ABoVE Science Cloud:
An orientation for proposers to the 2014 A.4 Research Announcement

During the solicitation please address all questions about the Science Cloud to peter.griffith@nasa.gov

Peter Griffith
Chief Support Scientist, Carbon Cycle & Ecosystems Office

Mark McInerney
CISTO Data Services Lead

Dan Duffy
NCCS HPC Lead
Contributors

High Performance Computing
- Scott Sinno, System Architect and System Administrator
- Hoot Thompson, System Architect and System Administrator
- Garrison Vaughn, System Architect and Applications Engineer

Climate Model Data Services
- Steven Ambrose, Principal Systems Engineer
- Julien Peters, Software Developer
- Eric Winter, Software Developer

Carbon Cycle & Ecosystems Office
- Liz Hoy, Support Scientist
The Carbon Cycle & Ecosystems Office is responsible for implementation and management of ABoVE

Selected science team members should plan to work closely with the CCEO and rely upon our guidance for field operations and safety, communications with local and regional stakeholders and authorities, and utilization of ABoVE cyberinfrastructure.

The ABoVE Science Cloud combines high performance computing with emerging technologies and data management with tools for analyzing and processing geographic information to create an environment specifically designed for large-scale modeling, analysis of remote sensing data, copious disk storage for “big data” with integrated data management, and integration of core variables from in-situ networks. The ABoVE Science Cloud is a collaboration that promises to accelerate the pace of new Arctic science for researchers participating in the field campaign. Furthermore, by using the ABoVE Science Cloud as a shared and centralized resource, researchers reduce costs for their proposed work, making proposed research more competitive.
Satellite and airborne remote sensing will be critical to the spatial and temporal scaling of observations made from field studies and in-situ measurements.
CCEO will assist the Science Team throughout the Data Management Lifecycle

Augmented from Rüegg et al 2014 in Front Ecol Environ
ABoVE will contribute to and extend in-situ networks.
- Researchers should anticipate sharing their data using ABoVE’s cyberinfrastructure and/or partnering networks
- Storage in the ASC will be tailored to meet Science Team needs
- Using ORNL DAAC best practices will facilitate integration
ABoVE Concise Experiment Plan

The Arctic-Boreal Vulnerability Experiment (ABoVE) Science Definition Team has completed the ABoVE Concise Experiment Plan. This report refines the science questions and rationale for ABoVE, poses a next tier of science questions, and lays out a top-level study design. The study design defines the ABoVE study domain and research areas, identifies the types of studies and observations needed to address the ABoVE science questions, and provides guidance for study implementation.

NASA Terrestrial Ecology will solicit for ABoVE research in 2014 through NASA ROSES Appendix A.4 TERRESTRIAL ECOLOGY.

- **ACEP Erratum, November 2014**
- **More About Research Areas for ABoVE**
  The ABoVE Science Definition Team (SDT) identified 15 conceptual Research Areas within the Study Domain that offer the range of environmental conditions necessary to address most of the research questions and objectives for ABoVE. [more...](#)

- **ABoVE Concise Experiment Plan** (10MB pdf)
- **Figures and Appendices as interactive web maps**
Plan to use data products from pre-ABoVE investigations

<table>
<thead>
<tr>
<th>Project Lead</th>
<th>Project Title</th>
<th>(view profile)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bourgeau-Chavez, Laura</td>
<td>Bourgeau-Chavez-02: Planning and Collection of Data on Boreal Wildfire Effects: Studies of broad-scale 2014 Wildfires in NWT, Canada</td>
<td></td>
</tr>
<tr>
<td>Carroll, Mark</td>
<td>Carroll-01: Determining the Extent and Dynamics of Surface Water for the ABoVE Field Campaign</td>
<td>(view)</td>
</tr>
<tr>
<td>Loboda, Tatiana</td>
<td>Loboda-01: Long-Term Multi-Sensor Record of Fire Disturbances in High Northern Latitudes</td>
<td>(view)</td>
</tr>
<tr>
<td>Munger, J. (Bill)</td>
<td>Munger-03: Development of a Data-Assimilation Framework for Integrating 25 Years of Surface and Airborne observations to assess patterns of net CO2 Exchange from Arctic Ecosystems</td>
<td>(view)</td>
</tr>
<tr>
<td>Schaefer, Kevin</td>
<td>Schaefer-03: Remotely-Sensed Active Layer Thickness (ReSALT) Product Derived from InSAR Data Over North American Arctic Regions</td>
<td>(view)</td>
</tr>
<tr>
<td>Walker, Donald (Skip)</td>
<td>Walker-01: Recovery and Archiving of Key Arctic Alaska Vegetation Map and Plot Data for Long-Term Vegetation Analyses</td>
<td>(view)</td>
</tr>
</tbody>
</table>
Some ABoVE data products available now: Schaefer

<table>
<thead>
<tr>
<th>Contribution</th>
<th>Available by: Feb 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Product Title</strong>: Active Layer Thickness derived from InSAR for the Barrow, Alaska</td>
<td></td>
</tr>
<tr>
<td><strong>Description</strong>: We evaluated the Remotely Sensed Active Layer Thickness (ReSALT) product that uses the interferometric Synthetic Aperture Radar technique to measure seasonal surface subsidence and infer ALT around Barrow, Alaska. We compared ReSALT with ground-based Active Layer Thickness (ALT) obtained using probing calibrated, 500 MHz Ground Penetrating Radar at multiple sites around Barrow. We used a total of 9 PALSAR scenes between June and September and formed 20 interferograms with good coherence and without contamination by obvious ionospheric artifacts. ReSALT accurately reproduced observed ALT within uncertainty of the GPR and probing data in ~76% of the study area. However, ReSALT was less than observed ALT in ~22% of the study area with well-drained soils and in ~1% of the area where soils contained gravel. ReSALT was greater than observed ALT in some drained thermokarst lake basins representing ~1% of the area. This dataset include Subsidence trends, ALT estimates, and the corresponding uncertainty map.</td>
<td></td>
</tr>
<tr>
<td><strong>Expected Users</strong>: scientists and researchers interested in environmental change in Barrow, Alaska and the use of InSAR for Active Layer Thickness</td>
<td></td>
</tr>
<tr>
<td><strong>Metadata URL(s)</strong>:</td>
<td></td>
</tr>
<tr>
<td><strong>Data Server URL(s)</strong>:</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Contribution</th>
<th>Available by: Feb 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Product Title</strong>: Active Layer Thickness derived from InSAR for the Prudhoe Bay Region of Alaska</td>
<td></td>
</tr>
<tr>
<td><strong>Description</strong>: The InSAR Prudhoe Bay product includes 1992–2000 average Active Layer Thickness (ALT) for an 80x100 km study area of continuous permafrost on the North Slope of Alaska near Prudhoe Bay. ALT is about 30–50 cm over moist tundra areas, and a larger ALT of 50–80 cm over wet tundra areas. The InSAR estimated ALT values match in situ measurements at Circumpolar Active Layer Monitoring (CALM) sites within uncertainties. This dataset include Subsidence trends, ALT estimates, and the corresponding uncertainty map.</td>
<td></td>
</tr>
<tr>
<td><strong>Expected Users</strong>: scientists and researchers interested in environmental change to Alaska's North Slope and the use of InSAR for Active Layer Thickness</td>
<td></td>
</tr>
<tr>
<td><strong>Metadata URL(s)</strong>:</td>
<td></td>
</tr>
<tr>
<td><strong>Data Server URL(s)</strong>:</td>
<td></td>
</tr>
</tbody>
</table>
Some ABoVE data products available now: Walker

<table>
<thead>
<tr>
<th>Contribution</th>
<th>Available by:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Product Title:</strong></td>
<td>Arctic Alaska Geoecological Atlas</td>
</tr>
<tr>
<td><strong>Description:</strong></td>
<td>Abundant ground-based information will be necessary to inform the planned Arctic-Boreal Vulnerability Experiment (ABoVE) activities. The Atlas is comprised of archives of maps and plot-based vegetation data, and associated information. The Map Archive contains map products at several scales and numerous themes. The maps range from detailed geocological maps, which are polygon-based integrated terrain maps at relatively fine scales, to raster-based map products derived from satellite data and digital elevation models. The Vegetation Plot Archive contains vegetation-plot data, associated environmental data, and other related information from over 3,000 plots in Arctic Alaska.</td>
</tr>
<tr>
<td><strong>Expected Users:</strong></td>
<td>ABoVE Science Team</td>
</tr>
<tr>
<td><strong>Metadata URL(s):</strong></td>
<td><a href="http://geobotanical.portal.gina.alaska.edu/search">http://geobotanical.portal.gina.alaska.edu/search</a></td>
</tr>
<tr>
<td><strong>Data Server URL(s):</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Archived Data Citation:</strong></td>
<td></td>
</tr>
</tbody>
</table>
Plan to use data products from pre-ABoVE investigations

<table>
<thead>
<tr>
<th>Project Lead</th>
<th>Project Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bourgeois-Chavez, Laura</td>
<td>Bourgeois-Chavez-02: Planning and Collection of Data on Boreal Wildfire Effects: Studies of broad-scale 2014 Wildfires in NWT, Canada (view)</td>
</tr>
<tr>
<td>Carroll, Mark</td>
<td>Carroll-01: Determining the Extent and Dynamics of Surface Water for the ABoVE Field Campaign (view)</td>
</tr>
<tr>
<td>Loboda, Tatiana</td>
<td>Loboda-01: Long-Term Multi-Sensor Record of Fire Disturbances in High Northern Latitudes (view)</td>
</tr>
<tr>
<td>Munger, J. (Bill)</td>
<td>Munger-03: Development of a Data-Assimilation Framework for Integrating 25 Years of Surface and Airborne observations to assess patterns of net CO2 Exchange from Arctic Ecosystems (view)</td>
</tr>
<tr>
<td>Schaefer, Kevin</td>
<td>Schaefer-03: Remotely-Sensed Active Layer Thickness (ReSALT) Product Derived from InSAR Data Over North American Arctic Regions (view)</td>
</tr>
<tr>
<td>Walker, Donald (Skip)</td>
<td>Walker-01: Recovery and Archiving of Key Arctic Alaska Vegetation Map and Plot Data for Long-Term Vegetation Analyses (view)</td>
</tr>
</tbody>
</table>
Next up:
High Performance Computing at Goddard Space Flight Center

Computational and Informational Science and Technologies Office (CISTO)

<table>
<thead>
<tr>
<th>CISTO</th>
<th>NASA Center for Climate Simulation (NCCS)</th>
</tr>
</thead>
</table>
| **Climate Model Data Services (CMDS)** | • Mark McInerney, Lead  
 • Tools and services to visualize, analyze, compare, and publish climate model data                  |
| http://cds.nccs.nasa.gov     | • Daniel Duffy, Lead  
 • High performance computing solutions for NASA science, specifically engineered to meet the needs of climate science |
|                            | http://nccs.nasa.gov                                                                             |
|                            | https://twitter.com/NASA_NCCS                                                                       |
Climate Model Data Service (CDS) Support for the Arctic-Boreal Vulnerability Experiment (ABoVE)
CISTO (NCCS/CDS) Support for ABoVE?

Computational & Information Sciences and Technology Office (CISTO / 606)
Dedicated Technical Support for ABoVE

System Support / Science Cloud

Data Management / Application Support

Climate Model Data Service (CDS) is a team of dedicated “science technical” experts providing data access, data visualization, data distribution, and data management tools and services for ABoVE Science Cloud users

NASA Climate Model Data Services (CDS) – http://cds.nccs.nasa.gov
Computational and Information Sciences and Technology Office’s (CISTO) Service Layers...

Brings together the tools, data storage and high-performance computing to for timely analysis over large-scale data sets, where the data resides, to ultimately produce societal benefits

Services layers managed by CISTO Units:
- Climate Informatics (CI)
- NASA Center for Climate Simulation (NCCS)
- Climate Model Data Services (CDS)
Climate science data is a Big Data domain that is experiencing unprecedented growth.

The size of NASA’s space-based observational data is growing dramatically as new missions come online.

The bigger challenge is the work of data intensive climate scientists whose models regularly ingest observations, produce dynamic petabyte size data sets, and require more efficient data access to enable scientific discovery. CDS is working to fill this existing gap.

**Sample Data Sets:**

- Nature Run Simulation (GMAO): 5 PB
- Climate Model Output for Intercomparisons, published in Earth System Grid Federation (ESGF): 94.2 TB
  - IPCC CMIP5 Data (GISS ModelE, GEOS5 Decadal): 68 TB
  - NEX-DCP30 Downscaled Climate Projections: 13 TB
  - GeoMIP Geoengineering Data: 13 TB
  - Reanalysis (e.g. MERRA, CFSR): 200 GB
- MERRA/AS: 80 TB (or 160 TB uncompressed, 480 TB in Hadoop)
- NASA Land & Cryosphere Science Very-High Resolution Data Delivery: 1 PB
CDS is partnered with NASA Center for Climate Simulation (NCCS) which provides high performance computing services for climate scientists on “Discover” and where generated model output is then accessed and published through CDS.

Discover Compute Summary:

- 1.121 petaflops peak = 1,121,000,000,000,000 floating-point operations per second
- 3,354 compute nodes
- 43,048 Intel Xeon processor cores (603 teraflops)
- 64 NVIDIA Tesla M2070 GPUs with 28,672 “streaming GPU” CUDA cores (33 teraflops)
- 480 Intel Phi Many Integrated Core (MIC) co-processors with 28,800 cores (485 double precision teraflops)
- 102.088 terabytes main memory
- InfiniBand interconnect: 20, 40, or 56 gigabits per second

HEC computing services generate the data for Data Services....
NASA Climate Model: Data Services

Climate Model Data Services

NASA’s Climate Model Data Services (CDS) provides IT infrastructure, tools and services, for climate science community to visualize, analyze, compare, and publish large volumes of climate model data.

Allows climate scientists to focus on science and not scientific data management and supporting technology infrastructure.

Efficient access to climate model data supporting broader use. E.g., Understanding multiple data formats or location of data not required

Providing end-to-end data management of large-scale climate model repositories

Sample Services:

• CDS API for direct access to climate model data
• Large-scale data federation
• Large-scale publication and distribution Services
• 3D Data Visualization and Analysis
• Digital Object Identifier (DOI) Services
• Climate Ontology Services

Data Services infrastructure then feeds Analytic Services ....

NASA Climate Model Data Services (CDS) – http://cds.nccs.nasa.gov
Climate Model Data Analytics Services

Climate Analytics-as-a-Service (CAaaS) brings together the tools, data storage and high-performance computing to perform timely analysis over large-scale data sets, where the data resides, to ultimately produce societal benefits.

Example CDS CAaaS: MERRA Analytic Services (MERRA/AS):

- Enabling MapReduce analytics over NASA’s Modern-Era Retrospective Analysis for Research and Applications (MERRA) data collection.
- MERRA reanalysis integrates observational data with numerical models to produce a global temporally and spatially consistent synthesis of hundreds of climate variables.
- MERRA/AS brings together end-to-end CAaaS Capability:
  - High performance, data proximal analytics
  - Scalable data management
  - Software appliance virtualization
  - Adaptive analytic
  - Domain-harmonized API

Analytic Services infrastructure then feeds Knowledge Services...

NASA Climate Model Data Services (CDS) – http://cds.nccs.nasa.gov
Open services that perform timely analysis over large-scale data sets, where the data resides.

Are there knowledge services to develop with the defined tighter integration of climate science model data, compute services, data services, and analytic services?

Knowledge Management: Achieving objectives for the climate science community by acquiring, creating, and sharing climate model data, information, and knowledge across multiple disciplines and a supportive technical foundation.

Knowledge Organization: Activities that document, index, and classify climate model data, by subject matter experts, with supportive computer algorithms and infrastructure.
Climate Model Data Service (CDS) Support for the Arctic-Boreal Vulnerability Experiment (ABoVE)
Access to Common Datasets in Science Cloud

Common Datasets “Staged” for ABoVE Investigators in NCCS Science Cloud

• Staged and available for direct use
• Individual investigators don’t have to locate and download
• Additional datasets can be added, requirements driven by ABoVE
• Data Management Services
  • Location service, metadata support

Existing Staged Datasets (Pre-ABoVE)

• Landsat (study domain), surface reflectance, 50 TB (250 uncompressed) [temporal coverage 1990-2012 by time projects are funded]
• MODIS (circumpolar), daily surface reflectance (MOD/MYD09GA) & 8-day surface reflectance composites (MOD09A1) 248TB
• MERRA, GEOS-5 reanalysis, 81 TB
• Hourly Weather Observations (METAR), 350 stations, ~38 MB per week
• NGA (study domain), Very-High Resolution Commercial Satellite Imagery, E1.8 PB
CDS Support for Investigator Generated Data

**Investigator Generated Data**

- Short-term storage and distribution services provided
- Folded into science cloud staged or common datasets for use by other investigators
- Long-term archive and distribution outside the Science Cloud (DAAC/ORNL, other)
  - CDS data management services assistance provided
### NASA Climate Model Data Services

Data Publication and Distribution Services

<table>
<thead>
<tr>
<th>Data Publication Services</th>
<th>Protocol</th>
<th>Download</th>
<th>Subsetting</th>
<th>2D Visualization</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Web Access</strong></td>
<td>HTTP</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>For downloading small files</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>File Transfer Protocol (FTP)</strong></td>
<td>FTP</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Anonymous FTP supporting wget</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>GRads Data Server (GDS)</strong></td>
<td>OPENDAP</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Data subsetting and analysis services</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Live Access Server (LAS)</strong></td>
<td>OPENDAP</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Data subsetting and analysis services</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>THREDDS Data Server (TDS)</strong></td>
<td>OPENDAP</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>subsetting, analysis, &amp; visualization</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Earth System Grid Federation (ESGF)</strong></td>
<td>OPENDAP</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Data access to IPPC CMIP data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Web Map Service (WMS)</strong></td>
<td>OPENDAP</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Data publication to IPPC CMIP Format</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NASA Climate Model Data Services (CDS) – http://cds.nccs.nasa.gov
Science cloud does have NASA enterprise licensed ESRI products installed

- ArcGIS for Desktop
- ArcGIS for Server
- ArcGIS for Portal

USE 1: ArcGIS services to stage select data by Carbon Cycle & Ecosystem office
- NGA, other

USE 2: ESRI application / system level support for ABoVE projects
- EG. Grant scientist / project ArcGIS server account to for data publication (scientist/project managed)
NGA High Resolution Commercial Satellite Imagery Support for ABoVE: Overview / Background

National Geospatial Agency (NGA) has licensed all DigitalGlobe > 0.5 m satellite imagery for US Federal use, i.e., NSF, NASA and NASA funded projects...and so ABoVE.

- Archive of 4.2 billion km² of data from 2000 to present
- Data from five different satellites; Worldview-1, Worldview-2, Ikonos, Quickbird, and Geoeye-1
- Access to NGA imagery (~3.4/ km²) at no cost to NASA

<table>
<thead>
<tr>
<th>Satellite</th>
<th>Bands</th>
<th>Nadir Panchromatic Resolution (m)</th>
<th>Nadir Multispectral Resolution (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ikonos</td>
<td>Pan, R, G, B, Near IR</td>
<td>1.0</td>
<td>4</td>
</tr>
<tr>
<td>GeoEye</td>
<td>Pan, R, G, B, Near IR</td>
<td>0.5</td>
<td>2.0</td>
</tr>
<tr>
<td>Quickbird</td>
<td>Pan, R, G, B, Near IR</td>
<td>0.6</td>
<td>2.4</td>
</tr>
<tr>
<td>WorldView-1</td>
<td>Panchromatic only</td>
<td>0.5</td>
<td>N/A</td>
</tr>
<tr>
<td>WorldView-2</td>
<td>Pan, R, G, B, Near IR 1, Near IR 2, Coastal, Red Edge, Yellow</td>
<td>0.5</td>
<td>2.0</td>
</tr>
</tbody>
</table>

* Worldview 3 images will be offered once available

NGA (DigitalGlobe) Satellite Fleet

NASA Climate Model Data Services (CDS) – http://cds.nccs.nasa.gov
NGA High Resolution Commercial Satellite Imagery:
High Level Goals for ABoVE

✓ **GOAL 1: (OBTAIN)** Establish access to NGA data through strong partnerships with NGA, Digital Globe, PGC, and NSF

**GOAL 2: (STAGE)** Collect and store ABoVE domain specific NGA data into the NCCS Science Cloud

**GOAL 3: (ACCESS)** Provide short-term data access, system compute, processing and analytics services for ABoVE scientists
• **Obtain Existing Imagery**: Alaska and entire Arctic north of 60N already tasked by NGA in stereo-panchromatic and multispectral, obtain from PGC

• **Direct tasking**: For ABoVE core and extended domain below 60N and 60N circumpolar (2-3 year activity)

• **Bulk transfer** and store raw NGA (NTF) images in ABoVE Science Cloud
  - Ordering process
  - Networked connection direct from Science Cloud to NGA

• **Value Added NGA Imagery**: Create ~0.5m panchromatic, orthorectified Arctic Mosaic of ABoVE core and extended domain including south of 60N

• **Elevation model** “first pass” of the ABoVE core and extended domain
  - “first pass” meaning models will not be corrected for vegetation and built structures and so should be used as first pass product
  - DEMS from stereoscopic imagery using Ames Stereo Pipeline (ASP)
11 Month Collection of NGA/DG Images at PGC
Current NGA Data holdings in ABoVE Science Cloud as of January 2015

- > 100 TB
- Image is zoomed in, numerous gaps remain (EG Alaska)
Questions

CISTO Data Services Lead

Mr. Mark A. McInerney
Phone: 301-286-1491
Email: Mark.McInerney@nasa.gov
URL: http://cds.nccs.nasa.gov
ABoVE Science Cloud
January 2015

Daniel Duffy (daniel.q.duffy@nasa.gov and on Twitter @dqduffy)
High Performance Computing Lead
NASA Center for Climate Simulation (NCCS)
http://www.nccs.nasa.gov
High-End Computing Capability (HECC) Project
NASA Advanced Supercomputing (NAS)
NASA Ames
Dr. Piyush Mehrotra

NASA Center for Climate Simulation (NCCS)
Goddard Space Flight Center (GSFC)
Dr. Daniel Duffy
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)
  - High Performance Science Cloud
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](http://www.nccs.nasa.gov)
- Located in Building 28 at Goddard
**Takes in small input and creates large output**

- Using relatively small amount of observation data, models are run to generate forecasts
- Fortran, Message Passing Interface (MPI), large shared parallel file systems
- Rigid environment – users adhere to the HPC systems

**Example: GEOS-5 Nature Run (GMAO)**

- 2-year Nature Run at 7.5 KM resolution
- 3-month Nature Run at 3.5 KM resolution
- Will generate about 4 PB of data (compressed)
- To be used for Observing System Simulation Experiments (OSSE’s)
- All data to be publically accessible
- ftp://G5NR@dataportal.nccs.nasa.gov/

• High performance computing capabilities within the NCCS are available for ABoVE research projects.
• The NCCS currently has a large amount of high performance computing and storage available for large-scale simulations
  – Petaflops of peak computing capability with 1,000’s of compute nodes
  – Petabytes of high performance storage
  – Help desk, ticketing system, applications support, etc.
• For more information about the high performance computing capabilities, please go to our web site
  – www.nccs.nasa.gov
Takes in large amounts of input and creates a small amount of output

- Use large amounts of distributed observation and model data to generate science
- Analysis applications are typically 100’s of lines of code
- Python, IDL, Matlab, custom
- Agile environment – users run in their own environments

Examples

- Decadal water predictions for the high northern latitudes for the past three decades (Mark Carroll)
  - Requires 100,000+ Landsat images and about 20 TB of storage
- Evaporative transport (Wei)
  - Requires monthly reanalysis data sets for four different spatial extents
Advanced Data Analytics Platform (ADAPT) “High Performance Science Cloud”

High Performance Science Cloud is uniquely positioned to provide data processing and analytic services for NASA Science projects

**Adjunct to the NCCS HPC environment**
- Lower barrier to entry for scientists
- Customized run-time environments
- Reusable HPC/Discover hardware

**Expanded customer base**
- Scientist brings their analysis to the data
- Extensible storage; build and expand as needed
- Persistent data services build in virtual machines
- Create purpose built VMs for specific science projects

**Difference between a commodity cloud**
- Platform-as-a-Service that comes close to matching HPC levels of performance
- Critical Node-to-node communication – high speed, low latency
- Shared, high performance file system
- Management and rapid provisioning of resources
ABoVE Science Cloud

Solution for the ABoVE Computing and Storage Services Requirement is the innovative *High Performance Science Cloud*

**Partnership between the CCE, CISTO, and NCCS**
- Provide compute, storage, data management, and data publication for the ABoVE campaign using the HPSC
- Reduces technical overhead for ABoVE scientists
- Allows scientists to focus on science in an optimized computing environment

**The Conceptual Architecture:**
- Data analysis platform collocating data, compute, data management, and data services
- Ease of use for scientists; customized run-time environments; agile environment
- Data storage surrounded by a compute cloud
- Large amount of data storage
- High performance compute capabilities **
- Very high speed interconnects

**Repurposes NCCS HPC Discover Compute for routine growth**
### System Components/Configuration

<table>
<thead>
<tr>
<th>Capability and Description</th>
<th>Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Persistent Data Services</strong></td>
<td>Nodes with 128 GB of RAM, 10 GbE, and FDR IB</td>
</tr>
<tr>
<td>Virtual machines or containers deployed for web services, examples include ESGF, GDS, THREDDS, FTP, etc.</td>
<td></td>
</tr>
<tr>
<td><strong>DataBase</strong></td>
<td>Nodes with 128 GB of RAM, 3.2 TB of SSD, 10 GbE, and FDR IB</td>
</tr>
<tr>
<td>High available database nodes with solid state disk.</td>
<td></td>
</tr>
<tr>
<td><strong>Remote Visualization</strong></td>
<td>Nodes with 128 GB of RAM, 10 GbE, FDR IB, and GPUs</td>
</tr>
<tr>
<td>Enable server side graphical processing and rendering of data.</td>
<td></td>
</tr>
<tr>
<td><strong>High Performance Compute</strong></td>
<td>~100 nodes with 32 to 64 GB of RAM, and FDR IB</td>
</tr>
<tr>
<td>More than 1,000 cores coupled via high speed Infiniband networks for elastic or itinerant computing requirements.</td>
<td></td>
</tr>
<tr>
<td><strong>High-Speed/High-Capacity Storage</strong></td>
<td>Storage nodes configured with a total of about 3 PB of RAW storage capacity</td>
</tr>
<tr>
<td>Petabytes of storage accessible to all the above capabilities over the high speed Infiniband network.</td>
<td></td>
</tr>
</tbody>
</table>
Software

• Operating Systems
  – Debian Linux
  – CentOS Linux
  – Windows (4 licenses)
• Commercial Tools
  – Intel Compilers (C, C++, FORTRAN)
  – IDL (4 licenses)
• Open Source Tools
  – Python
  – NetCDF
  – Almost anything you need
• How to install software?
  – Users can install software in your home and nobackup directories
  – Users can request support to help install and configure software on your virtual machines
• Can reach out to external license servers
  – We can do this and make external software licenses available to your virtual machines; takes time to poke holes through firewalls
### Data Currently In the Cloud

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Amount</th>
<th>Path</th>
<th>Environment Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Landsat – Surface reflectance</td>
<td>123 TB</td>
<td>/att/hdvault/LANDSAT</td>
<td>$LANDSAT</td>
</tr>
<tr>
<td>MODIS – Daily 500m surface reflectance</td>
<td>57 TB</td>
<td>/att/nobackup/MERRA</td>
<td>$MODIS</td>
</tr>
<tr>
<td>MERRA – GEOS-5 reanalysis</td>
<td>192 TB</td>
<td>/att/nobackup/MERRA</td>
<td>$MERRA</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>372 TB</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**NGA Data**
- NGA data is available only on specific virtual machines
- Contact Mark McInerney and the Climate Model Data Services for access

Additional data sets to be moved in the cloud will be dependent on the funded science research. Ownership and management of these large data sets will be a shard responsibility between the system (NCCS) and the project (ABoVE).
## User Requirements Table

What we will ask you before you get access

<table>
<thead>
<tr>
<th>Category</th>
<th>Options</th>
<th>Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating System</td>
<td>Debian Linux, Centos Linux, Windows</td>
<td></td>
</tr>
<tr>
<td>Number of Cores per Virtual Machine</td>
<td>1 to 20</td>
<td></td>
</tr>
<tr>
<td>Amount of Memory per Virtual Machine</td>
<td>24 GB to 64 GB</td>
<td></td>
</tr>
<tr>
<td>Number of Virtual Machines</td>
<td>1 to 48</td>
<td></td>
</tr>
<tr>
<td>Data Sets</td>
<td>LANDSAT, MERRA, MODIS, NGA, Other</td>
<td></td>
</tr>
<tr>
<td>Data Storage</td>
<td>Depends on request</td>
<td></td>
</tr>
<tr>
<td>Software</td>
<td>Compilers (gnu, Intel), IDL, python, NetCDF, perl, etc.</td>
<td></td>
</tr>
</tbody>
</table>
What works best in the cloud?

• Not designed for MPI (use HPC systems for that)
  – Highly coupled processes performing large amounts of message passing
• Designed more for inherently parallel processing of big data
  – Independent processes written to analyze large data sets
  – We have tools to assist users in submitting independent parallel scripts across multiple virtual machines
• Publishing of data
  – Persistent data services created to provide a capability for NASA scientists to share large data
  – Climate Model Data Services (Mark) supports this
• The system is not an archive
  – Users will have to adhere to their data management plan
• Large data stored in the system will be stored for the life of the project
  – ABoVE related data, such as Landsat and NGA, will be there for a long time
  – Other data may only be there for months and can be deleted after the project is completed
    • Dependent on the project and research
    • Will work with the ABoVE team and the researchers to decide how best to do this
• Where do you archive your results?
  – Outside the cloud, NASA DAACs, other
Moving Large Data

• The NCCS can help
  – High speed network connection into the ABoVE Science Cloud (10 GbE)
  – Will discuss this during our initial requirements gathering with each group
• Synchronizing data
  – Again, the NCCS can help
• Automated data transfers
  – Can set up automated (cron) jobs to move data into the cloud
• Public data only at this time!
How Much Storage Can You Use

• Home Directories
  – Quota of 15 GB
  – Replicated within the system, but not backed up
  – User code, applications, etc.

• Scratch Space
  – Quota of 2.5 TB
  – Not backed up
  – Temporary files, larger scratch space
  – Not a long term archive

• Large Data Sets
  – Shared storage environment (petabyte) for large datasets to be shared across the ABoVE research
The Future

• More capabilities
  – More compute nodes
• More service offerings to come on line
  – Remote Visualization
  – Database
  – Remote Desktop (maybe)
  – More Windows licenses
• Self Service Portal
  – Users will be able to launch their own virtual machines
The Team – The people that do the real work!

- System Engineers/System Administrators
  - Hoot Thompson
  - Scott Sinno
  - Garrison Vaughan
  - Ramon Rameriz-Linan
  - Brittney Wills

- NCCS Help Desk Contact Information
  - support@nccs.nasa.gov
  - 301-286-9120