SC26 Network Research Exhibition: Demonstration Preliminary Abstract

High Value, Petabyte-scale Data: Our Safe Place for Academic Research in the Era of AI

Joe Biondo, Lance Joseph, Alex Szalay, The Johns Hopkins University

The challenge

The success of AlphaFold demonstrated the importance of AI in sciences. However, the project could not have succeeded without the pdb, a large, open database of genomic and spatial structure of proteins. This database took almost 20 years and may billions of dollars to assemble. Google would never have considered to create such a dataset for the sake of performing AlphaFold.

This illustrates an important point: in the era of AI it is impossible for universities, academic institutions to compete with the hyoerscalars, who invest hundreds of billions every year in hardware. For this reason, creating unique, high value, long lived data sets like the pdb remains a safe place for academics. Yet, universities do not have 20 years or 20 $B foir this task, either.

The question is, how can we create a new, agile environment at our research institutions which can accelerate and support the creation of such data sets, but at a much lower cost.

The Johns Hopkins Platform

Two years ago Johns Hopkins started a new central effort to build up a large body of faculty specializing on AI. Given limited resources we were faced with the question of how should we build out our infrastructure that is capable of supporting academic AI research, can evolve, and can be sustainable, and do this is as short a time as possible.

Satisfying all these criteria simultaneously seemed impossible. We only had a 1.1MW container with air cooling for HPC and two other locations with a few hundred kW for lower powered storage.

On the other hand, we had a major unique advantage, JHU owns essentially all the dark fiber connecting our Baltimore campuses, and the long haul fiber connecting us to the Mid-Atlantic Crossroads (MAX)at College Park and the Ashburn, VA, where all the large cloud providers were.

The Board of Trustees authorized a $22M expenditure to have a quick-start effort towards an AI-aware framework. Getting GPUs was pretty hard already then, but what we have seen is that many other universities, we have focused on buying servers with a few hundred H100 cards did not have either the speed or capacity in their storage infrastructure to feed the System with data, thus the machines were running idle.

So, as our first priority, we have purchased several large Arista 7800 sqitches, with multiple line cards, supporting 400G ethernet. With these we quickly upgraded our 2x100G campus backbone to 4x400G. We have signed a 10 year licencing agreement with Weka, who built an extremely fast storage caching layere that can connect multiple deep silos. This fitted the JHU situation very well, where the different schools so far have built their own storage systems. Suddenly these all could be connected at extreme speeds.

We have purchased three NVMe clusters, with a total capacity of about 10PB. Weka installed on these systems provided an extreme performance: over 640 GiB/s (~5.5 Tbit/s) of aggregate read bandwidth, more than 20 million read IOPS at ~150 µs latency, measured with up to 48 clients, each on a 200 Gb/s link and reaching up to 94% of line rate. We have also added about 100PB of HDD storage, luckily still at the old prices. In order to cut costs, we did pretty unusual things, like

taking a 100 node cluster with 24 disk SMC servers, that was at the of their life, tearing out their motherboards, replacing them with brand new Ryzen systems, bought by parts on eBay, at a fraction of the cost of a new system. Now this system forms a 30PB+ Cepth cluster.

The we started an upgrade of the cooling system in our existing containers, with liquid cooling. This essentially double the capacity of the facility, and we could fill the rest with GPUs, mostly H100 and H200 systems. We have also added about 50,000 CPU cores to the systems. The GPUs were connected to an existing Infiniband system, running on NVIDIA CX8 cards. We will convert to convergent ethernet when we do our next generation purchase.

Large Scale Science Data

We have long promoted the open and public use of large data sets. We have been serving the data from the Sloan Digital Sky Survey (SDSS) for the whole world. The total data with raw files is about 1PB. It is still serving about 1.5M SQL queries per day. We have an interactive private cloud environment around this system, the SciServer, that has about 30,000 registered users from around the worlkd, who can work on petascale dataset, in an interactive and collaborative fashion, with resources sharing beyond tier Jupyter notebooks.

The system now hosts many petabytes of data from about 70 different science projects, both form JHU and beyond. Some of the larger collections include the JHU Turbulence Database which stores the world's largest supercomputer simulations of turbulence. We provide an interactive service where users can place virtual sensors into the simulation that can send back a data stream about their trajectory. The system contains about 1.1PB of data, and has served over 700 trillion data points to the world. It is also par tof the NAIRR pilot projects, and Nvidia has based their turbulence proxy simulation demo on our data.

JHU is also the host of the Poseidon project, a 4.5PB simulation for one year of the Earth's ocean circulation, at a 1.5km resolution. This is in active interactive use by the oceanography research community. Soon this will be augmented by a newer system that will include ocean-air interactions.

We are also building AstroPath, a project combing expertise from SDSS and cancer pathology, trying to collect 1000 more data than previously available about cancer immunotherapy. We are building spatial database of several billion cells, with detailed spatial features, showing the physical interactions between tumor and immune cells. The database contains today over a billion cells., In order to get there, we had to process over 65 trillion pixels of images.

AI-Ready Data

All of these datasets are now actively converted into AI-ready format. We are involved in several projects building multimodal foundation models of the different data.

It is clear that traditional ways of looking at data are superseded at the speed of light, and we have to reinvent how scientific discoveries will be taking place in the next few years.

We can only do this if we have significant amounts (Petabytes) of data, stored n AI ready formats (tensors, various embeddings) and in a way that they can be streamed into an IA system.

Large language models can increasingly cowork, access data in more varied formats. They are nonw capable of abstract reasoning, and they are launching a hierarchy of agents who can perform various computations on our own system where our data reside.

In this world it is the speed of the data that matters. This will not require billions of dollars at the universities, just a strong team with a deep understanding of low-level IO systems an extreme networking.

Demonstrations:

We will present two demonstrations, one with the Turbulence database and other with the Poseidon project. In both cases we will first show how to perform an interactive quick-look exploration of petabyte scale data, then we select a significant chunk of the data, transfer it in real-time to the floor and perform a more intense visualization. The demonstrations will be driven by domain scientists from JHU.