SwiftStack

CASE STUDY

VAPOR at Georgia Tech

University Academic Private Cloud for Researchers

The Georgia Institute of Technology, or Georgia Tech, a premier public research university in the US with over 21,000 undergraduate and graduate students. The largest of Georgia Tech's six colleges, the College of Engineering (CoE) offers more than 50 different degree tracks at the bachelor's, master's, and doctoral levels, and its schools are consistently ranked among U.S. News & World Report's top 10.

From VDI and HPC to Private Cloud

Georgia Tech has been deploying "private cloud like systems" since 2007 using a federated condominium approach where departments / researchers purchase their own servers and leverage a shared hosting infrastructure that includes the data center or HPC fabric, storage, monitoring, and other key components.

The projects supported by this shared hosting infrastructure include:

- The Virtual Lab Project (Vlab) supports a broad range of virtual desktops running specialized engineering applications.
- The PACE environment supports an extensive federated High Performance Computing (HPC) environment consisting of over 32,000 CPU cores and 2 Petabytes of storage.

The next evolution is the creation of a distributed and federated academic cloud (VAPOR). Initiated in early 2014, this project is led by multiple academic units including the College of Engineering, College of Science, College of Computing, Library and the central HPC Pace Group.

The goal for VAPOR is to support instruction and research on cloud computing, and other research that relies on cloud computing infrastructure. The Federation and Distributed design elements bring promising benefits such as experimental risks compartmentalization (being able to test the latest bleeding edge technologies without taking down the entire cloud), preservation of research and instructional autonomy while sharing common investments and infrastructure, leveraging multiple tier 2 / tier 3 server rooms, increased resiliency to failure and facilitated inter-disciplinary collaboration beyond just the existing HPC environment.

Flexibility was a key requirement for the VAPOR cloud. In order to support experimentation and rapid iteration, the cloud's design and supporting components needed to be able to adapt and adjust as both the researchers' needs and available storage technology evolved.

Data Everywhere

Behind any computational activity performed at Georgia Tech, whether on cloud systems like Vlab, PACE and soon VAPOR, or on detached storage like laptops, you will find valuable research data. Indeed data is the life blood of research and education. Every day Georgia Tech's researchers and students acquire, create, exchange, receive, and archive research data.

The challenge of long term curation and making research data created under federally funded grants

SPOTLIGHT

USE CASES

- Federated storage pooled across multiple departments, colleges and facilities
- Automated replication ensures DR backups
- Mix of long-term and short-term storage requirements

RESULTS

- Simple to deploy, turnkey manageability
- Single pool of storage serving all needs
- Low TCO from legacy storage reuse

HARDWARE

- Both new storage and existing hardware to reuse prior investments and lower TCO
- Numerous data centers scatter across a central campus and multiple satellite facilities

STORAGE

- Mix of existing legacy NAS/SAN and new node and storage hardware
- Over 2 petabytes needed just for HPC facility
- Authentication via existing LDAP and AD

easily accessible to the public is significant. This difficulty has increased significantly over the past few years, especially following the publication in 2013 of a Memorandum for Increasing Access to the Results of Federally Funded Scientific Research from the Office of Science and Technology Policy.

Existing practices often had research data "stashed" away on USB drives or older forms of media storages, consumer cloud services like Dropbox, or in isolated and obsolete enterprise storage that was far too costly to scale up. Current estimates are that 5 to 6 PB of research data is stored centrally at Georgia Tech. Plus an additional 1-2 PB exists scattered across laptop, workstation and USB drives.

The nature of this data varies greatly due to so many scientific and engineering research activities ranging from jet engine combustion data to interstate video for transportation research. Any storage foundation for VAPOR needed to accommodate all of these needs, both now and in the future.

SwiftStack Object Storage: the nexus of next gen data oriented services

During early brainstorming on the design requirements for VAPOR it became clear that this was the perfect opportunity to create an open and scalable storage infrastructure. Ideally this infrastructure would not only be capable of storing a vast amount of data but would also serve as a foundation on top of which new data oriented services could be developed. While these next generation services will evolve over time, an initial set included data analytics, research data repositories, and long-term data storage and curation.

By leveraging de-facto standards with broad industry support like the open source OpenStack Swift project, along with SwiftStack's management suite, the VAPOR project has been able to quickly deploy an object storage infrastructure that is both distributed and resilient. The combination provided multiple ways to store and access data via Swift's object HTTP APIs, pluggable services, and SwiftStack's filesystem gateway, enabling easy integration with current practices for research data capture and exchange.

SwiftStack's support for commodity storage node hardware provides benefits beyond just lower costs and no vendor lock-in. Faculty groups and research teams can contribute to VAPOR's storage expansion by converting their existing and planned storage capacity to the project and adding them into the new storage clusters. This bring-your-own-drive model makes supporting VAPOR easier within current budgets and also eliminates the prior problems of isolated and inaccessible data.

> *"As part of our new distributed and federated academic cloud initiative we needed an object storage platform with which we could experiment, scale up, and iterate quickly."*
>
> — Didier Contis, Director Technology Services, College of Engineering, Georgia Institute of Technology

Future Plans

Looking forward, the next set of challenges Georgia Tech faces with VAPOR is migrating both manual data management practices and automated data capture systems from file-based methods to SwiftStack's native object storage capabilities and HTTP APIs. Storing and accessing research data directly as objects will bring clear benefits from indexing, metadata, scalability, and especially performance. Units like the College of Engineering are helping to ease this file to object transition by providing technical assistance to researchers as well as an initial pool of free storage capacity.

In addition, the IT group is evaluating using Apache Spark cluster computing to process research data stored in SwiftStack to support a material genome initiative.

Find Out More

For more information on SwiftStack's features, support, pricing, and product documentation, visit www.swiftstack.com.