Agenda

• Ames Overview
• Big Data Challenges for Users
• NASA Supercomputing – HECC project
• Big Data related projects @ NAS
Ames Research Center

- Occupants: ~1130 civil servants; ~2,100 contractors; 1,650 tenants; ~1344 summer students in 2015
- FY2016 Budget: ~$915M (including reimbursable/EUL)
- ~1,900 acres (400 acres security perimeter): 5M building ft²
- Airfield: ~9,000 and 8,000 ft runways
Partnerships at Ames

- Partnering with external organizations to access capabilities under collaborative agreements
- Entering into reimbursable agreements for partner access to NASA capabilities
- Expanding overall landscape of space activity (maximizing public and private sector growth)
- Spurring innovation

**International**

**Commercial**

**Virtual Institutes**

**Interagency**

**Military**

**NASA Research Park**

**Academia**

**NASA Centers**

[Image with logos and symbols representing various international, commercial, virtual institutes, interagency, military, NASA Research Park, Academia, and NASA Centers]
Core Competencies at Ames Today

- Air Traffic Management
- Entry Systems
- Space and Earth Sciences
- Intelligent/Adaptive Systems
- Cost-Effective Space Missions
- Aerosciences
- Astrobiology and Life Science
- Advanced Computing & IT Systems
National Strategic Computing Initiative

Executive Order -- Creating a National Strategic Computing Initiative, July 2015

Objectives:

1. Accelerate delivery of a capable exascale computing system delivering approximately 100 times the performance of current systems across a range of applications.

2. Increase coherence between the technology base used for modeling and simulation and that used for data analytic computing.

3. Establish a viable path forward for future HPC systems even after the limits of current semiconductor technology are reached (the "post-Moore's Law era").

4. Increase the capacity and capability of an enduring national HPC ecosystem.

5. Develop an enduring public-private collaboration to ensure that the benefits of the research and development advances are shared among government, industrial, and academic sectors.
High-End Computing Capability (HECC)

**NASA’s Premier Supercomputer Center**

Resources have broad mission impact across all of NASA’s Mission Directorates
Over 500 science & engineering projects with more than 1,500 users
(hosted by the NASA Advanced Supercomputing (NAS) Division at Ames)

- **Pleiades** – 7.25 PF peak
  - Distributed memory cluster SGI Altix ICE
  - 246K-core; 11472 nodes; 4 Xeon generations
  - #15 (#7 in US) on TOP500; #9 in HPCG list (06/2016)

- **Specialized Hardware**
  - **Endeavour**: shared memory nodes – 1024 core 4 TB & 512 core 2 TB
  - **GPGPU nodes**: 64 nodes NVIDIA Tesla K40
  - **Xeon Phi**: 20 many-integrated core nodes
  - **NVIDIA DGX-1**: 8 Tesla Pascale GPUs for machine learning

- **Storage**: ~30 PB disk; ~500 PB tape capacity
- **Networking**: 10 Gb/s external peering

2 bldgs, 4 floors
26,600 sq. ft.
6.25 MW elec. power
2160 tons cooling
Integrated Spiral Support for MS&A

Develop and deliver the most productive integrated supercomputing environment in the world, enabling NASA to extend technology, expand knowledge, and explore the universe.

**NASA Mission Challenges**

Scientists and engineers plan computational analyses, selecting the best-suited codes to address NASA’s complex mission challenges.

**Performance Optimization**

NAS software experts utilize tools to parallelize and optimize codes, dramatically increasing simulation performance while decreasing turn-around time.

**Data Analysis and Visualization**

NAS visualization experts apply advanced data analysis and rendering techniques to help users explore and understand large, complex computational results.

**Computational Modeling, Simulation, & Analysis**

NAS support staff help users to productively utilize NASA’s supercomputing environment (hardware, software, networks, and storage) to rapidly solve large computational problems.

**Outcome:** Dramatically enhanced understanding and insight, accelerated science and engineering, and increased mission safety and performance.
Big Data Challenges for NASA Users

NASA supports enormous collections of big data sets:

**Observational Data**
- Estimate 100+ active satellites producing 50PBs per year

**Model Data**
- NAS has 30 PBs of online storage- MITGcm run produced > 3PBs

**Experimental Data**
- Wind tunnel tests projected to produce 100 TBs per test

---

**Data Discovery** – finding what data is available and where
- Indexing, federated metadata service and semantic reasoning

**Data management** – transferring very large data sets from archives to computational resources
- Increased WAN bandwidth
- Fault tolerant and resilient hardware/software infrastructure

**Tools/models/algorithms** – developing analytics/analysis software at scale
- Mechanisms for sharing software to reduce duplication

**Analysis workflow** – increasing complexity of processing pipelines have multiple components requiring heterogeneous resources
- Software for workflow description and management to tie all components together and facilitate re-use

**Analysis/Analytics infrastructure** – inadequacy of available resources
- I/O infrastructure
- Large memory spaces for in-core analysis
- Support for the heterogeneous resources in an integrated environment: distributed memory & shared memory systems, hadoop cluster, accelerators, FPGAs etc.

**Data Dissemination** – difficult to share knowledge across a wider community
- Support for dissemination and sharing of code, data products, results, etc.....
Big Data Challenges for NASA Users

Data Discovery – finding what data is available and where
• Indexing, federated metadata service and semantic reasoning

Data management – transferring very large data sets from archives to
• Increased WAN bandwidth
• Fault tolerant and resilient hardware/software infrastructure

Tools/models/algorithms – developing analytics/analysis software at scale
• Mechanisms for sharing software to reduce duplication

Analysis workflow – increasing complexity of processing pipelines have multiple
• Software for workflow description and management to tie all components together and
  facilitate re-use

Analysis/Analytics infrastructure – inadequacy of available resources
• I/O infrastructure
• Large memory spaces for in-core analysis
• Support for the heterogeneous resources in an integrated environment: distributed
  memory & shared memory systems, hadoop cluster, accelerators, FPGAs etc.

Data Dissemination – difficult to share knowledge across a wider community
• Support for dissemination and sharing of code, data products, results, etc.

NASA supports enormous collections of big data sets:

Observational Data
• Estimate 100+ active satellites producing 50PBs per year

Model Data
• NAS has 30 PBs of online storage- MITGcm run produced > 3PBs

Experimental Data
• Wind tunnel tests projected to produce 100 TBs per test

Fun Fact: The term “Big Data” was first used by Michael Cox & David Ellsworth of the NAS Division at Ames in their paper: “Visualizing flow around an airframe” Visualization 97, Phoenix AZ.

- Biggest data set considered 7.5 GB; high-end analysis machines had less than 1GB memory

Merging HPC and Data Analysis @ NAS: Data Intensive Supercomputing Environment

Collaboration / User Portals

Cloud-based data servers, web servers

Data Resources

• User & community Datasets
• Shared across platforms

Analytics/Analysis Platforms

• Optimized for analytics/analysis
  - Visualization engines
  - Accelerators
  - FPGAs
• Tools for data/workflow management; machine learning

Large-scale Compute

Pleiades
Endeavour

Optimized for large-scale simulation and analysis

External Data Resources

High-speed WAN connection

• Data Repositories
• Experimental sources (labs, wind tunnels, etc.)
Big Data Related Projects @ NAS

Applications
- Mining network flows for cyber-security risk assessment
- Tree cover classification for continental US
- ...

Analytics
- Machine learning algorithms using GPUs/Xeon Phis/FPGAs
- Large scale analytics on distributed and shared memory architectures

Support Software
- Collaboration platform for scientists (NASA Earth eXchange - NEX)
- Ontology-based data search environment for observational data
- Data tagging for security and data discovery

System Software
- Optimization of Lustre calls to improve QOS for applications
- Support for caching using SSDs
  - Lustre - target metadata and/or user access

Hardware Infrastructure
- SSDs for I/O optimization – linux-based block device software
- Specialized clusters for data analytics
  - Xeon PHI, GPU, FPGA, Hadoop
Summary

- NASA has an abundance of big data: *Observational, Simulation and Experimental*
- NASA Big Data users face many challenges across the full workflow for analyzing such data:
  - Data discovery, data access & management, analytics/analysis algorithms and software, infrastructure, data dissemination
- Ames an the ideal location for merging HPC and Data Analytics since it hosts the Agency’s premier supercomputer
- Several of the Ames projects are aimed at filling the gaps in the integrated software/hardware environment for Big Data Analysis
Questions?

piyush.mehrotra@nasa.gov
Backup Slides
SSD Support for Data Analysis

**Goal:** Assess the benefits of utilizing Solid State Devices (SSDs) for handling Big Data in the HPC environment at NAS

**Hardware:**
- Hyperwall: 128 Intel NVMe P3500 2 TB medium-durability SSDs
- Lustre file system: 6 OSSs augmented with Intel NVMe P3600 1.6 TB high-durability SSDs

**Focus research areas – utilize SSDs for:**
- Caching for Lustre-based global file system (in collaboration with Intel)
  - Metadata on Lustre OSS
  - Data for specific job id or user
  - Sequential streams
- Caching for Applications on hyperwall - SSDs used
  - As local disk drives
  - As shared file system over 128 hyperwall nodes using RDMA access to remote SSD over Infiniband
NAS Situational Awareness System (NSAS)

**Goal:** to identify actionable security events that require human or automated mitigation based on an analysis of the mountain of network data that flows in and out of NAS.

- Data sources: Bi-directional network flow data, intrusion detection data, log data, Nessus vulnerability scanner data, Domain Name Server requests, etc.
- Analyst dashboard to keep track of and deep dive into information
- Utilizing data analytics and machine learning techniques on flow data along with user’s and system network behavior profiles to detect:
  - phishing attacks
  - Signs of possible exfiltration
  - Advanced persistent threats (APTs)
Data Tagging for Security and Discovery (DTSD)

**Goal:** to develop base requirements and prototype a data-centric approach to tag data so as to provide

- Information for protecting the data from a security perspective
- Information that describes the data from a semantic perspective.

- Security restrictions embodied in the data tags will allow NASA systems that handle the data to automatically
  - Enforce access to the data based on the tag
  - Enforce flow restrictions based on the tag, e.g., not releasing unencrypted ITAR data to Internet

- Semantic information associated with the data tags will describe the characteristics of the data
  - Support semantics-based data discovery tool
ODISEES & OlyMPUS

Goal: Ontology-based interactive framework for discovery of Earth science data
LaRC (Science Directorate), GSFC (NCCS), and ARC (NAS)

- ODISEES
  - enables parameter-level search with little knowledge about the data
  - extensible to address additional datasets by extending the ontology
  - implements a flexible architecture that can be adapted for other domains

- OlyMPUS: extends ODISEES with a metadata provisioning portal for data providers along with enhanced search capabilities for data consumers