A Case for Virtualized Arrays of RAID

André Brinkmann, Kay Salzwedel, Mario Vodisek
University of Paderborn, Germany
Email: brinkman@hni.upb.de, {nkz, vodisek}@upb.de.

Abstract

Redundant arrays of independent disks, also called RAID arrays, have gained a wide popularity in the last twenty years. Most of the disks used in the server market are currently based on RAID technology. The primary reason for introducing RAID technology in 1988 has been the fact that large disk systems have become much slower and more expensive than the connection of a large number of inexpensive disks and the use of them as an array.

The times seem to repeat themselves. Today, large scale RAID arrays have become incredible big and expensive. It seems that it makes sense to replace them by a collection of smaller and inexpensive arrays of JBODs or mid-ranged RAID arrays. In this paper we will show that combining these systems with state-of-the-art virtualization technology can lead to a system that is faster and less expensive than an enterprise storage system, while being as easy to manage and as reliable. Therefore we will outline the most important features of storage management and compare there realization in enterprise class storage systems and in current and future virtualization environments.

I. Introduction

The capacities of single disk systems are growing exponentially since more than 20 years. Nevertheless, the demands for storage capacity are growing even faster. While disk capacities increased from 2 GByte in 1985 to more than 200 GByte today, the demands for capacity have even broken away the one PByte frontier at single installation sites [1]. These installations require even without considering the overhead for redundant data coding techniques and assuming the largest available disks the coordination of more than 5,000 disks.

A first step to enable the management of such a large number of disks has been taken in 1988 by Patterson, Gibson, and Katz [10]. They introduced five different RAID levels to overcome the limitations that have been imposed by a single, large, and expensive disk. These single disks did not only lack capacity scaling requirements but also were nor able to deliver the required I/O-rate and bandwidth (see table I). The combination of a large number of inexpensive disks with the technology to evenly distribute the data and redundancy blocks among these disks enabled the storage industry to seek new frontiers.

Today, RAID has become a multi-billion dollar industry. Most of the disks sold for the use in servers are build into one of the many forms of RAID arrays. The RAID technology can be implemented in software, in the host bus adapter (HBA) of the server, or inside dedicated hardware arrays. The costs for RAID technology vary from the free usage of

TABLE I. Comparison of a large disk with an array of small disks in 1988 [10].

	IBM 3390K	*IBM 3.5" 0061*	*x70*	*Factor*
Capacity	20 GByte	320 MByte	23 GByte	
Volume	97 ft^2	0.1 ft^2	11 ft^2	9x
Power	3 KW	11 W	1 KW	3x
Data rate	15 MB/s	1.5 MB/s	120 MB/s	8x
I/O-rate	600 I/Os/s	55 I/Os/s	3900 I/Os/s	6x
MTTF	250 KHrs	50 KHrs		
Costs	$250,000	$2,000	$150,000	

open source software solutions up to million dollar enterprise storage systems, like the IBM ESS 800.

The benefit of enterprise storage systems compared to standard RAID systems are their

- reliability,
- scalability,
- performance, and
- manageability.

Enterprise storage systems do not only enable the administrator to scale beyond 100 TByte in a single system, but also provide him with the necessary administration tools and can guarantee him a nearly unlimited reliability. Low-end and mid-ranged RAID systems have not been able to deliver similar features to the customer on their own for a long time.

Nevertheless, the situation is quite similar to 1988. Enterprise class storage systems have become incredible big and expensive. They force customers into a vendor-locking, taking from them the freedom and flexibility to choose between different technologies when expanding their storage infrastructure. And similar to the introduction of RAID in 1988, the introduction of network based storage virtualization combined with products supporting the *Storage Management Initiative-Specification SMI-S* [11] enables the customers to shift from a single and expensive RAID array to a large number of mid-ranged RAID arrays that are able to deliver an even better performance, scalability, and reliability for a better price.

In this paper we will show that the limitations of low-end and mid-ranged storage systems can be overcome by coupling many of them with the help of an open system virtualization solution. This coupling does not only impose a new flexibility in combining storage systems and breaks off the traditional vendor locking, but is also able to decrease the initial hardware costs while keeping an optimal scalability, performance, and manageability. To compare enterprise class storage systems with an array of simple disks, also called *Just-a-Bunch-Of-Disks or JBODs*, and mid-ranged RAID arrays, we have chosen publicly available information about different storage systems from Dell EMC, HP, and IBM. These disk systems are similar in the network interconnection and disk technology to each other, but differ in properties like caching and software support. The information can only give a snapshot about the relation between the different storage systems at the time of the writing of this paper. Nevertheless we assume that comparing different classes of storage at different times will lead to similar results.

In section II we will compare the initial hardware costs for JBODs, mid-ranged RAID arrays, and enterprise class storage systems to get an impression for the potential hardware cost savings. These cost savings can only be beneficial to the customer if the resulting virtualized storage systems can guarantee the same performance, reliability, and manageability as enterprise class storage systems. These aspect are covered in section III, IV, and V, showing the potential of open systems virtualization, but also outlining drawbacks of current solutions and areas for future research.

II. Financial Aspects of Storage Virtualization

In this section we give a cost comparison between the hardware investment into an enterprise class storage system and the hardware investment into the corresponding number of JBODs and mid-ranged RAID arrays.

Before the introduction of network based storage virtualization, low-end and mid-ranged RAID arrays lacked the ability to scale to more than a few TByte. A consolidated management of a large number of these arrays has not been possible; features like snapshot, remote-copy, or mirroring over arrays of different vendors have not been offered to the customer. Storage virtualization now gives a new freedom to the customer. One is able to combine storage systems from different vendors inside a single management environment and he can rely on enterprise class functionalities. Virtualization environments enable the customer to scale up to hundreds of TByte without being forced to invest into the technology of a single vendor. The customer can scale his environment by purchasing the storage that exactly fits his demands.

Table II summarizes the costs for a JBOD, different mid-ranged RAID arrays and the IBM ESS 800 enterprise class storage array [8]. The prices have been taken from the internet presentation of the manufacturers or their distribution channels between February and September 2004. The prices are

TABLE II. Price comparison between a JBOD, mid-ranged RAID arrays and an enterprise class storage system.

	transtec 3000 JBOD	*transtec 6600 RAID-System*	*Dell EMC CX200*	*IBM FAStT 200*	*IBM ESS 800*
Capacity	2.1 TB	2.3 TB	2.1 TB	9.6 TB	55.9 TB
FC ports	1	4	2	2	16
Cache Size / GByte	–	2	1	0.25	32 (max. 64)
Disks	15	16	15	66	380
Processors	–	2	2	2	Dual Cluster SMP
Price	$7,750	$18,250	$18,999	$220,660	$1,848,350
Price per GB	**$3.52**	**$7.75**	**$8.84**	**$22.45**	**$32,29**

calculated for the designated configuration shown in the table. These configurations do not always represent the maximum expansion of the storage systems. Furthermore, we have chosen not to include the costs for the infrastructure, like storage switches or SAN appliances, and the software costs. Storage switches are used to connect servers with JBODs or simple RAID arrays, but also to connect the servers with enterprise storage systems to decrease the failure probability. The number and costs for SAN appliances that are used in virtualization solutions depend on the virtualization technique and vary from 2 simple off-the-shelf computers that act as meta-data servers up to clusters of workstations. The costs for virtualization software are comparable to the costs for software features in enterprise storage systems.

Comparing the hardware costs for scaling them to a size of 55.9 TB, the prices range from 201,500 $ for the JBOD, 493,974 $ for an DELL/EMC CX200 based system, and up to 1,850,000 $ for the IBM ESS 800 (see also table III). Up to now, this price difference has been paid by the customer due to the lacking management capabilities and the lower reliability of mid-ranged storage arrays. In the following sections we will show that the introduction of network based storage virtualization enables the customer to overcome these limitations of JBODs and mid-ranged RAID arrays and enhance them beyond the capabilities of enterprise class storage systems for a much lower costs.

III. Flexibility, Manageability, and Scalability of Large Scale Storage Environments

In this section we will discuss the technological differences between enterprise class storage systems and tightly coupled JBODs or RAID arrays by using a storage virtualization solution. The comparison will focus on scalability and flexibility, cost aspects and the ability to satisfy future demands.

For today's companies it is quite common to store an amount of data ranging from a few TByte up to several PByte. Usually, such data volume is organized within a homogeneous SAN environment. Up to now, the main drawback of heterogeneous storage environments has been that the customer has to cope with non-uniform interfaces or protocols that can lead to integration problems and hence, rising costs. According to a recent study by Hitachi Data Systems (HDS), 55 % of the IT managers identify uniform storage management of heterogeneous systems as an important future topic [6]. Therefore, enterprises currently tend to build their storage infrastructures by solutions from a single vendor to fulfill their needs of uniformity and manageability.

A. Enterprise Storage Systems

At first glance, it seems sufficient for a customer to invest into a high end enterprise storage cabinet from a vendor like EMC, HDS, HP, or IBM. These vendors offer highly integrated solutions with a huge amount of storage space and additional state of the art management technologies, e.g. different *RAID levels* and virtualization inside the cabinets, *Snapshot*, and *Remote Copy*.

The technological advantage of these solutions is based on the optimization capabilities of the vendor who is aware of all technological components and their interactions and is able to bundle its storage system with additional management technologies as mentioned above. The close coupling of all relevant methods and features within either one or a few devices connected by proprietary interfaces can be of great benefit to a customer in terms of stability and reliability as well as by appropriate installation know how and support.

Therefore enterprise storage systems seem to be sufficient for many companies which do not want to cope with enhanced storage adminstration including the handling of a heterogenous storage environment. Furthermore, cabinets offer enough storage capacity for most applications and make the investment expensive, but calculable.

The main advantages of enterprise storage systems arise from the fact that the main part of administrating both the devices and the data takes place directly inside the storage systems. This is considered today as being safer and easier to handle than a heterogeneous storage environment but hinders the customer from changing their storage politics or upgrade to new storage technologies.

This change in storage politics may become necessary, if

- the exponentially growing amount of data even outgrowths the biggest available storage systems,
- the customer wants to keep the ability to react on significant changes of the integrated technology or include technological innovations, which are not offered from his storage vendor, or
- the customer wants to integrate additional disk subsystems from competing vendors, which offer a better price/performance or price/capacity ratio.

Generally, a company should be given the freedom of choice in designing its storage infrastructure not only according to its present but also to its future demands. This will lead to an easy and almost unlimited *infrastructure scalability* and a high degree of *technological flexibility* that correspond to the customer needs and should become a main topic for an enterprise to be able to react properly to rising costs and technological progress.

B. Open Systems Virtualization

A key demand for the acceptance of storage virtualization is the way of how the storage systems are administrated. Besides installation and further system costs the administration costs have crucial impact on the TCO [5]. This shift is due to increasing complexity of storage management that growth exponentially in the size of the storage environments.

Heterogeneous SAN environments, build from simple JBODs or mid-ranged RAID arrays, always possessed the capability to store huge amounts of data. Their drawback has been the overwhelming management burden that has been imposed by the lack of management software that spanned more than a single cabinet. This missing management software did not only include advanced features like snapshot, remote copy, or the support for RAID sets that span multiple cabinets, but also a unified software interface to administrate storage cabinets from different vendors.

These drawbacks are overcome by the introduction of open systems storage virtualization and SAN management environments.

- Open systems storage virtualization as an abstraction layer between the representation of the storage system and their physical implementation enables the customer to couple multiple heterogeneous storage systems and to become independent from specific device vendors, interfaces, and connectivity protocols.
- SAN management environments enable the customer to administrate their storage area network from a single console. This management includes the administration of storage switches, the partitioning of the storage subsystems, and the LUN masking. The development of SAN management environments is leveraged by the introduction of the *Storage Management Initiative Specification, SMI-S* that has been defined by the Storage Network Industry Association SNIA. SMI-S defines interfaces that enable the centralized management of different storage devices under global control and has already been integrated by many storage hardware and software manufacturers like IBM, Brocade, SUN, HDS, and VERITAS.

When comparing today's network based storage virtualization solutions we can recognize that virtualization solutions have evolved from a simple pooling of storage subsystems to complete storage management environments. Management techniques like snapshot, remote copy or the support for different RAID levels have shifted as core components into most available virtualization solutions, thus, being decoupled from a certain hardware cabinet into a hardware independent software layer ensuring the customers freedom of architectural design. Therefore storage virtualization is able to keep enterprises immune from becoming captive to a specific vendor or a proprietary technology.

Virtualization solutions have reached a maturity concerning their feature list that easily outperforms high end storage systems. These features are not limited to expensive storage devices but can also be used for simpler storage devices, like JBODs or entry-level RAID arrays. Furthermore, these different classes of storage devices inside a single environment can be managed under a single management console.

This feature list is coupled with a nearly unlimited scalability and therefore heterogeneous storage from different vendors can be coupled to form very large storage pools, delivering the storage capacity for the next decade.

The main advantages of storage virtualization are the following:

- Virtualization can span disk subsystems from different vendors and enables the customer to purchase the storage systems with the best price/performance or price/capacity ratio.
- Virtualization can easily cross location boundaries.
- Virtualization enables unlimited scalability of the capacity over the boundaries of single cabinets or storage clusters.
- Features like mirroring, snapshots, or replication can be applied to simple storage devices like JBODs or simple RAID arrays.

IV. Reliability

In section III we have shown that storage virtualization is able to deliver the same or even better functionality than incorporated in enterprise storage systems. These functionality is not only available in a proprietary but also in an open environment and strongly simplifies the management of enterprise wide storage environments.

Besides scalability and functionality, reliability and availability are two other strengths of enterprise storage systems. All critical components inside an enterprise storage system are fully redundant, starting from the power supply up to the internal data pathes. Often, these components are available more than twice inside the system, even increasing the reliability of them.

The disk failure probability of a JBOD or a simple RAID array is equal to the failure probability inside an enterprise class RAID array, but the internal architecture of a JBOD or even of a mid-ranged RAID system does not contain their sophisticated and expensive redundancy. All JBODs and mid-ranged storage systems in this study contain at least redundant power supplies and fans but can not cope with arbitrary failures of their internal data pathes. This includes broken links between their single subsystems. These kind of failures appear seldom in productive environments but nevertheless makes them more vulnerable against critical failures then enterprise storage systems.

This drawback of arrays of JBODs or mid-ranged RAID systems can be overcome by a simple trick introduces in [10]. By encoding the data according to one RAID level, each data block can be distributed about multiple storage units. This data encoding can be done in software or hardware. If one of the storage units fails, the data still can be accessed and reconstructed.

These encoding schemes can be used in two ways:

1) Each storage unit is used as a JBOD. All redundancy is generated in the virtualization environment.
2) Each storage unit is a simple or mid-ranged RAID array. The virtualization environment puts another level of redundancy on top of these RAID arrays to ensure a protection against the failure of one of the storage enclosures.

The first case ensures the best use of the available storage capacity by only using one level of redundancy inside the storage environment. The drawback of this approach is that the failure of a single disk can put the whole virtualization environment in a degraded mode. This overall available performance in this reconstruction phase is strongly degraded the and assuming systems with hundreds of disks, single disk failures may occur quite often.

In the second scheme, the recovery from single

TABLE III. Performance and cost parameters scaled for 55.9 TByte capacity.

	transtec 3000 JBOD	*transtec 6600 RAID system*	*Dell EMC CX200*	*IBM FAStT 200*	*IBM ESS 800*
Arrays	26	25	26	6	1
Raw Capacity	55.9	57.5	55.9	57.6	55.9
FC ports	26	100	52	12	16
Cache Size / GByte	–	50	26	1,5	32 (max. 64)
Disks	390	400	390	396	380
Processors	–	50	52	12	Dual Cluster SMP
Total costs	$201,500	$456,250	$493,974	$1,323,960	$1,848,350

disk failures is handled by the hardware RAID units. Therefore, a single disk failure has only a local impact on the performance.

The used scheme depends on the performance and reliability demands of the applications running on top of the storage systems. In many cases, even the use of JBODs as foundation of a large scale storage environment seems to be sufficient. Applications with higher reliability and availability demands will put an additional RAID level on top of the hardware RAID inside the storage units.

The performance of the system furthermore depends on the data encoding. A parity RAID technique is able to minimize the additional capacity wastage but will have a direct impact on the write performance. Using data mirroring as additional RAID level imposes nearly no additional performance penalty and can even increase the performance in many cases, but needs the double data capacity.

In most cases, data mirroring on the virtualization side will be the first choice. Taking into account that most high availability applications running on enterprise class storage systems are using mirroring about two (similar) machines, there will occur no additional capacity waste.

In all cases, standard RAID encoding schemes recover from failures by regenerating the lost data from the remaining disks inside the RAID group. This recovered data is saved on a single spare disk in a sequential process. If a complete subsystem fails in the virtualized environment, this recovery may take a long time. Here, it is sought for new encoding schemes which are able to decrease this recovery time up to a theoretical minimum to further increase the benefits of virtualization in high availability environments.

V. Performance comparison

The performance of a fibre channel based SAN architecture compared to traditional direct attached storage architectures has been one of the main reasons for the successful development of storage area networks. This performance benefit of a SAN is the key to many commercial applications, like streaming media and database applications.

Enterprise class storage systems contain a number of techniques to enhance the performance benefits like aggressive caching, a high fibre channel port number, and an optimized spindle utilization of the included hard disks. The caching techniques help to reduce the number of accesses to the hard disks and the high port density enables the systems to deliver a high peak performance for streaming media access. One of the most sophisticated features of enterprise storage systems is their ability to seamlessly adapt to changing access patterns and to a changing number of used disks by dynamically replacing data blocks [12].

As shown in table III, the accumulated fibre channel port number of a large number of mid-ranged storage systems can deliver an even better peak performance than a single high end storage system and the accumulated cache size is even bigger than the cache size of an ESS 800. Nevertheless there remains the question whether the virtualization environment is able to utilize the given peak performance and caching size of the storage systems.

The evaluation of a number of commercial available open systems storage virtualization systems has shown that most of them lack the ability to adapt to a changing access pattern or require a manual relayout of the data blocks, which is very time consuming and error prone [3].

Not considering the ability to adapt to changing access patterns can significantly decrease the possible performance of a distributed storage environment. Both, caching and port density are only helpful, if the load is evenly distributed about all units in the system. If hot spots are mapped on a limited number of units, only the cache and fibre channel ports of these units can contribute to the overall performance. In contrast, enterprise storage systems are (in principal) able to make the total cache size and port number available to all disks.

To overcome these limitations of commercial virtualization solutions, new data distribution techniques have been developed which are able to evenly balance the load about a heterogeneous SAN [4], [7], [9] and which already have been integrated in first solutions [2].

In the following subsection we will give an example for the relevance of performance optimization inside a storage virtualization environment.

A. Optimization of databases

In this subsection we introduce an example how important performance optimization can become. The performance of our example database architecture depends on

- the setup of SQL queries and indices and
- the layout of database tables concerning the underlying storage system.

The first point mainly depends on the database designer and cannot be optimized by a storage virtualization but the second point can. Conventionally, the table layout needs to take the underlying physical properties of the storage infrastructure into account to achieve a good performance. This is only possible if the database designer and the storage system architect work together. The setup of a database with a performance optimized virtualization architecture is much easier and it enables the designer to put the necessary number of storage devices into a storage pool and use the load balancing feature of the virtualization technology to achieve the best possible performance.

Furthermore, the load balancing features ensure an easier administration of databases in case of a changing data volume. This is, quantitative as well as qualitative, measurable. The following table assumes a database with a data volume between 8 TB. In

TABLE IV. Cost Comparison concerning database table layout maintenance for an initial database size of 8 TByte.

	Manual table layout	*Optimized virtualization solution*
Setup	10,000 EUR	2,000 EUR
Annual Maintenance	90,000 EUR	2,000 EUR
3 year total	100,000 EUR	8,000 EUR

case that no performance optimization is used on the virtualization layer, the initial table layout takes at least 10 days for the database expert and the storage virtualization expert. The expected costs are about 10,000 Euros. This time and therefore the costs can be reduced to 2 days by applying a performance optimized virtualization software.

Even more critical are the costs for the maintenance of the database. Assuming an annual growth of about 50%, both experts are required for about 15 days after each year, so the personal maintenance costs are estimated to 30,000 Euros after the first year. These costs scale with each further used gigabyte of database capacity.

Table IV summarizes the 3 years costs associated with the maintenance of the table layout for a database with an initial size of 8 TByte. Comparing these cost savings with the costs for purchasing a virtualization solution does not seem to be very impressive. This changes if one takes into account that the complexity and therefore the costs for maintaining a table layout increases exponentially with the size of the database.

Based on the the results from table IV and including the exponential growth of complexity, equation 1 describes the cost development for maintaining an optimal table layout depending on the size of the underlying database:

$$\text{Maintainace costs} = K \cdot c^{\frac{\text{Database Size}}{\text{8TByte}} - 1}, \qquad (1)$$

where in the following the constant K has been chosen to 100,000 EUR and the constant c has been chosen to 1.2.

Figure 1 shows the influence of this exponential cost development on different database sizes. Considering that many commercial applications, like MP3-databases, already contain more than one PByte of data, the requirement for automated performance optimization becomes even more obvious.

VI. Conclusions

In this paper it has been shown that open systems virtualization environments have evolved to an interesting alternative to enterprise storage systems. In terms of scalability and functionality, they do already outperform expensive high-end RAID environments.

Drawbacks of commercially available virtualization environments coupled with JBODs or midranged RAID arrays are still their performance in dynamically changing environments and the reliability and availability concerning failures of whole storage units. These drawbacks can be overcome by the integration of new adaptive data distribution schemes inside the virtualization environment, as being done in the storage virtualization environment V:DRIVE [2].

Acknowledgements

We would like to thank Richard Diehl and Jürgen Zender from the CONET AG for helpful and stimulating discussions. Furthermore we would like to thank the anonymous reviewer for the helpful suggestions.

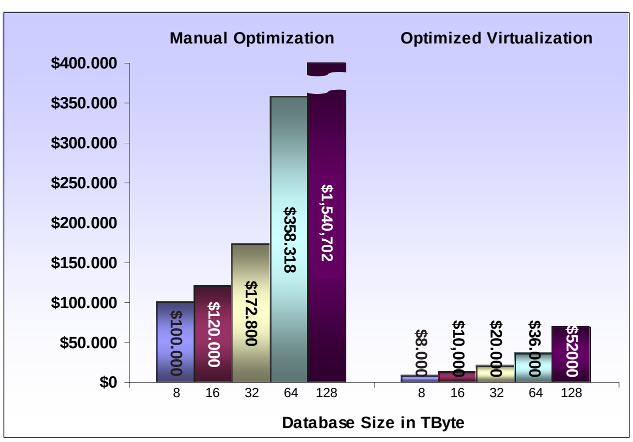

Fig. 1. Cost development for maintaining different sized database environments.

References

[1] O. Bärring, B. Couturier, J.-D. Durand, E. Knezo, and S. Ponce. Storage Resource Sharing wit CASTOR. In *Proceedings of the 12th NASA Goddard, 21st IEEE Conference on Mass Storage Systems and Technologies*, pages 345–359, Apr. 2004.

[2] A. Brinkmann, M. Heidebuer, F. Meyer auf der Heide, U. Rückert, K. Salzwedel, and M. Vodisek. V:Drive - Costs and Benefits of an Out-of-Band Storage Virtualization System. In *Proceedings of the 12th NASA Goddard, 21st IEEE Conference on Mass Storage Systems and Technologies (MSST)*, pages 153 – 157, College Park, Maryland, USA, 13 - 16 Apr. 2004.

[3] A. Brinkmann, M. Heidebuer, K. Salzwedel, and M. Vodisek. Dynamic data distribution schemes. Technical report, University of Paderborn, Sept. 2004.

[4] A. Brinkmann, K. Salzwedel, and C. Scheideler. Compact, Adaptive Placement Schemes for Non-Uniform Distribution Requirements. In *Proceedings of the 14th ACM Symposium on Parallel Algorithms and Architectures (SPAA)*, pages 53–62, Aug. 2002.

[5] G. A. Gibson and R. V. Meter. Network attached storage architecture. *Comm. of the ACM*, 43(11):37–45, Nov. 2000.

[6] HDS. Storage Index Spring/Summer 2004 - Independent IT director research on the EMEA data storage market. Technical Report, 2004.

[7] R. J. Honicky and E. L. Miller. A Fast Algorithm for Online Placement and Reorganization of Replicated Data. In *Proceedings of the 17th International Parallel and Distributed Processing Symposium (IPDPS 2003)*, Apr. 2003.

[8] IBM. Enterprise-strength storage for today's on demand environments. Technical report, IBM, 2004.

[9] D. Karger, E. Lehman, F. Leighton, M. Levine, D. Lewin, and R. Panigrahy. Consistent hashing and random trees: Distributed caching protocols for relieving hot spots on the World Wide Web. In *Proceedings of the 29th ACM Symposium on Theory of Computing (STOC)*, pages 654–663, 4–6 May 1997.

[10] D. A. Patterson, G. Gibson, and R. H. Katz. A Case for Redundant Arrays of Inexpensive Disks (RAID). In *Proceedings of the 1988 ACM Conference on Management of Data*, pages 109–116, 1988.

[11] SNIA. SMI-S Brings Compelling Value To Storage Management: The Opportunity for Vendors to Take Action Has Arrived. Technical report, Storage Networking Industry Association SNIA, Sept. 2003.

[12] J. Wilkes, R. Golding, C. Staelin, and T. Sullivan. The HP AutoRAID hierarchical storage system. In *Proceedings of the 15th ACM Symposium on Operating System Principles*, Dec. 1995.