GETTING READY FOR (NI)SAR IN THE CLOUD

Contributors:

F.J. Meyer[1)2)], S. Owen[3)], C. Stoner[2)], H. Hua[3)], S. Arko[2)]

[1)]Geophysical Institute, University of Alaska Fairbanks

[2)]Alaska Satellite Facility (ASF), University of Alaska Fairbanks

[3)]Jet Propulsion Laboratory, Pasadena, California

UAF Alaska Satellite Facility
Making remote-sensing data accessible since 1991

JPL Jet Propulsion Laboratory

6/2017 HGF Alliance Meeting

Table of Contents

The Get Ready For NISAR (GRFN) Project

THE CLOUD AND ITS RELEVANCE FOR LARGE VOLUME SAR MISSIONS

A Few Words on the Cloud

Public versus Private (On-Premises) Cloud

- **Public Cloud [Amazon, Google, Microsoft, …]**
  - Cloud vendors typically have “infinite” resources available
  - Virtual machines handle processing → **spinning-up and terminating of VMs provides full performance while only paying for what you use**
  - Web Object store for data storage & distribution → **full capability but pay-as-you-go**

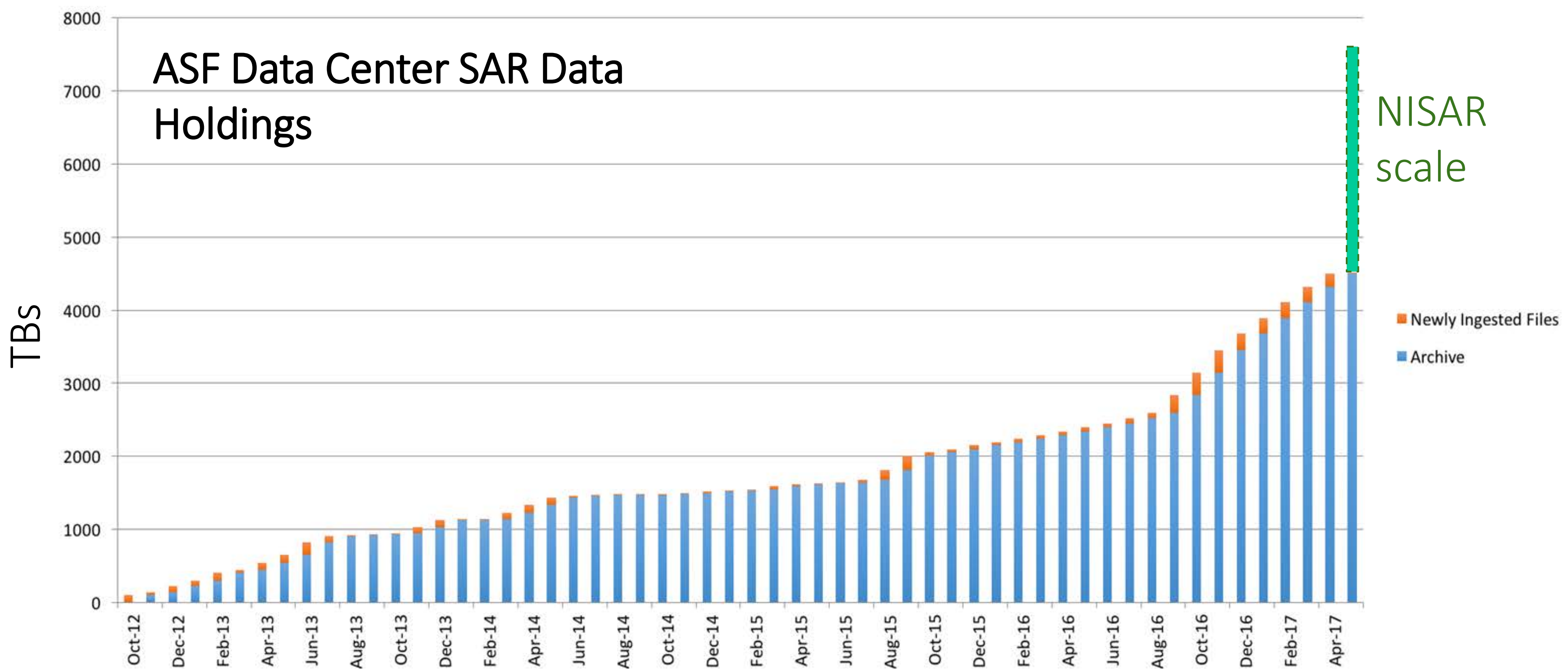
- **Private Cloud [your Typical Data Center]**
  - Can also be virtual machines for processing, but pay for machines upfront as a sunk cost
  - Build out of data distribution capabilities takes time and must be paid for upfront

- **Potential Benefits of the Cloud for Large Volume Remote Sensing Systems**
  - Cloud allows you to scale up as you need instead of a big up front sunk cost
  - Cloud also allows researchers to bring their processing to the archive - no more waiting to download 100s or 1000s of scenes first!

(NI)SAR And The Cloud

The Relevance of Cloud Processing for Modern SAR Missions

- **Modern SAR Sensors such as Sentinel-1 and NISAR will produce massive amounts of data at high data rates**
  - **Sentinel-1A & B**: ~5GB per frame (SLC) & 1PB/year data volume
  - **NISAR**: ~25+GB per frame (SLC) & 45PB/year data volume



(NI)SAR And The Cloud

The Relevance of Cloud Processing for Modern SAR Missions

- **Modern SAR Sensors such as Sentinel-1 and NISAR will produce massive amounts of data at high data rates**
  - **Sentinel-1A & B**: ~5GB per frame (SLC) & 1PB/year data volume
  - **NISAR**: ~25+GB per frame (SLC) & 45PB/year data volume

- **Traditional Architecture Won't Scale**
  - **For Data Centers**, too expensive (in cost and time) to scale up to 100s of PB, mostly due to data movement (i.e. processing to storage, user downloads, etc.)
  - **For Researchers**, a typical two year deep stack will be hundreds of TBs in size

- **The Big Data Challenges of Large Volume SAR Missions**
  - How to cost effectively ingest and store large volumes?
  - How to scale up and serve large volumes?

Why Should You Care About The Cloud

Relevance of Cloud Concepts for Mission Operators

- **Voluminous SAR Data Becoming a Forcing Function**:
  - Becoming too large to process SAR in traditional ways e.g. download L0/L1 to process to L2 interferograms and L3 time series

- **Advantages for Processing & Storage**
  - Bring processing to storage
    - → Fewer data movements
    - → More efficient data handling and processing

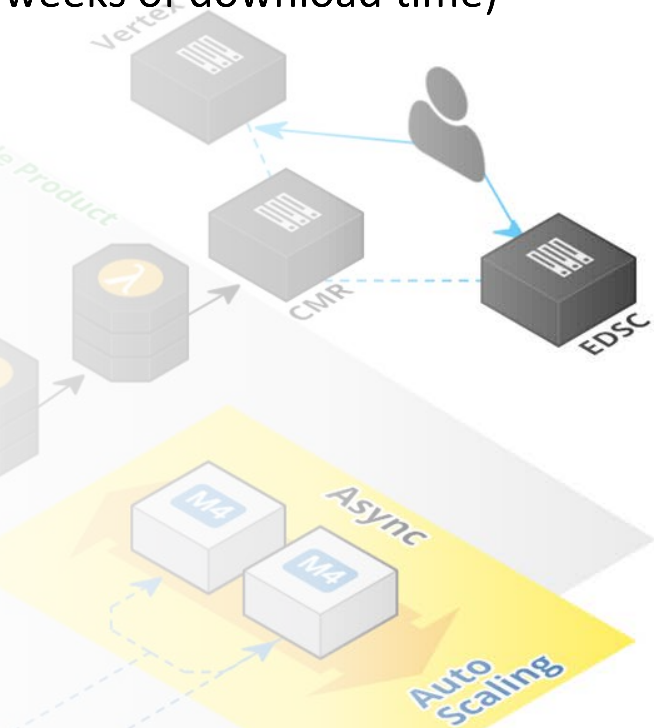
- **Advantages for Serving out Data**
  - Move user processing to storage
  - Spend your time processing, instead of downloading hundreds or thousands of large volume data
  - Examples:
    - estimated size of NISAR SLC is 25GB → 5 hours download time per file on 10mbps line → one-year stack (30 SLCs) requires >150 hours of download time

Why Should You Care About The Cloud

Relevance of Cloud Concepts for Science and Applications Communities

- **Infrastructure and Logistics Savings**
  - Avoid downloading large data volumes (e.g. days to weeks of download time)
  - No local storage infrastructure needs
  - No expensive compute infrastructure needs

- **Massively Parallel Processing Capabilities**
  - Opportunity to harness tens of thousands of compute machines
  - Process in hours what used to take days or weeks
  - Limited only by amount you can pay, but pay only for hours that you use



- **Easier Fusion Products**
  - Opportunity to create products with other datasets in the cloud
  - Without the cost to store
  - Without the time to download

What the Cloud is Not

The Cloud Does Not Solve All Problems

- **Processing in the Cloud is Plagued by Same Familiar Issues Related to …**
  - Metadata and data formats
  - Interoperability
- **Cloud Resources still have same "Hardware" Failures**
  - network timeouts and storage failures
- **The Cloud Is not Necessarily Cheaper ...**
  - Cloud cost models designed to be similar to total cost of ownership (TCO) of on-premise solutions
- **But … the Cloud Provides …**
  - "infinite resources"
  - pay-as-you-go options
  - Computing resources closer to the archive

THE GOALS OF THE GET READY FOR NISAR (GRFN) PROJECT

The Get Ready For NISAR (GRFN) Project

Project Goals

- **Project Goals:**
  - Understand **cost implications of various cloud-based and hybrid architectures** for NISAR science data system (processing) and data center (storage)
  - Get science community **familiar and comfortable early on** with interacting with and working on large SAR datasets in a cloud environment.
- **Approach:**
  - Build a **prototype NISAR processing system in the cloud based on Sentinel-1 SAR data**
  - Derive Sentinel-1 SAR data products to **socialize SAR data products to science communities**
    - L2 products (interferograms; covariance information) for expert SAR community
    - L3 products (deformation rates; time series; …) for broader science community
    - On-demand processing
- **SAR products used to understand how scientists will interact with NISAR data in the cloud to accurately estimate cost implications**

The Get Ready For NISAR (GRFN) Project

Expected Key GRFN Science Data System Capabilities

- **Prototype NISAR processing system with Sentinel-1 as proxy**
- **Up to L3 science data products for science focus areas** [solid earth; ecosystems; cryosphere; applications]
- **Simulate NISAR processing scenarios in the cloud**
  - Forward stream processing (“keep up”)
  - Bulk reprocessing
  - On-demand processing
  - Urgent response
- **Cloud-based collocation of processing system with ASF data center**
  - Establish cloud-based high-performance data delivery of L2 Sentinel-1 science data products from processing system to the ASF data center.
- **Costing – cloud economics**
  - Perform analysis necessary to produce costing reports needed for NISAR.

The Get Ready For NISAR (GRFN) Project

Key GRFN ASF Data Center Investigations

- **Ingest at NISAR Scale (bandwidth and volume)**
  - Achieved through shared storage by collocation of SDS and data center
- **Archive and Distribution (storage & distribution costs, bandwidth)**
  - Lifecycle and storage temperature (hot → cold → hot)
  - Data distribution via Earthdata Search Client & Vertex
- **Cost and Performance Implications of On-demand Processing Scenarios**
  - Standard product creation on-demand (virtual archive)
  - Bulk re-processing from various storage types
  - End-user processing system (bring processing to the data)
- **Science Community Outreach**
  - L2 product usability and convenience packaging
  - Fully Public Data access as a Beta product

CURRENT GRFN STATUS AND FINDINGS

Current GRFN Status and Findings

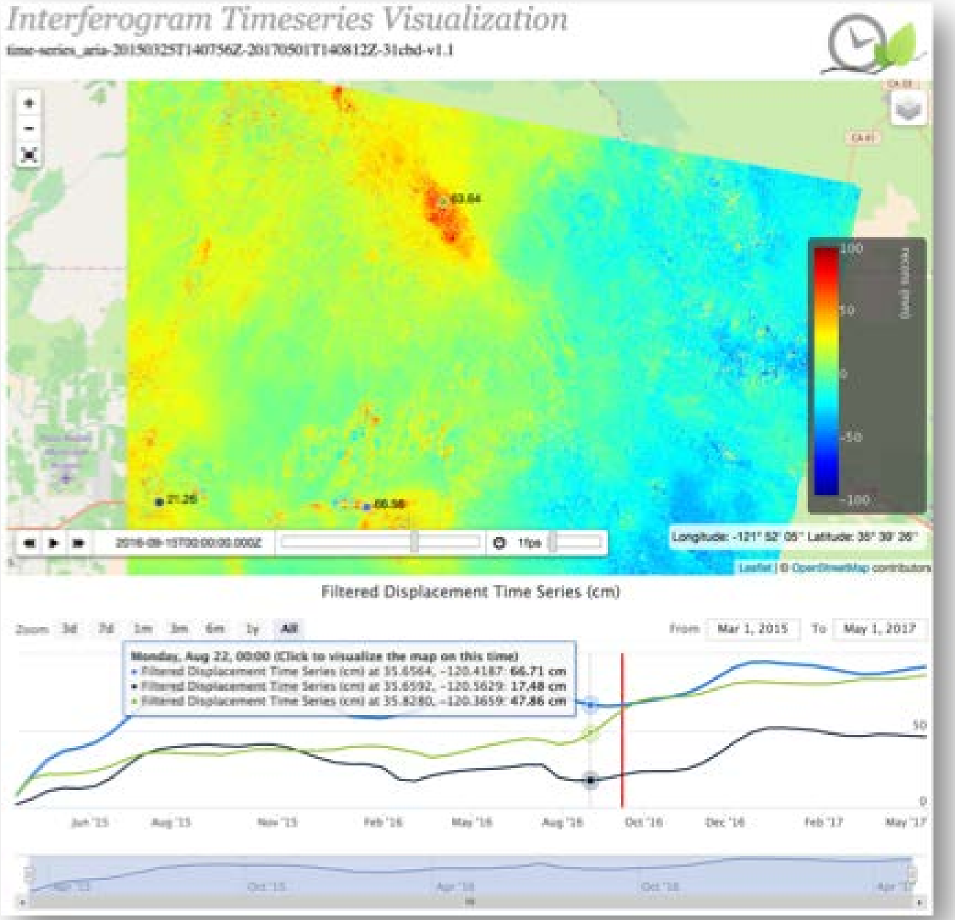
What Has Been Built/Done So Far

- **GRFN SDS and Storage System Built in Public Cloud Environment**
  - Now co-located with data center in AWS us-east region
- **Collocated SDS and Data Center to Avoid Large L0-L2 Data Volume Movement**
  - Capable of ingesting 10 Gbps at forward processing rates and 50 Gbps forward processing plus bulk reprocessing load
- **Automatic L2 (Solid Earth) Processing, Ingestion, and Distribution**
  - Products available via GRFN website, Vertex, and Earthdata Search client
- **On-Demand L2 (Solid Earth) Processing, Ingestion, and Distribution**
  - Limited scale, but also available via GRFN website, Vertex, and Earthdata Search client
- **Multi-Temperature Storage Prototype**
  - Used AWS native lifecycle policies and tracking to determine best (lowest cost) mix of storage classes based on Sentinel-1 distribution activity

Current GRFN Status and Findings

What Has Been Built/Done So Far

- **Cloud-Based Prototype of On-Demand L3 Displacement Time Series**
  - SBAS-type time series solution computed in AWS cloud
  - Time series analysis collocated with L2 interferogram stack

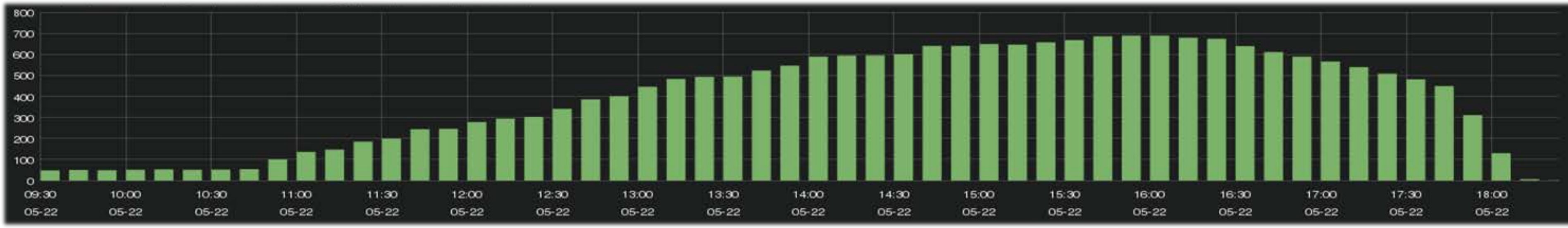


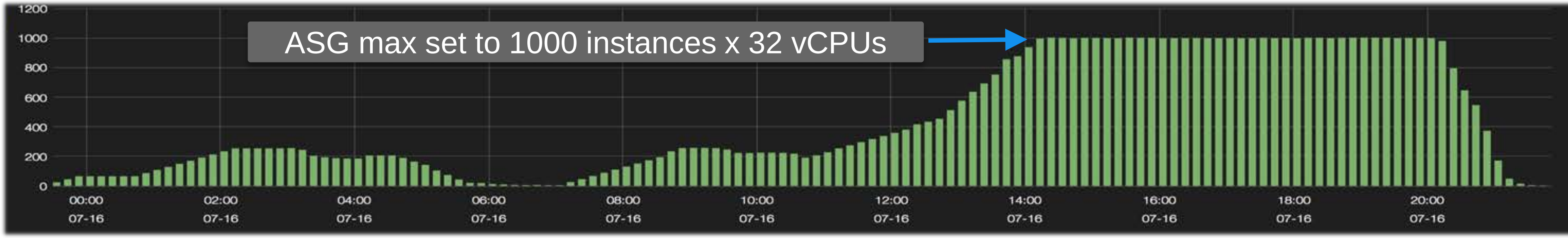
Load Testing: An Example of Cloud Scalability

- **The size of the science data system compute nodes can automatically grow/shrink based on processing demand**

- **Auto Scaling group policies can be added**
  - E.g., setting max scaling size to curb costs

Auto scaling enabling runs of over 100,000 ***vCPUs***





Current GRFN Status and Findings

Studying Performance and Cost Implications of Cloud Architectures

1. Egress Costs:

- Every file downloaded from the cloud will incur costs
- Depending on the download behavior of users, egress can be major cost factor for cloud-based architectures!

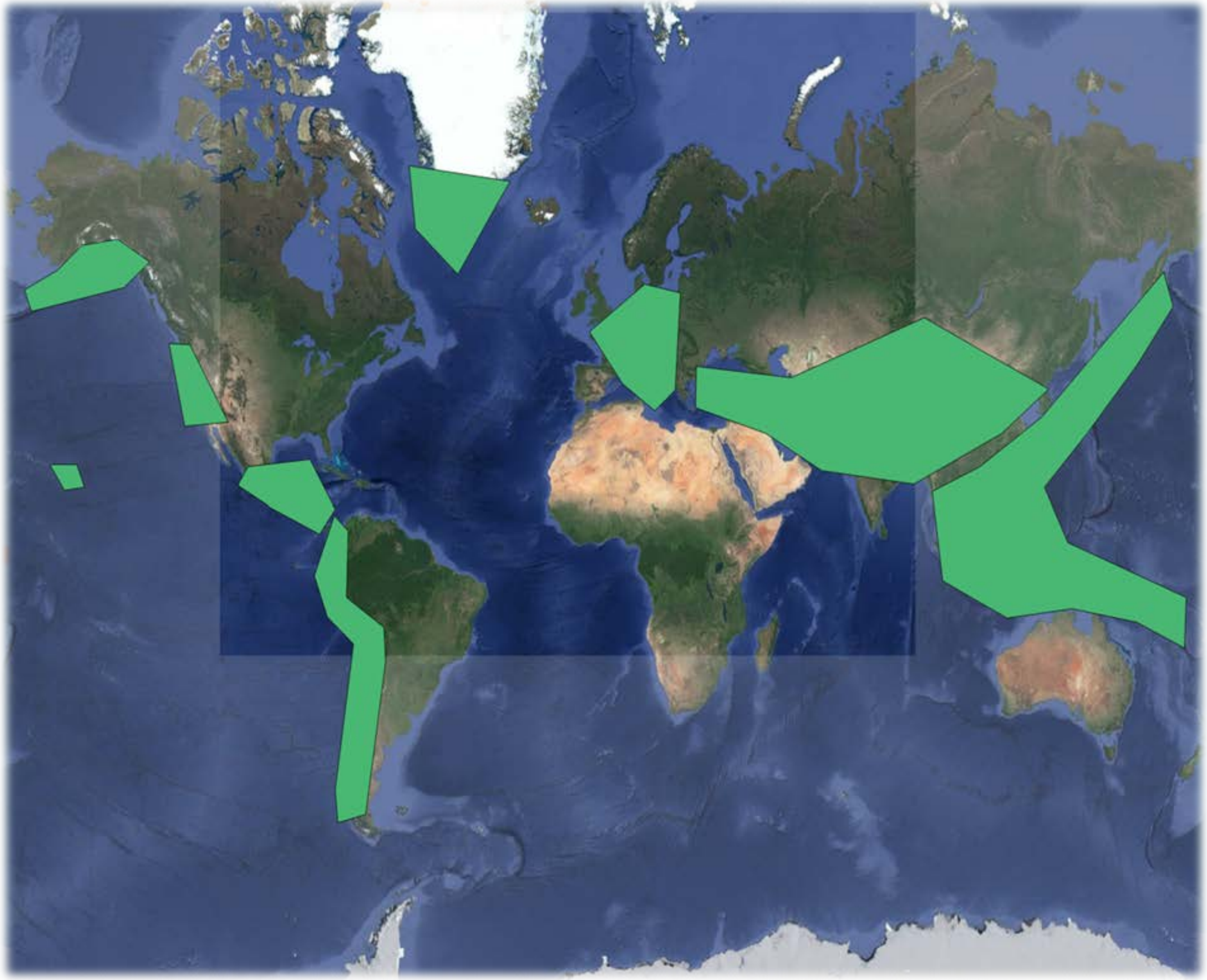
→ To Save Egress Costs, Hybrid Architecture Including Cloud, Edge Locations, and On-Premise Components is Recommended

- Store infrequently used data in the cloud (deep [cheap] storage)
- Serve "hot data" from on-Premise or cached edge locations → cheaper, even incl. direct connect & Hardware costs
- **Distribution of data among cloud, edge, and on-premise locations may be learned from user behavior**

Cloud Architecture and Egress Costs

Learn From Sentinel-1 SLC Download Patterns

Sentinel-1 SLC usage largely localized around Major Hazard Zones



Hot data zones (green areas) take care of >70% of Sentinel-1 SLC data downloads

→ Serving out data in green areas from on-premise or edge location can lead to large savings in egress costs

Current GRFN Status and Findings

Studying Performance and Cost Implications of Cloud Architectures

2. Compute Costs

→ Mix Fast-Path and Slow-Path Processing Streams to Curb Costs

- Spend reserved instance compute time on Minimum Viable Product requirements to get product to serving
- Queue extra compute to the slow path on Spot Market instances to be cheaper but possibly later (i.e. Browse generation)

3. Storage Costs

→ Make storage decision during Ingest

- Per file, likely based on Area of Interest (AOI), decide on most appropriate temperature storage class during ingest (i.e. Glacier [cheapest storage] for products that will likely never be touched)

Current GRFN Status and Findings

Identified Opportunities of the Cloud for End-Users

- **Advantages of Moving End-User Processing to Storage**
  - Fewer data moves [e.g. egress costs]
  - Faster results!
  - Process massively scalable next to storage
  - No hardware requirement/maintenance for end user
  - Fusion products with other datasets in the Cloud

How can we help you get started in the cloud?

Conclusion

- GRFN has shown the viability of cloud solutions for SAR data missions
- Various cloud architectures have been tested for performance and cost implications
- Some first findings have been presented today

- Not all functionality has been completely built and not all cost analyses are completed but progress is being made on all fronts

- **We just finished year #1 of GRFN – two more project years to come!!**

Looking for Beta Testers to Assess GRFN Products & Services

- **Earthdata Search**
  - Search for “GRFN”
  - https://search.earthdata.nasa.gov/search?q=GRFN&ok=GRFN
- **Vertex**
  - Missions Tab, Beta Products
  - https://vertex.daac.asf.alaska.edu/
- **GRFN/ARIA Science Data System**
  - https://aria-search.jpl.nasa.gov/ (data products, account sign up)
  - https://aria.jpl.nasa.gov/ (general info)

Questions?

Contact: fjmeyer@alaska.edu

AGU FALL MEETING

New Orleans | 11–15 December 2017

Session ID: 26762

Session Title: G015. Recent Advances in SAR and InSAR Processing, Analysis, and Cloud Computing

Section/Focus Group: Geodesy

Link: https://agu.confex.com/agu/fm17/preliminaryview.cgi/Session26762

Submit abstracts latest by Wed., August 2, 2017

A FEW ADDITIONAL LESSONS LEARNED

Summary and Lessons Learned

- **Engineering**
  - Development in cloud environment is more efficient and agile than traditional means.
  - Some cloud vendors (e.g. AWS) have strong community following that builds capabilities in and around cloud vendor (e.g., swiftstack storage)
  - Cloud vendors provide suite of hardened and ready-to-use data system tools and services

- **Infrastructure**
  - Collocation of compute and storage in same cloud region removes unnecessary data movements (and associated costs)
  - Moving processing to data archive improves efficiency for large data volume scenarios such as L3 time series from deep stacks
  - Data life-cycle policies can help to lower storage costs