

A Brief by Bill Mottram
Verdictus Associates Inc

Category: Storage Technology
Subject: Pergamum – “New Age” Storage Clustering takes aim at Disk Based Archival Storage.

Disk based archival storage is increasingly becoming identified as a class of storage that breaks the mold of traditional thinking in storage array design. Disk based solutions have already established a position over tape and optical as the choice for active archive data storage and when the design criteria for cost effective disk based-archive storage is met, disk based solutions will increasingly replace tape and optical as the broader archive storage solution of choice. Considering that 70% to 80% of data in a data center is normally unstructured, persistent or archival data and with growing availability or access demands, the market for such a solution is huge.

Clustered disk storage technology is the latest to take aim at solving the problem of storing massive amounts of archive data and replacing tape as the primary archive storage medium. However, to be successful it must meet the principles of being easily scalable in capacity, performance and time, it must deliver on long term system and data reliability, it must be cost effective, be energy efficient and minimize its data center footprint with a high storage density and of course it must be easy to manage.

However today’s historically biased, conventional thinking, dominant within the leading storage vendor community has created disk array designs driven by the need for low latency responses, high IOP’s and optimized to provide access to data at all times. These high level, guiding functional characteristics were established when most disk based data was transactional and by necessity drove expensive architectures, each subject to frequent refresh cycles. While many vendors are positioning traditional disk array’s as archive solutions but using SATA drives rather than FC drives, employing spin down features for power conservation and reducing cost and functionality by limiting cache size and redundancy features the bottom line is that conventional disk solutions do not have the economics (TCO) needed for the long term storage of archive class data. A disk based archival storage solution that will resonate with the end user must be cost comparable with tape, should have low operational costs (energy, space and management), must have a high storage density (>70TB/sq ft raw) and to avoid expensive upgrades and data migration issues it must be easily scalable in performance, capacity and time (evolvable).

Such an approach was discussed in an article published in the latest issue of :login;, the USENIX journal¹. The article is titled “Pergamum; energy efficient archival storage with disk instead of tape” – sound familiar?

¹ “Pergamun; energy-efficient archival storage with disk instead of tape”; Storer, Greenan, Miller, Voruganti; ;login;, Vol 33, Number 3, June 2008

A Brief by Bill Mottram
Verdictus Associates Inc

The authors are a group of researchers from the University of California, Santa Cruz and interestingly enough a technical director from the Advanced Research Group at NetApp. This paper is reminiscent of the paper published by the early MAID innovators² from the University of Colorado, which launched the concept of MAID, later developed and productized by COPAN Systems. Are we having an early look at some future storage technology that NetApp intends to develop? Would make sense.

Pergamum, as it is described in the white paper, is a distributed network of independent, file based, storage appliances which have been labeled as “tomes”. Each tome is a self contained sealed unit, interconnected via inexpensive switches. Presented as very power efficient, each tome draws 1 watt in the spin down state and 13 watts when running, a power requirement within the capabilities of power over Ethernet.

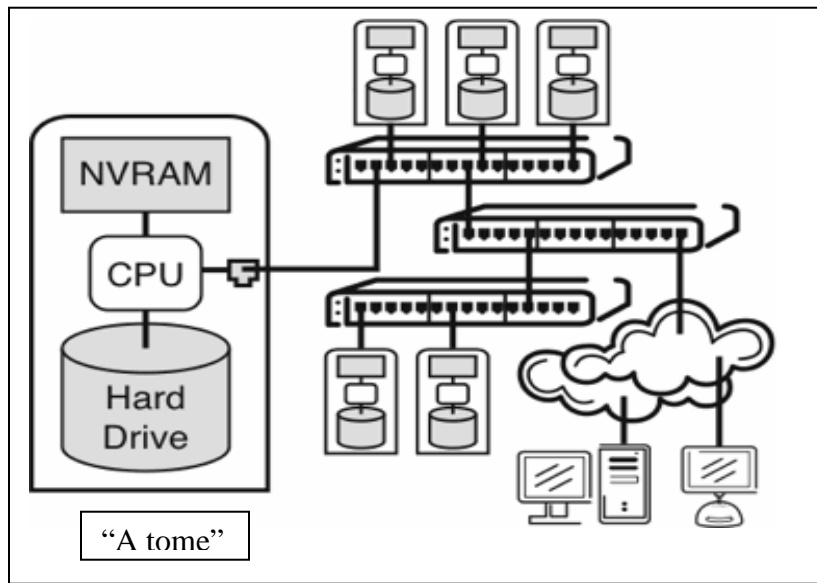


Figure 1: High-level system design of Pergamum. Individual Pergamum tomes, described in Section 3.1 are connected by a commodity network built from off-the-shelf switches.

Tomes are made of four main components:

1. A low power processor
 - a. This is the control and command center for the tome managing all client facing communications, tome to tome communications, metadata management, manages intra and inter disk reliability protocols. etc.
By dedicating a low cost processor to each tome performance and capacity scale concurrently, expectation is that maintenance will be simplified and interestingly overall power consumption is expected to be less.
2. SATA Drive
 - a. Standard SATA drive for persistent data storage. These drives can be spun down ala MAID.
3. NVRAM

² The Case for Massive array of Idle disks (MAID); Colarelli, Grunwald and Neufeld; Dept of Computer Science, University of Colorado, Boulder; January 2002.

A Brief by Bill Mottram
Verdictus Associates Inc

- a. A repository for metadata such as device index, data signatures and pending writes facilitating a response to metadata searches etc without the necessity of spinning up drives.
4. Ethernet controller and Network Port.
 - a. A standard Ethernet interface that should support long term client facing communications while masking any back-end technology evolution from the client community.

To give an idea of configuration topology consider that a 48–port switch could support 46 tomes leaving two ports free for inter-switch communication. Increase the switch count enables the capacity to be increased in one to 46 or tome increments. A 50 switch configuration could support over 2200 drives or tomes (2200 TB). Replicate this configuration and add 10GE inter-switch links and you can see the potential. It would appear that there is still a lot of work to finish and some invention to occur before this technology becomes a viable candidate for commercial exploitation but it is at least focused on the right issues and if NetApp become convinced of its viability they are the guy's to make it happen.

This approach is somewhat similar to the FAB (Federated Array of disks) proposed by the folks from HP Labs back in 2003 and what appears to be the underlying technology for their recently announced HP StorageWorks 9100 Extreme Data System (ExDS) a massively scalable NAS (read archive) product.

So it looks like we have two companies who have declared realistic disk based archive solutions, COPAN with their MAID technology and HP with their soon to be available, ExDS/Polyserve. However I wonder if COPAN has missed their chance. They have been in the market for 3 to 4 years but failed to drive significant traction. One reason could be the lack of a file based client interface which I understand is a failing that will shortly be resolved.

Until I learn otherwise, and whether Pergamum turns out to be commercially viable or not, I will give NetApp the credit for realizing that traditional thinking not going to produce the revolutionary design needed for a cost effective, disk based archival storage solution that can realistically replace tape..

Verdictus Associates Inc

*a pragmatic consultancy focused on data storage technology, management and
the information technology marketing*

email any comments regarding this brief to- bill@verdictusassociates.com