Putting the Cloud to Work for Seismology

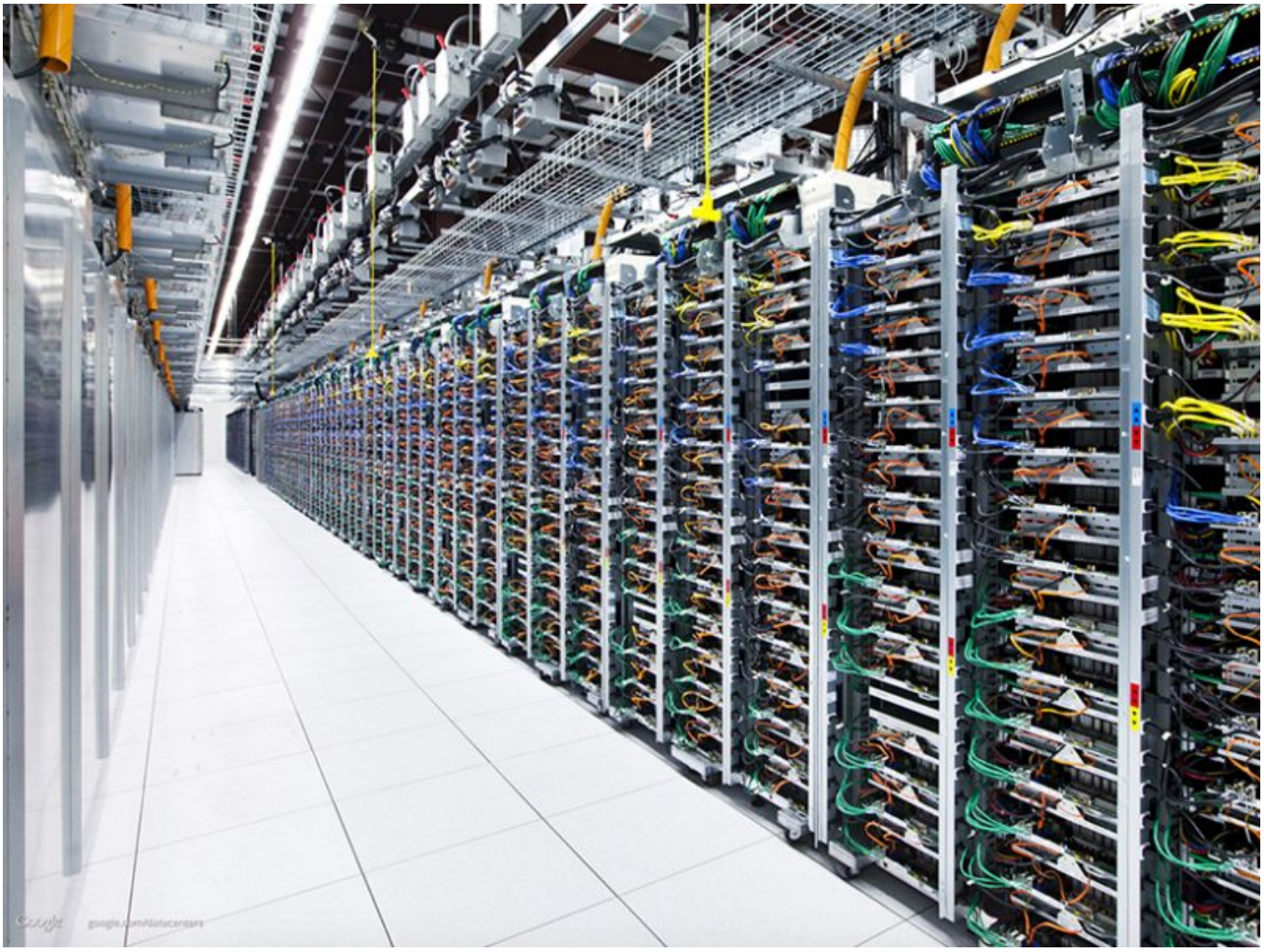
MacCarthy, Jonathan K.
Marcillo, Omar Eduardo
Trabant, Chad

To be published in: Eos

DOI to publisher's version: 10.1029/2019EO119741

Putting the Cloud to Work for Seismology

The cloud infrastructure developed in the business community has made access to cluster computing possible for even the smallest research groups, enabling new kinds of research workflows in geophysics.



Data centers such as those built by Google offer a number of cloud services, like on-demand computational nodes of various hardware specifications and operating systems. Credit: Google

By Jonathan MacCarthy, Omar Marcillo, and Chad Trabant 5 April 2019

By now, many researchers have heard about “the cloud,” the distributed network of data centers and computing facilities where, for a fee, online photos are stored and start-ups scale

up their business with demand. Far fewer researchers, however, have had the opportunity or even considered using the cloud to scale their research, yet cloud-based approaches are emerging as a new and cost-effective approach to accelerate large computations and significantly reduce time to science.

Until recently, the on-demand computing infrastructure provided by companies like Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure had been used almost exclusively by businesses or government (https://aws.amazon.com/solutions/case-studies/healthcare-gov/), but as the tools for using computer clusters in the cloud become easier to use, it is more feasible for scientists (https://eos.org/articles/tracking-global-change-with-a-cloud-based-living-atlas) to use them for data-intensive or computationally intensive (https://eos.org/research-spotlights/managing-radio-traffic-jams-with-the-cloud) research. This accessibility is particularly beneficial for applications that don't require the fast intercomputer communication already present in high-performance computing clusters.

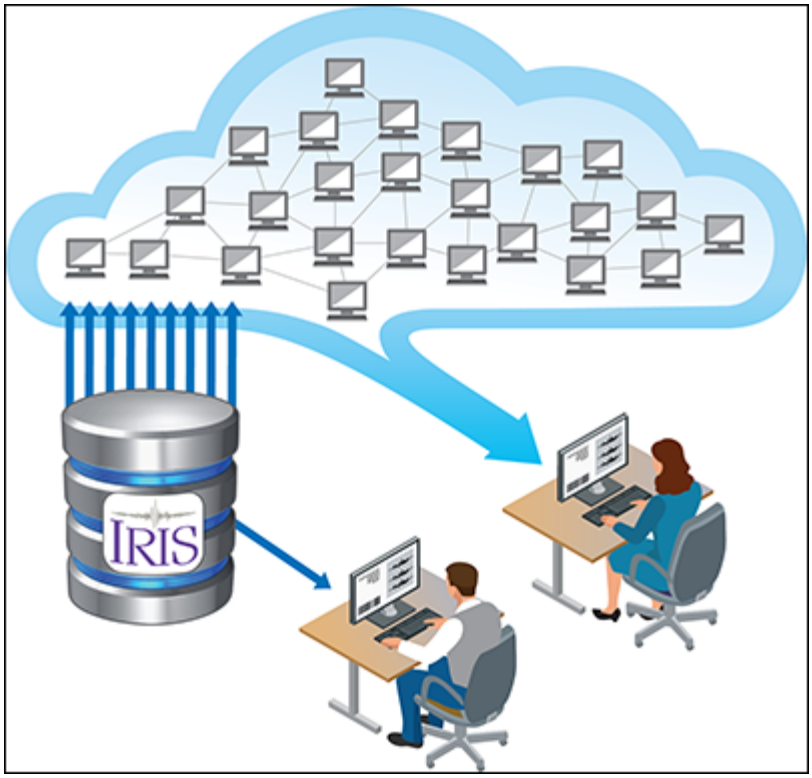
We rented 50 computing nodes in the cloud and configured a cluster for analysis of publicly available continuous seismic data.
This approach has been embraced by some research groups already facing problems of scale, such as the members of the Pangeo (http://pangeo.io/) collaboration [*Abernathey et al.* (https://doi.org/10.6084/m9.figshare.5361094.v1), 2017], who use AWS and GCP to analyze large quantities of gridded multidimensional data. Research groups such as these are lighting the way for the use of cloud-native tools like Docker and Kubernetes (which facilitate construction and orchestration of self-contained software environments on distributed systems) in the atmospheric sciences, but these tools have not yet gained adoption in the seismic community.

Our research seeks to change that. We rented 50 computing nodes in the cloud and configured a cluster for analysis of publicly available continuous seismic data. Initial results show our cluster (compared to a single computer's performance) could accelerate our analysis almost 2 orders of magnitude for as little as $100 per day.

Streaming Seismoacoustic Analysis

Although most seismological research problems still fit comfortably on a laptop, many require much larger resources. Analyses where large volumes of continuous waveform data are scanned or monitored to detect signals [*Bergen and Beroza (https://doi.org/10.1093/gji/ggy100)*, 2018; *Li et al. (https://doi.org/10.1038/s41598-018-19728-w)*, 2018] or data quality problems [*Casey et al.*

(https://doi.org/10.1785/0220170191), 2018] or used to characterize persistent background noise [*Marcillo and Carmichael* (https://doi.org/10.1785/0220170271), 2018] are examples of "trivially parallel" workloads that could benefit from adding more computing nodes.



On-demand clusters in the cloud allow researchers to overcome single-machine limitations and to quickly turn excess service capacity at data centers like the IRIS DMC into research products. Credit: LANL

Traditionally, the large volumes of data required for these survey-style analyses are downloaded from a central repository or data center such as the Incorporated Research Institutions for Seismology (https://www.iris.edu/hq/) (IRIS) Data Management Center (https://ds.iris.edu/ds/nodes/dmc/) (DMC) before processing, which can take days, weeks, or longer. These data can also become unwieldy, as the responsibility for storing, indexing, querying, and serving the data shifts from the data center to the researcher.

To sidestep these challenges, we—a group of researchers from Los Alamos National Laboratory, in collaboration with staff of the DMC—are conducting an experiment to test the feasibility of using AWS Elastic Compute Cloud (EC2) to support streaming, on-the-fly, seismic analysis, where data are requested from the data center as needed and not stored. The goal is to assess technical requirements, expose limitations and challenges, and outline the costs and benefits of

doing geophysical research using a streaming workflow in the cloud.

A Large-Scale Harmonic Tonal Noise Survey

We are testing our cloud-based workflow on a large-scale seismoacoustic background noise analysis. Our specific application is the detection of discrete spectral components of seismic noise generated by large machinery (harmonic tonal noise; see *Marcillo and Carmichael (https://doi.org/10.1785/0220170271)* [2018]). The characterization of the spatiotemporal structure of noise in seismic data is usually an onerous process. But we decided to see whether computing in the cloud could reduce the time it takes to both acquire and process the data.

We analyzed 13 years of continuous seismic data from two stations in Texas and detected noise from multiple wind farms within 100–150 kilometers (small inset map of Texas in Figure 1a). It took up to 1 hour of processing time per station per year to retrieve the data from the DMC repository using their International Federation of Digital Seismograph Networks Web Services interface (http://service.iris.edu/fdsnws/), apply our detector, and write the results on a single workstation. This non-cloud-based approach, in total, took over 30 hours.

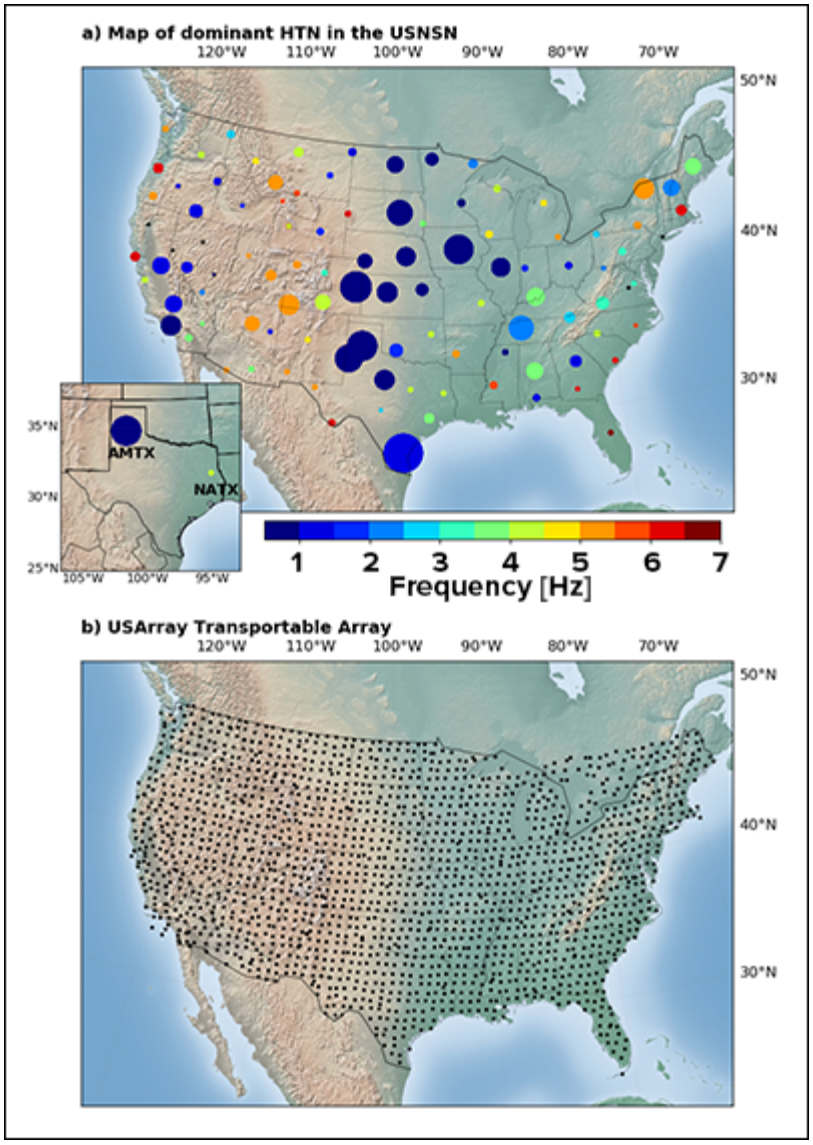


Fig. 1. Scaling up a seismic noise analysis. (a) A streaming harmonic tonal noise (HTN) survey was run on 100 stations of the USNSN over 10 years and run in the cloud, resulting in a processing rate of 1–2 minutes per station per year. Circles are located at seismic stations, colored according to the frequency content of the HTN detections, and scaled in proportion to the number of detections. This analysis is the natural progression of an initial two-station survey with stations in Texas (bottom left inset) conducted on a desktop computer, which had a processing rate of 1 hour per station per year. (b) A full survey of 1,690 stations of the

contiguous Transportable Array would not be feasible without a new architecture. Credit: Omar Marcillo/LANL

We expanded our analysis to cover a larger geographical area using the U.S. National Seismic Network (USNSN) [*Albuquerque Seismological Laboratory (https://doi.org/10.7914/SN/US)*, 1990], which has roughly 100 stations, each with approximately 10 years of continuous data. Processing the entire USNSN would take nearly 41 days on the basis of our single-station processing time.

To cut this time interval, we moved our analysis to a small 20-node cluster in the cloud and processed at a rate of 1–2 minutes per station per year. The entire survey was finished in 30 hours, at a cost of just $30 (Figure 1a). The reduction in processing time is even more impressive when you consider that our 20-node cluster can be turned into a 200-node one with almost the same level of effort.

We are now deploying our analysis over the 1,690 stations of the contiguous USArray Transportable Array using a medium-sized cluster of 50–100 nodes (Figure 1b). Initial results suggest that processing times can be further reduced to nearly 20 seconds per station per year using just 50 nodes, but with worse performance scaling at 100 nodes. This degradation is likely due to limitations or controls on the DMC's service capacity.

Engaging Data Centers and Building for Future Capacity

The ability to access data on demand from cloud systems represents a significant change for the data center, where usage has traditionally been limited by a researcher's ability to orchestrate requests for and manage large data volumes. With compute, storage, and bandwidth resources larger than many data centers, cloud systems make it easy for a single researcher to significantly affect a data center's ability to service requests. Furthermore, not all centers are well protected against an unexpected onslaught of requests, which could cause outages. For these reasons it will be important for researchers to follow posted usage guidelines and possibly consult with these data centers before undertaking a large data collection exercise, such as facilitated by cloud systems.

In anticipation of increasing data requests, the IRIS DMC is completing its part of the GeoSciCloud (https://www.earthcube.org/group/geoscicloud-deploying-multi-facility-cyberinfrastructure-

commercial-private-cloud-based-systems) project, funded by the National Science Foundation's (NSF) EarthCube (https://www.earthcube.org/) program, in which the concept of running the DMC's services directly in a cloud resource is being explored. Hosting a data repository and related access services in a cloud system has two potentially significant benefits: (1) allowing data access capacity to scale as needed, supporting many more users and/or concurrent requests than currently possible, and (2) hosting the data adjacent to large computing capacity. Initial results are very positive, with the DMC demonstrating that its core data services can run in the cloud and support a much higher data access load than possible with the current data center.

Toward Cloud-Native Research

Although offering many benefits, performing research using clusters in the cloud has drawbacks similar to those of traditional cluster computing, as well as new ones. In addition to users having to learn more specialized tools and skills, new issues such as data and code ownership and security are introduced, as the researcher assumes the role of cluster manager. These must be considered when performing research in the cloud.

The final component in large-scale seismic research, cluster computing, is now available in the commercial cloud and ready for scientific use.

In the seismology community, research instrumentation centers like the Portable Array Seismic Studies of the Continental Lithosphere (PASSCAL (https://www.passcal.nmt.edu/)) Instrument Center have helped to generate large volumes of data, and data centers like the DMC have expanded access to it. The final component in large-scale seismic research, cluster computing, is now available in the commercial cloud and ready for scientific use. These advancements, together with initiatives like the NSF's Exploring Clouds for Acceleration of Science (E-CAS (https://www.nsf.gov/news/news_summ.jsp?cntn_id=297193&org=NSF&from=news)) program, are making large-scale research possible for anyone in the age of data.

Acknowledgments

The authors would like to thank the Los Alamos Information Science and Technology Institute and the Office of the Chief Information Officer for initial support of this work. We also thank Robert Weekly, Robert Casey, and Inge Watson at the IRIS DMC for their collaboration during the experiment. This article has been authored by Triad National Security, LLC under contract 89233218CNA000001 with the U.S. Department of Energy and is approved for unlimited release as LA-UR-18-30252. The IRIS DMC's core functions are supported by NSF award

EAR-1724509, and the cloud deployment project (GeoSciCloud) is supported by NSF's EarthCube program ICER-1639719.

References

Abernathey, R., et al. (2017), Pangeo: An open source big data climate science platform, figshare, https://doi.org/10.6084/m9.figshare.5361094.v1 (https://doi.org/10.6084/m9.figshare.5361094.v1).

Albuquerque Seismological Laboratory (1990), United States National Seismic Network, International Federation of Digital Seismograph Networks, Dataset/Seismic Network of Entry, Int. Fed. of Digital Seismograph Networks, https://doi.org/10.7914/SN/US (https://doi.org/10.7914/SN/US).

Bergen, K. J., and G. C. Beroza (2018), Detecting earthquakes over a seismic network using single-station similarity measures, ***Geophys. J. Int., 213***(3), 1,984–1,998, https://doi.org/10.1093/gji/ggy100 (https://doi.org/10.1093/gji/ggy100).

Casey, R., et al. (2018), Assuring the quality of IRIS data with MUSTANG, ***Seismol. Res. Lett., 89***(2A), 630–639, https://doi.org/10.1785/0220170191 (https://doi.org/10.1785/0220170191).

Li, Z., et al. (2018), High-resolution seismic event detection using local similarity for large-N arrays, ***Sci. Rep., 8***(1), 1646, https://doi.org/10.1038/s41598-018-19728-w (https://doi.org/10.1038/s41598-018-19728-w).

Marcillo, O. E., and J. Carmichael (2018), The detection of wind-turbine noise in seismic records, ***Seismol. Res. Lett., 89***(5), 1,826–1,837, https://doi.org/10.1785/0220170271 (https://doi.org/10.1785/0220170271).

Author Information

Jonathan MacCarthy (jkmacc@lanl.gov (mailto:jkmacc@lanl.gov); @jkmacc (https://twitter.com/@jkmacc)) and Omar Marcillo, Los Alamos National Laboratory, N. M.; and Chad Trabant, Data Management Center, Incorporated Research Institutions for Seismology, Seattle, Wash.

Citation: MacCarthy, J., O. Marcillo, and C. Trabant (2019), Putting the cloud to work for seismology, *Eos, 100,* https://doi.org/10.1029/2019EO119741. Published on 05 April 2019.