Thesis Research

Distributed/Anonymous Data Backup

Chris Kuske
COMP 599 - Spring 2016
California State University Channel Islands

What will be covered

- Discuss the concept for this thesis
- Review the problems associated with this topic as they exist today
- Discuss potential technologies/methods to solve these topic problems

Thesis Concept

- Thesis summary: Research effective methods for ‘backing up’ critical data in a secure, reliable, and anonymous fashion.
- Model data redundancy requirements (how many nodes does data need to be sent to)
- Optimize backup/recovery speeds by geolocation
- Goal is secure data backup without the risks of data disclosure or data loss.

Problems To Solve

- **Anonymity** - How can a individual's data not be recognized on other machines
- **Security** - How can we ensure the data in only readable by us
- **Data Redundancy** - How can we ensure the data is able to be recovered with little possibility of errors during recovery
- **Ease of Use** - The solution needs to be easy to use

Anonymity

- How can data be safely distributed to ‘untrusted’ nodes in a distributed network

- When presented to the untrusted node, how can we ensure that the data does not contain identifiable data?

- When communicating with the remote nodes, is the communication secure

Security

- Is the data being transferred securely and safely to other nodes on the network? (Integrity)
- Can we prove that the data stored remotely cannot be tampered with or manufactured by a third party. (Non-repudiation)
- How can we prove both integrity and non-repudiation

Data Redundancy

- How can we guarantee that the data stored on the network is sufficiently distributed to allow for nodes being offline?
- Excessive redundancy causes slow backups
- Insufficient redundancy causes data loss

Ease of Use

- The code developed should not require extensive setup for have many software dependencies
- If this software is to be used outside of this research, ease of use is important.

Anonymity Ideas

- Ensure that the data is anonymous as possible (the storage protocol/data format should not have identifying data in it)
- Use hashes rather than file names to store the data on remote nodes so users on the remote nodes cannot tell what data came from where just by looking at the disk.

Security Ideas

- The obvious concept to apply: Encryption
- What encryption algorithms are viable and have long-term potential? (algorithms that haven't had practical or theoretical attacks for instance)
- Use asymmetric encryption schemes such as RSA (remote nodes get data signed with the client node's public key)

Redundancy Ideas

- Divide each file into ‘slices’, distribute those slices among multiple (or many) nodes
- How many nodes must the slices be sent to in order to ensure sufficient redundancy?

Redundancy Ideas

- Develop a mathematical model to determine the optimal ‘slice’ distribution count.
- Geolocation can be used to ensure that data is sent to nodes in geographically diverse areas, which can in turn alter the mathematical model

References

- Distributed Internet Backup System. (n.d.). Retrieved April 10, 2016, from http://web.mit.edu/~emin/Desktop/ref_to_emin/www.old/source_code/dibs/index.html
- CCL Research Publications - Cooperative Computing Lab. (n.d.). Retrieved April 10, 2016, from http://ccl.cse.nd.edu/research/papers/

Thank You!