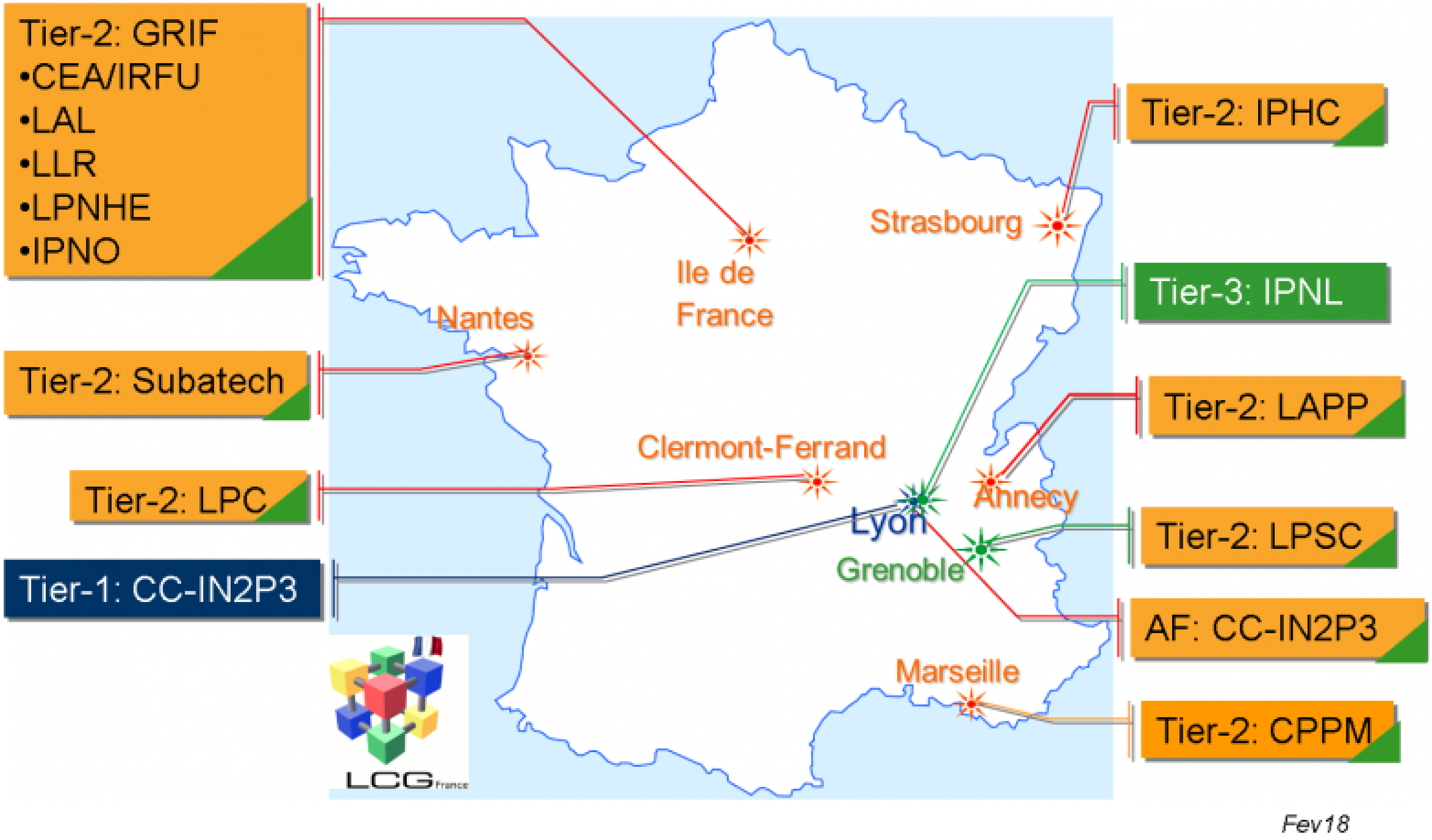
Les activités pour le LHC et les développements pour le HL-LHC

L. Duflot

Introduction

- Two main actors: LCG-France and DOMA-FR
- LCG-France
  - IN2P3 project
  - French component of Worldwide LHC Computing Grid
    - Represent France in WLCG bodies and meetings
    - Our sites are part of the WLCG infrastructure and must implement WLCG services and technical solutions as a production facility
    - We participate to WLCG (or HEPiX) working groups and task forces
- DOMA-FR
  - IN2P3 project
  - Data Organisation Management and Access is a working group setup by WLCG and HSF (open to other organisation or disciplines) for HL-LHC computing R&D
  - DOMA-FR coordinates French DOMA contributions
- HEP Software Fundation, HEPiX, France-Grille, ….
- DPM collaboration: France is part of this collaboration focused on a grid storage solution

LCG-France



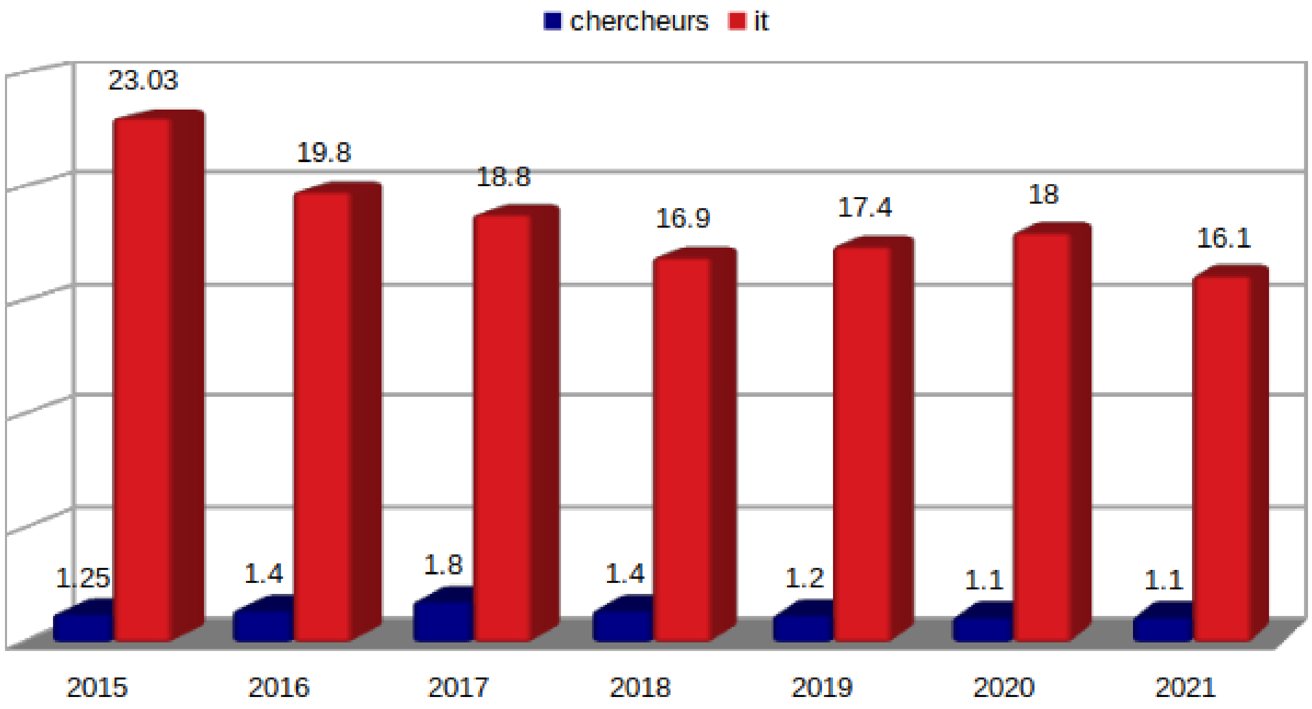
Tier 1: WLCG MoU ~98-99% availability/reliability requirements

Tier 2: WLCG MoU ~95% availability/reliability requirements

Tier 3: not part of MoU but in practice good availability/reliability

LCG-France : HR

- As a production infrastructure, the service relies on the site and service administrators
- Relatively stable overall but some Tier 2 sites are at the limit of what we consider sustainable, i.e. ~ 1-1.5 FTE with two or more persons.
- Several persons will retire in the coming years



LCG-France : protocole d'accord 2018-2022

- Tier 1 CC-IN2P3: funded by IN2P3/CNRS et IRFU/CEA
- Protocole d'accord 2018-2022 for IN2P3 sites funding

Texte du PA

Objet:

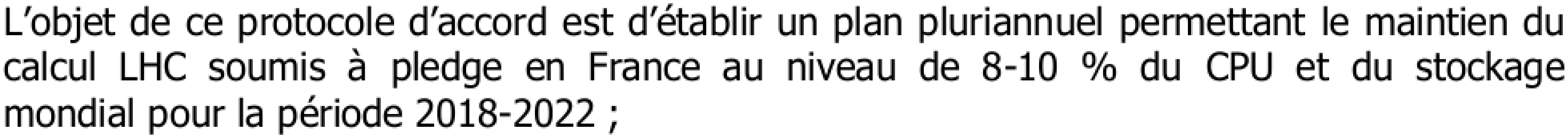
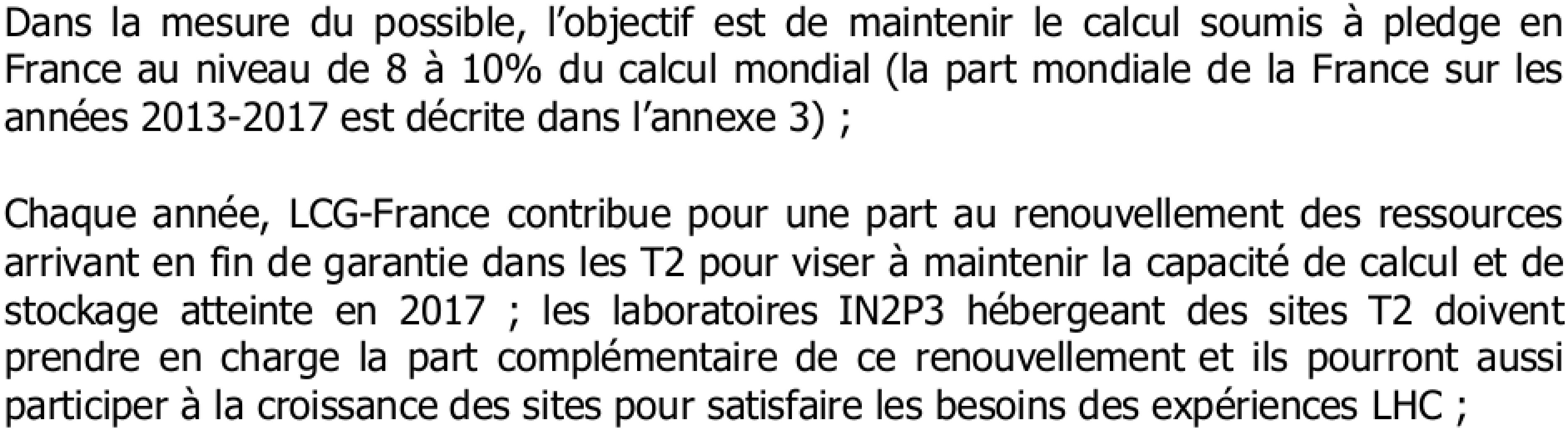

L'objet de ce protocole d'accord est d'établir un plan pluriannuel permettant le maintien du calcul LHC soumis à pledge en France au niveau de 8-10 % du CPU et du stockage mondial pour la période 2018-2022 ;

Schéma financier


Dans la mesure du possible, l'objectif est de maintenir le calcul soumis à pledge en France au niveau de 8 à 10% du calcul mondial (la part mondiale de la France sur les années 2013-2017 est décrite dans l'annexe 3) ;

Chaque année, LCG-France contribue pour une part au renouvellement des ressources arrivant en fin de garantie dans les T2 pour viser à maintenir la capacité de calcul et de stockage atteinte en 2017 ; les laboratoires IN2P3 hébergeant des sites T2 doivent prendre en charge la part complémentaire de ce renouvellement et ils pourront aussi participer à la croissance des sites pour satisfaire les besoins des expériences LHC ;

LCG-France : protocole d'accord 2018-2022

Les priorités accordées aux objectifs LCG-France par site seront respectées :

Priorité 1 : maintien des ressources du Tier1 et de l'AF ;

Priorité 2 : croissance minimale des ressources du T1 de 10% ;

Priorité 3 : contribution au maintien des T2 hors CC-IN2P3 ;

Priorité 4 : croissance additionnelle du T1 et de l'AF.

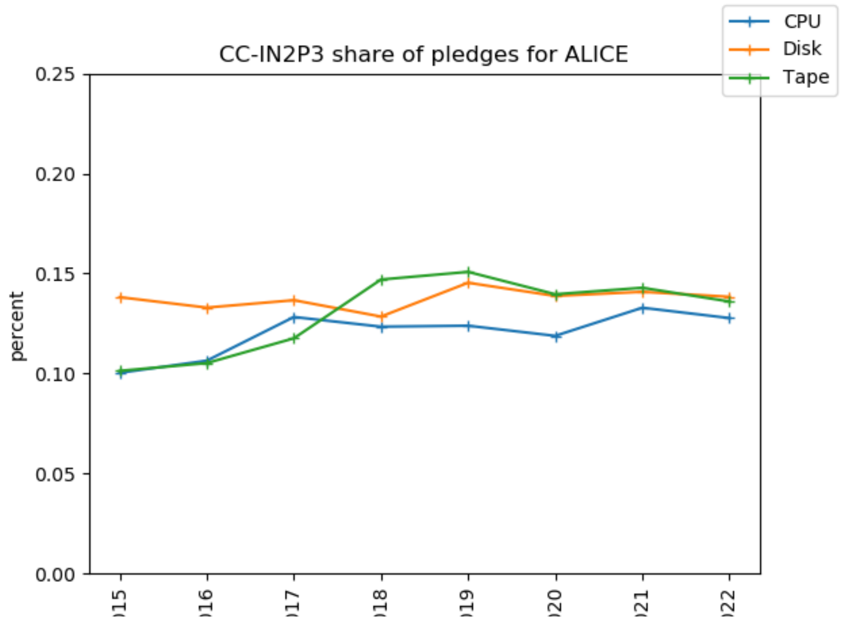
Les ressources au CC-IN2P3 doivent autant que possible être distribuées entre les quatre expériences proportionnellement aux contributions françaises aux expériences LHC, ce qui donne une clé de répartition de 45% pour ATLAS, 25 % pour CMS, 15 % pour ALICE et 15% pour LHCb, réf. [2] ;

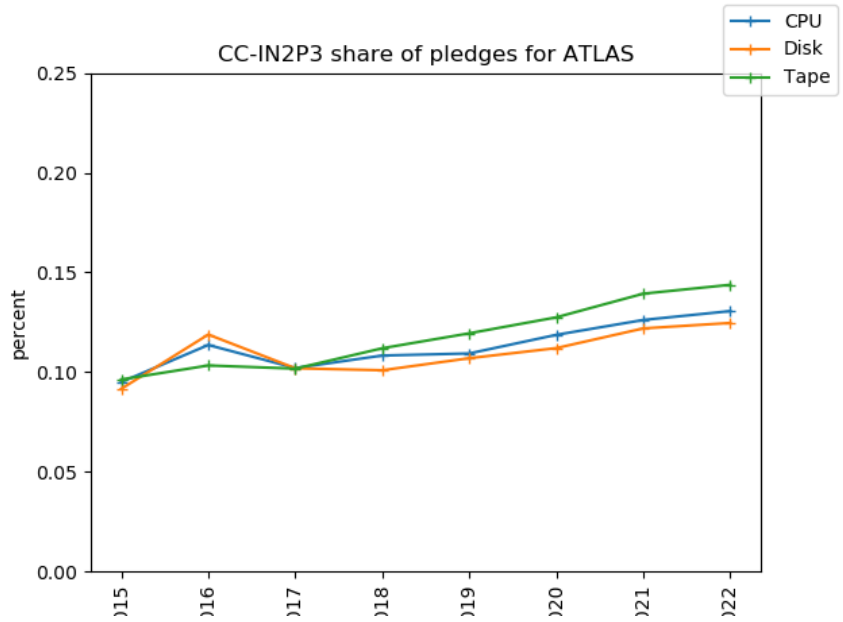
Dans ces conditions, 70% du coût estimé pour le renouvellement matériel dans les T2 hors CC-IN2P3 seront à la charge de LCG-France et 30% à la charge des sites (laboratoires).

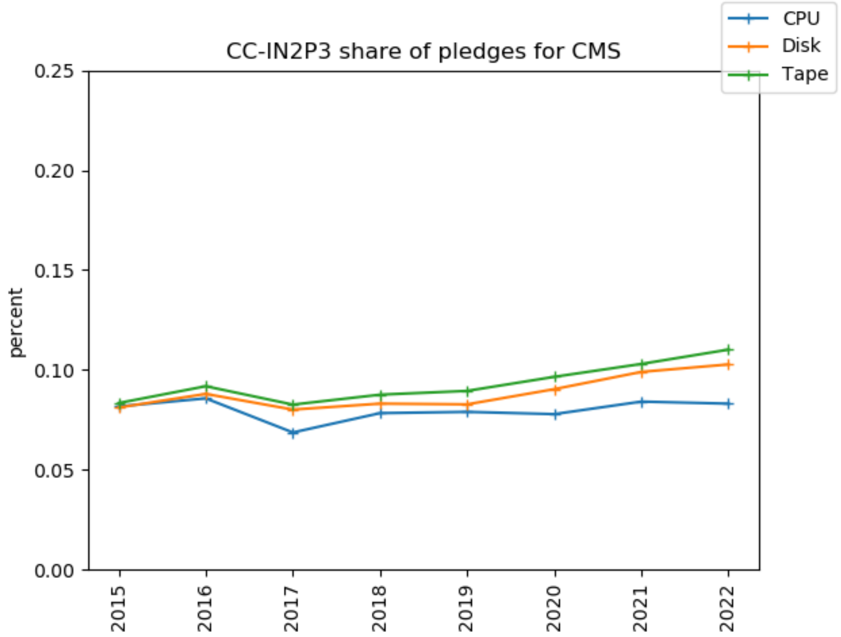
LCG-France

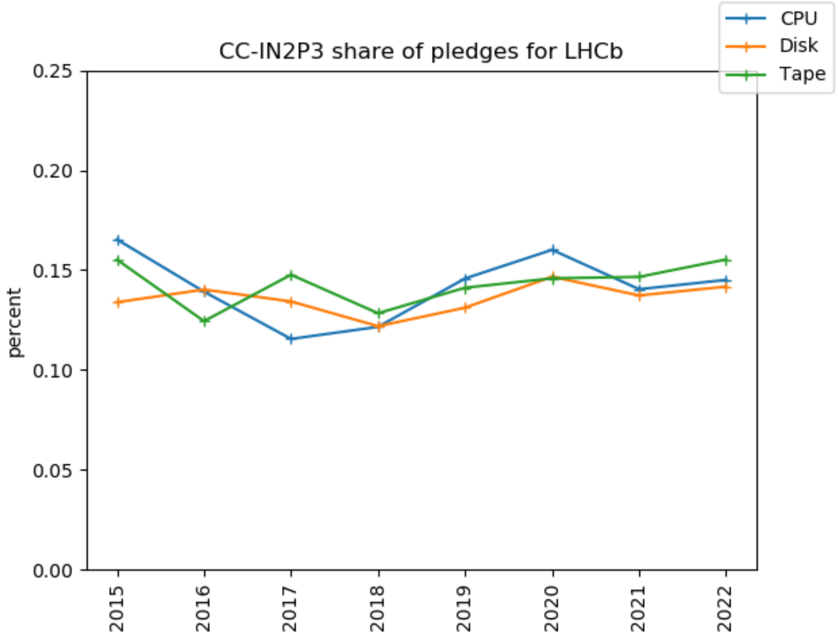
- Protocole d’accord 2018 – 2022 – summary
  - A Tier 2 site can have a one time funding (e.g. CPER) and later renew hardware at 30% of the cost – to encourage seeking funding
  - The budget split by experiment at Tier 1 is ~ fixed, Tier 2s define their policies
- Main results from this period:
  - No real funding problems
  - External funding worked well for some of the Tier 2s: total of 250-300k€ / year. Fluids sometime paid by hosting entity: ~ 150k€ / year.
  - Sites work well, better than average and well within the MoU requirements
  - We have maintained our global share of resources and complied with the goals in the “protocole d’accord”
  - Good participation in working groups / task forces : Middleware Readiness, Network and Transfer Metrics, CREAM CE migration, French perfsonar task force….

Global share of Tier 1s

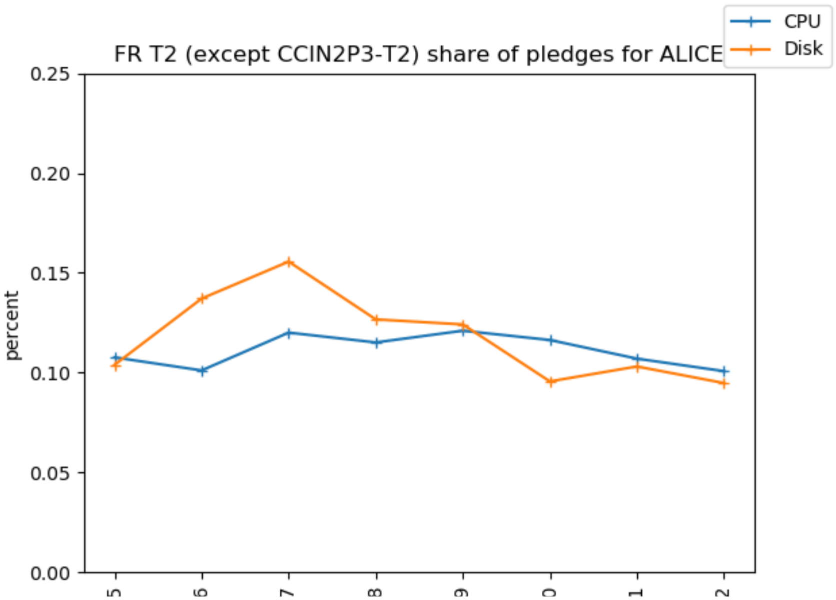

CC-IN2P3 share of pledges for ALICE

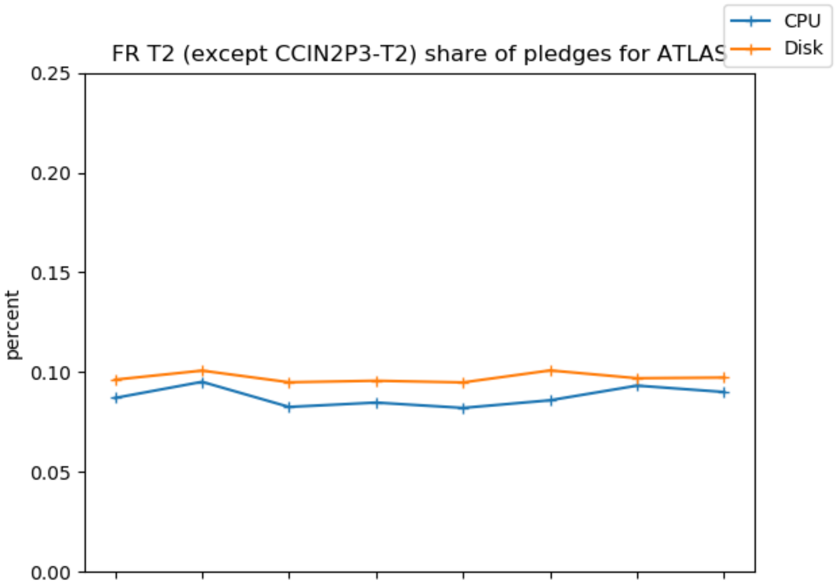

CC-IN2P3 share of pledges for ATLAS

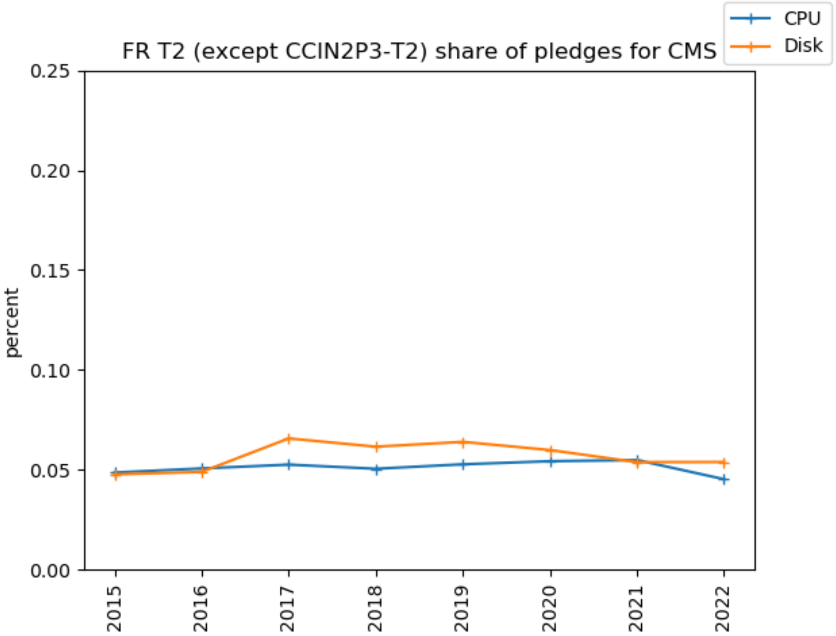

CC-IN2P3 share of pledges for CMS

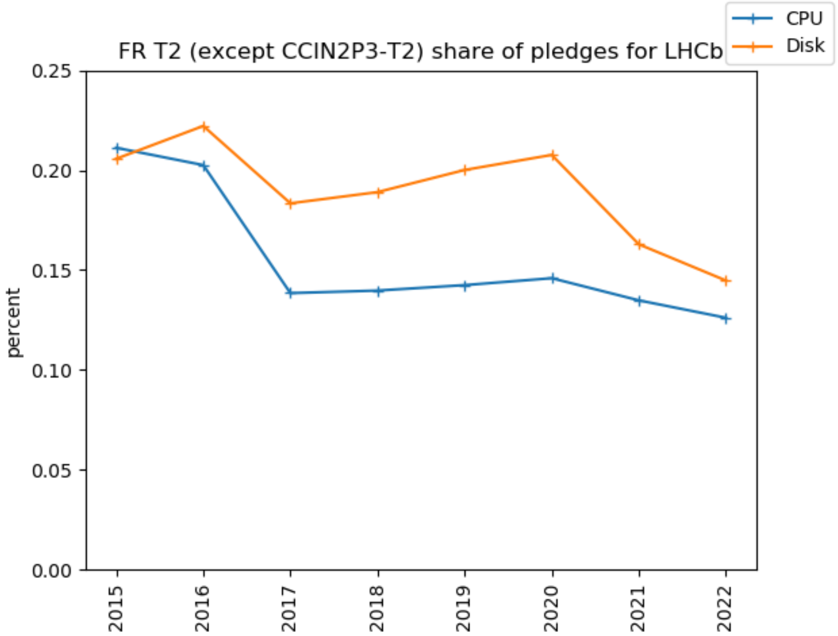

CC-IN2P3 share of pledges for LHCb

Global share of Tier 2s









LCG-France : immediate future

- We have to define / negociate a new “protocole d’accord”
  - Experiment shares at Tier 1?
  - Lab support for the Tier 2s ? (funding, personpower)
- Expected experiment requests for Run 3 are not compatible with flat budget
  - In particular LHCb
  - Arbitration?
- Two French sites have decided to stop:
  - LPSC: retirements / leaves in IT division, end of external funding: ALICE + ATLAS
  - SUBATECH: retirements, change in ALICE team

Networking

- Our network is mostly provided by RENATER, with the “last mile” sometime provided by local network
  - except for LAPP that is connected to CC-IN2P3 via regional network
- Two dedicated networks:
  - LHCOPN: connects Tier 0 (CERN) and Tier 1s, provides dedicated links e.g. for data export → 100Gb/s for CC-IN2P3
  - LHCOne: more general LHC (and beyond) private network with Tier 0, Tier 1s and Tier 2s → 2x100Gb/s for France
- Traffic is expected to grow significantly in the next few years and very much for HL-LHC. We have shared a calendar of required upgrades with RENATER.
- RENATER has severe financial difficulties, not clear if the entire required upgrade program can be achieved in a timely manner.

R&D towards HL-LHC

Working group participation

- System performance and Cost Modeling
  - Evaluate cost of HEP workflows and TCO of the computing infrastructure for a given Computing Model. Two of the main contributors from France.
- DOMA – Access
  - Data distribution and access: Data Lake, caches
  - ALPAMED test bed (CPPM, LAPP, LPC, LPSC)
  - Convenership
- Archiving – Data Carousel – DOMA Tape Challenges
  - Intensive use of tape for production campains, optimisation of tape usage
  - Optimisation of tape system at CC-IN2P3
- Benchmarking and Accounting
  - Replace "SPEC" based performance evaluation of compute servers by community workflow based score
  - Two contributors + sites to run benchmarks
-
- DOMA – Third Party Copy
- DOMA Data Challenge

France-Grille

- FG is the National Grid Infrastructure within EGI
  - Most FG sites are WLCG sites
- WLCG has been using EGI services and solutions since the beginning
- Cooperation:
  - Common workshop, presentations at LCG-France workshops
  - Participation and presentations at Journées SUCCES and JCAD co-organised by FG
  - Merged FG operation meeting with LCG-France technical meetings

Link with HPC in France

- WLCG uses many HPC machines in the US and EU, either via dedicated allocation or opportunistically.
- We did not manage to establish this connection in France.
  - There was an ATLAS prototype at IDRIS but usage restrictions did not allow for complete automation. We could not work around the policy of resource allocations that does not have the concept of year-long allocation
- FITS project: see Éric’s talk
- We presented the HEP workflows in the hexascale HPC project working groups on “applications”, in particular the specific needs of remote data access at large bandwidth

“Feuille de route du numérique”

- MESRI and CNRS have developed their computing roadmap in the last several year.
- One particular worry is that they push for Regional Data Centers that would host the computing resources of Education and Research for an entire region and only projects hosted there would in the end be eligible for state funding (e.g. CPER)
  - It would be very impractical to “nurse” a Tier 2 that would be located dozen of km away !
  - There are labeled DC in Marseille and Strasbourg, not located on the same campuses as the lab but not too far away.
- CC-IN2P3 is labeled as a National Center

Perspectives and risk assessment

- Need to sign a new “protocole d’accord” but uncertainties:
  - Fixed exp. Share makes difficult to follow actual needs
  - Cost of computing hardware (currently stable or rising, while Computing Model planned on decrease), delivery delays (short term)
  - Cost of electricity
  - Funding of network upgrades
  - Level of funding for Tier 2s, mainly rely on external fundings for growth (CPER, FEDER, …)
- Several retirements of site administrators in the next few years, need support from the lab and IN2P3/CNRS to maintain the level of efforts
- R&D for HL-LHC did not make any real breakthrough in term of cost reduction
- The Computing Model relies on tape as a cheap(er) storage but the global market has lost key players and so is no longer very competitive
- Possible divergence between WLCG and EGI (solutions, timescales)

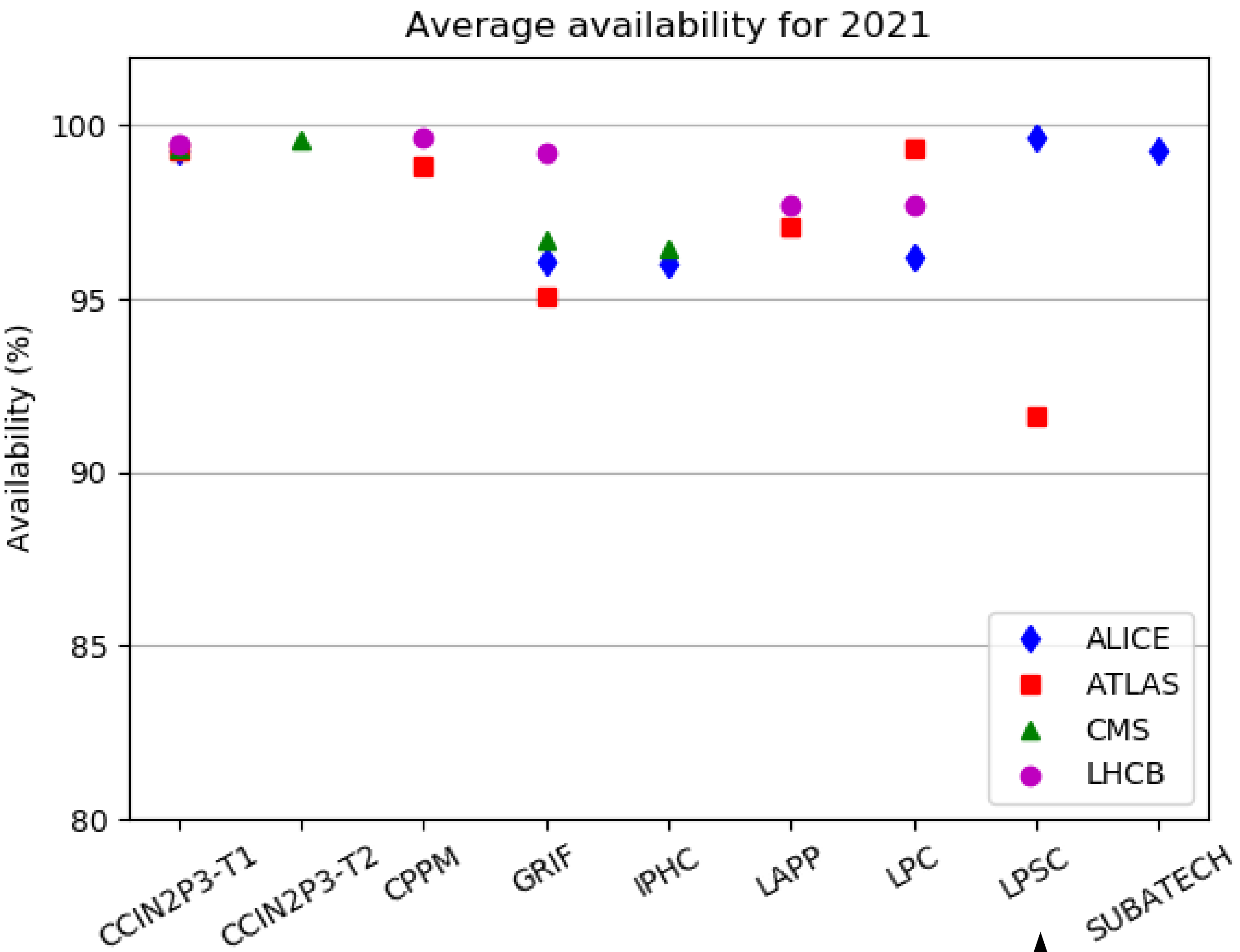
Questions ?

Backup slides

LCG-France : experiments supported

Site	ALICE	ATLAS	CMS	LHCb		Site	ALICE	ATLAS	CMS	LHCb
T1 CC-IN2P3	✔	✔	✔	✔		GRIF_IPNO	✔			
						GRIF_IRFU	✔	✔	✔	
T2 CPPM		✔		✔		GRIF_LAL		✔	(✔)	✔
T2-GRIF	✔	✔	✔	✔		GRIF_LLR		(✔)	✔	✔
T2 IPHC	✔		✔			GRIF_LPNHE		✔	(✔)	✔
T2 LAPP		✔		✔						
T2 LPC	✔	✔		✔						
T2 LPSC	✔	✔								
T2 Subatech	✔									
T3 IPNL	✔		✔							

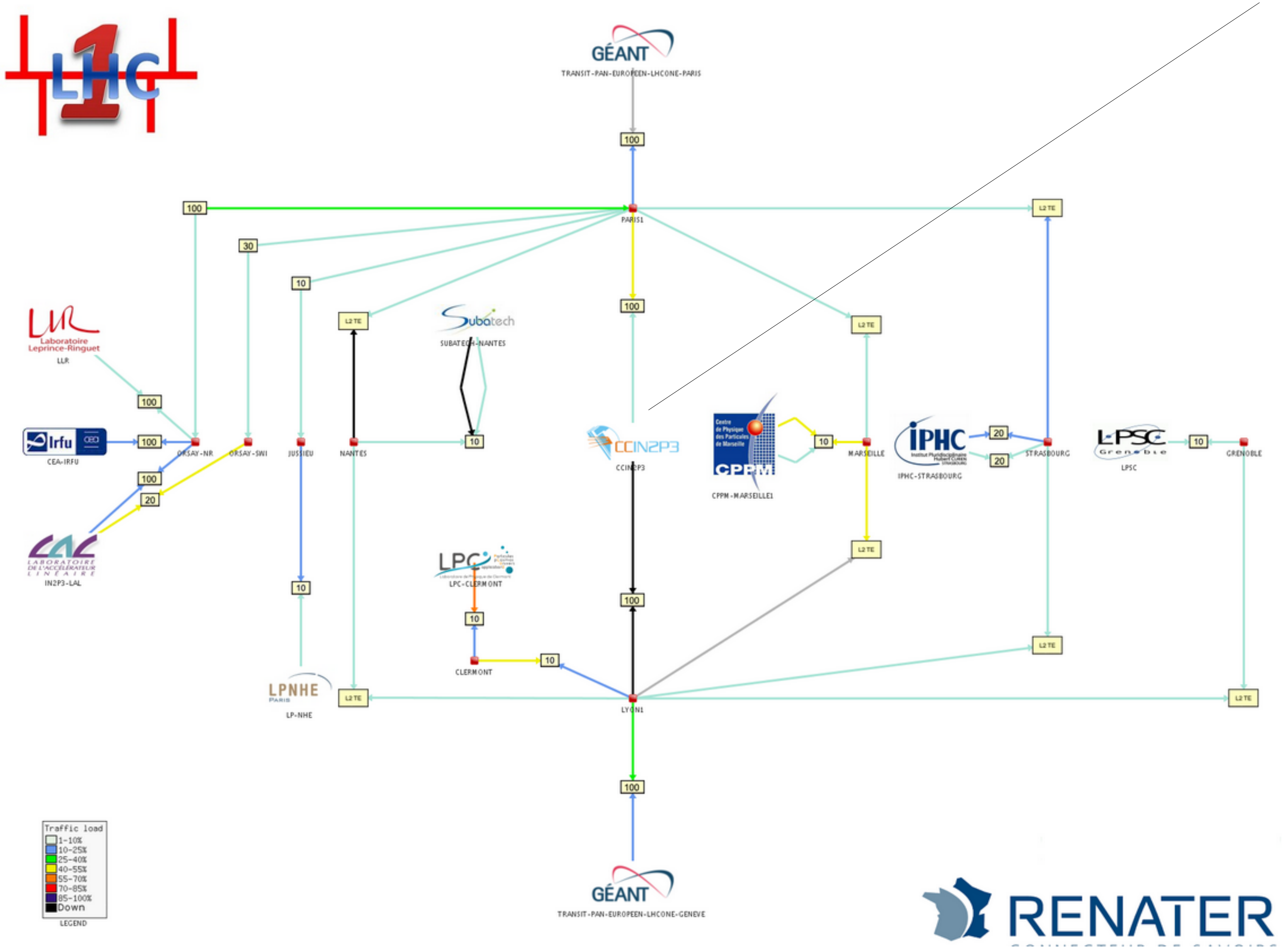
Reliability / availability : average for 2021



Being decommissioned

Network

LAPP



LCG-France : management

- Bi-weekly meetings (staggered):
  - Technical coordination: relays informations from WLCG, EGI and the experiment computing operations, organises middleware migrations, setup specific interest groups, etc…
  - Steering board: computing policies and organisation, budget, WLCG and experiment policy, long term evolutions of computing models, R&D, French and European political context, etc…
- Executive Board: very infrequent
- Journées LCG-France: 4 half-days, twice a year

System Performance and Cost Modelling Working Group

The goal of the working group was to evaluate the requirements (memory, CPU, storage) of each major HEP workflow (generation, simulation, digitisation, reconstruction) and to estimate the Total Cost of Ownership of the computing infrastructure so as to develop a model to evaluate the cost of a given computing model for HL-LHC.

A simulator was produced, initially based on the CMS simulator that was further refined and made more generic. Several workflow were measured. The actual cost of computing hardware, infrastructure and running cost (in particular electricity) was found to vary significantly from one country to the other.

Two of the main contributors of the group were French and made presentation at international conferences (WLCG workshop 2018 and CHEP 2019), and several other French people made contributions.

DOMA – ACCESS

During the first phase of DOMA project the ACCESS working group was charged at exploring solutions of data distribution and access. One of the convenors was French. The idea of a Data Lake[4], a mix of data storage sites and data processing sites (accessing data remotely, possibly through caches), emerged. The storage being identified as the most administratively heavy task compared to processing, it was thought that CPU-only sites would save a significant amount of person-power. The group studied in particular the performance of applications with local, remote and remote-with-cache data access.

The main French contribution was the implementation of a test bed processing infrastructure called ALPAMED federating storage and computing from four sites with a single entry point. That in particular implied that a given job can process data from any of the sites. ALPAMED was included in the ATLAS and ESCAPE computing. Initial tests were made with specific jobs in order in particular to measure the efficiency of specific applications when the data is accessed remotely. ALPAMED was then included as an ATLAS production site. This work was presented in an international conference (CHEP 2019).

Even if implementation, just starting, was not done during the DOMA-ACCESS group time, the evolution of the storage of the sites under the GRIF federation follows a similar philosophy: unifying the storage of four sites with a technology that will present the geographically separated storage hardware as a single entity from the point of view of the user.

Archiving – Data Carousel – Tape Challenges

CC-IN2P3 is a long standing member of the HEPiX "Archival" working group but during Run 2 of the LHC the computing models have shifted toward a more dynamic use of tapes in data processing activities. In particular, processing using a "Data Carousel" model (where data is staged from tape to disk buffers in order to be processed and quickly replaced by the next chunk of data from tape) has been commissioned and then integrated as standard workflows. This required a significant amount of work to optimise the tape systems for this kind of much more dynamic workflows, in coordination within WLCG.

In its second phase, the WLCG DOMA activities have switched from pure R&D to implementation of some of the solutions which lead to the definition of "Challenges" as demonstrators and stress tests, one of which being the Tape Challenge as a follow-up to the Data Carousel activities. So far two Challenges have been organised, in the autumn of 2021 and spring of 2022, with the specific goal of testing the readiness of the tape systems in Tier 0 and Tier 1 for Run 3. CC-IN2P3 has shown good performances, exceeding the requirements. Further Challenges will be organised in the coming years, raising progressively the bar to build the infrastructure for HL-LHC.

Data Challenge

Similarly to the Tape Challenge, DOMA is organising bulk data transfer challenges in order to test networks and disk storage systems. The first one took place in autumn 2021 and involved CC-IN2P3 and several French Tier 2 with good performances. As for the Tape Challenge, further Challenges are foreseen, roughly every other year, with increasing goals to pave the way towards an HL-LHC ready infrastructure. This schedule has driven in particular our requests for upgrade of the French network infrastructure for WLCG mostly operated by RENATER.

Benchmarking and accounting

Properly measuring the performances of the hardware in terms of processing power and translating it into an agreed-upon unit is necessary both for accounting of each site contributions but also for them to pledge a given level of resources and for experiments to express their need. Several years ago, a set of benchmarks based on the SPEC generic toolkit were defined and turned into a global performance measurement named HS06.

Along the years, some HEP workflows have shown different scaling than the SPEC benchmark. The Benchmarking Working Group has been charged to define and implement a different solution that would reflect better the actual experiment workflows and that could be extended to measure the performances for non CPU-only hardware, e.g. CPU+GPU. The working group has developed a toolkit based on containers encapsulating actual workflows from the LHC experiments (generation, simulation, digitisation and reconstruction). The working group is now in the process of defining the set of tools that would be combined in a new official overall performance index named HEPSCORE. A French person is a member of the steering group and other French people have participated in the implementation and test of the toolkit.

DOMA – Third Party Copy

Under this relatively cryptic name, the working group covered the activities related to the protocol (providing the new Third Party Copy feature) and authentication - authorisation for data transfer. One of the first task was to deal with the obsolescence of a particular toolkit for data transfer and its replacement by a more standard-based tool (gridftp to http-based transfer). The second task was the replacement of the authentication – authorisation system. This work is still ongoing in the second phase of DOMA and its Bulk Data Transfer group. People from CC-IN2P3 actively participated to DOMA-TPC.

DPM Collaboration

DPM is the storage solution used by most French Tier 2. It's a product developed mostly by CERN people around which a collaboration was formed. The French contribution consisted into providing test beds for testing new versions of the product and migrations. Our two representatives to the Collaboration Board have coordinated the writing of DPM Community Whitepaper of 2019-2020. The development for this product has stopped in 2020.

Other working groups

We have participated in other working groups, some of them formed for a specific task like transition from one solution (or version of a production) to another:

- Middlware Readiness WG
- Network and Transfer Metrics WG
- CREAM CE migration task force

and created a Perfsonar LCG-France Task Force (following a significant effort in investigating network monitoring).

France-Grilles

First and foremost, our natural partner project in France is the GIS France-Grilles, acting as the French National Grid Infrastructure against EGI. WLCG has been using EGI solutions and services since its inception (although this is less and less the case now) and there is a large overlap between France-Grilles sites and LCG-France sites.

We have organised one common workshop and had France-Grilles presentations during LCG-France workshops. We have encouraged participation to the Journées SUCCES (co-organised by France-Grilles, Groupe Calcul and “coordination mésocentres”) that became JCAD (Journées Calcul et Données, co-organised by the same partners with the addition of GENCI), where we have presented LCG-France or DOMA topics four times.

The participation to Journées SUCCES and JCAD was also an attempt to reach out to other communities in the French computing landscape, in particular the mésocentres[5] (some of our sites are mésocentres) and the computer scientists. At one of DOMA workshop we had people from CNES, INSERM and Nantes. We had contacts with INRIA scientists but this did not turn into a concrete collaboration.

As mentioned earlier, the France-Grilles operations meeting and LCG-France technical meeting have merged this year since there was a large overlap between the two (as LCG-France technical meetings include a report on WLCG operations).

High-Performance Computing

As the LHC computing is constantly starving for resources, there have been efforts to enable access to HPC centres from all over the world for several years. One of the difficulties is that the access restrictions and technical situations vary a lot from one HPC site to another (operation of border services, internal and external network access and bandwidth, allocation policies, etc…).

A first attempt to use the IDRIS machines was made with people from CC-IN2P3 and ATLAS. There were technical difficulties but the demonstration of feasibility was made, although the service would have required permamnent human interventions. Another hurdle was to accommodate the policy of resource requests which were designed for larger computations over a short time period while we would like a constant minimum allocation during the entire year to make the human investment worthwhile.

CC-IN2P3 leads for FITS project, which involve IDRIS and GENCI (the actual proprietary of IDRIS supercomputer), aiming to favor usage of both facilities by relevant users and the possibility for long duration requests. We hope that this will turn into an opportunity for LCG-France to give WLCG access to those resources.

In the last couple of years, France has express the intent to host one of the PRACE hexascale HPC computers. In particular a working group has been put together in order to reach out to all the communities that could use such a facility and assess their readiness to use a heavily GPU based architecture. IN2P3 is one of the communities represented there and LHC computing one of the topics. The requirements of LHC computing is quite different from the usual requirements of an HPC machine, in particular in terms of wide area network access at high speed. One specific project detailed in the working group was the usage of such a facility for LHCb trigger processing: it would require Tb/s networks. The working groups were also the opportunity to reach out to other communities in need of high data volume processing and in particular remote processing. The working group is now closed and the report being written but we hope that the discussions, in particular around distributed data, will continue as promised in a different forum.

Feuille de route du numérique

In the last few years, the French ministry of Higher Education Research and Innovation and CNRS have developed roadmaps for their computing activities, from hosting services or hardware to scientific computing. One item of potential direct consequence for LCG-France was the goal of drastically reducing the number of Data Centres (DC). The arguments were that small local DC were proliferating, not very energy efficient and in constant demand of human resources for their operation. The main idea was that a single DC per French region would be labelled (with possible exceptions e.g. for larger or more populated regions like Paris – Île de France) and that national agencies would then only fund computing projects with hardware hosted in a labelled DC. That would directly affects most of our external sources of funding, in particular CPER.

This caused great concern in our community since, except for CC-IN2P3 (already considered as a National Centre), essentially none of our DC would likely be labelled. We wrote a short document summarising the problems that a move to a different DC would cause. For example we argued that Tier 2 sites have signed an MoU with requirements on the availability of the site which implies that quick and easy access to the DC is needed. We estimated that Tier 2 admins would need physical access to the hardware once or twice a month at least for repair and installation of hardware. While this could be doable if the DC is located in the same town, it would be problematic if it is located dozen of kilometres away (or more, given the size of regions in France).

The process of site selection has since been very slow and the “one DC per region” not strictly followed. There are labelled DC in Marseille and Strasbourg, so not too far away from our DC. In Marseille, the DC is even one that was involved in common projects with the CPPM site. The IPHC site anticipates having to host at least part of their new hardware to the labelled DC. The situation is less clear for the other LCG-France Tier 2 sites.

International Context

As a production infrastructure within WLCG, LCG-France is constrained by the overall model and strategies and part of the technical solutions decided at that level although it is very much part of the decision process. The overall weight of CERN in the decision process is very large although in practice the management seek for a wide consensus. So far no controversial decision have been taken[6] with the possible exception of the withdrawal of CERN from development of the DPM storage solution, effectively bringing it to an end.

Other HEP and non HEP experiments will be faced with large volume of data, and it seems important that the LHC experiments and WLCG share their experience and, as much as possible, tools. In many countries, these experiments will be hosted in the same data centres (e.g. Tier 1) and having to support different tools for different experiments would be very costly human wise. One example is that the LHCOne network, initially designed for the LHC experiments, has been opened to more and more experiments that use the same data centres (Belle II, Pierre Auger Observatory, NOvA, XENON, JUNO), perhaps a bit too enthusiastically as it becomes difficult to identify the source of traffic in case of congestion. WLCG and CERN have also started high level discussions with SKA. Likewise, some EU projects like XDC and ESCAPE where French labs from WLCG were/are involved use tools developed in WLCG. The ALPAMED test bed has been used in WLCG for ATLAS and also in ESCAPE. We anticipate that such tools would also be used in other projects related to EOSC.

Perspectives and Risks

We are now in the process of defining a new agreement ("Protocole d'Accord") between the IN2P3 direction, sites and experiments project leaders for the next 4-5 years. One of the main discussion points is how the budget is shared between experiments at CC-IN2P3: the current fractions are based in the IN2P3 participation in LHC experiments but the needs of the experiments have changed for Run 3 (with a huge increased of request by LHCb due to their Phase I upgrade) and will change again approaching HL-LHC (where ATLAS and CMS needs will dominate).

As already explained two of our Tier 2 already announced that they would stop by the end of Run 3 so they already do not receive funding for hardware renewal and won't be part of the new agreement. For the other Tier 2 sites, continued participation make sense if they intend to run during the HL-LHC era. Most laboratory directors indicated that they rely on external funding for their Tier 2 sites and that funding is not guaranteed beyond the next few years, which makes them hesitant to commit.

The current funding model rely on the budget coming from CNRS IR for support to hardware renewal at Tier 2 and renewal and growth at CC-IN2P3. The target budget has been raised by 10% a few years ago but an additional increase would probably be necessary to follow the huge increase in requested resources for HL-LHC.

In addition to budget concerns, there are other sources of uncertainties for the next few years and even more at medium term like the start of HL-LHC. First, the DOMA R&D did not uncover any technical solution that would significantly reduce the cost of disk storage. It seems the baseline now for ATLAS and CMS is to achieve reduction of the analysis format tier to ~ 10kB per event in order to be able to store it permanently on disk. Second, the trend of constant reduction of cost of a unit of disk or CPU is no longer guaranteed or at least the reduction is less that what was common a few years ago. On a short term basis, it is expected that the current "silicon crisis" will continue to cause delays in delivery and shortage of supplies.

The ATLAS and CMS computing models rely heavily on the use of tapes as a cheaper storage than disk. They are also attractive in the current context of steeply rising cost of electricity. However, prices are driven by the global market and the market for tape usage in enterprise computing is shrinking (or at least not strongly growing). As a consequence key vendors have left the tape market so we are at risk of relying on single vendors with less investment in R&D and no incentive for price drops.

Another uncertainty we have already mentioned is due to the financial difficulties of RENATER, the French Research and Education Network provider. It is currently unclear how the cost of upgrading the capacities to the level of our anticipated needs will be funded.

As already mentioned, the success of our projects is built on the efforts and dedication of our people. We indicated that some sites have reached what we consider the minimum level of effort to maintain a healthy Tier 2 site. In addition we know of several retirements of key site administrators in the next few year. We must carefully plan for their replacement and this needs strong support for the laboratory directions and from the IN2P3 direction.