Really Big Storage: A 10,000 Petabyte Storage Cloud

Chris Gladwin
Cleversafe, Inc.
Cleversafe Overview

- **Market Opportunity**
  - Growing data increases storage system capacity requirements
  - Current data storage systems not designed for large scale capacity

- **Cleversafe Background**
  - Founded in 2004
  - Received $31.4 million in additional equity financing in Q4 ‘10
  - 100 people located in Chicago, plus Federal office in Virginia
  - Provide limitless scale data storage systems
    - Petabytes to Exabytes to Zettabytes
    - Most Reliable, Secure and Cost-Efficient
  - Multiple commercial deployments
    - Social media, financial services, media & entertainment, health care
  - Technology Leader
    - Multiple Industry Awards: Wall Street Journal, Business Week, etc.
    - Innovation Leader: fastest growing new storage patent portfolio
Data Storage Growth

**All Stored Data in 2008:**
800 Exabytes

**All Stored Data in 2020:**
35,000 Exabytes
= 35,000,000 Petabytes
= 35,000,000,000 Terabytes

IDC projecting a 44X increase in stored data
Very Large Scale System Requirements

How do you analyze and store data at this scale?

Very Large Scale Processing Requirements
- Analyze Gigabytes to Terabytes per second

Potential Solutions:
- Massively parallel, distributed pioneered by Google, Yahoo, etc.

Very Large Scale Storage Requirements
- Store Petabytes to Exabytes
- Gigabytes to Terabytes per second of I/O
- Growing 30%+ per year

Traditional data storage systems are not capable of this scale

>> Cleversafe Focus <<
Key System Challenges

Limitless Scalability
- Petabyes to Exabytes to Zettabytes

Reliability and Integrity at Scale
- Data integrity when the system size is 10,000,000,000 times larger than the bit error rate of a hard drive
- Reliability in a storage system where > 100 drives fail every day with week-long rebuild times
- 24x365 availability when devices are always being repaired, replaced and moved
- Data security in a system with millions of devices in multiple locations

Performance at Scale
- Data concurrency with multiple, simultaneous data writers and readers

Cost Effectiveness and Manageability at Scale
- Life-cycle cost-efficiency requirements for capital, electricity, cooling, space and system management
How Dispersed Storage Technology Works

1. Data is divided into slices using *Information Dispersal Algorithms*

   \[ \text{Total Slices} = \text{"width"} = N \]

2. Slices are distributed to separate disks, storage nodes and geographic locations

3. A threshold number of slices are retrieved and the original content is regenerated

\[ \text{Subset required to read} = \text{"threshold"} = K \]
Tunable Information Assurance

10 nodes needed to break confidentiality or integrity

10-of-16 Configuration

7 nodes needed to break availability

5 nodes needed to break confidentiality or integrity

5-of-9 Configuration

5 nodes needed to break availability

3 nodes needed to break confidentiality or integrity

3-of-8 Configuration

6 nodes needed to break availability
System Components

Dispersed Storage Infrastructure

Client(s)

Accesser(s) - SO HTTP Server

Slicestors - Storage

LAN / Internet

Manager

Embedded (SDK)
Traditional Storage - Assuming application-level metadata

Data (Objects)
- Replication required for reliability
- Physical storage 3X or more times data stored
- Multiple copies with RAID parity is expensive & difficult to manage
  - Decreased reliability as hard drives capacities increase
  - Decreased data integrity as system grows

Metadata
- 1 record per object
- 1 metadata update per data update
- Requires large amounts of memory & I/O
- Massive metadata database is impractical in large scale
  - Slower performance as system grows
  - I/O demands limits system scale
  - Single point of failure

Difficult to scale to Petabytes, cannot support Exabytes
Yottaspace Addressing

Approach
- Namespace is a mapping from Slicestor server to Slice Name Range
- Used by the client and store to determine where each Slice goes
- Objects are randomly named and evenly distributed across servers
- No Master controllers, no databases, no name node, etc.

Benefits
- Limitless scale storage
  - Mapping “Database” size is only ~120 bytes per server and cached by every client
  - Only one record per server
  - Only updated when servers are added or changed
  - Support quadrillions of object or more
- Each I/O transaction is independent
  - Performance scales linearly as parallel I/O occurs across multiple storage servers
Storage System Scale

Traditional Storage

Data (Objects)
- Replication required for reliability
- Physical capacity is 3X (or more) than actual data

Metadata
- 1 record per object
- 1 metadata update per data update
- Requires large amounts of memory & I/O

Dispersed Storage

Data (Objects)
Dispersal provides reliability of multiple copies with physical storage costs of a single copy

Yottaspace™ Addressing
- Requires only 2MB of memory per 10 Exabytes of data
- Limitless System capacity and performance
- Reads and writes continue even if not available
- Add, change or remove nodes while operating

3-6 X less storage with improved reliability

1 Billion times less memory required
10 Billion times fewer transaction I/O required
Example Use Case

What if you wanted to store and analyze 10 Exabytes in 2014 growing to 1,000 Exabytes in 2025?

Assumptions:

- Ingest: Uniform 24 x 7
- Retention: 6 months
- Object Size: 1 MB (avg.)
- Read / Write: 1:1
- Data Vaults: 1

Storage Requirements (Capacity):

- Actual: $1.5 \times 10^{19}$ B (10 EB usable)
- Ingest Rate: $\sim 938$ GB / sec
Very Big Data Sources

- 7 Terabytes per second growing 100X over 10 years

Near Real-time Parallel Data Analyzers (and filters)

- Multiple Simultaneous Writers
- 1 Terabyte/sec total growing 100X over 10 years

Very Large Object Storage Cloud

- 6 months of data (rolling)
- 10 Exabytes growing to 100’s of Exabytes
  - Ten million to One billion Terabytes

Secondary (Parallel) Data Analyzers

Secondary (Parallel) Data Analyzers

Metadata Database
- Data Characterization
Notional Example

Flat Architecture
Massively Scalable
Shipping Container Housing
Multi-Site Deployment

Ingest: 1.5 EB/mo.
Raw: 16 EB

Host Application

“Integrated Accesser” 1.5X expansion

~ 938 GB/s

~ 11 GB/s

Container 1
Container 2
... Container 88

663 GB/s

Standard shipping container
176 PB

~ 11 GB/s
Notional System Configuration

**Concept**
- Container (ea) 176.4 PB
- Ingest / Container 10.7 GB/s

**Container (ea)**
- Racks 21 (+1 Network)
- Servers 210
- Ingest / Rack 508 MB/s
- Power ~ 315 kW

**System**
- Containers 88
- Racks 1,848 (+88 Network)
- Servers 18,480
- Drives 1.6 Million
- Power ~11 MW
- Capacity 15 EB raw (10 EB usable)

**Server:**
- 5 U
- 84 Drives
- 840 TB
- 10 TB Drives [SATA or SAS]
- Ingest demand – 51 MB /s
- Drive demand – 0.6 MB/s
System Grows while Operating...

- 3 year Implementation – 72 Containers
- Rolling 3-year Technology Refresh
- Drive growth: 5TB, 10TB, 20TB, 40TB, 100TB, etc.
Example Deployments

- The Museum of Broadcast Communications has stored 100,000 hours of digital video for the past Two Years with Cleversafe serving over 5 Million online visitors
  - Enables large scale content distribution
  - Most cost-effective approach
  - Provides centuries-long storage

- Cleversafe Selected by In-Q-Tel for Strategic Deployments Supporting U.S. Intelligence Community

- Largest photo sharing site deployed 10+ PB to store billions of digital photos
  - Eliminates frequent file loss and file corruption
  - 70% cost reduction over prior single copy

- Large cable operator has deployed 400 TB of Cleversafe as origin storage for video-on-demand
  - Most flexible and easiest to manage storage
  - More cost effective method to provide high reliability
THANK YOU