CERN-RRB-2018-024

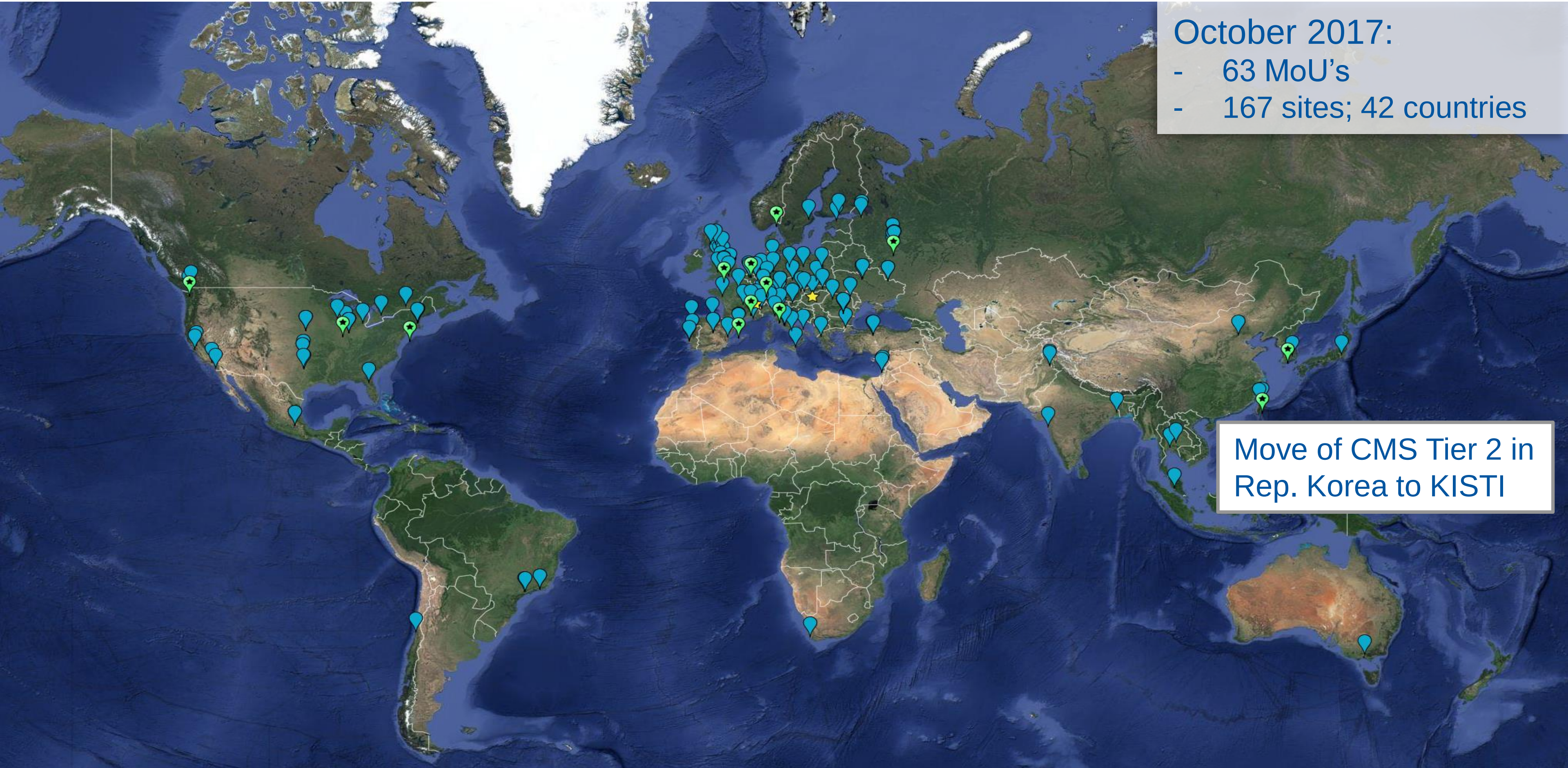
WLCG Status Report

Ian Bird

Computing RRB

CERN, 24th April 2018

WLCG Collaboration

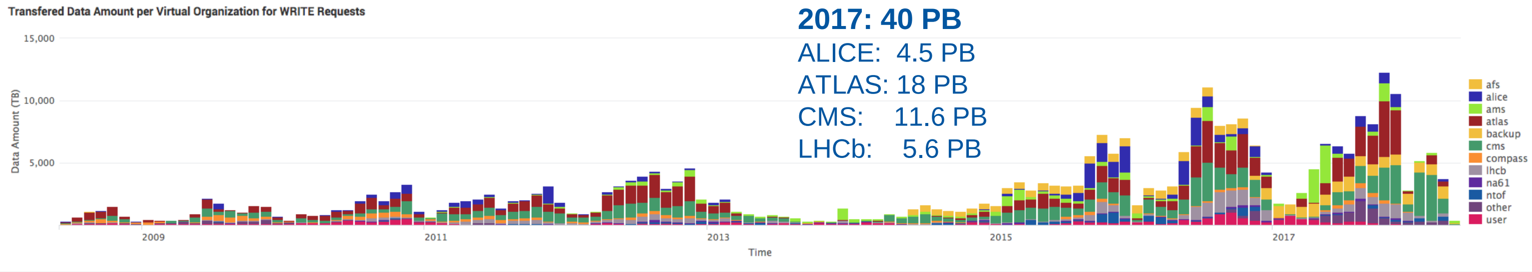


October 2017:

- 63 MoU’s
- 167 sites; 42 countries

Move of CMS Tier 2 in Rep. Korea to KISTI

Data



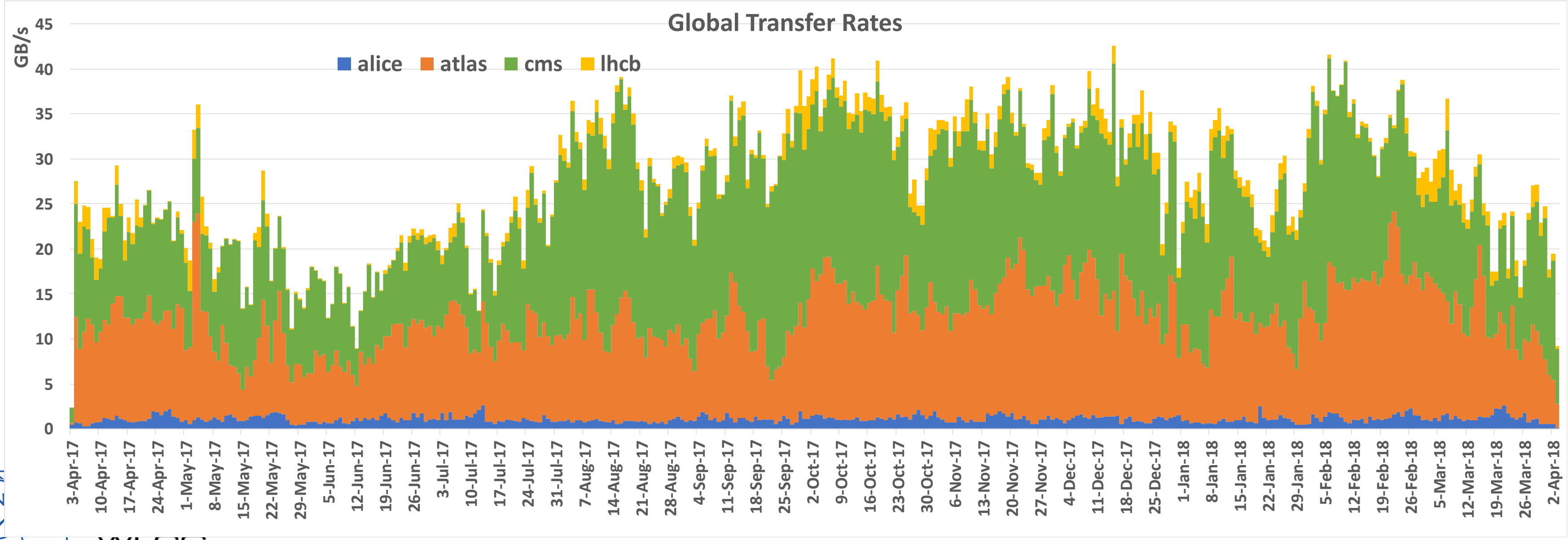
2017: 40 PB

ALICE: 4.5 PB

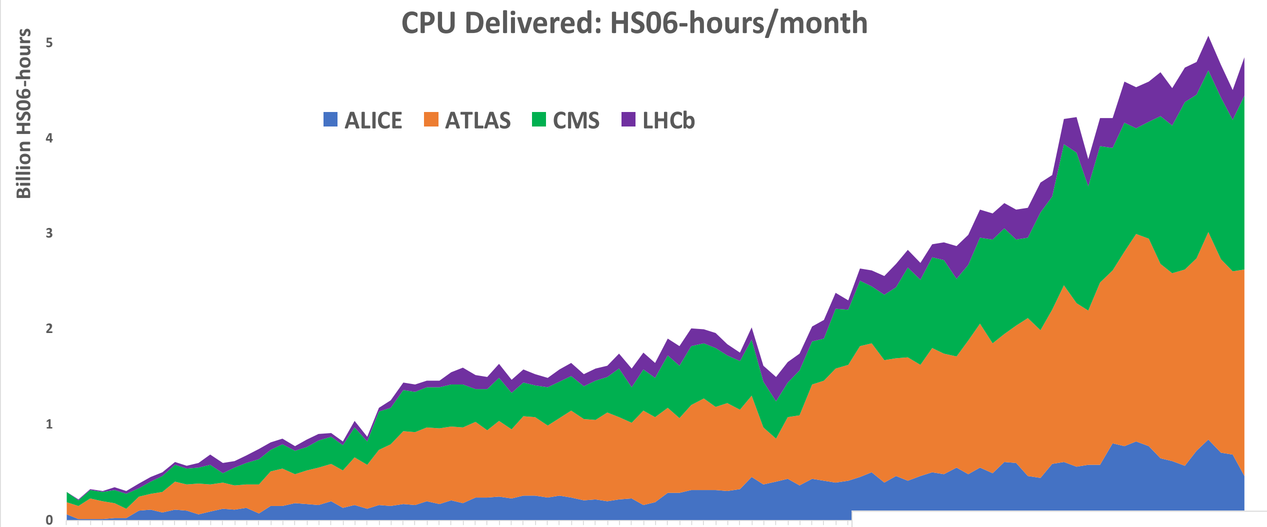
ATLAS: 18 PB

CMS: 11.6 PB

LHCb: 5.6 PB

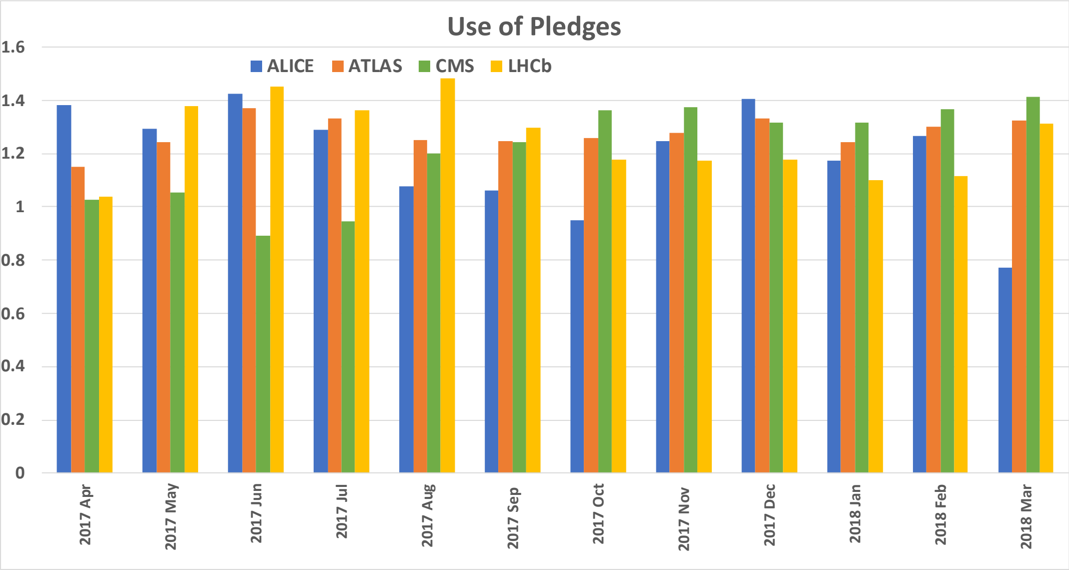


CPU Delivered



New peak: ~210 M HS06-days/month

~ 685 k cores continuous



CNAF Flood; impacts, mitigations

- Nov 9, water main burst and flooded CNAF Tier 1
  - Damage to electrical equipment, lower parts of equipment racks, and tape library
  - Loss of 15% CPU farm, 136 tapes damaged
- CNAF was down until ~ Feb 2018
- Luckily not during data taking
- LHCb worst affected
- Other Tier 1s, CERN provided some contingency
  - Missing (LHC) data accessed from other sites, or recreated
  - Some derived data was unavailable
- Tapes recovered by specialist company
- Tier 1 now back in production with full resources for 2018
  - Some equipment hosted in CINECA
  - No power redundancy for the moment

Other operational items

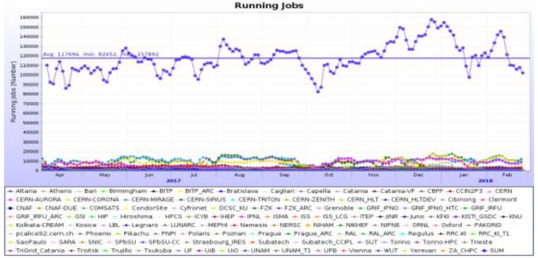
- IPV6 deployment
  - Tier 1s all in dual-stack mode for (almost) all services
  - Campaign started for Tier 2’s; >25% already done
- Meltdown & Spectre problems
  - Caused significant campaigns of firmware updates, patching, rebooting
    - Disruptive, luckily during end-of-year stop
    - Fear that would cause significant loss of performance was unfounded –
      - worst case is <5%, most workflows saw no effect

Experiment updates

ALICE

- ❑ Total data collected in 2017 4.5 PB
  - All replicated to Tier 1s
  - Continuously improving HLT compression resulting in reduction of recorded data volume 2018 tape requests
- ❑ Finished reprocessing of 2015,16 data, calibrated and processed 2017data
  - Full MC sample for 2015-17 has been generated
  - Processed and analysed special Xe-Xe run
- ❑ LHCC has reviewed and approved ALICE RAW to MC ratio
  - 1:1 for pp, 1:0.3 for PbPb
- ❑ Use of CPU is ~70% MC, 10% reconstruction and 20% analysis workloads
  - Successfully preparations for Quark Matter conference in May
  - Increased analysis load efficiently absorbed by the analysis trains
- ❑ ALICE is still facing the most significant Run 2 data taking year
  - 60% of Pb-Pb statistics will be recorded at the end of 2018, mostly central events
- ❑ ALICE 2018 computing resource request was reduced to match the existing pledges and avoid growing gap between pledges and request
  - Expecting to stay under 20% overall growth until 2021

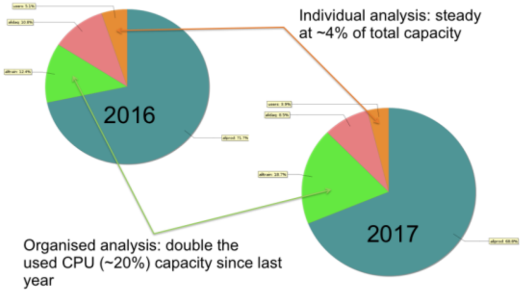
- Good overall performance, 30% capacity increase, including opportunistic resources



WLCG LHCC 27/02/2018 | Predrag Buncic

3

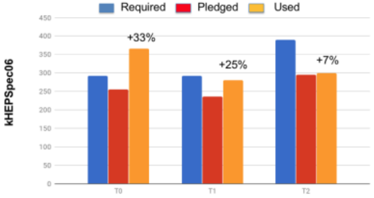
Analysis



Individual analysis: steady at ~4% of total capacity

Organised analysis: double the used CPU (~20%) capacity since last year

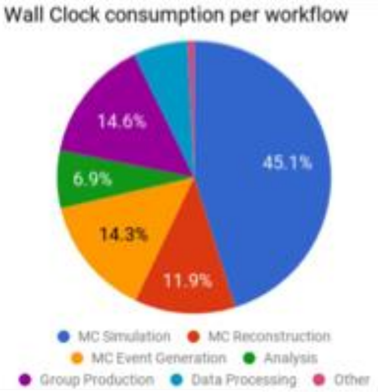
CPU UTILIZATION IN 2017

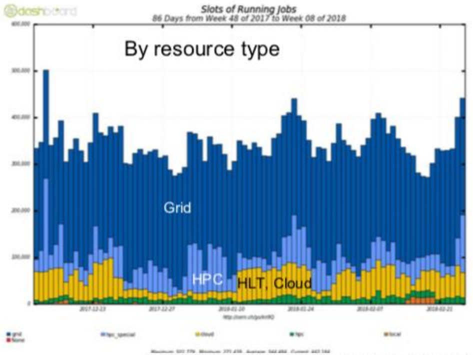


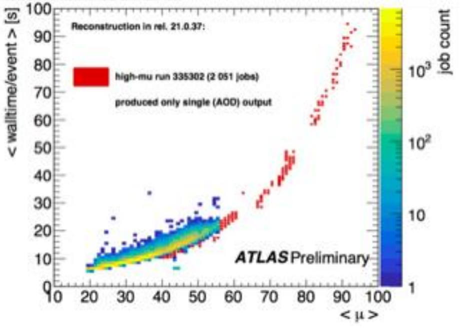
- Good utilization of opportunistic CPU resources

ATLAS

- ❑ Total recorded data was 47 fb^{-1} in 2017
  - ▪ Due to higher instantaneous luminosity and levelling average pile-up was close to x2 wrt design values
- ❑ Tier 0 kept up with data taking, but in 2nd part of year relied on LHC down time and special runs to be able to manage backlogs
  - ▪ The Tier 0 worked at high efficiency, spill over to grid was commissioned but not used
- ❑ MC productions – full sample for 2017, as well as for 2015-16.
  - ▪ New version of GEANT used for 2017 is 30% faster
- ❑ Fast chain for simulation will be validated in 2018. Expected to be factor 10 faster where it can be used; also expect storage savings
- ❑ Mitigation of storage space: AOD's are now 30% smaller; strict control of data lifetimes also important
- ❑ Progress with multi-threaded software: AthenaMT is expected to be in production in LS2
- ❑ Significant opportunistic use –of non-traditional resources for MC: 9.3% from HPCs, and 14% from clouds (including HLT)
  - ▪ Efficiency of opportunism is aided by use of the event service

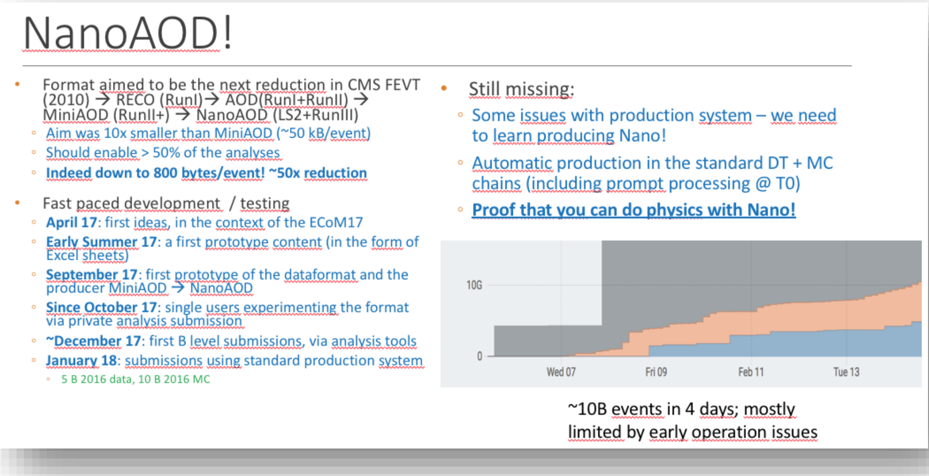






CMS

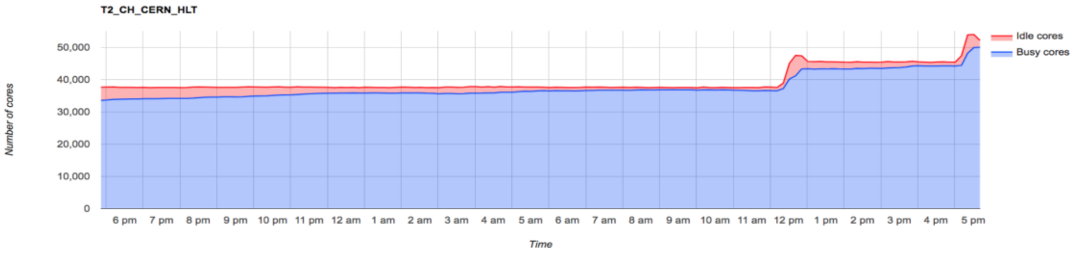
- ❑ The 2017 data and simulation samples have been processed in time for physics results at Winter conferences
- ❑ The higher luminosity and levelling gave much higher average pile-up:
  - 55-60 at start of fill, dropping to ~30 –
  - average of 45 rather than 35 used for resource estimates
  - This increased the load on the Tier 0 and Tier 1s, and gave a delay in prompt reconstruction >48hr design
  - Used CERN analysis resources to supplement Tier 0, lots of mitigation of use of storage
- ❑ In December finished processing MC sample for Phase 2 TDR for HGCAL: 10 PB of storage could then be recuperated
- ❑ Usage level is continuous at 170-200k cores, with ~50k used for analysis
  - During YETS Tier 0 (25k cores) and HLT(50k cores) were also used
- ❑ The CMS Tier 0 has been migrated to HTCondor for 2018



NanoAOD!

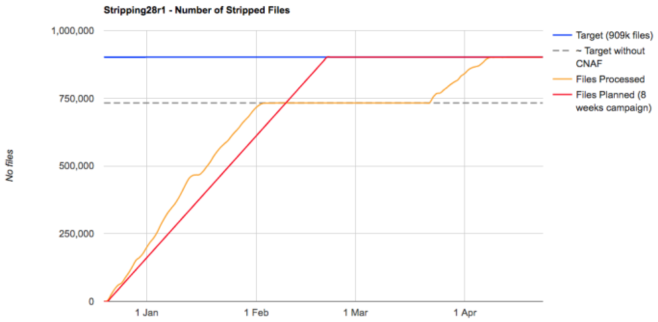
- Format aimed to be the next reduction in CMS FEVT (2010) → RECO (RunI) → AOD(RunI+RunII) → MiniAOD (RunII+) → NanoAOD (LS2+RunIII)
  - Aim was 10x smaller than MiniAOD (~50 kB/event)
  - Should enable > 50% of the analyses
  - **Indeed down to 800 bytes/event! ~50x reduction**
- Fast paced development / testing
  - **April 17**: first ideas, in the context of the ECoM17
  - **Early Summer 17**: a first prototype content (in the form of Excel sheets)
  - **September 17**: first prototype of the dataformat and the producer MiniAOD → NanoAOD
  - **Since October 17**: single users experimenting the format via private analysis submission
  - **~December 17**: first B level submissions, via analysis tools
  - **January 18**: submissions using standard production system
    - 5 B 2016 data, 10 B 2016 MC
- Still missing:
  - Some issues with production system – we need to learn producing Nano!
  - Automatic production in the standard DT + MC chains (including prompt processing @ T0)
  - **Proof that you can do physics with Nano!**

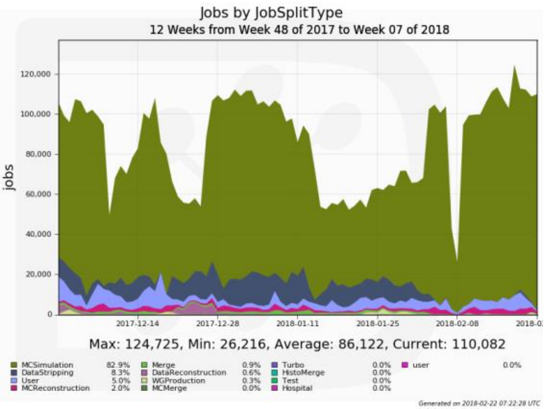
~10B events in 4 days; mostly limited by early operation issues

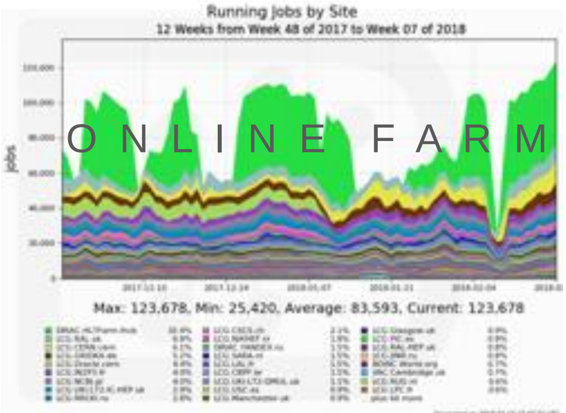


LHCb

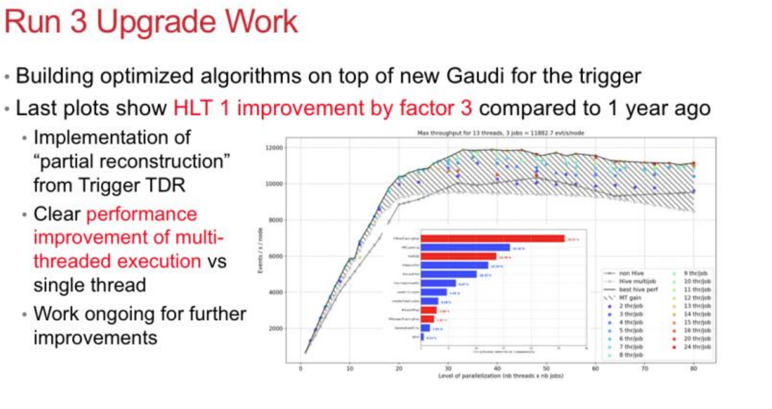
- ❑ During the YETS managed up to 124k concurrent jobs including use of HLT
  - ▪ Most (80%) is simulation, then re-stripping and analysis
- ❑ Re-stripping for 2015,16,17 has all been processed following end of data taking, CNAF data was last part
- ❑ Fast simulation being developed, re-use of underlying events with re-decay of signal in production gives x10 speed up where it can be used.
  - ▪ More fast options in progress







ONLINE FARM

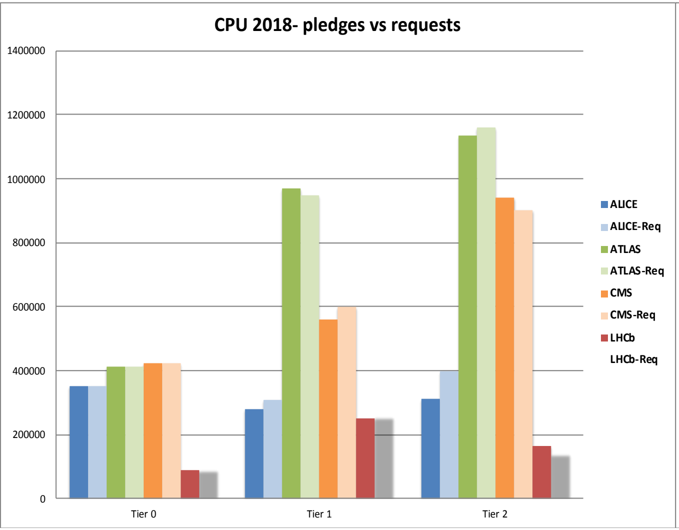


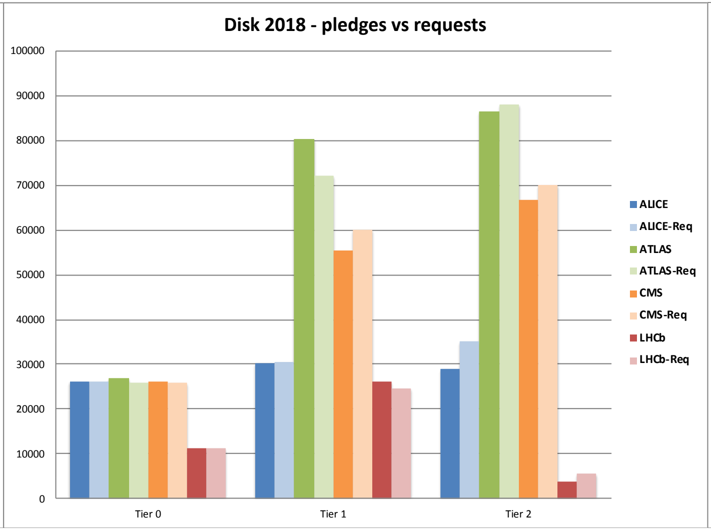
Run 3 Upgrade Work

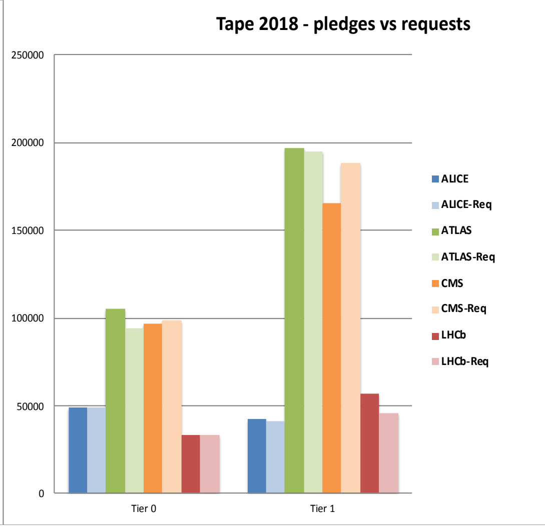
- Building optimized algorithms on top of new Gaudi for the trigger
- Last plots show HLT 1 improvement by factor 3 compared to 1 year ago
  - Implementation of "partial reconstruction" from Trigger TDR
  - Clear performance improvement of multi-threaded execution vs single thread
  - Work ongoing for further improvements

Max throughput for 13 threads, 3 jobs = 11882.7 evt/s/node
Events / s / node
12000
10000
8000
6000
4000
2000
0
10
20
30
40
50
60
70
80
Level of parallelization (nb threads x nb jobs)

2018 Pledge situation

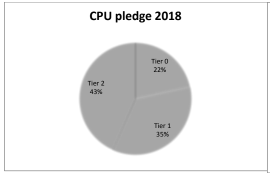


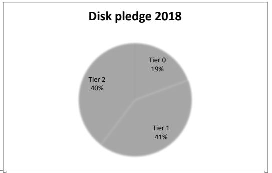


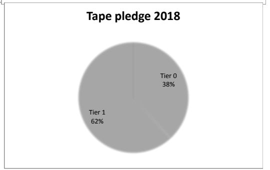


2018 pledges wrt requests:

As given in REBUS







Planning for Run 3

Running conditions – 2018

- Anticipated:
  - Luminosity: 2.0 x 10^{34}
  - 25 ns spacing - BCMS
  - 2544 bunches, with 1.15-1.3x10^{11} protons/bunch
  - Luminosity levelling at pile-up of ~55
    - →average is now ~45 (cf 35 in 2017)
    - →Reconstruction CPU needs 20-25% increased
  - Integrated luminosity: 60 fb^{-1} (cf 45 in 2017)

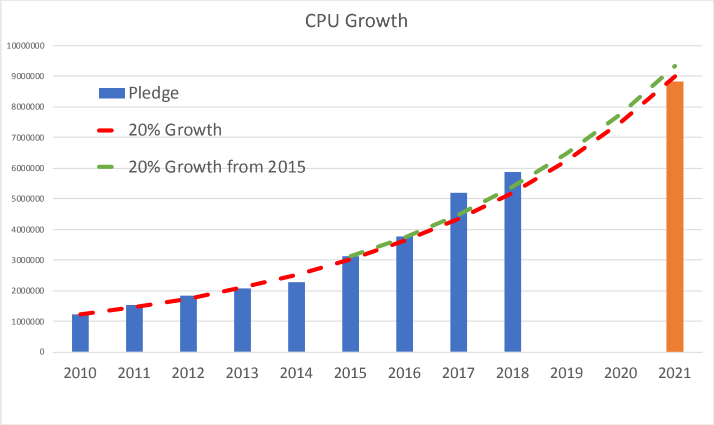
Run 3 running conditions – 1

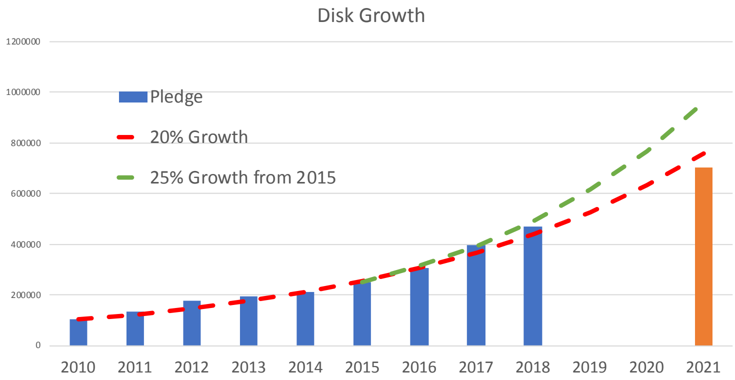
- Following discussion with LHC operations
- Still many unknowns
  - E.g. experiment planned trigger rates are tbd
- Expected conditions:
  - 7 TeV per beam, gives small reduction in beam size
  - The main limitation is the heat load in the cryogenics
  - Expect BCMS filling scheme; 25ns
    - 2544/2556 bunches, $\beta^* = 27$cm
    - 1.3 $\times 10^{11}$ protons/bunch
  - 2×10^{34} (could be a bit higher) is the limit due to the inner triplet cooling
    - This will not change in LS2
    - This is a pile up of ~60

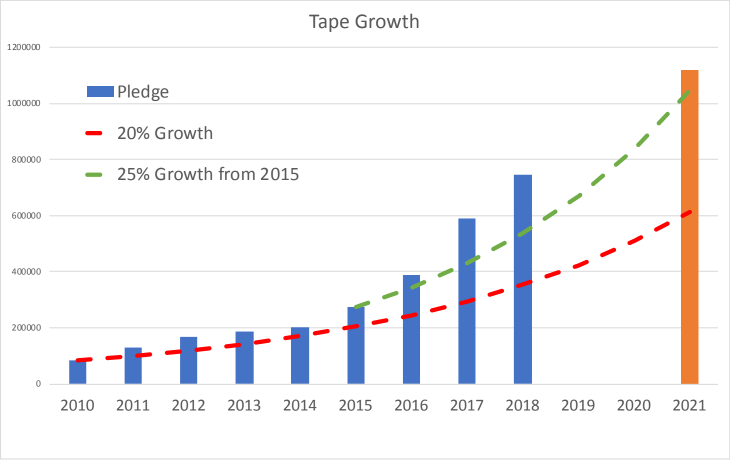
Summary – Run 3:

- ❑ Similar to 2018
- ❑ If the experiments luminosity level at a higher pile-up and for longer ➔
  - Potentially higher average pileup
  - Non-linear increase in CPU time
- ❑ Possibly less time between fills – more live time
- ❑ Overall the best estimate is 30% (50% conservatively) more resources needed than in 2018
  - But we have not seen 2018 yet
- ❑ For 2021: 1st year after LS2, could be only half-year live time but ramp up to optimal conditions rapidly
- ❑ Unknown:
  - Still need plans for experiment trigger rates
  - And plans for luminosity levelling

Resource evolution





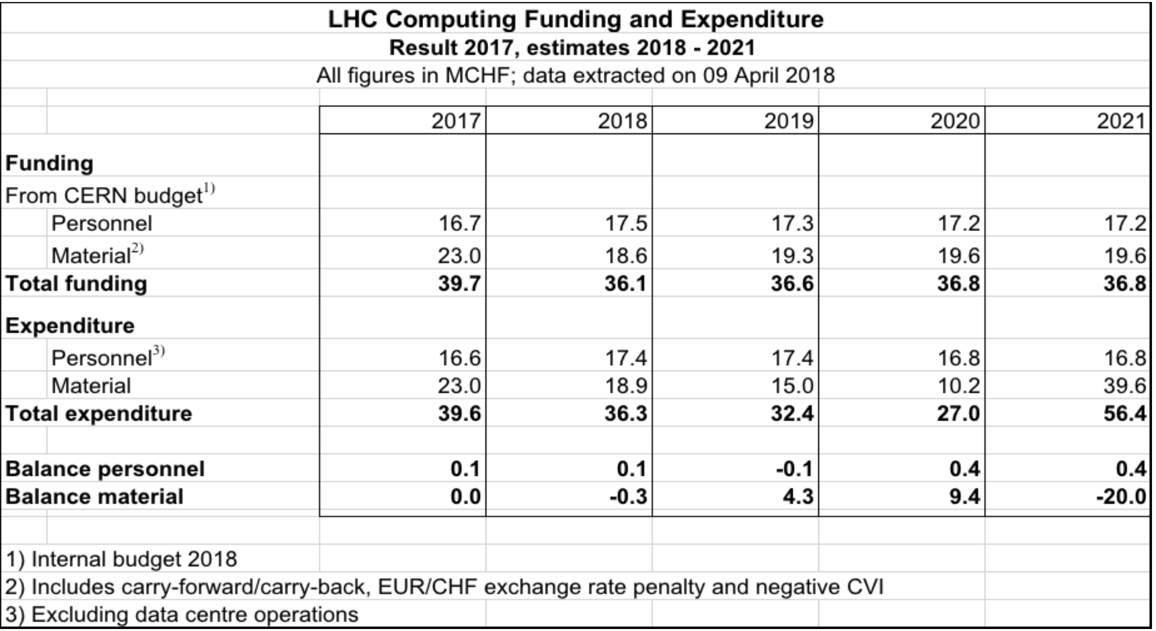


- 2010-2018 – pledges
- 2021 assume 1.5 x 2018

However …

- ALICE and LHCb are upgrading during LS2, so the expectations of their needs do not follow the assumptions in the previous slides:
  - LHCb:
    - luminosity and pileup increase by factor 5.
    - Major changes in computing model result in higher trigger rate and HLT output bandwidth.
    - LHCC milestone for computing model in Q3/2018, together with engineering TDR – currently under review
  - ALICE:
    - Factor 100 increase in readout rate (50 kHz)
    - Data volume increase mitigated by online reconstruction and raw data compression in new O2 facility
    - O2 TDR is approved; summary needs are:
    - Increases in 2021 wrt 2018: CPU: 48%, disk: 74%, tape 90%

WLCG Funding & Expenditure



LHC Computing Funding and Expenditure

Result 2017, estimates 2018 - 2021

All figures in MCHF; data extracted on 09 April 2018

	2017	2018	2019	2020	2021
Funding					
From CERN budget[1]					
Personnel	16.7	17.5	17.3	17.2	17.2
Material[2]	23.0	18.6	19.3	19.6	19.6
Total funding	**39.7**	**36.1**	**36.6**	**36.8**	**36.8**
Expenditure					
Personnel[3]	16.6	17.4	17.4	16.8	16.8
Material	23.0	18.9	15.0	10.2	39.6
Total expenditure	**39.6**	**36.3**	**32.4**	**27.0**	**56.4**
Balance personnel	**0.1**	**0.1**	**-0.1**	**0.4**	**0.4**
Balance material	**0.0**	**-0.3**	**4.3**	**9.4**	**-20.0**

1) Internal budget 2018

2) Includes carry-forward/carry-back, EUR/CHF exchange rate penalty and negative CVI

3) Excluding data centre operations

Personnel: balanced situation

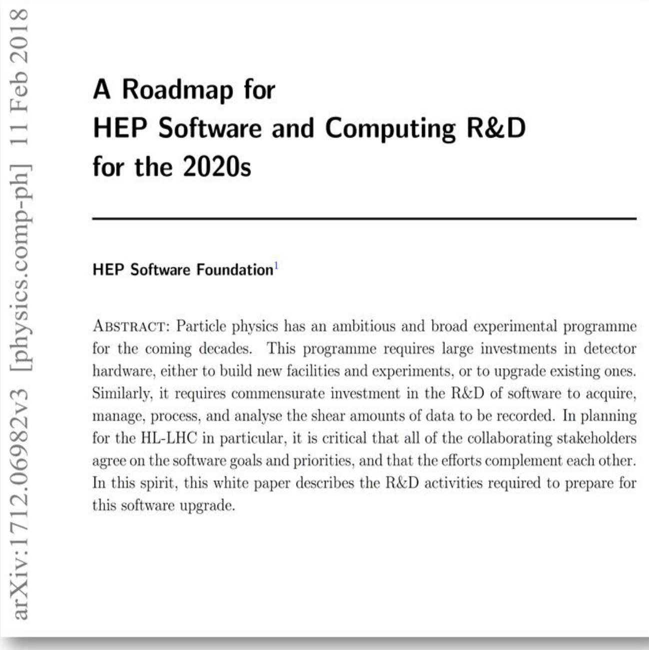
Materials planning based on currently understood parameters:

- CERN plan for 2019,20 is minimal purchases –
- 2021 assumes 1.5x2018
- Cost extrapolations based on recent experience;
- Large uncertainties and variations
- Overall balance in 2019-2021 is ramp up to Run 3

Planning for HL-LHC
CWP & Strategy document

Strategy

- In 2017, the global HEP community has produced the Community White Paper (CWP), under the aegis of the HEP Software Foundation (HSF).
  - A ground-up gathering of input from the HEP community on opportunities for improving computing models, computing and storage infrastructures, software, and technologies.
  - It covers the entire spectrum of activities that are part of HEP computing.
  - While not specific to LHC, the WLCG gave a charge to address the needs for HL-LHC along the lines noted above.
  - Published: https://arxiv.org/abs/1712.06982



arXiv:1712.06982v3 [physics.comp-ph] 11 Feb 2018

A Roadmap for HEP Software and Computing R&D for the 2020s

HEP Software Foundation[1]

ABSTRACT: Particle physics has an ambitious and broad experimental programme for the coming decades. This programme requires large investments in detector hardware, either to build new facilities and experiments, or to upgrade existing ones. Similarly, it requires commensurate investment in the R&D of software to acquire, manage, process, and analyse the shear amounts of data to be recorded. In planning for the HL-LHC in particular, it is critical that all of the collaborating stakeholders agree on the software goals and priorities, and that the efforts complement each other. In this spirit, this white paper describes the R&D activities required to prepare for this software upgrade.

- Strategy document – prioritise a program of work from the WLCG point of view:
  - A focus on HL-LHC, building on all of the background work provided in the CWP, and the experience of the past.

Strategy – outline

Themes

1. Software performance
2. Algorithmic improvements/changes
   - E.g. reco, fast MC, event generators
3. Reducing data volumes
4. Managing operations costs
5. Optimizing hardware costs

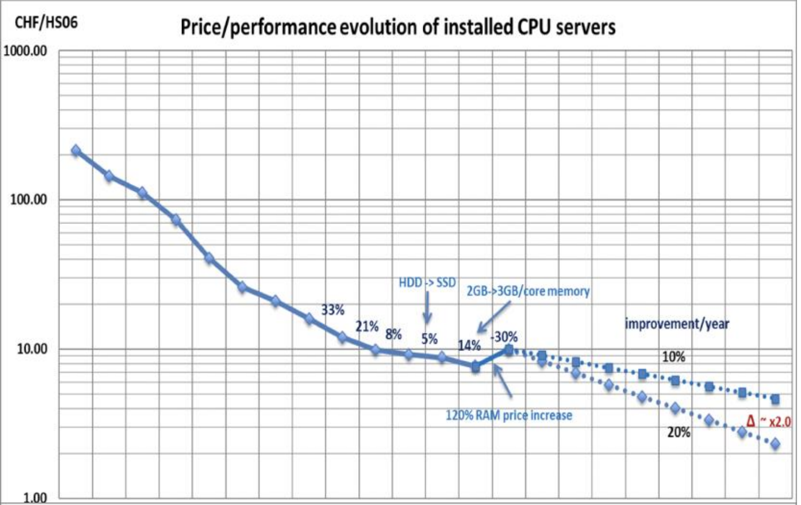
Demonstrate that we are in control of costs, while maximizing physics output

1. Introduction
2. Computing Models
3. Experiment Software
4. System Performance & Efficiency
   - Cost Model
   - Software performance
   - I/O performance
5. Data & Compute Infrastructure
   - Storage consolidation
   - Caching
   - Storage, access, transfer protocols
   - Data Lakes
   - Network
   - Processing resources
   - Cloud analysis
6. Sustainability
   - Common solutions
   - Security infrastructure
7. Workplan
8. Appendix: technology evolution
9. Appendix: likely benefits

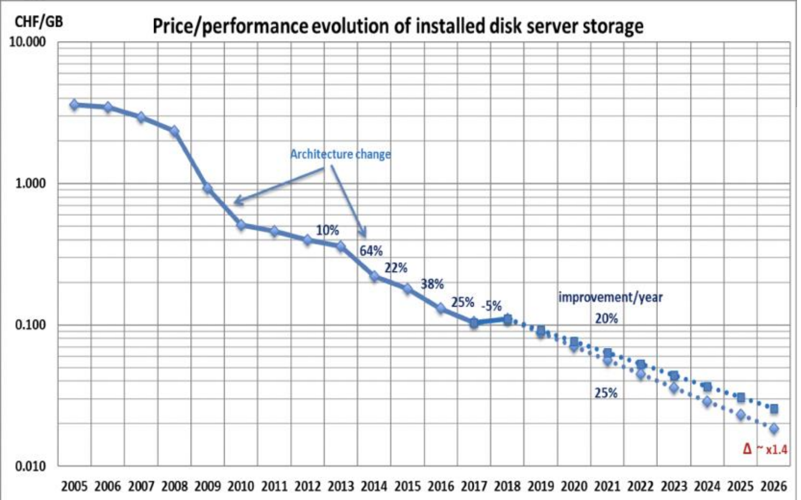
Status

- Draft is being reviewed by the LHCC
- Has been provided to the WLCG MB
- Following discussion at the next LHCC meeting will be published
- R&D efforts
  - Specific R&D projects are being proposed
    - Will have explicit timelines, goals, metrics, etc.
    - Focus on software, cost models, and data management
      - These are all being organized now
    - As well as ongoing work on reconstruction, simulation, etc.
  - Integrate with existing working groups where practical

Update on market evolution



Price/performance evolution of installed CPU servers



Price/performance evolution of installed disk server storage

Disk Server cost:
- Very hard to estimate real costs of HDDs – 70 different 6 TB models with price range of x2.5
- CERN observed x2 between ”street” prices and purchase prices

Current assumptions server annual price/performance improvement:

CPU servers: 15%

Disk server: 20%

Overview

- Technology progress is good, but obstacles are CPU, RAM, NAND
- Markets dominated by a few companies in all areas
- Price/performance advances are slowing
- Memory prices will increase, but price reductions expected when new fabs come online
- New processors/architectures focused on ML
- HDD still important for us, SSD not cost effective at scale
- Tape market and development is a concern
- Server market still 99% Intel, no convincing alternatives yet

Open access, data preservation, etc.

- Current situation is ad-hoc
  - Open data portal – backed by EOS disk storage
  - Different experiments have different scales of use
  - Cost of resources (~5 PB disk) is coming from the WLCG budget (on top of pledges)
  - Only CERN is currently doing this (?)
- Need a better medium term outlook for what is likely to be needed – scale of resources
  - This needs some statements from the experiments on their plans and the scale
- Need to understand how it is funded –
  - In future it will have to come out of the pledges (at CERN) – needs agreement
  - Or specific budget line for this
- Need a policy as to whether this a a responsibility of CERN alone or of the WLCG collaboration
  - And how a distributed archive would be managed

Conclusions

- ❑ Very efficient and heavy use of WLCG during the winter stop, new peak usage reached
- ❑ Major incident at CNAF accommodated by other centres
- ❑ Resources and infrastructure in place for 2018
- ❑ Community White Paper published and WLCG Strategy document drafted –
  - R&D activities aimed at HL-LHC beginning