LHCb Computing

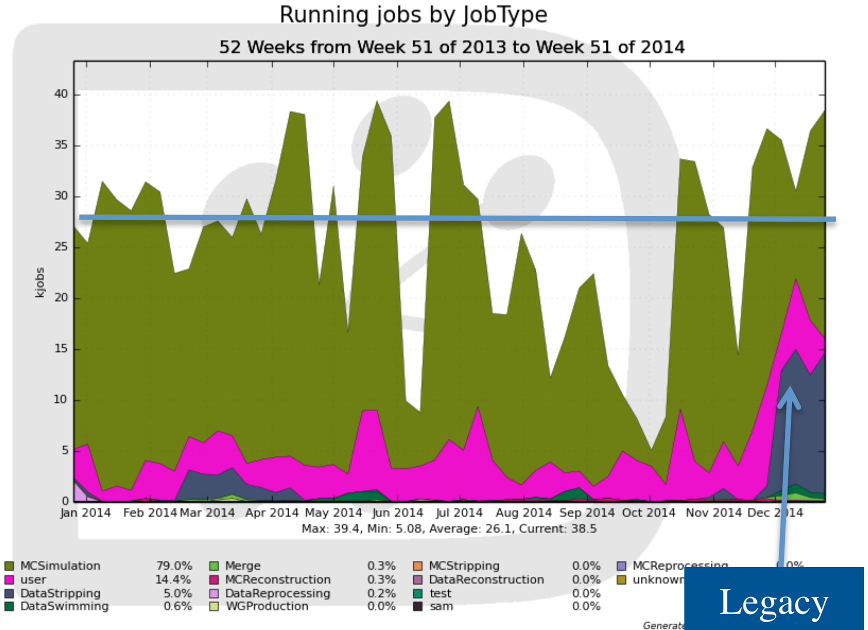
Resources: 2014 usage
2015 status
2016 requests

Concezio Bozzi

Bologna, May 25th 2015

Usage in 2014

- Main activity: MC production
- Other activites:
  - Incremental stripping in spring 2014
  - legacy stripping of Run1 data (2 months in Dec 2014-Jan 2015)
  - "swimming" production in Spring 2014
  - User jobs



WLCG Tier1

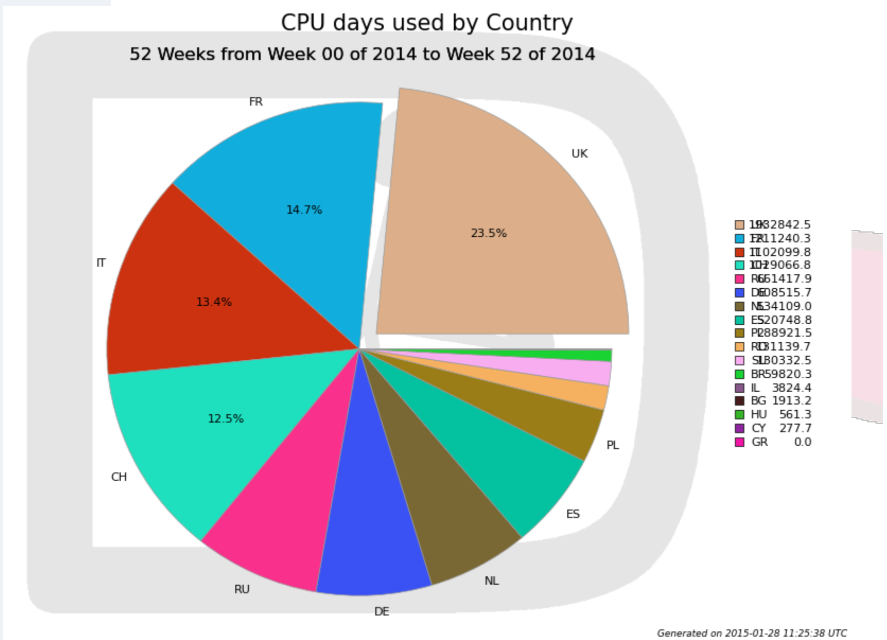
<Power>	Used (kHS06)	Pledge (kHS06)
CH-CERN	15.6	34.0
DE-KIT	14.6	19.2
ES-PIC	7.8	7.1
FR-CCIN2P3	20.6	21.7
IT-INFN-CNAF	23.2	19.8
NL-T1	13.7	13.8
RRC-KI-T1	15.2	10.8
UK-T1-RAL	20.4	34.7
Total	**131.2**	**161.1**

WLCG Tier 2

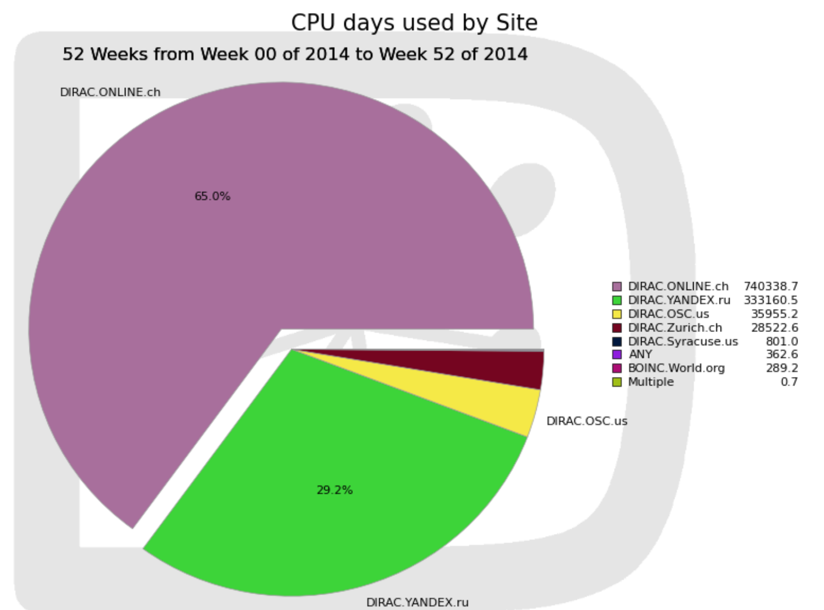
<Power>	(kHS06)	(kHS06)
Brazil	1.2	8.0
France	12.0	11.0
Germany	0.2	3.2
Italy	3.9	8.5
Netherlands	2.7	0.0
Poland	7.5	3.2
Romania	2.3	4.9
Russia	8.9	0.1
Spain	10.4	2.8
Switzerland	4.2	5.2
UK	30.6	10.1
Total	**84.0**	**57.0**

Resource usage: CPU

WLCG resources, by Country



Non-WLCG resources, by Site

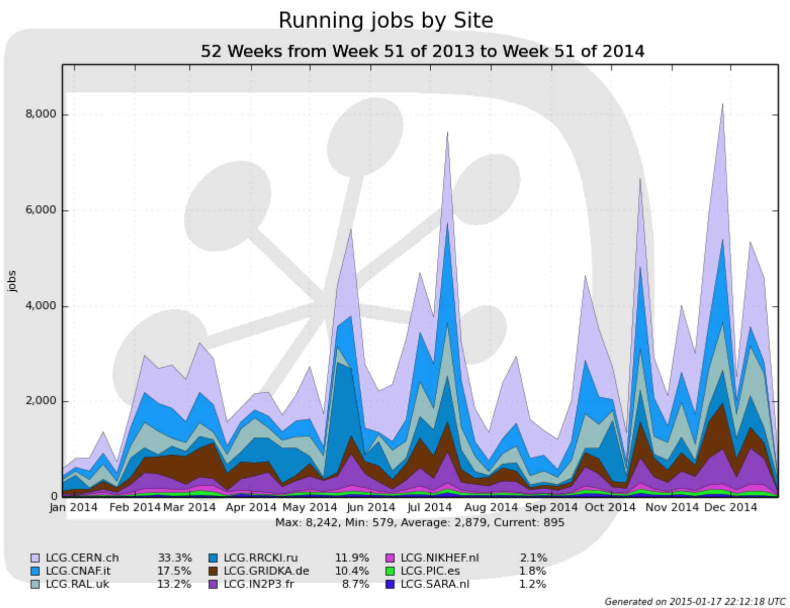


Ohio Supercomputing Center (OSC) providing ~ 4kHS06

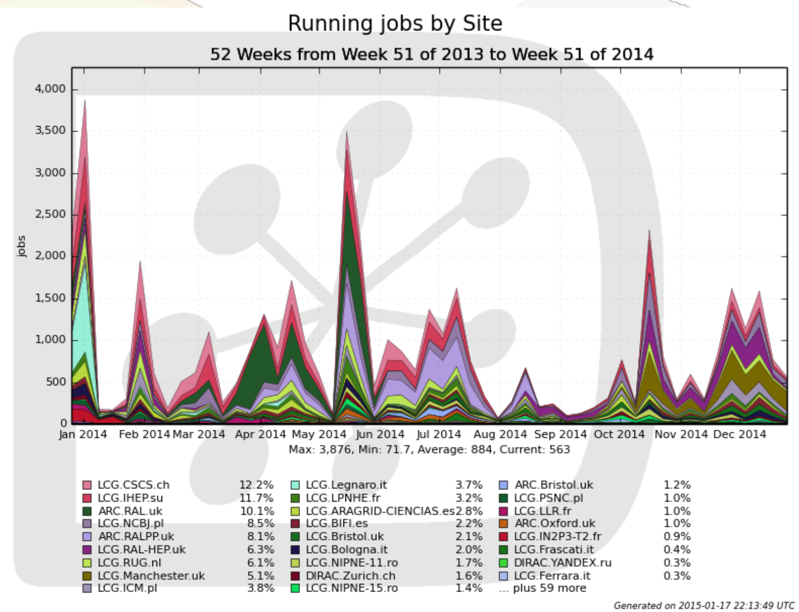
~50% at Tier0 + Tier1s
~50% outside

Resource usage: CPU, user jobs

Tier0 + Tier1s



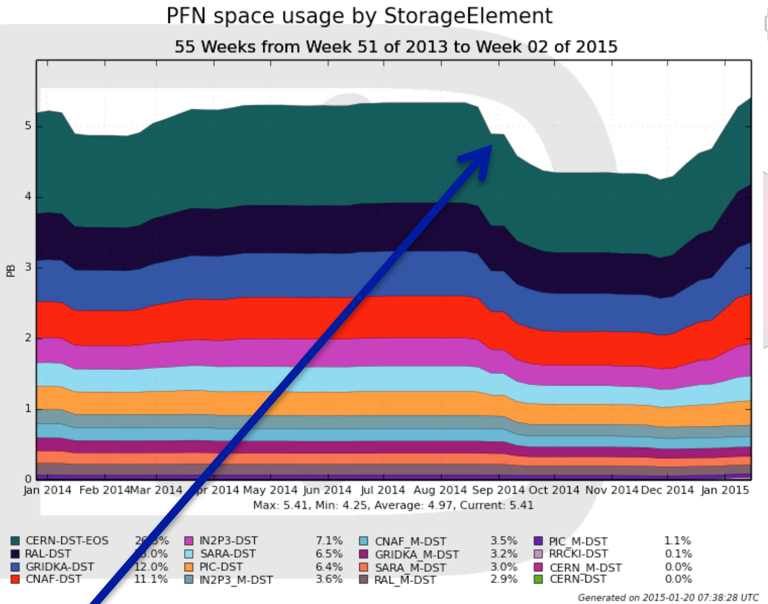
Outside Tier0 + Tier1s



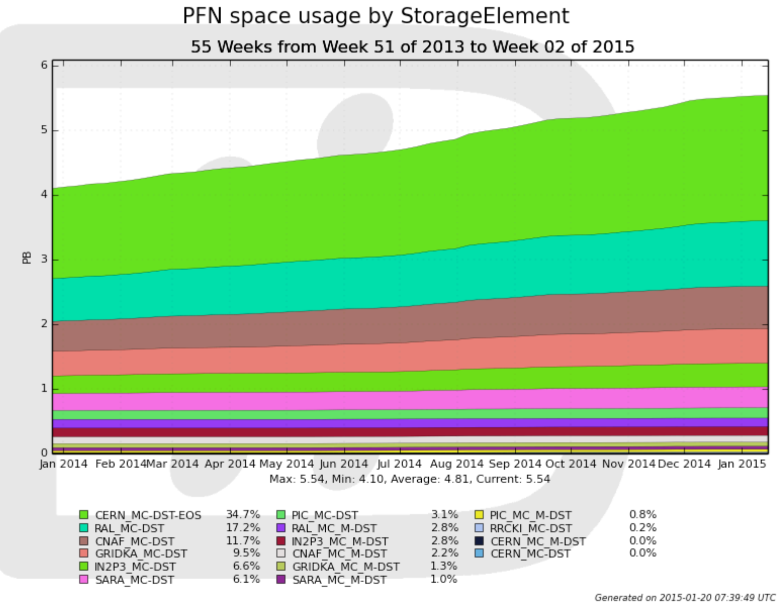
~3/4 at Tier0 + Tier1s
~1/4 outside

Resource usage: Disk

Tier0 + Tier1s: real DATA



Tier0 + Tier1s: MC simulation

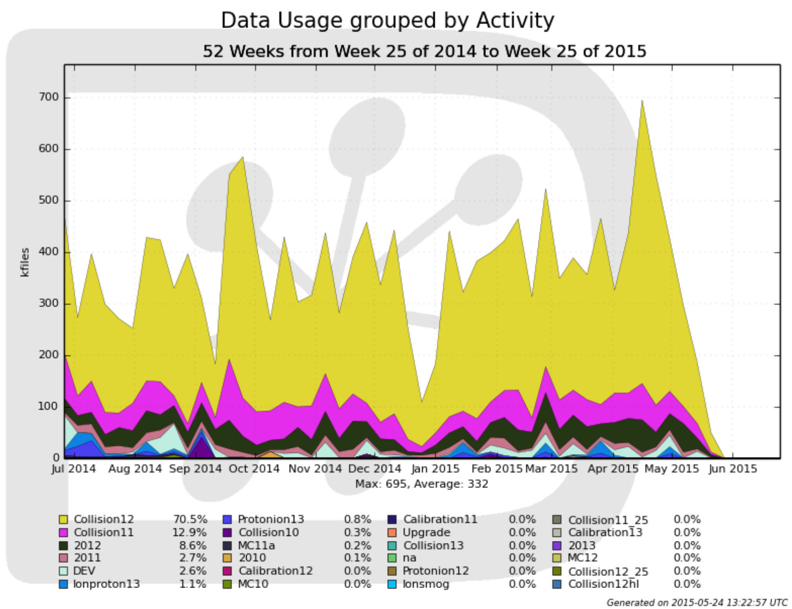


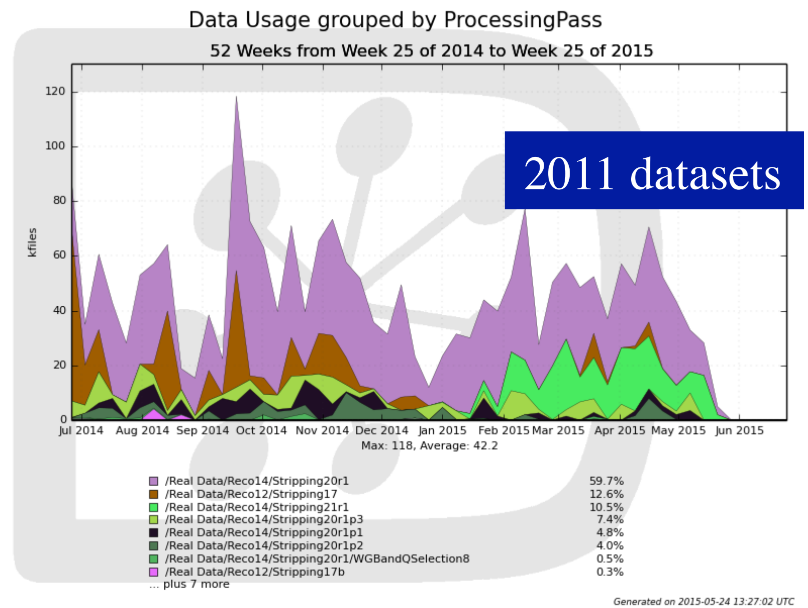
Cleanup after analysis on Data popularity

Data popularity

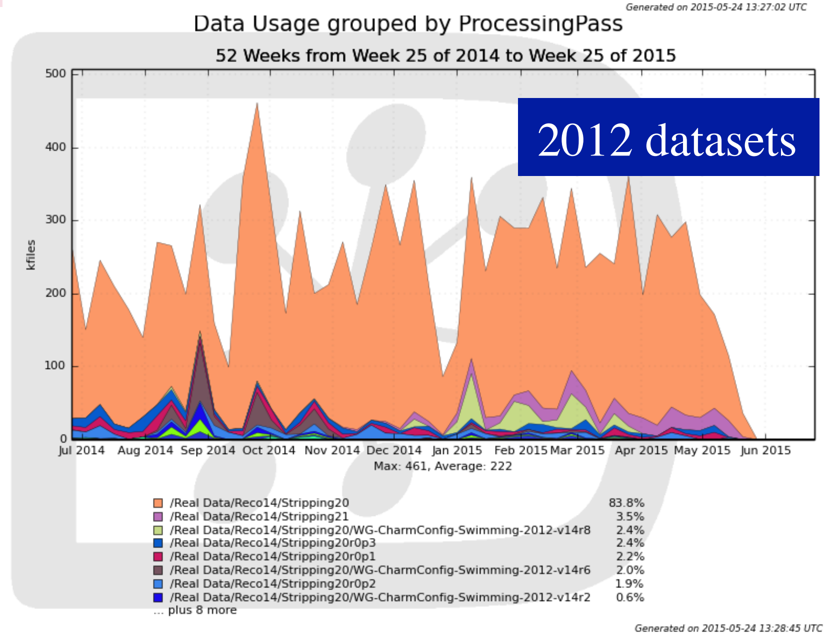
- Enabled recording of information as of May 2012
- Information recorded for each job:
  - Dataset (path)
  - Number of files for each job
  - Storage element used
- Allows currently by visual inspection to identify unused datasets
- Plan:
  - Establish, per dataset:
    - Last access date
    - Number of accesses in last (n) months (1<n<12)
    - Normalise number of dataset accesses to its size
  - Prepare summary tables per dataset
    - Access summary (above)
    - Storage usage at each site
  - Allow to trigger replica removal when space is required

Examples of popularity plots

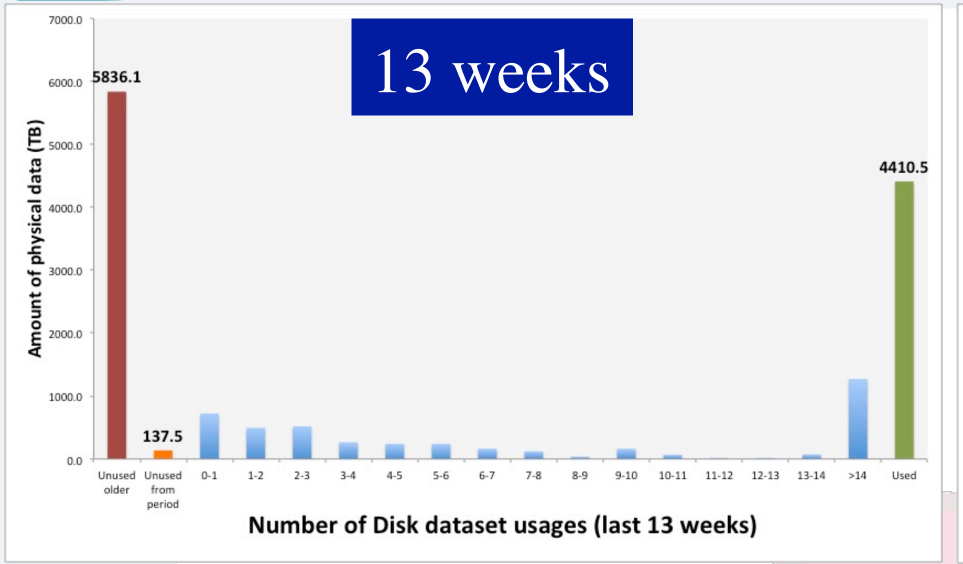


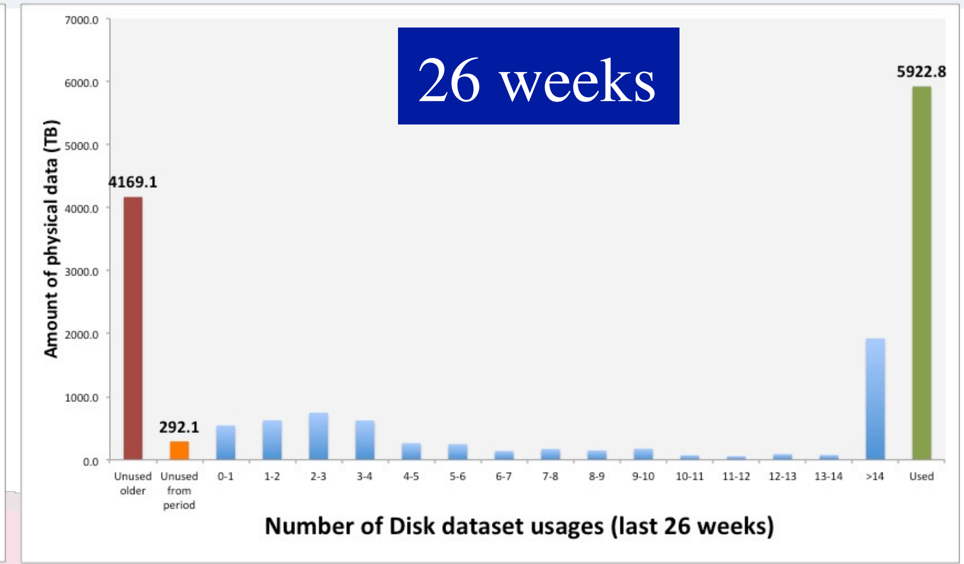


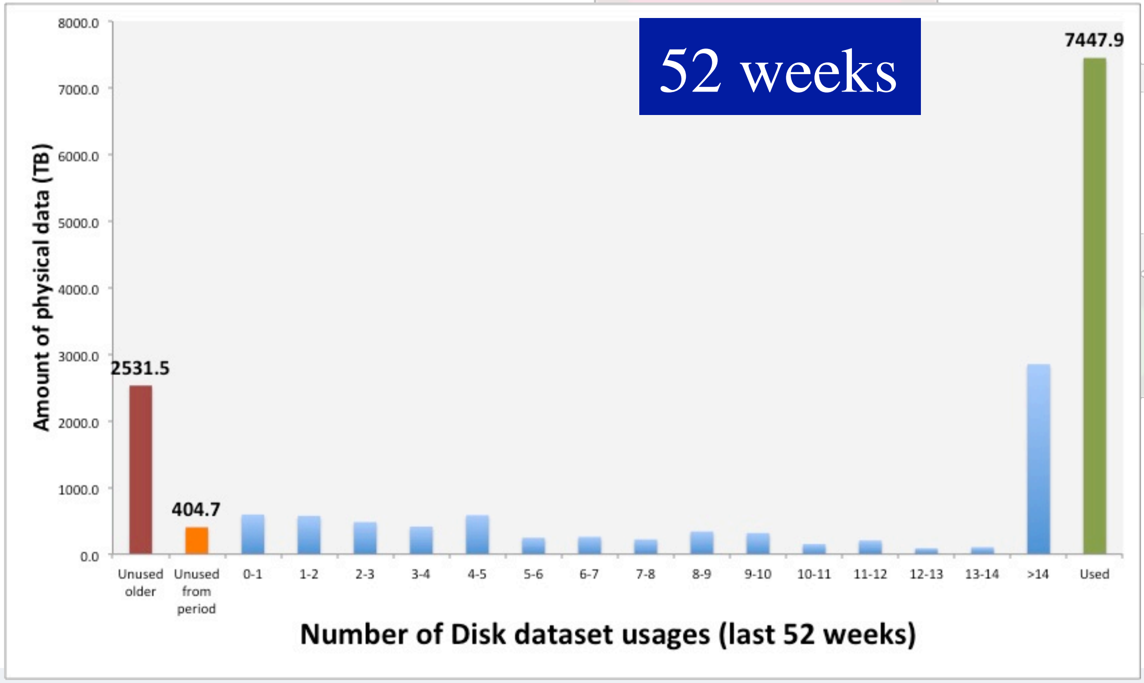
- We are also working on a classifier that, based on all metadata and popularity history, allows to classify datasets into those that are likely to be used within the next *n* weeks and those that are not.



Usage of datasets







~1PB disk space recovered by purging old data

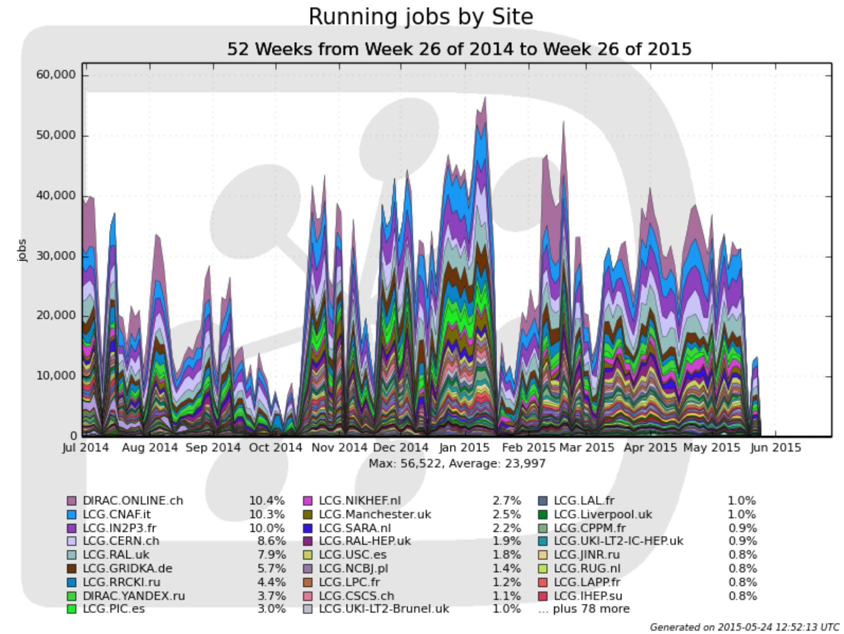
2015: suppression of reprocessing

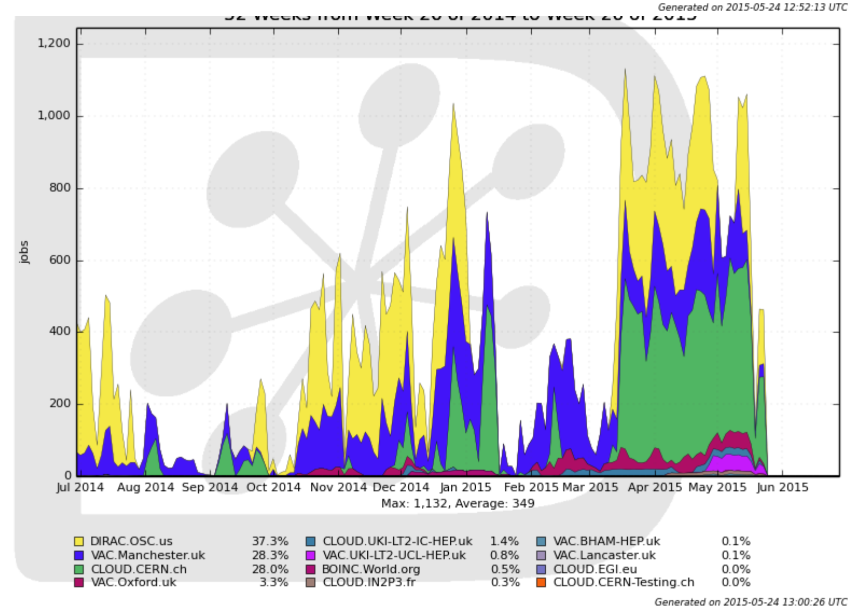
- During LS1, major redesign of LHCb HLT system
  - HLT1 (displaced vertices) will run in real time
  - HLT2 (physics selections) deferred by several hours
    - Run continuous calibration in the Online farm to allow use of calibrated PID information in HLT2 selections
    - HLT2 reconstruction becomes very similar to offline
- Automated validation of online calibration for use offline
  - Includes validation of alignment
  - Removes need for "first pass" reconstruction
- Green light from validation triggers 'final' reconstruction
  - Foresee up to two weeks' delay to allow correction of any problems flagged by automatic validation
  - No end of year reprocessing
    - Just restripping
- If insufficient resources, foresee to 'park' a fraction of the data for processing after the run
  - Unlikely to be needed before 2017 but commissioned from the start

Going beyond the Grid paradigm

DIRAC allows easy integration of non WLCG resources

- In 2014, ~10% of CPU resources from LHCb HLT and Yandex farms
- Vac infrastructure
  - Virtual machines created and contextualised for virtual organisations by remote resource providers
- Clouds
  - Virtual machines running on cloud infrastructures collecting jobs from the LHCb central task queue
- Volunteer computing
  - Use the BOINC infrastructure to enable payload execution on arbitrary compute resources





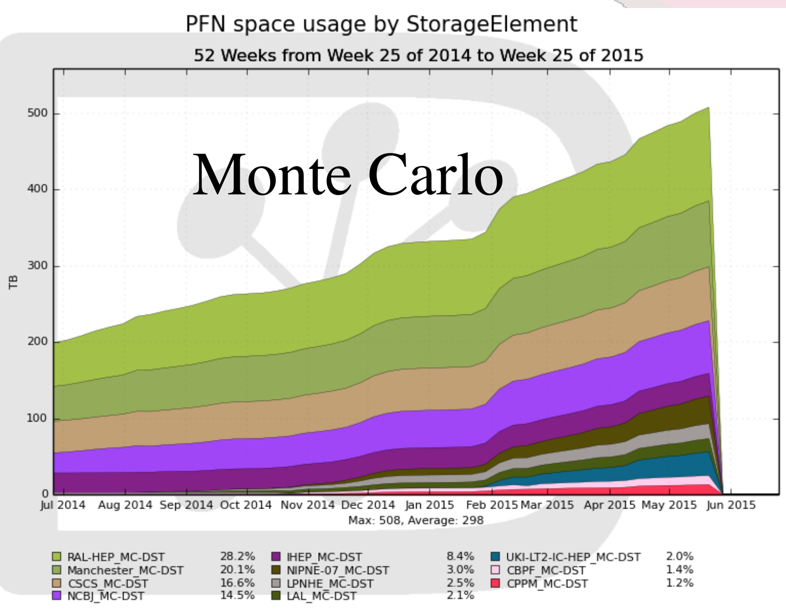
Changes to data management model

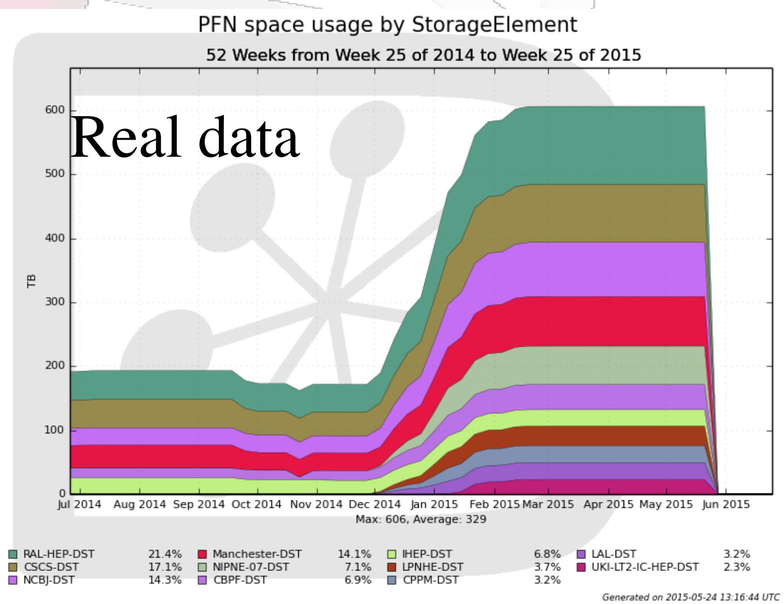
- Increases in trigger rate and expanded physics programme put strong pressure on storage resources

- Tape shortages mitigated by reduction in archive volume
  - Archives of all derived data exist as single tape copy
    - Forced to accept risk of data loss
  - Re-introduce a second tape copy in Run2, to cope with data preservation "obligations"
    - Re-generation in case of data loss is an operational nightmare and an overload of computing resources

- Disk shortages addressed by
  - Introduction of Disk at Tier 2
  - Reduction of event size in derived data formats
  - Changes to data replication and data placement policies
  - Measurement of data popularity to guide decisions on replica removals

- Tier2Ds are a limited set of Tier2 sites which are allowed to provide disk capacity for LHCb
  - Introduced in 2013 to circumvent shortfall of disk storage
    - To provide disk storage for physics analysis files (MC and data)
    - Run user analysis jobs on the data stored at the sites
- Blurs even more functional distinction between Tier1 and Tier2
  - A large Tier2D is a small Tier1 without Tape
- Status (Jan 18th 2015): 2.4 PB available, 0.83 PB used





Data Formats

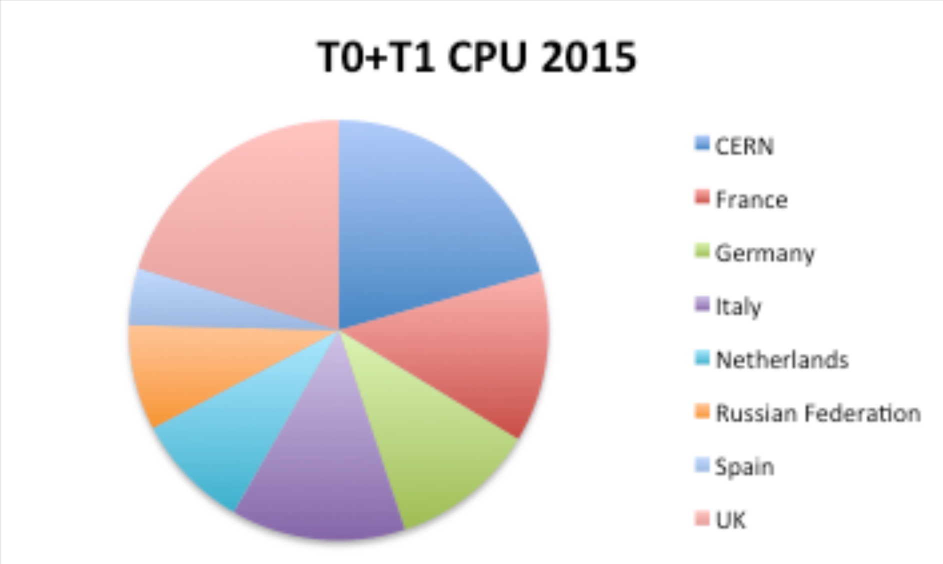
- Highly centralised LHCb data processing model allows to optimise data formats for operation efficiency

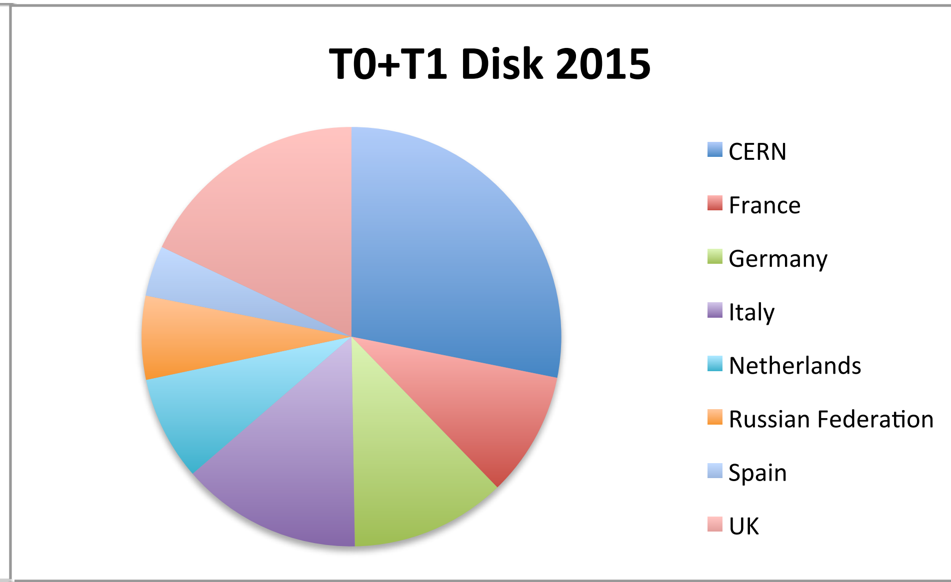
- Large shortfalls in disk and tape storage (due to larger trigger rates and expanded physics programme) drive efforts to reduce data formats for physics:
  - DST used by most analyses in 2010 (~120kB/event)
    - Contains copy of RAW and full Reco information
  - Strong drive to μDST (~13kB/event)
    - Save information for signal only
    - Suitable for most exclusive analyses, but many iterations required to get content correct
    - User-defined data can be added on demand (tagging, isolation,…)
  - "Legacy" stripping campaign of Run1 data just completed
    - Will allow to test μDST
    - MDST.DST == FULL.DST of all events passing a μDST stream. Temporary format (2015-2016) to allow regeneration of μDST in case of missing information without running the stripping again

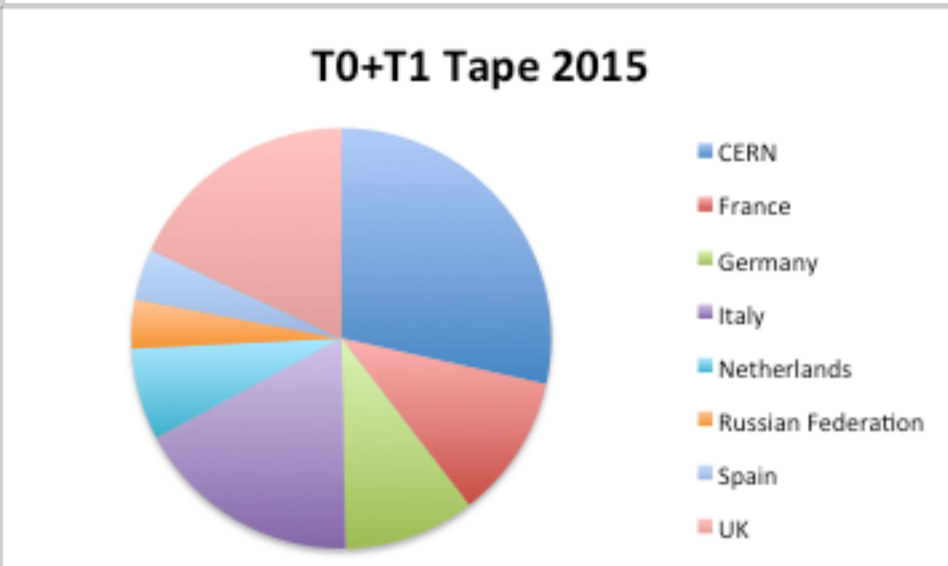
Data placement of (μ)DSTs

- Data-driven automatic replication
  - Archive systematically all analysis data (T1D0)
  - Real Data: 4 disk replicas, 1(→2) archives
  - MC: 3 disk replicas, 1(→2) archives
- Selection of disk replication sites:
  - Keep together whole runs (for real data)
    - Random choice per file for MC
  - Chose storage element depending on free space
    - Random choice, weighted by the free space
    - Should allow no disk saturation
      - Exponential fall-off of free space
      - As long as there are enough non-full sites!
- Removal of replicas
  - For processing n-1: reduce to 2 disk replicas (randomly)
    - Possibility to preferentially remove replicas from sites with less free space
  - For processing n-2: only keep archive replicas

Tier0 + Tier1 pledged resources in 2015

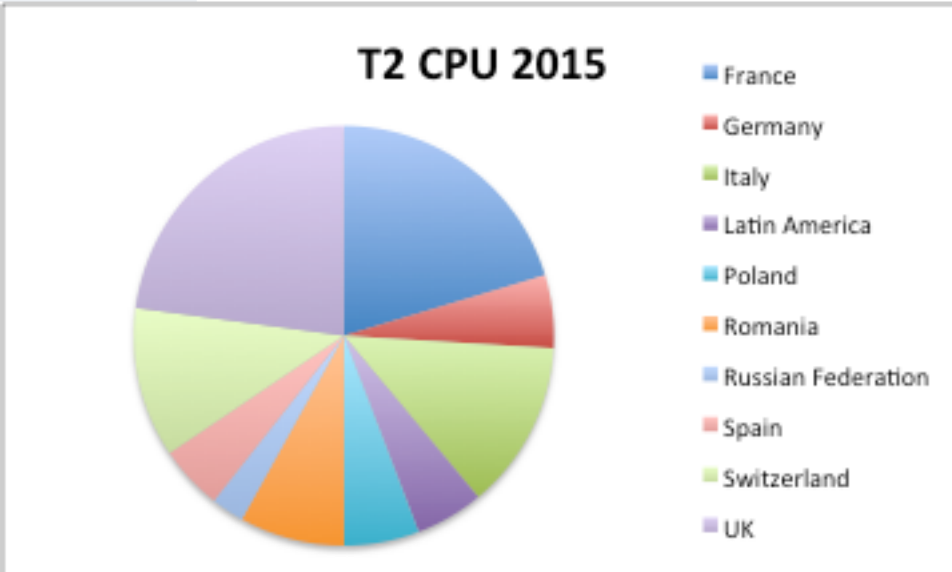


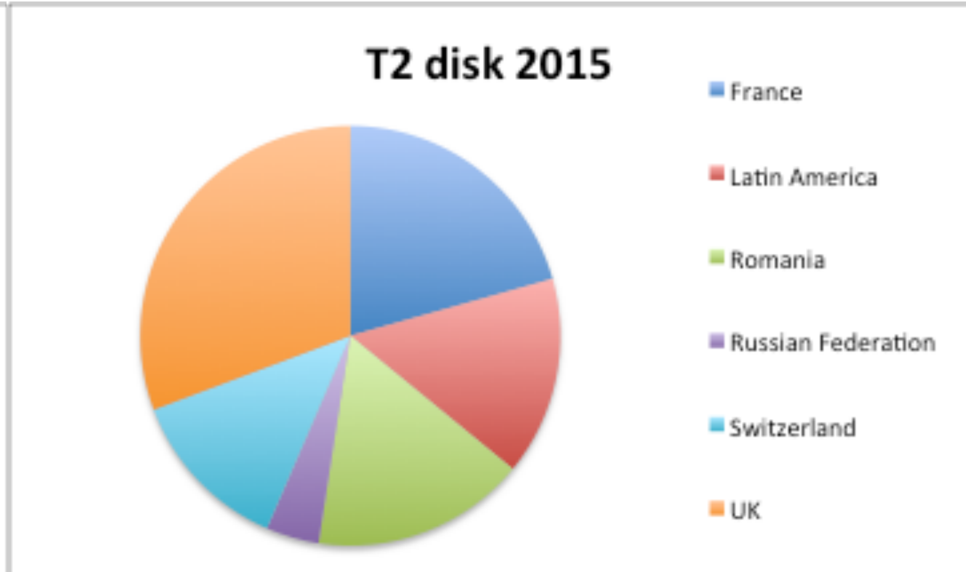




2015 T0+T1	CPU HS06	Disk Tbytes	Tape Tbytes
CERN	36000	5500	11200
France	23000	1880	4360
Germany	19600	2340	3960
Italy	23600	2720	6870
Netherlands	15661	1570	2773
Russian Federation	14200	1260	1480
Spain	7670	761	1541
UK	35400	3510	7110
Total	**175131**	**19541**	**39294**
Requested	**154000**	**17200**	**34900**
Difference	**13.7%**	**13.6%**	**12.6%**

Tier2 pledged resources in 2015





2015 Tier2	CPU HS06	Disk Tbytes
France	12323	404
Germany	3400	4
Italy	7875	0
Latin America	3183	300
Poland	3500	0
Romania	4900	323
Russian Federation	1539	80
Spain	3000	1
Switzerland	7000	250
UK	13861	602
Total	**60581**	**1964**
Requested	**66000**	**1900**
Difference	**-8.2%**	**3.4%**

Significant contribution from other sites not pledging resources to WLCG

- Yandex: 10000 HS06
- OSC: 2000 → 4000 HS06

Summary of 2015-2017 requests

Running assumptions

		LHC schedule		
Proton physics	LHC start date	01/05/2015	01/04/2016	01/04/2017
	LHC end date	31/10/2015	31/10/2016	15/12/2017
	LHC run days	183	213	258
	Fraction of days for physics	0.60	0.70	0.80
	LHC efficiency	0.32	0.39	0.39
	Approx. running seconds	3.0 10^6	5.0 10^6	7.0 10^6
Heavy Ion physics	Approx. running seconds	-	0.7 10^6	0.7 10^6

CPU

Power (kHS06)	Request 2015	Request 2016	Request 2017
Tier 0	36	51	62
Tier 1	118	156	191
Tier 2	66	88	107
Total WLCG	**220**	**295**	**360**
HLT farm	10	10	10
Yandex	10	10	10
Total non-WLCG	**20**	**20**	**20**
Grand total	**240**	**315**	**380**

Disk

Disk (PB)	**2015 Request**	**2016 Request**	**2017 Request**
Tier0	5.5	7.6	9.1
Tier1	11.7	13.5	15.0
Tier2	1.9	4.0	5.5
Total	**19.1**	**25.2**	**29.6**

Breakdown of CPU requests

pp Running

CPU Work in WLCG year (kHS06.years)	2015	2016	2017
Prompt Reconstruction	19	31	26
First pass Stripping	8	13	11
Full Restripping	8	20	11
Incremental Restripping	0	4	10
Simulation	134	153	207
VoBoxes and other services		4	4
User Analysis	17	20	24
Total Work (kHS06.years)	**186**	**246**	**293**
Efficiency corrected average power (kHS06)	**220**	**291**	**348**

HI Running

Resources for heavy ion running	2015 Request	2016 Request	2017 Request
CPU (kHS06)	0	24	32

Breakdown of DISK requests

pp Running

Disk storage usage forecast (PB)	2015	2016	2017
Stripped Real Data	7.3	13.1	15.3
Simulated Data	8.2	6.9	10.4
User Data	0.9	1.0	1.1
MDST.DST	1.5	1.9	0.0
RAW and other buffers	1.0	1.2	0.9
Other	0.2	0.2	0.2
Total	**19.1**	**24.3**	**27.9**

HI Running

Resources for heavy ion running	2015 Request	2016 Request	2017 Request
Disk (PB)	0	0.9	1.7

Tape

Tape (PB)	2015 Request	2016 Request	2017 Request
Tier0	11.2	20.6	30.9
Tier1	23.7	42.1	62.2
Total	**34.9**	**62.7**	**93.1**

pp Running

Tape storage usage forecast (PB)	2015	2016	2017
Raw Data	12.7	21.7	34.5
FULL.DST	8.7	15.2	20.7
MDST.DST	1.8	5.2	7.9
Archive – Operations	8.6	11.6	15.0
Archive – Data preservation	3.1	6.0	9.2
Total	**34.9**	**59.7**	**87.3**

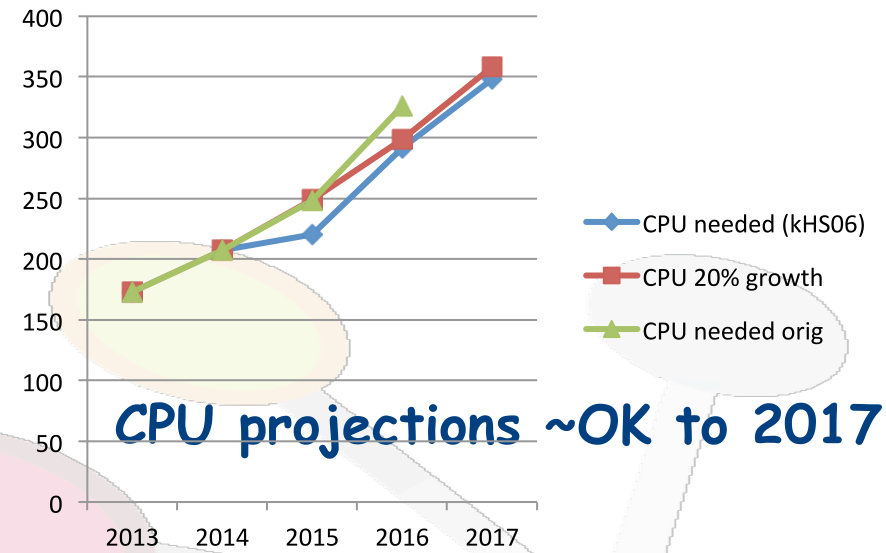
HI Running

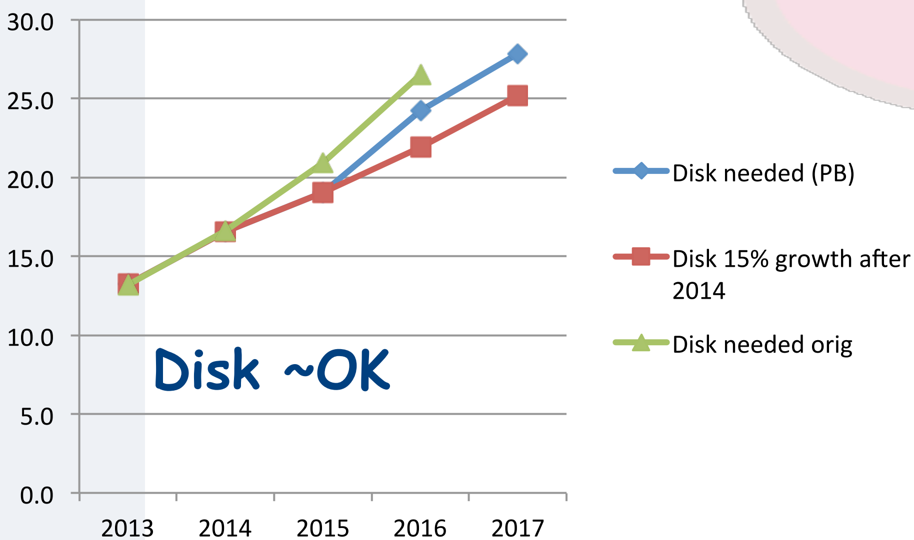
Tape (PB)	0	3.0	5.7

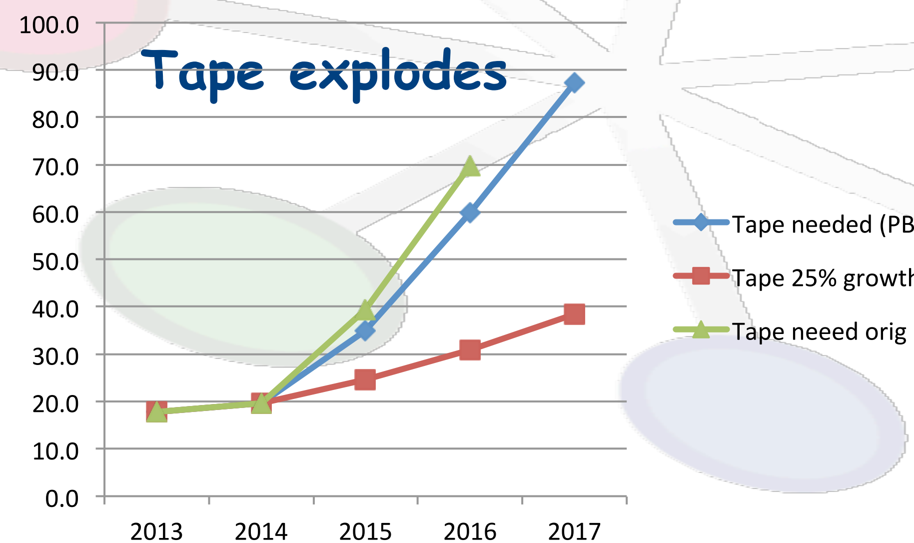
Please note:
WLCG estimates of tape costs include a 10% cache disk.
This is too large for our purposes.

Comparison with "flat budget"

- Definition of flat budget: same money will buy
  - 20% more CPUs
  - 15% more disk
  - 25% more tape







Mitigation strategies

- Increase in tape request is beyond flat-budget expectation
  - Ask resource providers for advance purchases in order to ease ramp-up
  - Trade some other resources for tape
    - But lever arm is short!

- Remove second tape copy of derived dataset?
  - Regeneration of even a small portion of data implies massive tape recalls and computing load, which might jeopardize other production activities

- Continue developing data popularity algorithms and data placement strategies
  - Potential significant savings on disk space

- Continue using available CPU resources "parasitically"

INFN requests: Tier1

- The resources requested for Tier1 CPU & disk are obtained by scaling the global requests with the INFN fraction: 21%
- For tape, we take the fraction with respect to countries hosting a Tier1: 27%
  - New entry: Russia (Kurchatov Institute, RRCKI)
- These fractions are taken from April 2015 RRB

  https://cds.cern.ch/record/2002286/files/CERN-RRB-2015-045.pdf
- This gives the following requests:

CNAF Tier1	2015 pledge	2016	Diff	Replace	tot	kEUR
CPU (kHS06)	23.6	32.5	8.9	?	8.9	**124**
DISK (TB)	2720	2811	91	?	91	**22**
TAPE (TB)	6870	8766	1896		1896	**47**
					Total:	**194**

INFN requests: Tier2

- The resources requested for Tier2 CPU are obtained by scaling the global requests with the INFN fraction: 21%
- This gives the following requests:

CNAF Tier2	2015 pledge	2016	Diff	Replace	tot	kEUR
CPU (kHS06)	7.88	18.3	10.4	?	10.4	**146**

INFN requests: Mitigation

- The increase of CPU is partly due to the trading of CPU for tape which was done for 2015
- We could do the same for 2016 and reduce CPU accordingly. For instance

CNAF Tier1	2015 pledge	2016	Diff	Replace	tot	kEUR
CPU (kHS06)	23.6	27.8	4.2	?	4.2	**59**
DISK (TB)	2720	2811	91	?	91	**22**
TAPE (TB)	6870	8766	1896		1896	**47**
					Total:	**128**

CNAF Tier2	2015 pledge	2016	Diff	Replace	tot	kEUR
CPU (kHS06)	7.88	15.7	7.8	?	7.8	**110**