Emerging Cyber Infrastructure for NASA’s Large-Scale Climate Data Analytics

AGU
12 December 2016

Daniel Duffy¹, Carrie Spear¹, Michael Bowen¹, John Thompson¹, Fei Hu², Chaowei Yang², Daniel Pierce¹

¹NASA Center for Climate Simulation (NCCS), NASA Goddard Space Flight Center, Greenbelt, MD, USA and ²George Mason University
Provides an integrated high-end computing environment designed to support the specialized requirements of Climate and Weather modeling.

- High-performance computing, cloud computing, data storage, and networking technologies
- High-speed access to petabytes of Earth Science data
- Collaborative data sharing, publication, and analysis services

**Primary Customers (NASA Science)**

- NASA funded science projects can get access to these resources
- Global Modeling and Assimilation Office (GMAO)
- Land Information Systems (LIS)
- Goddard Institute for Space Studies (GISS)
- Variety of other Research and Development (R&D) and Engineering
  » ABoVE, HiMAT, CALET, WFIRST

**High-Performance Science**

- [http://www.nccs.nasa.gov](http://www.nccs.nasa.gov)
- Funded by the High End Computing (HEC) program under SMD
  » Dr. Tsengdar Lee, Program Manager
- Code 606.2 at NASA Goddard Space Flight Center in Greenbelt, MD.
FV3 Dynamical Core uses a Cubed-Sphere which maps the Earth onto faces of a cube
- There are 6 faces of the cube and multiple vertical layers
- Total number of grid points
  - X * Y * Z * 6 Faces of the Cube

Current GMAO Research
- Operational research forecasts are running at 27 KM resolution using about 27 million grid points
- Target operational research forecasts at a resolution of 12 KM in the very near future
- Reanalysis (including chemistry)
- Nature Runs at 7 KM and 3.5 KM
- Dynamic downscaling of reanalysis and forecasts down to 6 KM
- Highest resolution research runs are at 1.5 KM global resolution
6 KM GEOS-5 Outgoing Longwave Radiation (OLR) (Global Modeling and Assimilation Office)
Increasing the GEOS-5 Model Resolution for Research

<table>
<thead>
<tr>
<th>X and Y Values</th>
<th>Grid Points</th>
<th>Resolution (Meters)</th>
<th>Cores</th>
<th>Memory (GB)</th>
<th>1 Simulation Year of Data (PB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>360</td>
<td>1.02E+08</td>
<td>27,776</td>
<td>118</td>
<td>0.50</td>
<td>0.13</td>
</tr>
<tr>
<td>720</td>
<td>4.06E+08</td>
<td>13,888</td>
<td>472</td>
<td>2.00</td>
<td>0.50</td>
</tr>
<tr>
<td>1,440</td>
<td>1.63E+09</td>
<td>6,944</td>
<td>1,888</td>
<td>8.00</td>
<td>2.00</td>
</tr>
<tr>
<td>2,880</td>
<td>6.50E+09</td>
<td>3,472</td>
<td>7,552</td>
<td>32.00</td>
<td>8.00</td>
</tr>
<tr>
<td>5,760</td>
<td>2.60e+10</td>
<td>1,736</td>
<td>30,208</td>
<td>128.00</td>
<td>32.00</td>
</tr>
<tr>
<td>11,520</td>
<td>1.04E+11</td>
<td>868</td>
<td>120,832</td>
<td>512.00</td>
<td>128.00</td>
</tr>
<tr>
<td>23,040</td>
<td>4.16E+11</td>
<td>434</td>
<td>483,328</td>
<td>2,048.00</td>
<td>512.00</td>
</tr>
<tr>
<td>46,080</td>
<td>1.664E+12</td>
<td>217</td>
<td>1,933,312</td>
<td>8,192.00</td>
<td>2,048.00</td>
</tr>
<tr>
<td>92,160</td>
<td>6.656E+12</td>
<td>109</td>
<td>7,733,248</td>
<td>32,768.00</td>
<td>8,192.00</td>
</tr>
</tbody>
</table>

The data sets needed for science are getting too big to move outside of the computing center or to simply serve through traditional data services.
Recent NCCS Evolution of Major Systems

In addition to increasing compute and storage, the NCCS is evolving its services to perform *in-situ, large-scale data analytics*. 

FY15: Data Portal, Mass Storage, HPC - Discover

FY16: ADAPT, Mass Storage, HPC - Discover

FY17: ADAPT, DASS, HPC - Discover

FY16: Creation of a High Performance Science cloud (virtual environment) designed for traditional data services, data analytics, and web services: move the data to the analysis.

FY17: Creation of the Data Analytics Storage Service (DASS), a combined High Performance Computing and Data environment to enable emerging analytics: move the analytics to the data.
Climate Analytics as a Service (CAaaS)

Data
Relevance and Collocation
Data have to be significant, sufficiently complex, and physically or logically co-located to be interesting and useful …

Convenient and Extensible
Capabilities need to be easy to use and facilitate community engagement and adaptive construction …

Model Data Sets
• Forecasts
• Seasonal Forecasts
• Reanalyses
• Nature Runs
• Climate Data

Observation Data Sets
• Science Driven

Interfaces
• APIs
• Web Services
• Python Notebooks
• Zeppelin Notebooks

High-Performance Compute/Storage Fabric
Storage-proximal analytics with simple canonical operations
Data do not move, analyses need horsepower, and leverage requires something akin to an analytical assembly language …

DASS
• 1,000’s of cores
• TF’s of compute
• PB’s of storage
• High Speed Networks
• Operational Spring 2017

Exposure

Data Sets
• Forecasts
• Seasonal Forecasts
• Reanalyses
• Nature Runs
• Climate Data

MERRA Analytic Services: Meeting the Big Data challenges of climate science through cloud-enabled Climate Analytics-as-a-Service, Schnase, et al., Computers, Environment, and Urban Systems

Emerging Cyberinfrastructure for NASA's Large-Scale Climate Data Analytics
Data Analytics Storage System (DASS)

Data movement and sharing of data across services within the NCCS is still a challenge

Large data sets created on Discover (HPC)
- On which users perform many analyses
- And may not be in a NASA Distributed Active Archive Center (DAAC)

Create a true centralized combination of storage and compute capability
- Capacity to store many PBs of data for long periods of time
- Architected to be able to scale both horizontally (compute and bandwidth) and vertically (storage capacity)
- Can easily share data to different services within the NCCS
- Free up high speed disk capacity within Discover
- Enable both traditional and emerging analytics
- No need to modify data; use native scientific formats
DASS Concept

Read access from all nodes within the ADAPT system
- Serve to data portal, web services (FTP, HTTP, OpenDAP, ESGF, etc.)
- Serve data to virtual machines for additional processing
- Mixing model and observations

Analysis request is sent to a service.

ADAPT

Climate Analytics as a Service

DASS (~20 PB)

Mass Storage

Read and write access from the mass storage
- Stage data into and out of the centralized storage environment as needed

HPC - Discover

Source of Data: Write and Read from all nodes within Discover – models write data into GPFS which is then staged into the centralized storage (burst buffer like).

Note that all the services will still have local file systems to enable local optimized writes and reads as needed within their respective security domains.
Initial DASS Capability Overview

- Initial Capacity
  - 20.832 PB Raw Data Storage
  - 2,604 by 8TB SAS Drives
  - 14 HPE Apollo 4520 units in 5 racks (2 servers per unit)
  - 28 Servers with a total of 896 Cores
  - 14,336 GB Memory
  - 16 GB/Core
  - 37 TF of compute
  - 10 GbE and 40 GbE networks
- Roughly equivalent to the compute capacity of the NCCS just 6 years ago!
- Designed to easily scale both horizontally (compute) and vertically (storage)
DASS Software Stack

**Traditional**
Data moved from storage to compute.

- POSIX Interface
- RDMA and TCP/IP
- Shared Parallel File System (GPFS)

**Native Scientific Data** stored in HPC Storage or Commodity Servers and Storage

**Emerging**
Analytics moved from servers to storage.

- MapReduce, Spark, Machine Learning, etc.
- RESTful Interface, Custom APIs, Notebooks
- Cloudera and SIA
- Shared Parallel File System (GPFS) & Hadoop Connector

Open Source Software Stack on DASS Servers
- Centos Operating System
- Software RAID
- Linux Storage Enclosure Services
- Failover: Pacemaker, Corasync, Stoneth

Emerging Cyberinfrastructure for NASA's Large-Scale Climate Data Analytics
Spatiotemporal Index Approach (SIA) and Hadoop

Use what we know about the structured scientific data
Create a spatiotemporal query model to connect the array-based data model with the key-value based MapReduce programming model using grid concept

Build a spatiotemporal index to
- Link the logical to physical location of the data
- Make use of an array-based data model within HDFS
- Developed a grid partition strategy
- Keep high data locality for each map task
- Balance the workload across cluster nodes

A spatiotemporal indexing approach for efficient processing of big array-based climate data with MapReduce
Zhenlong Lia, Fei Hua, John L. Schnase, Daniel Q. Duffy, Tsengdar Lee, Michael K. Bowen and Chaowei Yang
http://dx.doi.org/10.1080/13658816.2015.1131830

Emerging Cyberinfrastructure for NASA's Large-Scale Climate Data Analytics
Analytics Infrastructure Testbed

Built three test clusters using decommissioned HPC servers and networks.

- **Test Cluster 1**
  - SIA
  - Cloudera
  - HDFS
  - 20 nodes (compute and storage)
  - Sequenced data
  - Native NetCDF data
    - Put only

- **Test Cluster 2**
  - SIA
  - Cloudera
  - Hadoop Connector
  - GPFS
  - 20 nodes (compute and storage)
  - Sequenced data
    - Put and Copy
  - Native NetCDF Data
    - Put and Copy

- **Test Cluster 3**
  - SIA
  - Cloudera
  - Hadoop Connector
  - Lustre
  - 20 nodes (compute and storage)
  - Sequenced data
    - Put and Copy
  - Native NetCDF Data
    - Put and Copy
Comparison of Performance HDFS versus GPFS/HDFS

- Compute the average temperature for every grid point (x, y, and z)
- Vary by the total number of years
- Sequenced MERRA Monthly Means (Reanalysis)
  - Not using SIA
  - Approximately 20 GB of data
- Compare the performance of the same MapReduce query on
  - Cloudera HDFS
  - GPFS and Hadoop Connector
- Performance is very similar
Initial Serial Performance with SIA

- Compute the average temperature for every grid point \((x, y, \text{ and } z)\)
- Vary by the total number of years
- MERRA Monthly Means (Reanalysis)
- Comparison of serial c-code to MapReduce code
- Comparison of traditional HDFS (Hadoop) where data is sequenced (modified) with GPFS where data is native NetCDF (unmodified, copy)
- Using unmodified data in GPFS with MapReduce is the fastest
- Only showing GPFS results to compare against HDFS
Initial Parallel Performance with SIA

- Compute the average temperature for every grid point \((x, y, \text{ and } z)\)
- Vary by the total number of years
- MERRA Monthly Means (Reanalysis)
- Comparison of serial c-code with MPI to MapReduce code
- Comparison of traditional HDFS (Hadoop) where data is sequenced (modified) with GPFS where data is native NetCDF (unmodified, copy)
- Again using unmodified data in GPFS with MapReduce is the fastest as the number of years increases
- Only showing GPFS results to compare against HDFS

Performance Comparison of Parallel MPI C-Code versus HDFS with 42 reducers
Future of HPC and Big Data at Exascale

**ADAPT**
Virtual Environment
HPC and Cloud
Existing Size
~1,000 cores
~10 PB of storage

Designed for Big Data Analytics

**Mass Storage**
Tiered Storage
Disk and Tape
Existing Size
~10 to 100 PB of storage

Designed for long-term storage and recall; not compute

**DASS**
Tiered Storage
Memory, SSD, Disk
Existing Size
~20 PB of storage

Designed for both compute and longer term storage.

**HPC/Discover**
HPC Cluster
Existing Size
~100,000 cores
~50 PB of storage

Designed for Large-Scale Climate Simulations

**Future Exascale Environment**
Merging of HPC and Big Data Analytics Capabilities

Ability for in-situ analytics throughout the environment … known analytics and machine learning