Centre de Calcul de l’Institut National de Physique Nucléaire et de Physique des Particules

Retour Workshop ALICE @ Bucarest

LCG FR @ LAPP Annecy
2019-05-23
Jean-Michel-Barbet, Renaud Vernet

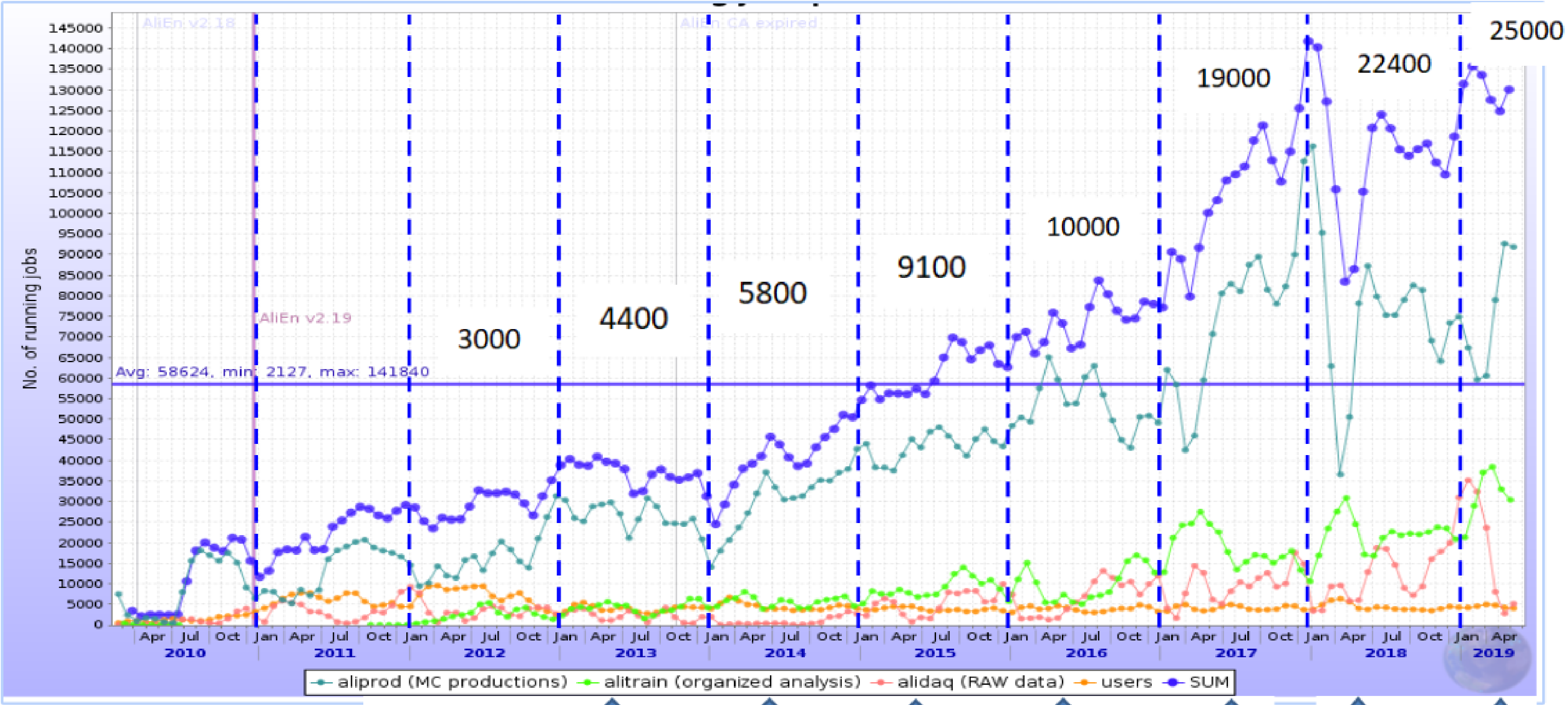
Worshop ALICE T1/T2

- Workshop annuel entre ALICE central et sites

- Accueilli cette année par Universite Politechnique de Bucarest (UPB)

- 2,5 jours
  - 50 % site reports
  - 50 % talks dedies
    - Bilan
    - Operations
    - Middleware
    - Stockage (EOS)
    - Securite

Evolution of tasks - share of organized analysis



3000

4400

5800

9100

10000

19000

22400

25000

Year on year increase

+47%

+32%

+57%

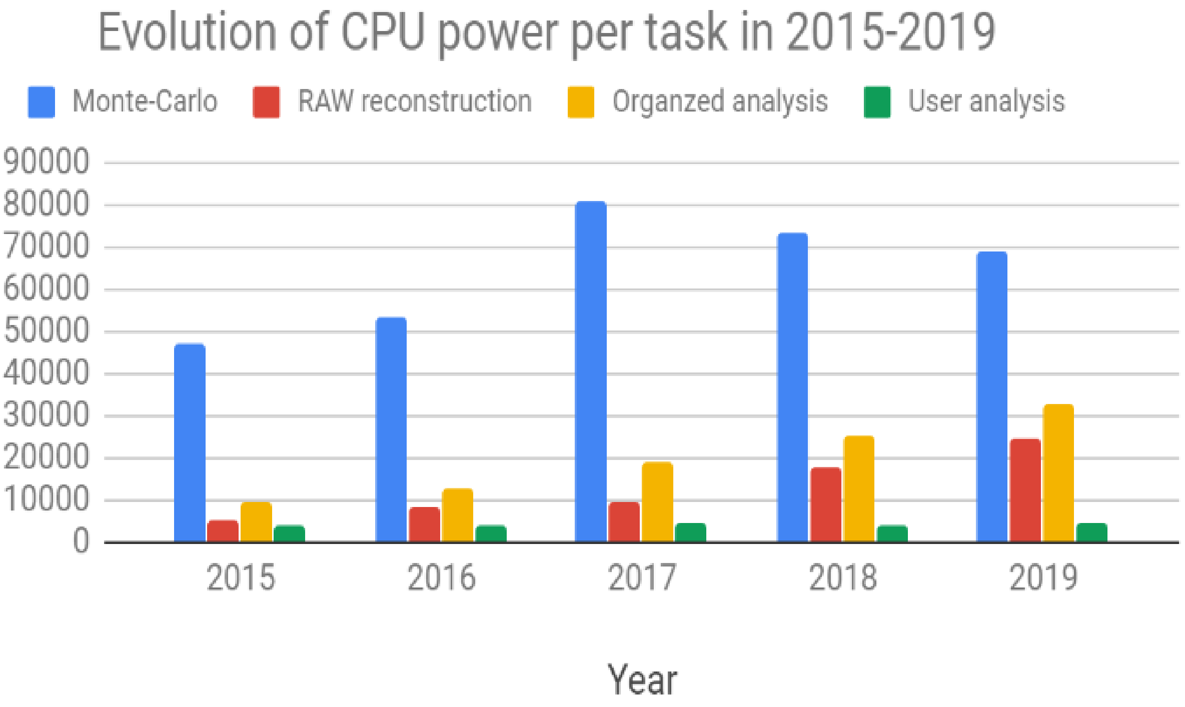
+10%

+90%

+20%

+45%

Type de tâches

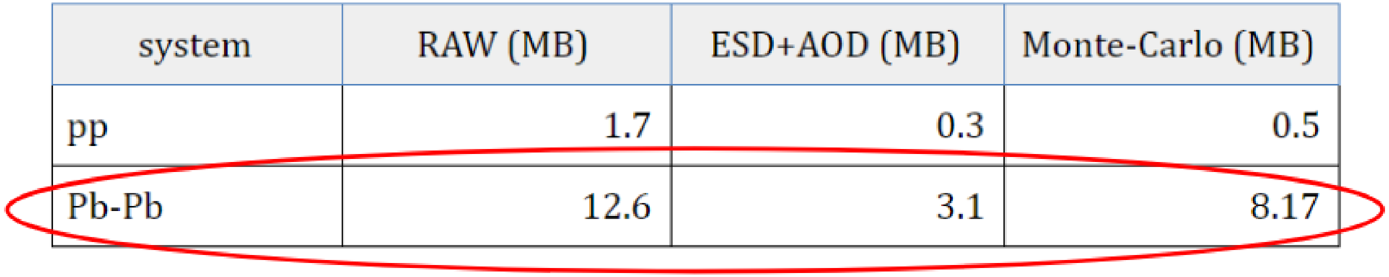


Moins de MC
De plus en plus de ressources dediees à l'analyse

Delivered / pledged = 124 %

Resources adjustment for Pb-Pb

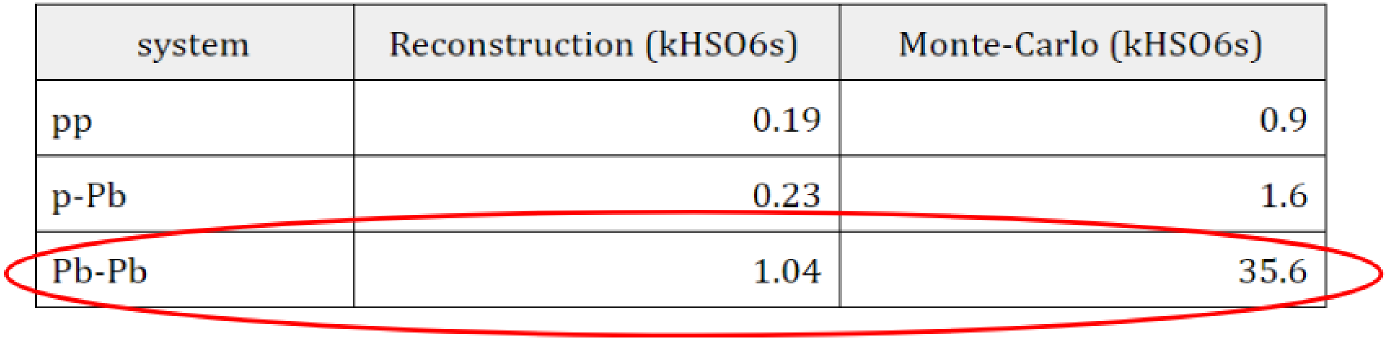
- Average event sizes



system	RAW (MB)	ESD+AOD (MB)	Monte-Carlo (MB)
pp	1.7	0.3	0.5
Pb-Pb	12.6	3.1	8.17

25% lower

- Computing power per event

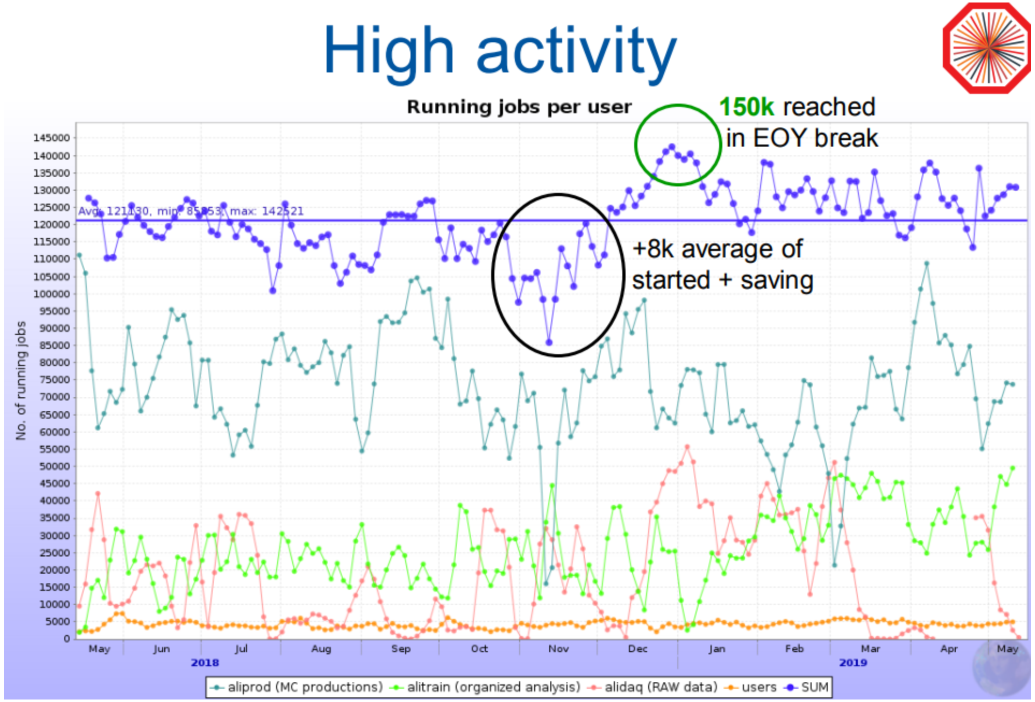


system	Reconstruction (kHS06s)	Monte-Carlo (kHS06s)
pp	0.19	0.9
p-Pb	0.23	1.6
Pb-Pb	1.04	35.6

25% lower

Operations centrales

- CA reorganization in AliEn
  - Some new CA’s incompatible with old SSL of AliEn
- Improvements
  - Trafic stability
  - Task queue host performance
- Transfers OPN non critiques cette annee
  - Buffer EOS DAQ suffisant
  - Export T1 apres prise de données
- EOS space name → QuarkDB
  - *(cf Jean-Michel)*
- Lost data (1 server) : 110 TB @ IN2P3



High activity

150k reached in EOY break

+8k average of started + saving

Bilan Run 2

- Objectifs atteints

- LS2
  - Reprocessing 2017 & 2018
  - Preparation Run3

- Demandes 2020 vs 2019

CPU	+ 4 %
Disque	+14.5 %
Tape	- 4 %

O^2

- Changements importants RUN3 - 2021
  - Lecture en continu TPC
  - Triggering et compression des données avant enregistrement
  - 100 fois plus de collisions à analyser
  - Refonte du format des AOD et de la procedure d’analyse

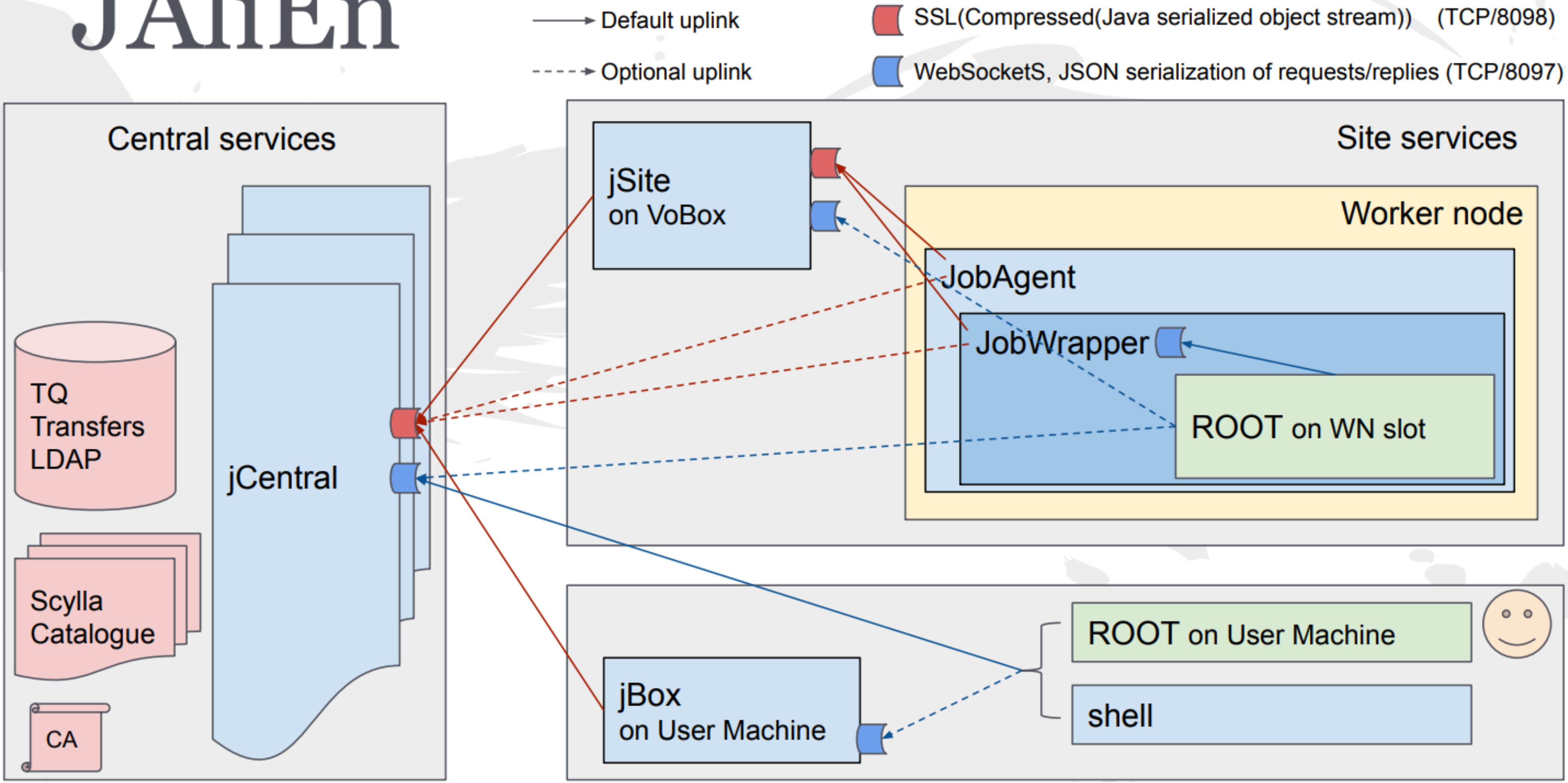
- Nouveau framework de compression O^2
  - Utilisation GPU
  - Purpose-built facility with balanced CPU/GPU components & large storage (60 PB on EOS)

https://indico.cern.ch/event/778465/contributions/3245315/attachments/1844447/3025854/ALICE_upgrde.pdf

Middleware

- Cream CE supported until Dec. 2020
  - EGI/WLCG will help provide guidelines for alternatives
    - ARC
    - HTCondor CE
    - Solutions with no CE
    - Summary GDB : https://indico.cern.ch/event/739878/contributions/3380144/attachments/1840722/3017853/CREAM_Migration_WS_report.pdf

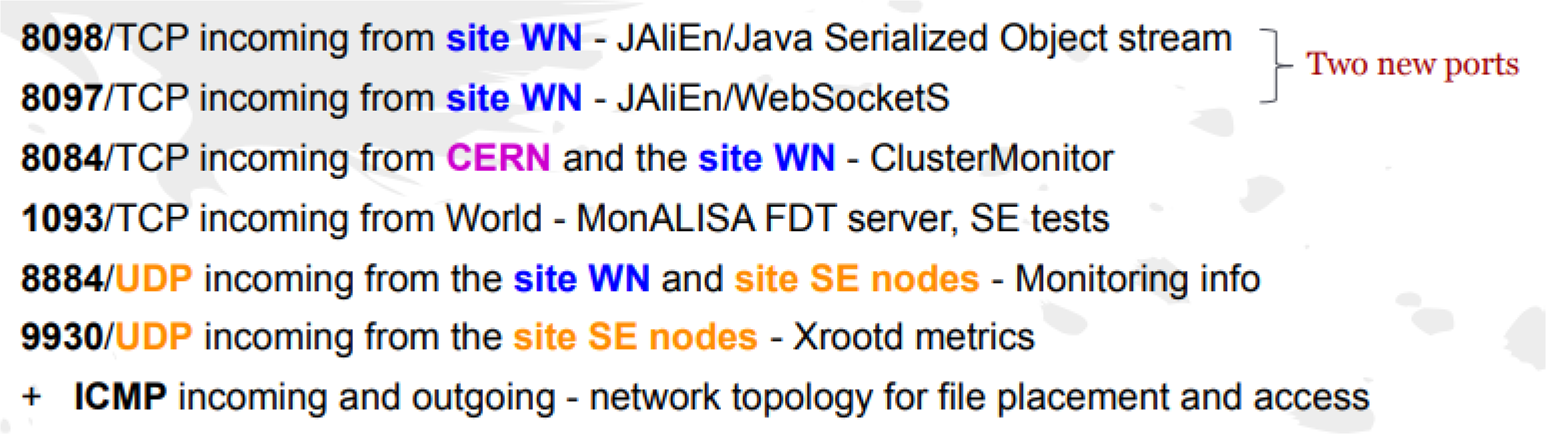
JAliEn


Default uplink
Optional uplink
SSL(Compressed(Java serialized object stream)) (TCP/8098)
WebSocketS, JSON serialization of requests/replies (TCP/8097)
Central services
TQ
Transfers
LDAP
Scylla
Catalogue
CA
jCentral
Site services
jSite
on VoBox
Worker node
JobAgent
JobWrapper
ROOT on WN slot
jBox
on User Machine
ROOT on User Machine
shell

jAliEn

- Scalability
- Improved user interface to AliEn services

- Changes required (soon)



8098/TCP incoming from **site WN** - JAliEn/Java Serialized Object stream

8097/TCP incoming from **site WN** - JAliEn/WebSocketS

Two new ports

8084/TCP incoming from **CERN** and the **site WN** - ClusterMonitor

1093/TCP incoming from World - MonALISA FDT server, SE tests

8884/**UDP** incoming from the **site WN** and **site SE nodes** - Monitoring info

9930/**UDP** incoming from the **site SE nodes** - Xrootd metrics

+ **ICMP** incoming and outgoing - network topology for file placement and access

Containers

- WLCG containers
  - WLCG intends to provide a guideline for deployment of singularity
- Linked to deployment of jAliEn
  - JobWrapper run in container for improved isolation
  - Final tests
- Sites
  - Install an RPM build (with underlay support)
  - Run through CVMFS (enable user namespace) : requires EL7 with 7.6+ kernel
- Most sites providing singularity not configured with underlay
  - → workaround
  - Bind dirs to preexisting dirs of container
- Tests with workaround successful

Security

- jAliEn security model
  - https://indico.cern.ch/event/778465/contributions/3378340/attachments/1843679/3023974/nhardi_jalien_security_model.pdf

- Deep Learning to prevent/detect intrusions
  - https://indico.cern.ch/event/778465/contributions/3239141/attachments/1843670/3023958/ALICE-tier-2019.pdf

IPv6

- Deployment on storage services
  - T0+T1 : 73 %
  - T2 : 46 %
- All storage should be in dual stack before ALICE uses IPv6-only compute resources

Recommendations

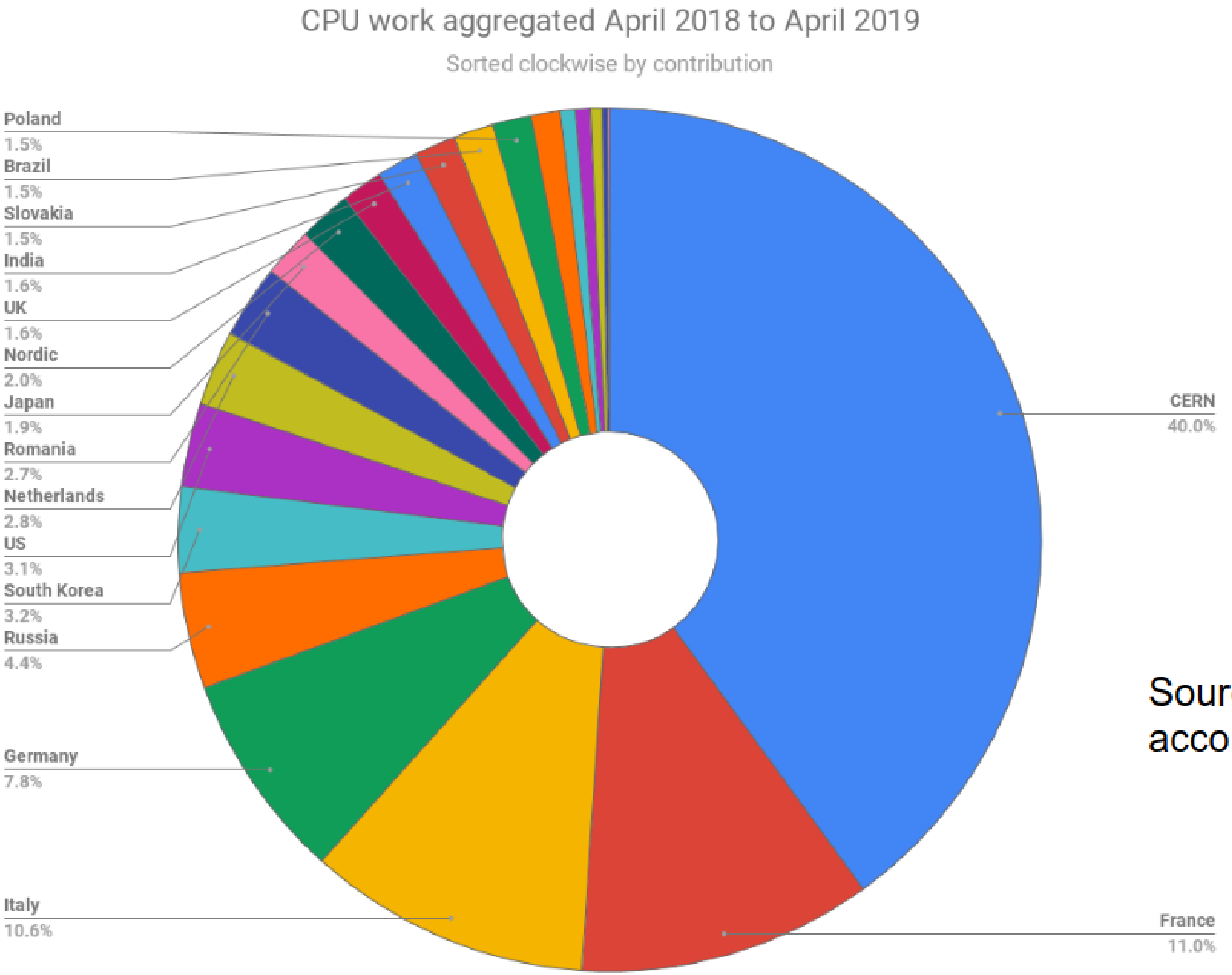
- Move to EOS
- Xrootd 4.1+

Sites

- Green IT cube (GSI)
  - PUE = 1.07
  - Fonctionne bien
  - Simple a mettre en place
  - Chaud a l'interieur

- KISTI met son « stockage froid » sur disque
  - Problemes de procurement avec fournisseur solution robotique
  - Appel d'offre pour solution disque
    - 20 PB – 1M$
    - Conso electrique ?

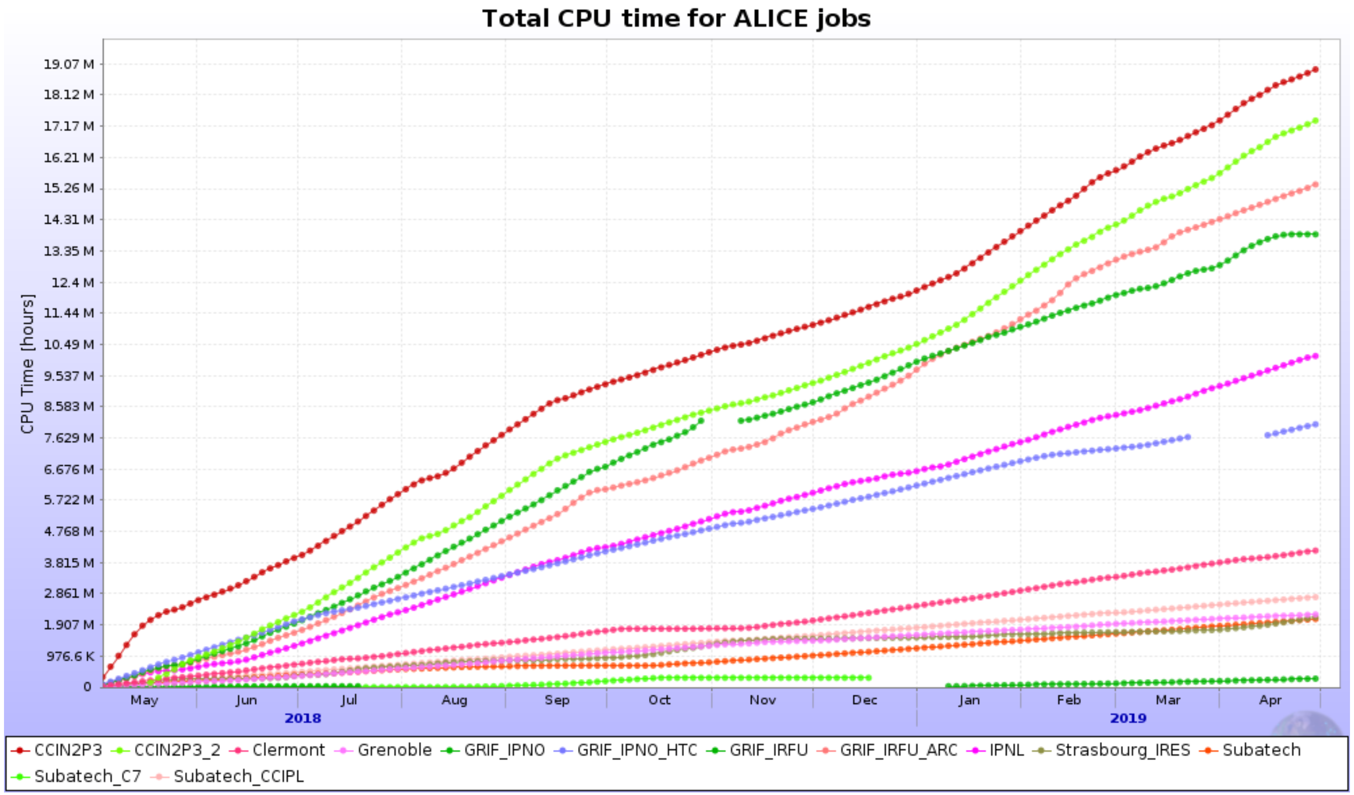
FRANCE



Source - WLCG accounting portal

French contribution to ALICE computing

- 11.7 % of total ALICE CPU time
  - was 9 % last year



CCIN2P3

IRFU

IPNL

IPNO

Clermont

Subatech
Strasbourg
Grenoble

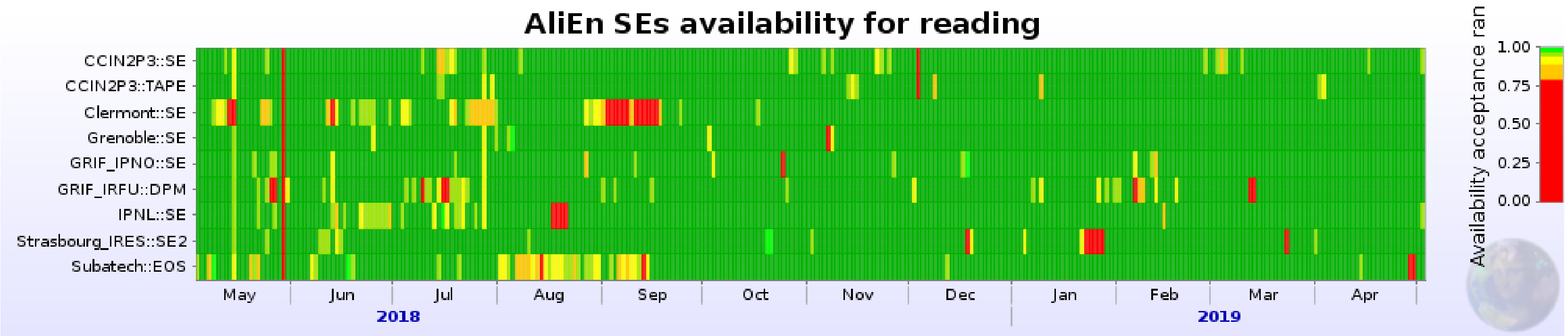
Pledges 2019

	T1			T2 (*)	
	capacity	vs T1 requ.		capacity	vs T2 requ.
CPU	41 k	11 %		45 kHS	12 %
Disk	5.1 PB	11 %		4.2 PB	12 %
Tape	6.2 PB	11 %			

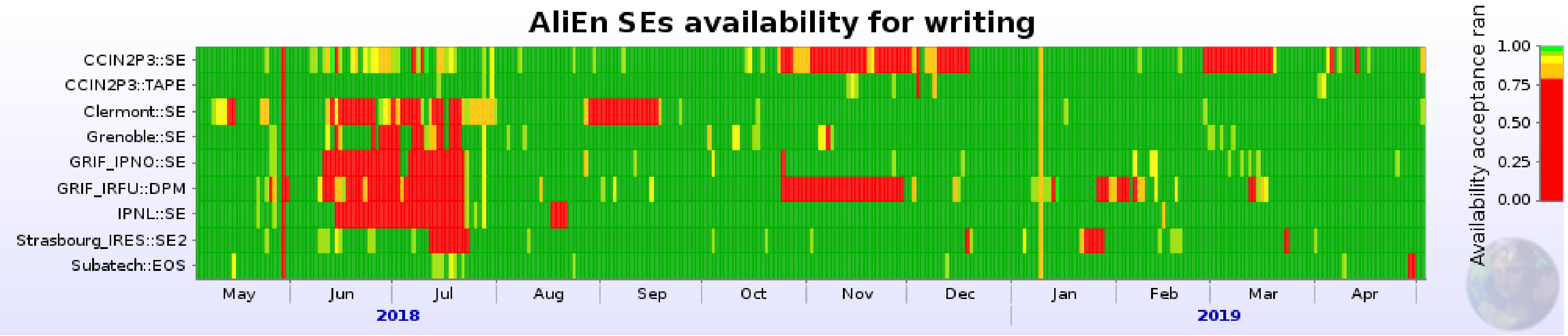
(*) IPNL T3 not accounted for

Significant budgetary support from FA maintained

Storage

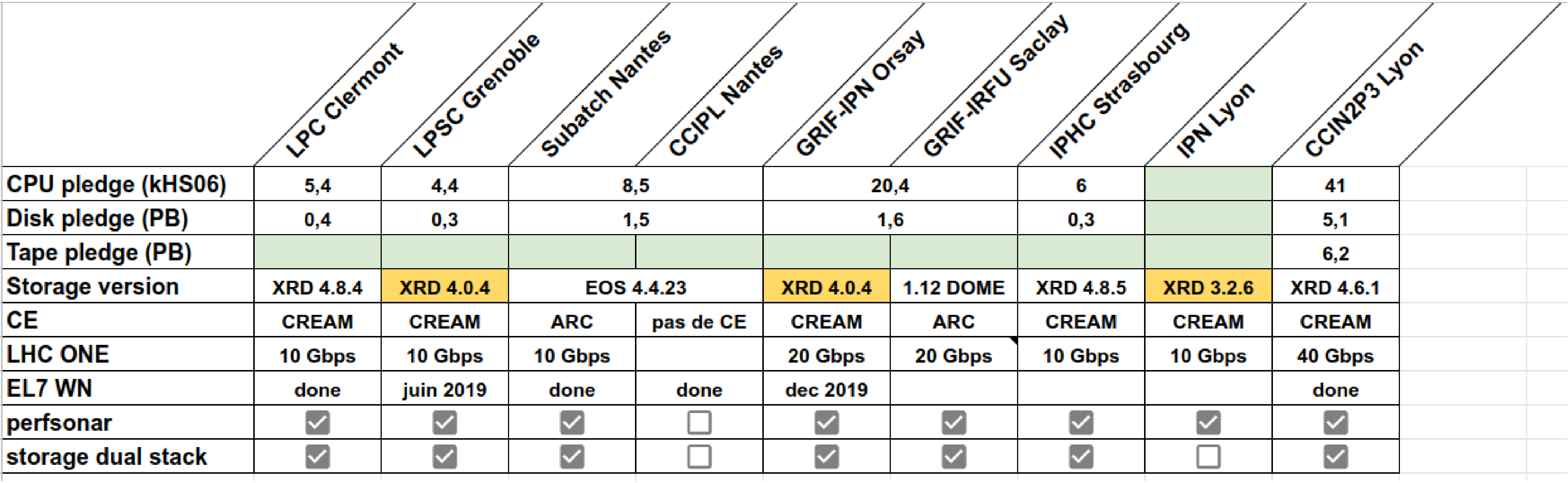


> 97 %



~ 93 %

(almost) all sites provide dual stack storage



	LPC Clermont	LPSC Grenoble	Subatech Nantes	CCIPL Nantes	GRIF-IPN Orsay	GRIF-IRFU Saclay	IPHC Strasbourg	IPN Lyon	CCIN2P3 Lyon
CPU pledge (kHS06)	5,4	4,4	8,5		20,4		6		41
Disk pledge (PB)	0,4	0,3	1,5		1,6		0,3		5,1
Tape pledge (PB)									6,2
Storage version	XRD 4.8.4	XRD 4.0.4	EOS 4.4.23		XRD 4.0.4	1.12 DOME	XRD 4.8.5	XRD 3.2.6	XRD 4.6.1
CE	CREAM	CREAM	ARC	pas de CE	CREAM	ARC	CREAM	CREAM	CREAM
LHC ONE	10 Gbps	10 Gbps	10 Gbps		20 Gbps	20 Gbps	10 Gbps	10 Gbps	40 Gbps
EL7 WN	done	juin 2019	done	done	dec 2019				done
perfsonar	☑	☑	☑	☐	☑	☑	☑	☑	☑
storage dual stack	☑	☑	☑	☐	☑	☑	☑	☐	☑

Sites : changements importants

- IRFU
  - CREAM to be decommissionned, moving to ARC6 (with SSD)
  - 100 Gbps deployed (problems NREN level)
- IPNO
  - Fusion laboratoires : Ch. voudrait garder xrootd natif
- IPNL
  - CPU contribution to drop
- Subatech + CCPIL
  - Fermeture prevue 2023
  - Ou ira le materiel prochainement installé ?
- Grenoble
  - Futur du site en consideration. Diskless ?
- CCIN2P3
  - 40(?) Gbps to LHC ONE

BACKUP

Split pilot - implementation summary

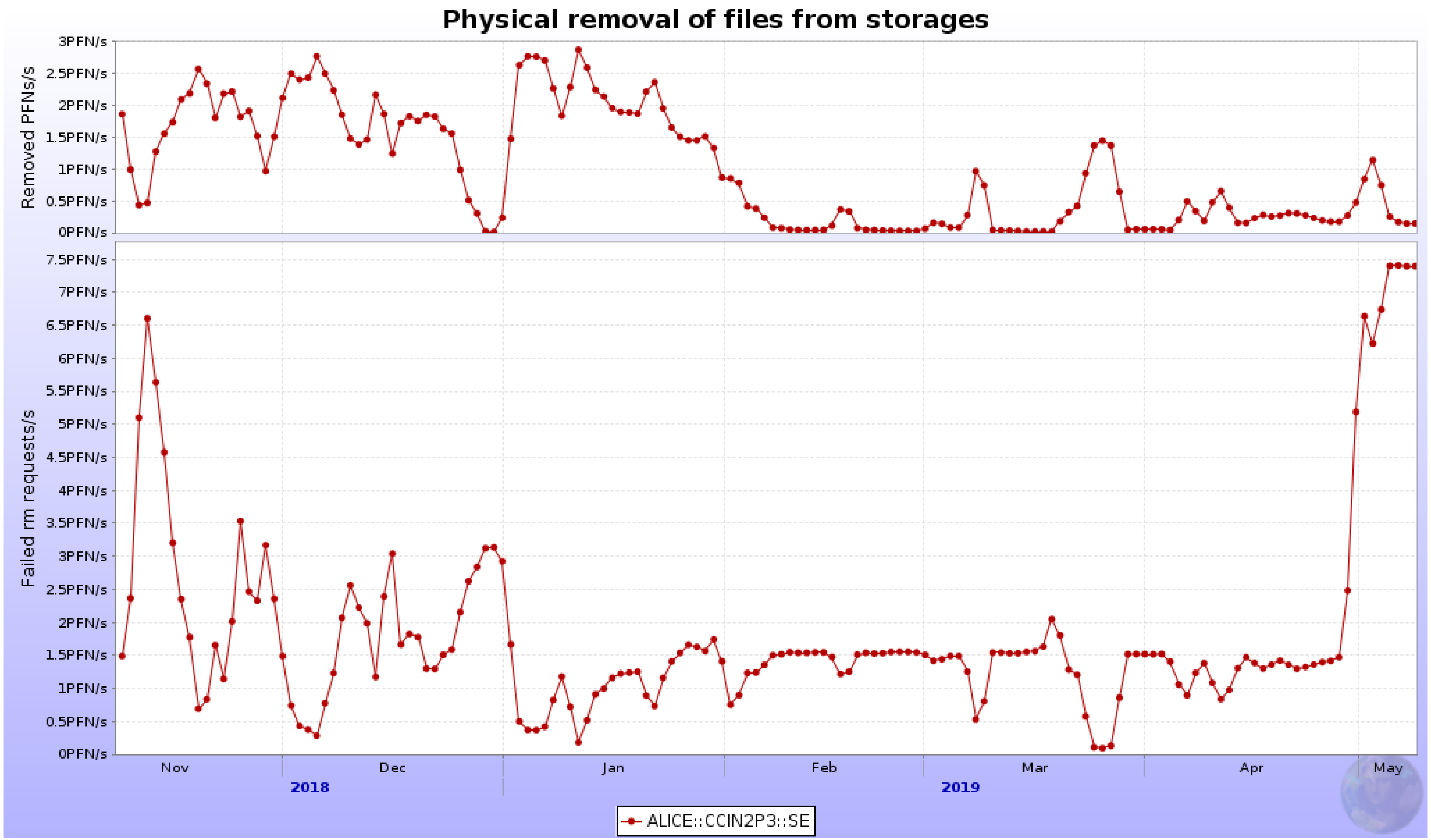
- JobWrapper has no other responsibility other than running each specific job
    - Everything else is handled by the JobAgent
    - JobWrapper sends messages to JobAgent with updates on status of the job
    - Status for the job is then changed within JobAgent
- Messages are handled by a listener process within the JobAgent and JobWrapper
    - Each received message is echoed back to confirm
- Logging options for the JobAgent are also applied to the JobWrapper, if available

ALICE::CCIN2P3::SE

- 4 PB Storage Element
  - Operations OK with jobs
- Many files to be deleted
  - Dark data (not registered in catalog)
- Deletion rate not good
  - ~ 2Hz
  - Dark data stacks up
  - Early 2019 : 180M files total, 100 Mfiles to delete
- 2 symptoms observed by Costin
  - Xrootd takes time to return answer (why?)
  - Large number of errors during deletion (why?)
- Temporary solution
  - Files deleted manually on site
  - Need to solve deletion speed in future

https://doc.cc.in2p3.fr/intranet:lcg:coordination:problem:aliceperformancesuppression

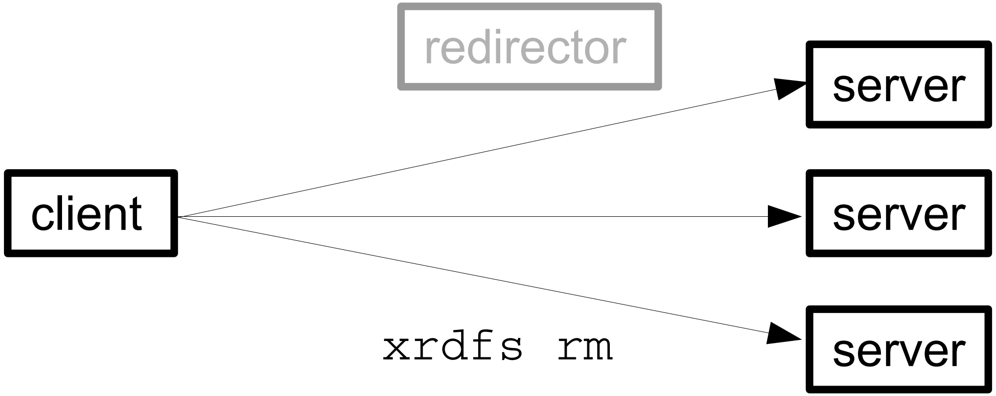
ALICE::CCIN2P3::SE



Deletion speed

Error rate

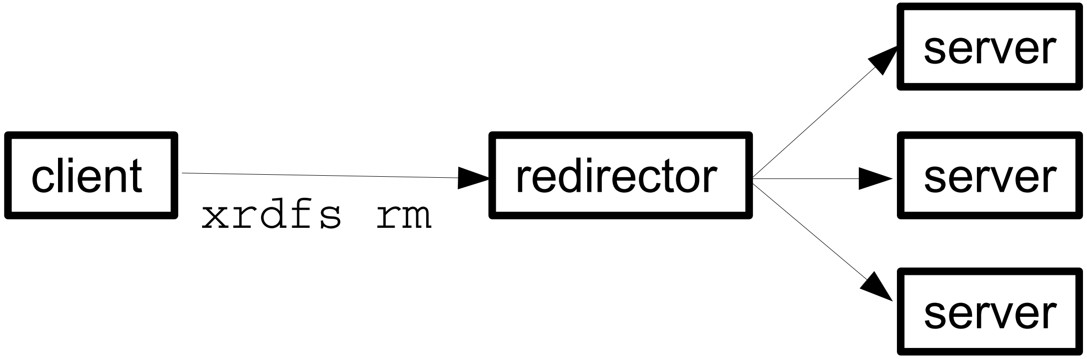
Bypassing redirector



τ ~ 10 ms

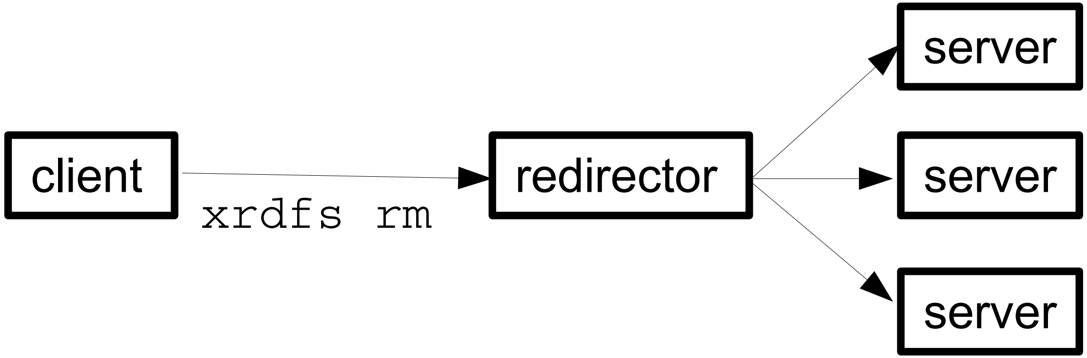
Through redirector

Files freshly written :



τ ~ 10 ms

After 'some time' :



τ = 5 s

ALICE::CCIN2P3::SE

- Many email exchanges to understand the reason
  - Cern ↔ ccin2p3 ↔ xrootd

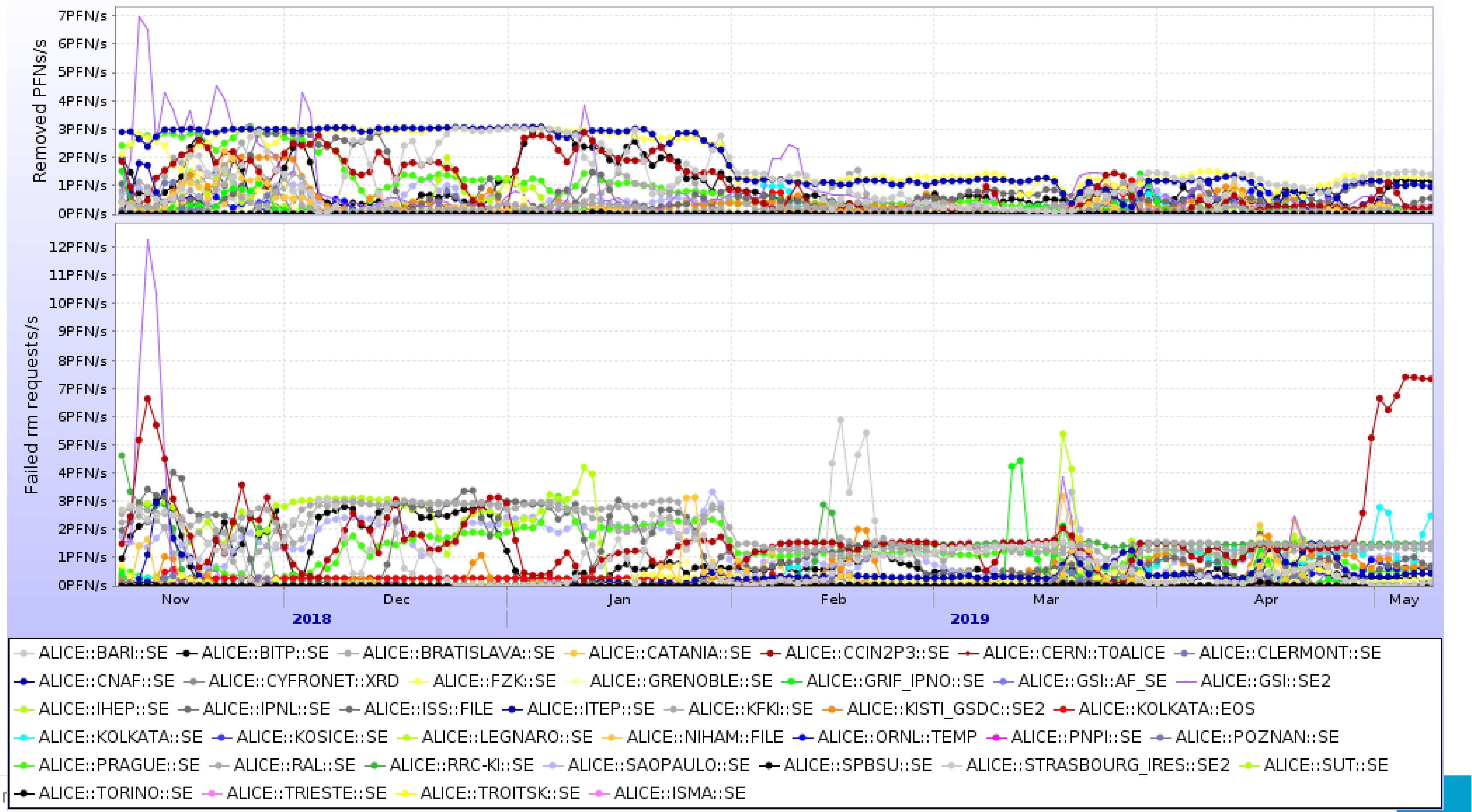
- (my personal) current conclusions
  - Cache effects
  - If file not in cache, **`cms.delay`** drives response time (default is 5 s)
  - Is that normal ? we don't know

- Xrootd support not conclusive yet

- Need more support from experts (who ?)

Is CCIN2P3 the only site in trouble ?

Physical removal of files from storages


Removed PFNs/s
7PFN/s
6PFN/s
5PFN/s
4PFN/s
3PFN/s
2PFN/s
1PFN/s
0PFN/s
Failed rm requests/s
12PFN/s
11PFN/s
10PFN/s
9PFN/s
8PFN/s
7PFN/s
6PFN/s
5PFN/s
4PFN/s
3PFN/s
2PFN/s
1PFN/s
0PFN/s
Nov
Dec
Jan
Feb
Mar
Apr
May
ALICE::BARI::SE
ALICE::BITP::SE
ALICE::BRATISLAVA::SE
ALICE::CATANIA::SE
ALICE::CCIN2P3::SE
ALICE::CERN::T0ALICE
ALICE::CLERMONT::SE
ALICE::CNAF::SE
ALICE::CYFRONET::XRD
ALICE::FZK::SE
ALICE::GRENOBLE::SE
ALICE::GRIF_IPNO::SE
ALICE::GSI::AF_SE
ALICE::GSI::SE2
ALICE::IHEP::SE
ALICE::IPNL::SE
ALICE::ISS::FILE
ALICE::ITEP::SE
ALICE::KFKI::SE
ALICE::KISTI_GSDC::SE2
ALICE::KOLKATA::EOS
ALICE::KOLKATA::SE
ALICE::KOSICE::SE
ALICE::LEGNARO::SE
ALICE::NIHAM::FILE
ALICE::ORNL::TEMP
ALICE::PNPI::SE
ALICE::POZNAN::SE
ALICE::PRAGUE::SE
ALICE::RAL::SE
ALICE::RRC-KI::SE
ALICE::SAOPAULO::SE
ALICE::SPBSU::SE
ALICE::STRASBOURG_IRES::SE2
ALICE::SUT::SE
ALICE::TORINO::SE
ALICE::TRIESTE::SE
ALICE::TROITSK::SE
ALICE::ISMA::SE