Implementation and performances of a DPM federated storage and integration within the ATLAS environment

S. Jézéquel,
C. Adam-Bourdarios, M. Gougerot, F. Chollet-Le Flour, P. Seraphin (LAPP)
S. Crepe-Renaudin, C. Gondrand (LPSC)
J-C. Chevaleyre (LPC), E. Knoops (CPPM)

5 November 2019

CHEP 2019

Introduction

- Motivation for storage federation
- Existing DPM federations
- Description of components of testbench
- First results
- Conclusion

Motivation for integrated storage federation

- Site admins
  - Share of responsibilities → tighter coordination
  - Reduce manpower to operate : especially head-node
- Funding agency
  - Look bigger in site ranking
  - Smooth local funding variations
- LHC experiments :
  - Single point of contact
  - Potential component for HL-LHC datalake

- *Challenges :*
  - *Share of informations and responsibilities*
  - *Always maintain same quality of service among sites*

Existing federated storage federations

- NDGF : dcache storage federation existing since LHC start
  - Integrate Arc-cache to preplace data and broker jobs to minimise transfers

- DPM :
  - Bern-Geneva DPM federation :
    - Bern : Headnode+ disks used for production
    - Geneva : Disk server only to store files for local users (LOCALGROUPDISK)
  - Italian federation (Napoli, Frascati, Roma) :
    - Storage federation functionnality tested
    - DPM caching mechanism evaluated for local analysis

- *New : Evaluating DPM storage performances integrated in the ATLAS production system*
  - *Demonstrate feasibility/reliability storage federation with DPM technology*
  - *Measure Wallclock/CPU eff. by replacing asynchronous transfers (FTS) with direct remote access/copy (no xcache component included yet)*

Why DPM Federation in France?

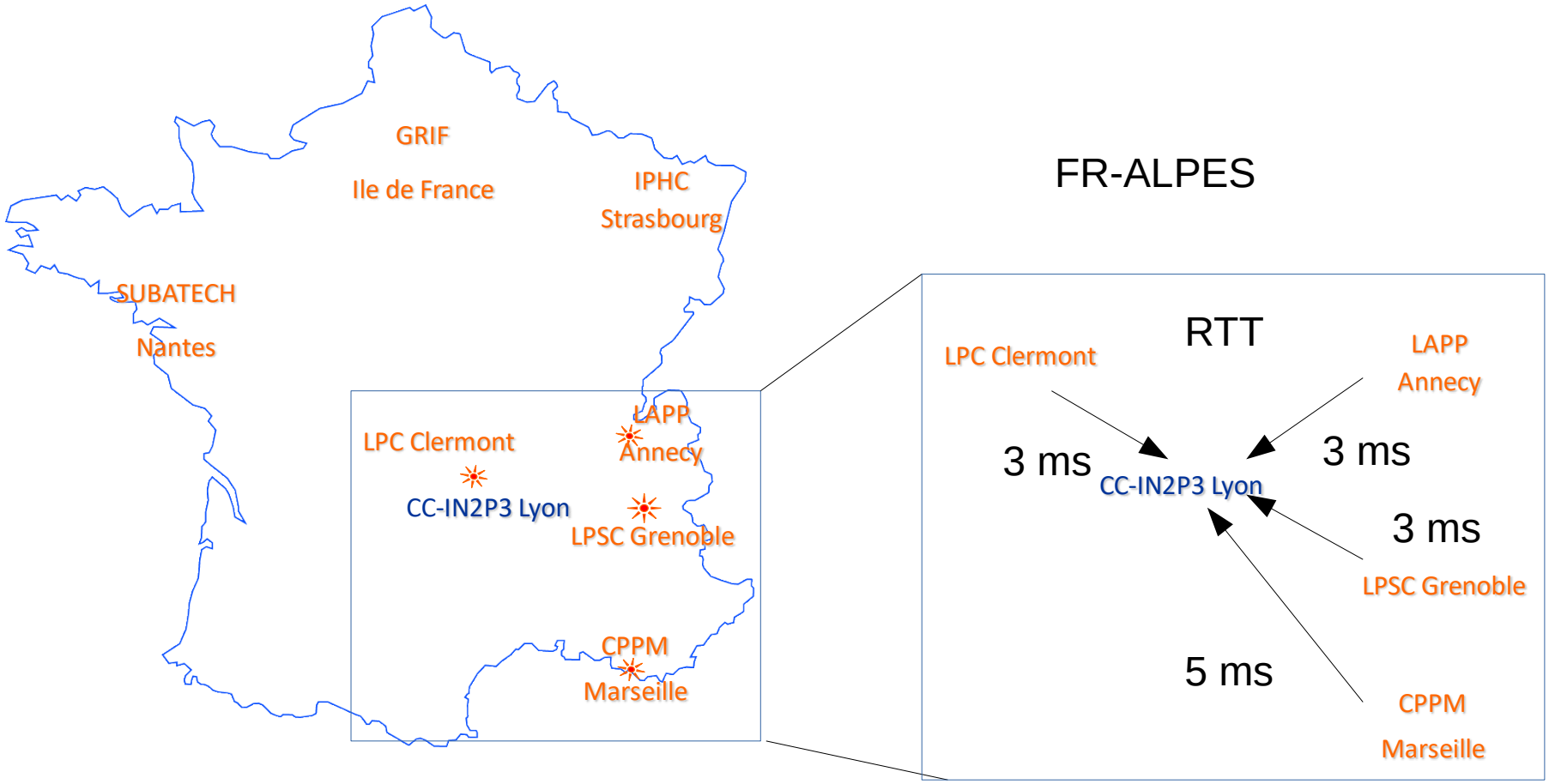
- Strong expertise in operation and maintenance of DPM storage in France
  - Recent effort to upgrade DPM to DOME+srm-less → DPM federation
- Local expertise in ATLAS Distributed computing tools
  - Integration in ATLAS information system (AGIS)
  - Integration in ATLAS HammerCloud
  - Collecting results through monitoring tools
- Crossing site admin and ATLAS contact expertise

  → Bring new ideas for R&D and evaluate them on dedicated testbed

- R&D projects
  - DOMA-FR : French forum within DOMA
  - DOMA
  - ESCAPE : Consolidate HL-LHC R&Ds for other communities : Astroparticle and other HEP

FR-ALPES in France

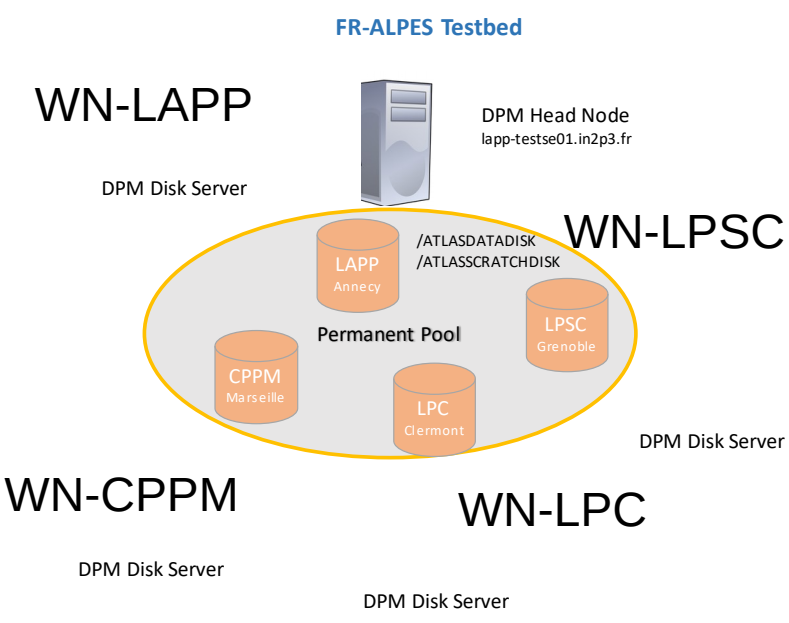
FR-ALPES : Started sites within french alps : LAPP/LPSC

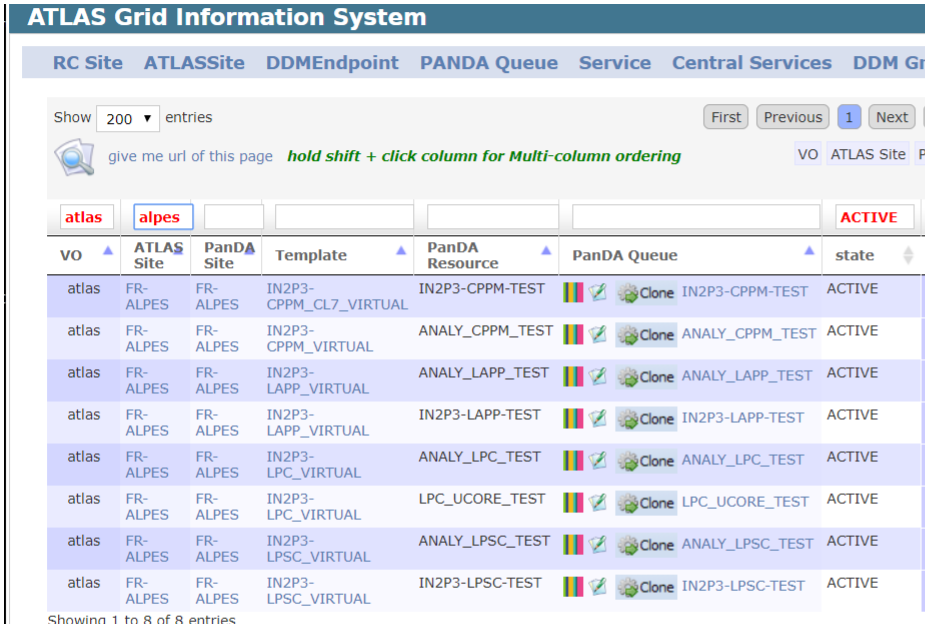


All sites connected through Lyon with LHCONE 10 Gb/s connections

FR-ALPES : South-East France DPM testbed

* Testbed built in spring 2019 integrated in ATLAS Grid (as test)





Web link

* Component connectivity :
  - WN connection : 1 Gb/s
  - Disk server connection : 10 Gb/s

Performance measurements : HammerCloud

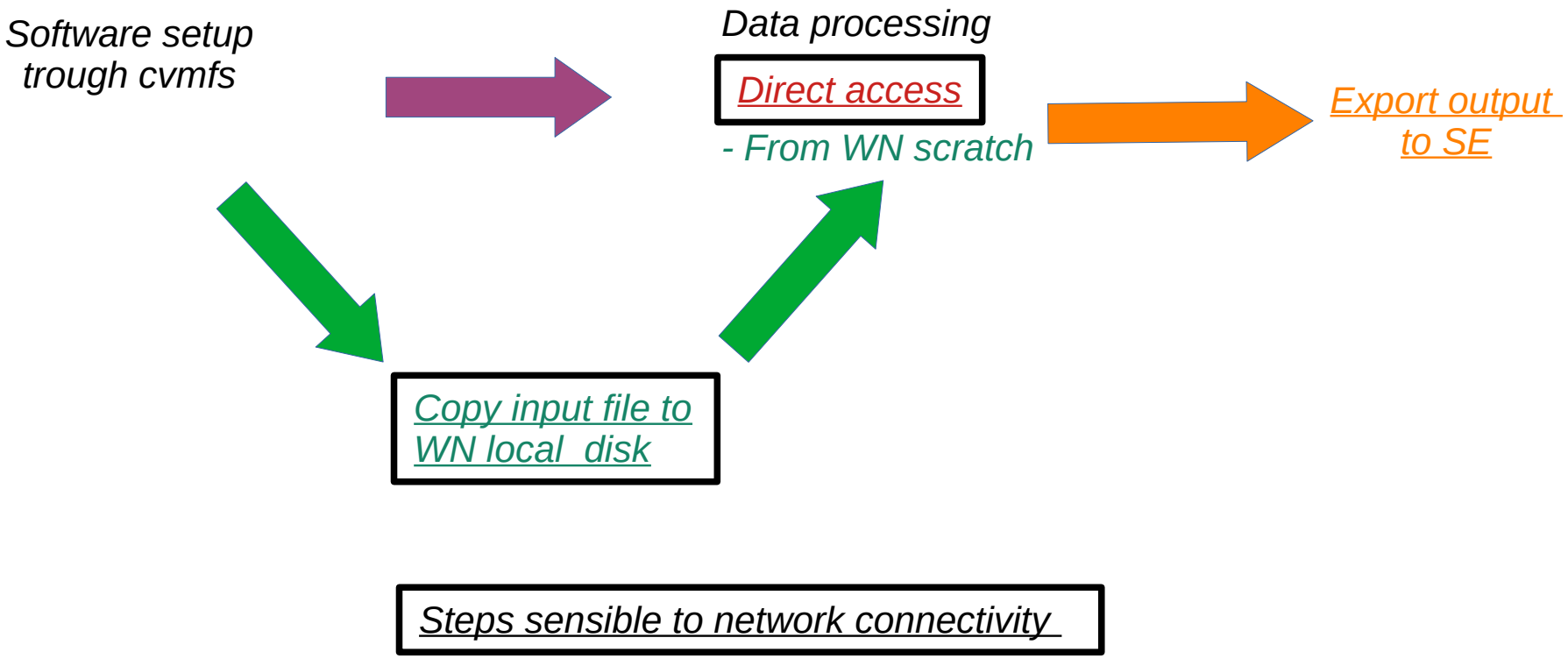
Running Tests backed by the WLCG Data Lake, group wlcgdatalakes_abc

State	Id	Host	Template	Start (Europe/Zurich)	End (Europe/Zurich)	Sites	subm jobs	run jobs	comp jobs	fail jobs	fail %	tot jobs
running	20147583	hammercloud-ai-11	1082: P.F.T. benchmark digi+reco derivation Athena/21.0.53 5 events - WLCG Data Lakes - copy2scratch folder LPSC ABC	24/Oct, 17:18	25/Oct, 18:31	IN2P3-LAPP-TEST, IN2P3-LPSC-TEST, LPC_UCORE_TEST, 1 more...	5	3	81	0	0	89
running	20147594	hammercloud-ai-11	1061: P.F.T. digi+reco derivation Athena/21.0.53 5 events - WLCG Data Lakes - copy2scratch folder LAPP/JINR ABC	25/Oct, 0:20	25/Oct, 23:57	IN2P3-LAPP-TEST, IN2P3-LPSC-TEST, JINR_UCORE-TEST, 4 more...	21	3	88	0	0	112
running	20147599	hammercloud-ai-11	1062: P.F.T. benchmark derivation AthDerivation/21.2.8.0 1k events - WLCG Data Lakes - copy2scratch folder LAPP/JINR ABC	25/Oct, 3:00	26/Oct, 4:22	IN2P3-LAPP-TEST, IN2P3-LPSC-TEST, JINR_UCORE-TEST, 3 more...	3	6	73	0	0	82
running	20147601	hammercloud-ai-11	1059: P.F.T. mc16 Sim_tf 21.0.16 - WLCG Data Lakes - copy2scratch folder LAPP/JINR ABC	25/Oct, 3:36	26/Oct, 2:11	IN2P3-LAPP-TEST, IN2P3-LPSC-TEST, LPC_UCORE_TEST, 4 more...	9	1	42	0	0	52
running	20147602	hammercloud-ai-11	1060: A.F.T. AtlasDerivation 20.7.6.4 direct accesss folder LAPP ABC	25/Oct, 3:44	26/Oct, 4:25	ANALY_LAPP_TEST, ANALY_LPSC_TEST, ANALY_LPC_TEST, 1 more...	7	0	6	0	0	13

Web link

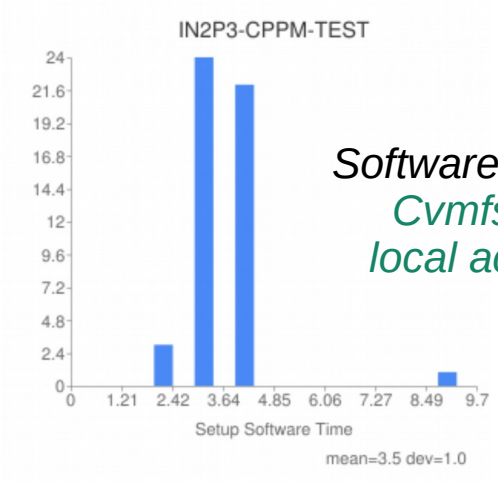
- Test campains managed through Hammercloud Jobs
  - Recycle setup built for WLCG DataLake
- Different datasets created and pre-positionned on disk servers at LAPP/LPSC
  - Procedure : All other disk servers set in read-only mode during data placement
- Different job types and configuration :
  - Low vs high IO
  - Direct access vs copy on scratch WN (copy2scratch)

Steps in job processing

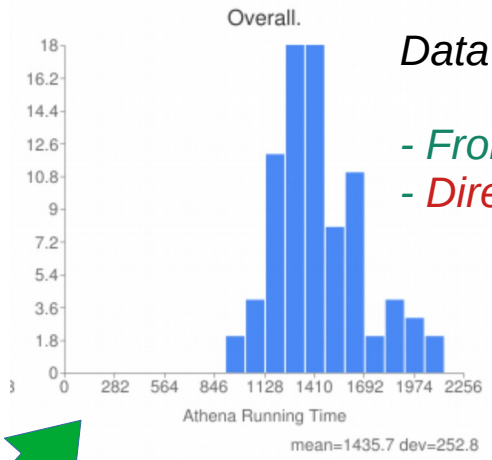


* Ge

Result displays

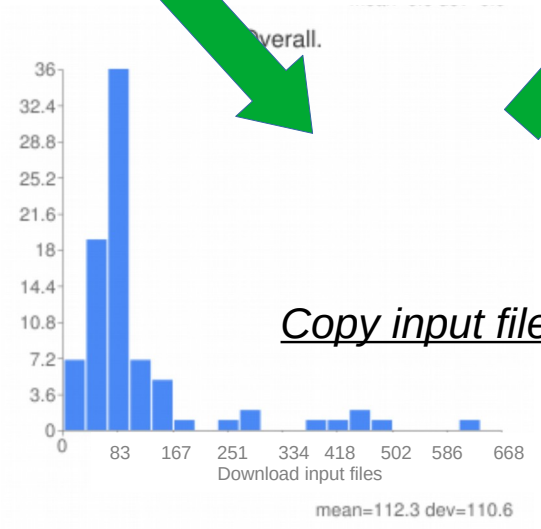


Software setup :
Cvmfs →
local access

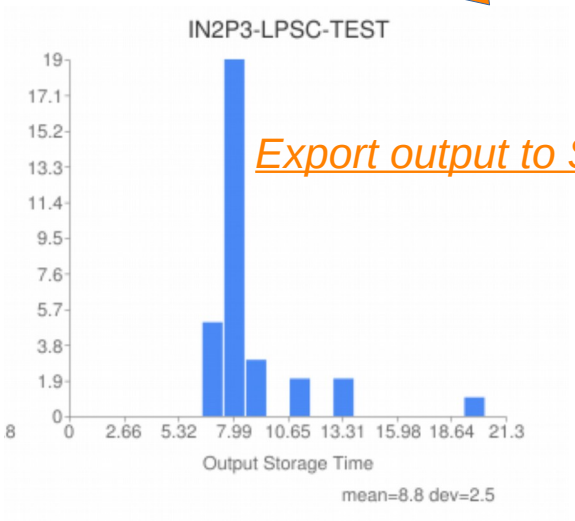


Data processing

- From WN scratch
- Direct access



Copy input file



Export output to SE

Plots/Results only for illustration

Simulation jobs : Low IO

- Benchmark of network access (copy2scratch) + WN processing speed
- Processing different samples of 2 events @LAPP and @LPSC
- Input file size (EVNT file) : LAPP/LPSC : ~38 MB
- Low IO → Should minimise impact from remote vs local

Preliminary

		Duration (s)							
	Action	Data @ LAPP				Data @ LPSC			
WN location		CPPM	LAPP	LPC	LPSC	CPPM	LAPP	LPC	LPSC
Copy2scratch + processing from WN	Download Input	O (5-10 seconds)				O (5-10 seconds)			
	Processing	O (500-700 seconds)				O (700-900 seconds)			
Direct access from SE	Processing	O (500-800 seconds)				O (700-1000 seconds)			

- Debugging of configurations and results not complete → Only orders of magnitude available

Digi+reco jobs : High IO

- Digitisation+reconstruction jobs : Heavy IO job
- Input file size (HITS files): LAPP/LPSC : 3 GB (HS) + (2+1) GB (Pile Up evts)
- Processing different samples of 2 events @LAPP and @LPSC

Preliminary

		Duration (s)							
	Action	Data @ LAPP				Data @ LPSC			
WN location		CPPM	LAPP	LPC	LPSC	CPPM	LAPP	LPC	LPSC
Copy2scratch + processing from WN	Download Input	O (100-1000 seconds)				O (200-1000 seconds)			
	Processing	O (1500-2000 seconds)				O (2000-3000 seconds)			
Direct access from SE	Processing	O (1600-3000 seconds)				O (2000-2500 seconds)			

- Results still very sensitive to difference of configurations at site → Significant variants between expected similar results : Reported ATLAS devs

Conclusion

- DPM storage federation among 4 sites operationnal for testing
  - Operationnal at small scale and short duration
  - Should be tested and long jobs (keep SE connection for long)
- Measurement infrastructure built based on Grid tools used in ATLAS
  - Discussions with experts to improve speed and precision of measurements
  - Similar infrastructure used to measure xcache performances
- First results obtained but still require detailed understanding of interaction between ATLAS tools and site configuration
- Next steps
  - Improve display of informations provided by HC (under discussion)
  - Follow-up with similar analyses and other data placement (Arc,...)
  - Disentangle network from local effects
  - Dedicate WN to ensure so degradation from uncontrolled jobs with same WN

  → Reach conclusion by early 2020 presented in future DOMA meeting

Backup