

The Preservation Storage Network

David Tarrant and Ben O'Steen

University of Southampton, Oxford University,
davetaz@ecs.soton.ac.uk, benjamin.osteen@sers.ox.ac.uk

The Preservation Storage Network is a cross platform server and client that enables the construction of a private cloud storage network. The PSN provides a single piece of software which acts much like a BitTorrent client but with data security, self healing abilities and an Amazon S3 API on the front. Taking influences from RAID and recent work on the Sun Honeycomb the PSN software provides a way of constructing an efficient trusted, multi-site, multi-node cloud storage network.

Current cloud storage provides unlimited¹ storage on a cost per Gigabyte basis and while this is ideal for short term and on demand services, using the currently available cloud providers to provide a long-term preservation store is not advisable due to lacking retention policies and legal reasons surrounding the actual location of the data. Handing control of the network over to the institution, or set of institutions, helps avoid such legal and policy based issues.

Upon ingestion of a file into the PSNs Preservation Bucket the PSN will split this file into pieces of data that are then stored on different geographically based nodes around the network. Much like how RAID stores parts of files over many disks the PSN stores data blocks over many PSN nodes. The advantage here however is that the PSN does not rely on proprietary hardware to manage this splitting of the original file and the amount of resilience can be defined on a per object basis.

The lightweight PSN client can be run on any platform and can even be installed on desktop machines that can donate free space to the network. Nodes can be added to the network at any point in order to expand it and advanced Markov based algorithms control any redistribution of data in order to maintain efficient performance.

Finally, while the PSN is designed from the start to be a preservation network, providing reliable and efficient cloud storage for valuable objects, there is a buffer in each node to enhance performance when retrieving files. The PSN keeps track of popular downloads from the network and if possible the most popular items get cached on their home node for improved provision and lower loads on other nodes.

Technology changes fast however always wanting tomorrows solution doesn't help solve today's problem. The PSN software provides a simple yet effective solution that can be deployed on any system, running on any underlying hardware platform, to enable that platform to become part of a much bigger cloud storage system. With the network able to rebuild itself, upgrades to nodes can now simply be done through forced failure meaning you can take advantage of the next generation technologies, without the headaches involved in data migration.

¹ Based on not knowing how much space is left and the assumption that if it gets near full and the company is making a profit they will expand to fill demand.