A Storage Architecture for Resilient Assured Data

Paul Manno

Georgia Tech / PACE

Date May 2019

Research Computing at Georgia Tech

- Georgia Tech: Founded 1885
- PACE- Partnership for an Advanced Computing Environment
  - 14 years (almost)
  - 1200+ Researchers
  - 50,000+ x86 cores
  - 10PB storage
  - 14 FTEs (and hiring!)
  - OSG, NSF, Big Data Hub, etc.
- Many research areas
  - LIGO
  - NSF
  - OSG
  - Health

Georgia Tech
CREATING THE NEXT

Georgia Tech: The New

- John Portman & Associates
  - CODA tower
    - 645,000 sq-ft office tower
    - Opened March 2019
    - Tallest “spiral” staircase in the world
    - First dual-cab elevators in North America
  - Collaborative space
- Databank, Inc.
  - Data Center
    - 60,000 sq-ft usable
    - 10+ MW
  - Open June 2019

Georgia Tech
CREATING THE NEXT

SOME Definitions

- What is ”…Resilient Assured Data”
  - We want it all: Speed, Availability, Accuracy, and Low cost!
  - Probably expect availability as top priority
  - Followed by speed vs cost and accuracy?

- What about security?
  - Do you need data secured at rest
  - Do you need data secured in flight

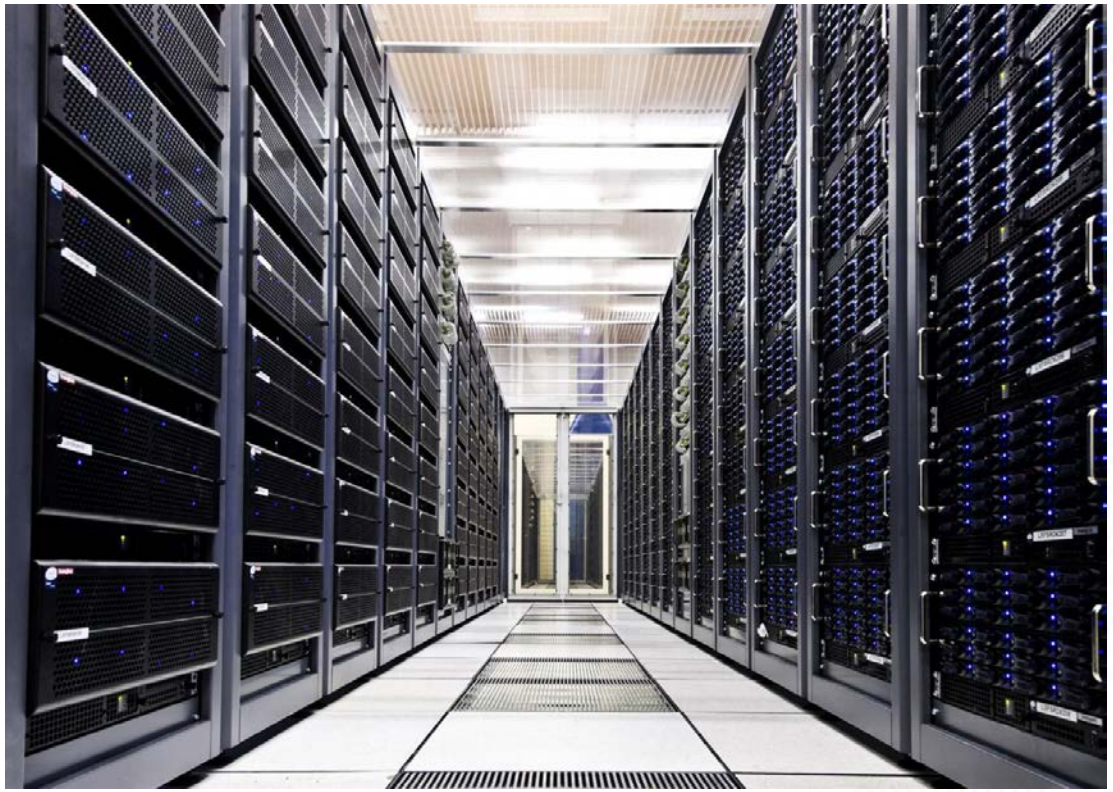
- Do you require geo-diversity?
  - Across a campus / town / country / world

Design Thoughts

- Simple example: Archive Tier of Storage
  - We have a need to store a bunch of cool or cold data for “a while”
  - Cost should be low
  - Maintenance requirements should be low or minimal
  - Convenient for multiple operating systems, platforms
  - Speed needs to be “acceptable”
  - Data could be recalled even after several years
- Types of information to be kept
  - POSIX files?
  - Objects?
  - Metadata?

More Design Considerations

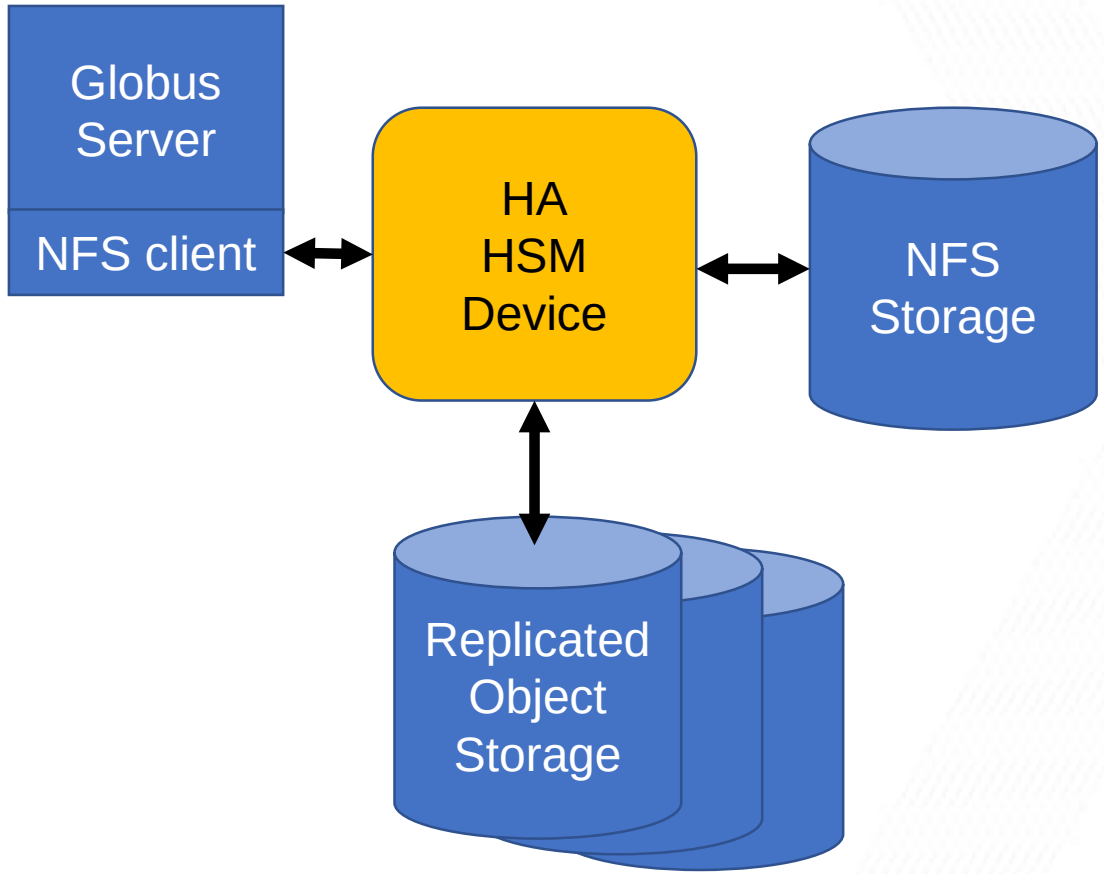
- Method(s) of access
- Computing platforms to support?
- Automation opportunities
- Long term options
  - On-Prem “cloud”
    - Data Center
    - Maintenance
  - Public cloud
    - Networking
    - Cost



Google-searched image used without permission

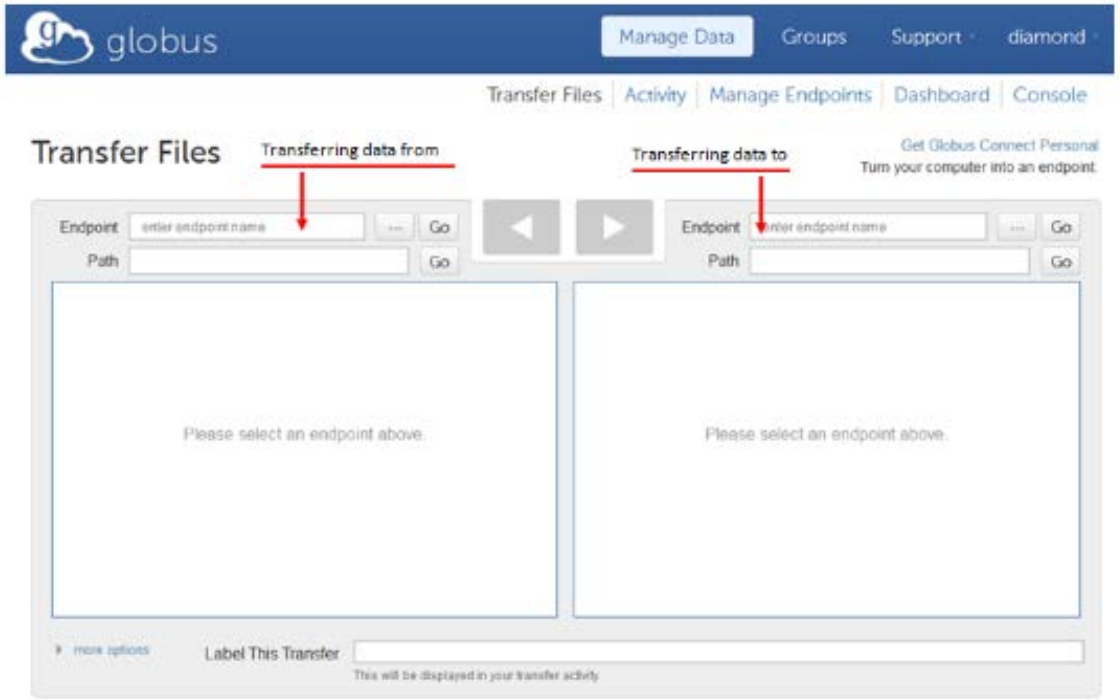
One Archive Solution (There Are Several)

- User Interface: Globus
  - Common across all platforms
  - Capable, extendable, reliable
- NFS Client and Storage
  - Inexpensive, reliable, efficient
- Highly Available HSM
  - … more on this in a moment
- Replicated Object Storage
  - Commonly available
  - On-Prem, Off-Prem, Hybrid



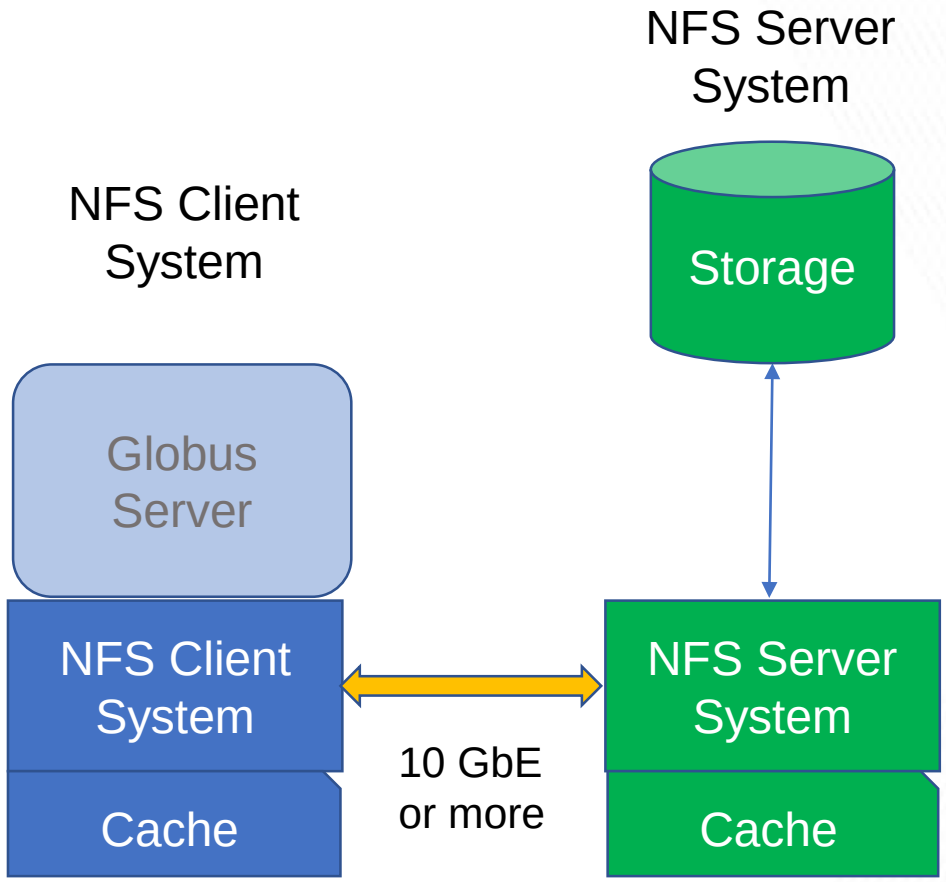
The Archive Parts – Globus User Interface

- Why Globus?
  - Long history of reliable transfers
    - XSEDE standard
    - Parallelizes transfers (configurable)
    - Auto-resume on interrupted transfers
    - Local and Wide-Area network support
    - Notification of success/failure
  - Platform agnostic
    - Transfers available via web front-end
    - Works to/from local system
    - Works to/from 3 rd party systems
  - Agnostic authentication
    - Just about anything
    - Shibboleth included



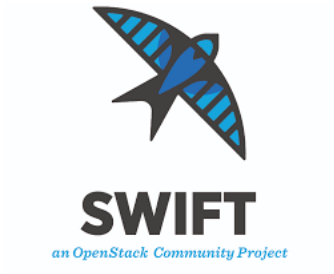
The Archive Parts – NFS Storage

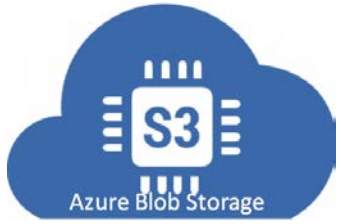
- Network File System (NFS)
  - NFS Service v3 or v4
    - Caching (can be important)
    - POSIX-based
    - Not seen by user (in this design)
    - HA service available
  - NFS Client v3 or v4
    - Caching (can be important)
    - POSIX-based
    - Not seen by user (in this design)
    - Multiple clients can use one server
- Caches help some operations



The Archive Parts – Replicated Object Storage (part 1)

- Why object storage?
  - Binary Large OBjects (BLOB)
  - Easy storage add / delete / move
  - Geographic Dispersion
- On-Prem Object storage
- Off-Prem Object storage
- Hybrid Object storage
- Speed considerations
- Objects known by
  - Object ID, Version, etc.





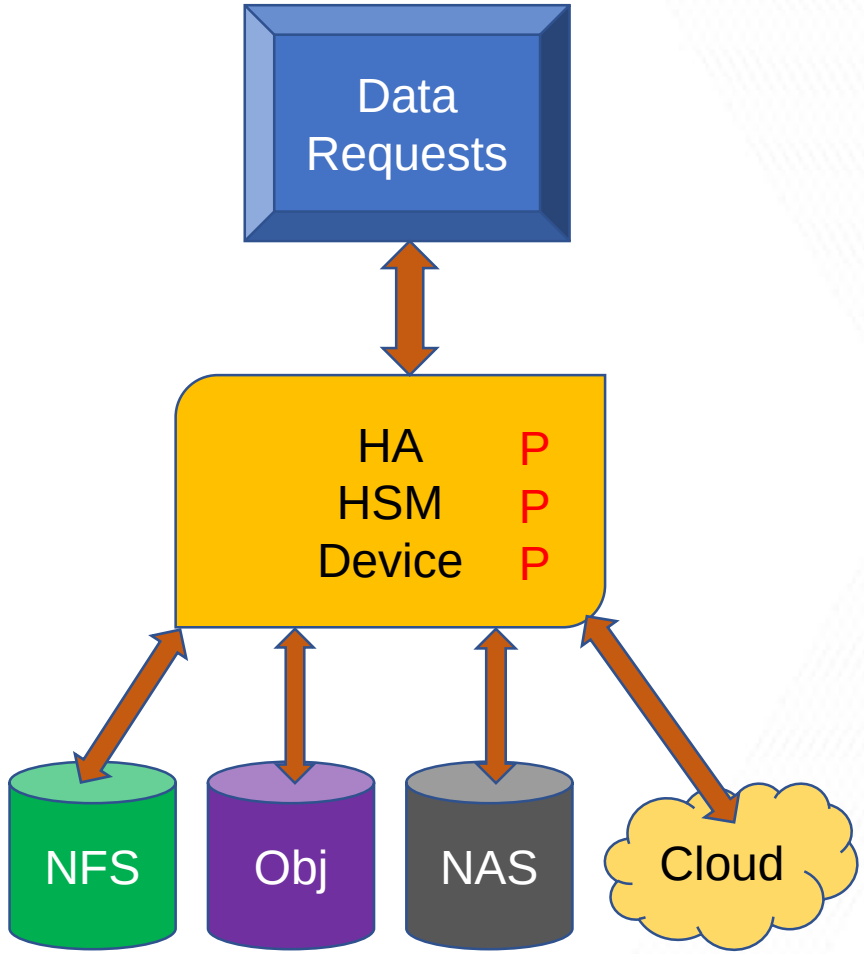
And Many More!

The Archive Parts – Replicated Object Storage (part 2)

- Object push
  - New object ID
  - Typically, versioning is on
  - Object-id re-use?
- Replications
  - Accepted is 3 copies but …
- Object read
  - Get object from wherever available
    - Source optimization
    - Size doesn't really matter
- Do you know data is "good"?
  - Metadata attributes
    - POSIX information
    - Versions
    - User information
  - Checksums, et. al.
  - 3 copies, compare data
- Encryption (many options)
  - At rest
  - In flight
  - In memory

The Archive Parts – HA HSM Device

- Some last definitions
  - Primary Storage
  - Secondary Storage
  - Tiered Storage
- Highly Available
  - Virtual IP addresses
  - Multiple units must synchronize
- Hierarchical Storage Management
  - The ”magic” happens here
  - Policy-based decisions
  - Multi-tier storage options
  - **Transparent to users**



The Archive Parts – What About Scale?

- Depends on the HSM
  - Some can be clustered
  - Some are built-into file system
  - Some are “bump in the wire”
- One HSM (Infinite IO) claims
  - Clustered operation
  - 3,000,000 MD requests per second
  - Many billions of files
- What about performance?
  - Secondary storage varies latency
  - Performance varies by network
  - Objects are relatively quick
- Archive vs Backup
  - Archive
    - Long Term Retention (years)
    - Versions are helpful
    - How to “refresh” technology?
  - Backup
    - Think business continuity
    - Versions are essential
    - Backup is not just copy
- Size of “things” to stored
  - Scans, Videos, Source Data
  - Becomes PB very quickly

How is this massive?

- Sizes of data to be stored
  - Grow to 100s of PB of storage
  - Many billions of objects
- Replication of objects
  - Can be any geography
  - Clustered HSM update lag
- Built-in HSM solutions
  - May work better
  - May be less-flexible
- Data Lakes (vs. Data Swamps)
  - Flexibility is key

Lessons Learned (so far)

- Change is ”bad”
  - Users don’t want things to change
  - Procedures are often rigid
  - Transparency is key
- Change is “good”
  - Accept technology updates
  - Newer / Faster / Better / Stronger
  - Transparency is key
- Pricing of off -prem storage
  - Pricing models vary considerably
  - Ingress/egress charges vary
  - Be sure to ask carefully
- Users like Globus ok
  - The GUI is intuitive
  - There is support
  - Users like point-and-click
- Data Management
  - Requirements vary
  - Inspect terms carefully
  - Often locations can’t change

Questions and Discussion

Many options to discuss …

What are your thoughts?

Paul Manno
Cyberinfrastructure Lead
Georgia Institute of Technology
756 West Peachtree Street, Northwest
Atlanta, GA 30332-0700

Georgia Tech
CREATING THE NEXT