Lattice QCD

Hubert Simma

DESY

DMA ST1 Synergy Workshop

November 9, 2023

Overview

1. Workflow and Data Lifecycle
2. Requirements
3. ILDG
4. Synergies

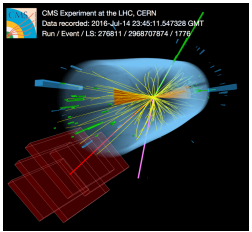
QCD on the Lattice

Quantum Chromodynamics (QCD)

Computation of hadronic observables, Q, (masses, formfactors, ...)
from first principles and beyond perturbation theory: "path integral"

where
$$\langle Q \rangle = \int_C e^{-S(C)} \cdot Q(C)$$

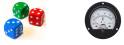
- **C** = gluon (and quark) field at each point in 4d space-time
- **S** = classical action ($\leftrightarrow$ field equations)



Lattice QCD

- discretization: fields defined only on a finite lattice e.g. $V \sim 50^3 \times 200 = 25$ million sites
- integration in $O(10)\times$ V dimensions: importance sampling of field configurations C_i with weight $e^{-S(C_i)}$ by a Markov Chain Monte-Carlo (MCMC) simulation

$$\langle Q \rangle \approx \frac{1}{N} \sum_{i=1}^{N} e^{-S(C_i)} \cdot Q(C_i)$$



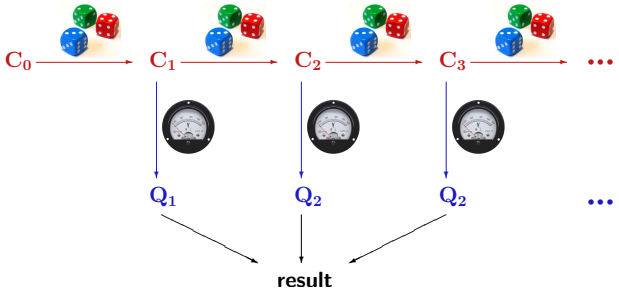
$$\begin{pmatrix} \textit{sample size } N & \to & \infty, \\ \textit{lattice spacing} & \to & 0, \\ \textit{phys volume} & \to & \infty, \\ \textit{masses } m & \to & m_{phys} \end{pmatrix}$$

Lattice QCD Simulations and Data

Simulation
HPC / performance

"Measurement"
HPC / throughput

Analysis



Raw data = samples ("ensembles") of "configurations" C_i

- low "event rate" (1 config / 30'000 core hours)
- large volume (1 ... 100 GB $\times$ 1000 ... 10000 configs $\rightarrow$ O(PB))
- expensive to generate (1 ... 100 million core hours / ensemble)
- re-usable in multiple projects / collaborations for different "measurements"

Workflow and Data Lifecycle

Computing

Generation of ensembles:
sequentially on single HPC system

↓ ↓ after up to O(10) years

"Measurement" on each config:
possibly in parallel at different HPC sites,
internally two computational steps:

- "propagators" expensive iterative solvers (as in generation), usually not stored
- "contractions" possibly expensive, specific for each project

➜ Analysis on workstations

Storage

keep at HPC center or
→ move to home institutions
↔ sharing via ILDG (?!)

← share via accounts / ssh-keys on common machines

(RAM)

→ home institutions
← share via accounts . . .
→ store individually

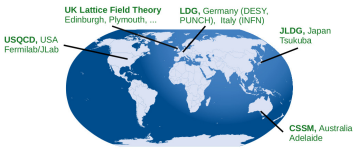
Requirements

- Software for generation (and analysis workflows):
  optimal use of expensive and heterogeneous HPC architectures
- Data management and curation: non-trivial effort
- Rich and flexible metadata: generic standards (e.g. DataCite) + community-specific
- Persistant and accessible storage $\geq$ 10 y: for re-use and RDM policies (e.g. DFG)
- Compliance with FAIR principles and modern data repository and publishing standards
- Training

Example: International Lattice Data Grid (ILDG)

- Effort by world-wide Lattice community [Lattice Conference 2002, 2004, ... 2022]
  - proposed in 2002, operational since 2007 (10 years before FAIR was termed)
  - usability and availability had severely degraded by 2020
  - re-activation and modernization since 2022 (with funding by PUNCH4NFDI)

- Basic concepts and elements
  - ☞ Single VO-wide AAI
  - ☞ Community-wide agreed standards for metadata, data format, and APIs
  - ☞ Autonomous "Regional Grids", each operating:
    - Metadata and File Catalogue (+Authorization Attribute Service)
    - Federated Storage Elements



An interesting (and not Lattice-specific) approach, but unlikely to be efficient and sustainable if realized just by physicists from within (individual and small) communities

Synergies: Computing

- algorithms (solvers, ML)
- optimization (e.g. GPU, FPGA)
- micro benchmarks and performance models (LQCD like "harmonic oscillator" for CS)
- analysis tools (autocorrelations)
- workflows (compatible with HPC centers)

Synergies: FAIR Data Storage and Management

- ❑ distributed storage infrastructure → file catalogue + FTS
- ❑ rich and searchable metadata → metadata catalogues
- ❑ embargoed and shared data → fine-grained AAI
- ❑ support for publishing process → definition of (meta)data content, DOI minting, LP, MD harvesting