

Fake it 'till you make it

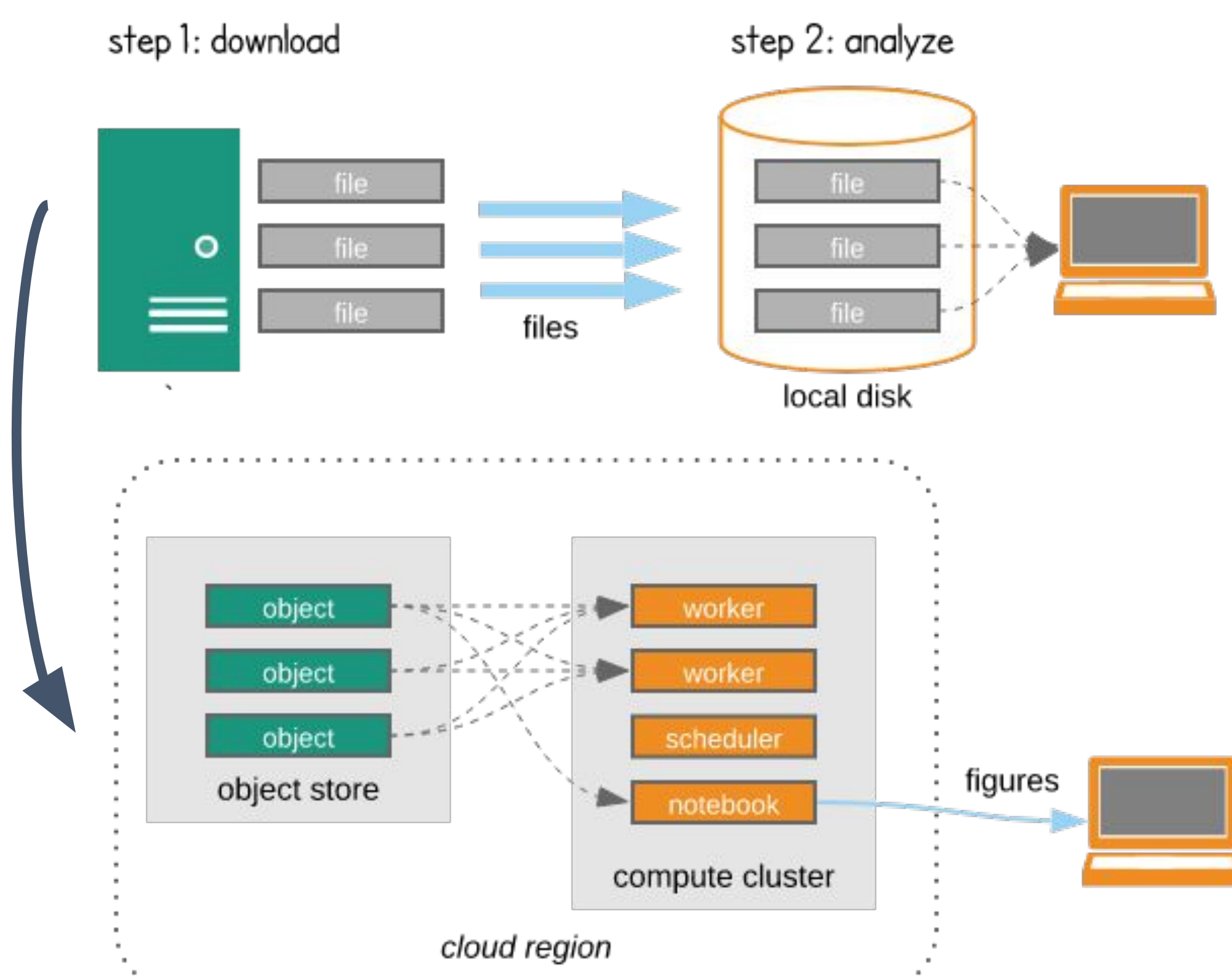
Zarr-like Access of Existing NetCDF4 Datasets



Lucas Sterzinger, Chelle Gentemann, Julia Kent, Kevin Paul



Paradigm Shift

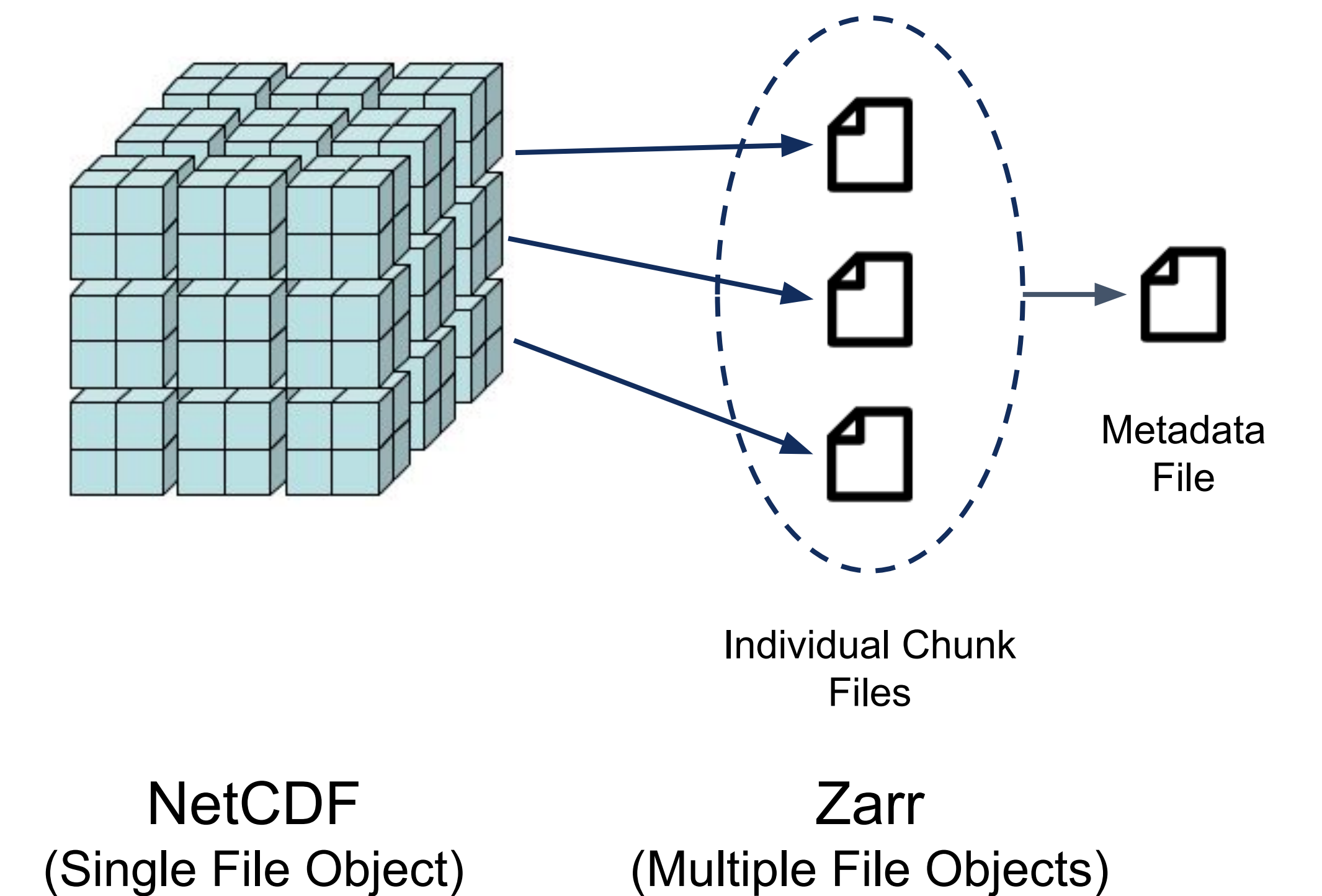


Motivation

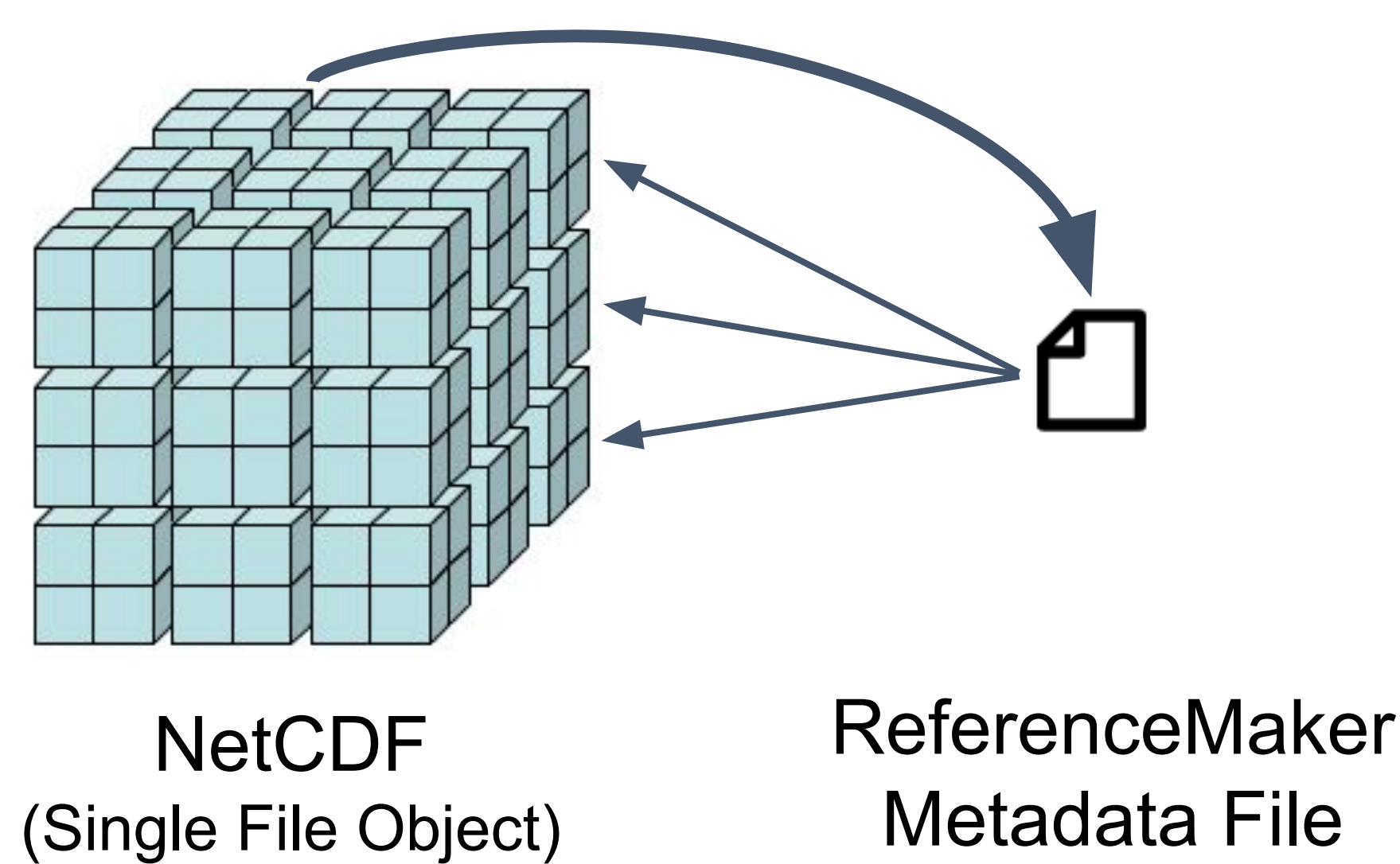
- Many institutions moving data to the cloud
- Most of this data is still in its native NetCDF4/HDF5 format, which is not cloud optimized
- The Zarr format is focused on optimizing large datasets for cloud access
 - Difficult to convince institutions to change data formats

Is it possible to access cloud-hosted NetCDF4 data in a Zarr-like fashion?

NetCDF vs Zarr



ReferenceMaker



Instead of individual chunk files, ReferenceMaker creates metadata files that point to byte offsets in the original NetCDF4 file

Introducing ReferenceMaker

- Creates metadata JSON files using Zarr specification
- Metadata describes remote file location, variable shape, and chunk information
- Allows Zarr engine to access netCDF chunks as if they were Zarr chunk files
- Metadata file is only a few MB per data file and easily shareable
- Can be generated/hosted by 3rd parties

Code available at <https://github.com/intake/fsspec-reference-maker>

Results/Conclusions

Workflow consisted of 24-hour CONUS imagery loop, RGB composite imagery, spatial gradients, timeseries, and histogram. See QR code link for details

| Format | Preprocess Time | Data Open Time | Workflow Time | Extra Storage |
|----------------|-----------------|----------------|---------------|---------------|
| Native netCDF4 | 0 min | 10 minutes | 46 min | 0 GB |
| Zarr | 1 h 38 min | 30 seconds | 4 min 10 s | 52 GB |
| ReferenceMaker | 1 h 25 min | 35 seconds | 4 min 30 s | 416 MB |

Huge speed boost with low storage cost!

- Harness the cloud optimization of Zarr without needing to convert any data
- Reference files can be also created and hosted by third parties



Workflow products, interactive examples, tutorials, and more resources

ACKNOWLEDGMENTS

Special thanks to Rich Signell (USGS) and Martin Durant (Anaconda) on the ReferenceMaker team for their help learning and contributing to this project, to Chelle Gentemann (Farallon Institute), Julia Kent (NCAR) and Kevin Paul (NCAR) for their outstanding mentorship, and to the SIParCS program for a highly productive and insightful summer internship.