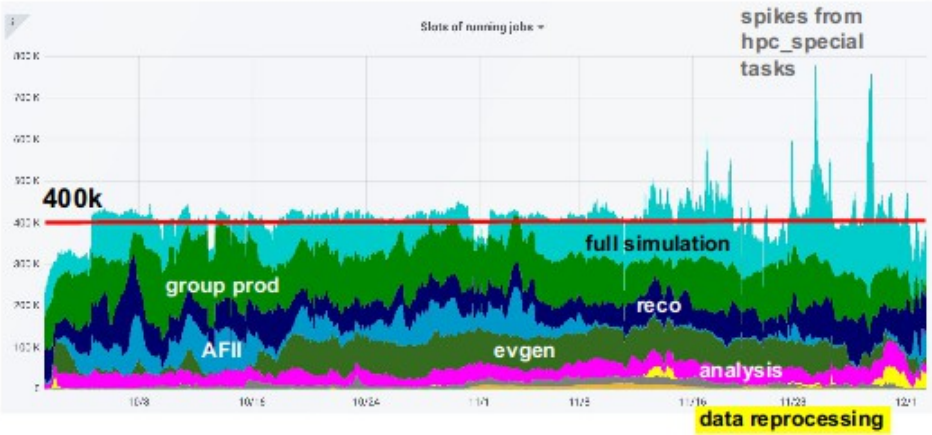
News from the ATLAS experiment

F. Derue, LPNHE Paris

Réunion des sites LCG France
11-13th December 2019, CC-IN2P3 Lyon

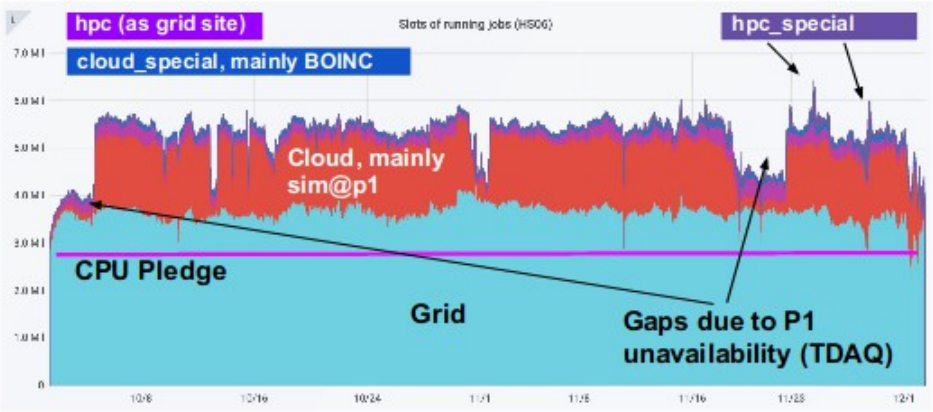
ATLAS CPU resource usage

Slots of running jobs Oct-Nov 2019

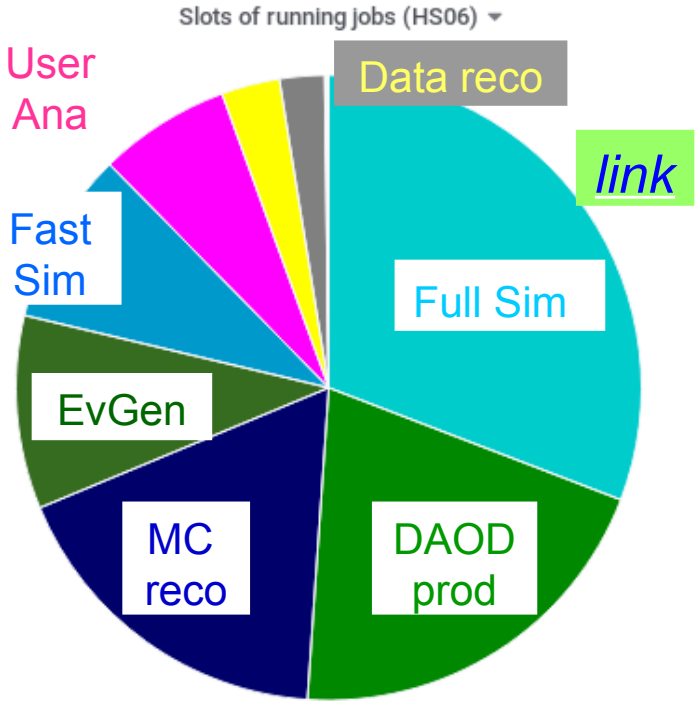


Slots (HS06) per resource type Oct-Nov 2019

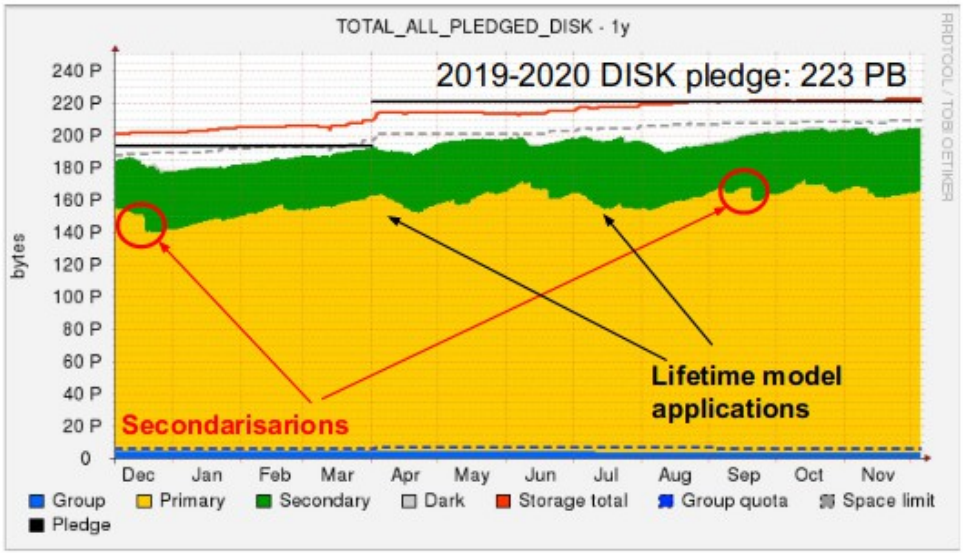
link

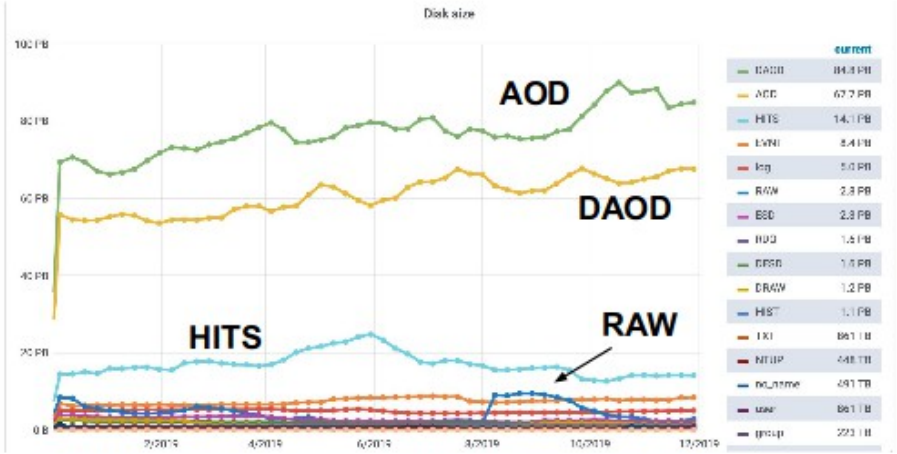


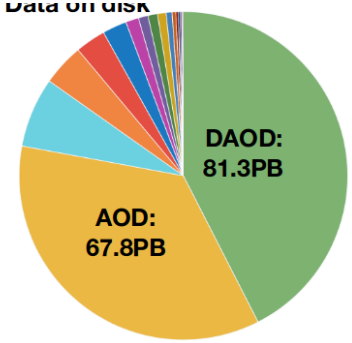
- **Smooth operation**
  - on the grid, HLT farm, T0, cloud, HPC
  - ~400 k jobs per day
  - ~65% MC simu, reco, evgen
- **Analysis**
  - 10-20% share on the Grid/Cloud - not HPC
- **Very large DAOD re-production campaign tailing off**
  - Next: big simulation campaig to run over Christmas, then reprocessing the RPVLL data
- **Expect to run in a similar mode for the next 18 months**

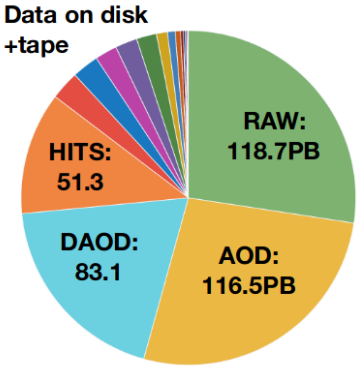


Storage resource usage









- **Storage evolution**
  - usual situation with storage, using full capacity to limit; 165PB primary, 40PB
  - majority of data on disk is DAOD and AOD (recent increases in RAW/AOD from repro)
- **Lifetime lodel**
  - applied lifetime model deletion on disk before the summer (next one planned for December)
  - applied lifetime model deletion on tape last summer (so tapes can be repacked before Run-3)
- **Data movement (per day)**
  - moving 1-2 PB (1.5M files) @15-20 GB/s
  - deleting 1.5 PB

LHCC/RRB : ATLAS request 2021

RRB approved ATLAS' provisional computing resources requests for 2021

- Summary: increases of 10% tape, 15% CPU, 20% disk

ATLAS		2019		2020			2021	
		CRSG recomm.	Pledged	Request	2020 req. /2019 CRSG	C-RSG recomm.	Request	2021 req. /2020 CRSG
CPU	Tier-0	411	411	411	100%	411	550	134%
	Tier-1	1057	1083	1057	100%	1057	1230	116%
	Tier-2	1292	1293	1292	100%	1292	1500	116%
	HLT	n/a	0	0	n/a	0	0	n/a
	Total	**2760**	**2787**	**2760**	**100%**	**2760**	**3280**	**119%**
	Others			0		0%		
Disk	Tier-0	27.0	26.0	27.0	100%	27.0	30.0	111%
	Tier-1	88.0	94.4	88.0	100%	88.0	107.0	122%
	Tier-2	108.0	101.2	108.0	100%	108.0	132.0	122%
	Total	**223.0**	**221.6**	**223.0**	**100%**	**223.0**	**269.0**	**121%**
Tape	Tier-0	94.0	94.0	94.0	100%	94.0	97.0	103%
	Tier-1	221.0	216.8	221.0	100%	221.0	249.0	113%
	Total	**315.0**	**310.8**	**315.0**	**100%**	**315.0**	**346.0**	**110%**

Software status and todo list

- **Software / release 22**
  - Multithreaded: Simu/Digi/Reco using the athenaMT/GaudiHive infrastructure to reduce memory consumption and makes it more feasible to investigate heterogeneous computing (e.g. use GPUs, FPGAs etc)
  - new job configuration with massive simplification of existing system, which has become unmanageable
  - detector Upgrade for Run-3 and beyond
  - Recently added: Port tracking CPU improvements developed in ITk studies to run 3 detector (and software)
    - Gains could be as big as a factor two for tracking - but not proven yet

Database status and todo list

- **Oracle version upgrade to 19c** *link*
  - the testbed INT8R is available for validation of applications

- **High priority R&D work**
  - monitoring and analytics of Frontier (conditions data) accesses
    - smooth operation in Lyon since October
    - about to replace the Oracle backend storage, next year (2020)
  - new WLCG squid operations support team
  - RESTful Conditions development ongoing
  - Ensuring the volume of DCS data is reasonable

- **From 2023, Oracle will change its cost model for its licenses - changing from site-wide campus to per-processor (and per product)**
  - The Oracle Golden Gate product is used only by ATLAS and from 2023 CERN IT will not be in a position to pay for it
    - Oracle Golden Gate: service enabling partial replication (e.g. selected DB schemas) of a database rather than the whole DB
    - used for synchronising the offline with the online databases within CERN as well as the Oracle-hosting Tier-1 sites
  - implies very substantial costs for ATLAS
  - CC-IN2P3 is a backup for CERN for Oracle ….

Computing status and todo list

- **Activities in 2019**
  - Reliable infrastructure and performance throughout the year, despite of (or enhanced by) successful commissioning of new features and workflows: CentOS7 migration, Pilot2, containers/Singularity, DPM/DOME migration, Rucio mover and reaper..
  - big load of work on syst admins !
  - French sites have made all these upgrades in due time, when not being among the firsts (e.g DOME)

- **Further projects in progress: Grand unified queues, DDM/DOMA projects, Data Carousel, User containers, HPC.**

- **Miscellaneous**
  - Future removal of SCRATCHDISK, to be replaced by DATADISK and user quotas
  - Rucio and kubernetes: offer benefits due to isolation of software within container environment. Some issues and lack of experience so far
    - use early 2020 to explore this and aim for full deployment in Q2 2020
    - pre-GDB on kubernetes next week

Panda queues Grand Unification

- **Panda job types**
  - single or multi-core, low/high memory, long/short
  - In late-binding pull-model, we must quantize these to a few categories
    - each needs PandaQueue(PQ) with stream of pilots requesting matching batch slot
    - submit pilots when activated jobs an a PQ
    - Local Batch System schedules them as it pleases
      - FIFO, % S/MCORE - not configurable by ATLAS. Not dynamically.
  - Problem is this does not follow ATLAS priorities/gshare
  - Stop submitting other pilot types, then BS can only run the ones we do submit

link

PQ = ABC_UCORE	Single core(SCORE)	Multi-core(MCORE)
RSS 2GB per core	2GB	PQ.corecount
>2GB RSS per core(HIMEM)	PQ.maxrss/corecount	PQ.maxrss

- **Grand unified queues (CERN-PROD_UNI)**
  - include analysis in the prod Unified Queues
  - allows highest priority work to use all cpu resources at a site
  - maximize data co-location
  - move less input data around
  - automatically add MCORE/HIMEM resources for Analysis
  - must remove any local batch system partitioning (and shares

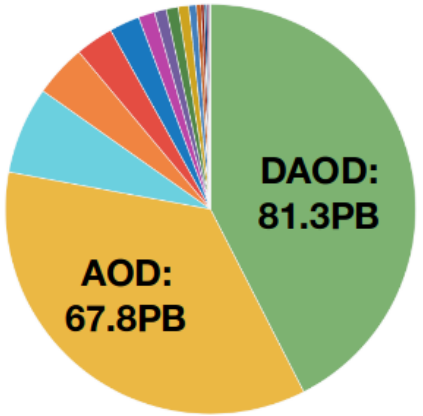
First tests/setup are ongoing (see *link*)

Production planning for 2020-2021

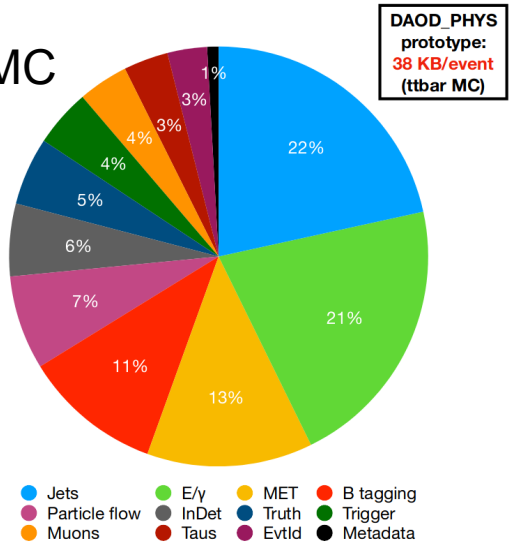
Sample	Activity	2020				2021			
		Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Run 2 data	DAOD production for ongoing analyses								
	Production of DAOD_PHYS for transition to new model								
	Reprocessing in release 22 + DAOD(_PHYS)(LITE)								
Run 2 MC	New production for ongoing analysis								
	Production of DAOD_PHYS for transition to new model								
	Reprocessing in release 22 + DAOD(_PHYS)(LITE)								
Run 3 data	Reconstruction at Tier-0 + DAOD(_PHYS)(LITE)								
2021 MC	Generation/simulation								
	Reconstruction + DAOD(_PHYS)(LITE)								
	Reprocessing with improved conditions								
2022 MC	Generation/simulation								
Upgrade MC	Generation/simulation/reconstruction + DAOD(_PHYS)(LITE)								

Changes for Run-3 analysis model

- **Run-2 analysis model is too expensive**
  - DAODs made from MC heavily overlap with each other
  - many DAODs are larger than originally anticipated
  - tracking, trigger and truth strongly dominate the size of most formats
  - analysis formats (AOD,DAOD) dominate the disk



- **Main AMSG3 recommendations**
  - Collect options to save at least 30% disk space overall, harmonise analysis and give directions for further savings for the HL-LHC.
  - replace most DAOD formats with a single unskimmed format (DAOD_PHYS) containing sufficient informations to support ~80% of analyses, with a target size of 50 kB/event
  - centrally run skims of this format foreseen for data, not for MC
  - exceptions foreseen for CP groups, and some physics use cases
  - introduce a very small calibrated format (~10 kB/event) for fast analysis and as a Run-4 prototype
  - apply lossy compression where possible, both to AOD and DAOD (up to ~25% gain)
  - store AODS primarly on tape rather than disk, using a data carousel to access them when needed



DDM/DOMA projects (2020)

- **Third Part Copy**
  - commission non-gsiftp transfers for at least 2 production sites (using multihop if necessary)
  - XrootD Tests
  - WebDAV Tests
- **Access & Caching**
  - commission XCache for storage-less sites
  - scalability, easy deployment, easy AGIS config
- **QoS**
  - Data Carousel and BNL MAS are two concrete use cases
  - for MAS: Do tests and measure performance/impact

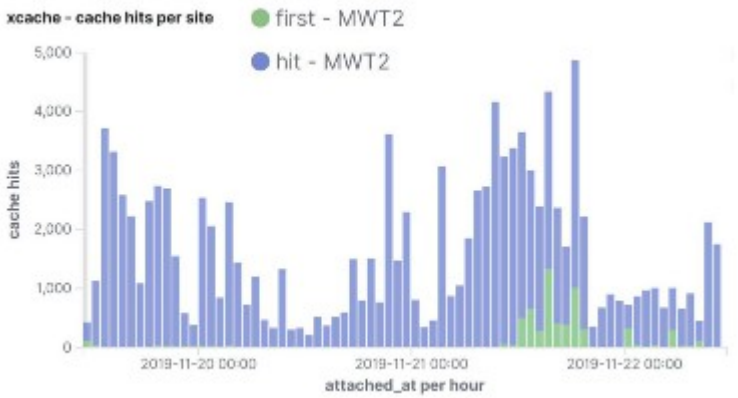
Data Carousel

- **The idea**
  - store most or all the AOD data on tape storage, which is much less expensive than disk. DAOD production campaigns would read the input AODs directly from tape
  - if successful, will lead to ~2X savings in storage costs
  - It will also have an impact on analysis workflows as physicists will not have direct access to the input AODs, except as part of centrally scheduled DAOD productions

- **Summary of information from sites [*link*], after pre-meeting survey and chat with T0 and T1s**
  - highlighted many issues, problems, concerns, differences and challenges but also potential benefits and improvements as a result of data carousel preparation and throughput stress tests
  - dCache was designed to get data to tape as fast as possible, not necessarily optimised to recall quickly as needed by Data Carousel workflow VM/container workflows
  - does the staging profile needs to be changed/adjusted, thresholds adjusted, before next tests?
- **Preparation for Data Carousel mode reprocessing of full Run 2 RPVLL in January 2020**
- **Development to continue, particularly in lines of communication between the involved parties**

Data Placement and Networks

- **Virtual Data Placement: Utilising xCache and Data Lakes to reduce stress on FTS**
  - Dataset assigned to N sites within Cloud (local Data Lake), but only files only moved (copied) if accessed
  - reduces stress on FTS and creates fewer rucio rules
  - deployed at 4 sites, future test on large site with only xCache
  - xCache stability required and care with additional transfers !

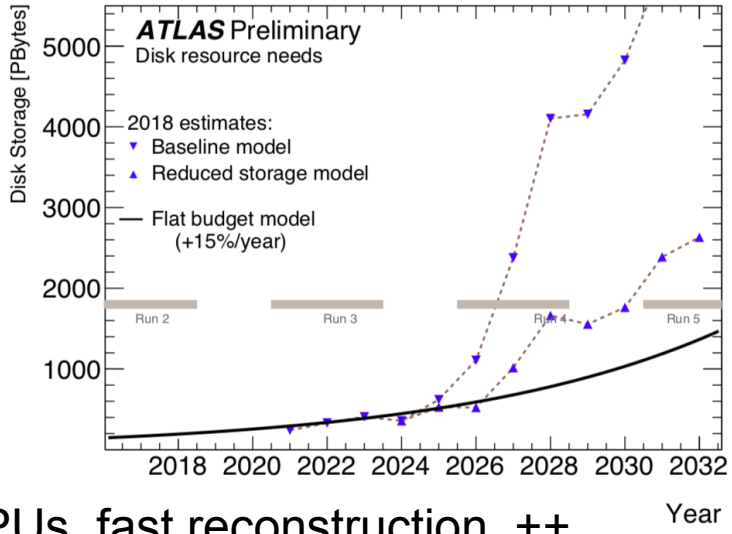


- **Networks: Preparation for upcoming *LHCOPN/LHCONE meeting* on future requirements (in January 2020)**
  - not only bandwidth: services, orchestration, visibility. ATLAS twiki, NVF WG report
  - how cloud-native networking looks in future: paradigm shift to current VM/container workflows
  - how to massage the network to the needs: Now, Run 3, HL-LHC. Marking network jobs with IDs?

Run-4 preparation

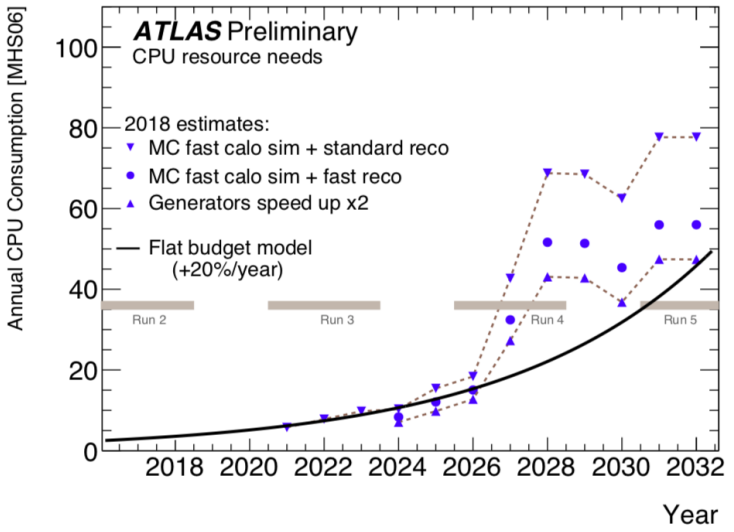
- **HL-LHC extrapolation**
  - HL-LHC will see up to 10x more luminosity and 4x as much pile-up, with concomitant simulation requirements
  - higher precision modelling needed
    → more computing resources for event generation

ComputingandSoftwarePublicResults



- **Technology evolution**
  - won't be enough - we need to think differently. Affects computation and storage
  - tiny data formats, fast chain simulation, HPCs, GPUs, fast reconstruction, ++

- **LHCC review**
  - HL-LHC computing in 2022/23
  - in preparation ATLAS is writing a conceptual design report, « WritAthon » during S&C week (last week),
    For ATLAS in Feb,
    For LHCC reviewers in May



Meetings CAF-ATLAS FR users

We organized the annual meeting on 28th Nov. 2019 (*agenda*)

- about 20 persons came to IPNL Lyon
  - mostly group leaders, CAF representatives + ATLAS France resp.
  - a few connected by visio
- morning focused on
  - review on computing/storage usage
  - feedback from each laboratory
- afternoon focused on
  - Machine Learning
    - for detector, simulation and reconstruction studies
    - for analyses
    - discussion : need for training/tutorial
  - preparation of S&C TDR and other activities
    - R&D activities : DOMA-FR, tracking for Run-3/4
    - other activties : AMI, Databases, etc ...

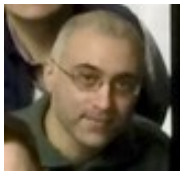
The hard news

- **Foreseen end of LPSC Tier-2**
  - news ~2 months ago that LPSC Tier-2 will close by 2023
  - first ATLAS Tier-2 site which will close

- **Change of person as Tier-1 support**
  - after many years of involvement as our Tier-1 manager Manoulis Vamvakopoulos is taking over

a great thank for all his work !

  - a lot of specific technical expertise was needed
  - we expect of course same level of support to ATLAS Tier-1 from CC-IN2P3
  - new person in charge (since a month), Aresh Vedaee, previously at PIC (and CC before)



Conclusion

- **CAF group and activities**
  - some ongoing changes in the list of CAF representatives
  - also involved in support for FR-cloud !
  - new person in charge of support for Tier-1 activities

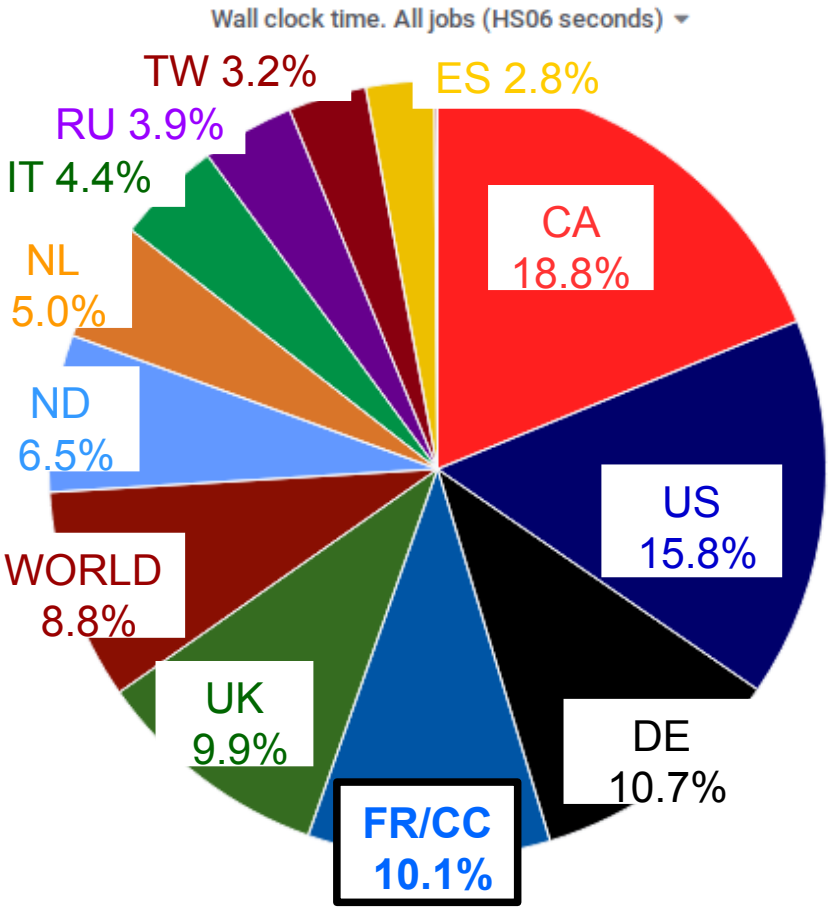
- **S&C in ATLAS, preparation of Run-3 and Run-4**
  - for Run-3 new analysis model will be used, in particular fewer/smaller DAODs
  - preparation of Run-4 within WLCG(-FR) and DOMA(-FR)
  - many tools/projects developped in // for ATLAS or LHC are now used in non-LHC collaborations : new organization of WLCG (less LHC specific), ESCAPE (vs DOMA), some particular tools (e.g AMI)
  - **No planning for ATLAS for the coming changes**
    We will keep you informed through LCG-FR Tech
  - **Next ATLAS Software and Computing Week, with a focus on sites:**
    10-14 February 2020

- **S&C in France**
  - CC-IN2P3 as a Tier-1 represents ~10-11% of ATLAS cpu and storage on grid
  - French Tier-2s represent ~8% of cpu and 12% of storage of Tier-2s on grid

CC-IN2P3 as a Tier-1 (CPU)

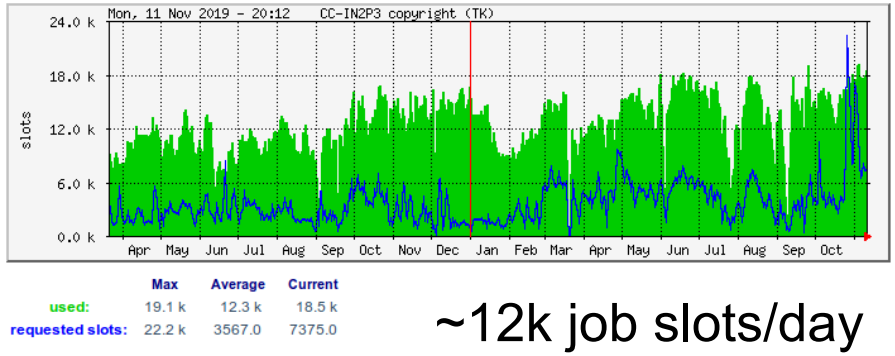
CC-IN2P3 is one of the leading ATLAS Tier-1

T1 only *link*



	Pledge 2019 (HS06)	Pledge 2020 (HS06)
CC-IN2P3	118500	134000
USA		
UK	132125	156436
Germany	132125	132125
Canada	105700	105700
Italy	90270	95130
Netherlands	80332	80074
Nordic	63470	63190
Taiwan	43000	44300
Spain	42300	42300
Russia	32800	32800

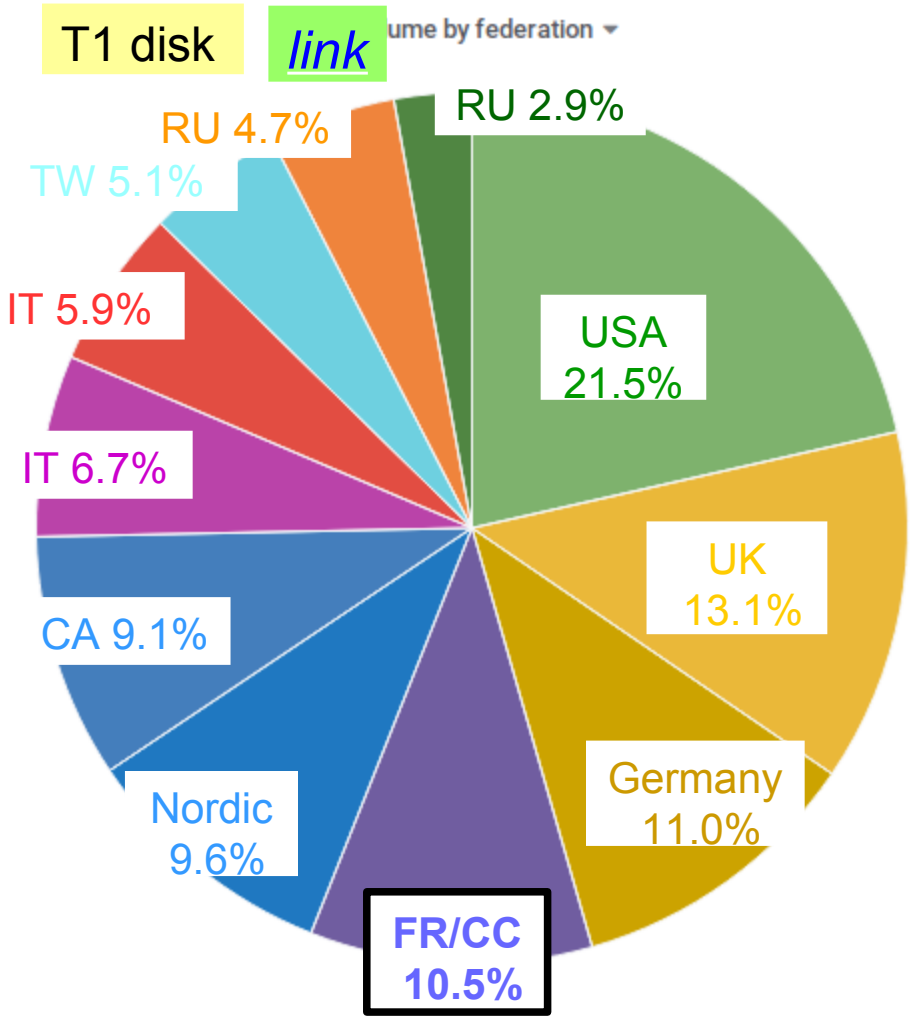
cctools view [*link*]



~12k job slots/day

CC-IN2P3 has delivered 110-20% of its pledge since a year

CC-IN2P3 as a Tier-1 (Disk/Tape)



	Disk Pledge 2020 (TB)	Tape Pledge 2020 (TB)
CC-IN2P3	11100	28700
USA	19000	39500
UK	13000	32700
Germany	11000	27600
Canada	8800	22100
Italy	7900	19900
Netherlands	6500	16100
Nordic	5400	12500
Taiwan	7100	0
Spain	3500	8500
Russia	4500	5700

CC-IN2P3 is one of the leading ATLAS Tier-1s

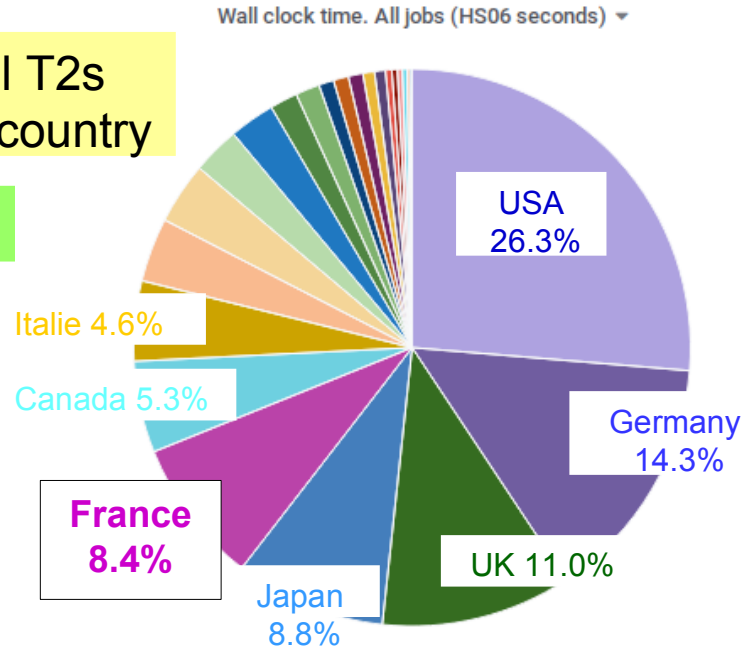
FR-cloud and French Tier-2s (CPU)

- **Tier-2 operations are hold by cloud**
  - FR-cloud corresponds to France, Japan, Romania, China
    ~18 % of CPU in Tier-2s
- **Looking at French sites (only)**
  - French Tier-2s represent ~8% of ATLAS CPU in all Tier-2s

	Pledge 2019 (HS06)	Pledge 2020 (HS06)
CPPM	14000	18000
GRIF	46930	47530
GRIF-IRFU	22600	22600
GRIF-LAL	15000	15000
GRIF-LPNHE	9400	10000
LAPP	21000	24000
LPC	11800	11800
LPSC	13325	14500

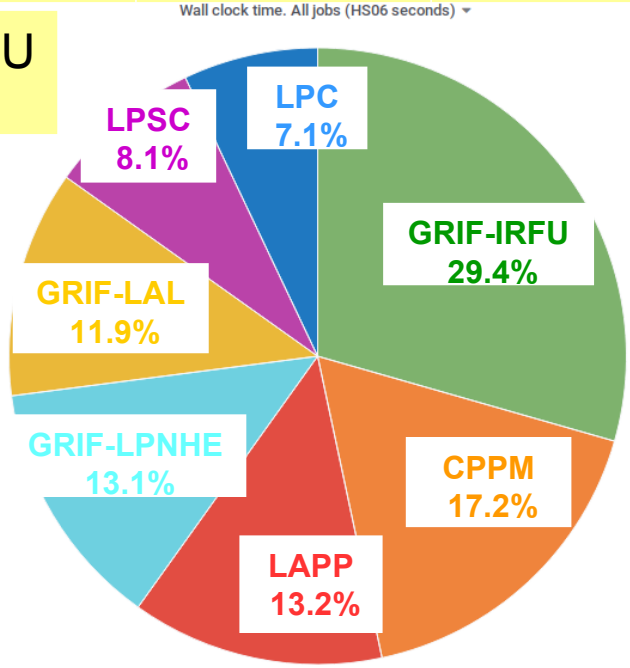
All T2s per country

link



Fraction of CPU used in T2s

link



French Tier-2s (Disk)

- **Tier-2s in France**
  - pledges for DATA+SCRATCHDISK ~12% of available disk in Tier-2s

- **LOCALGROUPDISK in Tier-2s**
  - local resources not pledged
  - large differences in volumes between sites [*current*, table]

	Pledge 2020 (TB)
CPPM	1800 (16.4%)
GRIF	3950 (36.1%)
GRIF-IRFU	1800 (16.4%)
GRIF-LAL	1100 (10.0%)
GRIF-LPNHE	1050 (9.6%)
LAPP	2550 (23.3%)
LPC	1500 (13.7%)
LPSC	1150 (10.5%)

	Total (TB)	Free (TB)
CPPM	245	90
GRIF-IRFU	396	25
GRIF-LAL	26	10
GRIF-LPNHE	320	30
LAPP	88	50
LPC	22	12
LPSC	82	22

Software & computing involvement in France

- **Involvement in computing**
  - ~10 FTE for FR-T1/2 (Class 4)
    CC (3.30), CPPM (0.70), IRFU (1.15),
    LAL (0.30), LAPP (1.30), LPC (0.75),
    LPNHE (1.10), LPSC (0.90)
  - ~0.6 FTE for FR-cloud support (Class 3)
    CPPM (0.05), LAL (0.15),
    LPNHE (0.3), LPSC (0.1)
  - ~0.75 FTE for FR-cloud management (C3)
    CPPM (0.05), IRFU (0.1) LAL (0.1),
    LAPP (0.05), LPNHE (0.2), LPSC (0.05)

- **Involvement in software (based on OTP pages report)**
  - ~10 FTE (Class 3)
    SW core, detector (upgrade), analysis
    CPPM (1.5), IRFU (0.7), LAL (2.4),
    LAPP (0.9), LPNHE (1.9), LPSC (2.5)
  - details on SW core
    - AMI
    - EventIndex
    - ROOT compression
    - b-tagging, Run-3, Multithreading

CommitmentFunding Task

- **Responsibilities**
  - International Computing Board (ICB)
    - members : A. Formica, F. Derue
    - Software & Computing Resource Scrutiny Group (SCRSG) : S. Jézéquel
  - Computing shift organization
    S. Crépé-Renaudin

- **Talks and publications** *[link]*