WP5: Architectural Support for Theoretical and Experimental Physics Data Management on the Distributed CN infrastructure

inputs from WP1

Mattia Bruno

October 13, 2022
Kick-off meeting Spoke 2 Centro HPC

WP1

1. identification of use cases

2. implementation of numerical strategies, and codes

   → testbed needed for R&D

   → production of data

3. validation

   → (global) access of data

   → production of additional data

Main players:

Condensed matter and low dimensional systems
Lattice field theory
Collider physics phenomenology
Gravitational waves, cosmology and astroparticle physics
Physics of complex systems
High energy nuclear physics

WP1

1. identification of use cases

2. implementation of numerical strategies, and codes

   → testbed needed for R&D

   → production of data

3. validation

   → (global) access of data

   → production of additional data

Main players:

Condensed matter and low dimensional systems
Lattice field theory
Collider physics phenomenology
Gravitational waves, cosmology and astroparticle physics
Physics of complex systems
High energy nuclear physics

Overall impression from WP1

Underlying common denominator: Monte Carlo methods/Markov processes to simulate behavior of physical systems

Development of codes capable of using GPUs is primary goal

- common to all involved participants
- still collecting requirements in terms of testbeds

Majority of requirements for testbeds

- storage: O(10 TB) up to O(100 TB)
  - outlier is Lattice QCD can reach O(1 PB)
- computing nodes for R&D both CPU and GPU

Interactions w/ data-lake infrastructure involves less participants

- but could easily evolve and become more inclusive

WP1 requirements

1. guidelines for access/movement of data across HPC and data centers (both small and big centers)
2. creation of databases of primary data generated from (expensive) HPC simulations
3. creation of databases of secondary data produced from processing primary data

Possible tasks for WP5:

- creation of guidelines, identification of tools for data movement and storage
- identify storage points within old/new data-lake
- support requests to board for HW testbeds, e.g. servers for searchable databases, and possibly analysis tools

Example of use case from LQCD: Storage

1. Primary data (e.g. field configurations) → large amount of data

   from 200TB up to 1PB

   searchable databases (see next slide)

   tapes more likely as primary architecture

   guidelines for optimal access/movement of data

   preservation on time scales of 10yrs

Example of use case from LQCD: Database and Analysis

2. “Derived” data, e.g. measurements on field configurations

   from 100GB up to 10TB

   contain information of primary data for searchable databases

   disks more likely as primary architecture

Idea for deployment of database: dedicated mini servers for
automatic deployment of analysis as new data is produced
hosting of jupyter-hub to directly manipulate data
hosting of searchable database of primary and secondary data

From WP5 to WP1

For example: many groups/activities use or plan to use Python

Possible "use cases" for WP5:

- creation of containerized applications shared by several groups
- guidelines for deployment on HPC centers and local in-house clusters
- link w/ WP4: optimized containerized software

Publicity of ideas like this one may inspire further interactions among WP1,2,3 ↔ WP4,5,6