Funding Agency: National Science Foundation, Division of Advanced Cyberinfrastructure
Award number: 1826967
Project title: CC* NPEO: Towards a National Research Platform

Project Senior Personnel: Larry Smarr, PI;
Ilkay Altintas, Co-PI
Tom DeFanti, Project Manager;
Phil Papadopoulos, Co-PI;
Tajana Rosing, Co-PI
Frank Wuerthwein, Co-PI;

Submission date: July 1, 2019
Recipient organization: University of California San Diego, Office of Contracts & Grants Admin., La Jolla, CA 92093-0934
Project/Grant period: 10/01/2018–9/30/2021
Reporting period end date: 10/01/2018–9/30/2019
ACCOMPLISHMENTS

1 What are the major goals of the project?
Academic researchers need a simple data sharing architecture with end-to-end 10-to-100Gbps performance to enable virtual co-location of large amounts of data with computing. End-to-end is a difficult problem to solve in general because the networks between ends (campuses, data repositories, etc.) typically traverse multiple network management domains: campus, regional, and national. No one organization owns the responsibility for providing scientists with high-bandwidth disk-to-disk performance. Toward the National Research Platform (TNRP), addresses issues critical to scaling end-to-end data sharing. TNRP will instrument a large federation of heterogeneous national-regional-state networks (NRSNs) to greatly improve end-to-end network performance across the nation.

The goal of improving end-to-end network performance across the nation requires active participation of these distributed intermediate-level entities to reach out to their campuses. They are trusted conveners of their member institutions, contributing effectively to the “people networking” that is as necessary to the development of a full National Research Platform as is the stability, deployment, and performance of technology. TNRP’s collaborating NRSNs structure leads to engagement of a large set of science applications, identified by the participating NRSNs and the Open Science Grid.

TNRP is highly instrumented to directly measure performance. Visualizations of disk-to-disk performance with passive and active network monitoring show intra- and inter-NSRN end-to-end performance. Internet2, critical for interconnecting regional networks, will provide an instrumented dedicated virtual network instance for the interconnection of TNRP’s NRSNs. Cybersecurity is a continuing concern; evaluations of advanced containerized orchestration, hardware crypto engines, and novel IPv6 strategies are part of the TNRP plan.

The Pacific Research Platform (PRP 2015–2020, NSF # 1541349) project has been focused for the past 4 years on academic researchers’ needs for a functioning high-speed big data sharing architecture enabled by end-to-end 10-to-100Gbps connections. TNRP, this project, offers a fundamentally top-down approach to address the issues of science engagement and scaling successful PRP strategies to a national level, working directly with network leaders, campus CIOs, and supercomputer center directors. The TNRP on a functional level, however, conducts its technical engagements at the grass roots level, so to speak, nurturing interactions between regional/national network technicians, cyberinfrastructure support staff, and researchers and their students, with the fundamental goal of achieving better understanding of detailed network performance through measurement, monitoring and analysis. Previous end-to-end programs tended to focus on the “ends” being campus gateways, rather than the “ends” being the computers holding and using the shared data.

(The NRP Pilot is an experimental, TNRP-complementary collaboration led by Internet2 with overlapping partners and goals. The NRP Pilot capitalized on the momentum of The First National Research Platform Workshop in Bozeman, MT in August 2017, where a strong national community emerged followed by the Second National Research Platform Workshop in Bozeman, MT in August 2018. The NRP Pilot is very actively leading and planning the planned
TNRP has 5 goals over three years, the first of which includes 11 specific objectives specified in the solicitation, and 4 more that are overarching. This Annual Report covers the first 9 months of the TNRP award.

Goal#1: Address the 11 NPEO specific objectives in the NPEO solicitation:

[NPEO#1] Identifying potential bottlenecks in data transfers for scientific collaborations.
[NPEO#2] Providing analysis and objective recommendations on tools, technologies, and integrated systems related to data network movement and transfer performance.
[NPEO#3] Serving as a centralized point of expertise, advice, and engagement for distributed scientific collaborations in identifying and solving scientific data movement inefficiencies adversely affecting time-to-science.
[NPEO#4] Disseminating current practices across the NSF community on achieving highly efficient end-to-end data transfer performance, especially in high-bandwidth/high-delay environments.
[NPEO#5] Providing online and in-person training and education on network performance analysis and troubleshooting.
[NPEO#6] Engaging with CI Engineer/Advanced Cyberinfrastructure-Research Education Facilitator/Cyberteam award personnel, and other relevant groups comprising the collective CI engineering expertise in the community in which network performance plays an important role.
[NPEO#7] Investigating, evaluating, and testing data movement systems and tools, transport protocols, and integrated platforms.
[NPEO#8] Selectively investigating and solution engineering for end-to-end performance.
[NPEO#9] If appropriate, re-engineering data transfer protocols and tools (note that any software developed in the CC* program require open source licensing).
[NPEO#10] Analyzing network performance and data transfer metrics, including periodic analysis to demonstrate trends and assist scientific collaborations in optimizing CI resources.
[NPEO#11] Defining a role in improving the community’s effectiveness in leveraging the significant distributed deployment of PerfSonar and associated active performance testing.

These specific objectives are discussed in detail in Sections 2 and 3.

Goal#2: Address issues critical to scaling PRP strategies

Toward the National Research Platform (TNRP) Goal#2 addresses issues critical to scaling PRP strategies toward a time when a financially viable National Research Platform (NRP) might be solicited. TNRP aims to prototype the NRP of 2 or so years hence. In addition, the TNRP is working closely with the Open Science Grid (OSG) to adopt/adapt its very scalable high-throughput computing techniques.
Goal#3: Build facilities and support for TNRP measurement and monitoring
This TNRP goal involves adopting nationally-applicable tools to instrument and measure performance using active and passive techniques.

Goal#4: Gain participation by a federation of regional and national networks
The goal of improving end-to-end network performance across the nation requires active participation of regional distributed intermediate-level entities to reach out to their campuses, just as the PRP has relied on its four regions. In addition, the participation of Internet2 has been strong (described in detail below). TNRP’s 3-year plan involves participation by a number of NRENs: Internet2 and ESnet, and state and regional networks: LEARN, KINBER, MCNC, the NM Tribal Colleges Consortium, Great Plains Network and Southern Crossroads, as well as the PRP regional partners: CENIC, Pacific Wave, Front Range GigaPop, and the Chicago-based MREN.

Goal#5: Identify national-scale science drivers
The TNRP partnership and the Open Science Grid (OSG) are working together to identify new science drivers; the leadership of the TNRP and the regional and national networks work with national labs and supercomputer center directors to identify national-scale science drivers, as described below.

Next, we describe how we have made progress in these specific objectives in the first 9 months of this award.

2. What was accomplished under these goals? (Must provide an answer to at least one of the subheadings below.)

   a. Describe the major activities (8000 max) (7893 so far)

Goal#1 has 11 objectives to be met over the 3 years. We list four more, Goals#2–#5, below.

Goal#1 activities: Addressing the 11 NPEO specific objectives:
See Sections 2b-d, 3, and 4, below; these 11 specific objectives’ activities are described.

Goal#2 activities: Addressing issues critical to scaling PRP strategies
TNRP is not an extension of the PRP, but rather a national-scale collaboration of regional and national networks that includes the PRP’s regional networks (CENIC, Pacific Wave, The Front Range GigaPoP, and MREN) as well as Internet2, LEARN, KINBER, NYSERnet, and the Great Plains Network. ESnet is also a participant. A major activity is making network measurements on a national-scale version of a PRP-like collaboration of regional networks, but at greater geographic scale than the PRP.

TRNP’s key scaling approach is to help participating regional organizations engage with their own scientists as drivers. In the first 18 months (the first half of which are part of this annual report), we are focusing on CENIC, Front Range GigaPop, Pacific Wave, MREN, LEARN, KINBER, NYSERnet, and the Great Plains Network. Weekly, TNRP/PRP researchers and technical staff provide analysis and objective recommendations on tools, technologies, and integrated systems related to data network movement and transfer performance through Zoom.
meetings. We maintain data capture and archiving for MaDDash, Elastiflow, and Prometheus/Grafana visualizations.

**Goal#3 activities: Building facilities and support for TNRP measurement and monitoring**

PRP and TNRP’s teams aim to prototype a new national-scale facility for network measurement and monitoring. Our hyper-converged Nautilus infrastructure consolidates software-defined compute, storage, and networking into one highly-distributed cluster to provide high-performance, cost-effective, and easily scalable virtualization. As of June 2019, Nautilus consists of 122 nodes distributed among 23 PRP partner sites, including >3800 CPU cores, >370 single-precision GPUs, as well as two Xilinx U200 Alveo FPGAs for computing and 100G P4/SDX networking. See [http://pacificresearchplatform.org/nautilus/cluster-map/#bynode](http://pacificresearchplatform.org/nautilus/cluster-map/#bynode) for details.

Nautilus uses Google’s open source Kubernetes (K8s), a now widely-adopted (by the major cloud providers and by over half of the Fortune 500 companies) way to manage containerized software. It is constantly being enhanced by open source packages to manage shared storage, which TNRP-paid staff implements eagerly. It supports current TNRP measurement and monitoring of distributed computing and shared storage (see the supplement for visualizations). K8s not only measures and monitors data transfer speeds using perfSONAR (memory-to-memory) over networks between Science DMZs, but also CPU, GPU, and (in Year 2) storage usage per application user group. Nautilus’ best practices are meant to be adopted by TNRP partners.

Visualization tools (e.g., MaDDash, Elastiflow, and Grafana) for analyzing Nautilus node-to-node performance with network monitoring are being exploited to show end-to-end performance. The TNRP team is currently experimenting with m-lab’s tcp-info ([https://www.measurementlab.net/blog/category/tcp-info/](https://www.measurementlab.net/blog/category/tcp-info/)) and Tstat, which analyzes either real-time captured packet traces or previously recorded packet-level traces.)

Internet2, critical for connecting regional networks, provides DTNs as part of Nautilus at three locations (New York, Chicago, and Kansas City). Please see the supplement for visualizations, maps, spreadsheets, and other supportive documentation that are not possible to include in these text-only sections.

**Goal#4 activities: Gaining participation by a federation of regional and national networks**

Several regional (the Great Plains Network, LEARN, KINBER, and NYSERNet) and national networks (Internet2, ESnet) in TNRP have already installed ESnet-like Data Transfer Nodes (DTNs) and/or FIONAs (Flash I/O Network Appliances, DTNs we optimize and support). These DTNs enable the network measurement and monitoring we are reporting on by creating “smart” endpoints to the optical networks. There are several regional-level MaDDash displays, and there are participants in TNRP’s national GridFTP MaDDash. Engineers participate in weekly technical Zoom calls and are adopting PRP strategies. Their set of willing networks covers the country spatially and structurally (but not fully), with much diversity.

After the TNRP proposal was submitted, but before it was awarded, the NRP Pilot was launched at a meeting at Internet2’s Global Summit in San Diego in May 2018 in order to maintain momentum among the many TNRP partners. The *NRP Pilot* is an experimental, TNRP-complementary collaboration led by Internet2 with overlapping partners and goals. The NRP
Pilot capitalized on the momentum of The First National Research Platform Workshop in Bozeman, MT in August 2017, where a strong national community emerged. A Second National Research Platform Workshop in Bozeman, MT was held in August 2018. The NRP Pilot is very actively leading and planning the planned Third National Research Platform Workshop in Minneapolis in September 2019. In fact, TNRP and the NRP Pilot have almost entirely overlapping goals and partners, by design; the NRP Pilot, however, has no specific reporting obligations and no specific end date.

The NRP Pilot continues a year later as a multi-partner, collaborative effort with these objectives and expected outcomes:

- Engage NRP teams and collaborators and facilitate in-person interactions.
- Gain a common understanding of the science engagement process.
- Explore and capture data storage, computing, and networking requirements across scientific domains.
- Identify common science-driven technical requirements for the NRP.
- Exchange technical ideas for scaling the NRP.

The NRP Pilot is helmed by the following Steering Committee, which e-convenes 4 times a year:

- Howard Pfeffer, Jim Bottum, Dana Brunson, Cathy Chaplin, Rob Vietzke, Internet2
- Larry Smarr, Tom DeFanti, UCSD
- Pankaj Shah, LEARN
- James Deaton, Great Plains Network
- Wendy Huntoon, KINBER
- Louis Fox, CENIC
- Jen Leasure, The Quilt

**Goal#5 activities: Identifying national-scale science drivers**

An NRP Pilot Engagement Team and an NRP Pilot Technical Team were also formed. The Engagement Team’s goal is to expand science engagement through personal networking. It focuses on discipline-specific solutions and seeks buy-in from campus leadership. The Engagement Team e-convenes bi-weekly (nrp-pilotengagement@internet2.edu) and consists of the following members:

- Jim Bottum, Dana Brunson, Cathy Chaplin, Internet2
- Wendy Huntoon, KINBER
- James Deaton, Great Plains Network
- Jim Kyriannis, NYSERNet
- Jen Leasure, The Quilt
- Louis Fox, CENIC
- Larry Smarr, Tom DeFanti, UCSD
- Pankaj Shah, LEARN
- Camille Crittenden, UC Berkeley

Now merged with the weekly PRP technical call, the following NRP Pilot Technical Team members e-convene weekly (nrp-pilottechnical@internet2.edu):

- Akbar Kara, LEARN
- Jim Bottum, Dana Brunson, Cathy Chaplin, Rob Vietzke, Internet2
The Technical Team addresses scaling challenges: integrated routing architecture, federated container infrastructure, DTN lookup service, additional data transfer tools, improved measurement and visualization, as well as network interruptions and other failures (software version mismatches, etc.). Many technical networking and cyberinfrastructure support personnel join the weekly zoom calls. Tom DeFanti takes and archives notes from each hour-long meeting, now in their 4th year. The participants on these calls are noted. About 150 people are invited and get the notes every week, about 20–24 call in every Thursday, spanning a much larger group since it is not always the same people. To join the PRP Listserv visit: https://mailman.ucsd.edu/mailman/listinfo/prp-l. It is very active, with >150 subscribers.

b. Describe the specific objectives

[NPEO #1 specific objectives] Identifying potential bottlenecks in data transfers for scientific collaborations.

TNRP is a combination of PRP/ESnet technology (like the FIONA DTNs and MaDDashes) and Open Science Grid (OSG) technology adoption. OSG’s experience is that some researchers, particularly those whose research requires large data transfers or computation, are already well aware of the network limitations in their applications, while many others use cyberinfrastructure in their lab/campus/centers, simply accepting the available performance rather than trying to identify and unplug bottlenecks. Science engagement efforts help identify those users whose network bottlenecks adversely affect their large data-transfers.

We developed a MaDDash, which helps spot and then improve underperforming connections and 10GB GridFTP data transfers among our 17 Year-1 NRP Pilot sites (see https://perfsonar.nrp-nautilus.io/maddash-webui/, Section 2c, and the supplement for corresponding visualizations).

[NPEO#2 specific objectives] Providing analysis and objective recommendations on tools, technologies, and integrated systems related to data network movement and transfer performance.

TNRP staff intimately work with CENIC, Internet2, and ESnet staff who connect TNRP’s regional networks to troubleshoot networks end-to-end. Internet2 works with TNRP and regional staffs to establish and manage operational practices.

[NPEO#3 specific objectives] Serving as a centralized point of expertise, advice, and engagement for distributed scientific collaborations in identifying and solving scientific data movement inefficiencies adversely affecting time-to-science.

TNRP disseminates information through websites (nationalresearchplatform.org), online documentation and code repositories (see Products/Section 4 below). It offers expertise, advice,
and engagements with Zoom calls and an active, real-time chatroom. See https://rocket.nautilus.optiputer.net. TNRP also intends to cooperate with the Engagement and Performance Operations Center (EPOC) as much as possible in Year 2.

**[NPEO#4 specific objectives]** Disseminating current practices across the NSF community on achieving highly efficient end-to-end data transfer performance, especially in high-bandwidth/high-delay environments.

Our RocketCHAT chatroom gives networking and cyber engineers a quick way of getting and giving guidance, or help, or reach out to others en masse.

TNRP resource users access Nautilus via CIlogon authentication that works with campuses’ single sign-on or Google’s credentials. We worked with CIlogon to bring the token refresh service up to standards to work with K8s directly.

The MaDDash visualizations and the mouseover data are an easy way for users to set expectations for end-to-end data transfer performance as demonstrated between campuses in the NRP Pilot’s national footprint. Users can then determine if poor performance is between the campus Science DMZ and their lab’s computers, or at the other end, since the middle is shown to be at a proven functioning state.

**[NPEO#5 specific objectives]** Providing online and in-person training and education on network performance analysis and troubleshooting.

For remote sites added to the Nautilus cluster, especially for OSG ones, we provide support for working with local network admins to bring the network up to our standards. The requirements of K8s networking bring up problems immediately upon node joining, indicating if troubleshooting is needed to establish the connection.

**[NPEO#6 specific objectives]** Engaging with CI Engineer/Advanced Cyberinfrastructure-Research Education Facilitator/Cyberteam award personnel, and other relevant groups comprising the collective CI engineering expertise in the community in which network performance plays an important role.

See Goals 4/5 activities, above.

**[NPEO#7 specific objectives]** Investigating, evaluating, and testing data movement systems and tools, transport protocols, and integrated platforms.

Custom performance dashboards show high-level views of current and historical data movement performance. TRNP uses widely-deployed perfSONAR test endpoints with esmond software, Cassandra, and custom dashboards infrastructure, as well as more general tools that can investigate, evaluate, and test data movement systems.

While network performance is a key goal, exploiting commonalities to achieve interoperability among extant infrastructures is just as critical. We look at uses that cross between OSG-controlled HTCondor/StashCache hosts and K8s-controlled TNRP hosts. We also work on the UCAR/UVA CC* Integration: Enhancement and deployment of LDM7 for scientific data distribution (award #1659174). We will be experimenting with the K8s SDN overlay network for transport of the multicast streams in Nautilus. The TNRP StashCaches are in Chicago, Kansas City, New York City, San Diego, Amsterdam, Cardiff, and Daejon.
[NPEO#8 specific objectives] Selectively investigating and solution engineering for end-to-end performance.
During weekly engineering Zoom calls involving at least 20 network engineers and cyberengineers, we selectively investigate and provide solution engineering for end-to-end performance. The notes of these meetings are taken and sent to a list of >150 signed up individuals. In addition, build specs for variants of FIONAs are updated regularly on the PRP website.

[NPEO#9 specific objectives] If appropriate, re-engineering data transfer protocols and tools.
While all software developed by TNRP will be open-source, we do not anticipate re-engineering data transfer protocols. Instead, our entire structure enables others to investigate specific tools and protocols within TNRP’s physical infrastructure by launching containers that hold their unique software configurations. Passive metrics gathered during these user-driven experiments will be provided as accessible TNRP databases. Discussions with the NRP Pilot are ongoing.

[NPEO#10 specific objectives] Analyzing network performance and data transfer metrics, including periodic analysis to demonstrate trends and assist scientific collaborations in optimizing CI resources:
Active and passive measurements are used together to iteratively narrow down the likely location of performance issues for a given distributed application through a standard bisecting “search” process. This narrows the focus to a particular network or network segment. To fully root cause a problem, more detailed information about internals of a specific network infrastructure is required. The internal physical topology of large networks is opaque to most users/applications (e.g., details of switch ports and intermediate connections). While a bisection search isolates to “there is likely some problem within this part of the network,” the in-depth view of network internals requires a trouble ticket to bring a network engineer into the loop. To support the TNRP and the NRP Pilot, we set up an NRP Federation K8s Cluster, where the regionals exchange federation credentials. It is also the home of the NRP perfSONAR MaDDash.

[NPEO#11 specific objectives] Defining a role in improving the community’s effectiveness in leveraging the significant distributed deployment of PerfSonar and associated active performance testing:
The automated creation of testing meshes and dashboards using our existing K8s perfSONAR and MaDDash containers as templates has already greatly improved regionals’ and other’s ability to quickly build active (and passive) monitoring infrastructures on their networks and at their institutions. We work with ESnet’s Andy Lake, principal developer of perfSONAR, and via Internet2’s John Hicks, with Edward Colone and Nathan Shepherd at the University of Michigan to adopt their GridFTP pscheduler plugin.

Specific objectives for Goal#2-5 are included in Goal#2-#5 activities above.

c. Describe any significant results (8000 max) (7955 now)

[NPEO #1 significant results] Identifying potential bottlenecks in data transfers for scientific collaborations.
A MaDDash has been constructed (see supplement for corresponding visualizations) that monitors GridFTP performance among the following 17 sites:
dtm1-v6.nysernet.org
ec2-3-19-82-180.us-east-2.compute.amazonaws.com
ec2-13-52-191-107.us-west-1.compute.amazonaws.com
ec2-54-200-110-185.us-west-2.compute.amazonaws.com
dtn0.lsanca.pacificwave.net
gpn-fiona-mizzou.scidmz.rnet.missouri.edu
hcc-fiona-v6.unl.edu
fiona01.tacc.utexas.edu
hcc-fiona.unl.edu
osg.chic.nrp.internet2.edu
dtn1.nysernet.org
osg.kans.nrp.internet2.edu
ec2-3-213-155-48.compute-1.amazonaws.com
ps-40g-gridftp.calit2.optiputer.net
osg.newy32aoa.nrp.internet2.edu
dtn0.uog.edu
fiona.sce.pennren.net

The Great Plains Network (GPN) has itself implemented 13 FIONAs with a MaDDash. In June, Amazon gave us a grant of $40,000 in AWS to support ongoing TNRP MaDDash testing. See the Supplement for details.

[NPEO#2 significant results] Providing analysis and objective recommendations on tools, technologies, and integrated systems related to data network movement and transfer performance.

TNRP’s propagation of PRP’s best practices has helped our partner regional and national networks to establish, test, and manage operational practices that support the dedicated science purposes of the partner best effort networks.

[NPEO#3 significant results] Serving as a centralized point of expertise, advice, and engagement for distributed scientific collaborations in identifying and solving scientific data movement inefficiencies adversely affecting time-to-science.

The PRP/TNRP website (pacificresearchplatform.org) features up-to-date hardware and networking specifications, network/server/storage procedures and configurations. Developments for network monitoring, network debugging (e.g., traceroute analysis), and analysis of network data collected so far are documented and made available online (for instance: http://pacificresearchplatform.org/userdocs/development/gitlab/).

To join the PRP RocketChat visit: https://rocket.nautilus.optiputer.net/

Weekly Zoom calls with over 20 techs and the bi-weekly science engagement calls serve as reliable shared points of expertise, advice, and engagement for distributed scientific collaborations. The primary topic is identifying and solving scientific data movement inefficiencies adversely affecting time-to-science.

[NPEO#4 significant results] Disseminating current practices across the NSF community on achieving highly efficient end-to-end data transfer performance, especially in high-bandwidth/high-delay environments.

See NPEO#4 in Section 2b above.
[NPEO#5 significant results] Providing online and in-person training and education on network performance analysis and troubleshooting.
See Section 3.

[NPEO#6 significant results] Engaging with CI Engineer/Advanced Cyberinfrastructure-Research Education Facilitator/Cyberteam award personnel, and other relevant groups comprising the collective CI engineering expertise in the community in which network performance plays an important role.
See Goals 4/5 activities, above.

[NPEO#7 significant results] Investigating, evaluating, and testing data movement systems and tools, transport protocols, and integrated platforms:

TNRP gathers metrics (active and passive), stores them in scalable, non-SQL time-series databases (in Nautilus: Elasticsearch, Prometheus, Cassandra), and then with existing tools, processes the information for alerts or other diagnostics. Collectively, these are termed TNRP databases; these will be more formalized in TNRP Years 2 and 3. We plan to experiment with graph databases Neo4j and TigerGraph.

[NPEO#8 significant results] Selectively investigating and solution engineering for end-to-end performance.

Our generalized approach of containers + instrumentation + databases + dashboards is a very productive template, but sometimes details need to be worked out in person-to-person engagement. We use active, end-to-end and disk-to-disk, monitoring and visualizations via ESnet-style MaDDashes and our own traceroute visualization to pinpoint many issues (note: the traceroute visualization software needs to be rewritten due to the much larger group of entities we are tracking, and changes in perfSONAR. We have started to do this). MaDDashes are typically quite human-intensive to construct and maintain. The test endpoints must be created, a testing plan and a measurement archive must be defined, and all this must be coordinated through person-to-person interactions. However, we now automatically create dashboards based on node labels, which allows very complex configurations without config editing, a real time saver.

[NPEO#9 significant results] If appropriate, re-engineering data transfer protocols and tools.

No significant results to date. Discussions regarding engineering data transfer protocols and tools are ongoing in the NRP Pilot/PRP technical Zoom calls; everyone prefers open source tools.

[NPEO#10 significant results] Analyzing network performance and data transfer metrics, including periodic analysis to demonstrate trends and assist scientific collaborations in optimizing CI resources.

TNRP staff have built visualizations and corresponding query services that enable the instrumentation and report a specific time-series for just the flow of interest. sFlow is a sampling technology that meets the key requirements for a network traffic monitoring solution. sFlow provides a network-wide view of usage and active routes. It is a scalable technique for measuring network traffic, collecting, storing, and analyzing traffic data (see https://sflow.org/sFlowOverview.pdf). Elastiflow helps us identify malicious actors or out-of-control software, and get the general statistics for data volumes being transferred for each project/field of science/university/etc. sFlow is running on all of the nodes and they stream their
data to the ElastiFlow/Elastic Search instance running in Nautilus (see Supplement). ElastiFlow provides network flow data collection and visualization (see https://github.com/robcowart/elastiflow).

[NPEO#11 significant results] Defining a role in improving the community’s effectiveness in leveraging the significant distributed deployment of PerfSonar and associated active performance testing.

Goal#2 significant results: Address issues critical to scaling PRP strategies
TNRP is focusing on measuring and monitoring end-to-end science data transfers. In order to do this, there have to be science engagements that result in end-to-end science data transfers that can be monitored. TNRP extends the range of CHASE-CI and PRP efforts by involving new regional networks and their campuses in two phases: first 18 months, engaging the Great Plains Network, LEARN, KINBER, and NYSERnet, and in the second 18 months involving other regionals, to be determined.

Goal#3 significant results: Build facilities and support for TNRP measurement and monitoring
Visualization tools to analyze disk-to-disk performance with network monitoring are being offered to show end-to-end performance. Internet2, critical for connecting regional networks, has provided equipment at three locations (NYC, Chicago, and Kansas City).

Goal#4 significant results: Gain participation by a federation of regional and national networks
TNRP and partners exhibit significant results:
- New MaDDashes in regionals connecting Science DMZs and in science collaborations
- Implemented science-focused containerized applications as pods/containers; We can measure everything regarding containers, but need to focus in Year 2 on what is meaningful.
- Contributed containerized services to the Open Science Grid (OSG)
Please see the supplement for visualizations, maps, spreadsheets, and other supportive documentation that are not includable in these text-only sections.

Goal#5 significant results: Identify national-scale science drivers
See Goal#5 activities above.

d. Describe key outcomes or other achievements

[NPEO#4 outcomes] Disseminating current practices across the NSF community on achieving highly efficient end-to-end data transfer performance, especially in high-bandwidth/high-delay environments.
TNRP technical staff are integrating OSG and the PRP models, with initial success. OSG sends background jobs to Nautilus, which do not interfere with real-time ML training. IceCube and LIGO currently use Nautilus (see supplement). In May 2019, IceCube used 151,861 CPU core hours and 83,502 GPU hours in Nautilus, the equivalent of running 14 FIONA8s 24x31.
OSG’s Stashcache has delivered more than 1.4 Petabytes in the first 3 months of 2019. The most positively affected were physics communities outside of the LHC.
[NPEO#7 outcomes] Investigating, evaluating, and testing data movement systems and tools, transport protocols, and integrated platforms: Extensive PRP staff time has already been spent on this topic and TNRP will build on this work.

Active metrics are used to display the performance of perfSONAR iperf3 and GridFTP transfers in MaDDashes. We can also record locally a set of passive metrics on every connection; switches/routers record interface errors, total traffic through ports, dropped packets and more. sFlow is a widely-deployed system that gives visibility into the use of networks, thereby aiding performance optimization and defense against security threats. Endpoints can provide more than 100 measures on every TCP network conversation via Tstat, a passive “sniffer” whose data can provide deep insights into the traffic patterns on both the network and the transport levels. While many organizations take advantage of passive metrics for internal monitoring, these have not yet been widely exploited for diagnosis on multi-owner networks like TNRP’s.

Elastiflow lets us see campus traffic grouped by Autonomous System Number (ASN) (see the supplement for usage).

In Year 1, TNRP worked with OSG to extend the StashCache Data Federation, using the K8s mechanisms supported by TNRP. This makes the system much more dynamic, allowing for adding additional containers for improving cache hit rates, as the StashCache functionality sees more and more adoption. Additionally, since caches can be configured to serve only parts of the total namespace in the Data Federation, a given partner of TNRP can decide to make a FIONA available for only its community’s data, or use a FIONA as an “origin” to inject its community’s data in the StashCache federation. The collaboration between OSG and TNRP allows communities across TNRP and its partners to join the StashCache Data Federation using their own resources, and setting their own policies of data and resource use.

OSG and PRP provided the software integration and user support. A first proof-of-concept was demonstrated in July 2018 as part of the IceCube experiment (see www.nsf.gov/news/news_summ.jsp?cntn_id=295955), leading us to future computational/storage cyberinfrastructure concepts to support analysis of the data flows produced by NSF Large Instruments. Specifically, an NSF CESER collaborative proposal (NSF-OAC 1841530) between LIGO (Caltech & GATech), IceCube (UW Madison), Internet2, and OSG personnel at the University of Chicago and UC San Diego was recently funded. It focuses on turning these initial prototyping ideas into a cohesive, federated, national-scale research data infrastructure for large instruments, initially for LIGO and IceCube, addressing production needs while serving as an exemplar for multi-messenger astrophysics and beyond. This is a redefinition of the role the academic Internet plays in supporting science. The collaboration will use the PRP/TPRP Kubernetes infrastructure and the NRP Pilot’s network coordination.

[NPEO#8 outcomes] Selectively investigating and solution engineering for end-to-end performance.

We have focused on building perfSONAR and MaDDash containers (instead of bare-metal hosts) and then executing these containers in Nautilus. Using K8s features and policy expressions, we run containers on specific hosts on the PRP (and next, the TNRP) to automatically enroll a perfSONAR test endpoint into a testing mesh with data reported to a specific MaDDash. The container layer itself is quite efficient in terms of I/O performance, yielding near bare-metal performance, but with the flexibility of running virtualized software environments. The container infrastructure allows us to mix periodic tests (like perfSONAR)
with on-demand experiments and application-based data transfers on the same physical hosts. Internal (passive) instrumentation through K8s cAdvisor reports performance metrics to capture usage of user-level applications. We teach others how to configure and launch containers across the regional and national networks to implement user-level experiments while capturing appropriate metrics that are placed in TNRP databases (e.g., we look for underperforming containers and work with users to optimizing them in half the cases). We also optimize scheduling by studying typical container behavior, and then optimizing resource requests.

Specific science end-to-end networking performance is often brought about by deep engagement with application communities. For example, the UCSD Center for Western Weather and Water Extremes, which studies the atmospheric rivers that episodically drench the West Coast, needed much faster access to simulation output produced by their HPC simulations running at multiple national supercomputer centers. A FIONA was configured to be a UCAR Unidata Thematic Real-time Environmental Distributed Data Services (THREDDS) data server, serving as a high-speed cache with data presented in the format needed by CW3E applications. Data was transferred from remote centers to the FIONA. We defined a THREDDS container and a FDT/GridFTP client container. We have a THREDDS Data Server container that is mounted to CephFS. When we need to move data into it, we can spin up a Globus container or a FDT container or wget just as we pull the MERRA2 data from NASA: [https://gitlab.nautilus.optiputer.net/unidata/thredds-nautilus/blob/master/copy.yaml](https://gitlab.nautilus.optiputer.net/unidata/thredds-nautilus/blob/master/copy.yaml)

Recasting the existing FIONA as a node of the PRP K8s cluster has allowed us to exploit common identity (CILogon) and K8s’ rich container instrumentation to report metrics like disk access times, TCP window negotiation, per-connection latency, container stability, and more. Recent improvements have led to a >500x decrease in wall clock time in one workflow’s analysis cycle, from 19 days to less than one hour, a key improvement on usability given it yields analysis of atmospheric rivers based on weather feeds.

Weekly joint TNRP/NRP Pilot/PRP engineering Zoom conferences, with >20 network engineers, engage a focused subset on specific problems, some of which reoccur periodically with rarely the same detailed cause. Changes in routing tables, buffer settings, memory errors, endpoint software changes, transceiver issues, slow disk arrays, and previously unknown firewalls are examples that have caused degradation in the networks, but once identified are usually readily fixed. For instance, a June 2019 NCAR fiber cut dropped the path we use to a 100Mb/1500MTU reserve path until we worked with them to get back to the 10Gbit/9000MTU path.

[Outcome10] Analyzing network performance and data transfer metrics, including periodic analysis to demonstrate trends and assist scientific collaborations in optimizing CI resources.

Elastiflow was updated this year to the recent version with new improved dashboards. It lets us see campus traffic grouped by Autonomous System Number (ASN). sFlow is running on all of the nodes and they stream their data to the ElastiFlow/Elastic Search, instance running in Nautilus. ElastiFlow provides network flow data collection and visualization (see [https://github.com/robcowart/elastiflow](https://github.com/robcowart/elastiflow) for more details on Elastiflow).
3. What opportunities for training and professional development has the project provided?

[NPEO#4 training] Disseminating current practices across the NSF community on achieving highly efficient end-to-end data transfer performance, especially in high-bandwidth/high-delay environments.

As one example, an active NRP-Pilot member at Clemson University is running deep learning oncogenomics workflows on the Nautilus (K8s) cluster, in Alex Feltus’ words: “allowing us to scale up our analyses by moving large genomics datasets between FIONA nodes and then screening tens of thousands of genes on GPUs for tumor biomarker discovery.”

As another, IceCube GPU users disseminate useful knowledge, see https://gracc.opensciencegrid.org/dashboard/db/gpu-payload-jobs-summary?orgId=1&from=now-30d&to=now&var-VOName=All&var-Project=All&var-Facility=All&var-User=All&var-ExitCode=All&var-Probe=All&var-interval=1d

Of concern is deciding what is/can be collected in Nautilus. The NRP Pilot technical community, led in this case by John Hicks (I2), has engaged the following people in the discussion: Dana Brunson and Paul Howell (I2), James Deaton (GPN), Tom DeFanti, John Graham and Dmitry Mishin (UCSD), John Hess, Sean Peisert and John Dundas III (CENIC), Jimmy Kyiannis (NYSERNet), Shannon Spurling and Jerry Heath (MORE), and Kristy Hamm (KINBER).

On 6/21/19, there was a Zoom call discussing this issue. Paul Howell discussed I2 flow analytics team/network services, which have the goal of protecting the network. His team collects anonymized Netflow from routers, telemetry, but no payload—they mask IP addresses (low order bits are removed). I2 uses no sniffers or splitters, hence no full data capture.

John Graham noted that Nautilus has sflow-->Elastiflow on all the nodes. No anonymization at this point, so the data collected is password protected. Paul suggested that we talk to the CISO at UCSD to find out how we might best inform people about Nautilus data collection. John Hicks suggested that for now, we should collect raw data, analyze it, but don't publish or share it. Tom noted that if we put analysis in NSF reports, these are public and are essentially publishing.

Dana Brunson noted that XSEDE publishes all sorts of use/user data, but not network data. James Deaton suggested we ask IU/EPOC what they do with NetSAGE. John Hicks asked what is acceptable? If one joins the NRP Pilot, is there agreement that DTN data is available? We can state that we will publish traffic stats. Tom DeFanti asked who at what campus level needs to agree to this? Is click through enough?

John Graham noted that we have Tstat data on every node in Nautilus. James Deaton said there is no scalable analysis framework available. Paul Howell said we want to build an ecosystem of use to data professionals so the data has to be published. John Graham commented that from an operational point of view, we want tools to fix crusty networks. Tstat, traceroute, Elastiflow are examples. Paul Howell said that I2 doesn't publish IPs, ASNs are deemed OK. I2 complies with GDPR regarding IP addresses. John Graham agreed that we want to comply with GDPR too.
John Graham suggested we measure NRP data transfer protocols beyond GridFTP, e.g., Aria2 (https://aria2.github.io) and FDT, and that we continue doing sFlow with a separate Elasitflow for the NRP Pilot. James Deaton said that one could opt in via login, as long as we focus only on DTNs. He said Nautilus should adopt Artemis (see https://www.caida.org/publications/papers/2018/artemis/) to help stay resilient. He said it’s not a privacy issue because it is all public data.

The meeting ended with the following suggested action items:
- Create a separate Elasticsearch for the NRP Pilot GridFTP.
- Share what we've done with sFlow setups ("in the repo already" says John Graham).
- Look into Artemis on Nautilus.
- Direct DTNs to a Nautilus collector.
- Adopt Artemis.

sFlow is a sampling technology that provides a network-wide view of usage and active routes. It is a scalable technique for measuring network traffic, collecting, storing, and analyzing traffic data. It can help us identify malicious actors or out-of-control software. It is running on all of the Nautilus nodes, and they stream their data to the ElastiFlow/Elastic Search instance running in Nautilus (see Supplement). It provides network flow data collection and visualization.

Fermilab-based experiments like DUNE and Minerva, as well as the transient science program in DES, have delivered 300 TB each of data through the jobs using the PRP mounted caching infrastructure on the network. In addition, the gravitational wave searches that OSG and PRP have enabled have consumed roughly 200 TB via the PRP-enabled OSG caching infrastructure. For the latter, we support both the analysis of private data by the LIGO-Virgo collaboration, and public LIGO data by the astrophysics community at large. As international collaborations with significant EU collaborators, LIGO and DUNE benefit from the integration of EU and US resources via OSG. TNRP supports this via a cache at the University of Amsterdam. The resulting reduction in transatlantic network traffic is roughly a factor 20 for LIGO and a factor of almost 30,000 for DUNE.

[NPEO#5 training] Providing online and in-person training and education on network performance analysis and troubleshooting:

As stated above, online documentation for training is being integrated with and available on fasterdata.es.net and posted on nationalresearchplatform.org. We share JupyterLab notebooks, and use Rocketchat channels. There are regional workshops (supported mainly by non-NPEO funds) targeted to IT/network engineers from the NSF community to train them on high-performance network analysis and troubleshooting, like the GPN workshop in May 2019 and the NYSERNet one planned for the last week of September, 2019. We are active at Quilt meetings in sessions addressing NRP issues and participation. In addition, TNRP is discussing with EPOC at Indiana University ways share and complement NPEO#5 efforts.

The Global Research Platform (GRP) is now formally organizing. A Global Research Platform workshop is set for 9/17-18 at UCSD. Active networking national and regional groups in the PRP/GRP efforts include CANARIE, CESnet, StarLight, SURFnet, NORDUnet, and APRP (Asia Pacific Research Platform).
In Year 1, TNRP researchers and staff gave invited talks at the meetings of CENIC, Internet2, GPN, NEREN, and others. We have also organized and delivered a Kubernetes hands-on workshop at the GPN Annual meeting in May 2019, and had a workshop accepted at SC19, to be delivered in November 2019. Our weekly engineering Zoom calls have >20 participants, and the notes go out to >150 subscribers. Our bi-weekly science engagement Zoom calls are also a popular engagement venue for research education facilitators and Cyberteam award personnel: Akbar Kara; Alex Feltus, Ana Hunsinger, Camille, Cathy Chaplin, Chris Hoffman, Dana Brunson, Dima Mishin, Eli Dart, Jack Shnell, James Deaton, Jason Zurawski; Jeff Weekley, Jennifer Griffin, Jen Leasure, Jen Oxenford, Jen Schopf, Jim Bottum, Jimmy Kyriannis, John Graham, John Hicks, Kate Adams, Kate Mace, Ken Lutz, Larry Smarr, Lauren Rotman, Lee Ann Weber, Mark Yashar, Meggan Levitt, Melissa Lucas, Pankaj Shah, Shawfeng Dong, Tim Boerner, Tom DeFanti, Tracy Smith, and Wendy Huntoon.

TNRP Manager Tom DeFanti continues to work with Florida A&M University, a public, Historically Black University in Tallahassee, Florida to coordinate training opportunities and participation. FAMU is currently assessing its cyberinfrastructure and working toward creating a Southeast Regional Network.

4. How have the results been disseminated to communities of interest?

TNRP leadership promotes the project and shares information and early results with communities of interest through its active participation in relevant conferences and meetings. For example, the bi-weekly science engagement meeting keeps its running notes at https://docs.google.com/document/d/1JfQO_odoUyYaNfCzRWz_cvrN2gtQj9ovbq9GJUkZyG0/edit

During the most recent project periods, PI Smarr and the project’s Co-PIs gave invited talks on the TNRP to a diverse set of audiences at several meetings, including the Bright Internet Global Summit, the Internet2 Global Summit, and the HEPiX Spring 2019 Workshop.

In particular, we continue to have a strong presence at the CENIC annual conference, which typically attracts close to 400 attendees. One panel focused on the NRP: The National Research Platform: An Update on Progress Toward Scaling PRP. This panel, led by TNRP/PRP leaders, described the efforts of TNRP/PRP partners and participants to scale the framework nationally and internationally. It also provided an update on presentations and discussions from the Second National Research Platform Workshop, held in Bozeman, Montana, in August 2018.

TNRP had a strong presence at the GPN Annual Meeting 2019, with Frank Würthwein delivering an invited talk, and Igor Sfiligoi conducting a Kubernetes workshop. A Kubernetes workshop led by Igor has been accepted for the upcoming SC19 conference in November.

In addition to publicizing TNRP information and data results at conferences, we maintain a
public website: http://nationalresearchplatform.org . As the NRP Pilot and TNRP progress, the website is updated with relevant information, news, special events, and presentations. For example, video links are available to all the presentations given during the first and second NRP workshops so interested parties can view the talks and download related PowerPoint slides: http://pacificresearchplatform.org/reports/.

Build specifications for all FIONAs are available in tabs at pacificresearchplatform.org. This is acknowledged as a very valuable community resource since these are not off-the-shelf items: http://pacificresearchplatform.org/admindocs/fiona/builds/ .


The Great Plains Network (GPN) has implemented 13 FIONAs. GPN Engineer George Robb III (as transcribed by Tom DeFanti during the joint PRP/NRP Pilot tech call on 4/4/19) said: “We are on rev3. Our environment is good for containerization, layering Kubernetes on top. We are putting out production-quality FIONA nodes to campuses. Local admins can stand up a virtualization. If they colo, they can share resources with Kubernetes. They can add other kinds of processors/processes. Our 3rd prototype is stable. We expect the Dakotas and Nebraska to join soon. If we want to bring a national center to work on a problem, we can use Singularity to bring 20,000 cores to bear. We love the federation. We could use 6000 cores for an hour, for example, federated. How to get the storage to be friction free is something to attack. GPN moving slow but surely. 13 FIONAs are deployed. Institutions are getting work done. We have EPSCoR states and need to get more regional network teams involved--so far, it’s been really great.”

Numerous talks have been given at national and international venues (e.g., OSG All Hands Meeting and WLCG Grid Deployment Board Meeting) on this project. Users include scientists from UC Santa Barbara, UC San Diego, Columbia University, and elsewhere. Fields of science include biomedical informatics and experimental particle physics, among others.

TNRP and OSG engaged Internet2 in the NRP Pilot project to cleanly delineate the scope of support for Big Data science, from the sinks and sources of the data, through the local, regional, and national network providers, to cyberinfrastructure teams, thence to the actual science teams. Internet2, in collaboration with the Great Plains Network, deployed FIONA storage hardware at three locations in the network backbone. That hardware was integrated into the Nautilus cluster. OSG uses that cluster to deploy its federated caching infrastructure (StashCache). The science user community of OSG is thus provided with additional caching at strategic locations in the backbone. The PRP functioned as enabler by showing OSG personnel in Würthwein’s group how to use K8s as a service deployment vehicle on the FIONAs in the network.

NYSERNet is very active in the TNRP/NRP Pilot. Jim Kyriannis on 4/1/19 released the following e-memo:

*In this next phase of our participation in the National Research Platform (NRP) Pilot, we’re inviting member institutions with well-suited research projects to participate. This year, we can*
accommodate up to six participants in total: three from large institutions and three from mid-sized/smaller institutions.

Please find some additional background information below. If you are interested in learning more about Pilot participation, please let me know. Our target date for finalizing our list of participants is end of April.

NYSERNet is participating in a pilot to grow a distributed compute environment built by University of California researchers, called The Pacific Research Platform (PRP - https://ucsd-prp.gitlab.io), into a national testbed: The National Research Platform. The technology consists of Data Transfer Nodes (DTNs), Compute Nodes, and a distributed Ceph storage environment acting in concert to enable inter-institutional high-speed research data transfers on a federated computing platform offering Intel and GPU processing capacity. Not designed to compete with multi-million core Supercomputers, this is an environment based on federated access into Kubernetes Linux containers providing a temporary virtual compute and storage environment, suitable for smaller research workloads.

Thus far, NYSERNet has deployed the DTN component of the NRP Pilot in both our Syracuse Data Center and at Stony Brook University. We're collaborating with Stony Brook on implementing a Science DMZ network and supporting research needs through the Pilot, such as improving data transfers to XSEDE via the DTN. We'll soon be exploring NRP compute possibilities as well. We've also begun collaborating with the American Museum of Natural History on an NSF-funded project to design a Science DMZ which supports NRP Pilot activities, including facilitating an Astrophysicist's data transfers with the Palomar Observatory and providing a compute platform.

We would like to invite additional institutions and their researchers to participate in the NRP Pilot as well. Due to the unique nature of its technology, not all research projects are well-suited for the Pilot. Those which are a good match for the Pilot will have need for collaborations and data transfers which rely upon the NYSERNet network to reach Internet2 and can leverage a virtual Linux compute environment. Projects requiring GPU support or discipline-specific software are welcome.

Brad Smith at UCSD asked for our help, which we are pleased to provide:

An Astrophysics faculty member (Brant Robertson) is hosting the Kavli Summer Program In Astrophysics starting July 8th on campus. He was awarded an NSF grant (~$1.5M) for a cluster that is being delivered soon (appears now to be July 8th) that he was hoping to have the participants use for projects they do as part of the workshop. The workshop is focused on machine learning (and his cluster is heavy in machine learning). He's worried, with the cluster showing up the day the workshop starts, that it won't be ready for early project work, and is interested in checking out the PRP ML nodes. Does it seem workable to have this workshop use the PRP ML nodes for early work on their project, and migrate to Brant's cluster (to be called lux.ucsc.edu) when it is ready?

Brad
5. What do you plan to do during the next reporting period to accomplish the goals?

We plan, of course, to continue working on the 5 Goals (including the 11 NPEO items) during the next reporting period.

We also plan to follow up on the results of The 3rd National Research Platform Workshop (September 24-25, Minneapolis), and participate in SC’19 in Denver in November. We are gaining experience every day with sharing the installed base of FIONA8s, FIONA4s, and FIONA2s GPUs, CPUs, storage, and newly added FPGAs. We have several Xilinx U200 FPGA add-in boards for a planned nationwide 100G SDN experiment for the GRP workshop and SC’19 with SDSU, Northwestern/StarLight, Ciena, and others. In addition, we are experimenting with Nvidia Jetson Xaviers and Google Edge TPUs, which may prove very useful for machine learning in low-power installations, and provide interesting challenges to network monitoring and measurement.

Another expansion of the PRP/TNRP, The Global Research Platform (GRP), is now organizing among the Global Lambda Integrated Facility (GLIF) membership. There will be a Global Research Platform workshop at UC San Diego September 17-18, 2019, with expected GLIF member participation and a GNA/GLIF meeting following on September 19-20.

We plan to extensively test the cloud (AWS, Google, Azure) data transfer capability against all the NRP-Pilot campuses and each other, maintaining a GridFTP MaDDash to show performance. In June 2019, we received a grant of $40,000 worth of time from AWS to pursue this testing (see Supplement for initial MaDDash results), and we will pursue similar grants from Google and Microsoft.

The NRP Pilot project is active and engaging a team of members from the PRP as well as four regional networks this year (GPN in the Great Plains, LEARN in Texas, KINBER and NYSERNet), and Internet2. Each of these regional networks is now working with selected campuses in their regions. The goal is to extend the infrastructure in a scalable way, taking into consideration the operational challenges inherent in increasing the number of intermediate players and end site institutions. The Pilot is continuing to deliver results. As with the PRP, the major thrust of the effort is to provide value to the research community.

We will continue to integrate our efforts with the Open Science Grid in Year 2.

One possible outcome of workshops and community experience with actual volumes of data is that multiple regional databases exist as anticipated in this proposal, but data is replicated to a nationally-aggregated version. Retention policies for this kind of “national” data will be developed. In all of this, TNRP will periodically analyze these datasets and publish the resulting trends, assisting all of our scientific collaborations in optimizing TNRP resources.

Taking advantage of tools like Kapacitor (an alerting platform) with InfluxDB (a time series database), one can craft automated alerts that are sent to cognizant network engineers when there are changes that should be perceived as indications of end-to-end
and/or network performance issues. We believe that TNRP’s database approach can readily support neural networking/machine learning approaches to make better inferences about network state and this will be pursued in Years 2 and 3. Rob Gardner (LHC, U Chicago) and John Graham showed a working prototype at SC17, while ESnet’s Mariam Kiran at the 2017 Internet2 Technology Exchange suggested this kind of approach to perceive anomalies and identify best paths in real time.

We need to look at alternatives to GridFTP because Globus is deprecating support for it. We will do this in conjunction with OSG and Internet2. FDT and Aria2 are being considered.

The NRP Pilot community strongly suggests Nautilus adopt Artemis (see https://www.caida.org/publications/papers/2018/artemis/) using a simple lightweight instance on each campus. Documentation to streamline new technical staff on-boarding is important. Artemis will help Nautilus stay resilient; it’s not a privacy concern—all the data collected is data.

The security of TNRP, as a whole, is quite complex. As best we can, all physical systems – K8s cluster nodes, TNRP Databases, Rook/Ceph storage, web-servers, and others – are regularly maintained with security updates, log processing, and other routine practices. In addition, we will need to establish the policy for running user-defined containers and the mechanisms for validating that data stored in TNRP databases has not been corrupted. There is much activity in academia and industry in this area due to the overwhelming acceptance of Kubernetes (over 50% of the Fortune 500 companies, for example).

Further, given the large volume of data transfers that cross the network as a whole, we need to analyze the interfaces (people and application), policies, and procedures when working with existing campuses and network intrusion-detection systems (such as those based upon the Bro network security monitor). Taken together, these policies and processes must be written down and be reviewable by external entities for transparency. We will develop TNRP security practices and policies as an open source document.

In order to explore ways to increase security on the PRP FIONAs, UCSD’s John Graham worked with CoreOS developers to provide a Distributed Trusted Computing environment that can scale from lab to international deployments. Trusted software distributions can be securely deployed and verified, allowing the underlying hardware to remain in a secure stable state. These application pods running any number of operating systems can be launched on any Kubernetes minion in the cluster that the user has privileges on. Within our infrastructure, we use CILogon for federated identity management for user authentication. We also plan to exploit physical Trusted Platform Modules (TPMs) in K8S physical nodes in combination with BIOS-based secure boot to ensure that software running on K8S nodes is unchanged. (TPMs are standard hardware components that are widely available, but not widely utilized in HPC/academic environments). CoreOS retired its work on the TPM admission controller for Kubernetes, but made it available for future upstreaming. We intend to bring this code forward to the current release of Kubernetes.

TNRP databases will be regularly validated so that “past data” has not been changed. But we see these databases as yet another source for security monitoring. For example, it is not unreasonable to use learning techniques to observe “usual” conversations among endpoints over TNRP and then alert when an unusual (and high-volume) transfer shows up. We believe that making these
databases open access will encourage others outside of this funding to process this data in interesting and novel ways.

Some additional Year 2 TNRP goals being considered include:

- Harvesting application usage patterns on PRP/OSG
- Add testing with network protocols beyond GridFTP
- Federating commercial & NSF clouds into the TNRP
- Continuing migration toward IPv6 while maintaining IPv4
- Adopting NextGen software-defined networking/storage Tools
- Making security more robust in the PRP/NRP/GRP
- Using TNRP as a national-scale “honeypot” to collect data for ML analysis of malicious network attacks.
- Continuing to address Goals#1-5.

PRODUCTS

1. Publications (journal articles, book chapters, conference papers, etc.)


28. Paolini C, Soselia D, Baweja H, Sarkar M. Optimal Location for Fall Detection Edge Inferencing (in review). IEEE Global Communications Conference (GLOBECOM); December 9-13, 2019; Waikoloa, HI.


31. Srinjoy Das and Dimitris Politis, Nonparametric estimation of the conditional distribution at regression boundary points, Accepted to The American Statistician (2018).


46. S. Gupta, M. Imani, T. Rosing “Exploring Processing In-Memory for Different Technologies”, ACM Great lakes symposium on VLSI (GLSVLSI), 2019 (Acceptance rate 29%).

47. J. Sim, S. Gupta, M. Imani, Y. Kim, T. Rosing “UPIM : Unipolar Switching Logic for High Density Processing-in-Memory Applications”, ACM Great lakes symposium on VLSI (GLSVLSI), 2019 (Acceptance rate 29%).


2. Technologies or techniques
We have developed the Nautilus hyper-converged cluster spread over 17 campuses for researcher access and provided means to join the cluster with researcher-provided hardware. We have also implemented CIlogon for federated access. We are helping Internet2 and regional partners join Nautilus and create their own Kubernetes hyper-converged clusters for measurement, monitoring, and moving science data.

We have built and documented specific workflows for projects like WIFIRE and CONNECT.

We have extensively prototyped running Open Science Grid High-Throughput Computing (HTC) jobs on Nautilus GPUs, CPUs, and storage as a way of testing background job running and stress-testing Nautilus/Kubernetes/Ceph.

We have created a means to migrate jobs to Google (for TPU access provided under a grant from Google) and AWS (users provide their own paid or granted accounts; a 40,000-unit ($40,000) grant was received in June 2019).
3. Inventions, patents, or licenses
Nothing to report.

4. Websites
PacificResearchPlatform.org is the project’s main website. Our Grafana visualizations can be found at grafana.nautilus.optiputer.net. Our Rocketchat is heavily used by hundreds of users of the PRP and TNRP: rocket.nautilus.optiputer.net

Five active gitlab repositories were built in conjunction with the CHASE-CI and PRP efforts:

https://gitlab.nautilus.optiputer.net/connect/segment-code
https://gitlab.nautilus.optiputer.net/connect/segment-tpu
https://gitlab.nautilus.optiputer.net/jtatar/matlab-docker-image
https://gitlab.nautilus.optiputer.net/eshearer/arc
https://gitlab.nautilus.optiputer.net/m.mcdermott/michael_gpu-pod_test

5. Other products, such as databases, audio or video products, software, etc.
None.

PARTICIPANTS

1. What individuals have worked on the project?
Provide the following information for PI/Co-PIs and each person who has worked at least one person month per year on the project during the reporting period, regardless of the source of compensation:

Name: Larry Smarr
Email address: lsmarr@ucsd.edu
Project role: PI
Nearest person month worked: 2
Contribution to the project: Director
Funding support: UCSD Qualcomm Institute

Name: Tom DeFanti
Email address: tdefanti@eng.ucsd.edu
Project role: Project Manager
Nearest person month worked: 3
Contribution to the project: Project Manager
Funding support: UCSD Qualcomm Institute; NSF
<table>
<thead>
<tr>
<th>Name</th>
<th>Email address</th>
<th>Project role</th>
<th>Nearest person month worked</th>
<th>Contribution to the project</th>
<th>Funding support</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Hess</td>
<td><a href="mailto:jhess@cenic.org">jhess@cenic.org</a></td>
<td>Network Engineer</td>
<td>1</td>
<td>CENIC Network Engineering</td>
<td>CENIC</td>
</tr>
<tr>
<td>John Graham</td>
<td><a href="mailto:jgraham@eng.ucsd.edu">jgraham@eng.ucsd.edu</a></td>
<td>Developer</td>
<td>2</td>
<td>Technology implementation</td>
<td>UCSD Qualcomm Institute; NSF</td>
</tr>
<tr>
<td>Molly Wofford</td>
<td><a href="mailto:mwofford@eng.ucsd.edu">mwofford@eng.ucsd.edu</a></td>
<td>Report writer</td>
<td>1</td>
<td>Support in report preparations</td>
<td>UCSD Qualcomm Institute; NSF</td>
</tr>
<tr>
<td>Dmitry Mishin</td>
<td><a href="mailto:dmishin@ucsd.edu">dmishin@ucsd.edu</a></td>
<td>Science Engagement Program Analyst</td>
<td>4</td>
<td>SDSC liaison and programmer</td>
<td>San Diego Supercomputer Center</td>
</tr>
<tr>
<td>Igor Sfiligoi</td>
<td><a href="mailto:isfiligoi@ucsd.edu">isfiligoi@ucsd.edu</a></td>
<td>Data Science Research Specialist</td>
<td>4</td>
<td>OSG liaison and network programmer</td>
<td>San Diego Supercomputer Center/Physics Dept</td>
</tr>
<tr>
<td>Joel Polizzi</td>
<td><a href="mailto:jpolizzi@eng.ucsd.edu">jpolizzi@eng.ucsd.edu</a></td>
<td>Cluster Engineer</td>
<td>3</td>
<td>Cluster implementation</td>
<td>UCSD Qualcomm Institute</td>
</tr>
<tr>
<td>Shava Smallen</td>
<td><a href="mailto:ssmallen@ucsd.edu">ssmallen@ucsd.edu</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Project role: Programmer
Nearest person month worked: 1
Contribution to the project: Cluster user support
Funding support: UCSD Qualcomm Institute

Name: Nadya Williams
Email address: nwilliams@ucsd.edu
Project role: Programmer
Nearest person month worked: 4
Contribution to the project: Cluster software
Funding support: UCSD Qualcomm Institute

Name: Charles Erwin
Email address: cerwin@ucsd.edu
Project role: Website developer
Nearest person month worked: 1
Contribution to the project: Website development
Funding support: UCSD Qualcomm Institute

2. What other organizations have been involved as partners?

Type of organization: Academic Institution
Name: UC Berkeley
Location: Berkeley, California
Contribution to project: Collaborative Research

Type of organization: Academic Institution
Name: UC Davis
Location: Davis, California
Contribution to project: Collaborative research

Type of organization: Academic Institution
Name: UC Irvine
Location: Irvine, California
Contribution to project: Collaborative research

Type of organization: Academic Institution
Name: UC Los Angeles
Location: Los Angeles, California
Contribution to project: Collaborative research

Type of organization: Academic Institution
Name: UC Merced
Location: Merced, California
Contribution to project: Collaborative research

Type of organization: Academic Institution
Name: UC Riverside
Location: Riverside, California
Contribution to project: Collaborative research

Type of organization: Academic Institution
Name: UC San Francisco
Location: San Francisco, California
Contribution to project: Collaborative research

Type of organization: Academic Institution
Name: UC Santa Barbara
Location: Santa Barbara, California
Contribution to project: Collaborative research

Type of organization: Academic Institution
Name: UC Santa Cruz
Location: Santa Cruz, California
Contribution to project: Collaborative research

Type of organization: Academic Institution
Name: Stanford University
Location: Stanford, California
Contribution to project: Research collaboration

Type of organization: Academic Institution
Name: California Institute of Technology
Location: Pasadena, California
Contribution to project: Collaborative research

Type of organization: Academic Institution
Name: University of Southern California
Location: Los Angeles, California
Contribution to project: Collaborative research

Type of organization: Academic Institution
Name: San Diego State University
Location: San Diego, California
Contribution to project: Collaborative research

Type of organization: Academic Institution
Name: University of Washington at Seattle
Location: Seattle, Washington
Contribution to project: Collaborative research

Type of organization: Academic Institution
Name: University of Illinois at Chicago
Location: Chicago, Illinois
Contribution to project: Collaborative research

Type of organization: Academic Institution
Name: University of New Mexico
Location: Albuquerque, New Mexico
Contribution to project: Collaborative research

Type of organization: National Laboratory
Name: NASA Ames/NASA Research and Engineering Network (NREN)
Location: Moffett Field, California
Contribution to project: Collaborative research
More detail on partner and contribution (optional):

Type of organization: National Laboratory
Name: Lawrence Berkeley National Laboratory
Location: Berkeley, California
Contribution to project: Collaborative research

Type of organization: Industry
Name: CENIC
Location: La Mirada, California
Contribution to project: Collaborative research

Type of organization: Industry
Name: Energy Sciences Network (ESnet)
Location: Berkeley, California
Contribution to project: Collaborative research

Type of organization: Academic Institution
Name: University of Utah
Location: Salt Lake City, Utah
Contribution to project: Collaborative research

Type of organization: Academic Institution
Name: Indiana University
Location: Bloomington, IN
Contribution to project: Collaborative research

Type of organization: National Laboratory
Name: National Supercomputing Centre
Location: Singapore
Contribution to project: Collaborative research

Type of organization: Academic Institution
Name: University of Amsterdam
Location: Amsterdam, Netherlands
Contribution to project: Collaborative research

Type of organization: Academic Institution
Name: Tokyo University
Location: Tokyo, Japan
Contribution to project: Collaborative research

Type of organization: Research Institute
Name: Korea Institute of Science and Technology Information (KISTI)
Location: Daejeon, South Korea
Contribution to project: Collaborative research

Type of organization: Industry
Name: Australia’s Academic and Research Network (AARNet)
Location: Australia
Contribution to project: Collaborative research

Type of organization: Academic Institution
Name: University of Hawaii at Manoa
Location: Manoa, HI
Contribution to project: Collaborative research

Type of organization: Academic and Research Institution
Name: UCSD Scripps Institute of Oceanography
Location: San Diego, CA
Contribution to project: Collaborative research

Type of organization: Academic Institution
Name: Montana State University
Location: Bozeman, MT
Contribution to project: Collaborative research

More detail on partner and contribution (optional):

3. What other collaborators or contacts have been involved?

Avi Yagil, UCSD (Particle Physics)
Edgar M Fajardo Hernandez UCSD (Particle Physics)
Andrew Lankford, UCI (Particle Physics)
Anyes Tafford, UCI (Particle Physics)
Daniel Whiteson, UCI (Particle Physics)
Harvey Newman, Caltech (Particle Physics)
Maxwell Chertok, UCD (Particle Physics)
John Conway, UCD (Particle Physics)
Robin Erbacher, UCD (Particle Physics)
Michael Mulhearn, UCD (Particle Physics)
Mani Tripathi, UCD (Particle Physics)
Claudio Campagnari, UCSB (Particle Physics)
Robert Clare, UCR (Particle Physics)
Owen Long, UCR (Particle Physics)
Stephen Wimpenny, UCR (Particle Physics)
Peter Nugent, LBNL (Astronomy and Astrophysics)
Tony Tyson, UCD (Astronomy and Astrophysics)
James Bullock, UCI (Astronomy and Astrophysics)
Joel Primack, UCSC (Astronomy and Astrophysics)
Piero Madau, UCSC (Astronomy and Astrophysics)
David Reitze, Caltech (Astronomy and Astrophysics)
David Haussler, UCSC (Biomedical Data)
Brad Smith, UCSC (Biomedical Data)
Rob Knight, UCSD (Biomedical Data)
Marla Meehl, NCAR/UCAR (Earth Sciences)
Dan Cayan, UCSD/SIO (Earth Sciences)
Christopher Paolini, SDSU (Computer Science)
Falko Kuester, UCSD (Scalable Visualization, VR, and UHR Video)
Tom Levy, UCSD (Scalable Visualization, VR, and UHR Video)
Jurgen Schulze, UCSD (Scalable Visualization, VR, and UHR Video)
Maxine Brown, UIC (Scalable Visualization, VR, and UHR Video)
Jason Leigh, UHM (Scalable Visualization, VR, and UHR Video)
Nicola Lecari, UCM (Scalable Visualization, VR, and UHR Video)
Cees de Laat, UvA (Scalable Visualization, VR, and UHR Video)
Frank Vernon, UCSD SIO (HPRWEN)
Greg Hidley, UCSD and CENIC (HPWREN)
Scott Sellars, UCSD SIO and NSF (CONNECT)
Marty Ralph, UCSD SIO (CONNECT)
Hong Zhou, UCLA (Cryo-EM)
Peng Ge, UCLA (Cryo-EM)
Matthew Settles, UCD (Cryo-EM)
Michael Cianfrocco, University of Michigan (Cryo-EM)
Machine Learning:
   Ramakrishna Akella, UCSC
   Charless Fowlkes, UCI
   M. Chandraker, UCSD
   A. Roy-Chowdhury, UCR
   Gary Cottrell, UCSD
   Nikil Dutt, UCI
   Rajesh Gupta, UCSD
TNRP Annual Report Year 1

Ken Kreutz-Delgado, UCSD  
Jeffrey Krichmar, UCI  
Falko Kuester, UCSD  
Arun Kumar, UCSD  
Anshul Kundaje, Stanford  
Walid Najjar, UCR  
Ravi Ramamoorthi, UCSD  
Tajana Rosing, UCSD  
Jurgen Schulze, UCSD  
Padhraic Smyth, UCI  
Nuno Vasconcelos, UCSD

IMPACT

1. What is the impact on the development of the principal discipline(s) of the project?

There has been significant impact on the tested performance of intercampus data transfer over the CENIC, Pacific Wave, Internet2, and international 100Gbps networks. Use of PRP’s distributed storage and compute resources is increasingly accepted and replicated.

2. What is the impact on other disciplines?

The primary impact at this time is the greatly increased performance of networked data transfers, remote mounting of data, and the distribution of visualization-based debugging technology to bring up new network circuits, and compute/storage systems and monitor their performance. In addition, lab-to-lab groups and regional networks participants are forming and will have this same performance measurement technology available to assure the health of their connections over campus DMZs and the wide area. The Nautilus distributed hyper-converged cluster is now a model for use and replication on various Research Platforms around the nation and globally. The TNRP has also greatly improved the quality of service for the OSG StashCache project, by allowing for both doubling the number of caching nodes, and by placing those in the high-performance network backbone. Furthermore, the additional monitoring infrastructure has helped keep the achieved performance high at all times.

3. What is the impact on the development of human resources?

Science engagement efforts have included student participation in many cases. Engaging directly with network monitoring and application development tools has provided an opportunity for students to develop new skills in network engineering and research facilitation. Students conduct monitoring tests, debug connections, and write up documentation of their efforts to benefit team members coming after them on the project.
The success of the JupyterHub FIONA configuration at UC Berkley and the popularity of the UCSD JupyterHub FIONA inspired the idea of replicating this capability for routine student use, which will improve access to classroom materials and instructional assignments in the sciences.

TPRP leadership also met with CENIC, Internet2, and ACI-REF representatives and plans were made to share and distribute the PRP’s FIONA technology and encourage broader science engagement among minority serving institutions (MSIs) and institutions in the Experimental Program to Stimulate Competitive Research (EPSCoR) states. Project Manager Tom DeFanti is working with Florida A&M University, a public, historically black university in Tallahassee, Florida, to coordinate training opportunities and participation in the PRP. FAMU is currently assessing its cyberinfrastructure and working toward creating a Southeast Regional Network. Outreach plans include a visit by FAMU’s CIO and networking team to UC San Diego for training. The American Museum of Natural History, a major research facility in NYC will send a team to UCSD to study TNRP techniques so they can work more closely with NYSERNet. More than half the TNRP’s partner institutions are MSIs or located in EPSCoR states.

4. What is the impact on physical resources that form infrastructure?

FIONA DTNs and SDX development equipment has been installed and updated to the latest Kubernetes versions at partner universities. CPU, GPU, IOT, and FPGA computing resources have been added, all as part of Nautilus, a 14-Campus hyper-converged cluster. Nautilus has:
- 3300 CPU cores on 122 hosts connected at 10Gbps or faster
- 1.7 PB Storage connected at 40Gbps or faster
- >380 GPUs offering 30 million GPU fp32 core hours per day.

Nautilus has also inspired campuses to clone similar GPU nodes for instructional lab use (mainly for machine learning, which is not yet served cost-effectively by commercial cloud providers). 1Gbps FIONettes (about $270 each) were built and distributed to show campus network engineers how to monitor and measure data flows.

FIONAs (to PRP/TNRP specs) or DTNs (to ESnet specs) are being replicated at many universities and regional network hubs, achieving 10–1000x speedups as measured disk-to-disk with GridFTP. The NRP Pilot project uses FIONAs/DTNs at UCSD, TACC, UNL, NYSERNet, UMissouri, Pacific Wave (LAX), Internet2 (Kansas City, New York City, and Chicago), and UGuam for daily tests. Five 100Gbps FIONAs are deployed at SDSU, UCSD, UCSC, Caltech and StarLight (in Chicago) for very high-speed networking and data transfer tests.

5. What is the impact on institutional resources that form infrastructure?

The UC IT Leadership Council (a group of all the UC CIOs) is actively supportive of the PRP and stays informed by TNRP Project Manager Tom DeFanti who attends the ITLC meetings, and PI Larry Smarr who presents updates yearly. In particular, the support of campus Science DMZs is crucial and yet difficult, due to the demand for higher security as attacks become more severe and frequent. The Science DMZs need to be treated specially and carefully, because often the campus firewall technology available prevents large-scale transfers at full speed.
In addition to the larger CENIC group of attached universities, and those attached via the Pacific Northwest Gigapop and StarLight, The Great Plains Network, NYSERNet, LEARN, KINBER, The Front Range Gigapop, and Internet2 are adopting ways to support Science DMZs as “outside” of internal campus networks. Similarly, the Asia Pacific Research Platform (APRP) and a European Research Platform were started to complement the PRP/TNRP. This is an extremely important outcome of the TNRP—it has become a model for deployable and testable regional, national, and global science DMZs. Annual National Research Platform Workshops bring together various stakeholders to discuss strategies for building out the TNRP. In addition, ancillary workshops are held to “train the trainers” at partner universities and regional network meetings, such as CENIC’s annual conference. See pacificresearchplatform.org for the latest details.

6. What is the impact on informational resource that form infrastructure?

FIONA data transfer nodes are being built with up to 200GB of storage for various disciplines (microscopy, archaeology, machine learning, etc.). They are also being installed in the DMZs of cooperating universities so that data can be made available to colleagues at very low cost. The TNRP is also working with various supercomputer centers (SDSC, NERSC, NCAR, TACC, NCSA, KISTI in Korea, NSCC in Singapore, and AIST in Japan) to provide fast access to their data by US scientists and others. A large project with the UCSD Library complements the PRP in the area of data preservation.

7. What is the impact on technology transfer?

Companies, government, and non-profit entities are adopting the DMZ/DTN concept, significantly due to the Department of Energy’s excellent tech transfer efforts and the PRP/TNRP’s various campus/PoP implementations. We are also popularizing Kubernetes as a working model for cluster resource sharing for specific domain sciences, including data science.

8. What is the impact on society beyond science and technology?

We expect to have significant impact on society and science and technology, evidenced by the quick uptake and cloning of our technologies and science engagement leadership. The fact that TNRP approaches full network speed is quite dramatic—high-speed data transfers until recently usually required Layer2 networks and a lot of engineering resources. TNRP is proving that Layer3 networks – which are dramatically easier to setup, expand, and maintain than Layer 2 vlans – actually work at just about full line speed between campuses with tuned DTNs. Kubernetes plug-ins will soon allow SDN when and where available, a long-term goal of our networking research and development partners worldwide.

Science engagement efforts are also expanding the professional field of research facilitation for data-intensive science. Including student teams in developing and testing new methods for data
transfer, monitoring, and visualization is helping to develop a cyber-workforce that will be necessary to meet technical opportunities and challenges in the future.

**CHANGES/PROBLEMS**

1. **Changes in approach and reasons for change.**

The ending of support for the Globus Toolkit by the Globus alliance is pushing several communities, including many in our partner OSG, to look for alternatives to GridFTP. We thus are actively investigating the addition of alternative protocols as part of our end-to-end network monitoring and testing suites.

Science driver projects are moving forward as planned with results generally as expected. New science driver projects continue to be added, significant to the scientists involved, but not otherwise affecting the measurement and monitoring objectives and scope of the overall TNRP project.

2. **Actual or anticipated problems or delays and actions or plans to resolve them.**

None.

3. **Changes that have significant impact on expenditures.**

None.

4. **Significant changes in use or care of human subjects.**

N/A

5. **Significant changes in use or care of vertebrate animals.**

N/A

6. **Significant changes in use or care of biohazards.**

N/A
APPENDIX A: The TNRP Networks and Visualization Tools

Figure 1a: PRP evolution Towards a National Research Platform (TNRP). TNRP uses Pacific Wave and Internet2 to connect the Quilt regional R&E networks shown. The background map is OSG’s and indicates significant overlap, which is advantageous as we build out the TNRP with the regionals, Internet2 and ESnet.

Figure 1b: The Nautilus hyper-converged cluster also connects three regionals and U Hawaii plus
three Internet2 NRP Pilot sites (Kansas City, Chicago, and New York City).

**Figure 1c**: Great Plains Network MaDDash, June 2019. From James Deaton, Executive Director

**Figure 1d**: Great Plains Network perfSONAR Dashboard, June 2019. From James Deaton, Executive Director
Table 1 below shows the Nautilus nodes and their funding source, GPU-type and GPU count. 370 GPUs are on-line as of June 28, 2019.

<table>
<thead>
<tr>
<th>GRANT</th>
<th>NAME</th>
<th>GPU-TYPE</th>
<th>370</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHASE-CI</td>
<td>fiona8-0.calit2.uci.edu</td>
<td>1080</td>
<td>8</td>
</tr>
<tr>
<td>CHASE-CI</td>
<td>fiona8-1.calit2.uci.edu</td>
<td>1080</td>
<td>8</td>
</tr>
<tr>
<td>CHASE-CI</td>
<td>fiona8-2.calit2.uci.edu</td>
<td>1080</td>
<td>8</td>
</tr>
<tr>
<td>CHASE-CI</td>
<td>k8s-chase-ci-01.calit2.optiputer.net</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>CHASE-CI</td>
<td>k8s-chase-ci-01.noc.ucsb.edu</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>CHASE-CI</td>
<td>k8s-chase-ci-02.calit2.optiputer.net</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>CHASE-CI</td>
<td>k8s-chase-ci-03.calit2.optiputer.net</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>CHASE-CI</td>
<td>k8s-chase-ci-04.calit2.optiputer.net</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>CHASE-CI</td>
<td>k8s-chase-ci-05.calit2.optiputer.net</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>CHASE-CI</td>
<td>k8s-chase-ci-06.calit2.optiputer.net</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>CHASE-CI</td>
<td>k8s-gpu-1.ucr.edu</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>CHASE-CI</td>
<td>k8s-gpu-1.ucsc.edu</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>CHASE-CI</td>
<td>k8s-gpu-2.ucsc.edu</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>CHASE-CI</td>
<td>k8s-chase-ci-07.calit2.optiputer.net</td>
<td>2080Ti</td>
<td>8</td>
</tr>
<tr>
<td>CHASE-CI</td>
<td>k8s-chase-ci-08.calit2.optiputer.net</td>
<td>2080Ti</td>
<td>8</td>
</tr>
<tr>
<td>CHASE-CI</td>
<td>k8s-chase-ci-09.calit2.optiputer.net</td>
<td>2080Ti</td>
<td>8</td>
</tr>
<tr>
<td>Calit2/UCOP</td>
<td>k8s-gpu-02.calit2.optiputer.net</td>
<td>Titan-x</td>
<td>8</td>
</tr>
<tr>
<td>UIC EVL</td>
<td>lyra-01.evl.uic.edu</td>
<td>1070</td>
<td>1</td>
</tr>
<tr>
<td>Institution</td>
<td>System Information</td>
<td>GPU Type</td>
<td>GPUs</td>
</tr>
<tr>
<td>---------------</td>
<td>--------------------</td>
<td>----------</td>
<td>------</td>
</tr>
<tr>
<td>Calit2/UCOP</td>
<td>k8s-gpu-01.calit2.optiputer.net</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>MRI-SCOPE</td>
<td>suncave-25 to 37</td>
<td>1080Ti</td>
<td>26</td>
</tr>
<tr>
<td>MRI-SENSEI</td>
<td>suncave-0 to 24</td>
<td>1080Ti</td>
<td>50</td>
</tr>
<tr>
<td>UC Merced</td>
<td>ucm-fiona01.ucmerced.edu</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>UC Merced</td>
<td>wave01-10</td>
<td>1080</td>
<td>20</td>
</tr>
<tr>
<td>Calit2/UCOP</td>
<td>fiona8.ucsc.edu</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>Calit2/UCOP</td>
<td>hydra.gi.ucsc.edu</td>
<td>1080Ti</td>
<td>4</td>
</tr>
<tr>
<td>UCSD CSE</td>
<td>k8s-bafna-01.calit2.optiputer.net</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>UCSD ECE</td>
<td>k8s-bharadia-01.sdsc.optiputer.net</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>UCSD ECE</td>
<td>k8s-bharadia-02.sdsc.optiputer.net</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>UCSD ECE</td>
<td>k8s-bharadia-03.sdsc.optiputer.net</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>UCSD ECE</td>
<td>k8s-bharadia-04.sdsc.optiputer.net</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>Calit2/UCOP</td>
<td>k8s-gpu-03.sdsc.optiputer.net</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>OSG</td>
<td>prp-gpu-1.t2.ucsd.edu</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>OSG</td>
<td>prp-gpu-2.t2.ucsd.edu</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>OSG</td>
<td>prp-gpu-3.t2.ucsd.edu</td>
<td>1080Ti</td>
<td>8</td>
</tr>
<tr>
<td>UCR Engr</td>
<td>kube01.engr.ucr.edu</td>
<td>2080Ti</td>
<td>8</td>
</tr>
<tr>
<td>UCR Engr</td>
<td>kube02.engr.ucr.edu</td>
<td>2080Ti</td>
<td>8</td>
</tr>
<tr>
<td>UCSD Calit2</td>
<td>knuron.calit2.optiputer.net</td>
<td>K40</td>
<td>2</td>
</tr>
<tr>
<td>UCSD Calit2</td>
<td>patternlab.calit2.optiputer.net</td>
<td>M40</td>
<td>2</td>
</tr>
<tr>
<td>UIC EVL</td>
<td>evldtn.evl.uic.edu</td>
<td>M4000</td>
<td>1</td>
</tr>
<tr>
<td>Korea KISTI</td>
<td>dttn-gpu2.kreonet.net</td>
<td>titan-x</td>
<td>6</td>
</tr>
<tr>
<td>UCSD CSE</td>
<td>k8s-ravi-01.calit2.optiputer.net</td>
<td>titanxp</td>
<td>8</td>
</tr>
</tbody>
</table>

Figure 2: 2017: PRP Connected 70 UCSD SunCAVE and 20 UCM WAVE 4K Screens to Share VR
2018: CHASE-CI added their 90 Game GPUs to TNRP/OSG for Machine Learning Computations
Figure 3: FIONAs being installed in late 2018 and early 2019 for CPU and GPU computing, and data posting and data transfer analysis (tbd) at UC Merced, Stanford, UC Santa Barbara, UC Santa Cruz and UC Riverside.

Figure 4: CHASE-CI experiments with OSG’s HTCONDOR show that significant 32-bit GPU computational loads can be run in the background, but not at the risk of degrading the performance of the walk-in virtual reality rooms called the SunCAVE (at UCSD) and the WAVE (at UC Merced). Notes Igor Sfiligoi, SDSC, on March 13, 2019: “This morning there was a big demo in the SunCAVE. The IceCube pods were kicked out automatically when the SunCAVE GPUs were in use, and restarted when the demo was over. No admin intervention needed.” We see the merging of CHASE-CI, PRP and the Open Science Grid (OSG) to be a major accomplishment that will enhance machine learning access significantly.
Figure 5: IceCube using 120 GPUs in Nautilus via OSG in the background March 8-14, 2019

Figure 6: Snapshot of details of use of some of the SunCAVE’s GPUs
Figure 7: Top 34 Nautilus GPU users in May 2019. GPU allocated hours, GPU core hours, and CPU hours are shown, and in the FIONA8 column, the equivalent of using an 8 GPU FIONA8 for the whole month, 24 hours a day, is shown for each user. Note that OSG jobs run in the background, so all the other researchers’ jobs have priority.

Science Engagement Tools
Dima Mishin and Nadya Williams built Grafana visualization dashboards for monitoring the storage build out on the PRP and GPU availability and usage, among many other measurements: see https://grafana.nautilus.optiputer.net/. The new dashboards allow switching from a broad cluster-wide view of GPU usage to per-namespace and per-pod, which allows monitoring resource consumption (or waste) in real time and it alerts users if the automated monitoring tools are not enough. See (https://grafana.nautilus.optiputer.net/d/fHSeM5Lmk/k8s-compute-resources-cluster-gpus?refresh=30s&orgId=1) and Figures 8 and 9.
Figure 8: Grafana chart of initial GPU activity on 65 GPU FIONAs on 1-17-19.

Figure 9: Grafana chart of GPU utilization by namespace and total in the cluster for February and March 2019. See https://grafana.nautilus.optiputer.net/d/fHSeM5Lmk/k8s-compute-resources-cluster-gpus?orgId=1&refresh=1m&from=now-90d&to=now
We illustrate in Figures 10–12 the ability to follow individual application projects. One can see how Grafana can enable us to see the history of CPU, memory, and GPU usage on UCI’s Jeff Krichmar’s computational neuroscience project “carl-uci.” In Figure 13, one sees how we can track each pod within the “carl-uci” namespace for % utilization of requested GPUs.

Figure 10: Details of CPU usage by the carl-uci namespace belonging to Jeff Krichmar’s group at UCI. See https://grafana.nautilus.optiputer.net/d/85a562078cdf77779eaa1add43ccec1e/kubernetes-compute-resources-namespace-pods?orgId=1&from=now-90d&to=now&var-datasource=prometheus&var-cluster=&var-namespace=carl-uci
Figure 11: Details of memory usage by the carl-uci namespace belonging to Jeff Krichmar’s group at UCI (scroll down on https://grafana.nautilus.optiputer.net/d/85a562078cdf77779eeaa1addccec1e/kubernetes-compute-resources-namespace-pods?orgId=1&from=now-90d&to=now&var-datasource=prometheus&var-cluster=&var-namespace=carl-uci)
Figure 12: Details of GPU usage by the carl-uci namespace belonging to Jeff Krichmar’s group at UCI (scroll down on https://grafana.nautilus.optiputer.net/d/dRG9q0Ymz/k8s-compute-resources-namespace-gpus?orgId=1&from=now-90d&to=now&refresh=30s&var-namespace=carl-uci). Note Spring break usage drop off.
Figure 13: More details of GPU usage by the carl-uci namespace belonging to Jeff Krichmar’s group at UCI (scroll down on https://grafana.nautilus.optiputer.net/d/dRG9q0Ymz/k8s-compute-resources-namespace-gpus?orgId=1&from=now-90d&to=now&refresh=30s&var-namespace=carl-uci).

Also, a new Ceph dashboard was installed to simplify the administration of storage (Figure 14):
Figure 14: A new Ceph dashboard was installed to simplify the administration of storage.

*sFlow* is a sampling technology that meets the key requirements for a network traffic monitoring solution. sFlow provides a network-wide view of usage and active routes. It is a scalable technique for measuring network traffic, collecting, storing, and analyzing traffic data (see [https://sflow.org/sFlowOverview.pdf](https://sflow.org/sFlowOverview.pdf) for more details). Elastiflow can help us identify malicious actors or out of control software.

sFlow is running on all of the nodes and they stream their data to the ElastiFlow/Elastic Search, instance running in Nautilus (Figure 6a-e). ElastiFlow provides network flow data collection and visualization (see [https://github.com/robcowart/elastiflow](https://github.com/robcowart/elastiflow) for more details on Elastiflow).

Elastiflow lets us see campus traffic grouped by Autonomous System Number (ASN). In Figure 15 below, you can see that UCSC exhibited a surge in traffic, swamping its 100G network (as compared with Figure 16). This was (we think) due to a problem with automated iPerf testing to a FIONA node at UCSC that was very busy running another job. When this bad behavior was literally seen by engineers watching Elastiflow output, iPerf was halted and the situation returned...
to normal (see Figure 16). This problem, now that it has been observed, can be addressed by PRP engineers and their colleagues at other campuses and ESnet.

Figure 15: UC Santa Cruz seeing its traffic peaking due to iPerf testing misbehaving.

Figure 16: Back to normal

Elastiflow also lets us see the major users (Figures 17a-d) and threats (Figure 18).
Figure 17a: The 7-day overview showing top flows on the upper left and tag clouds on the lower right.

Figure 17b: Mousing over the top flow shows it is IceCube at 48 TB.
Figure 17c: Clicking on the ddos in the lower right tag cloud gets the filtered view.

Figure 17d: This shows the top talkers from idc.com.
Figure 18: The Threats Dashboard shows more detail on timing, protocol and rate from each source.

In order to support the TNRP and the NRP Pilot, John Graham set up an NRP Federation Kubernetes Cluster, where the regionals exchange federation credentials. It is also the home of the NRP perfSONAR MaDDash (Figures 19-21).
Figure 19: NRP GridFTP MaDDash January 3, 2019 at https://perfsonar.nrp-nautilus.io/maddash-webui/ although some sites are still off line. Note beyond-PRP (Calit2, Pacific Wave) participation in Missouri, Nebraska, Pennsylvania, New York, and several OSG/Internet2 sites (Chicago, Kansas, New York).
Figure 20: NRP GridFTP MaDDash April 29, 2019 at https://perfsonar.nrp-nautilus.io/maddash-webui/ showing improvement in 3 months and more sites, although some are still off line. Note beyond-PRP (Calit2, Pacific Wave) participation in Missouri, Nebraska, Pennsylvania, New York, Texas, Guam, Nevada, and several OSG/Internet2 sites (Chicago, Kansas, New York).
NRP _GridFTP Dashboard

NRP _GridFTP - Throughput

- Throughput >= 7500Mbps
- Throughput < 7500Mbps
- Throughput <= 5000Mbps
- Unable to retrieve data

⚠️ Found a total of 11 problems involving 9 hosts in the grid

Figure 21: NRP GridFTP MaDDash June 16, 2017 with AWS sites disconnected to minimize costs.
As a direct result of cross-pollination of PRP ideas while preparing for the 18-month review of the PRP at NSF in April 2017, Co-PI Frank Würthwein raised $75,000 in non-PRP funds in collaboration with Open Science Grid (OSG), Calit2, and the San Diego Supercomputer Center (SDSC) to build a 48-GPU machine learning cluster prototype using low-cost game GPUs. The hardware was migrated into the Nautilus cluster after the
latter was made accessible to OSG users as described in reference 4: HEPIX Sfiligoi. The rational here was for both sets of communities, OSG and TNRP users, to be able to use the hardware via familiar mechanisms. OSG and TNRP provided the software integration and user support, getting

- OSG onto Kubernetes for its services.
- Science from OSG onto PRP infrastructure
- Science from PRP onto OSG infrastructure

The IceCube experiment, described in www.nsf.gov/news/news_summ.jsp?cntn_id=295955, shown in Figures 23 and 24 is the first proof of concept.

---

**Figure 23:** IceCube using 120 GPUs at UCSD and UCM March 8-14, 2019; On March 13, 2019, Igor Sfiligoi, of SDSC reported: “This morning there was a big demo in the SunCAVE. The IceCube pods were kicked out automatically when the SunCAVE GPUs were in use, and restarted when the demo was over. *No admin intervention was needed.*”

---

**Figure 24:** IceCube using 120 GPUs in Nautilus via OSG in the background March 8-14, 2019
PRP and OSG engaged Internet2 (Figure 25) in a pilot project to cleanly delineate the scope of support for Big Data science, from the sinks and sources of the data, through the local, regional, and national network providers, to cyberinfrastructure teams, to the actual science teams (Figure 26). As an exemplar of this paradigm, Internet2, in collaboration with the Great Plains Network, deployed FIONA storage hardware at three locations in the network backbone. That hardware was integrated into the K8s Nautilus cluster of the PRP. OSG uses that cluster to deploy its federated caching infrastructure (StashCache). The science user community of OSG is thus provided with additional caching at strategic locations in the backbone. TNRP functioned as enabler by showing OSG personnel in Würthwein’s group how to use K8s as a service deployment vehicle on PRP’s FIONAs in the network.

![Cache at I2 Peering Point With Chicago Cloud Providers](image)

**Figure 25:** Internet2’s network backbone in the US indicating the locations of the initial cache deployment. OSG Data Federation is built on 9 data caches to reduce network traffic and hide data access latencies. There are ~200,000 cores of compute federation across 100 compute elements.

![Science Collaborations (LSST, LIGO, SKA, etc.)](image)

**Figure 26:** Schematics depicting the layered division of responsibility from organizations that support sources and sinks of data to the science collaborations that perform science with that data.
APPENDIX B: PRP/TNRP/OSG Prototype for National-Scale Research Data Infrastructure for Large NSF Instruments

This activity led to a major new use of PRP/TNRP/OSG to prototype future computational/storage cyberinfrastructure to support analysis of the data flows produced by NSF Large Instruments. Specifically, an NSF CESER collaborative proposal (NSF-OAC 1841530) between LIGO (Caltech & GATech), IceCube (UW Madison), Internet2, and OSG personnel at the University of Chicago and UC San Diego was recently funded. This new 2-year proposal focuses on turning the initial prototyping ideas developed between PRP and OSG into a cohesive, federated, national-scale research data infrastructure for large instruments, focusing initially on LIGO and IceCube. The resulting data infrastructure platform addresses production needs of LIGO and IceCube while serving as an exemplar for multi-messenger astrophysics and beyond. In the process, this effort prototypes a redefinition of the role the academic internet plays in supporting science. The collaboration will work closely with PRP in that it uses the PRP Kubernetes infrastructure, and the network coordination via the NRP Pilot.

In addition, LIGO benefits from the strong PRP connection with Cees de Laat’s group at University of Amsterdam. The LIGO/Virgo Scientific Collaboration (LSC) uses significant computing resources in Europe as part of their operations on OSG. The data origin for LIGO/Virgo private data used by all of the LSC operations on OSG is at the University of Nebraska, Lincoln (UNL). Having a cache in Amsterdam significantly reduces transatlantic network traffic, and corresponding delays waiting for data to arrive at compute nodes in Europe. In addition to the private LIGO/Virgo data, the OSG data federation also started serving public LIGO data. This collaboration between PRP and OSG hopes to identify additional public astrophysics datasets to allow multi-messenger astrophysics to be performed on a single platform. LIGO is starting to use Nautilus CPUs as an experiment.

Another expansion of the PRP, The Global Research Platform (GRP), is now formally organizing among the Global Lambda Integrated Facility (GLIF) membership. There will be a Global Research Platform workshop at UC San Diego September 17-18, 2019, with expected GLIF member participation and a GNA/GLIF meeting following on September 19-20 (it will be preceded by the PRAGMA Conference the week before in San Diego, and the e-Science meeting will follow it, also in San Diego).

Active networking national and regional groups in the PRP/GRP efforts are:

- CANARIE in Canada
- CESnet in the Czech Republic
- StarLight in Chicago
- SURFnet in the Netherlands
- NORDUnet in the Nordic countries
- APRP (Asia Pacific Research Platform—APAN members) including:
  - National Supercomputing Centre (NSCC) in Singapore
  - AIST in Tsukuba, Japan
  - KISTI in Korea
  - AARNet consortium in Australia
  - REANZ in New Zealand

All are connecting with FIONAs and similar devices (see Fig. 27). SCAsia (March 11-14. 2019)
in Singapore hosted discussions regarding the APRP. PRP Partner Joe Mambretti (StarLight/MREN/iCAIR) won the SCAsia Data Mover Challenge in the Most Innovative category and is chairing the September GRP meeting, along with UIC’s Maxine Brown.

In 2019 we are [re]implementing a distributed high availability Kubernetes cluster and storage system based upon EdgeFS and the above [re]design (see Figure 28). This is a work in progress.

**Figure 27.** Details of 100G FIONA CPU and memory usage at UCSD 4/4/19 showing pods (containers running on this node) (be patient when clicking this URL since it has to generate this in real time): [https://grafana.nautilus.optiputer.net/d/UM28btNiz/node-pods?orgId=1&var-node=k8s-epyc-01.sdsc.optiputer.net](https://grafana.nautilus.optiputer.net/d/UM28btNiz/node-pods?orgId=1&var-node=k8s-epyc-01.sdsc.optiputer.net)

**Figure 28:** Proposed EdgeFS architecture for HPWREN
Pacific Wave – Western Region Network: GXP Route-Servers with RPKI

As part of the NSF IRNC-funded Pacific Wave SDN/SDX grant, we are exploring a new project that will build upon on the route-server infrastructure instantiated as part of the earlier PRPv2 BGP pilot (Figure 29). The goals of the project include exploring more optimal routing models for data-intensive traffic between Pacific Wave international partners and regional networks and implementing MANRS (Mutually Agreed Norms for Routing Security) guidelines for IXPs:

https://www.manrs.org/ixps/#actions

The initial scope of the pilot will include Pacific Wave WRN participants (ABQG, CENIC, FRGP, PNWGP, U. Hawaii), regional networks (GPN, LEARN, OneNet), and a selection of international NRENs and entities connecting through the Pacific Wave exchange points.

This effort leverages the Juniper vMX route-server instances at the Pacific Wave exchange points in Seattle, Sunnyvale, and Los Angeles. Each vMX instance will have a separate logical-system/logical-router operating as AS397052. The route-servers will each have a session with a RPKI cache server (aka RPKI Validator) to perform Route Origin Validation.

Figure 29: GXP Route-Server with RPKI pilot -- high-level topology diagram v0.10. Please zoom in for details, or browse the on-line source diagram, here: https://docs.google.com/drawings/d/1G_v1x5oDw-_TOtfs7iUQoIoAL0aqxqTYyFbG1B47lk/edit?usp=sharing
• RPKI Validator instances synchronize their local ROA database with the (IRR) trust anchors (Figures 30 and 31).

• Route Servers interact with RPKI Validators, using ROA validation status as a hook for determining BGP policy. Route Servers facilitate BGP policy for routing platforms which do not support for RPKI, and provide routing telemetry and other data to the Looking Glass instances.

• Looking Glass instances provide monitoring and debugging tools to network operators and participants.

Figure 30: Overview of BGP origin validation.
Additionally, the project will deploy Looking Glass functionality to allow pilot participants to analyze and debug the active route announcements in comparison with their BGP policies. OpenBMP is one of the candidate applications for collecting telemetry and providing a ReST/JSON API for the Looking Glass UI (Figure 32).
We expect the Pacific Wave/WRN project will catalyze activity to advance the PRPv2 BGP pilot, with the additional dimension of exploring routing security models that do not impede high-speed, data-intensive workflows across the PRP.

San Diego State University is now using the PRP Kubernetes cluster as a distributed compute-cluster for simulating the injection of CO\textsubscript{2} in brine-saturated reservoirs. Distributed and multi-threaded (i.e., MPI and OpenMP) computation works on our microcloud development cluster and tests have been run on SDSC Comet Virtual Clusters through Nautilus, the PRP Multi-Institution Kubernetes Science DMZ Infrastructure (Figure 33). Researcher Chris Paolini believes this is the first successful use of Open MPI over such a high-speed distributed hypercluster.
Paolini states that as a result of achieving high-speed network data transfers, his CO₂ subsurface modeling team sees requirements that go well beyond the current ~100 Gbps high-end networks up to 0.4/1.6 Tbps networking bandwidth. Also, because of the PRP success in integrating GPUs into FIONAs, his team is investigating how to use compute accelerators (such as GPUs and FPGAs) in FIONAs for his codes. Further, Paolini reported at the NRP 2\textsuperscript{nd} Workshop held in Bozeman MT in August, 2018 that his team is now investigating

- Distributed-Parallel (data acquisition, rendering) large-dataset visualization with ParaView using their client-server architecture on a PRP/NRP Kubernetes network.
- Integrating a PRP/NRP Kubernetes network with SDSU’s new NSF-funded 2.4PB BeeGFS storage cluster to support parallel I/O for I/O bound computation (MPI-IO, parallel NetCDF, parallel HDF5)

The PRP/NRP is also being extended to include IoT (distributed sensors and wearable devices using LPWAN on LoRa PHY) projects and projects involving embedded ML, OD, and DL on μControllers and FPGAs.
APPENDIX C: Integrating Open Science Grid (OSG) with SDX Projects Workshop - June 5-6, 2019 (A Workshop Supported in Part by TNRP Participant Support Funds)
by Calit2 & FIU-CIARA

The goal of this workshop is to explore the technical requirements for integrating the Open Science Grid (OSG) with the SDX projects of the IRNC program. This workshop is scheduled for June 5-6, 2019. The venue will be in The California Institute for Telecommunications and Information Technology (Calit2) at the University of California San Diego (UCSD) campus in La Jolla, CA.

Please register (space is limited) at the following page and indicate if you will be participating in person or remotely. A program, logistics for travel and lodging, and other relevant information will be announced soon.

Wednesday, June 5, 2019
08:30 Registration (coffee kiosk is conveniently located outside the building--no breakfast is served)
- Room 5302, Atkinson Hall, UC San Diego
  09:00 Welcome Address (15 minutes presentation, 5 minutes Q&A) - Larry Smarr
  09:20 Context and Goals of the workshop (15 minutes presentation, 5 minutes Q&A) - Julio Ibarra
  09:40 OSG Keynote (40 minutes presentation, 10 minutes Q&A) - Frank Wuerthwein
  - What are challenges OSG has in its landscape and foresees in its roadmap?
  10:30 Refreshment break (30 minutes) - (coffee kiosk is conveniently located outside the building)
  11:00 NSF IRNC Software Defined Exchange Panel (Introduction, 10 minutes) - Julio Ibarra (Moderator)
    - Introduction about each of the IRNC SDX projects: Scope, Goals, Accomplishments, etc.
    11:10 StarLight-SDX (20 minutes presentation) - Jim Hao Chen
    11:30 PacifiWave-SDX (20 minutes presentation) - John Dundas
    11:50 AtlanticWave-SDX (20 minutes presentation) - Julio Ibarra
  12:10 Q&A session (30 minutes)
  12:30 Lunch (1 hour) - (wear comfortable shoes we will walk to the food court)

13:30 Integration Panel (15 minutes presentation, 5 minutes Q&A for each presenter) Tom DeFanti (moderator)
- Architecture, Components, Interfaces and other relevant information to reveal what resources are available, and also not available, for potential integration
  13:40 OSG Layers and components - Edgar Fajardo
  14:00 Kubernetes Layers and components - John Graham
  14:20 AtlanticWave-SDX layers and components - Yufeng Xin
  14:40 StarLight-SDX layers and components - Jim Hao Chen
  15:00 Refreshment break (30 minutes) - (coffee kiosk is conveniently located outside the building)
15:30 PacificWave-SDX layers and components - John Hess
15:50 Integration panel Q&A and wrap up (20 minutes)
16:10 Roadmap to the integration with OSG: Plenary

- Describe and represent what OSG and SDXs want to accomplish. The aim of this session is to provide input to Thursday’s design and integration session.

17:10 Plans for day 2 of the workshop Julio Ibarra, Tom DeFanti, Frank Wuerthwein
18:00 Dinner (on your own). Please see below some suggested restaurants. A full list can be found here.

Thursday, June 6, 2019

8:30 Registration
09:00 Integration, Design, and Planning Jam Session - Moderator: Julio Ibarra
The goal of this session is to design how to integrate the SDXs with Kubernetes and OSG. Detailed diagrams and descriptions about how to accomplish the integration should be an outcome.

Figure 34: Joining Chameleon to experiment with and integrating it and Nautilus.
APPENDIX D: TNRP Network Engineering Activities—Toward Meeting NPEO#1-#8 Solicitation Requirements (E-mail correspondence)

The next 60 pages are a selection from hundreds of e-mails between engineers relevant to TNRP and its first 8 NPEO goals, listed here:

[NPEO#1] Identifying potential bottlenecks in data transfers for scientific collaborations.
[NPEO#2] Providing analysis and objective recommendations on tools, technologies, and integrated systems related to data network movement and transfer performance.
[NPEO#3] Serving as a centralized point of expertise, advice, and engagement for distributed scientific collaborations in identifying and solving scientific data movement inefficiencies adversely affecting time-to-science.
[NPEO#4] Disseminating current practices across the NSF community on achieving highly efficient end-to-end data transfer performance, especially in high-bandwidth/high-delay environments.
[NPEO#5] Providing online and in-person training and education on network performance analysis and troubleshooting.
[NPEO#6] Engaging with CI Engineer/Advanced Cyberinfrastructure-Research Education Facilitator/Cyberteam award personnel, and other relevant groups comprising the collective CI engineering expertise in the community in which network performance plays an important role.
[NPEO#7] Investigating, evaluating, and testing data movement systems and tools, transport protocols, and integrated platforms.
[NPEO#8] Selectively investigating and solution engineering for end-to-end performance.

-----------------

May 31, 2019,

Hello John and Dima,

Currently, the StartLight node is relying on proxy for public access but the pods on top still cannot reach outside. The main reason for that is the default network using 100G link which is not routed to public. We found a solution for this issue and this would probably be useful for other private nodes as well.

Based on our discussion in rocketChat, your network plugin Calico is selecting interfaces whichever access 8.8.8.8.(IP AUTODETECTION METHOD=can-reach=8.8.8.8). But our high speed interface cannot access 8.8.8.8. So our proposed solution is to try reach your master node(67.58.53.147) instead. In this case, Calico will chose our 100G interface for pod-to-pod network while our public interface can still remain as default network for outside connection. This should work for every node and prerequisite is just setup a route to your master with the interface we desire. In addition, proxy is not needed any more.
We have tested the proposed solution in our cluster and it worked. Please consider this and let us know your thoughts. Thanks!

Regards,
Shawn,
iCAIR, Northwestern University

May 29, 2019
We were the 4th largest provider of CPU to IceCube, too.
https://gracc.opensciencegrid.org/dashboard/db/payload-jobs-summary-iccube?orgId=1&from=now-7d&to=now

And the biggest in the USA!

Igor
IceCube wall clock hours 5-23-19 to 5-29-19. See https://gracc.opensciencegrid.org/dashboard/db/payload-jobs-summary-icecube?orgId=1&from=now-7d&to=now

Kibana dashboard showing Nautilus flows using Elastiflow on 6-6-19. See https://esk.nautilus.optiputer.net/app/kibana
user admin *ask for password*
John Graham, May 30, 2019: I updated my tstat daemonset and its now deployed on all nodes. [https://gitlab.nautilus.optiputer.net/prp/tstat](https://gitlab.nautilus.optiputer.net/prp/tstat)
Sflow has been running smoothly [https://gitlab.nautilus.optiputer.net/prp/sflow](https://gitlab.nautilus.optiputer.net/prp/sflow)

Hi Tom.

A recent exercise in trying to add a OSG site (UMD) to the Nautilus k8s cluster hit a wall...
They do not operate a Science DMZ, so the networking was “not good enough” for the current Nautilus Calico setup.

Discussing with Dima, it seems it would be hard to add such nodes to the Nautilus; it just add too much operational painpoint.
We would be better off/ less ops load if we had a separate k8s cluster for such nodes.

After discussing some more with Frank, too, we figured out it would actually make a lot of sense to move all of the OSG services in a k8s cluster with “lower Calico grade” networking.
All our services use the host networking anyway, so even for the network critical service we don’t really care about Calico tuning.

Would this be OK with you if we went ahead with creating a separate k8s cluster for the osg services?
(not opportunistic OSG use… that would stay in the “main Nautilus”).
Frank and Dima would be happy to go that path.

Please let us know.

Thanks,
Igor

---

_Dima Mishin_  
May 22, 2019, 5:07 PM  
Reply to all

Another advantage is that those nodes not controlled by us (me and JohnG) will not hold off the updates, and can be updated at the pace comfortable for OSG

--

Best regards,
Dima Mishin
John Graham

to Thomas, Igor, John, frank, Dima

Yup it sounds great. I can use one of the Phil hand-me-down 1U boxes in the datacenter :)

Frank Wuerthwein

to jjgraham, Thomas, Igor, Dima

Many thanks to you all!

I’m hoping that longer term, OSG will take this over so its out of your hair eventually. But for the next 6-12 months I much prefer to keep it with PRP.

Am counting on Dima, John, and Igor to keep an eye on “upgrades” and suggest them as you think they are necessary. So that the OSG K8s operated by PRP does not get stuck completely in old K8s versions.

Many thanks, frank

Dima Mishin

to Frank, John, Tom, Igor, Dima

Am counting on Dima, John, and Igor to keep an eye on “upgrades” and suggest them as you think they are necessary. So that the OSG K8s operated by PRP does not get stuck completely in old K8s versions.

No problem with that, I'll be managing it the same way I do with Nautilus

Larry Smarr

to Carlos, frank, Tom

Carlos-as we discussed at CENIC 2019, I talked to PRP co-PI Frank Wuerthwein about where a good first PRP LHC faculty in Mexico would be. Note Frank is one of our senior UCSD Physics professors, is Executive Director of OSG, and has a position in SDSC.
Puebla seems to be the answer and as luck would have it, Frank has hired a CS person from Puebla into his UCSD group. So the stage is set on both sides of the border! Frank—as I mentioned Carlos Casasus is one of our long-time collaborators. He is head of CUDI (I2 for Mexico) and has helped us with prior efforts to extend 10Gbps into Mexico.

Carlos—can you follow up with Frank on this and keep me and Tom DeFanti cc-ed? It sounds like you could help out the Puebla people with networking right away. The key is to get them coupled into our Kubernetes Nautlius hypercluster, which also couples with OSG.

Hope we can get this to work so Mexico joins our international PRP partners!

Carlos—here is some background from Frank to me after you and I talked at CENIC:

Larry—I looked into this and there is a computing center at Puebla. The BUAP CMS group runs that. I know one of the faculty there reasonably well. And we just hired Diego Davila, one of the CS people from Puebla who spent the last 2-3 years at CERN working in computing at CERN.

Together with the Puebla faculty [Isabel Pedraza www.linkedin.com/in/isapedraza/ ] I wrote a proposal to the Mexican government’s US-Mexico collaborations solicitation a couple years ago, but it wasn’t funded. As a result. Not much happened. Diego joined us March 15th. The plan is that he will have some time on TNRP to make budget work out because Igor is too expensive.

Bottom line, we can certainly do this.

I talked with Diego and apparently, UNAM has an ALICE effort, the heavy ion physics experiment at the LHC. Puebla has a 10Gbps connection to UNAM, and from there to the rest of the world. So, the CMS T3 at Puebla is network wise connected via UNAM to San Diego. CMS@Puebla and ALICE@UNAM, so I’m told.

I asked Diego to get me a trace route. He has root access in Puebla, and wants to continue helping them.

I’m happy to support this, especially if we do it via the nautilus cluster, i.e. have the necessary site services operated from UCSD because they are deployed in containers available via Nautilus.

Also, I want the Puebla cluster to show up in OSG, and get all of its CMS data from my cluster. It makes no sense for them to operate their own storage infrastructure. If we do this, then we guarantee network use from Puebla to us.

Diego also said that presently they have network connectivity trouble that they need to sort out.

So there’s more useful work for PRP/TNRP.

Are you ok going ahead with this from a PRP/TNRP perspective?
Thanks, Frank

On Fri, Apr 5, 2019 at 9:53 AM Carlos Casasús <ccasasus@cudi.edu.mx> wrote:
Larry,

How do we follow up on this? UNAM. Puebla and Guadalajara could be interested.

Please let me know.

Carlos
Carlos Casasús
Corporación Universitaria para el Desarrollo de Internet, A.C.
Director General
Parral 32
C.P. 06140
México, D.F.
www.cudi.edu.mx
ccasasus@cudi.edu.mx
+52 (55) 52-11-18-32

From Alex Feltus' group:

Sfiligoi, Igor
May 21, 2019, 2:19 PM
Reply to all

Hi Alex.

Can you let us know if your workflow can run as “a regular batch job”? Or do you really need k8s? And if you do, what parts of it you need that are not available to “plain batch systems”?

Thanks,
Igor

PS: If “good old batch system” is all you need, would having a batch interface to k8s help?

John Graham
May 21, 2019, 5:56 PM
Reply to all

Alex
Happy to support direct Nautilus experiments using our EdgeFS. Got any use for a FPGA with 512GB RAM? :)

---

**Ben Shealy**

to jjgraham, Igor, Alex, Dmitry, me, John, Frank

Hey everyone,

Lots of good questions!

Nextflow can interface with both traditional HPC systems and k8s. We have tested our workflows already with PBS and SLURM, because our university clusters (Clemson and WSU) use those. I am now trying to map our workflows to k8s because, from what I can tell, **nextflow seems to be the best way for me to use the NRP as a batch system.** Kubernetes alone does not have very good support for batch jobs, but nextflow has mechanisms for creating and deleting pods much in the same way as submitting jobs to an HPC scheduler.

@Frank The code for interfacing with k8s is part of nextflow itself, and the code for moving data is in the kube-runner repo, there are two scripts called kube-load.sh and kube-save.sh. Dmitry and I set up a PVC on the Ceph shared filesystem, so when running a pipeline on the NRP I first copy input data to the PV and copy output data from the PV afterwards.

GEMmaker does download SRA files from NCBI, currently iRODS is not involved. It uses a custom NCBI tool called Aspera, which uses UDP for data transfers instead of TCP, and so far we've been able to use it at scale on our university clusters.

Nextflow can also interface with HTCondor, which may make an OSG run feasible. However, nextflow needs to run on some sort of "head node" so that it can submit jobs. Essentially it must take the place of Pegasus. Assuming we can overcome these hurdles, we would love to run some large-scale experiments on OSG.

@John If I can find another undergrad minion I think we might be able to use that FPGA :)

---

**Sfiligoi, Igor**

to Ben, John, Alex, Dmitry, me, John, Frank

Hi Ben.

OSG submission through HTCondor submission always assumes you have a "head node" (of some sort) to submit from.
If we want PRP involved, we could instantiate that as a Pod in PRP k8s itself. I would be happy to help there.

Cheers,
Igor

---

**Ben Shealy**
May 22, 2019, 10:25 AM
Reply to all

to Igor, John, Alex, Dmitry, me, John, Frank

Hi Igor,

Excellent! I had assumed that there was only one "global" head node for OSG, but if we can provide our own head node then all the better. Nextflow only needs the condor submit command to be available. Can you provide a pod spec / usage example for submitting a job to OSG from a pod?

---

**Stiligoi, Igor**
May 22, 2019, 11:15 AM
Reply to all

to Ben, John, Alex, Dmitry, me, John, Frank

Nope… OSG is all about “distributed and federated” :-(

All we need is a schedd running on the node, and then you (locally) submit to it standard vanilla jobs. You can add additional requirements to steer where the job can run, but I (think) it is optional.

We will of course need to properly configure the schedd to talk with the rest of the OSG infrastructure. Not that hard, but I am not the one who is in charge with these details. Full disclosure: We never done it from a k8s Pod, but it is on a to-do list and we would be happy to have you as our first guinea pig :)

Please let me know if you are interested in following this path, and I will put you in touch with the right people.

Thanks,
Igor
Well we would be happy to be your guinea pig for this, I'm quite used to it at this point.
TNRP MaDDash with AWS tests running (above) and not (below). A request for a grant of AWS time is in process so this work can continue. From Kate Adams who works for Great Plains Network: For right now we just want to test our member's nodes against each other, and against the GPN node.

Hi all.

The GPN workshop has a hands-on session where I will want people to experiment with launching real pods. I have created a dedicated namespace for them, but now we need to plan for where they will run.

Given the audience, I would like to give the access to at least the host interface. And possibly to the local filesystem, too. Those are privileged operations not available to regular users. And for good reason.

So it would be good if those pods were running on nodes not used by other users and not used for critical services. The suncave nodes seem the best choice to me.

Alo, Richard

to Thomas, Larry

Dear Tom and Larry

I hope all is well

FAMU is definitely on the move here as we are recognizing our campus needs in computing.
Is it still possible to send 2/3 of our IT and a faculty to UCSD for training under the NRT?

1. We all did submit a NSF CC* TEAM proposal to create Southeastern Education and Research Network [SERN].
2. In the interest of improving our HPC capabilities, Florida Congressman Lawson is identifying $2M to encourage a partnership [probably with HP] and a proposal submission requiring better computing resources. I am working closely with our VPR on this. He is recognized for developing the Plasma research and PhD in Physics that we have.
3. We will be starting a PhD programs in Biology and in CDS&E in fall 2020
4. We have a partnership with the USDA Forestry Service in creating a Transdisciplinary/Translational Data Science Research Center [serving as an initial set of CDS&E research problems]. On the agenda:
   a. Drone Data of the Western Florida Panhandle from pre/post Hurricane Michael [U Florida partnership with their drones and their new unique sensors]
   b. Intrusion of salt water in Florida Aquifer
   c. Land subsidence
5. Further enhancing our TDSRC is our partnership with FIU[brain images of patients with neuro degenerative diseases] Mt Sinai Hospital, UF and FIU Med School to study social determinants of patient data [ADNI- Alzheimer’s]
6. We are in discussions with Dow who wants to sponsor a joint PhD in Chemistry /Chem Engr. Our joint Engineering College [FAMU/ FSU] is now producing the most African American PhDs
7. We have completed the SCOPE Walls and Classroom without Borders.
8. We hope to take a particular direction in our Bio program, viz., Applications of Mechanistic Modelling and Machine Learning in Biology. In preparation we are working with James Glazer [PI] and Geoffrey Fox of IU Bloomington on a NSF preproposal for a STC in the discipline.

Thus you can see that we are building an infrastructure and at same time working to change our teaching /research model as we move to increase our PhD production by at least 12 per year making us eligible to apply for Tier1 Status [our rival Howard needs 33]. Hence we have significant opportunities ahead of us.

Thomas DeFanti

How deep a dive do you want for these 4 people? What background do they have in network administration?
It might take more time if you want a lot of hands-on.
Let’s iterate.
..tom..

Alo, Richard

Thanks so much Tom for the quick reply.
Since Ron Henry, our CIO/ Assoc VP is copied here, I would like him to reply to the depth we need.
I have included Ron and his team in the NSF CC*TEAM proposal to create the Southeastern Education and Research Network so they are well aware of the direction we need to go.

Also since we have installed the NSF SCOPE Viz Walls here [Please inform Falko for us] it would be great that while they are on the California Institute of Telecommunications and Information Technology campus, it would be great for them to see the Viz Walls there but also the STAR CAVE allowing them to walk through the Coral Reefs.

For informational purposes I have cc'ed FAMU upper administration.

Thank you for considering training out Team.

Best wishes
Richard

From: Edward Colone
Subject: perfSONAR and GridFTP
Date: June 4, 2019 at 3:35:53 PM EDT
To: John Hicks <jhicks@internet2.edu>, Mark Feit <mfeit@internet2.edu>
Cc: Nathan Shepherd <nateshep@umich.edu>, Andrew Lake <andy@es.net>

Thanks to Andy Lake for the helping hand and the info:

Here's a page on how to get the D2D plugins installed and configured:

https://github.com/perfsonar/project/wiki/pScheduler-Disk-to-Disk-Plugin-Installation

Here's an example GridFTP psconfig json file:

https://raw.githubusercontent.com/perfsonar/perfsonar-dev-mesh/master/psconfig/gridftp.json

Here's what a dashboard for GridFTP looks like:

http://ps-dashboard.es.net/maddash-webui/index.cgi?dashboard=GridFTP%20Development

Thanks,

-Ed
Marla Meehl

to Tom, John, John, Lee, Joe, Thomas

All,

We are looking for a 30 minute presentation on kubernetes/containers - what are they, what are they used for, what are they good for. This is for the Westnet meeting at the University of Utah 6/25 - 6/27. It could be a remote presentation. Any suggestions?

Thanks

Marla, We use your Cloudstream containers for all kinds of applications.
https://github.com/Unidata/cloudstream
This is AWIPS CAVE running on our Nautilus cluster

Joe Breen

to Thomas, Alex, Scott, Marla, John, John, Lee, Thomas, Westnet

If desired, one of my students also has a few basic slides on Kubernetes and containers that we are using to onboard other students. Depending on which day/time of Westnet (I have to step in/out), I would also be willing to add comments from various lessons learned. FWIW, --Joe

Jim Kyriannis

to prp-nrppilot-engagement@internet2.edu, Jim, Dana

* A question that's come up a number of times in conversations with our members is what distinguishes the NRP Pilot from compute resources available from XSEDE or OSG. Having documentation and a link we can share with campuses and interested researchers that describes the unique value the NRP Pilot offers over other available resources will be very helpful in attracting participation.

* Related to the above, there are also several links available which describe the compute loads which are well-suited to OSG (HTC optimized jobs) and which are not (e.g., https://swc-osg-workshop.github.io/2017-05-17-JLAB/novice/DHTC/01-IntroGrid.html). Something similar for the NRP Pilot would be helpful as well.

From: John Hicks <jhicks@internet2.edu>
Subject: routing to/from AWS
Date: May 16, 2019 at 6:47:35 AM PDT
To: "Sfiligoi, Igor" <isfiligoi@sdsc.edu>
FYI - concerning the traffic between the I2 Stashcache machines and AWS, we (I) need to add a community to the OSG prefixes when they are exported out of the VRF and going to bleed I2PX routes into the OSG VRF. This will happen next week and should fix the issue to/from regular AWS images. Igor - I have a meeting tomorrow (2pm (ET)) to discuss getting you the “express route” L3VPN peering setup. We are still working on google. Also, working on machine deployment and routing in Amsterdam.

- John

For now, I need what you think the path forward for the AWS testing should be. Can you test and get useful results (with GrifFTP, say) for the NRP Pilot MaDDash? Yes, but the async nature of the current setup means we have to keep the VMs alive basically all day long.

Tom: Can you measure AWS both ways? If not, what do we need to do?
Igor: Only one way right now… outgoing traffic only.
The limit is in the globus-url-copy tool… it does not like the private-IP-only setup. I think this could be fixed by a small patch to that software… I have the source code, so I should be able to do it.
If you say it is high priority, I may be able to do it by Jun 15th.

Can you do tests with FDT and/or xrootd? If so, when?
Not much experience with FDT, but i am pretty sure I can get xrootd transfers going. Not sure how to integrate the results into a MadDash, but getting the raw numbers should be doable.

For FDT… I guess it would be doable, but need to first to familiarize with it.

I just need to say in the report that we are on it, and will have whatever done by Sept. 30, done but more to do by Sept. 30, or something we'll do after Sept. 30. I'm not picky about the timing, but I want to say something!
I think we can do all tests by Sept 30.
i don’t see any showstoppers.

Now, we may find problems, like the bad routing via Internet2… That may take longer to resolve… but we can only speculate at this point.
APPENDIX E: 2019 PRP/NRP PILOT ENGAGEMENT ~ RUNNING MEETING NOTES

From https://docs.google.com/document/d/1JfQO_odoUyYaNfCzRWz_cvrN2gtQj9ovbq9GJUkZyG0/edit (18 pages of notes as of 6-6-19):

MEETINGS: EVERY OTHER MONDAY ~ 10 PT/ 11 MT/ 12 CT/ 1 ET

EMAIL DISTRIBUTION LIST:
- prp-nrppilot-engagement@internet2.edu

CONNECTION DETAIL:
Join from PC, Mac, Linux, iOS or Android: https://internet2.zoom.us/j/7203992979

The following people are invited to the bi-weekly Zoom calls
- Akbar Kara - Learn
- Alex Feltus - Clemson
- Ana Hunsinger - Internet2
- Camille Crittenden - CITRIS and the Banatao Institute
- Cathy Chaplin - Internet2
- Chris Hoffman - Berkeley
- Dana Brunson - Internet2
- Dima Mishin - UCSD
- Eli Dart - ESnet
- Jack Shnell - Berkeley
- James Deaton - GPN
- Jason Zurawski - ESnet
- Jeff Weekley - UC Merced
- Jennifer Griffin - The Quilt
- Jen Leasure - The Quilt
- Jen Oxenford - Kinber
- Jen Schopf - Indiana University
- Jim Bottum - Internet2
- Jimmy Kyriannis - NYSERNet
- John Graham - UCSD
- John Hicks - Internet2
- Joseph Keefe - UCSD
- Kate Adams - GPN
- Kate Mace - ESnet
- Larry Smarr - Calit2
- Lauren Rotman - ESnet
- Lee Ann Weber - CENIC
- Mark Yashar – Berkeley
- Meggan Levitt - UC Davis
- Melissa Lucas - CENIC
- Pankaj Shah - LEARN
- Patrick Schmitz – Berkeley
- Scott Sellars – Dept. of State
- Shawfeng Dong - SLAC
- Steve Kankus - NYSERNet
- Tim Boerner - XSEDE
- Tom DeFanti - UCSD
I ran a traceroute inside AWS, between a node in aws-east-1 and one in aws-east-2.

Igor

[root@ec2-3-19-82-180 centos]# traceroute ec2-3-213-155-48.compute-1.amazonaws.com
trace route to ec2-3-213-155-48.compute-1.amazonaws.com (3.213.155.48), 30 hops max, 60 byte
packets
1  ec2-52-15-0-14.us-east-2.compute.amazonaws.com (52.15.0.14) 25.734 ms ec2-52-15-0-10.us-
est-2.compute.amazonaws.com (52.15.0.10) 11.908 ms ec2-52-15-0-6.us-east-
est-2.compute.amazonaws.com (52.15.0.6) 25.723 ms
2  100.64.0.78 (100.64.0.78) 21.488 ms 100.64.0.140 (100.64.0.140) 16.961 ms 100.64.3.10
   (100.64.3.10) 29.423 ms
3  100.66.2.36 (100.66.2.36) 18.018 ms 100.66.3.38 (100.66.3.38) 19.193 ms 100.66.2.200
   (100.66.2.200) 21.114 ms
4  100.66.7.203 (100.66.7.203) 14.035 ms 100.66.6.105 (100.66.6.105) 15.021 ms 100.66.7.239
   (100.66.7.239) 21.455 ms
5  100.66.4.197 (100.66.4.197) 17.841 ms 100.66.4.47 (100.66.4.47) 16.302 ms 100.66.4.251
   (100.66.4.251) 22.885 ms
6  100.65.8.129 (100.65.8.129) 0.576 ms 100.65.8.193 (100.65.8.193) 0.258 ms 100.65.10.225
   (100.65.10.225) 0.718 ms
7  52.95.1.7 (52.95.1.7) 11.630 ms 52.95.3.127 (52.95.3.127) 11.547 ms 52.95.3.129
   (52.95.3.129) 11.606 ms
8  52.95.1.110 (52.95.1.110) 14.130 ms 52.95.2.114 (52.95.2.114) 19.348 ms 52.95.1.138
   (52.95.1.138) 16.562 ms
9  52.95.2.209 (52.95.2.209) 12.340 ms 52.95.2.99 (52.95.2.99) 11.729 ms 52.95.3.33
   (52.95.3.33) 11.620 ms
10 100.91.39.20 (100.91.39.20) 28.311 ms 100.91.41.18 (100.91.41.18) 27.786 ms
100.91.39.102 (100.91.39.102) 16.447 ms
11 52.93.130.140 (52.93.130.140) 11.294 ms 52.93.130.134 (52.93.130.134) 11.264 ms
52.93.130.148 (52.93.130.148) 27.210 ms
12 100.91.163.76 (100.91.163.76) 11.399 ms 100.91.163.58 (100.91.163.58) 11.651 ms
100.91.163.24 (100.91.163.24) 11.765 ms
13 100.91.168.143 (100.91.168.143) 11.358 ms 100.91.163.19 (100.91.163.19) 12.641 ms
100.91.163.43 (100.91.163.43) 12.027 ms
14 100.91.164.40 (100.91.164.40) 11.778 ms 100.91.165.136 (100.91.165.136) 11.656 ms
100.91.160.20 (100.91.160.20) 17.057 ms
15 100.91.159.99 (100.91.159.99) 11.806 ms 100.91.160.5 (100.91.160.5) 17.199 ms
100.91.159.11 (100.91.159.11) 12.349 ms
16 100.91.177.211 (100.91.177.211) 11.108 ms 100.91.177.173 (100.91.177.173) 11.077 ms
100.91.177.69 (100.91.177.69) 11.386 ms
17  * * *
18  * * *
19  * * 52.93.28.78 (52.93.28.78) 11.226 ms
20  * 100.65.14.16 (100.65.14.16) 24.160 ms *
21  100.66.5.158 (100.66.5.158) 22.416 ms 100.66.5.104 (100.66.5.104) 24.374 ms *
22  52.93.28.94 (52.93.28.94) 18.802 ms 52.93.28.102 (52.93.28.102) 11.504 ms 52.93.28.74
   (52.93.28.74) 11.512 ms
23  100.65.15.48 (100.65.15.48) 25.023 ms 100.65.12.64 (100.65.12.64) 28.128 ms
100.66.14.179 (100.66.14.179) 23.715 ms
24  100.66.13.171 (100.66.13.171) 31.427 ms 100.66.5.192 (100.66.5.192) 33.116 ms
100.66.5.248 (100.66.5.248) 27.844 ms
25  ec2-3-213-155-48.compute-1.amazonaws.com (3.213.155.48) 31.313 ms 100.66.6.18
(100.66.6.18) 26.276 ms ec2-3-213-155-48.compute-1.amazonaws.com (3.213.155.48) 11.315 ms
Sfiligoi, Igor

to me, John, Larry, Frank, John

Hi all.

And here are the traceroutes for Los Angeles (west2) Google region.
Only premum IP avaiable.

UCSD looks good, internet2 sites all sport the same RTT, which is unexpected.

Cheers,
Igor
# Premium

[root@ps-40g-gridftp ~]# traceroute 35.236.48.53
traceroute to 35.236.48.53 (35.236.48.53), 30 hops max, 60 byte packets
1  3826-router.optiputer.net (67.58.50.65) 0.099 ms 0.141 ms 0.134 ms
2  67.58.48.38 (67.58.48.38) 3.146 ms 3.153 ms 3.197 ms
3  dc-sdg-agg4--ucsd-100ge.cenic.net (137.164.23.176) 0.468 ms 0.501 ms 0.528 ms
4  dc-tus-agg3--sdg-agg4--100ge-2.cenic.net (137.164.11.10) 2.023 ms 2.052 ms
dc-tus-agg3--sdg-agg4--100ge.cenic.net (137.164.11.8) 2.079 ms
5  dc-lax-agg6--tus-agg3--100ge-1.cenic.net (137.164.11.22) 2.896 ms 2.921 ms
dc-lax-agg6--tus-agg3--100ge-2.cenic.net (137.164.11.24) 2.830 ms
6  72.14.222.56 (72.14.222.56) 2.693 ms 74.125.49.165 (74.125.49.165) 2.748 ms
74.125.49.165 (72.14.222.56) 2.715 ms
3.752 ms
8  216.239.48.232 (216.239.48.232) 3.793 ms 209.85.247.8 (209.85.247.8) 3.668 ms
209.85.251.112 (209.85.251.112) 2.818 ms
9  * * *
10 * * *
11 * * *
12 * * *
13 * * *
14 * * *
15 53.48.236.35. bc.googleusercontent.com (35.236.48.53) 4.102 ms 4.100 ms *

[isfiligoi@osg.chic ~]$ traceroute 35.236.48.53
traceroute to 35.236.48.53 (35.236.48.53), 30 hops max, 60 byte packets
1  gateway (163.253.70.1) 0.407 ms 0.371 ms 0.361 ms
2  UCAID.bar1.KansasCity1.Level3.net (4.53.32.126) 11.457 ms 11.327 ms 11.544 ms
3  ae5-6.bar1.KansasCity1.Level3.net (4.53.32.125) 11.181 ms 11.208 ms 11.200 ms
4  * * *
5  72.14.242.34 (72.14.242.34) 35.092 ms 35.675 ms 35.063 ms
6  74.125.37.20 (74.125.37.20) 35.961 ms 209.85.246.100 (209.85.246.100) 35.091 ms
108.170.252.130 (108.170.252.130) 35.107 ms
8  108.170.228.82 (108.170.228.82) 35.413 ms 216.239.63.207 (216.239.63.207) 35.723 ms
35.789 ms
9  209.85.251.183 (209.85.251.183) 67.824 ms 67.580 ms 67.659 ms
67.343 ms
108.170.247.225 (108.170.247.225) 66.818 ms
12 209.85.247.8 (209.85.247.8) 67.900 ms 216.239.51.136 (216.239.51.136) 67.949 ms
216.239.46.134 (216.239.46.134) 67.598 ms
13 * * *
14 * * *
15 * * *
16 * * *
17 53.48.236.35. bc.googleusercontent.com (35.236.48.53) 67.752 ms *

# From inside the Los Angeles vm out
#
[root@instance-w3 isfiligoi]# traceroute ps-40g-gridftp.calit2.optiputer.net
traceroute to ps-40g-gridftp.calit2.optiputer.net (67.58.50.66), 30 hops max, 60 byte packets
Here are the paths into the Ohio (central1) Google region; Internet2 from Chicago still routes via Kansas!

Igor

# # Standard IP
# [root@ps-40g-gridftp ~]# traceroute 35.208.57.50
traceroute to 35.208.57.50 (35.208.57.50), 30 hops max, 60 byte packets
1 3826-router.optiputer.net (67.58.50.65) 0.118 ms 0.141 ms 0.135 ms
2 67.58.48.38 (67.58.48.38) 0.313 ms 0.321 ms 0.330 ms
3 dc-sdg-agg4--ucsd-100ge.cenic.net (137.164.23.176) 0.488 ms 0.524 ms 0.545 ms
4 dc-tus-agg3--sdg-agg4-100ge-2.cenic.net (137.164.11.18) 2.042 ms 2.070 ms dc-tus-agg3--
   sdg-agg4-100ge.cenic.net (137.164.11.8) 2.097 ms
5 dc-lax-agg6--tus-agg3-100qe-1.cenic.net (137.164.11.22) 2.901 ms dc-lax-agg6--tus-agg3-
   100qe-2.cenic.net (137.164.11.24) 2.849 ms dc-lax-agg6--tus-agg3-100qe-1.cenic.net (137.164.11.22) 2.919 ms
6 dc-lax-agg8--lax-agg6-100qe-2.cenic.net (137.164.11.7) 4.238 ms dc-lax-agg8--lax-agg6-
   100qe-4.cenic.net (137.164.11.37) 5.146 ms dc-lax-agg8--lax-agg6-100qe-2.cenic.net (137.164.11.7) 4.420 ms
7 et-1-0-2.0.rtsw.losa.net.internet2.edu (64.57.20.82) 2.817 ms 2.789 ms 2.818 ms
8 ae-1.4079.rtsw.salt.net.internet2.edu (162.252.70.115) 15.548 ms 15.614 ms 15.476 ms
9 ae-5.4079.rtsw.kans.net.internet2.edu (162.252.70.144) 35.262 ms 35.348 ms 35.412 ms
10 ae-3.4079.rtsw.chic.net.internet2.edu (162.252.70.140) 46.304 ms 46.424 ms 46.439 ms
11 lo-0.8.rtsw2.ech.net.internet2.edu (64.57.29.130) 46.465 ms 46.526 ms 47.180 ms
12 72.14.220.117 (72.14.220.117) 52.774 ms 74.125.49.146 (74.125.49.146) 52.760 ms
14 108.170.244.2 (108.170.244.2) 48.613 ms 108.170.244.15 (108.170.244.15) 71.855 ms 46.884 ms
15 209.85.250.35 (209.85.250.35) 47.236 ms 209.85.251.241 (209.85.251.241) 47.139 ms 47.206 ms
16 209.85.247.5 (209.85.247.5) 48.244 ms 48.115 ms 48.330 ms
17 64.233.175.107 (64.233.175.107) 46.829 ms 216.239.41.73 (216.239.41.73) 46.857 ms
18 216.239.63.211 (216.239.63.211) 47.187 ms
19 216.239.58.139 (216.239.58.139) 47.853 ms 216.239.58.193 (216.239.58.193) 48.480 ms *
20 * * *
21 * * *
22 * * *
23 * * *
24 * * *
25 * * *
26 50.57.208.35.bc.googleusercontent.com (35.208.57.50) 63.659 ms 63.657 ms *
traceroute to 35.208.57.50 (35.208.57.50), 30 hops max, 60 byte packets
1 gateway (163.253.70.1) 0.372 ms 0.346 ms 0.336 ms
3 ae-0-1.bar2.KansasCity1.Level3.net (4.69.159.202) 11.407 ms 11.402 ms 11.397 ms
4 kanc-b1-link.telia.net (213.248.90.109) 11.478 ms 11.550 ms 11.448 ms
5 chi-b21-link.telia.net (213.155.130.176) 29.647 ms 29.734 ms 29.559 ms
6 google-ic-326155-chi-b21.c.telia.net (213.248.66.127) 29.260 ms 29.321 ms 29.323 ms
7 216.239.56.7 (216.239.56.7) 29.382 ms 216.239.51.189 (216.239.51.189) 30.270 ms
8 108.170.231.175 (108.170.231.175) 29.072 ms
9 108.170.243.197 (108.170.243.197) 29.638 ms 108.170.244.2 (108.170.244.2) 30.098 ms
10 216.239.63.33 (216.239.63.33) 30.031 ms 209.85.254.94 (209.85.254.94) 64.850 ms
11 209.85.250.35 (209.85.250.35) 29.585 ms
13 * * *
14 * * *
15 * * *
16 * * *
17 * * *
18 * * *
19 * * *
20 * * *
21 * * *
22 50.57.208.35. bc.googleusercontent.com (35.208.57.50) 39.714 ms * 39.859 ms

# # Premium IP
# [root@ps-40g-gridftp ~]# traceroute 35.239.24.178 # iowa google premium
traceroute to 35.239.24.178 (35.239.24.178), 30 hops max, 60 byte packets
1 35.239.24.178 (35.239.24.178) (35.239.24.178) 0.118 ms 0.142 ms 0.135 ms
2 67.58.48.38 (67.58.48.38) 0.299 ms 0.307 ms 0.351 ms
3 dc-sdg-agg4--ucsd-100ge.cenic.net (137.164.23.176) 0.424 ms 0.452 ms 0.466 ms
4 dc-tus-agg3--sdg-agg4-100ge-2.cenic.net (137.164.11.10) 2.094 ms 2.115 ms dc-tus-agg3--sdg-
agg4-100ge-2.cenic.net (137.164.11.10) 2.036 ms
5 dc-lax-agg6--tus-agg3-100ge-2.cenic.net (137.164.11.22) 3.190 ms dc-lax-agg6--tus-agg3--100ge-
2.cenic.net (137.164.11.24) 3.210 ms
6 72.14.222.56 (72.14.222.56) 2.651 ms 74.125.49.165 (74.125.49.165) 2.725 ms 72.14.222.56
(72.14.222.56) 2.753 ms
7 108.170.247.180 (108.170.247.180) 3.248 ms
9 108.170.230.131 (108.170.230.131) 3.416 ms
10 216.239.54.139 (216.239.54.139) 29.767 ms 209.85.143.221 (209.85.143.221) 28.995 ms
11 216.239.43.227 (216.239.43.227) 38.103 ms 216.239.43.220 (216.239.43.220) 38.676 ms
12 74.125.253.181 (74.125.253.181) 38.958 ms 216.239.58.137 (216.239.58.137) 38.900 ms
13 * * *
14 * * *
15 * * *
16 * * * 
17 * * * 
18 * * * 
19 * * * 
20 178.24.239.35.bc.googleusercontent.com (35.239.24.178) 38.935 ms 38.920 ms 38.922 ms 

[isfiligoi@osg.chic ~]$ traceroute 35.239.24.178 # iowa google premium 
traceroute to 35.239.24.178 (35.239.24.178), 30 hops max, 60 byte packets 
1 gateway (163.253.70.1) 0.410 ms 0.388 ms 0.376 ms 
2 UCAID.bar1.KansasCity1.Level3.net (4.53.32.126) 11.391 ms 11.458 ms 11.526 ms 
3 ae5-6.bar1.KansasCity1.Level3.net (4.53.32.125) 11.197 ms 11.228 ms 11.210 ms 
4 * * * 
5 72.14.242.34 (72.14.242.34) 35.082 ms 35.055 ms 35.157 ms 
74.125.37.20 (74.125.37.20) 35.824 ms 
7 108.170.240.146 (108.170.240.146) 35.571 ms 35.757 ms 36.382 ms 
8 216.239.63.207 (216.239.63.207) 36.468 ms 108.170.228.84 (108.170.228.84) 36.669 ms 
216.239.63.253 (216.239.63.253) 37.193 ms 
9 209.85.250.37 (209.85.250.37) 41.491 ms 209.85.250.47 (209.85.250.47) 41.980 ms 
209.85.250.37 (209.85.250.37) 41.087 ms 
10 72.14.233.238 (72.14.233.238) 41.711 ms 209.85.246.84 (209.85.246.84) 40.058 ms 
11 216.239.43.17 (216.239.43.17) 39.822 ms 216.239.63.211 (216.239.63.211) 39.626 ms 
216.239.41.99 (216.239.41.99) 40.229 ms 
12 209.85.243.7 (209.85.243.7) 40.534 ms 74.125.253.171 (74.125.253.171) 40.855 ms 
209.85.241.221 (209.85.241.221) 41.990 ms 
13 * * * 
14 * * * 
15 * * * 
16 * * * 
17 * * * 
18 * * * 
19 * * * 
20 178.24.239.35.bc.googleusercontent.com (35.239.24.178) 40.750 ms * * 

# from inside the instance out 
#
[root@instance-c2 isfiligoi]# traceroute osg.chic.nrp.internet2.edu
traceroute to osg.chic.nrp.internet2.edu (163.253.70.2), 30 hops max, 60 byte packets
1 osg.chic.nrp.internet2.edu (163.253.70.2) 36.074 ms
2 * * osg.chic.nrp.internet2.edu (163.253.70.2) 36.074 ms
[root@instance-c2 isfiligoi]# traceroute osg.kans.nrp.internet2.edu
traceroute to osg.kans.nrp.internet2.edu (163.253.71.2), 30 hops max, 60 byte packets
1 osg.kans.nrp.internet2.edu (163.253.71.2) 39.949 ms 39.485 ms *
[root@instance-c2 isfiligoi]# traceroute ps-40g-gridftp.calit2.optiputer.net
traceroute to ps-40g-gridftp.calit2.optiputer.net (67.58.50.66), 30 hops max, 60 byte packets
1 ps-40g-gridftp.calit2.optiputer.net (67.58.50.66) 63.449 ms !X 63.286 ms !X 63.305 ms !X
[root@instance-c2 isfiligoi]# we definitely get a different path for the premium external IP address.
We seem to enter the google network way faster this way.

Note: Chicago still routes through Kansas.

Igor

```
[72@ps-40g-gridftp ~]$ traceroute 35.245.179.214 # premium east4 IP
trace route to 35.245.179.214 (35.245.179.214), 30 hops max, 60 byte packets
1 3826-router.optiputer.net  (67.58.50.65)  0.124 ms 0.147 ms 0.141 ms
2 67.58.48.38 (67.58.48.38)  0.299 ms 0.310 ms 0.349 ms
3 dc-sdg-agg4--ucsd-100ge.cenic.net (137.164.23.176)  0.363 ms 0.376 ms 0.397 ms
4 dc-tus-agg3--sdg-agg4-100ge-2.cenic.net (137.164.11.10)  2.029 ms dc-tus-agg3--sdg-agg4-
100ge.cenic.net (137.164.11.8)  2.064 ms 2.091 ms
5 dc-lax-agg6--tus-agg3-100ge-1.cenic.net (137.164.11.22)  3.003 ms 3.019 ms dc-lax-agg6--
72.14.222.56 (72.14.222.56)  2.655 ms 74.125.49.165 (74.125.49.165)  2.768 ms 2.736 ms
8 216.239.54.139 (216.239.54.139)  28.473 ms 209.85.143.221 (209.85.143.221)  28.950 ms
9 216.239.54.139 (216.239.54.139)  28.467 ms
10 216.239.43.220 (216.239.43.220)  128.694 ms 128.700 ms 128.689 ms
11 209.85.143.102 (209.85.143.102)  47.365 ms 72.14.234.8 (72.14.234.8)  50.582 ms
12 209.85.250.9 (209.85.250.9)  62.448 ms 216.239.40.164 (216.239.40.164)  62.102 ms
13 216.239.35.163 (216.239.35.163)  62.633 ms 216.239.63.233 (216.239.63.233)  62.612 ms
14 216.239.49.196 (216.239.49.196)  63.173 ms
15 216.239.48.8 (216.239.48.8)  62.448 ms 216.239.49.196 (216.239.49.196)  63.173 ms
16 * * *
17 * * *
18 * * *
19 * * *
20 * * *
21 * * *
22 * * *
23 * 214.179.245.35. bc.googleusercontent.com (35.245.179.214)  63.014 ms 62.830 ms
```

```
[isfiligoi@osg ~]$ traceroute 35.245.179.214
trace route to 35.245.179.214 (35.245.179.214), 30 hops max, 60 byte packets
1 gateway (163.253.70.1) 0.328 ms 0.309 ms 0.325 ms
2 UCAID.bar1.KansasCity1.Level3.net (4.53.32.126) 11.478 ms 11.536 ms 11.563 ms
3 ae5-6.bar1.KansasCity1.Level3.net (4.53.32.125) 49.406 ms 49.401 ms 49.420 ms
4 * * *
5 72.14.222.56 (72.14.222.56)  2.655 ms 74.125.49.165 (74.125.49.165)  2.768 ms 2.736 ms
8 209.85.246.100 (209.85.246.100)  35.094 ms 108.170.225.174 (108.170.225.174)  37.052 ms
9 216.239.54.139 (216.239.54.139)  28.467 ms
10 216.239.43.220 (216.239.43.220)  128.694 ms 128.700 ms 128.689 ms
11 209.85.143.102 (209.85.143.102)  47.365 ms 72.14.234.8 (72.14.234.8)  50.582 ms
12 209.85.250.9 (209.85.250.9)  62.448 ms 216.239.40.164 (216.239.40.164)  62.102 ms
13 216.239.35.163 (216.239.35.163)  62.633 ms 216.239.63.233 (216.239.63.233)  62.612 ms
14 216.239.49.196 (216.239.49.196)  63.173 ms
15 216.239.48.8 (216.239.48.8)  62.448 ms 216.239.49.196 (216.239.49.196)  63.173 ms
16 * * *
17 * * *
18 * * *
19 * * *
20 * * *
21 * * *
22 * * *
23 * 214.179.245.35. bc.googleusercontent.com (35.245.179.214)  63.014 ms 62.830 ms
```
On May 10, 2019, at 1:05 PM, Sfiligoi, Igor <isfiligoi@sdsc.edu> wrote:

Hi All.

I have spun up a VM instance in Google’s us-east4 zone (N Virginia).

The traceroutes look a bit weird to me.

From UCSD, we go over Internet2:

```
[isfiligoi@ps ~]$ traceroute 35.212.116.3
traceroute to 35.212.116.3 (35.212.116.3), 30 hops max, 60 byte packets
  1 3826-router.optiputer.net (67.58.50.65) 0.124 ms 0.151 ms 0.144 ms
  2 67.58.48.38 (67.58.48.38) 0.312 ms 0.321 ms 0.361 ms
  3 dc-sdg-agg4--ucsd-100ge.cenic.net (137.164.23.176) 0.886 ms 0.916 ms 0.939 ms
  4 dc-tus-agg3--sdg-agg4-100ge.cenic.net (137.164.11.8) 2.314 ms 2.345 ms dc-tus-agg3--sdg-
     agg4-100ge-2.cenic.net (137.164.11.10) 2.268 ms
  5 dc-lax-agg6--tus-agg3-100ge-2.cenic.net (137.164.11.24) 3.183 ms 3.209 ms dc-lax-agg6--
     tusat-agg3-100ge-1.cenic.net (137.164.11.22) 3.246 ms
  6 dc-lax-agg8--lax-agg6-100ge-4.cenic.net (137.164.11.37) 4.766 ms 4.438 ms 31.942 ms
  7 et-1-0-2.0.rtsw.losa.net.internet2.edu (64.57.20.82) 2.859 ms 2.857 ms 2.859 ms
  8 ae-1.4079.rtsw.salt.net.internet2.edu (162.252.70.115) 15.444 ms 15.589 ms 15.459 ms
  9 ae-5.4079.rtsw.kans.net.internet2.edu (162.252.70.144) 35.245 ms 35.294 ms 35.395 ms
 10 ae-3.4079.rtsw.chic.net.internet2.edu (162.252.70.140) 46.440 ms 46.550 ms 46.550 ms
 11 et-1-1-5.4079.rtsw.ashb.net.internet2.edu (162.252.70.60) 62.892 ms 62.924 ms 63.043 ms
 12 lo-0.8.rtsw2.ashb.net.internet2.edu (64.57.29.131) 64.114 ms 64.012 ms 63.834 ms
 13 72.14.217.16 (72.14.217.16) 64.007 ms 64.034 ms 64.008 ms
 14 108.170.246.33 (108.170.246.33) 63.265 ms 108.170.240.97 (108.170.240.97) 63.668 ms
     108.170.238.169 (108.170.238.169) 62.319 ms 172.253.64.46 (172.253.64.46) 64.018 ms
 15 216.239.41.228 (216.239.41.228) 64.203 ms
 16 * * *
 17 * * *
 18 * * *
 19 * * *
 20 * * *
 21 * * *
 22 3.116.212.35.bc.googleusercontent.com (35.212.116.3) 65.072 ms * *
```

From Chicago, we go over Kansas???:

```
[isfiligoi@osg ~]$ traceroute 35.212.116.3
traceroute to 35.212.116.3 (35.212.116.3), 30 hops max, 60 byte packets
  1 gateway (163.253.70.1) 0.381 ms 0.361 ms 0.370 ms
  2 UCAID.bar1.KansasCity1.Level3.net (45.33.126.1) 11.442 ms 11.590 ms 11.498 ms
```
Unfortunately, getting a traceroute (or tracepath) from inside the VM is pretty much useless:

[root@instance-1 isfiligoi]# traceroute ps-40g-gridftp.calit2.optiputer.net
traceroute to ps-40g-gridftp.calit2.optiputer.net (67.58.50.66), 30 hops max, 60 byte packets
1 ps-40g-gridftp.calit2.optiputer.net (67.58.50.66) 64.515 ms !X 64.623 ms !X 64.464 ms !X
[root@instance-1 isfiligoi]# traceroute osg.chic.nrp.internet2.edu
traceroute to osg.chic.nrp.internet2.edu (163.253.70.2), 30 hops max, 60 byte packets
1 osg.chic.nrp.internet2.edu (163.253.70.2) 38.326 ms 38.281 ms 38.292 ms
[root@instance-1 isfiligoi]# traceroute osg.kans.nrp.internet2.edu
traceroute to osg.kans.nrp.internet2.edu (163.253.71.2), 30 hops max, 60 byte packets
1 osg.kans.nrp.internet2.edu (163.253.71.2) 37.686 ms * 37.515 ms
[root@instance-1 isfiligoi]# traceroute osg.newy32aoa.nrp.internet2.edu
traceroute to osg.newy32aoa.nrp.internet2.edu (163.253.72.2), 30 hops max, 60 byte packets
1 osg.newy32aoa.nrp.internet2.edu (163.253.72.2) 6.596 ms * 6.619 ms

Will try the central and west coasts soon, too.

Cheers,

Igor

PS: I was using the “standard external IP”. I may try the “premium” one, and see if I get any better results.

Yep, we definitely get a different path for the premium external IP address.
We seem to enter the google network way faster this way.

Note: Chicago still routes through Kansas.

Igor

[root@ps-40g-gridftp ~]# traceroute 35.245.179.214 # premium east4 IP
traceroute to 35.245.179.214 (35.245.179.214), 30 hops max, 60 byte packets
1 3826-router.optiputer.net (67.58.50.65) 0.124 ms 0.147 ms 0.141 ms
2 67.58.48.38 (67.58.48.38) 0.299 ms 0.310 ms 0.349 ms
3 dc-sdg-agg4--ucsd-100ge.cenic.net (137.164.23.176) 0.363 ms 0.376 ms 0.397 ms
traceroute to 35.245.179.214 (35.245.179.214), 30 hops max, 60 byte packets
1 gateway (163.253.70.1) 0.328 ms 0.309 ms 0.325 ms
2 UCAID.bar1.KansasCity1.Level3.net (4.53.32.126) 11.478 ms 11.536 ms 11.563 ms
3 ae5-6.bar1.KansasCity1.Level3.net (4.53.32.125) 49.406 ms 49.401 ms 49.420 ms
4 * * *
5 72.14.242.34 (72.14.242.34) 35.127 ms 35.060 ms 35.109 ms
6 209.85.246.100 (209.85.246.100) 35.094 ms 108.170.225.174 (108.170.225.174) 37.852 ms
7 108.170.229.24 (108.170.229.24) 34.966 ms
8 216.239.149.108 (216.239.149.108) 35.625 ms 72.14.234.226 (72.14.234.226) 36.026 ms
9 209.85.243.161 (209.85.243.161) 35.202 ms 108.170.228.84 (108.170.228.84) 36.470 ms
10 216.239.40.139 (216.239.40.139) 36.052 ms 209.85.249.44 (209.85.249.44) 34.560 ms
11 108.170.232.198 (108.170.232.198) 35.187 ms 216.239.63.253 (216.239.63.253) 36.365 ms
12 108.170.228.79 (108.170.228.79) 35.625 ms 209.85.243.161 (209.85.243.161) 35.202 ms
13 108.170.228.79 (108.170.228.79) 35.625 ms 216.239.63.253 (216.239.63.253) 35.635 ms
14 108.170.228.79 (108.170.228.79) 35.625 ms 209.85.243.161 (209.85.243.161) 35.202 ms
15 * * *
16 * * *
17 * * *
18 * * *
19 * * *
20 * * *
21 * 214.179.245.35. bc.googleusercontent.com (35.245.179.214) 63.014 ms 62.830 ms

[isfiligoi@osg ~]$ traceroute 35.245.179.214
traceroute to 35.245.179.214 (35.245.179.214), 30 hops max, 60 byte packets
1 gateway (163.253.70.1) 0.328 ms 0.309 ms 0.325 ms
2 UCAID.bar1.KansasCity1.Level3.net (4.53.32.126) 11.478 ms 11.536 ms 11.563 ms
3 ae5-6.bar1.KansasCity1.Level3.net (4.53.32.125) 49.406 ms 49.401 ms 49.420 ms
4 * * *
5 72.14.242.34 (72.14.242.34) 35.127 ms 35.060 ms 35.109 ms
6 209.85.246.100 (209.85.246.100) 35.094 ms 108.170.225.174 (108.170.225.174) 37.852 ms
7 108.170.229.24 (108.170.229.24) 34.966 ms
8 216.239.149.108 (216.239.149.108) 35.625 ms 72.14.234.226 (72.14.234.226) 36.026 ms
9 209.85.243.161 (209.85.243.161) 35.202 ms 108.170.228.84 (108.170.228.84) 36.470 ms
10 216.239.40.139 (216.239.40.139) 36.052 ms 209.85.249.44 (209.85.249.44) 34.560 ms
11 108.170.232.198 (108.170.232.198) 35.187 ms 216.239.63.253 (216.239.63.253) 36.365 ms
12 108.170.228.79 (108.170.228.79) 35.625 ms 209.85.243.161 (209.85.243.161) 35.202 ms
13 108.170.228.79 (108.170.228.79) 35.625 ms 216.239.63.253 (216.239.63.253) 35.635 ms
14 108.170.228.79 (108.170.228.79) 35.625 ms 209.85.243.161 (209.85.243.161) 35.202 ms
15 * * *
16 * * *
17 * * *
18 * * *
19 * * *
20 * * *
21 * 214.179.245.35. bc.googleusercontent.com (35.245.179.214) 63.014 ms 62.830 ms
Edgar Fajardo Hernandez

to jhmoon, John, Tom, Frank, Igor, Dmitry

Dear Jeonghoon Moon,

OSG in collaboration with LIGO are working on ways to improve data distribution to jobs in the Grid. We recently commissioned KISTI to receive LIGO pilots via OSG. Our next step is to setup a cache somewhere close.

In collaboration with PRP we had setup caches using kubernetes at three locations on the Internet 2 backbone and one at the University of Amsterdam. Dima made me aware of an existing node in prp dtn-gpu2.kreonet.net that we could use for this effect.

We are pretty much ready to deploy a pod for this effect there but since all access to LIGO data are authenticated we need host certificates on this machine. I wonder if you could requests some host x509 certificates (hopefully InCommon IGTF certs) for this node and place them on `/etc/grid-security/` dir on that node or in any partition on that node.

IF you want to see more about stashcache deployment you can check these slides:

https://docs.google.com/presentation/d/1D0Zsv1MC7PjZfj1qZQ5mXHrGqr2B2Qi0g3NrttAvo0/edit?ouid=102515549915421709848&usp=slides_home&ths=true

Thanks,

Edgar M Fajardo Hernandez
emfajard@ucsd.edu

Dima Mishin

to Frank, Larry, Igor, John, Dmitry, me

Is there a way we can actually measure the network flows for an application?
So how can we directly measure the bandwidth used, as well as the CPU, GPU, Memory?

Yes, we have elastiflow (https://github.com/robcowart/elastiflow) installed which does exactly that
Dima Mishin  

to Frank, Tom, Igor, Larry, John

I thus think that the disk space as done in nautilus makes no sense. It’s a centralized model just like a supercomputing center, and its worse than a supercomputing center because while being centralized, it penalizes itself with large latencies between servers by distributing them across the country.

I think what you want is a federation model. A model where anybody can come with their space and data and join a global namespace. Data is then moved around within the federation of compute clusters, and cached in the network backbone as required, and without the researchers having to care.

This is exactly what we're addressing with EdgeFS - a set of distributed regions with low latency having common namespaces and exchanging data/metadata on demand. We're making great progress on that.

Our ceph cluster shrunk to California and surrounding states, which is ok for large spinning drives speeds and brings us good capacity. Anything needing better access times will go to EdgeFS.

First region is already deployed and being tested, the common namespace and regions support both inside the k8s cluster and between k8s clusters (federated) is also in progress.

EdgeFS guys are currently working in our cluster towards our requirements.

John Graham  

to Dima, Tom, Frank, Igor, Larry, John

We assist the open-source providers in evolving their software to meet our needs. This has been a very successful collaboration among many projects.

Alex Feltus  

to prp-nrppilot-engagement@internet2.edu, Stephen, Ben

Hi Dana et al.

I see use cases on the agenda.
The Feltus/Ficklin labs have git repo where we are plugging in our K8s ready Nextflow workflows that I think have all been tested on PRP/TNRP:

https://github.com/SystemsGenetics/kube-runner

These workflows process raw RNA gene expression data (GEMMaker), detect condition-specific dependencies between all gene pairs (KINC), and finds gene group biomarkers using a deep learning framework (GeneOracle). #1/#2 leverage GPUs. We are adding more workflows as time permits.

Thank you for providing this valuable computational environment!

Alex

---

Alex Feltus

to Thomas, John, Dima, Igor, Frank, Alex

Hi all.

I have $30K and interest from Clemson CCIT to host a TNRP K8s cluster node in our data center. I kindly request guidance on these 3 specific action items:

1. Is there a node specification guide/template sheet that fits a $30K budget so I can buy the node?
2. Is there technical documentation on folding in this node into Nautilus so I can share with CU sysadmins?
3. Are there any plans for a formal scheduler/policy engine for Nautilus that deals with walltime?

I am very excited about helping to build TNRP!

Alex

---

From: Thomas DeFanti <tdefanti@ucsd.edu>
Sent: Wednesday, April 24, 2019 1:08 PM
To: prp-nrppilot-engagement@internet2.edu; Alex Feltus <ffeltus@clemson.edu>
Cc: John Graham <jjgraham@ucsd.edu>; Dima Mishin <dmishin@ucsd.edu>; Sfiligoi, Igor <isfiligoi@sdsc.edu>; Frank Wuerthwein <fkw888@gmail.com>
Subject: Re: [prp-nrppilot-engagement] Feltus TNRP Usage For NSF Reports and 2 Questions

Thanks for the text, Alex! I’m going to try to assemble quantitative information about your usage of Nautilus from our data and let you know how that goes.
Q1: We are expanding our pot luck 32-bit supercomputer weekly. 24 new GPUs (3 FIONA8s) just came on line at UC Irvine last week (total of 306 today), 16 more tomorrow (one FIONA8 will have a TB of RAM for certain jobs that need it). We invite more FIONA8s (or FIONA2s or FIONA4s) as directly part of Nautilus (which means John needs ILOM and ACLs need to be opened), or as federated entities (Dima is working on Admiralty, I believe, as a mechanism). We also appreciate storage additions (we're up to 1.7PB) for fast sharing of datasets. Chris Paolini is also making his 2PB BeeGFS storage available.

See https://grafana.nautilus.optiputer.net/d/KMsJWWPiz/cluster-usage?orgId=1 and https://grafana.nautilus.optiputer.net/d/r6loPJmz/ceph-cluster-new?orgId=1

Q2: Our TNRP staff funding is half devoted to OSG integration. Frank Wuerthwein notes the following for our upcoming PRP quarterly report:

UCI: Prof. A. Taffard’s group continues to use the PRP as envisioned in the PRP proposal. This means her students and post-docs submit workflows to process the official data of the ATLAS experiment to produce their private reduced format data. The latter data is transferred as part of the workflow to UCI, and stored there on the PRP FIONA for subsequent processing. This implies that we have interfaced PRP to OSG in such a way that Prof. Taffard’s group can take full advantage of the wider OSG resources from the PRP FIONA. This continues to be their modus operandi for all of their physics analysis results.

In addition, PRP also facilitates other technology exploration and benchmarking for the LHC. E.g. the entire LHC community today depends on the open source gridftp software. The Globus team has discontinued support for this software last year. OSG and EGI together are presently maintaining it for as long as necessary to decide on, and commission a replacement technology. This work is done within one of the WLCG working groups. There are presently two candidate protocols as replacements, xroot and http. The FIONAs from Internet2 and GPN in Chicago and Kansas City respectively are being used for some of this benchmarking. These caches were chosen because PRP’s Nautilus cluster provides a very convenient deployment platform to switch server architecture by simply switching containers. We thus take out two servers for a few hours from the OSG Data Federation for benchmarking tests, and return them to the production data federation after the tests are done. This is a minimal perturbation to the production system with maximal benefit to the R&D efforts towards the HL-LHC. This sort of dual use is made possible by the PRP Nautilus cluster’s K8s container orchestration. At this point, we have demonstrated 20-30 Gbps transfers between Chicago and Kansas City using http. More detailed testing in 1-N and N-M configurations of both xroot and http are forthcoming. The ultimate goal of these tests is to establish whether or not http has significant protocol overheads that would make its choice as default protocol for the HL-LHC era undesirable.

The FIONAs at UCD and UCSB are presently envisioned to be primarily submission hosts, facilitating interactive login, submissions of workflows etc. Fundamentally, the CMS group at UCSB does most of its work at the UCSD T2 center. This includes some Machine Learning projects by Nick Amin (UCSB) and Sam May (UCSD) that used GPUs across PRP, as well as TPUs in Google Cloud that were integrated into Nautilus to make them easily accessible.
On the technical front, Igor can better explain how OSG and PRP/Nautilus are interacting, so please engage him. We are currently supporting Nautilus use by IceCube (as the grafana charts show) and plan to support two LIGO projects as experiments so we can learn how to accommodate users on Nautilus. If you look at the grafana pages carefully, you’ll note that our GPU usage is significant but our CPUs are way underutilized. This is a result of us wanting to encourage GPU jobs (and the experience of grad students grabbing only CPUs by the 100s without using GPUs). We will learn more about how to balance GPU/CPU as we get other users (like you?) on.

I’m sure you have more questions, Alex. I do too!

..tom..

Dima says:
Admiralty is federation controller superior to Kubernetes Federation V2 in terms of functions and readiness status, but still lacks a number of needed features, at the same time providing the ones that will not be useful for us.

Admiralty provides a multi cluster scheduler, which is supposed to run separate from the federating cluster and which provides scheduling across them all. This brings up first set of issues:
1. The computing power requirements - the scheduler should be able to process all pods running in all clusters and watch all the nodes, which limits the scaling capabilities of those
2. The presence of a single "top" component is not desired in scientific community, and peer-to-peer relationships are preferred.

Both can be fixed with special tuning (ignoring the scheduler and directly addressing the jobs, and running scheduler in federated clusters, but will not align with admiralty paradigm). Instead we’re proposing users to choose the cluster they want to run in, since this in most cases will be known by users.

3. The admiralty scheduler makes decisions based on it’s limited understanding of destination cluster's available resources, which can be quite complex (in case of Nautilus). To fully understand the possibility of running specific pod in a specific cluster, the full functionality of kubernetes scheduler should be re-implemented, and even in that case race condition will make it not guaranteed => the admiralty scheduler is not useful at all.

4. Admiralty provides credentials delegation from federated clusters to the top cluster. Same functionality was already implemented in [https://gitlab.com/ucsd-prp/nrp-controller](https://gitlab.com/ucsd-prp/nrp-controller) developed by PRP, and NRP implementation is doing it in more desirable peer-to-peer way - directly cluster-to-cluster, which improves security and builds the trust network in a more flexible way.
5. Admiralty provides "proxy" pods, services, deployments, etc abstractions, which is the most useful for us part of it. But since that's the only part that's missing currently in NRP controller, it might be better to implement it in NRP one and not deal with all issues with admiralty.

On Wed, Apr 24, 2019 at 6:26 AM Alex Feltus <ffeltus@clemson.edu> wrote:
Hey Tom.

1. **Feltus lab TNRP Activities Statement for NSF Reports.** The Feltus lab at Clemson University is running deep learning oncogenomics workflows on the National/Pacific Research Platform Kubernetes (K8s) cluster. The PRP K8s is allowing us to scale up our analyses by moving large genomics datasets between FIONA nodes and then screening tens of thousands of genes on GPUs for tumor biomarker discovery. Since many of our workflows are controlled with the Nextflow workflow manager (https://www.nextflow.io/), we are working with the Nextflow team to enable GPU resource wrangling on K8s clusters which opens a new job control opportunity for TNRP.

2. **Q1.** I want to build a small K8s cluster at Clemson and link it to TNRP. Still in negotiations with CCIT, but are you guys into growing the fabric?

3. **Q2.** What’s up with OSG/TNRP? Are y’all planning to flock K8s jobs onto OSG? If so, where do I buy tickets?

Alex

---

**John Graham**

to Jeronimo, John, Dima, me

Jeronimo

Whats the specs on the hardware and network attachment ? Are you L3 routed ? We would be happy to add a node or two from AmLight.

Are you interested in the 100G P4 using Xilinx U200s ?
On Wed, Apr 24, 2019 at 10:55 AM Dima Mishin <dmishin@ucsd.edu> wrote:
Hi John!

He said, for now, the nodes are not very busy (until the LSST come on line). I can help him but what is the best way to proceed?

Use our install guide (https://docs.google.com/document/d/1-d8GnQYnUT5NYRgnsOiyoMcoeyYGHze67nDZcZZ_kpDVg/edit?usp=sharing), install the nodes, get token from me, join

Any significant storage in those nodes?

--
Best regards,
Dima Mishin

On Apr 23, 2019, at 12:51 PM, John Hicks <jhicks@internet2.edu> wrote:

Hi John and Dima,
I talked with Jeronimo at FIU (AmLight) the other day and he has two nodes (one in South America and on in Miami) that I thought would be interesting to add to the k8s cluster (he said he already talked to you all about this). He said, for now, the nodes are not very busy (until the LSST come on line). I can help him but what is the best way to proceed?

-John

____________________________

Sfiligoi, Igor

Wed, Apr 17, 1:45 PM
to Frank, me, Larry

FYI: My talk has been accepted for Condor week. Currently scheduled for Tue May 21st at 3:45pm.

Cheers,
Igor

____________________________

Laurie Robinson

Tue, Apr 16, 10:04 AM
to Thomas, UMASS-JGriffin, MGHPCC-JGoodhue

Good afternoon Tom (or morning when you get this):
Thank you again for participating as a presenter at the recently held NEREN Seminar on April 5, 2019. Your presentation was well received by those attending in person and remotely.

We greatly appreciate that you were able to take time out of your busy schedule to join us.

The email we sent out following the presentation to attendees is listed below along with the recorded webcast link that includes your presentation.
http://demo.mediasite.oshean.org/Mediasite/Play/ecb220d266fb4b338cf8f95a843b8e121d

Laurie

---

**Jim Kyriannis**

Mon, Apr 1,
2:06 PM
Reply to all

to Tom, Jen, Dana, James, Camille, Cathy, John, Stephen, Jim

Good afternoon,

In this next phase of our participation in the National Research Platform (NRP) Pilot, we're inviting member institutions with well-suited research projects to participate. This year, we can accommodate up to six participants in total: three from large institutions and three from mid-sized/smaller institutions.

Please find some additional background information below. If you are interested in learning more about Pilot participation, please let me know. Our target date for finalizing our list of participants is end of April.

NYSERNet is participating in a pilot to grow a distributed compute environment built by University of California researchers, called The Pacific Research Platform (PRP - https://ucsd-prp.gitlab.io), into a national testbed: The National Research Platform. The technology consists of Data Transfer Nodes (DTNs), Compute Nodes, and a distributed Ceph storage environment acting in concert to enable inter-institutional high speed research data transfers on a federated computing platform offering Intel and GPU processing capacity. Not designed to compete with multi-million core Supercomputers, this is an environment based on federated access into Kubernetes Linux containers providing a temporary virtual compute and storage environment, suitable for smaller research workloads.

Thus far, NYSERNet has deployed the DTN component of the NRP Pilot in both our
Syracuse Data Center and at Stony Brook University. We're collaborating with Stony Brook on implementing a Science DMZ network and supporting research needs through the Pilot, such as improving data transfers to XSEDE via the DTN. We'll soon be exploring NRP compute possibilities as well. We've also begun collaborating with the American Museum of Natural History on an NSF-funded project to design a Science DMZ which supports NRP Pilot activities, including facilitating an Astrophysicist's data transfers with the Palomar Observatory and providing a compute platform.

We would like to invite additional institutions and their researchers to participate in the NRP Pilot as well. Due to the unique nature of its technology, not all research projects are well-suited for the Pilot. Those which are a good match for the Pilot will have need for collaborations and data transfers which rely upon the NYSERNet network to reach Internet2 and can leverage a virtual Linux compute environment. Projects requiring GPU support or discipline-specific software are welcome.

Thank you,

Jim

Jim Kyriannis  
CHIEF TECHNOLOGY OFFICER  
NYSERNet  
100 S. Salina St. Suite 300  
Syracuse, NY 13202  
(315) 413-0345 ext. 5421  
jkyriannis@nysernet.org  
nysernet.org

---

Edgar Fajardo Hernandez  
to Frank, Tom, Brian, Igor

Hi Frank,

I have been in constant communication with Dima about this he said he has never seen more than 20Gbps from inside a container. I mentioned this to John Hicks cause I would imagine the first thing to check is those nodes can do 100Gbps to their gateways.
On Mar 21, 2019, at 4:47 PM, Frank Wuerthwein wrote:

Hi Edgar,

Please report on this also to PRP. I cc-ed Tom Defanti. Maybe John Graham or Dima have any ideas of what's going on.

I believe they succeeded in filling the full 100Gbps pipe at various times between certain hosts.

Thanks, frank

On Mar 18, 2019, at 2:48 PM, Edgar Fajardo Hernandez <emfajardohernandez@physics.ucsd.edu> wrote:

Hi Brian, ALl,

I have been working on getting the 100Gbps tests setup. I started by making sure I can lperf from one to the other hosts that have 100Gbps capabilities. Right now I tried between three hosts:

1. Kansas City I2
2. UCSD
3. Chicago I2

All of them have 100Gbps network cards. But it doesn't matter how I slice it the most I can get is 28Gbps even with multiple parallel streams. I am trying to involve the network folks and the I2 folks. For now I will setup the TPC tests and in parallel work with John Hicks if he can at least test in the bare metal that those two hosts in I2 can talk at least 50 Gbps. I am not sure if this is part of what the I2 folks want to talk in the meeting of March 29th.

Edgar M Fajardo Hernandez
emfajard@ucsd.edu
About two hours ago Dima reclaimed the suncave nodes, because they are needed for a local demo. As expected, OSG workload was removed virtually immediately, so PRP did not suffer at all for letting OSG jobs run when resources were not in use.

Cheers,
Igor

---

Jim Kyriannis

Wed, Mar 13, 3:06 PM
Reply to all
to Tom, Larry, Stephen, Jim

Larry and Tom,

Thank you for taking the time to speak with us yesterday and help us better understand the vision for the PRP and TNRP projects going forward. As a follow-up to our conversation, below please find the list of NYSERNet institutions which we'll be engaging with for participation in the NRP Pilot. We'd like to take you up on your offer for assistance with identifying any activity these institutions may be incurring at UCSD, TACC, and OSG. (As we've not yet had discussions with many of these institutions, we ask that this list be kept confidential.)

Again, we look forward to continuing the collaboration, and please let us know if NYSERNet can be of additional assistance as the project evolves.

Thanks again,

Jim & Steve

---

NYSERNet NRP Pilot Candidates
Albert Einstein College of Medicine
Binghamton University
City University of New York (CUNY)
Clarkson University
Columbia University
Cornell University
Hofstra University
Icahn School of Medicine
Le Moyne College
Marist College
New York Genome Center
New York Structural Biology Ctr.
All,

What an amazing series of conversations in La Jolla! It was beneficial to hear…

Today, the NRP Communications Committee reviewed the, “One-Pager” document Melissa Lucas (CENIC) and Jen Leasure (The Quilt) have been working on, specifically the call to action element.

This document is meant to be a high level overview for the PRP effort that can be shared with our community. It will also be used as a base for the continued communication efforts for the NRP.

We would appreciate any feedback by Wednesday, February 27th… CENIC and Internet2 have upcoming meetings and having this on hand would be helpful.
Thank you!

John Graham

to Akbar, John, Amy, me

Akbar

We have several years of FIONA builds listed on our website. http://ucsd-prp.gitlab.io/resources/

The Classic 3U 40Gbps FIONA DTN is in here https://docs.google.com/spreadsheets/d/1wTdLzf1xLui1Dj0Gh_dGjfxxI5CWIrjvxggdR3Btfcg/edit#gid=1062495698

We have upgraded most of these to hold 16x 12TB SAS3 drives and 2x M.2 NVMe for metadata docker and kubelet. There are 1G 10G 40G and 100G versions we have tested. I am leaning heavily towards Epyc systems these days.

Happy to answer more questions.

John

On Wed, Feb 13, 2019 at 5:41 PM Akbar Kara <akbar.kara@tx-learn.net> wrote:

Hi John and John,

LEARN is looking at building a 40G DTN, do you have any pointers to server SKU or BOMs that we can look at? Where do you source them from?

Best,

/ak

Akbar Kara, CTO | +1 214-392-2717
LEARN: Lonestar Education & Research Network
John Graham <jjgraham@eng.ucsd.edu>  
to Patrick, jjgraham@ucsd.edu, John, Dima, me, Larry

TACC is on the MaDDash  

clean performance. we should see tests appearing in the dashboard soon.

[root@ps-40g-gridftp ~]# /usr/local/bin/timeout.sh -t 2000 globus-url-copy -vb -fast -p 16 
ftp://fiona01.tacc.utexas.edu:2811/export/data/10G.out file:///export/data/10G.out
Source: ftp://fiona01.tacc.utexas.edu:2811/export/data/
Dest:  file:///export/data/
10G.out
  8799649792 bytes   1049.00 MB/sec avg   1185.95 MB/sec inst From: Larry Smarr <lsmarr@ucsd.edu>

Dan and Pankaj-UTexas and therefore LEARN is on the MaDDash!

Thanks to Patrick Storm and his colleagues for working with PRP’s John Graham to create this milestone. With this anchor DTN at TACC, we can begin to expand into UTAustin and LEARN more generally.

As you know, the UTA Dell Medical Center is our next target. We will be having a meeting tomorrow at Calit2 with some technical folks from UTA to work on this.

Things are moving!
FYI…. Thought we ought to ask before we do ….

If you have other hosts to suggest, let us know.

E.g. what about the one at UC Santa Cruz? I believe there is a Fiona at 100Gbps there, isn’t there?

Begin forwarded message:

From: Frank Wuerthwein <fkw888@gmail.com>
Subject: Using some of the storage at the POPs for a scalability test
Date: February 8, 2019 at 3:01:43 PM PST
To: Howard Pfeffer <hpfeffer@internet2.edu>, John Hicks <jhicks@internet2.edu>,
"Fajardo Hernandez, Edgar"
<emfajard@ucsd.edu>, jed@greatplains.net, dbrunson@internet2.edu, "Sfiligoi, Igor"
<jisfiligoi@sdsc.edu>, Ana Hunsinger <ana@internet2.edu>

Dear Howard and James,

The global LHC community, and probably with it many others, is going through a gridftp replacement R&D phase.
We have storage system artifacts now that speak HTTPS, and want to do some scalability testing.

We have one host in San Diego and are looking for a second host that has 100Gbps or something close to it connectivity.

Would you be willing to have us use one of the systems, either in Chicago or Kansas City for the testing?
(Edgar said Chicago has 100Gbps but I don’t know what Kansas City has)

Thanks, frank

P.s. Here’s the testing in a nutshell. More details will be written up later.
All the testing is about third party transfer capabilities with HTTPS.

Ascertain protocol overheads
We deploy an Xrootd origin in San Diego, and a fast https server (forgot the product name) on the other host that writes data to /dev/null
The xrootd initiates the third party transfers. Here we have enough NVME in the box to push IO out from it.
My personal goal would be to show few 10’s of Gbps via third party HTTPS from an xrootd servers talking HTTPS to some other HTTPS receiver. At UCSD I would probably not want to go above 60Gbps outgoing to not affect the production networking unduly. We have done tests at that level before.

We’d be doing the tests “behind your back” so to speak, by just switching containers on K8s when we are ready, and switching back after we are ready. The tests would be fairly brief. Details to be worked out.

John Graham  
Fri, Feb 8, 3:44 PM  
to Frank, Larry, Tom, John, Dmitry, Edgar  

There are 9x 100G connected nodes in Nautilus  
https://wiki.nautilus.optiputer.net/wiki/Cluster_Map

You can deploy jobs that use a redis queue to feed workers on several of the nodes and write to /dev/null

We are using Aria2 for pulling HTTPS data from our THREDDS server  
https://github.com/aria2/aria2
We saw ~3GBps writing to the rook storage from a single 40Gbps THREDDS server.

You need to be logged in to view this repo.  

This shows the CPU and RAM usage during a test run  
The run on the left is 10 worker nodes the one on the right is 20 worker nodes

Robert Gardner  
Feb 1, 2019, 3:06 PM  
Repl  
to John, Dima, Thomas, Edgar, Frank, isfiligoi@ucsd.edu, Brian, Tim, Susan, Z y to alak, Jeff, Brian, Tom, Eric, John, Doug, Alessandra, James, Kevin, Oliver, Christopher, Lincoln, Shawn, Joe
Dear Colleagues,

We would like to share with you some ideas we've been developing for secure containerized application deployment on edge clusters in a federated k8s context. Could you have a look at http://bit.ly/app-sec-edge and provide us with any commentary and/or feedback? You can comment on the Google doc directly. Our hope is that we could iterate on this and come to a trusted model that moves our research platform communities forward. Your help is greatly appreciated!

Regards,

Chris*, Lincoln, Joe, Shawn, Rob
— for the SLATE (http://slateci.io) team

*corresponding author

---

Dima Mishin

Feb 4, 2019, 9:47 AM

Hi Rob,

I like the idea of trusted repo, but I don't like the idea of using helm for deploying apps. In my experience it's working ok for google and amazon clouds, but most charts are not working for our bare metal cluster and our policies. Also helm charts usually need same or even bigger amount of configuring compared to simply deploying yaml files.

Version 3 of helm was supposed to address many issues we're currently hitting, I'm not sure if that will happen and when. Until then I (successfully) try to use as little of Helm as possible.

My view is that operators are the future of complex applications deploy, also kubeless/serverless is on the rise, and I wouldn't count on Helm.

Could we instead limit the scope on certain trusted containers repo?

--

Best regards,
Dima Mishin
Susan E. Sons

to Mike, Robert, John, Dima, Thomas, Edgar, Frank, isfiligoi@ucsd.edu, Brian, Tim, Zalak, Jeff, Brian, Tom, Eric, John, Doug, Alessandra, James, Kevin, Oliver, Christopher, Lincoln, Shawn, Joe

Adding Mike Stanfield from my office to the list...he'll have a chance to go through this before I will.

Susan

John Graham

to Susan, Robert, John, Dima, Thomas, Edgar, Frank, isfiligoi@ucsd.edu, Brian, Tim, Zalak, Jeff, Brian, Tom, Eric, John, Doug, Alessandra, James, Kevin, Oliver, Christopher, Lincoln, Shawn, Joe, Mike

Another bit not to overlook is curated yum, apt, pip ... repos.
Having a custom tuned kernel running on all nodes can also improve performance and reliability.

Robert Gardner

to John, Susan, Dima, Thomas, Edgar, Frank, isfiligoi@ucsd.edu, Brian, Tim, Zalak, Jeff, Brian, Tom, Eric, John, Doug, Alessandra, James, Kevin, Oliver, Christopher, Lincoln, Shawn, Joe, Mike

Thanks John,

Dima, this is an excellent point, and the extent we can make the document technology-agnostic the longer it will be of value. We'll discuss generalizing since obviously the k8s ecosystem will continue to move rapidly beneath us.

Cheers,
- Rob

Lincoln Bryant

to Robert, John, Susan, Dima, Thomas, Edgar, Frank, isfiligoi@ucsd.edu, Brian, Tim, Zalak, Jeff, Brian, Tom, Eric, John, Doug, Alessandra, James, Kevin, Oliver, Christopher, Shawn, Joe, Mike
Hi Dima,

Thank you for the feedback.

Indeed, we should not be specifically tied to Helm and I think it's good to clarify that.

In our use case today, we use our own self-contained Helm setup that is deployed in the SLATE namespace via your NRP controller. The purpose is two-fold: First, to be a lightweight templating option for deployments across multiple clusters. Specifically, we're targeting bare metal clusters that meet (as of yet not formally specified) standard requirements, and we have created our own limited-scope catalog of charts for that purpose. We have also struggled with the charts provided by the Helm maintainers, and I also think sufficiently complex applications ought to use Kubernetes YAML files directly (since that is its purpose) or, as you point out, develop Operators. However I think that Operators will require upstream commitment from application developers that we shouldn't assume will be there.

The second, more important purpose for us, is that we use Helm to try to provide a reasonably secure boundary between the Kubernetes objects that are specified by application developers, which (in our use case) must be explicitly approved by the curators, and the run-time configuration options that can be freely changed. For example, for OSG's StashCache we want the Docker Image specification to be guarded by our approval process, but we don't care what an individual application administrator chooses for high and low water marks of the cache itself.

Happy to discuss more at length!

Cheers,
Lincoln

---

Dima Mishin
Mon, Feb 4, 4:17 PM
Repl

to Lincoln, Robert, John, Susan, Tom, Edgar, Frank, isfiligoi@ucsd.edu, Brian, Tim, Zalak, Jeff, Brian, Tom, Eric, John, Doug, Alessandra, James, Kevin, Oliver, Christopher, Shawn, Joe, Mike

I'd say Kubernetes provides super-rich capabilities for specifying what containers can do (RBAC and PSP), and whatever is not covered by those can be set by custom admission controllers - that's what I do in nautilus to better handle resources sharing
between groups. So there's no real need to restrict the yamls - I'd rather use standard options. Unless there's something you know that I didn't think about :)

Kevin Lannon  
Mon, Feb 4, 4:41 PM

Repl y to

all

to Kenyi, John, Dima, Thomas, Edgar, Frank, isfiligoi@ucsd.edu, Brian, Tim, Susan, Zalak, Jeff, Brian, Tom, Eric, John, Doug, Alessandra, James, Oliver, Robert, Christopher, Lincoln, Shawn, Joe

Hello Rob,

I haven’t had a chance to look through this carefully, nor is it clear to me that I could personally provide much useful feedback, so I shared it with Kenyi Hurtado here at ND (added to the e-mail list) and here’s what he said:

I like the model. One thing that I didn't explicitly see in the document was describing the authentication method between the kubernetes edge cluster and the platform (although they mentioned InCommon and Globus auth). I think in practice, the platform talks directly to the kubernetes cluster via HTTPS/TLS and requires at least one TCP port open in the kubernetes head node on the edge cluster side. Since all the platform does is talk to Kubernetes / Helm, firewall configuration to make sure the an external interface in the node can be contacted from the outside runs on the application administrator / edge cluster system administrator, as well as patching docker on all machines in the Kubernetes cluster for security patches regularly. It looks like maintaining the docker images for the different applications is up to the application administrator too, which I think is good in the sense that this doesn’t look like a blackbox from that perspective, but requires learning the dockerfile language.

Basically, you own and maintain your kubernetes cluster and the platform act as a user of such cluster with minimal privileges, and users can only get access to curated applications. This is easier to handle than having some federated-like infrastructure with full control of the cluster via kubectl in terms of security.

The workflow to get an application approved seems clear. Since there is a platform reviewer per application, I guess manpower could become an issue if this becomes really popular, but I guess that’s the kind of problem the team would like to have.

So, bottom line, it looks good to us and we’re looking forward to seeing more.

Sincerely,
Kevin

Oliver Gutsche  
Tue, Feb 5, 1:00 PM

Repl y to

all

to Anthony, David, Kevin, John, Dima, Thomas, Edgar, Frank, isfiligoi@ucsd.edu, Brian, Tim, Susan, Zalak, Jeffrey, Brian, Tom, Eric, John, Doug, Alessandra, James, Robert, Kenyi, Christopher, Lincoln, Shawn, Joe
Hi all,

I would like to add two FNAL people who think a lot about containers at FNAL and work on it actively. Tony Tiradani and Dave Mason. Tony is adding comments to the google doc as we speak.

Thanks,

OLI

---

**Christopher Weaver**

to Mike, Dima, Lincoln, Robert, John, Susan, Tom, Edgar, Frank, isfiligoi@ucsd.edu, Brian, Tim, Zalak, Jeff, Brian, Tom, Eric, John, Doug, Alessandra, James, Kevin, Oliver, Shawn, Joe

Hi Dima,

I just wanted to chime in with a few more ideas for the discussion.

Related to a desire not to be overly tied to Helm, we are considering user workflows that are not specifically tied to Kubernetes either, although of course that is the technology to use for the foreseeable future. We are implementing specific Kubernetes capabilities as needed while maintaining a robust high level architecture which works generally.

The focus of the document here is to examine security and policy layers in an environment of independent edge clusters choosing to participate in federated platforms supporting multi-institution, collaborative science. The key question we've grappled with is managing privilege such that (diverse) local policies are respected. For that reason (large numbers of diverse organizations) we've introduced a layer above native K8s to maintain policy and access rules to more easily identify and enforce the trust relationships. Of course as you've pointed out all of this can be implemented cluster-by-cluster and user-by-user via native K8s policy concepts and controllers. In federations or sub-federations where this can be efficiently managed, this gives the most power and flexibility to users (and in natural fashion for those already familiar with kubectl). For some HPC providers, including US leadership facilities or European sites, this level of integration may not be practical, though it is early days and perhaps we're too pessimistic. Regardless we don't see an issue doing both, and there may be differing classes of access-capabilities afforded to different user community-resource trust pairings. In one sense Nautilus and Slate are developing the platform model from both ends of the privilege spectrum.

Chris Weaver

---

From: Dima Mishin <dmishin@ucsd.edu>
I'd say kubernetes provides super-rich capabilities for specifying what containers can do (RBAC and PSP), and whatever is not covered by those can be set by custom admission controllers - that's what I do in nautilus to better handle resources sharing between groups. So there's no real need to restrict the yamls - I'd rather use standard options. Unless there's something you know that I didn't think about :)

--
Best regards,
Dima Mishin

On Feb 4, 2019, at 12:15 PM, John Graham <jjgraham@eng.ucsd.edu> wrote:

Another bit not to overlook is curated yum, apt, pip ... repos. Having a custom tuned kernel running on all nodes can also improve performance and reliability.

On Mon, Feb 4, 2019 at 10:10 AM Susan E. Sons <sesons@iu.edu> wrote:

Adding Mike Stanfield from my office to the list...he'll have a chance to go through this before I will.
Susan

On 2/1/19 6:06 PM, Robert Gardner wrote:

Dear Colleagues,

We would like to share with you some ideas we’ve been developing for secure containerized application deployment on edge clusters in a federated k8s context. Could you have a look at http://bit.ly/app-sec-edge and provide us with any commentary and/or feedback? You can comment on the Google doc directly. Our hope is that we could iterate on this and come to a trusted model that moves our research platform communities forward. Your help is greatly appreciated!
Regards,

Chris*, Lincoln, Joe, Shawn, Rob
— for the SLATE (http://slateci.io) team

*corresponding author

--

Susan Sons
Chief Security Analyst
IU Center for Applied Cybersecurity Research
sesons@iu.edu<mailto:sesons@iu.edu>

...
deploys (https://gitlab.cern.ch/explore/groups), and I'd be careful counting on a single such repo.

For some HPC providers, including US leadership facilities or European sites, this level of integration may not be practical, though it is early days and perhaps we're too pessimistic. Regardless we don't see an issue doing both, and there may be differing classes of access-capabilities afforded to different user community-resource trust pairings.

We could provide a recommended policy (and its implementation, in form of code or config files) that would be easy to modify if needed on site. Adoption of it should not be tied to the ability to join the federation.

---

Christopher Weaver
Feb 8, 2019, 2:08 PM

Hi Dima,

Responses inline below, but if we want to keep discussing in greater detail, feel free to start a new thread with us so we don't have to spam everyone else on this list.

> Reading the document from security and policy perspective (as opposite to technical implementation part), it sounds like there will be some team making decisions on inclusion of packages in the repo, and all members of the federation are expected to agree to those decisions. Am I reading it correctly?

Yes, the review would be performed by some subset of the team operating the platform/federation. This is intended to be a baseline check for suitability which would meet the concerns of most of the resource providers; those who have substantially more strict concerns could use an additional level of whitelisting within SLATE to authorize particular groups to run particular applications.

> The goal of "serve a valid science purpose and not contain security vulnerabilities" can be accomplished by utilizing one of container vulnerability scanning software, implemented by cluster policies (e.g. gitlab containers repo with automates scan as one of many examples), and list of approved apps/containers can vary significantly between clusters. A federation of such repos curated by multiple clusters might build a good platform for it, fe. cern already has gitlab and I'd suppose they're using it for their cluster deploys (https://gitlab.cern.ch/explore/groups), and I'd be careful counting on a single such repo.

Such automated scanning is exactly one of the components we had in mind for the
review step of the process. You are certainly correct that differing cluster policies are a concern; this is where we have to balance the usefulness of the platform (application administrators being able to easily deploy at many sites) against the security requirements of the sites. The approach we have been taking is to try to establish a baseline of security through the curated catalog which will satisfy many/most sites and then allow sites with stricter needs to apply them optionally, as mentioned above. Our current approach is focused on how to make one repository which contains applications which can have a baseline level of trust, but making more than one repository, possibly with different trust levels, available within the platform is something we can explore.

> We could provide a recommended policy (and its implementation, in form of code or config files) that would be easy to modify if needed on site. Adoption of it should not be tied to the ability to join the federation.

That would be a possible approach. I think it's worth noting that the single, curated catalog serves part of this anyway, as its contents are public and can be used by anyone who considers standards of the curation suitably aligned with their own wishes. It's true that this doesn't provide the component of generally restricting installs on that cluster to that set of applications; our thinking had been that cluster admins who want such a restriction can get it by joining the federation and having users access the cluster under the restrictions of the SLATE API, while they would still have the freedom to use the cluster unrestricted themselves, or provide unrestricted access to their local users. This is an important part of why we don't want membership in a federation to be exclusive, i.e. to inhibit other ways of using the cluster.

Thanks again for your feedback,
Chris Weaver

From: Dmitry Mishin <dmishin@ucsd.edu>
Sent: Wednesday, February 6, 2019 4:11 PM
To: Christopher Weaver
Cc: Lincoln Bryant; Robert Gardner; John Graham; Susan E. Sons; Tom DeFanti; Edgar Fajardo Hernandez; Frank Wuerthwein; isfiligoi@ucsd.edu; Brian Lin; Tim Cartwright; Shah, Zalak Sunilkumar; Jeff Dost; Brian Bockelman; Tom Barton; Eric Lancon; John R. Hover; Doug Benjamin; Alessandra Forti; James Von Oehsen; Kevin Lannon; Oliver Gutsche; Shawn McKee; Joe Breen; Mike Stanfield
Subject: Re: Application Security for the Edge

The focus of the document here is to examine security and policy layers in an environment of independent edge clusters choosing to participate in federated platforms supporting multi-institution, collaborative science. The key question we've grappled with is managing privilege such that (diverse) local policies are respected.

Reading the document from security and policy perspective (as opposite to technical implementation part), it sounds like there will be some team making decisions on
inclusion of packages in the repo, and all members of the federation are expected to agree to those decisions. Am I reading it correctly?

The goal of "serve a valid science purpose and not contain security vulnerabilities" can be accomplished by utilizing one of container vulnerability scanning software, implemented by cluster policies (e.g. gitlab containers repo with automates scan as one of many examples), and list of approved apps/containers can vary significantly between clusters. A federation of such repos curated by multiple clusters might build a good platform for it, fe. cern already has gitlab and I'd suppose they're using it for their cluster deploys (https://gitlab.cern.ch/explore/groups), and I'd be careful counting on a single such repo.

For some HPC providers, including US leadership facilities or European sites, this level of integration may not be practical, though it is early days and perhaps we're too pessimistic. Regardless we don't see an issue doing both, and there may be differing classes of access-capabilities afforded to different user community-resource trust pairings.

We could provide a recommended policy (and its implementation, in form of code or config files) that would be easy to modify if needed on site. Adoption of it should not be tied to the ability to join the federation.

--
Best regards,
Dima Mishin

On Feb 6, 2019, at 8:35 AM, Christopher Weaver
<cnweaver@uchicago.edu> wrote:

Hi Dima,

I just wanted to chime in with a few more ideas for the discussion.

Related to a desire not to be overly tied to Helm, we are considering user workflows that are not specifically tied to Kubernetes either, although of course that is the technology to use for the foreseeable future. We are implementing specific Kubernetes capabilities as needed while maintaining a robust high level architecture which works generally.

The focus of the document here is to examine security and policy layers in an environment of independent edge clusters choosing to participate in federated platforms supporting multi-institution, collaborative science. The key question we've grappled with is managing privilege such that (diverse) local policies are respected. For that reason (large numbers of diverse organizations) we've introduced a layer above native K8s to maintain policy and access rules to more easily identify and enforce the trust relationships. Of course as you've pointed out all of this can be implemented cluster-by-cluster and user-by-user via native K8s policy concepts and controllers. In federations
or sub-federations where this can be efficiently managed, this gives the most power and flexibility to users (and in natural fashion for those already familiar with kubectl). For some HPC providers, including US leadership facilities or European sites, this level of integration may not be practical, though it is early days and perhaps we're too pessimistic. Regardless we don't see an issue doing both, and there may be differing classes of access-capabilities afforded to different user community-resource trust pairings. In one sense Nautilus and Slate are developing the platform model from both ends of the privilege spectrum.

Chris Weaver

From: Dima Mishin <dmishin@ucsd.edu> Sent: Monday, February 4, 2019 6:17 PM
To: Lincoln Bryant
Cc: Robert Gardner; John Graham; Susan E. Sons; Tom DeFanti; Edgar Fajardo Hernandez; Frank Wuerthwein; isfiligoi@ucsd.edu; Brian Lin; Tim Cartwright; Shah, Zalak Sunilkumar; Jeff Dost; Brian Bockelman; Tom Barton; Eric Lancon; John R. Hover; Doug Benjamin; Alessandra Forti; James Von Oehsen; Kevin Lannon; Oliver Gutsche; Christopher Weaver; Shawn McKee; Joe Breen; Mike Stanfield
Subject: Re: Application Security for the Edge

I’d say kubernetes provides super-rich capabilities for specifying what containers can do (RBAC and PSP), and whatever is not covered by those can be set by custom admission controllers - that's what I do in nautilus to better handle resources sharing between groups. So there's no real need to restrict the yamls - I'd rather use standard options. Unless there's something you know that I didn't think about :) 

--
Best regards,
Dima Mishin

On Feb 4, 2019, at 12:15 PM, John Graham <jjgraham@eng.ucsd.edu> wrote:

Another bit not to overlook is curated yum, apt, pip ... repos. Having a custom tuned kernel running on all nodes can also improve performance and reliability.

On Mon, Feb 4, 2019 at 10:10 AM Susan E. Sons <sesons@iu.edu> wrote:
Adding Mike Stanfield from my office to the list...he'll have a chance to go through this before I will.
Susan

On 2/1/19 6:06 PM, Robert Gardner wrote:
Dear Colleagues,

We would like to share with you some ideas we’ve been developing for secure containerized application deployment on edge clusters in a federated k8s context. Could you have a look at http://bit.ly/app-sec-edge and provide us with any commentary and/or feedback? You can comment on the Google doc directly. Our hope is that we could iterate on this and come to a trusted model that moves our research platform communities forward. Your help is greatly appreciated!

Regards,

Chris*, Lincoln, Joe, Shawn, Rob
— for the SLATE (http://slateci.io) team

*corresponding author

--

Susan Sons
Chief Security Analyst
IU Center for Applied Cybersecurity Research
sesons@iu.edu

On Wed, Jan 23, 2019 at 8:54 AM Robert Gardner <rwg@uchicago.edu> wrote:
Hi Tom

We have a number of potential joint technical issues and/or projects to discuss with you guys. Three areas —
- automated provisioning for new sites
- federation roadmap — interoperability (prp-way, slate-way)
- application management on the edge (curation workflow, security)

Perhaps we can have a phone meeting in the near term. We’ll chat and get back to you.

- Rob
Nathaniel, Tom and Akbar:

As you are aware, CENIC, OneNet and LEARN have been funded by NSF to connect our states using PacWave, which helps us create the NRP platform and explore projects. Good news is that we now have physical connectivity established and would like to begin with the next phase. Talking to the California team just now, a logical next step will be to install a Fiona box at TACC/LEARN in early 2019. I would like to request all three of you (and your team) to get together on a technical call in early January and discuss feasibility and logistics to get the Fiona installation underway.

Please let us know if you need anything else from the leadership in order to move forward. LEARN Program Manager Amy Santana is cc here as well, and she can help coordinate calls and follow up as necessary.

Wish you'll happy holidays and be well. See you all in 2019.

Thanks.

Regards,
Pankaj

Pankaj Shah
President and CEO
LEARN: Lonestar Education and Research Network
Mobile: 614-354-9309
E-mail: pankaj.shah@tx-learn.net

Hi Ben and Doug

Let me introduce you to the prp/osg crowd. Dima & John can help you sort out your networking issues. Edgar can help you get IPAC data onto the OSG Data Federation, assuming that this is still the plan.

Edgar can also point you to an appropriate person to get an account on a submit host, though I recall that you Ben went to the osg summer school, didn’t you? So you have an account all set up, don’t you?

Personally, I suggest to go the K8s route and let us run higher level services for you first while you figure out any necessary network improvements.
Tom runs a weekly meeting w network engineers where you can bring up any network performance issues and ask for help from wan engineers to get them resolved.

Let me know if there’s anything else you need. And if we should have a phone call.

Thanks frank

Sent from my iPhone

> On Dec 12, 2018, at 09:58, Rusholme, Benjamin (Ben) <rusholme@caltech.edu> wrote:
> > Hi,
> > >> On Apr 17, 2018, at 5:05 PM, Rusholme, Benjamin (Ben) <rusholme@caltech.edu> wrote:
> >> >> Action: Ben to contract Frank once new IPAC network engineer onboard.
> >> > IPAC has hired a network engineer specifically to work on our Science DMZ/WAN connection:
> >> > Doug White <dwhite@ipac.caltech.edu>
> >> > Could you please put him touch with PRP and anyone else you think helpful?
> >> > Thanks, Ben

Kottmann, Brett

to tdefanti@ucsd.edu

I've been reading about the Pacific Research Platform and think a regional research platform is a great way for institutions in Ohio to move forward.

Are you aware of any resources that might help us map out a similar path?

Thank you,

--

Brett Kottmann
NSF Cyberinfrastructure Engineer & Educator
Email: brett.kottmann@uc.edu Phone: 513-558-4195
UCIT R&D: https://www.uc.edu/ucit/services/research-development.html
Hi Brett,
My first thought is to put you in contact with Jen Leasure of The Quilt. She works closely with us to help connect people like you to your regional/state R&E network (in your case, OARnet). See for example: https://www.oar.net/calendar/events/quilt_fall_member_meeting

I'm happy to talk to you about working directly with you for testing and technology deployment (we do this with Clemson, for example). Anyone can join our rocketchats and weekly phone call, participate in science engagement calls, and so on. Let me know what interests you.

Thanks for contacting me!
..tom..

---

Hello Tom and Brett,
I appreciate being brought into the email conversation about state/regional research platform efforts. As Tom mentioned in his note, OARnet, the state R&E networks in Ohio is actively engaged in national research platform efforts and leads the Ohio community effort on friction-free networking for scientific research. Paul Schopis is the executive director at OARnet and is a Quilt board member. I would be pleased to connect you with Paul and his staff to get involved in the Ohio institution efforts toward a regional research platform and will send an introductory note to Paul under separate cover.

Best Regards,
Jen

Jen Leasure
President, The Quilt
206-403-3188
jen@thequilt.net

---

Kottmann, Brett

to Jen, Thomas

Fri, Sep 21, 2018, 6:11 AM

---
Thank you for your help in getting us started. We've worked with Paul in the past and OSC is one of the entities we have listed as a key component of any regional platform for our area.

As our efforts get underway we'll certainly want to reach out to existing entities like the Quilt as well.

--

Brett Kottmann
NSF Cyberinfrastructure Engineer & Educator
Email: brett.kottmann@uc.edu Phone: 513-558-4195
UCIT R&D: https://www.uc.edu/ucit/services/research-development.html

---

Larry Smarr
Sun, Dec 2, 2018, 10:48 AM
Reply to all

to sergei, Frank, sgleyzer, Tom

Sergei-I am following up on our discussion after our panel on Friday. I believe you know Frank Wuerthwein, who works closely with me on distributed cyberinfrastructure. He also know you.

What I would like to do is get you on our Pacific Research Platform Nautilus hypercluster, so you could install some of your machine learning tools that you and Frank could try out on CMS data.

I am attaching some ppts to give you an idea of what our PRP CI looks like. It contains several hundred 32-bit GPUs and several thousand CPU cores. We also have been experimenting with Google TPUs, as you have been.

Frank-I will leave it to you as PRP co-PI to work with Sergei on this, but keep me and Tom in the loop by cc.

Of course, John and Dima will be available as needed.

I am excited about this opportunity!

---

Sergei V. Gleyzer
Sun, Dec 2, 2018, 1:00 PM
Reply to all

to Larry, Sergei, Frank, Tom

Dear Larry,

It was pleasure to meet you on Friday! Thank you for following up so quickly.

I read through the slides, and PRP sounds like a fantastic resource for ML, as you have
described.

I am certainly interested in this opportunity and would be happy to follow-up with Frank on the next steps to get started.

I have a project in mind that can be used as a starting point.

Best regards,

Sergei

---

On Nov 28, 2018, at 12:02 AM, Tajana Simunic Rosing <tajana@ucsd.edu> wrote:

I’d be happy to get together. I’ve cc’ed my team’s expert in power/performance analysis, Yeseong. How about Friday, Dec 7th, at 10:30am or early afternoon Dec 11th?  
Tajana Simunic Rosing  
Full Professor & Fratamico Endowed Chair  
Dept of CSE and ECE, UCSD

On Mon, Nov 26, 2018 at 9:42 PM Larry Smarr <lsmarr@ucsd.edu> wrote:  
Tajana—we are finally ready for your expertise on this project. 

Let us know when you can meet with us...

On Mon, Nov 26, 2018 at 2:44 PM Thomas DeFanti <tdefanti@ucsd.edu> wrote:  
Hi Tajana,

We now have about 100 CHASE-CI GPUs online, and we will soon add another 100 or so. We have been collecting data about usage (see: https://grafana.nautilus.optiputer.net/d/f_qakuSiz/k8s-nvidia-gpu-cluster?refresh=15m&orgId=1 )

We think that there’s enough data for you to dive into.

We want to see if our GPU/CPU/RAM/10G networks are being used or if something is waiting for something else. How can we optimize these boxes?

Larry has some detailed questions he’d like addressed. Dima and Nadya are proficient at Grafana, and can teach your students how to grab the data and generate charts.

It would be good for us all to frame the questions before the holiday break in 4 weeks and then focus on getting answers next year. What do you think?
Looking forward, ..tom..

APPENDIX F: Collaboration Plan - EPOC and TNRP May 1, 2019

Schopf, Jennifer M

Thu, Oct 18, 2018, 5:59 AM Reply

Hi Tom-

I’m at Tech Ex, and finally had a chance to get clarification about the NPEO awards from Kevin (who only became our PO in the last week for this project). He suggested to me (and I’m assuming to folks on your side) that we come up with a coordination plan for the two projects. In addition, several folks who were on the *RP engagement call Monday (which John M knew I wouldn’t be on since I was flying here) have made a point in approaching me so that we could clarify how the efforts overlap – the initial concern as expressed to me was coordination for user engagement (such as our Roadside Assistance program), but overall a better understanding of how the two projects might go forward. We’ve been trying to get information on our pieces public as we can - I’m attaching the 2-pager we’ve been using to describe he overall center, and we’ve got more info on the website as well – http://epoc.iu.edu. Anything you could share on your side would be helpful.

... So please let me know if we could try to find time tomorrow. Thanks.

-jennifer

Dr. Jennifer M. Schopf
Director, International Networks
Director, Engagement and Performance Operations Center
Indiana University

Summary
In 2018, the NSF CC* program funded two proposals in response to its request for a Network Performance Engineering and Outreach (NPEO) project submissions (https://www.nsf.gov/pubs/2018/nsf18508/nsf18508.htm). As the solicitation originally specified that there would be only one awardee, neither project plan included a plan for the two funded centers to collaborate. This document gives a brief overview of the two projects, identifies their areas of overlap, and details for collaborative areas going forward.

Engagement and Performance Operations Center (EPOC) - NSF Award#1826994
The Engagement and Performance Operations Center (EPOC) is a production platform for operations, applied training, monitoring, and research and education support. Over the last decade, the scientific community has experienced an unprecedented shift in the way research is
performed and how discoveries are made. Highly sophisticated experimental instruments are creating massive datasets for diverse scientific communities and hold the potential for new insights that will have long-lasting impacts on society. However, scientists cannot make effective use of this data if they are unable to move, store, and analyze it.

The Engagement and Performance Operations Center (http://epoc.global) is a collaborative focal point for operational expertise and analysis and is jointly led by Indiana University (IU) and the Energy Sciences Network (ESnet). EPOC provides researchers with a holistic set of tools and services needed to debug performance issues and enable reliable and robust data transfers. By considering the full end-to-end data movement pipeline, EPOC is uniquely able to support collaborative science, allowing researchers to make the most effective use of shared data, computing, and storage resources to accelerate the discovery process.

EPOC supports five main activities
1. Roadside Assistance and Consultation via a coordinated Operations Center to resolve network performance problems with end-to-end data transfers reactively;
2. Application Deep Dives to work more closely with application communities to understand full workflows for diverse research teams in order to evaluate bottlenecks and potential capacity issues;
3. Network Analysis enabled by the NetSage monitoring suite to proactively discover and resolve performance issues;
4. Provision of managed data services via support through the IU GlobalNOC and our Regional Network Partners;
5. Coordinated Training to ensure effective use of network tools and science support.

Toward the National Research Platform (TNRP)- NSF Award #1826967
The NSF-funded Pacific Research Platform (PRP) provides an Internet platform with 100-1000 times the bandwidth of today's commodity Internet to major research universities on the West Coast and several locations in the US and Europe/Asia. Now in its fourth year, it has been designed to serve specific application needs of researchers. The disciplines which are engaged in partnering with the PRP range from particle physics to climate to human health, as well as archaeology, digital libraries, and social media analysis. The next stage, now underway, is trying to understand how to scale this prototype cyberinfrastructure to National and Global Research Platforms.

Toward the National Research Platform (TNRP) addresses issues critical to scaling end-to-end data sharing and computing. TNRP instrumenting a large federation of heterogeneous "national/regional/state" networks (NRSNs) to greatly improve end-to-end network performance across the nation, which requires active participation of these distributed regional-level entities to reach out to their campuses, and the help of Internet2 and ESnet. Regionals are trusted conveners of their member institutions, contributing effectively to both the people networking and intercampus networking that are as necessary to the development of a full National Research Platform as is the stability, deployment, and performance of technology. TNRP's collaborating NRSNs structure leads to engagement of a large set of science applications, identified by the participating NRSNs, Internet2, ESnet and the Open Science Grid. This project offers the Kubernetes-managed PRP Nautilus Cluster as a replicable model. Nautilus includes equipment purchased from many fund sources and is coordinating closely with OSG on technological and science engagement practices.

TNRP is highly instrumented to directly measure performance. Visualizations of disk-to-disk performance with passive and active network monitoring show intra- and inter-NSRN end-to-end performance. Internet2, critical for interconnecting regional networks, provides several
instrumented dedicated virtual network instances for the interconnection of TNRP's NRSNs and OSG’s use. Cybersecurity is a continuing area of research and deployment; evaluations of advanced containerized orchestration, hardware crypto engines, and novel IPv6 strategies are part of the TNRP plan.

Areas of Overlap
In practice, although the goals of the two projects are similar, the day-to-day work by the two groups differ somewhat. In general, the TNRP project is more focused on advocacy to move towards container orchestration solutions as well as encapsulated and mobile storage and compute infrastructure to enable the use of more advanced network and compute infrastructures, including but not limited to the PRP Nautilus cluster. Their primary science use cases to date (http://pacificresearchplatform.org/nautilus/namespaces/) tend to be associated with the PRP or an R1 institution. They are planning to address scalability issues for user support in part via AI techniques with their RocketChat support room.

EPOC is focusing more on working with application scientists in their current environment and debugging existing infrastructure with limited changes to the underlying technological basis. Their service is available to anyone working in research or education. They are also promoting the use of advanced infrastructure in a supported way for organizations that don’t have the inhouse capability to offer these services. Their focus is more strongly on underserved (non-R1) institutions and small to medium sized research collaborations. They are focusing on an operations center approach for support.
<table>
<thead>
<tr>
<th>EPOC</th>
<th>TNRP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roadside Assistance helpdesk for any research or education project experiencing problems with data transfers; Consultation helpdesk to assist network engineers on design of DMZs, DTNS, etc</td>
<td>Rocket chat to assist end users of Nautilus cluster and to help with deployment of PRP-like regional/national infrastructure; weekly technical engagement of regional and national network engineers and application engineers who work with their users.</td>
</tr>
<tr>
<td>Advocacy to end users to understand future cyberinfrastructure needs through use of Deep Dives</td>
<td>Advocacy to end users to make use of advanced cloud technologies through the use of Kubernetes and Jupyter notebooks and high-speed shared storage technologies.</td>
</tr>
<tr>
<td>Science engagement to R&amp;E community, primarily through six science community partners, six regional network partners, and infrastructure partners including Internet2, Quilt, CACR, Science Gateways, CACR, and XSEDE, and CARCC. Engagement with the network engineering community through the cyberinfrastructure engineering list, campus champions, and strong ties to partner regional networks.</td>
<td>Science engagement to R&amp;E community, primarily through PRP, regional network, Internet2, ESnet, Quilt, and OSG partnerships. Engagement with CIOs through Internet2, CENIC, and regional/national networks.</td>
</tr>
<tr>
<td>NetSage and perfSONAR deployment for measurement and monitoring of network infrastructures supported by partners (For example, <a href="http://gpn.netsage.global">http://gpn.netsage.global</a>)</td>
<td>perfSONAR and DTN deployments for measurement and monitoring of network infrastructures supported by partners: <a href="https://perfsonar.nrp-nautilus.io/maddash-webui/">https://perfsonar.nrp-nautilus.io/maddash-webui/</a> and numerous Kubernetes-based open source measurement and monitoring tools for networks, computing, and data sharing.</td>
</tr>
<tr>
<td>Development and advocacy of advanced services focused on small and medium sized institutions: perfSONAR in a box, DTN in a box, DMZ in a box, Testing Service in a box</td>
<td>Development and advocacy to support science use flows: Kubernetes, Jupyter notebooks, Nautilus cluster and similar federated clusters nationwide</td>
</tr>
<tr>
<td>Training for the community in Deep Dives,</td>
<td>One-on-one instruction at Calit2 to support</td>
</tr>
</tbody>
</table>
Points of Collaboration
The two projects held a multi-hour meeting following the Quilt spring meeting in February to discuss how the two groups could collaborate going forward. The following items were identified:

- Continued and ongoing meetings to take place at shared venues (I2, Quilt, NRP, etc)
- Joint participation in the PRP/*RP Engagement calls on Mondays
- Shared notification of training opportunities
- Continued cooperation with partners in common, especially in the area of the *RP Pilot Studies

<table>
<thead>
<tr>
<th>Points of Collaboration</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>How IT can talk to scientists, Finding researchers on your network/campus, perfSONAR installations and support - all primarily classroom settings. Additional consultation one-on-one as part of Roadside assistance and consultation track.</td>
<td>use of Kubernetes, Jupyter notebook, nautilus cluster, FIONA construction and deployment.</td>
</tr>
<tr>
<td>Participation in alternate Monday PRP/NRP Pilot Engagement Team Meetings (led by Internet2), leading the weekly cyberinfrastructure engineering brownbag calls, Quilt meetings, and NRP annual meetings.</td>
<td>Participation in alternate Monday PRP/NRP Pilot Engagement Team Meetings (led by Internet2), weekly PRP/NRP Pilot Thursday Technical Engagement Meetings (led by CENIC/PRP/TNRP and Internet2), Quilt meetings, and NRP annual meetings.</td>
</tr>
<tr>
<td>Presentations at Internet2 and regional network conferences and workshops on basic and advanced topics. Demonstrations at SC conferences and other venues.</td>
<td>Presentations at Internet2 and regional network conferences and workshops on basic and advanced topics. Demonstrations at SC conferences and other venues.</td>
</tr>
</tbody>
</table>
APPENDIX G: Thursday Engineering Call Minutes
These mostly unedited minutes were taken mostly by Tom DeFanti. They include PRP, CHASE-CI, NRP Pilot, TNRP and GRP discussions from October 2018 through June 2019, in reverse chronological order.

June 20, 2019
John Hess, John Hicks, John Graham, TomD, Greg, Dima, James, George, Chad, Guaravadeep, Kevin, Jeff, Stephen, Mark, Stephen, Kate, Chris, JimK, Celeste, Hans.

Here's a draft agenda for the NRP Pilot portion of the call:

- Update on GridFTP pscheduler plugin
- What are the successes, opportunities, and challenges we should talk about and work on at the 3NRP workshop?
- Updates from NRP participants
  - Updates on node deployments
    - Current dashboard status - https://perfsonar.nrp-nautilus.io/maddash-webui/
- Action item review

JohnG: HA Proxy 2.0 is new, for K8s ingress. Need to figure out how to do multi-domain load balancing for HA masters. Will do first with HIPPA HA cluster, then HPWREN, then Nautilus HA. K8s is at v.15, we aren't yet. Dima: Cuda 10.2 Nvidia now. Tensorflow not even 10.1 yet. Updated Jupyter image to Tensorflow 2. beta. 2 regions of EdgeFS in Nautilus--one happy one unhappy user. Benefits in speed, permission issues, seemingly stable. JohnG: HPWREN a struggle--hardware upgrades bumpy. Took back some gear from SDSU while ChrisP is away (HPWREN server and U200 host box). Should be attaching HPWREN at 100G to new deeper buffer Nexus top of rack switch. Tom: Cisco building an 8-GPU box we will test.

JohnG: FPGAs: Sharing U200 Vivado page. 25 licenses. Can expose the devices through the containers. Azher has P4 SDnet dev environment running.

Dima: got a deep dive into perfSONAR. Can config dashboards by assigning labels to nodes in the cluster, and can scale it.

From Dmitry Mishin to Everyone: (10:14 AM)
  https://perfsonar.nautilus.optiputer.net/maddash-webui/
Tom: Dashboard looking good!
JohnG: showed what happened when the primary path came back on from NCAR:
From John Graham to Everyone: (10:17 AM)
  https://perfsonar.nautilus.optiputer.net/maddash-webui/details.cgi?url=maddash/grids/Nautilus+Mesh+-+Latency+general+-+Loss/k8s-gpu-1.ucr.edu_internal/fiona.nwsc.ucar.edu_internal/Packet+Loss

James: Dima and John--did you talk to Alberto/Alistair at caida? GPN and I2 working on refining Artemis (BGP hijack protection). Folks in Greece using K8s now. Big win for community--highly scalable, NSF funding, want to share my excitement--can we do this within Nautilus for enhancing BGP security. JohnG/Dima: met with their team last week. James: I2 all in. Very very very very very great.

From James Deaton to Everyone: (10:24 AM)
A set of 4-5 containers, demands low. Just consuming a bmp stream from ripe and doing analysis. Informs about BGP table handling. Program officers fond of Artemis. JohnG: BGP is one of the few ports exposed in a K8s cluster. James: my goal is to see this going asap. TomD: will put in the TNRP Annual report as new/ongoing effort.


From Jeffrey Weekley to Everyone: (10:37 AM)
John Hicks: have you looked at GridFTP plug in for pscheduler. J&D: nope. Nadya was going to do that, but she's gone to UCI.
John Hicks: anyone else in NRP Pilot needing this?
John G: Ansible light enough, should not concern anyone.
John Hess: Two new nodes at AMNH--will join Nautilus?
Jim K: U Delaware has ML/AI/IOT researcher. She wants to get involved. Hook into Nautilus. John G: use our IOT devices through K8s. Jim K: She has mobile devices, local FIONAs we set up. John G: looking for users of Javiers, Nanos, Google Edge TPUs--will be able to send containers to any, schedule. Some have cameras. Create collaborative namespaces, add your own. Jim K: any researcher wants to know that resources are available. John G: can prototype before they get their own gear in place. John Hess: invite her to this call sometime.
Tom: UCSC will use Nautilus FIONA8/2s for a course in July pending their cluster coming up.
James: do we have information on node comparisons? Prometheus and Grafana have that data per namespace and per node. Need to run active measurements. Will take it to rocket chat.

End...next meeting June 27. John Hicks lead off then.

June 13, 2019
John Hess, John Hicks, John G, Tom D, Dima, Kate, George, Chris, Jeff W, Gauravadeep, Cathy, Greg, Igor, James, Shawn, Mark, Kevin, Shash, Kristy, Dana, Hans, Hervey, Jim-C

Here's a draft agenda for the NRP Pilot portion of the call:

- Agenda bash
- Go over networking issues with fiona.nwsc.ucar.edu
- Igor Sfiligoi to give a brief on testing to commercial cloud providers
- Update on GridFTP pscheduler plugin
- What are the successes, opportunities, and challenges we should talk about and work on at the 3NRP workshop?
- Updates from NRP participants
  - Updates on node deployments
• Current dashboard status - https://perfsonar.nrp-nautilus.io/maddash-webui/
  • Action item review
  • Adjourn

JohnHicks:
James: I want to talk about crosscutting stages people are at in the pilot. Elastiflow in Nautilus great. One of the things we all have is the ability to get flow data/tstat data out of all DTNS, even in not in the K8s cluster. Doesn't matter if stand up the right set of containers within Nautilus to receive the data. Resonates with folks, shows collaboration. JohnHicks: Good for 3NRP meeting. Nautilus will need to implement this into K8s infrastructure, but also allows others in. Gives good data from their machines. James: JohnG, you OK with this? JohnG: sure. also tcpinfo, etc. People can send us, fine. Don't know how to viz what we have now, so need a workgroup to collaborate on the viz. JohnHicks: REUs? JohnG: yes. Looking for matching NSF REUs. JohnHicks: any problem with sharing flow data from DTNs? JohNG: can work with CAIDA to de-identify. Tcpinfo Meas. Labs suite de-identifies. James: if just the DTNs, don't need to de-identify.
Guaravadeep(Ciena): does this tie into the U200 cards too? Just got our system, discussing telemetry extraction from these Xilinx cards. JohnG: getting some data from the pod the apps run in, not sure what we'll see from the internals. Been working to set up the working environment. Upgrades are not yummy. Don't do 2019 toolset yet. Guaravadeep: just starting. JohnG: have sdnet license, 20 seat license for novato, etc. that you can use. Sdnet has p4 examples. Need to work with Xilinx on this more today. I'll invite you to our gitlab so you can have access to our notes and install tree, and check out one of our licences.
JohnHicks: how do we move forward to engage everyone to work on deliverables. Start in rocketchat? Another meeting? NRP participants will need to do some work, need engagement & commitments. Always flowdata privacy, but we can focus on DTN traffic. James: grab the metadata John is collecting already. Get the ones not contributing flow data talking to us, work through flowdata privacy concern. Let's get to the heart of it. JohnG: collecting is easy--want to use it via viz. Worried that someone might take it off the system and be a bad actor. Until we know what's in it, we shouldn't proceed. I don't know what's in the tons of Nautilus info. Need to look at it. James: want to make progress across domains by 3NRP. MaDDash gives info, but want more details. George: can't we specify what's unacceptable. Can we opt-in, ack flowdata is OK. JohnHicks: flowdata doesn't expose payload. George: if I have a DTN and blast data, is that ok? JohnG: gotta look at it. James: based on def. of flowdata and ingest process... JohnG: Nautilus has a much bigger set: tcpinfo is collecting a lot from all the DTNs. Need to get people together to see what's acceptable, ok for public, or password. Global collection going on? JohnHicks: can we narrow focus to just NRP Pilot? John: after we know what's there. James: so there is a framework to ingest. JohnG: we are collecting sflow, tstat, tcpinfo. Sflow only thing feeding into Elastiflow, the others don't. James: narrow the focus back down, will Nautilus ingest data? JohnG: if curated by a NRP consortium. James: let's meet next week and step through what's being collected. JohnG: Invite Sean Pisart? James: could have e2e

JohnHicks: UCAR problem
JohnG: UCAR starting dropping packets last night. Sometimes 100% percent packet loss. Still broken. Combination of a fiber cut in El Paso and a flappy regional link. JohnHicks: why didn't the fiber cut just reconfig itself? Dima: problem with Ceph--UCAR node part of Ceph. Ceph ok if total loss, but gets unhappy if a node can see some nodes but not other ones. UCAR went onto a lower MTU path--could see some nodes. New data written but Ceph can't update. JohnHicks: can this be fixed? Dima: shut down immediately? I missed that point. Ceph can't deal with partial communications. Really hard to find it. Dana: impact on end users? Dima: not whole Ceph is stuck, just some volumes. Gitlab, prometheus got upset. Lost all monitoring. CephFS most likely active. Things should come back. JohnG: Still flapping? Dima: 1Gb now, didn't try. JohnH: Calren doesn't have a router at NCAR. FRGP peering. Probably terminating in Sunnyvale but maybe LA too. JohnG: UCAR now says: shut down BGP between FRGP and CalREN at about 10:07 MST this morning. John Graham, do you still see network issues when you try to reach UCAR? The latest information I have is that Internet2 is waiting on CENIC, who is working with a vendor in El Paso. They have isolated the problem to a single card, which they reset at 10:00am MST. We haven't heard what the result of resetting the card was, but I2 hasn't seen problems since then. We just aren't ready to declare victory until we get some kind of message. Of course, as soon as that happens, I will restart BGP and notify this list. JohnHicks: what can we do? JohnG: once we eliminate the Ceph vulnerability, we can get messages from nodes and shut down. Use smarts on the controller backplane. What Facebook and Google do every day. JohnHess: maybe do a small group post mortem with Calren engineers to scrub backup path between FRGP and CalREN. Make fails more graceful, and stay failed until path is better. This is an isolated event, but worth clearing up for the future. So many devices in the failure path increases chance of this problem reoccurring. JohnHicks: a great idea. JohnHess, can you coordinate this meeting? Are there other groups vulnerable? Dima: we are most sensitive to MTU changes. Doesn't affect most people. Need to detect when the path isn't MTU 9000. JohnG: our network is still broken--need to fix this now. Who from CENIC can look at this? JohnHess: the CENIC NOC needs to drive this now. They are engaged with fiber provider (Century Link). JohnHess: most likely BGP sessions without alternate virtual layer2 path. Could look at existing sessions and how they are
provisioned, have a northern route too besides via El Paso. Future discussion. For now, CENIC NOC is the way to go.

JohnHess: Can John and Dima figure an alternative way of draining the node.

Igor: cloud testing with GridFTP: Started up new instances in AWS, installed MaDDash endpoints there. Problem initially with traffic one-way only. Can have external node get data out. Hard to get into AWS. Get a private IP with forwarding NAT for ingress. Globus URL copy binds to this private IP address. Fixed that by an environment variable you can set and things just work—not documented anywhere—dug into source code. Have it working between AWS and the other MaDDash nodes. Some of the routes were a little funky, but look good. Larry: UCSD CIO and SDSC Director has worked out Cloudbank with AWS—a way to make credits go further.

JohnHicks: initial routing problem being worked on—a tagging issue fixed. NAT issue on perfSONAR side is well documented. Igor: MaDDash runs outside pscheduler. JohnHicks: perfsonar works with pscheduler. Globus URL copy supports with the undocumented env. variable. JohnHicks: need to figure out how to run things in the cloud.

May 30, 2019
JohnHicks, TomD, Azher, Kristi, Dana, Cathy, GregH, Gauravdeep, JohnG, JeffH, Stephen, Chad, George, TomH, James, Dima, Kate, Mark, Shash, Jim, JeffW, Alan

John Graham gave a brief visual tour of his NM spot and talked about his excellent fiber connection. Talked about his Mobotix camera installation to test picking out wild fires. We have a visitor from Ciena, Ontario today: Gauravdeep Shami.

Gauravdeep: setting up VPNs through StarLight. Have ordered Xilinx U200s for P4 cards. Demonstrated P4 at SC18. Objective is to integrate P4 and the U200s for high-speed data transfers for SC19. Azher: heard that compiling P4 requires special software, not available in general. Gauravdeep: we rely on patches. You could design your own bitstream and get it working. Azher: the P4 compiler is not available to users. JohnG: also Xilinx-based Micron with 512GB. Good development environment. Having some problems with Centos 7.6. Bugs reported last night. Hope to get resolved soon. Got the 2 new single socket Gigabyte chasses, have physical problems with the lid—need to grid. Need to install another NVMe stick for software. Need to work with TomH to get another 100G machine at SDSC. TomH: e-mail me the power requirement. JohnG: not a big machine. Do we need a separate set of vlans for SC19? TomH: yes, easy to do. JohnG: will Ciena have 100G to your booth, Gauravdeep? Gauravdeep: not sure yet, should be possible, I’ll ask Mark. JohnG: Harvey has been sending e-mails, as has Joe Mambretti.

JohnH: anything else on Nautilus? JohnG: Igor’s been doing testing with NRP-Pilot sites and AWS. Asking AWS for some time.
Dima: Nautilus users required to add information into their namespaces. Nadya's been working on the namespace displays on the website. JeffW: do you have a quickstart guide?
From James Deaton to Everyone: (10:20 AM)
   Can we have the ability to change the namespace name within the portal? :)
Dima: no. Create another one. James: will do.

From Dmitry Mishin to Everyone: (10:20 AM)
   https://nautilus.optiputer.net/namespaces-q
Main portal page has a quickstart guide. Users are required to read the guide.
JeffW: I'll give a presentation to the help desk student workers.

Dana: researchers and students hate to read. Dima: I tell users to read twice a week.
JohnG: the docs have been getting better; want chatbot automation. Now provide direct links to needed information.

Dima: 4 new nodes in the cluster. 2 nodes have no internet access. 2 others having problems.
Gauravdeep: is one of the new nodes a MREN one? John: maybe. Dima: probably the VM. Gauravdeep: Ciena has VPN that is reachable but not routable. We call it the Ottawa DTN and want to connect it to PRP, via the MREN cluster. John: we will try to do Admiralty federation with MREN. Dima: Admiralty installed on 2 clusters, need to run some jobs next. Gauravdeep: we have a spare DTN and 40G connection perfect for the NRP. I will send screenshots.
JohnG: we've learned a lot about Ceph and EdgeFS regarding localization pools. Data is available on your namespace on other clusters with EdgeFS.

JohnG: HPWREN devices sent out with quad NVMes. Should be installed very soon. Greg: looking forward!

JohnHicks: Dana: update on 3NRP?
Dana: registration e-mail went out. James should get all complaints.
James: working on "What is the NRP?"
JohnHicks: do you want content from this group? Dana: yes. Collating list of great ideas for the committee to go through. Logistics well underway, focusing on content.

JohnHicks: where are we on the stretch goals? Experiment with other data transfer tools? Nautilus has a network file system, difficult to measure performance. Other than GridFTP, are there other tools and protocols. TomH: FDT is popular. Unless researchers are using Globus online, they find GridFTP too complex, so use SCP. JohnG: we've been using open source Aria2 built into nextcloud. TomD: Igor is testing AWS and wanting to measure xrootd. James: Aria2 is wonderful, but is rough on other users.
   https://aria2.github.io/
aria2 is a lightweight multi-protocol & multi-source command-line download utility. It supports HTTP/HTTPS, FTP, SFTP, BitTorrent and Metalink. …
From Dmitry Mishin to Everyone: (10:46 AM)  
  https://rclone.org/  
Azher: for SCAsia data transfer--award for most innovative solution went to MREN. Company called Zettar won the contest. Anyone working with Zettar?  PB in 20 hours.  
From George Robb III to Everyone: (10:48 AM)  
  https://www.zettar.com/  
From James Deaton to Everyone: (10:48 AM)  
  And JohnH, here is the other one: https://aria2.github.io/ you can feed it multiple ways to access a single file/directory and it'll combine them and stream them across all the access techniques  
From George Robb III to Everyone: (10:49 AM)  
JimChen: Zettar group came out of SLAC, from the R&E community.  
JohnHicks: Is Nadya working on the GridFTP plug in.  Leaving for UCI/PhilP.  Nobody from UCSD working on this? Have reached out to I2 people (Nate, Mark Fite (?), etc.) and their interns to work on this summer.  
JohnG: Nadya has clear notes, we will try to keep it going.  

May 23, 2019  

JohnH: Data mover challenge for SC Asia?  
JohnG: we can spin up globus connect servers on any node. FDT can be spun up again with Caltech (would be easy). Want to focus on FPGA ourselves. 
Shash: Harvey is planning something.  JohnH: Harvey has submitted NREs.  Also a document for I2 TechX.  


JohnG: 2 new FIONA8s at UCR now up.  Need to get UCSF's ACL opened.  Dima: people are now using perfSONAR namespace and solving MTU issues themselves.  
JohnG: Giving iperf toolset to anyone who needs it is powerful. Dima working hard on yum repose.  5.1.2 freshest. ZFS 5.1.2 package single install--takes 40 minutes.  JohnG: doc has recipe for workaround.  Latest testing repo.  Follow up on rocketchat.  

Dima: EdgeFS getting debugged.  Fixed/redeployed.  CSI drive magic to automatically map the closest region.  JohnH: I want to know how that magic works.  Dima: me
too. Throw pods in cluster or federated clusters and once a pod is assigned a node, it's attached to the EdgeFS region closest to that node, and works like a cache. Dima: I have Admiralty installed. Working with the developer. Can see resources from other clusters. Scheduler has heavy resources consumption. We'll see how that works. JohnG: will run on our HA Nautilus cluster. Dima: permissions need to be granted. Dima: Admiralty agent talks to top level scheduler. JohnG: can it run HA? Dima: yes.

From john hess to Everyone: (10:20 AM)
https://github.com/admiraltyio/multicluster-scheduler

From john hess to Everyone: (10:22 AM)
https://github.com/admiraltyio/multicluster-service-account

All regionals will have accounts in the top-level. Place holders for jobs, pods, deployments you can query: a metadata proxy.

JohnH: want to get PacificWave DTNs in. JohnG: Dima's repos are helpful for machines with only L2 paths. Nvidia drivers. Helped get some MREN nodes up. MREN all green. JimChen: working with multiple nodes in StarLight (6) and UIC, plus two new ones. JohnG: good to do Admiralty tests. Guys from Ciena have one coming online. Dima: I had firewall issues. JohnG: NJedge came on. Adam Bathard on the rocket chat. Have VM from GPN coming up w/beefy machine. Dima: not there yet. Have MTU 9000. Dima: will slice our cluster to separate OSG/stashcache nodes from Nautilus; will be its own K8s cluster, won't use our policies and we can update more easily.

Dima: started collecting metadata of our users. Have a page for users to type in metadata. Today I'm going to enforce--deny on submission. TomD: Lots of namespaces not being used.

From Dmitry Mishin to Everyone: (10:20 AM)
https://nautilus.optiputer.net/namespaces-g

JohnG: Chris Paolini giving us 12 dual core machines--more compute at SDSU. Delivered the 100G FPGA to Chris. Now In Tyan 1U. Chris is ordering a Tofino 32x100G switch. Still having problems with Tyan IU ILOM. Won't take the settings, despite 2 bios updates. A pain. Still need to solve that problem.

JohnG: got 32 edge TPUs, had to pay 25% customs duty ($1,400). Haven't received them yet. U200 development environment needs Windows, bloaty stuff. Not data-center usable for us yet.

TomD: texting NRP DTNs against 4 AWS sites. JohnG: curious routing choices between Amazon and campuses. Will look at the traceroutes. TomD: part of getting AWS and the other national resources into Nautilus.

JohnG: Micron has a brand new FPGA product. Following up. Micron has 512GB of memory! The CONNECT algorithm can hold 38 years of data in that memory. Micron
has own dev environment, uses same tools as we use for Xilinx FPGA. Micron uses Xilinx FPGA with Micron board and memory. Crosspoint from Micron. Compute-in-memory orchestrated by an FPGA could be a cool hybrid way to play.

From john hess to Everyone: (10:44 AM)


JimChen: still buy passive? JohnG: Xilinx recommends that for a server with airflow. Not much difference in price from blower version w/fans. Xilinx can tell you which if you tell them the chassis. Azher: allows daughter boards. JohnH: single PCIe 16--but there are other links off the FPGA.


From John Graham to Everyone: (10:53 AM)

https://www.servethehome.com/dell-emc-dss-8440-10x-gpu-4u-server-launched/

JohnG: deploying 4 cameras in NM with inference at the edge for high framework neural nets for smoke detection. Have 1G to my pole. Have 1G 60KM link. Fastest fire tower on the planet. Local training gpu cluster for these cameras. Mobotix camera streaming into Nvidia nano. Tensorflow light model. Nvidia digits workflow. Got a /27 IPv4 and a chunk of IPv6 (/48) from the local ISP. Should be able to walk into the SunCAVE and see it live. Cameras also come with microphones.

Conflict with CalREN HPR next week (May 30) but John Hicks will run the meeting, according to Cathy.

No call Thursday morning June 6 due to OSG/SDX workshop.

**May 16, 2019**


JohnG: kernel bug below 5.9.08--needs repatching immediately, Dima has repo. Have ZFS problems will do XFS now. JohnHicks needs to do his I2 boxes. Joe Keefe has RPMs we can put in our ZFS repo, but it's a major upgrade. Dima: Terrence says Docker is recommending an overlay2, not ZFS. (Tom: see https://docs.docker.com/storage/storagedriver/overlayfs-driver/).

JohnG: all in a state of limbo. XFS may have journal delete issues. Still like ZFS. Dima: ZFS very stable at SDSC. Slower but resilient.

JohnG: we need to build a good release and put it in our yum repo. GregH: will EdgeFS be stable for production? JohnG: yes. Am deploying NVMe in HPWREN so will have an
evenly distributed storage pool. ZFS used for docker and kubelet. Dima: region map is a work in progress.


JimK: we are staying with PRPv1. Kate: there's a Kansas City PoP GPN FIONA. Want perfSONAR to the out state networks. JohnG: do you want your own MaDDash for your specific sites? Kate: yes. JohnHicks: aren't we doing that through OSG/Shawn's config. Dima: we use 9443 port so we don't conflict with the regular perfSONAR port. JohnH: GPN FIONA node is storage and stashcache node for OSG. JohnG: you guys could launch your own MaDDash on your external network. JohnG: that node is very busy with stashcache though. JohnHicks: two instances of perfSONAR can compete, use pscheduler to avoid that. Kate: would prefer not to buy another box. Speedy died--was doing traceroute, hosting the MaDDash, perfSONAR. Want to test against 15 nodes 10/100G. JohnG: could host esmond in Nautilus node, spin up another MaDDash. JohnH: maybe add a separate dashboard for GPN, use same esmond. JohnG: simple, good. JohnG: already doing GridFTP testing (but not perfSONAR) on NRP dashboard. Not iperf, bwctl. JohnHicks: the one that Igor did to test external hosts? Shawn McKee has OSIRIS with all the configs doing testing and dashboards. We have external hosts using the same instance for stashcache. Dima: Igor is using it now to test against Amazon. JohnHicks: what John Hess says is very viable. I want access to the perfSONAR namespace. JohnG: done. Testpoint -h gives host attached. Testpoint is the overlay network, inside or outside. Will follow up with you Kate. Kate: I'll get the config files from github, and will need info for the 15 perfSONAR people. JohnG: we'll help. JohnHicks: I'll follow up.

JohnHicks: OESS layer2 provisioning work from IU (discussed at Global Summit). Can map addresses to google/amazon and get 100G express route straight into the cloud. Otherwise bounce around the Internet.

JohnHicks: PRPv1 origins. PRPv2 is compute/storage/K8s. PRPv1 is DTNs and traffic between Science DMZs. JohnG, Dima: how did PRPv1 evolve? Did you solve all the problems in PRPv1? JohnG: we solved a lot of issues. We were doing regular testing with 3TB DTNs. People with data wanted larger storage for actual use....originally for Scott Sellars' CONNECT. Copying it around was not the right answer--attaching is. Evolved Globus/GridFTP into K8s. JohnHess: trying to address end2end problem space. Built a regional dashboard for CENIC. Celeste and Alan Whinery started it. Greg Bell said to get a dashboard and you will find things. Found a border router at UCR that was dropping packets.

Tom: PRPv1 tech for NRP-Pilot is being extended to Amazon, Google, Azure. Spending $200/day. Can we use I2's allocation? JohnHicks: probably. JohnG: through out virtual SONAR buoys. Containerized, spawned, join the mesh.

JimK: wanted to get a market data feed. Didn't have any luck at NYU Biz school. That data is $300K/yr (5-yr old pricing).
JohnH: GPN cluster a viable cluster for federation. Kate: had a call with JohnG, Dima, and Tom on Monday. Adding nodes to Nautilus, federate later mainly? JohnG: the easy way to go. Nautilus could evolve to be a social layer on top of clusters. Dima: we need to break up Nautilus and create HA. Every campus has own master. Evolving ideas.

JohnHess: re-transmits are the killer long distance.

Tom: from the PRP Y4 Q2 report just submitted to NSF yesterday (corrections, discussion welcome!)

The National Research Platform (NRP) Pilot project is now underway with Internet2 and several regional networks in addition to the Pacific Research Platform (PRP) original network. This has led to ongoing Zoom conference calls for technical (weekly, on Thursdays at 10am Pacific time), science engagement (bi-weekly, on Mondays at 10am Pacific time), and leadership activities (monthly).

The Nautilus hypercluster is now the focus of new facilities and support for PRP measurement and monitoring as well as distributed computing and shared storage. It is synonymous with “PRPv2” which was described in the original program execution plan the latter years of the cooperative agreement, Years 4 and 5.

Nautilus is a distributed cluster orchestrated by Google’s open source Kubernetes, a now widely-adopted way to manage containerized software. It is constantly being enhanced by open source packages to manage shared storage, which the PRP implements eagerly, as needed.

Most of the PRPv1 Linux nodes bought by PRP (which we named FIONAs) have been converted to Nautilus storage nodes under Ceph. Nautilus can be defined by the set of such nodes that we have root access and remote management privileges. There are FIONAs and data transfer nodes (DTNs) provided by other funds that are not in Nautilus in many places; these are largely being used for PRPv1-type measuring and monitoring activities. The taxonomy is evolving as organizations add FIONAs with various capabilities (storage, CPU, GPU, FPGA) under non-PRP administrative control, but may be in Nautilus. For example, FIONAs with CPUs/GPUs do not need have lights-out management under our access; if they go off-line, the computing just goes elsewhere until they are restarted. The situation is similar for FIONAs that are just used for monitoring—the MaDDash visualizations just indicate they are off-line. However, shared storage cannot afford to be down very long, so we need to bring it back as quickly as possible, and thus we request root and restart capabilities. Root also allows us to install software patches without getting the attention of the host facilities IT staff. DTNs outside of Nautilus are capable of sending and receiving files at high speeds, of course, which is what the National Research Platform Pilot is accomplishing, in effect, a national PRPv1. Nodes at NERSC, NCSA, TACC, and at several regional network locations (notably to date the Great Plains Network and NYSERNet) perform these functions. Nautilus-like hyperclusters are growing around regional networks and the software to federate these Kubernetes clusters is just now being tested. Internet2 has its nodes in New York, Kansas City, and Chicago joined to Kubernetes, but since they manage the storage as independent OSG StashCache, we do not need root access. A
challenging goal of our efforts is to clarify the distinctions, given PRP’s fast technological evolution.

PRP’s Nautilus now includes >1.7 petabytes of storage, all on the same PRPv2 multi-campus DMZ-connecting network (CENIC HPR). This storage, deployed on PRP FIONA PC termination devices is made up of groups of sixteen 10TB or 12TB disk drives (160 or 192TB per FIONA). Soon to be added is a 2PB BeeGFS storage system sited on San Diego State University (SDSU), bought by PI Professor Chris Paolini with a NSF grant (OAC 1659169) to SDSU. That will bring the total storage on PRP’s Nautilus hypercluster to 4PB.

The Nautilus hypercluster FIONA end-points are located and operational at these partner campuses: UC San Diego, UC Davis, UC Irvine, UC Riverside, UC Los Angeles, UC Santa Barbara, UC Santa Cruz, UC San Francisco, UC Merced, U of Washington, San Diego State University, CSU San Bernadino, Stanford, Caltech, U Southern California, the Naval Postgraduate School (NPS), U Hawaii, Montana State University, UCAR/Wyoming, and U of Illinois at Chicago (see Figure 9 in this report for details). There are FIONAs in Nautilus, three at Internet sites (New York, Kansas City, Chicago), UCAR, at KISTI in Korea, University of Queensland in Australia and the University of Amsterdam in the Netherlands. The University of Guam is joining Nautilus soon.

On the compute side, Nautilus included (as of April 1, 2019) >300 single-precision GPUs, mainly Nvidia GTX 1080s and 1080-Tis, but also 8 RTX 2080-Tis (56 more of which are on order). Two Xilinx U200 Alveo FPGAs have been received and are getting PRP partner attention as well for computing and SDX networking. Of the current FIONAs on Nautilus, 26% have come from PRP funding. Less than 3% (one FIONA8’s worth) of the >300 GPUs were paid for by the PRP, the remainder being from other NSF grants and academic community sources.

JohnHicks: would be interesting to have a mechanism to identify other people’s DTNs in DMZs. Tag that traffic, identify it. JohnG: could experiment with Sflow emitters. Lots of metadata. Peel out labels and graph with Elastiflow. I'll be going to the GPN meeting next week.

JohnHess: next meeting May 23.

May 2, 2019

JohnHess, John Hicks, Tom, Cathy, Chris, JohnG, Chad, Shash, Dima, Greg, Hervey, Shawn, Jeff, Kate, JimK, James, Shava, Mark, Stephen, JohnD, Jim-c, Azher, James, Celeste, JimChen, JimM

JohnHicks:
JimK: interested in P4. Who is doing this? Chris: I have a proposal pending with the Navy and an internal edge-core wedge 32x100G switch, pisa architecture, programmable ASIC with barefoot networks SDK to flash P4 programs to it. Using
Xilinx FPGAs for accelerated computing. Also using Xilinx's P4 development environment—to control forwarding plane, pass packets through an action pipeline for line speed deep packet inspection, finding malicious flows. Can combine P4 with FPGAs do pattern matching/ML. Using RNN for classifying. Happy to send you my proposal. Saturday, there's a NSF workshop in Marietta, GA for ML for intrusion detection.

JimK: are there licensing costs for the SDKs? Xilinx(Vivado) and Intel (Stratus10 board) both have their own SDK, need licenses. Xilinx has U200s and U250s. Two U200s being integrated into Nautilus by JohnG. Azher interested in using SDnet to do forwarding plane manipulation. Also YOLO, CNN models. JimK: want to see some use P4 cases that expose all costs, sw & hw. Chris: want to demo with a geologist with hydrodynamic computation at SC19. Nothing to do with P4.

JohnG: use P4 to throttle wide area throws to give guaranteed bw (say 10G over 100G path). Harvey's group doing this.


JimM: I have colleagues working on security architecture, interested in this.

JohnG: have the 2 U200s in an older chassis. Can do helloworld apps fine, broken bits on CentOS. Xilinx needs to figure out how to mount drivers—encapsulate them internally, and have some Ubuntu hard coding. We met with Xilinx and they will get their devs linked into us to resolve K8s issues. Will take one board to Chris at SDSU. Xilinx came here for a FPGA conference on Tues/Weds. Low-end IOT FPGAs, JupyterLab interfaces, deployment to ARM-64 K3s. Micron has a Xilinx double width card with high-performance Micro memory, will give us a bunch of cards to seed in the community, work with them at SC19. Show interoperability between two vendors. Xilinx would like deploy racks of these U200s at a few universities—we'd like to be one. Gigabyte willing to provide servers for SC19 experiments.

JohnG: using Chris's license server at SDSU. I will ask about a community license for Nautilus. Working on Matlab licensing automating similar to this. Would allow anyone to call back to their university for other similar licensing things.

Chris: can set up a machine that acts as a hw board server with the Xilinx boards you gave us. Can make accessible to anyone in PRP with a FIONA. JohnG; happy to connect them. Azher showed mezzanine card with P4 and 4x10G. JohnG: we've distributed 6 of them. Chris: they draw a lot of power, need extra power supply, works. Azher: I'm using a 12v high-power supply. Using the same bitware card. If you have boards in a DTN at UCSD or SDSU, don't have to worry about licensing.

JohnHicks: please document your use cases and put them on the PRP/NRP web sites.


Dana: how much of Scott's 500x speedup was compute vs network? John: mostly I/O at first, caching onto a FIONA, then containerized him, then parallel GPUs and TensorFlow implementations. Queuing using reddis and scripting with THREDDs server (which reduces how much HDF5 files you need as raw data).
JohnHicks: onboarding new parties. Getting more and more requests. What do I need to do to be part of this? NJEdge wants to be a part, for example, bring some resources. Don't have any true use cases, just the first step. Dima: working on changes in policies for admission. Now accepting CPU jobs in Nautilus via OSG. JohnG: LIGO using lots of CPU cores, the right way. JohnHicks: LIGO can run on OSG, why Nautilus? JohnG: next gen 2U FIONA8 can have 128 CPU Cores (ROMA). JohnHess: is NJEdge building out infrastructure? JohnHicks: they just want to find the sandbox. Tom: new gigabyte $17,500 FIONA8 has gigabyte 2080-Ti cards. JohnG: we have EdgeFS working which allows regionally isolated storage, limiting latency. JohnHess: have they talked to Kinber? JohnHicks: Barr at the ERN? the Rutgers group? JohnG: should engage Barr. Need more scrutiny. JohnHicks: he said we use SLURM, not K8s. Need to figure out how to share.

JohnHicks: Federation?

Dima: Open Source Gravitational Wormhole. New VPN standard with modern encryption. Secures connections between nodes--can build a distributed cluster with security without performance degradation. Can do VPN from local machine into the cluster, make it part of the network. see:

From John Graham to Everyone: (10:48 AM)
https://gravitational.com/blog/announcing_wormhole/

JohnHicks: Admiralty? Dima: still struggling. Working with EdgeFS now. Won't matter where pods run, can use storage they need. When this is done, will use admiralty to spawn pods in different clusters. Need to get storage done first.

JohnHicks: need federated approach. GPN has a cluster. If, for example, I2 put up a small cluster, what do we need to do to make our resources available to you. Dima: need external IPs. TomH here needs to help us with that then EdgeFS will work, maybe next week. Need external IPs for that. JohnG: targeting BeeGFS at SDSU. JohnG: put cilium tools on your cluster, JohnHicks, and can try with our dev K8s cluster here.

TomD: MaDDash GridFTP to AWS from NRP-Pilot DTNs. Igor was able to do this: see From John Graham to Everyone: (10:56 AM)
https://perfsonar.nrp-nautilus.io/maddash-webui/

Need info on how Igor did this. Can set up EdgeFS storage pool at AWS. JohnHicks: used to be bare metal at AWS to test.

April 25, 2019


John Hess: Following CI Brownbag talk series webinar run on April 19th about using perfSONAR with Ansible the Project team is pleased to announce the release of this webinar's recording on Project's YouTube channel. This talk covers use of a perfSONAR Ansible playbook and roles to install and configure a perfSONAR deployment. Such infrastructure will include Archivers, MadDash dashboards, and
JSON schedule publishers as well as multiple testpoints. View the video on perfSONAR Project YouTube Channel https://youtu.be/9OyJsQB59Yg

JohnG: Received two Xilinx U200 FPGAs, installed for testing. Will test device plug in for integration with K8s. Will install in new boxes for SC19 P4 experiments. HPWREN storage cluster: EdgeFS developers helping. Dima: conference call at 11 today to relate our needs and expectations. Adjusting their product for our use case because we are a big customer, focus on throughput for storage over wide area for Nautilus and HPWREN. Moved NVMes from Ceph, will create 20-30TB EdgeFS fast pool. Will have separate regions, with background synching. One big pool, removing problem with high latency. Spinning disks stay same as one big Ceph pool, but each university will have a region of EdgeFS, one namespace. Pods that run at UCSD with storage region at UCSD, will be superfast. Data will be cached in that region, but can pull from outside the region seamlessly. Caches metadata, hopefully all on NVMe. Called the EdgeNautilus project--focusing on HPWREN's need for active most recent camera feeds and a cached archive (not as fast but available). S3x (S3 and Posix simultaneously).

From Dmitry Mishin to Everyone: (10:11 AM) https://nextcloud.nautilus.optiputer.net/s/TiKQJaKLkSNGRxW

JohnG: Have 2 new FIONA8s with 2080-Tis. Will power up today, another coming next week with different vendor and GPUs. Looking forward to passive GPUs. UCR has their 10G/FIONA8 up with ILOM. Found a MTU 1472 problem, will fix. UCI has three FIONA8s up and 1 FIONA with 192TB. Will fix the original DTN and re-install. Plan to reorg the HPWREN/ScienceDMZ uplink to Lightpath at UCI (https://sites.uci.edu/rcna/uci-lightpath/). C3 is part of the new HPWREN HA cluster, will go into the Arista with 20G asap.

Talking with PhilP about creating a backup service for PRP campus researchers. Another K8s stand-alone cluster like HPWREN, no users, just storage backup. Federate it. JimK: how about persistence of data? Say a researcher is doing analytics, done for now, but don't want to re-do the loading six months hence. Could this be used for that? John: with a SLA? Dana: with HPC on campus, NSF gave us money for compute but jobs stop. Storage is forever. James: is HPWREN's methodology available? JohnG: we're doing that with EdgeFS guys now. James: have archive needs. JohnG: Apache has new open source data lake tech. JimK: just want researchers to be able to come back to work on stuff easily. JohnG: we are not deleting anyone's storage yet. Tom: I'm talking with Phil about Disaster Recovery, not research data management. Dana: Alex's storage (OSN) and Michigan State's OSIRIS are worth considering.

Dima: 800GB of RAM and other storage at NPS connected to Nautilus (Maserati). JohnG: will install some FPGAs probably there. JohnG: 2 Nvidia Jetson nanos for inference detection with pre-trained models. Making it a K3s cluster, federate with K8s. Nvidia released an update to their Isaac framework,
allows you to 3D print a robot for the Nano, a Xavier, or two Xaviers. Good for zone controlled robotics. Main idea is for HPWREN edge TPUs on cameras--run up to 600FPS to see smoke behavior in ambient wind situations--need high frame rates that are too much data to push along wireless. Microphones already in place, could detect wind gusts, etc. Greg: have all the ultrasonic equipment at the same location as the cameras, emit over multicast stream constantly.

From James:
And I might’ve misheard did you say k3s or k3OS? I really love the latest news about the k3OS: [https://rancher.com/blog/2019/announcing-k3os-kubernetes-operating-system/](https://rancher.com/blog/2019/announcing-k3os-kubernetes-operating-system/)

JohnHess: higher bw connectivity between Nevada (UNR) and California? GregH: I asked Brian Court about that--it’s a work in progress. with CENIC.

JohnG: favorite thing: trying to get REUs--had 40 undergrads interested, 34 of whom submitted applications. Want to put traceroute viz into Unity AR. Also AI chatbots and chatops. IOT edge bits got interest too. Google has a TPU segmentation algorithm just released today. JohnHess: could this become eventually presentations for JasonZ's Friday CI Brownbag talk series? JohnG: 40% of the students were interested in NLP for rocketchat mining.


JohnG: Ciena trying to figure out how to get IP address space via L2 at StarLight--want to plumb those over (getting IP addresses) to Nautilus for P4 work. Will work with Jim Chen and Linda Winkler. Azher: Ciena has 300G into StarLight. Tom: Part of GRP/SC19 project that we are looking for partners. JimK: there are app-specific FPGA tools. Can we load them into the environment as scientist attractors?JohnG: Xilinx has K8s plug in.

Next week, May 2, lead with NRP Pilot agenda

April 18, 2019
John Hicks, JohnG, JohnHess, Dana, James, JeffH, Cathy, Azher, GregH, Kate, Jim, Stephen, Mark, Kristy, Dima, Shava, Alan, Shash, Than, Hervey, Dana, JohnD, TomD
JohnHicks: Want to work on a document that technically describes namespaces. JohnG: good docs in the site already. Dima: need portal to identify people and give them clues. TomD: Shava and Nadya working on this too. Shava: working to put a workflow into the portal and rocketchat to be more semi-automated about approving users. JohnHicks: that would be great. Noticed that Alex Feltus has a namespace and is running code. Want to point people at the Nautilus page, see cluster users, but how to sell that on a technical level--you have access to so much storage, so many cpus. Dima: we need a landing page with the resources identified. James: landing page should have examples that allow people to see how people use Nautilus. JohnG: Need to know how to read/write a yaml. Dima: pre-baked examples needed that can be tweaked. JohnG: we give a job example from the beginning. Dima: a couple of screens of yaml.

James: there is a case for deployments to share results and distributed data. JohnG: absolutely--have lots of permits for long-term runs to get waivers and protection from the kill bot. James: need to understand the difference between deployments and other stuff. Found that github can be used for bite-sized questions. JohnHicks: rocketchat not sufficient. James: start tracking byte-sized issues that people. JohnG: have nautilus cluster group. Dima:

From Dmitry Mishin to Everyone: (10:13 AM)
https://gitlab.nautilus.optiputer.net/groups/prp/-/boards

JohnG: Already a nautilus portal group which can be added to. Want Shava's work to glean out the metadata from having a chatbot discuss with a new user what campus, projects, etc., and forward the namespace admin to the new user.
Shava: my effort is complementary--when someone wants to use Nautilus, what's the process to approving and onboarding. Need a step zero first.
Stephen: I can't make a connection between JupyterHub and K8s. My own experience: people won't go through kubectl. Don't want to do CS, just science. JohnG: storage can be attached to a namespace that it can be linked to. Can execute these commands in a notebook and never leave the workbook. UCSC’s Rob Curry has great examples. We can do more in our JupyterLab host, config credentials can be in your Jupyter home directory without installing kubectl on your laptop. Dima: Igor is putting Condor access into Nautilus. Stephen: Jupyter way to go. When I hit the "download the config file" I didn't know where to put it.
James: can campus champions help? Given a JupyterHub with the right package loaded.

Dana: I'm smiling. Yes, yes yes. Great to hear that people want this now. I'm willing to help with this. JupyterHub is great for researchers. Sales pitch important--has to be worth their time. I want to gather up workflows. Claire Mizumoto and Sid Burrows-Shilling at UCSD are excellent campus champions. JohnG: new features in K8s allow throttling. Dima: can now limit the number of system processes one can spawn.
Dana: please share your stuff. Will help NRP-Pilots. JohnHicks: how to turn into an action item, turn into a coherent effort? Dana: I can be a clearing house. Shava: Nadya and I should show you what we have. Dana: great. JimK: marketing presence how to
match their research to what's available. Great topic...I want to help out too. JeffH: I will help as well. I've been going through the quickstart, well written but steep learning curve. I'm the liaison here at NPS between researchers and IT. I will participate. Dana: great. JohnG: use:

From John Graham to Everyone: (10:31 AM)
https://gitlab.nautilus.optiputer.net/prp/nautilus-cluster

Dima: There's a button to request membership. Tom is smiling too.
JohnG: good place to collect the documentation.
JohnHicks: deployment updates?
JohnG: UQ Brain Institute, Jake has 100G DTN, another box with GPUs, will work on EdgeFS federations.
JohnG: see spreadsheet for build with rough estimate of cost/node.
https://docs.google.com/spreadsheets/d/1tR9_MuDbo2xImXaAulUNRxKnb2VGEryWhlKT31kBgPE/edit#gid=1751263691

JohnHicks: remote out-of-band access? JohnG: we use IPMI with ACLs limiting access. Also using Open BMC to harden that interface. Will start putting an ACL in the BMC hardware itself--limiting range of mac addresses that can be responded to. Need to kick stuck stuff sometime, better than a human, could do power pdus, but powering down is kinda rude. EdgeFS working around a 132TB target limit with a multi-container pod. Now limited with how big a target--delaying the HPWREN project. Dima: EdgeFS will allow several targets per node. Stephen: please put up a link to your JupyterHub build.

From John Graham to Everyone: (10:45 AM)
Here are the jupyterlab bits https://gitlab.nautilus.optiputer.net/prp/jupyterlab

JohnHess:
TomD: GRP (with GLIF-Americas) Sept 17/18 followed by GNA/GLIF Sept 19/20 here at UCSD. Ciena, Jupyter helping with P4 presentations. PRAGMA is the week before, E-science the week after in San Diego.

Dana: 3NRP also the week after GRP but in Minneapolis between NSF PI and Quilt Sept 25/26. Collecting content ideas. Material to help people learn how to use Nautilus. Want to say what's going on below the Smarr level--more for technical and research facilitator level. JohnG: example of IceCube workflow. Dana: Scott Sellars too. Resonate with campus facilitators, researchers.

JohnG: getting Gigabyte servers for Xilinx, will support SC19 P4 experiments. Chris Paolini ready to run.
JohnG: UWash got the upgrade kit for their FIONA. 192TB SAS3.
Azher: Arista has 7170 with barefoot P4 chip--need to put all code on it yourself. Having issues with global intellectual property licensing. JohnG: will ask Tom Hutton to ask Arista for help.
Dima: BGP in Calico now separated from K8s. Changed to 3 reflector nodes for a smaller routing table for the cluster. Waiting for TomH to do external IPs. ElastiFlow updated--may be able to designate users. Let everyone view.


April 11, 2019
Celeste, TomD, ChrisP, JohnHess, JohnHicks, JohnGraham, Chad, Chris, Dana, Dima, Nadya, GregH, Cathy, Shava, Stephen, Alan, bmzq person, Kisty, Kate, JohnD, Jim, Azher, James, JimW, JimK, Shash, JamesM, TomH, Mark, JamesD

JohnG: working hard on the HPWREN cluster. 5th node at UCI needed to come up--it's been updated but there may have been a power failure. 2 storage servers at UCI. C3 plus the original UCI FIONA that has 192TB now. PhilP is the contact at UCI. GregH: having problems getting permissions onto UCI's ScienceDMZ.

Dima: new storage node. One user on Nautilus ran 1000s of processes in parallel, should use Node-GS or Go. Teaching him how to do parallel programming. Telling us that Ceph is slow, but they have load 1000. James: Doing system threads trashing the box?
Dima: talked to Igor about schedulers. May need to write our own for CPU jobs from OSG. Hope to get much better throughput for CPU/memory.
Dima: EdgeFS is rocking. 3 regions: Hawaii, Korea, UCSD. Want to integrate with K8s in Google cloud. Will need virtual IPs and announce with BGP. Google already supporting. TomH: deciding what address space makes most sense. JohnHess: important. Dima: originally looking at Metalb, supports L2 and L3. Calico is simpler with BGP router. Needs to be done in every region. JohnHess: need to negotiate with each site. Dima: once TomH is done and we understand it, we can advise other places. JohnHess: trying to map even within California, we have multiple connections with Google in the state, EdgeFS nodes go through Sunnyvale, etc. Dima: shouldn't matter if in one cluster or several clusters--should be able to sync and communicate, via hidden overlay network. JohnHess: still working on EdgeFS DTN storage nodes in CENIC. Use Admiralty to announce the address space. JohnG: IPv6? Dima: might be using private IPv6. TomH: none of the examples talk about IPv6. JohnHicks: we are discussing this internally in I2. Dima: may go back to Metalb, but prefer Calico at the moment. JimK: EdgeFS or Ceph in 6 months? Dima: Ceph probably better for big like 2PB. JohnG: more knobs to turn in Ceph. Dima: I like Ceph's shared file system, never blocking. EdgeFS uses stateless, never blocks, uses NFS. JohnG: EdgeFS does encryption at rest. Also S3x. Object and posix system optimized. Dima: EdgeFS's ISG interservice gateways good to synch data between region. JohnG: use external IPs to attach storage, then can mount directly. Will encourage Rook to add BeeGFS support. Dima: fixed CERN VMFS. Chris can switch to EdgeFS (although the grant is BeeGFS).
JohnG: FPGAs: call with Gigabyte and Ryan E. Will stage 6 of the gigabyte chasses in time for SC19 experiments. Discussed dual-socket Epyc ROMA mobo coming out. 162 PCIe lanes for the GPUs. 128 CPU cores. Will still make the single-socket Epyc for FIONA8s. Two U200 FPGAs coming soon. Put one at SDSU, SDSC, need to upgrade Calit2's connection to 100G (soon).

JohnG: Joel and Isaac have kickstart server in the SunCAVE for Paraview and CalVR. Google TPUs working nicely, want to integrate into Mobotix camera system that HPWREN uses. Coral dev board with edge TPU. Designing some plastic housing. Doing inference at 300 FPS. Will put inside Mobotix, instead of thermal camera slot. Need to procure one. GregH: have limited spares, can't share, sorry. Should be getting Nvidia Jetson Nanos.

JohnHicks/NRP: no items for today.
JimK: any progress with MaDDash? JohnG: no. Need to think about it. Jim: I Was on a call with Andy Lake, working with Michigan, will bring uop won dashboard. Nadya: group of grad students working on disk-to-disk plug-in, should hear in a few weeks. Dima: Shawn from NU was installing our K8s perfsonar version on their K8s cluster, successful. Have document on how to install. Also OpenNSA work with us. He's repeating our exercises. JohnG: Cilium install all new...much more simple: 2-lines for a HA cluster. JimK: going away from Globus? JohnG: plug-in will support Globus with pscheduler instead of chron. Dima: still need to filter the nodes, one node from each university max. JohnHess: disjoint would have advantages. Dima: hard to automate. JohnHess: could rotate nodes every month. Dima: will get tricky when we have new version of traceroute--need all-to-all. But don't want all-to-all for clusters, maybe once every other day. For throughput measurements, do all-to-all, see underperforming ones. If we disable cache, don't need to send 10Gigabytes. JohnHess: let's look at what size of data set is best. Different size files? JohnG: distance dependent. Could be tweaked by AI? Dima: dream is AI looking at graph and when sees underperforming segment, schedules tests, for example, on loss. Prove good or bad. Could be automated, then we don't need all to all. JohnG: working on getting REU undergrads to do AI. JohnHess: Jim Chen/ICAIR have pipeline of visiting scholars. JohnG: all volunteers welcome!

JamesM: we have a DMZ, want to use it. Stuck in the mud, will document it for others to learn.

Next meeting: April 18, start with NRP Pilot half hour.

April 4, 2019

John Hess, John Hicks, TomD, Dana, Cathy, JohnG, Kate, Stephen, Chad, Alan, Shawn, Shava, Matthias, Mark, Hervey, Chris, James, Jeff, JohnD, JimK, Kristi, Nadya, Chris
John Hicks:
Here's a draft agenda for the NRP Pilot portion of the call:

- Updates from NRP participants
- Management/Operational scaling issues.
- Updates on node deployments
  - Current dashboard status
  - [https://perfsonar.nrp-nautilus.io/maddash-webui/](https://perfsonar.nrp-nautilus.io/maddash-webui/)
- Action item review

Management/Operational scaling issues. John Hicks: at what point does this become unmanageable? # of users, usecases, etc. When will you become overwhelmed, JohnG? JohnG: two years ago? That's why we went to K8s. Will federate, will not do stuff for everybody. Just CHASE-CI, PRP, TNRP. Need to get more federated clusters talking to ours. Dana: does the tech exist to federate K8s clusters? JohnG: yes. One Slate uses that Dima wrote, FederationV2 with a SIG in the K8s tree. Also Admiralty. Supporting a lot of users. Trying to capture the conversations in rocketchat, plan to data mine it. JohnHicks: scheduling and resource allocation. Say there are more and more clusters federating with the PRP. There's got to be a level. JohnG: we have no automation for campus-to-campus policy setting for federation. Tom: described IceCube usage. JohnHicks: when this progresses to a national-scale infrastructure, scaling will be an issue.

JohnG: can set priorities for users. Coordinating job submission through K8s. Dima: integrating HTCondor now. JohnHicks: is IceCube using CONDOR? Dima: yes from outside, we will have that for the inside. Regular OSG jobs are going into Nautilus. Dana: that is cool.

Stephen: Does OSG understand SLURM? Dima: no. SLURM is not dynamic. Dana: they're working on it in ERN. Dima: I'm doing it for Comet, a pain. JohnHicks: A lot for you guys to run and develop at the same time. Trying to find the peak where it becomes too much. When the model goes national. Tom: PRP has 18 months. What then? Need to write proposals in 6 months. Dana: how do we replicate John and Dima to the regionals. Dima: two ways to go. Federating the scheduling (Admiralty) and federating the data (EdgeFS). Would be happy to have help. JohnG: EdgeFS developers want to help. Dima: they made changes we needed for global use in one day.

George Robb: Leaving Mizzou/GPN, going to work at ESnet/Berkeley, with Kate Mace, Jason Zurowski, SCInet group. Will continue to support NRP/PRP/GRP. JohnHicks: please give an update on GRP. George: we are on rev3. Our environment is unstable, good for containerization, layering K8s on top. Putting out production-quality nodes to campuses. Local admins can stand up a virtualization. If they colo, they can share resources with K8s. Add other kind of processors/processes. 3rd prototype is stable. Expect Dakotas and Nebraska to join soon. If we want to bring a national center to bear, can use Singularity to bring 20,000 cores to bear. Love the federation. Could use 6000 cores for an hour, federated. How do we get the storage friction free? GPN
moving slow but surely. 13 FIONAs deployed. Institutions getting work done. Have EPSCoR states, need to get regional network teams involved--so far, really great.

From John Hicks to Everyone: (10:23 AM)
https://docs.google.com/document/d/1DBjH-_OKIoixlnZeLTHZm6SZn6gJmRkmVnAI-P4hZwc/edit#heading=h.gi0qb2lpt8bd
From celeste anderson to Everyone: (10:42 AM)

JohnHess: U of Guam. Dashboard looks better. Sysadmins there corresponding with us, rerunning scripts. Their Science DMZ has a wide routing asymmetry but progress is being made. JohnG: want to try BBR to see if it helps (helped path to Korea), as soon as storage is stable. Dima: tried BBR 3 times. UCSD Science DMZ routers updated last night. Will try again.

JohnHicks: planning on deploying a couple more nodes placement geared toward OSG stashcache--one in Europe (need to talk about Ceph to there). JohnG: Have a node in Amsterdam--could have a regional cache. Cees de Laat has a node working. JohnHicks: maybe another in Paris, London? JeffW: Matthias did his PhD in Paris. JohnHicks: I'm working with ANA on this. JohnHess: SURFnet is helping placing a testing DTN at CERN.


JohnG: U200s should be arriving soon--going into 2U gigabyte chassis (see Spring 2019 build sheet FPGA tab).
From jjgraham to Everyone: (10:47 AM)
https://docs.google.com/spreadsheets/d/1tR9_MuDbo2xImXaAulUNRxnB2VGyWhlKT31kBgPE/edit#gid=914865956
Supermicro put out new products with Intel Optane. Puts data close to cores. Should be great for ML if they are CPU I/O bound.

Xilinx has a K8s device plug in. Development environment all set up. At SDSU, I removed T4 GPUs from the 1U chassis, put in K40s. Updated storage. Will use HA/EdgeFS instead of CEPH.

Dima: UCI back to old ACL problem from 2 weeks ago. JohnG: change is good.
JohnG: edge TPUs and Nvidia Jetson Xaviers going nicely. Putting one on a water tower here with a PTZ camera here in New Mexico. Showing TPU camea, Jetson. Prototype an in-camera AI for smoke detection. Jeff: Yosemite Connect project-please send me a link. JohnG: alertwildfire.org site. Jeff: haven't landed on a platform yet, so interested in cameras that can be upgraded with Edge TPUs. Chris: look at Lattice semiconductor...has CNN built in, free. Great for object detection, bounding boxes. EC5 board is $135. JohnG: can fit the camera/ribbon cable w/TPU stacked to do lightfield work. New toys. Chris: Sony Mx214 cmos camera. JohnG: going after REUs here. Want to use the undergrads to help with traceroute viz, chatroom bots, edge TPUs, etc.

Next meeting April 11.

This from John and Hervey to Marla this morning: as it happens Hervey and i updated & distilled the FIONA perfSONAR and GridFTP materials from the 4-day workshop in Bozeman for a 2-day workshop we held last month at Calit2 in La Jolla. here is a link to the La Jolla WS agenda:

https://ucsd-prp.gitlab.io/lajolla2019-agenda/

which has links to the presentations and lab exercises.

here is a direct link to the GitLab repo of materials:

https://gitlab.com/ucsd-prp/presentations/lajolla-2019

this is the direct link to the updated slide deck for pSConfig for MaDDash


this is the direct link to the lab exercise for setting up the perfsonar-centralmanagement bundle


and, if you are interested in other materials from the Bozeman workshop we can share pointers to those, as well.

cheers,

john
Also see JimK's slides, attached, by permission which pertain to this message from 4/1/19:

In this next phase of our participation in the National Research Platform (NRP) Pilot, we're inviting member institutions with well-suited research projects to participate. This year, we can accommodate up to six participants in total: three from large institutions and three from mid-sized/smaller institutions.

Please find some additional background information below. If you are interested in learning more about Pilot participation, please let me know. Our target date for finalizing our list of participants is end of April.

NYSERNET is participating in a pilot to grow a distributed compute environment built by University of California researchers, called The Pacific Research Platform (PRP - https://ucsd-prp.gitlab.io), into a national testbed: The National Research Platform. The technology consists of Data Transfer Nodes (DTNs), Compute Nodes, and a distributed Ceph storage environment acting in concert to enable inter-institutional high-speed research data transfers on a federated computing platform offering Intel and GPU processing capacity. Not designed to compete with multi-million core Supercomputers, this is an environment based on federated access into Kubernetes Linux containers providing a temporary virtual compute and storage environment, suitable for smaller research workloads.

Thus far, NYSERNET has deployed the DTN component of the NRP Pilot in both our Syracuse Data Center and at Stony Brook University. We're collaborating with Stony Brook on implementing a Science DMZ network and supporting research needs through the Pilot, such as improving data transfers to XSEDE via the DTN. We'll soon be exploring NRP compute possibilities as well. We've also begun collaborating with the American Museum of Natural History on an NSF-funded project to design a Science DMZ which supports NRP Pilot activities, including facilitating an Astrophysicist's data transfers with the Palomar Observatory and providing a compute platform.

We would like to invite additional institutions and their researchers to participate in the NRP Pilot as well. Due to the unique nature of its technology, not all research projects are well-suited for the Pilot. Those which are a good match for the Pilot will have need for collaborations and data transfers which rely upon the NYSERNET network to reach Internet2 and can leverage a virtual Linux compute environment. Projects requiring GPU support or discipline-specific software are welcome.

Thank you,

Jim
Toward an NRP multi-Regional Pilot: near-term and stretch goals

draft v0.05, last revised 20190131

Participants: GPN, KINBER, LEARN, NYSERNet, Internet2, PRP, CENIC

Planning Assumptions & Expectations -
Near-term goals will include outcomes which we believe have broad value and may be feasible to accomplish by (?), but may extend beyond that. In any event, we should prepare to report on progress toward near-term and stretch goals at relevant meetings.

Interconnection / Connectivity
The pilot will use the BGP-routed Layer3 connectivity among participants. Connectivity may include inter-Regional traffic traversing Internet2 R&E / AS11537, and/or “private” interconnects between Regionals. Depending on circumstances it may be interesting and feasible in the near-term to explore other interconnection methods.

Multi-protocol: both IPv6 and IPv4 are expected to be supported, with separate reporting for each

MTU: jumbo-clean, 9000 Byte
Routes tagged with NRP community so we can do analytics? (RPV)

Near-term Goals: Functional elements
NOTES: Default in **bold**, other choices may be feasible where:
- functionally equivalent
- Offers open API
- interoperable with default / baseline (through API)
- Series of containers that are connecting netflow and push it into elastiflow cluster
- Update NRP maddash to latest version

GridFTP plugin for pScheduler (in-progress at U.Michigan)
Federated Regional K8s-orchestrated clusters (Most likely based on Cilium: 
https://cilium.io/blog/2018/12/10/cilium-14-preview/ or perhaps Admiralty https://admiralty.io/)
Testing with multiple disk-to-disk tools (FDT, mdmFTP, ...)
Lookup-Service for DTNs
Repository / KB for Training & Reference Materials
Github and/or other central repository or Knowledge Base
PRP Workshop Materials Github in-progress
Could consider folding in with ESnet Fasterdata KB
FIONA build specs
Jupyter notebooks
Docker files
Other recipes

Stretch Goals
Traffic Identification - How do we find the traffic on our networks
Traffic prioritization - Is there an expectation for E2E headroom? priority? scavenger?
Measurement archive (MA)

esmond
http://software.es.net/esmond/index.html

Central, standalone. May be deployed as part of a pS toolkit bundle.

perfsonear-centralmanagement bundle
http://docs.perfsonar.net/install_options.html
Provides the dependencies for both esmond and MaDDash (see below) meta-packages as a separate, standalone installation from pS toolkit bundles.

Others: RabbitMQ
Dashboard for visualization

MaDDash
http://software.es.net/maddash/

Central, standalone. May be deployed as part of a pS toolkit bundle.

**perfsonar-centralmanagement** bundle
http://docs.perfsonar.net/install_options.html

Provides the dependencies for both esmond (see above) and MaDDash meta-packages as a separate, standalone installation.

Others: Grafana

Grids should include:

- **Throughput** disk-to-disk separate from memory-to-memory
- **Packet Loss & Latency**
  - (optional) Paths e.g. traceroute
  - Separate set of grids for **IPv6-only** and **IPv4-only**
  - Time-series graphs, a separate axis for each of Throughput (with disk-to-disk and memory-to-memory on same axis, bonus for reflecting retransmits)
  - Latency
  - Packet Loss

Traceroute Archive Viewer integration (Dale Carder / U.Wisconsin)

**DTN (Data Transfer Node)**

Minimum one, but two would be preferable for inter-Regional regular testing. More are welcome for intra-Regional regular testing.

**10Gbps-connected**. 1Gbps, 40Gbps, 100Gbps DTNs welcome

**GridFTP:**
https://fasterdata.es.net/data-transfer-tools/gridftp/

GridFTP as baseline for disk-to-disk throughput testing. Additional toolsets welcome, including FDT, mdmtFTP, nuttcp/nuttscp.

Expecting multiple, concurrent streams

**ESnet esmond-client** package:
http://software.es.net/esmond/perfsonar_gridftp.html

for parsing GridFTP transfer logs and registering results to a (central) esmond measurement archive.

**perfSONAR node**

Minimum one, two would be preferable for inter-Regional regular testing. More are welcome for intra-Regional regular testing.

pS node deployed topologically close to participating DTN(s)

**10Gbps-connected** for throughput tests. 1Gbps, 40Gbps, 100Gbps pS nodes welcome.

( optional) Separate interface for latency tests

iperf3 as default memory-to-memory throughput

Single-stream

pScheduler (as opposed to bwctl)

owping / owamp for latency & loss

Traceroute (vs tracepath)

**perfsonar-testpoint bundle**
http://docs.perfsonar.net/install_options.html

The perfsonar-testpoint bundle provides the minimal package elements required for the pS node to participate in mesh-orchestrated tests, and register results to a (central) measurement archive. Other acceptable bundles include perfsonar-core, and perfsonar-toolkit.

**Mesh configuration (regular testing test_spec parameters)**

Separate tests (test_spec) for **IPv6-only** and **IPv4-only**, as in:

```bash
ipv6_only 1; and,
ipv4_only 1

force_bidirectional 0 (do not force a given node to initiate bidirectional tests)
```

Throughput (10G-connected devices)

- Disk-to-disk: GridFTP (globus-url-copy); 10GB ESnet reference file; 4x / day; 4 parallel streams
- Memory-to-memory: iperf3; 30 second duration; 4x / day; single-stream; TCP

Other suggested default values in sample test_spec later in document
Thresholds:
Acceptable: >= 7.5Gbps
Warning: < 7.5Gbps
Critical: < 5Gbps

Latency & Loss

ICMP
Other defaults in sample test_spec later in document

Thresholds:
Acceptable: = 0
Warning: > 0
Critical: >= 0.01

Traceroute
Traceroute (as opposed to tracepath): 4x / day; UDP
Other suggested default values in sample test_spec later in document

Threshold / expectations for given pairwise test:
symmetric forward and reverse paths
IPv6 and IPv4 are congruent
A path which changes more than 4x / day is probably experiencing foo

Groups
Preference for grouping nodes by uplink speed, as in 10G-connected nodes will test against other 10G-connected nodes

Repository / KB for Training & Reference Materials
Github and/or other central repository or Knowledge Base
PRP Workshop Materials Github in-progress
Could consider folding in with ESnet Fasterdata KB

FIONA build specs
Jupyter notebooks
Docker files

Other recipes

Sample test_spec definitions
<test_spec iperf3_6h_tcp_v6>
  # Define a test_spec for testing achievable bandwidth once every 6 hours
  type perfsonarbuoy/bwctl  # Perform a bwctl test (i.e. achievable bandwidth)
  tool iperf3               # Use 'iperf' to do the bandwidth test
  protocol tcp
  interval 21600           # (21600) Run the test every 6 hours
  ipv6_only 1              # force ipv6 only
  duration 30              # Perform a 20 second test
  force_bidirectional 0     # do not force bidirectional test
  random_start_percentage 25 # randomize start time
  omit_interval 5          # ignore first few seconds of test
  window_size 134217728    # set 128MB TCP window
</test_spec>

<test_spec owamp_v6>
  # Define a test spec for doing jitter, loss and latency testing
  type perfsonarbuoy/owamp  # Perform a constant low-bandwidth OWAMP test
  packet_interval 0.1       # Send 10 packets every second (i.e.
pause 0.1 seconds between each packet)
loss_threshold    10                   # Wait no more than 10 seconds for a
response
session_count     18000                # Refresh the test every half hour
(once every 18000 packets)
sample_count      300                  # Send results back every 30 seconds
(once every 300 packets)
packet_padding    0                    # The size of the packets (not
including the IP/UDP headers)
bucket_width      0.0001               # The granularity of the
measurements
force_bidirectional 0                  # do not force bidirectional test
ipv6_only     1    # force ipv6 only
</test_spec>

<test_spec traceroute_6h_udp_v6>
# Define a test spec for running traceroute tests every 6 hours

type              traceroute            # Perform a traceroute test
test_interval     21600                # (21600) Run the test every 6 hours
protocol          udp                   # Run a UDP traceroute (could also
be 'icmp')
timeout           30                    # Wait 30 seconds
pause             0                    # Number of seconds to pause between
each traceroute probe
waittime          5                     # Number of seconds to wait for a
response before assuming its lost
first_ttl         0                    # The initial value for the TTL
max_ttl           64                   # The maximum value for the TTL
packet_size       40                   # The size of the packets (not
include IP/UDP headers)
ipv6_only     1    # force ipv6 only
as                1                    # Perform AS path lookups in routing
registries
force_bidirectional 0                  # do not force bidirectional test
</test_spec>

See
https://docs.google.com/spreadsheets/d/1tR9_MuDb02xlmXaAulUNRxKnB2VGeryWhlKT31k
BgPE/edit#gid=914865956
For Spring 2019 FIONA builds
Thursday, March 28
a reminder that we will have a PRP Engineering call tomorrow: Thursday, March 28 — 10:00AM Pacific

Join from PC, Mac, Linux, iOS or Android: https://cenic.zoom.us/j/997313186

Or iPhone one-tap (US Toll): +14086380968,997313186# or +16465588656,997313186#

Or Telephone:
Dial:
+1 408 638 0968 (US Toll)
+1 646 558 8656 (US Toll)
Meeting ID: 997 313 186
International numbers available: https://cenic.zoom.us/zoomconference?m=r7s8NEzC2mtDrdrE5OmKefYoUkP
WICLwa

Agenda
  o Bash the agenda
  o Nautilus K8s
    Rocket.chat https://rocket.nautilus.optiputer.net/
  o FPGA; (Edge) TPU
    Xilinx Alveo U200 (2x100G QSFP28, PCIe 3.0 x16):
    BittWare (Xilinx UltraScale+ 3/4-Length PCIe Board with 4x 100GbE and up to 512GB DDR4):
      https://www.bittware.com/fpga/xupp3r/
    Google Coral Edge TPU — Dev Board, USB Accelerator, … more coming soon
      https://coral.withgoogle.com/
  o Interconnection & collaboration
    NRP technical pilot (GPN, KINBER, LEARN, NYSERNet, with Internet2)
    Inter-domain SDX (Atlantic Wave, Pacific Wave, StarLight, with regional, national, and int’l partners)
  o DTNs / FIONAs / perfSONAR nodes
    John Graham’s new Fall 2018 FIONA build google sheet:
      https://docs.google.com/spreadsheets/d/1b6EzbwMB36T9ndAmCsGMqcThJDsnVItcjq2RT6RwQ2o/edit?usp=sharing
also available on the PRP website: prp.ucsd.edu —> Solution Resources tab

o Network

HPWREN
http://hpwren.ucsd.edu

PRP diagram (v0.40.16)
https://docs.google.com/drawings/d/158wd7pAWg5mUnyppJJHxGBOxP-H0UJmvTRQ70V2sGJs/edit?usp=sharing

PRP BGP pilot: route-servers, operating as PRP/AS395889 (in deployment), with a high-level diagram of two pilot sites:
https://docs.google.com/drawings/d/1xCE HPd7bPJhwnEOH3dOxpRolx8SAF4sQy9kVTHq16w/edit?usp=sharing

o Dashboards & measurement visualization

PRP Nautilus GPU Dashboard:
https://grafana.nautilus.optiputer.net/d/f_qakuSiz/k8s-nvidia-gpu-cluster?refresh=15m&orgId=1

PRP Nautilus pS Mesh - K8s pS dashboard
https://perfsonar.nautilus.optiputer.net/maddash-webui/

Dima’s Traceroute Visualization Tool on K8s:
https://traceroute.nautilus.optiputer.net/

CENIC pS (work-in-progress, 10G-connected grids reflecting results to central esmond MA)
https://ps-dashboard.cenic.net

Pacific Wave pS (very much a work-in-progress)
https://ps-dashboard.pacificwave.net

o Summary / logistics

— Upcoming Conferences, Workshops, Events —
— Calit2 events calendar: http://calit2.net/events/index.php
— CITRIS and the Banatao Institute events calendar: http://citris-uc.org/events/

— WRNP 2019: Gramado — 6-7 May
— TNC19: Tallinn — 16-20 June
— APAN48: Kuala Lumpur — 4-8 August
— GLIF Americas / GRP International Architecture, Applications, … : La Jolla
Here's a draft agenda for the NRP Pilot portion of the call:

- Agenda bash
- Discuss 90 minute call format
- Discuss federation and sharing draft presentation (John G)
  - https://docs.google.com/presentation/d/1umPnhG8MBI1bmM-GqAJ-a_5XU8_k8xhntldT3L-YYE/edit#slide=id.g46aaa3cd3f0_0_0
- Updates on node deployments
  - Current dashboard status
  - https://perfsonar.nrp-nautilus.io/maddash-webui/
- Action item review
- Adjourn

JohnHess, John Hicks, TomD, JohnG, Dana, MattH, Greg, Kate, Kristi, Shannon, Jeff, Kathy, Stephen, Jim-C, JimK, Al, Joe, Mark, Shava, James, John D, Geoff, Chris, Nadya, Chad, George,

JohnHess: GRP/GLIF and GLIF-NA meeting Sept 17-23 at UC San Diego. NRP following week in Mpls, MN.
Dana: NRP in beginning stages; talking to Wendy and JenL, putting together program committee. Jeff Weekley: I volunteer. Dana: try to have common content with NRP Pilot, NSF PI, Quilt people. JohnHess: John Hicks, use progress toward stretch goals and new goals as a basis for a session? JohnHicks: good idea. Dana: Quilt research roundtable suggestion about how people can join the NRP. JohnHess: JohnG and Dima are working on federation of clusters--talking about this/demoing would be good. JohnHicks: attached JohnG's document; I agree that we could have a demo. JohnG: we are happy to do that. EdgeFS for geographically distributed storage pools--one in Korea, one in Hawaii, one at UCSD. De-duplication and meta data coming from authors of EdgeFS to improve bandwidth utilization. We can provide simple yaml code -- admiralty examples of exposing this cluster to that cluster. JohnHess: cool. Granularity for cluster namespace resource. JohnG: stretch goal I suggest: update nrp.nautilus.io for v6 participation. Dima has enough v6 working in Calico. Hybrid both v4 and v6. Dima: no hybrid in Calico. OSG putting stashcache on host network. I figured out how to do network policy in v4 and v6. JohnG: can use Cilium to do both. JohnHess: MaDDash for v4 & v6? Dima: if map to a host network, can do this. JohnG: would like to get to a v6-only, like U Missouri. JohnHicks: this is already one of our goals.
JohnHess: What's new? Dima: really stable, but taking longer to add a new node. Policies: nagware working on pod level regarding gpu use. Switched to namespace level. Less than 10% sending an alert, less than 40%, sending a warning. Most are above 40%--may go to 60% for GPU utilization. No pod notification now. Have a pod killer. Letting users have more control, but looking after novice users who abandon jobs. All pods cleaned after two weeks, except for services that have to run all the time. Some zombies, but pods are dying for storage reasons. Jeff: can we capture the use cases? PRP science engagement team and NSF reports need this info. Dima: half of them are in rocketchat. People tell us what they are doing. JohnG: automating that. Shava: my chatbot is in early testing, looks for keywords, checks user e-mail against authorized list, then we can eventually approve them automatically. JohnG: chatbot can ask more questions. Captured in rocketchat--can give students access. Tom: Camille might be able to get students involved. JohnG: we have a namespaces page. Jeff: enforce some rigor on the namespace morphology. Link request to project proposal, title, something. JohnG: easier to just label and annotate the namespaces. Dima: labels are limited. Jeff's idea is a good one. Can give them a button to extend past two weeks, for example.

From nadya to Everyone: (10:26 AM) Current info on most active namespaces https://ucsd-prp.gitlab.io/nautilus/namespaces/ From Jeffrey Weekley to Everyone: (10:27 AM) That's awesome Nadia…but it doesn't look dynamic. Nadya: been contacting users by e-mail, very slow response, and not quality cutnpastables. Need verification beyond hubot.

JeffW: WAVE room warm. JohnG: WAVE should drain. Jeff: one node hangs. Dima: CERN VFS sometimes hangs--reboot. Jeff: when we rebooted, Ice Cube got back in before Joel could type. Dima: if not requesting GPUs, don't need to kill pods. So make sure you're requesting GPUs and Ice Cube will go away right away.

JohnG: storage upgrade to UWash. Going to SDSU to replace T4 GPUs with k40s. Tyan 1U box will not cool T4s. Ducting & bracket problems. Will put T4s into Kneuron box. Jeff: I have a $12,000 expansion chassis available if anyone wants to add cards. JohnG: use to host FPGAs. JohnG: new product available--PCIe Gen4 expander chassis. Will make Azher happy. Guys working on Pixi booting the SunCave. Will share kickstart files for auto installation of graphical CAVE nodes to simplify viz cluster building. Got confirmation that our U200 is on order. Can invite others to this program. Big Xilinx workshop here next week.

GregH: HPWREN cluster? JohnG: Centos 7.6 fully re-installed. Will replace a bad drive at SDSU, will upgrade all HPWREN storage nodes with NVMe sticks just received. C3 will be used to get to high-availability--talking to Geoff. Geoff: I have the WD drives--need to go to UCI. GregH: additional storage clusters needed? JohnG: just need one--thinking about attaching NPS machine. Need to clean up UCM path. Jeff: will go to 100G sometime,
have one 10G to HPR. If you can quantify the need, will help me advance compelling use cases, e.g., the cameras on the western Sierras. JohnG: new TPUs on Mobotix HPWREN cameras to do smoke plume inference in camera. GregH: HPWREN storage busting at the seams. Need UCM storage! Adding 10TB/month. JohnG: software side HPWREN still using CEPH, will zap and do EdgeFS with posix. Updated 3 master and 3 slave nodes. Will add C4 when Geoff is ready.

JohnHicks: every other week 90 minute session idea. JimK: the need to go beyond the hour is sporadic. Jeff: a 90 minute call is really long. JohnHicks: can we have an extra 1/2 hour every other week? JimK: sounds reasonable. Tom: let JohnHicks start every other week for max of 1/2 hours. JohnHicks: start next week. Do 1/2 hour, let PRP go for an hour if needed. JohnG: maybe need a focused call for federation issues. JimK: NRP pilot does not yet need its own call. JohnHicks: need to know what the PRP is doing. JohnH: keep PRP to 1/2 hour every other week instead. JohnHicks: sounds great, will start & go 30 minutes. PRP next 30 minutes.

JohnHicks: JohnG, please talk about federation document. JohnG: reading slides: CIlogon, Rook/Ceph/EdgeFS, K8s, Elastiflow, Grafana, etc. Speaker notes have talking points. Nautilus scale to Amazon, Google, regionals. K8s federation v2. JohnHicks: encryption (slide 14)? JohnG: comes out of the box Cilium with SDN. See https://docs.google.com/presentation/d/1umPnhG8MBI1bmM-GqAJ-a_5UX8_k8xhnqtldT3L-YYE/edit#slide=id.g46aaa3cd3f_0_0

See you next week, April 4.
..tom..
Thursday, March 21

John Hess, John Hicks, Tom D, John G, Dana, Greg, Kate, Kristi, Shannon, Jeff, Kathy, Stephen, Jim-C, Jim K, Al, Joe, Mark, Shava, James, John D,

John Hess: 18 new FIONSAR workshop graduates, thanks to Joe Keefe, Isaac Nealey, John Graham for help. Tom D will send out gitlab pointers to workshop materials, and CENIC ppts. Nadya, Dima, and John Graham had a 1-day tutorial on Sunday. Dima: went very well. Nadya started with general information and basics, I described problems I see users have in rocketchat. Also went through setting up Grafana. We showed Elastiflow, don't have public access to it yet. John Hess: is the work in the Nautilus repo? Dima: yes. James: I think there are options. Dima: maybe only in paid xpac. Amazon has open-source version.

From James Macdonell to Everyone: (10:05 AM)
https://ucsd-prp.gitlab.io/lajolla2019k8s/

From Dmitry Mishin to Everyone: (10:11 AM)
https://ucsdcloud-my.sharepoint.com/:f:/g/personal/dmishin_ucsd_edu/EsVGLQ7UK1dBntw3SOVHdABwWcfZnPU9yRI2OFvpXSXIQ

Tom: GRP, NRP details: GRP Sept 17-20 in San Diego, 3NRP during Sept 24-27 in Minneapolis, both inevitable but under construction. Other workshops on SDX and Secure PRP being planned.
James: so many success stories at CENIC, but I am stuck in the mud. Eli Dart wants to help me with Globus. Many people came to talk to me.
JohnG: been doing repairs. NTP issues. (John Hess unwrapping edge TPU). Dima and I have been balancing storage, swapping nodes. New Ceph version allows scaling of S3 gateway. Big performance increase. EdgeFS has Inter-segment Gateway--share storage between any cloud, extended even to your laptop. (John sharing slides.) Dima: lots of knobs. JohnG: evolution of ZFS work from years ago--into EdgeFS now wrapped in K8s. Pre-alpha. They had not considered our use case, but are engaged. Dima: I have already contributed to their repo. Ceph at 1GB/sec being delivered 10,000 iops (excluding Hawaiian node). Korea, Europe, Hawaii should use local Ceph. EdgeFS will also do this. i-scsei, NFS. JohnG: want posix too. EdgeFS replaces S3 in Ceph. Dima: easier to manage. Greg: ready to us in production? Dima: underlying storage tech is mature. K8s is new. John Hess: should we do this in dev cluster first? JohnG: HPWREN will be wiped and use EdgeFS. EdgeFS provides at rest encryption. Dima: can do sync encrypted or unencrypted. User interface great for configuration. (Sharing screen). Inter-segment gateway allows federation across clusters.

From jjgraham to Everyone: (10:24 AM)
https://docs.google.com/presentation/d/1umPnhG8MBI1bmM-GqAJ-a_5XU8_kxhnqtlT3L-YYE/edit#slide=id.g46aaa3cd3f_0_0
See From jjgraham to Everyone: (10:36 AM)
https://docs.google.com/spreadsheets/d/1tR9_Mudo2xImXaAulUNRxKnB2VGEryWhlKT31kBgPE/edit#gid=441990866
Found a 3-U chassis with Azher for the Xilinx U200 FPGAs, 16-core cpu, 100G ram, nice motherboard layout. Could put 3-4 FPGAs. Half-price accelerator deal. See 3rd tab. Also EdgeIoT tab. 4U Epyc DTN, Latest FIONA8, 4-core 1U designs.
JeffH: PCIe gen 4 boards appearing. Gen 5 might hit the market too. JohnG: OSS has Gen4 backplane.

JohnHicks: NRP-Pilot: I was at OSG all hands meeting all week. Good presentation on LIGO, resources needed. PRP came up quite a bit. Most interesting: Miron Livny closed out with the fact that OSG is now open. More than physics only now. Change to Open Academic Grid? Look forward to more cross-collaboration with PRP.

JamesM: At CENIC, John Delaney wanted more art. Be more inclusive.
JohnG: Stanford--Center for human-centered AI research. Our HDSI is embracing arts.

JohnHicks: looking at our dashboard, PennRen and Guam having problems. Need to follow up. Need contacts.
Having an IPv6 issue (will work with Dima).
JohnG: I will send out a list of contacts. PennRen has a congested path. We worked with Korea by turning BBR protocol which ran counters up where the problems are.
JohnHicks: Guam? John Hess: Guam’s physical connection is via Hawaii (100G) to West Coast US I2. Routed connections to U Guam, 10G peering to WRN. UH depreffing 10G in favor of 100G link into I2. Return path may take a (John Hess gave details) path.
Showing traceroute viz (which is back working now). JohnG showing it. Need to filter more.

From Dmitry Mishin to Everyone: (10:54 AM)
https://perfsonar.nautilus.optiputer.net/maddash-webui/

JohnHicks: talked to Shawn McKee, will talk to Andy about the iperf problem. Back on Guam. Traffic on one direction not good, other direction not getting any data. Timeout? JohnHess: trying to analyze. Will look in esmond. See what MA configs for the given tests are for GridFTP results. Maybe is not uploading the results. I will work with UoG on this.

JohnG: NYSERnet responding to 3 v6 nodes.

JimK: Spring 2019 build. 10G nodes fast enough? JohnG: we are using them for HA controllers in threes. Epycs don't have 10g on them, others have sfp sockets. Will hold 4 internal drives. NVMe too. Will post the builds soon on pacificresearchplatform.org under resources tab.

Next call on March 28.
Thursday, March 14

hi Rick,

for the FIONA Workshop we’re running CentOS7 and the GridFTP server provided by the Globus Connect Server package (v4.0.51), as in:

[root@gigiot-203 psconfig]# yum info globus-connect-server
Loaded plugins: fastestmirror
Loading mirror speeds from cached hostfile
* base: mirror.hostduplex.com
* epel: mirrors.develooper.com
* extras: mirror.hostduplex.com
* perfSONAR: linux.mirrors.es.net
* updates: mirror.hostduplex.com

Installed Packages
Name : globus-connect-server
Arch : noarch
Version : 4.0.51
Release : 2.el7+gt6
Size : 4.2 k
Repo : installed
>From repo : Globus-Toolkit-6-el7
Summary : Globus Connect Server
URL : http://www.globus.org/
License : ASL 2.0
Description : The Globus Toolkit is an open source software toolkit used for building Grid
systems and applications. It is being developed by the Globus Alliance and many others all over the world. A growing number of projects and companies are using the Globus Toolkit to unlock the potential of grids for their cause.

The globus-connect-server package contains:

Globus Connect Server

[root@gigiot-203 psconfig]#

thanks,

john

JohnHess, Cathy, Dana, TomD, JohnG, Chad, Shannon, Shash, Jim, Nadya, Dima, JohnHicks, Ryan, Alan, James, Kate, Kristy, Stephen, Mark, Harvey,

JohnHess: Nadya-->Jim re: DTNs
Jim: I set up a couple of FIONAs with DTN capability, one for NYSERnet in Syracuse, another in Stony Brook. To support a researcher and do longer distance tests. Want to look at K8s too. Got the 40G FIONA at SB working. Wanted to put on the NRP MaDDash. Also an opportunity to help build a prototype DMZ given the edge firewall. Exceptions that circumvent firewalls are a great concern. We can't put a read-capability anonymous FTP server with global visibility at Stony Brook. How to use GridFTP, unencrypted, but to secure the environment with ssh host keys. Allows for automation to allow esmond and MaDDash to function with GridFTP. Want to work with others to figure this out. JohnH: this is outside K8s? JohnG: this is for NRP K8s Nautilus. Waiting for Nadya and the pscheduler disk to disk. Nadya: I can do xfers in one direction. Need to figure out how to add commands to pscheduler to automate the transfer. Waiting for finalized version. Website says this is "in doing." JohnH: config hooks need to be added to central server that hosts json file. Supposed to be added. JohnG: we could modify what we're doing with mesh config yaml emitter (Dima) to add tokens, etc. if we can't get pscheduler to modify.

From nadya to Everyone: (10:09 AM) 
this is an example of 

From John Hicks to Everyone: (10:10 AM) 
https://perfsonar.nrp-nautilus.io/maddash-webui/

Jim: just want to assure Stony Brook IT that they aren't exposed. SSH-based keys for URI within GridFTP might work.

Dima: easiest way is to control access to yaml in git repo; JohnG: let's see what OSG has done to their fork of the last free version of Globus Connect Server.

John: Shash--Caltech did a great github gist file to queue jobs at SC18. Could be a private repo.

Kim: Fermilabs new js mdmtFTP transfer protocol described at the Global Summit. JohnG: I want to see a bake-off with Aria2, getting heavy use on Nautilus.

Jim: Last day Global Summit: Intelligent Network Service Plane session. Wenji Wu presented.

From Jim Kyriannis to Everyone: (10:19 AM) 
[https://meetings.internet2.edu/media/medialibrary/2019/03/04/20190308-wu-netservices-bigdataex-v2.pdf](https://meetings.internet2.edu/media/medialibrary/2019/03/04/20190308-wu-netservices-bigdataex-v2.pdf)

JohnHicks: Global Summit more technical (outside exec track). Platforms and collaborations, engaging researchers using resources emphasized, from regionals to national infrastructure. Jim and James did nice presentations too.
Jim: AMNH joining CC* NRP Pilot. AMNH has astronomer who gets Palomar time. Needs to get the image data, analyze, and then ask for what to look at next. Transfer speed xfers dependent. Now 10s of Mb/s. She wants compute access too from NRP. Need to talk to Palomar team. JohnG: use inference at the edge at Palomar—Nvidia Jetson or edge GPU. Microwave link adds latency, so make the decision there. I'm personally interested in this. Jim: yes. Anybody writing FPGA code? JohnG: Chris Paolini at SDSU, Dom Meyer here, packages available. CHASE-CI starting to support FPGAs (U200s on order).

JohnG: what's new. K8s 1.14 soon. Dima: cluster federation in Admiralty, using GPUs in Google--run it Google's way and federate it. Container service for TPUs. Admiralty guys very responsive. Expect soon. Also, Ceph updated to allow storage pools to federate better. Rancher submarine. JohnG: I've been adding NVMes for journalling everywhere. NPS fixed their border router so we can now get the paths we need. They have a restrictive ACL (like UCSF), need better way to communicate--experimental clusters having trouble. Will work on that.

From Dmitry Mishin to Everyone: (10:52 AM)


JohnH: great progress at UCI: ILOM available, 3 FIONA8s (two now, one later today online) and a 192TB FIONA. One old dead FIONA needs attention. Had an emergency request from UCSC about FIONA8 noise!

JohnH: low priority queue for GPUs. Pushing OSG GPU requests at us. Can kick out OSG jobs on SunCAVE and WAVE and they restart without admin help.

March 13, 2019: “This morning there was a big demo in the SunCAVE. The IceCube pods were kicked out automatically when theSunCAVE GPUs were in use, and restarted when the demo was over. No admin intervention needed.”—Igor Sfiligoi, SDSC
Still updates: venadi 40G FIONA at UCSC for de-commissioned Hyades cluster has been absorbed into Nautilus, previous users can use Nautilus storage.

From nadya to Everyone: (10:41 AM)
PRP setup for running globus-connect containers https://gitlab.nautilus.optiputer.net/prp/globus-connect

JohnHicks: NRP portion...
Have heard comments re:Global Summit. Anything else? James: took some liberties with PRP. Jim did a great job. Over Jimmed in that room. Kevin got to hear about all
the great things. [https://swipesforscience.org--Neuroscientist](https://swipesforscience.org) Postdoc: She gets containers at UW--created a namespace for her--she gets the idea. JohnG: introduce us! Jim: introduced her to Dana.

JohnHicks: Ajay from IU did a nice demo on L3 peering--build a network into multiple cloud providers (Azure, Google,AWS). Was netcast.

JohnHicks: old agenda item: architectural framework (being led by John and Dima). Is there any documentation? So many changes week to week: revs, features. Is there a framework diagram, or even a discussion. JohnG: I can't even imagine what to write. Everyday there's a new way to collaborate....bubbles with lines connecting them. Hard to visualize? What level of detail? JohnHicks: want to at least understand the overall emphasis. So many things, moving so fast, I don't get a sense of the direction. New things are great, but where are we going? What's important? JohnG: maybe examples of federation or sharing workflow examples. Dima: we're working with developers, go multiple ways and see what works. JohnHicks: are you driven by use cases and people, or by trying to get a new thing working? Dima: I'm driven by growing demand--if I do something twice, I automate it. Dana: is there a list of users, and what they are doing? JohnG: namespaces. Dana: not much info though. JohnG: I've been asking people to submit 1 page on what they're doing. JohnHicks: can we formalize that? JohnG: we have the hooks but haven't forced compliance. Can track the PI and the grant number. Can put in a block for descriptions, JupyterLab instances, etc. JohnH: have a group to oversee? JohnG: SLATE guys want to vet containers for us. Get known valid workflow components. Need for flexibility to discover hardened containers.

JohnHess: separate dev box cluster? JohnG: SLATE does that. Our users can break us. Dima: need to limit CPUs. One user can block all the GPUs by allocating all the CPUs. JohnHicks: what if an organization wants to bring resources? John: we approve attaching. don't require sharing, but encourage it, and want collaboration. Dima: we can take clusters off line for dedicated use. John: or like the SunCAVE with priority. Dana: OSG policy? Users can't request everything. John: OSG runs in Nautilus. Want to automate requests with AI Chatops. Stephen(UCSC): no limitations on resources. I can ask for RAM that's not available. Need to limit what people can grab. TomD: Nautilus for high-speed data sharing and GPUs (next FPGAs). CPU users should go to clouds, SC centers.

Newest version e of Singularity runs on Comet will allow us to help HPC. Singularity uses engine for containers. Dana: i'm interested in HPC. Trying to get help translating to Nautilus perspective. Where's SLURM, where does this function happen. Like Singularity more than Docker. Stephen: been pushing people toward JupyterLab interface to do ML. JohnG: Rob Currie at UCSD has nice JupyterLab that scales to S3 cluster with a button push.

JohnHicks: can I get a definitive list of contacts for each Dashboard site? Getting timeouts from some sites. Some nodes don't show up. Jim: we have two nodes on the
grid. IPv4 and IPv6 (might be inaccessible). JohnHicks: I can't get to either. Dima: we will shortly allow IPv6. JohnHicks: I'm the contact for all the OSG nodes. JohnG: my access control list should have names and addresses--I'll work on that.

JohnHess: CENIC next week. Thursday PRP call on the 21st.
Thursday, March 7
TomD, JohnHess, JohnG, Nadya, Amy, Chad, Chris, Dima, Geoffrey, Greg, John DV, Kate, Krisy, Mark, Scott, Joe, Scott, Shava, Stephen, Pavan, Jim Chen, Brad


From john hess to Everyone: (10:08 AM)
Google Coral beta dev board  https://coral.withgoogle.com/products/dev-board/
System on Module (SOM) coming later this year: https://coral.withgoogle.com/products/som/
Chris: fall detection project funded. Using Lattice ICE-40 FPGA to capture falling data--ideal place to put sensor is on shin bone. Trained a binarized neural network, uses file from Caffe, generates verilog description for the FPGA. JohnHess: Nvidia edge card might be able to accept a FPGA. JohnG: Nvidia Jetson is suitable to attaching a FPGA too.

Chris: hot topic: wildfire detection--CNN that looks at camera images looking for smoke. JohnG: we do that on WiFire cameras. Chris: package the Google TPU device inside the camera apparatus. JohnG: Greg Hidley has access to all HPWREN cameras. Chris: could do in situ inferencing. Scott: these dev boards are so cheap, why not just drop them from an airplane around a fire, send info from the field for real-time observation. JohnG: comms difficult if you want them to mesh together. Tom: UCSD group does this. Chris: can get LorA for this--I can send a presentation. Have a LorA tranceiever you bury in the ground. Good for this board, detect fungus, insects on crops. Capture images of crops, use comm net to detect with IR, etc.
JohnG: CO2 sequestration measurements too. Vernal pools study. UCM and SDSU sites have wireless.

Joe Hesse: can we have a strategy/planning conversation about Nautilus/Globus.
JohnG: Scott Sellars has a THREDDs server that pulls data in from GLADE at NCAR. Scott: new European analysis. Used Globus to get that from GLADE to THREDDs. (Talk to JohnG and Scott for details--Nadya has documented this at 2NRP--will put on website). Joe: want to expose CEPH file system in Nautilus to data producers and consumers, use GLOBUS endpoints to share with others. Want to try both models: mount ceph file system or use GLOBUS endpoint. Primarily for getting genetic data from NYC to here. JohnG: licensed GLOBUS supports data movement cluster for better data movement than one system. Personal GLOBUS connect is free. Joe: want to scale to multiple users. PRP best way to get datasets in and out of UCSF. JohnG: Could use Nexcloud as the interface to Ceph, or use Aria2 built into Nexcloud. Globus is really reliable but Aria2 is really fast (ask Scott Sellars). Joe: will start doing experiments with you John, and Matt Harrington, and Pavan. JohnDV: recommending faculty to use Nexcloud. Do we have long-term use? JohnG: I suggest you buy your own storage to attach to our cluster in case we lose funding. JohnDV: planning to do
that. JohnG: Dima and I are trying EdgeFS for hybrid local and remote storage pools--
but you want your own storage, always.

**From john hess to Everyone: (10:26 AM)**

ESnet Fasterdata listing of CLI transfer tools - including aria2, ...
https://fasterdata.es.net/data-transfer-tools/other/

JohnHess: 2-Day FIONA workshop with Herve Allen; will put new materials on PRP
website. New MaDDash esmond. PSconfig. Chris: still developing FIONettes? JohnG:
EL-20s done for, EL-30s hard to get. Several new supermicro 1U rackmount
devices. Not down to give-away range. Chris: got a grant to build machine vision
systems for traffic intersections in Chula Vista with microwave point to point
communication to pass video to fiber PoP. Would be nice to put a FIONette on each
box to run perfSONAR at each intersection. JohnG: consider Xavier platform instead.
Have master-slave mode. Chris: 3D print waterproof case. Backhaul is fiber to City of
Chula vista. Microwave form Ubiquity networks. Geoffrey: can run hypertext(?) between
Ubiquity radios, routers, switches. John: I have my own Ubiquity wireless we can
experiment. Geoffrey: mine has perfSONAR capability already.

**From Geoffrey Davis to Everyone: (10:45 AM)**

Here's a screenshot of the UNMS speed
test: http://anf.ucsd.edu/~davis/PRP/UNMS%20Speed%20Test.png

**From john hess to Everyone: (10:33 AM) ODROID links:**


JohnHess: Cas D’Angelo used Intel NUCs instead of EL-20s last fall. PerfSONAR dev
team interested in low cost small platforms.

Dima: new K8s provides block volumes/devices to pods. Expect additions from Rook
new file systems. We have two problems: have S3 gateway in Nautilus, in theory can
run 10 and load balance, but manager provides S3 metrics to Prometheus which breaks
if more than 1. Waiting for stable release to install the fix. Prolly end of March. Will
make our S3 speed faster.

John: we’ve been rebalancing nodes, caused some slowness, learning. Commit latency
down to 5ms, good. Dima: adding NVMe with journal--frontier of I/O operations, makes
Ceph happy, allows migration to slower spinning drives. Just upgraded more
nodes. Have to delete the node and then Ceph starts rebuilding, when done we add it
back, makes 3rd copy--at 2+Gb/s speed, but hogs I/O so user jobs see slowness.

JohnG: trying to analyze asymmetry on path to Korea. Starting using BBR. Went from
50Mbs to 7Gbs. Tons of retransmits, but worked. Something wrong in the path to
Korea. JohnH: working with Jim Chen. Jim-C: working with KISTI folks to debug from
StarLight PoV. John: try switching to BBR on both ends to drive error counters crazy--
might get attention. We see Seattle--Korea link problem. Happy to help. Jim: problem
for us is asymmetric too. JohnG: Something affecting all of us.
Dima: have 10 really active users, 450 people in rocketchat.
JohnG: after we put on Elastiflow, seeing storms of iPerf attacks--108Gb/s attack to UCAR, turned off iPerf, problem went away. Defaults in large mesh of systems doing other things, memory shortage causes problems. Needed to reboot hanging nodes. Have iPerf disabled. OWAMP testing enabled. Need to work with perfSONAR team on parameters. Elastiflow info really helped with this. Breaks out by protocol--really cool. Gets down to user level stuff. Will put up a bunch of screenshots since it's password protected (expert individuals contact John for the password). Dima: used Elastiflow to find problems. Threats dashboard shows Chinese IP addresses.

Next call March 14.

Bye.
..tom..
Thursday, February 28
Draft Agenda for today’s NRP engineering call.

Agenda bash
Introduce Communications team
Continue working on next set of short term and stretch goals and 2018 report.

New short term and stretch goals: https://docs.google.com/document/d/1DBjH_OKloxiIznZelTHZm6SZn6gJmRkmVnAI-P4hZwc/edit?usp=sharing

Draft 2018 goals report: https://docs.google.com/document/d/1UrQdBdWmjrNBEyLITJ8HrC2KOF49jvbKNB7Pay2cXk/edit?usp=sharing

Action item review
Adjourn

29 people: John Hicks, John Hess, TomD, Nadya, JohnG, Dima, JenL, Lincoln, Cathy, Chad, Glenn, Dima, Greg, Chris, James, Kate, Kristy, Mark, Stephen, Scott, TomH, JimK, JimC, 512, 831, Lee Ann, Alan, Bruce, Shava, Dana

JohnG: UQ (Australia) will join Nautilus soon. Will try federation experiments (DimaL Admiralty, for instance). UQ building out huge stacks of equipment. UCSF has HA clusters that we can experiment federation with. Working on IPv6--have 2 /64 addresses in operation.

John Leasure: NRP Communications group formed as a result of 2NRP in Bozeman. Pulled together communications experts--worked last fall defining two main projects. Need help with. First, NRP one-pager, what it is, how to get involved, etc. Asking for input. Goal: single mail list and website. Emerging level, leadership level being identified in a check list. Resources and planning. How to grow research support. Really want this group to help with accuracy and completeness.

John Hicks: are you taking on science engagement?
John Leasure: engagement group led by Camille. Identifying science drivers, engaging campus researchers, multi-institutional collaborations. Know from communications core group that we need input and review from users. NRP is a coalition of the willing--interesting to understand who the stakeholders are. Please help recommend places and people.

John Hicks: where will these documents live?
John Leasure: initial thought--use PRP/TNRP website. TomD: use Charley Erwin (Larry’s admin) for website support. Define TNRP and NRP Pilot scope.

John Leasure: want to better understand from this technical group: what this team will highlight as top 2-3 projects, what the current/upcoming challenges are.
TomD: I will send you our last year of PRP reports.

John Leasure: how are we capturing the stories? JohnG: namespaces page describe science in the cluster. We’ve developed patterns of using gitlab and containers for applications--very easy to replicate, simplifies workflows, gets code out there, provides secure copies and version controls, how to use cloud infrastructure (vs. classic HPC infrastructure). Look through our chatrooms for clues. TomD: I will send you a lot of apps stories, and...
From jjgraham to Everyone: (10:23 AM)
   http://ucsd-prp.gitlab.io/nautilus/namespaces/
John Hess: a lot of the apps stories are on this call!
Lee Ann: this is great! One repository is essential.
James: ditto. Hope everyone reviews and provides feedback.
Jen: please spell out acronyms, give urls, orient toward a general audience, very high
level.
JimK: we are struggling to explain to our users. This will be a huge help.
Jen: Tentative: 3NRP workshop collocating week of Sept 23 in Minneapolis with NSF CI
PI, and Quilt. Looking at hotel mechanics before we announce.
Stephen (UCSC): have you thought of taking advantage of XSEDE/campus champions?
Jen: we have that in the docs. Dana Brunson will help us with campus champions
pointers.
John Hicks: Would you like set a time to come back and report to this group? A target
date?
Jen: Getting one pager reviewed by Global Summit next week and then CENIC is the
goal. Week from tomorrow deadline.
John Hicks: come back to this group anytime.
From Jen Leasure - Quilt to Everyone: (10:33 AM)
   NPR Communications Team One-
Pager: https://docs.google.com/document/d/1HgnAQLhjswlkILcsILVifr5RUcbkr0f3uXEEb
VapmCs/edit?usp=sharing
From Jen Leasure - Quilt to Everyone: (10:34 AM)
   "What is Means to Participate in the NRP": needs attention to technical elements of
these levels. If additional levels should be added, please comment on this as
well: https://docs.google.com/document/d/1HgnAQLhjswlkILcsILVifr5RUcbkr0f3uXEEb
VapmCs/edit?usp=sharing

John Hicks: onto the NRP agenda. Current goals and stretch goals. What do other
people think we need to work on.
JohnG: NRP GridFTP nautilus.io cluster needs to be updated to IPv6. Want to do
federation experiments with several clusters.
John Hicks: how can we help, John or Dima?
JohnG: IPv6 is responding to James' IPv6 U of Mo goals.
TomH: NSF very interested in IPv6 deployment. Now that K8s supports it, move ahead.
JohnG: have 3 clusters now with IPv6 readiness. Performance will be better with IPv6
we think.

JohnG: been running several meshes (10G, 40G 100G, no WAVE or
SunCAVE). JohnG showing elastiflow, traffic by service last night. Have
hyperconverged system, so nodes chew up memory that iPerf wants, so iPerf goes after
the active node as if there's something wrong. iPerf acting like a DDOS. Had 108Gb/s
attack on one of the nodes--have turned off perfSONAR, will install other tools to
monitor packet loss, congestion (e.g., IPtables-netflow). Showing elastiflow geolocation.
Look at flows by AS.
James: this is great. We’ve been talking about this since last fall. This level of analysis is great. One of the big things: privacy concerns?
JohnG: 86.7% of our traffic was iperf before we turned it off. Out of the box.
TomD: need to work with perfSONAR folks (per NSF award deliverables) to use tools (e.g., netflow) to identify anomalies and then use iperf.
TomH: bwctl, make schedulers more aware of what K8s is doing. Iperf never meant to run on a machine that's running anything else. Never meant to be massive testing to all these boxes...just to border, maybe one hop in.
James: Peashooter targets iperf.
John Hicks: use when you see a problem.
John Hess: are there tools that are more appropriate for a cluster? Is it the underlying network?
TomH: most of UCs have policies about exporting flows. Talk to CAIDA about this. TomD: we are meeting with KC and CAIDA tomorrow, will bring up.
John Hess: Kevin Thompson suggests CAIDA/NRP collaborations
John Hicks: some sort of DTN lookup service needed? ESnet runs a perfSONAR lookup service. In our case, each node running perfSONAR is a DTN, so a directory would be helpful. Dima gave me a way to scrape the IPs for all the nodes, but it doesn't say what everything is doing.

John Hicks: moving items from our lists. Nadya, how is pscheduler coming? Nadya: Mark sent a couple of updates, still issues to fix, being sent to Ed whose students are working on the plug-in. Expect bug fixes...can transfer a file but not save it, for instance. Will follow up with Ed & students via a zoom call. Want to use at least the FTP portion.
John Hicks: testing different disk-to-disk tools. Talking to James about this. Some of the ceph literature says how to measure. John&Dima: is there a good way to measure end-to-end ceph transactions? Dima: Ceph as rich capabilities for testing. Can identify cluster-wide underperforming. Probably can't test individual files because of how they are split up. JohnG: can use redis queue and a tool called aria to write at 2GB/sec. Can use wall clock timing.
James: having trouble with feedback if file transfer fails. Dima: if caching data on clients, data is in memory. James: my pod got deployed on an I2 box in Chicago, need performance data. John Hicks: I want to compare ceph and e2e transaction to GridFTP transfer with timestamps.
JohnG: wall clock probably best tool...how big, how long. Can't see down to OSD unless it's a quiet cluster. Dima: I'll look at ceph's diagnostics.
John: been adding NVMe journaling which should improve writing a lot. Dima: slow nodes slow down everything. John: adding more nodes to upgrade other nodes. Dima: almost there.
John Hicks: What's important to NRP players? What are people asking you, what suggestion as to what we are working on. Target: 3NRP?
John Hicks: Alex Szalay spoke in Bozeman, and at Fall Quilt meeting.
TomD: also OSG HTCondor StashCache Internet2--lots of collaboration and cooperation regarding K8s, security.


JohnG: UCSF in the cluster! Very securely. Also working with Jeff at NPS on maserati (100G large memory DTN) adding to Nautilus soon. Jeff very interested in P4 scienceDMZ for SC19. Wants to buy a server of U200s FPGAs.

JeffW: been socializing use of Jupyter & Nautilus at NPS. Getting Jeff Haferman & LoPicolo onboard will help greatly with questions about persistence beyond PRP grants support.

John Hess: Jeff Wallace at NZ picking back up.

JohnG: found some fair queueing problems, flushing them.


John Hess: Alan Whinery long haul optical transport had MTUs mismatched at UH. Took a while to fix that.

Dima: can lower whole cluster to lower MTU number in calico. Have one 9000, one 1500.

JohnG: path to KISTI looks symmetrical now.

John Hess: KREONet upgrading path to Seattle from 10 to 100G. JohnG: find a tester on either side of the hop?

JohnG: shipping two NVMes to Stanford to fix ceph problems.

Also, Vince Hill at UCR adding 2 FIONA8 GPU nodes to Nautilus. UCR has a storage FIONA operating behind the Nexus--good so far.


JohnG: Ultralight node power failure, bumped wire. Will migrate into LHC Tier2 managed hardware with puppet or something like it, once Shash returns. Azher has volunteered to help. Still in Ultralight domain, improving monitoring is the goal. Allows nodes to co-exist in their infrastructure and ours, a good TNRP lesson.

JeffW: I have a ticket open with John Hess regarding flappiness on our two FIONA8s. Goes from 250Mb/s to 8Gb/s. Need to follow up. BGP routes? Observing behavior on southbound way to Riverside. Nathan here seeing low power alarms on switch when GPUs running. Will move one FIONA8 to the new research racks soon, but that really shouldn’t be the problem. JohnH: paths look symmetric, no flapping? JeffW: is this the same as the iperf? JohnG: probably. John Hess: these are tests across the cluster. JeffW: could be at the other end too. Will look at this perfSONAR00 which isn't seeing this problem. JohnG: can do some dedicated testing.

end of call 11:30.

Next call March 7.
Scaling a National Research Platform
March 2019

1. Does all content stay in the document?
2. How do we arrange so we get most important pieces in front?

Mission Statement
The National Research Platform’s primary purpose is to accelerate scientific research and discovery through building networks of campus Science DMZs that scale across geographies to create an end-to-end science-driven “big-data superhighway” system through partnerships in science engagement and technology adoption.

Background
Based on community input and on ESnet’s Science DMZ concept, the National Science Foundation (NSF) has funded over 100 campuses to deploy Science DMZs to build local big-data freeways starting in 2011. In 2015, NSF awarded a five-year, $5-million grant to fund the creation of a regional end-to-end science driven big-data highway system, called the Pacific Research Platform (PRP). Led by researchers at UC San Diego and UC Berkeley, PRP is a partnership of more than 50 institutions that includes all 10 University of California campuses, the National Science Foundation, Department of Energy, and multiple research universities in the US and around the world. The PRP’s data-sharing architecture of end-to-end 10-100 Gbps connections enables region-wide virtual co-location of data with computing resources. This regional big-data superhighway system enables researchers to move data between labs and collaborators’ sites, supercomputer centers, or data repositories without performance degradation. Today, dozens of top universities and research centers are doing work across several data-intensive fields, including particle physics, astronomy, biomedical sciences, earth sciences, and scalable data visualization research areas, positioning the PRP to be one of the most influential cyberinfrastructure coordination(?) projects in recent history.

What’s Next for National Big-Data Superhighway?
Now is the time to scale regional and national infrastructure across multiple test beds and connect state and regional research and education networks. This pilot is a multi-partner effort to drive and define the issues, challenges, and opportunities in scaling from a regional model toward a national research platform. The collaborative effort will provide new and transformative capabilities for high-performance science applications. It will offer a one-stop shop where stakeholders and users can go for support and guidance on infrastructure, process, architecture, and human networking. We will target outreach to researchers, network engineers, science engagers, funders, CIOs and CTOs, and other key audiences.

The National Research Platform (NRP) encourages a set of common activities so others can adopt activities and goals that are most appropriate for their own communities of users, stakeholders, and strategic interests. Some examples of common activities that create a big-data superhighway are offered below. These are examples only:

1. Establish minimum R&E network connection capacity
2. Deploy network performance measurement and monitoring capabilities
3. Provide or direct researchers to data transfer and/or storage capabilities (agnostic of application / protocol)
4. Designate personnel to assist in aspects of the project
5. Engage in the broader community on sharing and dissemination of results
6. Provide or direct researchers for access to computational tools and support
7. Support proactive discovery and engagement of researchers and their scientific individuals responsible for campus science drivers and collaborations (Always keep an eye out for ways to improve data transfer rates and new computational support?)

Current Activities
An NRP test pilot group was created and organized in the spring of 2018. The three committees include members that crossover and cross-pollinate the discussions: Steering Committee, Technical Committee, and Engagement Committee. A new Communications Committee was added in the fall. These groups include representatives of many national, state, and regional entities such as ESnet, Calit2, CITRIS/Banatao, CENIC, Great Plains Network, KINBER, LEARN, Merit, MOREnet, NYSERNet, San Diego Supercomputer Center, and others, as well as individual institutions and national project/program representatives.

How to Participate
Scaling the NRP is fueled by a “coalition of the willing” who are dedicated to our shared vision. We seek individuals and groups who are committed to actively participate in a national R&E effort to advance our collective ability to carry out high-performance end-to-end data transfers and storage to further scientific collaborations and discovery.

This is how to get involved:
1. Check out fasterdata.es.net to understand terms, tools, and capabilities common to R&E networking performance measurement and monitoring.
3. Familiarize yourself with tools that the National Engagement Performance and Outreach Center at IU (EPOC) has to offer.
4. Discover a regional Research & Education network in your area to collaborate with and plan for cyberinfrastructure capabilities in your region.
5. Deploy perfSONAR node and participate in at least one multi-institutional perfSONAR mesh that tests network performance
6. Engage with XSEDE campus champions, CI facilitator, and/or other campus resources to identify campus scientific research and education drivers
7. Participate in broader community conversations:
   a. Weekly PRP technical calls (open to all) every Thursday at 10:00 AM Pacific/1:00 PM Eastern.
      i. Sign up at https://mailman.ucsd.edu/mailman/listinfo/prp-l.
      ii. Also check out the PRP RocketChat at https://rocket.nautilus.optiputer.net/ and the PRP Slack Channel at https://prp-chat.slack.com.
   b. Campus Cyberinfrastructure Technical community calls and brown-bag lunches- Sign-up for cybinfo-engr@es.net mail list hosted by Jason Zurawski at ESnet.
   c. The NRP science engagement team will establish ~monthly calls soon. Stay tuned for scheduling from I2.
   d. ANdy’s Perfsonar-User mail list
8. NRP Mail List
Thursday, February 21

Agenda
  o Bash the agenda
  o Upcoming Workshops and Tutorials

  Calit2 - UCSD in La Jolla —
  — 2-day FIONA perfSONAR & GridFTP Workshop — March 16-17
  abstract & program (inc. how to participate):
  https://drive.google.com/file/d/12npiGbFzi166aHPfRSdqdlEPqTTjkeS/view?
  usp=sharing

  — 1-day Workshops Overview — March 17
  https://events.cenic.org/march-17-pre-conference-workshops-overview

  JupyterLab
  Kubernetes Policy Crafting
  CAVEkiosk and CAVEs Deep Dive

  1-day Workshop registration (via CENIC 2019 conference website)
  https://events.cenic.org/march-17-pre-conference-workshops-overview/workshop-registration

  o Nautilus K8s
    Rocket.chat https://rocket.nautilus.optiputer.net/

  o FPGA
    Xilinx Alveo U200 (2x100G QSFP28, PCIe 3.0 x16):

    BittWare (Xilinx UltraScale+ 3/4-Length PCIe Board with 4x 100GbE and up
    to 512GB DDR4):
    https://www.bittware.com/fpga/xupp3r/

  o Interconnection & collaboration
    NRP technical pilot (GPN, KINBER, LEARN, NYSERNet, with Internet2)

    Inter-domain SDX (Atlantic Wave, Pacific Wave, StarLight, with regional,
    national, and int’l partners)

  o DTNs / FIONAs / perfSONAR nodes
    John Graham’s new Fall 2018 FIONA build google sheet:
    https://docs.google.com/spreadsheets/d/1b6EzbwMB36T9ndAmCsGMgcThJ
    DSnVltcqj2RT6RwQ2o/edit?usp=sharing
also available on the PRP website: prp.ucsd.edu —> Solution Resources tab

o Network

HPWREN
  http://hpwren.ucsd.edu

PRP diagram (v0.40.16)
  https://docs.google.com/drawings/d/158wd7pAWq5mUnyppJJHxGBOxP-H0UJmvTRQ70V2sGJs/edit?usp=sharing

PRP BGP pilot: route-servers, operating as PRP/AS395889 (in deployment),
with a high-level diagram of two pilot sites:
  https://docs.google.com/drawings/d/1xCE HPd7bPJhwnEOH3dOxpRolx8SAF4sQy9kVTHq16w/edit?usp=sharing

o Dashboards & measurement visualization

PRP Nautilus GPU Dashboard:
  https://grafana.nautilus.optiputer.net/d/f_qakuSiz/k8s-nvidia-gpu-cluster?refresh=15m&orgId=1

PRP Nautilus pS Mesh - K8s pS dashboard
  https://perfsonar.nautilus.optiputer.net/maddash-webui/

Dima’s Traceroute Visualization Tool on K8s:
  https://traceroute.nautilus.optiputer.net/

CENIC pS (work-in-progress, 10G-connected grids reflecting results to central esmond MA)
  https://ps-dashboard.cenic.net

Pacific Wave pS (very much a work-in-progress)
  https://ps-dashboard.pacificwave.net

o Summary / logistics
— Upcoming Conferences, Workshops, Events —
  — Calit2 events calendar: http://calit2.net/events/index.php
  — CITRIS and the Banatao Institute events calendar: http://citris-uc.org/events/
  — APAN47: Daejeon — 18-22 February
  — Internet2 Global Summit: Washington, DC — March 5-8
  — PRP FIONA WS: La Jolla — March 16-17
  — CENIC 2019: La Jolla — March 18-20
  — WRNP 2019: Gramado — 6-7 May
Draft Agenda for today’s NRP engineering call.

- Agenda bash
- Continue working on next set of short term and stretch goals and 2018 report.
  - New short term and stretch goals: https://docs.google.com/document/d/1DBjH-OKloxiInZeLTHZm6SZn6gJmRkmVnAI-P4hZwc/edit?usp=sharingd
  - Draft 2018 goals report out: https://docs.google.com/document/d/1UrQdBDwMjprNBEyLIJT8HrC2KOF49jvbKNB7Pay2cXk/edit?usp=sharing
- Action item review
- Adjourn

Best, John Hicks

Report progress on 2018 near-term goals for the NRP multi-Regional Pilot.
draft v0.01, last revised 20190131

Participants: GPN, KINBER, LEARN, NYSERNET, Internet2, PRP, CENIC

Progress toward near-term Goals: Functional elements

Measurement archive (MA)
- esmond
  - http://software.es.net/esmond/index.html
- Central, standalone. May be deployed as part of a pS toolkit bundle.
- perfsonar-centralmanagement bundle
  - http://docs.perfsonar.net/install_options.html
  Provides the dependencies for both esmond and MaDDash (see below) meta-packages as a separate, standalone installation from pS toolkit bundles.
- Others: RabbitMQ

Dashboard for visualization
- MaDDash
  - http://software.es.net/maddash/
- Central, standalone. May be deployed as part of a pS toolkit bundle.
- perfsonar-centralmanagement bundle
  - http://docs.perfsonar.net/install_options.html
  Provides the dependencies for both esmond (see above) and MaDDash meta-packages as a separate, standalone installation.
- Others: Grafana
- Grids should include:
  - **Throughput** disk-to-disk separate from memory-to-memory
Packet Loss & Latency
(conditional) Paths e.g. traceroute
Separate set of grids for IPv6-only and IPv4-only
Time-series graphs, a separate axis for each of
  Throughput (with disk-to-disk and memory-to-memory on same axis, bonus for reflecting retransmits)
  Latency
  Packet Loss
Traceroute Archive Viewer integration (Dale Carder / U.Wisconsin)

DTN (Data Transfer Node)
Minimum one, but two would be preferable for inter-Regional regular testing. More are welcome for intra-Regional regular testing.
10Gbps-connected. 1Gbps, 40Gbps, 100Gbps DTNs welcome
GridFTP:
https://fasterdata.es.net/data-transfer-tools/gridftp/
  GridFTP as baseline for disk-to-disk throughput testing. Additional toolsets welcome, including FDT, mdtmFTP, nuttcp/nuttscp.
  Expecting multiple, concurrent streams
ESnet esmond-client package:
http://software.es.net/esmond/perfsonar_gridftp.html
for parsing GridFTP transfer logs and registering results to a (central) esmond measurement archive.

perfSONAR node
Minimum one, two would be preferable for inter-Regional regular testing. More are welcome for intra-Regional regular testing.
pS node deployed topologically close to participating DTN(s)
10Gbps-connected for throughput tests. 1Gbps, 40Gbps, 100Gbps pS nodes welcome.
(optional) Separate interface for latency tests
iperf3 as default memory-to-memory throughput
  Single-stream
  pScheduler (as opposed to bwctl)
ouping / owamp for latency & loss
Traceroute (vs tracepath)
perfsonar-testpoint bundle
http://docs.perfsonar.net/install_options.html
The perfsonar-testpoint bundle provides the minimal package elements required for the pS node to participate in mesh-orchestrated tests, and register results to a (central) measurement archive. Other acceptable bundles include perfsonar-core, and perfsonar-toolkit.

Mesh configuration (regular testing test_spec parameters)
Separate tests (test_spec) for IPv6-only and IPv4-only, as in:
  ipv6_only 1; and,
  ipv4_only 1
force_bidirectional 0 (do not force a given node to initiate bidirectional tests)
Throughput (10G-connected devices)
  Disk-to-disk: GridFTP (globus-url-copy); 10GB ESnet reference file; 4x / day; 4 parallel streams
  Memory-to-memory: iperf3; 30 second duration; 4x / day; single-stream; TCP
  Other suggested default values in sample test_spec later in document
- Thresholds:
  - Acceptable: $\geq 7.5\text{Gbps}$
  - Warning: $< 7.5\text{Gbps}$
  - Critical: $< 5\text{Gbps}$
- Latency & Loss
  - owping
  - ICMP
  - Other defaults in sample test_spec later in document
- Thresholds:
  - Acceptable: $= 0$
  - Warning: $> 0$
  - Critical: $\geq 0.01$
- Traceroute
  - Traceroute (as opposed to tracepath): 4x / day; UDP
  - Other suggested default values in sample test_spec later in document
  - Threshold / expectations for given pairwise test:
    - symmetric forward and reverse paths
    - IPv6 and IPv4 are congruent
    - A path which changes more than 4x / day is probably experiencing foo
- Groups
  - Preference for grouping nodes by uplink speed, as in 10G-connected nodes will test against other 10G-connected nodes

Repository / KB for Training & Reference Materials
- Github and/or other central repository or Knowledge Base
  - PRP Workshop Materials Github in-progress
  - Could consider folding in with ESnet Fasterdata KB
- FIONA build specs
- Jupyter notebooks
- Docker files
- Other recipes
ChrisP, JohnHess, TomD, JohnHicks, Cathy, Jim, Shannon, Chad, Kristi, JohnG, Nadya, Greg, Dima, Ryan, James, Kate, Pavan, Azher, Shannon, Scott, JohnD, Zach, Matthias, 760AC, JimChen, Mark,

JohnHess: TomD: what happened at the Quilt meeting?
TomD: learned about EPOC and TNRP complementarity even though both grants have the same set of deliverables.

JohnHess: does TNRP/PRP get involved in the data transfer for LSST via Brazil/FIU? Tom: yes. LSST is a PRP deliverable--we have a call with Julio, etc. at noon today.

JohnG: showed TPU for Jetson Xavier. Will put in the build docs.
Azher showed a gadget too. Capable of 200G--Mellanox connect x6.

JohnHess: 2-day tutorials almost full up. Tom: 1-day ones undersubscribed, will restructure them into more 1-1 sessions.

JohnG: docker exploit being addressed. normalizing the kernels now, one by one. no data loss, some hiccups. K8s bounced back. Going to UCLA and UCR tomorrow to install NVMEs. Getting a lot of use with CPU/RAM load with little GPU load--need to add on CPU nodes with lots of memory. Will add Comet nodes. Will add PatternLab nodes. Adding new FIONA8s (a 4U 8x2080-Ti 192GB RAM box dual intel scalable and a 8x1080-Ti 2U box). First benchmarking toolkit run on one 2080-Ti. Temp spiked to 80C. Had to abort. Went into fan profile, looked good, were slowly spinning up due to heat in the box. Fired it up again and held at 40C. May need to work on BIOS. Ran on 4 GPUs second run, was ok. Need to see if TPUs are available.


JohnHicks: re: runaway jobs. Is there a way to wrap policy around CPU/GPU forcing? Dima: a very rich policy engine, need to make sure each node is involved. Ceph biggest user of RAM in the cluster--hard to estimate its usage. Need to work with users to reserve what they need--one job wanted 250GB of RAM. Autoscalers exist. Maybe we can look at that. JohnG: hyperconverged nature of Nautilus allows compute on our storage, need compute-only nodes. We advertise it as a GPU cluster. OSG interested in cvmfs (Cern VM file system) drivers for K8s, Dima tweaked for newest K8s--can run low priority jobs in our cluster. JohnHicks: no interaction between cvmfs and ceph? Dima: right.

JohnHess: are there profiling tools that users can use to help in requesting resources. Dima: we have dashboards that show percent of utilization. Most are about 30%. Some
are 90%. Many don't use what they request. JohnG: Slate people have profiling. Could add hooks into gitlab to scan yamls for stupid settings. JohnHicks: can you look at the cluster as a whole? Dima: we have the data. Prometheus can show. Now we kill pods if they aren't being used. JohnHicks: can you do predictive load analysis? JohnG: we want to have undergrads work on ML on all the data we are collecting, which will help, we hope.


From Azher Mughal to Everyone: (10:20 AM)

JohnG: want to get undergrads to work on grants--TNRP traceroute viz improvement using AR/VR. CS, viz arts, humanities students. Have elasticsearch engine, can push data into other engines, pull back out. Have contacted author of elastiflow. Invited him to come into our rocketchat, we can use his backend experience with our front end VR/AR game experience. Get shared common view of our global network. James: can navigate. JohnHicks: Luke Fowler from IU and Doyle in perfSONAR were both REUs in TransPAC project.

James: group of folks in NRP Pilot talking about how to communicate. Want time in next call. Kate is a participant. Will ask fun questions next time. JohnHess: next call is Feb 28. James: will suggest it to them.

JohnG: auto renew on K8s cluster is broken. Talk to Dima to restore access to API if you are running more than a year.

JohnHicks: more bash. Dima, have you worked with federated admiralty? Need to start that. JohnG: have a /64 IPv6 working on High Availability HPWREN and our cilium-
based HIPAA-HA secure cluster. Pavan working on calico version. 500 physical cores on his HA cluster at UCSF. testing and rebuilding.

From Azher Mughal to Everyone: (10:20 AM)
Will give another 30% CPU cycles if you are running in a controlled environment. This is at the boot level.

JohnG: IPbv6 coming quickly. HIPAA HA--ACLs open to UCSF, need to attached UCSF DTN node into Nautilus.

JohnHicks: out of time. Can we do a DTN lookup service? Is there a plan, or a way? Get IP addresses, name, inventory of resources. JohnG: beyond Globus community? JohnHicks: is a stretch goal. JohnG: can describe how to spin up a DTN node with Globus, FDT, our cloud. Caltech did an automated service during the BW challenge. Scott Sellars using ARIA2 on thredds server, built into Next Cloud. James: way cool--saw it on the rocketchat. JohnHicks: show someone how to setup--how do they know other end? JohnG: node selector--wiki has every node, its storage (talk to JohnG). Could have gateway to Chris's 2PB. Ryan: is there a list of IPs so I can track traffic. Dima: getnodes from the PRP cluster. James: go from globus directory to get beyond Nautilus. Globus will get the majority of the DTNs, I2 community-wide. James: on OSG side, local data movement not using DTNs--xtransfer stats. JohnG: globus and FDT have gone to a generic log file. James: HDCondor needs to do that. JohnHicks: your cluster is a GPN cluster, not PRP. James: need to provide cross cluster traffic. JohnHicks: how much traffic goes out of your cluster. James: most of it. Thanks to PRPs work, it's easy to run GridFTP as a container. JohnG: having all the regional NRP data will be a pile. Can use nagios to access. Create the rules to wrapper that.

JohnHess: mapping resources and doing resource recovery (ESnet's SENSE project). SENSE orchestrator has an API. Inder Monga is the PI. CENIC/Pacific Wave is helping with dynamic provisioning at Sunnyvale and LA for ESnet and Caltech.

JohnHess: extended call on 2/28
Thursday, February 14
(no Minutes)

Thursday, January 31

JohnG, JohnHess, JohnHicks, TomD, Chris, Matthias, Kristy, Hans, Cathy, Jim, Joe Hesse, Joe Hill, Kate, Scott, JohnV, Jeff, Azher, Mark, Greg, Alan, Ryan, TomH,

FPGA item:
JohnG: P4 Xilinx U200 boards dual-100G QSPF ports: special accelerator program for 2 cards at $3,995 each. Have K8s, Jupyter notebooks, easy to integrate into Nautilus. Contact Tom <tdefanti@ucsd.edu> for further details and intro to Parimal. Cards can be used for ML--Have tensorflow use. JohnHess: use for NDN? JohnG: yes--AI on the core for P4 Hyperchannels. Product designed for data centers. Chris: I have P4 working with SDnet, API for C/P4 coding. SNORT pattern matching to scan packets for viruses. JohnHess: Reminds me of TechX ESnet presentation from Maryam Farahmand. Chris: I have articles on training FPGA-based network intrusion with RNNs. Jeff: I want to see those. We want to implement a Bro cluster--seems overlap here. Bro clusters start $25-$40K. Is there work that can be leveraged? Chris: yes--use the card as a K8s accelerator. We can offload linear algebra computations onto Xilinx card. BTW, I put in a link to the P4 whitebox switch (Edge-Core, program data plane ASIC). John: put them where we have 100G ports, e.g., StarLight. Chris: Intel stratus10 board is new. John: Intel Nauta has hooks into Nirvana. Needs its own cluster, not added to existing cluster. Git repo very active--probably will address that. Xilinx Vivado synthesis tool can be run in client-server mode. JohnG: we can spawn a container.... JohnHess: Joe Mambretti says Julio and Geronimo have NSF money to explore P4. Chris: also PISA devices (barefoot tofino) supports P4. JohnG: Dima right now at webinar regarding Arista and P4. Jim: would like to see PRP/NRP offering FPGA help? JohnG: this U200 card doesn't require hands-on--it's data-center oriented. Chris: FPGA's good for inferencing at the edge--low powered systems for ML at the edge. GPUs use 10x as much power. E.g., Iteris cameras at traffic lights. (Spelling might be wrong--Tom). JohnHess: will add FPGA to the standing agenda.

From jjgraham to Everyone: (10:23 AM)

From Christopher Paolini to Everyone: (10:05 AM)
From Azher Mughal to Everyone: (10:10 AM)
https://www.bittware.com/fpga/xupp3r/ Another very powerful board I have used this to work along with DPDK

From Jeffrey Weekley to Everyone: (10:28 AM)

https://events.cenic.org/march-2019/program

John Delaney is the keynote speaker.

Hotel room block closes soon. Pre-workshops on March 17: CAVE/WAVE, K8s/Grafana, JupyterLab, and 2-Day perfSONAR/FGridFTP workshop (not in program--contact Tom).

Please get the word out. More content than ever--very high quality. Not updates, new content.

John Hicks: Global Summit precedes CENIC conference. Dana, Celeste will attend. Agenda is already set.

Jim: Quilt ESnet meeting next week in La Jolla Shores.

Tom D: no plans for a third NRP yet. Committee discussing stand-alone or combined with another conference.

Tom H: Security conference NDSS is in San Diego Feb 27.

John G: have 3 high-availability nodes to join to Nautilus (needed to update K8s 1.13). Our 1U 10G machines don't recognize Arista-branded AOCs. Dima has updated the storage. See https://wiki.nautilus.optiputer.net/index.php/Cluster_Map

Need to upgrade USC, Stanford, UCR still. Jeff: I'm rehabbing FIONA-01, could have a big block in the Ceph Pool. John G: look at the MaDDash--we are seeing some problems--packet loss and retransmits--maybe some fiber cleaning?. Need a different port on the Science DMZ to compare. Jeff: Use WAVE. Also 2 new FIONA8s. I have 2 fiber cleaning kits.


John G: learning more and more about NextCloud. Scott will pull data out of Glade in Cheyenne. Scott: set up a globus container in Nautilus to connect to their globus at NCAR. John G: Also looking at shared ceph file system for MIRRA2 archive, config as personal globus connect endpoint. Need to automate more. NextCloud will help. It can have multiple back-end file systems (Ceph, AWS S3, etc.). Will update NextCloud to use Cilogon. Also doing issue tracking inside of Gitlab. Gitlab has a service desk feature we will deploy to help with support. Deep gitlab automation for K8s integration helps avoid typing errors. Step towards chat-ops automation. Still having issues with Caltech's ultralight node---their ScienceDMZ routing proxy been flaky (on a 10. network). Tom H: can we extend a vlan into Caltech? Can't get to docker.io--has to go though a proxy.CENIC can't see parts of the network--maybe it's a CENIC problem. Shash gone for a couple of months, I hear. John H: Multiple ASNs (ultralight has its own), how react to CENIC HPR vs CENIC DC. Azher: we had 2 sets of peerings, but probably has been changed. Shash said there was some DNS issues. I can try to help Harvey. John G: let's talk. Tom H: LHC1 routing might be getting in the
way. JohnHess: we worked through the LHC problem a few months ago. TomH: might be similar issues.

JohnG: working on UCSC 100g Soledad path--still some asymmetries. TomH: need to do TCP dump.

From jjgraham to Everyone: (10:55 AM)


5 horizontal orange blocks show the problem. I'll start iperfing.

John Hicks: Draft Agenda for today's NRP engineering call.

- Agenda bash
  - JohnG: will have elastic flow done soon.
  - JohnHicks: perfSONAR dev group needs info on the grafana, rabbitmq, etc., JohnG: MaDDash pulling data out of Esmond timing out if too many cells in the dashboard. JohnHicks: that's Andy Lake's.
- Introduce Dana Brunson (Executive Director for Research Engagement, Internet2)
  - Will be active in this group.
- Continue to work on next set of short term and stretch goals and 2018 report.
  - New short term and stretch goals: https://docs.google.com/document/d/1DBjH-OKIoxiInZeLTHz6SZn6gJmRkmVnAI-P4hZwc/edit?usp=sharingd
  - Draft 2018 goals report out: https://docs.google.com/document/d/1UrQdBDwMjprNBEyLIJT8HrC2KOF49jvbKNB7Pay2cXk/edit?usp=sharing

No call next week due to Quilt meeting. Next call is Feb 14.
Thursday, January 24

- Upcoming Workshops and Tutorials

  Calit2 - UCSD in La Jolla —
  — 2-day FIONA perfSONAR & GridFTP Workshop — March 16-17
  abstract & program (inc. how to participate):
  https://drive.google.com/file/d/12npIGbFzi166aHPfRSdgqIPEpTTjkeS/view?usp=sharing

  — 1-day Workshops Overview — March 17
  https://events.cenic.org/march-17-pre-conference-workshops-overview

  JupyterLab
  Kubernetes Policy Crafting
  CAVEkiosk and CAVEs Deep Dive

  1-day Workshop registration (via CENIC 2019 conference website)
  https://events.cenic.org/march-17-pre-conference-workshops-overview/workshop-registration

- Nautilus K8s
  Rocket.chat https://rocket.nautilus.optiputer.net/

- Interconnection & collaboration
  NRP technical pilot (GPN, KINBER, LEARN, NYSERNet, with Internet2)
  Inter-domain SDX (Atlantic Wave, Pacific Wave, StarLight, with regional, national, and int’l partners)

- DTNs / FIONAs / perfSONAR nodes
  John Graham’s new Fall 2018 FIONA build google sheet:
  https://docs.google.com/spreadsheets/d/1b6EzbwMB36T9ndAmCsGMgcThJDSnVltcqjZRT6RwQ2o/edit?usp=sharing
  also available on the PRP website: prp.ucsd.edu —> Solution Resources tab

- Network
  HPWREN
  http://hpwren.ucsd.edu

  PRP diagram (v0.40.16)
  https://docs.google.com/drawings/d/158wd7pAWq5mUnyppJHxGBOxPH0UJmvTRQ70V2sGJs/edit?usp=sharing

  PRP BGP pilot: route-servers, operating as PRP/AS395889 (in deployment), with a high-level diagram of two pilot sites:
o Dashboards & measurement visualization

  PRP Nautilus GPU Dashboard:
  https://grafana.nautilus.optiputer.net/d/f_qakuSiz/k8s-nvidia-gpu-cluster?refresh=15m&orgId=1

  PRP Nautilus pS Mesh - K8s pS dashboard
  https://perfsonar.nautilus.optiputer.net/maddash-webui/

  Dima’s Traceroute Visualization Tool on K8s:
  https://traceroute.nautilus.optiputer.net/

  CENIC pS (work-in-progress, 10G-connected grids reflecting results to central esmond MA)
  https://ps-dashboard.cenic.net

  Pacific Wave pS (very much a work-in-progress)
  https://ps-dashboard.pacificwave.net

o Summary / logistics
— Upcoming Conferences, Workshops, Events —
  — Calit2 events calendar: http://calit2.net/events/index.php
  — CITRIS and the Banatao Institute events calendar: http://citris-uc.org/events/
  — APAN47: Daejeon — 18-22 February
  — Internet2 Global Summit: Washington, DC — March 5-8
  — PRP FIONA WS: La Jolla — March 16-17
  — CENIC 2019: La Jolla — March 18-20

Chad, TomD, JohnG, Hans, Chris, Shash, Nadya, Kate, Kristy, Shannon, Hervey, JohnHicks, JohnHess, Ryan, Matthias, Scott, Mark, James, Dima, Joe, Celeste, JohnV,

From Christopher Paolini to Everyone: (09:59 AM) regarding "data center" FPGAs
Also Intel Stratix 10. JohnG: Intel has new K8s/Nirvana deep learning framework. Chris showed $199 FPGA board with CNN core. He also has $49 ones for distributed sensors.

From Christopher Paolini to Everyone: (10:08 AM)
Announcing: The Xilinx Alveo Device Plugin for Kubernetes (K8s)  
* Do you have familiarity with containers?  
* Do you already use docker?  
* Do you have the need to scale up containers from fewer servers to many more with minimal monitoring of nodes?  
If you said "yes" to the 3 questions above, you might be interested in using our Kubernetes Alveo Plug-In!  
Our plugin is in open beta and can:  
* Plugin automatically detects Xilinx cards via XRT  
* Run FPGA accessible containers in the k8s cluster  
* Keep track of the deployment and repair itself as needed  

External Github page:  
https://github.com/Xilinx/FPGA as a Service/tree/master/k8s-fpga-device-plugin/trunk  
If you have further questions or wish to include a customer in our beta program, please contact Melissa Sussmann via email: m\sussman@xilinx.com<mailto:msussman@xilinx.com>

JohnG: storage upgrade and FIONA8 UCSB/UCSC/UCSF/UCM run with Tom last two weeks. UCSC: added 20 12TB drives, lots of NVMeS, and another FIONA8 (8-GPU) (they have 3 FIONA8s and a FIONA4). At UCM, we put their 20 WAVE GPUs on Nautilus, NERSC 8-GPU, and a new FIONA8, and upgraded NVMeS. At UCSF, put in 12TB FIPS-encrypted drives for a security From Dmitry Mishin to Everyone: (10:31 AM)  

We're up to 232 GPUs in Nautilus and >1PB  
(see https://grafana.nautilus.optiputer.net/d/KMsJWWPiz/cluster-usage?orgId=1 )  
K8s newness: Dima upgraded to 1.13.2, need to update all nodes.  
Dima: new Ceph. Support newer kernels, enable rebalancer. Support more features (more filesystems, gateways). 600 parameters can be tuned(!), new update tunes them automatically. Playing with perfSONAR. Tuning Cassandra.  
JohnG: running into scalability issues with merging dashboards.  
Dima: for Calico, once we hit 50 nodes, a proxy is recommended. Did that, so can scale.  
JohnHicks: What didn't work? Dima: Edgar told me that performance was not good over the past years--they were using two databases--cassandra (good performance for write but not for read) and postgres. Apache is timing out. JohnG: could be stress on the network--it's leaving a lot of perfSONAR processes running. If only 32G of RAM on node, get delays.  
Dima: when apache building MaDDashes, says no data, but the data is there. JohnHess: NetSAGE uses Grafana.  
Dima: OSG has nice Grafana.  
JohnHess: Andy Lake moving away from graphing themselves, using grafana.  
JohnHicks: I'll reach out to the dev group, Andy and Mark.  
Dima: started using gitlab for all my issues/communications.  
Nadya: will create a new perfSONAR container offline before pushing it into gitlab. Setting up Globus personal endpoint, one active at a time, unless you have paid

JohnHicks: NRP:
From John Hicks to Everyone: (10:34 AM)
  Short term and stretch
goals: https://docs.google.com/document/d/1ZnL5Ksd6KYQhD6355zag_Z3awEmSnGCKR0bQmhdqNE/edit#heading=h.gi0qb2lpt8bd

James, GPN update?
James: Briefly, we have a similar group (Kate, Shannon, George Robb) meeting every Monday, HPC directors, campus champions, state networks staff, campus IT people. 1-2 dozen every week. CC* grant season. Tech resources match maker function. Have 13 GPN-funded FIONA nodes, plus one for the NRP. Making connectivity happen. Operating mini-clusters, getting nodes back into the cluster. Important to document what’s happening. Inspiring others. See:
Shows outputs of some of the discussions. Asked GlobalNOC to share data. Can base broader inquiries for E2E performance with traceroutes. Can see I2 and the RONs up to the campuses. Having the NRP discussions has motivate all of this.
JohnG: Nagios (www.nagios.com/) aggregators enabled, port enabled. Also have elasticsearch flow.
JohnHicks: James, are you using cilium? James: letting JohnG and Dima do this first. Been cautious with network overlays. Multis what the Measurement Labs folks use on their global K8s cluster--very substantial. See https://www.measurementlab.net/
JohnG: also Admiralty.
From Dmitry Mishin to Everyone: (10:48 AM)
  https://admiralty.io/
James: let me know if you run into anything not working.

JohnG: trying to get gitlab ultimate license.
Dima: most active service. Hard to update, but will do. JohnG: will migrate PRP wiki into gitlab wiki.

JohnHicks: will do a rough update of the goals list for the Quilt meeting in two weeks and then formalize more for the I2 GS meeting.
Will talk to Wendy and Jen at the Quilt meeting about a 3NRP.
Changes on interconnectivity? All L3.
JohnHess: discussion of High Performance DMZ?
JohnHicks: part of the discussion for NRP Pilot and OSG.
JohnG: add in BGP science domain tagging. UCSC scienceDMZ traffic reaching other sites (e.g., I2 sites) is very asymmetric. Looking for scienceDMZ tag? JohnHess, is there a BGP document. JohnHess: have discussed reserving a tag for scienceDMZ. Separate pipe for CC* scienceDMZ traffic. Gives a hook to distinguish the route.
JohnHess: Couple things for James and other CC*NIE funding--how often has that resulted in a different tagging approach.

In GPN, are there campuses with separate connections for scienceDMZs?

In GPN, state networks don't differentiate, but some campuses have separate connections for scienceDMZs, but they are mixed back in at the state level. Lots of logistics questions regarding routing. JohnHess: in CENIC's HPR network, which is separate, the campus tags scienceDMZ. James: just on the campus/isat mile? JohnHess: Joe Mambretti talks about superchannels. Honoring the community and passing it forward to differentiate traffic. JohnHicks: I2 routing report will be released, likely recommending honoring community strings to leverage--would be a fantastic thing; NRP should discuss this. Have discussed this with Shannon. James: Very valuable going forward for the larger number of regional connections. Each day more direct connections outside I2/ESnet.

JohnHicks: how can we move forward with a document to present to I2, others? JohnHess: I can co-lead that. JohnG: should include Tom Hutton. James: JohnHicks, do you agree? JohnHicks: yes, I will reach out to James, Jeff, JohnHess, Tom Hutton, JohnG, start the discussion. JohnG: move this into the short term goals to help with 100G asymmetries. JohnHess: a better understanding of our routing topologies will help.

JohnHicks: v4 & v6? JohnG: trying to get our v6 block plumbed to us by Tom Hutton, then we will have a v6-only cluster, along with all the name and git services, yum repositories. Still dual stack in the short term. George Robb at U Mo wants to be v6 only.

Shannon: direct v6 peer over GPN. Have scienceDMZ at 100G without affecting the campus resources.

JohnHicks: everyone is using jumbo frames. Dima: storage suffers badly when not jumbo frames (like when a switch is nuked).

JohnHicks: functional elements, lots already done. James suggestion from last week pushing netflow data to elastiflow containers--do we agree on this, someone to lead the effort? JohnG: we containerized netflow and sflow. Can be deployed on any node. Question: is this what Jen's group wants? Can send to Nagios. James: I agree. For NRP pilot, get through steps of policy for traffic going thru FIONA nodes--get into elastiflow to analyze. JohnG: can give campuses a window into what's happening. JohnHicks: providing data, filtered to go to certain people only. JohnG: policy in the form of code. JohnHicks: who wants to own the policy organization around this technical capability. James: JohnG, can you set something up so the policy people can work on it? JohnG: we have documented our policies internally on gitlab. All configs there. Can give people access to control their own chunk. Others write policies, I will handle the making the data available part, wait for complaints rather than ask for permission. James: sharing your containers will help a lot to fire up the policy people. JohnG: the HighAvailbility 3-node master cluster is clonable. HPWREN also a small cluster mostly storage with increasing compute. Better than all of Nautilus because of the services offered. Dima: 2 kinds of policies: K8s does quotas--set of yamls, and then there are active code policies.
JohnHicks: need policy lead after JohnG does tech.  James: I'll do that.  JohnG: will open it up, then dial it back. Need to know if anyone gets angry and fix it.  JohnHicks: next things mostly already done, claim success.
1. MA scaling issues, other data base tech.  JohnG: need to update NRP cluster to have latest perfSONAR.
JohnHicks: will add to short-term goals. For NRP nodes, each one has a version of K8s managed perfSONAR container reporting to the MA. Each node has a separate code outside K8s running GridFTP transfer heartbeat reporting to a separate dashboard.  JohnG: Not all NRP nodes (just I2 ones) currently in Nautilus. No GridFTP in K8s--Nadya working on that. Nodes can run GridFTP and perfSONAR in containers, but we don't run GridFTP. The three I2 nodes do GridFTP outside K8s. Dima: We don't do GridFTP in big MaDDash.  JohnG: Waiting for Michigan team to supply the plug-in--John Hicks gave Dima a new link for it, and Nadya is working on it. We have a few legacy GridFTP nodes and the I2 nodes and SCIdas nodes.

Thursday, January 17
Notes 1-17-19
JohnHess, TomD, Cathy, Shannon, Kristy Hamm, Hervey, Shava, JamesD, JohnHicks, Dima, Jim, JohnD, George, Kate, Mark, Ryan, Lincoln, Mark, 773 AC, TomH, 404 AC, JoeHesse.

Dima: Nautilus--now at 1PB. Ceph shuffling the data to balance it. Lost one node yesterday, hope we didn't lose anything. New nodes have fast journal devices and nvmes, big and fast now. Once migration is done, will have headroom. Recent updates in Rook have fixed issues. CephNFS users can mount now. Last bug is the K8s affects using large data sets in Jupyter. Added more GPUs (now 217). We are nagging & bugging users who waste GPUs (don't release them when job over) so they learn. Modified MaDDash OWAMP to show SunCAVE and MercedWAVE separately. Solved certificate for HPWREN on my side.

JohnHess: info posted about CENIC workshops. Opening FIONette workshop to XSEDE people.

CENIC Events overview page- https://events.cenic.org/2019
Link for workshop registration: https://events.cenic.org/march-17-pre-conference-workshops-overview/workshop-registration

John Hicks: Dana Brunson joining I2, had leadership role in XSEDE, taking John Moore’s place.
Short term goals and stretch goals for NRP Pilot, and the stretch goals:
From John Hicks to Everyone: (10:14 AM)

https://docs.google.com/document/d/1ZnL5Ksd6KYQhD6355zag_Z3awEmSnGCCKR0bQmhdqNE/edit#heading=h.gi0qb2lpt8bd
From John Hicks to Everyone: (10:14 AM)
Scaling
- https://internet2.app.box.com/notes/302913622687?s=2624I5bruejbh8hypad7phke5ggs1xz
Going through the goals document today:

Planning Assumptions & Expectations -

- **Near-term goals** should be accomplished in time to provide a minimum of two-weeks worth of data (test results) by the 2NRP Workshop (August 6-7, 2018). Stretch goals will include outcomes which we believe have broad value and may be feasible to accomplish by the 2NRP Workshop, but may extend beyond the Workshop. In any event, we should prepare to report on progress toward near-term and stretch goals at relevant meetings beyond 2NRP (I2TechEx, ...)

John Hicks: Quilt coming up in San Diego early February--can get updates regarding
NRP3 maybe this fall (JenL and Wendy)? John Hicks: Global Summit too close now. Jim: we are focusing on campus side first at NYSERnet. People struggling with using the bw, tuning, for sure. We want to identify the research and use cases, short-term DTN installs, later will provide K8s clusters. John Hicks: how do you organize this campus side activity? Jim: our board members are the willing CIOs of the institutions, interested in NRP. Working group supports sending out a 1-2 pager solicitations to our members. Asking campuses to provide researchers & projects. Making sure there is NYSENet connectivity, or perhaps put equipment in our data center if bw isn’t available on specific campuses. Hoping to have a dozen projects, 4-5 would be good to start. Stony Brook has a FIONA/DTN, American Museum of Natural History wants one. GridFTP not enough--need other data transfer protocols. TomD: try FDT. TomH: sometimes you have to tune Globus. Will pass on what we find out. Jim: good short term goal to experiment with other xfer software. TomH: look at front-end packages. SURFnet/SARA wrote a web-based wrapper over FDT to make it easy for users. TomH: user services folks to memory-to-memory globus transfers to see if it's the storage system. Jim: GridFTP our defacto standard. Would like to see the data repositories offer other services that operate faster than GridFTP. John Hess: test with new tools on engagement side.

John Hicks: form a subgroup to work on these topics? Set up a way to perform the tests, evaluate the ease of use, publish the results. John Hess: IRNC SDX group (StarLight, PacWave, CENIC) has had these discussions; could help form critical mass. Jim Chen/Joe Mambretti, me. Geronimo (Ampath), Shawn Donovan (GA Tech), and UCSD folks. We meet every other week. I can update this group, can add folks from this group to the discussion. TomH: also a task for the XSEDE data transfer group. John Hicks: let's make coordination a goal. Dima: have GridFTP plug-in and perfSonar co-existence solved. John Hicks: want to see if K8s interferes sometimes. Dima: haven't looked at this yet. John Hicks: have reached out to U Mich, no response so far. John Hess: bare metal vs. containerized--go back to original work with Fermilab, ESnet and see if they want to address this in tests. Containers/K8s shouldn't be a barrier. But we don't want to exclude bare-metal testing. John Hicks: trend seems to be toward containerization, but it shouldn't be a barrier. Need to find out if K8s slows things down. TomH: good to do some tests. John Hicks: also data transfer portal interface ease of use needs to be evaluated in the context of bare metal vs. K8s. James D: one of the short term goals would be for K8s people to agree on a data transfer methodology. Jim: run netflow? James: yes, a series of containers that collect netflow data, throw it into an elastic. TomH: UC would want flow data anonymized. CAIDA could advise. XSEDE sites have various requirements for data collection. John Hicks: I2 anonymizes data--can get to an organization, not to a host. TomD: we give resources to people and they expect us to measure. John Hicks: Jen Schopf has some experience with netflow and internationals. Could restrict ourselves to DTN to DTN.

- **Interconnection / Connectivity**
  - The pilot will use the BGP-routed Layer3 connectivity among participants. Connectivity may include inter-Regional traffic traversing Internet2 R&E / AS11537, and/or 'private' interconnects between Regionals. Depending on circumstances it may be interesting and feasible in the near-term to explore other interconnection methods.

- **Multi-protocol**: both IPv6 and IPv4 are expected to be supported, with separate reporting for each.

- **MTU**: jumbo-clean, 9000 Byte

From Kate Adams to Everyone: (10:20 AM) Were there 2 weeks of test results by August 6-7? John Hicks: I believe so...we are capturing data in perfSONAR and GridFTP.
From John Hicks to Everyone: (10:29 AM)

From John Hess to Everyone: (10:32 AM)

From George Robb to Everyone: (10:39 AM)
http://mdtm.fnal.gov/Releases.html (looks like there is a "docker" image) rapid testing in the K8 ? :) From Dung Vu to Everyone: (10:41 AM) Thank Dima for joining CSUSB's DTN back to cluster after CENIC fixed the MTU for new hardware

From Kate Adams to Everyone: (10:56 AM)
I added James's idea in the first block under near term goals, move if necessary.

Near-term Goals: Functional elements

NOTES: Default in bold, other choices may be feasible where:
• functionally equivalent
• Offers open API
• interoperable with default / baseline (through API)

Measurement archive (MA)
• esmond
  • http://software.es.net/esmond/index.html
• Central, standalone. May be deployed as part of a pS toolkit bundle.
• perfsonar-centralmanagement bundle
  • http://docs.perfsonar.net/install_options.html
  Provides the dependencies for both esmond and MaDDash (see below) meta-packages as a separate, standalone installation from pS toolkit bundles.
• Others: RabbitMQ

Dashboard for visualization
• MaDDash
  • http://software.es.net/maddash/
• Central, standalone. May be deployed as part of a pS toolkit bundle.
• perfsonar-centralmanagement bundle
  • http://docs.perfsonar.net/install_options.html
  Provides the dependencies for both esmond (see above) and MaDDash meta-packages as a separate, standalone installation.
• Others: Grafana
• Grids should include:
  • Throughput disk-to-disk separate from memory-to-memory
  • Packet Loss & Latency
TNRP Annual Report Year 1

- (optional) Paths e.g. traceroute
- Separate set of grids for IPv6-only and IPv4-only
- Time-series graphs, a separate axis for each of
  - Throughput (with disk-to-disk and memory-to-memory on same axis, bonus for reflecting retransmits)
  - Latency
  - Packet Loss
- Traceroute Archive Viewer integration (Dale Carder / U.Wisconsin)

DTN (Data Transfer Node)

- Minimum one, but two would be preferable for inter-Regional regular testing. More are welcome for intra-Regional regular testing.
- **10Gbps-connected**. 1Gbps, 40Gbps, 100Gbps DTNs welcome
- GridFTP: [https://fasterdata.es.net/data-transfer-tools/gridftp/](https://fasterdata.es.net/data-transfer-tools/gridftp/)
  - GridFTP as baseline for disk-to-disk throughput testing. Additional toolsets welcome, including FDT, mdtnFTP, nuttcp/nuttscp.
  - Expecting multiple, concurrent streams
- ESnet esmond-client package: [http://software.es.net/esmond/perfsonar_gridftp.html](http://software.es.net/esmond/perfsonar_gridftp.html)

**perfSONAR node**

- Minimum one, two would be preferable for inter-Regional regular testing. More are welcome for intra-Regional regular testing.
- pS node deployed topologically close to participating DTN(s)
- **10Gbps-connected** for throughput tests. 1Gbps, 40Gbps, 100Gbps pS nodes welcome.
- (optional) Separate interface for latency tests
- iperf3 as default memory-to-memory throughput
  - Single-stream
  - pScheduler (as opposed to bwctl)
- owping / owamp for latency & loss
- Traceroute (vs tracepath)

**perfsonar-testpoint bundle**

[http://docs.perfsonar.net/install_options.html](http://docs.perfsonar.net/install_options.html)

The perfsonar-testpoint bundle provides the minimal package elements required for the pS node to participate in mesh-orchestrated tests, and register results to a (central) measurement archive. Other acceptable bundles include perfsonar-core, and perfsonar-toolkit.

Mesh configuration (regular testing test_spec parameters)

- Separate tests (test_spec) for IPv6-only and IPv4-only, as in:
  - ipv6_only 1; and,
  - ipv4_only 1
- **force_bidiectional 0** (do not force a given node to initiate bidirectional tests)
- Throughput (10G-connected devices)
  - Disk-to-disk: GridFTP (globus-url-copy); 10GB ESnet reference file; 4x / day; 4 parallel streams
  - Memory-to-memory: iperf3; 30 second duration; 4x / day; single-stream; TCP
  - Other suggested default values in sample test_spec later in document
- Thresholds:
• Acceptable: $\geq 7.5\text{Gbps}$
• Warning: $< 7.5\text{Gbps}$
• Critical: $< 5\text{Gbps}$

• Latency & Loss
  • owping
  • ICMP
  • Other defaults in sample test_spec later in document
• Thresholds:
  • Acceptable: $= 0$
  • Warning: $> 0$
  • Critical: $\geq 0.01$

• Traceroute
  • Traceroute (as opposed to tracepath): $4x$ / day; UDP
  • Other suggested default values in sample test_spec later in document
  • Threshold / expectations for given pairwise test:
    • symmetric forward and reverse paths
    • IPv6 and IPv4 are congruent
    • A path which changes more than $4x$ / day is probably experiencing foo

• Groups
  • Preference for grouping nodes by uplink speed, as in 10G-connected nodes will test against other 10G-connected nodes

Repository / KB for Training & Reference Materials
• Github and/or other central repository or Knowledge Base
  • PRP Workshop Materials Github in-progress
  • Could consider folding in with ESnet Fasterdata KB
• FIONA build specs
• Jupyter notebooks
• Docker files
• Other recipes

Stretch Goals
• GridFTP plugin for pScheduler (in-progress at U.Michigan)
• Federated Regional K8s-orchestrated clusters
• Testing with multiple disk-to-disk tools (FDT, mdtmFTP, ...)
• Lookup-Service for DTNs

Next week, will be 90-minute call so we can have more time working on this document.
Thursday, January 10

Notes from PRP/NRP call 1/10/19
Attendees: John Hess, John Hicks, Tom D, John Graham, Pavan, Scott, Shava Smallen, Kristy Hamm, Shasha, Chad, Greg H, Alan, Dima, Jeff, Matthias, James Deaton, Mark Y, Dung Vu, Ryan Norbrega, Cathy Chaplin

John Hess - new things
  - John G
  - Dima

Greg H - updates from HPREN
John Hicks - NRP
  - Intro Kristy Hamm - KINBER
  - Deployment nodes
    - U of Guam - John G
    - John Hicks - is there a POC?
      - They are on Mad Dash & Rocket Chat
    - John Hess - TAC

John Hess
  - John Hess - John Hicks do you have a forecast for GS relevant for this group?
    - John Hicks - main workshop at TX
    - How many people on this call will be at GS?

  - John Hicks - to NRP folks…….looking at NRP as extension of PRP, working well so far
    - From the scaling perspective from NRP - specific asks
      - What services are we offering to researchers?
        - Specific science drivers looking for services from this platform - what are people asking for?
        - Are there compute functions, kubernetes needs…
      - John Graham - PRP has brought hyper-converged
        - Federated identity
        - Storage
        - Need for GPUs
      - John will follow-up with further questions specific to technical side
      - John Hess - is there overlap with this group and the engagement group
      - John Hicks - however you explain this project to folks, what does the NRP mean to you and how do you explain it to other groups
Notes:
Shava – will be tutorial at CENIC workshop
JohnG: Got U Guam on. Need to work on BGP peering
On road trip with Tom – deployed FIONA 8 – 2 ½ T NVME – 2 more FIONA 8 for Santa Cruz (rearrange) – SF replacing drive
- CENIC outage – lost connectivity but perhaps no lost data
- SF HIPPA compliant – HA cluster config – ansible to more simple config
- Dima working with new conferencing software(?)
Dima – apply to storage and K8s will scale – web odf
- Enable auth for file system – everyone has own folder and own key (CEPH FS)
- K8s changes all files in ROOK – Jupiter will take a long time if a lot of files
JohnH – is there a writeup for examples of workflows
Dima – once working example that will tell us what to do
- When new data arrives it triggers the workflow
- Enable lambda functions in K8s (open source)
- Google kubeless, serverless
JohnH – other FIONA updates from John G?
JohnG – preparing FIONette for workshop
- Isaac and joel(?)
Trip to UC Merced (audio breaking up) – Jeff – expanding GPU adding Jeff’s wave to the Nautilus cluster.
Greg (HPWREN) – ready to get a POSIX FS– reenable “get cam” – CEPH system
  Was using FUSE filesystem - Web dev – now S3
- HPWREN, UC Merced – drone based microwave link – challenges with gov shut down and weather. (Yosemite connect)
JeffW – not visiting Yosemite site due to no snow plows running
JohnG – U Guam new node – working on path – Rocket chat name “NOC” – in maddash gridFTP JohnH was going to help LEARN get a DTN (TACC)
JohnH – working with TACC – Tom Hutton (BGP) – Nathaniel Mendoza
John Hicks – What is the science community asking for from the NRP?
JohnG – federated identity – storage – hyper converged resources – take advantage of other resources – data sciences need GPUs
John Hicks – What is the NRP? Please think about what your “elevator pitch” for what the NRP pilot is and possibly share with the group on upcoming calls.
JohnH - IRNC – SDX – Chicago, Atlantic wave, - what are the priorities – interested in other transfer software:
  - mdtmFTP (developed by FNAL)
    o http://mdtm.fnal.gov
  - FDT (developed by CalTech)
    o http://monalisa.cern.ch/FDT/
  - BBCP (developed by SLAC)
    o https://www.slac.stanford.edu/~abh/bbcp/
Expand the international side with support for A-Star SCA19 inaugural Data Mover Challenge
Global Summit 2019 – using MANRs tut (more planned for Internet2 TechEx in December)
- Frank Wuerthwein and John Hicks will present about OSG and the NRP pilot
Thursday, December 20

JohnHess, JohnHicks, JohnG, TomD, Azher, Chad, Cathy, Greg, James, Shannon, Scott, Jim, Dima, Nadya, Shava, Alan, Celeste, Jeff, Shash, Mark, JoeHesse,


Dima: have been upgrading storage (new version of Rook), all restored. Can choose Ceph version we want (not tied to Rook version anymore), authentication fixed--will enable asap. All NFS is home. JohnG: still see mounts inside cluster. Dima: I can inhibit NFS. Been updating Nvidia drivers with Cuda10, TensorFlow 12, on Kubernetes 12.4. Internet2/Stash Cash updated one, waiting for Edgar. Will go to version 13 next. JohnHess: links to release notes?


JohnHess: use a separate channel just for release notes/updates?

Dima: Arista focused on commercial customers. JohnG: run a kubelet inside Arista--a Calico node could be attached to our cluster. Monitor usage? Azher, can you dig deeper.

From Azher Mughal to Everyone: (10:13 AM)


Dima: Advertising services (what we needed Metallb for) now in Calico. Everyone switching to Cilium. Uses Berkely packet filtering in kernel...super fast and secure; Will do overlay networking, packet filtering. JohnG: efficient to move overlay from applications into driver, closer to hardware. Dima: EVL is broken today. Ultralight can see nodes but not Amazon cloud where docker hub is located. Shash: problem with Internet2 peering with CENIC. Have contacted NOC.

JohnG: HA (high-availability) going good; see Fall 2018 Fiona build in the reminder e-mail--added a FIONA6 tab too (six T4 GPU in a 1U chassis--currently it doesn't work, though--need bios upgrade for T4s it seems) and a $1283 10G FIONA. https://docs.google.com/spreadsheets/d/1b6EzbwMB36T9ndAmCsGMgcThJD SnVltcqj2RT6RwQ2o/edit?usp=sharing

Have 60 12TB drives and 32 12TB FIPS encrypted drives to put up.

Samsung 970 NVMe M.2 sticks support two types of encryption. Will test performance. EdgeFS will do encryption at rest.

Dima: when you create a volume with current Rook, can choose encryption.

JohnG: will set up some drives for an experiment.

JohnG: want original FIONA from UCI back to fix it--won't boot anymore. UCI upgrading their Science DMZ. Want to put encrypted drives at UCSF and UCSD. Joe Hesse: we will see you with drives in hand at UCSF on Jan 14. Will coordinate with Joshua to install early that afternoon.

JohnG: got GPU server in Korea, fixed GFS. Need to clean up paths between nodes in Korea--can see us, but they can't see each other.
Tyan down until we can figure out the bios. Working with Tom Hutton to get 4 FIONA8s at SDSC on line. One is up, the other 3 don't get link light on the fiber path, could be switch-side SFPs.

Greg: HPWREN HA installation not accessible yet--John is waiting on us to register domain name doodads. Ball's in our court. Geoff and I will have a call today; should have progress in a few working days.

John Hicks, NRP:

- Updates on node deployments
  - Current dashboard status--still some issues. John Hess and JohnG fixed PacWave DTN problems (library conflicts). Esmond client uses a different flavor of the library. Google it for the solution--discussed in our FIONA workshop. JohnG: had a ZFS issue with old SPL, had to uninstall new kernel, rebooted, deleted, re-installed kernel then ZFS and it worked.
  - Jim: Two items from NYsernet: Kernel bug with DTN plus latest kernel, ZFS, plus Mellanox driver. Multicast was triggering error messages, posted the bug; non-fatal, doesn't effect performance, just annoying. JohnG: 4.19 or higher kernel with Mellanox drivers, so we stay below; Jim: could be Connect X3 not X4 card. Prefer latest kernel, myself. Looking good at MaDDash. JohnG: see From jjgraham to Everyone: (10:36 AM) [https://yumrepo.nautilus.optiputer.net/](https://yumrepo.nautilus.optiputer.net/).
  - Jim: have working 40G DTN. Have plenty of members that connect at 10G. Fall2018 config? JohnG: 1U dual-10G at $2400 a good choice. Jim: line rate with GridFTP? 40G older FIONA goes 13-14Gbps. JohnG: we haven't tested these boxes as DTNs, but could. Jim: Working with Mike at AMNH in NYC. JohnG: how much storage does Mike need? Could use SAS JBOD. Jim: Mike will find out. JohnG: Could do NAS with the other 10G. Or if it's only a few TB, could even do NVMe. Or an external PCIe extender to a bus extender. JohnHess: see

From john hess to Everyone: (10:41 AM)
Fermilab / ESnet comparison of transfer toolsets — mdtmFTP, GridFTP, FDT
[https://perfsonar.nrp-nautilus.io/maddash-webui/](https://perfsonar.nrp-nautilus.io/maddash-webui/)
JohnH: would like to re-do this comparison.

JohnG: I want to do virtual data mover deployments: pull data into scratch storage, dump into ceph pool, pods mount and play with it. Globus connect server docker works well, I also have FTD containers from SC18. Can move data to your app inside the cluster.
Jim: Researcher wants to get access to data--usually only one or two ways to get access. JohnG: HTTPS is painful--parallel mirror and cache.
JohnG: UCAR has massive data sets—if they can mount their node to data volumes and figure out who is attaching—maybe can forward identity from our cluster to their cluster. JohnHess: a NSF-PI workshop could address this? Jim: folks at NYU doing cancer genomic patient data with genetorrent GDC transfer tools. Years of work into GridFTP, would like to see how a new tool compares; sometimes it's what the data publisher offers. JohnG: get much better performance if don't rely on TCP, like FaceBook. Thinking beyond data transfer, need to access it; Nextcloud federates access to data, can transfer/mirror, has tools. We have that in our cluster—allows individual to push data into the cloud. Unless it's blocked by HTTPS. JohnH: like named data networking. OSG stashcaches have intelligence for certain data sets. JohnG: Alex Szalay’s storage can support things like genetorrent, geotorrent, etc. THREDDS subsetting for environmental data minimizes needs for transfers. Jim: are you involved with OSN? JohnG: will be nodes at SDSC. Need glue between us. JohnG: with release of K8s 1.13, much easier to implement (especially HA).

- Continued architectural framework discussion. JohnG: working on TPM, secure services (IBM cloud, for example)—see how much gets into open source. Dima: storage for secrets talk at KubeCon. Federation V2 doc released; good talk. Need to see implementation. Admiralty.io is creating startup doing federation the way we want it; will try. JohnG: looks like he wants to keep in step with V2 Federation. Looks really nice. He’s willing to allow us to play with him, become pre-commercial customers.

Next call: Jan 10.
Thursday, December 13

JohnHess, JohnHicks, TomD, JohnG, Alan, Chad, Josh, Kate, Mark, Rob, Greg, JohnM, Scott, Azher, Jim, Shash, JohnD, Matthias, TomH, JamesD, JamesM, Dima,

JohnG: 13.1.1 K8s out. New, slightly. Went to UCR yesterday, upgraded their FIONA with 2 half-TB NVME sticks, and racked a FIONA8. Waiting for paperwork to be completed. Met with 5 of their researchers. One is an astrophysicist (Bahram). Victor there has been in Nautilus, very clued in. Happy with how easy it is. https://gitlab.nautilus.optiputer.net/ucr-astro/cloud-astro

JamesD: want to follow up with this, and have a call to do it. JohnG: was really easy to get Bahram up. JamesD: need to have more people building repositories. Want to stress that again.
JohnG: Interest at UCR with Nvidia Xaviers.

JohnG: new Calico that does what Metallb was needed for, makes it easier. Cilium using Berkeley packet filters, will work on ARM64, will have SDN-capable multi-architecture cluster. Can add Nvidia Xaviers (ARM64). JohnH: maybe need a session on this.

From jjgraham to Everyone: (10:03 AM)
https://cilium.io/

Dima is at KubeCon, probing the developers in situ.

JohnHess: can do odroids!

JohnG: ready to plug in cables for four more FIONA8s at SDSC colo. TomH: should be ready. Didn't see the power cords, tho.

JohnHess: sat with prof. from KISTI. Wants a FIONA8 at KISTI. JohnG: they have a multi-GPU server there already. KISTI upgrading connection in February from 10 to 100G. Collaborating with NERSC on a 100G dat mover cluster. Probably GLOBUS.

JohnG: see video on cilium.io page. Has good part on byte-code injection and microcode enabling. XDP is the protocol enhancement that Facebook uses to accelerate throughout. Using Cilium will change protocols, more microservice control. Recommend watching this. Dima got HPWREN HA working with Calico. GregH: wrangled just right, can get at all the pieces, moving forward. JohnG: have EdgeFS that gives non-CEPH way, S3 and Posix. With latest Rook and K8s, can have both Ceph and S3 pool. Greg: we don't want to divide the pool, but have a single pool--want to do this with S3 and Posix. JohnG: Rook/S3/Posix might be faster than Rook/Ceph/Posix route. Supports backend S3 for NextCloud. Keep going the way you're going and we will work on making things faster. Greg: want to mount EdgeFS and do performance testing.

GregH: have permission to do drone RF modeling for Yosemite/UCM link. Have drone permit. Will be able to confirm bw on links from UCM to Yosemite field station. Also putting in more cameras on Orange County. May have found a solution on another tower on Santiago Peak--been taking way too long. Necessary for UCI connection.
JohnG: Have MaDDash movie running again, every 5 minutes, interesting behavior. Can see things blinking out. Can see when zones of machines turn yellow or red--that's caused by CPU starving with heavy use. One student started up a 600 CPU job--using some much CPU and memory that disks slowed down, then network slowed down. We can see that with Grafana dashboards, see which namespace has pods on nodes with bad paths.

From jjgraham to Everyone: (10:26 AM)
http://mad-movie.nautilus.optiputer.net
https://mad-movie.nautilus.optiputer.net/optiputer/optiputer-latency-40g-throughput.mp4

JohnHicks: NRP Pilot
JohnG: new Cilium supports IPv4 and IPv6, will upgrade Nautilus to be dual stack. Building a testbed. Australia guys have IPv6 at the edge. 100G inside their DMZ at UQ. Opens lots of possibilities with Nvidia V100s. Will put some simple hardware and the multi-GPU chassis. They are very interested in EdgeFS. I want EdgeFS over IPv6. Will allow IPv6 software transportable storage.

JohnG: IPv6 NYSERnet node added. Forced to be IPv6. JimK: looks good to UNL over IPv6. Brought up a FIONA at Stony Brook, working with their security folks. Do we have an authenticated way that doesn't use anonymous? JohnG: Nadya has a way for GridFTP globus connect server container--in gitlab repo:
From jjgraham to Everyone: (10:43 AM)
https://gitlab.nautilus.optiputer.net/prp/globus-connect
Ping Nadya for help.

JohnHicks: more?

JohnH: U Guam (Steven Merrill) looking for a place on one of our dashboards.
JohnHicks: regarding GridFTP dashboard, trying to get to UMich folks doing the GridFTP PerfSONAR plug-in. Hope to get a response this month still. Using GridFTP transfer code that John Hess did originally--outside K8s config as a 1-off, but want to roll that into the PerfSONAR container to make it cleaner. JohnG: will start recording a 4K Movie of that dashboard too. JohnHicks: as a public window into the NRP Pilot, want as much green as possible. Want to call out to all the sysadmins and network folks to solve the problems, will start to pester people to figure out if it's network, machine, or config. JohnG: found dependencies on workload affecting throughput. JohnHicks: would like to correlate. JohnG: with Nautilus we can see these things, learning pattern by staring at them. E.g., UCAR and USC had mirrored assymetric latency, unusual. Want to dig into that and see if any other node in the mesh has the same mirrored. JohnHicks: Traceroute?

JohnG: too much hunting and fishing in Traceroute now. JohnHicks: the movie will be good visibility. JohnG: can see the dashboard evolve.


JohnG: Admiralty multi-cluster scheduling good for NRP--discussion needed.
From jjgraham to Everyone: (10:51 AM)
https://github.com/admiraltyio/multicluster-scheduler

JohnHess: interested in looking at when the path was changed. JohnG: Dima's putting all the esmond data in a database so we can see time slices back and forward, train
From jjgraham to Everyone: (10:51 AM)
https://github.com/admiraltyio/multicluster-scheduler
https://kubernetes.io/blog/2018/12/12/kubernetes-federation-evolution/

JohnHicks: have action with GridFTP pschedule input. Other items to be addressed by next meeting? Hearing none, all done.

Will have next call on the 20th.
Thursday, December 6
John Hess, TomD, JohnG, JohnM, JohnHicks, Dima, Ryan, Cathy, Chad, Chris, Greg, James, Ivy, Matthias, Josh, Shannon, Scott, TomH, Pavan Mark, Jim, JohnD,


JohnHess: have FIONA workshop plan in attachments. Drafting the program. Focus on CSU campuses.

JohnG: want to use improved Traceroute to analyze a particular path between people on the call, interactively with screen sharing. Volunteers? see From
https://traceroute.nautilus.optiputer.net/ Dima will take over that.


From jjgraham to Everyone: (10:16 AM)
https://gitlab.nautilus.optiputer.net/patternlab/spark-base
Building docker images locally in our cluster.

From john hess to Everyone: (10:15 AM)
https://spark.apache.org/

JohnG: elastiflow, requires elasticsearch, got most of that set up in K8s. Dima: I have elasticsearch containers too. Can scale.
Josh: we have a lot of K8s expertise at UCSC in this area if anyone wants help.
James: Elasticflow containerization experience here. CS faculty excited. Easy to put in the Science DMZ. Let's have a zoom to align our efforts. Glad we're both on that direction.
JohnG: OK. JohnHess: perhaps during this call?

JohnG: need DNS names on switches so traceroute viz can help us figure whom to call. Need community standards for naming.
Jim: K8s security bug? From Jim Kyriannis to Everyone: (10:22 AM)
https://www.zdnet.com/article/kubernetes-first-major-security-hole-discovered
From Dmitry Mishin to Everyone: (10:18 AM)
https://gitlab.nautilus.optiputer.net/prp/monitoring/blob/master/.gitlab-ci.yml
Dima: we were on it. RocketChat has blue banner with secret so you can encrypt within. Know how to reset if you lose your key. From Dmitry Mishin to Everyone: (10:22 AM)

https://www.vaultproject.io/

TomH: not sure what happened to PRISM name servers. Corresponded to very large DNS attach on the campus. Infoblox servers stopped doing recursion. Started working again. Looked like google changed options. Flag in bind that IPv4 lookup gave only type A records, not quad records for IPv6 (ask Tom H). JohnG: any cluster with IPv6 couldn't get anything but IPv6 name. Had to disable. Had to do this on UCR, UCSC. Look to see if your name servers are getting confused. TomH: using 8.8.8.8 is an issue with making changes. JohnG: 9.9.9.9 is a curated version of 8.8.8.8. Jim: 9.9.9.9 is run by IBM to filter out some bad bots, etc. TomH: endpoints were getting a refused message coming from infoblox master. Running pure bind. Josh: we have a similar setup. TomH: harder than I thought to match bind variable to infoblox variable. All campuses running infoblox. JohnHess: do you have notes? TomH: will put in google drive folder.

JohnG: fixed the pacwave node this morning. Nysernet will have IPv6 today. Jim: have FIONA at Stony Brook which will come up today.

JohnHess: over to John Hicks’ NRP pilot agenda:

- Agenda bash
- Updates on node deployments
  - Current dashboard status
    - https://perfsonar.nrp-nautilus.io/maddash-webui/
  - Continued architectural framework discussion
- Action item review

JohnHicks: Dashboard looking more and more green (issue between our node and Kansas node):
From jjgraham to Everyone: (10:32 AM)
https://perfsonar.nrp-nautilus.io/maddash-webui/index.cgi

James: we are revisiting now that conflicting events are over. Need extra storage at Kansas so stashcache can work better.

Jim: seeing great performance to UNL in Nebraska, over 10G. JohnG; traceroute looking green.

Dima: going to use a graph database to store history. New javascript library. CAIDA has routing info for everyone we can reference.

Jim: please present on where you are going with that.

James: everyone should send BMP (bgp monitoring protocol) data to KC’s. OpenBMP is interesting.

From James Deaton to Everyone: (10:36 AM)
http://www.openbmp.org/

From Ryan Nobrega to Everyone: (10:36 AM)
https://github.com/OpenBMP/openbmp
From john hess to Everyone: (10:36 AM)

JohnG: have experimental guacamole service to get to management nodes and LOM consoles. Have also pulled in Unidata no-vnc streaming container. AWIPS-2 CAVE client for satellite, radar data. Also working on Unity server. Continuous integration allows snapshot, building, adding to docker registry. Spun up OpenNSA container with JohnHess on dev node.

JohnHicks: architectural framework?  JohnG: all new, haven't done ppts yet.  1.13 makes us rethink on how to deploy masters, in one data center or multiple.  JohnHicks: it's a moving target.  JohnG: use their Weave and it should come up.  We switched to Calico. JohnHicks: measurement folks using multus-cni multi adapter CNI. K8s supports a SDN plug in for other plug ins to plug in.

From James Deaton to Everyone: (10:42 AM)
https://github.com/intel/multus-cni

JohnHicks: will look at dashboard more.  JohnG: for NRP, I've been learning Ansible to simplify DTN/node deploys. Ansible playbooks.  Dima: Edgar is doing stashcache with Shawn McKee. Will register with lookup service. JohnHicks: meeting with them later today. Want it integrated into our 3 nodes.  Dima: hope it supports non-standard ports which will make it easy.  JohnHicks: will use pscheduler.  Also: want to get off the manual GridFTP testing outside K8s. pscheduler plug in from Michigan--need to get in touch with Nathan at UMich to get more information. Want everything in the K8s cluster--no external chron jobs running.

JohnG: Chris--can I come plug in some T4s in the 1U?  Chris: Friday the 14th 10:30 am? Confirmed. JohnG: will bring a 4U for the 5 GPUs and 3 V7 FPGAs and 1/2TB of memory.  This box has 3 flavors of NIC.

JohnHess: Dima traceroute viz?
Dima: sharing screen. UCD has campus assymetry. Wash to UCSC has 4 problems. Josh: will look at this more today. JohnG: something changed at the edge when you used 10G backup path. Josh: something maybe hasn't switched back. Rainy season. Mudslides.
Dima: more mousing around--Tom can't capture this Dima showing traceroute. Celeste: look at USC--we are making massive changes, unintended results--please let us know if problems. IPv6 changes. JohnG: need to push IPv6 testing...make sure it works. Need an IPv6 selector on traceroute. Dima: next major upgrade for Nautilus. TomH: paths on Ultralight to Caltech having problems. Celeste: Caltech and USC on Los Nettos, but FIONAs shouldn't use it--go directly to CENIC.
JohnHess; may be some things unwinding from special SC18 peerings (like to India). May have introduced problems at Caltech. Shash?

12/13 call will be 90 minutes to allow more NRP Pilot time. JohnG: please give suggestions on traceroute, and point out problem areas we can look at together.
Thursday, November 29
TomD, John Hess, TomH, JeffW, Curt Dodds (UHawaii Manoa), Wallace, Alan, Amy, Cathy, GeorgeR, Greg, James, Jim, JohnHicks, JohnM, Mark, Matthias, JohnD, Ryan, JohnD, somebody from 404AC.

Jeff Weekley briefed us on the CENIC March conference at the Estancia across from UCSD and March 16/17 tutorial/workshops at UCSD/Atkinson Hall. 4 Workshops: JupyterLab, CAVEkiosk and CAVEs deep dive, PerfSONAR/GridFTP, Kubernetes Policy Crafting.  
Lee Ann is on Global Summit program committee to help bridge the two conferences (Global Summit is two weeks early). We don't want to duplicate content.

Curt Dodds: CPU clock rate vs cores for DTNs; NSF funded astronomy data transfers--2 DTNs, 100G to the edge. spring 2018 FIONA build Epyc + 7451. Got an upgrade to 7551. More cores, but lower clock speed. JohnG: minimum clock rate needed--harder due to Spectre, etc. Need to figure out the work load. You can go to next gen CPU with faster clock, up to 64 cores. If you migrate this from a standard DTN to a Ceph one, need one core per drive. Lots of RAM. If you are using GPUs, need 2x RAM in the CPU per GPU core. Curt: what's the min GHz for 100G? JohnH: want 3 GHz, or more streams. Curt--will buy one of each and then compare. U Maryland involved.


JohnG: have 4 new FIONA8s installed, waiting for power and 10g ports. TomH: I need info from you about these.

Dima: Traceroute working again. Tightened up security. Uses cert manager now. Has I2 on it now. Make sure to hold shift when you load so it loads new code.  https://traceroute.nautilus.optiputer.net/
New dashboards in grafana for gpus.

JohnG: stable now--interactions with Rook/CEPH have paid off. Have split off the SunCAVE into its own cluster (not part of Nautilus anymore) so we can experiment with Dima's federation, after which we can consider federating other clusters. UCMerced WAVE next. JeffW: we're getting ready. Adding drives. JohnG: need small NVMe M2 device in each node. On sale $177 for 1/2 TB. Will do same with EVL once they get more nodes.
JohnM: have an agenda item about you doing an architectural map, want an organizing principle. JohnG: I've started, but wildfires and demos take priority. Will next focus on IPv6 federation.

JohnHess: Queensland Brain Institute are in contact. Maybe do something for next APAN in Dejong.

TomD: have attached new spreadsheet on Nvidia cards. Please give feedback!

JohnMoore: will follow JohnG's lead on the federation. JohnG: we need to develop examples of federation and share.

Tom/John/Dima had to leave for the I2 BoT demos.


**Thursday November 15**

Cathy, TomD, Matthias, Josh, JohnM, JohnHicks, JohnG, JohnD, JohnH, Mark

JohnM: It's SC18 and most people are busy. Look for Internet2's announcement. JohnG: deployed on 6x100G nodes as a "DDOS" service for SC18. Waiting for show floor to wake up and respond.  
[https://gitlab.nautilus.optiputer.net/prp/sc18](https://gitlab.nautilus.optiputer.net/prp/sc18)

JohnH waiting for vlan addresses for his 2 DTNs. JohnH: IPv6 is great.

JohnG: Scott Sellars deployed 70 GPUs on SunCAVE and other nodes during a demo Wednesday at 2pm Central. Did not get TPU inference on Google going yet. Will be pursued.

JohnG: using guacamole for ssh, etc. Have one admin instance in Nautilus. Everyone in the group has a log into the dashboard of html5 consoles. Tunnel through the network and get to LOMs with SDN. Opportunity to use TPMs to store keys for LOMs--using SSH keys stored on the TPMs.

JohnG: using NRP controller that Dima wrote to enable federated clusters. Can use golang coreDNS to automate name services, have secure control by individual universities. Can use software-definable storage, allowing all nodes to be IPv6 routable, touch each other.

TomD: need to figure out what to call the group of federated clusters (UCSD will have several). Nautilus is the soft(ware) name. OptIPuter was the name of the basic concept 17 years ago (I still have the t-shirts). It's important for groups to have identifiable names (like supercomputer centers have for their clusters) but also to be part of a name that "brands" the capabilities offered.

JohnM: good thing for NRP Pilot Exec. Committee discussion.

Tom: NSF likes IPv6 and SDN implementations that are actually used, which is what Nautilus is going after.

JohnHess: Tom Lehman had a successful demo of SENSE, ESnet's SDN/NDN end-to-end transfers (google it?).

From jjgraham to Everyone: (10:29 AM)  
[https://github.com/rcurrie/deepmarker](https://github.com/rcurrie/deepmarker)

S3 storage for genomics.

Check out THREDDs server from Unidata...has Jupyter notebooks integrated...allows huge geospatial data. Will update our current MIRRA2 archive and move one into our cluster storage to have s3 Jupyter notebook backend.

TomD discussed 4 pre-CENIC2019 workshop proposals just submitted (and subject to modification):
We are currently planning four pre-CENIC day-long workshops (one on March 16 & 17, three on March 17).

1. Kubernetes—use & policy crafting & security 5301
2. JupyterLab Apps VROOM
3. CAVE deep dive CAVEs & 5004
4. 2-day perfSONAR, MaDDash, traceroute 4004

#1, #2, and #3 start 9am, end 4:00 Sunday including lunch hour, breakfast 8-9am, in Atkinson Hall.
#4 is 2 days

----------

No meeting next week. Go to 90 minute meeting on Nov 29th. TomD will be helping to host the Internet2 Board visit to UCSD, but can be on the first 30 minutes of the call.
Thursday, November 1
JohnD, TomD, Cathy, John Hess, JohnM, John Hicks, Amy, Chris, Greg, James, JohnG, Kate, Matthias, Jim, Shannon, TomH, Pavan, Mark, Rob, Cеlete, Chad, Alan

JohnH: SC18 circuits & vlans to California in progress. Tomorrow morning at 10 call. TomD discussed several CENIC tutorials prior to CENIC2019 (K8s, JLab, perfSONAR. Security?). TomH: importance of inspecting fiber at 100G? Patching and installing SciNet saw a lot of errors from dirty 100G jumpers, cleaning fixed it. JohnH: part of the technical track. TomH: borrow a fiber microscope.
TomH: the fiber to the Estancia was last lit when it was installed. Need to work on it in December if we are going to use it. TomD: what's the planned use of dark fiber at the Estancia.
Rob: interest in revamping the Quilt fiber workshops. Raman re-introduction. Share stories.
JohnG: been working on HPWREN's high-available Ansible-deployed cluster. 2 nodes--100G connected with 160TB of storage each. GregH: C4 is ready to be wiped. New kernel repo with ZFS, ELRepo, 4.15.15 kernel. Still runs Mellanox. ConnextX4 and 5. Deployable by K8s job.
From john graham to Everyone: (10:57 AM)
https://docs.google.com/spreadsheets/d/1Ln3RMeL8Ze-lQI-AGftV8S 1UeZE1vpZFv85fiPZlVs/edit#gid=0
https://gitlab.nautilus.optiputer.net/prp/kubernetes-ansible

JohnH: KISTI to SC18, Singapore, NZ, etc. Call about Singapore data movement call distributed.
JohnM: Do we have specific questions about that? How to organize the response?
JohnH: I think the call is aimed at a broad spectrum of people. They want access to existing infrastructure like stashcache.
JohnM: we can identify infrastructure.
JohnG: seems like they want to have competing protocols.
Rob: They will all be at SC18. Should figure out what they want there.
JohnH: What assumptions are they making? Can they containerize?
JohnG: have root to 2 1G boxes in Singapore.
JohnH: Singapore 100G path unstable due to cuts in the Philippines. Another path via Tokyo. Don't know their primary path.
Rob: I agree--they are earnest but have a long way to go. The circuit was down half of last year. Have a redundant ring now.
TomD: just insist they join our cluster, even at 1G and start learning how to move/measure/use bytes.
JohnH: there is one driving demo at SC18 with NSCC.
TomD: working on RTX 2080-Ti and FIONA8 qualifying.

JohnMoore: NRP:
Updates on node deployments:
JohnHicks: working on CENTOS upgrade with OSG folks.
JohnG: IPv6 Missouri asymetrical, MaDDash getting more green, SciDas improving, USC did some routing changes on Science DMZ; Nysernet progress.
Celeste: want to defer to Azher--new design lets any building use Science DMZ--in process over the next year. Want to know when things don't work. Networks shifting. Arista core, HP in the buildings. 170 buildings, half done so far.
James: want to have a deeper dive on K8s, Have 8-9 nodes, having overlay issues which Dima has probably solved. Measurement folks want to collaborate. Want a fuller set of K8s docs.
JohnG: 90% MTU problems; should have a zoom meeting.
James: Maybe at SC18?
JohnG: need other folks to federate with us so we have a common platform to discuss. If you deploy what Dima has done for the MaDDash, get diagnostics with it. Gives lots of hints.
From john graham to Everyone: (10:57 AM) https://docs.google.com/spreadsheets/d/1Ln3RMel8Ze-IQI-AGftV8S_1UeZE1vpZFv85fiPZIVs/edit#gid=0
James: we have a big space at SC18; take it to rocketchat.
JohnG: can set up a private room for this.
JohnMoore: partial group at SC, then set up a separate zoom.rocket
James: we are working on an Ansible. John, you can clone ours.
From john graham to Everyone: (10:57 AM) https://docs.google.com/spreadsheets/d/1Ln3RMel8Ze-IQI-AGftV8S_1UeZE1vpZFv85fiPZIVs/edit#gid=0 https://gitlab.nautilus.optiputer.net/prp/kubernetes-ansible
James: let's continue on rocketchat
JohnG: will probably go Ansible deploy on our entire cluster. All I had to edit was the inventory list, and the group vars all (specific paths & hooks), tweaked for 1.12 K8s, Had to delete some unclean old things. Good baseline for HA cluster. JohnG: wiki growing.

Rob: ppts on 100G node deployments: JohnG: expectations on 100G with kernel patches for Spectre, etc. need to be dampened. Line rate not possible anymore. From Rob Vietzke to Everyone: (11:10 AM)
here is the TREX 100G testbed talk. https://meetings.internet2.edu/media/medialibrary/2018/10/22/20181016-Brochu-virtual-routers2.pdf

Last week, we targeted putting in a federation, milestones, end goals. Want this info for the NRP steering committee.
JohnG: have some basic concept, can put into graphics (now, soon, end); will be google slide deck. Filling in bullets regarding Scott Sellars' atmospheric river SunCAVE GPU challenge for SC18. Distributed GPU, running in Korea and the SunCAVE. Block diagram of one HA cluster in the repo is pretty complex. Will merge with other Nautilus. Need HA to have reliability. At the edge of experimental--not ready for prime time yet.
JohnM: will be looking for that, discuss when you've got it. Do people move to HA first? JohnG: that's what I want, Ansible script now makes it possible. Need another cluster to federate. Guys in Korea have enough nodes to play.
JohnM: want folks to do this within their region. JohnG: getting easier.
JohnM: invite Frank. TomD: Frank's new TNRP FTE can be charged with this.
JohnG: CHASE-CI has its specific policies. We label ownership and provenance of every object in the cluster.
Dima: two groups using all the GPUs, one user using all the remaining CPUs.
JohnG: need to embody fairness of access in code.
JohnM: need to translate the code to policy for the Steering Committee.
JohnG: new IPMI grafana, allows power consumption views.
From john graham to Everyone: (11:24 AM)
https://grafana.nautilus.optiputer.net/d/km9t3uxiz/ipmi-overview?orgId=1

From Dmitry Mishin to Everyone: (11:26 AM)
https://gitlab.com/ucsd-prp/prp_k8s_config/blob/master/ipmi-mon/ipmi.yaml
https://github.com/lovoo/ipmi_exporter

JohnM: want to learn from OSN and OSG regarding policy--is commonality possible? JohnG: some of these policies can be implemented in K8s.
Dima: need people to release GPUs when they are done. Will be able to judge efficiency.
JohnG: if use gitlab, can scrape code for this.
Dima: have admission controller--can see sleep infinity, etc., for running jobs after they are debugged.
JohnG: we can profile jobs with good and bad patterns.
Continued policy discussion next call, November 8.

Over&out 11:30.
Thursday October 11
JohnHess, JohnG, TomD, GeorgeRobb, Cathy, JohnD, Matthias, TomH, Amy, Alan, Dima, Shannon, Adam, GeorgePeek, JohnHicks, Greg, Shash, Chad, REANNZ, JohnMoore, Josh, Ryan, Kate, Pavan, Azher

JohnHess: Autogole NSI discussion/SC18--wanted a vm from SCInet for this, but not going to happen so will look for an external one.
JohnG: if it's just a controller, you can do a container in the cluster.
JohnHess: OpenNSI is skinny enough to be in a container.
JohnHess: JohnMcCauley is working to add SENSE resource manager for Harvey at Sunnyvale OpenNSI, with MAX and Fermilab/ANL. Sorting out the PacWave DTNs to get 100G connection to SC18.

JohnG: Gigabyte talking about giving us 2 AMD Roma systems for SC. 7.5nm 64-core. Azher wants to put ConnectX6 to do 200G. AMD wants to highlight Roma architecture at SC18. JoeM & Harvey & Azher could each use one. Azher: thank you very much.
JohnHess: should be approaching a terabit between Harvey and JoeM on the floor.

JohnHess: Planning workshops for CENIC. TomH: have fiber to the Estancia if we want to do demos there. Dark--light it with 100G if you want and we can borrow/buy small switch. Or run between two AMDS? TomH: will the NICs drive LR optics? JohnHess: can do LR4 okay. Maybe Gen4 better.

From shashwithaputtaswamy to Everyone: (10:09 AM)
https://docs.google.com/drawings/d/1c3vwoTR0Tc-t9zMEjJwv5U65DRmQokWS0_M5_nN4as8/edit https://docs.google.com/drawings/d/1-M0EwiM48cimf0fm3kE4MLE9umlm0EnBLXqGPjdzPNk/edit
From john hess to Everyone: (10:21 AM)
presentations from SURFnet workshop at ASTRON in NL:
From john hess to Everyone: (10:21 AM)
From john hess to Everyone: (10:22 AM)
JohnHess: SURFNet workshop folks discussing many of the same issues as faced by PRP-sponsored workshops, regarding science engagement and technical engineering.

TomD: designing High Availability 3 masters 1U 10G boxes with 3 nodes for storage, same 1U boxes with NVMe for UCOP CIO.
JohnG: FIPs420 ok for HIPAA? HDST makes a bunch. I'll get you some urls. Mobo has SAS, so may be able to put a 12TB disk in each. Could have a HBA for JBOD. Pavan: perfect. I want to help.
JohnG: Nautilus: UCSC nodes being updated (including original FIONAs there, including the Hyades DTN). Josh: right. JohnG: Rob Currie's application at UCSC is in the chat room--cool pipeline. Interactive debugger for deep learning sessions: From Dmitry Mishin to Everyone: (10:23 AM)

Dima created a reddis worker queue to scale to as many jobs as you want. Mehrdad here and Scott Sellars using this. Scott is pulling full planet layers from THREDDS server MIRRA2 data mirroring NASA. Once integrated into the reddish, kubectl job submission of Scott's flood fill/segmentation Connect algorithm, use as training data for a ML inference. Driven by Jupyter notebooks, command line. Grafana dashboards getting better and better. Figuring our how to manage memory in the containers with initd/teeny to clean up zombies from dead apps. Also memory allocation from unruly code--getting out of memory errors with hung NFS mounts and zombies. Figuring out new best practices. Have stood up S3 (Rob Currie uses S3 now). Also plugging Optane in to see if Ceph can use it, or if we have to create custom pools.

JohnHess: do you have a primer on Rook/Ceph access within Nautilus.
JohnG: our Wiki is growing. Dima: setup is just gitrepo for a bunch of yamls. From Dmitry Mishin to Everyone: (10:31 AM)
https://gitlab.com/ucsd-prp/prp_k8s_config


Dima: thanks to JohnHicks, got reply from perfSONAR4 guys, all deployed now. Can run several perfSONARs on the same host. Postgres uses local socket, so can disable port to avoid conflicts. OWAMP flag allows changing port. Should have 2 or 3 of them working today (including I2's custom one). Can't use labels for out stuff. Here's the pointer: https://perfsonar.nautilus.optiputer.net/maddash-webui/index.cgi?dashboard=Dashboard%201
Showing latency, but throughput broken. Much greener now. JohnG: maybe using less CPU? Invite perfSONAR people to look at our dashboards--they have access to every node log. JohnHess: memory to memory? Dima: yes, but I can also do GridFTP, maybe FDT.

Dima: talked to Rook about shared storage. Will release the fixes we need next. Dima: going to Kubecon in December in Seattle, they want us to give at talk. Link to yum repos: 4.15.15 kernel with no hacking install nvidia drivers, ZFS builds, etc. From Dmitry Mishin to Everyone: (10:39 AM)
https://yumrepo.nautilus.optiputer.net/ Dima: this is our working one with upper bound for Melanox and lower bound for Ceph.

JohnG: making progress on the 3-node High Availability cluster for HPWREN. Working on timeout issues. Community is aware of it. Very soon, will have 2 more storage nodes for HPWREN.
JohnMoore: hearing about people using Nautilus. JohnG: Matthias at UCM is from Jupyter team. Matthias: on it. Netflix has peppermill(?) to run jobs.
JohnG: will publish this stuff on gitlab. See Rob Currie's already there.
From Matthias Bussonnier to Everyone: (10:56 AM)
https://medium.com/netflix-techblog/notebook-innovation-591ee3221233 2 of the authors are also member of the Jupyter Team.

JohnG: LDM-7 experiment with UCAR and UVirginia, vlans extend through CENIC to AL2S in LA--will have reliable multicast distribution of weather data on 10 vlans. Perfect for hyperchannels and SDX. TomH: John Hess, what is CENIC's stand on putting private address L3 tags on vlans? LDM-7 project wants every switch to have L3 private addressing. JohnHess: we've done that on a case-by-case basis for trouble shooting, not routine. JohnG: probably for trouble shooting if we see problems and need breakout IPs. JohnHess: we only have 100 on the Brocade MLXCs. JohnG: can we set up one breakout at the CENIC AL2S border? JohnHess: really PacWave switch in LA. Could do that. TomH: I'll put in the request for one for debugging. LDM-7 folks should be ok with that. JohnHess: like vlan 1301. JohnG: but 10 vlans! Crying for SDX. JohnG: NYSERnet working on their PRPv1 node, their K8s cluster up and running though. JohnMoore: will see Jim at TechX.

No call on the 18th due to TechX. See some of you in Orlando. Next call on October 25.
..tom..
Thursday, October 4
Mark, JohnHess, JohnG, Chris, Dima, Cathy, Greg, Matthais, JohnHicks, JohnM, Azher, Shannon, Lincoln, Jeff, Shash, Harvey, Kate, JohnD

JohnG: we've built a new kernel with Nvidia drivers, 4.15 series. LTR. Latest Mellanox drivers. Pulled down with a K8s job in the cluster.
Dima: Ceph supporting the latest, Mellanox dragging. Pulled source from kernel.org, got spec from Kernel.ml
30 lines of yaml calls.
JohnHess: trying to wrangle the PacWave DTNs in to shape.

Harvey: noticed that John Hess, new DTNs in Sunnyvale, LA, Seattle...embedded in the network. We'd like to consider their use in some compatible way (SENSE), named data networking assisted by SDN. With many DOE labs, Caltech. Want to incorporate one or more in SC18 exercises. We need to specify which paths (dedicated? shared?) but also services and utilization levels.

JohnHess: discussion with John McCauley setting up SENSE RM to interact with Pacific Wave--100G nodes between Sunnyvale and Los Angeles for Caltech to use the fabric. Involves Tom at Max Gigapop.
Harvey: need to install a DTN RM (3 softwares: SENSE RM, orchestrator, DTN RM).
JohnHess: Use Sunnyvale and LAX ones, attached to PacWave testbed. LA one will get a Zeo link. Sunnyvale can be an in-network cache.
JohnG: should dockerize so it can be added to DTN easily.
Harvey: good idea. Need to know has this overlaps with FK Wuerthwein/OSG/I2. Caltech and UCSD work as a unified cache, thinking of growing it. Should overlap with SC, use of DTNs.
Dima: Stashcache: haven't heard much from Edgar recently.
Harvey: the production cache uses xrootd.
Dima: FKW has 3 nodes in I2.
JohnHicks: I deployed 3: Chicago, New York, Kansas City. StashCache containerized, deployed, problem with all I2 IPs set to Ann Arbor, but fixed that. Certificates working. Part of OSG and PRP/NRP K8s cluster. We are using a different OS (Scientific Linux), need to change to Centos7. Will bring the nodes down after TechX, want to know optimal disk config. JohnG: 500G NVMe for docker volume. Dima: Frank is creating one container per node. JohnHicks: I2 nodes in Chicago and NY are identical. Kansas one has different config.
Harvey: trying to automate everything with IPMI, etc.
JohnHicks: using Ansible to maintain OS and accounts, K8s for everything else; JohnG: we'll go with Ansible too.
Harvey: seeing saturation, using 200G link. Shash exercising the SDN links on Caltech campus. Will have 3 pairs of wave servers between Caltech booth and SCInet. Have talked to Dave Reese. Every CIO says need to move forward without increasing budgets. JohnHess: start a dialog with CENIC.
Harvey: Caltech CIO on CENIC board. Good time to start thinking about post SC18.
TomD: Tutorials at Global Summit and CENIC (both in March) being worked out now.
JohnM: they're here in Washington. From john hess to Everyone: (10:32 AM)

-- 2019 -- — Internet2 Global Summit: Washington, DC — March 5-8 — CENIC
2019: La Jolla — March 18-20

JohnG: working on UCLA 100G nodes--upgrading to 194TB. Companion is at UCSC has been upgraded to 194TB. Another at UCSC has 72TB ready to attach to Nautilus. Still having problems at UCM. Western Digital card issues with supermicro mobos. Jeff: on ours? JohnG: pull all the NVMe out of the mobo. Jeff: will check with Jeff Pedro. Maybe need to abandon that machine. JohnG: will work on it some more.
JohnG: Dima and I will Zoom into Rook meeting to up the urgency of fixes. Controller problems, in particular. Everyone click on the link: From Dmitry Mishin to Everyone: (10:39 AM)
https://github.com/rook/rook/issues/2112

JohnG: Samsung 970 Pros are good. WD Black is bad.
Harvey: SSD prices not falling anymore. JohnG: they are pretty good.
Harvey: who is using links from California to SC? Need to raise support--which campuses doing anything at SC? John Hess, can you get me a list. JohnHess: I don't have that, vlan requests come in near the event. Harvey: need to get things set early as possible. JohnHess: the list you shared a few days ago is good. Add UC Davis group working with NERSC (on 8K video) to deliver it to the venue. Harvey: UCLA? Stanford? Azher: USC--Shoah Fdn will do streaming of video. 500Mb/s. Maybe 3 of them. Not big BW. To StarLight or USC/Caltech--4 booths with displays. Harvey: important content. Azher: want to use Amazon S3 on west coast with I2 peering. Phil Reese wants to do presentations, not demonstrations. JohnHess: organizing mini-workshops w/out live demos with big BW.

JohnG: LDM7 project with UvA/UCAR 10 vlans plumbed across CENIC to AL2S to SDSC. Reliable multicast ingress for weather data. Coming along nicely. Bought 3 1U 10G boxes to be high-availability (HA) K8s controller. Will help other NRP members who want to do HA clusters. A lot of activity--yesterday all but 4 single GPUs in use.
JohnG: new single socket AMD Gigabyte is lovely. Harvey: please forward recommendations to me.
JohnG: will get some 7nm Gigabyte parts. Possibly before SC. Use 32 core. 64-core Epyc will plop into oldest 32-core sockets OK.
JohnHess: new build on google?
JohnG: Jetson Xavier being investigated. Great specs.

Azher: AutoGole VM for SC18. Need a VM to config NSI for anyone to use. 400G switch with dynamic vlans. Harvey: John McCauley working on SENSE services to look like AutoGole API. JohnHess: would be neat if openNSA instance could host SENSE RM. Harvey: and DTN RM. JohnHess: NSI UPA manage part of the infrastructure on the show floor. Azher: we can create this VM early on, let Harvey know.

JohnHess: on to NRP Pilot:
JohnMoore: Thanks. Here's a draft agenda for the NRP Pilot portion of the call:
• Updates on node deployments
  o Let’s work on getting to green on the dashboard for Tech Ex!
  o https://perfsonar.nrp-nautilus.io/maddash-webui/
• Mapping effort
  o Node map/diagram that includes campuses
  o Indicate current/target functionality?
• Action item review
• Adjourn

JohnMoore: not much movement to green in the dashboard. John Hicks, please reach our to Jimmy and George. JohnG: I have contact with all of them. Alex Feltus is in charge of ScIDAS group. Missouri IPv6 only, doing testing, but we can’t see them yet. NYSErnet on learning curve. JohnHicks: issue between I2 and UNL advertising on v6 just changed. JohnG: helps if I have root or it is containerized.

From John Hicks to Everyone: (11:03 AM)
https://perfsonar.nrp-nautilus.io/maddash-webui/

From john hess to Everyone: (11:06 AM)
https://docs.google.com/drawings/d/158wd7pAWg5mUnyppJJHxGBOxP-H0UJmvTRQ70V2sGJs/edit?usp=sharing

JohnMoore: PRP has a nice network diagram that overlays points that are connected. Need one for NRP Pilot. ---
https://www.google.com/maps/d/edit?mid=1dnNWZ9XnLX6DR58g1KqPcl5pvf8&ll=7.26005872497022%2C-113.42219799999998&z=3

JohnG: we have a live traceroute viz tool that reaches in the esmond data base. From jjgraham to Everyone: (11:07 AM)
https://traceroute.nautilus.optiputer.net/

JohnMoore: need a good diagram. Anyone on NRP Pilot team volunteering? If no takers, I can reach out to James and Jimmy.

JohnHicks: stashcache monitoring is very slow. working on it.
JohnG: DIma is working on clusterwide fluent dlogging. Dima: need to lower number of logs we generate, then it may work.

JohnHess: @JohnG, is this correct URL to the NRP nautilus mesh?
https://meshconfig.nrp-nautilus.io/cron-gridftp-transfer-mesh.sh

JohnG: see readme on top of that gitlab. We have 2 gitlabs!
Dima: playing with new perfSONAR psconfig release, is easier to use, but not working.

JohnHess: GregH/HPWREN updates?
Greg: UCM to Yosemite Park link--preparing for aerial surveys and getting permission to use tower 25 miles east of UCM from Cal Office of Emergency Services (CalOES), working on budgets and funding. Commercial towers close, if CalOES doesn't work out. Bench testing Cl Electronics at UCSD. K8s HPWREN cluster HA master nodes placement being worked out, deploying additional storage for HPWREN across California--looking forward to expanded Ceph pool for our testing. Can be relay point for other locations of interest. Santiago Peak: no optimistic news. Working on it, as always.

End of TNRP Year 1 Annual Report Supplementary Materials