04 Architecture of Storage Systems

www.huawei.com

Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Foreword

- This module mainly introduces the common system architecture of storage systems, the components of storage systems and also introduces the Huawei storage products.

Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

Objectives

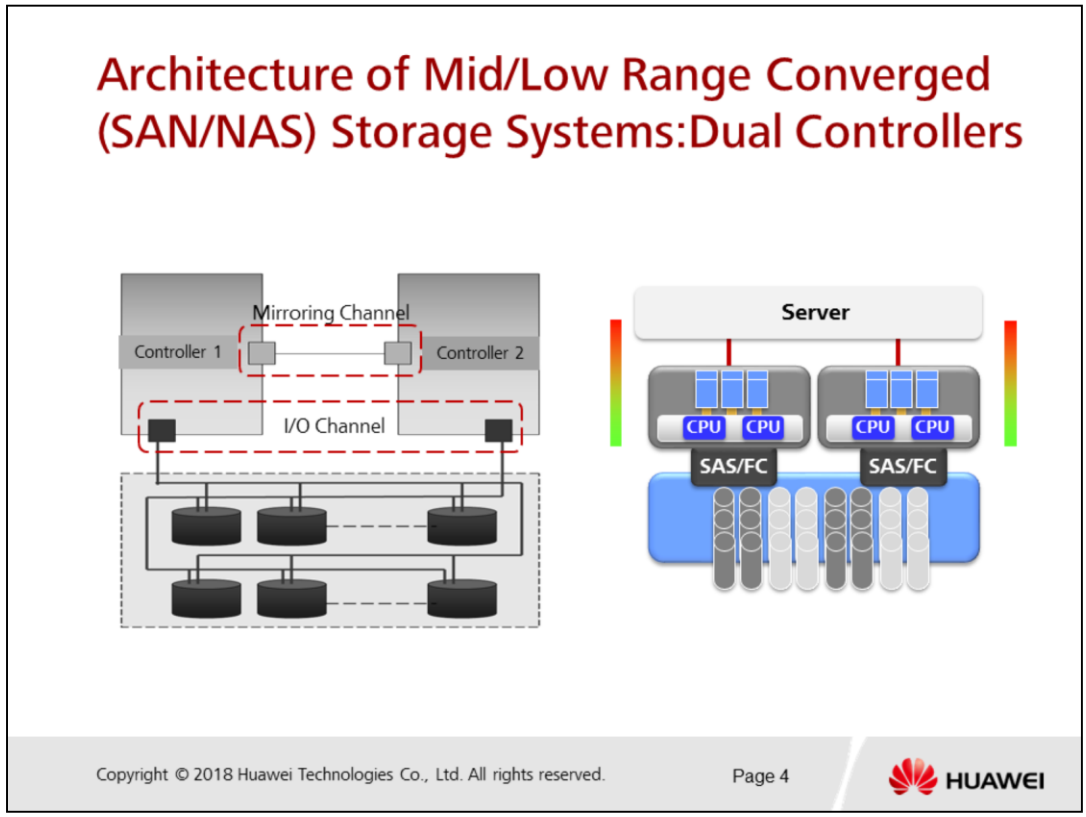
- Upon completing this module, you will be able to:
  - Understand the common storage system architectures.
  - Grasp the concepts of common storage components.
  - Understand the Huawei storage products.

Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

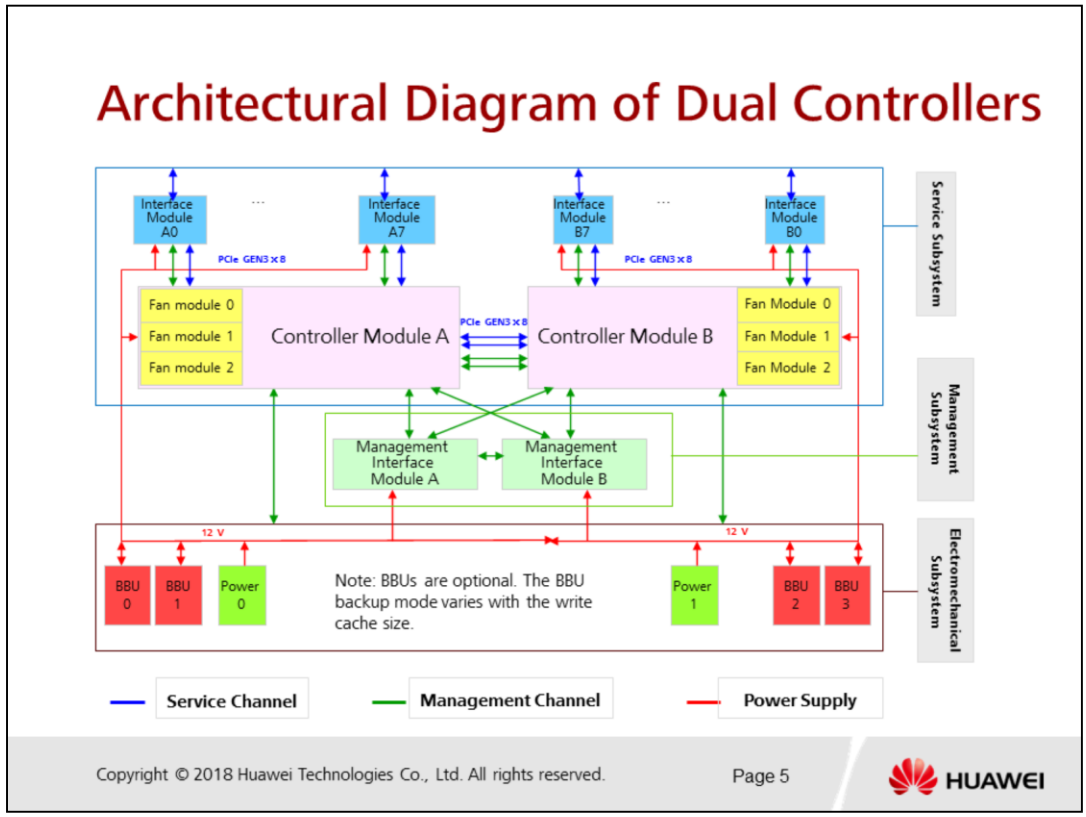
Contents

1. **Architecture of Storage Systems.**
2. Components of Storage Systems.
3. Introduction to Huawei Storage Products.

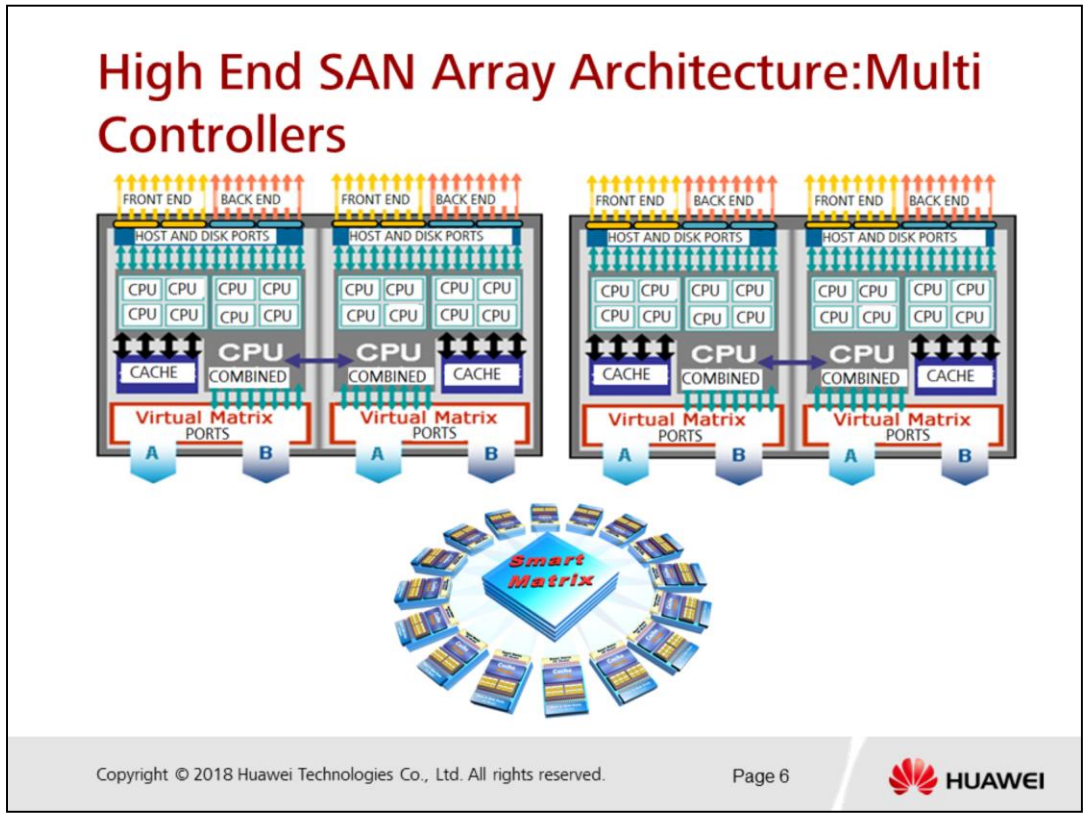
Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.



- Key features and benefits of Dual Storage controller system architecture:
  - A copy of all the data written during I/O process is kept in both storage controller's cache to avoid data loss when there is a single controller failure.
  - Each of the storage controller is connected via backend ports to the loop of physical disk enclosures for load balancing of the service workloads and increase the overall system performance.
  - Single or Multiple LUNs I/O service load is load balanced between the two storage controllers to avoid performance bottleneck.
  - Whenever a single controller is faulty, another storage controller can takeover the service workload with the help of the Multipathing software to ensure service continuity.



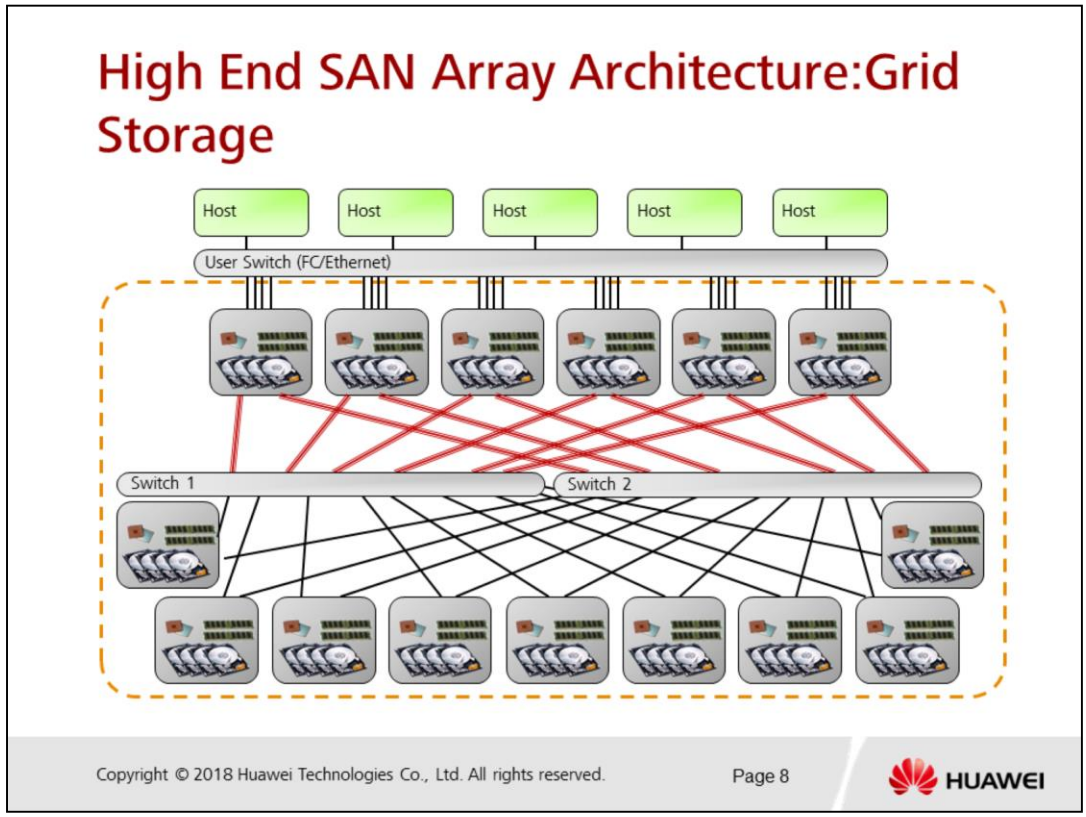
- The diagram above shows the interconnection between the storage controllers and other hardware components of the storage systems such as the power and fan modules, management interface modules and the interface modules.
- The diagram also shows the redundant design of the storage controllers and management interface modules. Hence, by having Dual controllers, the single point of failure faced by single controller storage architecture are eliminated.



Key features of High End Storage Systems:

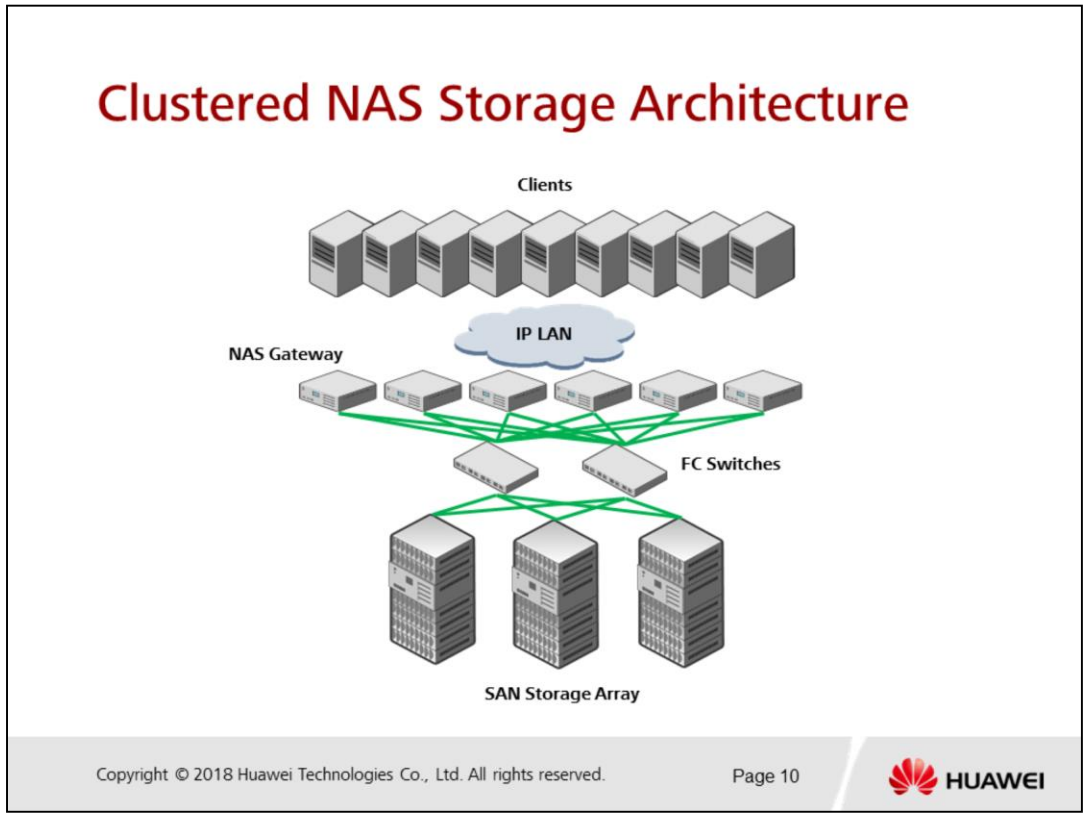
- Data plane is interconnected using PCIe connections, and the 4 controllers within the storage engine are interconnected using PCIe backplane which allows high bandwidth rate up to 128G/s.
- Storage engines uses full duplex switched network with PCIe optical cables for inter-engine connection up to 128G/s bandwidth rate.
- The data transmission and mirroring uses the internal backplane interconnection channel as prioritized transmission route and reverts to use inter-engine switched network channel if unable to transmit the data between the storage controllers using the backplane.
- The total bandwidth for data transmission (including data transmission within same storage engine and data transmission between different storage engines) is up to 640G/s.
- Management and control plane uses Gigabit Ethernet(GE) technology and the 2 ports of the SVP(Service Processor) effectively creates 2 management planes for redundancy, and all the storage engines and switches can be connected to both of the management planes at the same time. When the GE network is faulty, the management control plane still can use the PCIe channel in the data plane to transmit management and control messages. This ensures that service continuity and management of storage system is not affected during the faulty period of the GE network of the management plane.

- All 4 of the storage controllers within the storage engine is connected to the back end SAS interface cards. Whenever there is a faulty storage controller, it does not affect the connection of the remaining storage controllers to the storage hard disks.
- It supports persistent caching technologies which means that whenever a storage controller is faulty, the storage system will redistribute the cached I/O and Mirroring process to the other storage controllers to complete, ensuring that each cached I/O operation is completed with no data loss and with no service disruptions.
- The service processor (SVP), working with the keyboard, video, and mouse (KVM), is the core component for managing, configuring, and maintaining the OceanStor 18000 V3 series mission critical storage system. Maintenance and management tools are installed on the SVP and used for local or remote monitoring, management, configuration, and authentication.

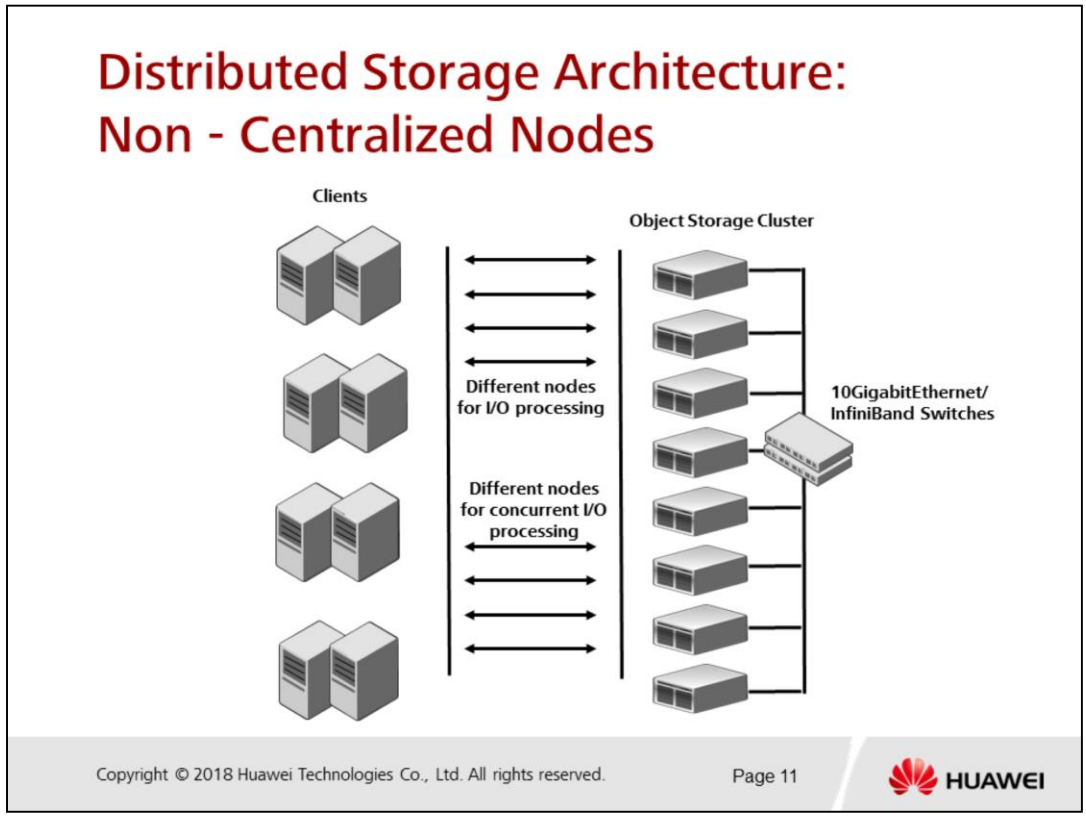


- Grid storage is considered as the new generation storage architecture by the industry, it employs the combination of latest large scale parallel distributed grid storage technologies with the Scale-Out(Horizontal Expansion) storage architecture. This architecture allows it to use a multipath grid network between the storage nodes for load balancing and fine-grained data distribution algorithm to ensure that the data is evenly distributed between the storage nodes. This increases the overall system reliability, usability and IO performance, and makes it easily scalable. Once a new storage node is added, the rest of the nodes in the grid will be able to detect the new node and use it for storage making scalability much easier to perform.
- In grid storage, there is no centralized control module as the storage system consists of grids with self-contained storage nodes. These self-contained storage nodes are interconnected in the storage grid and able to communicate to each other without a centralized control module, introducing a new level of fault tolerance and redundancy into storage system architecture. Each grid is also known as data modules, and each independent data modules has their own loosely coupled CPU processing capabilities, caching and disk storage capabilities.

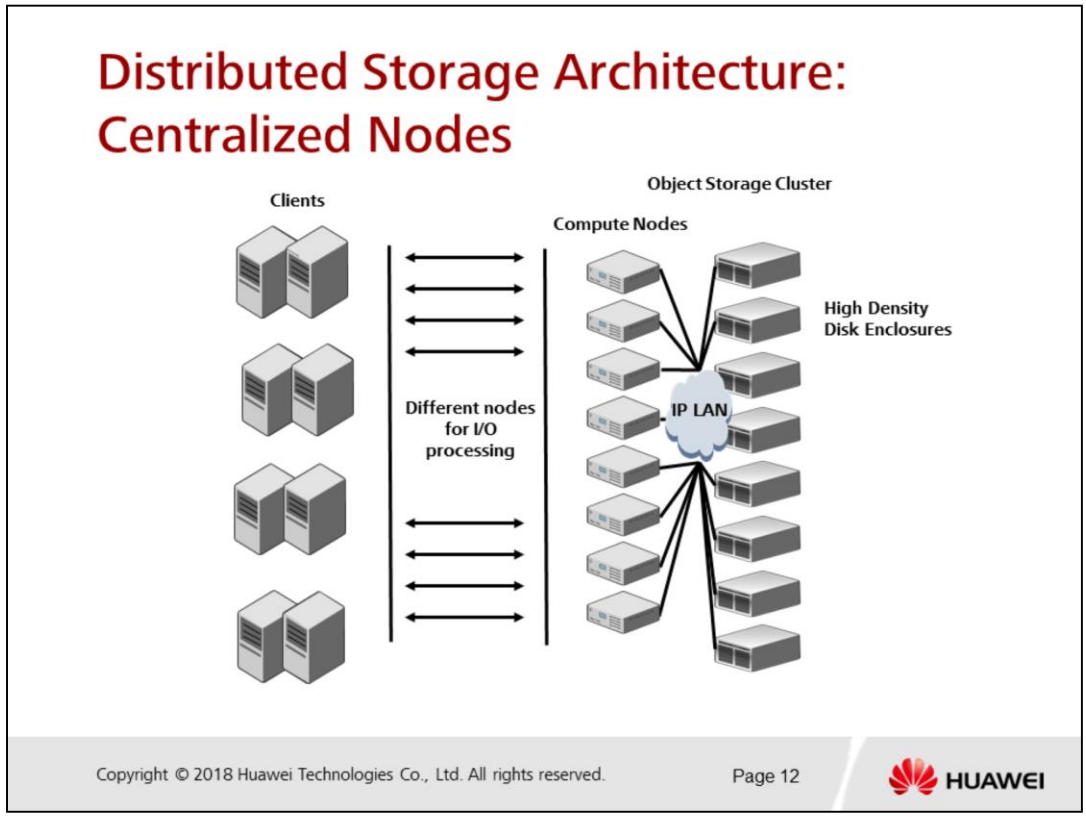
- The main component of the grid storage system is called modules. Each module provides the processing unit, caching, and host ports which are based on standard Intel and Linux systems. All the modules of the grid storage systems forms a grid based networked working entity by utilizing the internal redundant Ethernet switched network that exists between the storage nodes. Hence, this storage system is capable of parallel processing and has high computing power which is suitable for many integrated and hybrid computing scenarios.
- The interface modules in the grid storage system has similar structure with the data modules, however both of them have some differences in functionality. The difference between them is that interface modules not only provides the same functionality as data modules, they also provide support for connectivity such as FC ports and iSCSI ports for host connection, and support for remote mirroring. In concept, a module is the basic element of grid storage and it provides physical storage space, processing and caching capabilities. The relationship between these modules are equal with no priorities or order of importance.
- The grid storage system uses a "pseudo random"(Not fully random) distribution algorithm, which means that by using a unique calculation algorithm it is able to strategically distribute the specific workloads across multiple modules. This allows the overall performance of the grid storage to increase with the addition of more modules. The more storage module is added, higher performance and computing power can be attained.
- Grid storage system contains a redundant Ethernet switching architecture that act as the module for data transmission that includes the transmission of data between two interface modules, between interface and storage modules and also data transmission between two storage modules.



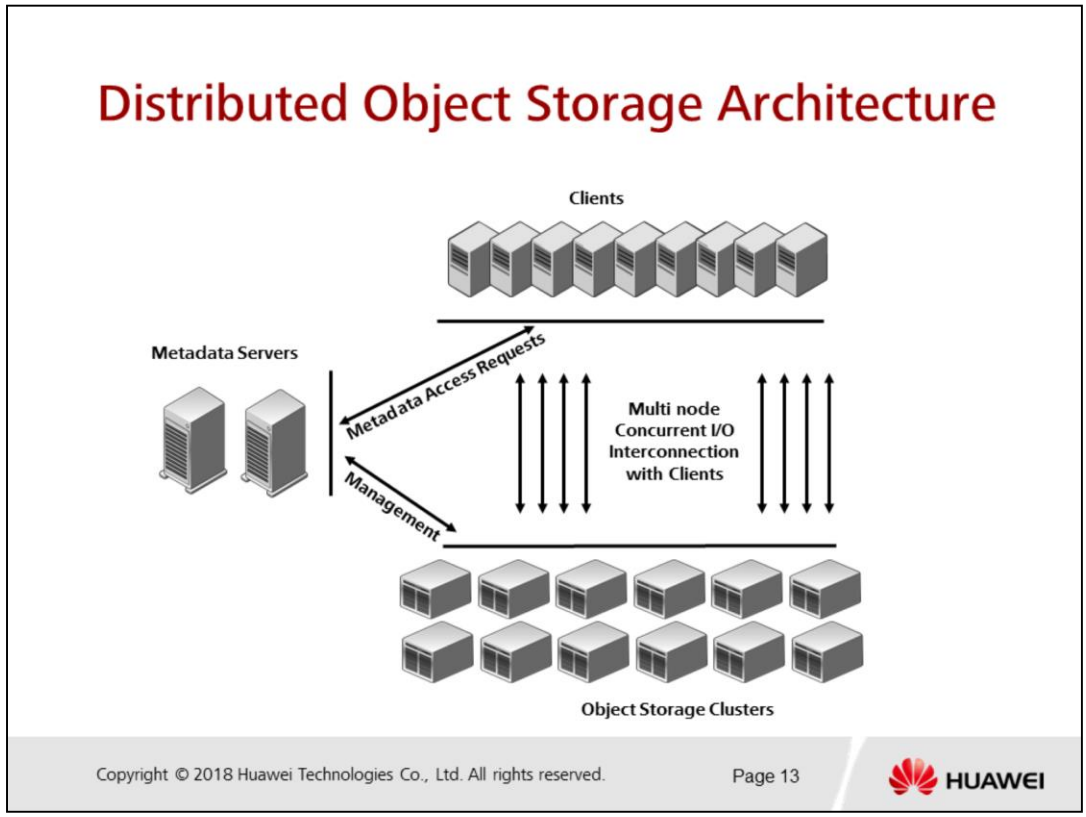
- Unified Global Namespace and Global Sharing.
  - A global namespace provides a consolidated view of multiple file systems or file servers in different location allowing data to be accessed without knowing where the data physically resides which is particularly beneficial for distributed implementation of storage. Without global namespace, each file system needs to be managed separately.
  - Global sharing allows the data to be shared globally and accessed from multiple locations making better data accessibility for the storage system.
- High Performance and Concurrent Multi Host IO Processing Array.
  - Clustered NAS allows IO process of multiple host to be processed in concurrent which in turn increases the overall performance of the storage systems.
- Online Expansion.
  - Expansion of the storage space and modules can be done without taking the whole storage system offline which eliminates downtime during storage expansions.



- Supports up to 10 billion files for storage.
- Global namespace and global sharing.
- High performance with concurrent I/O access to multiple nodes in the target storage cluster up to several GBs of combined I/O rate.
- Online expansion.
- Automatic load balancing.
- Low cost as it supports x86 servers.
- This distributed storage architecture has no need for centralized nodes for I/O processing as the I/O operation is sent directly from the client to the object storage cluster either concurrently or non-concurrently via different nodes.



- Supports Exabyte (1000 Petabytes) level file storage
- Global namespace and global sharing.
- Online expansion.
- Low cost as it supports x86 servers.
- This distributed storage architecture has a layer of centralized compute nodes that will process the I/O operation from the clients and directs them to the targeted storage within the high density disk enclosures.

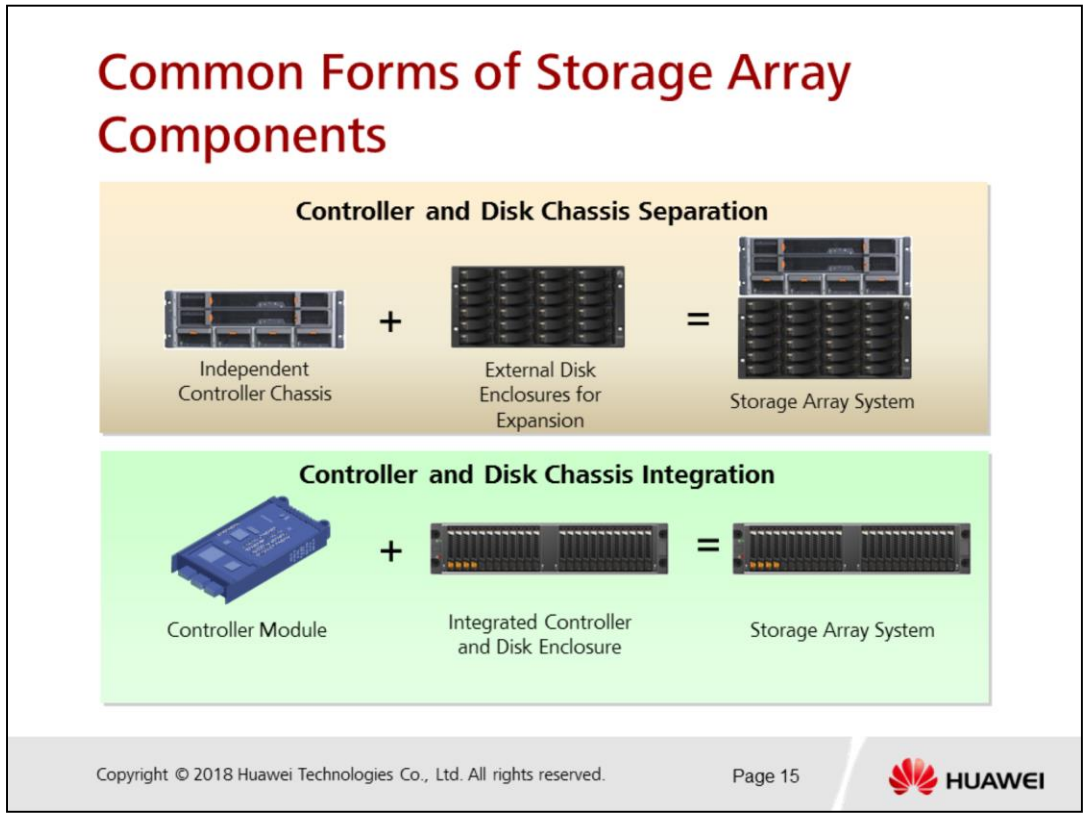


- Supports up to ten billion files for storage.
- Global namespace and global sharing.
- High performance with concurrent I/O access to multiple nodes in the target storage cluster up to several GBs of combined I/O rate.
- Online expansion.
- Low cost as it supports x86 servers.
- Object storage is a computer data storage architecture that manages data as objects, as opposed to other storage architectures like file systems which manage data as a file hierarchy, and block storage which manages data as blocks within sectors and tracks. It uses metadata for each object for indexing, data management and optimizing storage.
- In this distributed object storage architecture, we can see from the diagram above that there are metadata servers that manages the access request and manages the data while the object storage clusters has a multi node concurrent I/O interconnection with clients. In a standard I/O operation, the client can read/write the data from the object storage clusters with the help of metadata servers that pinpoints the location of the targeted data by using metadata information.

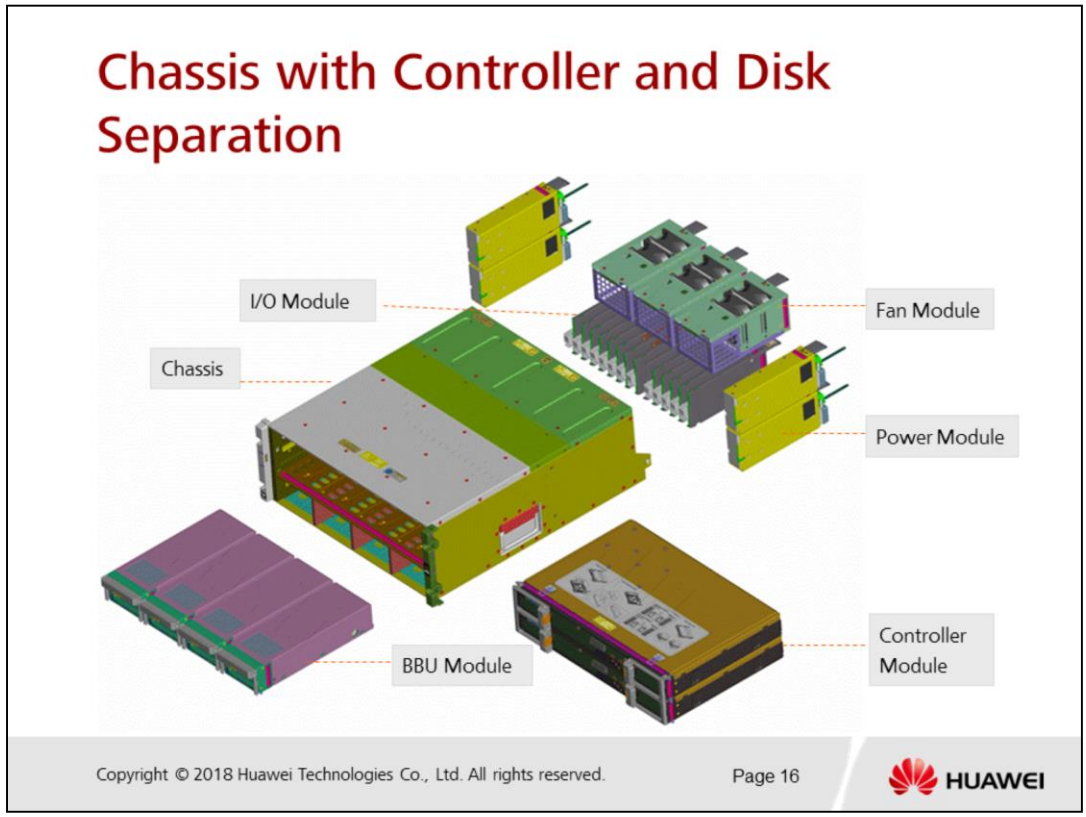
Contents

1. Architecture of Storage Systems.
2. **Components of Storage Systems.**
3. Introduction to Huawei Storage Products.

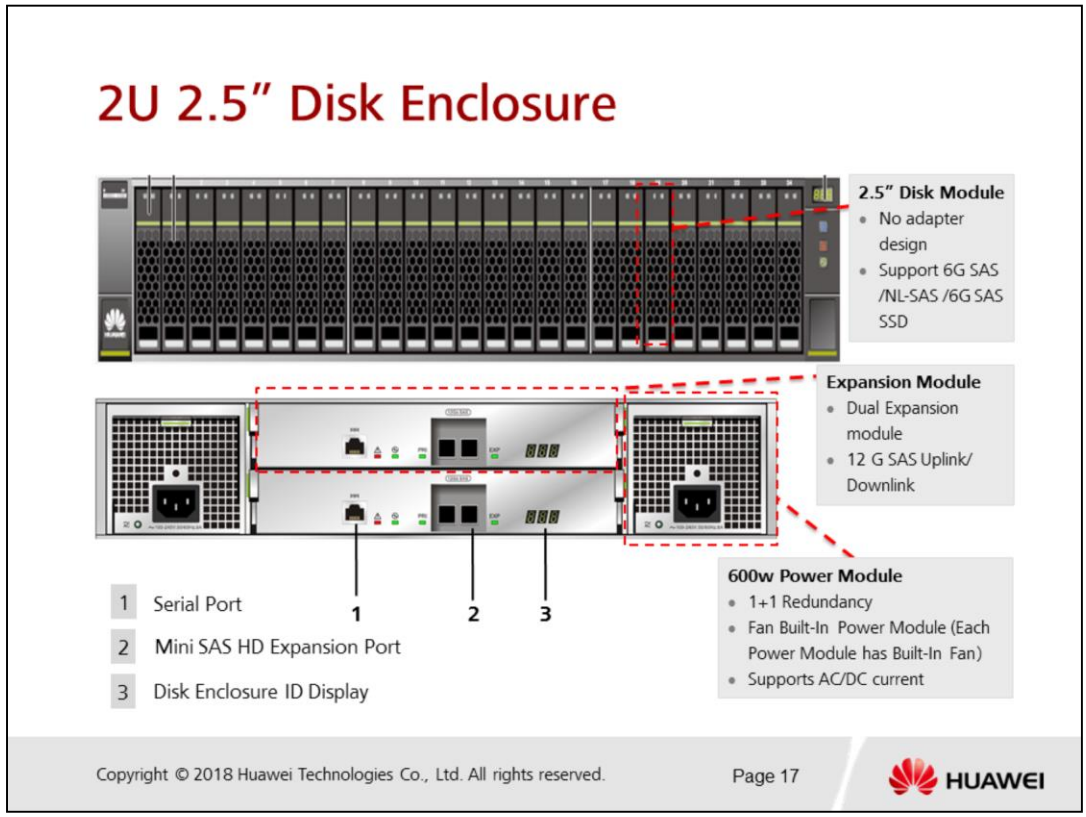
Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.



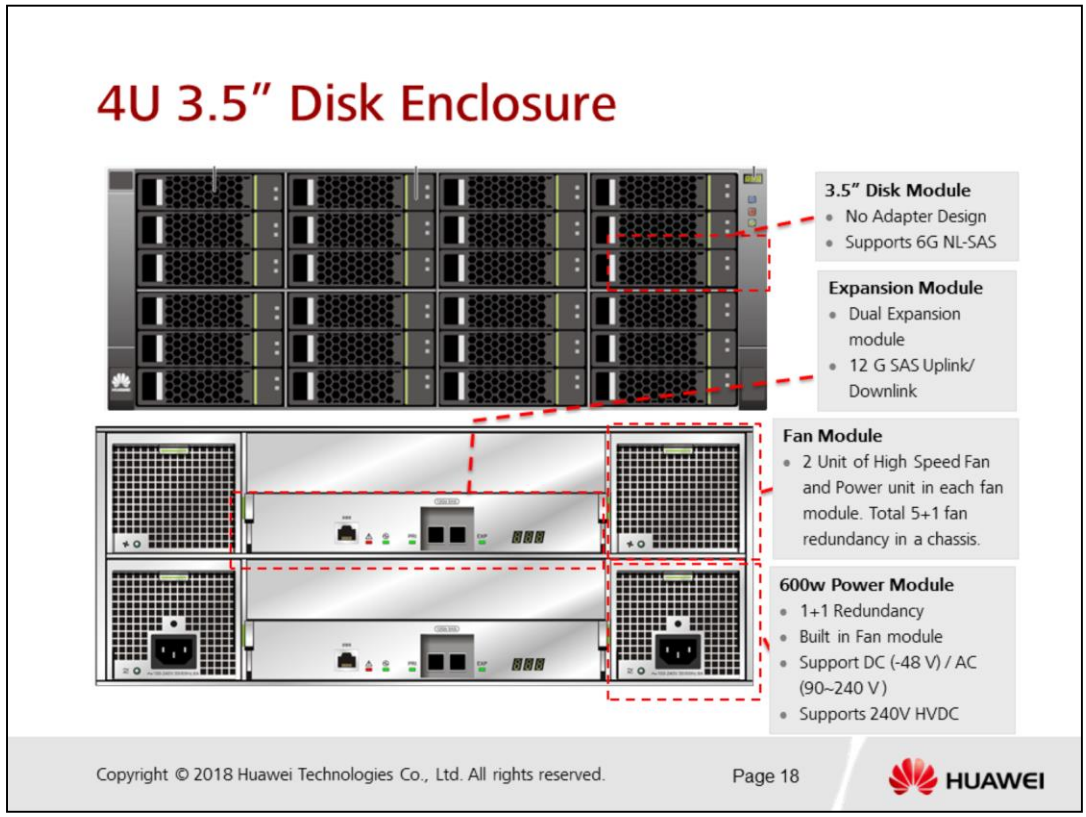
- The storage controller is the "brain" of the storage array, and it is the main component for processing and caching. It mainly manages the simple I/O operation and RAID management in the early days, but as technology progresses, it is now able to provide different kinds of advanced data management features such as snapshot, mirroring and cloning.



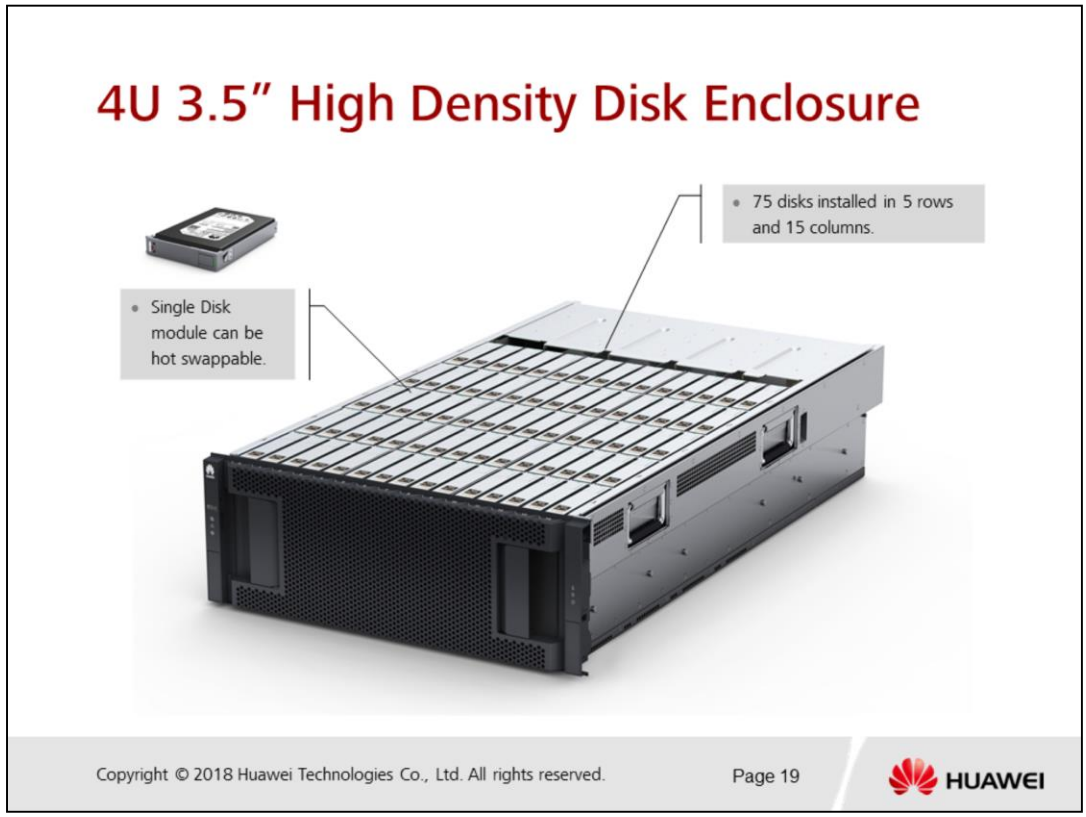
- The diagram above shows the different components that makes up the controller and disk separated chassis. The main components consist of the chassis, IO module, fan and power module and the controller and BBU (Backup Battery Unit) module.
- This chassis is connected to the external high density disk enclosures via the I/O module.
- BBU unit is the backup battery that provides power to the storage controller to save all the I/O cache into the coffer disks in the event of a power failure. Once the external power is restored, the cached data in the coffer disk will be flushed to the controller's cache and resume the I/O process to the physical disks in the external disk enclosures ensuring no data loss in the event of power interruptions.



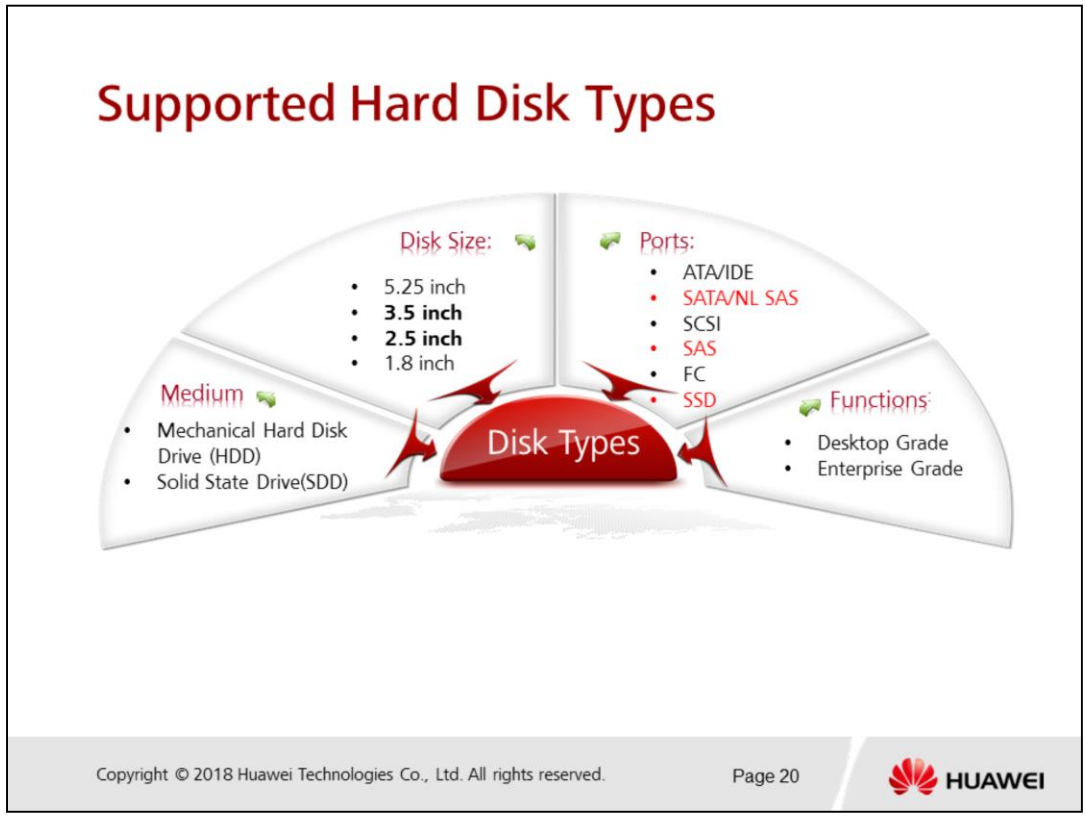
- The controller enclosure adopts a modular design and consists of a system sub rack(chassis), controllers, power-BBU modules, and disk modules.
- The disk module employs a no adapter design where the 2.5 inch hard disk are connected to the storage chassis directly via a Midplane without the need of an adapter board installed on each disks.
- This unit has a standard size of 2U and has a 1+1 redundant design for the power module and the controller module. The model shown above supports 25units of 2.5inch HDD in a single chassis.
- The controller module also acts as the expansion module to cascade multiple storage chassis together. The maximum number of storage controller in cascading mode differs if using either direct connection or switched network connection.
    - Cascading with Direct Connection between controllers(No switches) : Maximum number of cascaded controllers is 4.
    - Cascading in a switched network : Maximum number of cascaded controllers is 8.
    - The Mini SAS HD expansion port is used for expansion and cascading multiple storage chassis



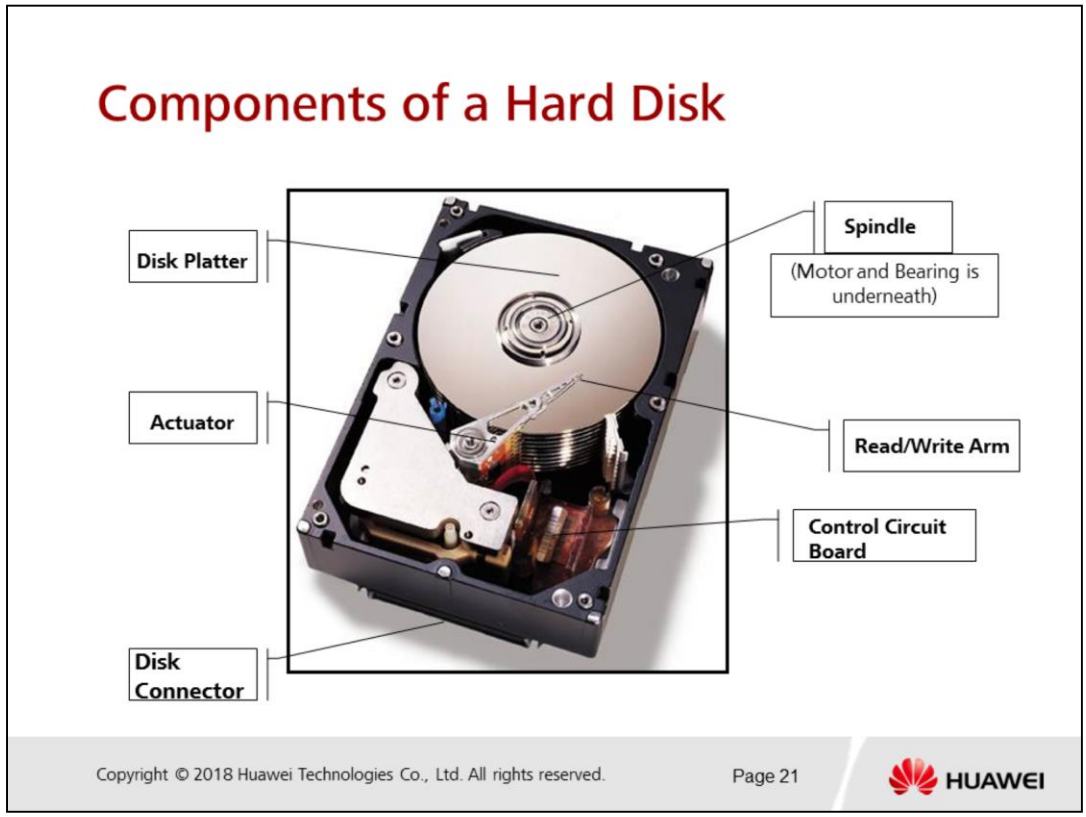
- The disk slots of a 4 U SAS disk enclosure are numbered 0 to 23 from left to right and from top to bottom.
- The expansion module has 2 Mini SAS HD expansion port (PRI/EXP) for cascading multiple chassis. There is also a Serial port for management console connection along with the disk enclosure ID display to show the ID for showing the ID of the storage chassis.
- The 600w power module supports both direct current (DC), alternating current (AC) and high voltage direct current (HVDC).



- The single disk module can be hot swappable without turning off the whole disk enclosure which minimizes the downtime during maintenance. However system performance might be impacted during disk replacement so it is suggested to do disk replacement during off-peak hours.
- The disk number of a high-density disk enclosure displayed on DeviceManager or CLI ranges from 0 to 74. These disks are numbered from left to right (15 columns) and from bottom to top (five rows).

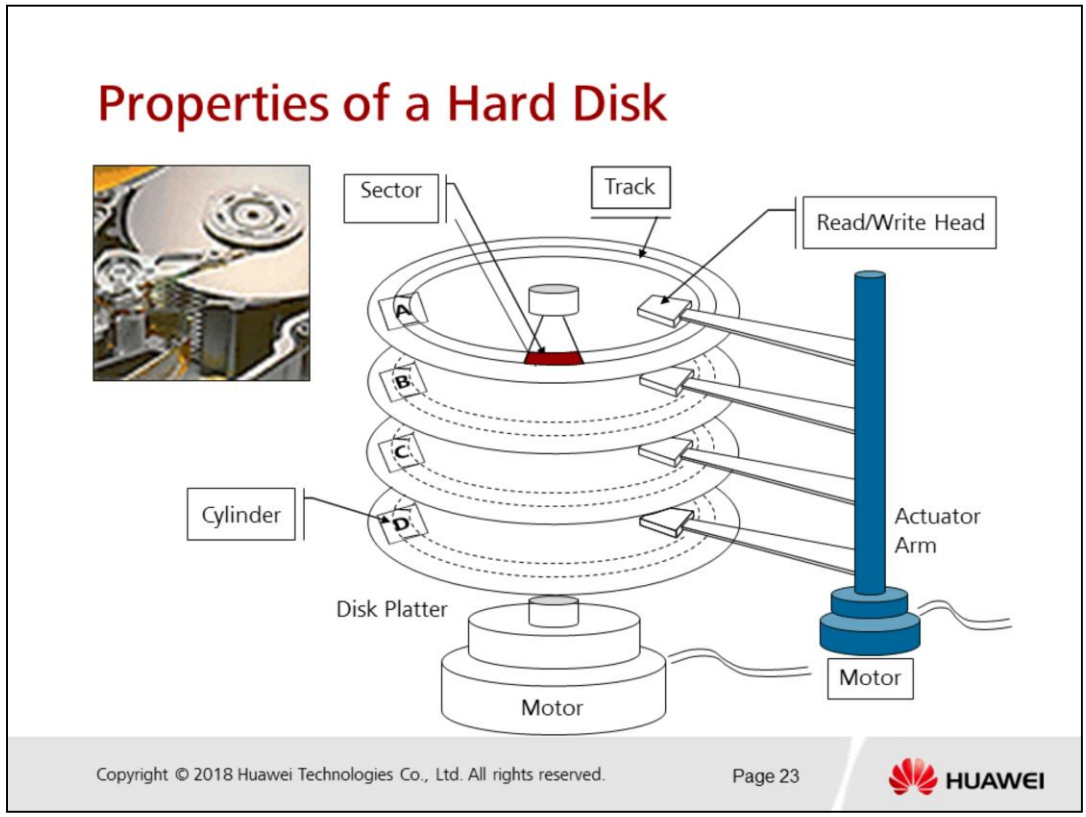


- The diagram above shows the different disk types that are supported by Huawei Storage Systems.
- Desktop grade hard disks are mainly for personal and home use and mostly used in desktop PC and laptops.
- Enterprise grade hard disks are mainly for enterprise level usage and mostly used in servers, storage arrays, and workstations.
- The difference between desktop and enterprise grade hard disks are as follows:
  - Capacity : Compared to desktop grade hard disks, enterprise grade hard disk have more storage capacity. Currently, single enterprise grade hard disk can have up to 4TB storage space or more.
  - Performance : Enterprise grade hard disk has higher performance mainly in rotation speed, caching and average seek time.
  - Reliability : Enterprise grade hard disk has higher mean time between failures (MTBF), commonly the desktop grade hard disk's MTBF is between 500, 000 hours whereas enterprise grade hard disks are above 1 million hours which makes enterprise grade hard disk much more reliable.



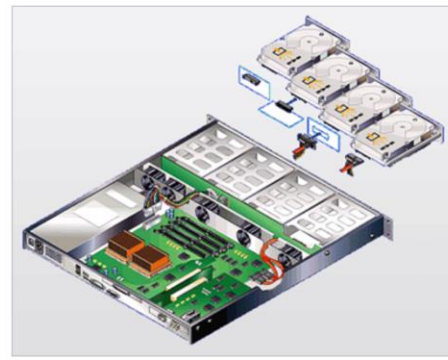
- All the mechanical hard disks has the same working principles where the platters are covered with magnetic materials and the small sub micrometer region on the platter is polarized to represent a message or bit in binary form.
- The Read/Write Arm : Handles the data reading and writing :
  - Actuator: Moves the Read/Write arm to the specified locations on the platter.
  - Platter: Preserves or saves the written data.
  - Spindle: Rotates the disk platter and moves the disc's specified location to under the Read/Write arm.
  - Control Circuit Board: Controls the disk speed, movement of the actuator and send commands to the disk head.
- A hard disk uses a high speed moving read/write arm to read data from the magnetic particles on a flat disk platter. Data from the disk is read and transmitted to the CPU through the read head. The hard disk drive (HDD) consist of few platters, read/write heads, control circuit board, spindle and other mechanical components. The magnetic region on the disk allows unlimited time of data recording and data erasure but it is subject to the wear and tear of the disk drive. In this module, we will be introducing in detail the connectors on the hard drive, the components of the hard drive and the data saving mechanism and also discuss the factors that might affect the performance of the drives.
- A standard disk contains one or more flat round discs called disk platters. Data is recorded in the form of binary (0 & 1) onto these platters. These platters are hard and coated with magnetic materials on both sides, and data can be read/written to both sides of the platter. The number of platter and the storage volume of these platters determines the final storage capacity of the hard drive.

- As the picture shown above, the spindle connects all the platters to the motor and the motor rotates the platters in a predefined speed. The rotation speed of the platter is calculated by rotation per minute (RPM), and the common RPM of the hard disks on the market is 7200,10000 or 15000rpm. Current storage systems commonly uses a 3.5inch (90mm) diameter disk platter but hard disk of other diameter are also used on the storage system within the market. When a disk drive reaches 15000rpm, the outermost region of the disk platter can have a movement speed of up to 25% of the speed of sound. Hence, with the advancement of technologies, the rotation speed of the disk platter increases over time. However due to the limitations on the rotational speed, there is limited room for improvement over this disk platter design of hard drives. Hence, for much higher performance, newer design of hard drives emerged such as solid state drives that forfeited the disk platter design and it does not have limitations of rotational speeds.
- Read/Write arm handles the data writing and data reading from the disk platters. Each disk platter has 2 read/write heads, which means that 1 head is used on 1 side of the platter for data read/write. Whenever data is written, it is done by changing the magnetic polarity of the region on the disk platter to save information in binary form. Whenever data is read from the disk, it is done by examining the magnetic polarity of the region on the disk platter. Additionally, whenever data is written or read, the read/write arm does not need to touch the surface of the disk platter to read or change the polarity of regions on the platter. When the spindle rotates, there is a small gap between the read/write arm and the disk platter which is also called as flying height/floating height/head gap. When the spindle stop rotating, the read/write head will spun down and be placed in a special area on the disk platter called the landing zone. This movement is called head parking where the read/write arm is moved to the landing zone where there is no data saved, this area has lubricants to minimize the friction between the read/write arm and the platter. The disks logic will ensure that the head is parked within the landing zone before starts spinning and ensures that it is set in the landing zone when it is stop spinning to avoid head crash. Head crash is an event the head of the read/write arm touches the disk platter surface outside of the safe landing zone, this will cause the magnetic coating of the disk platter to be scratched, the head might be damaged with data loss occurring at the same time. The scattered particles of the scratched disk platter might even cause more head crashes within the disk causing more data loss and eventually making the drive unusable.
- The control circuit board is a printed circuit board that is installed at the bottom part of the hard drive. It is made up of a microprocessor, internal storage, electric circuitry and other components. These components controls the power supply to the motors and controls the spindle rotation speed. They also manage the communication and data transmission between the disk and the host. Additionally, it help the drive to gain optimum data access through the interchanging use of the different read/write heads and movement of the actuator arm.



- A disk drive track is a circular path on the surface of a disk around the spindle where data in recorded magnetically. Track numbering starts from zero and the numbering starts from the edge of the platter to the innermost portion of the platter. We use the number of tracks per inch (TPI) to measure the density of the tracks alignment within a disk platter.
- Each track is divided into smaller units called sectors. Sectors are the smallest unit of storage that can be searched independently within a disk drive. The track and sector architecture is written to the disk by the disk manufacturers using a formatting tool on the disk platter. Different hard disk's track may contain different number of sectors. The first personal computer in the world had 17 sectors in each of its disk drive tracks, nowadays, our hard drives has much more sectors in each track compared to the past. The number of tracks within a disk drive differs based on the diameter and the TPI density of the disk platter, it may contain up to few thousand of tracks in a disk drive.
- Commonly, each sector can save 512 byte of user data, but some hard disk can be formatted to have a much bigger sector size such as 4KB sectors.
- Cylinders : All sets of the tracks that have the same track number on all the disk platter (Up and Bottom side) within a single disk drive forms a disk cylinder. The number of disk cylinder is equal to the number of disk tracks within the hard drive. However, the location of the magnetic head is determined by the disk cylinder number and not the track number which means that the disk controller moves the read/write head to any location on the disk platter based on disk cylinder number instead of track number. The addressing of individual sectors of the disk is traditionally done by referring to cylinders, heads and sectors (CHS).

SAS Drives

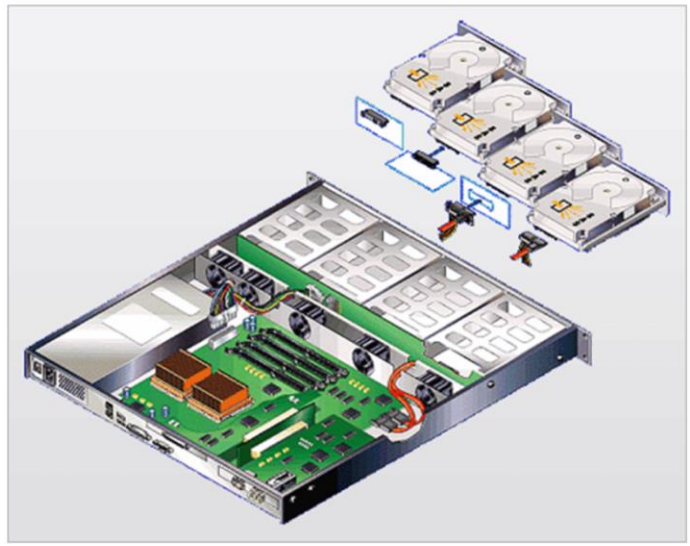


Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- SAS stands for Serial Attached SCSI (SCSI Stands for Small Computer System Interface, typically pronounced as “scuzzy”) and is a technology for transferring data from and to hard drives.
- SAS has provided a high performance, high reliability, scalable and easy operation solution for enterprise datacenters.
- SAS has provided unified support in terms of physical connector ports for ATA (Serial ATA) while is compatible with the parallel SCSI logics, providing more choices for servers and network storage for the disk types that can be used within their storage.
- SAS drives are faster and have more bandwidth throughput than SATA drives. SAS drives are also more reliable than SATA. However, SAS drives often comes with a higher cost range compared to SATA drives.

NL - SAS Drives



Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- Nearline-SAS (NL-SAS) are basically enterprise SATA drives with a SAS interface, R/W head and disk platters. In addition, they have the rotational speed (spin) of traditional enterprise-class SATA drives (7800 or slower RPM) with the fully capable SAS interface command sets. NL-SAS drives is a combination of SAS connectors and SATA drive platter which means that it has all the connectivity benefits of SAS while using a SATA disk platter that comes near to SAS drive level performance. Since, the internal disk structure uses SATA technologies, NL-SAS drive cannot reach full SAS drive level of performance and only come near to SAS level performance.
- NL refers to the abbreviation of Near Line, which means near line storage where it can be considered as the midpoint between the customer's online storage and offline storage. Here, online storage refers to data that has high access rate and frequently accessed by customer such as databases and offline storage where data access rate is very low and with very low access frequency such as customer data backup archives.

- The requirement for near line storage often doesn't require high performance drives but it need to have decent speed and high data transmission rate. For example, it is used for storing certain archive of customer data that that is not frequently accessed. It does not need a high performance drive to store these data but it needs drives with good data access capabilities. At the same time, these unfrequently accessed data usually have bigger file size which requires that near line storages need a bigger requirement on the storage capacity of the drives.
- NL-SAS drives can be considered as mid tier hard drives with good data access speed and high storage capacity which makes it a good choice in multiple different types of customer scenario where the customer wants to store large volumes of data and wants to have high data access performance with lower cost compared to investing in high tier drives. High performance drives such as solid state drives (SSD) often has higher cost and lower storage capacity and not suitable in scenario where data is not frequently accessed and data size is too large.

Solid State Drives (SSD)

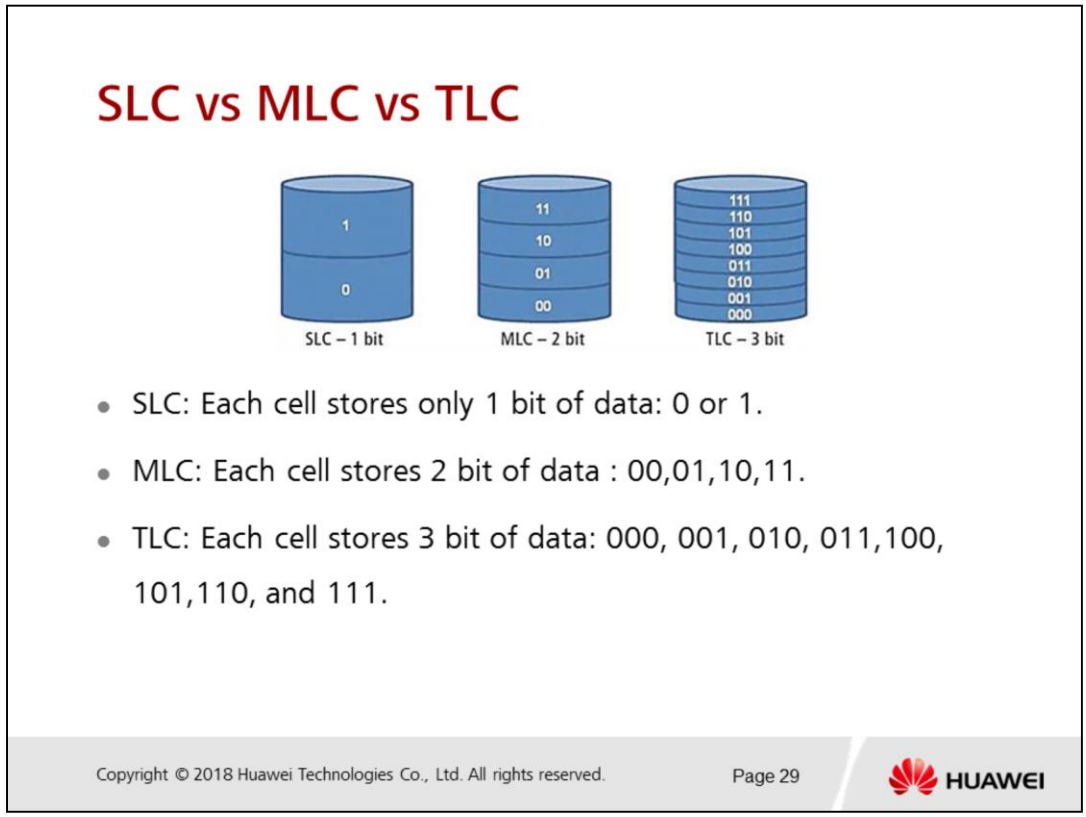
- As price lowers and storage capacity increases, SSD is becoming more and more popular in the market.
- SSD Principles:
  - Uses flash technology to store data.
  - No mechanical structure with lower energy consumption, low heat signature and low noise.
- Limited lifetime due to SSD usage frequency.
- 3 types of SSD:
  - SLC (Single Level Cell)
  - MLC (Multi Level Cell)
  - TLC (Triple Level Cell)

Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

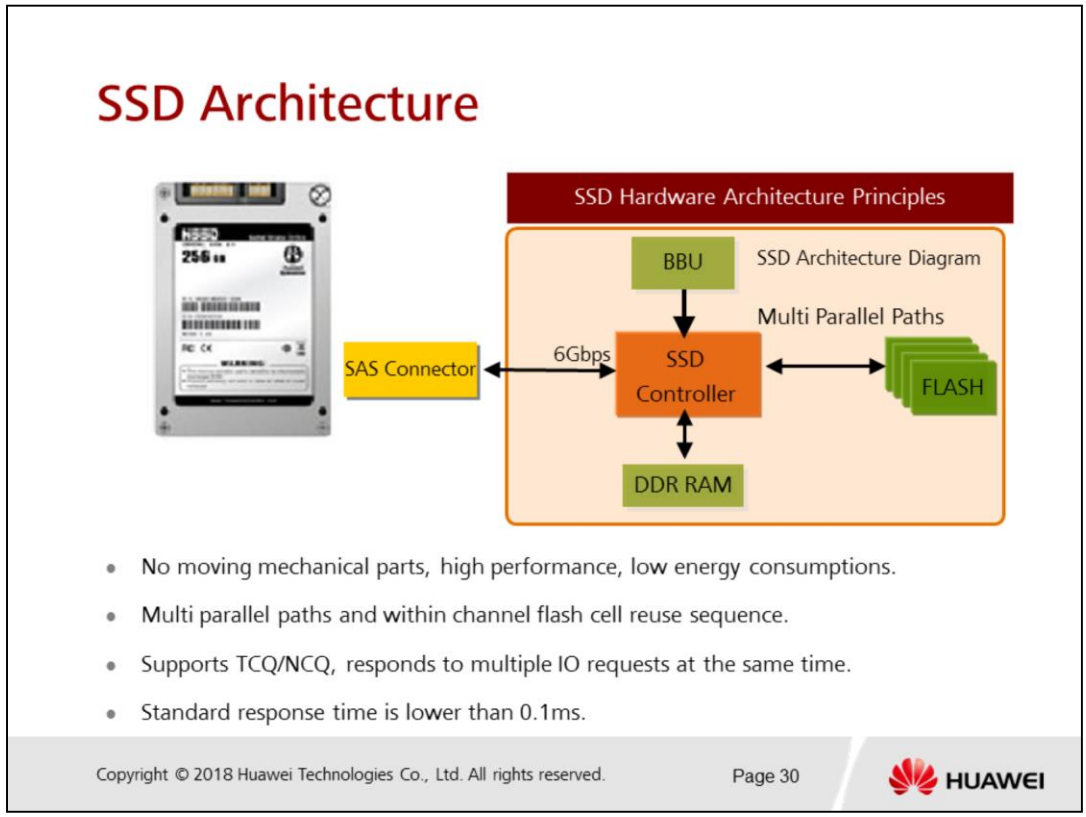
HUAWEI

- Although traditional mechanical drives will not be disappearing from the market in a short period of time, but its replacement which is the Solid State Drive (SSD) has become more and more popular and widely used in the market. SSD does not use magnetic materials to store data but uses basic units of NAND Flash called "cell" as storage units to store data. NAND Flash is a non-volatile random access memory storing medium which has a unique feature of not losing any written data even if the power is cut off. This technology allows data to be quickly written to SSD drives much faster than compared to mechanical drives. Another benefit of SSD is that it does not produce loud noise or generate high heat signatures like traditional mechanical drives such as SATA/NL-SAS/SAS drives.
- SSD does not have moving mechanical parts which greatly lowers the risk of mechanical part failures, but this does not mean that its lifetime is unlimited. Due to NAND Flash is a non-volatile medium, each block of data needs to wiped before a new block of data can be written or else it may cause error data. This increases the write/erase cycles of the SSD drive. However, NAND Flash has a limited write/erase limit which means that each cell of the SSD has a certain limited number of times it can be erased and written into, once the limit is exceeded, that particular cell can no longer be used for data read/write. However, these kind of wear and tear can be easily monitored and predicted so that preparations can be made in advance for disk replacement in time. On the other hand, mechanical hard drives failures have no early symptoms which means that disk failure can happen at any time and replacement disk must be prepared at all times.

- All flash memory suffers from wear, which occurs because erasing or programming a cell subjects it to wear due to the voltage applied. Each time this happens, a charge is trapped in the transistor's gate dielectric and causes a permanent shift in the cell's characteristics, which, after a number of cycles, manifests as a failed cell. It is important to understand this before making the option between choosing which type of SSD that you want to invest in. The types of the SSD can be differentiated as following:
    - SLC (Single Level Cell): Stores 1 bit of data in 1 cell. Lowest wear and tear. Longer lifespan. But expensive and available in small capacities only.
    - MLC (Multi Level Cell): MLC memory is more complex and can interpret 2 bits from a signal stored in a single cell. This makes it denser for a given area and so cheaper to produce, but it wears out faster.
    - TLC (Triple Level Cell): Storing 3 bits of data per cell, TLC flash is the cheapest form of flash to manufacture. Cells will survive considerably less read/write cycles compared to MLC NAND. This means that TLC flash is good for consumer use only.

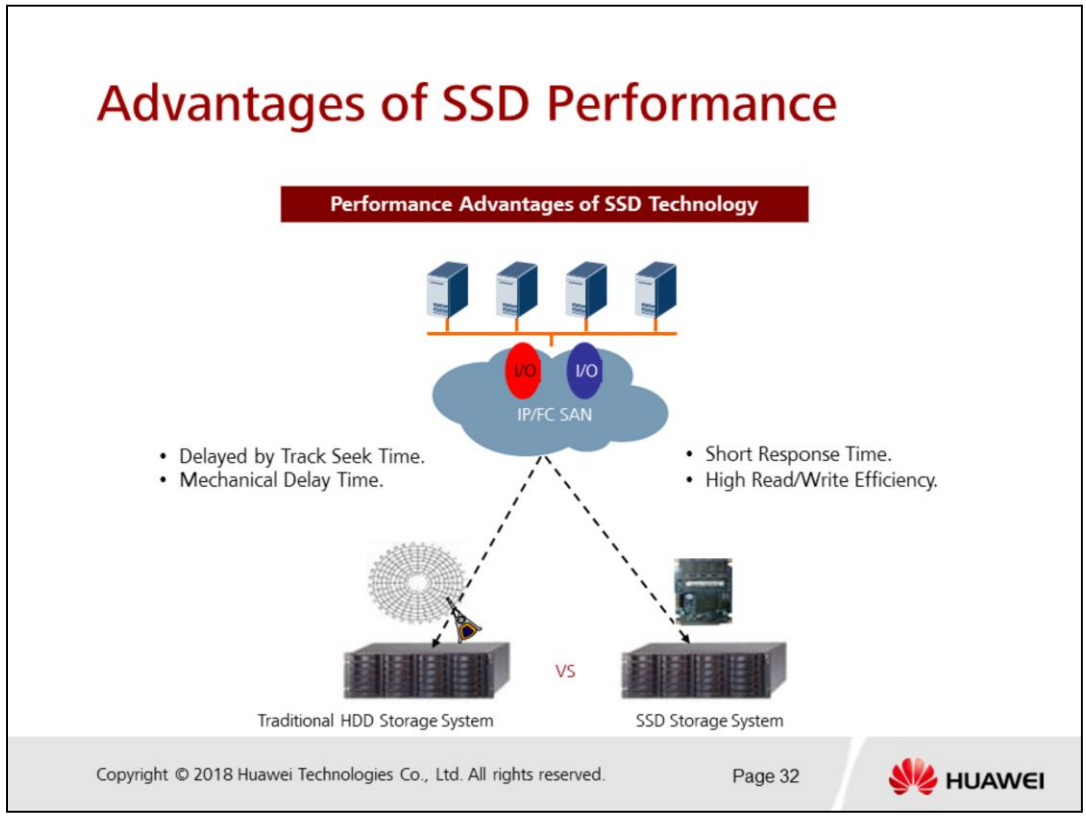


- Each cell in the NAND Flash of the SSD drive traditionally can store 1 bit of data, which is either 0 or 1. However, in newer generation of SSD drives, we are able to store more data in a cell using a special technologies which gave birth to the different types of SSD drives in the market today. The types of SSD drives can be categorized into Single Level Cell (SLC), Multi Level Cell (MLC) and Triple Level Cell (TLC).
- SLC only stores 1 bit of data which is in the binary form of 0 or 1. MLC on the other hand is able to store 2bits of binary data which means that it is able to store data in four states such as 00, 01, 10, 11. TLC is able to store 3 bits of data which means that it is able to store 8 different states of data which is shown on the diagram above.
- Three different types of cell although is able to store different amounts of data but their physical size remains the same, and this is also one of the reasons why SSD drive capacity are growing exponentially. Initially, SSD were only available in 64Gb or smaller capacity, but now the largest TLC SSD available on the market can store up to 2TB of data.
- However, different types of SSD has different level of wear resistance which results in different levels of drive reliability. The wear resistance of SSD disk is also an important factor in choosing a SSD disk.
- SLC has the highest reliability however comes at the highest cost, MLC has the middle level of reliability while TLC has the lowest reliability but with the lowest cost. Hence, choosing the right kind of SSD disk depends on the customer's requirement and the storage scenario and also the costing evaluations.

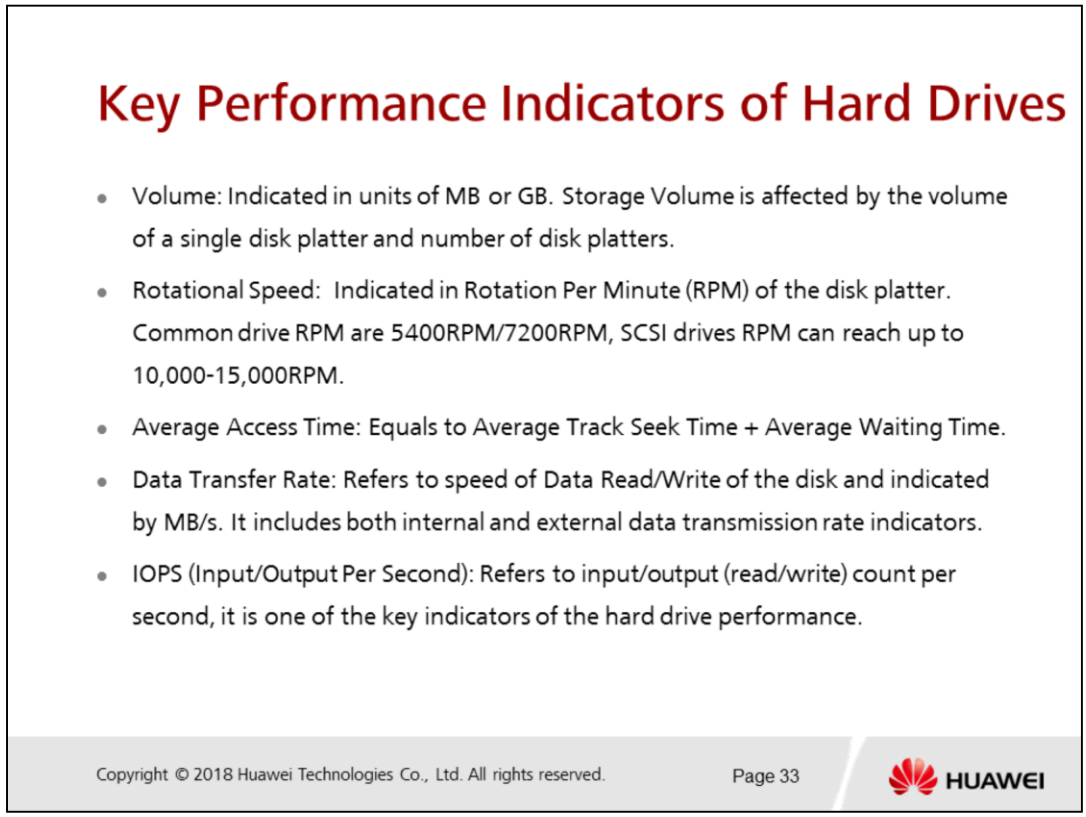


- The diagram above shows the architecture of the SSD and the technologies supported by it to optimize the read/write cycles of data for better performance.
- The SSD controller is able to process commands and IO requests in parallel due to the multi parallel path design that exist between the controller and the NAND Flash cells within the drive. The cell write/erase cycle is also optimized and allows the drive to reclaim and reuse any of the flash cells to store data efficiently.
- Native Command Queuing (NCQ) is an extension of the Serial ATA protocol allowing hard disk drives to internally optimize the order in which received read and write commands are executed. This can reduce the amount of unnecessary drive head movement, resulting in increased performance (and slightly decreased wear of the drive) for workloads where multiple simultaneous read/write requests are outstanding, most often occurring in server-type applications.

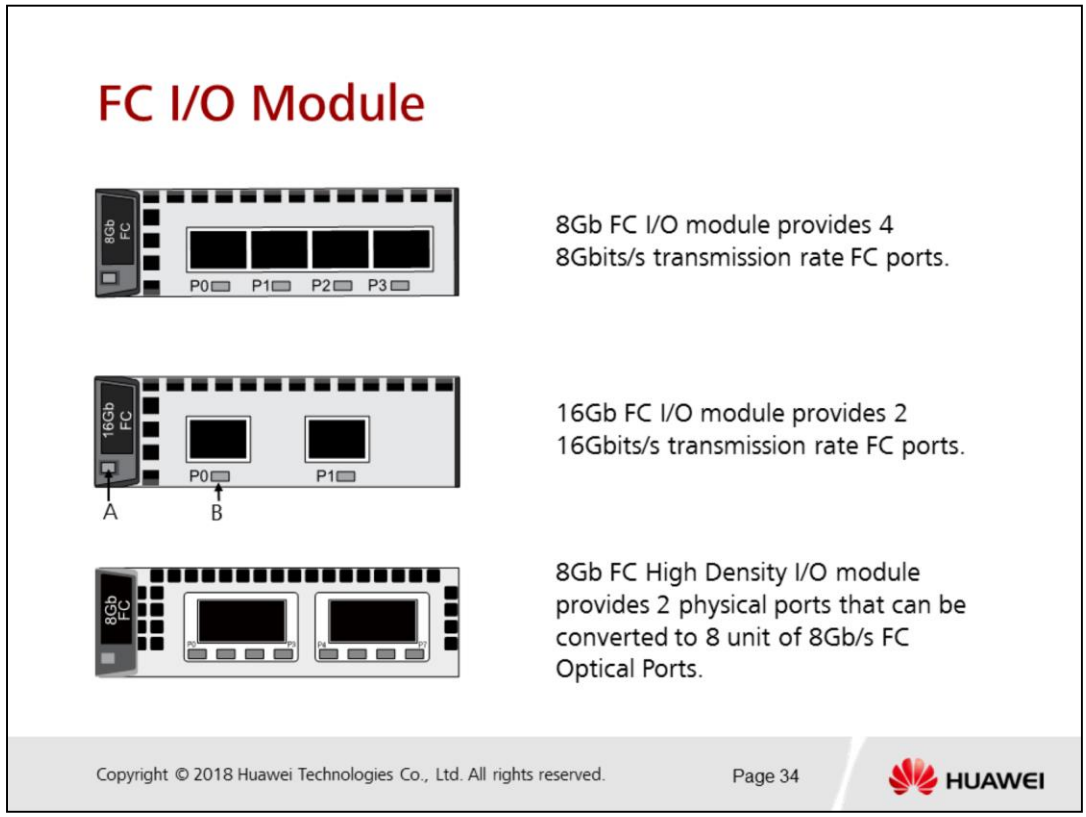
- Tagged Command Queuing (TCQ) is a technology built into certain ATA and SCSI[1] hard drives. It allows the operating system to send multiple read and write requests to a hard drive. Without TCQ, an operating system was limited to sending one request at a time. To boost performance, the OS had to determine the order of the requests based on its own – possibly incorrect – perspective of the hard drive activity (otherwise known as I/O scheduling). With TCQ, the drive can make its own decisions about how to order the requests (and in turn relieve the operating system from having to do so). Thus TCQ can improve the overall performance of a hard drive if it is implemented correctly.
- NCQ (Native Command Queuing) and TCQ (Tagged Command Queuing) are technologies designed to re-sort the commands from the host to drive and increase the performance of the drive. NCQ technology is implemented for 300 MB/s Serial ATA standards and focuses on mainstream hard drive products. Meanwhile, TCQ technologies is implemented for SCSI2 standards and (ATA-4 standards) and mainly focuses on servers and enterprise grade hard drive products.
- To use the NCQ or TCQ technologies, the disk connector chipset and the hard drive itself need to support those technologies for it to work. This means that if you bought a hard drive that does not support NCQ, even if your host motherboard supports NCQ, you will not be able to enable this feature and increase your hard drive's performance



- Traditional HDD storage system has more delays compared to SSD storage system due to its mechanical parts. Traditional storage has to uses a portion of time for track seeking and moving the mechanical parts for data read/write operations. SSD storage has no moving mechanical parts and uses NAND Flash technologies without the need for track seeking which saves more time compared to traditional storages.
- Short Response Time:
  - Due to the nature of the mechanical parts in traditional storages, most of time is wasted in track seeking and mechanical delay, hence data transmission speed is heavily restricted by these parts. SSD storages has no mechanical parts which saves the time for track seeking and mechanical delay, which in turns allows a much faster IO request response rate.
- High Read/Write Efficiency:
  - Mechanical drives needs to move the heads constantly when it is performing random read or write operations which lowers the efficiency of the read/write process. SSD in the other hand, calculates the data storage locations internally and performs the read or write process, which saves a lot of mechanical operation time that in turns increases the read/write efficiency greatly.

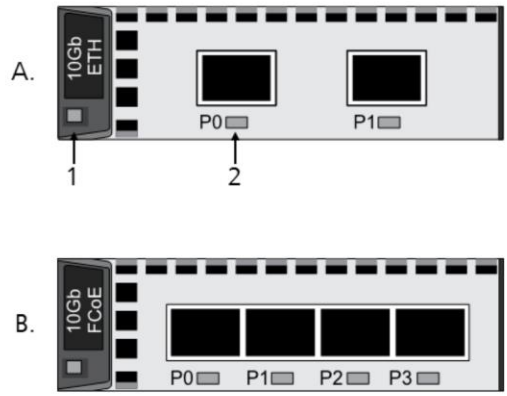


- IOPS is a key indicator for high frequency random read write applications such as OLTP (Online Transaction Processing). Another key indicator is **Throughput**, which refers the amount of data that can be successfully transferred in a single unit of time. For large amount of sequential data read write applications such as video editing of TV stations and video streaming, they are more focused on the indicator such as Throughput.



- FibreChannel Input/Output Modules (FC I/O Module) provides connectivity and expansion to storage system by providing different number FC ports with varying transmission rate based on the models of the I/O module.
- Indicator lights on the modules above shows status of :
  - Refer to A in diagram above: Module Power Indicator/ Hot Swap Button. This shows the power status of the module and also serves as a button to initiate the hot swapping process of the module for replacement.
    - Steady green: The interface module is working correctly.
    - Blinking green: The interface module receives a hot swap request.
    - Steady red: The interface module is faulty.
    - Off: The interface module is powered off or hot swappable.
  - Refer to B in diagram above: Link/Speed Indicator Light. This shows the current running link speed of the I/O module FC port.
    - Steady blue: The data transfer rate between the storage system and the application server is 16 Gbit/s.
    - Blinking blue: Data is being transferred.
    - Steady green: The data transfer rate between the storage system and the application server is 4 Gbit/s or 8 Gbit/s.
    - Blinking green: Data is being transferred.
    - Steady red: The port is faulty.
    - Off: The link to the port is down.

FCoE I/O Module

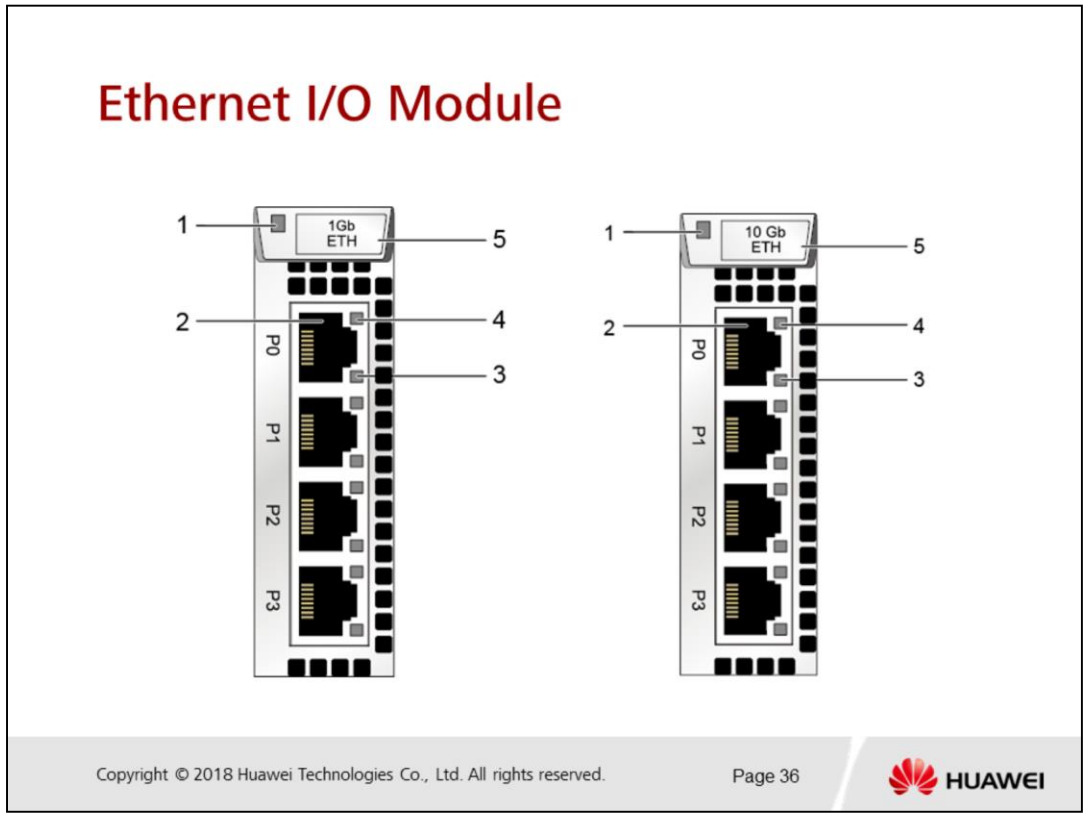


Note:
1 = Module Power Indicator Light / Hot Swap Button
2 = Port Link/Speed Indicator Light

Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

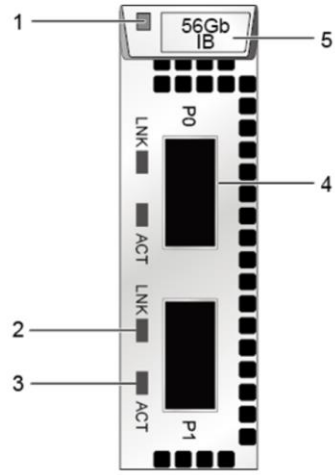
HUAWEI

- Refer to A in diagram above: 10Gb FCoE(FibreChannel Over Ethernet) I/O module provides 2 10Gbit/s data transmission rate FCoE port.
  - 10Gb FCoE I/O module (2 Ports) only supports Direct Network connection.
- Refer to B in diagram above: 10Gb FCoE(FibreChannel Over Ethernet) I/O module provides 4 10Gbit/s data transmission rate FCoE port.
  - 10Gb FCoE I/O module (4 Ports) only supports Switched Network connection.
- 10Gb FCoE I/O module is not recommended to run both iSCSI and FCoE protocols at the same time as it will lower the overall performance rate and have higher fluctuations in performance.



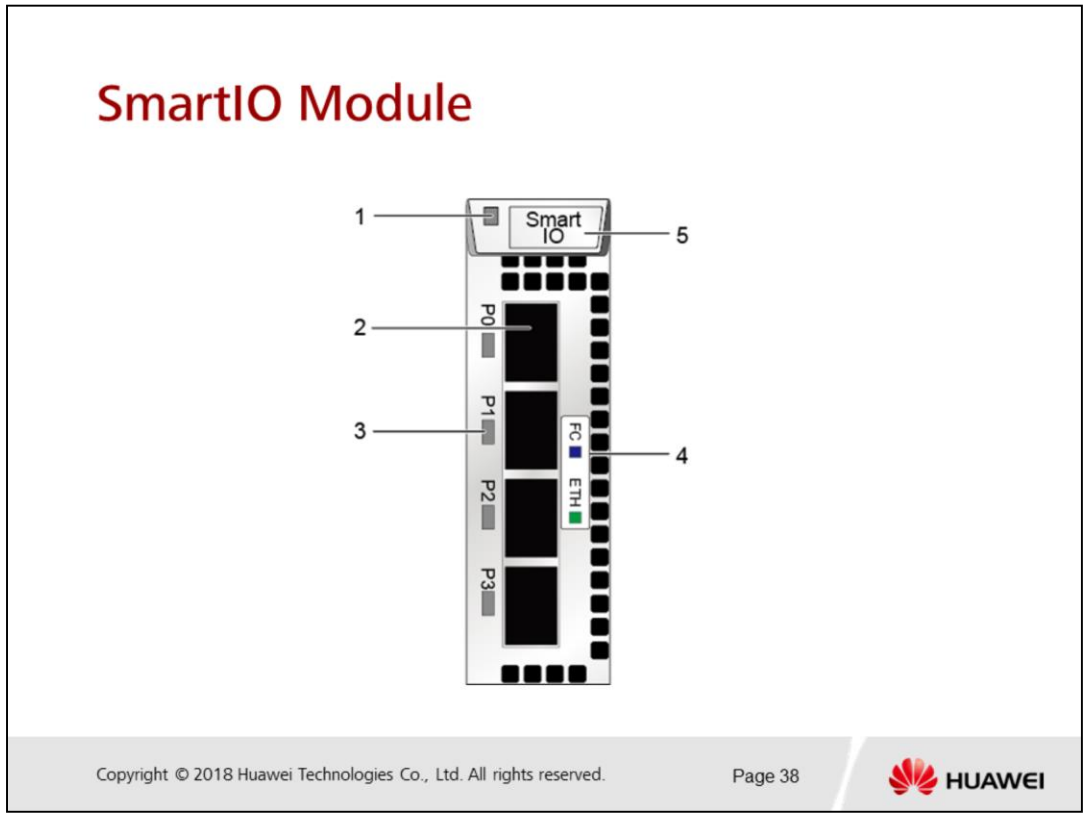
- GE (Gigabit Ethernet) Electrical I/O Module provides 4 1Gbits data transmission rate electrical Ethernet ports.
- 10GE Electrical I/O Module provides 4 10Gbit/s electrical Ethernet port, it supports auto speed negotiation between 10GE and GE for these ports.
- Indicator Lights:
  - Module Power Indicator Light/ Hot Swap Button.
  - 10GE Electrical Port.
  - 10GE Electrical Port Link/Active Indicator Light.
  - 10GE Electrical Port Speed Indicator Light.
  - Module Latch.

56Gb InfiniBand(IB) Module



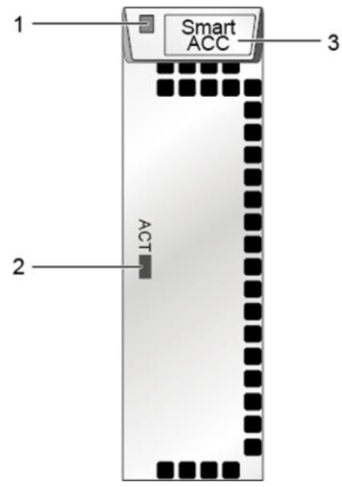
Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

- The 56Gb IB (InfiniBand) I/O module provides 2 (4x14Gbit/s) data transmission rate IB ports.
- Indicator Lights:
  - Module Power Indicator Light/ Hot Swap Button.
    - Steady green: The interface module is working correctly.
    - Blinking green: There is a hot swap request to the module.
    - Steady red: The module is faulty.
    - Off: The interface module is powered off or hot swappable.
  - 56Gb IB Port Link Indicator Light.
    - Steady green: The port is connected properly.
    - Off: The port link is down.
  - 56Gb IB Port Active Indicator Light.
    - Steady orange: Data is being transmitted.
    - Off: No data is being transmitted.
  - 56Gb IB Port.
  - Module Latch.



- SmartIO Module supports three types of optical transceiver with data transmission rate of 16Gb, 8Gb, 10Gb.
- Indicator Lights:
  - Module Power Indicator Light/ Hot Swap Button.
  - SmartIO Port.
  - Port Link/Active/Mode Indicator Light.
  - Port Mode Silk Screen.
  - Module Latch.
- If the mode of the SmartIO port is set to FCoE/iSCSI or Cluster on the software interface, the port indicator is in ETH mode and a 10 Gbit/s optical module is required. 10 Gbit/s optical transceiver does not support GE/10GE auto negotiation.
- If the mode of the SmartIO port is set to FC on the software interface, the port indicator is in FC mode, and an 8 Gbit/s or 16 Gbit/s optical module is required.
- If the mode of the SmartIO interface module is set to FCoE/iSCSI and the host uses the FCoE protocol, the module must connect to FCoE switches for networking and a 10 Gbit/s optical module is required.
- If the mode of the SmartIO interface module is set to FCoE/iSCSI and the host uses the iSCSI protocol, the MTU value of the SmartIO port must be the same as that of the host.

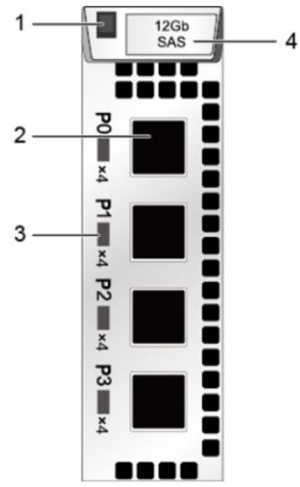
Smart Acceleration Module



Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

- Smart Acceleration Module can concurrently provide hardware acceleration capabilities of data deduplication, fingerprint calculation, gzip decompression and compression, offloading the CPU load and increase the overall performance of the storage system in data deduplication and compression.
- Indicator Lights:
  - Module Power Indicator Light/ Hot Swap Button.
  - Active Indicator Light.
  - Module Latch/ Silkscreen.

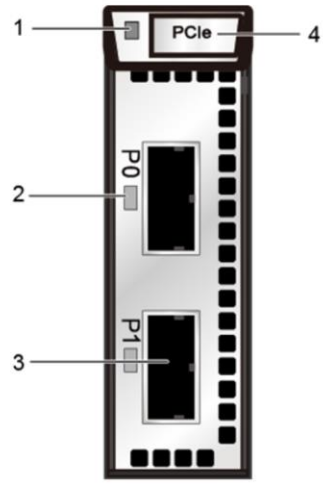
12Gb SAS Expansion Module



Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

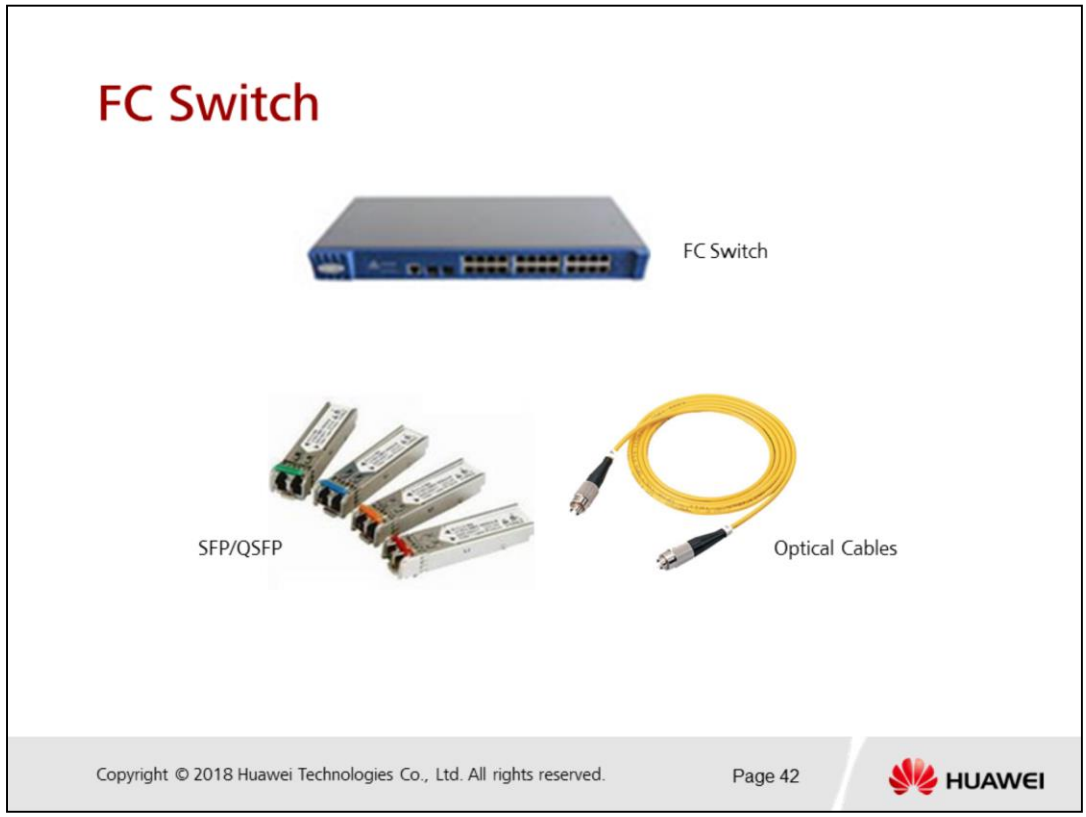
- A SAS interface module provides four 4 x 12 Gbit/s mini SAS HD expansion ports that provide connectivity to disk enclosures.
- The SAS interface module connects to the back-end storage array of the storage system through a mini SAS HD cable.
- When the transfer rate of the connected device is less than that of the expansion port, the expansion port automatically adjusts the transfer rate to that of the connected device to ensure the connectivity of the data transfer channel.
- Indicator Lights:
  - Module Power Indicator Light/ Hot Swap Button.
  - Mini SAS HD Expansion Port.
  - Mini SAS HD Expansion Port Indicator Light.
  - Module Latch.

PCIe Module



Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

- PCIe Module provides 2 PCIe ports that are the main service port between Oceanstor 18000 V3 and the data switches, it is used to exchange dataflow and control flow messages between storage engines within the network.
- Indicator Lights:
  - Module Power Indicator Light/ Hot Swap Button.
  - PCIe Port Link/Speed Indicator Light.
  - PCIe Port.
  - Module Latch.



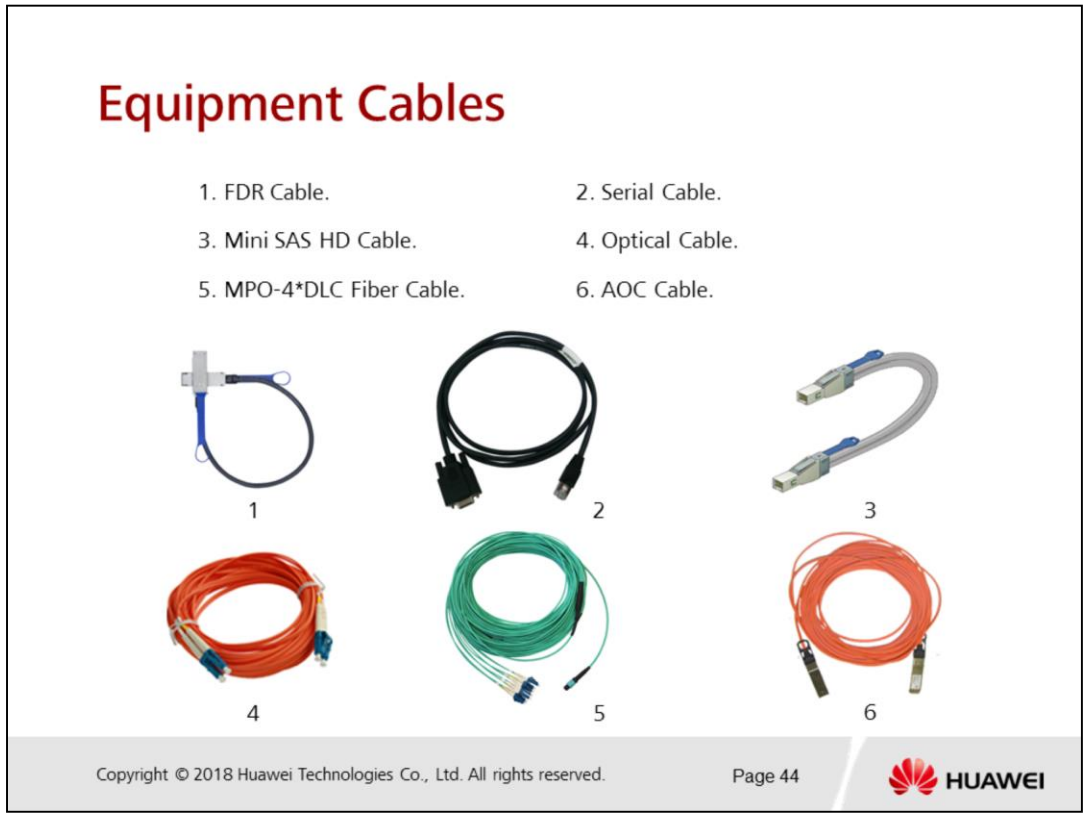
- Fibre Channel switch is a network switch compatible with the Fibre Channel (FC) protocol. It allows the creation of a Fibre Channel fabric, that is the core component of a storage area network (SAN). The fabric is a network of Fibre Channel devices which allows many-to-many communication, device name lookup, security, and redundancy. FC switches implement zoning, a mechanism that disables unwanted traffic between certain fabric nodes.
- The small form-factor pluggable (SFP) is a compact, hot-pluggable optical module transceiver used for both telecommunication and data communications applications. An SFP interface on networking hardware provides the device with a modular interface that the user can easily adapt to various fiber optic and copper networking standards such as FibreChannel and Ethernet.
- The Quad Small Form-factor Pluggable (QSFP) is a compact, hot-pluggable transceiver used for data communications applications similar to SFP. Certain technologies such as InfiniBand uses QSFP transceivers for the connection to networking hardware.
- A fiber optic cable is a network cable that contains strands of glass fibers inside an insulated casing. They're designed for long distance, very high performance data networking and telecommunications. Compared to copper cables, fiber optic cables provide higher bandwidth and can transmit data over longer distances.

Host Bus Adapter (HBA) Cards

Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

- A host bus adapter (HBA) is a circuit board and/or integrated circuit adapter that provides input/output (I/O) processing and physical connectivity between a host system, or server, and a storage and/or network device. Because an HBA typically relieves the host microprocessor of both data storage and retrieval tasks, it can improve the server's performance time.
- FC HBA Card: Commonly known as Fibre Network Cards. The transmission protocol used by FC HBA cards is FibreChannel protocol and usually is connected to the FC Switches using fibre optic cables.
- iSCSI HBA Card: Uses iSCSI protocol for data transmission and has same ports with Ethernet network cards. By integrating the hardware iSCSI initiator on the circuit board and using TCP/IP offload engine technology to do data processing on the HBA cards, we are able to lessen the host CPU's workload and provide a high usability hardware multi channel feature along with server remote booting feature. Booting from SAN is also possible if the iSCSI card supports remote booting or boot from LAN features.



- FDR (Fourteen Data Rate) cable is custom used for 56Gb InfiniBand I/O Module.
- Serial Cable is used for connecting equipment's serial port to the maintenance terminal. One end of the serial cable is RJ-45 connector used to the equipment and another end is the DB-9 connector used to connect to the maintenance terminal.
- Mini SAS HD cables is used for connecting disk enclosures with high density disk enclosures or connecting between 2 disk enclosures/storage chassis.
- Devices communicate between each other through the optical cables and FC switches. One end of the optical cables connect to the FC HBA card, while the other end connects to the FC switch or Application server. Both end of the optical cables are LC (Lucent Connector) type.
- MPO-4*DLC Fiber Cables is custom used by the 8Gb High Density I/O Module.
- AOC (Active Optical) Cable is used for connecting storage system controller module's PCIe port to the data switches.

Contents

1. Architecture of Storage Systems
2. Components of Storage Systems
3. **Introduction to Huawei Storage Products**

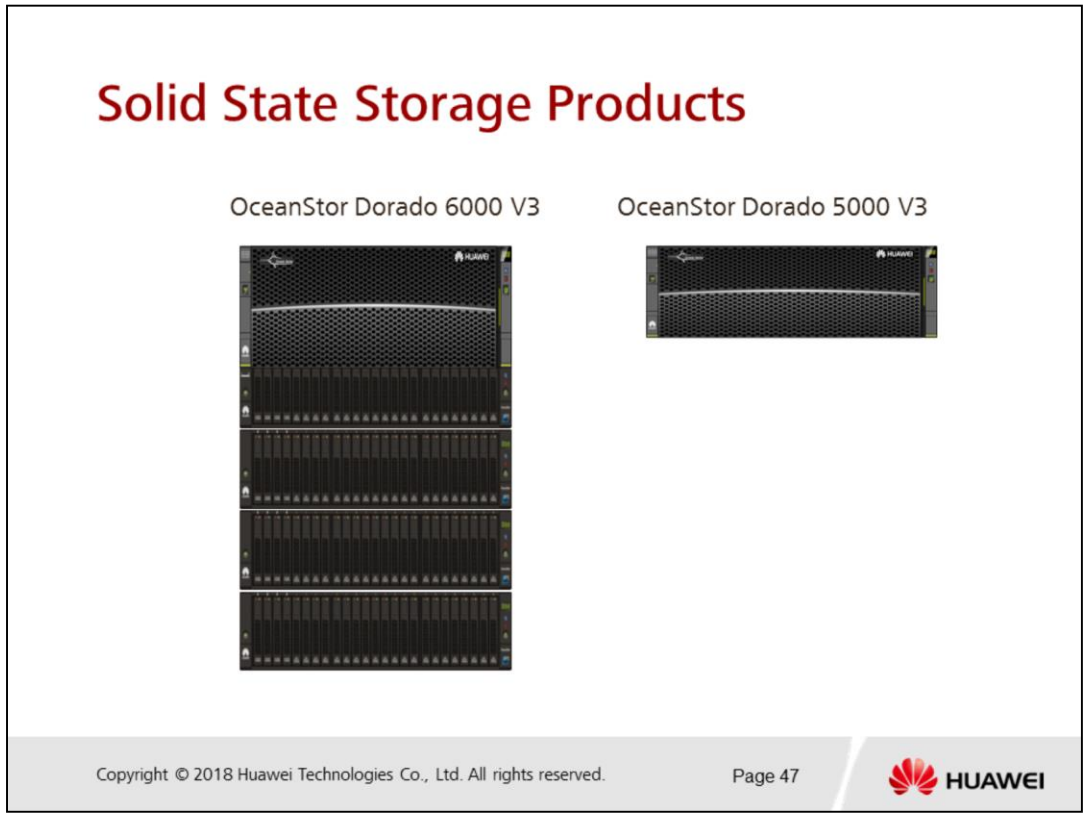
Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

Huawei Converged Storage Products

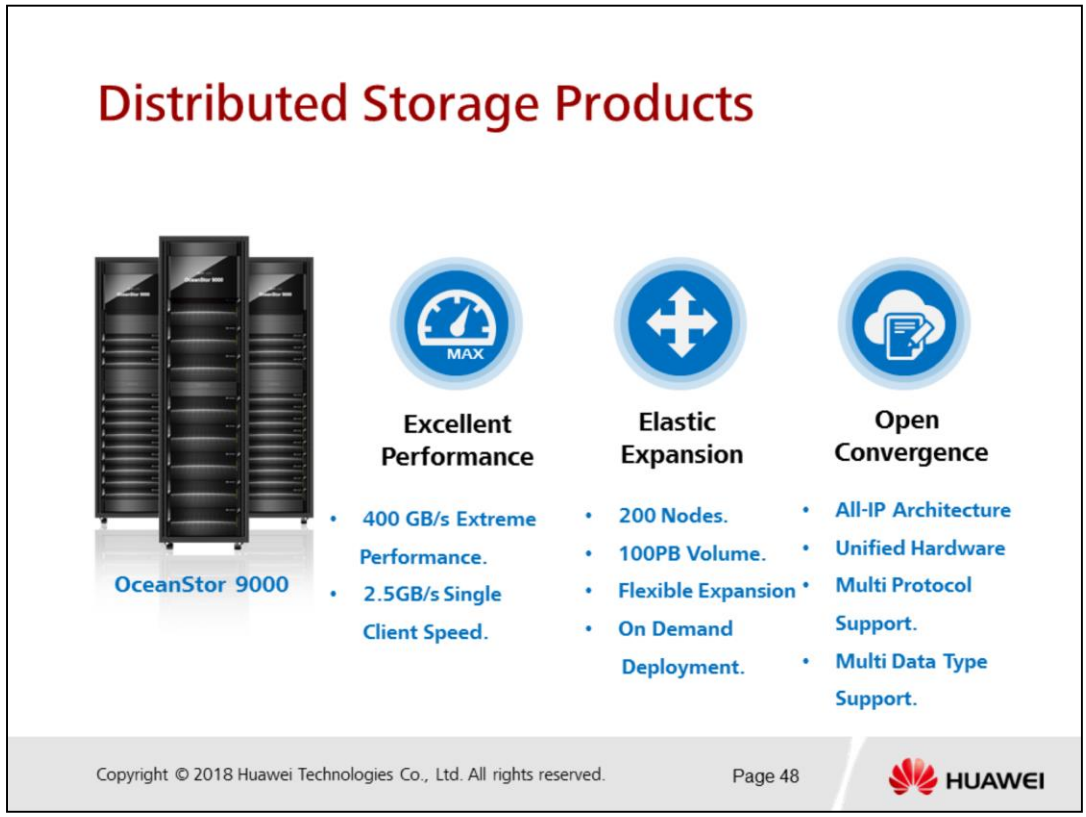
- Entry Level Storage:
  - OceanStor 2200/2600 V3 Series
- Middle Range Storage:
  - OceanStor 5300/5500 V3 Series
  - OceanStor 5600/5800 V3 Series
- High End Storage:
  - OceanStor 6800 V3 Series
  - OceanStor 18000 V3 Series

Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

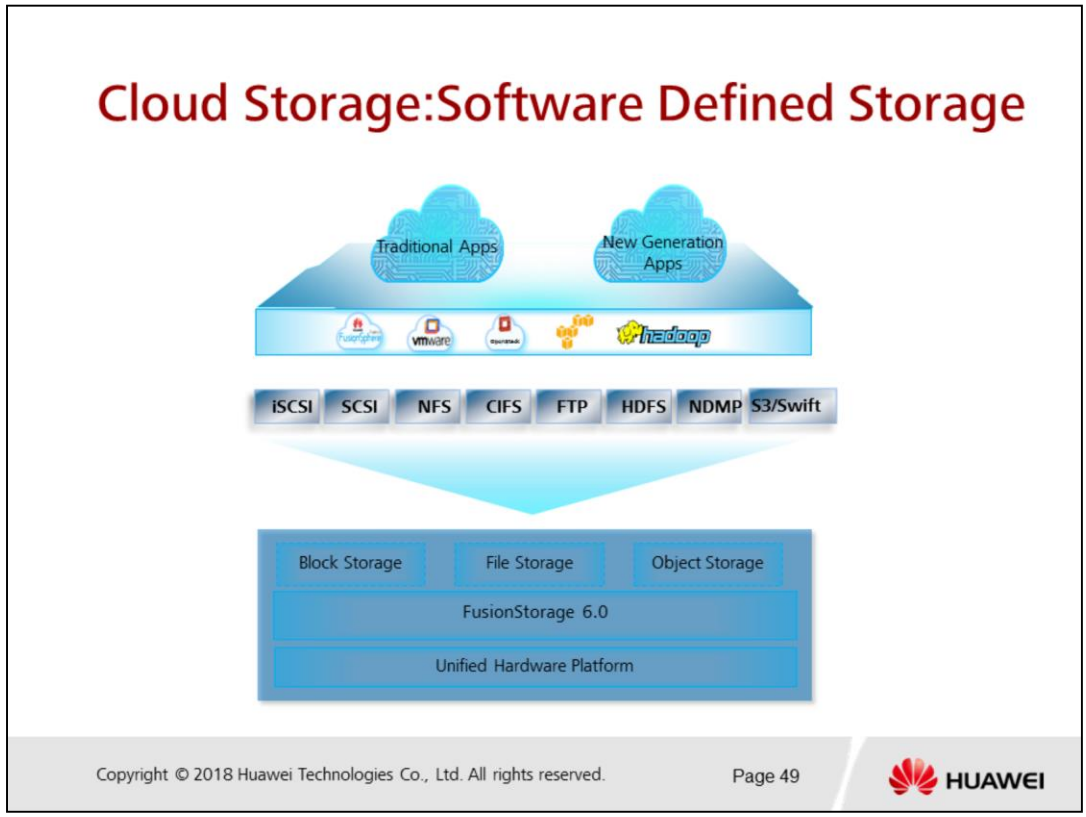
- OceanStor V3 series storage systems has high standard in terms of hardware architecture, unified block and file storage, combines multiple different advanced data application and data protection technologies, ensuring that Huawei storage systems has high performance, high scalability, high reliability and high usability to satisfy large and medium sized enterprise's different kinds of needs for storage products.
- OceanStor 2200 V3 only provides block storage services.



- Highest Speed: 500us stable time delay at millions IOPS, making database faster by 20 times.
- Highest Reliability: Multi Controller Fully Redundant Architecture ensuring no single point of failures in the storage system, all of the components supports online maintenance.
- Highest Savings: Converged storage, online deduplication and decompression saves up to 3 times the storage space and lowers total cost of ownership (TCO) by 70%.
- OceanStor Dorado V3 all-flash storage is the ideal choice for enterprises' mission-critical business. It is the industry's first commercial use of NVMe all-flash storage, and it delivers high-performing, reliable, and efficient storage services.



- As a big data storage product, the OceanStor 9000 provides advantages such as large capacity, high performance, and flexible scalability as well as a range of value-added features. It can be applied to various fields such as broadcasting, media and entertainment (M&E), high-performance computing (HPC), Internet operation, data centers, and large-sized enterprises.
- OceanStor 9000 is a Big Data storage platform that uses the general x86 architecture hardware and a scale-out NAS system. OceanStor 9000 adopts a fully symmetric architecture and supports seamless expansion of 3 to 288 nodes.
- The OceanStor 9000 employs a networking mechanism that features full mesh and redundancy, allowing any nodes to concurrently access any data. Multiple nodes can concurrently access different areas of the same data, achieving high-performance data read and write. The OceanStor 9000 provides standard NFS/CIFS/FTP/HDFS/object storage service compatible with Amazon S3 and OpenStack Swift APIs.



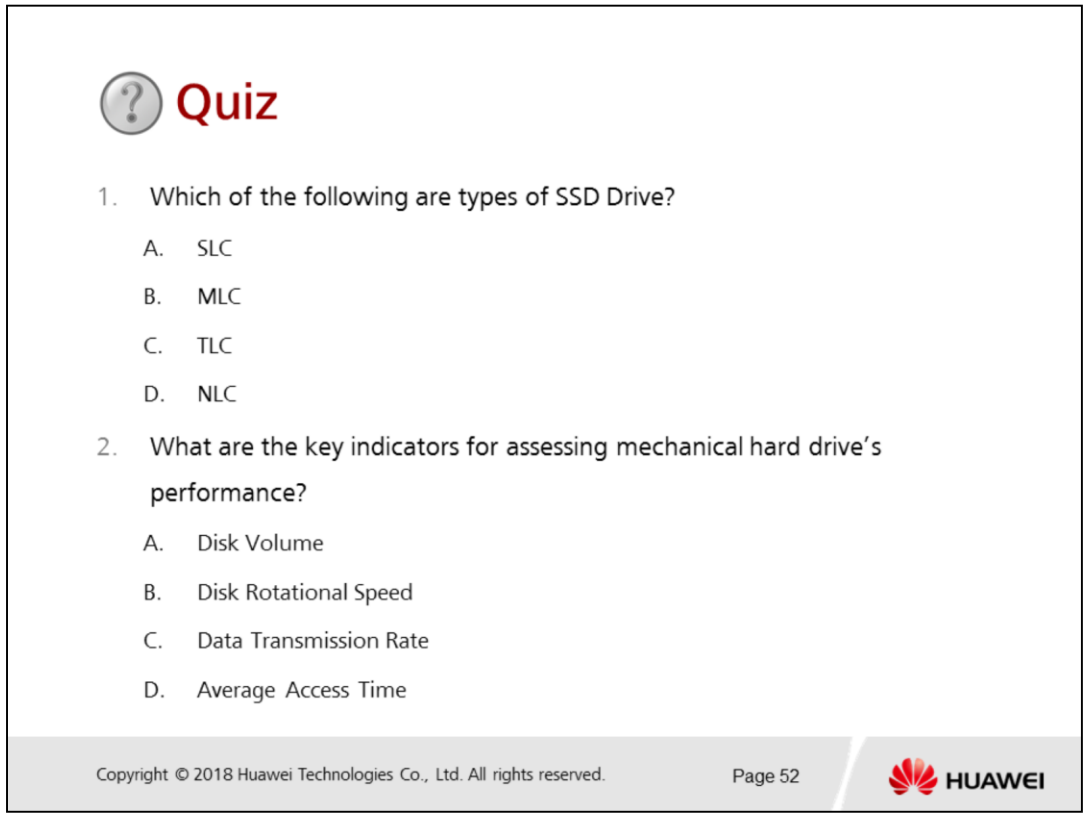
- FusionStorage Software Defined Converged Cloud Storage is based on standard X86 hardware platform for high performance, high scalability and open compatibility with storage resource pools to provide file/block/object based storage for finance, government, and new businesses.
- The product has the high performance, elasticity and flexibility of distributed storage architecture for meeting agile business needs, and also has the reliability and usability of enterprise grade products to easily face the challenges of globalization, cloudification and big data applications.
- All new products are built based on demands and Cloud Storages are no exception. Enterprise Cloud datacenters has hundreds or even thousands of application that presents data in different service formats and data structures such as structures data and unstructured data. Traditional and New generation applications along with Big Data applications also coexist at the same time in the cloud datacenters.

- Huawei FusionStorage6.0 cloud storage system consist of multiple types of resource pools converged together to provide support for multi protocols data access. It also provides agile resource supply and file services such as block/file/object storage to higher level applications.
    - If you are building a Cloud Datacenter, choosing Huawei Cloud Storage not only allows you to fulfill current storage resource pool needs, it also helps in fulfilling future storage needs, volumes and TCO requirements.
    - If your Datacenters has large amount of databases applications, Huawei Cloud Storage resource pools can provide a highly efficient data storing and reading services.
    - If your Cloud Datacenter has applications based on Openstack or Hadoop, Huawei Cloud Storage can fit in seamlessly due to its open architecture.
- Hence, software defined cloud storage can help protect your investments in datacenters and help realize the best match for resources and value.

Summary

- This module mainly introduces:
  - The different kinds of common storage system architectures and its differences.
  - The different kinds of storage components that exist within a storage systems and its functions.
  - The different ranges of Huawei Storage Products and briefly highlights the product's key features.

Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.



- Answers:
  - ABC.
  - BCD.

Thank You

www.huawei.com