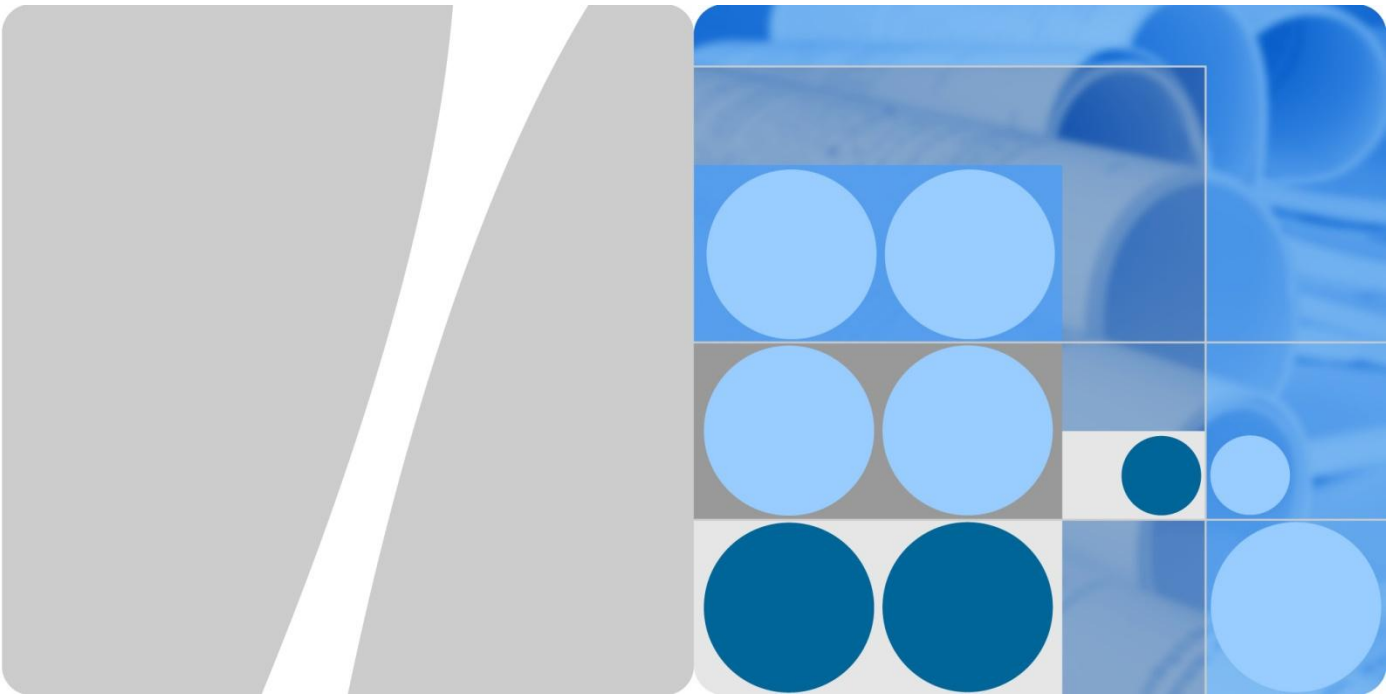


Huawei OceanStor S2600T V2 Technical White Paper

Issue 1.0

Date 2015-12

HUAWEI TECHNOLOGIES CO., LTD.

HUAWEI

Trademarks and Permissions


and other Huawei trademarks are trademarks of Huawei Technologies Co., Ltd.

All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Huawei Technologies Co., Ltd.

Address: Huawei Industrial Base
Bantian, Longgang
Shenzhen 518129
People's Republic of China

Website: http://ehnterprise.huawei.com

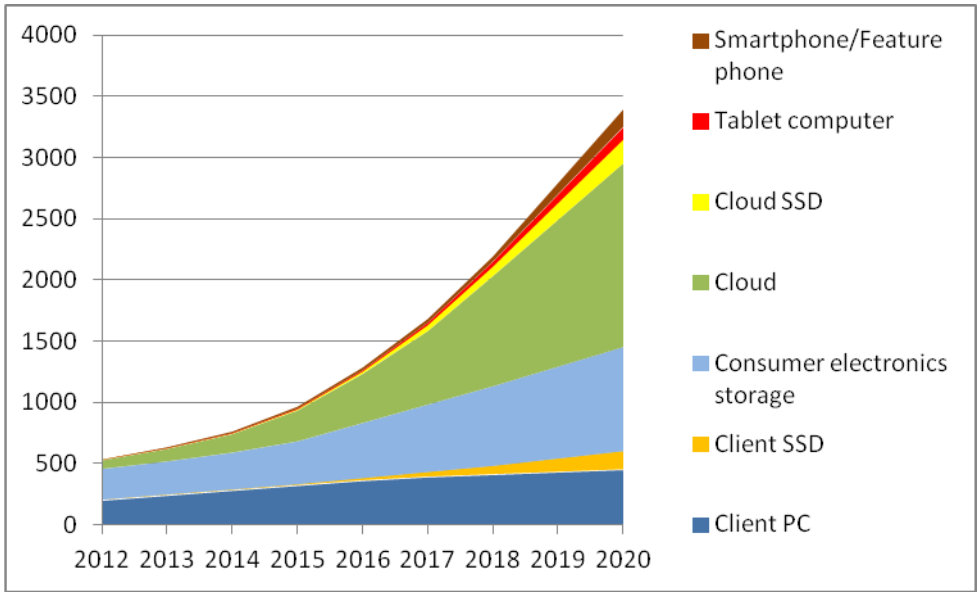
Contents

1 Overview

Evolving from mainframe servers to midrange computers, PCs, and desktop Internet, the information technology (IT) is penetrating into all walks of life. Nowadays, we are embracing the mobile Internet era. The change of application environments hastens data explosion. According to Gartner's statistics, about 2.6 EB of data was generated around the world in the era of midrange computers and 15.8 EB of data when PCs were popular. In the era of desktop Internet, the amount of data was almost quadrupled, reaching 54.5 EB. Up to 1800 EB of data may be generated in the era of mobile Internet. The skyrocketing amount of data not only requires super-large storage capacities but also imposes demanding requirements on other features of storage products.

Since data sources are increasingly diversified, clouds will gradually become the largest data sources, replacing PCs and consumer electronics (CE). The following figure shows predicted rankings of data sources.

Figure 1-1 Predicted rankings of data sources



Since data sources are changing constantly, data types change accordingly. Although the amount of critical service data, such as databases, increases continuously, it accounts for a decreasing percentage of the total data volume; whereas enterprise office data, such as emails

and large media files, once accounted for the highest percentage of the total data volume. In recent years, since the amount of personal data increases sharply, media and entertainment data replaces enterprise office data as the largest data sources.

In 1993, critical service data and enterprise office data accounted for 50% of the total data volume respectively, and the amount of personal data could be ignored. In 2002, 70% of data was enterprise office data, and 20% was critical service data. Since 2010, personal data accounts for 50% of the total data volume, whereas enterprise office data accounts for 40%, and critical service data accounts for only 10%.

Different types of data from diversified sources have different requirements on the performance, reliability, and costs of storage media. Critical service data requires high-performance and robust-reliability storage devices, whereas personal entertainment data requires inexpensive storage devices. The reality is that critical service data and personal entertainment data usually need to be stored in a single set of storage device. Such contradicting requirements impose new challenges. To keep with IT development, next-generation mid-range storage products must have:

- Integrated, simple, intelligent, and cost-effective system architecture
- High flexibility, meeting diverse storage needs
- Agile data planning and management
- Rich and practical functions

Considering customers' requirements and adopting the concept of "Integrated & Reliable, Intelligent & Efficient", HUAWEI OceanStor S2600T V2 storage systems (the S2600T for short) employ a brand-new software platform and deliver powerful flexibility and intelligent resource management capabilities, maximizing customers' return on investment (ROI).

2 Definition of "Integrated & Reliable, Intelligent & Efficient"

Inheriting the concept of unified storage, the S2600T integrates file- and block-level data storage and supports various storage protocols. Meanwhile, the S2600T uses industry-leading storage resource virtualization technologies, highly reliable software and hardware architecture, intelligent and efficient storage resource scheduling algorithms, and rich quality of service (QoS) assurance mechanisms, providing a high-performance and all-in-one solution that maximizes customers' ROI. The S2600T meets the requirements of various service applications including large online transaction processing (OLTP)/online analytical processing (OLAP) databases, high performance computing (HPC), digital media, Internet applications, central storage, backup, disaster recovery, and data migration.

Integrated: unified protocols and management

SAN and NAS storage protocols are integrated. A single storage system can store structured as well as unstructured data. Multiple storage network protocols such as iSCSI, Fibre Channel, Network File System (NFS), common Internet file system (CIFS), Hypertext Transfer Protocol (HTTP), and File Transfer Protocol (FTP) are supported.

The S2600T adopts RAID 2.0+ technology that virtualizes physical disks into small storage units for fine-grained space management.

Integrated: heterogeneous resource management

The S2600T can connect to storage systems from other vendors and provide the storage systems with new features, such as snapshot, remote replication, QoS, and cache partition optimization.

Reliable: decreased faults and quick self-healing

- The S2600T adopts the RAID 2.0+ technology to accelerate data reconstruction. The maximum reconstruction speed is 2 TB per hour.
- The S2600T employs the bad sector repair technology to proactively detect and repair bad sectors, reducing the disk failure rate by 50% and prolonging the service life of disks.
- The S2600T leverages multiple disk protection patented technologies to ensure that disks meet DC G1 to DC GX criteria in every aspect such as vibration and corrosion.

Reliable: second-level disaster recovery

The S2600T uses the innovative multi-timestamp cache technology. During replication and synchronization, data with specific timestamps is directly replicated from the production site

to the disaster recovery site, reducing latency. The minimum replication period is reduced to 3 seconds.

Intelligent: SmartTier

SmartTier automatically analyzes data access frequencies per unit time and migrates data to disks of different performance levels based on the analysis result. (High-performance disks store most frequently accessed data, performance disks store less frequently accessed data, and large-capacity disks store seldom accessed data.)

In this way, the optimal overall performance is achieved, and the IOPS cost is reduced.

Intelligent: SmartThin

SmartThin allocates storage space on demand rather than pre-allocating all storage space at the initial stage. It is more cost-effective because customers can start business with a few disks and add disks based on site requirements. In this way, the initial purchase cost and TCO are reduced.

Intelligent: SmartMigration

SmartMigration migrates host services from a source LUN to a target LUN without interrupting these services and then enables the target LUN to take over services from the source LUN without being noticed by the hosts. After the service migration is complete, all service-related data has been replicated from the source LUN to the target LUN.

Efficient: SmartQoS

SmartQoS categorizes service data based on data characteristics (each category represents a type of application) and sets a priority and performance objective for each category. In this way, resources are allocated to services properly, fully utilizing system resources.

Efficient: SmartPartition

The core idea of SmartPartition is to ensure the performance of mission-critical applications by partitioning core system resources. Users can configure cache partitions of different sizes. The S2600T ensures the number of cache partitions occupied by service applications. Based on the actual service condition, the S2600T dynamically adjusts the number of concurrent access requests from hosts to different cache partitions, ensuring the service application performance of each partition.

3 Integrated: Unified Protocols and Management

Support for Multiple Storage Protocols

As controllers and NAS engines of the S2600T are developed on the same hardware platform, the S2600T can provide IP SAN, FC SAN, and NAS networking modes simultaneously, and supports the iSCSI, FCP, NFS, CIFS, HTTP, and FTP protocols, as shown in the following figure.

Figure 3-1 Integration of various protocols and networking modes

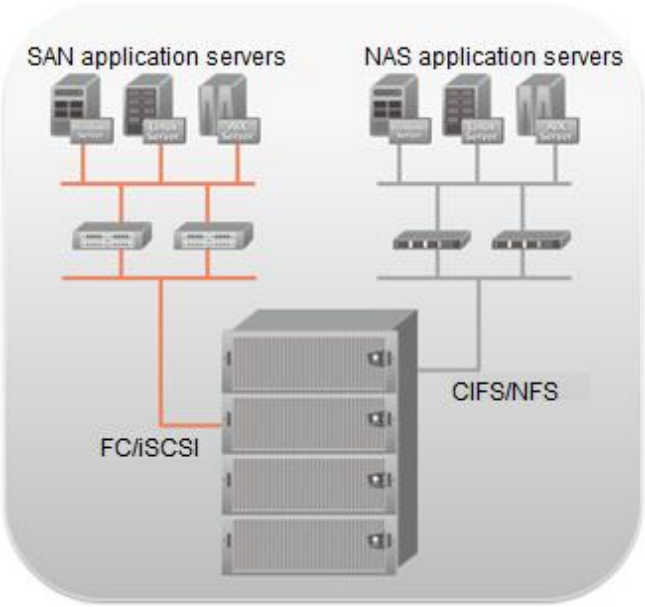


Table 3-1 Protocol specifications of the S2600T

Protocol Type	Specifications	Description
Fibre Channel	Supported protocols	FCP, FC-SW, FC-PH, and FC-PI
	Interruption aggregation	Supported, disabled by default

Protocol Type	Specifications	Description
	Port autonegotiation (rate/topology)	Rates: 8 Gbit/s, 4 Gbit/s, and 2 Gbit/s Topologies: Fabric, Loop, and P2P
iSCSI	Supported protocols	IPv4 and IPv6
	Port autonegotiation (rate/topology)	Rates: 1 Gbit/s and 10 Gbit/s
	iSCSI CHAP authentication	Unidirectional CHAP authentication, initiated by hosts
	Port aggregation type	Dynamic link aggregation (IEEE802.3ad)
	Jumbo frames	Supported, MTU ranging from 1500 bits to 9216 bits
CIFS	Supported protocol	SMB1.0
	Share types	Homedir and normal
	Number of normal shares	256
	Number of file systems shared in homedir mode	16
	Number of links shared in homedir mode	3000
	Number of active links shared in homedir mode	800
NFS	Supported versions	V2 and V3
	Number of shared links	800
FTP	Number of local users	1000
	Number of shared links	800
HTTP	Supported version	V1.0

Unified Management of Storage Resources

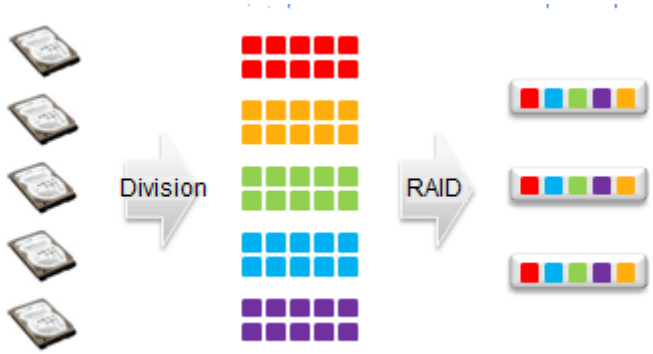
RAID 2.0+ is a brand-new RAID technology developed by Huawei based on the dedicated storage operating system to overcome the disadvantages of traditional RAID and keep in line with the storage architecture virtualization trend. RAID 2.0+ implements two-layer virtualized management instead of traditional fixed management. Based on the underlying disk management that employs block virtualization (Virtual for Disk), RAID 2.0+ implements efficient resource management that features upper-layer virtualization (Virtual for Pool).

RAID 2.0+ employs two-layer virtualized management, namely, underlying disk management plus upper-layer resource management. In a S2600T storage system, the space of each disk is divided into data blocks with a small granularity and RAID groups are created based on data

blocks so that data is evenly distributed onto all disks in a storage pool. Besides, using data blocks as the smallest units greatly improves the efficiency of resource management.

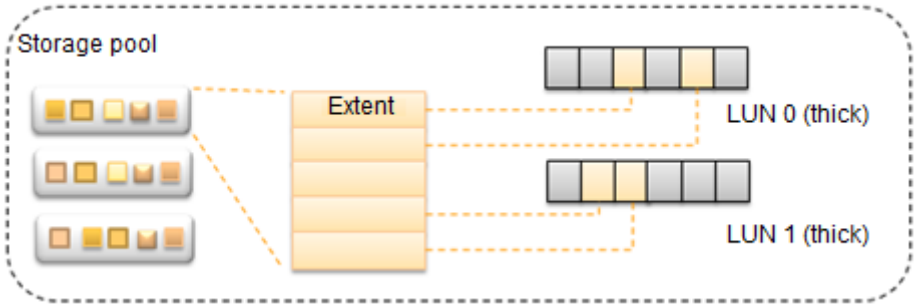
1. The S2600T supports SSDs, SAS disks, and NL-SAS disks. Each type of disks can form a tier. On each tier, every disk is divided into chunks (CKs) of 64 MB each. The S2600T employs the randomized algorithm to select disks and uses the RAID algorithm to combine CKs of different disks into Chunk Groups (CKGs).

Figure 3-2 CKs and CKGs



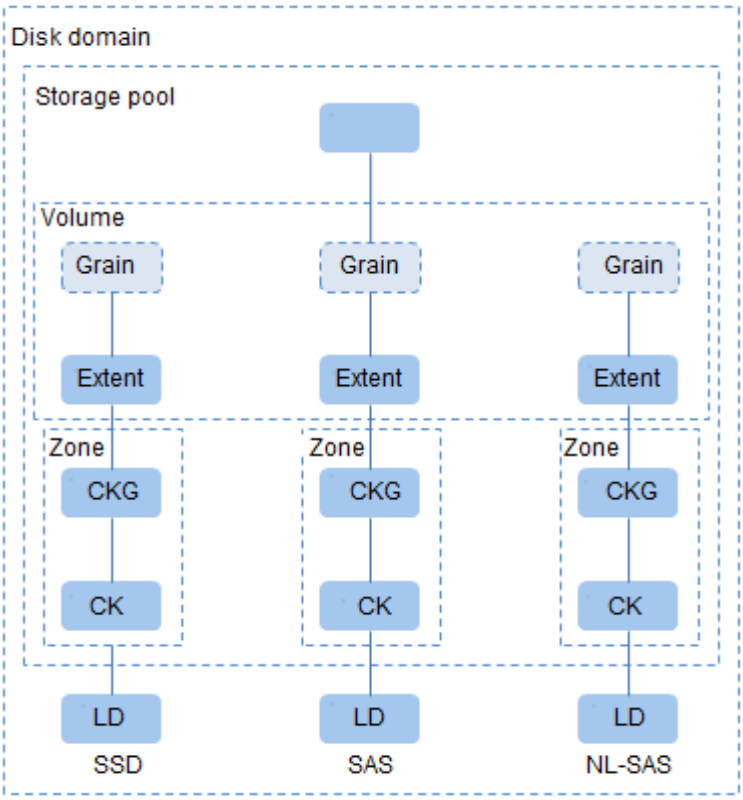
2. CKGs are divided into extents (logical storage space of fixed sizes). The size of an extent can be 1 MB, 2 MB, 4 MB, 8 MB, 16 MB, 32 MB, or 64 MB. The default extent size is 4 MB. Extents are basic units that constitute a LUN.

Figure 3-3 Data structures of LUNs



The following figure shows the RAID 2.0+ implementation architecture:

Figure 3-4 Data structure of RAID 2.0+



- Disk domain: A disk domain in the S2600T consists of disks from one or multiple tiers. Each tier supports disks of a specific type: SSDs compose the high-performance tier, SAS disks compose the performance tier, and NL-SAS disks compose the capacity tier.
- The disks on each storage tier are divided into CKs with a fixed size of 64 MB.
- CKs on each storage tier compose CKGs based on a user-defined RAID policy. Users are allowed to define a specific RAID policy for each storage tier of a storage pool.
- CKGs are further divided into extents. An extent is the smallest granularity for data migration and is the basic unit of a thick LUN. When creating a storage pool, users can set the extent size on the **Advanced** page. The default extent size is 4 MB.

  Multiple extents compose a volume that is externally presented as a LUN (a thick LUN) accessible to hosts. A LUN implements space application, space release, and data migration based on extents. For example, when creating a LUN, a user can specify a storage tier from which the capacity of the LUN comes. In this case, the LUN consists of the extents on the specified storage tier. After services start running, the storage system migrates data among the storage tiers based on data activity levels and data migration polices. (This function requires a SmartTier license.) In this scenario, data on the LUN is distributed onto the storage tiers of the storage pool based on extents.
- When a user creates a thin LUN, the S2600T divides extents into grains and maps grains to the thin LUN. In this way, fine-grained management of storage capacity is implemented.

Pool virtualization to simplify storage planning and management

Nowadays, mainstream storage systems typically contain hundreds of or even thousands of disks of different types. If such storage systems employ traditional RAID, administrators need

to manage a lot of RAID groups and must carefully plan performance and capacity for each application and RAID group. In the era of constant changes, it is almost impossible to accurately predict the service development trends in the IT system lifecycle and the corresponding data growth. As a result, administrators often face management issues such as uneven allocation of storage resources. These issues greatly increase management complexity.

The S2600T employs advanced virtualization technologies to manage storage resources in the form of storage pools. Administrators only need to maintain a few storage pools. All RAID configurations are automatically completed during the creation of storage pools. In addition, the S2600T automatically manages and schedules system resources in a smart way based on user-defined policies, significantly simplifying storage planning and management.

One LUN across more disks to improve performance of a single LUN

Server computing capabilities have been improved greatly and the number of host applications (such as databases and virtual machines) has increased sharply, causing the needs for higher storage performance, capacity, and flexibility. Restricted by the number of disks, a traditional RAID group provides only a small capacity, moderate performance, and poor scalability. These disadvantages prevent traditional RAID groups from meeting service requirements. When a host accesses a LUN intensively, only a limited number of disks are actually accessed, easily causing disk access bottlenecks and making the disks hotspot.

RAID 2.0+ supports a storage pool that consists of dozens of or even hundreds of disks. LUNs are created based on a storage pool, thereby no longer subject to the limited number of disks supported by a RAID group. Wide striping technology distributes data of a single LUN onto many disks, preventing disks from becoming hotspots and enabling the performance and capacity of a single LUN to improve significantly.

If the capacity of an existing storage system does not meet the needs, a user can dynamically expand the capacity of a storage pool and that of a LUN by simply adding disks to the disk domain. This approach improves disk capacity utilization. If the capacity of an existing storage system does not meet the needs, a user can dynamically expand the capacity of a storage pool and that of a LUN by simply adding disks to the disk domain. This approach improves disk capacity utilization.

Dynamic space distribution to flexibly adapt to service changes

RAID 2.0+ is implemented based on industry-leading block virtualization. Data and service loads in a volume are automatically and evenly distributed onto all physical disks in a storage pool. By leveraging the Smart series efficiency improvement software, the S2600T automatically schedules resources in a smart way based on factors such as the amount of hot and cold data and the performance and capacity required by a service. In this way, the S2600T adapts to rapid changes in enterprise services.

4 Integrated: Scalability of Resources

The S2600T adopts the industry-leading TurboModule technology to deliver excellent scalability. The TurboModule technology supports hot-swappable modules, flexible combination of host and expansion I/O modules, and flexible I/O modules and interfaces.

- Hot-swappable modules: The S2600T supports redundancy throughout the system. Redundant components, including controllers, power modules, fans, integrated BBUs, disks, and I/O modules, are all hot-swappable.

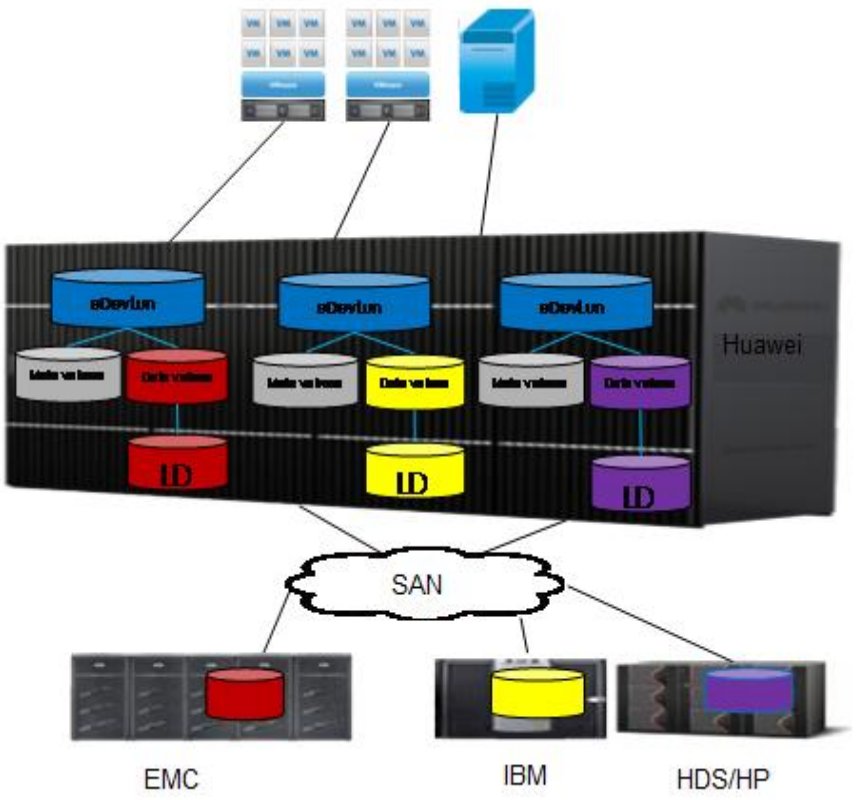
  Hot-swappable I/O modules are the most extraordinary design adopted by the S2600T in scalability. With this design, the S2600T allows users to add I/O modules online as service data explodes, cutting the cost and labor of adding switches. Furthermore, once an I/O module is faulty, it can be replaced online. This design greatly protects system reliability and service continuity.
- Flexible combination of front- and back-end I/O modules: The S2600T supports 2 I/O modules.
- Flexible I/O modules: The S2600T supports six ports types including 4 Gbit/s Fibre Channel, 8 Gbit/s Fibre Channel, 16 Gbit/s Fibre Channel, GE, 10GE, and 4 x 6 Gbit/s SAS.

5 Integrated: Heterogeneous Resource Management

The S2600T aims at providing rich virtualization functions for heterogeneous storage systems of customers:

- The heterogeneous takeover function reduces complexity in managing heterogeneous storage systems and improves LUN performance.
- The heterogeneous online migration function allows data to be smoothly migrated among LUNs of heterogeneous storage systems without interrupting services.
- The heterogeneous remote replication function implements disaster recovery for LUNs of heterogeneous storage systems.
- The heterogeneous snapshot function implements rapid backup for LUNs of heterogeneous storage systems. The heterogeneous virtualization feature provided by the S2600T is called SmartVirtualization.

SmartVirtualization uses LUNs mapped from heterogeneous storage systems to the local storage system as logical disks (LDs) that can provide storage space for the local storage system and create eDevLUNs that can be mapped to the host on LDs. LDs provide data storage space for data volumes, and the local storage system provides storage space for meta volumes of eDevLUNs. SmartVirtualization ensures data integrity of external LUNs.



eDevLUNs and local LUNs have the same properties. For this reason, SmartMigration, HyperReplication/S, HyperReplication/A, and HyperSnap are used to provide online migration, synchronous remote replication, asynchronous remote replication, and snapshot functions respectively for LUNs of heterogeneous storage systems. Meanwhile, SmartQoS, SmartPartition, and cache write back are used to improve the LUN performance of heterogeneous storage systems.

SmartVirtualization applies to:

- Heterogeneous array takeover

  As users' data centers develop, storage systems in the data centers may come from different vendors. How to efficiently manage and apply storage systems from different vendors is a challenge that storage administrators must tackle. Storage administrators can leverage the takeover function of SmartVirtualization to simplify heterogeneous array management. They need only to manage Huawei storage systems, and their workloads are remarkably reduced. In such a scenario, SmartVirtualization simplifies system management.

- Heterogeneous data migration

  A large number of heterogeneous storage systems whose warranty periods are about to be exceeded or whose performance cannot meet service requirements may exist in a customer's data center. After purchasing Huawei OceanStor storage systems, the customer wants to migrate services from the existing storage systems to the new storage systems. The customer can leverage the online migration function of SmartMigration to migrate data on LUNs of heterogeneous storage systems to the new storage systems. The migration process has no adverse impact on ongoing host services, but the LUNs must

be taken over before the migration. In such a scenario, SmartVirtualization ensures ongoing host services when data on LUNs of heterogeneous storage systems is migrated.

- Heterogeneous disaster recovery

  If service data is scattered at different sites and there are demanding requirements for service continuity, the service sites need to serve as backup sites mutually, and service switchovers can be performed between sites. When a disaster occurs, a functional service site takes over services from the failed service site and recovers data. However, as storage systems at the data site come from different vendors, data on the storage systems cannot be backed up mutually. The synchronous and asynchronous replication functions of SmartVirtualization enable data on LUNs of heterogeneous storage systems to be backed up mutually, achieving data disaster recovery among sites.

- Heterogeneous data protection

  Data on LUNs of heterogeneous storage systems may be attacked by viruses or damaged. SmartVirtualization leverages the heterogeneous snapshot function to create snapshots for LUNs of heterogeneous storage systems instantly, and rapidly restores data at a specific point in time using the snapshots if data is damaged.

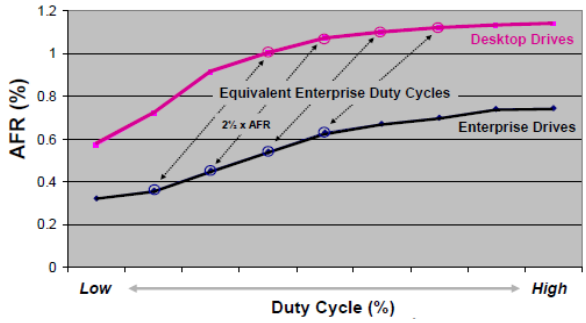
6 Reliable: Decreased Faults and Quick Self-Healing

Decreased Faults

Automatic load balancing to decrease the overall failure rate

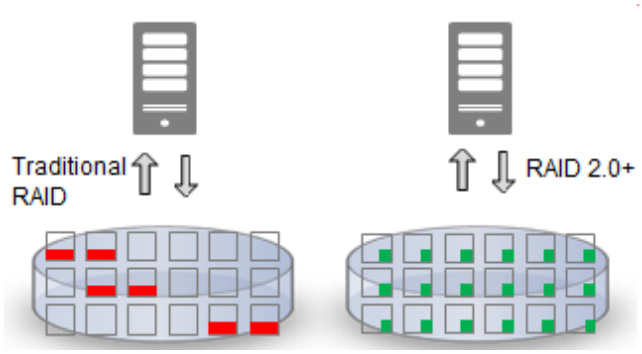
A traditional RAID-based storage system typically contains multiple RAID groups, each of which consists of up to 10-odd disks. RAID groups work under different loads, leading to an unbalanced load and existence of hotspot disks. According to the statistics collected by Storage Networking Industry Association (SNIA), hotspot disks are more vulnerable to failures. As shown in the following figure, **Duty Cycle** indicates the percentage of disk working time to total disk power-on time, and **AFR** indicates the annual failure rate. It can be inferred that when the duty cycle is high, the AFR is almost 1.5 to 2 times higher than the AFR in the low duty cycle scenario.

Figure 6-1 Mapping between Duty Cycle and AFR



RAID 2.0+ implements block virtualization to enable data to be automatically and evenly distributed onto all disks in a storage pool, preventing unbalanced loads. This approach decreases the overall failure rate of a storage system.

Figure 6-2 Data distribution in RAID 2.0+



Patented design to enhance adaptability

Adopting the industry-leading shockproof structure design to protect disks, fan modules, chassis, and guide rails, the S2600T was certified by China MII Communication Equipment Earthquake Resistance Performance Quality Inspection and Testing Center to have the 9-intensity earthquake resistance capability. The S2600T is the only professional storage system that meets the highest anti-seismic requirements stated in the *Specification for Seismic Test of Telecommunication Equipment* (YD5083).

1. Disk unit vibration isolation: Viscoelastic materials are installed inside disk trays to absorb the vibration energy generated by disk platter spinning. The viscoelastic washers used with captive screws effectively isolate external linear vibration energy. Besides, a Huawei's patented technology (patent number: China 200910221868.3) is used. This technology controls the diversion of adjacent disks to reduce disk resonance.
2. Multi-level fan vibration isolation: Fan modules are fastened by pilot nails made of thermoplastic and viscoelastic materials. The nails cover sensitive disk vibration frequencies. The fan vibrations are reduced by 40% based on vertical and horizontal multi-level vibration absorption among fans, supports, and disk enclosures.
3. Reinforced chassis, disks, and guide rails: The double-deck structure improves the strength of chassis and disks by more than 20% and ensures the consistency of disk slot dimensions. Guide rails are made of die-cast zinc alloy. The shock resistance quality of the material helps reduce the vibration passed from the enclosure to disks.

To protect hardware against corrosion, Huawei works with several suppliers to apply anti-corrosion techniques to multiple modules of the S2600T. Those anti-corrosion techniques ensure that the S2600T can work correctly when air pollutants of data centers range from DC G1 to DC GX.

1. Together with other disk vendors, Huawei has developed disk anti-corrosion techniques: Electroless Nickel/Immersion Gold (ENIG) and Solder Paste VIAs (SPV), greatly improving the service life and reliability of disks used in a contaminated environment.
2. The PCB anti-corrosion process along with temperature rise and voltage distribution provides local protection, prolonging the controllers' service life and improving controllers' reliability in contaminated environments.
3. Huawei's patented online corrosion monitoring devices (patent number: China 201210519754.9) can detect data center corrosion risks and professional test devices are available to quantify corrosion levels of data centers in 72 hours.
4. Huawei's patented anti-corrosion strainers (patent number: China 201110314472.0) are installed on chassis to prevent corrosion. To prevent equipment room-level corrosion, Huawei provides a chemical filtering solution.

Quick Self-Healing

Fault detection and self-healing to ensure system reliability

The S2600T employs a multi-level error tolerance design for disks and provides various measures to ensure reliability, including online disk diagnosis, disk health analyzer (DHA), bad sector background scanning, and bad sector repair. Based on a hot spare policy, RAID 2.0+ automatically reserves a certain amount of hot spare space in a disk domain. If the S2600T detects an uncorrectable media error in an area of a disk or finds that an entire disk fails, the S2600T automatically reconstructs the affected data blocks and writes the reconstructed data to the hot spare space of other disks, implementing quick self-healing.

1. Disk fault diagnosis and warning by the Disk Health Analyzer (DHA): As important mechanical components of a storage system, disks become aged gradually due to long-term uninterrupted operation. The disk failure rate increases as time goes by. The DHA subsystem of the S2600T establishes disk fault modules to monitor critical disk indexes and assess disk health using advanced algorithms. The subsystem allows users to set proper performance thresholds based on application scenarios of disks. When the values of disk health parameters are lower than preset thresholds, related disks can be replaced to eliminate risks.
2. Background bad sector scan: Data on some sectors cannot be read during normal data access. Those sectors encounter Latent Sector Errors (LSEs). If a disk has bad sectors, it cannot report information about them to hosts. Bad sectors can only be detected during data reads and writes. The background bad sector scan function of the S2600T can detect and recover LSEs without affecting services or disk reliability, reducing data loss risks. This function allows users to set proper scan policies in specific scan periods based on physical parameters of disks.
3. Bad sector repair: After bad sectors are detected, the system tries to read data on them again, or uses the RAID mechanism to reconstruct data on bad sectors, writes the reconstructed data to another disk, and then employs the remapping function of disks to recover the data, preventing disk failure and reconstruction.

Fast reconstruction to reduce dual-disk failure probability

In the recent 10 years of disk development, disk capacity growth outpaces performance improvement. Nowadays, 4 TB disks are commonly seen in enterprise and consumer markets. 5 TB disks will come into being in the second quarter of 2014. Besides, even a high-performance SAS disk specific to the enterprise market can provide up to 1.2 TB capacity.

Rapid capacity growth confronts traditional RAID with a serious issue: reconstruction of a single disk, which required only dozens of minutes 10 years ago, now requires 10-odd hours or even dozens of hours. The increasingly longer reconstruction time leads to the following problem: A storage system that encounters a disk failure must stay in the degraded state without error tolerance for a long time, exposed to a serious data loss risk. It is common that data loss occurs in a storage system under the dual stress imposed by services and data reconstruction.

Based on underlying block virtualization, RAID 2.0+ overcomes the performance bottleneck seen in target disks (hot spare disks) that are used by traditional RAID for data reconstruction. As a result, the write bandwidth provided for reconstructed data flows is no longer a reconstruction speed bottleneck, greatly accelerating data reconstruction, decreasing dual-disk failure probability, and improving storage system reliability.

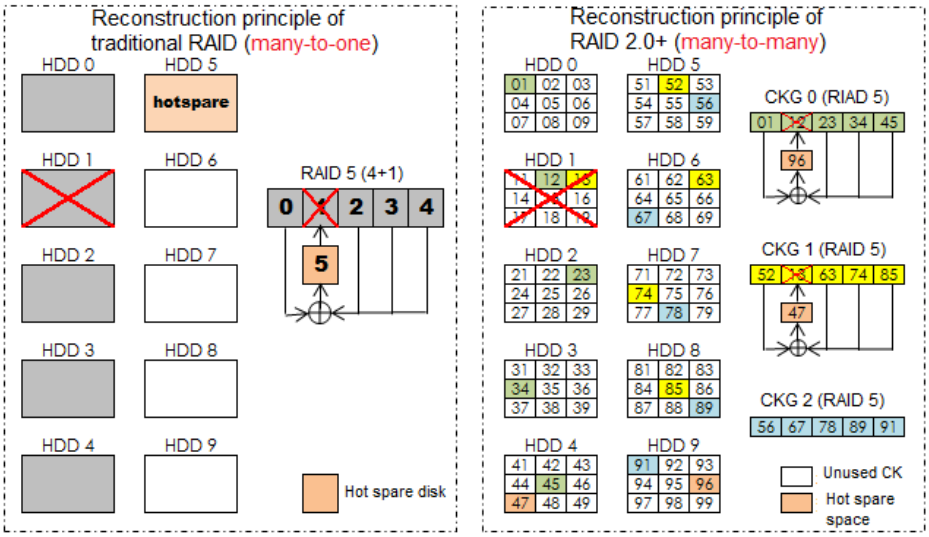
In the schematic diagram of RAID2.0 +, if HDD 1 fails, its data is reconstructed based on a CK granularity, where only the allocated CKs (CK12 and CK13 in the figure) are

reconstructed. All disks in the storage pool participate in the reconstruction. The reconstructed data is distributed onto multiple disks (HDDs 4 and 9 in the figure).

The following figure compares the reconstruction principle of traditional RAID with that of RAID 2.0+.

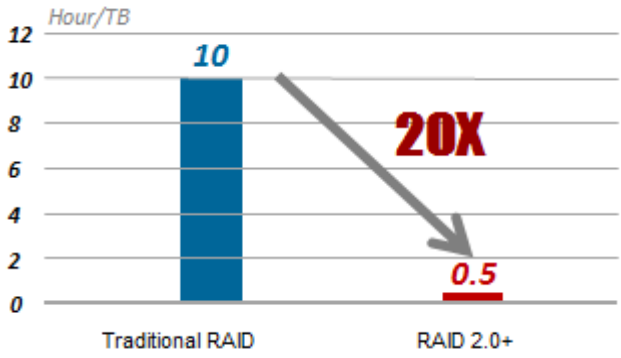
- In the schematic diagram of traditional RAID, HDDs 0 to 4 compose a RAID 5 group, and HDD 5 serves as a hot spare disk. If HDD 1 fails, an XOR algorithm is used to reconstruct data based on HDDs 0, 2, 3, and 4, and the reconstructed data is written onto HDD 5.
- In the schematic diagram of RAID2.0 +, if HDD 1 fails, its data is reconstructed based on a CK granularity, where only the allocated CKs (CK12 and CK13 in the figure) are reconstructed. All disks in the storage pool participate in the reconstruction. The reconstructed data is distributed onto multiple disks (HDDs 4 and 9 in the figure).

Figure 6-3 Reconstruction principle comparison between traditional RAID technology and RAID 2.0



With great advantages in reconstruction, RAID 2.0+ enables the S2600T to outperform traditional storage systems in terms of reconstruction. The following figure compares the time that a traditional storage system and the S2600T spend in reconstructing 1 TB data in an NL-SAS large-capacity disk environment.

Figure 6-4 Reconstruction time comparison between traditional RAID and RAID 2.0+



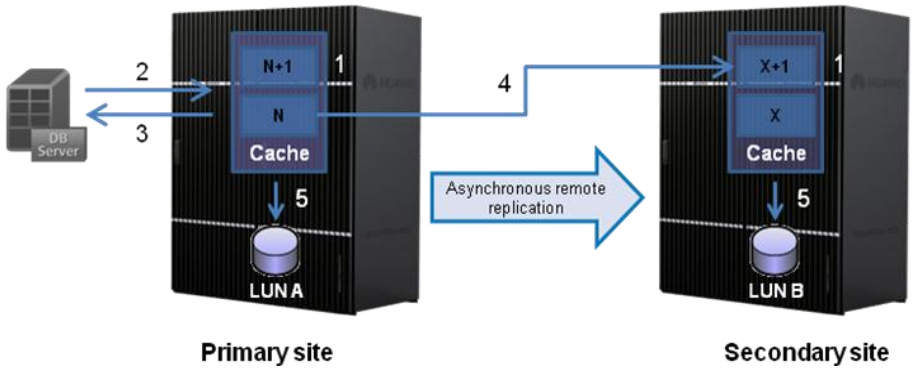
7 Reliable: Second-Level Disaster Recovery

The S2600T adopts the innovative multi-timestamp cache technology in asynchronous remote replication. Both data in caches and I/Os interacting with caches have timestamps. During replication and synchronization, data with specific timestamps is directly replicated from the primary LUN cache to the secondary LUN, reducing latency and decreasing impact on storage performance imposed by asynchronous remote replication snapshots. In this way, the second-level synchronization is achieved.

Asynchronous remote replication does not synchronize data in real time. Therefore, the RPO is determined by the synchronization period that is specified by a user based on an actual application scenario. The synchronization period ranges from 3 seconds to 1440 minutes. The working principle is as follows:

1. After an asynchronous remote replication relationship is set up between a primary LUN at the primary site and a secondary LUN at the remote replication site, initial synchronization is implemented to copy all the data of the primary LUN to the secondary LUN.
2. If the primary LUN receives a write request from the production host during the initial synchronization, data is written only to the primary LUN.
3. After the initial synchronization is completed, the data status of the secondary LUN becomes **Synchronized**. Then, I/Os are processed as follows:

Figure 7-1 Asynchronous remote replication with the multi-timestamp technology



- Incremental data is automatically synchronized from the primary site to the secondary site based on the user-defined synchronization period that ranges from 3 seconds to 1440 minutes. (If the synchronization type is **Manual**, a user needs to

manually trigger the synchronization.) When a replication period starts, data parts with new timestamps (TP_{N+1} and TP_{X+1}) are respectively generated in the caches of the primary LUN (LUN A) and the secondary LUN (LUN B).

- The primary site receives a write request from a production host.
- The primary site writes the requested data to the part with the TP_{N+1} timestamp and sends an acknowledgement to the host.
- During synchronization, data of the part with the TP_N timestamp in the cache of LUN A in the previous synchronization period is replicated to the part with the timestamp of TP_{X+1} in the cache of LUN B.
- After the data is synchronized, parts with the TP_N and TP_{X+1} timestamps are moved from caches of LUN A and LUN B to disks based on disk flushing policies and wait the next synchronization period.

NOTE

- Part: logical space in a cache that manages data written within a specific period of time. (Data size is not restricted.)
- In application scenarios where the RPO is low, the asynchronous remote replication period is short. The cache of the S2600T can store all data of multiple parts with timestamps. If the replication period is lengthened or the replication is interrupted because the bandwidth for host services or disaster recovery is abnormal, data in the cache is automatically moved to disks based on disk flushing policies and data consistency is protected. During replication, the data is directly read from disks.

Split mirror, switchover of primary and secondary LUNs, and rapid fault recovery

The asynchronous remote replication supports splitting, synchronization, primary/secondary switchover, and recovery after disconnection.

A split asynchronous remote replication session will not be periodically synchronized. Users can manually start synchronization. Then the session is synchronized based on a preset synchronization policy (manual or automatic).

Asynchronous replication supports three synchronization modes:

- Manual: Users need to manually synchronize data from a primary LUN to a secondary LUN. In manual synchronization, users can update data to the primary LUN as desired. That is, users can determine that the data of the secondary LUN is the copy of the primary LUN at a desired time point.
- Timed wait when synchronization begins: When a data synchronization process starts, the system starts timing. After one synchronization period, the system starts synchronization and timing again. After a specified period of time since the start of the latest synchronization process, the system automatically copies data from the primary LUN to the secondary LUN.
- Timed wait when synchronization ends: The system starts timing for the next synchronization session after the last synchronization session ends. In this mode, when a data synchronization session ends, the system waits for the duration preset by users. When the duration elapses, the system automatically synchronizes data from the primary LUN to the secondary LUN again.

You can choose a synchronization type that best fits your needs.

Continuous protection of data on secondary LUNs

The asynchronous remote replication supports continuous protection of the data on the secondary LUN. At the primary site, hosts have limited permission to read and write data on the secondary LUN, fully protecting data on the secondary LUN. If synchronization is interrupted, data in the TPx period can be restored to the primary LUN to overwrite data in

the TP_{X+1} period, rolling back the primary LUN (LUN B) to the point in time before the last synchronization session starts.

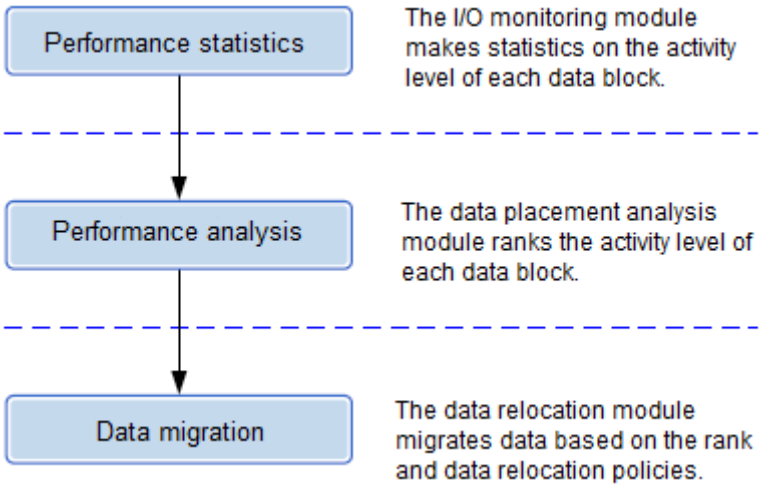
After a primary/secondary switchover, determine whether to recover data on the original secondary LUN based on the data availability of the original secondary LUN. If the data on the original secondary LUN is available, do not roll back the original secondary LUN; if the data on the original secondary LUN is unavailable, roll back the original secondary LUN to the point in time before the last synchronization session started. The data recovery is implemented in the background. After the recovery is complete, a message is displayed to prompt users.

8 Intelligent: SmartTier

The S2600T supports Huawei's self-developed SmartTier feature. This feature is used to implement automatic storage tiering. SmartTier stores right data onto right media at right time. SmartTier improves storage system performance and reduces storage costs to meet enterprises' requirements on both performance and capacities. By preventing historical data from occupying expensive storage media, SmartTier ensures effective investment and eliminates energy consumption caused by useless capacities, reducing TCO and optimizing cost-effectiveness.

SmartTier performs intelligent data storage based on LUNs and identifies LUNs based on a data migration granularity from 512 KB to 64 MB. The data migration granularity is called extent. SmartTier collects statistics on and analyzes the activity levels of data based on extents and matches data of various activity levels with storage media. Data that is more active will be promoted to higher-performance storage media (such as SSDs), whereas data that is less active will be demoted to more cost-effective storage media with larger capacities (such as NL-SAS disks). The data migration process of SmartTier consists of performance statistics collection, performance analysis, and data migration, as shown in the following figure:

Figure 8-1 Three phases in data processing by SmartTier



Performance statistics collection and performance analysis are automated by the storage system based on users' configuration, and data migration is initiated manually or by a user-defined scheduled policy.

The I/O monitoring module collects performance statistics.

SmartTier allows user-defined I/O monitoring periods. During the scheduled periods, it collects statistics on data reads and writes. Activity levels of data change throughout the data life cycle. By comparing the activity level of one data block with that of another, the storage system determines which data block is more or less frequently accessed. The activity level of each extent is obtained based on the performance indicator statistics of data blocks.

The working principle is as follows:

1. During scheduled I/O monitoring periods, each I/O is recorded to serve as data sources for performance analysis and forecasting. The following information is recorded based on extents: data access frequency, I/O size, and I/O sequence.
2. The I/O monitoring module records the I/Os of each extent based on memories. Each controller can monitor a maximum of 512 TB storage space.
3. The I/O monitoring module performs weighting for I/O statistics on a daily basis to weaken the impact of historical services on current services.

The data placement analysis module implements performance analysis.

The collected performance statistics are analyzed. This analysis produces rankings of extents within the storage pool. The ranking progresses from the most frequently accessed extents to the least frequently accessed extents in the same storage pool. Note that only extents in the same storage pool are ranked. Then a data migration solution is created. Before data migration, SmartTier determines the direction of relocating extents according to the latest data migration solution.

The working principle is as follows:

1. The data placement analysis module determines the I/O thresholds of extents on each tier based on the performance statistics of each extent, the capacity of each tier, and the access frequency of each data block. The hottest data blocks are stored to the tier of the highest performance.
2. Extents that exceed the thresholds are ranked. The hottest extents are migrated first.
3. During data placement, a policy is made specifically for SSDs and another policy is made to proactively migrate sequence-degraded extents from SSDs to HDDs.

The data relocation module migrates data.

Frequently accessed data (hotspot data) and seldom accessed data (cold data) are redistributed after data migration. Random hotspot data is migrated to the high-performance tier and performance tier, and non-hotspot data and high-sequence data are migrated to the capacity tier, meeting service performance requirements. In addition, the TCO of the storage system is minimized and the costs of users are reduced.

SmartTier has two migration triggering modes: manual and automatic. The manual triggering mode has a higher priority than the automatic one. In manual triggering mode, data migration can be triggered immediately when necessary. In automatic triggering mode, data migration is automatically triggered based on a preset migration start time and duration. The start time and duration of data migration are user-definable.

In addition, SmartTier supports three levels of data migration speeds: high, medium, and low. The upper limits of the low-level, medium-level, and high-level data migration rates are 10 MB/s, 20 MB/s, and 100 MB/s respectively.

The working principle is as follows:

1. The data relocation module migrates data based on migration policies. In the user-defined migration period, data is automatically migrated.

2. The data relocation module migrates data among various storage tiers based on migration granularities and the data migration solution generated by the data placement analysis module. In this way, data is migrated based on activity levels and access sequences.
3. The data relocation module dynamically controls data migration based on the current load of a storage pool and the preset data migration speed.
4. The minimum unit for data migration is extent. Service data can be correctly accessed during migration. Relocating an extent is to read data from the source extent and write the data to the target extent. During data migration, read I/Os read data from the source extent while write I/Os write data to both the source and target extents. After data migration, the metadata of the source and target extents is modified. Then read and write I/Os access the target extent. The source extent is released.

9 Intelligent: SmartThin

The S2600T supports Huawei's self-developed SmartThin feature. This feature is used to implement thin provisioning. SmartThin allows users to allocate desired capacities to LUNs when creating LUNs. When LUNs are being used, storage capacities are allocated on demand to improve storage resource utilization and meet the requirements of growing services. SmartThin does not allocate all space out, but presents users a virtual storage space larger than the physical storage space. In this way, users see larger storage space than the actual storage space. When users begin to use storage space, SmartThin provides only required space to users. If the storage space is insufficient, SmartThin expands the capacity of the back-end storage unit. The whole expansion process is transparent to users and causes no system downtime.

If the actual amount of data is larger than expected, LUN space can be adjusted dynamically. Free space can be allocated to any LUN that needs space. In this way, storage space utilization and effectiveness are improved. In addition, LUN space can be adjusted online without affecting services.

SmartThin creates thin LUNs based on RAID 2.0+ virtual storage resource pools. Thin LUNs and thick LUNs coexist in a same storage resource pool. Thin LUNs are logical units created in a thin pool. They can be mapped and then accessed by hosts. The capacity of a thin LUN is not its actual physical space, but only a virtual value. Only when the thin LUN starts to process an I/O request, it applies for physical space from the storage resource pool based on the COW policy.

SmartThin allows a host to detect a capacity larger than the actual capacity of a thin LUN. The capacity detected by a host is the capacity that a user can allocate to the thin LUN, namely the volume capacity (virtual space) displayed on the host after a thin LUN is created and mapped to the host. The actual capacity of a thin LUN refers to the physical space actually occupied by a thin LUN. SmartThin hides the actual capacity of the thin LUN from the host and provides only the nominal capacity of the thin LUN.

In addition, SmartThin allows users to create a thin LUN whose capacity is larger than the maximum available physical capacity of a storage resource pool. For example, if the maximum physical capacity of a storage resource pool is 2 TB, SmartThin allows users to create a thin LUN larger than 10 TB.

SmartThin uses the capacity-on-write and direct-on-time technologies to respond to read and write requests from hosts to thin LUNs. Capacity-on-write is used to allocate space upon writes, and direct-on-time is used to redirect data.

1. Direct-on-time

   When a thin LUN receives a write request from a host, the thin LUN uses direct-on-time to determine whether the logical storage location of the request is allocated with an actual storage location. If the actual storage location is not allocated, a space allocation task is triggered with a minimum grain of 64 KB. Then data is written to the newly allocated actual storage location.

2. Capacity-on-write

   Because capacity-on-write is used, the relationship between the actual storage location and logical storage location of data is not calculated using the formulas, but is determined by mappings based on capacity-on-write. Therefore, when a thin LUN is read or written, the relationship between the actual storage location and logical storage location must be updated based on the mapping table. A mapping table is used to record mappings between actual storage locations and logical storage locations. A mapping table is dynamically updated in the write process and is queried during the read process. Therefore, direct-on-time is divided into read direct-on-time and write direct-on-time.

   - Read direct-on-time: After a thin LUN receives a read request from a host, it queries the mapping table. If the logical storage location of the read request is assigned an actual storage location, the thin LUN redirects the logical storage location to the actual storage location, reads data from the actual storage location, and returns the read data to the host. If the logical storage location of the read request is not assigned an actual storage location, the thin LUN sets data at the logical storage location to all zeros and returns all zeros to the host.
   - Write direct-on-time: After a thin LUN receives a write request from a host, it queries the mapping table. If the logical storage location of the write request is assigned an actual storage location, the thin LUN redirects the logical storage location to the actual storage location, writes data to the actual storage location, and returns an acknowledgement to the host indicating a successful data write. If the logical storage location of the write request is not assigned an actual storage location, the thin LUN performs operations based on capacity-on-write.

SmartThin supports the online expansion of a single thin LUN and a single storage resource pool. The two expansion methods do not affect services running on a host.

- The expansion of a single thin LUN is to expand the nominal storage space of the thin LUN. After the nominal storage space of a thin LUN is modified, SmartThin provides the new nominal storage space of the thin LUN to the host. Therefore, the volume capacity (virtual space) displayed on the host is the capacity after expansion. In the expansion process, the original storage location is not adjusted. If new data needs to be written to the newly added thin LUN storage space, the thin LUN applies for physical space from the storage resource pool based on capacity-on-write.
- The expansion of a storage resource pool is a capability provided by RAID 2.0+ storage virtualization. Storage capacities are expanded without affecting services running on hosts. In addition, SmartMotion balances data among all the disks including newly added disks in the storage resource pool.

SmartThin provides two methods of space reclamation: standard SCSI command (**unmap**) reclamation and all-zero data space reclamation. The working principles of these two methods are described as follows:

- Standard SCSI command reclamation: When a virtual machine is deleted, a host issues the **unmap** command using the SCSI protocol. After receiving this command, SmartThin uses direct-on-time to search for the actual storage location that corresponds to the logical storage location to be released on a thin LUN, releases the actual storage location on the thin LUN to a storage resource pool, and removes the mapping from the mapping

table. To use this space reclamation method, applications on hosts must be able to issue the **unmap** command. VMware, SF, and Windows 2012 support the **unmap** command.

- All-zero data space reclamation: When receiving the write request from a host, SmartThin determines whether data blocks contained in the write request are all zeros. If the logical storage location that issues the all-zero data space is not allocated with an actual storage location, SmartThin returns a message indicating a successful data write to the host without space allocation. If the logical storage location that issues the all-zero data space is allocated with an actual storage location, SmartThin releases the actual storage location from the thin LUN to the storage resource pool, removes the mapping from the mapping table, and returns a message indicating a successful data write to the host. This space reclamation method does not require any special commands from hosts.

10 Intelligent: SmartMigration

The S2600T employs LUN migration to provide intelligent data migration. Services on a source LUN can be completely migrated to the target LUN without interrupting ongoing services. In addition to service migration within a storage system, the LUN migration feature also supports service migration between a Huawei storage system and a compatible heterogeneous storage system.

The LUN migration feature provided by the S2600T is called SmartMigration.

SmartMigration replicates all data from a source LUN to a target LUN and uses the target LUN to completely replace the source LUN after the replication is complete. Specifically, all internal operations and requests from external interfaces are transferred from the source LUN to the target LUN transparently.

Implementation of SmartMigration has two stages:

- Service data synchronization

  Ensures that data is consistent between the source and target LUNs after service migration.
- LUN information exchange

  Enables the target LUN to inherit the WWN of the source LUN without affecting host services.

SmartMigration applies to:

- Storage system upgrade

  SmartMigration works with SmartVirtualization to migrate data from legacy storage systems (storage systems from Huawei or other vendors) to new Huawei storage systems to improve service performance and data reliability.
- Service performance tuning

  SmartMigration can be used to improve or reduce service performance. It can migrate services either between two LUNs that have different performances within a storage system, or between two storage systems that have different configurations.
- Service migration within a storage system

  When the performance of a LUN that is carrying services is unsatisfactory, you can migrate the services to a LUN that provides higher performance on the same storage system to boost service performance. For example, if a user requires quick read/write capabilities, the user can migrate services from a LUN created on low-speed storage media to a LUN created on high-speed storage media. Conversely, if the priority of a type of services decreases, you can migrate the services to a low-performance LUN to release the high-performance LUN resources for other high-priority services to improve storage system serviceability.

- Service migration between storage systems

  When the performance of an existing storage system fails to meet service requirements, you can migrate services to a storage system that provides higher performance. Conversely, if services on an existing storage system do not need high storage performance, you can migrate those services to a low-performance storage system. For example, cold data can be migrated to entry-level storage systems without interrupting host services to reduce operating expense (OPEX).

- Service reliability adjustment

  SmartMigration can be used to adjust service reliability of a storage system.

  To enhance the reliability of services on a LUN with a low-reliability RAID level, you can migrate the services to a LUN with a high-reliability RAID level. If services do not need high reliability, you can migrate them to a low-reliability LUN.

  Storage media offer different reliabilities even when configured with the same RAID level. For example, when the same RAID level is configured, SAS disks provide higher reliability than NL-SAS disks and are more often used to carry important services.

- LUN type change

  SmartMigration enables flexible conversion between thin LUNs and thick LUNs without interrupting host services.

11 Efficient: SmartQoS

The S2600T supports Huawei's self-developed SmartQoS feature. This feature is used to ensure the QoS. SmartQoS intelligently schedules and allocates computing resources, cache resources, concurrent resources, and disk resources of a storage system, meeting the QoS requirements of services that have different priorities.

SmartQoS uses the following technologies to ensure the quality of data services:

- **I/O priority scheduling:** Service response priorities are divided based on the importance levels of different services. When allocating system resources, a storage system gives priority to the resource allocation requests initiated by services that have the high priority. If resources are in shortage, more resources are allocated to services that have the high priority to maximize their QoS. Currently, three priorities are available: high, medium, and low.
- **I/O traffic control**: Based on a user-defined performance control goal (IOPS or bandwidth), the traditional token bucket mechanism is used to control traffic. I/O traffic control prevents specific services from generating excessive large traffic that affects other services.
- **I/O performance assurance**: Based on traffic suppression, a user is allowed to specify the lowest performance goal (minimum IOPS/bandwidth or maximum latency) for a service that has a high priority. If the minimum performance of the service cannot be ensured, the storage system gradually increases the I/O latency of low-priority services, thereby restricting the traffic of low-priority services and ensuring the lowest performance goal of high-priority services.

The I/O priority scheduling is implemented based on storage resource scheduling and allocation. In different application scenarios, the performance of a storage system is determined by the consumption of storage resources. Therefore, the system performance is optimized as long as resources, especially bottleneck resources, are properly scheduled and allocated. The I/O priority scheduling technique monitors the usage of computing resources, cache resources, concurrent resources, and disk resources. If a resource bottleneck occurs, resources are scheduled to meet the needs of high-priority services to the maximum. In this way, the QoS of mission-critical services is ensured in different scenarios.

The I/O priority scheduling technique employed by SmartQoS schedules critical bottleneck resources on I/O paths. Those resources include **computing resources**, **cache resources**, **concurrency resources**, and **disk resources**. Scheduling policies are implemented based on user-defined LUN priorities. The priority of a LUN is determined by the importance of applications served by the LUN. There are three LUN priorities available: **high**, **medium**, and **low**.

The I/O priority scheduling technique controls the allocation of front-end concurrency resources, CPU resources, cache resources, and back-end disk resources to control the response time of each schedule object.

- Priority scheduling of **front-end concurrency resources** is implemented at the front end of the storage system to control concurrent access requests from hosts. A storage system's capability to process concurrent host access requests is limited. Therefore, when the maximum number of concurrent host accesses allowed by a storage system is reached, SmartQoS restricts the maximum number of concurrent accesses for each priority based on the number of LUNs of each priority running under each controller. The restriction principle is as follows: High-priority services and large-traffic services are allocated a larger number of concurrent access resources.
- Priority scheduling of **computing resources** is implemented by controlling the allocation of CPU runtime resources. Based on the weight of each of high, medium, and low priorities, SmartQoS allocates CPU runtime resources to services of each priority. When CPU resources become a performance bottleneck, priority scheduling ensures that high-priority services are allocated more CPU runtime resources.
- Priority scheduling of **cache resources** is implemented by controlling the allocation of cache page resources. Based on the weight of each priority, SmartQoS first processes page allocation requests initiated by high-priority services.
- Priority scheduling of **disk resources** is implemented by controlling the I/O delivery sequence. Based on the priorities of I/Os, SmartQoS enables high-priority I/Os to access disks first. If most I/Os are queuing on the disk side, priority scheduling of disk resources reduces the queuing time of high-priority I/Os. In this way, the overall latency of high-priority I/Os is reduced.

The priority scheduling technique employed by SmartQoS is implemented based on LUN priorities. Each LUN has a priority property, which is configured by a user and saved in a database. When a host (SCSI target) sends an I/O request to a disk array, the disk array gives a priority to the I/O request based on the priority of the LUN that will process the I/O request. Then the I/O carries the priority throughout its processing procedure.

The I/O traffic control technique of SmartQoS:

- Restricts the performance of some applications in the system by limiting the total IOPS or bandwidth of one or multiple LUNs in the storage system. This technology prevents those applications from generating high burst traffic that may affect the performance of other services in the system.
- Limits data processing resources available for data services on specific LUNs. The objects of traffic control consist of the I/O class (read, write, or read and write) and traffic class (IOPS or bandwidth). Based on the two classes, a 2-tuple (I/O class and traffic class) is obtained for traffic control specific to a certain LUN.
- Controls traffic based on the I/O class and the obtained 2-tuple. Each I/O class corresponds to one traffic control group. Each traffic control group contains a certain number of LUNs and LUN groups whose maximum traffic is restricted. The I/O class-based traffic control function is implemented based on the I/O class queue management, token allocation, and dequeue control.

The I/O latency control technique of SmartQoS ensures the minimum performance requirements of some critical services by restricting low-priority services. Users can set minimum performance requirements for high-priority services. If those requirements cannot be met, the storage system restricts low- and medium-priority services in sequence to ensure the minimum performance set for the high-priority services.

SmartQoS restricts the performance of low- and medium-priority services by gradually increasing their latency. To prevent performance instability, SmartQoS does not prolong the latency after the latency reaches the upper limit. If the actual lowest service performance becomes 1.2 times of the preset lowest performance indicator, the storage system gradually cancels the increased latency of low- and medium-priority services.

12 Efficient: SmartPartition

The S2600T supports Huawei's self-developed SmartPartition feature. This feature is used to optimize cache partitioning. The core idea of SmartPartition is to ensure the performance of mission-critical applications by partitioning core system resources. An administrator can allocate a cache partition of a specific size to an application. The storage system ensures that the application uses the allocated cache resources exclusively. Based on the actual service condition, the storage system dynamically adjusts the front- and back-end concurrent accesses to different cache partitions, ensuring the application performance of each partition. SmartPartition can be used with other QoS technologies (such as SmartQoS) to achieve better QoS effects.

Caches are classified into read caches and write caches. The read cache pre-fetches and retains data to improve the hit ratio of read I/Os. The write cache improves the disk access performance by means of combination, hitting, and sequencing. Different services need read and write caches of different sizes. SmartPartition allows users to specify read and write cache sizes for a partition, meeting cache requirements of different services.

The read cache configuration and the write cache configuration affect the I/O procedure differently. The impact on the write I/Os shows up in the phase of cache resource allocation. In this phase, the host concurrency and write cache size of a partition are determined. The reason for determining the two items in this phase is that it is the initial phase of a write procedure and the cache of the storage system is actually not occupied in this phase. The impact on read I/Os involves two aspects. The first aspect is similar to the write I/O scenario. Specifically, the storage system determines whether the host concurrency meets the requirement. If not, the storage system returns the I/Os.

The read cache is intended to control the size of the cache occupied by read data. The size of a read cache is controlled by the read cache knockout procedure. Therefore, the second aspect of the impact shows up in the read cache knockout procedure. If the read cache size of the partition does not reach the threshold, read cache resources are knocked out extremely slowly. Otherwise, read cache resources are knocked out quickly to ensure that the read cache size is below the threshold.

Compared with host applications, the processing resources of a storage system is limited. Therefore, a storage system must restrict the total host concurrency amount. For each partition, the concurrency is also restricted to ensure the QoS.

Regarding SmartPartition, the host concurrency of a partition is not fixed but calculated based on the priority weighted algorithm with the following factors taken into account:

- Number of active LUNs in the partition in the last statistics period
- Priorities of active LUNs in the partition in the last statistics period

- Number of I/Os completed by each LUN in the last statistics period
- Number of I/Os returned to hosts because the partition concurrency reaches the threshold in the last statistics period

Weighting the preceding factors not only fully uses the host concurrency capability but also ensures the QoS of a partition.

After one statistics period elapses, the concurrency capability of a partition may need to be adjusted based on the latest statistical result. The SmartPartition logic adjusts the concurrency capability based on a specific step to ensure a smooth adjustment, minimizing host performance fluctuation.

Similar to host concurrency control, back-end concurrency control is also intended to fully use system resources while ensuring the QoS of a partition. The back-end concurrency is calculated based on the priority weighted algorithm with the following factors taken into account:

- Amount of dirty data on high-priority LUNs in a partition in the last statistics period
- Disk flushing latency of LUNs in a partition in the last statistics period
- Actual disk flushing concurrency of LUNs in a partition in the last statistics period

The adjustment period and approach are similar to those of host concurrency control.