The O2 project and future of ALICE computing

ALICE T1/T2 Workshop

Predrag Buncic

ALICE

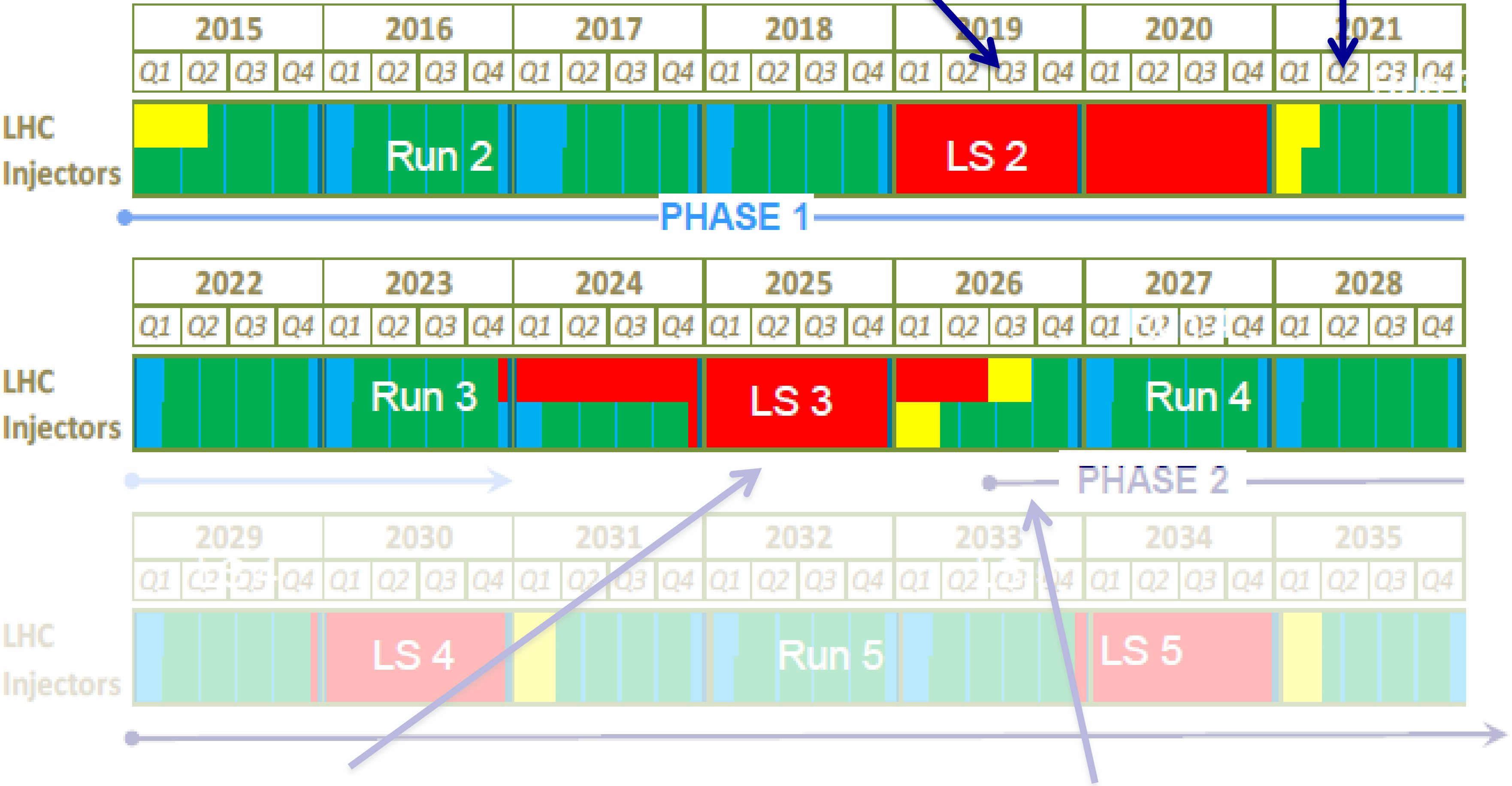
We are here

PHASE I Upgrade

ALICE, LHCb major upgrade

ATLAS, CMS ‚minor' upgrade

Heavy Ion Luminosity from 10^{27} to 7 x10^{27}

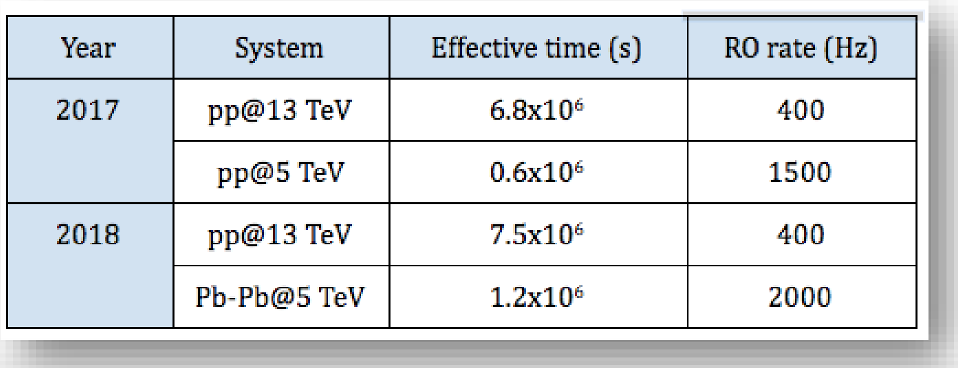


PHASE II Upgrade

ATLAS, CMS major upgrade

HL-LHC, pp luminosity from 10^{34} (peak) to 5 x10^{34} (levelled)

Data taking plans 2017-2019



Year	System	Effective time (s)	RO rate (Hz)
2017	pp@13 TeV	6.8x10^{6}	400
	pp@5 TeV	0.6x10^{6}	1500
2018	pp@13 TeV	7.5x10^{6}	400
	Pb-Pb@5 TeV	1.2x10^{6}	2000

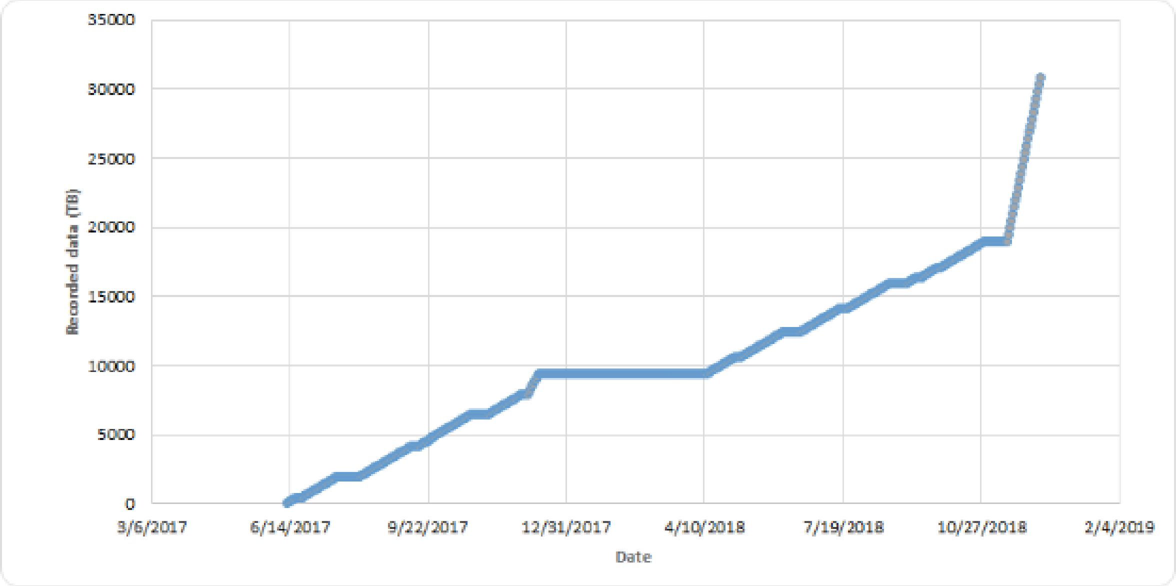
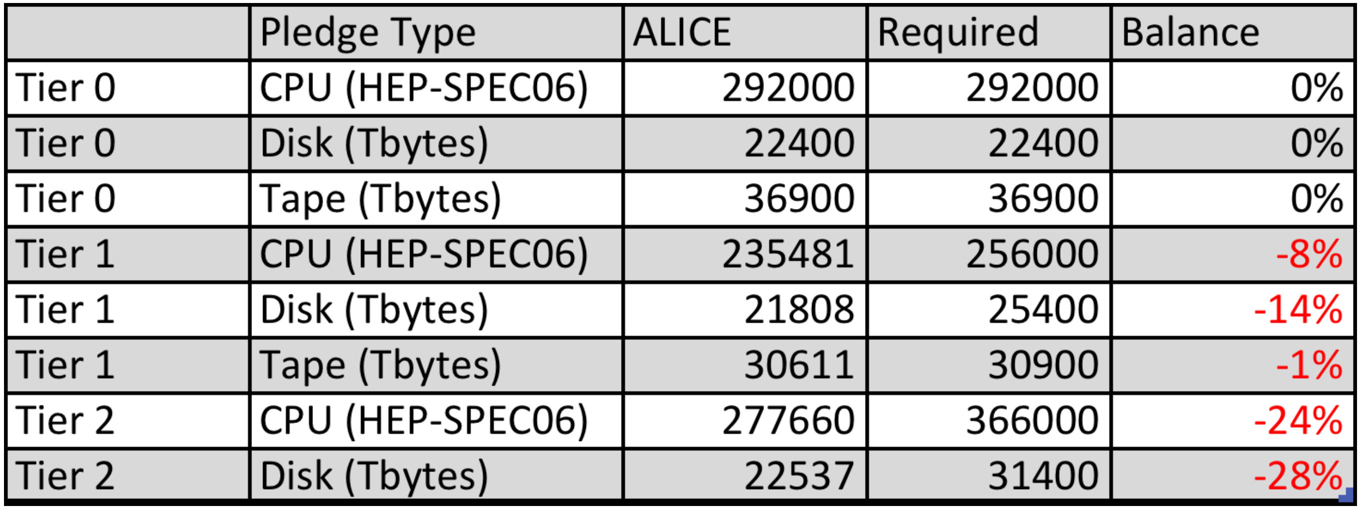


Figure 1: A graphical representation of the expected raw data accumulation based on the 2017-2018 data taking scenario

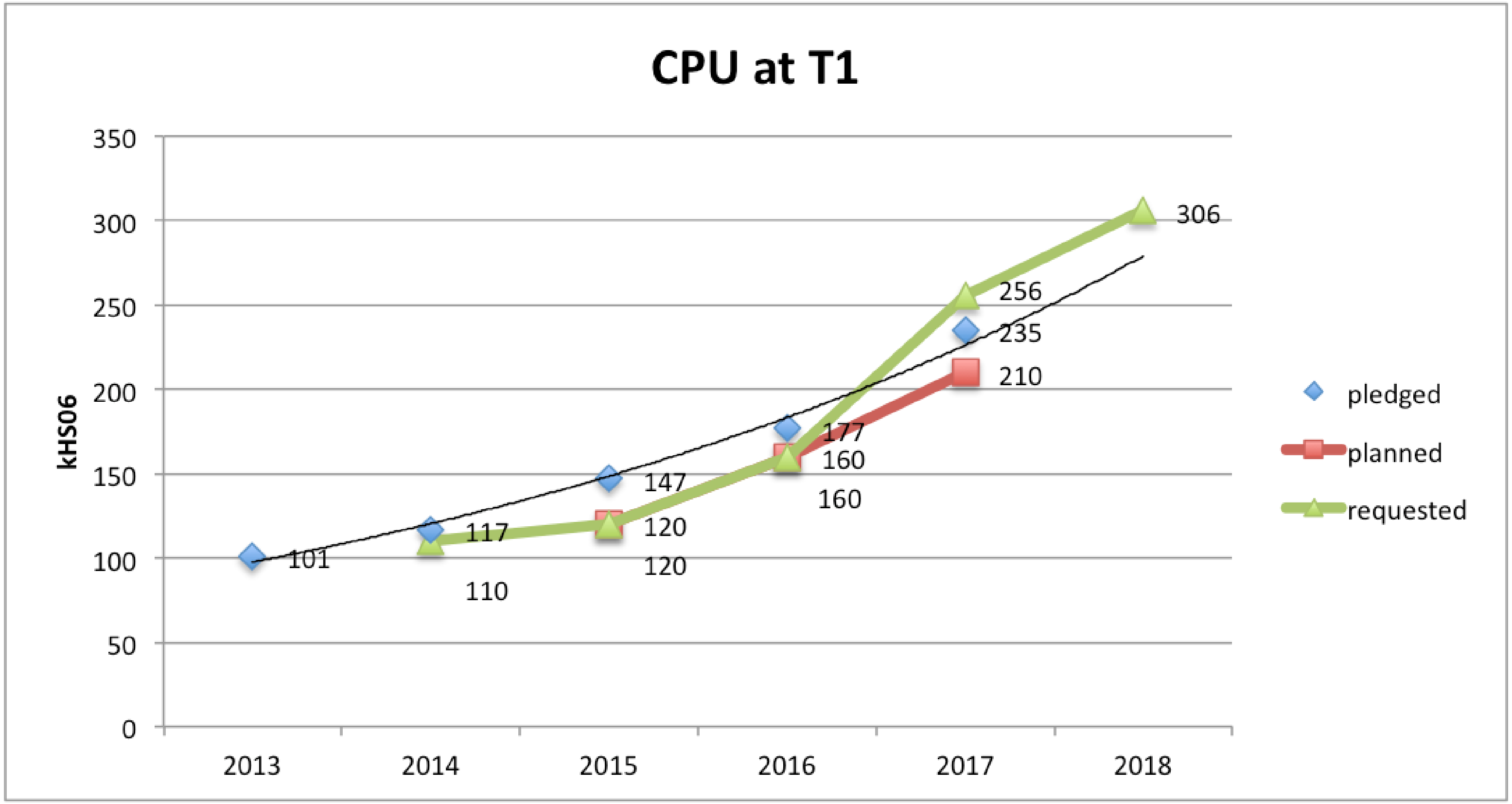
C-RSG Exercise and “flat budget”



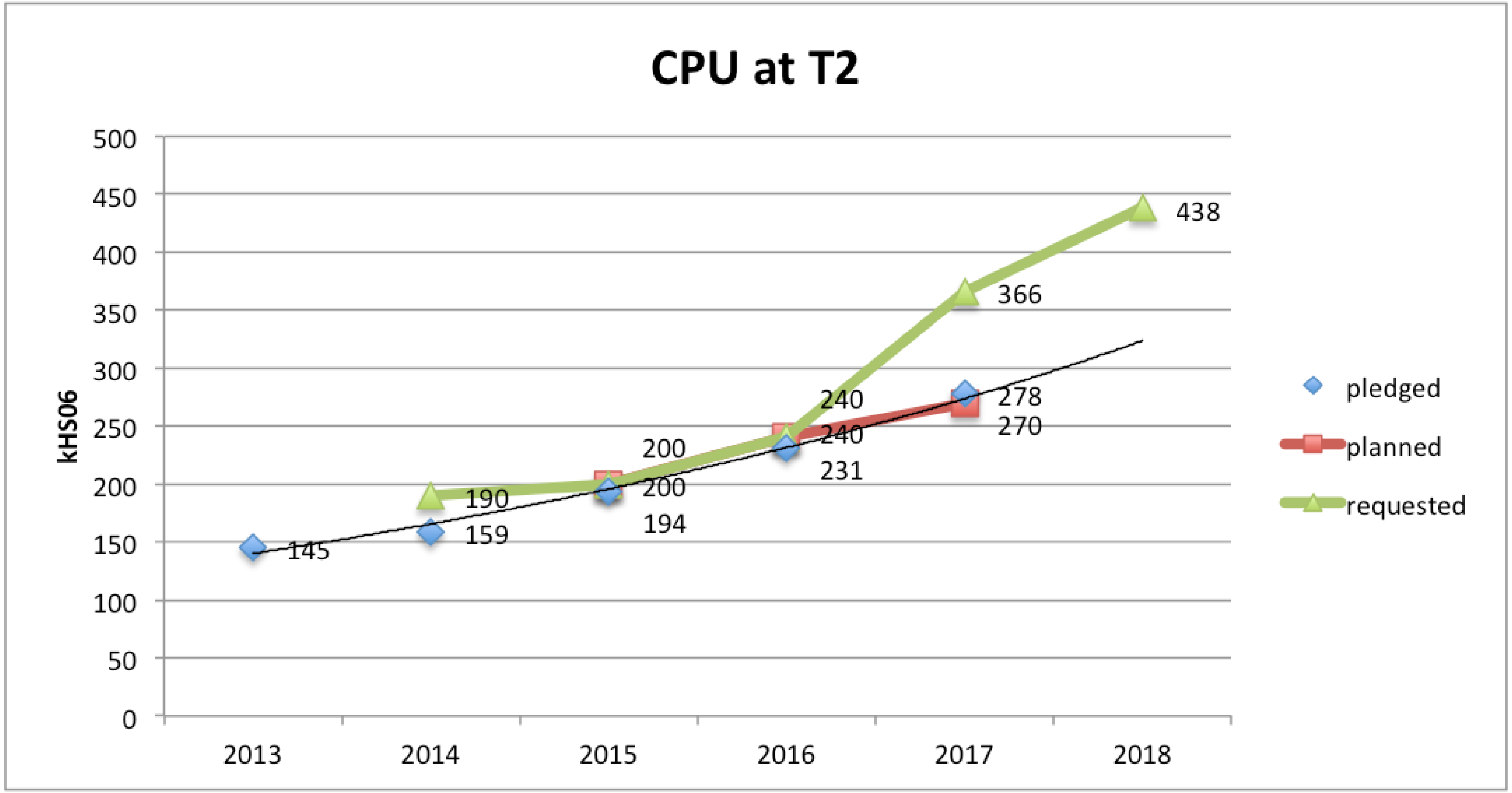
	Pledge Type	ALICE	Required	Balance
Tier 0	CPU (HEP-SPEC06)	292000	292000	0%
Tier 0	Disk (Tbytes)	22400	22400	0%
Tier 0	Tape (Tbytes)	36900	36900	0%
Tier 1	CPU (HEP-SPEC06)	235481	256000	-8%
Tier 1	Disk (Tbytes)	21808	25400	-14%
Tier 1	Tape (Tbytes)	30611	30900	-1%
Tier 2	CPU (HEP-SPEC06)	277660	366000	-24%
Tier 2	Disk (Tbytes)	22537	31400	-28%

“The requirements for 2018, specifically: T1-disk (47% increase), T2-disk (82%) and T2-cpu (58%) are well beyond "flat budget" and we noticed that T2 pledges are about 20% less than the request. CRSG is not in a position to accept these requests and asks the experiment and the LHCC to take the actions necessary to reduce the resource needs in order to be scrutinized in October. CRSG is willing to follow and to help in this process if asked.”

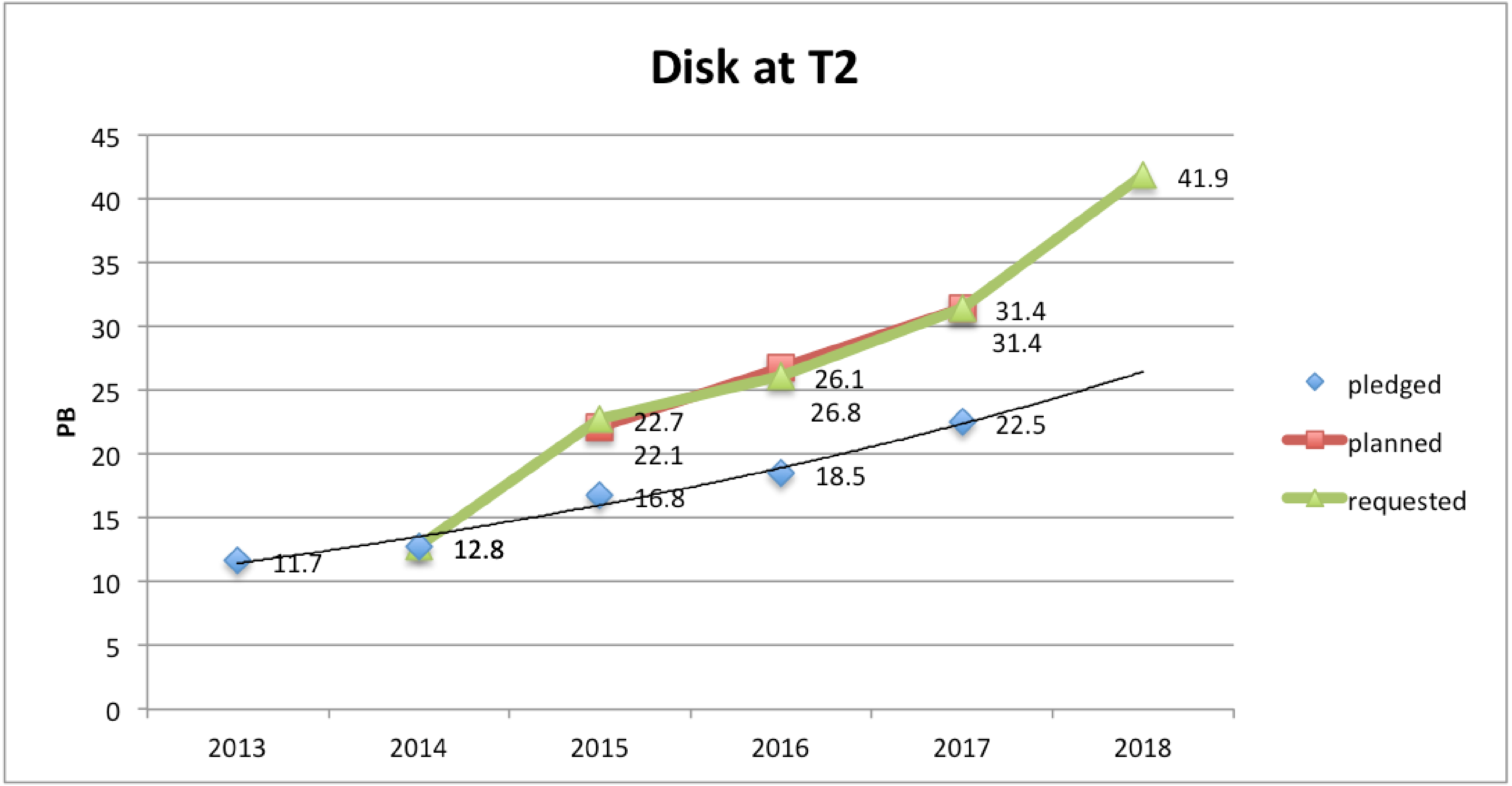
ALICE: Plans, requests and pledges



ALICE: Plans, requests and pledges



ALICE: Plans, requests and pledges



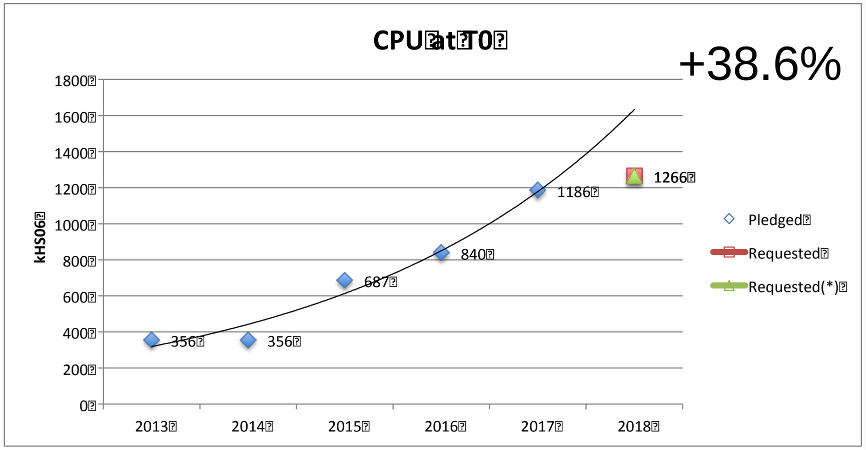
Projection 2017-2020 assuming 20% growth

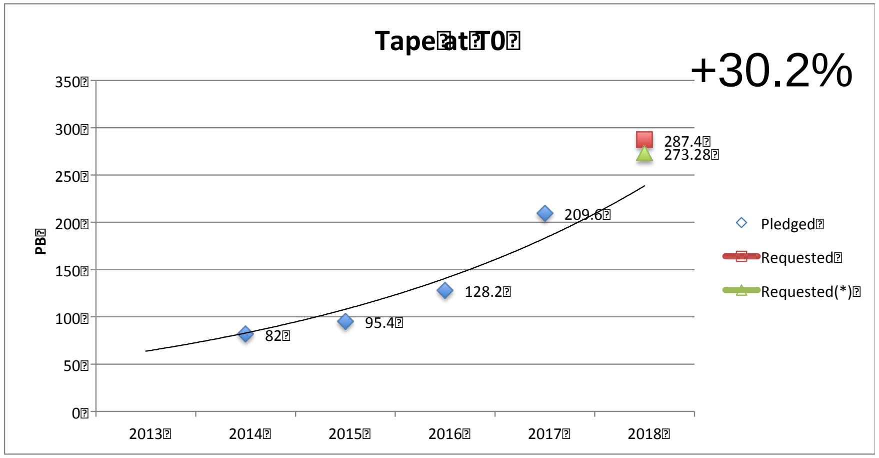
		CPU			Disk			Tape		
	Tier	requested	pledged	difference	requested	pledged	difference	requested	pledged	difference
2017	0	292.00	292.00	0.00%	20.60	22.40	8.04%	34.5	36.9	6.50%
	1	256.00	235.48	-8.71%	24.40	21.81	-11.90%	26.60	30.60	13.07%
	2	366.00	318.62	-14.87%	31.20	26.01	-19.95%			
	Total	914.00	846.10	-8.03%	76.20	70.22	-8.52%	61.10	67.50	9.48%
2018	0	350.00	350.40	0.11%	27.00	26.88	-0.45%	55	55	
	1	306.00	282.58	-8.29%	32.00	26.17	-22.30%	41.00	41.00	
	2	438.00	382.34	-14.56%	41.00	31.21	-31.35%			
	Total	1094.00	1015.32	-7.75%	100.00	84.26	-18.68%	96.00	96.00	0.00%
2019	0	534.00	420.48	-27.00%	33.60	32.26	-4.17%	55	55	
	1	501.00	339.09	-47.75%	39.90	31.40	-27.07%	49.50	49.50	
	2	635.00	458.81	-38.40%	51.10	37.46	-36.42%			
	Total	1670.00	1218.38	-37.07%	124.60	101.11	-23.23%	104.50	104.50	0.00%
2020	0	534.00	504.58	-5.83%	33.60	38.71	13.19%	55	55	
	1	501.00	406.91	-23.12%	39.90	37.68	-5.89%	49.50	49.50	
	2	635.00	550.57	-15.34%	51.10	44.95	-13.69%			
	Total	1670.00	1462.05	-14.22%	124.60	121.33	-2.69%	104.50	104.50	0.00%

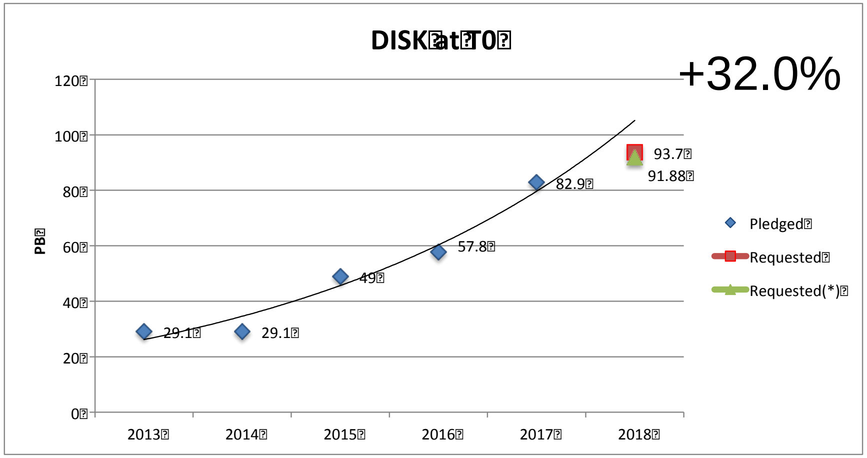
What next?

- To restore the expected resource growth, we must insist and follow-up on the installation of the **pledged** resources:
  - The deficit of disk in the T2 centres, to the **approved** levels for 2016 must be installed;
  - T1 and T2 **pledges** correspond to the **required** and approved resources levels in 2017. Presently, there is a **pledges** deficit of 8% and 24% (CPU) and 14% and 28% (disk) at the T1s and T2s respectively;
  - The computing resource sharing between countries and FAs should be respected and enforced
- We will calculate 2018 shares between the countries participating in ALICE based on our actual request and not on C-RSG approved resources.
- The additional resources would have to be over-pledged or provided as unpledged contribution, which however will be accounted by ALICE.

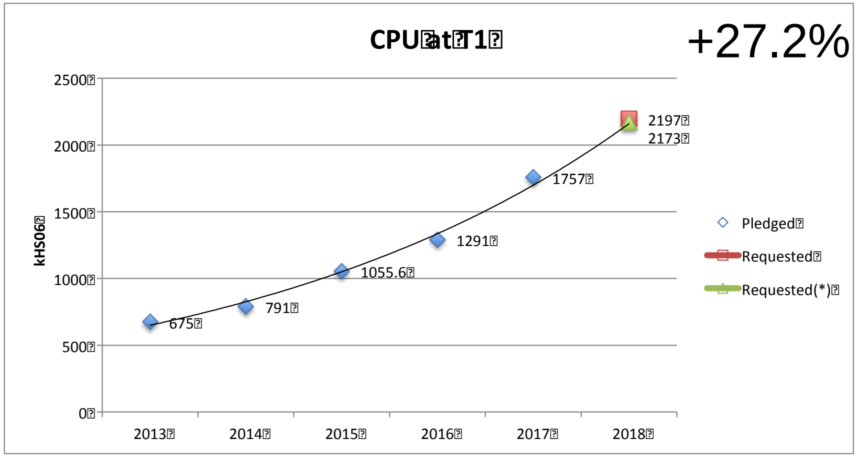
Flat budget?


CPU at T0
+38.6%
1800
1600
1400
1200
1000
800
600
400
200
0
kHS06
356
356
687
840
1186
1266
2013
2014
2015
2016
2017
2018
Pledged
Requested
Requested(*)

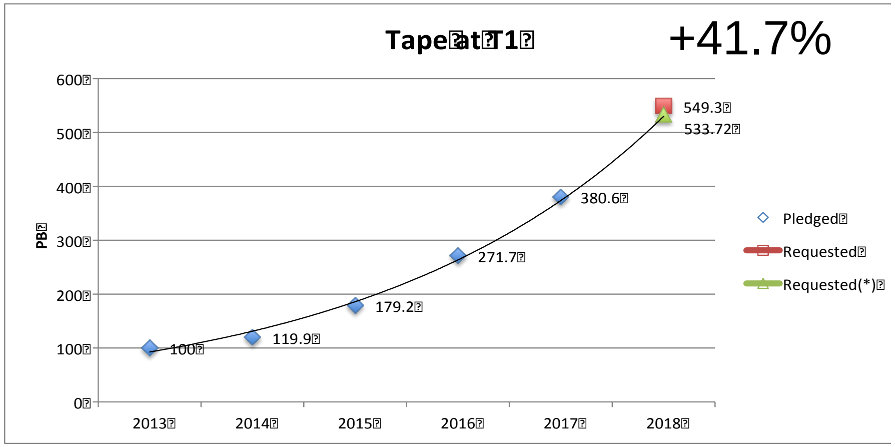

Tape at T0
+30.2%
350
300
250
200
150
100
50
0
PB
82
95.4
128.2
209.6
287.4
273.28
2013
2014
2015
2016
2017
2018
Pledged
Requested
Requested(*)


DISK at T0
+32.0%
120
100
80
60
40
20
0
PB
29.1
29.1
49
57.8
82.9
93.7
91.88
2013
2014
2015
2016
2017
2018
Pledged
Requested
Requested(*)

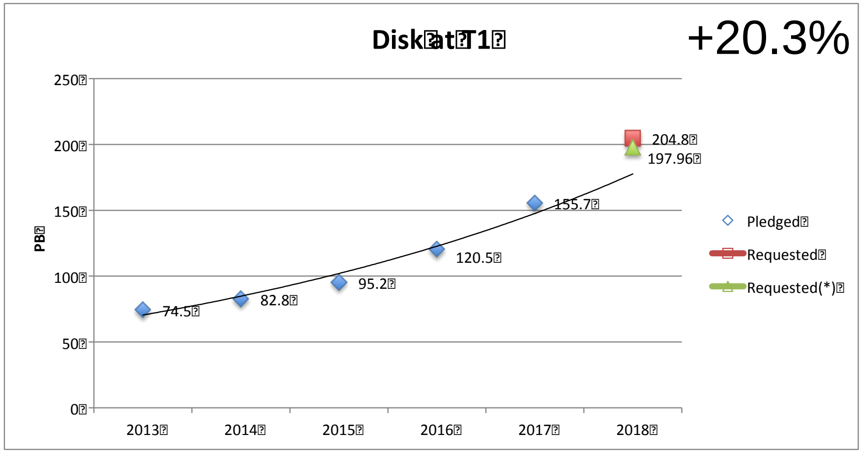
Flat budget?



+27.2%

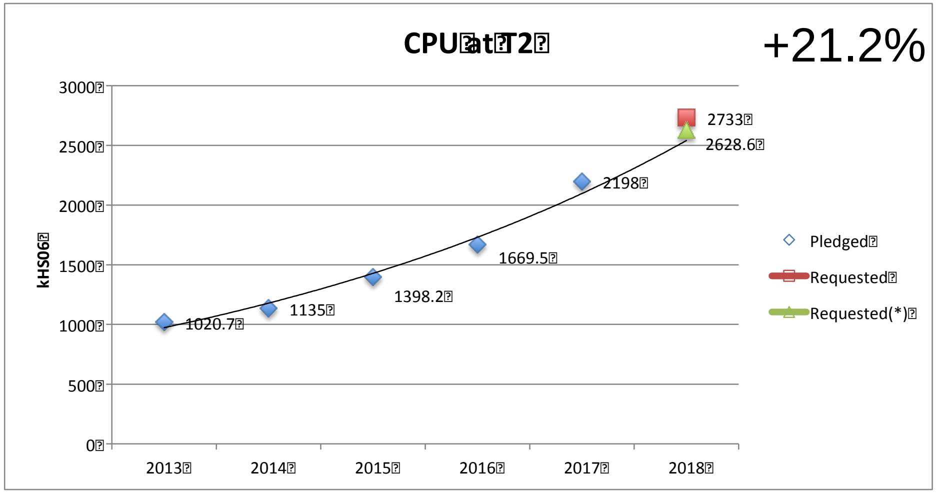


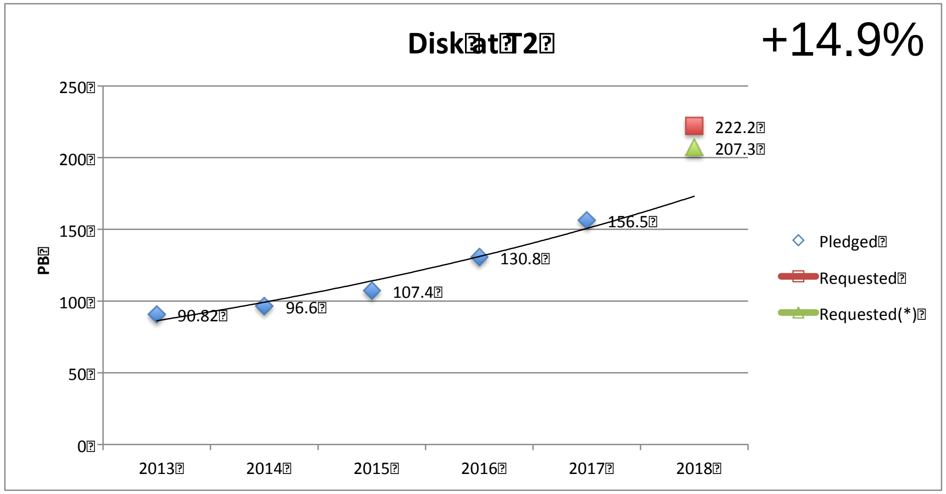
+41.7%



+20.3%

Flat budget?


CPU at T2
+21.2%
kHS06
3000
2500
2000
1500
1000
500
0
2013
2014
2015
2016
2017
2018
1020.7
1135
1398.2
1669.5
2198
2733
2628.6
Pledged
Requested
Requested(*)


Disk at T2
+14.9%
PB
250
200
150
100
50
0
2013
2014
2015
2016
2017
2018
90.82
96.6
107.4
130.8
156.5
222.2
207.3
Pledged
Requested
Requested(*)

- It is totally unclear what flat budget means on different tiers

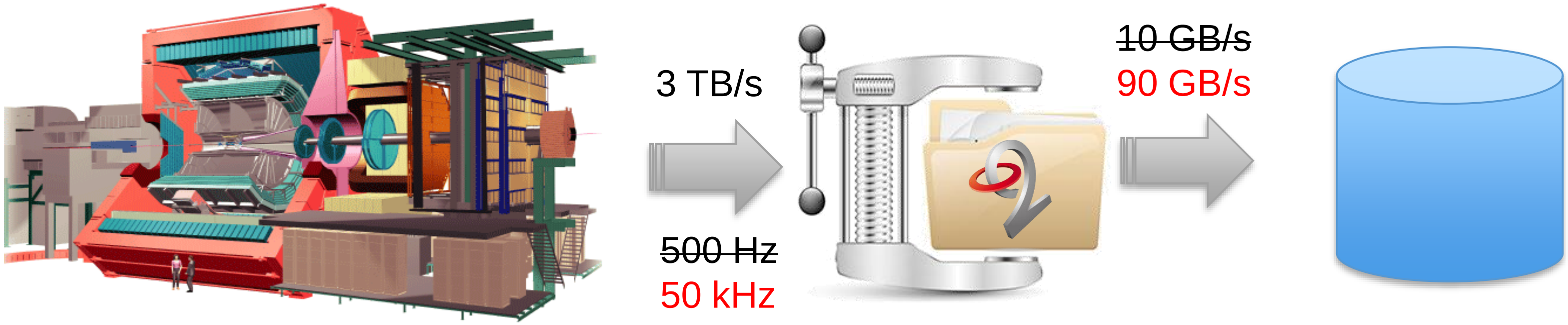
- However, we are forced to make our plans for Run 3 assuming that flat budget means

  - maximim 20% year on year growth
  - applied equally on all resources categories
  - applied equally to all Tiers

Run 3 data taking objectives

- For Pb-Pb collisions:
  - Reach the target of ~~1~~ 13 nb^{-1} integrated luminosity in Pb-Pb for rare triggers.

- The resulting data throughput from the detector has been estimated to be greater than 1TB/s for Pb–Pb events, roughly two orders of magnitude more than in Run 1



3 TB/s

~~500 Hz~~

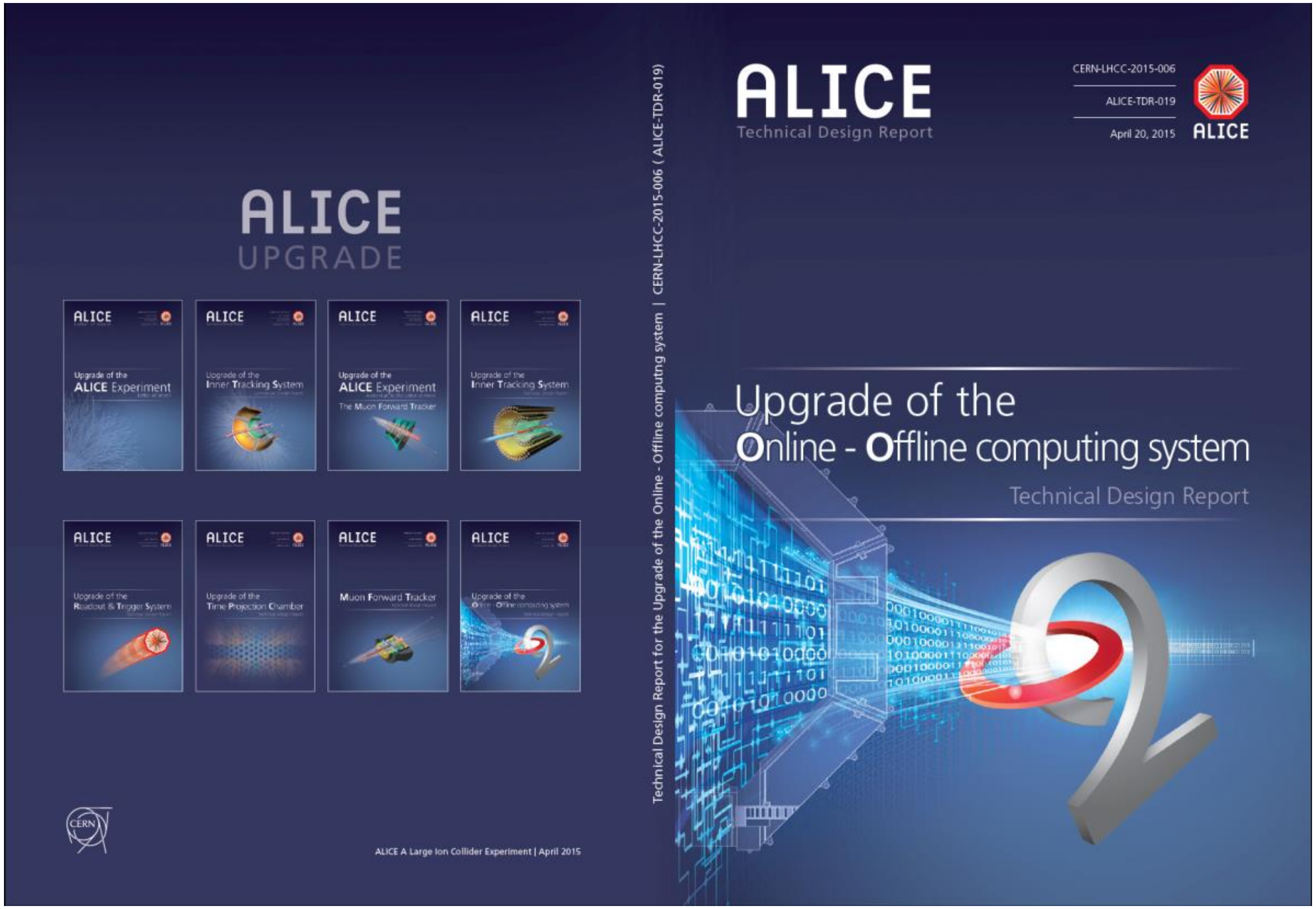
50 kHz

~~10 GB/s~~

90 GB/s

Factor 90 in terms of Pb-Pb events (x 30 for pp)

ALICE O^2 Technical Design Report



ALICE
UPGRADE

ALICE A Large Ion Collider Experiment | April 2015

Technical Design Report for the Upgrade of the Online - Offline computing system | CERN-LHCC-2015-006 (ALICE-TDR-019)

ALICE
Technical Design Report

CERN-LHCC-2015-006
ALICE-TDR-019
April 20, 2015

Upgrade of the
Online - Offline computing system
Technical Design Report

CERN
Upgrade of the ALICE Experiment
Upgrade of the Inner Tracking System
Upgrade of the ALICE Experiment
The Muon Forward Tracker
Upgrade of the Inner Tracking System
Upgrade of the Readout & Trigger System
Upgrade of the Time Projection Chamber
Muon Forward Tracker
Upgrade of the Online - Offline computing system

https://cds.cern.ch/record/2011297/files/ALICE-TDR-019.pdf

O2 Facility

+ 463 FPGAs
  - Detector readout and fast cluster finder
+ 100’000 CPU cores
  - To compress 1.1 TB/s data stream by overall factor 14
+ 3000 GPUs
  - To speed up the reconstruction
  - 3 CPU[1)] + 1 GPU[2)] = 28 CPUs
+ 60 PB of disk
  - To buy us an extra time and allow more precise calibration

----

= Considerable (but heterogeneous) computing capacity that will be used for Online and Offline tasks

  ✧ Identical s/w should work in Online and Offline environments

Containter@P2 or Prevessin Computing Centre (PCC)?

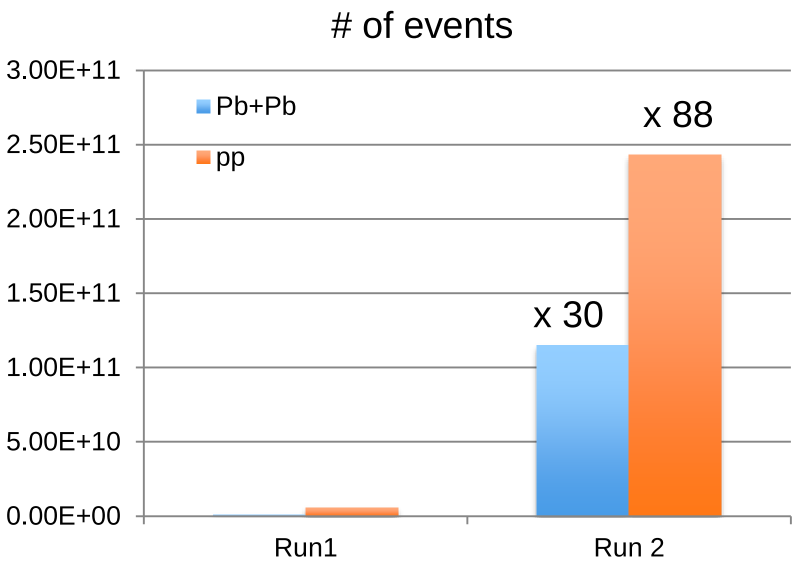
- Original plan: Commercial Container Data Center
  - 2015-16 Market Survey → Invitation Tender → Purchase lab (20-30 % of the total capacity)
  - 2018 Move lab to P2 and implement first slice of CR0 (10% EPN and DS)
  - 2020 Purchase CR0 addition (70-80 %) and full deployment

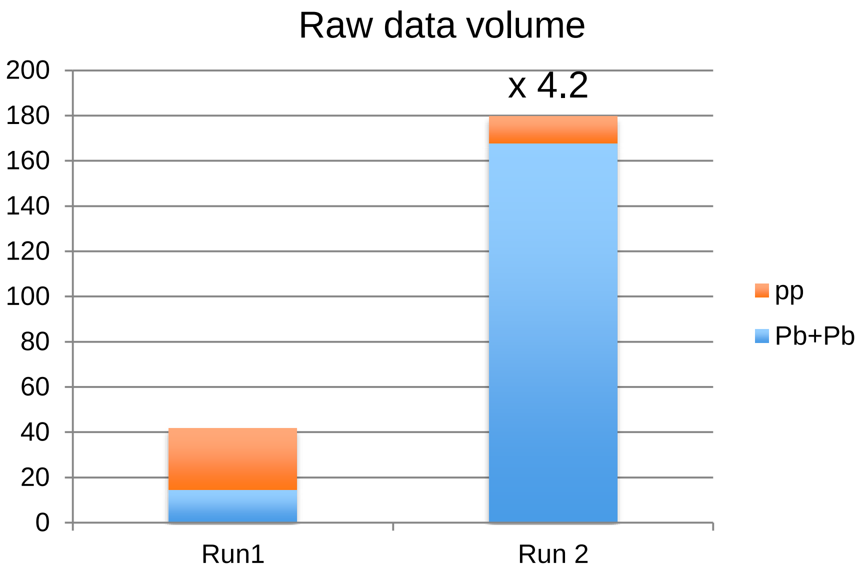
- We will know the answer in June…

Computing Model in Run 1&2

- Computing model built on top of WLCG services
  - Any job can run anywhere and access data from any place
  - Scheduling takes care of optimizations and brings "computing to data"
  - Every file and its replicas are accounted in the file catalogue
- Worked (surprisingly) well during Run 2 and Run 2

Run 2 vs Run 3: Data volume and # of events



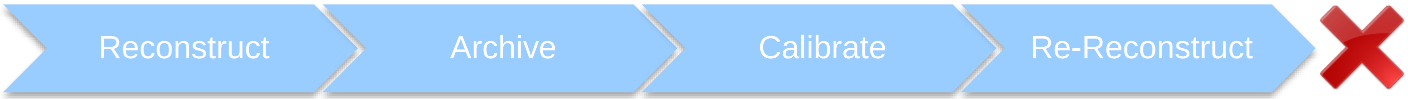


- While event statistics will increase by factor 30in Pb-Pb (x88 in pp), data volume will increase by factor 4.2
  - Thanks to data reduction in O2 facility
    - Online tracking that allows rejection of clusters not associated with tracks
    - Large effect in case of pileup (pp)

Data management

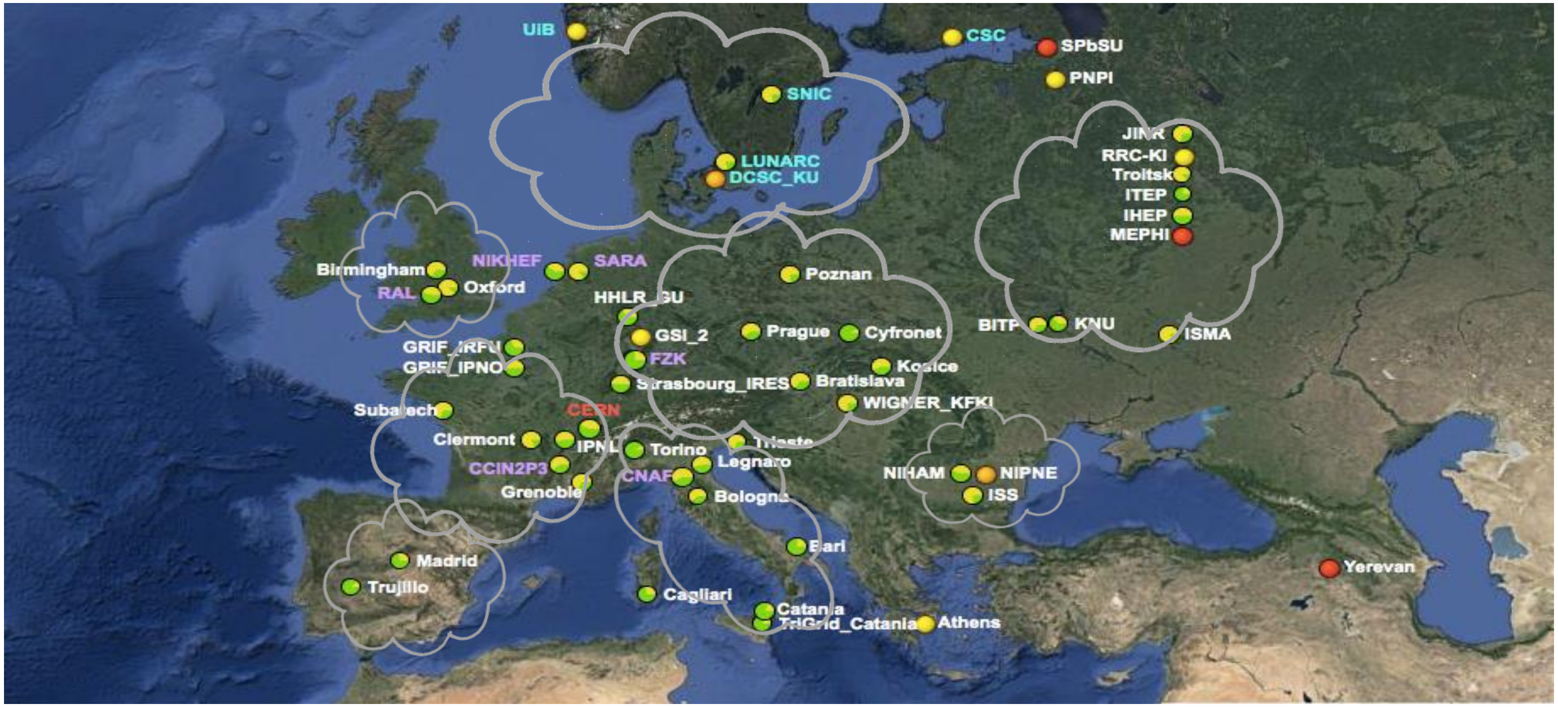
- Only one instance of each raw data file (CTF) stored on disk with a backup on tape
  - In case of data loss, we will restore lost files form the tape
  - O2 disk buffer should be sufficient accommodate CTF data from the entire period.
  - As soon as it is available, the CTF data will be archived to the Tier 0 tape buffer or moved to the Tier 1s

- All other intermediate data created at various processing stages is transient (removed after a given processing step) or temporary (with limited lifetime)
  - Only CTF and AODs are archived kept on disk to tape



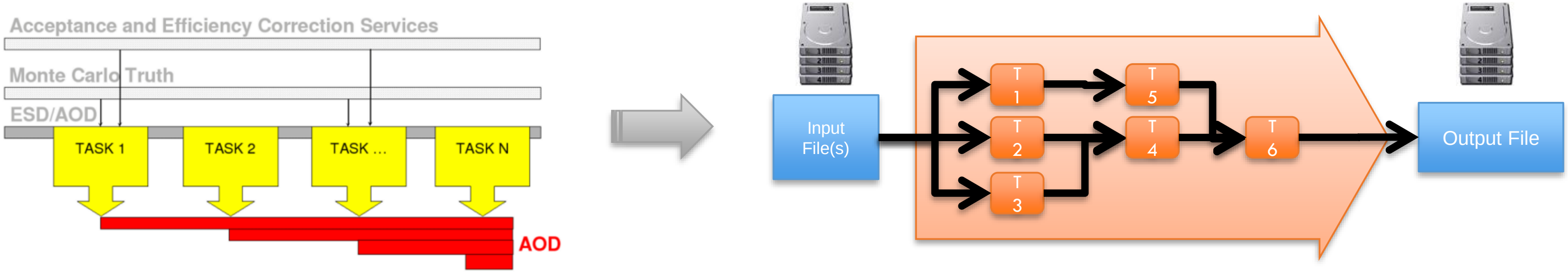
- Given the limited size of the disk buffers in O2 and Tier 1s, all CTF data collected in the previous year, will have to be removed before new data taking period starts.

Complexity management



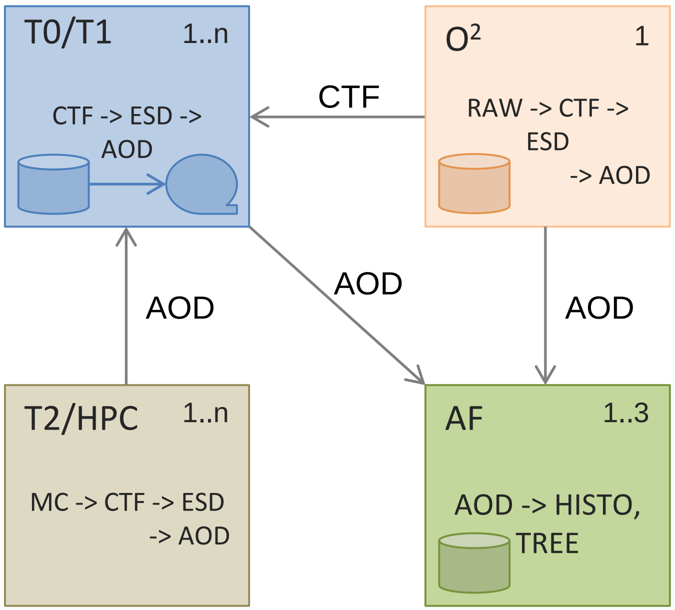
- We need to transform (logically) 100s of individual sites to 10s of clouds/regions
- Each cloud/region should provide reliable data management and sufficient processing capability to simplify scheduling and high level data management

Analysis Facilities



- Motivation
  - Analysis remains /O bound in spite of attempts to make it more efficient by using the train approach
- Solution
  - Collect AODs on a dedicated sites that are optimized for fast processing of a large local datasets
  - Run organized analysis on local data like we do today on the Grid
  - Requires 20-30’000 cores and 5-10 PB of disk on very performant file system
  - Such sites can be elected between the existing T1s (or even T2s) but ideally this would be a purpose build facility optimized for such workflow

Run 3 Computing Model

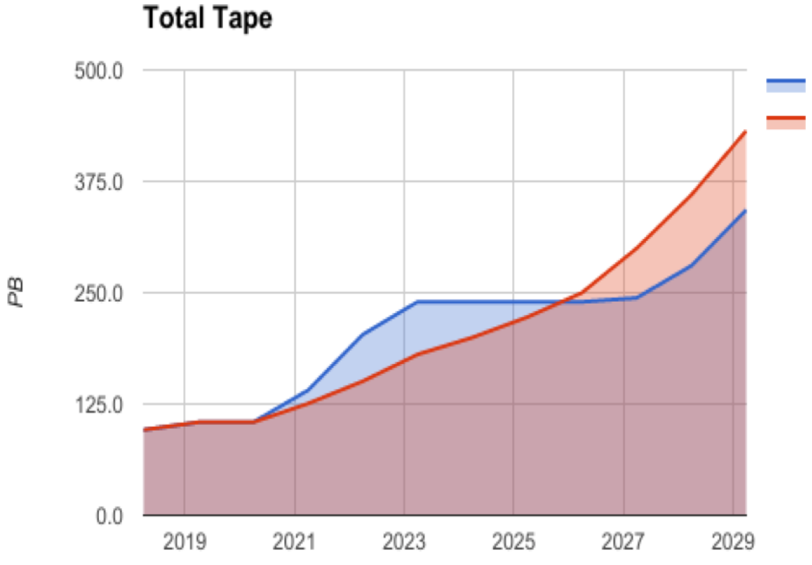


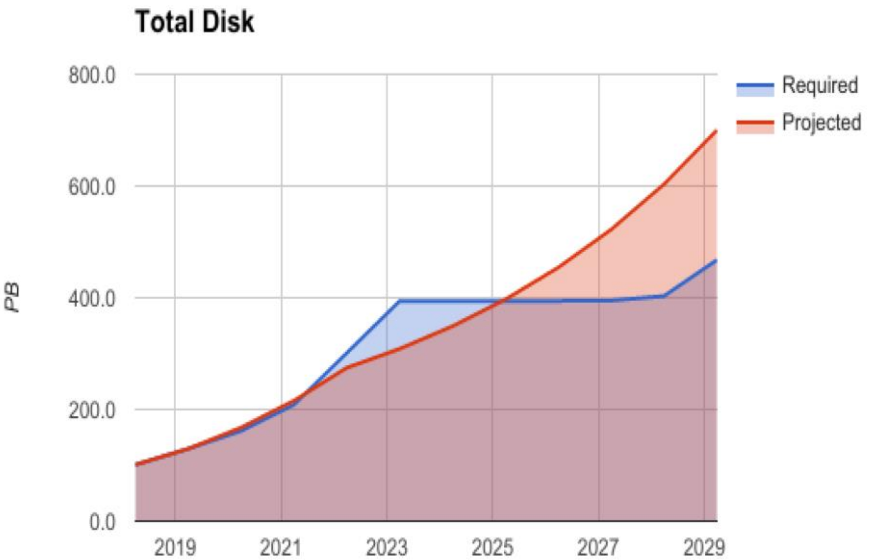
Grid Tiers mostly specialized for given role

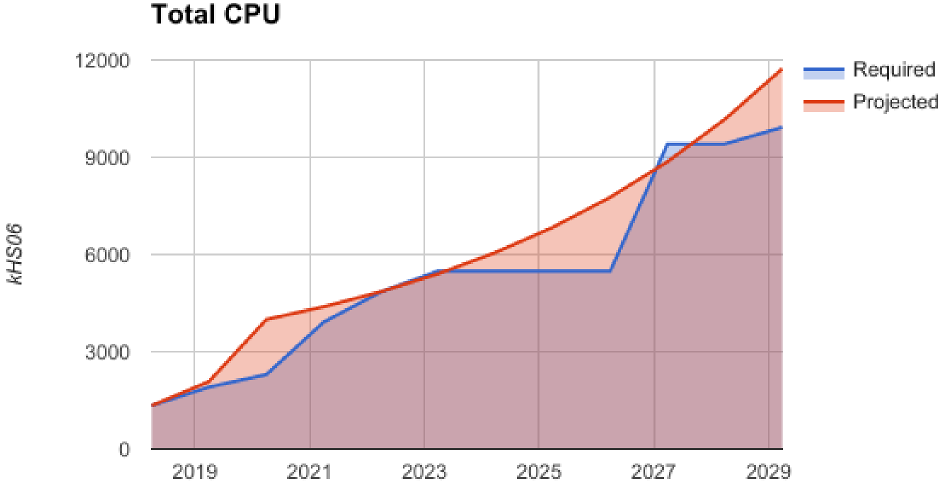
- O2 facility (2/3 of reconstruction and calibration), T1s (1/3 of reconstruction and calibration, archiving to tape), T2s (simulation)
- All AODs will be collected on the specialized Analysis Facilities (AF) capable of processing ~5 PB of data within ½ day timescale

The goal is to minimize data movement and optimize processing efficiency

“Flat budget” scenario







- Overall, we seem to fit under projected resource growth curves
  - Still doable under the flat budget scenario (+20%)

Summary (and caveats)

- If everything works as expected, ALICE will fit under flat budget envelope in Run 3
- Still large uncertainties
  - Run 2 resource evolution
  - Compression efficiency (assumed to be x14)
  - CPU and AOD size estimates (assumed to scale with raw data as in Run 2)
- Availability of Analysis Facilities and ability to grow
  - Depending on AOD size, AFs needs could grow faster than expected under flat budget
- Software readiness
  - We are already accumulating delays
- All this can distort our elegant Run 3 computing model
  - Force us to run analysis on T1s
  - Run simulation of all Tiers as a backfill