HEP use cases for HPC

Maria Girone, CERN
Tommaso Boccali, INFN

Based on
[Strategy for HPC Integration in WLCG/HEP]

Strategy for HPC Integration
in WLCG/HEP

Input to the ESPP24-26

Submitted by Maria Girone, CERN (maria.girone@cern.ch) on behalf of

M. Girone[1], B Bard, D. Benjamin, T. Boccali, A.F. Boyer, C. Bozzi, P. Calafiura, S. Campana,T. Childers, A. Filipcic, I. Fisk, O. Gutsche, J. Hernandez, D. Hufnagel, A. Klimentov, E. Lançon, M. Lassnig, V. Martinez Outschoorn, G. Merino, S. Piano, L. Sexton-Kennedy, D. Spiga, F. Wuerthwein

Executive Summary

High-Energy Physics (HEP), and particularly the sector supported by the WLCG Collaboration at the LHC experiments, is entering a new era of data-intensive research. This shift is driven by the High-Luminosity LHC (HL-LHC) that will generate exabyte-scale datasets each year. Fully realizing the physics potential of this massive volume of data will require a significant increase of resources. High Performance Computing (HPC) centers are crucial partners in this effort, offering either pledged resources or for providing additional opportunistic resources that enhance the physics output. Integrating HPC systems into HEP workflows offers transformative benefits, like expanding computational power, accelerating simulations, and enabling more sophisticated AI/ML algorithms. However, realizing these benefits demands a concerted effort to address technical, organizational, and policy-related barriers.

[1] corresponding author, mail: maria.girone@cern.ch

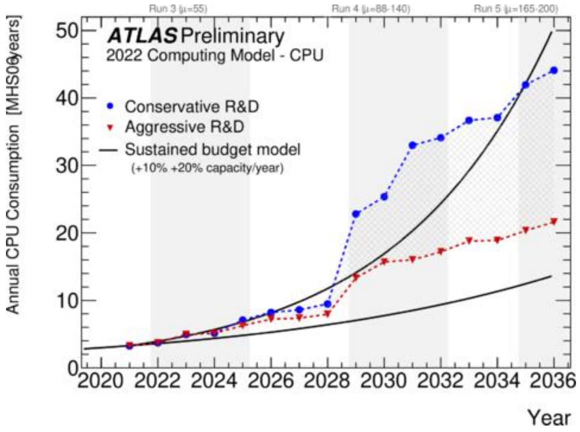
Introduction

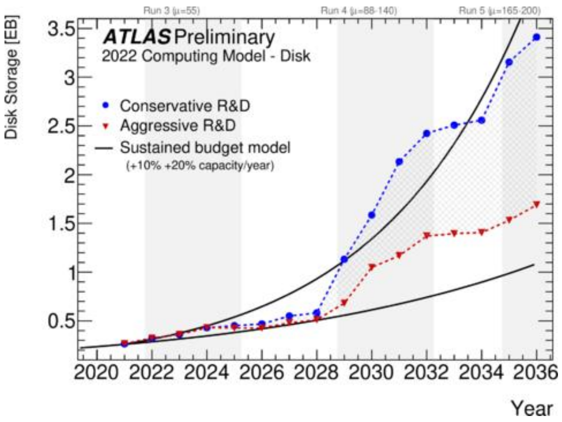
High-Performance Computing (HPC) presents a transformative opportunity for the HEP community

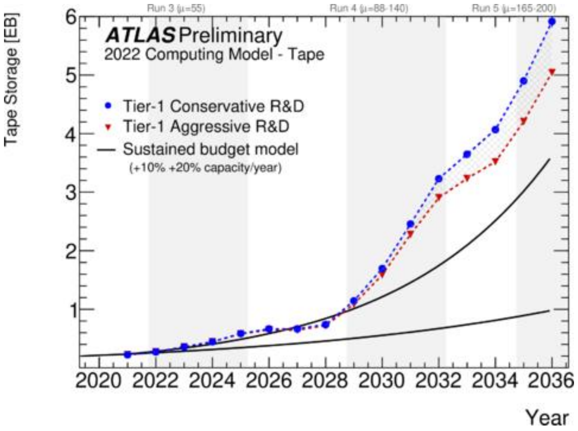
- Modern HPC systems, with exascale capabilities and advanced hardware, from CPUs to GPUs and AI accelerators, offer vast computational power, relevant to HEP activities

- Fully harnessing this power requires significant adjustments to our computing workflows
  - GPUs and AI-driven architectures promise higher performance and energy efficiency but also necessitate costly redesigns of legacy software and the development of AI-focused workflows – eventually more than once
  - Integrating HPC systems into HEP workflows offers transformative benefits but demands a concerted effort to address technical, organizational, and policy-related barriers

Computing Needs and R&D





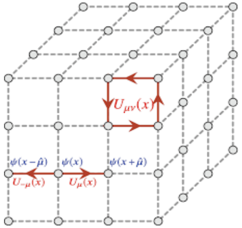


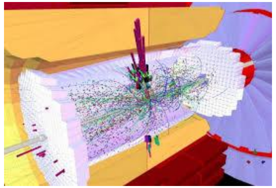
- High-Energy Physics (HEP) continues on its path towards a more and more data-intensive research, driven by the High-Luminosity LHC (HL-LHC) that will produce exabyte-scale datasets each year.
- This has spurred aggressive R&D programs across all experiments to meet evolving requirements within constrained budgets and uncertain technology gains.
- These efforts focus on reworking algorithms, adoption of AI workflows, and data management practices to fully exploit future hardware and achieve our physics programme.

Traditional HTC/HPC Use-Cases

Lattice QCD

- Precision calculations of hadronic observables (e.g., anomalous magnetic moment, CKM matrix elements, neutron dipole moment).
- Generation of gauge field configurations and quark propagators demanding large-scale GPU resources and low-latency networks.



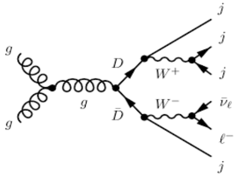


Detector Simulation (Geant4)

- GPU-accelerated electromagnetic shower simulations significantly reduce computational load.
- Projects AdePT (CERN) and Celeritas (US) integrating GPU-based physics modules into Geant4 for HL-LHC detector simulations.

Monte Carlo Event Generation

- Simulating initial particle collisions with highly parallelizable workflows, ideal for GPU acceleration.
- Projects like Madgraph5_aMC@NLO and Sherpa/Pepper advancing GPU implementations for NLO simulations.



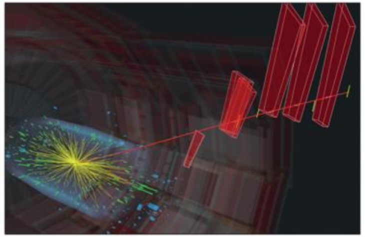
WLCG (distributed) and Tier-0 (CERN) Workflows

- HTC-like processing leveraging HPCs as elastic expansions of WLCG resources.
- Burst execution for large-scale sample production and rapid dataset reprocessing

AI/ML Use-Cases on HPC (1/2)

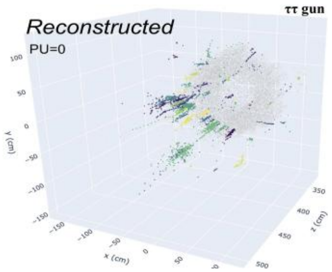
Next Generation Triggers (NextGen)

- Optimize rare-event selection through GPU-powered ML model training at large scale.
- HPC resources enable more complex trigger algorithms for HL-LHC physics analysis.



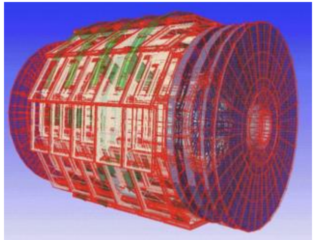
Distributed AI Model Training

- Large-scale hyperparameter optimization (HPO) and distributed training, e.g., Particle Flow (MLPF) models.
- Significant physics performance improvements demonstrated using HPC.



Digital Twins

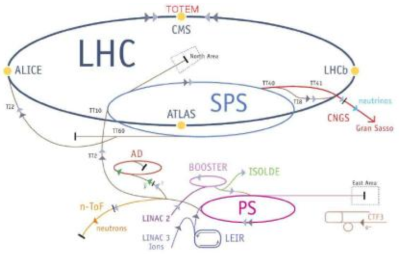
- High-performance training and simulation of digital twins for real-time replication of scientific instruments and processes.
- Essential for large-scale hyperparameter optimization and numerical simulation requiring HPC-scale computation.



AI/ML Use-Cases on HPC(2/2)

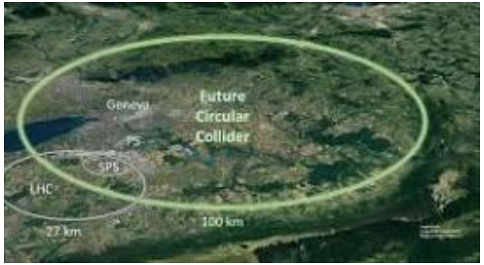
Foundation Models & Large Language Models

- Exploring transformer-based architectures (Foundation Models) for particle flow reconstruction and fast simulation.
- Large-scale HPC resources critical for training and evaluating these sophisticated ML architectures.



CERN Accelerator (ATS) AI/ML Activities

- Leveraging AI/ML for optimization, predictive maintenance, fault analysis, and operational efficiency of accelerators.
- HPC essential for training large-scale models (e.g., AccGPT chatbot, GPT-based automation).



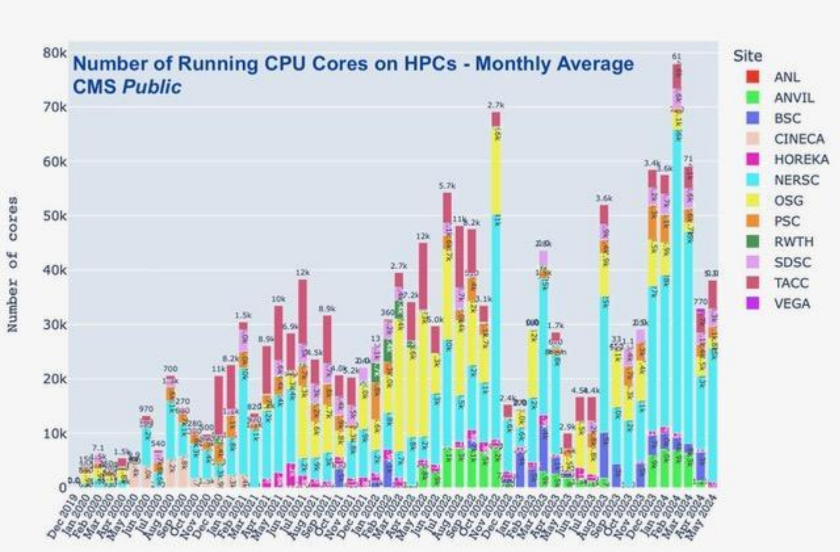
FCC (Future Circular Collider) Simulations

- Intensive detector simulations, ML training, and GPU-accelerated MC event generation benefiting from HPC resources.
- HPC accelerates iterative ML training cycles, reducing turnaround time significantly.

Technical Challenges (1/6)

Integration & Interfaces

- **Lack of Standard Interfaces**
  - Diverse HPC center-specific access policies complicate integration.
  - Workflows must be customized per center, hindering scalability.
- **Impact**
  - Manual integrations are inefficient, costly, and error-prone.
- **Next Steps**
  - Establish standard HPC-WLCG interface protocols.
  - Collaborate with HPC Federation Initiatives, WLCG, and resource providers to define standards.



Technical Challenges (2/6)

Software & Architectures

- **Software Adaptation for HPC**
  - HEP codebases (millions of lines, legacy optimized for x86_64 CPUs) not tailored for GPUs or new accelerators.
- **Challenges with Portability**
  - GPU architectures require specialized coding techniques.
  - Current software limits effective use of HPC accelerators.
- **Next Steps**
  - Adopt portability libraries and unified programming models.
  - Sustain pilot projects to transition critical workflows to GPU-compatible code.

Technical Challenges (3/6)

Data Management & Networking

- **Data Movement Limitations**
  - HPC facilities don’t normally support permanent data storage; transient storage solutions are required.
- **Network and Storage Constraints**
  - High-throughput data transfers to/from HPC sites are constrained by limited connectivity and policies.
  - Absence of standardized automated data transfer solutions.
- **Next Steps**
  - Develop scalable WLCG-HPC data transfer standards.
  - Run data transfer and workflow integration tests at large scale.
  - Engage with HPC Federation initiatives (EFP and IRI) to build federated HPC-WLCG storage and networking solutions.

Technical Challenges (4/6)

Authorization and Authentication

Lack of Federated Authentication

- HPC facilities don't have standard federated identity management compatible with WLCG's Virtual Organization (VO) trust model.

Operational Impact

- Current HPC authentication policies restrict access to known individuals, limiting scalability for large HEP collaborations.
- Complex management of user roles and permissions, complicating workflow integration and reducing efficiency.

Recommended Next Steps

- Develop a federated Authentication and Authorization Infrastructure (AAI) interoperable between WLCG and HPC.
- Collaborate with regional initiatives (e.g., EuroHPC JU, IRI, EFP) to create standardized frameworks for secure, federated access.
- Establish common security standards and VO-based access models at HPC centers, facilitating broader usage.

Technical Challenges (5/6)

Workflows & Edge Services

- **Workflow Integration Complexity**
  - HPC centers have limited compatibility with existing HEP/WLCG workflow management systems.
  - We need to have custom interfaces to submit pilot-based workflows on HPC.
- **Connectivity Constraints**
  - Restrictions on outbound connectivity from HPC worker nodes limit workflow flexibility.
  - We can make fully-contained applications that don't need outgoing access for a subset of workflows, but many need the ability to query edge services
- **Impact on HEP**
  - Reduced resource utilization and efficiency due to workflow integration challenges.
  - Increased complexity in workload management and scheduling decisions.
- **Recommended Next Steps**
  - Establish standard workflow integration protocols between HPC and WLCG.
  - Identify and optimize workflows best suited for HPC resources.
  - Implement "Edge Services" (e.g., CVMFS caches, local proxies, data-staging services) at HPC facilities to support flexible and efficient workflow execution.

Technical Challenges (6/6)

Resource Provisioning & Policies

- **Mismatch in Allocation Models**
  - HPC resources typically allocated via short-term, peer-reviewed proposals.
  - LHC experiments require guaranteed, multi-year resource commitments.
- **Impact on HEP**
  - Short-term HPC allocations limit long-term scientific planning and investments.
  - Frequent application cycles conflict with stable, multi-decade HEP programs.
- **Recommended Next Steps**
  - Advocate for multi-year or guaranteed annual allocations aligned with HEP goals.
  - Engage with HPC funding agencies to recognize HEP as a strategic, long-term science priority.
  - Establish agreements (MoUs) for continuous access to HPC resources, ensuring stable scientific output.

Conclusions

- Integrating HPC with the existing HEP and WLCG distributed computing infrastructures is critical to accelerate scientific discoveries and strengthen international competitiveness. Key challenges remain:
  - Software optimization for HPC architectures
  - Efficient workflow and data orchestration
  - Securing strategic, long-term access to HPC resources
  - Standardizing interfaces and policies across HPC centers

- A coordinated, interdisciplinary effort—including talent development in scientific software—is crucial for maximizing the benefits of HPC systems
  - HEP and RA work closely together in EU-funded initiatives such as SPECTRUM and ODISSEE

- By building upon suc es such as WLCG, the HEP community is in an excellent position to collaborate with strategic HPC programmes (e.g., AI Factories, Federation Platforms), fostering a robust, interconnected, and data-centric research ecosystem.

Thank you!