First Global Research Platform Workshop

September 17-18, 2019

Hosted by Qualcomm Institute-Calit2
University of California, San Diego

Chair: Joe Mambretti
Senior Advisor and Host: Larry Smarr
GRP Coordinator: Maxine Brown
Local Arrangements Chair: Tom DeFanti

Report by Richard Moore
Editing by Maxine Brown and Tom DeFanti

Table of Contents

1   Introduction

International scientific initiatives collaborate on building, accessing and analyzing data from one-of-a-kind advanced instruments in unique locations around the globe, connected to remote computational, data storage, and visualization systems at speeds of Gigabits and Terabits per second. The Global Research Platform (GRP) is an evolving effort focused on design, implementation, and operation strategies for next-generation distributed services and network infrastructure on a global scale, including interoperable Science DMZs, to facilitate data transfer and accessibility.

The GRP is a natural evolutionary step forward, rooted primarily in three existing initiatives with overlapping participants and several macro trends in evolving cyberinfrastructure based on recent innovations in architecture and technologies. First, the Global Lambda Integrated Facility (GLIF), is an international consortium that promotes the paradigm of lambda networking.[1] The GLIF participants are National Research and Education Networks (NRENs), consortia and institutions working with lambdas. GLIF participants provide lambdas internationally as an integrated facility to support data-intensive scientific research and middleware development for lambda networking. It brings together some of the world's premier networking engineers who work together to develop an international infrastructure by identifying equipment, connection requirements, and necessary engineering functions and services.

Second, the US National Science Foundation (NSF) is building on its major investments in the "CC*" program[2] to upgrade campus networking for science data access by funding an award to UC San Diego and UC Berkeley to establish the Pacific Research Platform (PRP), a science-driven high-capacity data-centric "freeway system" on a large regional scale. Its goal is to give data-intensive researchers at participating universities and collaborating institutions the ability to move data 1,000 times faster compared to speeds on typical inter-campus shared Internet. The PRP does this by federating campus Science DMZs[3], a concept developed by ESnet in 2010 and adopted by the NSF in its CC* solicitations, into a regional Science DMZ. Documentation, including reports and materials resulting from all PRP workshops, are published.[4]

Third, when the PRP was proposed and awarded, it was anticipated that successes and lessons learned on a regional scale would inform extensions to a national scale. The National Research Platform (NRP) in the US has been explored in a series of annual workshops[5] and pilot projects[6]. The workshops have involved diverse stakeholders, including domain scientists, network and system administrators, campus CIOs, regional network leaders, and representatives of ESnet, Internet2, the Quilt, XSEDE, and NSF, as well as international networks and universities.

---

[1] See GLIF website https://www.glif.is/.
[2] See, for example, the history of Campus Cyberinfrastructure (CC*) programs.
https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504748&org=OAC&from=home .
[3] See http://fasterdata.es.net/science-dmz/.
[4] See PRP website http://www.pacificresearchplatform.org.
[5] See first and second NRP workshop reports at http://www.pacificresearchplatform.org/reports/. See information for third workshop at https://meetings.internet2.edu/2019-924-third-national-research-platform-workshop/ .
[6] For example, see https://meetings.internet2.edu/media/medialibrary/2018/10/16/20181016-moore-deaton-nrp-pilot.pdf

Several pilot projects engaging different campuses and regional networks are actively being pursued. Internationally, there are evolving efforts by the Asia Pacific Research Platform[7] as well as in Europe and South America focused on these issues.

Macro trends are also driving changes in infrastructure, providing higher-level abstractions for functionality. The virtualization made possible by these changes provide opportunities for migrating from static "one-size-fits-all approaches" to dynamic capabilities for real-time responses to conditions and for high degrees of customization. This trend enables infrastructure environments to be "sliced" into segmented resource integrations, each with their own attributes.

This synthesis of efforts, from the international networking experience of GLIF participants, to the regional and national research platforms being developed in the US, Asia, Europe and South America, brings us to this initiative for the GRP. This First GRP Workshop highlights global science drivers and their requirements, high-performance data fabrics and distributed cyberinfrastructure, including advanced networks customized to support scientific workflows. The Workshop brings together researchers, scientists, engineers, and network managers from U.S. and international research platform initiatives, to share best practices and advance the state of the art.-It is clear that a successful GRP will require a successful intersection of two "networks" – not only high-performance data networks, but also productive human networks among diverse stakeholders.

Ninety-eight people registered for the workshop and 94 were able to attend, with 13 countries represented (Australia, Brazil, Canada, Czech Republic, Denmark, Germany, Japan, Korea, Netherlands, Poland, Singapore, Taiwan, US). A complete list of registrants with affiliations is provided in the Appendix to this report.

---

[7] https://apan.net/wg/aprp .

Figure 1: Group Photo of First GRP Workshop Attendees

High-level descriptions of the presentations and discussions are provided chronologically in Section 2 of this report. These notes are not intended to replicate the presentation material and should be read in conjunction with the presentation slides. Findings and Recommendations are summarized in Section 3, followed by Acknowledgments. Four appendices provide information on accessing the presentation material, the initial workshop agenda, a list of workshop participants, and biographies of speakers and program committee members.

2   Workshop Sessions and Discussions

The day-and-a-half workshop included three keynote talks (by Larry Smarr, Tomohiro Kudoh and Inder Monga) at the beginning of each half-day session, six panel sessions with 3-6 speakers on common themes, one round of lightning talks, an evening poster session with 10 posters/ demonstrations, and a final wrap-up discussion. The panel sessions covered topics such as science application drivers, existing networking initiatives and research platforms, data transfer techniques, and networking technology. Ample break time was allocated to enable the social networking that is key to the success of workshops.

2.1   Welcome and GRP Overview – Joe Mambretti, Northwestern Univ. (NU)

The Workshop Chair, Joe Mambretti, opened by thanking the audience for attending the launch of this initiative to build a global fabric for data-intensive science. The most important resource for this initiative is people, with many of them in this room. He then thanked Larry Smarr and Tom DeFanti as the hosts who have made this workshop a reality, Rod Wilson from Ciena for

sponsoring the evening reception, and JJ Jamison of Juniper Networks for sponsoring the writing of the workshop report.

Mambretti then introduced the first keynote speaker, Larry Smarr, by noting that he has been creating cyberinfrastructure for computational science for a long time and "needs no introduction."

2.2   Keynote – Research Platforms: Past, Present, Future – Larry Smarr, Calit2/UC San Diego

Larry Smarr greeted the participants – "this is a big international family." His talk reviewed his journey of network-related research over the last ~20 years, with an emphasis on the global context, concluding with the PRP and its extensions nationally and globally.

This journey started with the OptIPuter project[8], a large NSF grant from 2002-2009. The project's focus was to explore changing the central element of cyberinfrastructure from computers to networks, with the driving question being: "How can we take advantage of the amazing bandwidth of optical networks?" Many of the things that are being doing now were present as elements of OptIPuter, including international collaborations with Japan, Australia and other countries. Several NSF-funded projects followed OptIPuter, including Quartzite and PRISM[9], which built out a high-speed research network on the UCSD campus. UCSD's experiences provided input to the Campus Bridging Task Force report[10] to NSF, and this report, along with the Science DMZ concept developed by DOE's ESNet[11], were key elements in NSF's creating the Campus Cyberinfrastructure (CC*) program to upgrade campus networking capabilities across the US. The NSF has made over 250 CC* awards from 2012-2018, and this massive investment has transformed the digital fabric for university research in the US.

While the CC* program created a large number of high-bandwidth Science DMZs at individual campuses, the next step was to *connect* the individual campus Science DMZs and ensure that users could easily transfer data end-to-end from labs on one campus to labs on another campus. That was the primary objective of the PRP (see figure below), which is now entering its fifth year. The program was built on a large number of specific science application drivers, along with the substantial political work of getting support from the CIOs and other stakeholders across PRP's partner campuses. CENIC has also played a key role in providing the high-bandwidth connections and working closely with the PRP team and partners. The initial technical focus was on developing and deploying Flash I/O Network Appliances (FIONAs) as Data Transfer Nodes (DTNs) in the Science DMZs, and making sure that those nodes could keep up with the 100 Gbps networks. The second technical step was to deploy diagnostic tools (e.g., Monitoring and Debugging Dashboard or MaDDash displays) to monitor end-to-end network transfers and work through the numerous potential impediments along the paths between PRP partners.

---

[8] See www.optiputer.net .
[9] See http://prism.ucsd.edu/ .
[10] Available at https://www.nsf.gov/cise/oac/taskforces/TaskForceReport_CampusBridging.pdf .
[11] See https://fasterdata.es.net/science-dmz/ .

Source: John Hess, CENIC
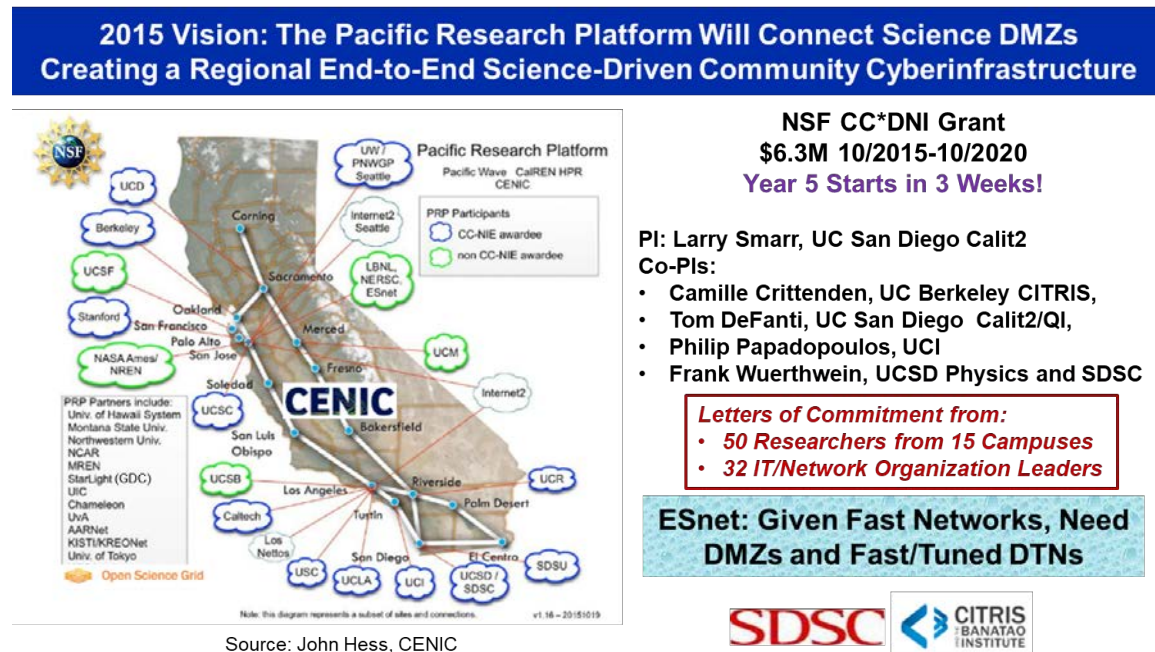
Figure 2: Graphical Overview of the PRP

While the word Kubernetes was never mentioned in the PRP proposal, it became a game-changer for the project when Google open-sourced the software several years ago. Kubernetes is an open-source system for automating deployment, scaling and management of containerized applications, and it was built by Google to operate at planetary scale. Essentially all cloud vendors adopted Kubernetes, and it quickly transformed how PRP operated with its partners. The individual systems that initially acted as Science DMZ DTNs and perhaps as local compute systems are now part of a hypercluster available to anyone on the network. Storage was also part of this revolution, and Rook[12] allowed storage to be coordinated with Kubernetes. John Graham on the PRP project has developed the Nautilus hypercluster across the PRP partnership, and it now includes systems on 15 campuses, with 4360 CPU cores, 134 hosts, 407 GPUs, and 1.7 PB storage. The Nautilus hypercluster has also been expanded to include systems on regional US networks like GPN, LEARN, MREN, KINBER, and NYSERNet, as well as the Open Science Grid (OSG) and national NSF supercomputers (see figure below). And globally, it has expanded to include systems at KISTI, Guam, U Queensland, and U Amsterdam. In addition, Kubernetes simplified moving software containers from the PRP into commercial clouds; Igor Sfiligoi (UCSD) is working on this and had a poster at the evening reception.

---

[12] rook.io/ .
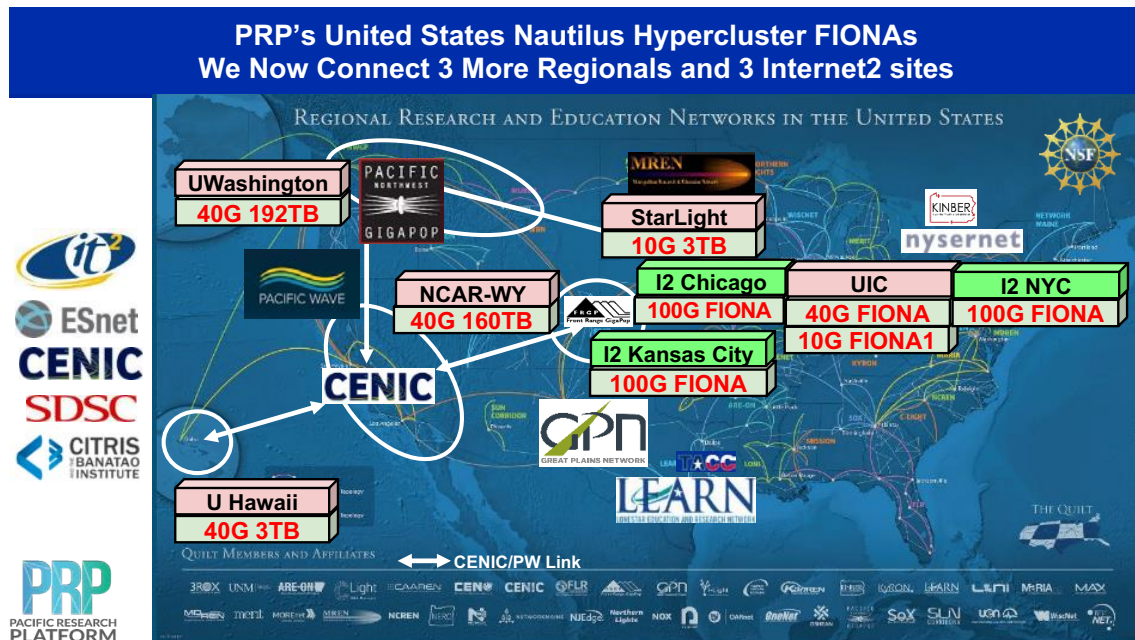
Figure 3: Expanding the PRP towards a National Research Platform

With Kubernetes, the Nautilus hypercluster is very flexible and can host a wide variety of computing architectures. There are various flavors of single-precision and double-precision GPUs, including tapping into systems on UCSD/SDSC's Comet supercomputer and AIST's 'AI Bridging Cloud Infrastructure (ABCI)' supercomputer in Japan. There are plans to host FPGAs (e.g., by Xilinx, Micron) within the PRP, as well as Google's Cloud Tensor Processing Units.

While the project has made great progress in the technical networking/hypercluster arena, it remains rooted in scientific applications, and a number of examples were detailed of how scientists are using the PRP. A project with Scripps Institute of Oceanography and UC Irvine had >500X speed-up in their workflow. Chris Paolini of San Diego State University has used OpenMPI on resources distributed across the PRP for his application, including UCSD/SDSC's Comet supercomputer and systems at KISTI in Korea. One of the major users of GPUs on Nautilus has been the IceCube project, which runs simulations of an ice scattering model to assist in determining directions for neutrinos. Other heavy users of Nautilus GPUs are artificial intelligence researchers working on robotics, autonomous driving, deep learning, reinforcement learning, computer graphics, machine learning, and biological quantitative modeling.

Why is all this important? Projects like the SKA or LSST are global projects that have to move huge amounts of data for processing to many different locations. "What we've prototyped with PRP is essential for these classes of instruments." There is a need to establish global platforms, including by people in this audience and beyond, and to build the digital infrastructure necessary to advance data-intensive science.

Question & Answer Discussion

Bill Johnston (ESNet): Is the PRP operating at Layer 2?

Larry Smarr: No, this is all Layer 3 routing. Layer 2 would have been too labor-intensive. It was amazing that it worked at Layer 3, but it's worked out beautifully.

2.3 Panel 1 Theme: Global Research Platform Cyberinfrastructure – Moderator: Shawn McKee, University of Michigan

This opening panel included four talks describing current cyberinfrastructure programs, at campus/regional/national/international levels, that are foundational for addressing various elements of a GRP.

2.3.1 Science DMZ Global Considerations — Tom DeFanti, UC San Diego/QI-Calit2

Tom DeFanti described the technical paths that the PRP project has taken over its four-year history, and looked ahead to more global participation in the platform.

Phase 0 (pre-funding) of the PRP project was in early 2015 and indeed, as Bill Johnston just asked Larry Smarr, it was done as a demonstration of capabilities at Layer 2. In the PRP's conception, Ron Johnson (U Washington) advised that we should operate the PRP at Layer 3 – we have and it's worked.

Phase 1 of the project, from 2015-2017, focused on moving to Layer 3, developing and deploying FIONA boxes as DTNs, working through the myriad of challenges to routinely get line bandwidth with essentially no TCP backoff on long-distance best-effort networks (i.e., "going green" on MaDDash plots), and recruiting scientists on PRP partner campuses to use the newly-capable system.

Phase 2 transitioned emphasis from network diagnosis to application support. The PRP award from NSF is a data grant, not a networking grant, and scientists need more than bandwidth/ latency tests. The project also decided to leverage all the FIONA boxes sitting at the border routers beyond just their data transfer function, and, for example, added a large amount of temporary storage to the systems.

As Larry Smarr described, Kubernetes was a game-changer for the program, and Kubernetes has been used to couple all the PRP FIONAs into the Nautilus hypercluster. Kubernetes enables central management of distributed systems, as long as there is a trust relationship, and it allows scaling to heterogeneous platforms – e.g., to commercial clouds and non-CPU systems. With the flexibility of management and usage policies offered by Kubernetes, a number of partners have contributed their resources to Nautilus, leveraging the grant-paid resources by a factor of four … for "pot-luck supercomputing."

Various monitoring and diagnostic tools have been developed and/or adopted by the program. A new traceroute visualization is shown below and was demonstrated at the poster session by Dmitri Mishin, as well as a mock-up of the 'NOC of the Future' developed with EVL's SAGE2.
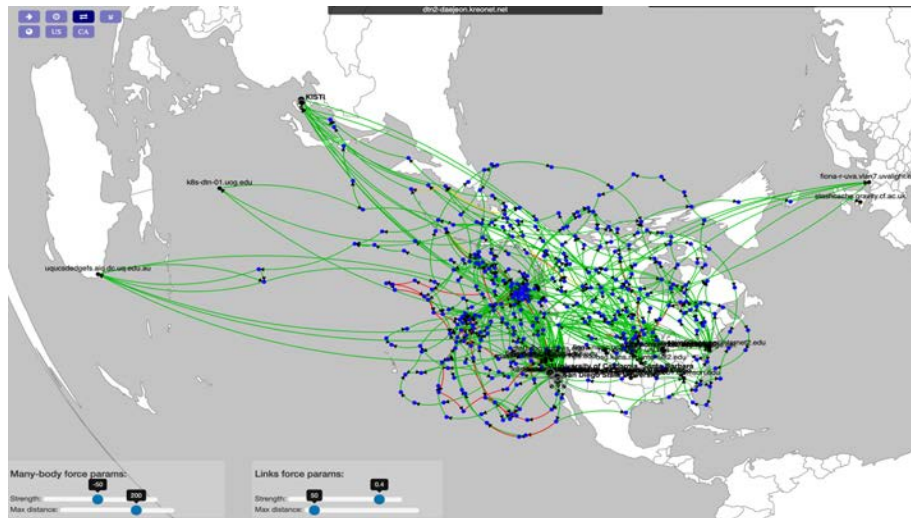
Figure 4: International Traceroute Visualization

Below are the bullets from DeFanti's talk about what the PRP has learned and its extensibility to a GRP:

Going Global in Manageable Ways: The Experiment and the Challenge
- Great Networking with 10-100Gbps Science DMZ Performance is a Necessary but not Sufficient Condition to Enable Data-Driven Researchers
- They need Science DMZs & DTNs with Low-Cost Storage, Encryption, Large RAM CPUs, GPUs, TPUs, FPGAs, Sensors, and High-Availability Computing
- Measuring and Monitoring at all Levels is Key to Better Usage and Security
- Compatibility with Google, Microsoft, Amazon Clouds, and NSF/DOE Supercomputers Helps Ensure Scalability and Continuation—CloudBank
- Kubernetes makes it a sane and extensible platform
- Open Science Grid and Internet2's NRP Pilot Brings in Global Experience
- More Global Partners are Welcome to Join our Pot Luck Supercomputing

2.3.2 Large Hadron Collider Open Network Environment (LHCONE) – Bill Johnston, ESnet

Bill Johnston of ESnet described the Large Hadron Collider (LHC) Open Network Environment (LHCONE), a global overlay network for LHC and the high-energy physics (HEP) community. As a prodigious data-intensive experiment, LHC is a pathfinder for next-generation experiments, such as the Square Kilometer Array (SKA), the Large Synoptic Survey Telescope (LSST) and the Linac Coherent Light Source (LCLS). The LHC networking experience can inform planning for these and other future experiments.

ESnet serves ~25% of traffic to/from the LHC data storage and analysis sites, and LHCONE traffic within ESnet is >1 PB/day. (LHCONE traffic within GEANT is ~1.5 PB/day.)

LHCONE is a 'private' network overlay that connects globally distributed data and compute facilities. It was created as a response to congestion on research and education transatlantic network links due to the onset of LHC traffic in 2010. A Layer 3 Virtual Private Network (VPN) was implemented in 2011 to allow network operators to manage the LHC traffic in their network.

The environment has much in common with a distributed Science DMZ, and the experiences and lessons learned can contribute to the path leading to a GRP.

The scale of LHCONE currently includes 209 IPv6 and 313 IPv4 destinations, 25 NREN providers (including many represented at this workshop, such as CANARIE, ESnet, GEANT, Internet2, KREONET and SINET), and 127 connected sites. LHCONE represents about half the total ESnet traffic (recently ~50 PB/month out of ~100 PB/month total).

The map below shows links, exchange points, provider networks, and locations of routing instances involved in LHCONE. While it is a complicated wiring diagram, it is important that all the participating sites are recognized as being "on the map."
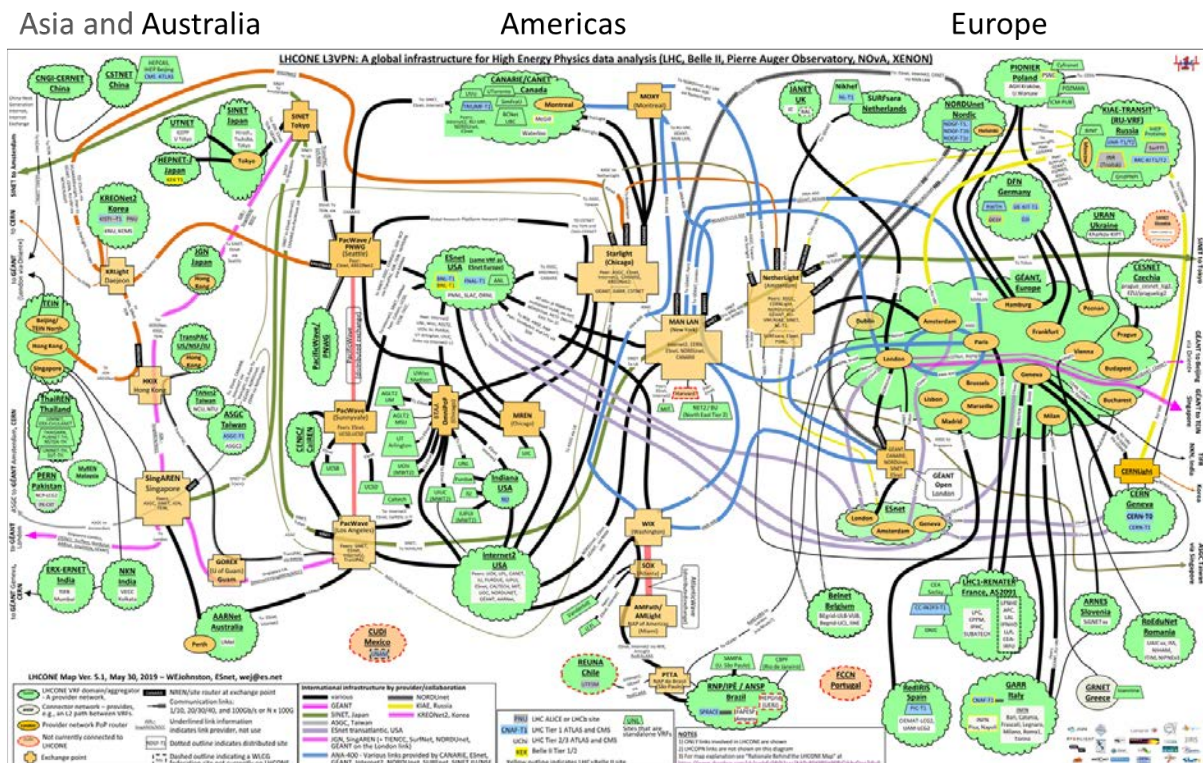
Figure 5: Overview of the global infrastructure of LHCONE

LHCONE uses Layer 3 VPN technology that tags traffic on shared links, allowing network operators to direct LHCONE traffic into parts of their network infrastructure that are not in their core network (e.g., transatlantic links). There is a centrally-managed record of signed Appropriate Use Policy statements (e.g., only physicists in high-energy physics).

In its early days, LHCONE provided for better security than the general internet, which allowed sites to experiment with resources such as Science DMZs and DTNs outside the site firewall. This is less true now, but the fact that the LHCONE community is relatively small allows for the use of Access Control Lists, which are an effective security tool for a Science DMZ.

Some of the key lessons learned regarding why the LHCONE "works" include:
- The stakeholder communities want it to work, including network operators, network politics (e.g., usage policies), and the LHC user community.
- There's a (loose) central management at CERN.
- It uses the same protocols, routing, processes and procedures as the general internet.
- It is significant enough that network provider engineers pay attention.
- It's closely monitored (e.g. perfSONAR and network NOCs).

How does the LHCONE experience inform other science communities on how to overlay networks? There is consensus within LHCONE that it works because it's a "relatively small" community of people with something in common (high-energy physics). There have been discussions of extending LHCONE to support other scientific communities, and several non-LHC particle physics communities are now using LHCONE. However, there are technical reasons why the current size is at the upper end of what this approach will support, and extensions beyond selected physics communities would lose the social advantages of supporting a community with a shared problem. Large-scale projects like SKA or LSST may well consider taking a similar approach, but with an independent network environment.

2.3.3 Nautilus & IceCube/LIGO – Igor Sfiligoi, UC San Diego/SDSC

Igor Sfiligoi has been working with the IceCube and Laser Interferometry Gravitational Observatory (LIGO) experiments to host their applications on the PRP/Nautilus resource[13] via the Open Science Grid (OSG) portal.

IceCube is a neutrino experiment located at the South Pole, using natural ice as the detector media. The primary compute application is a simulation workload running photon propagation (on GPUs), which is crucial for proper calibration and pointing of the instrument. The LIGO instrument has low signal-to-noise ratio and requires significant computer power to filter noise; some parameter fitting workloads are a good match for GPUs.

Both the IceCube and LIGO experiments had already been using OSG for their data analysis, and Sfiligoi's team has added Nautilus as an additional OSG resource that could be available to these experiments (and other OSG users). Nautilus delegates trust to OSG, and then OSG handles job distribution and authentication, authorization, and allocating resources to the experiment users. OSG deployed three layers as Kubernetes pods: an HTCondor batch system, a CernVM File System (CVMFS) driver for software distribution, and an OSG portal into the HTCondor batch system.

Nautilus provides opportunistic use of its resources and jobs can be pre-empted without warning. (The OSG infrastructure provides for recovery of pre-empted jobs.) Even with this pre-emption status, IceCube and LIGO are the largest users of the ~400 GPUs in Nautilus. And, Nautilus provides a significant fraction of the total demand by these applications – ~10% for IceCube and >80% for LIGO.

---

[13] See nautilus.optiputer.net for more info on the Nautilus hypercluster.

The team has worked out multi-level containerization (see figure). With recent Linux kernels, Singularity can run from inside an unprivileged Docker container, like in Nautilus Kubernetes. And the system is flexible – e.g., because of a procurement delay for storage, OSG has been operating a gridFTP server into CEPH storage for IceCube jobs running on Nautilus.

This talk demonstrated the successful interface of the Open Science Grid, with all its infrastructure and broad adoption, to the PRP/Nautilus hypercluster, and that Nautilus can provide significant resources to key experimental compute requirements, even on an opportunistic basis.
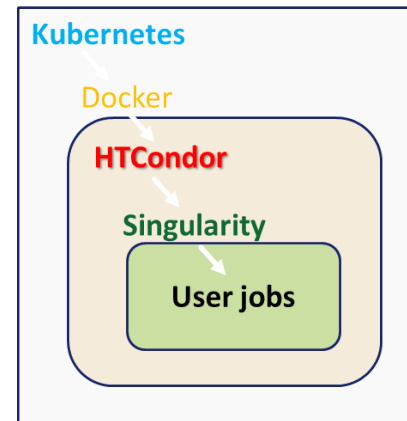
Figure 6: Multi-level containerization

2.3.4 Campus Cyberinfrastructure, as Implemented on a UC campus – Valerie Polichar, UC San Diego

Valerie Polichar, the Director of Academic Technology Services for the UCSD campus, described the historical context for UCSD's campus cyberinfrastructure.

In 2009, a committee of UCSD faculty and technology experts published the "Blueprint for a Digital University" report[14]. Six focus areas were identified, as shown in the figure below.
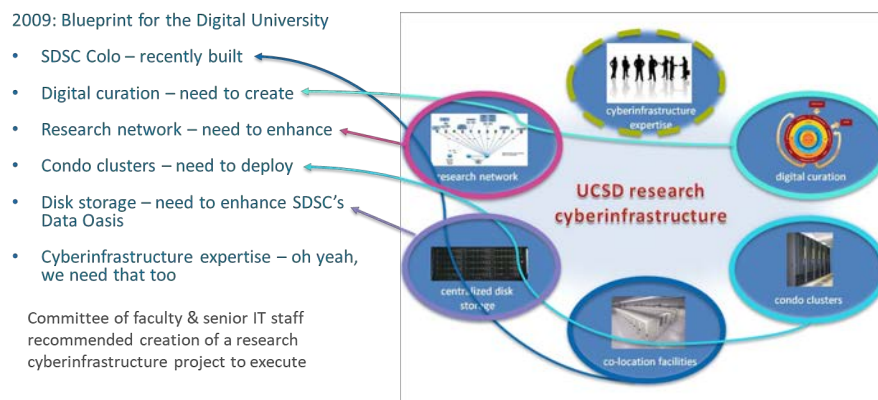
Figure 7: Original 'Blueprint for the Digital University' Concept (2009)

A number of efforts were made to realize this blueprint, building on some existing capabilities (e.g. SDSC's colocation facility) and developing new capabilities (e.g., digital curation by the library). As one example in networking, 10+ Gbps connectivity was extended to all research buildings on campus, and two NSF awards (CHERuB and PRISM) provided external 100 Gbps connectivity to campus and expanded an internal network for research elephant flows. One of the long-lasting benefits of this Research CyberInfrastructure (RCI) program is that it brought together expertise from a number of different organizations on campus – central campus IT,

---

[14] Available at sites.uci.edu/fasrc/files/2015/05/UCSD-Blueprint-for-the-Digital-University.pdf .

SDSC, Calit2, and the Library – and has transformed how those organizations operate in supporting research cyberinfrastructure for the campus. There were some challenges – e.g., a central campus data storage solution was not really solved, providing technical expertise to research groups was a 'late entry,' and better outreach to faculty and researchers was required.

In ~2014, Larry Smarr proposed the new Integrated Digital Infrastructure[15] (IDI) program, which was funded for several years. The focus was to enable transformational digital projects on campus by having technical experts work closely with key research groups on campus to develop semi-permanent shared digital research platforms. A number of high-capacity DTNs were deployed, and the Library expanded its storage capacity for curating research data. The research facilitation program is now permanently funded under a mature program that serves faculty and all researchers on campus, providing "concierge-style" support, technical expertise, and "we'll help create your demo or proof-of-concept" at no cost to researchers.

In summary, UCSD has made great strides in developing a campus cyberinfrastructure, with high-speed networking and a number of critical services available to its researchers. Some of the remaining challenges are to provide easy-to-use storage of many flavors, and storage/compute for different kinds of restricted (e.g., HIPAA-protected) data.

2.3.5 Panel Question and Answer Session

Cees de Laat (U Amsterdam – UvA): (For Tom DeFanti) Is everything in PRP Layer 3? In the early days, we struggled getting data between UCSD and UvA.

Tom DeFanti: PRP is working at Layer 3, but we are still having some problems getting to Amsterdam.

Igor Sfiligoi: The connection is working, but not at optimal performance.

Cees de Laat: (For Valerie Polichar) If one searches for RCI and UCSD, one gets UCSD's Office of Research Compliance and Integrity, not research cyberinfrastructure.

Valerie Polichar: After the RCI program was transformed to IDI, that office officially asked for use of that acronym on campus.

Marek Michalewicz (U Warsaw): What are the similarities and differences between the Chameleon system and Nautilus?

Tom DeFanti: They certainly have different funding mechanisms. Chameleon is a cloud system (led by U Chicago but with a number of other hardware sites and collaborators) and is designed for computer science experiments, perhaps with novel architectures and applications. Nautilus is an outgrowth of the PRP project and is focused on the machine learning community, although it has branched out to other domains.

---

[15] http://senate.ucsd.edu/media/186129/integrated-digitial-infrastructure-initiative.pdf

Michal Krsek (CESNET): It sounds like the IP level we are talking about is default IPv4. But what is the future – when do we go to IPv6?

Igor Sfiligoi: The PRP has both IPv4 and IPv6 already.

Michal Krsek: What is the balance between IPv4 and IPv6? (Neither Igor nor Tom knew off-hand). Also, PRP uses GEO-IP CVN, and there's no reliable GEO-IP on IPv6; we should switch to a more modern technology.

Tom DeFanti: John Graham (UCSD) in the audience is agreeing, and will be glad to discuss this offline.

Richard Moore (UCSD): (For Bill Johnston) What specific elements limit scaling in LHCONE, when you said that you are at the upper limit of what can be done with this approach?

Bill Johnston: From a technical perspective, the key thing is that there currently is no central route repository; routes are communicated informally to the big NRENs. Whether scaling would be fixed by a central route repository is unclear, but that's the current limit. There is nothing in the Level 3 VPN technology that's inherently limiting. Also, LHCONE is a small enough homogeneous community for people to know each other, communicate effectively and work out problems.

Shawn McKee (U Michigan): Scaling of trust is a key part of the size of the community.

Igor Sfiligoi: With an OSG hat on, he believes that the only way to scale is to federate across systems and delegate trust.

Jim Kyriannis (NYSERNet): In a computational workflow, when a DTN is involved, there's an atomic step where a transfer needs to take place. In the interest of efficiency, can you get to the point where that atomic step is not required, and the transfer is more built into the workflow or file system? Can we get to a point where the DTN is there in concept, but data movement and the computer are brought together more naturally?

Igor Sfiligoi: While this doesn't answer your question directly, in a StashCache implementation, one runs a cache on the DTN with storage. If required data is not at the computer, it's fetched; if it's already cached, it's there and no transfer is required. For those applications that can effectively use caching, the data movement is already hidden from the user.

Bill Johnston: The LHCONE architecture group has talked about this. One needs to be careful what you mean by DTN. DTNs have data models associated with them, normally on a server with GridFTP servers. But as applications become more sophisticated, they may have more application-specific I/O and storage models and then you have to implement that server also on the DTN and its storage. DTNs are not a file system; they are raw disks and you have to put something on top of them.

2.4   Panel 2 Theme: GRP Application Drivers – Moderator: Maxine Brown, UIC

This panel focuses on science applications that will drive the requirements and usage of the GRP.

2.4.1   The Square Kilometer Array: Data Transport, Processing, Archiving and Access – Shaun Amy, Australia Telescope National Facility

The Square Kilometer Array (SKA) is a large multi-national project to deploy arrays of radio telescopes in Australia and South Africa. The final Critical Design Review for SKA Phase 1 will occur in late 2019, with construction slated for 2020-2028. Pathfinder instruments are being built now to prove various technologies.

The Australian arrays, SKA1-LOW (50-350 MHz, with a maximum baseline of 65 km), are located in an isolated region of western Australia, with very little radio frequency interference. The site has recently been connected by ~400 km of optical fiber to a coastal town Geraldton, where it then connects to the computing facility in Perth. The South African arrays, SKA1-MID (350MHz-14 GHz, with a maximum baseline 120 km) are also located in an isolated region, the Karoo, Northern Cape, with connections to a computing facility in Cape Town.

Shaun Amy has been involved in the "Signal and Data Transport" work package. There are three networks designated for signal and data transport: Synchronization and Timing, Non-Science Data, and Science Data. On the science data flows, the SKA1-LOW raw bandwidth from individual antennas is ~2 Pbps, with the Central Signal Processor reducing that to ~9 Tbps to the Science Data Processor, which further reduces bandwidth by producing data products (e.g., lists of sources, multi-wavelength image cubes, time series data) for broader distribution to SKA regional centers. SKA data rates are so large that Science Data Processing is considered part of the telescope; for example, if fiber to Perth (or Cape Town) breaks, the system stops observing.

The project plans a global collaborative model for SKA Regional Centers (SRCs) as the point of access for the astronomy user community. The SRCs will maintain the data flow of observatory products, provide compute resources for data processing by users, establish a science archive, and provide technical expertise to users. There will be a proprietary use period for research investigators, after which data will become public. Details are being worked out, including the funding model (e.g., SRCs are not included in the construction budget). It is interesting to look to examples regarding regional center models from the LHC and high-energy physics community.

A projected look at the international network to support the distribution of observatory products is shown below (estimated prices in 2024 USD). There will be a need for strong partnerships with a number of NRENs for this network to be successful.
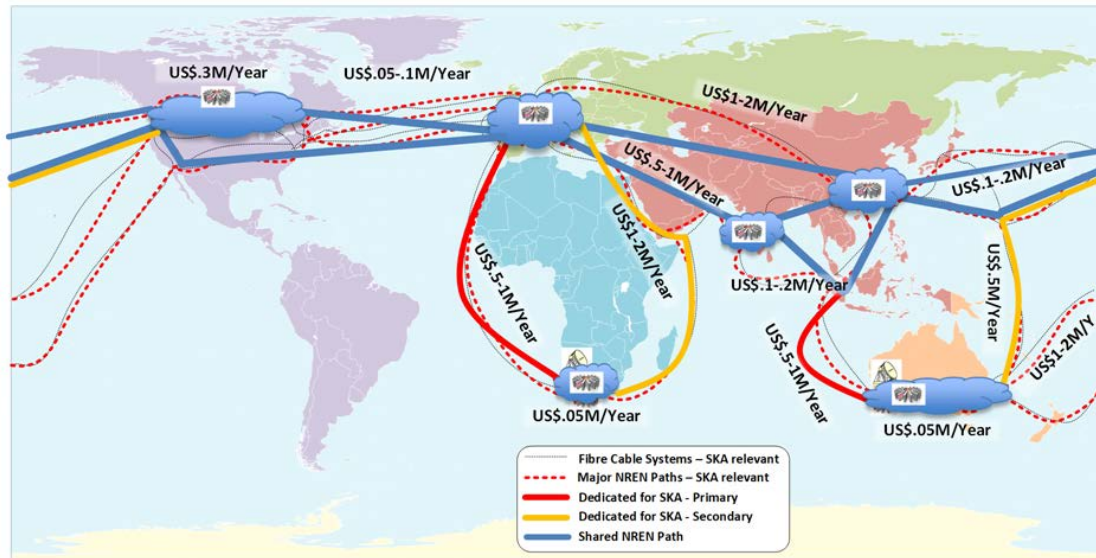
Figure 8: SKA International Network: Dedicated Links and Overlay   (see presentation for details)

There are a host of networking and data processing issues yet to be resolved. For example, locating processing equipment at the arrays not only requires power and facilities in remote regions but also creates RFI which then must be shielded at substantial cost; long-haul extreme-bandwidth networking to Perth may in fact be less expensive than the local shielding. The team is also evaluating FPGAs versus GPUs for specialized processing, and high-performance computing versus cloud-based services for data processing.

2.4.2   LSST Distributed Computing and Networks – Jeff Kantor, LSST

The Large Synoptic Sky Survey Telescope (LSST) will be an 8.4M telescope in La Serena, Chile that will conduct a 10-year survey of the sky, beginning operations in late 2022. The instrument is being built to address four major science areas: dark matter and dark energy, hazardous asteroids and the remote solar system, the transient optical sky, and the formation and structure of the Milky Way. It is a survey instrument, taking prescribed pictures of the sky, in contrast to most telescopes that schedule requests from researchers for observations of specific objects. Each image will be ~10 deg^2, with 3.2 Gigapixels and 6.4 GB of data. With a pair of images every 15 seconds (to eliminate transient effects), the data rate will be ~15 TB/night and 7 PB/yr. Over 10 years, the images and data products will be ~500 PB, and the catalog will contain ~37 billion stars and galaxies (each with thousands of measurements).

Long-haul high-capacity networks will be vital to LSST operations (see figure below). Every night, data will be transferred from the telescope itself to the base site in Chile and then via redundant 100 Gbps paths to Florida and on to U Illinois/NCSA for processing and dissemination. Additional LSST sites on the network include the project headquarters and science operations in Tucson (Arizona), camera support at SLAC (California), and a processing center in Lyon (France). Copies of the data will be maintained in Chile, NCSA and Lyon. There is a 60 second requirement on transient detection and alerts (to be met 98% of the observing time), which includes the time it takes to transfer data to NCSA and process it for transients.

Once per year, the accumulated dataset from initial operations through the present will be processed at NCSA and Lyon for "deep detection."
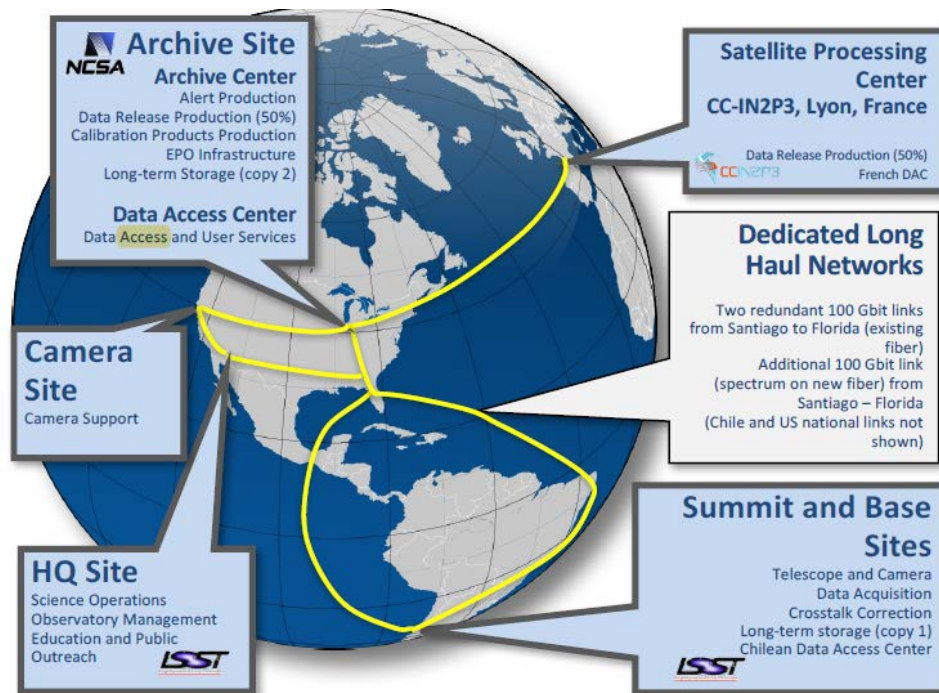
Figure 9: LSST Sites and Long-Haul Networks

The LSST workflow is relatively straightforward and simpler than the LHC. The challenge is in processing the data and providing access to the scientists. LSST will stand up computing resources at NCSA (and Chile) to either (a) process data locally at those centers, or (b) move big data to other sites for processing. He noted that there are distinct funding sources for construction and operations; however, funding separate from LSST construction would be needed, for example, for network links from universities to NCSA to access data.

The system will be dependent on the long-haul network that is implemented with diverse providers. They are establishing a Virtual Network Operations Center with the goals to provide (a) complete visibility to all participants into all links and sites, with a single-entry point for information and assistance, and (b) a single, integrated operational capability for end-to-end engineering, performance monitoring, security, maintenance, and all other operations.

2.4.3 Key Global Application Drivers in Asia and Korea – Buseung Cho, KISTI/KREONET

Buseung Cho, a senior researcher at KISTI and the Korean Research Environment Open Network (KREONET), provided an overview of KREONET's support/requirements from a variety of scientific applications, including high-energy physics, fusion energy, astronomy, biomedical science/genomics, and an AI-enabled datacenter. KREONET2 is a 100 Gbps national research network, with 17 domestic GigaPoPs and four international GigaPoPs; they currently provide connectivity to 200 national research institutes and support ~500K users. KREONET also has extensive international connectivity and collaborations, as shown in the figure below.
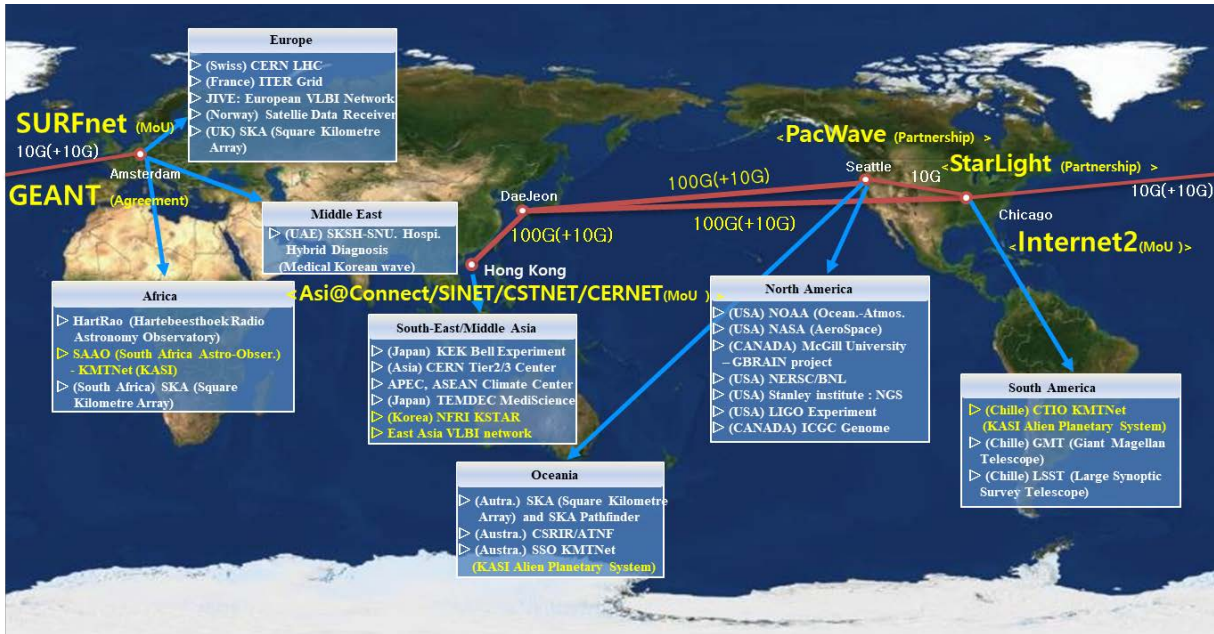
Figure 10: KREONET2's Global Research Networking and Collaboration

In the high-energy physics community, KREONET is part of the LHC Optical Private Network (LHCOPN) connecting LHC Tier 0 and Tier 1 sites, as well as part of the LHCONE (described earlier in this workshop by Bill Johnston of ESnet). They also are collaborating on the Worldwide LHC Computing Grid (WLCG) Data Lake initiative for distributed regional storage.

In fusion energy, KREONET supports both the Korean KSTAR project (see next talk by Si-Woo Yoon) and the international ITER program. It is estimated that the data volume per shot for ITER experiments will be ~1 TB initially and ~50 TB in its final configuration, with a potential archive capacity >200 TB/day. KREONET plans to support remote experimentation to Korean sites.

There are a large number of international Very Long Baseline Interferometry (VLBI) projects for astronomy and geodesy, such as the Very Long Baseline Array (VLBA) in North America, the European VLBI Network (EVN) in Europe (and Asia, South Africa, and Puerto Rico), the East Asian VLBI Network (EAVN) in China, Japan and Korea (correlator in Korea), and the Long Baseline Array (LBA) in Australia and New Zealand. There is an initiative for a Global VLBI Alliance that would bring together the capabilities of these separate systems to enhance sensitivity, resolution, and response time for event-driven observations. This would put even more strenuous demands on an end-to-end high-performance network, such as contemplated with the GRP, with various observing nodes requiring 40-350 Gbps across the various instruments. An initial step would be a collaboration across the four current correlation centers (New Mexico US, Daejeon South Korea, Dwingeloo Netherlands, and Perth Australia), and eventually linking individual telescopes for real-time correlation. The East Asian VLBI Network may be used as a pathfinder to demonstrate real-time correlation across individual observatories.

2.4.4 Introduction of Korean Fusion Program: KSTAR, ITER and K-DEMO and International Collaborators – Si-Woo Yoon, National Fusion Research Institute

Si-Woo Yoon described three fusion energy projects that Korea is involved in: the Korea Superconducting Tokamak Advanced Research (KSTAR), the International Thermonuclear Experimental Reactor (ITER), and the Korean Fusion Demonstration Reactor (K-DEMO). This was primarily a science talk, with some indications of the networking requirements. A comparison of the key parameters for the three systems is shown in the figure below.
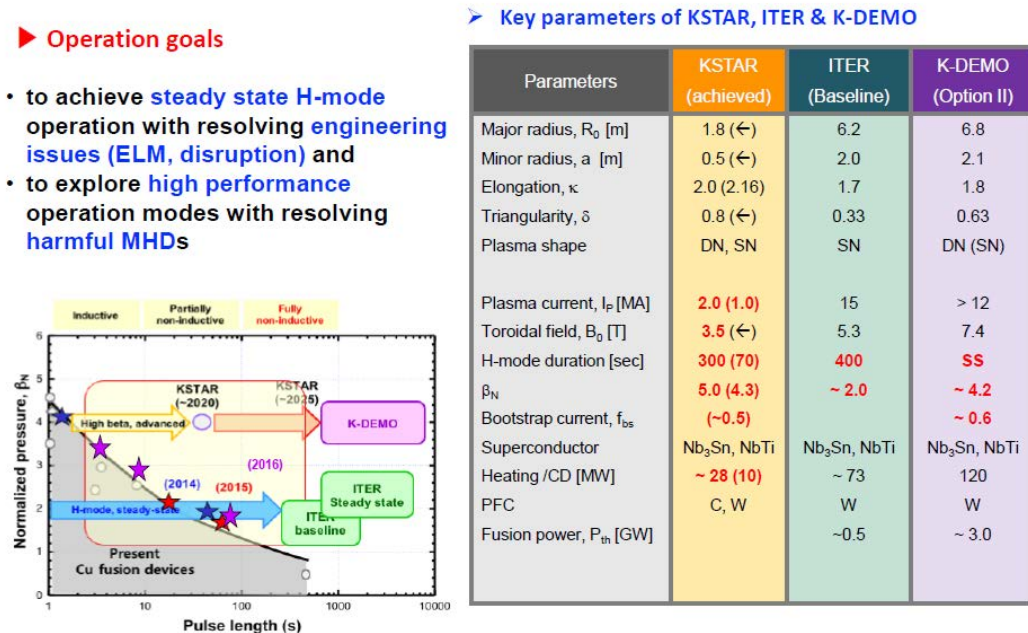
Figure 11: KSTAR Goals & Comparison of Key Parameters for KSTAR, ITER & K-DEMO

KSTAR was one of the first fully superconducting Tokamaks in the world, and has made a number of significant achievements in plasma fusion research. Many of the crucial technology issues for ITER are being addressed with KSTAR.

ITER is a large-scale international program, planning first plasma in 2025, with a target of achieving 500 MW thermal power and 10-fold energy amplification. The facility is located in southern France, with all partner countries building critical components of the reactor.

Assuming ITER is successful, the K-DEMO program would design a Korean fusion demonstration reactor that would be a step beyond ITER towards commercially viable reactors. Meanwhile, work is being done on a "Virtual K-DEMO" that is a simulation suite that includes physics and engineering, validated using data from KSTAR, ITER and other facilities. For the virtual demonstration, high-speed networking is key, with end-to-end performance of 10/100 Gbps required to KSTAR and ITER, as well as collaborators at worldwide institutions.

2.4.5 Next-Generation Cyberinfrastructures for LHC, High-Luminosity LHC and Data Intensive Sciences – Harvey Newman, Caltech

Harvey Newman has been involved with CERN, LHC and networking for decades, and his talk addressed next-generation challenges and approaches not only for the high-luminosity LHC (HL-

LHC) but also instruments such as the SKA and LSST, and more generally data-intensive sciences such as bioinformatics and earth observation.

The hierarchical (tiered) data flow system that was devised for LHC has generally worked well, although it has transitioned from the original concept. The availability of 100 Gbps networks about eight years ago changed the landscape significantly, and LHC is now tapping into cloud resources and elastic use of HPC systems and cloud systems. But it's important to note that the LHC has only produced ~2% of the total data that is anticipated over its lifetime (thru 2029). Cost projections to meet HL-LHC's networking, compute and storage requirements significantly exceed available funding, and the project needs to adapt. These adaptations will include machine learning algorithms to assist in data processing (especially high-bandwidth processing), a new "language" of neural networks, classical artificial intelligence, integration of GPU and FPGA accelerators both online and offline, and a focus on resource savings rather than scope expansion. They have done machine learning experiments with GPUs and FPGAs for top quark jet classifiers, with speed-ups of 100-1000X.

Newman commented that there needs to be a new character in our approaches to these challenges, beyond normal Moore's law-driven technology improvements; the figure below is a slide on new approaches to address these challenges.

Figure 12: Responding to the Challenges: New Overarching "Consistent Operations" Paradigm

He then described a series of demonstrations that Caltech and its collaborators have conducted at Supercomputing conferences over the years, which demonstrate not only ever-increasing network bandwidth, but also address various technologies associated with the next generation of solutions.

2.4.6 Panel Question and Answer Session

Michal Krsek (CESNET): Two comments. First for Dr. Yoon (KSTAR), in 2014-2016, in Singapore and with collaborators (some of whom are here), we worked on InfiniCortex and implemented 100 Gbps from Singapore to US using InfiniBand, demonstrating connectivity to places like PPPL for projects like KSTAR and ITER. Second, in 2016, using InfiniCortex fabrics, we demonstrated RCuda to simulate GPUs that are not in your location for GPU loads.

Bill Johnston (ESnet): (For Shawn Amy and Jeff Kantor) ESnet started seeing LHC data ramping up in 2004-2005, well before LHC turned on in 2007-2008. ESnet saw a continuous ramp-up with no discontinuity when LHC finally turned on – and the reason was that the data processing was being built and tested in continuous fashion well before data were available. Are SKA and LSST doing something similar in their project plans, so that data processing is not only operational but also operating on data prior to first light?

Shawn Amy: It's not being done as much as Amy might like to see. They have had one "data challenge," but that was with precursor instruments (not simulated SKA1 data). It would also be good to have a "network data challenge." There is still time to put this kind of plan into place, as the full SKA1 will not be fully operational until 2028.

Jeff Kantor: The first response is that LSST is over-provisioned, e.g., the requirement is 40 Gbps compared to a capacity of 100 Gbps. Second, we are testing with synthetic data as we bring each link up. We will have 3 stages of commissioning where the data rate ramps up: small-scale (calibration telescope), intermediate-scale (1/20th scale commissioning camera), then full-scale for a year before operational. The software will continue to evolve. Unless the cadence is changed, people have considered shorter exposure time to press the processing timelines.

Jason Haga (AIST): (For Jeff Kantor) How are you handling identifiers for persistent data products?

Jeff Kantor: He has stepped away from this area, but at the time he stepped away, the plan was that all the data products will have full provenance and metadata (data, telemetry, calibration, processing).

2.5 Demo Posters

The following posters and live demonstrations were hosted during an evening reception. The posters/slides/images presented are available on the workshop website.

- Brian Balderston (UCSD/SDSC) - U.S. National Data Service: Labs Workbench: A Scalable Platform for Research Data Access, Education and Training.

- Jim Chen (Northwestern/iCAIR) - DTN-as-a-Service at Starlight: A GRP Prototype Service.

- Alvin Chiam (NSCC) - Data Mover Challenge 2019/2020.

- Buseung Cho (KISTI)- SC18 NRE Demo: Global Petascale to Exascale Science Workflows, Accelerated by Next Generation SDN Architectures and Applications.

- John Hess (CENIC) - SDX interoperability.

- Christine Kirkpatrick (UCSD/SDSC) – The Open Storage Network: Distributed Storage Architecture for Data-Driven Science.

- Dima Mishin (UCSD/SDSC) - PRP Trace Route Visualization (real-time demo; screen shot provided on workshop website).

- Harvey Newman (Caltech) - Next-Generation Cyberinfrastructures for LHC and Data Intensive Sciences.

- Igor Sfiligoi (UCSD) - Network benchmarking of cloud providers and NRP/GRP resources.

- Wenji Wu (Fermilab) - BigData Express: Toward Predictable, Schedulable, High-performance Data Transfer.

2.6 Keynote – Infrastructure SINET 100G Global Ring and Data Exploitation – Tomohiro Kudoh, University of Tokyo

Tomohiro Kudoh, Professor and Division Head of the Network Research Division at the University of Tokyo, described Japan's high-performance academic research and education network, as well as plans to develop a Data Exploitation Platform around SINET with a focus on supporting "every-day" applications by a broad community of users.

More than 3 million users use the current SINET5 network, which connects more than 900 universities and research institutions with 100 Gbps lines to all 47 prefectures in Japan (see figure below). A 400 Gbps connection should be operational later this year between Tokyo and Osaka. Internationally, SINET5 has a 100 Gbps international ring from Tokyo-Amsterdam-New York-Los Angeles-Tokyo, as well as a 100 Gbps connection to Singapore. And SINET5 is deploying a mobile capability to support Internet of Things (IoT) research and mobile applications in the 5G era; the mobile virtual network created for SINET over a commercial mobile network is directly combined with the SINET VPN plane through gateways for secure communication. SINET always tries to over-provision the capacity relative to user demand. The next-generation SINET6 is being designed now and should be deployed in 2022.
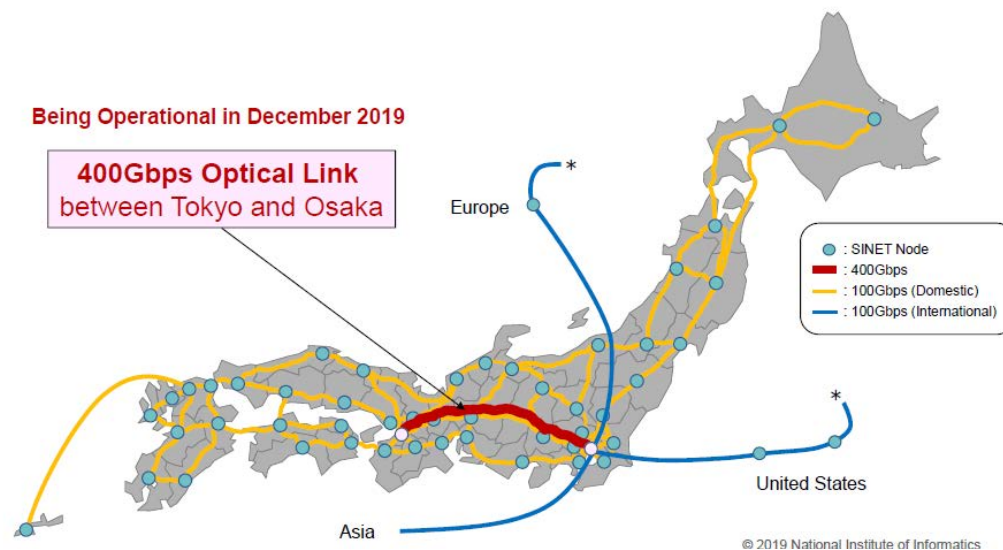
Figure 13: Map of SINET5 national links and international connections
(note that Japan is of comparable geographical size to California to compare density of connections)

SINET has demonstrated a number of high-bandwidth data transfers within Japan and internationally. An SC'17 demonstration showed 231 Gbps file transfer from Tokyo to Denver (using 3 100 Gbps paths), and 321 Gbps was achieved during TNC'19 between Tokyo and Estonia (using 4 100 Gbps paths). The system has been used for backups between HPC centers and for 8K pixel video transmission.

More than 2,700 virtual private networks (VPNs) are currently provided for collaborative research projects, access to cloud computing resources, and across multiple campuses.

A number of current usage examples were described, including high-energy physics, nuclear fusion, seismology, astronomy (including connectivity with the ALMA facility in Chile, as well as the worldwide VLBI network), geodesy, space observation, medical image data analysis, medical diagnosis by remote video, and remote medical education.

The second part of Kudoh's talk addressed plans for a Data Exploitation Platform (tentative name), shown in the figure below. It will provide a rapid Proof-of-Concept environment for data exploitation, including industry-academia collaborations, as projects can use wide-bandwidth, low-latency slices, with dynamic provisioning of real-time data collection/storage/analysis infrastructure leveraged by SINET. The platform will help solve "regional disparity problems"; for example, there are challenges with infrastructure and resources in rural areas, but users should be able to get access to significant resources using this platform.

A small-scale testbed of the data exploitation platform will be available in FY2019, and the first deployment system will be operational by the end of FY2020. AIST's AI Bridging Cloud Infrastructure (ABCI) system will participate in the program, and the platform could serve as a streaming, data-gathering infrastructure for ABCI or other supercomputers. In addition to software-based security (authentication, authorization, and encryption), the data exploitation platform will provide isolation by the network (VLAN).
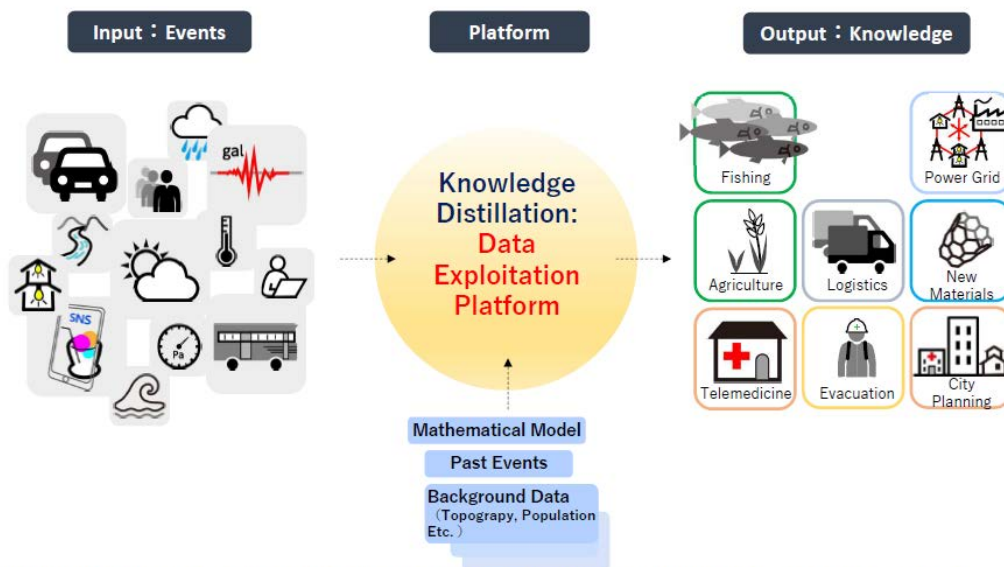
Figure 14: Platform for Data Exploitation to Realize Knowledge-Intensive Society

Question and Answer Session

Shawn McKee (U Michigan): In a slice, do users make a copy of the internet that lets them keep current addresses, or do they need to re-address all their infrastructure so that it exists within the slice?

Tomiro Kudoh: The user can manage addresses – users can assign private or global addresses.

Unidentified: How do users authenticate on a virtual mobile network? Are there SINET SIM cards, or looking at eSIM, or other solutions?

Tomiro Kudoh: Authentication is done by SIM. Each SIM assigned to a certain VPN. Assignment of SIM to VPN is static.

Unidentified: You mentioned isolation from the internet. How do you prevent someone from attaching a router to one of your edge nodes and using that as a gateway to the rest of the internet?

Tomiro Kudoh: The SINET profile VPN and the router connects the slice inside the infrastructure to the VPN, so no packet goes out to the internet and no packet can come in from the internet.

Follow-up: So, for your end users that may be connected with a 5G device, what keeps them from using that as a gateway?

Tomiro Kudoh: Edge devices should also be in the slice, using SINET mobile infrastructure. But you are right, a user may want to have a gateway to the slice and the user can provision such a gateway within a slice, and security would be the user's responsibility.

2.7 Panel 3 Theme: Globally Distributed Data Fabrics – Moderator: Marek Michalewicz, University of Warsaw

This panel included three talks describing different distributed systems for sharing and exchanging data.

2.7.1 Globally Distributed Secure Data Exchange Fabrics – Cees de Laat, University of Amsterdam

Cees de Laat, Chair of the System and Network Engineering (SNE) laboratory at the University of Amsterdam, opened his talk with the three musketeers' phrase "all for one and one for all," with the meaning that many cyberinfrastructures center around computers and workflows, and we need to create a fluid data layer that frees data to be shared and used by any applications. Networking efforts such as Science DMZ/DTN fabrics, as well as FAIR[16] (findable, accessible, interoperable and reusable) data policies help address this latter element of sharing/using data.

The main challenge is that many data (99%) are not shared, and even under FAIR policies, data is often restricted by policy or law. Organizations that compete often also have common goals where data sharing – to address those common goals and yet not the competing goals – is mutually beneficial. Are there ways to organize data alliances, with policies and technical mechanisms to implement agreed-to safeguards, so that data can be more freely shared?

He described several use cases that his group is involved in – e.g., medical data enabling personal interventions and airlines collaborating on sharing/exploiting aircraft maintenance data.

One would like to create a digital marketplace ecosystem. To accomplish this, it is necessary to flow down strategic issues (legislation, data use policy, trust) to tactical issues and finally to operational technical issues (encryption schemes, VM and container implementations, SDI/SDN technology, and blockchain ledgers). His group has been conducting research in this field (see figure below), including developing draft policies, developing archetypes, implementing a proof of concept (using DTNs and Kubernetes), and conducting demonstrations (e.g. container-based remote data processing and data harbors).

The PRP/NRP/GRP platforms are built on the concept of interoperating Science DMZs, but how does the organization resolve potential conflicts if various Science DMZs operate with different policies? Imagine that data that is to be shared amongst participants, but restricted from non-

---

[16] See for example, "The FAIR Guiding Principles for Scientific Data Management and Stewardship," M. D. Wilkinson et al. 2016, https://www.nature.com/articles/sdata201618 .

participants, 'leaks' from the system – the person that may be fired is the person that is likely to say no to this collaborative environment; the policies and procedures must be in place across the collaboration to meet that level of concern and satisfy both users and operators of the system.
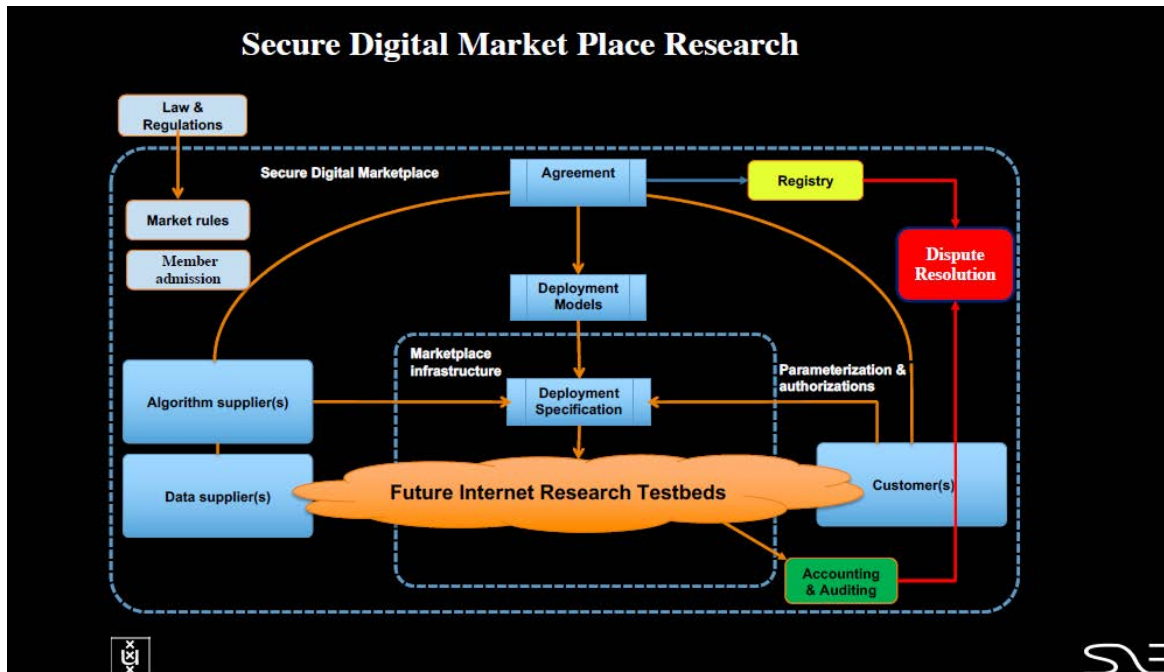
Figure 15: Schematic of Secure Digital Market Place Research

2.7.2   Data Lakes, Data Caching for Science: OSiRIS Distributed Storage Systems – Shawn McKee, University of Michigan

Shawn McKee works in U Michigan's Institute for Research and Innovation in Software in High Energy Physics (IRIS-HEP), which focuses on preparing for the High-Luminosity LHC (HL-LHC). As stated earlier by Harvey Newman, HL-LHC presents compute and storage requirements a factor of 5-10 beyond the available budget, and new approaches to computing/storage must be developed to stay within the budget constraints.

Two projects out of IRIS-HEP were discussed. The first is the OSiRIS program to deploy a distributed storage infrastructure, built with inexpensive commercial off-the-shelf (COTS) hardware, combining the Ceph storage system with software defined networking to deliver a scalable infrastructure to support multi-institutional science. The primary use case is sharing working access to data, and scientific applications range from HEP to bioinformatics, Cryo-EM and social science. The system currently spans five institutions in Michigan and Indiana, with connectivity from 4-100 Gbps, and a storage capacity of 7.4 PiB.

The second project is part of the HEP Data Organization, Management and Access (DOMA) research area and focuses on "Data Lakes" (see figure below), a reorganization of the current grid tiered infrastructure.

Data Lakes have been discussed in many contexts. Within DOMA we used the concept to provide a boundary between how we store and manage our data **cost-effectively** and how we access and consume that data.

The primary attributes

- Cost-effective (reduce copies of data)
- Provide internal intelligence
- Optimize QoS (performance/cost)

Our current HEP grid infrastructures typically uses various redundancy mechanisms at each site AND we store multiple copies across sites. This is expensive.

Data Lakes provide a way to optimize

Data Factory/source (e.g. T0 or sim)

Data Store/Lake

Intelligent Data Delivery Service (iDDS)

Data Cache

Compute Nodes/ Data Sinks

Rucio/ FTS

Lake Border

Areas where DOMA team is working

HPC   Analysis Facility   Grid Site
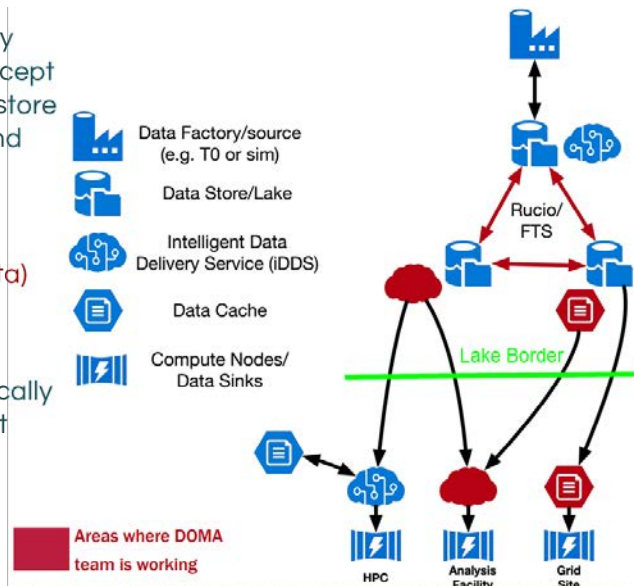
Figure 16: Data Organization, Management and Access (DOMA) Data Lakes

Caching can reduce storage requirements and provide acceptable performance. The group has analyzed current ATLAS traffic and found that 40% of accesses and 55% of traffic could have been served from 50 TB cache. While GridFTP is still the most used transport mechanism, they are exploring http and Xrootd. DOMA, in conjunction with the Xrootd and OSG teams, has developed XCache (think "Squid" for Xrootd) to help provide caching capability for the current grid and future data lake infrastructure.

They have used the OSiRIS system as a prototype for testing and evaluating various techniques. Early testing of quality-of-service from Ceph showed that shaping from sites towards bottleneck can improve client performance ~5-10%. Furthermore, while the regional-level OSiRIS system (with networking RTT $\sim< 10$ msec) is working well, they have evaluated how the Ceph-based system would work at higher latencies for more distributed systems. A demonstration at SC'16 added a new site to the system with 42 msec RTT; the infrastructure continued to work well, but the benchmark workflow dropped from 1.2 GBps to 0.45 GBps. In later experiments, they artificially added delay into the network. With 160 msec delay, the Ceph cluster stopped working, and latency had to be backed down to ~80 msec for the system to recover. Therefore, this experience indicates that for widely-distributed systems, one would need to use Ceph federations or employ caching to hide the latency.

The DOMA lessons learned to date show that the strawman concept of data lakes helps optimize use of storage while maintaining performance. In their research, they were able to effectively trade network use for storage by decreasing the number of copies and provide WAN access to specific data when needed, and they have demonstrated that caching allows one to both recover performance for WAN data and reduce the amount of network bandwidth required for typical workflows.

2.7.3   Open Storage Network — Christine Kirkpatrick, UCSD/SDSC

Christine Kirkpatrick is a co-Principal Investigator on the Open Storage Network (OSN) award, and is representing Alex Szalay (Johns Hopkins University, the OSN PI) who was unable to attend the workshop.

At least in the US, while computational and networking resources tend to be well-funded and planned and organized, storage facilities are not as mature. Storage systems tend to be built into large projects, more balkanized, with no standards requirements, and not the same level of central funding. Especially if you are not at a large-scale research university, storage can be a challenge.

The Open Storage Network is a distributed national-scale storage system for research data. A guiding philosophy is to "keep it simple" so that the system can easily scale out to a large number of sites. It is a distributed system made up of standardized and scalable systems, with unified administration. The initial deployment of prototype systems will be at six sites (JHU, Massachusetts Green HPC Center, StarLight, UCSD/SDSC, UIUC/NCSA and UNC/RENCI). The prototype scalable units (see figure below) have ~1 PB usable object storage (using Ceph) and 40 Gbps I/O with a price point ~$140K.
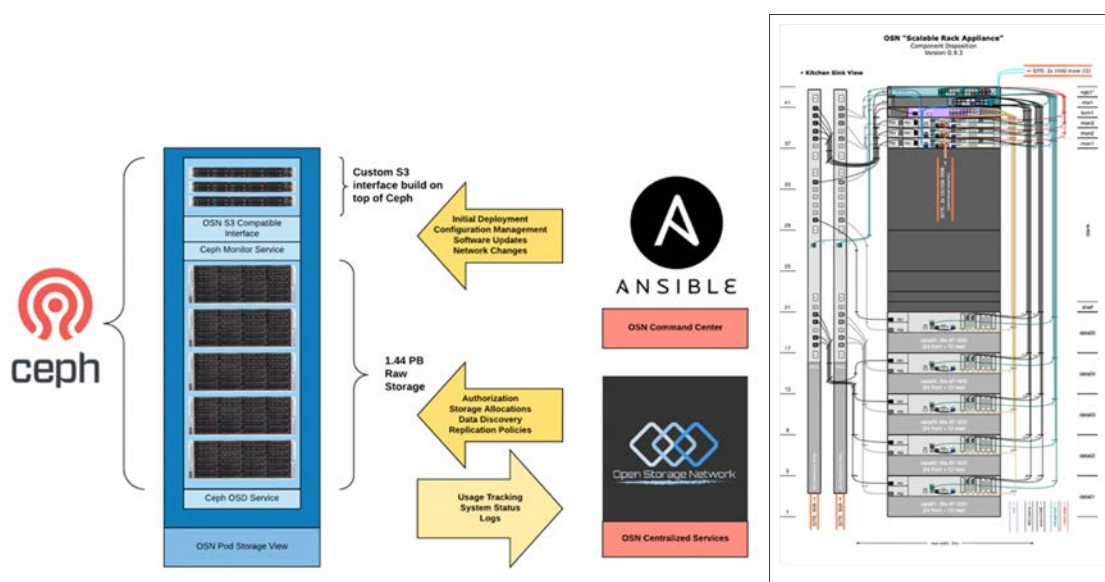
Figure 17: Open Storage Network Pod Storage View

A number of diverse use cases for the storage system were described, including experiment-to-site data dissemination, slice-and-compute, workflow staging, common resource access, community long-tail data, and dataset as benchmark.

Although the project is relatively new, there have already been some valuable lessons learned. First, while the initial prototypes are all homogeneous hardware, this is not realistic long-term as technology evolves and various institutions follow their independent procurement processes. Second, the project manager on a distributed project like this plays a crucial role – and can be a single-point failure (their initial project manager was hired away by Cray). Third, many

organizations have expressed interest in purchasing units and participating in the project, which indicates that the OSN is addressing a real community need.

Near-term plans include working on the application layer (with tools such as Clowder, Globus, Dataverse, iRODS, and WholeTale), refining governance of the OSN, making progress on policies and procedures, becoming an XSEDE Service Provider, partnering with AWS, and convening a workshop at TACC in October.

2.7.4   Panel Question and Answer Session

Gauravdeep Shami (Ciena): (For Christine Kirkpatrick) Am I correct in the bottleneck being 40 Gbps I/O?

Christine Kirkpatrick: No, that's not the bottleneck. To be honest, they assumed that an object store was the best way to go for users, but those protocols are slowing things down, not the networking. They will be working on this, perhaps revisiting the object store approach. And they are interested in working with OSG.

Harvey Newman (Caltech): (For Shawn McKee and Christine Kirkpatrick): There is a missing element in your talks, which is the agility in storage systems. There have been radical changes in performance and pricing in SSDs, which should be storage system front-ends. Commodity SSDs currently support 3 GBps, PCIe 4 GBps in 4 lanes, and next-generation 6-7 GBps. Can get 1 Tbps/rack-unit, and with pure commodity, can get 100 Gbps per rack unit. The science community would like to do things in much shorter times. We should design storage services with not just volume capacity, but also bandwidth. This is also highlighted in the case of AI (e.g. Optane SSDs).

Christine Kirkpatrick: They did look at SSDs, in fact this is where they encountered an end-of-life issue as there were delays in specifying their system. They would love to have the advantages of SSDs, but were constrained with their self-imposed $140K/PB price point.

Shawn McKee: He's been involved in this for some time. When they designed OSiRIS nodes, they did build in SSDs. In new nodes, they have better ratio of disk and SSD. They use the SSDs not just for metadata but also for caching. They try to work closely with each science application – e.g. throughput and IOPS versus resiliency and capacity – and can tune their systems.

Christine Kirkpatrick: As people use the infrastructure, they are likely to develop updated designs which may emphasize performance over capacity.

Shawn McKee: They plan to put 1PiB of OSiRIS into the Open Storage Network.

Marek Michalewicz: (For Cees de Laat) How do you see work on policies being combined with other efforts?

Cees de Laat: There is a lot of emphasis on storage capacity, but you need to have focus on data policies. If you don't have policies in place, there will remain risks of sharing data. They are looking primarily at programmability aspects.

Christine Kirkpatrick: In data-centric forums like Research Data Alliance, there are lots of diagrams about data life cycle, but for those with a production system perspective, lifecycle may look more like users put data in and the data stays there until the system goes away. You need to look at data policies matching user expectations to the system capabilities. For example, how long can you host data?

Cees de Laat: They are working with UvA business school to examine the value of intermediate steps in data processing. What are the incentives for various parties along the chain of processing so that they want to contribute to the end products?

2.8   Panel 4 Theme: Data Movement Services – Moderator: Tom DeFanti, UC San Diego/QI-Calit2

This panel session included six talks that focus on the data movement process – various software tools and DTN-based systems for moving data, and research into high-performance data transfers.

2.8.1   'Move That Data!' Data Mover Challenge Judging Reflections – Andrew Howard, Australian National University

The National Computational Infrastructure (NCI) of Australia, with the National Supercomputing Centre (NSCC) of Singapore, hosted a 'Data Mover Challenge' in conjunction with Supercomputing Asia 2019. The rationale for this Challenge was not only to test performance of various transfer tools, but to facilitate succession planning with engaging new network engineers and having various NOCs work together, to build trust networks, to containerize tools, and to validate network capabilities. Seven groups participated in the challenge: SEAIP/NCHC+StarLight, Fermilab, Zettar Inc, The University of Tokyo, Argonne National Laboratory, iCAIR/ Northwestern University, and JAXA/Fujitsu.

The testbed was based on existing DTNs on production research and education networks between 100G-connected centers at NSCC (Singapore), NCI (Australia), NICT (Japan), KISTI (Korea), and StarLight (US). Participants were not given root access, and all workflows and tools had to be built within Singularity containers. The test dataset was 2 TB with 1431 files of actual scientific data (lots of small files). Most participants requested NIC ring buffers (Mellanox), pacing, and NUMA socket affinity. Each participating team was allocated a week on the testbed to run their tests.
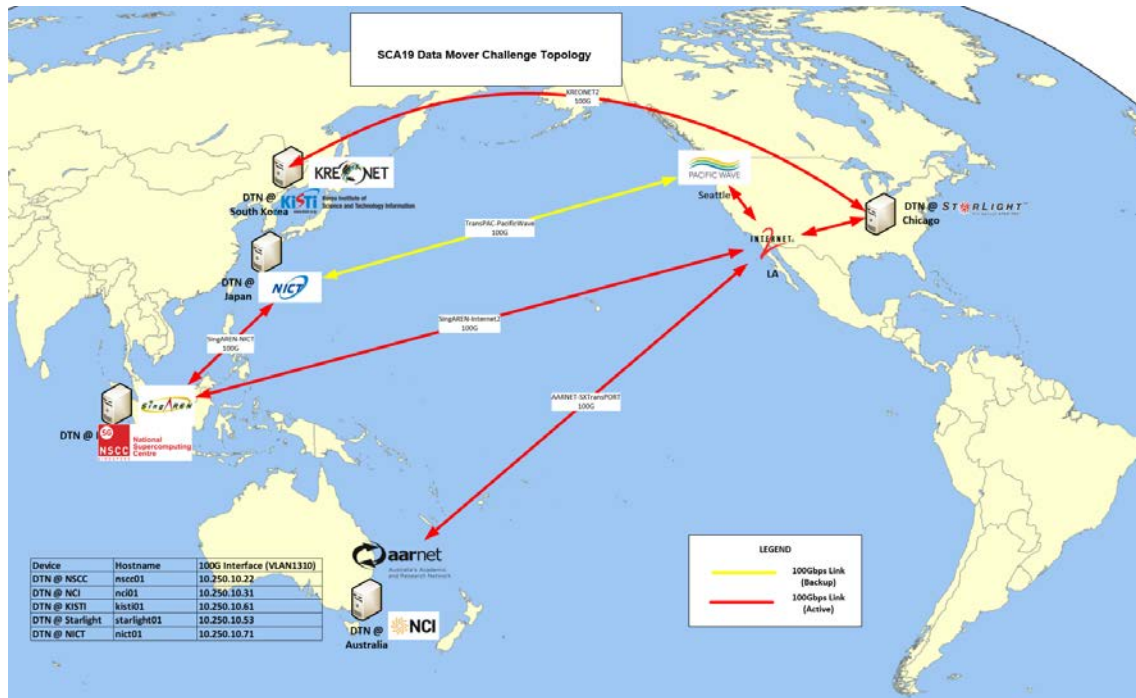
Figure 18: The Data Mover Challenge 2019 network

The SEAIP/NCHC+StarLight team developed an innovative DTN-as-a-Service approach (see talk by Jim Chen from the StarLight/iCAIR/Northwestern team) using a Jupyter Notebook controller and user interface, and the NUTTCP transfer tool. The Fermilab-led team used BigData Express (see next talk by Wenji Wu). Zettar Inc. used their "zx" hyperscale data distribution software platform. U Tokyo used their secure data reservoir approach. The Argonne team used GridFTP and Globus Online. And the JAXA/Fujitsu used a smart communication optimizer, with a proprietary protocol to accelerate standard tools.

In terms of data transfer rates, the Zettar team 'won' the competition. However, considering all criteria, Australia chose BigData Express as the transfer mechanism between their supercomputer centers.

2.8.2 BigData Express: Toward Predictable, Schedulable, and High-Performance Data Transfer – Wenji Wu, Fermilab

BigData Express[17] (BDE) has been developed by a team at Fermilab as a schedulable, predictable, and high-performance data transfer service with the following characteristics:

- A peer-to-peer, scalable, and extensible data transfer model
- A visually appealing, easy-to-use web portal
- A high-performance data transfer engine

---

[17] See http://bigdataexpress.fnal.gov .

- A time-constraint-based scheduler
- On-demand provisioning of end-to-end network paths with guaranteed QoS
- Robust and flexible error handling
- CILogon-based security (certificate version)

A diagram of the BDE software stack is shown in the figure below. There is an easy-to-use web portal. The BDE Scheduler is a constraint-based scheduler and co-schedules the DTN, storage and network. There are four agents for the DTN, SDN, storage and data transfer. The team has developed mdtmFTP[18] as their high-performance data transfer engine. The software structure is flexible (e.g., can establish multiple data transfer federations), scalable by managing site resources via agents, and extensible (e.g., the data transfer agent can host plug-ins for other transfer mechanisms such as GridFTP or Xrootd).
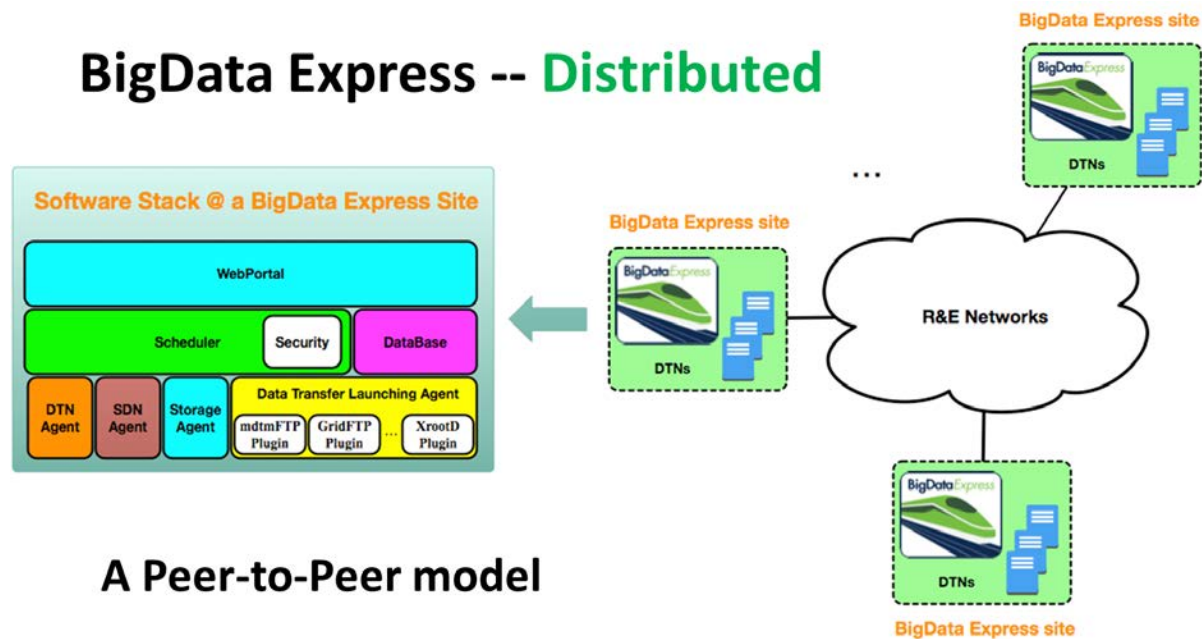
Figure 19: Schematic of the Distributed Nature of BigData Express

The data transfer engine, the multicore-aware data transfer middleware (MDTM), optimizes use of underlying multi-core systems and is efficient in transferring large numbers of small files.

The team conducted a demonstration of BigData Express at SC'18 with a large number of US and international partners.

In terms of current deployment of the software, it is being used at KISTI and KSTAR in South Korea, SURFnet and UvA in Europe, Fermilab, StarLight/Northwestern, U Maryland and Ciena in North America, and NCI in Australia. The software and documentation are available. Looking ahead, their roadmap includes REST APIs for scientific workflows, support for Kubernetes to automate docker-based deployment, and support for CILogon with tokens.

---

[18] See mdtm.fnal.gov for software.

2.8.3   DTN-as-a-Service at Starlight – Jim Chen, NU/iCAIR

The iCAIR/StarLight group has been working on a number of projects in support of a GRP, with the objective to provide services that can be used by many scientific groups.

The International P4 Experimental Network (i-P4EN) is to be demonstrated at the EuroP4 workshop in September 2019 with multi-tenant support prototypes for the international P4 testbed (P4MT) and sketch-based entropy estimation for network traffic analysis using a programmable data plane.

Starlight is developing a DTN-as-a-Service (DTNaaS) platform with the following characteristics:

- NUMA-aware task and process autonomous configuration
- Autonomous optimization for the underlying hardware and software system
- Modular data transfer system integration platform
- Support data access with NVMe over Fabrics
- Science workflow user interface for network provisioning with NSI/OpenNSA
- A monitoring system for high-performance data transfer

This system should provide many benefits including enabling users to move data without any knowledge of the underlying infrastructure, a platform for autonomous configuration and optimization for data transfer using DTNs, supporting operation in Docker, Singularity and Kubernetes with Docker, supporting NVMe over Fabrics for accessing remote storage as a local device, allowing users to evaluate the data movement in real time, reconfiguring the system, changing the transfer tools as required, and having a modular design, implemented on Jupyter + Python, well-suited for science research workflow integration. The current software stack architecture is shown in the figure below.
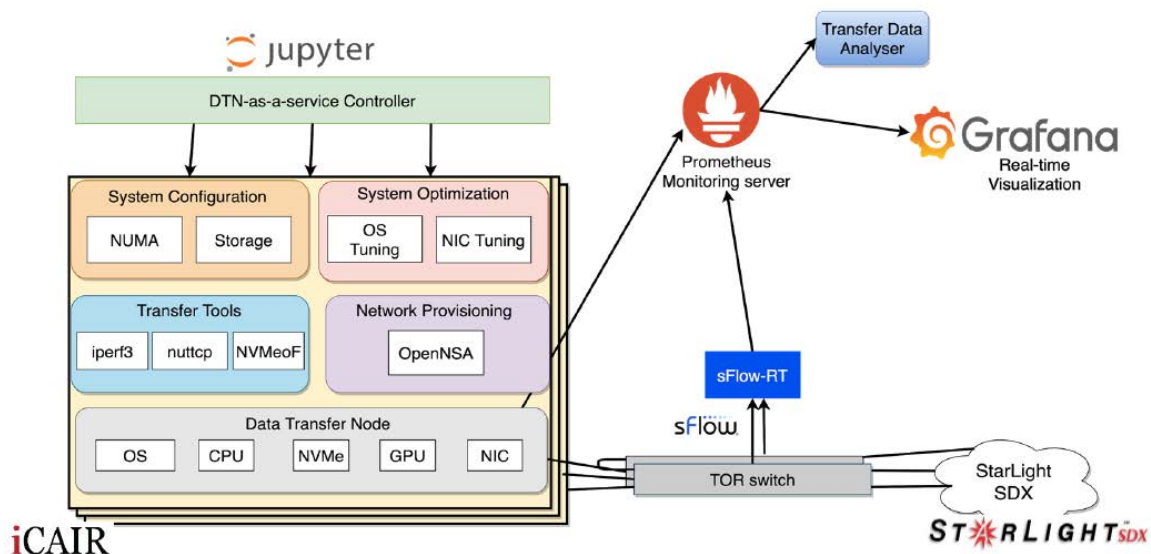

Figure 20: Current Starlight DTN-as-a-Service Software Stack Architecture

The team has developed a series of hardware/software packages to support DTN functions at

Supercomputing demonstrations, as well as working with the Southeast Asia International Program (SEAIP) on the Data Mover Challenge in 2019 (see Andrew Howard's talk earlier in this panel) and 2020 (to include US, Asia and Europe). They won an award for innovation at SCAsia 2019.

Future work includes developing an OSG DTNaaS prototype and participating in the national/international OSG Cache trial, partnering with the science community and SDXs to establish LAN/WAN packet loss troubleshooting reference workflow and procedure, integrating Xrootd and other protocols in the prototype, developing an SDX/MeoF service prototype, and working on DTNaaS clustering and federation.

2.8.4   Next-Generation DTN Architecture/Kubernetes – John Graham, UC San Diego/QI-Calit2

John Graham of UCSD described two primary projects he is working on related to the next-generation DTN architecture. The first is a project for the High-Performance Wireless Research and Education Network (HPWREN)[19], which connects remote sites across southern California via microwave network links. HPWREN operates an independent high-availability cluster, managed under Kubernetes. They use EdgeFS, with Rook enabling EdgeFS storage systems to run on Kubernetes with Kubernetes primitives. The public-facing side of the storage system is NextCloud.

With the Nautilus cluster, they use Ceph Block/CephFS/S3, which allows them to use both POSIX and S3 interfaces simultaneously. They have had very positive feedback from deep learning neuroscience researchers at UC Santa Cruz.

2.8.5   Advancing Open Science through Distributed High Throughput Computing – Frank Würthwein, UC San Diego

Frank Würthwein is a physics professor at UCSD and Executive Director of the Open Science Grid (OSG). The OSG is a consortium "dedicated to the advancement of all of open science via the practice of Distributed High Throughput Computing, and the advancement of its state of the art." OSG serves four groups: individual researchers/small groups on OSG-Connect, campus research support organizations, multi-institutional science teams (many of them international), and four "big science" projects (US-ATLAS, US-CMS, LIGO and IceCube). "We integrate clusters/storage from anywhere in the world that can be used coherently by science teams."

Federation, meaning distributed control, is a central philosophy of OSG and works on three principles: (a) resource owners determine policy of use, (b) resource consumers specify the types of resources they are willing to use, and (c) via HTCondor, jobs are submitted locally, queue centrally, and execute anywhere after a resource that matches the job requirements becomes available.

---

[19] See http://hpwren.ucsd.edu/.

The OSG Data Federation (see figure below) currently includes 6 data origins and 12 globally-distributed data caches. OSG operates a "Data Origin Service" for people with data on their storage systems to export their data into the federation, and the data owner can set rules for who can access what data. The OSG caches help to maintain performance by hiding latencies, reducing unnecessary network traffic via data reuse, and protecting the data origin sites from overloads. The data federation is "mainly sociology, a little technology."
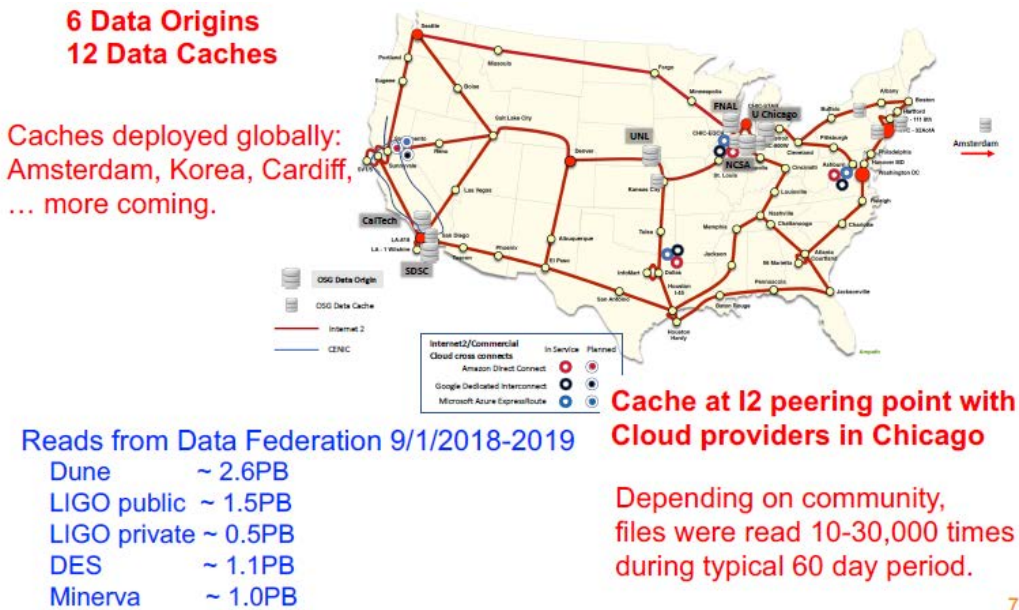
Figure 21: The OSG Data Federation

Their vision for a research platform like the PRP is to develop a packaged system, where the provider provides the hardware system, operating system and Kubernetes, and OSG provides all the details inside that. In the long-term, there will be capacity providers (clouds, on-premise, in the network, with both computing and storage), service providers (software, consulting, training, content), and science teams at all scales (from individuals to multi-thousand global collaborations). Containerization and container orchestration (e.g. Kubernetes) allow for a new division of labor to support science globally and lower barriers to adoption of new software and services.

2.8.6 Machine Learning for Research Networks – Anna Giannakou, Lawrence Berkeley National Lab

Anna Giannakou of LBNL described research she is conducting on automatically detecting and mitigating performance degradation in network transfers.

Scientific discovery depends greatly on data availability, often meaning large WAN transfers. Unfortunately, these transfers are susceptible to performance degradation events, such as packet loss and abnormal transfers. Packet loss may be due to congested paths, simultaneous competing transfers, or network configuration of end hosts. Abnormal transfers may be due to user errors or security breaches.

How can these performance degradation events be mitigated? The first step is to detect them. They are developing an accurate lightweight framework to detect such events and then integrating that framework in real network environments. These events can then be used to reconfigure the network to optimize data transfers. The proposed architecture is shown in the figure below. NetPLoP is used for predicting network packet loss, and FlowZilla for anomaly detection.
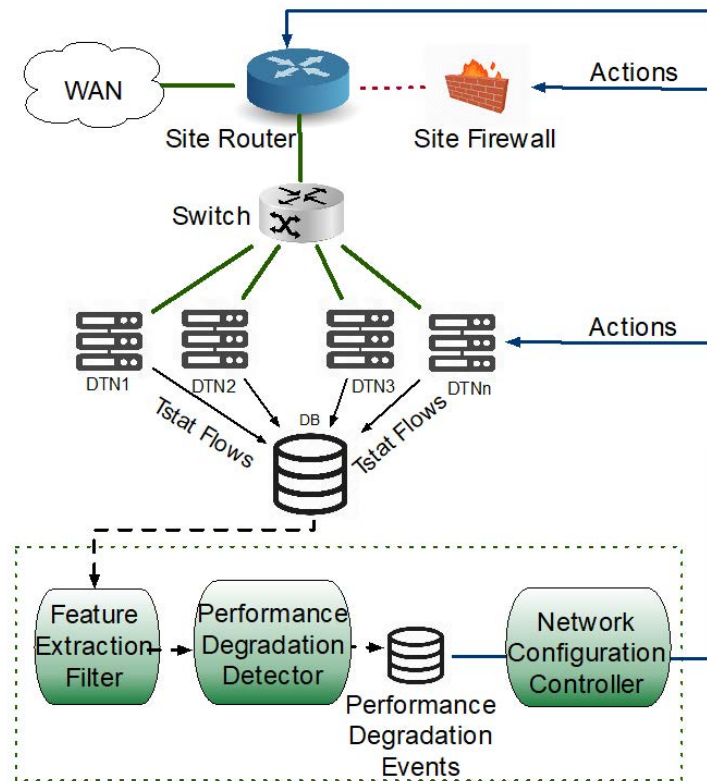
Figure 22: Proposed architecture for performance degradation detection and feedback

It is difficult to accurately predict packet loss, as there are many issues that impact packet loss. NetPLoP selects an appropriate set of flow properties (e.g. average RTT, throughput, TCP congestion window, size and duration) and trains the model using linear regression on only faulty transfers (those with packet loss). They have done experiments on the accuracy of NetPLoP and found that noise removal can significantly improve prediction accuracy.

With respect to anomaly detection using FlowZilla, it's hard to define a 'normal' profile as network traffic is very dynamic. This leads to a high number of false positives, which is unappealing for real-time operational deployment. They have applied the framework to research networks, which have less feature variability than general internet traffic. FlowZilla has two phases – training to determine what features represent abnormal events, and then detection using previously unseen flow data. The model is trained using linear regression. They use adaptive thresholds because of seasonal variations in traffic characteristics. Using artificially-injected anomalies, FlowZilla detected >80% of the anomalies.

In conclusion, machine learning/artificial intelligence provides the opportunity to realize smart networks and reduce human involvement in monitoring/tuning networks. The early results show reasonable accuracy with relatively low training cost. They would like to deploy this framework in real-time environments, and expand to additional types of performance degradation events.

2.8.7 Panel Question and Answer Session

Inder Monga (ESnet): (For Frank Würthwein and Wenji Wu) DTNs and distributed infrastructure are OK, but it is a challenge to get new data transfer software adopted. What is the best approach for the community to look at data transfer software?

Frank Würthwein: What we do is limited by social issues, for example the requirement that resource providers or users install something. There are problems of 'we don't do that here,' and adoption of anything new is hard. A transition to the model he described in his talk can change everything, because you separate people who own hardware and operate services (containers). You can use production infrastructure.

Wenji Wu: This is a long-lasting problem and within the DOE, there was a recent workshop that discussed this issue. There is a desire to get new products into production, but there are barriers to get individuals to adopt new technology. If we want to accelerate new processes, government program managers and principal investigators need to push the process.

Harvey Newman (Caltech): Solutions, including best solutions, get re-invented over and over. (For Anna Giannakou) Regarding machine learning, you have an interesting dataset, but how do you model the data before training the machine learning algorithms?

Anna Giannakou: They performed feature and correlation analysis prior to generating the dataset, then fed this data to regression models and it worked reasonably well.

Unidentified: The term federation appeared in several talks, also something called inter-segment gateway. What does federation mean, how are federations managed, what are resource discovery and access policies, governance issues, different approaches?

John Graham: Kubernetes federation is an evolving standard. They are currently experimenting with something called Admiralty, which is a 'master' federation for a federation and other clusters join that – the Admiralty developer has made changes in response to their feedback. This is for Kubernetes cluster-to-cluster federation.

Follow-up: That is federation at the infrastructure level, rather than at the application or the science level.

Frank Würthwein: In data (and compute) federations, there are multiple systems that use different software. In terms of governance, people vote with their feet. Federations service somebody – owners join federations because they want to (or not), and users participate in federations (or not). Sometimes federations use common API and it's easier to do, others not so much.

2.9   Keynote – Global Friction-Free Data Exchange – Inder Monga, ESnet

Inder Monga, Executive Director of ESnet, opened the keynote with a simple summary of his talk: "There's Big Data, there's Big Compute and you need a Big Network in between." He wanted to move from talking about transporting bits to the broader issue of data. He described three steps of 'bits to data' – data generation, data logistics, and data analysis and storage. This talk focused on the middle layer of data logistics, including DTNs, networks to transport data, data streaming, resource orchestration, etc., with the talk structured around four areas that are significantly changing within data logistics:

        Network performance -> End-to-end workflow performance
        Human manageable   -> Automation
        Experience-driven   -> Analytics driven.
        Fixed/scheduled   -> Flexible/interactive

The first area is that there is an evolution from looking at network performance to end-to-end workflow performance. An isolated Science DMZ doesn't accomplish anything – it needs to enable users to exchange data with other DMZs. This user-driven, end-to-end performance was exactly the thrust of the PRP project.

Second, there is a move from human-managed networks and systems towards automation. Kubernetes is scaling up the application layer management at DTNs. And the SENSE project automates the provision and resource allocation of the network and DTNs end-to-end. Orchestration and automation are key components of the next-generation ESnet6 design, and there is substantial work being done within DOE on using automation (including artificial intelligence) to advance facility integration across the workflow (see figure below).
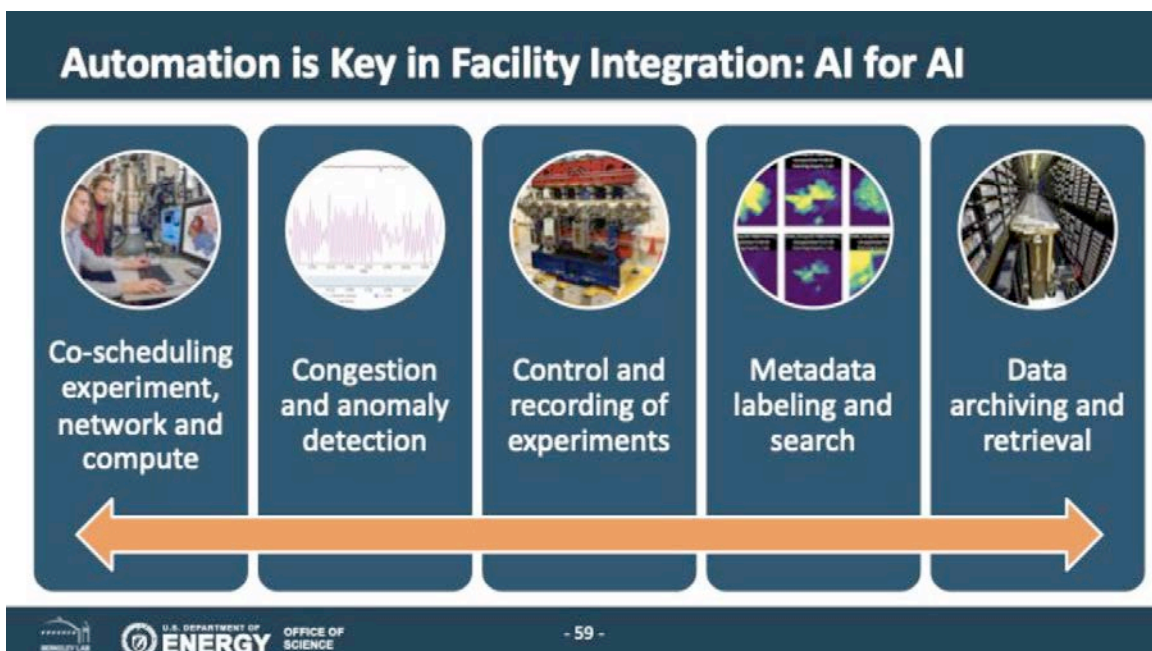

Figure 23: Automation is Key in Facility Integration: AI for AI

The third area is a transition from experience-based to analytics-driven planning and design. Analytics are playing a key role in the design of ESnet6, and he described a number of sophisticated techniques for using past data and current experiments to project future capacity requirements. A research group with Oak Ridge and Argonne collaborators has developed profiles of data throughput infrastructures and defined a metric (the utilization-concavity coefficient) that captures the performance of the entire infrastructure and is useful in predicting good transfers. Finally, there are early investigations into advanced machine learning techniques applied to network traffic data. For example, Google uses these kinds of techniques in predicting automobile traffic/congestion in the future, and perhaps this can be done with networks.

Finally, networks are moving from being fixed/scheduled to being flexible/interactive. The "hollow core" architecture planned for ESnet6 (see figure below) incorporates this flexibility. From a packet's perspective, one can consider "high-touch", "low-touch" and "no-touch" services, analagous to first-class/economy/cargo transport on airplanes.
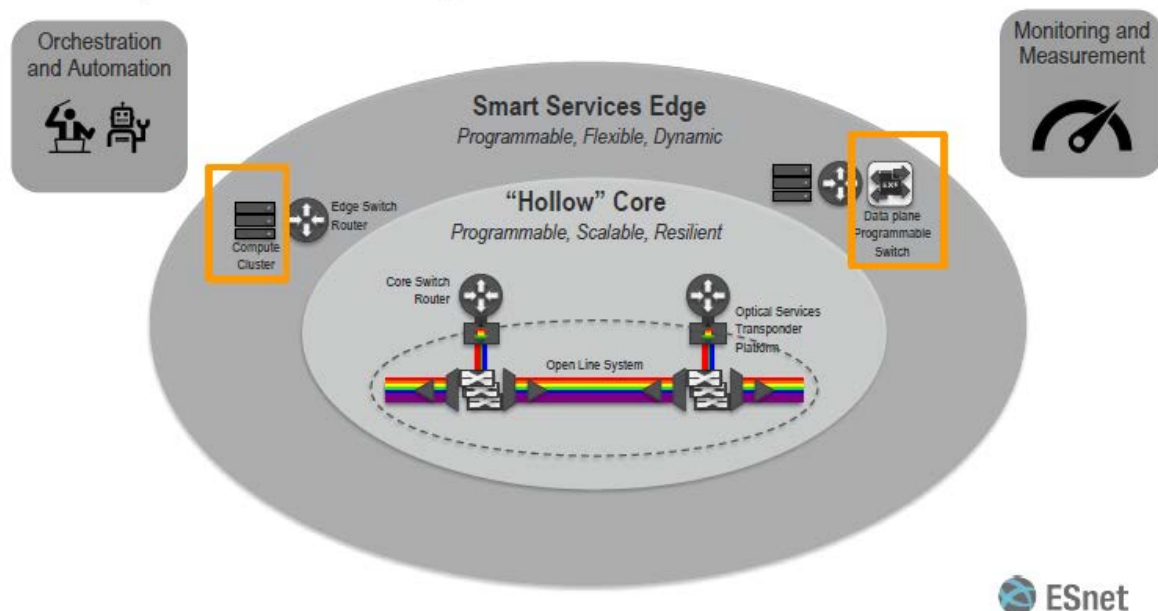
Figure 24: ESnet6 "Hollow Core" Architecture Overview

There is a recent award by NSF called FABRIC, led by Ilya Baldin at UNC/RENCI, which is an "opportunity for the research and education community to focus on the holistic integration of bits, bytes and CPUs." There are a large number of university partners on the award, along with national organizations such as ESnet and Internet2.

In summary, the key elements of a global friction-free data exchange are:
- Multi-domain for R&E needs to extend beyond the boundaries of our connected networks
- Science DMZs, PRP, NRP, and GRP play an important role to promote this view, but we need to keep pushing the boundaries of the workflows and integration
- Without automation of this integration, cannot succeed in managing the complexity
  - APIs, model-driven etc. are important

- ESnet6 and FABRIC are forward-looking infrastructures that the community can use to solve the next-generation Data and Application challenges

Question and Answer Session

Bill Johnston (ESnet): Do you think that SENSE is ready for a production environment – e.g. user authentication/authorization, policies on bandwidth, etc. … like a supercomputer center, there's a fixed pool of resources that needs to be allocated and managed.

Inder Monga: We need to do integration beyond just the network. For example, we can't look at the network bandwidth alone; we need to look at DTNs, storage system, etc. Without answering the question directly, there is a federated identity management system at supercomputing centers that is not discussed much, but could be valuable for this purpose. We need to get this deployed across the complex. There was a lab working group on distributed computing that recommended this project.

Harvey Newman (Caltech): (Comment, referring to the "AI for AI" figure above.) We have one photo of a future system. We've discussed this slide, and have discussed aspects of successful systems. There is no reason to respond to small files because you can combine small files; there is no reason to take a congestion algorithm that is unfriendly to everyone else when someone else can have one that adapts to what's available along the path. Good customers should get preferred service. Those organizations should get on the same side as the network and facility providers. There should be a consistent outcome. It's good to do machine learning on what exists, but it doesn't mean that's part of the solution in the longer-term.

Inder Monga: Several thoughts regarding this comment. Predictability and consistency are critical for applications and we're not there yet as a community. We are trying to get there, building prediction algorithms. It's a complex system that needs to be managed. The FABRIC project can help motivate some of the research we need to get there.

On other hand, building a complex infrastructure that is hard to use to get that predictability is also not going to succeed. For example, we've talked about Layer 2 or Layer 3 circuits. Whatever solution we come up with needs to be easy to use by scientists/applications.

Cees de Laat (UvA): On network predictability, if you add a car to traffic, it doesn't change Google's traffic prediction. If you use a weather prediction, it doesn't change the weather. But if you put an elephant flow on a network, it does impact the performance prediction. So how predictable will these algorithms really be?

Inder Monga: Interesting but we need to try to address predictability for networks. How do we build an infrastructure to try to maintain predictability?

2.10 Panel 5 Theme: Programmable Networking – Moderator Andrew Howard, Australian National University

This panel session included four talks describing international efforts that emphasize Software Defined Networking (SDN) and Software Defined Exchanges (SDX).

2.10.1 KREONET-Softwarization: Virtually Dedicated and Automated Networking over SDN-WAN – Dongkyun Kim, KISTI

KREONET is the NREN for South Korea, and KREONET-S (Softwarization) is the national and international SDN WAN environment for the advanced R&E community over KREONET. The goal is to provide new user services based on virtually dedicated networking and orchestrations. There were originally six deployment locations across Korea and in the US (StarLight), and they have recently added two additional locations in China and the US (Pacific GigaPoP).

The key designs and features of the KREONET-S Virtual Dedicated Networking (VDN) are
- User-based Dynamic On-demand Virtual Network Management
- Logically Isolated and Dedicated Networking w/ High Performance (~100Gbps) and Network Security Provisioned
- ONOS-oriented Event Detection (e.g. Link up/down) and Recovery
- GUI-based Intuitive Virtual Network Creation, Update and Deletion
- Additional App Development: vNAC, vFed, vSlicing, vDHCP, and VDNO

A schematic of the architecture is shown in the figure below. They are using Kubernetes with VDN to orchestrate resource requests and management.

Figure 25: KREONET-S Virtual Dedicated Networking
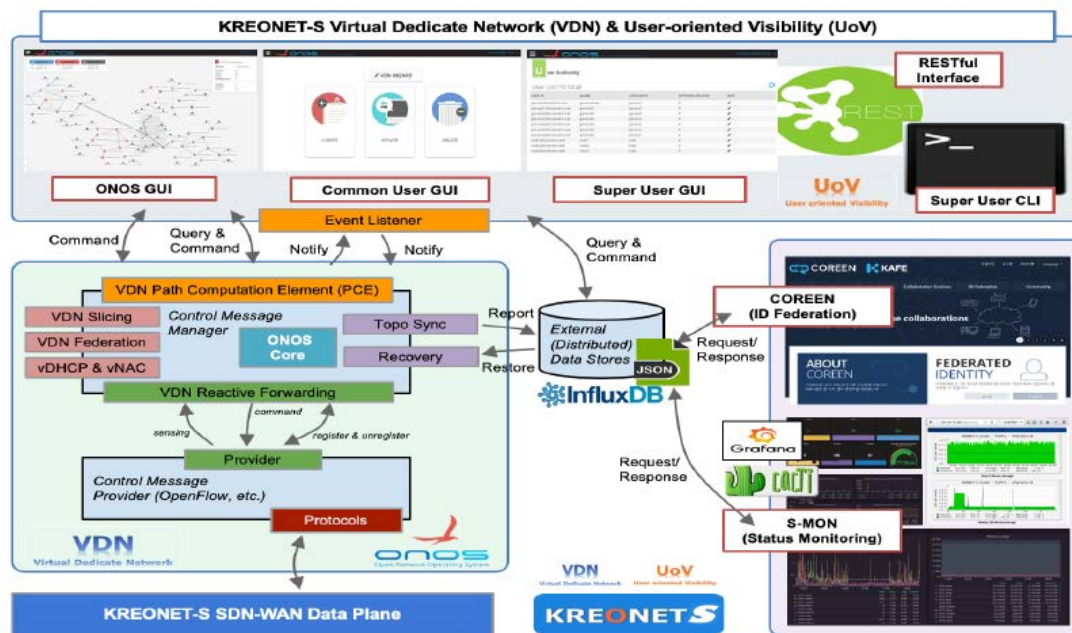
A real-time demonstration of the VDN Orchestrator was shown during the talk. A number of demonstrations over the last two years were described at the Open Network Foundation Connect 2018 meeting and the Asia Pacific Advanced Network 47 (APAN47) meeting in 2019. They are developing collaborations with various international NRENs (Internet2, ESnet, SURFnet, NORDUnet, and CSTNET).

.
A number of use cases were described, including a virtualized edge Science DMZ, Auto-VDN on Globus Online, a VDN for the Internet of Things Datacenters and Gateways, International Federation and Orchestration, and Network Intelligence,

In summary, KREONET-S is moving forward well, further work is planned with new software releases and collaborating sites, and they would like to pursue working with the GRP and other elements of advanced R&E networks.

2.10.2 PacificWave SDN/SDX – John Hess, CENIC

John Hess of CENIC first provided background information on CENIC and PacificWave. CENIC and its California Research and Education Network (CalREN) provides networking to ~20 million individuals across California's public and private universities, libraries, and K-12 schools. PacificWave is a collaboration between CENIC and the Pacific NorthWest GigaPoP (PNWGP), and is an international peering exchange currently serving 31 countries across the Pacific in connecting to the western US. PacificWave enables science-driven high-capacity data-centric projects, such as the PRP, enabling researchers to move data between collaborator sites, supercomputer centers, and campus Science DMZs without performance degradation.

The second element of the talk was a forward-looking view of SDX Technologies and Services. The core capacity of networks is moving from N*100 to 200 and 400 Gbps. Alternative approaches to segment routing were described, including SR-MPLS and SRv6. In orchestration, there are many projects (e.g., AutoGOLE / NSI + MEICAN (RNP), SENSE & BigData Express) for dynamic circuit and services provisioning. Containerization facilitates working across cloud-providers, within-network and on-premises service providers. And there is increasing federated access to infrastructure-attached resources such as heterogeneous compute systems, storage and data repositories.

Third, the routing ecosystem was briefly overviewed, including IANA/ICANN, internet routing registries (IRRs) and regional internet registries (RIRs), network operators, internet exchange points (IXPs), Border Gateway Protocol (BGP), and infrastructure services and tools.

Finally, there was a focus on security for routing and infrastructure services, including Mutually Agreed Norms for Routing Security (MANRS). MANRS is a global initiative to reduce the most common threats to the routing ecosystem, and includes actions for network operators/ISPs (e.g., filtering, IP source validation, coordination and global validation) and actions for internet exchange points (e.g., prevent propagation of incorrect routing, protecting the peering platform, coordination across network operators, and monitoring/debugging tools). Resource Public Key Infrastructure (RPKI) is a key element of implementing routing security and trust models. In general, there has been slow adoption of RPKI across CENIC. CENIC is conducting a pilot with ESnet and others that focuses on MANRS adoption and validation of routing information by implementing RPKI on a regional scale, first among select CENIC and PacificWave research universities (see diagram in figure below), and second to include research universities in other western regional networks.
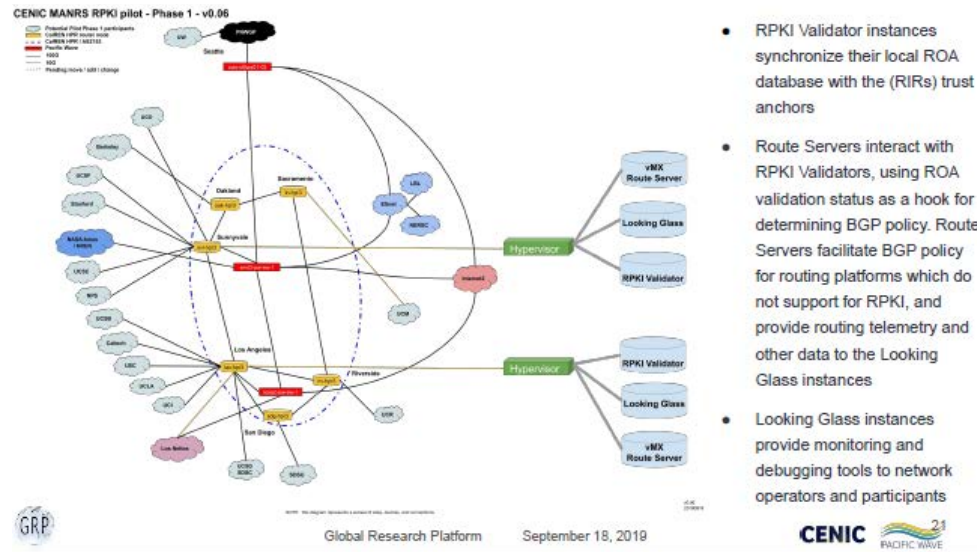
Figure 26: CENIC MANRS RPKI Pilot Project - Phase 1

2.10.3 AutoGOLE/MEICAN/NSI – Gerben van Malenstein, SURFnet

The Automated GLIF Open Lightpath Exchange (AutoGOLE) fabric delivers dynamic network services between open exchanges and networks. It is based on the Network Service Interface (NSI) Connection Service v2.0. They can deliver VLANS, but may also incorporate Layer 3 VPNs. The tool Management Environment of Inter-domain Circuits for Advanced Networks (MEICAN) has been developed by Rede Nacional de Ensino e Pesquisa (RNP) in Brazil as a front-end user interface to AutoGOLE. The figure below illustrates how AutoGOLE operates.
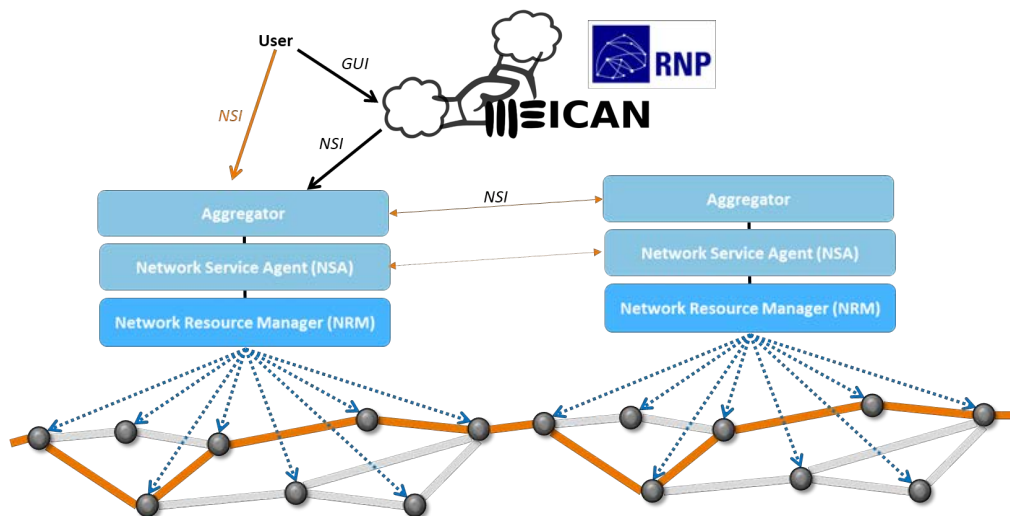
Figure 27: AutoGOLE: Multi-domain network services on-demand

The consortium has grown from initially 100 Gbps in 2013 to 300 Gbps and now 900 Gbps (including ESnet links). Many requests initially came via email to his Inbox and could be handled; still ~250 emails are typically required to set up a circuit, and the system is being targeted to be automated.

Page 45

Activities during 2019 include connecting more networks around the globe using NSI, a dynamic ANA planned for this year (starting with segment Chicago-Montreal-Amsterdam), and further development of MEICAN by RNP. In 2020, the dynamic ANA work will continue to include all ANA links, it may be used as a substrate for the Data Mover Challenge 2020, they are expanding with connectivity to DTNs (perhaps thru SENSE), and finally there is an urgent need to research and implement open multi-resource standard(s) to achieve inter-domain compatibility.

A one-day AutoGOLE workshop is planned to immediately follow this GRP workshop.

2.10.4 Federated International Network Research Testbeds – Joe Mambretti, Northwestern University

Joe Mambretti leads the International Center for Advanced Internet Research (iCAIR) at Northwestern University, which does basic research in transport networks, runs a variety of production networks, and operates approximately 30 national and international research testbeds. The testbeds are the focus of this talk.

One general point is that in something like the GRP, network testbeds can be considered a science research domain in itself, not just support for other science domains.

Networks are the lifeblood of modern economies. Innovations in networks have been driven by computational science. Their efforts are driven by science. As Bill Johnston pointed out, the LHC networking didn't just happen – it took a lot of work to make that a success. Testbeds and prototypes are a vital part of this process (e.g., ITER, LHC, telescopes, etc.).

There are advantages to federated testbeds, and SDN/SDX helps this process substantially. Exchanges are innovation platforms and one can slice them and put testbeds next to production networks.

StarLight is a foundation for the testbed efforts, with 60+ 100 Gbps connections and ~130 private networks. Another foundation is the Global Lambda Integrated Facility (GLIF), which is in the process of migrating to some new form. The focus now is more programmability, software and slicing rather than capacity *per se*. John Hess talked about the SDX interoperable fabric and the exchanges that are being integrated, and Gerben van Malenstein has talked about the AutoGOLE. The AutoGOLE is both a utility and a testbed. There are many international networks coming into StarLight, and SDN is used to slice networks to support various testbeds.

iCAIR has been supporting NSF's Global Environment for Network Innovations (GENI) which has shown that it's possible to slice an environment into hundreds of different tenant networks, each managed by individual research groups. (Similarly, in Europe, there is a series of federated testbeds, the EU FED4FIRE project.) iCAIR has also been supporting the NSF's Chameleon Cloud project, another testbed with tenant networks, and has worked with the LHC communities such as LHCOPN and LHCONE. Experimental capabilities have been developed within these projects. A number of non-LHC particle physics communities are coming on-board into the LHCONE environment, and the model used could be applied to other large-scale projects. They

have a long history of demonstrations using SCinet at Supercomputing conferences, and have participated in the Data Mover Challenge discussed earlier.
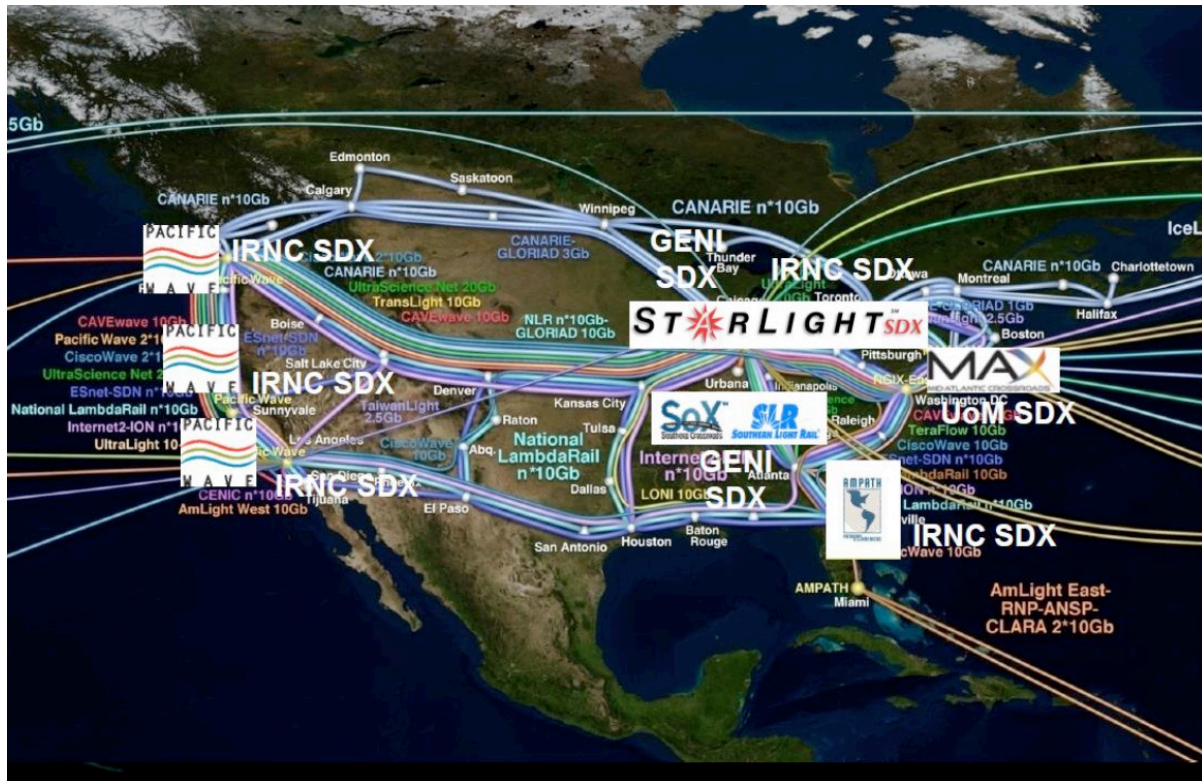
Figure 28: Emerging US SDX Interoperable Fabric

The US has launched a national quantum science initiative, and several federal agencies are starting quantum science programs. iCAIR is organizing with other collaborators in the Chicago area (Fermilab, Argonne, Northwestern) to establish a local quantum communications and networking testbed. With other organizational partners, Northwestern researchers have established an initiative in quantum science. As part of that activity, they are designing a testbed to investigate challenges in quantum communications and networking and integrating quantum networks with traditional networks.

StarLight will be participating in the new FABRIC testbed, an important new program led by Ilya Baldin at RENCI.

The KREONET-S SDN/SDX work described earlier by Dongkyun Kim (KISTI) is excellent.

The GRP should support network research. StarLight has been supporting network research since its inception, and will continue those efforts through the GRP, e.g., supporting the GRP 100 Gbps network.

2.10.5 Panel Question and Answer Session

Frank Würthwein (UCSD): (For Dongkyun Kim) In gravitational wave science, there are a number of interferometers sharing data. OSG supports LIGO and Virgo, but the Korean instrument KAGRA is coming online. Does it make sense to deploy Open Science Grid caches into his Kubernetes infrastructure to support computing access at KISTI?

Dongkyun Kim: Yes, that make sense. They are deploying resources in Korea orchestrated by Kubernetes. He would like to talk with Frank offline on this.

Harvey Newman (Caltech): There was the point raised earlier about the difficulty of organizations working together. We heard about a number of interesting systems being built, but how do we do better in coming together and making the most of the whole picture?

Joe Mambretti: A good start is this forum where we're talking. Some of these projects are siloed projects, but we don't want participants to just go back to their labs and not do anything differently. We need to bring together the resources we have (including people) and we have terrific software stacks and experiences, but we need a mechanism to bring those together.

Gerben van Malenstein: We have many good approaches. The challenge is to scout where the big requests are, where the needs are.

John Hess: We should look at specific science applications that are crossing continents, where people are collaborating across continents on specific projects. Maybe it's not just one approach, but rather combining elements of various approaches.

Cees de Laat (UvA): We used to have the Open Grid Forum where we did NSI and other things. Does this community need a standardization process and if so, where would we find it?

Gerben van Malenstein: Yes, we need it. Not sure where to find it, either in existing communities or create something new.

Bill Johnston (ESnet): (For Joe Mambretti) NASA wanted to do a quantum testbed 5-7 years ago. ESnet had dark fiber between JPL and NASA-Ames, but dark fiber wasn't the issue. The challenge was all the repeaters that had to be replaced from optical amplifiers to quantum amplifiers. Is that why you are just talking about a "Chicago" testbed at this point?

Joe Mambretti: Yes, that's exactly the reason the testbed is currently local. Quantum repeaters don't exist. (See talk by Eden Figueroa in the next panel session.) It is a challenge. But there is another trick. For secure information, you need a highly-guarded 'hut' where quantum state can be changed. But for the testbed, you don't require the same security levels, and the same techniques can can be used to transit information across the ocean. This testbed will be used to investigate how to connect quantum computers for computational science, not security.

Inder Monga (ESnet): Thomas Ndousse-Fetter (DOE Program Manager) just funded 5 proposals for $35M for quantum networking starting October 1. Quantum repeaters are a major challenge, as well as transduction. There is also a quantum institutes proposal being developed, thanks to Congressional funding.

Unidentified (KISTI): (For Joe Mambretti) Is there any activity to make a new protocol like TCP or IP for quantum communication?

Joe Mambretti: There is no quantum TCP or quantum IP. It's an exciting open-ended field. With regard to international activities, there are a number of challenges and it's good to have a large community to address them. For example, not all types of fibers support quantum communications to the same degree. As an anecdote, he received a note from a town in Italy; because they've had an earthquake, they needed to replace all fiber, and they chose quantum-friendly fiber for their town – and offered part of it to researchers to use as a testbed..

2.11 Panel 6 Theme: Next Generation Optical Networking – Moderator: Yves Poppe, NSCC

2.11.1 CESNET Developments in Optical Networking – Michal Krsek, CESNET

CESNET is the Czech national research and education network, and in addition to national networking, it provides international connectivity via GEANT and GLIF. They provide applications and services (e.g. grids, clouds, storage) to their user community and conduct optical research (CzechLight Family of Devices, remote sensing). This talk focuses on the optical research.

With respect to the "CzechLight Family of Devices," CESNET conducts research and development on devices, sometimes in collaboration with industry, develops devices to the prototype stage, then licenses the design to vendors who would productize the devices. Over the years, they have developed EDFA amplifiers, photonic switches, V-MUXes, fixed-grid legacy ROADMs, and WSSes.

One of the latest development 'products' is SDN Reconfigurable Optical Add-Drop Multiplexers (ROADMs), a dynamic optical cross-connect that can be managed by the SDN. These are packaged in 1-U boxes, and can split the signal up to 8 ways. The devices are entirely optical and work on Layer 0. A schematic diagram of the functionality is shown in the figure below. There are various ways to manage/configure the system. Prototypes are available on loan from CESNET and international academic/commercial partners are welcome to evaluate the systems and various management protocols.

They have been doing some 'remote sensing' research in exploiting the fact that physical perturbations to optical fiber (e.g. motion, shakes, pressure) affect the light propagation/ backscatter in the fiber, which can be measured to determine the perturbation magnitude and distance. Potential applications of this include fiber maintenance (e.g. fiber usually is shaken prior to a cut and detecting where that shake occurs can identify the location of the cut), 'military' applications (e.g. detecting whether fiber shakes are due to automobiles versus tanks), or fire detection/location (e.g., if cable is not underground and being heated).
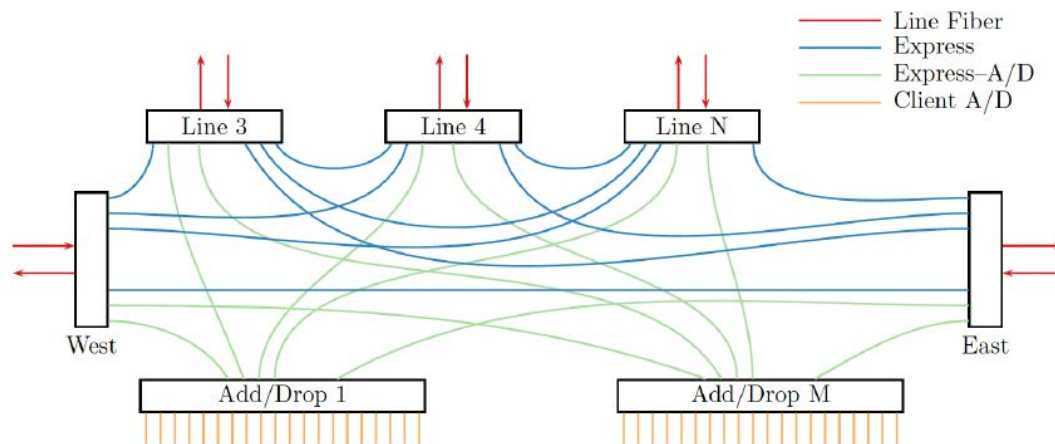
Figure 29: Schematic of the CzechLight SDN ROADM

2.11.2 Introduction to Quantum Networking – Eden Figueroa, Stony Brook/Brookhaven

The talk was in two sections – the first to motivate the importance of quantum computing, and the second to review the activities he and his group have been conducting in Long Island.

In quantum communication networks, imagine sending optical pulses over networks, and instead of many photons carrying one bit of info, you want to have single photons carry the information. There is a difference between bits and qubits. One can have quantum superpositions of two states, zero or one, or any phase in between, so multiple bits are represented in the form of a Qubit. State superpositions in photons can be represented by vertical or horizonal polarization (or left/right circular polarization).

The second tool in quantum communications is quantum entanglement, meaning that quantum states of two or more objects have to be described with reference to each other. They exist only in superposition, and if one measures one of the photons, the other photon collapses, even if they are separated in distance. (Einstein called this 'spooky interaction at a distance.') Optical systems can produce entangled photons (via lasers and nonlinear crystals).

The first application of these principles is quantum cryptography – in particular, one can detect if someone is eavesdropping on the communication. So far, quantum cryptography is the only element of quantum computing that is 'real' now. Another potential application is quantum teleportation – it should be possible to communicate classified information at higher bandwidth.

Research and demonstrations are being conducted in quantum communication, with a focus on limitations to the distances over which quantum networks can work using entanglement (currently ~100 km where quantum entanglement degrades). The Chinese space quantum network is the most advanced, along with work in Hefei, Calgary, Tenerife, and Vienna. Conventional fiber optics use amplifiers but in quantum networks, repeaters are required that transmit entanglement. There are three elements to quantum repeaters needed to transmit over

Page 50

long distances: entangled sources, entanglement 'swapping', and quantum memory/buffers; these are all difficult challenges and repeaters have not yet been successfully built.

The second part of the talk focused on work being done at Stony Brook and Brookhaven National Lab under the umbrella of the Long Island Quantum Information Distribution Network (LIQuIDNet). The focus is to work towards practical quantum communication networks by developing and demonstrating lower-cost, room-temperature quantum repeaters in a regional testbed. They have developed initial prototypes of room-temperature quantum memories and buffers (patent pending), with storage times of ~100 µsec. They are also developing portable rack-mounted entangled sources. They are using existing fiber loops over ~10 km distances at Stony Brook and Brookhaven to evaluate quantum entanglement distribution. The figure below shows the nodes they are trying to connect at Stony Brook, Brookhaven, a remote Stony Brook campus, and eventually a Manhattan station that could be used for further distribution.
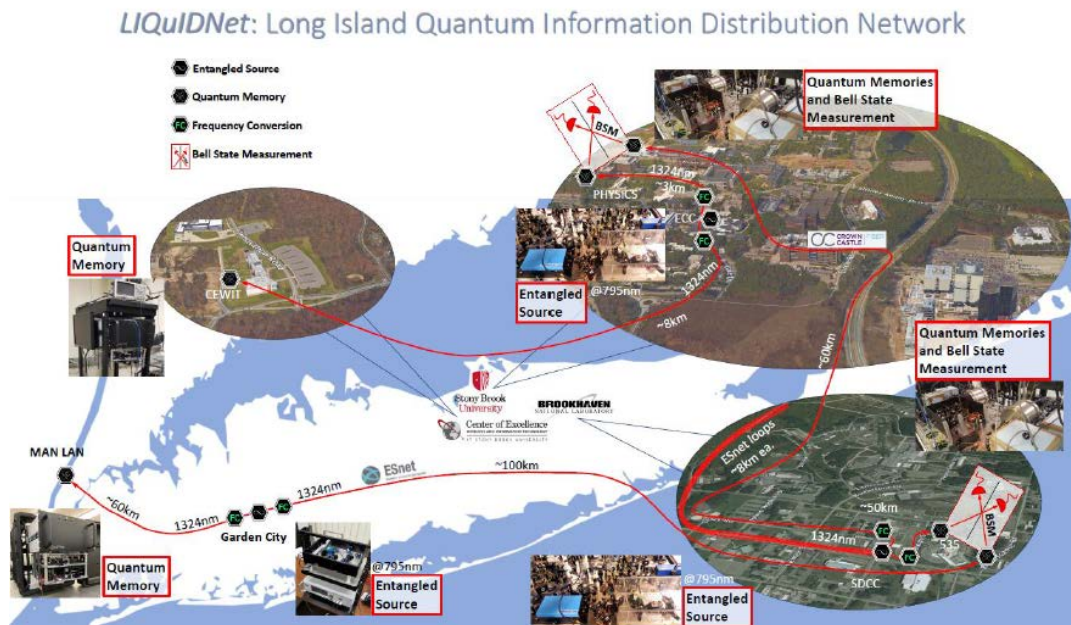
Figure 30: Pictorial Diagram of the Long Island Quantum Information Distribution Network (LIQuIDNet)

2.11.3 ESnet6: SDN-enabled for Big Data Science – John MacAuley, ESnet

There have been several talks already that have discussed ESnet6, including Inder Monga's keynote. This talk focuses on SDN in ESnet6. (ESnet6's optical equipment is currently under bid so it cannot be discussed at this time.)

The "SDN for End-to-end Networked Science at Exascale" (SENSE) program is the external API available to researchers and experimenters to do their orchestrations. ESnet6 provides the high-touch services that Inder Monga described in his keynote talk, including OSCARS. In the ESnet6 'hollow-core' architecture overview (figure included in Monga's keynote above), the

Hollow Core is an open-line system with multi-level transponders and core routers, and is programmable, scalable and resilient. The Smart Services Edge is heavily driven by SDN and is programmable, flexible and dynamic; there are some limitations here with SDN due to scalability.

In the current ESnet operations support suite, the fact is that the network has grown organically over time, with multiple pieces of software. ESnet6 would like to reduce the number of software packages while at the same time using community/third-party software as appropriate. They would like to implement improved orchestration and automation. Five key components are: workflow management (e.g., SURFnet orchestration engine), automated provisioning (e.g., using Cisco Network Services Orchestrator (NSO), network intent (ESDB, their inventory system), network discovery, and network topology. From an automated provisioning perspective, each domain typically has a best-in-class solution; NSO allows them to develop network service models across compute and network resources.

The SENSE architecture is shown in the figure below, and represents a new paradigm in application-to-network interactions. There is an intent-based API for interactive resource discovery, negotiation, service lifecycle and monitoring. There is a real-time Resource Manager (RM) with developed infrastructure and service models. And SENSE represents an end-to-end architecture to address multi-domain networks, end sites and the network stack inside the end systems.
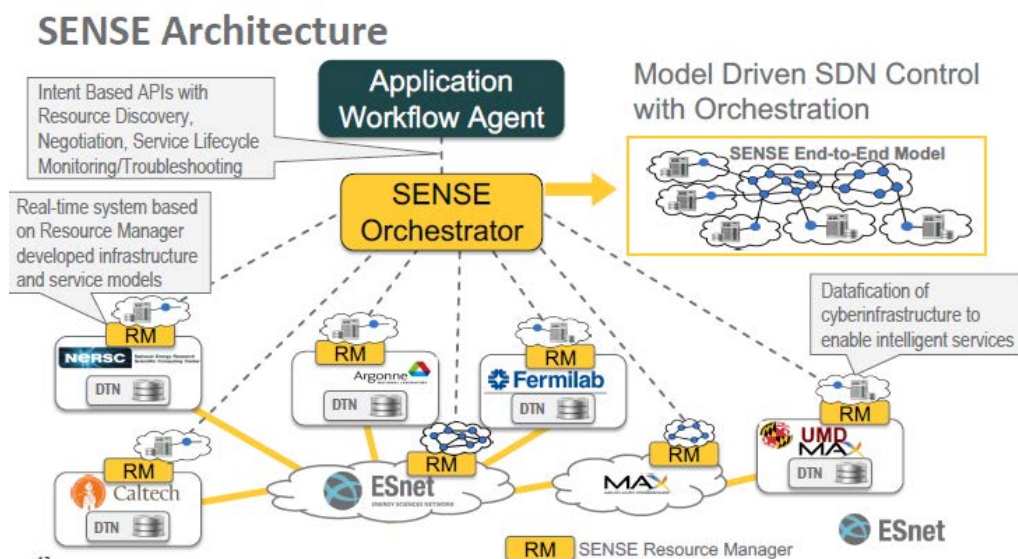
Figure 31: SENSE Architecture

As part of ESnet6's high-touch services, SDN devices will be used to do Precision Network Telemetry (PNT). Modern FPGAs/ASICs can operate counters to identify individual small packets and determine where packets are within ~10 feet. Using the mirroring service on the router, one can touch every packet in the flow, locate packets accurately, and measure buffer fill levels at single-packet resolution. They are picking off specific flows to monitor and examining all individual packets, parsing them and dispatching them to analyzers. To do end-to-end analysis, this information can be forwarded to a centralized collector. Storage will be an issue to

keep up with flows. A simulation-based prototype is being developed of the 'high-touch' analysis and algorithms for real-time analysis of TCP flow properties such as rate monitoring, retransmission, SRTT estimates, bytes-in-flight, and congestion window estimation.

2.11.4 Agile Optical 400G-800G Optical Networking – Marc Lyonnais, Ciena

Marc Lyonnais opened his talk by noting how many times he's heard the word 'more' during the workshop – and he wants to talk about how they will deliver 'more' from an optical perspective.

In ~2009, the transition from 10 to 100 Gbps network rates was facilitated by moving from amplitude modulation to phase modulation. Current projections are that phase-modulated optical components will reach 800 Gbps by ~2021, although it will be challenging to keep up with growing bandwidth in terms of processing, analytics, and management. There is a softwarization of the network. Projections indicate that by 2024, there will be as many 400 Gbps ports deployed as 100 Gbps ports. So capacity is going up, but the cost of networking is stabilizing as the capacity increases.
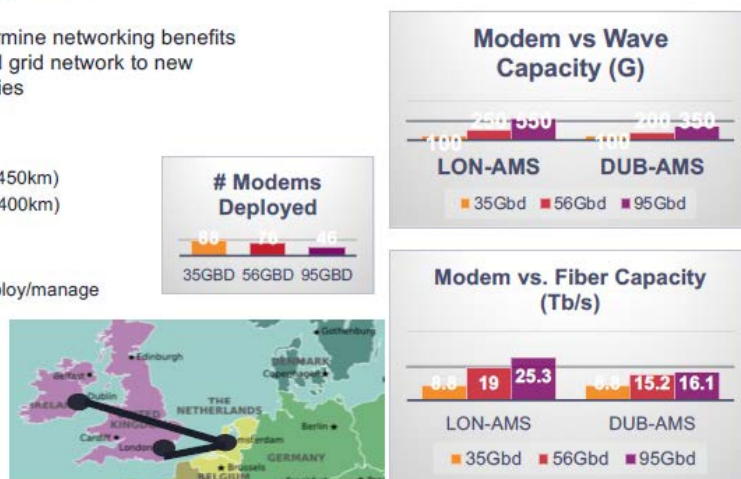
There are two types of solutions for reducing transmission costs: performance-optimized approaches (maximum distance) that are designed for the lowest $/bit-km (but require high-bandwidth electro-optical components), or footprint-optimized with low power (but require small footprint digital signal processing). So, in either approach, one needs the best of both worlds – electro-optical components and digital signal processing. Design choices include CMOS (e.g., 7 nm FinFET) or SiP or InP photonic integration (increases dies/wafer); but system requirements (e.g. automation with variable capacity and baud rates, encryption, latency) remain important.

In selecting the baud rate, one looks at what provides a good line rate for a good distance. There are trade-offs to be evaluated, such as the single-wave baud rate, the number of lambdas used per fiber, and the wavelength band. They have analyzed the potential system benefits in moving from 100 Gbps to new coherent technologies (see figure below).

Figure 32: Network benefits example of next-generation coherent technology

Page 53

The summary points are:
- Upcoming advancements in both DSP and electro-optics will continue to drive down cost/bit: 7 nm FinFET, miniaturization of electro-optics with SiP and InP promise significant benefits.
- Increasing baud reduces networking costs: design and time-to-market depend on high-speed converters and high-bandwidth electro-optics.
- At single-wave 400 Gbps and above, flexible grid network is required. Moving forward, software applications for simple spectral assignment and routing are essential.
- Co-design of both DSP and electro-optics provides advantages in both system performance and time-to-market.

2.11.5 Panel Question and Answer Session

Wenji Wu (Fermilab): (For Eden Figueroa) The quantum memory can currently store ~100 μsec. But processing takes time.

Eden Figueroa: Yes, but progress is being made. Simulations of repeaters show you need ~1 msec.

Follow-up: For quantum memory, in 5 years, how long will the memory last?

Eden Figueroa: In five years, maybe a few hundred msec.

Unidentified: Are you looking at fiber intrusion detection?

Michal Krsek: Yes, this application is similar to the "car versus tank" application.

Follow-up: What about measuring seismic activity?

Michal Krsek: We would need a more earthquake-active area to evaluate this. (Maybe collaborating with California would be good.)

Celeste Anderson (PacificWave): What does room temperature mean for quantum components? In real installations, temperatures vary widely.

Eden Figueroa: For example, room temperature in Phoenix, like 35 deg C. They are close to operating at room temperature, especially compared to most quantum systems.

Eden Figueroa: I am happy to talk with this community.

Harvey Newman (Caltech): (For Michal Krsek) Another use of fiber is acoustic detection, such as seismology or biological signals in the ocean. Are you working on this in CESNET?

Michal Krsek: No, we don't have a sea, but it's a great application. For example, we know things are being tapped on underwater cables.

Harvey Newman: (For Mark Lyonnais) What are the barriers that you see with CMOS and when will more fundamental physics changes happen?

Marc Lyonnais: Quantum networking is a relevant area. If we don't find a new way of populating a fiber, we can always throw another fiber at it. Two fiber pairs could give you 50 Tbps. This is throwing money at the problem but it scales. In the sub-marine market, those cables are very expensive – that's why L band is so popular. He would not be surprised to see optical line rates likely to be InP or SiP.

Unidentified: The barrier is Shannon's limit, where the signal-to-noise ratio impedes the ability to pass information thru fiber. Practically, the solution is per segment optimization; some of the software tools can be a part of this.

Cees de Laat (UvA): Regarding the physics, the stone age didn't end because they ran out of stones. In computing we've hit physical barriers (like frequency), and threw more parallelization at it. With networks, we add more channels/fiber, more fibers. In quantum communication, it presents a large number of states, but we only measure one and then entanglement is gone.

Harvey Newman (Caltech): Just like computers, technology will change as we improve it. We will run out of line size at 3-5 nm. The nanoscale era is between us and quantum era.

Michal Krsek: We're living with an IP stack. We are trying to avoid a TCP barrier. But the window barrier is a feature we created before. Think about stacks we already created.

2.12 Lightning talks: Global Research Platforms – Moderator: Gerben van Malenstein, SURFnet

This panel session consisted of five shorter lightning talks on a variety of national and international research platform efforts.

2.12.1 Asia Pacific Research Platform – Yves Poppe, National SuperComputing Centre (NSCC) Singapore

As a smaller country, it is a challenge for Singapore and other countries to keep up with international research efforts. Collaboration is key to meeting this challenge, and research platforms are a mechanism to facilitate collaboration. Researchers do not want to stitch together VLANs or fine-tune GridFTP in research networks; the next-generation infrastructure will be an easier-to-use research platform.

The Asia Pacific Research Platform (APRP) was initiated in August 2017, and is now an established APAN working group. Since being formed, the collaboration has conducted a series of meetings and demonstrations at APAN and Supercomputing conferences. Several participating countries are establishing national Science DMZs. Next year's Data Mover Challenge (see figure below) demonstrates considerable expansion; it is satisfying to see the

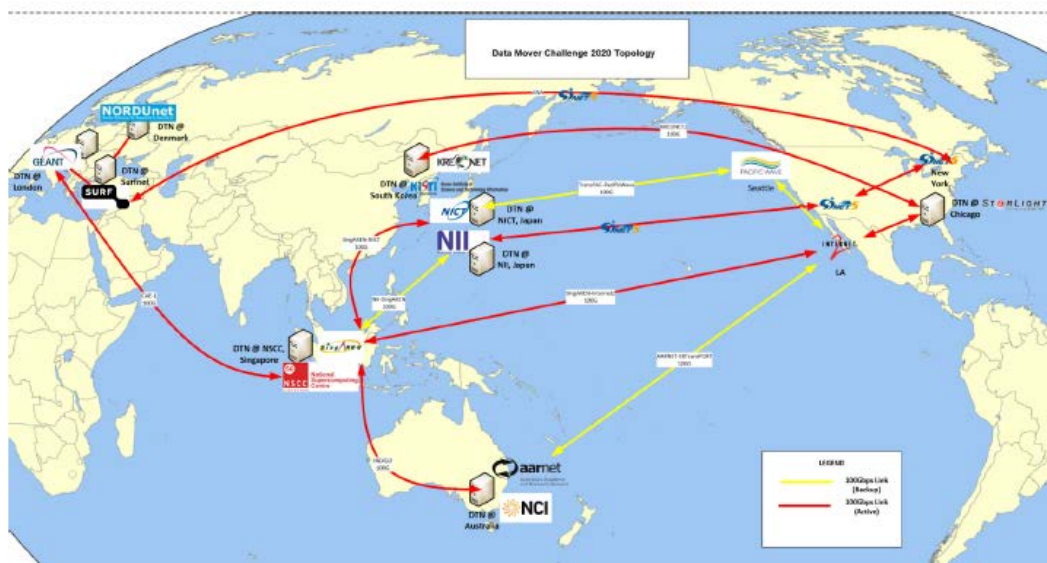collaboration on international links, including explicit cost-sharing, amongst international participants.

Figure 33: Data Mover Challenge 2020 Topology

There are a number of scientific projects that will leverage the outcomes of the APRP, including the Siam light source in Thailand, a shared supercomputer amongst ASEAN countries, the GenomeAsia 100K project, precision medicine, a platform for deep learning, and the Data Wormhole with iCAIR and UCSD.

2.12.2 Australian National Research Platform – Andrew Howard, Australian National University

There are two major supercomputer centers in Australia – the National Computational Infrastructure (NCI) in Canberra in eastern Australia, and Pawsey in Perth in western Australia. NCI is a collaboration between foundational partners ANU, CSIRO, Bureau of Meteorology, Geoscience Australia, and major research universities and institutes, and provides world-class computing resources, cloud resources and data storage to Australian researchers. There are a number of research clouds in Australia, including NeCTAR, operated by NCI and several universities.

A profound change in Australia's networking is that AARNet has recently introduced Indigo, the first system where the NREN owns spectrum; there is now a lot of bandwidth under sovereign control. AARNet announced this week that Indigo has 1.2 Tbps capability, which has been demonstrated to Singapore. Indigo service is based on a sub-sea cable system, touching the Australian coast at a number of points, and terminating in Singapore. This capability is also supplemented by the CAE-1 network providing shorter paths to Europe.

The Australian National Research Platform (ANRP) is a collaboration between NCI and Pawsey, supported by AARNet. They are looking at data movement, file replication, and data stores, all

building on Indigo's advanced network capabilities. Friction-free data movement is the central focus of this work; researchers "just want it to work." The requirements for the ANRP include: regional connection, federated access, data capacitor capabilities (with local storage), container provisioning, and VM provisioning. Containers are central to these functions and Kubernetes will play an increasing role. They have looked at the European Open Science Data Cloud as a good model for cloud collaboration, and have also looked at the Extreme Data Cloud. They appreciate the leadership by the PRP, and are active participants in the Asia Pacific Research Platform (APRP) described above by Yves Poppe.

The Australian National Research Platform (ANRP) uses dedicated 100 Gbps links over AARNet between facilities. The architecture is shown in the figure below. One of the key application drivers for the ANRP is the National Biosciences Cloud Pathfinder, a national facility for processing "omics" data. NCI will soon have a new $70M HPC system and the current HPC system will be re-purposed for the biosciences cloud. The objective is to eventually provide an ambitious set of resources under the ANRP, and they "have started the journey." He concluded by thanking the workshop participants for their presentations and ideas, and he looks forward to bringing back these ideas to the ANRP and the APRP.
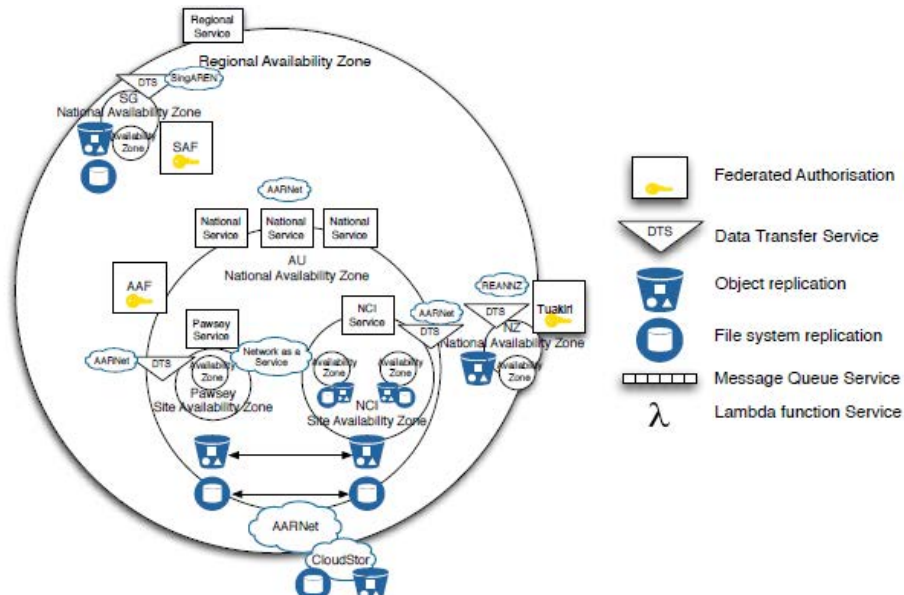
Figure 34: Australian National and Regional Research Platform Architecture

2.12.3 Canadian Research Platform – Florent Parent, Compute Canada

Florent Parent manages an HPC team that is part of Compute Canada, and is also in research IT on the Université Laval campus.

CANARIE manages the Canadian national research and education network, and handles the international peering. Canada also has provinces, each with their autonomous provincial networking organizations.

Compute Canada has seen many changes in recent years, including an effort to consolidate from 25 HPC centers to five centers. The national advanced research computing (ARC) platform (Compute Canada) includes four regional consortia, five sites, 35 member institutions – and across the system includes ~200K compute cores, 290 cloud nodes, 1900 GPUs, 50PB filesystem, and ~210PB tape storage.

All five supercomputing sites have established Science DMZs and are connected via CANARIE at 100 Gbps. But just as happened in the US in earlier years, end-to-end performance for researchers is what matters, and most campuses aren't at 100 Gbps. There is no NSF-equivalent CC* program to get campuses up to speed, but rather there are local initiatives at various campuses. By fortunate circumstances, there was an opportunity to build a Science DMZ at Université Laval (see figure below).
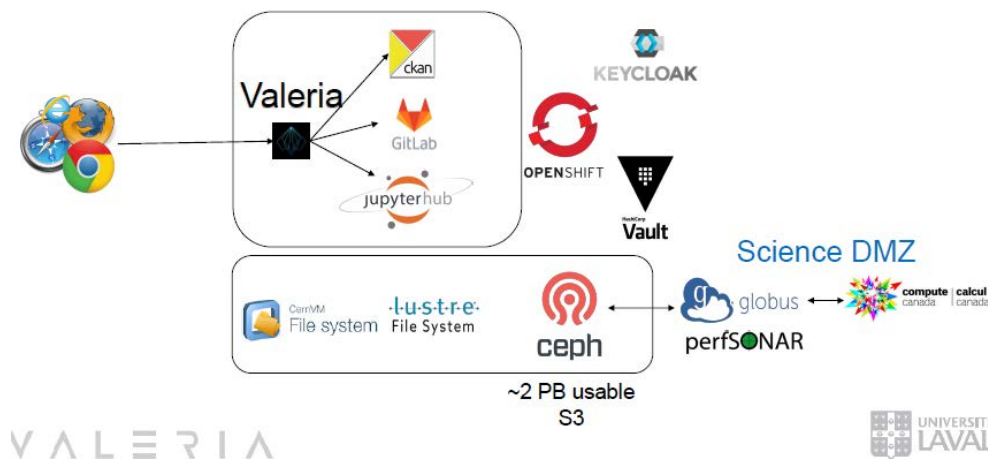

Figure 35: High-Level Architecture Diagram of Science DMZ at Université Laval

He concluded with a description of a new governance mechanism for digital research infrastructure in Canada, including CANARIE and Compute Canada.

2.12.4 University of Warsaw ICM Activities – Marek Michalewicz, University of Warsaw

The original title of "EU Activities" for this talk was "hugely oversized" as there are many different activities within the EU, well beyond his scope. This talk focused on activities at the Interdisciplinary Center for Mathematical and Computational Modelling (ICM) at the University of Warsaw activities.

The Poznan Supercomputing and Networking Center (PSNC) operates Poland's PIONIER NREN, connecting more than 700 centers (see figure below). Most of the main links are 100 Gbps. There are five major HPC centers across Poland, each with ~1 PF computing power.

The University of Warsaw has 42K students, including 5000 graduate students. The ICM has a large datacenter, with two substantial HPC machines. ICM is the main source of support from Poland to the LHC Alice and CMS experiments, and is active in the Research Data Alliance (RDA), and other international programs. There are ~200M/year visits to the website for weather predictions and data.
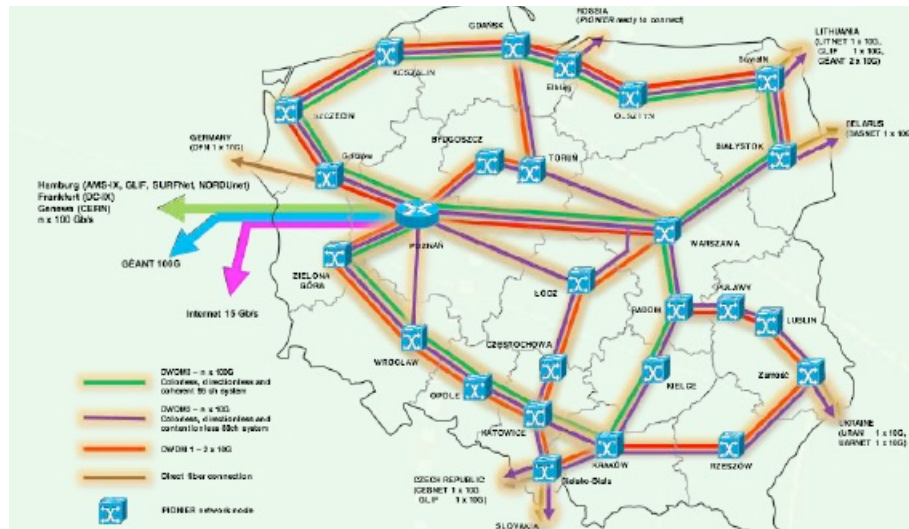
Figure 36: Map of the PIONIER NREN

ICM has established DTNs since 2017, and collaborated in various demonstrations at Supercomputing conferences. They have demonstrated high-speed connectivity between two centers in Poland ~40 km apart, and are working to establish similar capability to Gdansk (400 km). They are beginning a project to test long-haul communication between Warsaw and Pawsey/NCI in Australia. A number of other application drivers to establish a number of other international connections were briefly described.

2.12.5 Korean Research Platform – Buseung Cho, KISTI/KREONET

Korea has a robust NREN, the Korean Research Environment Open Network (KREONET), which was described earlier in Panel 2. KREONET has been active in demonstrating connectivity and supporting science applications with partners on virtually all continents except Antarctica.

Over the years, KREONET has looked at examples of other research platforms and how various institutions have deployed Science DMZs. They are an active partner with the PRP, and have established routine disk-to-disk transfers > 5 Gbps to PRP sites in the US. They are also participating in the US National Research Platform (NRP) effort, the Asia Pacific Research Platform (APRP), the Asian PRP with other APAN members.

Domestically, KREONET has already connected a large number of institutions (also described in Panel 2), with 30 Science DMZs established, and plans to expand this to 50 sites by 2020. An exemplary artificial intelligence initiative is shown in the figure below. The research platform supports many different applications, including astronomy (VLBI, solar observatory, as well as LSST and SKA pathfinders), high-energy physics, weather forecast and climate research, biomedical science and genomics, and fusion energy.
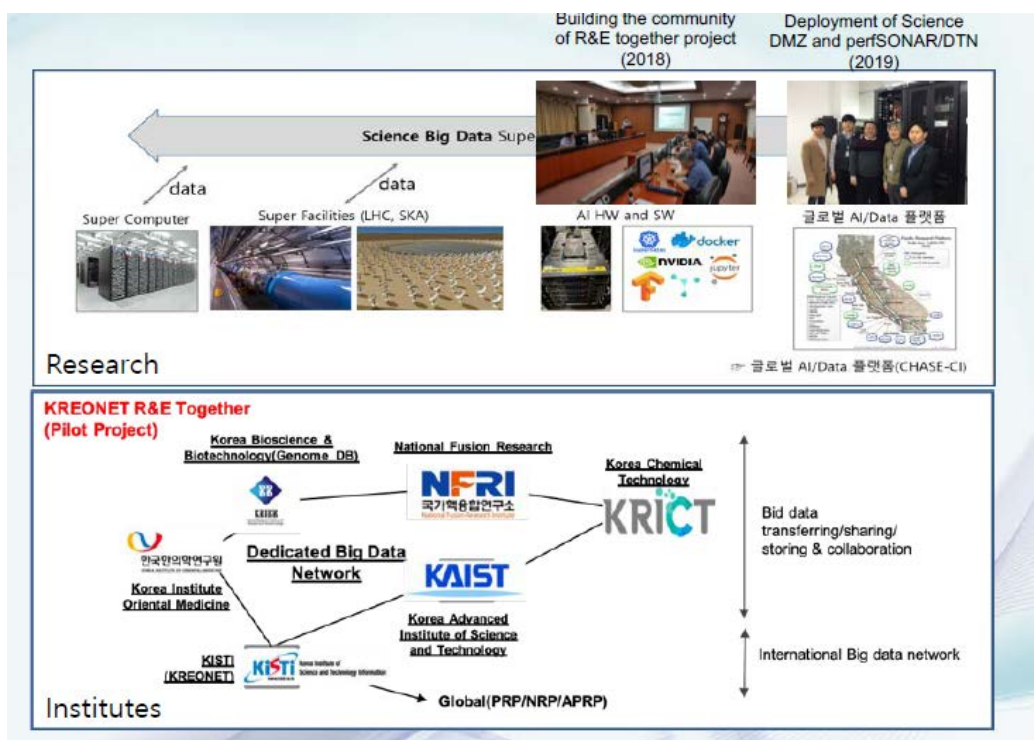
Figure 37: Building AI Platform on big data super-highway

2.13 Closing Session: Next Steps – Larry Smarr, Joe Mambretti, Inder Monga

Joe Mambretti opened the closing session by thanking everyone who contributed to the workshop, especially Larry Smarr as senior advisor and host, Maxine Brown as GRP coordinator, Tom DeFanti as advisor and local host, Charley Erwin as local administrator, the program committee (listed in Section 4), and sponsors UC San Diego, Qualcomm Institute, Ciena, and Juniper. Additional workshop support staff are included in Section 4.

Maxine Brown reported that she will have all workshop presentations posted to a website in a couple weeks, and that a workshop report will be published in the coming months. There are two follow-up meetings planned for the GRP – a BOF session at SC'19 and a second workshop to be co-located with the e-Science meeting in Osaka, Japan in September 2020. Workshop participants will be notified by email with more information about both meetings.

Larry Smarr thanked Joe and Maxine for leading the effort to organize this workshop, Tom DeFanti for his invaluable contributions, and the long list of other contributors mentioned by Joe Mambretti. He also thanked Kevin Thompson (NSF) for attending, and acknowledged Kevin's long support of related programs going back to the OptiPuter program in the early 2000's.

Inder Monga thanked all the participants because it's been an interesting two days of talks! There is a lot going on internationally in research and education networking, and it's interesting to hear what others are doing. These sorts of meetings are where collaborations and new initiatives are often kicked off.

Larry Smarr remarked that he was very impressed with all the work going on worldwide and is hoping that the community can piece together a number of these efforts. We have heard about a number of research platforms around the world, we've spent decades putting in place various network systems, ESnet codified and published the Science DMZ concept, and Google generously open-sourced Kubernetes. We are at a moment when it's relatively simple to plug in DTNs into a global network. Joe Mambretti has been great in pulling together partners from all over the world, and we are in a place where the number of international partners could significantly expand. And we all learn a lot when people join a research platform and face the reality of MaDDash plots and all the technical issues required to make this all work like it can work. Then there are the science applications – at the end of the day, that is why we're moving all this data around. We need to make sure it is simple for end users to do their science. In listening to the science talks, while particle physics and the LHC have led the way, it is clear that we are at a revolutionary moment, with completely digital instruments like SKA and LSST being built at unbelievable scale, and areas like genomics exploding. Climate research will become increasingly important – we are all in this together. Researchers in all countries will be interested in these massive datasets. And while much data is available in repositories, we have to make that data readily accessible. It is relatively low-hanging fruit for this community to make the vision a reality over the next few years to have a GRP where scientific discovery is an ordinary phenomenon.

Joe Mambretti urged the participants to take advantage of the opportunity in front of us. The core resource for this effort is the people in this room. By working collectively and collaboratively, we can build great infrastructure for next-generation science.

Inder Monga commented that as new instruments come online, the instruments and our use of them will keep changing. We need to be ready for new things, to adapt, to share use cases and technologies, to share these changes with one another.

The speakers then opened the floor for questions and comments from the participants.

Harvey Newman (Caltech) wanted to highlight the issue of "convergence." This meeting has been excellent, with application development, technology development, and many countries participating. How do we bring this together, not just for the science community, but for all of society? The implications beyond science are fantastic.

Marek Michalewicz (U Warsaw) said that his perception is that the European Union tends to be inward-looking, and wonders how we can break that mold. He would like to participate in more global efforts, perhaps by participating in testbeds.

Larry Smarr replied that there is a history of networking collaborations with Eastern Europe, such as the 4K transmission with CESNET. Larry continued by noting that Supercomputing as a conference brings together a lot of energy to do one-off demos – that last for a day. While these are great for driving innovation, we need to harvest those demos into ongoing collaborations and operational systems. Furthermore, there's not a big administrative structure in initiatives like the PRP or NRP. For example, we have worked with Cees de Laat for decades, and there is a long

tradition with CineGrid and other international projects. The PRP has only existed four years and has expanded considerably through collaborations. There certainly is a willingness to link up, and we don't need to let the world's political issues stop it.

Inder Monga pointed out that because of different funding agencies and other factors, we sometimes re-create things; it helps to be collaborative and leverage what each other does rather than re-create efforts.

Michal Krsek (CESNET) noted that there tends to be dedication/monopolization of international links for certain projects. While he has supercomputers and FIONAs/DTNs, it is an issue if he needs access to ANA for one day due to policy reasons; there needs to be more democratic access.

Joe Mambretti pointed out that Gerben van Malenstein is working to automate this in ANA, precisely for this reason.

Gerben van Malenstein (SURFnet) replied that these things are possible today, and we need to extend it to other venues.

Inder Monga added that there is modernization and policy work that needs to be done, but we need NRENs to help and there is some onus on NRENs to deploy systems/software that have been developed.

Cees de Laat (UvA) wanted to share his 'philosophical thoughts' on what's happening now. We created GLIF because there were problems back then (e.g., telcos, needed open exchanges). The GLIF structure was not bureaucratic at all – e.g., the charter was two pages. We are at a similar crossroads now. The game is at the data layer. There are automated AI systems that need data. RDM (Research Data Management) is the focus of researchers. Users don't see computers or networks anymore, and students see things as accessible. The PRP and NRP and GRP are moving to that data layer. The GRP is in ideal position to be an unbureaucratic organization/ project to solve that data fabric problem, across the world guided by a number of principles. Thinking about that long-term goal/vision brings us forward, not individual projects from different funding agencies. Let's focus on the long-term goals/visions.

Joe Mambretti concluded the workshop by encouraging everyone to continue on this journey to build cyberinfrastructure for advanced science.

3   Workshop Findings

This section summarizes findings from the workshop, drawn from the presentations and participant comments during the workshop.

- **The global research community will benefit from a broad-based initiative to bring Global Research Platform(s) (GRPs) into existence.**
- **There are numerous examples of current/future, large-scale data-intensive international projects that require GRPs with high-speed end-to-end networking.** The Square Kilometer Array (SKA) and Large Synoptic Sky Survey (LSST) instruments are widely-known examples, but there are many others in astronomy, physics, biology, geophysics, and computer science.
- **The Large Hadron Collider is a successful pathfinder for establishing a dataflow and processing architecture, with networking capabilities and tools to support its experiments.** LHC provides lessons learned for future data-intensive experiments such as the SKA or the LSST.
- **It is the networking community's responsibility to enable data-intensive workflows for scientific applications, and to 'make it easy' for scientists to implement and use these capabilities.** A GRP can provide the tools, personnel, knowledge, and active engagement of researchers to build a seamless platform that will be used in day-to-day scientific processes of discovery. As multiple scientific domains bring new applications, this is an opportunity for disciplines in computing (compute, storage, network/software/middleware) to come together, and build a seamless, science-accessible platform.
- **The 'research platform' concept, focused on implementing high-speed end-to-end networking into scientific workflows, is finding increasing global adoption.** The research platform can connect researchers to other collaborators and a variety of resources such as instruments/facilities, data repositories, HPC centers, cloud compute/storage facilities, and distributed storage systems.
- **The Science DMZ/DTN architecture is an effective means of enabling high-performance end-to-end networking, balancing researcher requirements and network security concerns**. Science DMZ/DTN designs routinely support disk-disk transfers approaching line-speed on 100 Gbps links (Layer 3). Tools such as perfSONAR, GridFTP/Globus and other data movement tools have wide adoption in this community.
- **A GRP cyberinfrastructure needs to be more than interconnected Science DMZs**. A set of Science DMZs interconnected by a high-speed network fabric is a platform to build on; a successful GRP will enable day-to-day scientific applications.
- **New paradigms are required for cost-effectively handling the data movement, processing and storage requirements for next-generation data-intensive instruments and research.**
- **Novel approaches to managing and automating networks are required,** including machine learning and artificial intelligence.
- **Trust is a key element for success of research platforms and systems that share data and resources**. There are trust issues associated with security, especially across institutions, but in most contexts, the reference is to human trust, a vital element in the

effectiveness of collaborative teams, both within a campus and across different institutions. Trust is a human-intensive endeavor, one relationship at a time, and is not readily amenable to scaling.

- **The GRP initiative is a social engineering project as well as a technical networking/IT project**. Many of the stakeholders often do not work together and almost all have separate funding sources and management structures. Stakeholders include scientists/users, science team experts who can work with users and solve end-to-end IT ecosystem issues, campus network engineers and IT personnel from campuses and regional/national networks. Projects like the LHC and its networking/data infrastructure indicate that "shared interests" around a specific domain/experiment facilitate collaboration and trust.
- **Many of the initial use cases for high-speed end-to-end international networks have centered around 'volunteer' efforts or one-off demonstrations** (e.g., at Supercomputing or Supercomputing Asia). These are important efforts to demonstrate and advance technology and establish collaborations, but supporting scientific applications on a day-to-day basis will require sustained systems and funding.

Technology Findings

- **Kubernetes is a "game-changer" in establishing a hypercluster of DTNs in a research platform across trusted partners,** enabling easy access to resources by users and centralized, flexible management.
- **For multiple applications, caching has been demonstrated to be effective in reducing network traffic and latency, and the "data lake" concept reduces storage requirements.**
- **The PRP has successfully shown that DTNs can be multi-purposed to provide storage and compute resources in the Science DMZ.**
- **Software Defined Networking (SDN) and Software Defined Exchanges (SDX) are gaining increasing adoption and proving valuable in advanced networks.**
- **Optical networking technology continues to advance,** with 800 Gbps and Tbps links expected to be available in select locations in 2020.
- **Quantum communication networks are in their infancy, but significant progress is being made.**

4   Workshop Recommendations

This section highlights key recommendations for actions that emerged from the workshop. Note that this workshop represents international participants and hence the recommendations represent the shared interests across the international partners, not any one country or experiment/ collaboration.

- **This Global Research Platform (GRP) forum should be continued and expanded** as a mechanism to facilitate international research platforms by sharing experiences and lessons learned, discussing best practices, conducting outreach and providing use cases to scientific applications, and building a community that advances trust and collaboration.
- **Participants should integrate and leverage technology and lessons learned** from the PRP project, the US NRP, and the myriad of regional, national and international research platform initiatives, in order to effectively move towards a GRP.
- **This community should conduct outreach to scientific collaborations** (science engagement) that can be advanced by improved international networking, particularly those amongst NREN participants that want to advance the GRP concept. It is important to engage scientists who will be users of the system at the outset, identify their requirements, design and build a system that responds to those requirements, and work on an ongoing basis with the science teams.
- **Funding agencies should provide opportunities to secure funding for GRP development**, particularly for sustained support of applications, development of key hardware/software, and advancing international collaborations.

5   Acknowledgements

This workshop was organized by the following individuals:

Senior Advisor and Host: Larry Smarr, Calit2/UC San Diego

Chair: Joe Mambretti, Northwestern University

GRP Coordinator: Maxine Brown, University of Illinois at Chicago

Local Arrangements Chair: Tom DeFanti, QI-Calit2/UC San Diego

Program Committee

Joe Mambretti, Northwestern University, U.S. (Chair)

Buseung Cho, KISTI, KREONET, Korea

Cees de Laat, University of Amsterdam, The Netherlands

Andrew Howard, Australian National University, Australia

Francis Lee, Nanyang Technological University, Singapore

Marek Michalewicz, University of Warsaw, Poland

Inder Monga, Department of Energy, ESnet, U.S.

Yves Poppe, National Super Computing Centre, Singapore

Local Administrator: Charley Erwin, Calit2/UC San Diego

VR/Vis Technical Support: Joel Polizzi, QI-Calit2/UC San Diego

Events Management: Max Carreon, Megan Eastin, and Sara Fam, QI-Calit2/UC San Diego

Registration and Finance: Karen Riggs-Saberton, QI-Calit2/UC San Diego

AV Support: Areli Alvarez (student), Hector Bracho, and Ruben Huerta, QI-Calit2/UC San Diego

6 Appendices

6.1 Workshop Presentations

Documentation for this workshop is available at the GRP workshop website.
http://grp-workshop-2019.ucsd.edu/

This report is posted and can be downloaded from that website.

All workshop presentations are also individually available on that website or can be downloaded as a single Zip file – keynotes (K), panel presentations (#1,2, etc), posters (P) and lightening talks (L) – at https://uofi.box.com/s/u4ls15n42xkamtne47z3qumpa4716x2d .

6.2 Workshop Agenda (Original)

The agenda below is the *planned* workshop agenda, with the original titles for presentations, sequence of speakers and time allotments. The workshop notes above (Section 2) reflect actual titles and speakers.

TUESDAY, SEPTEMBER 17

8:30-12:00   Americas' Research Platform (AmRP) – formerly GLIF Americas (optional, no separate registration)

12:00-1:00   Lunch – AmRP and GRP attendees

1:00-1:15   Welcome and GRP Overview – Joe Mambretti, Northwestern Univ. (NU)

1:15-1:45   Keynote – Moderator: Joe Mambretti, NU

- Research Platforms: Past, Present, Future – Larry Smarr, Calit2/UC San Diego (30 min)

1:45-3:00   Panel 1: GRP Cyberinfrastructure – Moderator: Shawn McKee, University of Michigan

- Science DMZ Global Consideration — Tom DeFanti, UC San Diego/QI-Calit2 (15 min.)
- Large Hadron Collider Open Network Environment – Bill Johnston, ESnet (15 min.
- Nautilus & IceCube/LIGO – Igor Sfiligoi, UC San Diego, SDSC (15 min.)
- Campus Cyberinfrastructure, as implemented on a UC campus – Valerie Polichar, UC San Diego (15 min.)
- Panel Q&A (15 min.)

3:00-3:30 Coffee Break

3:30-5:00   Panel 2: GRP Application Drivers – Moderator: Maxine Brown, UIC

- SKA – Shaun Amy, Australia Telescope National Facility (15 min.)
- LSST – Jeff Kantor, LSST (15 min.)
- Key Global Application Drivers in Korea and Asia – Buseung Cho, KISTI/KREONET (15 min.)
- KSTAR and International Collaborators – Si-Woo Yoon, Korea Superconducting Tokamak Advanced Research (KSTAR) (15 min.)
- High-Luminosity LHC – Harvey Newman, Caltech (15 min.)
- Panel Q&A (15 min.)

5:00-7:00   Reception and Demo Posters

WEDNESDAY, SEPTEMBER 18

8:30-9:00   Keynote – Moderator: Joe Mambretti, NU

- Infrastructure SINET 100G Global Ring and Data Exploitation – Tomohiro Kudoh, University of Tokyo (30 min.)

9:00-10:00   Panel 3: Globally Distributed Data Fabrics – Moderator: Marek Michalewicz, University of Warsaw

- Globally Distributed Secure Data Exchange Fabrics – Cees de Laat, University of Amsterdam (15 min.)
- Data Lakes, Data Caching for Science: OSIRIS Distributed Storage Systems – Shawn McKee, University of Michigan (15 min.)
- Open Storage Network — Christine Kirkpatrick, UCSD/SDSC and National Data Service (15 min.)
- Panel Q&A (15 min.)

10:00-10:30   Coffee Break

10:30-12:15   Panel 4: Data Movement Services – Moderator: Tom DeFanti, UC San Diego/QI-Calit2

- SCAsia Data Mover Challenge – Andrew Howard, Australian National University (15 min.)
- BigData Express – Wenji Wu, Fermilab (15 min.)
- DTN as a Service – Jim Chen, NU (15 min.)
- Next-Generation DTN Architecture/Kubernetes – John Graham, UC San Diego/QI-Calit2 (15 min.)
- The OSG Data Federation – Frank Würthwein, UC San Diego (15 min.)
- Machine Learning for Networking – Anna Giannakou, Lawrence Berkeley Lab (15 min.)

• Panel Q&A (15 min.)

12:15-1:15   Lunch

1:15-1:45   Keynote – Moderator: Cees de Laat, U Amsterdam

- ESnet6 as an International Science DMZ Fabric – Inder Monga, ESnet (30 min.)

1:45-3:00   Panel 5: Programmable Networking – Moderator Andrew Howard, Australian National University

- KREONET-S: Virtually Dedicated and Automated Networking over SDN-WAN – Dongkyun Kim, KISTI (15 min.)
- SDN/SDX – John Hess, CENIC (15 min.)
- AutoGOLE/MEICAN/NSI – Gerben van Malenstein, SURFnet (15 min.
- Federated International Network Research Testbeds – Joe Mambretti, NU (15 min.)
- Panel Q&A (15 min.)

3:15-3:30   Coffee Break

3:30-4:45   Panel 6: Next Generation Optical Networking – Moderator: Yves Poppe, NSCC

- CESNET Developments in Optical Networking – Michal Krsek, CESNET (15 min.)
- Quantum Networking – Eden Figueroa, Stonybrook/Brookhaven (15 min.)
- ESnet SDN-enabled for Big Data Science – John MacAuley, ESnet (15 min.)
- Agile Optical 400G-800G Optical Networking – Marc Lyonnais, Ciena (15 min.)
- Panel Q&A (15 min.)

4:45-5:35   Lightning talks: Global Research Platforms – Moderator: Gerben van Malenstein, SURFnet

• Asia Pacific Research Platform – Yves Poppe, NSCC (10 min.)
• Australian Research Platform – Andrew Howard, Australian National University (10 min.)
• Canadian Research Platform – Florent Parent, Compute Canada (10 min.)
• EU Activities – Marek Michalewicz, University of Warsaw (10 min.)
• Korean Research Platform – Buseung Cho, KISTI/KREONET (10 min.)

5:35-5:55   Closing Session
Next Steps – Larry Smarr, Joe Mambretti, Inder Monga

6.3   Workshop Registrants

A list of registrants, with affiliations and email addresses, is included below.

Biographies for speakers and the program committee are provided in Section 5.4.

First Name	Last Name	Institution
Shaun	Amy	CSIRO
Celeste	Anderson	Pacific Wave/CENIC/USC
Brian	Balderston	UC San Diego
Gavin	Biffar	UC San Diego
Zion	Brewer	Microsoft Corporation
Maxine	Brown	University of Illinois at Chicago
Dhusdee	Chandswang	California State Univ., Fullerton
Jim	Chen	Northwestern University/iCAIR
Xiaowei Alvin	Chiam	NSCC Singapore
Buseung	Cho	KISTI
Kevin	Coakley	UC San Diego
Amir	Dabirian	California State Univ., Fullerton
Susumu	Date	Osaka University, Japan
Cees	De Laat	University of Amsterdam
Thomas	DeFanti	UC San Diego
Eden	Figueroa	Stony Brook University
Dale	Finkelson	Internet2
Antonio	Francisco	ANSP
Anna	Giannakou	LBNL
John	Graham	UC San Diego
Jason	Haga	AIST
Michal	Hazlinsky	CESNET
John	Hess	CENIC and Pacific Wave
Rommel	Hidalgo	California State Univ., Fullerton
Paul	Hiew	NSCC Singapore
Bruno	Hoeft	Karlsruhe Institute of Technology
Andrew	Howard	NCI Australia
Yihsuan	Huang	UC San Diego
Thomas	Hutton	UC San Diego
Rogerio	Iope	Sao Paulo State Univ. (UNESP)
JJ	Jamison	Juniper Networks
Ron	Johnson	U Washington – Pacific Wave
William	Johnston	ESnet/Lawrence Berkeley Lab
John	Kan	A*STAR
Jeff	Kantor	LSST
Akbar	Kara*	LEARN
Jonah	Keough	Pacific Northwest GigaPoP
Dongkyun	Kim	KISTI
Ki-Hyeon	Kim*	KISTI
Christine	Kirkpatrick	UC San Diego
Michal	Krsek	CESNET
Li-Chi	Ku	NCHC Taiwan
Tomohiro	Kudoh	The University of Tokyo
Jim	Kyriannis	NYSERNet
Craig	Lee	NIST Cloud Federation WG
Francis	Lee	SingAREN
Richard	Lethin	Reservoir Labs

Jen Ngee Jerry	Lim	NSCC Singapore
Vincent	Lim	NSCC Singapore
Lance	Long	University of Illinois at Chicago
Luis	Lopez	Florida International University / University of Sao Paulo
Marc	Lyonnais	Ciena
John	MacAuley	ESnet
Joe	Mambretti	Northwestern University/iCAIR
Shawn	McKee	University of Michigan Physics
Kevin	Meynell	GLIF
Marek	Michalewicz	University of Warsaw
Dima	Mishin	UC San Diego
Inder	Monga	Lawrence Berkeley National Lab
Jeong-Hoon	Moon*	KISTI
Richard	Moore	UC San Diego (retired)
Heidi	Morgan	USC Information Sciences Institute
Azher	Mughal	USC
Harvey	Newman	Caltech
Florent	Parent	Université Laval
Chanjin	Park	KISTI
Bram	Peeters	GEANT
Willie	Peng	California State Univ., Fullerton
Valerie	Polichar	UC San Diego
Yves	Poppe	A*STAR NSCC Singapore
Brett	Rosolen	AARNet
Marcos	Schwarz	RNP
Igor	Sfiligoi	UC San Diego
Gauravdeep	Shami	Ciena
Jerry	Sheehan	San Diego State University
John	Silvester	USC
Shava	Smallen	UC San Diego
Larry	Smarr	UC San Diego
Jerry	Sobieski	NORDUnet
Alex	Szalay*	Johns Hopkins University
Ryousei	Takano	AIST
Atsuko	Takefusa	National Institute of Informatics
Thomas	Tam	CANARIE Inc.
Tin Wee	Tan	National Supercomputing Centre Singapore
Kevin	Thompeson	National Science Foundation
Gerben	Van Malenstein	SURF
Rick	Wagner	Globus
Wassapon	Watanakeesuntorn	Nara Institute of Science and Technology
Rod	Wilson	Ciena
Linda	Winkler	Argonne National Laboratory
Lawrence	Wong	National Supercomputing Centre, Singapore
Stephen	Wong	NSCC Singapore
Wenji	Wu	Fermilab
Frank	Würthwein	UC San Diego
Yufeng	Xin	RENCI, UNC
Xi	Yang	University of Maryland / MAX
Si-Woo	Yoon	NFRI/KSTAR
Juan Sebastian Aguirre	Zarraonandia	Osaka University

* Registered for workshop but unable to attend.