ATLAS data access performances among ALPAMED sites

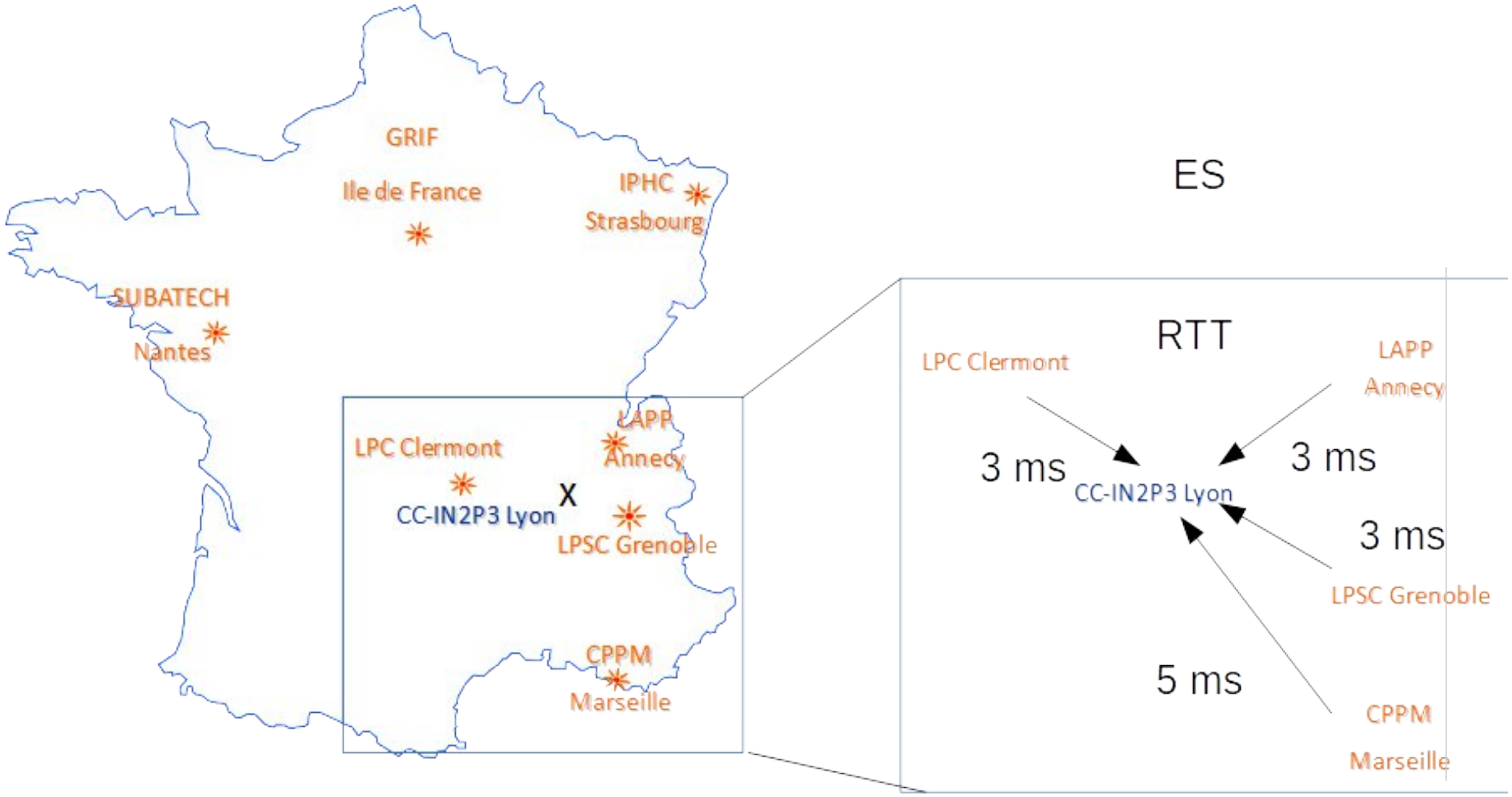
C. Adam-Bourdarios, S. Jézéquel (LAPP)

on behalf of E. Knoops (CPPM), F. Chollet, P. Seraphin (LAPP),

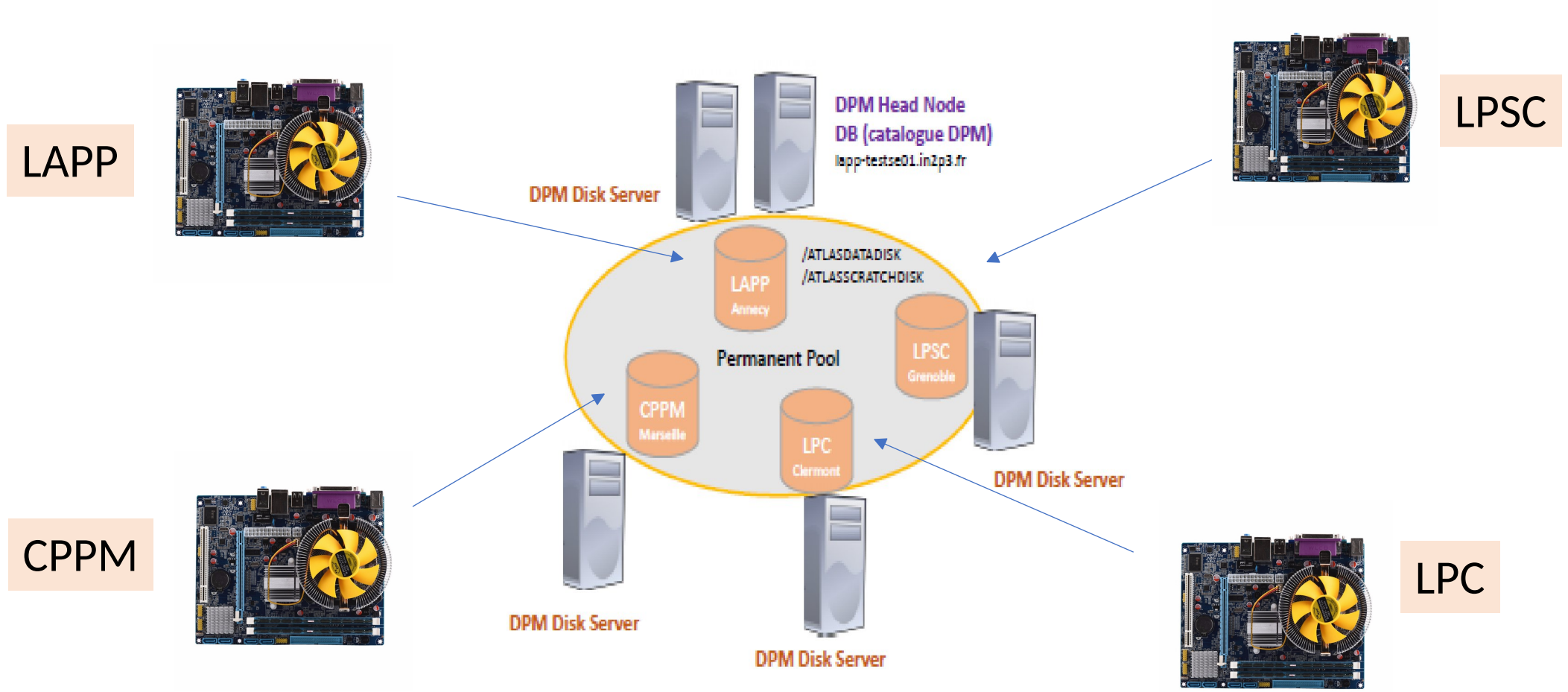
J-C Chevaleyre (LPC), S. Crepe, C. Gondrand (LPSC)

Storage Workshop : 19 November 2020

ALPAMED sites



ALPAMED infrastructure



Motivation : Reduce Total Cost of Ownership to operate storage infrastructure (FTE, headnode hardware)

Target : Data access performance

Laboratoire d'Annecy de Physique des Particules

- Primary goal : Measure data access rate vs short distance (N*100kms)
  - Direct access to SE : Default for user jobs but rare for production jobs (copy to local disk)
  - Penalty to loose locality of data over DPM storage federation (ALPAMED : South East France T2)
    - Comparison files already on local WN disk scratch or local SE
- By-product : Access efficiency of CPU_only sites to remote storage within
  - Regional distance (< 10 ms RTT)
  - European Data Lake (< 20 ms RTT)
- Calibrated jobs submitted through HC infrastructure
  - Single job reading 1000 events out of 7800 (total file size : 4.4 GB) : 10 minute jobs
    - 1 input file prepositioned per location
  - Evaluate direct access performances vs local copy

Beyond ALPAMED

- Add IN2P3-CC CPU/SE in performance matrix (called CC)
- Measure data access from far storages :
  - CERN (2 ms latency from CC)
  - GRIF-LNPHE (6 ms)
  - Freiburg (23 ms - Not in LHCONE)
  - Prague (17 ms)

CPU efficiency : files already on WN disk

Laboratoire d'Annecy de Physique des Particules

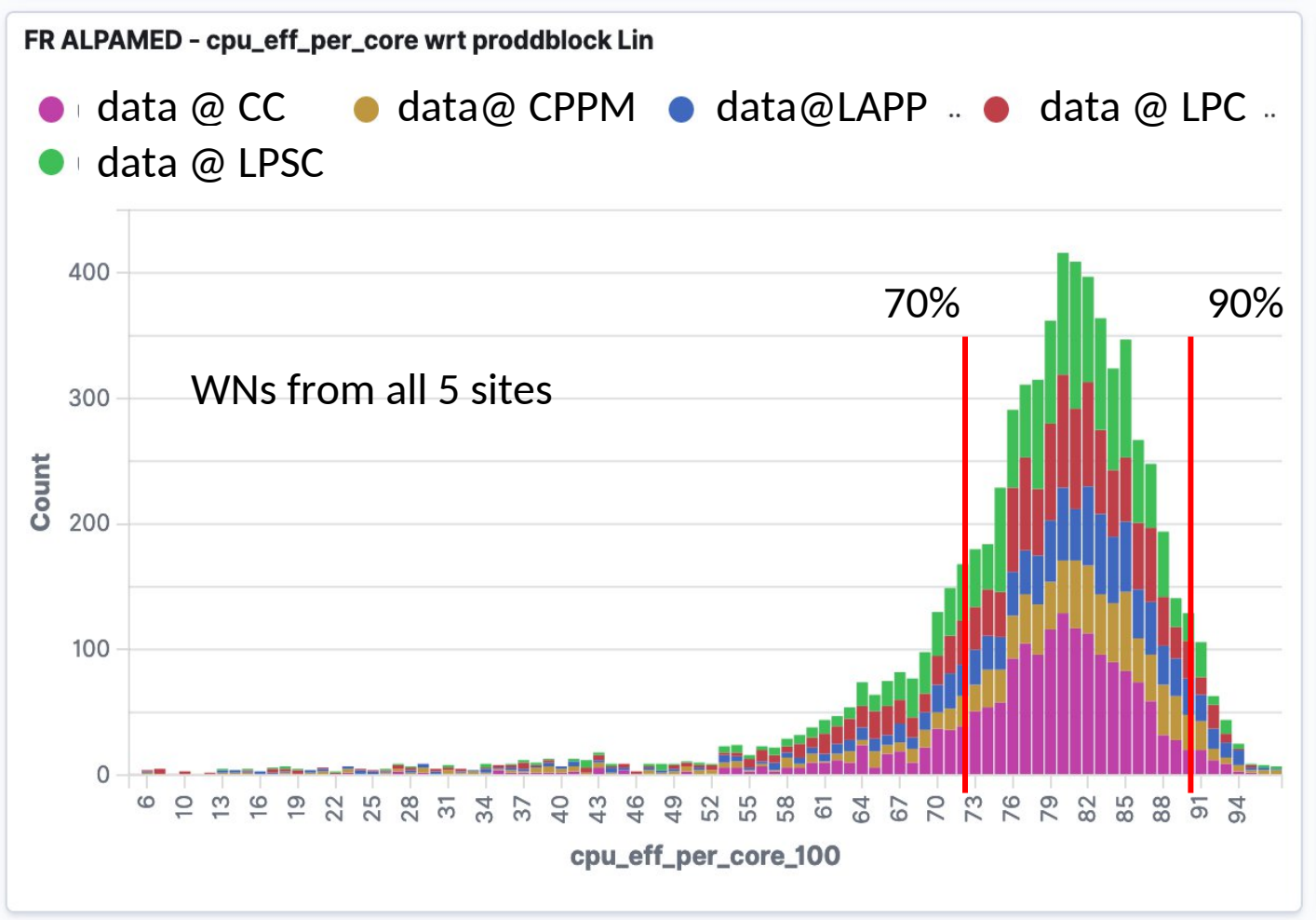
- AOD --> DAOD derivation

Mimic analysis jobs with large IO

CPU efficiency = CPU_time/(CPU_time+IO_time)

- Mean ~80 % : Large IOs --> challenging for network

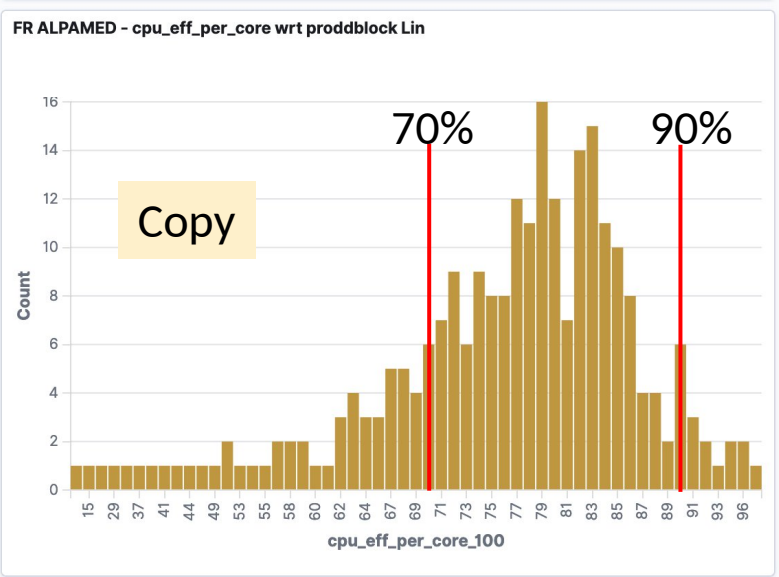
- Tails at low efficiency :
  - Memory swap
  - Disk heavily accessed

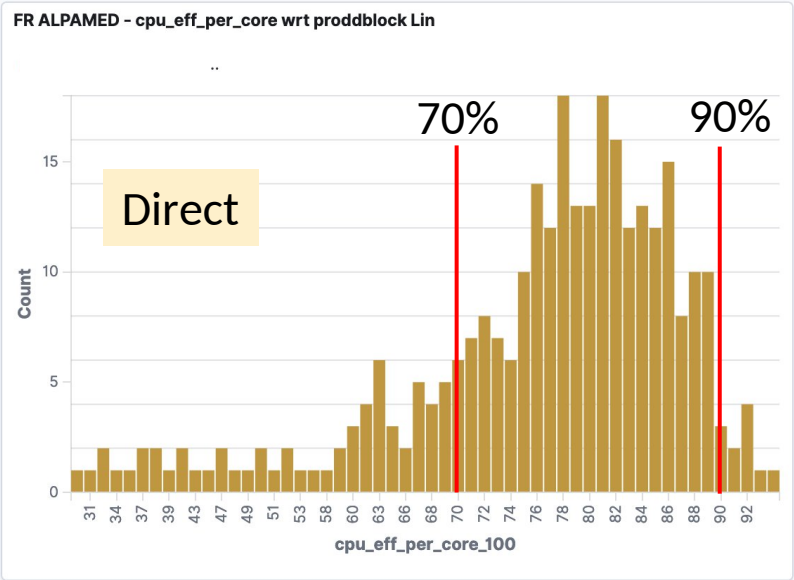


CPU efficiency : Local access (xcheck)

Laboratoire d'Annecy de Physique des Particules

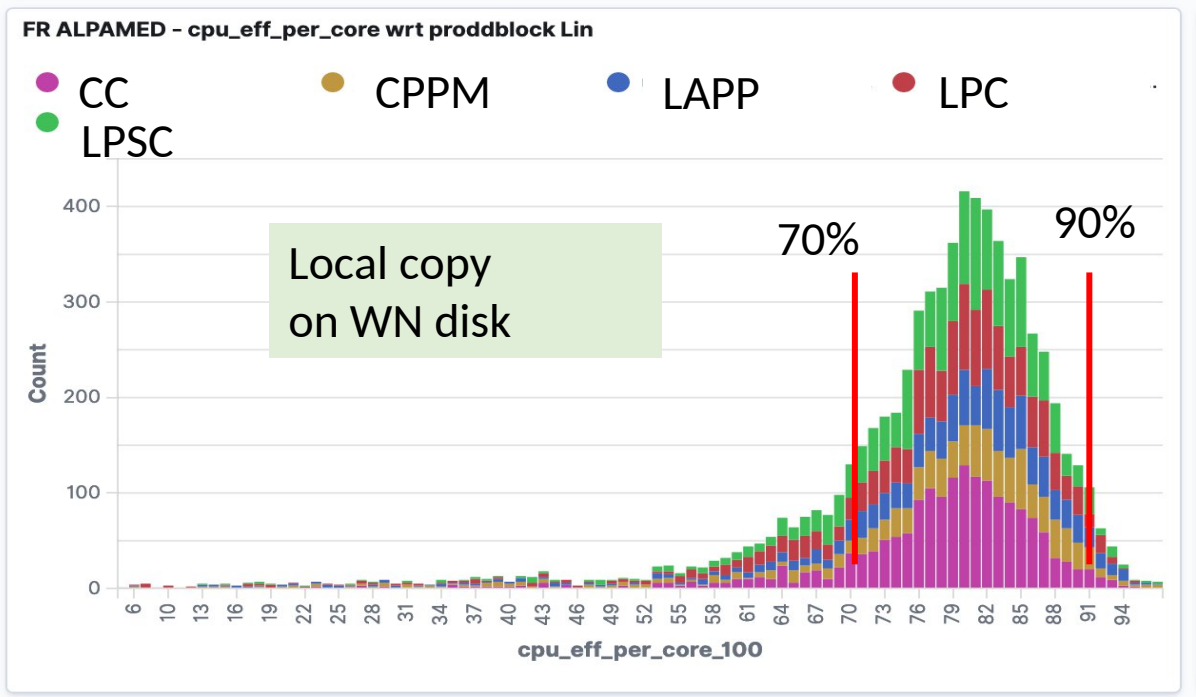
LAPP WN access LAPP storage (WN disk or local SE)

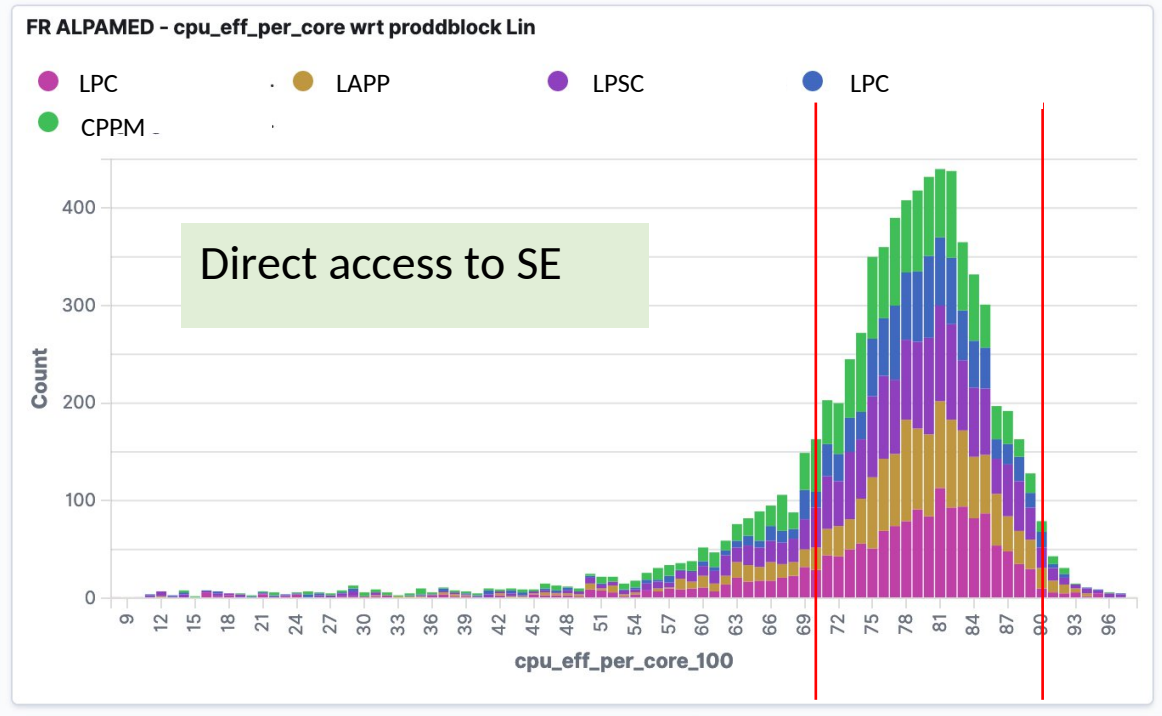




No visible degradation

CPU eff. : direct access : Close sites

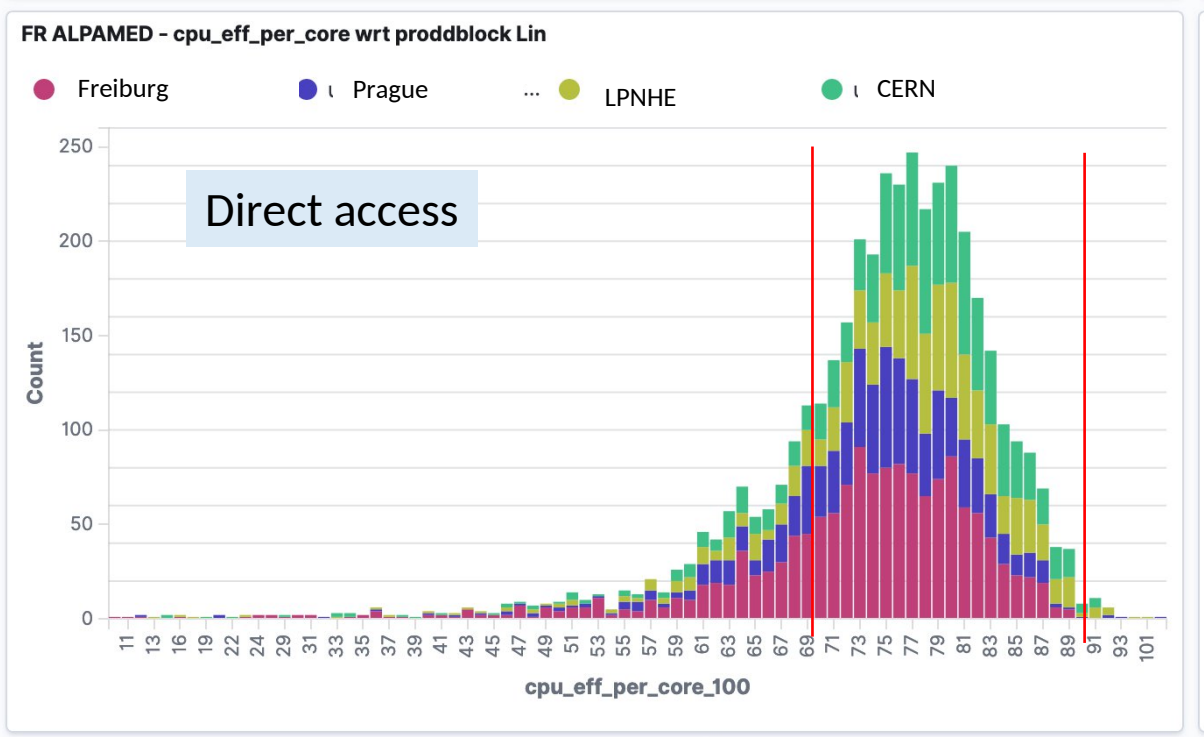




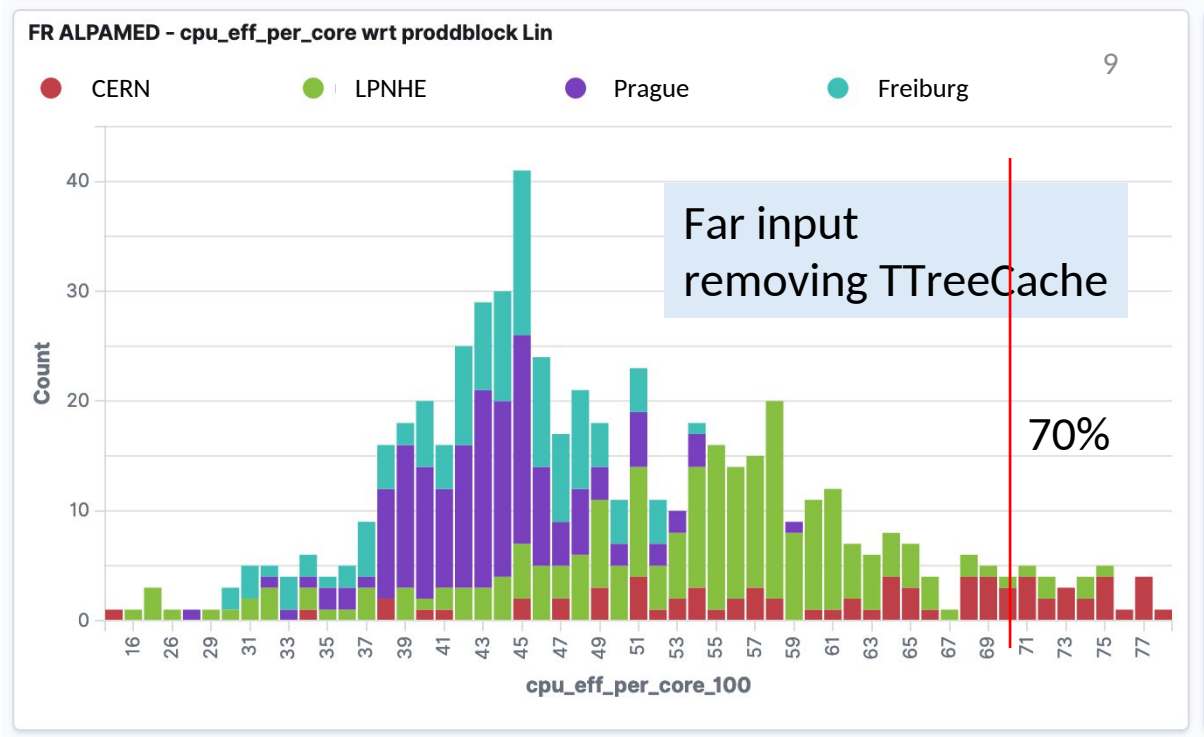
No significant loss of performances (compared to shape width)

CPU eff. : direct access : Far sites

Laboratoire d'Annecy de Physique des Particules

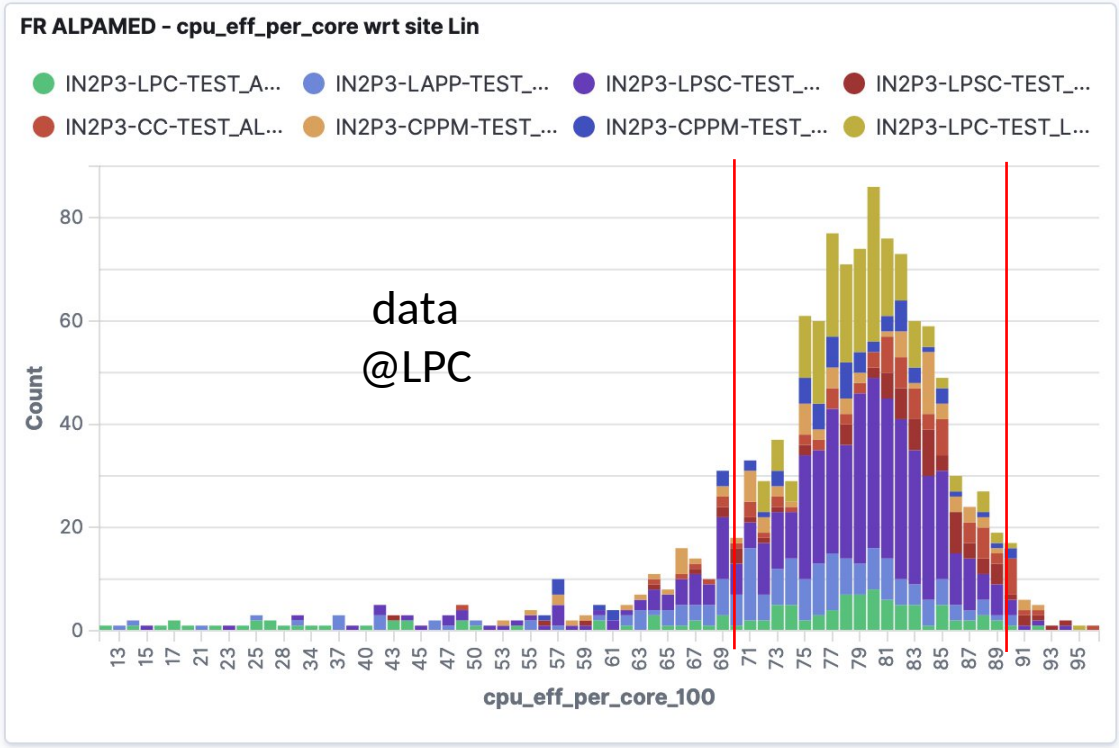


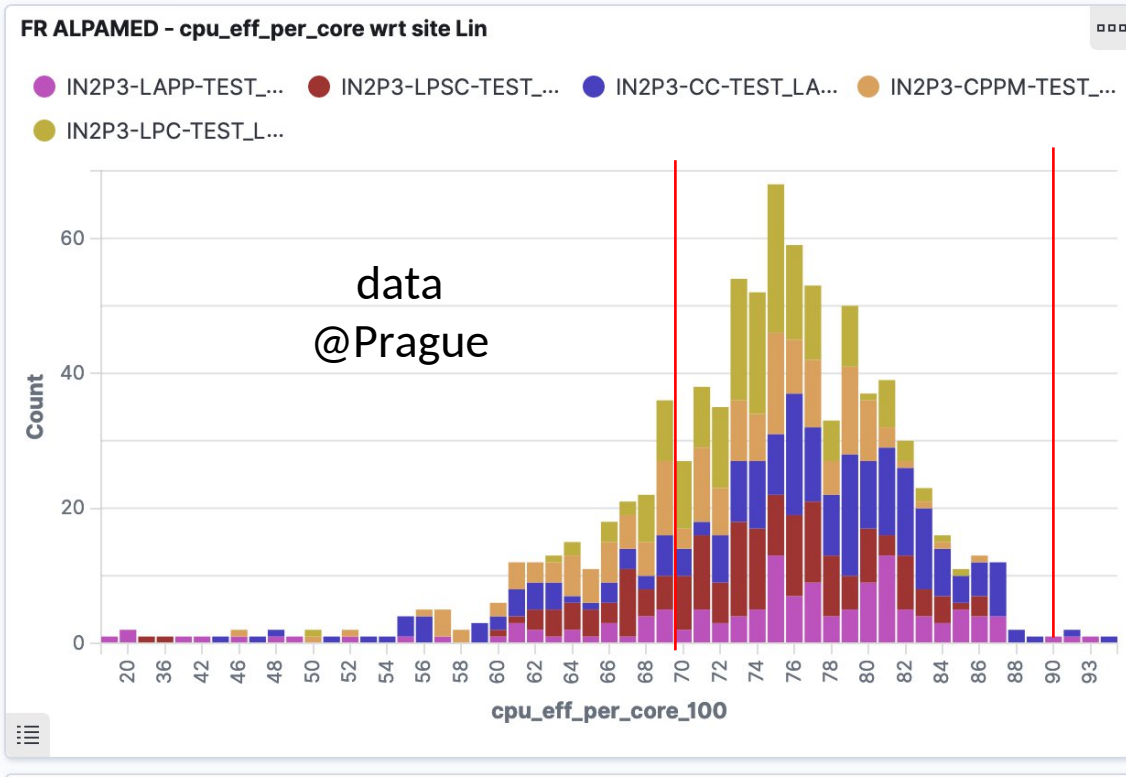
Slight degradation of processing efficiency



TTreeCAche mechanism useful
at European Data Lake level

Remaining latency impact





Small tail : Minimal network layers within NREN ?

Degradation starts being visible

DPM Storage federation Ops

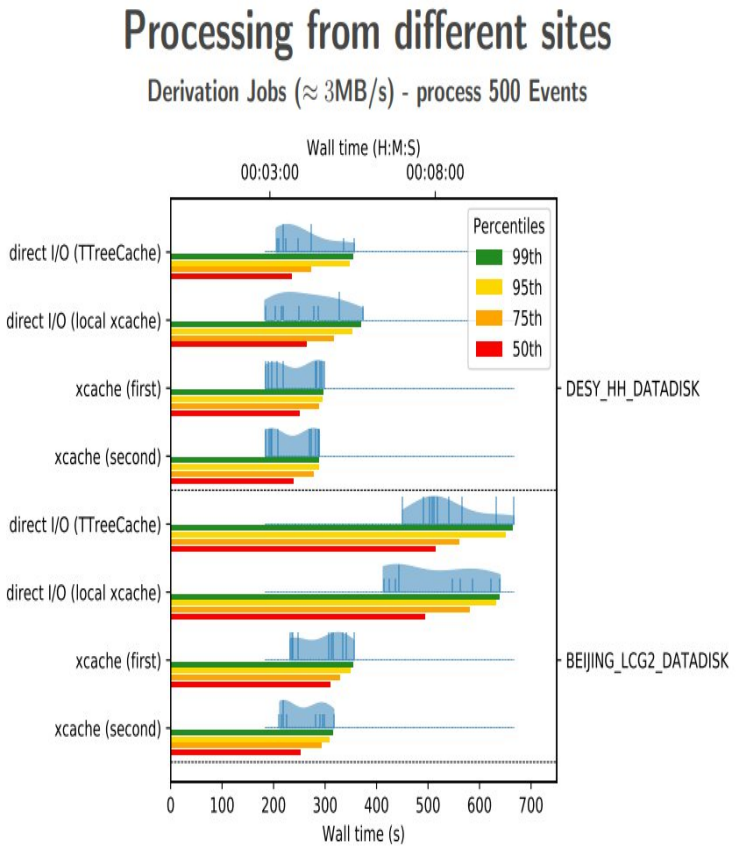
Laboratoire d'Annecy de Physique des Particules

- Team of site admins operating prod storages : First priority
- Creation of test-bed storage federation
  - No major technical issue to deploy
    - Each site admin deploys its disk servers
    - Head-node managed by hosting site

- Feedback from operation
  - Opportunity for Grid site admins to reinforce collaboration
    - Site admins administered others
  - Shared access to disks but dynamic share of responsibilities to be improved --> Ideas for improvement
    - Share responsibilities as function of time between site admins
    - Tool to centrally deploy new versions

Conclusion

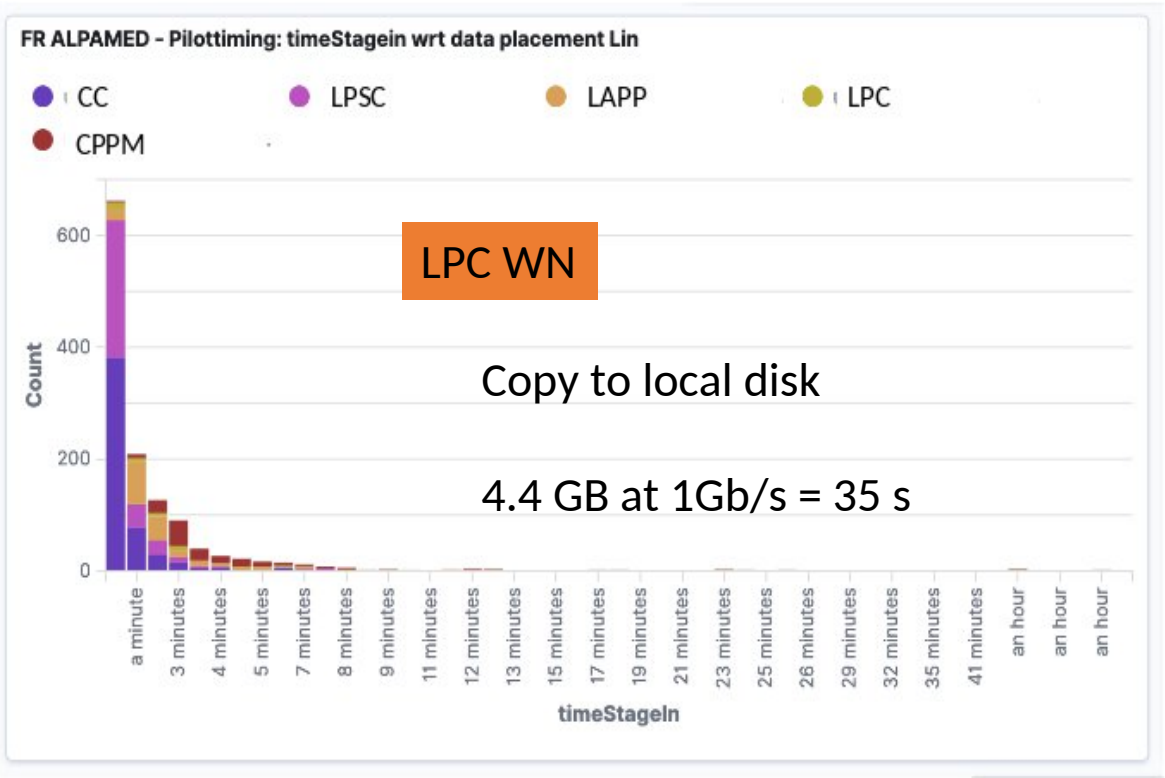
- ATLAS with heavy IO
  - Intelligent root branch filtering : Optimal to read fraction of files
  - Example of efficient recovery for long distance
- Next steps :
  - Add Xcache layer : Benefit from Caching
  - Ensure scalability (1000s jobs) and long term reliability (> 1 hour)
  - Resilience against saturated network
- Operation : Ideas to optimize share of responsibilities

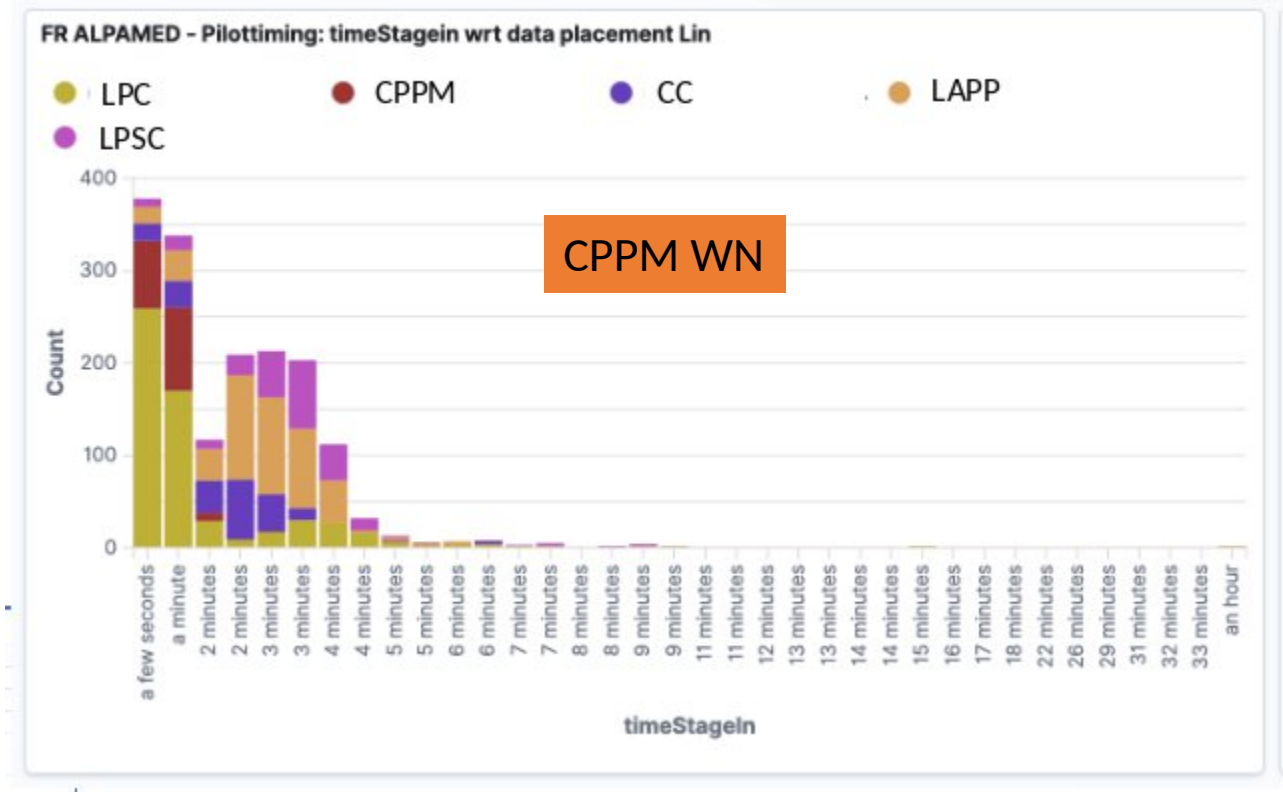


N. Hartmann : Link

Backup

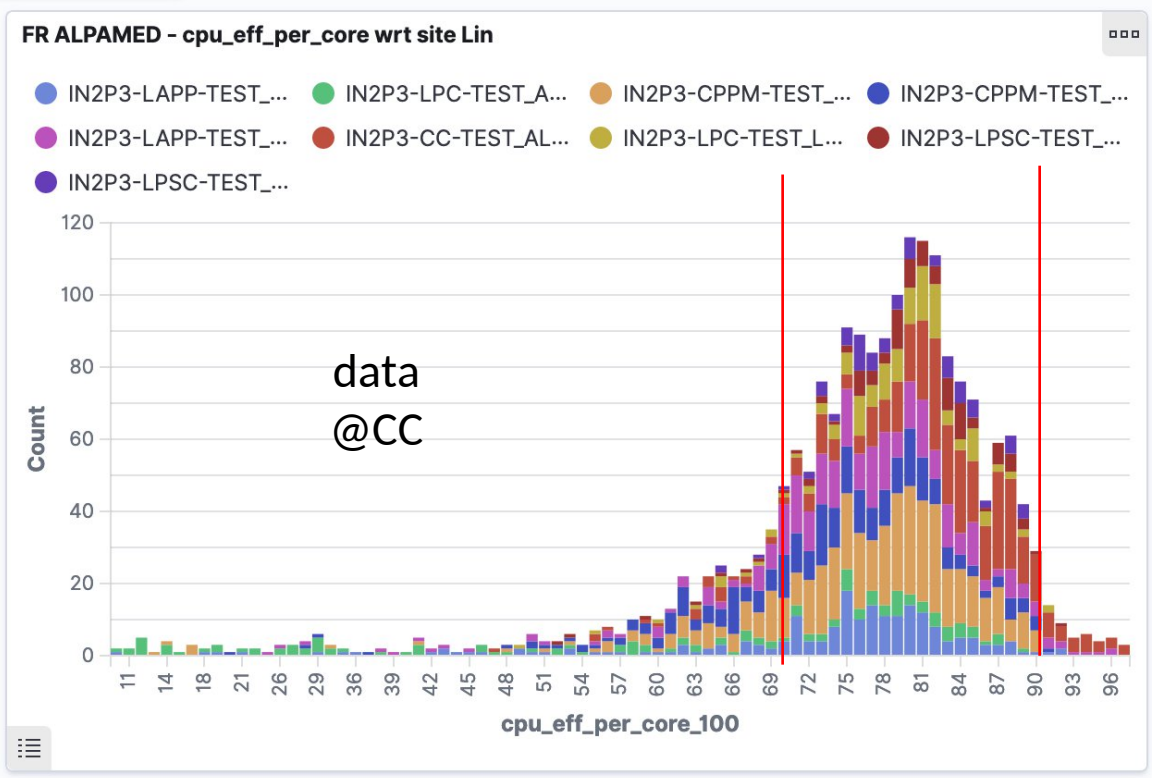
Simply copy remote files

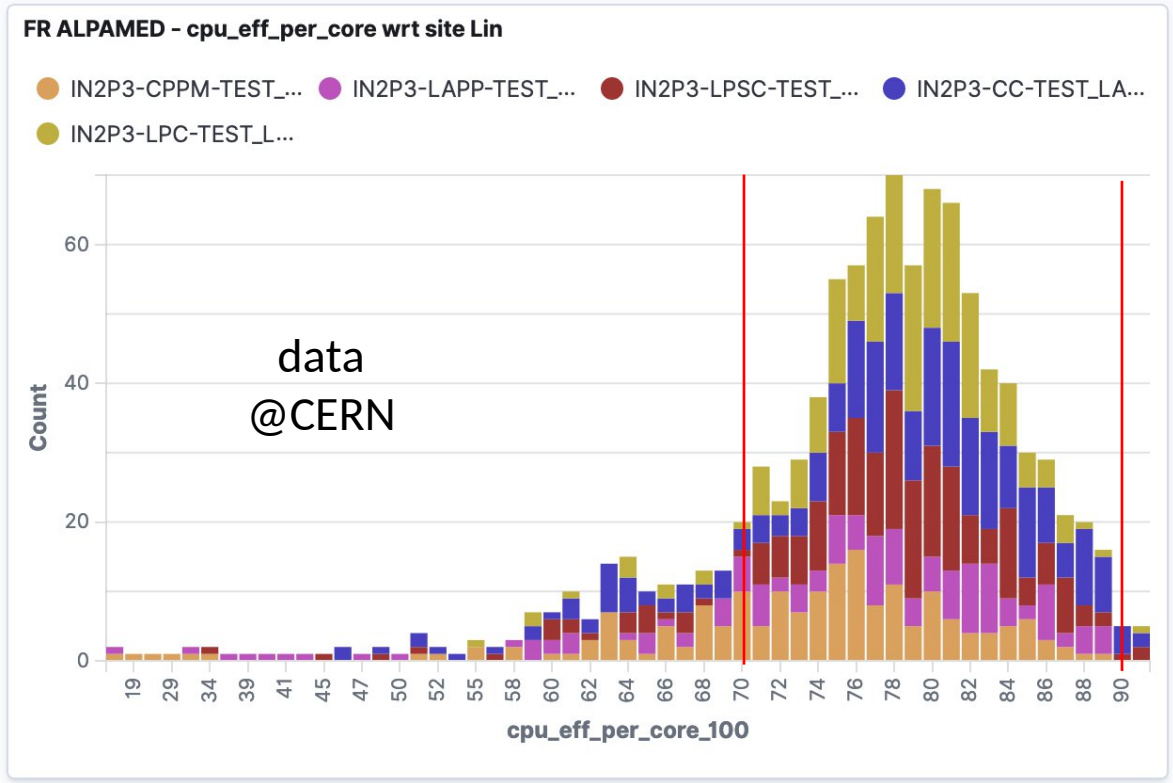




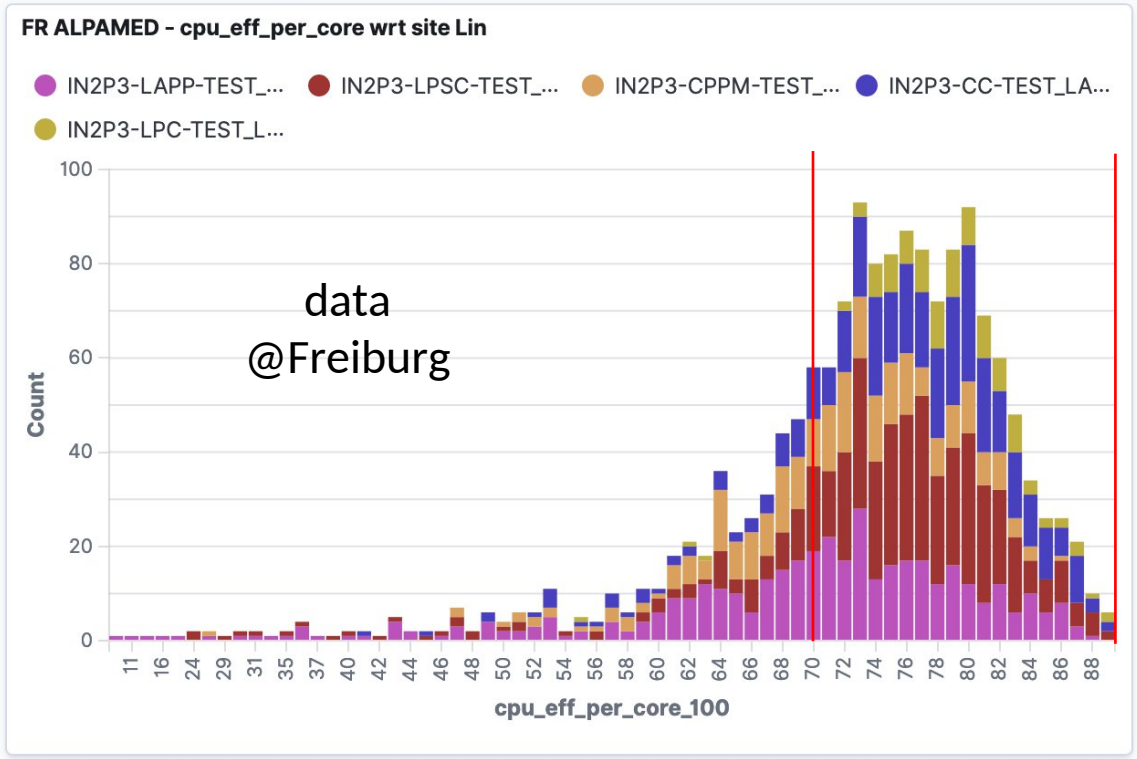
Indication for possible network limitations : CPPM already planed for network improvement

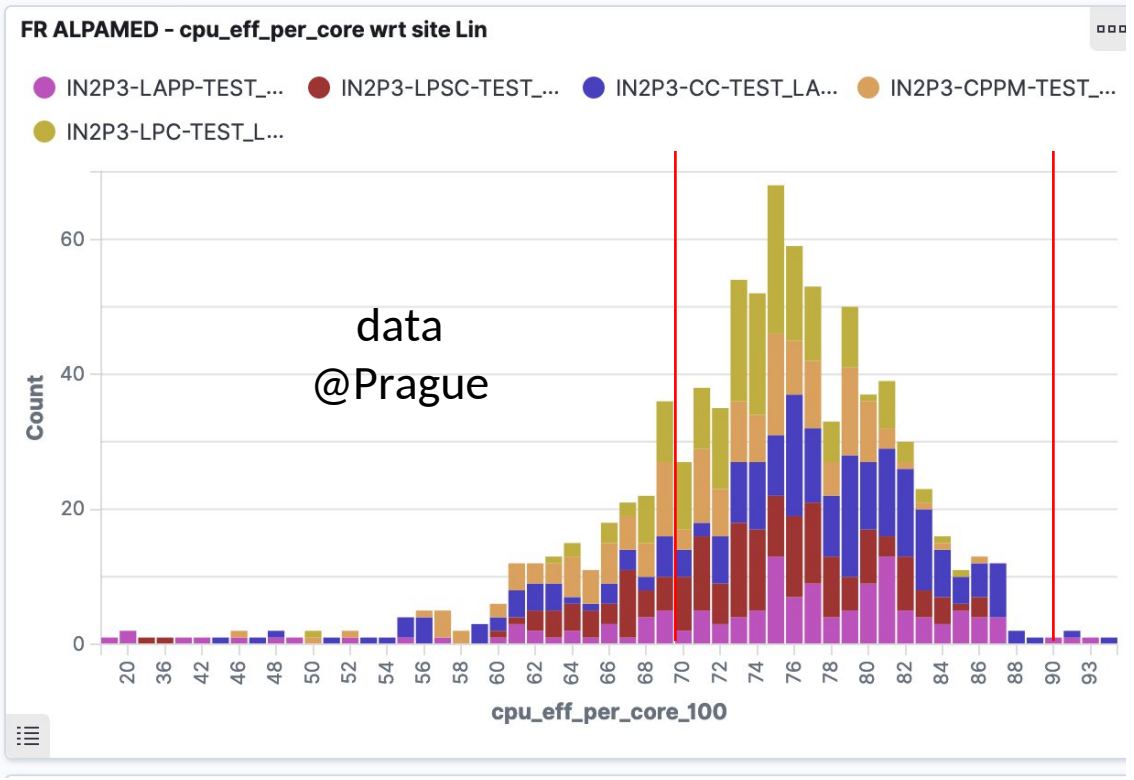
Connectivity stability : Medium distance





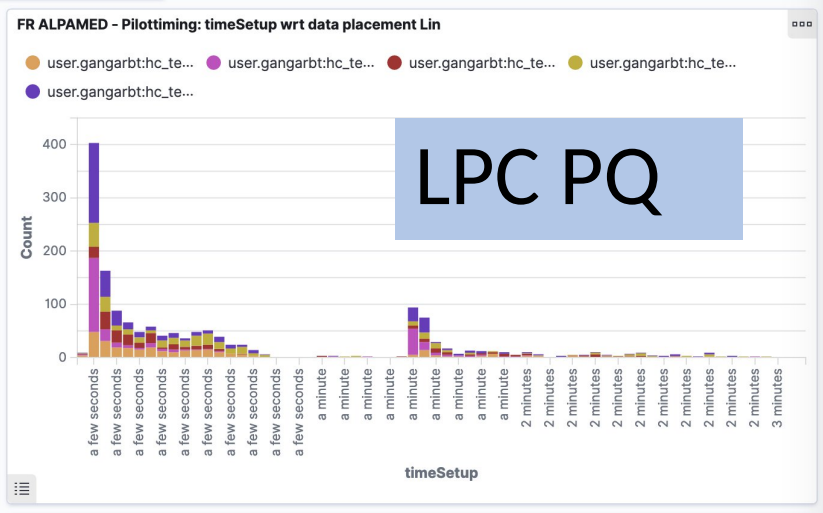
Connectivity stability : Long distance

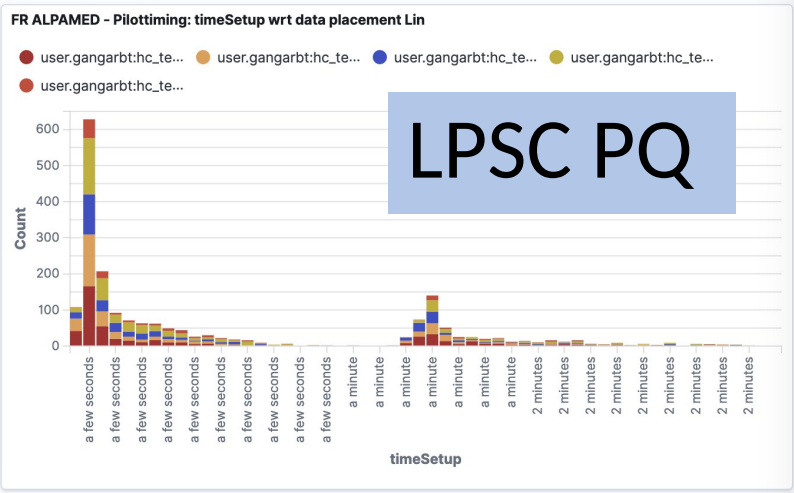


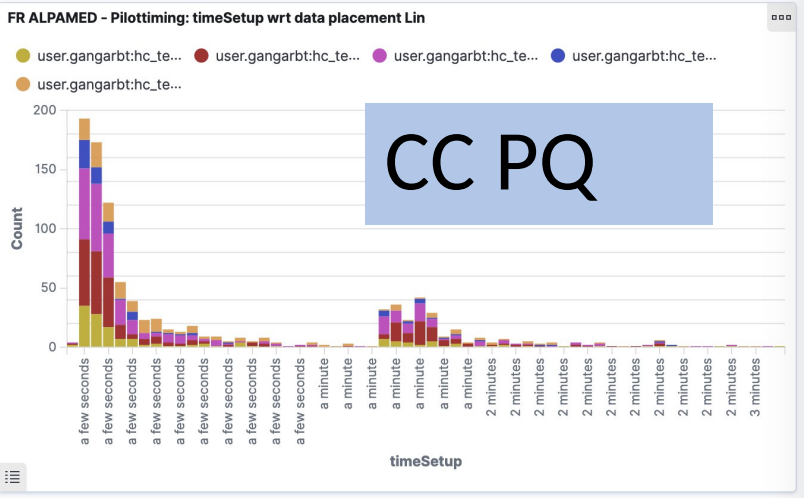


Time to setup software (backup)

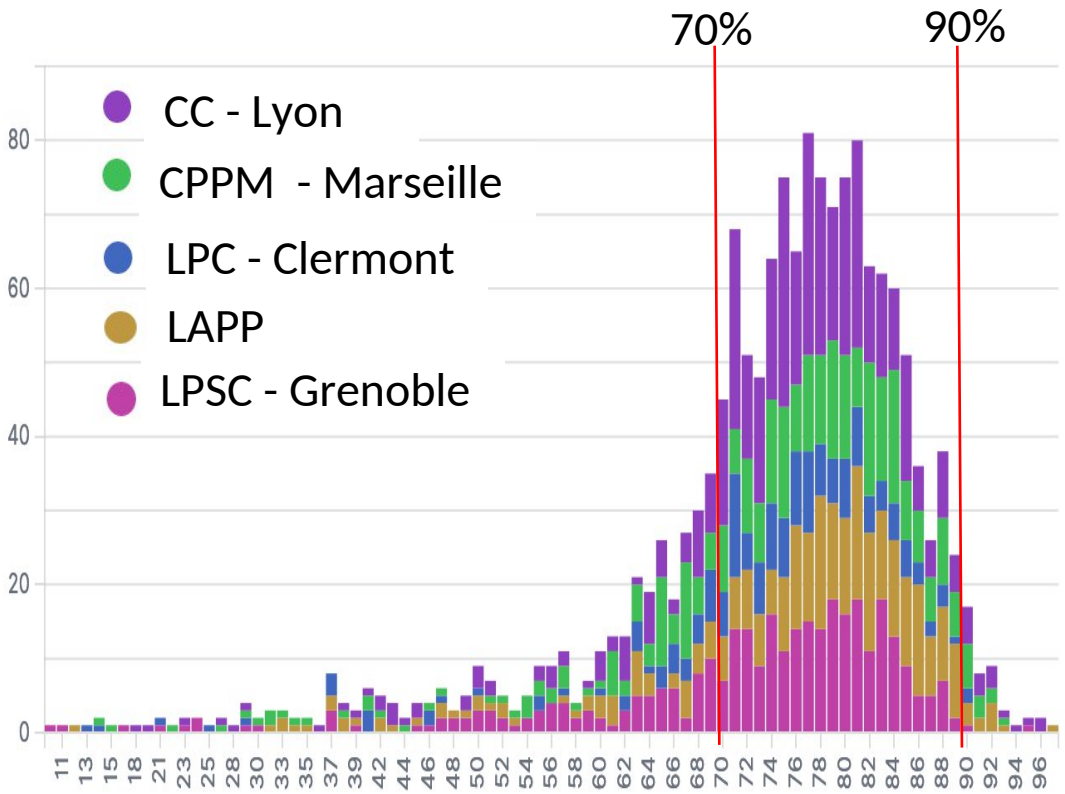
- Still second peak (small impact of global performance)
  - None for CPPM and LAPP
- Renewal of cached files ?

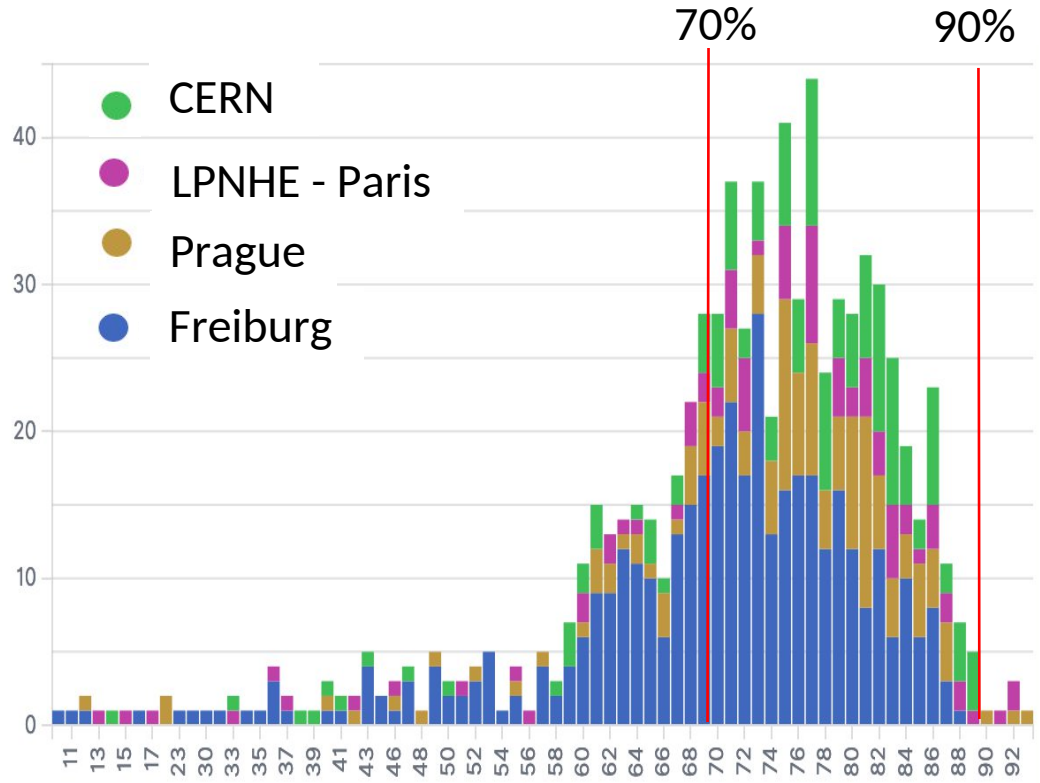




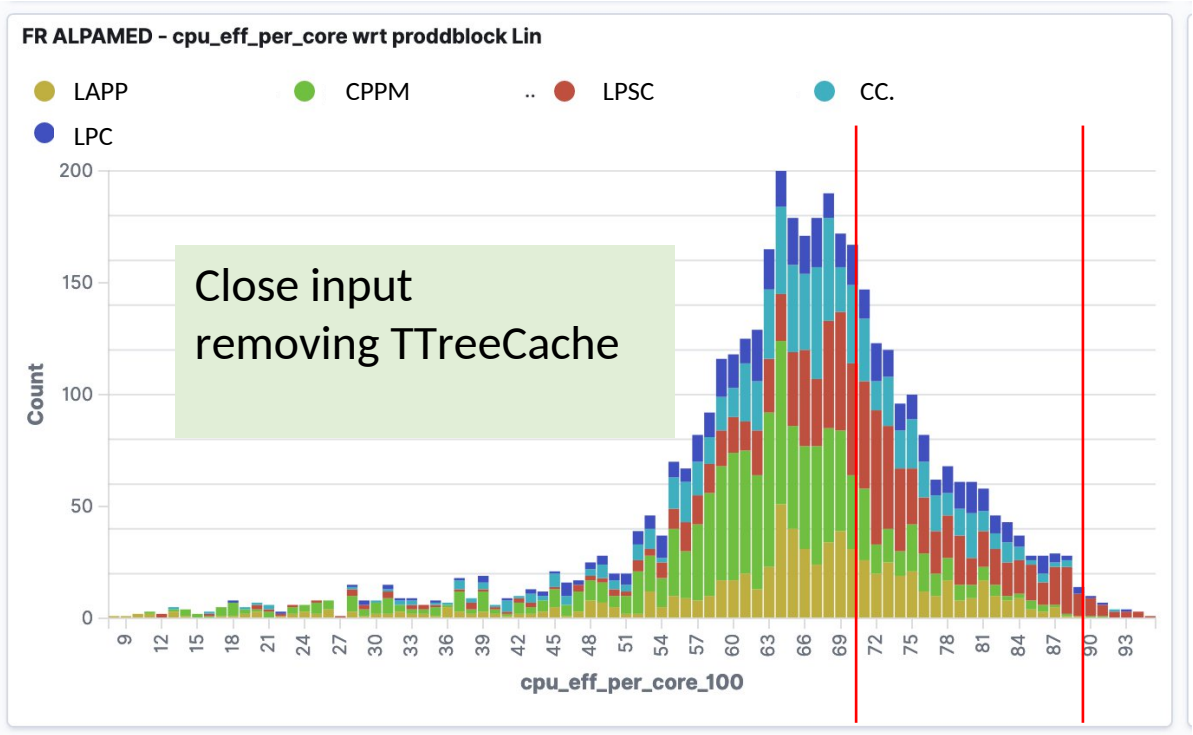


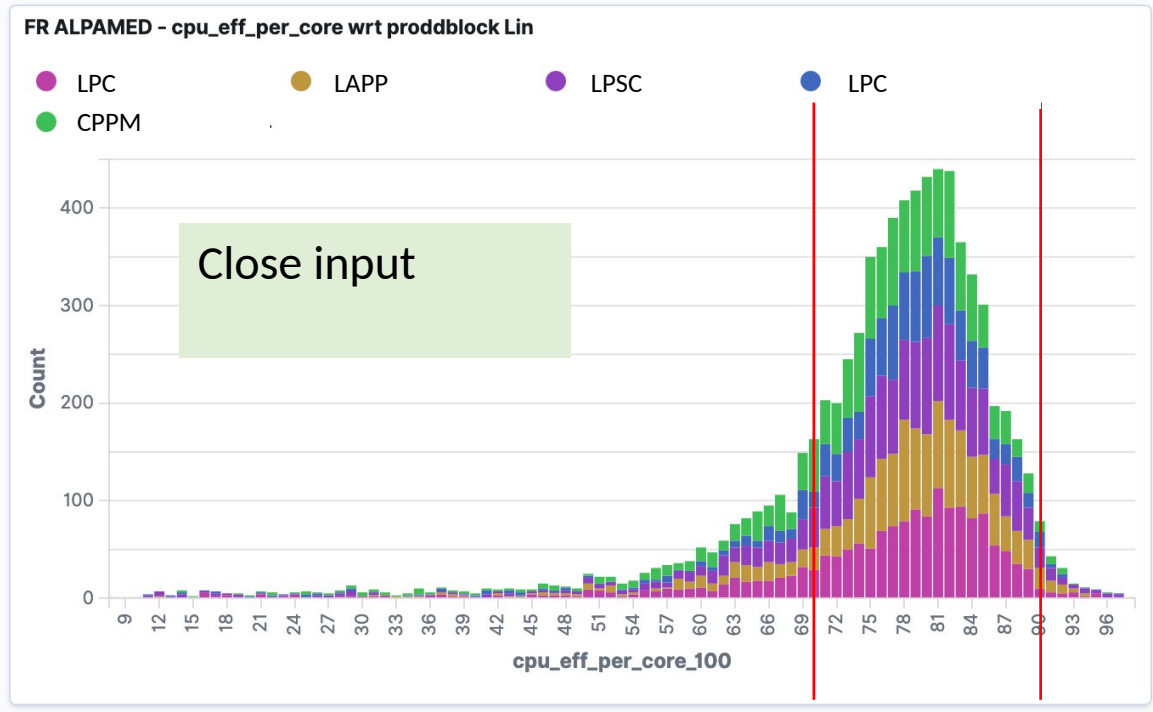
LAPP Panda Queue, direct access to input files





CPU eff. : direct access : Close sites





Significant gain within TTreeCAche already for close sites

19