Bridging the gap between HPC and High Performance Data Analysis

Ben Evans
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/

<table>
<thead>
<tr>
<th>Data Collections</th>
<th>Approx. Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMIP5, CORDEX, ACCESS Models</td>
<td>5 Pbytes</td>
</tr>
<tr>
<td>Earth Obs: Himawari-8, LANDSAT, Sentinel, MODIS, INSAR</td>
<td>2 Pbytes</td>
</tr>
<tr>
<td>Digital Elevation, Bathymetry, Onshore/Offshore Geophysics</td>
<td>1 Pbytes</td>
</tr>
<tr>
<td>Seasonal Climate</td>
<td>700 Tbytes</td>
</tr>
<tr>
<td>Bureau of Meteorology Observations</td>
<td>350 Tbytes</td>
</tr>
<tr>
<td>Bureau of Meteorology Ocean-Marine</td>
<td>350 Tbytes</td>
</tr>
<tr>
<td>Terrestrial Ecosystem</td>
<td>290 Tbytes</td>
</tr>
<tr>
<td>Reanalysis products</td>
<td>100 Tbytes</td>
</tr>
</tbody>
</table>
NCI’s Integrated Scientific HPC/HPD Environment

Integrated HPC-HPD Environment

Puppet DevOps Framework

3000 Core Cloud

Data Services

THREDDS
ESGF
OPeNDAP
GeoServer

Server-side analysis and visualization

VDI: Cloud scale user desktops on data

Web-time analytics software

10PB+ Research Data

© National Computational Infrastructure 2016

Ben Evans, WGISS, March 2016

nci.org.au
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’s National Environmental Data Interoperability Research Platform (NERDIP)

Workflow Engines, Virtual Laboratories (VL’s), Science Gateways

Tools

Data Portals, Mobile Apps

Fast “whole-of-library” catalogue

Direct Access

Metadata Layer

Services Layer (expose data models & semantics)

Data Library Layer 1

Data Library Layer 2

HP Data Library Layer 2

Lustre

Other Storage (options)

HDF5 MPI-enabled

HDF5 Serial

ISO 19115, RIF-CS, DCAT, etc.

FITS

Airborne Geophysics Line data

SEG-Y

LAS LiDAR

BAG

NetCDF-4 Climate/Weather/Ocean

NetCDF-4 EO

Libgdal EO

HDF5 MPI-enabled

HDF5 Serial

Lustre

Other Storage (options)

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

NetCDF

HDF

HDF-EO

NetCDF-CF

HDF-EOS

RDF, LD

Open DAP

OGC WCS

OGC WMS

OGC WFS

OGC WPS

OGC SOS

OGC WS

Other Storage (options)

American Environmental Data Interoperability Research Platform (NERDIP)

Workflow Engines, Virtual Laboratories (VL’s), Science Gateways

Tools

Data Portals, Mobile Apps

Fast “whole-of-library” catalogue

Direct Access

Metadata Layer

Services Layer (expose data models & semantics)

Data Library Layer 1

Data Library Layer 2

HP Data Library Layer 2

Lustre

Other Storage (options)

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

NetCDF

HDF

HDF-EO

NetCDF-CF

HDF-EOS

RDF, LD

Open DAP

OGC WCS

OGC WMS

OGC WFS

OGC WPS

OGC SOS

OGC WS

Other Storage (options)
NERDIP: Enabling Multiple Ways to Interact with the Data

Infrastructure to Lower Barriers to Entry

Workflow Engines, Virtual Laboratories (VL’s), Science Gateways

Ace Users

Tools

Data Portals

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

DAP

Open

OGC

WFS

OGC WCS

OGC WPS

OGC SOS

ISO 19115, RIF-CS, DCAT, etc.

Metadata Layer

Data Platform

NetCDF-CF

HDF-EOS

NetCDF-4 Climate/Weather/Ocean

NetCDF-4 EO

Libgdal EO

FITS

Airborne Geophysics Line data

SEG-Y

LAS LiDAR

BAG

HDF5 MPI-enabled

HDF5 Serial

Lustre

Other Storage (options)

Digital Bathymetry & Elevation Portal

Other Storage (options)

Workflow Engines,
Virtual Laboratories (VL’s),
Science Gateways

Ace Users

Tools

Data Portals

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

DAP

Open

OGC

WFS

OGC WCS

OGC WPS

OGC SOS

ISO 19115, RIF-CS, DCAT, etc.

Metadata Layer

Data Platform

NetCDF-CF

HDF-EOS

NetCDF-4 Climate/Weather/Ocean

NetCDF-4 EO

Libgdal EO

FITS

Airborne Geophysics Line data

SEG-Y

LAS LiDAR

BAG

HDF5 MPI-enabled

HDF5 Serial

Lustre

Other Storage (options)

Digital Bathymetry & Elevation Portal

Other Storage (options)
NERDIP: Enabling Ace Users to Interact with the Data

Infrastructure to Lower Barriers to Entry

Workflow Engines, Virtual Laboratories (VL’s), Science Gateways

Data Portals

Tools

Fast “whole-of-library” catalogue

Direct Access

Services Layer (expose data models & semantics)

Metadata Layer

Data Library Layer 1

Data Library Layer 2

HP Data Library

HDF5 MPI-enabled

Lustre

Data Platform

ISO 19115, RIF-CS, DCAT, etc.

FITS

Airborne Geophysics Line data

SEG-Y

LAS

LIDAR

BAG

Digital Bathymetry & Elevation Portal

Other Storage (options)

Data Portals, Mobile Apps

NERDIP:
Enabling Ace Users to Interact with the Data

Data Portals, Mobile Apps

Metadata Layer

HDF

HDF-EOS

CDF-CF

NetCDF-4

HDF5 MPI-enabled

Lustre

Data Platform

ISO 19115, RIF-CS, DCAT, etc.

FITS

Airborne Geophysics Line data

SEG-Y

LAS

LIDAR

BAG

Digital Bathymetry & Elevation Portal

Other Storage (options)

Data Portals, Mobile Apps
NERDIP: Applications Replicating Ways of Interacting with the Data

Workflow Engines, Virtual Laboratories (VL's), Science Gateways

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

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

Data Library Layer 2

HDF5 MPI-enabled

HP Data Library Layer 2

Lustre

HDF5 Serial

Other Storage (options)

Ferret, NCO, GDL, GDAL, GRASS, QGIS
Models
Fortran, C, C++, MPI, OpenMP
Python, R, MatLab, IDL
Visualisation Drishti

AGDC VL
All Sky Virtual Observatory
VGL
Globe Claritas
VHIRL
Open Nav Surface

HDF5 MPI-enabled

HDF5 Serial

FITS

Airborne Geophysics Line data

SEG-Y

LAS LiDAR

BAG

LAS

LiDAR

Lustre

Other Storage (options)

Ferret,
NCO,
GDL,
GDAL,
GRASS,
QGIS

Fortran,
C,
C++,
MPI,
OpenMP

Python,
R,
MatLab,
IDL

Visualisation
Drishti

AGDC VL
All Sky Virtual
Observatory
VGL
Globe
Claritas
VHIRL
Open Nav Surface

HDF5 MPI-enabled

HDF5 Serial

FITS

Airborne Geophysics Line data

SEG-Y

LAS LiDAR

BAG

HDF5 Serial

NetCDF

- CF

HDF

- EOS

ISO 19115,
RIF-CS,
DCAT,
etc.

VGL

Globe Claritas

VHIRL

Open Nav Surface

Data Portals, Mobile Apps

National Environmental Research Data Interoperability Platform (NERDIP)

Services Layer (expose data models & semantics)

Direct Access

Fast “whole-of-library” catalogue

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

Data Library Layer 2

HDF5 MPI-enabled

HP Data Library Layer 2

Lustre

HDF5 Serial

Other Storage (options)
NERDIP: Loosely coupling Applications and Data via Services

Infrastructure to Lower Barriers to Entry

APPLICATION

Workflow Engines, Virtual Laboratories (VL’s), Science Gateways

FOCUSED DEVELOPERS

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

DAP

Open

OGC

WFS

OGC WCS

OGC WPS

OGC SOS

OGC W*S

DATA MANAGEMENT

Data Platform

FOCUSED DEVELOPERS

Metadata Layer

Data Library Layer 1

Data Library Layer 2

HP Data Library Layer 2

NetCDF-CF

HDF-EOS

ISO 19115, RIF-CS, DCAT, etc.

HDF5 MPI-enabled

Airborne Geophysics Line data

SEG-Y

LIDAR

BAG

NetCDF-4 Climate/Weather/Ocean

NetCDF-4 EO

HDF5 Serial

Lustre

Other Storage (options)
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
• 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

<table>
<thead>
<tr>
<th>Data file layouts and performance analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Contiguous</strong> (default)</td>
</tr>
<tr>
<td>Buffer in memory</td>
</tr>
<tr>
<td>Data in the file</td>
</tr>
<tr>
<td>Data elements stored physically adjacent to each other</td>
</tr>
</tbody>
</table>

| **Chunked**                                |
| Buffer in memory                          |
| Data in the file                           |
| Better access time for subsets; extendible |

| **Chunked & Compressed**                   |
| Buffer in memory                          |
| Data in the file                           |
| Improves storage efficiency, transmission speed |

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Serial IO</th>
<th>Parallel IO</th>
</tr>
</thead>
<tbody>
<tr>
<td>IO interfaces</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GDAL/GeoTIFF</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>GDAL/NetCDF(4) Classic</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>NetCDF4/HDF5</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User application tuning</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Transfer size</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>File size</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Subset selection</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Concurrency</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Local access</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Remote access DAP server</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Serial IO</th>
<th>Parallel IO</th>
</tr>
</thead>
<tbody>
<tr>
<td>NetCDF4/HDF5 tuning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chunk pattern</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Chunk cache</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Compression</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

| MPI/O 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)

---

c/- Rui Yang

© National Computational Infrastructure 2016

Ben Evans, WGISS, March 2016
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

Read Throughputs (MB/s)

- 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

<table>
<thead>
<tr>
<th>Library</th>
<th>Default</th>
<th>Dynamic Filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>NetCDF4</td>
<td>Deflate (Zlib)</td>
<td>N/A</td>
</tr>
<tr>
<td>HDF5</td>
<td>Deflate (Zlib)</td>
<td>Bzip2, mafisc, spdtp, Blosc (blosclz,lz4hc,lz4,SNAPPY,ZLIB)</td>
</tr>
</tbody>
</table>

### Source File Attributes

<table>
<thead>
<tr>
<th>File Name</th>
<th>LC80990772015066LGN00.nc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>Band1</td>
</tr>
<tr>
<td>Data type</td>
<td>float</td>
</tr>
<tr>
<td>Dimension (elements)</td>
<td>7701*7591</td>
</tr>
<tr>
<td>Dataset Size</td>
<td>233,833,164 bytes</td>
</tr>
<tr>
<td>Chunk</td>
<td>1*7591</td>
</tr>
<tr>
<td>Shuffle</td>
<td>True</td>
</tr>
<tr>
<td>Deflate Level</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Write Parameters</th>
<th>Read Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Type</strong></td>
<td><strong>Hyperslab</strong></td>
</tr>
<tr>
<td>Float</td>
<td>1<em>1</em>7591</td>
</tr>
<tr>
<td><strong>Chunk</strong></td>
<td><strong>Chunk Cache Size</strong></td>
</tr>
<tr>
<td>1<em>1</em>7591</td>
<td>1MB</td>
</tr>
<tr>
<td><strong>Compression Level</strong></td>
<td><strong>Shuffle</strong></td>
</tr>
<tr>
<td>0-9</td>
<td>Blosc/Byte shuffle</td>
</tr>
<tr>
<td><strong>Shuffle</strong></td>
<td><strong>Blosc/Bit Shuffle</strong></td>
</tr>
<tr>
<td>Disabled/Enable</td>
<td>Blosc/Bit Shuffle</td>
</tr>
<tr>
<td><strong>Compressor</strong></td>
<td><strong>Compression Level</strong></td>
</tr>
<tr>
<td>As above</td>
<td>0-9</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Layout</th>
<th>tyx  (6,7851,7761)</th>
<th>yxt (7851,7761,6)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chunk size</strong></td>
<td>(1,256,256)</td>
<td>(256,256,1)</td>
</tr>
<tr>
<td>Full access</td>
<td>469.52</td>
<td>179.58</td>
</tr>
<tr>
<td>T=6,Y=7851,X=7761</td>
<td>597.01</td>
<td>399.82</td>
</tr>
<tr>
<td>Along yx</td>
<td>483.95</td>
<td>217.30</td>
</tr>
<tr>
<td>T=1,Y=7851,X=7761</td>
<td>239.92</td>
<td>165.77</td>
</tr>
<tr>
<td>Along t</td>
<td>365.16</td>
<td>159.82</td>
</tr>
<tr>
<td>T=6,Y=2048,X=2048</td>
<td>430.04</td>
<td>333.94</td>
</tr>
<tr>
<td><strong>Chunk size</strong></td>
<td>(1,512,512)</td>
<td>(512,512,1)</td>
</tr>
<tr>
<td>Full access</td>
<td>647.8</td>
<td>185.99</td>
</tr>
<tr>
<td>T=6,Y=7851,X=7761</td>
<td>816.0</td>
<td>436.95</td>
</tr>
<tr>
<td>Along yx</td>
<td>607.01</td>
<td>267.55</td>
</tr>
<tr>
<td>T=1,Y=7851,X=7761</td>
<td>267.71</td>
<td>164.02</td>
</tr>
<tr>
<td>Along t</td>
<td>408.26</td>
<td>173.13</td>
</tr>
<tr>
<td>T=6,Y=2048,X=2048</td>
<td>679.47</td>
<td>400.51</td>
</tr>
<tr>
<td><strong>Chunk size</strong></td>
<td>(1,1024,1024)</td>
<td>(1024,1024,1)</td>
</tr>
<tr>
<td>Full access</td>
<td>776.78</td>
<td>191.02</td>
</tr>
<tr>
<td>T=6,Y=7851,X=7761</td>
<td>720.51</td>
<td>391.51</td>
</tr>
<tr>
<td>Along yx</td>
<td>617.40</td>
<td>396.57</td>
</tr>
<tr>
<td>T=1,Y=7851,X=7761</td>
<td>263.45</td>
<td>163.45</td>
</tr>
<tr>
<td>Along t</td>
<td>560.33</td>
<td>163.50</td>
</tr>
<tr>
<td>T=6,Y=2048,X=2048</td>
<td>596.83</td>
<td>396.87</td>
</tr>
</tbody>
</table>
IOR Benchmark: MPI size = 16; Stripe size = 1M; Block size = 8G; Transfer size = 32M;
Serving Maps

THREDDS Server  WMS Server  Client (Browser)

1  2  3  4

Web Map Tile Servers instead of WMS
Serving Maps

THREDDS Server → Dynamic WMTS Server → Client (Browser)

1. THREDDS Server
2. Dynamic WMTS Server
3. GDAL
4. Client (Browser)
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