WLCG-ES resources and exploitation

Santiago González de la Hoz

Institut de Física Corpuscular (IFIC) – València

on behalf of the WLCG-ES sites

1st Workshop de Computing y Software de la Red Española del LHC

(28-29 April 2021)

Outline

- WLCG-ES sites
- Site Resources & Performance (Tiers, HPCs, Clouds, ….)
- Next Years (Run3)
- Future Perspectives (HL-LHC)
- Conclusions

Red LHC España

WLCG-ES sites

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

UB Universitat de Barcelona

IFCA

Ciemat

IFAE Institut de Física d'Altes Energies

IFIC INSTITUT DE FÍSICA CORPUSCULAR

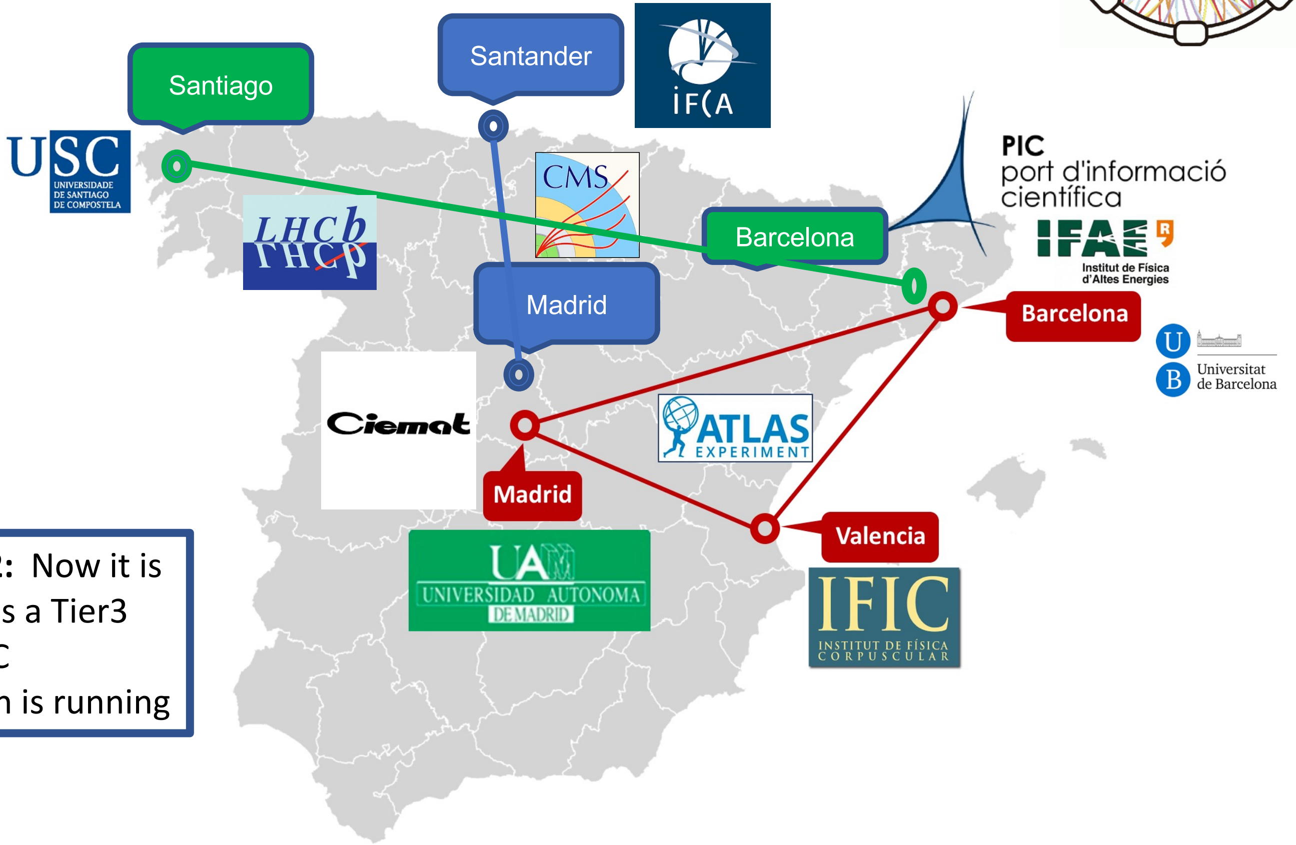
UAM UNIVERSIDAD AUTONOMA DE MADRID

PIC port d'informació científica

WLCG-ES sites (Spanish Tiers for the LHC)

- Tier 1 (ATLAS, CMS, LHCb):
  - **PIC-Barcelona**
- Federated Tier2s
  - **60% IFIC-Valencia**
  - **25% IFAE-Barcelona**
  - **15% UAM-Madrid**
  - **75% Ciemat-Madrid**
  - **25% IFCA-Santander**
  - **50% USC-Santiago**
  - **50% UB-Barcelona***

***UB Tier2:** Now it is working as a Tier3 where MC simulation is running


Santiago
Santander
IFCA
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
CMS
LHCb
Barcelona
Madrid
PIC port d'informació científica
IFAE Institut de Física d'Altes Energies
Barcelona
Universitat de Barcelona
Ciemat
ATLAS EXPERIMENT
Madrid
Valencia
UAM UNIVERSIDAD AUTONOMA DE MADRID
IFIC INSTITUT DE FÍSICA CORPUSCULAR

- LHC sites pledges in the last 5 years:

  (https://wlcg-cric.cern.ch/core/federation/list/)

- **Integrated in the WLCG project (World Wide LHC Computing GRID) and strictly following the experiments computing models.**

- **We represented the 4-5% of the total Tier-2s and 5% of the total Tier-1s resources, with the budget reduction now the 3% for Tier2s and 4% for Tier1!!!!!**

Site Resources & Performance (Tiers, HPCs, Clouds, …)

PIC resources & performances

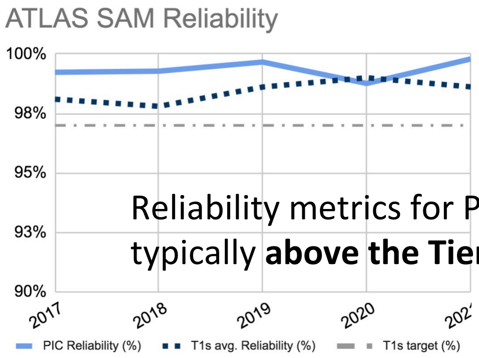
PIC
port d'informació
científica

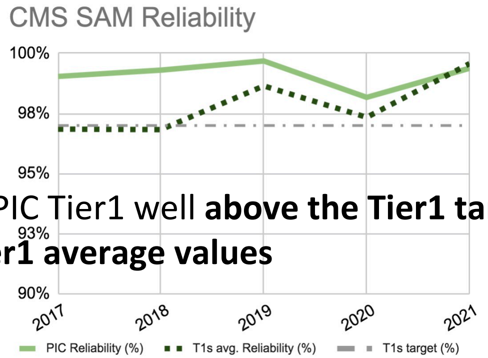
Resources usage

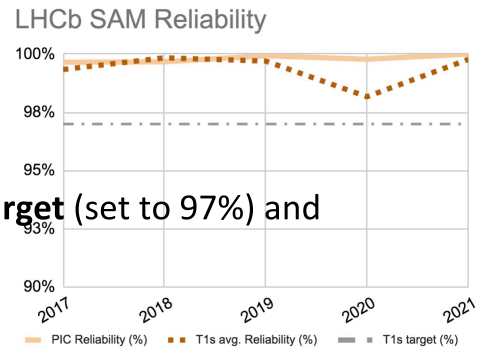
ATLAS	2017	2018	2019	2020	2021
CPU (MHS06·hours) - Grid	245.31	349.20	334.15	330.29	114.63
CPU (MHS06·hours) - BSC	-	3.32	36.06	174.04	51.80
Disk (TB)	2,428	3,266	3,404	3,500	3,473
Tape (TB)	5,683	8,124	8,889	8,184	6,687

CMS	2017	2018	2019	2020	2021
CPU (MHS06·hours)	208.32	292.19	279.48	250.36	73.12
Disk (TB)	2,118	2,803	2,799	2,894	2,691
Tape (TB)	6,299	6,677	8,902	8,981	8,438

LHCb	2017	2018	2019	2020	2021
CPU (MHS06·hours)	79.29	112.38	140.34	152.87	36.14
Disk (TB)	1,286	1,383	1,358	1,369	1,373
Tape (TB)	1,328	1,643	1,672	2,400	2,071


ATLAS SAM Reliability


CMS SAM Reliability


LHCb SAM Reliability

Reliability metrics for PIC Tier1 well **above the Tier1 target** (set to 97%) and typically **above the Tier1 average values**

Before 2017: PIC at **5.1%** share of experiment resource requests at Tier1s **(6.5% for LHCb)**

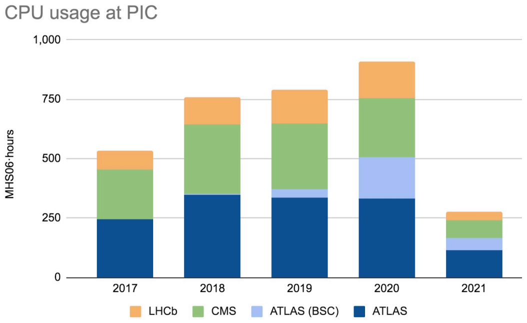
Period 2018-2020: PIC Tier1 reduces its share to **4% as requested by the Spanish funding agency**

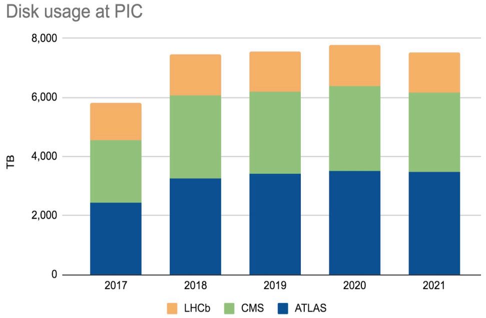
Performance

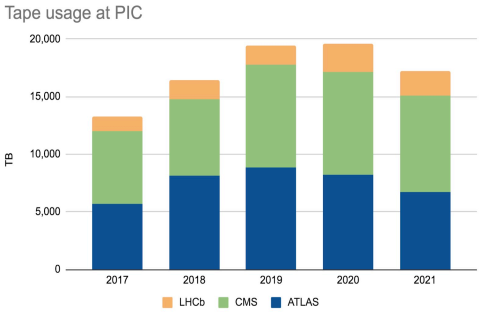
ATLAS	2017	2018	2019	2020	2021
PIC Availability (%)	98.97%	99.24%	99.61%	98.72%	99.09%
T1s avg. Availability (%)	97.21%	98.65%	97.96%	98.47%	98.43%
PIC Reliability (%)	99.23%	99.28%	99.66%	98.76%	99.79%
T1s avg. Reliability (%)	98.10%	97.80%	98.61%	99.00%	98.62%

CMS (Grid)	2017	2018	2019	2020	2021
PIC Availability (%)	98.85%	99.27%	99.60%	98.12%	99.38%
T1s avg. Availability (%)	95.57%	95.51%	97.91%	97.05%	99.56%
PIC Reliability (%)	99.04%	99.30%	99.68%	98.17%	99.38%
T1s avg. Reliability (%)	96.85%	96.83%	98.65%	97.34%	99.57%

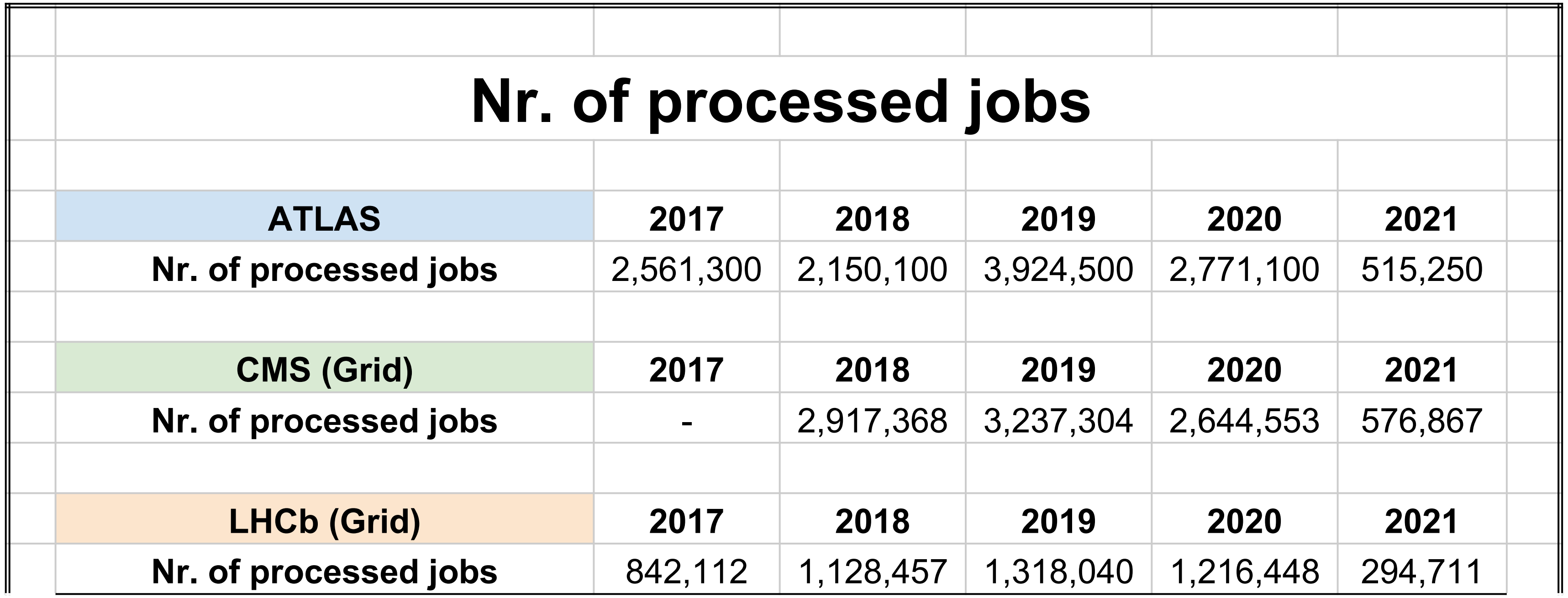
LHCb (Grid)	2017	2018	2019	2020	2021
PIC Availability (%)	99.56%	99.45%	99.69%	99.04%	99.36%
T1s avg. Availability (%)	97.02%	97.02%	98.59%	97.81%	99.30%
PIC Reliability (%)	99.65%	99.67%	99.93%	99.78%	100.00%
T1s avg. Reliability (%)	99.34%	99.84%	99.70%	98.17%	99.77%


CPU usage at PIC


Disk usage at PIC


Tape usage at PIC

PIC resources & performances



Nr. of processed jobs

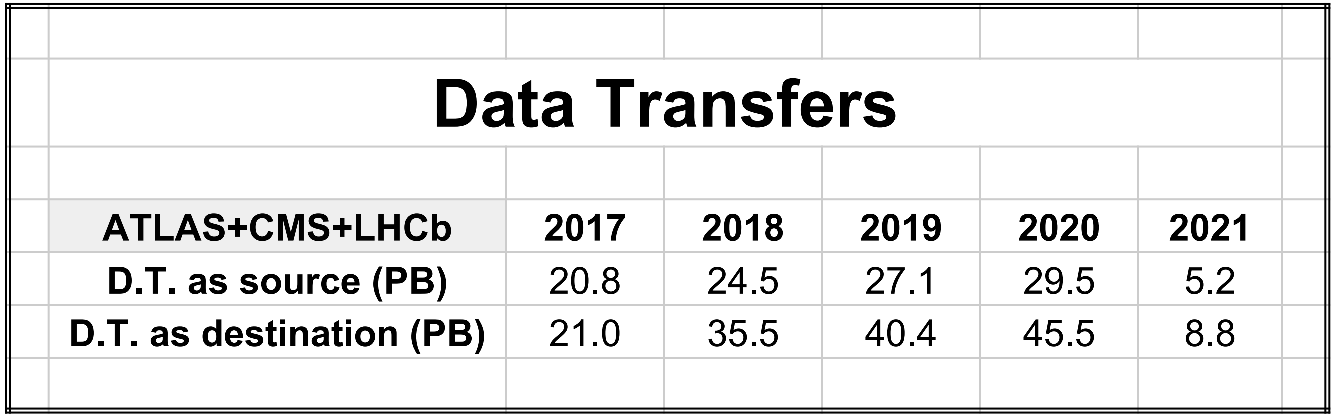
ATLAS	2017	2018	2019	2020	2021
Nr. of processed jobs	2,561,300	2,150,100	3,924,500	2,771,100	515,250

CMS (Grid)	2017	2018	2019	2020	2021
Nr. of processed jobs	-	2,917,368	3,237,304	2,644,553	576,867

LHCb (Grid)	2017	2018	2019	2020	2021
Nr. of processed jobs	842,112	1,128,457	1,318,040	1,216,448	294,711

~10 PB running on **dCache 5.2.35**

Old disk pools from 2014 to be retired next year (~1.3 PB)
Other disk pools from 2015 extended (again) warranty +1 year



Data Transfers

ATLAS+CMS+LHCb	2017	2018	2019	2020	2021
D.T. as source (PB)	20.8	24.5	27.1	29.5	5.2
D.T. as destination (PB)	21.0	35.5	40.4	45.5	8.8

Expansion of the new Tape Library

NEW

IBM TS4500

SL8500

IBM TS4500: 2 frames (L55+D55) + 8 LT08 drives

→ 4.8 PB capacity installed with cartridges LT07 M8
→ 750 TB capacity installed with cartridges LT08

This library is expected to grow to host future data

→ It will host new data and data migrated from SL8500 library (ongoing)
→ Dedicated drives, frames and cartridges installed to handle this

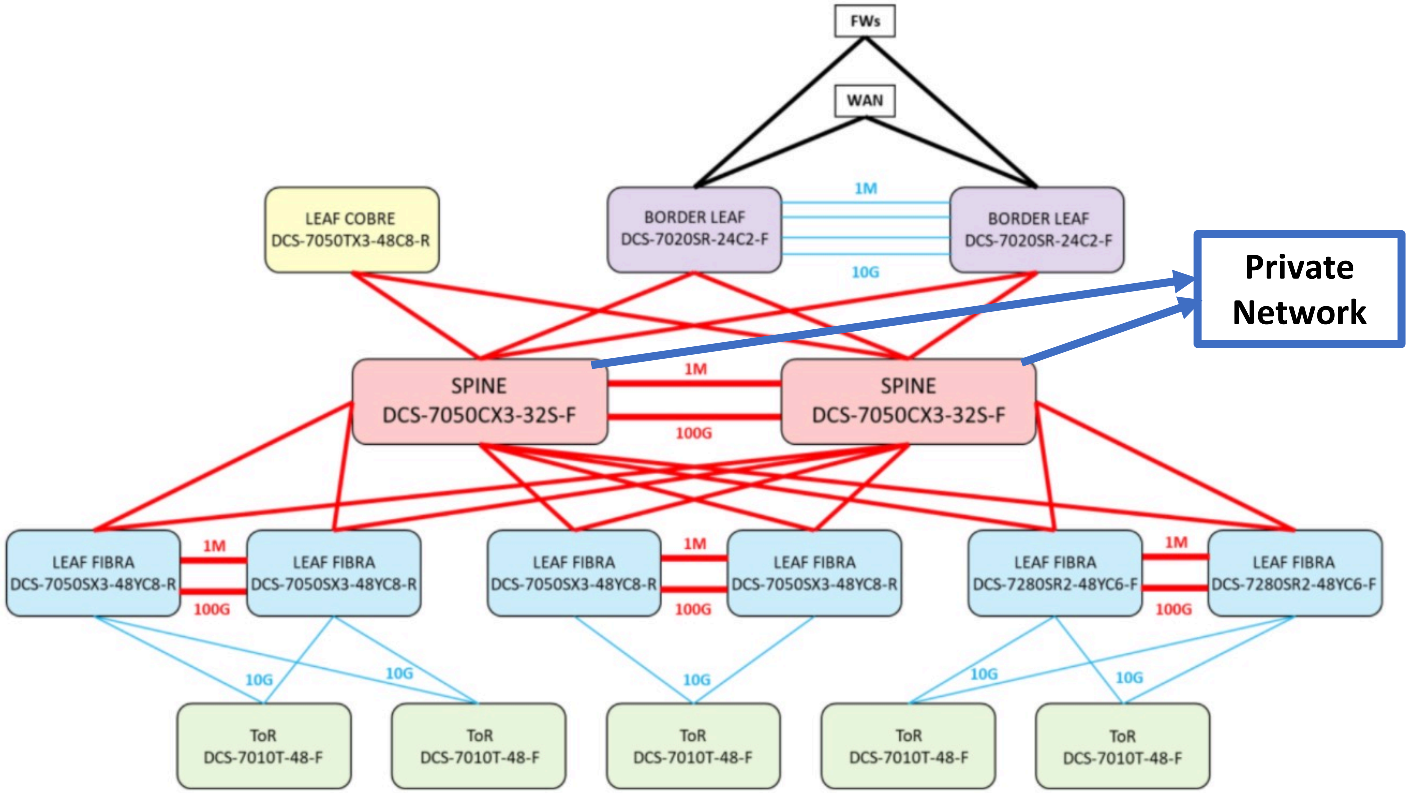
All new **CMS**, **LHCb** and **MAGIC** data go to the IBM

PIC currently runs **Enstore 6.3.4-2** (CentOS7)

All LHC data (20 PB) being migrated to the new system (for the next 2 years)

Network

- **Current 10Gbps** core network (NEXUS 7009) being upgraded to 2x100 Gbps (ARISTA)



PIC current WAN at **20 Gbps**

RedIRIS-Nova at 100 Gbps available

PIC is currently increasing its WAN connectivity to **200 Gbps** (by mid-2021)

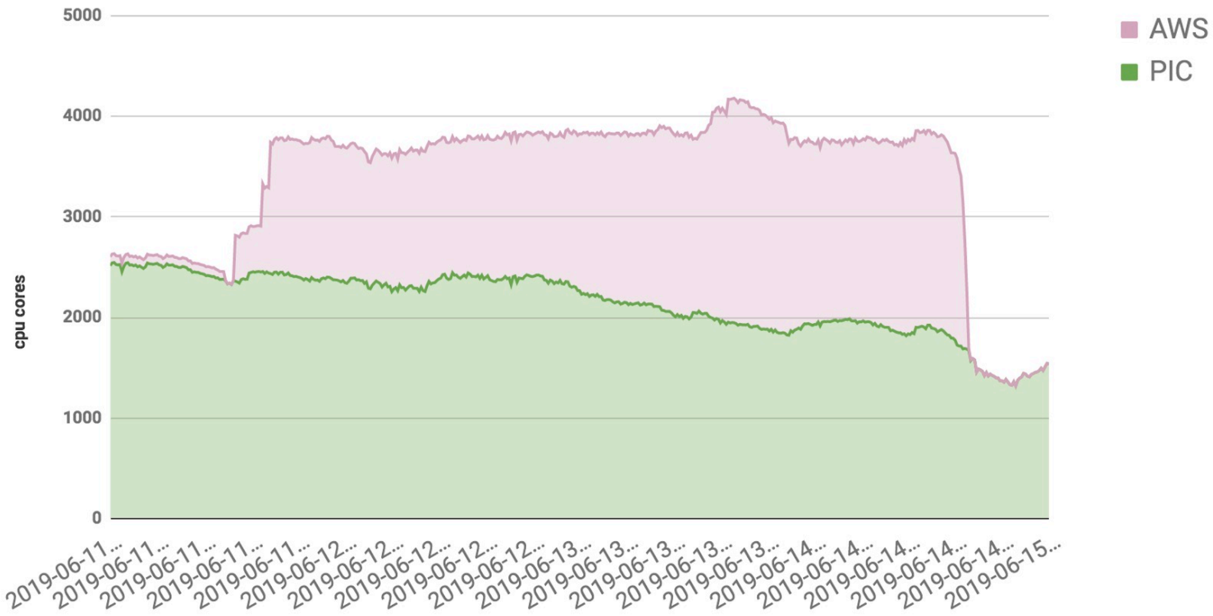
Proposal based on Leaf-Spine Network Topology

- 2x Spine → total of 64x 100 Gbps ports
- 6x Leaf → total of 288x 25 Gbps ports, and 8x uplinks of 100 Gbps

Keeping us busy for the next months - all elements expected in place before Summer 2021

Amazon - cloud bursting tests

- We tested **AWS** (Amazon Web Services) for a week (June 2019), doubling PIC compute power
- Integration of a cloud environment with the local batch system - sporadic increase of resources
- Special interest in a spot instance based scenario
- Data center in Frankfurt (**~40 ms**) - used Condor_Annex
- Set up HTCondor Connection Brokering (CCB)
- **Bridge** server to connect the local system to the outside nodes
- HTCondor-CE routing modified so only **ATLAS** and **CMS** send jobs to AWS
- Custom **WN image** deployed in AWS servers, + CVMFS, + access to Squids
- Configuration of **spot instances requirements** during the test



Good option to increase computing resources sporadically

Flexible and easy to deploy through HTCondor

Not very good for data intensive jobs

IFIC resources & performances

- Resources @ site

	2017	2018	2019	2020	2021
CPU(MHS06 .horas)	117	182	318	535	161
Disk(TB)	1872	2112	2429	2592	2236

- Performance

	2017	2018	2019	2020	2021
Availability	98.22%	97.59%	99.32%	98.78%	92.52%
Reliability	98.53%	97.99%	99.61%	98.78%	92.81%

- Number of processed Jobs

	2017	2018	2019	2020	2021
Nb proc. jobs	1,418,384	3,679,384	5,354,357	10,413,384	15,210,623

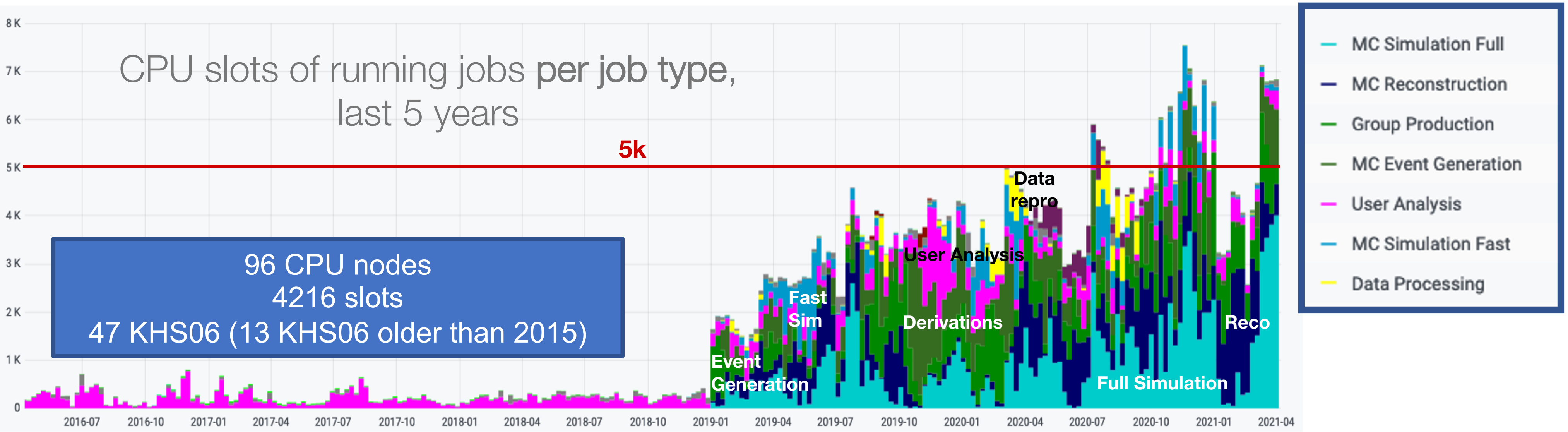
- Data transfers

	2017	2018	2019	2020	2021
D.T. as source	0.643 PB	1.874 PB	4.855 PB	10.800 PB	16.178 PB
D.T. as destination	1.44 PB	2.87 PB	5.31 PB	11.7 PB	17.44 PB

Network

- **At the top of availability and reliability ranks: IFIC Tier2 is a Nucleus**
- **Tier2s with a big amount of storage and very good network connection get elected "Nucleus",** passing job production on to smaller Tier2s (Satellites)
  - **Current network bandwidth is 2x10 Gbits/s. University of Valencia (UV) backbone is working at 100 Gbits/s.**
  - **IFIC infrastructure to work at 100 Gbist/s is ready. In the next weeks we will be ready to connect to the UV backbone.**
  - **Before summer to 100 Gbits/s expected this year thanks to REDIRIS (Spanish Academic network provider) → RedIRIS-Nova at 100 Gbps.**
  - **IFIC will increase its WAN connectivity to 100 Gbps before summer 2021.**

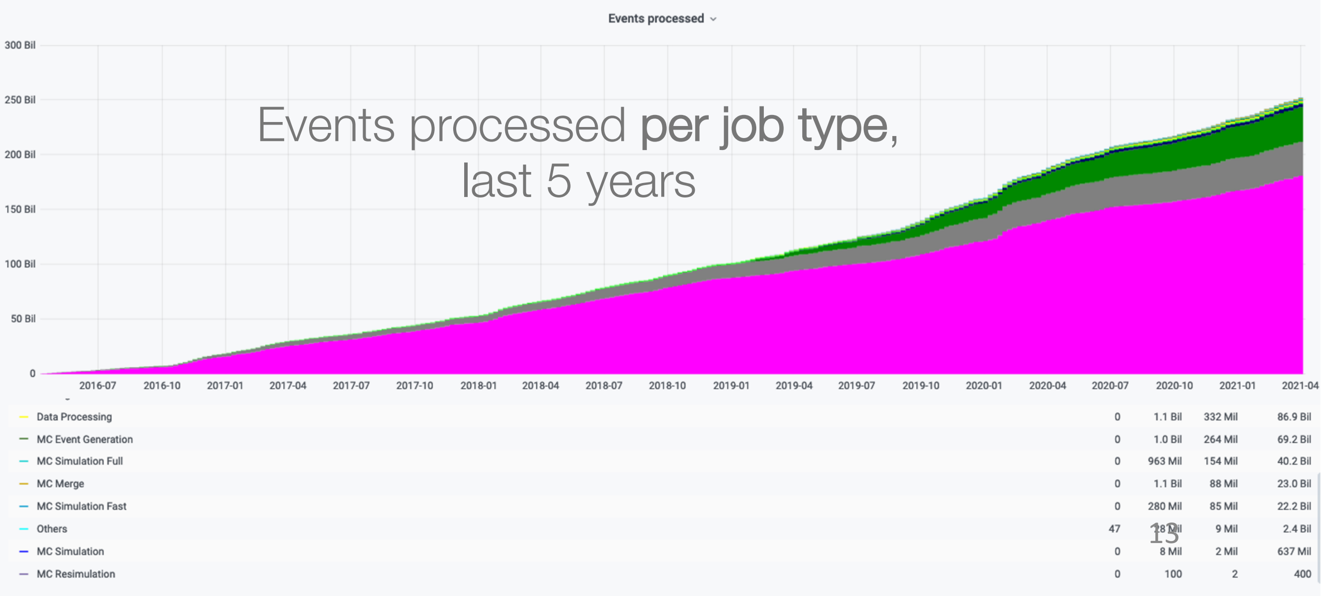
Number of jobs from 2017 to 2021



96 CPU nodes
4216 slots
47 KHS06 (13 KHS06 older than 2015)

- Steady state of more than 5.000 running job slots since 2019, typically using 2GB per job slot
- Mainly running with either 8 or 1 cores ("multi-core" or "single-core") per job, depending on type of job
- Variety of job types, where number depends on the current focus of ATLAS activities

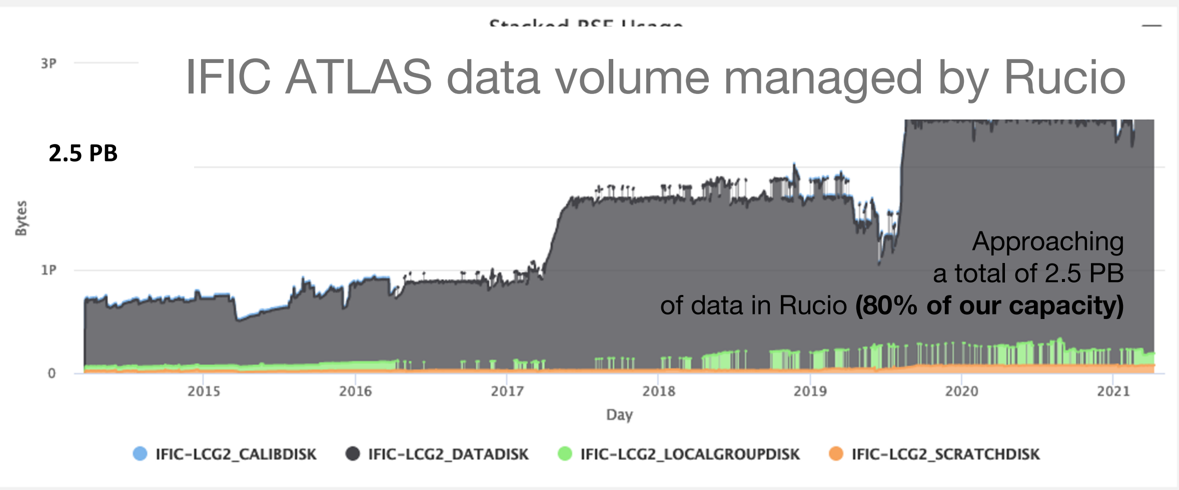
- We have processed more than 250 Bill. events in these last 5 years



Data Processing	0	1.1 Bil	332 Mil	86.9 Bil
MC Event Generation	0	1.0 Bil	264 Mil	69.2 Bil
MC Simulation Full	0	963 Mil	154 Mil	40.2 Bil
MC Merge	0	1.1 Bil	88 Mil	23.0 Bil
MC Simulation Fast	0	280 Mil	85 Mil	22.2 Bil
Others	47	28 Mil	9 Mil	2.4 Bil
MC Simulation	0	8 Mil	2 Mil	637 Mil
MC Resimulation	0	100	2	400

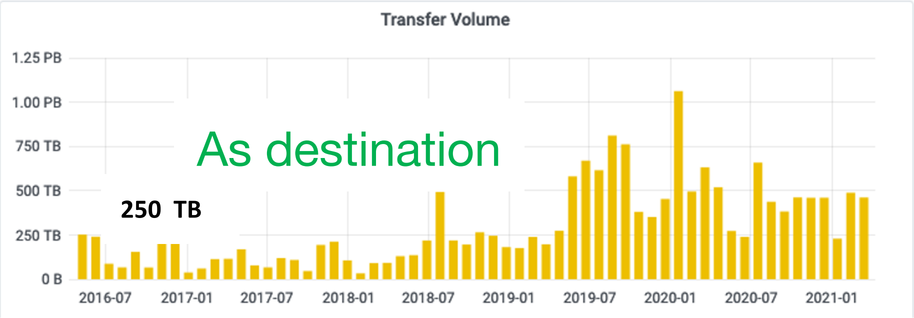
Data Transfers from 2017 to 2021

- ATLAS has a large amount of data – and we move it a lot

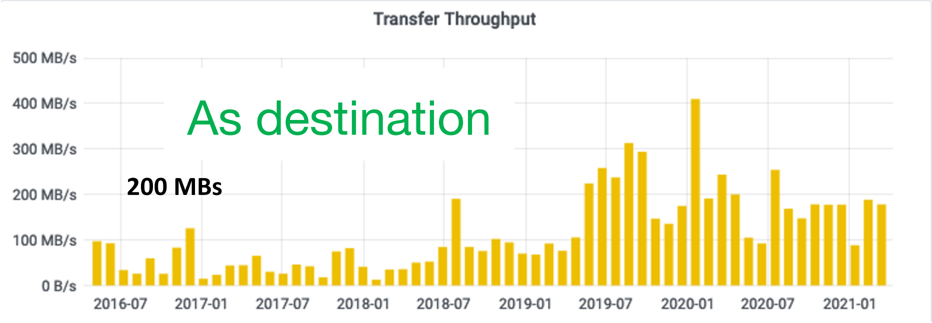

IFIC ATLAS data volume managed by Rucio

2.5 PB

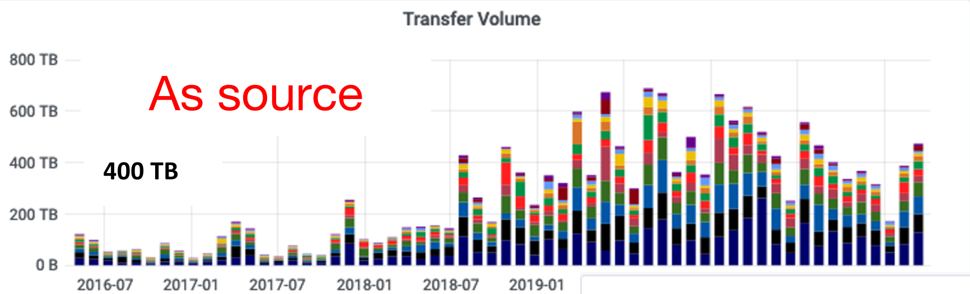
Approaching
a total of 2.5 PB
of data in Rucio **(80% of our capacity)**


As destination

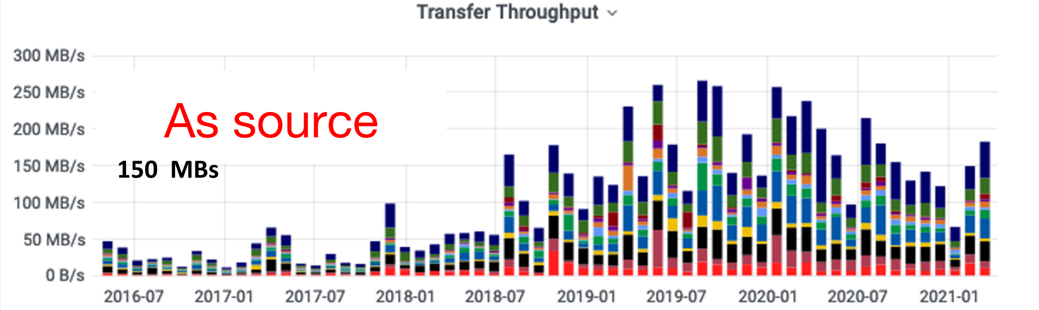
250 TB


As destination

200 MBs


As source

400 TB


As source

150 MBs

- Average transfer throughput as destination/source 200/150 MB/s, with peaks up to 400/250 MB/s, at a rate of > 0.5/0.3 Hz (50k/30k files/day)
  - Consistently transferring more than 250 /300 TB/month

IFAE resources & performances

- Resources @ site

	2017	2018	2019	2020	2021
CPU(KHS06. horas) Accounting MHS06	8.46 69	11.6 75	13.92 115	12.92 92	12.92 26
Disk(TB)	976	976	996	996	996

- Performance

	2017	2018	2019	2020	2021
Availability	98.67%	97.12%	99.50%	98.16%	98.46%
Reliability	98.99%	99.19%	99.57%	98.24%	99.16%

- Number of processed Jobs

	2017	2018	2019	2020	2021
Nb proc. jobs	1,686k	2,253k	2,568k	960k	234k

- Data transfers

	2017	2018	2019	2020	2021
D.T. as source	0.193 PB	0.465 PB	1.128 PB	1.726 PB	0.556 PB
D.T. as destination	0.964 PB	1.821 PB	3.37 PB	3.21 PB	0.67 PB

Overview

- The ATLAS IFAE Tier-2 is co-located with the ATLAS Tier-1 resources at PIC sharing the same infrastructure. See slides for PIC Tier-1 for details.
- The **same team** supports the **Tier-1, Tier-2 and Tier-3** for ATLAS at PIC.
- This is a unique case in Spain where all components are available in the **same site sharing all resources and data.**
- This Tier-2 has no **HPC resources** as they **deployed at PIC** under the ATLAS Tier-1.
- The base capacity of this T2 is 25% of the Federated Spanish Tier-2.
- The Tier-2 hosts the Tier-3 analysis facility of the IFAE-ATLAS group and provides 12% (120 TB) for for local data analysis repository (IFAE_LOCALGROUPDISK).

Capacity in 2021

- The current capacity of the IFAE Tier-2 is of the order of **12.000 HepSpecs2006 of CPU** capacity and **1 PB of data storage**.
- As collateral, IFAE T2 has access to the 3.5 PB data on disk and 10 PB on tape of the PIC ATLAS Tier-1, and 0.5 PB of the Tier-3 disk.
- The CPU capacity share is adapted monthly as a function of the pledge and the delivered computing resources.
- If the **ES-Tier2 pledge is underperforming for some circumstances, IFAE could take CPU capacity** from the excedent of PIC resources. In more than ten years, this mechanism was never needed.

Upgrades in 2021

- This month the **Spanish Data Supercomputing Network** (RES-DATA) has provided a multi-year grant (DATA-2020-1-0024) to **increase the local data analysis repository with 200 TB on disk and 200 TB on tape** as an extension of the disk.

- The IFAE will benefit from the 200 Mbps upgrade of the PIC network, multiplying the network capacity to the **LHCOPN and LHCONE** networks.

- IFAE is planning to provide access from **grid analysis jobs** to heterogeneous resources to the **GPU capabilities of the HTCondor system at PIC**, given the rising interest in these types of resources for analysis. Tier3 already has this capability through the Jupyter notebooks.

UAM resources & performances

- Resources @ site

	2017	2018	2019	2020	2021
CPU(HS06. hours)	76 M	72 M	100 M	63 M	8 M
Disk(TB)	1000	1000	1000	1000	1000

- Performance

2020-2021

- ARC Issues
- dCache Pinmanager issue
- Electricity shutdown

	2017	2018	2019	2020	2021
Availability	99.93 %	99.19 %	98.39 %	86.22 %	80.57 %
Reliability	99.67 %	99.65 %	98.64 %	90.25 %	80.57 %

- Number of processed Jobs

	2017	2018	2019	2020	2021
Nb proc. jobs	354,596	551,908	1,338,589	2,603,346	3,802656

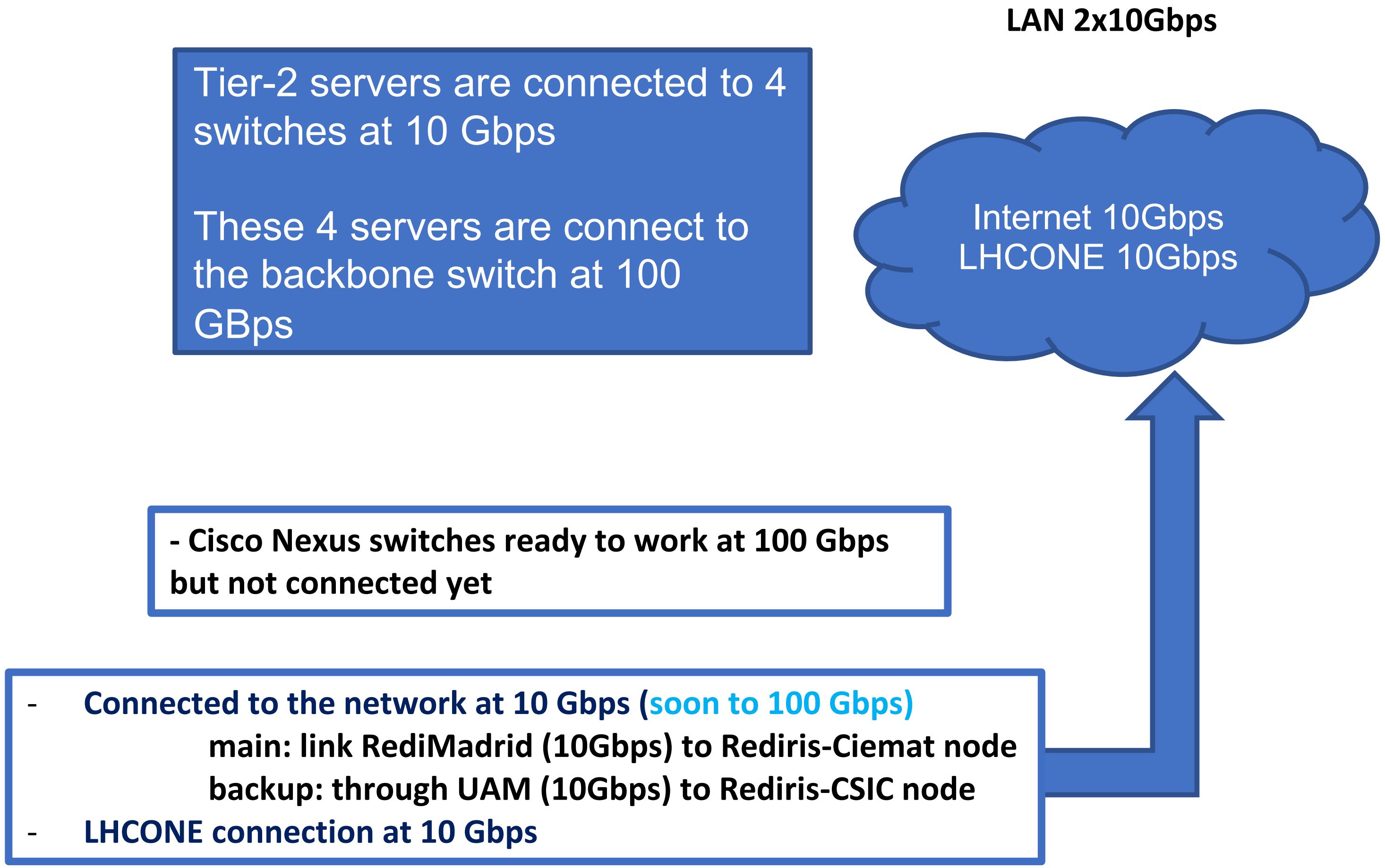
- Data transfers

	2017	2018	2019	2020	2021
D.T. as source	0.160 PB	0.281PB	1.214 PB	2.7PB	4.044 PB
D.T. as destination	0.36 PB	0.72 PB	1.33 PB	2.93 PB	4.36 PB

UAM network

- New switches
  - Cisco Nexus 93180 (header): 28 * (40 Gbps) fibre + 4 * (100 Gbps) fibre
  - Cisco Nexus 93108 (4 slaves of the header): 48 * (10 Gbps) copper + 6 * (100 Gbps) fibre.
  - Servers not connected yet to switch slaves. When done, cabling will be cleaner
  - Cisco Nexus 5548 (slave):

    32 * (10 Gbps) fibre



LAN 2x10Gbps

Tier-2 servers are connected to 4 switches at 10 Gbps

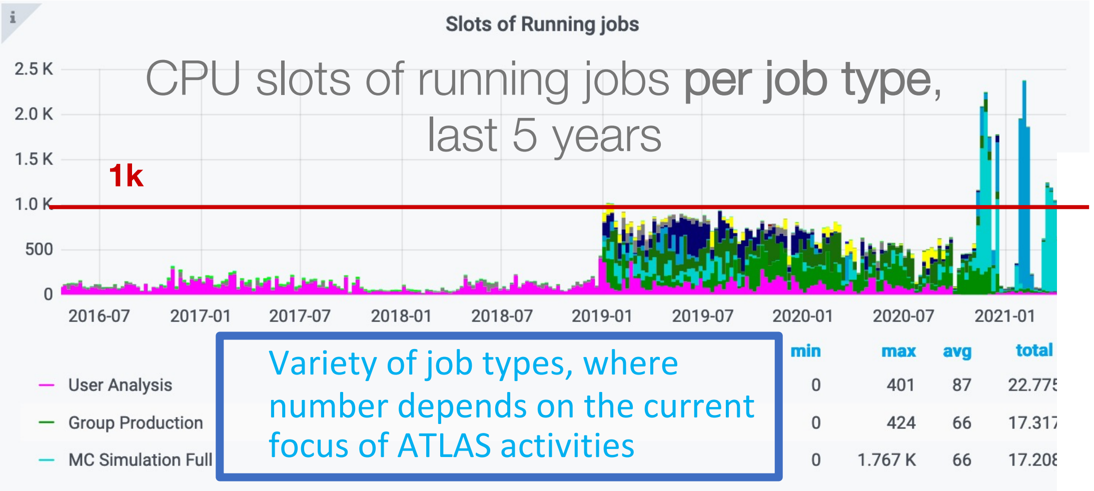
These 4 servers are connect to the backbone switch at 100 GBps

Internet 10Gbps
LHCONE 10Gbps

- Cisco Nexus switches ready to work at 100 Gbps but not connected yet

- **Connected to the network at 10 Gbps (soon to 100 Gbps)**
  main: link RediMadrid (10Gbps) to Rediris-Ciemat node
  backup: through UAM (10Gbps) to Rediris-CSIC node
- **LHCONE connection at 10 Gbps**

Number of Jobs & Data Transfer (2017 – 2021)

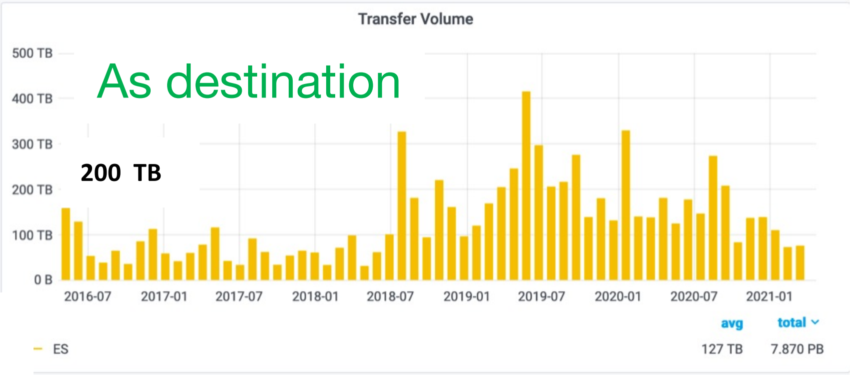


CPU slots of running jobs **per job type**, last 5 years

Variety of job types, where number depends on the current focus of ATLAS activities

	min	max	avg	total
User Analysis	0	401	87	22.775
Group Production	0	424	66	17.317
MC Simulation Full	0	1.767 K	66	17.208

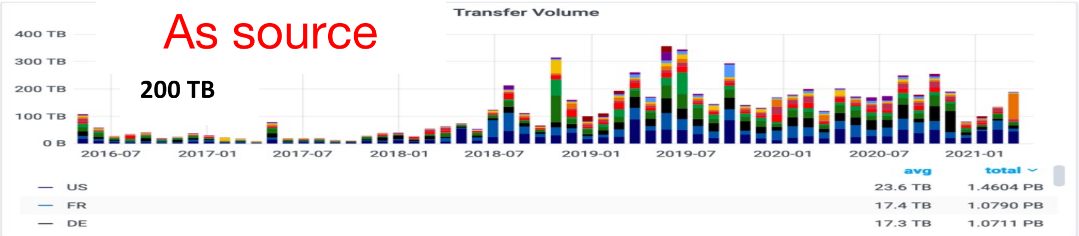
- Steady state of more than 1.000 running job slots since 2019, typically using 2GB per job slot
- Mainly running with either 8 or 1 cores (“multi-core” or “single-core”) per job, depending on type of job



As destination

200 TB

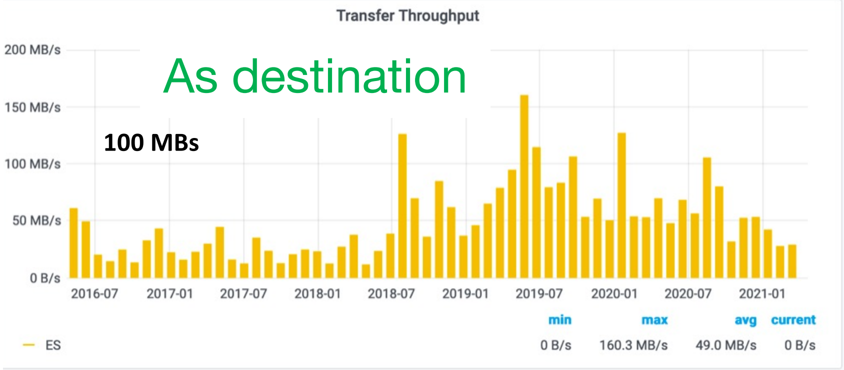
	avg	total
ES	127 TB	7.870 PB



As source

200 TB

	avg	total
US	23.6 TB	1.4604 PB
FR	17.4 TB	1.0790 PB
DE	17.3 TB	1.0711 PB

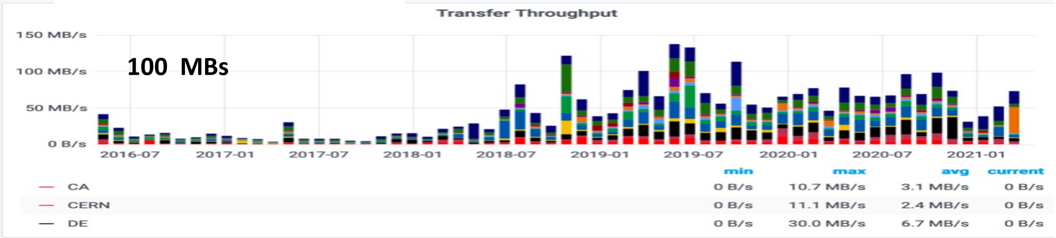


As destination

100 MBs

	min	max	avg	current
ES	0 B/s	160.3 MB/s	49.0 MB/s	0 B/s

As source



100 MBs

	min	max	avg	current
CA	0 B/s	10.7 MB/s	3.1 MB/s	0 B/s
CERN	0 B/s	11.1 MB/s	2.4 MB/s	0 B/s
DE	0 B/s	30.0 MB/s	6.7 MB/s	0 B/s

- Average transfer throughput as destination/source 100/100 MB/s, with peaks up to 400/150 MB/s
- Consistently transferring more than 200 /200 TB/month

Ciemat resources & performances

- Resources @ site

	2017	2018	2019	2020	2021
CPU(HS06 hours)	241 M	247 M	281 M	312 M	91 M
Disk(TB)	1600	2100	2340	2340	2550

- Performance

	2017	2018	2019	2020	2021
Availability	98.77 %	99.77 %	98.52 %	98.38 %	99.70 %
Reliability	99.55 %	99.80 %	98.67 %	98.85 %	99.99 %

- Number of processed Jobs

	2017	2018	2019	2020	2021
Site proc. jobs (pilots)	604,425	546,672	252,156	282,887	55,889
VO proc. jobs (tasks)	No data	2,943,000	4,283,000	4,013,000	956,000

- Data transfers

(total SE input/output)

	2017	2018	2019	2020	2021
D.T. as source (PB)	No data	15.8	19.0	14.1	3.7
D.T. as destination (PB)	No data	9.5	15.9	9.2	0.89

CIEMAT CMS Tier-2

75% of CMS Spanish Tier-2 federation

Also providing computing services and support to other local research communities (astroparticle physics, cosmology, neutrinos…)

CPU

- **~150 CPU nodes, ~2800 slots**
- HTCondor (v8.8.10) and 2 HTCondorCEs (v3.2.1)

Storage

- **~2.6 PB**, dCache v2.27
- dCache pools in dual-stack IPv4/IPv6
- TPC enabled for HTTPs (already moving to production)

Network

- **2x10 Gbps WAN** (LHCOne + Internet connections)
- **Upgrade to 100 Gbps WAN** pending deployment by RedIRIS and CIEMAT

CIEMAT CMS Tier-2

People

- J.M. Hernández (CMS contact person, CRB co-chair)
- A. Delgado Peris, J. Rodríguez Calonge (Tier-2 site managers)
- R. Fernández Pérez, J.J. Rodríguez Vázquez (technicians)

Ongoing R&D activities

- Test instance of **XCache** deployed and running since July
- Collaborating in several efforts on **Machine Learning** (aiming for future application to CMS activities)
- Requested project for an improved **analysis facility** *(see AF overview talk)*
- Collaborating in several R&D projects with PIC *(see additional contributions)*
  - Data access studies
  - Sites resource federation
  - HPC (BSC) resources integration

IFCA resources & performances

- Resources @ site

	2017	2018	2019	2020	2021
CPU(MHS06. horas)	62	63	69	85	33
Disk(TB)	1100	1100	1100	800	900

- Performance

	2017	2018	2019	2020	2021
Availability	90%	90%	96%	97%	90%
Reliability	91%	94%	97%	97%	92%

- Number of processed Jobs

	2017	2018	2019	2020	2021
Nb proc. Jobs (Mjobs)	1.4	1.9	2.4	2.8	0.8

- Data transfers

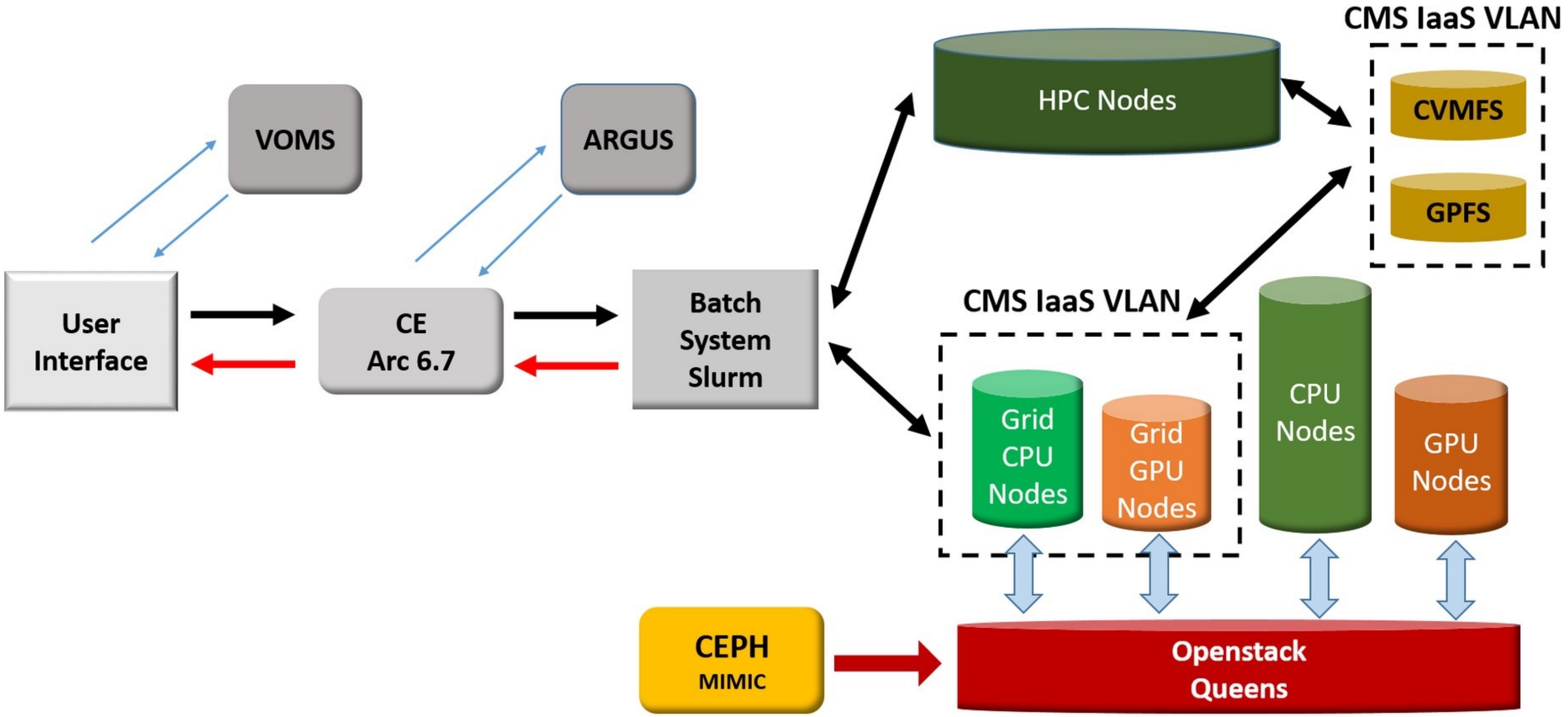
	2017	2018	2019	2020	2021
D.T. as source	0.643 PB	1.874 PB	4.855 PB	10.800 PB	16.178 PB
D.T. as destination	1.44 PB	2.87 PB	5.31 PB	11.7 PB	17.44 PB

IFCA Tier2 running on a cloud service

- The IFCA Tier2 is implemented on the Opensource Suite of Cloud OpenStack.
  - Integrated with the rest of the IFCA computing infrastructure.
- IFCA provides a IaaS (Infrastructure as a Service) to the Tier2 project of CMS.
  - Allows to easily benefit from already deployed services.
- Different resources can be used and shared through the BatchSystem:
  - Grid Worker nodes (IaaS).
  - GPU nodes can also be served by the cloud system (IaaS).
  - Opportunistic running on the HPC Altamira node.
- Worker Nodes are cloud machines building singularity containers to run CMS jobs.
  - CMS software loaded through cvmfs cache.
  - Output is stored in GPFS distributed file system.
  - Containers deleted after execution.

General Workflow

- CE takes care of the User Subject, Group or Role, and mapping to a defined queue at arc.conf file.



USC resources & performances

- Resources @ site

	2017	2018	2019	2020	2021
CPU(MHS06.horas)	68	69	59	55	9
Disk(TB)	0	0	0	0	0

- Performance

	2017	2018	2019	2020	2021
Availability	97.06%	96.93%	64.42%	98.94%	100%
Reliability	98.20%	96.93%	64.42%	98.94%	100%

- Number of processed Jobs

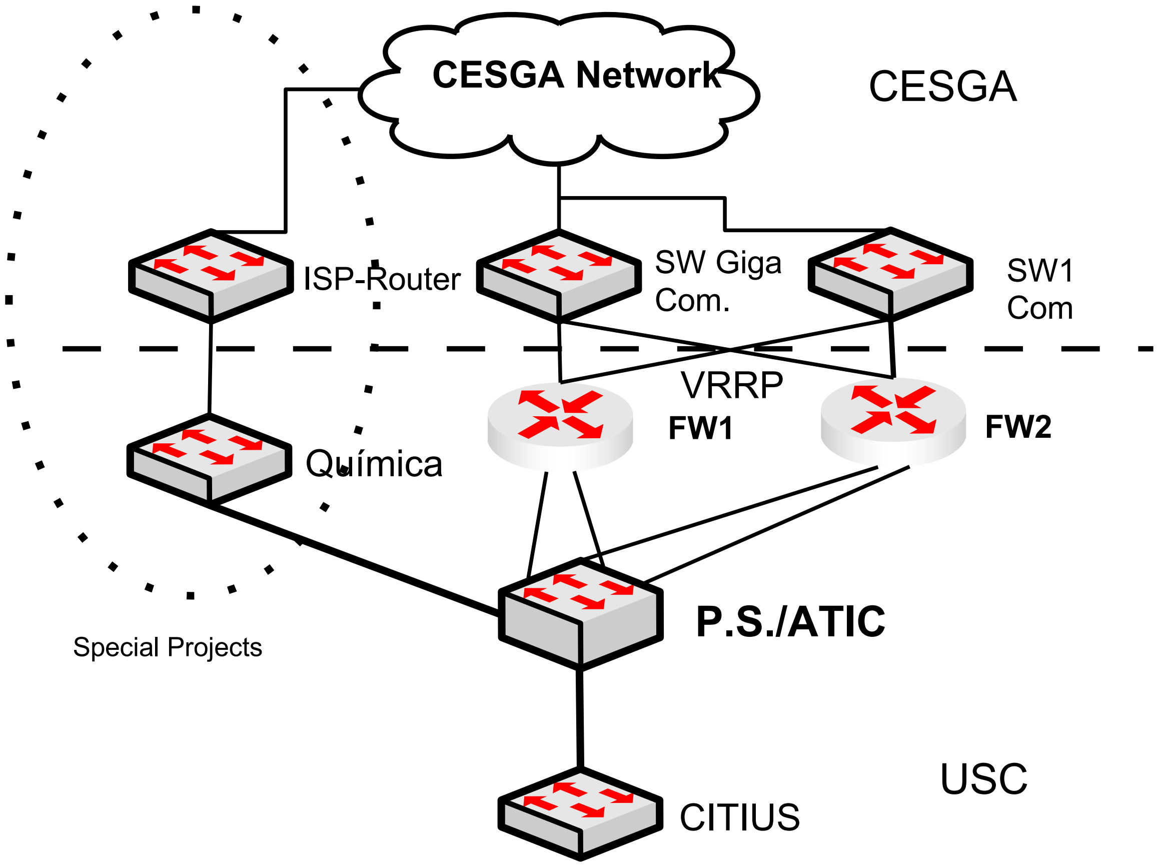
	2017	2018	2019	2020	2021
Nb proc. jobs	469,747	465,842	526,288	706,047	116,001

- Data transfers

	2017	2018	2019	2020	2021
D.T. as source	0	0	0	0	0
D.T. as destination	0	0	0	0	0

Our site only provides resources to the LHCb VO which does not use our SE.

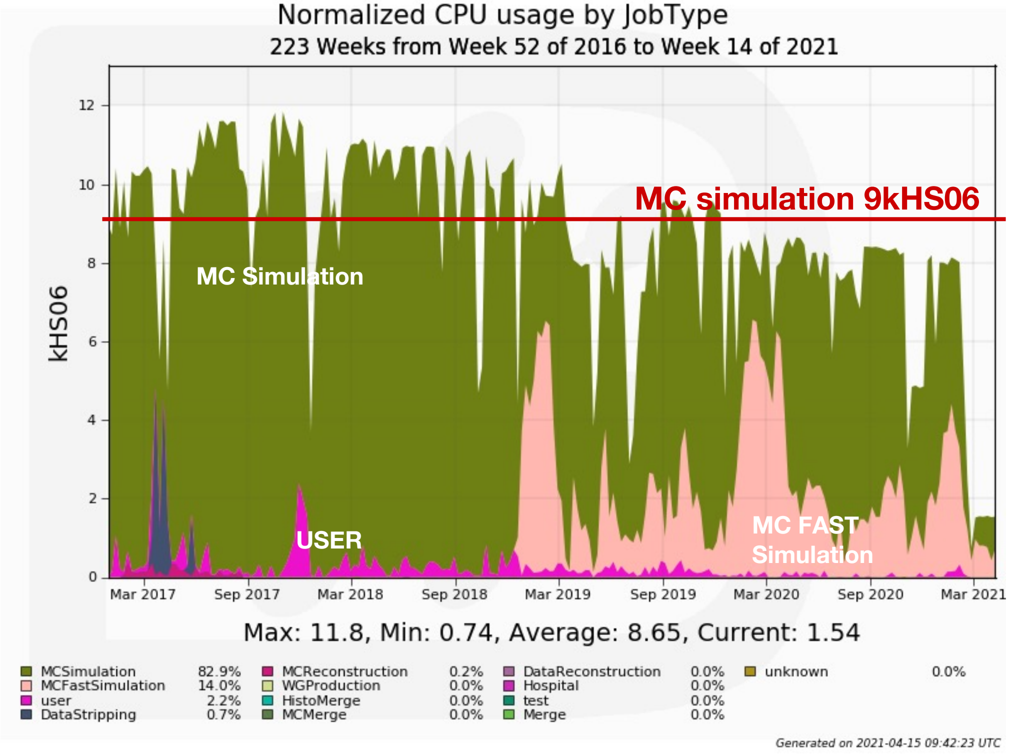
Network connection to Cesga



- Traffic filtering at the Perimetral Firewall
- ACLs on some internal routers
- Special Projects Network is not filtered and is directly connected to CESGA.
- Control nodes are all connected to the Special Projects Network
- All paths are at least 10Gb/s optical fibers

Normalized CPU and Job rates

Normalized CPU (HS06), 01/01/2017 – 01/04/2021



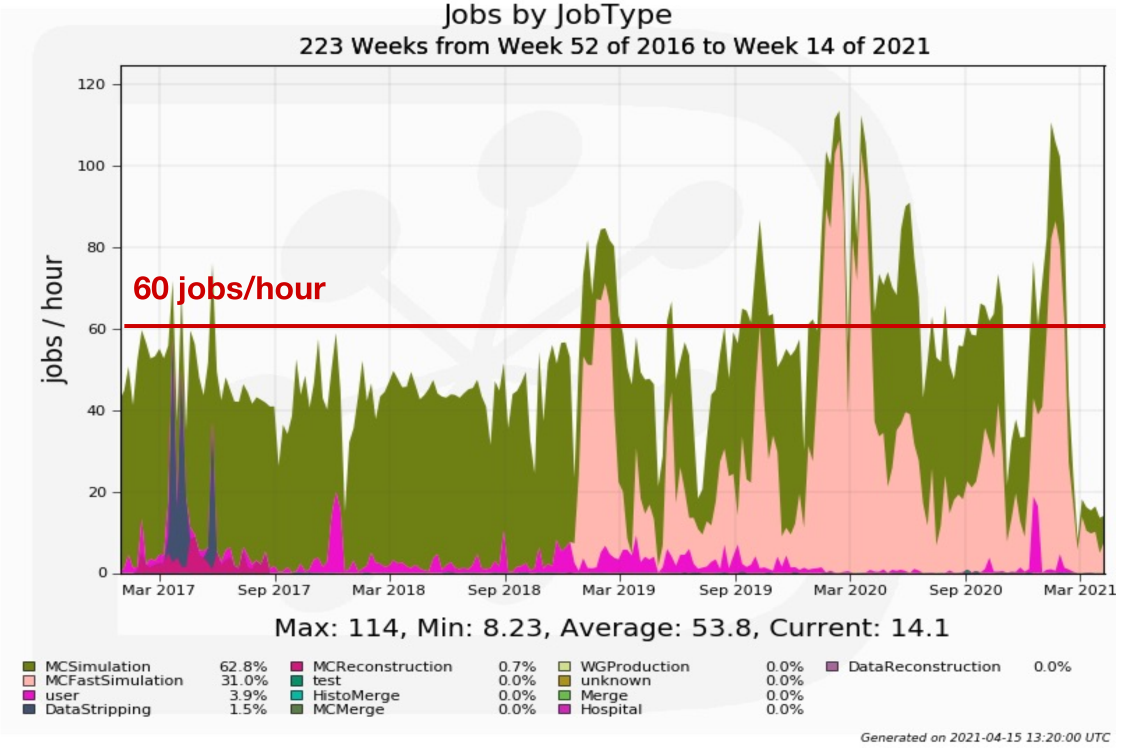
MC simulation 9kHS06

MC Simulation

USER

MC FAST Simulation

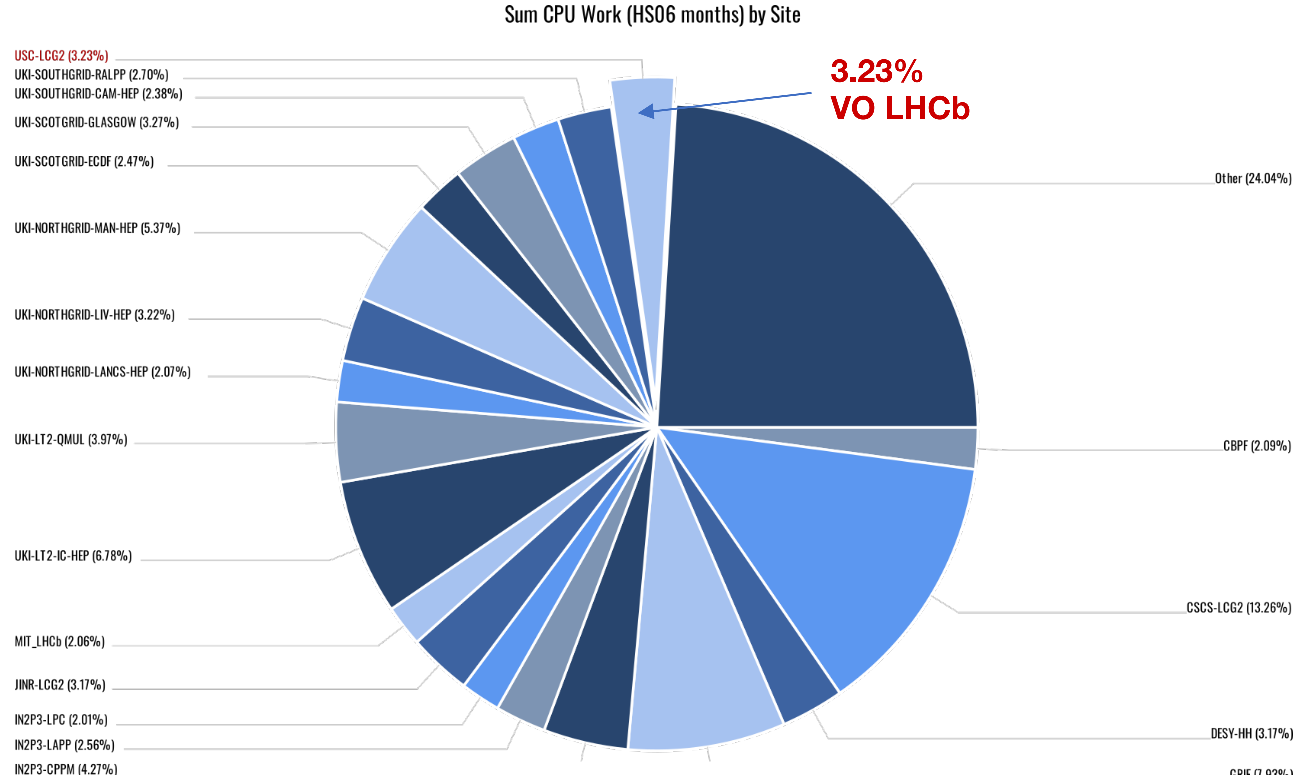
Job Rate, 01/01/2017 – 01/04/2021



60 jobs/hour

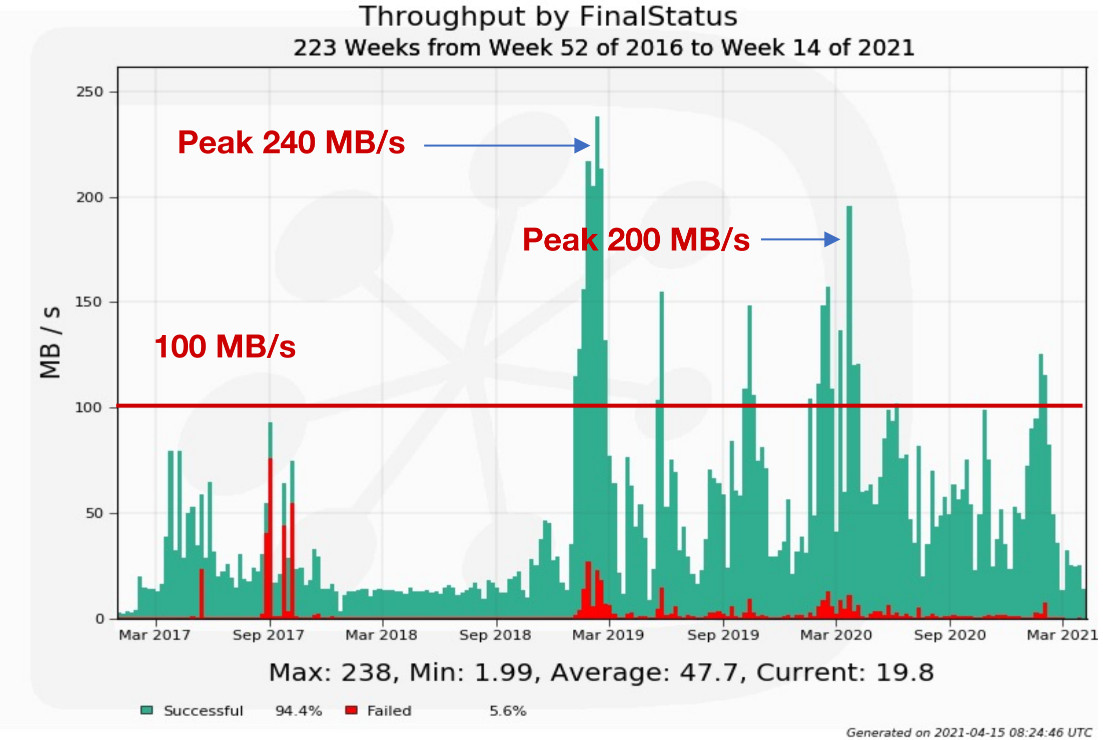
Total production and Throughput

Total Production, 01/01/2017 – 01/04/2021



3.23%
VO LHCb

Throughput



Peak 240 MB/s

Peak 200 MB/s

100 MB/s

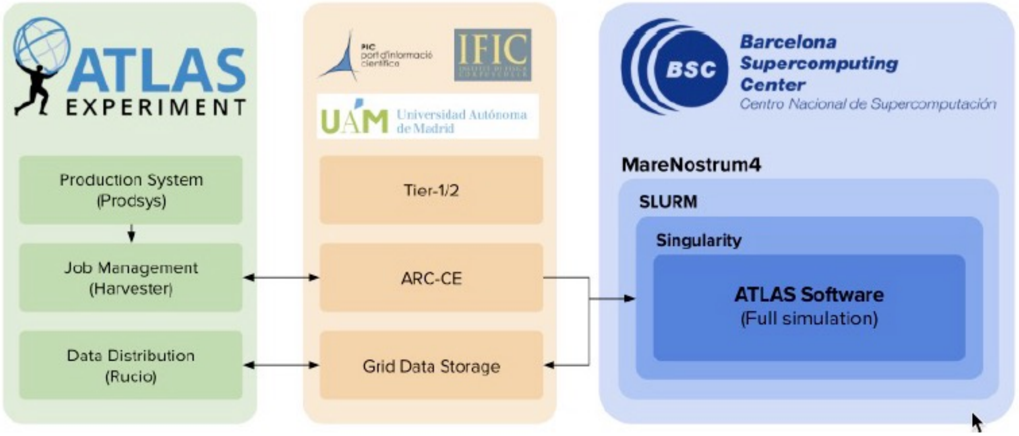
- **Improved connectivity (2018)**
  - **Migration to a NEW CPD**
  - **Top of the Rack Routers with 10Gb/s connection to Building Routers**

HPC ATLAS (IFIC, IFAE-PIC, UAM)

Use of HPC resources

- A large effort that is paying back
  - Started as an opportunistic resource **now it is a backbone of our computing contribution to simulation.**
- The access to HPC CPU time has been through the RES open calls.
  - From 2018 to mid 2020 as standard calls.
  - Starting in mid 2020 within the Ministerio-BSC agreement ("Proyecto Estratégico de Acceso al Marenostrum 4 para su utilización en la Computación del LHC").
- Three HPCs have been used Lusitania, Cibeles and MareNostrum4
- **LHCb** testing similar technical implementations in the same grant

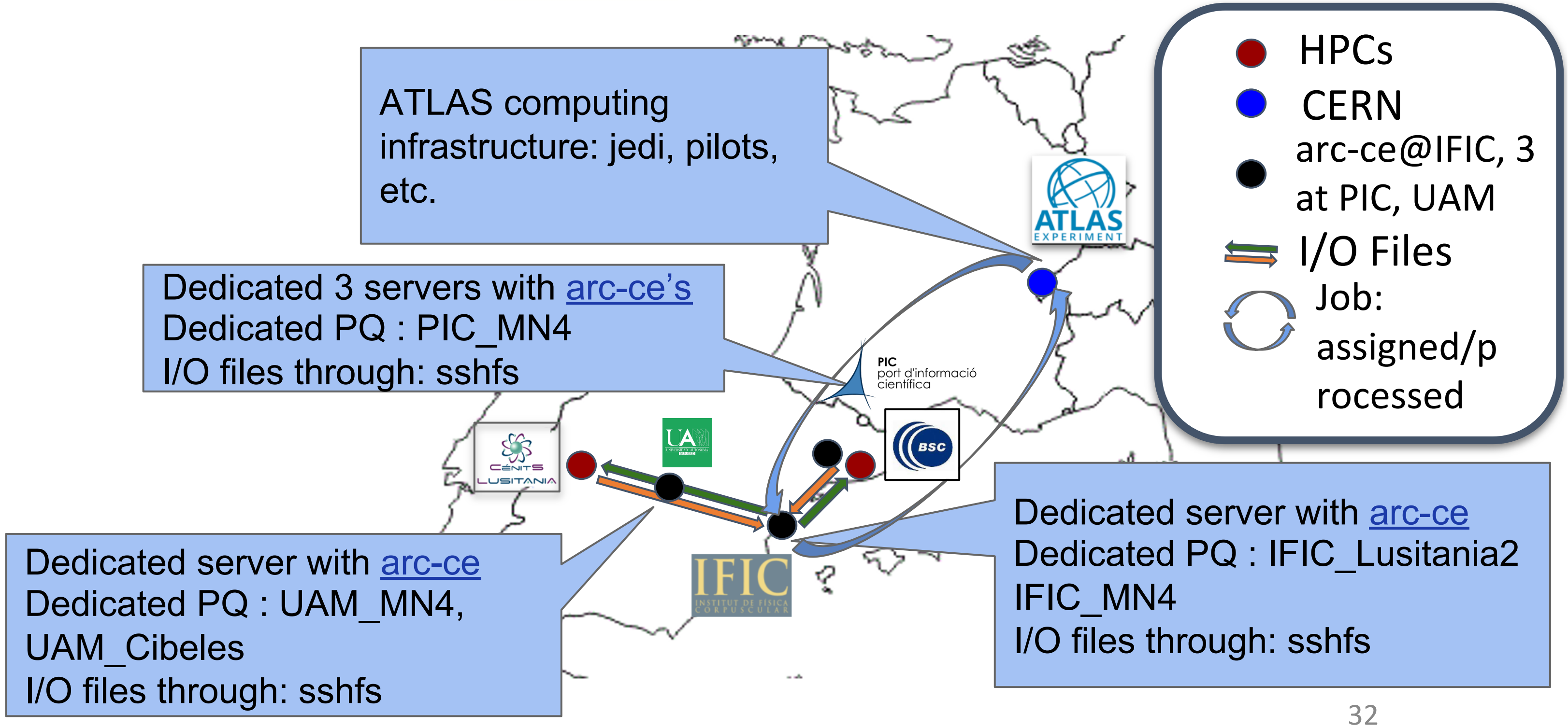
PIPELINE



ATLAS EXPERIMENT
Production System (Prodsys)
Job Management (Harvester)
Data Distribution (Rucio)
PIC port d'informació científica
IFIC
UAM Universidad Autónoma de Madrid
Tier-1/2
ARC-CE
Grid Data Storage
BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación
MareNostrum4
SLURM
Singularity
ATLAS Software (Full simulation)

- Only simulation workflow validated - singularity containers, pre-placed at MareNostrum GPFS

- MareNostrum accepts only SSH protocol for job submission and data transfer



ATLAS EXPERIMENT
PIC port d'informació científica
CénitS Lusitania
UAM
BSC
IFIC INSTITUT DE FISICA CORPUSCULAR

● HPCs
● CERN
● arc-ce@IFIC, 3 at PIC, UAM
I/O Files
Job: assigned/processed

ATLAS computing infrastructure: jedi, pilots, etc.

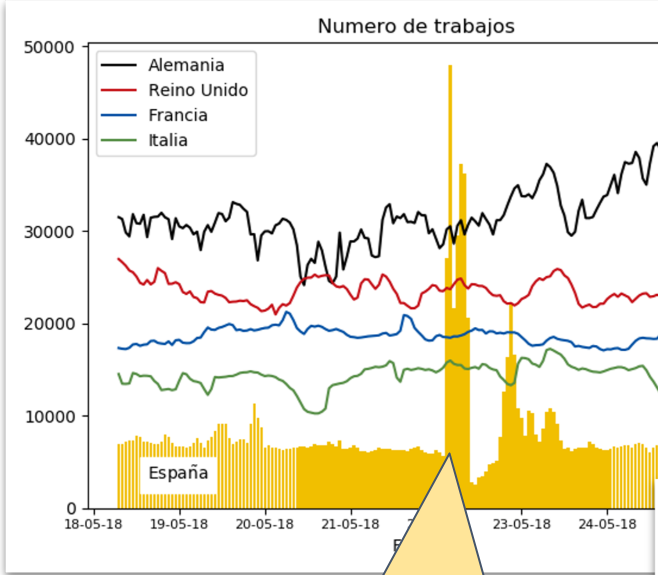
Dedicated 3 servers with arc-ce's
Dedicated PQ : PIC_MN4
I/O files through: sshfs

Dedicated server with arc-ce
Dedicated PQ : UAM_MN4, UAM_Cibeles
I/O files through: sshfs

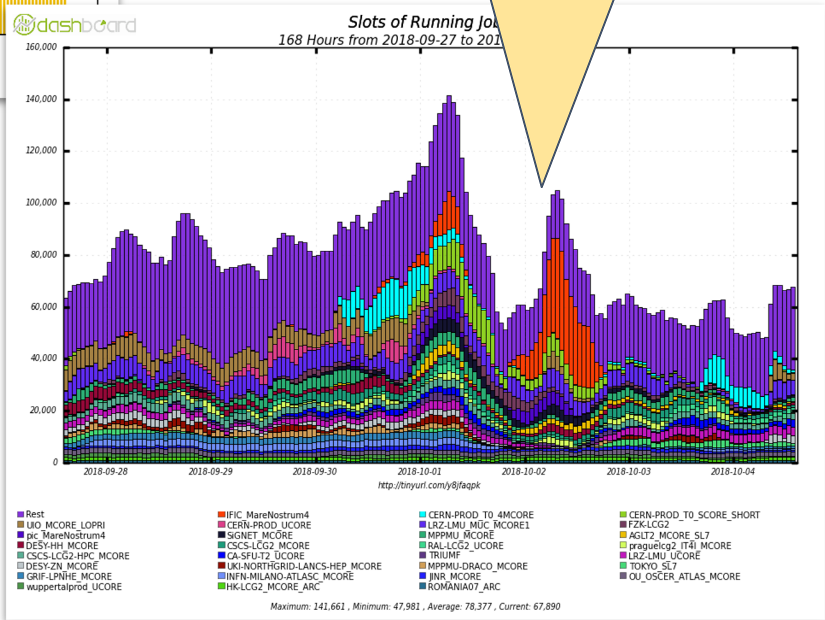
Dedicated server with arc-ce
Dedicated PQ : IFIC_Lusitania2
IFIC_MN4
I/O files through: sshfs

HPC ATLAS achievements

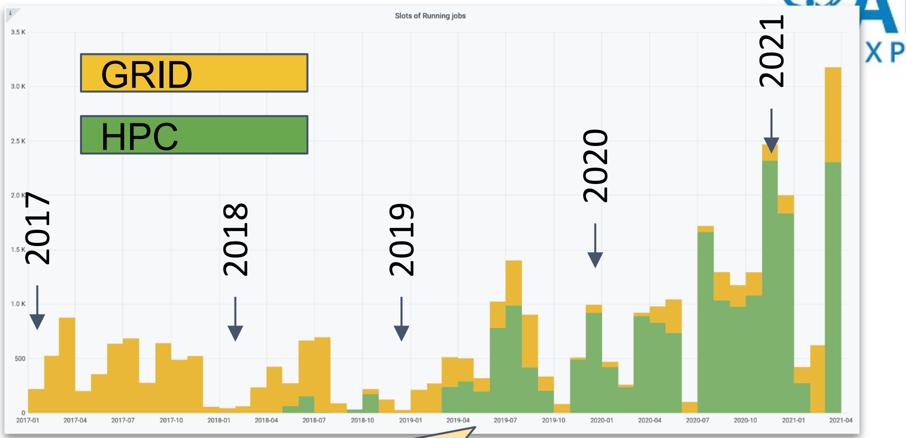
- > 500k jobs processed
- > 500M events simulated
- >30Mh CPU consumed



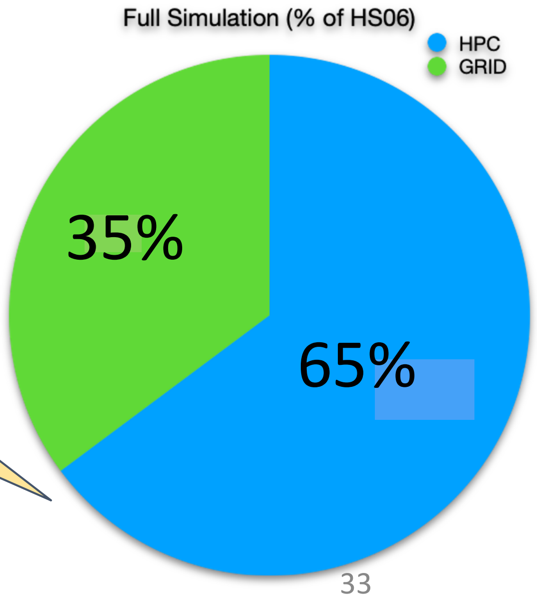
First week of HPC use by PIC and IFIC, Spain leads the ATLAS computing effort in Europe!



Number of slots running ATLAS jobs. In red Spanish Contribution (IFIC_MN4, PIC_MN4)



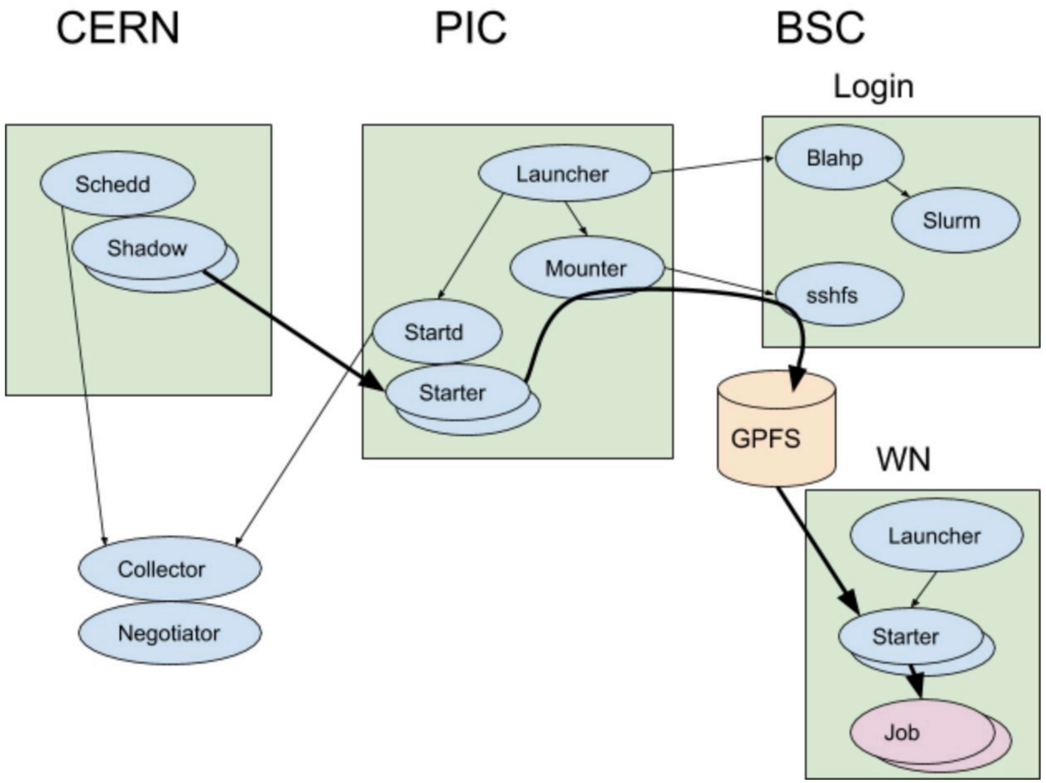
Evolution of slots of running jobs GRID-HPC from 2017 until now



Percentage of HS06 provided by GRID y MN4 since the agreement Ministerio-BSC. ONLY SIMULATION JOBS

HPC (BSC) CMS (PIC, Ciemat)

PIC and HTCondor team collaboration to **use a shared FS as control path for HTCondor**



Setup that interconnects all of the HTCondor daemons for the CMS Global Pool, PIC Tier-1 center and the BSC

Current status

- **An HTCondor-bridge has been deployed at PIC to interact with BSC** execute nodes through the login node, mounting the shared FS through **sshfs** and sending jobs to the Slurm scheduler via **ssh**

- Ran a self-contained payloads which **do not require external connectivity** connected to the CMS global pool (application packaged inside **Singularity container**, and **conditions data** read at run time dumped into a **sql file**, no I/O)

- **CMS Software modified** to accept sql files for conditions data at runtime

- **Allocations for CMS**

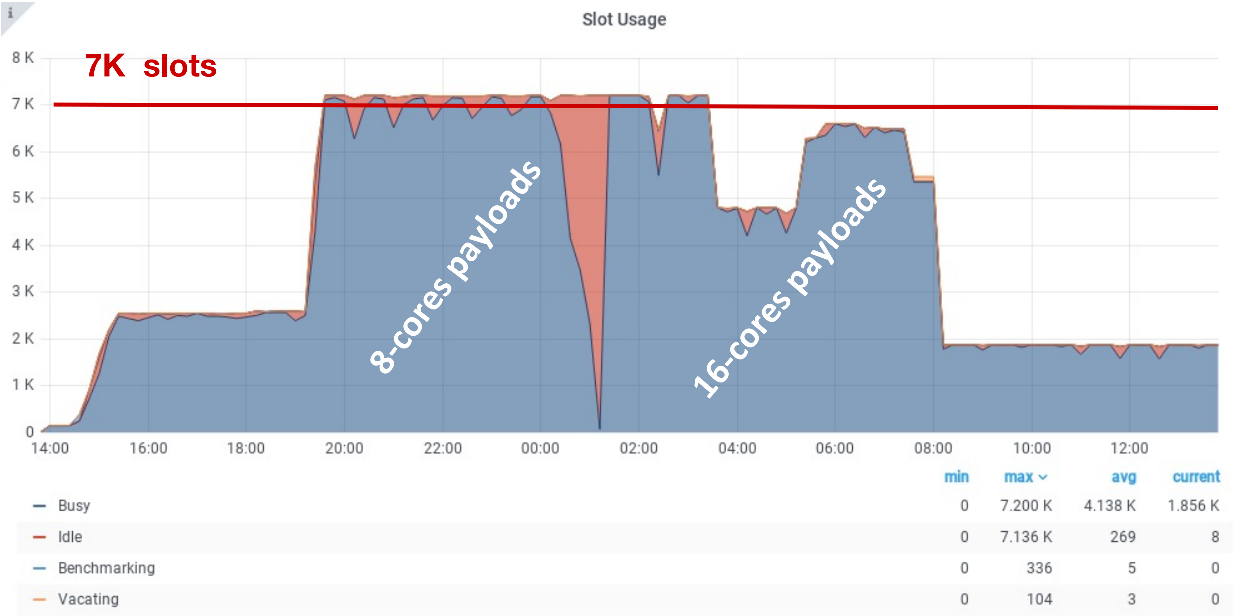
  1M CPUHrs: Nov 2020 - Feb 2021
  6M CPUHrs: Mar 2021 - Jun 2021
  Standing allocation of ~6M CPUHrs every 4 months

Use of the BSC by CMS PIC Tier-1

Integration status:

- Work done with allocation Nov 2020 - Feb 2021 ➡ **Proof of concept**
    - HTCondor flow successfully tested at scale
    - CMS Flow tested with SIM workflow
        - Custom-built singularity images
        - Custom-generated sqlite conditions data file
        - Manually pre-placed input and manual stageout
        - No WM layer involved yet

- Work being done with allocation Mar 2021 - June 2021 ➡ **Fully automation**
    - Connecting CMS WMS (pilot and payload handling)
    - Optimization of bridge service at PIC (scalability, coupling of resource request to workload demands)
    - Replication of CVMFS CMS tree to BSC (avoid building custom images)
    - Central generation and distribution of conditions data files (via CVMFS)
    - Handling of input and output data files (copy from/to PIC SE as pre/post job steps)
    - Consume 6Mhours allocation with CMS production simulation workflows



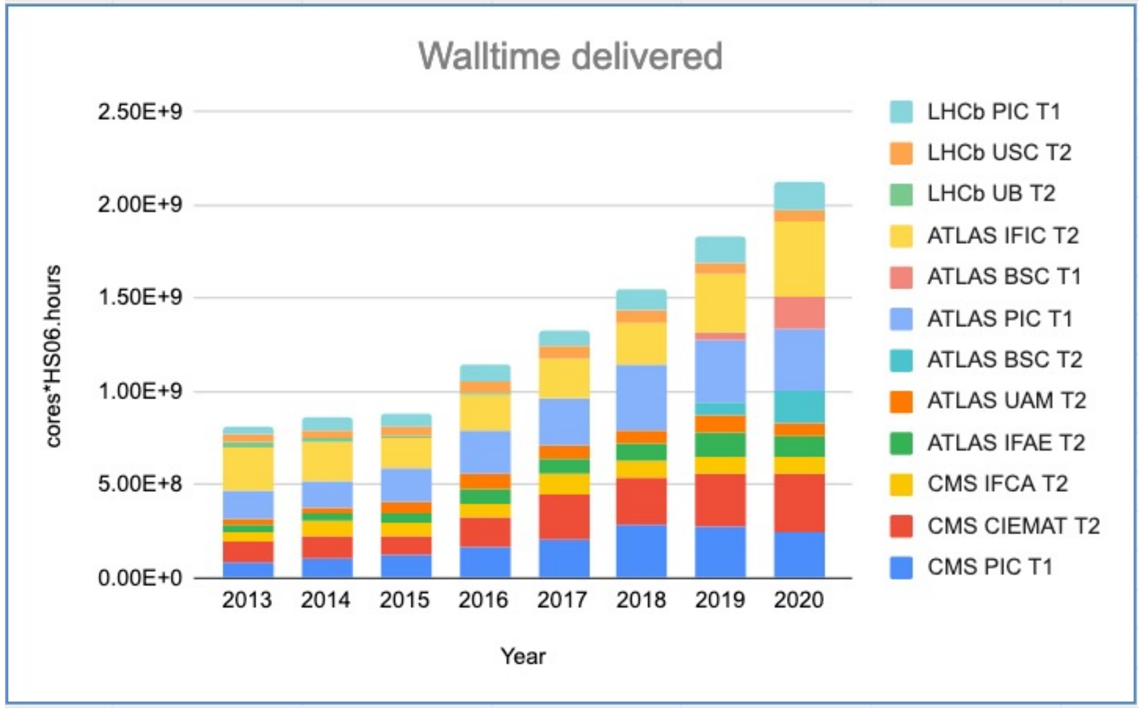
	min	max	avg	current
Busy	0	7.200 K	4.138 K	1.856 K
Idle	0	7.136 K	269	8
Benchmarking	0	336	5	0
Vacating	0	104	3	0

Scale tests: running singularity images for CMS simulation on ~7k slots (running in 48 cores machines, tuning payload core usages to maximize global CPU efficiency), plugged into the CMS Global Pool (test instance) through PIC HTCondor infrastructure, using the shared FS at BSC

WLCG-ES CPU usage 2013-2020

- **Wall time delivered in cores*HS06 hours**

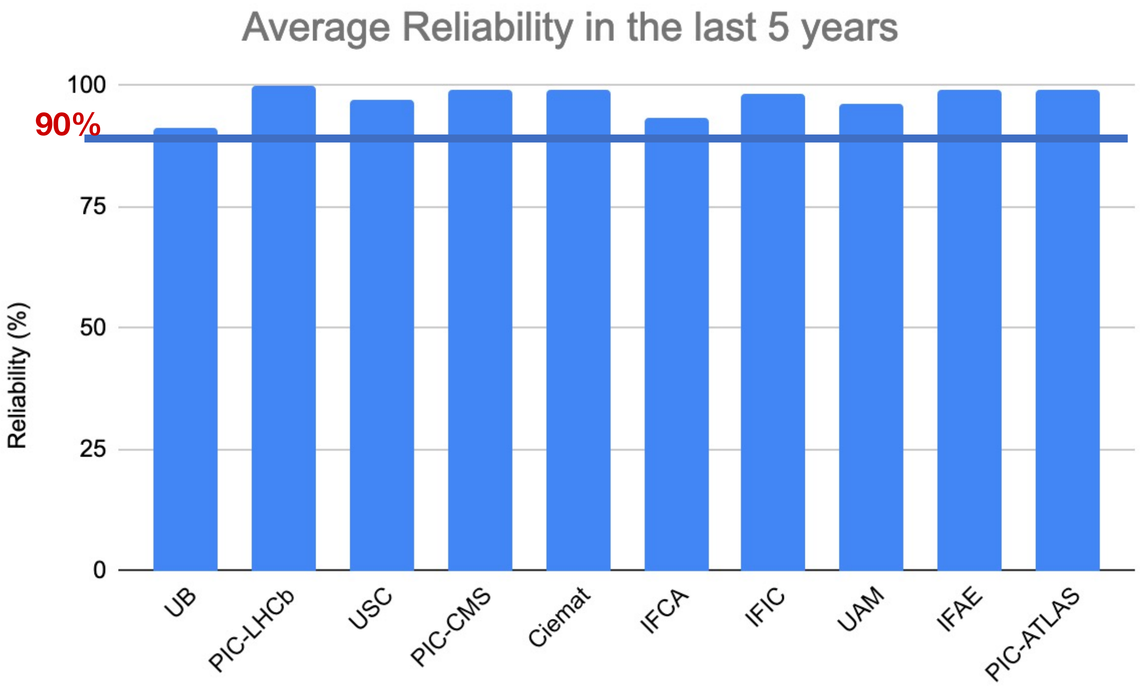
	2013	2014	2015	2016	2017	2018	2019	2020
CMS T2	159227040	200933672.4	168929194	228883033.8	352135962.3	342636012.9	370651318.6	403758260.6
ATLAS T2 + BSC	302920246.5	277432120	270987435.2	361355442.4	368447483.6	384301504.1	611370465.8	758373112.5
LHCb T2	72194181.71	66269701.17	66232507.11	69577920.94	69981581.65	69223184.64	60168744.58	56731011.73
Total T1 + BSC	280403504	317122022.5	381782310.9	488697378.6	532924029.1	757087017.7	790043784.6	907556386.7
ATLAS BSC T1						3322840	36064391	174035674
ATLAS BSC T2						2616631	70446462	173424113
ATLAS PIC T1	149307996	143553113.4	181037224.6	233818253.7	245308640.5	349196091.1	334150820.6	330290682.1
ATLAS IFIC T2	230958798.5	202283160.6	160632757.4	194551842.2	212430924.8	224914233.5	317957278.4	403643089



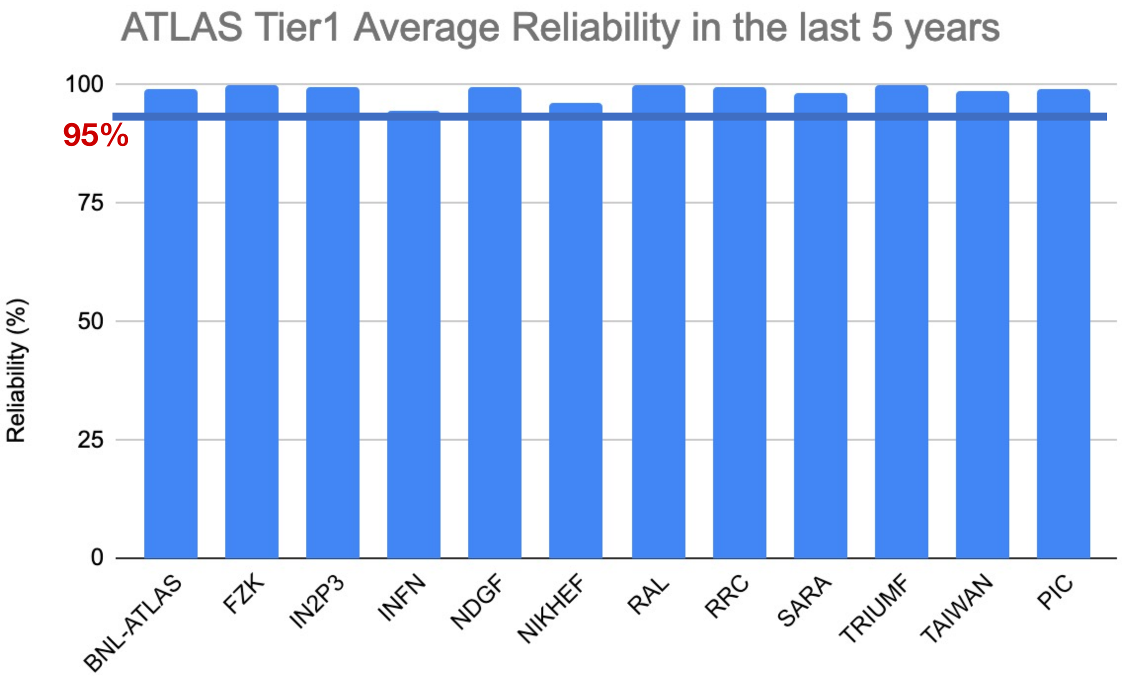
WLCG-ES Average Reliability 2017-2021

From 2017 to 2021 average reliability **greater than 91%**!!!!

PIC Tier1 greater than 99%!!!!



WLCG-ES sites

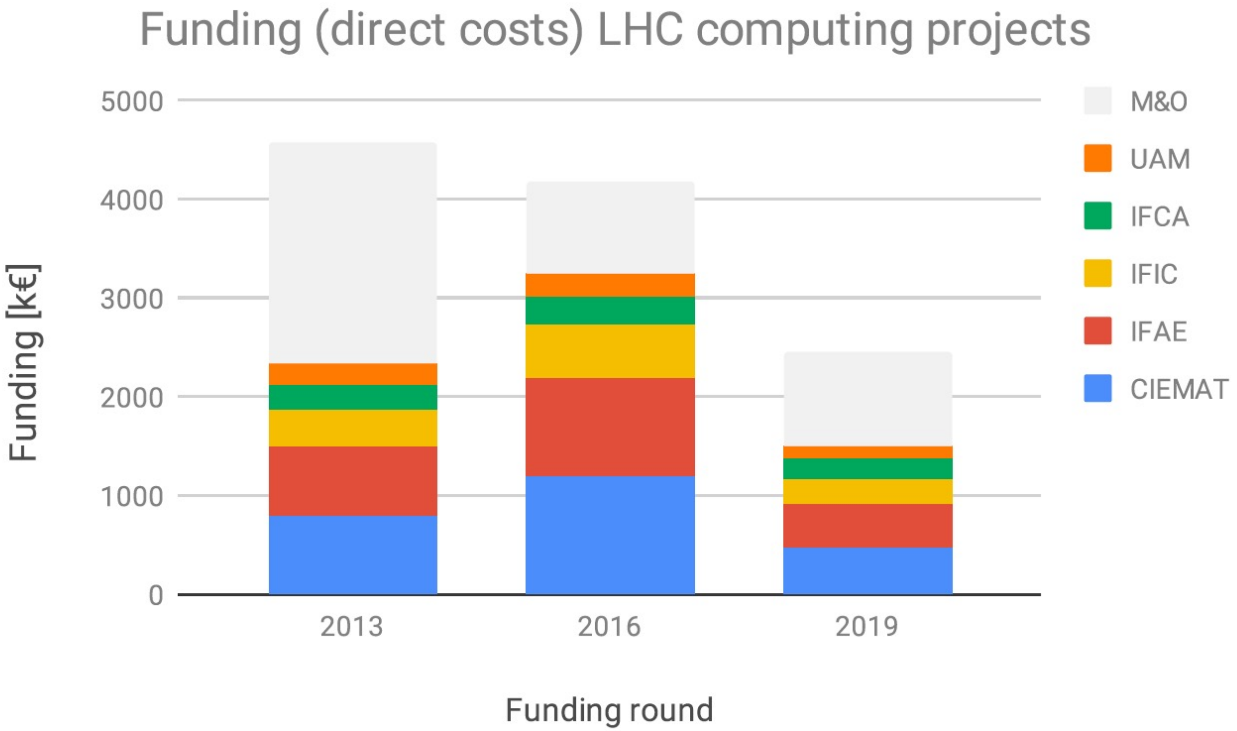


ATLAS Tier1 sites

Budget in Spain for Tiers activities in the last 6 years

- From 2013 to 2019 there is a budget **reduction around 54%**!!!!

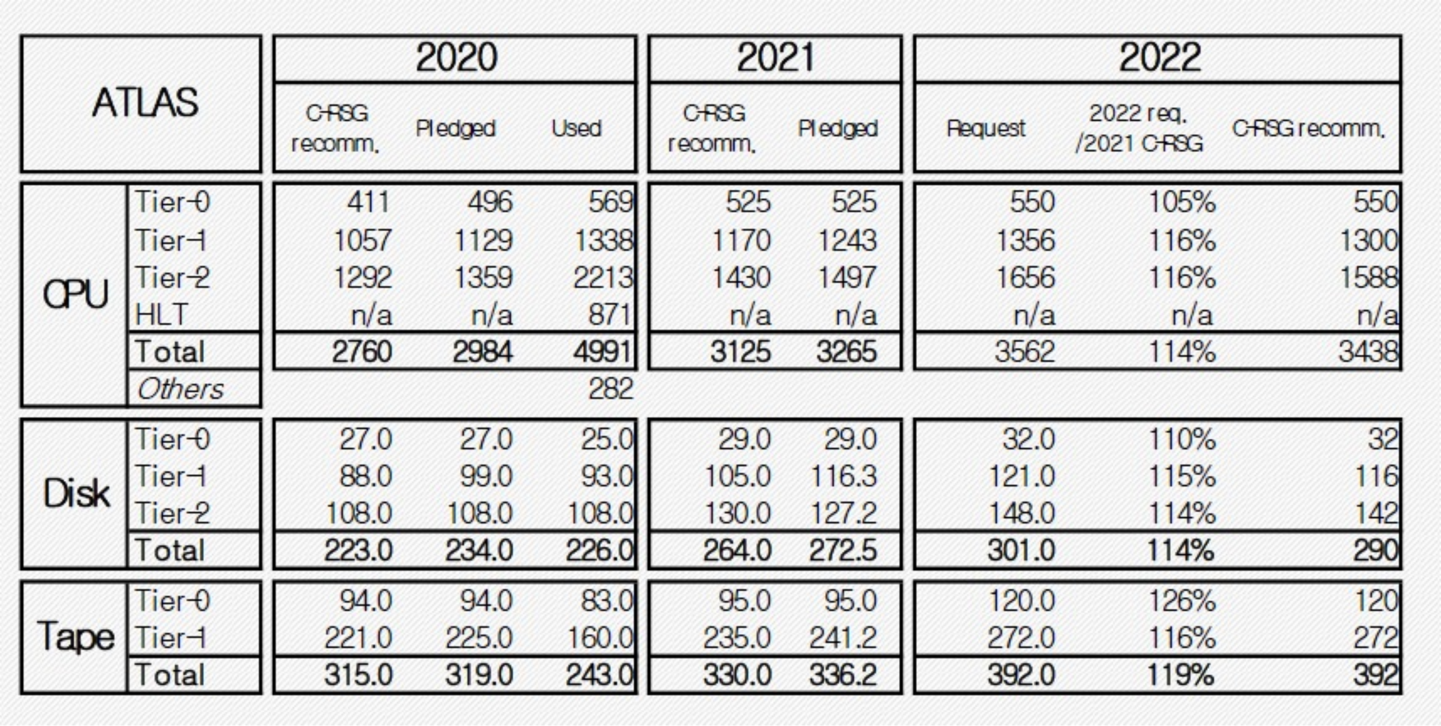
Site	FPA2013 (euros)	FPA2016 (euros)	PID2019 (euros)
PIC	1106000	1420000	702500
IFIC	1035000	765000	499000
IFAE	830000	1230000	702500
UAM	210000	245000	119000
Ciemat	700000	PIC-Ciemat	PIC-Ciemat
IFCA	700000	510000	435000
Total:	**4581000**	**4170000**	**2458000**



Next Years (Run3)

Next Years (Run3) ATLAS

- Last C-RRB in April 2021

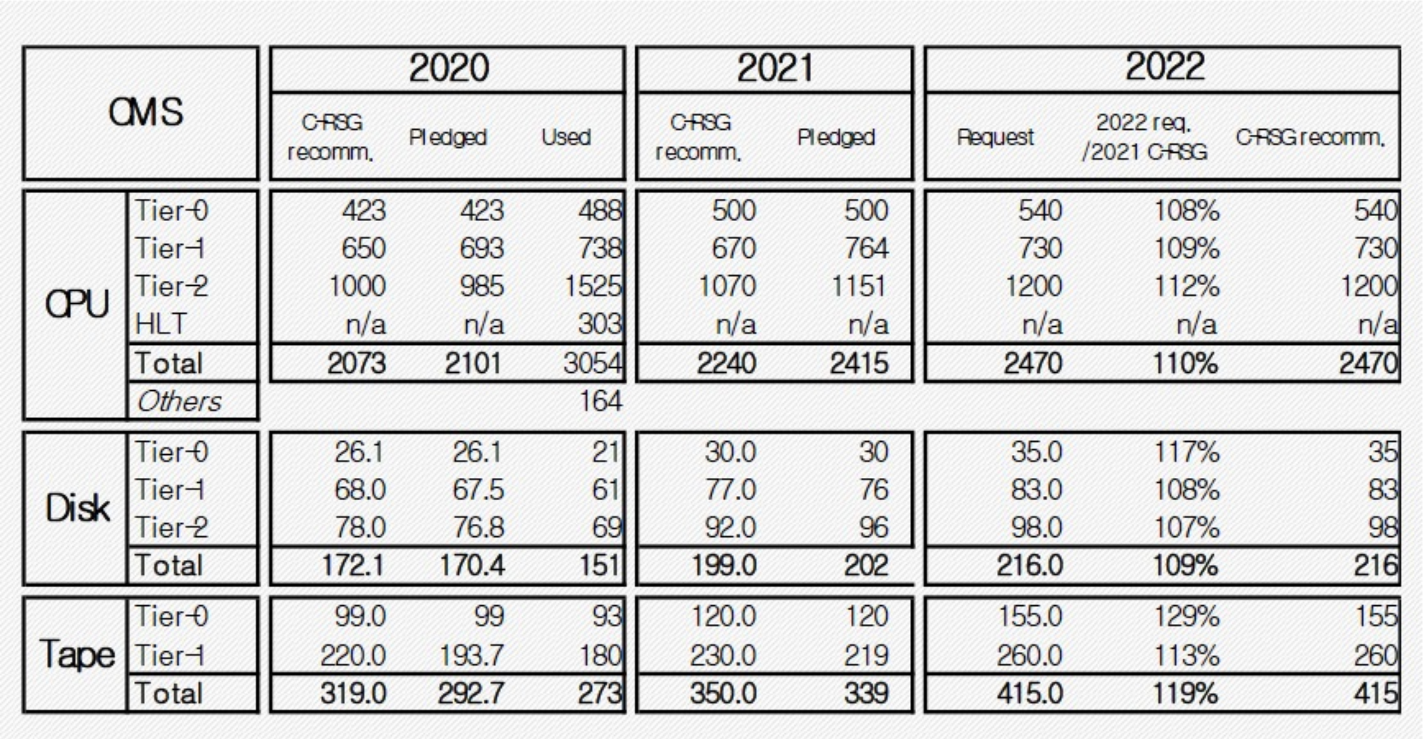


ATLAS		2020			2021		2022		
		C-RSG recomm.	Pledged	Used	C-RSG recomm.	Pledged	Request	2022 req. /2021 C-RSG	C-RSG recomm.
CPU	Tier-0	411	496	569	525	525	550	105%	550
	Tier-1	1057	1129	1338	1170	1243	1356	116%	1300
	Tier-2	1292	1359	2213	1430	1497	1656	116%	1588
	HLT	n/a	n/a	871	n/a	n/a	n/a	n/a	n/a
	Total	**2760**	**2984**	**4991**	**3125**	**3265**	3562	114%	3438
	Others			282					
Disk	Tier-0	27.0	27.0	25.0	29.0	29.0	32.0	110%	32
	Tier-1	88.0	99.0	93.0	105.0	116.3	121.0	115%	116
	Tier-2	108.0	108.0	108.0	130.0	127.2	148.0	114%	142
	Total	**223.0**	**234.0**	**226.0**	**264.0**	**272.5**	**301.0**	**114%**	**290**
Tape	Tier-0	94.0	94.0	83.0	95.0	95.0	120.0	126%	120
	Tier-1	221.0	225.0	160.0	235.0	241.2	272.0	116%	272
	Total	**315.0**	**319.0**	**243.0**	**330.0**	**336.2**	**392.0**	**119%**	**392**

- C-RRB provide the request and C-RSG recommendation for the next 2 years (2021-2022).
- With a flat budget, we (Spanish Federated Tier2) want to represent the 3% of the total ATLAS Tier2 resources and Spanish Tier1 the 5-4% of the total ATLAS Tier1.
- Expect to increase **ATLAS CPU around 14%, ATLAS Disk around 14% and TAPE around 19% per year for Run3** from 2022 to 2024.
- 50% ATLAS Spanish simulation will be done in Spanish HPC (like Mare Nostrum at BSC).
- With a flat budget and the Spanish HPC contribution, we will be able to achieve the computing challenges (CPU & Disk) for the Run3 period (2022-2024).

Next Years (Run3) CMS

- Last C-RRB in April 2021



CMS		2020			2021		2022		
		C-RSG recomm.	Pledged	Used	C-RSG recomm.	Pledged	Request	2022 req. /2021 C-RSG	C-RSG recomm.
CPU	Tier-0	423	423	488	500	500	540	108%	540
	Tier-1	650	693	738	670	764	730	109%	730
	Tier-2	1000	985	1525	1070	1151	1200	112%	1200
	HLT	n/a	n/a	303	n/a	n/a	n/a	n/a	n/a
	Total	**2073**	**2101**	3054	**2240**	**2415**	**2470**	**110%**	**2470**
	Others			164					
Disk	Tier-0	26.1	26.1	21	30.0	30	35.0	117%	35
	Tier-1	68.0	67.5	61	77.0	76	83.0	108%	83
	Tier-2	78.0	76.8	69	92.0	96	98.0	107%	98
	Total	**172.1**	**170.4**	**151**	**199.0**	**202**	**216.0**	**109%**	**216**
Tape	Tier-0	99.0	99	93	120.0	120	155.0	129%	155
	Tier-1	220.0	193.7	180	230.0	219	260.0	113%	260
	Total	**319.0**	**292.7**	**273**	**350.0**	**339**	**415.0**	**119%**	**415**

- C-RRB provide the request and C-RSG recommendation for the next 2 years (2021-2022).

- With a flat budget, we (Spanish Federated Tier2) want to represent the 3% of the total CMS Tier2 resources and Spanish Tier1 the 5-4% of the total CMS Tier1.

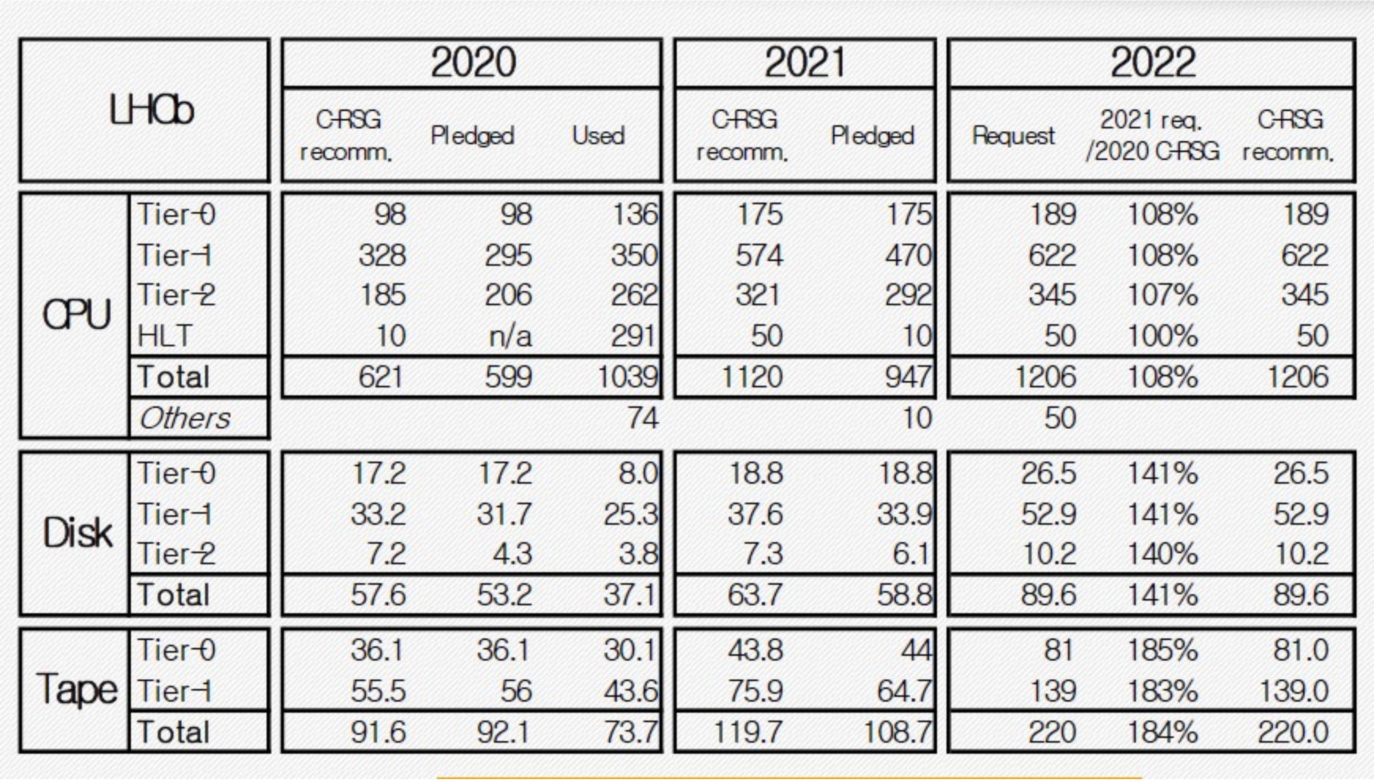
- Expect to increase **CMS CPU around 10%, CMS Disk around 9% and CMS Tape around 19% per year for Run3** from 2022 to 2024.

- 50% CMS Spanish simulation will be done in Spanish HPC (like Mare Nostrum at BSC)

- With a flat budget and the Spanish HPC contribution, we will be able to achieve the computing challenges (CPU & Disk) for the Run3 period (2022-2024).

Next Years (Run3) LHCb

- Last C-RRB in April 2021



LHCb		2020			2021		2022		
		C-RSG recomm.	Pledged	Used	C-RSG recomm.	Pledged	Request	2021 req. /2020 C-RSG	C-RSG recomm.
CPU	Tier-0	98	98	136	175	175	189	108%	189
	Tier-1	328	295	350	574	470	622	108%	622
	Tier-2	185	206	262	321	292	345	107%	345
	HLT	10	n/a	291	50	10	50	100%	50
	Total	621	599	1039	1120	947	1206	108%	1206
	Others			74		10	50		
Disk	Tier-0	17.2	17.2	8.0	18.8	18.8	26.5	141%	26.5
	Tier-1	33.2	31.7	25.3	37.6	33.9	52.9	141%	52.9
	Tier-2	7.2	4.3	3.8	7.3	6.1	10.2	140%	10.2
	Total	57.6	53.2	37.1	63.7	58.8	89.6	141%	89.6
Tape	Tier-0	36.1	36.1	30.1	43.8	44	81	185%	81.0
	Tier-1	55.5	56	43.6	75.9	64.7	139	183%	139.0
	Total	91.6	92.1	73.7	119.7	108.7	220	184%	220.0

- C-RRB provide the request and C-RSG recommendation for the next 2 years (2021-2022).

- With a flat budget, we (Spanish Federated Tier2) want to represent the 3% of the total LHCb Tier2 resources and Spanish Tier1 the 5-4% of the total LHCb Tier1.

- Expect to increase **LHCb CPU around 8% more, LHCb Disk around 41% and LHCb Tape around 84% per year for Run3** from 2022 to 2024.

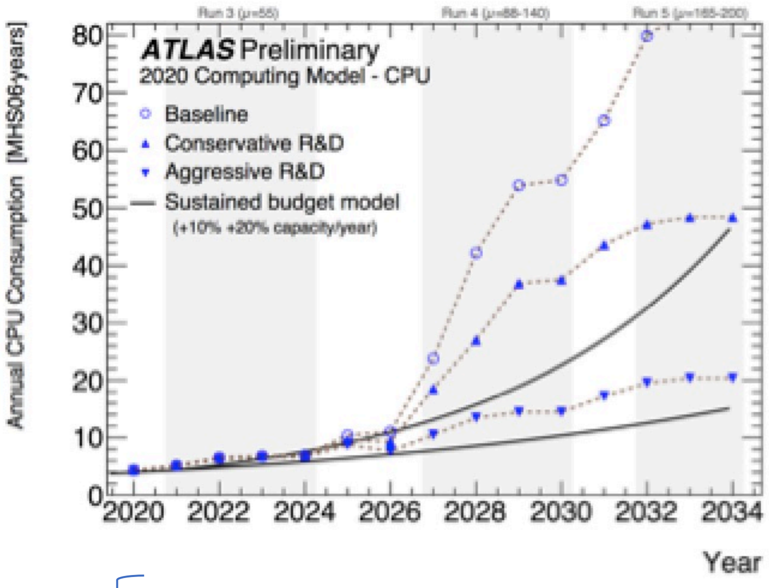
- 50% LHCb Spanish simulation will be done in Spanish HPC (like Mare Nostrum at BSC)

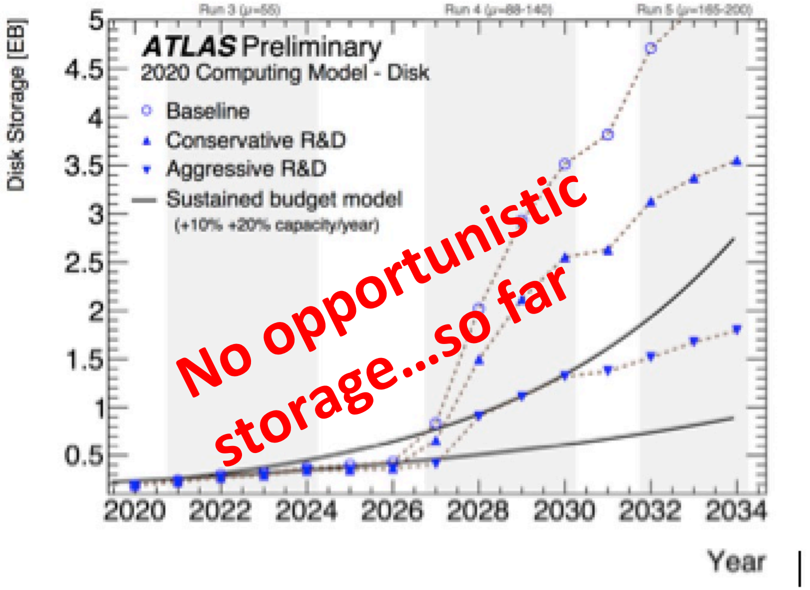
- **With a “flat budget” and the Spanish HPC contribution, we don’t know if we will be able to achieve the computing challenges (CPU & Disk) for the Run3** period (2022-2024).

Future Perspectives (HL-LHC)

Future Perspective (HL-LHC) ATLAS

Profile of resource increasing for Run 3 (2022-2024) and HL-LHC (2026-2030)





No opportunistic storage…so far

Approaches to solve CPU shortfall

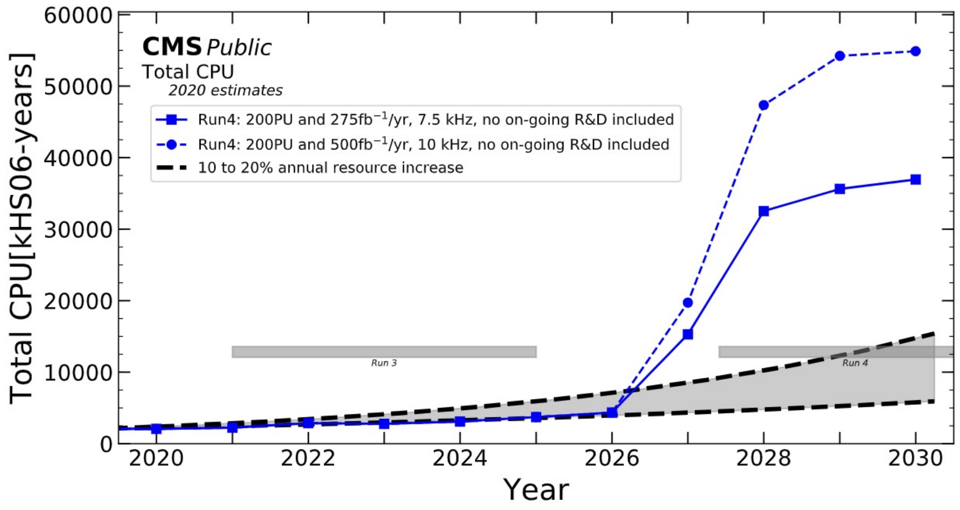
- There are a few options to face this challenge: **HPC's, cloud computing and High Level Trigger Farm.**
- Further options: use **fast simulation** instead of full one. And **speed up the MC generators** by a factor two.
- **Running on GPU's** is also feasible, but needs significantly time and effort to adapt our software to new architecture

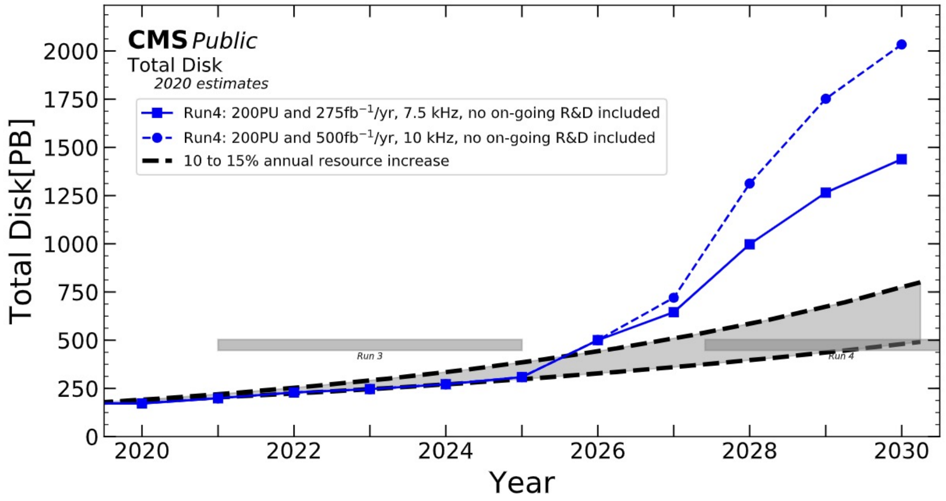
Approaches to solve Storage shortfall

- **Increase investment** in computing
- New file formats (to **reduce filesize**, many data formats for physics analysis)
- "**Less data**"
- **Use of tapes**. But this option slows down the workflow
- **Data Lakes / DOMA**

Future Perspective (HL-LHC) CMS

CMS computing resource needs projections

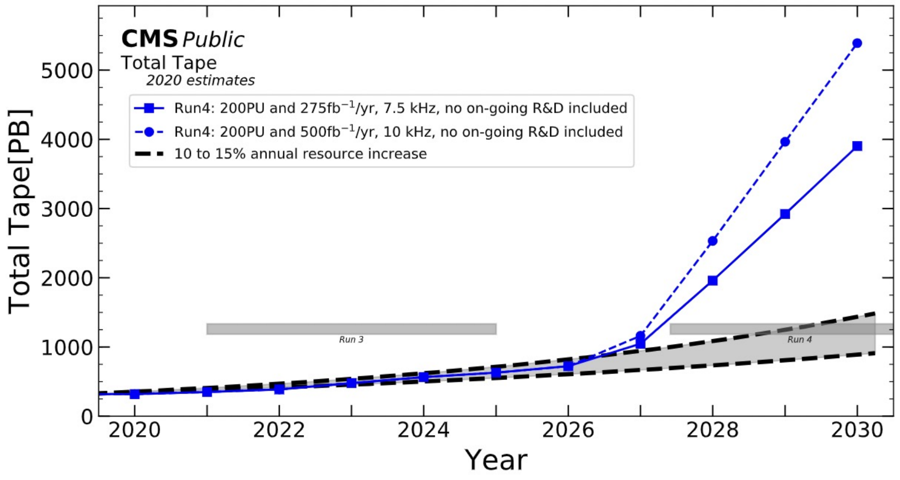




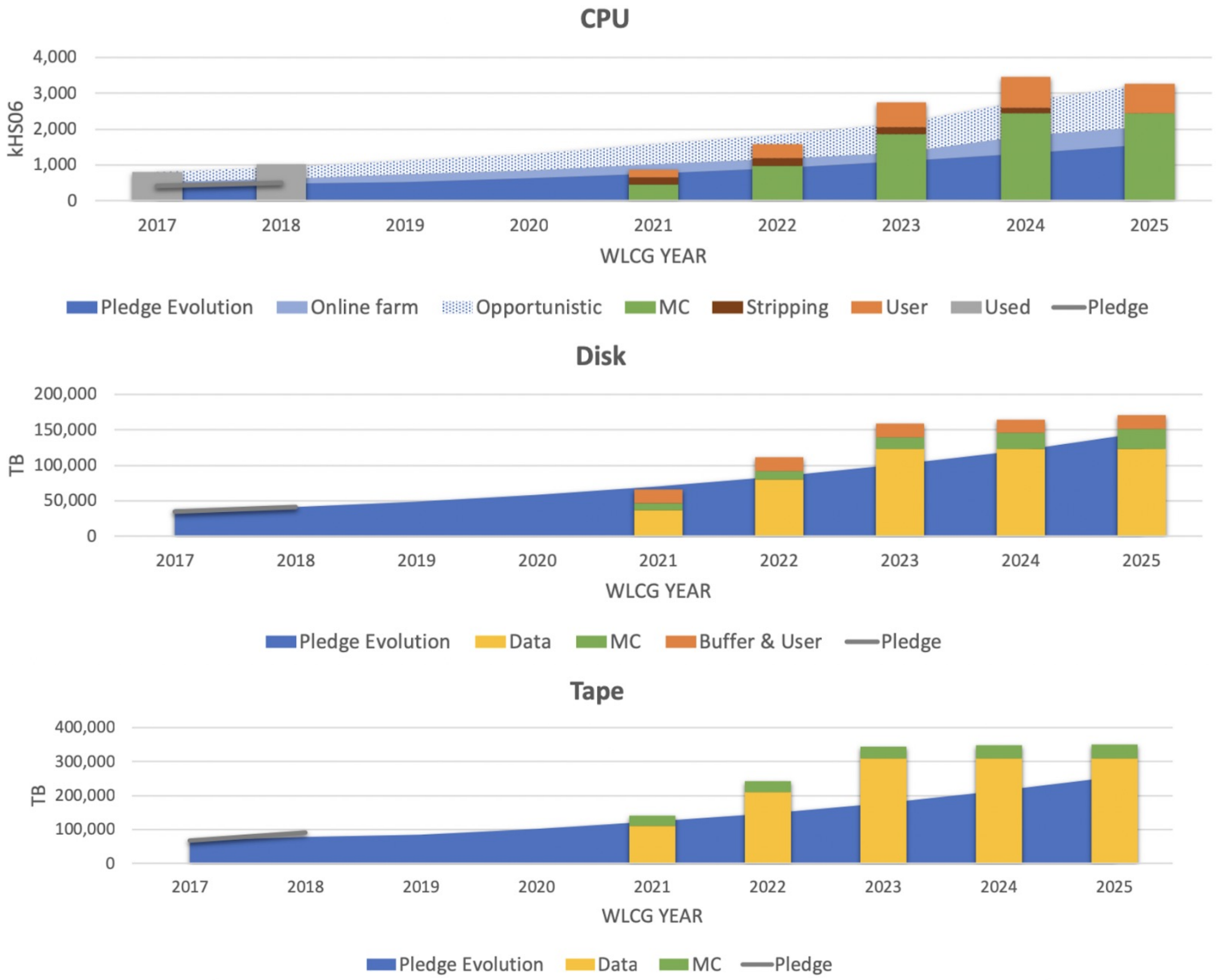
- CMS projections are not directly comparable to ATLAS one, as they are calculated with a different pileup level (200 for CMS for Run4 and 88-140 in case of ATLAS).

- Regarding the actions to face the CPU and Storage challenges for Run4 in CMS, it is essentially the same as ATLAS.

- A way to reduce storage needs, is to avoid dataset replications by accessing the data through caches (it is one of the DOMA activities) and process the data through buffers (the ATLAS data carousel model), in addition, of course, to use reduced formats for the analysis.



Future Perspective (HL-LHC) LHCb



- Resource diagrams that LHCb include in its TDR for Run3.

- For LHCb, the challenge is the Run3, not the HL-LHC. They have upgraded the detector and are going to collect a factor 10 more data in Run3 than in Run2.

Conclusions

- **Our perspective/objective is to have an infrastructure at the different sites for the contribution to computing at the Spanish level that has CPU and disk, and Tape for Tier1.**
  - **Spanish WLCG Tier1 provides ~5% of Tier1 data processing of CERN's LHC detectors ATLAS, CMS and LHCb (since 2018 is around 4% and it will be keep since 2022)**
  - **Spanish WLCG Tier2s provide ~5% of Tier2s data processing resources (since 2018 is around 4% and since 2020 will be the 3%)**
- **And also complement with the use of additional resources from:**
  - **HPC resources (BSC will host one of the first pre-exascale supercomputer in EU: ~200 peak Petaflops)**
    - **Collaboration agreement between Barcelona Supercomputing Center (BSC) and LHC Computing Spain to exploit a fraction of their resources for ATLAS, CMS and LHCb**
    - **LHC computing designated as one of the BSC strategic projects, to provide CPU time required for LHC simulation Spain (~55 Mhours in 2021)**
    - **ATLAS and CMS have opted for different solutions to overcome the lack of internet connectivity from the execute nodes @ BSC**
    - BSC has **some limitations** to run WLCG jobs:
      - Execute nodes do not have internet connectivity, hence it breaks late binding models used in WLCG
      - Not possible to install edge services (Squids for conditions and CVMFS [VO software], …)
      - Access to input data and/or handling output data is challenging
  - **Cloud Computing**
    - **Amazon-AWS (PIC) , but not very good for data intensive jobs**
    - **Cloud openstack (IFCA)**
- **in order to provide a consistent and appropriate contribution according to the overall Spanish participation in LHC for Run3 and Run4 (HL-LHC) periods.**

Special Thanks to WLCG-ES community:

- José Salt, Andreu Pacheco, José Hernández, Juan José Saborido, José del Peso, Esteban Fullana, Fco. Javier Sánchez, Josep Flix, Eugeni Graugés, Xavier Vilasís, Francisco Matorras, Antonio Delgado, ….

THANKS. QUESTIONS?

Kein Plan überlebt die erste Feindberührung

“Ningún plan sobrevive al primer contacto con el enemigo”

Mariscal Helmuth von Moltke
(1819-1888)