National Research Platform

Frank Wuerthwein
Director
San Diego Supercomputer Center

NRP aims to address 2 challenges

- Provide an AI Education & Research Infrastructure to all of academia that academia can afford.
- Provide a “technology playground” where CS and Domain Science researchers can meet and work together to accelerate technology adoptions in light of the end of Moore’s Law

AI Education Infrastructure For All

Frank Wuerthwein
Director
San Diego Supercomputer Center

Let's play with some numbers …

- 20M students attend college in the USA
- 50% of students at San Diego State University want some sort of formal training in AI (survey result based on 8,000 student responses)
- 22% of UC San Diego undergraduates have used data & compute in their classroom in AY24

=> There are 4-10M college students in the USA that need data & compute resources for their classroom education

Cloud is too expensive for academia

Unless the NSF pays the cloud bill ….

- Renting an NVIDIA HGX from AWS for 100 days costs same as buying it.
- When we buy it, we typically operate it for at least 6 years, or so
  - ⇒ 6 years / 100 days = 20 … or so …

When you rent, and run out of money you become homeless.
When you own, and run out of money, you still have a roof over your head.

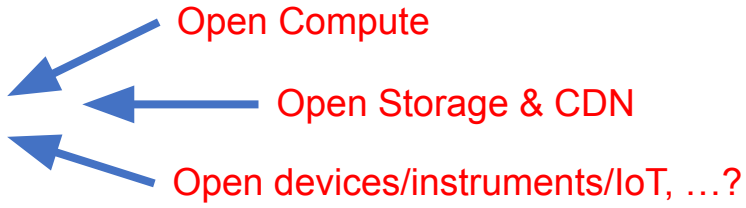
Operations is expensive …
… unless it is amortized across large scales

- Of the ~4,000 accredited, degree granting higher education institutions only less than 200 are research intensive (R1), and have the scale to afford a group of sysadmins, cybersecurity, user support, … professionals.

- **What if we aggregated the system administration, cybersecurity, and user support across 1,000 colleges ?**
- Colleges own their AI hardware, but we centralize the operations to benefit from economies of scale, just like the cloud providers.

Long Term Vision

- Create an Open National Cyberinfrastructure (CI) that allows the federation of CI at all ~4,000 accredited, degree granting higher education institutions, non-profit research institutions, and national laboratories.
  - Open Science
  - Open Data
  - Open Source
  - Open Infrastructure

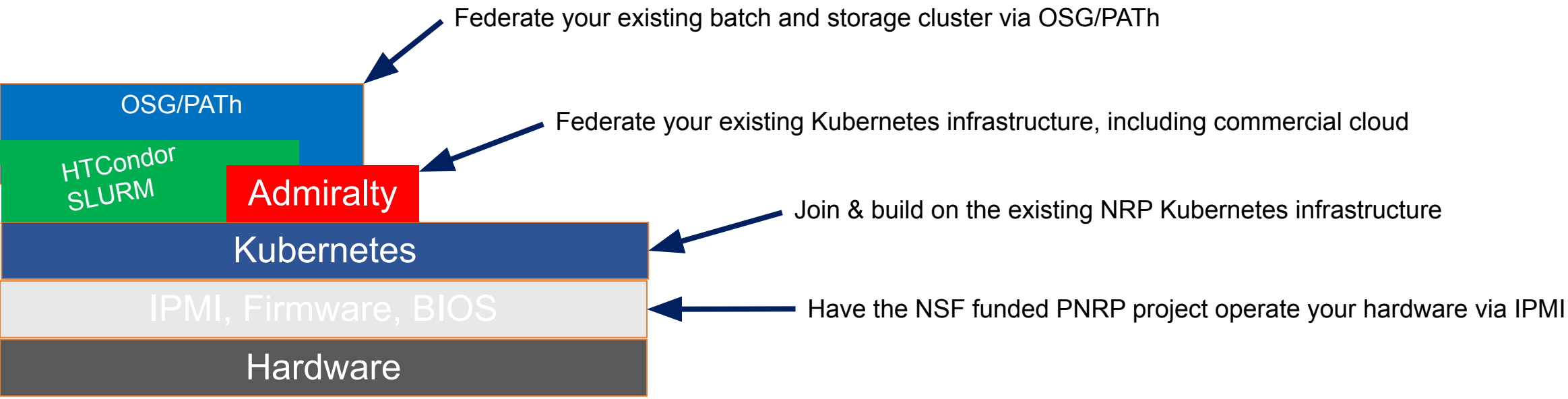


NRP offers the community

- **To run your hardware from IPMI up**
  - OS maintenance, security monitoring, …
- **Researchers, Educators, and students see a global scale Kubernetes cluster**
  - While the cluster is shared, we restrict use of individual hardware to owners only when necessary
  - Documentation and training for the community
- **Lots of software to reuse**
  - E.g. we operate JupyterHub for the community & show people how to deploy & customize their own
- **Maintain lots of topical Chat channels for the community to learn from each other and interact with each other**
  - Chat channels are monitored by professionals but community is encouraged to interact and help each other
  - Starting to explore AI chatbots trained on the chats
- **LLM as a service**
  - We run about a dozen popular LLMs and provide popular APIs to upload any AI model from Hugging Face onto our resources.

Flexible Architecture to build on horizontally & vertically

- Depending on effort available and control desired, you can built on NRP both vertically and horizontally at different layers of the stack.



Mental Picture of 3 types of “AI” Services

- Compute, Storage, Jupyter, Matrix
  - The initial set we started with
- LLM as a Service
  - We offer any LLM that is available via Hugging Face
  - We can integrate your ChatGPT offering via LiteLLM to facilitate comparisons between different LLMs
- AI workflows as a service => leverage NDP
  - Imagine curated data, curated AI agents, curated RAG workflows, … all available via a simple intuitive interface to compose custom AI services that fundamentally require workflow execution to instantiate the service.

https://nrp.ai/documentation/userdocs/ai/llm-managed/

NRP Nautilus

Search ⌘ K

Auto

Start Here

User Guide
- Start
- Tutorials
- Running
- Jupyter
- Coder
- AI/LLMs
  - NRP Managed LLM
  - LLM as a Service
  - LLM in JupyterHub
  - Cloud AI 100 Cards
- Networks
- FPGA
- Storage
- Development

Admin guide

NRP-Managed LLMs

The NRP provides several hosted open-weights LLM for either API access, or use with our hosted chat interfaces.

Chat with NRP LLMs →
Use the LibreChat interface to chat with NRP hosted LLMs

Get an API token for NRP LLMs →
Get an API token to programically interact with the LLMs or use the LLMs in other apps

Chat Interfaces

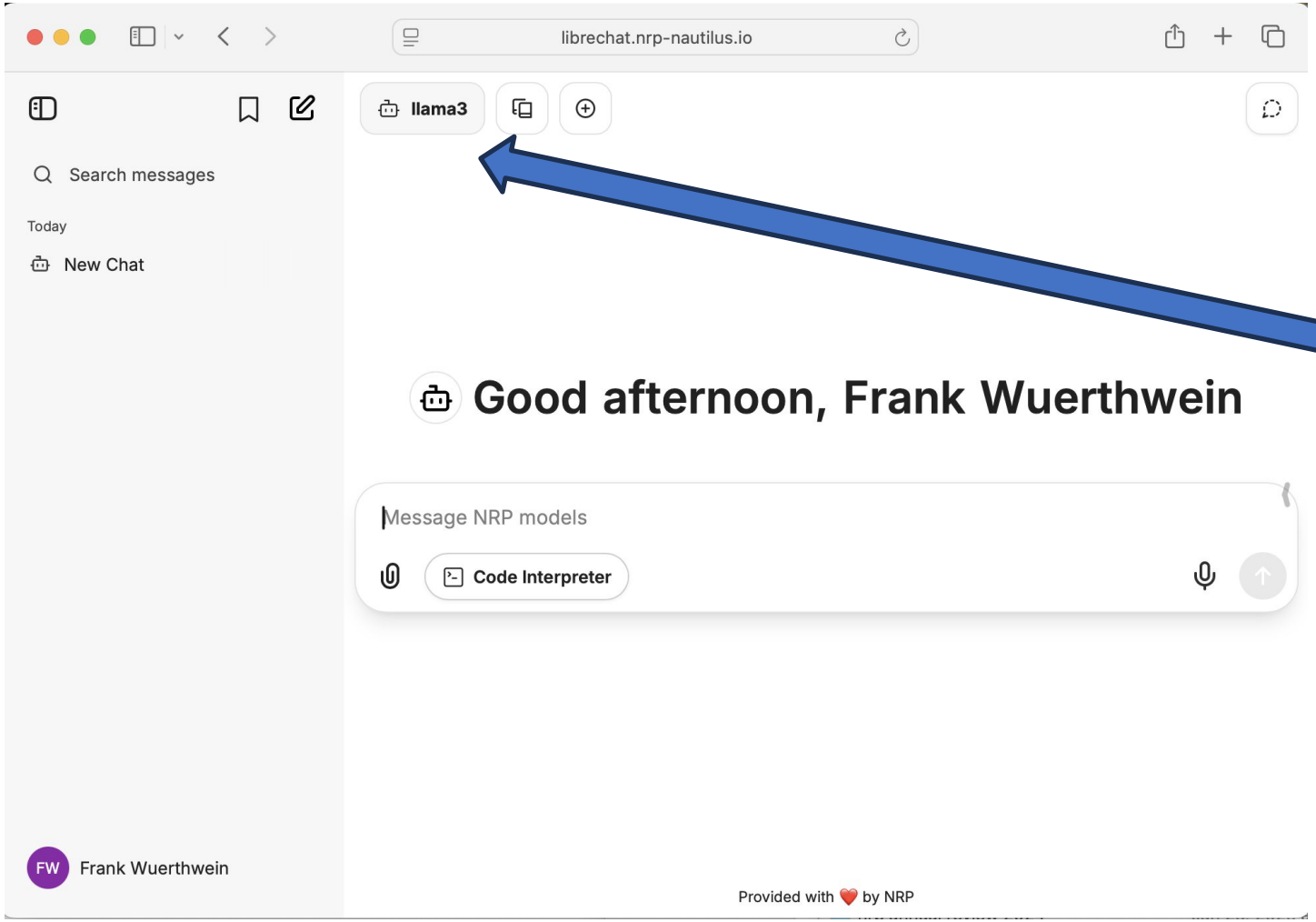
If you are looking to chat with an LLM model similar to the interface provided by ChatGPT, we provide LibreChat, based on the LibreChat project. This is a simple chat interface for all of the NRP hosted models. You can use it to chat with the models, or to test out the models.

Visit the LibreChat interface

On this page
- Overview
- Chat Interfaces
- API Access to LLMs
- Example Python Code
- Available Models

Click here ….

Then login via your institutional login



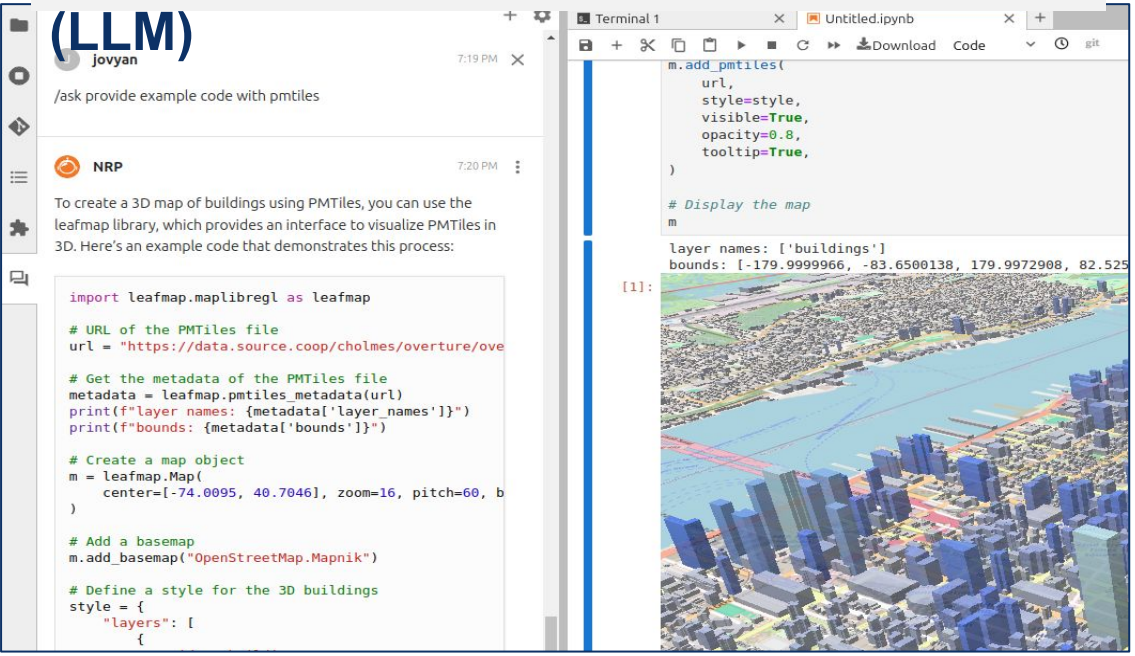
You can pick different models, and explore.

E.g. a fun thing to do is ask deepSeek about Tiananmen Square, and compare what you get with llama3.

Example for complex classroom use

Students learn how to visualize a 3D map of buildings in NYC.
They start with a vanilla LLM, that has no clue how to do this.
Then teach that LLM about a few useful software packages.

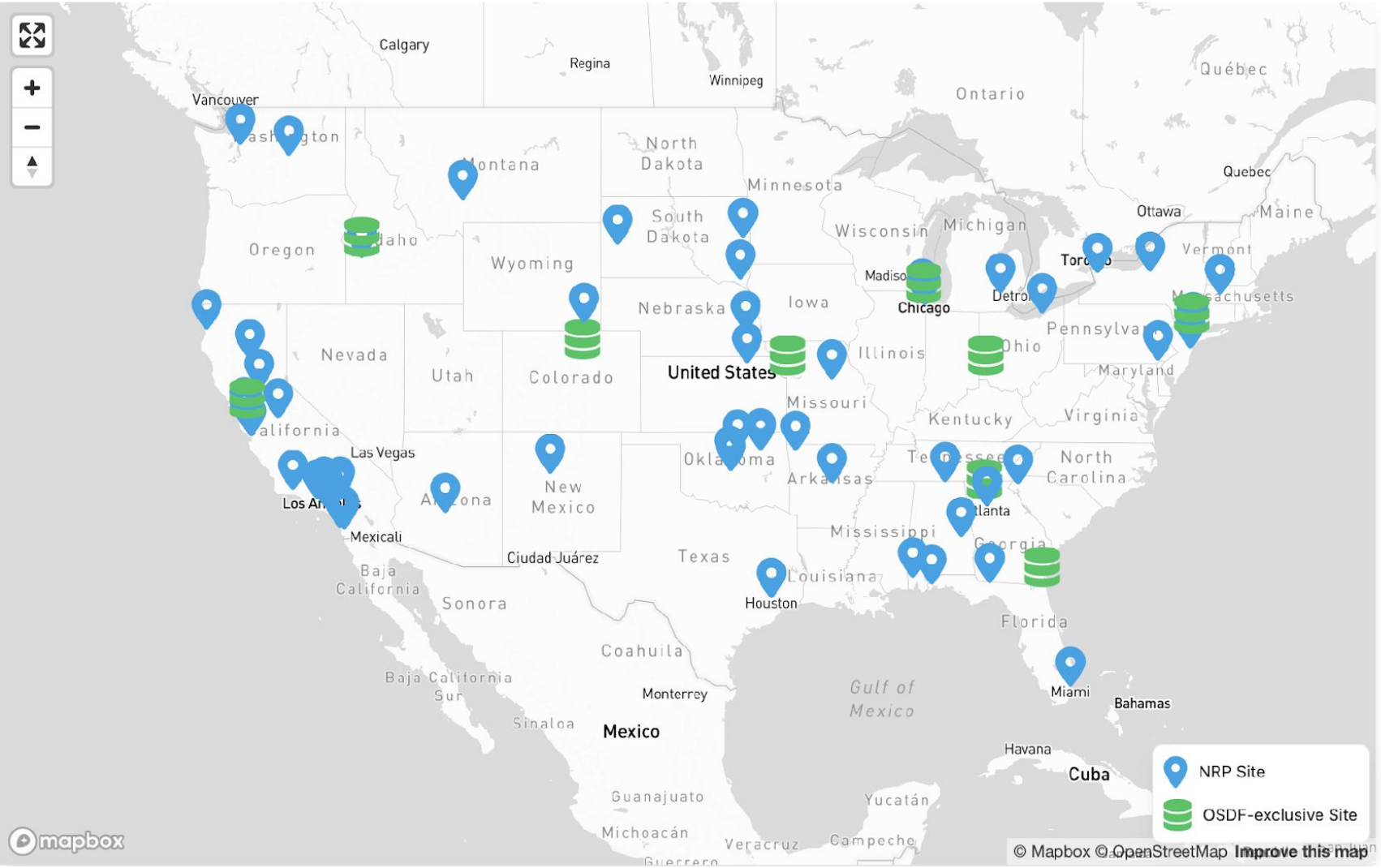
Coding with Large Language Models (LLM)



Voila, the LLM has learned how to provide working code for Jupyter@NRP combining multiple packages.

Classroom of **Carl Boettiger**
University of California, Berkeley

- 122 students
- Active learning classroom



NRP NATIONAL RESEARCH PLATFORM

The National Research Platform is a partnership of more than 50 institutions, led by researchers at UC San Diego, University of Nebraska-Lincoln, and Massachusetts Green High Performance Computing Center and includes contributions by the National Science Foundation, the Department of Energy, the Department of Defense, and many research universities and R&E networking organizations in the US and around the world.

Select a site or click on a site in the map

Select...

Sites	Nodes
84	**439**
Sites hosting NRP nodes	Nodes registered in Kubernetes
GPUs	**CPU Cores**
1,496	**29,254**
Total GPUs across all nodes	Total CPU cores across all nodes

In January we were at 72 locations and 22 PB of disk space

https://dash.nrp-nautilus.io

NRP as of yesterday …

We operate internationally for US Science Collaborations

NRP NATIONAL RESEARCH PLATFORM

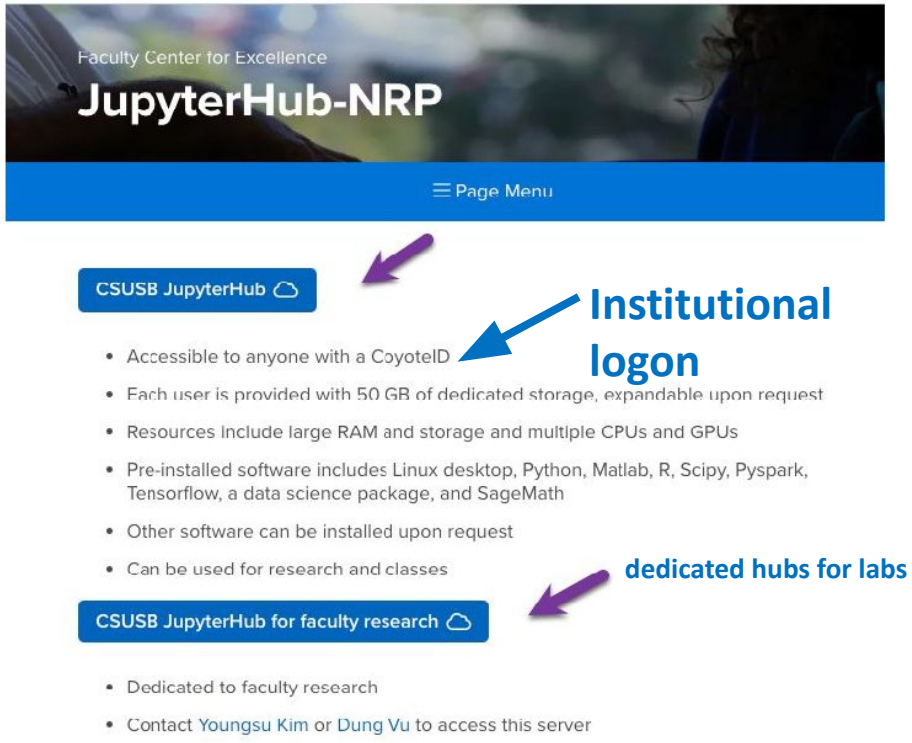
UCSD

Example: CSU San Bernardino

- 1,100 faculty members
- 19,000 students
- Serves 2 of CA's largest counties
- Hispanic Serving Institution
- 57% Pell Grant recipients
- Many student oriented projects

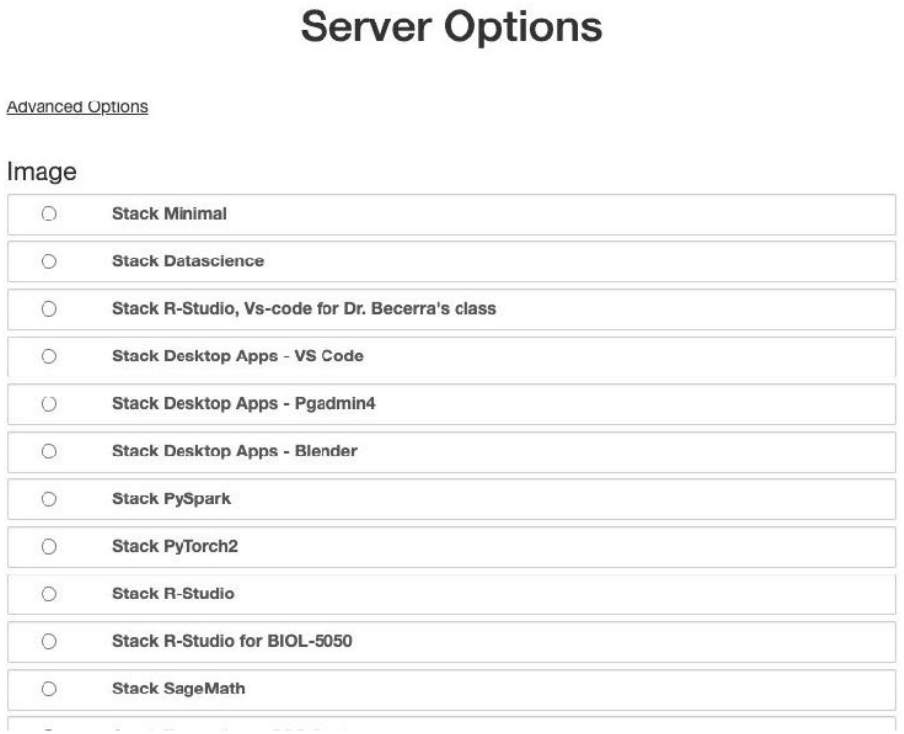
Non-R1 Institution

JupyterHub User Interface at CSU, San Bernardino



Institutional logon

dedicated hubs for labs



Wide range of dedicated hubs:

https://csusb-metashape.nrp-nautilus.io: 3D modeling
https://csusb-vasp1.nrp-nautilus.io Viena Ab initio Simulation package (VASP)
https://csusb-cousins-lab.nrp-nautilus.io: VASP simulation
https://csusb-becerra.nrp-nautilus.io AI/ML project
https://csusb-biol-5050.nrp-nautilus.io: Biology course
https://csusb-cse-salloum.nrp-nautilus.io Summer Research
https://csusb-drhamoudahub.nrp-nautilus.io Data Analytics
https://csusb-ratnasingam.nrp-nautilus.io Data Analytics
https://csusb-zhang.nrp-nautilus.io AI/ML project

63 Campuses had active namespaces in 2024 on NRP

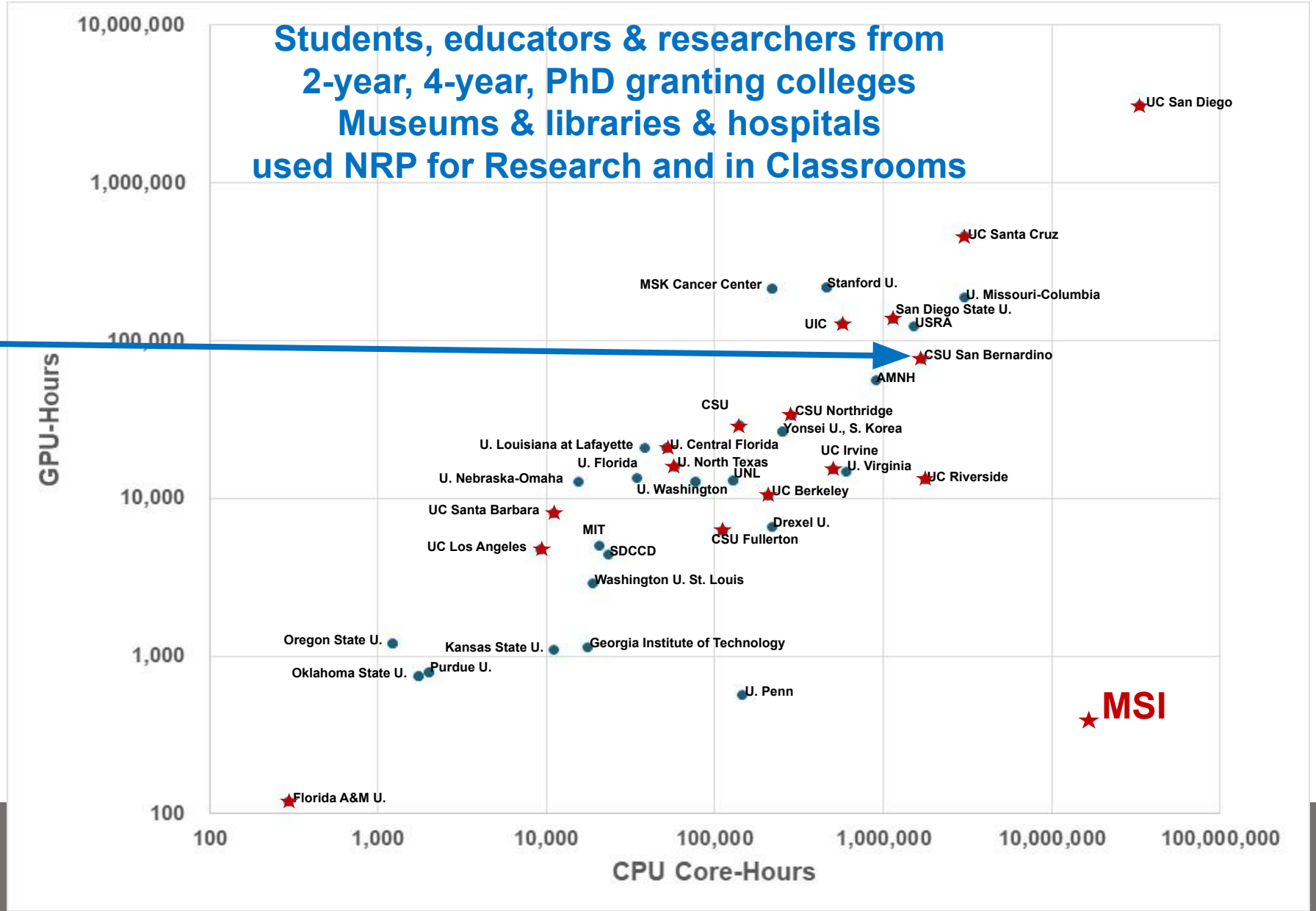
CENIC connected campuses:

San Diego Community College District

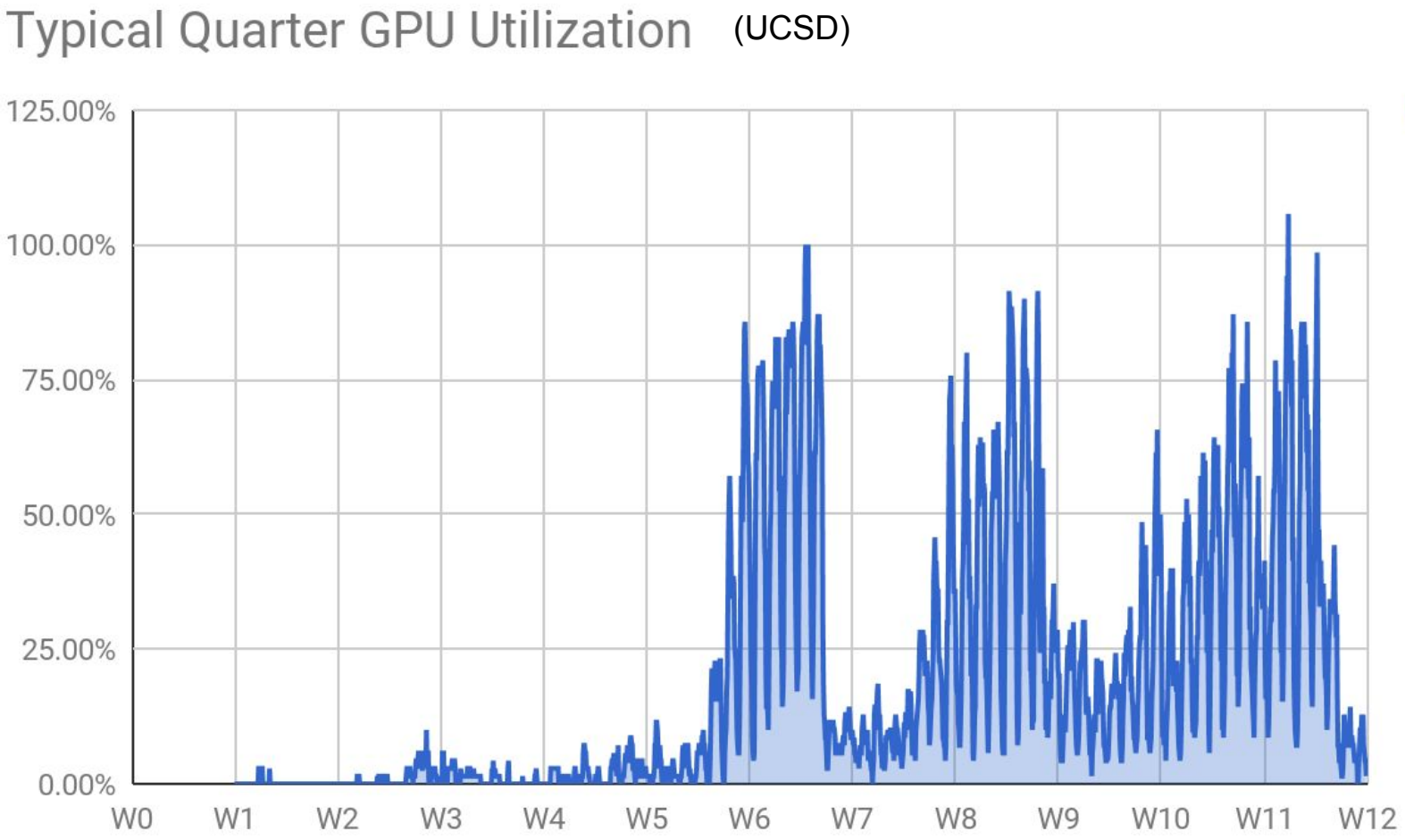
San Diego State University
CSU San Bernardino
CSU Northridge
CSU Fullerton
Cal Poly Humboldt

UC San Diego
UC Santa Cruz
UC Irvine
UC Riverside
UC Berkeley
UC Santa Barbara
UC Los Angeles

Stanford University



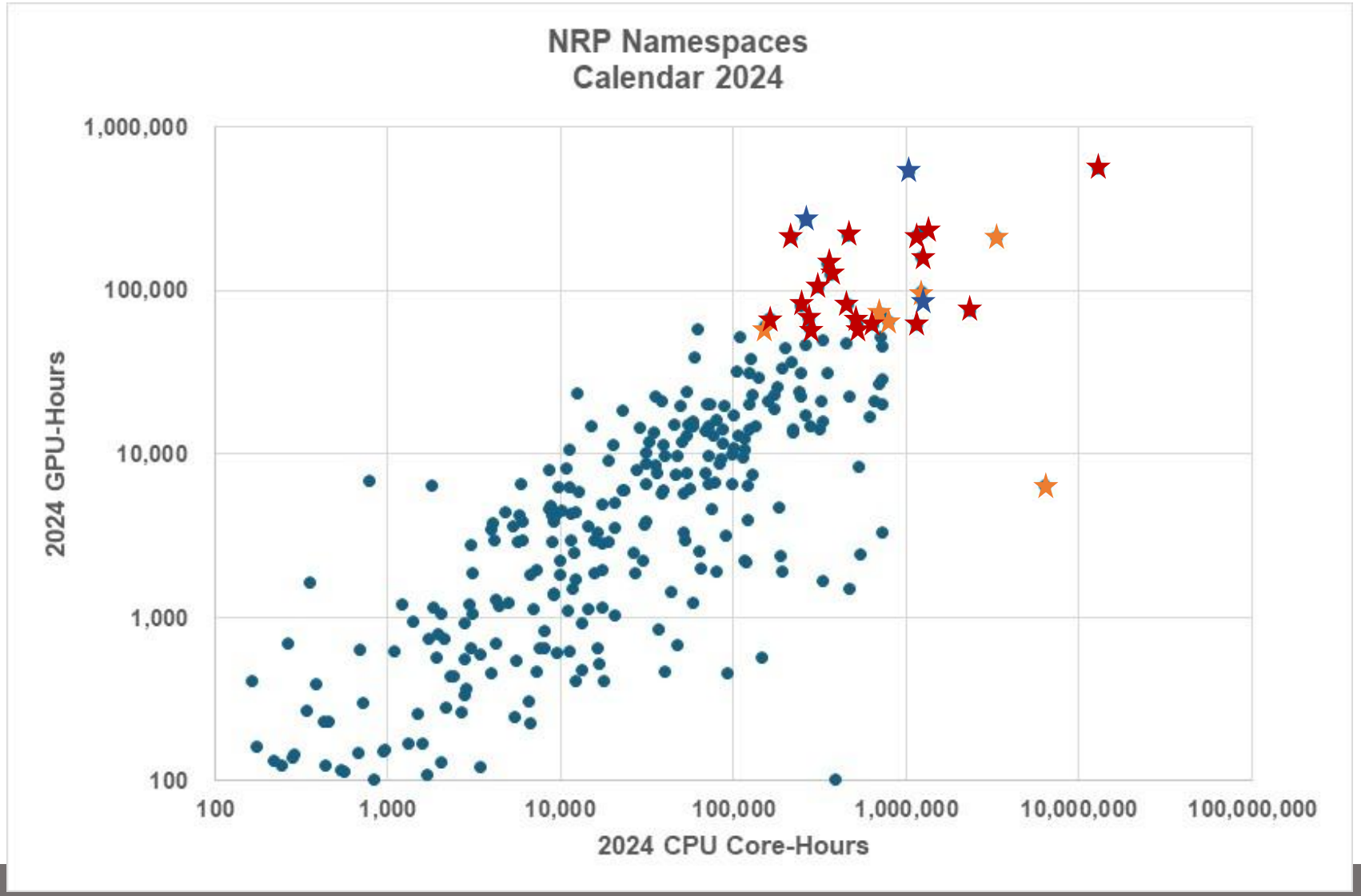
Coursework Activity Patterns



There will always be a lot of left over capacity for research.

Education must provision for peak demand.
Research mops up the left-overs.

In 2024, 285 NRP Namespaces Used More Than 100 GPU-Hours & 100 CPU-core hours [2 GPU-Hrs/Week]



We are in phase 1 of 3 phases

- **Phase 1:**
  - **Establish Concept and scale out to 100 colleges and 10,000 students/year**
  - Today we operate hardware at 53 colleges
  - Expecting to complete phase 1 in 2-3 years.

- **Phase 2:**
  - **Scale out to 1,000 colleges and 1 Million students/year**
  - Scale out to 24x7 support for system & cybersecurity, but only 9-5 for user support & training
  - Expecting this to be a 5-10 year program
  - Estimating 10,000 GPUs needed to serve this population

- Phase 3: Sustainability through 501(c)(3) and membership fees
  - **At 1 Million students per year, even a modest fee of $8/(student*year) sustains the program**

Longer term: include high schools and public libraries

Technology Playground

Frank Wuerthwein
Director
San Diego Supercomputer Center

NRP NATIONAL RESEARCH PLATFORM

UCSD

NRP brings CS R&D and Domain R&D onto the same platform

NRP blurs the lines between ”testbed” and “production” CI

Create social cohesion to accelerate domain science adoption of new programming paradigms & architectures

NRP NATIONAL RESEARCH PLATFORM

UCSD

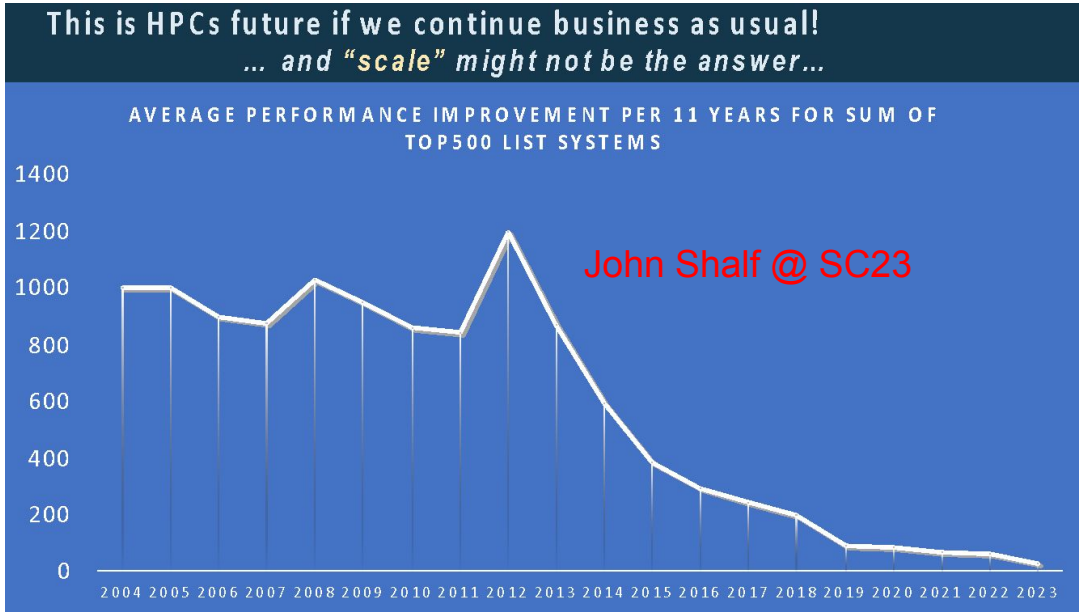
Algorithm x Hardware = Science

For decades, scientific progress was exponential in part because hardware performance per $ increased exponentially due to Moore's Law.

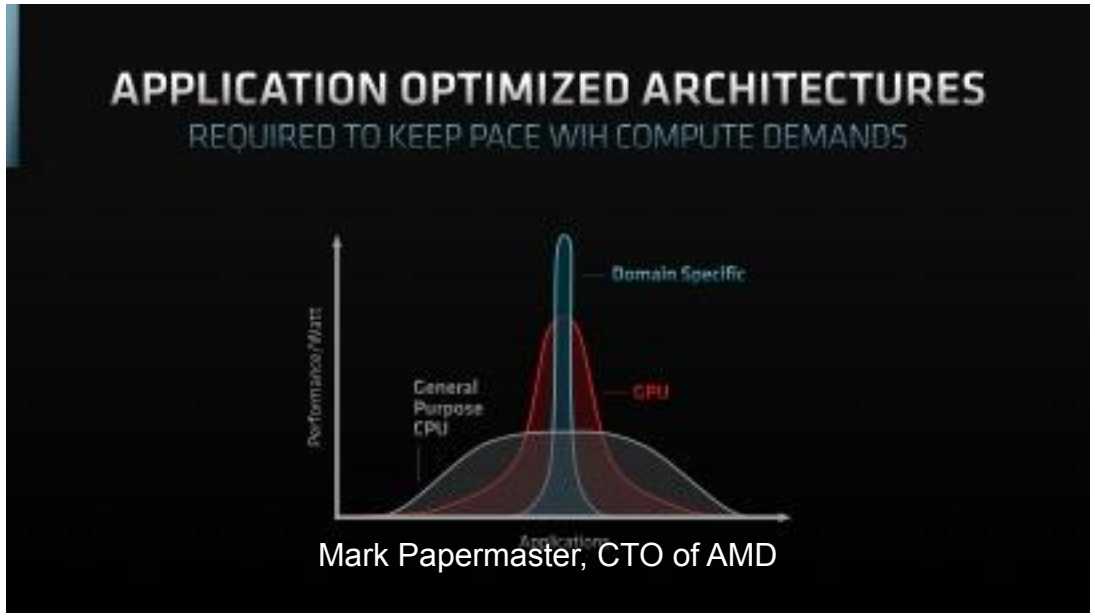
Algorithmic performance did not scale nearly as fast, in most cases.

The end of Moore's Law threatens scientific progress.

“end of Moore’s law” motivates new architectures



Performance improvements vs time slowed down by O(100)



More details Thursday 1:30-3:30pm

PI, Tajana Rosing

PRISM, a Jump 2.0 project funded by SRC is early user of FPGAs@NRP

NRP supports FPGAs, P4 switches, NVIDIA DPUs & DGXs

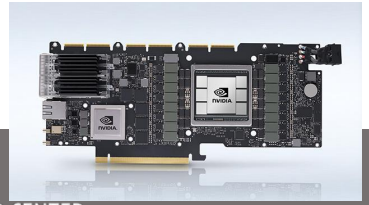
Committed to be a “Playground” of technologies, easily deployed & operated via BYOR and BYOD.

Advanced Technology Laboratory on NRP

- **Programmable computational capabilities emerged in devices of all kinds**
  - Storage devices with embedded FPGAs => ”Computational Storage”
  - GPUs on Network Interface Cards => “Data Flow Programming”
  - Programmable switches, down to individual ports => “Programmable Networks”
- We innovate nextGen systems in NRP to solve grand challenges of science
- Innovations made available to all of open science via our Open Infrastructure

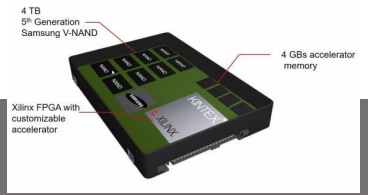
Strategic Objective is to bring CS Research closer to Domain Research in the hope of decreasing time to adoption of new technologies & ideas

NVIDIA BlueField DPU

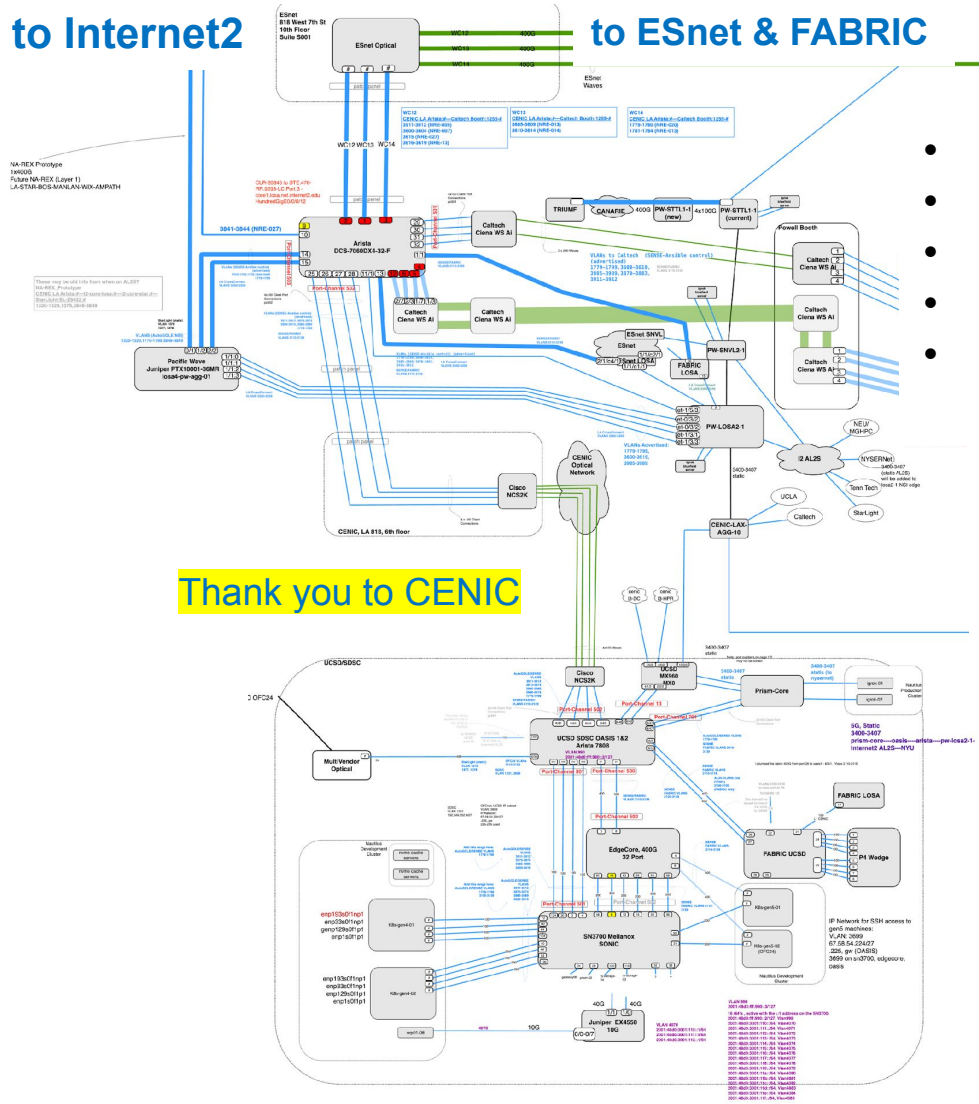


P4 programmable switches

Xilinx SmartSSD



400G WAN Infrastructure



Infrastructure at SDSC:

- FPGAs: 32 U55C, 24 Bitware 520
- 400G P4 programmable switches
- 8 NVIDIA HGX w 8xA100 80G each
- 400TB of NVMe
- FABRIC node

We own 400G capable nodes at MGHPCC, CERN, SDSC

We peer at 400G in LA with multiple networks via our 400G Arista switch

Real-Time In-Network Machine Learning on P4-Programmable FPGA SmartNICs with Fixed-Point Arithmetic and Taylor

Mohammad Firas Sada, John J. Graham, Mahidhar Tatineni, Dmitry Mishin, Thomas A. DeFanti, Frank Würthwein

As machine learning (ML) applications become integral to modern network operations, there is an increasing demand for network programmability that enables low-latency ML inference for tasks such as Quality of Service (QoS) prediction and anomaly detection in cybersecurity. ML models provide adaptability through dynamic weight adjustments, making Programming Protocol-independent Packet Processors (P4)-programmable FPGA SmartNICs an ideal platform for investigating In-Network Machine Learning (INML). These devices offer high-throughput, low-latency packet processing and can be dynamically reconfigured via the control plane, allowing for flexible integration of ML models directly at the network edge. This paper explores the application of the P4 programming paradigm to neural networks and regression models, where weights and biases are stored in control plane table lookups. This approach enables flexible programmability and efficient deployment of retrainable ML models at the network edge, independent of core infrastructure at the switch level.

Comments: To appear in Proceedings of the Practice and Experience in Advanced Research Computing (PEARC25)
Subjects: **Distributed, Parallel, and Cluster Computing (cs.DC)**; Networking and Internet Architecture (cs.NI)
Cite as: arXiv:2507.00428 **[cs.DC]**
(or arXiv:2507.00428v1 **[cs.DC]** for this version)
https://doi.org/10.48550/arXiv.2507.00428

Just one example use of our FPGAs, P4 Programming, and AI/ML.

Qualcomm Cloud AI 100 Ultra

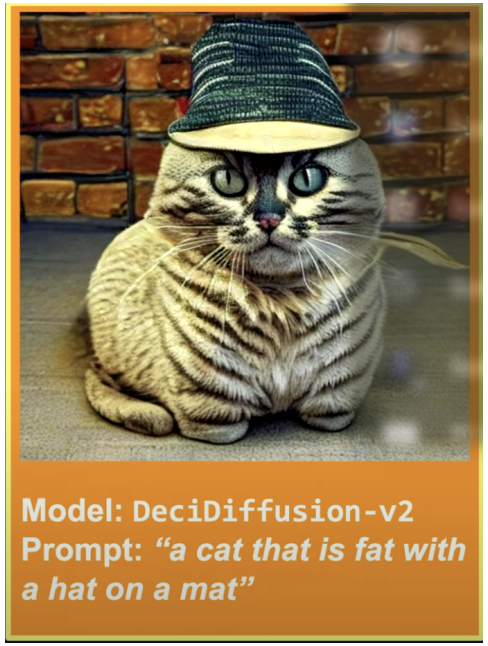
- Multi-SoC PCIe Cards
- 100B GenAI models on a single card
- Software tools
- 8 cards per server (2x4, w/ 4 on PCIe switch)
- Fully programmable + C++ SDK
- Finetuning capabilities

Work in progress: Understand the tokens/$ and tokens/(sec*power) performance for different models.

We are looking for more cost-effective inference hardware with which we can provide LLM service for a million students.

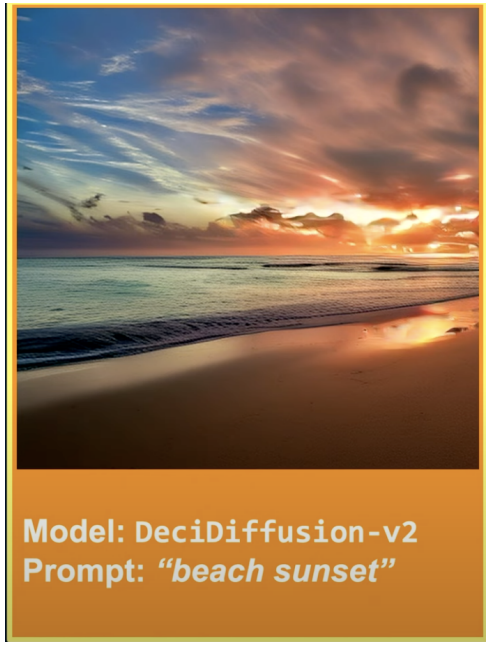
Qualcomm Cloud AI 100 Ultra

- **Models tested:**
  - **LLMs (varying sizes)**
  - **OpenAI Whisper (speech recognition model)**
  - **Stable Diffusion**



Model: `DeciDiffusion-v2`
Prompt: *"a cat that is fat with a hat on a mat"*

Model: `DeciDiffusion-v2`
Prompt: *"nature"*



Model: `DeciDiffusion-v2`
Prompt: *"beach sunset"*

Model: `DeciDiffusion-v2`
Prompt: *"a dinosaur in front of a tree"*

Summary & Conclusions

- **Opportunity is a requirement for Social Mobility**
  - Education is the most effective guarantor of social mobility
  - Data & Compute is increasingly required for Education

We provide a Data & Compute Platform for Higher Ed that is cost-effective and aspire to make it available to all colleges in the USA

- **Algorithm x Hardware = Science**
  - AI is the dominant paradigm for exponential growth in algorithm performance.
    - As a predominantly GPU infrastructure, we are supporting this by default.
  - As Moore's Law is stalling, we are exploring new architectures and programming paradigms.
  - Aspiring to offer a "digital watercooler" for the CS community to meet the domain sciences.

Acknowledgements

The NRP is supported by NSF grants OAC-1541349, OAC-1826967, OAC-2030508, OAC-1841530, OAC-2005369, OAC-21121167, CISE-1713149, CISE-2100237, CISE-2120019, & OAC-2112167

And by CENIC, Pacific Wave, MREN, GPN, NYSERNet, FLR, NEREN, SunCorridor, OARnet, SCLR, the Albuquerque GigaPoP, and Internet2

As well as a long list of Universities and colleges that host hardware, and share their hardware with the community.