The International Lattice Data Grid

Hubert Simma (DESY) and Carsten Urbach (Uni Bonn)

PUNCH4NFDI workshop on “open data”

Feb 11, 2021
For a given set of physical parameters $\alpha_{\text{phys}}$ (quark masses, ...)

- a Markov Chain ensemble $\{C_i\}$ is generated using MCMC, $i = 1, \ldots, N_{\text{cfg}}$
- each element $C_i$ (configuration) up to $O(200)$ Gigabytes large
- ensemble ($N_{\text{cfg}} \sim 10^4$) takes years to generate on largest HPC systems
- $C_i$ to good approximation random numbers, no compression possible
- many ensembles $O(20)$ needed to obtain physical results

⇒ ensembles highly valuable (and costly)!
⇒ ensembles the basis for a plethora of physical observables $Q$

Community concluded ensembles need to be shared!
LQCD Workflow (and jargon)

Simulation
HPC / performance

Measurement
HPC / throughput

Analysis

(1) "Simulation" or "Gauge field generation" by MCMC

(2) "Measurement" and average of (primary) observables $Q$

(3) "Statistical Analysis"

- extract physical quantities e.g. $\langle Q \rangle \sim e^{-tm}$
- combine different ensembles, fit, extrapolate, …
In practice, workflow more complex

→ tune, calibrate, cross-check, optimise, ...
Motivation

Highly valuable configurations $C_i$ should be shared internationally!

Time Line

$\approx 2001$ Plans to organize and set up infrastructures for community-wide sharing of configs

$\approx 2003$ ILDG architecture as a grid of (regional) grids

. Development of common metadata schema

. ” ” standard data format for configs

. Setup of regional grids and Virtual Organization

. Specification of interoperable web services

2008 Fully operational infrastructure and services

2013 Last (minimal) revision of QCDmlEnsemble schema
ILDG vs. Regional Grids (RG)

ILDG

- defines a grid of inter-operable (data-) grids
  (initially 5 regional grids: Japan, EU, UK, US, Australia)

- formally consists of
  - Virtual Organization (VO)
  - Specification of Services (File Catalog, Metadata Catalog)
  - Specification of Data Formats (Lime container)
  - Specification of Metadata Standards (Ensembles, Configs)
  - URLs of Services of each Regional Grid
  - Metadata and Middleware working groups, governance

- is implemented by
  - VOMS (VO)
  - Website (specifications and docu)
  - Board and working groups
Regional Grids

e.g. the Latfor Data Grid (LDG) in continental Europe

Implement and operate the following services

- Webpage = RG-specific info
- Metadata Catalog (MDC) = LDG-specific implementation
- File Catalog (FC) = WLCG middleware
- Data Storage (SE) = shared with WLCG

Backward compatibility is a must for ILDG!
Distributed Architecture of ILDG

Services

of ILDG

of each RG

website

VOMS

url

url

attr

MDC

FC

SE

SE

...
Distributed Architecture of ILDG

Services of ILDG:
- MDC
  - ensemble:
    - MCID
    - mgmt
    - phys
    - algo
  - config:
    - mcid
    - cid
    - mgmt
    - crc
    - prov
    - sim
- FC
  - config:
    - CID
    - turl
    - turl:
  - raw data:
    - fmt
    - CID
    - cfg
- SE
  - raw data:
    - fmt
    - CID
    - cfg

Metadata of each RG:
- acl
- url schemata
- url attr

Access:
- fully public
- possibly restricted
MD Schemata

Rich and extensible (20...∞ elements)

- Unique ID’s
- Revisions (who, when, why)
- Provenance (where, when, who)
- Integrity [and access]
- LQCD-specific info

MD Catalog (LDG Implementation)

- searchable (Xpath) → Findable
- publicly accessible (https) → Accessible
- standard SOAP interface (WSDL/axis) → Interoperable
- free MD schemata (XSD) → Reusable
- scalable data base (eXist)
  \[O(10^3)\] ensembles, \[O(10^6)\] configs
Summary

10 years ago ILDG has

- **realized** an idea which 2003 may have been ahead of the time
- completed a long and difficult community-wide process to **converge** to rich and flexible MD schemata (still adequate and FAIR compliant)
- implemented **stable** services with state-of-the-art technology (now becoming out-dated)

Today ILDG

- needs a thorough **redesign** based on modern data-lake concepts and up-to-date (cloud and web service) technologies
- but can provide a **working prototype** for building blocks and architecture of research data infrastructure
Outlook

Future (I)LDG directions within PUNCH

- replacement of grid certificates and VOMS by a more common AAI
- development of schemata for additional classes of research data
- modern interfaces and modules to easily set up (additional) interoperable services
- integration of newer cloud storage technologies
- support and integration of data transformations and workflows
- convenient user interfaces and tools to control workflows, to perform data curation, and to monitor services
- active coordination (board) and working groups
- integration into a well-designed and multi-layered architecture beyond the LQCD community