Evaluación de las tecnologías object storage para almacenamiento y análisis de datos climáticos

Autor: Ezequiel Cimadevilla Álvarez

Director: Antonio S. Cofiño González
Codirector: Aida Palacio Hoz

Máster en Ciencia de Datos
Universidad de Cantabria
10 Julio 2019

Grupo de Meteorología de Santander
Grupo de Computación Avanzada

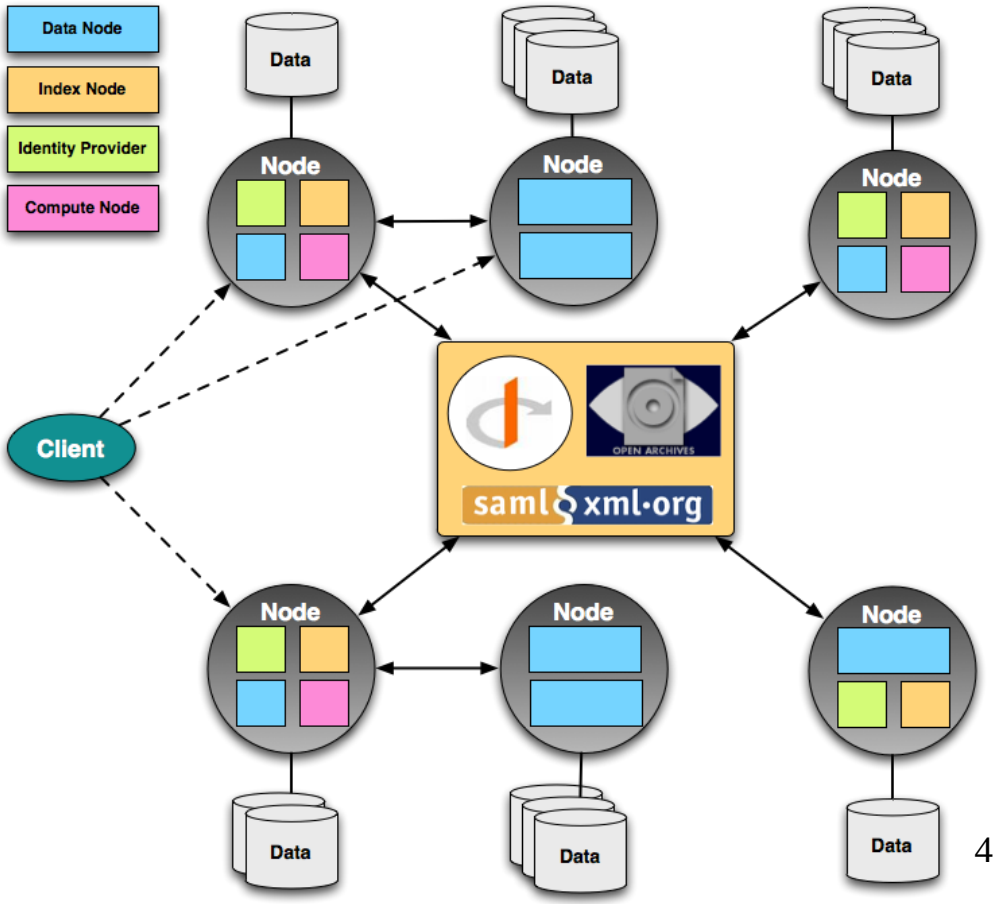
Índice

- Motivación y objetivo
- Datos climáticos
  - Introducción
  - NetCDF
  - Chunking
- Sistemas de almacenamiento
  - Sistemas de ficheros POSIX
  - Object storage
- Almacenamiento en object storage de datos climáticos
- Evaluaciones HPC y cloud
- Conclusiones y trabajo futuro

Motivación y objetivo

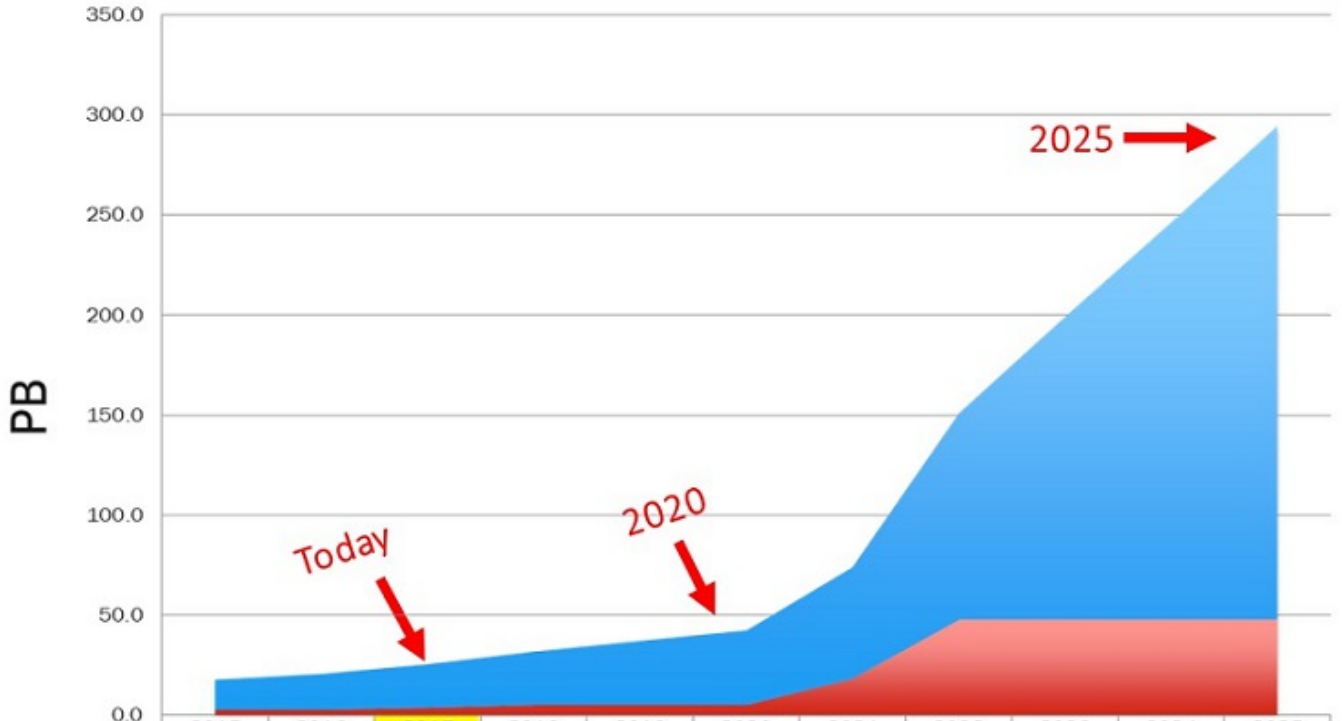
Motivación

- CMIP6 – Sexta fase del marco de trabajo para la mejora del conocimiento sobre cambio climático
  - CMIP3 – 36 TB
  - CMIP5 – 3,3 PB
  - CMIP6 – ¿100 PB?



Motivación

- EOSDIS - NASA's Earth Observing System Data and Information System
- 2020 - 37 PB, 2025 - 246 PB



	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
Cumulative Archive Size (PB)	15.0	17.7	21.6	26.8	32.0	37.2	55.6	103.4	151.1	198.9	246.6
Archive Growth Rate (PB)	2.6	2.8	3.9	5.2	5.2	5.2	18.4	47.7	47.7	47.7	47.7

Archive Growth Rate (PB) Cumulative Archive Size (PB)

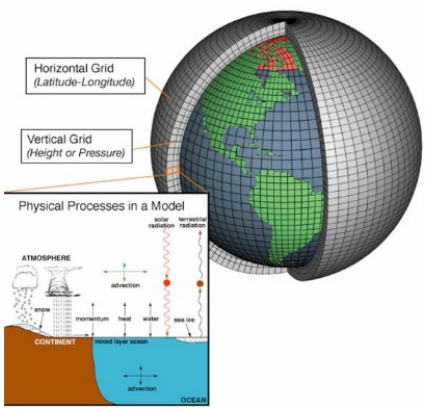
Objetivo

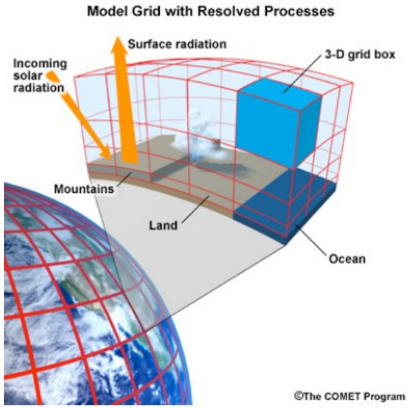
- Análisis del estado del arte sobre almacenamiento y análisis de datos climáticos
- Evaluación de almacenamiento en object storage de datos climáticos
- Evaluación de entornos compartidos para realizar análisis de datos climáticos
- Diseño y despliegue de infraestructuras cloud y HPC para realizar análisis de datos mediante distintas librerías

Datos climáticos

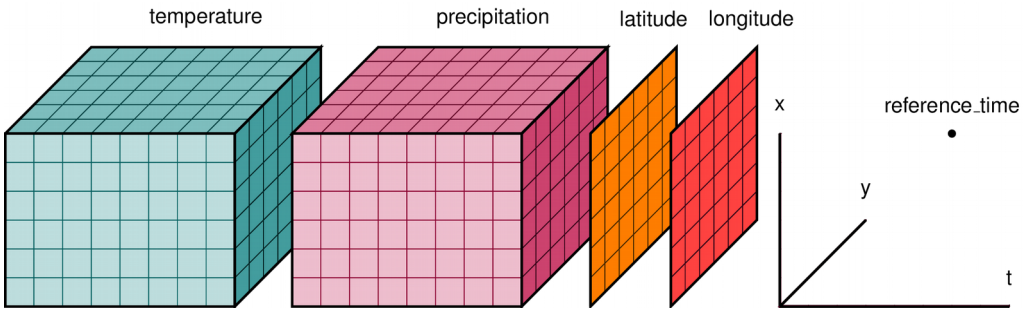
Introducción

- Provienen de observaciones dentro de las ciencias del clima o son producidos por ESMs, modelos del sistema terrestre
- Son datos multidimensionales



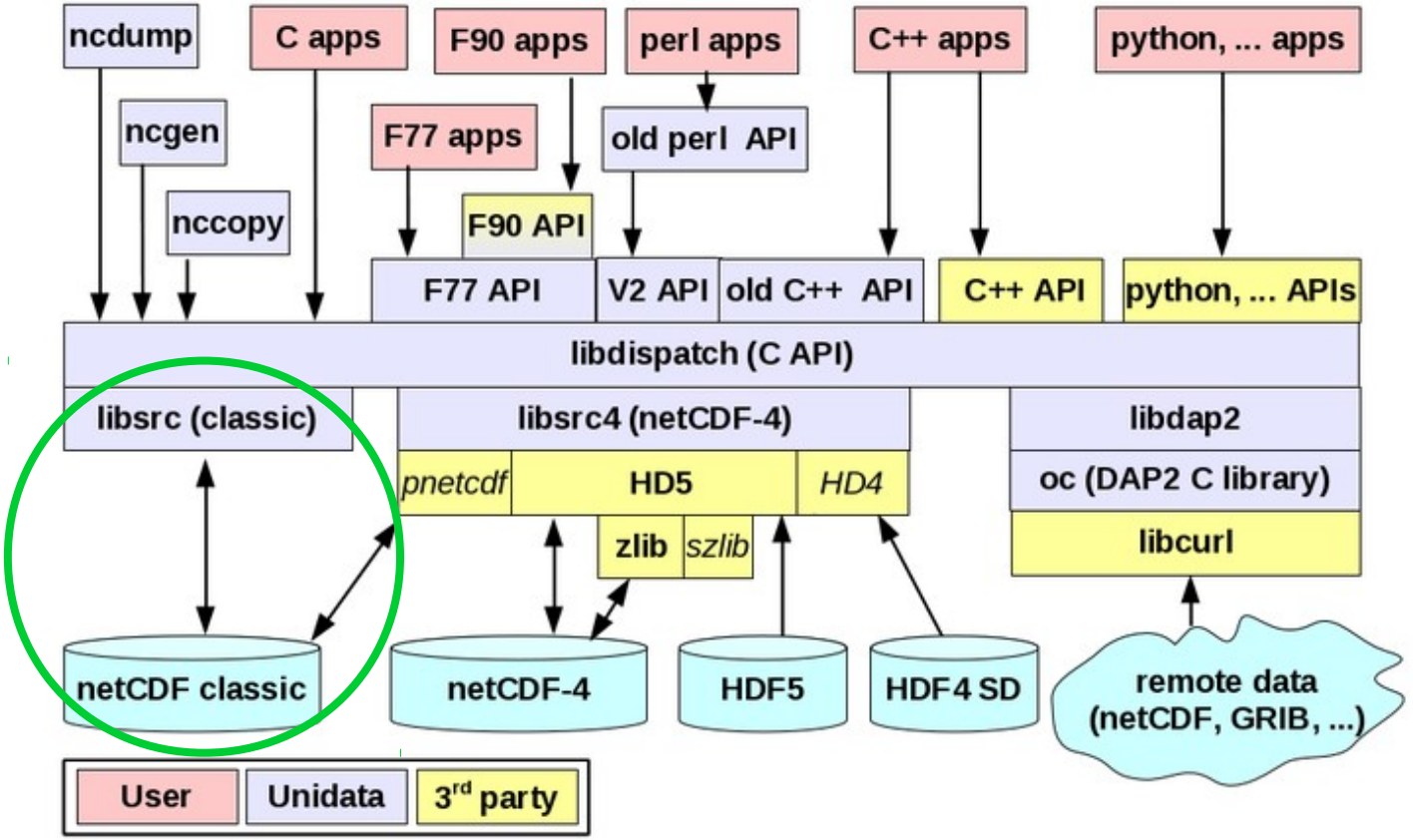


(Image: Maslin and Austin, Nature, 2012, 486, 183)



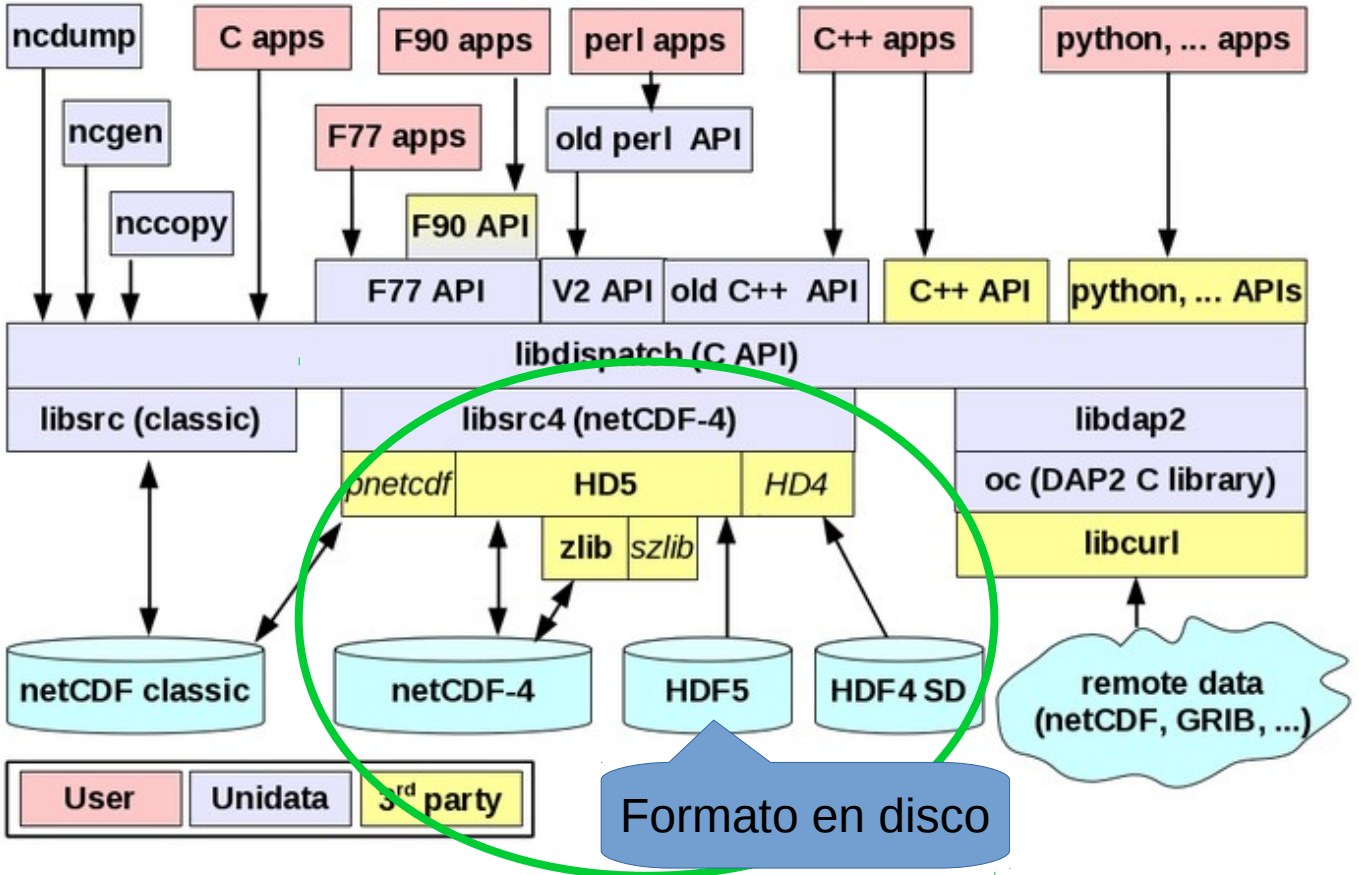
NetCDF

- Librería de referencia para leer y escribir datos climáticos, escrita en lenguaje C
- Múltiples formatos del almacenamiento y APIs



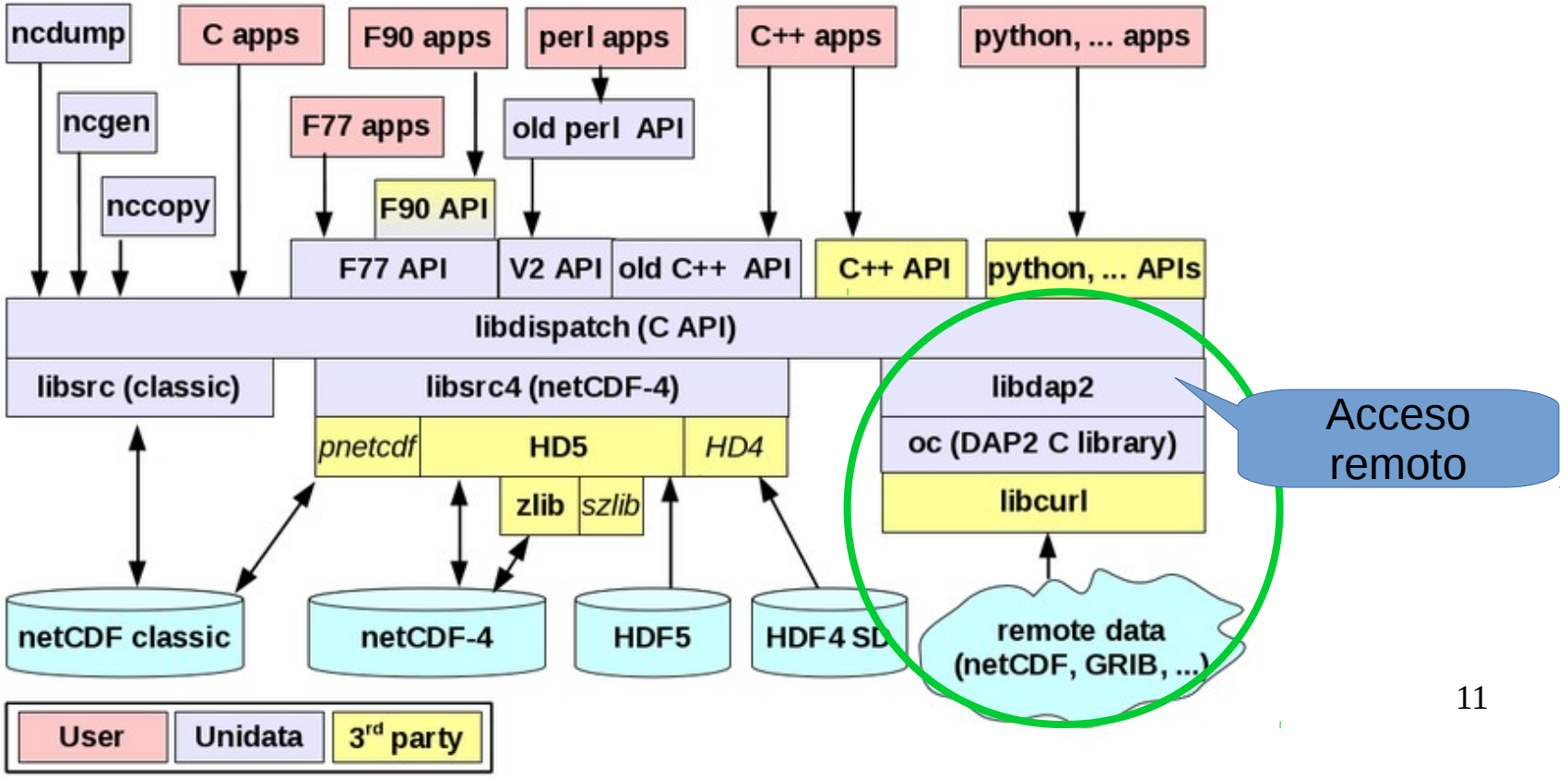
NetCDF

- Librería de referencia para leer y escribir datos climáticos, escrita en lenguaje C
- Múltiples formatos del almacenamiento y APIs



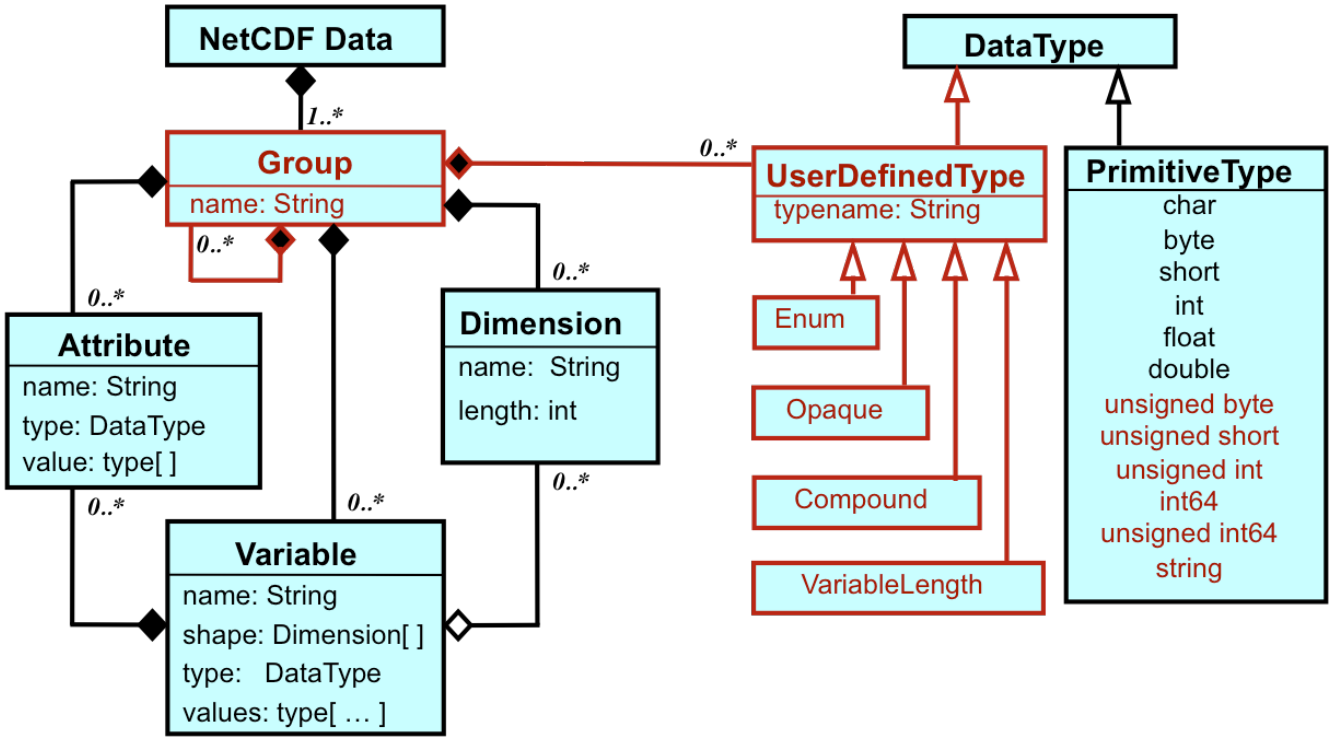
NetCDF

- Librería de referencia para leer y escribir datos climáticos, escrita en lenguaje C
- Múltiples formatos del almacenamiento y APIs

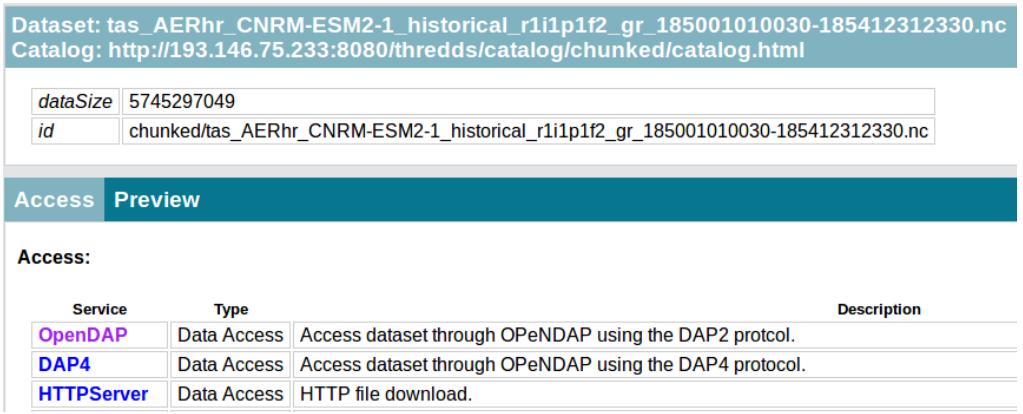


NetCDF

- Modelo de datos formado por grupos, variables multidimensionales, atributos y tipos de datos



NetCDF y DAP



Dataset: tas_AERhr_CNRM-ESM2-1_historical_r1i1p1f2_gr_185001010030-185412312330.nc
Catalog: http://193.146.75.233:8080/thredds/catalog/chunked/catalog.html

dataSize	5745297049
id	chunked/tas_AERhr_CNRM-ESM2-1_historical_r1i1p1f2_gr_185001010030-185412312330.nc

Access **Preview**

Access:

Service	Type	Description
OpenDAP	Data Access	Access dataset through OPeNDAP using the DAP2 protcol.
DAP4	Data Access	Access dataset through OPeNDAP using the DAP4 protocol.
HTTPServer	Data Access	HTTP file download.

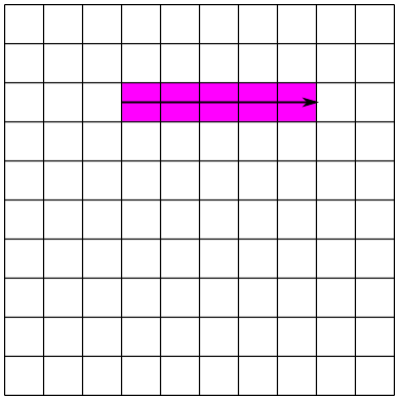
```
(netcdf) [zequi@hera tap] $ ncdump -h "http://193.146.75.233:8080/thredds/dodsC/chunked/tas_AERhr_CNR
M-ESM2-1_historical_r1i1p1f2_gr_185001010030-185412312330.nc"
netcdf tas_AERhr_CNRM-ESM2-1_historical_r1i1p1f2_gr_185001010030-185412312330 {
dimensions:
        time = UNLIMITED ; // (43824 currently)
        axis_nbounds = 2 ;
        lat = 128 ;
        lon = 256 ;
variables:
        double lat(lat) ;
                lat:axis = "Y" ;
                lat:standard_name = "latitude" ;
                lat:long_name = "Latitude" ;
                lat:units = "degrees_north" ;
        double lon(lon) ;
                lon:axis = "X" ;
                lon:standard_name = "longitude" ;
                lon:long_name = "Longitude" ;
                lon:units = "degrees_east" ;
        double height ;
                height:name = "height" ;
                height:standard_name = "height" ;
                height:long_name = "height" ;
                height:units = "m" ;
                height:axis = "Z" ;
                height:positive = "up" ;
        double time(time) ;
                time:axis = "T" ;
                time:standard_name = "time" ;
                time:long_name = "Time axis" ;
                time:calendar = "gregorian" ;
                time:units = "days since 1850-01-01 00:00:00" ;
                time:time_origin = "1850-01-01 00:00:00" ;
                time:bounds = "time_bounds" ;
                time:_ChunkSizes = 2739 ;
        double time_bounds(time, axis_nbounds) ;
                time_bounds:_ChunkSizes = 2739, 2 ;
        float tas(time, lat, lon) ;
                tas:online_operation = "average" ;
                tas:cell_methods = "area: time: mean" ;
                tas:interval_operation = "900 s" ;
                tas:interval_write = "1 h" ;
                tas:_FillValue = 1.e+20f ;
                tas:missing_value = 1.e+20f ;
                tas:coordinates = "height" ;
                tas:standard_name = "air_temperature" ;
                tas:description = "Temperature at surface" ;
                tas:long_name = "Surface Temperature" ;
                tas:history = "none" ;
                tas:units = "K" ;
                tas:cell_measures = "area: areacella" ;
                tas:_ChunkSizes = 2739, 8, 32 ;
```


Problema 1: Grandes volúmenes de datos

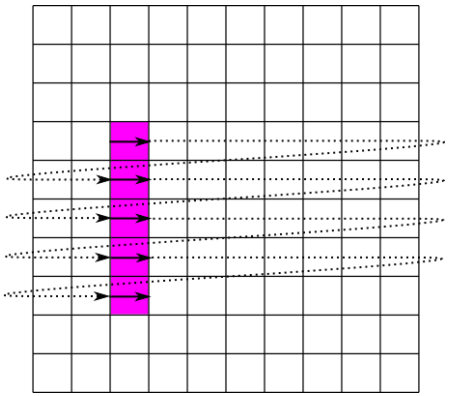
Chunking

Problema 2: Altos tiempos de acceso

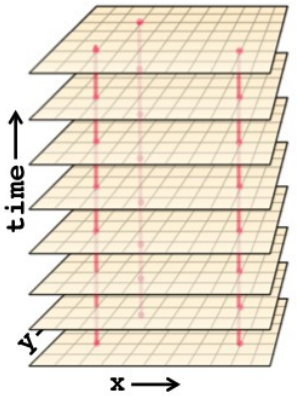
- Disposición en disco de los arrays multidimensionales
- Enorme variabilidad en los tiempos de acceso
- La librería HDF5 usa estructuras de datos para almacenar los chunks dentro del fichero

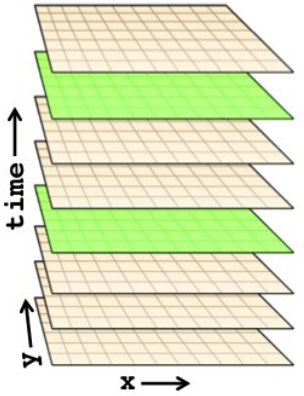


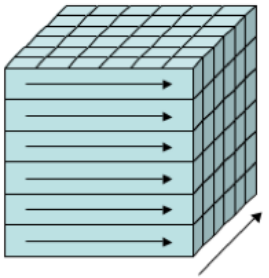
Time series access



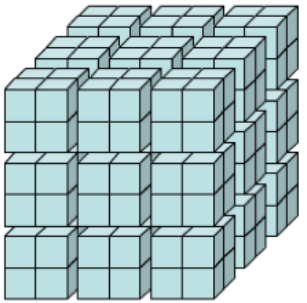
Spatial access







index order

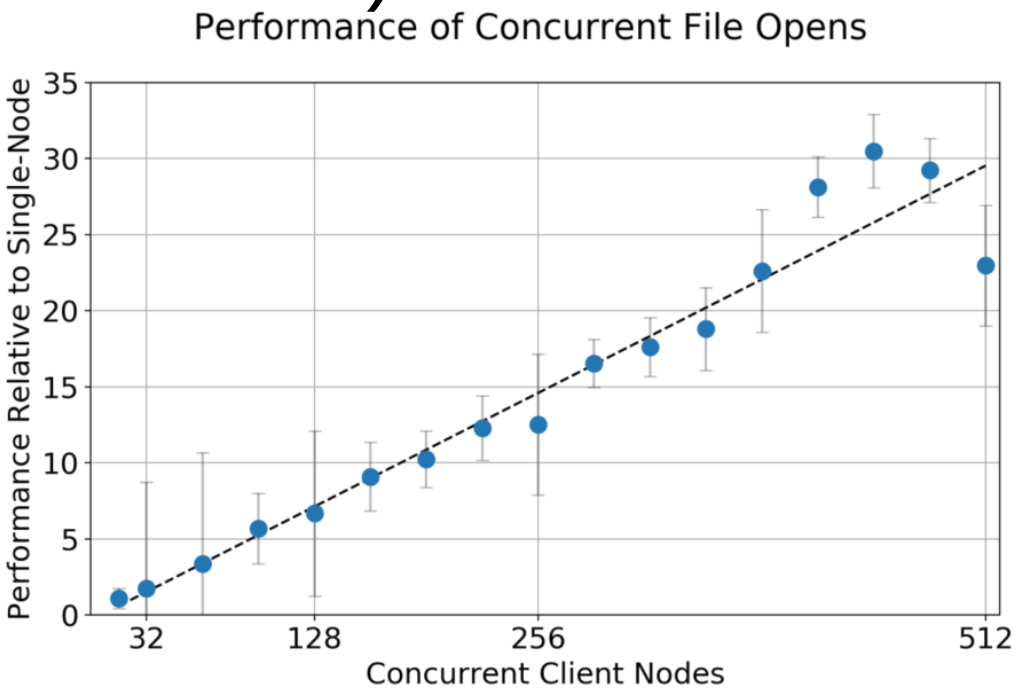


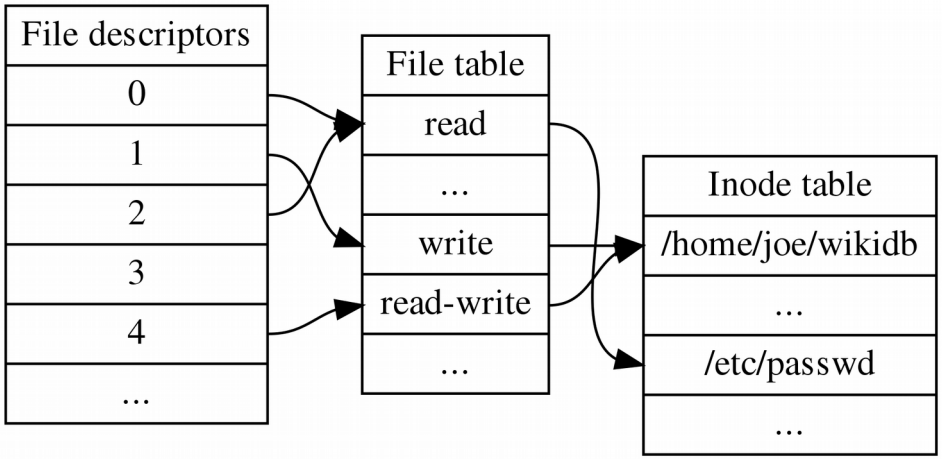
chunked

Sistemas de almacenamiento

Sistemas de ficheros POSIX

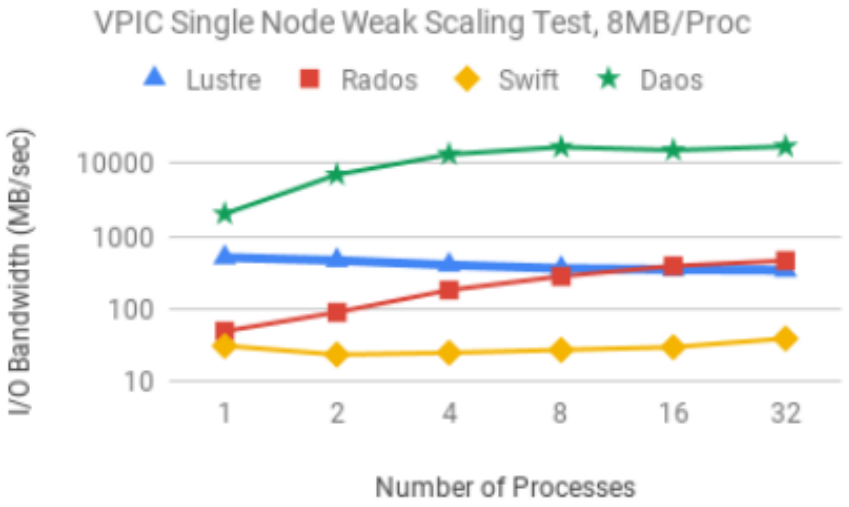
- Los procesos y el sistema operativo mantienen el estado mediante descriptores de fichero
- Bloqueo entre procesos paralelos debido a las semánticas de fuerte consistencia (Lustre, GPFS)

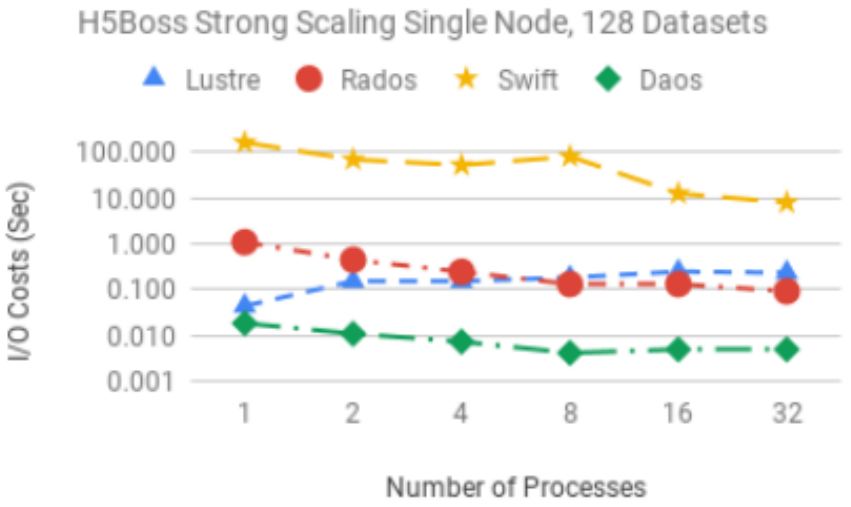




Object storage

- Evitar los problemas de bloqueo POSIX
- Espacio de nombres plano sin metadatos basado en clave-valor
- Acceso mediante operaciones atómicas sin estado
- Inmutabilidad de los objetos





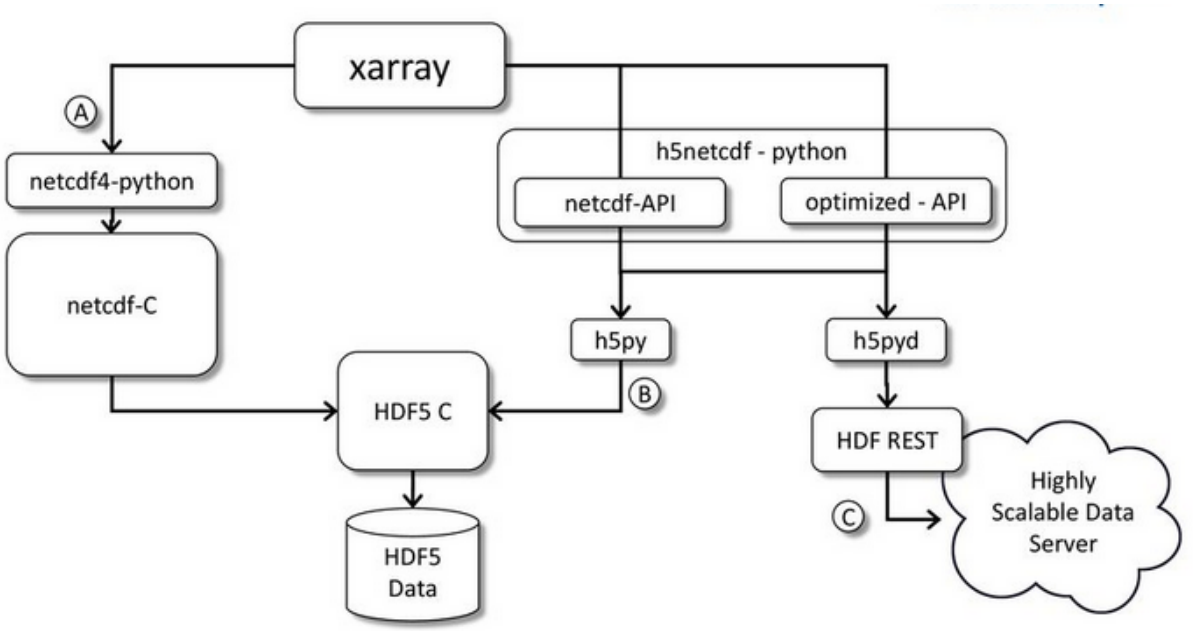
Almacenamiento en object storage de datos climáticos

Almacenamiento en object storage de datos climáticos

- NetCDF y HDF5 no pueden acceder a datos almacenados en object storage
- Posibles soluciones para HDF5/netCDF
  - HSDS
- Nuevas librerías
  - Zarr

HDF5 - HSDS

- API REST/HTTP que representa el modelo de datos HDF5
- Problemas de escalabilidad y consistencia ya resueltos en los object stores



Zarr

- Librería escrita en Python orientada al almacenamiento de arrays multidimensionales
- Modelo de datos similar a netCDF
- Orientado a sistemas clave-valor
- Concepto de ‘store’ o almacén frente a fichero
- Chunks y metadatos se mapean a objetos

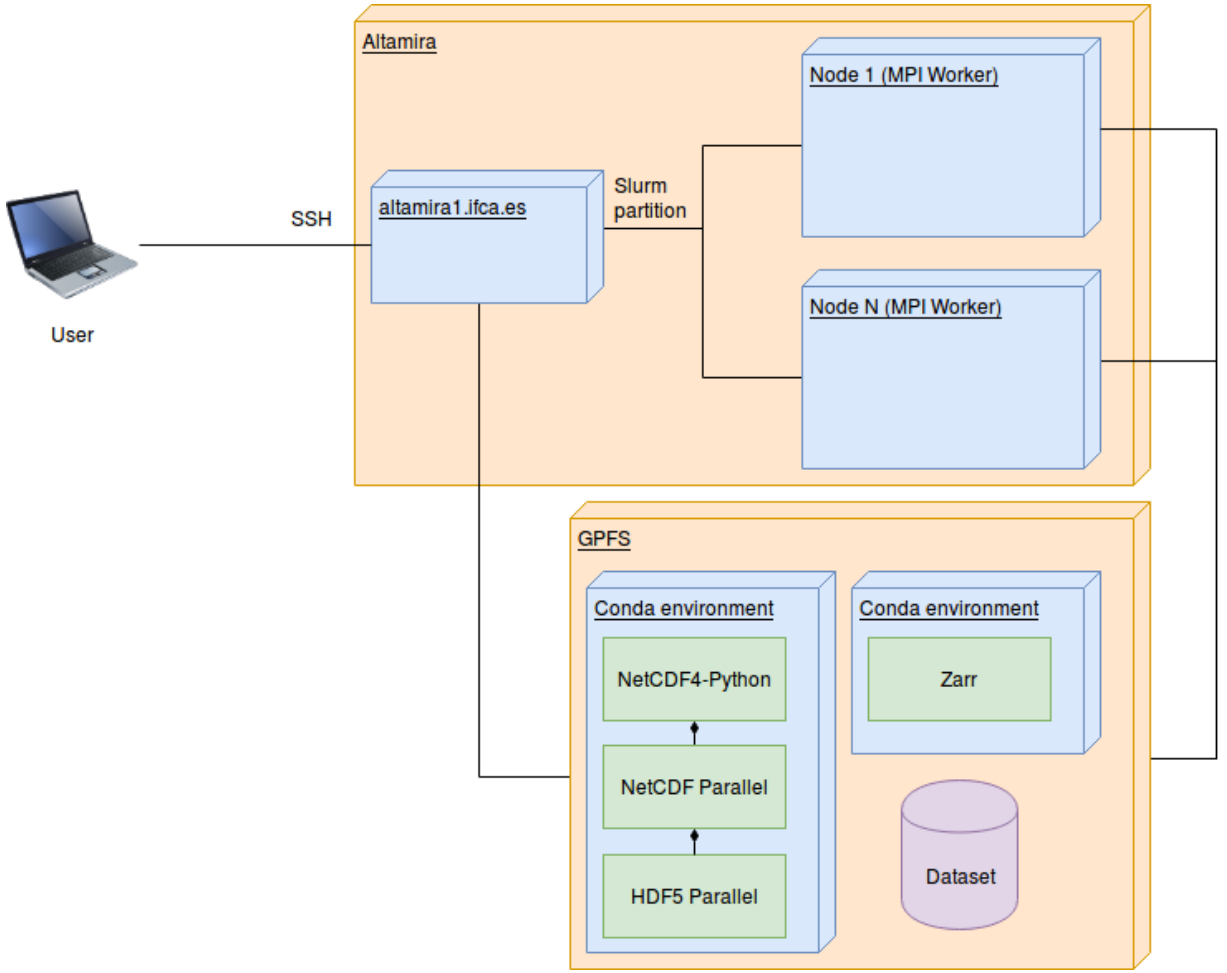
```
[ecimadevilla@altamira1 tas_AERhr_CNRM-ESM2-1_historical_r1i1p1f2_gr_185001010030-185412312330]$ ls -a
.  ..  height  lat  lon  tas  time  time_bounds  .zattrs  .zgroup
[ecimadevilla@altamira1 tas_AERhr_CNRM-ESM2-1_historical_r1i1p1f2_gr_185001010030-185412312330]$ ls -a tas/
.       0.4.4   10.13.4  10.9.0   1.11.6  11.7.4   1.2.2    13.10.5  1.3.6    14.14.5  15.0.1   15.4.3  2.0.6
..      0.4.5   10.13.5  10.9.1   11.1.6  11.7.5   12.2.0   13.10.6  13.6.0   14.14.6  15.0.2   15.4.4  2.0.7
0.0.0   0.4.6   10.13.6  10.9.2   1.11.7  11.7.6   12.2.1   13.10.7  13.6.1   14.14.7  15.0.3   15.4.5  2.1.0
0.0.1   0.4.7   10.13.7  10.9.3   11.1.7  11.7.7   12.2.2   13.1.1   13.6.2   14.1.5   15.0.4   15.4.6  2.10.0
0.0.2   0.5.0   10.1.4   10.9.4   1.1.2   11.8.0   12.2.3   13.11.0  13.6.3   14.15.0  15.0.5   15.4.7  2.10.1
```

Evaluaciones HPC y cloud

Evaluaciones HPC y cloud

- Evaluaciones de análisis de datos en entornos compartidos usando distintos tipos de almacenamiento
- Diseño y despliegue de infraestructuras HPC y cloud
- Caso de uso - Media temporal de la variable temperatura en superficie para cada celda del grid espacial, 5,4 GB

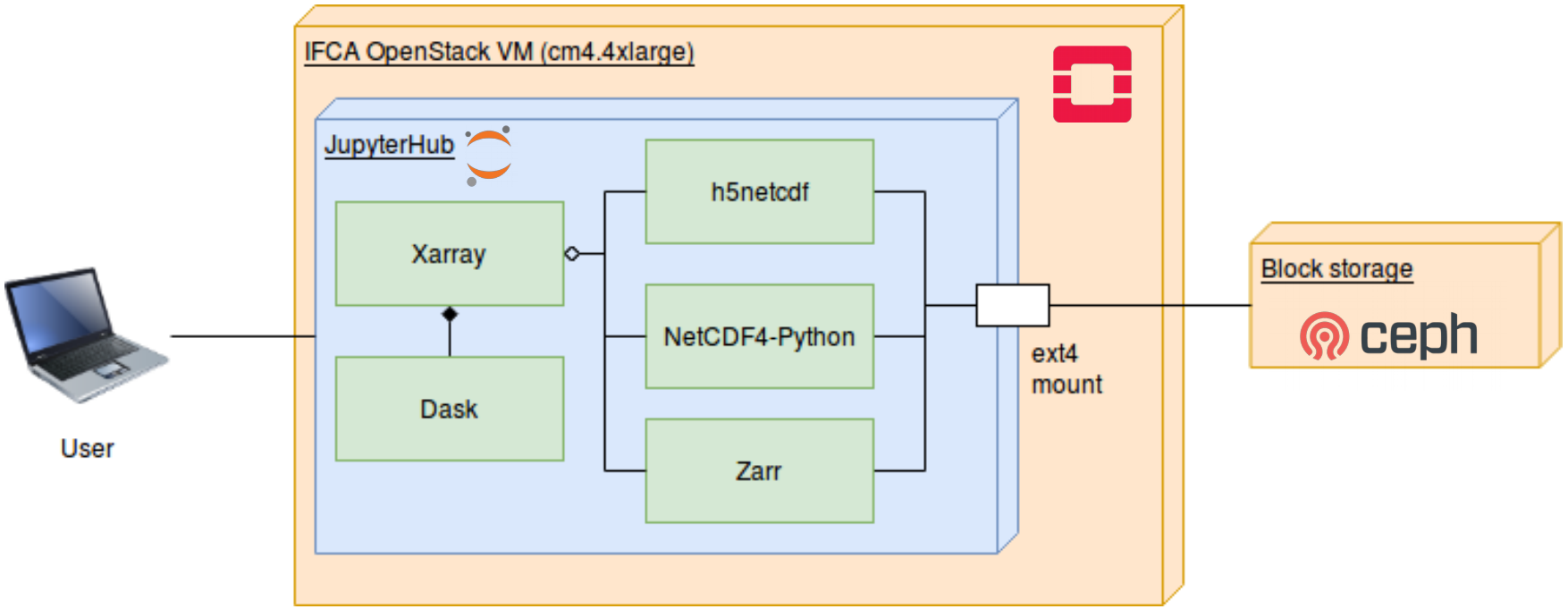
Acceso local HPC



Librería / Tasks	NetCDF4 MPI Independent	NetCDF4 MPI Collective	Zarr
2	145,04s	-	86,82s
4	80,22s	29,52s	37,35s
8	39,73s	17,75s	14,43s

https://github.com/zequihg50/tfm/tree/master/hpc

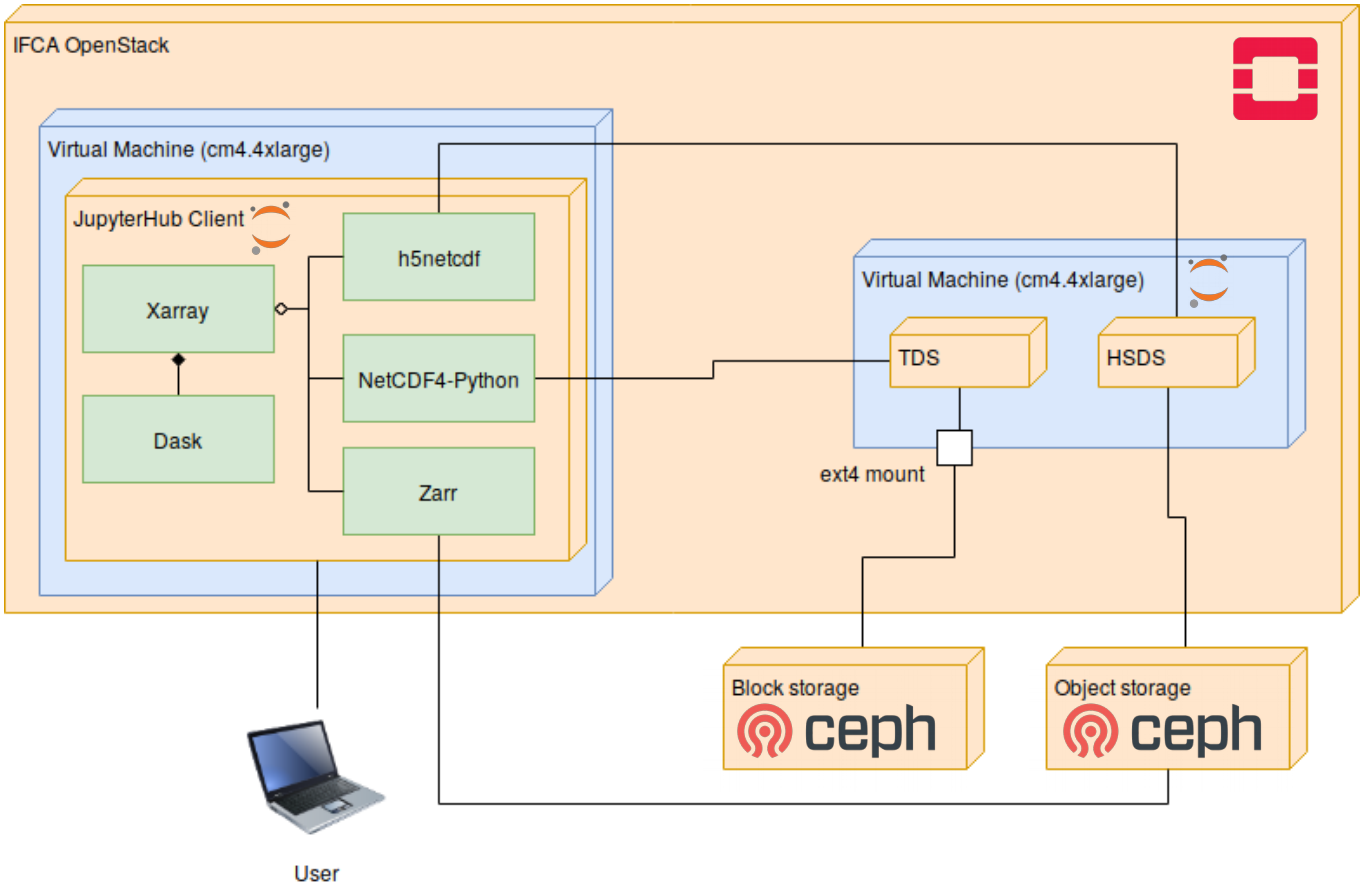
Acceso local cloud



Librería / Acceso	netCDF4	Zarr	h5netcdf
Serie	146,2s	148,8s	128,7s
Threads	90,5s	49,1s	109,1s
Speed up	1,61	3	1,17

https://github.com/zequihg50/tfm/tree/master/local

Acceso remoto cloud



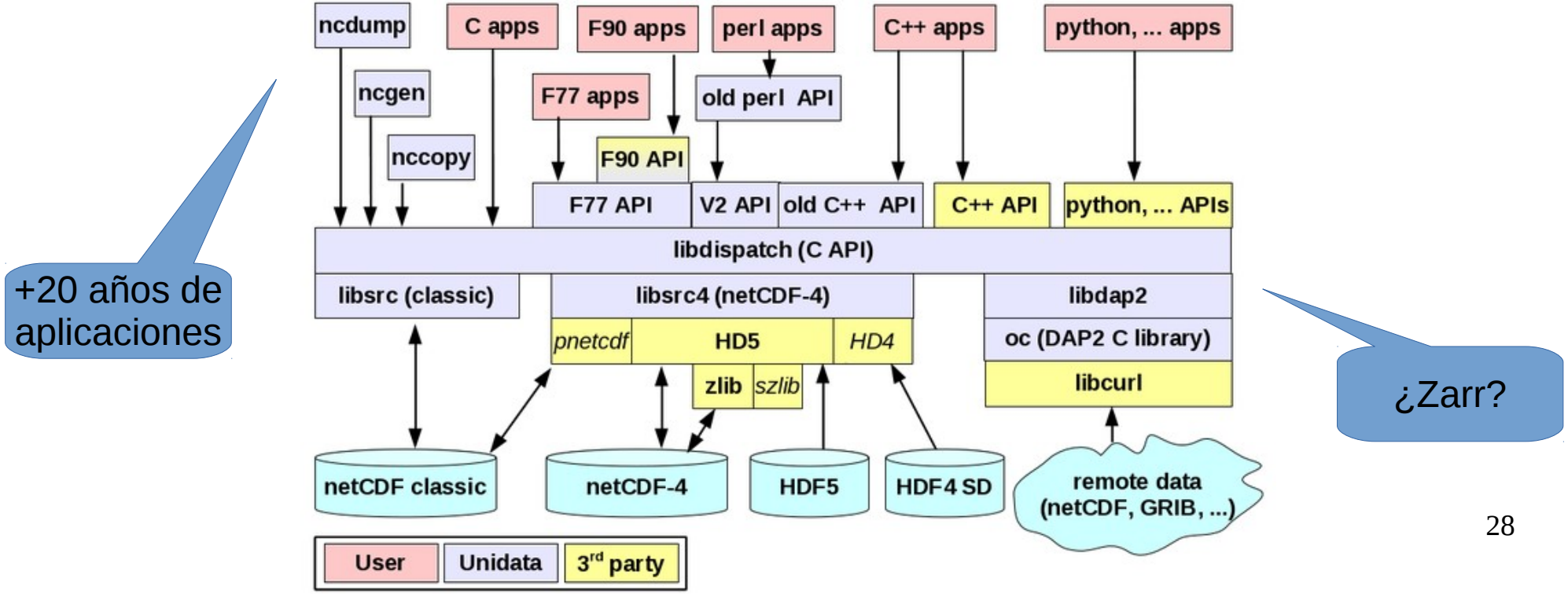
Librería / Acceso	netCDF4 - TDS	Zarr	h5netcdf - HSDS
Serie	330,6s	477,4s	-
Threads	287,9s	60s	-
Speed up	1,14	7,95	-

https://github.com/zequihg50/tfm/tree/master/remote

Conclusiones y trabajo futuro

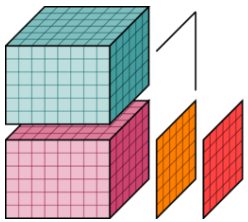
Conclusiones

- El movimiento de los datos a object storage (cloud) forma parte del presente
  - ¿Adaptar netCDF a object storage?
  - ¿Adoptar una nueva librería en la comunidad?



Conclusiones

- Estudio del estado del arte de análisis climáticos (librerías y protocolos)
- Estudio del estado del arte de los sistemas de almacenamiento, tanto sistemas de ficheros como object storage (ver [8] J. Lieu et al.)
- Diseño y despliegue de infraestructuras HPC (compilación y MPI)
- Diseño y despliegue de infraestructuras cloud, en la línea de las prácticas curriculares
- Valor añadido del TFM - “Build your own Pangeo”




Trabajo futuro

- Explicación de las diferencias en los tiempos de acceso
- Extensión del entorno cloud a un clúster en el que realizar paralelismo distribuido
- Evaluación de casos de uso más complejos de minería de datos o machine learning
- Uso de un dataset de mayor tamaño en las pruebas

Evaluación de las tecnologías object storage para almacenamiento y análisis de datos climáticos

Autor: Ezequiel Cimadevilla Álvarez

Director: Antonio S. Cofiño González
Codirector: Aida Palacio Hoz

Máster en Ciencia de Datos
Universidad de Cantabria
10 Julio 2019

Grupo de Meteorología de Santander
Grupo de Computación Avanzada