Bridging the gap between HPC and High Performance Data Analysis

Ben Evans

NATIONAL COMPUTATIONAL INFRASTRUCTURE

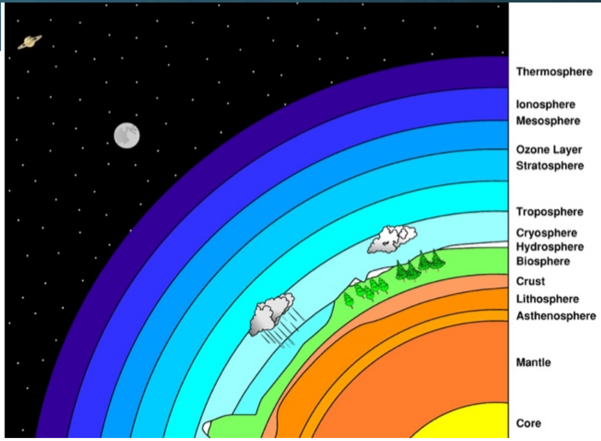
NCI: Some of the HPC and HPD integrated activities

NCI REI Teams (that are directly relevant to this work)

- C. Richards, L. Wyborn – stakeholder engagement and mgt
- J. Wang, K. Gohar, W. Si – Data Collections Team
- J. Antony, P. Larraondo – High Performance Data Team
- D. Roberts, M. Ward, R. Yang – HPC and scaling analysis Team
- C. Trenham, K. Druken, A. Steer – Data Services Team
- J. Smillie, C. Allen, S. Pringle – Virtual Labs Team

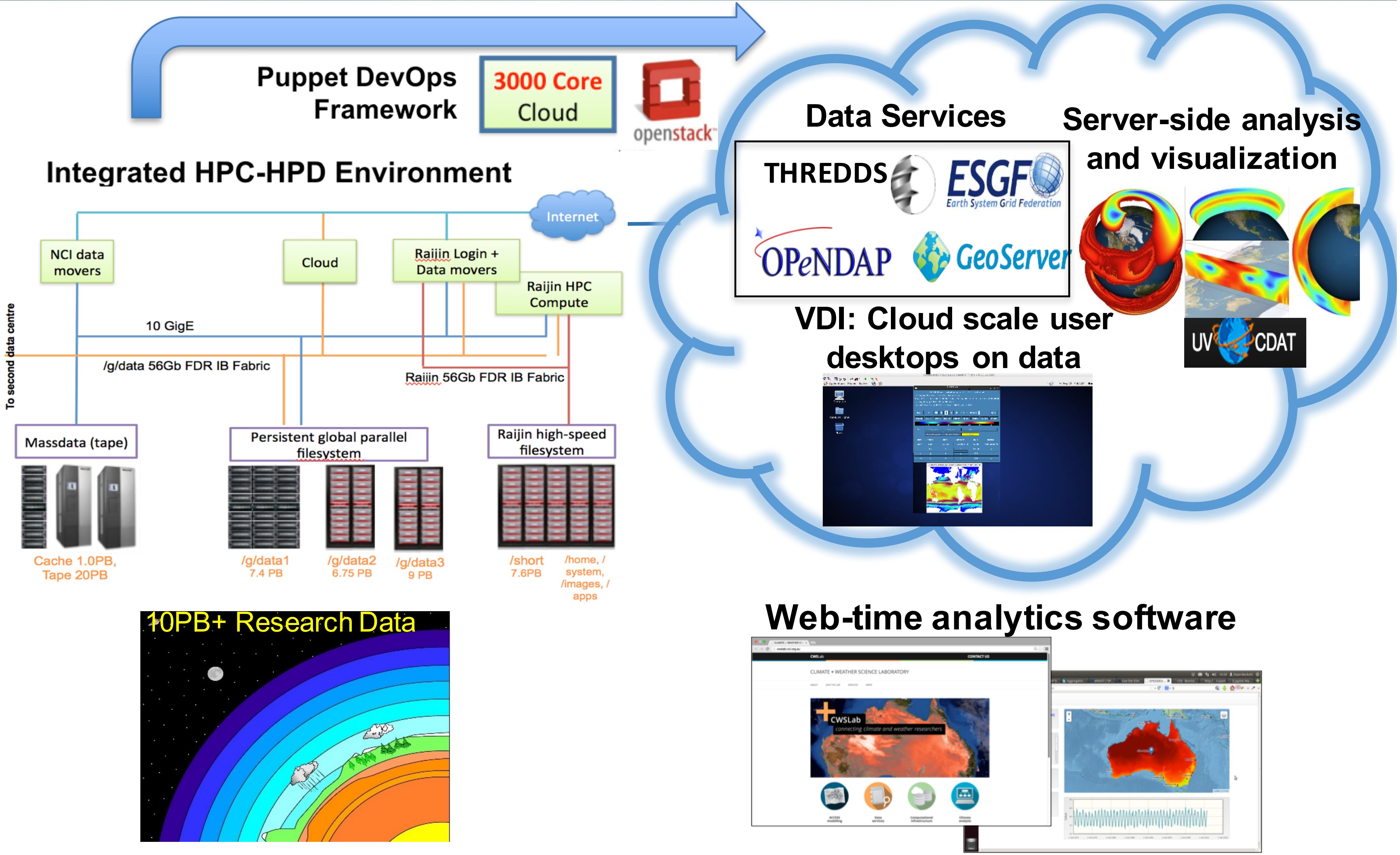
NCI High Performance Data Collections

1. Climate/ESS Model Assets and Data Products
2. Earth and Marine Observations and Data Products
3. Geoscience Collections
4. Terrestrial Ecosystems Collections
5. Water Management and Hydrology Collections
   http://geonetwork.nci.org.au/

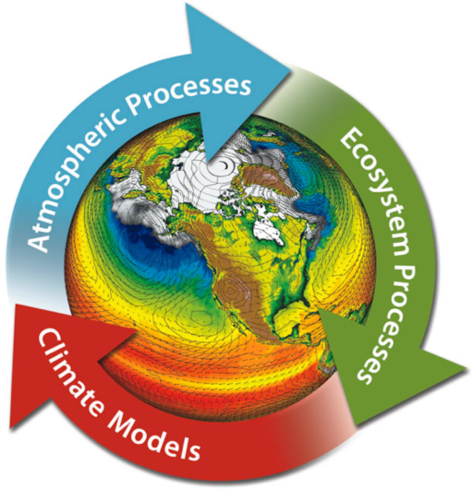


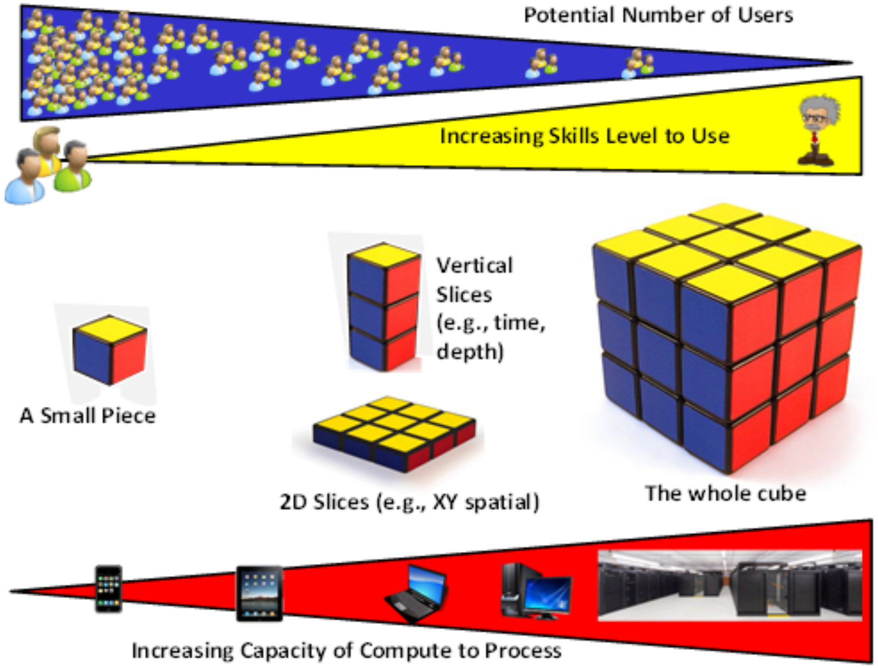
Data Collections	Approx. Capacity
CMIP5, CORDEX, ACCESS Models	5 Pbytes
Earth Obs: Himawari-8, LANDSAT, Sentinel, MODIS, INSAR	2 Pbytes
Digital Elevation, Bathymetry, Onshore/Offshore Geophysics	1 Pbytes
Seasonal Climate	700 Tbytes
Bureau of Meteorology Observations	350 Tbytes
Bureau of Meteorology Ocean-Marine	350 Tbytes
Terrestrial Ecosystem	290 Tbytes
Reanalysis products	100 Tbytes

NCI's Integrated Scientific HPC/HPD Environment



Enable global and continental scale as well as scale-down to local/catchment/plot scale





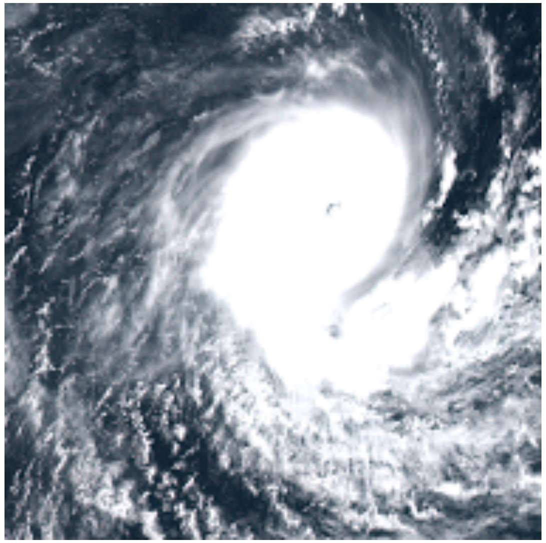
- NWP and Forecasts
  UM, APS3 (Global, Regional, City), ACCESS-TC
- Coupled Seasonal and Decadal Climate
  ACCESS-GC2/3 (GloSea5)
- Data Assimilation
  3D-VAR, 4D-VAR (Atmosphere), EnKF (Ocean)
- Ocean Forecasting and Research
  OceanMaps, BlueLink, MOM5, CICE/SIS, WW3, ROMS
- Fully-Coupled Earth System Model
  ACCESS-CM, ACCESS-ESM, CMIP5/6

- Water availability and usage over time
- Catchment zone
- Vegetation changes
- Data fusion with pt-clouds and local or other measurements
- Statistical techniques on key variables

Some of the foci for Satellite Imagery

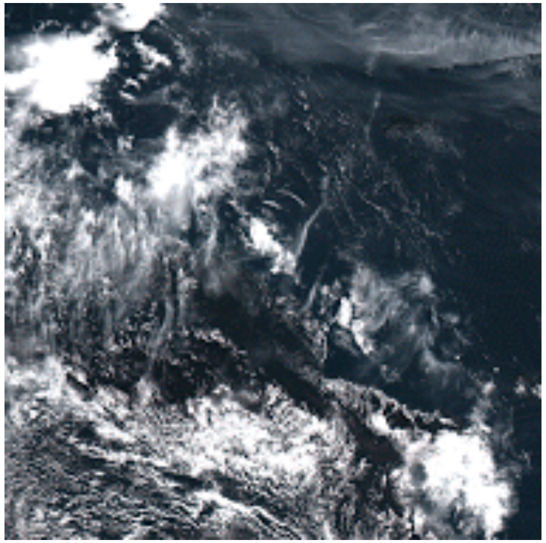
- Modelling Extreme and High Impact events – BoM
- NWP, Climate Coupled Systems and Data Assimilation – BoM, CSIRO, Uni's.
- Hazards - Geoscience Australia, BoM
- Monitoring the Environment and Ocean – ANU, BoM, CSIRO, GA, IMOS, TERN

Tropical Cyclones



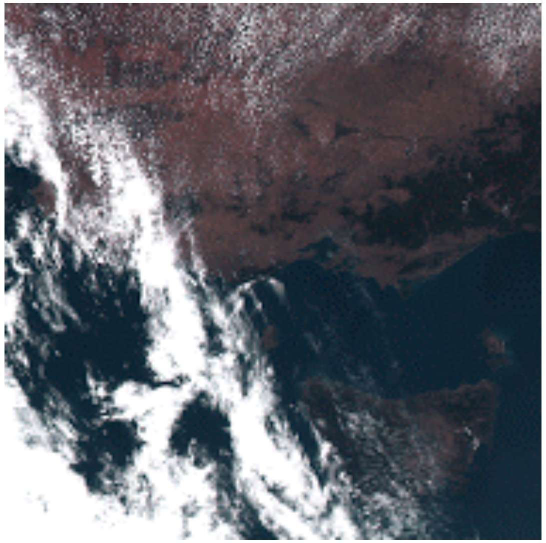
Cyclone Winston
20-21 Feb, 2016

Volcanic Ash



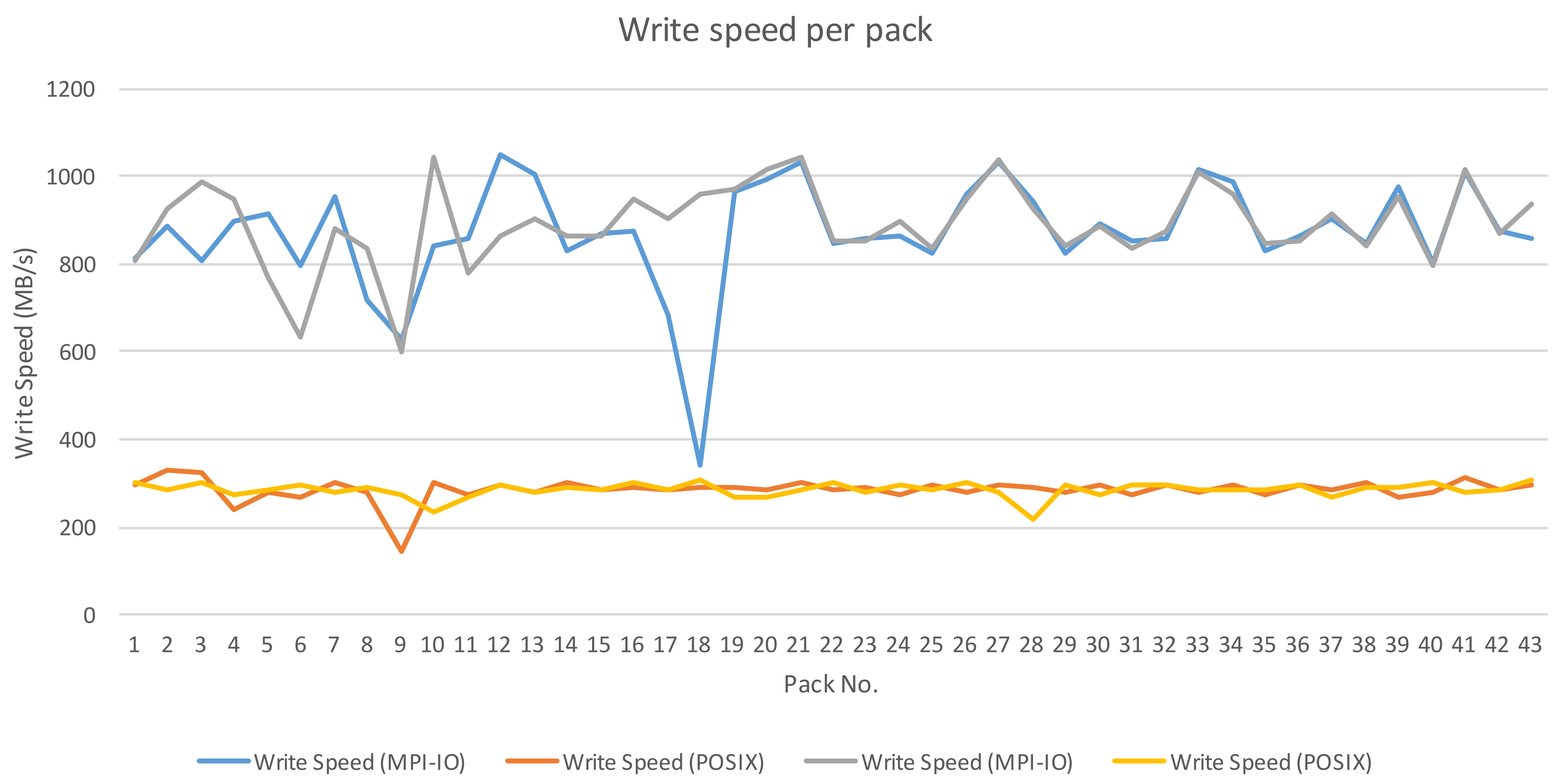
Manam Eruption
31 July, 2015

Bush Fires



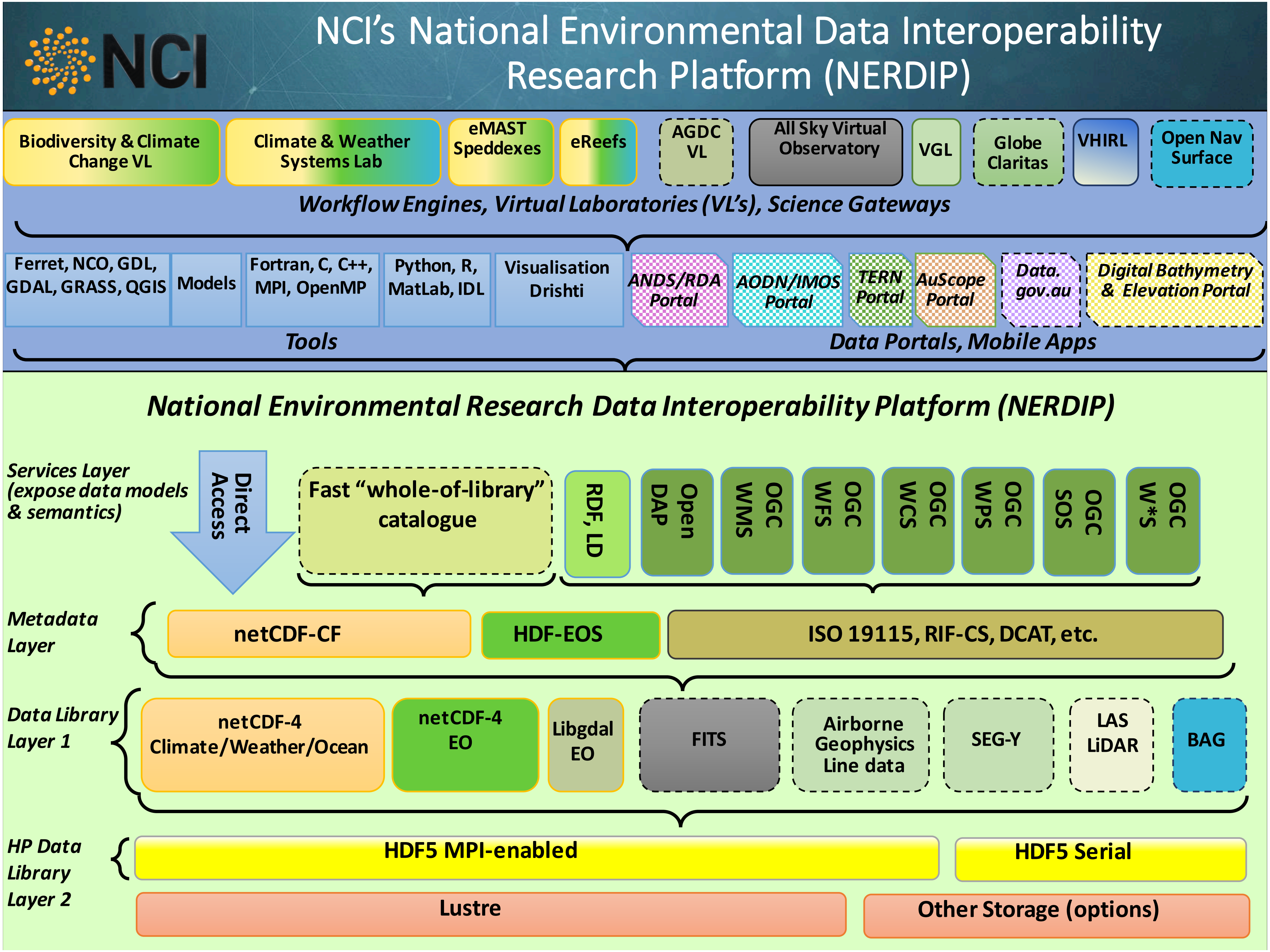
Wye Valley and
Lorne Fires
25-31 Dec, 2015

Example of I/O for large scale HPC:
- UM (atmosphere) with and without parallel IO



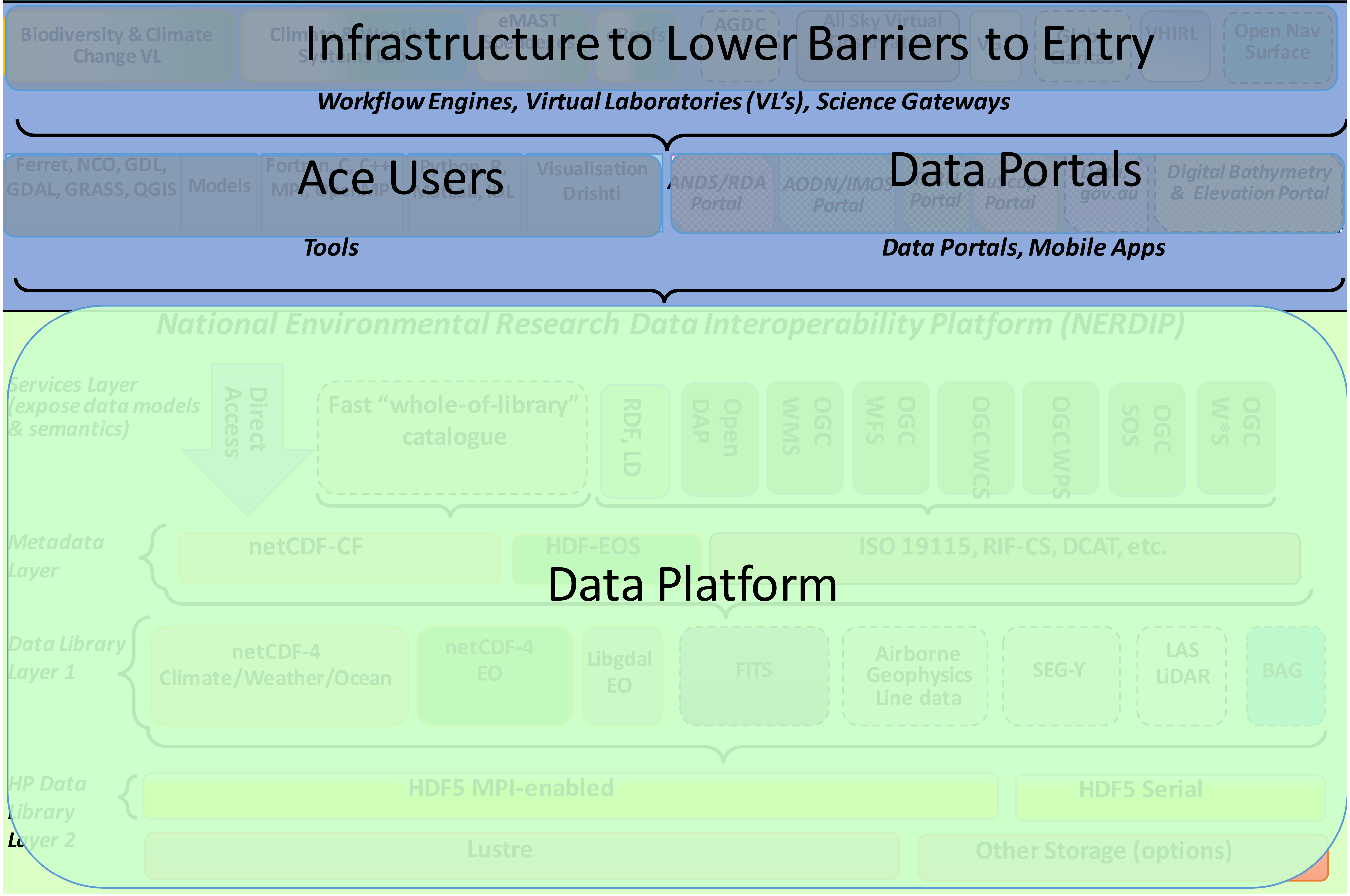
I/O speeds per pack for UM files with and without MPI-IO.

c/- Dale Roberts

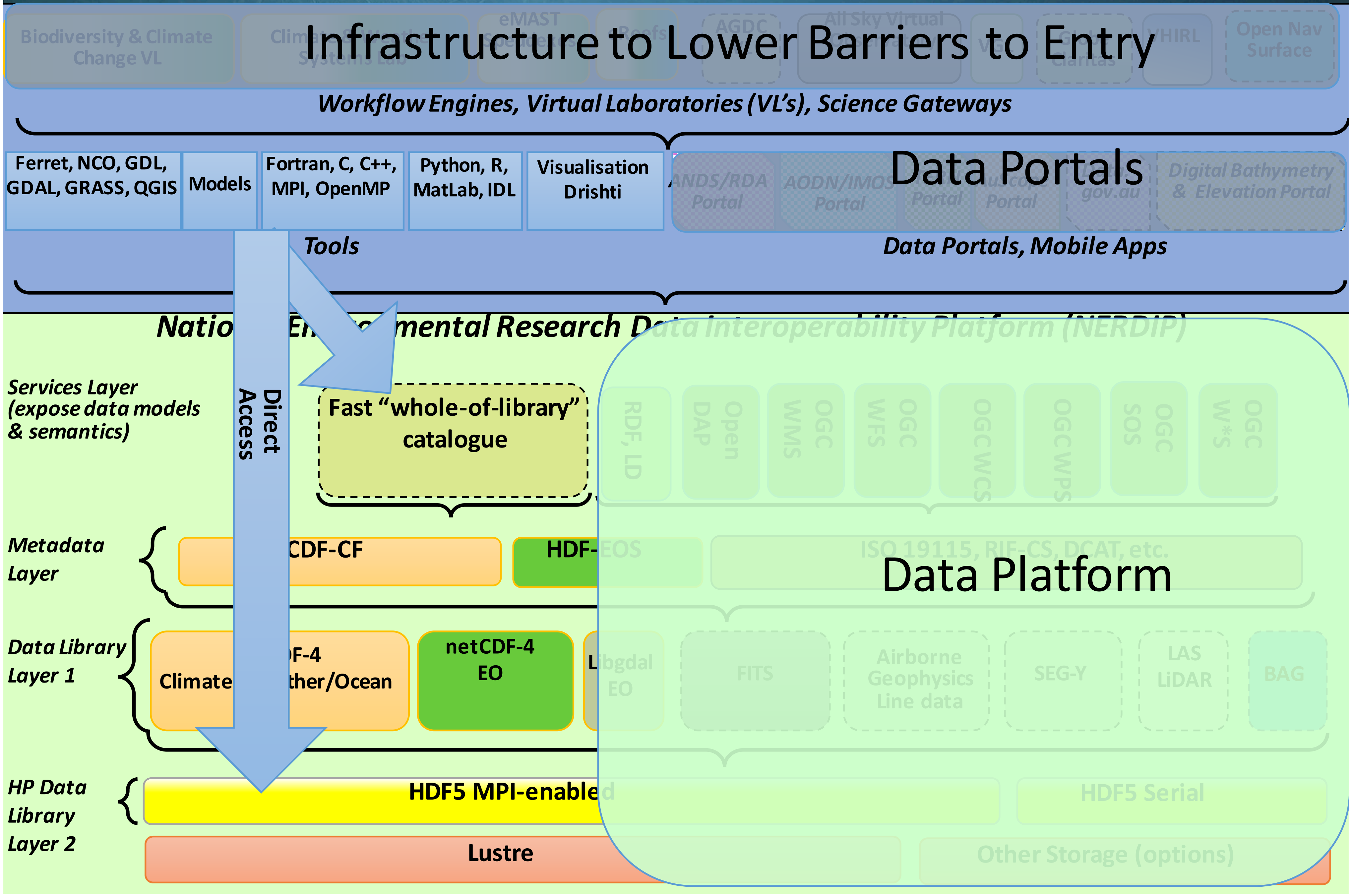


NCI
NCI's National Environmental Data Interoperability Research Platform (NERDIP)
Biodiversity & Climate Change VL
Climate & Weather Systems Lab
eMAST Speddexes
eReefs
AGDC VL
All Sky Virtual Observatory
VGL
Globe Claritas
VHIRL
Open Nav Surface
Workflow Engines, Virtual Laboratories (VL's), Science Gateways
Ferret, NCO, GDL, GDAL, GRASS, QGIS
Models
Fortran, C, C++, MPI, OpenMP
Python, R, MatLab, IDL
Visualisation Drishti
ANDS/RDA Portal
AODN/IMOS Portal
TERN Portal
AuScope Portal
Data. gov.au
Digital Bathymetry & Elevation Portal
Tools
Data Portals, Mobile Apps
National Environmental Research Data Interoperability Platform (NERDIP)
Services Layer (expose data models & semantics)
Direct Access
Fast "whole-of-library" catalogue
RDF, LD
Open DAP
OGC WMS
OGC WFS
OGC WCS
OGC WPS
OGC SOS
OGC W*S
Metadata Layer
netCDF-CF
HDF-EOS
ISO 19115, RIF-CS, DCAT, etc.
Data Library Layer 1
netCDF-4 Climate/Weather/Ocean
netCDF-4 EO
Libgdal EO
FITS
Airborne Geophysics Line data
SEG-Y
LAS LiDAR
BAG
HP Data Library Layer 2
HDF5 MPI-enabled
HDF5 Serial
Lustre
Other Storage (options)

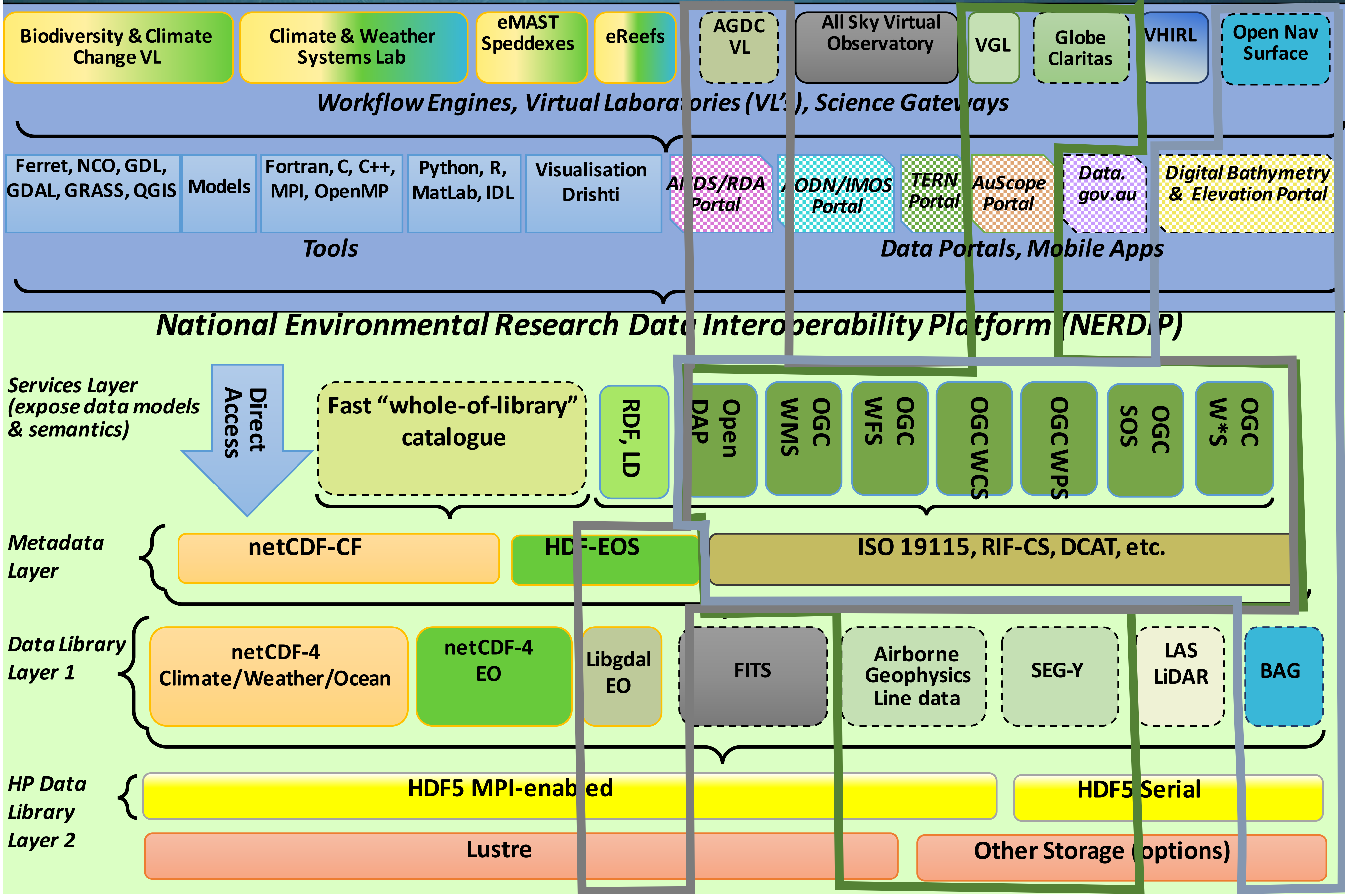
NERDIP: Enabling Multiple Ways to Interact with the Data



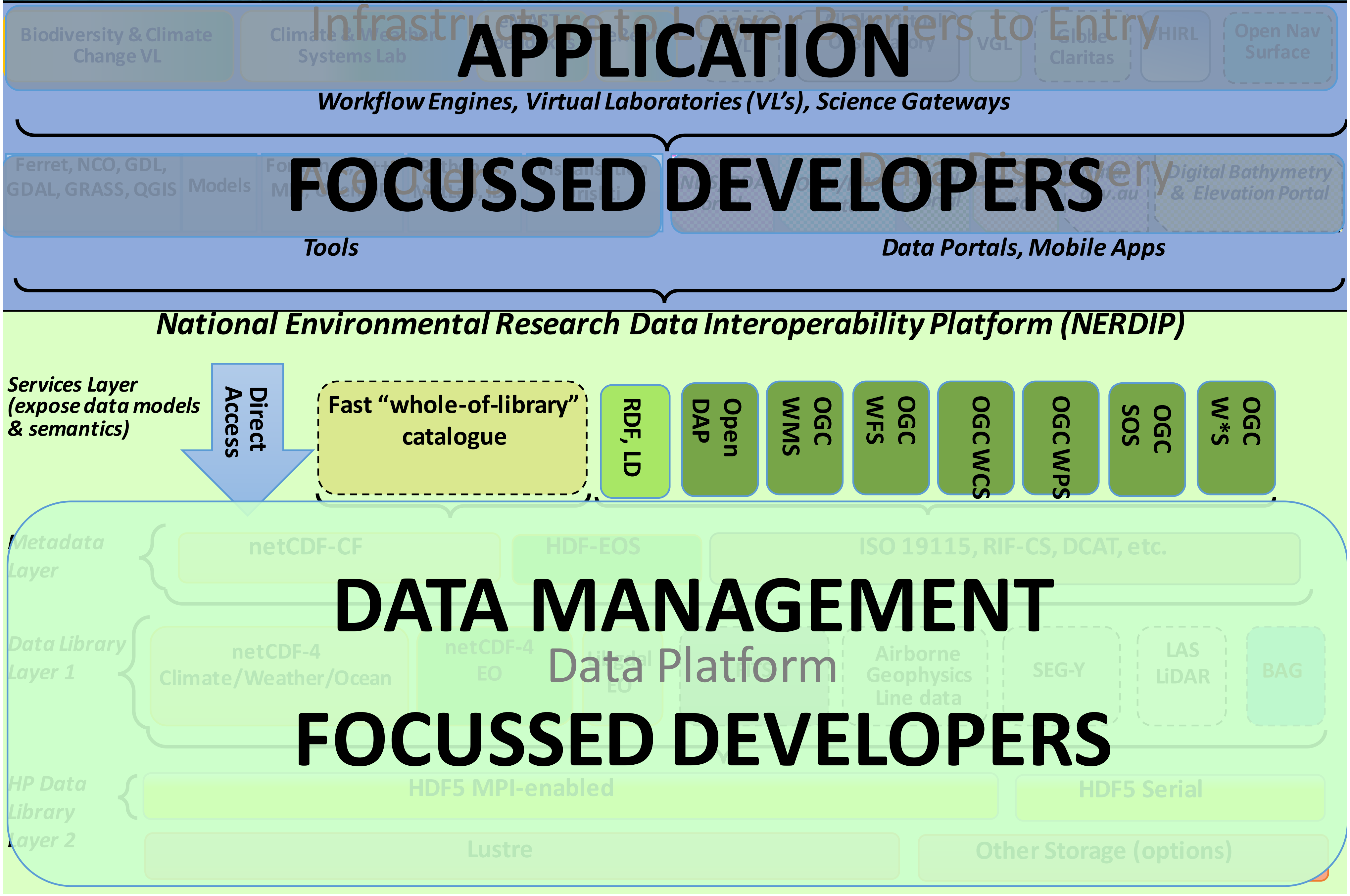
NERDIP: Enabling Ace Users to Interact with the Data



NERDIP: Applications Replicating Ways of Interacting with the Data



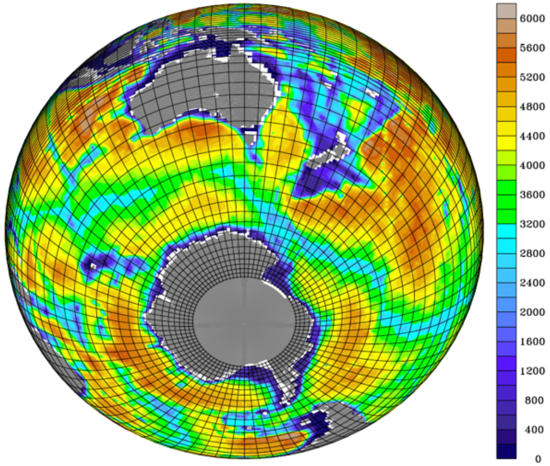
NERDIP: Loosely coupling Applications and Data via Services



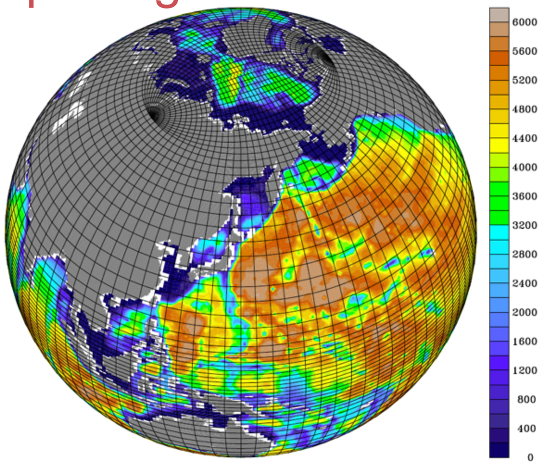
Grid Diversity in CMIP5

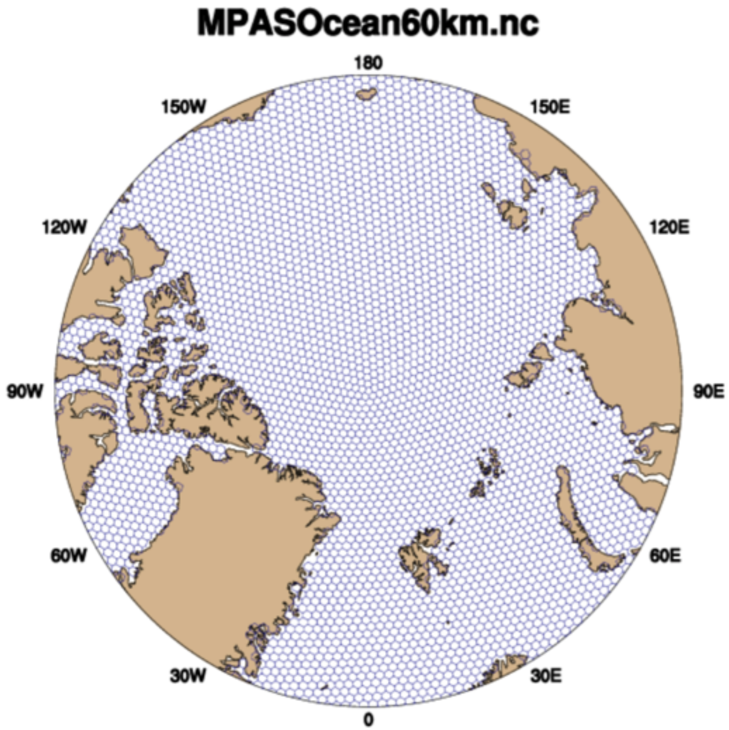
Downstream communities may not wish to deal with different grids, but the modelling communities generate data appropriate to them.

Mercator grid in south



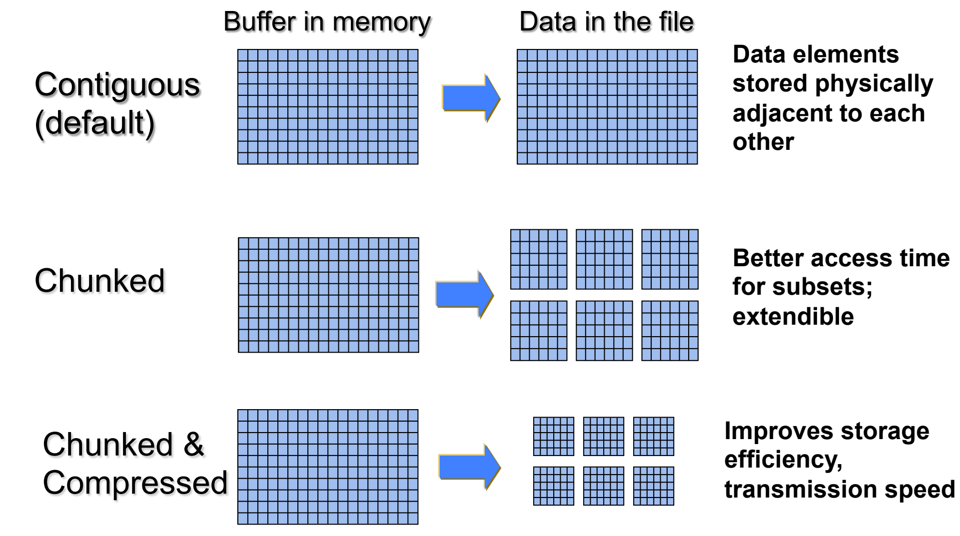
Tripolar grid in north





Data file layouts and performance analysis

- Global profiling tools focused on IO
- Compare to baselines
- Data is stored in chunks of predefined size
  - Two-dimensional instance may be referred to as data tiling
  - Matched chunking to cache size on the processor



Metrics	Serial IO	Parallel IO
IO interfaces		
GDAL/GeoTIFF	✔	
GDAL/NetCDF(4) Classic	✔	
NetCDF4/HDF5	✔	✔
MPIIO		✔
POSIX		✔
User application tuning		
Transfer size	✔	✔
File size	✔	✔
Subset selection	✔	✔
Concurrency		✔
Local access	✔	✔
Remote access DAP server	✔	

Metrics	Serial IO	Parallel IO
NetCDF4/HDF5 tuning		
Chunk pattern	✔	✔
Chunk cache	✔	✔
Compression	✔	✔
MPIIO tuning		
Independent & Collective		✔
Collective buffering		✔
Data sieving		✔
Lustre file system tuning		
Stripe count	✔	✔
Stripe size	✔	✔
IO profiling & tracing	✔	✔
total	14	17

Benchmark testing a Landsat scene in various file formats

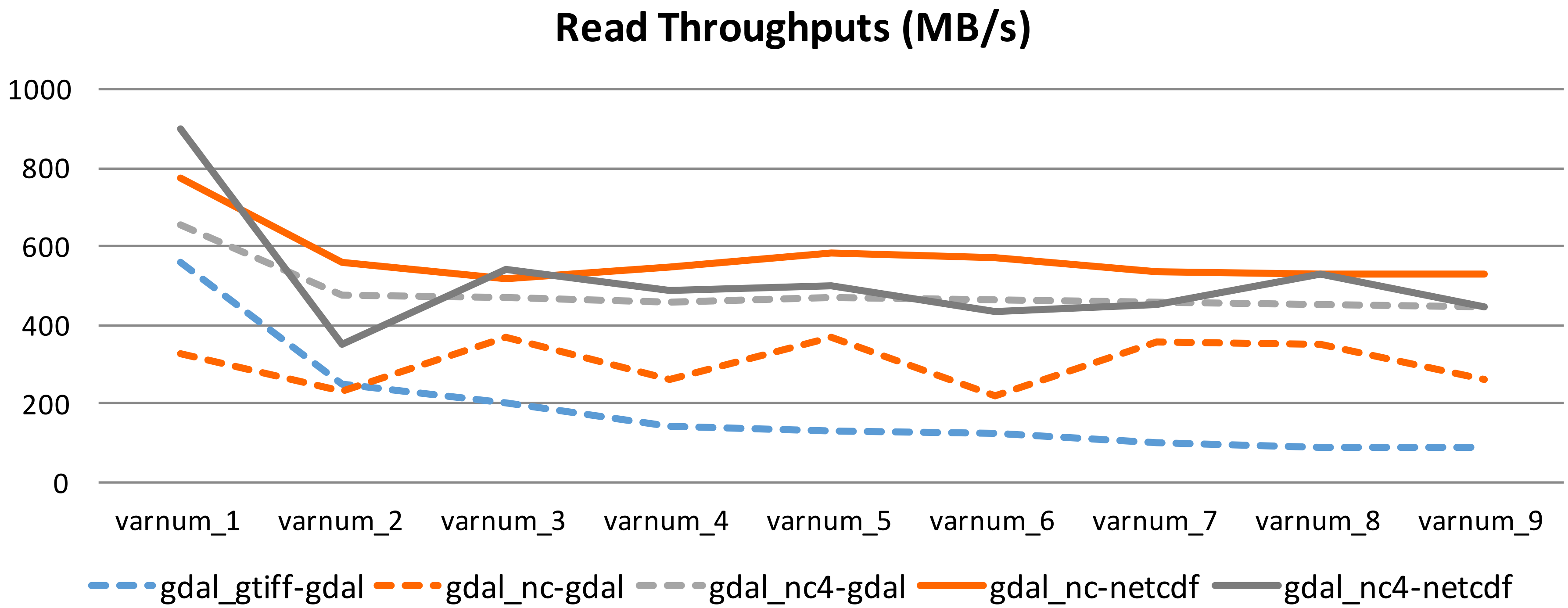
- Read Source File
  - LC80771182015023LGN00_B1.nc
  - Block size: 7771*7841
  - Data type: Short (2 Bytes)
  - Libraries: GDAL, NetCDF, HDF5
  - 1~9 Variables/Bands
- Write Target Files
  - Library: Formats
    - GDAL: GeoTiff, NetCDF Classic; NetCDF4, NetCDF Classic
    - NetCDF: NetCDF Classic, NetCDF4, NetCDF4 Classic
    - HDF5: HDF5
  - Data: 1~9 Bands
- IO Libraries
  - GDAL 2.0.2 (GTIFF,NC,NC4,NC4C) 2D array
  - NetCDF (4.4.0) (NC,NC4,NC4C) 2&3D array (for this study)
  - HDF5 (1.8.16) (NC4, HDF5) 2&3D array array (for this study)

	APIs		
Formats	GDAL	NETCDF	HDF5
GDAL created GTIFF (GDAL_GTIFF)	✔		
GDAL_NC	✔	✔	
GDAL_NC4C	✔	✔	✔
GDAL_NC4	✔	✔	✔
NC	✔	✔	
NC4C	✔	✔	✔
NC4	✔	✔	✔
HDF5	✔	✔	✔

c/- Rui Yang

Full access to 2D contiguous datasets

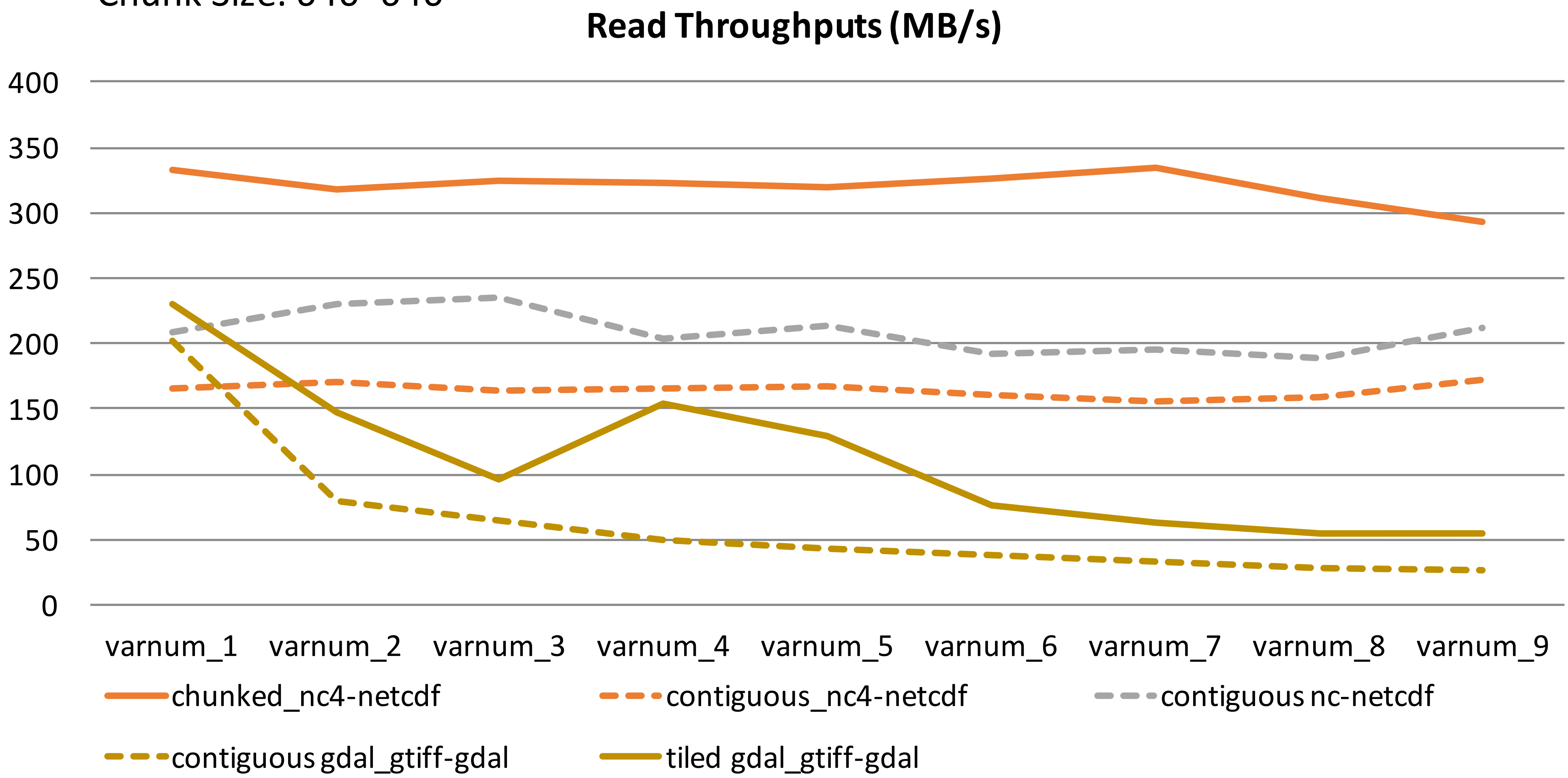
full dataset 7771*7841



- Geotiff performance impacted by number of variables (reads the whole file for each variable)
- GDAL creates overhead on NetCDF3 Classic file (requires additional mem_copy op.)
- GDAL and NetCDF/HDF5 library access NetCDF4 file with similar performance

Per Block Access of contiguous vs chunked datasets

- Subset size: 2560*2560
- Chunk Size: 640*640



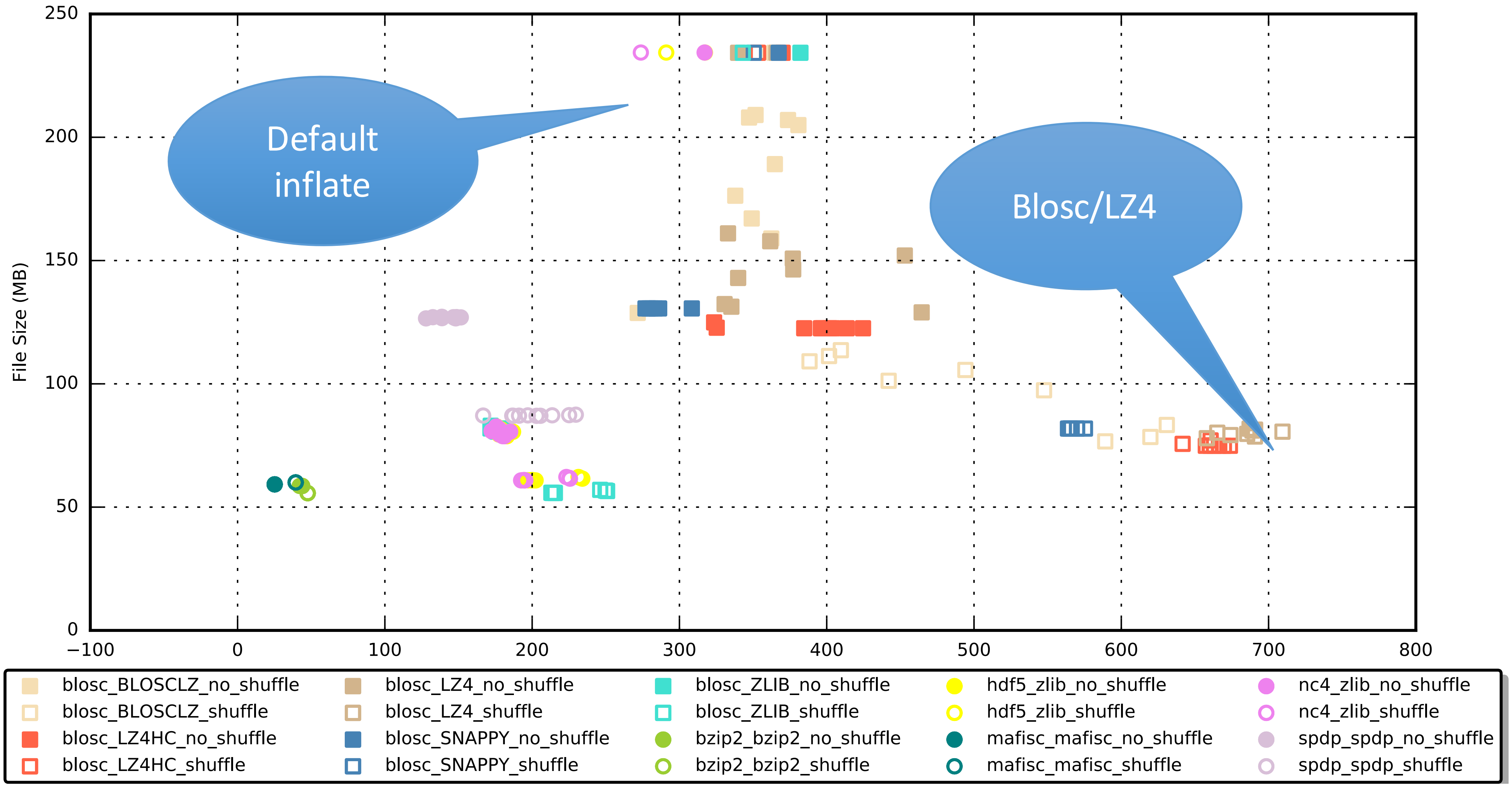
- Access is slower than full access to the previous benchmark of contiguous datasets.
- But … accessing chunked/tiled dataset is faster than contiguous dataset

Benchmark Configurations with Compression

Library	Default	Dynamic Filter
NetCDF4	Deflate (Zlib)	N/A
HDF5	Deflate (Zlib)	Bzip2,mafisc,spdp, Blosc(blosclz,lz4hc,lz4,SNAPPY,ZLIB)

Source File Attributes	Write Parameters	Read Parameters
File Name LC80990772015066LGN00.nc **Dataset** Band1 **Data type** float **Dimension (elements)** 7701*7591 **Dataset Size** 233,833,164 bytes **Chunk** 1*7591 **Shuffle** True **Deflate Level** 1	**Data Type** Float **Chunk** 1*1*7591 **Compression Level** 0-9 **Shuffle** Disabled/Enable **Compressor** As above	**Hyperslab** 1*1*7591 **Chunk Cache Size** 1MB **Shuffle** Blosc/Byte shuffle Blosc/Bit Shuffle **Compression Level** 0-9

Read Performance vs File Size



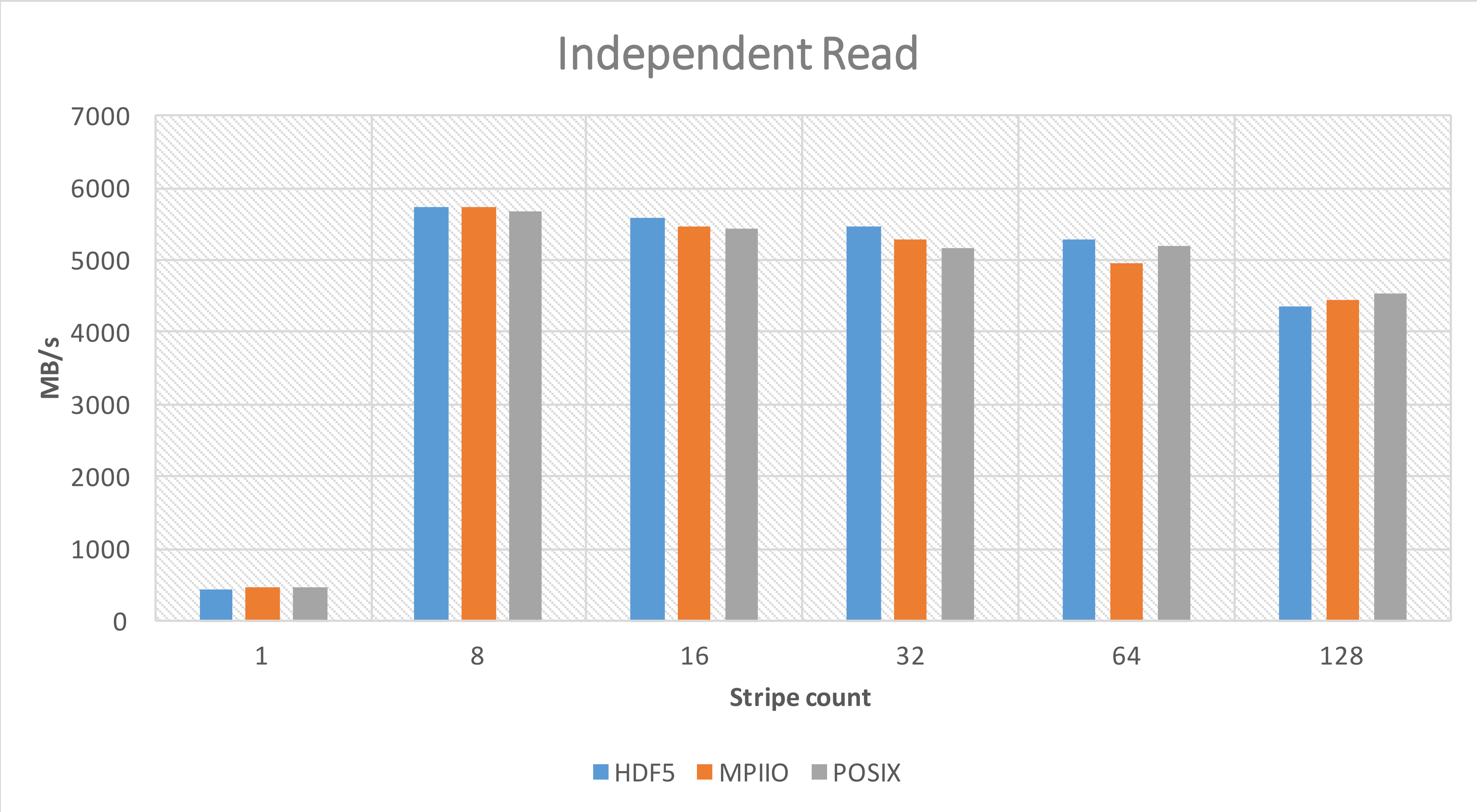
Another benchmark - internal data organisation

- Data Layout used to write file
  - Coordinate y, Coordinate x, Time t
  - Time t, Coordinate y, Coordinate
- Chunking
  - Along 2D (yx) or 3D (t,y,x)
- Read Access
  - Along yx or time t
  - Block subsets
  - Choose appropriate data layout and chunk shape to provide satisfied performance for any subset selection

Layout, Chunking and Subset

Layout	tyx (6,7851,7761)			yxt (7851,7761,6)		
Chunk size	(1,256,256)	(3,256,256)	(6,256,256)	(256,256,1)	(256,256,3)	(256,256,6)
Full access T=6,Y=7851,X=7761	469.52	597.01	691.31	179.58	399.82	783.90
Along yx T=1,Y=7851,X=7761	483.95	239.92	133.14	217.30	165.77	104.05
Along t T=6,Y=2048,X=2048	365.16	430.04	493.11	159.82	333.94	539.49
Chunk size	(1,512,512)	(3,512,512)	(6,512,512)	(512,512,1)	(512,512,3)	(512,512,6)
Full access T=6,Y=7851,X=7761	647.8	816.0	823.3	185.99	436.95	870.40
Along yx T=1,Y=7851,X=7761	607.01	267.71	150.78	267.55	164.02	110.47
Along t T=6,Y=2048,X=2048	408.26	679.47	642.62	173.13	400.51	710.93
Chunk size	(1,1024,1024)	(3,1024,1024)	(6,1024,1024)	(1024,10241)	(1024,1024,3)	(1024,1024,6)
Full access T=6,Y=7851,X=7761	776.78	720.51	738.95	191.02	391.51	811.89
Along yx T=1,Y=7851,X=7761	617.40	263.45	150.13	396.57	163.45	103.57
Along t T=6,Y=2048,X=2048	560.33	596.83	701.69	163.50	396.87	663.34

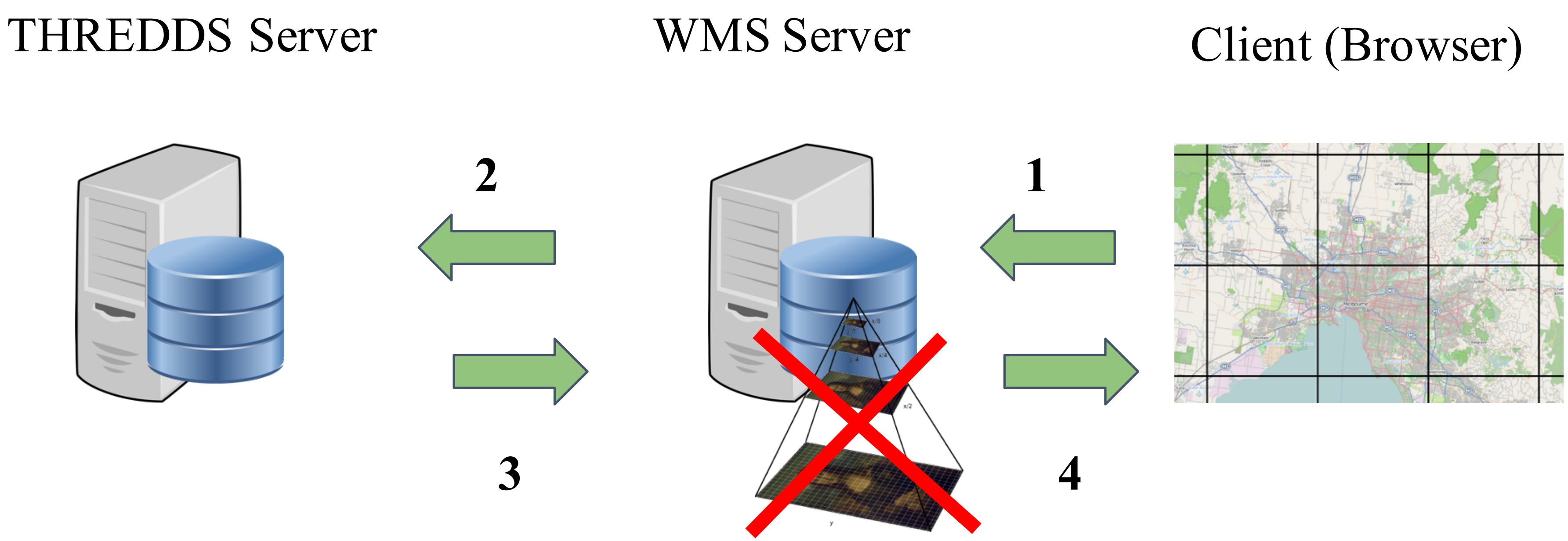
Parallel IO: HDF5 based on MPI-IO



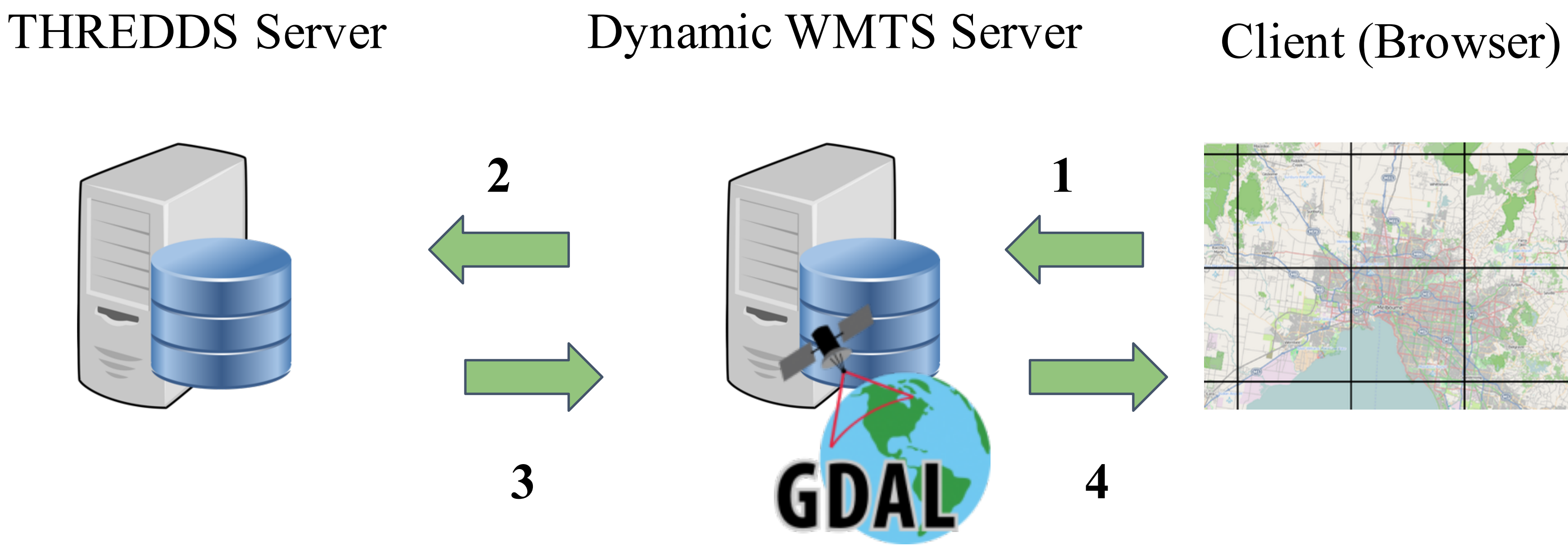
IOR Benchmark: MPI size = 16; Stripe size =1M; Block size = 8G; Transfer size = 32M;

Web Map Tile Servers instead of WMS

Serving Maps



Serving Maps



Reprojecting raster data on-the-fly from multiple satellites

Landsat8:
- 2015, 25 meters resolution, 11 Bands, revisit period 16 days
- UTM projection
- Original USGS L1T scenes packed in HDF5 (chunked & compressed)
- Local API and CEPH access

Himawari-8:
- 500, 1000, 2000 meters (depending on the band), 16 Bands, image every 10 mins
- Geostationary projection
- BoM NetCDF4 files
- Access through NCI TDS (THREDDS) subsetting

ERA Interim:
- 2015, 75 km resolution, 45 different atmospheric variables, one field every 3 hours
- WGS84 projection
- ECMWF netCDF4 files
- Local API and CEPH access

c/- Pablo Larraondo, Joseph Antony

Reprojecting raster data on-the-fly from multiple satellites