Hybrid Computing Environments:

Implementations, Issues, Futures

Bob.Ciotti@nasa.gov
Chief Technologist
NASA Advanced Supercomputing
NASA Ames Research Center
www.nas.nasa.gov

SOS-25 Workshop
March 2023

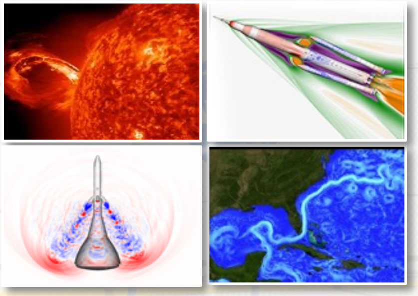
High End Computing Capability (HECC) Project

HECC Hardware

4 Compute Clusters

- Pleiades 158 Racks / 11,215 nodes / 7.85 PF / 8,222 SBU/hr
- Electra 24 Racks / 3,422 nodes / 8.32 PF / 4,815 SBU/hr
- Aitken 12 Cells / 3,200 nodes / 9.07 PF / 10,204 SBU/hr
- Endeavour 3 Racks / 2 nodes / 32 TF / 44 SBU/hr

1 Visualization Cluster	128 node viz 'wall'
15 Lustre File Systems	**130 PB**
6 NFS File Systems	**1.5 PB**
1 VAST NFS	**5 PB**
2 BeeGFS SSD	**2 PB**
1 BeeGFS HDD	**1.5 PB**
Archive Capacity	1,000 PB



~10 MW power

NAS Facility Extension

- A 1-acre site with 30 MW power to house HPC systems in modules

HECC Services

HECC provides a suite of complimentary services to the user community to enhance the scientific and engineering results obtained from the hardware assets.

- Systems – Customized solutions including compute and storage solutions to meet specific project or mission requirements. Cloud access for immediate or non-standard computing.
- Application Performance and Productivity – Software solutions provided to research/engineering teams to better exploit installed systems.
- Visualization and Data Analysis – Custom visualization during traditional post-processing or concurrent during simulation to understand complex interactions of data.
- Network – End-to-end network performance enhancements for user communities throughout the world.
- Data Analytics – Exploitation of data sets through neural nets and emerging new techniques.
- Machine Learning – Custom environments to enable learning through advanced data techniques.
- Custom Data Gateways – Custom data portals to support diverse programs and projects

Problem/Systems Evolution

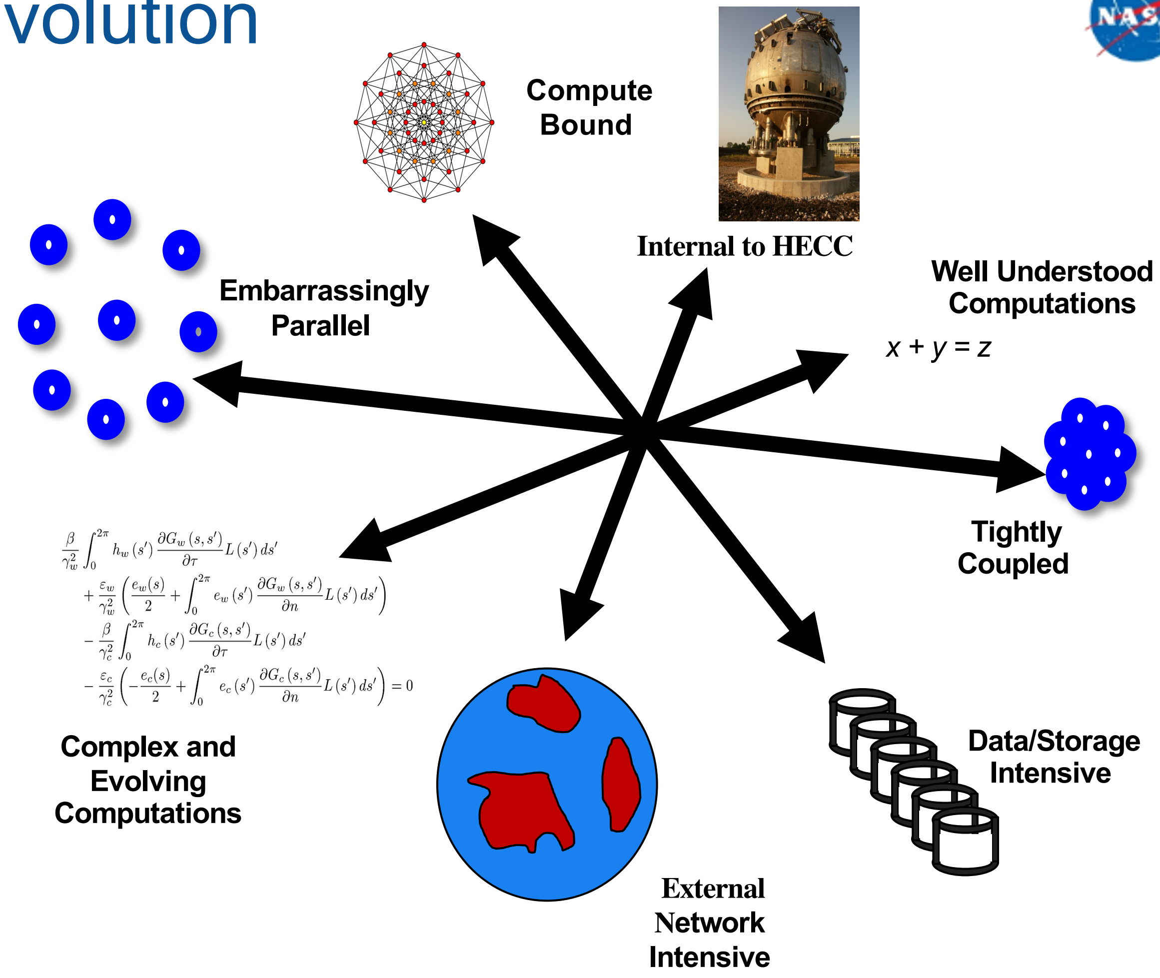
Classic view of HPC

- Large/sometimes singular computing problem
- Standalone/self contained - few external dependencies
- Detailed construction of static inputs
- Results interpreted locally
- Compute the more critical resource

Major shift to data analysis and its distribution

- Logarithmic increases in computing capabilities - 'Moore's law' slows
- Exponential increases in volumes of data produced and/or analyzed
- Wide geographic/geopolitical distribution of ***compute providers*** and ***users***
- Wide geographic/geopolitical distribution of ***generators*** and ***consumers*** of ***data*** products
- Shift from only compute being the most critical resource to adding ***distributing*** and ***understanding data***

How can HECC adapt to ingest, compute, save, and distribute de-centralized datasets?



2018 Cloud Report - Findings

NAS Technical Report:

- https://www.nas.nasa.gov/assets/nas/pdf/papers/NAS_Technical_Report_NAS-2018-01.pdf

- Finding 1: ***Tightly-coupled, multi-node applications*** *from the NASA workload* ***take somewhat more time*** *when run on cloud-based nodes*

- Finding 2: *The per-hour* ***full cost of HECC resources is cheaper*** *than the (compute- only)* ***spot price*** *of similar resources at AWS*

- Finding 3: ***Commercial clouds do not offer a viable, cost-effective approach*** *for replacing in-house HPC resources at NASA.*

- Finding 4: ***Commercial clouds provide a variety of resources not available at HECC****. Several use cases, such as machine learning, were identified that may be cost effective to run on commercial clouds.*

2018 Cloud Report - Findings - 2023 Update

NAS Technical Report:

- https://www.nas.nasa.gov/assets/nas/pdf/papers/NAS_Technical_Report_NAS-2018-01.pdf

- Finding 1: ***Tightly-coupled, multi-node applications*** *from the NASA workload* ~~***take somewhat more time***~~ ~~*when*~~ *run on cloud-based nodes (I/O often faster)*

- Finding 2: *The per-hour* ***full cost of HECC resources is cheaper*** *than the (compute- only)* ***spot price*** *of similar resources at AWS (costs improved/requires efficiency – data mgmt)*

- Finding 3: ***Commercial clouds do not offer a*** ~~***viable***~~***, cost-effective approach*** *for replacing in-house HPC resources at NASA.*

- Finding 4: ***Commercial clouds provide a variety of resources not available at HECC****. Several use cases, such as machine learning, were identified that may be cost effective to run on commercial clouds. (add high availability, content delivery, hybrid processing, multi-organization teams, role based security controls, cloud optimized software, …)*

2018 Cloud Report - Actions

NAS Technical Report:

- https://www.nas.nasa.gov/assets/nas/pdf/papers/NAS_Technical_Report_NAS-2018-01.pdf

- Action 1: ***Get a better understanding of the potential benefits and costs*** *that might accrue from running a portion of the HECC workload in the cloud.*

- Action 2: ***Define a comprehensive model that allows accurate comparisons of cost*** *between HECC in-house jobs and jobs running in the cloud.*

- Action 3*:* ***Prepare for a broadening of services offered by HECC to include*** *a portion of its workload running on commercial* ***cloud*** *resources.*

2018 Cloud Report - Actions - 2023 Update

NAS Technical Report:

- https://www.nas.nasa.gov/assets/nas/pdf/papers/NAS_Technical_Report_NAS-2018-01.pdf

Action 1: ***Get a better understanding of the potential benefits and costs*** *that might accrue from running a portion of the HECC workload in the cloud. (broadly applicable)*

Action 2: ***Define a comprehensive model that allows accurate comparisons of cost*** *between HECC in-house jobs and jobs running in the cloud. (incomplete for cloud)*

Action 3*:* ***Prepare for a broadening of services offered by HECC to include*** *a portion of its workload running on commercial* ***cloud*** *resources. (actively running)*

Why Hybrid/Multi-Cloud Architectures?

- Public Cloud is todays dominant IT solution
    - ***Incredibly robust and hardened – Easy to design High Availability (HA) solutions***
    - ***Software/Services (lots of it) are “cloud optimized” – minimal use is often free (turn key)***
    - ***Hardware is virtualized and easily provisioned***
    - ***Robust Security***
    - Best ”Distributed” Performance
    - Massive Reach – Worldwide – Persistent
    - High Capitalization - worldwide - multi billion industry investments
    - ***Fast Turnaround when needed***

- Private HPC resources and clouds still have a role
    - ***Users understand the environment – short runway to productivity***
    - ***Cost optimized - 1/3 to 1/8 the cost = SBU/Instance – more when including storage and staffing***
    - Minimal set of software tools can be fine tuned to your advantage
    - Some common use cases may be easier to operate and maintain
    - Problem Resolution
    - ***Predictable Cost***

Success Combines the Best of Both – Best of All

features, access, scale, resilience, disaster recovery, lower cost, platform independence, increased competition

Access requires your workloads to be distributed and easily moved around

SBU Cost Breakdown

SBU rates listed at https://hec.nasa.gov/user/policies/sbus.html

NAS/HECC

AWS

1 x 28 core Broadwell for one hour <= 1 x 64 vCPU m4.16xlarge

$3.20 on-demand

$1.98 1-year reserved

$1.38 3-year reserved

$1.33 spot

Annual HECC Budget / 75% of potential hours delivered

- Equipment Capitalization
- Scientific Consulting
- Visualization Services
- Networking
- Big Data Analytics
- 24/7 monitoring and helpline
- Staffing/Management
- Facility Maintenance
- Power/Cooling
- User on-boarding
- Project Allocation Management
- etc

SBU - 56 hyperthreaded vs AWS - 64 vCPU

SBU - 128GB vs AWS 256GB

AWS SKUs not on arc.intel.com

Historical Cost per HECC SBU2	
FY2023	**$0.30**
FY2022	$0.44
FY2021	$0.47
FY2020	$0.47
FY2019	$0.55
FY2018	**$0.65**
FY2017	$0.97
FY2016	$1.05
FY2015	$1.13
FY2014	$1.54
FY2013	$2.26
FY2012	$3.23

SBU vs Instance comparison

SBUs/Instance accounts for:

- > 18,000 compute nodes
- Systems Maintenance
- Facility Maintenance
- 10 MW Power/Cooling
- Some of the Systems Team

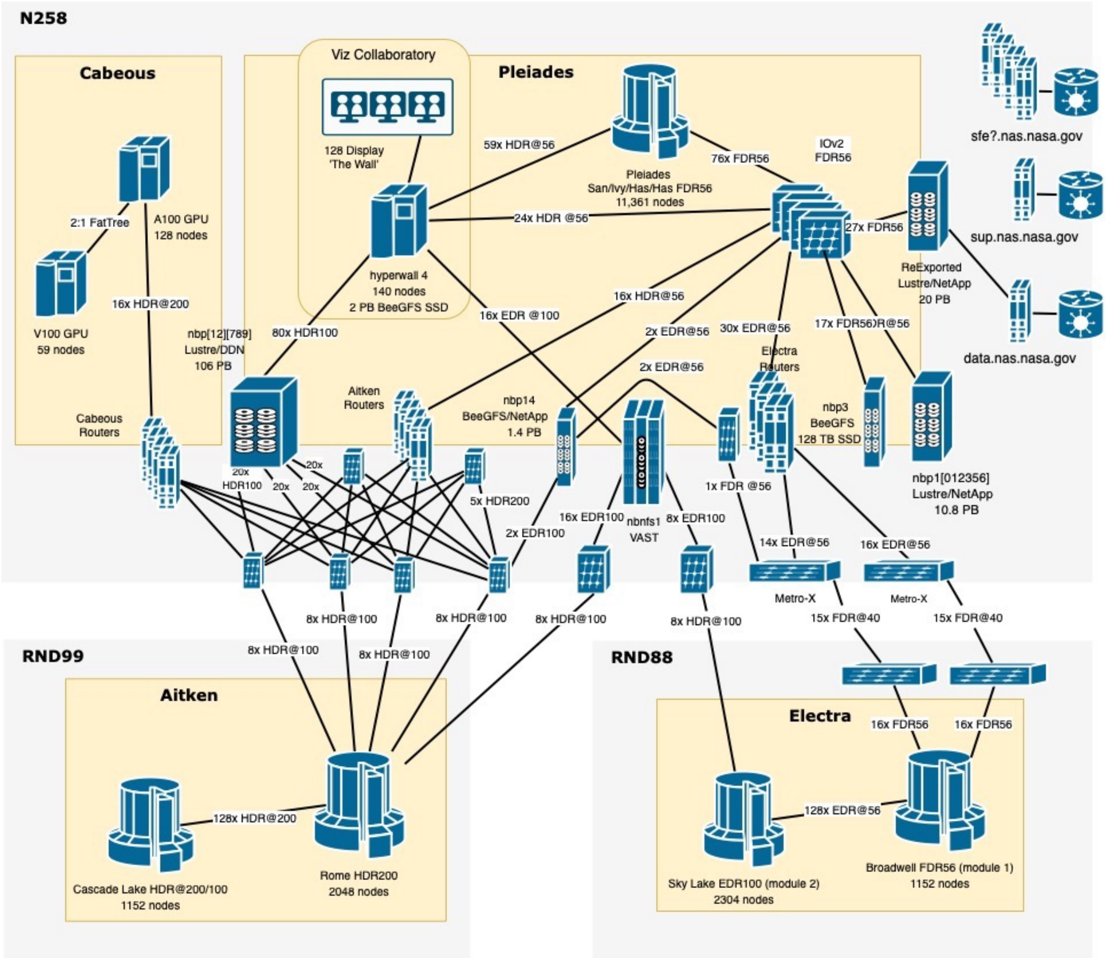
What's missing?

Storage/Network architecture

- >
- >
- >
- >
- >

Support Staff for:

-
-
-
-
-
-
-
-
-
-
-
-
-
-



SBU vs Instance comparison

SBUs/Instance accounts for:

- > 18,000 compute nodes
- Systems Maintenance
- Facility Maintenance
- 10 MW Power/Cooling
- Some of the Systems Team

What's missing?

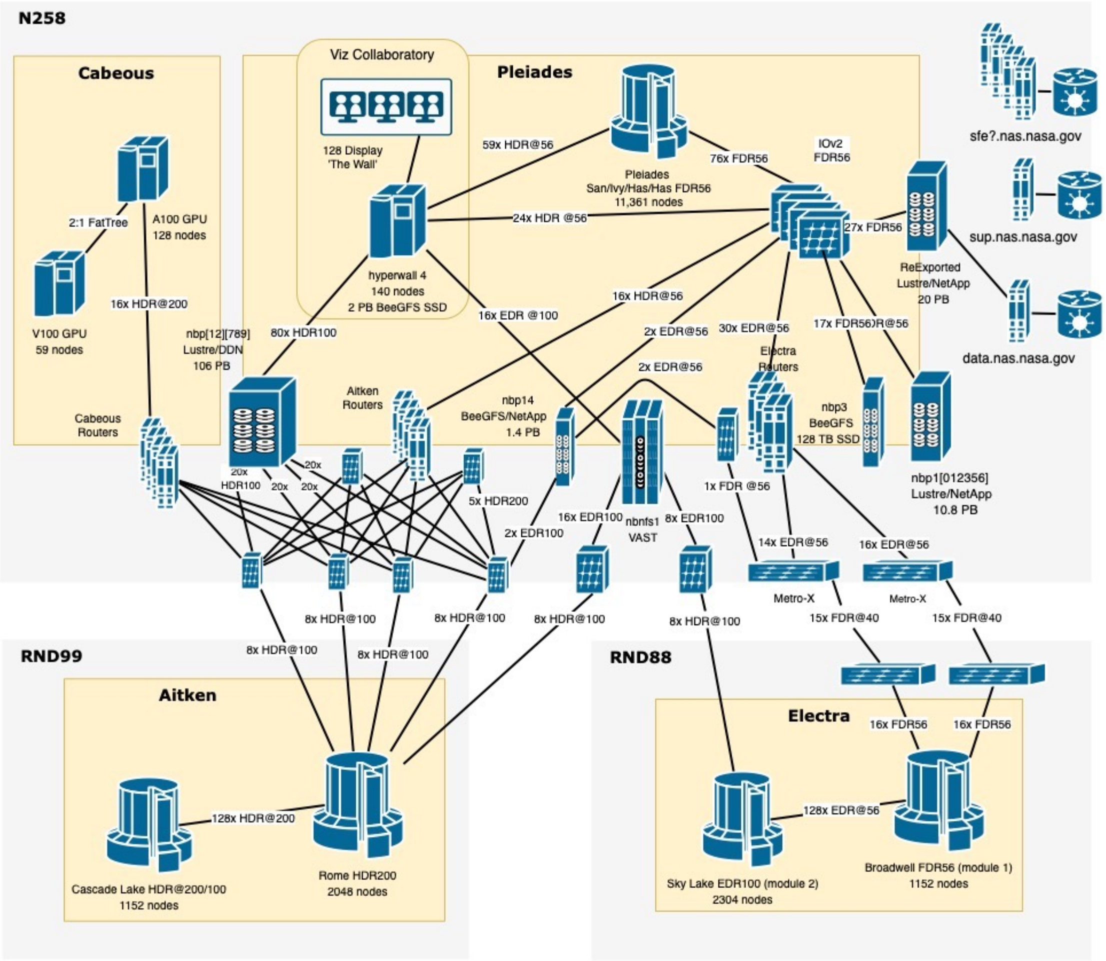
Storage/Network architecture (~$160 M retail list)

- > 6 PB SSD (~$11 M FSx)
- > 130 PB scratch (~$133 M FSx)
- > 250 PB tape archive (~$6 M glacier)
- > 20 PB public data export (~$5 M S3)
- > 50 PB WAN downloads (~$3 M egress)

Support Staff for:

- User on-boarding – 1,899 users (FY23)
- Project Allocation Process
- Resource Tracking/Enforcement
- Problem Resolution
- Scientific Consulting
- Visualization Services (hyperwall)
- Cloud Implementation
- Networking
- Desktops
- Big Data Analytics
- 24/7 monitoring and help desk
- Management
- Documentation
- Auditing

Software Licensing



SBU vs Instance comparison

SBUs/Instance accounts for:

- > 18,000 compute nodes
- Systems Maintenance
- Facility Maintenance
- 10 MW Power/Cooling
- Some of the Systems Team

What's missing?

Storage/Network architecture (~$160 M retail list)

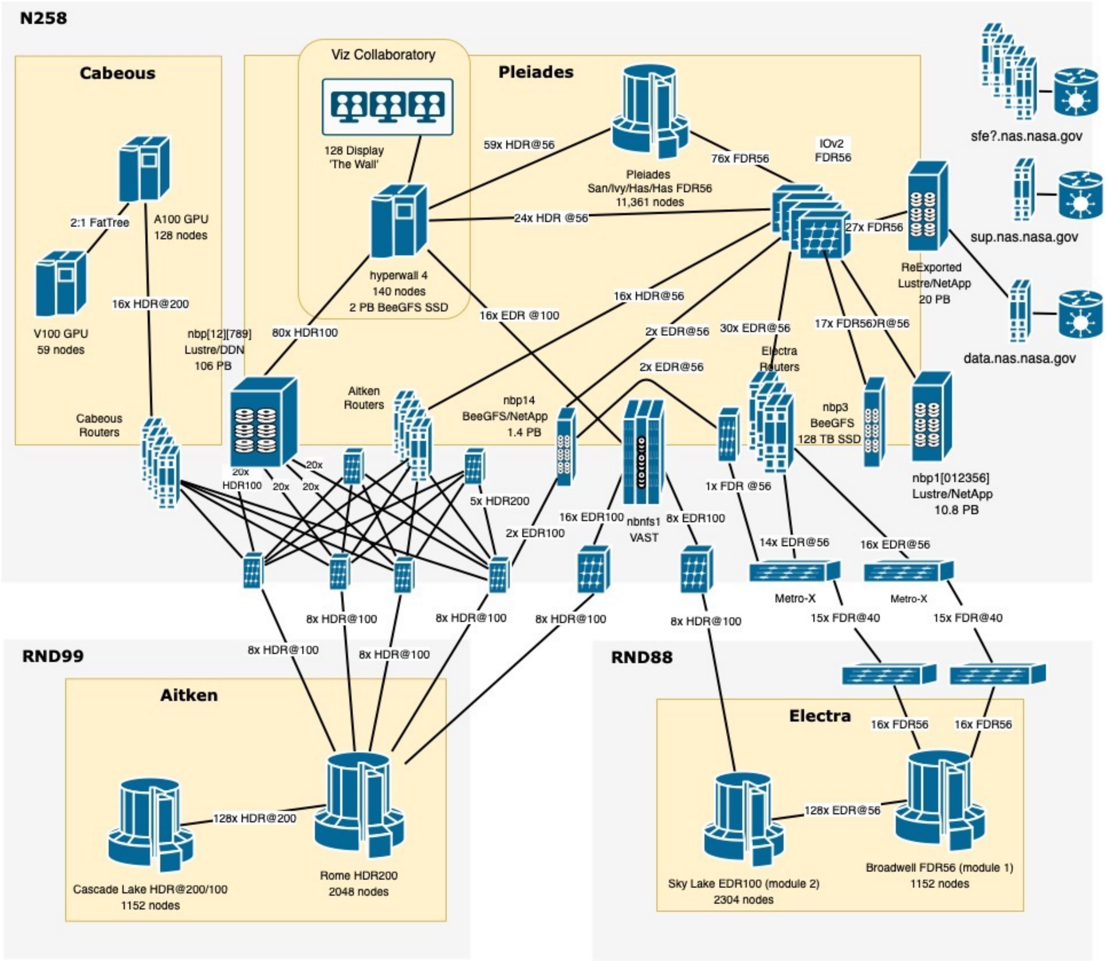
- > 6 PB SSD (~$11 M FSx)
- > 130 PB scratch (~$133 M FSx)
- > 250 PB tape archive (~$6 M glacier)
- > 20 PB public data export (~$5 M S3)
- > 50 PB WAN downloads (~$3 M egress)

Support Staff for:

- User on-boarding – 1,899 users (FY23)
- Project Allocation Process
- Resource Tracking/Enforcement
- Problem Resolution
- Scientific Consulting
- Visualization Services (hyperwall)
- Cloud Implementation
- Networking
- Desktops
- Big Data Analytics
- 24/7 monitoring and help desk
- Management
- Documentation
- Auditing

Software Licensing

Less Costly but Less Capable

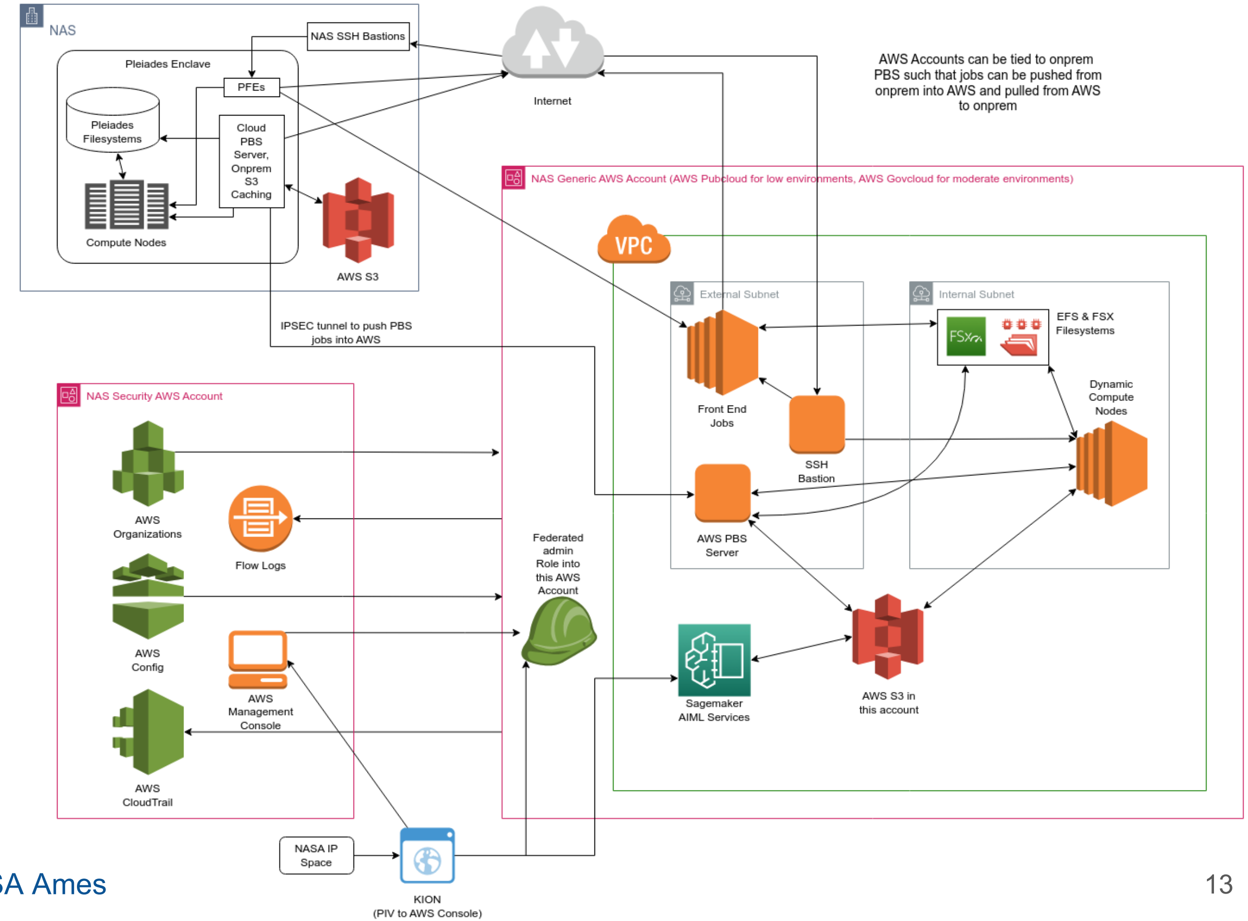


Implementation #1 -- NAS Hybrid Cloud

Three modes of operation

- Standalone AWS environment
- Hybrid burst out
  submit batch and interactive jobs from our on prem PFE to run in AWS
- Hybrid burst in
  submit batch and interactive jobs in AWS to run on-prem

Can support up to EAR99/ITAR.



NAS Cloud Use Cases

- Burst out high priority work
  - LAVA
    - Walled Model LES

- Burst in – (use case in development)
  - Air Quality/Fire Spread - USFS fire sciences lab
    - Ensemble Atmospheric modeling + Fire spreading models
    - Need to run many scenarios - scale out
    - Higher Fidelity
    - ***Affordability a big issue***
    - HECC Capability – Role in Regional/National scale emergency?

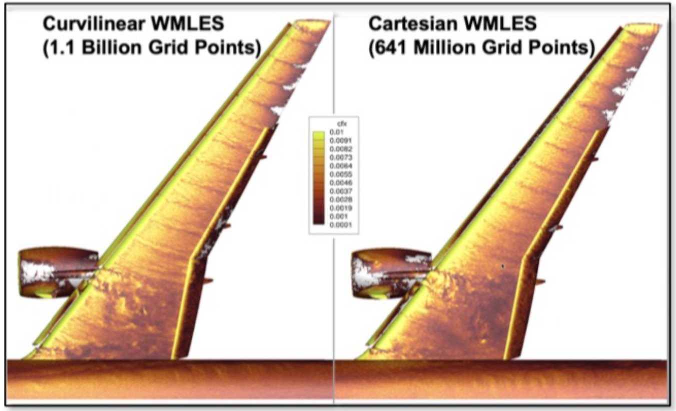
- Student Cluster Competition
  - Broaden community engagement and promote NASA Open Science Goals
  - Quickly build/deploy isolated infrastructure
  - Foreign Nationals/Young Researchers
  - Short term engagement

All fall under a FISMA moderate plan
  - Comes with additional controls and scrutiny

LAVA WMLES Studies of the High-Lift CRM on HECC Cloud

- Utilizing the HECC Cloud to access Amazon Web Services (AWS) resources, in addition to in-house HECC resources, the LAVA team carried out Wall-Modeled Large Eddy Simulation (WMLES) studies of the High-Lift Common Research Model (HL-CRM) and met important deadlines for the AIAA Scitech CFD 2030 special session and the High-Lift Prediction Workshop data submission.
- The team found that per-node performance and interconnect of AWS resources are comparable to or even better than HECC systems. In addition:
  - At most, 160 c5n-type AWS nodes were utilized, with the queue wait time never exceeding five minutes.
  - Visualization and post-processing can both be completed in the cloud, eliminating the need for large data transfers and dedicated post-processing workstations.
  - On-demand cloud resources are approximately 5-8 times more expensive than in-house HECC resources.
- Evaluation of other node types, such as p4d (with 8xA100 GPUs) is planned for 2022 to facilitate rapid development and debugging of LAVA GPU capabilities, and to perform scaling studies and provide feedback to HECC management.

IMPACT: Availability of on-demand HECC Cloud resources enables important NASA applications to be performed in a timely matter in order to meet critical milestones.



Instantaneous surface skin friction (c_{fx}) at CLmax (α=19.57°) predicted from LAVA for two WMLES cases utilizing AWS resources and in-house HECC resources. *Gaetan Kenway, Aditya Ghate, NASA/Ames*

PBS - Integrated Data Staging/Caching Prototype

- In anticipation of NASA's Distributed Active Archives Centers (DAACs) move to the cloud and to meet the agency's future need for hybrid computing and data processing environments .

- Team prototyped a PBS scheduler-integrated data caching mechanism.

- The effort enables HECC users to use distributed datasets more efficiently with the on-prem HPC systems.

- The prototyping effort for this milestone included:
  - PBS job prolog directive enhancements: users can specify new directives that will trigger a data pull from select external cloud sources automatically before a job begins.
  - Stage and Cache requested data
  - Cache Aging and Recovery

- A facility-wide world readable cache
  - multiple independent users/teams transparently leverage cloud-based data
  - reducing egress, storage duplication, and overall costs
  - reducing personnel data-wrangling tasks.

- **IMPACT**: The scheduler-integrated cache capability 1) can reduce agency cloud-based data egress and local storage costs by minimizing duplicate data movements and storage of data standard products 2) can reduce external data access latency for HECC users, 3) can improve local resiliency for higher-order workflow systems that drive data intensive processes which may leverage off-prem data sources 4) reduce data wrangling for both data staging and local data cleanups with data processing, product generation, and analytics efforts.

User Story: In early December, initial feedback from a NASA Earth Exchange (NEX) user was very positive. A considerable reduction in overall task times, from three weeks to just five days, was achieved using the prototype. The NEX user test utilized 87 terabytes of data and 6,628,350 file objects of Landsat collection 2 data, with the source from an s3://usgs-landsat bucket within the AWS Oregon us-west-2 region, which the USGS EROS Center maintains

PBS Staging/Caching Directives

Directives of the form:

#CLOUD -s3cache=s3://noaa-goes17/ABI-L2-SSTF/2019/358/c20193582204141.nc

Will show up as the file:

/nobackup/s3cache/noaa-goes17/ABI-L2-SSTF/2019/358/c20193582204141.nc

Supports stage-in and stage-out from:

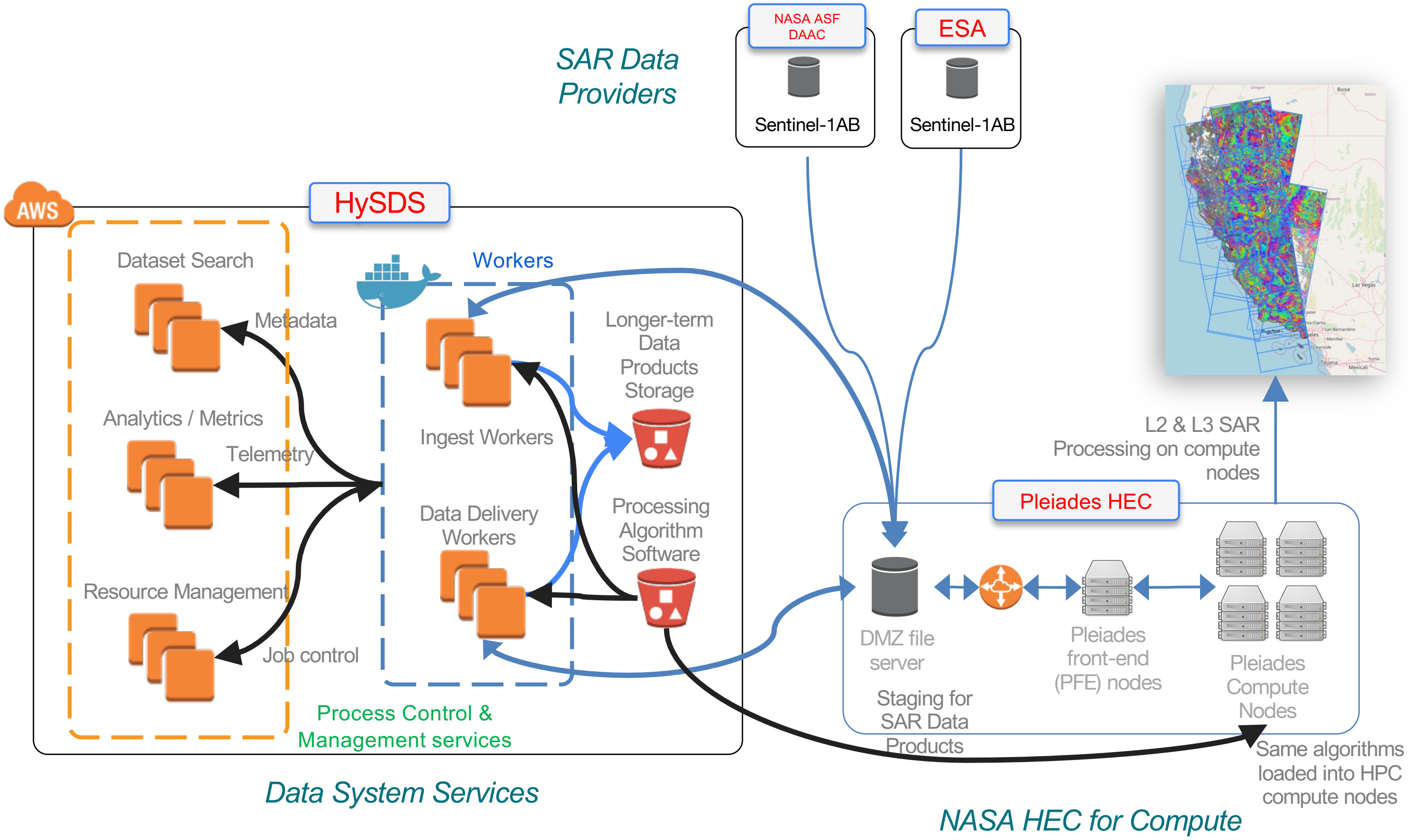
NAS Local filesystems to/from cloud environments

S3 (does integrity checks based on boto calls)

http

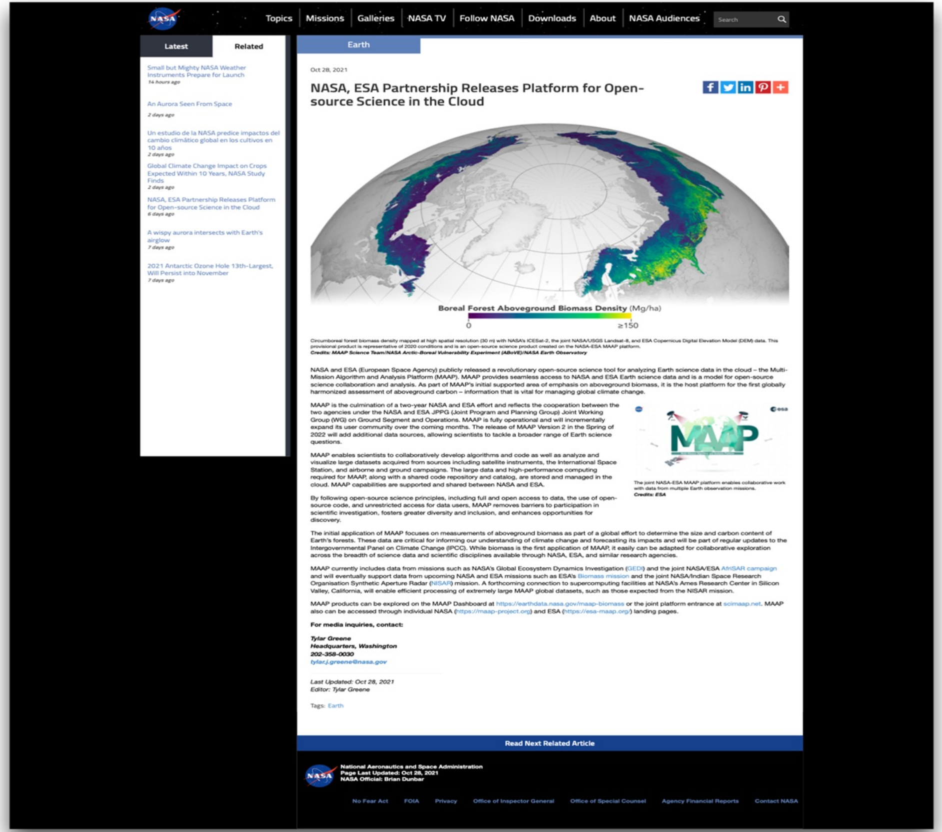
Working to implement same staging to/from NAS-Cloud

Implementation #2 -- HySDS Hybrid Cloud



Collaboration: Hook Hua, HySDS architect @ NASA JPL

NASA Open Science Initiative



Topics | Missions | Galleries | NASA TV | Follow NASA | Downloads | About | NASA Audiences

Latest | Related

Small but Mighty NASA Weather Instruments Prepare for Launch
14 hours ago

An Aurora Seen From Space
2 days ago

Un estudio de la NASA predice impactos del cambio climático global en los cultivos en 10 años
2 days ago

Global Climate Change Impact on Crops Expected Within 10 Years, NASA Study Finds
2 days ago

NASA, ESA Partnership Releases Platform for Open-source Science in the Cloud
6 days ago

A wispy aurora intersects with Earth's airglow
7 days ago

2021 Antarctic Ozone Hole 13th-Largest, Will Persist into November
7 days ago

Earth

Oct 28, 2021

NASA, ESA Partnership Releases Platform for Open-source Science in the Cloud

Boreal Forest Aboveground Biomass Density (Mg/ha)
0 ≥150

Circumboreal forest biomass density mapped at high spatial resolution (30 m) with NASA's ICESat-2, the joint NASA/USGS Landsat-8, and ESA Copernicus Digital Elevation Model (DEM) data. This provisional product is representative of 2020 conditions and is an open-source science product created on the NASA-ESA MAAP platform.
Credits: MAAP Science Team/NASA Arctic-Boreal Vulnerability Experiment (ABoVE)/NASA Earth Observatory

NASA and ESA (European Space Agency) publicly released a revolutionary open-source science tool for analyzing Earth science data in the cloud – the Multi-Mission Algorithm and Analysis Platform (MAAP). MAAP provides seamless access to NASA and ESA Earth science data and is a model for open-source science collaboration and analysis. As part of MAAP's initial supported area of emphasis on aboveground biomass, it is the host platform for the first globally harmonized assessment of aboveground carbon – information that is vital for managing global climate change.

MAAP is the culmination of a two-year NASA and ESA effort and reflects the cooperation between the two agencies under the NASA and ESA JPPG (Joint Program and Planning Group) Joint Working Group (WG) on Ground Segment and Operations. MAAP is fully operational and will incrementally expand its user community over the coming months. The release of MAAP Version 2 in the Spring of 2022 will add additional data sources, allowing scientists to tackle a broader range of Earth science questions.

The joint NASA-ESA MAAP platform enables collaborative work with data from multiple Earth observation missions.
Credits: ESA

MAAP enables scientists to collaboratively develop algorithms and code as well as analyze and visualize large datasets acquired from sources including satellite instruments, the International Space Station, and airborne and ground campaigns. The large data and high-performance computing required for MAAP, along with a shared code repository and catalog, are stored and managed in the cloud. MAAP capabilities are supported and shared between NASA and ESA.

By following open-source science principles, including full and open access to data, the use of open-source code, and unrestricted access for data users, MAAP removes barriers to participation in scientific investigation, fosters greater diversity and inclusion, and enhances opportunities for discovery.

The initial application of MAAP focuses on measurements of aboveground biomass as part of a global effort to determine the size and carbon content of Earth's forests. These data are critical for informing our understanding of climate change and forecasting its impacts and will be part of regular updates to the Intergovernmental Panel on Climate Change (IPCC). While biomass is the first application of MAAP, it easily can be adapted for collaborative exploration across the breadth of science data and scientific disciplines available through NASA, ESA, and similar research agencies.

MAAP currently includes data from missions such as NASA's Global Ecosystem Dynamics Investigation (GEDI) and the joint NASA/ESA AfriSAR campaign and will eventually support data from upcoming NASA and ESA missions such as ESA's Biomass mission and the joint NASA/Indian Space Research Organisation Synthetic Aperture Radar (NISAR) mission. A forthcoming connection to supercomputing facilities at NASA's Ames Research Center in Silicon Valley, California, will enable efficient processing of extremely large MAAP global datasets, such as those expected from the NISAR mission.

MAAP products can be explored on the MAAP Dashboard at https://earthdata.nasa.gov/maap-biomass or the joint platform entrance at scimaap.net. MAAP also can be accessed through individual NASA (https://maap-project.org) and ESA (https://esa-maap.org/) landing pages.

For media inquiries, contact:

Tylar Greene
Headquarters, Washington
202-358-0030
tylar.j.greene@nasa.gov

Last Updated: Oct 28, 2021
Editor: Tylar Greene

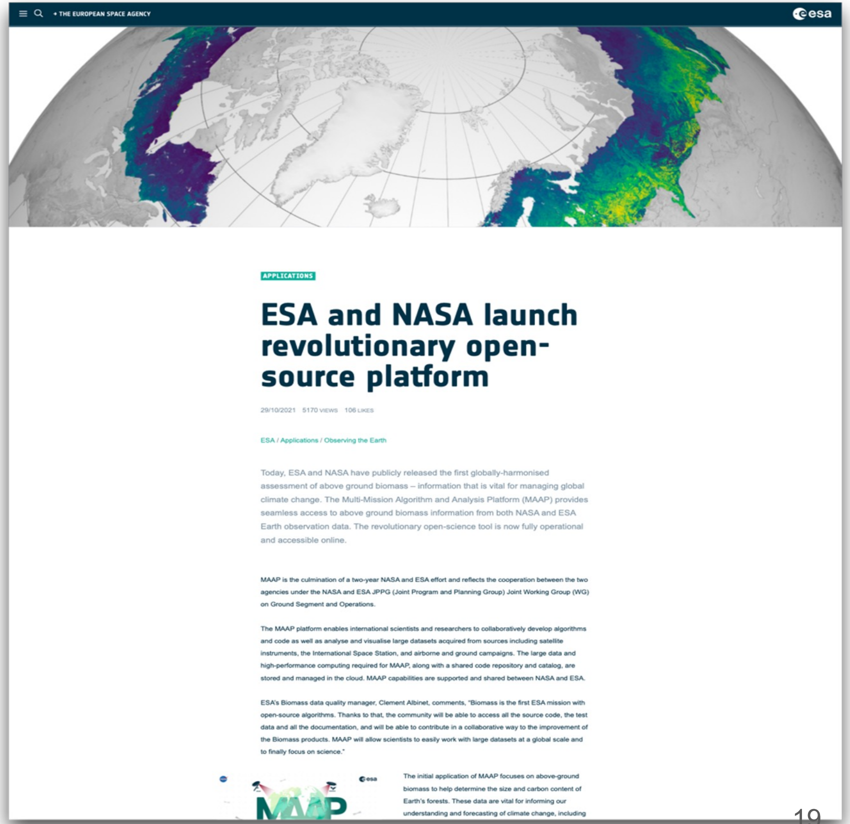
Tags: Earth

Read Next Related Article

National Aeronautics and Space Administration
Page Last Updated: Oct 28, 2021
NASA Official: Brian Dunbar

No Fear Act | FOIA | Privacy | Office of Inspector General | Office of Special Counsel | Agency Financial Reports | Contact NASA

https://www.nasa.gov/feature/nasa-esa-partnership-releases-platform-for-open-source-science-in-the-cloud



THE EUROPEAN SPACE AGENCY

APPLICATIONS

ESA and NASA launch revolutionary open-source platform

29/10/2021 5170 VIEWS 106 LIKES

ESA / Applications / Observing the Earth

Today, ESA and NASA have publicly released the first globally-harmonised assessment of above ground biomass – information that is vital for managing global climate change. The Multi-Mission Algorithm and Analysis Platform (MAAP) provides seamless access to above ground biomass information from both NASA and ESA Earth observation data. The revolutionary open-science tool is now fully operational and accessible online.

MAAP is the culmination of a two-year NASA and ESA effort and reflects the cooperation between the two agencies under the NASA and ESA JPPG (Joint Program and Planning Group) Joint Working Group (WG) on Ground Segment and Operations.

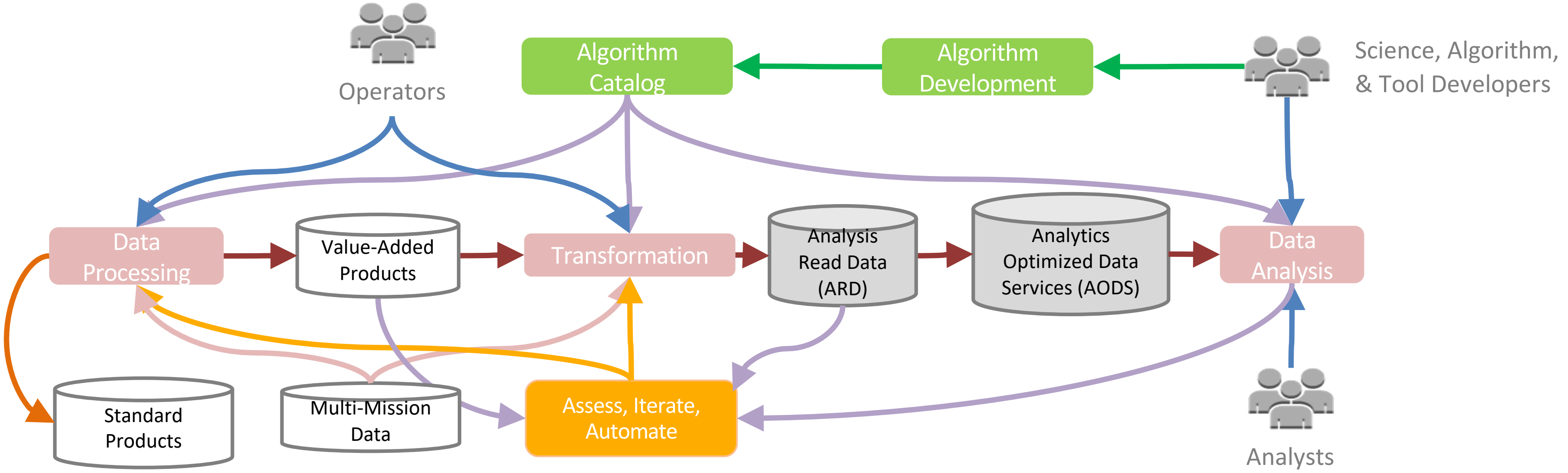
The MAAP platform enables international scientists and researchers to collaboratively develop algorithms and code as well as analyse and visualise large datasets acquired from sources including satellite instruments, the International Space Station, and airborne and ground campaigns. The large data and high-performance computing required for MAAP, along with a shared code repository and catalog, are stored and managed in the cloud. MAAP capabilities are supported and shared between NASA and ESA.

ESA's Biomass data quality manager, Clement Albinet, comments, "Biomass is the first ESA mission with open-source algorithms. Thanks to that, the community will be able to access all the source code, the test data and all the documentation, and will be able to contribute in a collaborative way to the improvement of the Biomass products. MAAP will allow scientists to easily work with large datasets at a global scale and to finally focus on science."

The initial application of MAAP focuses on above-ground biomass to help determine the size and carbon content of Earth's forests. These data are vital for informing our understanding and forecasting of climate change, including

https://www.esa.int/Applications/Observing_the_Earth/ESA_and_NASA_launch_revolutionary_open-source_platform

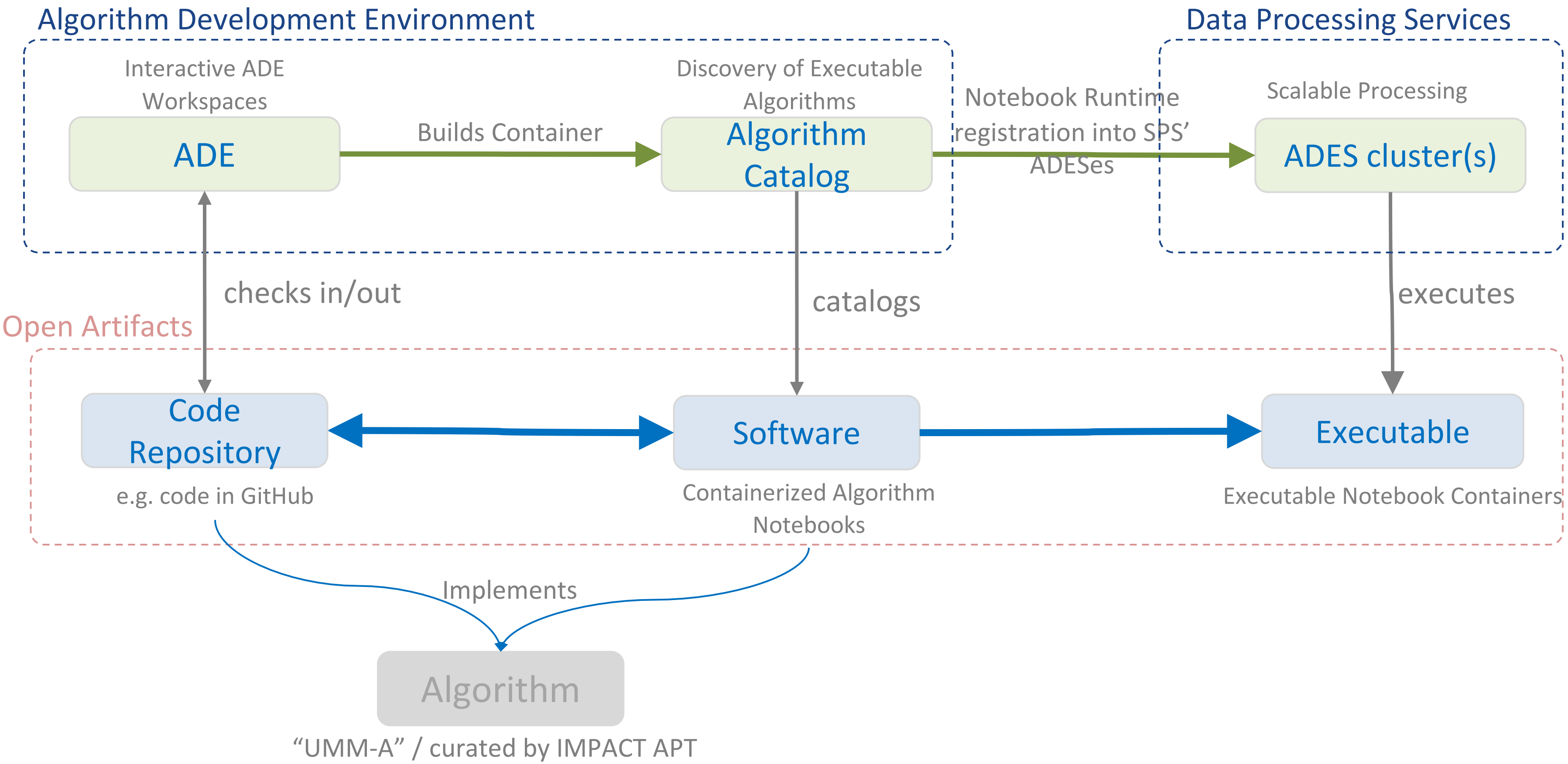
Mission Science Data Systems
Development/Operations Key Concepts



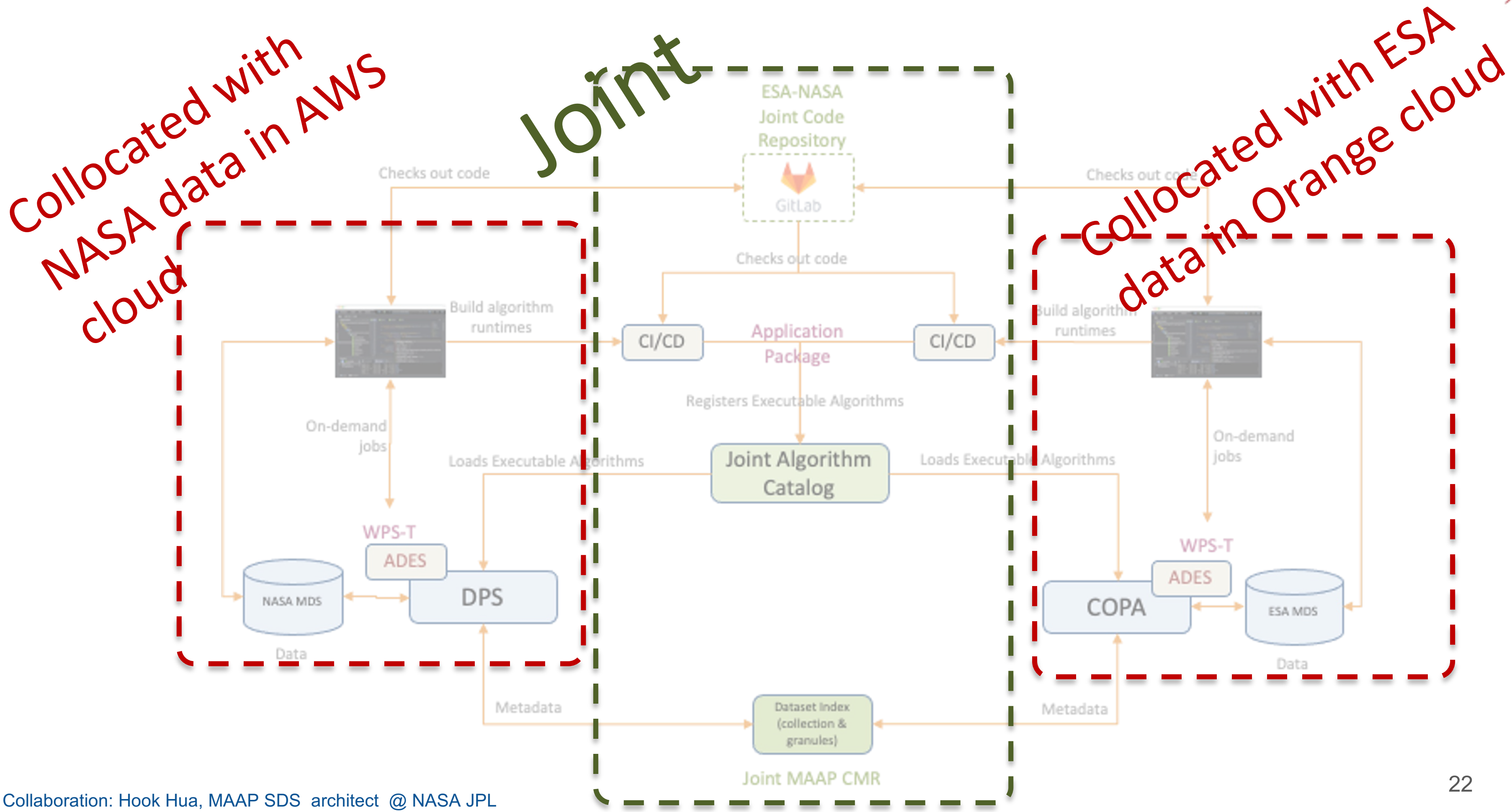
- Algorithm development environment (notebooks)
- Collocated in cloud: algorithm development, science data processing, analysis
- Algorithm testbed –at scale
- **Harmonized same approach for (1) development of algorithms for production, and (2) development of algorithms for cal/val, product validation, on-demand testbeds, (3) R&A analysis**
- ARDs/AODS approach for analysis

Collaboration: Hook Hua, MAAP SDS architect @ NASA JPL

Algorithm Development Environment

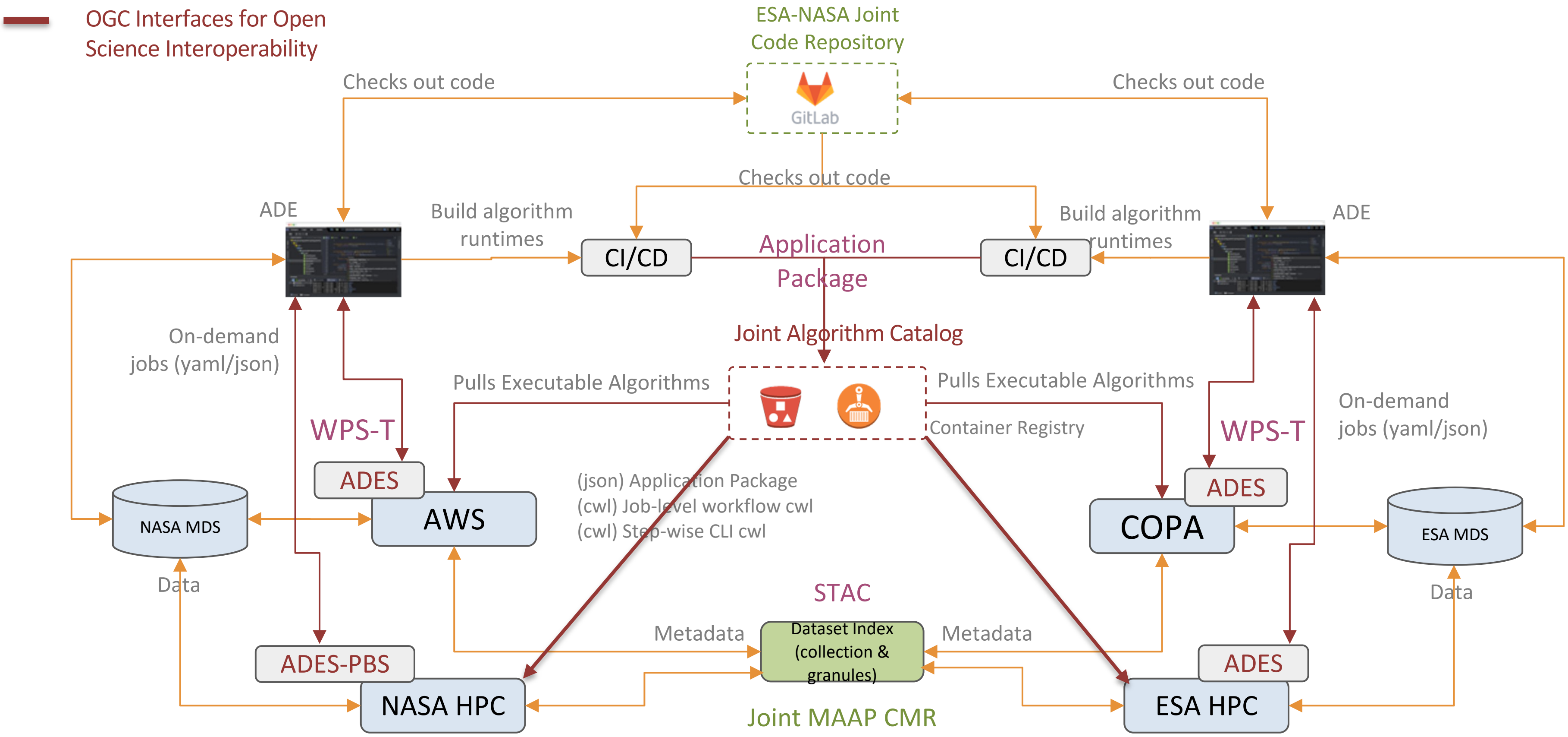


Joint While Being Collocated with Data



Collaboration: Hook Hua, MAAP SDS architect @ NASA JPL

Implementation #3 -- MAAP Hybrid Cloud (proposed)



Collaboration: Hook Hua, MAAP SDS architect @ NASA JPL

* *Interoperable Orchestration (OGC EMS) can be added later*

A&A Approach from MAAP into NAS/HECC

Release 3.0 - Multi User Design (dec 22)

Users interact primarily in AWS MAAP instance

Authenticate the user running in AWS

Bind known user to a specific instance/container

Establish trust between container in AWS and HECC

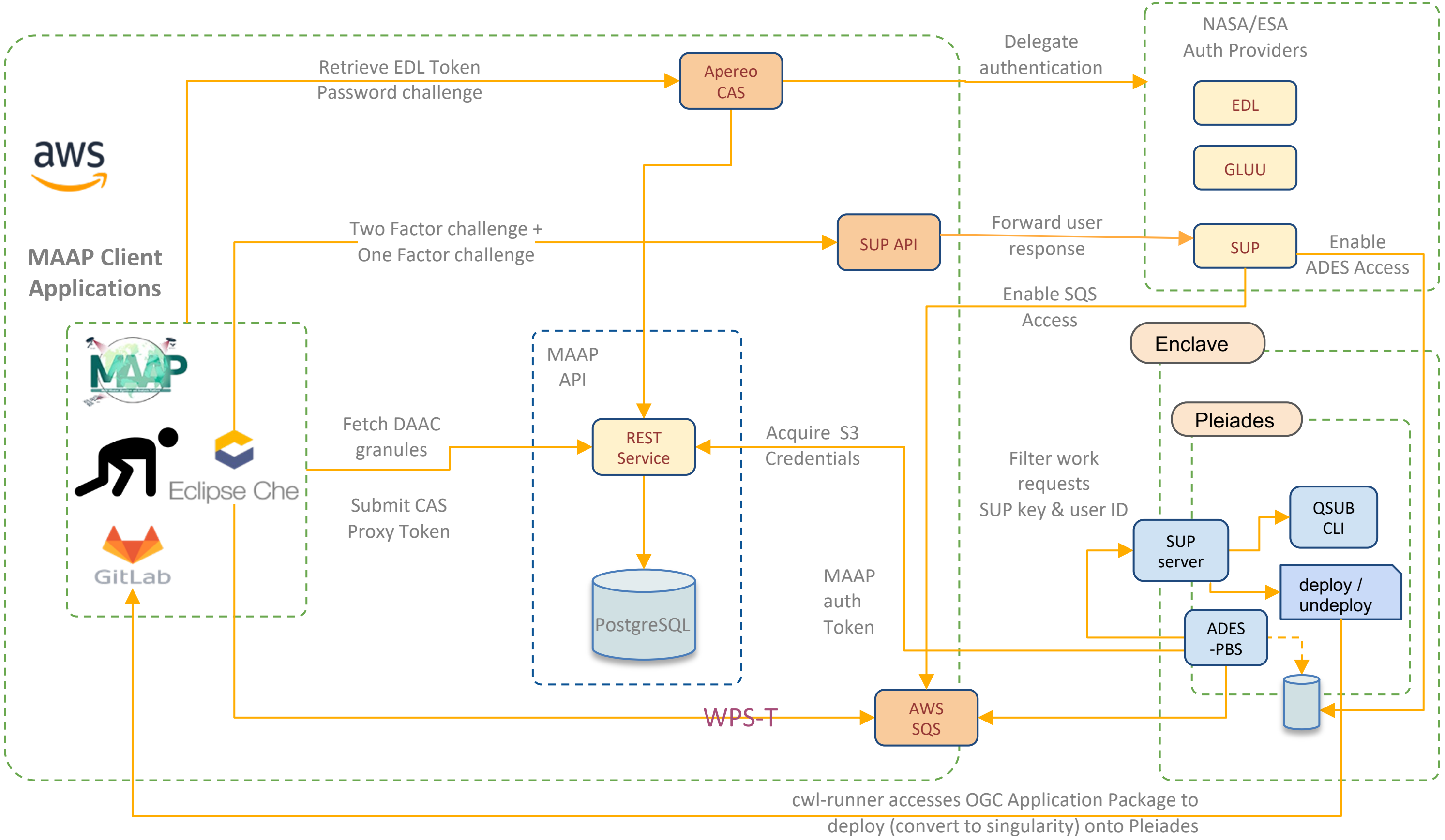
Assurance of authorized users submitting work

System mediates container staging

System provides a data caching mechanism

Authenticates per user work submission requests

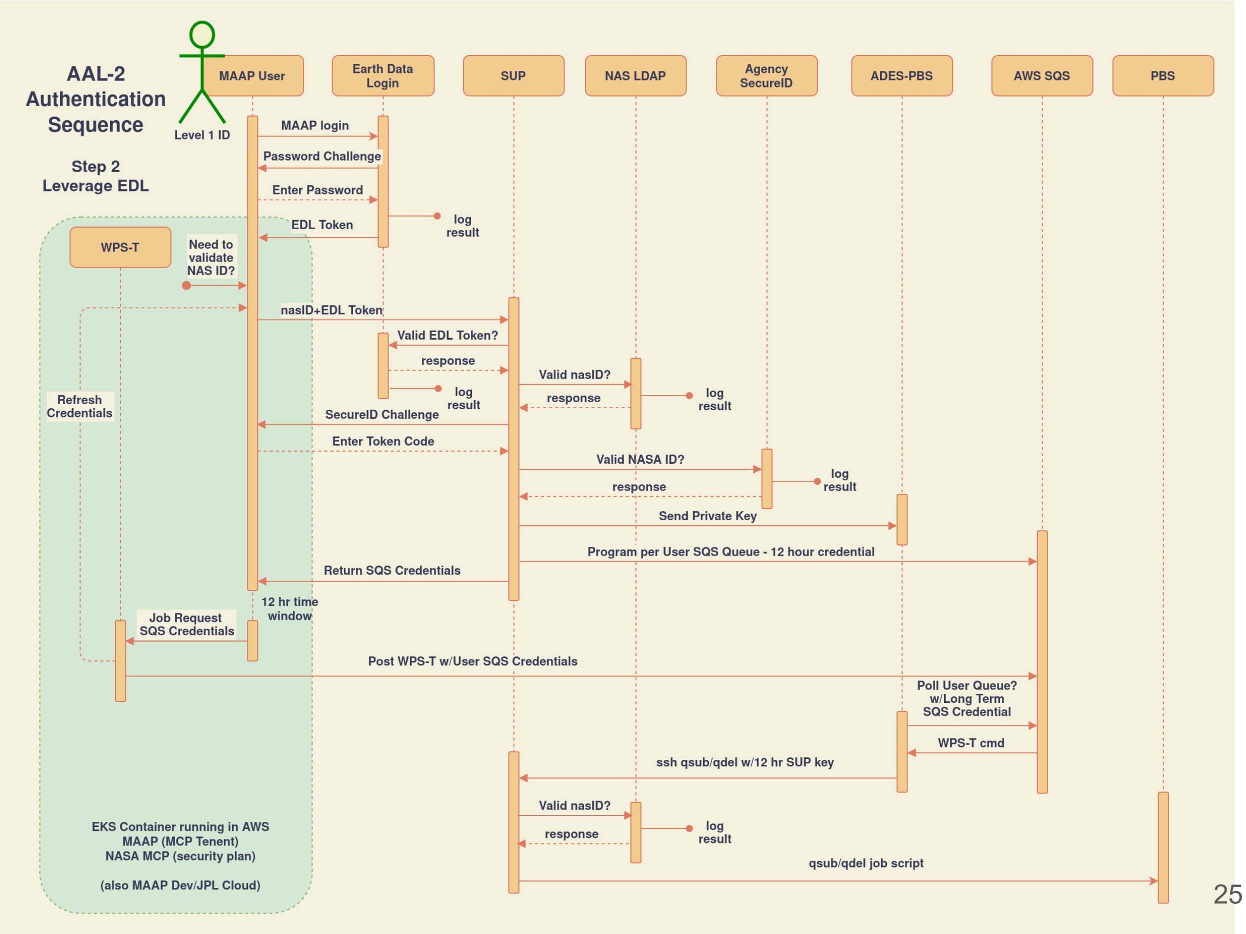
System mediates batch work submission requests



Security Team Expectations

Security Actions:

- Provide an explanation that we can understood this sequence diagram ->
- Establish a connection agreements between system owners
- Insure leaked credentials are invalid
- Leverage AWS Role Based Access Controls



Caveat Emptor

Competition

Congressional ***Investigation of Competition in Digital Markets***
In July 2022 – findings: certain practices of Cloud Service Providers (CSPs) have a chilling effect on competition or are anti-competitive:

https://www.govinfo.gov/content/pkg/CPRT-117HPRT47832/pdf/CPRT-117HPRT47832.pdf

- How does government structure procurements to encourage competition?
- How to create competitive boundaries around specific capabilities
  - E.g. distribution vs. compute

Unfortunately

- This matters when we consider hybrid computing designs
- Risk that the Agency effectively becomes an Agent promoting anti-competitive practices

Policy

NASA Science Policy Directive 41a (aka SPD41a)

https://science.nasa.gov/science-red/s3fs-public/atoms/files/SMD-information-policy-SPD-41a.pdf

Section V:

A. Repositories designated as appropriate for archiving of SMD-funded scientific information shall align, to the extent practicable, to the National Science and Technology Council document entitled: "Desirable Characteristics of Data Repositories for Federally Funded Research"

https://www.whitehouse.gov/wp-content/uploads/2022/05/05-2022-Desirable-Characteristics-of-Data-Repositories.pdf

Section VII:

A. At the end of the period of performance of a research award, scientifically useful data associated with the award that has not already been made public shall be made publicly available to the extent allowed by applicable law and existing NASA policies.

B. Scientifically useful data from models and simulations developed using SMD funding shall be made available at the time of publication of the peer reviewed manuscript that describes the scientific results. Each SMD division shall provide further guidance on the requirements for the sharing of outputs from models and simulations.

C. To achieve reproducibility, scientific software developed using SMD funding and used in support of a scientific, peer-reviewed publication shall be released as open source software no later than the publication date.

Business Cycle Risk

Continuity/Longevity – support for long running missions

- Cost of a service is spread over a large pool of customers. That can shrink and make a particular service business model unsustainable.
- Availability may change quickly if a cloud based capability becomes financially unsustainable
- Industry moves on from using a specific capability
- Government may move more slowly

- Centos
- Ceph
- AWS Amazon Linux 2 -> 2022 -> 2023
- Google Earth Enterprise

Summary

Lack of Data Interoperability abstractions/standards

Locating/Validating against Authoritative Copies

- Staging Data
- Caching Data
- Integrity checks (hashes, parallel hashes)

‘Cloud Optimized’ formats

Versioning

Cloud Data Management Interface (SNIA/CDMI)

- https://www.snia.org/cdmi

Named Data Networking

- https://named-data.net

AWS S3

Azure Blob

Google Cloud Storage

S3 as a ‘defacto standard’ (not really, but)

- MinIO, Ceph, Swift, Scality
- AWS, wasabi, cloudflair r2, backblaze b2

Potential Futures

- Cloud Data Management Interface (CDMI)
  - www.snia.org/cdmi
- Named Data Networking
  - named-data.net
- Open Storage Network
  - www.openstoragenetwork.org
- Open GeoSpacial
  - www.ogc.org
- Common Workflow Language
  - www.commonwl.org

Hybrid Computing Project Outcomes

What are or will we do differently as a result of these projects?

- Development of new authentication and credential management process
  - Changes to our 'secure unattended proxy' mechanism
  - Interface to AWS infrastructure and Security Control features
  - Extend to allow for non-NASA users
- Re-Architecture of our internal security design and controls
  - Development of container management infrastructure
  - What's permitted, bind mounts, external IP NAT
  - Potential to use virtualization, vlan, or other containment measures
  - Alternate UI (docker -> singularity, podman)
- Data Staging – in and out
  - Locate, Reference, Cache, and Validate
  - Credential Management
  - Egress Management
- Changes to the batch scheduling (PBS)
  - Efficiently splitting up nodes
  - Rapid scheduling decisions
  - Supporting interactive work
  - Staging remote data in and out (e.g. satellite sources)
  - Caching remote data
- Reassessment of how we connect to Amazon
  - Impacts physical network design
  - Impacts to peering agreements
  - Impacts to cost tracking (egress)
- Performance and Logging
  - How do you track this work flow
  - How do you identify performance issues
  - Can you identify suspicious activity

Challenges:

- Security
- Identity and Credential Management
  - Differences in desired level of verification
- Security Plan Connection Agreements
- Staging Data
- Caching Data
- Publishing Data
- Data Integrity
- Computational Integrity
- User Acceptance

Generalized capability that can be applied to numerous other use cases

Acknowledgements

Thank you! Any questions?

NASA ARC Team	NASA MSFC Team	NASA JPL Team
Bob Ciotti Jeff Becker Greg Matthews Paul Kolano Steve Heistand Andy Michaelis Jon Jenkins Matt Lesko Chris Ishisoko Ted Bohrer	Kaylin Bugbee Alex Mandel Jamison French Chuck Daniels	Joe Jacob Gerald Manipon Marjorie Lucas Mohammed Karim Max Zhan David Moroni Hook Hua George Chang Brian Satorious

http://www.nas.nasa.gov/hecc