Challenges and Solutions in Allocating Data in a SAN Environment

Bruce Naegel
VERITAS Software
1600 Plymouth Street
Mountain View, CA 94043
bnaegel@VERITAS.com
Phone +1 650 527 4038
FAX +1 650 527 8659

1.0 Challenges in Allocating Data

1.1 The Storage Explosion

The amount of data used for applications has grown dramatically over the last few years. Market research firms suggest this will continue with estimates of 50% to 100 % per year capacity growth. This growth has occurred across a range of applications including the scientific, database, and Internet and client server marketplace, applications where NASA has data storage.

This storage explosion shows little sign of slowing down. The 72 GB capacity of our largest disk drives today will grow to 1 TB within a few years. In addition to data density on disk drives, we have faster processor speeds to use all this data and increasingly faster networks to access and share the data. This means we need methods of allocating storage for all these uses that can respond to rapidly changing environments in an intelligent fashion.

1.2 Host Based Storage Allocation Methods (Storage Islands)

Storage in many cases is tied directly to the host using it. While this method currently serves the majority of the market, this configuration has limits in a rapidly changing environment. This problem is even more challenging if one needs to move storage facilities from one host to another, especially if this has to be done across operating systems. As an example, one customer visited by the author was supporting 4 different operating systems (Solaris, NT, HP-UX and AIX). This customer was consolidating their applications on Sun and NT, and wished to re deploy their storage resources accordingly.

In addition, a project may require 100 GB for a single month, and then have that capacity moved to another use on another set of hosts. Test deployments for applications may have different capacity requirements from the final production requirements. All of these challenges suggest another method for storage allocation is required.

1.3 Centralized Storage and the Cost of Storage Allocation

IDC (International Data Corporation), a leading IT market research firm, performed a study of Fortune 100 companies in 1997. IDC looked at 3 storage topologies and found a single person could manage 7 times the capacity in a fully centralized case (centralized storage, servers in the same room) as they could with a fully distributed case (storage tied to individual servers, each in separate rooms). This means consolidated storage can save money by saving a precious resource, skilled IT staff. If one looks at recent figures, a highly available storage system may cost $0.10 per MB in hardware, but cost up to 10 times that much in ongoing management costs. Reducing the management costs by making storage allocation easier to manage is key to using all that data with controlled management costs.

1.4 Types of Storage Networks Enabling Centralized Storage.

To centralize storage, one needs to connect multiple hosts to a central disk system using a network, The computer industry today has two types of storage networks in general use, SAN (Storage Area Network) and NAS (Network Attached Storage). Both have a standard network structure with:

- Storage Clients, the computer systems which use or consume the storage
- A network infrastructure consisting of switches, hubs and HBAs (Host Bus Adapters).
- A storage pool, where the storage systems (disk drives, RAID systems), etc. are addressed.

A picture of direct attached storage, SAN storage and NAS storage is shown below:

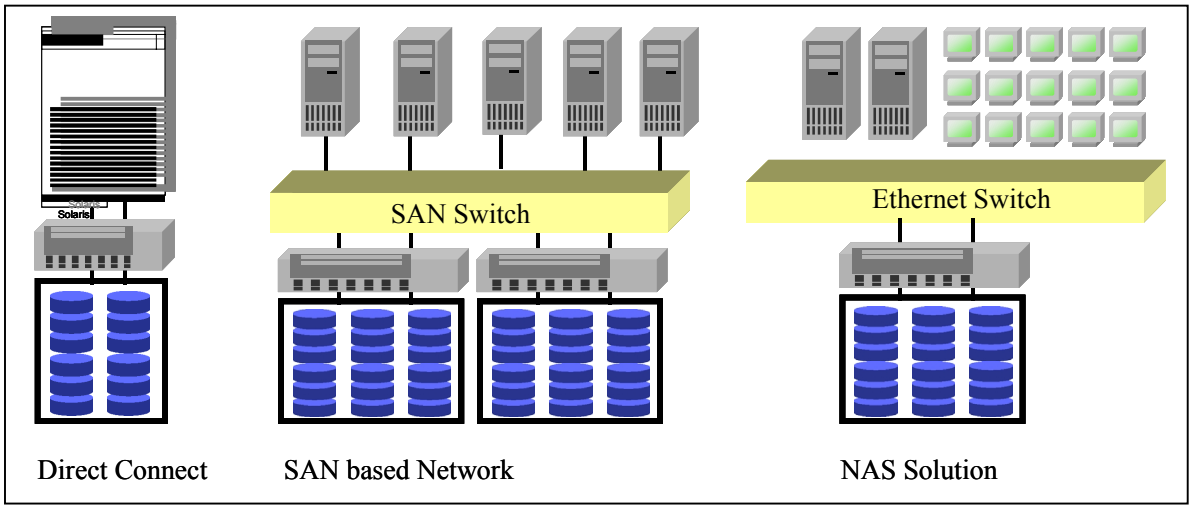

Figure 1: Direct Connect, SAN and NAS Environments

SAN and NAS are a bit different in the type of network and the connection of the storage to the OS itself.

In a SAN, the disk storage is seen by the OS in the Storage Client as a virtual disk drive, The storage appears to the OS much as if one took a physical disk drive and added it to the computer system directly (e.g. a virtual disk drive). The protocol connecting the SAN storage to the storage client is usually SCSI over Fibre Channel, a lightweight protocol designed to work within confined environments, with high performance. One should also note that there are applications that require or favor direct connection to the storage by the OS. Databases using "raw partitions" and Microsoft Exchange fit this category.

SANs can also communicate over longer distances than direct connect storage. The current protocols support up to 10 KM of distance using standard single mode fibre. This distance can be extended via special hardware to 100 KM with somewhat less performance.

Network Attached Storage (NAS) is different than a SAN since the network connection is a standard IP network. Note on the diagram above this is Ethernet, but ATM and Token Ring are also used. Instead of virtual disk drives, the storage is presented as files. This form of data presentation works for a number of applications and allows for data sharing based on the locking provided by the file system used. One should also note since standard networks have developed network resources for security and manageability today, these services are available in NAS environments.

1.5 Is there a Unified Storage Network?

A network with the best attributes of both the SAN (high throughput, low overhead) and NAS (defined network protocols, management) would seem to be a highly desirable goal. One protocol, iSCSI has been proposed to allow for block or virtual disk networked storage over a standard IP network. If iSCSI can minimize the additional overhead of the TCP/IP protocol, iSCSI will offer the performance of direct connected storage with the management facilities in an IP network. Expect to see this in product deliverables in the next few years. It will be available first as a bridge between SANs which are separated by distances greater than the SAN supported 10 to 100 kilometer maximum distance.

There are also advanced forms of file systems to provide file sharing in a Fibre Channel SAN environment, available today. These file systems provide SAN users with file sharing and higher performance than standard NAS solutions. They are discussed in the section "Sharing Data between the Data Acquisition System and the Data Analysis System".

2.0 SAN Storage Networks

2.1 SAN Benefits:

The focus of this paper will be on the SAN environment since this is the one generating considerable market attention and where benefits can be shown today. The high sustained throughput of SANs also lends itself well to the scientific data environments at NASA. SAN technology as viewed by the industry and customer successes provides the following:

- Storage Consolidation and Allocation
- Data Sharing (using SAN based File Systems today)
- Improved Backup and Recovery
- Disaster Tolerance
- Enabling higher availability
- Easier migration to new technologies
- Centralized policy based management

2.2 Challenges in Managing Storage

SAN based storage is planned to grow to an almost 10 Billion dollar marketplace worldwide in 2003, up from 2 Billion in 1999. These projections from IDC are based in on SAN solutions being developed to provide:

- Asset Management
- Capacity Management
- Configuration Management
- Performance Management
- Availability Management
- Policy Management

2.3 QoSS Definitions:

QoSS (Quality of Storage Service) is defined by the following capabilities:

- Availability (usually determined by RAID level and replication capabilities)
- Performance (transfer rate, transaction rate or latency)
- Capacity
- Cost per MB for the storage

2.4 Basic and Fine Grained QoSS Definitions:

In general the terms for describing QoSS (e.g. RAID level) are specific to the storage industry. With many choices for configuration, one has a "fine grained" QoSS definition. However, today customers are defining QoSS within their accounts with a more basic QoSS. This more basic level may just be the level of choosing which storage system to place the data on. This basic approach is a start to understand more fine-grained choices in QoSS and will be used in the illustrations below.

2.5 Using QoSS with Real Life Examples:

Customers need different qualities of storage related to the application.

Consider first the acquisition of new data. This represents the most important part of the data chain. One should ensure this data is safeguarded to the highest level. In the scientific world, this may represent the fly-by of a planet such as the data from the Shoemaker Levy comet -- a once in a millennium type of event. In the business world this is equivalent to a buy or sell order, with revenue (and therefore its loss) as part of the equation. This type of data should be kept on a system with a QoSS which is high in:

- Data Reliability (so no data is lost)
- Sufficient Performance (no loss of data capture is allowed)
- Cost (the features required are the important consideration) Note, this will raise cost over non-highly available solutions.
- Capacity (needs to be easily expanded to continue data downloads).

This defines a high performance (for the workload) and very high reliability disk system.

Once the "new data" has been stored and backed up, it may be analyzed from a safe repository. Since this data is a copy of the "new data", it can be recovered from a primary source if lost. This "new data" then becomes "analyzed data" which should be backed-up to eliminate the need to run the same analysis twice.

Then there is data that is kept on line for historical purposes (either "new" or "analyzed' data"). This data is not referenced often and acceptable access times are much less stringent.

These three data stores represent three different Quality of Storage Service factors. With each different factor, one can use the difference in these factors to choose different storage systems to keep the data.

Since data acquisition and data analysis usually have different I/O patterns, the "new data" may be replicated to a second computer and storage system to assure the data

acquisition systems are not impacted by data analysis. The last phase of data storage may be on tape, where data is retrieved only in rare instances.

3.0 Example System Introduction:

The best way to show how a system works is to provide an example system. Let's assume the following system for better understanding capacity allocation in a scientific SAN environment:

The data acquisition system for "new data" consists of multiple data modules from separate telemetry channels. This creates specific transactions on multiple channels with a large sustained transfer rate and a large transaction rate.

The two-stage data analysis system creates "analyzed data" with two separate compute systems. The primary system does the first level of analysis, preparing the data for the second level of analysis. . The secondary system supports the next level of analysis based on a bank of workstations accessing the data. The storage systems could either be split or part of a larger SAN storage pool.

The data repository system is designed for future analysis of the previous results. This data is designed to be on line, but with a lower level of access performance.

The overall system will be optimized if data can be moved easily from one section to the other (e.g. data acquisition to data analysis). So the first requirement is that all three sections need to be interconnected. A SAN (Storage Area Network) would be ideal for this interconnection since it has:

- High speed of transfer (currently up to 100 MB per second per fibre connection)
- High resiliency and availability (full resiliency including remote site can be built into the SAN
- High capacity (extensive address space means one can scale the number of hosts and storage elements on the networks to millions)

4.0 Data Acquisition System:

This system has the highest requirements for data availability. It is unacceptable for this system to go down during data acquisition, especially since the data may never be available again. This means:

- All data storage systems need to be built with highest level of resiliency (no unplanned downtime)
- Transactions as they occur from the data collection should not be interrupted by any administrative requirements (no planned downtime during data acquisition)

Requirements of allocating storage to this part of the system:

- It is important that one never over run the data being acquired. Therefore growing the data space on line is an important consideration

4.1 Constructing such a SAN Storage System for Data Acquisition

To ensure the data is properly received, we will ensure the following type of system for full resiliency:

- Data acquisition is presented to a high availability and high performance disk system in the primary location (Availability / Performance Management)
- Data acquisition is kept within a separate SAN zone (Asset /Configuration Management) with extra storage for expansion.
- Performance of the entire storage system is configured for the known rate of data acquisition. (Performance / Capacity Management).
- For ultimate reliability, data can be replicated via mirroring to a separate location under 10 km away with separate power and environmental protection. This ensures a failure in the first site does not affect the second site (Availability Management).
- File Systems are grown on the storage at periodic times and if capacity on a specific entity gets to within 90% full (Capacity / Policy Management).

Expanding and allocating storage on this system requires the following

- Easily expanded storage systems, based on real time parameters
- As a next step, automatically expand available storage based on policy (quotas) based on triggers (e.g. file system 90% full).
- Storage which can be expanded in concert first on the remote side and then on the local side to assure that one does not over run the remote disaster recovery site
- Provisions for resynchronization between the storage sites should one side fail.

4.2 SAN based Methods for Allocating Storage on a SAN for the Data Acquisition System:

The main requirements for data acquisition include high availability and the ability to ensure that storage does not run out for the data being acquired. This allocation can be quick (in real time if necessary) depending on how the storage pool and the system have been set up. Some SAN based storage systems today have the ability to quickly (within minutes) present new storage to a specific host. Some of these systems have been called "Storage Appliances".

To expand this picture, let's assume the storage for data acquisition is actually a series of computers, each acquiring part of the data. Let's also assume that some of the data acquisition computers are Windows NT based. Windows NT 4.0 has a property that

multiple servers on the same SAN segment will each try to mark all the storage seen at its own. Refer to the picture below:

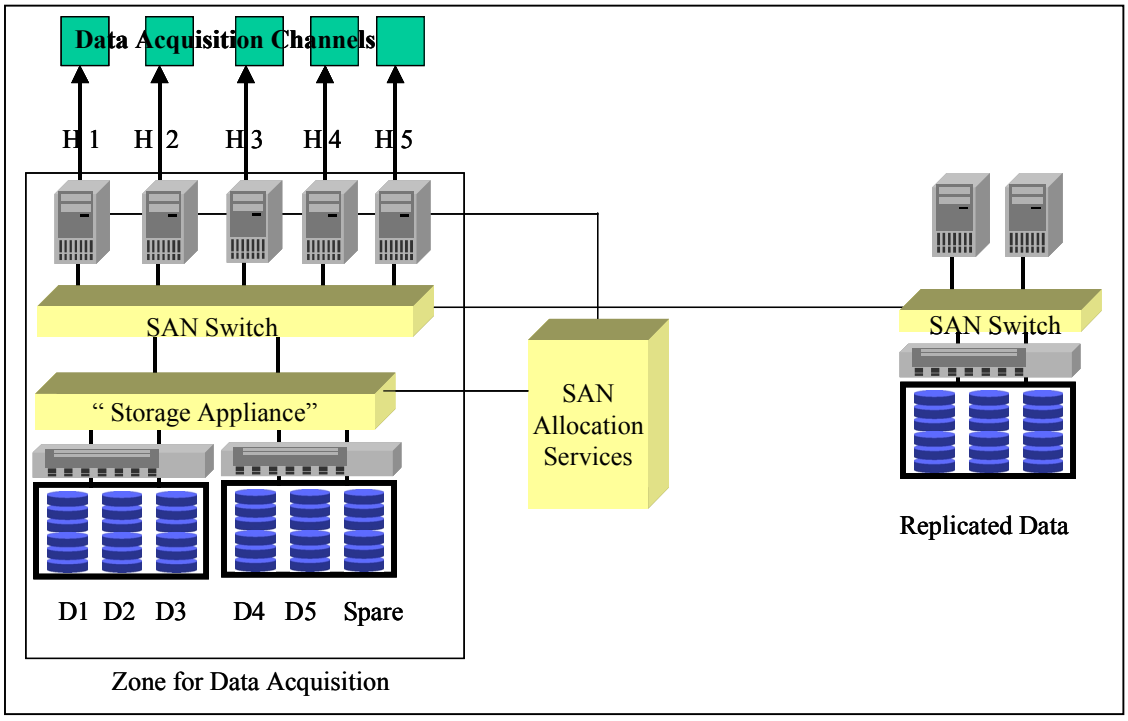

Figure 2: High Availability Data Acquisition System

The configuration addresses these issues:

1. A zone was created around the entire Data Acquisition system to separate it from the rest of the system. This keeps the spare drives to allocate within this data acquisition framework. A zone is similar to a VLAN and is created by the SAN switch (or multiple SAN switches).
2. Each NT host was "bound" to a specific chunk of storage to ensure there would be no crossover from one to another using LUN binding or the Access Control List . This means host 1 (H1) was bound to disk set 1 (D1), host 2 (H2) was bound to disk set 2 (D2) etc. Note that the disk set could be a LUN on the disk system, a virtual disk, etc.
3. Data is replicated from the primary site to the secondary site using the SAN.
4. When more storage needs to be allocated to the primary site, this can be allocated from the "spare" disks in the disk system.
5. There is a "storage appliance" (in band) shown as one method of allocating data automatically and quickly to the servers on the SAN.
6. A storage allocation server is also shown in this function, but shown as an out of band device. Both the "storage appliance" and the storage allocation server can aid in allocation of storage and may be used together.

5.0 Data Analysis System:

The data analysis system in this example is one with two stages of analysis (2 types of analysis computers). A large single server with a large disk system performs the first set of analysis. This is SAN connected to the Data Acquisition Server and Storage. This storage contains the results of the first level analysis of the data (perhaps the conversion from telemetry data to first level graphical data to be processed). We will call this "Stage 1 Analysis".

The second part of the analysis is done on a set of workstations. Note that either a configuration with the workstations being driven by human interaction or ones run in parallel would work for this analysis stage. We will call this "Stage 2 Analysis".

To "off load" the I/O load from the mass storage for the Data Acquisition system, data is copied from the Data Acquisition System to the Data Analysis System over a SAN. A number of different methods for linking the two parts of the Data Analysis system (Stage 1 and Stage 2) are shown including SAN Zone based, SAN File System and as a comparison NAS connected.

A picture of the data analysis system connected to the data acquisition system is shown below:

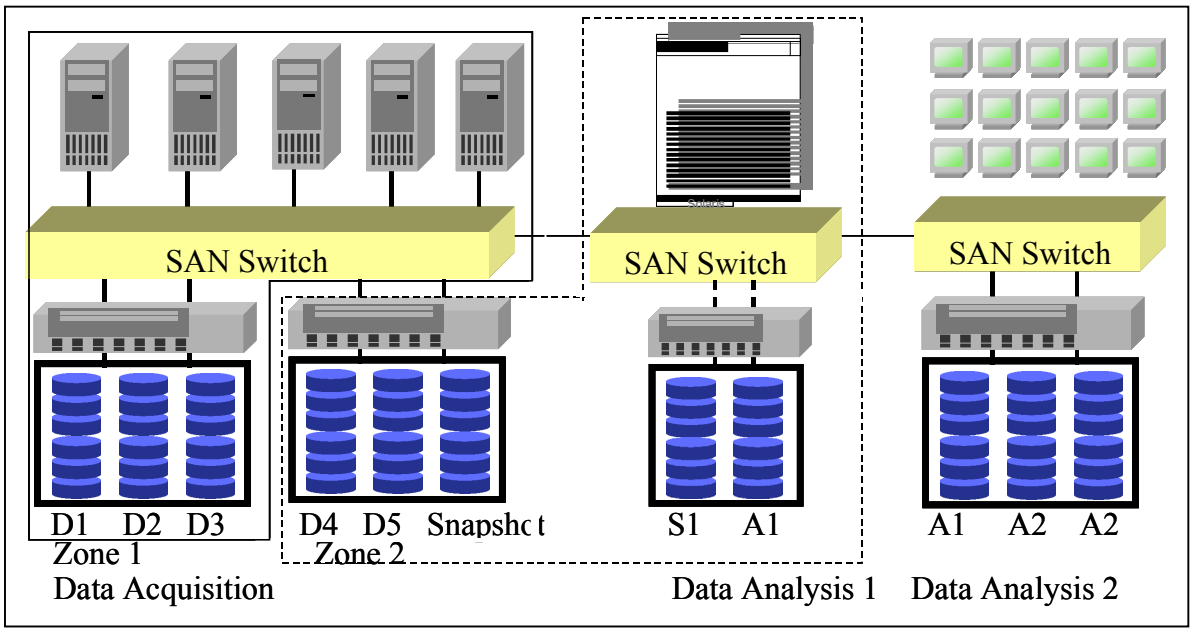

Figure 3: Combined Data Acquisition System and Data Analysis System.

5.1 Sharing Data between the Data Acquisition System and the Data Analysis System:

Once one has the data from the data acquisition phase, and then one has to share it with the Data Analysis system. This can either be done:

- After the completion of the data acquisition phase (data acquisition stops, and analysis starts)
- While data acquisition is occurring, with some way to assure that the data taken for analysis is consistent within itself for analysis.

Satisfying the first requirement means one has to switch the data from data acquisition to data analysis. This can easily be done by rezoning the storage. In this case, we will assume the data acquisition zone is Zone A and the data analysis zone is Zone B. We will also assume 2 different storage systems (SS1 and SS2) being used for data acquisition and one system (SS2, with disks D4, D5 and Snapshot) is now full. The first system (SS1 with D1, D2 and D3) is left in Zone A to continue data acquisition, but the second storage system (SS2) is rezoned into the data analysis zone. Note this is the configuration shown in the diagram.

This means that data acquisition can continue while analysis is done on the data in SS2. Data can then be copied from SS2 to the Data Analysis system.

Often data analysis is required during the continued acquisition of data. In these cases, the storage must in effect be allocated to both the data acquisition server and the data analysis server at the same time. To address this, the storage should:

- Have a stable image for a fixed point in time so the data set being analyzed is consistent and self contained
- Allow for the continued addition of data to the data acquisition database.

The facility normally used for this is called a snapshot. Note that one of the groups of disks in SS2 is configured for snapshot. Two mechanisms are normally used to create this snapshot, break away mirrors or “copy on write logs” A break away mirror could be for example a 3 way mirror. When a “snapshot” of the data is required for data analysis, one of the 3 elements of the mirror is separated from the rest of the mirror and exposed to the other server as a separate volume. This separation is chosen for a specific time when the data is in a consistent state. Given the current example, this might be on the completion of the data acquisition for one full rotation of a planet.

A second method for creating snapshots is based on “copy on write” technology. This technology allows one to write a log of the changes to the data after the point when the snapshot is being taken, In the example above, this would be complete for a rotation of the planet, so the snapshot would be initiated at that point. After the snapshot is started, changes roll into the database as we continue to acquire data. Whenever added data

comes into the database to add or change the data, the previous state of the data is kept in a log.

As an example, one might have a data set with address spaces from 1 to 10 GB. With this data set, one sets a snapshot using a copy on write log, Let's say the first set of changes in the data is for data in the address space above 10 GB. A subsequent read of the log will contain no updates. Therefore the current state of the data also equals the previous state of the data for the address space below 10 GB. A simple read of the total data set will get the information for the state of the data before the snapshot.

A second example would show the snapshot data after an update in the 5-6 GB address space. A subsequent read request comes in for data in the 5 to 8 GB space. Now the log has information about the previous state of the data. The snapshot read of the 5 to 8 GB range will then contain:

- Data from the log showing the previous state of the data in the 5 to 6 GB space
- Data from the rest of the data in the set for the 6 to 8 GB space.

Combining the data from the log (5 to 6 GB) and from the actual file system (6-8 GB) will give a complete picture of the state of the data set at the time of the snapshot.

5.2 Paths between the Data Acquisition System and the Data Analysis System

We have established a number of methods to make the right data available from the acquisition system to the analysis system. In general, if one wished the ultimate safeguard to the acquired data, one has two choices:

- Keep the access to the data from the acquisition system in read only mode
- Copy the data over to separate storage systems in the data analysis section for the best data safety.

The choice between these two methods depends on a number of factors:

- Time to copy over the data for analysis
- If there is a need to change the data in the data set to do analysis
- The amount of storage available within the system.

Allocation of the data in the SAN between the two systems therefore can depend on the following mechanisms:

- Zoning of the data acquisition data into the zone with the data analysis systems
- Use of a possible snapshot to stabilize the data set during acquisition, exposing the data to the data analysis system either as:
    - A broken mirror snapshot (either block or file system)

- A write to log snapshot

Note that with the snapshot mechanism, one may not have to switch the zoning at all to make the storage available. The snapshot provides the isolation between the data acquisition systems and the data analysis systems.

5.3 Allocating More Storage to the Data Analysis System:

Data Analysis Storage needs (both stage 1 and stage 2 analysis) can grow in a much less predictable fashion than the storage for Data Acquisition. The Stage 1 analysis output will normally be fairly well defined, but one might have multiple new analyses being performed at Stage 2. One may for example have an extended set of analyses required at Stage 2, done for a month to finish a project, requiring 100 GB of data to do the analysis. Once the analyses are finished, the storage can be allocated for other uses.

All of this suggests the allocation of storage for Data Analysis needs to be able to support more dynamic changes than need to be supported with the Data Acquisition Section. A suitable system for this might be a high performance RAID system with high sequential throughput to quickly scan and rescan the data for performance. This would be used first for the storage for Stage 1 Analysis. The resiliency of this system would allow for one to keep the results for Stage 1 analysis available for all the computers running Stage 2 analysis.

Distributing this information to multiple workstations doing Stage 2 Analysis could be through 3 methods:

- A Storage Appliance type device, presenting the data as raw partitions and allocating storage to multiple systems. This can either be based on hardware like an EMC or Hitachi subsystem or based on software.
- A global allocation method on a SAN, enabling zoning and disk administration on a SAN.
- A SAN based file system connected through the SAN to make the storage distributed as files in a file system, using the Fibre Channel to distribute data at high speed.

Stage 1 and Stage 2 storage can either be on separate disk subsystems or on the same disk subsystems, depending on the amount of I/O throughput available on the storage systems. Some pictures of alternates are shown below.

How well would each of these 3 solutions address the 6 needs presented earlier (Asset Management, Capacity Management, Configuration Management, Performance Management, Availability Management, Policy Management)

Capacity Management: Storage Appliances can have easy interfaces
Global Allocation could be very good

	SAN file System would have some level of automation
Asset Management Configuration Mgt.	Storage Appliances do well at this Global Allocations do well at this SAN File Systems (N/A, Global Allocator required)
Performance Mgt.	Storage Appliances (depends on the internal architecture, can be slower or faster than the underlying storage) Global Allocation (as fast as underlying storage) SAN File Systems (as fast as underlying storage)
Availability Management	All three solutions can excel or be less than adequate, depending on configuration.
Policy Management	Global Allocator with associated hooks into the software is best for this.

Another possibility would be to transfer the data to a large scalable NAS system and use this as the distribution point to the multiple workstations. This is traditionally how a lot of this type of analysis is done today. The current downside is that the throughput of a NAS system is generally less than a SAN system. In addition, at full throughput, network connected file systems can impose more CPU load on a server than a direct connection like Fibre Channel or SCSI.

6.0 Storage for Archived Records:

Many organizations find that they are in the position of desiring a large number of records on line, but wishing to do so ONLY if the cost is relatively low in providing this facility. In some cases, a NAS solution may be sufficient IF the records can be stored on a shared file system. In other cases where access over a SAN is important, a SAN based storage appliance may be the right solution. A SAN based storage appliance can be built from a standard server and specialized software, presenting data on a SAN as blocks. If constructed this way, this storage appliance can be built from pre existing hardware in the account, allowing re purposing of this existing disk hardware to allow presenting the storage on a SAN.

Data transfer to this repository (which could either be from the Data Acquisition server or the Data Analysis systems (either 1 or 2) would be through the SAN if the rest of the system were SAN based. A NAS solution would be useful for records stored as files, where other parts of the solution are NAS based.

7.0 Summary of Allocation Methods:

This paper has traced a number of methods of allocating storage between various parts of a computer system focused on scientific data gathering / analysis. Discussed within this paper are the following methods:

- Creating a Fully Redundant System with Remote site capability for highly resilient data capture
- Supporting File Based quota notification when capacity is close to being used up
- Supporting Storage Appliances or Global allocation to provide more volume / block based storage as required.
- The data acquisition system is in a separate zone to ensure no conflict with other resources
  - Providing input from the Data Acquisition System to the Data Analysis System
  - Use of Zoning to separate out phases of use
- Using Snapshot techniques to allow on line data capture with on line analysis of the data
- Providing Storage Allocation to the Data Analysis System
  - Use of Global Allocation
  - Use of Storage Appliances
  - Use of SAN Based File Systems in conjunction with Global Allocation

This list indicates there are a number of allocation techniques for storage in a scientific computing system based on SAN storage. As these techniques get integrated into suites of software, one will have a set of tools that will make management of all the parts of the system much easier.

8.0 Products Available from VERITAS:

VERITAS Software has a wide range of software products for managing storage which address many if not all of the requirements including:

1. VERITAS ServPoint Appliance Software for SANs for quick and easy allocation of storage on a SAN (Note, a NAS version is also available)
2. VERITAS SANPoint Control for Discovery and Zoning control. This product is being extended to provide storage allocation.
3. VERITAS Volume Manager enabling mirroring between hosts to allow one to make an exact copy of the data from one host to another. Note this standard Volume Manager works in a SAN environment from a host perspective today. The Volume Manager also has “snapshot” ability.
4. VERITAS File System enabling growth on line (and shrinkage) if required.
5. SAN based file systems for Solaris (SANPoint Foundation Suite) and NT (SANPoint Direct) as one of the methods to distribute data to the workstations
6. SAN based Backup systems (VERITAS NetBackup with Vertex) to create high performance backups on a SAN.