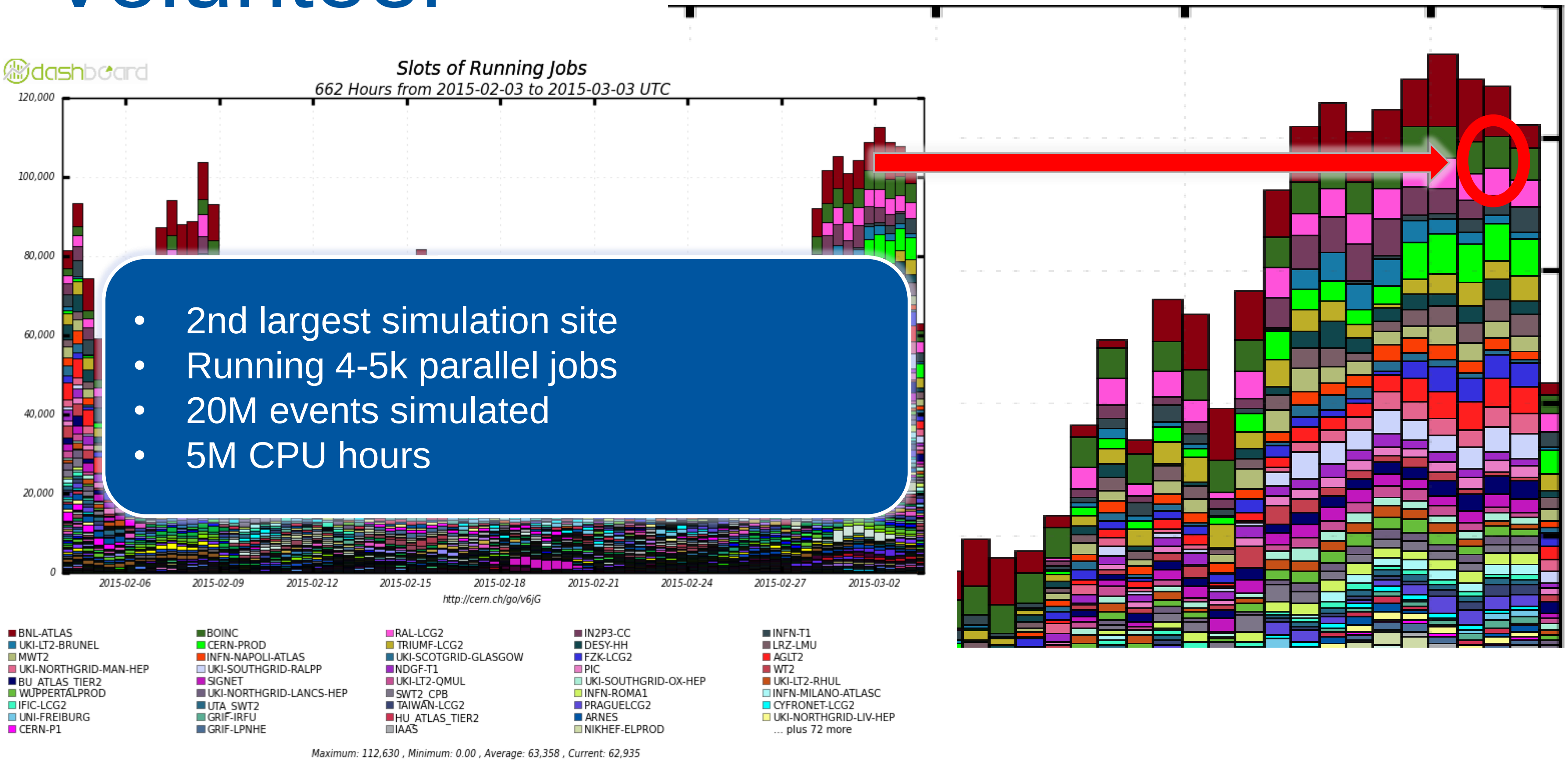
WLCG Stack

Applications	Analysis, reco	ROOT
Community Services	Workflow, Data Management	Atlas, Alice, CMS, LHCb
Infrastructure Services	Transfer, Identity, Mon	FTS, X509, monit
Site services	Storage, Compute	DPM, CREAM
Site, Cloud, HPC, …	Storage, CPU, Network	Ceph, Torque

Not (only) a grid

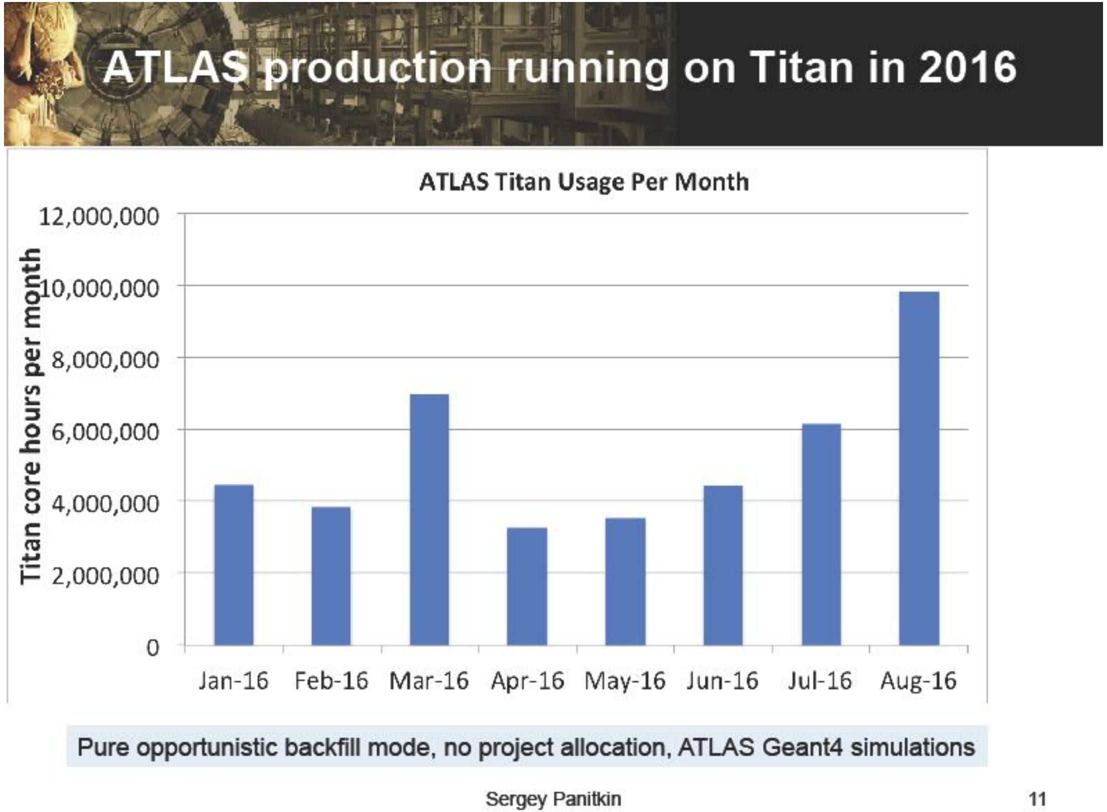
- The WLCG infrastructure comprises
  - Grid - pledged
  - Cloud - rented
  - HPC - allocated
  - Volunteer - donated
  - Concepts
    - Opportunistic resources
    - Pre-emptibility

Volunteer



- 2nd largest simulation site
- Running 4-5k parallel jobs
- 20M events simulated
- 5M CPU hours

HPC – backfill on Titan



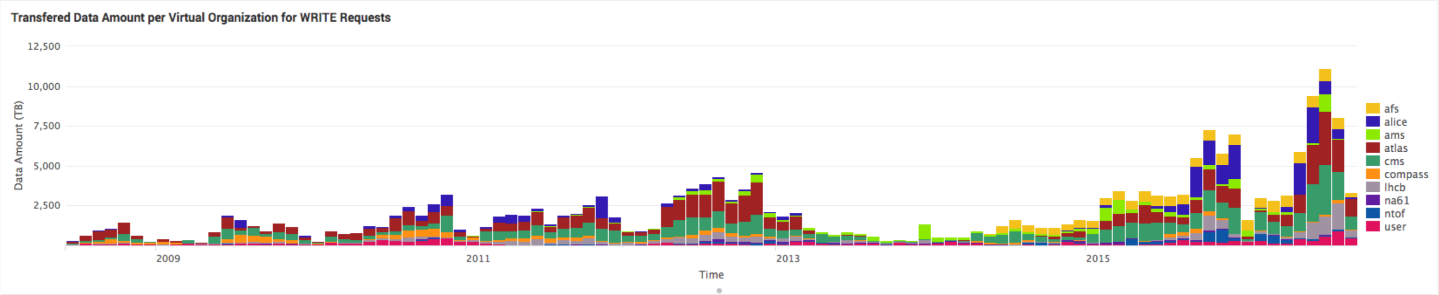
ATLAS production running on Titan in 2016

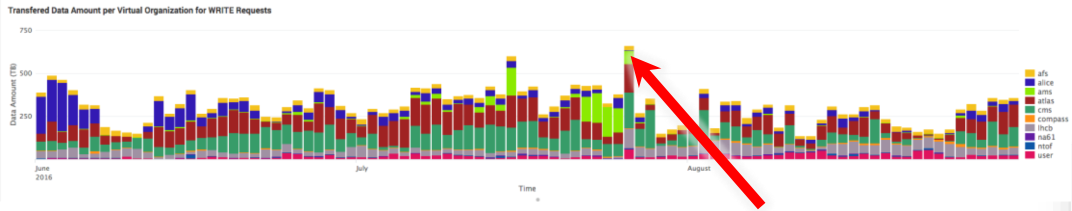
ATLAS Titan Usage Per Month

Pure opportunistic backfill mode, no project allocation, ATLAS Geant4 simulations

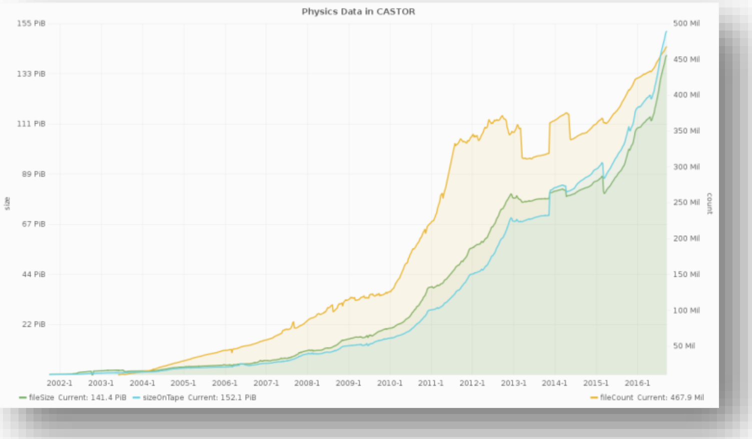
Sergey Panitkin 11

2016 data volumes





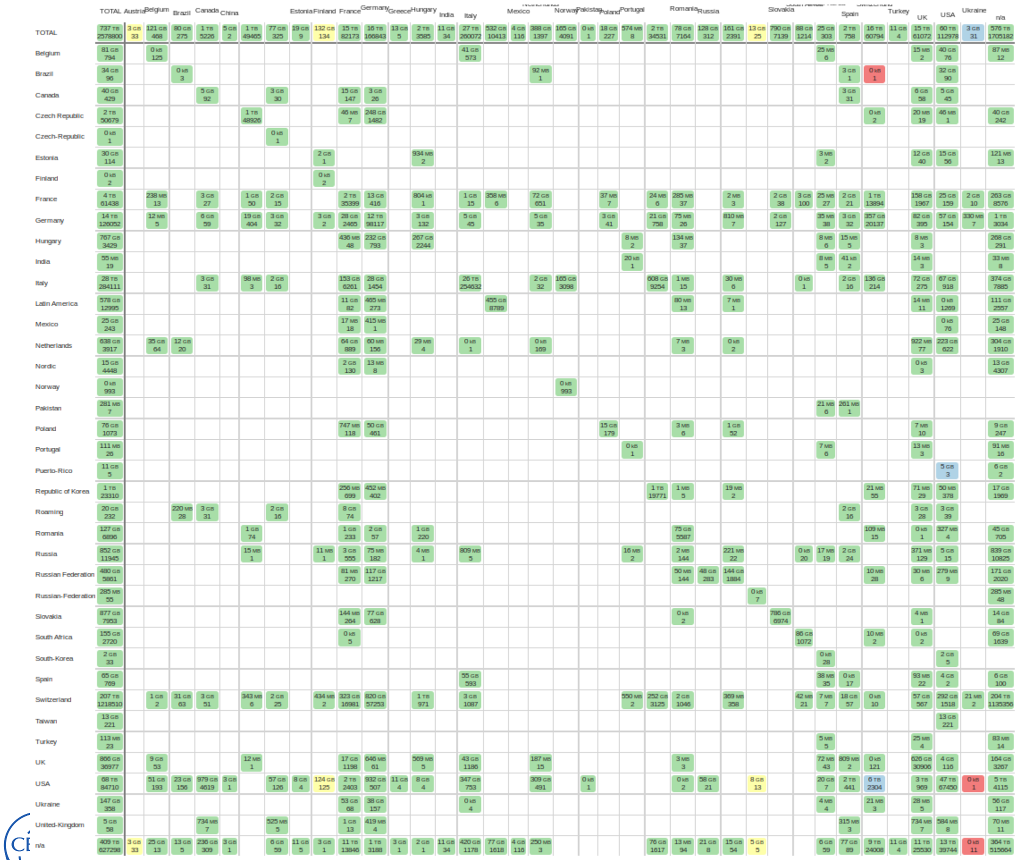
~160 PB on tape at CERN
500 M files



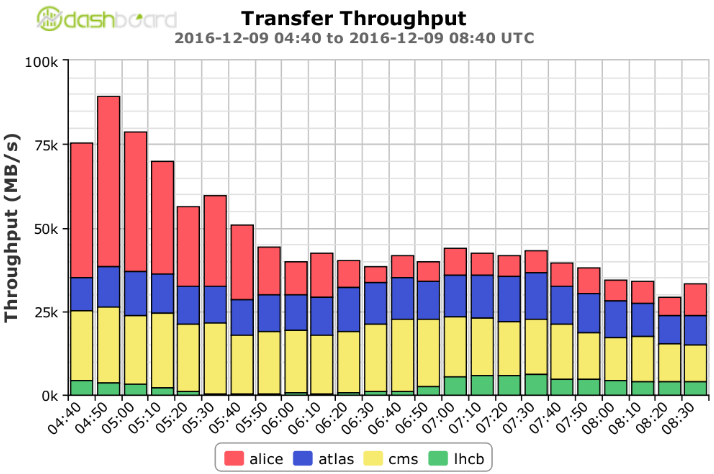
LHC data – Continue to break records:
10.7 PB recorded in July
CERN archive ~160 PB

June-Aug 2016
>500 TB / day
(Run 1 peak for HI was 220 TB)

Data - transfer



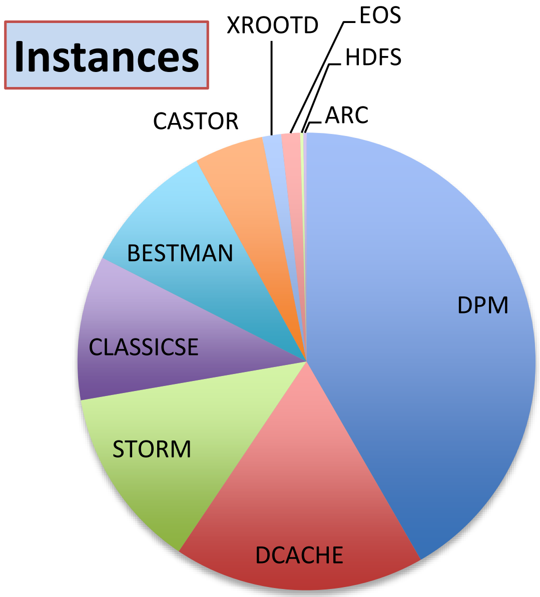
	TOTAL	Austria	Belgium	Brazil	Canada	China			Estonia	Finland	France	Germany	Greece	Hungary	India	Italy		Mexico	Netherlands	Norway	Pakistan	Poland	Portugal		Romania	Russia			Slovakia	South Africa	South-Korea	Spain	Switzerland	Turkey	UK	USA	Ukraine	n/a
TOTAL	737 TB 2578800	3 GB 33	121 GB 468	80 GB 275	1 TB 5226	5 GB 2	1 TB 49465	77 GB 325	19 GB 9	132 GB 134	15 TB 82173	16 TB 166843	13 GB 5	2 TB 3585	11 GB 34	27 TB 260072	532 GB 10413	4 GB 116	388 GB 1397	165 GB 4091	0 kB 1	18 GB 227	574 MB 8	2 TB 34531	78 GB 7164	128 GB 312	161 GB 2391	13 GB 25	790 GB 7139	88 GB 1214	25 GB 303	2 TB 758	16 TB 60794	11 GB 4	15 TB 61072	60 TB 112978	3 GB 31	576 TB 1705182
Belgium	81 GB 794		0 kB 125													41 GB 573															25 MB 6				15 MB 2	40 GB 76		87 MB 12
Brazil	34 GB 96			0 kB 3															92 MB 1													3 GB 1	0 kB 1			32 GB 90		
Canada	40 GB 429				5 GB 92			3 GB 30			15 GB 147	3 GB 26																				3 GB 31			6 GB 58	5 GB 45		
Czech Republic	2 TB 50679						1 TB 48926				46 MB 7	248 GB 1482																					0 kB 2		20 MB 19	46 MB 1		40 GB 242
Czech-Republic	0 kB 1							0 kB 1																														
Estonia	30 GB 114									2 GB 1				934 MB 2																	3 MB 2				12 GB 40	15 GB 56		121 MB 13
Finland	0 kB 2									0 kB 2																												
France	4 TB 61438		238 MB 13		3 GB 27		1 GB 50	2 GB 15			2 TB 35399	13 GB 416		804 kB 1		1 GB 15	358 MB 6		72 GB 651			37 MB 7		24 MB 6	285 MB 37		2 MB 3		2 GB 38	3 GB 100	25 MB 27	2 GB 21	1 TB 13894		158 GB 1967	25 GB 159	2 GB 10	263 GB 8576
Germany	14 TB 126052		12 MB 5		6 GB 59		19 GB 404	3 GB 32		3 GB 2	28 GB 2465	12 TB 98117		3 GB 132		5 GB 45			5 GB 35			3 GB 41		21 GB 758	75 MB 26		810 MB 7		2 GB 127		35 MB 38	3 GB 32	357 GB 20137		82 GB 395	57 GB 154	330 MB 7	1 TB 3034
Hungary	767 GB 3429										436 MB 48	232 GB 793		267 GB 2244									8 MB 2		134 MB 37						8 MB 6	15 MB 5			8 MB 3			268 GB 291
India	55 MB 19																						20 kB 1								8 MB 5	41 kB 2			14 MB 3			33 MB 8
Italy	28 TB 284111				3 GB 31		98 MB 3	2 GB 16			153 GB 6261	28 GB 1454				26 TB 254632			2 GB 32	165 GB 3098				608 GB 9254	1 MB 15		30 MB 6			0 kB 1		2 GB 16	136 GB 214		72 GB 275	67 GB 918		374 GB 7885
Latin America	578 GB 12995										11 GB 82	465 MB 273					455 GB 8789								80 MB 13		7 MB 1								14 MB 11	0 kB 1269		111 GB 2557
Mexico	25 GB 243										17 MB 18	415 MB 1																								0 kB 76		25 GB 148
Netherlands	638 GB 3917		35 GB 64	12 GB 20							64 GB 889	60 MB 156		29 MB 4		0 kB 1			0 kB 169						7 MB 3		0 kB 2								922 MB 77	223 GB 622		304 GB 1910
Nordic	15 GB 4448										2 GB 130	13 MB 8																							0 kB 3			13 GB 4307
Norway	0 kB 993																			0 kB 993																		
Pakistan	281 MB 7																														21 MB 6	261 MB 1						
Poland	76 GB 1073										747 MB 118	50 GB 461										15 GB 179			3 MB 6		1 GB 52								7 MB 10			9 GB 247
Portugal	111 MB 26																						0 kB 1								7 MB 6				13 MB 3			91 MB 16
Puerto-Rico	11 GB 5																																			5 GB 3		6 GB 2
Republic of Korea	1 TB 23310										256 MB 699	452 MB 402												1 TB 19771	1 MB 5		19 MB 2						21 MB 55		71 MB 29	50 MB 378		17 GB 1969
Roaming	20 GB 232			220 MB 28	3 GB 31			2 GB 16			8 GB 74																					2 GB 16			3 GB 28	3 GB 39		
Romania	127 GB 6896						1 GB 74				1 GB 233	2 GB 57		1 GB 220											75 GB 5587								109 MB 15		0 kB 1	327 MB 4		45 GB 705
Russia	852 GB 11945						15 MB 1			11 MB 1	3 GB 555	75 MB 182		4 MB 1		809 MB 5							16 MB 2		2 MB 144		221 MB 22			0 kB 20	17 MB 19	2 GB 24			371 MB 129	5 GB 15		839 GB 10825
Russian Federation	480 GB 5861										81 MB 270	117 GB 1217													50 MB 144	48 GB 283	144 GB 1884						10 MB 28		30 MB 6	279 MB 9		171 GB 2020
Russian-Federation	285 MB 55																											0 kB 7										285 MB 48
Slovakia	877 GB 7953										144 MB 264	77 GB 628													0 kB 2				786 GB 6974						4 MB 1			14 GB 84
South Africa	155 GB 2720										0 kB 5																			86 GB 1072			10 MB 2		0 kB 2			69 GB 1639
South-Korea	2 GB 33																														0 kB 28					2 GB 5		
Spain	65 GB 769															55 GB 593															38 MB 35	0 kB 17			93 MB 22	4 GB 2		6 GB 100
Switzerland	207 TB 1218510		1 GB 2	31 GB 63	3 GB 51		343 MB 6	2 GB 25		434 MB 2	323 GB 16981	820 GB 57253		1 TB 971		3 GB 1087							550 MB 2	252 GB 3125	2 GB 1046		369 MB 358			42 MB 21	7 MB 7	18 GB 57	0 kB 10		57 GB 567	292 GB 1518	21 MB 2	204 TB 1135356
Taiwan	13 GB 221																																			13 GB 221		
Turkey	113 MB 23																														5 MB 5				25 MB 4			83 MB 14
UK	866 GB 36977		9 GB 53				12 MB 1				17 GB 1198	646 MB 61		569 MB 5		43 GB 1186			187 MB 15						3 MB 3						72 MB 43	809 MB 2	0 kB 121		626 GB 30906	4 GB 116		164 GB 3267
USA	68 TB 84710		51 GB 193	23 GB 156	979 GB 4619	3 GB 1		57 GB 126	8 GB 4	124 GB 125	2 TB 2403	932 GB 507	11 GB 4	8 GB 4		347 GB 753			309 GB 491		0 kB 1				0 kB 2	58 GB 21		8 GB 13			20 GB 7	2 TB 441	6 TB 2304		3 TB 969	47 TB 67450	0 kB 1	5 TB 4115
Ukraine	147 GB 358										53 GB 68	38 GB 157				0 kB 4															4 MB 4		21 MB 3		28 MB 5			56 GB 117
United-Kingdom	5 GB 58				734 MB 7			525 MB 5			1 GB 13	419 MB 4																				315 MB 3			734 MB 7	584 MB 8		70 MB 11
n/a	409 TB 627298	3 GB 33	25 GB 13	13 GB 5	236 GB 309	3 GB 1		6 GB 59	11 GB 5	3 GB 1	11 TB 13846	1 TB 3188	3 GB 1	2 GB 1	11 GB 34	420 GB 1178	77 GB 1618	4 GB 116	250 MB 3					76 GB 1617	13 MB 94	21 GB 8	15 GB 54	5 GB 5			6 GB 59	77 GB 89	9 TB 24008	11 GB 4	11 TB 25530	13 TB 39744	0 kB 11	364 TB 515664

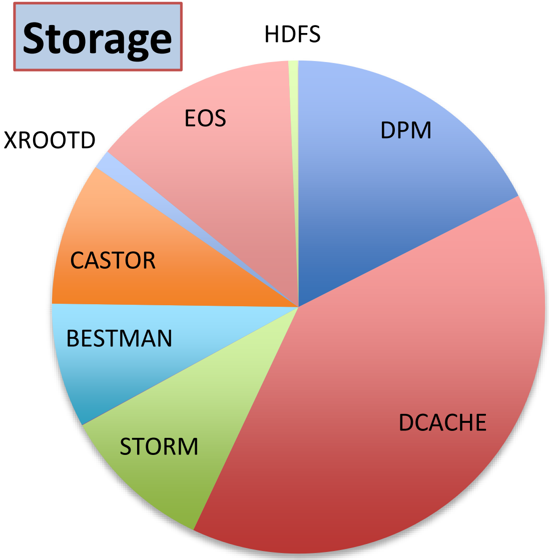


Most LHC transfers are managed by the File Transfer Service (FTS)

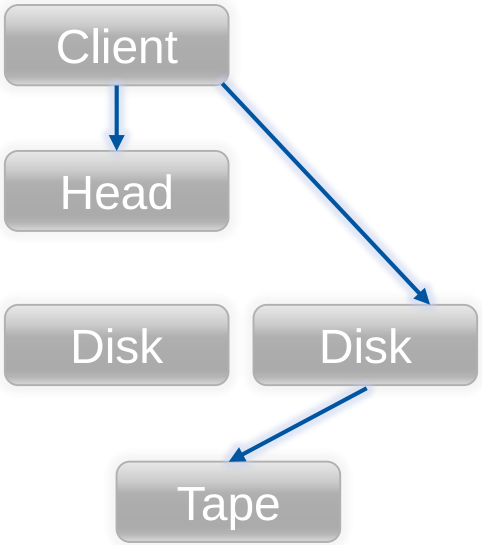
Try it at https://webfts.cern.ch

Data – storage systems



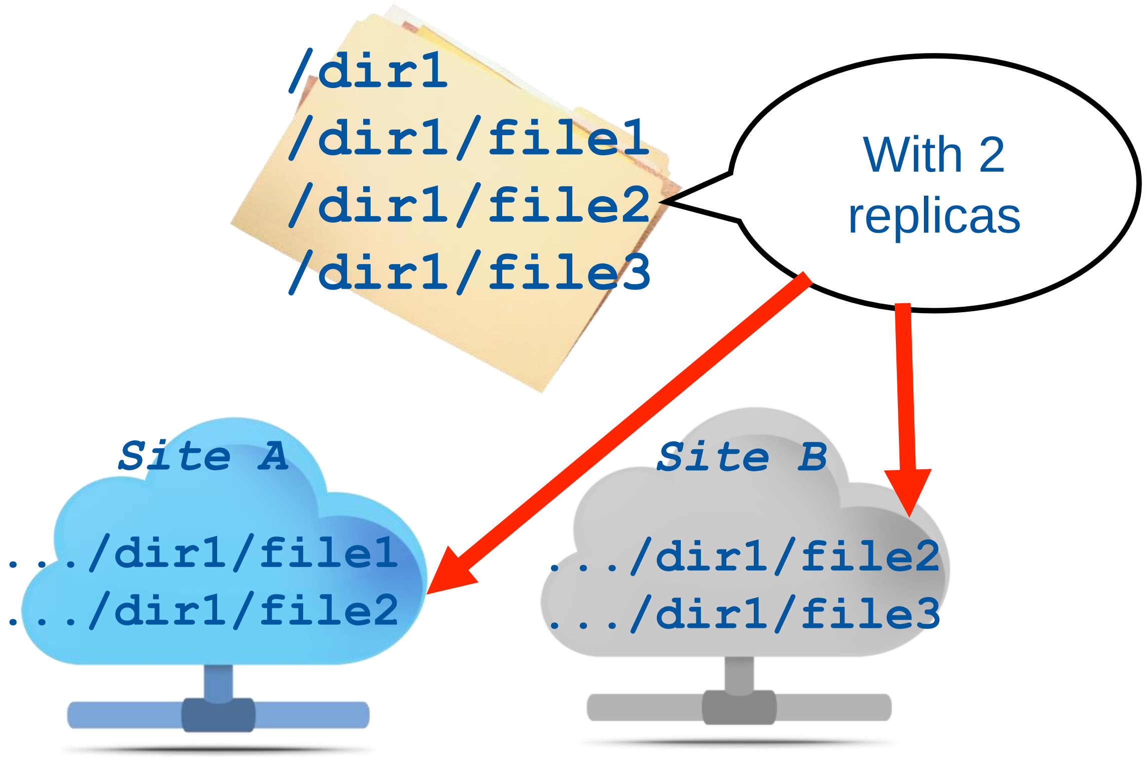


Basic Architecture



Systems are evolving towards standards. All now offer HTTPS access.

Global Data Federation



```
/dir1
/dir1/file1
/dir1/file2
/dir1/file3
```

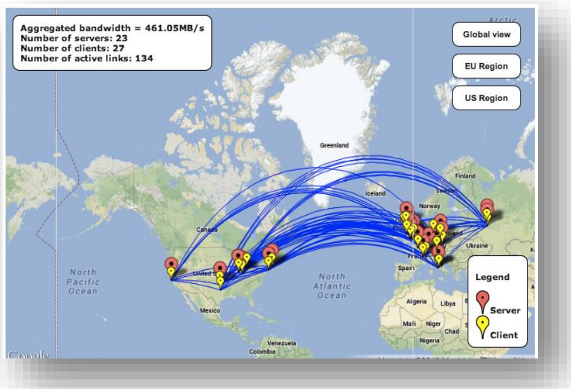
With 2 replicas

Site A

```
.../dir1/file1
.../dir1/file2
```

Site B

```
.../dir1/file2
.../dir1/file3
```



Aggregated bandwidth = 461.05MB/s
Number of servers: 23
Number of clients: 27
Number of active links: 134

In use by

- Atlas (FAX)
- CMS (AAA)
- Main uses
  - Failover
  - Overflow
  - Diskless sites

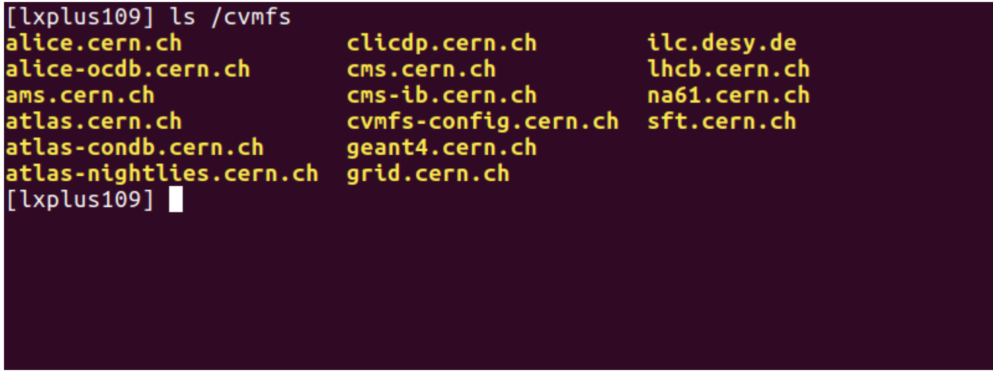
Data – what is it?

```
Open https://server/data.root
While (next event) {
  do stuff;
}
```

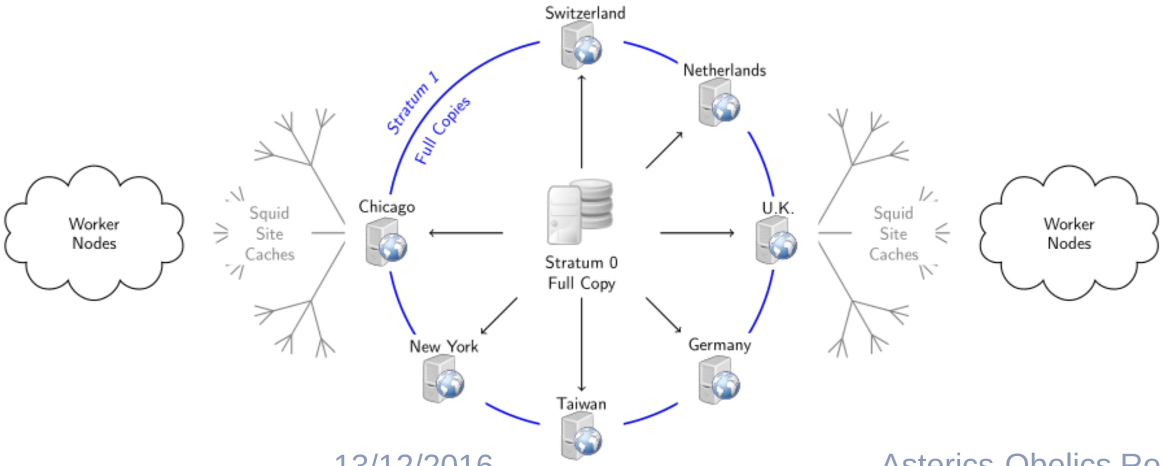
- ROOT files
- Typically a few GB each
- Column-like structured storage
- Lots of I/O optimisation
  - WAN access

Software distribution

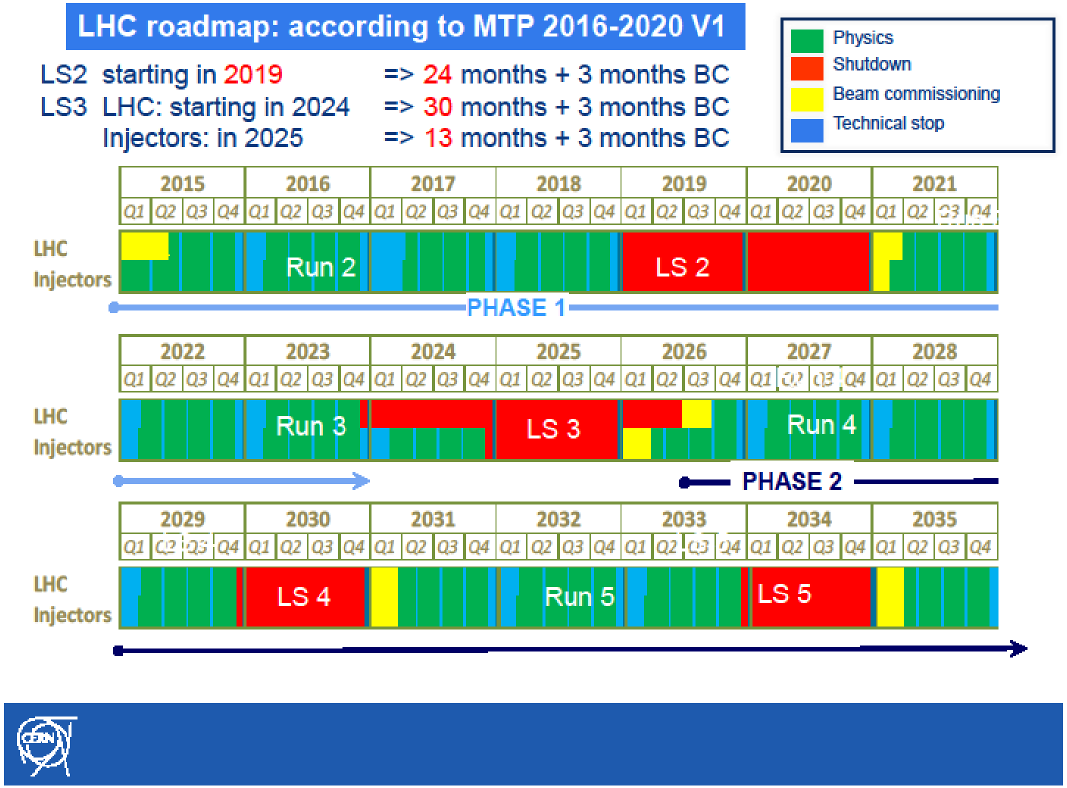
- CVMFS
- r/o cached fs
- >350M files



```
[lxplus109] ls /cvmfs
alice.cern.ch              clicdp.cern.ch        ilc.desy.de
alice-ocdb.cern.ch         cms.cern.ch           lhcb.cern.ch
ams.cern.ch                cms-ib.cern.ch        na61.cern.ch
atlas.cern.ch              cvmfs-config.cern.ch  sft.cern.ch
atlas-condb.cern.ch        geant4.cern.ch
atlas-nightlies.cern.ch    grid.cern.ch
[lxplus109] 
```



The road ahead

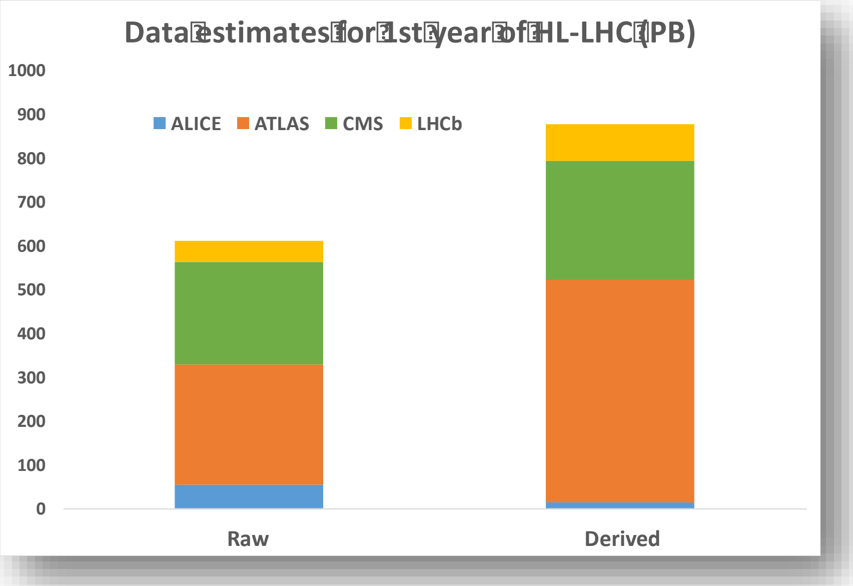


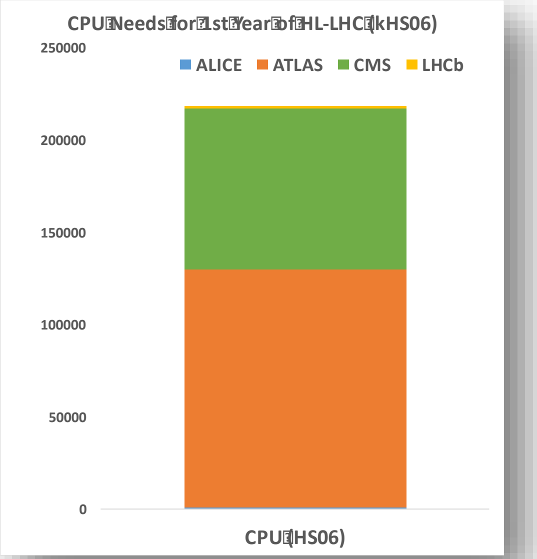
LHC roadmap: according to MTP 2016-2020 V1

LS2 starting in 2019 => 24 months + 3 months BC
LS3 LHC: starting in 2024 => 30 months + 3 months BC
Injectors: in 2025 => 13 months + 3 months BC

🟩 Physics
🟥 Shutdown
🟨 Beam commissioning
🟦 Technical stop

High-Lumi LHC resource estimates





Data:

- X10 from 2016
  - Raw 2016: 50 PB → 2027: 600 PB
  - Derived (1 copy): 2016: 80 PB → 2027: 900 PB

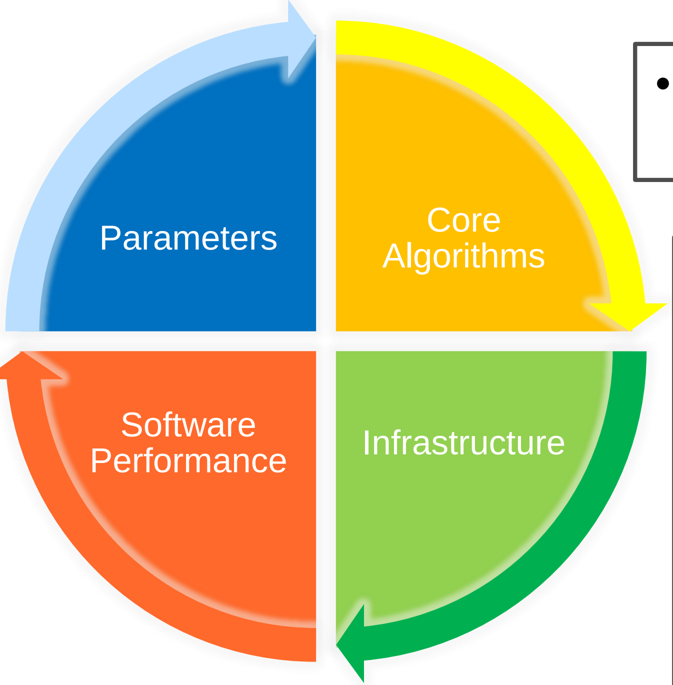
CPU:

- x60 from 2016

Technology at ~20%/year will bring x6-10 in 10-11 years

HL-LHC Solutions

- Raw data
- Triggers
- Detector design
- …



- Reconstruction and simulation algorithms

- Performance/architectures/memory
- Tools
- Concurrency
- Vectorisation
- Collaboration with externals – via HSF
- …

- New grid/cloud models
- Optimise CPU/disk/network
- Economies of scale via clouds, joint procurements etc.
- Opportunistic resources
- Pre-emptible jobs
- Storage consolidation
  - WAN access
- Data strategies
- Caching solutions
- …

Summary

- WLCG is the production offline computing platform for the 4 LHC experiments
- Can process multiple PB / month
- In 2025 we will have a new accelerator with new experiments
  - Order of magnitude more load at fixed cost
  - Technology sharing = sustainability & reduced costs

H2020-Astronomy ESFRI and Research Infrastructure Cluster (Grant Agreement number: 653477).

Acknowledgement

H2020-Astronomy ESFRI and Research Infrastructure Cluster (Grant Agreement number: 653477).

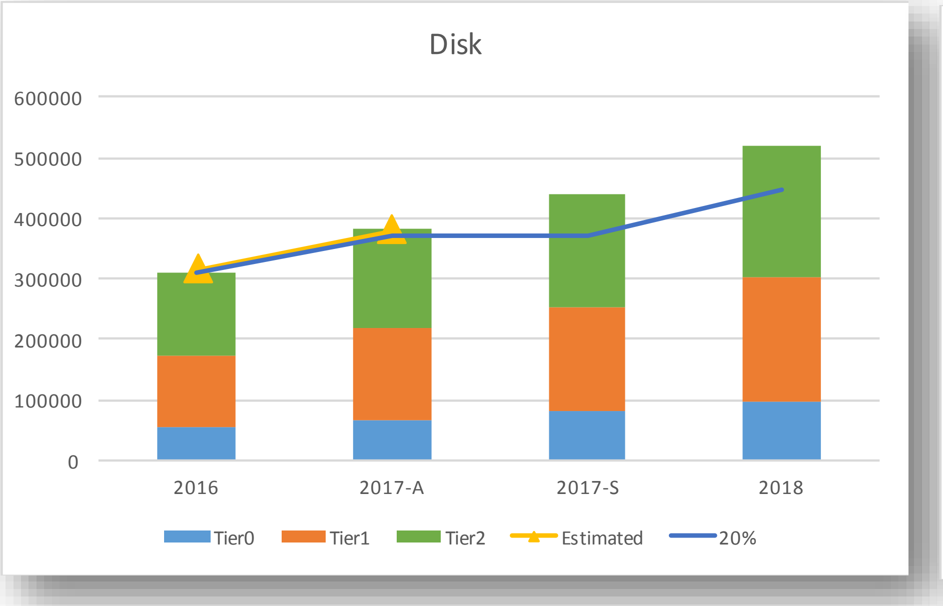
Supplementary Slides

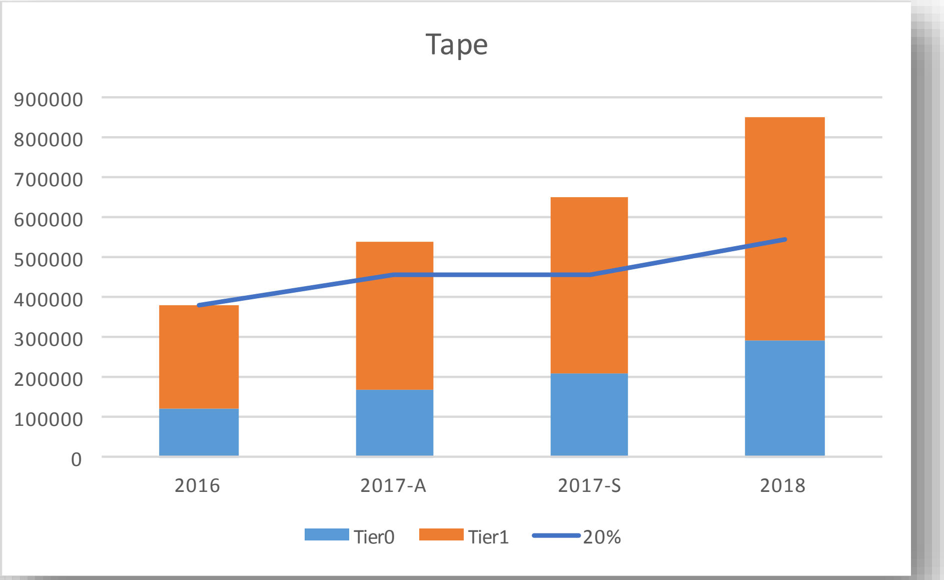
Data Management Directions

- Reduce cost/volume
  - cost of storage management
    - integrating standard (non HEP) solutions e.g. ceph
    - protocol zoo, SRM-less operation
    - T2 storage as cache
    - multi-site storage
    - regional federations
    - cloud storage
    - system manageability
  - storage overheads
    - redundancy
      - replication, erasure, RAID levels etc
    - reduce system reliability requirements?
      - reduce cost/impact of data loss
    - component technology
      - shingled disks
      - consumer/enterprise disks

- Reduce volume used
  - reduced number of global replicas
    - remote access
    - latency hiding
      - applications, overcommitting
    - global federations
    - CPU-only resources (inc cloud)
  - data formats and lifecycle, intermediate products
  - resource reporting
    - monitoring usage
    - eliminating dark data
  - data "enrichment"
    - popularity
    - caching, avoiding unused data
      - promoting locality in workflows
  - trading disk for…
    - tape
      - data parking
    - CPU
      - maintain metadata enabling regeneration of data on demand

LHC: “outstanding performance”





Estimated: Estimates made in 2014 for Run 2 up to 2017

20%: Growth of 20%/yr starting in 2016 (“flat budget”)