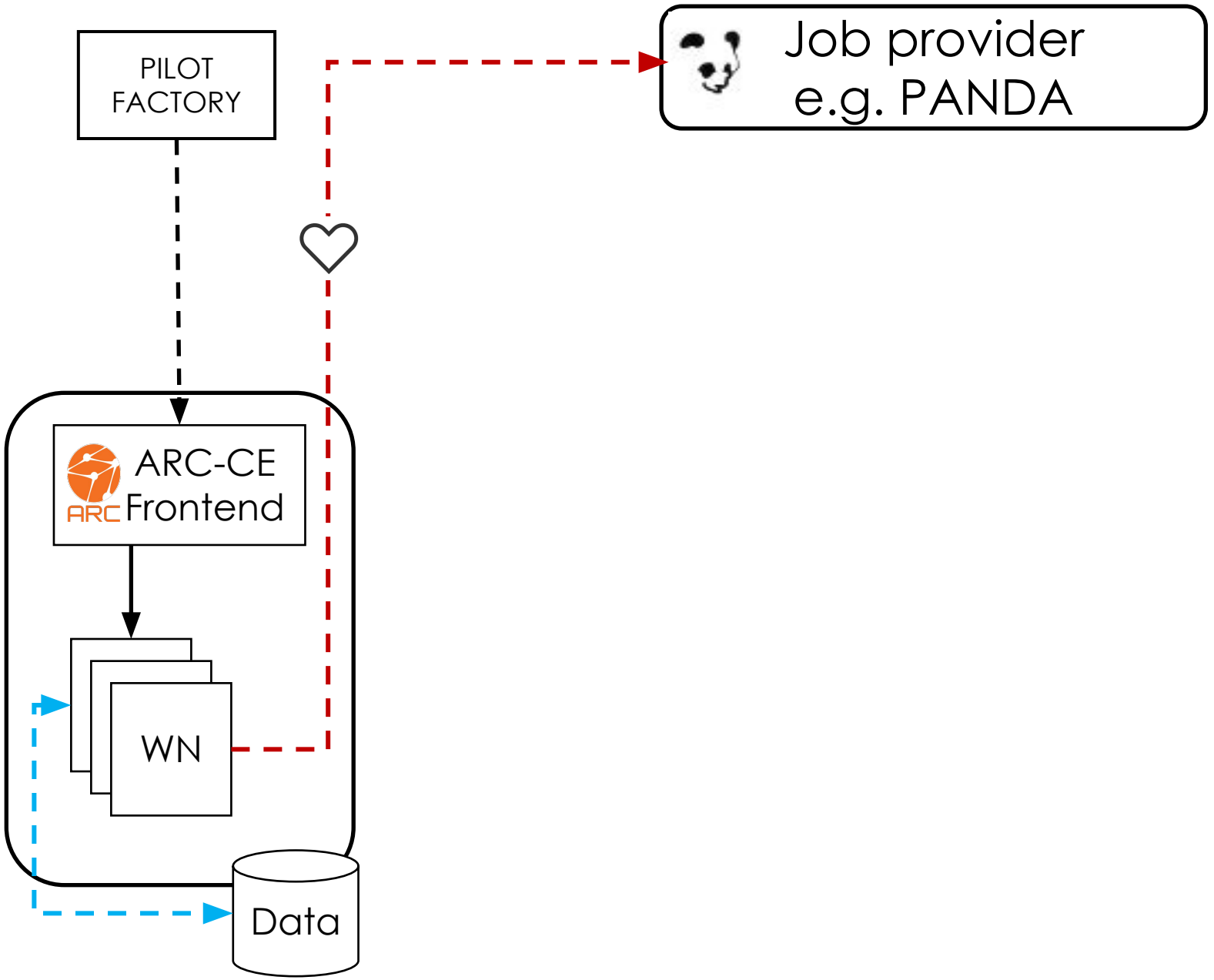
Nordugrid ARC datastaging and cache

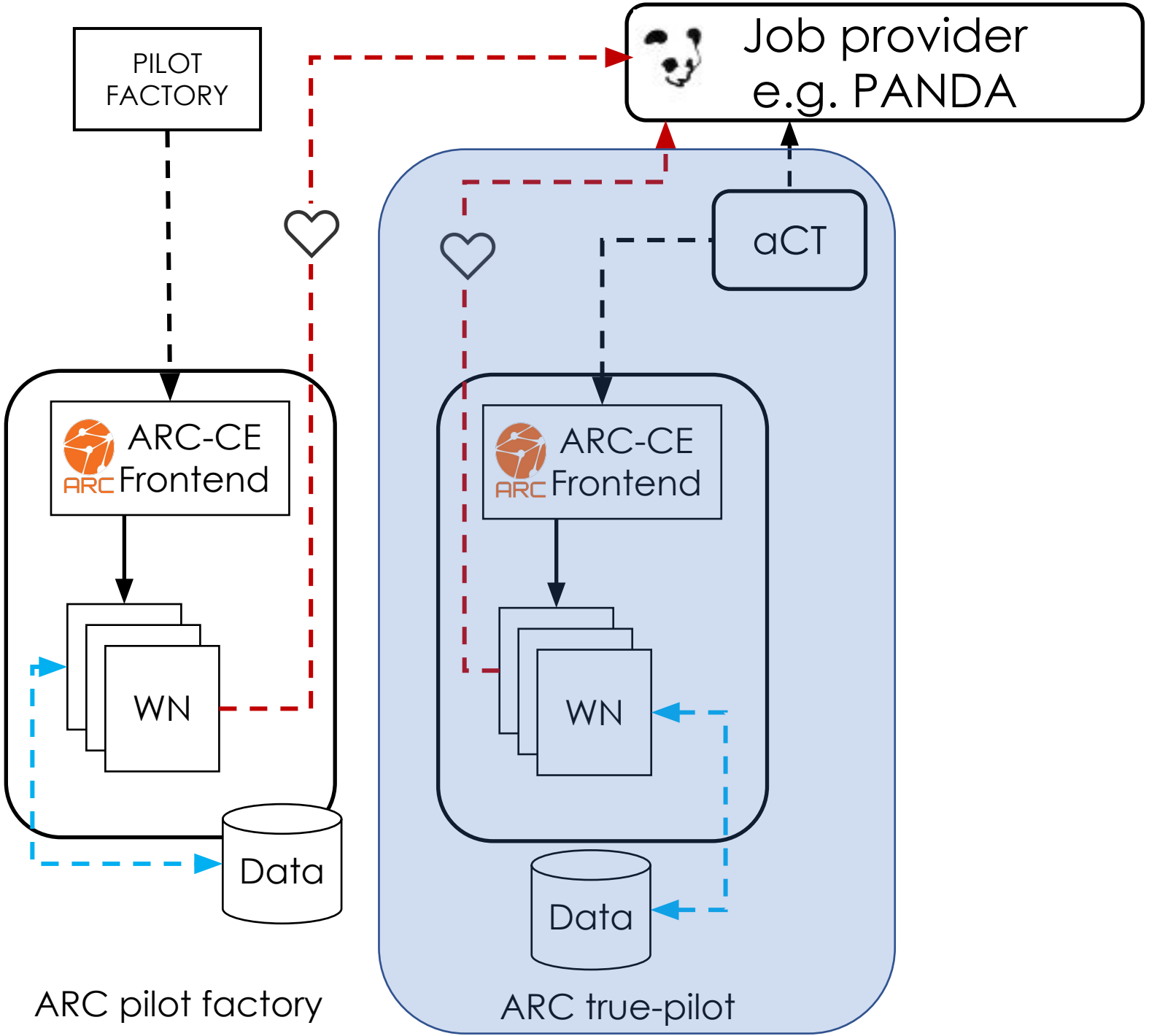
Efficiency gains on HPC and cloud resources

NDGF AHM 2020-2
20.10.2020-23.10.2020
Zoom-meeting

Maiken Pedersen University of Oslo/NeIC
Nordic Tier 1
Nordugrid Collaboration (Balazs Konya)

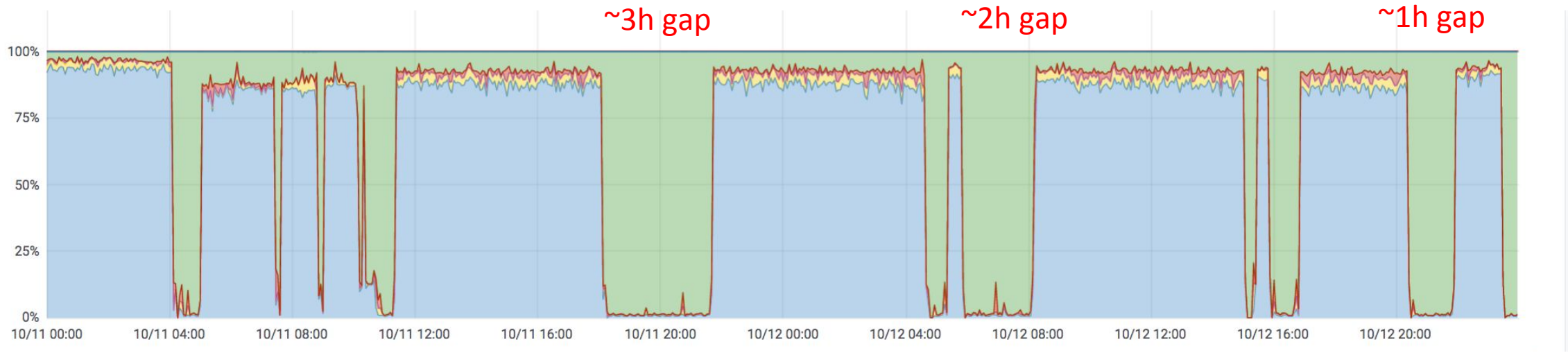


ARC pilot factory

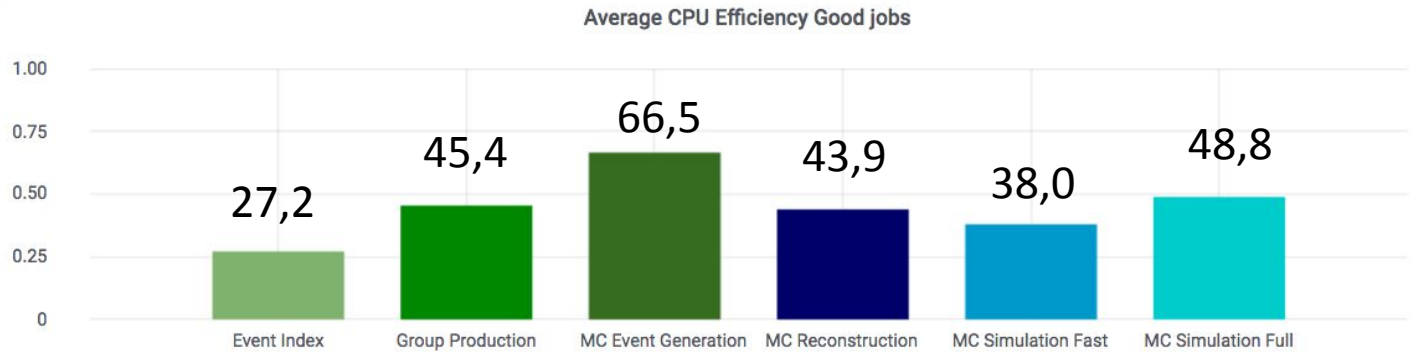

PILOT FACTORY
Job provider e.g. PANDA
aCT
ARC-CE Frontend
WN
Data
ARC-CE Frontend
WN
Data
ARC pilot factory
ARC true-pilot

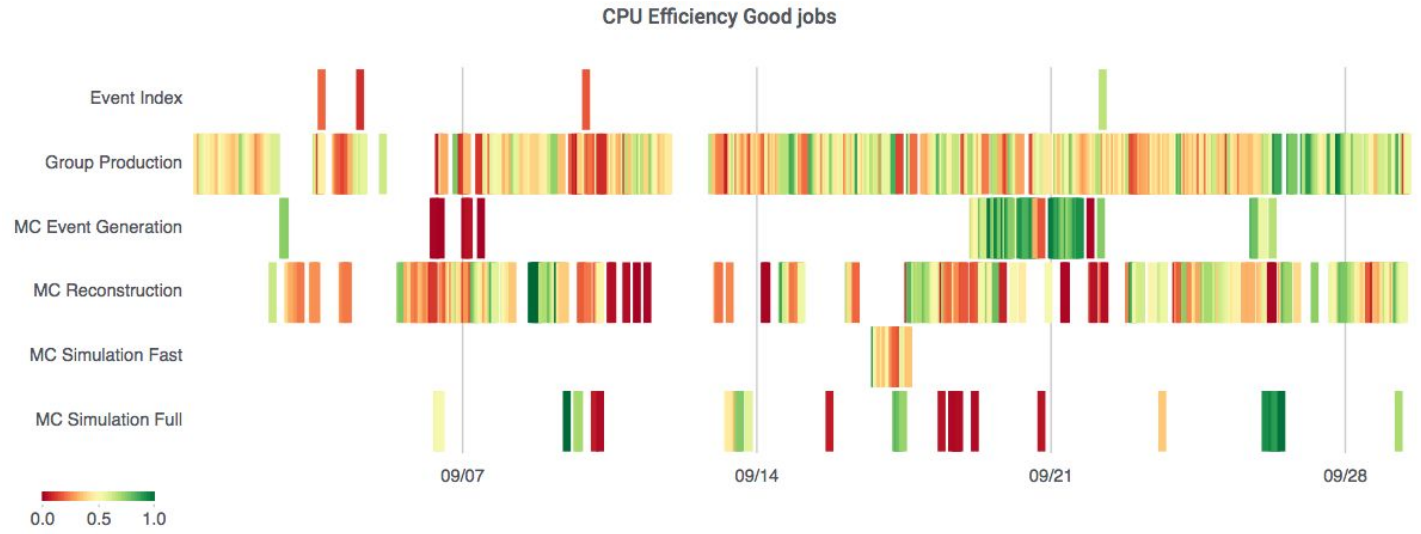
CPU pattern on Openstack cluster ARC true-pilot

- Shows a typical worker node and its CPU pattern
- Green: idle
- Long idle periods on the worker nodes (up to hours) as the pilot collects the needed input files

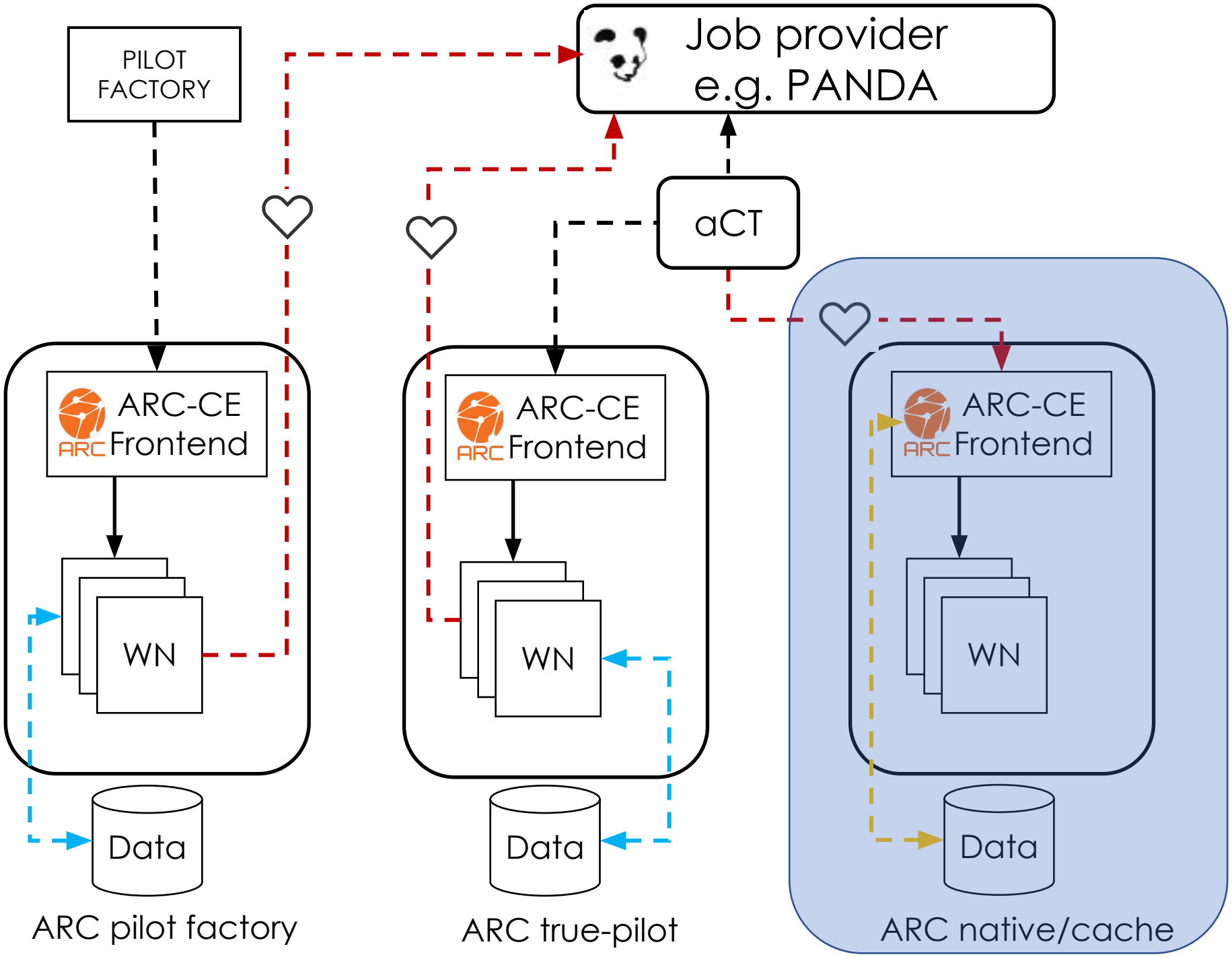


Panda Job Accounting metrics
CPU Efficiency
ARC Pilot mode



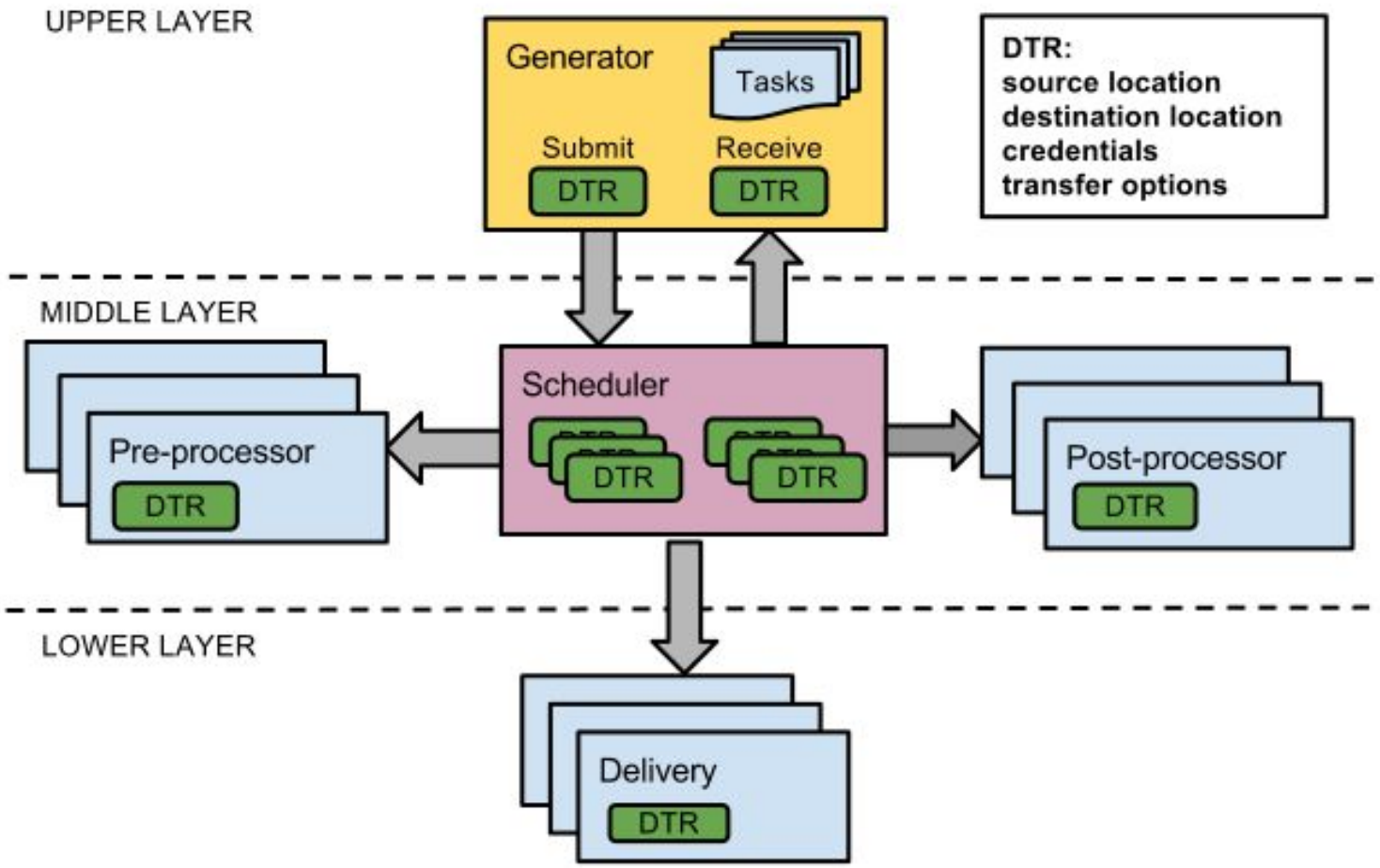


- UIO_CLOUD CPU Efficiency running in pilot mode
- Low efficiency < 50 % for data intensive jobs
- Compute nodes idle while downloading data


PILOT FACTORY
Job provider e.g. PANDA
aCT
ARC-CE Frontend
WN
Data
ARC pilot factory
ARC-CE Frontend
WN
Data
ARC true-pilot
ARC-CE Frontend
WN
Data
ARC native/cache

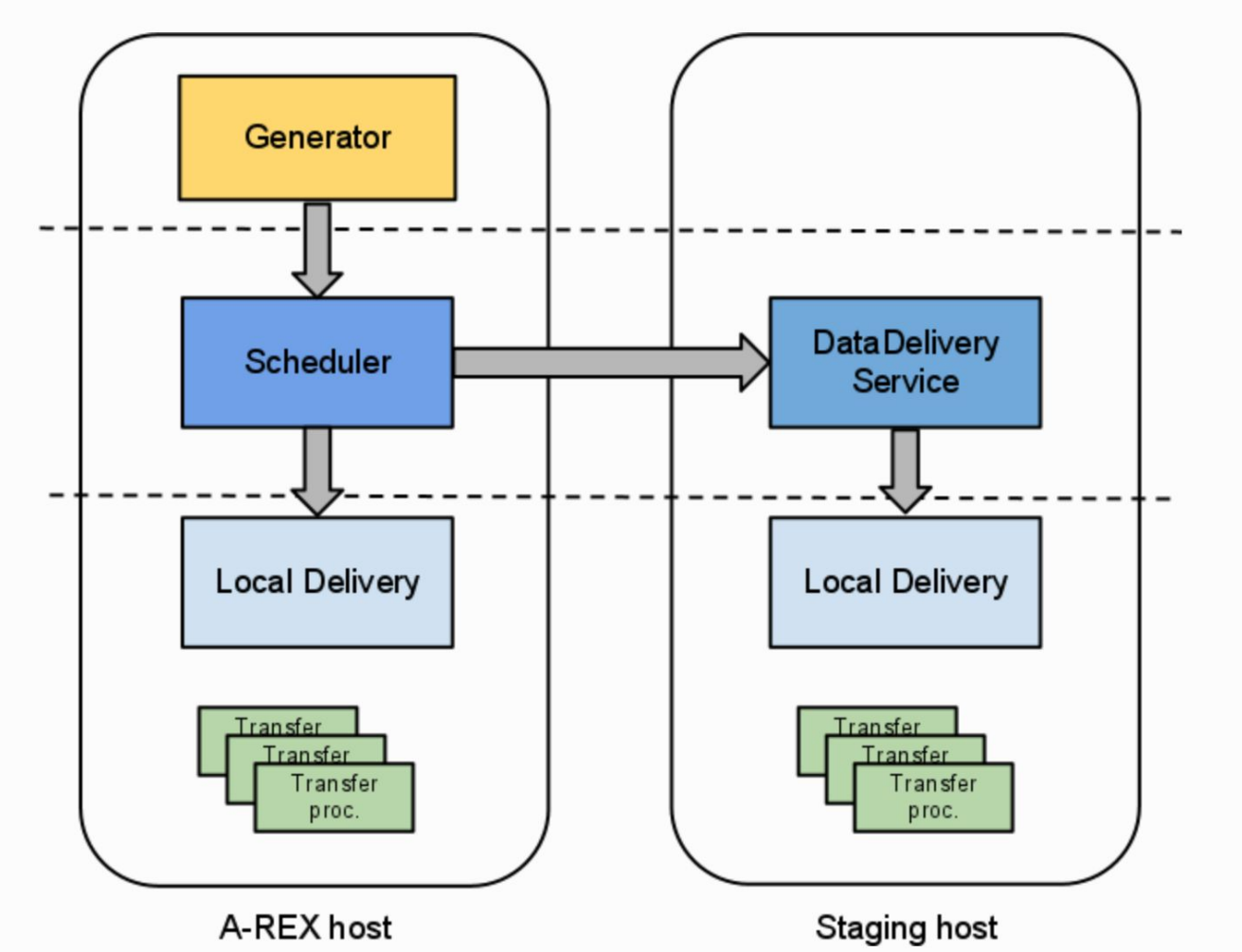
ARC does data staging – ARC Data Transfer (DTR)

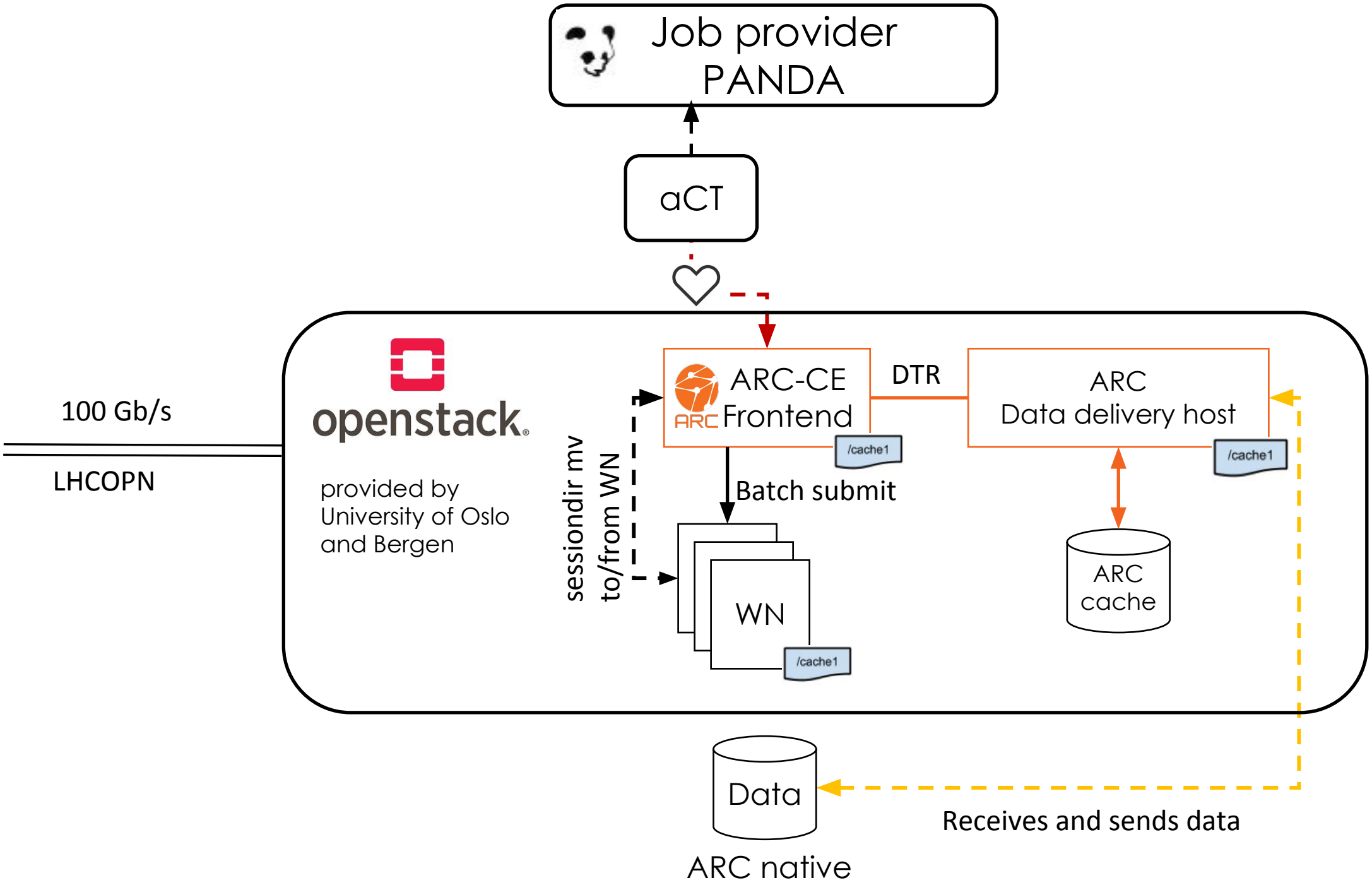
- ARC sends the job to the underlying batch system **only once all input files are staged**
- Improves the CPU efficiency considerably compared to running in pilot mode
- ARC cache for data-recycling
  - Allows caching of frequently used files
  - Minimizes bandwidth
- ARC Data Transfer framework: Generator, Scheduler, pre- and post-processor and Delivery.
  - Pre-processor for cached files
  - Delivery for remote transfer



Multiple data staging hosts

- To distribute the load on the main ARC-CE one or more remote data delivery service hosts can be set up
- All the logic is handled by the A-REX host
- The delivery hosts do just that – deliver the data requested by the DTRs it receives

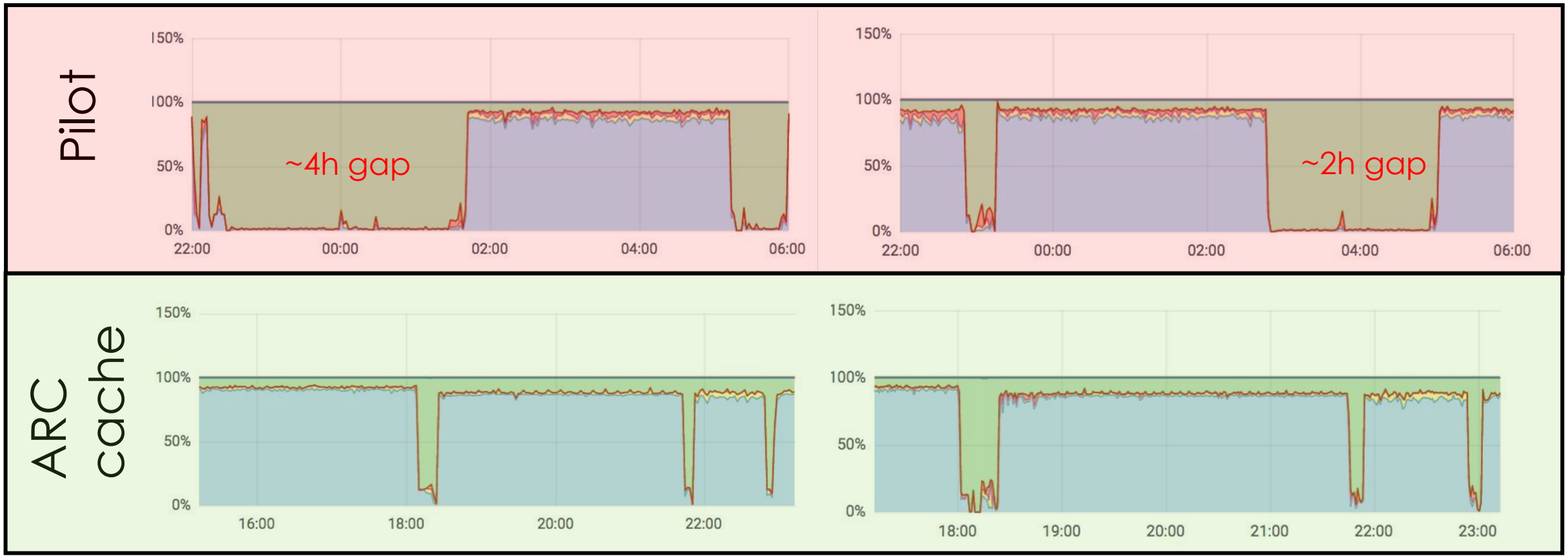



Job provider
PANDA
aCT
100 Gb/s
LHCOPN
openstack
provided by
University of Oslo
and Bergen
ARC-CE
Frontend
DTR
ARC
Data delivery host
/cache1
/cache1
Batch submit
sessiondir mv
to/from WN
WN
/cache1
ARC
cache
Data
Receives and sends data
ARC native

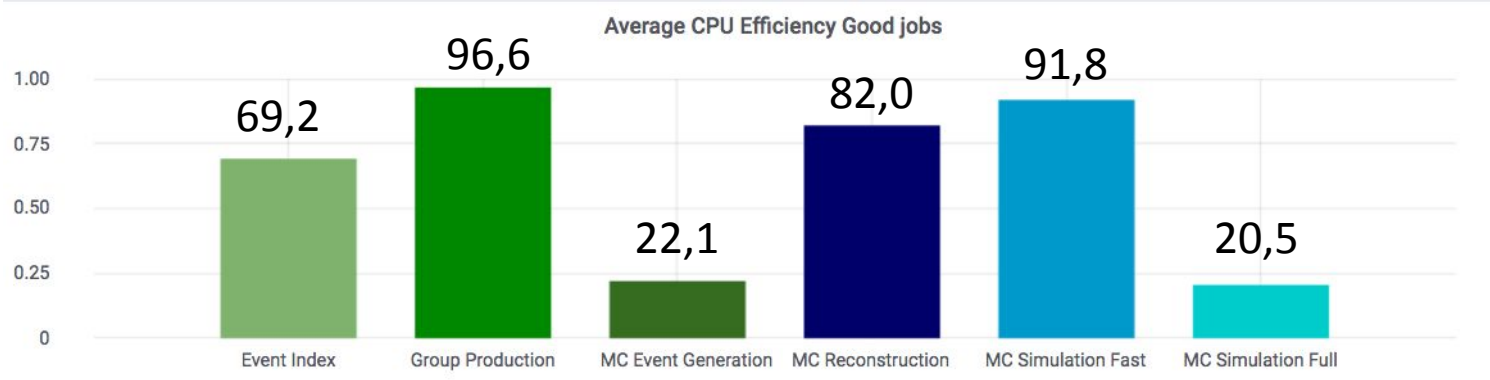
Using ARC datadelivery service and ARC cache – "NDGF mode"

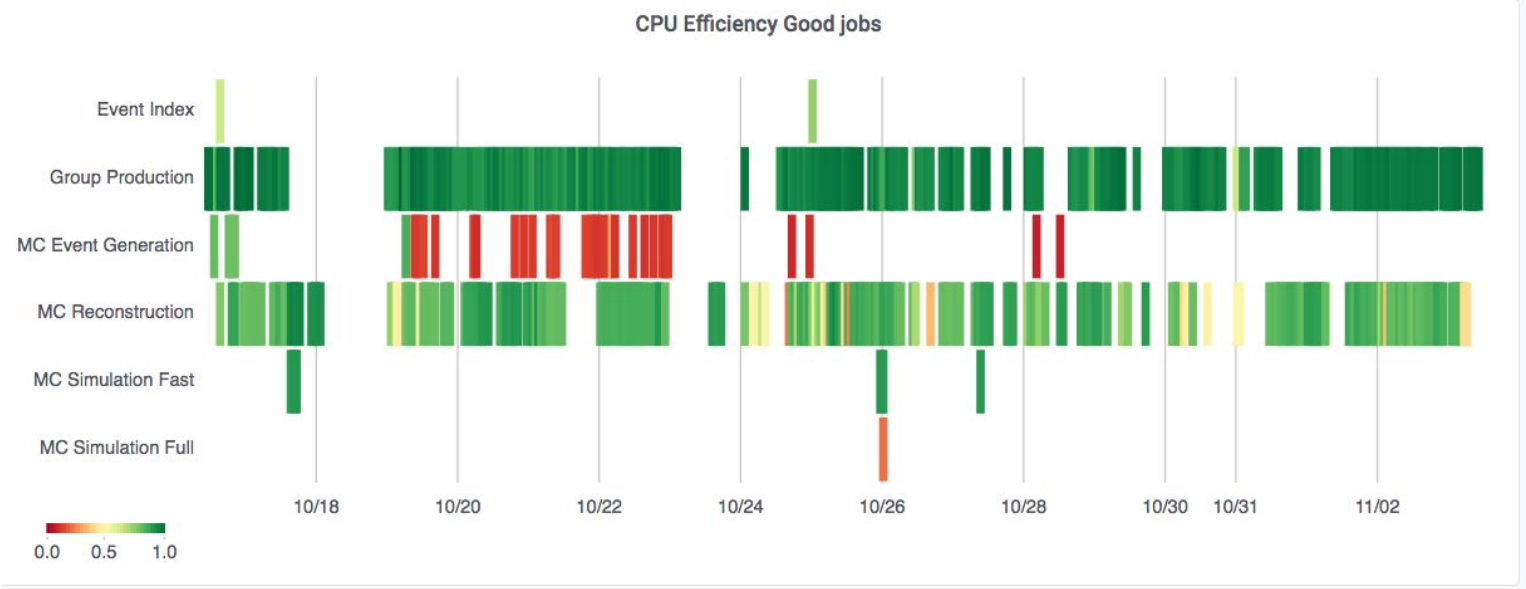
- Compute nodes idle-time greatly reduced
- Jobs run at expected and optimal CPU utilisation (unlike with ATLAS@home running in parallell)

 CPU efficiency increases substantially



Panda Job Accounting metrics
CPU Efficiency
ARC cache mode





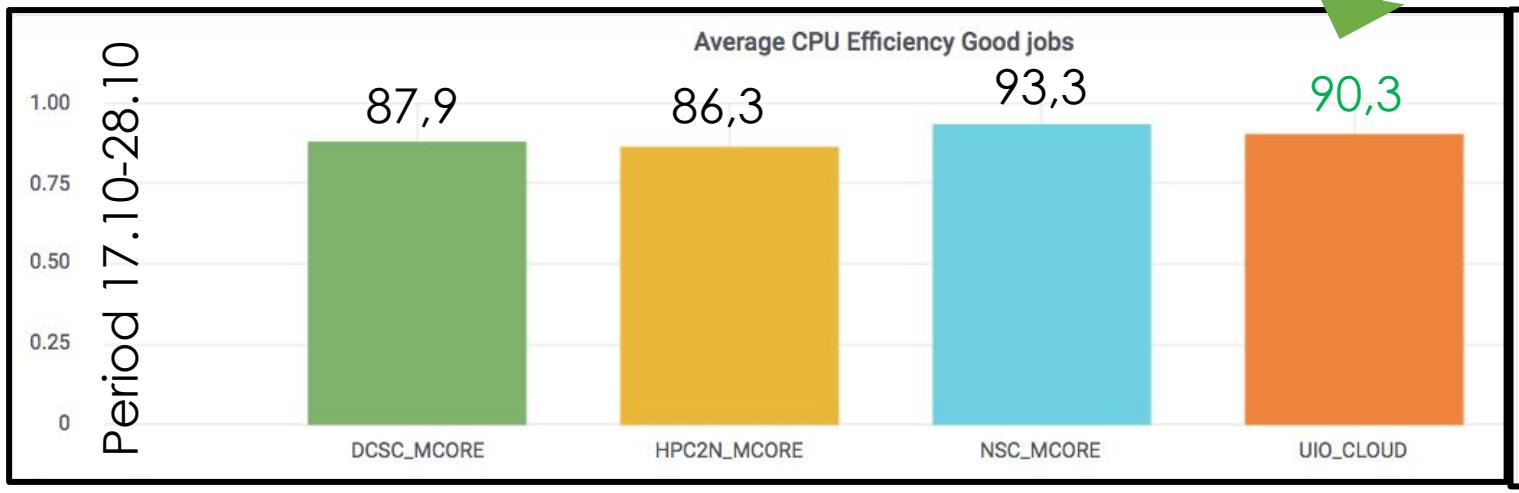
- UIO_CLOUD CPU Efficiency running in NDGF ARC datastaging mode
- High efficiency > 80 % compare to <50% in pilot mode
- Only showing a selection of jobs since other types hardly ran in this period
- With ARC doing datastaging, practically 0 CPU idle time on the worker nodes

Sites using the ARC datastaging increase their CPU efficiency compared to running in pilot mode

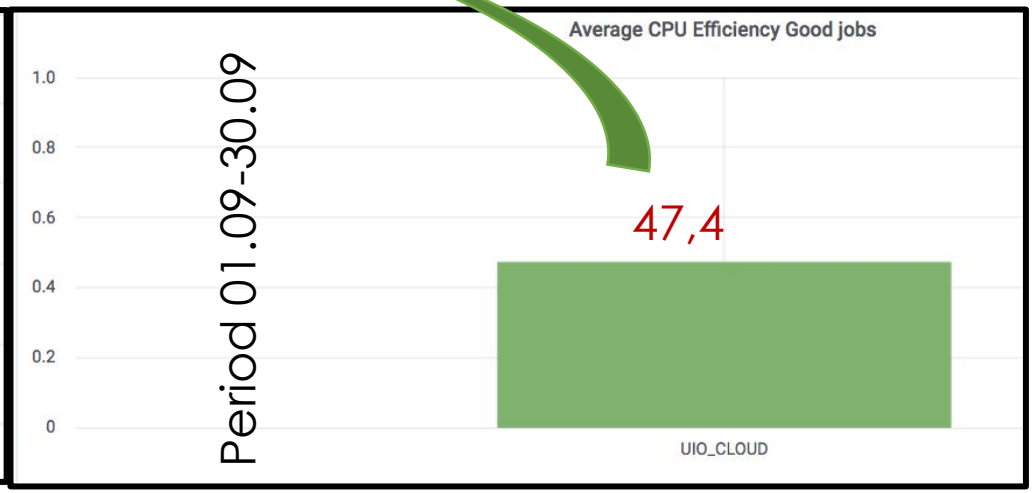
- CPU efficiency for all jobs combined for Nordic Tier 1
  - (not Analysis as UIO_CLOUD currently does not receive those)
- Left: CPU efficiency for NDGF sites – all in NDGF mode (ARC cache and datastaging)

- Right: UIO_CLOUD running in pilot mode

Increase from 47,4% to 90,3 %



NDGF ARC cache MODE – NDGF sites



UIO_CLOUD Pilot mode

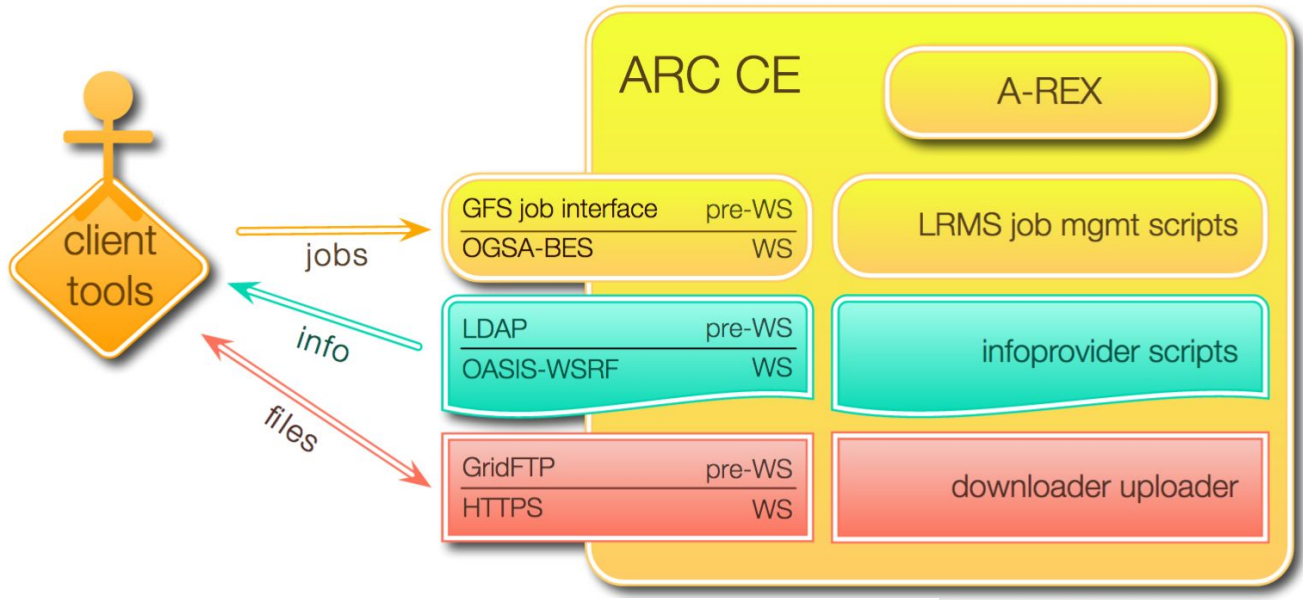
Relevant links

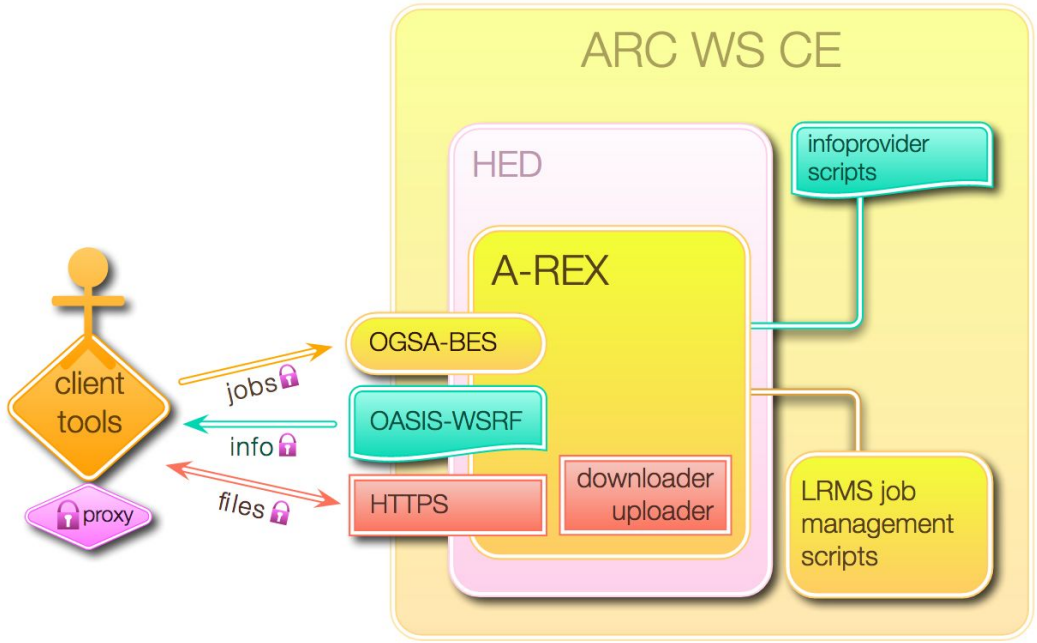
- Documentation and information ARC: http://www.nordugrid.org/arc/arc6/
  - ARC 6 released summer 2019 – much improved sys-admin user experience
- ARC 6 on GitLab: http://www.nordugrid.org/
- Nordugrid collaboration: https://source.coderefinery.org/nordugrid/arc
- NeIC: https://neic.no/, https://neic.no/nt1/
- University Center for Information Technology: https://www.usit.uio.no/english/
- http://www.nordugrid.org/arc/arc6/tech/data/arex_cache.html
- http://www.nordugrid.org/arc/arc6/tech/data/datastaging.html

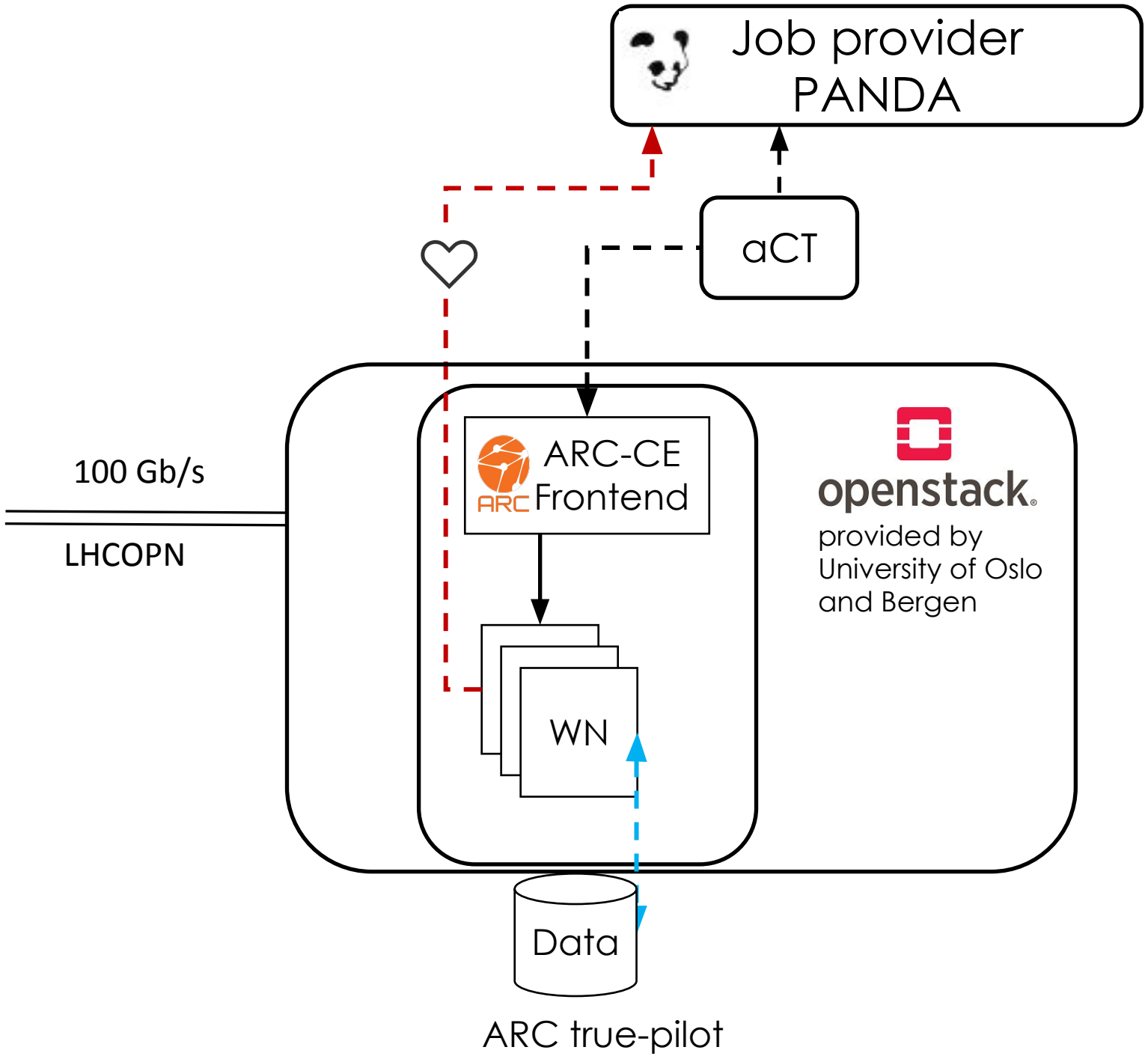
Removed for NDGF AHM

ARC overview

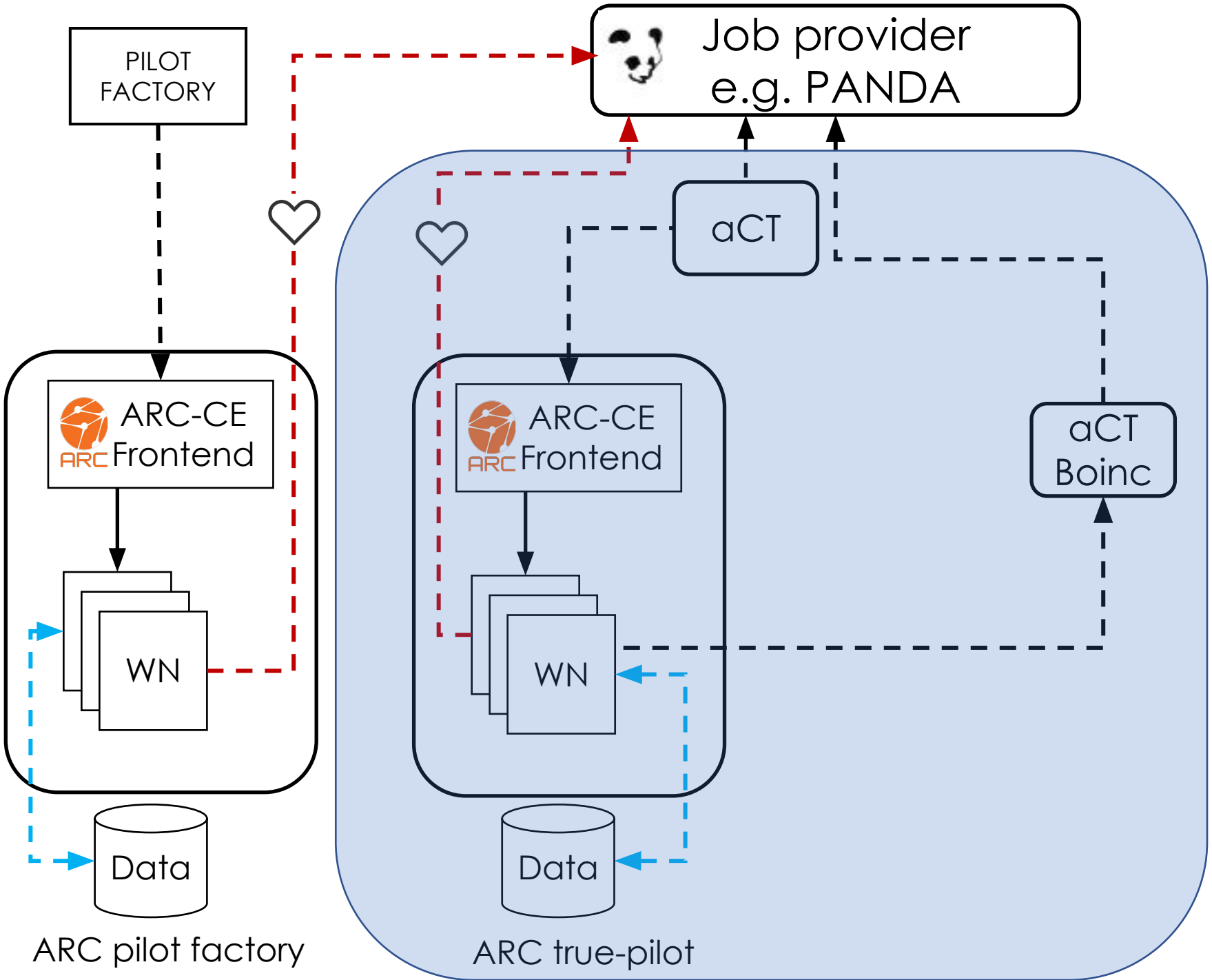
- ARC: middleware to enable computing grids
- Used since 2002, and currently ~180 sites worldwide use ARC
- Job submission
  - ARC Gridftp server, web-service (OGSA-BES)
- Information exchange
  - Ldap/bdii, web-service (OASIS)
- File access
  - Gridftp, https
- Web-services provided by A-REX (no separate service)
- An ARC-CE can provide both interfaces in parallell





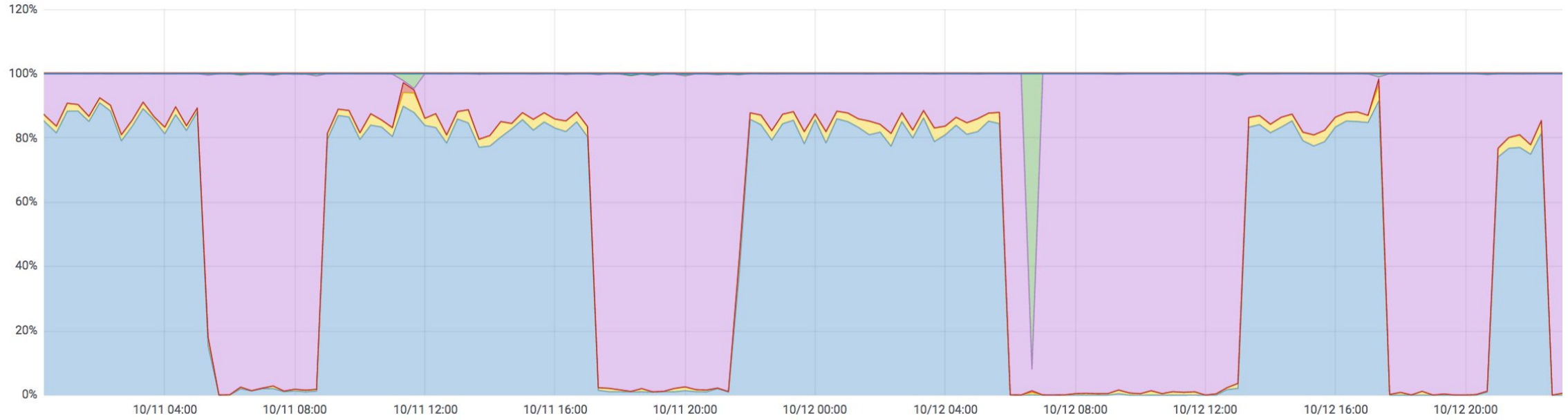


- ATLAS jobs to OpenStack
- 100 Gb/s link to LHCOPN
- 10 Gb/s link within OpenStack
  - will be increased to 25 Gb/s
- /scratch disk on the WN for the jobdirectory
  - CEPH block devices
- Data fetched remotely by the worker nodes


PILOT FACTORY
Job provider e.g. PANDA
aCT
ARC-CE Frontend
WN
Data
ARC pilot factory
ARC-CE Frontend
WN
Data
ARC true-pilot
aCT Boinc

One solution: fill up with opportunistic job slots ARC true-pilot & ATLAS@home Boinc

- The idle periods are now instead filled up by ATLAS@home jobs (simulation)
- Purple: Nice/Steal – ATLAS@home boinc jobs
- No idle nodes
- Cost: slightly less cpu utilisation for normal job – typically around 85-90 % compared to around 90-95% w/o boinc.



Conclusions/summary

- ARC is flexible and can serve sites running traditional WLCG pilots, on traditional grid facilities, as well as HPC and cloud.
- ARC's power is in the data staging and cache
- Sites using the ARC data staging increase their CPU efficiency compared to running in pilot mode

- With the increasing data intensity from the LHC, efficient use of compute power will just become more important