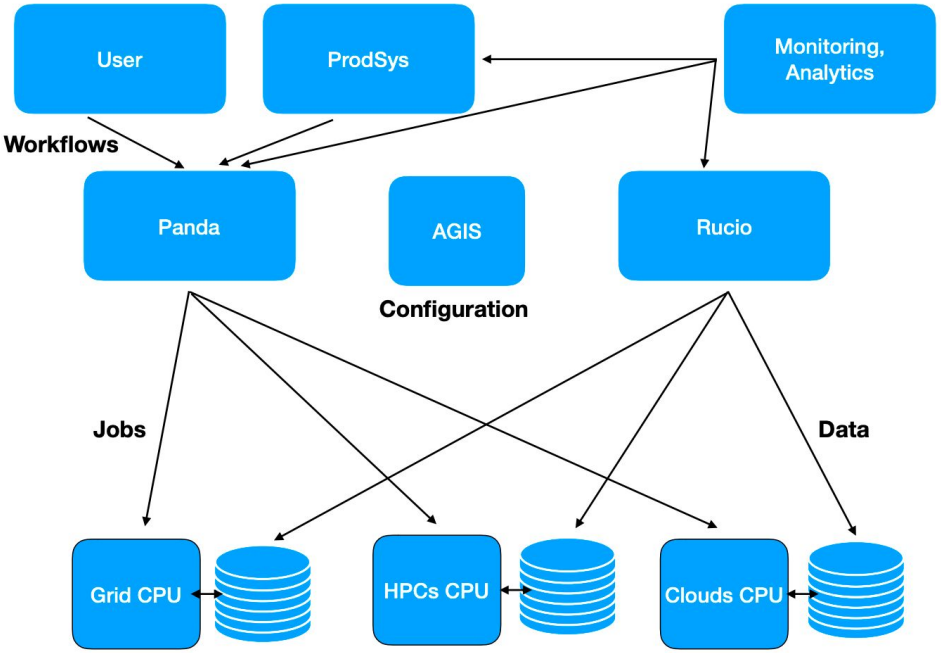
ATLAS HPC Perspective

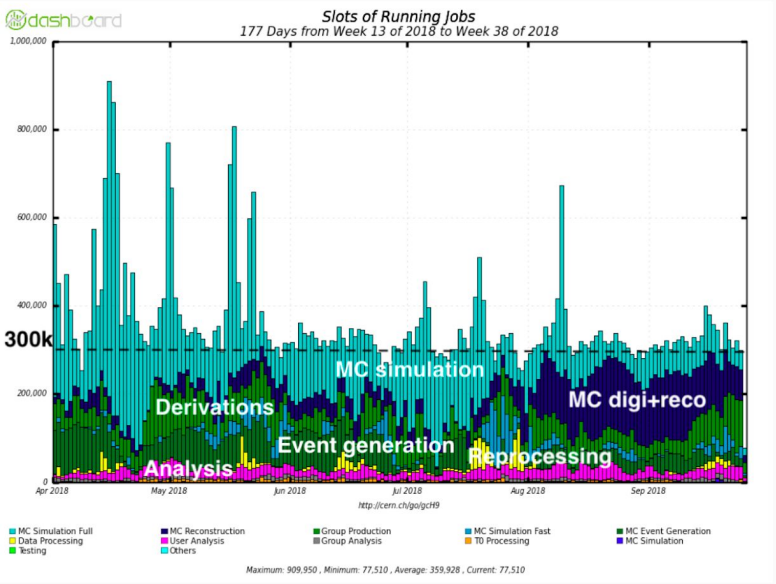
Andrej Filipcic

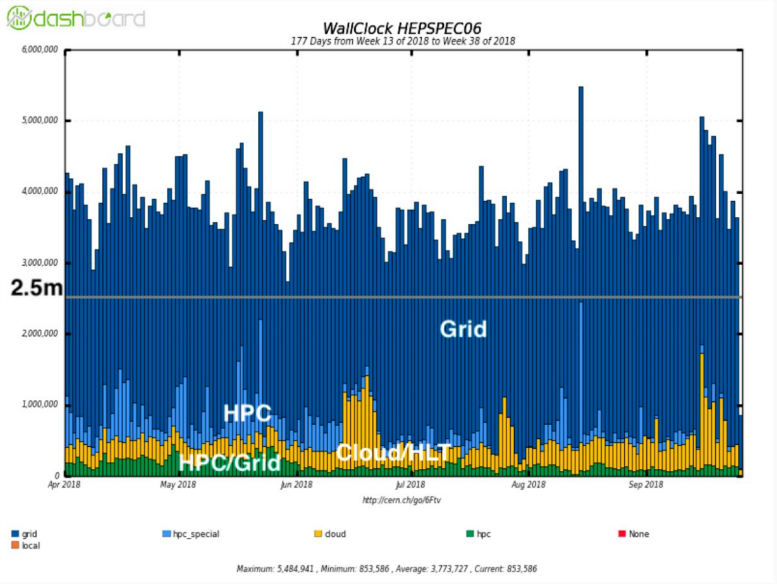
ATLAS Computing Infrastructure



- Central infrastructure
  - Resource configuration - AGIS
  - Job execution - PanDA
  - Data management and transfer - Rucio
- Data placement and payload execution on
  - Grid - data intensive computing centers, reserved resources for ATLAS
  - HPC - typically opportunistic CPU usage
  - Clouds and Boinc - opportunistic cloud and volunteer CPUs

Payload execution - jobs



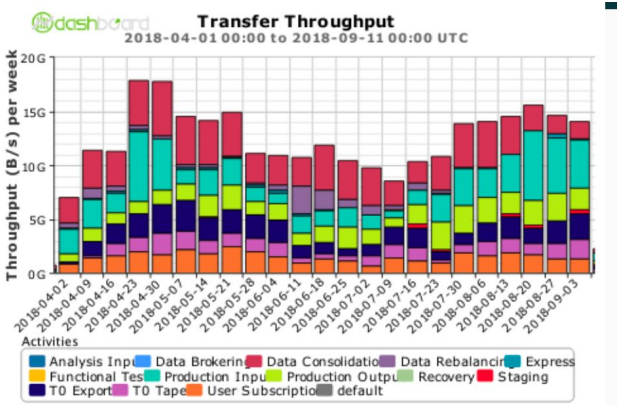


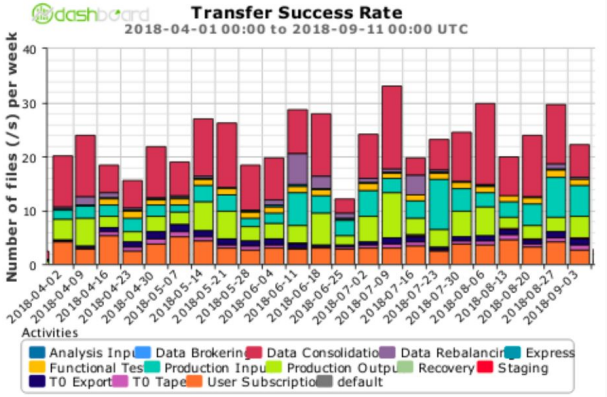
- ~300k CPU cores used all the time on ~100 centers
- CPU intensive (simulation, event generation) and data/CPU intensive jobs (the rest)
- 2.5B core hours/year

Processing campaigns

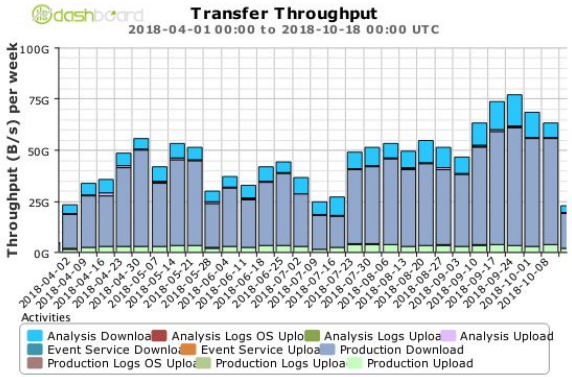
- Campaigns are periods of time when particular processing is ongoing
  - Several in parallel
- Data taking period, when LHC is active
  - Online High-Level Trigger Farm - 100k cores reducing the data rate from ~100kHz to 1kHz, few s latency
  - Data processing within 24h after data taking - full reconstruction on ~25k cores
- Data reprocessing campaign
  - Delayed reconstruction of real data with improved calibration and algorithms/software
  - Few months after data taking
  - A couple of months on 150k cores
- Monte-Carlo event (collision) generation and detector simulation
  - All the time, with an effective share of ~50%, using ALL available resources
- Monte-Carlo event reconstruction
  - Similar to data reprocessing, but more frequent, several times a year
- Derivations - following reconstruction of Monte-Carlo and data - preparing ready-to-analyze datasets for physics groups
  - Very frequent - every few weeks - analysis models and software evolve rapidly

Data transfers between computing centers

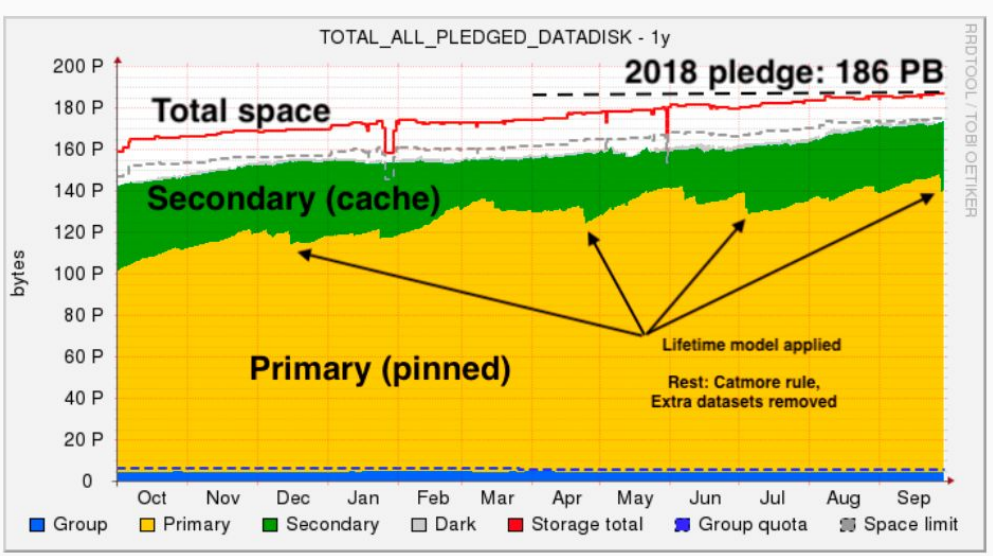




- Moving >1 PB, >20 GB/s, 1.5-2mio files per day between the sites
- Transfers managed by Rucio + File Transfer Service
  - Multiprotocol: gsiftp, https, xrootd, webdav, ...
- Jobs process up to 100GB/s, ~10PB/day
  - 203PB last month
    - 800B events

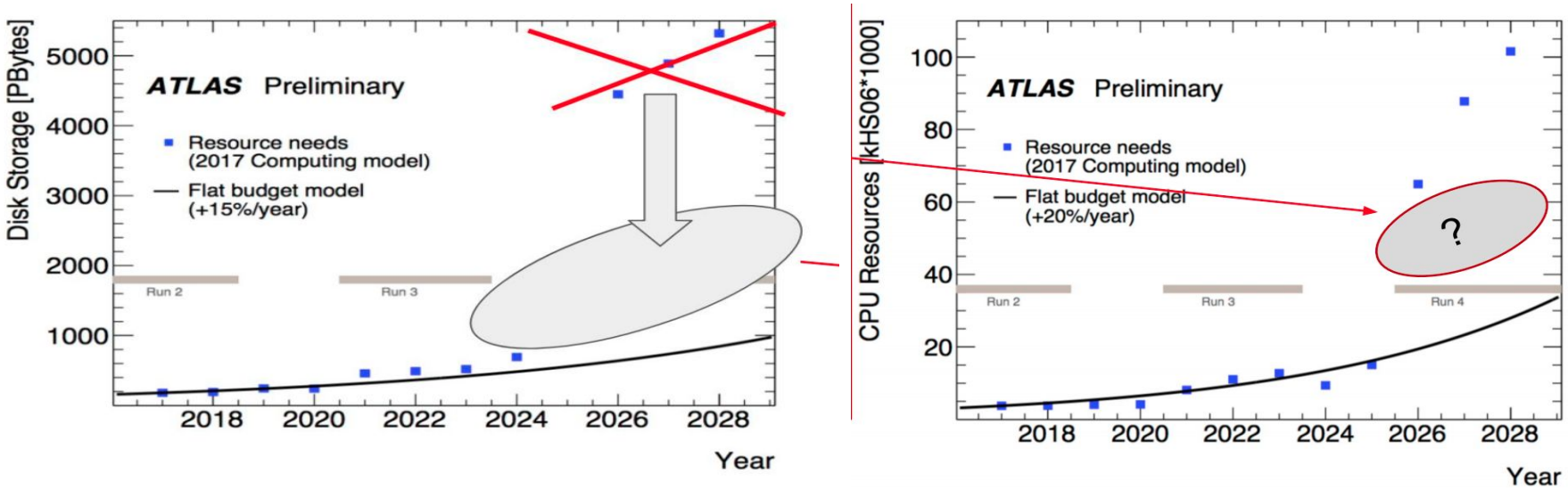


Disk and Tape

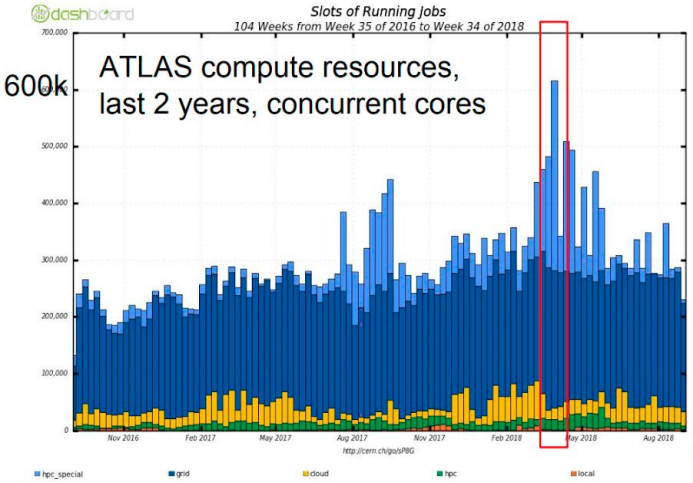


- Storage space:
  - ~200PB of disk
  - ~250PB of tape
- Replication factor ~1.3
  - Not enough space to have full redundancy

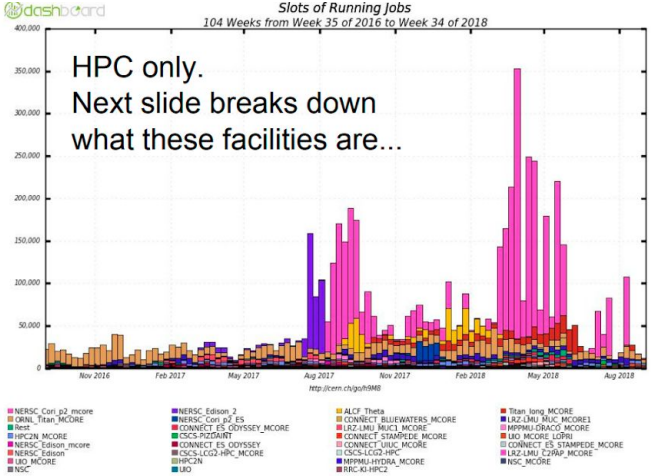
ATLAS in the next 10 years



- NOTE: preliminary plots, under discussion for upcoming LHCC
- Run-4 requirements (2026-2028) much higher than what we can expect from the past technology scaling
  - Disk: 6EB, but ~2EB more realistic
  - CPU: 5M today's cores, but maybe we should cope with 2-3M, 100-150 PFlops



A long history but a new era in the last year: very large facilities, so far in the US



Light blue: "special" HPCs, where special means big, difficult to use, US DOE

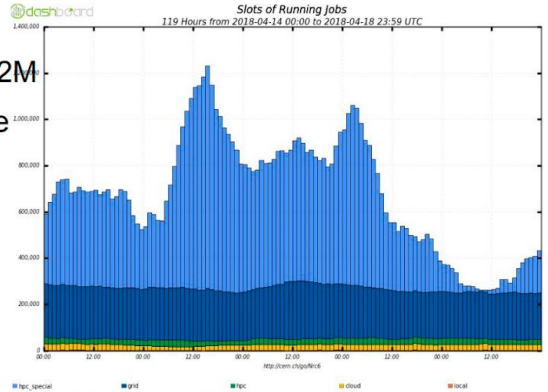
Dark blue: the grid

Yellow: cloud resources including (dominantly) HLT

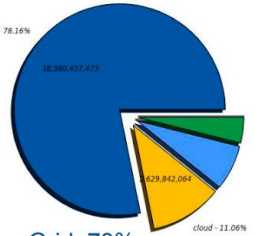
Green: "regular" HPCs, meaning easier to use, operate like a grid site, European or US NSF

Zoom showing full size of scaling peak: 1.2M concurrent cores.

Our workload management system is highly scalable!



CPU HS06 shares, last year



Grid: 78%
Cloud, HLT: 11%
HPC special: 7%
HPC regular: 4%

T. Wenaus October 2018

9

US ATLAS HPC resource allocations

US DOE has ASCR Leadership Computing Challenge (ALCC).

For many years we have gotten awards.

In 2016 we were awarded 13M hours at NERSC and 93.5M hours at ALCF (ANL)

In 2017 OLCF was added. 58M hrs at ALCF, 58 Mhrs at NERSC and 80M hrs OLCF (Titan) . We also got 10M hrs at NERSC through ERCAP program.

In 2018 - We received 100M hrs at NERSC through ERCAP program. We have submitted an ALCC proposal for 100 Mhrs ALCF, 70M hrs NERSC, 80 Mhrs OLCF

...and got 80M hours each at ALCF and OLCF from ALCC in 2018

Then there is the backfill time on Titan 195 Mhrs were used in 2017

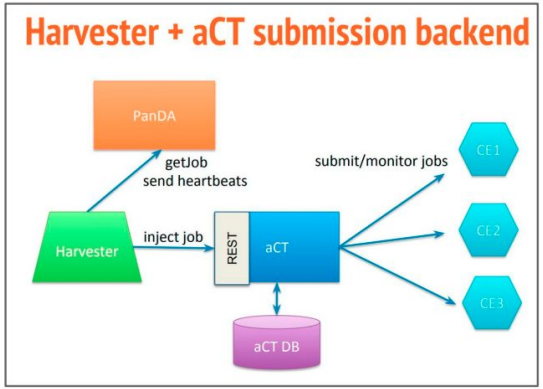
Doug Benjamin, March 2018

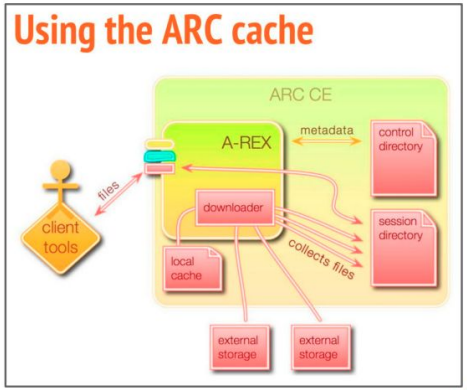
EU/PRACE HPC allocations and usage

- Nordugrid: Abisko, Abel, NSC pledge ~5k cores
  - Many more used on Abisko (opportunistic) and Abel (preemptive mode)
- PizDaint: decided to move the Phoenix WLCG T2 to HPC resources
  - ~10k cores, more in the future (ATLAS 40%)
- SuperMuc, Hydra
  - opportunistic/preemptive usage on ~6000 cores, or ~50M hours/year
- MareNostrum4:
  - Testing started this summer, with ~2M core hours for production
- IT4I:
  - Backfill usage of 2k cores
- Cineca: 20M hours planned per year

Nordugrid's ARC software

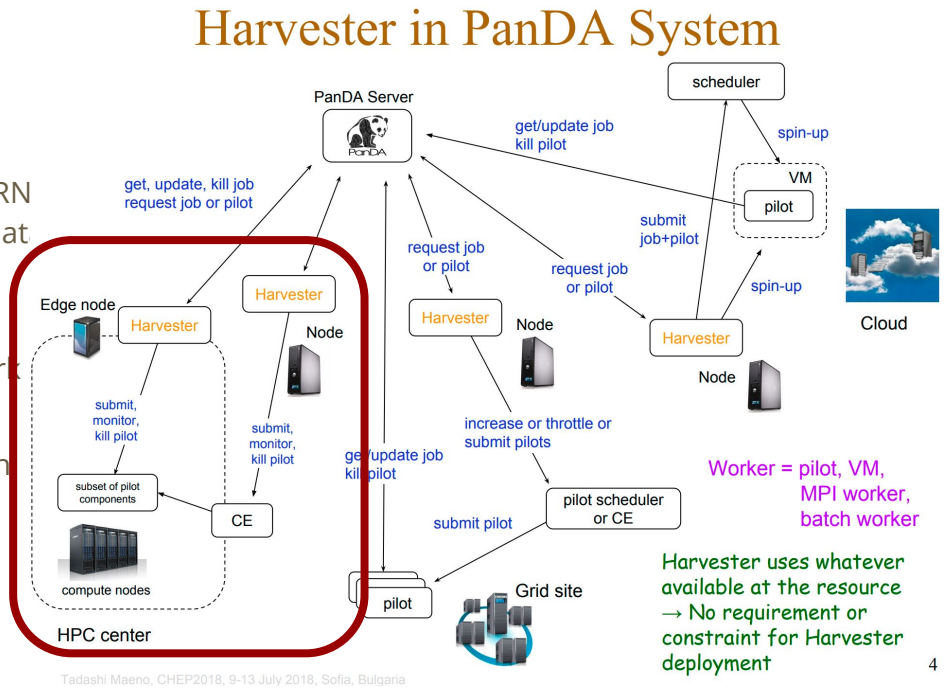
- Has long been the backbone of HPC integration in Europe
- Considered using ARC for US DOE HPCs but not seriously enough to make it happen
- Integrates workload and data management
- Isolates internal details from external users
- Integration requires little manpower for each system
  - But some policy dependence, friendly = easier
- Integrated with Event Service
- Integrating with Harvester to support advanced, dynamic workflows





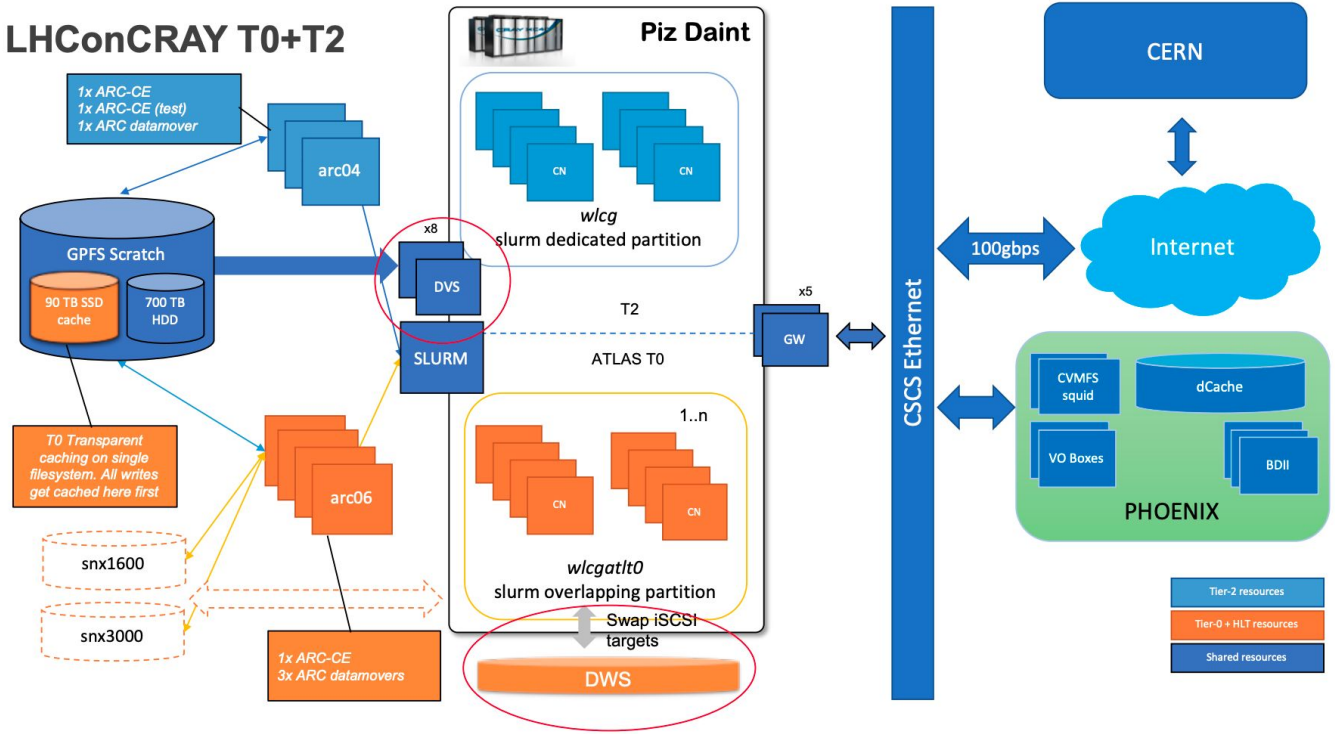
Transparent HPC integration in ATLAS

- EU HPCs
  - SuperMUC, IT4I, MareNostrum4, PizDaint, Draco, Nordugrid HPCs, IDRIS, Archer
  - Central arcControlTower payload submission at CERN
  - Local ARC Compute Elements on HPCs to transfer dat and submit jobs to batch
- CN HPCs
  - as EU but using SCEAPI to submit to CN HPC network
- US HPCs
  - Running pilot on the login node, submitting to batch
- All now being integrated with Harvester - payload and job submission controller
  - Running at CERN using aCT/ARC-CE for EU HPCs
  - Running at edge node using HPC batch in US HPCs



ARCHITECTURE HIGHLIGHTS

Elastic Tier-0 & HLT reprocessing - On-demand remote computation



Difficulties

- Authorization
  - on some US HPCs two-factor with hw keys
- Middleware and job submission frontends
  - Need for dedicated edge service - workarounds with userspace job management
- Data throughput
  - Need to use caches
- Access to external services (frontier db access, ATLAS services eg panda, rucio...)
  - No outbound connectivity
  - Can partially be solved by http(s) proxy
- Software installation, cvmfs
  - ATLAS release is big: 12GB, many releases used
  - Fat images (400GB)

Containers

- Using containers:
  - Independent on OS and system software (apart from kernel)
  - Negligible overhead in terms of CPU efficiency
  - Embedding experiment software
- Using singularity where available:
  - Through CVMFS (Nordugrid and PizDaint)
  - Or with fat images containing all the software where cvmfs is not available
- Using shifter
  - PizDaint
  - NERSC

ATLAS job requirements, 100k cores for one day

- Monte-Carlo simulation: CPU intensive and low I/O
  - 0.5 to 1GB memory per core, depending on no of cores/node
  - No outbound connectivity on nodes
  - 30TB needed to store input (10%) and output (90%)
- MC reconstruction and data processing: still CPU intensive, ~5 time faster than simulation and ~5 times more I/O per event
  - 2GB memory/core
  - Outbound connectivity to access calibration database
  - 750TB needed to store input (80%) and output (20%)
- software development efforts to make significant speed up in simulation (10 times faster)
  - I/O requirements will be 10 times higher

Conclusions

- ATLAS is successfully using many large HPCs
  - >10% contribution to overall CPU consumption
- More and larger allocations expected in the next few years
- Integration in ATLAS computing model is transparent, but we would all benefit with more standard approach
  - Eg common HPC access interface
- ATLAS software is being optimized/developed to leverage the HPC technologies
  - Interconnects, GPUs, non x86 architectures
- HPCs are opening up for BigData and cloud technologies
  - Many centers including WLCG grid sites are planning to provide high-performance, BigData and cloud
- Much larger contribution of such modern HPC centers is foreseen to be used by ATLAS in the future