The NetSage Measurement Framework: Design, Development, and Discoveries

Katrina Turner
University of Hawai'i at Mānoa
khj@hawaii.edu

Mahesh Khanal
University of Hawai'i at Mānoa
mkhanal@hawaii.edu

Tyson Seto-Mook
University of Hawai'i at Mānoa
tmook@hawaii.edu

Alberto Gonzalez
University of Hawai'i at Mānoa
agon@hawaii.edu

Jason Leigh
University of Hawai'i at Mānoa
leighj@hawaii.edu

Andrew Lake
Lawrence Berkeley National Laboratory
andy@es.net

Sartaj Singh Baveha
Lawrence Berkeley National Laboratory
ssbaveja@es.net

Samir Faci
Lawrence Berkeley National Laboratory
samir@es.net

Brian Tierney
Lawrence Berkeley National Laboratory
bltierney@es.net

Daniel Doyle
Indiana University
daldoyle@globalnoc.iu.edu

Lisa Ensman
Indiana University
lensman@globalnoc.iu.edu

Jennifer M. Schopf
Indiana University
jmschopf@iu.edu

Douglas Southworth
Indiana University
dojosout@iu.edu

Edward Balas
Lawrence Berkeley National Laboratory
ebalas@es.net

***Abstract*— Data sharing to support research collaborations has increased exponentially in the last ten years, but effective data transfer performance continues to be hard to achieve. The NetSage Measurement and Analysis framework was developed to support understanding research data movement by collecting a broad set of monitoring data from various resources, and visualizing that data using performance Dashboards which are specifically designed to address the analysis needs of stakeholders. This paper describes the design methodology, the resulting architecture, development, and deployment approach, and a set of discoveries that NetSage Dashboards made possible.**

Keywords— Network Measurement and Monitoring, Visualization, Analytics, R&E Networks

I. Introduction

Scientific investigation is highly collaborative and requires the ability to seamlessly share data between institutions to meet the goals of the research. However, effective data sharing, especially for large data sets, can be challenging. For example, a common astronomy workflow involves a telescope producing very large data sets that are then analyzed at multiple international sites, which must complete before the next data collection window in order to refocus the telescope. Delays in data transfers can lead to researchers shipping disks instead of using the network for data delivery. Or in another case, it took over three months to transfer data from a set of climate science experiments for a centralized analysis [1].

The ability to measure and interpret network behavior is critical to understanding data transfer performance and ensuring stakeholders are getting the expected throughput. Information about the end-to-end data path makes it possible to identify problems with resources or potential delays to data transfers.

This paper details our design and development approach for NetSage, an open source measurement framework used to understand data transfer performance. We describe our stakeholder-focused methodology to design performance Dashboards to respond to specific questions. The software architecture and implementation were constructed to use multiple data sources and to take advantage of related approaches. We then walk through several use cases to show the types of analyses and discoveries that NetSage enables.

II. NetSage Overview

NetSage [2] is a unified, open, privacy-aware measurement, analysis, and visualization service designed to address the needs of today's research and education (R&E) data sharing collaborations. NetSage is unified in that it combines data from a variety of sources into a single unified view. NetSage is open in that the data collected are meant to be widely accessible, with performance Dashboards open to the public at http://portal.netsage.global, and also that the software developed by the team is open source. NetSage is privacy-aware, meaning that it contains no personally identifiable information (PII) about individual hosts or users of the network. Also, if required by the data provider, the Dashboards can be secured by password or Shibboleth [3] as needed, although the Dashboards described in this paper are all open to the public.

The innovative aspect of NetSage is not in the individual pieces but rather in the integration of data sources to support objective performance observations as a whole. NetSage deployments can collect data from routers, switches, active testing sites, and science data archives, which are common for collaborative research. NetSage uses a combination of passive and active measurements to provide longitudinal performance visualizations via performance Dashboards. The Dashboards can be used to identify changes of behaviors over monitored resources, new patterns for data transfers, or unexpected data

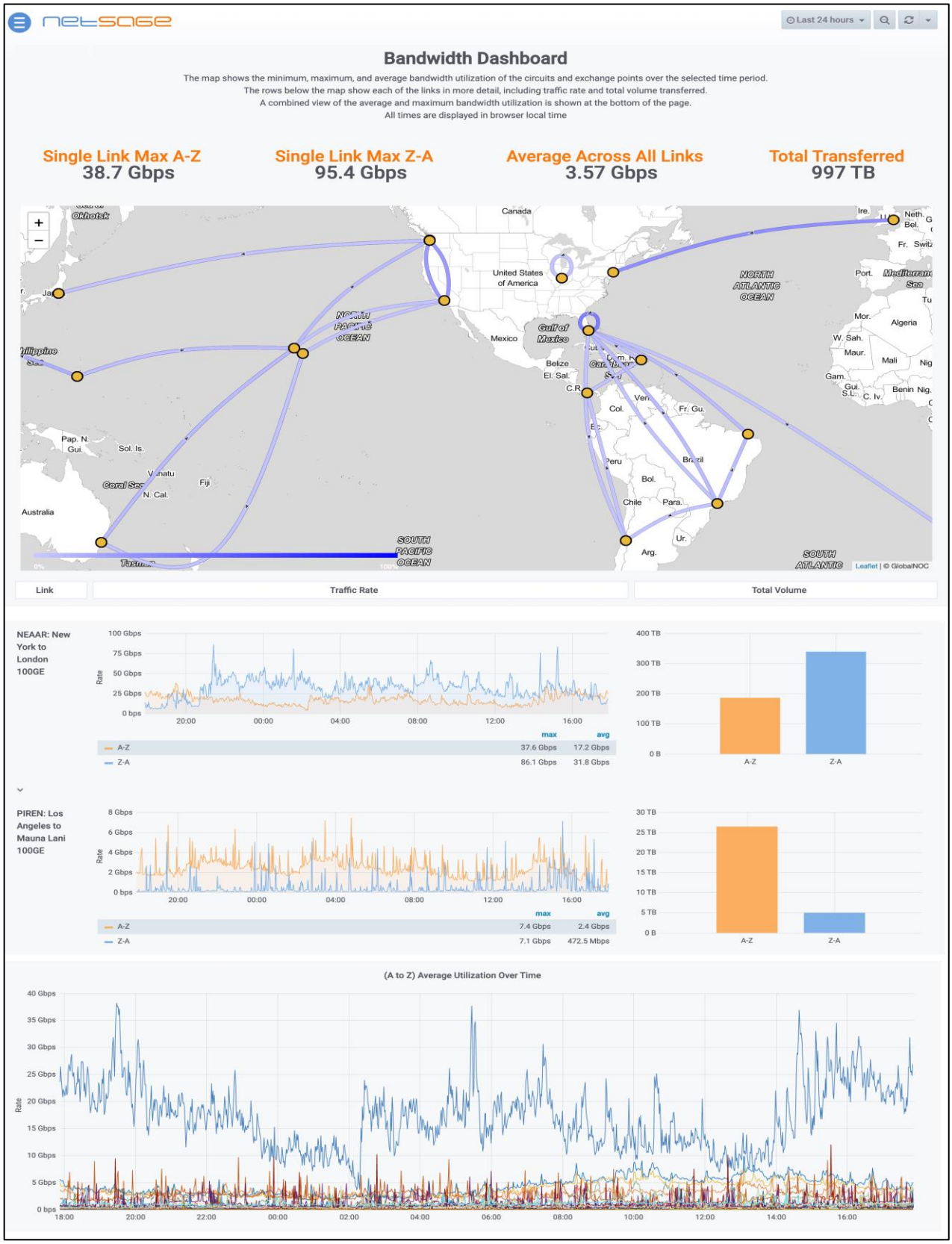

Figure 1: A partial screenshot of the IRNC Bandwidth Dashboard [12], which uses SNMP data to show the status of the IRNC resources on a map, line and bar charts for throughput and transfer volume for each individual circuit, and a set of summary line charts for all of the circuits. Orange indicates traffic flowing from endpoint A to endpoint Z, and Blue indicates traffic flowing in the opposite direction.

movement to help researchers achieve better performance for inter-institutional data sharing.

NetSage is different from other approaches (see Section IV), as it was designed specifically to meet a set of end user questions through innovative Dashboards, not just to supply measurement data to a Network Operations Center (NOC). It was designed to enable further insight by combining multiple data sources to create a result larger than the sum of its parts, and to make that data available to a broad set of end users. NetSage is used for data analysis to understand longer term trends and behaviors, so unlike data used specifically for operations, is flexible if specific data is not available for short time periods.

NetSage was originally funded as part of the NSF International Research Network Connections (IRNC) [4] program to develop and deploy advanced measurement services to understand how the science and engineering community was taking advantage of this NSF-funded research network connections. This included working with, and gathering data from, the seven funded IRNC projects: TransPAC4 [5], America's Lightpath Express and Protect (AmLight ExP) [6], Pacific Islands Research and Education Networks (PIREN) [7], Networks for European, American, and African Research (NEAAR) [8], Atlantic Wave Software Defined Exchange [9], StarLight [10], and Pacific Wave [11], in addition to a set of science data archives. These projects support the majority of data sharing between US researchers and their collaborators all over the world. Fig. 1 shows these projects as part of the IRNC Bandwidth Dashboard [12].

In its first five years, the NetSage project has focused on:

- Understanding the traffic patterns across the IRNC-funded resources;
- Understanding the main sources and destinations for large data transfers, or flows;
- Identifying and visualizing information about the science disciplines and projects that use the IRNC-funded resources;

Displaying patterns of behaviors for data movement between organizations.

NetSage usage statistics show that the project is currently reaching a global community, not only in the number of views but also in how it is being used. Between July 2019 and May 2020, over 3,200 unique users in 71 countries visited the NetSage Dashboards. In addition, over the last 5 years, NetSage team members have made over twenty presentations at meetings including the Internet2 Global Summit, TNC, the NOAA NWave Annual Meeting, the Great Plains Network Annual meeting, the Front Range GigaPop Annual Meeting, and the Internet2 Technical Exchange [13].

III. Design Methodology

The NetSage team adapted the Immersive Empathic Design Methodology [14] for developing visualizations. This process, and other similar techniques such as Design Thinking [15], is standard practice among visualization experts and has been used successfully to produce effective visualizations for many decades. This methodology has eight stages:

1. Create profiles for representative stakeholders to understand their visualization needs.
2. Sketch storyboards to characterize the type of visualization to answer their identified needs.
3. Present storyboards to stakeholders for feedback, which is often accomplished by recording storyboard presentations for stakeholders to view and comment on.
4. Update the storyboards based on the feedback from Stage 3, and reiterate, as time and resources allow.
5. Develop prototypes based on the storyboards.
6. Give early working prototypes to stakeholders for them to try out in their own workflows.
7. Elicit feedback from the stakeholders.
8. Iterative development using the feedback to produce a successively better system, as well as to introduce additional requested features.

For the initial IRNC NetSage deployment, the profiles for representative stakeholders were defined by identifying a set of end users for the NetSage Dashboards and the types of questions they might ask of the data. The initial Dashboard users included:

- IRNC resource owners and operators who wanted to know the status of the resources;
- Collaborative research teams trying to understand resource use and how their data transfers would behave;
- Engineering staff to ensure effective resource use; and
- Funding staff to understand who uses the resources.

After discussions with representatives from each audience, sets of use case questions were identified, including:

- What is the present state of the IRNC resources?
- What are the top sources or destinations for data flows using the IRNC resources?
- What are the top science domains that use the resources?
- What is the maximum, minimum, and average duration of large data transfers?
- Which countries are sharing data using the resources?
- Are there patterns of behaviors that can be identified for how the IRNC resources are used?
- Which sources or destinations have transfers that are not effectively using the IRNC resources.

A series of hand-drawn graphical storyboards were then produced to describe the proposed Dashboards. The Stage 3 feedback enabled the NetSage development team to identify commonalities across the stakeholders and to adapt the Dashboard designs accordingly.

These questions were used by the NetSage development team to design Dashboards with visualizations to provide the answers. This approach not only verified that user goals were being addressed, but also that each Dashboard was focused on addressing the response to a particular question.

IV. NetSage Architecture and IRNC Deployment

The NetSage software consists of a set of open source tools that follows a basic monitoring tool architecture, as shown in Fig. 2. NetSage TestPoints are a collection of software and hardware components that gather active and passive data into records that are then sent to the Data Ingest Pipeline. The five-step Pipeline filters those records and adds additional tags before de-identifying the data. The records are then stored in the NetSage Archive, a centralized storage framework consisting of

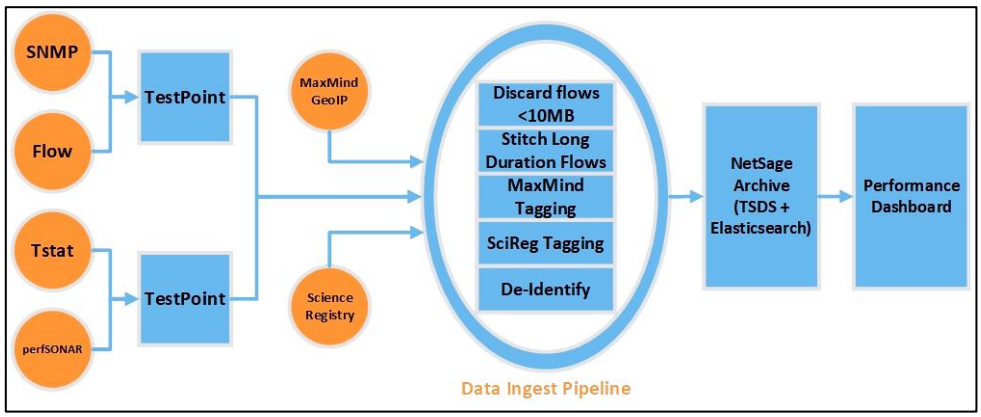

Figure 2: The current logical NetSage network monitoring architecture.

two different databases, a Time Series Data System (TSDS) archive [16] and an Elasticsearch archive [17]. Performance Dashboards built using the open source Grafana [18] analysis and visualization engine, access the records from the NetSage Archive to present visualizations to answer the questions identified by the stakeholders. The effort to build the software has been roughly 5.5 FTE per year over five years.

A. NetSage Data Collection TestPoint

The core of the data collection process for NetSage is the set of hardware and software that make up the logical NetSage TestPoint. TestPoints use both active and passive measurement techniques to gather data for a broader understanding of network behavior. NetSage TestPoints use the Simple Network Management Protocol (SNMP) [19], an application–layer protocol for information about managed devices on IP networks, to passively collect data from routers or switches, generally with a polling rate of once every 60 seconds. This data set includes the interface name, the number of input and output bits, any errors or discards, and the use of unicast or multicast.

The second type of passive data collected by the TestPoint is flow data from routers using tools such as NetFlow [20], sFlow [21], or IPFIX [22]. Flow data is typically sampled at between 1:100 and 1:1000 packets. This data includes information for sampled flows including the source and destination, the number of bits and packets transferred, the duration of the flow, the flow type, and the protocol and port used.

The third type of passive data collected by the TestPoint comes from packet header inspection tools running on science data archives using Tstat [23, 24], which was developed as part of the European Union (EU) Measurement Plane (mplane) FP7 project [25]. Tstat examines the packet headers for the data flowing in and out of instrumented science archive and reports TCP statistics for each flow, including the congestion window size, the number of packets retransmitted, the source and destination, the number of bits and packets transferred, the duration of the flow, the flow type, and the protocol and port used. Unlike similar data collected using a standard flow tool, Tstat data is not sampled.

The fourth dataset collected by the TestPoint is from active measurements using perfSONAR [26], an open source network measurement suite designed to provide end-to-end performance metrics. There are currently over 2,000 perfSONAR nodes deployed worldwide [27]. The NetSage project uses perfSONAR for active measurements of throughput, latency, and loss. Tests are only run briefly four times per day to ensure minimal impact on the production data transfers.

Each of these data sets can be used in multiple ways. In the examples and figures, we highlight which data sets are the source of the information given. Note also that the use cases that NetSage addresses are for long-term trends, so if any particular data source is not available for a period of time, it generally does not affect the analysis or understanding.

B. Data Ingest Pipeline

Records from the TestPoints are sent to the Data Ingest Pipeline, which consists of five stages, as shown in Fig. 2.

In the first stage, records for flows smaller than a threshold (10 MB over 5 minutes) are discarded. The primary goal of NetSage is to understand large-scale data transfers, so records related to small flows are not retained. This filtering increases the level of privacy, as records related to emails or web page downloads are under this threshold and discarded. The filtering also decreases the computation requirement to run the pipeline software, since less data is processed, as well as the storage requirements.

In the second stage of the pipeline, records for longer flows are stitched together into a single unit. Most flow collection techniques share data at specified time intervals, generally 5 minutes. A single data transfer may occur over several time intervals, which results in multiple records that need to be combined together so each record represents a full transfer.

At the third stage, tags are added to the record to map the source and destination of the data transfer to their Autonomous System (AS) Numbers and Names using the MaxMind GeoIP database [28]. This allows the NetSage Dashboards to list the sources and destinations in a more user-friendly way.

In the fourth stage, tags are added to the record to identify the science domain and project information using the NetSage Science Registry [29]. 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

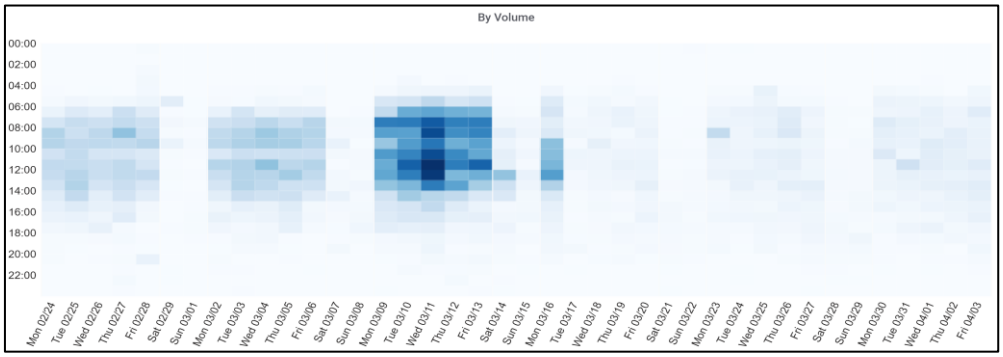

Figure 3: A Heatmap showing data transfers by volume (using flow data) at hourly intervals, with darker colors indicating higher transfer volumes. The x-axis is the day of the month and the y-axis is the time of day. This graph is part of a Dashboard that displays data related to transfers with one endpoint at the Zoom video conferencing facility that crossed the IRNC Pacific Wave Exchange Point to an academic institution source or destination [36].

data entry and is a key component for presenting higher fidelity information about endpoints than what existing MaxMind database can provide. In particular, data transfers with endpoints in the Science Registry can be tagged with information about science projects based on the IP address of the endpoint. This information can include 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, as well as the science discipline, associated organizations, and other project data as available.

In the fifth stage, the low order bits of the IP addresses are stripped off to de-identify the data. One of the system design goals of NetSage was to avoid storing personally identifiable information (PII), and this stage of the pipeline addresses this requirement. Full details are given in the NetSage Data Privacy Policy [30], which was developed to balance the need for user privacy with the practical value of the data. Note that since NetSage does not include full IP addresses, it does not reference data related at a personal level, so it is compliant with the European Union General Data Protection Regulation [31].

C. NetSage Data Archive

After passing through the decentralized Ingest Pipeline, data is stored in the centralized NetSage Archive. The Archive consists of a Time Series Data System (TSDS) [16] archive and an Elasticsearch [17] archive, hosted by the OmniSOC [32].

TSDS, developed by the Indiana University GlobalNOC [33], provides well-structured and high-performance storage and retrieval of time series data and metadata. TSDS is most effective for storing data with constant time intervals. NetSage uses TSDS to store SNMP and perfSONAR data.

The Elasticsearch, Logstash, and Kibana (ELK) Stack is open source software that forms a scalable system used to flexibly ingest, store, and analyze sporadic event data. The Elasticsearch archive stores data as JSON documents and indexes it for quick searching and retrieval. NetSage uses the Elasticsearch archive to store flow data and data from Tstat. One of the features of Elasticsearch is that it is designed to be horizontally scalable, meaning that one can increase both performance and capacity by adding more nodes to the cluster.

D. Dashboard Components

NetSage Dashboards are used to visualize the answers to the stakeholder questions that were identified as part of the design methodology. The Dashboards are built using the open source Grafana analysis and visualization engine and contain sets of widgets that show different aspects of the data in response to a query. In cases where Grafana did not have a ready-made widget for a visualization, new ones were developed in D3.js [34].

We use basic line and bar charts to answer stakeholder questions such as “what is the present state of a resource?” in several dashboards, as shown in Figs. 1, 5, and 7.

We use Heatmaps to show changes in values over time and to easily identify behavior characteristics, and can answer stakeholder questions such as “Are there patterns of behaviors that can be identified for how the IRNC resources are used?”. Fig. 3 shows a Heatmap for the volume of data transferred over 6 weeks with one end of that transfer as the site that hosts the Zoom video conferencing service. This type of display can accentuate changes of behavior, in this case, caused by multiple universities shifting policies in response to the COVID-19 pandemic, as described in Section VI.F.

Sankey graphs [35] show relationships between items using a ribbon graphic, where the width shows the quantity proportionately. We use Sankey graphs as a visual way of answering stakeholder questions such as “Which countries are sharing data using the IRNC resources?”, as shown in Fig. 4 for data transferring to and from an endpoint in the US.

E. Deployment

The NetSage deployment for IRNC is spread across resources at multiple institutions that send data to a central Archive which is used by a centralized Grafana deployment for performance Dashboards. The TestPoint deployment consists of SNMP data for approximately 950 interfaces, with a polling rate of once every 60 seconds; TCP flow data from 11 routers using NetFlow or SFlow with a sample rate of 1:512, unsampled flow statistics using Tstat at 4 science archive sites, and active perfSONAR data running tests four times a day between 13 sites using various servers with approximately 4GB RAM, dual-core 2GHz CPU and a 10Gbps NIC, per perfSONAR requirements [37]. Tstat can be run either directly on an archive head node (as is done at TACC and NERSC), or by splitting the data and

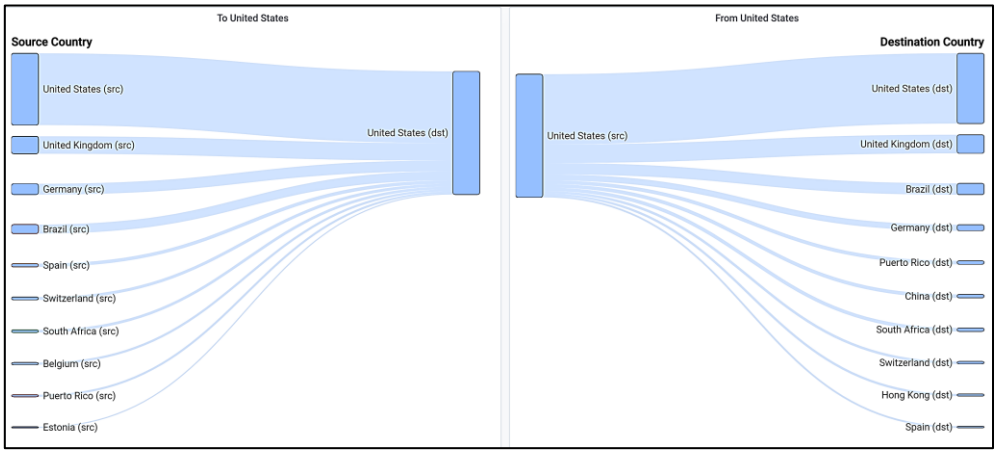

Figure 4: A Sankey graph uses flow data to show the volume of data to and from the United States over 7 days. This graph is part of the Flow Data by Country Dashboard [38]. Wider traces indicate larger data volumes.

collecting it on a separate server (128GB ram, 24cores@2.7GHz, and 1.8TB), as is done at Hawaii and NCAR.

The Ingest Pipeline is also distributed across multiple institutions allowing individual organizations to anonymize and enrich data before sending it to the central archive. The Pipeline throws out approximately 90% of the data by flow count which is under 10M, but this preserves information about 90% of the data by volume transferred using these resources. The NetSage Archive, hosted by the OmniSOC, consists of 6 larger servers (2x PowerEdge R640, 2x Intel(R) Xeon(R) Silver 4116 CPU @ 2.10GHz (24core), 192 GB memory DDR4 2666MHz, 3x 1.92 TB SSD) running ElasticSearch and LogStash, supporting roughly 36TB of SSD. This set up supports archiving over 7.5 million measurement events per day for the IRNC deployment. The Dashboards run separately on a larger server using Grafana 7.0, but this software can run on very slim machines.

NetSage has been designed to run as a managed service, not as independent software, in part because the initial users who approached us to deploy the software did not have the hardware or expertise to support the ELK stack, which is a base requirement. We have focused on ease of deploying the decentralized components, and while the full suite of software is available and open source, the definition of minimum requirements and documentation is incomplete. Currently, NetSage as a Managed Service is supported by the IU GlobalNOC for a modest fee [39], including estimated 0.25 FTE for basic support and maintenance.

V. Design Lessons Learned

In the course of developing any large-scale pragmatic software framework, plans change and lessons are learned. Three of the major lessons as we have experienced while working on NetSage have been to adapt when necessary, to leverage other people's work as much as possible, and that what users request will change as soon as they have a prototype to work with (sometimes referred to as "No plan survives contact with the enemy"-Helmuth von Moltke).

Lesson 1: Adapt when Necessary. NetSage had originally planned to collect both sampled and unsampled flow data from routers by using a packet-header inspection software tool such as Argus [40], Zeek (formerly called Bro) [41], tcptrace [42], or Tstat. Part of the reason behind this approach was to be able to compare the data collected from the different measurement approaches to evaluate any differences. Part of how packet-header inspection tools function is that they track both sides of the "conversation" between a source and a destination for each data transfer. In order to track both sides of the conversation, data sent from the source to the destination has to take the same path through the network as when data is sent from the destination to the source. However, many if not most, international data transfers experience asymmetric routing, in other words, the network path from the source to the destination is not the same as from the destination to the source. Because of this, none of the currently available packet heading inspection tools could be used in the middle of the path at a router.

However, this finding resulted in our realization that we could adapt this approach if the TestPoint collecting the data was located at the endpoint of the path, that is, at the actual source or destination. So, we did an evaluation of common sources and destination of data sets and identified several major data centers that we then approached to see if we could add a TestPoint to their infrastructure to gather data for NetSage. We are now collecting data about transfers to and from science archives at the National Energy Research Scientific Computing Center (NERSC), the Texas Advanced Computing Center (TAAC), the National Center for Atmospheric Research (NCAR), and the University of Hawai'i's Institute for Astronomy (IFA).

Lesson 2: Leverage when Possible. As an NSF-funded project, NetSage planned to leverage other open source projects as strongly as possible in order to maximize project resources. For example, the initial NetSage archive implementation used the existing TSDS database instead of building our own. When the data collection expanded to include flow data, the NetSage Archive was updated to include the Elasticsearch archive as well, as opposed to building a new one for this data type on our own. Our initial implementation of the Data Ingest Pipeline used NF Dump and bespoke scripts, which overtime have been transitioned to taking advantage of a logstash pipeline writing to Elasticsearch. Similarly, the initial NetSage Dashboards were written using custom software, because when the project started

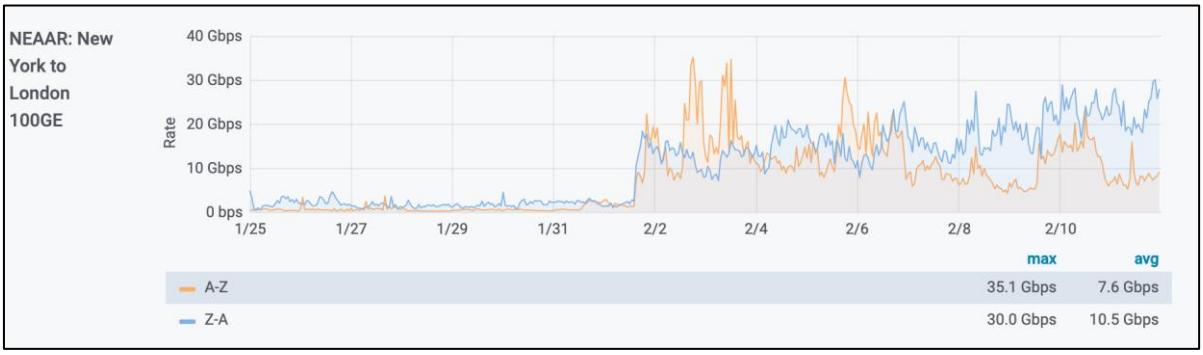

Figure 5: Traffic volume graph using SNMP data to show the increase in network traffic on the NEAAR link between New Yok and London for January-February 2018 [44]. Note the two colors indicate traffic in different directions on the circuit.

there was no clear best toolkit approach to building them. In Year 3, we shifted to using Grafana, which has saved countless hours of development and decreases our support burden.

Lesson 3: Changing Requests. The NetSage development team, like most builders of pragmatic software, has also discovered there are successes and opportunities when working directly with an active user base. For each new Dashboard we have storyboarded and designed, once we have successfully met the stated requirements, the stakeholders take the opportunity to define additional aspects and functionality that is also needed. An ongoing challenge has been to keep the Dashboards focused and simple enough for use by a wide audience, but still delivering the functionality that has been requested. We continue to expand the use cases we address, and the visualizations used to meet user needs.

VI. Discoveries Made Using NetSage

NetSage has been used in practice to find a variety of networking and data transfer behaviors. These include, among others, understanding the resource use and identifying both possible erroneous use as well as a variety of changes in behaviors.

A. *Understanding Traffic Using the NSF-funded International Network Resources*

The use case that NetSage was originally developed to address was to better understand how NSF's multi-million dollar investment in international networks was being used by the US research and education community. NetSage has two dashboards that specifically address this. The first is the Bandwidth Dashboard [12], shown in Fig. 1, which uses SNMP data to generate and display a map for the NSF-funded resources, details about the use of each circuit, and summary line graphs for the average and maximum bandwidth utilization for all of the circuits. The second is the Summary Statistics Dashboard [43], not shown as a figure, which uses SNMP, Flow, Tstat, and Science Registry data to give useful statistics about the IRNC-funded resources as a whole.

B. *Detection of Unexpected Traffic*

Resource owners use NetSage to track typical behavior and to identify when behavior changes occur. For example, Fig. 5 shows part of the Dashboard for the NEAAR project, the IRNC project that supports a link between the US and Europe, which experienced a significant change in behavior [44]. A review of the NetSage Dashboard for Top Flows for the circuit showed that many of the source or destination organizations for the increased traffic were associated with high energy physics research. This behavior shift was due to US Department of Energy network operators adding the NEAAR circuit to the set of network resources that support data transfers related to the Large Hadron Collider (LHC). In this particular case, Dashboards were able to exhibit this change before the email notification was sent to the resource owner. This example shows how NetSage can be used to observe unexpected changes in network usage and prompt further investigation.

C. *Detection of Erroneous Traffic Behaviors*

It is common when additional network capacity is added that the paths used by data transfers will change, often in unexpected

Source	Destination	Total Vol. ▾	Largest Flow	# Flows
KISTI	Computer Network Information Center	35.0 TB	72.3 GB	101.0
University of Hawaii	Indiana University	27.4 TB	14.4 GB	3.1 K
Indiana University	University of Hawaii	18.6 TB	12.9 GB	2.0 K
The Chinese University of Hong Kong	Jisc Services Limited	12.0 TB	36.5 GB	6.6 K
University of Pennsylvania	The University of Hong Kong	10.9 TB	45.5 GB	98.0

Figure 6: A partial listing, from collected flow data, of the top pairs of organizations transferring data over the TransPAC4 connectivity between Guam and Hong Kong for September-December 2018 [45]. This table shows that over 35 TB of data was incorrectly moving from KISTI, in South Korea, to the Computer Network Information Center, in China.

Figure 7: Partial Flow Analysis [46] Dashboard showing SNMP and flow data that identifies recurring data transfers that were part of a VLBI astronomy research project between Italy and Japan over the TransPAC4 link between Seattle and Tokyo in October 2018.

ways. One example of this was seen when staff members for the TransPAC4 project used NetSage to understand how adding a new 20G connection between Guam and Hong Kong would affect traffic. The partial Dashboard [45], shown in Fig. 6, shows that the top pair of organizations transferring data over the new connection was the Korea Institute of Science and Technology Information (KISTI) in South Korea and the Computer Network Information Center (CNIC) at the Chinese Academy of Science. A deeper investigation showed the traffic path to include South Korea -> Hong Kong -> Guam -> Hawai'i -> LA -> Seattle -> Japan -> China, which was clearly not intended. In other words, traffic that should have been able to go directly from China to South Korea was crossing the Pacific Ocean twice, resulting in significantly decreased performance. A discussion with the network engineers overseeing different parts of the path determined that the routing preferences were incorrect, and the problem was resolved.

D. Detection of Unusual Data Transfer Patterns

TransPAC4 is the IRNC project that supports connections between the US and Asia. Engineers for TransPAC4 used NetSage to identify an unusual pattern of behavior where every 10-12 days there was a significant increase in the data volume over the resource, as shown in the partial screen shot of the Analysis Dashboard in Fig. 7. This Dashboard [46] shows the advantage of being able to combine both SNMP and flow data into the same dashboard to easily identify the source of spikes in performance. When using it interactively, the time frame can be

zoomed in on a specific SNMP behavior, and the Top Talkers and Individual Flows related to that time frame are displayed.

Investigation indicated that the periodic data transfers were taking place between the Instituto di Radioastronomia, in Italy, and the Kashima Space Technology Center in Kashima, Japan, and that the traffic was related to an astronomy very-long-baseline interferometry (VLBI) project. The workflow for VLBI applications involves several geographically distributed radio telescopes that are all aligned on the same celestial object, in this case, all located in Italy, sending their data to a collector site, in this case in Japan. This example shows how NetSage can be used to identify patterns of behaviors for how the resources are being used.

E. Understanding a Universities International Data Movement

On occasion, an institution may be asked to provide details about how it interacts with other educational institutions internationally. NetSage can provide this data, as shown in the example for Emory University in Fig. 8 [47]. During the six-month timeframe, data shows that Emory receives more data

Figure 8: Partial Flow Data by Organization Dashboard [47] showing flow data for Emory University over the IRNC resources for six months.

than it sends, and that it works with a wide range of internaitonal collaboration sites.

F. Understanding Resource Use

In March 2020, many US universities put in place travel restrictions due to the COVID-19 pandemic. Fig. 3 shows a Heatmap from a Dashboard [36] for data transfers to and from the Zoom video conferencing hosting site during the timeframe where R&E institutional use of Zoom changed radically. The Heatmap shows data volumes starting in February that increase on/around March 12 when many US universities declared that researchers couldn't travel. This was followed 10 days later by a decrease, likely caused by a combination of institutions shifting to Spring Break, institutions issuing "Work from Home" directives (so the traffic shifted to home networks, not R&E networks), and Zoom shifting some of its hosting to use cloud services, not at their IP space. This is one example of how longitudinal data collection can provide broader context for changing resource use especially in unusual circumstances.

VII. Related Work

Over the past 20 years, the research and education network community has developed numerous monitoring portals similar to NetSage, including my.es.net for ESnet [48], the Gloriad Portal for the Gloriad network [49], the IU Global NOC's WorldView [50], CERN's monitoring portal for the LHCOPN network [51], and the GÉANT tools portal [52]. Each of these portals were developed primarily for use by a network operations center to provide a view for a single network provider, primarily to understand and address network outages. NetSage was developed for a broader set of users and to be able to analyze network performance related to data integrated from multiple networks and resources, primarily to understand performance issues and performance degradation, not outages. NetSage was influenced by or leverages prior work from some of these portals, such as the flow analysis capabilities of the ESnet portal and the Science Project database used in the Gloriad Portal.

There is also a large set of measurement and monitoring tools that are not full R&E portals but were also developed primarily to support network operations center staff and include some of the functionality also supported by NetSage, as shown in Table I. None of these include all of the features supported by NetSage, for example, being able to identify a science resource. Only Kentik and SolarWinds use both flow data and perfSONAR data, similar to NetSage, but neither of these are Open Source. The most common feature supported by these tools that is not included in NetSage is alerting, which is planned as part of the next year's development cycle.

TABLE I. Comparison of Related Work

	Functionality/ Attribute	Graphical User Interface	SNMP Data	PerfSonar Data	Flow Data	Science Registry Data	Alerting	Open Source
Tool	NetSage [4]	Y	Y	Y	Y	Y	N	Y
	Argus [40]	Y	N	N	N	N	Y	Y
	Cacti [53]	Y	Y	N	N	N	N	Y
	Deep Field [54]	Y	N	Y	N	N	N	N
	Elastiflow [55]	Y	N	N	Y	N	N	Y
	InMon [56]	Y	N	N	Y	N	Y	N
	Kentik [57]	Y	N	Y	Y	N	Y	N
	WLCG Grafana Dash [51]	Y	Y	Y	N	N	N	Y
	Nagios [58]	Y	N	N	N	N	Y	Y
	NFSEN [59]	Y	N	N	Y	N	Y	Y
	Ntop [60]	Y	Y	N	Y	N	Y	N
	SolarWinds [61]	Y	Y	Y	Y	N	Y	N

VIII. Conclusion and Future Work

In this paper, we have described the NetSage design methodology, its architecture and the current IRNC deployment, as well as several use cases and discoveries. Within the space of Research and Education networks, we believe it is the most comprehensive open source approach to date that enables insight into underlying resource behaviors, and as such, differs significantly from other approaches which have been developed for network operations. NetSage is in current use today by not only the NSF-supported IRNC resources described here, but by several US domestic regional and state networks.

Future work will continue to be driven by stakeholder requests using our design methodology. In the short term we are developing additional visualizations as well as continuing to add more data to the Science Registry to be able to better reveal network use patterns of scientific applications. Longer term work includes adding in alarms and alerts and exploring adaptations needed to use NetSage in a campus environment to better meet the needs of research collaborations, which will require a different privacy model, as the dashboards will not be able to be public, and the granularity of information will need to be much finer. Overall, the core NetSage development will continue to be funded in part by NSF through IRNC and other projects at least until 2022, but is also partially supported for the foreseeable future via third-party deployments with the IU Global NOC.

Acknowledgment

The authors would like to thank S. Peisert, A. Giannakou, and D. Dwivedi at Lawrence Berkeley National Laboratory, and D. Kobayashi, N. Kirshenbaum, B. T. Wooton, and P. Karjala University of Hawaii at Mānoa's Laboratory for Advanced Visualization and Applications. This project was funded by the National Science Foundation award ACI #1540933.

REFERENCES

[1] Dart, Wehner, and Prabhat. An Assessment of Data Transfer Performance for Large-Scale Climate Data Analysis and Recommendations for the Data Infrastructure for CMIP6. 2018. http://www.escholarship.org/uc/item/91z9m2sm

[2] IRNC: AMI: NetSage - An Open, Privacy-Aware, Network Measurement, Analysis, and Visualization Service, NSF-1540933, 05/2015-04/2021, PI Jennifer Schopf, coPIs Jason Leigh, Andrew Lake (former coPI Sean Peisert), https://www.nsf.gov/awardsearch/showAward?AWD_ID=1540933

[3] Shibboleth – Internet2. https://www.internet2.edu/products-services/trust-identity/shibboleth/

[4] NSF International Research Network Connections Program, https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503382

[5] IRNC-BackBone- TransPAC4 - Pragmatic Application-Driven International Networking. NSF-1450904. 2015-2021. PI Jennifer Schopf, CoPI Hans Addleman. http://www.nsf.gov/awardsearch/showAward?AWD_ID= 1450904

[6] IRNC: Backbone: AmLight Express and Protect (ExP). NSF-1451018. 2015-2021. PI Julio Ibarra, CoPI Heidi L. Morgan, Donald Cox. http://nsf.gov/awardsearch/showAward?AWD_ID=1451018

[7] IRNC Backbone: SXTransPORT Pacific Islands Research and Education Network (PIREN). NSF-1451058. 2015-2021. PI David Lassner, CoPI Gwen Jacobs, Ronald Johnson, Louis Fox. https://www.nsf.gov/awardsearch/showAward?AWD_ID=1451058

[8] IRNC: Backbone: NEAAR: Networks for European, American, and African Research. NSF-1638863. 2016-2021. PI Jennifer Schopf, CoPI Edward Moynihan, Cathrin Stover. http://www.nsf.gov/awardsearch/showAward?AWD_ID=1638863

[9] IRNC: RXP: AtlanticWave-Software Defined Exchange: A Distributed Intercontinental Experimental Software Defined Exchange (SDX). NSF-1451024. 2015-2021. PI Julio Ibarra, CoPI Russell Clark, Heidi L. Morgan. https://www.nsf.gov/awardsearch/showAward?AWD_ID=1451024

[10] IRNC: RXP: StarLight SDX A Software Defined Networking Exchange for Global Science Research and Education. NSF-1450871. 2015-2021. PI Joel Mambretti, CoPI Maxine Brown, Thomas DeFanti, Jim Hao Chen. http://www.nsf.gov/awardsearch/showAward?AWD_ID= 1450871

[11] IRNC: RXP - Pacific Wave Expansion Supporting SDX & Experimentation NSF-1451050. 2015-2021. PI Louis Fox, coPI Ronald Johnson. https://www.nsf.gov/awardsearch/showAward?AWD_ID=1451050

[12] Netsage Grafana Portal. https://portal.netsage.global

[13] NetSage Project website. http://www.netsage.global

[14] Yu-Chung Chen, Sangyoon Lee, HyeJung Hur, Jason Leigh, Andrew Johnson, Luc Renambot. Design an interactive visualization system for core drilling expeditions using immersive empathic method. CHI EA '09. 2009. Pg. 2671–2674. https://doi.org/10.1145/1520340.1520382

[15] M. Carroll, L. Britos, S. Goldman. Becoming a Design Thinker. London: Berg Publishers Open University. 2012.

[16] Time Series Data System (TSDS). https://docs.globalnoc.iu.edu/software/measurement/tsds.html

[17] Powering Data Search, Log Analysis, Analytics. https://www.elastic.co/products

[18] Grafana - The open platform for analytics and monitoring. https://grafana.com/

[19] J.d. Case, M. Fedor, M.l. Schoffstall, and J. Davin. Simple Network Management Protocol (SNMP). 1990. https://doi.org/10.17487/rfc1157

[20] Introduction to Cisco IOS NetFlow - A Technical Overview. 2012. https://www.cisco.com/c/en/us/products/collateral/ios-nx-os-software/ios-netflow/prod_white_paper0900aecd80406232.html

[21] Phaal, Panchen, and Mckee. 2001. InMon Corporations sFlow: A Method for Monitoring Traffic in Switched and Routed Networks. (2001). https://doi.org/10. 17487/rfc3176

[22] Benoit Claise, Stewart Bryant, Simon Leinen, Thomas Dietz, and Brian H. Trammell. 2013. Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of IP Traffic Flow Information. https://tools.ietf.org/html/rfc5101

[23] Rossi. 2008. TCP STatistic and Analysis Tool. http://tstat.polito.it/

[24] Mellia, Locigno, and Neri. 2005. Measuring IP and TCP behavior on edge nodes with Tstat. Computer Networks 47, 1 (2005), 1–21. https://doi.org/10.1016/ s1389- 1286(04)00201- 4

[25] Brian Trammell, et. al. 2014. mPlane: an intelligent measurement plane for the Internet. IEEE Communications Magazine 52, 5 (2014), 148–156.

[26] Brian Tierney, Joe Metzger, Jeff Boote, Eric Boyd, Aaron Brown, Rich Carlson, Matt Zekauskas, Jason Zurawski, Martin Swany, Maxim Grigoriev, "perfSONAR: Instantiating a global network measurement framework." In Proceedings of the SOSP Wksp. Real Overlays and Distrib. Sys, 2009. http://www.perfsonar.net/

[27] perfSONAR nodes: http://stats.es.net/ServicesDirectory/

[28] GeoIP, MaxMind, Inc. 2019. https://www.maxmind.com/

[29] NetSage Science Registry. https://scienceregistry.netsage.global/rdb/about.html

[30] NetSage Privacy Policy. 2017. http://www.netsage.global/home/netsage-privacy-policy

[31] General Data Protection Regulation (GDPR) https://gdpr-info.eu/

[32] OmniSOC, 2020, https://omnisoc.iu.edu/

[33] GlobalNOC at Indiana University. https://globalnoc.iu.edu/

[34] Mike Bostock. D3 Data Driven Documents. 2019. http://d3js.org

[35] Sankey Diagram. https://en.wikipedia.org/wiki/Sankey_diagram

[36] PacificWave Portal - Individual Flows Dashboard for Zoom Video Communications. https://pacwave.netsage.global/grafana/d/-l3_u8nWk/individual-flows?orgId=2&var-src=Zoom%20Video%20Communications,%20Inc&from=1582538400000&to=1585994399000

[37] PerfSONAR Hardware Requirements, http://docs.perfsonar.net/install_hardware.html

[38] IRNC NetSage Flow Data by Country Dashboard. https://portal.netsage.global/grafana/d/fgrOzz_mk/flow-data-per-country?orgId=2

[39] "IU's GlobalNOC offers network monitoring service, improving research collaboration", IU Press Reelease, January 31, 2020, https://itnews.iu.edu/articles/2020/GlobalNOC-offers-network-monitoring-service-improving-research-collaboration%20-.php

[40] ARGUS - Software Solutions for the CRE Industry. https://www.altusgroup.com/argus/

[41] Zeek - An Open Source Network Security Monitoring Tool. https://zeek.org/

[42] Shawn Ostermann. Tcptrace. http://www.tcptrace.org/

[43] IRNC NetSage Flow Data Summary Statistics Dashboard, https://portal.netsage.global/grafana/d/CJC1FFhmz/other-flow-stats?orgId=2

[44] Netsage Dashboard for IRNC circuits, Januaery 25-February 11, 2018. https://portal.netsage.global/grafana/d/000000003/bandwidth-dashboard?orgId=2&from=1516856400000&to=1518411599000&var-links=All

[45] NetSage Flow Data Dashboard for TransPAC Guam-Hong Kong circuits for September 1-December 31, 2018. https://portal.netsage.global/grafana/d/xk26IFhmk/flow-data-for-circuits?orgId=2&var-Sensors=TransPAC%20Hong%20Kong%20sFlow&var-source_scope=meta.src_organization.keyword&var-dest_scope=meta.dst_organization.keyword&from=1535774400000&to=1546232399000

[46] Netsage Flow Analysis Dashboard. https://portal.netsage.global/grafana/d/VuuXrnPWz/flow-analysis?orgId=2&from=1538366400000&to=1541131199000&var-sensors=TransPAC%20Seattle%20sFlow&var-links=TransPAC:%20Seattle%20to%20Tokyo%20100GE&var-src_scope=meta.src_organization.keyword&var-dst_scope=meta.dst_organization.keyword

[47] Netsage Dashboard for Flow Data by Organization for Emory University, December 2, 2019-May31, 2020.

https://portal.netsage.global/grafana/d/QfzDJKhik/flow-data-per-organization?orgId=2&var-Sensors=All&var-Organization=Emory%20University&from=1575176400000&to=1590983999000

[48] MyESnet Portal. http://my.es.net

[49] Global Ring Network for Advanced Applications Development (GLORIAD). NSF-0441102. 2005-2012. PI Greg Cole. http://www.gloriad.org/

[50] GlobalNOC World View. 2019. https://docs.globalnoc.iu.edu/worldview.html

[51] Babik, M., McKee, S., Bockelman, B., Hernandez, E., Martelli, E., Vukotic, I., Weitzel, D., and Zvada, M.. (2019). Improving WLCG Networks Through Monitoring and Analytics. EPJ Web of Conferences. 214. 08006. 10.1051/epjconf/201921408006.

[52] GÉANT Tools Portal. 2019. https://tools.geant.net/portal/

[53] Cacti - The Complete RRDTool-Based Graphing Solution. https://www.cacti.net/

[54] Deepfield, Nokia. https://www.nokia.com/networks/solutions/deepfield/

[55] Rob Cowart. Elastiflow - Network flow Monitoring (Netflow, sFlow and IPFIX) with the Elastic Stack. 2019. https://github.com/robcowart/elastiflow

[56] InMon. https://inmon.com/

[57] Kentik - AIOps for Network Professionals: Network Flow Analytics, Network Monitoring & DDoS Detection. https://www.kentik.com/

[58] Nagios - The Industry Standard In IT Infrastructure Monitoring. 2020. https://www.nagios.org/.

[59] Peter Haag. User Documentation nfdump & NfSen. https://www.first.org/conference/2006/papers/haag-peter-papers.pdf

[60] ntop – High Performance Network Monitoring Solutions Based on Open Source and Commodity Hardware. https://www.ntop.org/.

[61] SolarWinds. IT Management Software & Monitoring Tools. https://www.solarwinds.com/