



HPC usage and perspectives in ALICE

Latchezar Betev

Disclaimer

This presentation covers the ALICE HPC integration and operation principles

Many of the aspects and conditions are common for all LHC experiments

The KISTI Nurion integration - talk by Hyeonjin Yu

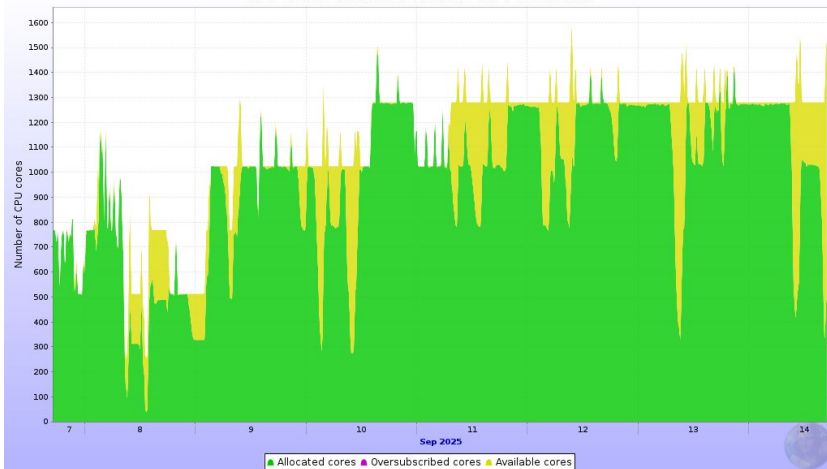
General principles for HPC integration and use

- HPCs are treated as integral part of the ALICE Grid infrastructure
 - Subject to same rules as the other Grid sites
 - Made (to the extent possible) into a Grid node
 - ***All users and workflows are shielded from HPC specifics***
 - Manual support is not favoured - the Grid should be automatic
- Individual discussions with HPCs are necessary to reach a compromise on the infrastructure limitations
 - There is always something new and different from the other HPCs
 - ... necessitating a customization of interfaces or use principles

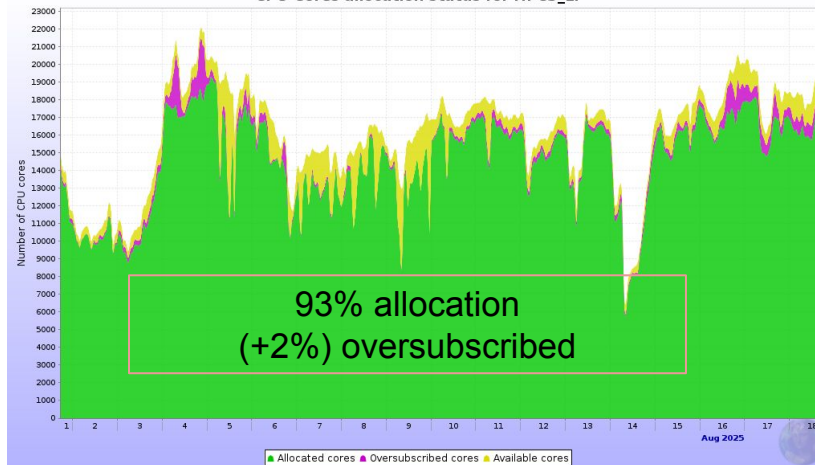
jAliEn middleware enhancements

- Whole-node scheduling
 - Compatible with HPC operation - the smallest quanta offered is a node
 - Internal brokering of arbitrary resources size (within one node) - CPU cores, memory, disk
 - External brokering of arbitrary payload type - compatible with HPC capabilities
 - Allows for adaptation to specific resources availability, for example GPUs

CPU Cores allocation status for Perlmutter

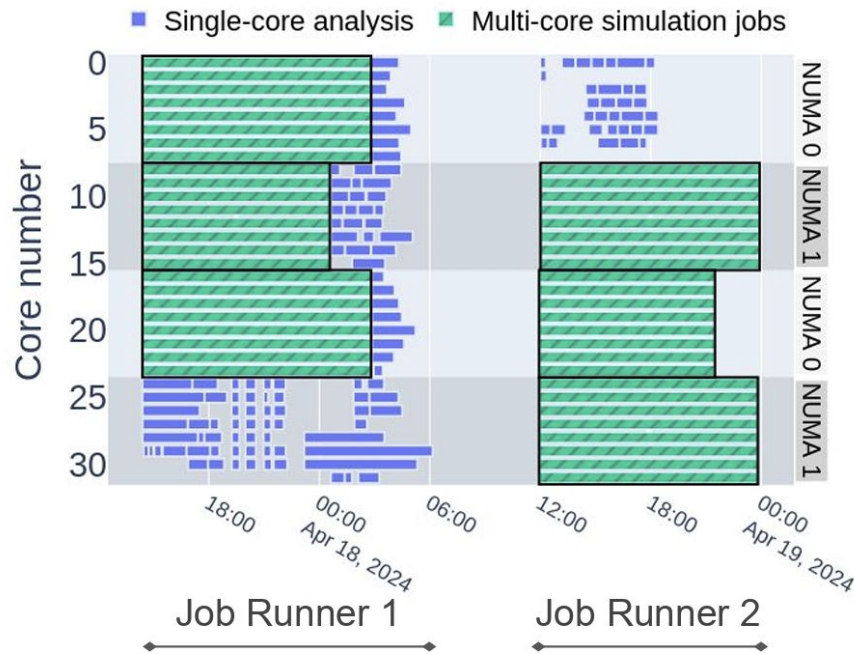


CPU Cores allocation status for HPCS_Lr

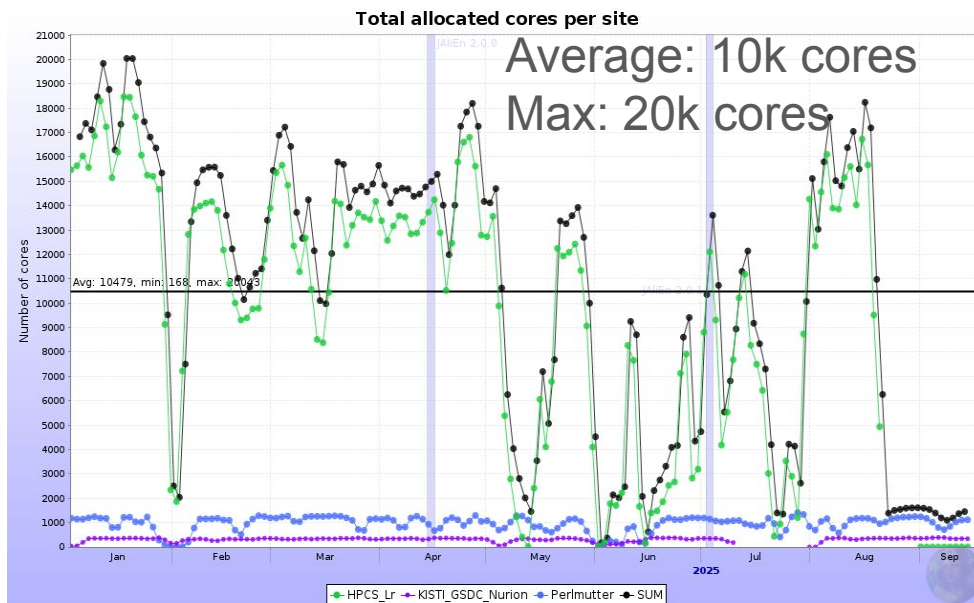


Optimising job placement in whole node scenario

- Job-to-core mapping algorithm starts by jobs with larger allocations
 - The slot is filled with payloads with different core requirements
 - Transparently balancing resource loads
- NUMA-aware pinning leads to improved execution efficiency
- Predictable execution time fostering better scheduling decisions



Current utilization of HPCs in ALICE



- 5% in average, 10% max contribution to total CPU allocation for ALICE

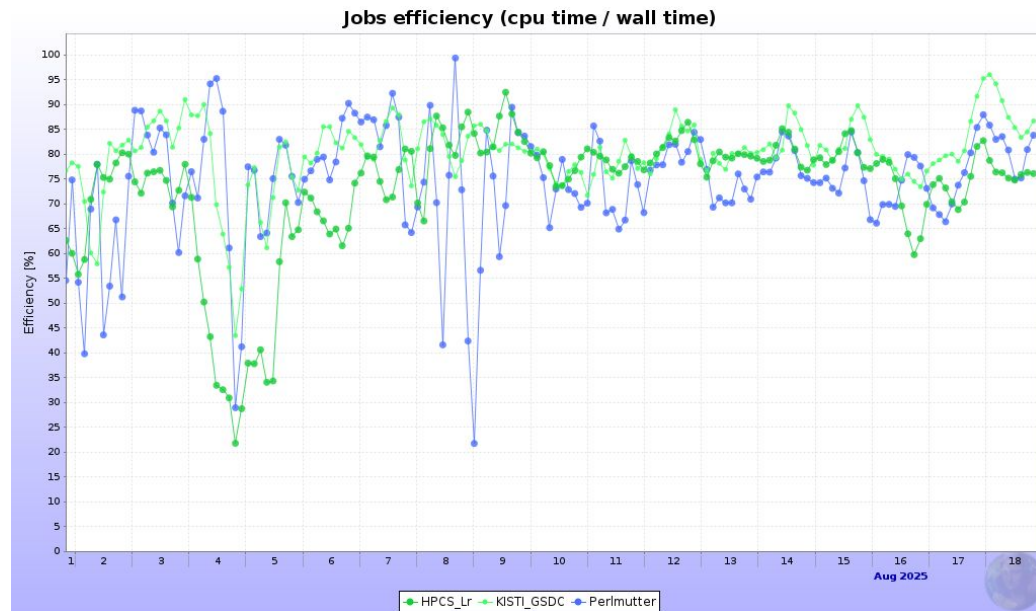
- 3 HPC integrated through JALiEn in the ALICE Grid, following the principles from the previous slide
 - Lawrence Livermore and Perlmutter at LBNL
 - Nurion at KISTI
- Co-located at a T2 and T1
 - Direct connection with the local storage elements
- Running all types of ALICE workflow - reconstruction, simulation, analysis
 - Closeness of storage is an important factor

HPC resources allocation

HPC	Allocation	Remarks
Nurion@KISTI	Static, 370 cores	10 nodes, 'alice' queue
Perlmutter@LBNL	Dynamic, ~1200 cores	CPU queue; From 2025, GPU allocation
Lawrencium@LBNL	Dynamic, no limit on resources	Preemptable

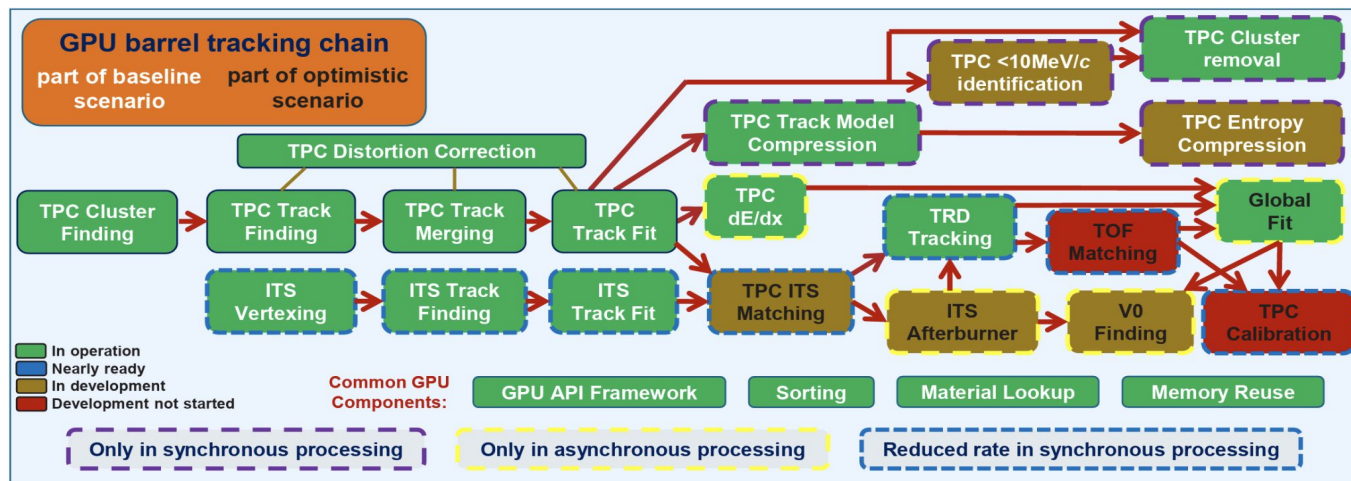
HPC location - resources use efficiency

- Payload fractions (MC, analysis) are similar on all resources
 - The HPC we have access to are well connected with network both to the internal storage on the sites and to outside
 - Ideal condition for payload-agnostic use
 - The above is not always the case
- I/O factors play the same role on HPCs and standard Grid resources



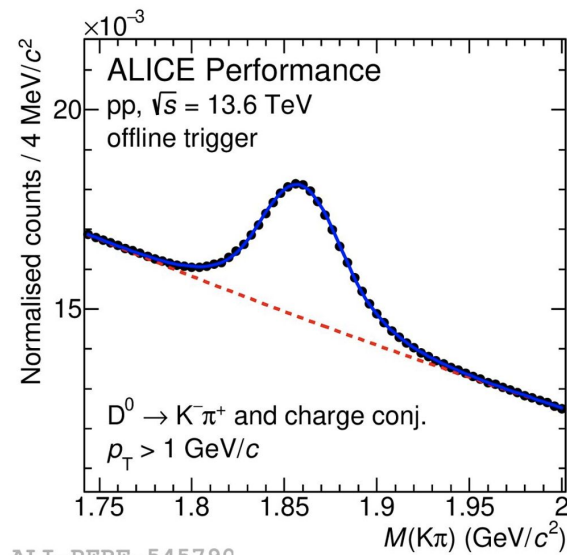
RAW data processing on Lawrencium, Perlmutter and Nurion

- A path to scale out a T1 resources and to run T0/T1 workflow on a T2
- Requires storage for RAW data close to HPC (tens of PBs in ALICE's case)
 - Not easily solvable problem
 - Worth pursuing, since with GPUs we achieve x4 speed-up and they are 'free' on HPCs



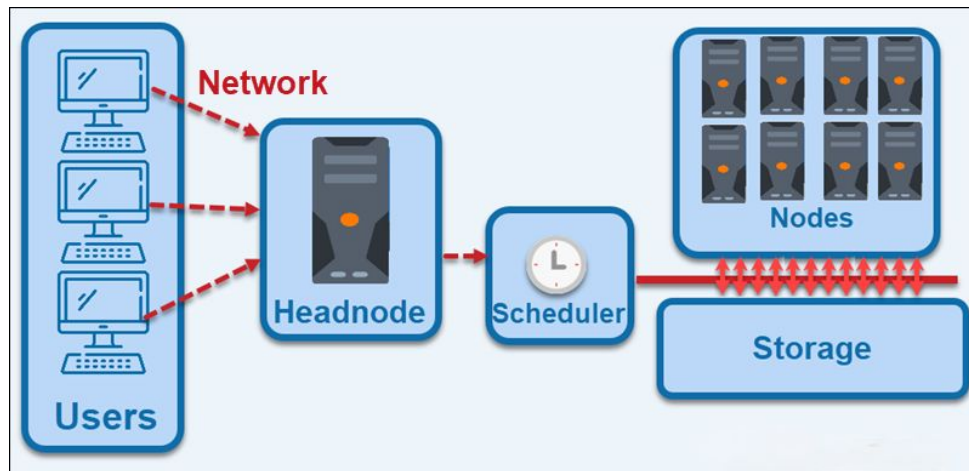
More workload examples and possibilities

- Detector calibration with NN
- Particle identification
- Offline triggers
- All of the above is used in production, but also presents a lot of development opportunities for interested people
- Expected to have soon GPU-enabled common MC software (G4)



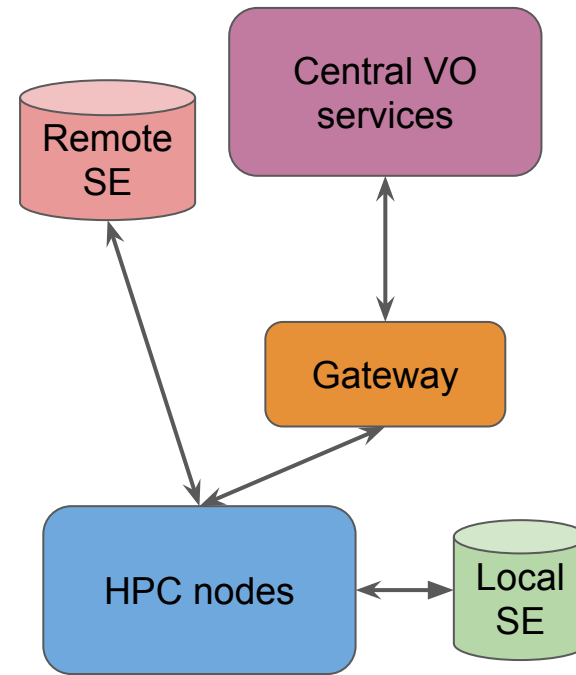
HPC - integration limitations

- High-Performance Computers pose unique integration challenges for WLCG
- By design, most of these are self-contained systems, use specialized software and have no outside network which clash with the Grid's interconnected, standardized environment
- These unique characteristics require individual approaches and non-trivial amount of development and support
- Today, their integration is up to the individual VOs efforts



HPC - key elements for adoption

- Common and standardized access mechanisms (similar to Grid gateways)
- Standardized authorization and authentication tools and protocols
- Sufficient external network capabilities for remote storage and services access
- Common methods for software delivery (e.g., CVMFS)
- Allows the use of modern techniques for software isolation and control (e.g., containers)
- All of the above is easier to achieve if taken into account in the HPC design phase



HPC - future projects - Pegasus@Tsukuba

- Perfect internal architecture for Grid payload
- jAliEn is ready for the integration at all levels
- Could be a solid basis for the re-establishment of the Tsukuba T2
 - Only a VO-box is needed
 - ...and the fulfillment of certain limiting conditions
- 1 Node - 48 cores, 128GB RAM, i.e. 2.6GB/core



Pegasus@Tsukuba status

Condition	Status
Kernel modules for CVMFS	FUSE available
Batch manager	PBS derivative
Whole node or limited CPU queues allocation	24h allocation, whole node
Scratch space on WNs	6TB/node, 12GB/core
Access protocol	2FA, need to have local user
Containers	Apptainer
Outgoing network	HTTP proxy (or similar)
Allocation principle	Paid allocation

Pegasus - project timeline

- With the exception of networking - all other critical components are compatible with the Grid requirements
- A standard yearly allocation (10 nodes, 90k node-hours) is equivalent to a medium-sized T2 and will be equivalent to the Perlmutter HPC allocation
 - T2 focus is shifted to storage procurement and operation
- If successful negotiations and setup, ***Tsukuba will be the first (and only) WLCG site*** which is entirely based on HPC
 - This may be the future of more sites, especially those co-located with HPC
 - Could be a model for WLCG resources provisioning, if compatible resources and capacity exist

Summary

- HPC resources play an important role (up to 10%) in the ALICE Grid computing
 - 3 HPCs in operation
- Adopting HPCs - made simpler over the past 2-3 years
 - Number of additional modules included in the jAliEn middleware
 - Payload encapsulation to isolate from the HPC specifics
- HPC resources (unless entirely opportunistic) can be the building block for a site
 - Focus is shifted predominantly to storage operation
- ALICE is very interested in prototyping this @Tsukuba with the Pegasus HPC
- HPC integration is challenging and a good project for computing science majors (MsC and PhD)
 - Participation from Asian technical universities is welcome and supported