Notices

SANique CVM is copyright 2002 by MacroImpact Inc. This includes all software, documentations, and associated materials.

MacroImpact and SANique are trademarks or registered trademarks of MacroImpact Inc. in the United States and in Korea. Linux is a trademark of Linus Torvalds in the United States, other countries, or both. Qlogic is a trademark of Qlogic Corporation in the United States, other countries, or both. Java is a trademark of Sun Microsystems in the United States, other countries, or both. Pentium is a trademark of Intel Corporation in the United States, other countries, or both. GFS is a trademark of Sistina Inc. in the United States, other countries, or both. Other company, product, and service names may be trademarks or service marks of others.

© Copyright MacroImpact Incorporated 2002. All rights reserved. Printed in Korea.
Table of Contents

1 Introduction ......................................................................................................................................................... 1

2 SAN – A New I/O Subsystem Paradigm .............................................................................................................. 2

3 SANique CVM Technology Overview ............................................................................................................... 7

3.1 SANique Working Environment ..................................................................................................................... 7

3.1.1 Symmetric Cluster Architecture .................................................................................................................. 7

3.1.2 Interconnection Mechanism of SAN ............................................................................................................. 8

3.1.3 Shared Storage ............................................................................................................................................ 8

3.1.4 Operating System Support .......................................................................................................................... 8

3.2 SANique Software Architecture .................................................................................................................... 9

3.2.1 SANique Cluster Volume Manager (CVM) ................................................................................................. 9

3.2.2 SANique Cluster System Manager (CSM) ................................................................................................. 11

3.3 Volume Hierarchy ........................................................................................................................................... 13

3.4 Support for Cluster Environments ................................................................................................................ 14

3.4.1 Volume Sharing ........................................................................................................................................ 14

3.4.2 Scalability .................................................................................................................................................. 15

3.5 Support for SAN Environment ....................................................................................................................... 16

3.5.1 64-bit Address Space ................................................................................................................................ 16

3.5.2 Centralized Storage Management ................................................................................................................ 16

3.6 Support for High Availability ......................................................................................................................... 18

3.6.1 Downtime Prevention ................................................................................................................................. 19

3.6.2 Downtime Tolerance .................................................................................................................................. 21

3.6.3 Downtime Minimization ............................................................................................................................ 22

3.6.4 Snapshot ................................................................................................................................................... 23

3.7 Support for High Performance ....................................................................................................................... 24

3.7.1 High Aggregate I/O Bandwidth ................................................................................................................... 24

3.7.2 High Per-application I/O Bandwidth ........................................................................................................... 25
3.7.3 Data Striping .............................................................................................................................26
3.7.4 Partitioned Metadata Management ............................................................................................28
3.7.5 Fine-grained Resource Sharing ..................................................................................................29
3.8 Support for Customizability ........................................................................................................29
3.8.1 Custom-tailored Storage Layout ..............................................................................................30
3.8.2 Selection of Building Block Size ..............................................................................................30
3.8.3 Switching and Tuning Options ................................................................................................31
3.9 Integrateability with File Systems ...............................................................................................31

4 Summary and Future Work ...........................................................................................................32
1 Introduction

SANique Cluster Volume Manager (CVM) is a storage virtualization solution for a cluster environment supporting Storage Area Network (SAN). SANique CVM virtualizes physical storage space and provides virtual or logical volumes for the use of location-independent and restriction-free storage space. The granularity of volume sharing provided by SANique CVM is very fine-grained; file-level and block-level data sharing can be implemented on top of SANique CVM. The main features of SANique CVM include storage virtualization and fine-grained logical volume sharing on top of SAN, heterogeneous cluster environment support, on-line storage reconfiguration, on-line service reconfiguration, and centralized storage management. SANique CVM also provides high aggregate/per-application I/O bandwidth via parallel disk I/O (software RAID 0) and high availability via storage mirroring (software RAID 1). SANique CVM is designed to meet the current market demands for High Availability (HA), High Performance (HP), and High Manageability (HM).

The current demand for storage virtualization and logical volume management is necessitated by a number of practical needs. The size of a single physical volume is limited by current storage technology and a meaningful information-base can hardly be built with such a limited physical volume; virtually unlimited and highly scalable volume space is required to satisfy the current demand. The I/O bandwidth of a single physical storage device is limited and too low to support current computation horsepower requirements; support for simultaneous or parallel disk accesses is inevitable in order to increase both aggregate and per-application I/O bandwidth. Hardware devices, especially storage devices, are subject to physical damage, which in turn may result in disastrous system failures. A mechanism for backing up possible system failures or avoiding unnecessary service interruptions should be embedded into the solution in order to accomplish the goal of 24x7 uninterrupted service.

This technical whitepaper provides a high-level technical and functional overview of SANique CVM and is organized as follows:

- Section 2: SAN – A New I/O Subsystem Paradigm
  Storage Area Network (SAN) is a new paradigm for I/O subsystem and this section provides an overview of SAN features.

- Section 3: SANique CVM Technology Overview
  SANique CVM is an innovative storage virtualization software solution for cluster servers and this section describes its technical and functional features, including:
  - Section 3.1: SANique Working Environment
  - Section 3.2: SANique Software Architecture
Section 3.3: Volume Hierarchy

Section 3.4: Support for Cluster Environment

Section 3.5: Support for SAN Environment

Section 3.6: Support for High Availability

Section 3.7: Support for High Performance

Section 3.8: Support for Customizability

Section 3.9: Integrateability with File Systems

Section 4: Summary and Future Work

This section summarizes the features and functions of SANique CVM with a brief discussion of future work.

2 SAN – A New I/O Subsystem Paradigm

Over the last few decades, computer technology has made great progress in terms of computing horsepower and networking technology. Major technological challenges now seem to lie in the arena of I/O performance where most applications are extremely I/O intensive in terms of storage requirements, available I/O bandwidth, and corresponding I/O response time for retrieving or storing data. In order to develop scalable solutions for servers that are capable of serving different applications, it is critical to develop scalable and cost-effective I/O subsystems. Furthermore, the dependency upon computing services in virtually all business subsections keeps increasing and, therefore, the demand for 24×7 uninterrupted service is higher than ever.

In general, the following demands have been identified to be satisfied in the earliest possible timeframe in order to meet the increasing requirements in both areas of high-performance and commercial-based computing environments:

- Scalable and highly available server solution in commercial service area in terms of computing horsepower, storage capacity, and redundancy
- Centralized storage management support in enterprise environments
- LAN-free serverless on-line backup support for data integrity without consuming service computing power or communication network bandwidth
- Cost-effective alternative to expensive supercomputers for high-performance and parallel computing environments
- Direct data transfer between client and storage subsystems to provide clients with scalable and high I/O bandwidth
Demand-specific storage mechanisms in the file system architecture

Network-striping of data to increase aggregate per-application I/O bandwidth and to support parallel processing

Both computer industry and universities have been searching for feasible solutions to meet these requirements and have suggested a number of alternatives. Clustering is one of them. Clustering is not a new concept; the technology has been known for many years but considered experimental because its core technologies have not been powerful enough to deliver the performance required for real world applications. During the last decade, advances in microprocessor technology have demonstrated a promising future not only performance-wise but cost-wise also; a 2.5 GHz CPU is currently available on the market at the price of approximately $300, and its price tends to go down rapidly. A small-scale system clustered with such CPUs far exceeds the computing power of an expensive mainframe of a few years back, and a large-scale cluster system with hundreds of such CPUs can now even replace a multi-million dollar supercomputer. In networking technology, general purpose networks such as Gigabit Ethernet or Myrinet can now play an active role as an interconnection network, replacing expensive special-purpose private networks used in conventional supercomputer machines. Above all, clustering technology itself is all about scalability and easily eliminates a single point of failure while providing a high degree of availability at a fraction of the cost.

Again, major technological challenges remain in the area of I/O. Even though most applications in these days are extremely I/O intensive, I/O performance is not the only challenge; other I/O related challenges include capacity, scalability, manageability, cost, and availability.

I/O Performance

While all other computer operations are electronic, only disk operations are mechanical. An electronic operation can be performed in nanoseconds while a mechanical operation occurs in milliseconds. Therefore, it is obvious that I/O performance is eventually upper-bound by disk performance. However, the mechanical aspect of disk operations cannot be eliminated unless there is a new cost-effective, non-volatile storage device that works electronically. Therefore, improvement in I/O performance should be made in some other areas than disk performance. Parallelism is one such area in which a significant performance improvement can be expected. Hardware RAID technology improves I/O performance as well as availability by allowing multiple simultaneous accesses to multiple disks and simple mirroring or parity management, but its bandwidth is not only narrow but also unscalable; software RAID or parallel file system approaches should be applied in order to increase per-application I/O bandwidth. Another big improvement can be made by altering the current I/O subsystem architecture; the conventional SAD – Server Attached Disks – is not able to meet the current commercial service requirements in terms of scalable bandwidth, availability, and manageability.
• **Capacity**

A radical shift from text-based data to multimedia data has resulted in the explosion of data to be stored or processed. A single logical data unit is often too large to be stored in a single physical disk device and multi-terabytes of data storage is often not enough for even a small business.

• **Scalability**

An explosive increase in the amount of data affects not only storage capacity but also scalability. A considerable amount of new data storage should be purchased and added to the system in very short timeframes in order to keep up with the increasing business demands.

• **Cost**

Based on a blind bytes-per-dollar calculation, storage is relatively cheap and tends to be getting less expensive. Nevertheless, the amount of money invested in storage is enormous and tends to keep increasing because the number of bytes to be stored is huge and the increase in data use far exceeds the decrease in storage price. Furthermore, in current I/O subsystem architectures, the number of storage devices attached to a single host is physically limited; for instance, 16 SCSI devices per host in general. As applications grow, a whole new server system might have to be purchased to replace the old system in order to provide more storage space even though no more computing power is necessary; the old system can hardly be reused.

• **Manageability**

Managing multiple storage devices and maintaining their integrity while providing uninterrupted service is not only critical but almost painful, especially in a physically distributed computing environment. Managing such systems is labor-intensive and time-consuming while requiring a high degree of expertise. Most organizations maintain one or more management teams to handle storage administration service. Within the current computing environment, the size of the storage management team linearly increases with that of the computing facility. Storage management should be centralized and free from the physical locations of storage devices.

• **Availability**

Most systems providing public or commercial services are required to operate in a 24×7 mode. System availability is heavily dependent upon data availability. In general, the probability of disk failure in a system with \( n \) disks, where the probability of single disk failure is \( p \), is \( n \times p \). In other words, the more disks in a system, the more probable failures. Various levels of mirroring can be applied to provide uninterrupted data service at the price of possible performance degradation. In terms of failure recovery, the conventional `fsck` approach can be no longer applied to multi-terabytes of disk space; it may take days to check and fix a multi-terabyte file system. A journaling approach should be considered for a fast
recovery process at the price of slight performance degradation. Data backup should be continued for recovery from catastrophic system failures, but the amount of data to be backed up is far beyond the range of data that any server can handle while simultaneously providing 24×7 service; on-line LAN-free serverless backup based on snapshot technology is inevitable. In addition, the size of applications tends to grow faster than ever these days and systems are highly likely to be extended within a short timeframe. However, the service should not be interrupted even during such a system expansion; on-line system reconfiguration should be supported to reduce the service downtime.

Storage Area Network (SAN) is a new architecture paradigm for I/O subsystems and has been positively evaluated and supported by the industry and universities. Traditionally, storage devices are connected to the host via a point-to-point connection mechanism, such as SCSI and IDE. Channels are hardware-controlled and therefore fast; neither address resolution nor routing is required. Their usage, however, is limited in terms of scalability and distance. Fibre Channel (FC) technology has been proposed and developed to provide “practical, inexpensive, yet expandable means of quickly transferring data between workstations, mainframes, supercomputers, desktop computers, storage devices, displays, and other peripherals.”¹ Via Fibre Channel technology, storage devices can now be directly attached to the network, and such a network composed of storage devices as well as a set of host computers is called a Storage Area Network. Figure 3-1 illustrates a symmetric cluster architecture built on top of a SAN.

A SAN by itself may not directly resolve all the I/O issues mentioned above, but it does provide an architectural infrastructure on top of which many feasible solutions can be built. In fact, the solutions for some problems mentioned earlier are embedded into its architecture. For instance, scalability in terms of both computing horsepower and storage capacity is no longer an issue in a cluster built on top of a SAN as shown in Figure 3-1. In this architecture, cluster nodes or storage devices can be independently added to or removed from the system on demand; there are virtually no physical limits on the number of disks or nodes that can be attached in this paradigm,”² and thus the cost of storage expansion can be minimized. In addition, there is a direct path between storage devices and the tape drive. Data can now move back and forth between storage and backup devices without going through a server node. An effective third-party transfer mechanism or a small microprocessor-controlled box (often called a mover) can now implement on-line LAN-free serverless backup.

With an appropriate means of software assists, then, SAN architectures can effectively address other I/O problems as well. As illustrated in Figure 3-1, a direct path exists between each cluster member node and storage device; thus, data

¹ “Fiber Channel Overview” ANCOR Communications, Inc., http://www.ancor.com/fcinfo.html#Overview

² The maximum number of storage devices is upper-bound by the given operating system.
can be transferred directly from a storage device to the node requesting data, and vice versa. The conventional performance bottleneck at file server nodes can be easily eliminated in this paradigm. In addition, any single node can easily acquire proper authorization to manage all storage devices in the system so that centralized storage management is possible. Well-designed storage management software should provide virtual storage volumes, which consist of all available storage space in the given system, and its centralized user-friendly interface should transparently manage all such physical storage devices regardless of their physical locations. RAID technology can be implemented in software to provide higher availability and data bandwidth. Due to the potential overhead of handling parity, not all RAID schemes may be feasible to implement in software. Nevertheless, mirroring is still effective since SAN architectures provide simultaneous multiple writes. A service hour per day is also an important factor in terms of availability. In conventional SAD architectures, a server must be brought down in order to add a new storage device and reconfigure the system. Such down time, scheduled or unscheduled, has a significant impact on both users and system administrators. Changes to the system, if any, should be done on-line in order to provide 24×7 uninterrupted service.

It is not possible to meet all such requirements at the same time because satisfying some requirements often comes into conflict with others. Obviously, high availability or journaling for fast recovery sacrifices performance, and performance offers no room for such redundancy. In general, safety and speed are in conflict. The answer is a custom-tailoring of system configuration for each different application domain and applying a different QOS to each mission-critical application. A system should be designed and built to offer as many on-off switches and tuning devices as possible while providing features to satisfy as many requirements as possible. Such a system, then, can be put in service after tuning and customization to meet the requirements with the highest priority of the target application first.
3 SANique CVM Technology Overview

3.1 SANique Working Environment

3.1.1 Symmetric Cluster Architecture

SANique Cluster Volume Manager (CVM) is designed and implemented for a SAN-based symmetric cluster architecture, as shown in Figure 3-1: SANique CVM Cluster Architecture, in which all member nodes provide the same functionality or service and are directly connected to shared storage devices so that the failure of any single component does not have any cluster-wise impact and is, therefore, transparent to the users.

Unlike most conventional distributed software based on a client/server model, SANique CVM employs a peer-to-peer model or symmetric architecture in which all member nodes are running the same piece of code just playing different roles based on effective configuration. Such a symmetric architecture provides an infrastructure for easy on-line system reconfiguration and transparent failover operation as well as uniform service distribution, eliminating both possible performance bottleneck and a single point of failure.
3.1.2 Interconnection Mechanism of SAN

SANique CVM requires each and every member node to have a direct access path to all shared storage devices. However, SANique CVM is independent from the interconnection mechanism of the underlying SAN; it can be Fiber Channel, Myrinet, iSCSI, or any other type of network as long as standardized device drivers are available and each networked device, such as storage devices or host computers, is individually addressable. Even conventional SCSI channel can be used to build a simple two node SANique CVM cluster with a dual-port SCSI RAID. Basically, the design of SANique CVM is hardware-independent.

3.1.3 Shared Storage

SANique CVM is a storage virtualization and volume sharing solution for multiple cluster nodes and assumes the existence of storage devices which are physically shared among the cluster member nodes as illustrated in Figure 3-1. Any storage device supporting Storage Area Network can be used, including high-end RAID devices or mid-range JBODs. NAS devices have their own data service unit and are not suitable for use with SANique CVM. When used with RAID devices, SANique CVM sees each LUN as a row storage space no matter whether it is internally mirrored or striped. Therefore, SANique CVM can support two-phase storage redundancy or higher I/O bandwidth transparently when combined with RAID storage systems because SANique CVM itself provides volume mirroring and data striping via software RAID 0 and 1.

3.1.4 Operating System Support

SANique CVM is essentially a device driver; it creates virtual (or logical) volumes and provides interfaces between the OS kernel and virtual volumes. Developing a device driver is relatively straightforward and usually independent from the underlying OS internals. SANique CVM, however, is a cluster solution, which requires inter-node communication between all cluster member nodes, and implementing the kernel-level communication package is heavily OS-dependent. SANique CVM relies on a TCP-based kernel level inter-node group communication package – SANique CCM. Even though TCP is a standard protocol, its actual implementation varies from system to system. SANique CVM is basically designed to be operating system independent at its highest level and abstracts such differences in the implementation. SANique CVM implements a filter driver in order to accommodate the differences between Unix-like operation systems and others wherever required. SANique CVM therefore supports various operating systems, including SUN Solaris, IBM AIX, HP-UX, SGI IRIX, GNU Linux, and even MS Windows. Please contact our sales department at sales@macroimpact.com for more information on the supported versions for each operating system.

3 SANique CVM 2.1 currently supports SUN Solaris 8 and Linux 2.4 kernel.
3.2 SANique Software Architecture

Figure 3.2 demonstrates the SANique CVM software architecture within the kernel stack. SANique CVM consists of two major components or modules: cluster volume manager (CVM) and cluster system manager (CSM). SANique CVM provides general volume management and storage virtualization while SANique CSM is responsible for managing the cluster system, including cluster inter-node communication, on-line system reconfiguration, and automatic failover. These SANique components are dynamically loadable kernel modules. When loaded, SANique CVM locates on top of device drivers and under the type-dependent file system layer. Raw virtual volumes created by SANique CVM can be directly accessed and managed by applications such as DBMS or other applications which implement their own data management. SANique CSM provides the system call interface to communicate with the command line user interface and coordinates the activity of SANique CVM. Accessing shared storage devices or volumes via SANique CVM is transparent to users – users can access shared storage or volumes as if they were local storage devices, without being aware that such volumes are actually shared and concurrently accessed by multiple nodes.

![SANique CVM Architecture inside Kernel](image)

3.2.1 SANique Cluster Volume Manager (CVM)

A cluster system is different from the conventional aggregation of network-connected multiple hosts in terms of the degree of resource sharing. Especially in a cluster built on top of a SAN, all member nodes physically share storage resources. This feature is the major difference between SANique CVM and other volume management software such as Linux LVM or Veritas VxVM. Unlike other conventional volume management software, SANique CVM allows multiple cluster nodes to physically and concurrently share virtual volumes at the block-level without corrupting data integrity. This type of sharing is far different from an NFS approach, which provides data sharing via file service. Each node in the cluster directly accesses storage devices as if they were their own local volumes. SANique CVM supports
block-level data sharing. SANique CVM controls concurrent accesses from multiple nodes in order to guarantee volume-level data integrity. Concurrency control can be provided through volume metadata service, which is evenly distributed among all participating cluster member nodes. If one or more service nodes fail, other active member nodes can take over the services of the failed nodes, and therefore there is no single-point of failure or a service hot spot.

SANique CVM provides the following major functionality:

- **Storage Virtualization**
  SANique CVM provides and manages virtual volumes, which are location and capacity independent. Users can create the storage volumes of any size up to \(2^{64}\) bytes and do not need to know how many physical storage devices are used or where those devices are physically located once a volume is created. When a volume runs out of space, additional space can be easily added to it on-line. SANique CVM, therefore, allows users to make the best use of storage space, free from any physical limitations of the storage devices.\(^4\)

- **Software RAID**
  SANique CVM supports RAID 0 (striping), 1 (mirroring), 0+1 (striped mirroring), and 1+0 (mirrored striping), providing a reliable recovery mechanism based on metadata journaling. A RAID volume created and managed by SANique CVM is not corrupted unless there is hardware damage. SANique CVM supports scalable N-way data striping. Currently, SANique CVM supports up to 32-way data mirroring and each mirrored volume image is hot-swappable, meaning that it can be taken out of service or placed back in service at any time. SANique CVM also supports striping of a mirrored volume or mirroring of a striped volume.

- **Volume Snapshot**
  Backing up a volume of data is not a trivial task, especially for large volumes which cannot be taken off-line because it consumes of a considerable amount of computing power and the data being backed

\(^4\) There is no clear distinction between virtual and logical volumes. One commonly used distinction is the scope of management. Logical volumes are traditionally managed by a single host and logical volume management is similar to in-band storage virtualization; there is no concept of a volume service provider or clients. Virtual volumes, on the other hand, can be further divided into two sub-categories: in-band and out-of-band virtualizations depending on the location of the service provider. The scope of a virtual volume, therefore, is somewhat wider than that of a logical volume.

Storage management provided by SANique CVM is more similar to in-band storage virtualization, but it is also clusterized logical volume management. In this document, the terms ‘logical volume’ and ‘virtual volume’ will be used interchangeably.
up keep changing during the processing. A snapshot technology resolves this problem by leveraging an indexing technique with the use of some additional space. SANique CVM supports a volume-level data snapshot so that on-line data backup of any size is possible. Since a file system is built on top of a volume, SANique CVM also supports a file snapshot; a snapshot image of a volume is a snapshot of a file system at a particular time and a file-level incremental backup can be implemented from a snapshot image of the given volume. In addition, a serverless and LAN-free backup can be also implemented from a volume snapshot image because a volume image resides within the SAN storage devices, which are physically separated from the service nodes and provide additional paths for other service providers, such as a mover.5

3.2.2 SANique Cluster System Manager (CSM)

Compared to a single host autonomous computing system, a cluster system provides two obvious advantages. First, it provides not only abundant but also scalable computing horsepower – multiple CPUs can cooperate to process a single job and extra horsepower can be easily added to the system when needed. Another advantage of a cluster system is that it is architected to provide better availability. Unlike a single host system, a cluster system rarely goes down completely unless the power fails within the whole building or a worse disaster occurs. Therefore, a service will rarely be interrupted if it is properly partitioned, distributed, or duplicated across multiple member nodes of the cluster. However, it is very difficult to build a reliable cluster system. The term ‘clustering’ implies resource sharing in its strictest definition and at least synchronized collaboration in its weakest definition. In a cluster environment built on top of a storage area network as shown in Figure 3-1, the only physically shared resources are storage devices. Managing these shared resources is not a trivial job, but SANique CVM is doing the job in cooperation with SANique CCM (see below). Other resources – such as CPUs, memory, or in-core data on each member node – should also be properly managed to work together in harmony, and SANique CSM performs this housekeeping for the given cluster system. SANique CSM basically provides three major functions: cluster collaboration support, on-line management support, and fault-resilience support.

- Cluster collaboration support

One of the basic requirements for cluster collaboration support is the existence of a communication mechanism between participating entities – cluster member nodes in this case. SANique CSM includes a cluster communication manager (CCM) which provides a set of TCP-based primitives for inter-node communication, including both synchronous and asynchronous point-to-point communication primitives, membership-oriented group communication primitives, and active message type remote function call primitives. SANique CSM itself uses such primitives to communicate with other SANique CSM instances on

5 Mover is a common name for a serverless or LAN-free backup appliance.
other member nodes while offering them to other SANique modules, such as SANique CVM, for their communication requirements. In fact, each SANique module provides its own cluster collaboration functionality; SANique CSM just provides the mechanism for modules to communicate with each other and coordinates cluster synchronization.

- **On-line management support**

  A major function of SANique CSM is resource management. The set of resources managed by SANique CSM includes SANique modules, devices such as NIC, HBA, or storage, and other member nodes in the cluster. SANique CSM monitors the status of these resources periodically or upon request and takes necessary actions as required to keep the cluster as healthy as possible. Performance monitoring is not included in the current SANique CVM release, but is planned for a future release. Application monitoring and failover service are planned for a separate user-level cluster software product (under development). At the heart of resource management is cluster membership management. SANique CSM manages and grants SANique cluster membership and is responsible for adding or removing SANique member nodes or storage devices on-line. Such membership information is also the basis for group communication provided by SANique CCM. Based on SANique membership information, SANique CCM provides a heartbeat service for each member node and raises an alarm if any member node stops responding.

- **Fault-resilience support**

  Upon receiving a report of a possible node failure, SANique CSM investigates the status of the suspected member node. A false alarm will be determined by multiple crosschecking and simply dropped out. When one or more actual failures are detected, SANique CSM initiates and coordinates a failover procedure by invoking the recovery procedure of each module in a defined order. The failover procedure includes the following steps:

  1. Identify and investigate failed resources
  2. Isolate failed resources
  3. Suspend current service
  4. Take over the services provided by the failed resources
  5. Recover corrupted resources
  6. Update membership and cluster configuration
  7. Resume the suspended service

  When one or more secondary failures occur during the failover procedure, the failover procedure in progress is immediately suspended and another failover procedure starts again from the beginning for both newly and previously failed resources. The entire failover procedure takes a few seconds in most cases and tens of
In summary, SANique’s storage virtualization solution consists of two major functional components. SANique CVM is designed and optimized to meet the requirements for storage virtualization and block-level data sharing among multiple cluster nodes. Volume sharing among multiple nodes is transparent to the users and can best be utilized in developing cluster-aware applications or in operating those applications with idempotent I/O activities. SANique CSM is tightly coupled with SANique CVM for on-line cluster management and fast recovery from system failures. SANique CSM monitors the activity of cluster system resources and invokes necessary failover procedures in case of any failure including the recovery of shared virtual volumes. SANique CSM also allows cluster system resources, such as computing nodes or storage devices, to be added or removed on-line without interrupting current services. As a whole, SANique’s storage virtualization solution provides highly available and robust block-level volume sharing among multiple member nodes of a SANique cluster.

### 3.3 Volume Hierarchy

SANique CVM provides volume management and good volume management relies heavily upon a well-defined volume hierarchy. A well-defined volume hierarchy reduces implementation cost, enhances manageability, and improves performance. A deep hierarchy may enhance flexibility, but results in management overhead and possible performance degradation. The hierarchy of SANique CVM volumes consists of three building blocks: physical disks, space chunks, and logical volumes. Figure 3-3 illustrates such a volume hierarchy.

![Figure 3-3: SANique CVM Volume Hierarchy](image-url)

A raw disk is first initialized to a physical disk. Physical disks can be grouped together to form a named storage (or disk) group. Logical volumes created within a named storage group cannot cross the boundary of the storage group, so that the entire storage group can be migrated to another system later on. Physical disks that do not belong to any named
storage group form an unnamed storage group. Each physical disk is then segmented into space chunks, which compose logical volumes. The attributes of a logical volume include volume size, space chunk size, a list of space chunks, a configuration scheme, the logical volume type, and scheme-dependent parameters. Configuration schemes include:

- Concatenation - simple concatenation of space chunks
- RAID-0 or striping
- RAID-1 or mirroring
- Chained-declustering or RAID-1+0

Logical volume type is one of:

- Data type: stores user data and metadata
- Log type: stores file system metadata log (used by journaling file systems)
- Snapshot type: stores snapshot information

### 3.4 Support for Cluster Environments

Clustering is not a new solution – the concept was first introduced back in the early 80’s and then ignored due to the infeasibility of its technical requirements. Network solutions at the time could not meet the given requirement and the implementation cost was too high. The situation has changed over time. Networking technology has advanced significantly during the last few decades, mainly due to the popularity of the Internet, so that the time lag between remote and local access is now within an acceptable window. The cost of building a cluster is still relatively high, but negligible when compared to the benefits that can be obtained by achieving high availability and scalability. In the current market, clustering is gaining people’s attention back as an unfair choice for HA solution. SANique CVM is a cluster solution for SAN storage virtualization and management.

#### 3.4.1 Volume Sharing

In a cluster environment built on top of a SAN, storage devices are physically exposed to all nodes in the cluster, allowing any node to read or write directly from/to any storage devices. Therefore, without an appropriate concurrency control mechanism, data can easily be corrupted. SANique CVM virtualizes physical storage and provides a single logical view of storage space while providing block level volume sharing among multiple cluster nodes. Applications distributed across multiple nodes can share such virtual volumes without corrupting them.

---

6 This configuration scheme is supported starting from the SANique CVM 2.2 release.
3.4.2 Scalability

The cluster architecture itself provides an infrastructure for scalable solutions in terms of the hardware structure, as does a SAN. SANique CVM coordinates and completes a scalable system solution by supporting interrupt-free on-line operations for any change in the system configuration. Any node or storage device added or removed to/from the cluster system can be easily reconfigured, and any service required to reflect such changes can be also reconfigured on the fly.

Addition/Removal of Cluster Nodes

From SANique CVM’s point of view, adding or removing nodes to/from the cluster is related to the storage membership and metadata services. When a newly added or removed node is a storage client, the node is simply checked in or out of storage membership on the fly. When the node is a storage service provider, the system metadata service is dynamically reconfigured to reflect the change on-line. Access to the corresponding data blocks can be partially and temporarily blocked while the service is being redistributed, but such inaccessibility is so minimal that it is unrecognizable.

Addition/Removal of Storage Devices

After a storage device is newly added to the system, the device can be initialized to a physical disk and included in the system storage lists without interrupting any system operation. When a logical volume is extended to include the new physical device, data re-striping can be performed while data service to the corresponding logical volume is continued. Slight performance degradation is expected while data blocks are being re-striped because moving physical data blocks is an expensive operation, especially when the size of the volume is large. Nevertheless, the duration of blocked access to a certain portion of storage is relatively temporal and such blocking may or may not affect user operations.

The scalability of storage software is critical in supporting a SAN environment as described in the next section. The housekeeping information related to logical volumes should be non-volatile and retrieved as needed. When SANique CVM is first loaded into the kernel, it retrieves all the metadata related to logical volumes and reconstructs them from where they were shutdown the last time. Each storage device configured under SANique CVM is assigned a unique ID and stores its own housekeeping information for logical volumes. However, simple migration of a storage device from one cluster environment to another may or may not work; the storage device can be recognized, but may not be used in constructing logical volumes unless it maintains all the information of related logical volumes. On the other hand, the migration of a whole storage (or disk) group from one cluster environment to another is always guaranteed to work. SANique CVM defines the concept of disk group and a logical volume cannot cross the boundary of a disk group.
Performance Scalability

SANique CVM also provides scalability in terms of I/O performance by striping data across multiple storage devices. Data mirroring can also provide scalable and enhanced performance for read-oriented applications by distributing concurrent read accesses to the same file among duplicated copies. Such data mirroring and striping are revisited in more detail in Section 3.6 and 3.7, respectively.

3.5 Support for SAN Environment

The mechanical aspect of storage devices has been an “untouchable” barrier for I/O performance to make a quantum leap to the next generation. Many efforts have been made to eliminate the mechanical aspect from the storage structure – some have made meaningful progress (e.g., non-volatile organic material), but the magnetic storage devices based on mechanical movement are still the only cost-effective solution. It seems that no other feasible alternative to the magnetic device will be available for at least another few decades. People have begun to look at other possibilities to by-pass this barrier, and SAN is the current solution that can compensate for the drawback of mechanical devices.

3.5.1 64-bit Address Space

SAN architecture is scalable; the number of storage devices that can be recognized by the system is currently 128 (due to the restriction on the device minor number of the system) and eventually unlimited. SANique CVM internally uses 8-bytes or 64 bits in order to address each storage unit. The maximum number of storage devices is upper-bound by the underlying OS, not by SANique CVM.

64-bit operations are not supported in some operating environments, and address translation may cause possible performance overhead due to emulated address operations (e.g., Linux on Intel machine). Some popular operating systems use 32-bits for address handling inside their kernels. Linux defines a 32-bit variable type for storage addresses in its buffer cache implementation. Such type-dependent kernel code should be patched out until a full 64-bit kernel is released. Many other commercial Operating Systems such as Solaris and AIX support a full 64-bit address space.

3.5.2 Centralized Storage Management

The management of a large-scale storage subsystem is a time-consuming and labor-intensive job. When storage devices are physically distributed, as in SAD-based cluster environments, it is almost painful. The unorganized management of
storage subsystems may frequently cause an irritating interruption of business transactions. SAN architectures provide an infrastructure for a single point of storage management and SANique CVM coordinates centralized storage management. All the storage devices under SANique CVM control can be initialized, configured, allocated, monitored, and even traced through a single management interface. System administrators no longer need to log into each machine individually to manage storage subsystems. SANique CVM classifies storage devices into three categories: uninitialized disk list, ungrouped CVM disk list, and named disk group. SANique CVM maintains a separate list for each category.

- **Uninitialized Disk List**

  Storage devices not initialized for SANique CVM belong to this list. When SANique CVM detects uninitialized storage devices, they are listed with their specifications and initialized to physical disks as needed. One or more raw disks can be designated as a hot spare in order to backup unexpected device failures.

- **Ungrouped CVM Disk List**

  When a raw disk is initialized to a physical disk, it belongs to the available storage list. SANique CVM then starts to use the new physical device for creating a new disk group or maintaining existing disk groups.

![Figure 3-4: Storage Management of SANique CVM](image-url)
Named Disk Group

A logical volume can be created out of a single disk group. Each disk group is named when created and maintained by SANique CVM as a list, as shown in Figure 3-4. Such a named disk group can be migrated as a whole to another SANique CVM cluster system.

System administrators can group storage devices in accordance with their hardware specifications, such as I/O bandwidth or overall capacity, in order to synchronize access performance and storage usage. Each group has its own set of attributes and SANique CVM refers to these attributes when creating a logical volume. SANique CVM creates a logical volume out of a single storage group without crossing the boundary of a storage group. A physical disk does not necessarily have to belong to a storage group. Physical disks not belonging to any storage group form a special group, which can also be used to create logical volumes.

3.6 Support for High Availability

High Availability (HA) is the highest priority requirement for service providers. One way to ensure HA is to reduce or eliminate downtime. Yet, serious performance degradation (especially at peak time) must also be avoided in order to provide a high quality of service. This section focuses on reducing downtime; performance degradation is discussed in the next section – High Performance.

The highest possible availability can be accomplished by ensuring the shortest possible downtime or no downtime at all if possible. The best way to reduce or remove downtime is to prevent it completely. Nevertheless, systems fail from time to time. Once the system fails, the issue becomes how quickly to get the system back in action. Providing software and hardware redundancies can mask some failures. Other failures can never be prevented or masked; the system has to be physically fixed and recovered. The question again is how fast the recovery can be accomplished.

The term “downtime” is used to indicate any period during which service is interrupted. Downtime can be classified into two categories: controllable and uncontrollable downtimes.

- Controllable Downtime

Downtimes that fall into this category can be anticipated and, therefore, scheduled in advance. According to recent statistics, about 80% of overall system downtime is controllable or scheduled downtime. Operations that may require scheduling a service interruption include:
SANique Cluster Volume Manager Whitepaper

- Hardware installation, replacement, and maintenance
- Software (including OS) installation and upgrade
- System configuration change

* Uncontrollable Downtime

Any event resulting in uncontrollable downtime is abnormal and not known in advance. Such events can be either physical failures or human-related incidents and include:

- Power Failure
- Hardware failures (node, storage, and network)
- Software failures (including OS)
- Human mistakes (accidental operation or incorrect configuration)
- Human mischief (hacking or virus)

3.6.1 Downtime Prevention

The first step to reduce downtime starts from preventing all possible interruptions and failures in advance. SANique CVM has no control over preventing other software or hardware failures. However, SANique CVM provides redundancies and backup plans in order to deal with these situations; these are discussed in detail in the next section. SANique CVM trusts the underlying operating system and follows the semantics of the OS as far as security is concerned. Avoiding controllable downtime depends mainly on the degree of redundancy and is discussed in the next section. In this section, the reliability, safety, and on-line features of SANique CVM are described.

Highly Reliable Software

The reliability of a software product is critical in satisfying HA requirements. MacroImpact Inc. has its own high-end software testing technology and this code-based software testing technique has been applied to the entire process of designing and implementing SANique CVM. Over 720,000 test runs have been applied and passed with various test configurations and data for the following test techniques:

- Functional (or API) test
- Integration test

Please refer to SureSoft Technology’s CodeScroll Technical Guide or other general SE books for more information.
Sanity Check

SANique CVM is designed to prevent human errors that may result in serious damage. Every irreversible operation that may result in a catastrophic disaster must go through a confirmation process before being committed (executed) and any change in the system configuration must pass a sanity check before actually being applied.

On-line System Reconfiguration

SANique CVM supports the on-line addition and removal of system resources, such as cluster nodes or storage devices. Once a new node with SANique CVM installed is added to the cluster, the acting master node grants SANique CVM membership and reconfigures necessary services. When a member node is removed, the acting master again reconfigures services and then the node checks out. When a new storage device is detected, SANique CVM initializes the device and puts it in the available storage list. If the device is assigned to an existing storage group, on-line storage reconfiguration is initiated. The removal of a storage device can also be done on-line.

On-line Storage Reconfiguration

The attributes of a logical volume can be changed in three different ways: it can grow or shrink, the degree of its distribution may change, and its distribution scheme can be altered. Some changes require a simple addition or removal of storage space (e.g., resizing a concatenated volume) and some require the physical movement of data blocks (e.g., changing the striping degree or scheme of logical volumes). SANique CVM performs such operations while continuing to provide storage services. While a volume is being reconfigured, access to a certain portion is temporarily blocked, but the service continues for the rest of the volume.

On-line Service Reconfiguration

SANique CVM employs a partitioned resource management policy. Metadata service is fully distributed among all member nodes of a SANique CVM cluster. When a member node is newly added to or removed from the cluster, the metadata service is automatically reconfigured so that the service is evenly redistributed among all cluster member

---

8 The automatic on-line detection of a newly added storage device is currently not supported.
nodes.

3.6.2 Downtime Tolerance

With current technology, it is not possible to completely prevent system failures. Mechanical devices especially, such as magnetic disks, are easily damaged and subject to arbitrary failures. In reality, every system component including the OS is susceptible to failure. Yet, most of these failures can be hidden with an appropriate degree of redundancy. A fault-tolerant architecture with redundancy can effectively absorb a single point of failure and significantly reduce system downtime.

Storage Redundancy

Hardware RAID technology is a well-known solution for high availability and high performance. The problem with hardware RAID technology is its high cost and low scalability. SANique CVM provides a scalable software RAID solution for SANs. Software RAID technology within a SAN environment provides high availability, high performance, and high manageability. Reliability has been an issue for software RAID solutions, but SANique CVM provides a reliable solution through a metadata logging technique. Currently, SANique CVM supports RAID-0, RAID-1, and RAID-0+1 on top of logical volumes.

* RAID-0

Software RAID-0 stripes files across multiple disks. This technique helps to scale up I/O bandwidth because a single I/O operation is performed on multiple disks simultaneously. Ideally speaking, when a file is striped across N disks, $B_d \times N$ per-I/O bandwidth is expected. But, due to difficulty in synchronizing disk accesses, variations in disk seek time, and different queuing delays, it is not feasible to achieve $B_d \times N$ bandwidth with N-way striping.\(^9\) With well-designed software support, up to 80% of $B_d \times N$ bandwidth can be achievable. Unlike hardware RAID systems, SANique CVM has no explicit limit on the degree of data striping. A file can be striped across as many physical disks as the maximum number of disks that the given SAN can accommodate, and the striping scheme and degree can be reconfigured on-line.

* RAID-1

Software RAID-1 maintains one or more mirrored images of each data block on separate physical disks. A logical volume configured for RAID-1 is trouble-free from a single disk failure and provides high availability, but it may suffer from performance degradation due to synchronous data writing. Applications with frequent

\(^9\) $B_d$ = maximum bandwidth of a single disk, $N$ = number of disks
updates may not be appropriate for this configuration. SANique CVM allows the degree of redundancy to be expended up to 32 so that the maximum 32 copies of a data block can co-exist at the same time. Applications with read-oriented access patterns can benefit from this configuration; the maximum of 32 parallel reads on the same data block can be possible even without concurrency control overhead.

- **RAID-0+1 or RAID-1+0**

  Software RAID-0+1 is a combination of RAID-0 and RAID-1; files are first striped across multiple disks and then mirrored (or the other way around for RAID-1+0). SANique CVM ensures that a single disk failure does not result in data inaccessibility when allocating logical storage units for stripping and mirroring.

**Service Redundancy**

Service is an abstract object and the concept of service redundancy is a bit different from other hardware redundancy. Service redundancy can be achieved in two different ways. One approach uses a hot-standby technique. As in hardware redundancy, two or more images of a service are up and active at the same time, but only one – usually called the “primary server” – actually provides service. The other image or “secondary server” resides on the other node and monitors the activity of the primary server. When the primary server is not able to serve, the secondary server immediately takes over the service job. This is a typical auto-fail-over architecture in a centralized service-providing environment. SANique CVM employs a partitioned resource management policy and therefore takes another approach. Instead of a hot-standby, SANique CVM provides redundancy on-demand. The main service provided by SANique CVM is a metadata service. When a node providing metadata service fails, either one of the metadata client nodes takes over or the corresponding service objects are redistributed among other server nodes as determined by the system configuration setting or system administrator’s choice. Service reconfiguration is done automatically by SANique CSM. 

3.6.3 **Downtime Minimization**

It is not possible to eliminate all failures. If it is not possible to prevent a failure, the next choice is to minimize the damage when the failure occurs. A disk access is the most expensive operation in a computer system and also the most vulnerable to data loss when there is a system failure. In order to preserve high system performance, current I/O technology relies on a deferred disk access (such as write-behind policy) in order to minimize the number of disk accesses. The fresh copies of data blocks are kept in core – usually in the system buffer cache – as long as possible. Depending on the hit ratio, such a caching technique improves system performance significantly. The drawback of this technique is possible data loss. When a system suddenly crashes, all data and metadata blocks kept in the system buffer

---

10 Refer to 3.2 SANique Software Architecture
SANique Cluster Volume Manager Whitepaper

cache are lost, the integrity of the corresponding file systems is destroyed, and the file systems are corrupted. The conventional approach of recovering file system consistency – *fsck* – may still work, but is not feasible in a cluster environment with multi-terabytes or more of storage space; it may take a whole day to scan all the storage systems. The situation is same with logical volume management. SANique CVM supports metadata logging in order to avoid disk scanning and provide a fast volume recovery in the event of system failures.

**Metadata Logging**

A journaling file system can recover the consistency of the corrupted file system with a fixed timeframe regardless of the file system size. The recovery process ensures data integrity in a storage subsystem environment built on top of physical devices. In a virtualized storage environment, especially one built on top of a SAN, there might be single or deeper indirection (or mapping) between file metadata and physical blocks. The recovery of file system alone may guarantee a consistent file system, but with possible scrambled data mapping. For example, a file may be recovered with the right access time and correct size, but with incorrect data blocks. SANique CVM journals metadata logs for logical volumes. SANique CVM has an internal recovery module that reads metadata logs, analyzes them, and re-applies them in order to recover data integrity. SANique CSM, when necessary, can automatically initiate SANique CVM’s recovery process in accordance with file system recovery. When SANique CSM is not available, the volume recovery process can be launched manually by the system administrator.

- A single file operation – say a single write – may launch multiple block operations in the logical device level; yet the file operation can only be completed upon the completion of the last block operation. If the system crashes in the middle of multiple block operations, the file operation is incomplete; it should either be aborted (undone) or completed (redone). In other words, the atomicity of the file operation must be guaranteed. For instance, a single write on a logical volume configured for 2-way RAID-1 requires two separate writes on both mirrored physical devices. When the system crashes after the completion of one write and before the completion of the other, the consistency of the corresponding file system is lost when the system comes back up. In general, ACID properties are applied to all file operations in a virtualized storage environment as in a DBMS. SANique CVM applies the concept of a transaction to all file operations and guarantees atomicity by logging block operations.

### 3.6.4 Snapshot

HA solutions for downtime prevention, tolerance, and minimization all assume a normal everyday working environment. Unplanned failures do not occur frequently, yet can happen at any time. Typically, a backup solution is used to minimize the impact of these failures. Yet, even a good backup solution cannot protect data from a fire, a
bombed, or an earthquake. An off-site data replication solution in which data are backed up onto a remote site that is physically far away is recently earning people’s attention. Snapshot technology is critical in order to support such HA solutions for large volumes of data. SANique CVM supports snapshot at the logical volume level so that data can be backed up on-line. In addition, due to the characteristics of a SAN, a LAN-free or serverless on-line backup can be easily implemented with this snapshot volume for both volume-level and file-level incremental backups.

A snapshot at file system layer creates an instant image of a file system and includes file information and directory structures. With such information, data can be selectively restored when necessary – a single file or directory may be selected and restored. A volume manager, however, should be file system independent and does not know anything about file or directory structures; that is the file system’s jurisdiction. A snapshot at the volume manager layer creates an instant image of a logical volume. The entire logical volume should be restored or nothing can be; files or directories cannot be selectively restored\(^{11}\). In a sense, this type of snapshot is better for replication than for normal backup. Snapshot technology is also useful to freeze the current state of a volume or the entire storage system when making a management decision.

3.7 Support for High Performance

The requirements of high performance create another dimension different from the HA line in the requirement matrix. System performance can be defined in many different ways: overall throughput, response time, the number of service connections, no drastic performance degradation, and so on. All of these performance metrics are somehow related to each other and heavily depend upon the I/O performance; no system output can be generated without data being properly supplied. SANique CVM provides an infrastructure for the highest possible system performance through high I/O bandwidth, data sharing, direct data transfer, and demand-specific storage layout.

3.7.1 High Aggregate I/O Bandwidth

The I/O bandwidth of a physical disk is upper-bound by its mechanical aspect; it lineally improves, but can never catch the exponential improvement of electronic operations. The performance of the conventional storage systems based on a SCSI connection is also upper-bound by that of a physical disk due mainly to its channel attributes – one disk at a time no matter how many disks are connected. Hardware RAID technology was a breakthrough that made it possible for the performance of a storage device to go beyond the physical limit. The drawback of a hardware RAID system is its high cost and poor scalability. SAN can be described as a generalized RAID technology with good scalability and a virtually unlimited number of connection slots. The aggregate I/O bandwidth of SAN is now \(B_d \times N\), where \(B_d\) is the bandwidth\(^{11}\).

\(^{11}\) Backup software can implement file-level incremental backup from a snapshot volume.
of a single physical disk and $N$ is the number of disks connected to the SAN. In addition, $N$ is not physically fixed, but virtually unlimited and dynamically reconfigured. The bandwidth of the SAN is now upper-bound only by that of the connecting network, which is usually far higher than that of storage devices. SANique CVM is a software solution that can achieve an aggregate I/O bandwidth of $B_d \times N$.

![Diagram of parallel application access](image)

**Figure 3-5: Data Accesses of Parallel Processes**

### 3.7.2 High Per-application I/O Bandwidth

Effective I/O bandwidth from an application’s point of view is usually not the same as the overall aggregate I/O bandwidth of the system. Even when all storage devices are busy serving requests, only a portion of the service can be allocated for a given application unless the data file of the application is spread out across all storage devices. The turnaround time of an application could be multiplied by the factor of $p$ (where $p$ is the number of active processes in a node) due to the characteristics of an interleaved I/O service when the application is a parallel or distributed process. SANique CVM guarantees a high per-application I/O bandwidth by allowing the striping of a data file across all
available storage devices; an application can enjoy a full $B_d \times N$ I/O bandwidth exclusively. In addition, SANique CVM supports simultaneous accesses to the same file by multiple nodes or processes when associated with appropriate cluster or parallel file systems (e.g., SANique™ CFS) that support IX (Intentionally eXclusive) and IS (Intentionally Shared) concurrency control modes. Figure 3-5 illustrates such simultaneous accesses to the same file by multiple parallel processes.

### 3.7.3 Data Striping

Data striping is a way to increase I/O bandwidth. The maximum I/O bandwidth of a single disk, $B_d$, is physically limited and low. When a file is striped across $N$ disks, the possible I/O bandwidth of the corresponding file can be $B_d \times N - N$ times higher than that of a non-striped file. From an HA perspective, however, data striping without redundancy is dangerous. When the probability of a single disk failure is $P$, the probability of file inaccessibility becomes $P \times N$ when a file is striped across $N$ storage devices. Nevertheless, the need for higher I/O bandwidth may offset the potential danger. SANique CVM supports data striping with a number of different policies and user-defined degrees to allow the system administrator to select the optimal solution for each particular service.

The degree of data striping and its policy are specified when a logical volume is created with SANique CVM. If they are not specified, the default system values are used in creating logical volumes. Once a logical volume is created and a file system is initialized on top of it, all files created on the given file system inherit the attributes of the underlying logical volume; files are striped across $D_s$ storage devices according to the striping policy of the logical volume, where $D_s$ is the striping degree of the underlying logical volume and ranges from 1 to the total number of storage devices in the system.

---

12 IX and IS lock modes are currently not supported.
SANique Cluster Volume Manager Whitepaper

Figure 3-6: Striping Policies

SANique CVM currently supports the following striping policies:

- Distributed cyclic layout or simple round-robin (Figure 3-6 A)
- Right-symmetric: Right-staggered distributed cyclic layout (Figure 3-6 B)
- Right-symmetric: Generalized right-staggered distributed cyclic layout (Figure 3-6 C)
- Left-symmetric: Left-staggered distributed cyclic layout (Figure 3-6 D)
- Left-symmetric: Generalized left-staggered distributed cyclic layout (Figure 3-6 E)
- Extended right-symmetric (Figure 3-6 F)
- Extended left-symmetric (Figure 3-6 G)

13 Extended right-symmetric (Figure 3-6 F) and extended left-symmetric (Figure 3-6 g) are not supported in SANique CVM 2.1.
These striping policies are illustrated in Figure 3-6. Different striping policies should be selected to distribute the access load as uniformly as possible.

**File Level Data Striping**

When a file system is first created on top of a logical volume, each logical extent (or block) of the file system is mapped onto the corresponding physical extent (or block) of the underlying logical volume in one of the modes illustrated in Figure 3-6. The logical extents then are allocated for files. However, the order of extent allocation is arbitrary and random from a file’s point of view. The overall distribution of extents might be uniform, but there is no guarantee that the adjacent logical extents of the same file are physically distributed among different storage devices; in the worst case, an entire file might be stored on a single disk. One such possible distribution is illustrated in Figure 3-7 (A). In Figure 3-7 (A), 9 logical extents are evenly distributed among 3 physical disks – 3 extents on each disk, but only two disks are accessed when three consecutive extents (e.g., extents [0, 1, 2] or [1, 2, 3]) are simultaneously accessed. The arbitrary interleaving of extent allocation for different files results in such a non-uniform distribution. SANique CVM provides an API that forces the ordered and uniform distribution of striped extents for a file as illustrated in Figure 3-7 (B). Any file system that uses such an API can guarantee the uniform distribution of file extents ensuring the maximum I/O bandwidth.

<table>
<thead>
<tr>
<th>0/0</th>
<th>1/1</th>
<th>2/1</th>
<th>3/2</th>
<th>4/2</th>
<th>5/0</th>
<th>6/1</th>
<th>7/0</th>
<th>8/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Overall Uniform Distribution of File Extents with Partial Density (File Extent Number / Disk Number)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>0/0</th>
<th>1/1</th>
<th>2/2</th>
<th>3/0</th>
<th>4/1</th>
<th>5/2</th>
<th>6/0</th>
<th>7/1</th>
<th>8/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(B) Uniform Distribution of File Extents with No Partial Density (File Extent Number / Disk Number)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 3-7: Striping Pattern of File Extents**

### 3.7.4 Partitioned Metadata Management

When there are multiple resources that can be accessed by multiple users, the complexity of management rises. In logical volume management, there are multiple storage objects to be managed and a number of nodes accessing those objects. A client/server model is the traditional approach. One node called a server node is responsible for managing resources and other nodes, typically called client nodes, access resources under the server’s control. The semantics of a client/server approach is simple and easy to implement. Its drawback is a possible performance bottleneck at the server node; all service requests are queued up at the server node and there is a non-deterministic waiting time when the server is not powerful enough to serve all requests in a given time window. Another problem with a client/server model is that

---

14 This feature is not supported in SANique CVM 2.1.
the server becomes a single point of failure. When the server fails, all services related to the given objects are interrupted. In conventional cluster architectures with SAD, there are not many choices besides a client/server approach because the server physically owns storage devices or objects; there is no direct path between clients and storage objects to allow clients to bypass the server. In a cluster environment built on top of a SAN, multiple paths exist between cluster nodes and storage devices and a direct path always exists between any node and any storage device. Such an environment enables another management methodology – serverless management – in which all nodes are responsible for the management of all storage objects. There is no single point of failure. The drawback of a serverless approach is a possible chaos, so a means of ensuring storage consistency must be provided. The most common solution is a concurrency control agent, often called a lock manager, and in fact, most cluster file systems employ a locking mechanism in order to prevent storage inconsistency. SANique CVM employs another approach – partitioned resource management.

Within SANique CVM, storage objects are partitioned and each group of partitioned objects is managed by a separate server. The degree of storage partitioning is a boot-time configuration parameter and can also be reconfigured at any time on-line. Depending on the degree of storage partitioning, SANique CVM provides a centralized (degree = 1), a fully distributed (degree = N), or a partitioned (1 < degree < N) storage service. The overall architecture may appear similar to that of a conventional client/server model, in which multiple file servers are providing file services on top of NFS, but there are two distinct features between SANique CVM metadata servers and NFS file servers. First, SANique CVM metadata servers do not provide data service; they do not access data on behalf of other clients – data are transferred directly between clients and storage devices. Second, the failure of a SANique CVM metadata server does not block clients from accessing the corresponding storage objects; the storage objects may be blocked for a negligible amount of time, but another metadata server takes over the corresponding service on the fly.

3.7.5 Fine-grained Resource Sharing

SANique CVM provides building blocks for resource sharing and the actual sharing mechanism is implemented in the upper layer – usually a file system layer. Yet, providing the building blocks for sharing is as important as implementing the sharing mechanism. Veritas™ cluster VxVM, for instance, provides a disk group as a sharing building block and only one cluster node can access the entire disk group at any given time. SANique CVM supports block level resource sharing; a single node can access each block at a time and multiple nodes can access different blocks at the same time.

3.8 Support for Customizability

HA (High Availability) and HP (High Performance) are the two most important requirements in the current market. Unfortunately, HP solutions and HA solutions generally conflict with each other – almost every HA solution degrades
system performance and HP solutions have no room for redundancy. The choice between HP and HA depends on the given requirements; the solutions can favor HP or HA, or stay somewhere in the middle depending on user requirements. SANique CVM provides a fine-grained tuning mechanism. When all HA options are turned off, it behaves as a high performance system. When all HA options are on, it becomes a high availability system.

3.8.1 Custom-tailored Storage Layout

Planning storage layout is important, especially for a large-scale storage system. SANique CVM provides on-line storage reconfiguration, so that storage can be reconfigured at any time. Nevertheless, moving physical data blocks is very resource intensive and should be avoided as much as possible. Hence, it is important to plan storage layout carefully from the beginning. SANique CVM provides the concept of storage groups; storage devices can be grouped together and logical volumes are created within the boundary of such storage groups. Such storage groups then can be detached from the system and moved to another system.

A logical volume (LV) can be created either top-down or bottoms-up. The system administrator specifies size, striping degree, and striping policy. Then SANique CVM creates an LV in such a way that storage resources are best utilized. This top-down approach is simple and easy, but hardly reflects the specific requirements of the target application onto the storage layout. In a bottom-up approach, the system administrator can build a logical volume by specifying which building blocks should be used and how they should be organized in order to custom-tailor the LV for the given requirements.

3.8.2 Selection of Building Block Size

The basic building block of logical volumes is an extent. An extent is a group of physically contiguous data blocks. The size of a data block is variable and is specified for the given logical volume. An extent is a basic building block for storage allocation and a block is the basic unit of I/O operations. In general, \( E = B \times n \) where \( E \) is the size of an extent, \( B \) is the size of a block, and \( n \) is the multiplying factor. Choosing appropriate values for \( E \), \( B \), and \( n \) is very important and affects the system performance. Some applications access a small amount of data randomly. Some access a large amount of data sequentially. The characteristics of I/O workloads include access granularity, favor, traffic, pattern, type, and orientation as listed below.

- Access granularity (small, large, variable)
- Access favor (locality, uniform)

\(^{15}\) SANique CVM 2.1 currently supports a top-down approach only.
3.8.3 Switching and Tuning Options

SANique CVM is first designed for a high performance system and then revised for high availability by inserting necessary redundancies into the corresponding modules. Every redundancy is optional and can be turned on/off or tuned for the given range. Hence, the system administrator can configure SANique CVM for the requirements of the target application. The switching options currently provided by SANique CVM include:

- Metadata logging (on/off per LV)
- File-level striping (on/off per file)\[16\]
- The degree of data striping (1 \(\sim\) D)
- Storage redundancy or mirroring (1 \(\sim\) 32)

3.9 Integrateability with File Systems

From file system’s point of view, SANique CVM is a device driver and is not different from ordinary device drivers such as a SCSI driver. The major difference between a logical volume of SANique CVM and an ordinary storage partition is that the logical volume is physically exposed to multiple nodes and therefore can be accessed by multiple nodes. In other words, the file system module must provide a concurrency control capability in order to preserve file system consistency as well as data integrity; any cluster file system can be integrated on top of SANique CVM. MacroImpact provides a cluster file sharing solution – SANique Cluster File System (CFS) – which has been fully integrated with SANique CVM to leverage its block-level data sharing and storage virtualization capabilities. Additionally, Ext2 – a native Linux file system – has been modified to act as a cluster file system and successfully (in terms of functionality, not performance) integrated with SANique CVM. The clusterization of Ext3 and xFS is currently in progress, and IBM JFS has been investigated for the possibility of integration with SANique CVM. For integration

\[16\] This feature is not supported in SANique CVM 2.1 release.
purposes, the FDISK functionality of JFS doesn’t appear to be distinct from a logical volume of SANique CVM. Those file systems, however, were designed without considering the shared aspects of cluster environments, and their locking granularities are coarse. Hence, the performances of clustered Ext3, xFS, and JFS are currently in doubt.

Another difference is an extra API set provided by SANique CVM. SANique CVM provides a set of advanced functionality, such as creating a file with file level striping or providing extent allocation hints. Such APIs are used to optimize storage performance. Currently, only SANique CFS is using this functionality, but any cluster file system can also take advantage of this advanced functionality. The APIs are open to users through the system calls of Type Independent File System (or VFS) and are integrated with the SANique runtime library for high performance parallel I/O.

4 Summary and Future Work

The current market demands a shift from a client/server model to a cluster solution due to computing horsepower, high availability, and cost-effectiveness. A shift from SAD to SAN is inevitable due to requirements for large storage volumes, high I/O bandwidth, data sharing, high availability, and cost-effectiveness. A cluster solution on top of a SAN is a new paradigm in storage subsystem architectures. While the hardware market for SAN solutions is already mature, there are as yet few software SAN solutions on the market for a cluster environment including cluster data sharing and cluster storage virtualization. SANique CVM is a cluster storage virtualization solution for SAN, satisfying commercial-grade I/O requirements in terms of performance, availability, and manageability.

SANique CVM is an innovative, concurrent, block-level data sharing and in-band storage virtualization solution with embedded fault-resilience capability in a multi-way tightly coupled cluster environment. SANique CVM creates logical storage volumes by virtualizing physical storage devices and allows multiple member nodes in the cluster to physically share such virtual volumes for concurrent accesses. Journaling and associated failover features ensure that any failure of cluster components can be transparent in terms of service continuity. On-line management allows the extension and reduction of the cluster system without interrupting current service, significantly reducing service downtime. SANique CVM also allows managing cluster system as well as storage devices through an easy-to-use centralized GUI (or console). Finally, its application specific block management capability, along with the high data transfer rate of SAN, enhances the overall system I/O performance a great deal. To summarize, the main features of SANique CVM are:

- In-band storage virtualization
- Software RAID (0, 1, 0+1) support
- Fast and robust block-level volume sharing among multiple nodes
SANique Cluster Volume Manager Whitepaper

- Large volume support based on a full 64-bit address space
- Variable size extent-based space pre-allocation
- Partitioned resource (including metadata service) management for high concurrency
- Fast recovery from system failure by metadata logging
- On-line metadata server reconfiguration
- On-line cluster reconfiguration
- Fully integrated and centralized cluster and storage management with Java-based GUI
- High availability with fast auto-failover

SANique CVM is the only commercially available cluster storage virtualization solution with block-level data sharing capability. Its high availability features meet the current market demand of 24×7 uninterrupted service. MacroImpact provides a cluster file sharing solution – SANique Cluster File System (CFS) – which has been fully integrated with SANique CVM to leverage its block-level data sharing and storage virtualization capabilities.