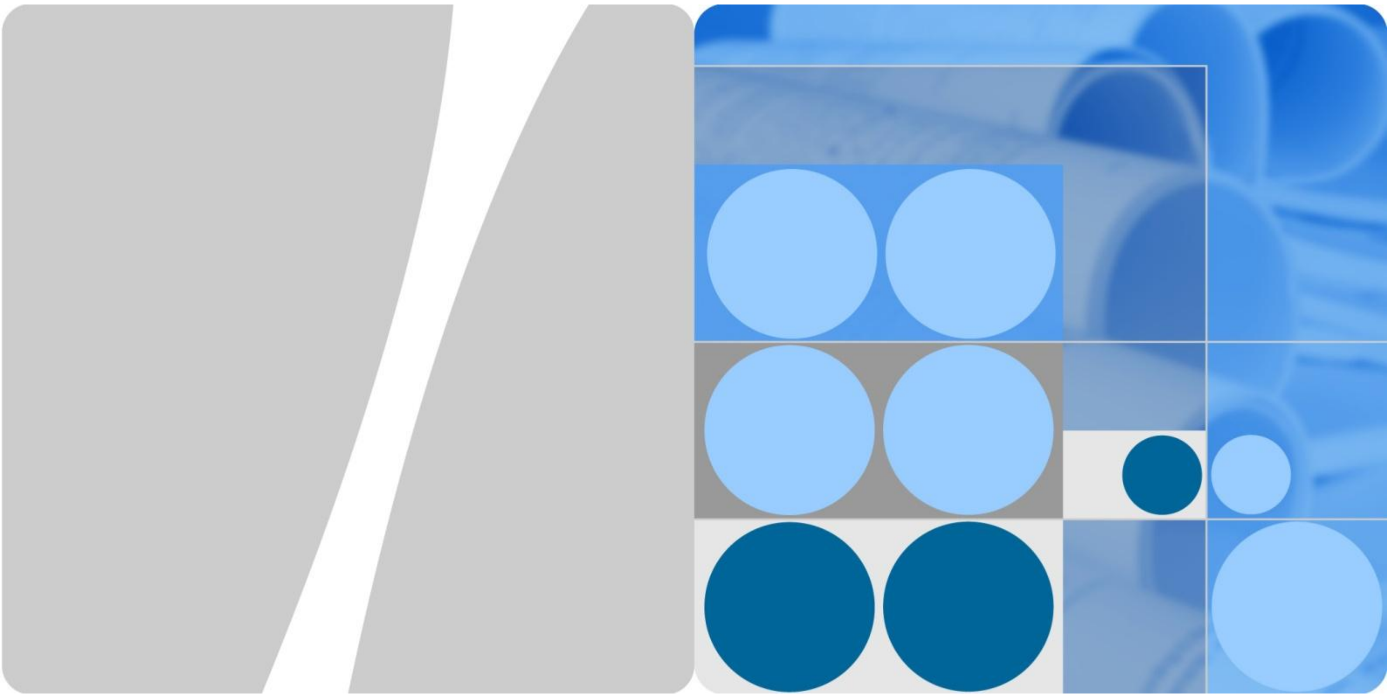


Huawei OceanStor 5300F, 5500F, 5600F, 5800F, 6800F, and 18000F V5 All Flash Storage Systems

Technical White Paper

Issue	01
Date	2017-11-30

HUAWEI TECHNOLOGIES CO., LTD.

Trademarks and Permissions

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://e.huawei.com

Change History

Date	Version	Description	Prepared By
2017-11-30	1.0	Completed the initial draft.	Li Xingliang/00335377

Contents

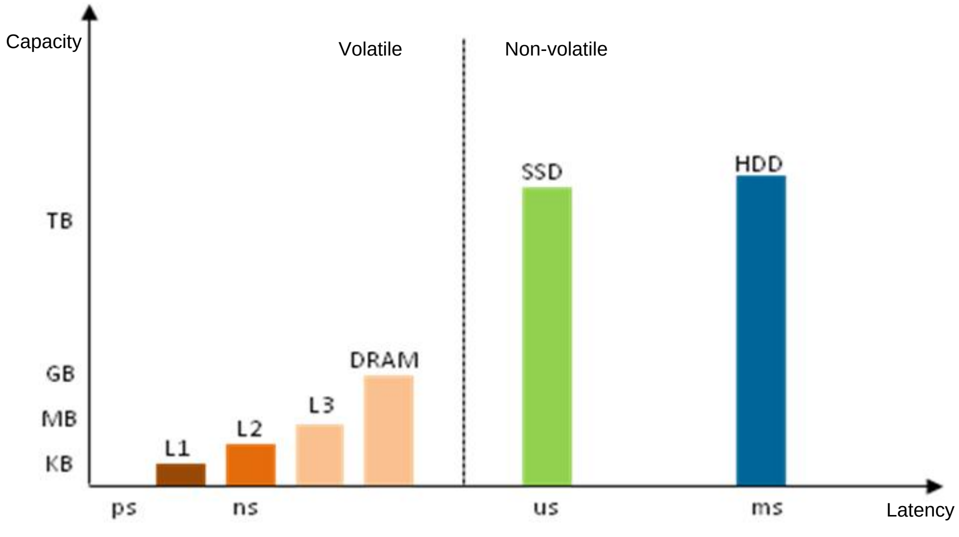
1 Background

1.1 Emerging and Advantages of SSDs

Computing, network, and storage are basic components of modern IT systems. The processor performance doubles every 18 months. The network speed increases by 10 times every 5 years (10 M/100 M/1 G/10 G/40 G/100 G). The development of hard disk drives (HDDs) stops, and enterprise-class SAS HDDs develop from 7200 rpm to 15k rpm only. However, the development of storage lags far behind that of computing and network. A large number of HDDs must be built into RAID groups to provide storage performance that matches computing and network resources, wasting storage capacity and increasing energy consumption.

Solid state disks (SSDs) provide storage capacity using semi-conductor transistors (flash chips) rather than magnetic media. Electronic read and write operations replace motor drives and mechanical seek of traditional HDDs, remarkably reducing access latency and increasing I/O access efficiency. Especially in reading and writing random small I/Os, the latency decreases from milliseconds to 100 microseconds. The following figure compares the access latencies of different levels of storage media. L1, L2, and L3 indicate cache levels in CPUs. Compared with HDDs, the SSD access latency reduces by at least two orders of magnitude.

Figure 1-1 Access latencies of different levels of storage media



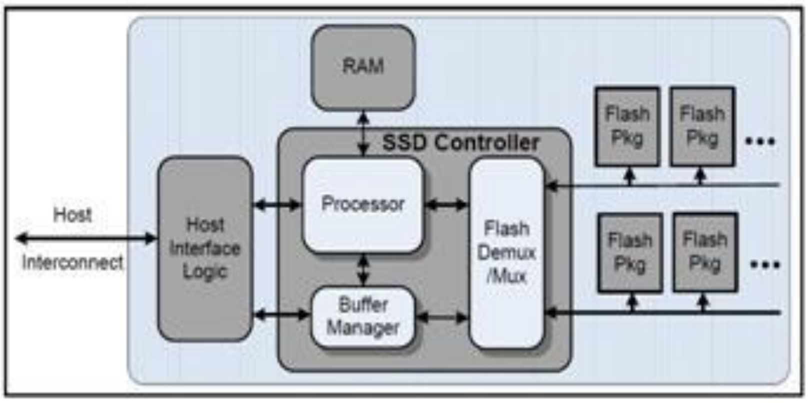
NOTE

Volatile: volatile medium; non-volatile: non-volatile medium

1.2 Architecture and Status Quo of SSDs

An SSD comprises a control unit and a storage unit (flash chips). The host ports of an SSD are the same as those of an HDD. An SSD consists of a controller, host ports, memories, and flash chips, as shown in Figure 1-2.

Figure 1-2 SSD architecture



Status quo:

The capacity ranges from hundreds of GB to several TB. 16 TB or larger SSDs enter the market.

The IOPS of a standard disk is higher than 100,000, and that of a high-performance disk is up to 300,000.

The disk price decreases. The cost per GB of SSDs is the same as that of 2.5-inch 15k rpm SAS disks.

1.3 Problems Facing SSDs in Enterprise-Class Storage Arrays

Although the introduction of SSDs improves the performance of traditional HDD arrays, increases IOPS, and lowers the latency, most vendors' converged storage products that support SSDs are designed for HDDs, and SSDs are used as high-speed HDDs. The storage products are not reconstructed and optimized based on characteristics of SSDs. Since the software is designed for HDDs, advantages of SSD performance cannot be maximized and the service life of SSDs cannot be effectively managed.

1.3.1 Performance Advantages of SSDs Cannot Be Brought into Full Play in HDD-based Storage Arrays

1. As SSDs deliver a latency of 200 microseconds, the ratio of the latency of array software processing to the I/O processing latency increases remarkably.

   The relationship among the IOPS, concurrent I/Os, and I/O latency is as follows: Outstanding/Latency.

   The latency of a high-performance enterprise-class SAS HDD is about 5 ms for 4 KB I/O random access. The latency of a SAS SSD is about 0.2 ms for 4 KB I/O random access.

   In Figure 1-3, an HDD and an SSD are respectively connected to a host with the same configurations. Then these hosts deliver single I/Os.

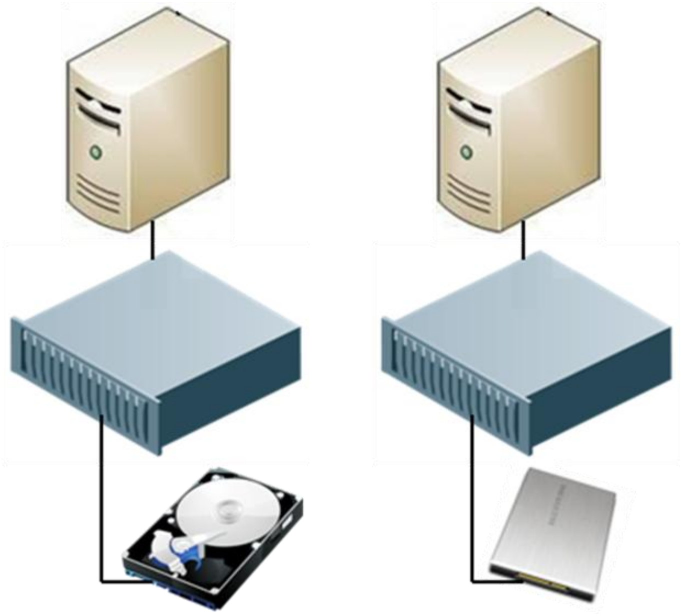
   Figure 1-3 An HDD and an SSD directly connected to hosts

   In the previous figure, the left drive is an HDD and its IOPS is calculated as follows: 1/5 ms = 200 IOPS

   In the previous figure, the right drive is an SSD and its IOPS is calculated as follows: 1/0.2 ms = 5000 IOPS.

   Use a controller to connect between the host and the HDD, and between the host and the SSD, as shown in Figure 1-4.

Figure 1-4 An HDD and an SSD connected to hosts through controllers



These controllers cause processing latency. Generally, the latency varies with workloads. The latency is about 0.2 ms for single I/Os.

The IOPS perceived by hosts is changed.

In the previous figure, the IOPS on the left side is calculated as follows: 1/(0.2 ms + 5 ms) = 192 IOPS.

In the previous figure, the IOPS on the right side is calculated as follows: 1/(0.2 ms + 0.2 ms) = 2500 IOPS.

The previous calculation shows that for an HDD, the latency slightly increases and the IOPS slightly decreases after you install a controller. For an SSD, the latency is doubled and the IOPS reduces by half after you install a controller.

Therefore, a storage array constructed by replacing HDDs with SSDs cannot bring the high performance of SSDs into full play.

2. The high IOPS of SSDs changes the bottleneck of a storage array

   For a traditional storage array, HDDs bottleneck its performance. Therefore, the IPOS and bandwidth of the entire storage array can be increased by adding HDDs. In addition, the system software of a conventional storage array is developed for eliminating the performance bottleneck caused by HDDs.

   For SSDs, the IOPS of an SSD reaches tens of thousands. The IOPS of a disk enclosure with 24 slots reaches one million. Therefore, the performance bottleneck of a storage array lies in the controller, including the CPU processing capability, system bandwidth, and system software designs and algorithms.

   For these reasons, the software and hardware designs of storage arrays using SSDs must be different from those of HDD-based storage arrays. Otherwise, the high performance of SSDs cannot be brought into full play.

3. The cache algorithm designed for HHDs does not apply to SSDs.

   An HDD supports hundreds of random IOPS, requiring an access bandwidth of 1 MB/s. Its sequential access bandwidth is around 200 MB/s. An SSD supports tens of thousands of IOPS, requiring an access bandwidth of up to 100 MB/s. Its sequential access bandwidth is around 300 MB/s.

For an HDD, the data throughput of sequential access is more than 100 times larger than that of random access. For an SSD, the data throughput of sequential access is only 2 to 4 times larger than that of random access.

Due to the remarkable throughput difference between an HDD's sequential access and random access, conventional storage arrays use various cache algorithms to ensure that data on HDDs can be accessed sequentially. However, for an SSD, the data throughput of sequential access is closer to that of random access. The cache eviction algorithm that applies to HDDs may not apply to SSDs.

1.3.2 Storage Array Software Designed for HDDs Cannot Ensure SSDs' Reliability

An HDD consists of such mechanical components as a magnetic disk, head, arm, and motor, while an SSD consists of such electrical components as a controller and flash chip. According to the data writing principle of flash chips, the service life of an SSD is determined by the number of program-erase cycles (P/E cycles).

In conclusion, HDDs and SSDs are different in principles, so their error types vary remarkably. Conventional enterprise-class storage array software is specifically designed to improve the reliability of HDDs instead of flash chips. It also lacks life management function designed for SSDs. Therefore, storage array software designed for HDDs cannot ensure SSDs' reliability.

1.4 Application of SSDs in Storage Arrays

1. Inherit the advantages of conventional enterprise-class storage arrays. Conventional enterprise-class storage arrays have accumulated experience in ensuring functions, system reliability, and maintainability, such as redundant and online replaceable active components, and abundant data protection software. Such functions of conventional storage arrays have been inherited.
2. Fully consider the performance difference between SSDs and HDDs. The performance of SSDs is two orders of magnitude higher than that of HDDs, which changes the system performance bottleneck. Therefore, the performance designed for HDDs must be reviewed.
3. Fully consider the service limit of SSDs. The systems are designed based on the service life characteristics of SSDs to effectively manage and prolong the service life, ensuring SSDs' reliability in enterprise-class storage arrays.

2 Overview

Mainstream flash storage arrays are divided into hybrid arrays, and all flash arrays based on how SSDs are used in storage arrays.

Huawei OceanStor F V5 all flash storage systems provide almost the same performance as all-flash-memory arrays to protect data security and business continuity. In the mean time, Huawei OceanStor F V5 all flash storage systems deliver six-nines reliability.

1. Inherit the advantages of conventional storage arrays. Conventional storage arrays present accumulated experience in ensuring system reliability and ease of maintenance. For example, all active components are redundant and replaceable online, and a full range of data protection software is available. All these features of conventional storage arrays have been inherited.
2. Fully consider the performance difference between SSDs and HDDs. The performance of SSDs is two orders of magnitude higher than that of HDDs, which changes elements that cause the system performance bottleneck. Therefore, the performance designed for HDDs must be reviewed.
3. Fully consider the service limit of SSDs. Huawei OceanStor F V5 all flash storage systems are designed and developed based on the service life of SSDs. The service life of SSDs is efficiently managed to ensure the reliability of SSDs in enterprise-class storage arrays.

To better serve customers and to meet customer requirements in different application scenarios, Huawei puts SSD development in storage systems to a strategic position, combining rock-solid reliability of enterprise-class storage with high performance of flash arrays.

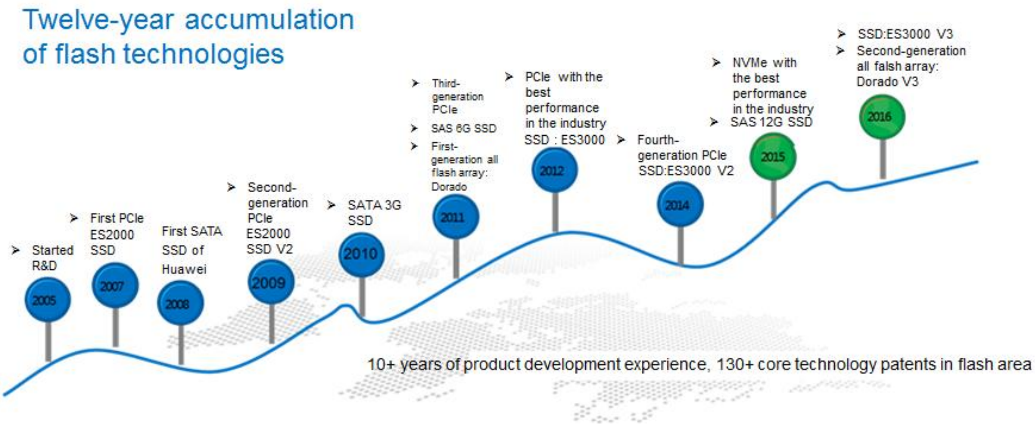
Considering SSD characteristics, Huawei optimizes and reconstructs storage systems in terms of the basic I/O process, value-added features, product hardware design, and SSDs to construct product differentiation advantages and maximize the integration of SSDs with OceanStor OS. As a result, Huawei OceanStor F V5 fully leverages SSD advantages to provide high performance, robust reliability, and rich data storage features.

3 Working Principles

Huawei OceanStor F V5 uses the RAID 2.0+ global virtualization technology to integrate with SSDs into one storage system. It is comprehensively optimized for SSDs to lower the latency to less than 1 ms, effectively manage the service life of SSDs, and integrate with verified reliability of conventional storage arrays, fully satisfying customers' requirements on performance, reliability, cost, and capacity.

3.1 Huawei's Technical Preparations in Flash Storage

Huawei is the only storage vendor that has R&D capabilities for both SSDs and storage systems. Huawei masters core technologies of SSD controller chips and has its own proprietary SSDs. The company has applied for 150 patents for its SSDs and obtained 130 plus patents.


Twelve-year accumulation of flash technologies
➢ Started R&D
2005
➢ First PCIe ES2000 SSD
2007
First SATA SSD of Huawei
2008
➢ Second-generation PCIe ES2000 SSD V2
2009
➢ SATA 3G SSD
2010
➢ Third-generation PCIe
➢ SAS 6G SSD
➢ First-generation all flash array: Dorado
2011
➢ PCIe with the best performance in the industry SSD : ES3000
2012
➢ Fourth-generation PCIe SSD:ES3000 V2
2014
➢ NVMe with the best performance in the industry
➢ SAS 12G SSD
2015
➢ SSD:ES3000 V3
➢ Second-generation all falsh array: Dorado V3
2016
10+ years of product development experience, 130+ core technology patents in flash area

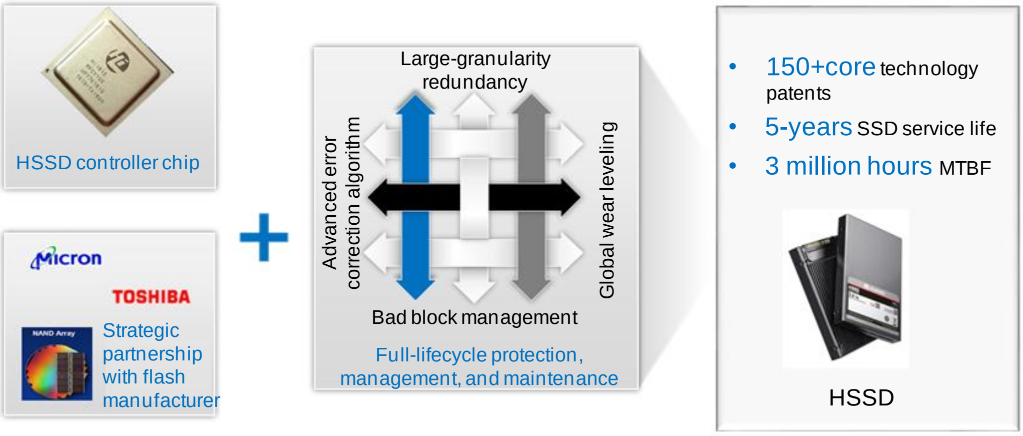
3.2 Storage Architecture Designed for Future All-Flash-Memory-based Data Centers

3.2.1 Huawei SSD

As the core component of an SSD, the SSD controller decides the performance and reliability of the SSD. Huawei SSD (HSSD) uses self-developed second-generation controllers which are designed for enterprise-class applications and provide industry-standard PCIe3.0x4, SAS3.0x2 interfaces. The controllers offer high performance with low power consumption and provide value-added storage service features. To mitigate the impact of media wear on SSDs' service life, various technologies are introduced, such as enhanced error checking and correcting (ECC) and digital signal processing, and built-in RAID, meeting the requirements for enterprise-class reliability. The controllers support the latest DDR4 and SAS 12 Gbit/s interfaces and hardware acceleration FTL, ensuring stability and short delay for enterprise-class applications.

HSSDs use the following methods to improve the service life and data reliability:

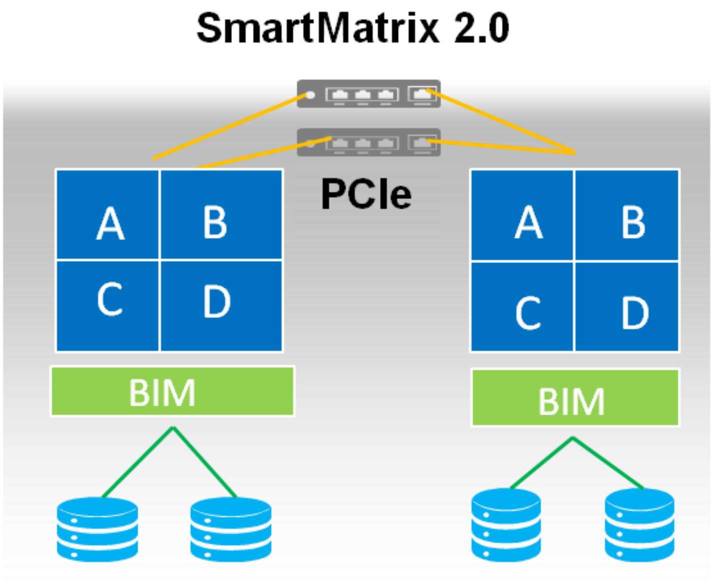
1. Enhanced ECC processing capability: As the flash process evolves from 2Xnm to 1Xnm, the proportion of error increases during program/erase cycles. The traditional widely-used BCH error correction algorithm cannot meet the requirement of flash storage. To ensure data reliability, HSSD uses LDPC and DSP algorithms that match with enterprise-class SSD controllers. The algorithms integrate hard decision, soft decision, and DSP. Huawei also cooperates with flash manufacturers for further optimization, extending the lifespan of flash by up to three to four times.
2. Precise internal processing of NAND flash: The NAND flash status is monitored in real time. Advanced technologies such as background prevention, medium-specific bad page management, category and level-based management, and component redundancy (to cope with DIE failure) further improve SSD reliability.
3. Optimized SSD management and scheduling algorithm: System algorithms like Garbage Collection (GC), wear leveling (WL), and bad block management are optimized to reduce the write amplification coefficient and wear times of SSDs.
4. System-level data protection for SSDs: The voltage monitoring module and backup power circuit protection system are used to lower data loss risks upon unexpected power failures.


HSSD controller chip
Micron
TOSHIBA
NAND Array
Strategic partnership with flash manufacturer
+
Large-granularity redundancy
Advanced error correction algorithm
Global wear leveling
Bad block management
Full-lifecycle protection, management, and maintenance
150+core technology patents
5-years SSD service life
3 million hours MTBF
HSSD

3.2.2 Brand-New SmartMatix2.0 Architecture

Huawei OceanStor 6800F V5, 18000F V5 high-end all flash storage systems boast all-PCIe 3.0 interconnection, back-end SAS 3.0, and high-speed channels and powerful computing capability of Intel Skylake CPUs, meeting increasingly demanding performance requirements.

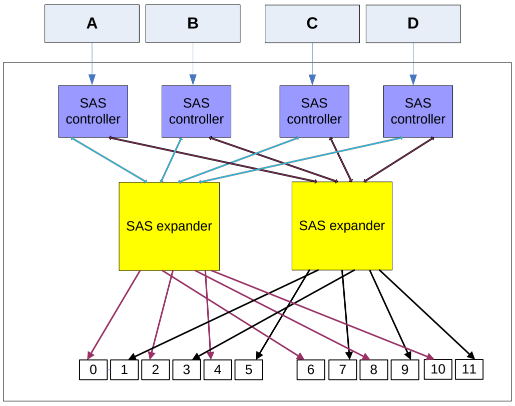
Each storage engine has four controllers, and no external switches and switch cables are required, simplifying deployment and improving reliability. The four controllers are fully interconnected through PCIe 3.0 high-speed channels on the backplane.



3.2.2.1 SAS 3.0 Back-End Full Interconnection

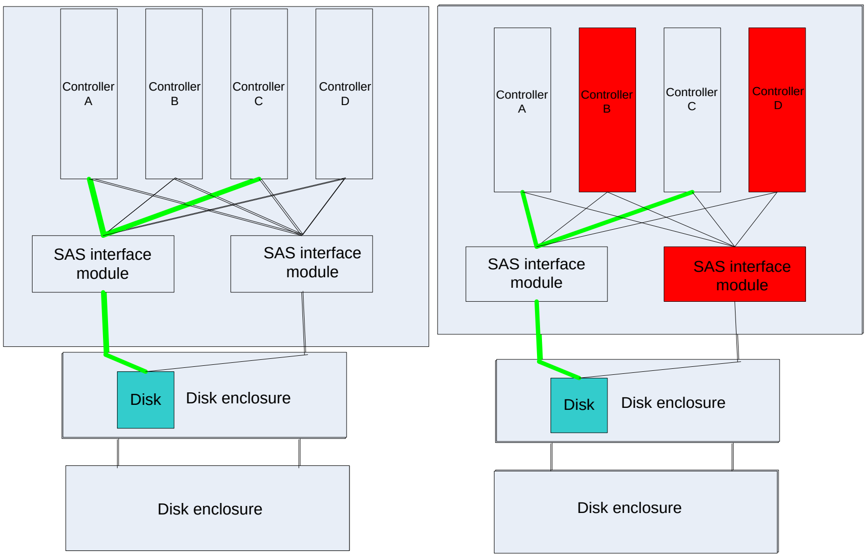
Back-end full interconnection interface modules enable controllers in engines to connect to disks or disk enclosures. If any two controllers in an engine fail, the links are up between the remaining two controllers and the disks or disk enclosures. The figure on the right illustrates the working principle of back-end full interconnection interface modules. Four SAS controllers reside in one interface module and connect to the four controllers in one engine by high-speed PCIe. In the mean time, there are two SAS expanders in one interface module. Each expander provides six 4 x 12 Gbit/s SAS 3.0 ports. The two SAS expanders are connected to four SAS controllers by high-speed cables.

Figure 3-1 SAS 3.0 back-end full interconnection interface module



As shown in the following figure, a pair of SAS interface modules is configured for a storage engine. Two SAS ports of each disk connect to two interface modules through the expansion module of the disk enclosure. As each controller is connected to each SAS interface module, if any interface module or 1 to 3 controllers are faulty, the remaining controllers are still connected to the disk. Even if controllers B and D are faulty and one SAS interface module malfunctions as shown in the figure on the right, controllers A and C can still access disks without inter-controller forwarding (the green line indicates the access path).

Figure 3-2 Back-end full interconnection

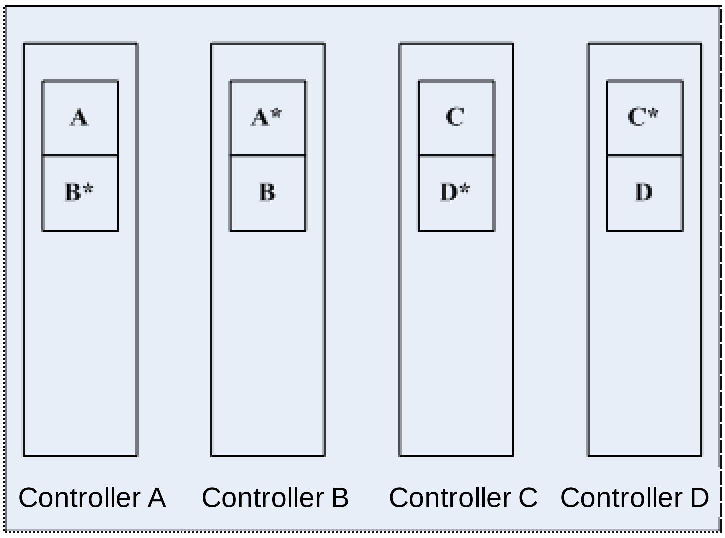


3.2.2.2 Persistent Cache

To improve performance, storage systems are configured with data caches. Mirroring is implemented between data caches of controllers so that no data in the cache will be lost even when a controller is faulty. In traditional design, two controllers mirror each other. If one controller is faulty, the image of data in the cache of the other controller will be lost. To ensure data integrity, the other controller must enter the write through mode, causing a performance drop.

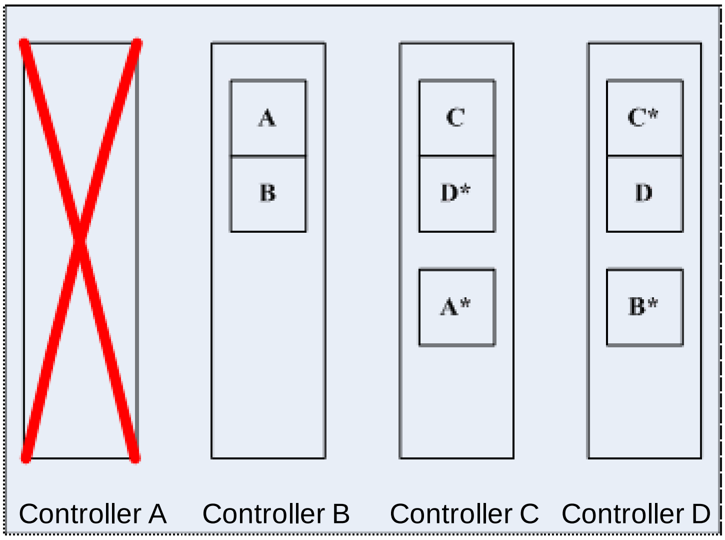
As illustrated in the following figure, each engine has four controllers, and each controller has a cache (the small squares in the following figure). To improve the reliability of cached data, each cache is mirrored to another controller. By default, caches of controllers A and B are mutually mirrored, and those of controllers C and D are mutually mirrored. If a controller is faulty, the mirroring relationship among the rest controllers is adjusted. The system can tolerate the successive failure of three controllers, which does not cause data loss or service interruption at all.

Figure 3-3 Working principle of persistent cache (1)



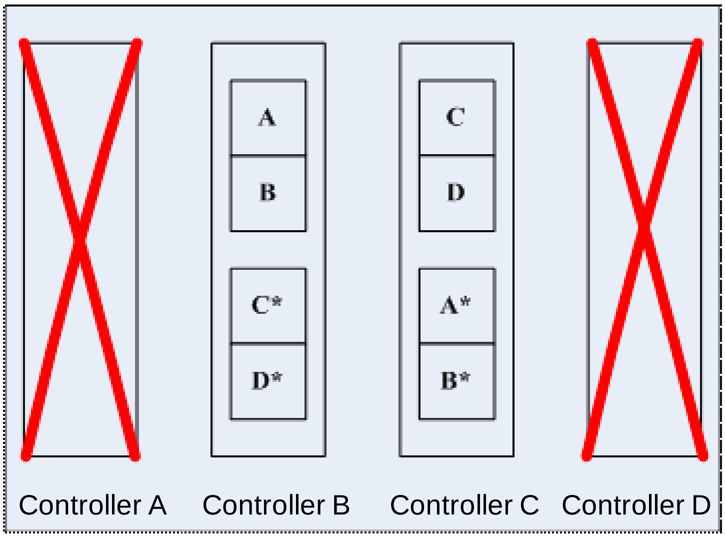
If one controller, controller A for instance, is faulty, its cache will be taken over by controller B, and then the caches on controller B will be mirrored to controllers C or D according to a controller selection algorithm. In this way, all controller caches are mirrored.

Figure 3-4 Working principle of persistent cache (2)



If another controller, controller D for instance, is faulty, its cache will be taken over by the controller that is mirrored to controller D. Controllers B and C mirror each other.

Figure 3-5 Working principle of persistent cache (3)



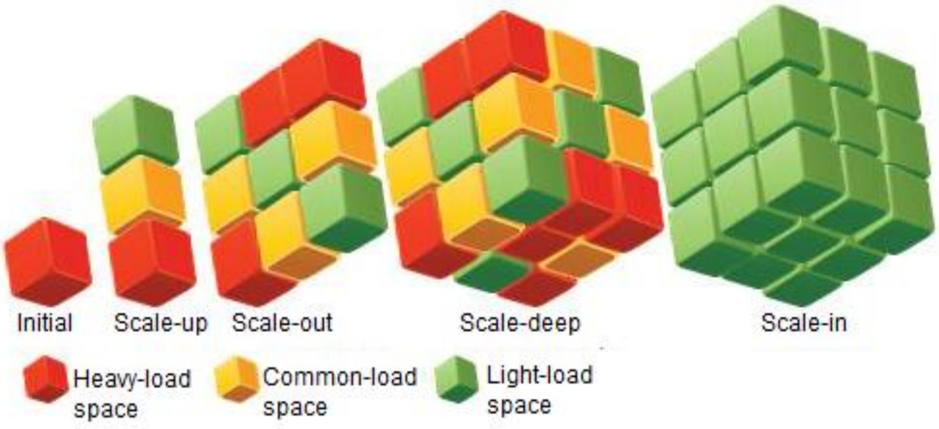
If the faulty controllers recover, the storage software distributes caches among the controllers based on the algorithm. The distribution must ensure that:

- All controllers are effectively used.
- All caches have mirrors. If all the controllers recover, the default cache distribution restores. Controllers A and B mirror each other, and controllers C and D mirror each other.

3.2.3 4S Flexible Expansion

Based on the Smart Matrix, OceanStor F V5 all flash storage systems deliver 4S elastic scalability and significantly improve system resource utilization.

Figure 3-6 4S elastic scalability



- **Scale-up**

  Expands system performance, storage capacity, and connection capability by adding memory, disks, and service ports.

- **Scale-out**

  Using the online expansion of system engines, the Smart Matrix couples related resources and adds storage resources on demand to ensure linear capacity and performance growth, addressing increasing service requirements.

  To expand to four controllers, backplane interconnection (high-end storage) or direct-connection networking (mid-range storage) is used. To expand to more than four controllers, two redundant data switches are used for data witching. This kind of cluster supports up to four engines. Dual switching links are used to ensure redundancy.

- **Scale-deep**

  Heterogeneous storage systems are integrated and managed by OceanStor F V5 all flash storage systems in a unified manner, eliminating information islands and protecting customers' investment.

- **Scale-in**

  Data is intelligently and evenly distributed onto intelligent volumes through automatic load balancing without adding storage resources, improving resource utilization and achieving automatic expansion.

Customers can expand the storage pool by simply inserting new disks, and then the system automatically adjusts data distribution to store data evenly among all disks. Also, customers can expand a volume by merely specifying the desired volume size. The system then automatically allocates the needed storage space from the storage pool and adjusts data distribution among the volume to balance volume data on all disks.

3.2.4 SSD-oriented RAID 2.0

The term redundant array of independent disks (RAID) was first defined by the University of California, Berkeley in 1987. The basic idea of RAID is to combine multiple independent physical disks based on a certain algorithm to form a virtual logical disk that provides a larger capacity, higher performance, or better data error tolerance.

As a mature and reliable data protection standard, RAID has always been used as a basic technology by storage systems since its existence. However, with rapid growth of data storage needs and emergence of high-performance applications in recent years, traditional RAID gradually exposes its defects.

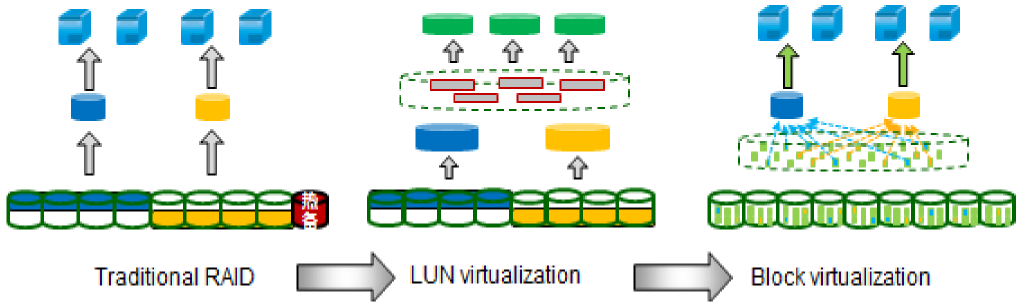
A large number of SSDs are used in storage arrays to meet high-performance storage requirements. Traditional RAID is subject to the number of disks. In the era of soaring data growth, traditional RAID fails to meet enterprises' needs for unified and flexible resource scheduling. Besides, as disk capacity increases, disk-based data management becomes increasingly inefficient.

The capacity of SSDs is gradually increasing. Reconstructing a 3.84 TB SSD in a RAID 5 group (8D+1P) takes about 20 hours. The reconstruction process consumes system resources, decreasing the overall performance of the application system. If a user restricts the reconstruction priority in return for timely application response, the reconstruction time will be even longer. During the time-consuming reconstruction, a large number of access operations may cause the failure of other disks in the RAID group, greatly increasing the disk failure probability and data loss risk.

To resolve the preceding issues of traditional RAID and follow the virtualization trend, many storage vendors adopt alternatives for traditional RAID as follows:

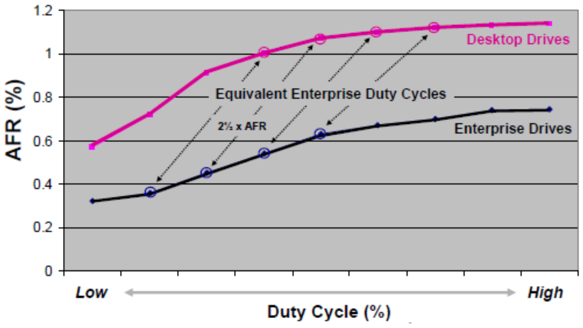
- LUN virtualization: Based on traditional RAID, some storage vendors such as EMC and HDS divide RAID groups into more fine-grained units and combine these units to form storage space accessible to hosts.
- Block virtualization: Some storage vendors such as Huawei and HP 3PAR divide the space of disks that belong to a storage pool into small-granularity data blocks and create RAID groups based on these data blocks. This approach allows data to be evenly distributed onto all disks in the storage pool and enables resources to be managed in the form of data blocks.

Figure 3-7 RAID technology evolution

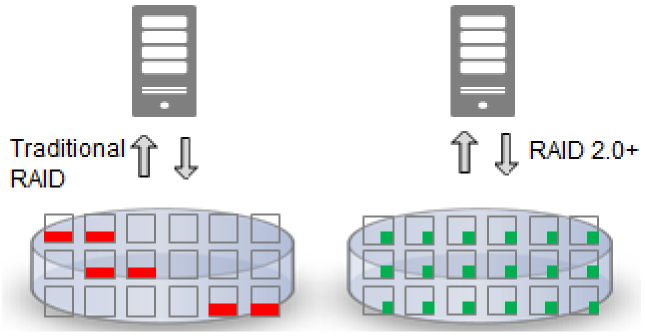


3.2.4.1 Dynamic Load Balancing

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. In the following figure, **Duty Cycle** indicates the percentage of disk working time to total disk power-on time, and **ARF** indicates the annual failure rate. It can be inferred that when the duty cycle is high, the ARF is almost 1.5 to 2 times higher than that in low duty cycle scenarios.



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.



3.2.4.2 Fast Thin Reconstruction, Reducing Dual-Disk Failure Probability

In the recent years of disk development, disk capacity growth outpaces performance improvement. Nowadays, 15 TB large-capacity SSDs are commonly seen in enterprise market. In the near future, 32 TB SSDs will be used in storage systems.

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.

3.3 In-Depth All Flash Storage Optimization

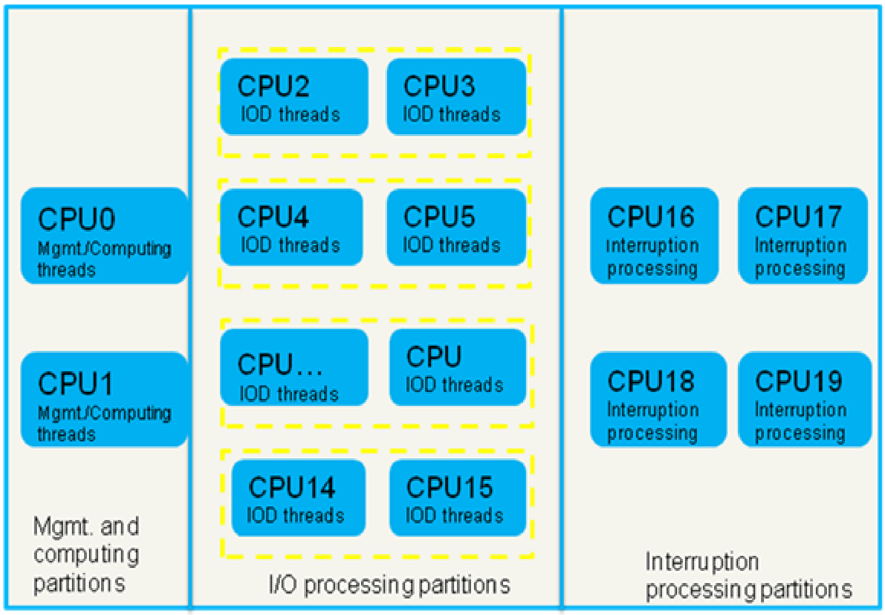
3.3.1 Optimizing the I/O Process to Reduce the Processing Latency of Storage Arrays

The HDD-based I/O process mainly considers I/O combination and sequence and fewer HDD head movements. As SSDs do not involve head movements, the key consideration of the SSD-based I/O process is to give play to each component. High IOPS performance of SSDs makes the processing capability of CPUs on storage arrays the performance bottleneck. OceanStor F V5 all flash storage systems adopts multiple technologies to improve the CPU efficiency:

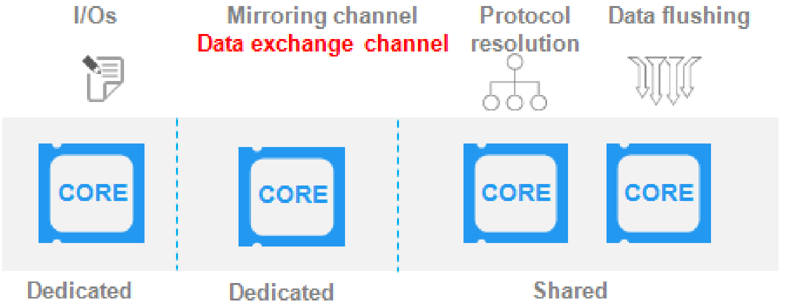
1. Intelligent CPU partition algorithm

   The I/O processing latency in storage arrays is determined by many factors. The efficient scheduling framework ensures that high-speed media I/Os are preferentially processed and prevents lock conflicts, CPU core interference, and hardware interruptions during I/O processing. As a result, the latency for processing each I/O is fixed and the system provides a stable I/O processing latency. The intelligent CPU partition algorithm of OceanStor F V5 all flash storage systems solves these problems as follows:

- Partitions CPU computing resources based on service requirements to isolate the interference of the operating system, control plane, computing plane, and hardware interruptions with service I/Os.


CPU2 IOD threads
CPU3 IOD threads
CPU0 Mgmt./Computing threads
CPU4 IOD threads
CPU5 IOD threads
CPU16 Interruption processing
CPU17 Interruption processing
CPU1 Mgmt./Computing threads
CPU… IOD threads
CPU IOD threads
CPU18 Interruption processing
CPU19 Interruption processing
CPU14 IOD threads
CPU15 IOD threads
Mgmt. and computing partitions
I/O processing partitions
Interruption processing partitions

- Groups CPUs in I/O processing partitions based on the characteristics of service I/Os and schedules I/Os among CPU groups to achieve lock-free I/O processing in groups.

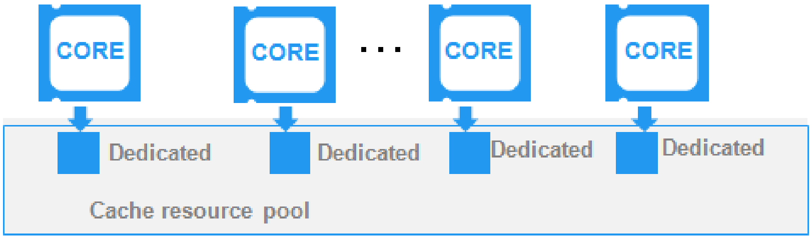


Intelligent CPU partitioning

Dedicated cores, providing sufficient resources for critical services and reducing latency

Shared cores, contributing to load balancing of services

- Cache resources are reserved based on the number of CPU cores to ensure that each CPU core has its own memory resource pool. A common memory resource pool is provided to ensure that memory resources can be applied if the memory resources of a CPU core are exhausted.

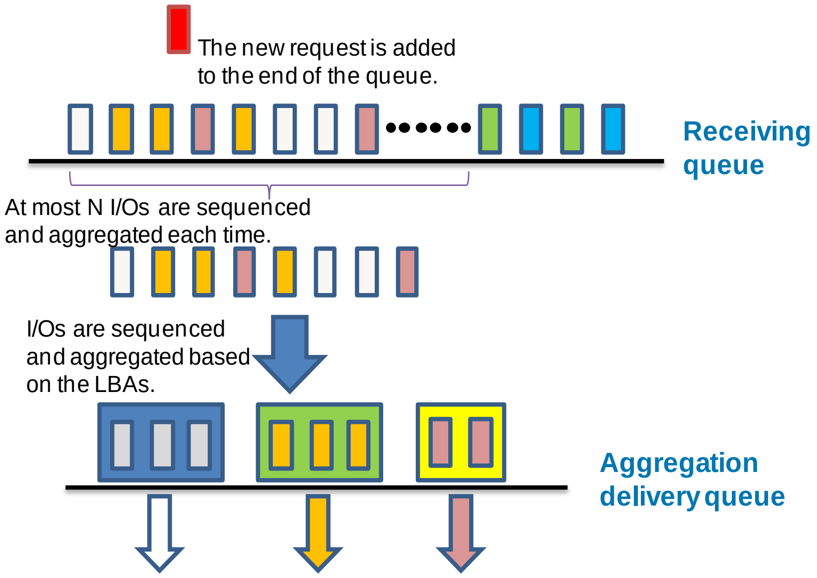


Core reservation

Cache resources are reserved based on the number of CPU cores, reducing the conflicts in multi-CPU environments and improving memory application efficiency by 50%.

2. Dynamic image I/O aggregation algorithm

   The dynamic image I/O aggregation technology of OceanStor F V5 all flash storage systems ensures a stable image latency, reduces the CPU consumption during the mirroring process, and improves random write IOPS. The mirroring process uses two queues: receiving queue and aggregation delivery queue. During aggregation, only image I/Os that first arrive at the receiving queue are sequenced and aggregated, reducing the mirroring latency and improving mirror channel utilization.

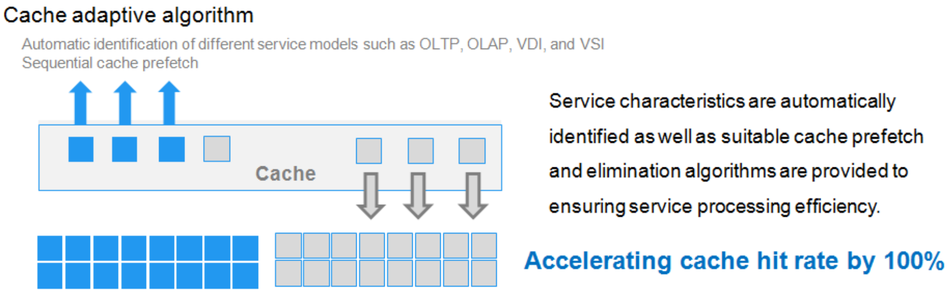

The new request is added to the end of the queue.
Receiving queue
At most N I/Os are sequenced and aggregated each time.
I/Os are sequenced and aggregated based on the LBAs.
Aggregation delivery queue

3. Cache adaptive algorithm

   Storage systems face various service models. For example, the OLTP and OLAP service models are totally different. Ensuring the excellent performance of storage systems in different service models becomes a problem. With the optimized cache prefetch algorithm and elimination mechanism, Huawei OceanStor F V5 all flash storage systems provide prefetch and elimination mechanisms that match the front-end services, ensuring the continuously high hit rate. The principles are as follows:

   - Service models of the front-end I/Os are identified. Group-based I/O management is implemented. A prefetch policy is provided for each group of I/Os, achieving accurate I/O prefetch and improving I/O hit rate.

- The access frequency is introduced in the elimination algorithm, helping to save hotspot data, reducing the probability that prefetch data (not accessed) is eliminated, and improving the prefetch efficiency.



Cache adaptive algorithm

Automatic identification of different service models such as OLTP, OLAP, VDI, and VSI
Sequential cache prefetch

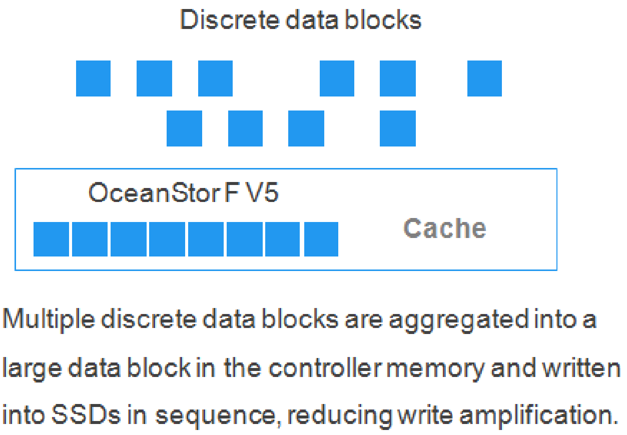
Service characteristics are automatically identified as well as suitable cache prefetch and elimination algorithms are provided to ensuring service processing efficiency.

Accelerating cache hit rate by 100%

4. I/O delivery technology

There are two phases for writing host data into Huawei OceanStor F V5 all flash storage systems. In the first phase, data is written into cache. In the second phase, data is written from cache into disks.

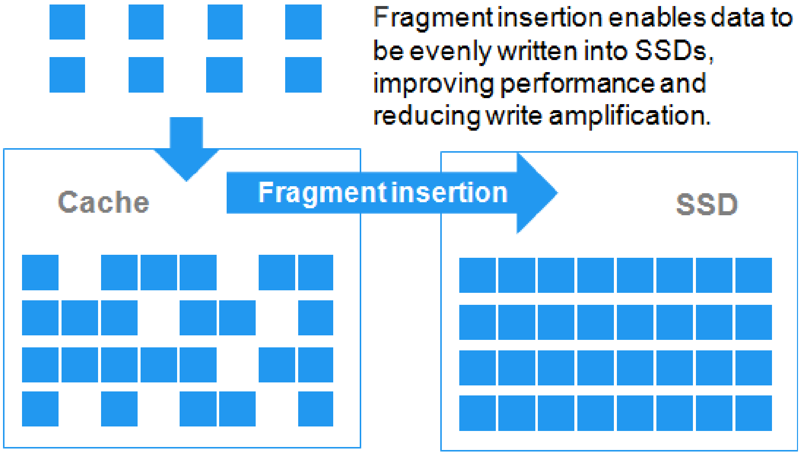
In the first phase, the storage systems use the transaction writing technology. The discrete data that is written into the cache will be aggregated into a logical large I/O (from several MB to hundreds of MB) logically called a transaction. Subsequently, data in the transaction can be used as a logical unit for centralized data flushing, achieving the purpose of combining discrete data blocks and reducing write amplification when data is written into SSDs.



- Multiple discrete data blocks are aggregated into a large data block in the controller memory and written into SSDs in sequence, reducing write amplification.

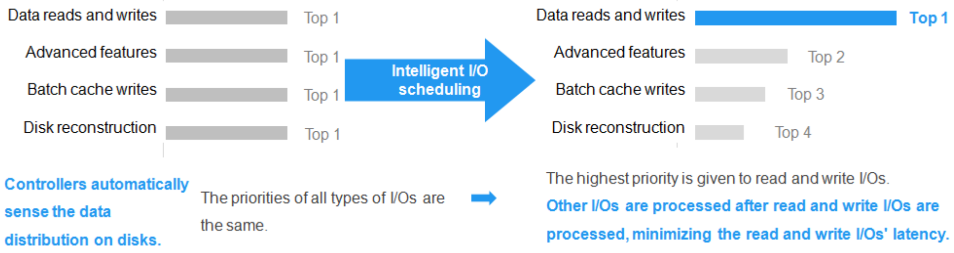
Huawei OceanStor F V5 all flash storage systems use the redirect-on-write (ROW) technology. When the transaction data is written into disks, the storage systems use new physical space to store updated data instead of overwriting the data. Therefore, the physically continuous space will be allocated to store the data when transaction data is written into disks. By using transaction-based data writing and ROW write policy, data will become continuous I/Os when it is written into disks, reducing the write amplification of SSDs.

The ROW technology enables Huawei OceanStor F V5 all flash storage systems to flexibly adjust the space allocation policy based on disk conditions. In addition to converting discrete I/Os of hosts into continuous I/Os, the ROW technology enables fragment insertion to evenly distribute data if data is unevenly distributed to SSDs.


Fragment insertion enables data to be evenly written into SSDs, improving performance and reducing write amplification.
Cache
Fragment insertion
SSD

5. Intelligent I/O scheduling

   To reduce the I/O delay of hosts, the intelligent I/O scheduling system of Huawei OceanStor F V5 all flash storage systems provides scheduling policies with different priorities for different types of I/Os. The highest priority is given to read and write I/Os, ensuring the low delay of host services. The secondary priority is given to I/Os of Smart and Hyper features, ensuring that feature performance meets the requirements. The lowest priority is given to I/Os of background tasks such as disk reconstruction, pool formatting, and space recycling, minimizing data read and write I/O latency.


Data reads and writes
Top 1
Advanced features
Top 1
Batch cache writes
Top 1
Disk reconstruction
Top 1
Intelligent I/O scheduling
Data reads and writes
Top 1
Advanced features
Top 2
Batch cache writes
Top 3
Disk reconstruction
Top 4
Controllers automatically sense the data distribution on disks.
The priorities of all types of I/Os are the same.
The highest priority is given to read and write I/Os.
Other I/Os are processed after read and write I/Os are processed, minimizing the read and write I/Os' latency.

3.4 Value-Added Service Optimization Oriented to the Flash Storage Architecture

3.4.1 SmartQoS

With the digitization of many industries, data has become critically important for the efficient operation of enterprises and public institutions, and customers require increasingly high data storage performance and stability. Many storage vendors provide high-performance storage systems. However, storage systems carry increasing types of services and face increasingly complex application scenarios. How to address the QoS requirements of different services has become a challenge for storage systems. Storage QoS, a technology that intelligently schedules and allocates various resources of a storage system to meet the QoS requirements of different services, come into being in such a background.

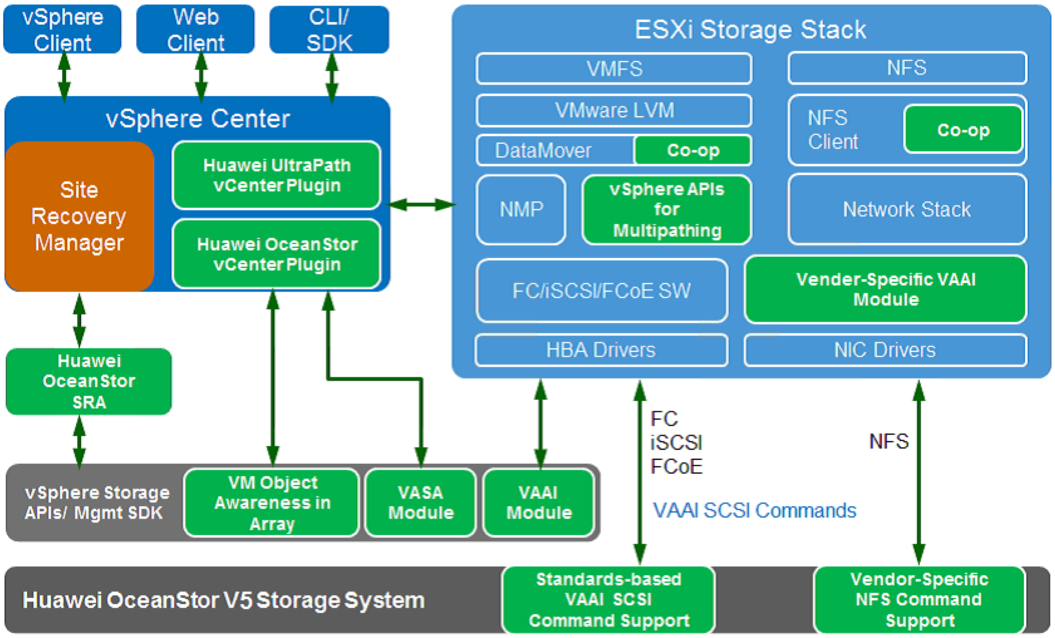
Enterprise services have various complex requirements for data response. SmartQoS ensures the QoS of data services based on the following techniques:

- **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 a 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.

3.4.2 Optimizing VMs to Improve the Efficiency of All Flash Storage

As server virtualization technology constantly develops, an increasing number of customers prefer virtual machines (VMs) to traditional servers when deploying service systems. VMs notably decrease the initial investment costs and facilitate subsequent service system integration, service data migration, and backup.

However, VMs also have their flip side: most I/O operations on a virtual machine are completed by software. This requires lots of CPU, memory, and network bandwidth resources. The widespread application of virtual machines makes this demerit increasingly pronounced. To resolve this problem, VMware proposed the idea of hardware acceleration, in which storage arrays with certified interoperability utilize a special plug-in called VMware vStorage APIs for Array Integration (VAAI) to complete operations that used to be performed by the virtual machine. By doing so, the overall system performance increases remarkably.


vSphere Client
Web Client
CLI/ SDK
vSphere Center
Site Recovery Manager
Huawei UltraPath vCenter Plugin
Huawei OceanStor vCenter Plugin
Huawei OceanStor SRA
vSphere Storage APIs/ Mgmt SDK
VM Object Awareness in Array
VASA Module
VAAI Module
ESXi Storage Stack
VMFS
NFS
VMware LVM
DataMover
Co-op
NFS Client
Co-op
NMP
vSphere APIs for Multipathing
Network Stack
FC/iSCSI/FCoE SW
Vender-Specific VAAI Module
HBA Drivers
NIC Drivers
FC
iSCSI
FCoE
VAAI SCSI Commands
NFS
Huawei OceanStor V5 Storage System
Standards-based VAAI SCSI Command Support
Vendor-Specific NFS Command Support

The integration of the storage array and VMware delivers the following functions:

- Block zeroing

  The most common operation on a virtual machine is virtual disk zeroing. This operation consumes abundant shared resources on the virtual machine such as CPU and direct memory access (DMA) resources. By interworking with VMware, the OceanStor F V5 all flash storage systems takes over block zeroing from virtual machines. With powerful CPUs, the T series can swiftly zero out blocks. This interworking function minimizes the I/Os between the OceanStor F V5 all flash storage systems and the ESX Server by over 10 times. In addition, it quickens virtual disk initialization, uplifting the overall performance of service systems.

- Full copy

  When a virtual machine migrates and clones virtual disks, it copies large amounts of file blocks. If the file block size reaches several GB, the replication may take hours to complete. The lengthy process consumes abundant server resources and occupies bandwidth for a long time. As a result, the overall system performance deteriorates. By interworking with VMware, the T series takes over block replication operations from virtual machines. With replication being optimized by the hardware system, the replication operations that used to take hours are completed in seconds. Moreover, the interworking mode reduces ESX Server CPU pressure, enabling virtual machine resources to focus on software services of application servers. This function minimizes I/Os between the T series storage system and the ESX Server by over 10 times. In addition, it speeds up operations such as Storage vMotion, and simplifies VM deployment.

- Hardware assisted locking

  To ensure data consistency on virtual machines in a cluster, locking mechanisms are implemented to properly allocate resources in concurrent accesses. A traditional way is to lock an entire LUN when it is accessed by an ESX Server so that other ESX Servers are stopped from writing I/Os to the LUN. This greatly compromises write performance. In addition, a series of commands need to be executed to obtain and release the lock when the LUN is being locked. This increases I/O latency. By interworking with VMware, the T series locks blocks in a LUN rather than complete LUNs when multiple VMs are writing I/Os to the LUN. This ensures faster concurrent write, shorter write latency, and higher VM density. These improvements further boost the overall system performance.

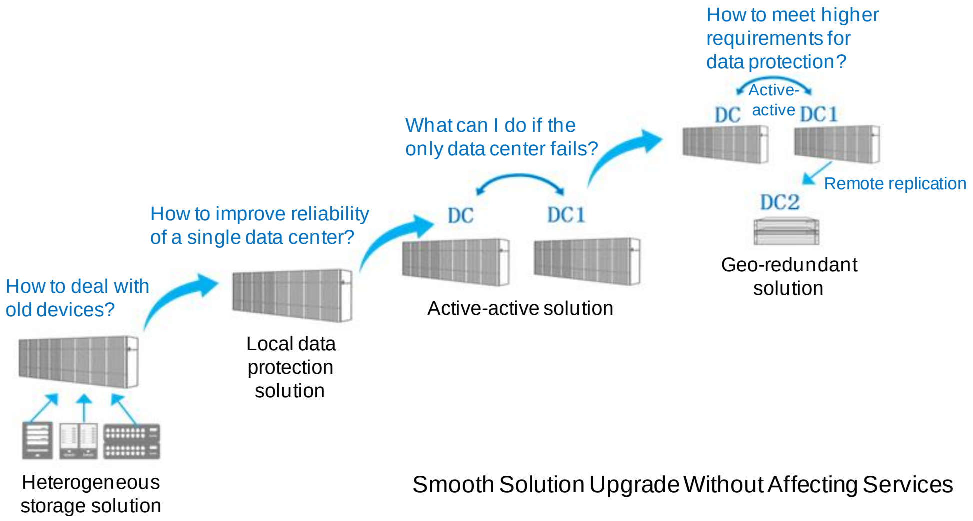
- Space reclamation of thin LUNs

4 Enterprise-Class Features Supported by All Flash Storage Systems

The mainstream flash arrays are optimized enterprise-class arrays and all-flash-memory arrays that focus on hardware architecture resolution. As all-flash-memory arrays are not widely used in the enterprise DCs, they cannot support most enterprise-class features, such as local data protection, cross-DC data protection, and business continuity guarantee.

Huawei OceanStor F V5 all flash storage systems support all enterprise-class storage features while providing high performance. Conventional storage arrays can be converted into all flash arrays with ease without changing use habits.

Huawei OceanStor F V5 all flash storage systems provide complete Hyper and Smart series data protection software to cope with disasters, viruses, or commissioning, ensuring system reliability and stable running of core services.


How to meet higher requirements for data protection?
Active-active
DC
DC1
What can I do if the only data center fails?
Remote replication
DC2
How to improve reliability of a single data center?
DC
DC1
Geo-redundant solution
How to deal with old devices?
Active-active solution
Local data protection solution
Heterogeneous storage solution
Smooth Solution Upgrade Without Affecting Services

5 Smooth Migration to All Flash Storage Systems

With the decrease of SSD prices, an increasing number of all flash storage systems are used by enterprises. IT system administrators must tackle the following challenges brought by this trend:

- How to manage heterogeneous storage systems in a unified manner
- How to migrate service data from legacy storage systems to all flash storage systems with the minimum impact on services
- How to reuse legacy storage systems to improve service reliability

5.1 Service Performance Tuning

5.1.1 LUN Migration

With the evolution of storage technologies, the need for service migration arises as a result of storage system upgrade, service performance tuning, and data center migration. In the era of fierce business competition, however, service suspension means loss of business opportunities. In the mean time, mission-critical services must be migrated without being interrupted. LUN migration is an easy-to-use data migration solution.

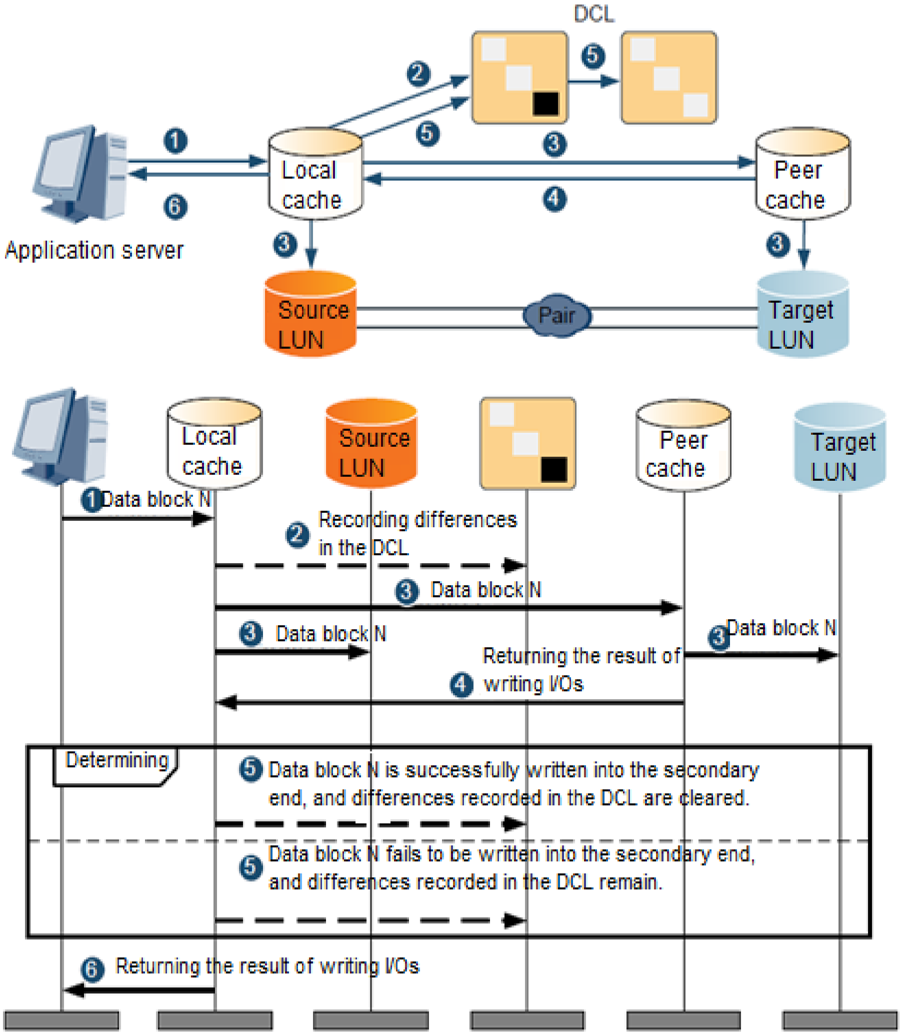
LUN migration migrates host services from a source LUN to a target LUN without interrupting these services and then enables the target LUN to take over services from the source LUN without being noticed by the hosts. After the service migration is complete, all service-related data has been replicated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

Service migration is a necessary approach to resource optimization. Users can migrate data from low-speed storage media to high-speed storage media to improve data read and write speeds, or migrate infrequently accessed data to low-speed storage media for backup to optimize the service capability of storage devices.

In addition to service migration within a storage system, LUN migration also supports service migration between a Huawei storage system and a compatible heterogeneous storage system.

The LUN migration feature provided by OceanStor F V5 all flash storage systems 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.

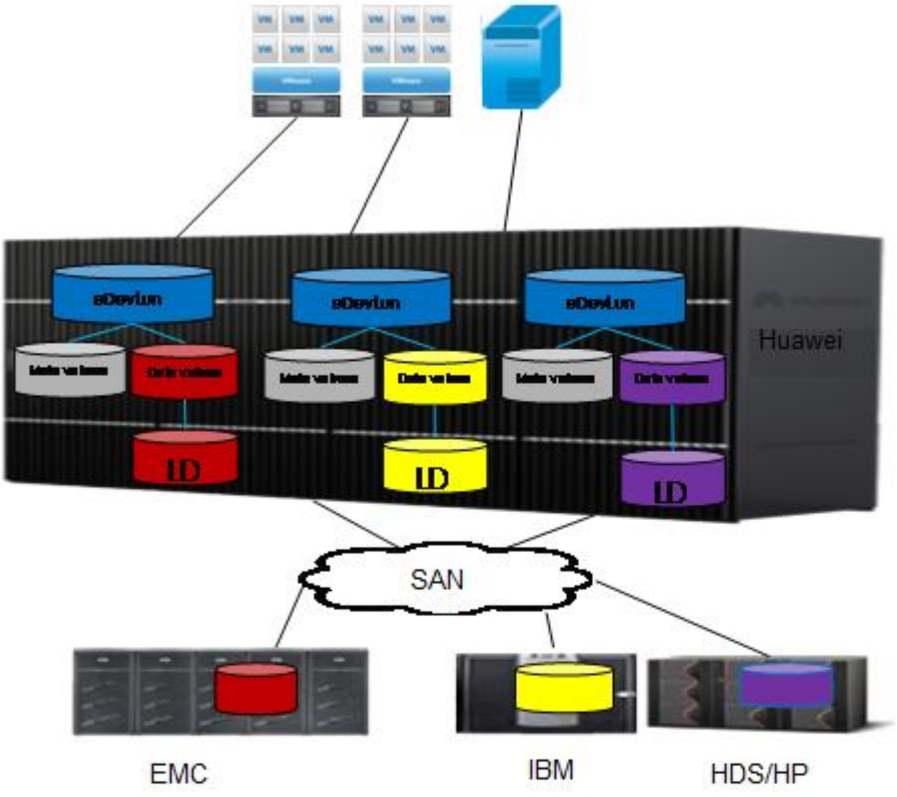

DCL
2
5
5
1
6
Application server
Local cache
3
4
Peer cache
3
3
Source LUN
Pair
Target LUN
Local cache
Source LUN
Peer cache
Target LUN
1 Data block N
2 Recording differences in the DCL
3 Data block N
3 Data block N
3 Data block N
4 Returning the result of writing I/Os
Determining
5 Data block N is successfully written into the secondary end, and differences recorded in the DCL are cleared.
5 Data block N fails to be written into the secondary end, and differences recorded in the DCL remain.
6 Returning the result of writing I/Os

5.2 Migration Between Heterogeneous Storage Systems

The SmartVirtualization feature of Huawei OceanStor F V5 all flash storage systems is committed to eliminating incompatibility issues between different storage systems and working with other value-added features to achieve unified management of heterogeneous storage systems, safe and complete application migration, and improvement of application reliability.

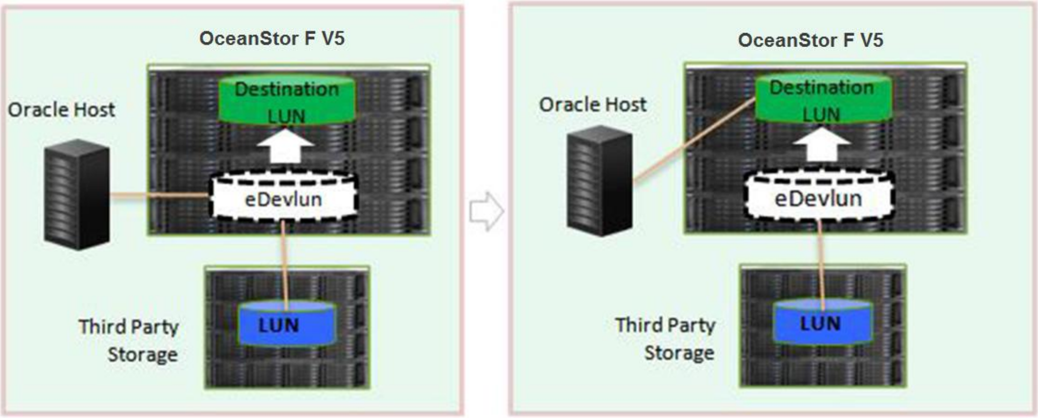
5.2.1 Heterogeneous Virtualization

As users' data centers develop, storage arrays in the data centers may come from different vendors. How to efficiently manage and apply storage arrays 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 arrays, and their workloads are reduced. In such a scenario, SmartVirtualization simplifies system management.

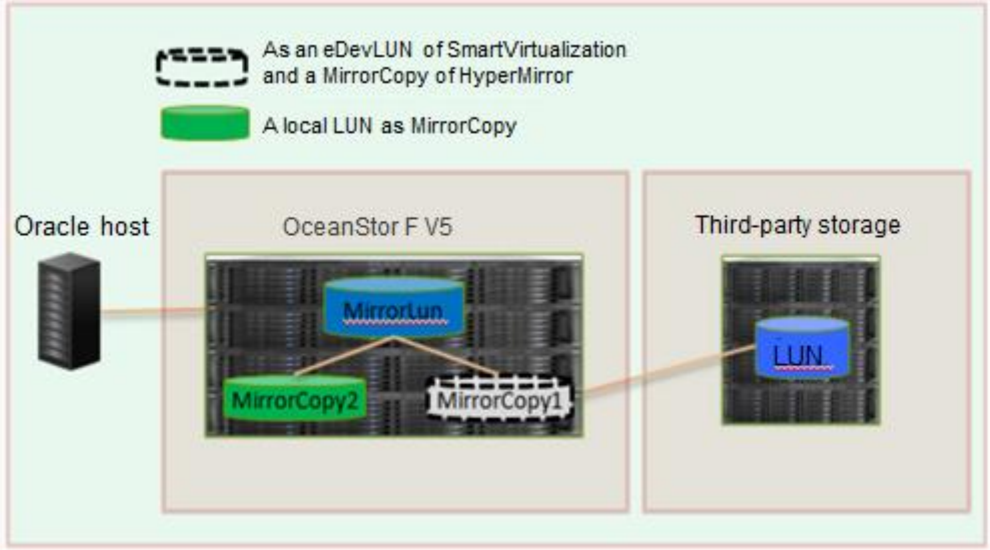


5.2.2 Migrating Data From Third-Party Storage to Huawei OceanStor F V5 All Flash Storage

Oracle databases can be migrated from third-party storage to Huawei OceanStor F V5 all flash storage systems using Huawei storage's SmartVirtualization and SmartMigration features.



SmartVirtualization and HyperMirror are leveraged to implement dual-write on two heterogeneous storage arrays, improving service reliability.


As an eDevLUN of SmartVirtualization and a MirrorCopy of HyperMirror
A local LUN as MirrorCopy
Oracle host
OceanStor F V5
MirrorLun
MirrorCopy2
MirrorCopy1
Third-party storage
LUN

6 Acronyms and Abbreviations

Figure 6-1 Acronyms and abbreviations

Acronym or Abbreviation	Full Spelling
AFR	Average Failure Rate
BIM	Back-End Interconnect I/O Module
GC	Garbage Collection
HDD	Hard Disk Drive
LUN	Logical Unit Number
LRU	Least Recently Used
MTBF	Mean Time Between Failures
OLTP	Online Transaction Processing
RAID	Redundant Array of Independent Disks
ROI	Return on Investment
SSD	Solid State Disk
TBW	Total Bytes Written
TPM	Transactions Per Minute
UBER	Uncorrectable Bit Error Rate