Chapter 2: Related Works

2.1 The Trend of Storage technology

From 1957 the first magnetic disk equipment named *RAMAC* developed by IBM to 1980's, storage technology was just meaning the large capacity and high performance disk technology. As the same of Moore's law for CPU, there is also a breakthrough as MR Head technology for disk large capacity. The technology is keeping the trend of be twice every 1.5 years for a long time. So in 2010, the 3.5 inch disk capacity would be 2 TByte[1].

In1988 RAID (Redundant Arrays of Inexpensive Disks) technology was proposed and enterprise system was commonly used in 1980's. RAID technology is not the former expensive disk like 14 inch used for enterprise system (SLED: Single Large Expensive Disks), but the technology having high performance connected to many 3.5 inch disks used on PC and high reality with redundant. RAID became the storage system core with large capacity cache control technology. In 1992, RAB (RAID Advisory Board) was established for popularizing RAID. Now there are about 40 vendors continuing activities (http://www.raid-advisory.com).

The later half of 1990's was the drawn of SAN and from 2000's it was the spread period in USA. For managing the large data efficiently, some usage like backup and disaster recovery is introduced into storage system. Choose which function how introduce into system and cooperate with server to provide solution for making use of data makes the vendors different from others. About this trend, the storage vendors who provided the same functions to main frame before went ahead, but now lots of vendors like middle range storage vendor, Fibre Channel switch vendor and storage software vendor are developing the technology. For spread and standardization the storage networking technology including SAN, in 1997 SNIA (Storage Networking Industry Association, http://www.snia.org/) was established in USA, and the branch office in Europe and Japan were established in 2001 (http://www.snia-europe.org/, http:// www.sina-j.org/).

With the spread of SAN, Fibre Channel network and the software managing the storage connecting to this network are developed, and are planning to cooperate with the software managing the whole system including IP network. And the more attention is being paid to the virtualized storage technology, the higher the storage management technology importance on software layer is.

With the preparation of broadband infrastructure, it is expected that storage technology will be developed on the field of large region spread of SAN and IP network fusion. And it is also believed that kinds of solutions will be constructed on the infrastructure of storage management software.

2.2 SAN

2.2.1 The definition

SAN is the network using Fibre Channel to connect servers and storages in a narrow sense. The definition of SAN by SNIA is as below[2]:

CONTEXT [Fibre Channel] [Network] [Storage System] [iSCSI]

- Acronym for storage area network. (This is the normal usage in SNIA documents.)
- Acronym for Server Area Network which connects one or more servers.
- Acronym for System Area Network for an interconnected set of system elements.

In this definition, it isn't just limited in the Fibre Channel. If Ethernet is used for the above goals, it would also be defined as SAN. But it is suggested that SAN means Fibre Channel SAN in common such as in this paper. So we will give a more details about Fibre Channel SAN.

2.2.2 The Fibre Channel SAN Environment

Historically in storage environments, physical interfaces to storage consisted of parallel SCSI channels supporting a small number of SCSI devices. With Fibre Channel, the technology provides a means to implement robust storage area networks that may consist of 100's of devices. Fibre

Channle storage area networks yield a capability that supports high bandwidth storage traffic on the order of 100MB/s, and enhancements to the Fibre Channel standard will support even higher bandwidth in the near future.

Depending on the implementation, several different components can be used to build a Fibre Channel storage area network. The Fibre Channel SAN consists of components such as storage subsystems, storage devices, and server systems that are attached to a Fibre Channel network using Fibre Channel adapters. Fibre Channel networks in turn may be composed of many different types of interconnect entities. Examples of interconnect entities are switches, hubs and bridges.

Different types of interconnect entities allow Fibre Channel networks to be built of varying scale. In smaller SAN environment, Fibre Channel arbitrated loop topologies employ hub and bridge products. As SAN increasing in size and complexity to address flexibility and availability, Fibre Channels switches may be introduced. Each of the components that compose a Fibre Channel SAN must provide an individual management capability, and participate in an often complex management environment.

Due to the varying scale of SAN implementations described above, it is useful to view a SAN from both a physical and logical standpoint. The physical view allows the physical components of a SAN to be identified and the associated physical topology between them to be understood. Similarly, the logical view allows the relationships and associations between SAN entities to be identified and understood.

2.2.3 The Physical View

From a physical standpoint, a SAN environment typically consists of four major classes of components. These four classes are:

- End-user platforms such as desktops and/or thin clients;
- Server systems;
- Storage devices and storage subsytems;
- Interconnect entities.

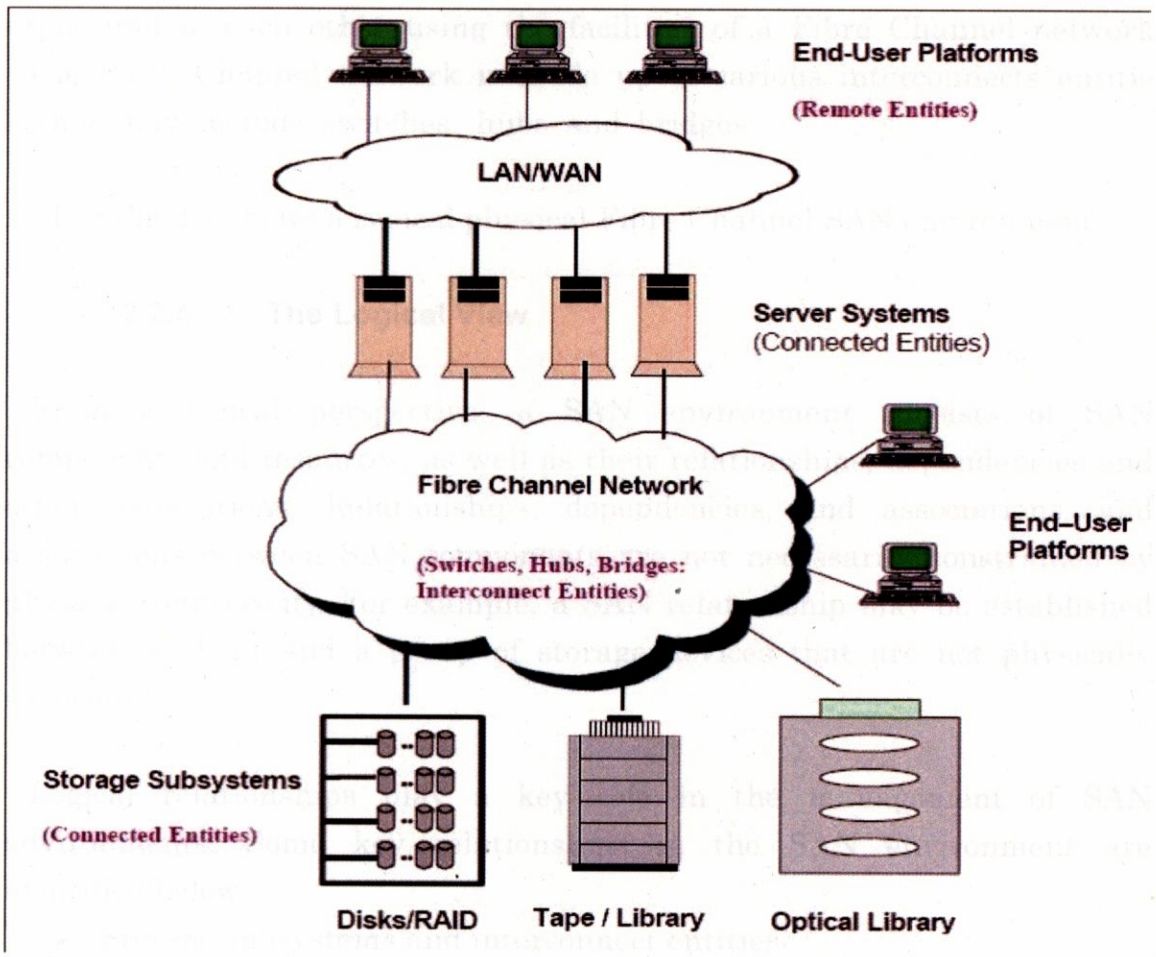

Figure 2-1 The Physical View of SAN

Typically, network facilities based on traditional LAN and WAN technology provide connectivity between end-user platforms and server system components. However in some cases, end-user platforms maybe attached to the Fibre Channel network and may access storage devices directly.

Server system components in a SAN environment can exist independently or as a cluster. As processing requirements continue to increase, computing clusters are becoming more prevalent. A cluster is defined as a group of independent computers managed as a single system for higher availability, easier manageability, and greater scalability. Server system components are interconnected using specialized cluster interconnects or open clustering technologies such as the Fibre Channel – Virtual Interface mapping.

Storage subsystems are connected to server systems, to end-user platf

orms, and to each other using the facilities of a Fibre Channel network. The Fibre Channel network is made up of carious interconnects entitie s that may include switches, hubs and bridges.

The Fig.2-1 shows a typical physical Fibre Channel SAN environment.

2.2.4 The Logical View

From a logical perspective, a SAN environment consists of SAN components and resources, as well as their relationships, dependencies and other associations. Relationships, dependencies, and associations, and associations between SAN components are not necessarily constrained by physical connectivity. For example, a SAN relationship may be established between a client and a group of storage devices that are not physically co-located

Logical relationships play a key role in the management of SAN environments. Some key relationships in the SAN environment are identified below:

- Storage subsystems and interconnect entities;
- Between storage subsystems;
- Server systems and storage subsystems (including adapters);
- Server systems and end-user components;
- Storage and end-user components;
- Between server systems;

As a specific example, one type of relationship is the concept of a logical entity group. In this case, server system components and storage components are logically classified as connected components because they are both attached to the Fibre Channel network. A logical entity group forms a private virtual network of zone within the SAN environment with a specific set of connected entities as members. Communication within each zone is restricted its members.

In another example, where a Fibre Channel networks is implements using a switched fabric, the Fibre Channel network may further still be broken down

into logically independent sections called *sub-fabrics* are again divided into *regions* and *extended-regions* based on compatible service paraments. Regions and extended regions can also be divided in to partitions called zones for administrative purposes.

2.2.5 The Management of SAN

With SAN appeared, the Fibre Channel network was commonly used. But it was not used in the traditional data center, it needed special SAN management software. And it is lack of the idea of storage management in the traditional network management, so the traditional network management software is not enough to introduce IP-SAN, and has to introduce new one. Some of the functions of SAN management can be mentioned as below:

- The management for SAN construction (discovery, topologies, error check, etc.)
- The security management (Zoning, LUN[Logical Unit Number] security)
- Capacity planning
- Performance management
- The management for apply data (apply backup, apply discovery, apply moving data)

We will show more details about the management for SAN construction and security management because they are the necessary items for making use of SAN. The others are not necessary but important for managing the data with high performance.

2.2.6 The Management for SAN Construction

For making using of SANs, it is important to grasp how are servers, storages and switches, etc. are being connected. In SAN, the function to grasp the connection details named *discovery function*. There are the method of requesting Fibre Channel Switch for construction details, the method of finding the equipments connecting to the servers Fibre Channel ports, the

method of finding the equipment connecting to the SAN management servers Fibre Channel ports and using some protocols like SNMP (Simple Network Management Protocol) with LAN to ask Fibre Channel Switch for construction details in *discovery*. But because there are some complex constructions like one server connecting sever HBAs, one RAID with tens of Fibre Channel port, and one port with tens of volumes, it is almost impossible to grasp these complex constructions using the methods above. For example, it is possible to grasp HBA ports that connecting to switch ports from Fibre Channel switch, but can't grasp the server hostname what connecting to the HBAs. So it is impossible to make a difference between single server or multiple server if there are multiple HBAs. It is need multiple methods mentioned above to specify the constructions.

2.2.7 The Security Management (including Zoning, LUN security)

Multiple servers couldn't share the same storages in the DCS with traditional SCSI. Server operation can't act unless all the storages resources connected are only for itself specially.

Comparing to that, in SAN the data may be broken for multiple servers sharing the same data in the above condition. It is better that if the server operation would keep up with SAN and hold the function to solve address competition. But in fact there is almost none holding this function. The reason is that to the server operation, it need to be trusted the same as the traditional SCSI to introduce SAN easily.

For solving the address competition, there is a special function can't find in Ethernet but in Fibre Channel layer named Zoning. There are special ID named WWN (World Wide Name) for every ports in Fibre Channel. With this WWN base, multiple ports are divided into groups named zone, and the exclusive function between zones is hold by the Fibre Channel switch. For example, there are server A and server B, and storage C for A, storage D & E for B. Now set server A as zone 1, server B as zone 2, storage C as zone 3, and storage D & E as zone 4. Set the switch so that packets can be transmitted between zone 1 and zone 3 but there is no permission to transmit packet between zone 1 and zone 4. With above, server A could access storage C but can't access storage D & E, so the data belong to server B can't be destroyed

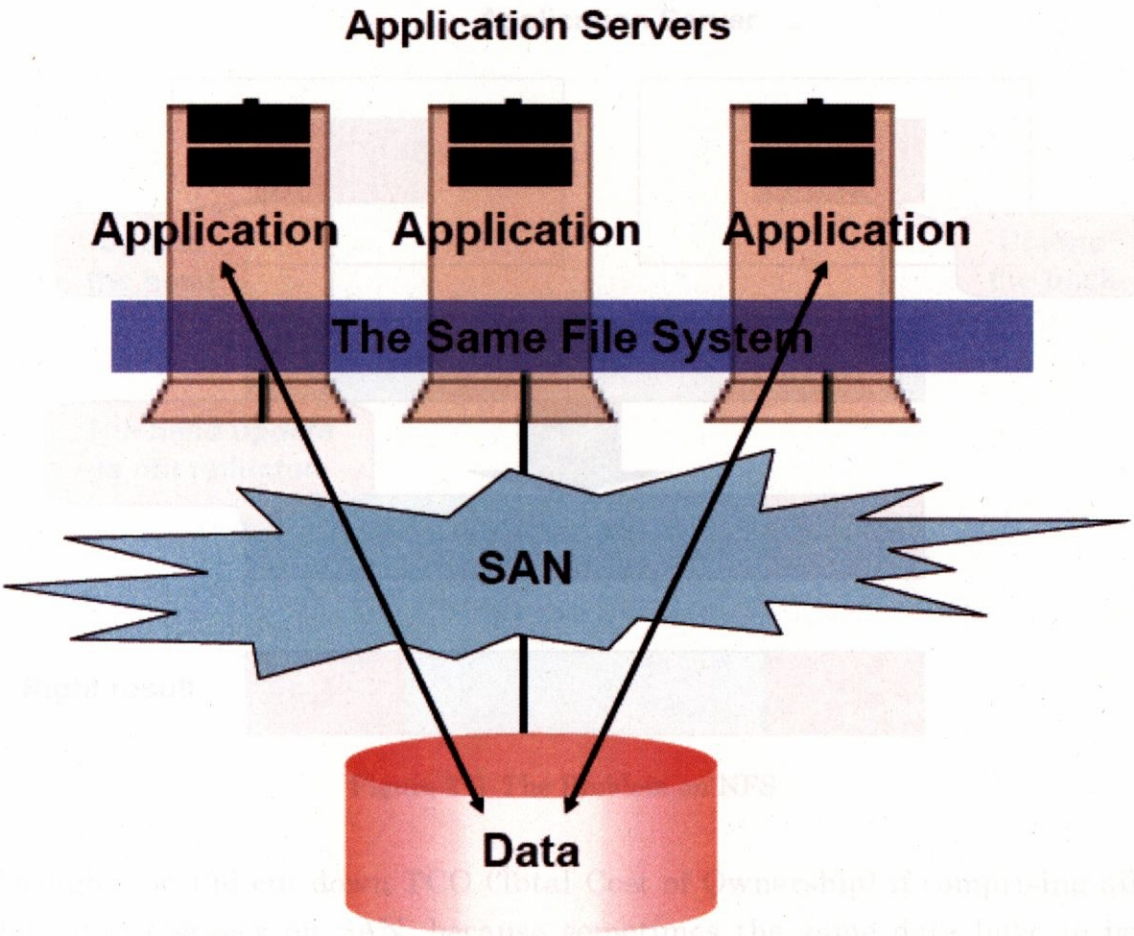

Figure 2-2: The Construction of Cluster System

by server A. And of course, it is set that server B can't access storage C.

Though the zoning function could restrict the access in port level, if there are several volumes connecting to one port, it wouldn't restrict access in volume level. So there is a function named LUN security function holding the access exclusive function in volume level in storage equipments. WWN is the Fibre Channel's identifier while LUN is SCSI's. So for Fibre Channel, this function can't be acted in the switch can't be explained by high level protocol of SCSI protocol. But the storage equipments could combine the WWN for Fibre Channel and LUN for SCSI protocol to execute the access restriction. With the combination of LUN security function and zoning function, it would be executed access exclusive restriction in a high practical level.

2.3 Cluster File Systems

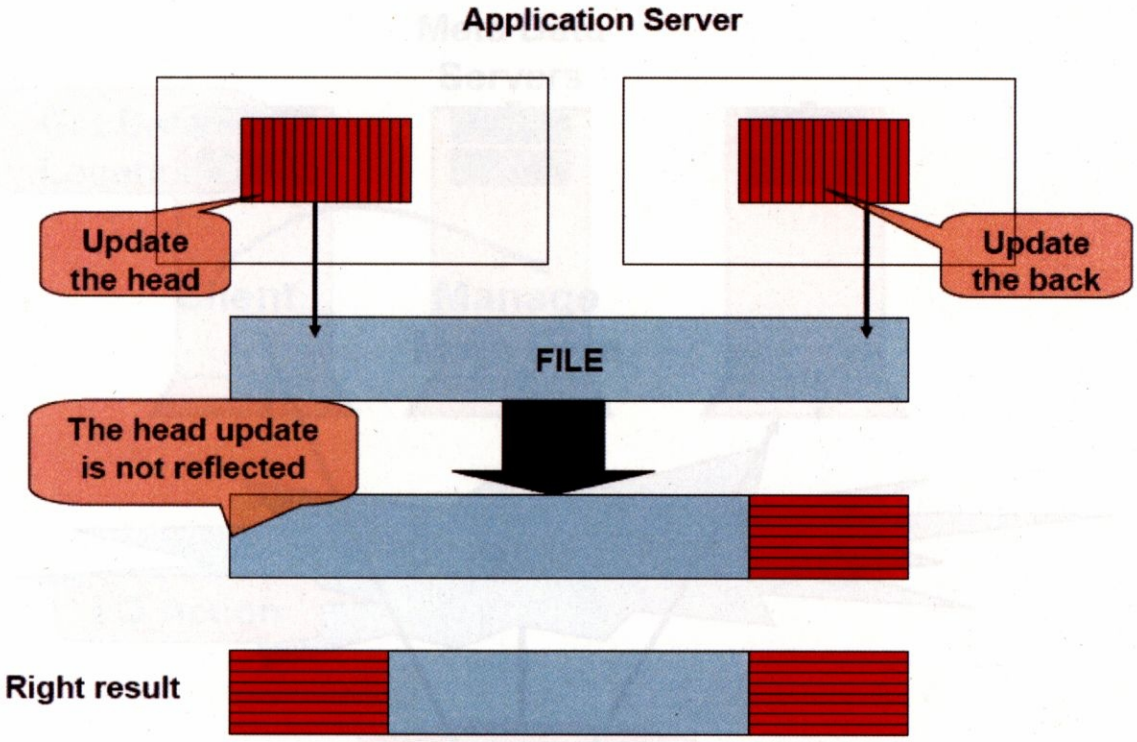

Figure 2-3: The Problem on NFS

Though it would cut down TCO (Total Cost of Ownership) if comprising all data into storages on SAN, because sometimes the same data have to be shared by kinds of works, the data can't be used efficiently if the data can't be shared during multiple servers on SAN. It will give a view of the method for using data efficiently.

2.3.1 The technology of Cluster File System

As the solution for sharing data, there are some products that can copy data without network and instantly copy in storage have been developed. At the same time, the file systems for the servers on SAN to share the data in storage directly are also developed. As Fig.2-2, with multiple servers mounting one file system at the same time and exclusive controlling, it is possible to access consistent files. And the servers are clustered.

These file sharing products have to guarantee consistency. For example, like Fig2-3, if multiple servers update the file at the same time in NFS (Network File System), one update would not be reflected even with different

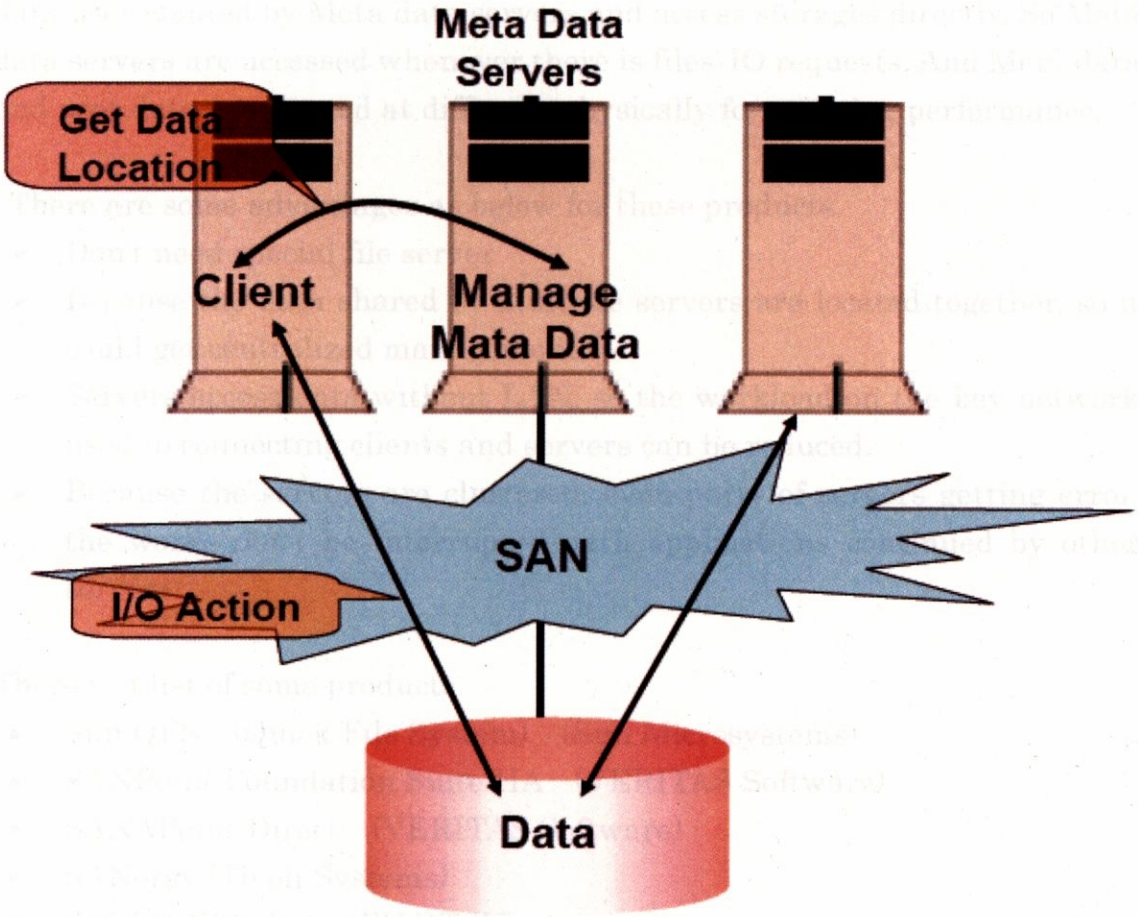

Figure 2-4 Meta Data Control

update place and the file may be destroyed.

And every server has to hold data cache for guaranteeing performance. When one server updating its data, with cooperation during servers, all servers have to update their cache and need the control for keeping consistency. For example, in NFS, because the information will left on client's cache after being read once, the updated information can't reflected on other clients.

Like the Fig.2-4, the file systems introduced in this part hold Meta data servers. The Meta data is managing that where is the user data keeping on storages or where is storage having free space. The programs with the client function that can cooperate with Meta data server are introduced into each server. When servers get the files IO requests, the client programs are started up via kernel. With the programs, servers are noticed if there is the

data user wanted by Meta data servers and access storages directly. So Meta data servers are accessed whenever there is files' IO requests. And Meta data and user data are located at different physically for securing performance.

There are some advantages as below for these products.

- Don't need special file server
- Because the data shared by multiple servers are located together, so it could get centralized management.
- Servers access data without LAN, so the workload on the key network used to connecting clients and servers can be reduced.
- Because the servers are clustered, even parts of servers getting error, the works don't be interrupted with applications continued by other servers.

There is a list of some products.

- Sun QFS (Quick File System) (Sun Microsystems)
- SANPoint Foundation Suite/HA (VERITAS Software)
- SANAPoint Direct (VERITAS Software)
- SANergy (Tivoli Systems)
- SafeFILE/Global (FUJITSU)

2.3.2 The Product Example ---- SafeFILE/Global

SafeFILE/Global is listed here as the product example materialized the cluster file system for sharing data. It is worked on SPARC/Solaris and is developed by FUJITSU Institute using HAMFS technology.

① **The Exclusive Control on Block level**

This product is using exclusive control on block level to protect data from being destroyed. The use right to each server cache is controlled by publishing file exclusive token to each file access. This information is operated by the next 2 modules.

- ✓ The Meta data server (MDS) accessing Meta data and controlling token
- ✓ Access Client (AC) requested token by each server file system

The file name and the information of using block limit are include in read and write token, so different blocks can parallel access the same file. And user programs don't need to know MD and AC because they are included into file system software.

② **The Cache Control**

Management information and user data are cached on each server memory, so the access to data volume and the communication between servers is controlled to hold to a minimum. And for the file being written on cache, if there is write or read requests from other servers, the cache contents is copied to opponents using the either way below to keep the cache consistency.

- ✓ Using the read token as a chance to reflect the data on disk and read again. The hardware function like cache bind of high performance storage will protect form performance deterioration.
- ✓ Transmit directly via network. Performance deterioration will be avoided by speedy inter-connector.

③ **Meta data control**

Usually if Meta data is directly shared by servers, the performance will be declined for competition by the same time update from servers. To get speedy update process of Meta data and high reliability, the transaction control using on database is materialized in SafeFILE/Global.

- ✓ MDS just update the Meta data on cache but not update the Meta data volume during the transaction process
- ✓ At the moment finished all processes without error, the Meta data had been changed on cache are picked up in byte unite, written into log volume together in one I/O process and returned to client as reply.
- ✓ If there is a certain data being kept in log volume, the changed Mata data left on cache will be written into Meta volume. This writing is done asynchronous with Meta data update, and the process will done at once unless the log volume is full
- ✓ When system rebooting, Meta volume will be restored by using the restore data in log volume

With this way, the Meta data volume writing will be reduced, and the Meta data consistency will be kept with the Meta data update process done great speedy.

④ **Multi-Volume Support**

Multiple disks are controlled independently and the disk will be allotted with the workload on each disk is the same. So it is difference from the traditional file system and not need to adjust the workload between disks by user.

And if new disk is need for the data is creasing, the new data volumes can be added into file systems and get the workload averaged without stopping the applications under execution, so the data transitions formerly used is not necessary.

2.4 NAS

What is NAS? From the name, it's Network Attached Storage. The network for SAN means special network constructed by Fibre Channel. While the network of NAS means IP network that are commonly used. As storage, there should be hosts using NAS via network and the data is exchanged in file level. So from the view of basic function, it would be said NAS is just file system. Maybe it is confused by the name of Network Attached Storage, from the traditional idea before the technology of storage network spread, NAS should be thought as server (computer) rather than as storage. But now it is acknowledged as one form of storage equipment because the form as an exclusive file server machine that delete everything except necessary function. This part will give a review of NAS.

2.4.1 The History

As said above, the basic NAS method, what is sharing the file on network, is commonly used for a long time. So the equipments classified as NAS have existed for a long time. But of course, it was not called NAS and used just in a limited way for without network spread.

But after that, the network was commonly used, and the transmit speed raised rapidly. The infrastructure was prepared. And more, with the background of data increased massively and storage market expanded suddenly, NAS was paid more and more attention and established the definition. Also, with this process, the technology got great developed.

2.4.2 NAS Technology

From the idea that basically NAS is file server, the basic function of NAS can be materialized by the traditional server technology. But NAS holds some advantages of using exclusive file servers. This part will show the advantages of NAS as exclusive machines and the functions with these advantages.

- **The server technology supporting NAS**

The way of thinking file servers on network has exist a long time, and the general-purpose servers were acted as file servers at most cases. So if thought NAS as file servers, the basic NAS functions can surely be materialized by using the traditional server software technology.

In fact, NAS is making using of the traditional server technology. Above the NAS materialized the basic function, a very important thing for the files sharing technology like NFS and CIFS is that it should be OK even the servers and clients are different architecture computers. So what is the structure?

This function is introduced in server-client machine no matter NFS or CIFS. What means that the function that allow other computers access the file system itself who provide the file service and the function that make the file systems accessed by the computer itself who getting the file service. And the interface is clearly defined. So if it is fit the interface, files could be exchanged even though with the difference of operation systems, hardware and file systems. If we have a well thinking, we will find that even as UNIX operation systems, in face there are some kinds with different hardware interface and everyone is supporting multiple different file systems. In NFS

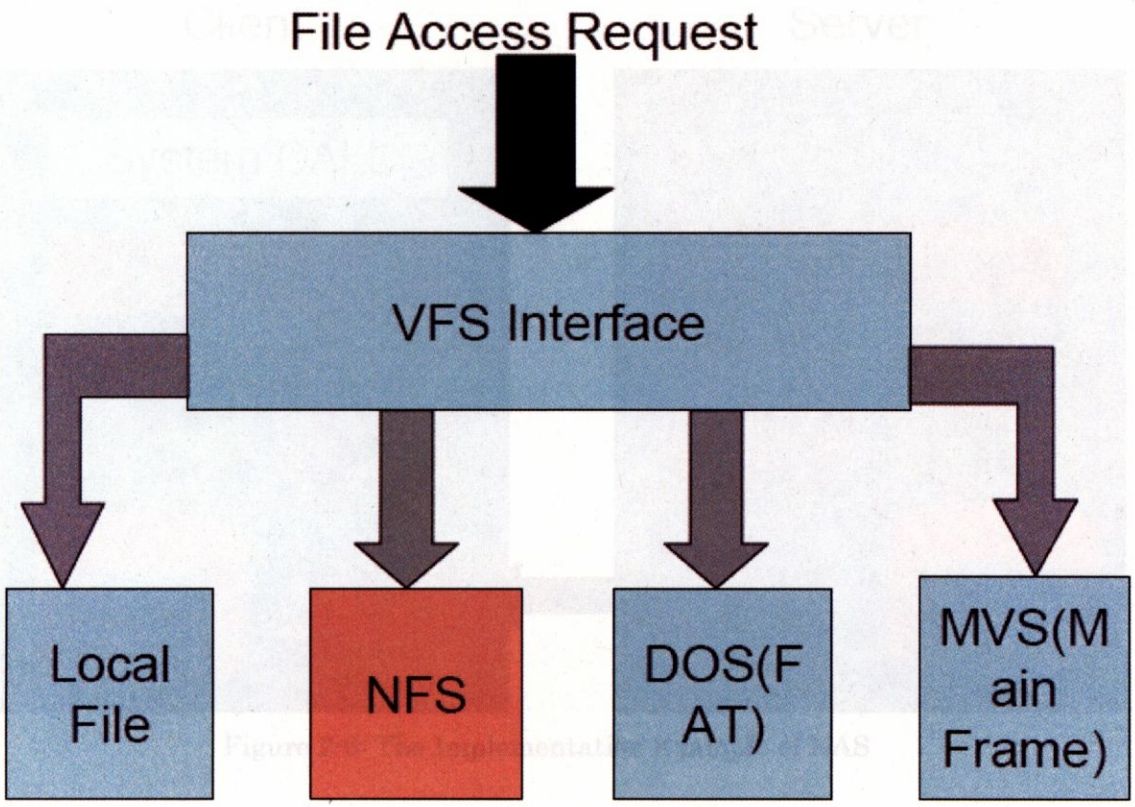

Figure 2-5: The Action of VFS Interface

different files could be shared is also for this structure.

Especially in UNIX operation systems, a common interface named VFS (Virtual File System) is originally used to access file systems, and would be used on different file systems (Fig.2-5). This structure is also used NFS. If it is UNIX system, the basic idea is the same no matter what kind the operation system is. So files would be shared easily between different operations via VFS.

Like explained above, if the client machine is UNIX system with introduced NFS interface, the computer could be a file server no matter what architecture it is in the computer (Fig.2-6).

The basic thinking of using CIFS is the same as that of using NFS. Files could be shared between different kinds of computers via CIFS interface. There is software named Samba as example. It means introducing the file sharing protocol for Windows on UNIX operation systems. With introducing

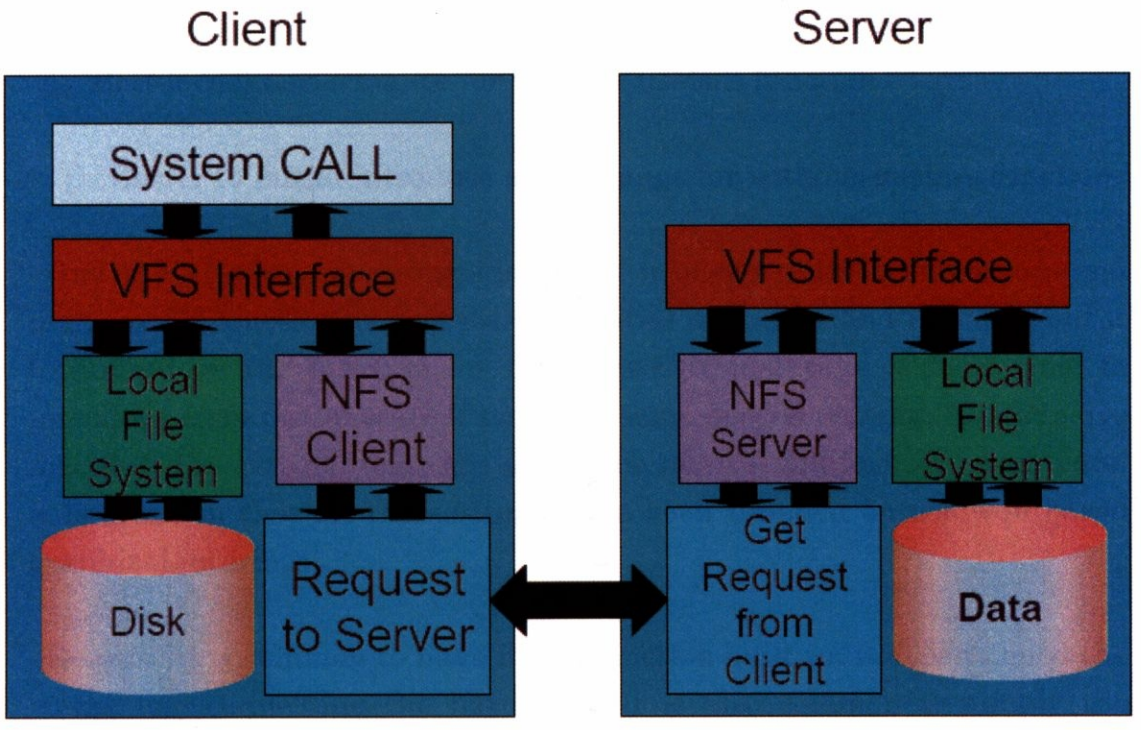

Figure 2-6: The Implementation Example of NAS

this Samba into UNIX servers, the file systems on UNIX servers could be used by Windows client machines. This means that the UNIX servers are able to be used as the file servers for Windows client machines.

Now thinking about NAS again, NAS are using lots of conjunction could cover the difference between the architecture of servers and clients via the file sharing interface on network. At most case, NAS could provide the same file service to UNIX systems and Windows systems though itself is not UNIX systems or Windows systems. At the same time, with turning performance and introducing kinds of function, NAS have the liberty couldn't get from the traditional technology.

- **The advantages comparing NAS to general server**

What are the advantages comparing using exclusive servers of NAS to using general servers? They will be listed mainly as 4 points as below:

① Easier to initial introduce and manage for without useless functions

② Easier to get high performance with special turning

③ Higher reliance for the structure simplification of hardware and

software

④ Better cost performance for without useless function

① Easier to initial introduce and manage for without useless functions

One of NAS's big advantages is that it is easier to initial introduce and manage. Of course it is necessary to initial introduce and manage, but it is limited as the function for file server. It just take a short time to introduce from initial state if not need some special setting. As exclusive machine, the setting items are able to be more simplification. And more, it is common that all the settings have been finished on the web based graphical interface.

Because it is common to have RAID function built-in, it doesn't need the works like connecting the hardware. It is also common to simply the settings on RAID level.

Comparing to above, it is necessary to initial introduce the server and do some management when using general servers. It is not only just to set the file server function, but also need to set everything for file service. Of course it needs special knowledge for server management. And with the idea of preserving data, it still need RAID function. For this, there are the methods of using software RAID or the method connecting external RAID equipments, but both need the works for adding hardware, setting RAID function and the management.

② Easier to get high performance with special turning

As exclusive machine, it is easier for NAS to get higher performance with turning through all the necessary processes for file service. It is necessary for some processes to cooperate to provide file service. These processes are getting request from client; the network I/O and the protocol process for sending data; file system process to access file; actual disk I/O and RAID process if RAID is being used. With all these above, it is possible to do optimization.

While about the general server, it is certainly impossible to get

optimization only focus on file service. Because no matter network I/O or file service, they have to be considered as being used independently.

As exclusive machine, it could get better optimization than general server even the optimization is difference depend on products. The result through optimization is that comparing to general server, there are some advantages like better tolerance to get centralized access, and less response time with more clients on NAS.

③ Higher reliance for the structure simplification of hardware and software

The necessary hardware and software are much limited if just considering file service. That means all the equipments are able to get simplification as exclusive machine. Simpler to create, less possibility to have bugs and have accident on hardware. So it could have higher reliance.

④ Better cost performance for without useless function

It's also much benefited by attempting the whole equipments simplification. Reducing the useless compositions certainly has the benefit of cutting down the cost. And because the turnings for special service, the machines having the hardware with lower specifications than general servers are able to get the same performance.

In the NAS's advantages mentioned above, easier to manage is a very important point. Like clamored loudly every year, the management of storage system is more and more difficult with the data (storage capacity) rapidly increasing. In this environment, easy management is a very important character. More, the high scalability connecting the network is a big advantage for the several service supporters using network, so NAS is commonly welcomed by this field.

2.4.3 The Functions as the NAS advantage

Using special journal file system is usually mentioned as one of NAS technical advantages. Usually in journal file system, it is not only making sure the file consistency, but also interested in the speedy reboot at the same time when the abnormal stop. The normal file will be checked file consistency when rebooting, but it will take a long time if the size of file system is large, even more the file system consistency can't be backup sometimes. For the character of file server, it is usually using journal file system to secure the conservation of data in NAS.

Usually in journal file system, the Meta data of file system is materialized by saving the Meta data in other field named journal before the Meta data is actually being updated. But the file itself may be lost when abnormal stop because it just save the Meta.

To solve this problem, it is even said exclusive NVRAM is loaded to save the write data in NAS. In most case, the technology watched in RAID is loaded to make sure not lost file when abnormal stop. With loading this structure, NAS is concerned to hold the effect to minimum to clients in case abnormal stop.

The other technology as NAS advantages is snapshot. The basic function of snapshot is to save the image of file system at one certain time point. It is mostly used for data backup. If it needs to backup data saved in NAS, it is enough to take the snapshot at that time point. Even the file service is continued after that, the image of snapshot wouldn't be affected, so the image for backup is being kept. The online backup is possible with copying the data from the snapshot image to a second device such as a magnetic tape later. The other use of snapshot is that it is possible to roll back the former contents in file level in usual use. That means, even the file is lost for user mistake or software reason, it would be restored instantly using the snapshot image.

How is the technology materialized? The famous NAS of product from Network Appliance, Inc. will be listed as an example (Fig.2-7).

In the case of Network Appliance's product, it is used a special journal file

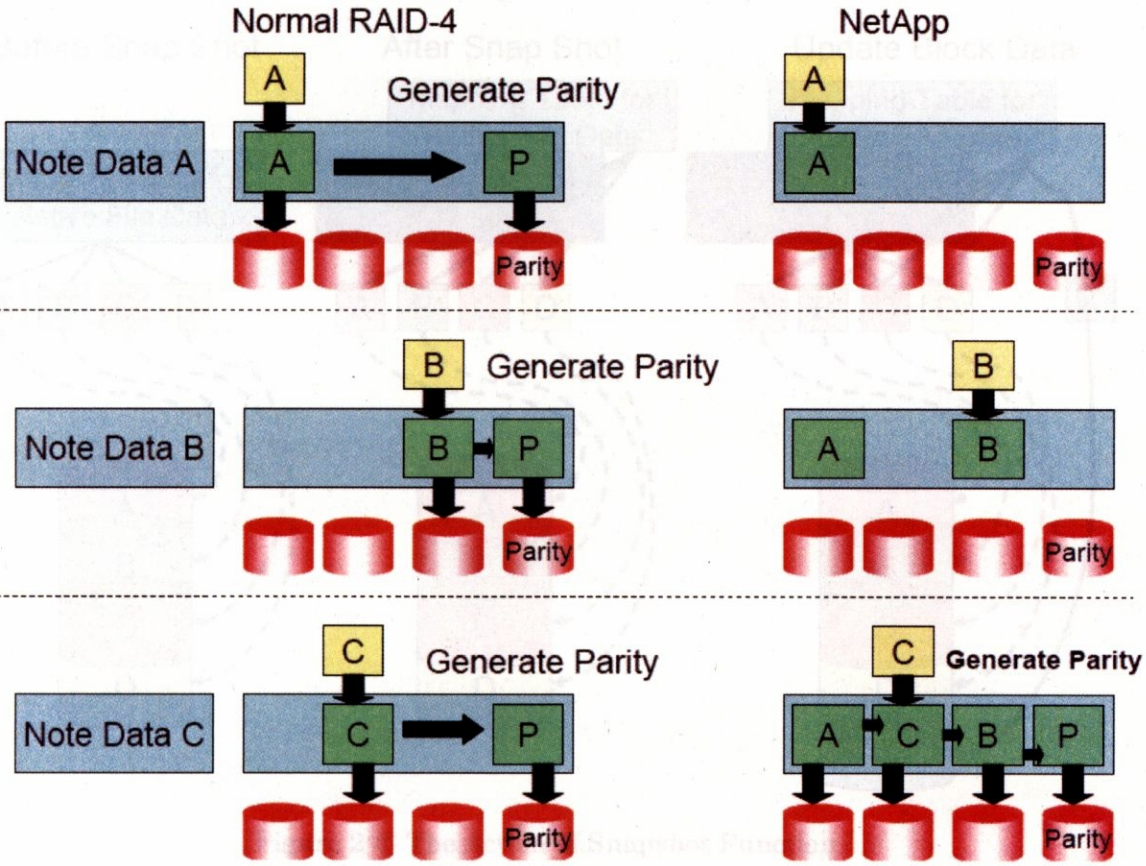

Figure 2-7: The RAID-4 Theory of Network Appliance Product

system named WAFL (Write Anywhere File Layout). The big character of this system is that it is loaded with closed cooperating with RAID structure. Using RAID-4 that is rarely used as RAID is its peculiarity. In RAID-4 and RAID-5, the write data must be updated with parity data together. In the case of RAID-4, the parity data is limited into the parity disk, so the accesses are centralized on the parity disk when writing data. If making the structure of RAID-4 with 4 disks (3 data disks and 1 parity disk) as an example, the workload on the parity disk is 3 times as it 1 disk average. For this drawback on RAID-4, it is usually to use RAID-5 with dispersed parity data.

The data on disk wouldn't be overwritten even being updated is the other peculiarity of WAFL.

This method is very important to materialize the function of snapshot. With his method, the former data is left on disk in fact though it should be overwritten as file data.