NetSage
Award #1540933
Year 4 Annual Report
1 Feb 2018 through 31 Jan 2019

**PIs: Jennifer Schopf (IU), Sean Peisert (UCD, LBNL),
Andrew Lake (LBNL), Jason Leigh (UHM)**

Summary
The goal of the IRNC NetSage project is to collect data from the IRNC-funded backbone and exchange points to better understand the use of the resources. In addition, this collected data is also made available for use by the NOC for day-to-day operations and to support end-to-end performance troubleshooting. Highlights of Year 4 included releases of 11 dashboards of which 7 were entirely new, and featured flow data, Sankey graph visualization, Tstat data, and the science registry, third party deployments of NetSage and a collaboration with the new Engagement and Performance Operations Center (EPOC), and significant work with Tstat to measure science archive behaviors.

1. NetSage Overview
NetSage is building and deploying advanced measurement services that will benefit science and engineering communities, focusing on:

- Better understanding of current traffic patterns across IRNC links;
- Better understanding of the main sources and sinks of large flows to know where to focus outreach and training; and
- Better understanding of where packet loss is occurring, whether or not the loss is caused by congestion or other issues, and the impact of this on end-to-end performance.

NetSage services provide a combination of passive measurements (including SNMP data, flow data, and deep packet header inspection), as well as active measurements (mainly perfSONAR) for longitudinal network performance data visualization.

Year 4 of the project focused on additional analysis of the data being collected, especially flow data, expanding flow data collection across the IRNC backbones and exchange points, and collecting Tstat data from archives. Going back to our guiding list of questions (https://docs.google.com/spreadsheets/d/1BFdS32mwOv9g2CknOL7kWkKCDj4rrCwDqDblZxBaeD0/edit#gid=1801119320), this included:

- Which links are experiencing packet loss;

- What are the top sites that use the IRNC links?
- What are the top science projects that use the IRNC links?
- What is the nature of elephant flows that use the links?
- What is the max, min, and average duration of elephant flows?
- How can we best identify a list of top talkers for each link?
- How many flows experiencing issues also have small buffer sizes?

Year 5 will focus on extending the data collected from instrumented archives and the science registry, and will address the following questions:

- 4.c - What are highest retransmit rates between organizations/subnets over a timeframe?
- 6.g- Who have been the top talkers each year (longitudinal study)
- 7.f - Is retransmit data a proxy for packet loss?
- 9.b - What level of retransmits is an archive experiencing during data flows?
- 9.c - What are the patterns for retransmits on an archive?
- 9.d - For a pair of endpoints, what is the retransmit behavior for the individual flows?
- 11.a- What is the bandwidth of a GridFTP file transfer between two backbone end points?
- 12.a - How are active tests between two sites performing? (replacement for perfSONAR MaDDash)

This report details the staffing, collaboration, tool development, deployment, and planning for the project.

2. Staffing

At the beginning of Year 4, funded staff included:

- Jennifer Schopf, IU, PI - overall project director
- Ed Balas, IU, system architect - collection and reporting
- Dan Doyle, IU, developer - collection and reporting
- Michael Johnson, IU, developer - collection and reporting
- Sangho Kim, IU, system engineer - collection and reporting
- CJ Kloote, IU, developer - collection and reporting
- Ed Moynihan, IU, Science Registry Data support
- Lisa Ensman, IU, developer - Science Registry
- Jonathan Stout, IU, developer - Science Registry
- Sean Peisert, UC Davis and LBNL, co-PI - security, privacy, design
- Jon Dugan, LBNL/ESnet, senior personnel - monitoring architecture
- Anna Giannakou, LBNL Post Doc, measurement analysis
- Dipankar Dwivedi, LBNL Post Doc, measurement analysis
- Jason Leigh, UH Mānoa, co-PI - visualization oversight
- Alan Whinery, UH System, senior personnel - perfSONAR, PIREN coord.
- Alberto Gonzalez, UH Mānoa, graduate research assistant - viz developer

- Tyson Seto-Mook, UH Mānoa, graduate research assistant - viz developer

Sreemuka Taduru, IU, was integrated into the project in April with a focus on the Grafana map and data source development. During Quarter 3, the IU software team was restructured, and Johnson, Kloote, and Stout were shifted off the project. In August, Scott Chevalier was added to assist with the overlap with supporting the IRNC perfSONAR mesh and for work in adding additional data from that monitoring system. In October, Andrew Lee was added to reflect his work adapting the current NetSage dashboards for additional use in analyzing network traffic. In Year 5 we expect to bring on at least one possibly two more developers to the IU team.

Year 4 saw significant changes in the staffing at LBNL, including a change in leadership from Sean Peisert to Andy Lake. The team spent considerable time getting a better understanding of the project deliverables going forward, and how to map the current staff resources to those efforts. As a result of this, most of the staff at the beginning of Year 4 were shifted to other projects. Mariam Kiran, LBNL/ESnet research scientist, briefly joined the project in April 2018 to support data analysis, but needed to subsequently drop out, due to demands from other projects. Anna Giannakkou and Jon Dugan rotated off the project at the end of July 2018. Dipankar Dwivedi rotated off the project at the end of August 2018, in part due to the change of role that significantly increased his salary rate in August. It is expected that LBNL will bring on additional software developer resources in Year 5.

At University of Hawaii, Tyson Seto-Mook was on summer internship at the Jet Propulsion Lab June through August. Mahesh Kanal joined the team as a new graduate student also working in visualization in September. He took over the duties of Alberto Gonzalez, who is now focusing on his dissertation.

At the end of Year 4, funded staff included:

- Jennifer Schopf, IU, PI - overall project director
- Ed Balas, IU, system architect - collection and reporting
- Scott Chevalier, IU, IRNC perfSONAR mesh support
- Dan Doyle, IU, developer - collection and reporting
- Lisa Ensman, IU, developer - Science Registry
- Heather Hubbard, IU - Staff support
- Sangho Kim, IU, system engineer - collection and reporting
- Andrew Lee, IU, network data analysis development
- Ed Moynihan, IU, Science Registry Data support
- Sreemuka Taduru, IU, Grafana map development
- Andy Lake, LBNL, co-PI
- Sean Peisert, LBNL, Security advisor
- Jason Leigh, UH Mānoa, co-PI - visualization oversight
- Mahesh Kanal, UH Manoa, graduate research assistant - viz developer
- Tyson Seto-Mook, UH Mānoa, graduate research assistant - viz developer
- Alan Whinery, UH System – perfSONAR, PIREN coordination

3. Collaborations, Travel, and Training

NetSage staff participated in various meetings to support ongoing deployment, collaboration, and training. Note that several of these were funded by other sources but relevant to NetSage. The travel for the first 3 quarters of the year, detailed in those project reports, included:

- Schopf attended the Quilt Winter member meeting in La Jolla, CA, February 6-8. https://www.thequilt.net/public-event/2018-winter-member-meeting/.
- Schopf attended the CENIC spring member meeting in Monterey, CA, on March 6-8 https://cenic.org/conference.
- Johnson attended the perfSONAR face to face in Amsterdam, March 7-8.
- Schopf and Moynihan attended Internet2 Global Summit, San Diego, CA May 6-9 https://meetings.internet2.edu/2018-global-summit/.
- Schopf, Moynihan, Balas, Doyle, Leigh, Gonzalez and Tierney attended NetSage All Hands Meeting July 10-11, 2018 at LBNL.
- Schopf attended the National Research Platform meeting held in Bozeman, Montana, on August 6-7, http://www.cvent.com/events/national-research-platform-conference-toward-a-national-big-data-superhighway/event-summary-48a69b9807bd46ecb5d4343bcbfa61c5.aspx.
- Jared Schlemmer and Jeff Terzino, members of the GlobalNOC team who provide engineering services for TransPAC and systems support for NetSage, installed equipment in Hong Kong to support the Hong Kong-Guam circuits on August 5-8.
- On August 17, International Networks at Indiana University (IN@IU) celebrated its Twentieth Anniversary.
- Lee and Moynihan attended the GNA Technical meeting, Nordunet meeting (https://events.nordu.net/display/NDN2018/Welcome) and GLIF Americas and Annual meetings (https://www.glif.is/meetings/2018/) on Sept 18-22 in Denmark.
- Schopf attended the Quilt/CC* PI meeting on Sept 24-27 at the University of Maryland, College Park, Maryland https://www.thequilt.net/public-event/2018-nsf-esnet-quilt-workshops-meetings/.
- Schopf, Doyle, Lake, and Balas attended the Internet2 Technical Exchange, on October 14-19 in Orlando, Florida https://meetings.internet2.edu/2018-technology-exchange/.
- Schopf visited NOAA in Boulder, Colorado, on October 22-25, 2018, to meet with the NOAA and EUMetSat weather satellite groups who are interested in using NetSage.

During Quarter 4, travel included:

- Schopf, Lee, Chevalier, Leigh, and Mahesh attended the SC18 Conference in Dallas, TX, November 11-18, 2018 https://sc18.supercomputing.org/. Meetings were held with partners, and several presentations on the exhibition floor were given. The SCiNet Operations Team had NetSage displayed as part of their monitoring system.

- Schopf visited NSF and the American Geophysical Union annual conference (https://fallmeeting.agu.org/2018/) in December in part to discuss NetSage status and collaborate on possible future projects.
- January 23-25 all members of the team attended the Winter All Hands meeting in Hawaii, which consisted of a one day Hack-a-thon, followed by two days of face-to-face meetings to plan for Year 5.

During Quarter 2, a publication was written in response to our initial positive reviews for a submission to the PEARC 18 conference (https://www.pearc18.pearc.org/). We submitted "Wide Area Network Monitoring with NetSage" with author list Schopf, Doyle, Kloote, Balas, Peisert, Martinez, Seto-Mook, and Leigh. This paper was not accepted for the final conference.

During Quarter 3, a paper related to the retransmission prediction work was submitted to the INDIS 2018 workshop (https://scinet.supercomputing.org/workshop/) but it was not accepted. It is expected that the paper will be resubmitted to the journal Future Generation Computer Systems' (https://www.journals.elsevier.com/future-generation-computer-systems) and their Special Issue on "Innovating the Network for Data Intensive Science" for 2018.

4. Project Coordination

4.A Internal Coordination

Internal project coordination continued with weekly meetings of the majority of the team. We have also implemented a weekly technical call to be able to dive-down into more detailed topics with those NetSage members who are interested. These are complementary to the twice yearl face-to-face meeings that concentrate on more strategic planning.

During the year, we did a refactoring of the effort from LBNL, and Andy Lake is now the LBNL PI. Andy has a depth of experience with monitoring systems through his work with perfSONAR, and will be bringing on valuable expertise. Sean Peisert will continue on the project in a reduced role as a security policy consultant.

On January 23-25, the full team met at the University of Hawaii for the annual winter All Hands Meeting to plan Year 5. This year, the first day of the meeting was spent in a Hack-a-Thon, where 14 tickets were closed out, and the basis for a late January release was established. The partners then spent two days planning on Year 5 activities, including how the project could be handed off if necessary, at the end of its fifth year.

4.B Work with IRNC Partners

Work with the IRNC-funded backbones continued, and we now have SNMP and perfSONAR data from all of the original circuits for the IRNC backbone projects, as

well as the added links that took place this year for TransPAC (Guam-Hong Kong) and PIREN (LA-Hawaii-Guam). Sampled flow data is being collected from NEAAR, TransPAC (both links), and AmPath. We continued conversations with PIREN to collect flow data, but did not make forward progress this year- the discussion remains with the AARNet lawyers. For the exchange points, we have sampled flow data from AmLight and CENIC, but not StarLight. We are also continuing discussions to get Tstat deployed with several archive owners to collect that data.

We continue ongoing discussions and coordination with the IRNC NOC. We had several discussions to understand the NOC use cases for alarming and alerts on the NetSage data to support NOC activities, however no use cases were identified in the meetings held between the teams, and this work was dropped.

In addition, prior to September, the primary support for the IRNC perfSONAR mesh was done by the IRNC NOC, namely the Performance Engagement Team. However, it was decided by the IRNC NOC PIs that this was outside of the remit of that project at this time. Since NetSage depends on that data source, staff was brought on to help maintain the resource.

4.C External Partners

The first third-party deployment of a NetSage dashboard was in Quarter 4, featuring the SNMP data for the core Advanced North Atlantic consortium (ANA), as shown in Figure 1 and available online at http://ana.netsage.global. Next steps for this deployment include expanding the dashboard to include data from the ESnet trans-Atlantic circuits and beginning the discussions to collect and display flow data.

In addition, collaboration began with the US domestic-focused Engagement and Performance Operations Center (EPOC) (http://epoc.global). That project is funded in part to use the open source NetSage code base to deploy dashboards for their six regional networking partners, which include the Indiana State Network (I-Light), the Ohio State R&E Network (OARnet), the Keystone Initiative for Network Based Education and Research (KINBER), the Great Plains Network (GPN), the Texas State R&E Network (LEARN), and the Front Range Gigapop (FRGP). An initial bandwidth dashboard for GPN was developed and deployed by EPOC staff members, and presented at the Internet2 Technical Exchange, as shown in Figure 2.

In Year 5 we expect the number of third-party deployments to grow, and resources will be spent not only ensuring that process can happen smoothly but fully documenting all aspects of the system for a complete setup of the archive and Grafana backend resources in light of a possible hand-off activity at the end of Year 5.

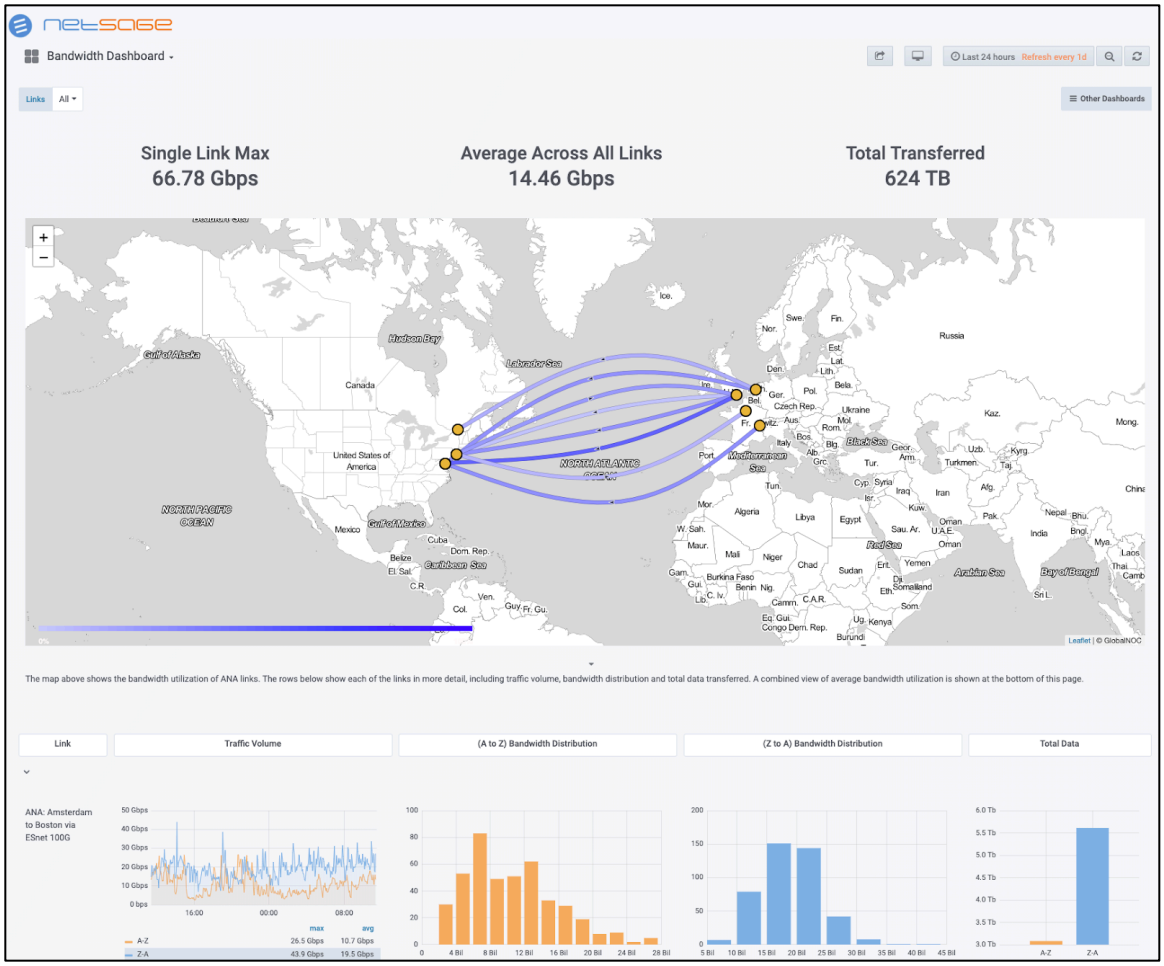


Figure 1: Screenshot of the bandwidth dashboard for the Advanced North Atlantic consortium.

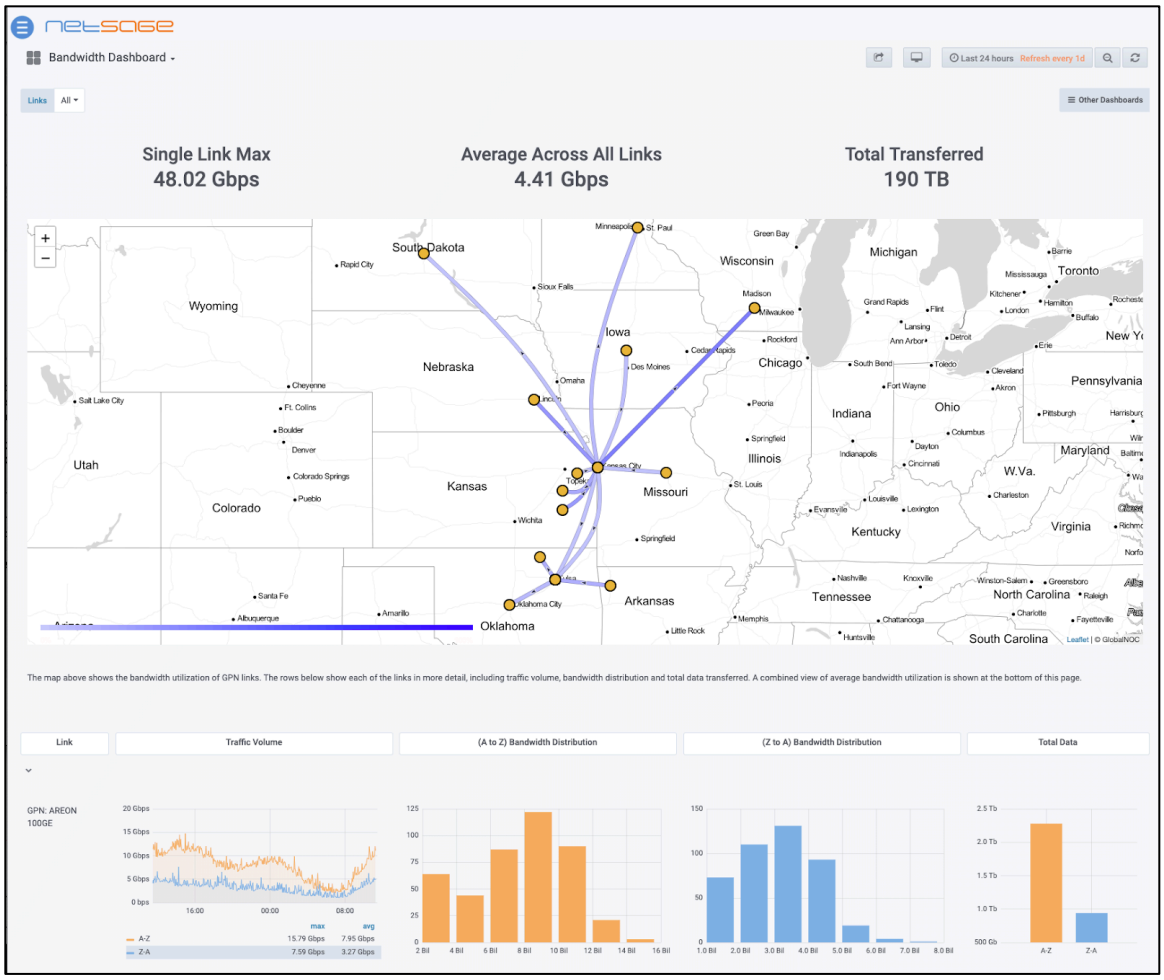


Figure 2: Screenshot of the bandwidth dashboard for the Great Plains Network collaboration.

5. Data Collection

NetSage staff are involved in the development and deployment of various pieces of software to support collecting active and passive measurements. This section details that work. In Year 5 we will continue to investigate possible additional data sources to respond to the questions asked by our end user community.

5.1 System Architecture

For Year 4, we continued to refine the architecture with an emphasis on leveraging existing open source components, including Elasticsearch, Logstash, and Grafana. We continued to make increasing use of Grafana.

During a review of our existing data ingest pipelines, we identified a few cases where we could replace early written custom code with the deployment and configuration of Logstash (part of the ELK stack). Substantial work was done to implement and test replacing sections of the pipeline with standard logstash configuration. During this time, we also identified and resolved an issue that could sometimes cause duplicate flow entries to be recorded.

We now support the use of longest prefix matching when tagging flows. With the Science Registry data, this allows us to provide overlapping entries of increasing specificity. For example, a large block of IP addresses might represent an institution, but a small subset of those IP addresses may represent a specific science division within the institution. Flows from that specific science division will now be correctly tagged as such even though they also match the more general institution record.

During Year 4, an evaluation of effort at IU identified a possible cost savings and opportunity for expanded archival capabilities by leveraging existing resources through the OmniSOC project, managed by the GlobalNOC. We have entered a contract with this group to support the Elastic search cluster for NetSage, all of the NetSage data was migrated in January. As expected, we saw considerable performance gains by moving to the new hardware, allowing for longer range queries and analysis. Additionally, this opens up the possibility of ingesting more data by allowing us to reduce the elephant flow threshold which we intend to pursue in the coming quarter.

5.2 Time Series Data System (TSDS)

The Time Series Data System (TSDS) (http://globalnoc.iu.edu/software/measurement/tsds.html) is a software suite that provides well-structured and high performance storage and retrieval of time series data, including interface throughput rates, flow data, CPU utilization, and number of peers on a router. Along with the raw data, the TSDS suite is capable of tracking and reporting based on metadata, for example viewing interface throughput from the viewpoint of a VLAN or BGP peer session of a particular ASN.

In the first three quarters of Year 4, we made seven releases of the TSDS Grafana integration driver to continue to support the needs of the NetSage project. These releases were largely focused around keeping pace with the Grafana environment and its 5.X release series, searching and templating, query efficiency, and overall polish and bug fixing.

In addition to the Grafana integration component, we also had a single release of the core TSDS code (1.6.1) that contained significant query optimizations and several bug fixes that were discovered through NetSage work, such as the handling of very large numbers in subqueries. Support was added for extrapolation queries to predict future link usage based on historical data. Support was also added to perform math operations using metadata fields, such as observed usage divided by total link capacity in order to represent percentage in use.

In Year 4 Quarter 4, we continued this trend of improving our Grafana TSDS integration software. Three releases (0.2.8 to 0.3.0) took place during this time period. These releases were primarily bugfix releases to handle specific issues, such as special character encodings or bugs introduced with specific combinations of

features, such as sub querying and aggregation. For more information, please see: https://github.com/GlobalNOC/tsds-grafana/releases.

Additionally, during Year 4 Quarter 4, a substantial release to the core TSDS code was made (1.6.2). This release was largely focused on optimizations such as reducing the memory footprint when running large queries, reducing the overall number of database connections, and in some cases, removing unneeded queries entirely due to new features available in MongoDB. For more information, please see: https://github.com/GlobalNOC/tsds-services/releases.

We expect that Year 5 will include the same types of maintenance releases as seen in Year 4.

5.3 Simple Network Management Protocol (SNMP)

The Simple Network Management Protocol (SNMP) is an application–layer protocol defined in RFC1157 for collecting and organizing information about managed devices on IP networks. SNMP is used by routers and switches to monitor networks for conditions that warrant administrative attention. This data is commonly collected and openly archived by most R&E networks.

In the first three quarters of Year 4, we added SNMP collections for the new circuits for the TransPAC router in Hong Kong. In Quarter 4, we added SNMP collections for two new 100G PIREN links from Guam to Hawaii, and from Hawaii to Los Angeles.

5.4 perfSONAR

perfSONAR (http://www.perfsonar.net/) is a network measurement toolkit designed to provide federated coverage of paths, and help to establish end-to-end usage expectations. The NetSage project uses perfSONAR for its active measurements of bandwidth and throughput, and archives them in the NetSage archive using TSDS. The IRNC suite of projects participate in the IRNC perfSONAR mesh, available at http://data.ctc.transpac.org/maddash-webui/index.cgi?dashboard=IRNC%20Mesh.

In the first quarters of Year 4, multiple releases of the perfSONAR toolkit were released (4.0.2.* to 4.1.3) and deployed onto the NetSage infrastructure. Several team members are also involved in the production perfSONAR consortium and contributed to these releases directly. This included going to the perfSONAR face-to-face All Hands Meeting March 7-8 in Amsterdam, Netherlands, to assist in plotting the course of development for the coming year.

New perfSONAR testpoints in Guam and Hong Kong were deployed and integrated into the IRNC test mesh. These new perfSONAR hosts provided active testing across the new TransPAC4 Guam-Hong Kong circuits and the PIREN Hawaii-Guam circuits. These deployments were able to immediately identify performance issues in the network that required troubleshooting, showcasing the value in using perfSONAR.

During Quarter 3, there were several long-term outages for IRNC-associated perfSONAR nodes, which led to a series of discussions about maintenance and support for these resources. As part of this, the CENIC-supported Sacramento and Los Angeles nodes were replaced. In addition, these outages highlighted a long-term issue between the IRNC NOC PET team, which was nominally supporting the mesh although only as best effort and with very little time to give, and the IRNC NetSage project, which relied on the data. Subsequently, the NOC project determined perfSONAR support was out of scope for them, so the NetSage team picked up the effort given the importance of the data to the tool. Additional staff time was added to help maintain the resource.

In Quarter 4, we began the process of re-integrating the perfSONAR test points in the AmLight network, which had been made private to their own network. This process is still ongoing at this time with AmLight network engineers to define a suitable plan and implement it.

We were also able to identify and fix an issue with overreporting on latency data based on some changes in the perfSONAR code. These changes were causing impossibly high latency values to be reported and were observed in the latency dashboard https://portal.netsage.global/grafana/d/000000005/latency-patterns?orgId=2.

5.5 Flow Data (sFlow, NetFlow)

Network Flow data collected using NetFlow or sFlow data consists of IP traffic information to better understand network traffic is coming from and going to and how much traffic is being generated.

During the first three quarters of Year 4, we added flow collections for the new TransPAC Hong Kong-Guam deployment. This is in addition to the existing flow data for the NEAAR New York-London circuit, the AMPATH circuits, and the TransPAC Seattle-Tokyo circuits, as well as the AmLight and CENIC exchange points.

Conversations continued with the PIREN team, who need to have sign off from their Australian contacts in order to gather flow data on the Hawaii-Australia links. Due to limits in project bandwidth and a general unwillingness to communicate, we are no longer pursuing the collection of flow data from the StarLight project.

During Quarter 3, prior to the release of the flow data dashboards, significant data cleaning took place to make the flow data clean enough for representation. The largest effect was through an update in the GeoIP DataBase to update the WhoIS records used, correcting many mislabelings. This clean up was implemented and all prior data was processed to update the correctness.

Currently, we only collect information on flows greater than 500M. We re-evaluated that limit and have decided to adjust this cap to 1M. In doing so, we will be able to collect information about 98% of the volume of data on the links, which turns out to be only 2% of the flows. In Quarter 4, additional equipment was purchased to support this extension, along with shifting the support of our flow archives to be a subcontract to the OmniSOC contract. This has enabled a more professional support service for the archive, and also freed up internal resources. Data was migrated into this new cluster and finalized during the end of Quarter 4. As expected, this new cluster has enabled longer range queries and analysis on the data already present in the flow dashboards. We anticipate being able to make progress on reducing the threshold of 500M elephant flows shortly.

5.6 Tstat on Backbones and Exchange Points

Tstat (http://tstat.polito.it/) is part of the EU Measurement Plane (mplane) FP7 project developed by Munafó and Mellia at Politecnico di Torino. Tstat can be used to analyze either real-time or captured packet traces, and rebuilds each TCP connection by looking at the TCP header in the forward and reverse direction. Tstat reports a number of useful TCP flow statistics, including congestion window size and number of packets retransmitted, which can be used to analyze the health and performance of the link.

In Year 4, we had discovered that Tstat does not collect data on flows that have asymmetric paths, which unfortunately includes over 90% of the flows on the NEAAR New York-London circuit and a significant portion on other backbone lines. We are currently discussing options in this space. As such, discussions for additional Tstat deployments on circuits or exchange points have been put on hold. We expect to complete this evaluation early in Year 5, which may include the deployment of an alternative to Tstat for circuits.

5.7 Tstat Data Collection from Archives

We are experimenting with collecting Tstat directly on a number of archives to provide additional insight on the overall health and performance of data transfers. This work is not impacted by the asymmetry issues we see with the circuit instrumentation since one end point is the archive itself.

In Year 4, we had planned to expand the deployment of Tstat to additional archives associated with UH Mānoa, CENIC, and PRP. While conversations between teams took place, this did not move forward. Through other channels, we began a conversation with the Texas Advanced Computing Center (TACC), and at the end of Quarter 4, Tstat was successfully deployed on several DTN boxes at TACC. A dashboard was provided to them using that data, and initial feedback was very positive. We expect this to be released publicly early in Year 5.

Work with the Tstat archive data will be highlighted strongly in Year 5. This will include extended analysis and visualizations but also additional deployments. We

are moving forward with the UH Astronomy group as of January, and are in the process of purchasing hardware for this set up.

5.8 Science Registry

The Science Registry is a system we have developed to document known network endpoints, organizations, and science projects that are users of network resources. The system supports collaborative and crowd sourced data entry and is a key component for finding higher fidelity information about endpoints than what existing WhoIs databases can provide. The science registry data is a key part of the de-identification pipeline that lets us tag flow data without retaining personally identifiable information (PII) in the form of IP addresses.

Work in the first three quarters consisted of continued development resulting in three releases of the underlying software as well as continued low level efforts to manually manage content based on TransPAC top talker activity. More information is available at: https://github.com/netsage-project/resourcedb/releases.

The pages describing resources and organizations were made available to the public, while also adding authentication requirements for making changes to the data. A series of administrative pages were created to help with managing the data. In addition, an internal dashboard was created to track the data coming into the Science Registry to help understand the extent of the coverage of current flows. There was also work done in preparation for the conversion to using logstash for the flow ingestion pipeline, which uses the science registry data. In particular we added proper utf8 support for international characters and added a new type of export format that can be ingested by logstash.

During Quarter 4, the science registry tagging piece of the flow ingestion pipeline was converted to using logstash, making use of the earlier preparatory work. Now that we have a basic threshold of data in place, we are evaluating ways to represent multi-use facilities, how to represent which organization owns the location vs the data, and a more concrete list of science disciplines based on the NSF recognized ones located here https://www.nsf.gov/about/research_areas.jsp.

In order to better track and understand our progress on this, we created an internal use Grafana dashboard that examines all of the flows ingested by the project and shows how they were tagged by the science registry. Substantial progress has been made on adding entries into the database as illustrated by the Figure 3.

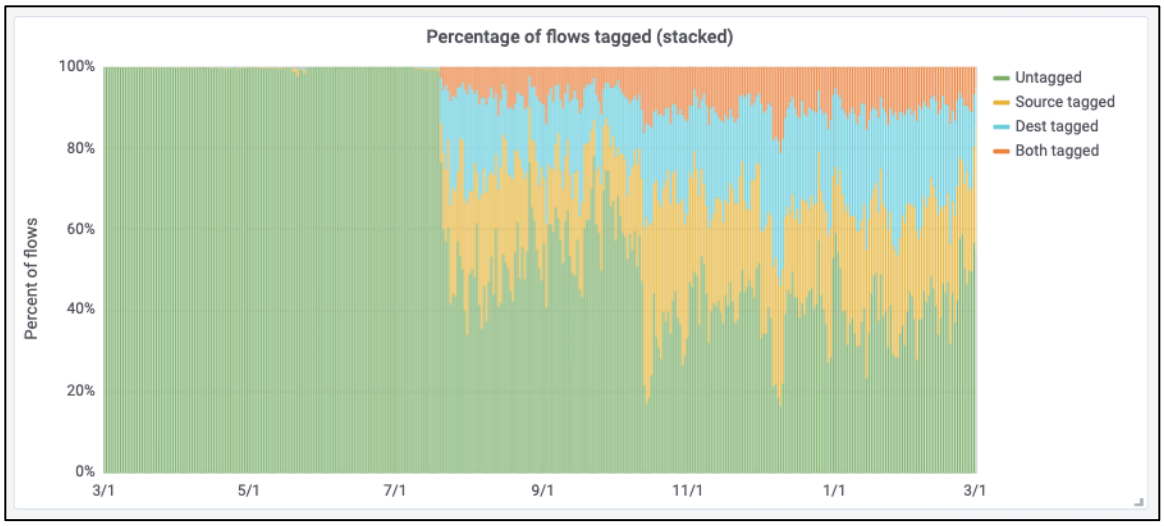


Figure 3: Screen shot showing coverage of science registry for data ingested by NetSage.

We anticipate that Year 5 will see similar release to support additional features as we saw in Year 4. We will continue to work with the end projects for additional data. This will include working directly with science groups as well, for example, the UH Astronomy team. In part, the data ingest portion of the science registry is dependent on crowd-sourced resources from the other IRNC projects.

5.9 File Transfer Performance Data

In Year 4, the project is investigating the inclusion of information from actual data transfers as an additional data source to be added to the testpoint. Currently, some of the Pacific Research Platform (PRP) are using Fiona nodes to understand larger-scale GridFTP data transfer behaviors, which are then displayed in a mesh, similar to that used with standard perfSONAR measurements. We are tracking the work by the perfSONAR consortium to expand the current set of perfSONAR tests (using the pScheduler extension) to begin gathering data from actual file transfers.

This work was originally expected to be part of the standard perfSONAR release in September 2018 in version 4.1 but due to time constraints has been moved back to 4.2, which is expected in early 2019. We plan to include updates of testpoints to collect this data in Year 5.

6. Visualization And Analysis

With the move to a Grafana-based front end, the line between visualization and analysis has become blurred, so we discuss these items together in a single section and list them by dashboard type.

6.1 General Infrastructure and Maintenance

As part of the January release of dashboard, we simplified the navigation bar interface significantly by removing unneeded space and Grafana widgets that were once required by Grafana but aren't any longer. We developed a spinning

"hamburger" button to enable the user to unfold a menu of questions that NetSage can answer, so that the selection of an individual question would take the user to the relevant NetSage visualization page. This question interface was designed to be flexible and extensible so that it can be easily added to as more dashboards are developed. These combined improvements are show in Figure 4.

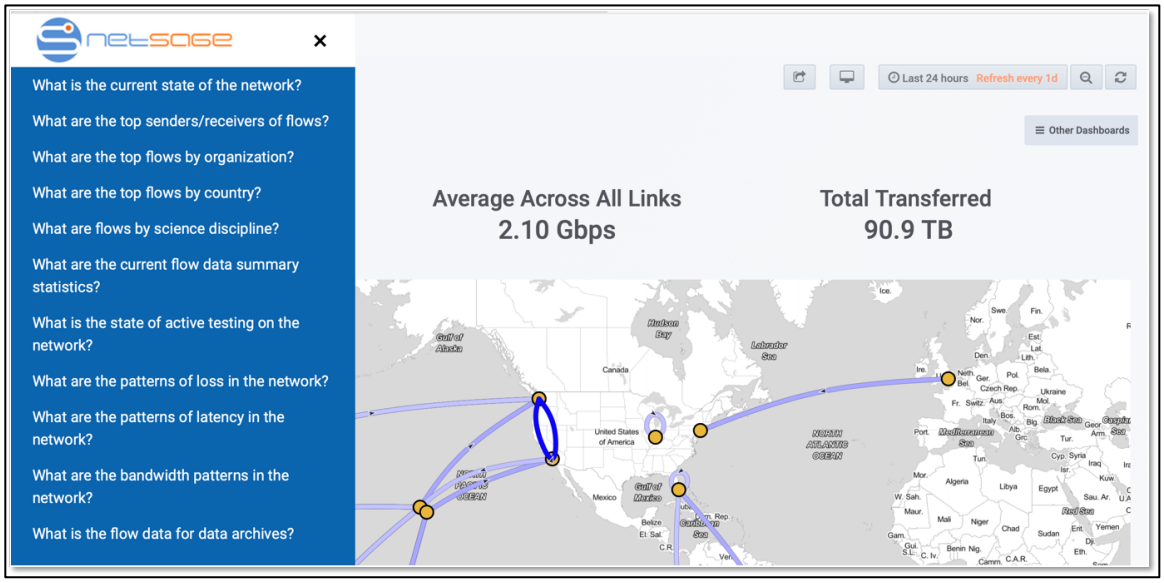
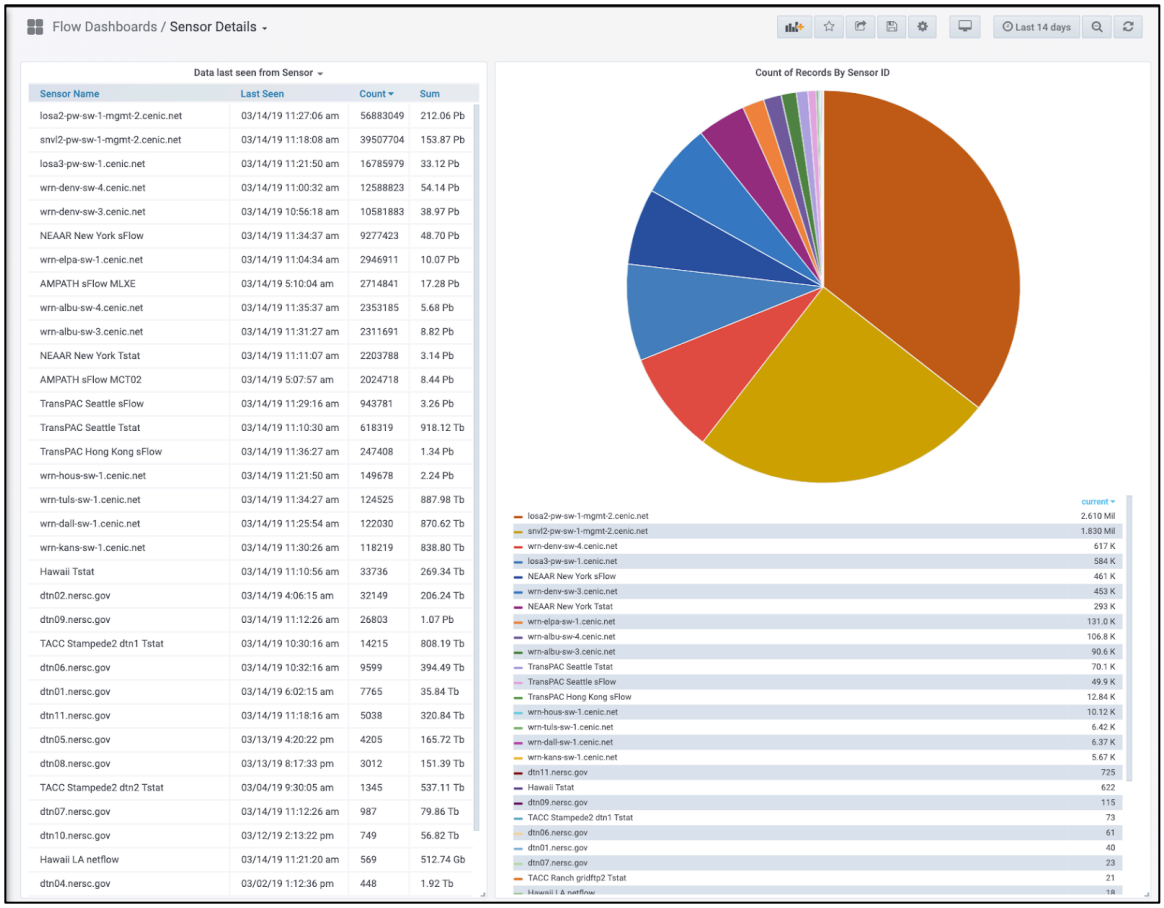


Figure 4: NetSage showing the new user interface to improve navigation between queries.

In Year 5, we will develop a few overall tools to be used across the dashboards. This will include a template, so that when updates are made to a menu or footer, it will only need to be done in one place. This feature is currently not available in Grafana.

We have created an internal dashboard to ensure that we understand the volume and number of sensors and this functionality as the project continues to scale up. Figure 5 shows an early version of this, which we anticipate will be expanded and possibly made public in the future. This dashboard provides at a glance views for when data was last seen from a sensor, which, in addition to active monitoring checks, provides an easy way to see whether there may be problems or when particular sensors may have gone away. It also provides some information on the volume of data we are seeing from sensors.

We are also in the process of developing a general IRNC Statistics dashboard which will enable IRNC PIs to see information about coverage data, for example how many countries, sources, or destinations are associated with a set of links or how many unique AS pairs for flows have been seen.



Sensor Name	Last Seen	Count	Sum
losa2-pw-sw-1-mgmt-2.cenic.net	03/14/19 11:27:06 am	56883049	212.06 Pb
snvl2-pw-sw-1-mgmt-2.cenic.net	03/14/19 11:18:08 am	39507704	153.87 Pb
losa3-pw-sw-1.cenic.net	03/14/19 11:21:50 am	16785979	33.12 Pb
wrn-denv-sw-4.cenic.net	03/14/19 11:00:32 am	12588823	54.14 Pb
wrn-denv-sw-3.cenic.net	03/14/19 10:56:18 am	10581883	38.97 Pb
NEAAR New York sFlow	03/14/19 11:34:37 am	9277423	48.70 Pb
wrn-elpa-sw-1.cenic.net	03/14/19 11:04:34 am	2946911	10.07 Pb
AMPATH sFlow MLXE	03/14/19 5:10:04 am	2714841	17.28 Pb
wrn-albu-sw-4.cenic.net	03/14/19 11:35:37 am	2353185	5.68 Pb
wrn-albu-sw-3.cenic.net	03/14/19 11:31:27 am	2311691	8.82 Pb
NEAAR New York Tstat	03/14/19 11:11:07 am	2203788	3.14 Pb
AMPATH sFlow MCT02	03/14/19 5:07:57 am	2024718	8.44 Pb
TransPAC Seattle sFlow	03/14/19 11:29:16 am	943781	3.26 Pb
TransPAC Seattle Tstat	03/14/19 11:10:30 am	618319	918.12 Tb
TransPAC Hong Kong sFlow	03/14/19 11:36:27 am	247408	1.34 Pb
wrn-hous-sw-1.cenic.net	03/14/19 11:21:50 am	149678	2.24 Pb
wrn-tuls-sw-1.cenic.net	03/14/19 11:34:27 am	124525	887.98 Tb
wrn-dall-sw-1.cenic.net	03/14/19 11:25:54 am	122030	870.62 Tb
wrn-kans-sw-1.cenic.net	03/14/19 11:30:26 am	118219	838.80 Tb
Hawaii Tstat	03/14/19 11:10:56 am	33736	269.34 Tb
dtn02.nersc.gov	03/14/19 4:06:15 am	32149	206.24 Tb
dtn09.nersc.gov	03/14/19 11:12:26 am	26803	1.07 Pb
TACC Stampede2 dtn1 Tstat	03/14/19 10:30:16 am	14215	808.19 Tb
dtn06.nersc.gov	03/14/19 10:32:16 am	9599	394.49 Tb
dtn01.nersc.gov	03/14/19 6:02:15 am	7765	35.84 Tb
dtn11.nersc.gov	03/14/19 11:18:16 am	5038	320.84 Tb
dtn05.nersc.gov	03/13/19 4:20:22 pm	4205	165.72 Tb
dtn08.nersc.gov	03/13/19 8:17:33 pm	3012	151.39 Tb
TACC Stampede2 dtn2 Tstat	03/04/19 9:30:05 am	1345	537.11 Tb
dtn07.nersc.gov	03/14/19 11:12:26 am	987	79.86 Tb
dtn10.nersc.gov	03/12/19 2:13:22 pm	749	56.82 Tb
Hawaii LA netflow	03/14/19 11:21:20 am	569	512.74 Gb
dtn04.nersc.gov	03/02/19 1:12:36 pm	448	1.92 Tb

Figure 5: Screen shot of internal dashboard for sensor health.

6.2 Bandwidth Dashboard (Basic Dashboard)

The Bandwidth Dashboard (https://portal.netsage.global) is the initial dashboard users see for the IRNC NetSage work. It answers the basic questions:

- How heavily used are the IRNC Backbones at this time, or for a set time period?
- What is the single link maximum throughput and the combined performance for the IRNC backbone links?
- What is the performance over a circuit, both incoming and outgoing, over time?
- How much data has been transferred for the time period selected for each link and for all the links combined? (Default 24 hours)

In Year 4, incremental improvements were made to the dashboard, such as incorporating the PIREN and new TransPAC Hong Kong-Guam links into the map and updating related charts to include data from those links. We re-arranged some of the data to be more consistent with other dashboards and also added code so the coloring in the bottom comparison graphs for the different circuits was consistent between the graphs.

6.3 Heatmap Dashboards

We use heatmap visualizations to show changes in values over time and to easily identify patterns of behavior. We currently have the dashboards that use this technique to answer the following questions:

- What are recurring patterns of network latency using active data- available at https://portal.netsage.global/grafana/d/000000005/latency-patterns?orgId=2.
- What are the recurring patterns of network loss using active data- available at https://portal.netsage.global/grafana/d/000000006/loss-patterns?orgId=2.
- What are the recurring patterns of bandwidth, using SNMP data - available at https://portal.netsage.global/grafana/d/000000004/bandwidth-patterns?orgId=2.

In Year 4, a problem relating to unresponsive perfSONAR measurement archives that affected the perfSONAR Heatmap dashboard was resolved. The problem was caused by scripts failing to complete in time when queried.

For the general heatmap approach, during the January Hack-a-thon we spent time adjusting the color differential to make these more effective within their context. For example, loss can be a value between 0 and 100%, but even 1% loss needed to be much more strongly highlighted as that indicated a significant problem. These updates were part of the January releases.

In Year 5, we plan to use this technique to show patterns of behavior for transfers between countries as well as for retransmit behavior.

6.3 Sankey Graph Dashboards

Sankey diagrams are a specific type of flow diagram in which the width of the arrows is shown proportionally to the flow quantity. We are continuing the adaption of Sankey graph techniques for the Grafana framework and as a novel way of showing how data flows over the IRNC networks. Specifically, the graphs are intended to answer these networking questions:

- What is the distribution of the volume of data among flows between countries?
- Which ports and protocols are most used?
- What are the sources and destinations of these flows?

In Year 4, Sankey diagrams went from design to full implementation. We used this opportunity to create a generalized Grafana framework to facilitate the rapid integration of future visualization tools. (https://github.com/uhmlavalab/netsage-grafana-boilerplate-plugin).

Figures 6, shows the first conceptual designs of the Sankey diagram and explain how they can be used to depict 5 dimensions of network data on a single chart- such as source country and continent, destination country and continent, traffic volume.

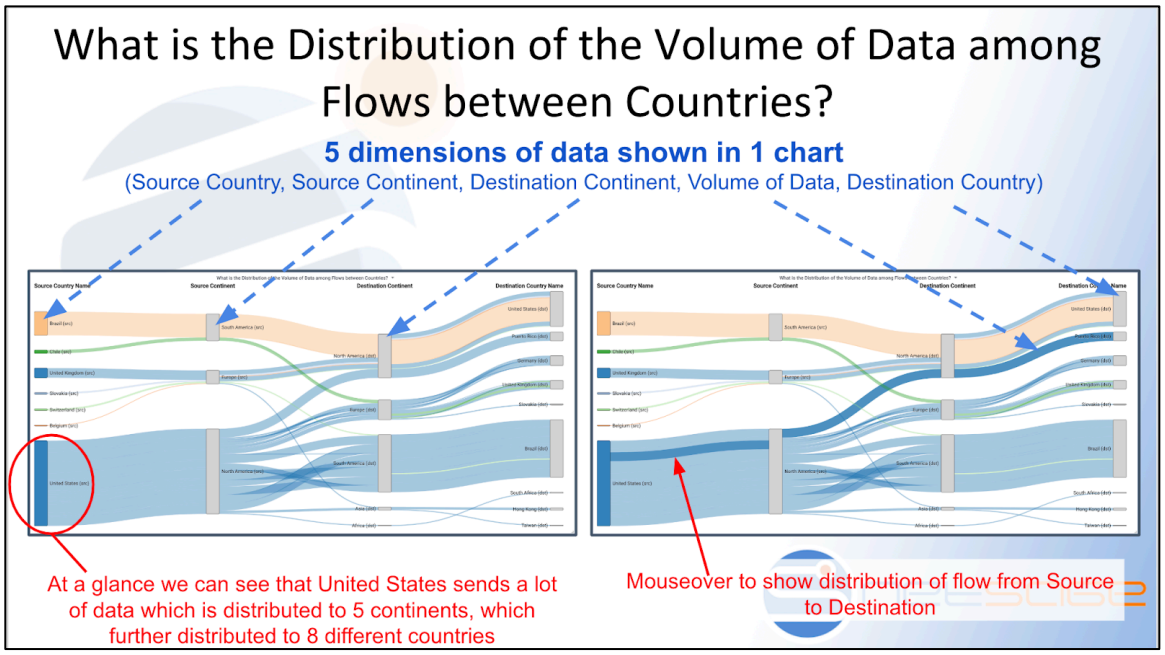


Figure 6: Sankey prototype shows 5 dimensions of network flow data.

The fully working Sankey Dashboard for showing flows by Science Discipline, depicted in Figure 7, was demonstrated at SC'18, and then released publicly in January. It can be viewed at: https://portal.netsage.global/grafana/d/WNn1qyaiz/flows-by-science-discipline?orgId=2.

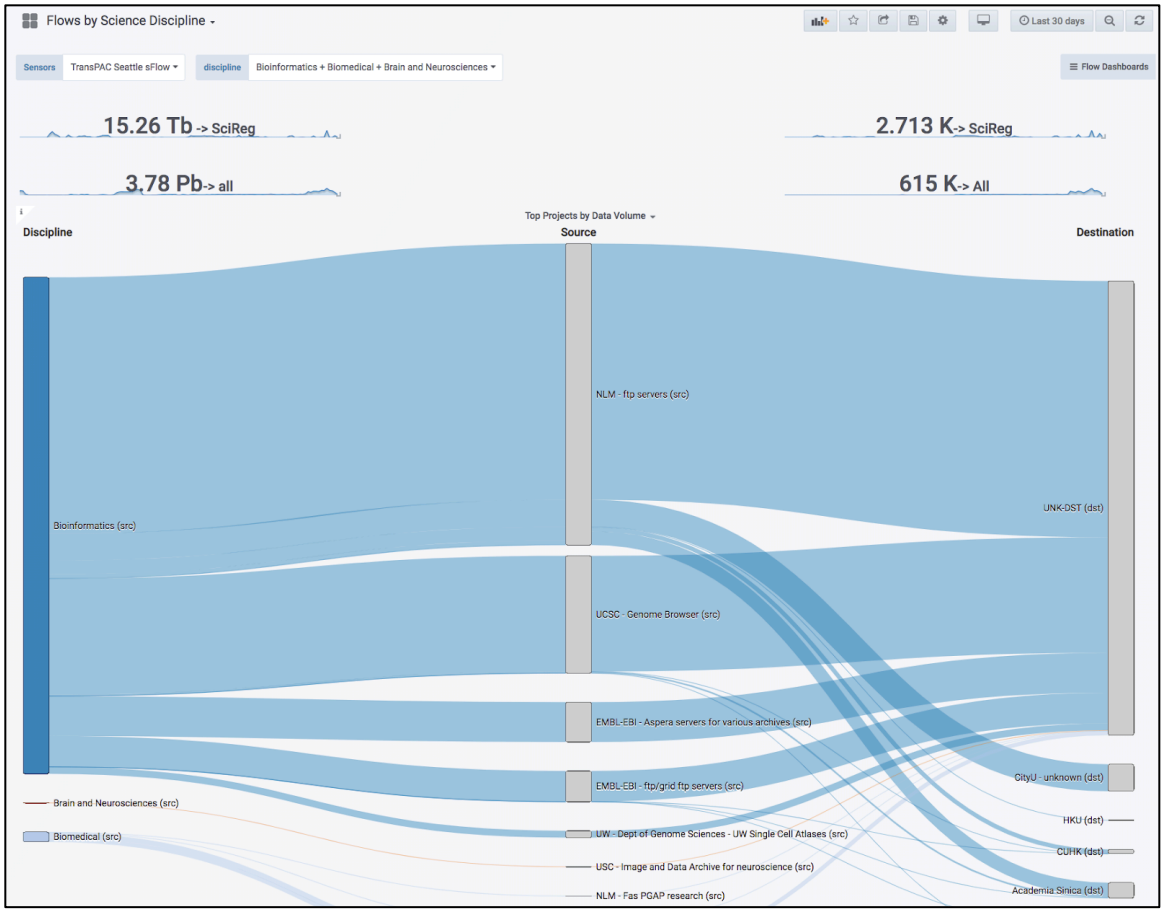


Figure 7: Screen shot for Flows by Science Discipline dashboard.

All codes were transferred to the official NetSage github repository: (https://github.com/netsage-project/grafana-sankey). In Year 5, we expect to release additional dashboards featuring this visualization technique.

6.4 Basic Flow Data Dashboard(s)

The initial Flow Data Dashboards were released in October, with a follow-on release in January. They answer the basic questions:

- Who are the top ten senders/receivers of flows by institutions, countries, or ASN (ranked by volume or rate) by source and destination - shown in Figure 8 and available at https://portal.netsage.global/grafana/d/xk26IFhmk/flow-data?orgId=2
- What are the top flows by organization - shown in Figure 9 and available at https://portal.netsage.global/grafana/d/QfzDJKhik/flow-data-per-organization?orgId=2
- What are the top flows by country? https://portal.netsage.global/grafana/d/fgr0zz_mk/flow-data-per-country?orgId=2
- What are the current flow data statistics - shown in shown in Figures 10 and 11, and available at

https://portal.netsage.global/grafana/d/CJC1FFhmz/other-flow-stats?orgId=2

- What is the flow and retransmit data for archives - shown in Figures 12 and 13, and available at https://portal.netsage.global/grafana/d/mNPduO8mz/flow-data-for-data-archives?refresh=15m&orgId=2

In Year 4, work focused on innovative displays of flow behaviors. A soft release of flow dashboards was shown to IRNC project staff at the Internet2 Global Summit in May, after which the feedback was used finalize a public release in early October, and publicly announced at the Internet2 Technical Exchange meeting. A presentation given by Schopf highlighted the flow data dashboard's functionality, which was also used by the IN@IU team in several other meetings with application science groups to highlight their use of the IRNC-funded international links.

During Quarter 4 at the All Hands Meeting Hack-a-thon, a number of quality of life improvements were made to the dashboards. For the flow dashboards, this consisted of re-arranging the way the panels were presented to be more consistent across dashboards, expanding the flows by country view to include top organization pairs within that country, and the addition of two new dashboards to show Flows by Science Discipline based on science registry data as well as Flow Data for Data Archives to make use of the Tstat instrumentation on DTNs.

In Year 5, we will extend our visualizations using flow data in response to the additional questions requested by the IRNC PIs and other users.

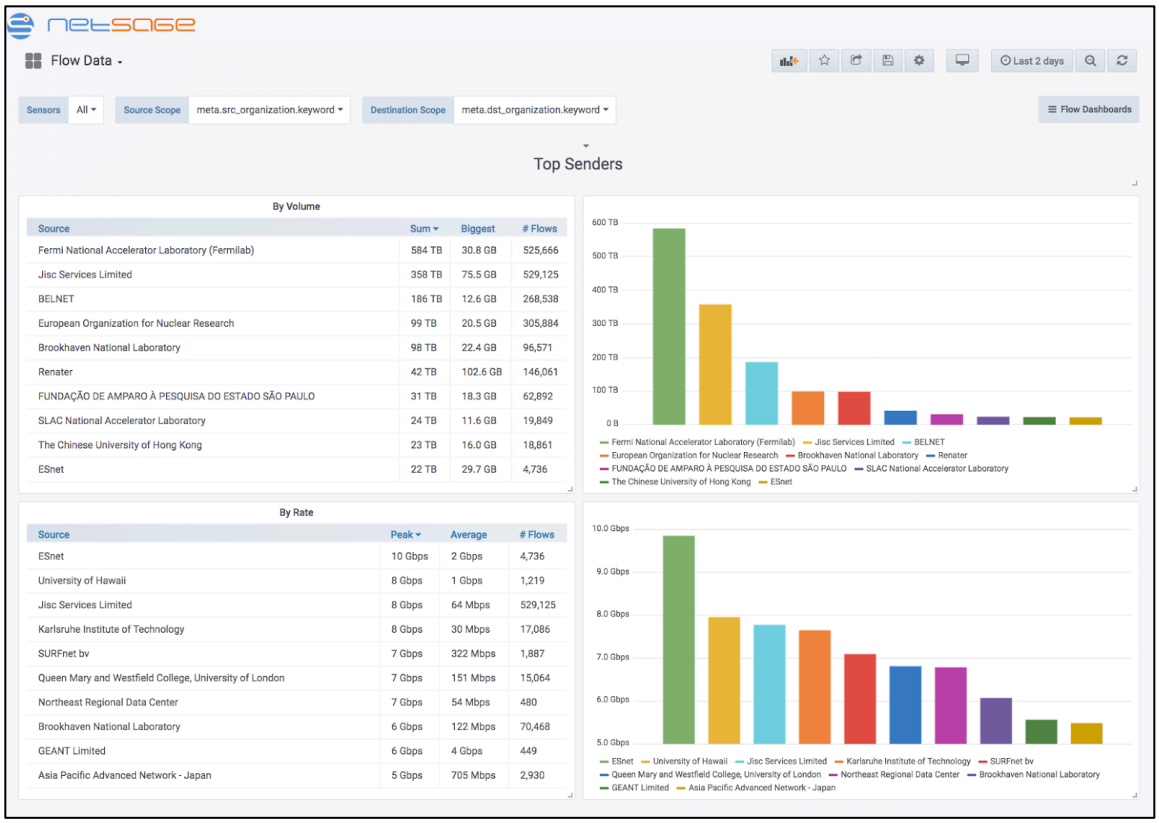


Figure 8: Screen shot of IRNC circuit Flow Data Dashboard.

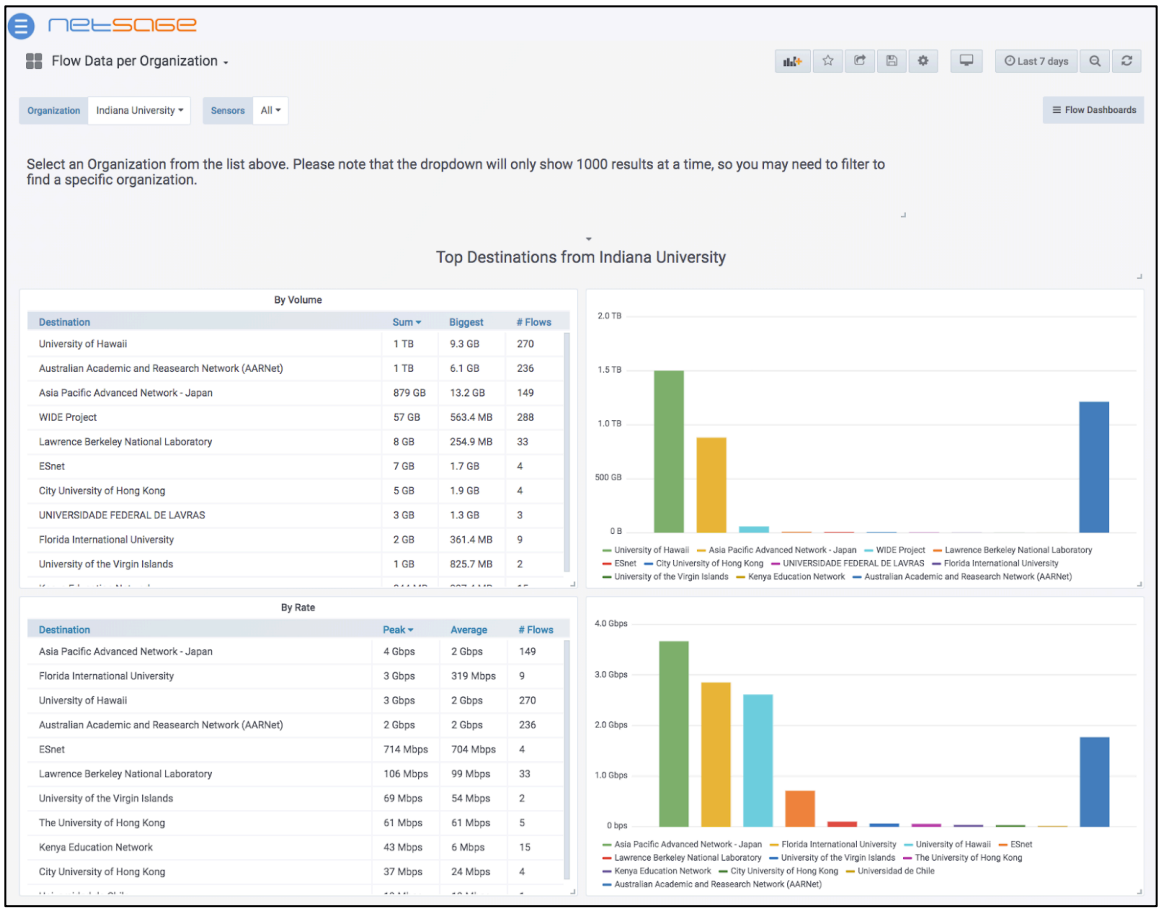


Figure 9: Screen shot of the Flow Data by Organization Dashboard, for Indiana University as the selected organization.

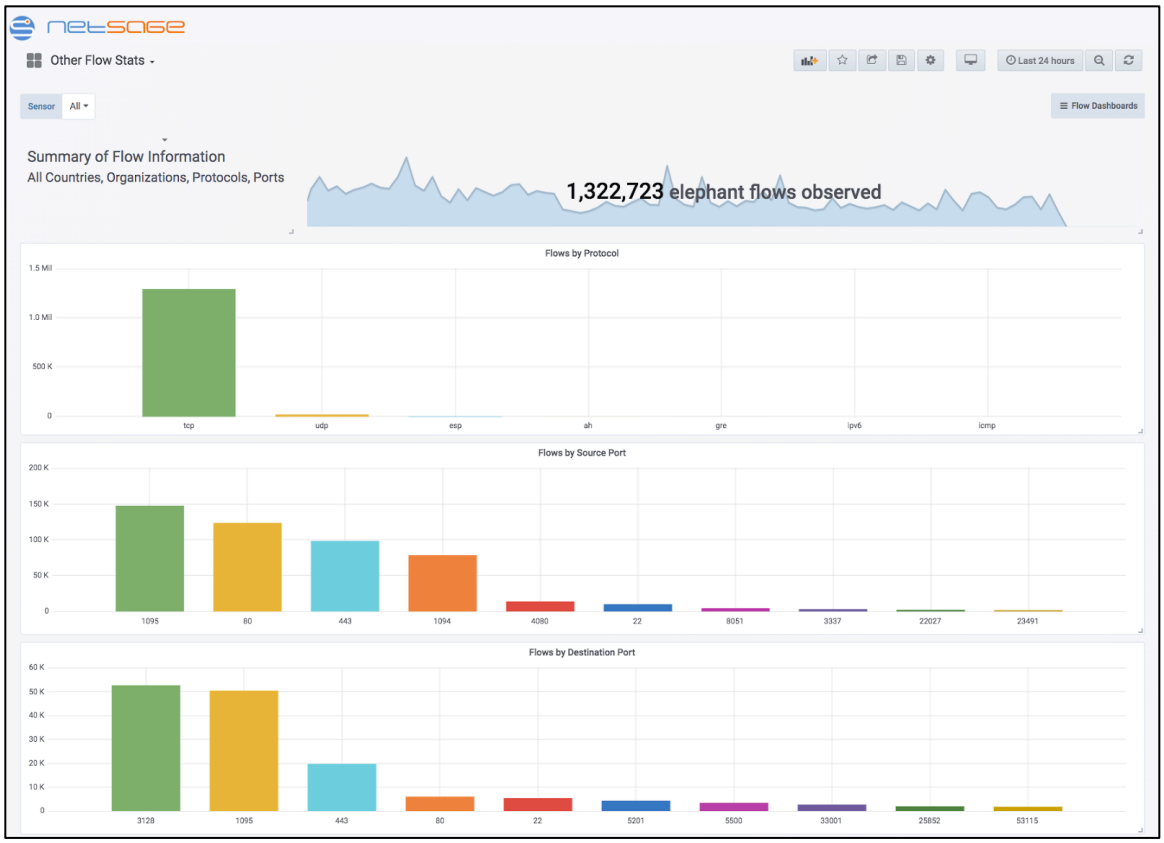


Figure 10: Screen shot for the Other Flow Statistics Dashboard.

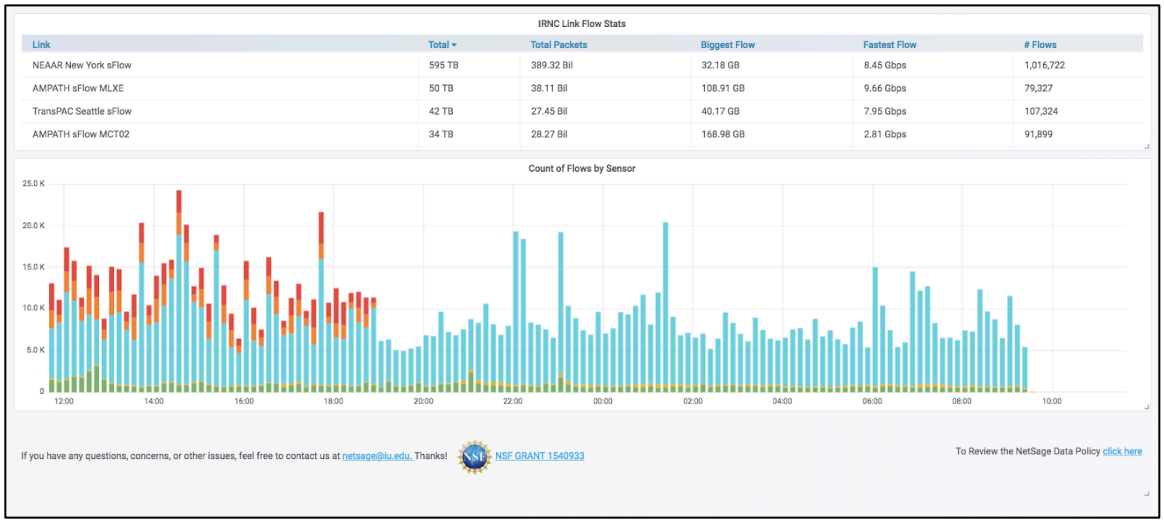


Figure 11: Screen shot for the Other Flow Statistics Dashboard.

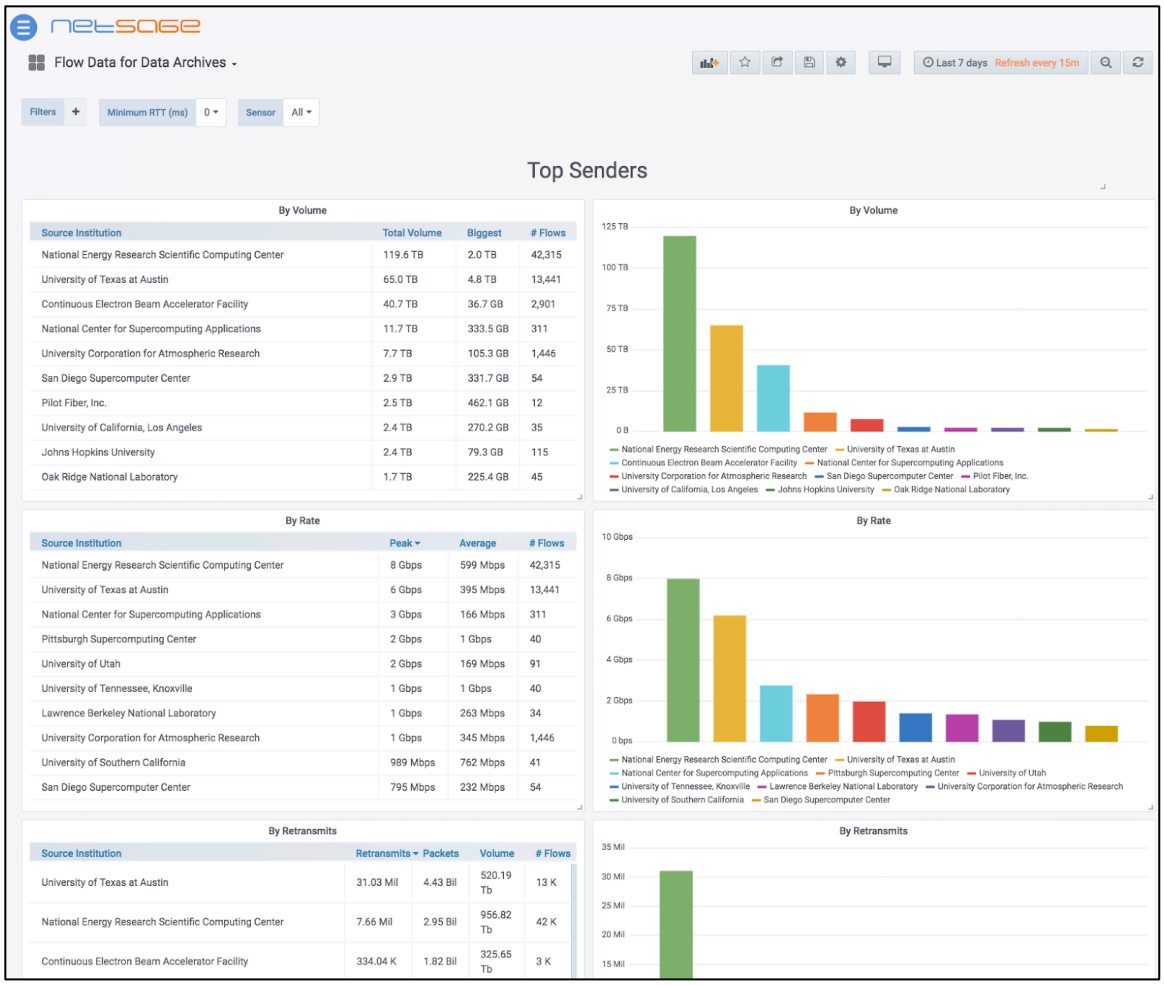


Figure 12: Screen shot for Flow Data for Data Archives Dashboard.

Figure 13: Screen shot for Flow Data for Data Archives Dashboard

6.5 Additional Map Work

Work continued on the map visualization widget for Grafana with two major releases. Many quality of life enhancements were made, including the ability to apply instantaneous configuration changes to allow users to immediately obtain feedback on their changes to the map.

Additional improvements were made to enable the map's reporting capabilities. For example, map links can be configured to show not only data averages but minimums and maximums, or instantaneous values. The map's legend was also improved to allow the easy creation of custom color palettes and the inversion of legend values if necessary. Lastly, map definition information was moved into the Grafana graphical editor to enable full containment of all data for configuring the look and feel of maps, rather than relying on external files as in the past. Work also occurred to support non-geographical maps, such as logical views of campus networks, as well as supporting the visualization of bi-directional traffic.

A darker background scheme for the map was tested at SC'18, as shown in Figure 14, but ultimately replaced due to contrast challenges on some displays. A broad set of bug fixes were made to this dashboard as part of the January 2019 AHM Hack-a-thon, and the updated dashboard was included in the January release. We also added in a feature to explicitly show the IRNC exchange points and their live data.

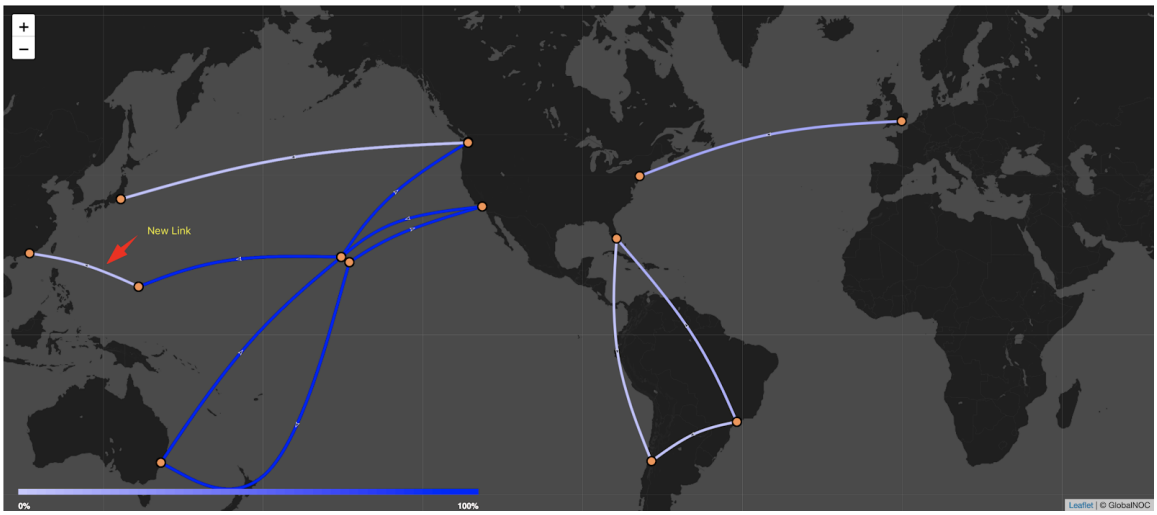


Figure 14: Prototype of color shifted map including added PIREN and TransPAC Hong Kong-Guam links that was rolled back due to issues with contrast.

In addition, a perfSONAR based map dashboard was also release in January. This is available at https://portal.netsage.global/grafana/d/000000033/loss-and-throughput-active-testing?refresh=1h&orgId=2.

In Year 5, we plan to explore the possibility of using this approach to show data transfers associated with the science registry.

6.6 Additional Analysis Work and Bespoke Dashboards

6.6.1 Alarms and Alerts for the NOC

After an additional set of conversations with the IRNC NOC about how they might use NetSage with alarms and alerts to help them support the IRNC networks, the

IRNC NOC stated this tool was not needed to support their mission. These work items are being retired.

6.6.2 Evaluation of Retransmit Predictions

We concluded our work analyzing Tstat data from the NERSC and ESnet archives. We developed a statistical method that predicts packet loss, including identification of the weighting of individual factors, and attempts to trace the origin of those factors (path, end host, network). Our method leveraged a random forest technique for the following Tstat fields: throughput, size of the file being transferred, source and destination IP addresses, round trip time, duration of the flow, and TCP congestion window. The resulted showed smoothing techniques substantially reduced noise and also helped to improve accuracy. As expected, our results also showed a number of seasonal trends, e.g., pertaining to paper deadlines for major conferences that involve moving data in and out of NERSC.

We developed a prototype Grafana dashboard as an example of how the results of this analysis could be displayed, shown in Figure 15. The actual and predicted network loss are juxtaprosed in both an overview view and a zoomed-in view simultaneously that can help end-users see both detail and context in their data at a glance, rather than having to zoom in and out constantly as is the case for typical Grafana line charts.

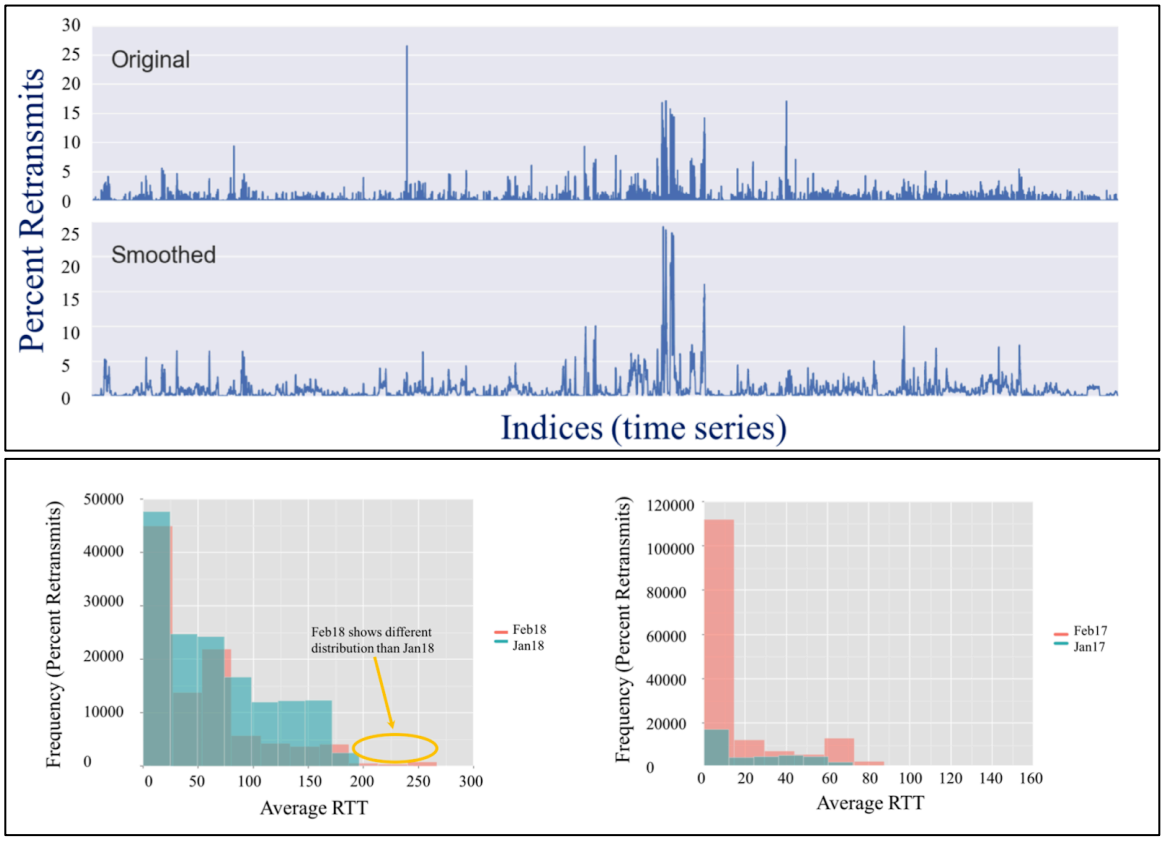


Figure 15: Chart juxtaposing actual and predicted network loss. Bottom of chart shows overview view of the time period under examination.

A paper on the Tstat analysis results was submitted to the 5th annual Innovating the Network for Data-Intensive Science (INDIS) workshop, on Nov. 11, 2018, co-located with SC18, on September 12, 2018. The paper was not accepted but reviews were positive and helpful and the paper is planned to be re-submitted to the journal *Future Generation Computer Systems' (https://www.journals.elsevier.com/future-generation-computer-systems)* Special Issue on "Innovating the Network for Data Intensive Science" for 2018 (https://www.journals.elsevier.com/future-generation-computer-systems/call-for-papers/innovating-the-network-for-data-intensive-science-indis-2018). All of the code for this work is available at https://github.com/netsage-project/tstat-dtn-analysis.

6.6.3 CENIC Dashboard

During Year 4, we were asked by IRNC-funded CENIC for a visualization of data related to several networks that they were starting to support as part of a supplemental grant. As the flow dashboards had not yet been put together, we were able to work with CENIC to design a temporary and private one off view of the data to meet their needs. This dashboard, shown in Figure 16, was provided at multiple times throughout the year to judge how CENIC's efforts were impacting overall traffic performance on the networks in question.

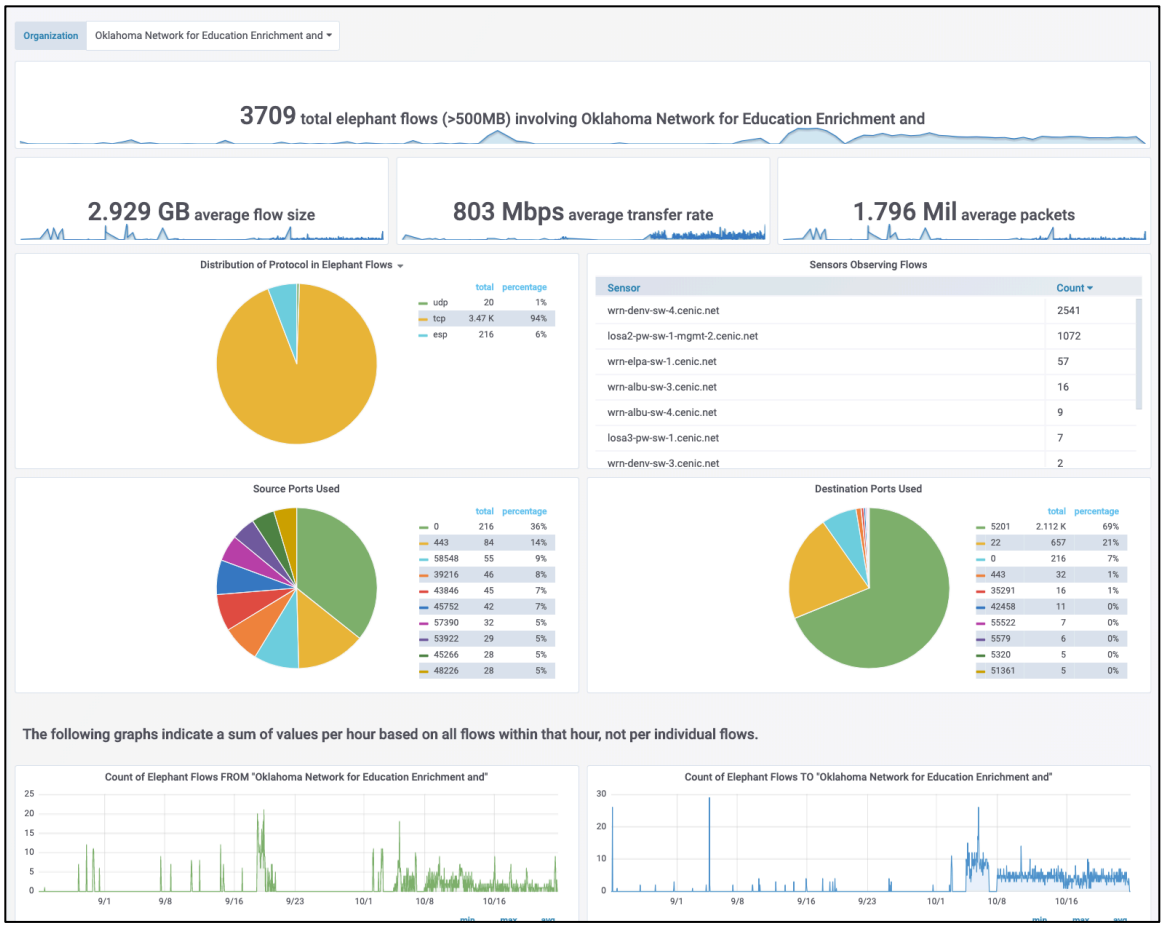


Figure 16: Example of custom dashboard design.

6.6.4 Third Party Dashboards

During Year 4, we received several requests to help groups outside of the IRNC collaborators to help them install NetSage or to support dashboards for them. This included staff from PRAGMA, the South African R&E Network (SANREN), the Asia Pacific Ring collaboration, and others. In Year 5 we will discuss ways to have a scalable approach to these requests.

7. Data Privacy and Security

Basic security measures are being maintained and there were no security incidents to report for any quarter in Year 4. As a reminder, NetSage does not collect PII. We began discussions with several groups about the upcoming role out in Europe of the General Data Protection Regulation (GDPR), but the consensus is that GDPR pertains to data about individuals, and NetSage data only gets down to the level of organizations, so GDPR constraints are not relevant.

We integrated Shibboleth and Grafana to enable 3rd party authentication to any dashboard that might collect sensitive information or that the PIs would wish to review before it became public. This integration work involved configuration modifications and testing against the GlobalNOC identity provider. In the end, no

code modification was required as the support provided within Grafana was adequate. However, many of the institutions for the IRNC PIs do not support Shibboleth, so this approach was not used in the end.

8. Year 5 Plans

The high level plans for Year 5 will focus on extending the data collected from instrumented archives and the science registry, and will address the following questions:

- 4.c - What are highest retransmit rates between organizations/subnets over a timeframe?
- 6.g- Who have been the top talkers each year (longitudinal study)
- 7.f - Is retransmit data a proxy for packet loss?
- 9.b - What level of retransmits is the archive experiencing during data flows?
- 9.c - What are the patterns for retransmits on an archive
- 9.d - For a pair of endpoints, what is the retransmit behavior for the individual flows?
- 11.a- What is the bandwidth of a GridFTP file transfer between two backbone end points?
- 12.a - How are active tests between two sites performing? (replacement for MaDDash)

8.1 Project Coordination

In Year 5, we plan to continue to coordinate with our partners much as we did in Year 4. Of note, we will continue the practice of having a one-day Hack-a-thon prior to our face to face All Hands Meetings as an effective was to clear out low hanging fruit and bug fixes for the various dashboards. We will continue to explore work with external partners to broaden the use of the basic NetSage framework as well.

8.2 Data Collection

We will continue to investigate the opportunity for new data sources in Year 5. Most critically, we will continue with our evaluation of Tstat to fully understand the asymmetry and performance limitations it may be experiencing, and to evaluate if there is an alternative solution to consider deploying. We will deploy the updated perfSONAR file transfer test infrastructure when it is released to the community and work to incorporate that data into existing and new dashboards. Basic maintenance for our internal tools, TSDS, perfSONAR and the Science Registry, will continue.

We will also work to expand the collection of the existing data sources. The primary area of expansion we expect in Year 5 is with Tstat on data archives. We expect significant interest from data archive owners when we make the Tstat for Data Archives dashboard public.

8.3 Dashboard Development

One of the larger projects in Year 5 will address some of the issues we have now that the number of dashboards is growing. We will develop a few overall tools to be used across the dashboards, starting with a base-line template to make updates and changes to the basic framework easier and more consistent. This feature is currently not available in Grafana, so we will also work with that development team to submit it for public use.

We will of keep up our basic maintenance of the existing dashboards, but we are actively gathering comments and feedback to develop new dashboards to answer additional questions. Much of these we expect to involve the additional data we will gather from Tstat archive deployments and visualizations of the science registry data. We also plan to devlop additional dashboards using Sankey visualizations for different types of flow data.

8.4 Third Party Use and Public Access

Year 4 saw the first instance of third-party use of the NetSage infrastructure with our deployment of the ANA dashboard and work with the EPOC team to adapt it for their domestic use. We expect this to continue to expand, and as such will be increasing our documentation and adding ease-of-use and ease-of-deployment features throughout the year.

We are beginning to receive requests by groups who would like to do a fully-independent deployment of the full infrastructure. While some groups, such as PRP, have deployed sensor aspects on their own, right now we do not have the documentation for a complete setup of the archive and grafana backend resources. This will take place in Year 5 as part of a possible hand-off activity in light of the upcoming IRNC solicitation and a possible re-bid for measurement services.

One aspect of this is the preparation for a potential hand off at the end of the grant period. This will include additional documentation of the systems set up. We've already had experience in moving the full data set, so this process is well documented.

8.5 Papers and Presentations

In Year 5, we will work towards submitting the paper from INDIS 2018 to the Special Issue on "Innovating the Network for Data Intensive Science" in the journal Future Generation Computer Systems (https://www.journals.elsevier.com/future-generation-computer-systems). We will also explore submitting to the SuperComputer State of Practice session.

9. Updated WBS for Year 4 and 5

Item	Y3 WBS	Notes
Data Collection	1	
PerfSonar Related Tasks	1.4	Ongoing
Define and deploy PS test mesh for backbones	1.4.2	Ongoing
Add in node for Honolulu-LA link	1.4.2.10	Completed Q2
Add in node for Honolulu-Guam link	1.4.2.11	Completed Q3
Ongoing support for IRNC PS mesh	1.4.3	Ongoing
SNMP related tasks	1.5	Ongoing
SNMP data from Backbones	1.5.2	Ongoing
Input PIREN-LA SNMP data	1.5.2.5	Completed Q2
Input PIREN GUAM SNMP data	1.5.2.6	Completed Q3
Tstat/Flow deployment	1.7	Ongoing
Input Ampath Flow Data	1.7.11	Ongoing
Purchase and deploy equipment if needed to support TSTAT at Ampath	1.7.11.3	On hold due to Tstat issues
Incorporate unsampled flow data (tstat) from Ampath	1.7.11.5	On hold due to Tstat issues
Input PIREN HI-Australia Flow/tstat Data	1.7.12	Ongoing
Talk to Lassner/David Wilde about Tstat and sFlow data being available	1.7.12.1	Waiting on PIREN since Y2
Purchase/Deploy flow equipment if needed for Tstat at PIREN	1.7.12.2	Waiting on 1.7.12.1
Incorporate sampled flow data from PIREN into TSDS	1.7.12.3	Waiting on 1.7.12.1
Deploy PIREN tstat data collection	1.7.12.4	Waiting on 1.7.12.1
Incorporate tstat data from PIREN into TSDS	1.7.12.5	Waiting on 1.7.12.1
Talk to someone about Guam flow/Tstat data	1.7.12.6	Waiting on 1.7.12.1
Talk to someone about HNL-LA Tstat/flow data	1.7.12.7	Waiting on 1.7.12.1
Input CENIC Flow Data (Year 3)	1.7.14	sflow completed, TSTAT on hold
Purchase/Deploy flow equipment if needed for CENIC	1.7.14.2	Waiting on 1.7.14
Deploy CENIC tstat data	1.7.14.4	Waiting on 1.7.14
Incorporate tstat data from CENIC into TSDS	1.7.14.5	Waiting on 1.7.14
Input StarLight Flow Data (Year 4)	1.7.16	Canceled
Talk to StarLight team about Tstat and sFlow data being available	1.7.16.1	Canceled
Input TP Guam-HK Flow Data	1.7.17	Completed Q3
Talk to TP team about Tstat and sFlow data being available	1.7.17.1	Completed Y1
Purchase/Deploy flow equipment if needed for TP	1.7.17.2	Completed Q3
Incorporate sampled flow data for TP Guam-HK into TSDS	1.7.17.3	Completed Q3
Deploy TP Guam-HK tstat data	1.7.17.4	Completed Q3
Incorporate tstat data from TP Guam-HK into TSDS	1.7.17.5	Completed Q3
Input PIREN Guam-HI-LA Flow Data	1.7.18	Ongoing
Talk to PIREN team about Tstat and sFlow data being	1.7.18.1	Ongoing

available		
Purchase/Deploy flow equipment if needed	1.7.18.2	Completed Y4Q4
Incorporate sampled flow data for PIREN Guam-HI-LA into TSDS	1.7.18.3	Completed Y4Q4
Deploy PIREN Guam-HI-LA tstat data	1.7.18.4	On hold due to Tstat issues
Incorporate tstat data from PIREN Guam-HI-LA into TSDS	1.7.18.5	On hold due to Tstat issues
Lower flow data collection threshold from 500M	1.7.19 (NEW)	Ongoing
Instrumentation of Data Archives	1.8	Ongoing
Generate an RPM and/or better documentation on how to install tools on archives that will forwards tstat data to IU	1.8.2	Needs final documentation
Deploy Tstat on Hawaiian astronomy archives	1.8.3	Ongoing
Deploy Tstat on CENIC/PRP archives	1.8.4	Planned Year 5
Use top talkers list to identify likely DTNs	1.8.5	Planned Year 5
Instrument NCAR Archive	1.8.6	Discussions started Y4Q4
JMS to follow up with ESIP NASA guy for NASA DTN instrumentation	1.8.7	Planned Year 5
Other possible DTNs from IRNC partners?	1.8.8	Ongoing
TACC Data archives	1.8.8.1 (NEW)	Completed Y4Q4
Additional software framework Upkeep	1.12	Ongoing
TSDS maintenance	1.12.1	Ongoing
Add a keep alive notification for Tstat sensors (modify package)	1.12.5	Ongoing
Upgrade archival storage and hand off maintenance to OMNISOC team	1.12.6 (NEW)	Completed Y4Q4
Re-evaluate Tstat load and collection issues	1.12.7 (NEW)	Ongoing
Data transfer information (ie fiona) as additional data source	1.15	Delayed by late PS release, planned Y5
Investigate approaches to including data transfer info	1.15.1	Completed
Decision about inclusion	1.15.2	Completed
Find guinea pigs for data transfer inclusion	1.15.3	Waiting for release of PS
Analysis	**2**	
Data cleaning	2.2	Ongoing
Recreate AS to Science Project Database (Science Registry)	2.4	In progress
input data	2.4.3	Ongoing
Get TransPAC to add data to science registry	2.4.3.1	Ongoing
Get NEAAR to add data to science registry	2.4.3.2	Ongoing
Get Ampath to add data to science registry	2.4.3.3	Discussed again Q3, Q4
Get PIREN to add data to science registry	2.4.3.4	After flow data collection
Get CENIC to add data to science registry	2.4.3.5	Planned Year 5
Get StarLight to add data to science registry	2.4.3.6	Canceled
Extensions to basic science registry framework	2.4.4	Ongoing

Add “short name” to flow tagging for organization names	2.4.4.1	Completed Q2
Science registry metadata exporter (for use by data pipeline)	2.4.4.2	Completed Q2
Flow tagging based on SR metadata	2.4.4.3	Completed Y4Q4
Read-only public mode for SR	2.4.4.4	Completed Q3
Form to submit changes to SR	2.4.4.5	Completed Y4Q4
More science disciplines and ability to edit list	2.4.4.6	Started Q3
More roles and ability to edit list	2.4.4.7	Year 5
Notes field for SR	2.4.4.8	Year 5
URL field for SR	2.4.4.9	Year 5
Admin section functionality for SR	2.4.4.10	Ongoing
Dashboard for SR data collection tracking	2.4.4.11 (NEW)	Completed Q3
Tstat Analysis scripts (non flow, retransmits etc)- walk through Kibana experiment	2.7	Completed Y4Q4
Dashboards for Tstat data from archive showing retransmit	2.7.1 (NEW)	Ongoing
Basic Tstat archive dashboard similar to flow data dashboard	2.7.1.1	Completed Y4Q4
Updates to Tstat archive dashboard	2.7.1.2	Ongoing
Heatmap for Tstat archive data	2.7.2 (NEW)	Year 5
In Depth analysis for packet loss	2.12	Q2 complete
Analysis for top X based on flow	2.13	Completed Q3
by organization	2.13.1	Completed Q3
By country	2.13.2	Completed Q3
By Protocol	2.13.3	Completed Q3
For each link	2.13.4	Completed Q3
Americas Greatest Networks - most reliable	2.13.5	Year 5
Soft release for IRNC PI meeting at I2 (May 2018)	2.13.6	Completed
Incorporate feedback from May 2018	2.13.7	Completed
Full release mid-2018	2.13.8	Completed Q3
Updates to top X based on flow	2.13.9 (NEW)	Ongoing
Filter out Australian university names from flow data	2.13.9.1	Ongoing
Analysis for top science projects	2.14	Completed Y4Q4
Analysis for elephant flows - min, max and duration	2.15	Duplicate of 2.13
Analysis of buffer size issues	2.16	Year 5
PIREN analysis for astronomy data	2.20.	Waiting on Tstat on UH archives
Evaluate moving average for additional smoothing in graphs	2.22	Part of 2.12 - Complete Q2
Evaluate Elastic X-Pack for data analysis and anomaly detection	2.23	OBE
Alarms and alerts for NOC	2.24	OBE
Meet with NOC for initial questions	2.24.1	Completed
Identify properties needed to alert on	2.24.2	Deemed unnecessary by

		IRNC NOC
Basic prediction	2.24.3	OBE
Hook into ticketing or email system	2.24.4	OBE
Feedback from NOC	2.24.5	OBE
Evaluation of PS tests and sampling	2.25	Year 5
Compare sampled and un-sampled flow data	2.27	Year 5
Compare active and passive measurement data	2.26	Year 5
Flow data dashboards with variable queries	2.28	Completed Q3
Dashboard for PS File Transfer data	2.29 (NEW)	Waiting on 1.15
Basic graph dashboard of file transfer data	2.29.1	Waiting on 1.15
Heatmap of PS file transfer data	2.29.2	Waiting on 1.15
Dashboard Tasks	**3**	
Dashboard management tasks	3.11	Ongoing
Navigation through views by question	3.11.1	Completed Y4Q4
Grafana - shib access for grafana, set all to read only	3.11.2	Completed Q3
Incorporate Shib extensions back to Grafana	3.11.3	Not needed (Q2)
Develop template for standard dashboard	3.11.4 (NEW)	Planned Year 5
Sensor health dashboard	3.11.5 (NEW)	Planned Year 5
IRNC Statistics dashboard	3.11.6 (NEW)	Planned Year 5
Documentation for full deployment by 3rd party	3.11.7 (NEW)	Planned Year 5
Map updates	3.12	Ongoing
Current map with PS data instead of SNMP data	3.12.1	Completed Y4Q4
Map for science registry data	3.12.2 (NEW)	Planned Year 5
Bugs and Fixes	3.15	Ongoing
Updates for bandwidth dashboard	3.22	Completed Y4Q4
Update SNMP Map on Bandwidth Dashboard	3.22.1	Completed Y4Q4
Add PIREN LA link to SNMP map	3.22.1.1	Completed Y4Q4
Add PIREN Guam Link to SNMP Map	3.22.1.2	Completed Y4Q4
Add exchange point info	3.22.1.8	Completed Y4Q4
work to get map added to mainline grafana widgets	3.22.1.9	Completed Q3
logarithmic scale	3.22.1.10	OBE
opacity based legends	3.22.1.11	Completed Q3
apply config changes without reload	3.22.1.12	Completed Q3
lines based on different functions	3.22.1.13	Completed Q3
invert legend	3.22.1.14	Completed Q3
additions to hover text box	3.22.1.15	Completed Q3
mapping improvements between datasource and text displayed	3.22.1.16	Completed Q3
dynamic wireframes, hide not present elements	3.22.1.17	Completed Q3
dynamically scale legend from dataset	3.22.1.18	OBE

wireframe editor in grafana instead of outside	3.22.1.19	OBE
get map added to grafana project	3.22.1.20	Completed Q3
investigate flows per country map (IU Communications style)	3.22.1.21	Completed Q3
Add TP Guam-HK link to map	3.22.1.22	Completed Q3
Check all A-Z and Z-A mappings	3.22.2	Completed Q3
Bottom Graph Updates	3.22.3	Completed Y4Q4
Make bottom graphs have same color for same network in each	3.22.3.1	Completed Y4Q4
Make bottom graph hover listing sort according to largest	3.22.3.2	Completed Y4Q4
Viz for Max sending vs retransmits	3.16	Dependent on research
Sankey Grafana Integration	3.20.	Completed Y4Q4
Add Sankey prototype (with mock data)	3.20.4	Completed Q1
refactor Sankey prototype to handle actual data	3.20.5	Completed Q1
integrate a specific query of flow data with Sankey	3.20.6	Completed Q3
generalize data processing for Sankey	3.20.7	Completed Q3
Review over the questions that guide Viz	3.21	Completed Q3
Which are still valid?	3.21.1	Completed Q2
Match questions with visualizations that we have	3.21.2	Completed Q1
Gather additional questions	3.21.3	Ongoing
Design additional dashboards in order to answer the questions.	3.21.4	Completed
Third party deployments	3.23 (NEW)	Ongoing
ANA Deployment	3.23.1	Ongoing
ANA SNMP bandwidth dashboard	3.23.1.1	Completed Y4Q3
ANA Flow deployment	3.23.1.2	Planned Year 5
Help with EPOC deployments	3.23.2	Ongoing
GNA	3.23.2.1	Completed Y4Q1
iLight	3.23.2.2	Planned Year 5
Asia Pacific Ring deployment	3.24. (NEW)	Planned Year 5
Project Coordination	4	
Project management and coordination	4.1	Ongoing
Weekly project meetings	4.1.1	Ongoing
Refresh NetSage website home page	4.1.2	Ongoing
REU funding for testers	4.1.3	Year 5
Coordinate with NOC	4.2	Ongoing
Year 4 reporting	4.8	Completed Y4Q4
44Q1 report	4.8.1	Completed
Y4Q2 report	4.8.2	Completed
Y4Q3 report	4.8.3	Completed Y4
Y4 annual report (with Q4)	4.8.4	Completed Y4Q4
Year 4 travel plans	4.15	Completed Y4Q4
CENIC March 2018	4.15.1	Q1

Internet2 PI meeting May 2018	4.15.2	Q2
Brian Tierney - NetSage for APAN March 2018	4.15.3	Q1
PEARC meeting - June 2018	4.15.4	Did not attend
Paper submission to PEARC	4.15.4.1	Completed - not accepted
SC '18	4.15.5	Q4
July 2018 AHM	4.15.6	Q2
October I2 Tech Ex	4.15.7	Q3
Jan 2019 AHM	4.15.8	Q4
Year 5 reporting	4.9	Ongoing
44Q1 report	4.9.1	Planned Year 5
Y4Q2 report	4.9.2	Planned Year 5
Y4Q3 report	4.9.3	Planned Year 5
Y4 annual report (with Q4)	4.9.4	Planned Year 5
Year 4 travel plans	4.16	Ongoing
Quilt Meeting - February	4.16.1	Planned Year 5
Internet2 Global Summit March	4.16.2	Planned Year 5
TNC June	4.16.3	Planned Year 5
AHM in July	4.16.4	Planned Year 5
PEARC meeting July	4.16.5	Planned Year 5
CC* meeting September	4.16.6	Planned Year 5
SC'19 november	4.16.6.1	Planned Year 5
Submit paper to SC'19	4.16.7	Planned Year 5
December I2 Tech Ex	4.16.8	Planned Year 5
Jan 2019 AHM	4.16.9	Planned Year 5

10. Financials

Table 2 shows the expenditures for Year 4 across the full team.

Table 3 shows a summary of expenditures Years 1-4, and our projected expenditures for Year 5. At this time, we are slightly underspent for several reasons which we are working to correct in Year 5. Each subaward is underspent at a slightly different level.

First, due to the lag in identifying a new PI for the LBNL portion of the project, there were several months in Year 4 where there were no charges by that institution. Lake has made it a priority as he is getting on board to hire and we expect additional staffing by that team early in Year 5, especially to support some of the dashboards related to perfSONAR and comparative network statistics.

In addition, the IU software development team underwent a restructuring, so the portion supporting NetSage was short staffed for much of Year 4. A new member of the team has already been hired (with a start date of July 2018 due to prior commitments), and additional positions are currently posted. We expect additional staffing in Year 5, especially to help with gathering feedback, the deployment of additional Tstat servers on archive, the development of additional dashboards for the retransmit data sets, and the productivity of the NetSage infrastructure to enable an easier hand off at the end of the project.

We plan to request a no-cost-extension, as currently we are projecting funding for the project will run approximately 6-8 months past the current end date of April 30, 2020. Separately from the main funding, due to significant cost savings in the running of the All Hands Meetings, our participant support budget is projected to be under spent by approximately $30,000. We will likely formally request a shift of budget from this spending category to salaries during Year 5.

Table 2: Expenditures for full project in Year 4.

Item	Univ	Feb '18	Mar '18	Apr '18	May '18	Jun '18	Jul '18	Aug '18	Sep '18	Oc '18	Nov '18	Dec '18	Jan '18	TOTAL
STAFF COSTS (INCLUDING BENEFITS, F&A)														
Schopf, Jennifer	IU	4,628	4,628	4,628	4,628	4,628	4,921	4,921	4,921	4,921	4,921	4,921	4,921	57,587
Lee, Andrew	IU									1,977	1,977	1,977	1,977	7,908
Moynihan, Ed	IU	1,543	1,543	1,543	1,543	1,543	1,549	1,549	1,549	1,549	1,549	1,549	1,549	18,558
Chevakier, Scott	IU						1,946	1,946	1,946	1,946	1,946	1,946	1,946	13,622
Balas, Ed	IU	1,370	1,370	1,370	1,370	1,370	1,469	1,469	1,469	1,469	1,469	1,469	1,469	17,133
IU Dev Team	IU	20,580	20,580	20,580	20,580	20,580	21,284	21,284	21,284	21,284	21,284	21,284	21,284	251,888
Hubbard, Heather	IU	1,923	1,923	1,923	1,923	1,923	1,923	1,923	1,923	1,923	1,923	1,923	1,923	23,076
Sean Peisert	UCD	921	3,349	5,262	0	4,957	3,868	0	0	0				18,357
Jon Dugan	LBNL	1,709	1,553	1,627	1,486	0	1,262	0	-111	0				7,526
Dwivedi, Dipankar	LBNL	3,948	5,742	6,016	4,806	4,423	4,926	11,341	-767	0				40,435

Giannakou, Anna	LBNL	9,328	6,784	8,884	8,112	6,097	2,890	0	-1,884	0				40,211
Kiran, Mariam	LBNL			1,277	0	0	-34	0	-12	0				1,231
Lake, Andrew	LBNL										0	5,660	6,460	12,120
Whinery, Alan	UH				22,427	22,427								44,854
Gonzalez, Alberto	UH	4,144	4,144	4,144	4,144	4,144	8,288	4,144	4,144	4,144	4,144	4,144	4,144	53,872
Kanal, Mahesh	UH							3,950	3,950	3,950	3,950	3,950	3,950	23,700
Seto-Mook, Tyson	UH	4,144	4,144	4,144	4,144			4,144	4,144	4,144	4,144	4,144	4,144	41,440
Leigh, Jason	UH					13,193	13,193							26,386
TOTAL STAFFING		**54,238**	**55,760**	**61,399**	**75,162**	**85,285**	**67,485**	**56,672**	**42,556**	**47,308**	**47,308**	**52,968**	**53,768**	699,909
TRAVEL, OTHER (INCLUDING F&A)														
Travel - Dwivedi - AHM Hawaii Jan 2018	LBNL	598												598
Travel - Schopf - I2 May 2018	IU	990				2,843								3,833
Travel - Schopf - NRP Aug 2018	IU			1,233	528	784								2,545
Meeting support (IU portion) HI feb 2018	IU	205												205
Meeting support (IU) GS May 2017 (yes 17)	IU				376									376
Travel - Leigh - AHM July 2018	UH						3,223							3,223
Travel - Gonzalez - AHM July 2018	UH						2,284							2,284
Travel - Schopf - AHM July 2018	IU						29	1,494						1,523
Travel - Balas - AHM July 2018	IU					2,124								2,124
Travel - Doyle - AHM July 2018	IU					2,124	391							2,515
Travel - Balas - I2 GS May 2018	IU					315								315
Travel - Schopf - I2 Tech Ex Oct 2018	IU							990						990
Travel - Schopf - SC'18 Nov 2018	IU									132		3,267		3,399
Dell warranty renewal	IU									2,794				2,794
Travel - Mahesh - SC'18 Nov 2018	UH										2,694			2,694
Travel - Andrew Lake - NetSage F2F Jan 2019 - Hawaii	LBNL												3,462	3,462
Travel - Jason Zurawski - NetSage F2F Jan 2019 - Hawaii	LBNL												2,263	2,263
TOTAL TRAVEL		**1,792**	**0**	**1,233**	**904**	**8,191**	**5,927**	**2,484**	**0**	**2,926**	**2,694**	**3,267**	**5,725**	35,143
EQUIP OVER $5K														
Dell Archival Storage	IU										68,978			68,978
Elastic SW	IU										49,950			49,950
TOTAL EQUIP		**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**118,928**	**0**	**0**	118,928
PARTICIPANT SUPPORT														
Meeting support HI Feb 2018	IU	311												311
Meeting support IRNC PI May 2017 (yes 17)	IU				2,495									2,495
Travel - Tierney GS May 2018	IU				275									275
Travel - Tierney AHM Berkeley JUly 2019	IU									350				350

Travel - Tierney - SC18 Nov 2018	IU									100				100
TOTAL PARTICIPANT SUPPORT		**311**	**0**	**0**	**2,770**	**0**	**0**	**0**	**0**	**450**	**0**	**0**	**0**	3,531
TOTAL EXPENDITURES		**56,342**	**55,760**	**62,632**	**78,836**	**93,475**	**73,412**	**59,156**	**42,556**	**50,684**	**168,930**	**56,235**	**59,493**	857,511

Table 3: Spending forecast for Year 5 and 6.

	Year 1	Year 2	Year 3	Year 4	Year 5 est	TOTAL		Year 6 estimates
Time span	May'15 - Jan'16	Feb'16-Jan'17	Feb'17-Jan'18	Feb'18-Jan'19	Feb'19-Apr'20	May'15-Apr'20		May'20-Sep'20
Months	9	12	12	12	15			6
Staff	295,681	644,631	567,155	699,905	1,550,000	3,757,372		656,167
Travel	40,227	85,110	85,629	35,143	107,000	353,109		45,297
Over 5K	25,723	43,810	65,987	118,928	250,000	504,448		105,833
IU Overhead on subs	16,000	8000				24,000		0
TOTAL SPENT	377,631	781,551	718,771	853,976	1,907,000	4,638,929		807,297
Budgeted	750,000	1,000,000	1,150,000	1,150,000	1,400,000	5,450,000		
Under	372,369	218,449	431,229	296,024	-507,000	811,071		3,774
Part Support ($48K)	2,549	1,357	2,763	3,551	5,000	15,220		