Alcom-FT Technical Report Series
ALCOMFT-TR-03-117

In Proceedings of the SSGRR 2003s, July 28 - August 3, 2003, L'Aquila, Italy

Storage Management as Means to Cope with Exponential Information Growth

A. Brinkmann, F. Meyer auf der Heide‡, K. Salzwedel‡,
C. Scheideler*, M. Vodisek‡, and U. Rückert

***Abstract*— The advances in Internet technology have led to tremendous improvements in business, education, and science and have changed the way we think, live, and communicate. Information exchange has become ubiquitous by the possibilities offered through modern technologies. We are able to offer information 24 hours a day through our web sites and can leave messages every time and from anywhere in the world.**

This change in communication has led to new challenges. Enterprises have to deal with an information amount that doubles every year. The technological foundation to cope with this information explosion is given by Storage Area Networks (SANs), which are able to connect a great number of storage systems over a fast interconnection network. However, to be able to use the benefits of a SAN, an easy-to-use and efficient management support has to be given to the storage administrator. In this paper, we will suggest new storage management concepts and we will introduce a new management environment that is able to significantly reduce management costs and increases the performance and resource utilization of the given SAN infrastructure.

***Keywords*— Storage Area Networks, Storage Management, Virtualization, PReSto**

I. Introduction

DATA has become the central asset for companies and organizations. This is not only true for companies of the new economy but also for most traditional enterprises. Many business processes became too complex to be managed by employees without the support of modern information technology, the design of new automobiles or plants is even unthinkable without computers. The quality of the data availability and the access speed have got a direct impact on the success of a company. Today, it is easier to cope with the loss of an important employee than to cope with the loss of a data center, which can even cause the downfall of a company.

According to this development, the strategy to manage and to store data is of great importance for companies and organizations. The hard- and software of the storage infrastructure have to ensure a fast and safe access to the corporate data. Due to the exponential increase of information, this task can only be accomplished, if the management of the storage infrastructure can be centralized and highly automated.

A first step in this direction is the consolidation of the storage systems that will be connected by a dedicated storage network that is called Storage Area Network or SAN (see Fig. 1). The connection between the servers and storage systems takes place via the high speed interconnects of the SAN. This makes it possible to disband the traditional tight coupling of the servers and storage systems and enables a real any-to-any communication inside the SAN.

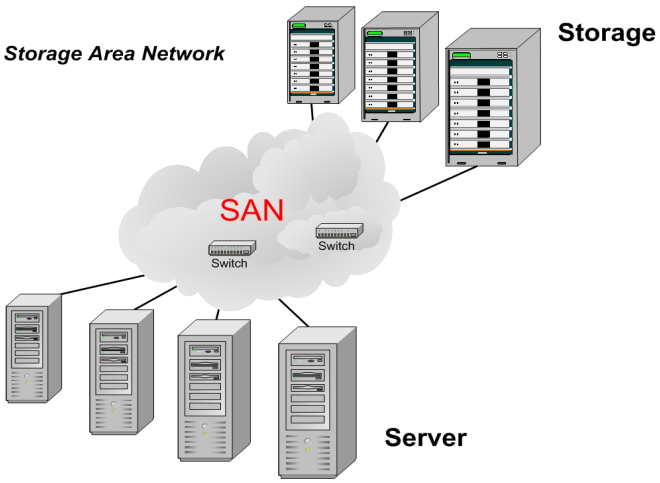

Fig. 1 Storage Area Network

To use the full potential of a SAN, it is necessary to integrate storage management solutions into the storage concept. Even after the consolidation of storage inside a SAN, the enormous growth of data and the heterogeneity of storage systems and operating systems makes it very difficult to efficiently manage the complexity without software support. The SAN management contains different tasks like the administration of switches, remote and local backup of data, and the partitioning of storage systems.

An often underestimated task is the distribution of data among the connected storage subsystems. The conventional approach is to tightly couple one or more logical partitions of a disk or raid system with a file system. As we will show in the following chapter, this strategy has serious drawbacks concerning the efficient use of the storage systems. Neither the capacity nor the bandwidth of the storage systems can be utilized in an adequate fashion. Furthermore, the

Faculty of Computer Science, Electrical Engineering and Mathematics, University of Paderborn, Germany (email: {brinkman, rueckert}@hni.upb.de, {madh, nkz, vodisek}@upb.de)

‡Partially supported by the Future and Emerging Technologies programme of the EU under contract number IST-1999-14186 (ALCOM-FT)

*Department of Computer Science, Johns Hopkins University, Baltimore, USA. (email: scheideler@cs.jhu.edu)

administration of this tightly coupled system is error-prone and limits the scalability of the storage infrastructure.

The efficient use of a storage system can be significantly enhanced by the integration of a virtualization solution. This software makes a transformation between a logical address space that is presented to the servers and the access to the physical disks. It is possible to use the available physical capacity of the storage systems as one or more big storage pools. Virtual disks are built from these storage pools without worrying about the limitations of the underlying hardware.

Storage virtualization can be used in various ways, e.g. to extend the capacity and performance of a file system. An immediate consequence of the use of a storage virtualization solution is a much better utilization of the available disk space. Further improvements are: reduction of administration costs, support for heterogeneous storage devices, better adaptivity, and higher availability of the storage systems.

Our contribution in this paper will be the presentation of the PReSto storage management software. The core component of PReSto is a virtualization engine that uses a patent-pending data distribution technology, the so called Share-strategy [2]. Advantages of the Share-strategy are an improved utilization of the available bandwidth and a much better protection against the failure of subsystems inside the Storage Area Network.

The main focus of this paper will be on virtualization technology as the core technology for storage management. The paper will be structured as follows. In the next chapter we will present different approaches to the virtualization of storage area networks. These approaches differ mainly in respect to the support for heterogeneity and the location of the virtualization software. After this, we will take a closer look on the main component of a virtualization engine, the distribution of data among the different storage devices. In the last chapter, we will introduce the PReSto storage management environment.

II. VIRTUALIZATION TECHNOLOGY

Storage virtualization is often seen as the key technology in the area of storage management. But what actually is storage virtualization? A very good definition is given by the Storage Networking Industry Association SNIA [8]:

"[Storage Virtualization is] an abstraction of storage that separates the host view [from the] storage system implementation"

This abstraction includes the physical location of a data block as well as the path from the host to the storage subsystem through the Storage Area Network. Therefore, it is not necessary that the administrator of a SAN is aware of the distribution of data elements among the connected storage systems. Generally, the administrator only creates a virtual disk and assigns it to a pool of storage systems. Then, a file system or a database can be mounted on this virtual disk and the virtualization software provides a consistent allocation of data elements on the storage systems. It is even possible that a large number of virtual disks share a common storage pool.

The use of a virtualization environment has many advantages compared to the traditional approach of assigning an address space to a fixed partition. An obvious advantage is that a virtual disk can become much larger than the size of a single disk or even than a single RAID-system [5]. When using virtualization software, the size of a virtual disk is only limited by the restrictions of the operating system and the total amount of available disk space.

An important feature of virtualization software is a much better utilization of disk space. Without the help of virtualization software, the size of a file system has to be estimated and allocated in advance. If a file system reaches its limit, it is nearly impossible to enhance its size. It has been shown that in the traditional storage model only 50% of the available disk space is used. The disk utilization can be increased to 80% through using the central and more flexible administration of virtualization software [3]. Thus, the required storage capacity and, with it, the hardware costs of a Storage Area Network can be significantly reduced.

Virtualization software offers new degrees of flexibility. Storage systems can be added to or removed from storage pools without downtime, thus enabling a fast adaptation to new requirements. These storage systems do not have to be from a single vendor, so that the traditional vendor-locking of customers can be avoided.

Virtualization can be done at different places in the SAN; it can be implemented in the connected servers, in the storage subsystems or distributed above the network. In the following we will present the different approaches:

A. Host-based Virtualization

The simplest form of virtualization is the host-based virtualization, also known as logical volume management. The virtualization software offers available partitions which can be connected to virtual volumes.

The flexibility of a logical volume manager, or short *lvm*, is limited. An lvm normally does not allow to share partitions between different virtual disks and the information about the volume management is kept inside the server and is not shared with other hosts. Therefore it is difficult to keep a consistent image of bigger SANs with many servers.

B. Virtualization inside the Storage Subsystems

Many high-end RAID-systems integrate virtualization software inside their storage cabinet. The virtualization of the disks inside the storage cabinet has got a similar drawback like the host-based virtualization. The information about the virtualization is not shared with arbitrary other storage systems and the virtualization is therefore limited to the virtualization of a single cabinet or to the virtualization of storage systems of a single vendor.

C. Network-based Virtualization

The most flexible virtualization solution is the network-

based virtualization. It is able to virtualize all connected storage device, independently of their size, speed, and manufacturer. Furthermore the network-based virtualization is able to coordinate the access of an arbitrary number of servers to the storage devices and therefore eliminates the disadvantages of the host-based virtualization and the virtualization inside the storage subsystems. The customer is neither bound to one manufacture nor has he got to manually coordinate the access of the hosts to the storage devices. Using a network-based virtualization technology, the customer is able to build a cost-efficient storage infrastructure that is "easy" to administrate.

The network-based virtualization is subdivided in two different categories, the in-band virtualization and the out-of-band virtualization. Both methods require the use of one or more SAN-appliances to ensure a consistent view of the Storage Area Network. These SAN-appliances usually are off-the-shelf computer systems with a dedicated software running on top of a standard operating system like Linux or Microsoft Windows.

In-band and out-of-band virtualization differ in the use and placement of the SAN-appliances.

1) In-Band Virtualization

In case of the in-band virtualization, the SAN-appliances are placed between the hosts and the Storage Area Network. The SAN-appliances appear to the hosts as storage systems and the hosts directly address read and write commands to the appliances. Then the appliances transform the requests and forward them to the real storage subsystems.

The advantage of this solution is the independence of the host operating-systems. This is especially important in a mixed environment with many different operating systems and in environments with not so widespread operating systems.

The drawback of an in-band virtualization is the poor scalability. Because all hosts are connected to at least one SAN-appliance, the SAN-appliances can become a bottleneck of the Storage Area Network. They have got to transform each request of the hosts into a request to a storage subsystem and they have got to transfer the bulk data from the storage subsystems to the hosts. If the number of hosts or the number of storage systems increases, also the number of SAN-appliances has to increase to ensure an efficient use of the available bandwidth and to ensure a small latency of disk requests.

Furthermore, the SAN-appliances can be outdated very fast. They have to be expanded or exchanged with nearly every new server and network technology to keep track with ever growing demands.

2) Out-of-Band Virtualization

In an out-of-band implementation, the SAN-appliance is only used to collect and distribute metadata about the Storage Area Network. The communication between the SAN-appliance and the hosts can be done via a local network and has not to take place via the SAN. Therefore even in case of a failure inside the SAN, the appliance can reconfigure the connected hosts.

The virtualization itself takes place inside the hosts. A driver module is integrated inside each server that translates the requests to virtual disks into requests to physical storage devices. The performance of the SAN directly scales with the number and the capabilities of the connected servers. The performance of the SAN does not depend anymore on the performance of the SAN-appliance.

The integration of an out-of-band virtualization is less costly than the integration of an in-band solution. Two SAN-appliances are sufficient to ensure a proper and fail-safe functionality of the SAN. Due to the fact that the appliances are not in the data path, they can be much less expensive than in-band appliances. Furthermore the number of SAN-appliances has not to scale with the size of the Storage Area Network.

A disadvantage of an out-of-band solution is that its integration requires the provision of a driver module for every used operating system. Therefore it is not possible to virtualize a whole Storage Area Network if just one host operating system is not supported. It is important for a customer to be aware of the ability of a vendor to support all necessary operating systems.

An elegant way to provide this is to mix the advantages of in-band and out-of-band virtualization. This can be done by providing out-of-band drivers for the most commonly used operating systems *and* an in-band-appliance that can virtualize hosts with other operating systems.

D. Limitations of the virtualization technology

Virtualization software is a powerful tool to enable an efficient use of the resources in a Storage Area Network. Using the wrong virtualization technology can lead into performance, cost, and management problems that cannot be resolved by other software management layers. Anywhere it is important to embed the virtualization solution into a management strategy, that considers all aspects of storage management. This strategy has to start with the incorporation and partitioning of new storage systems, the detection and administration of switches, and does not end with the backup process. Nevertheless, virtualization technology is the ideal starting point to take a first step to this holistic Storage Area Management *SAM*.

III. THE CORE VIRTUALIZATION ENGINE

A. Preliminaries

The global data distribution in a storage area network has to ensure scalability, data reliability, and fast data access. This is a complex task if the system is *heterogeneous*, i.e. the disks have different capacities, the resources are limited, and it should adapt to changing user demands.

The efficient use of the storage capacity is supported in quite a simple way by traditional virtualization solutions: The available disks are just concatenated in the address space. This

solution has a serious drawback: only the number of data elements stored on a disk and not the number of requests to that disk are evenly distributed (see Fig. 2). Therefore it is likely that hot spots appear on some disks. These hot spots limit the scalability of the Storage Area Network, because the latency strongly increases with the number of storage systems inside the SAN. Other conventional approaches, like striping the data over the disks (similar to RAID 0), have the disadvantage that it is nearly impossible to increase the number of disks participating in a stripe.

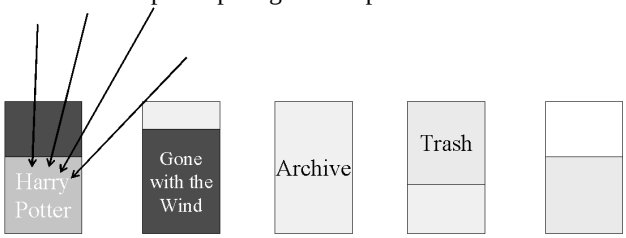

Fig. 2 Data distribution in conventional virtualization environments.

A fast data access can be implemented by distributing the data elements evenly over all participating disks with a pseudo random hash function. In such a distribution the probability of hot spots is minimized. Furthermore, all disks are equally utilized and the data access are served by different disks resulting in a parallel, and therefore, faster answering of data requests (see e.g. [6]). In a heterogeneous system, the disks have different capacities and hence, need to get different loads. In general, the larger disks are faster (because they are newer) and more load mapped onto them does not result in a bottleneck.

We call a data distribution *adaptive* if the volume of data that needs to be redistributed to ensure a distribution according to the heterogeneous capacity distribution is minimized. Adaptivity is measured by *competitive analysis*. We compare an algorithm that has to adapt to new situations (e.g. new disks enter the system) with an (possibly not known) algorithm that ensures the same properties but knows all changes in advance. A strategy is said to be c-competitive if it is only a factor of c worse then such an algorithm. Furthermore, we say that a strategy is *faithful* if it ensures a distribution according to the capacity requirements of all disks after any change in the system parameters and it is *compact* if the needed resources (space) to access any given data element is small, e.g. depend only logarithmically on m, the number of data elements currently stored in the system. Let n denote the number participating disks.

B. Previous Data Distributions

Compact, adaptive placement strategies are relatively new. So far, only good strategies are known for *uniform* settings, i.e. all disks have the same capacity. In this case, it only remains to cope with situations in which new disks enter or old disks leave the system. Karger et al. [4] present an adaptive hashing strategy that is faithful and *2*-competitive. In addition, the computation of the position of a data element takes only an expected constant number of steps. However, their data structures need at least $n \log^2 n$ bits to ensure that with high probability the distribution of the data elements does not deviate by more than a constant factor from the desired distribution. Brinkmann et al. [1] present an alternative placement strategy for uniform capacities. Their scheme requires $\mathbf{O}(n \log n)$ bits and $\mathbf{O}(\log n)$ steps to evaluate the position of a data element. Furthermore, it keeps the deviation from the demanded number of data elements on a disk extremely small with high probability: if m fulfills $m \gg n \ln n$, then the maximum number of data elements per disk is bounded by $m/n + \mathbf{O}(\sqrt{(m \ln n/n)})$, w.h.p. (The scheme in [4] only achieves $\mathbf{O}(m/n)$ with high probability if $\mathbf{O}(n \log^2 n)$ bits are used, even if $m \gg n$.) Another adaptive placement strategy was proposed by Sanders [7]. He considers the case that disks fail and suggests to use a set of forwarding hash functions h_1, h_2, ..., h_n where at the time h_i is set up, only disks that are intact at that time are included in its range. From his description it seems that this strategy can cope reasonably well with failed disks, but it runs into problems when the number of disks grows.

C. The Share-Strategy

The *Share* strategy is a randomized data distribution strategy that works in two phases. In the first phase, the algorithm reduces the heterogeneity problem to a uniform one. This operation will be performed whenever a data element is accessed. The result is a number of disk drives which are equally likely to store the requested data element. In the second phase, we use any faithful distribution strategy (see the previous section) to map the data element to one of the suitable disks.

The reduction phase is based on two hash functions h: $\{1, ..., M\} \rightarrow [0,1)$ and g: $\{1, ..., N\} \rightarrow [0,1)$ where M is the maximal number of data elements in the system and N is the maximal number of disk that are allowed to participate, respectively. In general, those number may be unrestricted but fixing them to an arbitrary large number simplifies the analysis significantly. Furthermore, any practical system can only use a limited representation for N and M.

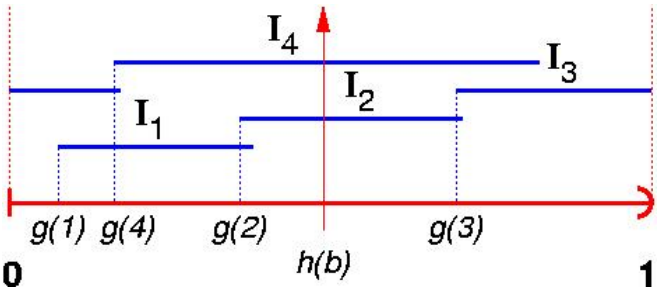

Fig. 3 Hashing scheme in Share.

The reduction phase works as follows: First, the data elements are hashed into an [0,1) interval using h where the quality of h ensures an even distribution of all elements over

the whole interval. Then we map the starting points of sub-intervals of certain length into the same interval using the second hash function g. The length of these sub-intervals I_i corresponds to the capacity of a disk i. To ensure that the whole interval is covered by at least one sub-interval we need to stretch each of the sub-intervals by an factor s. In other words, the sub-interval I_i starts at $g(i)$ and ends at $(g(i) + s \cdot d_i) \bmod 1$, where d_i is the normalized capacity of disk i (d_i is defined by the ratio of the disk capacity and the sum of all the capacities of disks currently in the system). Now, the data elements can be accessed by calculating its hash value and then deriving all sub-intervals that value falls into. Any faithful uniform strategy can be applied to get the correct disk out of the number of possible candidates. It can be shown that *Share* is $(2 + \varepsilon)$-competitive for arbitrary small ε to any change in the system parameters (see [2] for more detail).

IV. PReSto

PReSto is a storage management solution that is based on a powerful virtualization engine. The core component of this virtualization engine is the Share-strategy, guaranteeing an optimal distribution of data over all storage systems connected to the SAN. The virtualization engine can be supplemented by additional tools, like snapshot technology, synchronous and asynchronous remote copy, or backup support tools.

The modular design of PReSto supports both out-of-band and in-band virtualization. Currently we directly support different Linux versions with an out-of-band solution. Further out-of-band solutions are planned for the most common server operating systems, like Windows NT, Windows 2000 and XP, Sun Solaris, and IBM AIX. Hence, a great part of the hosts can directly access data without being forced to take the detour over an in-band appliance. Furthermore, not directly supported servers and network attached storage *(NAS)* devices can participate to the virtualization by an additional in-band solution. This mixed assembly enables to take part on the advantages of both worlds, the very good scalability with low costs of an out-of-band solution with the unlimited support for each operating systems of an in-band solution.

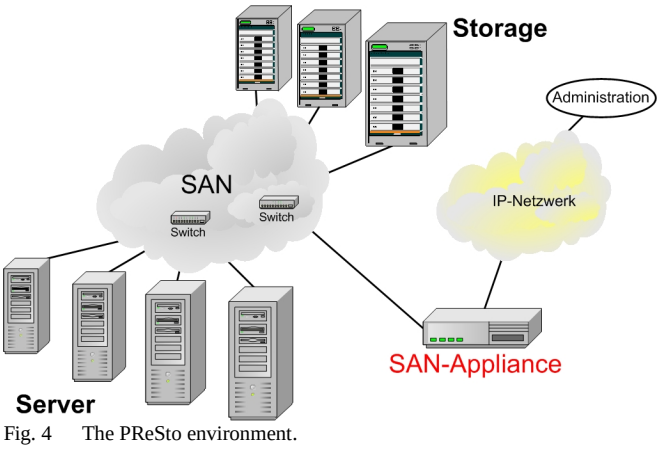

Fig. 4 The PReSto environment.

PReSto has neither restrictions to the used storage systems nor to the underlying infrastructure. Hence, the infrastructure of a SAN can be a fibre channel based network as well as an IP network. By the design of our solution we are given the capability to appropriately support today-in-use fibre channel infrastructure as well as future systems basing on the iSCSI standard and to combine both under the cover of our storage management software.

Information about the construction of our storage network, the so called *meta-data*, are administrated by at least one dedicated SAN appliance, storing state information within an internal database. The number of these control instances does not rely on the size of the network, but on the desired degree of reliability. The data path, i.e. the way the data takes for reading and writing data elements between the hosts and the storage systems, is independent of the SAN appliances. Hence, the number of appliances does not effect the performance of the storage network. Communication between a server and the SAN appliances is necessary only when the number of connected storage media changes, i.e. meta data about the new state of the network has to be exchanged.

If a server uses one of the directly supported operating systems, the virtualization functionality will be integrated within the server by an appropriate device driver. Other servers will be connected to a in-band appliance performing the virtualization for the server.

The administration of PReSto is handled by an easy to use GUI built as java application. This enables the administration of the Storage Area Network from any point in the world via internet.

The administrator is given the ability to cluster the connected storage systems in storage pools that can be combined according to their age, speed, or protection against failures. Each storage pool has its specific storage policy that describes individual aspects like its logical and physical block size or its software protections against failures of individual disks. The assignment of different storage policies to different storage pools creates the possibility to appropriately react to the different requirements concerning safety and speed of an application. Hence, PReSto supports the individual aspects of different applications in an enterprise, like data warehousing and multimedia applications within one storage solution.

The capacity of each storage pool can be changed by simply adding or removing storage systems without system downtime. The data blocks are replaced once according to the Share-strategy. The resulting data placement is optimal concerning the distribution of blocks and the expected distribution of requests to the storage systems, independently of the access pattern of the hosts. The changes are completely transparent to the hosts; hence, they can continue their tasks undisturbed.

The access of a server to the storage system does not occur in a direct way, it will be masked by the concept of virtual disks. A virtual disk will be allocated by the GUI (Fig. 3) and is then being assigned explicitly to one storage pool. Each storage pool can incorporate millions of virtual disks and the

virtual size of any virtual disk can be bigger than the size of the storage pool. Therefore, it is possible to create a file system on a virtual disk whose capacity is beyond the available storage within the data center. The required disk capacity will be allocated on demand. To avoid changes of the file system size in case of changes of the real storage capacity, it is possible to set the initial size of a file system to the maximum possible capacity that is supported by the file system.

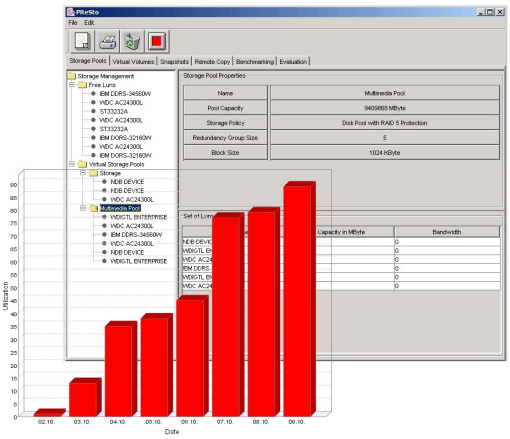

Fig. 5 Graphical User Interface of PReSto.

PReSto contains a rich spectrum of additional support tools to ease the administration of the SAN. An important method is the snapshot technology that generates virtual copies of the virtual disks within a few seconds. After synchronizing the data, both versions can be used independently of each other. The most popular application of the snapshot technology is to ease the backup process within large storage systems. During the backup process, the data has to be kept unchanged, leading to a backup window of many hours for a large file system or data base. With the support of the snapshot technology this backup window can be reduced to a few seconds, because the backup can be made on the virtual copy. The original data can be changed during the backup without interfering the backup process.

A strength of PReSto is the strong protection against the loss of data. Besides the offering of different remote-copy technologies, PReSto can recover very fast from the failure of storage systems inside the SAN. Using the Share-strategy together with a Software-RAID protection, PReSto can reconstruct the data in case of a disk failure without the use of an additional spare disk. If the remaining disk capacity is big enough, PReSto can reconstruct the lost data on the remaining disks. This reconstruction can be done with a minimum number of data movements, returning very fast to a state with full data protection and access speed.

Being a complete storage management solution PReSto contains all features customers will expect by virtualization software. As a result of the simple administration of the storage resources and their optimized use, PReSto helps to reduce the total cost of ownership *TCO* of a SAN. In addition, the reduction of host downtime, e.g. by backup windows, will increase the usability of a SAN with simultaneous cost saving.

REFERENCES

[1] A. Brinkmann, K. Salzwedel, C. Scheideler: *Efficient, Distributed Data Placement Strategies for Storage Area Networks*. In Proc. of the 12th ACM Symposium on Parallel Algorithms and Architectures (SPAA), pp. 119-128, 2000.

[2] A. Brinkmann, K. Salzwedel, C. Scheideler: *Compact, adaptive placement schemes for non-uniform distribution requirements*. In Proc. of the 14th ACM Symposium on Parallel Algorithms and Architectures (SPAA), pp. 53-62, 2002.

[3] R. Graefen. *Wenn zwei sich streiten, … (Storage-Virtualisierung, Teil 2).* Information Week, Dezember 2002.

[4] D. Karger, E. Lehman, T. 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 Proc. of the 29th ACM Symposium on Theory of Computing (STOC), pp. 654–663, 1997.

[5] D. Patterson, G. Gibson, R. Katz. *A Case for Redundant Arrays of Inexpensive Disks (RAID)*. In Proc. of the 1988 ACM Conference on Management of DATA (SIGMOD), pp. 109-116, 1988

[6] M. Raab, A. Steeger. *Balls into Bins – A Simple and Tight Analysis*. In Proceedings of the International Workshop on Randomization and Approximation Techniques in Computer Science, pages 159-170, 1998.

[7] P. Sanders. *Reconciling simplicity and realism in parallel disk models*. In Proc. of the 12th ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 67–76, 2001.

[8] The Storage Networking Industry Association (SNIA). *Storage Virtualization I: What, Why, Where and How*.