

Oceanspace S2000 Series Storage System Technical White Paper



Huawei Symantec Technologies Co., Ltd.

All rights reserved.

Contents

1 Overview	3
2 Advanced Architecture	3
2.1 Advanced Multi-Core Processing Chip	4
2.2 Leading Bus Technology.....	4
2.3 Independent Storage Space for the Operating System	4
2.4 Fully Redundant Channel for the Interconnection between Dual Controllers	5
3 All-Round Data Security Protection	5
3.1 Active-Active Dual Controllers Technology	5
3.1.1 Fully Redundant Channel Design.....	6
3.1.2 Cache Mirror Technology.....	7
3.2 Data Safe Box	9
3.3 Pre-Copy Technology with Extra Safe Hard Disk.....	12
3.4 Global Hot-Spare Disk	14
4 Advocating Energy Saving Design	15
4.1 Increasing Space Usage through Careful Design	15
4.2 Integrated UPS	16
4.3 Integrated Controller and Hard Disk Subrack.....	16
4.4 Intelligent Hard Disk Power-on	17
5 Concerning Your Power Supply Environment	17
6 Acronyms and Abbreviations	18

1 Overview

Through years of continuous investment and hard research, Huawei Symantec has accumulated rich experience in the storage field. It carries out thorough analysis on customer requirements by tracking the development trend of storage technologies and the IT industry. It successfully launches the new storage series product, Oceanspace S2000, which combines the energy saving, high reliability, high availability, and easy-to-manage design concepts based on the mature IT technology architecture.

The Oceanspace S2000 is designed in two product forms: single control and dual control. It provides iSCSI, SAS, and FC host ports, and supports SAS and SATA at the back end. Through flexibly configured host port or hard disk type, it supports FC SAN, IP SAN, and SAS networking modes to meet various storage requirements.

The Oceanspace S2000 series storage system includes S2100 and S2300, where S2100 includes S2100 single controller and dual controller, and S2300 includes S2300 single controller and dual controller. The Oceanspace S2000 dual controller ensures the data availability by using perfect data protection technologies, such as Active-Active dual controller, Cache mirror, data safe box protection, hard disk pre-copy, global hot spare disk, and integrated UPS. Meanwhile, the VTL3605 supports both AC and DC power supply to accommodate your equipment room.

The Oceanspace S2000 series storage system guarantees the data with the following features:

- Architecture based on an open standard
- Fully redundant hardware design, ensuring the reliable operation of the equipment.
- Integrated UPS with the controller, ensuring that data can be written and saved in hard disks in the case of power failures.
- Hard disk pre-copy function, avoiding risks of RAID group failure and data loss.
- Available iSCSI, SAS and FC host ports, supporting SAS and SATA hard disks.

2 Advanced Architecture

The Oceanspace S2000 series storage system is designed in the Active-Active dual-controller architecture. It uses the full 64-bit system platform, provides greater system bus bandwidth, supports 64-bit operating system, and is compatible with 32-bit operating system. Its full redundancy design ensures high reliability. Figure 2-1 shows the logic architecture of the S2000 series storage system.

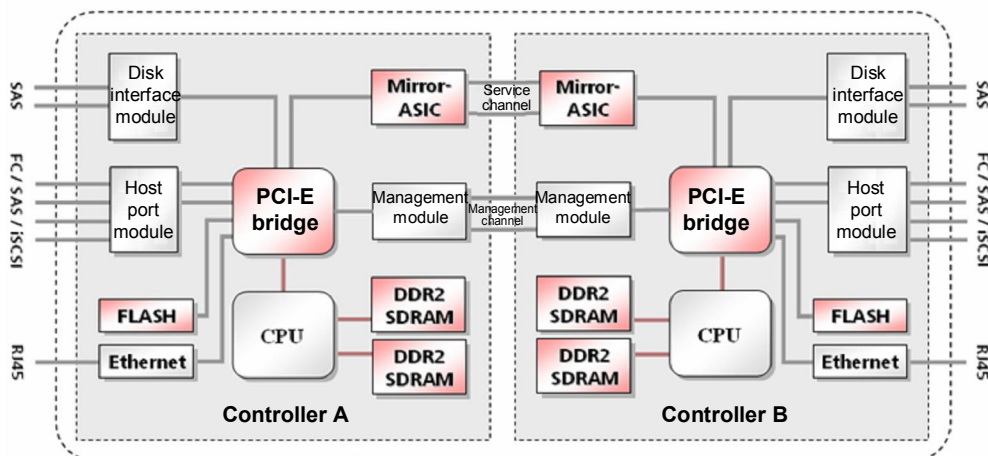


Figure 2-1 Logic architecture of the S2000 series storage system

2.1 Advanced Multi-Core Processing Chip

The core processing unit of the controller in the Oceanspace S2000 series storage system uses a multi-core processing chip and integrates a memory controller. The DDR2 SDRAM memory featuring dual channels and high performance exchanges data with the CPU directly, making the traditional bridging circuit unnecessary. This expands the access bandwidth between the memory and the CPU, and improves the access efficiency through shortening the bridging path.

The application of multi-core processing chip provides the system with good software extension capabilities, for example, flexible support for RAID levels and continuous expansion of value-added software.

2.2 Leading Bus Technology

The Oceanspace S2000 series storage system uses an advanced PCI-E technology. Compared with the last generation PCI-X bus technology, the PCI-E technology provides the system with a higher bus bandwidth, which reaches 10 GB/s.

2.3 Independent Storage Space for the Operating System

The operating system of the controller is stored in the FLASH memory that enjoys a special bandwidth. This improves the data transmission efficiency between the system and the CPU and ensures the fast response of the system. In addition, this makes it unnecessary to occupy additional storage space in the hard disk. Fast system startup speed minimizes the impact of system restart on services. The system uses an industrial FLASH memory, which can support tens of millions of read and writing operations within the life cycle. This ensures the high availability of the controller.

2.4 Fully Redundant Channel for the Interconnection between Dual Controllers

The design of each component in the S2000 series product shows the high availability. The Cache mirror data service channel and management information channel between dual controllers adopts fully physical redundant design. The Cache mirror channel adopts a special mirror-ASIC chip for acceleration. This redundant design in physical class guarantees the security of the channel between dual controllers.

3 All-Round Data Security Protection

3.1 Active-Active Dual Controllers Technology

In the storage system with two or more controllers, FC, iSCSI, SCSI, SAS or NAS port are available. The following two working modes are available for the controllers.

- **Active-Passive working mode**

The Active-Passive mode is also called active-standby mode, in which only one of the two controllers is in active state and acts as the active controller to process I/O requests from application servers, and the other controller is in idle state and acts as the standby controller. In case the active controller is faulty or in offline state, the standby controller takes over its services.

- **Active-Active working mode**

In AA (Active-Active) mode, both controllers are in active state and process I/O requests from application servers (ASs). In case one controller is faulty or in offline state, the other controller can take over its services and the services are not affected. Thus, the AA mode can guarantee the high reliability of the system through mutually redundant backup of the two controllers. In addition, the AA mode can balance services, fully utilize resources, and improve the system performance.

The Oceanspace S2000 series storage system fully supports the AA working mode.

If a controller fails, for example, the link connected to the controller fails, the service on the failed controller can be switched over to the other controller. After the link recovers and then the failed controller works properly, the controller can continue to control the previous services. In the entire process, the switchover of services is transparent to you. You see a link failure for a short time and then link recovery on the host. In addition, the continuity of the storage services is guaranteed, and the requirements of the storage services for reliability and data consistency are met.

Storage services are shared on two controllers. This prevents the condition that excessive load is on one controller but the other controller is idle for a long time. Thus, the load on one controller is reduced, system resources are used more effectively, and the working efficiency and performance of the system are improved.

3.1.1 Fully Redundant Channel Design

In the design of dual controllers, the service data and management information between the two controllers are interacted through the service channel and the management channel. The available design schemes are as follows:

- No independent service channel and management channel is available, and the hard disk channel at the back end is used. This makes it difficult to ensure the performance of data exchange between the two controllers and brings about certain impacts on the access to the hard disk.
- Independent service channel and management channel are available, but no physical and logic redundancy is available. This reduces the reliability of the system.
- Independent service channel and management channel are available, and fully physical redundancy is available.

The communications between the storage devices become a new bottleneck as the storage device manufacturers make efforts in improving the internal processing capacities of the storage devices. Thus, fully redundant channel is designed for the Oceanspace S2000 series storage system during the implementation of dual controller design scheme, which improves the security of the channel and guarantees sound communication performance.

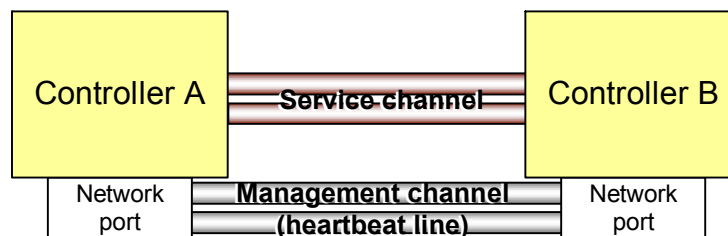


Figure 3-1 Fully redundant channel design

- **Management channel**

Through the management channel based on TCP/IP protocol, the two controllers can check the state for each other, for example, one controller checks whether the other controller is online or is faulty. Thus, the management channel is also called heartbeat line. In addition, the controllers can send and receive some control commands through the management channel, such as load sharing and I/O service sharing.

- **Service channel**

The FC-based service channel is responsible for service processing and I/O reading and writing operations. When a controller receives a new host service request, it can judge whether there is a need to switch the service request to the other controller for sharing according to the load balancing algorithm. If determining that there is a need to switch the service request to the other controller for sharing, the controller forwards the request to the other controller through the service channel. When a controller is faulty, the host request may be automatically transferred to the other controller through the service channel.

3.1.2 Cache Mirror Technology

The Active-Active working mode raises higher system design requirements, for example, it requires the system to support cache mirror technology. To ensure that the dual controllers can better guarantee the data integrity, the cache mirror technology needs to be used. That is, the cache data of both controllers is backed up through the mirror of each other, so that the host can read the cache data in other controller if one controller is faulty. This avoids the risk of data loss and ensures the data security and integrity.

The Oceanspace S2000 series storage system provides two mirror channels, which feature fully physical redundant and fully duplexing. Each mirror channel is based on FC port, with the unidirectional bandwidth of 4 Gbps and total bandwidth of 16 Gbps. This meets the requirements for high traffic of data and high IOPS communication performance.

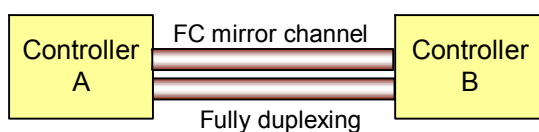


Figure 3-2 Fully redundant FC mirror channel

The cache where the cache mirror is available is classified into three parts, each of which performs different functions:

- Read cache: saves the data that the host reads from the cache.
- Write cache: saves the data that the host writes to the cache.
- Mirror cache: maps the write cache of the peer controller. In case one controller is faulty, the other controller directly obtains the contents of the write cache in the faulty controller from its own mirror cache. This ensures the data security and integrity and improves the availability of the system.

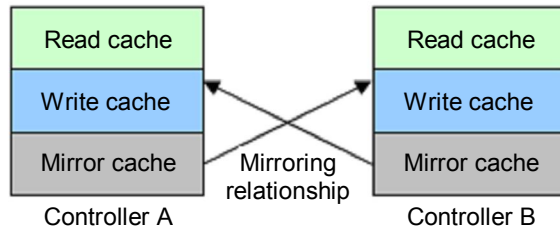


Figure 3-3 Schematic diagram of mirror cache

The cache mirror technology ensures the data integrity and completeness through mirror writing and mirror deletion.

I. Mirror Writing

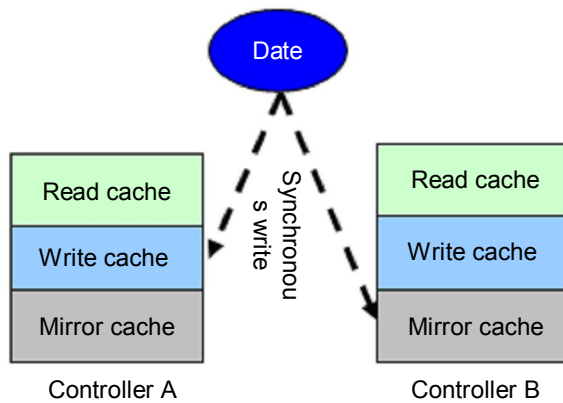


Figure 3-4 Schematic diagram of mirror writing

When the host writes data, the data is written to the mirror cache of the other controller while the data is saved in the write cache of the current controller. Thus, the data of the write caches in both controllers is mutually redundant, which improves the data security and the availability of the system.

II. Mirror Deletion

To synchronize the process of writing the data in the cache to the hard disk, the data is deleted from the cache after certain data in the write cache is written to the hard disk successfully. To ensure the data consistency between the mirror caches, there is a need to delete the same data in the mirror caches in time.

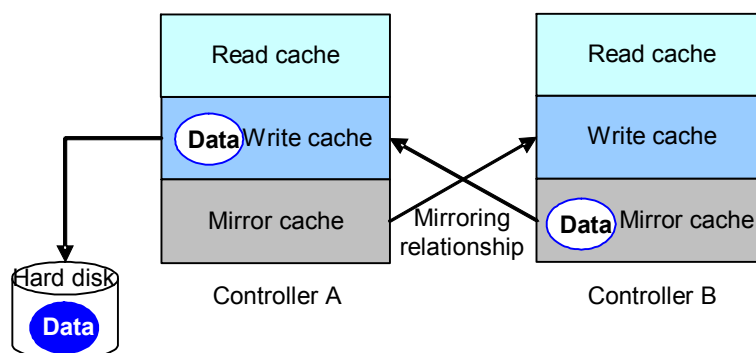


Figure 3-5 Schematic diagram of mirror deletion

3.2 Data Safe Box

In general, the storage system uses the cache to improve the read/write performance of the host. The host data is written to the cache first instead of the disk directly. Two write cache technologies are available to write data to the hard disk:

- **Transparent write cache:** The data is written to the hard disk immediately after the data is written to the cache.
- **Write back cache:** To respond to the host faster, the data is written to the cache by batch, and then all the batches of data is written to the hard disk. Unlike the hard disk, the cache may lose all the data in the case of power failures. If certain data in the cache is not written to the hard disk due to sudden power failures, the data is lost.

The following two methods are available to ensure the data security and integrity in the system in the case of sudden power failures.

One method is using lithium batteries to provide the cache with continuous power supply, which is also called cache battery power failure protection. This ensures that the cache power supply is not disconnected in the case of external power failures, thus guaranteeing the availability of the data in the cache. This method is limited by a lot of factors, such as battery capacity, data amount in the cache, cost, volume, and unit design. The battery capacity is limited. When the data amount in the cache is small, the battery capacity may support two or three days, which may be greatly shortened when the data amount in the cache is large. Thus, the battery capacity may not meet the data security requirements in extreme conditions.

The other method is using data safe box. That is, certain sectors are specified in several hard disks of the system to save the data in the cache that is not written to the hard disk due to sudden power failures and some system configuration information. In the case of external power failures, the built-in batteries or external UPSs are used to provide power supply, so that the data in the cache can be written to the data safe box.

When the external power supply is recovered, the controller reads the data from the data safe box back to the cache to continue the data processing.

The lithium batteries prolong the data saving capability of the cache by providing continuous power supply, but the data safe box saves the data in the cache permanently. Thus, the data safe box is more reliable and secure.

Manufacturers implement the data safe box in different ways. Some manufacturers use a special type of hard disk as the data safe box; for example, they use the FC hard disk only. Instead, some manufacturers do not limit the type of hard disks; for example, they may use FC, SATA, and SAS hard disks as the data safe box. In addition, the manufacturers adopt different technologies to ensure the security of the data safe box. For example, some manufacturers adopt RAID 5 to protect the data in the data safe box, while some manufacturers adopt RAID 10 to protect the data at a higher security level.

The Oceanspace S2000 series storage system is compatible with the SATA and SAS hard disks, that is, it has not special requirement for the hard disk type which is used as the data safe box. The Oceanspace S2000 series storage system adopts RAID 10 for data protection at a high security level.

To prevent the potential impact on the system in the case of external power failures, the Oceanspace S2000 series storage system designs integrated redundant UPSs, which bring about such advantages as small battery volume in the cache and cost-effectiveness, and benefits of redundant design. In case the external double power supplies are faulty, both built-in UPSs can independently provide power supplies to the two controllers and multiple hard disks for the data safe box. Thus, the data in the cache can be written to the hard disk in the case of power failures, ensuring the data integrity and reliability. The built-in UPSs are designed in self maintenance mode, making it unnecessary to perform special or additional maintenance during the normal use.

The data safe box of the Oceanspace S2000 series storage system stores three types of data: cache data, configuration data, and alarm log. The cache data refers to the service data that is not written to the hard disk in time. In the case of failures, all the critical data is written to the data safe box after the cache receives a message indicating that there is a need to write data to the safe box. The safe box provides reliable hardware and sufficient capacity to ensure the write operation. After the system is recovered, the safe box receives a command for reading the data in the safe box. Then, the safe box is open, and all the data is submitted to the cache. This recovers all the data in the cache.

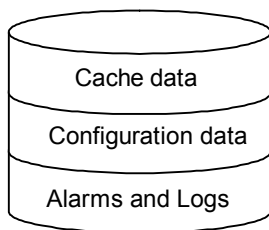


Figure 3-6 Data in the safe box

For the Oceanspace S2000 series storage product with dual controllers, the safe box is also designed to be fully redundant. The storage space of the safe box consists of four hard disks of the controller subrack. However, the safe box occupies only a small space at the end of each hard disk. The rest of space in each hard disk can also be used for storing data. For example, RAID or LUN is established to store service data required by the host.

As shown in Figure 3-7, the four hard disks of the safe box are divided in two groups, each of which consists of two hard disks. The two groups are named safe box 0 and safe box 1. There are link connections between the two controllers and safe box 0 and safe box 1. The data in the two hard disks in the same group is mutually backed up, that is, the data contents are consistent in the two hard disks. This avoids single point failures in the hard disks of the safe box and controllers, thus realizing high reliability and high availability of the data.

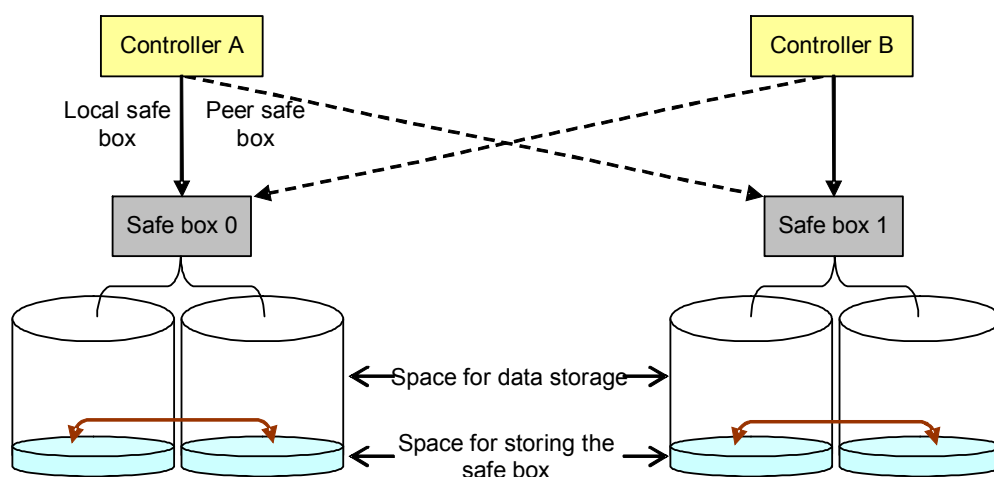


Figure 3-7 Mapping between the controllers and the safe box

In case a hard disk in the safe box is faulty, data loss may not occur in the Oceanspace S2000 series storage product. So long as you insert a new available hard disk in time, the safe box may recover the data completely to the new hard disk by using its data recovery mechanism automatically. The faulty hard disk can be replaced online, that is, you do not need to interrupt the system services, thus ensuring the service continuity.

3.3 Pre-Copy Technology with Extra Safe Hard Disk

Data security is the basic requirement for a storage system. The S2600 storage system uses the RAID technology to ensure system reliability. However, the security of RAID algorithms relies on the reliability of hard disks. When a disk runs for a long time, the probability of failure increases. Especially for a storage system using the disks of the same batch, when a disk fails, it indicates the failure probability of the entire system increases.

In addition, any RAID algorithm allows only a specific number of disks to fail at the same time. If you cannot find the potential faults in running disks and handle the faults in time, great risks are posed to data security. If a disk fails, it takes a certain time to reconstruct the data in the failed disk, which degrades the performance of the entire system. In this case, the redundant copy technology is introduced to prevent or reduce the impacts on the storage system caused by disk reliability.

The redundant copy technology allows you to obtain the information about hard disk status through the self-monitoring analysis and reporting technology (S.M.A.R.T). The redundant copy algorithm checks the running status of the hard disks to calculate the probability of potential failures of hard disks and copy the data from the hard disk with a potential failure to the hot-spare disk in advance. The entire copy process is performed when the system is idle to prevent impacts on host services. This prediction act shortens the reconstruction time after a disk failure and reduces the probability of further failures of the disk during reconstruction. In addition, it greatly improves storage security and ensures service continuity.

The S2000 series storage system uses the redundant copy technology with extra safe hard disk prediction.

The accuracy of disk status prediction is the key to the redundant copy technology. By recording traceable property items during disk running, it determines the health status of the disks. Common disk status prediction faces the following problems:

- Because the disk is a precise mechanic component, there is a small probability that some faults cannot be found in time through the S.M.A.R.T.
- According to statistics from professional organizations, up to 36% disks receives no alerts from the S.M.A.R.T before the disks fails. The S.M.A.R.T mainly detects mechanical problems, but disk damage is caused by the problems of electrical parts.
- A new idea in reliability shows that we have some misunderstanding about the reliability of disks.
- Specific environmental factors shall be considered in failure judgment in future.

The following innovation is introduced in the redundant copy technology of the S2000 storage system.

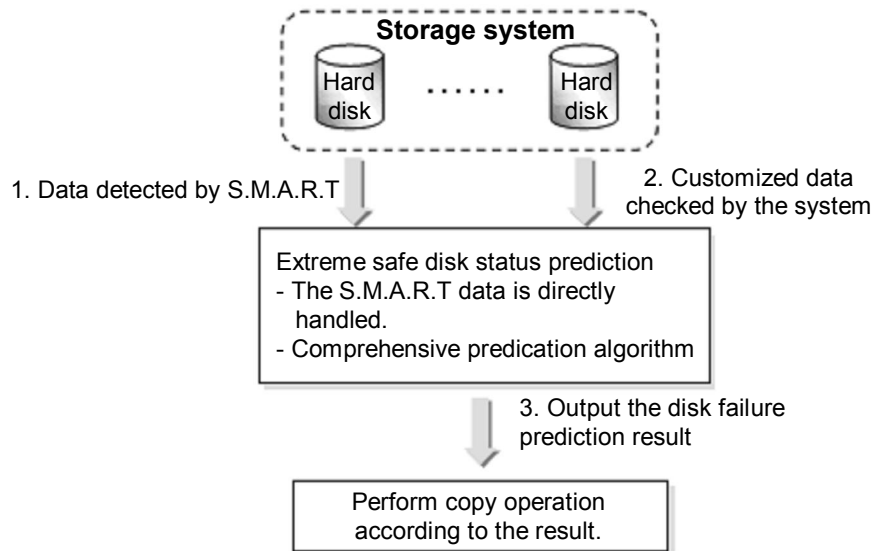


Figure 3-8 Schematic diagram of redundant copy technology with extra safe hard disk

- (1) The first part of data: disk items checked by the S.M.A.R.T and disk status predicted.
- (2) The second part of data: customized data collected and checked by the storage devices periodically, including special attributes of hard disks, running parameters of RAID groups, equipment management data, and environment factors.

The special attributes of hard disks include year of each hard disk, batch No., reliability index of the batch, and new or old features. The running parameters of RAID groups refer to the dynamically measured use details of hard disks, for example, continuous read and write details, random read and write details, bandwidth measurement, and measurement of some running features.

- (3) The preceding two parts of data are called extra safe data. The data is calculated by a specific algorithm of predicting disk status and then the result of the predicted disk failures is outputted.

The specific algorithms of the Oceanspace analyze and refer to the data based on different conditions and different weight and predication strategies. For example, magnetic calibration needs to be considered in the case of frequent random read and write; read and write errors need to be considered in the case of frequent sequential read and write; and running performance variation rate needs to be focused on in the case of a disk failure.

- (4) Enable redundant copy according to the predicted result.

The Oceanspace series storage system provides the redundant copy technology and the technology of completely and effectively predicting disk status, which ensure data security and service continuity.

3.4 Global Hot-Spare Disk

The hot-spare disk is a spare hard disk specified in the configuration of hard disk array system. The standby hard disk does not respond to the service requirements of the host during the normal operation. When a hard disk in the hard disk array is faulty, the hard disk array replaces the faulty hard disk with the spare hard disk, and reconstructs the data in the faulty hard disk in the spare hard disk.

The hot-spare disk is important in a large data processing center or control center where the operation cannot be stopped. It can avoid risk of data loss in the case of hard disk failure during the nighttime or unattended time.

The hot-spare disk includes the following two types:

- Global hot-spare disk: It is designed for the entire hard disk array, and functions for all the RAID groups in the hard disk array.
- Partial hot-spare disk: It functions for one RAID group only.

The Oceanspace S2000 series storage system uses intelligent global hot-spare disk technology, which does not specify the special location of the hot-spare disk and facilitates the configuration of hot-spare disk.

Besides hot-spare disk, the Oceanspace S2000 series storage system supports COPYBACK for the replaced hard disk. COPYBACK means that the system copies the data from the hot-spare disk to the new hard disk automatically if the faulty hard disk is replaced by a new hard disk after the data reconstruction is complete. Figure 3-9 shows the process of reconstruction and COPYBACK in the hot-spare disk.

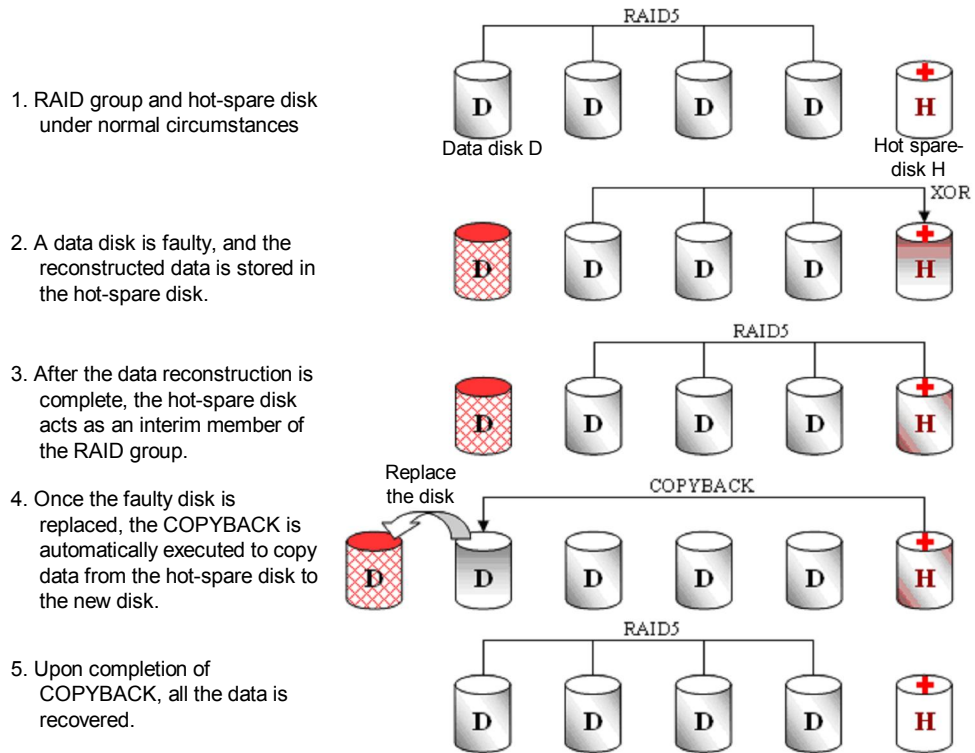


Figure 3-9 Process of reconstruction in hot-spare disk

If a hot-spare disk is already available in the hard disk array of the Oceanspace S2000 series storage system, the system may search the hard disk array for the hot-spare disk in the case of failure of a hard disk in the array. After finding the hot-spare disk, the system reconstructs the data in the hot-spare disk according to the information in other hard disks, such as check recovery. After the reconstruction process is complete, the system may perform a COPYBACK operation if a new hard disk is inserted to replace the faulty hard disk, that is, the system may read the data from the hot-spare disk to the new hard disk. It should be noted that all the operations in the whole process are performed without interrupting the system operations.

4 Advocating Energy Saving Design

4.1 Increasing Space Usage through Careful Design

With enterprise operations transforming towards refinement, you will surely consider the unit space input and earnings, whether for self-built or rented equipment rooms. The annually rising unit space cost will turn your focus on how to increase space usage.

With fast advance of information technology and dramatic increase of data amount in enterprises, the data centralization must be performed to meet the requirements for convenient management and maintenance, low operation costs, necessary data mining, and secondary development of enterprise information. However, such huge data amount requires large storage capacity, which brings about incredible number of hard disks and necessary equipment installation space.

At present, when a 2U typical storage disk subrack is designed, it may hold a maximum of twelve 3.5-inch hard disks. When a 3U storage disk subrack is designed, it may hold a maximum of fourteen, fifteen, or sixteen 3.5-inch hard disks according to the design capacities of different manufacturers. Sometimes one more hard disk may require one more hard disk subrack, thus wasting the space of the cabinet. Therefore, you certainly hope that a hard disk subrack can hold more hard disks to meet higher storage capacity requirements in the limited physical space. Nevertheless, subsequent problems such as power supply, heat dissipation and noise pose new challenges to the hardware design capabilities of the manufacturers.

By virtue of years of experiences in the telecom field, Huawei Symantec has developed excellent hardware design capability. The design technologies and experience contribute to successful design of 4U storage disk subrack which tops the industry by the support of up to twenty-four 3.5-inch hard disks in the same size of space. For example, a 750 G SATA hard disk subrack can provide a storage capacity of up to 18 TB. Compared with the 3U hard disk subrack, it saves 40% of cabinet space.

4.2 Integrated UPS

Two data power-failure protection modes are available. One is using built-in lithium batteries. Though the built-in lithium batteries can be integrated with the controllers, the battery capacity may not meet the data amount requirement in the cache. The other is using external UPSs. Configuring UPSs of a third party requires additional costs and equipment space, and more importantly, adds failure points, which brings about more risk exposures to the system.

The Oceanspace S2000 series storage system adopts an innovative technology—integrating UPS with the controllers, which can save space and protect data like external UPSs without additional space.

4.3 Integrated Controller and Hard Disk Subrack

To save space as much as possible, Huawei Symantec Oceanspace S2000 series storage system integrates a 4U hard disk subrack with two Active-Active controllers using iSCSI host ports. As a result, a 750 G SATA hard disk can provide a super large

capacity of 18 TB in 4U space, which proves the powerful hardware design capabilities of Huawei Symantec.

4.4 Intelligent Hard Disk Power-on

Because there are a lot of hard disks in the storage system, the power supply may be affected and even such serious accidents as current overload or tripping may occur if the hard disks are powered on at the same time. To solve such problem, the Oceanspace S2000 series storage system uses an intelligent hard disk power-on technology, that is, hard disk soft power-on technology, to lower the power supply requirements.

Upon power-on of hard disk subrack, all the hard disks enter soft start mode, that is, the system controls the power-on sequence of each hard disk by using a special algorithm for cyclic scanning to ease the pressure of power supply. During the process of normal system functioning, the power-on recovery of a single hard disk is performed in fast startup mode to ensure timely power-on of hot swap hard disk.

5 Concerning Your Power Supply Environment

Currently, two power supply environments are available in China:

- AC power supply (single phase 220 V or three phase 380 V, 50 Hz) environment
The AC power supply environment is common and easy to obtain. The AC power supply is provided directly by the power supply bureau, and can be used without conversion in an actual power using environment. The AC power supply fluctuates easily because multiple users may be connected to the same power supply line.
- DC power supply (- 48 V) environment
The DC power supply environment is complex and raises higher requirement. It is generally provided by the power supply bureau through a special line. In addition, battery strings may be used as standby power supply to provide a long term and stable power supply environment and ensure the normal operation of the equipment, thus ensuring the uninterrupted services of the service system.

Many enterprises, such as telecom operators, professional IDC equipment rooms, and data management centers, propose extremely high data security requirements. Thus, they may establish their own equipment rooms where the DC power supply environment is used. Because most storage device manufacturers can only provide AC power supply environment, their customers are hard to make choices when purchasing storage devices. Thus, they have to specify an interim AC power supply

area for the storage devices or other devices that cannot support DC power supply in their equipment rooms in order meet the installation requirement of these devices.

The Oceanspace S2000 series storage system fully considers the power supply environment for the users. With many years of experiences in designing DC power supply for the telecom equipment, Huawei Symantec provides AC and DC power modules for the users, making the power supply environment easier to plan in the equipment room.

6 Acronyms and Abbreviations

Acronyms and Abbreviations	Full Spelling
ATA	advanced technology attachment
FC	fibre channel
IDC	Internet Data Center
IP	Internet Protocol
iSCSI	Internet small computer systems interface
LUN	logical unit number
NAS	network attached storage
PCI	peripheral component interconnect
RAID	redundant array of independent disks
S.M.A.R.T	self monitoring analysis and reporting technology
SAS	serial attached SCSI
SATA	serial advanced technology attachment
SCSI	small computer system interface
TCP	Transmission Control Protocol
UPS	uninterruptible power source