LATTICE QCD IN THE IO-SEA ENVIRONMENT

IO-SEA Workshop @ HIPEAC 2024

18 January 2024 | Eric B. Gregory | Forschungszentrum Jülich / JSC

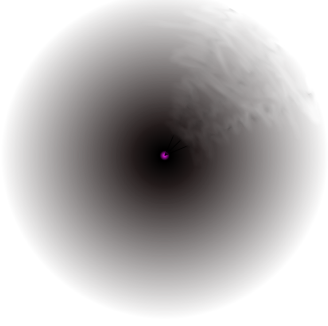
THE PLAN

- What is LQCD?
- LQCD workflow
- LQCD data challenges in the exascale era
- Running the LQCD workflow with the IO-SEA SBB service
- Results
- Summary

What is Lattice Quantum Chromo-Dynamics?

What is Lattice Quantum Chromo-Dynamics?

First a little physics:

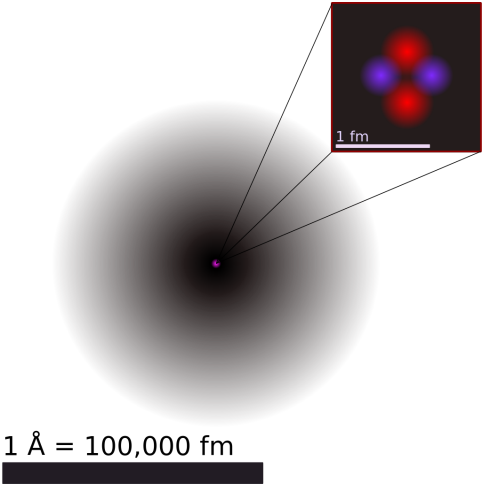


1 Å = 100,000 fm

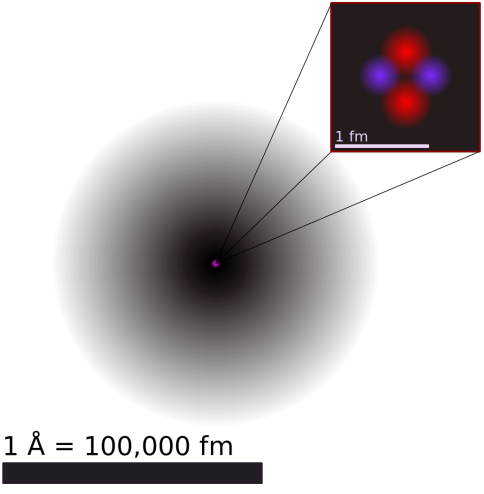
^{4}He

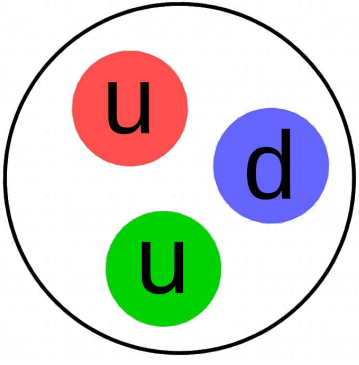
What is Lattice Quantum Chromo-Dynamics?

First a little physics:



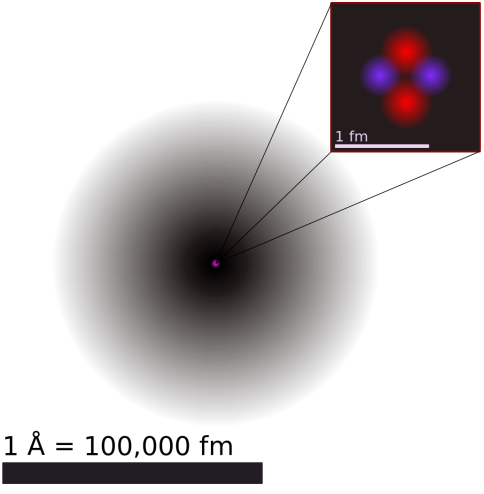
QCD

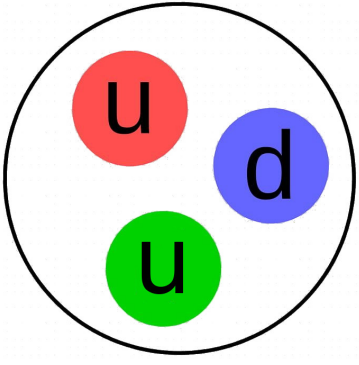




proton

QCD



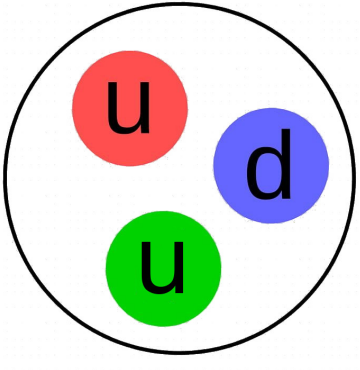


proton

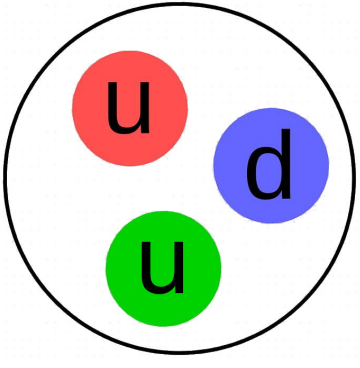
Proton is a *hadron*, a particle made of quarks bound together by the strong force.

QCD

We say protons have the *quantum numbers* of three quarks:



QCD

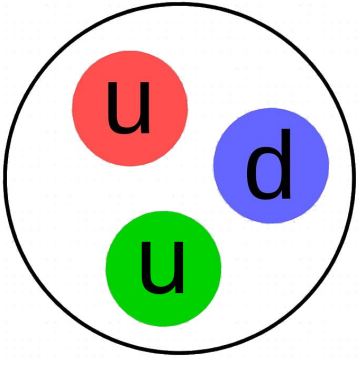


We say protons have the *quantum numbers* of three quarks:

Quarks have a color charge, so-called because:

red+blue+green= neutral

QCD

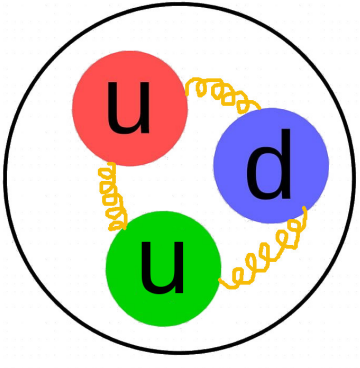


We say protons have the *quantum numbers* of three quarks:

Quarks have a color charge, so-called because:

red+blue+green= neutral

QCD



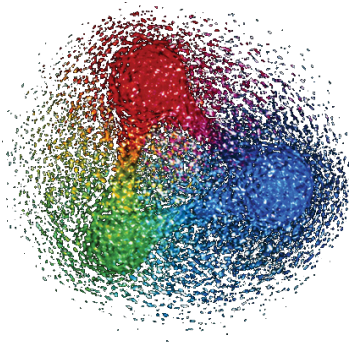
We say protons have the *quantum numbers* of three quarks....

Quarks have a color charge, so-called because:

red+blue+green= neutral

Hadrons also have gluons contributing to their properties.

QCD



We say protons have the *quantum numbers* of three quarks....

Quarks have a color charge, so-called because

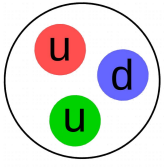
red+blue+green= neutral

Hadrons also have gluons contributing to their properties.

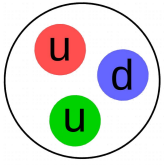
More useful picture includes *quantum* fluctuations: particle–anti-particle creation, annhilation, interaction, ...

QUANTUM FLUCTUATIONS ARE IMPORTANT!

QUANTUM FLUCTUATIONS ARE IMPORTANT!

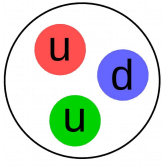


QUANTUM FLUCTUATIONS ARE IMPORTANT!



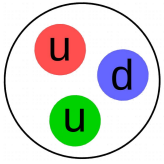
$$2 \times M_{\text{up}} \qquad + M_{\text{down}}$$

QUANTUM FLUCTUATIONS ARE IMPORTANT!



$$2 \times M_{\text{up}} \quad + M_{\text{down}}$$
$$2 \times (2.2\ \text{MeV}) \quad + (4.7\ \text{MeV}) \quad \approx 9\ \text{MeV}$$

QUANTUM FLUCTUATIONS ARE IMPORTANT!

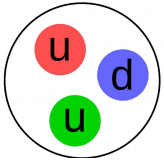


$$2 \times M_{\text{up}} + M_{\text{down}}$$
$$2 \times (2.2\ \text{MeV}) + (4.7\ \text{MeV}) \approx 9\ \text{MeV}$$

But ...

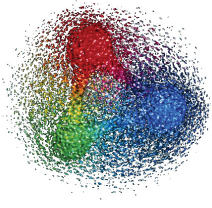
$$M_{\text{proton}} = 938\ \text{MeV}$$

QUANTUM FLUCTUATIONS ARE IMPORTANT!



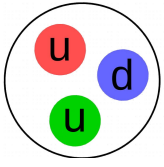
$$2 \times M_{\text{up}} \quad + M_{\text{down}}$$
$$2 \times (2.2\ \text{MeV}) \quad + (4.7\ \text{MeV}) \quad \approx 9\ \text{MeV}$$

But ...



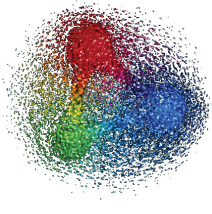
$$M_{\text{proton}} = 938\ \text{MeV}$$

QUANTUM FLUCTUATIONS ARE IMPORTANT!



$$2 \times M_{\text{up}} \quad + M_{\text{down}}$$
$$2 \times (2.2\ \text{MeV}) \quad + (4.7\ \text{MeV}) \quad \approx 9\ \text{MeV}$$

But ...



$$M_{\text{proton}} = 938\ \text{MeV}$$

To understand properties of hadrons, we must take quantum fluctuations into effect.

WHAT CAN WE HOPE TO CALCULATE?

WHAT CAN WE HOPE TO CALCULATE?

- **Properties of hadrons**
  - **mass**
  - **internal structure**
  - **decay probabilities**
  - **...**

WHAT CAN WE HOPE TO CALCULATE?

- **Properties of hadrons**
  - **mass**
  - **internal structure**
  - **decay probabilities**
  - **...**
- **Existence of unobserved states**

WHAT CAN WE HOPE TO CALCULATE?

- **Properties of hadrons**
  - **mass**
  - **internal structure**
  - **decay probabilities**
  - **...**
- **Existence of unobserved states**
- **BIG QUESTION:**

Does

$$\{\text{experiment}\} - \{\text{theory}\} \stackrel{?}{=} 0$$

WHAT CAN WE HOPE TO CALCULATE?

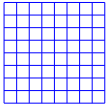
- **Properties of hadrons**
  - **mass**
  - **internal structure**
  - **decay probabilities**
  - **...**
- **Existence of unobserved states**
- **BIG QUESTION:**

Does

$$\{\text{experiment}\} - \{\text{theory}\} \stackrel{?}{=} 0$$

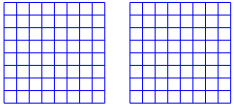
Physics beyond the Standard Model?

TYPICAL LQCD WORKFLOW



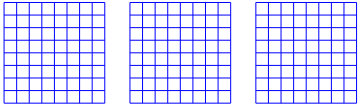
- Generate Markov chain of lattice gauge field configurations

TYPICAL LQCD WORKFLOW



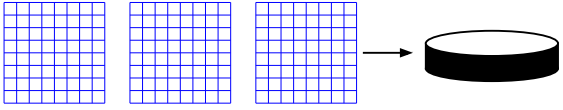
- Generate Markov chain of lattice gauge field configurations

TYPICAL LQCD WORKFLOW



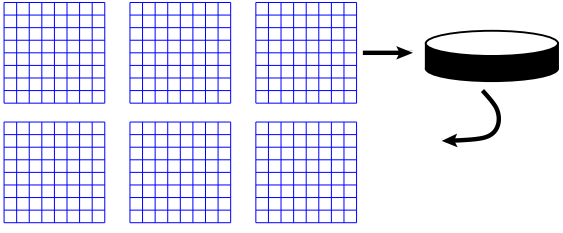
- Generate Markov chain of lattice gauge field configurations

TYPICAL LQCD WORKFLOW



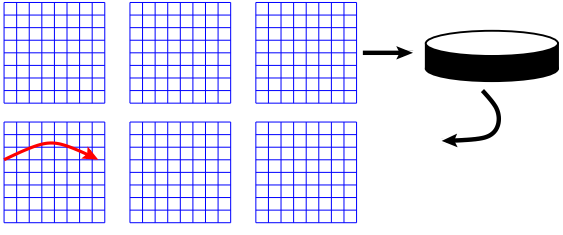
- Generate Markov chain of lattice gauge field configurations
- Save each to disk

TYPICAL LQCD WORKFLOW



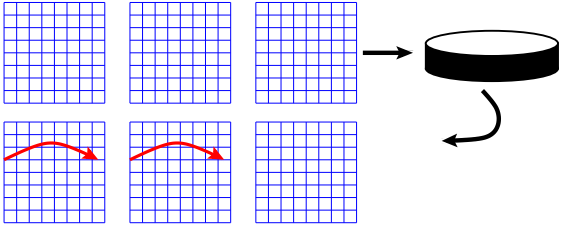
- Generate Markov chain of lattice gauge field configurations
- Save each to disk
- Load configuration files from disk

TYPICAL LQCD WORKFLOW



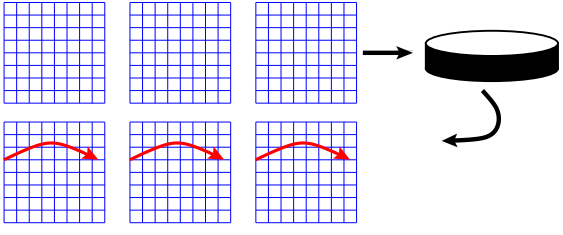
- Generate Markov chain of lattice gauge field configurations
- Save each to disk
- Load configuration files from disk
- On each, calculate quark propagators: $p(x) = M^{-1}(x, y)q(y)$

TYPICAL LQCD WORKFLOW



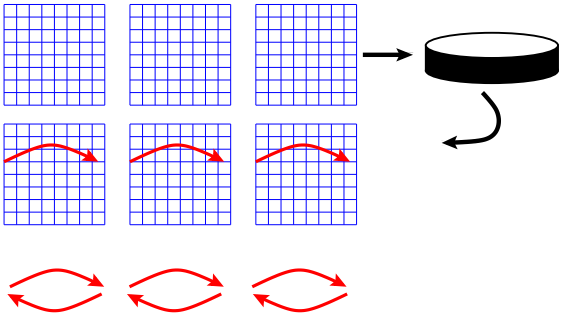
- Generate Markov chain of lattice gauge field configurations
- Save each to disk
- Load configuration files from disk
- On each, calculate quark propagators: $p(x) = M^{-1}(x, y)q(y)$

TYPICAL LQCD WORKFLOW



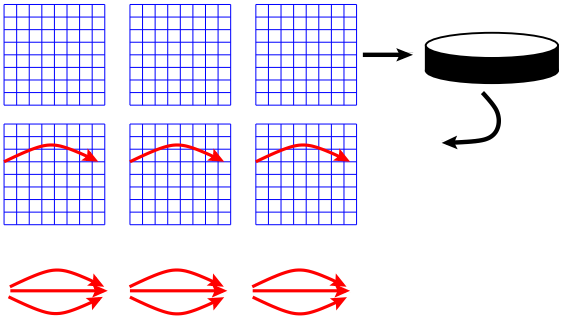
- Generate Markov chain of lattice gauge field configurations
- Save each to disk
- Load configuration files from disk
- On each, calculate quark propagators: $p(x) = M^{-1}(x, y)q(y)$

TYPICAL LQCD WORKFLOW



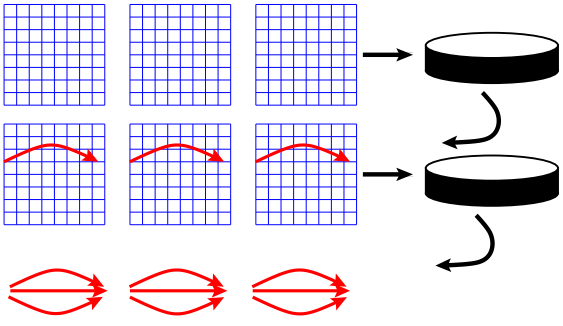
- Generate Markov chain of lattice gauge field configurations
- Save each to disk
- Load configuration files from disk
- On each, calculate quark propagators: $p(x) = M^{-1}(x, y)q(y)$
- Contract propagators to generate hadron correlators $C(t)$

TYPICAL LQCD WORKFLOW



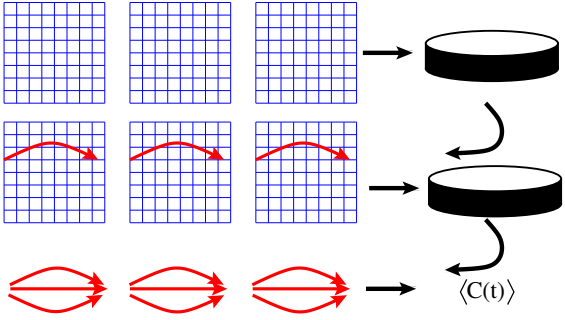
- Generate Markov chain of lattice gauge field configurations
- Save each to disk
- Load configuration files from disk
- On each, calculate quark propagators: $p(x) = M^{-1}(x, y)q(y)$
- Contract propagators to generate hadron correlators $C(t)$

TYPICAL LQCD WORKFLOW



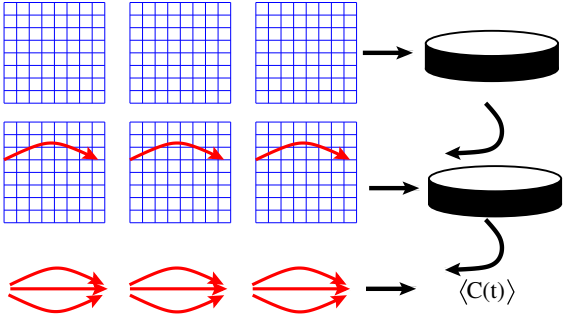
- Generate Markov chain of lattice gauge field configurations
- Save each to disk
- Load configuration files from disk
- On each, calculate quark propagators: $p(x) = M^{-1}(x, y)q(y)$
- Contract propagators to generate hadron correlators $C(t)$

TYPICAL LQCD WORKFLOW



- Generate Markov chain of lattice gauge field configurations
- Save each to disk
- Load configuration files from disk
- On each, calculate quark propagators: $p(x) = M^{-1}(x, y)q(y)$
- Contract propagators to generate hadron correlators $C(t)$
- Average over ensemble gives expectation value
- Fit correlators to extract physical quantities, e.g., hadron masses

TYPICAL LQCD WORKFLOW



- Generate Markov chain of lattice gauge field configurations
- Save each to disk
- Load configuration files from disk
- On each, calculate quark propagators: $p(x) = M^{-1}(x, y)q(y)$
- Contract propagators to generate hadron correlators $C(t)$
- Average over ensemble gives expectation value
- Fit correlators to extract physical quantities, e.g., hadron masses

Lots of data re-use!

LQCD I/O AND DATA CHALLENGES

Scaling motivation:

- Bigger lattice size: $V = N_x^3 \times N_t$ $\longrightarrow$ smaller systematic uncertainties
- More lattice gauge field configurations N_{conf} $\longrightarrow$ smaller statistical uncertianties
- More points in parameter space (m_q, β) $\longrightarrow$ smaller systematic uncertanties
- More propagators, contractions per config $\longrightarrow$ smaller statistical uncertianties, ...
  ... or new physics results

LQCD I/O AND DATA CHALLENGES

Scaling motivation:

- Bigger lattice size: $V = N_x^3 \times N_t$ $\longrightarrow$ smaller systematic uncertainties
- More lattice gauge field configurations N_{conf} $\longrightarrow$ smaller statistical uncertianties
- More points in parameter space (m_q, β) $\longrightarrow$ smaller systematic uncertanties
- More propagators, contractions per config $\longrightarrow$ smaller statistical uncertianties, ...
  ... or new physics results

Scaling opportunities:

- LQCD data and algorithms are very homogeneous
  many available levels of concurrencies.
- Algorithms are highly scalable.

LQCD I/O AND DATA CHALLENGES

If given the opportunity, LQCD researchers will use machine capacity in pursuit of more precise/accurate results.

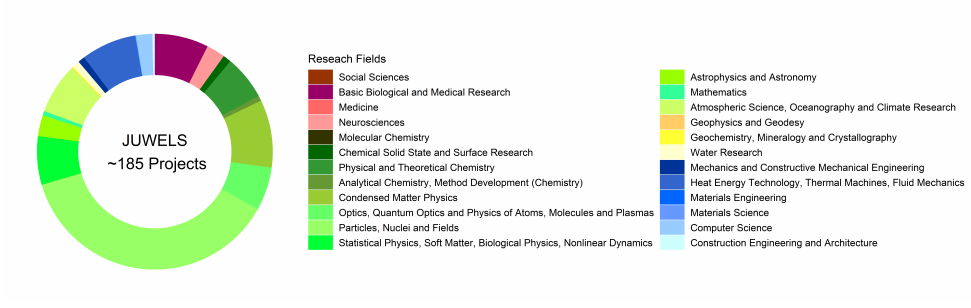
LQCD I/O AND DATA CHALLENGES

If given the opportunity, LQCD researchers will use machine capacity in pursuit of more precise/accurate results.

- Strong scaling $\longrightarrow$ I/O operations are more frequent.
- Weak scaling $\longrightarrow$ I/O operations are larger.

LQCD I/O AND DATA CHALLENGES

If given the opportunity, LQCD researchers will use machine capacity in pursuit of more precise/accurate results.



- Strong scaling $\longrightarrow$ I/O operations are more frequent.
- Weak scaling $\longrightarrow$ I/O operations are larger.

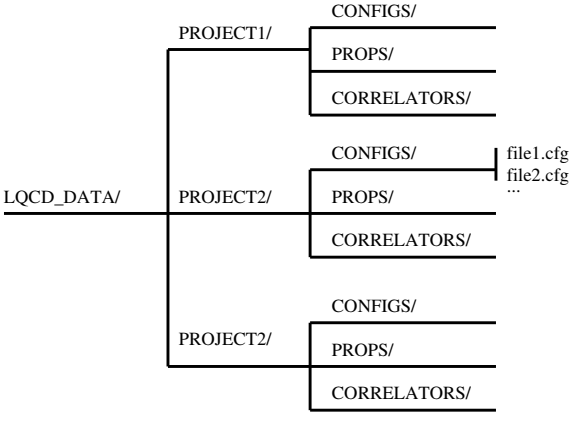
$\longrightarrow$ I/O and data management issues will be magnified

LQCD AND THE IO-SEA SOLUTION

The Smart Burst Buffer ephemeral service

- Tool for I/O optimization, rather then data-management
- Requires no code or input modification
- Can be applied to existing directory structure
- Minimizes data traffic to and from disk during workflow
- Maximizes “data re-use”
- Can isolate workflow from a busy storage system

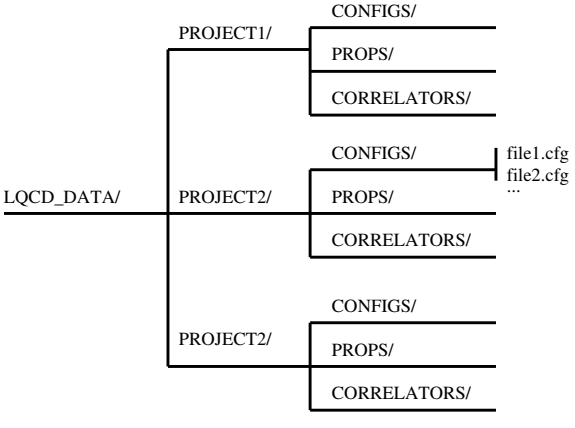
THE SMART BURST BUFFER EPHEMERAL SERVICE



THE SMART BURST BUFFER EPHEMERAL SERVICE

Specify a target directory, e.g.:

<full-path-to>/LQCD_DATA/PROJECT2/

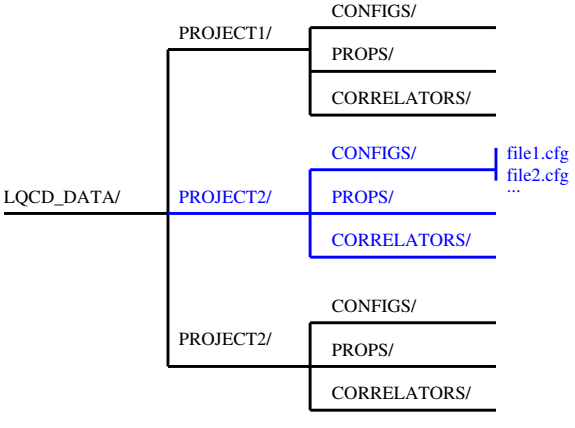


THE SMART BURST BUFFER EPHEMERAL SERVICE

Specify a target directory, e.g.:

`<full-path-to>/LQCD_DATA/PROJECT2/`

Data *reads* and *writes* in target directory and its sub-directories are intercepted by the SBB.



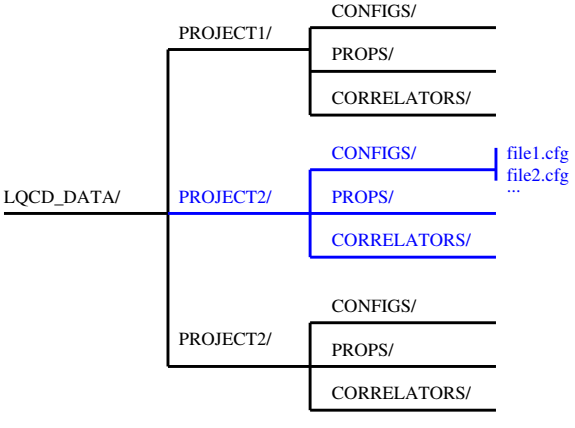
THE SMART BURST BUFFER EPHEMERAL SERVICE

Specify a target directory, e.g.:

`<full-path-to>/LQCD_DATA/PROJECT2/`

Data *reads* and *writes* in target directory and its sub-directories are intercepted by the SBB.

Output is stored in fast RAM or NVMe storage, ready for re-use or to be flushed to disk when convenient.



HOW DO WE DO IT?

HOW DO WE DO IT?

Assume I already have Slurm batch scripts prepared for three workflow steps:

HOW DO WE DO IT?

Assume I already have Slurm batch scripts prepared for three workflow steps:

STEP	READS IN	WRITES OUT

HOW DO WE DO IT?

Assume I already have Slurm batch scripts prepared for three workflow steps:

STEP	READS IN	WRITES OUT
A	initial configuration file	Markov chain of configs

HOW DO WE DO IT?

Assume I already have Slurm batch scripts prepared for three workflow steps:

STEP	READS IN	WRITES OUT
A	initial configuration file	Markov chain of configs
B	1 configuration file	multiple propagator files

HOW DO WE DO IT?

Assume I already have Slurm batch scripts prepared for three workflow steps:

STEP	READS IN	WRITES OUT
A	initial configuration file	Markov chain of configs
B	1 configuration file	multiple propagator files
C	1 configuration groups of propagators	correlation arrays

HOW DO WE DO IT?

Assume I already have Slurm batch scripts prepared for three workflow steps:

STEP	READS IN	WRITES OUT
A	initial configuration file	Markov chain of configs
B	1 configuration file	multiple propagator files
C	1 configuration groups of propagators	correlation arrays

Note that we have data re-use in the lattice gauge field configuration files and the propagator files.

HOW DO WE DO IT?

Assume I already have Slurm batch scripts prepared for three workflow steps:

STEP	READS IN	WRITES OUT
A	initial configuration file	Markov chain of configs
B	1 configuration file	multiple propagator files
C	1 configuration groups of propagators	correlation arrays

Note that we have data re-use in the lattice gauge field configuration files and the propagator files.

HOW DO WE DO IT?

Assume I already have Slurm batch scripts prepared for three workflow steps:

STEP	READS IN	WRITES OUT
A	initial configuration file	Markov chain of configs
B	1 configuration file	multiple propagator files
C	1 configuration groups of propagators	correlation arrays

Note that we have data re-use in the lattice gauge field configuration files and the propagator files.

Since you were wondering, in our tests:

configuration file	1.2GB
propagator	4.6 GB
correllator	few 10^2 kB

... but could be *much* bigger in modern production runs!

PREPARE A WDF

Workflow Description File is a YAML-format file describing how ephemeral services will be used by the workflow steps.

PREPARE A WDF

Workflow Description File is a YAML-format file describing how ephemeral services will be used by the workflow steps.

```yaml
workflow:
  name: LQCD_ABC

services:
  - name: lqcd-sbb1
    type: SBB
    attributes:
      targets: /afsm/iosea/LQCD_DATA/PROJECT2/
      flavor: high
      datanodes: 4
      location: dp-esb

steps:
  - name: step_A
    command: "sbatch ~/sub_LQCD_stepA.sh"
    services:
      - name: lqcd-sbb1

  - name: steps_B_and_C
    command: "sbatch ~/sub_LQCD_stepBC.sh"
    services:
      - name: lqcd-sbb1
```

PREPARE A WDF

Workflow Description File is a YAML-format file describing how ephemeral services will be used by the workflow steps.

```yaml
workflow:
  name: LQCD_ABC

services:
  - name: lqcd-sbb1
    type: SBB
    attributes:
      targets: /afsm/iosea/LQCD_DATA/PROJECT2/
      flavor: high
      datanodes: 4
      location: dp-esb

steps:
  - name: step_A
    command: "sbatch ~/sub_LQCD_stepA.sh"
    services:
      - name: lqcd-sbb1

  - name: steps_B_and_C
    command: "sbatch ~/sub_LQCD_stepBC.sh"
    services:
      - name: lqcd-sbb1
```

← Configures the SBB service

PREPARE A WDF

Workflow Description File is a YAML-format file describing how ephemeral services will be used by the workflow steps.

```yaml
workflow:
  name: LQCD_ABC

services:
  - name: lqcd-sbb1
    type: SBB
    attributes:
      targets: /afsm/iosea/LQCD_DATA/PROJECT2/
      flavor: high
      datanodes: 4
      location: dp-esb

steps:
  - name: step_A
    command: "sbatch ~/sub_LQCD_stepA.sh"
    services:
      - name: lqcd-sbb1

  - name: steps_B_and_C
    command: "sbatch ~/sub_LQCD_stepBC.sh"
    services:
      - name: lqcd-sbb1
```

⟵ Configures the SBB service

⟵ Instructions for launching step

PREPARE A WDF

Workflow Description File is a YAML-format file describing how ephemeral services will be used by the workflow steps.

```yaml
workflow:
  name: LQCD_ABC

services:
  - name: lqcd-sbb1
    type: SBB
    attributes:
      targets: /afsm/iosea/LQCD_DATA/PROJECT2/
      flavor: high
      datanodes: 4
      location: dp-esb

steps:
  - name: step_A
    command: "sbatch ~/sub_LQCD_stepA.sh"
    services:
      - name: lqcd-sbb1

  - name: steps_B_and_C
    command: "sbatch ~/sub_LQCD_stepBC.sh"
    services:
      - name: lqcd-sbb1
```

⟵ Configures the SBB service

⟵ Instructions for launching step

⟵ Associates ephemeral service with the step

STARTING THE WORKFLOW SESSION

In current implementation, each user must have an instance of the persistent API server running in the background:

STARTING THE WORKFLOW SESSION

In current implementation, each user must have an instance of the persistent API server running in the background:

```
> wfm-api &
```

STARTING THE WORKFLOW SESSION

In current implementation, each user must have an instance of the persistent API server running in the background:

```
> wfm-api &
```

Use the prepared workflow description file `lqcd_wdf.yaml` , to initialize the workflow:

```
> iosea-wf start -w lqcd_wdf.yaml -s lqcd_00
```

STARTING THE WORKFLOW SESSION

In current implementation, each user must have an instance of the persistent API server running in the background:

```
> wfm-api &
```

Use the prepared workflow description file `lqcd_wdf.yaml` , to initialize the workflow:

```
> iosea-wf start -w lqcd_wdf.yaml -s lqcd_00
```

Before running any steps, check that the service is started:

STARTING THE WORKFLOW SESSION

In current implementation, each user must have an instance of the persistent API server running in the background:

```
> wfm-api &
```

Use the prepared workflow description file `lqcd_wdf.yaml` , to initialize the workflow:

```
> iosea-wf start -w lqcd_wdf.yaml -s lqcd_00
```

Before running any steps, check that the service is started:

```
> iosea-wf status -s lqcd_00
SESSION                 WORKFLOW                    STATUS
lqcd_00                 LQCD_WFM_ABC                starting
```

STARTING THE WORKFLOW SESSION

In current implementation, each user must have an instance of the persistent API server running in the background:

```
> wfm-api &
```

Use the prepared workflow description file `lqcd_wdf.yaml` , to initialize the workflow:

```
> iosea-wf start -w lqcd_wdf.yaml -s lqcd_00
```

Before running any steps, check that the service is started:

```
> iosea-wf status -s lqcd_00
SESSION                 WORKFLOW                    STATUS
lqcd_00                 LQCD_WFM_ABC                starting

> iosea-wf status -s lqcd_00
SESSION                 WORKFLOW                    STATUS
lqcd_00                 LQCD_WFM_ABC                active
```

STARTING THE WORKFLOW SESSION

In current implementation, each user must have an instance of the persistent API server running in the background:

```
> wfm-api &
```

Use the prepared workflow description file `lqcd_wdf.yaml` , to initialize the workflow:

```
> iosea-wf start -w lqcd_wdf.yaml -s lqcd_00
```

Before running any steps, check that the service is started:

```
> iosea-wf status -s lqcd_00
SESSION                 WORKFLOW                    STATUS
lqcd_00                 LQCD_WFM_ABC                starting

> iosea-wf status -s lqcd_00
SESSION                 WORKFLOW                    STATUS
lqcd_00                 LQCD_WFM_ABC                active
```

Good to go!

RUNNING THE WORKFLOW STEP

RUNNING THE WORKFLOW STEP

```
> iosea-wf run -s lqcd_00 -t step_A
```

RUNNING THE WORKFLOW STEP

```
> iosea-wf run -s lqcd_00 -t step_A
```

Issues the slurm command described in the WDF for `step_A`:

```
sbatch ~/sub_LQCD_stepA.sh
```

RUNNING THE WORKFLOW STEP

```
> iosea-wf run -s lqcd_00 -t step_A
```

Issues the slurm command described in the WDF for `step_A`:

```
sbatch ~/sub_LQCD_stepA.sh
```

- Can run multiple instances of the same step concurrently.
- Current WMF version does not yet handle complicated job dependencies

RUNNING THE WORKFLOW STEP

```
> iosea-wf run -s lqcd_00 -t step_A
```

Issues the slurm command described in the WDF for `step_A`:

```
sbatch ~/sub_LQCD_stepA.sh
```

- Can run multiple instances of the same step concurrently.
- Current WMF version does not yet handle complicated job dependencies

```
> iosea-wf run -s lqcd_00 -t step_B_and_C
```

RUNNING THE WORKFLOW STEP

```
> iosea-wf run -s lqcd_00 -t step_A
```

Issues the slurm command described in the WDF for `step_A`:

```
sbatch ~/sub_LQCD_stepA.sh
```

- Can run multiple instances of the same step concurrently.
- Current WMF version does not yet handle complicated job dependencies

```
> iosea-wf run -s lqcd_00 -t step_B_and_C
```

Burst-buffer space is finite, so we should free the resource by stopping the session when the runs are complete:

```
iosea-wf stop -s lqcd_00
```

IO-INSTRUMENTATION

Running a workflow through the WFM activates IOI by default:

Many metrics available detailing I/O performance.

WHAT HAVE WE GAINED?

WHAT HAVE WE GAINED?

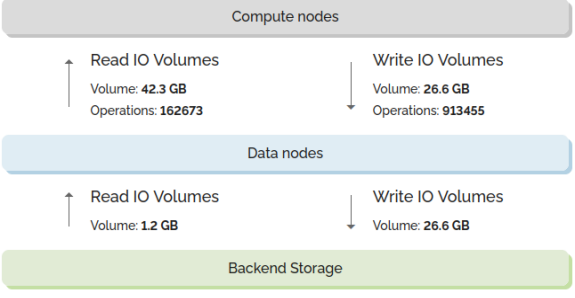
STEP	READS IN	WRITES OUT
A	1 config 1.2 GB	6 configs (6×1.2 GB) 7.2 GB
B	1 config 1.2 GB	4 props (4×4.8 GB) $\approx$ 19.3 GB
C	1 config 1.2 GB	
	8 props ($2 \times 4 \times 4.8$ GB) 38.7 GB	
TOTAL	42.3 GB	26.6 GB

WHAT HAVE WE GAINED?

Economy of data movement!

STEP	READS IN	WRITES OUT
A	1 config 1.2 GB	6 configs (6 × 1.2 GB) 7.2 GB
B	1 config 1.2 GB	4 props (4 × 4.8 GB) ≈ 19.3 GB
C	1 config 1.2 GB	
	8 props (2 × 4 × 4.8 GB) 38.7 GB	
TOTAL	42.3 GB	26.6 GB

IOI says:

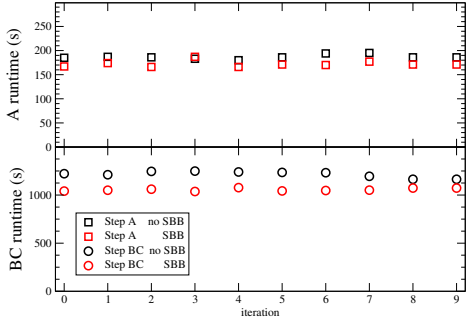


PERFORMANCE

Test setup is not for high performance runs:

- Proof-of-concept of IO-SEA environment
- Datanodes are ∼9 year old hardware
- Connected to DEEP by two long IB cables

Nevertheless, performance improvement is visible compared to direct accfess to JSC NFS storage.



Performance with the SBB service does not beat direct access to the all-flash storage module.

TESTS ON A BUSY STORAGE SYSTEM

TESTS ON A BUSY STORAGE SYSTEM

- Reserve the system

TESTS ON A BUSY STORAGE SYSTEM

- Reserve the system
- Start *N* instances of IOR benchmark —
  loads the storage with constant, repeated writes

TESTS ON A BUSY STORAGE SYSTEM

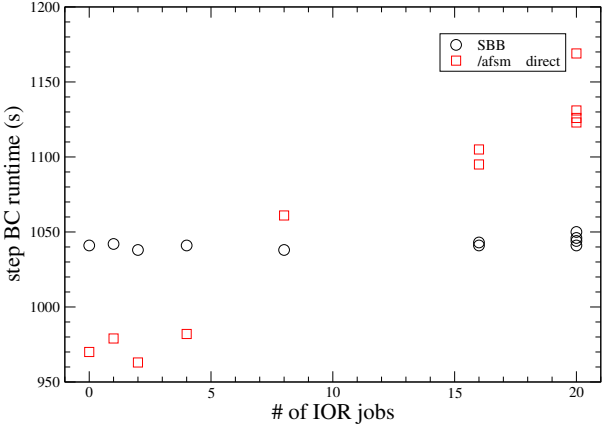
- Reserve the system
- Start *N* instances of IOR benchmark —
  loads the storage with constant, repeated writes
- Run the LQCD workflow with and without the SBB service

TESTS ON A BUSY STORAGE SYSTEM

- Reserve the system
- Start *N* instances of IOR benchmark —
  loads the storage with constant, repeated writes
- Run the LQCD workflow with and without the SBB service
- Repeat with increasing *N*

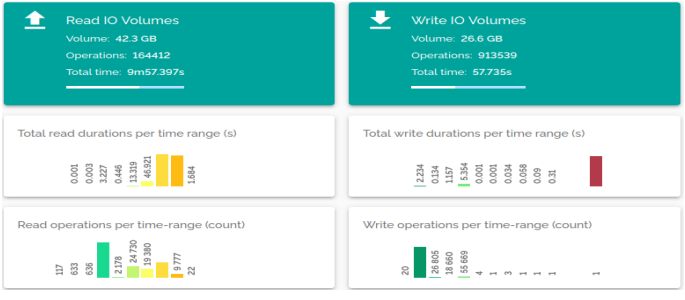
TESTS ON A BUSY STORAGE SYSTEM

- Reserve the system
- Start *N* instances of IOR benchmark —
  loads the storage with constant, repeated writes
- Run the LQCD workflow with and without the SBB service
- Repeat with increasing *N*

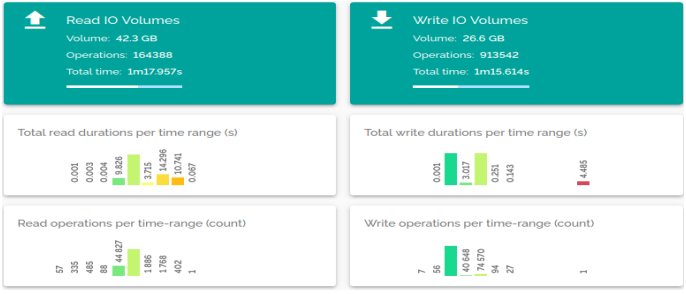


TESTS ON A BUSY STORAGE SYSTEM (20 × IOR)

Direct
AFSM



SBB



A WORD ABOUT OTHER IO-SEA SERVICES

NFS Service

- Early tests show surprisingly good performance
- Use of *datasets* will have data-management advantages

DASI

- Will not be able to use directly in the LQCD application
- May be useful in archiving systems for re-usable LQCD datasets

SUMMARY

- IO-SEA SBB service provides optimization of data movement when a workflow has significant data re-use.
- IO-SEA SBB service is a private storage service that can insulate a workflow from a busy back-end filesystem
- High data re-use in typical LQCD workflows make it ideal for SBB optimization.

EXTRA — Tests on a busy storage system

TESTS ON A BUSY STORAGE SYSTEM

Direct AFSM

SBB

unperturbed

perturbed (20x IOR)