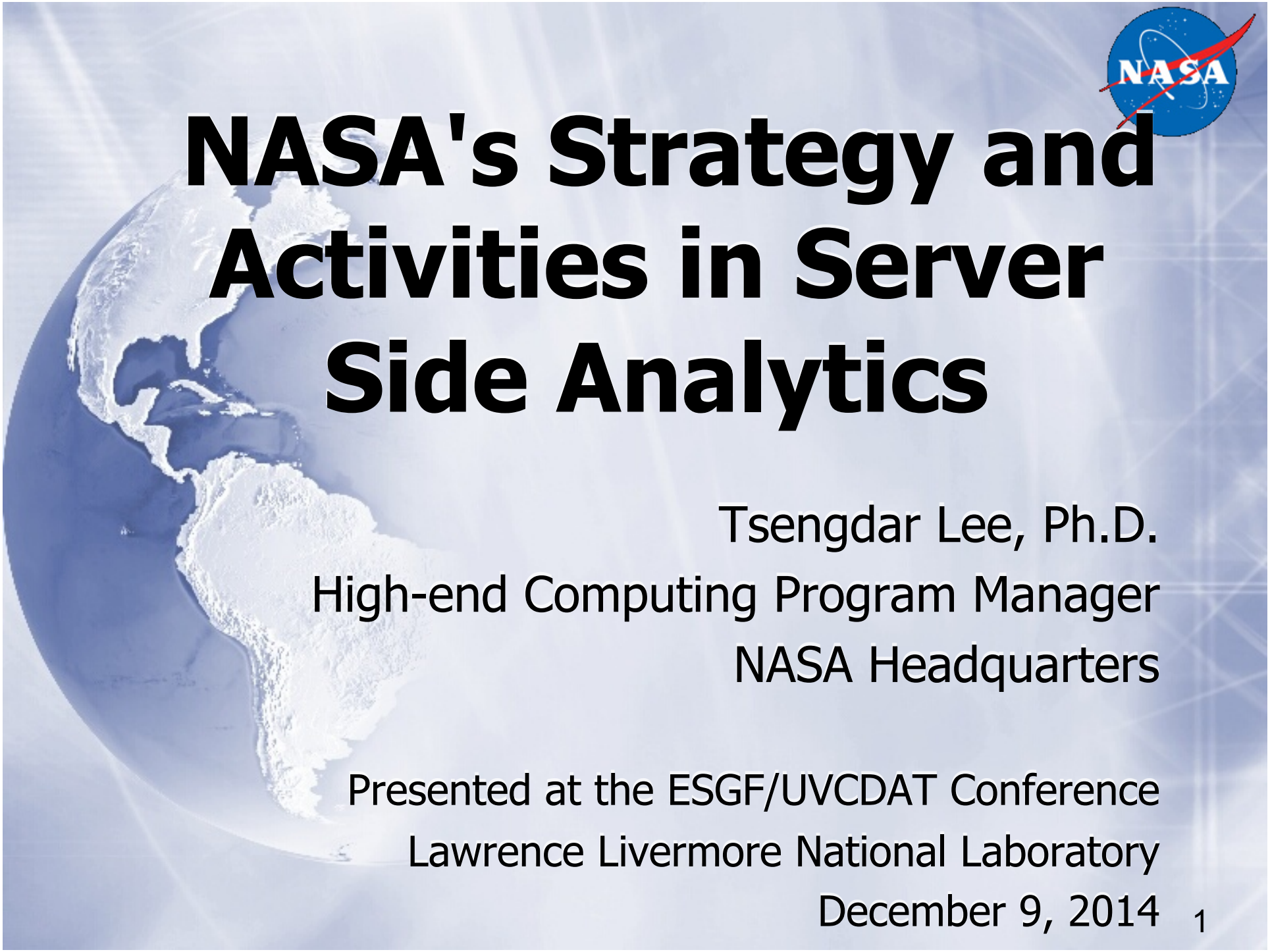


NASA's Strategy and Activities in Server Side Analytics

Tsengdar Lee, Ph.D.
High-end Computing Program Manager
NASA Headquarters

Presented at the ESGF/UVCDAT Conference
Lawrence Livermore National Laboratory
December 9, 2014

The Challenges ESGF is Facing

SCIENTIFIC

- Documenting the status and behavior of the Earth system and its multiple, interacting components
- Documenting the evolution of the Earth system and providing understanding of the sources of that evolution
- Supporting the projection of the future evolution of the Earth system
- Making Earth system science data easily available to users for both scientific and societal purposes

GROWNING DATA VOLUME

- *Satellite Data:* consider a global imager with 250 m resolution measuring once per day at 30 wavelengths for a year - ~ 10^{14} pixels/year
- *Model Output:* consider a chemistry/climate model, with 1^{o}x1^{o} resolution and 50 layers, writing out 30 parameters at hourly intervals for a year - ~ 10^{12} results written/year

COMMUNITY

- *Research Community*: scientific researchers looking to answer fundamental questions about the Earth
- *Assessment Community*: researchers of all types looking to document information about prior and future evolution of the Earth system to inform long-term policy and decision making
- *Forecasting Community*: operational scientists and others looking to provide forecasts to the general public
- *Applications Community*: research, corporate, and non-governmental organizations looking to inform nearer term decisions for management and planning

2

Typical ESGF Data Analysis and Data Processing Work Loads

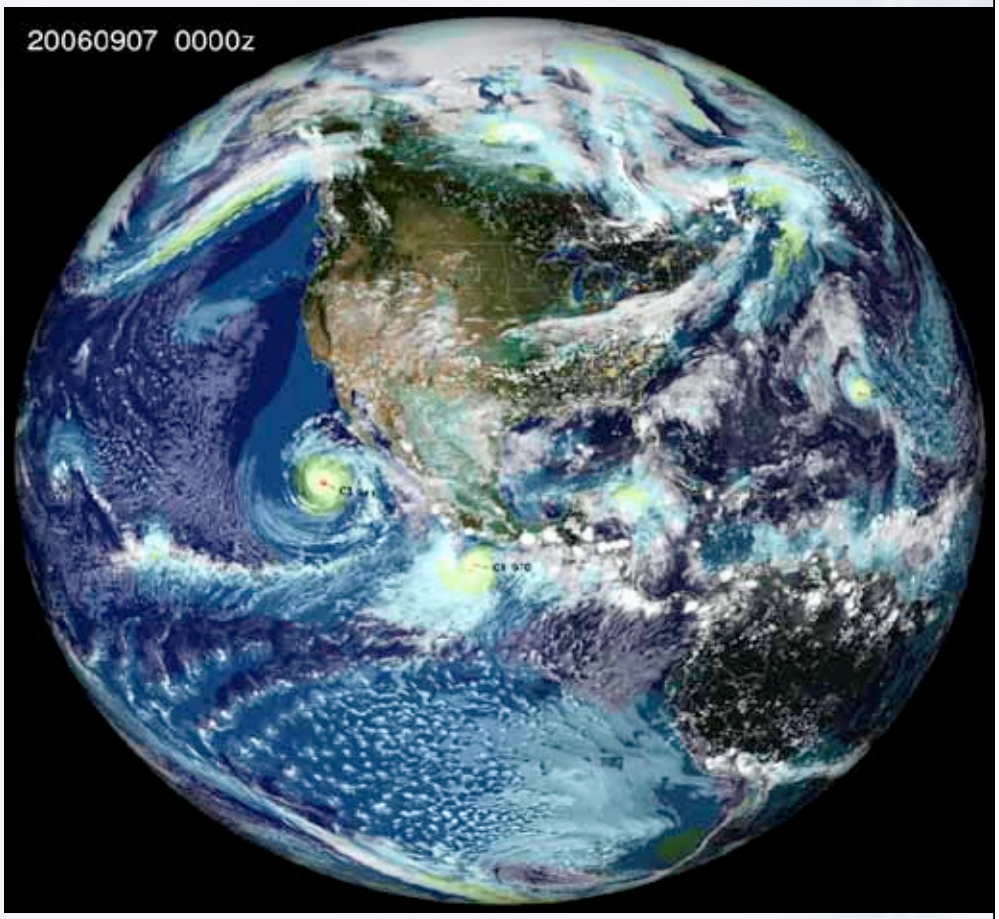
- A scientist or engineer queries a metadata server for the data and orders the data from a data center.
- The data center fulfills the order by preparing (subsetting, resampling, averaging etc.) the data and puts the result on a FTP server.
- After receiving a notification from the data center, the investigator goes to the FTP server and fetches the data.
- Data is transmitted to the investigator's institution and stored on a local storage.
- The investigator processes and analyzes the data locally using local computing resources.
- Some of the processed data will have to transmitted back to the data center.

7-km GEOS-5 Nature Run

Global Tropical Cyclones

- Nature run is used in Observation System Simulation Experiment. This GEOS-5 Nature Run successfully reproduces typical tropical cyclone activity in all basins including a large number of weak tropical storms as well as major hurricanes and typhoons.

- This 2 years simulation was done with 7200 cores for 75 days and generated 2 PB of simulation output.



In this short period from September 7-12, 2006 during the GEOS-5 7-km Nature Run, two hurricanes spin through the east Pacific basin while a major Atlantic hurricane develops in the Gulf of Mexico making landfall along the US Gulf coast as a category 3 hurricane with winds (color shading) in excess of 110 mph.

Use cases for Scientific Information System

- Scientists and engineers often use computing services to perform data analysis, theory verification, and predictions.
- Often move large volume of data to and from data centers and to and from compute centers.
- Need to communicate, collaborate, and share data with external (e.g. university) investigators.
- Require high speed connections and high speed computing platforms beyond business administration requirements for transferring large files.
- Require local disk storage and visualization HW and SW.

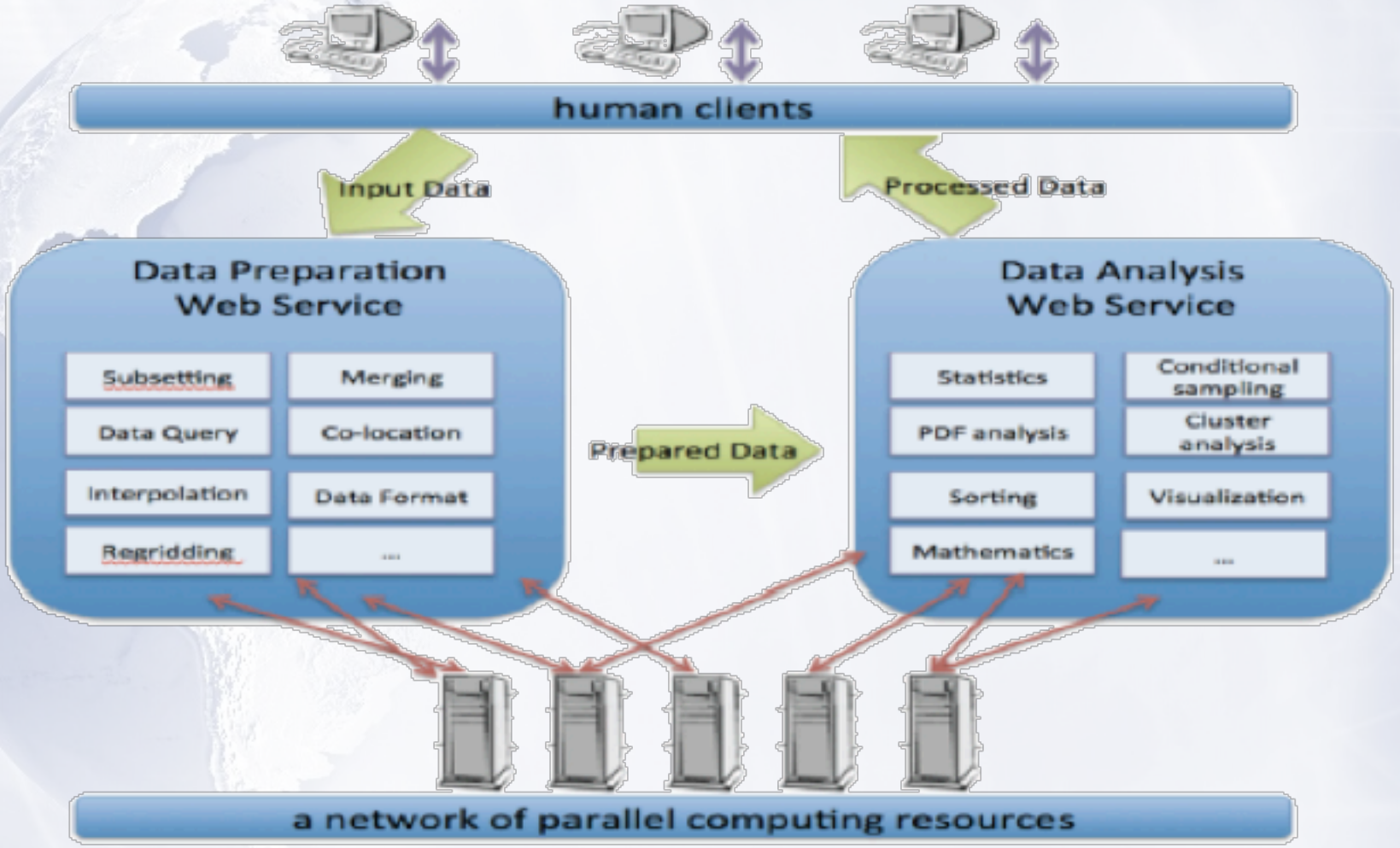
Analogy and Challenges

Analogy:

Challenges:

- Stewardship
- Curation
- Indexing
- Cataloging
- Searching
- Ordering
- Subsetting
- Provenance
- Lineage
- Data Mining
- Dissemination

Functional Architecture

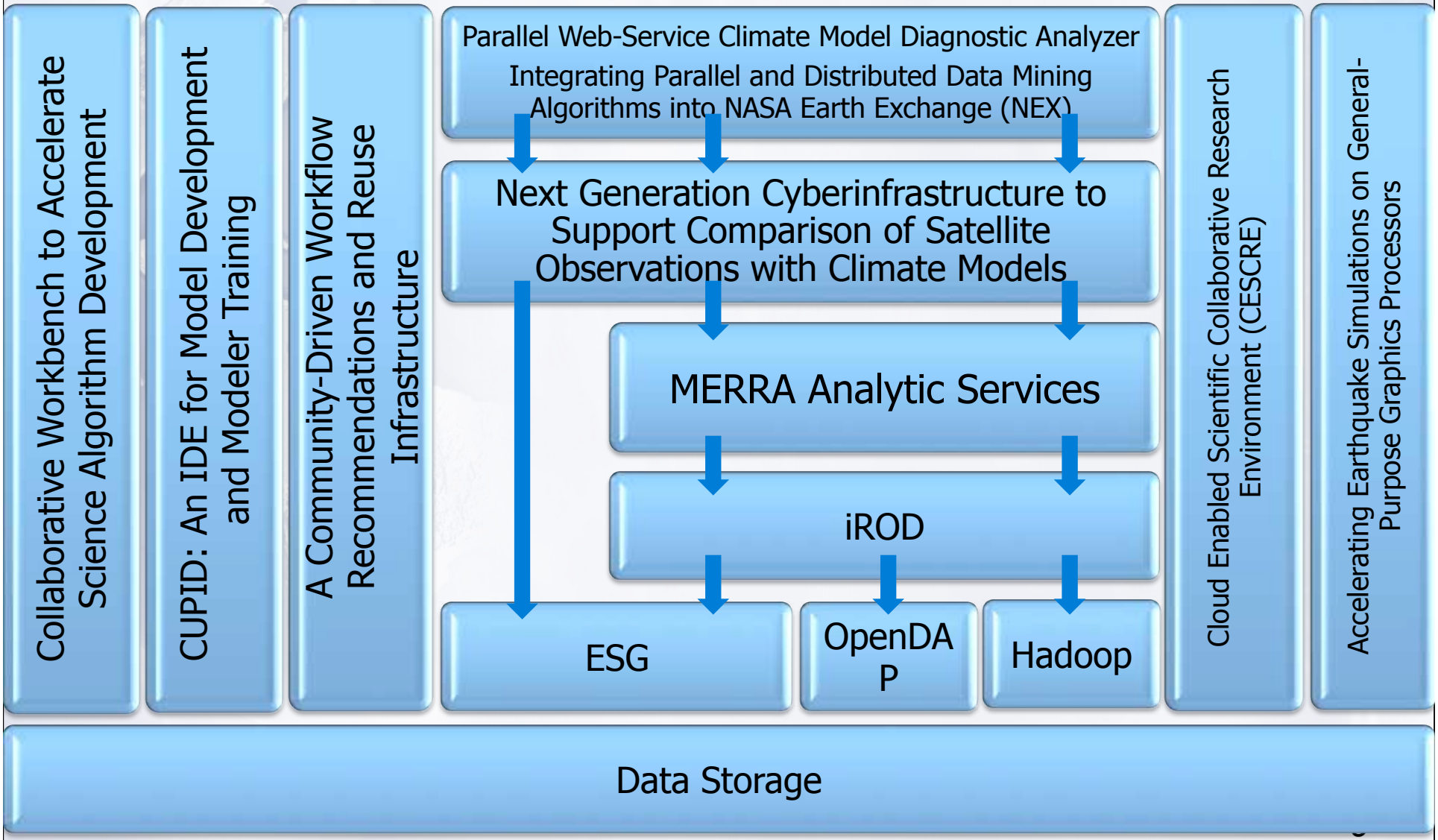


Challenges in Server Side Analytics

Challenges:

- Remote and local data visualization
- Server side processing capacity
- Data Mining
- Machine Learning
- Distributed data analysis
- Data on-boarding
- ETL
- High speed network
- Data management
- Data storage

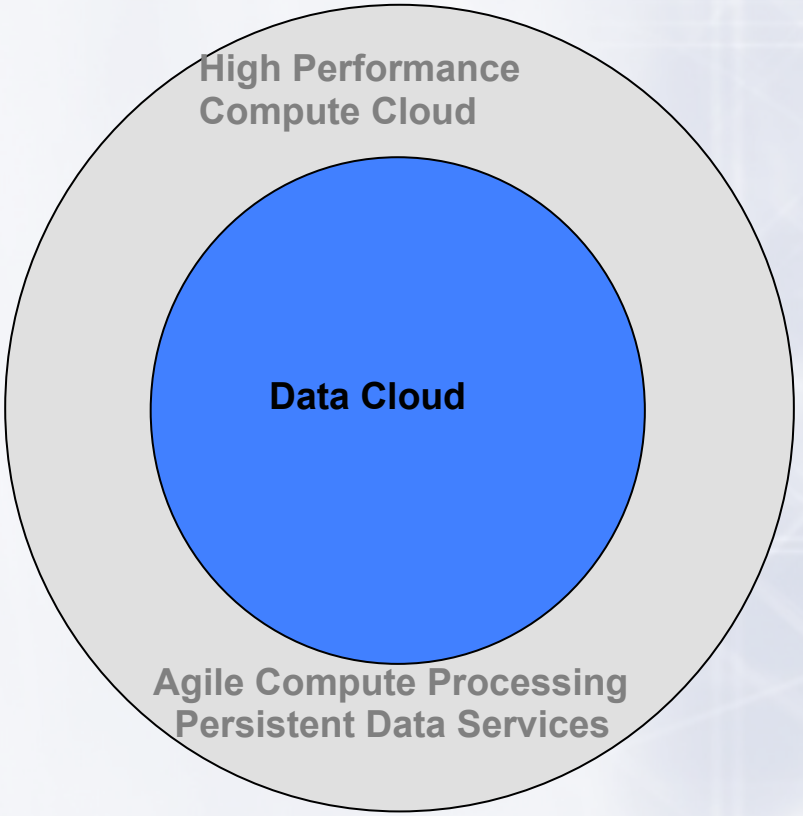
Notional Architecture



ABoVE Science Cloud

ABoVE Science Cloud

- Make use of the NCCS High Performance Science Cloud (HPSC)
  - Unified Data Analysis Platform that provides a colocation of data, compute, data management, and data services
  - Low barrier to entry for scientists; customized run time environments; agile environment
- Joint activity of the CCE, CISTO, and the NCCS
  - High performance data and compute for ABoVE scientists, algorithms, models, observations, analytics
- Data storage surrounded by a compute cloud
  - Large amount of data storage, high performance compute capabilities, very high speed interconnects



User Software Stack

Software Package		Comments
Operating System - Linux	Open Source	NCCS standard support
Operating System - Windows	Proprietary	May require licenses for long term support
IDL and Matlab	Proprietary	Will require licenses; how many and what packages are required?
ArcGIS Desktop, Server, Portal	Proprietary	NASA wide license
R, Python, HDF (4 and 5), GDAL, Geotiff, NetCDF	Open Source	NCCS standard support based on scientists needs
Data Management - Ramadda	Open Source	Could use the science cloud as a platform to perform an analysis of alternatives for data management
Data Management – iRODS	Open Source	NCCS supported platform

Example Software Stacks and Responsibilities

Application	Custom Science App	Application	Persistent Data Service
Commercial	IDL, Matlab	Commercial	IDL, Matlab
Libraries/Tools	e.g. R, Python, HDF, GDAL, Geotiff, NetCDF	Libraries/Tools	e.g. R, Python, HDF, GDAL, Geotiff, NetCDF
File System	RedHat Gluster	File System	RedHat Gluster
Interconnect	OpenIB and TCP/IP	Interconnect	OpenIB and TCP/IP
Guest VM	Linux (Debian, Centos, Windows)	Guest VM	Linux (Debian, Centos, Windows)

File System	RedHat Gluster
Hypervisor	KVM
Internal Interconnect	Open Fabrics Infiniband (OpenIB)
Operating System	CentOS (Linux)

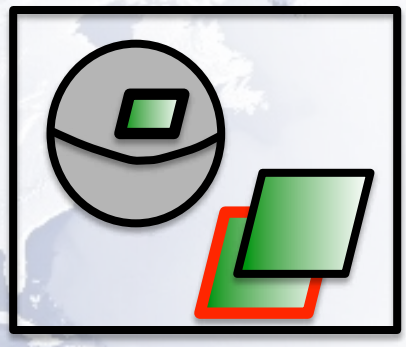
Responsibilities

- Scientists
- NCCS

Future Directions and Challenges

- **Scale with "Big Data" produced by higher resolution models, satellites, and instruments**
- **Expand server-side functionality**
  - Server-side processing through WPS (climate indexes, custom algorithms); GIS mapping services (for climate change impact studies at regional and local scale); Facilitate model to observations inter-comparison
  - Due to the distributed data centers, new algorithms will be necessary when there is only partial sample at each of the data centers
- **Expand direct client access capabilities**
  - Increased support for remote data access; Track provenance of complex processing workflows for reproducibility and repeatability
- **Package VMs for Cloud deployment**
  - Instantiate data grid nodes on demand for short lifetime projects; Environment with elastic allocation of back-end storage and computing resources

Open Source Strategy



RCMES

APACHE OPEN CLIMATE WORKBENCH

APACHE OODT

Thought about ESGF Governance

- If it is a federation, it needs to be governed by the members of the federation.
- It needs to allow autonomous processes independent of funding agencies.
- Current governance model is good in the VERY near term.
- The ESGF community and the funding agencies need to think about how this may be hand off to the community.
- We should probably take a look of some hybrid models (e.g. ESIP).

Final Thoughts

- ESGF needs to co-exist with other systems.
- WGCM needs to recognize there are other communities out there.
- We need to figure out how to leverage public cloud computing architecture.

Thank You!

Tsengdar Lee, Ph.D.
High-end Computing Program Manager
Weather Focus Area Program Scientist
NASA Headquarters
tsengdar.lee@nasa.gov