National Aeronautics and Space Administration

NCCS User Forum

02/05/2020

Laura Carriere, Dan Duffy, Mark Carroll, Ellen Salmon, Bruce Pfaff

NCCS
NASA CENTER FOR CLIMATE SIMULATION
HIGH-PERFORMANCE SCIENCE

Agenda

- **Introductions**
- **Trends**
- **Upcoming Plans**
  - Discover SLES 12, Slurm 19, GPFS upgrade
  - Compute Upgrade
  - Consolidated Cloud – Explore
  - GPU Cluster
- **Data Science Group**
- **Storage**
  - Centralized Storage, Data Management Plans
  - MSS going read-only, AWS Deep Archive

Introductions

- Sujay Srivastava
- Jim Dolan
- John Villa
- Kenny Peck
- Greg Factor
- Perry Asare
- Francis Chumbow
- Matt Stroud
- Bob Budden
- Sean Keefe
- Mary Aronne
- Ryan Forbes
- Vincent Wild
- Sam Ramos
- Jordan Caraballo-Vega

Emerging Trends

- **Convergence of HPC, AI, and Big Data**
  - Diversity of compute platforms has increased dramatically in the last few years
  - Along with the hardware, the software stacks are also becoming more diverse
  - Large number of cores, smaller memory per core, requires increase in concurrency
  - Introduction of new operations: half-precision, bfloat, tensor
- **Purpose Build Computing**
  - Design systems for specific applications; cost effective?
- **Accelerated computing**
  - GPU (Nvidia, Intel, AMD), Tensor Processor Units (TPUs), FPGA (reconfigurable), Neuromorphic, Quantum
- **Diverse hardware**
  - Intel, AMD, IBM, Arm
- **Domain specific languages**
  - Kokkos, GridTools, Intel OneAPI
- **Cloud computing**
  - Amazon Web Services, Microsoft Azure, Google

Accelerator Based HPC Systems

Summit: Oak Ridge National Laboratory (ORNL) - Operational

- Ranked as the fastest computer in the world: Top500
- Ranked as the 3rd most energy efficient: Green500
- IBM, Nvidia, Mellanox
- $325M

El Capitan: Lawrence Livermore National Laboratory (LLNL) – 2022

- Cray Shasta, Slingshot Interconnect
- 1.5 Exaflops
- $600M
- CPUs and GPUs (To be determined)

Aurora: Argonne National Laboratory (ANL) – late 201

- Cray Shasta, Slingshot Interconnect
- 1.0+ Exaflops
- $500M+
- Intel Xeon General Purpose, Intel Optimized Accelerators: X^e architecture (Intel GPU), Intel Open API Software

Frontier: Oak Ridge National Laboratory (ORNL) – 2021

- 1.5 Exaflops
- AMD Epyc and AMD Radeon GPU

This is just the DOE Exascale program. The Europeans are funding exascale projects that will develop a new processor, and the Chinese are doing the same.

Facilities Constraints

- **Staying with the traditional cluster computing route, the NCCS will have to double its power and cooling capabilities**
- **Current analysis and planning with Goddard Facilities Management Division (FMD)**
  - Expect that the cost for this will be in the multiple millions
  - Will have estimates for this later this fiscal year
- **What if the NCCS deployed accelerator based computing?**
  - To equal the peak computing capacity of the NCCS would take about 10 to 12 racks of GPU systems and ¼ the total power
  - Representative GPU node:
    - » Dual-socket Intel Xeon (20 core/socket, 1.6 TF/socket)
    - » Quad Nvidia V100s (7 TF/GPU)
    - » Peak computing 31.2 TF
    - » 20 Nodes/Rack – 624 TF/Rack

What does that mean for all of us?

- **NCCS will be integrating more and more accelerator based systems**
  - Starting with Nvidia GPU based systems this month and more later this year
  - Significant possibility that we will be deploying all GPU based computing within 5 years
- **Need software engineering investments in applications to take advantage of these compute capabilities**
  - How do we fund this? Combination of Scientific Computing and project funding
  - Need to start now; it could take 3 to 5 years to modify the applications
- **Benchmarking public cloud computing HPC capabilities to satisfy traditional cluster computing capabilities**
  - Need to understand the performance, costs, business case, and workflows
  - Also need to build more cloud-native HPC applications in order to be efficient and cost effective

We have to partner with our users to continue to be successful moving into this future!

Upcoming Plans

- **Discover SLES 12, Slurm 19, GPFS upgrade**
- **SCU16**
- **Consolidated Cloud – Explore**
- **GPU Cluster**
- **Consolidated Downtime**
  - Discover: 2/19/20 8:00 am to 2/20/20 8:00 am (24 hrs)
  - MSS, CSS, & Dataportal: 2/19/20 10:00am to 8:00pm (10 hrs)

Discover

- **Supercomputer environment**
  - 129,000 cores (CPUs)
  - 6.8 Pflops of performance
  - 45-50 PB of storage
  - Runs Linux (SLES 11 & 12), uses the Slurm job scheduler
  - 2 IB Fabrics, 1 OPA Fabric
- **Intended for workloads that can be parallelized and require communication between cores (e.g. message passing)**
- **Access requires an approved allocation request through the High-End Computing Capability (HECC) Project**

Discover Upgrades

- **Need to upgrade from SLES 11 SP3 – OS at end of life**
- **SLES 12 SP3 is installed on 400 IB nodes plus all OPA nodes (SCU14/15) - Please test and convert ASAP.**
- **Reducing the versions of supported software (compilers, libraries) to simplify Discover config & usage**
  - E.g. Intel compiler v19+ only, limiting # of versions available
- **Earlier compilers are not recommended on SLES 12**
- **List of supporting software – module avail, spider**
- **Lmod - new features, forces you to load the correct dependencies (compiler + mkl, etc.)**

Slurm 19

- ALL SLES 12 nodes being managed via Slurm 19
- As nodes move over to SLES 12, they become part of the Slurm 19 realm.
- New features:
  - Ability to stack jobs with small core counts on one node (one job owner only per node) - Contact USG for more info.
- Jobs will default to requesting Haswell nodes
- Contact the NCCS if you wish to test on Skylake nodes
- Skylake nodes ONLY exist in the OPA fabric, so it may require additional work to get your jobs to run there.

GPFS Upgrade

- **Need to upgrade to version 5.x**
  - Version 4.x ends support this fall
  - Will require building new filesystems and moving data
- **Need to have empty storage available for the move**
- **Working with the vendor to develop a migration plan**
- **New storage purchases will support 5.x migration**
- **Oldest storage systems will be replaced this year with larger capacity systems**

Current Private Cloud Environments

- **ADAPT – Advanced Data Analytics PlaTform:**
  - Virtual Linux environment in OpenStack, with Windows option
  - Per VM Maximum: 20 cores, 240 GB
  - Access to NASA data sets (CSS)
  - Intended for science workloads that are intrinsically parallel or benefit from collaboration in a customizable and extensible environment
- **GPC – Goddard Private Cloud:**
  - Virtual Linux environment in OpenStack
  - Per VM Maximum: 16 cores, 55 GB
  - Intended for engineering workloads

Explore – A Consolidated Cloud

- **ADAPT + GPC, best of both worlds**
- **OpenStack**
- **Self Provisioning**
- **New hardware**
- **Access to CSS data**
- **Migration to Explore:**
  - GPC users starting July 1, 2020
  - ADAPT users starting in January 1, 2021

GPU Cluster

- **Currently have 3 GPU nodes:**
  - Two have 4 NVIDIA V100s, one has 8 NVIDIA V100s
- **Cluster with 22 nodes arrives next week**
  - NVIDIA V100s with 32 GB of RAM
  - Systems have 20 cpu cores and 768 GB of RAM
  - SSD storage for fast local data access
- **Access through Slurm to start**
- **Will move to OpenStack later in the year**
- **Available May 1, 2020**

Science Managed Cloud Environment

- **The SMCE is** a managed Amazon Web Service (AWS) based infrastructure for NASA funded projects that can leverage cloud computing capabilities.
- Provide cloud access to NASA PIs with non-NASA team members
- Perform research using new computing capabilities without extensive start-up time
- Ability to try new tools and methods from AWS's product catalogue easily and affordably
- Easily scalable computing for high-demand, high-bandwidth needs

CISTO Data Science Group

- **Mission**
  - The Data Science Group (DSG) within the Computational and Information Science and Technology Office (CISTO) offers a unified approach to advance science through the application of advanced analytics including artificial intelligence and machine learning in a high-performance computing environment. We will accomplish this across disciplines at GSFC by exploiting the python ecosystem of tools, exploring the utility and cost effectiveness of commercial cloud resources, and technologies for integrating these tools to discover crosscutting results.

DSG – Current Activities

- Support science investigators by co-developing software optimized to run in ADAPT, on premises cloud and in commercial cloud
- Develop AI Center of Excellence website as a central location for hosting and collecting information about AI/ML at Goddard
- Lead Science Task Group for ML at GSFC *Oct. 2019 – Sep. 2020*
- Co-host AI workshop with JPL to be held *Mar. 24 – 26, 2020* at JPL
- 1st NASA Hackathon in support of Heliophysics *Jun. 8 – 12, 2020*
- 2nd NASA Hackathon in support of Climate science *Date TBD*
- Support data science academy at GSFC: Linked In/Lynda learning
- ASTG Python Courses

Storage Agenda

- Centralized Storage
- Data Management Plans
- MSS going read-only
- AWS Deep Archive

Storage

- Improve NCCS users' ability to access NASA data from all NCCS systems to provide new science opportunities (Machine Learning/Deep Learning, Big Data Analytics)
- Replace MSS with additional online storage, both Centralized Storage (CSS) and local to Discover and ADAPT
- Provide access to archive options for long term storage of data products
- Utilize Data Management Plans to better understand storage and archiving requirements

Rationale for Changes

- **Address issues with the current situation:**
  - Data on MSS is inaccessible for ML/DL/AI applications
  - Data on Discover is frequently duplicated because users don't know where it is
  - NCCS costs are higher due to duplication, science opportunities are lower
  - Limited data sharing between NCCS environments (CSS)

- **One science project's output becomes another project's input**

Centralized Storage (CSS)

- CSS has 17 PB of storage, 17 PB more arrives in March
- Provide storage of, and compute on, big NASA curated data sets to our HPC, Cloud, GPU, and Dataportal environments
- Provide data discovery and usage reporting to reduce data duplication and facilitate data deletion
- Manage the data lifecycle through Data Management Plans and policies

- Find data to move from Discover or MSS to CSS

Centralized Storage Data

- ABoVE
- CMIP5
- CREATE
- GeoMIP
- IceBridge
- ICESat-2
- Landsat
- LISdata
- MERRA-2
- GDDP
- NGA
- Obs4MIPs

Coming soon:

- CETB
- GMAO
- GOES
- MODIS
- Planet

Data Management Plans

- **Improving our storage resource utilization through Data Management Plans (DMPs)**
  - Input, intermediate, final data sets
  - Software
  - Ingest, access, sharing, disposition
- **Allows for planning for enough online storage for intermediate data**

Data Management Plan Concepts

- **Four types of data:**
  - Input – local or remote (to be brought into the NCCS)
  - Intermediate – data created during software runs:
    - Not permanent
    - Not to be shared publicly
    - Could be restart files, research results, temporary files
  - Final – used for publications, shared with the science community or collaborators, could be input to other science programs
  - Software – save in a Git repository for re-use

Data Management Plan Contents

- **Governs input, intermediate, and final data products**
- **Identifies workflow, diagrams encouraged:**
  - Ingest
    - Source, destination, volume
  - Access
    - Public, private, proprietary, business sensitive
    - Required systems
  - Sharing
    - Group access, data services
  - Disposition
    - Centralized storage, archiving, deletion

Example DMP – Model Production Runs

- **Input data is stored on CSS for sharing**
- **Develop charts showing saving for intermediate files:**
  - < 1 month – 100% - on Tier 1 storage (Discover)
  - 1-6 months – 25% - on Tier 2 storage (Discover)
  - > 6 months – 8% - Offsite Archive
- **Store final data products at a NASA archive and on CSS while in use**
- **Deletion of data:**
  - Input is determined by project life
  - Intermediate is determined by above schedule
  - Final (CSS) is determined by use of product locally

MSS – Read Only Mode

- **Fall 2020**
- **To get there:**
  - Need intermediate and archive storage requirements (DMPs)
  - Order online storage on Discover and ADAPT
  - AWS Deep Archive instructions and cost model
  - Changes to workflows to write to disk and not MSS
  - Integrate new online storage
  - NASA archive solutions for climate model output

AWS Deep Archive

- If data isn’t going to a DAAC, AWS Deep Archive may be an option
- Cost -> $0.99/TB/month
- Upload costs are minimal
- Download costs are $0.09/GB for the first 10TB per month and costs/GB drop with increased downloads
- Data must first be “restored” before downloads, a process that takes 12 hours
    - *This should be a relatively rare occurrence, else Deep Archive is not a good fit*

AWS Deep Archive Access

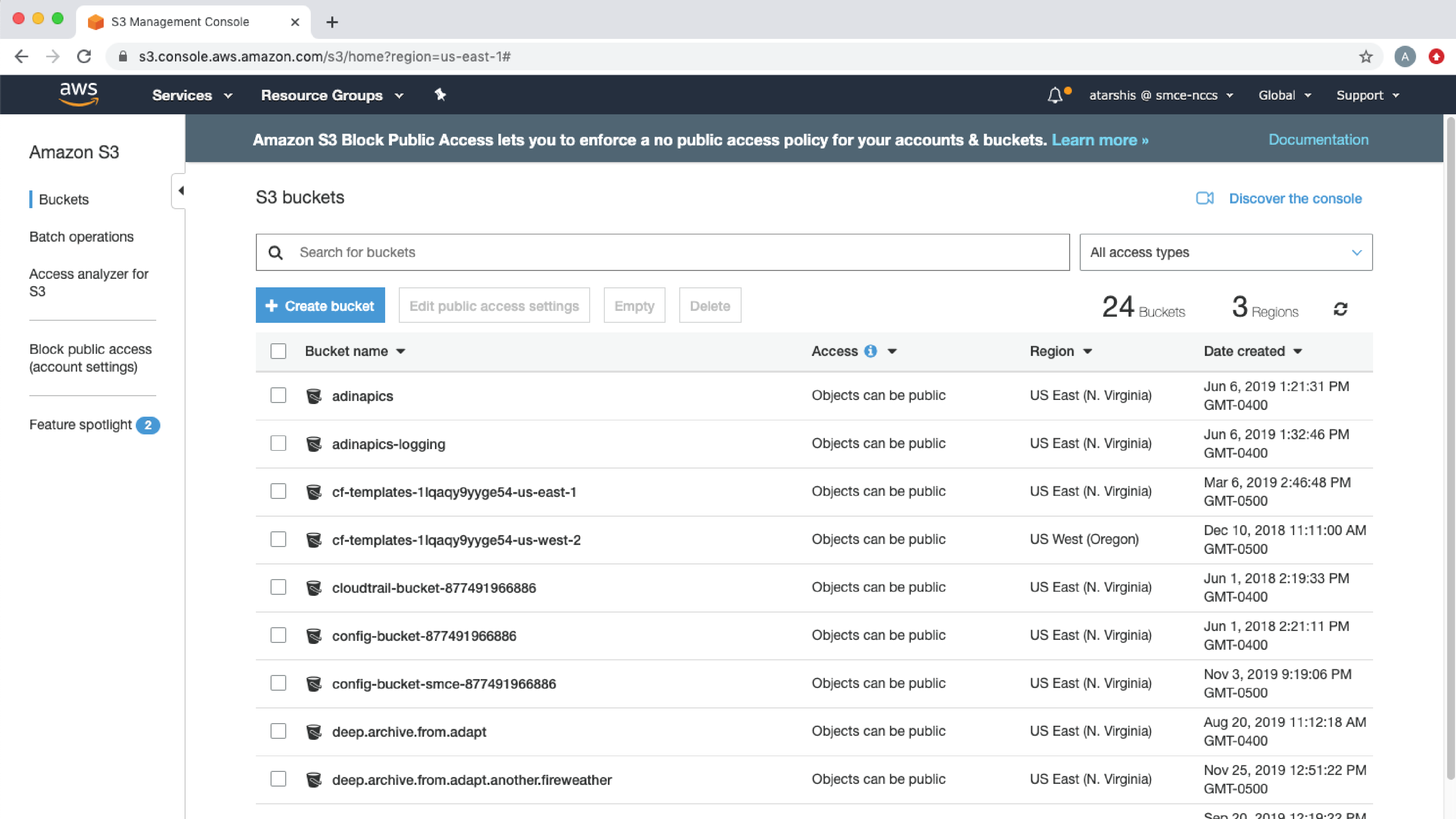
- **Console access and CLI access**
  - Console is easiest way to get started, but CLI allows for much more efficient operations on large numbers of files
- **Buckets (similar to a directory) are set up with appropriate permissions**
- **If data is currently located in the mass storage system the NCCS may upload the data on the user's behalf**
  - It will need to be done in tape order and controlled by admins due to the limited front-end disk cache available on Dirac

AWS: Example CLI Commands

- **Create a bucket:**
  - aws s3 mb s3://<new-bucket-name>
- **Copy a file to that bucket:**
  - aws s3 cp <filename> s3://<new-bucket-name> --storage-class DEEP_ARCHIVE
- **List buckets I can access:**
  - aws s3 ls
- **Start the 12-hour restoration of a deep archive file, then make it retrievable for 2 days**
  - aws s3api restore-object --bucket <bucket> --key <file> --restore-request Days=2

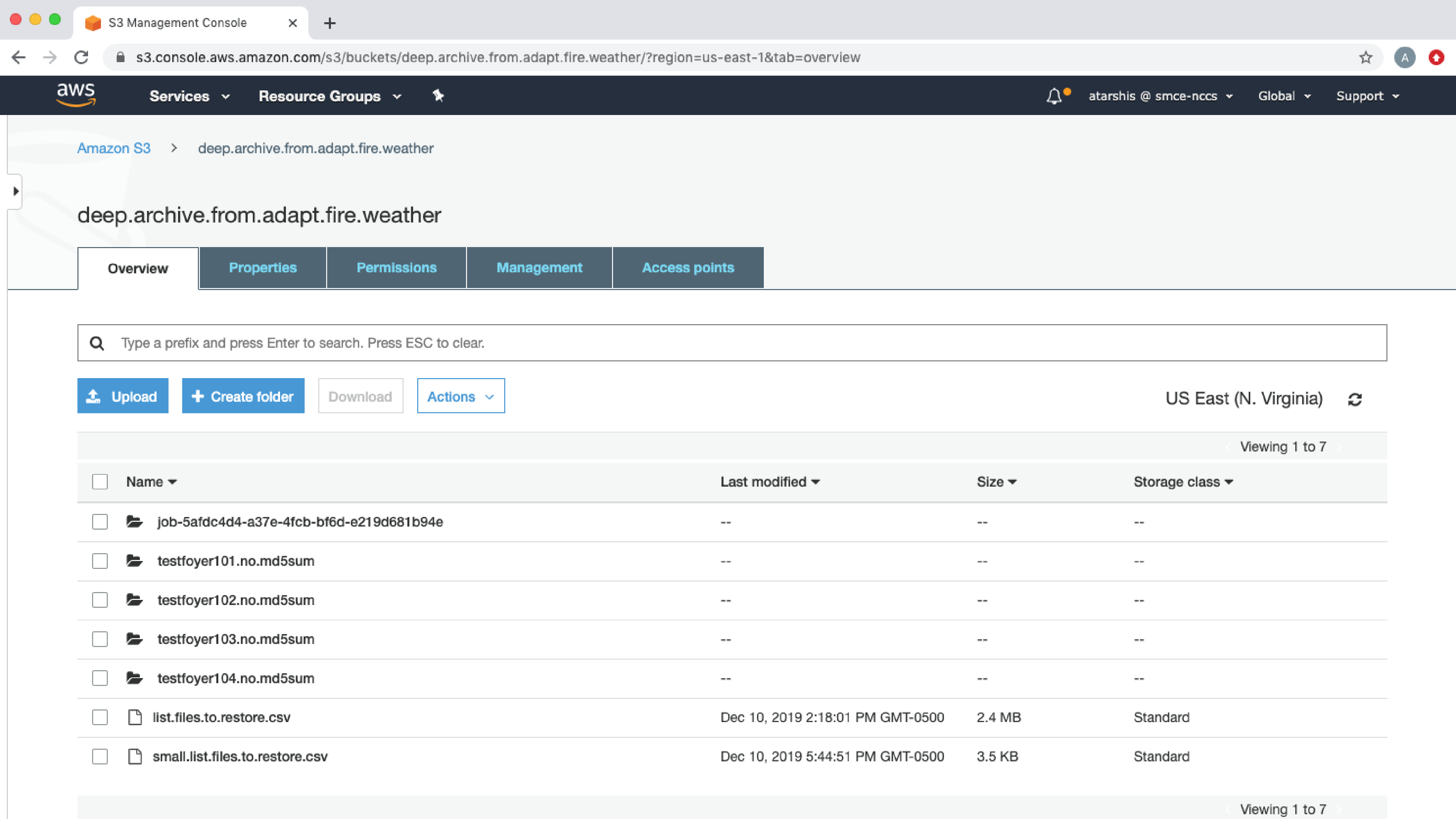
AWS: Getting an Account

- **Set up a meeting with the NCCS**
  - Create a Data Management Plan
  - Amazon script estimates upload and data storage costs
- **SMCE admins will grant access and restrict bucket permissions appropriately**
- **Multi-factor authentication is required and will be set up for you**
  - Requires the Google Authenticator app on your smart phone and a script to generate special environment variables for a Linux shell



S3 buckets

□	Bucket name	Access	Region	Date created
□	adinapics	Objects can be public	US East (N. Virginia)	Jun 6, 2019 1:21:31 PM GMT-0400
□	adinapics-logging	Objects can be public	US East (N. Virginia)	Jun 6, 2019 1:32:46 PM GMT-0400
□	cf-templates-1lqaqy9yyge54-us-east-1	Objects can be public	US East (N. Virginia)	Mar 6, 2019 2:46:48 PM GMT-0500
□	cf-templates-1lqaqy9yyge54-us-west-2	Objects can be public	US West (Oregon)	Dec 10, 2018 11:11:00 AM GMT-0500
□	cloudtrail-bucket-877491966886	Objects can be public	US East (N. Virginia)	Jun 1, 2018 2:19:33 PM GMT-0400
□	config-bucket-877491966886	Objects can be public	US East (N. Virginia)	Jun 1, 2018 2:21:11 PM GMT-0400
□	config-bucket-smce-877491966886	Objects can be public	US East (N. Virginia)	Nov 3, 2019 9:19:06 PM GMT-0500
□	deep.archive.from.adapt	Objects can be public	US East (N. Virginia)	Aug 20, 2019 11:12:18 AM GMT-0400
□	deep.archive.from.adapt.another.fireweather	Objects can be public	US East (N. Virginia)	Nov 25, 2019 12:51:22 PM GMT-0500
				Sep 20, 2019 12:19:22 PM



s3.console.aws.amazon.com/s3/buckets/deep.archive.from.adapt.fire.weather/?region=us-east-1&tab=overview

Amazon S3 > deep.archive.from.adapt.fire.weather

deep.archive.from.adapt.fire.weather

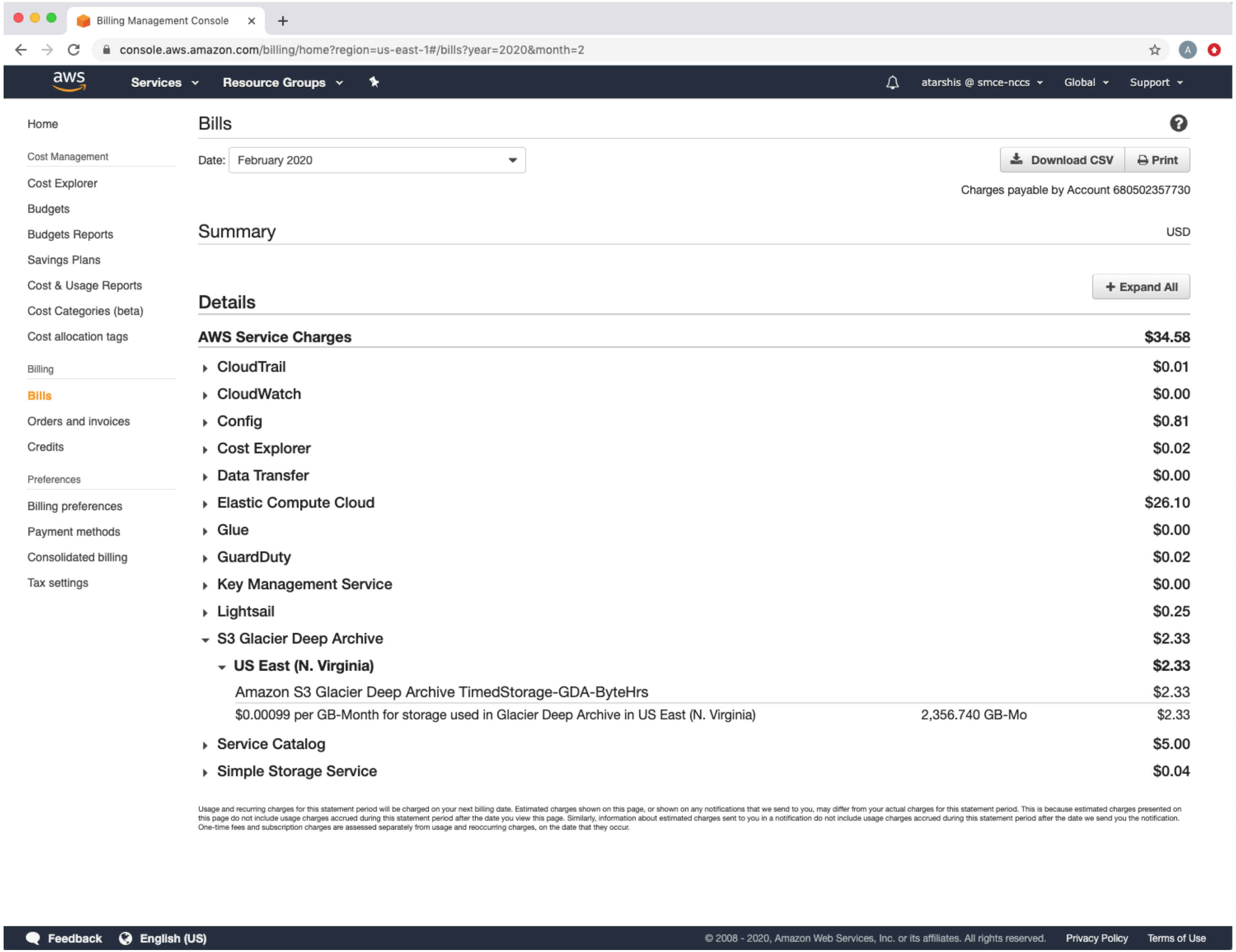
Type a prefix and press Enter to search. Press ESC to clear.

US East (N. Virginia)

Viewing 1 to 7

Name	Last modified	Size	Storage class
job-5afdc4d4-a37e-4fcb-bf6d-e219d681b94e	--	--	--
testfoyer101.no.md5sum	--	--	--
testfoyer102.no.md5sum	--	--	--
testfoyer103.no.md5sum	--	--	--
testfoyer104.no.md5sum	--	--	--
list.files.to.restore.csv	Dec 10, 2019 2:18:01 PM GMT-0500	2.4 MB	Standard
small.list.files.to.restore.csv	Dec 10, 2019 5:44:51 PM GMT-0500	3.5 KB	Standard

Viewing 1 to 7



Bills

Date: February 2020

Charges payable by Account 680502357730

Summary

USD

Details

AWS Service Charges		**$34.58**
CloudTrail		$0.01
CloudWatch		$0.00
Config		$0.81
Cost Explorer		$0.02
Data Transfer		$0.00
Elastic Compute Cloud		$26.10
Glue		$0.00
GuardDuty		$0.02
Key Management Service		$0.00
Lightsail		$0.25
S3 Glacier Deep Archive		$2.33
US East (N. Virginia)		**$2.33**
Amazon S3 Glacier Deep Archive TimedStorage-GDA-ByteHrs		$2.33
$0.00099 per GB-Month for storage used in Glacier Deep Archive in US East (N. Virginia)	2,356.740 GB-Mo	$2.33
Service Catalog		$5.00
Simple Storage Service		$0.04

Usage and recurring charges for this statement period will be charged on your next billing date. Estimated charges shown on this page, or shown on any notifications that we send to you, may differ from your actual charges for this statement period. This is because estimated charges presented on this page do not include usage charges accrued during this statement period after the date you view this page. Similarly, information about estimated charges sent to you in a notification do not include usage charges accrued during this statement period after the date we send you the notification. One-time fees and subscription charges are assessed separately from usage and reoccurring charges, on the date that they occur.

NASA Archives

- **NASA doesn’t currently provide a climate model output archive although there are requirements to archive results, particularly if associated with a DOI or a grant**
- **NASA archives provide data discoverability through the Common Metadata Repository (CMR)**
    - Note, the NCCS can help with CMR data entry even if data is stored at the NCCS
- **GES DISC is an option**

Questions?

Funding

- **SMD Strategy for Data Management and Computing for Groundbreaking Science 2019-2024**
  - Recommends requiring data derived from NASA-funded research to go to a NASA archive for long-term curation and public availability. Notes that this will require funding