

Optimizing Storage Utilization

January 2008



| Copyright 2008, Exanet Inc.

Executive Summary

Storage management has evolved considerably over the last few years. With increasing demands for online storage, IT decision makers are forced to constantly review their environments, looking for ways to optimize utilization levels.

Most organizations look into increasing the usable disk space in a storage system. However, a broader perspective of a storage solution and a detailed analysis of its components provide a clearer path toward an optimized storage environment.

This whitepaper describes the concept of storage utilization and its many facets, starting with capacity and continuing through issues such as utilization of hardware, power, cooling, and rack space.

Building on a “do more with less” mantra, Exanet is changing the paradigm of storage economics by offering a solution that dramatically optimizes overall utilization levels using off-the-shelf hardware and efficient software.

Understanding Storage Utilization

Most administrators relate the term storage utilization directly to capacity, or more specifically to the ratio between the free and total disk space in a given system.

However capacity is not the only factor - several other aspects affect overall storage utilization:

- Capacity utilization
- Hardware utilization
- Network utilization
- Power, cooling & rack space

Capacity Utilization

The most common aspect of a storage system is disk space utilization. There are two distinct metrics:

- **Used vs. free space** – Comparison between the actual space taken by user data and the total space still available for new data. Most administrators refer to this metric as the overall “storage utilization level.” The configuration of file servers and file systems and the sharp increase in disk capacity all affect this metric.
- **Raw vs. net capacity** – Comparison between the vendor-rated capacity (typically measured as the total number of disks in the storage system multiplied by their rated capacity) and the actual disk space available to the user for writing data. Storage subsystem overheads such as RAID protection and reserved areas affect this metric.

Customers need to consider various aspects of capacity utilization:

Too many file servers and file systems – Most network attached storage (NAS) systems are shipping with 32-bit file systems originating from designs dating as early as the late 1980s. As a result, it is very common to see file systems limited in capacity to 16TB ($2^{32} * 4\text{KB}$ blocks, which looked infinite in 1990). Storage administrators are forced to create and manage an increasing number of separate file systems and implement more file servers as a result. This phenomenon is best described on the street by the saying, “I loved my first filer, I hate my fifth one.”

In addition, administrators find that traditional NAS systems are inadequate for handling data sets reaching hundreds of millions of files within a single file system.

Increasingly large disk drives – As Serial-ATA moves into mainstream application use, disk capacities of 750GB and 1TB have become the new de-facto standard. While offering improved price/performance (especially when compared to Fibre Channel), larger SATA disks dramatically limit the physical configuration of RAID groups and volumes. Configuration best practices recommend larger RAID groups for better utilization (fewer disks are required for RAID protection). However, when reliability is considered, smaller groups are desirable to reduce the risk of concurrent media failures. Traditional NAS systems are therefore deployed with file systems each located on a separate RAID group, reducing the utilization level as described previously.

Disk vendors’ capacity measurement scale – Disk vendors have chosen to advertise disk capacities with a confusing scale, using a base-10. For example, a disk advertised

by the disk vendors as a 500GB disk actually contains only 476.8GB¹. Customers have to adjust their calculations accordingly to get to “real” utilization levels – ones in the more standard base-2 scale.

RAID reserved areas – Storage vendors typically use the RAID subsystem itself to store configuration data and status and monitoring data. This is done by allocating specific sections of either some or all the disks in the system and reserving them for system use only. These areas are completely invisible to the storage administrator and cannot be used to store user data.

RAID protection – RAID technology was invented in the 1980s to protect storage systems from magnetic disk failures by adding redundancy of data. Different RAID levels dictate the amount of redundancy and the overhead that the particular level requires.

The most common RAID levels are described in Table 1.

Table 1 – RAID Levels and Capacity Overhead		
RAID	Explanation	Capacity Overhead
RAID 1	Mirroring	100%
RAID 4/5	Single Parity	For RAID group with 4 disks = 25% For RAID group with 8 disks = 12.5%
RAID 6	Double Parity	For RAID group with 4 disks = 50% For RAID group with 8 disks = 25% For RAID group with 12 disks = 16.6% For RAID group with 16 disks = 12.5%

Some newer vendors are using unique protection schemes whereby parity information is striped across many storage nodes. Some of these schemes can become very inefficient; for example, in one of the solutions the overhead can be as high as 40%. In addition, since a single motherboard, single CPU and RAM are combined with disks, the whole storage node (sometimes referred to as a “brick”) becomes a failure unit. In such systems, a node failure triggers a RAID reconstruct for the total set of disks in the node – typically 10-16 disks – increasing the reconstruct time by 10x-16x times compared with standard RAID architectures. In another case, due to parity scheme optimization for large files, files 128KB or smaller can only be mirrored, decreasing the usable capacity by 100%.

¹ To calculate the actual capacity one has to multiply the advertised size by 1 million and divide by 1,048,576.

Block size – File systems efficiency is directly related to the ratio between the file sizes it stores and its pre-configured block size. For example, a 100TB filesystem with a block size of 96KB used to store 4KB image thumbnails will waste over 95TB of space – a whopping 95% waste.

Hardware Utilization

A storage system is, in essence, a very hardware-oriented solution. Utilization issues surround both the system hardware as a whole, as well as its parts. Many components build up such a system – CPU, memory, NICs, HBAs, FC switches, etc. Each of these can be drastically underutilized.

CPU utilization – In a traditional file server environment, particular applications are bound to particular servers. See Figure 1 for an example. Application App1 is served by file server FS1, and Application App2 is served by file server FS2.

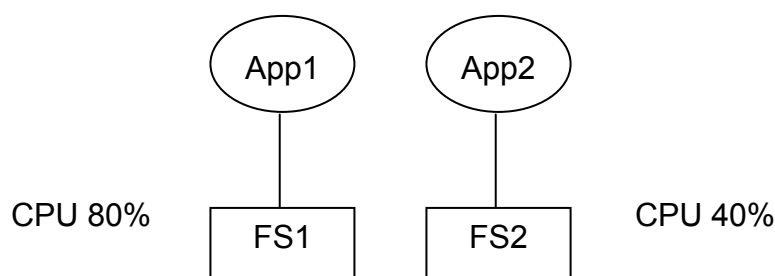


Figure 1 - Unbalanced CPU utilization

This unbalanced environment is causing non-optimal CPU utilization:

- **Over-subscription of resources** - Administrators have to plan for throughput peak times and therefore purchase faster file servers than normally required.
- **Applications are captive to file servers** – Thus, it is impossible to simply move applications between storage servers without actually migrating data. This is both costly and requires downtime.

Cache utilization – Traditional NAS systems are typically designed with a fixed hardware configuration. Read cache is implemented in RAM, and write cache is implemented separately in custom Non-Volatile RAM (NVRAM). The cache memory is therefore both fixed and static and cannot adapt to changing workloads. This issue is

observed during write-intensive I/O patterns, underutilizing read cache memory and at the same time “over-utilizing” the write cache and becoming a performance bottleneck.

Network Utilization

Connectivity is a key element of network attached storage. Both IP-based networks and Fibre Channel fabrics are at play, necessitating review of the utilization of links, ports and switches to shed light on the overall network utilization.

Ethernet link aggregation – Also known as “NIC teaming” in MS-Windows or “bonding” in Linux, port aggregation is a technology devised to transfer data in parallel thru multiple physical links. However the IEEE 802.3ad Link Aggregation² standard dictates that all packets associated with a given conversation are transmitted on the same link to prevent disordering of Ethernet frames. The typical outcome is uneven distribution across the links in the trunk, underutilizing switch ports, cabling and NICs.

Fibre Channel multipathing – Traditional NAS vendors leverage Fibre Channel’s ability to provide redundant connectivity from the file server to the disk enclosures. This enhances the availability of each single server. However, each multipath set is dedicated to a single server, requiring redundant FC links, FC HBA and FC switch ports. In an environment with many file servers, this redundancy comes at a considerable cost, while the overall utilization level of the FC fabric is very low. For example, in one of the traditional NAS systems with four 2Gbps FC ports going from the file server to the disks, the NAS throughput to NFS client is at most 500 MB/sec – merely 50% of the potential backend Fibre Channel link’s capability (4x2Gbps =~1,024 MB/sec).

Power, Cooling & Rack Space

As energy costs reach new heights and data centers are at maximum power consumption levels, the increasing demand for computing power continues. Power consumption, cooling requirements and rack space footprints have become the cornerstone of every IT project, even taking precedence over price and performance. Even Amazon’s CTO admitted in the Next-Generation Data Center 2007 conference in San Francisco that “the efficiency in (our) data centers translates to dollars per quarter.”

² See http://grouper.ieee.org/groups/802/3/hssg/public/apr07/frazier_01_0407.pdf

Today's data center requirements are challenging traditional storage architectures:

No power left – The rise in “pizza box” 1U and blade servers have raised the computing density of standard 19” racks. As a result many data centers have reached maximum power capacity, preventing IT departments from adding any more equipment. Some customers even report that they are over-subscribing racks, literally leaving half of the rack empty due to inadequate power.

Hardware innovation gap – A recent study done by Exanet compared CPU performance with product introduction dates from two proprietary NAS vendors. The study results shown in Figure 2 suggest a 12- to 18-month hardware innovation gap between vendors using standard off-the-shelf servers leveraging the latest available processor technologies and the delay of proprietary designs requiring additional development, integration and testing.

What does this gap mean to overall NAS efficiency? The green computing trend is bringing to the market new technologies that are saving energy and at the same time provide more computing power with a smaller foot print. However, outdated hardware cannot benefit from these improvements as they require more power and cooling as well as consume more floor space.

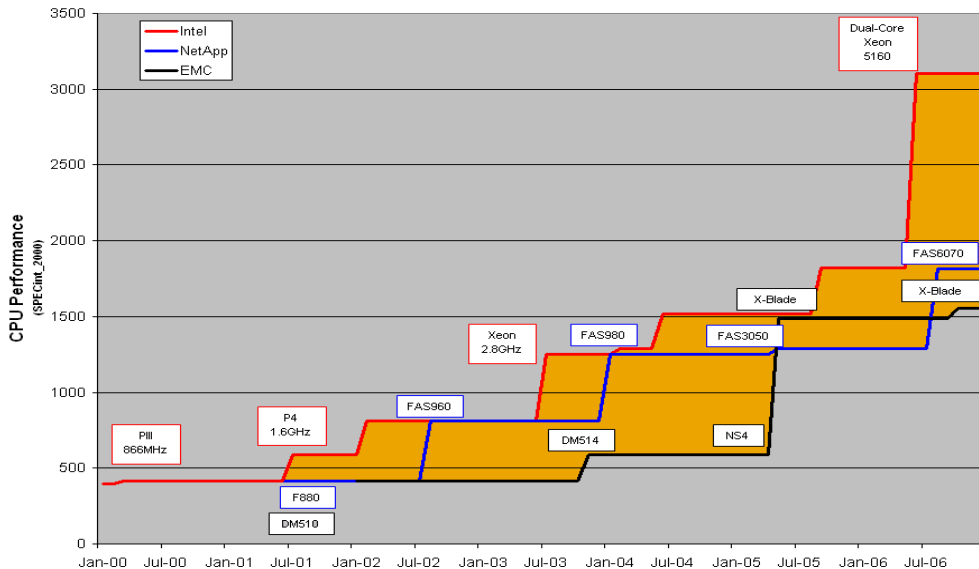


Figure 2 - Processor Innovation Gap Study

Optimizing Storage Utilization

ExaStore®, the industry-leading enterprise network-attached storage software from Exanet, is offering a new architecture based on off-the-shelf server and storage hardware. An innovative design helps customers to increase the overall storage utilization level to over 90%, and reduces power and cooling requirements by over 40%.

File virtualization – Customers that have implemented traditional NAS servers in their environments have found that they have to manage many file servers and file systems, wasting large amounts of free space. ExaStore is a file virtualization “engine” that provides to files what SAN virtualization controllers provide to block storage – a way to transparently consolidate many file servers into to a single file system solution, without impacting user access or limiting performance. See Figure 3 for an example of how ExaStore is used to consolidate 4 filers with 20 file systems, increasing utilization by more than 30%.

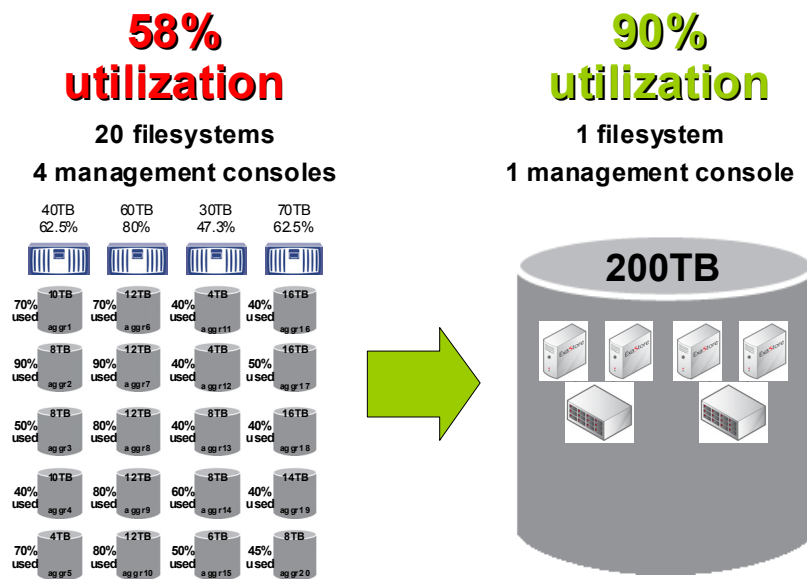


Figure 3 - Consolidate file servers to ExaStore

The ExaFS distributed file system, an integral part of the ExaStore cluster, is automatically striping data across all nodes and disks, thereby fully utilizing any free space (and CPU processing power) available in the LUNs.

Adaptive caching – ExaStore’s cache buffers are allocated dynamically for data depending on actual system workload. When the system receives read requests, it automatically allocates more cache buffers for storing read objects and respectively for writing. Read ahead using prediction algorithms is optimizing the disk I/O. Adaptive caching using standard RAM offer significantly more flexibility and efficiency than proprietary NVRAM-based architectures, since much larger memory portions can be allocated for write cache at peak times.

Striping on all disks – When clients write data to ExaStore, it gets distributed on all disks attached to the storage cluster nodes. An automatic write allocation policy helps ExaStore achieve 100% utilization of the disks available to the system.

Small block size – ExaStore uses a small data block size – 4KB. This enables ExaStore to store efficiently any file size – from small to large, without wasting precious disk space.

Save disk space with snapshots – The ExaCDP high-frequency snapshots feature of ExaStore was designed with efficiency in mind. It uses pointers between shared blocks in the ExaFS filesystem to reduce the amount of space required to store incremental changes. The ExaFS block size is only 4KB, reducing waste even further.

Reduce power with latest processor technologies – Advancements in processor technologies such as multi-core architectures and improvements in the silicon manufacturing process (with 65nm and 45nm) allow for further reduction in power consumption. For example, in a recent benchmark performed by Exanet, a server configured with two Quad-core Intel ® Xeon® processors delivered 30% more throughput than the same server configured with two Dual-core processors. The new server used 40% less kilowatts per hour (kWh) with the more efficient Quad-core technology.

Leverage multi-core microprocessors for distributed I/O processing – The ExaFS architecture is uniquely positioned to leverage from many CPU cores. ExaFS file system is a set of inter-related yet distributed set of processes running on separate cores. The benefits are overall scalability that is linear to the number of cores.

Shrink footprint with high-density packaging – New form-factors for servers and storage allow for more efficient use of rack space. For example, the use of latest processor and bus technologies allows ExaStore to provide up to 4x times the throughput per-node compared to a competing clustered NAS vendor.

High-density storage arrays provide an even more dramatic reduction in footprint. For example, a 4U enclosure containing 48 SATA disks yields a 9 TB per U ratio, while a 3U enclosure containing only 14 SATA disks yields only 3.5 TB per U.

No need for 802.3ad Link Aggregation (i.e. NIC Teaming/bonding) – ExaStore is capable of network failover and load balancing by simply using the ARP protocol. No configuration outside of the ExaStore solution is required.

No need FC multipathing – ExaStore reduces port burn and the number of links by removing the need for multipathing and a fabric altogether. Utilization of links is high since redundancy is node-based, not link-based.

Summary

Utilization of storage solutions goes beyond capacity. An overall approach is required to achieve considerable results, and all components are subject to review. This white paper offers an insight into many of the aspects to an optimization process.

The ExaStore enterprise clustered NAS solution from Exanet offers clear scalability and performance benefits over traditional NAS architectures. There are also significant improvements in utilization of hardware and network infrastructure.

Building on the “do more with less” mantra, Exanet is changing the paradigm of storage economics by offering a solution that dramatically optimizes the overall utilization levels using off-the-shelf hardware and efficient software.

Table 2 provides a summary of storage optimization aspects.

Table 2 – Storage Optimization Aspects

Capacity Utilization		
Issue	Traditional NAS	ExaStore®
Too many file servers and file systems	Outdated file system designs are limiting capacities to 16TB. Storage administrators are forced to create and manage an increasing no. of separate file systems and file servers.	A file virtualization engine that transparently consolidates many file servers into a single file system solution.
Increasingly large disk drives	Conflicts between reliability and utilization force physical configurations with many file systems each located on a separate RAID group, reducing the overall utilization level.	When clients write data to ExaStore, it gets striped on all attached disks. An automatic write allocation policy helps ExaStore achieve 100% utilization of the disks available to the system.
RAID Protection	Fixed or a limited choice of RAID configurations. In some solutions, RAID overhead can be as high as 40%.	Open choice of RAID configuration. Typical configuration with RAID6 will have an overhead of 12.5 to 25%.
Block Size	Some NAS solutions use very large block sizes that yield poor utilization when used to store small files.	ExaStore uses 4KB block size to efficiently store any file size, from small to large, without wasting precious disk space.
Hardware Utilization		
CPU utilization	In a traditional file server environment, particular applications are bound to particular servers, and CPU resources are over-subscribed to support peak throughput periods. This results in an unbalanced environment with low CPU utilization.	The ExaFS architecture is uniquely positioned to leverage from many CPU cores. ExaFS file system is a set of inter-related yet distributed set of processes running on separate cores. The benefits are overall scalability that is linear to the no. of cores.
Cache utilization	In traditional file serves cache memory is both fixed and static and cannot adapt to changing workloads. This issue is observed during write-intensive I/O patterns, underutilizing read cache memory and at the same time "over-utilizing" the write cache and becoming a performance bottleneck.	ExaStore's cache buffers are allocated dynamically for data depending on actual system workload. When the system receives read requests, it automatically allocates more cache buffers for storing read objects and respectively for writing. Read ahead using prediction algorithms is optimizing the disk I/O. Adaptive caching using standard RAM offer significantly more flexibility and efficiency than proprietary NVRAM-based architectures, since much larger memory portions can be allocated for write cache at peak times.
Networking Utilization		
Ethernet link aggregation	Can only be achieved with IEEE 802.3ad standard. The typical outcome is uneven distribution across the links in the trunk, or in other words underutilization of switch ports, cabling and NICs.	Capable of network failover and load balancing by simply using the ARP protocol. Works out of the box without external configuration.
Fibre Channel multipathing	Commonly used in traditional NAS, multipathing is configured in sets of FC links each dedicated to a single server, requiring redundant FC links, FC HBA and FC switch ports. This leads to a very low utilization level of the FC fabric.	ExaStore reduces port burn and the no. of links by removing the need for multipathing and a fabric altogether. Utilization of links is high since redundancy is node based, not link based.

Table 2 – Storage Optimization Aspects – cont.		
Power, Cooling & Rack Space		
Issue	Traditional NAS	ExaStore®
No power left	The rise in computing density has caused many data centers to reach maximum power capacity, preventing IT departments from adding any more equipment.	Reduces power using the latest processor technologies. In a recent benchmark performed by Exanet, a server configured with two Quad-core Intel® Xeon® processors delivered 30% more throughput than the same server configured with two Dual-core processors. The new server used 40% less kilowatts per hour (kWh) with the more efficient Quad-core technology.
Hardware innovation gap	<p>A recent study done by Exanet compared CPU performance with product introduction dates from two proprietary NAS vendors. The study results suggest a 12- to 18-month gap between the NAS vendors and the latest microprocessors available on the market.</p> <p>The outdated NAS hardware cannot benefit from green computing and new efficiency improvements and therefore requires more power and cooling as well as consumes more floor space.</p>	ExaStore is designed to run on 100% commodity hardware with no customization whatsoever. This allows Exanet to use the latest technologies available at any point in time. For example, latest generations of ExaStore support Quad-Core microprocessors, PCI Express bus, RAID 6, 4Gbps FC and SAS disk drives.