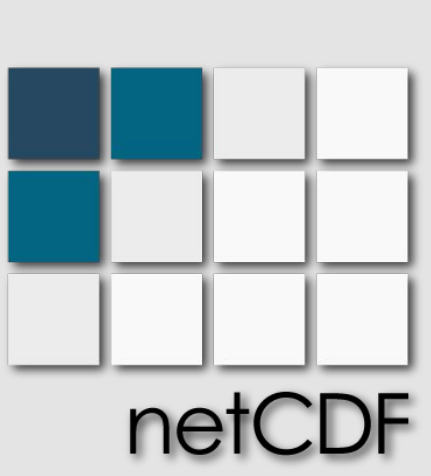
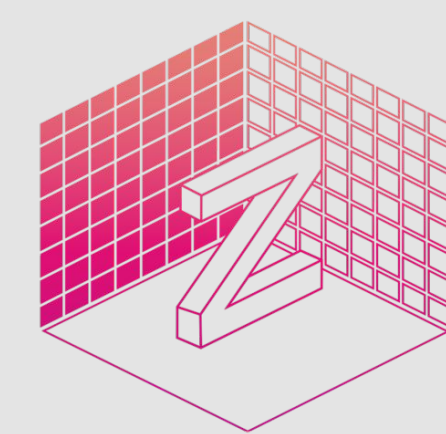




A look into the NetCDF Zarr implementation

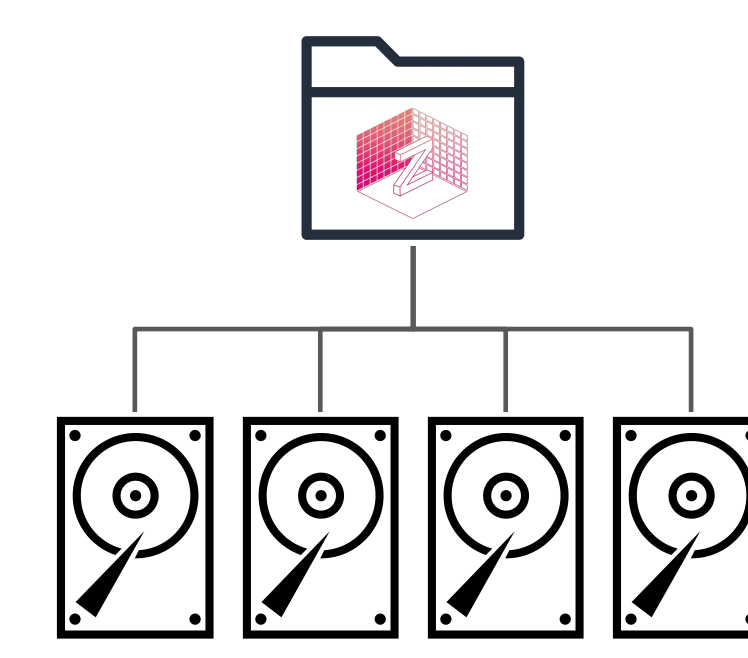
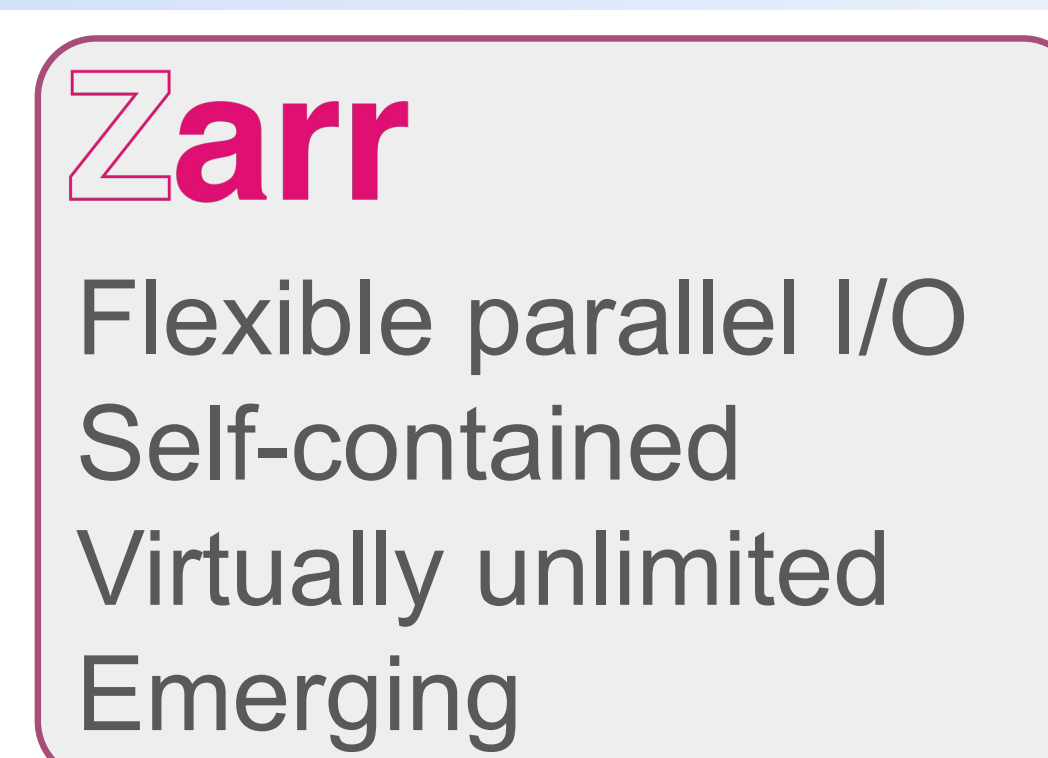
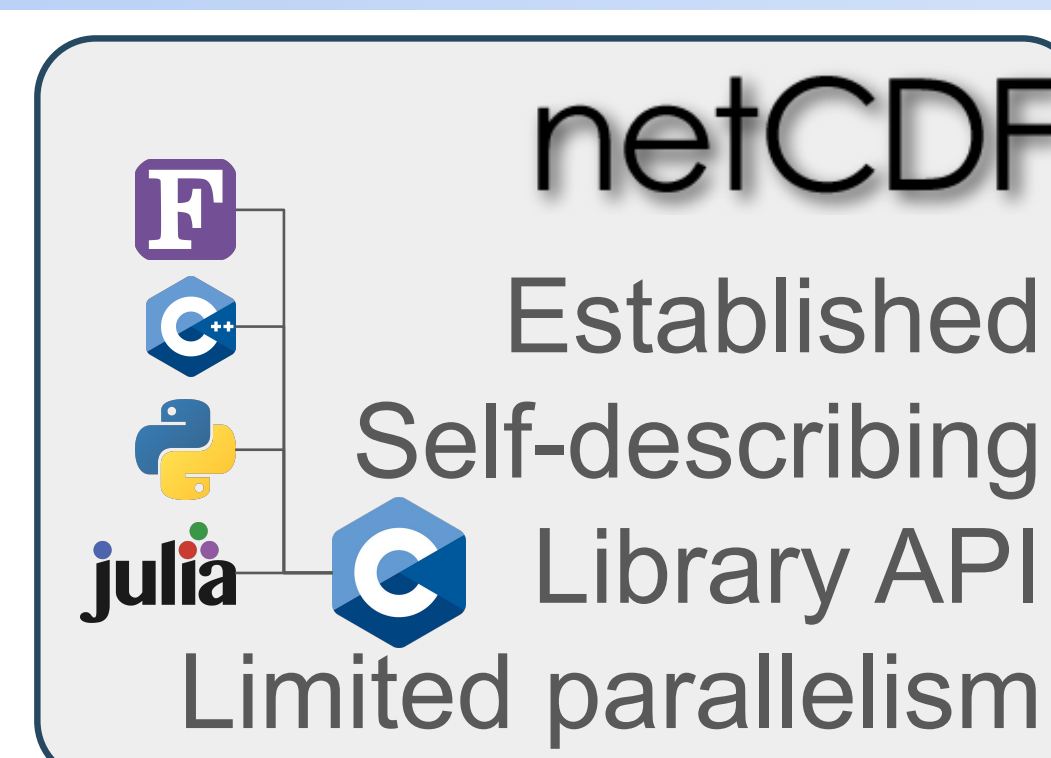
Manuel Reis - Deutsches Klimarechenzentrum - reis@dkrz.de

Ward Fisher - Unidata - wfisher@ucar.edu



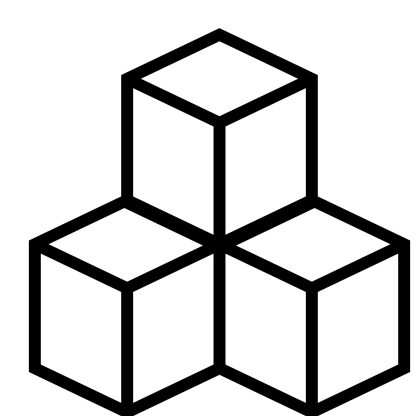
Data formats ecosystem

Ways to serialize/deserialize data are constantly revisited to serve particular needs and motivate new formats. From **Grib**, to **NetCDF**^[1] and now **Zarr**^[2].



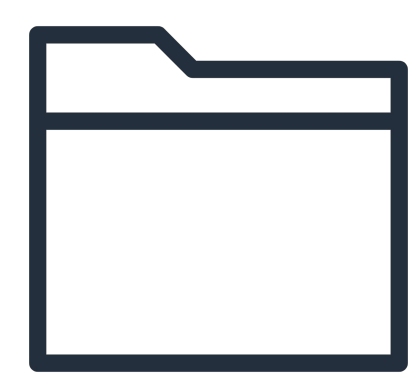
NetCDF Data Model and Architecture

NetCDF-C extended its **data model**, structure and metadata mappings to support Zarr.



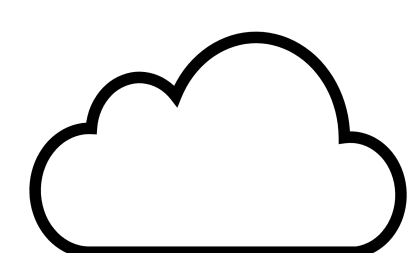
Zarr V2 data model

- Hierarchical
- Chunked



Local access

- Filesystem
- Zip



Remote access

- S3
- HTTP

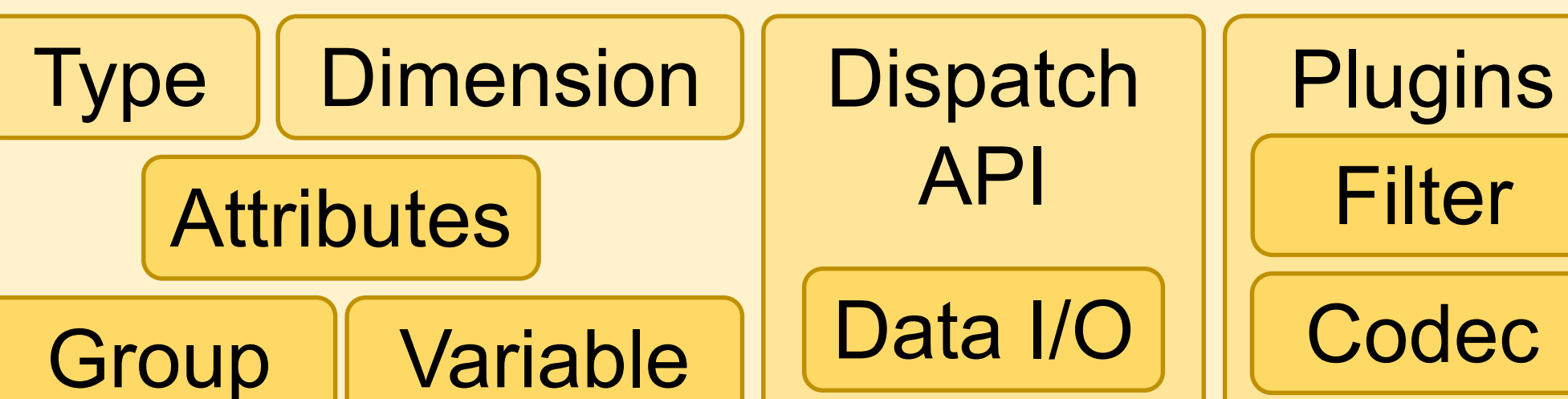


Applications

Analysis, Simulation & Visualization
(Fortran • C++ • Python • Julia • R)

NetCDF API (netcdf.h/libnetcdf.so)

libnetcdf Core



For consolidated datasets only

Interoperability remarks

The Zarr implementation allowed implementation divergences.

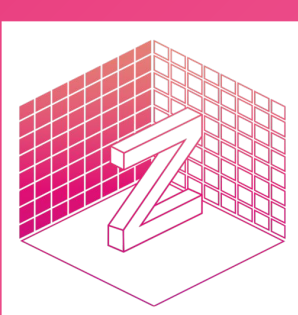
- **Consolidated metadata**  **Now** implemented in NetCDF - **20-40x speedup** on **open()**! ***Especially in remote data.***

Only basic types are supported.

- Temporal - **datetime**
- Non UTF-8 **unicode**
- Scalars - unspecified

Coming directly from **numpy** in **zarr-python** implementation

Special attributes may be needed



Zarr Interoperability & V3 format

```
$ cdo sinfon https://s3.waterpark.dkrz.de/reanalysis/healpix/era5/PT1H/level_7.zarr#mode=zarr,s3,consolidated
File format : NCZarr filter
-1 : Institut Source T Steptype Levels Num Points Num Dtype : Parameter name
1 : ECMWF unknown v instant 1 1 196608 1 F32f : pr
2 : ECMWF unknown v instant 1 1 196608 1 F32f : tas
Grid coordinates :
1 : generic : points=196608
cell : 0 to 196607 by 1
Vertical coordinates :
1 : surface : levels=1
Time coordinate :
time : 754752 steps
RefTime = 1940-01-01 00:00:00 Units = hours Calendar = proleptic_gregorian
YYYY-MM-DD hh:mm:ss YYYY-MM-DD hh:mm:ss YYYY-MM-DD hh:mm:ss YYYY-MM-DD hh:mm:ss
1940-01-01 00:00:00 1940-01-01 01:00:00 1940-01-01 02:00:00 1940-01-01 03:00:00
.....
2026-02-05 20:00:00 2026-02-05 21:00:00 2026-02-05 22:00:00 2026-02-05 23:00:00
cdo sinfon: Processed 2 variables over 754752 timesteps [0.97s 1003MB]
```

Enabling seamless data access across formats in the NetCDF library unlocks modern datasets to the existing software relying on it (e.g., **cdo**^[5]).

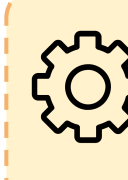
Similarities between **Zarr** and **NetCDF-4 data models** motivate full interoperability.

Challenges:

- Accommodating new **developments**
- **Ad-hoc** convention
- Sustaining the **legacy** and existing **restrictions**

Road ahead - Zarr **V3** brings interoperability improvements and flexibility.

- **Breaks** backward **compatibility**.
- **Clarifies** ambiguities from the previous **V2 specification**.
- Introduces **sharding** - improves the toll on IO and storage layer.
- **Extensions** with schemas.

 **Extending** the NetCDF Zarr layer to **V3** requires significant redesign, at a time when maintenance is costly.