5 NRP NATIONAL RESEARCH PLATFORM

# FIFTH NATIONAL RESEARCH PLATFORM WORKSHOP

March 19-22, 2024


ABSTRACT

This workshop was supported by NSF award #1826967 and was held at the University of California, San Diego in conjunction with the KNIT8 FABRIC workshop.



C. Anderson and T. DeFanti

# Table of Contents

Table of Figures

# 1 Executive Summary

Scientific exploration and discovery are enabled by increasingly specialized information technology infrastructure. The requirements of a wide array of scientific research domains provide challenges that regularly exceed the capabilities of existing infrastructure, necessitating a continuous rejuvenation of services, architecture, and technologies. This process must not simply provide incremental improvement but truly transform current capabilities.

The Fifth National Research Platform (5NRP) workshop focused on examining opportunities for designing and implementing future-oriented cyber-capabilities for our science research communities, and how to engage researchers and educators. Participants discussed expanding and replicating the National Research Platform (NRP) addressing the potential challenges, funding opportunities, and private sector interest in scaling the Science DMZ model to a national level with the NRP as a fully working, welcoming and expandable model. How the NRP can support Artificial Intelligence (AI) and Machine Learning (ML) was also presented and discussed.

Leading up to this initiative and driven by the high-speed networking needs of collaborative, big-data science, the National Science Foundation funded the Pacific Research Platform (PRP). The PRP created a researcher-defined and data-focused network whose requirements are driven by direct engagements with sophisticated, cyberinfrastructure-knowledgeable science teams chosen from several data-intensive fields, including particle physics, astronomy, biomedical sciences, earth sciences, and scalable data visualization. A partnership of more than twenty-five institutions initially, it has grown to a nationwide collaboration and is complemented by other allied NSF awards to carry on and expand NRP over the next ten years.

The 5NRP Workshop brought together more than one hundred representatives from existing PRP partners, the new NRP awardees, and many other institutions, who support domain scientists, network and system administrators, campus CIOs, regional network leaders (The Quilt), and representatives of ESnet, Internet2, ACCESS, the Open Science Grid (OSG), the Open Storage Network (OSN), the OpenAirInterface[1], NSF IRNC awardees, NSF's Office Director of the Office of Advanced Cyberinfrastructure (OAC), leaders of the Established Program to Stimulate Competitive Research (EPSCoR) section of the NSF Office of Integrative Activities (OIA), the Global Research Platform, and other research platforms.

[1] https://openairinterface.org/

# 2 5NRP Tutorials Track – Tuesday March 19, 2024

Several tutorials were offered aimed at helping attendees learn how to access NRP and allied resources. Registration had a limit of twenty per tutorial, but over forty people showed up for each. The slide decks for the tutorials are available on the NRP website.

## 2.1 Kubernetes for Scientists: Examples drawn from AI - Tutorial Leaders: Mahdihar Tatineni, Dima Mishin, Igor Sfiligoi, UCSD

This tutorial provided a Kubernetes architectural overview, an overview of typical job and workflow submission procedures, and examples of the various options available to enable optimal use of GPU, CPU, and storage resources for AI use cases. Theoretical information was paired with hands-on sessions operating on the Prototype National Research Platform (PNRP) production Kubernetes cluster which features a variety of compute and storage resources.

## 2.2 How to run AI/ML computations on SDSC's Voyager[2] - Tutorial Leaders: Paul Rodriguez, Mahidhar Tatineni, UCSD

In this tutorial we provided information on the Voyager system architecture with details on the Habana processing units (HPUs), provided information on containerized software stacks, file systems, examples using Kubernetes, and overall guidelines. Also, we discussed options and considerations for scaling training across multiple processors and multiple nodes, including brief introduction to parallelization options, like DeepSpeed[3] and other useful tools.

## 2.3 How to run AI/ML computations on SDSC's Expanse[4] - Tutorial Leaders: Paul Rodriguez, Mahidhar Tantineni, UCSD

In this tutorial we provided information on the Expanse system architecture, with details on the available GPU resources and scheduling, using containerized and Conda-based software stacks, and examples of batch and interactive use of TensorFlow[5], PyTorch[6]. Attendees were able to use the Expanse portal to run Jupyter notebooks with AI/ML examples. Additional information was provided on running multi-node TensorFlow workloads on Expanse.

## 2.4 Writing a successful CC* Grant - Tutorial Lead: Jerry Sheehan, Salk Institute

This tutorial explored best practices for developing NSF Campus Cyberinfrastructure proposals. Topics addressed included the 2024 CC* Solicitation, identifying science drivers, best practices for developing your proposal technology plan, strategies to meet your 20% resource sharing requirements, and tips to tell your broader impact story.

---

[2] https://sdsc.edu/support/user_guides/voyager.html
[3] https://www.deepspeed.ai/
[4] https://www.sdsc.edu/support/user_guides/expanse.html
[5] https://www.tensorflow.org/
[6] https://pytorch.org/

## 2.5 Demo/Poster Session

In the evening there were demonstrations and a poster session in conjunction with FABRIC's KNIT8 workshop.

# 3 5NRP Track – Wednesday March 20, 2024

## 3.1 5NRP/NAIRR Pilot - Dr. Kate Antypas, NSF OAC

### 3.1.1 Overview

Dr. Antypas gave an overview of the Office of Advanced Cyberinfrastructure (OAC) and its breadth of scope. There are six current priorities for the Office: defining, advancing and connecting the current cyberinfrastructure (CI) system; growing and developing communities and the workforce; enabling discovery through integrations of data, software, and cyberinfrastructure; infrastructure for artificial intelligence (AI); investing in new technologies; and developing partnerships for long-term US leadership in research CI. The role of OAC is to inform and transform science and engineering research through an integrated cyberinfrastructure ecosystem.

There are five OAC investment areas: Advanced Computing; Networking and Cybersecurity; Learning and Workforce Development; Software and Data; and Strategic Investments. Examples of these investments include an NSF-funded collaboration that exposed safety flaws in customizable pre-trained LLMs[7], a research project out of the University of Texas that used AI to analyze tens of thousands of X-ray images and genetic sequences using NSF-funded Frontera system to understand the genetics that shape our skeletons[8], and a cross-division collaboration between the Division of Materials Research and OAC to investigate accelerating discovery of design and materials[9].

Another highlight is the National Discovery Cloud for Climate (NDC-C)[10], a portfolio put together at the end of 2023. It aims to serve as a pilot for future efforts to enable equitable access to NDC-C across all fields. There are over thirty investments including Advanced Computing, Open Platforms, Platform Services, Focused Pilots, Cloud Resources, Data Resources, Training and Education, Security and Resiliency, and Broad Engagement and Outreach.

---

[7] Qi, Xiangyu; Zeng, Yi; Xie, Tinghao; Chen, Pin-Yu; Jia, Ruoxi; Mittal, Prateek; Henderson, Peter**. "**Fine-tuning Aligned Language Models Compromises Safety Even When Users Do Not Intend To!" October 2023 10.48550/arXiv.2310.03693 / arXiv:2310.03693 and "Researchers Say Guardrails Built Around A.I. Systems Are Not So Sturdy**"**, *The New York Times*, October 20, 2023. https://www.nytimes.com/2023/10/19/technology/guardrails-artificial-intelligence-open-source.html

[8] Eucharist Kun, Emily M. Javan, Olivia Smith, Faris Gulamali, Javier de la Fuente, Brianna I. Flynn, Kushal Vajrala, Zoe Trutner, Prakash Jayakumar, Elliot M. Tucker-Drob, Mashaal Sohail, Tarjinder Singh, Vagheesh M. Narasimhan. "The genetic architecture and evolution of the human skeletal form." *Science*, 2023; 381 (6655) DOI: 10.1126/science.adf8009 .

[9] Laasner et al., (2020**). "**MatD^3^: A Database and Online Presentation Package for Research Data Supporting Materials Discovery, Design, and Dissemination". *Journal of Open Source Software*, 5(45), 1945, https://doi.org/10.21105/joss.01945 .

[10] National Discovery Cloud for Climate: https://new.nsf.gov/cise/national-discovery-cloud-climate .

The Pelican project[11] is one example of bringing different communities together to help move the science forward. Using the data collected by NOAA's National Marine Fisheries Service, the project was able to adapt a workflow to process fifty-five research cruises comprising over 100,000 files in just 11 hours.

### 3.1.2 An Update on the National Artificial Intelligence Research Resource (NAIRR) Pilot[12]

The vision for the NAIRR is a national infrastructure that connects the research and education communities to the necessary computing, data, models, software, user support, and expertise to advance the AI ecosystem. Many potential contributors lack access to needed resources that can be expensive and difficult to navigate. We need NAIRR to ensure a strong public interest in the direction of AI, and to train the next generation of AI researchers and leaders.

NAIIR aims to fulfill the urgent national goals of spurring innovation, increasing diversity of talent in AI, improving US capacity and capabilities, and advancing trustworthy AI. To do so, NAIRR will facilitate national coordinated access to AI resources, assuring that the public interest is represented, and combine forces across the science and technology enterprise to increase capacity and expertise.

The National AI Initiative Act (NAIIA) of 2020[13] established the National AI Initiative Office (NAIIO) to coordinate key AI activities across the federal government. The NAIRR taskforce, charged with investigating the concept of a "NAIRR" and whether it was feasible published its findings in January 2023[14]. The program envisioned in the report is large in scale, but that is not where we are today. No funding has yet been appropriated for the NAIRR program. However, in October 2023, an executive order came out[15] that directed NSF and collaborating agencies to launch a Pilot for the NAIRR within ninety days.

NSF had to move quickly to launch a pilot. Using existing resources and in-kind contributions from non-governmental partners, the NAIRR Pilot aims to demonstrate the value and impact of the concept, support novel AI research and education, and gain experience to advance and refine the NAIRR design with early lessons learned.

NAIRR Pilot Users will be AI researchers, domain scientists applying AI, and students and educators. They must come from U.S.-based institutions. Some initial research thrusts are: 1) AI

---

[11] Pelican: Advancing the Open Science Data Federation Platform, NSF Award #2331480, https://www.nsf.gov/awardsearch/showAward?AWD_ID=2331480 .

[12] The website for the NAIRR Pilot is: https://nairrpilot.org/ .

[13] "Text - H.R.6216 - 116th Congress (2019-2020): National Artificial Intelligence Initiative Act of 2020." Congress.gov, Library of Congress, 12 March 2020, https://www.congress.gov/bill/116th-congress/house-bill/6216/text.

[14] National Artificial Intelligence Research Resource Task Force. "Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resource**."** January 2023. https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf .

[15] "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." October 2023: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ .

safety and reliability, 2) health, 3) resilience of public infrastructure including agriculture, water, and grid infrastructure, 4) advanced manufacturing systems, and 5) earth, environmental, and climate challenges. These thrust areas are intended to be broad umbrella areas for the first part of the NAIRR Pilot launch.

The NAIRR Pilot is providing infrastructure, it is not funding individual research. NAIRR Open will be providing access to resources through a portal. The Department of Energy (DOE) and the National Institutes of Health (NIH) will be leading the NAIRR Secure Thrust. NAIRR Software will allow the community to take a step back to think about how we would design a future NAIRR. Right now, we have a potpourri of solutions. NAIRR Classroom is about reaching and training new communities on how to use the resources and enabling educators and students to get hands-on access.

Contributions to this pilot come from a wide range of organizations including government and non-government.

Several activities were launched in January 2024 including a survey of the top need for research. Most researchers indicated that computing was their top need, followed by education and training. The second-place need was data resources and services. There is clearly a demand for resources. The Allocations Working Group will be handling getting processes up and running.

There are quite a few data and networking challenges along with some opportunities. You cannot do AI research without large amounts of data. There are challenges related to the size and complexity of the data, the source of the data, how to discover the data, and whether the datasets are appropriate for AI research or not. We have been thinking hard about how to handle these challenges in this pilot and what this pilot can do that will most inform the direction of the full NAIRR.

The strategy moving forward is to address the data challenges through specific use cases and demonstration projects. How do we ensure that we are doing this in a responsible way with transparent processes? What type of training should be provided?

We are not only piloting the processes, but also the policies. There are a lot of questions that need to be tackled. The Pilot is trying to adhere to the NAIRR guidelines for governance described in the Taskforce report, under the current guidelines of its authority. There are some differences. NSF is the administrative home for the NAIRR Pilot managed by the Pilot Program Management Office (PPMO).

The NAIRR Pilot Interagency Working Group (IWG) includes NSF, EOP, and interested federal agencies. The NAIRR Pilot Steering Committee includes NSF and Representatives from EOP and contributing partner federal agencies. The NAIRR Pilot Advisory Subcommittee advises on pilot policy, ethics, and technical issues. The Non-Government Partner Forum flows into Pilot Program Management Office (PPMO) under this the Pilot Operation Teams and WGs: NAIRR Pilot Portal, NAIRR Open, NAIRR Software, NAIRR Classroom, and NAIRR Secure.

There is no shortage of challenges and that is why the pilot is needed. We learn by doing. Community engagement and design is imperative to the success of the NAIRR pilot.

### 3.1.3 Questions and Discussion

Comment*:* I would like to note that although the high-end GPUs are hard to get, you can get any number of the gaming GPUs. They are perfect for training students, and they are not that expensive.

Question*:* What will help NSF measure the success of NAIRR pilot?

Response: At the end of the two-year pilot, NSF wants to look at whether we are reaching new communities, or piloting new communities. Has the pilot supported some new science that would not already have been done? In terms of the lessons learned, have we as a community identified key gaps on which to work or produce great blueprints? What are the documented lessons learned?

Question: I am curious about data egress costs for cloud storage. If we have data, how do we avoid the vendor lock and siloing of that data? Is there some solution to egress costs part of the agreements with the cloud provider contributions?

Response: Different agencies have taken different approaches to this problem. NIH and NASA have big cloud contracts. Whereas NSF has not negotiated anything different. The pilot will be getting into some of those details. It is critical that as a community we think about where we cache our data. If you put data in the commercial cloud, are you committing to doing all your computing there, or do we put caches around the country? We need a discussion about this.

Question: I recall at the Supercomputing Conference in November; you had a slide showing seven to eight different NSF projects that had eight different interfaces for the user community. One of the challenges that you said you would be working on is trying to bring that all together to create one interface for making is easier for users. Will you be doing that for the NAIRR pilot to make things easier for the user community?

Response: I have been a part of a lot of portability projects that try to get portability across architectures. We need to be careful here. We are certainly trying to make it easier for the research community to access resources. Are we trying to get a uniform environment across all these different providers? I would send this question back to the community. What is the priority around that? We want to try to move towards some common software stack at a limited level for people to start and get moving. But there is a huge diversity of resources that are contributed. We do want to have a portal that is easy for users to access. For the pilot it is more a potpourri of different options.

Question: There are 3,900 accredited institutions of higher learning. The history of allocations in NSF typically manages a dozen. How do we get from a dozen to a thousand? We know how to run applications programs. They have a finite throughput, and they are scratching the service of what is happening in the nation. Does NAIRR aspire to go deeper than just allocations to a few things and how would it be done to account at least or have a view of everything that is happening that is not allocated?

Response: My experience is that whether you are a beginning researcher or even the most advanced, most start with Jupyter-like notebook. For many communities that is going to be a sufficient resource and a very critical resource.  Many universities cannot handle a large data center. We need to move forward with Infrastructure-as-a-service. This is our opportunity to reach those universities that do not have the capital to invest in their own data center. They can and should be part of the cyberinfrastructure and AI research community. I think we can provide many communities with access to notebooks. They do not need to be through allocation programs at all. The scale of compute needs can be small enough that the overhead for many communities to apply for compute may be too high. But on a larger scale these are expensive resources, and we would want to make sure they are being used wisely and the resource requests are being reviewed and allocated appropriately.

Question: There exist many communities that have a broad outreach for education such as the Campus Research Computing Consortium (CaRCC), campus campions, and regional networks. How does NAIIR plan to leverage those communities?

Response: We are just starting, but those are going to be key communities not only in getting the word out but also being navigators on how to access the resources. We have just stood up a user experience working group. I think those communities that you identified are going to become critically important to reach into those pockets of the country that did not hear the first message.

Question: Is there an opportunity for AI startups?

Response: There is a requirement that there be a federal grant for startups to apply for Pilot use.

Comment: I run the HPC systems on my campus and what I get a lot is individual researchers who want to dabble in it, such as microscopy and they need a GPU with a lot of memory. Then I somehow need to find a way to provide that tool and it does not scale terribly well, but this ability to give people a way to easily experiment is an important thing to provide.

Response: I agree. I think that is critical.

Question: One of the things some of our researchers have been looking at and that we have been curious about with NAIRR is one of the goals or pieces of the research thrusts or the core trustworthy AI drive to combat implicit bias in large models? And if not, is there a place in the federal government or in other open calls, some additional consideration for that?

Response: You would put it in your proposal submission. I think addressing bias would be in the AI safety, privacy, trustworthy thrust area. I think that is in scope.

This presentation may be viewed at: [2024 5NRP - Katie Antypas](2024 5NRP - Katie Antypas) .

## 3.2 Status of the National Research Platform – Frank Wüerthwein, SDSC Director

One of the objectives is to talk a little bit about the larger vision where NRP is one element of this much larger vision. Within this talk, there is a distinction between the NRP community and the NSF-funded project “Prototype NRP” or PNRP.

There are three challenges we are trying to address as a community:

1) The gap between those who have and those who cannot afford is getting wider.
2) Cyberinfrastructure (CI) needs are growing for Education. AI is accelerating this.

3) The end of Moore's Law is leading to a proliferation of architectures. This is not something a domain scientist desires.

First, the overall vision is of an open infrastructure that is both vertically and horizontally open. Fundamentally the vision is to create an open national cyberinfrastructure that allows federation of CI at all the accredited degree granting higher education institutions, non-profit research institutions and national laboratories. Open infrastructure means open compute, open storage and CDN, and open devices and instruments. This, combined with open science, open data, and open sources leads to openness for an open society.

Why do we want to do this? The why is democratizing access. There are 3,900 institutions of higher learning. The US has an enormous range in academic institutions, from hundreds of students to many tens of thousands of students, from institutions that are a few buildings to campuses that have multi-billion-dollar research revenues. Being able to serve all is the ultimate goal, which seems an almost impossible goal.

A government agency would never fund something like this. The project is too big. The alternative is for a community with a shared vision to enable Principal Investigators (PIs) to respond to areas not already funded and that overlap the shared vision. Seen as a daisy with shared vision in the center and funded projects as the petals, some examples of the petals are PNRP, PATh, NDP, Pelican, Fusion, R&E platforms including GP-engine and TIDE. Open infrastructure is "owned" and "built" by the community for the community. The 20% shared resource requirement in some of the NSF solicitations is part of the shared vision and is part of the open infrastructure that is "owned".

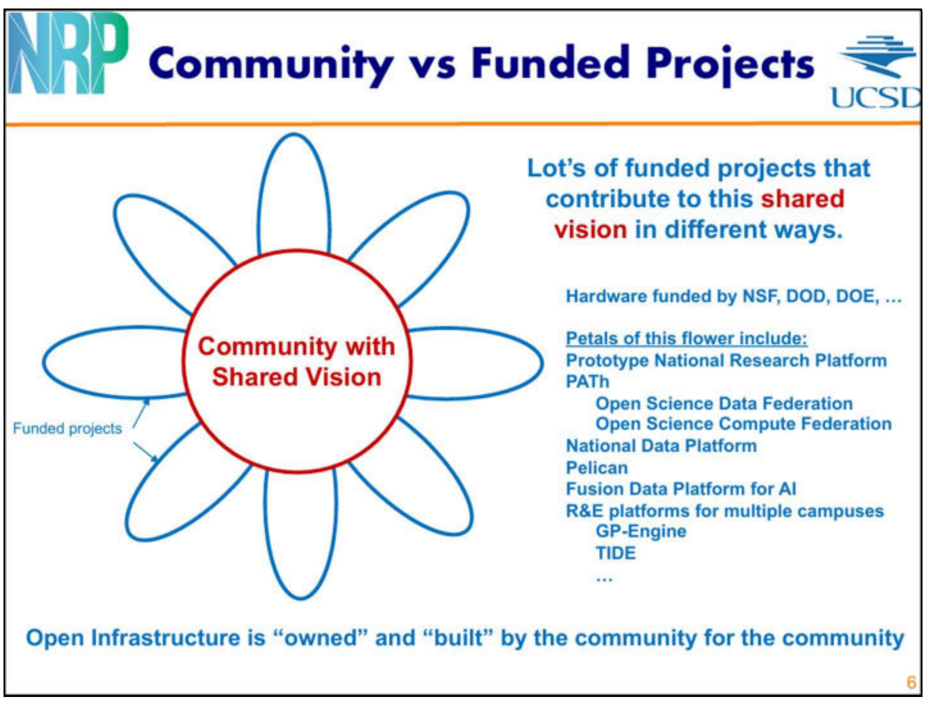


*Figure 1: Community vs. Funded Projects petal diagram.*

So how does it work? How do we execute this vision? To accomplish this, we need a flexible architecture to build on horizontally and vertically. One-size does not fit all.

NRP is a non-local extendable container deployment platform, thus allowing many uses unthinkable for a cluster in a data center. If you can build a containerized thing on top of Kubernetes, you can build vertically.

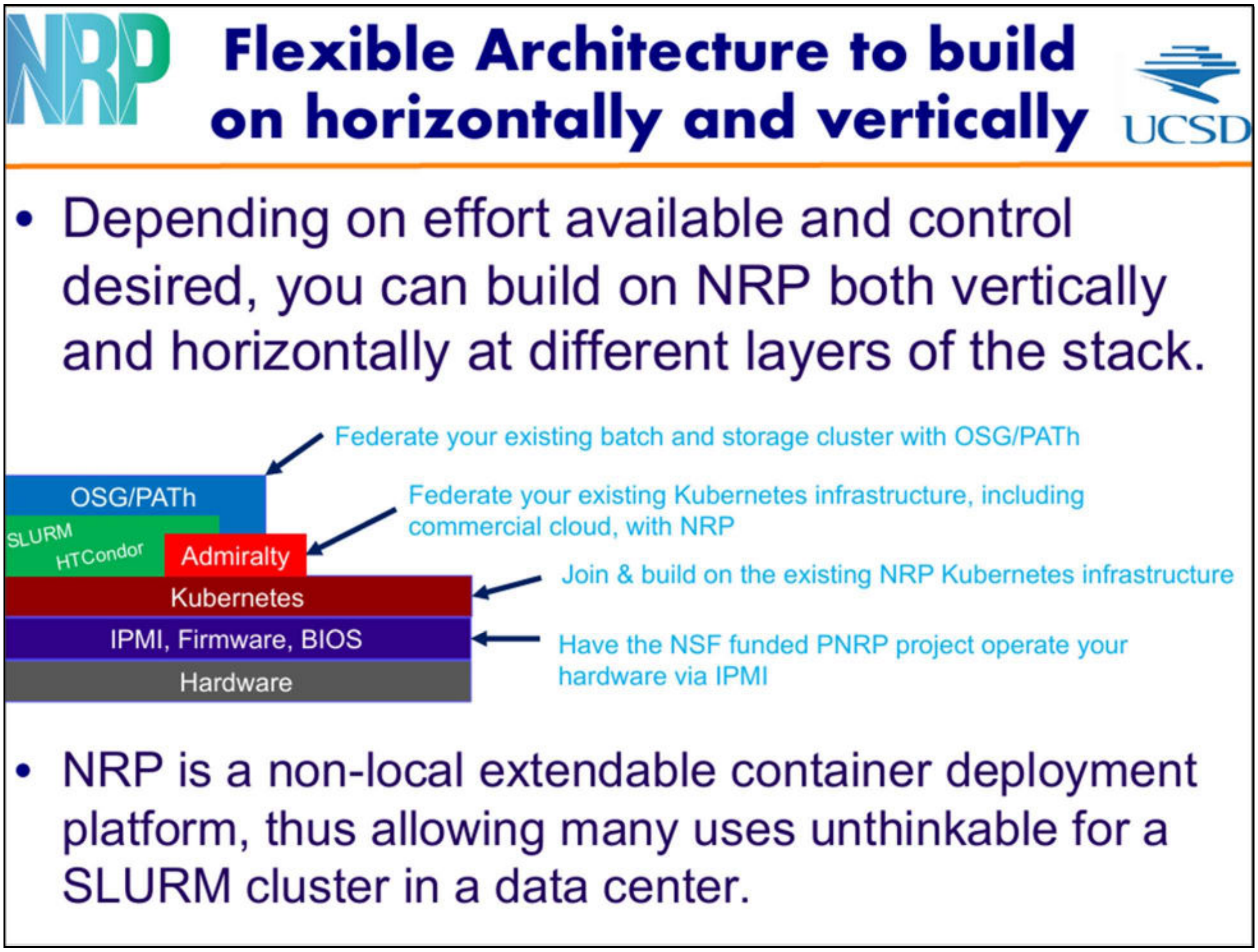


*Figure 2: NRP Flexible Architecture to build on horizontally and vertically.*

There are sixty-two institutions integrated into NRP. Forty-two integrated at the IPMI layer. Twenty integrated at the K8S layer. What drives it is the notion that it takes effort to exert control, an effort that not everyone has, or wishes to spend.

### 3.2.1 NRP as an AI platform

What is on the platform? There are thirteen petabytes of storage, 1,262 accelerators, 21,070 x86 CPU cores,7,406 of which are in nodes that have no accelerators. NRP can be viewed as the NVIDIA museum as it has every GPU produced since the 1080's except the H100's. In AI infrastructure the user can figure out what they should be buying and get a handle on how much money really needs to be spent to get the job done. This is one of the things that NRP provides as a side effect of being so diverse and heterogeneous. NRP could donate a few petabytes of storage to NAIRR. There is enough room to host all the NASA data desired.

Larry Smarr's talk on science in the NRP, highlights eight science speakers that used up to 240 GPUs and 5,000 CPU cores daily over six months on NRP.

The Open Science Data Federation (OSDF) as an "application" on top of NRP is very unusual because it is a distributed application that implements a content delivery network (CDN) of sorts. It is run by PATh[16]. It has the notion that data is stored in origins and access is via caches. There is a global namespace that is federated. OSDF uses only three components to bring order into the chaos of data: namespaces, caches, and origins so the data is always available. Six out of the sixteen origins are on NRP, and nineteen out of thirty-two caches are hosted on NRP. PATh combines all three layers, integrating over two hundred institutions over five continents.

### 3.2.2 NRP as a bridge to education

Many smaller campuses care more about educational use than research. Jupyter allows both to co-exist on one platform. Several talks discuss in more detail how the community is approaching this: GP-Engine, TIDE, and the CENIC-AI Resource (CENIC AIR).

There is much more to compute and data needs than just AI. Education needs cyberinfrastructure that goes beyond just AI. It is embedded into all the STEM fields at this point. UCSD's Data Science/Machine Learning Platform for Students (DSMLP) is one such effort aimed at providing AI and ML resources for classrooms and students in all disciplines.

### 3.2.3 NRP brings Computer Science and Domain Research and Development onto the same platform.

Bringing computer scientists together with domain scientists, we can blur the lines between testbed and production CI, allowing interesting things to happen. NRP tries to do this. The Chief Technology Officer of AMD noted that there was a tradeoff between the general-purpose CPUs, and more narrowly focused, but better performing GPUs and that in future the GPUs will be even more narrow per silicon used. The loss of Moore's Law drives architecture innovation and investment. The Semiconductor Research Corporation is giving out grants to develop next-generation architectures for science.

One of the objectives of NRP is to be a toy store, a playground of technologies to put weird stuff into the platform and play with interesting new technology. NRP can do this because it can grow horizontally and vertically. But what about all those dumb devices that can now be purchased with external compute capability, each with their own computing paradigm? This paradigm per device is scary because it necessitates the need to re-engineer software and architecture.

One of the reasons NRP is collocated with FABRIC is to connect the infrastructure at the networking layer and provide access to compute and storage.  NRP peers with FABRIC at 400G in Los Angeles with multiple networks via a 400G Arista switch. As an example, a student can select which loop they wish to use. Demonstrations have worked with ESnet to create various RTT (round-trip time) loops across FABRIC, NRP, ESnet, CENIC, StarLight, Internet2 and other networks. NRP would like to encourage FABRIC users to come up with ideas.

NRP has a very ambitious vision. It is open both horizontally and vertically. Given that the PNRP project that drives NRP is only a year old, we are off to an excellent start, as we are built on other people's work. There is more to come over the next five years.

---

[16] The PATh project webpage is at: https://path-cc.io/ .

### 3.2.4 Questions and Discussion

Question: A couple of questions related to the community aspect: 1) In chameleon we are seeing an increased need for infrastructure from education? Are you seeing this too? Are you aware of any studies that put any numbers to this phenomenon?

Response: I do not have a study that measures it. I see it at my own institution, and I see it across NRP. More and more institutions have discovered NRP to satisfy their needs for academic needs.

Comment: For tools, NRP should talk with Chameleon. They had a user meeting focused on education last year for the same reason.

Question: The second question relates to NRP as a vehicle for socializing computer science and domain science research. This is something that in which Chameleon is interested. Are you just doing it via the communities using the same platform, or do you have a specific social structure around it such as events? How do you promote that socialization?

Response: As with most successful things I have done in my life, I make it up as I go. We wrote into the PNRP proposal a matrix, and the idea was that PNRP had different thrusts that were technological and different science domains we wanted to serve and highlight. We hypothesized that the creation of matrix will result in the desired outcome. There is still work to be done in figuring out processes that we have not yet done, including the integration of NRP and Chameleon.

Question: For the next five years, do you foresee any system level development that will look after workflow? In the Global Network Advanced Group (GNA-G), we foresee a mismatch between the end-system's raw capability and the foreseeable affordable network infrastructure. So, do you foresee that NRP, and its partners can build a sort of end-to-end system, including workflow management, agent-based architecture, to help with executing a series of tasks over the next five years?

Response: The short answer, NRP cannot solve this problem. The funded PNRP cannot do that, but the NRP community may be able to solve this. This goes back to the petal diagram above. There is space for the petals to figure out which solicitations out of the NSF are the appropriate avenues to push these things and I am happy to write a letter for anybody that wants to do that and wants to build on what we have with NRP. Figuring out which of these paradigms makes sense is very hard and experimentation is required.

Response: Some new startups are succeeding well. They use some of the paradigms listed, but we cannot afford those products because they are too expensive.

Comment: We need a deeper discussion about this.

Question: What do you think about the future of RISC-V[17] in HPC?

Response: I am waiting to see the Europeans sort this out. The first machine of size will show up in Europe.

This presentation may be viewed at: [2024 5NRP - Frank Wüerthwein].

[17] https://riscv.org/about/

## 3.3 Turing meets Shannon: AI meets Wireless, Bringing AI to wireless sensing, communication, networking with NRP – Dinesh Bharadia, UC San Diego

The focus of this talk was all weather perception for autonomous systems. This project uses a lot of NRP resources: peaking at 65 GPU hours and 633 CPU-cores for the last six months. It also touched on autonomy at large and sensing for vehicles.

### 3.3.1 All weather for autonomous systems[18]

This project started in 2017 and is trying to build a perception system that works in all-weather environments. Both graduate and undergraduate students are part of this project.

Autonomous driving is the new "black." It includes everything from small cars to trucks, tractors, and construction vehicles. There is a local company called TuSimple[19] trying to do autonomous trucks as it is becoming harder to find truck drivers. John Deere[20] is working on autonomous tractors.

There are five levels of autonomy. Currently none of the self-driving vehicles are above Level 3.

- Level 1 – cruise control and lane tracking
- Level 2 – self-driving
- Level 3 – Human Supervision Required
- Level 4 – Full autonomy in limited areas
- Level 5 – final aim, full autonomy.

We need to get to Level 5 before you can truly rely on these vehicles. So, what is preventing them from getting to Level 5? If they cannot drive better than a human, what is the point?

Imagine driving in the fog or inclement weather. The most common sensor used is LiDAR (Light Detection and Ranging). It captures the environment using dots or point clouds. As the fog increases, visibility declines. The wavelength LiDAR uses is very small and cannot pass through fog, dust, and other particles. That is physics.

Radar is different from LiDAR in that it uses a much larger wavelength, and it can then penetrate the particles and only gets deflected by larger objects. There are lots of problems with radar but some of the good things are it can look behind a car and it can perform long range sensing.

Safe and reliable autonomy in adverse conditions can be enabled by active sensors at mmWave frequencies and by developing systems that can provide high quality data, semantic understanding, and contextual information of the scene. This is what this project investigated.

---

[18] https://ucsdwcsng.github.io/vehicle.html
[19] https://www.tusimple.com/
[20] https://about.deere.com/en-us/our-company-and-purpose/technology-and-innovation/autonomy?CID=SEM_Brnd_enUS_GGLE&creative=Corporate&gad_source=1&gclid=Cj0KCQjwir2xBhC_ARIsAMTXk86plht6V4eIcaod9r9PvPR-vgwaHz-RObmepia6RYkm2eY8k3GCcEwaAtP-EALw_wcB

There were three fundamental research questions: 1) how can we improve the data quality? 2) how can we get semantic understanding? and 3) how can we get contextual information?

The team produced three papers that were delivered on each of these aspects. The pointillism delivers high quality data[21], RadSegNet[22] combines radar and cameras[23], and R-fiducial[24] provides contextual understanding such as a stop sign or red light.

The team created a rig and attached it to a Mazda that collected lots of data. Many companies use low quality radar, and those datasets are not useful. We needed high quality radar, but it is hard to use because of the amount of data generated. We needed to build custom pipelining software to get the datasets we wanted and then label that data. It turns out that this is very expensive.

Once we had the data, we hired graduate students and put all the data on the NRP. We then launched VNC browser type desktop machines and assigned each student a folder or two. Using an in-house tool, the students labeled different LiDAR sets. The labeled datasets were used to train models. The tuning of the parameters was many times faster with many parallel instances.

### 3.3.2 Autonomy At Large: Perception, Pathfinding, and Digital Twins

When we went out to industry and asked if they needed 10-centimeter accurate boxes for a 5-meter by 5-meter car or 20- or 30-centimeter accuracy, no one knew the answer. Random numbers were used. The only way to evaluate this would be to build the sensing and drive the car using that sensing. We would then know how many collisions and hits occurred. Since this cannot be done in real life, we did this using a Digital Twin.

So how do we prove that we have built something that is reliable and autonomous? We created high-fidelity simulation tools. One problem with these tools is that on one side we can go to something called high-frequency structure simulation (HFSS), which can take a long time. For example, a 5x5 meter room would take a day to simulate accurately. It is a fundamentally hard problem at this frequency to do accurate simulations due to the wave and the diffusion.

Using Carla[25] models, we built our own sensor model. Taking a lot of real-world data, we built a tool that accurately mimics radar. This tool is being licensed. Do not use Carla's radar. It is very inaccurate, taking LiDAR and randomly dropping a few reflections. Shenron[26] is the tool we created which is more accurate. This is then plugged into the Carla Simulator. To train the model, we

[21] K. Bansal, K. Rungta, S. Zhu, D. Bharadi, "Pointillism: Accurate 3D Bounding Box Estimation with Multi-Radars." https://wcsng.ucsd.edu/files/Pointillism_paper.pdf .

[22] Bansal, Kshitiz et al. "RadSegNet: A Reliable Approach to Radar Camera Fusion." arXiv:2208.03849 https://arxiv.org/abs/2208.03849 (2022): n. pag.

[23] Cheng, B., Saggu, I.S., Shah, R., Bansal, G., & Bharadia, D. (2020). S3Net: Semantic-Aware Self-supervised Depth Estimation with Monocular Videos and Synthetic Data. ArXiv, https://doi.org/10.48550/arXiv.2007.14511 .

[24] M. Dunna, K. Bansal, S. A. Ganesh, E. Patamasing and D. Bharadia, "R-fiducial: Millimeter Wave Radar Fiducials for Sensing Traffic Infrastructure," *2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring)*, Florence, Italy, 2023, pp. 1-7, doi: 10.1109/VTC2023-Spring57618.2023.10199374. https://ieeexplore.ieee.org/document/10199374 .

[25] CARLA: An Open-source simulator for autonomous driving research. https://carla.org/

[26] Kshitiz Bansal, Guatham Reddy, Dinesh Bharadia, "SHENRON: Scalable, High fidelity and EfficieNt Radar simulation," https://wcsng.ucsd.edu/shenron/

needed to drive a lot of scenarios which would normally run very slowly if done with our local compute. So, we created a workflow that could spin up 201 jobs on the A10 GPUs and what used to take a week in our lab took only four hours or overnight. Once we had a trained model, what used to take a minimum of four days took only hours. The iterations became rapid. We are close to finishing this work and publishing results thanks to Nautilus.

### 3.3.3 Questions and Discussion

Question: My first Tesla had radar. They dropped it for LiDAR and now they have brought radar back.

Response: The first generation of Tesla had LiDAR, but they were very low resolution. Tesla then realized that ultrasonic sensors could do some of the same things. The initial radar was a single chip, with one transmit and two receive antennas. The next generation (FSD 4.0) will have a high resolution two-chip radar. It will have three six-transmit, eight-receive antennas. Tesla has also spent a lot of time using cameras. Tesla is now Level3; Waymo is almost Level4.

Question: What is the cost of millimeter wave radar?

Response: Millimeter wave radars cost roughly $70 today. LiDAR instruments are quite expensive. I am guessing that at scale that they would be at least $700-$800.

We should have every possible tool to make cars safer and more reliable. LiDAR, radar, and cameras should be used in unison. Our paper does argue that you can use radar and cameras well enough. See the website for more information – http://wcsng.ucsd.edu .

This presentation may be viewed at: 2024 5NRP - Dinesh Bharadia .

## 3.4 Knowledge Graph-based Industrial AI for Future Manufacturing - Bingbing Li, CSU Northridge

Cal State Northridge (CSUN) is one of twenty-three CSU campuses, known to be teaching intensive, not known as research intensive. The campus has limited resources and over thirty-six thousand students. This talk focuses on how CSUN leverages NRP Nautilus.

Back in 2018, CSUN received a grant from NASA, and a grant from DoE for working on AI for the manufacturing domain. They initially received a $6,000 credit from Amazon but used it up in six months. They were then introduced to Nautilus. Utilizing GPU and CPU resources, in the past six months they used14,000 GPU hours and 80,000 CPU-Core hours.

They have two namespaces: *nsf-maica* and *cesmii-scw* that support multiple research projects.

- Smart Connected Worker
- Energy Disaggregation for Manufacturing plant
- Medical Image Restoration through Optical and CT Scanning
- Graph Representation Learning for Material Prediction and Recommendation in CAD Automation
- Knowledge Graph Construction through the Potential of Large Language Models within Manufacturing
- Multi-Domain AI for Metal Additive Manufacturing

They used Large Language Models (LLMs) which require large memory for training and fine-tuning. They have multiple publications using results powered by NRP Nautilus.

One of the initial two-year projects was the DOE Smart Connected Worker Project. Large manufacturers have been unable to capitalize on the advancements in digital technology to dramatically increase productivity via smart manufacturing. The high capital and human cost of transitioning largely manual operations to a more data driven efficient manufacturing environment has left small and medium enterprises behind. To leverage the new technology, they worked on creating a Smart Connected Worker System (SCW) using affordable consumer grade hardware, open-source software, accessible machine learning intelligence, and portable wireless devices.

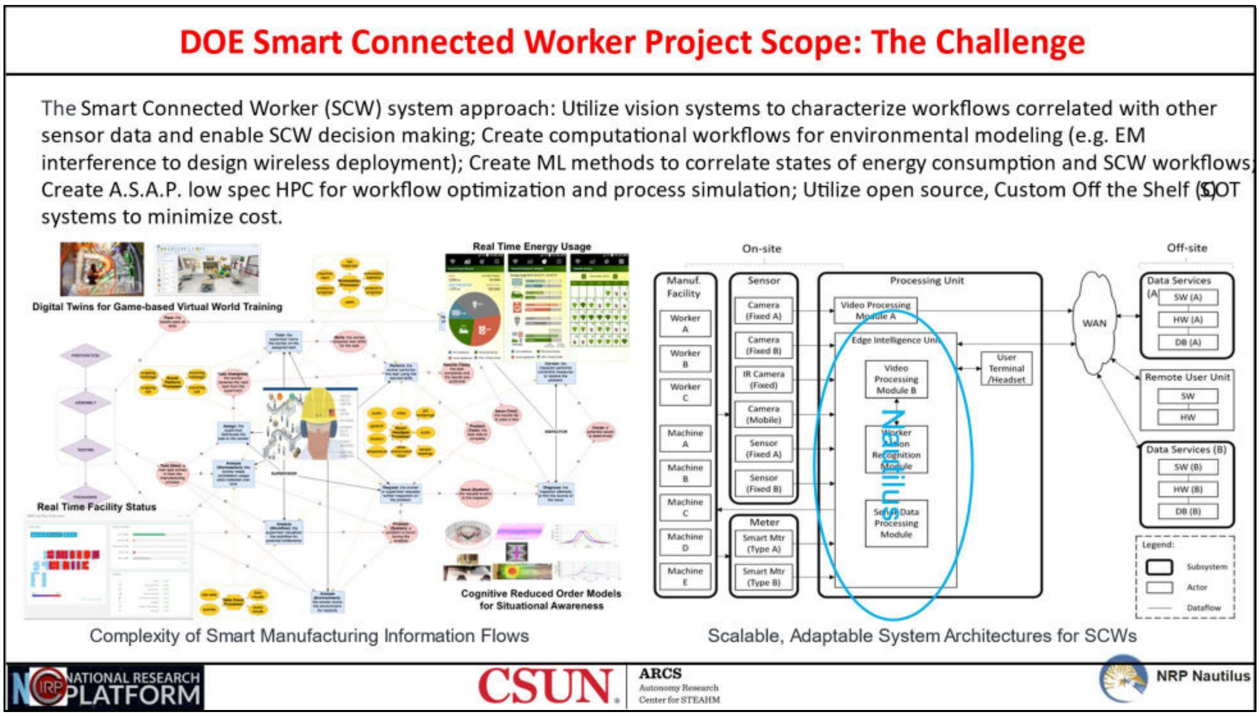


*Figure 3: DOE Smart Connect Worker Project - CSUN*

Another project was the Digital Modeling of Large-Scale Additive Manufacturing (LSAM) and Augmented Reality (AR) of 3D printing Operation. They created a digital model and a digital twin for the lab. Working with a small startup that only had basic step-by-step guidelines and basic training, the project team realized that they could not track finger gestures for good human-machine interaction. Using machine deep learning, they were able to develop an application that could be embedded into the company's existing platform, thereby improving the training experience.

The metal additive manufacturing machine is extremely expensive as well as the training to use the machine. They needed to pay $5,000 a year for the Autodesk engineer to come to campus to do annual maintenance service. To reduce the cost and get rid of human operation, they developed an immersive augmented reality training system using Nautilus. Users can interact with each part of the machine.

The California Energy Commission funded work on creating a framework for real-time energy disaggregation. The project explored how to use just one meter to be able to know each machine's state. Collecting all the data in the lab and then using long short term memory (LSTM) methodologies, they process the time series data using a lot of GPUs and CPU hours.

They received another award two weeks ago from NASA on a proposal for efficient deep learning in space for the space computer.

The most recent research work they are conducting in the manufacturing domain is how to leverage machine learning and the large language model to build a knowledge system, extract the knowledge, and then bridge the gap between the knowledge you know and the extraction. This is more important as senior personnel retire or are laid off, as their knowledge leaves the organization. In one case, the retirement of a senior manager led to the company not noticing dust accumulation under a conveyor belt, which caused a one-month stoppage of work and close to a one-million-dollar loss.

The project team used Apache Tika and ChatGPT to process more than 1,000 data file formats including text, images, x-ray images, thermal images, and CTIC signals. One of the key findings of their fined-tuned Alpaca-33B model was that it surpassed GPT-3.5 by 10.4%. The aim is to develop more specialized LLMs for various manufacturing domains.

### 3.4.1 Questions and Discussion

Comment: I took a picture of your various projects and am going to show it to our faculty when they complain about the amount of teaching they must do. I am impressed.

Response: It is hard to conduct research in a teaching-intensive university with limited resources. I really appreciate the NRP and the access to resources. NRP and Nautilus helped us in the past five to six years, otherwise we would not have been able to do the work.

This presentation may be viewed at: [2024 5NRP - BingBing Li](2024 5NRP - BingBing Li).

## 3.5 Learning Humanoid Robots - Xiaolong Wang, UC San Diego

Using several namespaces (*vision-rl, ece3D-vision, rl-dev, rl-work, and rl-multitask*), the GPU usage on Nautilus per day for the last six months peaked at 170 GPUs and the CPU-Core hours at 1,050 with totals of 216,000 GPU-hours and 2,021,000 CPU-Core-hours. This is due to the reinforcement learning research we are conducting.

Currently humanoid robots are the missing components for whole body control. Two additional essential components are perception and tactile sensing. Most of these are using reinforcement learning to train the robot in simulation and then later to transfer to the real robot.

### 3.5.1 Expressive Whole-body Control for Humanoid Robots

In a video, a robot interacts with a human using its upper body. It then moves along with the human at the same time, like dancing. Exploring the more expressive whole-body control, the robot can "express" itself. In another video, a robot mimics human motions, dancing to music.

It is difficult to make simulations work in the real world. Unlike the simulated motions seen in video games, real robots do not have that degree of freedom. Robots have very limited angles. In the simulation there is 60 to 70 degrees of freedom, whereas with the real robot only 19 degrees of freedom. The approach used is to split upper and lower body motions and use a different kind of reward to do the learning. The lower body does not have much mobility, so the lower body is just kept stable while the upper body does the imitation.

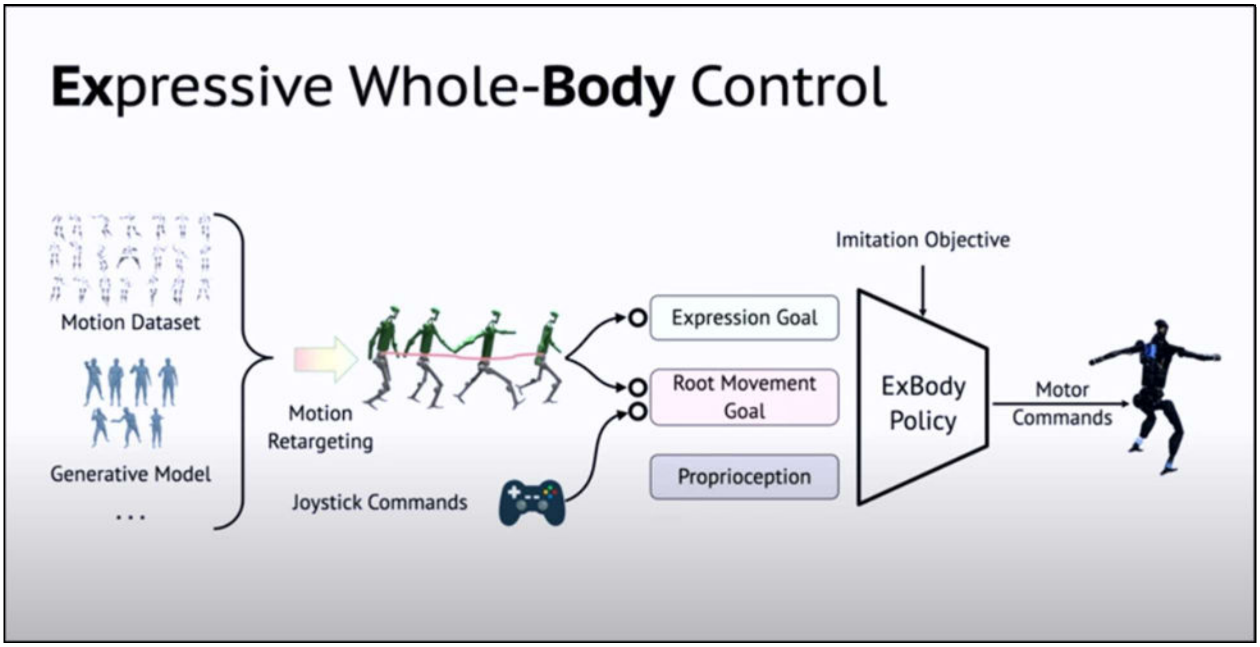


*Figure 4: Expressive Whole-Body Control for Humanoid Robots*

We are using a large motion capture dataset of about eight hundred moves of a human skeleton and then they convert those moves to robot motions. We train robots by making the environment very difficult. The video clips show balls being thrown at a robot, the robot being pushed by a student, walking through grass, etc.

### 3.5.2 Perception

Working in 3D perception, the robot as it moves has an egocentric view of what it perceives. An example is shown giving language instructions to a robot and watching it autonomously complete the task. We propose a representation containing 3D geometry for planning, semantics for answering human instructions[27] with colorful maps built in real-time. This is top down, so the robot keeps building the map as it is moving. If a person suddenly appears in front of it, the robot learns to avoid the person and go around them to complete the task.

[27] Ri-Zhao Qiu, Yafei Hu, Ge Yang, Yuchen Song, Yang Fu, Jianglong Ye, Jiteng Mu, Ruihan Yang, Nikolay Atanasov, Sebastian Scherer, Xiaolong Wang. "Learning Generalizable Feature Fields for Mobile Manipulation." https://doi.org/10.48550/arXiv.2403.07563 or for the project website: https://geff-b1.github.io/

### 3.5.3 Tactile Sensing

An important component of the humanoid is its hand. We want humanoids to move objects and manipulate them just by feeling. We are going to be using multi-finger hands, not just parallel grippers, as they provide more robust stable grasping and flexibility in manipulating objects. Can we enable robots to manipulate and "sense" an object without seeing?

Simple sensors can work. Sixteen low-cost force-sensing resistors attached to an Allegro hand, can have continuous signals or binarized signals. We binarized the signals both in the simulation and real world. The reason for this is that when your tactile sensor is too complicated it does not transfer to the real world. We made this extremely simple using just zero and one, which then becomes just geometry. We can then transfer with almost no gap. It works well.

A takeaway from this work is we need to think about hardware design and learning together. We can use simple hardware and use learning to do the rest.

### 3.5.4 Questions and Discussion

Question: With the autonomous robot that is picking up cups and things, have you explored the idea of using LLMs to train or natural language to give commands and then transfer into the API of the robot.

Response: Yes, this is something we are extending on. We have a simple semantic map, and we are just doing one step of action, just grasping. We are doing long horizon planning with multiple steps of tasks and are exploring using LLM to break down tasks into different subtasks, and then follow those subtasks using this feature. A more ambitious goal would be to clean up all the trash on campus, which would require many subtasks and some planning from the large language model. We do integrate this high-level planning into the current pipeline, but all the things introduced today are low level skills that can be smoothly integrated with the LLM.

Question: A quick comment. A couple of weeks ago at the Internet2 Community Exchange the Keynote Speak was Kate Darling of MIT and she talked about how people get attached to their robots. You mentioned that the students were happy because they got to dance with the robot, and I am wondering if you are observing whether your students are getting attached to these robots. When one falls over do they get sad?

Response: I am more worried if people get creeped out than if they get attached to the robot.

Comment: The creepy robot you showed always has his knees bent. Humans normally have knees straight.

Response: That might be a problem with our algorithm. Our current reinforcement algorithm leans toward walking and bent knees. The robot has only one degree of freedom. If it had two degrees, it would be more stable.

Comment: I can propose to you why his knees are bent. As a surfer your knees are always bent because it is much more stable. Maybe your robot learned something about the ability.

There are multiple videos embedded in this talk. This presentation can be viewed at: 2024 5NRP - Xiaolong Wang

## 3.6 Machine Learning on Secure Human Time Series - Benjamin Smarr, UC San Diego

This talk discussed an application of the NRP that is underutilized, and that is integration into health research. In 2020, the DoD asked Dr. Smarr to help figure out if commercially available wearable devices could be used to detect when people are getting COVID. That work allowed collaboration with many people and led to some Public Health implications about the ability of NRP to reach underserved communities that cannot afford their own CI infrastructure. There is a lot of transformative potential there.

People generate more data than hospitals. That is a big deal. We know distributions do not reflect the patterns of change. We knew this from wearables with temperature ten years ago. If we can train ML/AI on pattern features, they can be much more precise. If we can measure a person's physiology continuously, more interesting things could be determined from the patterns of change and could make predictions about health trajectories more intelligent.

When we train machine learning algorithms from large data objects that have a temporal structure, unless we make a point of explicitly specifying the temporal structure, the data can wash out unintentionally and we can lose all the information that is contained in that time series.

It is hard to get machines and clinicians to appreciate the differences. Perhaps there is a way where data from wearable devices, or near body devices that capture an individual's biology over time might be used, featurized, and trained to support medical decision making for people where it is harder to measure them.

So COVID started, and the DoD funded the project TemPredict (namespace *TemPredic*t) to investigate the use of wearables. The company Oura[28] allowed the project to request users to donate their data and tell them over the year whether they were experiencing symptoms, whether they were getting COVID, and so forth.

By November of 2020, we had a working algorithm that was able to detect if people were getting sick. On average we were able to detect COVID two or three days before confirmation. To do this we needed to preserve the rhythmic structure of the data. All the time-of-day information blurs out if we aggregate the data.

The challenge is to determine the right way to featurize all the signal processing that goes into pulling out parameters of these data so that they can be fed into machine learning models.

We are not all physiologically the same, and we need to grapple with what is the baseline, what is normal for an individual. Multiple factors can be at play including ovulation or pregnancy. A graduate student, Lauren Bruce, did work on being able to class some women as cyclical, showing that sex is not the right proxy. You do not need a proxy if you have time series data. She built a classifier tree that allows us to classify everybody by their dynamic time warping distance of their autocorrelative curve. If we look at somebody across time and take their time series and shift it a day, we can get a correlation of how similar that person is to themselves, and so forth. This led to us

[28] Oura's website is at: https://ouraring.com/

being able to remove false positives and increased accuracy into the high nineties. This turned out to be a much more equitable algorithm across the entire population.

Philosophically there is still much work to be done to know what is physiological diversity that makes people look different from each other not in terms of their genes, skin color, or sex, but by what is happening in their body day by day.

This trial, where we were alerting medical first responders about COVID exposure, was only possible with the resources of NRP. A slide of the CPU usage shows over a six-month period, 190 GPU hours and 178,000 CPU-Core hours.

Just showing that something is possible does not make it deployable in the world. Perhaps it should be more efficient, more parallelizable, or more scalable. Machine learning can fail if not given the feature sets, the right signal processing of rhythmic structure on which the physiology is based. And it needs to be coupled with unsupervised learning. Unfortunately, in physiology, we still do not know what the right parameters are. No one has ever seen physiology continuously over years like this and so this is a new field.

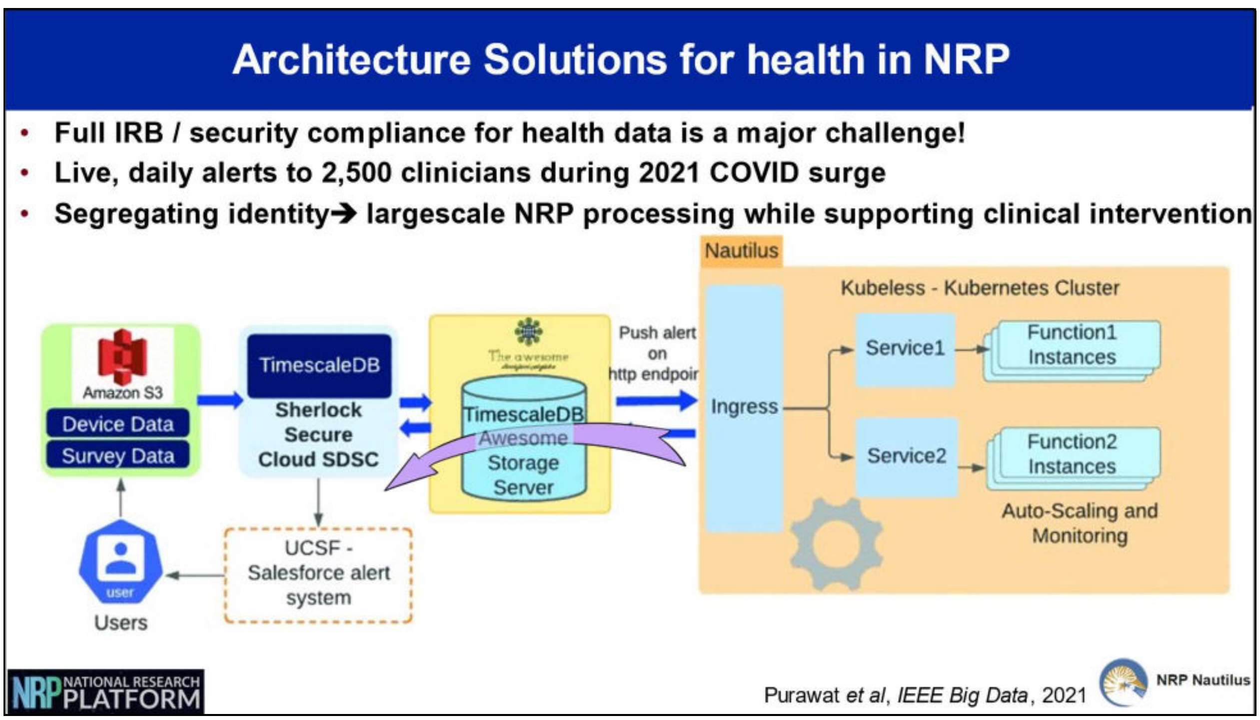


*Figure 5: TemPredict - Architecture Solutions for health in NRP.*

We needed to get full IB security compliance for health data, which is encrypted and hashed, but we needed to be able to look at the data fast enough across thousands of people to send out an email early in the morning alerting people who needed to test for COVID. It worked. We segregated the identifiable data from the data that was being de-identified, chewed up, and processed, encrypted, and coded. Only the IDs that looked like they were triggered by the algorithm were pushed back into a hashed version and put into the secure system Sherlock and from Sherlock were able to open some secure portals. The personal identifiable data was never violated.

This was thrown together in just two to three months.

Tech access is not equitable and there are some serious cultural trust barriers. It is going to take a long time to figure out what is the safe and equitable non-creepy way to provide this kind of support. The potential is transformative for public health and individuals and the science problems are coupled with some interesting social problems. We are now collaborating with a few minority serving institutions. Eastern Illinois University is working on getting onto NRP and their students will be able to learn how to do this cutting-edge medical research without having to have the infrastructure locally.

### 3.6.1 Questions and Discussion

Question: What happens next? Is COVID still relevant? What is the next problem you want to address?

Response: We have moved a lot of this towards behavioral issues. Our genes are not fundamentally changing day-to-day, but our physiology changes all the time. What does it mean to have a digital twin in some meaningful way that affects medical decision-making?

This presentation may be viewed at: 2024 5NRP - Benjamin Smarr .

## 3.7 Neural Radiance Fields for View Synthesis - Ravi Ramamoorthi, UC San Diego

The goal of this work is to enable virtual experiences of real-world scenes. A user should be able to click a few images with their phone, input those imagines, and then enable an immersive visualization that can be used in virtual and augmented reality, transforming the 2D images into 3D ones.

Formally known as view synthesis, images of a scene are viewed from new viewpoints in a full 3D representation leading to the need for a lot of compute. This work uses several different machine learning algorithms with multi-layer perceptron and convolutional neural networks, transformers, and diffusion models.

As seen in the GPU/CPU usage graph for the UCSD-ravigroup, some time is spent on generating synthetic datasets using computer graphics rendering that does not always run on the GPU. That is the primary reason for the spikes seen in the CPU core hours in the figure below.

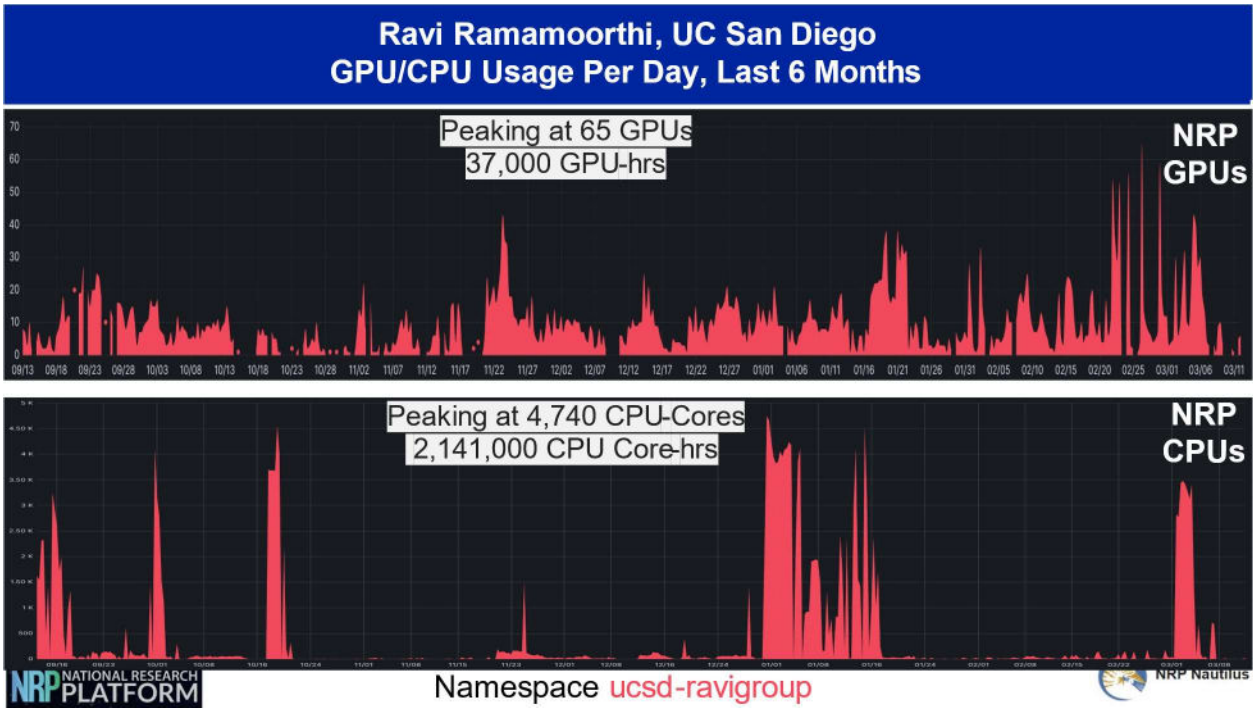


*Figure 6: Namespace ucsd-ravigroup GPU/CPU Usage Per Day, Last 6 Months*

The view synthesis problem goes back at least three decades, which is one of the reasons it is exciting now to use modern machine learning methods and the resources of the NRP to make a big advance in this area. In 1993, Shenchang Eric Chen and Lance Williams presented a paper on view interpolation for image synthesis at SIGGRAPH[29] that introduced the problem. In 1996, Lavoy and Hanrahan showed the basic idea of why light field rendering is possible in the first place in their SIGGRAPH 1996 paper.[30] Taking several images around an object or a scene, one can interpolate the individual images and can go beyond the images and look at a point looking towards the scene and resample images from different cameras. This gives you an idea of how with the right kind of scene representation, you can take these images to create a three-dimensional representation of the scene. Up until 2020, these works were using relatively simple geometric descriptions of the scene known as disparity maps, which is the apparent motion of objects between two stereo images.

The work on neural radiance fields from 2020 introduced a fundamentally new description of the volumetric representation of the scene, predicting these volume densities and colors using machine learning with a multi-layer perceptron. Neural Radiance Fields (NeRFs) have led to a revolution in the ability to take images and create representations of scenes. Examples include The New York Times using it to generate workflows for portraits, Meta investing in it for 3D object scanning, and Time Magazine listing NVIDIA's Instant NeRF as one of the best inventions of 2022. Additionally, work on 3D Gaussian Splats for Radiance Fields produces higher quality faster.

---

[29] Chen, Shenchang Eric, and Williams, Lance. "View interpolation for image synthesis." https://doi.org/10.1145/166117.166153 .

[30] Levoy, Marc and Harahan, Pat. "Light field rendering." https://doi.org/10.1145/237170.237199 .

Examples of practical applications of these techniques include the ability to create a digital twin of your entire showroom or factory or capturing landscapes from online postings. These large language models can be coupled to neural radiance fields providing a bridge between text-based generation of imagery and three-dimensional models.

Students who have worked on this have gone on to Google and Luma AI. In the context of Google StreetView these techniques are used to create photorealistic renderings of interiors allowing users to view restaurants and other locations prior to making reservations. A radiance field can be created in real time as a bridge from 2D to 3D such as creating an avatar for use in 3D videoconferencing.

IARPA, the Intelligence Advanced Research Projects Agency, has introduced the WRIVA[31] program to apply neural radiance fields to large-scale environments where you want to take a few images from different altitudes and be able to navigate and create a photorealistic model.

We have added resources to the NRP Cluster: 25 8-GPU nodes, compute nodes, and a $100K storage server.

### 3.7.1 Questions and Discussion

Question: Most phones can do 3D rendering. Why do you need so many GPU hours?

Response: If you have multimodal data where you have images as well as 3D that can be included and you can get better results but in many cases we find that just using the RGB information is adequate and in fact one of the algorithmic innovations in neural radiance fields is not to use a representation of depth or a mesh or a disparity because those representations actually are fairly error prone and therefore we optimize a new volumetric representation so for many problems in view synthesis you are better off just core looking at the RGB images. There might be other cases where the depth sensor benefits you.

The second question in terms of computation so there are certainly neural radiance fields algorithms such as instant NGP or even gaussian splatting that are designed for a given scene to run very efficiently on your desktop GPU and in fact I showed some of the examples with nearly real-time results, however there are a bunch of other research problems where we need a lot more computation. For example, in the REVA data where there are large scale environments and you may be trying to analyze the variation across different times of day, or different seasons or in some of the language model work when you are using diffusion models to generate general scenes. There is a range of problems, which can be done efficiently on a single desktop GPU and some that require the full power of the platform. Also, creating synthetic datasets requires a huge amount of computation.

Question: Is there still a role for classification and memory? For every model that you generate can you store it?

Response: One of the interesting things, one of the things we worked on was generating NeRF representations and 3D scenes from a single image. For a portrait, there is a lot of previous

---

[31] Walk-Through Rendering from Images Of Varying Altitude (WRIVA) project page is at: https://www.iarpa.gov/research-programs/wriva.

information about human faces. In fact, the dataset was created entirely from a GAN model, not a synthetic image, no acquisition on a light stage. You generate sixteen million images from a GAN model, and you train the dataset.  For general objects now that we have these kind of general five billion element diffusion models. We have seen a remarkable advance in progress where you can look at the single image of a general object and by knowing what it is, that prior information allows you to create a 360-degree flow around. So far, we have not had large scale datasets in 3D and large-scale neural radiance field datasets but a startup such as Luma AI, people can take a million NeRFs and do machine learning. You can imagine ways of making further progress as well as storing it with a full range of models.

Comment: The next step would be interactions. Once you classify objects in a very general sense you can generate knowledge about how they interact.

Response: Some of the work now deals with physical simulation. You may wish to not only go around an object but may want to be placed in the real scene with the ability to interact.

Question: You showed an example of semantic mapping. What would it take for scientific instrument data like microscopes or LiDAR to be using these techniques to map.

Response: There is a large scope for that. In the field of tomography, it is recovering a volumetric representation and in face some of the neural implicit representations that have been used in neural radiance fields have been and can be used in tomography even for things like black hole reconstruction. The idea of general volumetric representations and neural implicit representations of them have broad impact even in the scientific domain.

Question: In terms of resolution, what is the best resolution you can achieve? For example, for 3D printing, we can convert the representation using 3D printing to fabricate.

Response: One of the key technical advances in NeRFs is the notion of positional encoding which is the multi-layer perceptron taking the spatial location and the angular coordinates and outputting a color.

There are several embedded videos in this talk. This presentation may be viewed at: 2024 5NRP - Ravi Ramamoorthi.

## 3.8 National Data Platform as a NAIRR data platform - Ilkay Altintas, UC San Diego

This talk discussed some examples of AI at SDSC from exploring new architectures to teaching best practices in machine learning AI to building methods to integrate AI into society and the applications at that scale. Equity is a real problem. It is like an onion. As you peel it, more comes out. There is a real difference between equality and equity. Equality equals sameness, whereas equity equals fairness. There are reports and studies on this. We can architect equity through technical approaches. Some of the open questions for equitable open research are:

- What are the foundational data abstractions, catalogs, multipurpose services, and expandable workflows for data-driven and AI-integrated application patterns?
- How can everyone effectively access and utilize these abstractions and services?

- How can services and workflows be developed and deployed on top of production-ready CI?
- How can equity be ensured for all to access and use CI from storage to the edge-to-HPC computing continuum?
- What are the governance and open science, open data and open CI requirements and challenges?
- What are the required guardrails for protecting privacy, civil rights, and civil liberties that will ensure a more equitable use of such data systems and services for everything from education to new AI training and application development?

Think about the user experience, bring in those workflows, ensure all have access and can use resources. It is not everything goes; we need to create an environment that has safeguards built into it. The dream is that all the walls are broken down.

We can architect research workflows for equity ensuring that everyone has access to research resources and AI. We need to involve diverse users to architect around their access, use and expertise gaps. As we understand and work through those gaps, we need to change the experience of working with data so that as communities work with data, they have a meaningful, purposeful experience and they benefit from that experience to generate insights and solve their own problems. We can find and innovate around existing public-private partnerships.

Creating an ecosystem approach to capacity building with services, platforms, and education, we can incubate use-inspired solutions to scale. We can explore new models for allocation such as service units, credits, tokens, and aggregated workflow cooperation.

When we started with the National Data Platform this is what we had in mind. We knew building blocks or puzzle pieces existed but wanted to address the missing middle for AI enabled data-driven research and education workflows. We created the NDP pilot thanks to support from NSF, with a use-inspired approach. First, identify the gaps. Next incubate, innovate, and educate use-inspired workflows and interfaces. Using tools in CI data and knowledge, composable services, and composable systems and platforms, integrate them through use of inspired workflows and interfaces for users to use them within an extensible platform. From that develop sustainable and scalable use. The benefit of this approach is that it is distributed in nature, it aligns with NRP which sustains the backbone, and it is a lightweight sustainment model.

The NDP allows communities to develop and deploy their own services application workflows, etc. This hopefully will foster scientific understanding, decision-making, policy formation and societal impact. Bringing research practices and governance processes to this method, we can manage benefits and risk at the same time.

Complicated architectures and Reference architecture links data and cyberinfrastructure with domain specific platforms to enable value add services and open educational capabilities. On the NDP website landing page, there are some capabilities listed there that can be used.

NDP does not store data, it federates it from other sources. From a lifecycle point of view, what does it mean to bring data into NDP, what happens to the data? We need to understand this further.

Are you able to subset, index, visualize and work with the data? NRP is trying to change the user experience of finding and working with data.

NDP is creating reusable capabilities to amplify the value of existing data repositories for the benefit of science, education and society. It enables AI-integrated science workflows to foster discovery, decision-making, policy formation and societal impact. The National Discovery Cloud for Climate is funding Case Studies for Generalizable Workflows. These studies are generalizable workflows for fire, earthquake, and food systems.

The idea is not to be everything for everyone. We can work with educators and have them use NDP to build data challenges for their classrooms. Once they build the challenges, they could be developed as toolkits for use by others.

There is enormous potential for NDP Generative AI and LLM. A lot of research needs to go into it. There are shortcomings in domain expertise. There is a need for domain-specific LLMs with human-curated data and controlled knowledge if we are going to use it for research and call it trustworthy. LLM deployment is expensive so there is an equity issue. Integrating AI and LLMs into workflows is difficult and there is a huge community of social scientists who are not equipped to do it. Operating an LLM necessitates an infrastructure with multiple GPUs and substantial memory capacity. NDP LLM as a Service can be aligned with NAIRR objectives.

This presentation may be viewed at: 2024 5NRP - Ilkay Altintas - NDP.

## 3.9 Triton AI Platform & Triton GPT – Vince Kellen, UC San Diego

This talk shared what UC San Diego is doing with generative AI both as an enterprise and as a teaching and learning aid in academics.

### 3.9.1 Strategy, use cases, architecture, collaborations.

The strategy for TritonGPT is to leverage existing infrastructure and capabilities and do everything on premise in concert with the San Diego Supercomputer Center (SDSC), taking advantage of their deep knowledge and expertise in AI. For the last seven years UCSD has been running their machine learning platform for education that started as a clone of the architecture at SDSC. They have added more sets of DGX cards to power the generative AI on premise.

The more that is focused on UC San Diego content, the less need for a big sprawling model. They need language expressiveness and skill but not all the content. They are also pursuing smaller models versus larger ones and using open source everywhere so that it is possible to democratize the AI and make it affordable. An interesting article came out in 2023 describing the ability to figure out what use cases would make sense as a “jagged frontier”[32] line of cost and performance. Inside the frontier line means it would be cost effective and outside, perhaps not. This research pointed

[32] Dell'Acqua, Fabrizio, Edward McFowland III, Ethan Mollick, Hila Lifshitz-Assaf, Katherine C. Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." Harvard Business School Working Paper, No. 24-013, September 2023. https://www.hbs.edu/ris/Publication%20Files/24-013_d9b45b68-9e74-42d6-a1c6-c72fb70c7282.pdf

out that generative AI works best when the knowledge worker exhibits critical thinking skills. If they make a copy of the AI, they do worse. There is no easy button.

Horizontal AI is like giving a $30 dollar screwdriver to everyone and they can use it for many different purposes. It can be purchased and configured to be applicable to many organizations and industries. Vertical AI must be configured to be specific to an organization to be useful. The AI makes knowledge available to employees within the organization and is not replicable to other organizations.

There are lots of ways to think about how generative AI can help an enterprise. One way is the generation and classification of words, another as fact retrieval, and then reasoning. An example of AI generation that works well is writing a job description. Take a policy and summarize it for the user. Classify a document. Predict something. An issue with fact retrieval is hallucinations. Still a bit of a puzzle. Reasoning is where a lot of people are struggling, especially for more elaborate reasoning problems.

Triton AI and TritonGPT were developed over a year ago with dozens of IT staff, expanded to include members of the community and onboarded to staff in the last calendar quarter of 2023. The plans are to expand to all staff and faculty in Spring 2024. It will not be released to students yet.

A two-million-dollar fund supported academic and administrative AI initiatives, implemented on premise and free to departments. Various departments piloted several Teaching and Learning GenAI tools. To get access to the system, staff and administrators go through basic training to level set people on expectations and the use of AI. It is all tied to university business processes. It is not AI for AI sake, but because something can be accomplished better, faster, or cheaper using it, such as the Job Description Helper.

*The AI assistant* can answer questions that have been ingested into TritonGPT such as all the university website content, all the policies and procedures, job descriptions, etc. A significant amount of time was developed to data scrubbing, for example, uncovering documents that are old and archaic.

*The Job Description Helper*, based on what a user types in, will give a list of possible job titles. After the user selects one and types in a couple of sentences, the helper delivers a job description based on that content and the 1300 templates that are available.

*The Fund Manager Coach* for grant accountants assists people in understanding the policies and procedures for grant accounting. It focuses on novices, not on experts, as a democratizing technology. The intent is to raise up the novices and average people. It can provide information on how much grant funding is left without having to create a report.

Other assistants are AI Study Aide, Accessibility Enhancer, and Assessment Builder.

A diagram of the Triton AI Platform architecture was shown. The target architecture is multiple LLMs, all of them open source and on premises. It could be run in the cloud as it is a containerized architecture. They are talking to a UCSD faculty startup called Protopia[33] that has a way of encrypting a chat window nonreversibly. It may be a way to add extra layers of security.

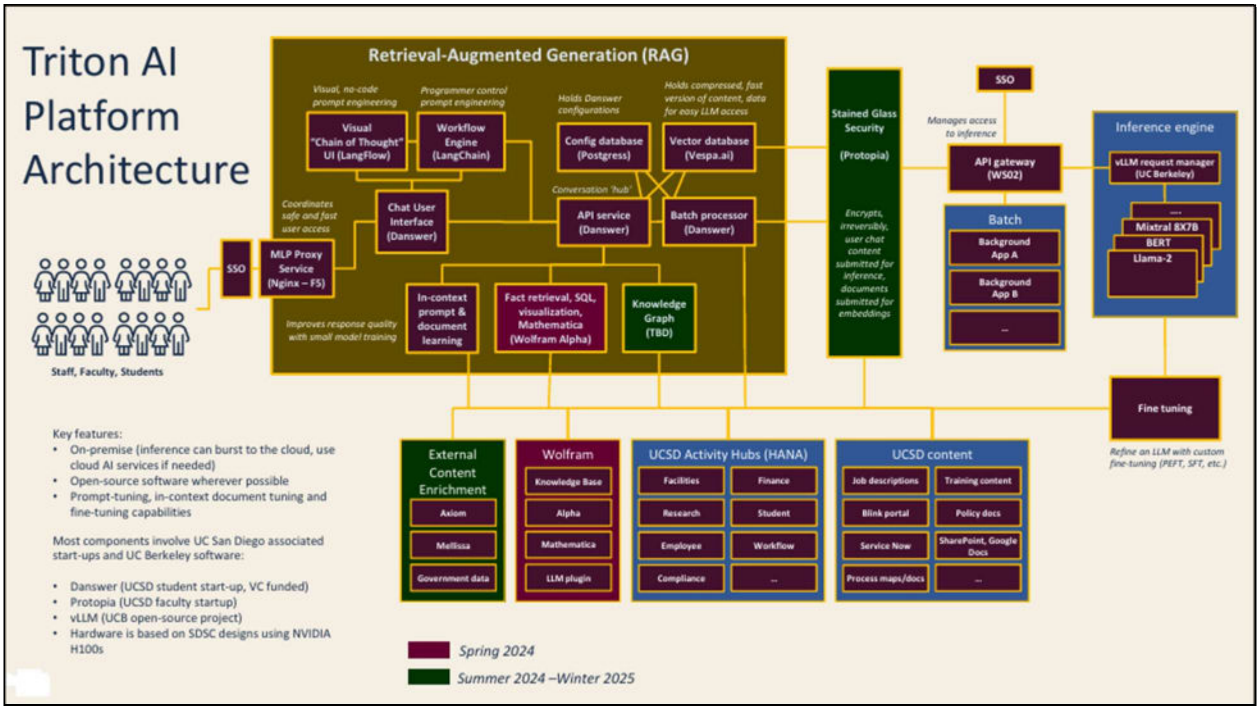


*Figure 7: Triton AI Platform Architecture*

UCSD is using LangChain[34] and LangFlow,[35] a GUI that is good for prototyping. They will hook it up to their data warehouse which is now being shared with UC Merced, University of Washington, and Alabama State University for student analytics. It can be thought of as a unified data warehouse in a very high-speed, in-memory analytical tool called SAP HANA, managing eight billion rows of data, and growing. It is all perfectly curated and joined up, so the user does not have to worry about joining data across multiple sources. This translates to a simpler natural language and the generative AI connection.

Since their students are using Wolfram[36] products such as Wolfram Mathematica and WolframAlpha, they would like to use them. This also allows student workers to help build the system and train other students reinforcing skillsets. It also means that biology students would not have to become programmers to do advanced analytics, they could just simply talk, and Mathematica could produce the right code for them and link to their libraries.

Additional platform development is needed into how this architecture can trigger other systems in a multi-agent framework such as the finance system or payroll system. They are looking at how to automate continuous quality assurance testing and keep tabs on the quality of what the system is

---

[33] The Protopia website is at: https://protopia.ai/ ,
[34] The LangChain website is at: https://www.langchain.com/ ,
[35] Langflow is a dynamic graph where each node is an executable unit. See https://github.com/langflow-ai/langflow .
[36] The Wolfram website is at: https://www.wolfram.com/,

producing. They are looking into the ability to leverage prompt engineering across multiple LLMs and are already supporting instructional faculty use cases like cloud bursting, and anything that saves power.

Currently their engineering cost target is 20-35% of commercial offerings. This includes the ongoing costs of hardware, power, facility rental and support costs, but does not include one-time project and start-up costs.

They have had a lot of interest from other educational institutions.

### 3.9.2 Questions and Discussion

Question: A lot of CSUs have the potential to use this. Already investigating extension to other campuses. Can scale nicely.

Question: Imagine surprise when asked by UCSC Vice Chancellor how AI can be applied. Can UCSD pile on?

Response: Yes. Tons of cases on the research workflow side. One example is all research offices that handle contracts and grants must look at the proposal and determine what compliance regimes apply. That is perfect for a rubric and generative AI with a simple API call from that tool out to a generative AI with a classification back and then another review to test the quality of that followed by a human review.

Question: Are you exposing any of your backend services?

Response: Absolutely. Right now, UCSD is carefully air gapping it.

Question: Is your deployment IS3 compliant and P3 P4?

Response: That is a complicated answer. Yes and no.

Question: How do you envision this to work with other universities? Do they bring the data here?

Response: The cluster of Config database, Vector database, Batch processor API service, and perhaps Knowledge Graph become a separate instance per location or institution. See Figure 4 for more detail.

This presentation may be viewed at: 2024 5NRP - Vince Kellen

## 3.10 NDP LLM-as-a-Service on NRP Tutorial - Ilkay Altintas, UC San Diego

This session aimed at having workshop participants log in to NDP, get a Jupiter Notebook and get a sense of how the service works. This live tutorial used a community LLM called ClimateGPT that has been trained on over ten billion webpages and open access academic journals on climate change. It is currently available on HuggingFace[37] and built on the Llama2 architecture.

[37] https://huggingface.co/

How do we add NAIRR context to ClimateGPT? The basic steps are shown below in a copy of the presentation slide. Step by step directions can be seen in the slides and the YouTube video.

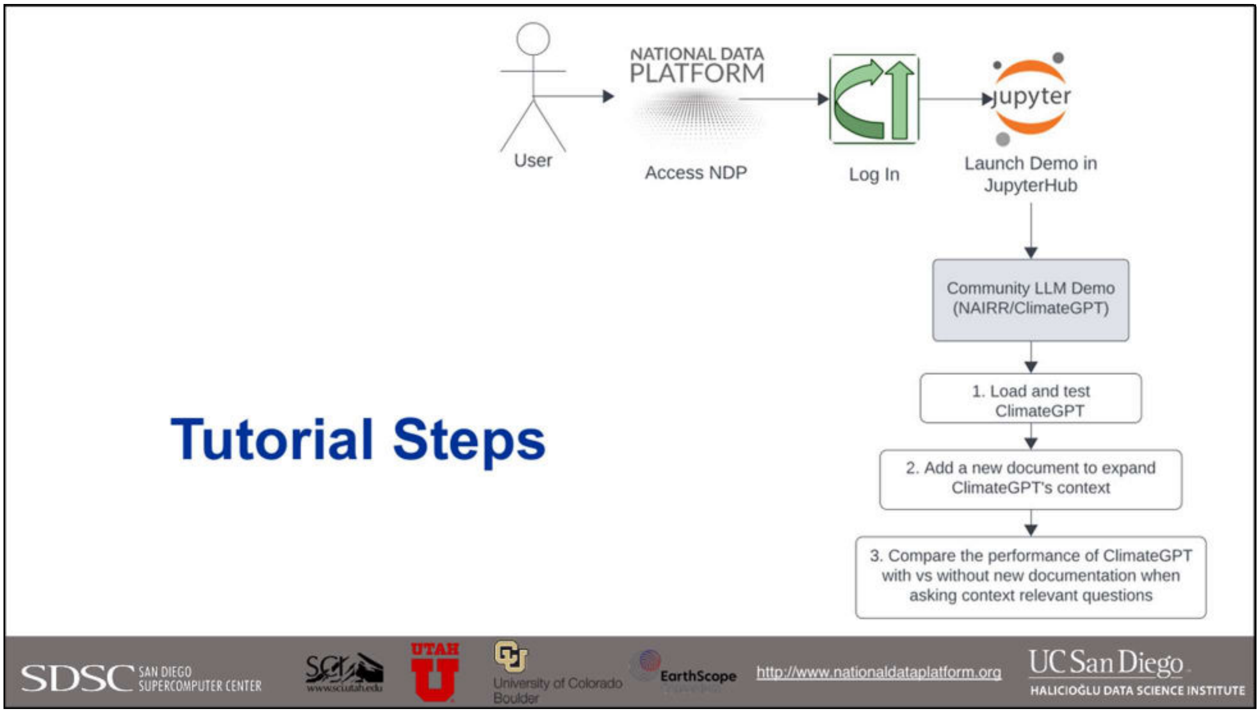


*Figure 8: NDP LLM-as-a-Service Tutorial*

The first step was to go to NationalDataPlatform.org and Log in. Unfortunately, the Wi-Fi at UCSD was not working well during the tutorial and it led to uneven results for participants.

Once logged in, starting with the JupyterHub, some participants were able to select the "LLM Service Client" and click "Start." The notebook had two parts. The first part helped set up the runtime environment for the notebook, including a conversational context function for asking questions. The second part will give a node in the cell that adds the NAIRR task force document to it. Examples were shown of pre-context and post-context.

To host an LLM within your own server, you would need at least 1 GPU instance. You could develop locally and then push to NRP and NDP later.

If the community wishes to participate, the team would love to work with others.

### 3.10.1 Questions and Discussion

Question: I am working with a few of our researchers on campus and we are trying to build a data commons for multiscale earth observations: everything from local LiDAR drone images to satellite images and longitudinal data. Are there converged datasets like this out there and is this something that would fit and dovetail in with the national data platform?

Response: NDP catalogues data in a storage facility.

Response: Yes, is the short answer. We want to be in a position to aggregate data like yours, like other institutions onto OSDF and then catalog them into NDP. Ilkay's work is about federating and being able to find the data whereas OSDF is a layer underneath it that wants to store the data. There are basically two options, we can bring you in as an origin of the data and you eternally will serve it, or if some of the data comes from elsewhere with an origin from a source like NASA or NCAR, we would rather keep the original copy from the data source.  However, if the original is in AWS or Google, we might opt for copying the data into OSDF due to the cost of data egress. Currently NRP has sufficient capacity to host this data.

Response: It is not just data that sits in an archive, but also includes streams and knowledge graphs that could be catalogued.

Response: Working with SAGE to make data available.  Another thing that Curt Dodds pointed out is that there are some things that have real value to the community that are not the primary data. We might wish to investigate organized pre-processing and only save the output of the pre-processing which then needs to be curated.

Question: What consideration are you giving generally to the idea of prompt engineering for quality answers?

Response: One of the things we are analyzing is the workflows for prompt engineering that can be built into the platform.

Question: Do you see citizen science as something that falls within the scope of democratization?

Response: For the concept of NDP that would be a yes. We look forward to integrating with the community in a more formal way with a formal engagement.  In the future, working with challenge providers and making sure the platforms and services that support the challenge are linked through NDP.

This tutorial may be viewed at: [2024 5NRP - Ilkay Altintas - Tutorial]

# 4 5NRP Track – Thursday March 21, 2024

## 4.1 FABRIC Keynote: Designing a Global Measurement Infrastructure to Improve Internet Security and How to get and use real data sets about Internet infrastructure for our ML/AI models – KC Claffy, UC San Diego

This talk was on how to get and use real data sets about Internet infrastructure for Machine Learning and Artificial Intelligence models. One of the goals of the talk was to communicate the state of what is currently available to study the global internet to do science on the global infrastructure: meaning the transport system, the plumbing, including the routing system, the naming system, and the certificate transparency system.

Why should data on the internet infrastructure play a key role in AI? The core of these models is the internet. Everything in these models works because they can get data freely off the Internet. GPUs may be the gold of AI, but the net is the oxygen. Ask the AI itself:

*"…if GPUs are the "gold" providing the raw material wealth or computational power that drives AI, then Oxygen (O) symbolizes the fundamental, life-sustaining environment – the Internet – that AI technologies require to function, connect, and flourish*." GPT4 (on its second try)

The Internet is the foundation to everything we are currently doing, yet we still have little understanding of what keeps it stable. When it is not stable, what are the vulnerabilities? What can we do that is relevant to AI. We can create data sets, and label them, but there is difficulty in doing so. The focus on data is coming, but data governance is a problem and is more than just a technical problem.

There is a national priority to train STEM students using infrastructures, including AI. However, it is hard to get good data for STEM students because much of the interesting data is owned by companies. One of the things that we can do is give students macroscopic global data sets about critical infrastructure that are freely available yet hard to understand.

Many Internet infrastructure problems are ripe for transformative AI solutions. So, what are the gaps? We need more comprehensive data sets and better labeling. We need to reduce the barriers on how to find, interpret and validate the data. And finally, we need to provide access to open LLMs that researchers can train at a reasonable cost.

Because the Internet evolved from a research project, every core transport service of the internet addressing the IP layer has serious, well-known vulnerabilities. After many decades of trying to retrofit some security onto these layers to mitigate these vulnerabilities, data is going to be a big part of the way forward. There is recognition that we need to find solutions.

Collecting data is expensive, complex, and subject to privacy issues. Every layer of the Internet from the IP address you put into the packet to the mapping from the host name to the IP address can be subverted. There are critical security flaws in Internet Protocol (IP) addressing, Border Gateway Protocol (BGP) routing, Domain Name System (DNS) domain naming, and Certificate Authority (CA). Nothing in these protocols prevent people from announcing addresses or domains they do not control.

Securing the Internet should be a national priority. The MSRI: Design: GMI3S project (OAC-2131987[38]) at UCSD is a design project in its third year to secure the internet, gather critical data, and design a global measurement platform[39]. The project focused on four types of data: active measurements, routing data, passive data capture and DNS.

CAIDA has spent a lot of time on active measurement since this does not require cooperation. The Archipelago platform deployed nodes all around the world. As part of the MSRI award, they are designing more software programmability into this platform to support complex internet measurement experiments that you cannot do anywhere else. There are other platforms out there such as RIPE ATLAS, but they cannot do complex measurements. There is also the Cloud if you want to pay for egress, but it will only give a Cloud vantage point.

[38] https://www.nsf.gov/awardsearch/showAward?AWD_ID=2131987
[39] https://gmi3s.caida.org/reports/Vulnerabilities-harms-dataneeds_V2.4.pdf

They want to operate in the middle sweet spot where researchers can do complex things on the nodes without having to program the nodes. To do this, they are creating a Domain Specific Language (DSL) to provide programmability, which supports measurements such as Ping, Traceroute, DNS Queries, HTTP, UDP Probes, Packet Capture, and Alias Resolution.

There are some infrastructure operational challenges to keeping systems like this going. Incentivizing people to host infrastructure, staffing changes, idiosyncrasies in automation all contribute to the operational risk of using crowdsourced infrastructure.

The project supports other tools to support access to vantage points not available through the platform. FANTAIL is a comprehensive query system designed to search raw Internet traceroute data. Periscope supports navigation of the data collected and the catalog of datasets.

Unlike active measurements, routing data requires cooperation with ISPs. The future of routing security is for global public measurements that provide a layer of transparency and accountability. Ongoing public measurements will guide the evolution of routing security by evaluating the effectiveness of policies and the inter-organizational auditing and accountability.

Current approaches to gathering BGP data are operator driven such RouteViews[40] and RIPE RIS[41] where route collectors are deployed at different exchange points and networks peer with the collector. Like active measurement challenges, this approach requires convincing others to deploy these collectors in their infrastructure. Scaling is also an issue, as is monitoring and managing data quality.

Hijackers can engineer BGP hijacks around the measurement infrastructure. To mitigate this, CAIDA has been working with researchers in Europe on designing an approach to scaling BGP data collection. GILL[42] BGP data platform is an architecture to manage a collection of routes from an order of magnitude more routers than existing platforms. Since a lot of data is redundant, storing just non-redundant data requires a framework for defining and detecting redundancy.

Other CAIDA tools to discover and analyze routing data are BGP2GO,[43] a graphical interface to explore and compile RV data (MRT files) and AS Rank[44], a tool for inferred routing relationships between Autonomous Systems (ASes) and orgs. LLMs could really help with the ranking system.

Traffic could be its own talk. It has the most privacy sensitivities. UCSD Network Telescope provides access to DARKNET traffic data in several ways so that researchers can gather insights. Traffic data is expensive to maintain. The total size of compressed historical raw packet capture data is about six petabytes stored at NERSC and this is only available due to NERSC providing storage at no cost. However, the data is very useful for security research.

A new approach to shortening the time to insight turned into a CICI project called Starnova[45] that expands the capability of identifying targeted attacks and hopes to make the data coming out of the

[40] https://routeviews.org
[41] https://ris.ripe.net
[42] https://bgproutes.quest/
[43] https://nids/caida.org:44444/
[44] https://asrank.caida.org/
[45] https://www.caida.org/funding/cici-starnova

darknets more actionable to the people who operate networks. The project is just getting started, but there are opportunities for others to get involved.

Distributed Denial of Service (DDoS) issues still loom. Data sharing for the metadata about DDoS is a big problem. It is very hard to get a common view of DDoS. Two-way traffic data is the holy grail to understanding what is really happening on the Internet. It took five years to upgrade the 10-gigabit packet capture connection to 100-gigabit, but now networks are upgrading to 400-gigabit. You cannot buy the hardware off-the-shelf.

They have proposed some approaches to overcoming limits on differential privacy. The real issue with differential privacy is how it can protect corporate proprietary information while still allowing reasonable queries of the data for research.

The Domain Name system has so many vulnerabilities. Many more things can go wrong in the DNS and do that any of the other infrastructures previously mentioned. It is resistant to oversight in governance. There are two big data sets studying macroscopic properties of the global DNS and while incomplete they are the best we have. One is called DZDB[46] and the other OpenIntel[47]. Researchers have used these data sets to find macroscopic security problems with DNS dependencies and misconfigurations that open domains up to hijack.

These daily snapshots miss short-lived domains which are probably the most interesting to security researchers. AI and ML could do a lot of work helping to detect the vast number of domains that are malicious. Twenty years ago, Verisign implemented something called rapid zone update, where every update to the zone would be visible in real-time. This would be useful for security researchers provided a policy agreement to establish a zone of trust with USG-controlled top-level domains (TLDs).

This talk concluded with some ways in which AI/ML could be used and is being used to enhance research and education using the CAIDA Resource Catalog.

---

[46] https://dzdb.caida.org/

[47] https://www.openintel.nl/

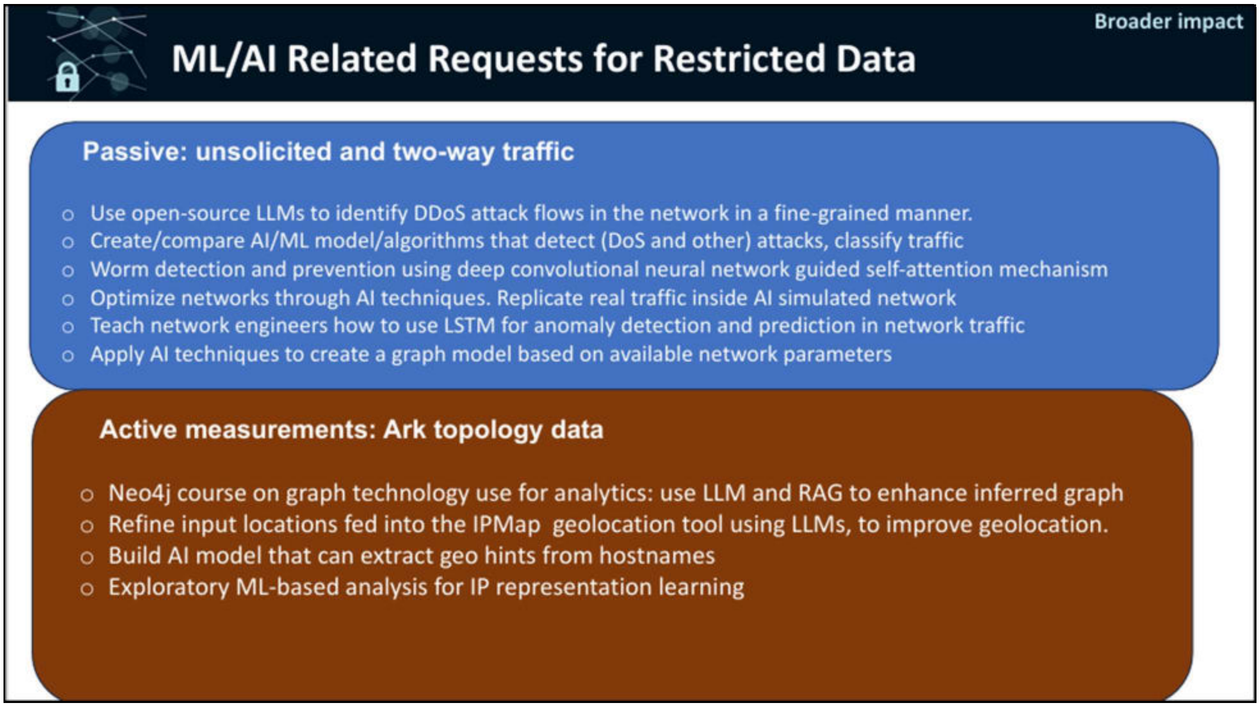


*Figure 9: CAIDA ML/AI Related Requests for Restricted Data*

### 4.1.1 Questions and Discussion

Question: For your measurement points, what are your requirements for a device or a location to be a good measurement point?

Response: It is good if your testbed is connected to the Internet, especially where CAIDA does not already have connectivity. Unlike RIPE Atlas, we cannot give them out to anyone at a conference and get three hundred new nodes. There is a cost for every deployment. CAIDA does not want one hundred nodes in one zip code unless it is 100 different ISPs. CAIDA wants topological diversity and geographical diversity, but it would be fine to have every node on all the testbeds. CAIDA likes to create teams of ISPs to see how much internet topology can be seen. So, you could have a FABRIC team, NRP team, Cisco team, or Comcast team.

In terms of hardware and power, the information is on the website, but it is a very low rate, no more than 200 packets per second.

Question: Do you want to use Large Action Models (LAM), one step beyond LLMs to interpret the data. Large Action Models do objects and relationships and that could give you information about causality.

Response: Absolutely. Researchers are working on this now.

Question: Could you say a few words about what you are doing with the list of industry collaborators on one of the last slides?

Response: Some of the industry partners have deployed infrastructure for us and some have used our telescope data and asked for help using the telescope data. DomainTools provides access to their data sets and uses the UCSD Network Telescope data. The slide is not a complete list. We have had good luck getting commercial data from data companies.

Kentik has been a partner with CAIDA all along and would love to see a state-of-the-internet report. Privacy of data is policy sensitive.

Question: What can FABRIC do to help?

Response: Deploy measurement vantage points. CAIDA is prototyping their deployment for R&E exchanges or nonprofit research friendly exchange points. If FABRIC is on any of these exchanges, it could peer with the node.

This presentation may be viewed at: [2024 5NRP - kc claffy]

## 4.2 FABRIC Keynote: In Search of a Networking Unicorn: Realizing Closed-Loop ML Pipeline for Networking - Arpit Gupta, UCSB

In the post pandemic era, there is more reliance on the Internet. It is important to make sure that everyone has access to a secure and performant internet. We have made significant progress in the last decade using software defined networking (SDN) to enable us to get better insights into what is happening in the network. This allows us to react with precise controls such as dropping DoS traffic. SDN can sense, infer, and react using network automation with limited resources while keeping our networks secure and performant with a limited set of resources.

However, the SDN network automation is still human-centric, relying on pre-existing heuristics and rules. This is ineffective against dynamic events such as coordinated multi-vector threats that are coming into the network. As was articulated in a report by the National Security Commission on AI, AI is a quintessential dual-use technology. It can empower us to make networks more performant but can also be used by malicious actors. AI-enabled tools will be the first resort in a new era of conflict. Countering AI-capable adversaries with human intelligence alone is recipe for disaster. There is a compelling case to be made for complimenting human intelligence with AI.

If we replace SDN with self-driving networks that have a bunch of different machine learning agents or models, capable of performing with minimal human interventions, we can do a better job of identifying complex subtle traffic patterns and synthesizing the actions that need to be taken. The most important requirement for self-driving networks is machine learning artifacts. We need production-ready ML models to handle various dynamic, coordinated, multi-vector threats. We need a lot of them.

We have invested over a decade building machine learning models for networks. There are at least 1,000+ research publications that talk about this. The government and venture capital have invested billions of dollars in this space. After all this investment, our expectation would be that it should be easy to develop ML models for any given problem and target environment, and that it should be ready for high-stake decision-making.

The reality is that we do not have this. The availability of public datasets dictates the choice of learning problem and environment. They do well in lab settings, but in the real world they are not that successful. Getting even a single ML model deployed in production is difficult.

There are fundamental roadblocks that are stopping us from achieving our desired goal. Using the standard ML pipeline, the responsibility to find the right data is placed on the users. For a given learning problem and target environment, users need to figure out what is the right data to use.

Another roadblock is data labeling. Labeling pre-existing data is hard. To get around this, people end up doing endogenous data collection where they collect and generate their own traffic. This is an additional step. Not only do users need to know what the right data is, but how they are going to collect the data.

Data Collection has challenges of its own. When it comes to data collection for machine learning in the context of networking, the efforts are very fragmented. There is little extensibility from one effort to another. Using public datasets dictates choice of problems and models.

Under specification issues are troubling and affect the generalizability of machine learning models. When training the machine learning model, you are unable to provide all the information that is required to produce the one best model. This leads to models being vulnerable to shortcut learning, or an OOD issue, spurious correlations, or lucky guesses and it will not be able to correctly generalize.

An example of this problem is trying to solve HTTP brute force attack detection. Do we know what the right data is, do we know how the model is making decisions?

A suggested closed-loop ML pipeline approach is discussed in detail in this talk. The building blocks for closed-loop ML pipeline are netUnicorn,[48] a data collection platform, Trustee[49], a post-hoc global model explainability framework, and PINOT, a programmable infrastructure[50]. With this, we have just scratched the surface and it is still a slow process requiring domain expertise.

The next frontier is a Network Foundation Model which is basically the same thing as what we know as a LLM for natural language processing.  The foundation model is perfect for networking. We have a lot of data, most of it unlabeled. If we can train a good foundation model, then we can fine-tune that for different tasks and we can expedite the process. These foundation models might be able to extract more value out of the data, leveraging the abundant telemetry data and catalyzing innovation and the development of production ready machine learning artifacts.

---

[48] Beltiukov, Roman et al. "*In Search of netUnicorn: A Data-Collection Platform to Develop Generalizable ML Models for Network Security Problems*." Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (2023), arXiv:2306.08853. And the project website: https://netunicorn.cs.ucsb.edu

[49] Jacobs, Arthur & Beltiukov, Roman & Willinger, Walter & Ferreira, Ronaldo & Gupta, Arpit & Granville, Lisandro. "AI/ML for Network Security: The Emperor has no Clothes," CCS '22: Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security. https://doi.org/10.1145/3548606.3560609 and the project website: https://trusteeml.github.io

[50] PINOT: Programmable Infrastructure for Networking, ANRW '23: Proceedings of the Applied Networking Research Workshop July 2023, Pages 51–53, https://doi.org/10.1145/3606464.3606485 and the project website: https://pinot.cs.ucsb.edu

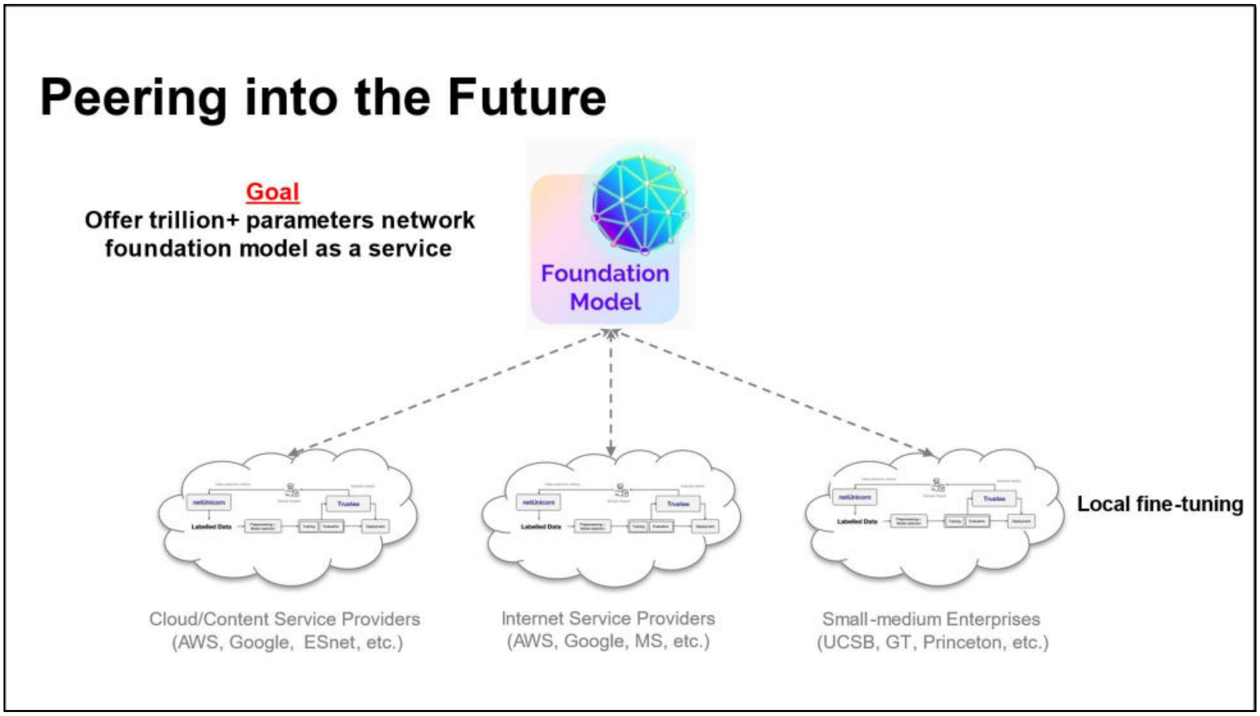


*Figure 10: Peering into the Future – Trillion-plus parameter foundation model as a service*

Peering into the future, if we get things right, we can think about offering this trillion-plus parameter network foundation model as a service to enterprises. We need to solve some fundamental research questions.

### 4.2.1 Questions and Discussion

Question: Trustee is based on decision trees and labeled data. Do you have near-term plans to start working on explainability for unsupervised data?

Response: Trustee does dive into how it can think about providing explanations for unsupervised models. It is non-trivial. So, yes, we are working on some issues where we can use Trustee to explain some decisions.

Question: How can FABRIC help?

Response: FABRIC can be extremely useful to the whole process of training this type of machine learning artifacts and providing data consolidating as well as building it and offering it as a service. This whole thing requires a lot of scaling up and FABRIC may be able to provide the support that we need looking forward.

This presentation may be viewed at: [2024 5NRP - Arpit Gupta .]

## 4.3 Keynote Demo – Leveraging FABRIC's Hardware Resources for Programmable Networking - Nik Sultana, Illinois Tech

This talk had two goals: one to give examples of using FABRIC's programmable networking hardware and the other to provide a starter kit.

Programmable networking is where the network plays a more active role in moving data across the network and making it possible to do some processing along the way to enable better performance. This talk presented concrete examples in two application domains: cybersecurity and research network infrastructure.

A demonstration running in a FABRIC Jupyter notebook illustrated that computation could take place in the network and not the notebook, in this case on a XILINX Alveo U280. There are many of these on FABRIC as a shared resource with a large geographical span. These heterogeneous resources can be used for networking and non-networking research applications. They are conveniently accessible through *authn* and *authz*, notebooks and SSH.

Peeling back the onion, so to speak, to get closer to where the action took place, the Jupyter notebook made a request to a node that was deployed on FABRIC and this node then interpreted the request and generated a packet. The request never enters the peer host. It crosses the Alveo card and is serviced in the data plane on the card, doing the calculation and sending the response back over the link to the host that originated the request. The reply reuses the frame, switching the source and destination MAC addresses.

This limited computation in the data plane, as distinguished from the control plane or the edge, gives guaranteed performance in terms of low jitter and predictable throughput and latency but has less expressiveness. Some examples of programmable networking in FABRIC are University of South Carolina's work with BMv2 and DPDK-based prototype software switches, Northeastern and University of Massachusetts' work with FPGAs and ESnet's use of Alveo-based smartNIC framework.

The smartNIC framework supports the P4 language, is extensive and open source, does not require hardware or engineering skills, has streamlined modular workflow based on using containers, and builds on an AMD/Xilinx toolchain. P4 is a good entry point. It is extensible. It has a streamlined modular workflow.

These can be and are being used for teaching, network monitoring, and cybersecurity. Another use case is network profiling where the goal is to understand the traffic composition and characteristics of FABRIC experiments then sharing this data and mechanism with the community of testbed users. This provides data and insight for research projects on the network. Profile is itself a FABRIC experiment.

Testbeds have unique characteristics that differentiate them from standard network profiles. Profiler over time collected a lot of data across the international FABRIC footprint. One of the challenges was to digest and analyze that data. By looking at the data over time, you can get a picture of resourcing and header diversity.

The role of the data accelerator card (Alveo in this case) is to offload sampling and do filtering, truncation and blinding to the smartNIC from the host. For this FABRIC experiment, they were sampling traffic over a 100-gigabit link, with performance bottlenecked at 8.5 gigabits per second. On the host side they were using DPDK (data plane development kit) because they did not want to go through the kernel network stack.

Another application is Remote Attestation which tries to get evidence about remote equipment and resources in which you are a stakeholder, but do not directly control. This is motivated by the early detection of misconfigurations and advanced persistent threats (APTs), mitigating attack surface from programmability.

One approach is to augment the fundamental aspect of reachability. These devices bleed a bit of the configuration that allows you as a peer or stakeholder to make inferences about what this configuration is and then reconcile that with what you through the configuration would be. You can get a third-party software switch and attach it to a sandbox. This allows you to get an image of the P4 program and determine if the program has been changed or is exhibiting different behavior. Another approach is to sandwich third party equipment for an external element of trust.

To deliver the starter kit, development will be done outside FABRIC, experiments will be done inside FABRIC. The plan is to use the smartNIC framework, allocating the necessary resources on FABRIC, including an Alveo, then to deploy the program, running and testing the program and finally to diagnosis.

The talk then described the process starting with writing P4 programs, then running the Alveo experiments, including how to create a slice, configuring the host, configurating the Alveo card and program's tables, then clearing the tables and swapping the rules. Finally, once the program is up it is already running. For testing, Scapy[51] is being used. For diagnosis there are different types of simulation that you can run to get quick feedback and test whether the logic is working the way it was intended.

The tutorial starter kit can be found at http://packetfilters.cs.iit.edu/esnet-smartnic-tutorial/ and additional information can be found on the FABRIC Forums: https://learn.fabric-testbed.net/forums/forum/all-about-fpgas/.

### 4.3.1 Questions and Discussion

Comment: There will be a full-day INDIS Workshop at SC24[52]. It gives students and faculty a great opportunity to publish papers and get talks and keynotes for your normal scientific Curriculum Vitae.

Response: FABRIC will be promoting this workshop for FABRIC users. Many previous INDIS papers were FABRIC-related.

Question: What is next? How can this capability work with scientists and users that would benefit from doing this in network computation using the FPGAs? How do we take the next step?

Response: I see two interrelated activities, one is capacity building and more on the education side, putting students into contact with resources and infrastructure they typically would not encounter in a typical university course, feeding that research, experience, knowledge, and energy to solve novel problems. I am primarily looking at two areas, cybersecurity and continuing to refine prototypes that allow stakeholders of a network to verify its integrity and get more transparency into the resource. The other direction is in science. How research networks can use programmable

[51] Website for Scapy is: https://scapy.net/

[52] The SC24 INDIS workshop: https://scinet.supercomputing.org/community/indis/

networking to make life easier in terms of maintenance, cost, and scale to meet the increasing demands of their instruments.

This presentation may be viewed at: 2024 5NRP - Nik Sultana .

## 4.4 Lunch Session: National Research Platform International Extensions, Including the Global Research Platform

The lunch session was not recorded. There are slide decks available on the NRP site for most of these quick presentations.

### 4.4.1 Joe Mambretti, Introduction to the session

Science is global and open information sharing is a cornerstone of the scientific process. Science Instruments produce and export exabytes of data necessitating an extensive global infrastructure. The global collaborative research communities openly communicate and share concepts, experiments, methods, techniques, data, technologies, and results. As part of this, the Global Research Platform (GRP) is an international collaborative partnership that is creating a distributed environment for international data intensive science.

The GRP is sharing information on tying things together. AutoGOLE/SENSE is one resource. Another is NA-REX. 400G is available for large-scale demonstrations. GRP meetings are once a year in conjunction with the IEEE eScience conference. The next one will be at Osaka University in Japan in September 2024. Global scale science has been highlighted at prior GRP workshops including the Square Kilometer Array (SKA), the Large Synoptic Survey Telescope (LSST), the Korean Fusion program, Belle-II, Deep Underground Neutrino Experiment (DUNE), and the KAUST Genomics Cloud.

Some of the themes that have been emphasized in the GRP are orchestration among multiple domains, large-scale high-capacity data WAN transport, high fidelity data flow monitoring, visualization, analytics, and event correlation.

International Testbeds for data-intensive science are important. Many of these have supported Network Research Exhibitions (NREs) at conferences such as Supercomputing, Supercomputing Asia and the Data Mover Challenge, and OFC conference. The new Ciena CENI Testbed is also important as are Open Exchanges, the FABRIC testbed, and other resources.

Demonstrations of extending data center services over 400G WAN have provided initial results and they are close to doing 400G end-to-end. A spinoff from Northwestern's Center for Photonic Communications and Computing, NuCrypt[53] is working on the distribution of quantum entanglement through fiber with co-propagating classical data.

### 4.4.2 Kevin Thompson, International Research Connections Program

NSF Networking and Cybersecurity CI Investments are supporting a set of programs CC* (Campus Infrastructure), CICI (Cybersecurity Innovation for Cyberinfrastructure), IRNC (International R&E

[53] http://www.nucrypt.net/

Network Connections) and Special Projects such as TrustedCI. Resilient, robust, and performant networking is a fundamental layer and underpinning of the CI Continuum.

Science is International and as such NSF is supporting awards in three IRNC areas.

Area 1: Backbone and Exchange Point International Networking (IRNC: Core) is funding Indiana University's Transpac5, University of Hawaii's PIREN, CENIC and PNWGP's Pacific Wave, and FIU's AmLight.

Area 2: International Testbeds (IRNC: Testbed) is supporting FAB (FABRIC Across Borders) to Europe (UK, CERN, The Netherlands), Japan (University of Tokyo), and South America (CPTEC/INPE Brazil).

Area 3: Engagement for Training and Human and Network Capacity Building (IRNC:ENgage) is supporting NSRC's ENgage: Strengthening Global Cyberinfrastructure Ecosystems to Advance International Science Collaboration.

### 4.4.3 AutoGOLE/SENSE WG and Infrastructure: Infrastructure and Services for Network and Domain Science Workflow Innovation- Tom Lehman, GNA-G AutoGOLE/SENSE WG

Tom Lehman gave a short update to the GNA-G AutoGOLE/SENSE working group efforts to create a worldwide collaboration of open exchange points and R&E networks interconnected to deliver network services end-to-end in a fully automated way. NSI is not the only way to implement these services.

The group is focused on control plane monitoring which is Prometheus/Grafana-based, data plane verification and troubleshooting service, AutoGOLE-related software (Kubernetes, Docker etc.), experiment, research, and use case support. A key focus now is on integrating edge resources such as the National Research Platform site, domain science workflow prototypes and development systems, and others.

The presentation slides contain multiple diagrams showing topology, work in progress, integration with FABRIC, and the NRP integration with advanced services. Some of the Internet2 AL2S paths are being replaced by NA-REX[54] paths. Some of the paths are dedicated for experimental use and others are production.

### 4.4.4 Asia-Pacific Research Platform of APAN and related work - Jeonghoon Moon, Asia Pacific Research Platform (APRP)

#### *4.4.4.1 APRP*

APAN initiated the Asia-Pacific Research Platform (APRP) working group in 2018 with the goal of sharing experience with members and joining with the Global Research Platform (GRP). During the last APAN meeting, there were two sessions, each with five presentations – one session focusing on infrastructure and the other on applications.

[54] The press release for the NA-REX collaboration can be found at: https://internet2.edu/network/initiatives-partnerships/global-networks-and-partnerships/north-america-research-and-education-exchange/

##### *4.4.4.2 KRP*

The Korea Research Platform (KRP) is working to establish a high-reliability and high-speed transfer system without boundaries between participants and support the research platform as a computing resource. In Korea, there are twenty-five government-funded research institutes and KREONET has dedicated optical network connectivity to those institutions. KRP's final goal is to cover all twenty-five institutes.

Researchers are using KRP to analyze urban flooding and solar visual data, running crop models for smart agriculture, establish smart decision making for climate scenarios, and establish a high-performance research platform for medical research.

##### *4.4.4.3 A3 Foresight Program, a Korea, China, Japan collaboration*

If approved "*Data Sharing Infrastructure across Northeast-Asia Supercomputing Centers for Open Science*" will be a five-year, $1.2 million budget involving KISTI (Korea), NSCC-GZ (Singapore), RIKEN and Osaka (Japan).

"*This multi-party research project aims to spearhead the development of a dynamic environment conducive to the utilization and sharing of research data, thereby establishing an advanced data-sharing infrastructure among Korea, China, and Japan. The project's core object is enhancing open science by improving data infrastructures and metadata standards to facilitate seamless data sharing and collaborative research across these nations*." – *From slide 17 of this presentation*.

##### *4.4.4.4 Future plans and conclusion*

Future plans include expanding infrastructure to the Asian region using APAN and TEIN networks, activating the APRP to further collaborations, and expanding third party research areas in smart agriculture, medical, climate, AI, bioscience, cloud, and wireless.

### 4.4.5 Julio Ibarra, AMPATH/AmLight

Julio Ibarra gave an update on the AmLight-ExP and Protect project (NSF #OAC-2029283) and the Atlantic Wave- SDX. AmLight-ExP is an International R&E network built to enable collaboration among Latin America, Africa, the Caribbean, and the US. There is 600G between Latin America and US, and100G to Africa. The connection to San Paolo will be upgraded to 400G in 2024.

Eleven Open Exchange points are interconnected: three in Florida, one in Atlanta, Georgia, two in Brazil, one in Chile, one in Puerto Rico, one in Panama, one in South Africa, and soon one in Argentina.

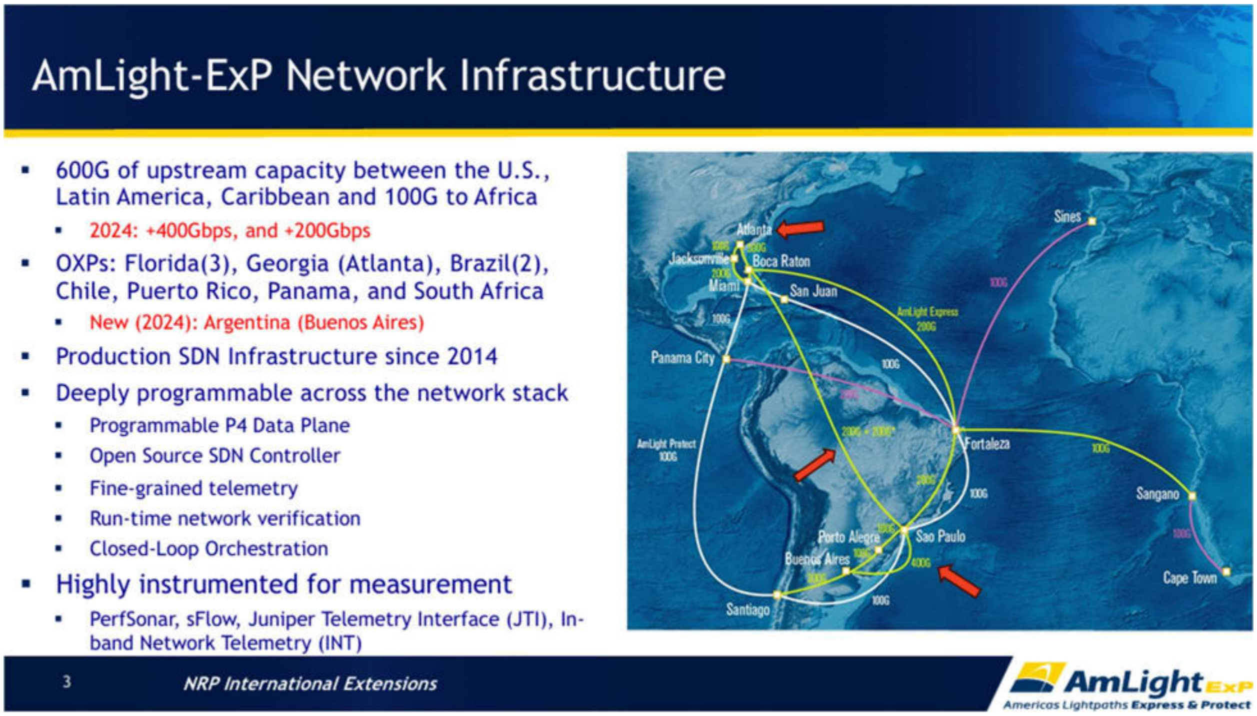


*Figure 11: AmLight-ExP Network Infrastructure*

AmLight has had production SDN since 2014, with deep programmability across the network. In 2023 they upgraded the network. Just this week they upgraded Santiago. Panama is the only site left to be upgraded. They use a home grown SDN controller that can collect fine-grain telemetry for runtime verification. They have been increasing automation of the network and their backbone is highly instrumented (INT, etc.).

Its closed loop orchestration and network verification and packet provenance reduces the need for operator intervention. It supports major facilities such as NOIRLab, ALMA, Vera Rubin, NRP, FABRIC, LHC, OSG (S Africa) and PATh (Santiago). Open Exchange Points provide the flexibility to place computation and storage closer to major facilities. Multiple diverse paths and bandwidth capacity translate to high availability.

A GEN4 NRP Node at AMPATH supports Nautilus and operates as an OpenNSA AutoGOLE resource. Two storage nodes were added in 2023 to help with the Ceph storage pool, helping universities in the Southeast.

There is a dedicated 100G path between Miami and Atlanta for FABRIC. They are enabling researchers to create and conduct experiments both on FABRIC and AmLight. They are working on creating a Jupyter notebook library for the AtlanticWave-SDX to integrate with FABRIC and providing FABRIC experimenters access to resources.

### 4.4.6 Gauravdeep Shami, CENI International Testbed

Ciena is building a testbed called Ciena Environment for Network Innovation (CENI) that is a persistent, international platform for Ciena and the R&E Community. It is an accelerator for applied network research and facilitator of state-of-the-art demonstrations.

The networks and internet exchanges involved are Utah Education and Telehealth Network (UETN), StarLight, ILight, CANAIRIE, Internet2, and Mid-Atlantic Crossroads (MAX). Ciena has a *tfNode* that is a traveling FABRIC node, capable of traveling to conferences and events. During SC23, in collaboration with UETN, iCAIR, and ESnet, the “Optimizing Data Transfers Through Real-Time AI” demonstration established real-time data transfer node (DTN) transmissions on the CENI testbed utilizing advanced analytics.

### 4.4.7 Global Network Advanced Group/ DIS WG: Towards a Next Generation Network-Integrated System - Harvey Newman, GNA-G

The GNA-G is an open volunteer group devoted to developing the blueprint to make using the Global R&E networks both simpler and more effective by pulling together a large and diverse group of institutions.

The principal aims of the GNA-G Data Intensive Sciences working group are to meet the needs and challenges faced by major data intensive science programs, but in a manner that is compatible with supporting the needs of individuals and small groups and to provide a forum for discussion and a framework for shared tools. A special focus of the working group is to address the growing demand for network-integrated workflows, cross institutional data management, automation, and federated infrastructures.

Using technology, infrastructures, and investments of its participants, the GNA-G brings together researchers, NRENs, Global Exchange Points, regionals, and other R&E providers to support the needs of its participants. Diagrams of the participants were shown including Pacific Wave, Rednesp[55] and RNP, the Global P4 Lab, CANARIE, ESnet, GEANT, and Internet2. Multiple Network Research Exhibitions (NREs) relied on these networks and exchange points during Supercomputing 2023.

NAIRR represents a major opportunity going forward.

You can get involved by joining the GNA-G mailing list, signing up for a working group, contributing your thoughts during one of GNA-G’s virtual conference calls, or reaching out to the GNA-G leadership team.

To join the mailing list: https://lists.gna-g.net/postorius/lists/all.lists.gna-g.net/

To join a working group: https://www.gna-g.net/working-groups/

To share your ideas: https://www.gna-g.net/meeting/

To send your ideas to the leadership team, send email to: leadershipteam@lists.gna-g.net

## 4.5 AI/ML Computing for Gravitational Waves - Phil Harris, MIT

This talk discusses AI/ML workflows for gravitational waves and how NRP is pushing this forward. The Computing Revolution happening now has led to the rise of heterogeneity enabled by deep learning. In the past few years, the success of deep learning has led to sophisticated algorithms

[55] More on Rednesp can be found at: https://fapesp.br/rednesp

that take advantage of parallel processing, sometimes by a factor of a thousand. Processors like GPUs and FPGAs give us a dramatic speedup in processing. This realization led to building an institute to develop real-time applications for large scale science experiments.

The creation of the fast machine learning for science community was followed by the creation of the A3D3 Institute to address the problem of real-time AI for science. The Institute is looking at developing machine learning and GPU integration for large throughput computing and in this case for physics experiments. They also want to develop machine learning with specialized FPGAs and Asics for low latency computing. Science benchmarks are competitive with other benchmarks in the rest of AI world.

A gravitational wave is produced from large catastrophic events in the universe such as when two objects merge resulting in radiation. The goal of LIGO is to identify and detect these gravitational waves. These waves can be measured by the distortions in space which are on the order $10^{-19}$ meter over a 4-kilometer length, many magnitudes smaller than the size of a nucleus. Using a large interferometer, scientists look for deviations in their space coordinates and these deviations give off oscillations in the intensity of light that can then be translated into a signal.

Each LIGO detector gives a single stream of data. To ensure that what is seen is not a glitch or noise, multiple detectors are used. There are four gravitational wave detectors on earth, two in the United States (Hanford, Washington and Livingston, Louisiana), one in Italy (Virgo) and one in Japan (Kaga). Identifying a gravitational wave takes an incredible amount of processing. The data then needs to be cleaned and the type of topology identified. Would doing this with AI speed things up and do it as fast as possible?

If there is an event, gravitational waves hit the earth first before x-rays and gamma rays, visible light, radio waves, and neutrinos. Identifying them quickly allows scientists to alert the rest of the world, other telescopes and other detectors and inform them where to look in the sky so they do not miss the event.

The LIGO workflow takes in the data and de-noises it as fast as possible. The data comes in about 10Khz. To calibrate the data, there are 100,000 auxiliary channels to remove the noises and get a clean stream. This is already a considerable amount of data at about 1PB per year. Once there is clean data, scientists look for ripples to identify the gravitational wave. Once identified, an alert is sent out. This workflow is not the fastest, but it works. However, they are convinced that they can do this faster on GPUs.

They are developing the kind of infrastructure end-to-end that can replace the current rule-based algorithms with AI-based algorithms. Almost all of the compute can be pushed onto the GPUs. The upgraded workflow, with almost 100% ML-based setup, includes a ML4GW toolkit to enable fast machine learning deployment for gravitational waves. The strategy is to use the end-to-end pipeline for development and deployment of these algorithms so that they can be deployed at each LIGO site and process data in real-time.

This is what industry calls MLOPS (machine learning operations) which starts with a scientist generating data, building an appropriate simulation, designing algorithms and training on various simulations and data sets. The machine learning requirements are training and inference. In training, the strategy is to leverage the data sets at hand and put everything on the GPU and train and keep the data on the GPU allowing various augmentations and simulations done of the fly. They have a full scheme to deploy this to scale allowing a researcher or scientist to run this on 10,000 years of data.

One insight garnered while developing this framework relates to how things are cached on the GPU. They developed a caching scheme where the time series is saved on the GPU and only new data is put onto the GPU thereby creating a kind of stateful caching on the GPU that reduces the IO to the GPU and speeds up the overall inference. This dramatically reduces network overhead and creates a more flexible way to do the whole processing of the data.

In keeping with using standard tools for deploying machine learning to scale, they relied on inference-as-a-service paradigm. Leveraging a GPU cluster such as NRP, these GPUs could be turned into a big server. If Kubernetes is available, dynamic load balancing and allocation becomes easy and efficient. To do this, A3D3 created a toolkit dedicated to inference-as-a-service named Hermes where everything is containerized.

To claim they can find anomalies, they need ten thousand years of pseudo data. Using an algorithm called GWAK that they developed and using the Nautilus Hypercluster, they ran one hundred years of data in two weeks, which is tenfold the speed of the conventional GPU.

To achieve these results some of the following algorithms were employed.

The DeepClean algorithm takes in detector and witness channels benchmarked for NRP and outputs a clean stream. This is a big improvement to previous algorithms and allows real time noise subtraction that goes beyond linear subtraction. The NRP was perfect for benchmarking this on the FPGAs.

A-Frame[56] was developed for rapid real time identification of binary coalescences, or black hole mergers. Nautilus is used heavily for this work.

GWAK is used for LIGO-based anomaly detection.

Parameter Estimation used the likelihood free inference scheme that has been developed recently. It uses normalizing flows to compute the probabilistic distribution of your gravitational wave and out of this you get sky localization.

The beauty of using ML algorithms is they can take seconds to process.

The current use case is the black hole merger detection work using Nautilus.

There is a white paper[57] written on scoping out the use of GPUs for physics experiments.

---

[56] Marx, Ethan; Benoit, William; Gunny, Alec; Omer, Rafia; Chatterjee, Deep; Venterea, Ricco C.; Wills, Lauren; Saleem, Muhammed; Moreno, Eric; Raikman, Ryan; Govorkova, Ekaterina; Rankin, Dylan; Coughlin, Michael W.; Harris, Philip; Katsavounidis, Erik. “A machine-learning pipeline for real-time detection of gravitational waves from compact binary coalescences.” https://arxiv.org/abs/2403.18661.

[57] Agarwal, Manan et al. “Applications of Deep Learning to physics workflows.” arXiv:2306.08106.

The critical computing needs for this work include connected infrastructure for machine learning development, efficiency, a software stack where all algorithms aim for a set of core software tools, containerization, and the awareness of diversity of problems.

### 4.5.1 Questions and Discussion

Question: Do these algorithms work better or worse in higher or lower mass ratios?

Response: This is an active discussion. When you go to the lower masses, you want to deal with longer time scales. The neural network architecture that we are dealing with seems to be more optimized for shorter periods like fast mergers. There are solutions but we are early in the exploration of this.

This presentation may be viewed at: 2024 5NRP - Phil Harris .

## 4.6 SmokeyNet for Wildfire Smoke Detection - Mai H. Nguyen, SDSC

The size and frequency of wildfires in the US have notably increased in the last few years. Since 1980 twenty major wildfire events exceeding a billion dollars in damages have occurred. Sixteen of those occurred in the last two decades. In California, seven out of the ten largest wildfires have occurred in the last six years. Wildfires can spread quickly; early detection is essential to minimize catastrophic damage.

Detecting smoke from wildfires sounds like a straightforward problem, but in fact is very challenging in real world scenarios. Smoke is transparent and amorphous. Smoke plumes can be small, faint, and dissipating. There are many sources for false positives such as clouds, haze, and fog.

An approach to wildfire smoke detection using a deep learning approach and deep learning model is called SmokeyNet. The architecture for SmokeyNet consists of three different neural networks: the convolutional neural network (CNN) that extracts features from each image tile; the LSTM that is the long short-term memory neural network that models the temporal aspect; and the Visual Transformer (ViT) that captures all the spatial correlations between the tiles within an image. So SmokeyNet is a spatial temporal model.

To develop and evaluate SmokeyNet, the data set called FigLib (fire ignition images) is used. FigLib consists of 3,155 sequences of wildland fire images from cameras mounted on top of weather stations mostly in Southern California. The FigLib data spans about five years of images taken at one-minute intervals. This data set has been curated so that each sequence has the ignition image forty minutes before and forty minutes after ignition. FigLib is part of HPWREN which is a high-performance wireless research network.

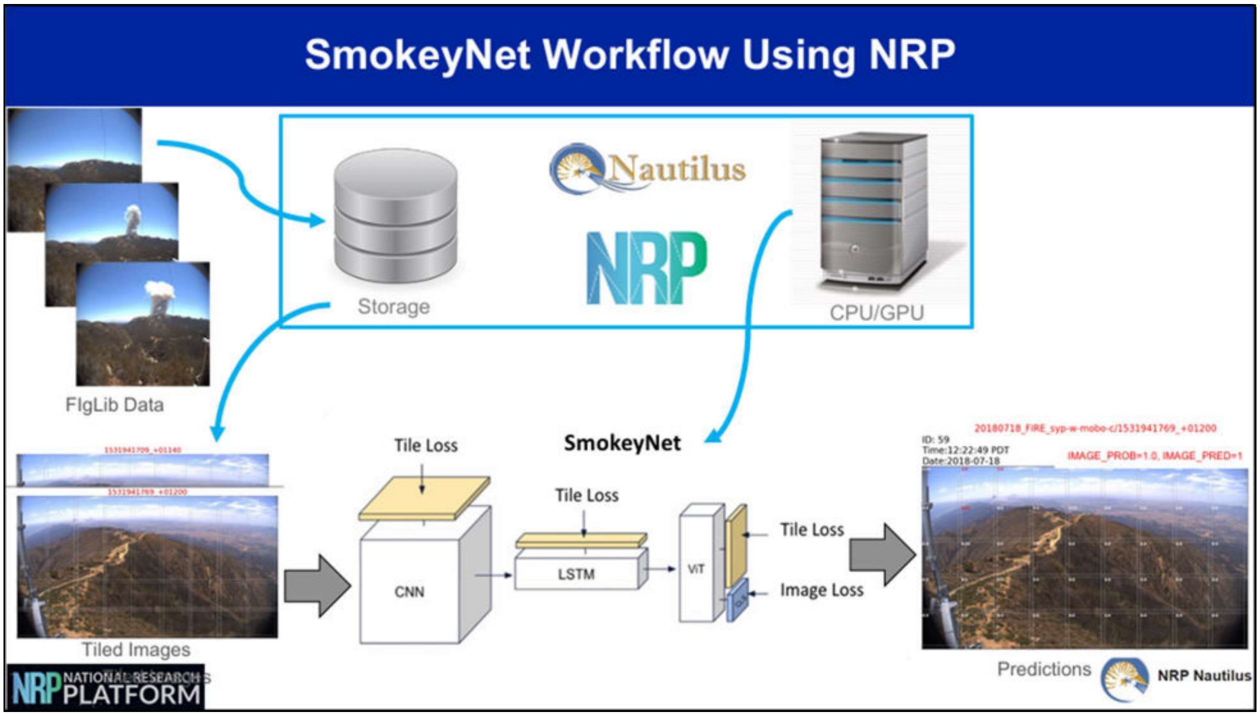


*Figure 12: SmokeyNet Workflow Using NRP.*

The workflow using NRP starts with storing the FigLib images in Ceph volume in Nautilus and then these images are processed for use by SmokeyNet. SmokeyNet takes these images as input and then outputs the tile and the image probabilities with the results being stored back into the Ceph volume. Many experiments were run to evaluate SmokeyNet against other deep learning approaches for object detection. Overall SmokeyNet outperformed the others and delivered on high detection performance and quick time to detect.

To help performance, they investigated other the use of multiple input data types from weather data and satellite-based detections in addition to the camera image. Weather data is captured from three different networks, HPWREN, San Diego Gas and Electric (SDG&E0, and Southern California Edison (SCE). For every image, they collect weather data from the three stations closest to the camera. The weather features that are used are air temperature, relative humidity, wind speed, wind gust, wind direction and dewpoint temperature.

Additionally, a satellite-based fire detection system developed by the University of Wisconsin Madison called WFABBA, which means Wildfire Automated Biomass Burning Algorithm, uses heuristics to determine thermal anomalies. It is a rule-based system not a machine learning system. The data comes from NOAA GOES-R satellites,

To incorporate these three different data sources, the camera images, the weather data and the satellite-based fire detections, an ensemble approach was used. The performance of the ensemble model was not much better than baseline. Given this result, they proceeded to look at another approach of integrating the weather data directly into SmokeyNet. The weather data was injected in two different places into the model, one to concatenate the CNN embedding and the other to concatenate the LSTM embedding.  Both input sources are integrated together and processed together by the model. This approach gave a little boost to detection performance.

For future work they would like to investigate unlabeled data, decrease false positives, test generality to other geographical areas and camera types, and deploy SmokeyNet as early notification system for effective real-time wildfire detection.

Other work that is part of the WiFIRE lab is BurnPro$^{3D}$ a platform that gives our public sector partners next generation fire science using data and AI to optimize prescribed burns at an unprecedented scale.

### 4.6.1 Questions and Discussion

Question: I would like to consider the time theme. How long does it take the fire people to get there? You also need a certain amount of time to get the resources there to put it out.

Response: I do not know. It depends on how remote the area is and how quickly they can get there and the resources that they have. There are fire scouts and there are also 911 calls.

Response (Larry): If we adopted a military response to a civilian threat, military bases with airlift capability would be able to provide an instantaneous response, getting fire retardant in the air and dropping it on the fire.

Question: Have you considered two different networks, first to identify the sky and remove the sky from the background and then try to identify the fire.

Response: We do some cropping to get rid of some of that stuff up in the sky, but I do not think that is the problem. The problem is low clouds that are where the smoke would be, especially if it is behind a mountain.

This presentation may be viewed at: [2024 5NRP - Mai Nguyen]()

## 4.7 Deep Learning for genomic discovery – Aman Patel, Stanford University, Kundaije lab

Human populations have genetic variants associated with traits and diseases which makes cataloging genetic variation across the human population an essential first step. We have done this through genome sequencing, the more genomes we have the more that we can uncover rate and not so rate variants. You cannot find the variance unless you know many genomes that have been sequenced.

Since the Human Genome project in the early 21st century, the number of genomes has skyrocketed due to the decrease in cost and the power of sequencers has increased. We have been able to uncover millions of common and rare generic variants and can perform powerful analysis to figure out what positions are associated with a variety of traits or diseases. These associations are useful to tell us where to look in the genome and what variants might be important, but they do not tell us what the actual biological mechanism is. Knowing this could help us discover what is causing a disease or trait and possibly develop a corrective action.

Not all genes are active all the time. A complex regulatory mechanism that determines whether a gene is active or repressed are segments of DNA called control elements. Control elements are more numerous than genes and the important parts of them are six- to ten-based pairs of DNA called motifs. Motifs are where protein bonding takes place and then the proteins then perform important regulatory functions. These motifs along with the control elements create a complex DNA syntax that defines regulation in that particular cell type. While this syntax is extremely important biologically, it is poorly understood.

Studies have found that less than five percent of genetic variants associated with disease are within the genes themselves. Ninety-five percent are within the control elements. It is important to study how control elements work, and how genetic changes could alter their function. To do so, biologists have developed excellent molecular mapping techniques that allow us to look across the genome and figure out where each control element is.

There are many cell types in the human body. Each cell type has a different syntax. To do accurate analysis regarding the role of a genetic variant or on gene regulation in general, we must consider each cell type individually. That is a lot of data. Over the past decade, multi-institutional consortia such as ENCODE[58] and the ROADMAP epigenomics project[59] have dedicated resources and effort to produce this data.

With this data, we have everything we need to study regulation on a very fine scale across all cell types and across the entire genome. This very large and powerful data set can be used to train machine learning models to decode the syntax behind regulation in all these cell types. Once that training is done, we can use those models to predict the functional effect of a genetic variant.

The deep learning framework for decoding regulatory DNA uses BPNet[60] to map the sequence to base-resolution profiles and DeepLIFT, FastISM, Yuzu and MoDISCo to infer the contribution of every base in each control sequence through the lens of model. Over 5,000 models have been trained across 1,100 different cell types. This has not been done before with any accuracy. It is enabling biological and genetic insights that would not have been possible without this kind of technology.

NRP is used to train a vast number of these models at scale. Training and analyzing over 5,000 models require a massive amount of compute. The workflow is both CPU-heavy for data preprocessing and GPU heavy for model training and evaluation. Nautilus served as an ideal platform to meet our compute needs. The availability of CPUs and GPUs allowed us to train and evaluate our models in a timely manner and the Kubernetes engine allowed for easy parallelization.

Once they have their models, they can do what is called Sili mutagenesis where they take DNA sequences, introduce mutations, and see how the predictions from their models change. There are many things that can be done with these deep learning models, we are just scratching the surface

---

[58] The ENCODE website can be found at: https://www.encodeproject.org/.
[59] The project website for ROADMAP is no longer active. There is some information available on Washington University at St. Louis' website: https://egg2.wustl.edu/roadmap/web_portal/
[60] BPNet is a Python toolkit to train base-resolution deep neural networks on functional genomics data and to interpret them.

on what can be learned. One example is congenital heart disease where a mutation disrupts an important control element in arterial endothelial cells.

It is important to democratize machine learning for genomics and make sure it is usable by as many people as possible.

### 4.7.1 Questions and Discussion

Question: There are natural genetic differences for humans. How do you figure out the meaningful signals from the noise?

Response: This is where the entire question of model explainability comes in. Where if you know a variant does not do anything right or does not have a very strong biological function, a substitution shows a drastic change and produces a difference in regulation.

Question: Let me make sure I understand. You have singular flips and those singular flips your model was trained across large populations and was told that these people are sick, these people are not sick and therefore it has learning that there are other parts of the genome that are massively different when this gets flipped.

Response: The model was not given labels of who is sick and who is not sick. All it has is the input DNA sequence and the output molecular maps. What is learns is what the syntax is. It is learning what are the important parts of DNA that are regulatory in nature.

Question: Where does the data come from that establishes the language?

Response: The data from the molecular maps is from consortia across the country and various donors.

## 4.8 The Rise of NRP AI/ML Computing across Diverse Disciplines - Larry Smarr, UCSD

This talk gave brief summaries of some of the work that is supported by NRP. Seven years ago, Tom DeFanti had the radical thought that computer scientists could compute on infrastructure. Tom also came up with the idea of getting gaming GPUs and interviewed computer science faculty to come up with a set of campuses for the CHASE-CI proposal on "Cognitive Hardware and Software Ecosystem Community Infrastructure". This was bolted onto the PRP.

Last year at the 4NRP workshop, we showed the new video on the PRP, but there was no mention of AI or machine learning. Showing the last six months of users on the NRP, almost all NRP namespaces use AI/ML. Only three of those are large scale non-AI/ML community efforts: Ice Cube which is using the GPUs for parallel computing of photon tracks and the ice under the South Pole; Open Force Field (OFF), and the Open Science Grid opportunistic.

There are many stories hidden in the NRP that we could reveal. Nautilus was designed to support six broadly defined families of information extraction and pattern recognition algorithms that are commonly used in AI/ML research.

Historically there are six ways to view the different kinds of algorithms used:

1. Deep Neural Network (DNN) and Recurrent Neural Network (RNN) algorithms
2. Reinforcement Learning (RL) and inverse-RL algorithms
3. Variational Autoencoder (VAE) and Markov Chain Monte Carlo (MCMC) stochastic sampling
4. Support Vector Machines (SVM) algorithms
5. Sparse Signal Processing (SSP) algorithms including Sparse Bayesian Learning (SBL)
6. Latent Variable (LVA) algorithms

It is easy to search for namespaces and descriptions by going to [https://portal.nrp-nautilus.io/namespaces.g](https://portal.nrp-nautilus.io/namespaces.g) . You can find faculty and students who are already using the NRP for their studies and research and can share their experiences with you. You can also use Matrix to ask the community for assistance if you are having trouble.

Today, the great majority of Nautilus namespaces are using some form of neural nets or reinforcement learning. For neural nets, PyTorch, TensorFlow and Keras are the preferred open-source deep learning frameworks. AI/ML researchers are using subtypes of DNNs including Deep Belief Networks (DBN), Quantum NNs (QNN), Graph NNs (GNN), and Long Short-Term Memory (LSTM). Transformer neural nets (TNN) have become the default architecture with large language models for applications involving images and sounds.

The OpenForceField namespace has surpassed namespace *osg-icecube* in GPU usage over the last six months. There was a huge jump when they revised their software. Both community efforts are using about the same amount resources. OpenForceField (OFF) is being used for drug discovery.

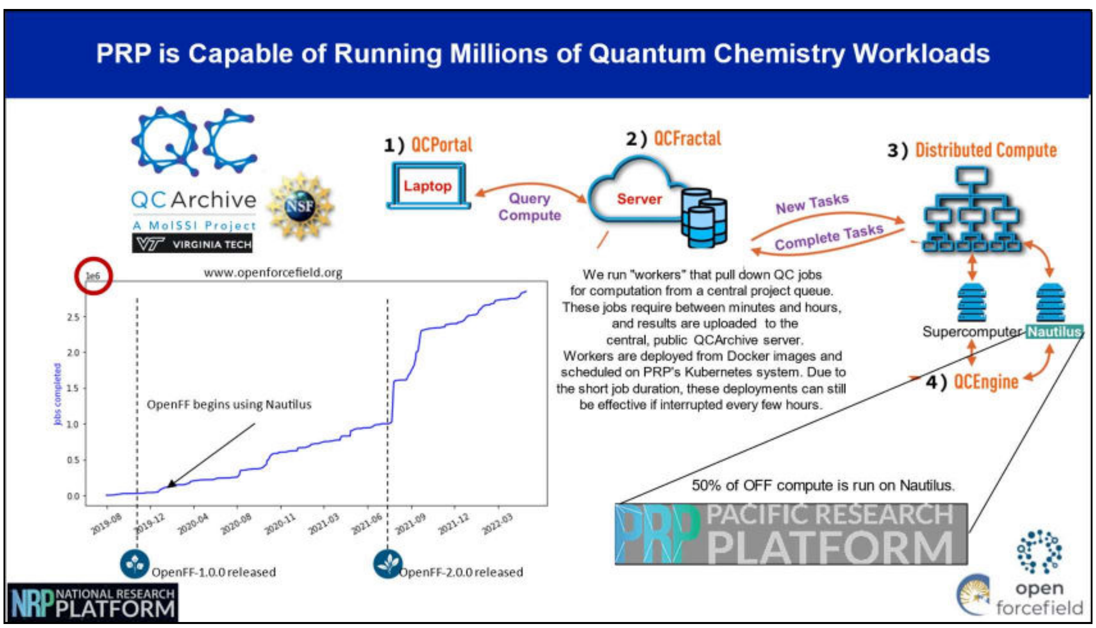


*Figure 13: PRP Quantum Chemistry Workloads.*

A newcomer to NRP, John Choderea (namespace *choderalab*) from Memorial Sloan-Kettering cancer center has been using Nautilus resources for the last six months. NIH funded a $68 million open science drug discovery effort called ASAP (Accessible Antivirals to Prevent Pandemics) and it uses AI and ML to accelerate drug discovery.

After OpenForceField, Hao Su (namespace *haosulab*), who is doing robot learning, is the second largest user. He used almost 5 million CPU hours and a quarter of a million GPU hours.

Frank Wüerthwein and Javier Duarte (namespace *cms-ml*), are taking the CMS on LHC and then using machine learning on the data for unsupervised anomaly detection using the public CMS dataset as a benchmark.

Rose Yu (namespaces *deep-forecast*, *deep-point-process*, *spatiotemporal decision*, *climate-ml*) is bringing AI and ML into Physics. She is looking at things where you have spatiotemporal data that are varying in space and time, and then applying them to areas such as biomedical engineering, quantum chemistry, mechanical engineering, transportation, climate science, public health, and sports analytics. She is also taking the first principle of physics differential equations and applying these new parallel learning techniques. This physics-guided AI aims to increase trust in AI, which is one of the core features of NAIRR.

The *Jupyterlab* namespace has 268 registered users on the NRP. Jupyter is completing the process for those in data science and machine learning that was started back when Tim Berners Lee created HTML and HTTP, “the web.”

CSU San Bernardino is an excellent example of how to help your faculty and students to use NRP. They created a group dedicated to helping faculty and students access high-performance computing resources. Last year they burned more CPU and GPU hours than eight of the ten University of California campuses. They used Jupyter to accomplish this. Use the easy button! This resulted in 450 users from their campus.

The next session will be on how we can use NRP-like capabilities to greatly increase through courses, the output of our campuses in terms of trained AI/ML people entering the workforce. By using courses, we can increase the number of students who can use AI and ML.

The last time we saw something like this was in 1985 when the US realized it needed a supercomputing program and five years after the NSF set up the program, we went from 100-200 researchers in the US that had logged onto a supercomputer to 50,000.

This is not business as usual. We have an elaborate higher education with multiple levels and can move at a speed no one can match. We need to get out and do what we can do better than any other country.

This talk can be found at: [2024 5NRP - Larry Smarr](#) .

## 4.9 Expanding the AI/ML Workforce Through Coursework - Michael Farley, SDSU, and Valerie Polichar, UC San Diego

### 4.9.1 Data Science/Machine Learning Platform for students – Valerie Polichar and Adam Tilghman - UCSD

The Data Science/Machine Learning Platform for Students (DSMLP) was inspired by the PRP and the CHASE-CI projects. In 2017, Larry Smarr challenged UCSD to put something like the low-cost GPU cluster into the hands of graduate students. When an IT department goes about doing something like this, the usual way is to go to a vendor and ask them to spec out a system. But when it is novel and innovative, vendors may not have the right fit and may not be ready to provide a cutting-edge solution. If we want something that is cutting edge, we are limited by the vendor's expertise in what may be a novel field. We will also miss out on learning from world-renowned researchers. By leveraging on-campus expertise, we can design and implement something that is innovative, cutting-edge, and very efficient, with the ability to share experiences and troubleshoot problems with others.

Typical CIOs would respond that this is not the normal way to do things, it has not been done before, you must design it, and there are no standard replacement support contracts. When it breaks, you will have to fix it. Enterprise IT tends to build things that have a lower risk. Research and Instruction have a higher tolerance for risk because there is a tradeoff to get greater capacity and more cutting- edge equipment and that is worthwhile.

UCSD IT accepted the challenge and invested 75,000 dollars to develop an initial system. Designed around the FIONA design specs, the DSMLP rolled out to ten undergraduate and graduate classes. In 2019 they added access for student independent study and researchers with entry-level needs. There are a lot of great offerings for people who know what they want and are ready to go and have larger needs at SDSC, but they did not have a good option for researchers in underrepresented fields who want to start working with machine learning.

The system has grown a little bit every year. Every year dollars saved by other areas have been used to support and expand this system. The system now has almost 72,000 student enrollments, offered 620 classes, touched by 170 instructors in all nine divisions.

What allows UCSD to support all these different disciplines is the variety of choices offered to the faculty. Jupyter Notebooks is the easy button for many disciplines. Shell accounts, terminal logins and a full development suite are available for computer science students. Faculty can install their own library packages, or even build their own customer Docker container. Students have training available in every division and many disciplines in the skills required to perform, design, or build machine learning and/or data science appropriate for that field. Their students graduate with a competitive background that prepares them for tomorrow's jobs.

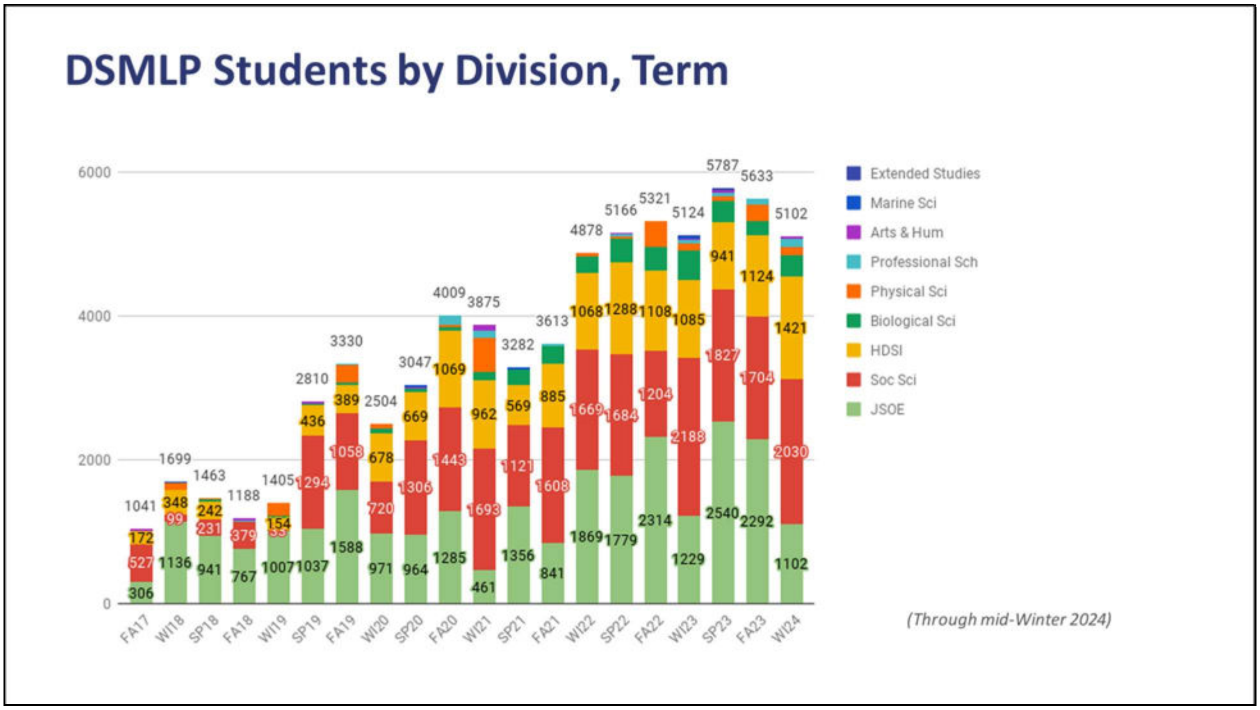


*Figure 14: UCSD DSMLP Students by Division, Term.*

A steady number of disciplines are using DSMLP including two courses in music using ML. Many more could be using the platform and there is every indication that usage will increase over time.

The system is designed based very closely on Nautilus and NRP architecture but has been adapted to fit local IT capabilities. Staff were already familiar with LINUX but not Kubernetes, so a traditional NFS-based storage system is used. For coursework NFS has worked fine and allows IT to align what people do within this environment with what they do on a desktop.

Each student gets a namespace, and this helps isolate their usage. Resources can be tailored to course requirements.

There are 132 GPUs, 23 nodes, 1,024 CPU cores, and 650TB storage via ZFS/NFS housed in two racks. Leveraging this environment, it is a natural place to pilot the campus-facing LLM. There are GPUs available for UCSD students that need them for a project, thesis, or Capstone. The hardware is aging and will need to be replaced. A collaboration with TritonGPT or NRP may be in the future.

Most of the usage is CPU hours, so taking this model as CPU only would be a valuable addition to any campus. The on-prem utilization model works.

To run this, UCSD has 4.5 FTE, 0.5 architect, 0.5 manager, 2 x 0.5 sys admin, 2.5 system engineers plus 5 student staff.

### 4.9.2 SDSU and NRP Nautilus for Instruction

San Diego State University (SDSU) is one of twenty-three campuses within the California State University system with about 36,000 students. They receive about 109,000 applications a year for admission. Since 2020 they have been operating under a new strategic plan where one of the focus areas is to become a premier R1 Hispanic Serving Institution (HSI). They do about 200 million dollars in external grant funding per year.

*"The mission of SDSU is to provide research-oriented, high-quality education for undergraduate and graduate students and to contribute to the solution of problems through excellence and distinction in teaching, research, and service*."

Last year, SDSU administered an AI student survey and received a 20 percent response rate. Some high level takeaways from the survey were that 86 percent of the respondents believe that AI will become essential to most professions, 51 percent are using AI in their work, and 34.4 percent of instructors encouraged the use of AI coursework.

They have developed an AI micro-credential geared towards faculty to prepare them to apply generative AI (gAI) technology and introduce them to terminologies as well as responsible use of AI. About six hundred people signed up for this.

Their journey to NRP started by creating a Cyberinfrastructure Committee comprised of faculty representatives from each college. They did not start with instruction, but instead sent representatives back to their colleges to survey their colleagues on what they needed in the cyberinfrastructure space: compute, networking, storage. The data that came back and indicated there were three clustered areas: JupyterHub (notebook lab), high performance computing and GPUs.

Those three things seemed like a natural fit for Nautilus especially since the SDSU staff did not have high performance computing experience. The ability of the NRP staff to provide that Administration was a key factor in their decision. They still needed to address the need for a Science DMZ so they reached out to EPOC[61] for assistance. They were able to repurpose a core route and switch from a campus refresh and add two CENIC HPR connections to provide connectivity for their Science DMZ.

Based on faculty recommendations and what they knew about NRP's "bring your own resource" nodes, they wrote a proposal called "Visionary Education Research Network Environment" (VERNE) and received funding from the University president for fifteen CPU, GPU storage servers that were added to NRP. JupyterHub is their "easy button." As of this talk, there were 422 JupyterHub users. Similar to CSU San Bernadino they maintain a research and an instructional JupyterHub with different settings and limits on capacity.

To support the creation of containers and support the instructional use case, they created the Software Factory that employs students to be part of the team and work on real issues alongside research engineers. In the process, students learn tools, DevOps practices, how to create container images and use GitHub. They learn critical skills that they can take with them when they graduate.

Some lessons learned include:

- Containers are not that easy.
- Utilize what is already there.

---

[61] The website for EPOC is: https://epoc.global/ .

- Nautilus is a research resource that doesn't always fit 24-7 uptime.
- Growth is slow but steady.
- Support has not been overwhelming.
- Not everyone needs GPUs.

We are better together – the NRP community is its real secret.

### 4.9.3 Questions and Discussion

Question: For Valerie. Let's suppose I am an optimist and would love to see a singular system that takes the DSMLP and turns it into a scalable resource for both research and instructional use and I see some organizational barriers to that and some financial barriers to that. How would you tackle some of these barriers?

Response: The color of money issue is not as bad as it used to be. We do have to be careful about the difference between core and grant funds, but it does not mean we cannot take advantage of having system administrators that work both for instruction and research.

Question for Michael: I love the idea of Software Factory – how do you find the students?

Response: We have not been doing this long enough where I could easily answer that. We do have a Faculty Fellowship where we buy time from faculty, and they become great resources within the college departments. This leads to referrals for both students and other faculty.

Question for both: There is a big push for administrative functions to move to the cloud. Beyond money, what made you decide to do this locally instead of going to the cloud?

Response Valerie: What UCSD has learned is that there are learning experiences with building things, collaborative opportunities. If faculty come up with an idea, and if you have the resources, you can easily meet their needs. By purchasing the equipment, we have established maximum cost – the cloud can be costly.

Response Michael: The cloud has some advantages to keeping hardware current, but OpEx is hard to sustain.

Response Frank: Another advantage to doing this in-house is the student involvement. It gives the students an education and valuable skillset and occasionally you find someone to hire.

Question: Have you hired any of the students?

Response Valerie: Yes, we have just hired one of our students into an open full-time position.

Response Michael: Yes, we hired Kyle who started out as a student.

Comment (Tom): UCSD and University of Washington have dropped the overhead on cloud services, but it is still 4-10 times cheaper to have GPUs on premises.

This presentation may be viewed at: [2024 5NRP - Polichar, Tilghman and Farley]

# 5 5NRP Track – Friday March 22, 2024

## 5.1 The Quilt - Jen Leasure, The Quilt

The Quilt is a patchwork of regional networks across the United States. The map below represents the backbones of Quilt members and their networks. What you do not see is the tens of thousands of last mile connections to higher education institutions, K-12 schools, public libraries, research centers, community non-profits, Indigenous communities, government, healthcare, public safety, and cultural institutions.

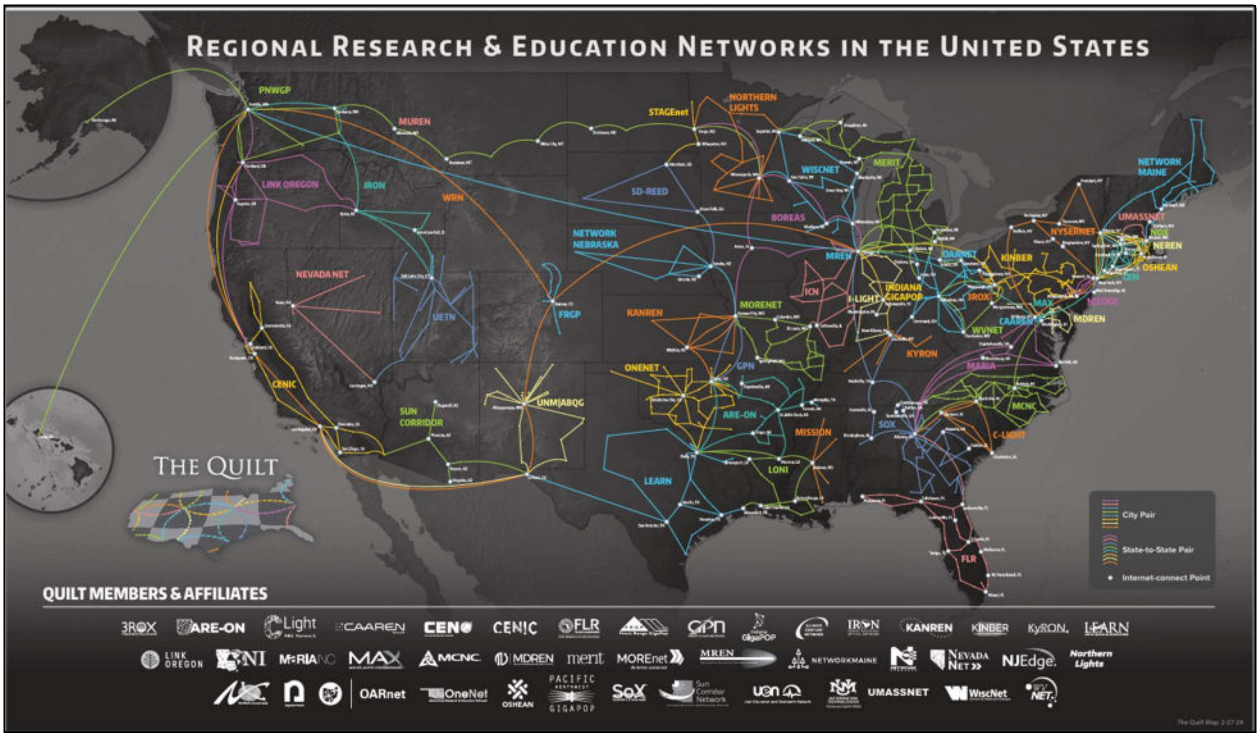


*Figure 15: The Quilt National Map - thequilt.net/maps/national-map/*

Reaching well beyond the roots in higher education, the Quilt members support the continuum between K-12, community colleges, higher educational institutions, and even adult learning. Ensuring that K-12 and beyond have access to cyberinfrastructure resources is one of the goals of the regional networks represented by the Quilt. More than just being advanced network providers, regional R&E networks (RENs) act as regional aggregators of cloud and CI resources providing needed technical support and facilitating awareness and access to the types of resources available through national CI programs. They are a convenor of human networks and serve as bridges from the local to the national resources.

The Missing Millions report found that exposure to the computing and data infrastructure of the country is a real critical element to having a diverse STEM workforce. The engagement of faculty and educators in the CI space ensures that they have exposure to these resources. As RENs bridge to help smaller institutions and ensure that faculty and students have access, they are working with other partners in this space such as the American Indian Higher Education Consortium (AIHEC) and the Minority Servicing Cyberinfrastructure Consortium (MS-CC) focusing on tribal colleges and universities (TCUs) and historically black colleges and universities (HBCUs). They are helping them with strategy, planning, and designing what their campuses need in the CI space and then hopefully secure funding to achieve those goals. There is a challenge in that these institutions may perceive that they do not need these resources or that they are only for large institutions. There is also a challenge in RENs communicating the value of research resources such as virtual labs or data sets. Some of the questions the smaller institutions ask are: how do I get to it? Can Nautilus only be used for one thing? What are the monthly recurring costs? What kind of support do they need? How can I support it when I do not have the staff?

NRP is the easiest path for any institution to participate in the larger compute environment. Once the hardware is up, it immediately shows up on maps. Once on NRP, you are now connected to Open Science Grid (OSG) – participating and contributing to the national compute environment. Once plugged in, you are part of this community. There is support from NRP and OSG as well as the community.

### 5.1.1 Regional Network Examples

Southern Crossroads (SoX) is relatively new to NRP. SoX is deploying nodes as part of CC* award that includes several HBCUs in Alabama and Georgia. They are considering redeploying their FIONA nodes purchased as DTNs as Nautilus nodes. They are offering space and connectivity for those interested in hosting nodes. They are looking for toolkits, demonstrations, and training for their participants that would help them participate in NRP. The goal is to enable research within their community and assist faculty who are unfamiliar with the platform.

NYSERnet has eleven nodes as part of NRP supporting many different disciplines, including immersive visual and augmented reality. It also supports Water in the Changing Coastal Environment of Delaware (WiCCED), an NSF funded collaboration of the University of Delaware, Delaware State University (an HBCU) and the Delaware Technical Community College working to understand the changing coastal environment.

### 5.1.2 EducationDMZ (EdDMZ)

The Quilt is exploring ways to enhance science access, delivery, and collaboration and is investigating the concept of an EducationDMZ which is a regional platform, consisting of servers, applications, connectivity, and support that will make access to science data, applications, instruments, and collaborations easier. It is different from a Science DMZ that aims to provide friction free access to science and focuses on high-performance data flows. With an EducationDMZ the emphasis is on stable hardware and software platforms, intuitive and easy-to-use applications that classrooms need, and accessible resources, where teachers do not have to be the technical expert. Regional networks are a place where this can happen. James McCabe of Sun Corridor has been leading the effort in the Quilt.

An example of this is Precision agriculture in Arizona. The University of Arizona has a research farm and is working with the Arizona Western College and the Yuma Union School District to use the agricultural data connected by the sensors. Each is doing something different with the data.

Scotty Strachan at NevadaNet has different architecture for their EdDMZ that is more complicated than the one Sun Corridor is considering. He is looking to simplify it so that it is easier to maintain and support. The EdDMZ will likely be regional and adaptive meaning differing things to each region based on needs and demographics. They will be collecting Use Cases, which is important to show other campuses how they might apply the tools. One goal is to share this concept more broadly and get feedback.

### 5.1.3 Questions and Discussion

Question: What is a good way of getting in touch with the community at large? A lot of effort to get people to share data.

Response: Start with Jen, glad to be the Point of Contact

Comment: It is a good idea to think about how you position the EdDMZ. The DMZ part of the name is not going to help; however, this is a great initiative because I think it would help many schools and universities if we had more standard environment for providing these features to these infrastructures because that is lacking everywhere.

Response: You bring up a good point about who owns it, where do you go for support. Thank you for your feedback. If educators and staff are going out to external resources that may be housed and accessed through the regional network, where do you go for support? Is that a one-stop shop, or are there other places for support? You may want to firewall some things and there is also the challenge of federated ID which still needs to be solved.

This presentation may be viewed at: 2024 5NRP - Jen Leasure.

## 5.2 Great Plains Network and GP-Engine- Alex Hurt, University of Missouri, and Derek Weitzel, University of Nebraska, Lincoln

The Great Plains Network (GPN) is a collaboration of Midwest universities for CI. It started as a collaborative for network access but grew significantly. Some of the successes have been the grants such as Cyberteams which trained the experts through the GPN states, and GP-Argo[62] which was originally designed to provide a gateway for OSG jobs to access local resources in the GPN. GP-Argo has contributed 23,400,392 core hours to OSG, contributed to 83 OSG projects and supported 57 institutions.

GP-Engine is comprised of six states, eight universities, and two regional network organizations under an NSF CC* Regional Computing award OAC#2322218. GP-Engine purchases hardware and places it in these different institutions that are optimized for computational and AI science. It can handle scaled automation of experiments using Kubernetes and the NRP platform which is important to a lot of institutions that do not have these resources to support computational research. GP-Engine helps researchers that have been using their laptops or local resources to grow first by using GP-Engine nodes and then growing organically out to the rest of the NRP.

[62] Gp-argo.greatplains.net

They are increasing high throughput computing infrastructure within the GPN as well as expanding the NRP with forty-four latest generation GPUs, with over 300,000 cores. There is emphasis on education and teaching educators how to use the NRP and GP-Engine nodes with GP-Engine experts going out to universities and teaching CI professionals and instructors.

There is now about a petabyte of storage in the central region, and this is growing. The University of Missouri is now hosting twenty-four A100's and recently added the first Grace Hopper special platform from Nvidia.

The GP-Engine project is just getting started but so far it has enabled sixty-one research groups, provided 85,000 GPU hours, and 1.3 million CPU Hours. Only half of the nodes are up and running with the other half waiting for institutions to set up the infrastructure. The nodes are quite beefy and can be used in advanced deep networks.

Another important part of the project is the GPU Workbench providing Jupyter containers for AI/ML Data Science, and STEM w/GUI as needed, training researchers to use GP-Engine, and learning to scale on Kubernetes. One of the core reasons this works so well is that researchers can develop interactively on the same containers they do their base prototyping and then convert it into a Kubernetes pod, link it to GPU hardware, and then link it to persistent storage to save their artifacts and outputs. Kubernetes tutorials are offered in-person and via Github: [https://github.com/orgs/MU-HPDI/repositories](https://github.com/orgs/MU-HPDI/repositories) .

While delivering nodes to members, GP-Engine hosts workshops on utilizing NRP for researchers. They also run tutorials during the GPN annual meeting. All resources used are published on GitHub to allow researchers to review them later and are continuously refined with feedback from workshop attendees.

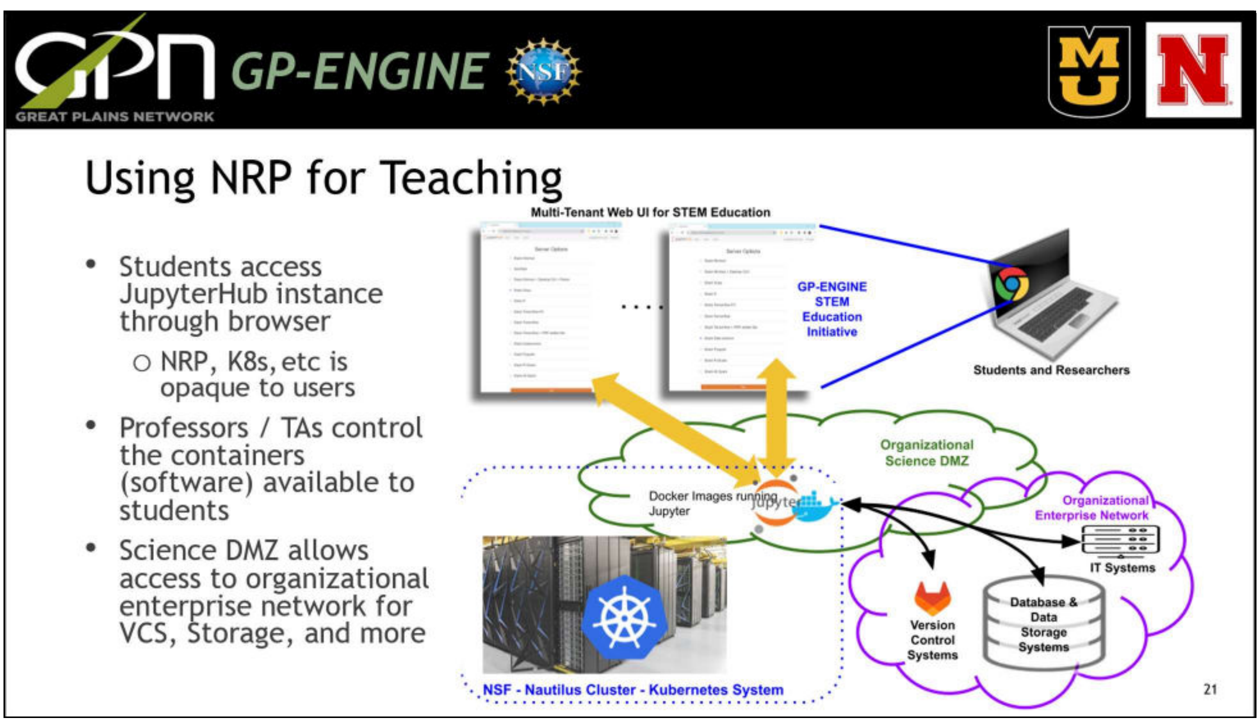


*Figure 16: GP-Engine using NRP for Teaching.*

The Great Plains Network is using Nautilus for computer science and AI education. Each student gets access to JupyterHub through a browser. The beauty of this is that it can remain opaque to the users and the students do not have to know that they are working with Kubernetes or containers, but for more advanced classes they can see the containers. Professors and Teaching Assistants (TAs) can control what software is available, so if the professors or TA needs to provide a library for a homework assignment, they can do so. The Science DMZ allows access to the organizational enterprise network for VCS, Storage, and more. At the University of Missouri courses like Cloud Computing for Data Analytics, or Computer Science Artificial Intelligence use NRP for the course. Parallel Programming for HPC and Parallel Hardware and Distributed Processing are other courses.

Courses taught at UNL include an AI filmmaking Hackathon using generative AI to create film. NRP hosts the AI models and the inference, i.e., standard diffusion. Led by their libraries group, the University of Oklahoma facilitates classes on the use of Jupyter notebooks.

Some research highlights include:

- Deep Learning on Nautilus Transformer Research
- Maasai Boma Mapping using DCNN and remote sensing data - very compute intensive, from the noise to useful data.
- Deep Seasonal Network for Multi-temporal Data Fusion – did all testing of model on NRP.
- Space-borne computer vision: wildfire burned area mapping - took 12 hours to set up, 21 days of wall clock time.
- Nan-Energetic Material Reaction Classification –Different models to understand approach. Not just for large scale batch experiments
- Grace Hopper Superchip: GH200 has been added to NRP. It is a very fast link – quick benchmarks training on ResNet.

### 5.2.1 Questions and Discussion

Question: This is stunningly good. It is needed by everybody. Is there a way that the Quilt can uplift the training? It is fine that it is in GitHub. It would be great if you could find a way to get it out to all the RONs. We don’t have time to do these one by one, need to move quickly.

Response: This is something we could look at. Appreciate the feedback.

Question: Yeah, really great work. Especially in engaging with new or naïve users. At UCSC we looked at existing documentation and studied how novices reacted. We stepped everybody through onboarding and the existing Nautilus tutorials and found there was a lack of context. People needed to switch in and out of different applications to complete a task. So, context switching was added to cognitive load so we are revising documentation – Would you be willing to participate in a second round of study?

Response: Yeah, we can talk at lunch about that.

## 5.3 The Technology Infrastructure for Data Exploration (TIDE) and the Regional Optical Network as an Enabling Partner for Digital Equity - Jerry Sheehan, Salk/SDSU

Everything Mike Farley talked about yesterday and the work that is being discussed today is possible because of everyone who has worked to build these resources. This started back in the 1990's when the supercomputing program developed the concept of the meta-computer. This was then followed by the establishment of CalIT2 and PRISM and then the PRP. Tom, John, and Dima built the PRP, transitioned it into the NRP and infused it with Frank's knowledge of OSG. All of it is contingent on the network.

The Technology Infrastructure for Data Exploration (TIDE) is a recently received NSF Campus CI award for regional computing, led by SDSU, and partnered with CSU San Bernadino, Cal Poly Humbolt, CSU Stanislaus and SDSC. The largest public university system in the United States, the California State Universities, has never had a computational core or anything to serve the science needs of the CSU across all campuses. This is in part due to the missions being instructionally focused but increasingly over the course of the last two decades to make students competitive moving into the research environment.

The new ecosystem contains $800,000 for new hardware for a computational core consisting of 616 cores, 73 GPUs and 14.5TB worth of RAM that will be integrated into the NRP. Equally important, the secret sauce is not just the hardware but the people who can take time to work with non-traditional users and translate the challenge which is containerization. Containerization is enough to put off most scientists from changing their methodologies regardless of their expertise. The critical component is the people to help with the translation and the network. Part of this award funds $200,000 for student hiring at four partner campuses to help onboard and support new users.

The role that regional optical networks (RONs) play in TIDE is foundational. It is the essential thing that allows things to happen. The beautiful thing about CENIC infrastructure is the ability to put a machine in a very large machine room at SDSU and make it available to other campuses. As an aside, SDSU has colocation space available. The research infrastructure ties back to CENIC and the science drivers that you see from biomedical to natural sciences to AI to archaeology and geology and mathematics, all of those get back to the CSU campuses on the standard CENIC connections. Science networks are useful, but not always necessary to get science done.

TIDE will be integrated into the National Research Platform Nautilus Hyper-Cluster. This allows CSU to let others manage the resource and concentrate on onboarding and engagement with users to drive science faster. It also allows CSU to have burstable resources on a scale that is a power law change. There are fifty institutions that are collaborators in 300 research groups, everything that anyone is contributing here is free. In the NRP at no point are you asked to enter a credit card number and that is a radical thought. Everyone is contributing to the common good and this is important when we think about equity.

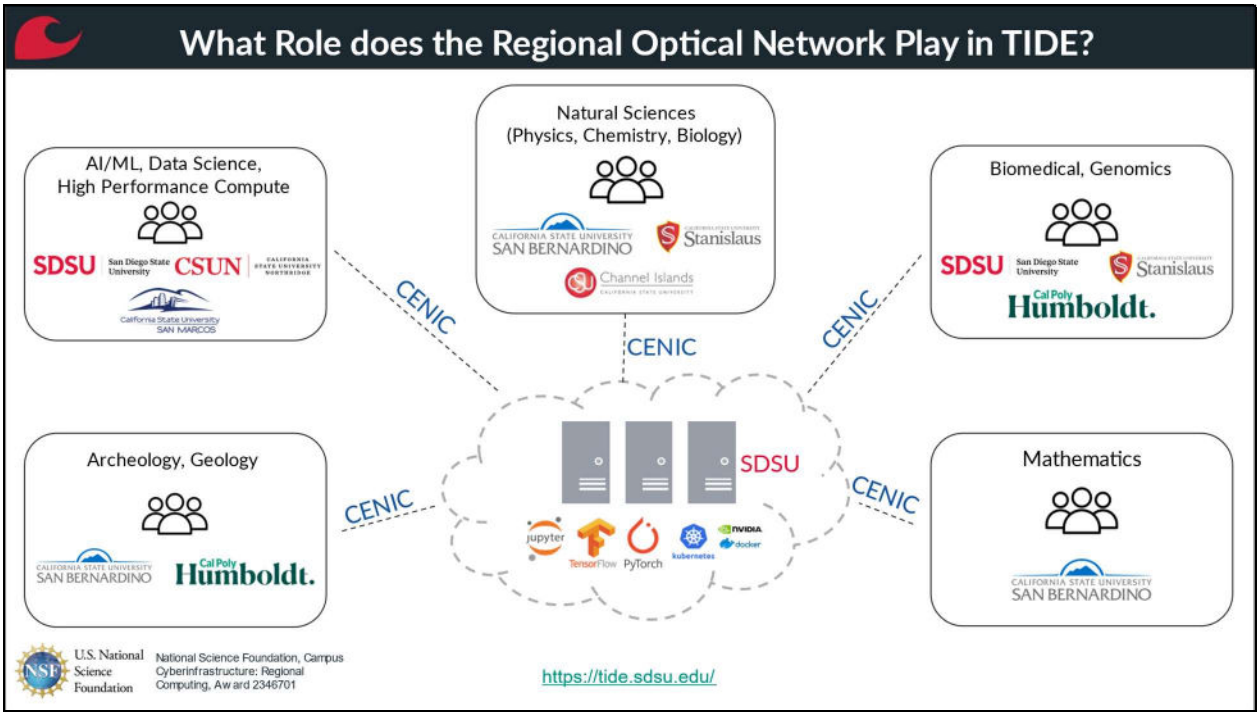


*Figure 17: What Role does the Regional Optical Network Play in TIDE?*

While being the largest public university network in the United States, there is limited deployment of science networks on any of the campuses. For the twenty-three campuses, the system funds the campus production networks, but has not spent a dime into a research network. That is left up to each individual campus. Three of the campuses, San Diego State, CSU San Bernadino, and Humbolt State have been able to build infrastructure with limited staffing.

Science networks are inherently bursty, bursts which are required for discovery and collaboration. It is hard to convince campuses to invest in a network that is only used every now and then but is essential. Research networks are never going to make it to the table when compared to better Wi-Fi on campus for everyone. It requires us to rethink the concept of hardware-based research networks. Software-defined architectures have allowed us to change that thinking and firewalls have gotten better at allowing research traffic to go through them. Software-based solutions need to be where the science is, or the science is going to be disabled.

The NRP is not about equal access. Not everyone needs the same thing. The value proposition of the NRP is that it is more equitable because you do not have to have the same resources that everyone else does. SDSU, on behalf of the largest university system in the United States was able to contribute 75 GPUs but get access to nearly 1,300 GPUs.

## 5.4 CENIC AIR and the Science DMZ Model – Christopher Bruton, CENIC

CENIC operates California's research and education network and serves the University of California system, the California State University system, independent institutions such as Caltech, Stanford, and USC as well as K-12 school districts, libraries, community colleges and other non-profit organizations. The network, called CALREN, plays an important role in the network infrastructure used by the NRP.

The Science DMZ model as viewed by CENIC is a dedicated high-throughput, low-latency network access for research computing. NRP nodes and other high performance computing applications require high performance networking access and in general, this cannot be satisfied effectively by campus enterprise networks. These networks are designed to provide internet access to as many campus users as possible but are not optimized for high throughput or low latency. Every device between the edge and the end user may introduce latency, congestion, or firewall restrictions making troubleshooting much more challenging.

The Science DMZ model bypasses as much of the campus infrastructure as possible. CENIC sees two Science DMZ models already in use by CENIC campuses, and there are two more that were in development and are newly offered to CENIC participants as part of CENIC-AIR.

Option 1 - Virtual Science DMZ is probably the fastest way to build a new Science DMZ since it does not require new circuits or new hardware. It makes use of the existing physical connection with a special VLAN for science. The campus is still responsible for their own routing and switching and the physical infrastructure on the campus side.

Option 2a – Physical Science DMZ requires an additional physical handoff at the CPE, is higher cost, and has longer lead time. It has a dedicated connection to the Science DMZ.

Option 2b – Physical connection is to a dedicated Science DMZ switch. It is higher cost but may better isolate traffic. The campus still owns and operates this Science DMZ border device. Some of the advantages are that traffic is completely separated from the main border router and the dedicated device might provide better performance with deeper buffers.

### 5.4.1 Interlude: Introducing CENIC AI Resource (CENIC AIR)

CENIC-AIR is a new branding of the California portion of the National Research Platform representing CENIC's commitment as an active participant in the NRP program beyond being a network provider. CENIC will provide design and engineering support to campuses to establish their Science DMZs and to facilitate access to the NRP.

With that in mind, there are two more Science DMZ models that reflect this active approach.

Option 3: Remote Science DMZ is geared toward campuses that cannot host their own physical NRP modes due to lack of space and power or insufficient funding to purchase and maintain the nodes. CENIC would extend the Science DMZ from one campus over the CENIC backbone on a layer2 level to another campus. It is essentially a virtual extension of a Science DMZ across the wide area network.

Option 4: Managed Science DMZ is the most radical model that CENIC is offering. CENIC will provide and manage a Science DMZ device on the campus and includes 24/7 production level support. The campus simply needs to connect their NRP modes to the switch, configure their IP addresses, set up default routing, CENIC handles the networking, so the member can focus on the computing. The campus must provide space, power, remote hands, fiber runs, etc. to support the CENIC device. No complex routing is required. This option is only offered for CENIC AIR use cases.

### 5.4.2 Questions and Discussion

Question: In option 5, I know that you have equipment on our campus, but do you have CPE means that you have equipment on every campus in the CSU system?

Response: Not necessarily. Some have a direct provider connection to the CENIC backbone.

Question: So, to implement this, in some cases you would have equipment at a location for the first time.

Response: Yes, and I think that is not reflected in the diagram. CENIC would need to work on the design. It might require a new circuit.

Next Question: If campus has a data producing instrument, would you be willing to plug it in?

Response: Officially, I can't say yes.

Comment: PRP wrote into the proposal instrument integration. There is a lot of interest in agricultural applications for NRP.

Comment: A common failure mode for UCSC is big instrument data streams that can produce data faster than can be handled.

These presentations may be viewed at:

(TIDE) 2024 5NRP - Jerry Sheehan and (CENIC AIR) 2024 5NRP - Chris Bruton .

## 5.5 Regional Networks and Support for Data Intensive Science - Joe Mambretti, Northwestern

MREN was founded in 1993 as NSF was getting out of the NSFNET business and was designed for data intensive science connecting supercomputers together. Tom DeFanti suggested bringing in international networks and so this talk will include them. MREN is international by default.

From NSF CIF21 Framework for the 21st Century (CIF21):

“This vision of the near future shows clearly the urgent need for a comprehensive, scalable, cyberinfrastructure that bridges diverse scientific communities and integrates high---performance computing, data, software, and facilities in a manner that brings theoretical, computational, experimental, and observational approaches together to advance the frontier.”

During this workshop we have seen a new generation of instrumentation leading to new knowledge discovery with deeper insights into the world of molecules, atoms, and the microworld of the universe. We are on the verge of the greatest scientific discovery era in the history of the world.

The National Research Platform democratizes that capability which allows many minds to make new discoveries, investigate and understand the tools of science. But science is global. Open information sharing is a cornerstone of the scientific process. The Global Research Platform[63] was created to share these concepts.

Large scientific instruments will produce more data than we can handle. The Large Hadron Collider, large telescopes in Chile and Western Australia all generate data which is computed around the world. Sharing infrastructure is key. The GRP is supporting some of the largest science projects in the world and almost all of them are creating next generation instruments that will generate orders of magnitude more data than current instrumentation. This data must flow somewhere.

Examples of these new instruments are the High Luminosity LHC, Square Kilometer Array, Vera Rubin Observatory, KSTAR Korea Superconducting Tokamak, Next Gen Advanced Photon Source, and Biomedical/Genomic sequencers.

Leveraging work done previously by the Global Lambda Integrated Facility (GLIF) which has evolved into the Global Network Advancement Group (GNA-G), the Global Research Platform (GRP) is designed to share infrastructure as well as concepts worldwide. There is an annual workshop held with the last few being collocated with the IEEE International Conference on e-Science conference[64].

An example of one of the areas we have become involved with is the Large Hadron Collider Optical Private Network (LHCOPN) which takes data from Tier0 at CERN and distributes worldwide for to Tier1 centers for compute. The Large Hadron Collider generates more data than any other science instrument in the world. There are also Tier2 and Tier3 centers supported by this infrastructure by the private overlay network Large Hadron Collider Open Network Environment (LHCONE), and it is so successful it is being used by other science groups such as the Belle II Experiment, Pierre Auger Observatory, LIGO and Virgo, Xeon Dark Matter project, and DUNE. Others have asked if they could join this network, and the answer is no. Not everybody can fit into the same boat, so we are conceptualizing one that can use the same architecture.

Another example is the SKA which will produce more data than the LHC from Western Australia and South Africa, with data going to compute centers and repositories worldwide.

Some of the key issues are orchestration across multiple domains. Any given project can cross eight domains. We also must deal with high-capacity WAN flows. This is one of the reasons we are moving toward 400 gigabit end-to-end disk-to-disk transfers, high fidelity data flow monitoring, visualization and of course, AI. This requires international testbeds. You must test the capabilities before you put them into production. This is something that we do with our partners.

There are a lot of building blocks to do this. One of them is AutoGOLE. Other partners are ESnet6, Internet2, Quilt and the new NA-REX[65] collaboration connecting North American R&E exchanges with 400 gigabit connections. At the StarLight facility, we now have twenty-four 400-gigabit connections and one hundred and ten 100-gigabit connections.

---

[63] The website for GRP can be found at: https://www.theglobalresearchplatform.net/

[64] The IEEE International Conference on e-Science will be held in Osaka, Japan, 16-20 September 2024. Information can be found at: https://www.escience-conference.org/2024/

[65] NA-REX Press Release: https://internet2.edu/network/initiatives-partnerships/global-networks-and-partnerships/north-america-research-and-education-exchange/

To create the future, we need creativity for next-generation science and new techniques and technologies. We need testbeds to do that. Currently they are supporting 28 network testbeds. We do the experiment and the testing and put them into prototype and if they are successful, we can move them to production and integrate them into operations.

Every year as part of the Supercomputing Conference, SCinet puts together a testbed to do experiments and demonstrations of emerging technologies. StarLight supported multiple Network Research Exhibitions (NREs) at SC23. One of the demonstrations was for a project called NOTED (Network Optimized for the Transport of Experimental Data). It is an AI-based model that looks at flows coming out of the Large Hadron Collider and based on the historical pattern of those flows predicts what flows will be in the future. It then does real-time provisioning with about 95% fidelity in prediction which is the superpower of AI.

Other examples of demonstrations were included in this talk and included descriptions of NREs from SC23, the Data Mover Challenge at SCAsia, and the planned demonstrations at OFCnet. Also mentioned is the emerging Chicago Quantum Exchange Testbed that will integrate the quantum computers that show up at Fermi Lab and Argonne Lab and other universities. It uses dedicated fiber.

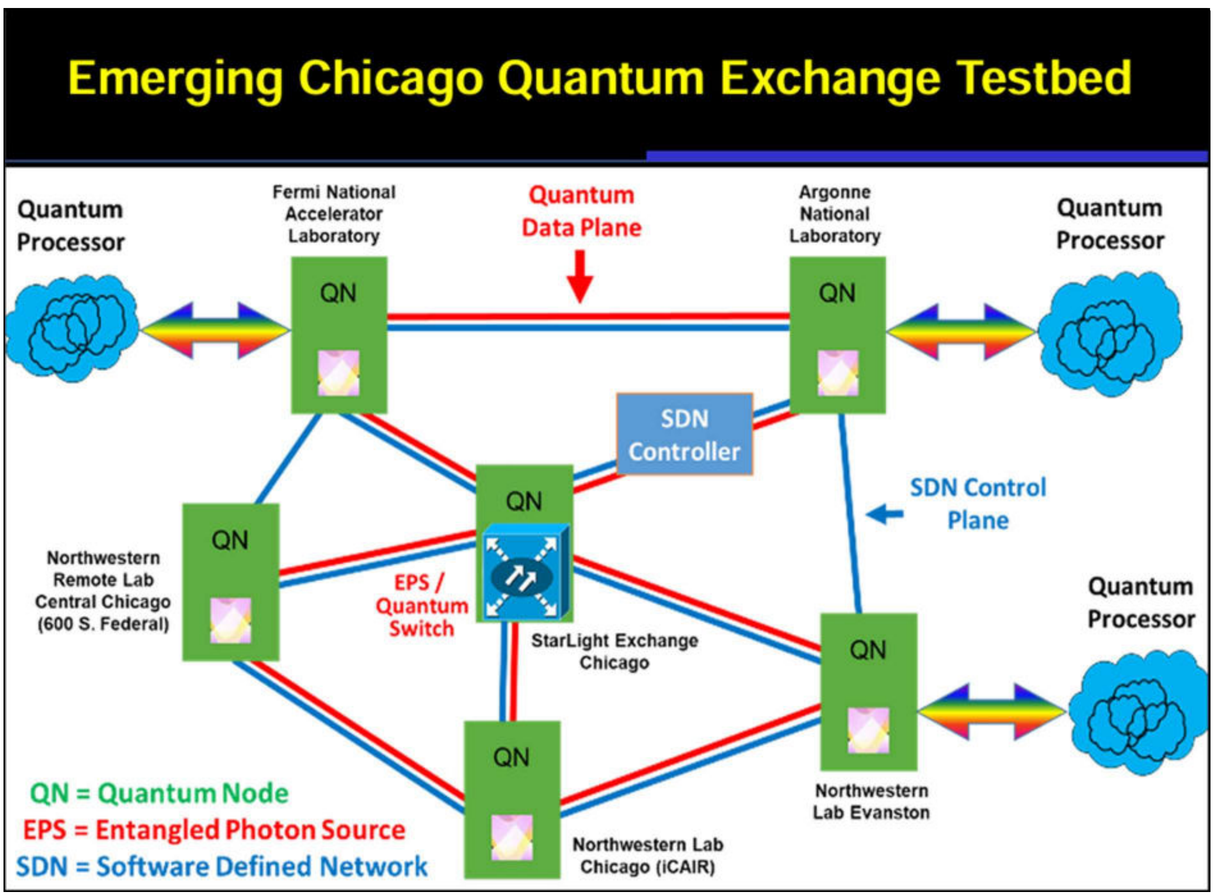


*Figure 18: Emerging Chicago Quantum Exchange Testbed.*

Last year during OFC23, they showed that you can co-propagate classical channels and quantum channels in the same fiber. They are connecting quantum computers in the same way the Internet was created to connect supercomputers. While currently they are limited to 150km due to a lack of quantum repeaters, this is not true of line-of-sight. You can go to satellites for thousands of miles and that may be how they connect the various NRP sites.

In preparation for the High Luminosity LHC which will require terabit connections end-to-end, they are building a prototype testbed with 400-gigabit data transfer nodes on either side to stand up the service. Ciena's CENI Testbed will provide another 1.2 terabits.

### 5.5.1 Questions and Discussion

Comment: I want to compliment you on all your maps. If you want to cause any of the generative AI programs to hallucinate, ask them to describe our networks. You get the weirdest responses.

Response: Yes, I have tried and seen the hallucinations. ChatGPT told me that iCAIR was founded by Argonne National Lab and Fermi Lab.

Question: Could you say a little more about NA-REX?

Response: It is a great resource. Looking at the NA-REX map, you see the partners on the right and the sites on the chart. It has a 400-gigabit backbone, but there are a couple of segments that will not be upgraded immediately and will be at 100-gigabt such as between Montreal and Chicago.

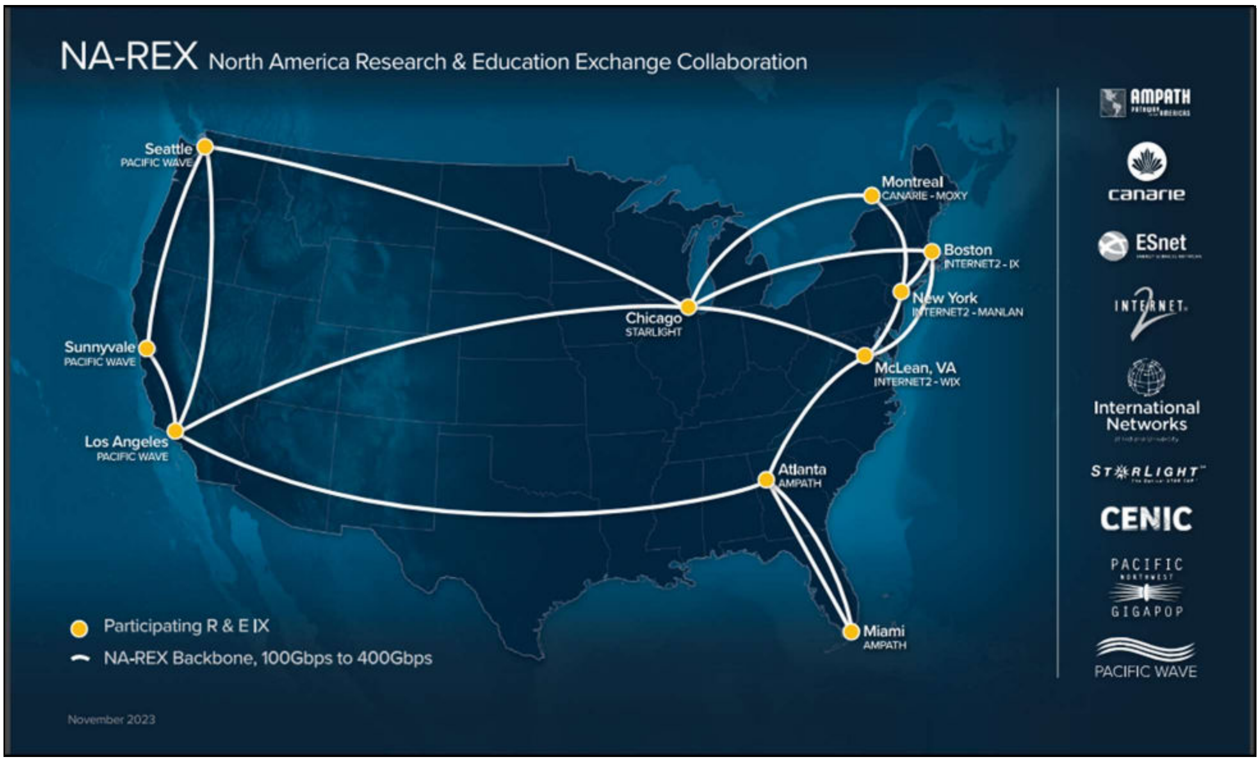


*Figure 19: NA-REX Map dated November 2023*

Question: What is the usage model? Who can use it?

Response: The target usage is research and in scientific workflow research, networking research, new innovations on the network plus production for science

Question: If I peer in Los Angeles on Pacific Wave, can I peer into it?

Response: Yes.

Question: What is the latency between those nodes?

Response: It varies by site, but ranges from a few tens to 120.

This presentation may be viewed at: 2024 5NRP - Joe Mambretti .

## 5.6 Expanding NRP Usage with Jupyter at CalState San Bernardino - Youngsu Kim and Dung Vu, CSUSB

California State University, San Bernardino (CSUSB) has more than 70 traditional degree programs as well as educational credential certificate programs, and a doctoral program in Education. They are currently designated as an R2 campus, but their primary focus is on education. They serve the two largest counties in California.

Nearly one thousand faculty serve 20,000 students, 70% of which are Hispanic and 57% are on Pell grants. They have a good male/female ratio of both faculty and students with more female faculty and students compared to male faculty and students. The faculty are enthusiastic about mingling and working with students.

Though their focus is on education, there is a high level of research activities on campus. They are in the top five CSU institutions for faculty-led research. For the academic year 2022-2023, they were awarded $44 million in federal funding.

Historically, most funded projects did not require high performance computing back in 2015, but when needed, faculty went on their own to garner resources from their colleagues and partner institutions. For those seeking HPC resources, they need to find the resources with little funding and lack of expertise to build an in-house infrastructure. They experimented with building an HPC cluster, hoping to secure NSF funding.

Meanwhile, their library connected their faculty with XSEDE members. This did not click and they transitioned from XSEDE to the San Diego Supercomputing Center. At a CENIC conference in 2017, they were introduced to PRP and joined. They were the second CSU after San Diego State to join the platform. They created a Science DMZ to connect to the platform and attended a FIONA workshop. Currently there are three nodes dedicated to NRP.

There are five people on the HPC Program team that support faculty-led projects, research groups, and classes that have HPC needs, most of which are using JupyterLab/Hub.

JupyterLab is a highly extensible, feature-rich notebook authoring application and editing environment and is part of Project Jupyter. A typical way of using it is the way the University of Missouri is using it. Keep in mind that everything is happening in the browser. It offers a lot of features. There is a file tab to upload/download files with an easy drag and drop feature and is easy to open files and share them. Jupyter is good with large CSV files, which is perfect for those people who need to open large data sets to analyze them and have difficult doing it in their local computer.

It supports Markdown, the default language on GitHub. It is excellent for collaboration. There is TensorFlow 2 QuickStart for beginners, PyTorch and Kagle offered in Jupyter notebooks. Instead of needing to learn secure FTP or STP to transfer a file, or Cube CTL copy command in Kubernetes, users can just drag and drop. JupyterLab is quite portable and wherever you work you see the same environment.

Currently, CSUSB has almost 400 users and several research groups with five to ten students in each.  They do not have the capacity to control or monitor the JupyterLab environment, but JupyterHub does this automatically. It deploys resources and allocates and controls user credentials through a cyberinfrastructure logon. They do not have to worry about controlling the resources, or how to distribute resources. They can customize the environment for each user and each faculty member is given admin access so they have full control of who can use it, who to add or delete.

To make sure everything is working for classes, they auto-start all student servers 30 minutes before class starts, so when the student walks into the classroom and opens their browser, they are ready to go. In addition to NRP's documentation, they are working on more customized documentation for their campus.

A research example:

Dr Cousins in the Biochemistry Department is studying how molecules are positioned and how they stay against each other. Using the Vienna Ab initio Simulation Package (VASP), a computer program for atomic scale materials modelling, as the main software, her team needs to do a lot of calculations. can choose GPUs, etc. Her lab is provided with a customized environment, and can select the number of GPUs, CPUs, and RAM needed.

She originally created 21 similar cases of a single experiment that she wanted to run and if done serially it would take a month.  The HPC team worked with her to create batch jobs, run it on JupyterLab as it can run with minimal resources.

### 5.6.1 How Jupyter Notebooks are Provided

The Jupyter approach is by far the most efficient and scalable way to provide HPC access to everyone. Over 450 distinct people have logged onto CSUSB's JupyterHub which is available to all CSUSB members, students, and other collaborative universities. They leverage NRP's resources to deploy research computing packages, including customized software stacks. Production readiness 24x7 availability with dedicated tech support.

Since NRP uses GitHub, Dong Vu uses the command line to convert codes and then copies samples and code from NRP to their JupyterHub. This accelerates the deployment of JupyterHub. GitLab allows them to develop and deploy the code simply and easily. NRP includes a huge rich library of the Jupyter stack making it possible to deploy JupyterHub in a short amount of time. However, it is very important that JupyterHub is in a production environment, with dedicated technical support 24/7 so it is available to faculty and students when they need to access it.

The first use case is for the public. For example, a student needs to access JupyterHub to do their homework. In this case, the resource requirement is low, but the readiness requirement is high. To provide this, CSUSB sets some limitations to preserve resources. A student's port is scaled down if they close their browser for 30 minutes or more.

The second use case is dedicated JupyterHub for research for faculty and researchers, with no limitations and including very large HPC resources and high-end GPU. This can run without interruption until the job is complete, even if the faculty or researcher closes their browser. For this use case, the resource requirement is high. Sometimes when the resource is not available at CSUSB, a dedicated NRP node helps provide the resources needed.

The third use case is for faculty, researchers and projects that need multiple dedicated JupyterHubs and dedicated systems. For example, if a project needs specialized software, specialized storage, or the faculty wants to control the user access to the hub, they deploy a dedicated research hub. To date, there are ten dedicated research hubs.

In addition to the repository at NRP, CSUSB learned how to deploy their own stack as shown in a typical HPC Desktop below:

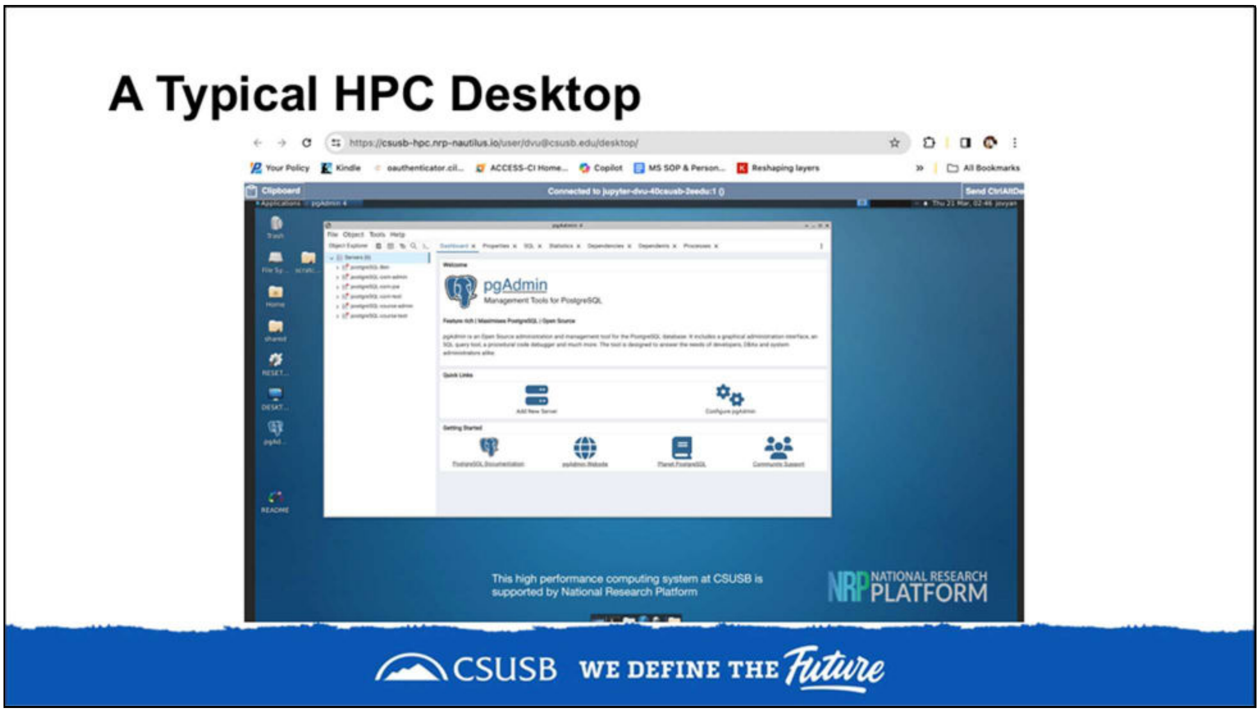


*Figure 20: A Typical HPC Desktop at CSUSB.*

The user can access and run a program with the GUI application and the power of HPC.

With so many pods running at the same time, they need to monitor the resources. There is a cost related to AI. There are 14 AI/ML projects currently and they all have their own research hub.

Recently SDSC and CSUSB received a $1 million grant for a huge infrastructure deployment which will help them accelerate and expand this program.

Their future plans include promoting HPC awareness, expanding HPC support to all domains, expanding collaborations with CSU sisters to share HPC services and provide AI/ML support. They also want to expand support for HPC related courses and the research groups that involve student participation. The students can learn the skills and become scientists in the future.

NRP has not only provided computing resources and hands-on experience but has also provided technical support and solutions. It is a model to advance HPC programs especially for minority serving institutions (MSI) that lack cyberinfrastructure and HPC expertise. It has allowed CSUSB to provide research computing capacity to its faculty and students previously only available to University of California campuses.

### 5.6.2 Questions and Discussion

Question: Given how horizontal the Jupyter is, what would it take to scale across all the CSUs? NRP and CENIC AIR are available to all of them. Is it extra money for staff? Coordination with other people?

Response: Some knowledge we learned. Building trust on campus is important. Security is an issue. Could NRP make a tutorial on how to build a JupyterHub?

Question: For the faculty who use this JupyterHub for their classes, how tech savvy did they have to be, and how much handholding do they need?

Response: The faculty needed the minimum knowledge, just like any other user. If a problem arises that is beyond our authorization or ability, we pass it to NRP.

Question: A related question. How hard would it be to scale this up? Would you be able to support it if there were 100 to 200 faculty that wanted to use this for their courses?

Response: When the demands scales, if we have another support staff person or graduate assistant to take care of certain tasks, it should not be a problem.

This presentation may be viewed at: 2024 5NRP - Youngsu Kim and Dung Vu

## 5.7 Toward NAIRR - Frank Wüerthwein, UC San Diego

One thing that has been clear the last few days. NSF has sent email around to its resources that weren't in the first allocation. What can you provide?

Within the last few days, it has become totally clear that what NAIRR does not have now is the educational platform. We do have something within the scope that NAIRR wants to offer. What NRP should be offering the NSF for the next round of allocations of NAIRR is an educational platform. We will see how we can figure this out and get other institutions on the bandwagon.

It is important to get students involved not just as users but as workers. This gives them valuable educational and work experience to do something that supports others.

The CENIC-AIR presentation was impressive as were those of the other regional networks. It is an incredible community. Maybe a real theme here is how to accelerate and scale out the integration of 4,000 institutions of higher learning via the RENs and RONs. The CENIC AIR offering, especially Option 4, is one way that networks can support their institutions. Perhaps other RENs will find out ways to do so and NRP will be there to help.

### 5.7.1 Discussion

It would be good to have a 24x7 support system instead of an 8x5 support system. A year and a half ago, Larry and John were pushing for AI integration into Matrix for user support, but the world has changed in the last year, and this seems much more natural now. A colleague at MIT has done this for support of their cluster and the cluster now has all the email tickets processed through an AI and the AI gives some possible responses.

Another dimension is looking at data. In the initial NAIRR pilot, there are a lot of data resources missing from an astronomy perspective. There should be a way to browse the OSDF registry of origins and discover the data collections that are available. Ideally, you would have an easy button so you could push a button and get the *yaml* file and then pick the stack you want for that data and then push a button and get the *yaml* file that you need to submit to NRP to launch a container for that data.  Something like that would be great. This would be a great resource for NAIIR.

NRP has a plan for this. NSF funded two projects as part of the National Discovery Cloud for Climate called Pelican, National Data Platform to build on each other. Pelican is the software that drives OSDF. Pelican is charged with connecting all these data sources via origins and NDP is charged to make it findable.

One of the things to consider with the sustainability and growth of NRP is making it easier to tell the story of the impact of the NRP. The science stories we have heard over the past few days could be short write-ups. Being able to get information on how your campus is using NRP should be automated. This would help people get the data they need to tell the story of why this is important to their campus. Another thing would be to calculate GPU offsets and the environmental savings of using NRP. For everyone who uses an NRP resource who does not need to extend a GPU-based machine room with the accompanying electrical and power needs, those offsets matter when explaining network costs. Move more people to the easy button but figure out how to customize JupyterHub in a way that we can see more usage data.

Very few people have access to all the information that has been recorded. There is a standing offer from James Deaton to use Internet2's PR machinery to put the stories out. If we manage to get the visibility, create the stories, we can put the stories out.

Thanks to the AV crew, and the support staff: Chante, Max, Susan, Tracy, and Cindy. We will probably do one of these workshops next year, so if you have any suggestions send them to Tom DeFanti.

Workshop Adjournment.

This concluding segment of the workshop  may be viewed at: [2024 5NRP - Frank Wüerthwein - Discussion]

# 6 Acknowledgements

General Chair:

- Frank Wüerthwein

Program Chairs:

- Gwen Jacobs
- Larry Smarr
- Tom DeFanti

Local Arrangements and SDSC Facilities:

- Susan Rathbun
- Cindy Wong
- Chante Powell
- Tom DeFanti
- Crystal Liu

Atkinson Hall (QI) facilities:

- Hector Bracho
- Maximino Carreon
- Tracy Cornish
- Tom DeFanti

Participant Support funding:

- NSF OAC Awards: 2112167, 2005369, 1928224

Workshop Report

- Celeste Anderson
- Tom DeFanti

NRP Operations Staff:

- Sam Albin, U Nebraska-Lincoln
- John Graham, UC San Diego
- Ashton Graves, U Nebraska-Lincoln
- Tom Hutton, UC San Diego
- Dima Mishin, UC San Diego
- Derek Weitzel, U Nebraska-Lincoln
- Mahidhar Tatineni, UC San Diego
- Huijun Zhu, U Nebraska-Lincoln