Executive Summary

One of the key problems in open science is unlocking efficient and equitable access to the massive amounts of data in archives, under production, and planned for future missions. The Zarr data specification and open-source implementations can solve this problem for multidimensional datasets. The excitement and expectations for Zarr has grown tremendously over the past year. For example, the European Space Agency proposed distributing Sentinel data products as Zarr, representing an likely investment of tens of millions of dollars, with other agencies considering similar approaches. The Zarr community can accelerate progress and adoption with a small investment to support a unique week-long meeting of Zarr developers and adopters to unblock core issues and facilitate adoption.

State of the Field

Many disciplines ranging from the Earth sciences to astronomy, biomedicine, plasma physics, and more are producing hundreds of petabytes of multidimensional data every year. At this scale, gaining impactful insights from the data requires cloud-optimized storage, unfettered and global access, and highly efficient computing. The Zarr data specification was designed with these patterns in mind, providing a simple, transparent, open, and community-driven solution for high-throughput distributed I/O on many different storage systems. Many groups including large international organizations such as the European Space Agency, technology giants like Google and Microsoft, non-profit organizations like CarbonPlan, and small companies like Scalable Minds are adopting Zarr for these benefits. However, unlocking the true potential of Zarr for scalable and accessible open science requires the complete implementation, extension, and adoption of the next generation of the Zarr specification - version 3.0. Our proposed meeting will bring together a core group of leaders, implementers, and adopters for the first time to meet this challenge.

Community Impact

This meeting will enable scalable open-science on multi-dimensional data by unblocking the implementation of core features required by several disciplines and ensuring that adopters fully benefit from improvements provided by Zarr version 3.

We have identified the following technical priorities to maximize community impact:

- Variable chunk grid: The requirement that all chunks in an array are equal size limits Zarr's impact. However, Zarr format 3 defines a mechanism for extending the specification to support variable chunk grids. Through this meeting, the Zarr leaders and developers will

settle on an extension definition for the variable chunk grid extension and prototype its implementation.
- Sharding: The addition of sharding in Zarr format 3 allows high-throughput distributed I/O while minimizing the number of files associated with a Zarr store. Through this meeting, the Zarr developers will finalize the design for the sharding implementation and demonstrate its use for the community.
- Codecs: It is vital for the Zarr community to maintain a shared set of codec definitions. This meeting will enable high-bandwidth codec development, with a key goal being the stabilization of codec specifications for common image compression routines.
- Cross-language support: Implementers from across languages have never before gathered to ensure Zarr is widely supported beyond Python. This meeting will bring together developers from each implementation to ensure cross-compatibility and feature parity.

We have identified the following priorities for maximizing the impact for Zarr adopters:

- Roundtable with Zarr leaders and implementors for an unprecedented opportunity for Zarr adopters to meet with a wide range of implementors.
- Virtual Zarr guidance to unlock cloud native data access for organizations with massive quantities of archival data.
- Migration guidance to walk potential Zarr V3 adopters through the process and address any gaps in the implementations.

Early List of Participants

The meeting will be organized by Dr. Max Jones, Dr. Joe Hamman, and Dr. Davis Bennett. We will invite Zarr Steering Council members, Zarr Implementations Council members, Zarr Python Core Developers, and GDAL Core developers as core developers. We will invite adopter representatives from ESA, Copernicus, CNES, EUMETSAT, NASA, USGS, NVIDIA, GAD Climate Prediction and Applications Centre, Deutches Klimarechenzentrum, Google Earth Engine, ESRI, Helmholtz Zentrum, Janelia Research Campus, Deutsches Krebsforschungszentrum, European Molecular Biology Laboratory, European Bioinformatics Institute, National Institutes of Health, Allen Institutes of Cell Science, RIKEN and CZI BioHub. We will aim for ~25 participants.