Tier-1 Configuration Evolution & Options

J. Flix – PIC/CIEMAT – jflix@pic.es

March 2017 GDB – ISGC2017 - Taipei

PIC
port d'informació
científica

Outline

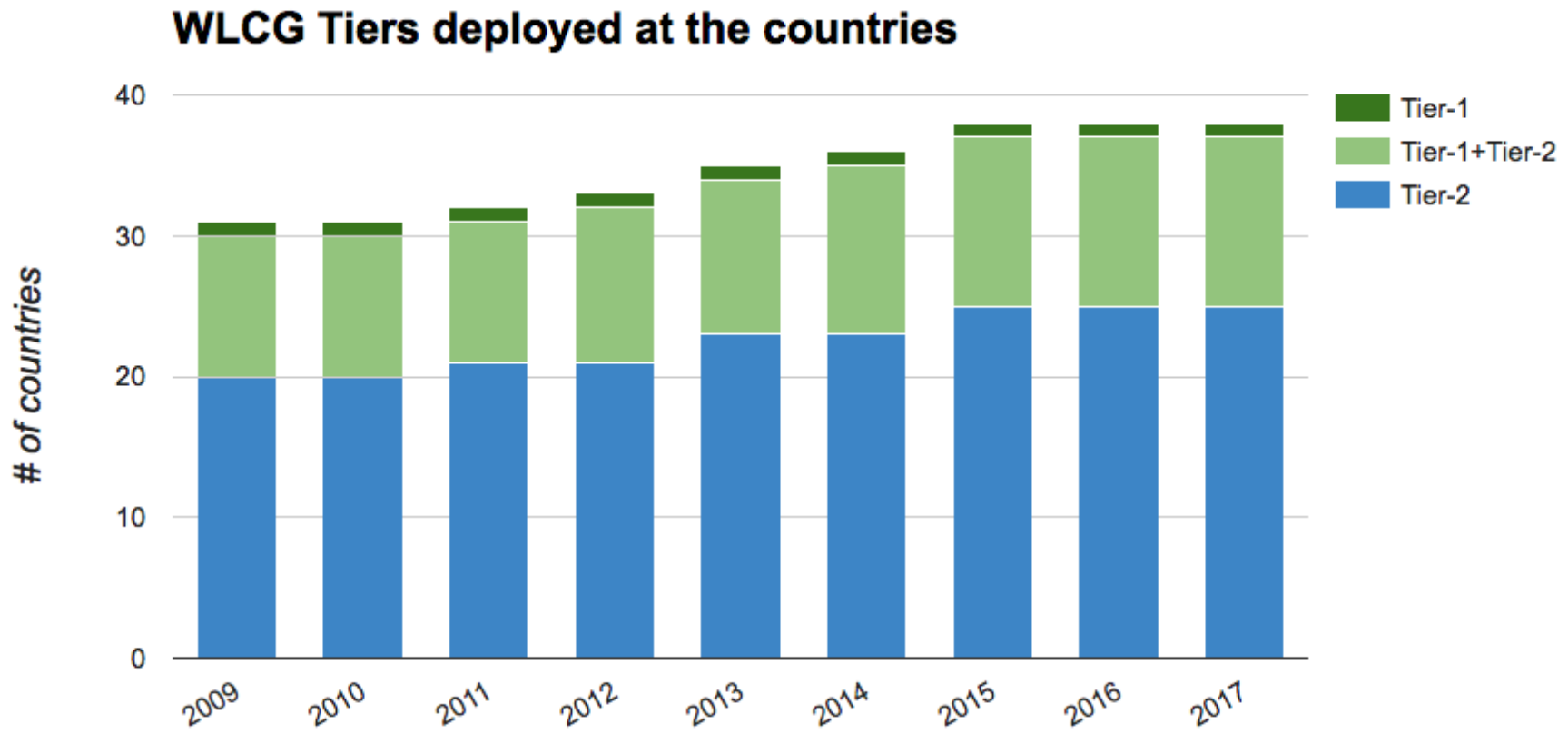
- Not going to explain (all of) the functions of a Tier-1, in detail
- Look at the evolution/usage of WLCG tiers in the last years
- Different modes of Tier-1 operation & current R&D activities
- Tier-1/Tier-2 activities and reliabilities
- The effect of flat-funding budgets in WLCG for 2017→
- Computing in Run3 and HL-LHC
- Modeling the current WLCG costs → My 'toy' model (cost scale issue)
- Personal thoughts on evolution

One can easily touch the 40k active cells limits in Google Sheets

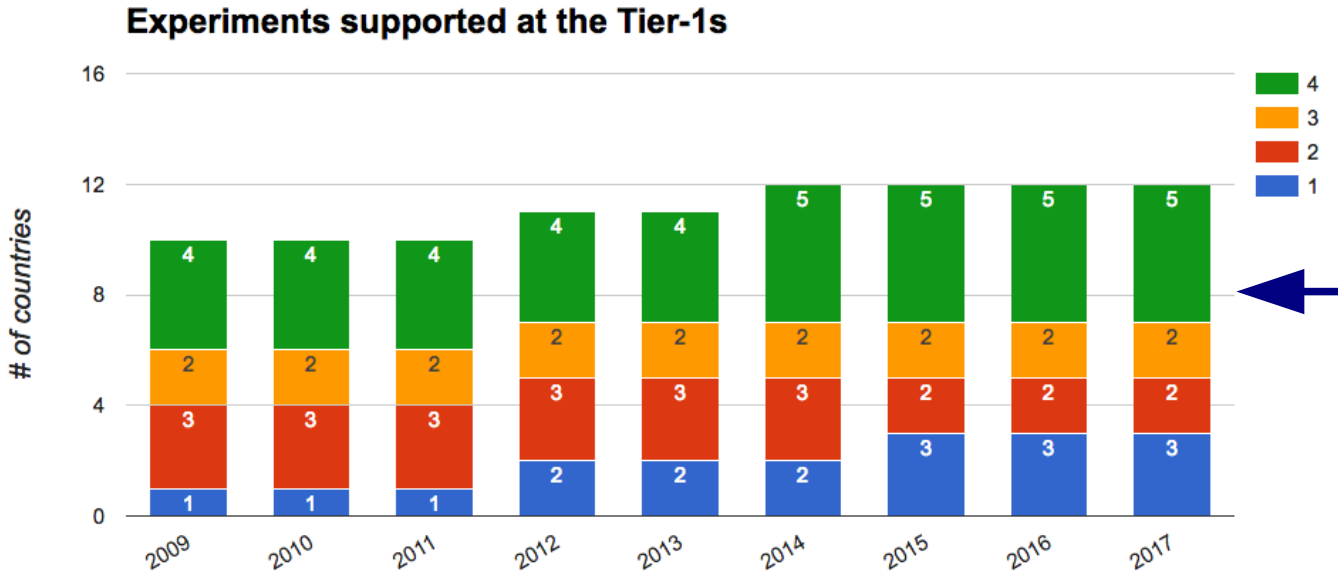
WLCG Tiers: countries participation

- As of today, WLCG has resources in ~40 countries:

→ The countries with Tier-1(s), offer Tier-2 resources as well (except NL)
→ The majority of countries offer Tier-2-only resources



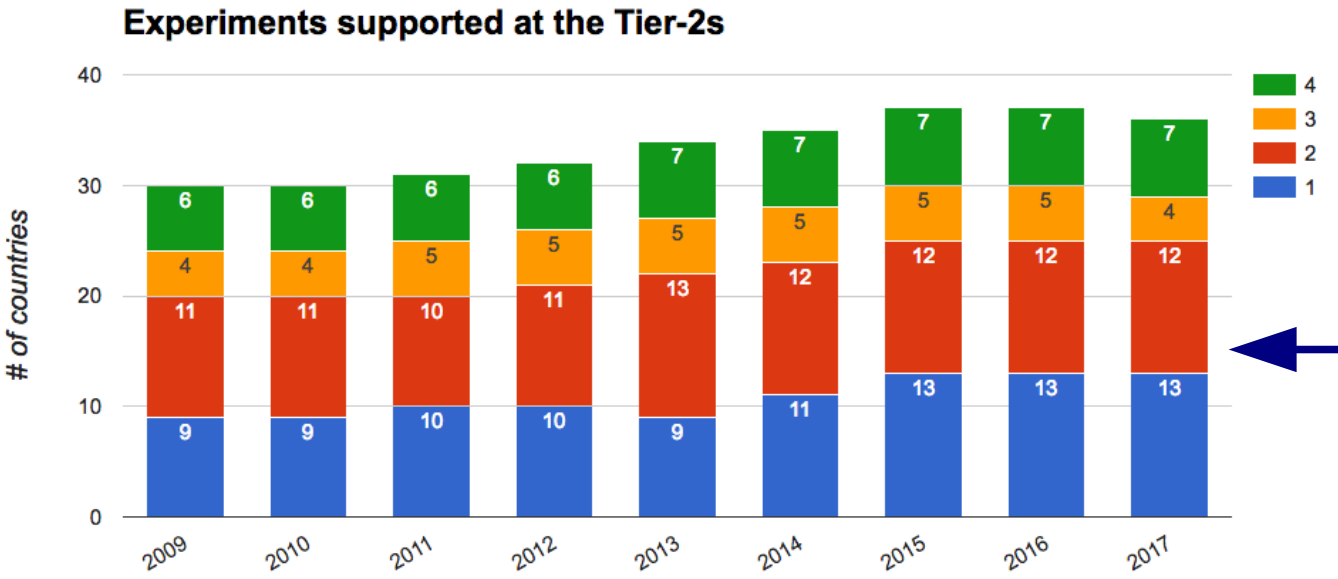
Experiments supported @countries



Countries with Tier-1s typically support most of the LHC exp. in the sites

→ via multi-VO T1s

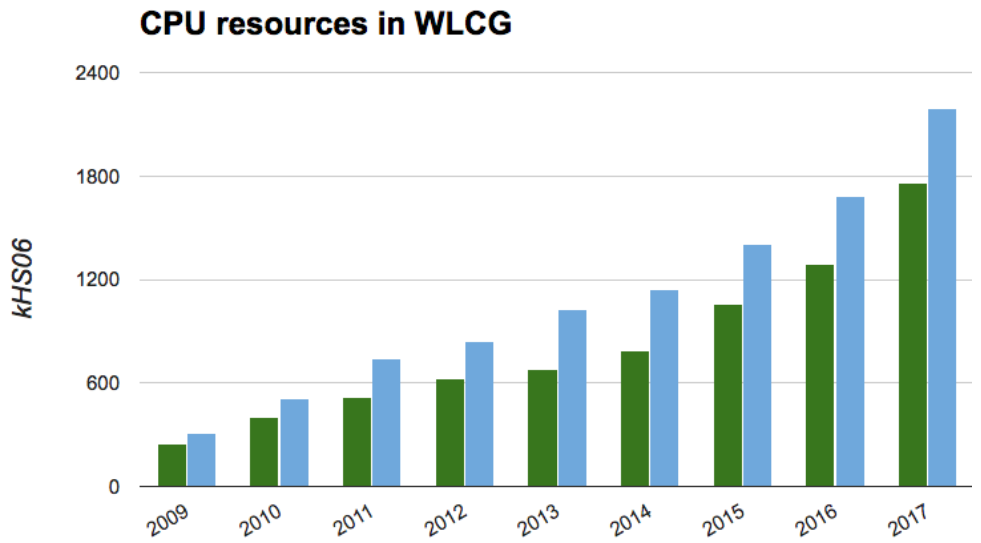
→ via independent T1s

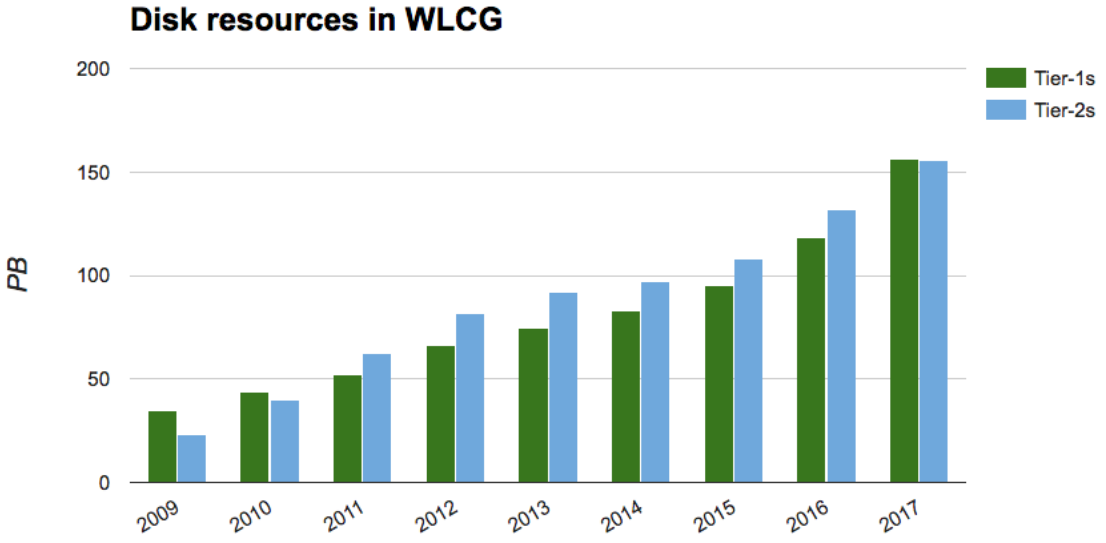


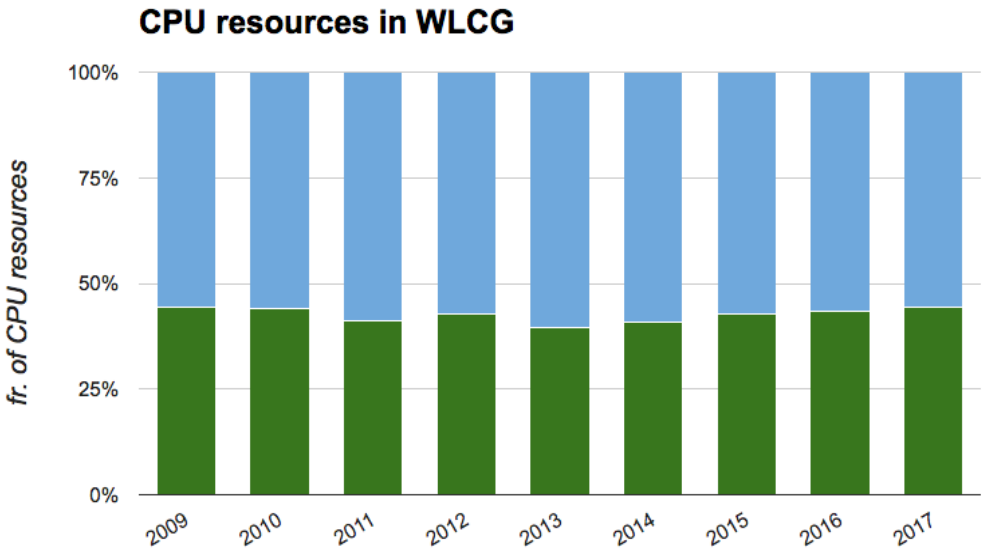
Tier-2s at the countries typically support 1 or 2 exps.

→ T2s typically support 1 exp.

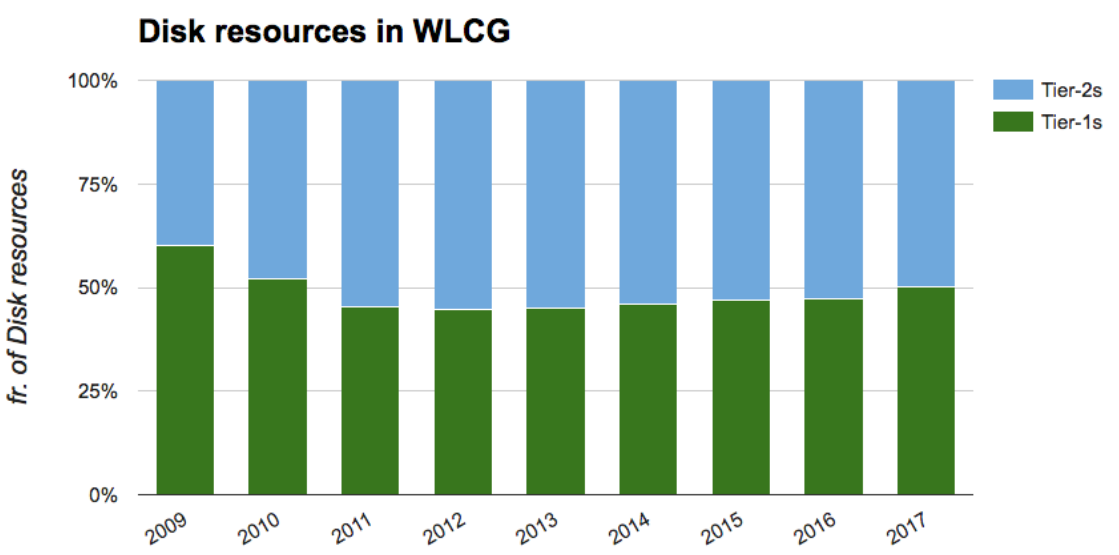
Deployed resources at Tier-1s and Tier-2s





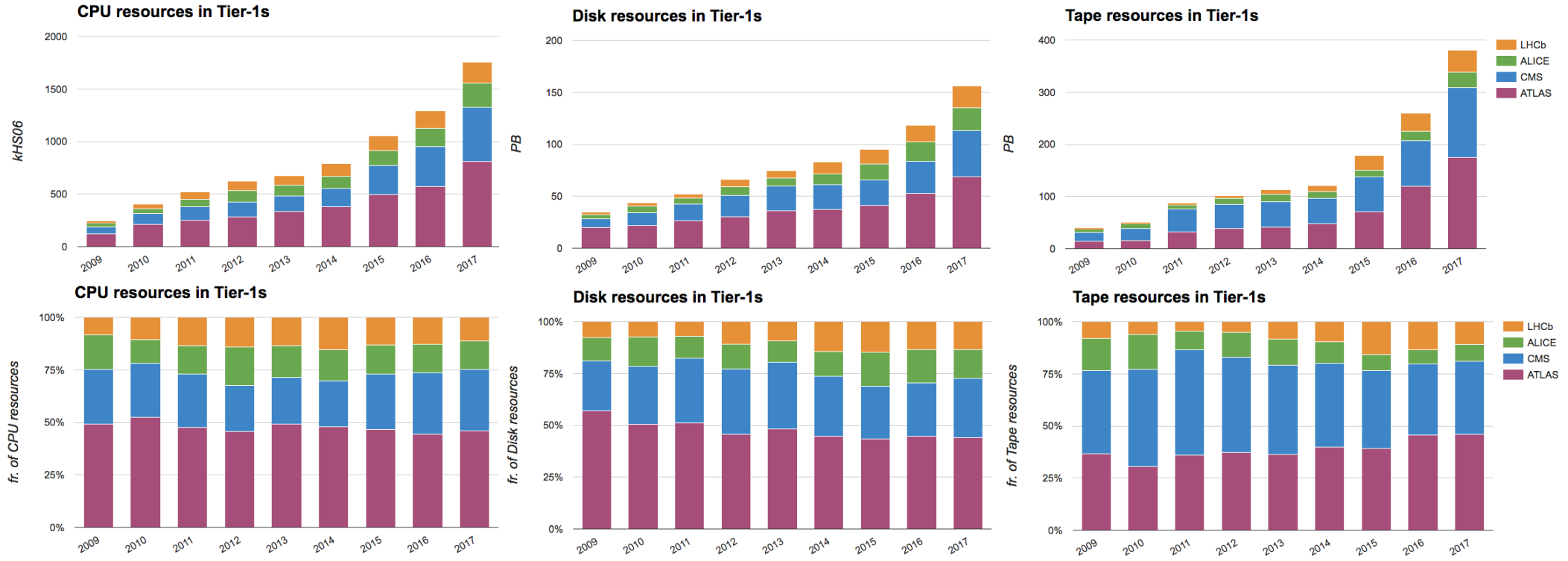


~45% of CPU is provided by Tier-1s



~50% of Disk is provided by Tier-1s

Experiment resources at the Tier-1s



- The majority of resources in WLCG Tier-1s are pledged/requested by **ATLAS and CMS**

  → ~73% (CPU), ~76% (DISK), and ~80% (TAPE) ← **Averages**

- Disk resources growth are more <u>contained</u> than in other resources

  → Asked/recommended by CRSG, since the disk is the most expensive resource

  - Development of new tools and procedures to optimize the disk usage
  - Changes in exps. computing models to contain growth

PIC
port d'informació
científica

Tier-1s in WLCG: modes of operation

“LOCALIZED”

- Resources deployed in one site
- Bare metal WNs attached to a batch system (CE Grid interfaces), or running VMs in private clouds or using Vacuum models

“DISTRIBUTED”

- Resources deployed in several sites – even trans-national collaboration **[NDGF]**
- HPC cluster resources or Grid sites exploited
- Distributed disk storage and eventual deployment of data caches

“ELASTIC”

- “localized” (or “distributed”) sites elastically growing using (more) HPC clusters and/or commercial Cloud providers **[see later]**

Tier-1s: (some) current changes/challenges

Computing

- **Dockers** used in production (allows SL7/CentOS7 Wns)
- Adoption of **HTcondor** and **HTcondor-CEs**
- **Oil-immersion** techniques for CPU resources **[PIC]**

Disk Storage

- Adoption of **Ceph**: recycling 'old' storage, or as an alternative to current storage

Tape Storage

- Several **migrations** from old to new technologies
- **T10K out of business**: some words from FNAL CIO: http://computing.fnal.gov/news/

Network

- **WAN increases** (LHCOPN/LHCONE) everywhere: multi-10Gbps/200Gbps
- **IPv6**: disk pools available; WNs soon available (dual-stack)
- **SDN** enabled routers deployed for R&D **[ASGC]**

Tier-1s: (some) current changes/challenges

Infrastructural/core

- **BNL** unification of all scientific computing (HPC/HTC) facility operations into one organization – plans for transitioning to a **new datacenter**
- **SARA** tape storage moved to **new datacenter**
- **TRIUMF** being integrated into **Compute Canada** to reduce infr. /op. costs
  → new hardware deployed in Simon Fraser University (SFU) – federated sites
  → TRIUMF-side services to be decommissioned in 2018
- **NDGF** underwent an audit to **improve operations and costs**
- **Spanish region** was audited to **optimize** the usage of deployed resources
  → Federation of CIEMAT/IFAE/**PIC** sites (~65% of LHC resources in Spain)
  → Elastic growth tests for peak demands or special requests foreseen
- **FNAL**: **HEPCloud project** to extend into commercial/community clouds, Grid federations, and HPC centers – peak demands or special requirements
- **BNL & FNAL**: **Amazon/EC2** and **AWS S3 storage** tests
- Several Tier-1s in **HNSciCloud**: joint procurement of comm. cloud services

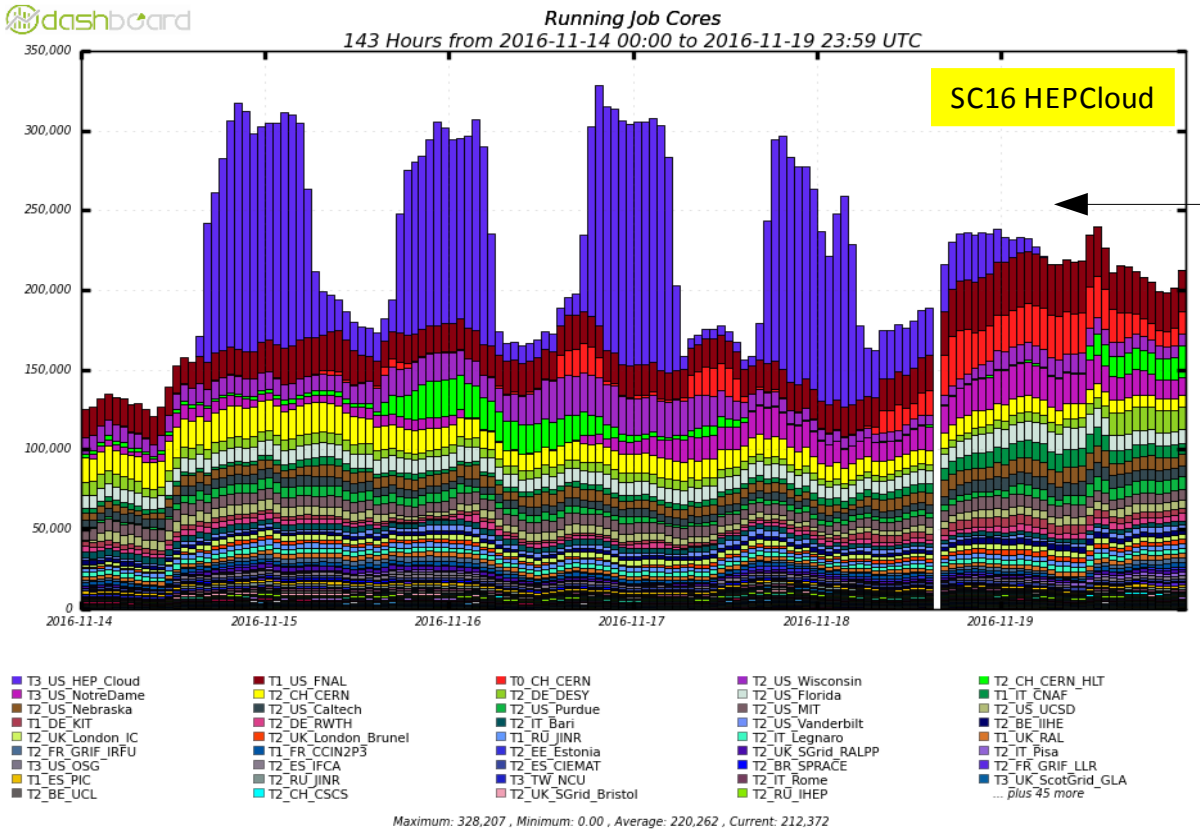
Opportunistic resources

- Exploitation of HPC centers and commercial clouds has been a priority in the WLCG Computing Program in the recent years

- CMS Experiment

  → Transparent use of NERSC resources @US (Edison, Cori-1, Cori-2)

  → AWS @US, Google Cloud Platform @US, Aruba @IT, ongoing Microsoft Azure



Using the FNAL HEPCloud facility w/HTcondor to send bursts of CMS simulation jobs to GCP

$100k credit

The bursts were approx. of the same size of the whole CMS Computing at all the Tiers!
(doubled the capacity of the CMS HTCondor global pool)

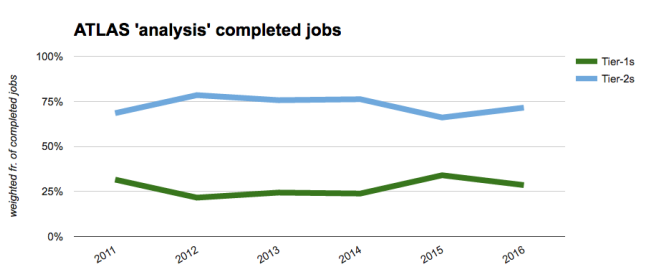
https://cloudplatform.googleblog.com/2016/11/Google-Cloud-HEPCloud-and-probing-the-nature-of-Nature.html

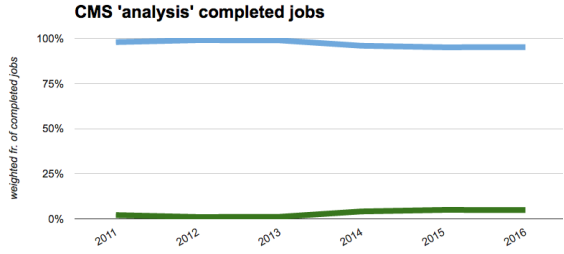
Activities run at the WLCG Tiers

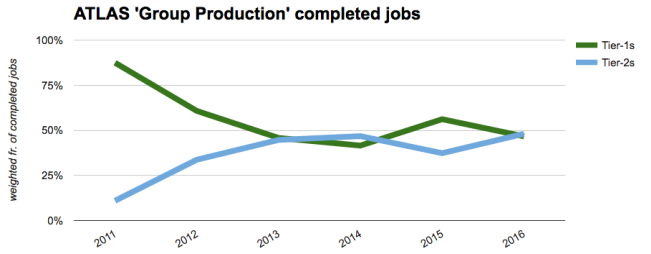
- The tiered structure to compute is **vanishing**:
  → Tools and procedures deployed to flexibly use all of the available computing resources
  → access of data through WAN

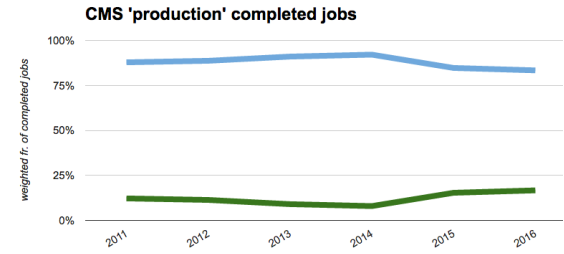
- Big and reliable T2s **growing**

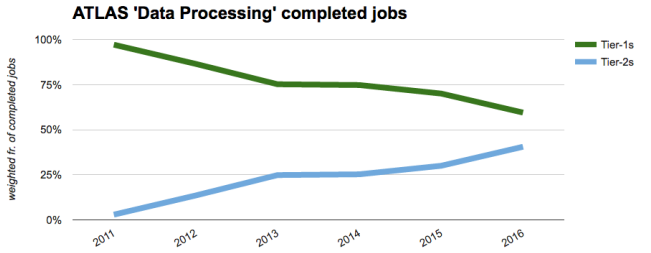
- Tier-1s play an important role for **long-term storage**, offer **24x7**, they are subject to **high reliability** levels, they can be instrumental as gateways for elastic growth

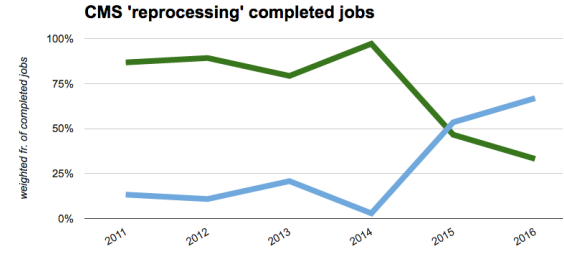


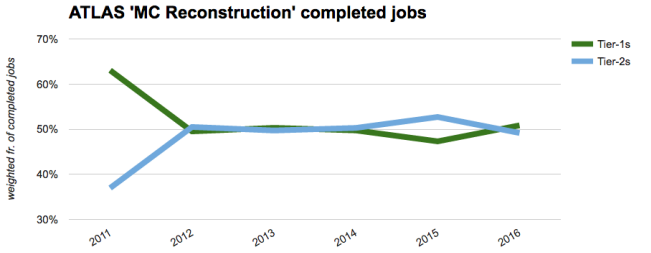


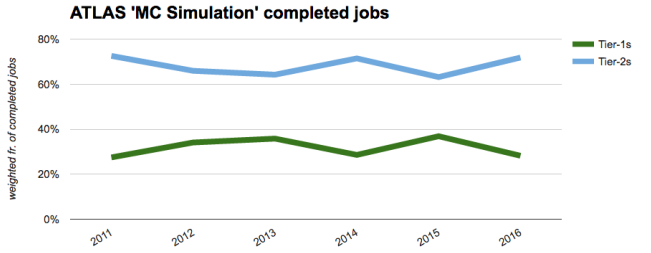






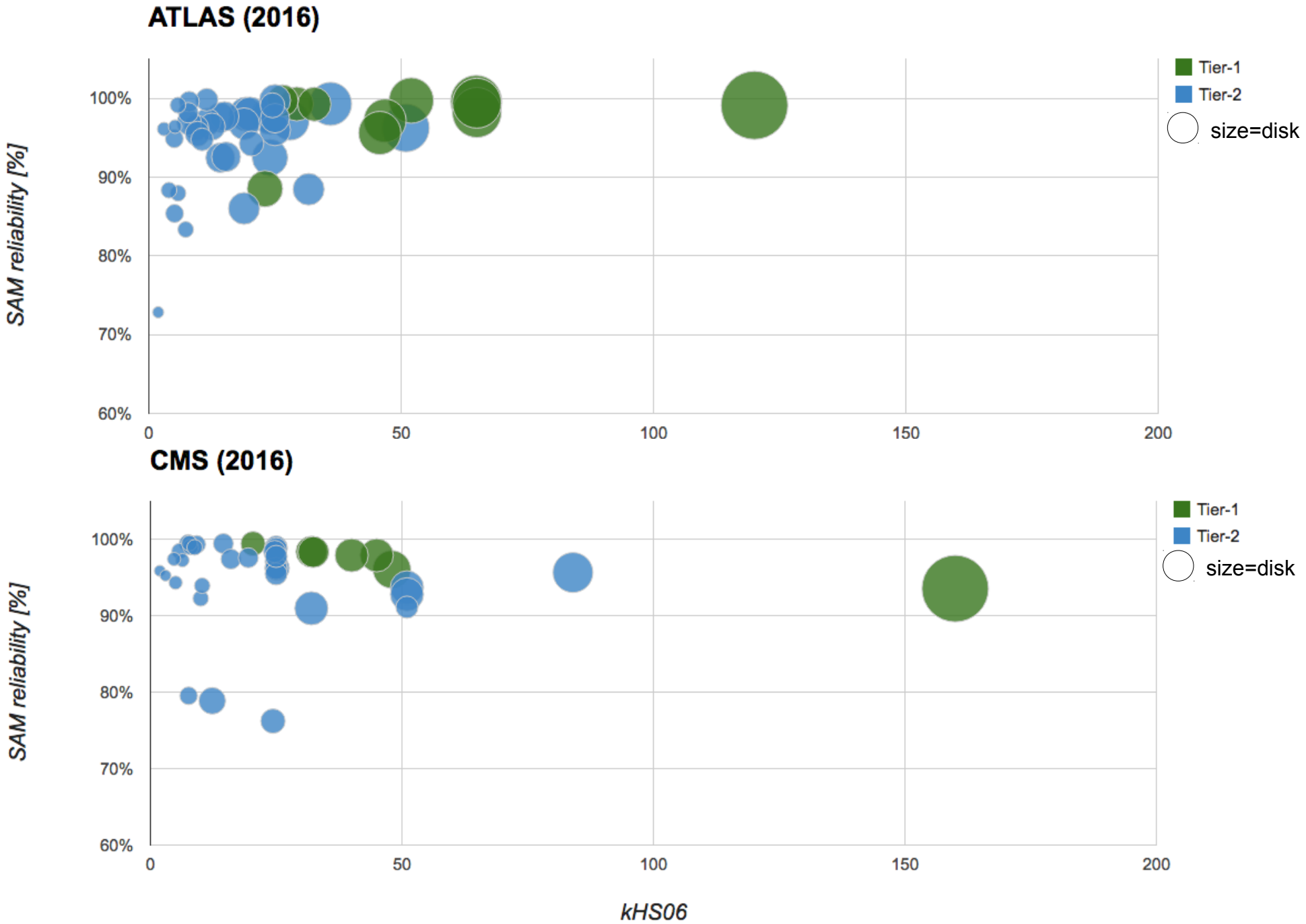




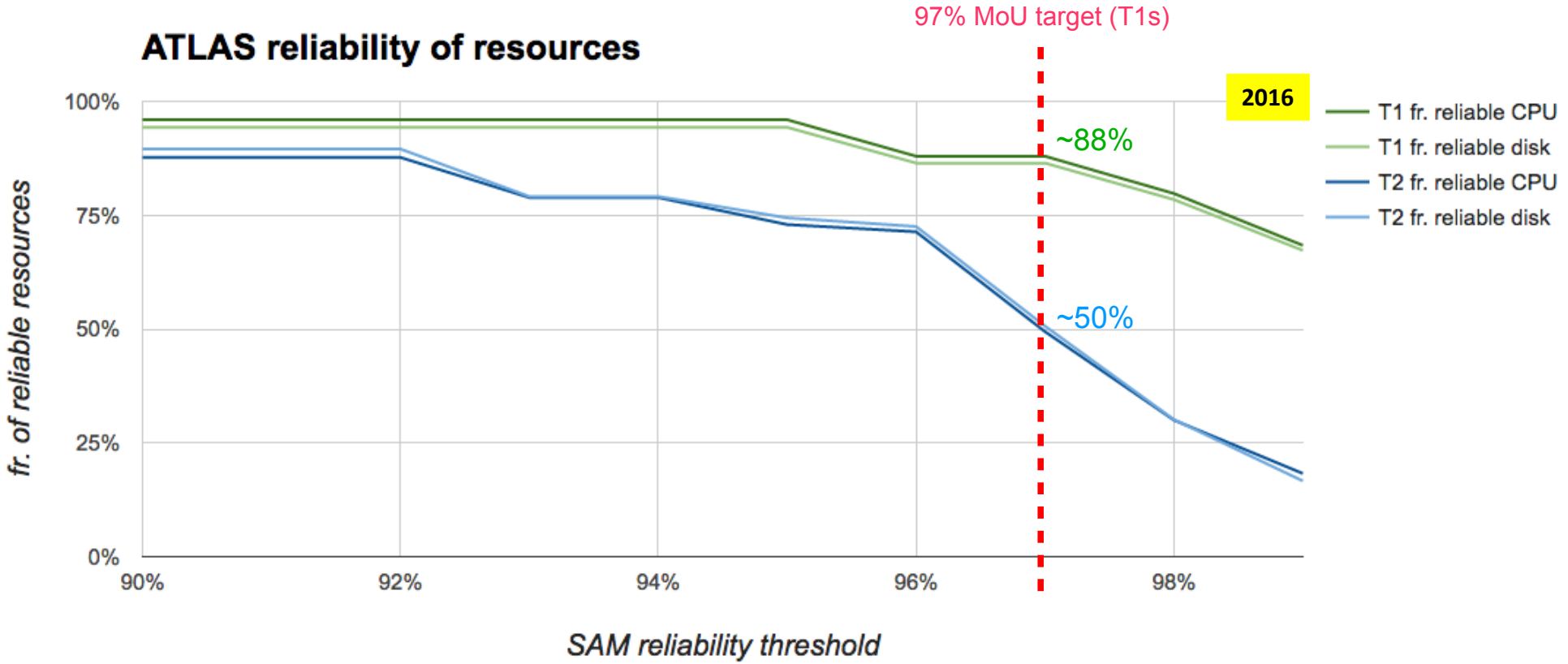


PIC
port d'informació
científica

Reliability of sites wrt. size 1/2

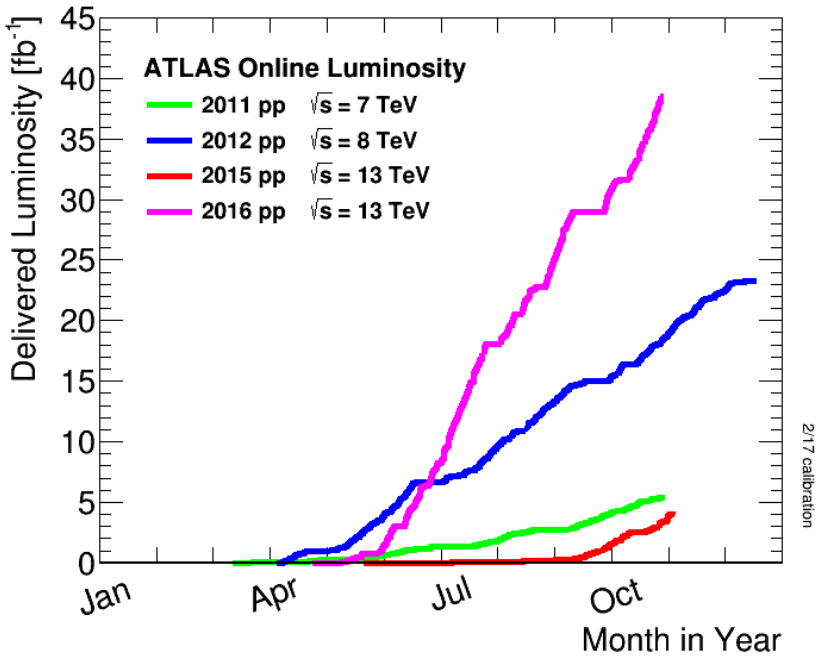


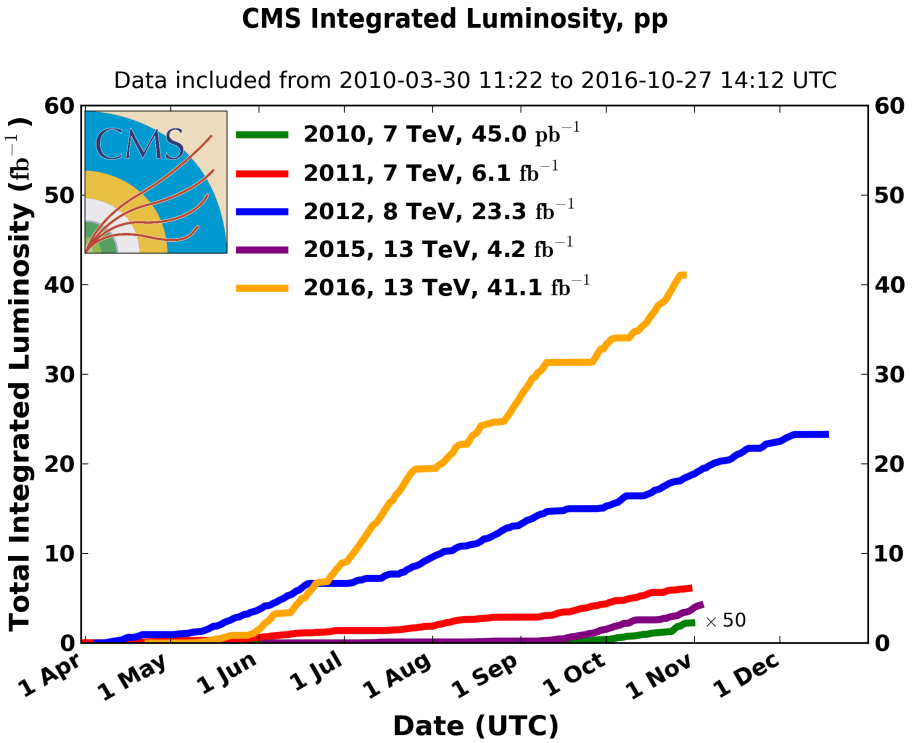
Reliability of sites wrt. size 2/2



- The Tier-1 sites are typically very reliable
- Reliable (big) Tier-2 sites around (not checked – but improved in time)

2016 LHC performance → 2017 requests



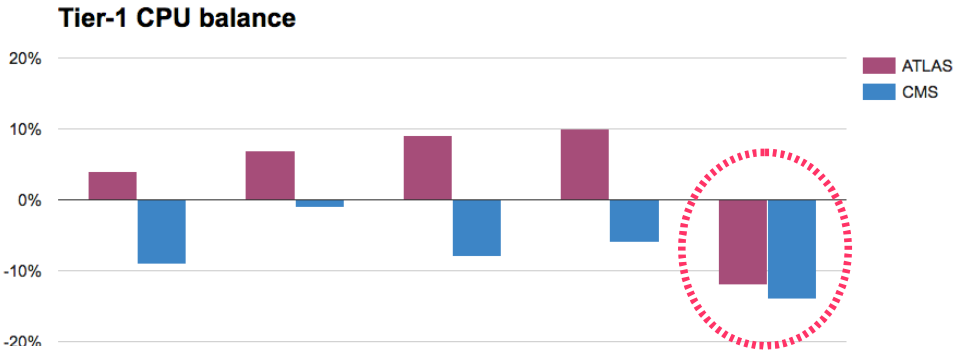


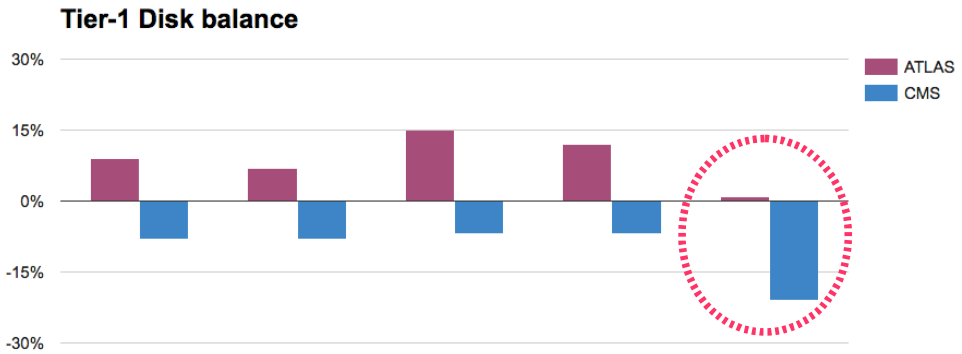
- In Summer 2016 LHC **exceeded** design luminosity by >30%

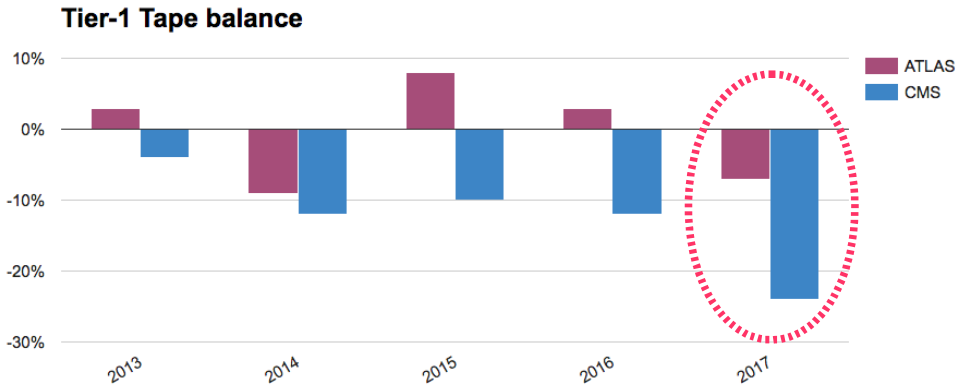
→ more data! :) → more computing requests needed! → more costs! :(
→ Mitigations done by the experiments → But, ~+20% additional requests 2017
→ Similar LHC performance expected for the rest of Run2 → impacts 2018

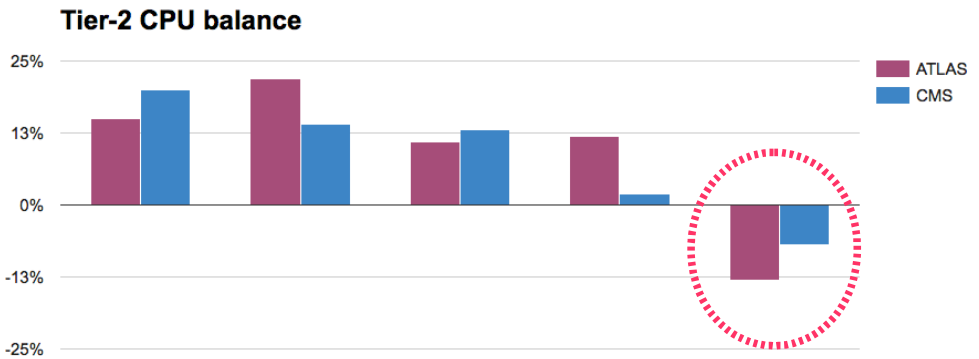
PIC
port d'informació
científica

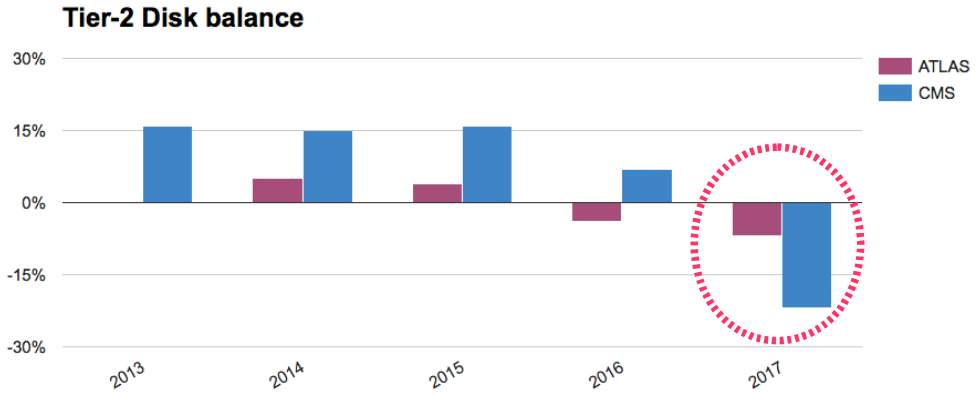
2017 site pledges wrt. Exp. requests







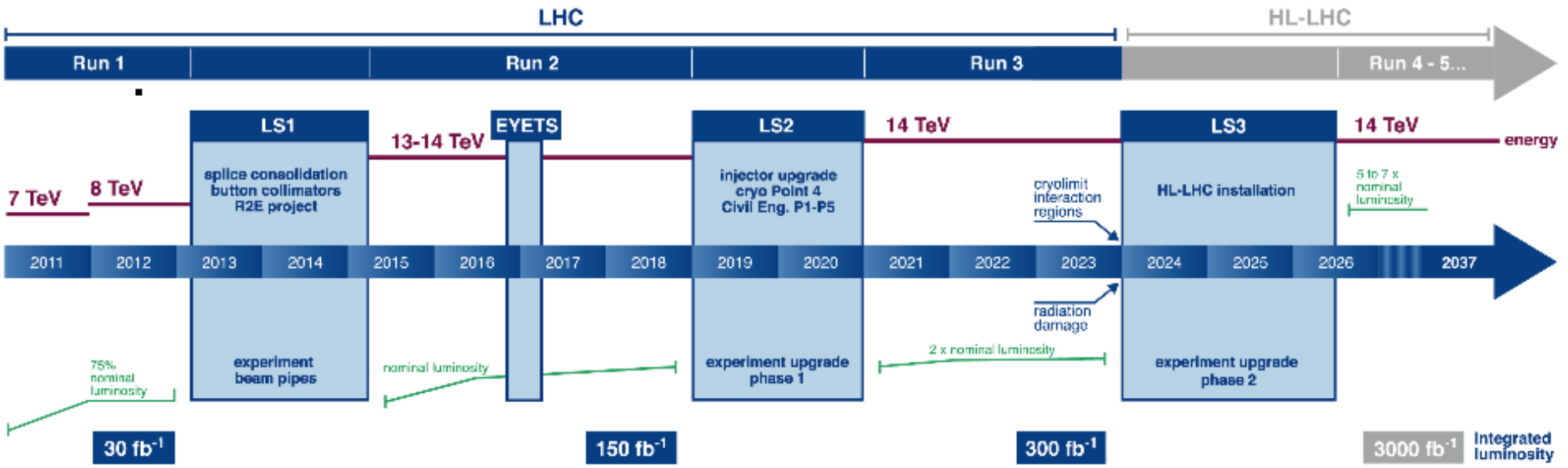




Flat budgets for computing are here... most likely to stay!

PIC
port d'informació
científica

Run3 and HL-LHC



Technology improvements (~20%/year) brings x6-x10 in 10-11 years
With the expected HL-LHC operating parameters and these improvements we expect needs ~x10 above the 'flat-budget' scenario

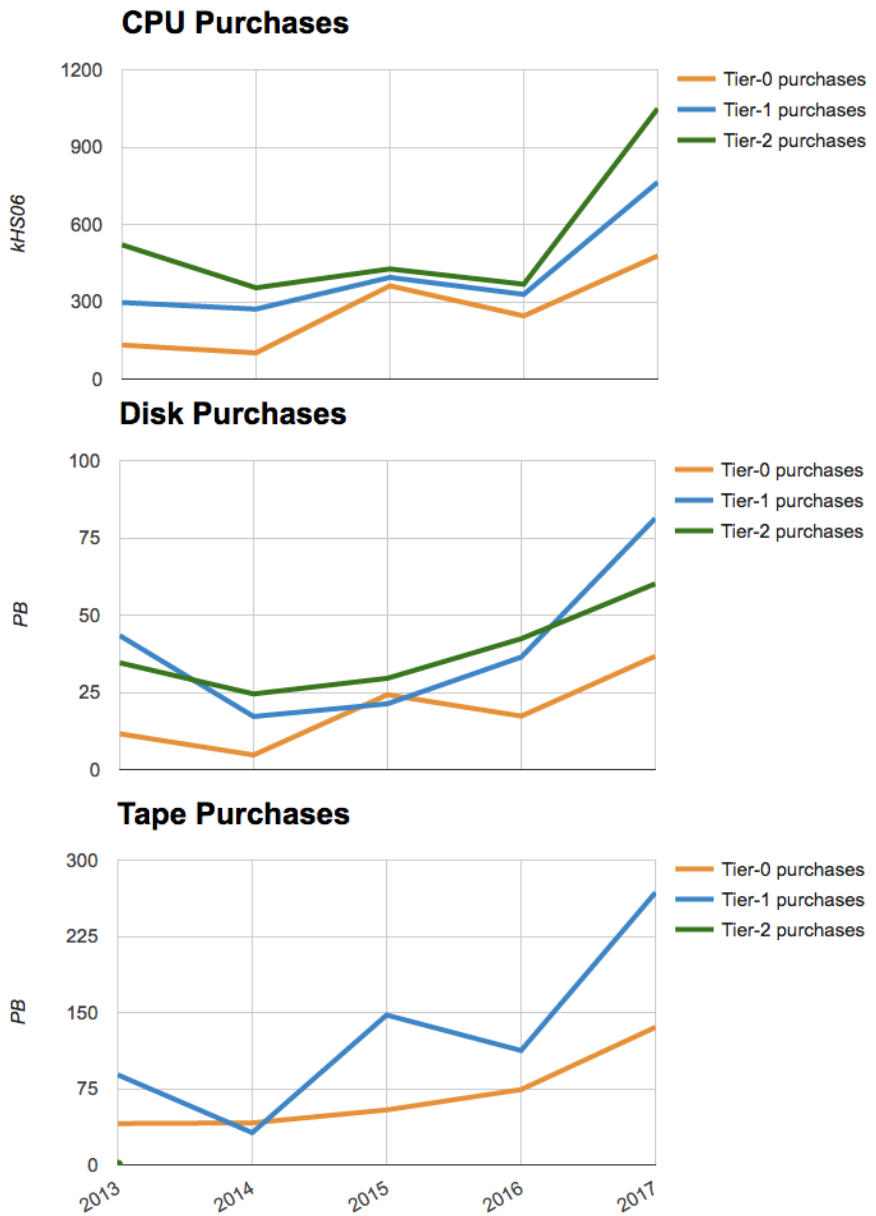
Big gap that won't be fulfilled by technology alone

I. Bird – 21/09/2016 (LHCC)

Disclaimer

Next slides describe my own
Toy model for WLCG costs
(Blame on me!)

Cost 'toy' model for WLCG 1/7



4 years equipment life-cycle (CPU/Disk)

No tape storage migrations

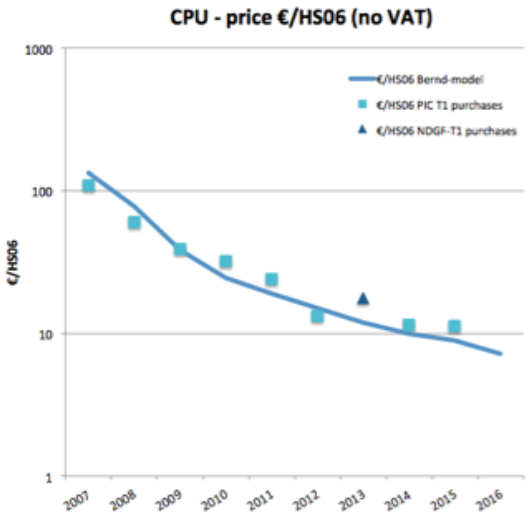
Pledges profiles growth

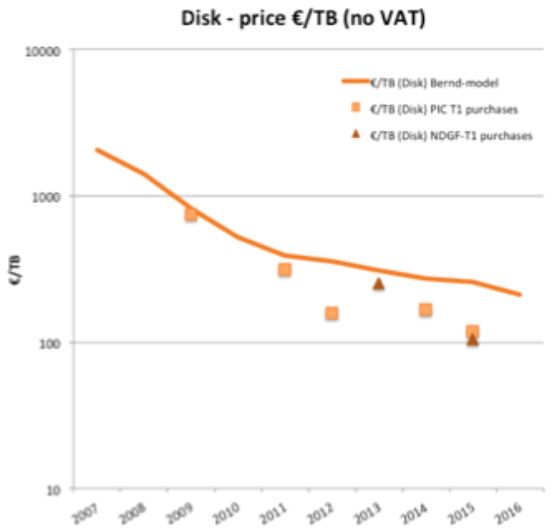
Resources purchases profiles

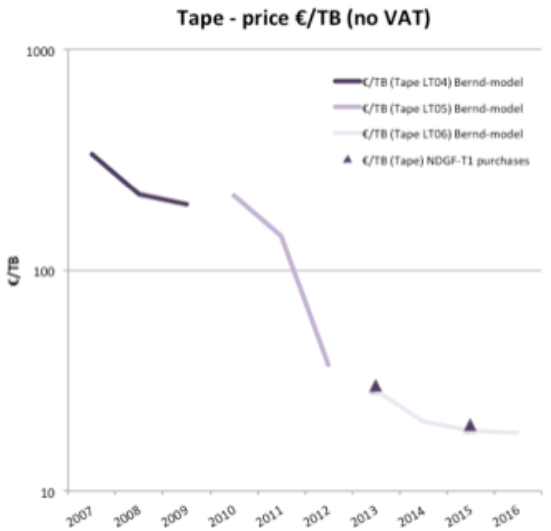
Cost 'toy' model for WLCG 2/7

- Technology evolutions: Bernd-Panzer models
- Resources costs estimations over time
  → combining with the purchases growth profiles → growth cost

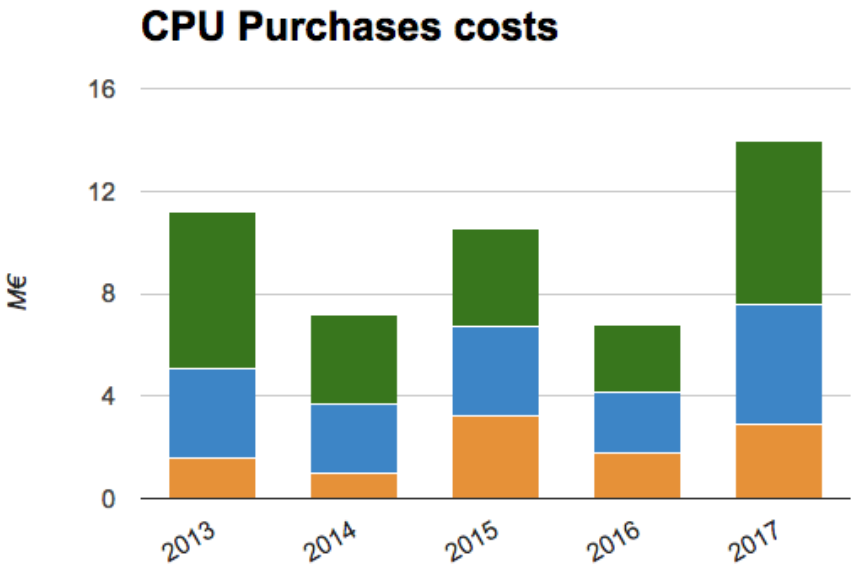
	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
€/HS06	133.4	78.3	38.2	24.4	19.1	15.1	11.8	9.9	8.9	7.2
€/TB Disk	2037	1412	824.5	518.1	390.2	355.4	308.9	272.7	257.0	211.0
€/TB Tape	335.4	220.1	198.7	217.4	142.3	37.2	28.5	20.7	18.7	18.3

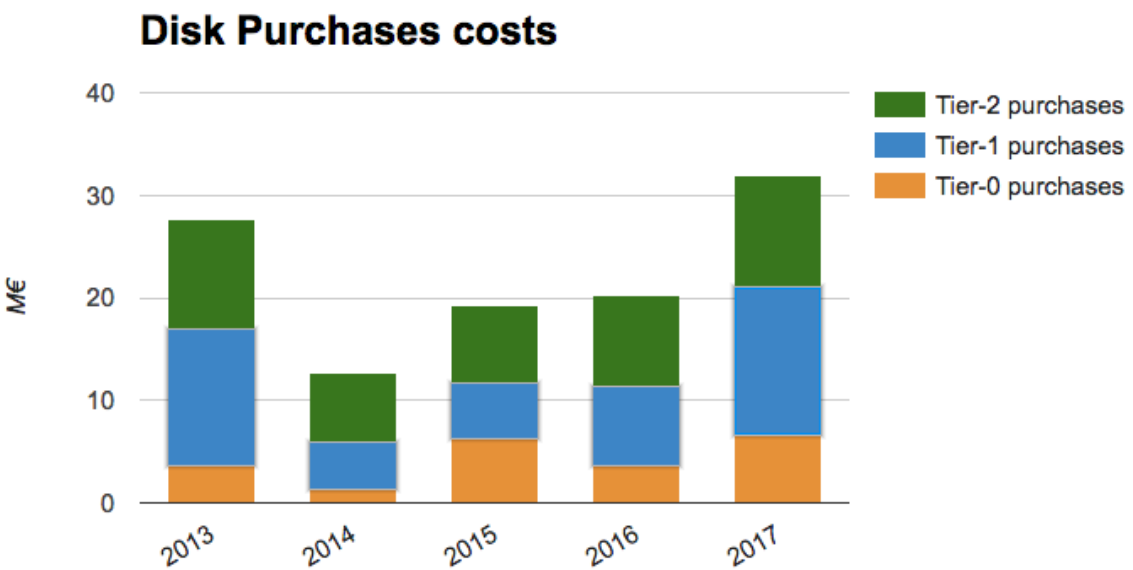


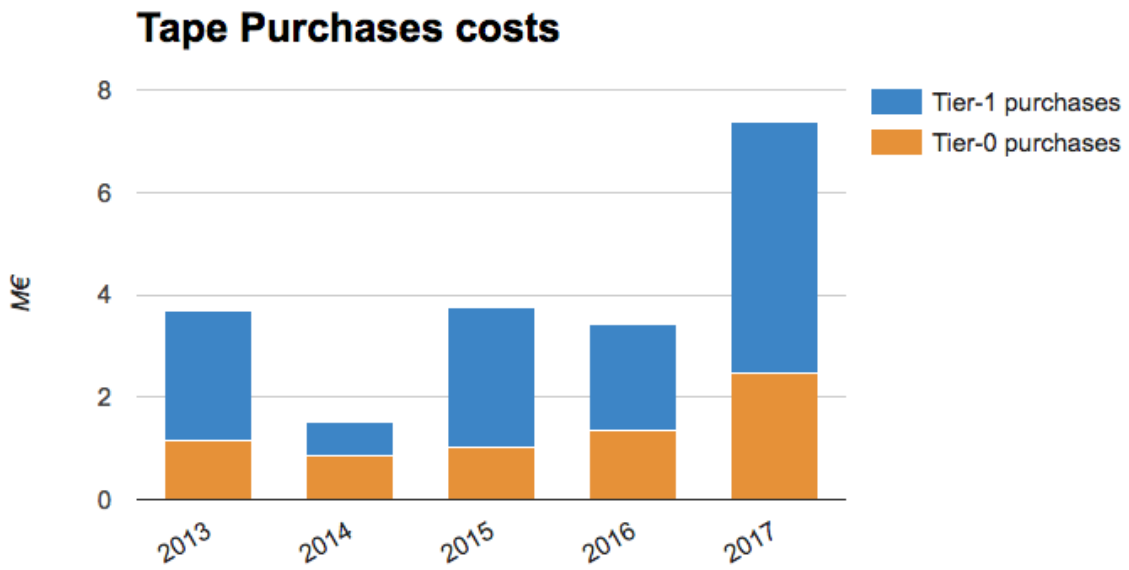




Cost 'toy' model for WLCG 3/7







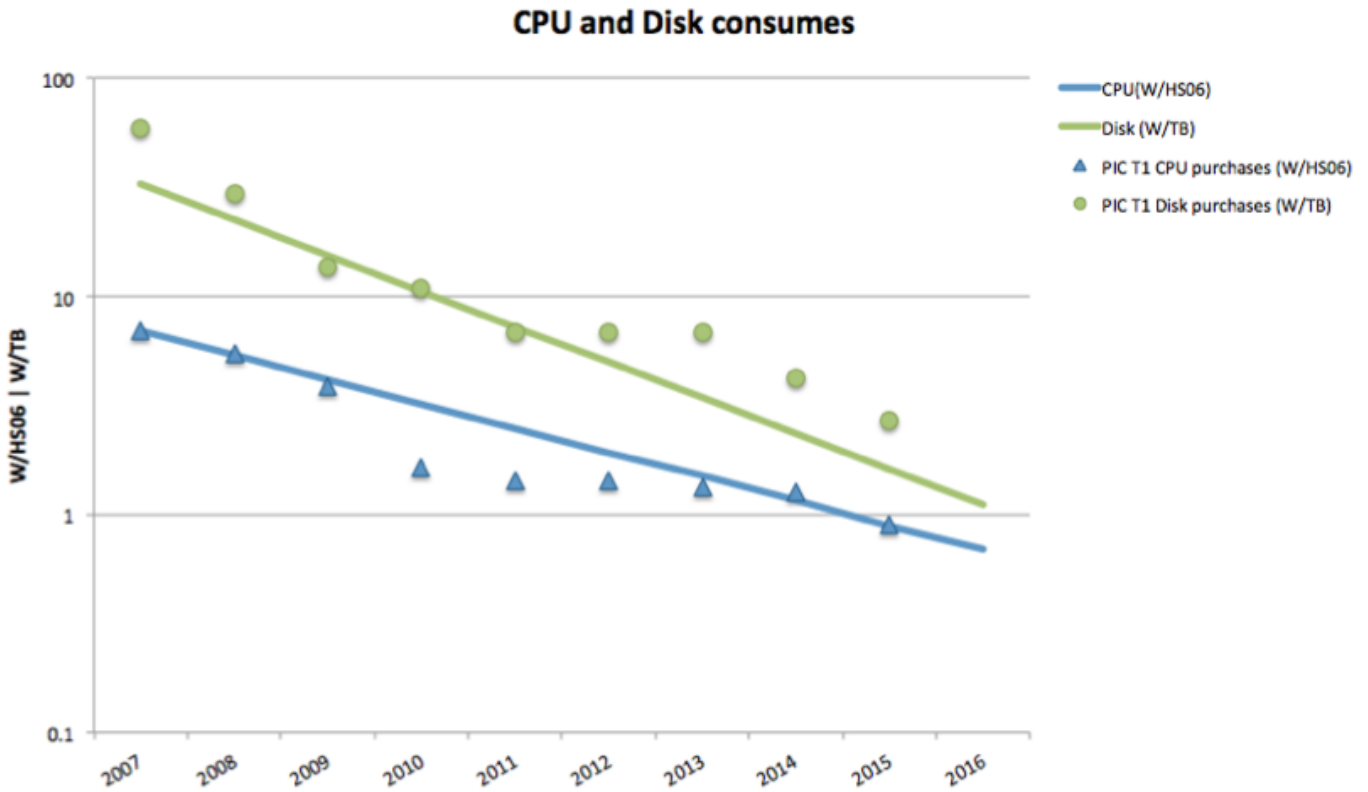
Tier-1

CPU: ~3.3 M€/year
DISK: ~9.2 M€/year
TAPE: ~2.6 M€/year

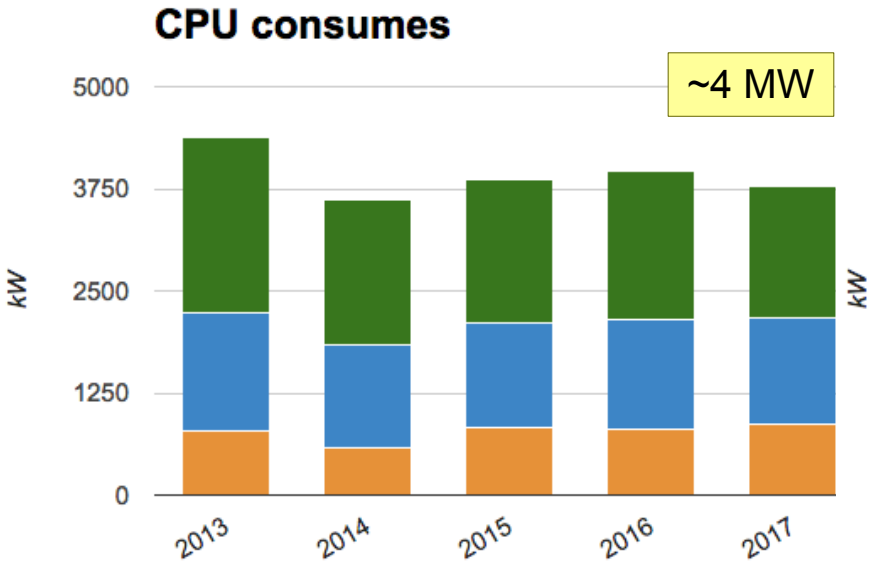
average

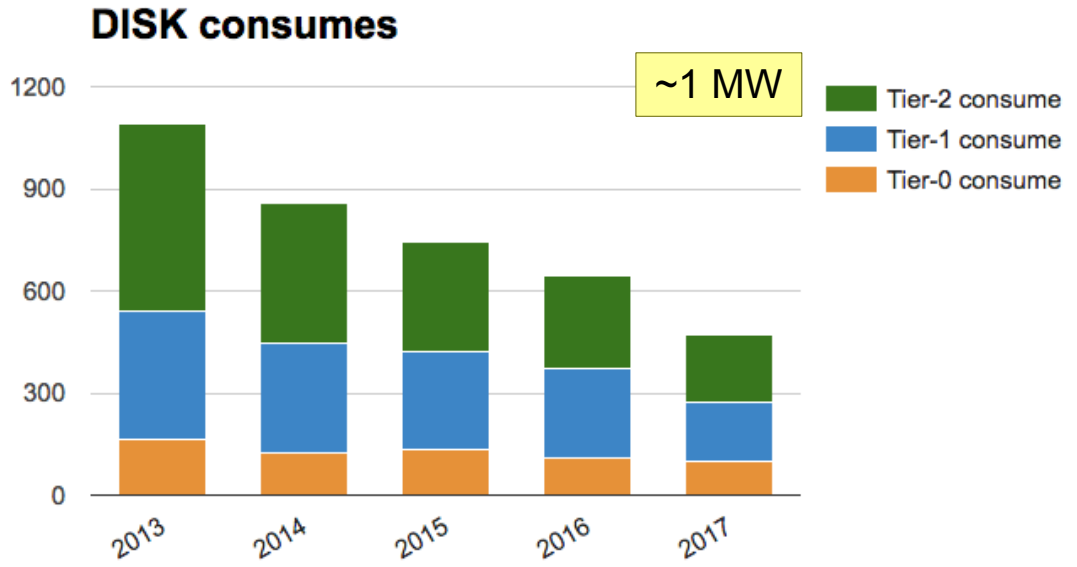
Cost 'toy' model for WLCG 4/7

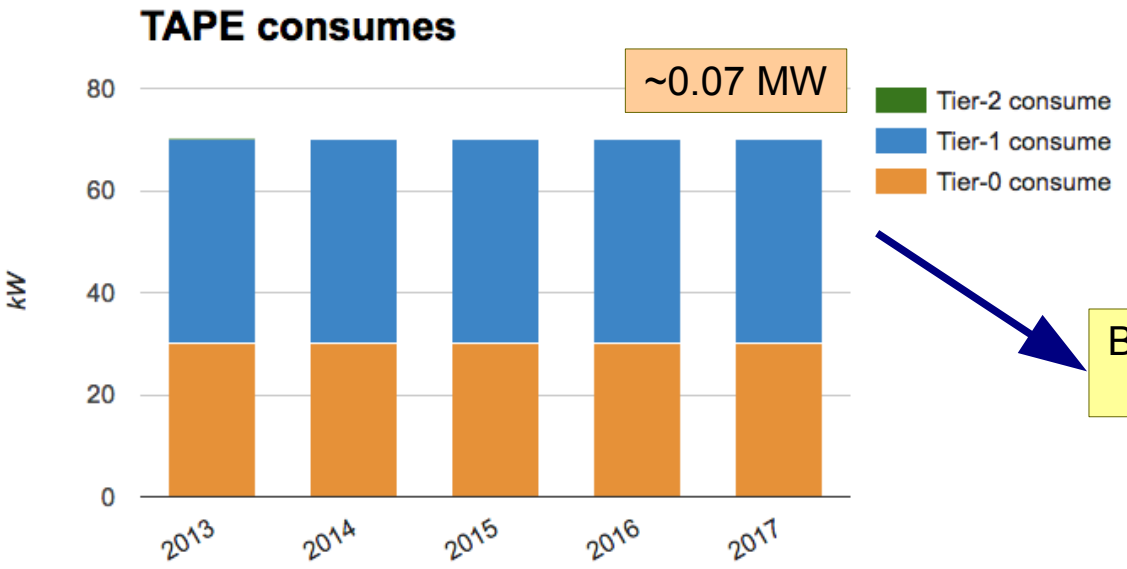
- Taking into account the purchases per year, and their consumes, we can estimate the total consume to operate CPU, Disk and Tape resources
  → Based on data from purchases made at PIC Tier-1...



Cost 'toy' model for WLCG 5/7



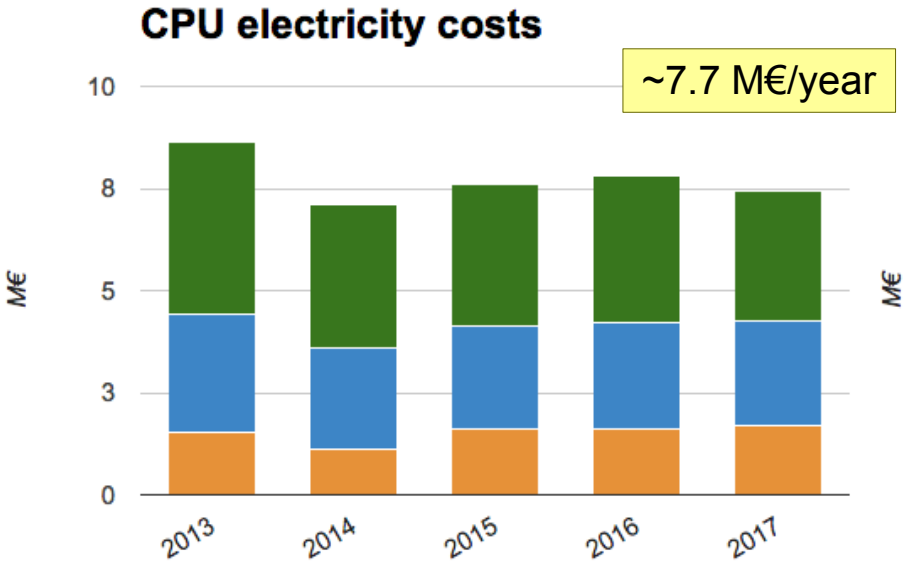


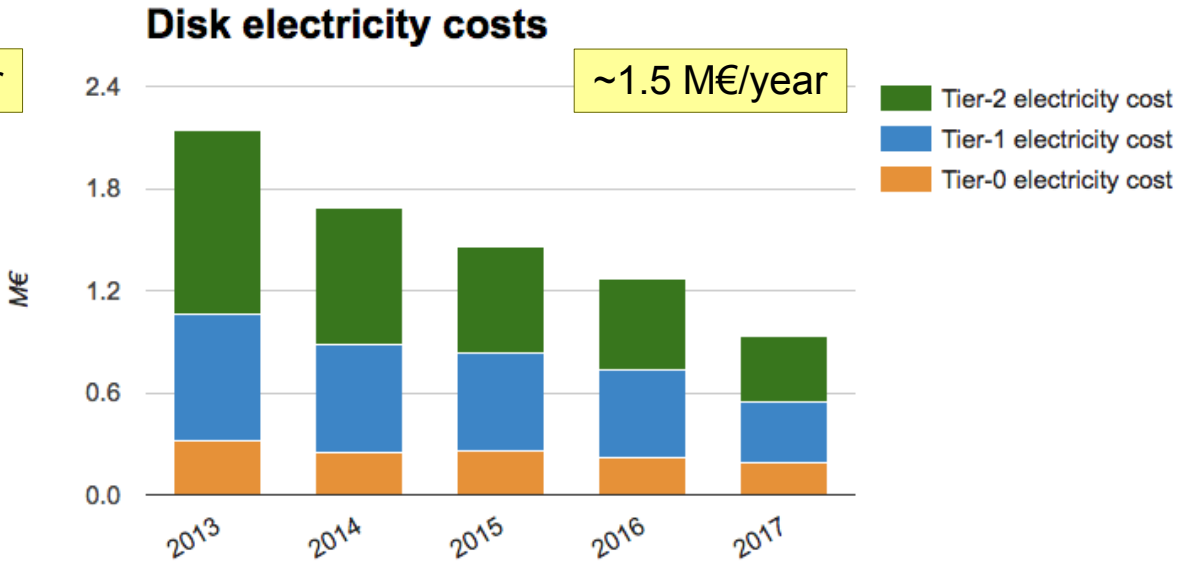


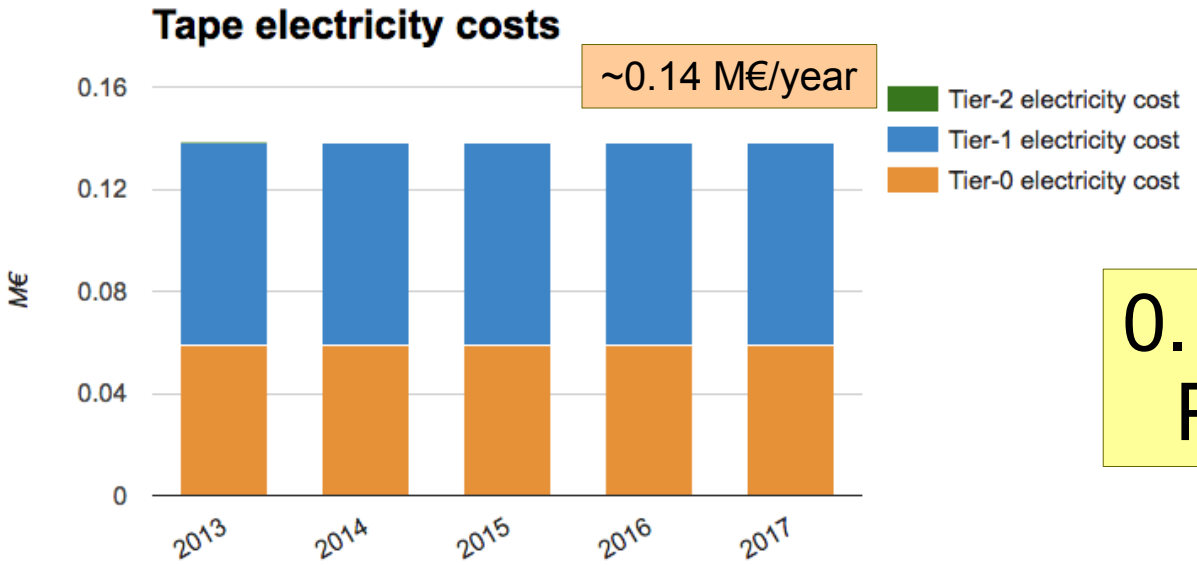
Rough estimation
Extrapolated from
PIC consumes...

But in any case, these are negligible...

Cost 'toy' model for WLCG 6/7

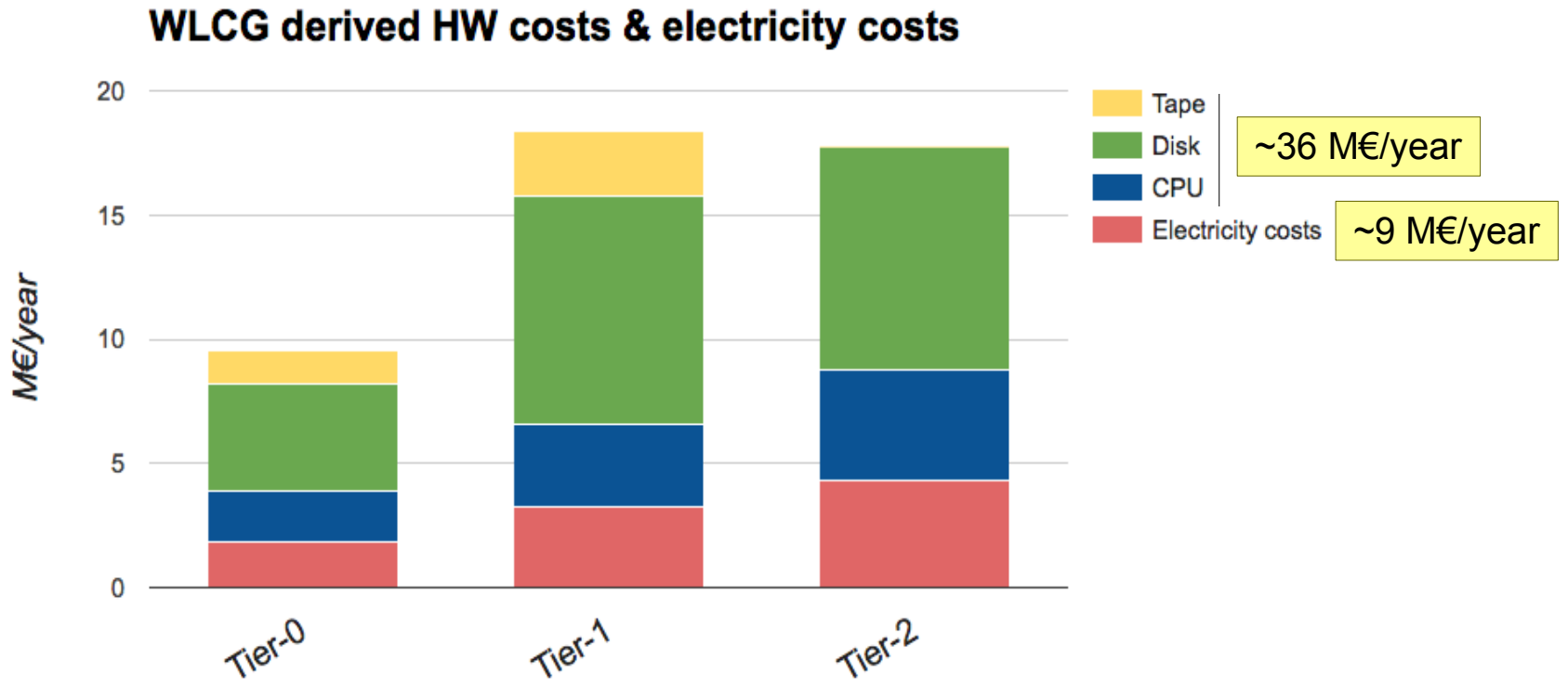






0.15 €/kWh
PUE 1.5

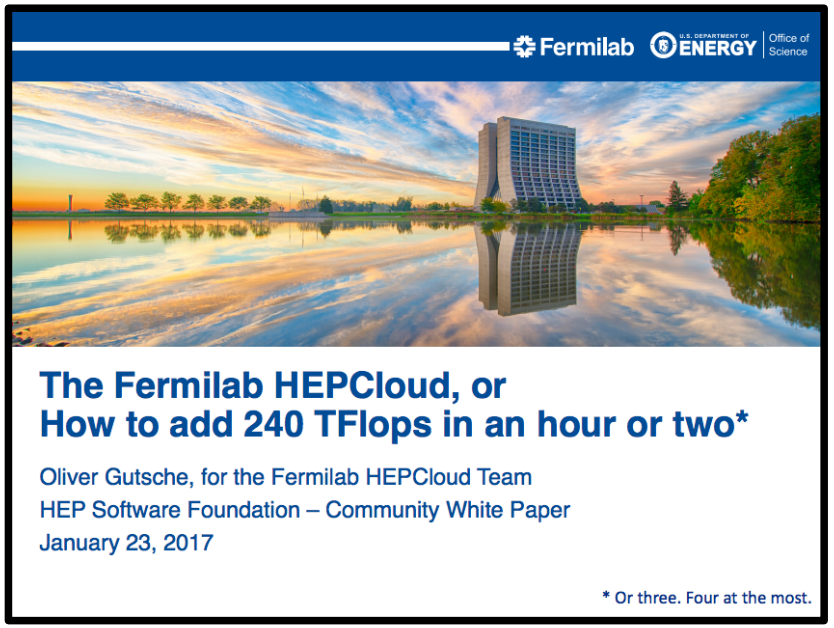
Cost 'toy' model for WLCG 7/7



- This 'toy' model does not include NREN/RREN costs
- From "Optimising costs in WLCG operations" (2015 J. Phys.: Conf. Ser. 664 032025)
  → 12.5 (3) FTEs to operate a Tier-1 (Tier-2)
  → Assuming 50 k€/FTE → manpower costs = 32 M€/year
- From EU e-FISCAL study: **1:1:1** (resources:infr./electricity/running costs:personnel)
  → This 'toy' model is yields **WLCG cost (excluding network) ~100M€/year**

Cost comparisons to Clouds

- Check O. Gutsche HEPCloud at the HSF Workshop @San Diego (January 2017):
https://indico.cern.ch/event/570249/contributions/2423184/



FNAL on-premises cost: **$0.009 core-hour**
AWS: **$0.014 core-hour**
GCP: **~$0.01 core-hour** (60h/150kcores/100k$)
(my rough estimation)

- Commercial clouds offering competitive resources at decreased cost compared to the past
- From the 'toy' model presented here → $core-hours for WLCG on-premises resources
  → taking into account the CMS CPU costs + infr./manpower shares
    - CPU consumes lot of electricity
    - less manpower needs than storage
  → toy-model: CPU cost ~$0.008 core-hour

Clouds are at <x2 factors (+50%/+75%)

PIC
port d'informació
científica

(personal) thoughts for evolution & challenges

Next 10 years

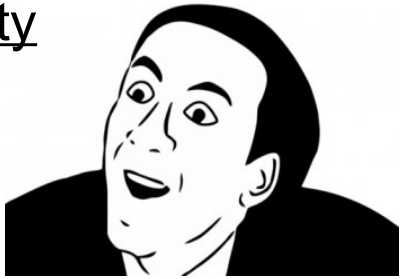
The first generation iPhone was released on June 29, 2007 (in US)

The original operating system for the original iPhone was **iPhone OS 1**, marketed as OS X, and included Visual Voicemail, multi-touch gestures, HTML email, Safari web browser, threaded text messaging, and YouTube. However, many features like MMS, apps, and copy and paste were not supported at release, leading hackers jailbreaking their phones to add these features. Official software updates slowly added these features.

iPhone OS 2 was released on July 11, 2008, around the same time as the release of the iPhone 3G, and introduced third-party applications, Microsoft Exchange support, push e-mail, and other enhancements.

iPhone OS 3 was released on June 17, 2009, and introduced copy and paste functionality

iPhone OS 3 was released on June 17, 2009, and introduced copy and paste functionality



(personal) thoughts for evolution & challenges

Next 10 years

Impossible to fit HL-LHC into the current model: **WLCG needs a (r)evolutionary solution**

Evolution to big sites (economies of scale, less manpower needs), well connected, holding the data (responsibility reasons)? Infrastructure capable to elastically growth into diverse commercial/community clouds, HPCs, HLT farms, other 'Grid' sites (with caches)
→ challenging for planning and procurement processes, indeed
→ Network to commercial cloud providers and HPCs might be an issue:
- effort for one NREN? Across global NRENs? Bandwidth? Costs? (shared - global)
→ we do science: many sociological aspects involved (and political) in this global challenge

LHC Computing = **Data Intensive Science** - not all of the workflows types could be outsourced

Trigger-less DAQs – data alignment, calibration, (even) fast data reprocessing close to the detectors? (real-time processing) **Reduced data from T0**? Simplifies data management needs

Adoption of **Big Data tools** for the users (Hadoop/Python Notebooks): PBs → TBs

Exponential increase of **network** bandwidth use (ESnet traffic ~1EB/month in 2021)
→ insufficient or unreliable network might severely impact workflows – Tbps connections
→ many technical challenges: not to provision for peaks (SDNs) (factor x6 improvement)

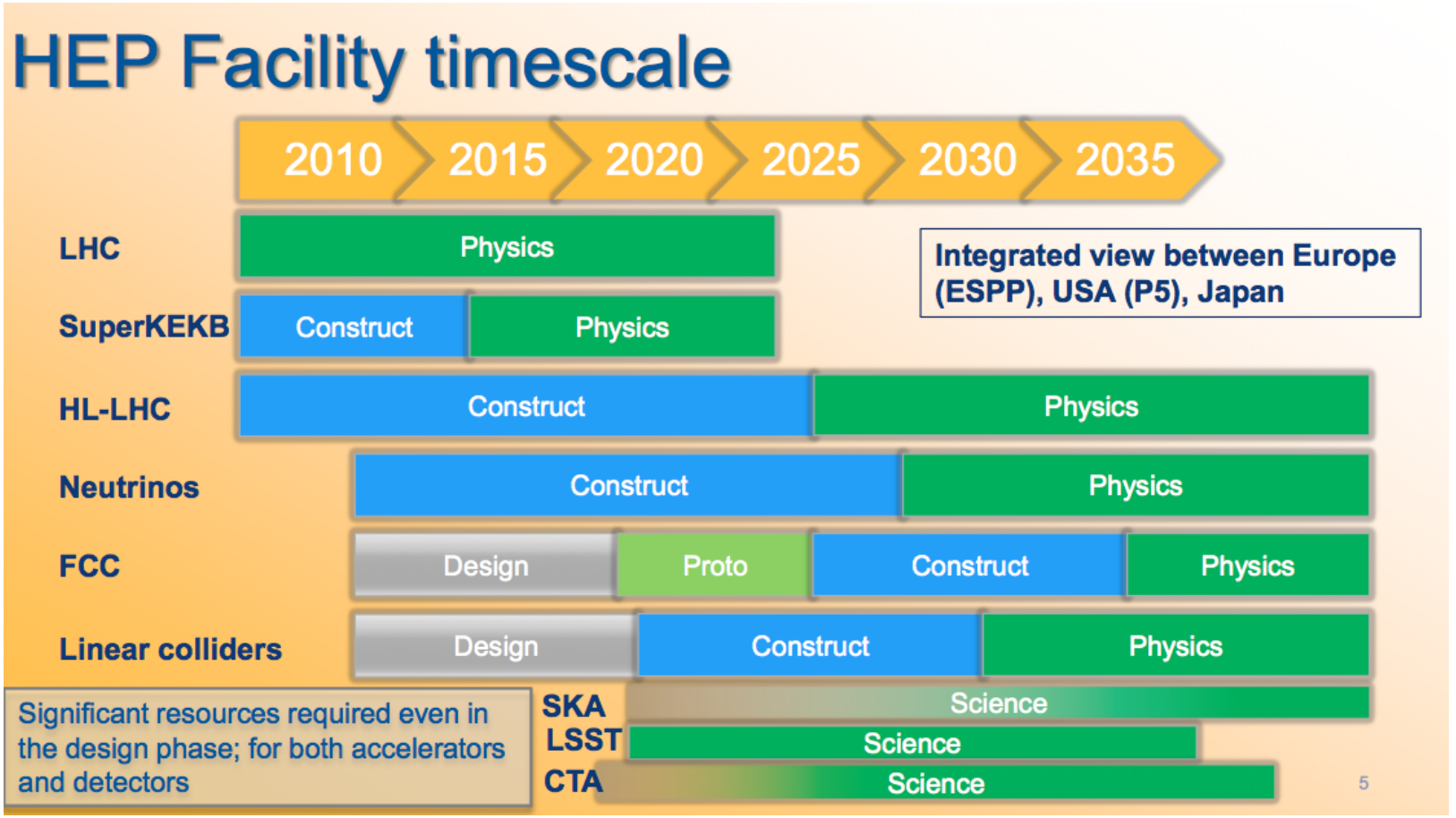
Tape market evolution? Adoption of **tiered storages**?

(personal) thoughts for evolution & challenges

Next 10 years

We would need to perform **many improvements** to reduce costs for the future

→ At all levels: software, tools/services, models, infrastructure...
→ HSF White Paper ; Computing TDR
→ Competition with other sciences to occur – HEP-wide computing collaborative environment?



PIC
port d'informació
científica

Conclusions

In June 2017 – 10 years since the first generation iPhone was launched, with built-in apps, and no copy/paste 'feature' available...

PIC
port d'informació
científica

Conclusions

In June 2017 – 10 years since the first generation iPhone was launched, with built-in apps, and no copy/paste 'feature' available...

PIC
port d'informació
científica

Conclusions

In June 2017 – 10 years since the first generation iPhone was launched, with built-in apps, and no copy/paste 'feature' available...

As of today, we have >2 million distinct apps in Apple Store and Google Play, and we have more mobile devices registered than human beings in the planet

PIC
port d'informació
científica

Conclusions

In June 2017 – 10 years since the first generation iPhone was launched, with built-in apps, and no copy/paste 'feature' available...

As of today, we have >2 million distinct apps in Apple Store and Google Play, and we have more mobile devices registered than human beings in the planet

2007

2017

I cannot answer what a Tier-1 (or WLCG) will look like in ten years from now, but for sure the path is going to be really interesting and challenging!