

Membrane - Distributed File Backup - Literature Review

Dominic Hauton

January 23, 2017

Abstract

An application in which users can join a large public network of untrustworthy intelligent storage agents to trade their available file space. This will allow the user to store files safely on hostile computers, increasing availability and redundancy. The hosts will act as intelligent agents, and trade space when they communicate with one another.

0.1 Introduction

Distributed storage is a well studied and explored domain with clear advantages over bare metal. In order to implement a peer-to-peer distributed agent based storage system we must first decompose the problem into several parts and explore the advances made in those fields.

With the advent of cheap high-speed internet users are now able to use monolithic cloud services to backup data. These tend to be expensive for anything but small amounts of storage and many people have expressed security concerns over holding their data in a data centre owned by another company. Especially if their data leaves the users country where data protection laws may be different.

Over time personal storage capacities for users have also increased. It is now common for computers to come with large amounts of hard drive space which is often not filled to capacity. The project proposed promises to swap this free hard drive space to back up other users' data in exchange for their surplus space to backup your data.

To simplify the process, the system will be able to negotiate contracts of varying complexity for space allocation on another machine, in exchange for space on itself. Unlike most distributed databases and cloud storage solutions frequent down-time will be expected across devices and contracts will have to account for this.

The challenges of such a system include a way of only copying parts of a file during updates to reduce bandwidth usage. An effective way of managing where file chunks are stored. A way to encrypt the files on the remote host and a way to authenticate that the host still holds the file.

Part of the required research also includes exploring the more recent technology of intelligent agents, in which we are most interested

in trust metrics. Intelligent agents will allow the software to make storage decisions based on the trust and reputation of other nodes in the network.

0.2 History of Problem

A simple file backup can be imagined as simply copying a file to another location. In order to keep the duplicated file in sync they must be compared. The program diff solves this through finding the longest common subsequence of bytes between files. In order to improve performance hashing, presorting into equivalence classes, merging by binary search, and dynamic storage allocation are used. [Hunt and MacIlroy, 1976]. This allows the user to view changes and copy the file over again if required.

0.2.1 Rsync

In a networked scenario, bandwidth from source to destination is at a premium. Rsync, introduced in 1996 presents a much better solution through copying changed file chunks (deltas). [Tridgell et al., 1996]. Rsync splits the file into shards and calculates a weak rolling checksum and strong MD4 checksum for each block that allows quick comparisons of shards along the file. When a discrepancy is found, we assume an extra byte or bytes have been added to the file. The weak checksum can be efficiency recalculated for the next offset and once there is a match, it is confirmed with the strong checksum. The new added chunk can now be transmitted. This results in a lot less data being copied than there would be with a diff file. [Tridgell et al., 1996] This combination of weak and strong checksums has been used across multiple distributed systems including low-bandwidth file systems [Muthitacharoen et al., 2001] and high performance

transactional object stores. [Stephen et al., 2000].

Multiround Rsync improves on the rsync algorithm by allowing for more communication to lower bandwidth. Blocks of smaller and smaller sized are used to find holes in the old file in each round and the file until the minimum block size is reached and a copy occurs. [Langford, 2001] This works better than standard rsync in situations where the source file has been changed in many places distributed around the file.

Rsync requires both old and new copies of a file to exist on the host system during an update. This issue has been addressed by creating an in-place rsync (ip-rsync) that uses techniques used in delta compression and Lempel-Ziv compression to move the areas of the file around. In ip-rsync file sender sends add and move commands to the destination in an order that guarantees no files will be overwritten. [Rasch and Burns, 2003]

0.2.2 Git

Git is an improvement on Rsync as it provides both version history and minimises data transfer. To keep storage simple, a copy of the whole file is stored and a reference is put into the version history. By storing old files locally operations are fast. This is also an important distinction from other version control systems and one of the reasons why Git was chosen as an example versioning system compared to other versioning systems like SVN. The systems can continue to operate without a centralised server. To reduce file duplication all files are referenced using their SHA-1 hash. This means you can be sure the contents of the file hasn't changed since you last read it. [Torvalds and Hamano, 2010]

Git also uses a 3 stage workflow. A working directory, where the current files are stored, a staging area and a .git directory. The staging

area prepares your next commit and then it is finally committed. When the staging is complete the change is irreversibly stored. This is a good approach that will be adopted in the final software solution. It will allow incrementally finding changed files, and assessing the need for a new version number to be saved.

0.2.3 Bittorrent

The BitTorrent protocol is a mechanism for sharing files between swarms of hosts. As BitTorrent splits files into parts, users start sharing data even before they have received the full file. Each file has a SHA-1 identifier, similar to Git. [Qiu and Srikant, 2004]

If a user wishes to download a file from the swarm, the user downloads a metadata file from the web and locates users sharing the data using a Tracker Server, Distributed Hash Table (DHT) or Peer Exchange (PEX). [Cohen, 2008]

A Tracker server is a centralised store of all current connected users along with how much of the file they hold. This approach is vulnerable to exploitation by authorities as all of the data about a swarm is stored on a single server and as a result cannot be used for the proposed system.

A DHT contacts other known users for information instead of a centralised server. The Mainline DHT as outlined in BEP No.5 is based on the Kademlia protocol that allows for decentralised peer discovery for a particular piece of content.

PEX is a method for two clients to share a subset of their peer lists. Coupled with DHT, PEX removes a vulnerability from the BitTorrent network by allowing fully distributed bootstrapping, tracking and peer discovery.

A DHT with a form of PEX a tried and tested way of successfully mapping and finding files on a network and will be used within the proposed project.

0.2.4 Resilio

Resilio Sync is an example of a distributed file storage system that utilises the BitTorrent protocol to automatically synchronise folders between a user's systems. It is not a cloud backup solution and not intended as a form of off-site storage. There is no distributed file system and as a result, no redundant data block algorithm adding complexity. [Farina et al., 2014]

As Resilio Sync uses DHT to transfer data, there is no central authority to manage authentication or log data access attempts. This makes it difficult to determine whether a file has been accessed by another user. [Farina et al., 2014] As a result in the project the assumption will be made that everyone in the network has access to all encrypted file chunks. To access and reassemble a file, a user will be required to request all of the file chunks individually and then locally reassemble them.

0.2.5 Storj

Storj is a peer-to-peer cloud storage network which aims to allow users to store files on a decentralised platform. The platform takes special care to provide protection against Sybil attacks and other forms of fraud. [Wilkinson et al., 2014]. To store files it stores encrypted hashed shards on the network. In order to provide proof of storage it uses Merkle Audits and pre-generated audits with hash-challenges to determine whether the client still holds the required data. By adding a seed to the hash-calculation the client can enforce the workers are still in possession of the data. It prevents the client cheating a farmer through using blockchain proof-of-existence to keep both parties honest.

The most efficient form for proof of storage is through using a deterministic heartbeat. Using Feistel permutations data can be verified with $n + 2\sqrt{n}$ challenges. Erasure encoding is

added to shards to detect any minor changes to the data. This is less I/O intensive than a full heartbeat, but still allows an attacker to complete heartbeats with only a data integrity of $1/n$, where n is the number of nodes holding the data.

In order to add extra protection to files, we can use erasure encoding to allow file recovery if one of our shard types is lost. This can be investigated in our software, however, as the shards are expected to change on a regular basis because of versioning, this may not be possible.

To prevent Sybil based attacks, Storj encrypts each shard segment with a different salt. This stops workers completing proof of storage on behalf of another node.

0.3 Peer Admission

The first step in designing the distributed file system is locating other peers within the swarm. This is accomplished through Peer Admission. Once the first peer is found data within the swarm can be located using a DHT which guarantees content can always be found.

0.3.1 Bootstrapping

There are two types of PTP networks which must be examined:

- *Asynchronous*
- *Synchronous*

Within Synchronous networks the number of nodes on the network is constant and all of the nodes are aware of each others existence. This does not allow storage networks to scale but it does allow data to be kept private. [Saxena et al., 2003] This is the simplest and first approach that will be taken in locating nodes within Membrane.

In most current P2P systems such as Gnutella [Klingberg and Manfredi, 2002], Chord and Tapestry as well cryptocurrencies such as in Bitcoin and Litecoin a bootstrapping node is contacted, which provides information about what clients are currently online. Once a bootstrapping node allows the client to find the edge of the swarm, more information can be found using peer exchange.

Within a local network we can also use Universal Plug and Play to find other nodes within the local network. This prevents an external call to a bootstrapping node and as a result is less prone to attack.

Through looking at availability metrics within Bittorrent systems Neglia et al. [2007] determined that both trackers and DHT should be used in creating a highly available distributed storage system such as BitTorrent. DHT tends to be slower at finding new data, however it is much more reliable.

Within Membrane, I plan to use a combination of Asynchronous and Synchronous techniques. Users will try to bootstrap from their last known neighbour nodes on the network, this takes advantage of the static nature of the backup data. Only if this fails, and with the user's permission will they contact a centralised bootstrapping node. This should only happen during a first install and if the user has no referrals. Throughout the lifetime of the application the centralised bootstrapping node would ideally be replaced with a referral system.

A further extension of this, would be to allow hosts to provide a DNS name along with their IP. Users that have setup Dynamic DNS (DDNS) [Bound and Rekhter, 1997] would be able to locate each other without the help of a bootstrapping server.

0.3.2 Peer Exchange

When bootstrapping is complete new Peers can increase their knowledge of the network through Peer Exchange. This is used by BitTorrent to help share swarm information with other nodes. As soon as a client connects to the swarm, peer information is collected using DHT or PEX.

There are two common extension protocols called AZMP and LTEP, which send at most one message a minute when a client leaves or exits the swarm. To reduce congestion at most 50 peers can be added or removed in one PEX message. [Vuze, 2010]

Shard Discovery

Bittorrent also uses the Mainline DHT to find other hosts in the network. This is a Kademlia DHT which now according to Jones [2015], now supports 25M users. It works through assigning each node and file a 160-bit string as an ID. We can work out which node is meant to store a file metadata and crawl in the direction of the node using a hill climb algorithm. Once the metadata is stored on the host, if a host wants to download a file, it can take the metadata on the known host to find the IP of hosts with the file.

Hosts in Membrane will store a metadata file with all of the information required for the specific account to run, including friends, encryption keys as well as local mappings for which host owns which shard and which IP belongs to each host. As a result a DHT will only be required if this initial metadata is lost and needs to be recovered. The proposed version of Membrane will make use of this metadata, but recovery can be added in further iterations of the project.

Dynamic IP Address

Dynamic IP addresses have proven to be problematic in distributed computing, as ISP typically charge more for users to have a static, unchanging IP. Bittorrent tackles this issue through using a DHT to dynamically find the IP address of the user that owns a file. This approach is robust, however, it is a complex solution to IP address resolution.

Another widely used approach is using Dynamic DNS (DynDNS) as described by Bound and Rekhter [1997] in RFC 2136. This allows a client to automatically update a nameserver with a new IP or other information, this allows clients to have a persistent addressing method for devices that change their location. This approach requires initial configuration by the user, however, it provides a reliable way to connect with a user when their IP is lost. There are several tools such as MintDNS, cURL and Iandyn that could be used to ease the development of a built in DynDNS. When setting up a relationship with another host, both an A/AAAA Address and CNAME could be provided, where the CNAME is a backup if the A/AAAA address does not work.

To resolve IP Address resolution within Membrane, I would like to take advantage of small-world networks. [Porter, 2012], in which the mean shortest-path between two nodes increases slowly compared to the number of nodes in a network. Within a group of users in Membrane hosts are likely to share multiple first, second and third degree connections. By storing a list of IP addresses from all of the hosts. It is highly unlikely that all connections within three hops will have changed IP address. We take inspiration from ARP [Plummer, 1982], and send broadcasts with a limited hop count in the network to see if anyone is aware of the current address of a host. The downside of this approach is that it relies on a node within your social network to be online. Measures will need

to be put in place to reduce broadcast spam.

A broadcast storms runaway broadcast events, common in networks that used broadcasting for communication, particularly when areas of the network overlap. [Tseng et al., 2002] These can be mitigated by reducing broadcast traffic, however, Membrane will rely on broadcasts to find hosts. The first step to limiting these broadcasts, is implementing a hop count on broadcasts. This is commonly seen in routing protocols such as IPv6 [Deering, 1998]. The Spanning Tree Protocol (STP) as seen in IEEE 802.1d [Group et al., n.d.; Sharma, Gopalan, Nanda and Chiueh, 2004] provides loop-free routing in LANs by pruning redundant links. Topology changes are dealt with by rebuilding the tree.

Within Membrane rebuilding a Spanning Tree would be an expensive operation. If broadcasts become a problem a 'block request' system can be used, similar to that of ICMP redirects. [Postel et al., 1981] If a node receives a duplicate broadcast message it sends a request back one hop to not send broadcasts from that source toward it for some time. The time limit would allow for corrections if the network topology changes. This preventative approach and could be improved by using the full Spanning Tree implementation. As a further step it could be improved by using a DHT 'closest jump' approach if required.

0.4 Data Allocation on External Nodes

In order to store data on another node Membrane must first have permission to store files on another node. In order to make a choice we must be able to look at trust information about other nodes on the network, and negotiate and trade space once a suitable candidate has been found. These two areas have been explored in the context of Multiagent Systems (MAS) in

the past. [Wooldridge, 2009]

0.4.1 Negotiation

There are three main areas that are important for negotiation. Negotiation protocols, negotiation objects and the node's reasoning models. [Beer et al., 1999] The sophistication of the negotiation is determined by each of these and can take different forms such as auctions, argumentation and protocols in the style of a contract net. The simplest negotiation uses fixed size items, which Membrane shall be initially using for this reason. More complex negotiation allows for counter offers and stronger guarantees.

A simple negotiation protocol issues a call for proposals to a number of nodes and waits for their bids. To formalise this a Agent Communication Language (ACL) such as KQML (Knowledge Query and Manipulation Language) [Finin et al., 1992] or the more modern FIPA (Foundation for Intelligent Physical Agents) [Fipa, 2002]. Beer et al. [1999] tells us that within KQML the agent sending the query is expected to decide how the receiver will handle it, which places limits on negotiation. On the other hand, FIPA is newer and as a result can be more error prone.

0.4.2 Communication

As Membrane is a distributed system communication is key for nodes on the network to interact. It is traditional to form a protocol that agents can use to share information with each other. These protocols are often very rigid and do not allow for expansion. Formalising communication using ACLs is a more flexible approach that aims to let agent share a common understanding of the domain, and allows hosts to reason about communication themselves. Instead of using a strict protocol we shall instead take an agent-based approach to

communication using ACLs.

Agent Communication Language (ACL)

On balance, we shall be using FIPA for communication. This is an ACL based on Speech Act Theory [Labrou et al., 1999]. It splits communication into a communicative act (CA), an ACL message structure and a set of communication protocols such as XML or OWL (Web Ontology Language). In FIPA communication should be rational, in that when sending a message:

- The sender believes the proposition.
- The recipient does not already believe the proposition.
- The recipient will believe the proposition after the proposition.

In the case of Membrane this could be put into the context of asking another node to store data, it would only be reasonable to send another node data to be stored, if the two nodes had negotiated storage of that block between them previously.

To keep communication simple for nodes, we shall use two CAs

- REQUEST
- INFORM

where *REQUEST* expects a reply of some sort and *INFORM* does not. This will enable easier implementation and can be expanded if required.

Ontology

An ontology will need to be created for negotiation. Within MAS ontologies allow agents to share a common understanding of the structure of information, Noy et al. [2001] gives us 5 steps for ontology creation.

1. Determine the domain and scope of the ontology
2. Consider reusing existing ontologies
3. Enumerate important terms
4. Define classes and their hierarchy
5. Define class properties (slots)
6. Define slot facets
7. Create Instance

Using this process a basic ontology for contract negotiation can be created. It is important to ensure the ontology goes into the right amount of detail for the requirements. An overly detailed ontology can make reasoning difficult, and an ontology that is too simple will limit expressiveness [Wooldridge, 2009].

0.4.3 Trust and Reputation

0.4.4 File Transfer Across NAT

Network Address Translation (NAT) is used extensively to solve the IP exhaustion problem, however, it also creates a lot of well documented problems for peer-to-peer communication as hosts can now potentially share IPs. In order to transfer data over NAT a NAT-Traversal Technique must be employed. There are four major techniques described by Ford et al. [2005] that we shall explore.

0.5 Distributed File System

0.6 Authentication

0.7 Untrustworthy Hosts & Encryption

0.8 Conclusion

Bibliography

- Beer, M., D'inverno, M., Luck, M., Jennings, N., Preist, C. and Schroeder, M. [1999], 'Negotiation in multi-agent systems', *The Knowledge Engineering Review* **14**(03), 285–289.
- Bound, J. and Rekhter, Y. [1997], 'Dynamic updates in the domain name system (dns update)'.
- Cohen, B. [2008], 'The bittorrent protocol specification'.
- Deering, S. E. [1998], 'Internet protocol, version 6 (ipv6) specification'.
- Farina, J., Scanlon, M. and Kechadi, M.-T. [2014], 'Bittorrent sync: First impressions and digital forensic implications', *Digital Investigation* **11**, S77–S86.
- Finin, T., Weber, J. et al. [1992], 'Specification of the kqml agent-communication language'.
- Fipa, A. [2002], 'Fipa acl message structure specification', *Foundation for Intelligent Physical Agents*, <http://www.fipa.org/specs/fipa00061/SC00061G.html> (30.6.2004) .
- Ford, B., Srisuresh, P. and Kegel, D. [2005], Peer-to-peer communication across network address translators., in 'USENIX Annual Technical Conference, General Track', pp. 179–192.
- Group, I. . S. W. et al. [n.d.], 'Ieee standard for local and metropolitan area networks: media access control (mac) bridges', *IEEE Std* **802**.
- Hunt, J. W. and MacIlroy, M. [1976], *An algorithm for differential file comparison*, Cite-seer.
- Jones, B. [2015], 'Bittorrent's dht turns 10 years old', *URL* <https://torrentfreak.com/bittorrents-dht-turns-10-years-old-150607/> .
- Klingberg, T. and Manfredi, R. [2002], 'The gnutella protocol specification v0. 6', *Technical specification of the Protocol* .
- Labrou, Y., Finin, T. and Peng, Y. [1999], 'Agent communication languages: The current landscape', *IEEE Intelligent Systems and Their Applications* **14**(2), 45–52.
- Langford, J. [2001], 'Multiround rsync'.
- Muthitacharoen, A., Chen, B. and Mazieres, D. [2001], A low-bandwidth network file system, in 'ACM SIGOPS Operating Systems Review', Vol. 35, ACM, pp. 174–187.
- Neglia, G., Reina, G., Zhang, H., Towsley, D., Venkataramani, A. and Danaher, J. [2007], Availability in bittorrent systems, in 'IEEE INFOCOM 2007-26th IEEE International Conference on Computer Communications', IEEE, pp. 2216–2224.
- Noy, N. F., McGuinness, D. L. et al. [2001], 'Ontology development 101: A guide to creating your first ontology'.
- Plummer, D. [1982], 'Ethernet address resolution protocol: Or converting network proto-

- col addresses to 48. bit ethernet address for transmission on ethernet hardware’.
- Porter, M. A. [2012], ‘Small-world network’, *Scholarpedia* **7**(2), 1739.
- Postel, J. et al. [1981], ‘Rfc 792: Internet control message protocol’, *InterNet Network Working Group* .
- Qiu, D. and Srikant, R. [2004], Modeling and performance analysis of bittorrent-like peer-to-peer networks, *in* ‘ACM SIGCOMM computer communication review’, Vol. 34, ACM, pp. 367–378.
- Rasch, D. and Burns, R. C. [2003], In-place rsync: File synchronization for mobile and wireless devices., *in* ‘USENIX Annual Technical Conference, FREENIX Track’, Vol. 100.
- Saxena, N., Tsudik, G. and Yi, J. H. [2003], Admission control in peer-to-peer: design and performance evaluation, *in* ‘Proceedings of the 1st ACM workshop on Security of ad hoc and sensor networks’, ACM, pp. 104–113.
- Sharma, S., Gopalan, K., Nanda, S. and Chiueh, T.-c. [2004], Viking: A multi-spanning-tree ethernet architecture for metropolitan area and cluster networks, *in* ‘INFOCOM 2004. Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies’, Vol. 4, IEEE, pp. 2283–2294.
- Stephen, Z. H., Blackburn, S. M., Kirby, L. and Zigman, J. [2000], Platypus: Design and implementation of a flexible high performance object store., *in* ‘In Proceedings of the Ninth International Workshop on Persistent Object Systems’, Citeseer.
- Torvalds, L. and Hamano, J. [2010], ‘Git: Fast version control system’, *URL* <http://git-scm.com> .
- Tridgell, A., Mackerras, P. et al. [1996], ‘The rsync algorithm’.
- Tseng, Y.-C., Ni, S.-Y., Chen, Y.-S. and Sheu, J.-P. [2002], ‘The broadcast storm problem in a mobile ad hoc network’, *Wirel. Netw.* **8**(2/3), 153–167.
URL: <http://dx.doi.org/10.1023/A:1013763825347>
- Vuze [2010], ‘Peer exchange’, *URL* http://wiki.vuze.com/w/Peer_Exchange .
- Wilkinson, S., Boshevski, T., Brandoff, J. and Buterin, V. [2014], ‘Storj a peer-to-peer cloud storage network’.
- Wooldridge, M. [2009], *An introduction to multiagent systems*, John Wiley & Sons.