NATIONAL CENTERS FOR
ENVIRONMENTAL INFORMATION

NOAA National Centers for Environmental Information Fisheries Acoustics Archive Network Deep Dive

August 19, 2021

U.S. DEPARTMENT OF
ENERGY
Office of Science

ESnet
ENERGY SCIENCES NETWORK

Disclaimer

NOAA National Centers for Environmental Information Fisheries Acoustics Archive Network Deep Dive

Final Report

August 19, 2021

The Engagement and Performance Operations Center (EPOC) is supported by the National Science Foundation under Grant No. 1826994.

ESnet is funded by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research. Benjamin Brown is the ESnet Program Manager.

ESnet is operated by Lawrence Berkeley National Laboratory, which is operated by the University of California for the U.S. Department of Energy under contract DE-AC02-05CH11231.

This is a University of California, Publication Management System report number LBNL-2001417[1].

[1] https://escholarship.org/uc/item/0wm478c4

Participants & Contributors

Hans Addleman, Indiana University
Charles Anderson, University of Colorado Boulder / National Oceanic and Atmospheric Administration
Alex Hsia, National Oceanic and Atmospheric Administration
Ken Miller, ESnet
Veronica Martinez, University of Colorado Boulder / National Oceanic and Atmospheric Administration
Doug Southworth, Indiana University
Carrie Wall, University of Colorado Boulder / National Oceanic and Atmospheric Administration
Jason Zurawski, ESnet

Report Editors

Hans Addleman, Indiana University: addlema@iu.edu
Ken Miller, ESnet: ken@es.net
Doug Southworth, Indiana University: dojosout@iu.edu
Jason Zurawski, ESnet: zurawski@es.net

Contents

1 Executive Summary

Deep Dive Review Purpose and Process

EPOC uses the Deep Dive process to discuss and analyze current and planned science use cases and anticipated data output of a particular use case, site, or project to help inform the strategic planning of a campus or regional networking environment. This includes understanding future needs related to network operations, network capacity upgrades, and other technological service investments. A Deep Dive comprehensively surveys major research stakeholders' plans and processes in order to investigate data management requirements over the next 5–10 years. Questions crafted to explore this space include the following:

- How, and where, will new data be analyzed and used?
- How will the process of doing science change over the next 5–10 years?
- How will changes to the underlying hardware and software technologies influence scientific discovery?

Deep Dives help ensure that key stakeholders have a common understanding of the issues and the actions that a campus or regional network may need to undertake to offer solutions. The EPOC team leads the effort and relies on collaboration with the hosting site or network, and other affiliated entities that participate in the process. EPOC organizes, convenes, executes, and shares the outcomes of the review with all stakeholders.

This Review

Between May 2021 and August 2021, staff members from the Engagement and Performance Operations Center (EPOC) met with researchers and staff from the National Oceanic and Atmospheric Administration (NOAA)'s N-Wave (the Enterprise network that supports the NOAA mission) and National Centers for Environmental Information (NCEI)'s Fisheries Acoustics Archive for the purpose of recording a Deep Dive into research drivers. The goal of these meetings was to help characterize the requirements for the research use case, and to enable cyberinfrastructure support staff to better understand the needs of the researchers they support.

Material for this event includes documentation on the research use case as well as other components related to the current state of technology support, and a write-up of the discussion that took place via e-mail and video conferencing. The Case Study highlights the ongoing challenges that NOAA has in supporting research use cases that are increasing in capability and are in need of advanced technology to increase productivity and output beyond the current levels .

The review produced several important findings:

- Data set volumes and quantities will grow in the coming years, from a number of different users and use cases, along with a need to support downloads from around the world.

- Acoustic data set sizes currently range between 100s of GB to 10s of TB, and a total archive size of 100s of TBs

- Acoustic data set sizes will grow to 10s of TB to 100s of TB, and a total archive size of 100s of PBs.

- The current set of local resources (computation and storage) will be insufficient in coming years, even with a migration to the cloud to offload some of the computational and storage resources. Growing data sets will still require local resources to handle portions of the scientific workflow related to data ingest.

- ”Human”-centric technology resources are always needed, and often in short supply, to integrate and maintain workflows and basic computational, storage, and networking needs.

- Metrics that track usage and value of data sets are required, even after the migration to cloud resources is complete. These will help the Fisheries Acoustic Archive understand usage patterns from the user community.

- Consideration for automating some parts of the workflow once cloud resources are integrated, while still maintaining some pieces that must be operated locally.

Lastly, EPOC and NOAA identified a number of actions that will be followed up on in subsequent years:

1) Upgrading local resources (computation, storage) to support Fisheries Acoustic Archive even with the migration to cloud, due to some parts of the workflow being performed manually.

2) Allocating “Human” technology resources that can help with software and hardware components of the scientific workflow.

3) While portions of the data ingest process involve importing data from contributor-submitted physical media, there are approaches to mitigate:
    a) Working with contributors that submit physical media to understand workflows and process
    b) Adopting certain hardware and software approaches. e.g., Data Transfer Nodes (DTNs) and advanced data mobility software
    c) Keeping local resources at the Fisheries Acoustic Archive up to date to facilitate the use of physical media data ingest

4) Making ways to track and report usage metrics for data accessed via the cloud.

5) Adopting approaches that improve data mobility for a number of use cases.

2 Deep Dive Findings

The Deep Dive process helps to identify important facts and opportunities from the profiled use cases. The following outlines a set of findings from the Deep Dive that summarize important information gathered during the discussions surrounding the C ase Study:

- Acoustic data set submissions will continue to grow in the coming years, from a number of different users and use cases. These will include NOAA sources, but also others from universities, other federal agencies, non-profits, and industrial collaborators.

- Acoustic data set download demands will continue to grow in the coming years from the same set of collaborators: NOAA sources, universities, other federal agencies, non-profits, and industrial collaborators.

- Acoustic data set sizes will continue to grow: both in terms of individual contributions, but also the entire archive that the Fisheries Acoustic Archive team manages.
    - The current ranges for individual data sets are best described as 100s of GB to 10s of TB, which results in a complete archive size that can be measured in 100s of TBs.
    - Individual submissions will grow to 10s to 100s of TBs in the coming years, implying that the archive will grow to 100s of PBs.

- The current set of local resources, defined to be computational and storage, for the Fisheries Acoustic Archive will be insufficient to adequately address the growing challenges in the coming years. While migration to the cloud is planned, local resources are still required to accomplish key aspects of the workflow: namely receiving, checking, and importing data submissions from collaborators.

- The Fisheries Acoustic Archive team still requires regular ”human”-centric technology resources to integrate and maintain workflows and basic computational, storage, and networking needs. The core of the team supports the scientific mission, but are not subject matter experts in networking, software development, or computational and storage integration.

- The Fisheries Acoustic Archive is requesting a more efficient, and transparent, mechanism to track metrics of use for the public-facing data archives. This information is accessible for managers of the cloud platform, but not immediately available to the users that are posting and maintaining the data sets.

- Fisheries Acoustic Archive would like to move toward a fully automated workflow during their process of science, which would mean working to simplify the ways that data ingest, processing, storage and retrieval are managed. Doing so is desirable for both the operations and end-user experiences.

3 Deep Dive Action Items

EPOC and NOAA recorded a set of action items that are a reflection of the Case Study report and discussions:

- The local resources that Fisheries Acoustic Archive uses to accomplish the process of science, namely workstations and virtual machines that provide computational support as well as temporary and long term storage capabilities, must be upgraded even with the planned migration to the cloud. Certain aspects of the workflow, namely data ingest, verification, and categorization, cannot be fully outsourced to the cloud and require upgraded local capabilities to keep up with the growing data submission volume and quantity.

- "Human" technology resources, namely experienced staff that can help with software and hardware components of the scientific workflow are required to assist the Fisheries Acoustic Archive team. In particular, a backlog of technical problems exists that is preventing progress for several aspects of the workflow, and the migration to the cloud.

- The Fisheries Acoustic Archive data ingest process often involves a manual step of importing data from contributor-submitted physical media. It is not expected that this part of the workflow will change in the coming years, due to the sophistication of some of the contributions, and the growing sizes of the data. Because of this, several steps are recommended:
    - Working with known groups that submit physical media to better understand their use case and workflow, and work to move to a more automated method
    - Adopting certain hardware and software approaches. e.g., Data Transfer Nodes (DTNs) and advanced data mobility software
    - Keeping local resources at the Fisheries Acoustic Archive up to date to facilitate the use of physical media data ingest

- As the Fisheries Acoustic Archive data migrates wholly to the cloud, a system to track and report usage metrics is required to better help the team understand the value of data products. Investing in this will help to guide resource allocation in the future.

- Data mobility has many forms in the Fisheries Acoustic Archive workflow. In particular there are numerous contributors that submit data for archival, there is the workflow to upload the data into a cloud resource, and lastly the user population downloading from the cloud. To keep performance at high levels for these use cases, it is recommended that NOAA consider adopting high performance software and hardware components to ensure fast data transfer.

4 Process Overview and Summary

4.1 Campus-Wide Deep Dive Background

Over the last decade, the scientific community has experienced an unprecedented shift in the way research is performed and how discoveries are made. Highly sophisticated experimental instruments are creating massive datasets for diverse scientific communities and hold the potential for new insights that will have long-lasting impacts on society. However, scientists cannot make effective use of this data if they are unable to move, store, and analyze it.

The Engagement and Performance Operations Center (EPOC) uses the Deep Dives process as an essential tool as part of a holistic approach to understand end-to-end research data use. By considering the full end-to-end research data movement pipeline, EPOC is uniquely able to support collaborative science, allowing researchers to make the most effective use of shared data, computing, and storage resources to accelerate the discovery process.

EPOC supports five main activities

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

Whereas the Roadside Assistance portion of EPOC can be likened to calling someone for help when a car breaks down, the Deep Dive process offers an opportunity for broader understanding of the longer term needs of a researcher. The Deep Dive process aims to understand the full science pipeline for research teams and suggest alternative approaches for the scientists, local IT support, and national networking partners as relevant to achieve the long-term research goals via workflow analysis, storage/computational tuning, identification of network bottlenecks, etc.

The Deep Dive process is based on an almost 10-year practice used by ESnet to understand the growth requirements of Department of Energy (DOE) facilities[2]. The EPOC team adapted this approach to work with individual science groups through a set of structured data-centric conversations and questionnaires.

[2] https://fasterdata.es.net/science-dmz/science-and-network-requirements-review

4.2 Campus-Wide Deep Dive Structure

The Deep Dive process involves structured conversations between a research group and relevant IT professionals to understand at a broad level the goals of the research team and how their infrastructure needs are changing over time.

The researcher team representatives are asked to communicate and document their requirements in a case-study format that includes a data-centric narrative describing the science, instruments, and facilities currently used or anticipated for future programs; the advanced technology services needed; and how they can be used. Participants considered three timescales on the topics enumerated below: the near-term (immediately and up to two years in the future); the medium-term (two to five years in the future); and the long-term (greater than five years in the future).

The Case Study document includes:

- ***Science Background***—an overview description of the site, facility, or collaboration described in the Case Study.
- ***Collaborators***—a list or description of key collaborators for the science or facility described in the Case Study (the list need not be exhaustive).
- ***Instruments and Facilities***—a description of the network, compute, instruments, and storage resources used for the science collaboration/program/project, or a description of the resources made available to the facility users, or resources that users deploy at the facility.
- ***Process of Science***—a description of the way the instruments and facilities are used for knowledge discovery. Examples might include workflows, data analysis, data reduction, integration of experimental data with simulation data, etc.
- ***Remote Science Activities***—a description of any remote instruments or collaborations, and how this work does or may have an impact on your network traffic.
- ***Software Infrastructure***—a discussion focused on the software used in daily activities of the scientific process including tools that are used locally or remotely to manage data resources, facilitate the transfer of data sets from or to remote collaborators, or process the raw results into final and intermediate formats.
- ***Network and Data Architecture***—description of the network and/or data architecture for the science or facility. This is meant to understand how data moves in and out of the facility or laboratory focusing on local infrastructure configuration, bandwidth speed(s), hardware, etc.
- ***Cloud Services***—discussion around how cloud services may be used for data analysis, data storage, computing, or other purposes. The case studies included an open-ended section asking for any unresolved issues, comments or concerns to catch all remaining requirements that may be addressed by ESnet.

- ***Resource Constraints***—non-exhaustive list of factors (external or internal) that will constrain scientific progress. This can be related to funding, personnel, technology, or process.
- ***Outstanding Issues***—Final listing of problems, questions, concerns, or comments not addressed in the aforementioned sections.

During a series of virtual meetings , this document is walked through with the research team (and usually cyberinfrastructure or IT representatives for the organization or region), and an additional discussion takes place that may range beyond the scope of the original document. At the end of the interaction with the research team, the goal is to ensure that EPOC and the associated CI/IT staff have a solid understanding of the research, data movement, who's using what pieces, dependencies, and time frames involved in the Case Study, as well as additional related cyberinfrastructure needs and concerns at the organization. This enables the teams to identify possible bottlenecks or areas that may not scale in the coming years, and to pair research teams with existing resources that can be leveraged to more effectively reach their goals.

4.3 Organizations Involved

The Engagement and Performance Operations Center (EPOC) was established in 2018 as a collaborative focal point for operational expertise and analysis and is jointly led by Indiana University (IU) and the Energy Sciences Network (ESnet). EPOC provides researchers with a holistic set of tools and services needed to debug performance issues and enable reliable and robust data transfers. By considering the full end-to-end data movement pipeline, EPOC is uniquely able to support collaborative science, allowing researchers to make the most effective use of shared data, computing, and storage resources to accelerate the discovery process.

The Energy Sciences Network (ESnet) is the primary provider of network connectivity for the U.S. Department of Energy (DOE) Office of Science (SC), the single largest supporter of basic research in the physical sciences in the United States. In support of the Office of Science programs, ESnet regularly updates and refreshes its understanding of the networking requirements of the instruments, facilities, scientists, and science programs that it serves. This focus has helped ESnet to be a highly successful enabler of scientific discovery for over 25 years.

Indiana University (IU) was founded in 1820 and is one of the state's leading research and educational institutions. Indiana University includes two main research campuses and six regional (primarily teaching) campuses. The Indiana University Office of the Vice President for Information Technology (OVPIT) and University Information Technology Services (UITS) are responsible for delivery of core information technology and cyberinfrastructure services and support.

National Oceanic and Atmospheric Administration (NOAA) is an agency that enriches life through science. NOAA's reach goes from the surface of the sun to the depths of the ocean floor to keep the public informed of the changing environment

around them. NOAA's mission to better understand the natural world and help protect its precious resources extends beyond national borders to monitor global weather and climate, and work with partners around the world. NOAA works to understand and predict changes in climate, weather, oceans, and coasts, to share that knowledge and information with others, and to conserve and manage coastal and marine ecosystems and resources.

National Centers for Environmental Information (NCEI) is responsible for preserving, monitoring, assessing, and providing public access to the Nation's treasure of geophysical data and information.

N-Wave is a NOA operated and managed highly scalable, stable and secure network consisting of a private carrier class network backbone that supports NOAA's scientific mission by providing high speed networking services to NOAA customer sites, programs, line offices, and research facilities. N-Wave is built in partnership with the Science, Research and Education (R&E) network community including: Internet2 (I2), the Global Research Network Operations Center (GlobalNOC) and several Advanced Regional and State Networks including: the Mid-Atlantic Gigapop (MAGPI), the Mid-Atlantic Crossroads (MAX), the North Carolina Research and Education Network (NCREN/MCNC), Florida LambdaRail (FLR), OneNet (Oklahoma's R&E network), the Lonestar Education and Research Network (LEARN), the Front Range GigaPop (FRGP), the Pacific Northwest Gigapop (PNWGP), the University of Hawaii, University of Utah/Utah Education Network (UEN), and the Corporation for Education Network Initiatives in California (CENIC).

5 Fisheries Acoustics Archive Case Study

NOAA's National Centers for Environmental Information (NCEI) hosts and provides public access to one of the most significant archives for environmental data on Earth. NCEI provides over 37 petabytes of comprehensive atmospheric, coastal, oceanic, and geophysical data, and is the nation's leading authority for environmental data. The NCEI archive is one of the largest known resources for atmospheric, coastal, geophysical, and oceanic research in the world. NCEI contributes to the National Environmental and Satellite Data Information Service (NESDIS) mission by developing new products and services that span the science disciplines and enable better data discovery.

NCEI helps NOAA meet the growing need for high value data by supporting projects like the Weather Research and Forecasting Innovation Act and the NOAA Blue Economy Initiative. These stewardship practices maximize the organization's investment in environmental research, converting scientific insights into dynamic, usable information that inform strategy and decision making in government, academia, and the private sector. NCEI data helps businesses and organizations across sectors operate more efficiently, safely, environmentally, and economically.

5.1 Science Background

NOAA, other federal agencies, and academic institutions collect acoustic data to support their research and management objectives. Theis data contain valuable information, even beyond their original collection purpose. To ensure the greatest value in the data and to follow NOAA policy on public access, the NOAA National Centers for Environmental Information (NCEI) Fisheries Acoustics Archives team archives water column sonar and passive acoustic data, and makes them publicly accessible. The archive team works directly with the data providers to complete the archive process. Once in the archive, all datasets are made publicly accessible through the archives' web-based map viewers and cloud buckets through the NOAA Big Data Program, along with an authoritative copy that remains within NOAA. The data can then be requested and delivered to the public for re-use.

Data that goes into the NCEI archive are gzipped and tarred, stored for at least 75 years, and are largely freely available to the public. The passive acoustic data are compressed using a free lossless audio codec for extra space saving prior to gzip and tar.

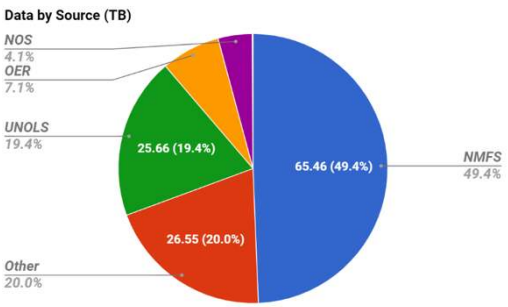

Figure 1: Water column sonar data archived. NMFS, NOS and OER are all NOAA.

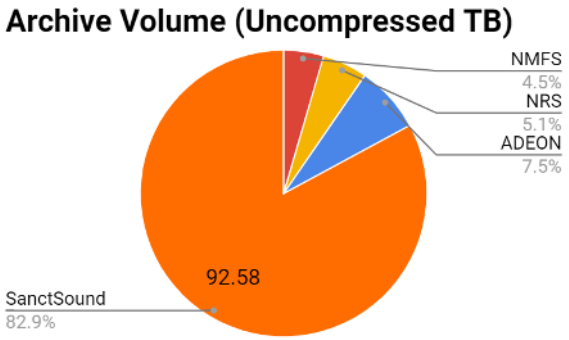

Figure 2: Passive acoustic data archived. NMFS, NRS, and SanctSound are all NOAA.

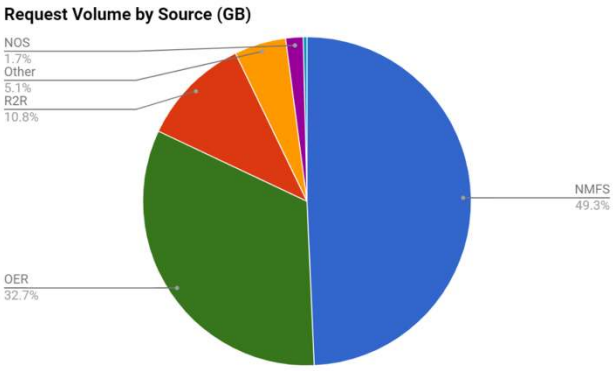

Figure 3: Water column sonar data requests by source.

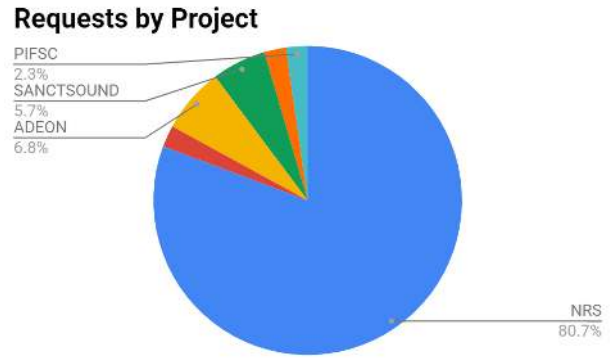

Figure 4: Passive acoustic datasets requested by project.

5.2 Process of Science

Data providers use custom packaging tools built by the Fisheries Acoustics Archives team to:

1. Describe the datasets
2. Copy the files
3. Create standardized data packages that include the standardized metadata, data files (often 100s GB to 10s TB), and ancillary data

These packages are then sent to the Fisheries Acoustics Archives team, typically on external hard drives. The data are verified and security scanned before being ingested by NOAA systems. The JSON-format metadata populate a domain-specific Oracle database that drives a web-based map viewer where the archived data are publicly available and discoverable. Requests for water column sonar data should be filled automatically by pulling from the archive and staging on FTP if the request is less than 100 GB - but this process has been broken since 2019. The Fisheries Acoustics Archives team directs customers to the cloud, or will pull data for them semi-manually. Requests for passive acoustic data are filled semi-manually if the customers don't go to the cloud where the data is hosted.

In collaboration with software developers, the Fisheries Acoustics Archives team is creating a cloud-based interactive web interface where users can explore the cloud-hosted water-column sonar data. Data is transformed from raw binary files to zarr format/zarr stores. Raw files are hosted on AWS S3 bucket.

Present-2 years
The Fisheries Acoustics Archives team budget over the next 3-4 years is increasing to support an additional passive acoustic data manager to support upcoming efforts to archive petabytes of passive acoustic data from NOAA, Bureau of Ocean Energy Management, U.S. Navy and National Park Service. Archival efforts will be focused on local pipeline and infrastructure.

Next 2-5 years
The Fisheries Acoustics Archives team will need increased throughput of the data pipeline, sufficient storage capacity on site at NCEI Boulder, and a pathway to upload the datasets to the cloud to support the incoming tsunami of data.

Beyond 5 Years
The Fisheries Acoustics Archives team aims to archive data directly to the cloud where it can be accessed directly and linked to associated oceanographic datasets also archived at NCEI or available elsewhere throughout NOAA. Even with the cloud, there will always be a need to maintain local resources to handle the data ingest portion of the workflow.

5.3 Collaborators

The main collaborator group is defined to be those that provide data sets. This group represents scientists throughout NOAA and other federal agencies, academic institutions, and to a small extent non-US institutions. Data is submitted to the archive through external hard drives. Collaborators could also include customers who request data from the archive. They vary in origin as well from U.S.-based agencies, universities, and businesses along with non-US universities. Internet capabilities vary widely from high speed users to a necessity for snail mail delivery as a result of very large data requests or difficulties using FTP.

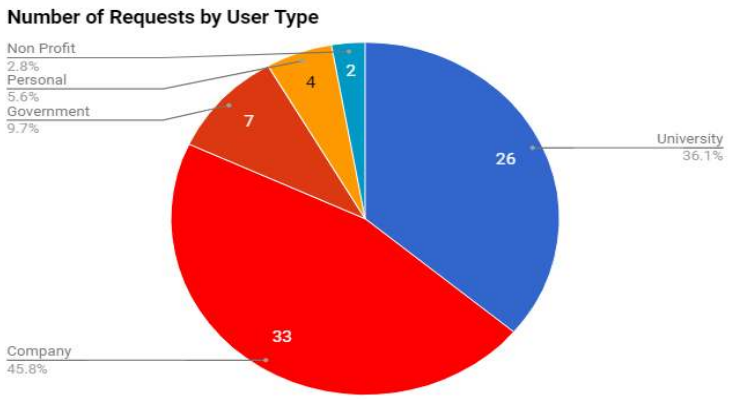

Figure 5: Passive acoustic data requests by user type (exclusive of those who download directly from GCP)

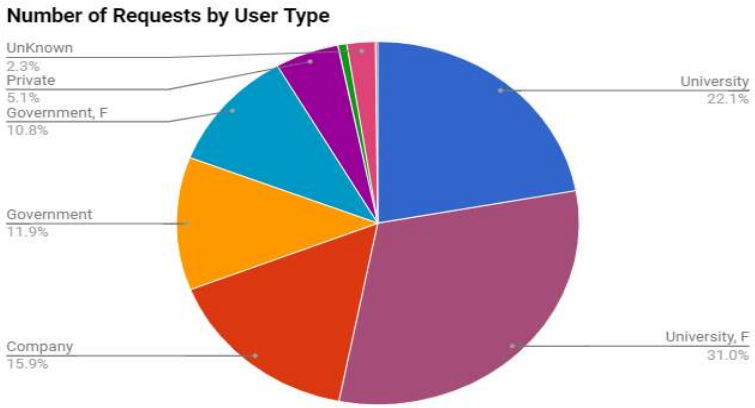

Figure 6: Water column sonar data requests by user type (exclusive of those who download directly from AWS). F next to a user type denotes non-U.S.

5.4 Instruments and Facilities

Data is collected by echosounders installed on ships and ROVs. The Fisheries Acoustics Archives team's focus is on the long term stewardship of the data in a digital archive to be made accessible now and in the future. The systems the Fisheries Acoustics Archives team interacts with for archiving data consist of servers and a tape storage system.

4.1 Passive Acoustic Data

Acoustic signals travel quickly and efficiently over long distances in the aquatic environment. Passive acoustic data is captures sound in underwater environments, enabling scientists to eavesdrop on the acoustic behavior of marine animals (e.g., whale song, fish chorusing, snapping shrimp), natural abiotic sounds (e.g., wind, earthquakes), and human generated sounds (e.g., cargo vessels). These data help to answer questions about marine habitats and provide important condition information to managers and policymakers.

- Total datasets archived: ***223***
- Average dataset size: ***420 GB***

- Total volume archived: ***29 TB*** compressed, ***94 TB*** uncompressed
- Files are FLAC compressed and gzipped on tape

4.2 Water Column Sonar Data

Water column sonar data, the acoustic backscatter from the near-surface to the seafloor, are used to assess physical and biological characteristics of the ocean including the spatial distribution of plankton, fish, methane seeps, and underwater oil plumes.

- Total datasets archived: ***1045***
- Average dataset size: ***134 GB***
- Max dataset size: ***16 TB***
- Min dataset size: ***6 KB***
- Total volume archived: ***140 TB***
- Files are gzipped on tape

4.3 Facilities

Archived data is stored on a tape system located on-premise. The hardware is rented and limited in space. The goal in the next 5 years is to transition to cloud storage or some hybrid storage solution. Currently not all data stored locally have been mirrored to the cloud, but will be migrated there as the technology pipelines are developed and implemented.

Present-2 Years

- Computer hardware
    - DELL Precision Desktop
    - 8 core Intel(R) Xeon(R) W-2123 CPU @ 3.60GHz
    - 32GB RAM
    - RHEL 7
- 2X vVirtual Servers
    - 2 core Intel Core Processor (Haswell, no TSX)
    - 16GB RAM
    - RHEL 7
- 2X vVirtual Servers (1 server dedicated to data cloud upload
    - 6 core Intel Core Processor (Haswell, no TSX)
    - 32GB RAM
    - RHEL 7

Next 2-5 Years

Planning for the middle time range is still in progress. NOAA NCEI and NESDIS are migrating resources to cloud-based solutions, but some processing activities will remain within the NOAA ecosystem (e.g., portions of the workflow that must deal with data ingest).

Beyond 5 Years

More adoption of cloud resources is likely, as well as trying to automate more portions of the overall workflow. Local resources will still be required for some ingest activities.

5.5 Non-Local Resources

The acoustic data submitted to NCEI is primarily done via external hard drives, due to large data volumes involved and the challenges is submitting electronically. However, some data providers routinely submit data via FTP, and more recently, the Fisheries Acoustics Archives team has received some data via the cloud. In the next 2-5 years the team plans to improve network speeds to better connect to data providers and to be able to connect across cloud buckets. Beyond 5 years, the team envisions a heavier focus on leveraging the cloud to receive data.

1) There will likely always be a need to maintain a 'local' presence, even with full cloud, due to the ingest via mailed hard drives from providers lacking the network bandwidth to load data to the cloud themselves.

2) All of the existing local workflows (ingest, egress) will need to be updated/converted to cloud. The existing ingest system was built with cloud migration in mind to lower the level of effort for migration to cloud-native services. This includes a highly modular design with communication between components handled via passing metadata/control data structures.

3) With the adoption of cloud for all forms of egress/download, users may not have to 'mail a hard drive' for data retrieval anymore. It is anticipated that a 'local' workflow will be in place where the team could extract data from the archive and send on a hard drive for those not able to access via the cloud

5.6 Software Infrastructure

The following represents the current state of the software infrastructure supporting this scientific workflow. The last 3 items (denoted in bold and italicized script) are packages that are supported and maintained by the higher-level NOAA organization, thus are not able to be modified.

Present-2 Years

- FTP, SCP, and RSYNC are used to move files between systems and storage locations though most data from data providers arrives on external USB drives.
- A collection of "Data packagers" that can be used by contributors to create the submission package.
- Common Package Ingest (CPI)
    - Automated workflow that Ingests data from external or internal drives, harvests metadata from files. tar/gzips files and moves them to StorNext tape archive, populated metadata database, publishes metadata, mints DOIs

    - Python 3.x system composed of custom classes and functions wired together to make workflows for water column sonar data and passive acoustic data.
    - Multithreaded and multiprocessing. Threads generally used with subprocess module to call processes on remote systems via ssh calls.
    - Run on Linux RHEL 7 desktop when ingesting from external media, RHEL 7 VM servers when ingesting data already on NFS mounts.
- Oracle 18
    - Custom schemas to hold metadata about datasets and files
    - Foundation for ESRI map services
    - Used by data managers to monitor and administer archives.
- ***ArcMap web map and map services***
- ***Custom NCEI code to add data to Quantum StorNext tape archive***
- ***Custom NCEI code to retrieve data from Quantum StorNext tape archive***

Next 2-5 Years
NCEI is working to migrate many servers (processing, storage, and sharing) to the cloud. In the 2-4 year timeframe the Fisheries Acoustics Archives team anticipates migrating CPI's water column sonar and passive acoustic data specific processing to run in cloud containers while interfacing with NCEI enterprise metadata, storage, and access systems.

Adoption of higher performance data mobility tools, such as GridFTP or Globus, can be considered if they become available from NOAA as a viable solution for data movement purposes.

Beyond 5 Years
This time frame has uncertainty due to the pending cloud migration plans. It is expected that much of the toolset will be converted, with some portions of the workflow still remaining local to support data ingest.

5.7 Support for Network and Data

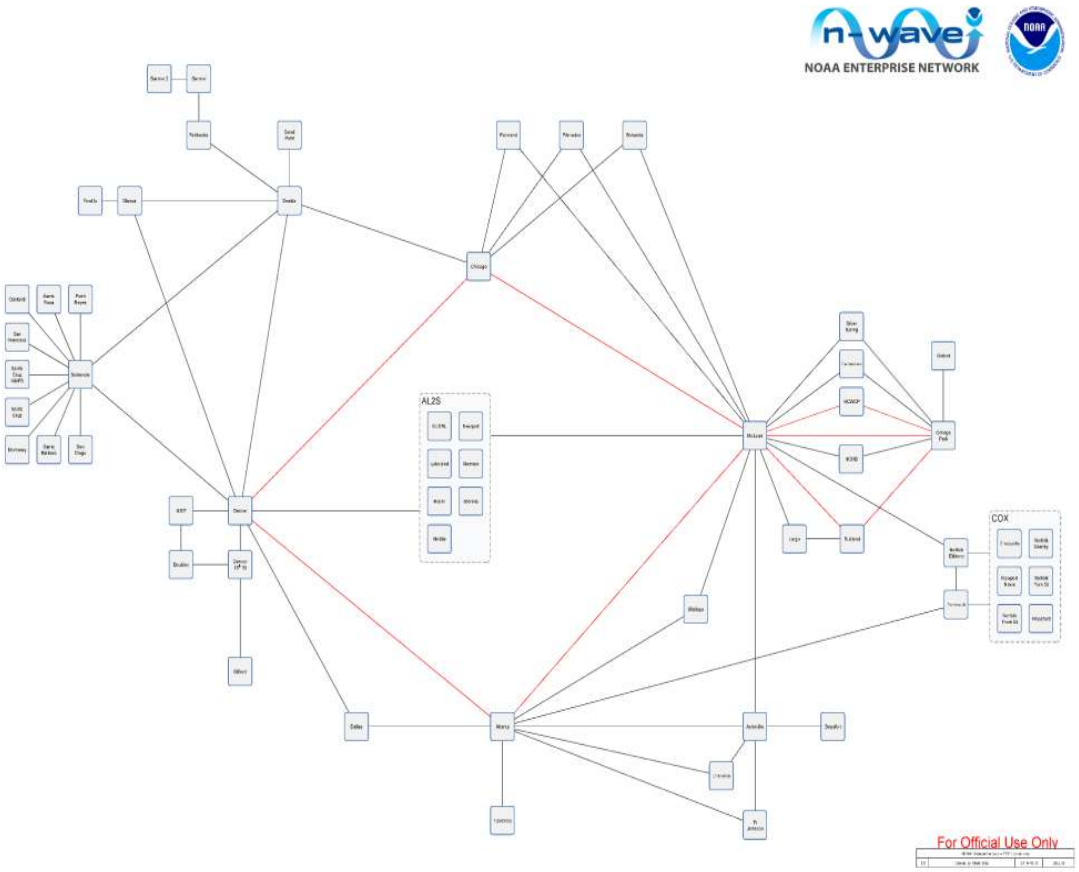

Figure 7: DSRC Network connectivity

The Fisheries Acoustic Archive is located under the US Department of Commerce, National Oceanographic and Atmospheric Administration (NOAA), and National Environmental and Satellite Data Information Service (NESDIS) for technology support. The N-Wave network provides wide area connectivity for the Fisheries Acoustic Archive, and access to both collaborators and users. Both of these downstream entities are subject to variable network connectivity, which complicates some portions of the ingest and download workflows. Major facilities (e.g., research universities) could be well connected and never report problems; individuals coming from secure locations (governmental, military) or smaller connections (home broadband) often resort to the use of slower mechanisms to share or request data.

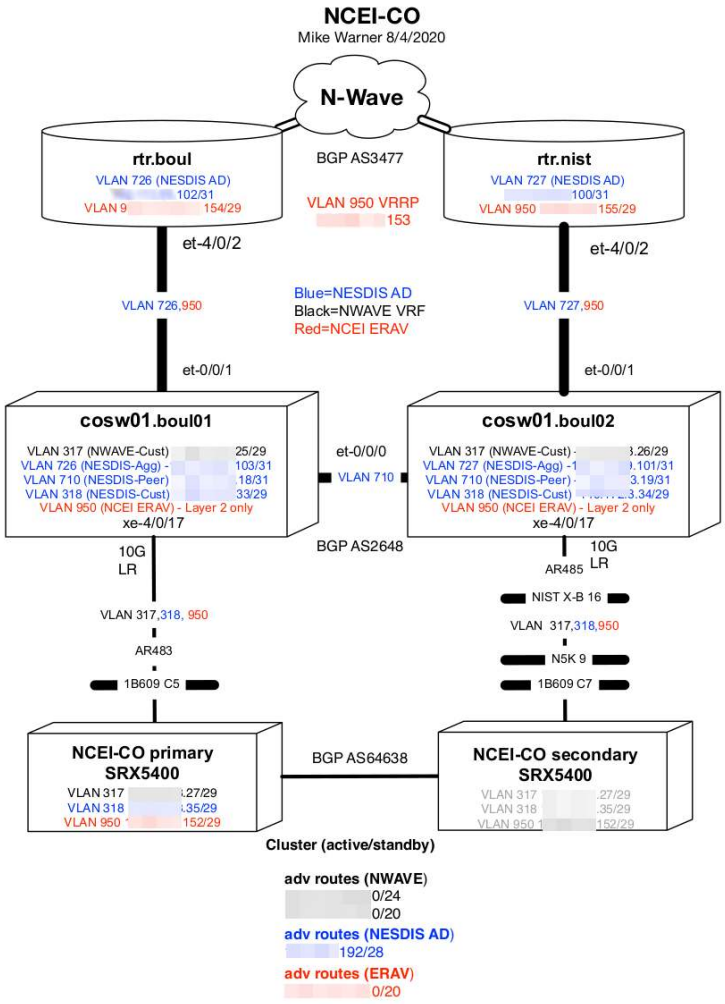

Figure 8: NCEI network connectivity

5.8 Cloud Services

The Fisheries Acoustics Archives team has access to cloud resources through the NOAA Big Data Program (BDP) which provides public access to NOAA's open data on commercial cloud platforms (AWS, GPC, IBM) through public-private partnerships to support managing and working with large datasets in the cloud. The team has explored and utilized GCP and AWS storage buckets as secondary data access points for the fisheries archives.

8.1 Storage

The Fisheries Acoustics Archives team currently hosts archived fisheries acoustic data in the cloud to serve as an additional data access point for the public as well as support processing large amounts of data in the cloud. The team has 105 TB of water column sonar data stored in a Amazon Web Services S3 bucket and ~23 TB of passive acoustic data in a Google Cloud Platform bucket. The data ingest pipeline was designed to be able to upload data to these cloud services as part of ingest.

However, as of 2021 this capability is turned off due to security constraints, and is undergoing a design process to create a more secure workflow.

8.2 Data discovery portal

As the Fisheries Acoustics Archives team loads more data into these storage buckets, there is a plan to develop a way for users to discover and access these data directly from online map viewers that currently deliver data from on-premise tape storage. Initially this will be a cloud-based version of an existing map-based discovery and filtering portal that facilitates access to the data in the cloud. In the next 2-3 years there will be an effort to incorporate data visualization tools as well as access to processed data products.

8.3 Analysis

In addition to storage, the Fisheries Acoustics Archives team plans to develop and deploy data processing and visualization tools in the cloud which can directly work with data in the storage buckets. This is not anticipated for 2-5 years.

8.4 Organization plans

NESDIS plans to be operational in the cloud within the next 5 years.

5.9 Resource Constraints

The following items are identified as risks to scientific progress for the Fisheries Acoustics Archives team. Addressing some of these may be solved with technology, or policy.

Present-2 Years

NCEI and its parent line office NESDIS are putting most of their resources into migrating to cloud-based archive and data processing. This focus, while good long term, is limiting resources available to support or upgrade on-prem systems. With over a petabyte of legacy fisheries acoustic data across NOAA and ~.5PB of new data being collected every year, the Fisheries Acoustics Archives team anticipates the on-prem limitations in data storage and processing resources will pose a significant challenge. For example, current on-prem tape storage for all NCEI data in Colorado is only 1.5PB. The limitation in support for on-prem systems, both hardware and personnel, significantly limits the resources that can be leveraged to maintain and operate the current on-prem ingest and data delivery systems.

Next 2-5 Years

As cloud resources are used more extensively for processing and storage, the impacts to contributors that submit data, and users that download it, will become more known. It is anticipated that some will experience issues as these two new workflows are adopted.

Beyond 5 Years

Audio and compression standards may change in future years, which could precipitate having to re-process the data archive. Such an operation would require significant storage and computational resources.

5.10 Outstanding Issues

Lastly, the Fisheries Acoustics Archives team has identified the following areas of friction to progress in their scientific process.

- IT resources are very limited with high turnover and the Fisheries Acoustics Archives team has no power to increase IT staff.
- The Fisheries Acoustics Archives team lacks a good way to track metrics due to having no ability to know the clicks on the map viewers and very little information coming from the cloud buckets. Both pertain to a need for better user metrics
- With respect to the eventual move to the cloud versus local, there is a potential change in workflow. The 'old' involves having to tar/compress files to support the requirements of local storage (namely, fewer but larger files). The 'new' would involve removing the tar/gzip step where more but smaller files would be archived. If the Fisheries Acoustics Archives team can archive in the cloud, but also need to maintain an on premises 'golden standard', then there may be a mismatch in files between the two repositories. This issue could be addressed within the database, which would keep track of the name/location for the file in the cloud and on premise.
- Requests for water column sonar data should be filled automatically by pulling from the archive and staging on FTP if the request is less than 100 GB - but this process has been broken since 2019.
- Archived data is stored on a tape system located on premise. The hardware is rented and limited in space. The goal in the next 5 years is to transition to cloud storage or some hybrid storage solution. The Fisheries Acoustics Archives team anticipates local storage and processing mechanisms will be needed to handle hard drive ingest and data request cases.
- The Fisheries Acoustics Archives team have practically zero control when it comes to IT resources and capacity, much less decision making and planning for the future. The majority of NCEI IT staff are contract employees, and are currently under-resourced and at great risk each time there is staff turnover due to knowledge loss and large workload left for remaining staff.

Appendix A – NOAA N-Wave Networking Diagram