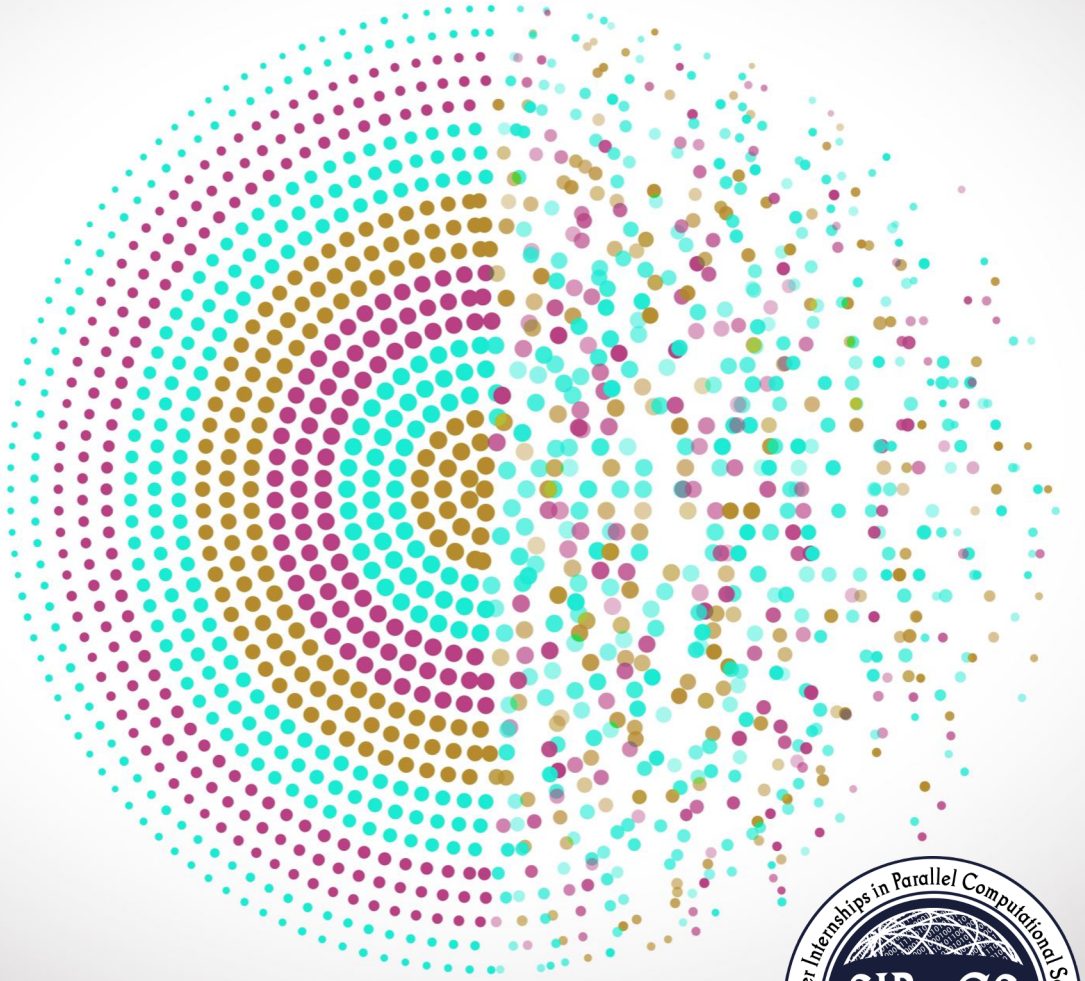
Introducing: Kerchunk

Cloud-performant reading of NetCDF4/HDF5/Grib2 using the Zarr library

Lucas Sterzinger - *UC Davis*

Martin Durant - *Anaconda, Inc*
Rich Signell - *USGS*
Chelle Gentemann - *Farallon Institute*
Kevin Paul - *NCAR*
Julia Kent - *NCAR*



NCAR | NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Summer Internships in Parallel Computational Science
SIParCS
2017
BOULDER • CO

The data is growing, and we can't keep up!

- Data is becoming too big to download!
- Solution: Move data to cloud
  - Great for storage-adjacent computing
- However, many datasets still in native NetCDF4/HDF5/GRIB format
  - Other formats (e.g. Zarr) are more optimized for cloud but require data conversion/duplication

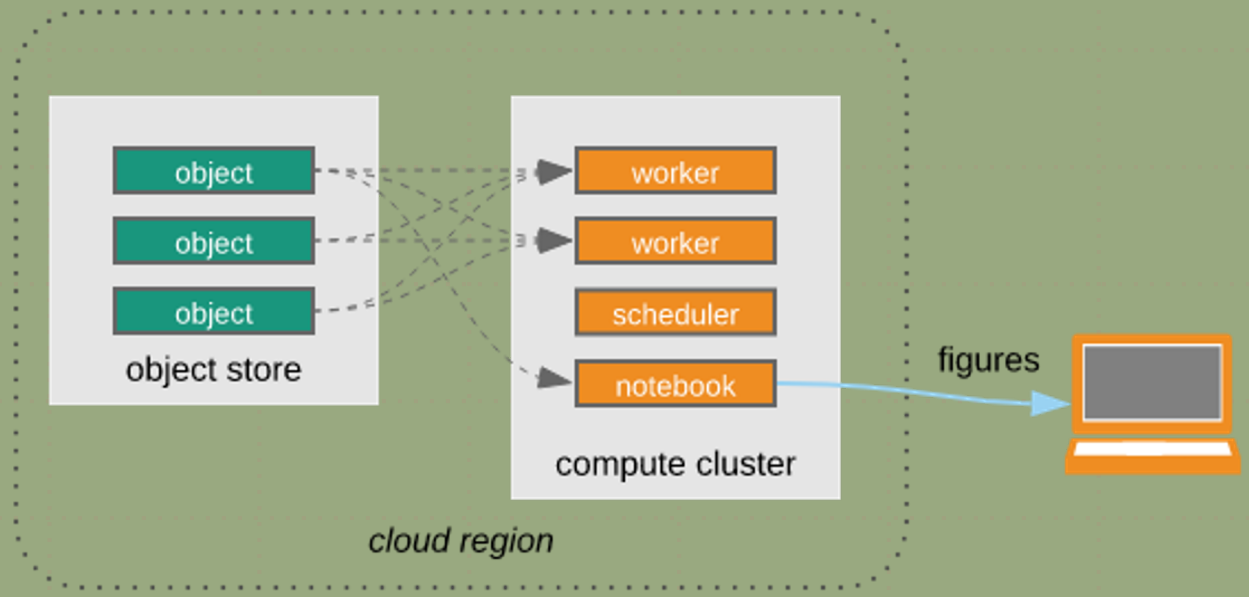
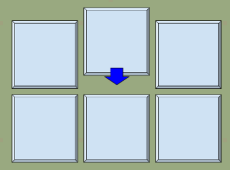


Image courtesy of Ryan Abernathey

Introducing:  Kerchunk

What?

- Kerchunk allows for simple and fast access to common data formats (e.g. NetCDF/HDF/GRIB) via fsspec's ReferenceFileSystem
- Cloud-optimized access without need to convert to more cloud-friendly data formats (e.g. Zarr)

How?

- Consolidate metadata into Zarr-spec JSONs

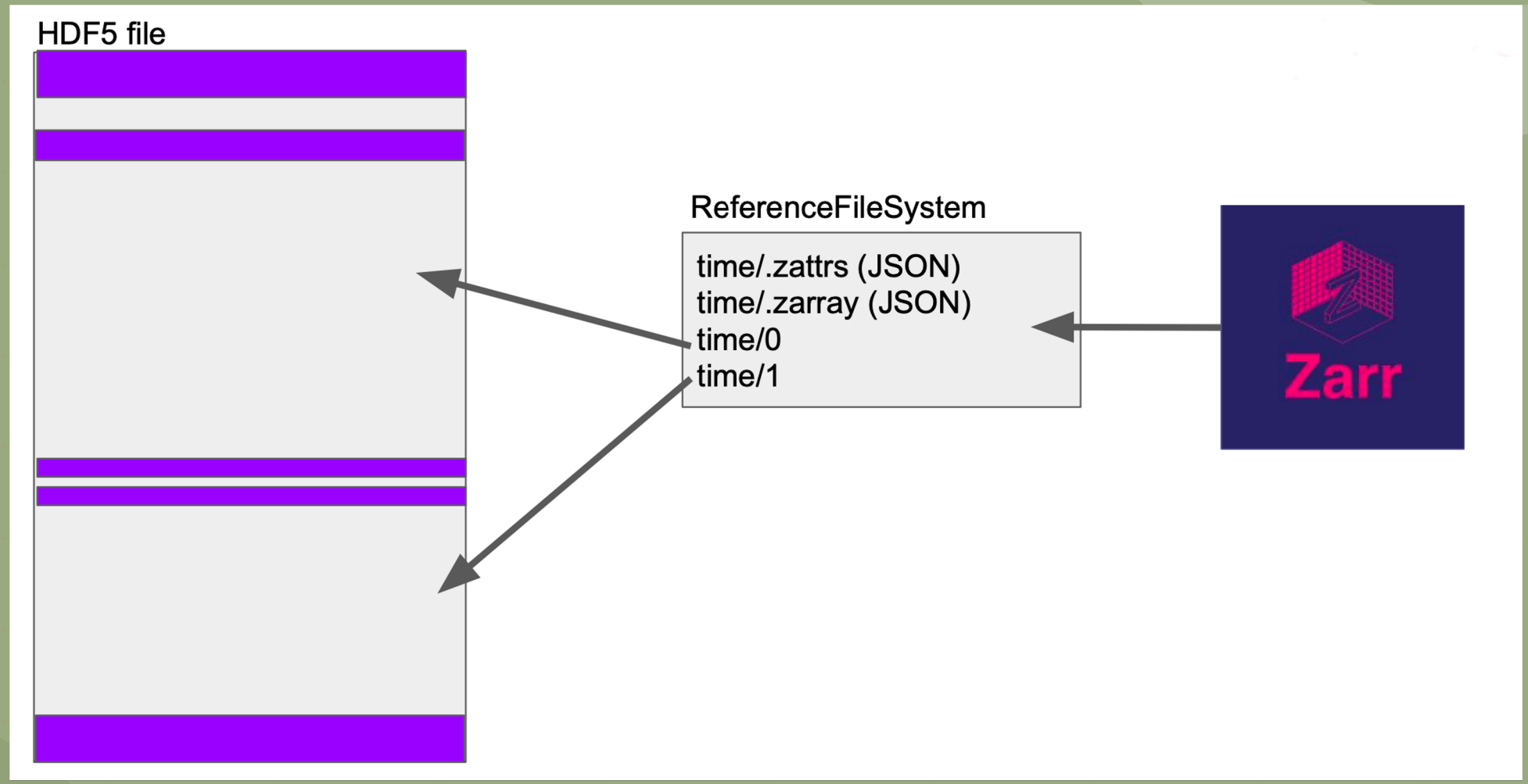


Image courtesy of Martin Durant

So what?

- Kerchunk allows for fast, easy access to existing datasets/formats without the need for data conversion
  - Asymptotically as fast as Zarr
  - Plug-and-play compatibility with xarray + fsspec
- Reference metadata JSONs are small, and can be easily (and cheaply) hosted and shared
- Metadata files can describe data spanning multiple files, representing them as a single dataset

More Info

- Contact me
  - Email: lsterzinger@ucdavis.edu
  - Twitter @lucassterzinger

- Links to Slides, AGU Poster, Kerchunk, and more!
  - https://lucassterzinger.com/2021-agu-poster