Integrating Remotely Sensed and Ground Observations for Modeling, Analysis, and Decision Support

**Andrea Donnellan, Jay W. Parker**

**Jet Propulsion Laboratory**

**California Institute of Technology**

**4800 Oak Grove Dr.**

**Pasadena, CA 91109**

**818-354-4737**

**Andrea.Donnellan@jpl.nasa.gov, Jay.W.Parker@jpl.nasa.gov**

**Jun Wang, Yu Ma, Marlon Pierce,Geoffrey Fox**

**Indiana University**

**2709 East 10th Street**

**Bloomington IN 47408**

**wang208@indiana.edu, yuma@iu.edu, marpierc@iu.edu, gcf@indiana.edu**

***Abstract*—** **Geodetic imaging data from Interferometric Synthetic Aperture Radar (InSAR) are used to measure crustal deformation related to tectonic motions and displacements on earthquake faults. NASA’s UAVSAR project and related efforts are creating large catalogs of data products. The user base of these data products is also growing, introducing the need for downstream tools to support computationally expensive individual research as well as access to voluminous and heterogeneous data products. Bundling data inside a virtual machine becomes impractical for load balancing and on-demand auto scaling. A possible solution is to separate the application services from the data service. The Amazon public cloud would be utilized for computation and analysis and private data would be served from a private cloud through Open Geospatial Consortium (OGC) cascading services. We are using Amazon’s Elastic Compute Cloud (EC2), which is a basic virtual machine service, for cloud deployment. Elastic Load Balancing (ELB) is used to seamlessly distribute incoming traffic among multiple instances. Auto scaling with CloudWatch results in on-demand scalability. ELB detects unhealthy instances and automatically reroutes traffic, while auto scaling replaces the unhealthy instances to maintain high availability. Content is distributed globally through CloudFront where distribution is via a global network of edge locations, which optimizes performance by routing content to the nearest edge location. Amazon web services (AWS) cloud infrastructure provides easy deployment of highly available and on-demand scalable applications. However, applications requiring instant access of relatively large datasets face certain limitations from Elastic Block Store (EBS). A single EBS volume is limited to 1 TB, and as well as the high costs, and EBS volumes cannot be shared among multiple instances. The current process UAVSAR Repeat Pass Interferometry data products are about 2 TB and the volume continues to expand.**

Table of Contents

1. Introduction 1

2. UAVSAR Data Products 3

3. Amazon Cloud Services 5

4. UAVSAR Application in the Cloud 7

5. Interfacing InSAR Processing and Analysis 9

6. Discussion 10

7. Downstream Cloud Computing 11

7. Conclusions 12

Acknowledgements 12

References 12

1. Introduction

NASA geodetic imaging observations of crustal deformation are improving understanding of earthquake fault processes. The geodetic imaging provides observations of surface motions from Global Positioning System (GPS) and Interferometric Synthetic Aperture Radar (InSAR) data. The observations measure seismic and aseismic motions related to tectonic strain accumulation and release and complement other earthquake related data, which measure seismic waves or long-term geologic offsets on faults (Figure 1).

Rapidly increasing volumes of geodetic imaging data and increasing complexity of data processing algorithms have posed a challenge for data providers and end-users. The QuakeSim project is investigating how to build advanced data-intensive computing infrastructure with emerging technologies and platforms, such as grids and cloud computing, to allow rapid exploration of large datasets, and identify subtle but important features in large datasets [1].

Interferometric Synthetic Aperture Radar (InSAR) is a radar technology used in remote sensing to measure deformation at the Earth’s surface using interference patterns from two or more images captured at different times from the same perspective [2]. Integrated with models and geodetic GPS, geologic, and seismic data, InSAR provides estimates of crustal deformation that are key to improving fault models. The models are used for forecasting, simulation, emergency planning and response.



Figure . Schematic of the earthquake cycle of strain accumulation and release with the long-term tectonic rate removed.

**Page break before sections until final formatting**

2. UAVSAR Data Products

NASA’s Uninhabited Aerial Vehicle Synthetic Aperture Radar (UAVSAR) is an airborne L-band (23.8 cm in wavelength) imaging radar system, fully polarimetric SAR (POLSAR) with repeat-pass interferometric (RPI) observation capability currently carried under a Gulfstream III Aircraft (Rosen et al., 2006). In a typical mode of operation, 36-look slant-range polarimetric images are produced at 5 m x 7.2 m pixel spacing, which typically cover 20 by 100 km of the Earth’s surface. The UAVSAR system adopts GPS technology to precisely re-fly flight lines. High-resolution georeferenced interferograms (with 6 x 6 m ground resolution) are released as Repeat-Pass Interferometry (RPI) products, which also include multi-look amplitude products, interferometric correlation products, unwrapped interferometric phase products, and digital elevation models used in producing ground-projected images. All the data products are open to the public through JPL UAVSAR data portal (http://uavsar.jpl.nasa.gov/) and Alaska Satellite Facility SAR Data Center (ASF SDC, https://vertex.daac.asf.alaska.edu/).

UAVSAR products are distributed as single-band binary files with a text annotation file; the size of a single image product ranges from several hundred megabytes to several gigabytes. UAVSAR products also supply pre-rendered images packed in KML/KMZ formats, which can be visualized in Google Maps and Google Earth. As of October 2013, around 3600 POLSAR and over 400 RPI data products (Figure 2) have been delivered with the size estimated to be more than 40TB data volume. When an image is selected for further analysis, including download to a local platform for detailed study with specialized tools, more than the deformation image is usually needed. A typical case requires 7 GB for a complete data image including layers such as correlation data. For California, most locations are covered by multiple images captured at different times. The current file-based bulk-download distribution model from data center to end-users has become impractical to meet user needs and is particularly difficult for rapid response after a major earthquake event.

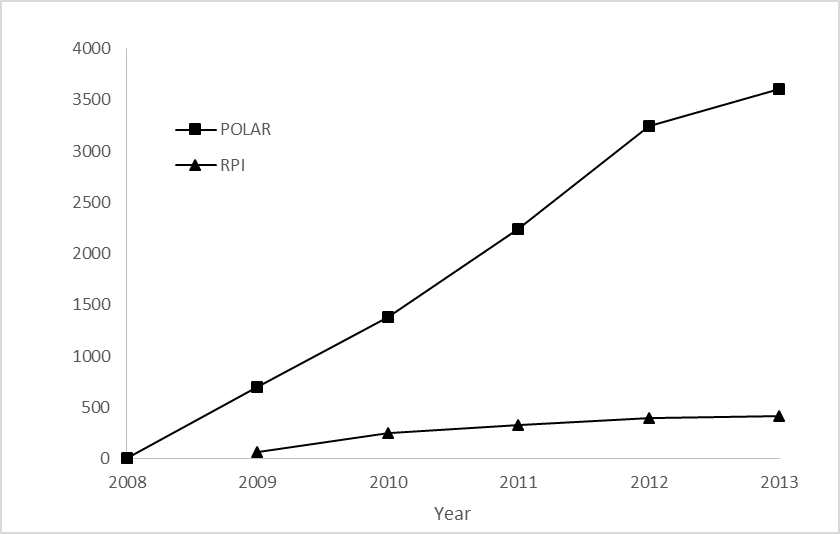


Figure . Total number of released UAVSAR datasets (source: ASF data portal and QuakeSim)

QuakeSim web service tools allow rapid exploration and feature extraction from UAVSAR RPI products. The UAVSAR RPI data products are converted into GeoTiff format and published with Open Geospatial Consortium (OGC) standards. The InSAR profile tool plots line-of-site changes of pixels on the ground relative to the flight instrument, referred to as ground range changes, for user-selected profiles across unwrapped UAVSAR interferogram products [3]. The user draws a profile line on the selected image and is presented with interactive plots that show values of the line-of-sight displacements and topographic height of the corresponding digital elevation model (DEM) along the profile line (Figure 3). Users can switch to different images through the list on the right. Pre-rendered images are displayed on Google Maps as part of an online interface to help users visually identify regions of interest.

Due to the increasing volume of UAVSAR data, it is necessary to keep the data and software service as close as possible. Commercial cloud services from Amazon and Microsoft offer Infrastructure as a Service (IaaS) for both computing and storage features. This study investigates the applicability of Amazon Elastic Compute Cloud (EC2) to geospatial earthquake study applications. Combining so-called “Big Data” with Platform as a Service (PaaS), which is constructed as a basic scalable infrastructure with higher level capabilities, is also discussed in this paper.

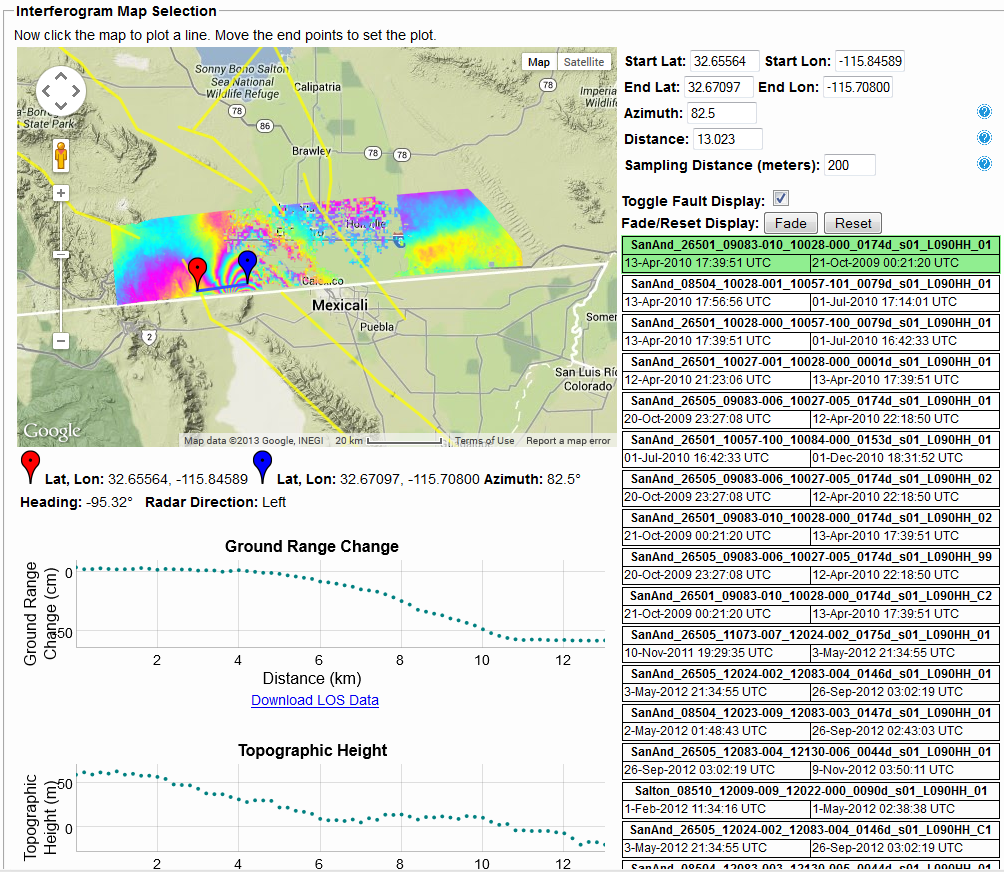


Figure . Main interface of InSAR Profile Tool (http://quakesim.org/tools/insar-profile-tool)

3. Amazon Cloud Services

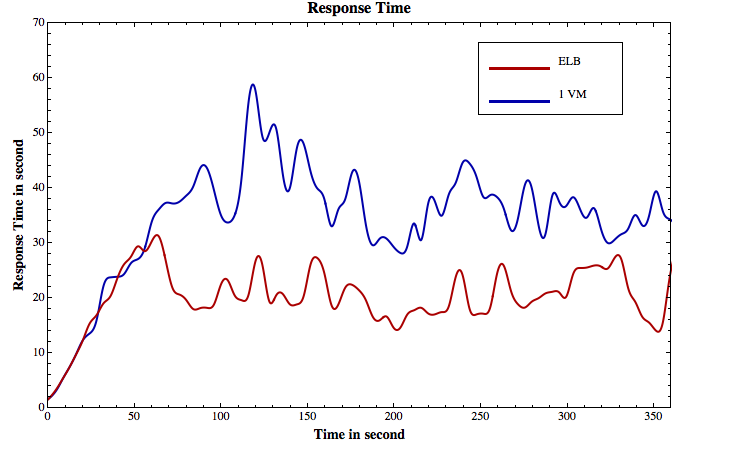


Figure . Response time of a single Virtual Machine (VM) vs. Elastic Load Balancer (ELB) with 2 VMs

Amazon Web Services (AWS) offer a comprehensive set of cloud infrastructure and application services. The following computing and networking products are of particular interest for deploying highly available and scalable applications on the cloud.

*Elastic Compute Cloud (EC2)* for application deployment provides basic virtual machine services such as creating images and launching instances. It offers a variety of instance types and configurations, highlighting flexible and elastic computing capacity for user applications.

*Elastic Load Balancing (ELB)* provides high availability of mirrored EC2 instances. While EC2 provides multiple regions and availability zones to launch instances and protect user applications from single point failure, ELB further enhances application availability by seamlessly distributing incoming application traffic across multiple instances. It also detects unhealthy instances and automatically reroutes traffic as necessary.

*Auto Scaling (AS)* with *CloudWatch* provide on-demand scalability that can automatically increase or decrease the number of EC2 instances according to predefined conditions such as Amazon CloudWatch and monitor metrics to respond to demand spikes and lulls. AS can detect and replace unhealthy instances, and be configured to work with ELBs.

*CloudFront* provides a global Content Delivery Network (CDN) that enhances application performance by delivering entire web contents through a global network of edge locations.

*CloudFormation* provides resource management through a collection of related AWS resources listed above. With CloudFormation, we can easily manage, provision, and update deployments through templates. The JSON-format, text-based file describes all the AWS resources needed to deploy and is instantiated as a stack.

Vector datasets are usually small enough (e.g. tens of GB) to be bundled within virtual machines. The cloud deployment of corresponding applications is therefore straightforward. Typical steps are:

1. Create customized Amazon Machine Image (AMI) based on a fully installed and loaded instance;
2. Create ELB for listeners and health check conditions;
3. Create auto scaling group using the ELB;
4. Configure scale up and scale down policies for corresponding CloudWatch alarms;

Third VM

Second VM

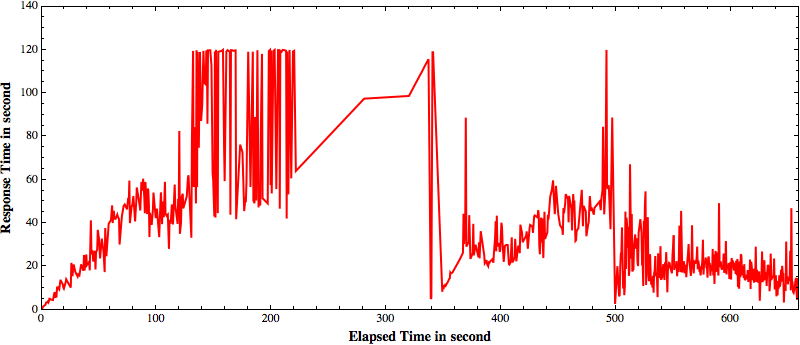


Figure . Changes of response time during the auto-scaling test. In 340 seconds, second VM started up by AS, reduces response time to about 40 seconds, and in 500 seconds, third VM started, reduces response time further to around 30 seconds

1. Configure CloudFront distributions;
2. Assemble all in a CloudFormation template.

To investigate functional features of ELB and AS, we use a sample application that deploys a large set of vector data of over 3 million polygons through PostgreSQL and Geoserver on Amazon EC2 m1.medium instances. Data are accessed via randomly generated spatial queries using WMS/WFS calls. Apache Jmeter (<http://jmeter.apache.org>) software is used to conduct the load test and simulate concurrent access of multiple users. The basic testing methodology is to ramp up 256 threads in 30 seconds, hold the load for 5 or 10 minutes, and then gradually stop all threads in 30 seconds. We use query response time as the performance measure.

ELB automatically distributes incoming application traffic across multiple EC2 instances. Figure 4 shows the performance improvement (average response time in seconds) during the 6 minutes testing period of two VM ELB over a single EC2 VM.

Auto Scaling allows users to scale the number of active EC2 instances up or down automatically based on predefined CloudWatch metrics. For testing purpose, our scaling policy is set to increase the number of instances by one when the CPU utilization is higher than 70% for 2 minutes, and decrease the number of instances by one when the CPU utilization is lower than 50% for 10 minutes. We also configured the auto scaling group behind an ELB to evenly distribute server loads. Figure 5 demonstrates auto-scaling effects on the response time over the 11 minute testing period: starting with a single EC2 instance, as the server load increases and about 3 minutes into the test, the query response time quickly hits the ELB default 60 second timeout, with one retry, capping the response time at 120 seconds; auto-scaling policy now enters the evaluation and about 2 minutes later, at 5 minutes into the test, a new instance was started to share the load, bringing down the response time to about 40 seconds; as auto-scaling evaluation continues, yet another scale up happened 2 minutes later, dropping the response time further down to around 30 seconds.

4. UAVSAR Application Services in the Cloud

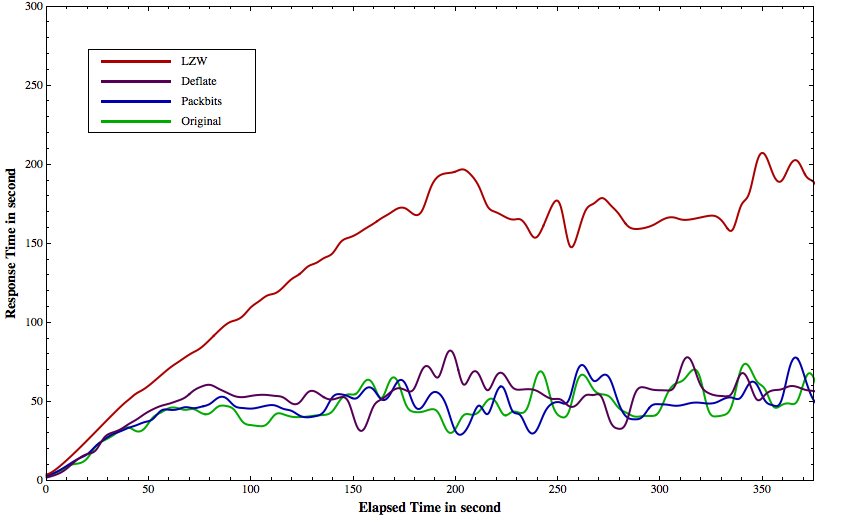


Figure . Response time of different compression methods.

The UAVSAR InSAR profile application handles much larger data volumes than the vector data setup in the previous section. Currently the QuakeSim InSAR profile tool uses unwrapped interferometric phase images and DEM data. New functions under development will extract information from the interferogram, amplitude (amp1 and amp2), and correlation images. This triples the total volume of images handled by the tool to nearly 8TB. We address this issue with two solutions. The first solution adds compression to the UAVSAR images to reduce the total data volume. The second uses a two-tier data server and application server architecture to scale up the service.

Compression has two benefits: one is reducing the network traffic to transfer UAVSAR images to the cloud, the other is decreasing the total storage requirement on the cloud. Geospatial Data Abstraction Library (GDAL) offers three lossless compression methods for GeoTIFF: LZW, Deflate and Packbits. Testing different compression methods shows that the deflate method has the best compression results for single band UAVSAR images (Table 1). Packbits is a simple run-length encoding compression n scheme, and it works for two images (UNW and COR) with lots of adjacent same values. For DEM data (HGT.grd), it does not provide any compression.

Table . File sizes for different compression methods

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Image | Original | LZW | Deflate | PackBits |
| UNW.grd | 507M | 117M | 113M | 197M |
| HGT.grd | 507M | 232M | 155M | 510M |
| COR.grd | 507M | 257M | 163M | 267M |

The performance impacts of compression are tested with randomly generated starting and ending points for the InSAR profile tool with 256 concurrent requests. Figure 6 shows the result, the average response time in second: uncompressed 41.96, Packbits: 44.25, Deflate: 52.31, LZW: 196.56. The packbits with 5% performance penalty is the best among three compression method, deflate has 25% penalty, and LZW has the worst performance with InSAR profile tool.

Bundling data inside the application server quickly becomes expensive and impractical for AS with large amount of raster data because EBS volumes cannot be shared among multiple instances. One simple solution is to separate the application services from data storage by having a dedicated high performance server for data hosting (Figure 7). Application servers can access the data through OGC cascading services while remain flexible for ELB and AS.

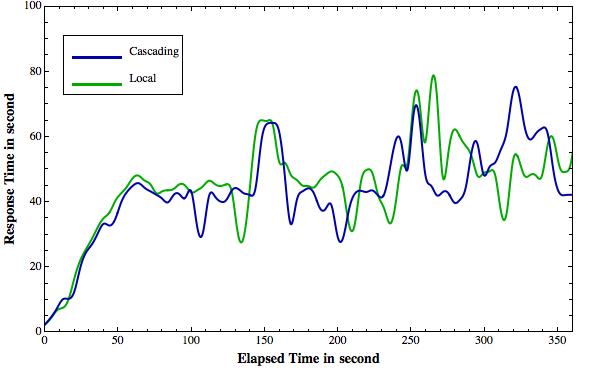
Two EC2 VM are configured for the performance testing: application server (m1.medium): 3.75 GB memory, 2 ECU; Data server (m1.large): 7.5 GB memory, 4ECU. OGC cascading tests query sent through application server to access cascaded data on data server. The figure 8 shows the response time of pulling data through cascading service for data server against the data hosed locally on application server. There is no noticeable performance penalty with two-tier setup.

Figure . Performance of cascading service vs. data hosted locally on application server.

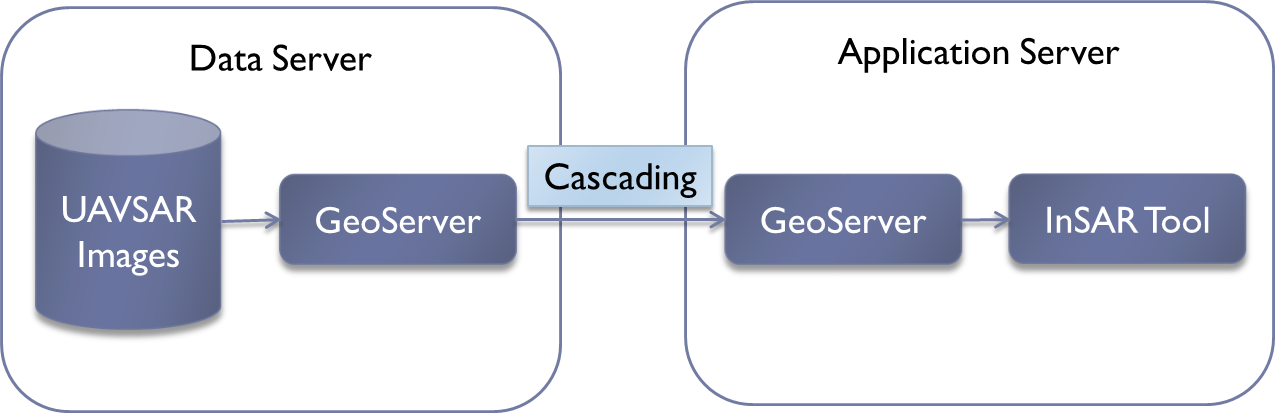


Figure . Two-tier architecture for InSAR application.

5. InSAR Processing and Analysis Chain

QuakeSim relies on data product providers and adds analysis tools for existing interferograms. QuakeSim’s current primary source of InsAR imagery is the UAVSAR data product collection. These data sets are currently integrated into the QuakeSim InSAR profile tool and services. The products are accessed from the Alaska Satellite Facility as they become available via an API (https://www.asf.alaska.edu/program/sdc/asf\_api). They are then converted to GeoTIFF, and stored in the profile tool’s Postgres databases for analysis by users. Metadata are also stored and users can download the complete data product set for further analysis.

CESCRE (Cloud-Enabled Scientific Collaboration Research Environment) is another candidate source for InSAR data sets. CESCRE processes spaceborne InSAR data on Amazon’s cloud using ISCE (InSAR Scientific Computing Software). Products are then handed off to the UAVSAR profile tool for further analysis.

InSAR data providers usually distribute data products and associated metadata in their own formats. As a down-stream data consumer, QuakeSim have encountered two particular problems when building a data processing service to convert InSAR data into GeoTiff.

The first problem is on how to mark out “null” value in a particular data product. There are generally two types of null values in an InSAR image; one is for pixels where no data were collected, and the other is used when the data were collected, but the return or the correlation is too low. In the second case, the pixel is regarded as a “bad pixel” or null value. It is necessary to clearly distinguish these two type of pixels.

The second problem is on metadata associated with InSAR products. Metadata is essential for both data processing and analysis. There are no general standards or guidelines on InSAR metadata. UAVSAR’s metadata is stored as parameter/value pairs in plain text format, while CESCRE distributes metadata in XML format. Also different terminology is used in these metadata files. For example “Ground Range Data Latitude Spacing” in UAVSAR metadata is equivalent to “Coordinate quantization” in CESCRE’s metadata.

The UAVSAR data processing pipeline is loosely coupled between ASF data center and QuakeSim. The new products are acquired either through JPL’s email notification or by periodically checking new product with ASF data API. The data has to be downloaded from the data center to QuakeSim server; it takes several hours to download a full data set. CESCRE processes InSAR data on Amazon Cloud, and it is possible to provide a tight integration with QuakeSim through Amazon Simple Notification Service (SNS). Amazon SNS enables users on the cloud to push messages to distributed services. This means once the new data set is available through CESCRE, it can send the notification to QuakeSim cloud service with the metadata information and data location on Amazon S3 storage, then QuakeSim data processing services can connect to data directly without downloading and start processing automatically.

6. Discussion

We have found hosting geospatial datasets services on Amazon’s cloud to be feasible for vector data. In this case, we were able to take advantage of the full range of Amazon features, particularly load balancing and autoscaling. With larger data sets, such as UAVSAR data, storage becomes a concern for both technical architecture and cost. We examined here a solution in which storage can be separated from the Web service using the cascading capabilities of OGC services.

Our analyses Identified the major cost was associated with hosting a server running (24x7). Amazon continues to lower prices on its AWS offerings. Based on the latest pricing, the cost of reserving a single m1.medium Linux server instance as used for our vector-data Geoserver test case is under $600 per year ($338 one-time upfront payment + $21 monthly fee = $590).

Amazon ELB and AS model serves extremely well for computing needs of emergency responses, where the application server usage is generally very light but spikes quickly in case of a natural disaster or emergency.

In addition to computing products described above, AWS also offers a variety of storage products. Elastic Block Store (EBS) provides highly available persistent storage for EC2 instances; Simple Storage Service (S3) is designed for Internet data storage and hosting using simple web interface (REST API) for I/O operations; Glacier is optimized for low cost data archival. Our experiment configuration fits best with vector data, where the dataset size is relatively small and can be easily bundled inside the application server for ELB and AS. The additional cost for corresponding EBS persistent storage is minimum. For example, the complete vector dataset we used for testing is under 10GB, which corresponds to $1 monthly fee.

AWS EC2 offers high storage instance hs1.8xlarge (48 Terabytes of storage) that can be used as the data server for just under $20K a year. The projected storage requirement in next three years, including both vector and UAVSAR/InSAR data accessible through QuakeSim tool, is about 30TB total, hence the additional archival cost using Amazon Glacier is $3000 annually.

Summing up, the total cost of hosting 30TB data product, with simple backup in Amazon Glacier and a separate Linux application server configured for elastic load balancing and auto scaling, is around $22K/year. Adding on actual on-demand usage fees, the annual cost of production server system should be under $25K.

7. Downstream Cloud Computing

[Could focus here on combining UAVSAR data products with Twister, Map Reduce, etc. Would need Geoffrey’s input].

7. Conclusions

Clouds are a promising new method for analysis and modeling of large data product sets and for multiple parallel runs. In some instances it may be better to use standard resources instead of clouds. Careful consideration should be given to what resources to use for various tasks. QuakeSim consists of several components interfacing with many different organizations, data products, and with several different types of applications for modeling and analysis. Collaboration between QuakeSim developers and data product providers and downstream consumers early on will result in improved efficiency across the board for earthquake studies and analysis of geodetic imaging observations.

Acknowledgements

This work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, and Indiana University under contract with NASA. The work was sponsored by NASA’s Advanced Information Technologies, Earth Surface and Interior, and Applied Sciences Programs. Part of this material is based upon work supported in part by the National Science Foundation under Grant No. 0910812 to Indiana University for "FutureGrid: An Experimental, High-Performance Grid Test-bed."

References

[1] A. Donnellan, J. Parker, R. Granat, E. De Jong, S. Suzuki, M. Pierce, G. Fox, J. Rundle, D. McLeod, R. Al-Ghanmi, and L. G. Ludwig, “QuakeSim: Integrated modeling and analysis of geologic and remotely sensed data,” in 2012 IEEE Aerospace Conference, 2012, pp. 1–9.

[2] R. Bürgmann, P. A. Rosen, and E. J. Fielding, “Synthetic Aperture Radar Interferometry to Measure Earth’s Surface Topography and Its Deformation,” Annu. Rev. Earth Planet. Sci., vol. 28, no. 1, pp. 169–209, May 2000.

[3] J. Wang, M. Pierce, Y. Ma, G. Fox, A. Donnellan, J. Parker, and M. Glasscoe, “Using Service-Based GIS to Support Earthquake Research and Disaster Response,” Comput. Sci. Eng., vol. 14, no. 5, pp. 21–30, Sep. 2012.

**Biographies**

|  |
| --- |
| ***Andrea Donnellan*** *is a principal research scientist at NASA's Jet Propulsion Laboratory, and Adjunct Assistant Professor of Research of Earth Sciences at the University of Southern California. She is Principal Investigator of NASA's QuakeSim project, which was co-winner of NASA’s Software of the Year Award in 2012. Donnellan was Deputy Manager of the JPL's Science Division, Pre-Project Scientist of an L-band radar mission, and NASA's Applied Sciences Program Area Co-Lead for Natural Disasters. Donnellan received a B.S. in geology from the Ohio State University in 1986, a master's and Ph.D. in geophysics from Caltech in 1988 and 1991 respectively, and an M.S. in Computer Science from the University of Southern California in 2003. She held a National Research Council postdoctoral fellowship at NASA's Goddard Space Flight Center. Donnellan has conducted field studies globally in tectonically active areas, and on ice sheets and glaciers, and has received numerous awards.* |
| ***Jay Parker*** *joined the Satellite Geodesy and Geodynamics Systems Group at JPL in 1996, and has been part of the JPL technical staff since 1989. He completed both a master's and PhD in Electrical Engineering from the University of Illinois (Urbana-Champaign), and graduated with a Bachelors of Science from the California Institute of Technology in 1981. His professional interests lie in applications of fast and accurate numerical models to geophysical remote sensing. Past modeling projects include vortex formation in the ionospheric D region, parallel supercomputer modeling of radar scattering and antenna power patterns, and high-fidelity modeling of clear-air infrared spectra for determining climate change and pollution sources. He is currently working on methods to invert SCIGN GPS data to determine earthquake and after-slip fault movements, finite element models of earthquake cycles, and new methods for GPS data processing on supercomputers. Jay has been inducted into Tau Beta Pi, and received a JPL Technology and Applications Programs Group Achievement Award. He is a member of the American Geophysical Union, and the IEEE Antennas and Propagation Society.* |
| Marlon_Pierce***Marlon Pierce*** *is the Assistant Director for the Science Gateways Group in Research Technologies Applications at Indiana University. Pierce received his Ph.D. Florida State University (Physics) in 1998 in computational condensed matter physics. His current research and development work focuses on computational sciences with an emphasis on Grid computing and computational Web portals. Prior to forming the Science Gateway Group, Pierce served as assistant director for the Community Grids Laboratory at Indiana University's Pervasive Technologies Institute. Pierce supervises the research activities of software engineering staff and Ph.D. students, and serves as principal investigator on multiple federally-funded research projects. Pierce leads research efforts in the following areas: the application of service-oriented architectures and real-time streaming techniques to geographical information systems and sensor networks; the development of open source science Web portal software for accessing Grid computing and data resources; and Grid-based distributed computing applications in computational chemistry and material science, chemical informatics, and geophysics.* |
| un Wang's picture***Jun Wang*** *is a GIS specialist in the Community Grids Lab, Indiana University. His current research interests are in the areas of large scale spatial data processing and visualization with Cloud computing technology. He joined QuakeSim team in July 2010, and develops the software components for QuakeSim visualization products.* |
| ***Macintosh HD:Users:andrea:Documents:Projects:Meetings:IEEE:2014:Cloud:Marie.jpegYu (Marie) Ma*** *is a Research Associate in the Science Gateways Group at Indiana University. She received a Ph.D. in Computer Science from Indiana University in 2006. She then joined the Research Technologies Applications Division at Indiana University. Her research focuses on data and metadata management framework for scientific applications. This includes distributed instruments in diverse areas, computational cyberinfrastructure for natural disaster management, web portals, and large scale data processing in cloud computing environments.* |
| ***Fox_G2_2004Geoffrey Fox*** *received a Ph.D. in Theoretical Physics from Cambridge University and is now professor of Informatics and Computing, and Physics at Indiana University where he is director of the Digital Science Center and Associate Dean for Research and Graduate Studies at the School of Informatics and Computing. He previously held positions at Caltech, Syracuse University and Florida State University. He has supervised the PhD of 62 students and published over 600 papers in physics and computer science. He currently works in applying computer science to Bioinformatics, Defense, Earthquake and Ice-sheet Science, Particle Physics and Chemical Informatics. He is principal investigator of FutureGrid – a facility to enable development of new approaches to computing. He is involved in several projects to enhance the capabilities of Minority Serving Institutions.* |