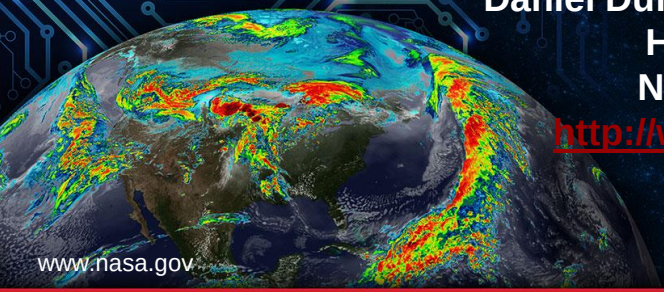
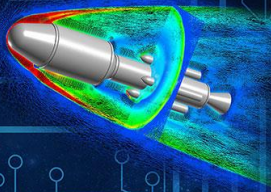
National Aeronautics and Space Administration

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

24 October 2016

Daniel Duffy (daniel.q.duffy@nasa.gov and @dqduffy)
High Performance Computing Lead
NASA Center for Climate Simulation
http://www.nccs.nasa.gov and @NASA_NCCS





Acknowledgements

Thank you to the conference organizing committee to allow us the opportunity to discuss our ongoing work with enabling data analytics at NASA Goddard.

Thank you to the people that do the real work!

- Carrie Spear/NCCS GSFC
- Hoot Thompson/NCCS GSFC
- Michael Bowen/NCCS GSFC
- John Schnase/GSFC
- Glenn Tamkin/NCCS GSFC
- Phil Yang/GMU
- Fei Hua/GMU
- Dan'l Pierce/NCCS GSFC
- Dave Kemeza/NCCS GSFC
- Garrison Vaughan/NCCS GSFC
- Scott Sinno/NCCS GSFC
- Phil Webster/GSFC
- Special thanks to our funding manager, Dr. Tsengdar Lee/NASA HQ.

NASA High-End Computing Program

HEC Program Office
NASA Headquarters
Dr. Tsengdar Lee
Scientific Computing Portfolio Manager
http://www.hec.nasa.gov/

High-End Computing Capability (HECC) Project
NASA Advanced Supercomputing (NAS)
NASA Ames
Dr. Piyush Mehrotra
https://www.nas.nasa.gov/hecc/

NASA Center for Climate Simulation (NCCS)
Goddard Space Flight Center (GSFC)
Dr. Daniel Duffy
http://www.nccs.nasa.gov/

NASA Center for Climate Simulation (NCCS)
High Performance Science

Provides an integrated high-end computing environment designed to support the specialized requirements of Climate and Weather modeling.

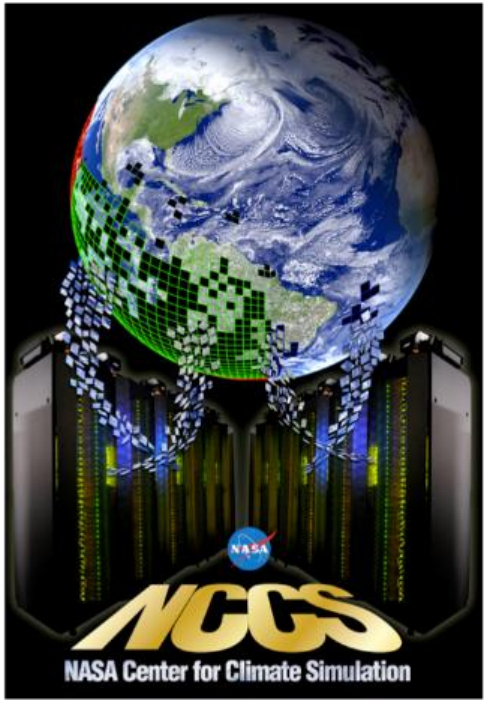
- High-performance computing, data storage, and networking technologies
- High-speed access to petabytes of Earth Science data
- Collaborative data sharing and publication services
- Advanced Data Analytics Platform (ADAPT)

Primary Customers (NASA Climate Science)

- Global Modeling and Assimilation Office (GMAO)
- Goddard Institute for Space Studies (GISS)

High-Performance Science

- http://www.nccs.nasa.gov
- Code 606.2
- Located in Building 28 at Goddard



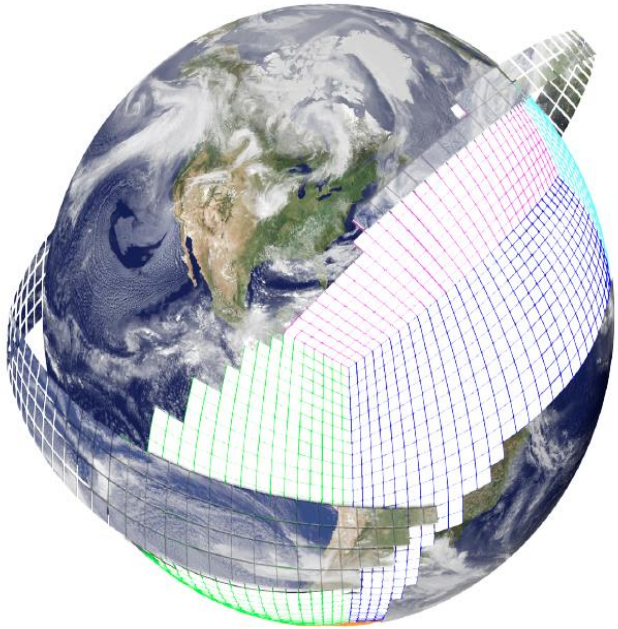
Goddard Earth Observing System (GEOS) Model
NASA Global Modeling and Assimilation Office (GMAO)

FV3 Dynamical Core uses a Cubed-Sphere which maps the Earth onto faces of a cube

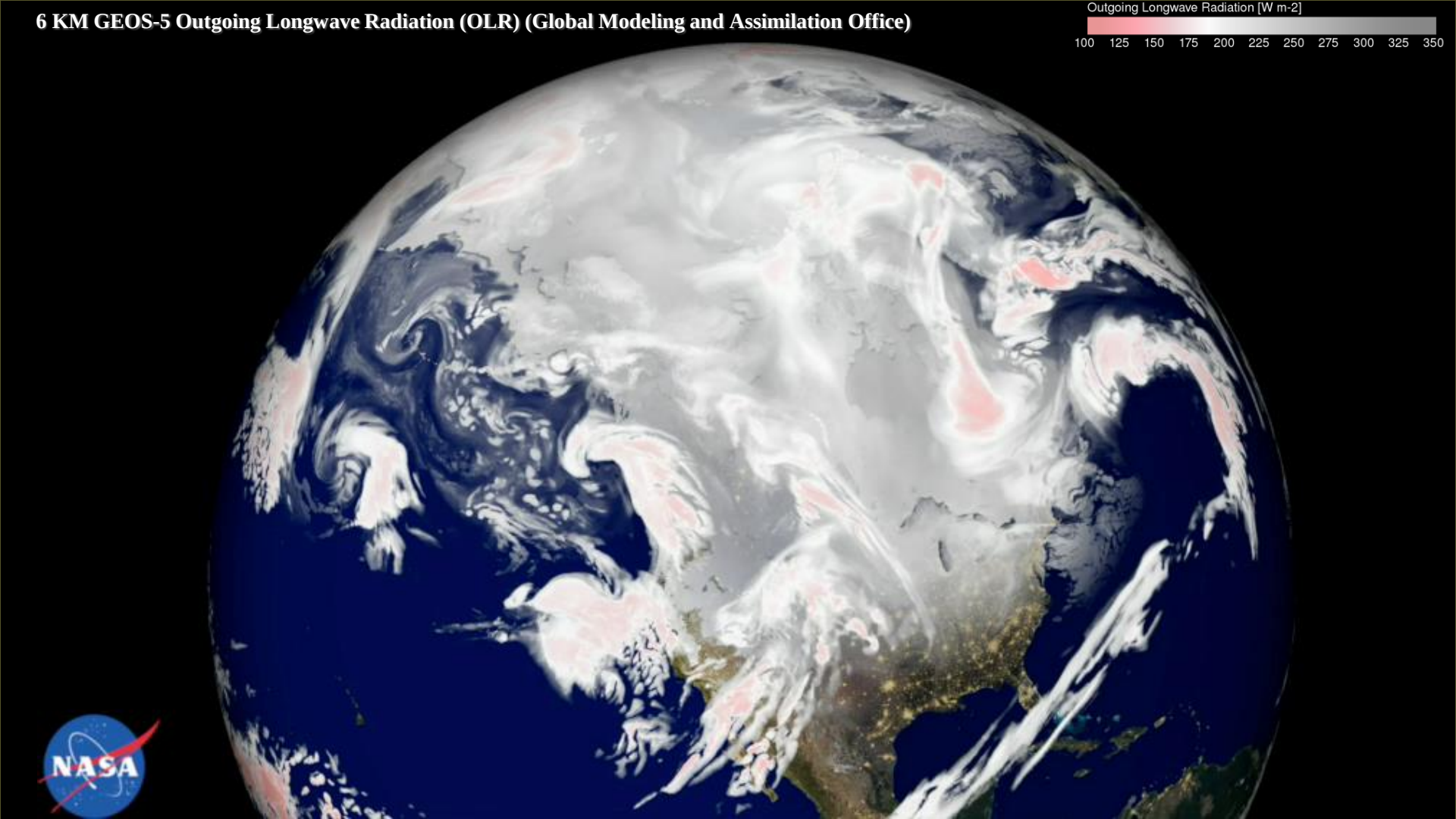
- There are 6 faces of the cube and multiple vertical layers
- Total number of grid points
  - X * Y * Z * 6 Faces of the Cube

Current GMAO Research

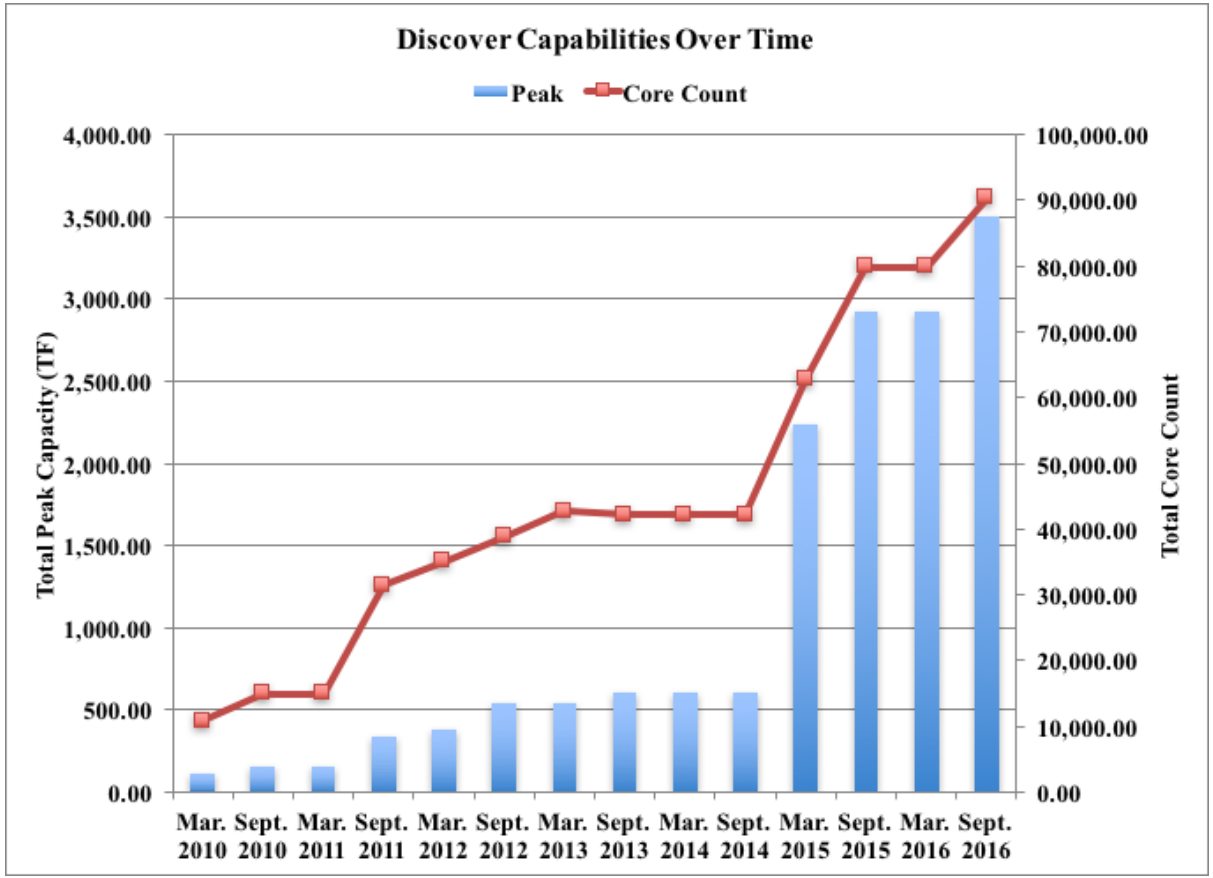
- Operational research forecasts are running at 27 KM resolution using about 27 million grid points
- Target operational research forecasts at a resolution of 12 KM in the very near future
- Reanalysis (including chemistry)
- Dynamic downscaling of reanalysis and forecasts down to 6 KM
- Highest resolution research runs are at 1.5 KM global resolution



Much more about the GEOS Model in the presentation by Bill Putman, NASA GMAO, this Friday, October 28!


6 KM GEOS-5 Outgoing Longwave Radiation (OLR) (Global Modeling and Assimilation Office)
Outgoing Longwave Radiation [W m-2]
100 125 150 175 200 225 250 275 300 325 350
NASA

Discover (HPC) Capacity Evolution

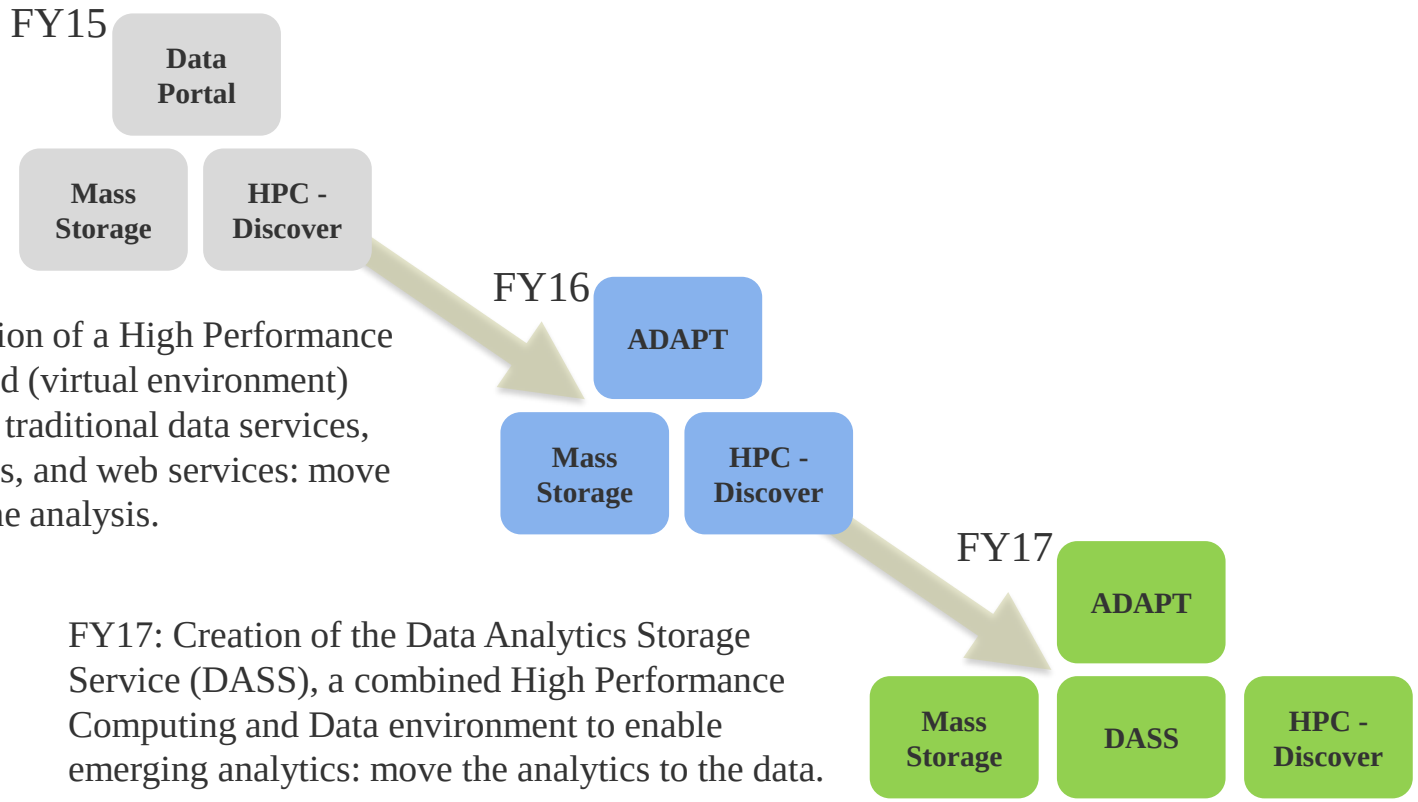


Discover (HPC) Scratch Disk Evolution

Calendar	Description	Decommission	Total Usable Capacity (TB)
2012	Combination of DDN disks	None	3,960
Fall 2012	NetApp1: 1,800 by 3 TB Disk Drives; 5,400 TB RAW	None	9,360
Fall 2013	NetApp2: 1,800 by 4 TB Disk Drives; 7,200 TB RAW	None	16,560
Early 2015	DDN10: 1,680 by 6 TB Disk Drives, 10,080 TB RAW	DDNs 3, 4, 5	~26,000
Mid 2015	DDN11: 1,680 by 6 TB Disk Drives, 10,080 TB RAW	DDNs 7, 8, 9	~33,000
Mid 2016	DDN12: 1,680 by 6 TB Disk Drives, 10,080 TB RAW	None	~40,000
Early 2017	13+ PB RAW	TBD	~50,000

- Usable capacity differs from raw capacity for two reasons. First, the NCCS uses RAID6 (double parity) to protect against drive failures. This incurs a 20% overhead for the disk capacity. Second, the file system formatting is estimated to also need about 5% of the overall disk capacity. The total reduction from the RAW capacity to usable space is about 25%.

NCCS Evolution of Major Systems



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.

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)

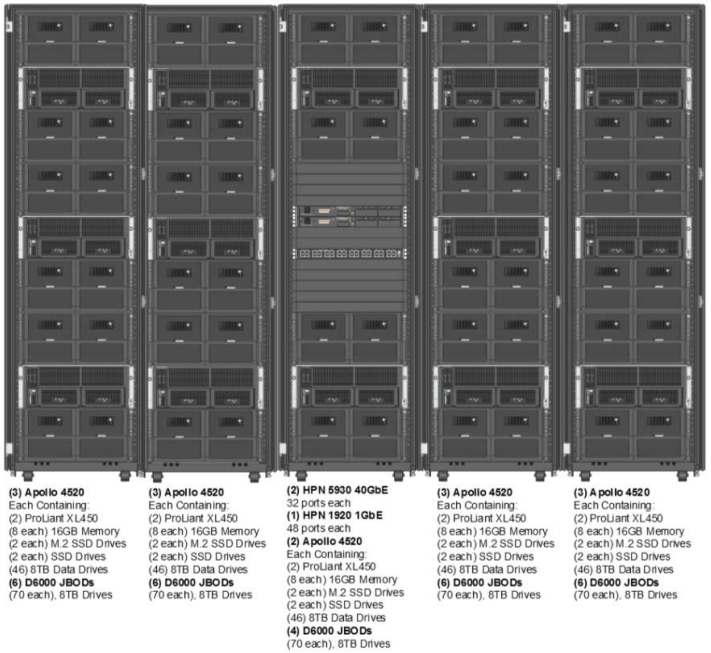
- On which users perform many analyses
- And may not be in a NASA Distributed Active Archive Center (DAAC)

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

Initial DASS Capability Overview

- Initial Capacity
  - 20.832 PB Raw Data Storage
  - 2,604 by 8TB SAS Drives
  - 14 Units
  - 28 Servers
  - 896 Cores
  - 14,336 GB Memory
  - 16 GB/Core
  - 37 TF of compute
- Roughly equivalent to the compute capacity of the NCCS just 6 years ago!
- Designed to easily scale both horizontally (compute) and vertically (storage)



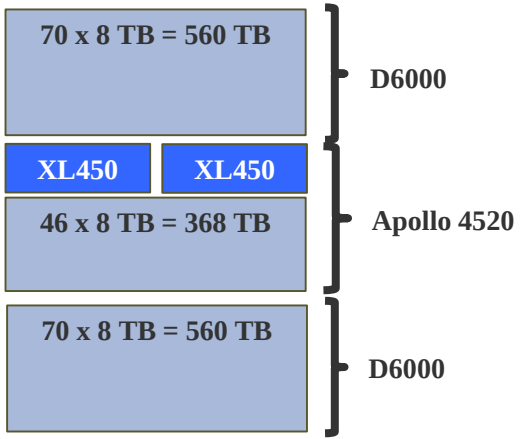
DASS Compute/Storage Units

HPE Apollo 4520 (Initial quantity of 14)

- Two (2) Proliant XL450 servers, each with
- Two (2) 16-core Intel Haswel E5-2697Av4 2.6 GHz processors
- 256 GB of RAM
- Two (2) SSD's for the operating system
- Two (2) SSD's for metadata
- One (1) smart array P841/4G controller
- One (1) HBA
- One (1) Infiniband FDR/40 GbE 2-port adapter
- Redundant power supplies
- 46 x 8 TB SAS drives

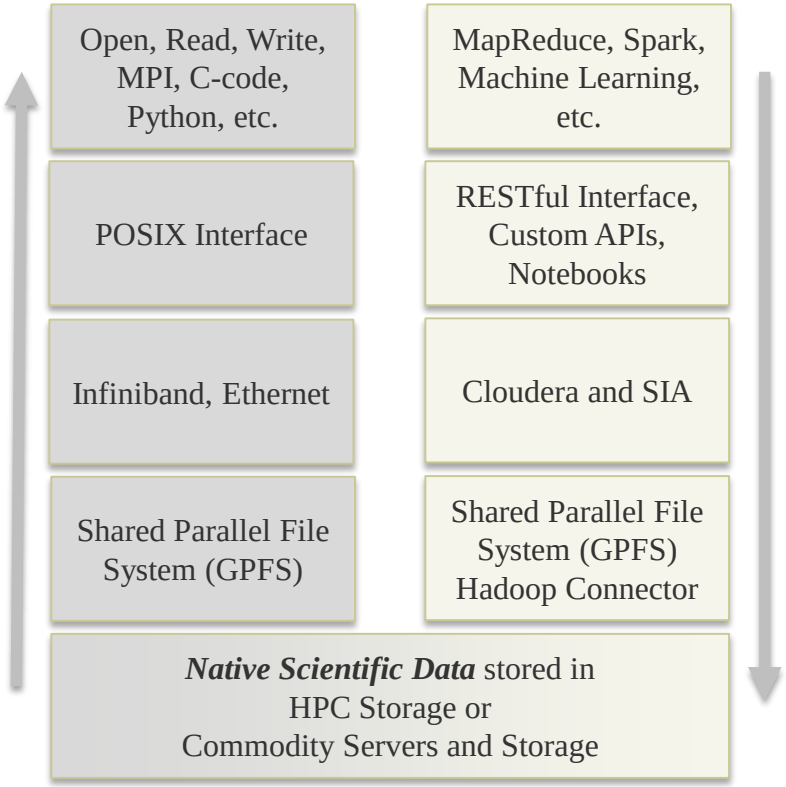
Two (2) D6000 JBOD Shelves for each Apollo 4520

- 70 x 8TB SAS drives



DASS Software Stack

Traditional
Data moved from storage to compute.



Emerging
Analytics moved from servers to storage.

Open Source Software Stack on DASS Servers

- Centos Operating System
- Software RAID
- Linux Storage Enclosure Services
- Pacemaker
- Corasync

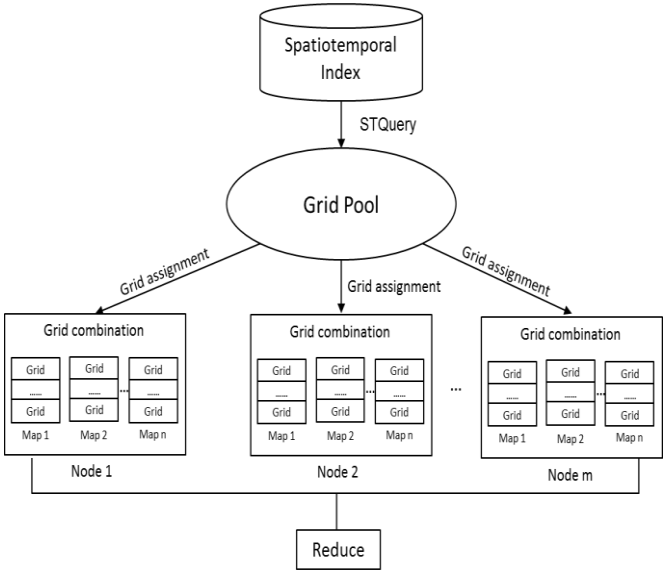
Spatiotemporal Index Approach (SIA) and Hadoop

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

Built 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 to
- 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
International Journal of Geographical Information Science, 2016
http://dx.doi.org/10.1080/13658816.2015.1131830

Analytics Infrastructure Testbed

Built three test clusters using decommissioned HPC servers and networks.

Test Cluster 1
SIA
Cloudera
HDFS

- 20 nodes (compute and storage)
- Cloudera
- HDFS
- Sequenced data
- Native NetCDF data
  - Put only

Test Cluster 2
SIA
Cloudera
Hadoop Connector
GPFS

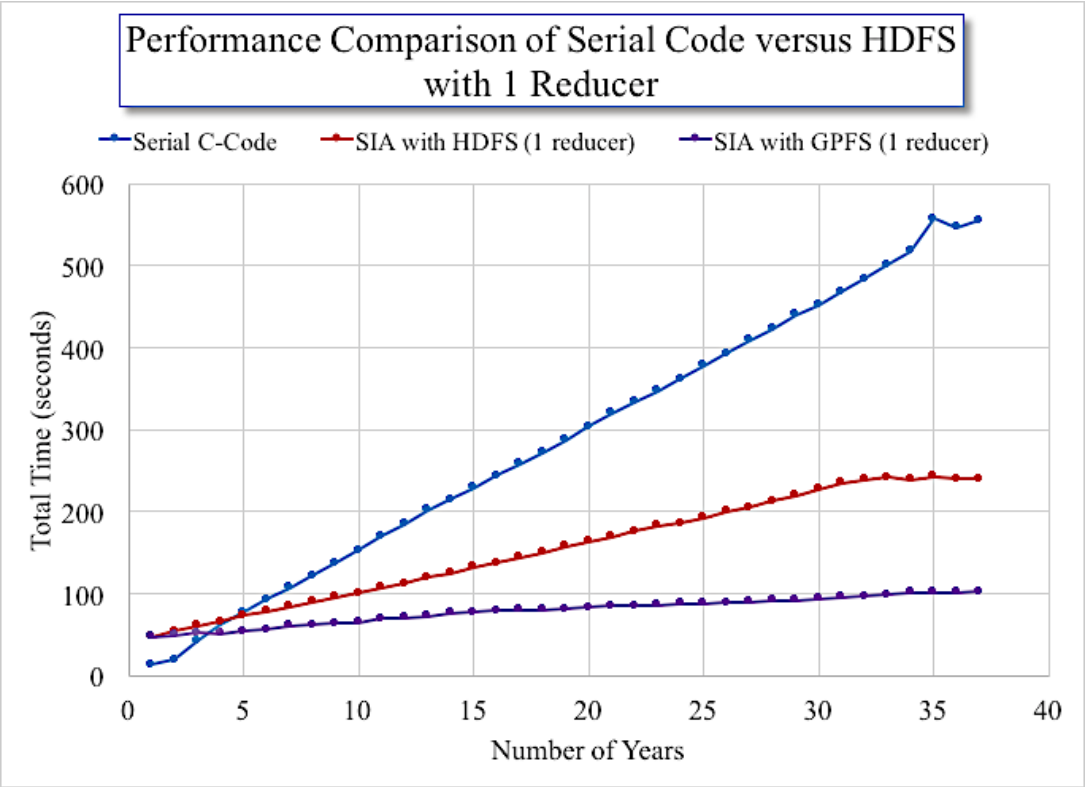
- 20 nodes (compute and storage)
- Cloudera
- GPFS
- *GPFS Hadoop Connectors*
- Sequenced data
  - Put and Copy
- Native NetCDF Data
  - Put and Copy

Test Cluster 3
SIA
Cloudera
Hadoop Connector
Lustre

- 20 nodes (compute and storage)
- Cloudera
- Lustre
- *Lustre HAM and HAL*
- Sequenced data
  - Put and Copy
- Native NetCDF Data
  - Put and Copy

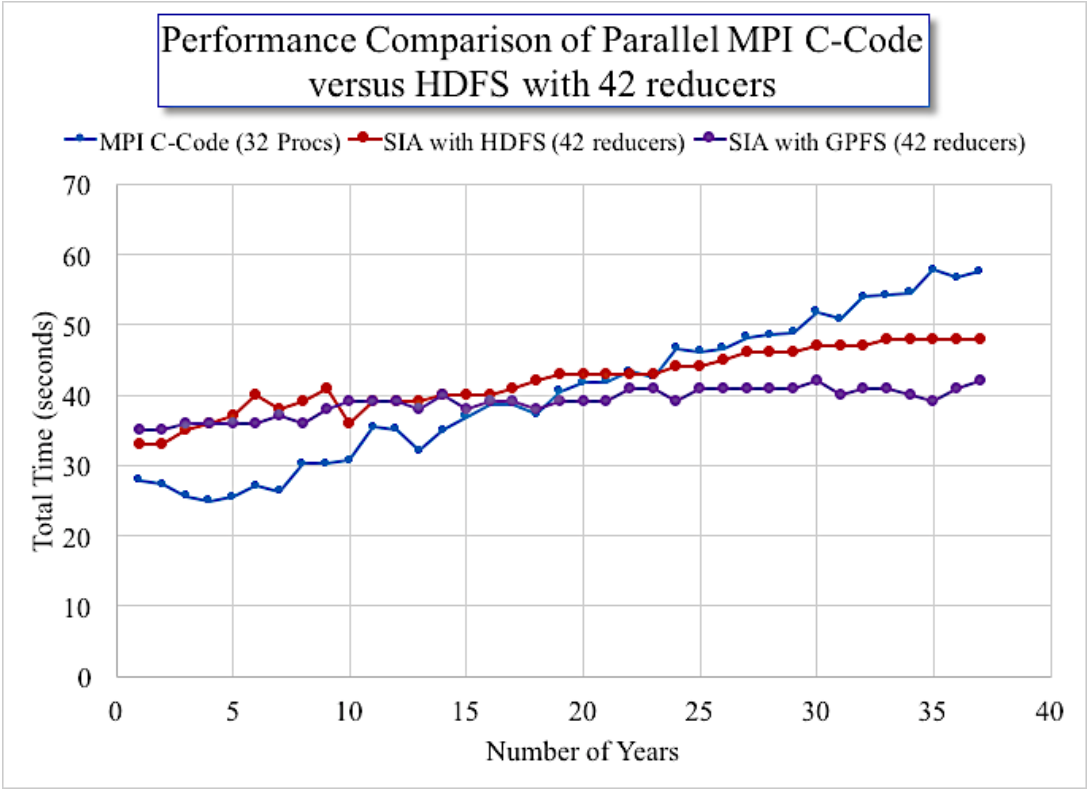
DASS Initial Serial Performance

- Compute the average temperature for every grid point (x, y, 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



DASS Initial Parallel Performance

- Compute the average temperature for every grid point (x, y, 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



Climate Analytics as a Service

High-Performance Compute/Storage Fabric

Data Sets
- Forecasts
- Seasonal Forecasts
- Reanalyses
- Nature Runs

Data

Relevance and Collocation

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

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***

Interfaces
- APIs
- Web Services
- Python Notebooks
- ***Zeppelin Notebooks (show demo)***

Exposure

Convenient and Extensible

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

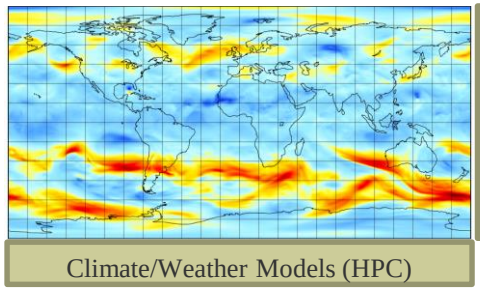
Increasing the GEOS-5 Model Resolution for Research

Target – Run approximately 100 meter global resolution research runs in 10 to 15 years
Each doubling of resolution requires 4x the grid points in the (x, y) direction; assume number of vertical layers are a constant at 132

Model	X and Y Values	Grid Points	Resolution (meters)	Cores	RAM (PB)
C1440	5,760	26×10^9	1,736	30,000	0.12
C2880	11,520	105×10^9	868	120,000	0.48
C5760	23,040	420×10^9	434	480,000	1.92
C11520	46,080	$1,682 \times 10^9$	217	1,920,000	7.68
C23040	92,160	$6,727 \times 10^9$	109	7,680,000	30.72

Bad News – This is only one component of the application (the atmosphere). GMAO is working on their coupled model including Atmosphere, Ocean, Waves, Ice, and More; We expect the model to require much more memory pushing us toward a higher memory to flop ratio.

Future of Data Analytics



- Future HPC systems must be able to efficiently transform information into knowledge using both traditional analytics and emerging *machine learning* techniques.
- Requires the ability to be able to index data in memory and/or on disk and enable analytics to be performed on the data where it resides – even in memory
- All without having to modify the data

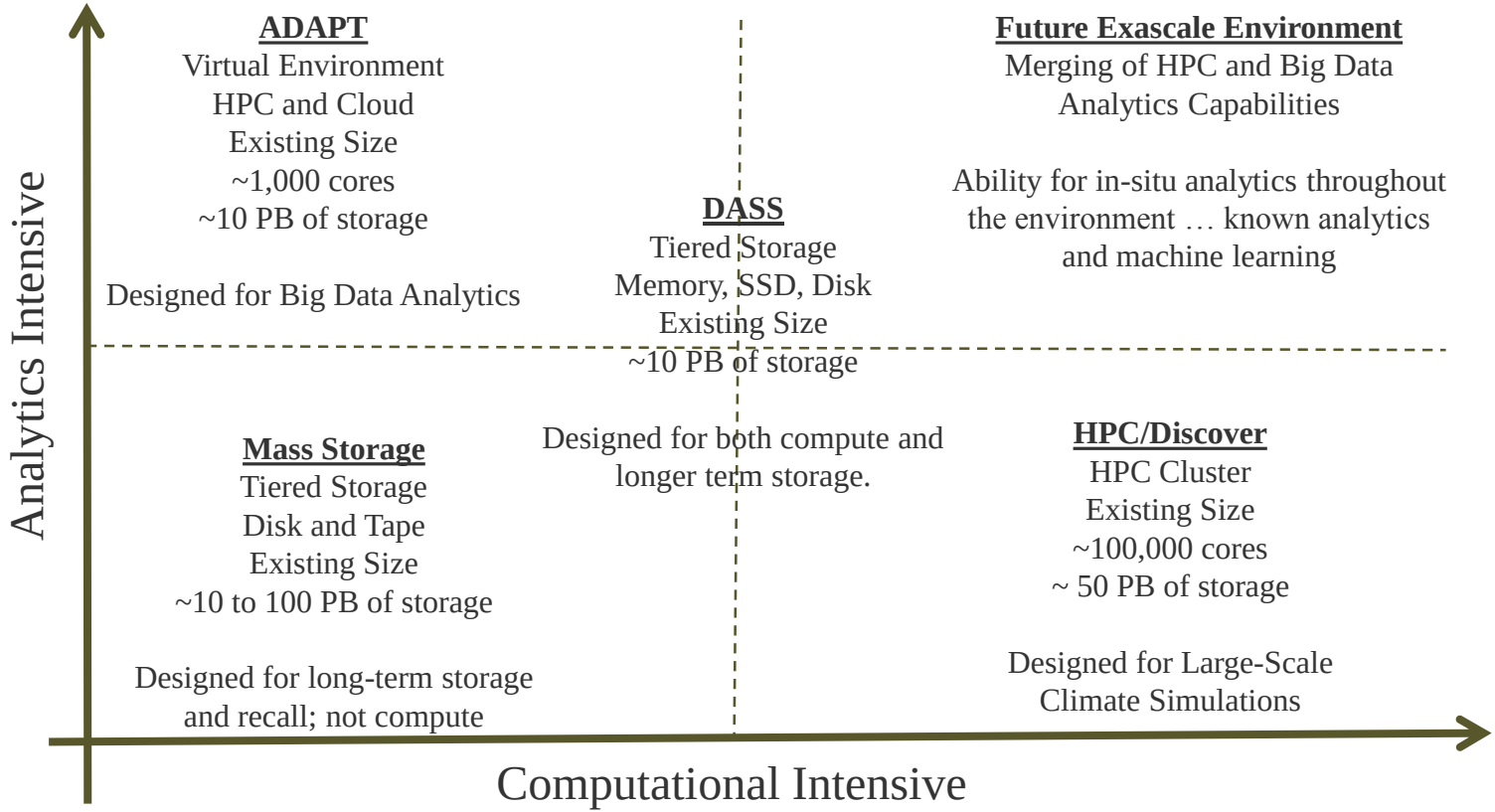
Compute Where the Data Resides

CPU	On chip high bandwidth memory – think NVIDIA GPUs and Intel Phi architectures.
In Package Memory	High Bandwidth Memory on the chip.
Node Memory	NVME, emerging technologies, etc. Large quantities of persistent storage close to the CPU.
File System in Memory	Technologies to enable shared memory across many nodes as well as collective operations at the network level.
Network	Very fast reads and writes into and out of the network and the HPC environment.
High Performance File System	Large aggregate space and throughput designed for longer term persistent storage.
Tiered Storage Subsystems	Hierarchy of storage systems from SSD's to spinning disks to …
Cloud	Ability to store data in the cloud and burst when appropriate for data analytics.

Major Challenges

- Hierarchy of memory and storage will make it even more difficult to optimize application performance.
  - Nothing we have not already heard before.
  - One of the most (perhaps the most) important optimizations for code performance.
  - Continuing to become harder as additional layers of memory and storage are being included in systems.
  - Quite possible that memory will become larger but slower as well with the inclusion of Non-Volatile RAM (NVMe).
- Must be able to perform data analysis at EVERY system layer.
  - Be able to efficiently move the data to the appropriate layer for computation, or
  - Move the thread to the data.

Future of HPC and Big Data at Exascale



Just for fun
The 5 V's of Big Data … and More

Let's start with the 5 V's of data that everyone knows…

- Volume, Velocity, Veracity, Variety, Value

Others are adding more V's …

- Visualization, Variability, Viability

Here are a few more that we are keeping in mind as we move forward …

Lifecycle of Data

- Viva La Data
- Vintage
- Vindictive
- Vicious

Data Security

- Vandalized
- Victimized
- Velociraptor
- Voldemort

Just for fun

- Vortex
- Vice
- Venomous
- Vivacious