NetSage
Award #1540933
Year 6 Annual Report
1 February 2020 through 31 January 2021

PIs: Jennifer Schopf (IU), 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 6 included four releases of the Dashboard sets, adding the ability to run the Flow Ingest Pipeline in a Docker container by a third party (which is required by some security policies), a shift to using the CAIDA AS to Organization Mapping Data set, and updating the science registry to include over 450 entries. Over the year, more than 3,500 unique users in 103 countries visited the NetSage Dashboards.

1. NetSage Overview
NetSage is building and deploying advanced measurement services to 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.

The NetSage software consists of a set of open source tools that follow a basic monitoring tool architecture, as shown in Figure 1. NetSage TestPoints are a collection of software and hardware components that gather data from SNMP, perfSONAR, and flow devices. The data from the TestPoints is sent to the Data Ingest Pipeline, where additional tags are added, including information from the MaxMind GeoIP database and the NetSage Science Registry. Data collection is discussed in Sections 5.1-5.5.

The data is then stored in the NetSage Archive, a storage framework consisting of several different databases, including an Elasticsearch archive and a Time Series Data System (TSDS) archive, discussed in Section 5.6.

A variety of Dashboards, built on top of the open source Grafana analysis and visualization engine, access the data from the NetSage Archive to visualize the results of queries. The Dashboards deployed for the IRNC resources are available online at http://portal.netsage.global and are discussed in Sections 6 (Dashboards) and 7 (Components).

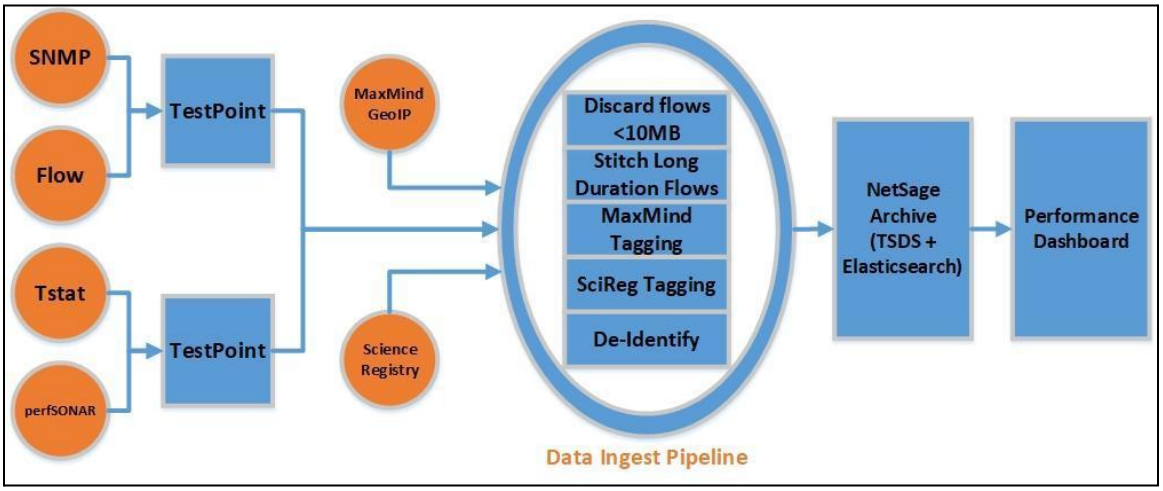

Figure 1: The current software architecture for NetSage.

Year 6 focused on hardening the code and adding requested functionality. As we enter the last year of the project, we will underscore documentation and ease-of-use aspects, contribute code back to the Grafana community as possible, and transition to a different support model for the software.

2. Staffing

At the end of Year 5, funded staff included:

- Jennifer Schopf, IU, PI - overall project director
- Scott Chevalier, IU, IRNC perfSONAR mesh support
- Dan Doyle, IU, developer - collection and reporting
- Lisa Ensman, IU, developer - Science Registry and data ingest pipeline
- Heather Hubbard, IU - Staff support
- Sangho Kim, IU, system engineer - collection and reporting
- Ed Moynihan, IU, Science Registry Data support
- Doug Southworth, analysis
- Andy Lake, LBNL, co-PI
- Sartaj Baveja, LBNL, developer
- Samir Faci, LBNL, developer
- Jason Leigh, UH Mānoa, co-PI - visualization oversight
- Mahesh Khanal, UH Mānoa, graduate research assistant - developer
- Katrina Turner, UH Mānoa, graduate research assistant - developer
- Alan Whinery, UH System – perfSONAR, PIREN coordination

At Indiana University, Aryan Deora joined the development team in June. He is working on an overhaul of the maps and Grafana updates. Chevalier was laid off in

December due to lack of continued funding for perfSONAR training. At LBNL, Sartaj Baveja was migrated off the project as part of internal project re-organization and Biju Jacob began ramping-up to help with Grafana plug-in development. At UH, Alan Whinery finished his portion of the project in early 2020 and graduate student Mahesh Khanal graduated in December 2020.

At the end of Year 6, funded staff included:

- Jennifer Schopf, IU, PI - overall project director
- Aryan Deora, IU, developer - mapping and visualization
- Dan Doyle, IU, developer - collection and reporting
- Lisa Ensman, IU, developer - Science Registry and data ingest pipeline
- Heather Hubbard, IU - Staff support
- Sangho Kim, IU, system engineer - collection and reporting
- Ed Moynihan, IU, Science Registry Data support
- Doug Southworth, analysis
- Andy Lake, LBNL, co-PI
- Biju Jacob, LBNL, developer
- Samir Faci, LBNL, developer
- Jason Leigh, UH Mānoa, co-PI - visualization oversight
- Katrina Turner, UH Mānoa, graduate research assistant - developer

We expect the LBNL and University of Hawaii teams to end their involvement in the main project in April 2021. Staff at IU will ramp down over the course of Year 7.

3. Collaborations, Travel, and Training

NetSage staff participated in various meetings to support ongoing deployment, collaboration, and training.

The following virtual events were attended:

- Schopf attended and presented at the Quilt Spring meeting in La Jolla, CA, on February 5-7, 2020 https://www.thequilt.net/public-event/2020-winter-member-meeting/. She jointly led a 4-hour session reviewing lessons learned from the EPOC Deep Dives and presented an in-depth live NetSage example using EPOC NetSage data.
- Doyle attended GrafanaCon 2020, May 13-29, 2020. The meeting was broken up in 1-2 hour segments each day. During this time, Doyle learned about the just released Grafana 7 and the changes the project would need to make to keep pace.
- Lake, Chevalier, and Southworth attended the perfSONAR Virtual Developer meeting June 15-18, 2020. Lake led a session where NetSage archiving strategies were discussed and how they may benefit the broader perfSONAR community.

- Schopf virtually attended PEARC 2020, July 27-31, 2020, https://pearc.acm.org/pearc20, to understand the measurement and monitoring needs and to evaluate how NetSage might help this community.
- Leigh attended initial meetings with CAIDA and UCSD in July 2020 to introduce existing visualization efforts.
- Moynihan attended the Virtual "Global Collaboration on Data beyond Disciplines" conference, September 23-25, 2020. He attended and participated in sessions on FAIR data principles, international COVID data sharing, and Japanese efforts to support Polar data sharing with an eye to how NetSage could help these efforts.
- Schopf attended the Quilt Fall Meeting, September 30-October 1, 2020, https://www.thequilt.net/public-event/2020-quilt-virtual-fall-member-meeting/. She gave an overview of EPOC and its NetSage deployments.
- Schopf attended the Internet2 TechExtra (virtual Tech Ex), on October 6-7, 2020, https://www.internet2.edu/news-events/events/techextra-2020. She learned about community events that may impact NetSage planning.
- Schopf presented at the Southern Crossroads (SoX) Fall Member meeting, October 8, 2020. She walked through the new NetSage deployment for SoX and got feedback from their members.
- Schopf attended the Fall Coalition for Academic Scientific Collaboration (CASC) Members Meeting, October 13-16, 2020, https://casc.org/event/casc-fall-2020-membership-meeting, to understand where NetSage might help this community.
- Lake, Schopf and Southworth attended Internet2's TechExtra perfSONAR Data, November 2, 2020, https://internet2.edu/events/event-one/techextra-perfsonar-day, to better understand future direction of the perfSONAR software and how ideas from NetSage are contributing to the evolution of that project. Southworth also moderated sessions regarding future perfSONAR development and training.
- Southworth attended the GÉANT Telemetry meeting, November 10, 2020, https://wiki.geant.org/display/PUB/Telemetry+and+Big+Data+Workshop. He presented on NetSage use cases and scalability to accommodate increasing scientific dataset sizes and bandwidth requirements.
- Katrina Turner, Mahesh Khanal, Dan Doyle, Andrew Lake, Jason Leigh, and Jennifer Schopf attended SC'20 (https://sc20.supercomputing.org/) and the INDIS workshop (https://scinet.supercomputing.org/community/indis/previous-editions/sc20-indis/program/) on November 11-19, 2020. Katrina presented at INDIS.
- Schopf attended and presented at the Trans-Pacific Research and Education (TPRE) Networking meeting, January 16, 2021. She gave a presentation featuring NetSage data for TransPAC4 and TransPAC5.
- Schopf and Tierney attended the NSF Workshop on Overcoming Measurement Barriers to Internet Research (WOMBIR 2021), January 11-12, 2021, https://www.caida.org/workshops/wombir/2101/. They had a white

paper accepted and took part in many discussions about what measurement data is needed for internet research.

A list of presentations and publications in Year 6 includes:

- Jennifer Schopf and Jason Zurawski, "Highly Interactive EPOC Deep Dive Outcomes and Researcher Engagement", Invited Workshop, Quilt Winter Meeting, La Jolla, CA, on February 5-7 2020.
- "Advanced flow analysis, new visualization headline latest NetSage release", Blog post for the 1.3 release, http://netsage.global February 25, 2020.
- "NetSage adds map of global science transfers", Blog post for 1.4 release, http://netsage.global, June 5, 2020.
- "NetSage update adds new Dashboards and visualizations", Blog post for 1.5 release, http://netsage.global, August 24, 2020.
- Jennifer Schopf, "Moving Data Faster with the Engagement and Performance Operations Center (EPOC)", Quilt Fall Meeting, September 30, 2020.
- Jennifer Schopf, "NetSage, EPOC, and SoX", Invited Presentation, Southern Crossroads (SoX) Fall Member Meeting, October 8, 2020.
- "Deep-dive into flows with new NetSage features", Blog post for 1.6 release, http://netsage.global, November 18, 2020.
- Turner, K., Khanal, M., Seto-Mook, T., Gonzalez, A., Leigh, J., Lake, A., Baveja, S. S., Faci, S., Tierney, B., Doyle, D., Ensman, L., Schopf, J. M., Southworth, D., Balas, E., "The NetSage Measurement Framework: Design, Development, and Discoveries", November 12, 2020, IEEE/ACM Innovating the Network for Data-Intensive Science (INDIS), DOI: 10.1109/INDIS51933.2020.00010. NSF-PAR ID: 10215022
- Jennifer Schopf, "TransPAC and NEAAR", invited presentation, Trans-Pacific Research and Education Networking meeting, January 16, 2021.
- Jennifer Schopf, Brian Tierney, Hans Addleman, Doug Southworth, "Routing Behaviors on R&E Networks", Invited White Paper, NSF Workshop on Overcoming Measurement Barriers to Internet Research (WOMBIR 2021), January 11-12, 2021, https://www.caida.org/workshops/wombir/2101/slides/wombir2021-paper31.pdf.

4. Project Coordination

4.1 Internal Coordination

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

Per NSF direction, in February we submitted a request for $750,000 of supplemental funding to continue the operation of NetSage through December, 2020.

In June we held a virtual All Hands Meeting to start to focus on how to wind-down the various development aspects. We walked through several in depth topics, including finding a substitute for MaxMind, how we wanted to approach data rollup and compression, general scalability and documentation aspects, and which components to focus on solidifying to be able to give them back to the greater Grafana community.

In October we held another virtual All Hands Meeting and Hackathon on finishing items for the 1.6.0 release, including a new full-details flow Dashboard and the ability for users to easily include or exclude known perfSONAR test traffic from the views. We continued conversations about wind-down on development efforts and timelines remaining for the different groups at UH, IU, and LBNL.

4.2 Coordination with IRNC Partners

Work with the IRNC-funded backbones continued, and we now have SNMP and perfSONAR data from all of the original circuits as well as their latter-year additions. Sampled flow data is being collected from NEAAR, TransPAC, Pacific Wave, PIREN (Hawai'i links), and AmPATH. StarLight and the PIREN Australia links will not share flow data.

We continued ongoing discussions and coordination with the IRNC NOC until their award ended, December 31, 2020.. The IU GlobalNOC is now also supporting NetSage as a managed service for deployments that were not part of NSF-supported partnerships.

The current status for the deployments for the IRNC projects is:

- Pacific Wave/CENIC became an EPOC partner, and their NetSage data is now being supported as part of the EPOC NetSage work.
- PIREN is still discussing options with the NetSage team.
- The AmLight-ExP IRNC project ended yesterday, March 31st, 2021. Because of this, we have suggested shutting down the data collection for both AmLight and AmPATH on April 15, 2021.
- StarLight will not continue to share data on a date of our choosing, so we will stop the collection of their data also on April 15.
- NEAAR will transition to NEA3R – and will continue to use NetSage via the IU GlobalNOC per the NEA3R proposal, exact timing depends on IRNC NetSage funding for Indiana.
- TransPAC4 will transition to TransPAC5 will continue to use NetSage via the IU GlobalNOC per the TP5 proposal, exact timing depends on IRNC NetSage funding for Indiana.

The IRNC-supported perfSONAR deployment will be ramped down starting April 15, 2021 as well.

4.3 Coordination with External Partners

Additional third party deployments for network data also took place this quarter. Deployments, mostly funded by other projects, include:

- SNMP data for the core Advanced North Atlantic consortium (ANA), available online at https://ana.netsage.global.
- SNMP data and flow data for the core Great Plains Network (GPN), available online at https://gpn.netsage.global.
- Flow data for the Indiana GigaPOP and I-Light networks, available online at https://ilight.netsage.global.
- Flow data for KINBER/PennRen, available online at https://pennren.netsage.global.
- Flow data for Front Range GigaPop, available online at https://frgp.netsage.global/
- Flow data for the Texas Advanced Computing Center (TACC) network, available online at https://tacc.netsage.global/
- Flow data for the Southern Crossroads (SoX), available online at https://sox.netsage.global/
- Flow data for the Sun Corridor Network (SCN), available online at https://suncorridor.netsage.global/. Note - this is not public yet, pending signoff from their technical committee likely in March.

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.

5.1 Collecting 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.

During Year 6, we finalized work with the AmLight engineers to update our collections and monitoring of their network changes. This resulted in increasing our SNMP collections on AmLight to 11 network devices representing 8 circuits, including a new trans-Atlantic link from Miami, Florida, to Cape Town, South Africa, by way of Angola. A new SNMP collection was established in mid-December 2020 for the new NEA3R circuit between New York and Amsterdam.

5.2 Collecting Flow Data from Routers (sFlow, NetFlow)

Flow data, collected from routers using NetFlow and sFlow, contains information about where network traffic is coming from and going to, as well as how much traffic

is being generated. This is vital, since with source and destination information, we can begin to ask questions about the sciences and organizations using the networks. Flow data processing relies on the Data Ingest Pipeline, shown in Figure 1, to filter data, add tags, and de-identify the flows.

5.2.A Data Ingest Pipeline Updates

During Year 6, there were 7 releases of the Data Ingest Pipeline in addition to several upgrades to supporting software Logstash. The upgrades included additional work to address scaling issues for larger deployments.

A major contribution for Year 6 was adding support to run the pipeline in a Docker container. This work was done to make third party deployments substantially more turn-key. The first of these distributions was with the Great Plains Network (GPN), whose security policy did not allow them to send flows outside of their network prior to being deidentified. We have since deployed this for several other networks including SoX, SANReN, and Sun Corridor.

Along with the Docker releases, the team developed improvements to the overall release process of the NetSage pipeline. This included more robust version tagging and change documentation as well as automation to ensure that every development check-in creates a new docker image for ease of testing and deployment. Based on user feedback configuration options were added to allow additional tuning of the data collection by end users to their particular environment, such as the number of collectors, memory usage, support for both NetFlow and sFlow, and options for data retention and filtering. In all releases, user facing documentation was expanded to include the additional features.

Functionality was also added to support data from the Shared Whois Project (SWIP) to identify institutions that do not have their own ASNs. The SWIP data is used to formally document cases where subsets of IP space need to be identified. For example, a regional network may allocate a piece of its own IP space to a member institution. SWIP enables the traffic to that IP-subset to be properly attributed to the institution and not the regional network, which in turn allows the NetSage Dashboards to list the sources and destinations more accurately.

In January 2020, we identified a need to replace the MaxMind GeoIP database that mapped IPs to organizations when their upgrade resulted in less readable naming. After extensive comparisons and tests, we selected moving to Center for Applied Internet Data Analysis (CAIDA) AS to Organization Mapping Data set. After extensive comparisons and testing, the migration was made in December 2020.

As we look forward, we are examining replacement options for the nfdump tool. While this tool has served us well for many years, it requires maintaining significant extra code and has had several bugs and limitations that we have had to work around.

Year 6 work on the Data Ingest Pipeline also included significant improvements to the development, testing, and release process. The team migrated from a VM-based to Docker-based development environment. This not only simplified bringing up local environments but also allowed for integration with continuous integration (CI) environments capable of bringing up Docker containers. A TravisCI (https://travisci.com) infrastructure was put in place to perform basic tests on the NetSage software when changes are submitted.

The team also put a focus on documentation throughout the year. A documentation infrastructure was put in place that uses Docusaurus (https://docusaurus.io). This provides a consistent way to generate developer documentation. This is important for maintaining the project as well as onboarding new developers.

5.2.B Flow Data Collection Updates

During Year 6, we started collecting five new flow datasets for the AmLight network. This new data corresponds to the increase in SNMP data collection for AmLight in Year 6 Quarter 1 and helps to provide insights into traffic across their new links, including the one to South Africa, instead of only at the exchange point in Miami. This brings the total number of sensors collecting flow data for AmLight to seven. We also incorporated filtering of the AmLight data so that only flows specific to the links were included, regardless of the actual flows exported to us.

Additionally, in December, the South African National Research and Education Network (SANReN), a partner of the NEAAR project, was added as a Docker installation and that flow data is now viewable in the Dashboards.

5.3 Collecting Tstat Data from Archives

We collect archive-based TCP flow statistics using Tstat, a tool that was developed as part of the EU Measurement Plane (mplane) FP7 project by Munafó and Mellia at Politecnico di Torino. Tstat examines all packet headers, similar to flow data collection on a router, only unsampled, and reports the source and destination IPs and ports, the number of bits and packets transferred, the duration of the flow, the flow type, protocol used, and TCP retransmits.

The current Tstat deployments include:

- TACC/LEARN: The TACC deployment remains active, though sometimes requires working with them to restart it based on changes in their environment.
- University of Hawai'i Astronomy: This work is running in a stable state.
- NCAR/FRGP: This work is running in a stable state.
- National Energy Research Scientific Computing Center (NERSC): This work continues to run in a stable state.

In Year 6, we worked on containerized deployments for Tstat using Docker in order to better support third-party deployments. Since we had seen success with the Ingest Pipeline Docker container, it was thought that making a similarly easy to install and run version of Tstat would be of benefit to the project. This may also open up more opportunities for data sharing with institutions that do not or cannot allow us the direct management existing Tstat deployments require. Work was completed but there are no current deployments using this approach.

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 throughput, latency, and loss, and archives the data in the NetSage archive using TSDS. The IRNC projects participate in the IRNC perfSONAR mesh, available at http://data.ctc.transpac.org/maddash-webui/index.cgi?dashboard=IRNC%20Mesh.

Members of NetSage team were active participants in the perfSONAR project through direct contributions, attending developer meetings and assisting in beta testing. There were five perfSONAR releases throughout the year, each of which NetSage upgraded to as part of its automatic update process. The largest of these was perfSONAR 4.3.0 released on November 2, 2020 with the most relevant change to NetSage being that it modernized the perfSONAR components to Python 3.

NetSage staff will end their support of the IRNC perfSONAR installation on April 1, 2021. The nodes will be used by the individual projects going forward.

5.5 Science Registry

The Science Registry is a system developed by NetSage to document known network endpoints, organizations, and science projects that are users of the resources. The system supports collaborative and crowd sourced data entry and is a key component for presenting information including the science domain, project name, university or institution, geo-location, and other related data. Science Registry data is generally collected from resource owners who identify the science project that is using specific address space.

Year 6 focused on extending the number of Science Registry entries and cleaning up the data fields for more consistent use. Several fields were dropped, and existing ones were clarified. Previously recorded data was cleaned up with these new changes as well. There are now over 450 entries in the Science Registry. After each release, we reindexed the older data to account for changes in field names, types, or values. This ensured historical data maintained consistency with newer data.

5.6 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 Year 6, there were two releases of the core TSDS code. These releases were focused on the efficiency of index usage and automatically reconnecting services in the event of an outage, as well as several bug fixes. There was also one release of the TSDS Grafana driver code to better support UTC times, along with addressing documented security issues in the dependencies.

In Year 7, we expect to convert the TSDS Grafana driver code from Angular into React in order to keep pace with the Grafana environment.

6. Visualization And Analysis Dashboards

6.1 Development Process and Environment

The project follows a development process to provide developers with consistent environments, track changes, and verify the software is working as expected. That process is constantly being refined to increase developer efficiency and help ensure high-quality software is released.

The project implemented templating features to allow easy adaptation of Dashboards to different deployments. Customizations can now be expressed in a per deployment configuration file instead of relying on a person to tune them all manually after install. This helps avoid human error in individual deployments and provides a foundation for scaling up to more deployments in the future.

6.2 Dashboard Releases

There were four major releases of the Dashboards in Year 6, one in each quarter. Across all the releases, there was the addition of three brand new Grafana visualization plug-ins, four completely new Dashboards, and updates to every existing Dashboard. Highlights from each release included:

- Version 1.3.0 in February 2020 introduced the Advanced Flow Analysis Dashboard and Slope Graph visualization plug-in.
- Version 1.4.0 in June 2020 saw the addition of a new map plug-in developed for the Flows by Science Discipline Dashboard as well as an upgrade to Grafana 7 which led to enhancements in the overall look and feel of all the Dashboards.
- Version 1.5.0 in September 2020 added the Flow Data by Projects Dashboard and Top Talkers Over Time Dashboard, the latter of which included the newly developed Bump Chart visualization plug-in.

- Version 1.6.0 in November 2020 added a new Individual Flow Information Dashboard, allowed filtering of network performance traffic from flow data, and added a map to the Flow by Projects Dashboard.

6.3 Currently Supported IRNC NetSage Dashboard Sets

NetSage Dashboards are designed to answer specific types of questions about the state of a resource. Each Dashboard, divided here into seven related Sets, is made up of a suite of complementary components that are detailed in Section 7.

The current Sets of NetSage Dashboards deployed for the IRNC projects include:

- **Set 1: Bandwidth Dashboard:** How heavily used are the circuits and exchange points?
    - http://portal.netsage.global

  Includes:
    - A *map* showing monitored circuits and exchange points, with information about the current use of the links from SNMP, as well as indications if the relevant pieces of the link are experiencing loss.
    - A *line chart* for each IRNC circuit showing capacity and usage information from SNMP throughput data.
    - A *line chart* showing the use of all of the circuits by average and maximum use, in both directions.
- **Set 2: Flow Data Dashboards (by circuit, archive, organization, or country):** What are the top ten senders/receivers of flows (ranked by volume or rate) by source and destination?
    - Top Sources/Destinations for all Flow data: https://portal.netsage.global/grafana/d/xk26IFhmk/flow-data-for-circuits
    - Top Flows by Archive: https://portal.netsage.global/grafana/d/mNPduO8mz/flow-data-for-data-archives
    - Top Flows by Organization: https://portal.netsage.global/grafana/d/QfzDJKhik/flow-data-per-organization
    - Top Flows by Country: https://portal.netsage.global/grafana/d/fgrOzz_mk/flow-data-per-country
    - Top Flows for Projects (using the NetSage Science Registry): https://portal.netsage.global/grafana/d/ie7TeomGz/flow-data-for-projects
    - Individual Flow Information: eg https://portal.netsage.global/grafana/d/nzuMyBcGk/flow-information?orgId=2&var-flow=DlbpXHgBuSt-Bick6GQq&from=1615918029031&to=1616522829033

  Includes:
    - A *table* and *bar charts* showing Top Talker data by volume and rate.

    - A *table* of Top Pairs of Talkers, if appropriate.
    - A *Slope Graph* for Top Pairs, if appropriate.
    - *Summary statistics* about ports, protocols, and sensor use.
    - A *Sankey graph* of Top Talkers, for some use cases.
    - A *Map for flows*, for some use cases
- **Set 3: Individual Flows Dashboards:** For a particular organization or country, what is the per-flow data?
    - Individual Flows by Organization: https://portal.netsage.global/grafana/d/-l3_u8nWk/individual-flows?orgId=2
    - Individual Flows by Country: https://portal.netsage.global/grafana/d/80IVUboZk/individual-flows-per-country

  Includes:
    - A *table* listing data specific to individual flows, including subnet for source and destination.
    - A *Heatmap* showing data transfers over time, by volume and rate.
    - *Summary statistics* about ports, protocols, and sensor use.
- **Set 4: Science Discipline Dashboards:**
    - What are the flows for a defined Science Discipline (using the NetSage Science Registry)? https://portal.netsage.global/grafana/d/WNn1qyaiz/flows-by-science-discipline

  Includes:
    - *Summary statistics* about how many flows were matched in the Science Registry.
    - A *Map* showing the flows between sources and destinations.
    - A *Sankey graph* of Top Talkers.
    - *Summary statistics* by discipline.
- **Set 5: Pattern Dashboards (for Bandwidth, Loss, Latency, Science Discipline)**: What are the recurring patterns of behaviors?
    - Bandwidth data from SNMP: https://portal.netsage.global/grafana/d/000000004/bandwidth-patterns
    - Loss data from perfSONAR: https://portal.netsage.global/grafana/d/000000006/loss-patterns
    - Latency data from perfSONAR: https://portal.netsage.global/grafana/d/000000005/latency-patterns
    - Science Discipline data from Flow data and Science Registry data: https://portal.netsage.global/grafana/d/ufIS9W7Zk/science-discipline-patterns

  Includes:
    - *Heatmaps* by relevant item (circuit, exchange point, discipline) showing data use over time.

- **Set 6: Flow Statistics Dashboard:** What are the current flow data statistics?
    - https://portal.netsage.global/grafana/d/CJC1FFhmz/other-flow-stats?orgId=2

  Includes:
    - *Summary statistics* for number of flows, source and destination information, and Science Registry use.
    - *Summary statistics* about ports, protocols, and sensor use
- **Set 7: Flow Analysis Dashboard:** What is this performance anomaly?
    - https://portal.netsage.global/grafana/d/VuuXrnPWz/flow-analysis?orgId=2

  Includes:
    - A *line graph* of SNMP data.
    - A *table* of Top Talker pairs.
    - A *table* and *bar chart* of top sources and destinations, by volume and by rate.
    - *Summary statistics* for the sensor.
    - A user selected timeframe on the SNMP line chart adjusts the timeframe for the corresponding Flow data graphs, including Top Pairs, Top Sources/Destinations, and Individual Flows information, to show additional detail to help determine what data transfers may be related to the performance change. Further query refinement is possible using supplemental filters to remove a specific value or view only that specific value.
- **Set 8: Statistics over Time:** How have the Top Talkers changed over time?
    - https://portal.netsage.global/grafana/d/b35BWxAZz/top-talkers-over-time

  Includes
    - A *Bump Chart* of Top Talkers.

In addition to these Dashboards, we support separate Dashboards to visualize the flow data for the Pacific Wave Exchange Point, specifically Sets 2, 3, 4, 6, and 8. This data is not currently included with the rest of the IRNC Flow Data Dashboards since it includes a large amount of domestic data that overwhelms the international data. These Dashboards are available, separate from the standard IRNC Dashboards, at https://pacwave.netsage.global/.

7. Dashboard Components

7.1 Navigator and Look-and-Feel

The Navigator is located at the top left of each NetSage Dashboard and looks like a spinning S. It enables users to easily shift between Dashboards by selecting the corresponding question that each Dashboard was developed to answer.

During Year 6, the Navigator code and documentation went through significant cleanup and hardening. Navigation items were added for the new Dashboards. The Navigator is now considered stable and the only expected changes are links to any additional Dashboards added to the project.

7.2 Flow Data Tables using sFlow, NetFlow, Tstat

Tables are used throughout the Flow Dashboards to display common statistics about the volume and rate of traffic of one or more flows. Additional statistics about retransmits and RTT are also displayed if the flow information is from Tstat. Tables are used in Sets 2, 3, and 7.

During Year 6, many of the tables were updated to have inline bar graphs to more easily highlight scale between elements. Additionally, there were several minor adjustments made to improve consistency of labels used across Dashboards. This component is in a stable state and we do not expect to make any changes here moving forward outside of fixing bugs.

7.3 Heatmaps

We use Heatmap visualizations to show changes in values over time and to easily identify patterns of behavior. Heatmaps are used in Sets 3, 4, and 5.

Year 6 accomplishments involved enabling the Heatmap legend to alternate between light and dark display modes, in accordance with modern trends in user-interface designs. In addition, coloration changes were made to more clearly reflect when data was not available versus when it was available and zero. Final work for the component will include contributing all the Heatmap improvements back to Grafana.

7.4 Sankey Graphs

Sankey Graphs show relationships between items using a ribbon graphic, where the width shows the quantity proportionately. We use Sankey graphs to show data flows over the IRNC resources in Dashboard Sets 2 and 4.

In Year 6, the Sankey plug-in was completed along with its accompanying documentation. The query for the Sankey graph on the Flows by Science Discipline Dashboard was updated to use the “preferred organization” field instead of the science registry organization, hence filling in the formerly “unknown” category associated with some flows. This resulted in the Sankey graphs ballooning into an over complicated visualization. The decision was made to favor showing fewer flows with greater accuracy. We plan to contribute the back to the Grafana community.

7.5 Slope Graph

A Slope Graph is a visual way of showing relationships between two lists. We use Slope Graphs in Dashboard Set 2.

Slope Graphs were included in Dashboards for the first time as part of NetSage 1.3 in February. We began the process of contributing the plug-in back to the Grafana community by converting the code base to React and modifying the code to meet Grafana coding and documentation conventions. This work is expected to be completed by the end of April 2021.

7.6 Maps for Arbitrary Locations

A new Map plug-in to answer the query "Where are science disciplines sending their traffic to and at what volume?" was included in the 1.4 release. The map uses data tagged by the Science Registry. Because this query requires arcs between arbitrary map locations to be drawn automatically, we could not use the same mapping tool as the one used to produce the main bandwidth map, which requires the specification of the end points to draw the arcs.

In Year 6, we adapted the mapping for instances where flows may have a known source but unknown destination (as determined by the Science Registry). Performance improvements were made to enable more complicated maps to be depicted. A legend was added to reflect both the source and destination information. As this map is specific to the needs of the NetSage project, we chose not to contribute this code back to the Grafana user community. No further work is planned for this component.

7.7 Bump Chart

A Bump Chart can answer queries such as "How has the use of a resource changed over time?" This type of chart provides a longitudinal view that can be used, for example, to show which science disciplines remain as the top users of the network over time.

Bump Charts are now part of the Top Talkers Over Time Dashboard, included in the 1.5 release. The only remaining work is to complete the port of the chart to React and provide the capability back to the Grafana community. This is anticipated for end of April 2021.

7.8 Heatmap Variant

A new plug-in was started as a variant of the current Heatmap plug-in intended to show bandwidth over time by countries. Grafana's built-in heatmap does not allow for the creation of such a chart therefore necessitating the development of this new plug-in.

In Year 6, a prototype was successfully completed using live network data. However no further progress was made due to the lack of staffing time and the prioritization of completing other development efforts.

8. Examining The Effects Of TCP Retransmissions In perfSONAR Test Results

In order to address the question: "How do TCP retransmissions correlate to degraded performance?", a large sample of perfSONAR throughput test results were collected, analyzed, and compared with a series of data transfer tests. It was observed that retransmissions can be linked to a cause of degraded performance, or as a side effect of near-limit performance, or unrelated to performance. Across a spectrum of throughput performance levels, retransmissions appear in distinctly different contexts, and we posited that it may be possible to examine different patterns of retransmissions on a transfer that can provide diagnostic insight when troubleshooting networks.

Several networking trends, such as the incipient change from traditional congestion control algorithms (CUBIC, HTCP), to innovative algorithms like BBR (https://cloud.google.com/blog/products/gcp/tcp-bbr-congestion-control-comes-to-gcp-your-internet-just-got-faster), and the upgrading of a majority of test and transfer nodes from 10 G to 40G and 100G interfaces was hypothesized to affect the study of TCP retransmissions at large.

Our study concluded that there did not appear to be any meaningful correlation between TCP retransmissions and reduced throughput, largely because the congestion avoidance algorithms in use (CUBIC, H-TCP) are not as likely to reduce rate as previous algorithms (Reno, Tahoe, Vegas). The root cause of retransmissions, as well as the vulnerability of an application's network throughput to loss or retransmission scenarios will be determined most often by configuration elements and the choice of data transfer application at the source and destination hosts.

9. Use of NetSage

The NetSage project collects data about people accessing and using the Dashboards using Google Analytics. Between February 1, 2020, and January 31, 2021, over 3,500 unique users in 103 countries visited the NetSage Dashboards, as shown in Figure 2.

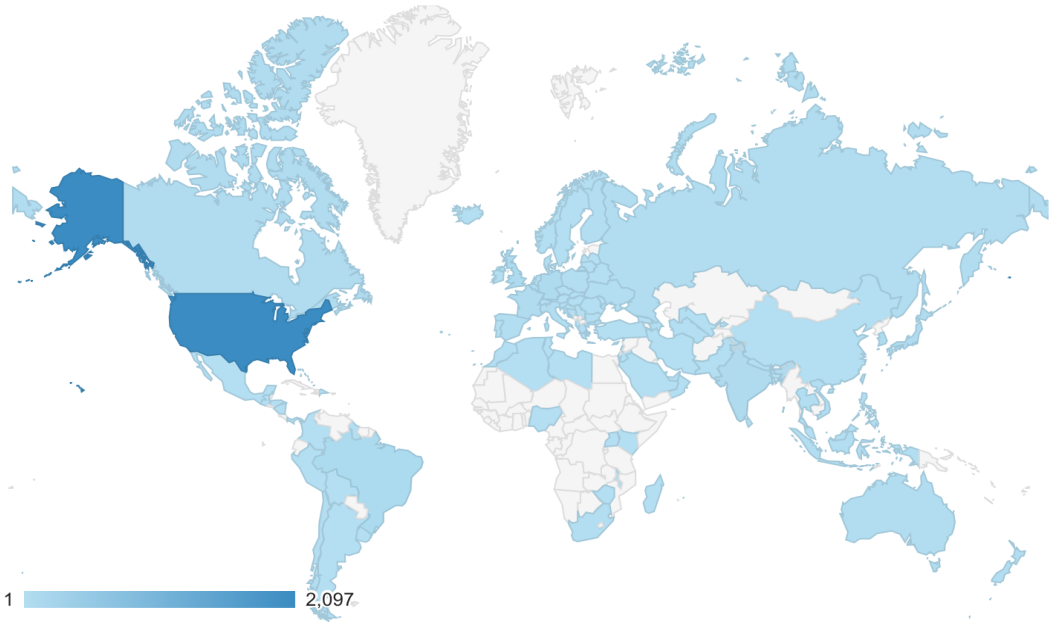

Figure 2: Map of the countries that accessed NetSage Dashboards between February 1, 2020, and January 31, 2021. Darker blue indicates higher usage within that country for individual users.

The United States led the way accounting for 59% of the Dashboard traffic, followed by Singapore (3.4%), Brazil (3.1%), and the United Kingdom (2.5%). The trend of the average user spending multiple minutes on the site held steady throughout the year and the Flow Dashboards continued to be the most popular Dashboards.

10. Data Privacy and Security

Basic security measures are being maintained and there were no security incidents to report. As a reminder, NetSage does not collect PII and therefore are compliant with the European General Data Protection Regulation (GDPR). No data privacy or security changes were made during the year.

11. Updated WBS for Year 6

Item	Y5 and Y6 WBS	Notes
Data Collection	1	
PerfSONAR Related Tasks	1.4	Ongoing
Define and deploy PS test mesh for backbones	1.4.2	Completed
Add AmLight resources directly to NetSage archive	1.4.2.12	Completed
Ongoing support for IRNC PS mesh	1.4.3	Ongoing
Update PS Ingest	1.4.4	OBE
SNMP related tasks	1.5	Ongoing
SNMP data from Backbones	1.5.2	Ongoing
Tstat/Flow deployment	1.7	Ongoing
Dashboard for exchange point flow data	1.7.14.6	Completed
Lower flow data collection threshold from 500M	1.7.19	Completed
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	Completed
Deploy Tstat on Hawaiian astronomy archives	1.8.3	Completed
Deploy Tstat on CENIC/PRP archives	1.8.4	OBE
Use top talkers list to identify likely DTNs	1.8.5	OBE
Instrument NCAR Archive	1.8.6	Completed
NASA DTN instrumentation	1.8.7	OBE
Other possible DTNs from IRNC partners	1.8.8	Completed
NOAA Data archives	1.8.8.2	Completed
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
Data transfer information (ie fiona) as additional data source	1.15	Completed
Find guinea pigs for data transfer inclusion	1.15.3	OBE
Hand off data sets to IRNC owners	1.16	Planned Y7
Dockerize Ingest Pipeline	1.17	Completed Y6Q1
Dockerize Tstat	1.18	Completed Y6
Additional tagging	1.19	Completed Y6
Investigate Ingest Scalability	1.20	Ongoing
Adapt ingest for additional scalability	1.20.1	Ongoing
Data Rollups	1.21	Planned Y6
Analysis	**2**	
Data cleaning	2.2	Completed

Recreate AS to Science Project Database (Science Registry)	2.4	Ongoing
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	OBE
Get PIREN to add data to science registry	2.4.3.4	Completed
Get CENIC to add data to science registry	2.4.3.5	OBE
Extensions to basic science registry framework	2.4.4	Ongoing
More science disciplines and ability to edit list	2.4.4.6	Completed
More roles and ability to edit list	2.4.4.7	Completed
Notes field for SR	2.4.4.8	OBE
URL field for SR	2.4.4.9	OBE
Admin section functionality for SR	2.4.4.10	Ongoing
Stitching with Elk Stack	2.4.4.12	Completed Y6Q1
Rework SR Data model	2.4.4.13	Completed Y6Q2
Transition plan for MaxMind	2.4.4.14	Completed Y6Q2
Dashboards for Tstat data from archive showing retransmit	2.7.1	Completed
Updates to Tstat archive dashboard	2.7.1.2	Ongoing
Heatmap for Tstat archive data	2.7.2	Completed
Americas Greatest Networks - most reliable	2.13.5	OBE
Updates to top X based on flow	2.13.9	Ongoing
Filter out Australian university names from flow data	2.13.9.1	Completed
Analysis of buffer size issues	2.16	Completed
PIREN analysis for astronomy data	2.20.	Completed
Dashboard Tasks	**3**	
Dashboard management tasks	3.11	Ongoing
Develop template for standard dashboard	3.11.4	Completed Y6Q1
Sensor health dashboard	3.11.5	Completed
IRNC Statistics dashboard	3.11.6	Completed
Discussions for running as managed service	3.11.7	Completed
Updated for Grafana 7.0 (May 2020)	3.11.8	Completed Y6Q2
Map updates	3.12	Ongoing
Map for science registry data	3.12.2	Released Y6Q2
Bugs and Fixes	3.15	Ongoing
Dashboard for PS File Transfer data	3.25	OBE
Basic graph dashboard of file transfer data	3.25.1	OBE
Heatmap of PS file transfer data	3.25.2	OBE
Viz for Max sending vs retransmits	3.16	OBE

Sankey Next Steps	3.20.8	Submitted, waiting on Grafana
Analysis Dash	3.26	Completed Y6Q1
Slope Graph	3.27	Completed Y6Q1
Dash for Volume >X, rate <y	3.28	Expected Y6
Navigator Plug-in	3.29	Ongoing
Add additional dash to Navigator	3.29.1	Ongoing
Submit Navigator to Grafana after cleanup	3.29.2	Submitted, waiting on Grafana
Dash - Top Talkers by Year (Bump Chart)	3.30	Released Y6Q3
Dash - Science Registry Projects	3.31	Released Y6Q3
Dash - Country vs Time	3.32	OBE
Dash - Show all data for a flow	NEW - 3.33	Completed Y6Q4
Review questions that guide Viz	3.21	Completed
Which are still valid?	3.21.1	Completed
Gather additional questions	3.21.3	Completed
Design additional dashboards in order to answer the questions.	3.21.4	Completed
Third party deployments	3.23	Ongoing
ANA Deployment	3.23.1	Completed
Add new link to ANA SNMP dashboard	3.23.1.3	Completed
ANA Flow deployment	3.23.1.2	OBE
Help with EPOC deployments	3.23.2	Ongoing
iLight- Flow	3.23.2.2	Completed
LEARN	3.23.2.3	Expected Y7
KINBER	3.23.2.4	Completed
FRGP	3.23.2.5 (NEW)	Completed
GPN - Flow	3.23.2.6 (NEW)	Completed Y6Q1
OARnet	3.23.2.7 (NEW)	Expected Y7
Asia Pacific Ring deployment	3.24.	OBE
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	OBE
Coordinate with NOC	4.2	Ongoing
Year 6 reporting	4.17	Ongoing
Y6Q1 report	4.17.1	Completed
Y6Q2 report	4.17.2	Completed

Y6Q3 report	4.17.3	Completed
Y6Q4 report	4.17.4	Completed
Y6 Annual report	4.17.5	Completed
Year 7 Reporting	4.17.6	Ongoing
Year 6 travel plans	4.18	OBE
AHM in June	4.18.1	OBE - Virtual
All Year 7 Travel	4.18.2	TBD - COVID
Google Analytics	4.19	Ongoing

12. Financials

Table 2 shows the expenditures for Year 6 across the full team. No funding was spent in foreign countries.

Item	Univ	Feb-20	Mar-20	Apr-20	May-20	Jun-20	Jul-20	Aug-20	20-Sep	Oct-20	Nov-20	Dec-20	Jan-21	TOTAL
STAFF COSTS (INCLUDING BENEFITS, F&A)														
Jennifer Schopf	IU	5,039	2,520	2,520	2,519	8,818	5,032	5,032	5,032	5,032	5,032	2,517	2,516	51,609
Ed Moynihan	IU	2,380	793	793	793	2,380	2,376	2,376	2,376	2,376	0	0	0	16,643
Scott Chevalier	IU	498	498	498	498	3,489	3,489	3,489	3,489	3,489	3,482	0	0	22,919
Doug Southworth	IU	2,722	2,722	2,722	2,722	2,722	2,722	2,722	2,722	2,722	2,718	2,718	2,718	32,652
Dan Doyle	IU	1,941	1,941	1,941	1,941	1,941	1,941	1,941	1,941	1,941	1,938	1,938	1,938	23,283
IU Dev Team	IU	21,132	21,132	21,291	21,373	21,805	23,117	23,117	23,117	23,117	23,561	23,757	23,561	270,080
Heather Hubbard	IU	1,318	1,318	1,371	1,977	1,319	1,316	1,316	1,316	2,334	960	1,154	987	16,686
Andrew Lake	LBNL	13,032	15,483	16,172	11,887	15,103	14,742	9,849	8,159	13,212	11,323	9,161	11,234	149,357
Sartaj Baveja	LBNL	12,675	14,664	0	0	14,806	14,458	0	11,600	12,531	10,891	9,198	0	100,823
Samir Faci	LBNL	29,118	21,066	20,489	17,089	19,420	19,477	19,166	16,261	26,797	21,715	20,466	0	231,064
Biju Jacob	LBNL										0	0	8,978	8,978
Jason Leigh	UH					16,986	16,986							33,972
Mahesh Kanal	UH	3,414	3,414	3,414	3,414	3,414	3,414	3,414	3,414	3,414	3,414	3,414	0	37,554
Katrina Turner	UH	4,450	4,450	4,450	4,450	4,450	4,450	4,450	4,450	4,450	4,450	4,450	4,450	53,400
TOTAL STAFFING		**97,719**	**90,001**	**75,661**	**68,663**	**116,653**	**113,520**	**76,872**	**83,877**	**101,415**	**89,484**	**78,773**	**56,382**	1,049,020
Travel and Other Costs (including overhead)														
Travel - Southworth - AHM J	IU	5,460												5,460
Travel - Ensman - AHM Jan'	IU	2,412												2,412
Travel - Doyle - AHM Jan'20	IU	4,425												4,425
Travel - Schopf - TNC	IU				73									73
AHM Dinner - IU portion Jan			556											556
Elastic SW	IU										65,950			65,950
Conf - Doyle INDIS@SC'20												66		66
Conf - Schopf SC'20												396		396
Domain name renewal												264		264
Conf - Leigh INDIS@SC'20												71		71
Conf - Turner INDIS@SC20												28		28
TOTAL TRAVEL		**12,297**	**556**	**0**	**73**	**0**	**0**	**0**	**0**	**0**	**65,950**	**825**	**0**	79,701
PARTICIPANT SUPPORT														
Hospitality - AHM dinner			1,137											1,137
TOTAL PARTICIPANT SUPPORT		**0**	**1,137**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1,137
TOTAL EXPENDITURES		**110,016**	**91,694**	**75,661**	**68,736**	**116,653**	**113,520**	**76,872**	**83,877**	**101,415**	**155,434**	**79,598**	**56,382**	1,129,858

According to our records, the IRNC NetSage budget is projected to be approximately $268,000 underspent at the current end date of May 1, 2021. This is the amount that Indiana University, the primary awardee, is underspent. Both subawardees are expecting to have their full budgets spent by the current May end date. IU's financial underspend is primarily due to delays in travel and meeting spending over the last 12 months due to COVID-19 related restrictions. Another contributing factor has been the IU hiring freeze over the last 12 months, also related to COVID-19 university regulations, so staff were not replaced from the prior year. Because of these issues, our planned interviews with current users did not take place.

We have requested and received a no cost extension through April 2022. Over the next year, we plan to use the funds for additional virtual outreach to potential and current users, to make up for the in-person interviews planned for last year that didn't take place, and hardening the software for handoff.