CompF4 Processing Summary

Ian Fisk (Flatiron Institute)
Meifeng Lin (Brookhaven National Laboratory)

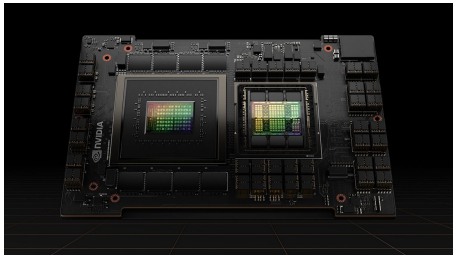
Introduction

- Definition of processing in the context of CompF4: **“Processing is the step that transforms raw data, simulation configurations or theoretical models into objects useful for analysis and discovery”**
- Processing can take place in a variety of environments under different constraints:
  - Dedicated online systems for experiments
  - Globally distributed processing sites, such as Grid
  - Tightly coupled High Performance Computing (HPC) resources (Perlmutter, Frontier, etc.)
  - Cloud resources (AWS, Google Cloud, etc.)
- The core of processing is the hardware compute units, which are getting more and more diverse – CPUs, GPUs (AMD, Intel, NVIDIA), FPGAs, etc.
- We address the processing challenges and R&D needs arising from the diversification of both the **computing environments** and **hardware architectures**.

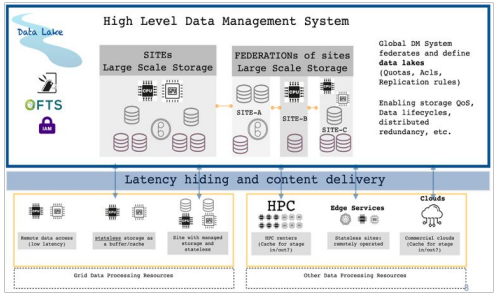
Relevant Whitepapers

- S. Campana, A. Di Girolamo, P. Laycock, Z. Marshall, H. Schellman and G.A. Stewart, Hep computing collaborations for the challenges of the next decade, 2022. 10.48550/ARXIV.2203.07237.
- B. Joo, C. Jung, N.H. Christ, W. Detmold, R.G. Edwards, M. Savage et al., Status and future perspectives for lattice gauge theory calculations to the exascale and beyond, European Physical Journal. A 55 (2019).
- Y. Kahn et al., Snowmass2021 Cosmic Frontier: Modeling, statistics, simulations, and computing needs for direct dark matter detection, in 2022 Snowmass Summer Study, 3, 2022 [2203.07700].
- P. Boyle, D. Bollweg, R. Brower, N. Christ, C. DeTar, R. Edwards et al., Lattice qcd and the computational frontier, 2022. 10.48550/ARXIV.2204.00039.
- D. Casper, M.E. Monzani, B. Nachman, C. Andreopoulos, S. Bailey, D. Bard et al., Software and computing for small hep experiments, 2022. 10.48550/ARXIV.2203.07645.
- FASER, ATLAS, LZ, Fermi-LAT, H1, T2K, SBND collaboration, Software and Computing for Small HEP Experiments, in 2022 Snowmass Summer Study, D. Casper, M.E. Monzani, B. Nachman and G. Cerati, eds., 3, 2022 [2203.07645].
- M. Girone, Common challenges for HPC integration into LHC computing, Feb., 2020. 10.5281/zenodo.3647548.
- M. Bhattacharya et al., Portability: A Necessary Approach for Future Scientific Software, in 2022 Snowmass Summer Study, 3, 2022 [2203.09945].
- C.D. Jones, K. Knoepfel, P. Calafiura, C. Leggett and V. Tsulaia, Evolution of HEP Processing Frameworks, in 2022 Snowmass Summer Study, 3, 2022 [2203.14345].

Challenges Identified



- Heterogeneous Hardware
  - HEP software for processing already requires significant maintenance efforts.
  - The diversity of heterogeneous hardware poses further software development and maintenance challenges.
- Resource Interfaces
  - The protocols and interfaces to connect to grid sites have functioned and scaled, but the integration of new resources like HPC and clouds sites is a new technical challenge.
  - The HPC facilities have stricter cyber-security requirements and Authentication and Authorization Infrastructure (AAI) needs.
- Provisioning and Policy
  - Both HPC and Clouds are fixed resources, either due to allocation or budget, and this places challenges on how to predict usage and enforce experiment priorities.
- Data Management and Delivery (see also **Storage and Networking**)
  - Traditionally, data is moved to dedicated storage and accessed locally with only a minority share, if any, of the data streamed.
  - The addition of non-dedicate processing resources like HPC and clouds places challenges on the data management system to be more dynamic.
- Impact of ML-based processing (see also **AI Hardware**)
  - The adoption of Machine Learning based workflows in processing intensive applications introduced challenges in the balance of resources and the types of computing needed.
  - ML training - suitable for HPC sites
  - ML inference - may benefit from dedicated hardware such as FPGAs



Research Directions

Use of heterogeneous architectures

- What is the future of architectures like GPUs, FPGAs, XPU, DPUs, TPUs, etc?
- Can we have a single code base with unified programming models and portability libraries?

Evolution of resource sharing and provisioning

- Use of annual HPC allocations and cost per use Cloud allocations requires rethinking how we budget and provision resources
- Can we have Workflow driven computing rather than constant programmatic computing?

Evolution of data access

- Separation of storage and processing drives needs in networking, caching, data placement
    - *DataLakes*

Evolution of interfaces

- Description of diverse hardware, costs and accounting, cyber security and increases in scale require rethinking of resource interfaces

Modifying programming models

- Will be able to afford large scale storage of raw data for eventual reprocessing? Do we need to reconstruct in the online and write analysis objects?

Findings 1/2

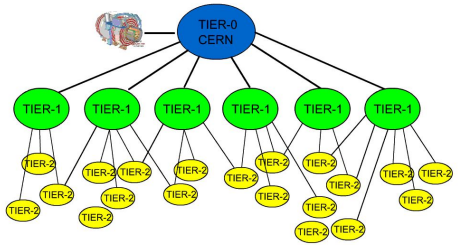
The environment is changing

- Investments from industry in more specialized processing architectures
- HPC sites and clouds are growing
- Expectations for how long a result or query should take are shrinking
- Growth in non-traditional processing techniques like ML, computer vision, etc.
- To make an impact in computing the breadth of expertise needed is increasing

Findings 2/2

The processing models deployed today have evolved adiabatically from a set of real and imagined constraints

- There has been very little discussion of what we are optimizing for and how to direct the evolution moving forward
- We should be identifying the guiding system attributes beforehand
    - Overall costs, use of existing resources, Time to results, Carbon Footp familiarity, synergies with other sciences or industry
- There is probably not single set of optimizations, but the process to identify them can be common



Recommendations

- ***Evolution of Resource Access:*** HPC facilities should revisit their resource access policies to allow more flexible allocations and job executions. This, coupled with new authentication and authorization models, will allow more HEP projects to benefit from the large computing facilities. Work internationally with science and HPC communities.
- ***Efficient Use of Heterogeneous Architectures:*** Investment in software development effort is key to maximize the efficient utilization of diverse processing resources. In particular, research and development of portable software solutions is critical for a sustainable software ecosystem in light of the evolving and increasingly diverse hardware architectures.
- ***Evolution of Computing Facilities:*** Research is needed to determine the tradeoff between dedicated HEP computing facilities and general-access computing facilities such as the HPC center, Grid and Cloud resources. Work within the current and future collaborations.
- ***Data Access and Management:*** Infrastructure development will be needed to support better data management frameworks across different types of facilities.

Questions and Comments?

Contact us:

Ian Fisk, ifisk@flatironinstitute.org

Meifeng Lin, mlin@bnl.gov