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)

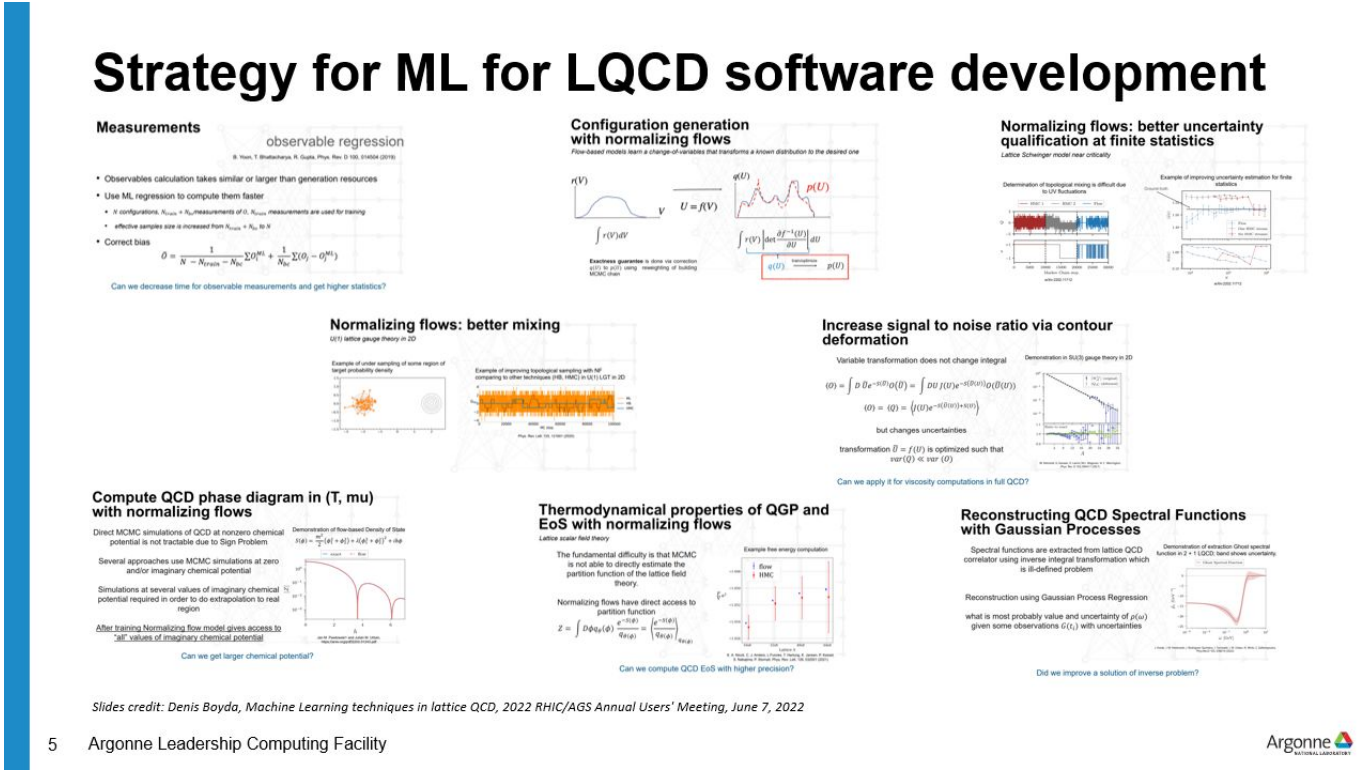
2022-2025 IAIFI Fellows

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



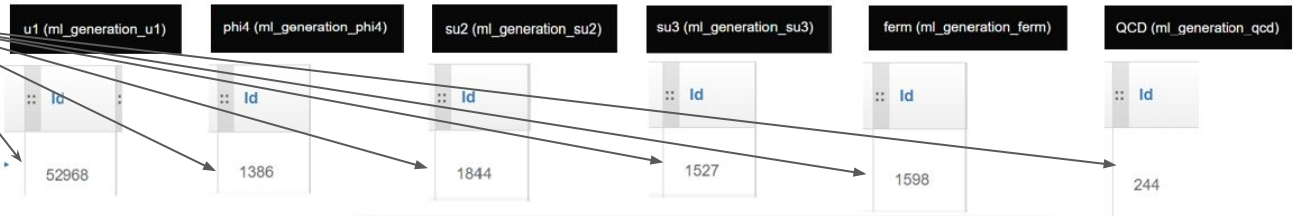
Slides credit: Denis Boyda, Machine Learning techniques in lattice QCD, 2022 RHIC/AGS Annual Users' Meeting, June 7, 2022

Needs for new services at supercomputer services - 2

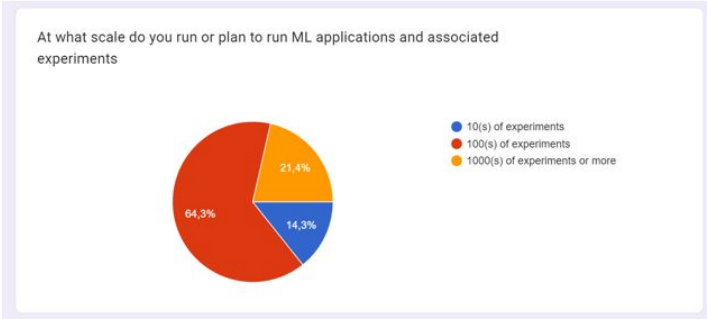
Number of ML experiments

Scaling the number of ML experiments / MLOPs

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



Recent review of users of Argonne Leadership Computer Facility



Argonne NATIONAL LABORATORY

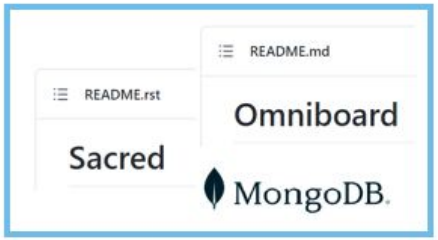
Scaling the number of ML experiments / MLOPs - 2

MLOps (Machine Learning Model Operationalization Management) provides

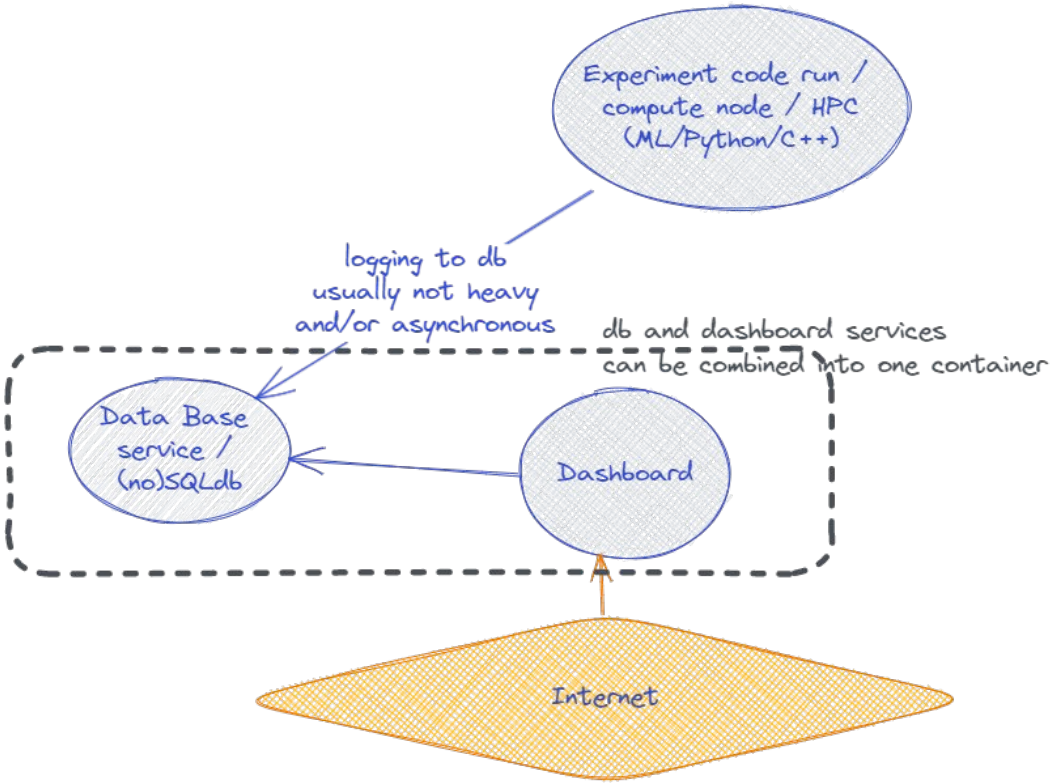
- Hyperparameters/configuration tracking
- Live information (stdout, stderr, results)
- Artifacts (models, datasets) control and versioning
- Code control and versioning
- Environment configuration
- Fail trace

and an efficient way of analyzing experiments though

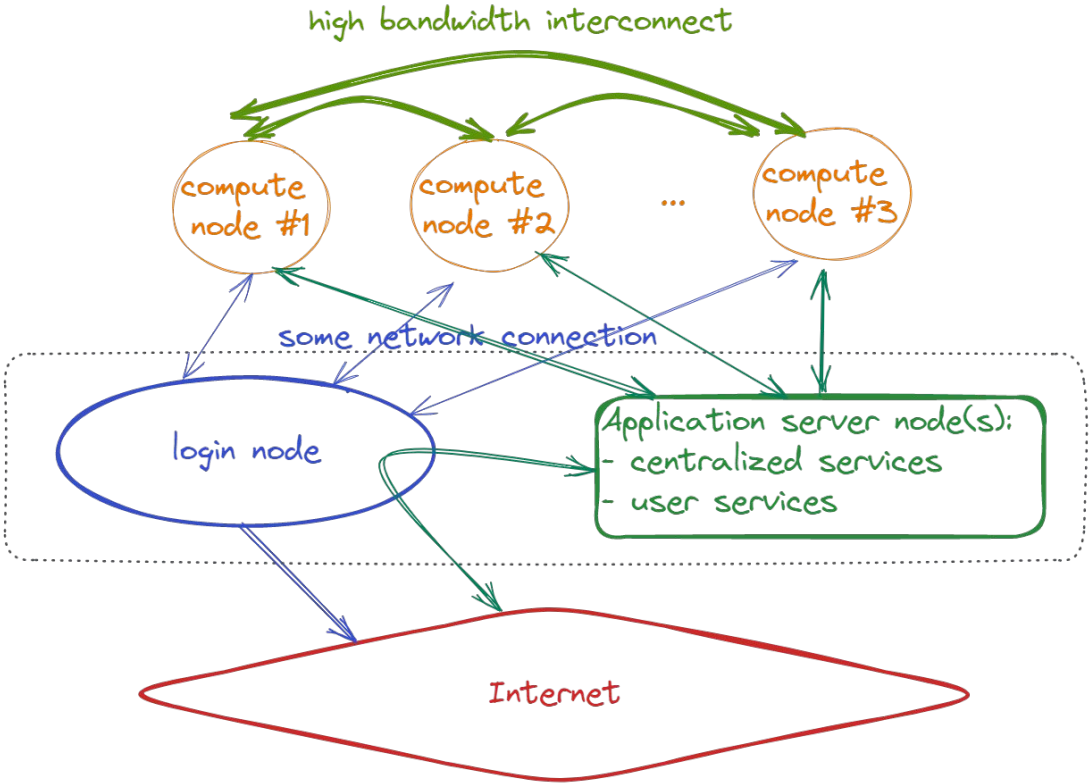
- Dashboard
- API to database



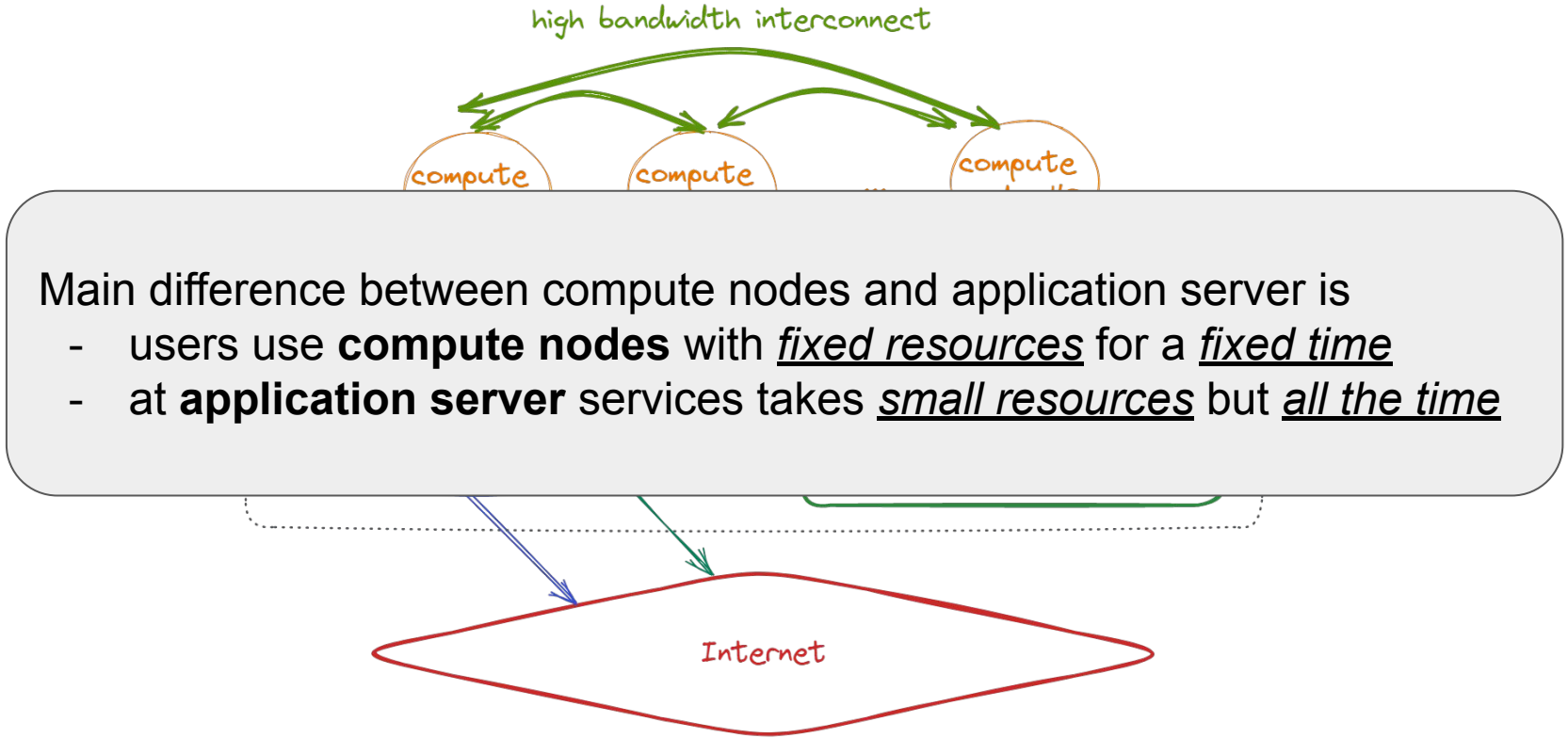
Experiment Organization Overview



Sketch of Cluster Organization



Sketch of Cluster Organization



Main difference between compute nodes and application server is

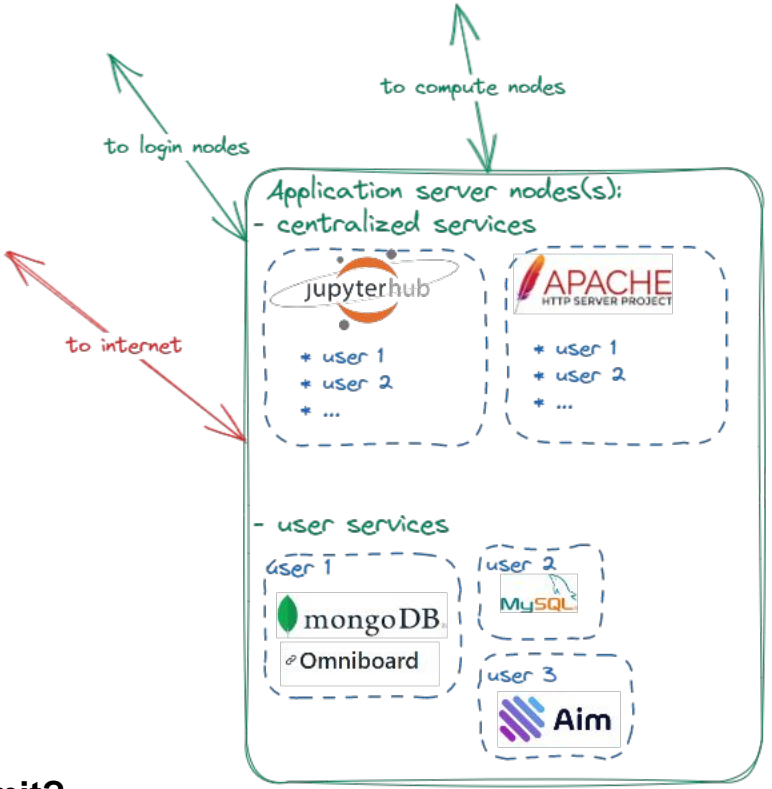
- users use **compute nodes** with *fixed resources* for a *fixed time*
- at **application server** services takes *small resources* but *all the time*

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?