FOURTH NATIONAL RESEARCH PLATFORM WORKSHOP

February 8-10, 2023

ABSTRACT

This workshop was supported by NSF award #1826967. The 4th NRP workshop was held at UC San Diego.

celeste anderson

Table of Contents

1. Executive Summary

The many advances in data-intensive science and engineering research and the underlying cyberinfrastructure are enabling new insights and discoveries. To further accelerate these advances, new system architectures are required. A cyberinfrastructure (CI) ecosystem designed to be open and scalable, that can grow with time, and advance through contributions of compute and data resources by the national science and education community may be seen as a collection of resources, tools and expertise that can be integrated and used to enable transformative discoveries across science and education.

Designed for growth and inclusion, the National Research Platform (NRP) is a partnership of more than 50 institutions, led by researchers and cyberinfrastructure professionals at UC San Diego, University of Nebraska-Lincoln, and UC Berkeley and includes the National Science Foundation (NSF), Department of Energy (DoE) and multiple research universities and networks in the United States and around the world. Evolving from the Pacific Research Platform (PRP), this novel, national scale, distributed testbed architecture is supported by several National Science Foundation grants.

As part of NSF Award 2112167, a workshop was held February 8-10, 2023, at UC San Diego in California. The fourth NRP (4NRP) workshop brought scientists, researchers, and technologists together to share their experiences using cyberinfrastructure and discuss future directions and aspirations. The workshop website is https://nationalresearchplatform.org/events/fourth-national-research-platform-4nrp/ and contains links to videos and presentation slide decks.

2. DAY ONE- Wednesday February 8, 2023

2.1 NetSage Tutorial

A well-attended pre-workshop NetSage [1]tutorial was conducted by Jennifer Schopf (TACC/UT Austin) and Jason Zurawski (ESnet). The goal of the tutorial was to walk participants through the basic architecture of NetSage and then focus on use cases. Using a hands-on approach, users found out about flows coming and going from an institution, between two sites, or interacting with a particular network and how the tool could be used to identify partner science drivers, identify questionable routes, and to better understand their data trends for future planning. The slide deck for this pre-workshop is available at "EPOC and NetSage for 4NRP": 4NRP Presentations – Slide Decks – National Research Platform.

2.2 Reception

In the evening, a tribute to Dr. Larry Smarr was held at the Birch Aquarium at Scripps. A video of this event may be viewed at Tribute to Dr. Larry Smarr - Fourth National Research Platform (4NRP) - YouTube

[1] The NetSage website is located at: https://sites.google.com/site/netsagensf/home#NetSage

3. DAY TWO – Thursday, February 9, 2023

3.1 **Amy Walton**, Deputy Office Director, Office of Advanced Cyberinfrastructure, National Science Foundation

Keynote: NSF's Computational Ecosystem for 21st Century Science & Engineering

Amy Walton, the Deputy Office Director of the Office of Advanced Cyberinfrastructure (OAC) at the National Science Foundation, gave the Keynote entitled "NSF's Computation Ecosystem for the 21st Century Science and Engineering.

Looking back at the start of the Pacific Research Platform (PRP) seven years ago, Amy and Kevin Thompson, then both program managers, collaborated between two OAC programs (CC*DNI and DIBBs) to solicit proposals that would foster multi-campus collaboration. Five awards were given in Area 1 of this solicitation, one of which was the Pacific Research Platform (NSF Award #1541349).

The program was looking for multi-campus collaborations with an emphasis on science-driven sharing of data beyond the walls of the institutions involved. Originally proposed to include half a dozen states, PRP expanded well beyond this to include many more locations, including international connections. Five technologies that were not mentioned in the original proposal emerged and were integrated into what became Nautilus: Kubernetes[2], containers, automation, Jupyter[3], and Ceph[4]. There were also five applications not mentioned in the proposal that have since become very important: Machine Learning (ML), Artificial Intelligence (AI), Neutrino Observatory, COVID, and wildfires. These application areas became some of the largest PRP CPU/GPU consumers. PRP was open and responsive to the science community, and it was possible to add new things, new ideas, new problems, and new possibilities.

There have been many contributors to this platform. TNRP (NSF Award #1826967) expanded the platform to regional networks. EXPANSE[5] provided additional resources for the community with testable, checkable pilots. Voyager[6] and Prototype NRP (PNRP) were awarded from the Office of Advanced Infrastructure (OAC), CHASE-CI (NSF Award #2120019) included integration with the cloud and added GPUS for faculty to train AI algorithms. Nautilus has nearly 20,000 CPU-cores and nearly 1,500 GPUs.

Looking forward, OAC will be addressing challenges to ensure that widely distributed infrastructure is available for large instruments, big data, big computing, across many scientific specializations while still ensuring workflow integrity, ease of use and adherence to regulatory requirements.

Federal guidance on Open Science and public access will be big issues as well as very valuable to society. There need to be ground rules that support all the teams. These are the challenges. How do you set up the rules so that all teams can work together and make things happen? How do you set up equitable

[2] The Kubernetes website is at: https://kubernetes.io/
[3] The Jupyter website is at: https://jupyter.org/
[4] The Ceph website is at: https://ceph.com/en/
[5] The EXPANSE website is at: https://www.sdsc.edu/services/hpc/expanse/
[6] The Voyager User Guide is at: https://sdsc.edu/support/user_guides/voyager.html

access? How do you support equitable access? What do you put in a data system? How do you hand off data to a data center? Basic principles will help with the commerce of data and data exchange.

There are barriers to the use of national resources as students who assist in setting up and running cyberinfrastructure graduate, limiting the expertise available to help scientists and researchers set up and use existing resources. The more people who have access, the more issues there are.

A future focus area is how to create career pathways for bright students to work in computing centers and facilities with scientific instruments. How do you acquire, train, and mentor CI professionals so that they choose a career in managing data and instruments? There are plans for mentoring and training where the mentors give back 20% of their time to the community to work on national issues.

Finally, democratizing science through cyberinfrastructure is a challenge we must meet. COVID made us realize that the work we have done prepared us for handling information quickly. Having a national computing capability only strengthens our ability to respond to whatever problems the future brings. One example is Ice-Cube[7], a highly flexible and usable system that positions us to respond effectively.

The wider the participation of the community and the ability to use data will make a difference. User friendly gateways and flexible access will help us reach the "Missing Millions."[8]

This keynote may be viewed at: [2023 4NRP - Amy Walton - YouTube]

3.2 **Frank Würthwein**, SDSC Director

NRP & the Path Forward: Progress in 2022-2023 and the next 10 years

The vision starts with democratizing access. This is not just a matter of fairness and justice. It is much deeper than that. If we want to remain competitive as a nation, we need to give equal access to all parts of the country and not completely exclude certain pockets. Beyond being the right thing to do, it is an economic driver.

Vision

The long-term vision is to create an open national cyberinfrastructure that allows federation of CI for everyone. We are adding Open Infrastructure that can federate data, computer, storage, data access, and anything that is bytes and flops capable, to Open Science, Open Data and Open Source. It should be open both horizontally (everyone can federate) and vertically (we can build new things on top of the infrastructure). Using a two-pronged approach, we can build communities of people who subscribe to the vision, working together, sharing resources, and building on each other. The new infrastructure must also have funding. This NRP community for the community will grow the NRP both vertically and horizontally.

Some people ask what the difference between NRP and OSG is, and why NRP is needed. NRP operates at all layers of the stack from the Intelligent Platform Management Interface (IPMI) up and is willing to run the Operating System (OS) for you, federate Kubernetes, federate at the Admiralty layer and then relies

[7] The Ice-Cube website is at: https://icecube.wisc.edu/

[8] The Minds We Need (https://mindsweneed.org)

on OSG to federate at the cluster layer. What results is a complementary implementation of "bring your own resource" model.

IPMI reduces the Total Cost of Ownership (TCO) and lowers the threshold to entry. This helps under-resourced institutions, network providers and their Points of Presence (PoP), Computer Science and Engineering faculty who all find it difficult to justify the required staff to support all layers. Kubernetes allows service deployments and a complementary vision of "bring your own resource." Many organizations do not want to or cannot support the staff required to run clusters, or storage systems since this is not their primary focus area. NRP does this on behalf of the community. NRP gives researchers the ability to share their resources with others and use resources of others. Since most users do not use their resources 24/7, other community members can use the shared resource.

Typical faculty support their research either by buying CPU cycles or by purchasing hardware and putting it under their desk. The first is not cost effective, and the second is only cost effective if you have at least 25% utilization. NRP gives these researchers the ability to integrate their resources into the national infrastructure, share those resources, and use other resources.

We had no intention of providing a global service deployment platform, but as we were deploying, we discovered that this platform was superb for running services on it. Services are packaged in containers; and containers are deployed by Kubernetes. If you have a Kubernetes deployment that is global, you can tell your collaborators to put their hardware into Nautilus, and then I can run the services I want you to provide. And then you have a global service platform. The two things that come to mind are the Open Science Data Federation (OSDF)[9] and Protein Data Bank (PDB).[10] NRP can support the PDB and replicate it in Asia and Europe.

Supporting Nautilus for the next decade

So how do we support this going forward? To build a community you must convince it that there is at least a ten-year runway. The NSF Cat-II program has a three-year testbed phase, a two-year allocation phase with an additional five-year renewal without re-competition if the system is successful. PNRP promises to build on PRP functionality and go beyond NSF Acceptance review of the system. The PNRP is planning to not only build out an open infrastructure, but also be able to bleed a little at the edges with architecture, systems, and engineering.

The PNRP proposed innovations:

1. Innovative network fabric that allows rack of hardware to behave like a single node connected via PCIe.
2. Innovative application libraries to expose field-programmable gate arrays (FPGAs) hardware to science applications at language constructs scientists understand (C, C++, rather than firmware).
3. A "Bring Your Own Resource" model that allows campuses nationwide to join their resources to the system.
4. Innovative scheduling to support urgent computing, including interactive via Juypter.

---

[9] The Open Science Data Federation website is at: https://osdf.osg-htc.org/

[10] The Worldwide Protein Data Bank website is at: https://www.wwpdb.org/

5. Innovative Data Infrastructure, including national scale Content Delivery System like YouTube for science.

Items 1 and 2 are totally new.

From the PRP we take the regional Ceph storage concept and from the OSG/PATh[11] we take the data origin and caching concepts and from that create a totally new feature: user-controlled replication of partial namespaces across regions. To build a scalable system that delivers to an infinite community, we built a matrix of science innovations that has three types of things in it: people, science driver themes. and NRP innovations. By building the community, the community to some extent can self-support. NRP uses matrix channels as a scalable user support mechanism backstopped by the professional at SDSC.

Wishlist for the future

- For the growth of NRP we grew the number of GPUs to over 1,000. Would like to grow to 50 PB of storage by the end of 2024, grow the community.
- Vertical things to add, machine learning at 100TB scale, expanding NRP into Wireless, edge and IoT, towards FAIR on OSDF.
- New directions initiative by the Community.
- New Data Origins – NSF CC* awarding nine campuses with $500K storage, 2023 another round. Some of these campuses may integrate their CC* storage with the OSDF as well as storage from other projects.

In conclusion, NRP has replaced PRP and has added and will add significant new capabilities: more GPUs, more FPGAs, more caches, and data volume served expected to grow substantially. NRP hopes to recruit new partners to build FAIR capabilities on top of OSDF within the next five years and expand into sensor networks using 5G and 6G in the next ten years.

Questions

(***UCSC planning to do storage as a campus service at UCSC, how do they integrate this with the vision of Open Storage***

(Frank) There is a simple way to do this. We have this notion of an origin server. The mental model is something like this. Imagine all your storage is behind a firewall. Now imagine you want to share only certain parts of your namespace and you do not want anybody to have access to other parts of the namespace. Now imagine you have a server that straddles that firewall and was deployed and accessible by Kubernetes in Nautilus. Then you give the PATH team the means to deploy the origin service such that it gets ingested into the OSDF system. Imagine you export out of your file system only out of the areas to which you want people to have access. You are guaranteed that even if Nautilus or PATH were to totally screw up, nothing would be exported other than what you exported to Nautilus. There are several options to do this (see video presentation for examples).

How do you see the interaction between NRP and FABRIC, Chameleon, and other testbed research?

[11] https://path-cc.io/

We are going to figure it out. The best bet is that we use a layer, like BRIGHT, that basically does a push button dynamic reinstall and then it can reinstall into the Kubernetes of NRP or into whatever testbed chosen. We will have to figure out how much pain we are in for to make this happen.

Question: ***There are a lot of data publication platforms for a wide variety of science domains that will not be easy to integrate directly into NRP. Have you thought about the external facing protocol set for ingesting data from external repositories?***

Everything is based on open protocols. The degree to which you can support open protocols allows you to figure out the ingestion and maybe even a federation without having to ingest. A first attempt may be to ask whether you can federate without having to ingest. Ingest means replication, replication means storage and storage costs money.

This presentation may be viewed at: [2023 4NRP - Frank Würthwein - YouTube]

3.3 Panel: Democratizing Access to CI and Democratizing Access to STEM via CI

3.3.1 Panel Chair: Lauren Michael, Minority Serving – Cyberinfrastructure Consortium (MS-CC)

The Missing Millions: Democratizing Computation and Data to Bridge Digital Divides and Increase Access to Science for Underrepresented Communities[12]

This session introduced in more detail a recent report "The Missing Millions" Democratizing Computation and Data to Bridge Digital Divides and Increase Access to Science for Underrepresented Communities referenced earlier. Lauren also provided some additional references to the National Science Board reports on this topic.[13]

The study showed that we are not reaching the numbers of women and minorities needed to provide diversity. A motivating question was how can NSF expand diversity and support development of new cohorts and communities of researchers to address pressing research, social and global issues in 2030?

There is urgency in addressing these matters given the broad applicability of computing and data to all fields and disciplines and the ever-increasing pace of change of technologies. The COVID-19 pandemic has brought into sharp focus the social disparities and digital divides that exist today.

Eighty-eight individuals across fifteen focus groups and six interviews found that there are still substantial barriers to access. Accessibility also includes the ability to use technology. There is insufficient engagement with underrepresented institutions. It is not just computation, but software and data that need investment. More diverse fields and disciplines need support to address access gaps. Facilitation successes need expansion so that they scale. CI capabilities are expanding as an important

[12] https://www.rti.org/publication/missing-millions/fulltext.pdf
[13] https://www.nsf.gov/pubs/2020/nsb20222/nsb20222.pdf and Diversity and STEM: Women, Minorities, and Persons with Disabilities 2023 | NSF - National Science Foundation

way to bring CI to communities rather than waiting for the communities to come to CI. Systemic change is needed and requires deliberate thought.

Vision of success was organized into 12 different paths:

1. Ensure inclusive access.
2. Experiments in NSF operations.
3. Elevate data and software to match computing investments.
4. Engage in long-term, inclusive community building.
5. Expand the scope of funded research.
6. Enable distributed edge technology.
7. Ensure racial, gender, and other forms of inclusion.
8. Accelerate broader impacts.
9. Advance social and environmental sustainability.
10. Bridge across directorates and federal agencies.
11. Foster consortia and partnerships.
12. Engage NSF leadership.

The Missing Millions report is a call to action and there are several initiatives in response to some of the recommendations.

This presentation may be viewed at: [2023 4NRP - Lauren Michael - YouTube]

3.3.2 Speaker/Panelist: **Manish Parashar** – National Science Foundation Democratizing Science through Cyberinfrastructure

OAC aims to support CI for enabling innovation, science and engineering research and education. OAC looks at CI broadly: big and little compute, networking, testbeds, services, and importantly, people. They are trying very hard to make sure that all the infrastructure pieces work together towards enabling science. Think of this as an ecosystem. Science moves forward, technology evolves, needs change so infrastructure cannot remain static.

OAC invests in an ecosystem of services that has been broadened to include capacity systems, and experimental systems. It is good to have these capabilities, but how do you make sure that everyone benefits from this ecosystem?

Broad, fair, and equitable access to advanced CI is essential to democratizing science in the 21st century. But significant knowledge, technical, and social barriers exist. People need to be aware of what resources are available and how to use them. They need to know how to connect with the right software to access data. They also need to be aware of the benefits of CI and how to obtain grants and funding to provide this access. There are complex tradeoffs.

To meet the growing need for increased CI capabilities and capacity we need to have the right building blocks. To implement a scalable federated ecosystem of diverse capabilities and capacity we will need to leverage institutional and commercial investments and incentivize partnerships.

Once built, we need to make sure everyone can access it and not just open access, but equitable access. Then we can move from open/equal access to democratized/equitable access. This vision is not unique to NSF. A memo came out this past summer about making data or the results of federally funded research open and equitable and accessible without any embargo[14].

Next steps:

1. Access: Collectively ensure, and advanced accessibility of data and shared CI needed across science.
2. Data: Collectively enable and accelerate compute- and data-intensive science across NSF, including scaling up for facilities and mid-scale RI needs.
3. People: Collectively develop and sustain a capable workforce of CI professionals who are key to science.

This presentation may be viewed at: [2023 4NRP - Manish Parashar - YouTube]

3.3.3 Speaker/Panelist: **Matthew Rantanen** – Southern California Tribal Chairmen's Association

Presenting a Tribal perspective, Matt discussed joining a project about 21 years ago in 2001 that was created when Russ Frank, an ethnics professor at UCSD, wrote a grant proposal with the Tribal Chairman's Association[15] to get funding from Hewlett Packard (HP) to get access to communications technology. They realized that this access was essential for the future of their children, education, and their community centers and libraries. Matt got into policy and advocacy and went to Washington, DC to find out why tribes were not getting access to technology and opportunities such as funding, spectrum, and subsidies to which everyone else had access.

The Einstein wizard types of minds could very well be situated on a reservation, but without access and opportunities, we may never hear about them. We may miss a very good opportunity to get the perspective of someone from a First Nation, with a different history and way of thinking.

The Tribal Digital Village network connected here in California includes 17 reservations, 13 libraries, 550 homes, 105 tribal municipalities, and 650 miles of microwave. They have been flirting with organizations that utilize NSF from the beginning. Hans-Werner Braun mentored them and taught them how to build a wireless network, and Michael Peralta, how to set up wireless on a mountaintop. CENIC, which is the headend of network now, gives them affordable access to the Internet and the potential to be a part of this community, bringing it back to the reservation. They are still trying to figure out how to tap into that.

At the end of the Obama administration in 2016, there were 8,000 missing middle connector miles, 320 in lower 48 states. There are 574 tribal nations with about half of them in Alaska, so that is about 320 in the lower 48 states. A ballpark estimation of what it would cost to get to each reservation was a seven billion problem in 2016, which would be a ten billion problem today. The Federal government has set aside three billion dollars for this, which is a good foundation start, but not nearly enough. Tribes are in the learning curve now, trying to figure out how to get things up overnight. It is hard to manage the funding that is coming in now and know who to trust. There are a lot of sharks in the broadband

[14] 08-2022-OSTP-Public-Access-Memo.pdf (whitehouse.gov)

[15] https://sctca.net/

industry and tribes get preyed upon. Relationships with tribal governments could be influential now. Information flow from the education community is important now.

So far, every attempt of the federal government to get the incumbent service providers to bring services to Indian country have failed. Tribes must figure out how to manage their own communications: how to provide electricity, how to build houses, how to manage water and waste, and now how to build out digital infrastructure. Until tribes get connected, they cannot participate. When working on the development of cyberinfrastructure of today, tribal minds need to be included and involved. They are very important to the future. Native Americans are some of the heaviest users of internet once they get access. They are geographically isolated from other resources which makes internet connectivity even more important. There are opportunities to include and integrate some of those great minds that are out there.

3.3.4 Panel Discussion

Question 1: NSF's investments have been based through high-ed and national labs – how can institutions of higher ed get involved?

(Matt) There is an opportunity for different funding sources such as FCC, NTIA, Commerce, USDA to provide funding for broadband tied to science projects. There are silos at the scientific level of funding. Different funding that supports broadband and infrastructure for Native American communities can also be tied to science and higher-education projects if you break down the silos. A lot of education projects could be integrated into a CI piece. Tribal communities manage everything themselves as sovereign nations. After-school resource programs such as the Pala reservation off highway 76 biology projects, water management projects, riverbed exploration and testing which can be integrated into the CI piece. Dragging these projects into technology beyond the hands-on collection of data, but also how to capture and collect data and analyze the data over time.

(Manish) NSF works through the community, bringing the best ideas for engagement forward. NSF has flexible mechanisms to facilitate that. An example of this is the Civic Innovation Challenge.[16] Instead of coming up with a great idea and then finding an application, this is the other way around. Come up with an important problem and bring a team together to solve it. The Convergence Accelerator[17] is another example where it is not just academic research but translating it into real problem solving. Regional Innovation Engines[18] addresses problems that are important to society regionally, leverages talent to solve those problems, and provides feedback results to address problems.

(Jen) If you lead with technology, you attract only a small subset of interested students in CI. If you lead with problems that have a social impact and address issues within underrepresented communities, you get more women and underrepresented minorities showing up for those types of training opportunities.

(Matt) Tribal communities want to be a part of the solution. They want to be the tech wizards and not rely on others to solve their problems. This creates a new ecosystem. Most children on the reservations have a phone or tablet, but learning is still being taught in a traditional way. Education has not embraced technology. If you do not figure out how to teach students their language using an app that is

---

[16] https://nsfcivicinnovation.org/

[17] https://new.nsf.gov/funding/initiatives/convergence-accelerator

[18] https://new.nsf.gov/funding/initiatives/regional-innovation-engines

fun, you are not going to have students learning their language in the future. If you can integrate the technology in a creative way, you will catch most of the students on the reservation.

Question 2: Given that we've got a panel tomorrow on Workforce Development for underrepresented minorities and workforce relevant to the CI field, what are some of the requirements you have encountered for workforce development to achieve cyberinfrastructure enhancements? What strategies are being applied to address these?

(Matt) Workforce development is a problem and was amplified during the pandemic. There is a lack of trusted partners. During COVID, there were many requests for broadband training. Reservations had closed borders. A Tribal Broadband Bootcamp was held in July 2021, the first of six. Multiple tribal representatives who know what they need to do, but not how to go about it attended. There were 64 people and 17 reservations at one time. The result is the creation of a human network, breaking down silos and creating a peer group which is now over 200 people and includes 35 different tribes.

(Manish) We need to build content for training, be very responsive to local, and societal needs. There has to be a continuous process of mentoring of professional development. There needs to be a community, not just instruction, but peer engagement where you are learning from your peers and are able to work with as part of the community.

Matthew's presentation and the panel discussion may be viewed at: [2023 4NRP - Matthew Rantanen - YouTube]

3.4 Panel: Open Infrastructure for an Open Society: OSG, Commercial Clouds, and Bring-Your-Own-Resources

3.4.1 Panel Chair: **James Deaton**, Executive Director Great Plains Network

Great Plains network is a consortium of six state networks and 20 plus research universities across the west. Great Plains Network was excited when PRP was announced and internally funded a bunch of nodes across the campuses. They put one node at each campus and challenged the state to provide additional funding. Since then, they have continued finding opportunities to expand the conversation to the CIOs and help people understand the value of the Open Science Grid and NRP resources.

This introduction to the panel may be viewed at: [2023 4NRP - James Deaton - YouTube]

3.4.2 Speaker/Panelist: **Derek Weitzel**, Research Assistant Professor, University of Nebraska-Lincoln, OSG, PATh, PNRP
Open Infrastructure

Derek Weitzel's presentation on "*Open Infrastructure*" had a quick overview of some of the projects and showed a US view of NRP. All components on the NRP are Open Source such as Kubernetes and containers. Anyone can contribute resources. Anyone can use resources. NRP has documented interfaces. Resources were seeded through grants but have grown over time.

OSG is also open infrastructure. All components are open source, such as HTCondor[19] and other various tools. Anyone can contribute. Anyone can utilize these resources. Interfaces are documented. Resources

[19] https://htcondor.org/

are seeded by large organizations such as LHC[20] and now CC* but as with NRP have grown through the contributions of users.

The Open Science Data Federation (OSDF) operates completely on open protocols. All components are open source. Anyone can contribute resources. Interfaces are documented. Resources were seeded by various grants and Internet2 by have grown organically by contributions from users and soon CC*. Many sites have installed caches just to limit the number of connections to the wide area networks. Syracuse is a great example of doing that.

This presentation may be viewed at: [2023 4NRP - Derek Weitzel - YouTube]

3.4.3 Speaker/Panelist: **Jeremy Evert**, Associate Professor, Computer Science, Southwestern Oklahoma State University

Leveraging NRP on a smaller campus

A four-university collaboration for a CC* grant, including Southwestern Oklahoma State University (SWOSU), took a DTN, a switch and a perfSONAR[21] node to build up what they already had. This summer they hope to connect to the Open Science Grid (OSG).

SWOSU is 10th in the state in enrollment behind 2 community colleges with 5,000 students across two campuses. They have a 200 square foot server closet. In terms of “bringing your own resource,” joining NRP as a contributor is harder than standing up a perfSONAR node, but not as hard as standing up a switch. Dima Mishin at UCSD helped them set up, James Deaton stepped up and suggested that they enable user authentication through github, and OneNet and SWOSU central IT provided the alias that allow them easy access to NRP.

Locally they engage and empower every SWOSU student by assigning them to join GitHub.com/swosu. Students are pointed to their server as soon as they start running codes. The students stay engaged. SWOSU invites area technology teachers for a weeklong camp, luring them in with esports, graphic design, Microsoft, and programming. They spend a full day teaching programming and have teachers run jobs on jupyter.swosu.edu. The aim is to engage and empower every elementary and high school student and researcher.

They plan for the next 10 years to enable more science drivers, partner with the education department to integrate more of the Campus Champions, and leverage mentors from NRP, the Great Plains Network, OneNet, and the OneOklahoma Cyber Infrastructure Initiative to keep growing.

Jeremy recommends setting up a weekly statewide call, setting up a mailing list, and encouraging key players to join. This allows staff to show up and make connections and looks for ways to add value to the individual and community. It does enable people at smaller institutions to know what is happening at the national level and helps them tie into the resources that might be available to them.

This presentation may be viewed at: [2023 4NRP - Jeremy Evert - YouTube]

---

[20] https://home.cern/science/accelerators/large-hadron-collider

[21] https://www.perfsonar.net/

3.4.4 Speaker/Panelist: Igor Sfiligoi, Lead Scientific Software Developer and Researcher, San Diego Supercomputer Center
Open Infrastructure for an Open Society: Commercial Clouds

Igor Sfiligoi's talk, "*Open Infrastructure for an Open Society: Commercial Clouds"* provided the pros and cons of using commercial clouds for research. With all the investments that NSF is making for the on-premise hardware, what would be the reasons to look at commercial clouds?

It seems like everyone in industry is moving to the cloud today. One reason to look at commercial clouds is that they have huge compute capacity. Igor has personally accessed 50K GPUs in the cloud. They have a large variety of compute resources, many GPU variants, all accelerators, FPGAs, and often have hardware available before you can buy it. They also have some cloud-exclusive hardware variants such as ARM CPUs, AI accelerators and FPGAs. The cloud is alluring. Commercial cloud providers could easily get 1Tbps four years ago. They have high-end stuff.

What are the pros and cons for commercial clouds? All you need is your credit card, but you need money, lots of it. Regular on-demand cloud computers are expensive. They are anywhere between 3 to 10 times more expensive than doing it on-prem. It is easy to get in, hard to get out. Pricing is optimized to get data in cheaply, but expensive to move out. There are no automatic price caps.

Who should consider commercial cloud? For flexible urgent computing, costs are acceptable for short spikes. For prototyping R&D there is instant access and no-contention which drastically raises productivity. If you need ultra-high availability services, commercial clouds have a proven track record with hundreds of data centers and cloud deployments.

Are commercial clouds easy to use? Yes and no. They do provide flexibility but can be daunting for non-IT users. There are lots of support services, but finding what you need can be a challenge. Cloud providers invest a lot into user interfaces, but each provider has their own flavor.

How do you mix on-prem and cloud resources? CloudBank does a good job with account management, spend and budget tracking and has extensive documentation and training. But it is not for the faint of heart. One of the easiest options is to integrate with the OSG/PaTh/HTCondor ecosystem. Another option is going the Kubernetes route. Cloud K8s feel like on-prem K8s and a Kubernetes federation can make it completely transparent.

3.4.5 Panel Discussion

(Inder) In the cloud, the moment you run out of money, you run out of resources. If you buy the resources, you can use them for some time, donate them, etc., which you cannot do if you are renting.

(Tom DeFanti) Universities typically charge overhead on cloud services, but do not charge on equipment purchased. This can make using the cloud very expensive for researchers. University of Washington and UCSD do not charge overhead on cloud.

(James Deaton) The last two bullets from Igor's presentation are exciting. When you do have the on-fire issue, the open infrastructure button makes it easier to scale up.

(Eli Dart) ***For a smaller institution, how do they attach to the national facility easily?*** *Understanding what it takes to connect is of interest.*

(Jeremy) One of the things that made it easy is that SWOSU is in a state that has a lot of collaboration that was already taking place. Connecting national resources to the end user in the future will need to be available via phone or tablet.

(James Deaton) Jeremy has been an early adopter in engaging his students. In doing so he has run into a lot of friction points. PNRP has lowered the barrier to entry. The support program within PNRP to onboard nodes could be completely turnkey. The educational aspects are still there. The exciting thing is having some layer of involvement so that there is a discovery process for how the underlying system works. CIOs and specially CISOs want to see the weeds.

(Jeremy) When Frank and Lauren were talking about getting resources, it is not enough to provide a national resource, and it is not necessarily enough that you provide this lens for how it can be used. You need to provide training in an accessible way. Free is good incentivized is better.

(James Deaton) When SWOSU had a catastrophic failure of an air conditioning system, they had to power down their system, but because they belonged to the national infrastructure, students could still get to resources.

(James) ***What kind of benefits do you see from contributing resources on-prem to these larger efforts, whether it is the Open Science Grid or to NRP?***

(Derek) The University of Nebraska has long been contributor to national cyberinfrastructure. We see the advantage of providing resources to these national organizations in two ways. One, we have users on campus that use these resources such as OSG. It makes sense to provide our own cluster and make it available on an opportunistic basis to the OSG. One of our graduate students finished her PhD, by performing work on the OSG. When she graduated, she had so much experience using the system, she was hired as a research facilitator at UNL. Two, by donating resources, we get a lot of different use cases that run on our clusters. The systems administrators learn more about the use cases, get to learn a lot of different scenarios, and this has led to locking down users for OSG and has helped with on-prem resources. Users cannot hurt others when competing for resources.

(Jeremy) SWOSU federated a $40,000 server to NRP, adding to shared resources and got better adoption with users since it was “trusted”.

(JamesD) ***What improvements need to happen to open-source infrastructure to scale up? Users need assistance in using the commercial cloud resources.***

(Igor) Cloud is not just about scale, it is also breadth of the kind of resources you can have. If you need the same resources that we already have in our cluster, it is a no-brainer. If you want to push the limits, you must be more careful. Now if you want to use exotic resources it is a bit more of a problem. You must help users use those resources.

(Derek and Igor) A barrier to use commercial clouds is money, and experience. IceCube does not care what you are doing, you can add to resources. A recommendation is do not send anyone to cloud unless they know what they are doing.

(Igor) If you can keep your machines busy 24-7, add on-prem costs including all electricity, all the sys admin, building costs, and compare to cloud costs, it is 3x to 10x more expensive in the cloud. Commercial clouds are there to make money. They do not run at 100%, that is their business model.

(Tom) There are two models that universities have: servers are set up so that only students can use them, the other model is to put them into Nautilus and then be able to use the resources.

(James Deaton) There are a lot of network researchers in this community and every one of these boxes at NRP has a subset of performance analysis tests. This vantage point has been inspirational to see the impact when you see a small campus hosting one of these boxes without having to go through the headaches of using a perfSONAR node. There are hundreds of nodes spread out across all these institutions and NRP has tools to see what is happening in the network, analyze traffic patterns, and take advantage of those spare CPU cycles.

Igor's presentation and the panel discussion may be viewed at: 2023 4NRP - Igor Sfiligoi - YouTube

3.5 **Ramesh Rao**, Qualcomm Institute Director

Welcome to Qualcomm Institute (QI)

Ramesh Rao gave a few remarks on the theme of infrastructure. Going back to 2000 when the proposal for CalIT2 was written, the boldest thing UCSD did was to make a case for the creation of new infrastructure. It was one of the best decisions the team made that holds up twenty-five years later. The CalIT2 building is still exquisite.

In 2000, the world of wireless and wireline were far apart in terms of data rates. Over many years the two have come together. There are many more opportunities for exploring what happens at that interface. That is the topic of the next panel, taking advantage of the NRP platform, extending it farther with wireless technologies.

5G, 6G and beyond have created new opportunities and challenges. Here at CalIT2, we are in the process of creating a metaverse signal observatory, a place where you can study electromagnetic propagation at the same time as you work on the creative content. One such application is to extend surgery to remote areas. On campus there is an increase in the student population, and it is the perfect opportunity to create new programs and align research activities to the experiential learning opportunities for students.

CalIt2 was initially a partnership with UC Irvine, with just a point-to-point link, but in recent months it has been extended to UC Riverside. Taking advantage of the geography it is positioned now for even bolder experiments. The most tangible one is Alert California[22] which takes advantage of the 1,000-camera network on mountaintops gathering hyperspectral imaging data to anticipate outbreak of fire and be able to fight it.

[22] The website for ALERTCalifornia is: https://alertcalifornia.org/

Al Pisano has just joined UCSD from UC Berkeley. He understands the opportunities and landscape and is steering new directions in which optical fiber technologies combined with wireless can take us.

You may view Ramesh Rao's talk at: 2023 4NRP - Ramesh Rao - YouTube

3.6 Panel: Future Wireless Extensions of Regional Optical Networks

3.6.1 Panel Chair: **Albert P. Pisano**, Distinguished Professor and Dean, Jacobs School of Engineering, UC San Diego

What if there were a national research innovation platform that could immediately span the country and the eastern hemisphere, that could stand up to the challenge of helping the United States and all its allies to start working on the next generation of wireless communication? Why would a group that moves large data sets around be interested in wireless? If you think about what happened between 2G and 3G, data became an important part of what gets moved, not just voice. And the system did not work, until the optical fiber community put in enough fiber to connect all those cellphone towers.

This panel is run as an implicit pitch. What would you think about finding some precious bandwidth and attention to allow the country to immediately stand up a research platform that is not dominated by one vendor or another, a platform that allows a wide variety of people to make experiments that lead to future wireless communication research? What if we could be that platform on which this grand new venture is prosecuted?

You may view Al's talk at: 2023 4NRP - Albert P. Pisano - YouTube

3.6.2 Speaker/Panelist: **Larry Smarr**, Distinguished Professor Emeritus, Computer Science and Engineering, UC San Diego

Al has spent time on this nationally because he sees correctly that there is an inevitability about wireless extensions of the optical backplane being what will be coming in the next 5-10 years. Following Ramesh Rao's comments, for twenty years he and Larry have been thinking about how to put together wireless and infrastructure. NSF-funded Hans Werner Braun 20 years ago to fund the High-Performance Wireless Research and Education Network here in California. All of this was off the CENIC optical backplane, so it was logical to use that backplane to connect the data servers to which all the continuous wireless are connected.

George Guilder, who was the prophet of this, would talk about the world of fiber wrapped in glass and then the fuzz of wireless stuff coming off this glass. This vision is what led to the creation of Optiputer.

When you are on a transformational theme that is inevitable, you cannot lose. It does not get any better than that. That is what this panel is about.

You may view this presentation at: 2023 4NRP - Larry Smarr Part 1 - YouTube

3.6.3 Speaker/Panelist: **Sujit Dey** – Professor, Electrical and Computer Engineering, and Director, Center for Wireless Communications, UC San Diego
Future Wireless Extensions of Regional Optical Networks

As cellular technology evolved, 3G brought about real internet browsing, 4G brought about video and supported social media. We have been waiting for 5G for some time. Before 5G, improvements in bandwidth and latency were key. 5G paid attention to concepts such as network virtualization, SDN, and slicing, so it could detect what flows need what kind of treatment. Slowly the 5G deployments will get to a point where the network will figure out what kind of treatment is needed and recognize that some applications need quality of service. The key 5G technologies are massive MIMO (up to 128 antennas), mmWave (24-60 GHz) in use indoors, work being done for outdoors, Edge Computing, Small Cell Densification, SDN/NFV Slicing, new radio, and C-RAN.

When we go from 5G to 6G or NextGen, we see quite a few things. One is that some interesting spectrum bands are opening up. Beyond the mmWave band there is a THz or sub-THz band that is opening up. There are challenges still to be resolved. The key 6G technologies are extreme MIMO (up to 1,024 antennas), sub-THz, edge intelligence, macros go away, IRS and wireless sensing, energy efficiency, and the use of renewables. The target QoS evolutions are 10x higher rates with 1Tbps Peak, 1Gbps average, 10x less latency. The NextG network should be the sensor, computer, and AI engine, and yes, also a communication pipe!

So how does this impact autonomous mobility, smart manufacturing, agriculture, metaverse, virtual physical systems, digital healthcare and so forth? How do we bring together these revolutions? It is impossible to achieve NextG goals without an ultra-high throughput, ultra-low-latency optical backhaul network. There are amazing computing capabilities at the edges. The networks have CPUs, GPUs and so on, with much higher capacity than the current cellular networks. NextG needs combined capabilities of cellular and optical networks. If optical and wireless can come together it becomes a SUPER-Edge.

Some of the optical networks have fixed wireless today, but if we can add mobile wireless extensions, we can enable different kinds of applications such as forest and wildfire monitoring and containment using drones. It is still a tall order, and a lot of work still needs to be done there. Similarly, what about all the tractors and sensors needed for smart agriculture? An optical network connecting with them would be tremendously useful for data collection. Optical networks can benefit from additional sensing capabilities provided by RF link extensions and cellular edge intelligence capabilities.

This national research infrastructure with wireless extensions will enable impactful research and development by both optical networking and cellular communities and will lead to true next generation network.

This presentation may be viewed at: [2023 4NRP - Sujit Dey - YouTube]

3.6.4 Speaker/Panelist: **Robert Kwon**, Associate Vice President of Engineering, CENIC
CENIC – NRP Panel on Future Wireless

CENIC is the research and education network (REN) for California and provides connectivity for public internet and research education ecosystem in the state through Internet2, Pacific Wave and other entities. Over 20,000,000 Californians use the CENIC network that has over 8,000 miles of fiber throughout the state of California and 12,000 connected sites from diverse organizations.

CENIC's membership is diverse, connecting everyone from libraries and cultural institutions to R1 universities. Since its inception, it has supported research and research initiatives such as the Pacific Research Platform. Conceived in 2014 to connect the science DMZs together, CENIC provided the backbone for the PRP.

Another project supported by CENIC's backbone is the High-Performance Wireless Research and Education Network (HPWHREN). Started in 2000 with NSF funding, it enables hundreds of cameras and sensors to collect data on wind direction, air temperature, precipitation, and other data and then stream that data to servers across the CENIC network. Users can connect and stream the data reliably.

Another project that leverages the CENIC network is WIFIRE[23]. It takes in data from multiple sources so the HPWHREN real-time meteorological sensors can feed the weather forecasts, landscape, and fire perimeter data to PRP and run certain workflows which output WIFIRE fire map data. This fire map data is critically important for policy makers and first responders, giving them the ability to make real-time decisions in the management of wildfire responses.

From 2016-2000 CENIC was able to upgrade data servers and extend locations and data replications sites to UCSD, SDSU, UCI and UCR improving data redundancy, disaster recovery, and high availability for HPWHREN and WIFIRE. During wildfires there is a 10x increase in traffic as the public looks for real-time information.

Many locations within California are remote and have limited internet access. Fifteen percent of California's population lives in the 60% rural areas. To reach these areas it is sometimes very costly and not generally feasible to connect them. CENIC has connected 81 sites via wireless. An example of this is Yosemite Park Elementary which is within a National Park.

There is a potential to scale fixed wireless using CENIC optical fiber backbone in California. A proposed CALFIRE[24] multi-hazard network would aid first responders to identify, analyze, and communicate during disasters statewide.

This presentation may be viewed at: 2023 4NRP - Robert Kwon - YouTube

3.6.5 Speaker/Panelist: Andrew Wiedlea, Computer Systems Engineer, Science Engagement Group, ESnet
Wireless FasterData and Distributed Open Compute

ESnet's job is to make sure scientists can do what they need to do, seamlessly and as easily as possible. ESnet6 currently spans the continent and an ocean supporting all the laboratories and facilities under the US Department of Energy. But there is a lot of science that is happening outside of user facilities and the laboratories. ESnet is developing a strategy to support the wireless edge. It is a work in progress as there is a wide range of possibilities. They are developing an understanding of where they would like to be in this space and what type of products and physical capabilities they should provide.

---

[23] https://wifire.ucsd.edu/
[24] https://www.fire.ca.gov/

Science use cases in the short term include the convenience of IoT and the ability to deploy instruments in non-lab settings, hostile or constrained environments, and easily within facilities. Many things are happening with non-terrestrial networks, including different millimeter waves, LoRa, Zigbee, free space lasers, etc. Scientists want their connectivity to be more mobile and flexible.

A longer-term use case driver is to develop self-driving, self-guiding field laboratories and an automation revolution for scientific activity. Use cases for this are developing in many disciplines such as Earth and Environmental sciences, Astronomy, and Physics.

The basic idea is that you have the combination of the ability to deploy autonomous and mobile sensors in an environment and to manage the reduction of uncertainty and sample planning more intelligently. Sensors will be able to move themselves to attain the next most piece of valuable data.

How do they do that? Well, they must be in comms and that is going to be wireless with some fast-running edge compute capability that helps them determine where they need to be and what they need to be doing, which in turn is developed into a workplan. This needs to be supported on the backend by ties to leadership class compute facilities and other HPC resources to high fidelity digital twin and other types of phenomenology modeling. And this all must be tied and integrated together through an enriched network capable of handling the distributed storage, compute, data management, workflow logic, uncertainty reduction activities that are necessary to make this work as a seamless system. If you could package this up, you could sell this to a whole lot of different communities. It involves a set of skills that are unique to the research and education community. It would be very difficult for other communities like commercial clouds to focus on creating this. There is a lot of hand skill, coordination, art, craft and understanding of science that go into creating this to work for different domains. Many of the capabilities needed may be provided by commercial cellular, so there is room for everyone to participate in the development.

Some examples of enriched network uses are:

1. A company called Saildrone[25] is sending intelligent drones across the oceans.
2. There are meteorological balloons developing high fidelity models of the stratosphere above the North American continent.
3. Lawrence Berkeley National Laboratory (LBL) has a project named EcoPOD[26] that is trying to understand how plants and plant metabolisms develop and are shaped by their environment and their genes. combination of soil, and atmosphere, etc. There are controlled fields at UC Davis where the project grows crops and collects data that is moved wirelessly to bio chambers. These bio chambers are used to simulate the exact combination of light, soil, nutrients, and other environmental factors to explore how crops respond to things like changes in climate.
4. In France, SOLEIL[27] is developing self-driving experiments on their beamlines in some of their facilities.
5. Colleagues at USC's Microwave Systems, Sensors, and Imaging Lab (MiXIL[28]) are working hard to autonomously allow drones to develop collection plans for sampling schema.

[25] Saildrone's website is at https://www.saildrone.com/
[26] EcoPOD's website is at https://ecopods.lbl.gov/
[27] SOLEIL's website is at https://www.synchrotron-soleil.fr/en/beamlines
[28] USC's MiXIL website is at: https://mixil.usc.edu/

The bottom line is the idea that we create a national federated wireless capability, not just 5G, but also ways for scientists to deploy LoRa and other mesh network capabilities particularly for low energy situations where there is insufficient power to run a full 5G. This should be integrated into edge compute with the ability of the RENs to share the data and coordinate the workflows and logic.

We do not know all the answers yet, since there is so much happening in this space. We need to be able to test things out and break things.

Some examples of use cases are:

1. Soil Atmospheric Integrated Laboratory (SOIL) in Colorado are working to develop an integrated view from space to bedrock of what is happening in the Rockies. There are some hardware challenges due to the weather and terrain. It would be great to put a FIONA there, but is there an affordable ruggedized FIONA for edge compute in the field that would satisfy power and mesh communication needs?
2. Smart Soils Testbed[29] at LBL is exploring self-driving components.
3. The EQSIM project's [30] earthquake simulation package, using sets of diode systems in downtown Oakland and Bay Area buildings, could use wireless to get data out of the building quickly to provide rapid estimate of area damage.
4. Urban and Radiation mapping is another area where you want self-driving capabilities. Urban radiation effects are a complex combination of soil and building materials, weather and atmosphere and human behaviors. The detection of anomalies is short-range so being able to quickly deploy drones to take measurements would help build a model of what is happening with the background radiation environment and more easily pick up on anomalies.
5. Work is underway to build an energy simulation environment that spans lab boundaries. Right now, this is all optical, but it would be advantageous to bring in wireless sensors and tie them into management models of new generation power, solar, and wind plants.
6. As a country the US is planning on deploying large offshore wind turbine farms. The larger the turbine blade, the more efficient it is. But bigger wind turbines need to be located further offshore, and that makes the comms problem harder. This might be an area where science practitioners would love to have a self-driving component.

Desired End State Thoughts

The development of capabilities is shaped by what customers want and what they think the network can deliver. That is one reason why it is important to co-design infrastructure and embed network professionals with scientific practitioners so that the shape of technology and scientific possibility expand together. It would be great to just hand a person a SIM card and it just works anywhere in the country. Now is a great time to take risks.

Let's think about PRP in Space!

This presentation may be viewed at: 2023 4NRP - Andrew Wiedlea - YouTube

[29] Smartsoils Testbed - Smartsoils Testbed (lbl.gov)

[30] EQSIM - Exascale Computing Project (exascaleproject.org)

3.6.6 Panel Discussion

(Al) ***Sujit, what are the special challenges of mobile vs. fixed?***

(Sujit) If the device is not moving, it becomes like a fixed wireless device, however, if a device is moving with that mobility comes all kinds of challenges including fading, speed, and environment.

(Al) –***Robert, Andrew can you think of any examples in which reconfigurability is currently happening on your networks today? Rerouting of dataflows through different channels based on relocating sensors.***

(Andrew) – It becomes more complicated when you add in mobile wireless that could be a mixture of commercial and your own pathways, across multiple bands and multiple channels. You need something that can manage the workflow.

(Robert) – From CENIC's perspective, we provide the backhaul for any connections that come into the network. With advancements in software-defined networking and telemetry. CENIC does review how it takes in information coming from routers/switches to steer traffic to an optimal path.

(Al) - ***Suppose we wanted to open a platform where people from amateur to expert could practice algorithms for redistributing their hardware assets. Some companies in the United States write as many patent applications as possible, and without ever interoperating or figuring out if it is coherent or not, they randomly collide into each other at the regulation bodies when the standards are set. Wouldn't it be good to have neutral ground to test out interoperability? What if we develop an open innovation platform?***

(Sujit) - Interoperability has always been a problem. Remember WiMAX – there was a lot of innovation, but it was killed due to interoperability issues. There is a movement in the commercial world happening. The Open Network Alliance shows we are moving in the right direction. To answer the question, yes, this is needed.

(Robert) – In the networking space we have learned to interoperate, so this issue is not as prevalent, but to support something like an open innovation platform, CENIC would need to make more strides into its automation for the network to be programmable from the edges.

(Andrew) – There is an innovation band that was created for this type of activity, the 3.5 to 3.75 gigahertz band. We should stay in our box and our band and demonstrate we can produce an interoperable system. Spectrum is worth a lot. There is no way our group is going to evangelize open source and open sharing and the kind of values we want to see in this space until we demonstrate that it works within that band and is desirable from a customer's standpoint.

(Glen RIcart, USIgnite) ***People have not been able to find spectrum to do this. Pal licenses taking up 3.5 Ghz band. With this grand vision, do we have the spectrum to do this?***

(Andrew) – At least with PAL licenses it is at least imaginable that you could afford to get in that space, but you need to fight for it.

(Sujit) – The recent success of Citizen's Broadband Radio Service (CBRS) and creating these private networks bodes well and gives hope that we can use the spectrum sharing and partner with those in the DoD and other communities to see what could be done.

(Andrew) – This may be an opportunity to partner with the tribal community.

(Glen) I believe this is an area where things can be done. Future research could open up more spectrum. Spectrum enlarging technologies could help in this area.

(Statement)– In terms of 3.5 Ghz, the ham radio community has demonstrated a couple hundred megabytes of point-to-point connectivity shown for close to ten years in Southern California. We can build upon this to do better things. Getting a building on a mountain top is very difficult, can be an obstacle to deploying. 3.5 Ghz was taken away from Ham radio by the Department of Transportation to be given away to develop autonomous vehicles.

(Andrew) – You can imagine deploying a sensor system and then visiting it once a month with a drone that has an eNodeB on it, flying over and doing the data collection. There are a lot of things happening to make it much more feasible to deploy sensors in weird parts of the world.

(Sujit) – We have seen this battle in the cellular world also. In the cellular world, the deployments are encroaching more and more on neighborhoods and city streets. There are also a lot of misconceptions around the deployments due partially to the lack of systematic studies. Universities can play a much more active role with social scientists and medical professionals to measure and analyze the impacts both negative and positive.

(Frank) - ***All the applications mentioned are where information is carried away. Can you think of an example where the optical network is in the middle?***

(Sujit) Tele-robotics and telemedicine are examples where it communicates with wireless to fiber, carries the data to another location, and then acts upon it at the other end.

(Frank) ***– So this in a sense in relatively small spaces where there is a robotics thing happening on one side and it communicates by wireless because it is mobile, and it communicates to a base station of some sort, fiber carries traffic to another place and stimulates some kind of activity on the other side.***

(Sujit) – So the application is, there is an expert on one side who is giving instructions and then those instructions are carried over the network, the wireless link, and the optical network. On the other end the robot is acting according to instruction on a patient and then the feedback is coming all the way to the other end.

(Andrew) – To add to that, the whole self-guiding field laboratory vision is to have distributed compute in the network, on the edge, backhauling the high-fidelity phenomenology codes to leadership class facilities.

(Frank) ***So playing back this example, there is a sensor in the environment that collects information that needs to be processed in ways where there is something at the edge, something in the middle, something at the data center and then a decision is made that informs the next movement.***

(Scotty) – You can have widely distributed sensor systems where you want to change the fidelity of the observations at the other end based on some disturbance somewhere else. That requires immediate processing, priority and capacity aware network automation, and delivery of a data product to systems that are changed automatically.

(Ramesh) – First an observation, optical fiber can be a sensor itself, not just a transfer medium. We can rethink the way we partition these things. ***When we open up these interfaces, on the one hand we are demonstrating technical feasibility, on the other hand there are different business interests at play. Maybe we need to model competing business interests better, and national security interests, and not oversimplify to just technology issues. Do you think our standards mechanisms capture this?***

(Andrew) – The future is inherently socio-technical, so you need to make progress on the technology side and adjust on the business side. We cannot know exactly how this plays out. Some of these things are beyond the event horizon and we cannot know how they turn out, but we can shape the possibility frontier by making it possible to do some of the self-guided field laboratory kinds of things.

(Sujit) Telecom operators are both fiber and wireless providers, so there is some convergence. There is a bigger clash of business models with satellite providers.

(Mariam) - ***The needs that scientists have are different than those the commercial providers use to develop their business models. Any comments on what science needs are versus commercial?***

(Sujit) - Future wireless networks are trying to become as configurable as possible so that they can fill many kinds of needs.

(Andrew) – Traffic evolves to fill any road that you build. Field sensing tends to involve low data rates, with low demands for anything responsive, although that is rapidly changing. Data flows from the field sensor community look different from someone streaming Netflix. In the longer term, the introduction of high-fidelity, physics-based code and the need for distributed different versions of code will have a lot more two-way traffic and will need predicable latency.

(Sujit) - We need to think of this network as it is evolving. We will need models for this type of computing.

(Larry) - We moved from a central supercomputer to a nationally and globally distributed system that acted just like it was in the machine room. We need to take this point of view to wireless over the next ten years, as we move into 6G and NextG. It is a once in a generation opportunity to rethink software infrastructure. We should start thinking about experiments or partnering with the first leaders.

(Al) ***Is this group interested in leading the rebellion that does the same thing for wireless as it did for creating the internet?***

The panel discussion video may be viewed at: 2023 4NRP - Future Wireless Extensions of Regional Optical Networks Panel - YouTube

3.7 Larry Smarr, Calit2 Founding Director

NRP Application Drivers

This presentation showed some of the newest and biggest users of PRP resources. Looking at the data using Grafana, NRP is almost up to 19,000 CPU cores today under Nautilus, the Kubernetes hyper-cluster, 1,300 GPUs and 4 petabytes. There are over 800 namespaces in Nautilus, meaning projects with a Principal Investigator (PI) and a CI logon. The Grafana graph shows the growth just in 2022 to over 900 GPUs and 7,000 CPUs that are used daily.

The largest user is more than one million times the smallest. The largest projects are OSG IceCube, UCSD-haosulab, and OSG Opportunistic to name a few. The Pacific Research Platform video[31] has beautiful imaging about Ice Cube, VR Systems, WIFIRE – everyone should watch this, it is a ten-minute video.

When the PRP proposal was submitted, UCSD concentrated on five major areas of science we were going to do:

1. Particle Physics
2. Biomedical 'omics
3. Telescope surveys
4. Earth sciences
5. Visualization, Virtual Reality, Collaboration

The Open Science Grid (OSG) was the model that PRP wanted to grow into and federate with and has been integrated with the PRP. OSG Grid delivers to over 50 fields of science, 2.6 billion core hours a year versus the 35,000 core hours in 1980 at NCSA. Just one of the namespaces, *osg-opportunistic*, supported a wide set of applications as the largest consumer of Nautilus CPU Core-Hours in 2022. It was able to grab cycles that were not being used, delivering 3.7-million CPU core-hours peaking at 3,500 CPU cores that would have been "thrown into basket." Because of the integration between OSG and PRP, PRP just looks like another cluster. It is transparent. It is delivered to a wide range of applications because of the architecture of the system.

For Particle Physics, the holy grail of the LHC is measurement of di-higgs[32] production to infer the hhh (3H)[33] coupling that determines the Higgs potential. They are running 1,000 CPU jobs, each one grabbing 10 CPUs over PRP/NRP to use machine learning and bringing that to particle physics. An example is the namespace *cms-ml* which is the fourth largest consumer of Nautilus GPU-hours in 2022.

Another big area is Telescopes. PRP/NRP supports some very large scientific instruments. NSF's south pole neutrino detector, IceCube, is exploring how a neutrino photon cascade goes through ice. It is a highly parallel simulation needing over 500 GPU simulations to improve the ice model. This leads to a significant improvement in pointing resolution for multi-messenger astrophysics. This *osg-icecube* namespace was the largest consumer of Nautilus GPU hours in 2022. Of all the GPU cycles available to OSG, IceCube is the largest consumer, and the biggest supplier of GPU cycles to OSG is NRP. It runs in low-priority mode using only GPU cycles that would be otherwise unused.

The Laser Interferometer Gravitational-Wave Observatory (LIGO) uses Nautilus/OSG data CI. LIGO does not use the CPU and GPU resources on NRP as much as the data management aspect of NRP. Rucio[34] is the de facto data management system for many large instruments. LIGO uses PRP-managed Kubernetes pods to run their production Rucio data management system on Nautilus, one of the major users of the OSG caching infrastructure.

[31] The video can be found at: https://nationalresearchplatform.org/media/pacific-research-platform-video/
[32] An update of di-higgs can be found at: https://atlas.cern/updates/briefing/new-milestone-di-Higgs-search
[33] Triple Higgs
[34] The Rucio website is: http://rucio.cern.ch/

PRP supports radio astronomy by partnering with the Collaboration for Astronomy Signal Processing and Electronics Research (CASPER)[35]. Headquartered at UC Berkeley, CASPER is a collaboration of over 1,000 members and 50 radio-astronomy instruments worldwide. This wholesale innovation to research develops open-source, signal processing and instrumentation pipelines. They use a lot of FPGAs and PRP has worked closely with CASPER to integrate their systems with PRP software and systems. PRP imported the tools and libraries to which the thousand-plus CASPER members have access and have added 32 FPGAs. Juypter-hub, where software, results and documentation are shareable, has been developed as well. And this is all optically connected.

A good example of this is the 2021 workshop talk and tutorial given by John Graham on how CASPER designs, compiles, tests, and evaluates instrumentation on the PRP, then deploys dedicated FPGA and GPU clusters at the observatories. The agenda for the workshop is available at: https://casper.berkeley.edu/index.php/casper-workshop-2021/agenda/ and the talk is available on YouTube at: https://www.youtube.com/watch?v=fevhSBnDH8U&t=10485s and the Tutorial at: https://www.youtube.com/watch?v=fevhSBnDH8U&t=17895s.

The new PNRP grant includes 32 new FPGAs. We may see a move to FPGAs like the move from CPUs to GPUs. UCSD has a new award for innovating FPGAs, the secret sauce for innovation. The first radio barometry image of a black hole used instruments that used these resources. Because you are using the whole earth as the scale of your radio telescopes, linking them together around the planet, as the planet moves, we can see fine-grained views of the universe.

OpenForceField[36] (OFF) uses OPEN Software, OPEN Data, OPEN Science and PRP to generate quantum chemistry datasets for druglike molecules in a way that takes all data and puts it into an NSF-supported archive. This makes it possible to do the same work that Folding@home[37] used to do. PRP can run millions of quantum chemistry workloads, through the PRP. One million jobs have been completed, using worker docker images that they throw into the PRP system, with all results going into an open science QCArchive[38]. OFF was the largest CPU core consumer. About 50% of OFF total compute is run on Nautilus.

The Nautilus namespace *tempredict* for Professor Benjamin Smarr utilized PRP to compute COVID-19 and vaccine responses for approximately 65,000 participants. An Oura ring was used to obtain medical data from frontline medical workers, including minute-to-minute temperature, respiratory rate, heart rate, and importantly heart rate variation. The Oura company[39] allowed other Oura ring users to opt into the experiment – a step forward for personal medicine. The data was then used to create individual timelines for 70,000 people using SDSC's HIPPA compliant system to make the data unidentifiable, and then run in the PRP.

NeuroKube is an example of why we need optical networks, taking electron microscopes of brains and using machine learning to use images and refine them to make very precise images, giving us a three-dimensional image.

---

[35] CASPER website is at: https://casper.berkeley.edu/.
[36] OpenForceField's website is at: www.openforcefield.org
[37] The website for Folding@home is at: https://foldingathome.org/?lng=en-US
[38] The QCArchive website is at: https://molssi.org/software/qcarchive/
[39] The Oura website is at: https://ouraring.com/

Taking machine learning to the edge, the Fire Ignition Library (FigLib) and SmokeyNet[40] (PRP namespace *digits*) are examples of combining publicly available datasets and deep learning architecture using spatiotemporal information from camera imagery for real-time wildland fire smoke detection. Training in neural nets is a very computationally intensive activity, but once you have the neural net trained, you can put it in the learning accelerator on the camera. The camera can do its own machine learning recognition. Intelligence is going to the edge. Tensorflow[41] is at the edge in the device.

Virtual Reality has brought machine learning to the movement of the physical world representation into highly accurate virtual worlds, making digital twins of everything. The namespace *ucsd-haosulab*, supports a major project at UCSD that consumed the second most Nautilus GPU hours in 2022. It is bringing machine learning to robots. For robots to manipulate objects, they need to learn more about the physical world. They build these digital twins and have these agents: specialists and generalists. Specialists learn specific skills by trial and error and that uplifts into the generalist who distills knowledge to solve arbitrary tasks.

A completely different, but also massive use of PRP's GPUs is UCSD Professor Ravi Ramamoorthi's work on transforming a series of 2-D images into a 3-D view synthesis using machine learning. Neural radiance field (NeRF)[42], which is a fully connected neural network that can generate novel views of complex 3D scenes based on a partial set of 2D images, is taking AIs directly into that process. The namespace *ucsd-ravigroup* has consumed the third most Nautilus GPU-Hours in 2022.

Time Magazine called NeRFs, one of the 200 biggest and most important inventions of 2022. The New York Times has a whole unit dedicated to this. Meta is investing in NeRF for 3D Object Scanning. Larry recommends viewing the Mark Zuckerberg Keynote at MetaConnect with the CEO of Microsoft and the CEO of Accenture: https://www.youtube.com/watch?v=hvfV-iGwYX8. Accenture has 700,000 people worldwide. They have bought tens of thousands of headsets, and they are building these virtual worlds. Accenture then advises the Fortune 500 on how to move work into a true hybrid space. Microsoft is taking all its production software into this space. Before long it will be hard to distinguish whether you are in a virtual or physical space.

This community has created something extraordinary in seven years. This kind of change is massive.

This presentation may be viewed at: 2023 02 09 4NRP - Larry Smarr Part 2 - YouTube

---

[40] The *paper FigLib & SmokeyNet: Dataset and Deep Learning Model for Real-time Wildland Fire Smoke Detection* by Anshuman Dewangan et al. can be found at: www.mdpi.com/2072-4292/14/4/1007

[41] The TensorFlow website is at: https://www.tensorflow.org/

[42] Information on NeRFs can be found at: https://datagen.tech/guides/synthetic-data/neural-radiance-field-nerf/

3.8 Panel: The Global Research Platform: An Overview

3.8.1 Panel Chair: **Joe Mambretti**, Director, International Center for Advanced Internet Research, Northwestern University, Director, Metropolitan Research and Education Network, Director, StarLight International/National Communications Exchange Facility

The Global Research Platform (GRP)[43] consists of communities around the world that share information on creating ecosystems for distributed computational science.

The NSF has a lot of documents like the Cyberinfrastructure Framework for the 21st Century (CIF21)[44]. One of the key takeaways is that science is global, and NSF should extend worldwide. A key piece of this is information sharing, a lot of the technologies that are emerging in the United States are also emerging in other parts of the world. There are concepts, experiments, instruments, methods, techniques, protocols, data, technologies, and results that are openly communicated and shared among collaborative science communities worldwide. Generally, these communities are in their own silos and do not have a lot of opportunities for interconnecting.

The themes that are being explored by these communities are large scale global science cyberinfrastructure ecosystems as a whole: architecture, services, and technologies. A research platform requires orchestration between multiple domains such as AutoGOLE/SENSE, large scale high-capacity data WAN transport, high fidelity data flow monitoring, visualization and analytics and introduction of AI and machine learning.

Some of the scientific driver applications come from the large-scale instruments that need to move exabytes of data. A nice model for this is the LHC as can be seen from looking at the LHCOPN map. There are a variety of communities that have joined the LCH1 such as the Bell II experiment, the Pierre Auger Observatory, and the XENON Dark Matter project.

This presentation may be viewed at: [2023 4NRP - Joe Mambretti - YouTube]

3.8.2 Speaker: **Jeonghoon Moon,** Principal Researcher, Korea Institute of Science and Technology Information (KISTI), Advanced KREONET Center Chair, APAN APRP (Asia Pacific Research Platform) WG

The Asia Pacific and Korea Research Platforms: An Overview

Asia Pacific Research Platform

The Asia Pacific Research Platform (APRP) was introduced to the Asia Pacific Advanced Network (APAN) in 2018. There have been 18 sessions during 9 APAN meetings. The EU government provided an

[43] The GNA-G website is at: https://www.gna-g.net/.

[44] Cyberinfrastructure Framework for 21st Century Science, Engineering, and Education (CIF21) (nsf.gov)

Asi@Connect[45] grant to help develop the platform. The goal is to share 'xRP' experience by promoting the HPC ecosystem in the Asia Pacific region with APAN members and Association of Southeast Asian Nations (ASEAN) countries and to establish the APRP as part of the Global Research Platform (GRP).

Some of the challenges the project faces are building the platform around the lack of IT infrastructure in Asia, the need for NRENs to upgrade their networks to create the big data super-highway, and applying third party applications (Weather/Climate, Environment, Agriculture (Smart Farm), AI, Bioinformatics, Cloud, etc.)

Asi@Connect participants are Korea, Australia, Malaysia, Pakistan. The data explosion in science and technology drives the demand for high-capacity networks for accessing computers, storage, and collaborations. Simultaneously it brings many restrictions on the generation, transmission, and utilization of research data.

The purpose of this project is to suggest a method to enable the public high performance computing resources to be delivered to underdeveloped countries through the TEIN network. The project will interconnect available computing resources in a distributed manner and the resources of the computing servers will be managed by container cloud technology.

To implement this, the building strategy is divided into three activities.

Activity 1 is to build the super-highway, interconnect participant NRENs via TEIN, install a DTN at participant locations, and measure the Science DMZ activity using perfSONAR.

Activity 2 is to build a distributed-HPC platform to manage CPU/GPU computing resources including storage by Kubernetes, build block-chain based ID federation, and build a user interface that is easy to use.

Activity 3 is presenting pilot application use-cases such as bioinformatics-based science and AI-based science use cases. Examples are Malaysia (Bioinformatics) and Pakistan (AI research).

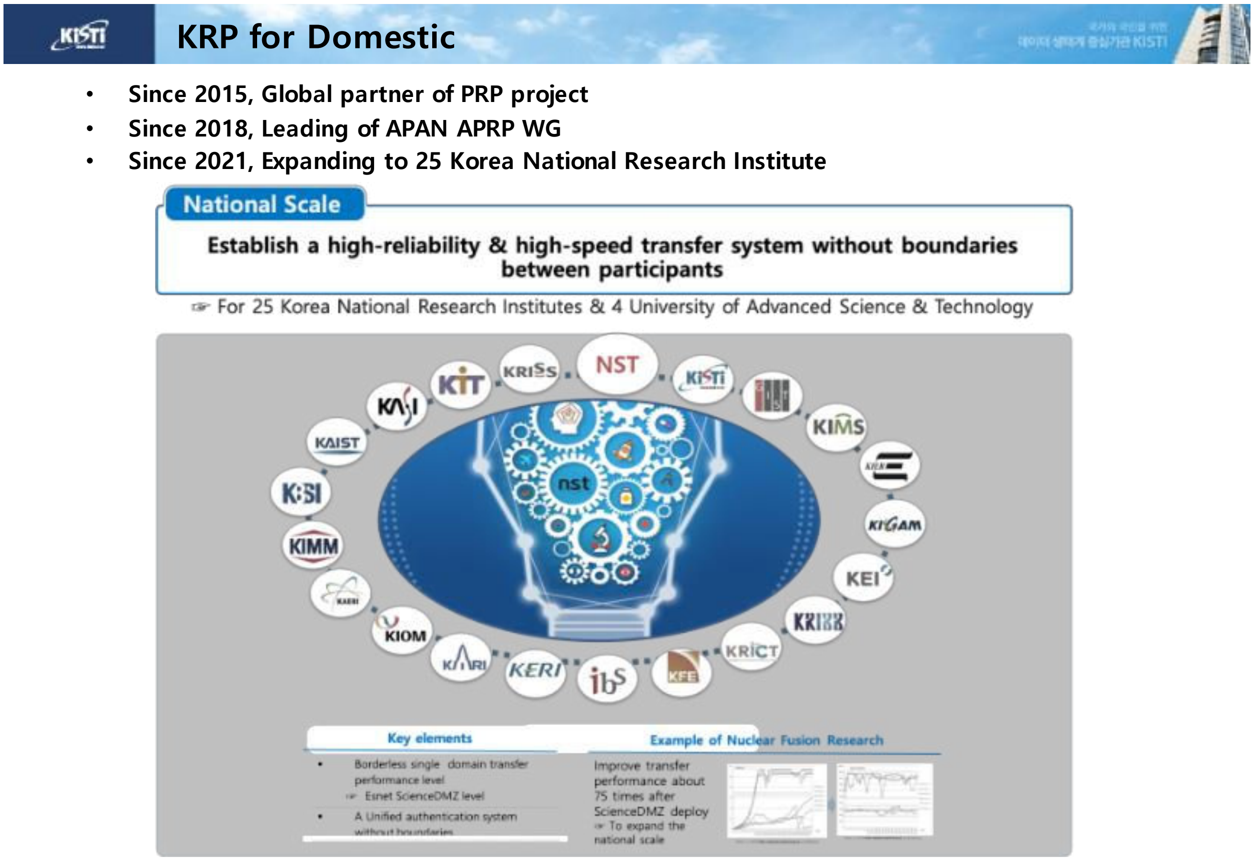
Korea Research Platform

The Korea Research Platform (KRP) established a national scale, high speed transfer systems without boundaries between participants. Korea has 25 government-funded national research institutes. Most of them are located near KISTI in Daejeon which means that KREONET already has a dedicated optical layer for those. Plans are to expand to the major research institutes such as aerospace, chemistry, nuclear fusion, medical, ICT, universities with advanced science and technology as a beginning step, then gradually expand to cover all.

The first one is the building dedicated to science big data transfer and AI computing environment. One of the biggest challenges is how to manage the obstacle of passing the security firewall.

---

[45] Asi@Connect is a project that provides dedicated high-capacity internet connectivity for research and education communities across Asia-Pacific. It is jointly funded by the European Union and Asian partners and is managed by TEIN*CC based in South Korea. For more information see TEIN 테인 협력센터.



New related challenges include the Smart Hospital development project, which was awarded in September 2022, the Korea Rural Development Administration project granted in March 2022, and the Korea Environment Institute project awarded in December 2022.

The Smart Hospital project aims to construct a patient-tailored, hyper-personalized digital medical ecosystem based on the OmniVerse system. Participating in this project are 20 institutes, hospitals, and companies. KISTI/KREONET will be providing the big data super-highway for inter-hospital and genomic data centers, the high-speed inter/intra wireless communication and IT technologies for the Smart Hospital.

The Korea Rural Development Administration project, entitled “Development of integrated linkage system for agricultural big data and utilization model,” will be using KISTI/KREONET resources to facilitate agriculture big data transfer, run crop models on the research platform (using CPU and GPU), and establishing a research platform for farmers and researchers.

The Korea Environment Institute project, entitled “Establishment and development of a smart decision-making system based on diverse environmental model,” is a 3–4-year project. KISTI/KREONET will be building a container-based research platform and Ceph storage, building a multi-institutional distributed research platform, and operating various models on the platform.

New challenges for APRP and KRP may expand to disaster/urban flooding and AI education areas.

This presentation may be viewed at: [2023 4NRP - Jeonghoon Moon - YouTube]

3.8.3 Speaker: Harvey Newman, Professor, Division of Physics, Mathematics and Astronomy. California Institute of Technology Chair, GNA-DIS Working Group
The GNA-G Data Intensive Science Working Group

The Global Network Advancement Group is fulfilling an important mission to provide the needs of not only the Large Hadron Collider (LHC), but of major science programs.

It is the long-term nature of our research to maintain and find solutions to problems such as the Higgs Boson discovery. We are making tremendous progress. This global system, which we first saw with initial developments by Caltech and UCSD together, has grown to over 170 Tier2 centers at universities, smaller laboratories and 13 national centers.

We are seeing challenges to both scale and complexity. The global data flow continues to grow exponentially. Against this backdrop, the GNA-G[46] brought together researchers, R&E networks, Global exchange providers, to share infrastructure and support science programs. Two working groups were formed: the AutoGOLE/SENSE WG and the Data Intensive Sciences WG.

Gna-G has two global testbeds. The AutoGOLE/SENSE testbed provides a global persistent testbed, and the P4 testbed supported by GEANT explores free router developments. At the last SC22, there was a watershed moment during the NRE Data Tsunami event with participants utilizing near peak capacity of 5 terabits per second.

We need to develop a new type of system to meet and address the needs and challenges faced by major data intensive science programs. It is the science mission that gives us the impetus to do this over decades. It is a new type of system that will change the way in which research and education networks manage and operate themselves in a new level of cooperation between networks and the science programs they serve.

Between the PRP, NRP, GRP, GNA-G and regional research platforms, this is a new grand new direction. There has been tremendous progress, and it is a wonderful time to bring this up. The key to success is partnerships.

This presentation may be viewed at: [2023 4NRP - Harvey Newman - YouTube]()

3.8.4 Speaker: Maxine Brown, Senior Research Scientist, University of Illinois Chicago
The Global Research Platform: Initiatives and Events

The Global Research Platform (GRP) is derivative from the PRP and the Global Lambda Integrated Facility (GLIF). In 2019, GLIF merged with GNA Tech to form a new organization: GNA-G. The GNA-G and the GRP are two things that arose from these early efforts.

The first GRP Workshop was at CalIT2 in 2019 with 95 attendees, between a PRAMA meeting and an E-Science conference. COVID hit in 2020, but in 2021, 140 people registered online for the 2nd GRP

[46] https://www.gna-g.net/

Workshop. In 2022, the GRP workshop co-located with the E-Science held in Salt Lake City and had 60 attendees.

Workshop topics have included:

- Large Scale Global Science
- Next-Generation Research Platforms
- Orchestration Among Multiple Domains
- Large-Scale Data WAN Transport
- Data Monitoring, Visualization, Analytics, Diagnostics, Event Correlation AI/ML/DL
- International Testbeds for Data-Intensive Science
- Updates on international networks and open exchanges

This year the 4th GRP will be collocated with E-science in Limassol, Cyprus October 9-13, 2023. See https://www.esience-conference.org/2023.

The full presentation may be found at: 2023 4NRP - Maxine Brown - YouTube

3.9 Panel: Building the NRP Ecosystem with the Regional Networks on their Campuses

3.9.1 Panel Chair: **Jen Leasure,** President and CEO, The Quilt

What you don't see on the Quilt map are the last mile connections to about a third of the nation's research and education institutions. The Quit community feels strongly about the "Minds we Need" vision.

The Quilt regional networks that have Nautilus nodes connected to NRP are CENIC, NYSERnet, OneNet, GPN, C-Light, MREN, FRGP, PNWGP, Eastern Research Platform, Merit, Florida LambdaRail, AREON, MOREnet, Albuquerque GigaPoP, KanREN, Network Nebraska and SD-REED. The Eastern Research Platform, which is a collaboration of Edge, OSHEAN, CEN, and NYSERNet, came out the first NRP workshop.

The three panelists represent regional networks at various stages of deploying Nautilus nodes on the research platform.

The full presentation may be viewed at: 2023 4NRP - Jen Leasure - YouTube

3.9.2 Speaker/Panelist: **Scotty Strachan,** Principal Research Engineer, NevadaNet Regional Perspective from NevadaNet

NevadaNet serves a state that is one of the most urbanized states, the most people in an urban zone, a geographically diverse landscape. Scotty monitors the oldest tree species in the world, with high elevation climate monitoring. He is running one of the highest elevations 24x7 weather stations in the Sierra Nevada which brings a whole different approach to the function of the network.

How can the National Research Platform meet the types of challenges seen in Nevada? How do we get to the rest of research or the research-curious, the people who want to teach science. Nevada has no large facilities, no flashy national scale research going on, but there is a wide variety of research going on as a land-grant based EPSCOR state. As the classic graph on how cyberinfrastructure gets used shows, a small group of people use the most, and then there is a long tail of users that have not been facilitated into scaling their science yet. From a cyberinfrastructure ecosystem viewpoint how do we get the long tail of users moved up a little bit. It comes down to asking the right questions such as do you spend a lot of time maintaining your technology? Have you had grants turned down because you cannot or don't think you have resources or capability to use this technology? Do you think additional support would contribute to your research success?

What you may find is that unless researchers are being directly facilitated, the national platform engagement is not found. For many it just takes a little time and effort. Many perceive that the bar to entry is too difficult, and they do not belong on the platform. This perception is one of the challenges to wider use.

The value proposition for the ecosystem and the components that is important to users:

1. Does it service a wide range of service workflows?
2. Is it easy to get access?
3. What is the learning curve to productivity?
4. What is the support model?
5. Is it a service commitment?

All of these things are related to time-to-science and stability.

The NevadaNet example starts with identity federation, federating everyone including the state and community colleges with InCommon for semi-automated onboarding and provisioning, then standing up new assets outside of the institutions so that they are exposed to the national fabric as if they are location agnostic.

Thinking about the research futures in the west, many of the regional challenges are geographical such as water and wildfires. How does the future of wireless as an extension of optical networks fit into a national research platform approach? The Internet of Wild Things (IoWT) will be driven by the regional networks as they can cross political and geographical boundaries to enable the kinds of work that need to be done for wide area science and engagement of those missing millions.

Scotty Strachan's talk can be found at: [2023 4NRP - Scotty Strachan - YouTube]

3.9.3 Speaker/Panelist: **Grant Scott,** Director, Data Science and Analytics Master's Program, University of Missouri

Grant Scott's background includes standing up a data science program at University of Missouri (Mizzou) using Kubernetes. He has done a variety of things using Jupyter and interactive computing and is pushing researchers to use Nautilus.

Mizzou has been using Nautilus for deep learning, computer vision, protein structure prediction and modeling with diffusion models, and robotics. Working with GPN, they have run many workshops and tutorials to help researcher-facing cyberinfrastructure professionals understand what Kubernetes is, how it can be leveraged, and how research codes can get migrated from workbenches into the Nautilus cluster. One of their post-docs ran mini tutorials at SC22 in the GPN booth.

As a case study, Dr. Alex Hurt is using Kubernetes to streamline the scientific processes. He was able to churn through and train 27 deep learning models, 8,100 Epochs, and 30 million iterations of training on these models. He has pulled through about 415GB of data and 124,740GB of neural model loading. The wall clock time was about 77 days, about three hours of human effort.

Another example is from PhD candidate Anes Ouadou, who trained 144 deep learning models to evaluate the best way to map areas burned by wildfires from space-based imaging. Wall clock would have been 22 days, but he was able to do it in 13 hours.

In both these examples they are getting more than enough data for multiple publications in less than 24 hours of human effort.

Dr. Scott has pushed the educational and interactive interface for STEM education including running an introduction to Python programming course for the state of Missouri, to upscale the capabilities of the workforce during the recent pandemic. He has also used this for computer science classes. One of the beautiful things about using Nautilus and Jupyter is that it facilitates STEM education. Students do not need to install special software. Everything is on one platform.

NSF Award OAC#1925681 is a CC* Team award helping the Great Plains region leverage collective cyberinfrastructure resources and allows GPN to make contributions to Nautilus and the NRP. GPN is working on a proposal to deploy NRP nodes into 8 locations across 6 states.

Grant Scott's presentation can be viewed at: [2023 4NRP - Grant Scott - YouTube]

3.9.4 Speaker: **Eric Buckhalt,** Senior Director, Technology Strategy, Architecture, & Governance, Georgia Tech
Georgia Tech / Southern Crossroads (SoX)

SoX is deploying nodes as part of CC* Area 2 regional networks that include several HBCUs in Alabama and Georgia. They have FIONA nodes purchased as DTNs or perfSONAR test points and are considering redeploying them as Nautilus nodes. They will be deploying new hardware purchased by the CC* award and repurposing some undermanaged perfSONAR test points to build a functioning MaDDash.

The goal is to enable research within their community and to assist faculty unfamiliar with the platform. SoX is looking for toolkits, demos, training, tutorials, and workshops that can be shared with their participants to help take advantage of the NRP.

SoX also has data center space and offers connectivity and space for those interested in housing nodes.

Hopefully, they will be able to share stories about deployment later.

Eric Buckalt's presentation can be viewed at: [2023 4NRP - Eric Buckhalt - YouTube]

4. DAY THREE – Friday, February 10, 2023

4.1 Panel: NRP Science Impacts

4.1.1 **Chair: Subhashini Sivagnanam,** Manager of CI Services and Solutions, SDSC/UCSD

4.1.2 **Dr. Jia Wan -** Statistical Visual Computing Laboratory (SVCL), Department of Electrical and Computer Engineering, University of California, San Diego Statistical Visual Computing with Nautilus

Data and computation power are growing every year. The computation platform has become a very important factor in research. This talk discussed some of the projects that are using the Nautilus platform.

The purpose of the *Audio-Visual Special Alignment project* is to learn video representation in a self-supervised way without any manual annotations. The project first collected a dataset which contains over 200 hours of 360-degree video collected from YouTube. All these videos contain spatial data in Ambisonics format so they could encode the sound along different directions. Based on this dataset, the team designed a self-supervised learning method motivated by human behavior. Humans can automatically identify and align the object and the sound of the object. The model is trying in a similar way to do this which can improve downstream video tasks.

Another project is *DISCO: Adversarial defense with local implicit functions*[47], which aims to counter an adversarial attack to a clean image. We can easily train a detector to detect objects in a clean image, but if we add specific noise to the image, the detector no longer works. Adversarial defense is used to protect against these types of attacks and used to recover the original image. This work might be used for other types of adversarial attacks.

Dense network expansion (DNE) for continual learning

The most widely used continuous learning methods are distillation-based model and sparse network expansion. The distillation model generally does not work very well, and the sparse network expansion suffers from a quickly growing model and feature size. We propose the Dense Network Expansion (DNE) to make full use of old features, the model size and feature size grows slowly and improves performance.

Deep Realistic Taxonomic Classifier – This work is to address the imbalanced and long-tailed issue which is a common problem in AI models. Some images are common in datasets. Other images are rare. This poses a problem for current classifiers. It is very hard to recognize those classes, so we reorganized the

[47] https://openreview.net/pdf?id=vgIz0emVTAd, DISCO: Adversarial Defense with Local Implicit Functions, Chih-Hui Ho, Nuno Vasconcelos, *Advances in Neural Information Processing Systems* (NeruIPS), New Orleans, Louisiana, United States, 2022.

class in a tree structure based on class taxonomy and we allow the decision to be made at intermediate levels.

The example used is a Mexican Hairless Dog. Maybe we do not know this picture is a Mexican Hairless dog, but we know it is a mammal and it is a dog. This allows us to preserve correct information and prevent performance degradation on tails or silent failures.

CoordGAN: identity preserved texture swapping – allows application of new texture while maintaining original. This is achieved by the disentanglement of structure and texture features and then allows us to apply different texture while preserving the structure. We can use this technology to apply different styles from the image. Another interesting application is the disentanglement of shape and articulation for 3-D models. We can use this to generate shapes with unseen articulations.

Visual Relationship Learning using Conditional Queries

In this project, we want to learn not only the object in the image but also the relationship between objects. Traditional methods have two steps, the first to detect the objects in the image and the second to find the relationship between them. In this work propose a one stage method that predicts object and relationship. Two examples were shown.

Prediction of visual field (VF) based on Optical Coherence Tomography Angiography (OCTA). Glaucoma is the leading cause of blindness, so it is very important to detect and monitor it in its early stages. In practice, the prediction of the visual field is a time consuming and complex process. We propose to predict the visual field based on optical coherence tomography angiography which is a more efficient way based on the vision processing method.

In conclusion, there are many directions for exciting progress in AI, however the requirements in data and compute are becoming astronomical. The main driver for the latest breakthroughs such as GPT-3 and ChatGPT is the size of the model and data. These systems do not use technology beyond what has already been shown. The difference is the memory and compute requirements are several-fold larger. Soon no single PI will be able to work on these problems without the support of platforms such as Nautilus. Even this will be too small soon, so we need to rethink the whole architecture for academic research to enable this kind of scalability.

This presentation may be viewed at: 2023 4NRP - Jia Wan - YouTube

4.1.3 Shih-Chieh Hsu, University of Washington and NSF A3D3 Institute[48] Accelerating Science Discover with AI inference as a Service

The NSF HDR[49] Institute Accelerated Artificial Intelligence Algorithms for Data-Driven Discovery (A3D3) started in 2021. It is a multi-disciplinary, multi-institute collaboration to push a particular area of science and engineering technology. The vision is to establish a tightly coupled organization of domain scientists, computer scientists, and engineers that unite three core components which are essential to achieve real-time AI. This allows us to transform science and engineering discoveries.

[48] https://a3d3.ai/

[49] Harnessing the Data Revolution (HDR)

A3D3 is part of a larger ecosystem that harnesses the data revolution system. A national-scale activity to enable new modes of data-driven discovery, HDR will address fundamental questions at the frontiers of science and engineering. There are three parallel tracks (Institutes, TRIPODS, DSC), 70 awards, and 200 million dollars that cover a wide range of the area over the size and engineering. A3D3 covers real-time AI and multi-messenger astrophysics.

Part of five major institutes, A3D3 is primarily composed of trainees, students, and post-docs. They are interested in passing this technology on to the next generation. The next generation experience will outpace industry data volumes by 2026. The Large Hadron Collider (LHC) data generation has already achieved 100 petabytes in 2019 and will be at an exabyte by 2026. LCH is not the only instrument challenged by large data volumes. The large telescopes coming online will also be a challenge.

Continued software research and development improvements fit into optimistic assessments of future computing power that might be available. The same holds true for disk and tape. It is not easy to predict if memory and network projects are sufficient, but they are undeniably finite resources.

The data volume increases together with the compressive data itself is a factor. It can be challenging to do some of the reconstructions for the CMS and LIGO-VIRGO-KAGRA projects. A cross-disciplinary challenge, not just for LHC, but also for other projects such as LIGO, IceCube, DUNE, the EM telescope ZTF, and neuroscience is the need for low latency and high throughput. Computer scientists who can lead the foundational AI evolution working together with hardware engineers who can optimize the algorithm to accelerate computation can make it available for science.

The evolution of AI is amazing in the way that it increases performance beyond the conventional algorithms, using deep learning to unlock more information from the data itself, but this means longer latency. If the direction is to use more AI, AI algorithms can be accelerated by using coprocessors. But it is difficult to scale up, very expensive, and users need to know coprocessor technology in some detail. Another option is as-a-service connection using a couple of powerful coprocessors to serve many clients. Domain scientists can focus on coding, and the hardware engineer and foundation AI expert can focus on the best way for that particular AI architecture to be accelerated in the coprocessor. Then we can bring different expertise together to tackle the final science problem.

This model is the simplest support for mixed hardware, is scalable and can support multiple clients. You can easily phase in any kind of coprocessor such as FPGA, GPU, and even IPU for Machine-learning-as a-service. Open source such as SONIC for HEP and Hermes for Gravitational Wave is used at A3D3. The Nvidia triton inference server is used for the GPU. This model can easily apply to high-level trigger or the offline reconstruction in particle physics.

To test the strength of this model, A3D3 collaborates with computing resources across the nation, CERN, and NAR Labs in Taiwan. Several benchmarks and results from tests were presented showing how NRP can be leveraged as a wonderful playground for aaS R&D. In a heterogeneous computing performance comparison, FPGA-aaST greatly outperforms GPU-aaS for FACILE in the smaller models. There is comparable performance between FPGA-aaST and GPU-aaS for ResNet.

In the CMS MiniAOD production, they are only using 10% of per event latency AI algorithms. For the full event processing, only 10 percent will be accelerated but for throughput performance 10% speed up is achieved relative to just running the old jobs on CPU.

For the Neutrino protoDune-single phase 1Kt LAr-TPC, they have processed 6.4 million out of 7.2 million events from 2018. They can speed up the pre-processing by a factor of 2.7, optimizing the CPU, GPU ratio 1 to 68.

With the Gravitational Wave, using the conventional GPU processing method takes a long time. Moving to the as-a-service, even using a CT4 CPU cluster, they can speed up the inference by a factor of ten. Adding a coprocessor gains an additional factor of five speed up. Trying different combinations of the set-up results in speed up proportional to the number of GPUs used.

NRP allows A3D3 to use the FPGA server and quickly test the FPGA-as-a-service. Being able to work with NRP allows them to understand how HBM memory can be efficiently used to transfer data to achieve better throughput. This work is only possibly with the physical power at NRP.

NRP is a wonderful playground for as-a-Service R&D.

As a summary, AI as a service shifts the paradigm of real time AI processing and offline processing. A3D3 has demonstrated promising acceleration using the Large Hadron Collider. The influence on systems like NRP is crucial for future tests. They would like to try multiple cards, different algorithms, and client server combinations to make R&D easy and available. HDR has an active training program – encourage those to apply.

The presentation may be viewed at: [2023 4NRP - Shih-Chieh Hsu - YouTube]

4.1.4 **Jeff Wagner** - Open Force Field Technical Lead
Open Force Field: Scavenging pre-emptible CPU hours in the age of COVID

Jeff gave an Impromptu talk about using the PRP/NRP platform. Open Force Field (OFF)[50] is an open-source consortium partially funded by NIH and industry partners in the pharmaceutical and materials science industries. It is their mission to make open-source tools and open-source datasets and models for modeling pharmaceuticals, for modelling other organic molecules, polymers with desired properties. OFF was the number one PRP CPU-hour user during 2020 and 2021.

OFF tried to be good citizens and working with Dima Mishin at UCSD scavenged pre-emptible CPU hours (and panel spots) in the age of COVID. They used the SunCave quite a bit. They are committed by contract not to create secrets. They cannot sign secret SLAs. They give fellowships to a number of post-docs and graduate students at UCSD and several other institutions.

PRP and now NRP is a great resource for OFF because they do not have to worry about the hosting costs, the engineering, or the administration of the platform. It is easy to set up tons of Kubernetes workers and take whatever is available.

Open Software, Open Data, Open Science

On the application side, the behaviors of molecules to a first approximation from quantum chemistry can be modeled using simpler things. When you add them all you can get good approximation. These models of physics may not be perfectly accurate and allow researchers to take time steps on the scale of femtoseconds. If you take a billion of those, you simulate a microsecond and that is a microsecond of a

---

[50] https://openforcefield.org/

drug revealing what it is going to do in the presence of a protein that can be very useful in a pharmaceutical context. Using cheap computer time, you can screen possible drugs and save expensive lab testing time for only viable options.

It requires large amounts of data to train these force fields. Training new force fields requires new data (SMILES) format, which is a linear way to encode any reasonable organic molecule. OFF has built a pipeline that can take a request from the simple line notation, pop it up to 3-D and come up with some bad guesses at what the geometry wants to be. Then it can be submitted to the QC archive project[51] which is run by MolSSI[52], an NSF center and refine these initial guesses using quantum methods to see what the equilibrium structure of this molecule wants to be, and then that becomes a training set for the force field.

PRP/NRP is helpful in generating enormous amounts of QM data. OFF created their first force field in 2019 and after that got onto PRP. The scale of what they knew they could do expanded, and it took the team scientists a while to realize that they could submit gigantic datasets. This meant they had to be disciplined with their datasets when making their second release in case they had an emergency before the second force field came out. They massively increased replicas being produced on Kubernetes.

The model they created was good. The commercial model is still best, but they are catching up. OpenFF 2.0 outperforms other public small molecule force fields and continues to improve. OpenFF 3.0 will be coming out soon.

The datasets on QCArchive are fully open!

The presentation can be viewed at: 2023 4NRP - Jeff Wagner - YouTube

4.1.5 Panel Discussion

Question: Why NRP and not other services?

(Jeffrey) Sloan Kettering has some resources, some at UCI, for most common workloads NRP works best.

(Shih)Would like to align 100 GPU and 10,000 CPUS for one day for scaling test, and would like to try different combinations, different co-processors have different results and can split AI inference to different co-processors.

(Jia) Some local machines in the lab, but they are hard to use if not expert, resource is limited and not able to upgrade, cannot upgrade all the time, so Nautilus works, easy to work on large scale problems, more than 10 students working on same cluster, so good to have Nautilus platform.

Question: Is there someone in A3D3 looking into low power AI inference?

The answer is yes, using a lower number of bits without losing the accuracy of the model is being investigated within the community, which is actively working on benchmarking these activities. A3D3 is working with a project called tinyML[53] that is actively involved in the tiny machine learning benchmark

---

51 https://molssi.org/software/qcarchive/
52 https://molssi.org/
53 https://www.tinyml.org/

activities. They are pursuing low latency low power AI and trying to investigate on each device more than just high throughput and scalability. As a service model it has the potential to have an impact.

The panel discussion can be viewed at: [2023 4NRP - Panel NRP Science Impacts - YouTube]

4.2 Panel: Engaging Underrepresented Technologists, Researchers, and Educators in the NRP ecosystem

4.2.1 Chair: **Brenna Meade,** Network Engineer, International Networks at Indiana University

4.2.2 Panelist: **Karen Lopez**, Automation Engineer, National Renewal Energy Laboratory (NRLEL)

4.2.3 Panelist: **Jennifer Kim,** Infrastructure Engineer, Montgomery County Community College

4.2.4 Panelist: **Maria Kalyvaki,** Assistant Professor of Marketing, Minnesota State University, Mankato

4.2.5 Panel Discussion

The panel participants were Brenna Meade, Jennifer Kim, Maria Kalyvaki, and Karen Lopez. All panelists are former Women in IT Networking at SC (WINS)[54] participants. WINS is a multi-year program funded by the National Science Foundation to address the prevalent gender gap in Information Technology (IT), particularly in network engineering and high-performance computing. A collaborative project between UCAR, ESnet and KINBER was introduced in 2015 as a pilot program. Now in its 7th year going into 2023, the percentage of women participating in SCInet has increased significantly from 15% of volunteers to 41%. SCinet women in leadership has grown from 8 to 15%.

First question: What about your current role initially attracted you to IT and what about it will likely persuade you to stay?

After undergraduate study in computer science, Karen tried software engineering, but did not like it. When she started working with automation, that is when she became interested in infrastructure, seeing how everything was working. She likes working in the R&E community, facilitating research, merging the best of two interests: coding and infrastructure.

Jennifer was a non-traditional student. She picked the networking certificate program because that was a track leading directly towards a Cisco CCNA. She enjoys the work and feels she is making an impact at community college level, not just for the campus but for the students that are similar non-traditional students trying to find their way.

Maria is also not a traditional IT person. Her bachelor’s degree is in agricultural engineering from Greece. She started programming automation for greenhouses for different types of equipment and that drew her into Information Technology. Later, she completed several other master’s degrees that required some programming including instruction technology. She realized she enjoyed being in the

[54] The Women in IT Networking at SC (Women in IT Networking at SC • SC23 (supercomputing.org))

STEM area advocating for other women because when she was in Greece there were only a few women in the agriculture engineering program. That made her question why there were not more women who wanted to be part of the program. In her role as a professor, she advocates for the students to explore career options.

Question 2: Based on your organization and demographic are there any diversity efforts you can describe as successes?

Karen's team has grown from one woman to three. Her organization has been working on diversity and her management has been very supportive.

In the context of Montgomery County Community College, 16% of the student population are women in computer science majors and certificate programs and of that percentage only 55-65% of those stay with their programs from Fall to Spring semesters. They have found that degree and certificate completion for women is even worse, with only 11-21 percent over the past five years completing their program. In response to this, they have created a group, Women in STEM (WIST) primarily focused on CIS and Technology majors to help provide some guidance on what options are available, and to help students not feel alone or siloed. These female students have a support group to help them persist.

At Minnesota State, there are different efforts, including providing scholarships, to recruit students from different minorities as well as international students. There are various committees such as one that supports women's professional development. They are trying to get some grants to help students and faculty to do more research about equity and diversity at the university.

With WINS, some of things that they do to keep retention and the program going are events at SC with other WINS women including with mentors. There is an alumni group that meets four times a year to talk about jobs, resume development, and advice. They are striving to build a community that can rely on each other. Today there are 40 women in that group. WINS is not only a wonderful group of women who are doings things that you do but it is an incredible hands-on experience. You can work with some of the greatest engineers in the R&E community. The professional development you receive is immense and the community that you build not only within the WINS group but within the R&E community is one of the reasons the program has been an incredible success. For young professionals this is a good place to find support. WINS is like a sorority – once you graduate you are still part of the community, hard for women to be in engineering, having someone who can relate with you is important.

Question 3: What are the biggest challenges and barriers to engaging your underrepresented groups in IT?

One barrier is visibility. There are so many different titles and paths within IT and computer science. Many women have fallen into the network engineering role unintentionally.

Another challenge is to determine reasons why students, particularly non-traditional students and women are not persisting in their studies. What do they lack? What support do they need? It is not always easy to get the answers when the students are not engaged. Trying to individualize support helps build rapport and helps serve the student best.

Communicating what IT is and how interdisciplinary it is can be difficult. Many students do not understand by getting a degree what they are going to be doing. Unless they know someone, who is in the field, they do not know what to expect, what the work entails, the different tasks they will do, whether the work is easy or hard for them. Every job now has some aspect of IT.

Once they find a position, they need support from their organization and a good work environment. Good supervisors let their people go to conferences, explore these support groups, and further their careers with professional development.

Question 4: What is one thing as a community we could do to make IT more accessible or desirable to underserved communities?

With underrepresented communities, it is hard and overwhelming to go into a space where you are the only minority. A community of people with similar ideologies or similar issues is one way to make IT more desirable. To retain people in IT we need to create a community for the underserved to feel they belong and there is a space for them in IT.

Mentorship and the ability to get hands-on-experience are important. SCInet gives a lot of experience. Giving young professionals hands-on-experience.

Try to find additional ways to expose the options and opportunities that are available. More intentional predetermined events may give this exposure.

Question 5: What impact has the emergence of remote or hybrid workplaces and events had on engaging underrepresented populations?

Remote work can help work-life balance. Remote work may allow professionals to continue their education. People can join events from anywhere, which makes them more accessible. On the flip side, hybrid remote work can sometimes keep people from engaging.

During the pandemic, online options were greatly beneficial to international students and others who needed to work while workplaces were shut down. Online meetings can be more productive.

Economic barriers such as travel to conferences and workshops benefit from the online options which are reduced in cost and in some cases free.

Question from audience (John Graham)***:*** How are you reaching out to the undergraduate students? Can you get grants?

WINS was designed as a program for young professionals. There is a potential for taking this pilot program and expanding it to undergraduates. SC does have a student volunteer track. There have been workshops for undergraduates and high school students.

Question from audience (Larry Smarr)***:*** Given that SC is a once-a-year event, and NRP is distributed across the entire country and is available 365 days a year, perhaps the S in WINS could be expanded to distributed computing. There might be opportunities to partner with things like NRP.

This might be the next step, but the most difficult thing is not the funding, but who is going to do the work and handle operating everything including logistics. This group is open to suggestions and collaboration.

The panel discussion is available at: [2023 4NRP - Brenna Meade - YouTube]

4.3 Panel: Reaching More Minority-Serving Campuses

4.3.1 Chair: **Richard Alo,** Dean College of Science and Technology, Florida A & M University

Reaching More MSI Campuses: Why?

How do we reduce the vast underrepresentation in STEM? Collaboration will help us all to equalize the playing field. 2019 data shows that the US population from ages 20 to 34 is forty-seven percent minority. China that has four times the US population is ready to pass the curve in the production of PhDs. From 2016 to 2021 our population grew, but our production of PhDs fell in math, statistics, and computer science. Despite all our efforts, we have not moved the needle enough.

Black and Hispanic workers remain underrepresented in the STEM workforce. Enrollment in STEM fields dropped 14.3% for Black men and 6.9% for Black women during the pandemic. We need to start looking at advancing not only Black women in STEM, but also Black men.

Minority Serving Institutions (MSIs) are Hispanic Serving Institutions (HSIs), Historically Black Colleges and Universities (HBCUs), and Tribal Colleges and Universities (TCUs). A relatively small percentage of MSIs serve a much greater proportion of underrepresented minority (URM) students. Engaging them is an efficient way to broaden the participation of our next generation of scientists and engineers.

How does the R&E ecosystem benefit from URMs inclusion? With declining enrollments, we must be innovative, and we must provide significant programs to increase the success of diverse underrepresented minorities in STEM. We are now in the world of digital professions such as cybersecurity, additive manufacturing, architecture, data science and so on. We are also in the world of cyber-enabled science, where we talk about merging market science with domain science. The workforce is demanding that we pay attention to this. Educational institutions must promote racial and ethnic minority students' educational attainment in STEM.

There are opportunities for the many regional networks to step in to help as suggested by the missing millions report such as providing access to resources, opportunities, and needed faculty development. But without the ability to use and execute the resources, the impact will be minimal. See the presentation slides for suggested opportunities and strategies for regional networks.

The Economist article "Creative Destruction"[55] tells us why we must reinvent higher education. It promises a better, more affordable education for many more people. The NRP helps with breaking down the wall between research and teaching. NRP provides new learning environments and modalities. It is hands-on experiential learning, providing much needed engagement.

An example of this is the HyperWall tiled display where students at Navajo Tech and University of Houston collaborate using the visualization walls to study the Mars surface.

---

[55] Creative destruction (economist.com)

Another example is undergraduate students from various disciplines at Elizabeth City on the East Coast are studying the melting glaciers in Greenland.

The FAMU CST research and learning lab collaborating with USC Viterbi School of Engineering bring students together in a flipped global classroom over the cyberinfrastructure from 22 different countries.

The NRP is a significant key and foundation for reinventing the university.

This presentation can be viewed at: [2023 4NRP - Richard Aló - YouTube]

4.3.2 Speaker: **Deborah Dent**, CIO, Jackson State University
Reaching More MSI Campuses: an HBCU Experience

Jackson State is an urban university in Jackson, Mississippi with 7,000 students. It is the only urban campus in Mississippi. They have seen decreasing student enrollment due to the pandemic and water shortages. They do have research going on across the campus, but it is in silos. They are working on their strategic plan and would like to move from an R2 to an R1 institution within the next 5-10 years. To do this, they need a very good CI infrastructure. Cybersecurity and Cyberinfrastructure education and research will be interwoven throughout these focus areas. When Coach Prime came to Jackson State, they got unwanted attention from cyber-attacks due to the national and international exposure.

Initially they planned to train the Enterprise team to support the research side and that is not working out as planned. Their CI roadmap started in 2013 with discussions on what needed to be done. They developed a CI plan under the guidance of an NSF Empower Grant. In 2018 they were awarded a CC* award to upgrade their network to 100G and established a science DMZ. They became a part of the MS-CC also in 2018.

Limited access to external HPC resources is still a challenge. They do have some small HPC Clusters in silos. Their target is to get access to cloud resources for HPC and CI engineers. They will utilize the MS-CC Tiger Team and work on getting access to additional computing resources.

This presentation may be viewed at: [2023 4NRP - Deborah Dent - YouTube]

4.3.3 Speaker: **Al Anderson,** CIO, Salish Kootenai College
Reaching More MSI Campuses: a TCU Experience

Tribal colleges are weird little entities because they are community colleges on one hand but striving to get into more four-year programs and even master's programs. They tend to have smaller groups of students, sometimes as little as six or eight students in a classroom. There are reasons they exist. Tribes are exercising their sovereignty to provide education for their community. The tragic history of Native Americans has led to are trust issues and deep generational trauma. Most TCUs started in the 1970's and 1980's. Academic programs are varied. Typical cyberinfrastructure at TCUs consists of basic infrastructure, data storage, computing services, and a high-speed campus network.

They face similar challenges to mainstream colleges and universities, but their locations are out in the hinterlands. There can be poor internet connections and difficulty finding professionals to support IT services.

In terms of collaboration with each other, all tribal colleges belong to American Indian Higher Education Consortium (AIHEC). Started as a lobbying group in 1973, it recently started branching out to help colleges get their infrastructure going. One such project was the Advanced Networking for Minority Serving Institutions. The new effort is the CI effort that was started a couple of years ago. AIHEC and some CIOs have been going around to different tribal colleges and helping them figure out what needs to be fixed. They came up with the following list:

- Not enough IT strategic planning,
- IT staffing storages,
- faculty LMS training,
- wireless coverage,
- technical equipment refresh rates,
- internet connectivity,
- centralized authentication,
- IT professional development,
- enterprise resource planning systems, and
- fiber-optic cabling issues.

Using MS-CC and AIHEC they are working through these issues.

There are resource challenges with budgetary limitations which lead to TCUs prioritizing other campus needs over those of the IT department. This in turn impacts technical IT infrastructure issues. Funded from the federal government directly, campuses are without funds to run the campus if the federal budget is delayed.

The MS-CC is something new. It is important for the TCUs. There have been five proposals put in towards the NSF Pilot program (PoCG) under the second MS-CC grant. In order to solve the problems TCUs, have, the key is shared resources: people and hardware. NRP fits right into the development of and access to shared technical resources such as clusters, cloud storage, training resources, people etc. Providing faculty the access to resources is pretty easy. It is the engagement piece that is more difficult. Training would help with this.

Specific to Salish, they are looking to come up with a Cyberinfrastructure plan and need to develop better internal relationships with faculty.

This presentation can be viewed at: [2023 4NRP - Al Anderson - YouTube]

4.3.4 Speaker: **Samuel Sudhaker**, CIO California State University San Bernardino Reaching More MSI Campuses: An HSI'' Experience, High Performance Computing at CSUSB

Cal State San Bernadino (CSUSB) is one of the top five institutions in the CSU for Faculty Led Research. They received an R2 Carnegie designation in 2022. Most of their research projects did not require high performance computing, but when it was needed, faculty would garner resources from their colleagues and partner institutions.

CSUSB discovered they needed to find resources for faculty who need high performance computing, but they did not have funding or the expertise to build the infrastructure needed. They experimented with building an HPC cluster with hopes of securing NSF funding to build an infrastructure. Over the years they explored partnering with XSEDE, but resources did not quite match up with faculty and researcher needs. They bought HPC resources from the San Diego Supercomputing Center. In 2017, CENIC introduced them to PRP. They worked with their ITS team to create a science DMZ and connected to the platform. They upgraded their connection to CENIC to 100G. They have a small team, but you do not need to have a large team to benefit from the NRP.

They want to get all CSU campuses involved with NRP.

The presentation can be viewed at: [2023 4NRP - Samuel Sudhaker - YouTube]

4.3.5 Speaker: Jill Gemmill, Executive Director, Research Computing, Clemson University
The R1/PWI Experience: Road to Empowerment is Slow, Steady and Deliberate

Why does a primarily white institution want to help local HBCUs and minority serving institutions? A lot of the why is driven by regional things. Clemson is in an EPSCOR[56] state which supports single point submission for collaboration inside the state. Clemson is interested in increasing diversity. The superstars for the next generation may be coming from the minority serving institutions.

In 2009 NSF awarded a grant for "Collaborative Research: An EPSCoR Desktop to Teragrid Ecosystem." The program was to increase use of HPC resources from local to national levels as appropriate. Looking for groups that did outreach to MCIs, Jill discovered a group called MSIC and ended up being connected to Richard Alo. Jill contacted Joey Brenn and Nick Panaslk of Claflin University[57] about HPC resources that would be available to them, but they needed a campus connect to the state C-Light to make use of HPC resources.

Under NSF OAC 1440659, C*IIE Region: Southern Partnership in Advanced Network (SPAN), a program of workshops, site visits, and best practices documentation to administrators, campus IT groups, and researchers was offered in partnership with Internet2, SoX and ESnet. In 2017 they came up with a new model for more effective engagement of HPCUs in the cyberinfrastructure and faculty under NSF ACI 1659297. They brought CIOs and faculty together, helped them write campus CI plans, and then submitted proposals to the CC* program. Several NSF grants came out of this: NSF OAC 1827127, NSF OAC 1827098, and NSF OAC 1925641.

After that series of experiences, they explained the process to other HBCUs, and held workshops. Setting up a timeline, "Slow, Steady, and Deliberate," there is an inflection point where outreach turned into people receiving their own funding, taking leadership in their own area, and facilitating a community that could move that forward.

This presentation can be viewed at: [2023 4NRP - Jill Gemmill - YouTube]

---

[56] https://new.nsf.gov/funding/initiatives/epscor
[57] https://www.claflin.edu/

4.3.6 Speaker: **Ana Hunsinger**, Vice President Community Engagement, Internet2 MS-CC: Advancing CI @ HBCUs, TCIUs and other MSIs

Awarded in 2021, the NSF CI Center of Excellence Demonstration Pilot created the Minority Serving Cyberinfrastructure Consortium (MS-CC)[58]. The vision of the MS-CC is to promote CI capabilities with HBCU, HIS, TCU and MSI campuses. It exists to increase access to CI capabilities, enhance communication between researchers, CI professionals and campus leadership, support CI-enabled professional and career development.

Driven and guided by the Consortium Leadership Council (CLC), they meet every week, except holidays. They are: Dr. Richard Aló, Florida A&M University • Al Anderson, Salish Kootenai College • Jim Bottum, American Indian Higher Education Consortium • Joey Brenn, Claflin University • Dr. Ming-Hsing Chiu, Dillard University • Bobby Clark, Clemson University • Dr. Damian Clarke, Alabama State University • Dr. Deborah Dent, Jackson State University • Tom Jackson, North Carolina A&T State University • Leah Kraus, North Carolina Central University • Dr. Kylie Nash, Alabama A&M University • Dr. Adebisi Oladipupo, Morgan State University • Dr. Urban Wiggins, University of Maryland Eastern Shore.

The workshops that they have conducted working closely with the leadership at the campuses. The community keeps growing.

Challenges that HBCUs and TCUs face include how to leverage research infrastructure to accelerate science, how to train and scale support staff to meet criteria for grants, and how to access resources such as IAM, HPC, clouds and other technology tools when they are not readily available on a campus.

A second award will fund activities through 2027 and is a phased approach to engage the Missing Millions. The MS-CC Proof-of-Concept Grant (PoCG) is a novel approach to CI planning and capacity development. It is a combination of funding and dedicated expert facilitation. Subawards will start soon.

This presentation can be viewed at: [2023 4NRP - Ana Hunsinger - YouTube]

4.4 Closing Session: Open Mic for Participants

4.4.1 Chair: **Kevin Thompson,** Office of Advanced Cyberinfrastructure, National Science Foundation

This workshop has been a great learning experience with very impressive content and great content. Amy, Larry, and Frank listed several awards, past and current, that have gotten us to this point.

NSF Award 1826967, Toward the National Research Platform (TNRP) started in October 2018 as part of the CC* program. This award had a key role in scaling up the PRP, with a key component in measurement and monitoring that included instrumentation for performance measurement, a visualization of end-to-end performance, and strengthening containerization capabilities.

The NSF award 0225642, Optiputer, ran from 2002-2009. It got awarded under the Information Technology Research (ITR) program which produced some huge impact award activities. The program was seminal in allowing HTCondor, and OSG to walk a path to where they are today.

[58] https://www.ms-cc.org/

Quoted from the project text, this research "exploits a new world in which the central architectural element is optical networking, not computers." Essentially the Optiputer is a virtual parallel computer in which the individual processors are widely distributed clusters. The "backplane" is provided by IP delivered over multiple lambdas and the "mass storage systems" are large distributed scientific data repositories.

This was a large project that included many partners and participants, several state and regional networks, and NRENs from other countries. The core team included six institutions. There was a list of industry players, including leading edge state-of-the-art optical networking companies and multiple science projects and researchers. At its heart, it was a networking R&D project, an experimental deployment project. One of the outcomes was the development and application of SAGE, a scalable adaptive graphics environment enabling scientific visualization and education applications.

The outstanding teams that are multi-institutional, multi-disciplinary, with diverse expertise and backgrounds drove these projects. It is one of the hallmarks of the activities led by Larry Smarr and Tom DeFanti. A common thread, a central theme, is the power of collaboration. The agenda of this workshop reflects this strong underlying current. It is the spirit of collaboration that will drive innovation and CI and time-to-science. On behalf of the National Science Foundation, thank you Dr. Larry Smarr for showing us the way and your leadership.

4.4.2 Open Microphone

(Larry) Applying the collaboration concept to the previous panel– Of all the universities that have made use of the PRP, CSUSB is the most successful in their method that could easily be put across the California State Universities, two-thirds of which are MSIs or within the MS-CC. They have just four people that are sort of what you were supporting as cyber-engineers in a previous CC* solicitation, but they are helping the faculty to figure out how to use a broad set of HPC resources, including OSG and supercomputer centers, PRP, NRP and the cloud. Just those four people have led to a huge overproduction and use of the PRP by one of 23 CSUs. There is nothing like it across the entire country. It is a model of a successful engagement. Perhaps it could be extended to the entire CSU system – take a model and move it out of the one campus and share with others. The same is true for the MS-CC.

It is sub-critical to take a campus that is already under-resourced and try to get them to get their faculty involved. It does not do any good to have a DMZ or network connections if no one uses it to do research. That is the nut that must be cracked - how to get the faculty to use the resources.

The relationship between CIOs and faculty is not strong and is sometimes problematic. If you get these "glue people" to show faculty how to use the resources. CSUSB uses more GPUs than seven of the UC campuses. You get vastly more leverage using the glue people.

(Kevin) Partnership between campus IT and researchers and educators. It is a requirement for every NSF award.

(James Deaton) Concurs with Larry.

(Kees) Remind everyone from international point of view that we have been able to collaborate effectively with the US internationally using International Research and Education Network Connections (IRNC)[59] program, have extended Optiputer to Europe, take away boundaries, walls, there is no

[59] https://new.nsf.gov/funding/opportunities/international-research-education-network

impossibilities, just another challenge, think what you can do if there are no limits. CineGRID, OnFactor, bringing telecom, networking, and research together.

(Harvey) Global perspective, high level of activity from industry (optical, wireless), end of CMOS, industry is working towards that, societal changes with truly intelligent networks/technology. Trying as a bigger community to bring forward the leaders of tomorrow. Invite people to join the GNA-G. When you promote people, it is not just to bring them up to speed, but to make them future leaders.

(?) A couple of talks mentioned staffing. Where do you see the NSF fitting in, to support all this staffing? How can you encourage training and professional development?

(Kevin) Programmatically there is Cyber-training to address this[60]. There are not easy solutions. The underserved institutions need to look with business perspective to tap students, pay them, put them on help desk, let them learn some of the operational elements.

(Jill) The idea of sharing staff can work but has some challenges. It may work better to have some people embedded in R1 or R2 institutions.

(Inder) Hope that collaboration extends and can be taken worldwide.

(Doug) Concerned that students are not coming into infrastructure to train. There are less students from which to pull.

(Jen) Like the idea of shared staffing. Many of the campuses that the MS-CC is serving have been left out of a lot of the collaborations. Please find a way to collaborate with institutions that are different from you.

(Jeffrey W) Sharing staff and expertise is easier when you are sharing the same infrastructure. Think about shared services, think about working with other institutions.

(Curt Dodds) How about pulling together domain scientists to form groups supporting the use of NRP and other shared resources? Use existing domain champions to build community. Curt commits to creating an Astronomy group for NRP.

(Dong) Would like to honor PRP/NRP staff who made their research possible at CSUSB. They stand on the shoulders of the giants. Encourage the MSIs to do it.

(Rod Wilson) Industry is an important player in NRP. We learn a lot from this community. At the ESnet-6launch, there were five industry partners. They spend 6.7 billion dollars a year on R&D. This group can tap into some of those investments.

This presentation and the Open Microphone session can be viewed at: 2023 4NRP - Kevin Thompson - YouTube

---

[60] Strengthening the Cyberinfrastructure Professionals Ecosystem | NSF - National Science Foundation

5.0 Acknowledgements

5.1 Program Committee

- Maxine Brown (University of Illinois at Chicago)
- James Deaton (Great Plains Network)
- Marla Meehl (UCAR)
- Lauren Michael (Internet2)
- Subha Sivagnanam (SDSC)

5.2 Local Arrangements

- Susan Rathbun
- Lara Sievert
- Maximino Carreo
- Lovella Cacho
- Milton Szyperko-Zuniga
- Crystal Liu
- Hector Bracho
- Charles Erwin

5.3 Special Thanks

- The National Science Foundation award #1826967
- Albert P. Pisano, Distinguished Professor and Dean, Jacobs School of Engineering, UC San Diego
- Ramesh Rao, Director, Qualcomm Institute, UC San Diego
- Frank Würthwein, Director, San Diego Supercomputer Center, UC San Diego
- Rod Wilson and Marc Lyonnais, Ciena
- Ryan Emmerich, Applied Data Systems