

Application server / Containers as a service / MLOps

Materials for discussion

Denis Boyda, IAIFI
January 06, 2023

Outlook

- Who am I / why am I here
- Needs for new services at supercomputer facilities
- Machine Learning Model Operationalization Management (MLOps)
- Experiment organization overview
- Sketch of Cluster Organization
- Application server scheme



Who am I? Why am I here?

- worked a few years at Argonne Leadership Computing Facility on scaling ML for Science
- *currently working on ML for LQCD ESP project for simulation at Aurora (~ 2 exaflops)*
- part of **MIT lattice QCD group**
 - working with Phiala Shanahan and William Detmold
 - Yin Lin is a representative of the group
- currently **IAIFI postdoc fellow!**
(just arrived)



2022-2025 IAIFI Fellows

We are excited for [Denis](#), [Carolina](#), and [Jessie](#) to join us as our second round of IAIFI Fellows to help spark vital interdisciplinary research at the intersection of Physics and AI!

Denis Boyda

Research Interests: Denis Boyda has been working on the application of the Machine Learning method to simulations of physical systems and bringing physical ideas to Machine Learning. His research is devoted to developing algorithms enabling simulations of nuclear and particle physics which are currently computationally intractable. Denis Boyda is interested in the Monte Carlo techniques and generation modeling. He develops equivariant models which respect the symmetry of a target problem and runs simulations at leading supercomputer machines.



Needs for new services at supercomputer services - 1

Strategy for ML for LQCD software development

Measurements

observable regression
© Thom T. Spharshakya, R. Gupta, Phys. Rev. D 100, 014004 (2019)

- Observables calculation takes similar or larger than generation resources
- Use ML regression to compute them faster
 - N configurations, N_{obs} measurements of O , N_{obs} measurements are used for training
 - effective samples size is increased from N_{obs} to N
- Correct bias

$$\hat{O} = \frac{1}{N - N_{\text{obs}}} \sum_{i=1}^{N_{\text{obs}}} O_i^{\text{ML}} + \frac{1}{N_{\text{obs}}} \sum_{i=1}^{N_{\text{obs}}} (O_i - O_i^{\text{ML}})$$

Can we decrease time for observable measurements and get higher statistics?

Configuration generation with normalizing flows

Flow-based models learn a sequence of operations that transform a known distribution to the desired one

$\int r(V) dV$ $U = f(V)$ $\int q(U) \left| \det \frac{\partial f^{-1}(U)}{\partial U} \right| dU$

Exactness guarantees is done via correlation in (U) using emerging of building MCMC chain

Normalizing flows: better uncertainty qualification at finite statistics

Lattice Schwinger model near criticality

Determination of topological mixing is difficult due to UV fluctuations

Example of improving uncertainty estimation for false observables

Normalizing flows: better mixing

U(1) lattice gauge theory in 2D

Example of under sampling of some region of target probability density

Example of improving topological sampling with NF

Increase signal to noise ratio via contour deformation

Variable transformation does not change integral

$$\langle O \rangle = \int \mathcal{D}\phi e^{-S(\phi)} O(\phi) = \int \mathcal{D}U f(U) e^{-S(U)} O(\phi(U))$$

$$\langle O \rangle = \langle O \rangle = \langle f(U) e^{-S(U)} O(\phi(U)) \rangle$$

but changes uncertainties

transformation $\tilde{O} = f(O)$ is optimized such that $\text{var}(\tilde{O}) \ll \text{var}(O)$

Can we apply it for viscosity computations in full QCD?

Compute QCD phase diagram in (T, μ) with normalizing flows

Direct MCMC simulations of QCD at nonzero chemical potential is not tractable due to Sign Problem

Several approaches use MCMC simulations at zero and/or imaginary chemical potential

Simulations at several values of imaginary chemical potential required in order to do extrapolation to real region

After training Normalizing flow model gives access to "all" values of imaginary chemical potential

Can we get larger chemical potential?

Thermodynamical properties of QGP and EoS with normalizing flows

Lattice scalar field theory

The fundamental difficulty is that MCMC is not able to directly estimate the partition function of the lattice field theory.

Normalizing flows have direct access to partition function

$$Z = \int \mathcal{D}\phi_{ij}(\theta) \frac{e^{-S(\theta)}}{\mathcal{Q}(\theta)} \left(\frac{e^{-S(\theta)}}{\mathcal{Q}(\theta)} \right)_{\theta_{\text{ML}}}$$

Can we compute QCD EoS with higher precision?

Reconstructing QCD Spectral Functions with Gaussian Processes

Spectral functions are extracted from lattice QCD correlator using inverse integral transformation which is ill-defined problem

Reconstruction using Gaussian Process Regression

what is most probably value and uncertainty of $f(\omega)$ given some observations $\xi_i(t_i)$ with uncertainties

Did we improve a solution of inverse problem?

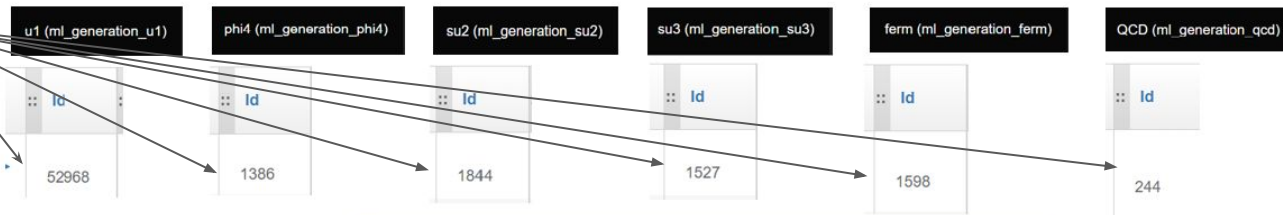
Slides credit: Denis Boyda, Machine Learning techniques in lattice QCD, 2022 RHIC/AGS Annual Users' Meeting, June 7, 2022

Needs for new services at supercomputer services - 2

Scaling the number of ML experiments / MLOPs

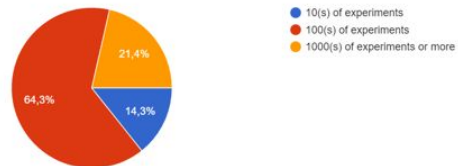
Number of ML experiments in our data bases. Development of flow-based models for

Number of ML experiments



Recent review of users of Argonne Leadership Computing Facility

At what scale do you run or plan to run ML applications and associated experiments



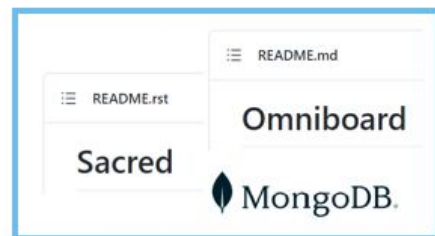
Scaling the number of ML experiments / MLOPs - 2

MLOps (Machine Learning Model Operationalization Management) provides

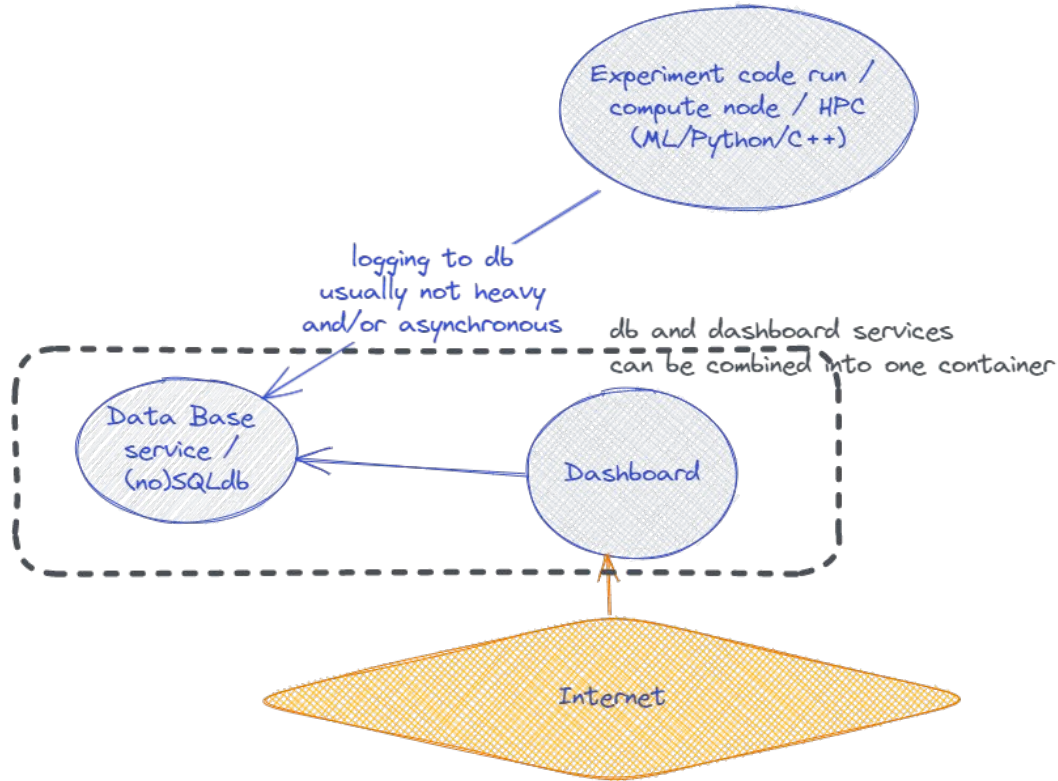
- Hyperparameters/configuration tracking
- Live information (stdout, stderr, results)
- Artifacts (models, datasets) control and versioning
- Code control and versioning
- Environment configuration
- Fail trace

and an efficient way of analyzing experiments though

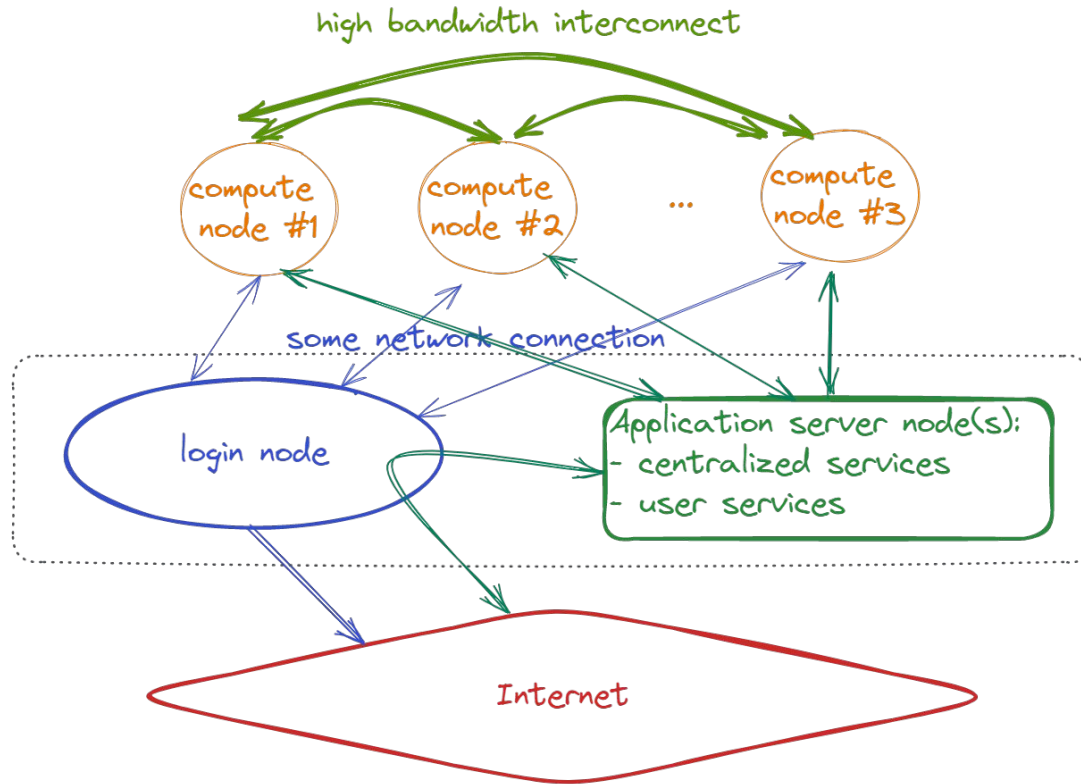
- Dashboard
- API to database



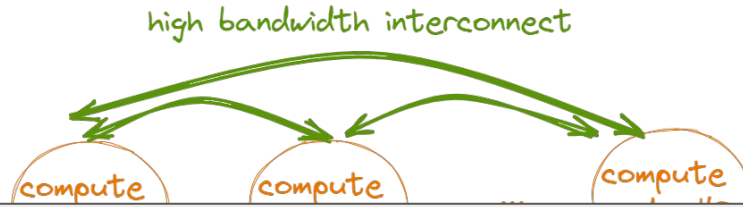
Experiment Organization Overview



Sketch of Cluster Organization

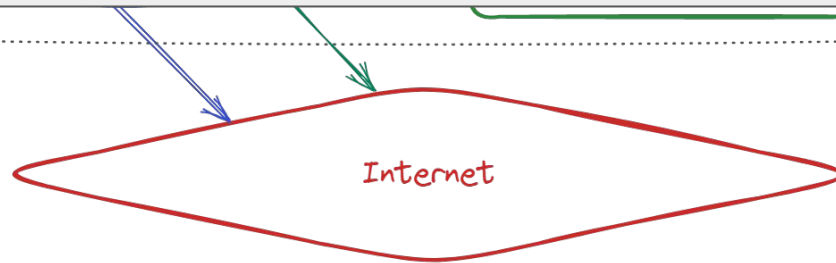


Sketch of Cluster Organization



Main difference between compute nodes and application server is

- users use **compute nodes** with fixed resources for a fixed time
- at **application server** services takes small resources but all the time



Application server scheme

In centralized fashion services/applications control users - aka “users under application”

In user fashion users control applications stack - aka “applications under users”

Idea for prototype: run a docker on dedicated nodes and allow groups/users to run certain number of containers

How can we set up application server at submit?

