GeoCloud Project Report – - USGS/EROS Spatial Data Warehouse Project

Description of Application

The Spatial Data Warehouse project at the USGS/EROS distributes services and data in support of The National Map. A prototype environment was established to determine the feasibility of moving dynamic map services into a virtual environment with some added layers of management to simulate cloud computing. The description below encompasses complete SDW requirements and the prototype infrastructure.

Operating Organization

USGS EROS Data Center Spatial Data Warehouse (SDW)

Community of Interest

General Public, Raster GIS Data Mapping Services

OS and software requirements

Windows OS, IIS, PHP, MySQL, ESRI ArcIMS, Apache HTTP Server, and Apache Tomcat Server, ESRI ArcGIS Server, ESRI ArcGIS Server Image Server Extension

Prototype ran 3 scenarios:

- Windows OS, Apache HTTP Server, Apache Tomcat Server, ESRI ArcIMS
- Windows OS, IIS, PHP, MySQL, ESRI ArcGIS Server
- Windows OS, IIS, ESRI ArcGIS Server, ESRI ArcGIS Image Server Extension

Operational Requirements

While the applications, services, and downloads are available 24 hours per day, 7 days per week, they are actually only supported from 8am to 4pm CST. (Local business hours)

Image type (RAM, local disk)

The servers usually have dual-CPU, with 4GB of RAM for the ArcIMS and Web services, and quad-CPU, with 8GB or 16GB of RAM for the ArcGIS Server machines. (Some have 32GB of RAM, but they do not really need that much if services are apportioned properly.) Generally, the local disk space is limited (maybe 100GB or less) as the data storage is separate from the front-end processing/serving machines.

Prototype used 2 Server Machine Types:

- Dual CPU, 4GB RAM, 100GB local disk
- Quad CPU, 16GB RAM, 100GB local disk

Data storage

IO Intensive attached storage: 150TB (We hope to reduce this to less than ½ down the road.)

Web Accessible Object Storage: 100TB
Database Storage: 10TB

Prototype used:
3 TB Database Storage, 2TB IO Intensive Storage, and no object storage.

Upload monthly

Depends upon data ingest and compression, but ranges from 1TB to 10TB.

Download monthly

Varies, was almost 16TB for September 2011. Peak downloads were in December 2010 with 27TB. The trend line is approaching 20TB per month. In addition, there are anywhere from 20 to 40 million HTTP requests handled by SDW monthly.

Elastic IP

Using one for each hosted site supported through SDW, so at least 8.

Redundancy and Load balancing

All front-end interfaces are setup with redundant hardware to allow for bringing servers up and down for maintenance. Access to the servers is through a DNS resolution service that averages the load across all available servers.

Deployment in the Cloud

As SDW already uses machine images and configuration scripts to control the infrastructure, the concerns related to deployment in the cloud appears minimal.

Using existing images

SDW currently uses machine imaging to transfer configurations from one hardware machine to another. These same images deploy onto VMWare-based virtual machines.

Loading application with data

Data storage is separate from the physical machines already. Therefore, the loading process from an application perspective is just setting up the connection to the data.

Customizing application suite

Most of the customization relates to security, certification, and accreditation. Configuration changes occur on one server and then the settings transfer to other servers using the imaging process.

Installation scripting

SDW uses custom scripts that localize the server parameters. These run when a server image is setup from an image.

Operations in the Cloud

The prototype effort shows how feasible it is to move SDW infrastructure to virtualized or pooled resources. Even though it is feasible, the architecture is not optimal when migrated to the cloud. Architecting a solution for the cloud still needs to occur.

Monitoring of operations

SDW uses a custom application for monitoring applications and services. This uses Python, PHP, and MySQL. In addition to that, XYMon is used to monitor the systems. Google Analytics and Web Log Expert generate reports based upon web usage. During the prototype, the VMWare vCenter application monitored the health of the virtual machines and underlying hardware. The vCenter software also set alerts based upon usage, and controlled the virtual machine infrastructure.

Monthly usage, costs, (tables and charts)

Hosting all of SDW in a cloud environment could potentially be very expensive. Setting up 24 large windows reserved instances with 75TB of EBS storage that has 500 GB input per month, 2TB output per month and 8 elastic LB's, plus 100 TB of Reduce Redundancy S3 storage that has 2TB input per month, 20 TB output is in the range of $30-45K per month. To run just the services tested in the prototype would require 6 large windows instances, 5 TB of EBS storage, 50GB input per month, and 500GB output with no S3, or elastic LBs. This would cost anywhere from $2,500 to $4,000 per month.

Discussion

The cost estimates come from using the Amazon Simple Monthly Calculator. These are very rough estimates based only upon running the calculator and have not been validated through testing. In addition, the GovCloud and some other regions have higher costs than the US-East used in the estimates (up to 20% higher). The majority of the cost relates to the data storage, in particular the IO intensive storage. Reduce the IO intensive (EBS) storage to 1/10th of the current requirement (and the S3 storage increased by equivalent amount) then the costs are roughly 30 to 50% smaller.

Operational cost comparison (extrapolate to one year)

Overall, the annual costs are potentially similar between a private cloud and a public cloud. The main differences revolve around the capacity available for scaling and the optional services (like 24x7 support, and higher availability SLA's). A public cloud potentially shortens the time to deploy considerably by minimizing the procurement to a contracted service rather than hardware.

Fixed (one-time) costs versus monthly costs

SDW currently spends about $350K - $400K per year on IT Infrastructure. The Amazon estimate for this is around $420K to $540K (+$42K – $100K if running on GovCloud). The original purchase for the prototype cost about $45K-$50K for the hardware and software. The same equipment and software today would cost $35-40K. (The update pricing here is for current cost to compare to the Amazon estimate because it uses current cost.) The prototype on Amazon has a cost range of $30K – $48K (+$3-8K if running on GovCloud).

Telecommunications

Telecommunications costs are part of the overhead and/or underlying infrastructure for the facility. Estimating the portion of this that is particular to SDW is difficult. Even if obtained, it may not be valid in a cloud vs. internal comparison. Once moved the cloud, the 20TB outgoing data bandwidth would move, but there would still be 2-10TB of update bandwidth that would be going from here to the cloud resources. Therefore, this cost is likely to change very little.

Operations and maintenance support

The software development/support and management requirements are not likely to change with the hosting environment. SDW currently has about $750-$800K of expenses necessary to support the databases, servers, and IT infrastructure not related to the software development/support and management requirements. The move to a cloud environment still requires local infrastructure to support product development. Therefore, the operations and maintenance costs still exist, but shrink. Using a reduction estimate of 25 to 50%, the potential cost savings are in the range of $180K - $200K. Those savings only occur if there is a reduction in the required amount of IT infrastructure (not guaranteed). The COTS and in-house software still needs maintenance, support, patching, and upgrades, which means operations and maintenance support must continue. Normal operations and maintenance covers the expenses associated with the prototype.

Security plan development and approval

The security requirements are not different in the cloud environment. Compliance with NIST specifications and the Certification/Accreditation process still occur with lots of corresponding paperwork. While the change to a cloud environment may reduce the paperwork for underlying infrastructure, the COTS and in-house software, plus the local infrastructure still need this process. SDW currently has expenses of $200K - $225K per year. A 10% savings (which is probably generous) would result in $20K to $22K per year. Normal security support covers the expenses associated with the prototype.

Issues and Lessons Learned

The SDW project has some static or fixed resource requirements that do not adapt entirely to a cloud-computing environment. Storage used in online services that need moderate to high IO interaction is easily the most important consideration in determining the cost. There are break points in the Amazon Web Services where a smaller data size allows considerable cost savings.

Security approval process

Security and C&A on the public cloud still seems to be somewhat variable. Each organization is pursuing it individually. As such, there is little difference between the processes for internal versus external resources. A private cloud uses internal resources, which means existing processes and documentation makes the process relatively painless.

Recommendations on C&A

C&A should use existing documentation whenever possible.

Software deployment

Where possible, replace IO-intensive data sources in the architecture with object storage and pre-generated content. Doing this will directly influence the costs. Also, design pieces to run on separate service-based, or loosely coupled, infrastructure rather than monolithic software constructions. These portions of the software also need scalability designed into them. That combination of loose coupling and scalability allows the cloud infrastructure to use the principle of paying only for what is used and scaling just portions of the infrastructure.

Time-to-deploy

Procurement takes an extremely large amount of time. The use of the cloud changes this from procurement to an IT process. That process needs some definition to ensure that costs are appropriate and controlled, but this process is one measured by days rather than months. If the contracting for the underlying resources (public cloud or private cloud) has capacity for scaling, then time to deploy is almost negligible.

Failover, redundancy

Properly implemented, cloud environments have virtually unlimited capacity. This means that for properly designed software, the infrastructure scales easily. The catch is that capacity of the underlying infrastructure limits scalability. Private clouds work fine, but the underlying infrastructure must account for the future capacity. A public cloud often has an extremely large base of infrastructure available in pools that allow lots of room for growth. In addition, public clouds often have 24x7 support available as an option because it is a shared cost already needed by other customers.

Project planned future environment

Consideration about future infrastructure is ongoing. Initially, it looks like object storage (IE: Reduced Redundancy S3) in a public cloud will be used to distribute some file based data. Internally, some sort of cloud (or at a minimum virtualization) software will allow greater efficiency.