How to Store Data to the Cloud without giving the Cloud anything

Jason Resch
Cleversafe
Topics

- Barriers to cloud storage adoption
  - Reliability concerns
  - Security concerns
- Conventional Solutions
  - Replication, Encryption
- Better Solutions
  - Information Dispersal
  - All-or-Nothing Transform
Why some don’t embrace the cloud

- According to recent business surveys:
  - 83% plan to or already use cloud storage
  - 42% had security and loss of control concerns
  - 24% were concerned about reliability

- In summary:
  - An issue of trust
  - Stems from a perceived loss of control
Some recent comments…

- “I really worry about everything going to the cloud, I think it's going to be horrendous. I think there are going to be a lot of horrible problems in the next five years.”

- “I want to feel that I own things. A lot of people feel, 'Oh, everything is really on my computer,' but I say the more we transfer everything onto the web, onto the cloud, the less we're going to have control over it.”

  -- Steve Wozniak
Are reliability concerns valid?

- “We are very sorry, but ultimately our efforts to manually recover your volume were unsuccessful. The hardware failed in such a way that we could not forensically restore the data.”
  - Amazon Web Services, EBS Support

- “Despite the very best efforts of our engineering and datacenter operations teams, unfortunately the host machine that your cloud server resides on was not able to be recovered. At this point, there is complete data loss on the server.”
  - The Rackspace Cloud
Data is the lifeblood of business

- 40% of all companies that experience a major disaster will go out of business if they cannot gain access to their data within 24 hours.
  - Gartner

- 93% of businesses that lost their data center for 10 days went bankrupt within one year.
  - National Archives & Records Administration
Legitimacy of reliability concerns

- Reliability concerns are legitimate
  - Data is critically important to businesses
    - Data loss jeopardizes an organization’s survival
  - Cloud storage providers are imperfect
    - Multiple instances of loss have already occurred
    - If provider closes down, what happens to your data
Are security concerns valid?

- In 2011, DropBox introduced a bug that left all 25 million customer accounts open to the world
  - Any storage locker could be accessed by entering any password
  - Data was in this state for four hours
The changing flavor of breaches

- 855 breaches, 174 million records stolen in 2012

Who?
- 98% came from external agents (+6% YOY)
- 4% involved internal employees (-13 YOY)

How?
- 81% involved hacking (+31 YOY)
- 10% involved physical attacks (-19% YOY)
Real security or “security placebo”

- On the topic of Amazon’s new server-side-encryption feature, Alan Williamson wrote:
  - This feature, at best, is a security placebo. A security blanket if you will! It protects you against nothing that is probable. If someone gains access to your Amazon access credentials, they can also gain access to your stored S3 data. The biggest omission from this potentially excellent feature is putting the key management into the hands of the customers.

- Real security requires that the storage provider never receive unencrypted information
Legitimacy of security concerns

- Security concerns are legitimate
  - Exposure of data can be damaging and costly
    - Most businesses have to store confidential data: employee information, customer lists, trade secrets
  - Cloud storage providers are imperfect
    - Multiple instances of breaches have occurred
    - If a provider exposes sensitive information, it is the customer who has to deal with the consequences
Solving the reliability problem

- If you can’t trust a cloud provider to maintain data reliably, there are a few solutions:
  - Keep copy of data locally
  - Backup to a second cloud storage provider
- But these approaches are burdensome:
  - Cloud storage was supposed to alleviate the burden of in-house data storage management
  - Storing to a second provider will double costs
Solving the security problem

- If you can’t trust a cloud provider to keep data private, there are a few solutions:
  - Encrypt it and store keys locally
  - Encrypt it and store keys at a second provider
- But these approaches are problematic:
  - Storing keys locally provides no DR
  - Storing keys at a second provider makes it twice as likely there will be data loss
The ideal cloud storage scenario

- Customer is in full control of data
  - Even if cloud provider disappears overnight, customer retains complete access to data
  - Even if cloud provider is completely breached, no customer information is revealed

- Even better: Same cost as existing providers
- This is possible!
Realizing the ideal scenario

- Two things are required:
  - Multiple cloud storage providers
    - Preferably using a common protocol
  - Clever application of two algorithms
    - Information Dispersal
    - All-or-Nothing Transform
An Information Dispersal Algorithm (IDA)
- Receives $K$ inputs, produces $N$ outputs
- Can recover input from any $K$ of the $N$ outputs
- Where $1 \leq K \leq N$

In essence: an IDA is forward error correction combined with a slicing and spreading operation
IDA example: $K = 10, N = 16$
The more cloud storage providers are used, the better the reliability and efficiency become:
- Asymptotically approach overhead of 0
- Become more resilient to provider error
Improved reliability, lower cost

- Example configurations:
  - Using 3 providers, $K = 2$, $N = 3$
    - 50% overhead, and we can tolerate 1 failure
  - Using 10 providers, $K = 8$, $N = 10$
    - 25% overhead, and we can tolerate 2 failures
  - Using 24 providers, $K = 20$, $N = 24$
    - 20% overhead, and we can tolerate 4 failures

- The task of creating such a client would be much easier if these providers were CDMI compliant!
What about security?

- Using so many providers, we become even more vulnerable to security breaches
  - Breach of any provider can expose data
- Assuming average provider suffers a breach only once every 10 years
  - If we use 24 providers, one of our providers will be hacked every 5 months!
So let's encrypt it..

- We could encrypt the data before we process it with the IDA, but there is a better solution.

- The All-or-Nothing transform will offer greater security than encryption, and it is easier to use.
  - Can tolerate many more security breaches.
  - No need to manage encryption keys.
The All-or-Nothing Transform

- Is a mode of encryption in which the decryption key is derived from the entire set of ciphertext
  - Invented by Ron Rivest (the ‘R’ in RSA, and inventor of RC4 and MD5 algorithms)
- By itself, AONT it isn’t very useful:
  - Anyone with the whole ciphertext can decrypt
- Combined with an IDA:
  - Becomes an efficient form of secret sharing
Below: a simple way to make the decryption key derivable only from entire set of encrypted data.
How AONT is used with IDA

- Data transformed first using AONT then IDA
- $K$ inputs are formed by splitting AONT package
- Finally, IDA uses $K$ inputs to produce $N$ outputs
Properties of AONT + IDA

- When AONT applied prior to dispersal, IDA turns All-or-Nothing into a K-or-Nothing Transform
  - With any $K$ outputs, we can recover the AONT package using the IDA, and thus recover data
  - Short of $K$ outputs, we have zero information about either the key or the data
- Using AONT in an 8-of-10 configuration
  - We can tolerate 7 breaches without exposure
    - Regular encryption only tolerates one breach
Ideal system realized?

- The IDA ensures that we can tolerate data loss from multiple providers without losing anything.
- AONT ensures no provider, or group of colluding providers (less than $K$) can see one’s data.

Moreover, the cost of this system is an incremental fraction above using one provider:

- Equal to: $(N / K) \times \text{average cost per GB}$
Where can I get it?

- Gert Liekens, as his master’s thesis has implemented a CDMI client that uses information dispersal, AONT, and multiple providers
- He created it to solve the trust issues users have with putting sensitive data in the cloud
- Java prototype to be available soon
Related work

- **Towards User Centric Data Governance and Control in the Cloud**
  - Stephan Groß and Alexander Schill
    - Created multi-provider storage layer that uses traditional encryption (PKI) to achieve security, implement access controls, and user management.

- **RACS: A Case for Cloud Storage Diversity**
  - Hussam Abu-Libdeh, Lonnie Princehouse, and Hakim Weatherspoon
    - Created IDA-based S3 compatible proxy that maps slices to different cloud storage providers. Performed real-world cost analysis for such a system.

- **A Security and High-Availability Layer for Cloud Storage**
  - Maxim Schnjakin, Rehab Alnemr, and Christoph Meinel
    - Applied RAID technology to distribute data across cloud providers. Found there are over 100 vendors offering cloud storage currently.
If you can’t think of any, here are some samples:

- How expensive is AONT?
- How expensive is the IDA?
- How does recovery work if a provider fails?
- How can XOR securely protect the key?
- Why not use secret sharing to store the key?
References

- http://www.rackspace.com/managed_hosting/services/proservices/disasterrecovery/
- http://www.wired.com/threatlevel/2011/06/dropbox/
- http://gigaom.com/cloud/dropbox-yes-we-were-hacked/
- http://alan.blog-city.com/amazon_s3_encryption.htm