



Building a scalable shared file infrastructure

Executive summary

Mezeo offers a storage services platform for storage service providers (SSP). As the platform is storage agnostic - the only requirement is a file system that Linux can mount - the SSP is free to choose the appropriate storage. This paper is intended to give service providers an introduction to:

- Cloud storage applications and customer drivers
- Mezeo's storage architecture and options
- Basic shared file storage reference designs

Mezeo supports a wide range of storage options. The common denominator is that the back end provides a Linux-mountable file system. The storage infrastructure behind that file system is an SSP decision. A desirable feature is a scalable file system that enables rapid capacity expansion.

Several high-level scalable shared file reference designs are described with a discussion of the trade-offs.

Introduction

The key to building a successful cloud storage infrastructure is scalability. Cloud storage requires a degree of scalability far greater than what passes for scalability in enterprise storage.

While typical enterprise definitions of scalability range from 2x to 10x, cloud storage scalability should be designed for 100x or even 1000x scalability. Enterprises typically make active use of only 30-40% of capacity so they have a large cushion to handle growth and demand spikes.

For economic reasons Storage Service Providers (SSPs) want higher utilization than the enterprise. And since they have a broad data portfolio any one customer's demand spike will usually be offset by another customer deletions. However, SSPs are more likely than the enterprise to see substantial user growth and accompanying data growth.

An SSP - unlike an enterprise data center - doesn't know if their customer base is going to grow 2x in 3 months or 10x in 12 months. Since each customer represents a profitable revenue stream it pays to be ready.

But not too ready: storage prices typically fall about 5% each quarter. Fast, simple infrastructure expansion means never having to turn a customer down, while keeping costs under control.

Cloud storage applications

Cloud storage (i.e., storage available on the Internet) comprises several market segments.

- **Backup.** The most widely used and requested application for online storage. With a local client that compresses, encrypts and de-duplicates the data, a home user or small-to-medium size business can create an automated disaster-tolerant storage infrastructure that five years ago would have been economically prohibitive.
- **File storage.** For example, in photo storage and sharing numerous files are stored, often in several resolutions, and may be displayed as thumbnails and in different sizes. This model can also apply to academic data, legal discovery files, public documents or anything where a large number of files are shared.
- **Online applications.** The server, software stack, application and storage are managed by the service provider and the entire application service is accessed over the web. Mezeo's architecture makes it easy to add customer facing storage services required for customer applications.

Cloud storage customer drivers

Why are cloud storage services popular?

- **Time.** Cloud storage can typically be provisioned by a customer with a credit card and delivered in minutes, not months. Compared to the time it takes to order storage, take delivery, unbox and install, cloud storage services are quick and easy.
- **Support.** Cloud storage services typically have simple interfaces that are easy to use and support. Unlike debugging newly installed storage on-site, as long as the customer is connected to the Internet they can use the storage.
- **Scalability.** Typically cloud storage can be expanded or contracted in minutes or hours. Excellent for handling temporary spikes or declines in demand.
- **Services.** Cloud storage providers serve many markets. Some serve bulk storage applications using low cost white box infrastructure and proprietary scale-out software to ensure data replication and availability. Other providers provide enterprise customers with enterprise-grade infrastructure that appeals to risk-averse, cost-insensitive customers.

- **Pay for use.** Cloud storage saves both money and time. There is no cash-intensive capital investment and multi-year depreciation. The customer makes the decision and the infrastructure is ready in hours. Typically capacity and bandwidth are billed as used.

The varied applications and high customer expectations place demands on SSPs that enterprise data centers rarely face. It is important to align customer expectations with infrastructure capabilities.

Data growth

Data growth is another special issue for SSPs. Enterprises maintain a large inventory - often 60%-70% of the total - of idle capacity. Fast growing apps are highly visible and extra storage is purchased well in advance.

But SSPs usually maintain a one or a few large pools of storage. All of their growth goes into those pools. And that growth can be substantial.

Take an example of an SSP with 10,000 customers. Let's say each customer has 80 GB of business data in a backup and archiving solution and a data growth rate of 50% per year.

Further, let's say 20% of the customers sign up for the service each year. At the beginning of the first year that SSP will have 160 TB under management. At the end of the 6th year that SSP will have over 9,000 TB to manage.

That is over 50x growth in six years, showing how scalable the storage infrastructure must be.

The components of scalable storage

The major components of any scalable file infrastructure include:

- File system software.
- Hardware to host the file system.

- Front-end client network.
- Back-end storage network.
- Protocol to communicate to the backend storage.

Let's look at each of these components in turn.

File system

There are many different kinds of scalable file systems available today. Likewise there are many words attached to scalable file systems including distributed, global, SAN, cluster and parallel, and many have overlapping meanings. For this paper the focus is on scale-out cluster file systems. Some have global names spaces, some handle parallel traffic and some are used in SANS, but those attributes are secondary.

Some file systems can only be deployed in an integrated package with specific vendor hardware. Others are available on commodity hardware sourced from a choice of vendors. Examples of fully integrated file system solutions include PanFS from Panasas, the file system built into the Permabit product line, the PolyServe cluster in HP's ExDS 9100 and NetApp's Data On Tap GX. Freestanding software implementations include Red Hat's GFS, IBM's GPFS, Symantec's (formerly Veritas) CFS and ParaScale.

For the purposes of a scalable shared file infrastructure there are two key parameters. First, is the file system's meta-data management running symmetrically across all the nodes in the cluster, or asymmetric, where one or

more dedicated metadata managers maintain the file system and it's disk structures?

Second, does the back end storage have the intelligence to manage itself or does it rely on the file system to manage the individual disks? The level of intelligence can vary from none in a JBOD

(Just a Bunch Of Disks) system, to some in a RAID array, to extensive in a commercial NAS system or storage server with a large number of disks.

Hardware

The advantage of a fully integrated solution is that you get the hardware at the same time you get the software. The disadvantage is that these systems tend to be more expensive to acquire, although prices have come down dramatically over the past few years and fully integrated solutions can now be found for less than one dollar a gigabyte.

The advantage of using commodity hardware is that you can shop for the best deal every time you expand your system or replace components. The critical issue is whether you have the time and expertise to handle issues when they come up. It is simpler if a single vendor is responsible for the entire system.

Front-end network

Gigabit Ethernet is the most common front end network because it is inexpensive and reliable. While faster networks can be used, such as InfiniBand or 10GigE, wide-area network latency is usually the limiting factor in online storage applications. It would be an unusual application that would require higher performance than gigabit Ethernet.

Back end network

The back end network offers more choices. Besides gigabit and 10 Gbit Ethernet, InfiniBand, Fibre Channel, and Fibre Channel over Ethernet (FCoE) may make sense. InfiniBand offers very low latency and low-cost high-speed switches and is a good interconnect for both storage and internal cluster communications and asymmetric cluster. FCoE makes the most sense for shops that already use Fibre Channel for their backend storage. The native Fibre Channel interconnect is also fast and

offers low latency, but at a greater cost than InfiniBand .

Back end protocol

The lowest cost way to connect block storage to a global file system is through direct attached storage (DAS). Whether it is a JBOD or a RAID array, DAS is fast and can be very cost effective. The potential downside to DAS is that the file system must manage replication and access across servers.

The iSCSI protocol, which enables block level storage over Ethernet, is growing in popularity because of its low cost and reasonable performance over gigabit Ethernet.

Mezeo's storage architecture

Mezeo offers great flexibility in configuring the underlying storage layer.

Any file system that can be mounted on Linux is supported. This includes Red Hat's GFS, Oracle's Cluster File System and any NFS file server.

Nor do the file systems need to run on the same hardware. One mount point could be backed up by slow bulk storage, while another could use a high-end filer for maximum performance and availability.

With Mezeo a single account is served by a single mount point. A customer may have as many accounts as needed. The ability to run multiple, heterogeneous back ends offers SSPs tremendous flexibility in meeting customer needs.

Storage capabilities are passed through Mezeo to the customer. For example, if the back end storage offers automatic tiering, that feature is passed through to the customer.

All customer data passes through the Mezeo services platform. This enables it to allocate bandwidth and storage capacity on a per-account basis. The Mezeo platform interfaces to SSP

provisioning systems through a standard Service Provisioning Markup Language API.

Reference configurations

Please note that the use of specific company products is not an endorsement of those products for use with Mezeo software. These are examples of possible configurations designed to help service providers understand the breadth of the storage options available today. Consult Mezeo for current information on storage options.

Reference configuration 1

Nexenta offers a software platform based on OpenSolaris and the innovative ZFS file system to provide both scale and exceptional data integrity on low-cost commodity platforms. OpenSolaris is based on the enterprise-grade Solaris OS while ZFS is a modern file system whose 128-bit address space, end-to-end data protection, unlimited snapshots and storage pools reduce management overhead and hidden data corruption.

Nexenta combines OpenSolaris and ZFS with a GNU/Linux userland. Users get a robust Solaris foundation and the option to use well-known Linux programs and management tools. Any storage that will mount on OpenSolaris is usable by NexentaStor.

- Systems: 2 servers running NexentaStor
- Processors: (4) Xeon CPU 8 GB RAM
- Network: Gigabit ethernet
- Storage: any - DAS to SAN - that can mount on OpenSolaris
- Network Switch: Mid-range managed 48-port switches

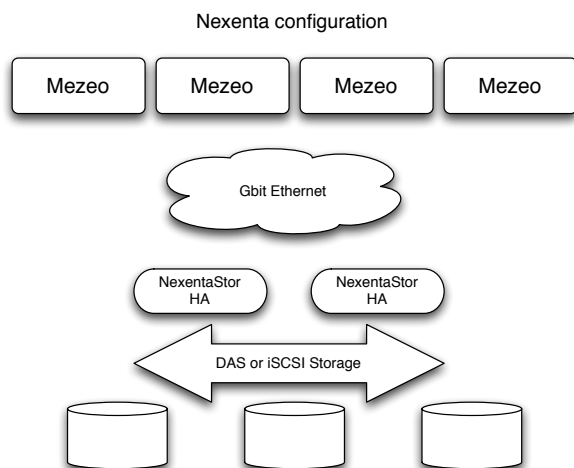
Like Permabit, NexentaStor virtualizes all storage capacity into storage pools. An unlimited number of file systems of unlimited size may be created. For example, it is fast and easy to give every user their own file system. If an account cancels, delete the file system and that capacity returns to the storage pool for re-use.

Reference configuration 2

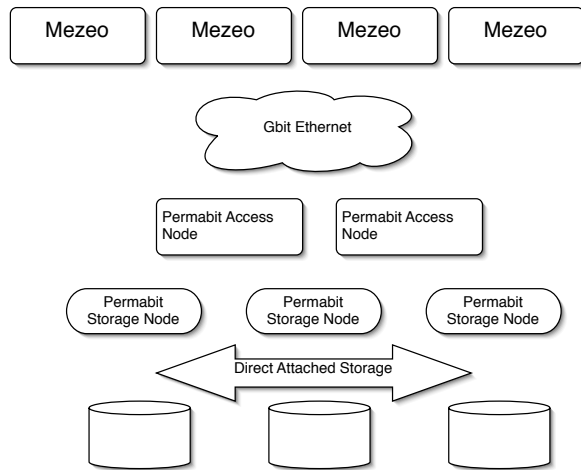
Archive customers are often driven by regulatory of legal considerations that make data preservation of utmost concern. As archives can grow very large, the storage economics are a concern as well.

Permabit offers a solution to this problem. Their system combines high data availability - beyond RAID 6 - with block level deduplication in a platform that can continue working even with multiple simultaneous failures of disks, network links, power supplies and storage nodes.

All data flows through the Permabit access nodes, which enables them to perform block level de-dup and apply advanced erasure coding techniques to ensure data availability.



Highly available archive configuration



- Access: Model 4012 access nodes.
- Storage: multiple Model 4010 storage nodes
- Replication: Model 4012 replication node
- Network: Gigabit ethernet
- Network Switch: Mid-range managed 48-port switches

Permabit virtualizes all storage capacity into one or more storage pools, eliminating LUN management overhead. As new capacity is added it is added to the pool automatically.

The requirement that all data go through an access node limits total bandwidth. I/O intensive apps may not be successful on this platform.

Reference configuration 3

Direct attached storage is the most economical way to connect storage and servers. Large scale out infrastructures often use white box rackmount servers that support 12, 24 or even 48 drives.

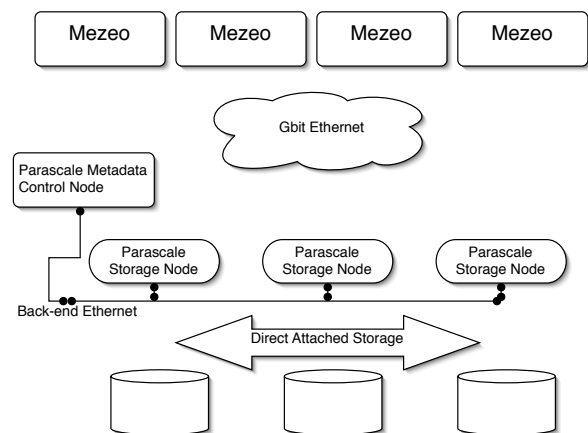
This configuration is powered by a Parascale cluster file system, software that runs on commodity servers running Linux or Windows. Parascale's software creates an NFS-exporting

storage cluster from 3 or more servers with local storage.

Any storage mounted on a server may be part of the cluster. High capacity disk drives mounted in the server are the most economical.

Parascale's asymmetric architecture simplifies management and expansion. File creation and access is controlled by the Metadata Node. When a file is requested, the Metadata Node directs the query to the least busy Storage Node with the data, who serves the data directly to the requesting node. No customer data passes through the Metadata Control Node.

Highly scalable white box configuration



- Systems: white box servers running RHEL 5.2
- Processors: (4) Xeon CPU 4 GB RAM
- Network: Gigabit ethernet
- Storage: multiple Parascale Storage Nodes
- Network Switch: Mid-range managed 48-port switches
- Cluster Interconnect 16-port Gigabit switch

These systems typically offer great flexibility at a low capital cost because the service provider can choose the most cost-effective hardware. Because

of the dedicated metadata server these clusters typically scale to the limits of the metadata server performance.

One downside of these systems is that the integration requires the customer to take extra responsibility for integration of the Parascale software.

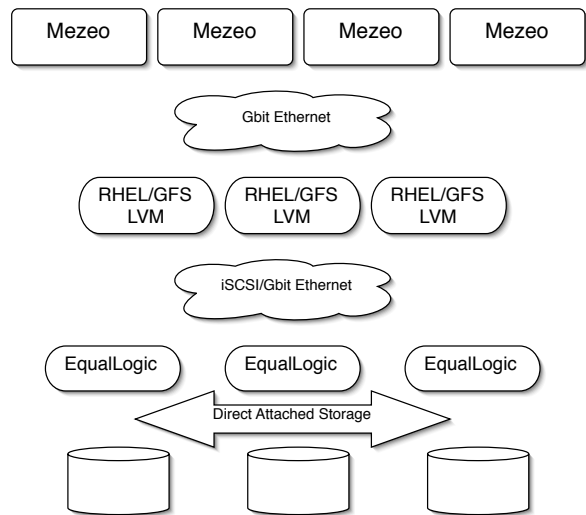
Another issue is that the metadata server works best with files that are 1 MB or larger. File creation overhead is constant, so a workload that creates many large files will have higher bandwidth and greater scalability than a small file workload.

Reference configuration 4

This configuration is an Red Hat Enterprise Linux (RHEL) cluster running their Global File System (GFS) that exports NFS to the Mezeo cluster. The RHEL cluster software provides a clustered Logical Volume Manager (LVM) to manage the back-end block storage.

The backend storage is an EqualLogic iSCSI-based Gbit Ethernet array. The iSCSI arrays offer shared block-based storage to the Linux GFS. The arrays have self-management features that reduce management overhead, including snapshots and high-availability features that enable the SSP to offer enterprise class availability to customers.

Highly scalable enterprise-grade configuration



- Systems: white box servers running RHEL 5.2
- Processors: (4) Xeon CPU 4 GB RAM
- Network: Gigabit ethernet
- Storage: (2) mid-range iSCSI RAID arrays
- Network Switch: Mid-range managed 48-port switches
- Cluster Interconnect 16-port Gigabit switch

Reference configuration 5

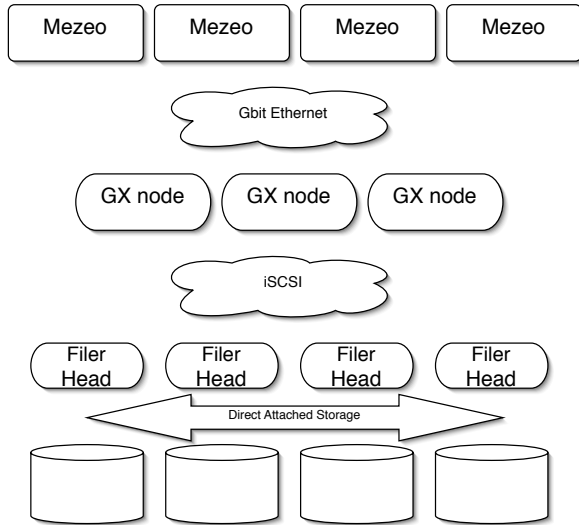
This configuration uses enterprise class hardware with a highly scalable front-end NFS server. The hardware consists of NetApp filers with a Data ON TAP GX cluster front end. The GX front-ends a group of standard NetApp 6000 series filers and provides a global name space and file and directory striping across the back-end NFS filers for maximum performance and enterprise-class availability.

In addition NetApp's added-value software is available to provide differentiating services for the service provider. Low-overhead snapshot copies make it easy to provide clients with older file versions. NetApp's free de-duplication software can

reduce capacity needs by up to 80%. NetApp's dual-parity RAID protects data from two disk failures.

This system's interconnect are multiple Gbit Ethernet links and switches. Redundant switches and links help ensure high network availability.

High-performance, enterprise-grade configuration



- Back end storage systems: NetApp 6000-series filers
- OS: Data ON TAP 7
- Front end cluster: NetApp 6000-series filers
- OS: Data ON TAP GX
- Network: Gigabit ethernet

Conclusion

There are multiple ways to build highly scalable storage for cloud storage applications. They vary in performance, availability, scalability, self-management and, of course, cost. In theory, a Mezeo configuration could incorporate all five of the reference configurations presented in this paper.

The Mezeo platform allows the special features of the storage to be delivered to customers, while giving SSPs a powerful platform on which to build a business. Understanding what storage choices will better meet target market needs is a critical success factor.

SSPs can differentiate their cloud services by careful selection of back end storage systems. The Mezeo platform gives SSPs great flexibility. Understanding how to use that flexibility will be key to growing a successful cloud storage service business.

About The Author

Robin Harris is the founder, editor and senior analyst for StorageMojo.com. He has over 25 years experience in the IT industry in product management and marketing, sales, business development, and strategic planning at companies large and small. He earned degrees from Yale University and the Wharton School of the University of Pennsylvania and lives in northern Arizona. Robin may be reached at robin [at] StorageMojo.com

Copyright ©2004–2009 TechnoQWAN LLC. All rights reserved. Reproduction of this publication without prior written permission is forbidden. All trademarks are the property of their respective companies and owners. TechnoQWAN LLC believes the statements contained in this paper are based on accurate and reliable information. However, due to the technology's velocity of change and the complexity of the applications we cannot warrant that this publication is complete and/or error free. TechnoQWAN LLC disclaims all implied warranties, including warranties of merchantability or fitness for a particular purpose. TechnoQWAN LLC shall have no liability for any direct, incidental, special, or consequential damages or lost profits. The opinions expressed in this paper are those of the author and are subject to change without notice due to new information.