

FR-ALPES status

S. Jézéquel,

C. Adam-Bourdarios, M. Gougerot, F. Chollet-Le Flour, P. Seraphin
(LAPP)

S. Crépe-Renaudin, C. Gondrand (LPSC)

J-C. Chevaleyre (LPC), E. Knoops (CPPM)

12 Décembre 2019

High Level Storage and Data Management (RUCIO)



Federates and define datalakes (quotas, acls, replication rules)

Datalakes formed by several storage centers with at least one archive center

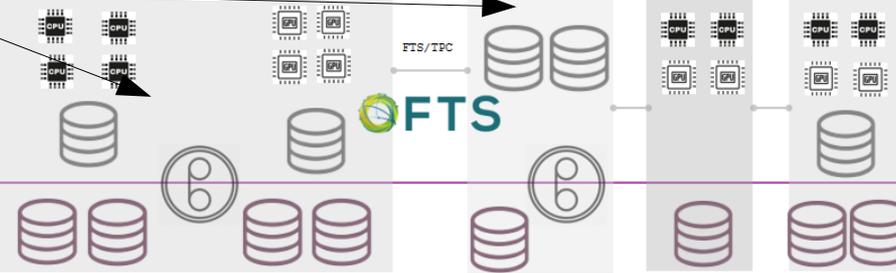
Worldwide datalake infrastructure formed by the different regional datalakes

Allow to fit different possibilities for data replication and data placement models

Regional datalake

Sites with large storage and/or archive facility and CPU

Federation(s) with large storage and/or archive facility and CPU



Datalake

Federated Storage
File IO: FTS, xroot, http

Remote IO

Caching and latency hiding

Sites with low latency providing CPU and accessing data via remote IO

Site with CPUs and stateless storage as a buffer/cache

Site with CPUs, managed storage and buffer/cache

HPC
HPC centers with a cache as a stage in/out area

Edge Services
Stateless sites: CPU clusters and caches remotely deployed and operated via K8s and Slate

Commercial clouds with a cache as a stage in/out areas

Several site topologies
Allow sites to chose the best model for local and pledged resources
Allow experiments to optimise the workload management in different scenarios

Possible Joint National Initiatives

Opportunistic resources

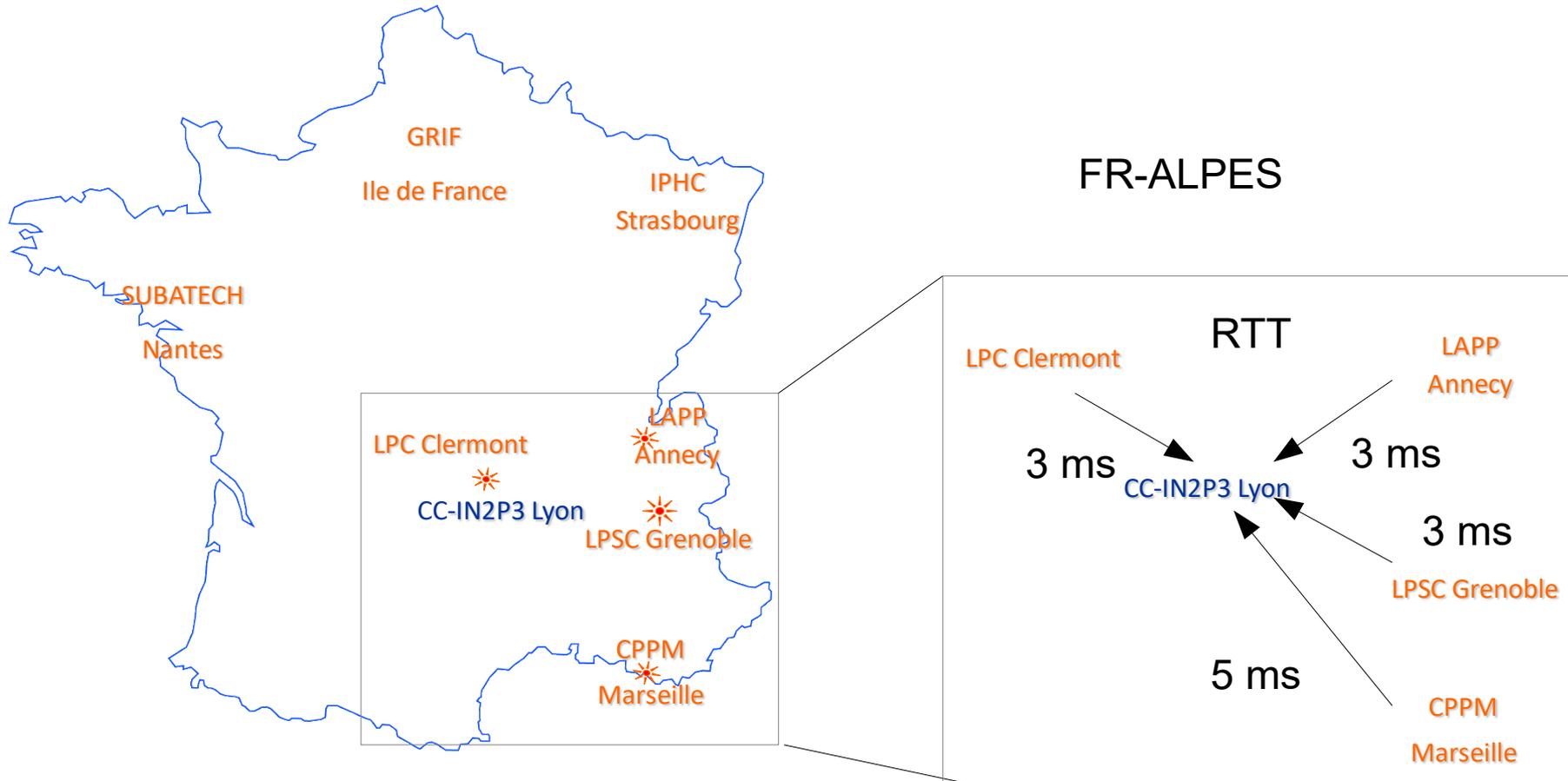
Diskless sites (~10-15 in ATLAS)

Datalakes, latency hiding and caching - Xa

- * Volonté d'équipes (physiciens + informaticiens) de participer au R&D HL-LHC Computing
- * Synergies entre laboratoires et physiciens ATLAS (LAPP-LPSC → FR-ALPES)
- * Compétences et intérêt sur le stockage : stratégie au-delà LHC
- * Datalake : restriction du nb d'entités de stockage : 'big is beautiful'
- * Temps d'accès aux données : plus optimal en minimisant Round Trip Time
< 10 ms entre sites ('regional datalake' < 20 ms)
→ FR-ALPES : premier jalon vers 'regional datalake' français/européen

FR-ALPES : collaboration CPPM, LAPP, LPC, LPSC

- testbed : Fédération de stockage DPM opérationnelle (100 TB)



* Administrateurs de site

- Réduire le travail de maintenance en particulier sur headnode
→ Permet de palier en partie la baisse des effectifs dans labos
- Partage des responsabilités/droits d'accès + expertises + idées
→ résoudre les problèmes et mettre en place de nouveaux développements
- Récent upgrade DPM → Facilite construction fédération stockage DPM

* Agence de financements

- Avoir des fédérations plus visible sur WLCG :
 - LAPP+LPSC+LPC+CPPM : 6 PB pledge pour ATLAS → un des plus gros site T2
- Lisser les variations de financement local

* Intérêt des expériences LHC :

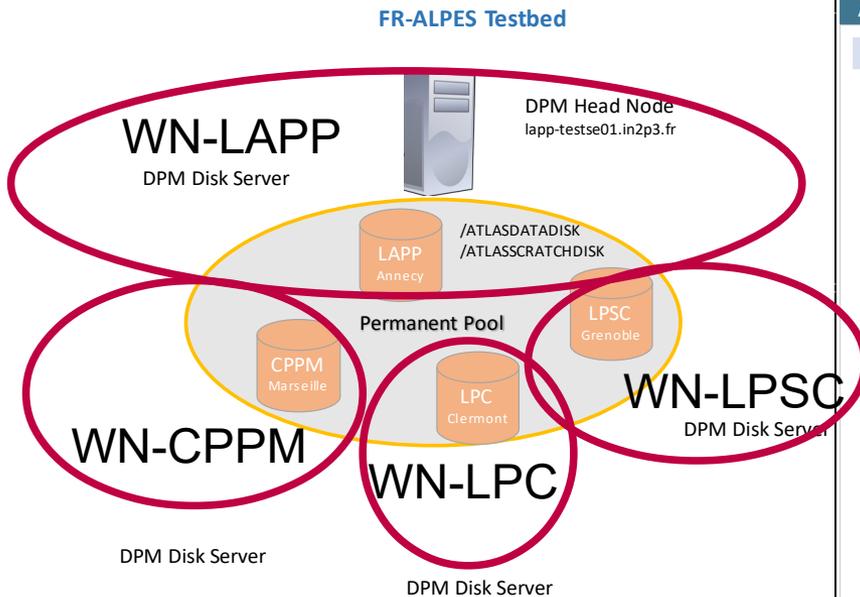
- Réduire le nb de points de contact

- * Dcache : NDGF (Scandinavie+Slovénie) : Depuis le début du LHC
 - Pas d'accès direct au SE mais utilisation de Arc-cache pour transférer les données près des CPUs

- * DPM :
 - Bern-Geneva DPM federation :
 - Bern : headnode+ stockage pour la production
 - Geneva : stockage pour utilisateurs locaux (LOCALGROUPDISK)
 - Fédération italienne (Napoli, Frascati, Roma) :
 - Testbed construit pour CHEP 18
 - Focalisation sur caching DPM pour analyse locale

- * Site admins :
 - Validation de la fiabilité/scalabilité d'une fédération de stockage avec techno DPM utilisant des organisations de stockage actuelles
 - Partage effectivement travail et informations en temps réel
 - Maintenir la qualité de service globale
- * Perte localisation des données sur 4 sites → Accès au stockage par WN plus souvent à travers WAN que LAN
 - Impact sur le trafic réseau entre sites
 - Impact sur taux d'erreurs des jobs et 'CPU/Walltime efficiency' ?
 - A mesurer en essayant de séparer accès distants et locaux (ATLAS phys.)
- * Suisse envisage de faire idem pour palier arrêt dcache à CSCS
 - Fédération de stockage DPM Bern-CSCS

- * Testbed démarré au printemps 2019 : LAPP+LPSC
 - Intégration de LPC et CPPM à l'été 2019
- * Utilise les outils de Grille ATLAS + HammerCloud



ATLAS Grid Information System

RC Site: atlas | ALPES | ... | ACTIVE

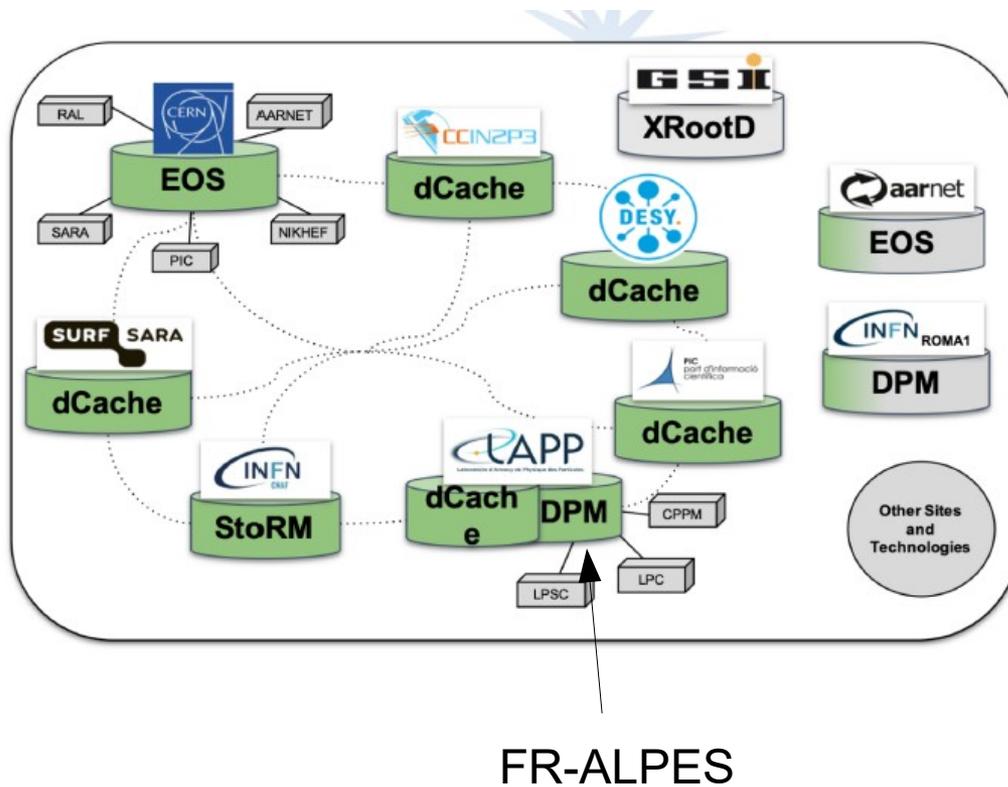
VO	ATLAS Site	PanDA Site	Template	PanDA Resource	PanDA Queue	state
atlas	FR-ALPES	FR-ALPES	IN2P3-CPPM-CL7_VIRTUAL	IN2P3-CPPM-TEST	IN2P3-CPPM-TEST	ACTIVE
atlas	FR-ALPES	FR-ALPES	IN2P3-CPPM_VIRTUAL	ANALY_CPPM_TEST	ANALY_CPPM_TEST	ACTIVE
atlas	FR-ALPES	FR-ALPES	IN2P3-LAPP_VIRTUAL	ANALY_LAPP_TEST	ANALY_LAPP_TEST	ACTIVE
atlas	FR-ALPES	FR-ALPES	IN2P3-LAPP_VIRTUAL	IN2P3-LAPP-TEST	IN2P3-LAPP-TEST	ACTIVE
atlas	FR-ALPES	FR-ALPES	IN2P3-LPC_VIRTUAL	ANALY_LPC_TEST	ANALY_LPC_TEST	ACTIVE
atlas	FR-ALPES	FR-ALPES	IN2P3-LPC_VIRTUAL	LPC_UCORE_TEST	LPC_UCORE_TEST	ACTIVE
atlas	FR-ALPES	FR-ALPES	IN2P3-LPSC_VIRTUAL	ANALY_LPSC_TEST	ANALY_LPSC_TEST	ACTIVE
atlas	FR-ALPES	FR-ALPES	IN2P3-LPSC_VIRTUAL	IN2P3-LPSC-TEST	IN2P3-LPSC-TEST	ACTIVE

Showing 1 to 8 of 8 entries

[Web link](#)

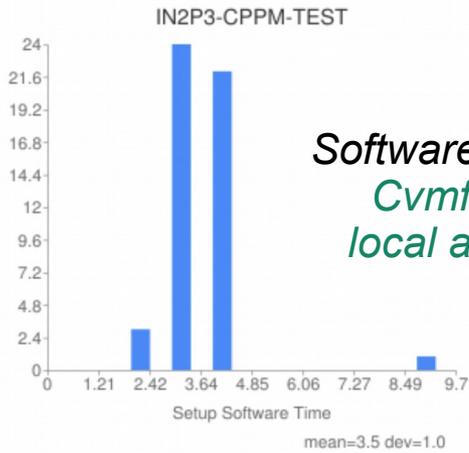
- * Connectivité :
 - WN : 1 Gb/s
 - Disk servers : 10 Gb/s

- * Fédération de stockage pour WLCG DOMA et ESCAPE
 - Réorganisation en single-pool pour toutes les Vos (EK)
 - Nouveaux modes d'authentification : macaron, IAA (FC)

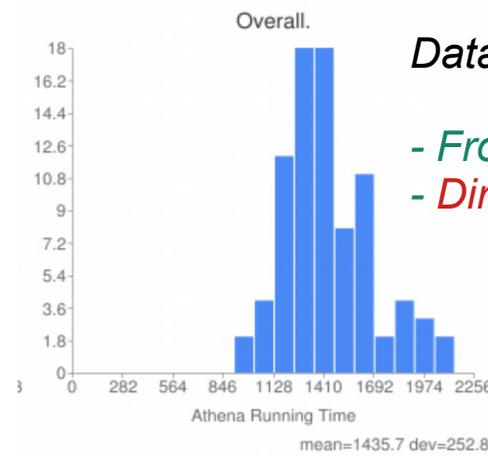


Fédération ESCAPE

- * Soumission automatique de jobs calibrés : CERN-IT tool
 - Blacklisting des sites pour prod ATLAS
 - Mesure de performances des sites
 - Evaluation des performances du EU-datalake
 - Donne détail des étapes d'un job :
 - Initialisation du soft avec cvmfs
 - Copie fichiers inputs : Grid SE → WN scratch
 - Processing
 - Copie fichier output : WN scratch → Grid SE
- Réutilisation pour FR-ALPES et ESCAPE

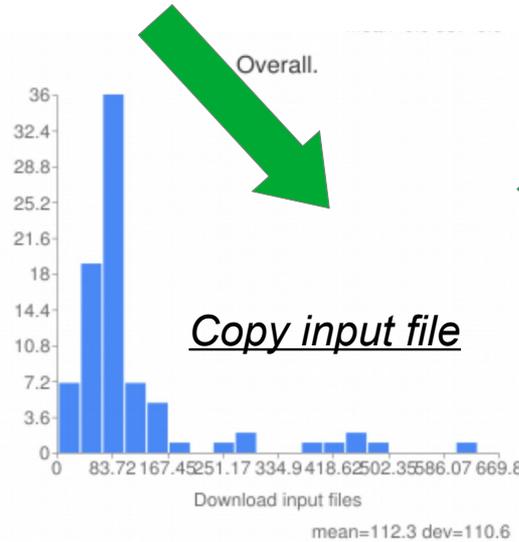


Software setup :
Cvmfs →
local access

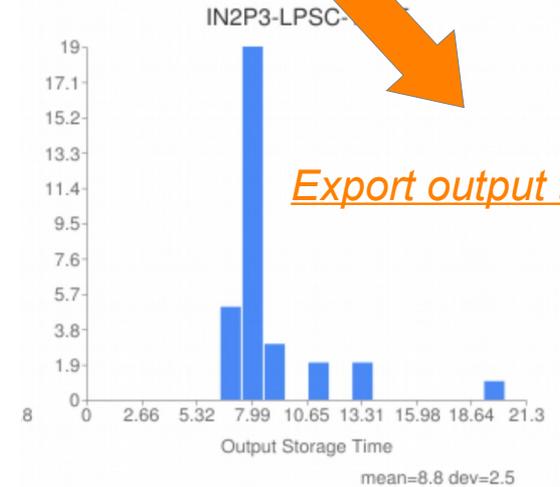


Data processing

- *From WN scratch*
- *Direct access*



Copy input file

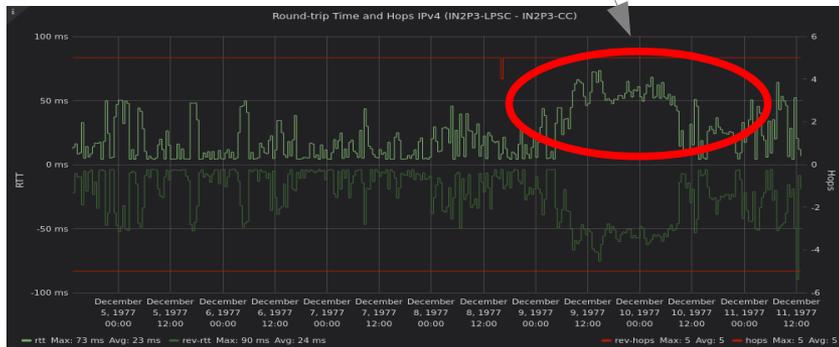


Export output to SE

- * Bug dans le monitoring ATLAS
 - WallClocktime mal évalué pour certains jobs ATLAS sur sites 'Arc-CE'
 - Problème ATLAS identifié et correctif à venir
- * HammerCloud
 - : encore difficile de trouver les jobs qui posent problème
 - Solution : devenir expert en Elastic Search (base à Chicago)
- * Saturation des réseaux géographiques → Instabilité dans les résultats
 - Compétition des ressources réseau avec FTS
 - Solution possibles
 - Avoir une vision de l'activité WAN des 4 sites → **demande monitoring (JCC)**
 - Augmenter le débit des liaisons réseaux : $N * 10 \text{ Gb/s}$
 - Bridger FTS avant saturation de 10 Gb/s
 - Xcache devant les WN : Réduction du trafic recurrent : **déploiement en évaluation (ED)**

HC jobs : 2 events
 AOD → DAOD
 Fichiers au LAPP sur FR-ALPES

Production HITS → AOD : 1000 events
 LAPP_ALPES... = accès FR-ALPES
 Autres = accès SE local

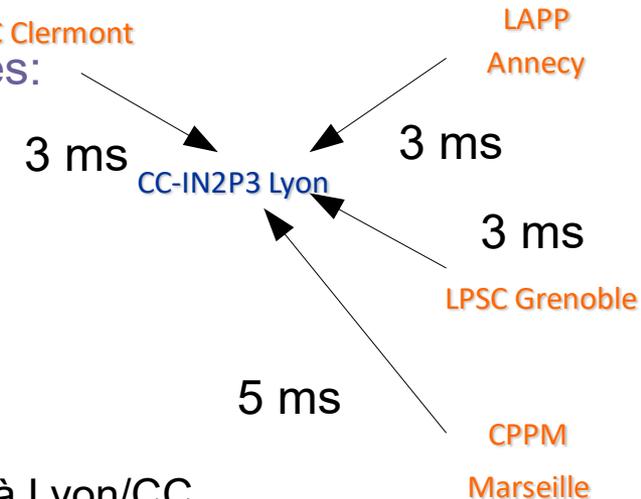


Premiers résultats encourageants

- * Inclure CCIN2P3 dans l'évaluation des accès directs extérieurs au stockage dcache/Grille du CC dans les mesures de performances d'accès distant → Nouvelle étape dans 'regional datalake' français puis européen

- * Idée de positionner Headnode DPM au centre des sites:

- Actuellement : En bout de réseau LHCONE
 - Pas de redondance réseau
- CCIN2P3 :
 - Point de passage obligé pour réseau
Headnode ↔ Disk servers
- Intérêt technique de l'hébergement du headnode DPM à Lyon/CC
 - CC fournirait une infra
 - Headnode installé et opéré par admins FR-ALPES



- * FR-ALPES : Testbed opérationnel pour fédération de stockage
 - Collaboration efficace entre administrateurs/phys. CPPM/LAPP/LPC/LPSC
 - Simplification maintenance/opération
 - Documentation dans DOMA-FR Wiki ([Lien](#))
 - Élément pour R&D HL-LHC : DOMA(-FR) et ESCAPE
 - Présentation à CHEP19 ([Lien](#))
 - Démarrage avec outils pour ATLAS
 - Premiers résultats encourageants
 - Limitation principale : Saturation récurrente du WAN avec utilisation actuelle
 - Une des première brique du ‘regional datalake’ européen
 - Futur : Possibilité d’évaluer nouveaux composants
 - Augmentation des performances/disponibilités ?
 - Ouverture stratégique pour l’avenir
 - Possible collaboration avec le CCIN2P3 ?
 - Mise à disposition pour manip astros :
 - Premier exemple : CTA/LAPP dans le cadre ESCAPE

Software setup
through cvmfs

Data processing

Direct access

From WN scratch

Export output
to SE



Copy input file to
WN local disk
: Copy2scratch

Steps sensible to network connectivity