Arecibo Data Preservation
Contact Point: Hans Addleman,
addlema@iu.edu
Audience: General
Last updated: October 28, 2021

ABOUT EPOC

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 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.

EPOC supports six main activities:

- **Roadside Assistance and Consultations** via a coordinated Operations Center to resolve network performance problems with end-to-end data transfers;
- **Application Deep Dives** to work more closely with application communities and 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;
- **Data Transfer Testing/ Data Mobility Exhibition** to check transfer times against known good end points;
- **Provision of managed services** via support through the IU GlobalNOC and our Network Partners;
- **Coordinated Training** to ensure effective use of network tools and science support.

Arecibo Collapse and Data Preservation

The Arecibo Observatory is located in Puerto Rico and over the last 50 years has collected more than 3 petabytes of astronomy data, which is primarily stored as 1 petabyte on hard drives and 2 petabytes on tape storage at the on-site data center. The observatory and data center are managed by staff at the University of Central Florida (UCF) , who approached EPOC and the Cyberinfrastructure Center of Excellence (CI COE) in August, 2020, to determine the best way to make a copy backup of this data.

This material is based upon work supported by the National Science Foundation under Grant No. 1826994

During this engagement, the facility suffered several failures which only served to emphasize the need to have a full backup of the data away from the observatory site. On November 6, 2020, one of the three primary support cables holding the equipment above the dish failed, which caused major damage to the dish itself. While options for repair were being considered, the entire Arecibo array collapsed on December 1, 2020. The on-site data center was not damaged in the collapse, but staff were removed during the damage assessment, which determined that the primary radio telescope and 305 meter dish were destroyed. Months later, only a reduced staff are working at the facility.

Finding a Data Store Location

A team was brought together to identify a location that could store the 3PB of data for the long term and the most expedient way to get the data to that site. Over the full engagement, more than 40 people were involved, including staff from EPOC, Arecibo, UCF, Texas Advanced Computer Center (TACC), CI COE, University of Puerto Rico, Globus, and the National Radio Astronomy Observatory (NRAO).

Over the engagement, several options were considered as places to home the replicated data set. The first option discussed was to move the data over the network to UCF for storage and processing. However, the campus data center at UCF did not have sufficient space for the full data set, so this was ruled out.

The second option considered was to move the data to a Microsoft Azure Cloud storage location. UCF had worked with Microsoft previously, and could easily extend their contracts to include an additional 3PB of storage. However, data transfer tests from Arecibo to Azure were only able to achieve 20Mbps, which means it would take over 40 years to transfer the complete 3PB data set. As a separate issue, it was determined that it would likely be cost prohibitive for researchers to download large data sets, due to the cost model used by most Cloud storage systems.

EPOC staff then reached out to the TACC Executive Director as well as the TACC Chief Information Security Officer. TACC is the first NSF Leadership Class Computing Center (LCCF) and currently supports more than 100 petabytes of dedicated user storage and hosts systems designed specifically for data analysis and computations. It was determined that they had both the infrastructure capable of storing data and staff support to put in place mechanisms for researchers to easily retrieve it in the future. So a home for the 50-years of data had been identified.

Moving the Data

However, there were still several problems to address in order for the data to reach TACC. In order to transfer the data over the network, the data would flow from Arecibo over a land lnk to a connection near the the University of Puerto Rico (UPR), where it would transit UPR's 10Gbps connection to AmLight, the NSF funded 100Gbps network the runs between South America and

Miami. From there, the data would transit the standard US research and education network to TACC, which consists of 10Gbps and 100Gbps components.

However, the connection between Arecibo and UPR was only a 1 Gbps commodity link, since the 10Gbps link that had existed previously was damaged in 2017 by Hurricane Maria. With a 1Gbps connection in the path, it would take almost 1 year to transfer the full 3 PB of data with ideal conditions including maximum line rate. In order to avoid this bottleneck, the decision was made to use portable network attached storage (NAS) devices, each holding 108TB of data, that would be driven to locations with 10Gbps network connections at UPR and also at Engine4, an on island commercial colocation facility. With a path that was 10Gbps enabled, the maximum line rate scenario meant that it would take only 30 days to transfer 3PB of data. Real world results were longer and varied depending on the time it took to load the NAS at Arecibo, and the drive time to the 10G connected sites.

Speed	Time to transfer 3PB
1Gbps connection - max for transfers from Arecibo to another site	305 days
10Gbps connection - max from UPR to another site	30 days
20Mbps - observed from Arecibo to Azure	40 years

Table 1. The estimated time to transfer 3PB of data under different conditions.

Two additional bottlenecks were identified and addressed. The first was that the NAS needed to be tuned to allow for faster, long distant data transfers. When the NAS were first deployed, they could only achieve transfer rates of near 200Mbps, mitigating the benefit of their use. EPOC engineers produced a detailed guide to optimizing the specific NAS, a Synology DS1819+, that allowed for an almost 10x increase in transfer speeds, achieving just over 3Gbps in a lab environment. And because there were two sites that could transfer data from the portable NAS, this rate could be doubled. Additional speedup was achieved by using the Global file transfer tool and enabling multiple data streams. The resulting real world file transfer rate from both Engine4 and the UPR data center was close to 1.5Gbps.

Speed	Time to transfer 3PB
Untuned NAS - 200Mbps observed	1,500 days
EPOC tuned NAS - 1.5Gbps observed	182 days

Table 2. The time to transfer 3PB of data from Arecibo to TACC, assuming the NAS is the bottleneck, before and after tuning the NAS transfer rates.

This material is based upon work supported by the National Science Foundation under Grant No. 1826994

Conclusion

In June 2021, Arecibo staff reported that their primary spinning media, over 1 Petabyte, had been fully transferred to TACC. Additional effort continues to copy the full data set still on tape storage, with a predicted completion time of December 2021.

Associated Press Releases:

- https://www.tacc.utexas.edu/-/continuing-arecibo-s-legacy
- https://itnews.iu.edu/articles/2021/Continuing-Arecibos-legacy-.php
- https://lightbytes.es.net/2021/04/21/arecibo-data-recovery-behind-the-scenes-with-jason-zurawski/

This material is based upon work supported by the National Science Foundation under Grant No. 1826994