Cloud-Optimized ASDF-H5 for Seismology

Yiyu Ni[1], Joseph-Paul A Swinski[1], and Marine Denolle[1]

[1]Affiliation not available

May 2, 2023

Cloud-Optimized ASDF-H5 for Seismology

Yiyu Ni[1] (niyiyu@uw.edu), Joe-Paul Swinski[2], Marine Denolle[1]

[1]Earth and Space Sciences, University of Washington
[2]Goddard Space Flight Center, NASA

[S15E-0297]

Background

Recent seismology analysis (e.g., ambient noise cross-correlation, full waveform inversion, distributed acoustic sensing, machine learning) necessitates powerful computation platforms and file management techniques. Traditional I/O systems, data exchange, and data processing workflows meet significant challenges to the community when the size of data grows to ~TB/PB.

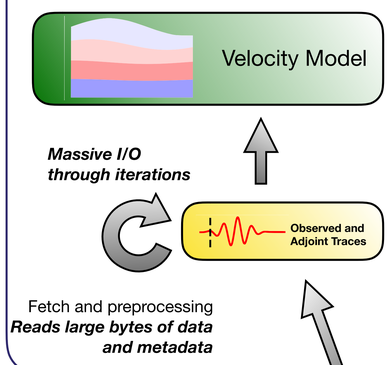
Data processing workflows have been developed for scalability on HPC or cloud platforms, e.g., SPECFEM, PyTorch, and SeisNoise.jl. But seismic data formats are not ready as hundreds of millions of standard data files (miniSEED, SAC, SEG-Y) will overload the file system and dramatically decrease job efficiency. HDF5-based ASDF is proposed to settle this problem and has been well tested on HPC platform (*Krischer et al., 2016*).

Motivation

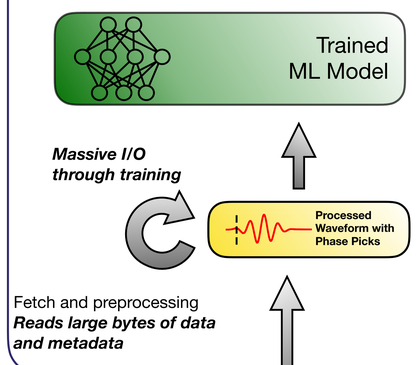
Native HDF5 C library is not optimized for the cloud platform. The way to read HDF5 file is different from reading a linearly stored file, as extra reads are required to go through HDF5 architecture and reach the position of the target dataset: this is fast with a file handle on a local file system but is slow on **cloud object storage of high latency** (*Guimarães et al., 2021*). The full advantage of cloud computing with HDF5 has not been exploited, and a reader designed for cloud seismology research is required.

Event-based Seismology

Full Waveform Inversion

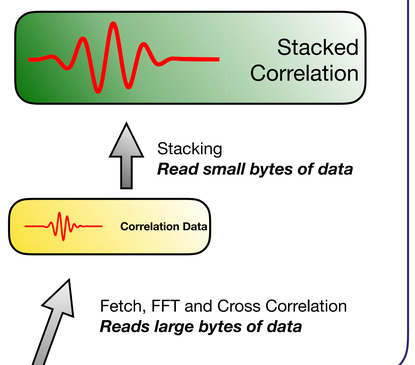


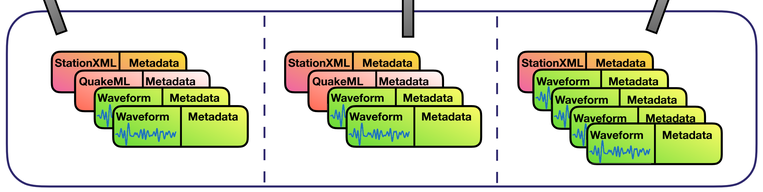
Machine Learning Seismology



Continuous-based Seismology

LargeN, DAS and Ambient Noise Seismology

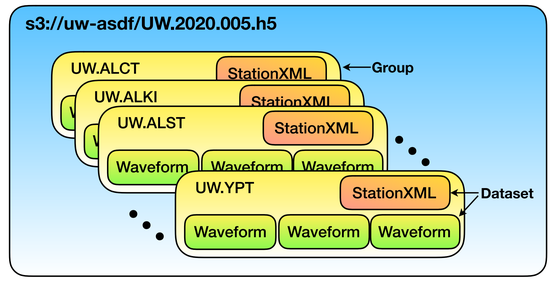




HPC File Server/Object Store

Performance Test

- ASDF (8.7 GB) on S3 contains 294 stations one-day mseed traces (UW.2020.005, 10.1 GB).
- Traces of 200 stations are iteratively read and loaded into memory directly from the cloud through CloudPyASDF.
- Read with CloudPyASDF, s3cp+h5py and s3fs+h5py, with different types of EC2.

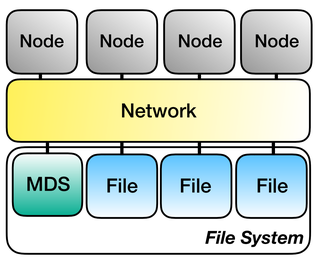


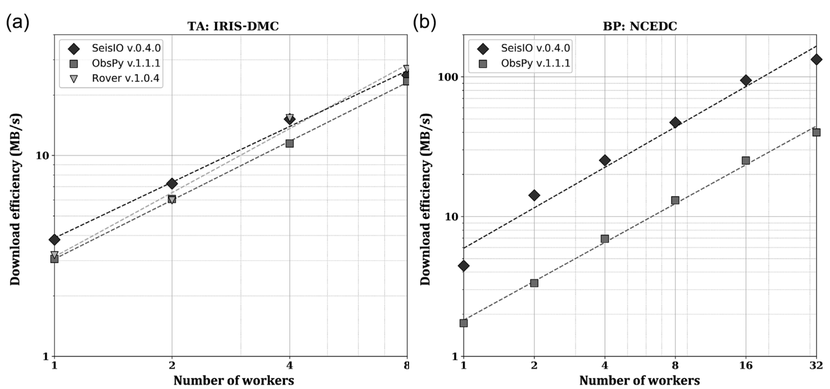
Reference

- Jones, J. P., Okubo, K., Clements, T., & Denolle, M. A. (2020). SeisIO: A Fast, Efficient Geophysical Data Architecture for the Julia Language. Seismological Research Letters, 91(4), 2368–2377. https://doi.org/10.1785/0220190295. https://github.com/jpjones76/SeisIO.jl
- Krischer, L., Smith, J., Lei, W., Lefebvre, M., Ruan, Y., de Andrade, E. S., Podhorszki, N., Bozdağ, E., & Tromp, J. (2016). An Adaptable Seismic Data Format. Geophysical Journal International, 207(2), 1003–1011. https://doi.org/10.1093/gji/ggw319
- Guimarães, A., Lacalle, L., Rodamilans, C. B., & Borin, E. (2021). High-performance IO for seismic processing on the cloud. Concurrency and Computation: Practice and Experience. https://doi.org/10.1002/cpe.6250
- University of Washington. (1963). Pacific Northwest Seismic Network—University of Washington [SEED data]. International Federation of Digital Seismograph Networks. https://doi.org/10.7914/SN/UW
- Clements & Schmidt personal communication
- CloudPyASDF available at https://github.com/niyiyu/CloudPyASDF

HPC Platform

- High computing performance (queuing needed).
- User-organized data on a local file system.
- Slow network connection, which also limits open data access. Complete regional data archive generally takes months to build.
- Inconvenient sharing of data products.

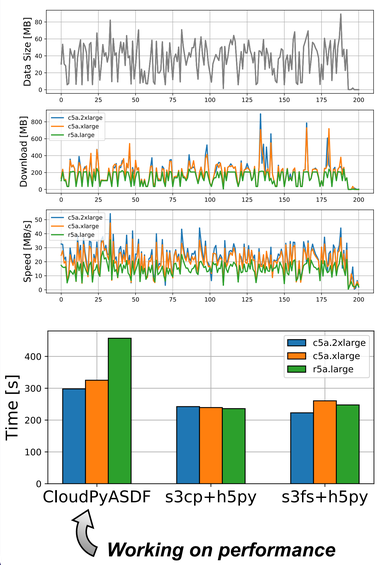




Jones et al., 2020

CloudPyASDF

Using a **HDF5 cloud-optimized read-only library (H5coro)**, we developed a new python module for direct and multi-thread ASDF dataset/attributes accessing from AWS S3 bucket to EC2. **No local cache file or metadata repository is required** for file access.



Working on performance

Results

- **Highly consistent reading** efficiency across all datasets.
- High overhead in network throughput.
- **10-30 MB/s** trace downloading and loading speed. Varies with instances.

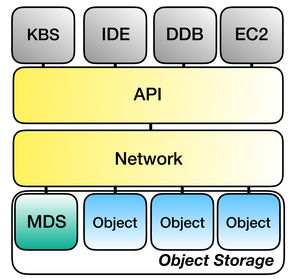
What is more:

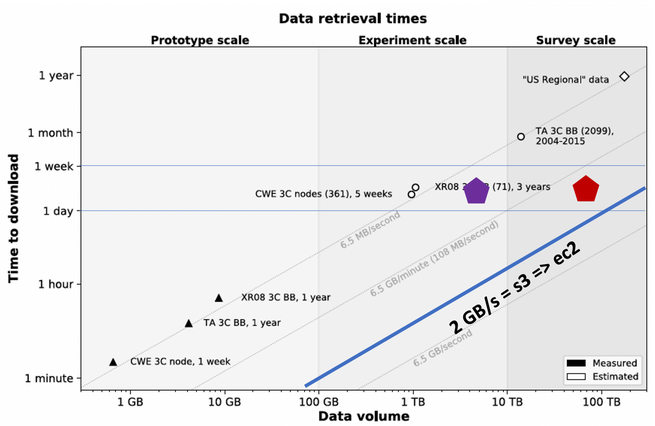
- ASDF is 15% smaller than the local mseed archive, with no compressing applied.

Instance	vCPU	Memory	Network	Price/h
c5a.2xlarge	8	16 GiB	Up to 10 Gigabit	$0.308
c5a.xlarge	4	8 GiB		$0.154
r5a.large	2	16 Gib		$0.113

Cloud Platform

- Flexible data organization on cloud object storage within the cloud ecosystem.
- Scalability of data storing, downloading, and processing within the cloud front.
- Fast and direct access to public storage on the cloud, e.g., SCSN, PoroTomo, IRIS PASSCAL.
- Convenient sharing of data products among the community.





Clements & Schmidt personal communication