News from the ATLAS experiment

F. Derue, LPNHE Paris

Réunion des sites LCG France
22-24th May 2019, LAPP Annecy-Le-Vieux

Introduction

- **ATLAS/LHC coming milestones**

- **HL-LHC**
  - 2019: LHCC Review in Spring on HL-LHC S&C
    ATLAS will prepare a short document to accompany the WLCG strategy document
  - 2022: ComputingTDR
    → postponed from 2019 to 2022
    → a priori WLCG. Not clear yet if ATLAS part of it or dedicated TDR

⇒ need to scrutinize efforts in Software/Computing to be put by FR labs

Meetings CAF-ATLAS FR users

There was no more meeting between CAF (Computing ATLAS France) and physicists/users since ~2012.
We organized such meeting on 27th Nov. 2018 (*agenda*)

- about 20 persons came to IPNL Lyon, a few connected by visio
- focused on feedbacks from each laboratory and some analysis
  - ⇒ ~10 FTE in computing (engineers and physicists) and ~10 FTE in software (mostly by physicists)
  - ⇒ ~200 TB of storage needed for all analyses (for one cycle/version) with usage of large variety of ressources (grid, CC, CERN, local)
  - ⇒ global satisfaction with available ressources
  - ⇒ larger use (than expected) of local ressources
  - ⇒ on coming needs for GPU ressources
- will organize such meeting on annual basis
  - → next meeting (28th November) will focus on machine leearning and tracking

Grid and non grid resources in FR labs

Some « statistics » of common effort from our labs putting all numbers together

T2	TB	HS06
CPPM	1500	14000
IRFU	1620	22600
LAL	1000	13000
LAPP	2100	21000
LPC	1400	14000
LPNHE	1000	9400
LPSC	850	13300
Total	~10000	~100000

- T2 resources are known from pledges,
- the « other grid » resources represent an additional typical T2 !
- local resources represent an additional typical T2 !

other grid	TB	HS06
CPPM	200	10000
IRFU	360	
LAL	26	up to 30000
LAPP	100	0
LPC	50	yes
LPNHE	300	12000
LPSC	85	
Total	1100	~30000 ?

Local	TB	cpu
CPPM	200	~100 cores
IRFU	200	~400 cores
LAL	10	~160 cores
LAPP	2	batch shared with grid
LPC	210	~300 cores
LPNHE	100	~30 cores
LPSC	45	~13 servers
Total	~800	1000 cores

Institutional commitments

- Inputs
  - Class 3: Management, Database, Distributed Computing , Software Project, Squad support
  - Class 4: Tier-1 & Tier-2 sites operations tasks
- Status
  - increase by 10% in required S&C Class 3 manpower since 2016
  - person-power allocated remains constant
  - lack of allocated vs required : ~85 % for Class 3

Class 3

	Alloc (FTE)	Req (FTE)	Alloc/Rq [%]
2016	188 (12.6)	197	95
2017	190 (12.0)	216	88
2018	189 (12.4)	217	87

all ATLAS

FR

Class 4

	Alloc (FTE)	Req (FTE)	Alloc/Rq [%]
2016	169 (17.2)	182 (12.0)	92 (143)
2017	153 (9.2)	155 (10.4)	99 (88)
2018	150 (9.3)	154 (10.4)	97 (89)

CommitmentFunding

Task

French S&C (2018)

Class 3 for FR-Cloud Op : ~1.3 FTE

Semester	Persons (FTE)
1st	E. Le Guirriec (0.15) ; S. Crépé (0.1), L. Poggioli (0.45)
2nd	J-P Meyer (0.1), F. Derue (0.1), L. Poggioli (0.3), C. Biscarat (0.05)

Class 4 for FR-T1&T2s : ~9.3 FTE

Institute	FTE	Fraction of ATLAS members
CC-IN2P3	3.30	E. Vamvakopoulos (0.8)
CPPM	0.60	E. Knoops (0.4)
IRFU	1.15	J-P Meyer (0.15)
LAL	0.30	
LAPP	1.25	S. Jézéquel (0.1), F. Chollet (0.1)
LPC	0.75	
LPNHE	1.10	F. Derue (0.1)
LPSC	0.85	S. Crépé (0.1)

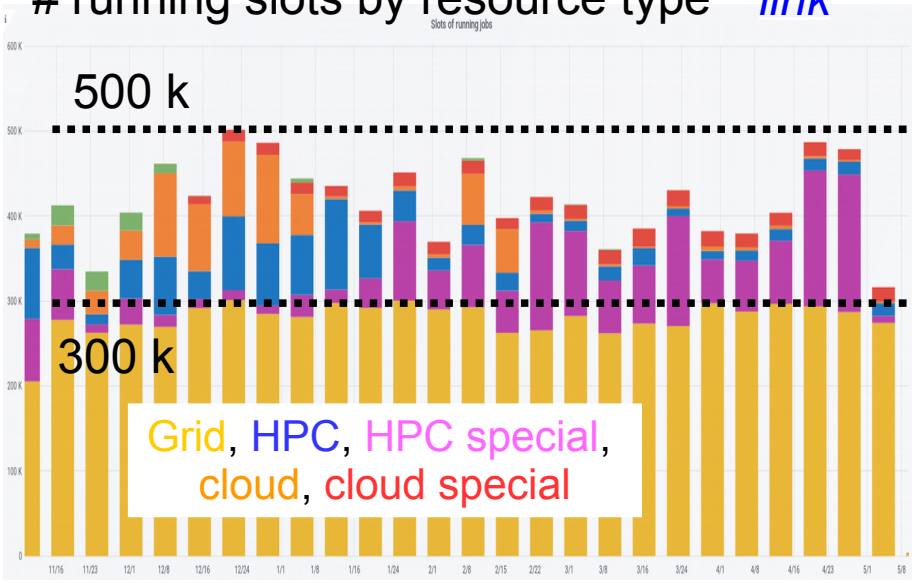
Class 3 for FR-software : ~11.2 FTE

Institute	FTE
CPPM	2.40
IRFU	0.80
LAL	2.75
LAPP	2.25
LPNHE	2.20
LPSC	0.80

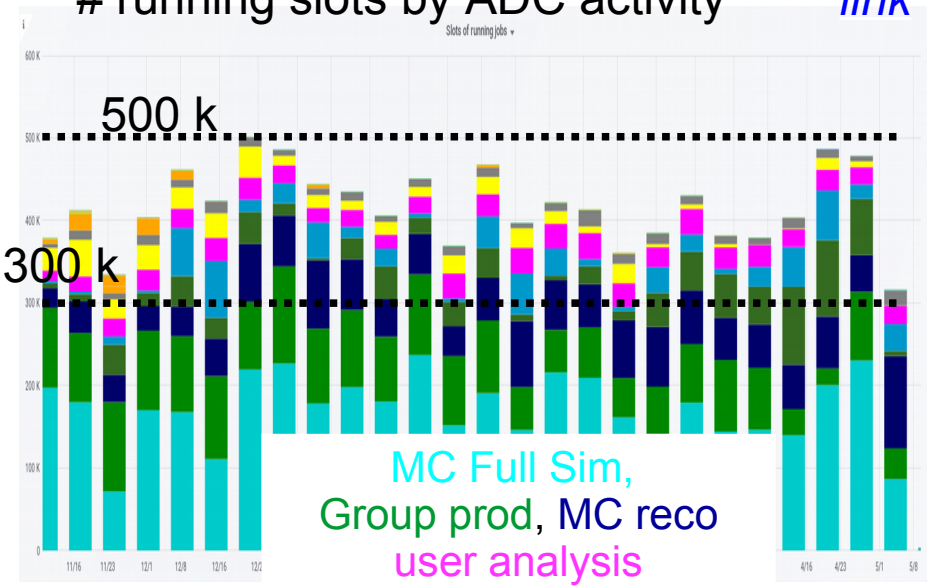
Implication of ATLAS French groups represents about 10 FTE in computing and 10 FTE in software (from few engineers in core software/application and many physicists spread in many different activities)

CPU delivered since last meeting / 6 months

running slots by resource type *link*



running slots by ADC activity *link*



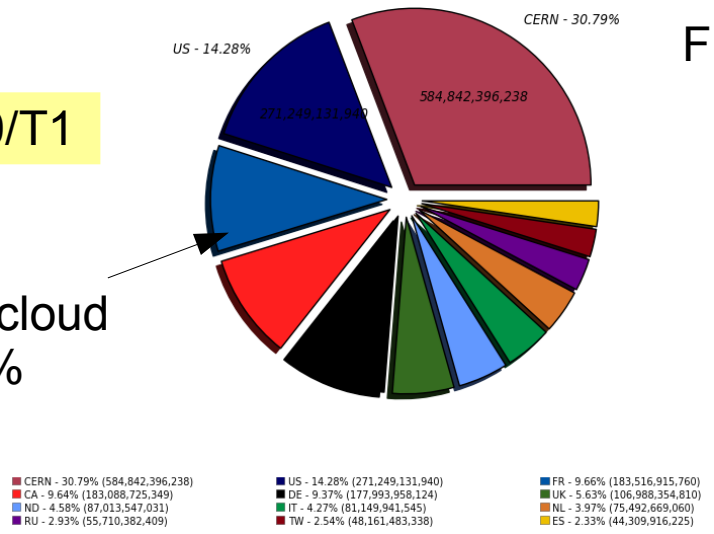
Good performance of ATLAS in usage of various computing resources

- >300 k jobs per day
- dominatedby MC (simu, reco, evgen)
- 1M jobs/day-45% isuser analysis

CPU delivered since last meeting / 6 months

Detailed and evolution of pie charts are shown at each CAF meeting (~2 months)

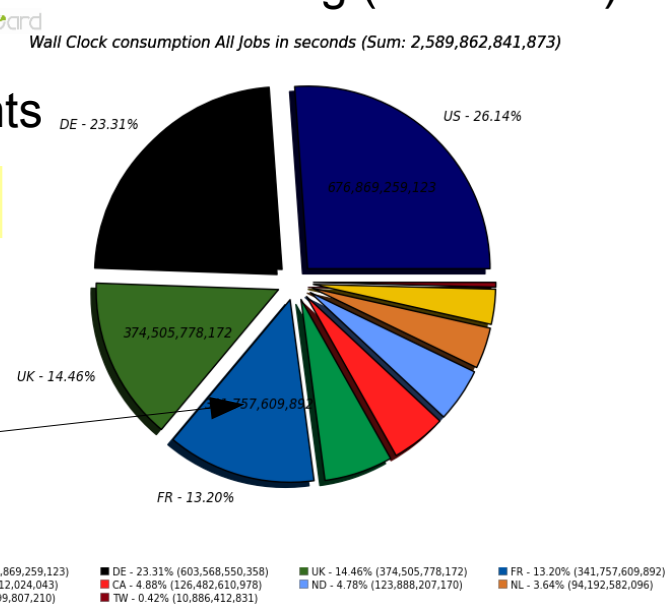
T0/T1



FR-cloud
9.7%

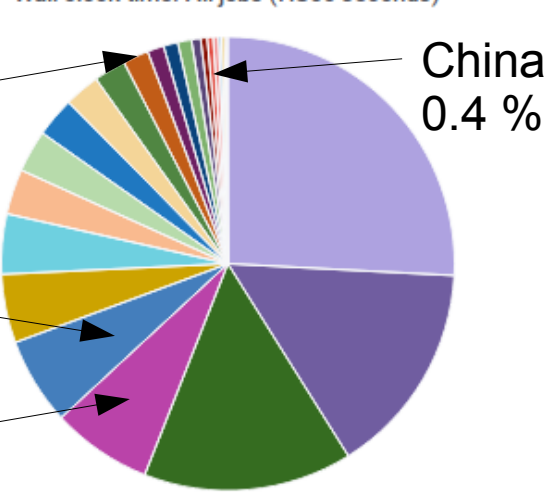
FR-cloud represents
- France
- Japan
- Romania
- China

T2



FR-cloud
13.2 %

T2 *link*



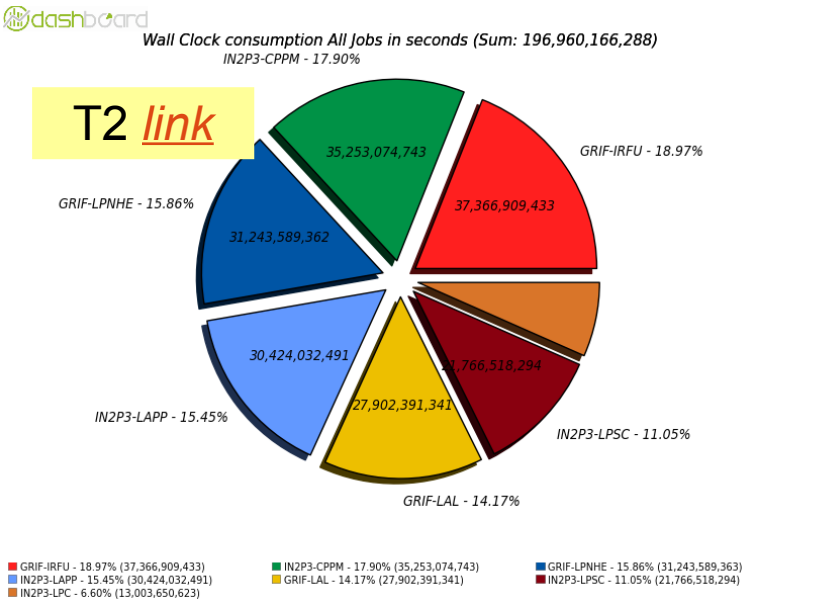
Romania
1.0 %

China
0.4 %

Japan
6.2 %

France
7.2 %

T2 *link*



Non standard computing resources

- **Cloud computing**
  - → see dedicated talk by Aurélien for overview in France
  - → little use of academic cloud, but could be used for local resources ?
  - → some sites have done tests on Google and other private clouds

- **Volunteer computing**
  - → easy use of BOINC even for production
  - → could be installed to use local machines from our labs ?

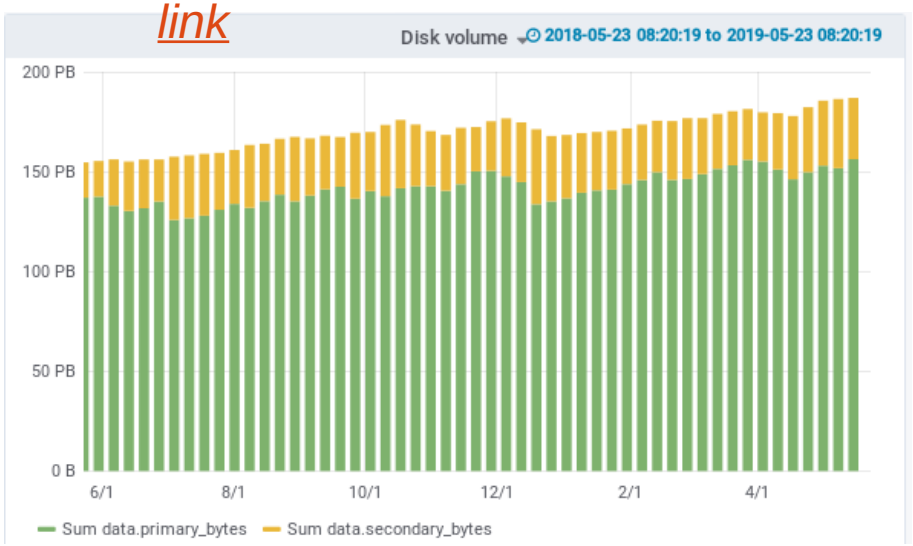
- **HPC**
  - → surprisingly (as reported during Jamboree), a large number of sites have deployed or could offer access to GPUs (18 sites from few to few 100)
  - → ATLAS encourages sites to provide GPUs
  - → GPU usage in ATLAS France
    - talk on usage of GPU in ATLAS France at GPU in April (*indico*)
    - resume of HPC/GPU resources in France at CAF meeting of 1st April [*slides*]
    - several analyses in France already used >1 year-GPU

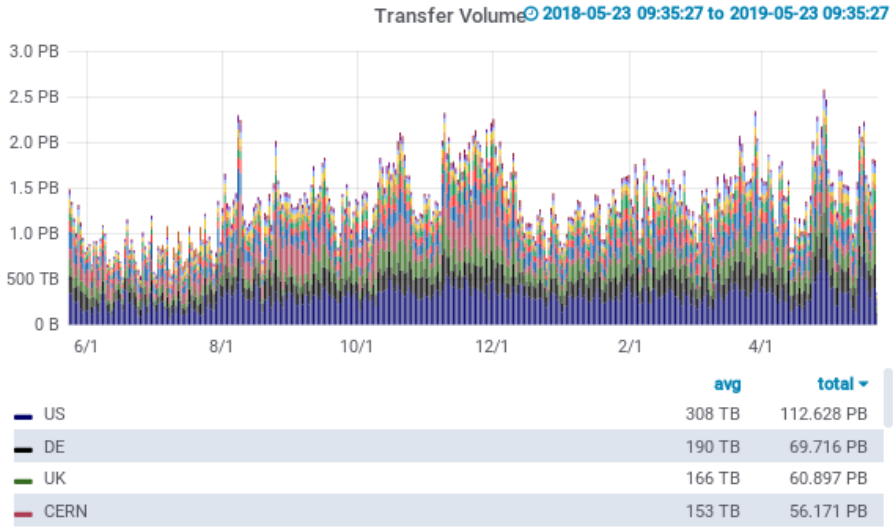
Usage of storage resources

Disk (~190 pB)
- almost full, 2018 pledge 186 pB
- dominated by AOD&dAOD

Tape (150 pB)
- RAW data & MC RDO

link





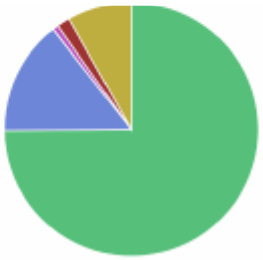
Per day
- moving > 1 pB/day
  1.5 M files, @20 GB/s
- deleting 1.5 pB

Run 2 analysis

- **Run 2 model was very successful**
  - many derived AOD (dAOD) formats : ~100
  - dAOD widely used

⇒ checked among French physicists thanks to survey done before CAF-user meeting

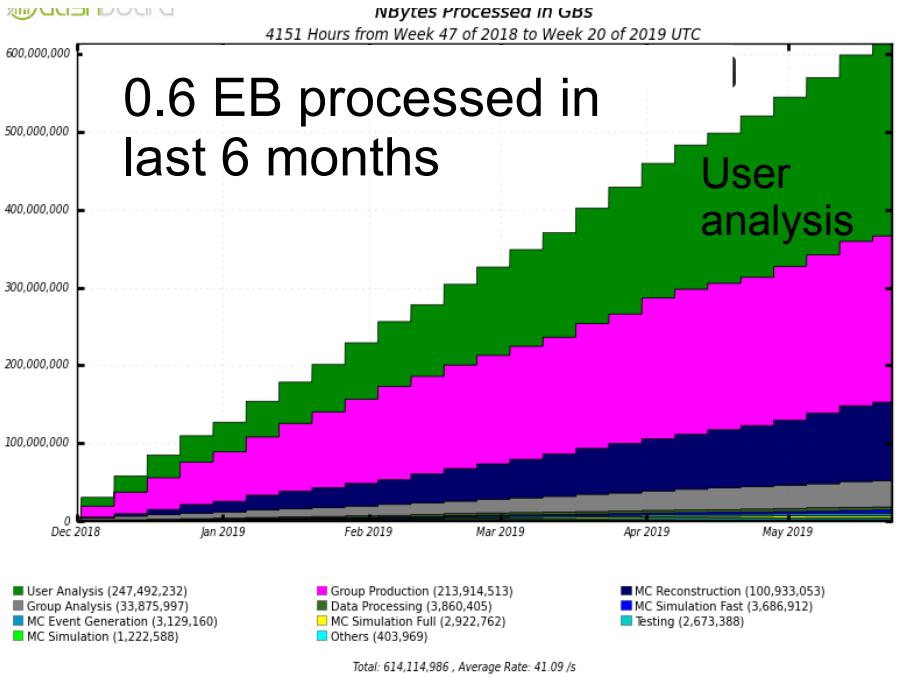
WallTime per input type



events per input type

- **Analysis**
  - constant 20k users jobs
  - processed data dominated by analysis

⇒ many/most of French physicists are using local resources to a non negligible extent



Run 2 analysis : issues and progress

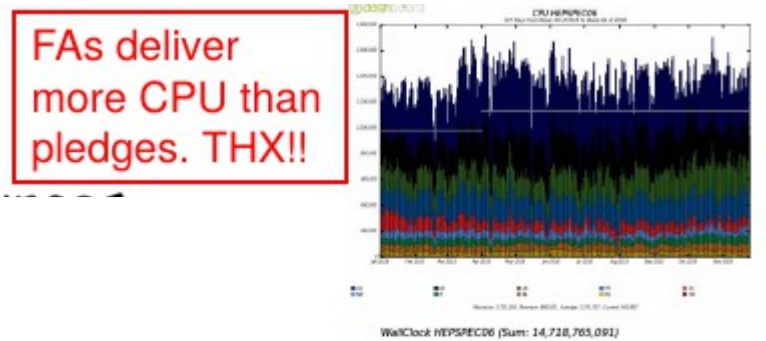
- **Assumed**
  - derivations run fast enough → able to run new full derivations every week if necessary
  - derivations output small → analysis jobs (on the grid) able to process all data/MC in about a day

- **In real**
  - ~6 weeks to reprocess all data into about 100 dAOD
    → run only major productions
  - analysis jobs on grid have tails >> 1 day
    → users write much larger outputs to reduce number of grid iterations
    ⇒ in French labs size of outputs for one iteration ~200 TB !

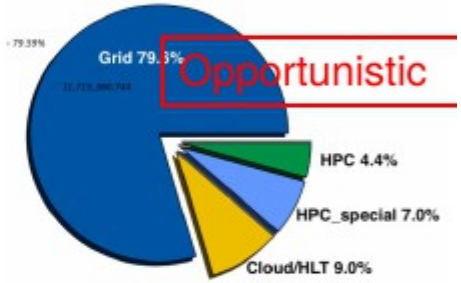
- **Progress**
  - dAOD production : more efficient merging, smaller dAOD
  - dAOD processing : improvements in distributed analysis (global shares, dynamic data placement)

Improvements for physics analyses

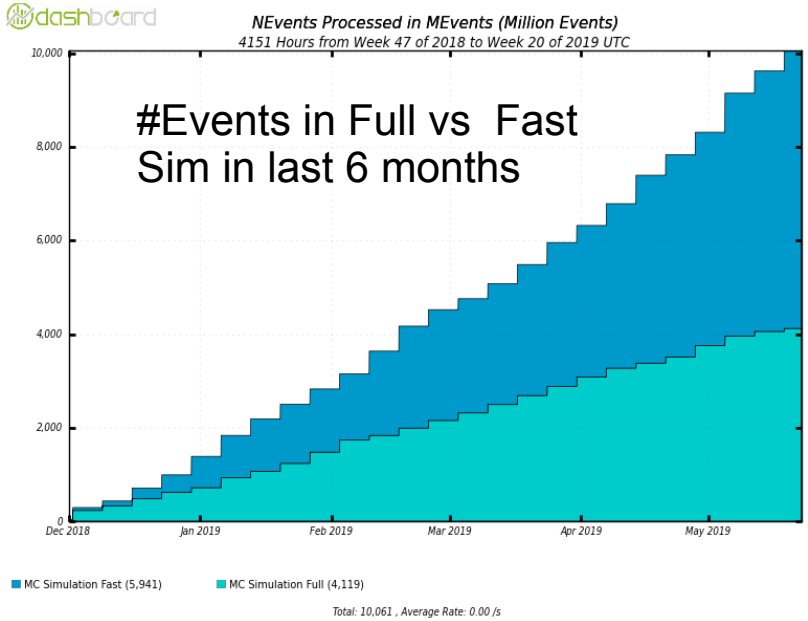
- **more MC statistics**
  - more grid resources
  - hunt for opportunistic resources
  - fast simulation



- **handling higher pileup**
  - improve recnstruction memory usage (AthenaMP for Run-2, Athena-MT for Run-3)
  - overlay for MC (needs new Condition Database scheme)



- **improve analysis**
  - more CPU resource
  - optimize workflows (tape usage)
  - reduce storage needs



ADC & sites configuration & recommendation

ATLAS sites jamboree and HPC strategy, 5-8th March, CERN [*indico*]

- **Migration to CentOS7**
  → most of the remaining RH6 sites plan to migrate to CentOS 7 by 1st June 2019, for France remain LPSC and LPNHE
  → for containers and pilot2 sites have to deploy singularity everywhere: (2.6.*)
    ⇒ as discussed in previous LCG-FR Tech more tests are needed from our side to check the exact required configuration

- **Nodes, batch**
  → most of the sites have migrated to HTCondor or SLURM
    ○ some are planning to do so this or next year - in some cases it is coupled to migration to CentOS 7
    ○ other batch systems are considered deprecated by ATLAS

- **Computing Elements**
  → the recommended CEs for ATLAS are CondorCE and ARC-CE

- **Storage Elements**
  → most of the sites use DPM, dCache and Storm
  → ATLAS is using gridftp, https and xrootd

Lightweight sites

- 4 presentations on this subject on thursday [*link*]:
- Recommendation (ICB-2018) to redirect funding from storage to CPUs for lightweight Grid site : 2019 limit is 520 TB
  - already implemented on volontary basis on 5 sites which also had small amount of CPUs (<200 cores)
    - 3 T2s with pledge CPU resources (HEP-UIBK, RO-14,RO-16)
    - support is now ~0 (and no worry about future storage migrations)

  Question : Can this setup be extended to larger Grid sites (200-1000 cores)
- Focus activity to low IO production jobs
  - Lightweight Grid site represents ~15k (5%) of Grid CPU capacity
    → Transfor to diskless would not affect significantly high IO jobs processing
  - Low IO → almost not affected by network occupancy
    → lower operational burden
  - No more analysis queue (currently~2%)→No interest to keep DATADISK
  - Migrating to ATLAS@Home/BOINC would be even simpler solution (monitoring report issue solved recently)
- Up to the cloud or country coordination to drive reorganisation
  - Convince FA that it is optimal usage of funding
    ⇒ on going discussion in LCG-FR for future of sites

Network

- **IPv6 Deployment** → done for all French sites
  → most of the Storage Elements are accessible through IPv6
  → some sites have difficulties with deployment

- **Network connectivity** : dedicated talk [link] → mostly on Run-3
  → most of the sites are connected with 10Gb/s or 20Gb/s WAN links or even faster, many considering upgrading to 100Gb/s in the next 2 years

Network Capacity Estimation for Run3

- Assume analysis on an HT-Core (job-slot) consumes 1.2 MBytes/sec
  - Implies job-slots need that level of network bandwidth to storage
  - WAN access to remote storage at 20% (ATLAS avg now)
    - 10 PBytes/day, 8 PB LAN, 2 PB WAN from Mario
  - **Minimal Tier-2**: 1000 job slots => 1.2 GBytes/sec, WAN 1.6 Gbits/sec
  - **Nominal Tier-2**: 5000 job slots => 6 GBytes/sec, WAN 9.6 Gbits/sec
  - **Leadership Tier-2**: 10000 job slots => 12 GBytes/sec, WAN 19.2 Gbits/sec
  - **NOTE**: Run-3 will have 3-4 times the data...have to either increase cores or improve average software throughput by that factor
- **Summary Network Capacity Recommendation:**
  - Average numbers above need a burst capability, assume x3
  - **Minimal Tier-2 WAN**: 1.6 Gbps x 3 = 4.8 Gbps => **10G link**
  - **Nominal Tier-2 WAN**: 9.6 Gbps x 3 = 28.8 Gbps => **40G link**
  - **Leadership Tier-2 WAN**: 19.2 Gbps x 3 = 57.6 Gbps => **80G link**

7

Monitoring and analytics

- 2 dedicated presentation by T. Beermann [*link*]
- **DDM Transfer Monitoring**
  - *DDM Transfers* is the main dashboard to be used in day-to-day to work
    - it includes the efficiency matrix, tables/ plots / details for transfer / staging / deletion
    - additionally, includes rucio and FTS submission queues
  - *DDM Transfers (Historical Data)*
    - includes the same plots as the main dashboard without the matrix and the queues
- **DDM Accounting**
  - global accounting for all pledged DATADISK and TAPE (all DISK is coming)
    - *DDM Global Accounting (Snapshot)* shows the current accounting numbers for the past week
    - *DDM Global Accounting (Historical)* shows the evolution of volume and files. Currently data available since beginning of last year
  - Site accounting is meant as a replacement for the old DDM accounting
    - *DDM Site Accounting* provides daily physical accounting per RSE. A lot of additional groupings like datatype, account, campaign, topology available
- **Job Accounting**
  - all plots available in one dashboard
    - provides plots for submitted, pending, running, finalizing and completed jobs
    - pledges have been added
    - additional plots available for cpu consumption / efficiency, processed data, success/failure rates and resource utilisation

Database

- Dedicated talk on subject [*link*] and parallel session - thanks Andrea for the summary !

1) Activities for Run3

COOL will be in use for Run3 but we are preparing for the future, coordinating with system experts to simplify the present COOL storage methods.
Progress in the tools for the migration to Crest.
Studies and activities on going (based on Frontier monitoring results) to improve the caching of our infrastructure (Frontier/SQUID) by introducing more cacheable requests on the client side.

2) Frontier infrastructure updates since Dec 2018

New Squid-4 deployed at CERN
New infrastructure at RAL, with 3 new services which will replace the old ones (RAL will dismiss in the future the Frontier infrastructure because they abandon Oracle)
Proxy autoconfig in progress: it is to avoid “wrong” choice from clients (a recent problem was seen on CI jobs in Meyrin, choosing Wigner SQUIDs) . Backup proxy is in production.

3) Oracle

New version 18c will be deployed in Q3 2019.
Goldengate support under discussion: evaluate the impact in view of renewal of Oracle licensing for 2023

⇒ **need to check implication at CC vs maintenance of AMI**

4) Eventindex

Working generally smoothly with lots of data now. Developments on going to redesign part of the infrastructure in view of Run3.
A dedicated workshop at LAL is planned for beginning of June.

Conclusion

- Successful ATLAS usage of computing ressources
  ⇒ good contribution of our Tier 1 and Tier 2 sites

- Sites configuration and recommendation
  - discussed in particular during the Jamboree in March
  - important on going changes : CL7, singularity, DOME ...

- Long term studies within LCG
  - future of LCG sites
  - DOMA ….

- Computing ATLAS France
  - forum for physics groups to get links/infos/… to S&C
  - many ATLAS France physicists participate also to LCG-FR meetings
  - since a year reinforce links with physics groups
    → annual meetings to get their inputs and
    to review « new » needs (GPU access, machine learning,
    preparation of Run-3 performance studies