05 Introduction to Common Storage Protocols

www.huawei.com

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

HUAWEI

Foreword

- This module mainly describes the common storage protocols:
  - Definition, Technical Principles, and Applications of SCSI\iSCSI Protocols.
  - Definition, Technical Principles, and Applications of SAS Protocol.
  - Definition, Technical Principles, and Applications of PCIe Protocol.
  - Definition, Technical Principles, and Applications of IB Protocol.
  - Definition, Technical Principles, and Applications of CIFS\NFS Protocols.
  - Definition, Technical Principles, and Applications of FTP\HTTP Protocols.

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

HUAWEI

Objectives

- Upon completion of this module, you will be able to:
  - Describe the definitions of common storage protocols.
  - Understand the technical principles of common storage protocols.
  - Understand the application scenarios of common storage protocols.

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

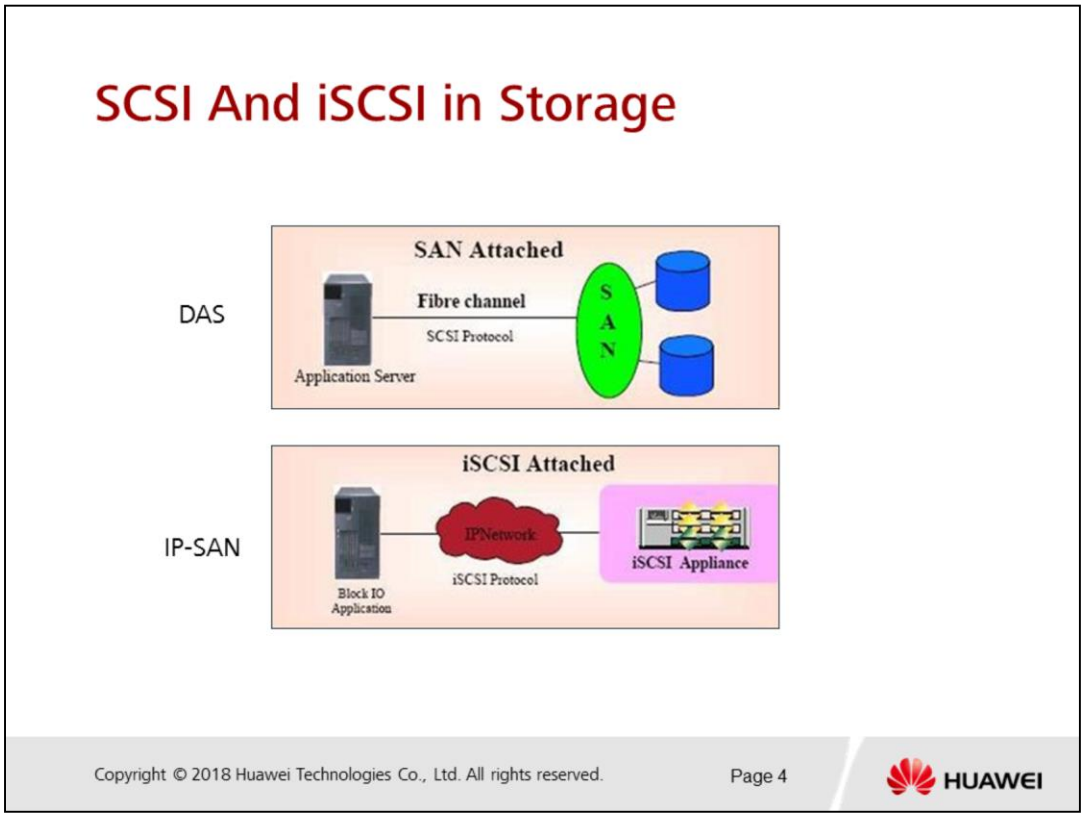
HUAWEI

Contents

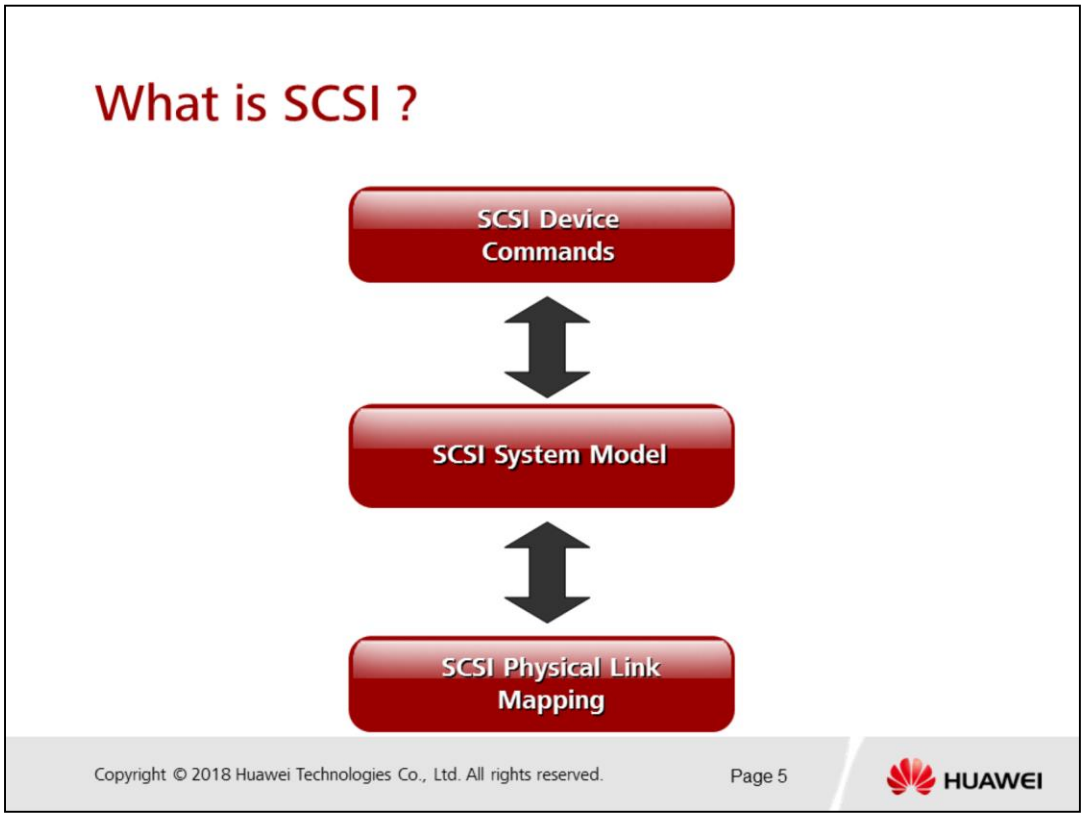
1. **SCSI/iSCSI**
2. SAS
3. FC/FCOE
4. PCIe
5. IB
6. CIFS/NFS
7. FTP/HTTP

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

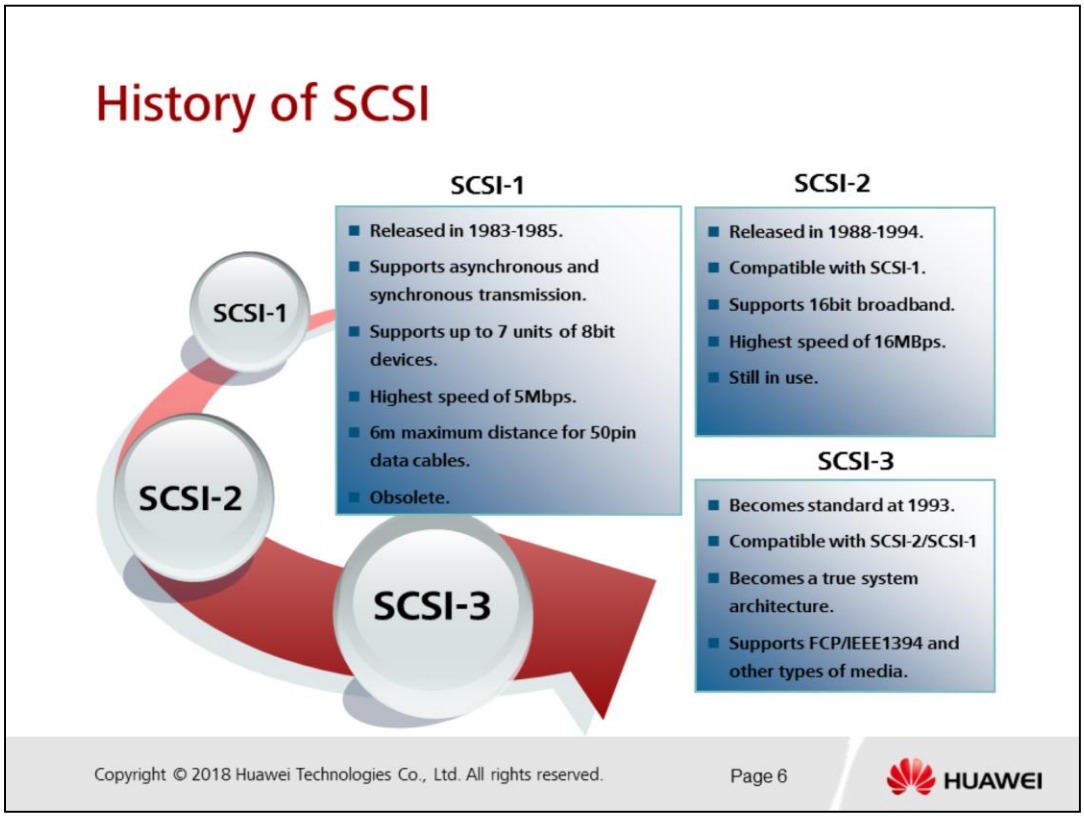
HUAWEI



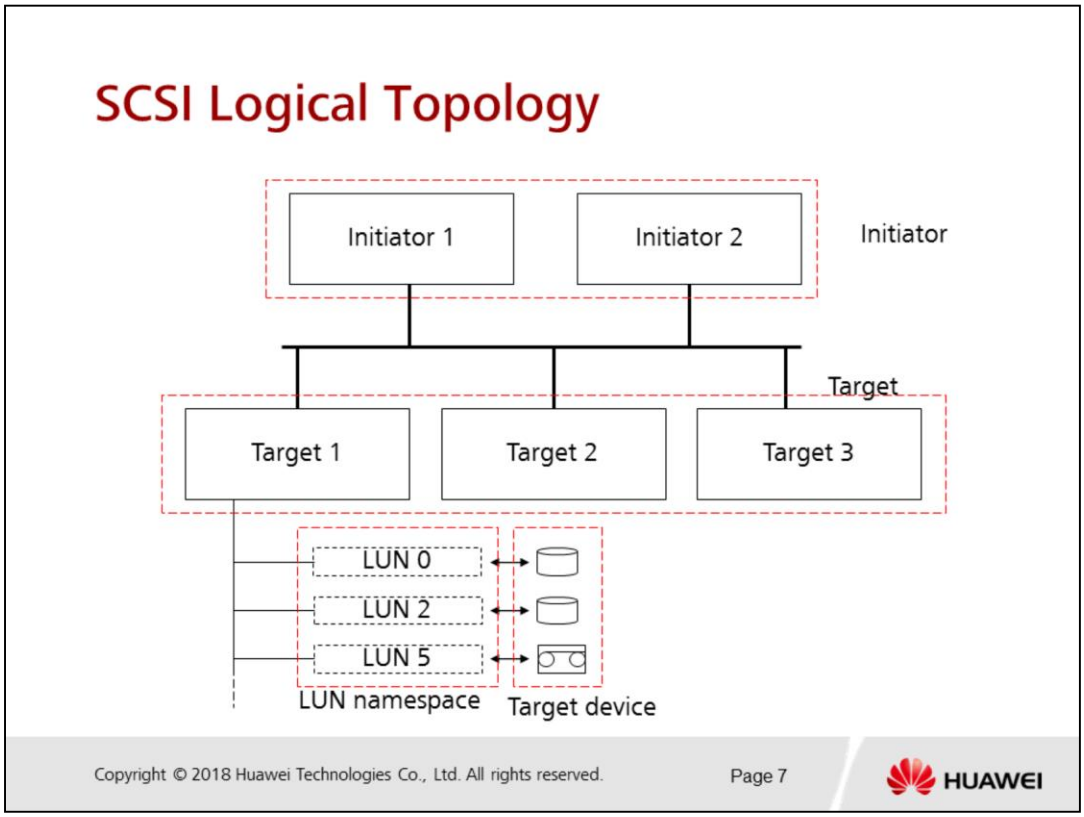
- SAN (Storage Area Network): It is a storage architecture that connects storage device and application servers through network, and this network is solely used for the data access between the host and the storage. When there is a data retrieval request, the data can be transmitted in high speed between the servers and the back end storage devices through the SAN.
- IP SAN: It uses Fast Ethernet/Gigabit Ethernet/ 10 Gigabit Ethernet network to connect the application servers and the storage systems. IP SAN transmits the SCSI commands and data blocks through high speed Ethernet network. It also inherits all the benefits of Ethernet network, and achieves the purpose of building an open, high performance, high reliability, highly scalable storage resource platform.



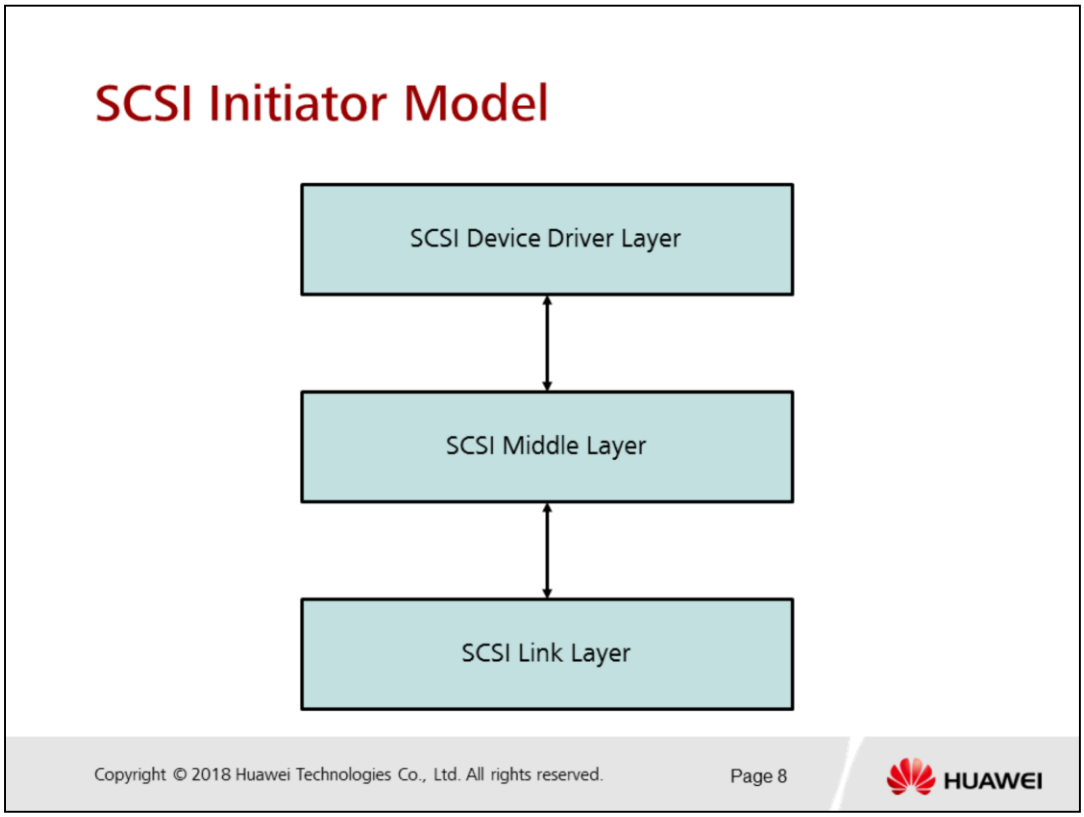
- SCSI, which is Small Computer System Interface, refers to a large protocol system, and has undergone evolution of SCSI-1/ SCSI-2/ SCSI-3 up till now.
- SCSI protocol defines the commands list and the system model for information communication within the architecture for different devices (disks, tapes, processors, optical devices, network devices etc.). This means that all these different devices can communication with each other using the SCSI protocol to work together in order complete tasks such as data read/write or data transmission.
- SCSI protocol in its essence is not related to the transmission media, and SCSI protocol can run on may different types of transmission medium or even virtual medium. For example, iSCSI protocol that is based on IP protocol or FCP (Fibre Channel Protocol) that is based on iSCSI protocol using underlying Fibre Channel connection. This means that SCSI protocol can run in different types of medium such as copper cables, optical cables or even in virtual connections in virtualization environments.



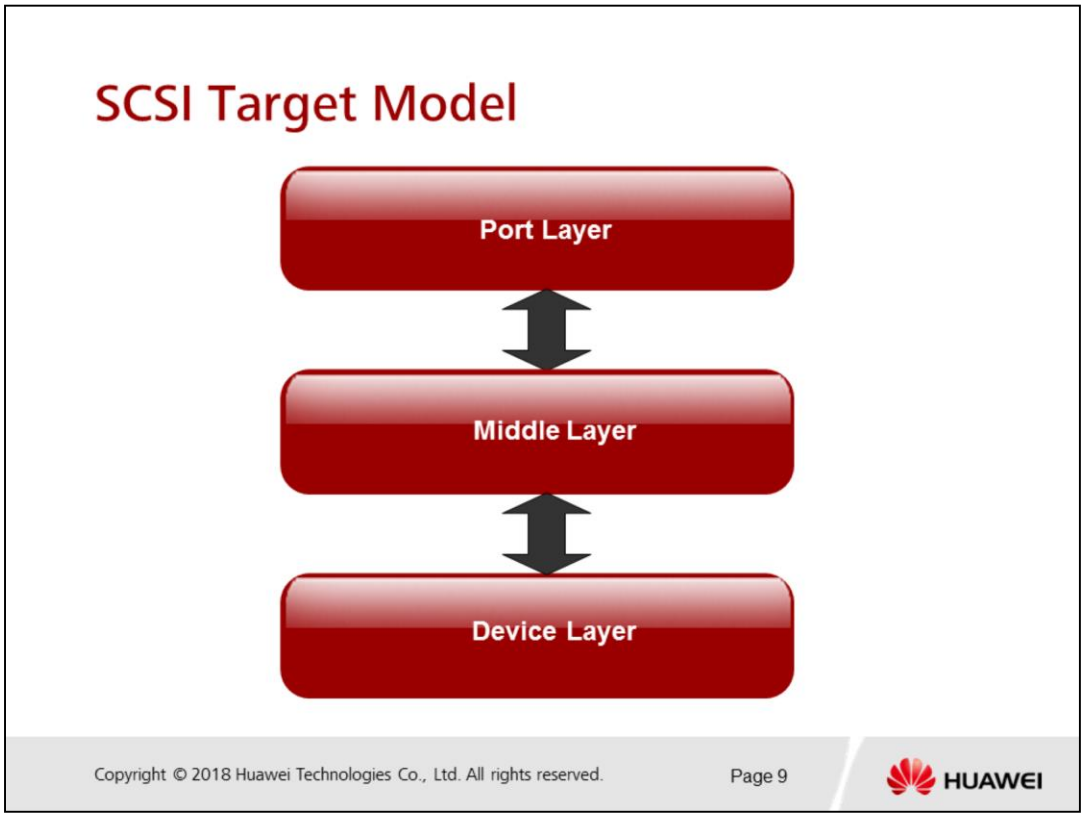
- The diagram above shows the history of SCSI technology ranging from its first version introduced in 1983 to the latest version introduced in 1993. The comparison above shows the changes that SCSI technology has gone through in order to keep up with the advancement of technologies and user requirements.



- LUN (Logical Unit Number): LUN refers to the logical storage resource connected to the target. An target can connect to multiple LUN and each LUN can have different properties. LUN#0 can be a disk drive, and LUN#1 can be other storage devices.
- Initiator: In reality, SCSI is a client/server(C/S) architecture, the clients are the initiator which is responsible to send commands to the SCSI target. Commonly, hosts plays the role of the initiators.
- Target: Target is the server end that processes the SCSI commands. It receives the commands from the host(initiator) and then analyze and processes those commands. An example of the target is the role of the storage array.
- SCSI initiator and target forms a standard C/S model, where each of the SCSI commands are run in “request/response” model.
    - Main functions of Initiator: Sends SCSI requests.
    - Main functions of Target: Respond to SCSI requests, and provide storage services through the LUN, and provide task management capabilities through the task processor.

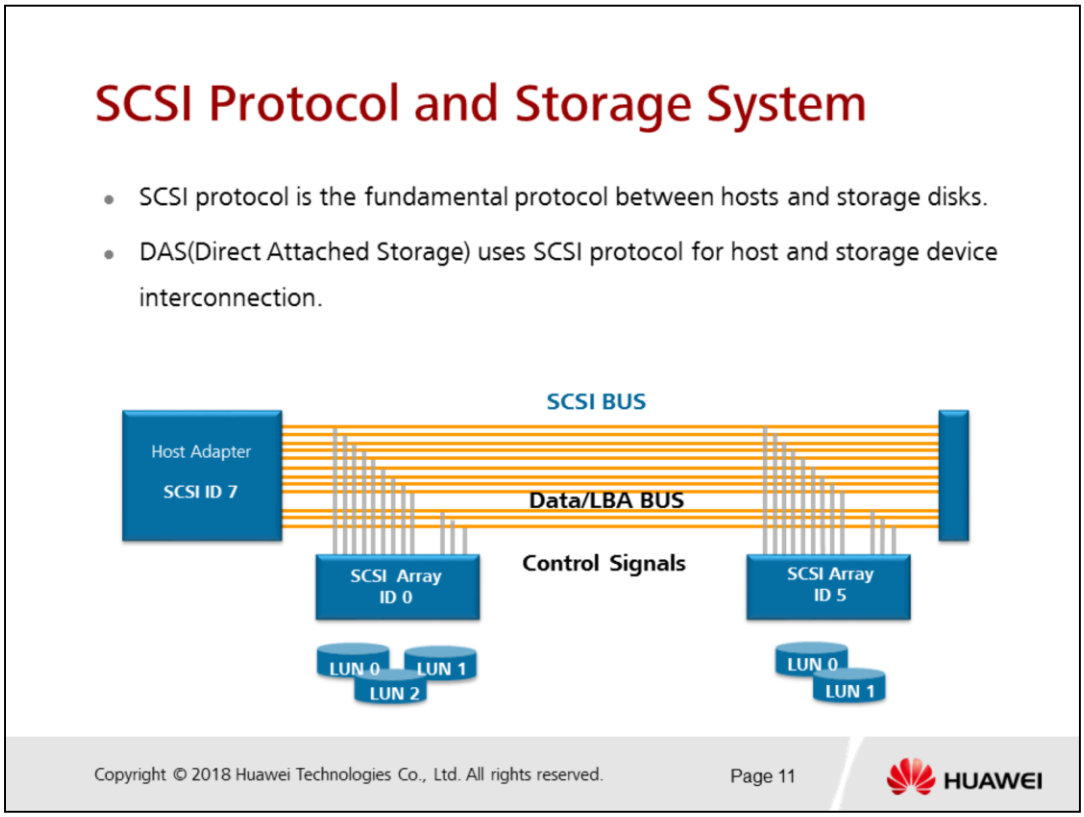


- The SCSI system in the host usually operates under the "Initiator" mode. From the perspective of SCSI system architecture, it is divided into "Middle Layer", "Devices Layer", and "Transmission Layer". Thus common operating systems including Windows, Linux, AIX, Solaris, BSD also divides SCSI into 3 layers.
- The initiator architecture under Windows: Windows separates SCSI initiator into 3 logical layers, in which SCSI Port is responsible for the basic SCSI processing such as device discovery, namespace scanning etc. To understand more details on the Windows drivers, you may refer to the DDK (Driver Development Kit) or WDM (Windows Driver Model) documentations.
- The initiator architecture under Linux: Linux also divides the SCSI initiator into 3 logical layers, in which scsi mod middleware layer handles the complex processing of process that doesn't involve SCSI devices and adapters such as abnormalities and namespace maintenance etc. On the other hand, HBA driver provides the encapsulation and decapsulation and transmission of SCSI commands. Device driver contains certain SCSI devices drivers such as some of the well known drivers for SD (SCSI Disk), ST (SCSI Tape) and SR (SCSI CDROM) etc.
- The initiator architecture under Solaris: The architecture for Solaris is similar to Linux/Windows, so we will not explain it further here.
- The initiator architecture under AIX: The AIX architecture is also divided into 3 layers which are SCSI Device Driver Layer, SCSI Middle Layer and SCSI Adapter Driver Layer.

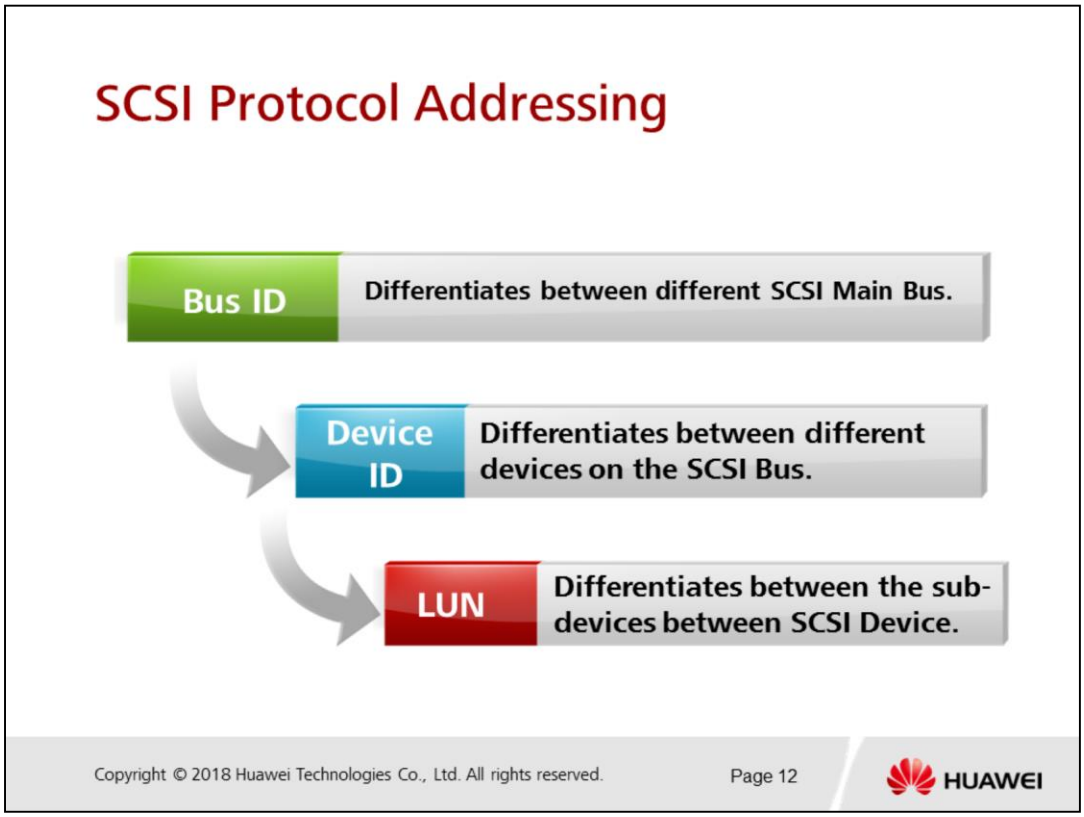


- Target is also divided into 3 layers which are Port Layer, Middle Layer and Device Layer based on SCSI system architecture. The most important layer is the Middle Layer which is designed using SAM(SCSI Architecture Model)/SPC (SCSI Primary Commands) specifications. The middle layer manages and maintains the LUN namespace, link ports, target device, tasks, commands and communication. Port layer drivers are loaded in registered mode whereas the device layer drivers are dynamically loaded.
- PORT Model in Target: PORT drivers are dynamically loaded, and the main functions of PORT is to complete the encapsulation and decapsulation of the SCSI commands carried over the link/connection. For example, encapsulating the SCSI commands into FPC or iSCSI or SAS packets, or decapsulate the FCP/iSCSI/SAS packets into SCSI commands. The target model drivers for iSCSI/FCP/SAS hardware belongs to the PORT category. The functions provided by the PORT includes transmitting responses (xmit_response), transmitting data (xfer_data), notification on completed management tasks, ending the processing task (cmd_done), port control (reset control) etc.

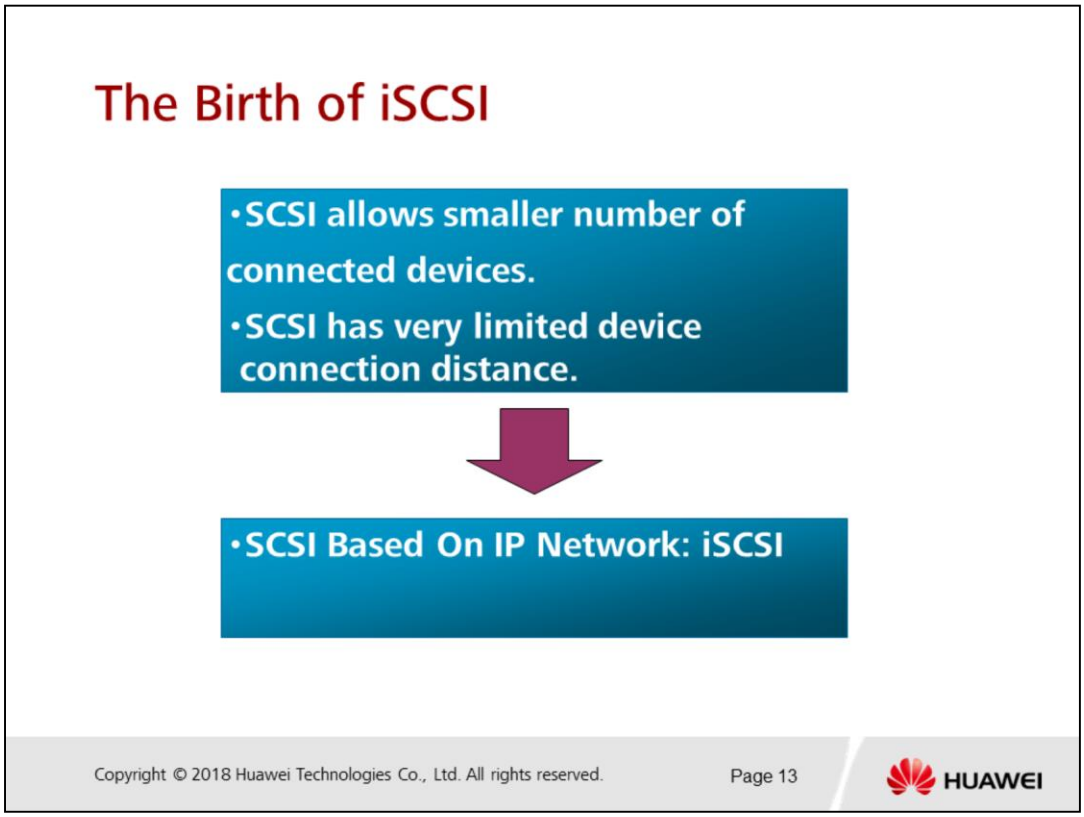
- Target Middle Layer: The middle layer provides maintenance for the models such as “LUN space”, “Tasks Sets”, “Task (Commands)” etc. The are two methods for LUN space maintenance, one is through the Global LUN which is viewable by all the Ports, and another method is to assign each port to maintain a certain LUN space. The first method is much easier to implement and for the following section we will be discussing on this method.
- Device Model in Target: In its essence, the device is an “analyzer” for SCSI commands, and by processing INQUIRY to notify the Initiator that the LUN resides on what device, then uses READ/WRITE commands to process all the I/O.



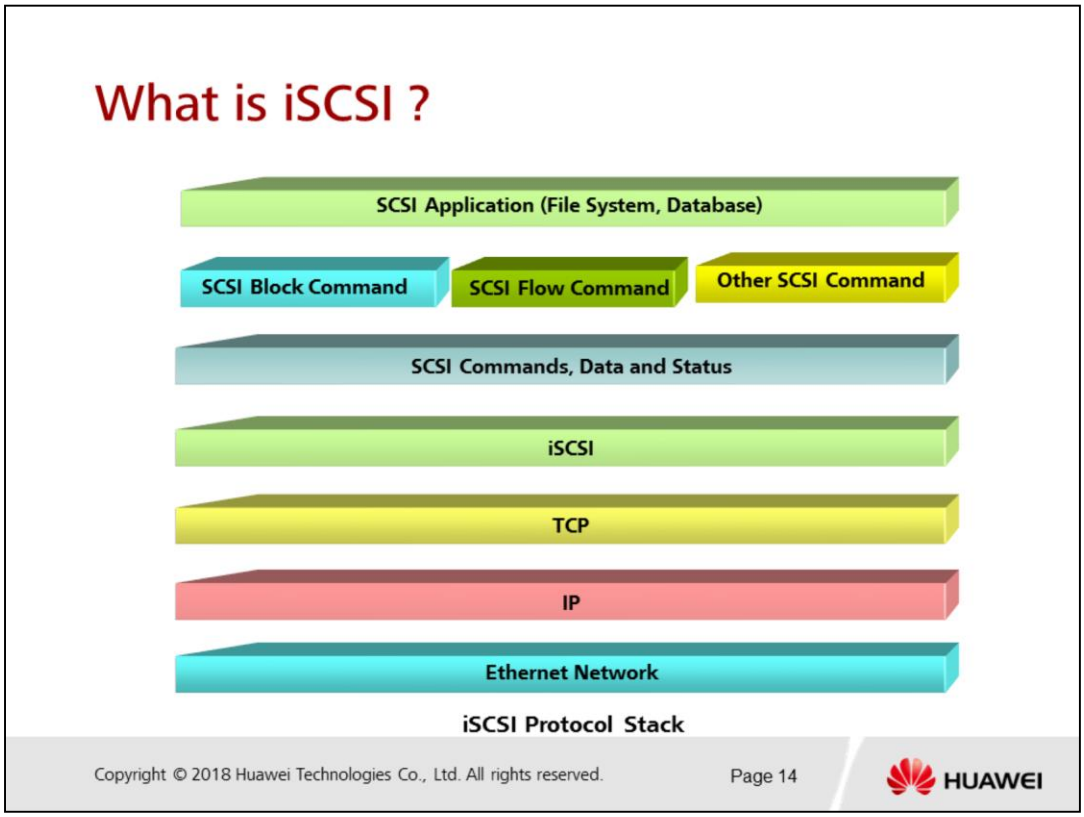
- The storage controller send a request signal to the Bus processor to use the Main Bus. When the request is accepted, the storage controller cache will begin to perform the send operation. In this process, the controller utilizes the main bus, and the other devices cannot use the main bus during that period. However, if the main bus has the Interrupt function, the main bus controller can stop the sending or transmission process at any time and pass the main bus control to other devices for a much higher priority operation.
- SCSI controller is similar to a small CPU, and has its own command sets and cache. SCSI is a special bus architecture, and can perform dynamic task allocation and task operation for multiple devices in the computer at the same time. This means that it can dynamically allocate and dynamically complete multiple tasks required by the system at the same time.



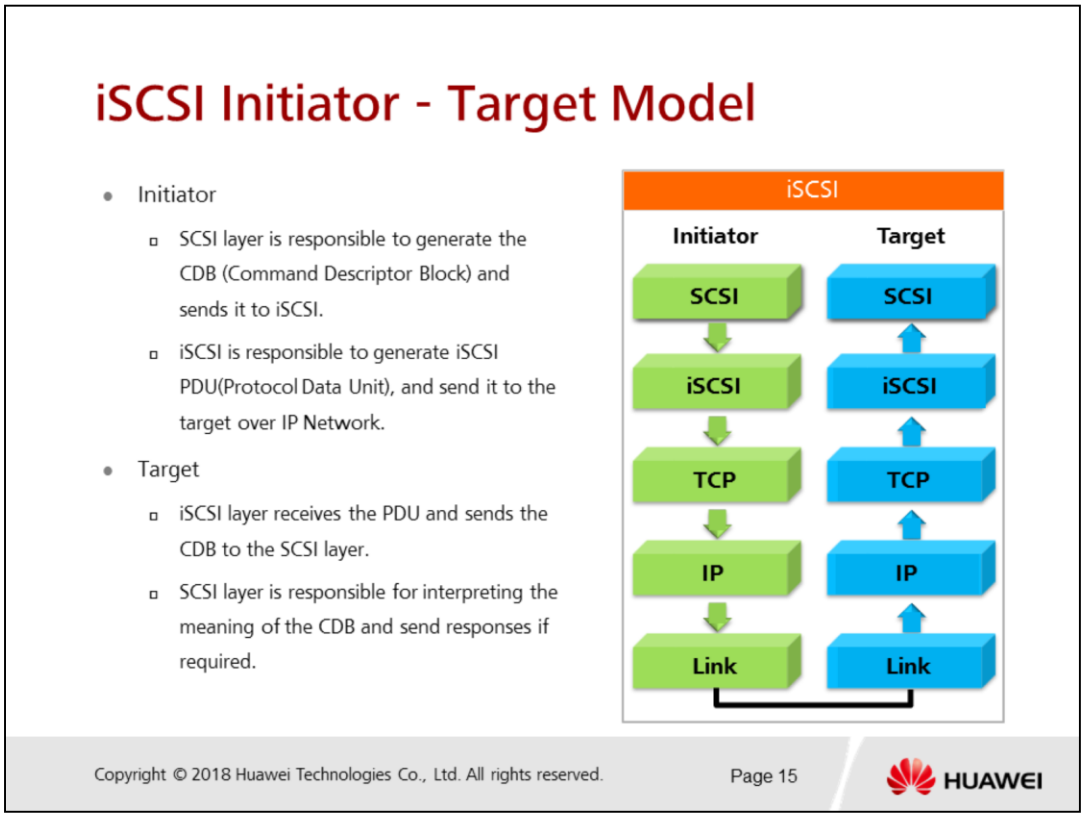
- In order to address the devices connected to the SCSI main bus, SCSI protocol introduced SCSI Device ID and LUN(Logical Unit Number). In the SCSI main bus, each device must have a unique device ID, and the HBA(Host Bus Adapter) in the servers also have their own device ID which is fixed at 7 by default. Each bus allows a maximum of 8 or 16 device IDs including the HBA. Device ID is not only used for addressing, but also used to identify the device priority on the bus. Additionally, all the different devices connected to the main bus must have different device ID to avoid addressing and priority conflicts.
- Each storage device includes multiple sub devices such as disks, tape drives etc. Hence, SCSI protocol introduced the concept of LUN ID for the addressing of all these sub devices within the storage devices.
- Traditional SCSI controller connects to a single bus, and only has a single bus ID. An enterprise level server may be configured with multiple SCSI controller, and thus have multiple SCSI bus. When SAN is introduced, each FC HBA or iSCSI network cards are connected to a bus, hence each bus needs to be assigned with a bus ID for differentiation between different buses. We can use a three element description to identify or label a SCSI target in the form of : Bus ID/Target Device ID/LUN ID.



- iSCSI protocol was initially introduced by IBM, CISCO and HP. It became the IETF (Internet Engineering Taskforce) standard by 2004, and the current iSCSI protocols are based on SAM2(SCSI Architecture Model 2).
- iSCSI (Internet SCSI) encapsulates the SCSI commands and data blocks into TCP packets and transmit them over the IP network.
- iSCSI became the SCSI transmission level protocol, and its initial starting point is to use mature IP network technologies to implement and extend Storage Area Network (SAN) capabilities.

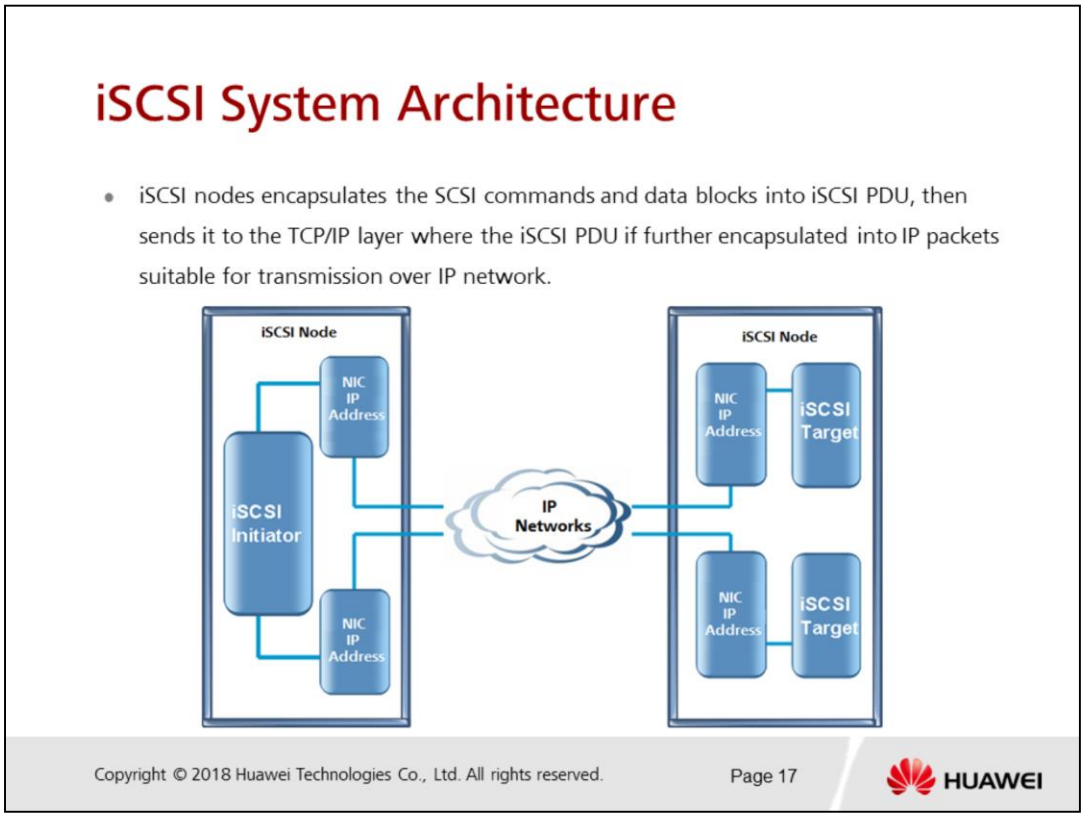


- iSCSI is the abbreviation of Internet Small Computer System Interface, and it is a standard in data blocks transmission over TCP/IP. It can be considered as SCSI over IP.
- iSCSI can be used to build IP based SAN to provide high speed, low cost, long distance storage solutions.
- iSCSI encapsulates the SCSI commands into TCP/IP packets which allows I/O data blocks to be transmitted over IP network.
- iSCSI is the mapping of SCSI remote procedure call to the TCP/IP protocol. SCSI layer is responsible to generate the CDB (Command Descriptor Block) and sends it to the iSCSI protocol later where it is further encapsulated into PDU(Protocol Data Unit) and transmitted over the IP network.



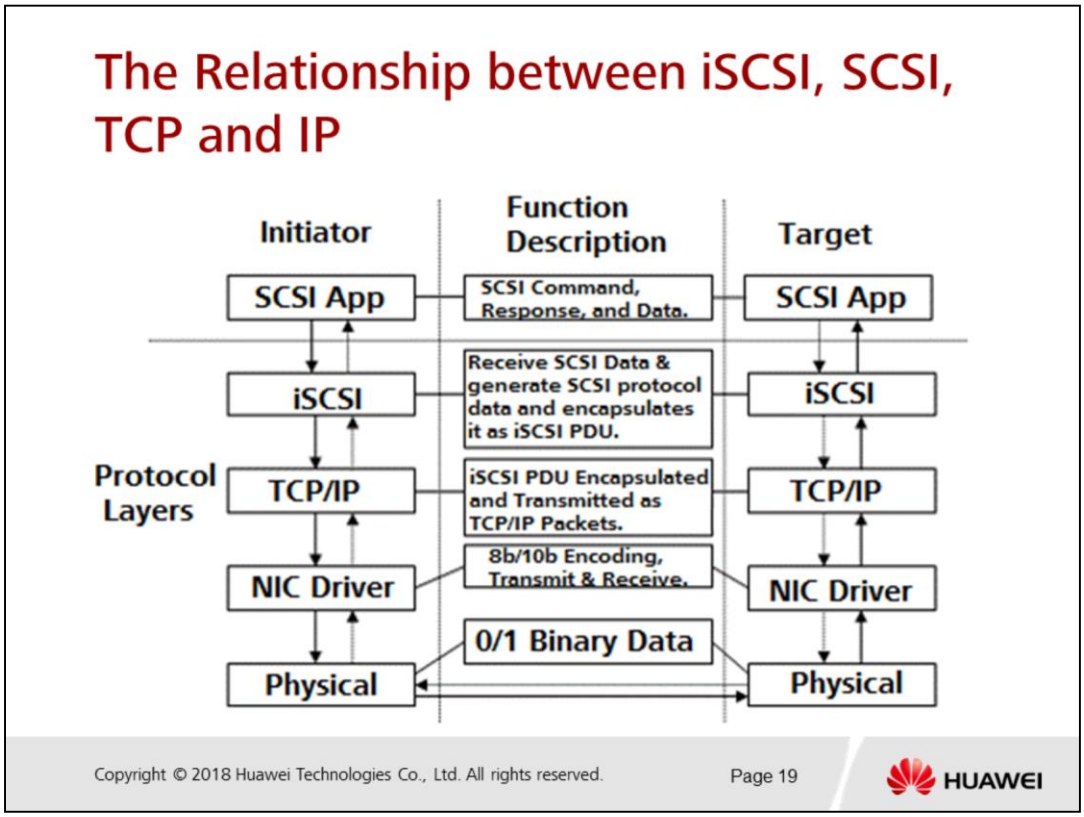
- iSCSI communication system still inherits certain characteristics of SCSI. In iSCSI communication, there is a Initiator device that sends the I/O request, and a Target device that responds and execute the actual I/O operations based on the I/O requests received. Once the Initiator and Target establish a connection, the Target will be the main device that controls the whole working process.
- iSCSI Initiator: It can be divided into 3 types, which are software Initiator driven program, hardware TOE (TCP Offload Engine) card and the iSCSI HBA card. Judging from performance, the software initiator has the lowest performance, TOE card has medium performance and iSCSI HBA cards has the best performance.
- iSCSI Target: The target are usually the iSCSI disk arrays or iSCSI tape libraries.
- iSCSI protocol has defined a set of naming and addressing method for Initiator and Target. All the iSCSI nodes are identified by iSCSI names. This naming convention avoids iSCSI names and host names to be confused with each other.

- iSCSI uses iSCSI Name to differentiate the initiator and target device. The address will change following a change in location for the initiator and target device, but the name will remain as the same. When a connection is initiated, the initiator device sends a request, and once the target receives the request, it will check whether the iSCSI name in the request is the same as the target iSCSI name. If the names match, then the connections will be successfully established. Each iSCSI node only allows 1 iSCSI Name, and 1 iSCSI name is used to establish connections from a single initiator to multiple target device. On the other hand, multiple iSCSI name can also be used to establish connections from single target device to multiple initiator device. Hence, the relationships between them can be summarized as 1 Initiator → Multiple Target , and 1 Target → Multiple Initiator.



- In a system that supports iSCSI, the users sends out instructions to send or fetch data from a SCSI storage device, the operating system processes that request and converts it into one or many SCSI commands which are then send to the target SCSI controller card. iSCSI node will do the encapsulation process on the SCSI commands and data blocks which forms a iSCSI PDU, which is sent to the TCP/IP layer. The TCP/IP layer will encapsulate the iSCSI PDU into the IP packets suitable for transmission over the IP network. It can also do secure encryption on the encapsulated SCSI commands, then send it over unsecure networks.
- Packets can be transmitted over the local area network or Internet. On the receiving storage controller, the data will be recombined and the storage controller will interpret the SCSI commands and data contained within those packets and send them to the corresponding disk drives. The disk drives will then execute the functions initially requested by the host or application. If the received request is for data retrieval, then the data will be fetch from the disks, encapsulated and transmitted to the requesting computer. This whole process is transparent to the user.
- Although the SCSI commands execution and data preparation can be completed by standard TCP/IP or current network card software, there are some caveats where if the encapsulation and decapsulation of the packets process are done by software, it will take a toll on the CPU and impact performance. This is due to the fact that host CPU will require a lot of CPU computation cycles to process all the data and the SCSI commands.

- If these command and data processing tasks are designated to a dedicated device, it can minimize the impact to the system to the lowest extent. Hence, it is crucial to develop dedicated iSCSI devices that can execute the SCSI commands and complete the data preparation process under standard iSCSI specifications.
- iSCSI HBA is a device that combines the functions of NIC and HBA. This iSCSI adapter fetches data using block access, uses TCP/IP processing engine on the card itself to complete the data splitting and processing, and then uses the IP network to transmit those IP packets. These functions of the iSCSI device allows you to build an IP based SAN that on the basis that it does not lowers the server performance by offloading those processing and data preparation tasks to the iSCSI adapters.



- PDU: Protocol Data Units. A protocol data unit is information delivered as a unit among peer entities of networks containing control information, address information or data. In layered systems, PDU represents a unit of data specified in the protocol of a given layer, which consists of protocol control information and user data. PDU is a significant term related to the initial four layers of the OSI model. In Layer 1, PDU is a bit, in Layer 2 it is a frame, in Layer 3 it is a packet and in Layer 4 it is a segment. In Layer 5 and above, PDU is referred to as data.
- TCP/IP: Transmission Control Protocol/ Internet Protocol. In simple words, it is the language a computer uses to access the internet. It consists of a suite of protocols designed to establish a network of networks to provide a host with access to the internet. TCP/IP is responsible for full-fledged data connectivity and transmitting the data end to end by providing other functions, including addressing, mapping and acknowledgment.
- 8b/10b encoding takes a group of continuous 8 bit of data and then splits it into 2 sets of data, the first set with 3 bits and the second set of 5 bits of data. After encoding, it forms a set of 10 bit transmission character with a group of 4bit character and another group of 6 bits of character, which are transmitted over the network.
- 8b/10b encoding: It is an algorithm for encoding data for transmission in which each eight-bit data byte is converted to a 10-bit transmission character. Invented and patented by IBM Corporation. 8b/10b encoding is used in transmitting data on Fibre Channel, ESCON, and Gigabit Ethernet. 8b/10b encoding supports continuous transmission with a balanced number of ones and zeroes in the code stream and detects single bit transmission errors.

Contents

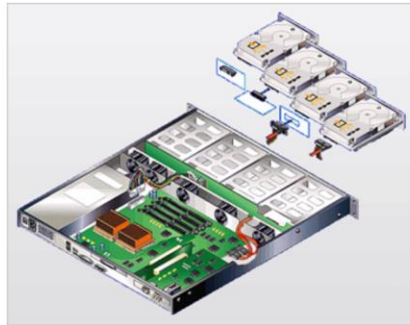
1. SCSI/iSCSI
2. **SAS**
3. FC/FCOE
4. PCIe
5. IB
6. CIFS/NFS
7. FTP/HTTP

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

HUAWEI

SAS in Storage System

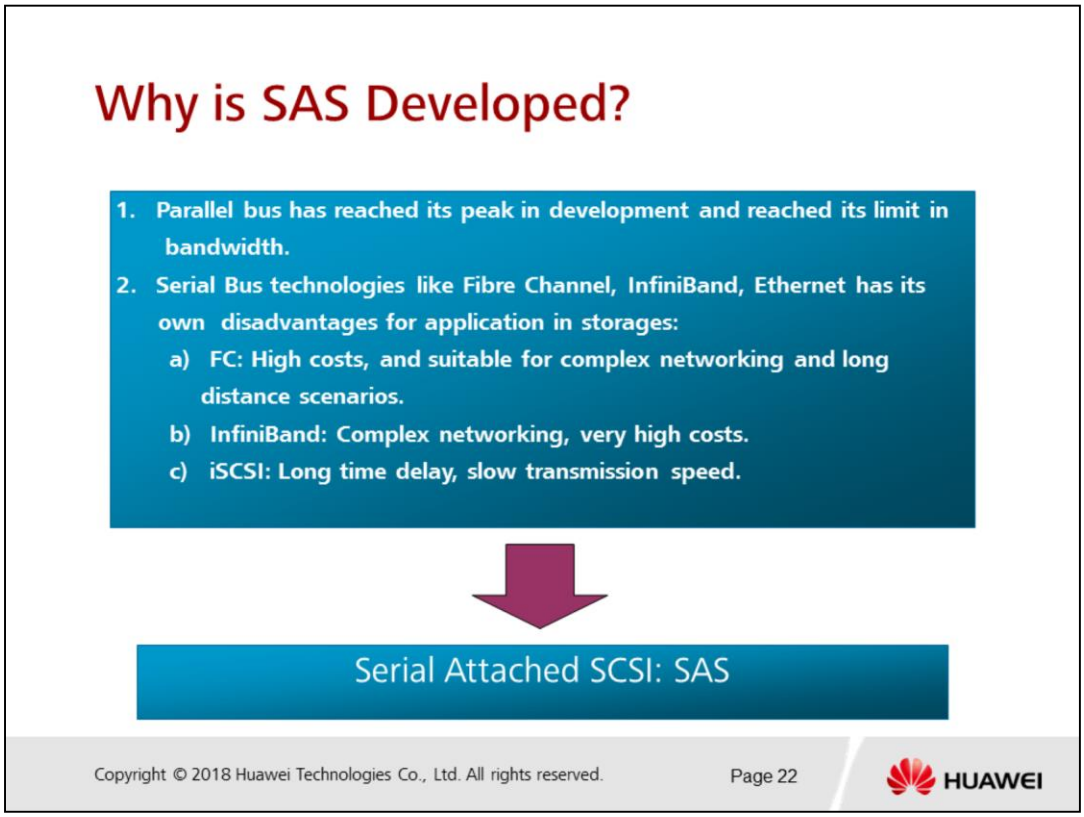
- In enterprise storage systems, SAS (Serial Attached SCSI) interface has already replaced SCSI and SATA interface.
- SAS uses point to point architecture, and its performance can go up to 300MB/s, 600MB/s or higher.



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

HUAWEI

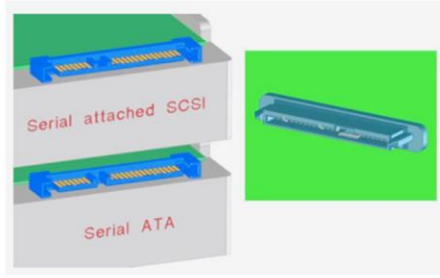
- SAS uses a point-to-point connection design to enable a dedicated link to be established between two devices in the communication. The multi point bus design in Parallel SCSI (often called as SCSI) allows multiple devices share the same bus, which means that the more devices are connected, the performance gets worse. With point-to-point connections in SAS, the communication speed is much faster because it is not necessary to detect whether the two devices communicating are allowed to use the connection link before communication. Each device is connected to the specified data path to increase the bandwidth.
- The serial interface has a simple structure, supports hot swap, has high transmission speed, and high operational efficiency. Commonly, large parallel cables can cause electronic interference but the SAS cable structure can solve this problem. The SAS cable structure saves space and improves the heat dissipation and ventilation capabilities of SAS hard disk based servers.



- SAS is the abbreviation of Serial Attached SCSI, which is a serial connection form of SCSI.
- SAS is the serial standard for SCSI bus protocol.

What is SAS?

- SAS: Serial Attached SCSI, is the serial standard for SCSI bus protocol.
- SAS uses serial technology to obtain higher transmission rate and better expansion capability, and it is compatible with SATA hard disks.
- SAS current transmission rates are up to 3Gbps, 6Gbps, 12Gbps or higher, and supports full-duplex mode connection.

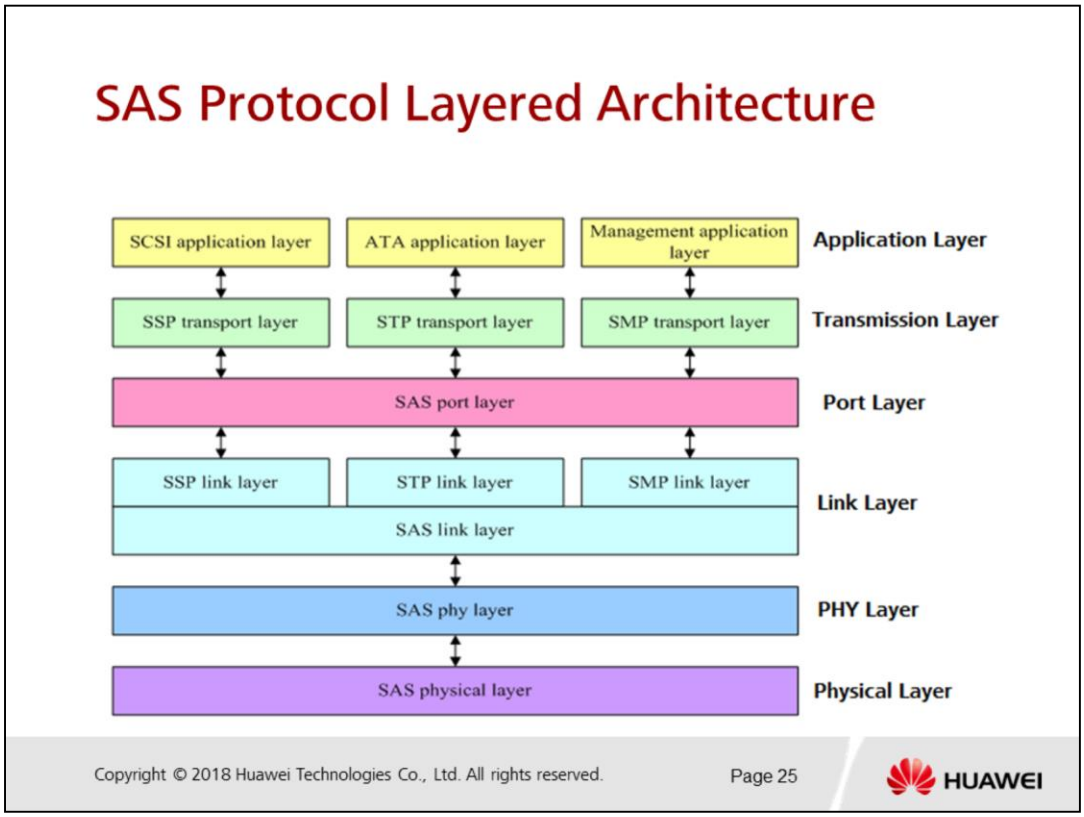


Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

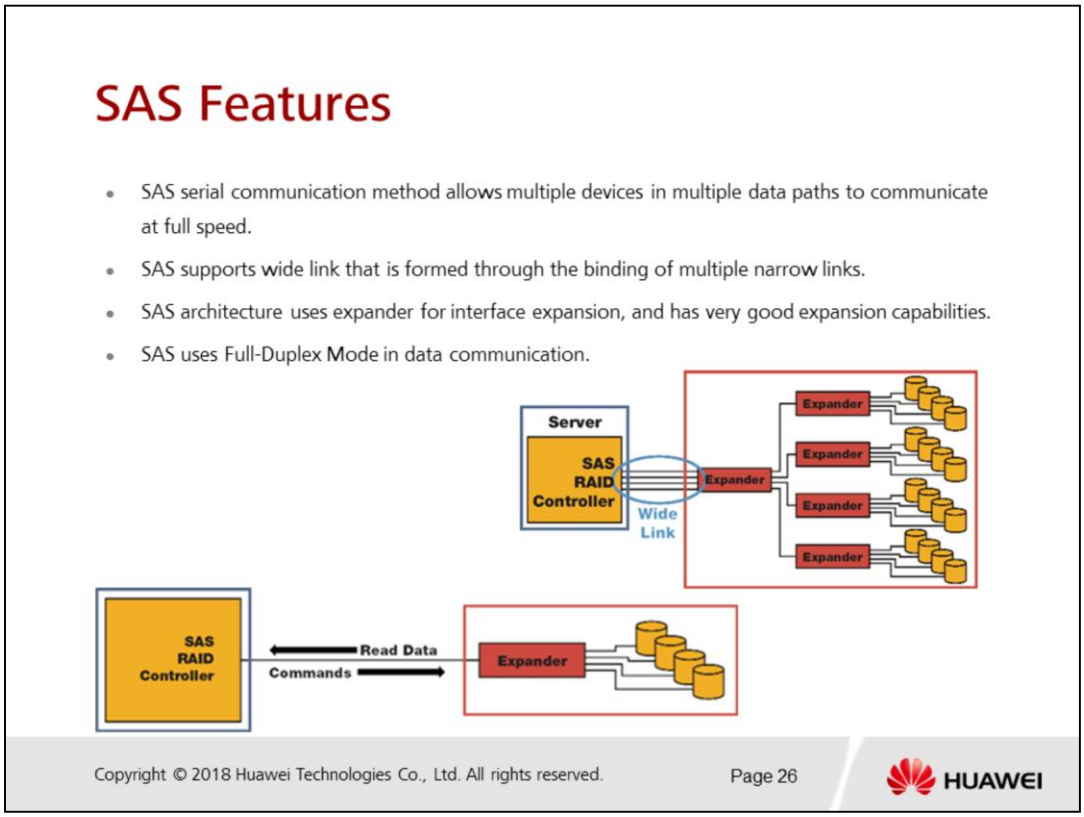
HUAWEI

- Lower Costs:
  - SAS backplane is compatible with both SAS and SATA hard disks, which means that there is no reinvestment required when using different types of hard disks and compatible with the existing SAS and SATA hard disks within the enterprise.
  - Lowers the cost of designing different products for SCSI and SATA standards, and at the same time lowers the complexity in cabling and layers in PCB(Printed Circuit Board) which saves cost in manufacturing.
  - System Integrators do not need to purchase different backplanes and cables to configure the different types of hard disks in the client environment.
- Connects More Devices:
  - SAS technology introduces SAS expanders that allow SAS systems to connect more devices, each of these allows multiple ports to be connected, and each port can connect to SAS devices, hosts, or other SAS expanders.

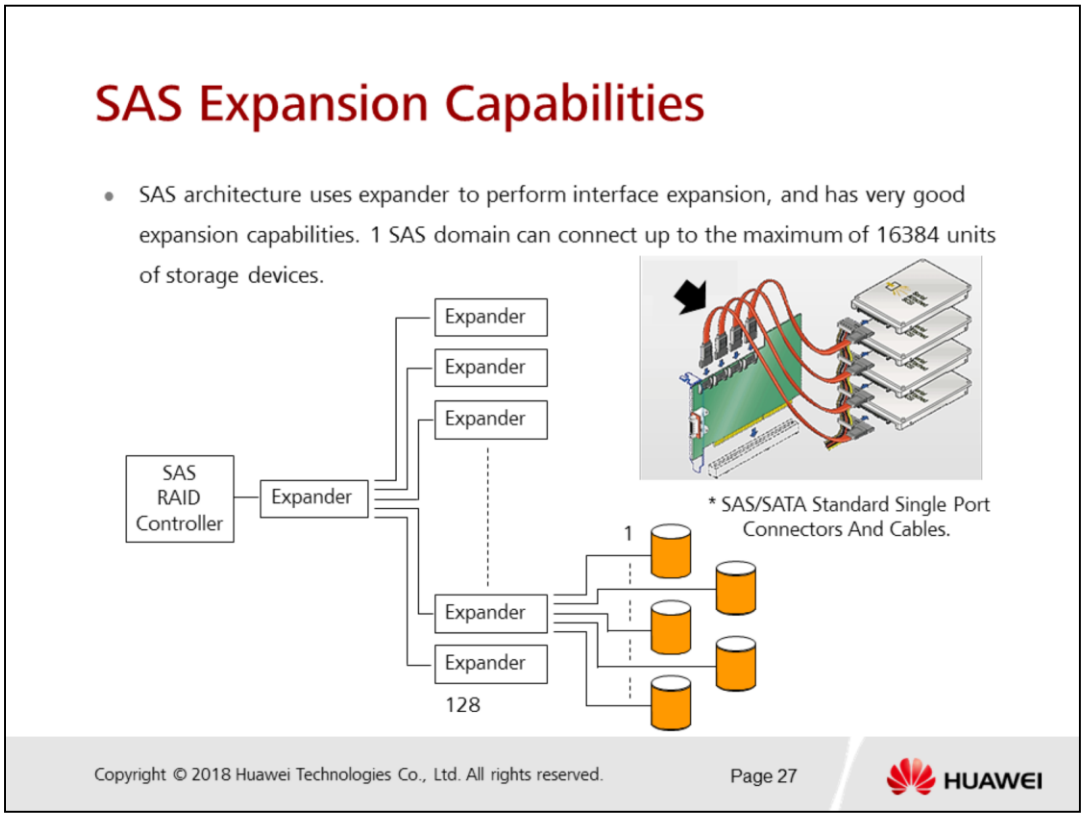
- High Reliability
  - The device reliability is the same as SCSI, FC disks and better than SATA disks.
  - Retains the verified SCSI commands.
- High Performance
  - High Unidirectional Port Rate
- Compatible With SATA
  - SATA drives can be directly installed in SAS environments.
  - Can use both SATA or SAS drives within the same system, which fits the currently popular storage tier policies in the storage systems.



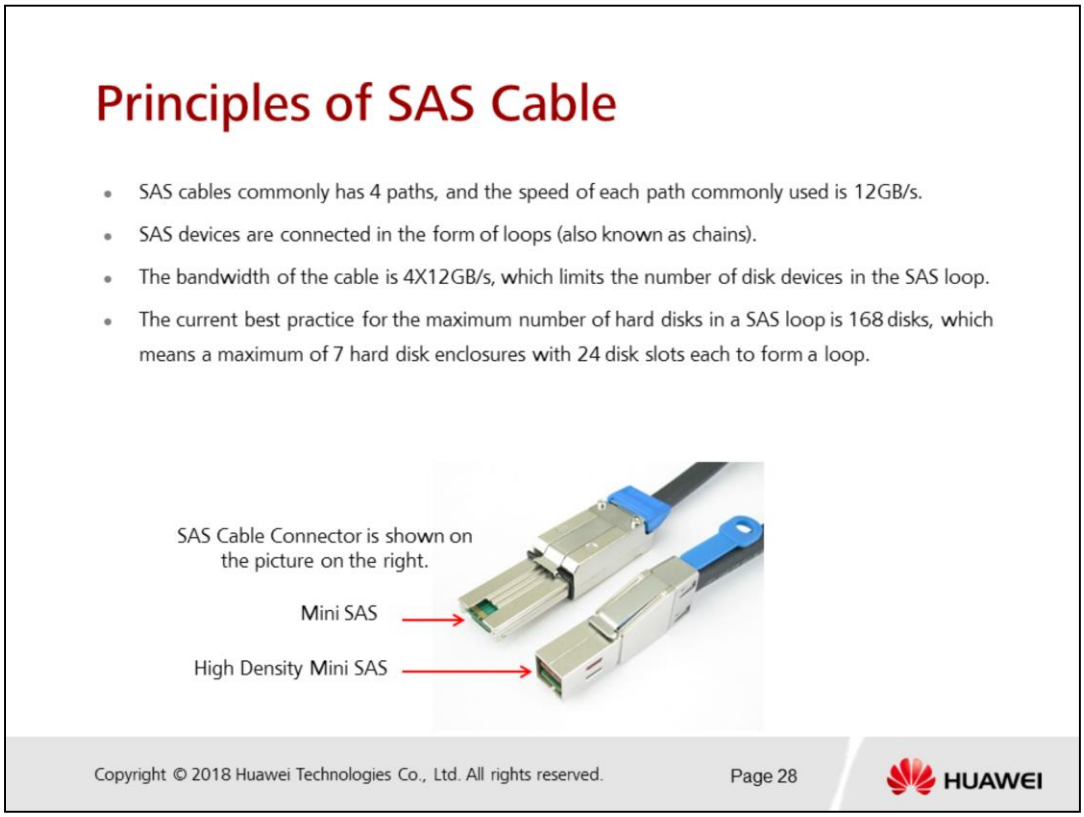
- SAS standard has divided the SAS architecture into 6 layers, which are in the order of lowest layer to the highest layer : physical layer, sas phy layer, link layer, port layer, transmission layer and application layer. Each layer is responsible for certain functions:
  - Physical Layer: Defines the hardware such as cables, interfaces, and transmitters.
  - SAS PHY Layer: Includes the low level protocols, data encoding (8b10b), signaling (Out of Band, OOB) and connection reset sequences.
  - Link Layer: Describes how to control phy layer connection management, as well as primitives, CRC checksum and descrambling, and transmission rate matching & processing.
  - Port Layer: Describes the interface between the link layer and the transmission layer, including how to request, break and building the connection.
  - Transmission Layer: Defines how to transmit all the commands, status and data encapsulation in the SAS frames and also how to perform decapsulation of the SAS frames.
  - Application Layer: Describes the details on how to uses SAS for different types of applications.



- Compared to SCSI, the main improvements of SAS are:
  - Provides higher throughput due to the usage of serial communication, and has possibility for higher performance in the future.
  - 4 narrow links can be bonded together to form a Wide Link and provide higher throughput.
- SAS uses a full duplex (two-way) communication mode instead of one-way communication. Traditional parallel SCSI can only communicate in one direction. In parallel SCSI scenario, when the device receives a packet, if the device wants to respond to the packet, it needs to reestablish a new SCSI communication link after the previous link is disconnected. With SAS, two-way communication is possible. Each SAS cable has 4 cables which is used for 2 inputs and 2 outputs. Thus, SAS can read and write data at the same time, and full-duplex data operations improve data throughput efficiency.



- SAS Expander: An interconnection device in the SAS domain similar to Ethernet switches. Through the cascading of Expanders, it can greatly increase the number of connected terminal devices, in which HBA(Host Bus Adapter) costs can be saved. Each expander can connect up to a maximum of 128 terminal devices or 128 expanders. A SAS domain consists of the following components which are: SAS Expanders, Terminal Devices and Connection Devices (SAS connection cables).
    - SAS Expander is equipped with address routing table for tracking and recording of the addresses of all the SAS drives connected.
    - Terminal Devices includes initiators (commonly known as SAS HBA cards) and targets (SAS/SATA disks or HBA cards in the Target mode).
- Loops cannot be formed with the expanders within the SAS domain to ensure the normal operation of the device discovery process.
- In actual use, due to the bandwidth factor, the connected terminal devices in the expanders are often lesser than 128 devices.
-



- Most current storage device vendors use SAS cables to connect the disk enclosures to the controller enclosures or use it for interconnection between disk enclosures. SAS cables usually bundle four separate channels (narrow ports) into one wide port to provide more bandwidth. Each of the 4 independent channels can operate at 12 Gb/s, so the entire wide port can provide 48 Gb/s of bandwidth. To ensure that the amount of data on the SAS cable does not exceed the maximum bandwidth of the entire SAS cable, we need to limit the total number of hard disks connected to a SAS loop.
- For Huawei devices, the maximum number of hard disks in a SAS loop is 168 disks, which means that up to seven disk enclosures with 24 disk slots can form a SAS loop. However, the prerequisites condition is that all the hard disks in the loop are traditional SAS hard disks. Now that SSDs are more commonly used, we must realize that SSDs are much faster than SAS disks in terms of speed. Therefore, for the SSD disks, the best practice for the maximum number disks in a loop is 96 disks, which is a loop consisting of 4 disk enclosures with 24 disk slots.
- The SAS cable interface is called a Mini SAS cable when the SAS cable is a single-channel 6Gb/s cable type. Now that the single-channel speed is increased to 12Gb/s, the corresponding SAS cable is now called a high-density Mini SAS cable.

Comparison Between SAS And Other Transmission Technologies

Types of Technology	Main Advantages	Main Disadvantages	Fields Of Application
ATA	Low Cost	Low Performance	PC
SCSI	Medium Performance	Drawbacks of Parallel Technologies	Enterprise Grade Storages
FC	High Performance, High Reliability	High Cost	High End Storages
SATA	Low Cost, High Capacity	Low Performance and Low Reliability	Mid and Low End Storages
SAS	High Performance, High Reliability	Medium Cost	Mid and Low End Storages

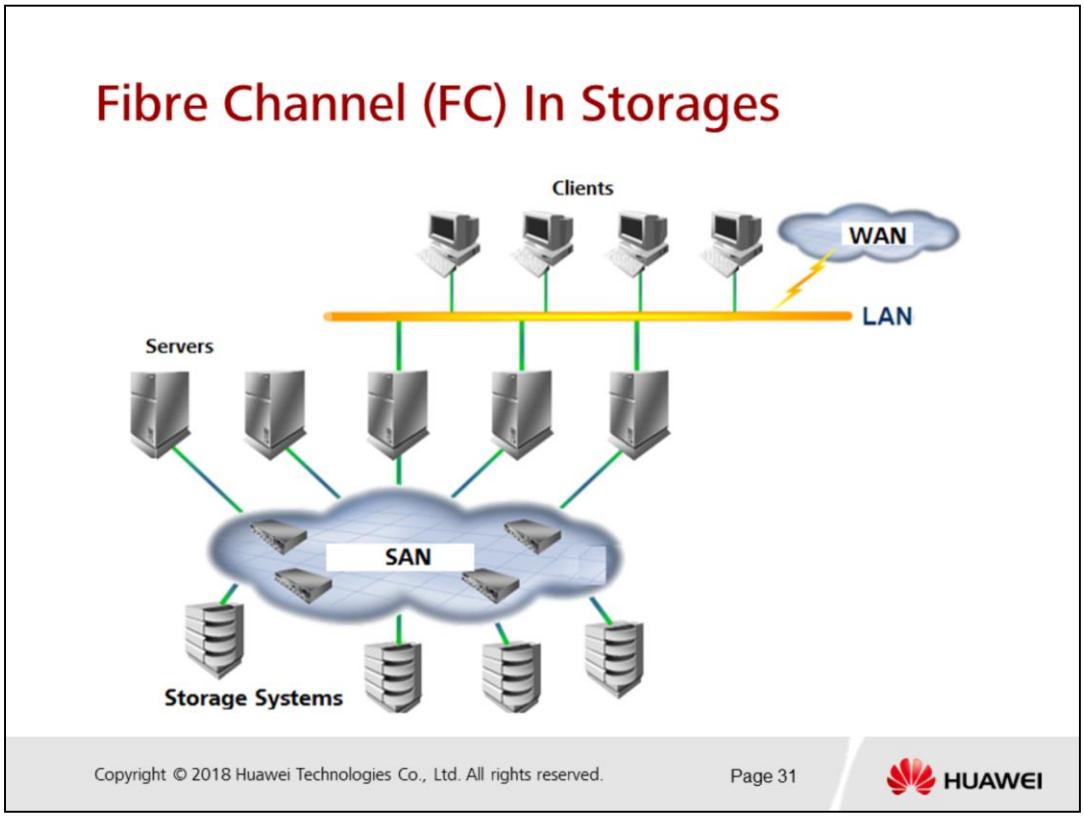
Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

- The table above shows the comparison different types of transmission technology with SAS with the main advantages and disadvantages highlighted.
- SAS has high performance and reliability and comes with a medium cost range which is suitable for mid tier and low tier storage systems.
- Although FC has the highest performance, but the higher costs of implementation makes it not the cost effective choice for small and medium enterprises.

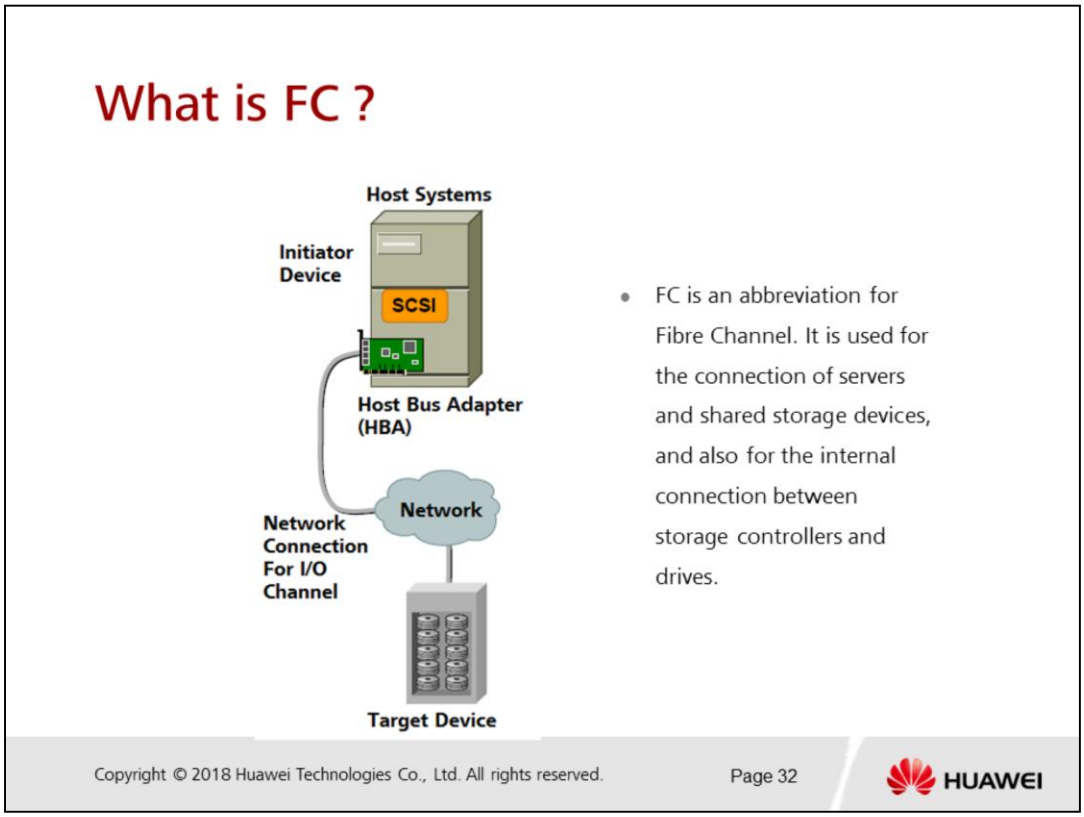
Contents

1. SCSI/iSCSI
2. SAS
3. **FC/FCOE**
4. PCIe
5. IB
6. CIFS/NFS
7. FTP/HTTP

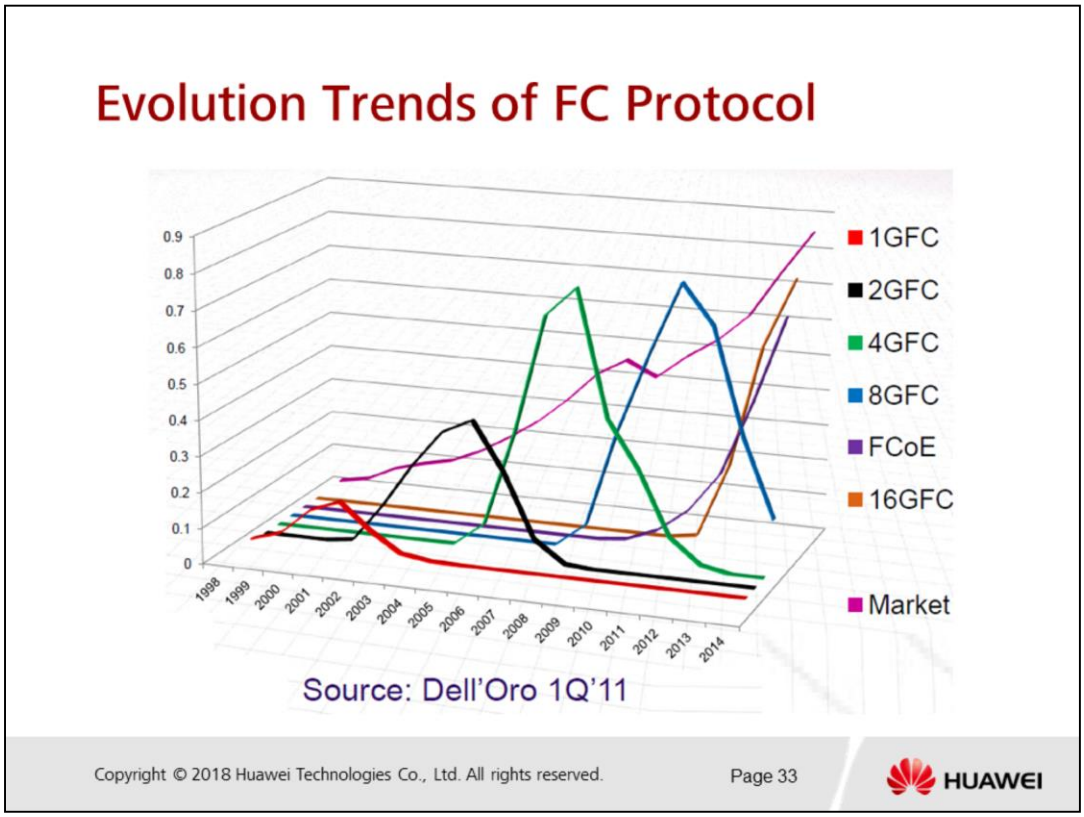
Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.



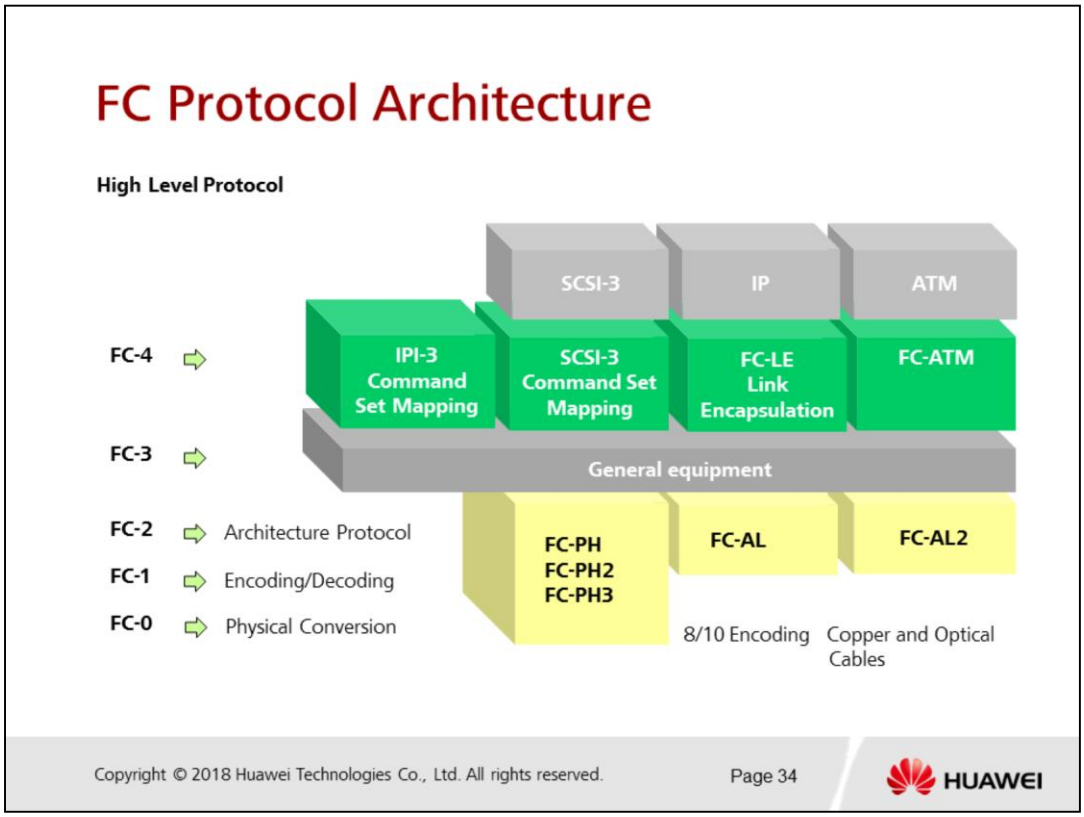
- Fabric is a term used to refer to an intelligent network composed of intelligent Fibre Channel switches with good system design. This network can provide enterprise-level performance, scalability, manageability, reliability, and availability.



- FC is a standard for high-performance serial connection. Its interface transmission rate is 4Gbps, 8Gbps, 16Gbps or higher. FC transmission medium can be chosen from the option of copper cable or optical fiber. It has long transmission distance, and supports multiple types of network topologies.
- The definition of Fibre Channel: FC refers to the meaning of "channel in the form of fibre", and it is often misunderstood with Fiber Channel due to the similarities of the word Fibre and Fiber. Fiber Channel can be understood as "optical channel". Both refers to different types technological context. Compared to TCP/IP, FC protocol stack contains a lot of similar concepts. For example FC switches, FC routers, SPF(shortest-path-first) algorithm etc. We can consider FC protocol as the TCP/IP protocol within SAN because both of them follow the OSI model in their architecture design.
- FC protocol cannot be considered as Fiber protocol because the FC connection medium can be optical fiber, twisted pair cables or coaxial cable. Due to FC protocols commonly adopt optical fiber as the transmission medium, thus a lot of people tend to consider FC as fiber channel protocol instead of the actual Fibre Channel Protocol.
- FC Protocol Advantages: It has high bandwidth, high reliability, high stability, low latency, and resistance to electromagnetic interference (EMI)etc. It also can provide very stable and reliable fiber connections, its easy to build large-scale data transmission and communication networks, and currently supports 1x, 2x, 4x and 8x bandwidth connection rate. As with the continuous development of technology, the bandwidth for FC is still expanding to meet the technical performance requirements of higher bandwidth data transmission.

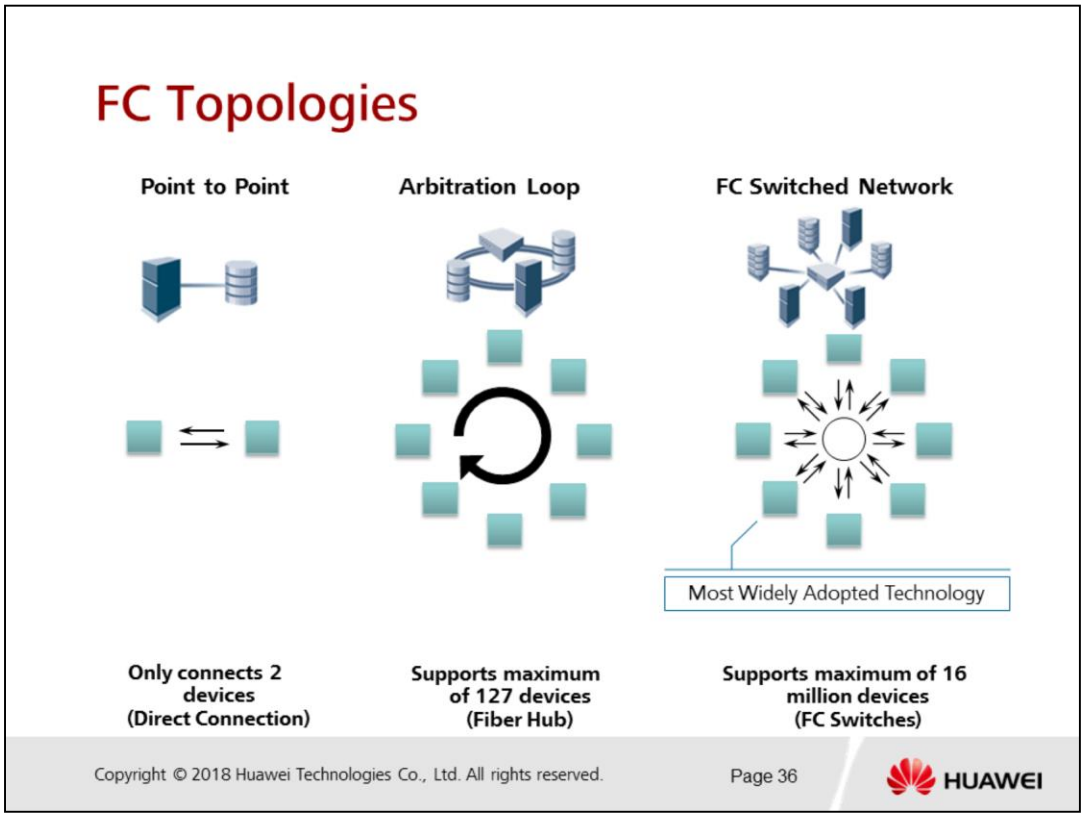


- FC originated in the year 1988, during that time, FC's main purpose was to simplify the connection and increase the data transmission distance instead of increasing the data transmission rate. As time progresses, it is used to increase the transmission bandwidth of hard disk protocols, with more inclination towards high speed, high efficiency and reliable data transmission. FC SAN started to gain wide application and adoption starting from the end of 1990s.
- Future Analysis:
  - From the perspective of trend analysis of research institutions, 16G FC has become an inevitable trend, and with the advent of PCIe3.0 and the increase in the density of virtual machines, higher I/O throughput is required, thus 32G FC will also gain more traction and has very promising future.
  - Standards update and commercialization pattern: The FC standard will be updated once every 36 to 48 months. After the new standard comes out, it will be commercially available within 12 to 24 months. Subsequently, the mainstream manufacturers will launch relevant products within 2 years. The standard life cycle of each generation is about 6 years, but from the point of view of 8G and 16G FC, the life cycle is obviously shortened, and the replacement cycle is obviously accelerated. Thus, FC technology has become an inevitable trend in the storage and SAN industry.



- The main part of the Fibre Channel protocol architecture is actually FC-2. Among the layers, FC-0 to FC-2 is called FC-PH, which is also called "physical layer". Fibre Channel mainly transmit data through FC-2 layer. Therefore, Fibre Channel is also often referred to as "Layer 2 Protocol" which is similar to the Ethernet Protocol.
- Fibre Channel data units are called frames. Even though the fibre channel itself consists of five layers, but most of the time Fibre Channel is referred as a Layer 2 protocol. A Fibre Channel frame has a maximum of 2148 bytes, and the head of the Fibre Channel frame is different from the Ethernet packet. FC uses only one frame format and performs various tasks on multiple layers.
- The function of the frame determines its format. The Fibre Channel frame starts with the start of frame (SOF) flag, followed by the frame header. We will discuss more details on the frame header later. Then it is followed by the data, or Fibre Channel content, and finally the end of frame (EOF) flag. In simpler words, the frame structure is SOF + Data + EOF. The purpose of this encapsulation is to allow Fibre Channel frames to be carried by other protocols like TCP when its needed.

- The relationship between Fibre Channel and SCSI: Fibre Channel is not a substitute for SCSI. In fact, Fibre Channel will use frames to transfer SCSI instructions and status information. In relevance, SCSI is the upper layer protocol in the FC4 protocol stack, and SCSI is a subset of the FC protocol.
- When transferring large amounts of data, there will be a large number of frames that need to be sent. When a group of frames is sent as a batch, we call it an exchange.
- The FC Protocol is mainly divided into 5 layer which has different functions as the following:
    - FC-4 – Protocol-mapping layer, in which upper level protocols such as SCSI, IP or FICON, are encapsulated into Information Units (IUs) for delivery to FC-2. Current FC-4s include FCP-4, FC-SB-5, and FC-NVMe.
    - FC-3 – Common services layer, a thin layer that could eventually implement functions like encryption or RAID redundancy algorithms; multiport connections;
    - FC-2 – Signaling Protocol, defined by the Fibre Channel Framing and Signaling 4 (FC-FS-4) standard, consists of the low level Fibre Channel protocols; port to port connections;
    - FC-1 – Transmission Protocol, which implements line coding of signals;
    - FC-0 – PHY, includes cabling, connectors etc.;



- FC protocol used a long period of time since its initial introduction in 1988 to become a mature technology. Nowadays, SAN has a few methods to interconnect the physical components:
  - Point to Point: 2 devices are directly connected to each other. This is the most simple topology and has limited connection capabilities.
  - Arbitration Loop: In this type of connection, all the devices are connected into a loop similar to a Token Ring. Any addition or removal of the device within this loop will cause all the operation in the loop to stop. A faulty device within the loop will cause the whole loop to malfunction. By using a Fiber Hub, multiple devices can be connected together forming a logical loop, and bypass the faulty nodes meaning that a faulty device will not affect the communication of the whole loop. Arbitration loop was used in small SAN environments, but are obsolete nowadays. The main reason for it to be obsolete is because an Arbitration Loop can only contain a maximum of 127 devices, and the amount of devices used in SAN environment nowadays are way more than 127 devices.
  - Switched Network: This is the method that builds modern FC SAN networks. It uses FC switches to interconnect host and storage devices. In the modern SAN, it is the recommended best practice to use 2 switches to connect the host and storage devices, in order to form a redundant link which increases the reliability of SAN. Switches are an intelligent devices, it not only can perform the interconnection between devices but can do many more functions related to the network.

Types of Ports on FC Switches

- D_Port (Diagnose Port): It is used to isolate the ISL (Inter-Switch Link) to diagnose the link level faults. It can only be used for diagnostics and testing, and does not carry any Fabric traffic or data flow.
- E_Port (Expansion Port): It is used for ISL with other switches to achieve Fabric expansion.
- EX_Port: It is a type of E_Port that is used to connect the FC Router with the outer fabric. The EX_Port is terminated by the Router and cannot be used to combine fabrics like done by the E_Ports of switches.
- F_Port (Fabric Port): The port which is connected to the Node Port (N_Port) of the FC devices. For example, when connecting storage devices, the switch port will be displayed as the F_Port.
- FL_Port(Fabric Loop Port): It is a type of L_Port that is able to perform the function of an F_Port, attached via a link to one or more NL_Ports in an Arbitrated Loop topology.
- G_Port(Generic Fabric Port): A generic switch port that may function either as an E_Port, A_Port, or as an F_Port.
- M_Port(Mirror Port): Used to replicate all the traffic flow between the source port to the destination port.
- U_Port / GL_Port (Universal/Generic Fabric Loop Port): The most basic FC port type, all unidentified or uninitialized ports belongs to U_Port or GL_Port.

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

HUAWEI

- The list above shows the different types of Ports in the FC fabric and their respective functions. The usage of correct port type can help prevent issues in accessing the fabric from the nodes.

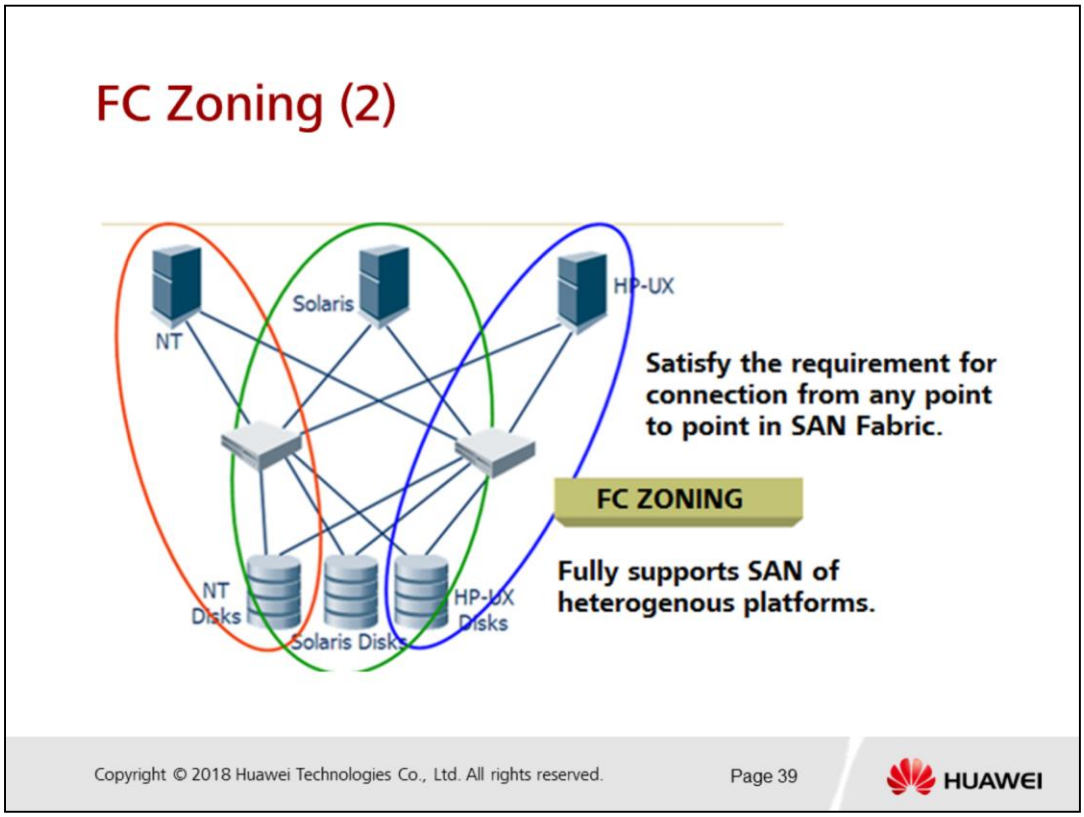
FC Zoning (1)

- Zoning & Grouping of Nodes within the SAN Fabric.
- Dynamically exchange core functions.
- Quarantine and Isolation:
  Supports Heterogeneous Device/Structure, and RSCN.

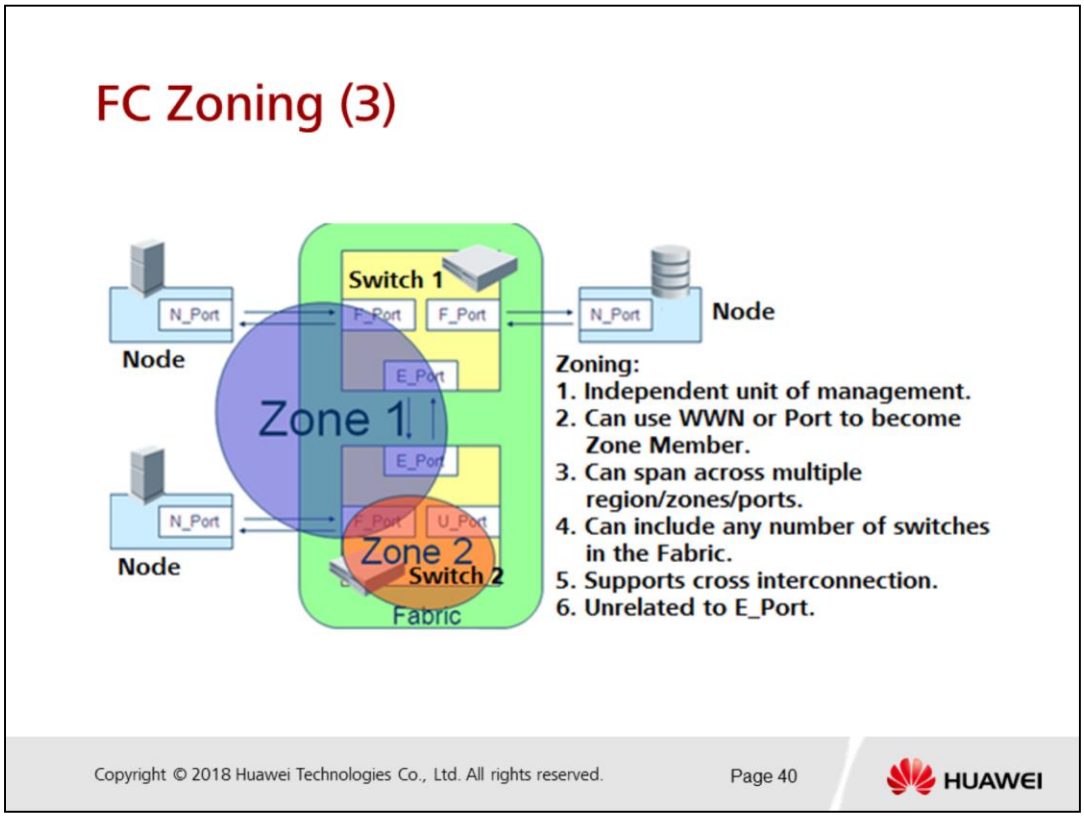
Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

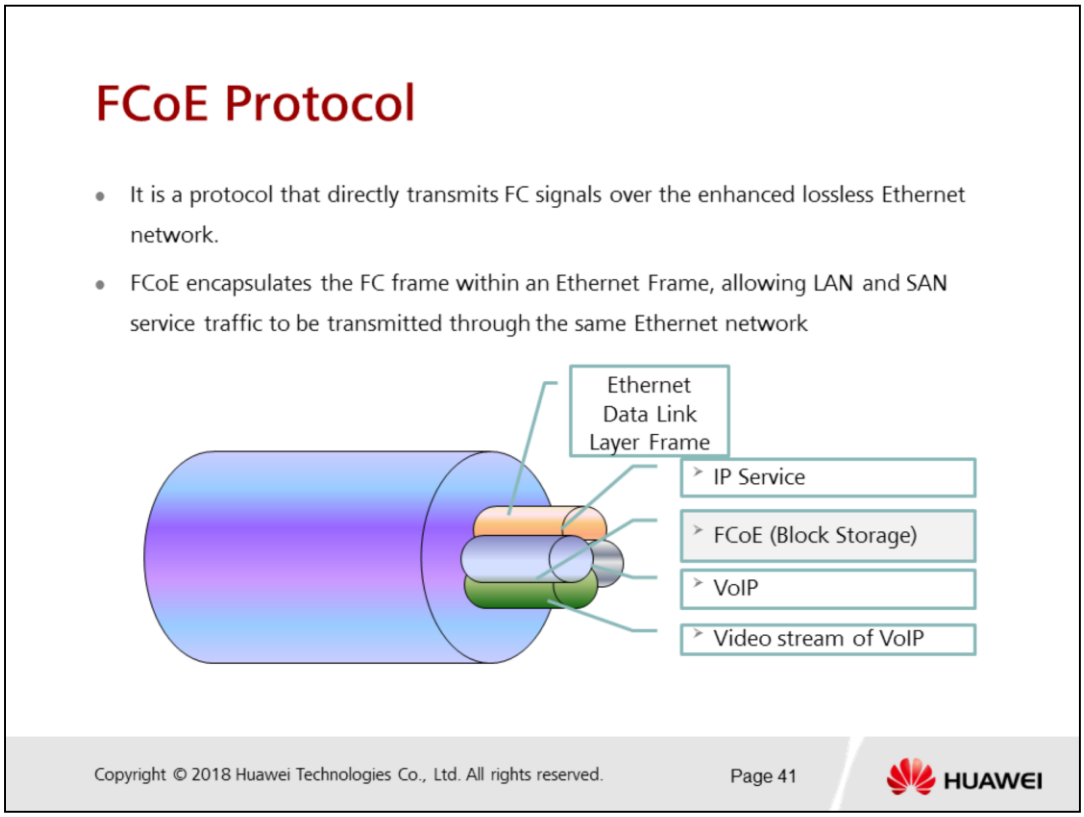
- FC switches Zoning function is similar to the VLAN functions of Ethernet networks, which is used to achieve interconnection of specific devices in the network and avoiding flooding of broadcast packets.
- Zoning function allows access control of the members of the same VSAN(Virtual SAN), providing further division of regions within the VSAN. Different N_Port members can be added into a Zone based on different purposes, and this also allows the isolation between the N_Port members of different zones.
- Registered State Change Notification (RSCN) : A feature in the FC switches that is responsible to notify the nodes of the changes within the network fabric or changes within the architecture itself.
- In Fibre Channel protocol, a registered state change notification (RSCN) is a Fibre Channel fabric's notification sent to all specified nodes in case of any major fabric changes. This allows nodes to immediately gain knowledge about the fabric and react accordingly.



- The diagram above shows the application of FC zoning in a heterogeneous SAN platform.
- FC zoning works across storage platforms in the fabric and can logically group the host nodes and storage nodes within the SAN fabric.

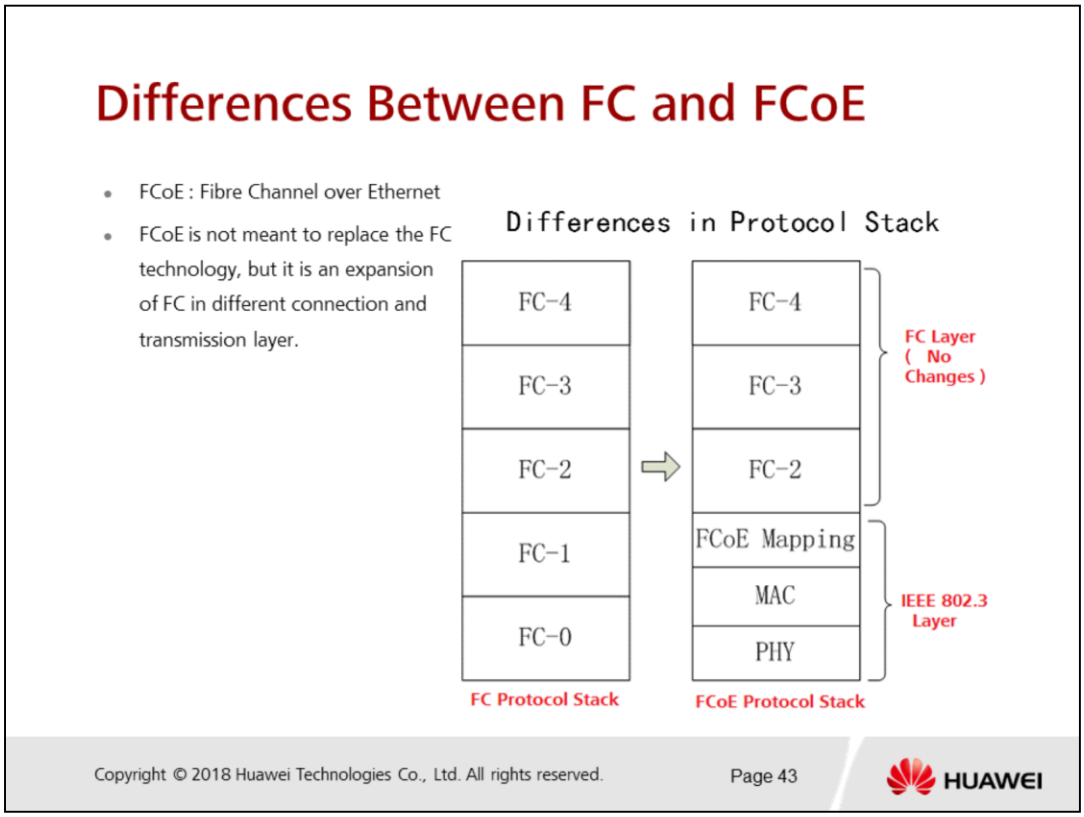


- The diagram above shows a scenario where FC zoning is applied by spanning across multiple switches and interconnection of two FC zones.
- The nodes are logically isolated in the fabric by the zones in the SAN fabric.



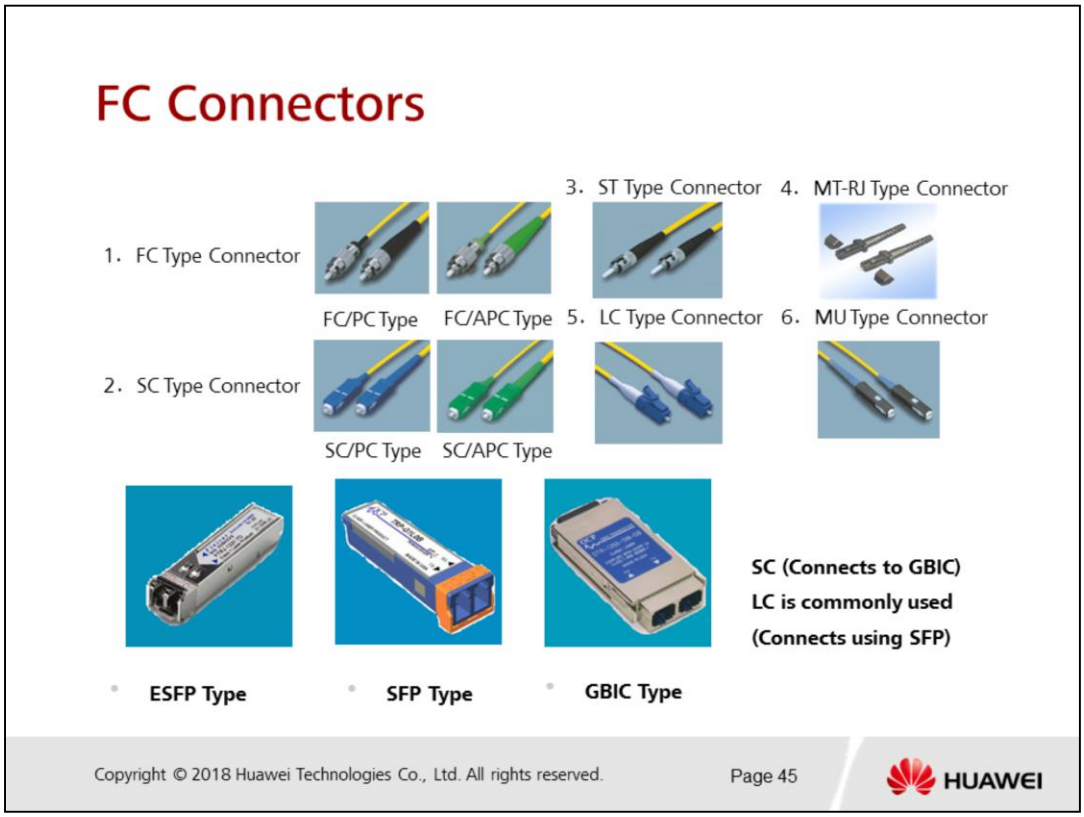
- FCoE (Fibre Channel Over Ethernet) can provide standard FC services in Ethernet networks such as device discovery, global namespace, zoning and these services can operate according to the actual original standards, retaining the low latency, high performance capabilities of FC.
- From the perspective of FC protocols, FCoE is just carrying FC on a new type of link which is the Ethernet link. One thing to be noted is that, this Ethernet network must be Enhanced Lossless Ethernet Network, in order to satisfy the transmission requirements of FC protocols towards the link layer.
- Features of FCoE:
    - Standards Organization: The protocol was submitted to ANSI (American National Standards Institute)T11 committee for approval in 2008, and it needs to work closely with IEEE (Institute of Electrical and Electronics Engineers). Thus, this protocol is verified and accepted globally.

    - Objective of the protocol: FCoE intends to utilize the expansion capabilities of Ethernet network, and at the same time retain the advantage of FC in terms of high reliability and high efficiency.
    - Other Challenges: The combination of FC and Ethernet, requires them to handle issues such as prevent the loss of packets, path redundancy, failover, frame splitting and recombining, and lossless transmission etc.
    - FC by default has poor compatibility and doesn't support long distance transmission. This are the 2 problems that FCoE also could not solve.
- Note: VoIP refers to Voice over IP, which is technology to transmit audio and video over the Ethernet network.



- FCoE retains the protocol stack above the FC-2 layer, and replaces the FC-0 and FC-1 layers with the Physical and Data Link Layer in the Ethernet protocol.
- Due to the fact that FC-0 main purpose is to define the types of transmission medium, and FC-1 was responsible for the encoding and decoding of the frames, both of these layers was required in the FC SAN network during data transmission. However, since FCoE runs on the Ethernet network, these 2 layers are no longer required and can be replaced with the Physical and Data Link Layer of the Ethernet protocol.
- The value of FCoE lies in the fact that the users have the choice of either having the whole logical network to function as the dedicated network for storage data transmission and signaling, or to have a shared hybrid network that has mixed traffic of storage data transmission, VoIP, video streaming and other data transmission. The objective of FCoE is to integrate storage data transmission with the Ethernet network in the preconditions that it continues to support users requirement towards high performance and features extensiveness of FC SAN.
- Different Operating Environment: FC protocol runs on tradition FC SAN network. Meanwhile, FCoE runs as the storage protocol in the Ethernet network.

- Different Operation Channels: FC protocols runs in FC network and all packets operates within the FC channel. In Ethernet network, there are multiple protocol operating at the same time such as traditional IP, ARP. Thus FCoE needs to create a virtual FC channel to carry all the FC packets.
- FCOE Has Added FIP: Compared to FC protocol, FCoE has the addition of FIP which is the FCOE Initiation Protocol(FIP). Since FCoE is a storage protocol that runs on Ethernet, for it to run normally in this network, it requires FIP to obtain the corresponding VLAN to build virtual FC channels with the FCF(FCOE Forwarders) and also for the maintenance of the virtual links. Hence, from this perspective, we safely say that FCOE has an addition of the extra FIP compared to the FC protocol.
- FCoE requires the support of other protocols: Due to the fact that Ethernet is a non-lossless network, it allows certain packets to be lost during transmission and retransmission of lost packets occurs afterwards. However, FC protocol does not allows the lost of packets, thus FCoE which runs on the Ethernet must also have to inherit the same feature of lossless transmission in FC protocol. Hence, for normal lossless operation of FCoE over Ethernet requires some enhancement to the Ethernet network to prevent packet loss. This enhancement to the Ethernet network is called CEE (Converged Enhanced Ethernet).



- FC type optical fiber connector: FC is the abbreviation of the Ferrule Connector, which indicates that the external reinforcement method uses a metal sleeve, and the fastening method is a turnbuckle. Such connectors has a simple structure, it is easy to operate, and it is easy to manufacture, but the fiber end is more sensitive to micro dust, and Fresnel reflection is easy to occur in these connectors, and it is difficult to improve the return loss performance of these connectors.
- SC type optical fiber connector: SC is the abbreviation of Square Connector, its named such because the connector housing is square in shape. The pin size and the coupling sleeve used in these connectors are the same as the FC type, in which the fiber end of the pin is mostly PC or APC type. The fastening method is a plug-and-pin latch type, which does not require rotation. These connectors are inexpensive, easy to insert and withdraw, have small fluctuations in insertion loss, high compressive strength, and high installation density.

- Based on the shape of the fiber end of these connectors, it is divided into the categories of FC, PC(including SPC or UPC) and APC.
    - FC type - Flat Connect: This is called a flat connector. The end of the ceramic pin used is a flat contact (FC). Such a connector has a simple structure, is easy to operate, and is easy to manufacture. However, the fiber end is sensitive to fine dust, and Fresnel reflection is likely to occur. Therefore, it is difficult to improve the return loss performance. Return Loss: 40dB
    - PC type - Physical Connect: It is also called the spherical connector, which refers to the spherical surface of the fiber end of the pin (PC), and the external structure has not changed, so that the insertion loss and return loss performance has been greatly improved. Return Loss: 40dB
    - APC type - Angle Physical Connect: It is also called abrasive spherical connector, which is similar to the PC type but with an angled surface for connection. Return Loss: 55dB
- ST Type Connector: ST and SC connectors are quite similar, but the difference between them is that the ST connector has an exposed fiber end, while the SC connector fiber end is enclosed and not exposed. For 10Base-F connections, commonly ST type connectors are used, and commonly SC type connectors are mostly used for 100Base-FX connections.
- MT-RJ Type Connectors: MT-RJ is the abbreviation for Multi-Transmit-Receive Joint which is a connector for multi core transmission. It has a latch mechanism similar to RJ-45 LAN connector, and it is aligned to the optical fiber through the guide pins installed on both sides of the small sleeve. For the ease of connection with the optical transceiver, the optical fiber of the connector end is double-core (Interval 0.75mm) alignment design. It is mainly used as the next generation of high density fiber optic connector for data transmission.
- LC Type Connector: LC is the abbreviation of Lucent Connector and as the name suggests it was developed by Lucent Corporation. It incorporates an easy to operate modular jack(RJ) latch mechanism. The used pins and sleeve specification is half the size of SC and FC connectors which is at 1.25mm. This helps to increase the density of the cables within fiber distribution frame. Currently, in the aspect of single mode SFF(small form factor) connectors, LC type connectors has taken a dominant position in the market, and has increased growth in applications in the multimode aspect of the market.
- MU Type Connectors: MU is the abbreviation of Miniature Unit Coupling. Based on the most widely used SC-type connector, MU is the world's smallest single-fiber connector developed by NTT (Nippon Telegraph and Telephone) Corporation. The connector uses a 1.25mm diameter sleeve and self-retaining mechanism, which has the advantage of enabling high-density mounting. With MU's 1.25mm diameter sleeve, NTT has developed MU series of connectors. With the rapid development of larger bandwidth and larger capacity of optical fiber networks and the wide application of DWDM technology, the demand for MU-type connectors will also increase rapidly.

Contents

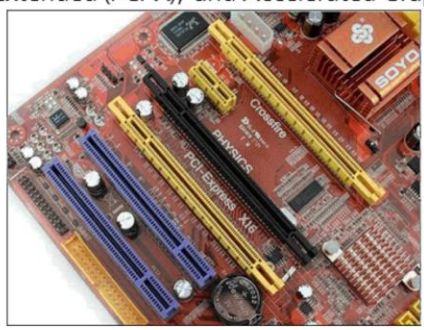
1. SCSI/iSCSI
2. SAS
3. FC/FCOE
4. **PCIe**
5. IB
6. CIFS/NFS
7. FTP/HTTP

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

HUAWEI

What Is PCIe ?

- PCI Express (also known as PCIe) is a high-performance, high-bandwidth serial communications interconnection standard first introduced by Intel and later developed by the Peripheral Component Interconnect Special Interest Group (PCI-SIG) to replace bus-based communications Architectures such as: PCI, PCI Extended (PCI-X), and Accelerated Graphics Port (AGP).

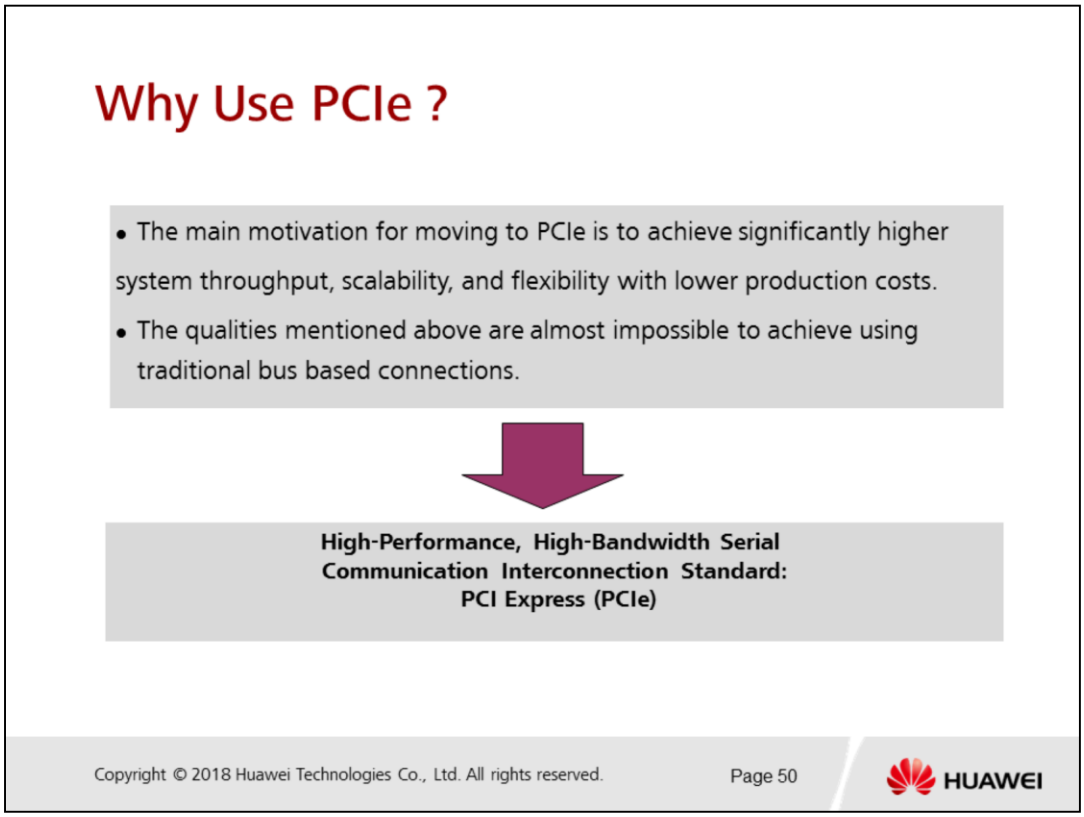


Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

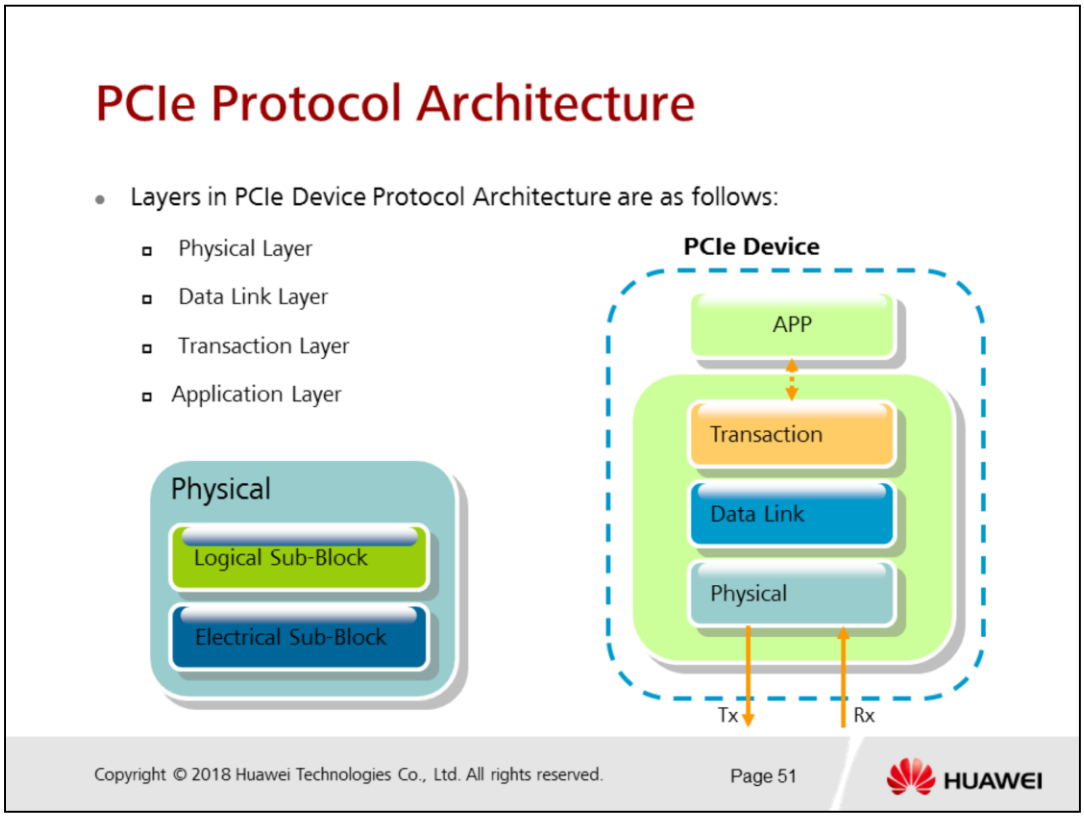
HUAWEI

- In the year 1991, Intel Corporation introduced the concept of PCI. With the development of modern processor technology, the replacement of the parallel bus with the high-speed differential bus is a general trend in the field of interconnection. Additionally, when compared with single-ended parallel signals, the high-speed differential signals can be used for higher clock frequencies, and thus induced the emergence of PCIe bus.
- PCI Express (aka PCIe) is a high-performance, high-bandwidth serial communications interconnection standard first introduced by Intel and later developed by the Peripheral Component Interconnect Special Interest Group (PCI-SIG) to replace bus-based communications. Architectures such as: PCI, PCI Extended (PCI-X), and Accelerated Graphics Port (AGP).
- PCI Express has the following advantages over the traditional PCI bus:
  - Dual-channel, high-bandwidth, fast transfer rate: Achieves a similar full-duplex transmission mode (RX and TX are separated). It has high transmission rate. The bandwidth of the first generation PCIe X1 is 2.5 Gigabits per second (Gbps), the second generation has up to 5.0 Gbps, and the recently released PCIe 3.0 standard is capable of supporting 8.0 Gbps. For larger bandwidths, it can be achieved by expanding the number of links, and the resulting bandwidth is the number of paths(N) multiplied by the bandwidth of each paths.

- Compatibility: PCIe maintains compatibility with PCI at the software level, and in version upgrades, and also backward compatible with PCI software.
- Ease of use: Supports hot-plugging. PCIe bus interface slots include "hot-swap detection signals" that can be hot swapped and can be hot swapped like USB.
- Equipped with error handling and advanced error reporting capabilities: Benefited from the PCI Express bus hierarchical structure, its software layer has error handling and error reporting capabilities.
- Each physical connection also has multiple virtual channels: multiple virtual channels are supported in each physical channel (in theory, 8 virtual channels are allowed for independent communication control) so as to support the QoS of each virtual channel, reaching very high level of traffic quality control.
- Saves IO, reduces board space, and reduces crosstalk: For example, a typical PCI bus data cable has to have at least 50 IO paths, but PCIe Xl only requires 4 IO paths. The reduction of IO paths saves board space and the direct distance of each IO path can be wider, thereby reducing crosstalk.



- The formulation of the PCI Express standard is based on the concerns for the future and it continues to evolve to provide higher throughput for the system. The throughput of the first-generation PCIe protocol was 2.5 gigabits per second (Gbps), while the second-generation PCIe protocol reached 5.0 Gbps, and the recently released PCIe 3.0 standard can support 8.0 Gbps. While the PCIe standard continues to utilize the latest technology to provide increasing throughput, the transition from PCI to PCIe will be simplified using a layered protocol by keeping the driver software compatible with existing PCI applications.
- The PCIe protocol features include:
  - Point-to-point connection
  - High reliability
  - Tree network
  - Full duplex
  - Frame-based transmission



- Physical Layer:
  - The physical layer in the PCI Express bus architecture mainly defines the physical characteristics of the bus. In future development, the performance of the PCI Express bus can be further improved by speeding up or changing the codec mode. These changes will only affect the physical layer and will not affect other structures, thus facilitating the upgrade process.
- Data Link Layer:
  - The important role of the data link layer is to ensure the correctness and reliability of data packets transmitted over the PCI Express bus. It will check if the data packet encapsulation is complete and correct, add the serial number and cyclic redundancy check code (CRC) to the data for detection and error correction, and use ack/nack handshake protocol for detection and correction.
- Transaction Layer:
  - There are 2 main role of the transaction layer: The first role is to accept read and write requests sent from the software layer, or to create a request encapsulated packet itself to the data link layer, this data packet is called "Transaction Layer Packet". The second role of this later is to accept the response packet (Data Layer Packet, DLP) sent from the data link layer and associate it with the relevant software request and send it to the software layer for processing.
- Application Layer:
  - This layer mainly consists of applications that sends requests and receive responses from PCIe devices.

Bandwidth Of PCIe Links

Specification	Bandwidth/IO Count	Work Frequency	Transmission Rate	Encoding
PCI 2.3	32 bit	33/66 MHz	133/266 MB/s	--
PCI-X 1.0	64 bit	66/100/133 MHz	533/800/1066 MB/s	--
PCI-X 2.0 (DDR)	64 bit	133 MHz	2.1 GB/s	--
PCI-X 2.0 (QDR)	64 bit	133 MHz	4.2 GB/s	--
AGP 2X	32 bit	66 MHz	532 MB/s	--
AGP 4X	32 bit	66 MHz	1.0 GB/s	--
AGP 8X	32 bit	66 MHz	2.1 GB/s	--
PCI-E x1	4 (Differential)	2.5 GHz	500 MB/s (Full Duplex)	8b/10b
PCI-E x2	8 (Differential)	2.5 GHz	1.0 GB/s (Full Duplex)	8b/10b
PCI-E x4	16 (Differential)	2.5 GHz	2.0 GB/s (Full Duplex)	8b/10b
PCI-E x8	32 (Differential)	2.5 GHz	4.0 GB/s (Full Duplex)	8b/10b
PCI-E x16	64 (Differential)	2.5 GHz	8.0 GB/s (Full Duplex)	8b/10b
PCI-E Gen2 x1	4 (Differential)	5.0 GHz	1.0 GB/s (Full Duplex)	8b/10b
PCI-E Gen2 x2	8 (Differential)	5.0 GHz	2.0 GB/s (Full Duplex)	8b/10b
PCI-E Gen2 x4	16 (Differential)	5.0 GHz	4.0 GB/s (Full Duplex)	8b/10b
PCI-E Gen2 x8	32 (Differential)	5.0 GHz	8.0 GB/s (Full Duplex)	8b/10b
PCI-E Gen2 x16	64 (Differential)	5.0 GHz	16.0 GB/s (Full Duplex)	8b/10b
PCI-E Gen3 x1	4 (Differential)	8.0 GHz	~2.0 GB/s (Full Duplex)	128b/130b
PCI-E Gen3 x2	8 (Differential)	8.0 GHz	~4.0 GB/s (Full Duplex)	128b/130b
PCI-E Gen3 x4	16 (Differential)	8.0 GHz	~8.0 GB/s (Full Duplex)	128b/130b
PCI-E Gen3 x8	32 (Differential)	8.0 GHz	~16.0 GB/s (Full Duplex)	128b/130b
PCI-E Gen3 x16	64 (Differential)	8.0 GHz	~32.0 GB/s (Full Duplex)	128b/130b

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

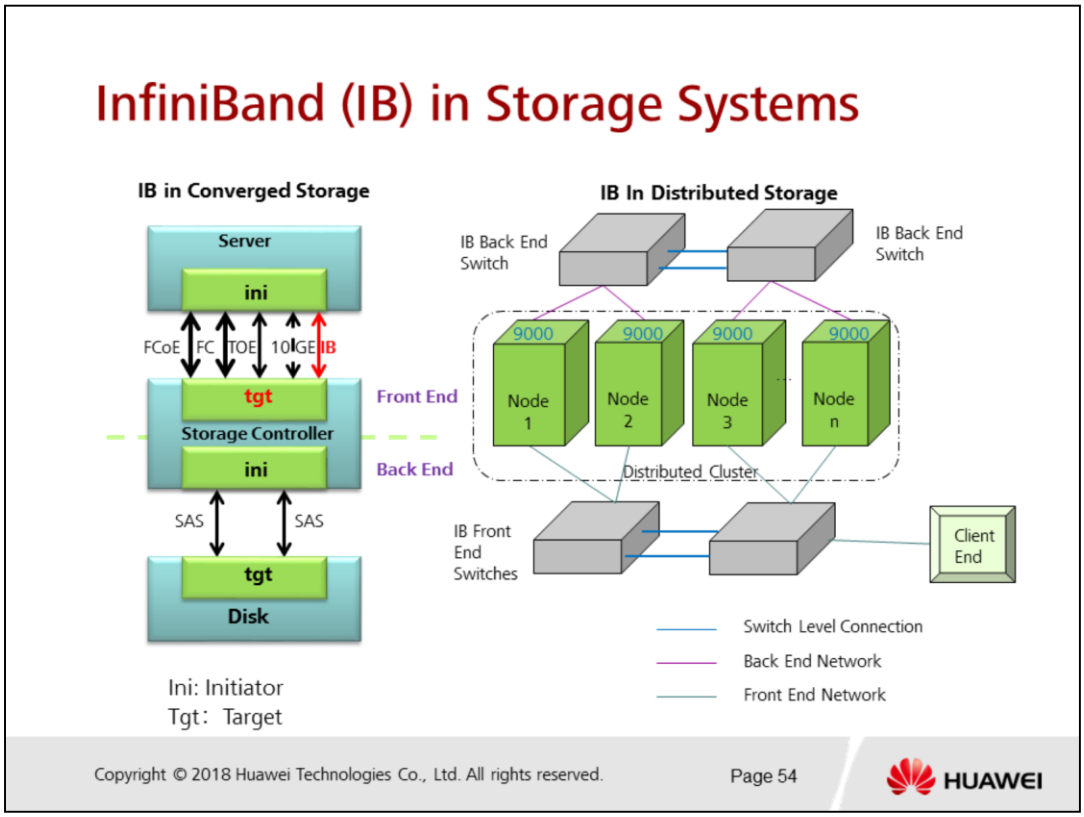
HUAWEI

- The table above shows the different bandwidth, working frequency, encoding method and the transmission rate for the specifications of PCI an d PCIe technology of different generations.

Contents

1. SCSI/iSCSI
2. SAS
3. FC/FCOE
4. PCIe
5. **IB**
6. CIFS/NFS
7. FTP/HTTP

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



- IPoIB (IP over IB), which is the adaptation layer of Linux kernel and IB driver, is responsible for constructing, destroying IP header, sending and receiving IP message.
- The IB front-end network is used to communicate with customers for data exchange networks, and transmits data based on the IPoIB protocol.
- The IB back-end network is used to store the data exchange network between the nodes inside the device. The RPC(Remote Procedure Call) module uses RDMA (Remote Direct Memory Access) to complete the data synchronization between nodes.

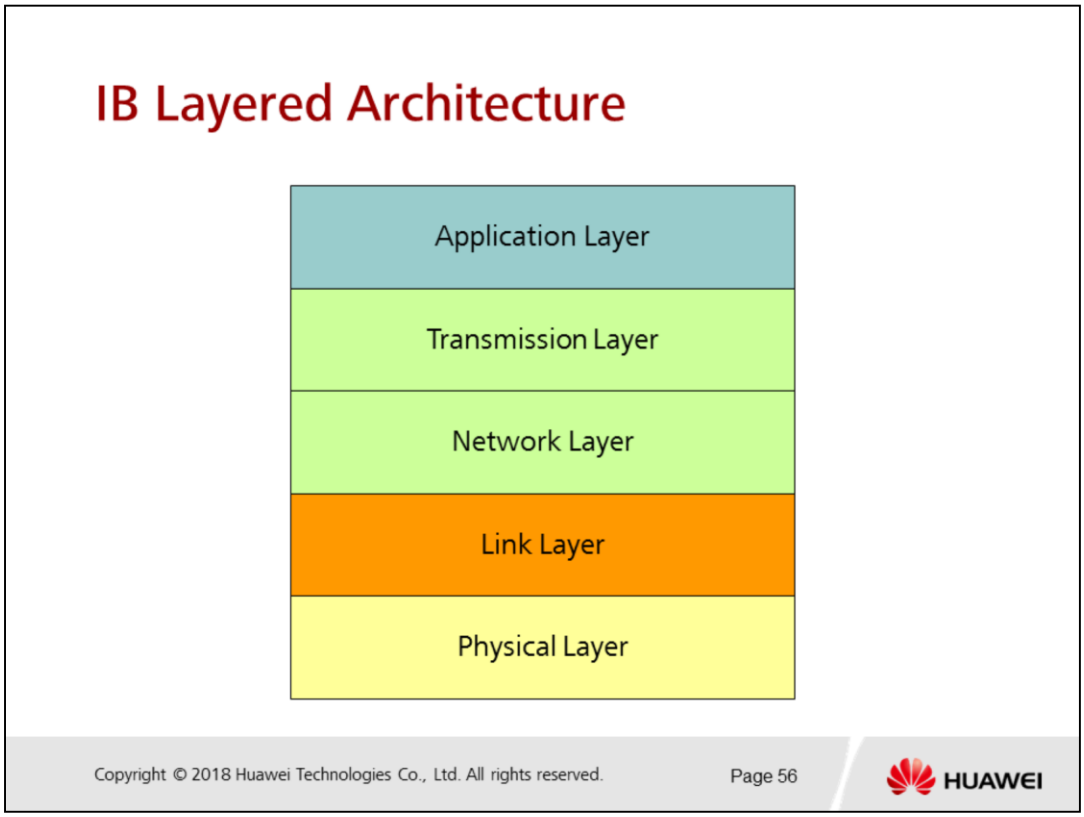
What Is IB ?

- IB (InfiniBand):
  - InfiniBand technology is not used for general network connection. Its main design purpose is to solve server-side connection problems.
  - InfiniBand technology is used for communication between servers and servers (such as replication, distributed work, etc.), between servers and storage devices (such as SANs and direct storage attachments), and between servers and networks (such as LANs, WANs, and the Internet).
- InfiniBand features:
  - Based on standard protocols
  - High bandwidth, low latency
  - Remote direct memory access
  - Transport offload
- The key to the InfiniBand architecture is solving the bottleneck problem of the shared bus by adopting a point-to-point switching architecture that is specifically designed to address fault tolerance and scalability problems. By adding switches to the InfiniBand system, I/O system expansion can be easily implemented, and thus allowing more terminal devices to access the I/O system.

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

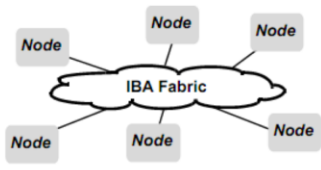
HUAWEI

- The InfiniBand standard defines a set communication system used by multiple devices, which includes channel adapters, switches, and router channel adapters for connection to other devices, including host channel adapters (HCA) and target channel adapters (TCA).
- The main features of the InfiniBand protocol:
  - Protocol Based On Standards: The InfiniBand Trade Association, which was established in 1999, consists of 225 companies that have jointly designed this open standard. The members that mainly control the association include: Agilent, Dell, HP, IBM, InfiniSwitch, Intel, Mellanox, Network Appliance and Sun Microsystems. More than 100 other members assisted in the development and promotion of the IB standard.
  - Speed: InfiniBand's performance of 10 gigabytes per second significantly exceeds the speed of 4 gigabits per second for existing Fibre Channel technology and also exceeds 1 gigabit per second speed for Ethernet technology.
  - Memory: InfiniBand-enabled servers use host channel adapters to convert protocols to the server's internal PCI-X or PCI-Xpress bus. HCA has RDMA functionality, sometimes also referred to as Kernel Bypass. RDMA is a good fit for a cluster because it can use a virtual addressing scheme to let the server know and use some of the other server's memory without having to refer to the operating system's kernel.
  - Transport Offload: RDMA can help offload transfers, which shifts packet routing from the OS to the chip level, saving the processor's processing workload. An 80 GHz processor is required to process data at 10 Gbps in the OS.

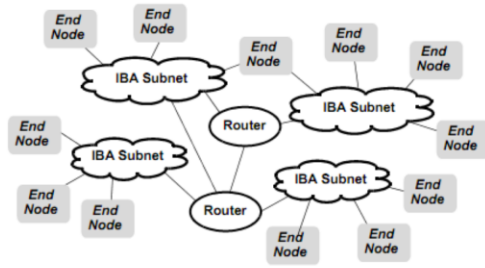


- Physical layer: Defines connections of three different rate of transmission which are, 1X, 4X, and 12X respectively, and their signal transmission rates are 2.5, 10, and 30 Gb/s respectively. This means that the IBA (InfiniBand Architecture) allows multiple connections until a 30Gbps connection speed is achieved. Due to the use of full-duplex serial communication, a single-speed two-way connection requires only four cables, and in the case of a 12-speed mode, it requires only 48 cables, which is very attractive in terms of cable usage.
- Link layer: The link layer provides functions such as packet design, point-to-point connection operation, and packet switching in the local subsystem. At the packet communication level, two special packet types are specified, which are data transmission and network management packets. The network management package provides functions such as device enumeration operation control, subnet indication, fault tolerance, etc. On the other hand, the data transmission package is used to transmit actual data information. The maximum length of each packet is 4 KB. Within each specific device subnet, the direction and switching of each packet is completed by the subnet manager through the usage of the local 16-bit identification address.
- Network layer: This layer provides a routing mechanism for information packets from one sub-structure to another sub-structure. Each of the source and destination nodes has a Global Routing Header (GRH) and a 128-bit IPv6 address. The network layer also embeds a standard global 64-bit identifier that is unique across all subnets. Through intricate exchanges between these identification values, data is allowed to travel across multiple subnets.

IB Architecture (IBA)



IBA Network



IBA Network Components

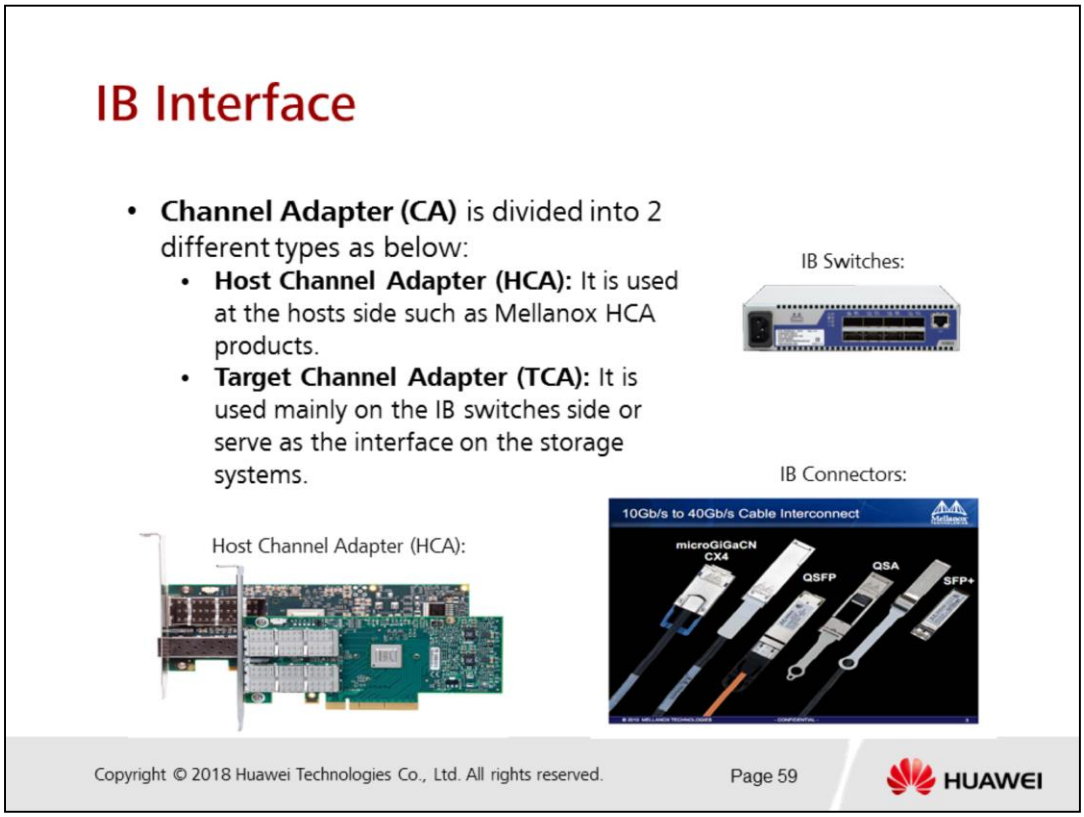
- Components of InfiniBand Architecture(IBA):
- Node: Host Channel Adapter (HCA), Host Target Adapter(HTA).
- Network: Switch, Routers.
- Physical: Links(Fiber or Cables), Repeaters.

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

HUAWEI

- IBA originally targeted to become the standard interface for internal and external connections for computing, communication and storage. However, IBA lost the opportunity for internal system operation in communication devices due to the implementation of Rapid I/O by multiple communication equipment manufacturers. So, IBA can only set its target towards computing and storage devices for its adoption. But coincidently the rapid growth of SAN, made FC technology to be widely adopted which marks that IBA lost its ground for adoption in storages. Afterwards, IBA sets its sights at computing devices, aiming to replace PCI standard, but PCI-X took the place in the replacement of PCI. After PCIe was proposed by Intel to replace PCI-X, it made it impossible for IBA become the standard for internal connections in computing devices. Even the main competitors of Intel such as Sun and AMD used HyperTransport/HTX and not IBA which made the chances for IBA adoption in internal system connections for computing devices much smaller.

- Since there is no chances for becoming the internal connection standard of devices, IBA could only move towards the fields of high speed transmission outside of machines. Although Rapid I/O, PCIe and PCI doesn't have any solutions for off board external connection scheme, unfortunately FC is developing at a rapid pace, thus making IBA to compete with FC in this field. Although FC loses to IBA in terms of performance, but it has much cheaper costs, and factors such as the increase in speed of Ethernet from 1Gbps to 10Gbps and the finalization of its optical fiber transmission standard (IEEE 802.3ae) and copper transmission standard (IEEE 802.3an) in the year 2002 and 2005 respectively causes IBA to not able to take over the market share in the high speed transmission field.
- IBA finally founds it purpose and fits in the field of clustered supercomputers. Since FC only has 2Gbps, 4Gbps transmission speed, and 10Gbps Ethernet is not mature enough with the delay control not yet to be refined, IBA became the most suitable option for application. IBA has low latency and can transmit in the speed of multiples of 10Gbps as the unit. This allows IBA to be used in as parts or mixed with other high speed technologies in considerable number of systems of the world's top 500 performance supercomputers.
- However, the top 500 supercomputers which is on top of the computing pyramid only serves as a minority, and the scale of usage could not sustain the long term development of the IBA and its ecosystem. This means that IBA need to venture into mid tier or low tier of the market to gain wider adoption, even if it means to compete and conflict with FC and GbE(Gigabit Ethernet) and also face the challenge of proprietary technology such as Quadrics' QsNet, or Myricom's Myrinet.
- When we talk about Infiniband now, the main application is in the server clusters, and the interconnection between systems and its application in the fields of storage, data centers and virtualization.



- Mellanox Technologies was founded in 1999 and has headquarters in Santa Clara, California, and Yokneam, Israel. Mellanox is a leading provider of server and storage end-to-end connectivity solutions. In the end of 2010, Mellanox completed the acquisition of the famous Infiniband switch manufacturer Voltaire, which enabled Mellanox to gain more comprehensive capabilities in the HPC, cloud computing, data center, enterprise computing and storage markets.
- Channel adapters are divided into:
  - Host Channel Adapter (HCA), which is used to control the NODE.
  - Target Channel Adapter(TCA), which is used for peripheral NODE, so that the IO of the device is offloaded from the host and transferred directly in the network.
- Channel adapters implement the physical layer, link layer, network layer, and transport layer functions.
- The channel adapter is an important part of the IB network interface. It is a programmable DMA device with specific protection features allowing local and remote DMA operations.

IB Signaling Mode

Transmission Rate (Gbit/s)	SDR	DDR	QDR	FDR	EDR
1X	2.5	5	10	14	26
4X	10	20	40	56	104
8X	20	40	80	112	208
12X	30	60	120	168	312

- SDR: single data rate
- DDR: double data rate
- QDR: quad data rate
- FDR: fourteen data rate
  EDR: enhanced data rate
- HDR : High Data Rate
  NDR : Next Data Rate

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

HUAWEI

- 8b/10b encoding for SDR, DDR, FDR-10, and QDR: This encoding method has 8 bits data per 10 bits transmitted.
- 64b/66b encoding for FDR and EDR: This encoding method has 64 bits of data per 66 bits transmitted.
- The remaining 2 bits in these encoding methods are at the start and end of frame.
- The effective one-way theoretical throughput (actual data rate, and non-signaling rate) will be lower than this value.

What Makes IB Performance So Fast?

The channel-based end-to-end switched interconnection structure does not share the bus and there are no related electronic restrictions, arbitration conflicts, and memory consistency issues.

Simple and efficient protocol, low overhead, protocol supports hardware offload.

QoS: 16 level VL(Virtual Lanes) and 16 level SL(Service Level), which helps in achieving efficient service quality management and credit-based dual-layer flow control mechanism.

RDMA + Zero Copying technology for data.

Supports multiple concurrent links: Theoretically more number of concurrent links equals to more speed, such as QDR: X1 10Gbps, X4 40Gbps, X8 80Gbps, X12 120Gbps.

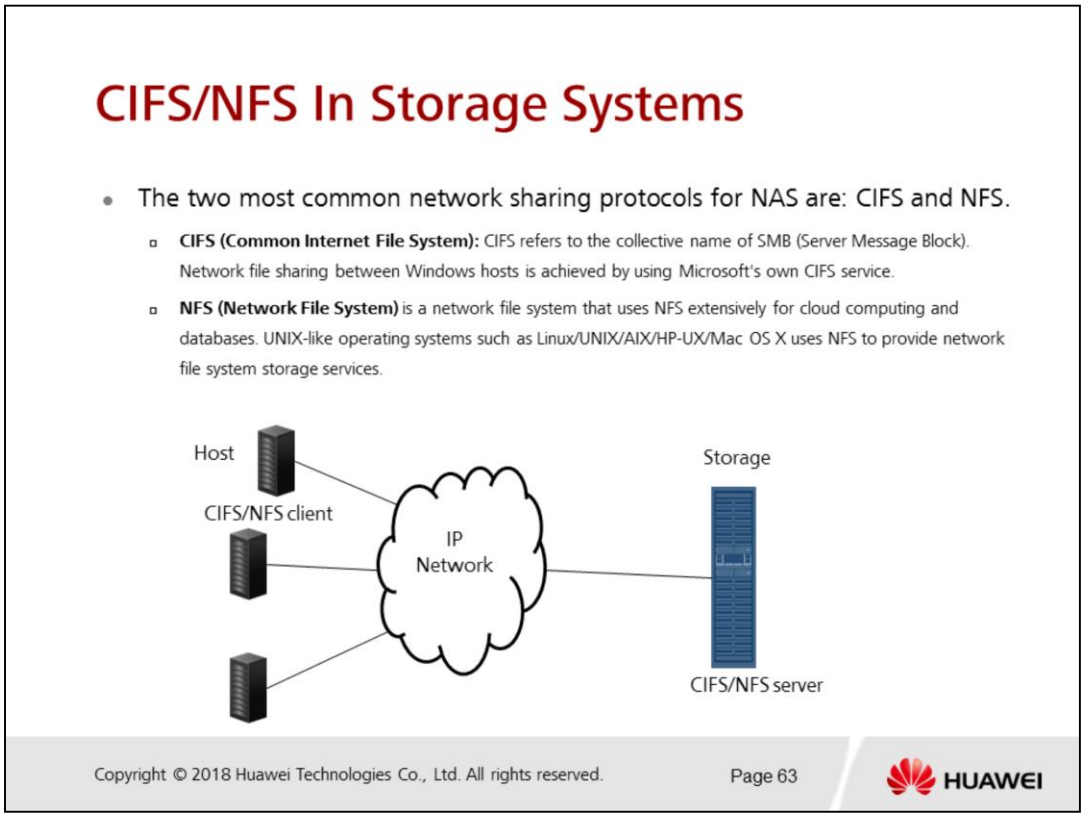
Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

Contents

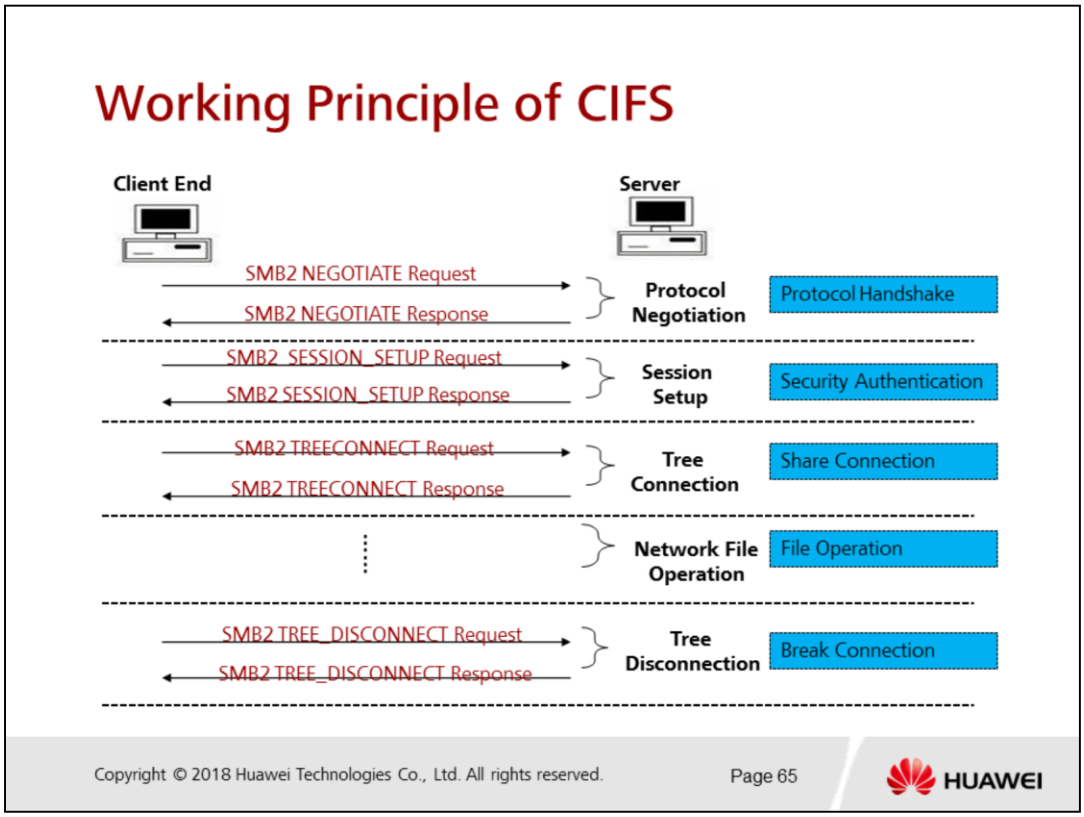
1. SCSI/iSCSI
2. SAS
3. FC/FCOE
4. PCIe
5. IB
6. **CIFS/NFS**
7. FTP/HTTP

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

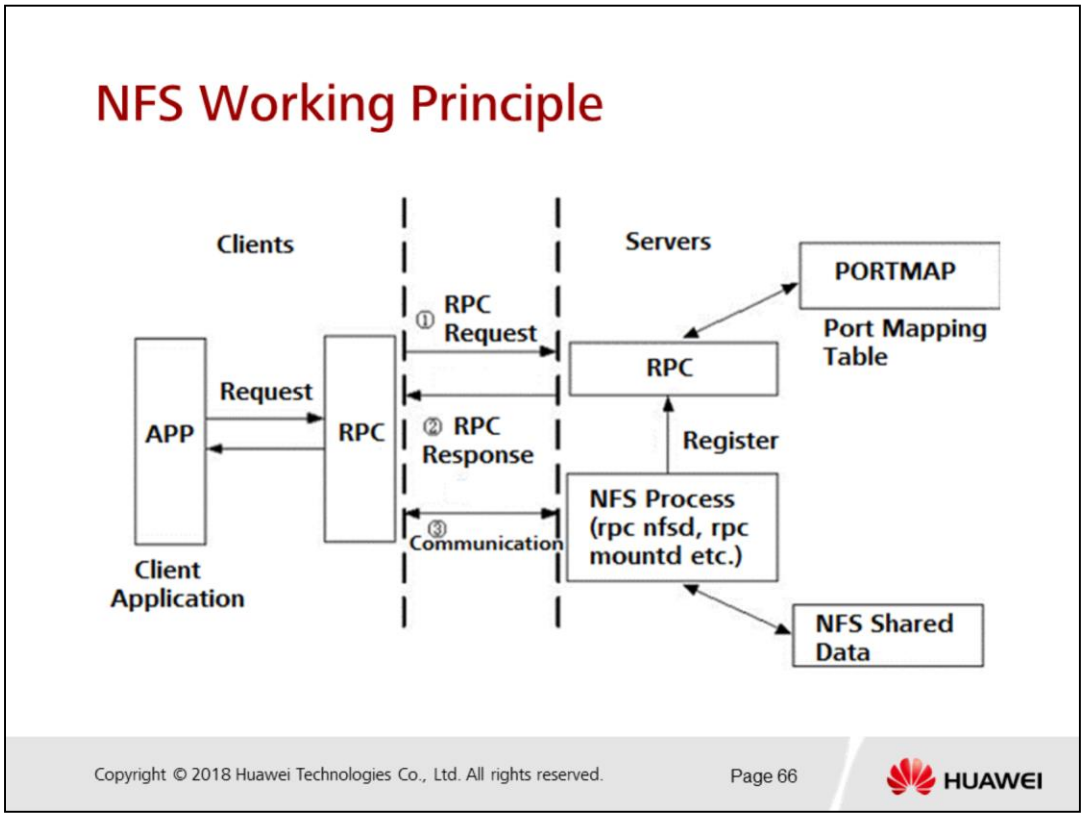


- CIFS:
  - In 1996, Microsoft proposed to rename the SMB as CIFS (Common Internet File System) and added many new features. Now CIFS refers to the general name of SMB, specifically each version is SMB1, SMB2, SMB3.0. SMB is a client/server, request/response mode protocol.
  - The Server Message Block (SMB) was originally developed by IBM's Barry Feigenbaum. Its purpose is to transform the local file interface "Interrupt 13" in the DOS operating system into a network file system. The SMB protocol is mainly used to share files, printers, serial ports, etc. between computers.
  - The SMB has been continuously improved since 1988 and has evolved from SMB to SMB2 (2007) and SMB3 (2012).
  - CIFS (Common Internet File System) is a file sharing protocol developed by Microsoft for connecting Windows clients and servers.
  - CIFS is a standard formed after the publication of the Server Message Block (SMB) developed and used by Microsoft. SMB is mainly used by computers on the network to share files, printers, and serial ports.
  - After redevelopment by UNIX server vendors, SMB can be used to connect UNIX servers and Windows clients and perform tasks such as printing and file sharing.

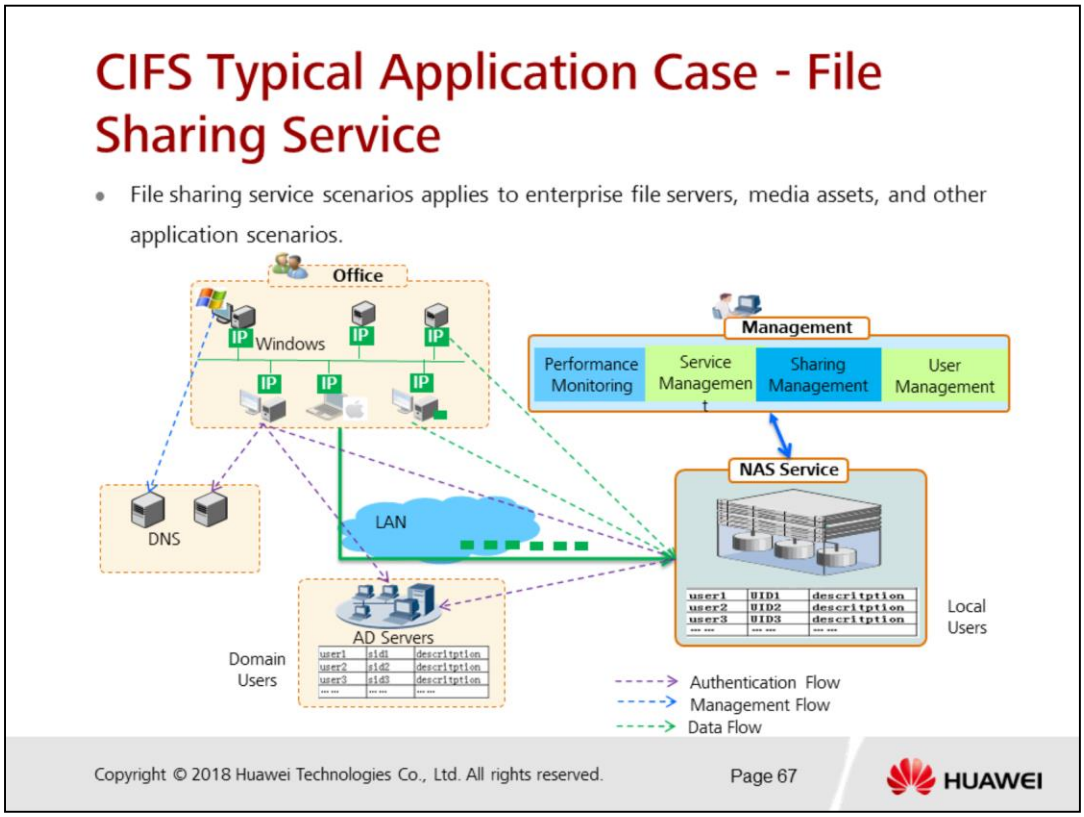
- CIFS service features:
  - Compatible with a variety of operating systems, and can achieve file resource sharing under heterogeneous network environments.
  - CIFS shared units are directories, and the shared directory can be accessed by multiple clients.
  - In a clustered manner, external shared services are provided. Nodes can monitor the service status of each other.
  - Load balancing can be implemented based on the status of services and nodes, and data access is evenly distributed within the cluster.
- NFS:
  - NFS (Network File System) is one of the current mainstream heterogeneous platform sharing protocols. It is mainly used in Linux and UNIX environments. NFS can be used in different types of computers, operating systems, network architectures, and transport protocol environments to provide network file remote access and sharing services.
  - NFS SERVER is deployed in a clustered mode.
  - The NFS SERVER is deployed on each storage server and is a distributed and fully symmetrical cluster architecture.
  - Provides network file system storage services on UNIX-like systems such as Linux/UNIX/AIX/HP-UX/Mac OS X. Allows users to access files on other systems as if they were local files. Provide support for diskless workstations to reduce network overhead.
  - Simplifying application access to remote files eliminates the need to invoke special procedures for accessing these files.
- NFS service features:
  - Compatible with a variety of operating systems, can achieve file resource sharing under heterogeneous network environments.
  - The NFS share unit is a directory, and the shared directory can be accessed by multiple clients.
  - In a clustered manner, external shared services are provided. Nodes can monitor the service status of each other.
  - Load balancing can be implemented based on the status of services and nodes, and data access is evenly distributed within the cluster.



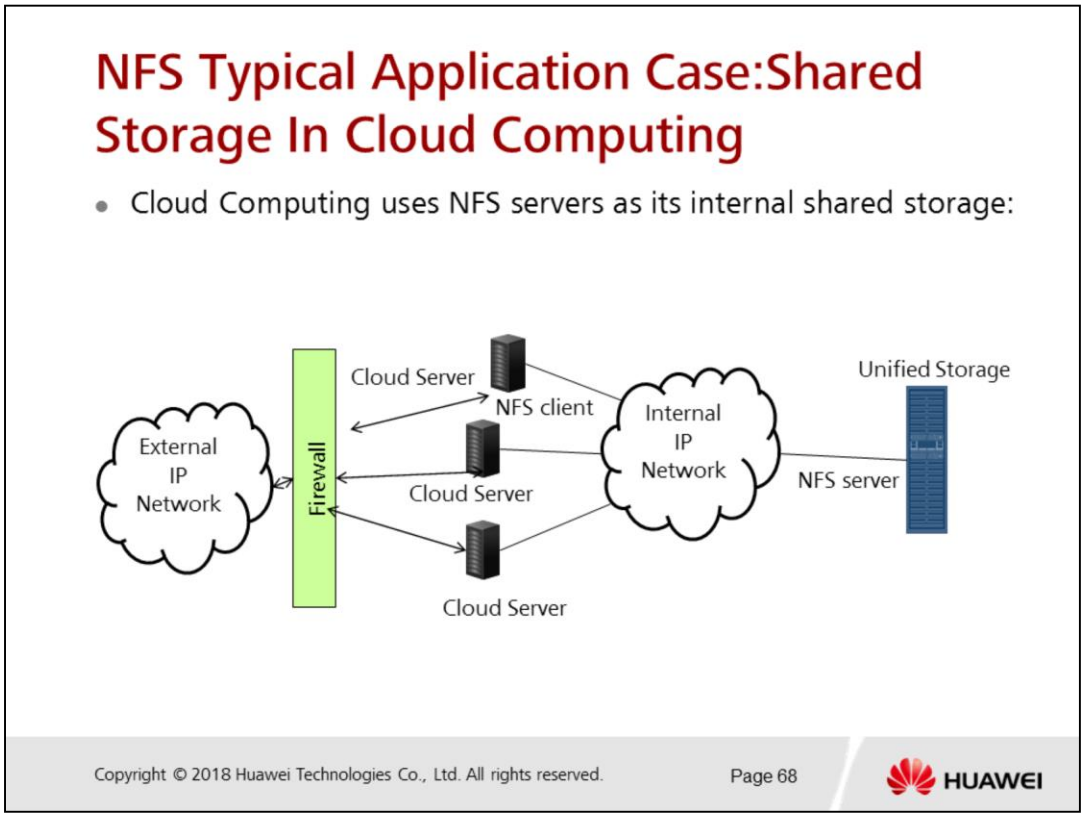
- Authentication method: NTLM, Kerberos etc.
- Network traffic can be carried over TCP or other ways such as: RDMA etc.
- Network file operations are processed and transferred to the actual file system.
- NTLM security supports provider services. NTLM means NT LanManger, one of the authentication methods provided under NT which uses 64-bit encryption.
- Kerberos network authentication. Kerberos is a network authentication protocol designed to provide powerful authentication services for client/server applications through a key system.



- RPC (Remote Procedure Call): A remote procedure call, which is a protocol for requesting services from a remote computer program over a network without having to understand the underlying network technology. The RPC protocol assumes the existence of certain transport protocols, such as TCP or UDP, to carry information data between communications programs. In the OSI network communication model, RPC crosses the transport layer and the application layer. RPC makes it easier to develop applications that include network distributed programming.
- RPC uses client/server mode. The requestor is a client, and the service provider is a server. First, the client that invokes a RPC process sends an invocation of the process parameter to the service process and then waits for a reply. On the server side, the process stays asleep until the RPC message arrives. When a RPC message arrives, the server obtains the process parameters, calculates the result, sends the reply message, and then waits for the next RPC message. Finally, the client-side RPC process receives the reply message, obtains the process result, and then invokes execution to proceed.



- DNS: Refers to domain name service, which is used to implement the mapping of domain names and IP.
- AD: Refers to Active Directory. An AD domain is used for providing directory services.
- The CIFS protocol is mainly used for file sharing. The following are two of its typical application scenarios:
  - File sharing service scenario
    - File sharing service is the most typical application scenario. It is mainly applied to enterprise file servers, media assets, etc., and provides users with file sharing services.
  - Hyper-v virtual machine application scenario
    - Microsoft promotes Hyper-V virtual machines and uses SMB to share virtual machine images. In this scenario, the failover feature of SMB 3.0 needs to be relied on. This feature ensures that the node switchover occurs in the event of a node failure and the service is not interrupted, thus ensuring the reliability of the virtual machine operation.



- Cloud virtualization software (such as VMware) is optimized specifically for NFS clients to create virtual machine storage space on the NFS server's shared space.
- This cloud-optimized NFS client provides better performance and reliability.

Contents

1. SCSI/iSCSI
2. SAS
3. FC/FCOE
4. PCIe
5. IB
6. CIFS/NFS
7. **FTP/HTTP**

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

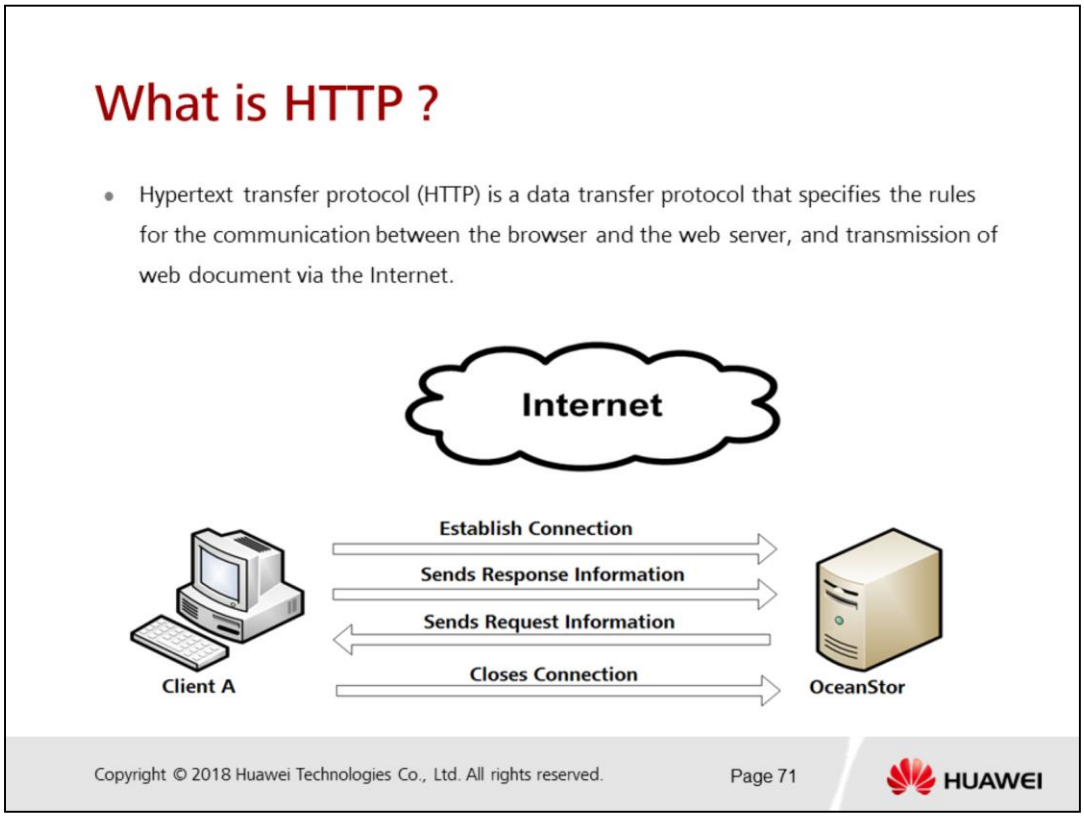
What is FTP ?

- FTP (File Transfer Protocol) is used to transfer files between a remote server and a local host. It is a common protocol for transferring files on an IP network.
- The FTP protocol is an application layer protocol in the TCP/IP protocol suite. It is used to transfer files between a remote server and a local client, and uses TCP ports 20 and 21 for transmission. Port 20 is used to transmit data, and port 21 is used to transmit control messages. The basic operation of the FTP protocol is described in RFC959.
  - FTP works in two different modes:
    - Active mode (PORT): The FTP server initiates a connection request when a data connection is established. This is not applicable when the FTP client is in the firewall (for example, the FTP client is on a private network).
    - Passive mode (PASV): The FTP client initiates a connection request when establishing a data connection. This is not applicable when the FTP server restricts the client from connecting to its upper port (generally greater than 1024).

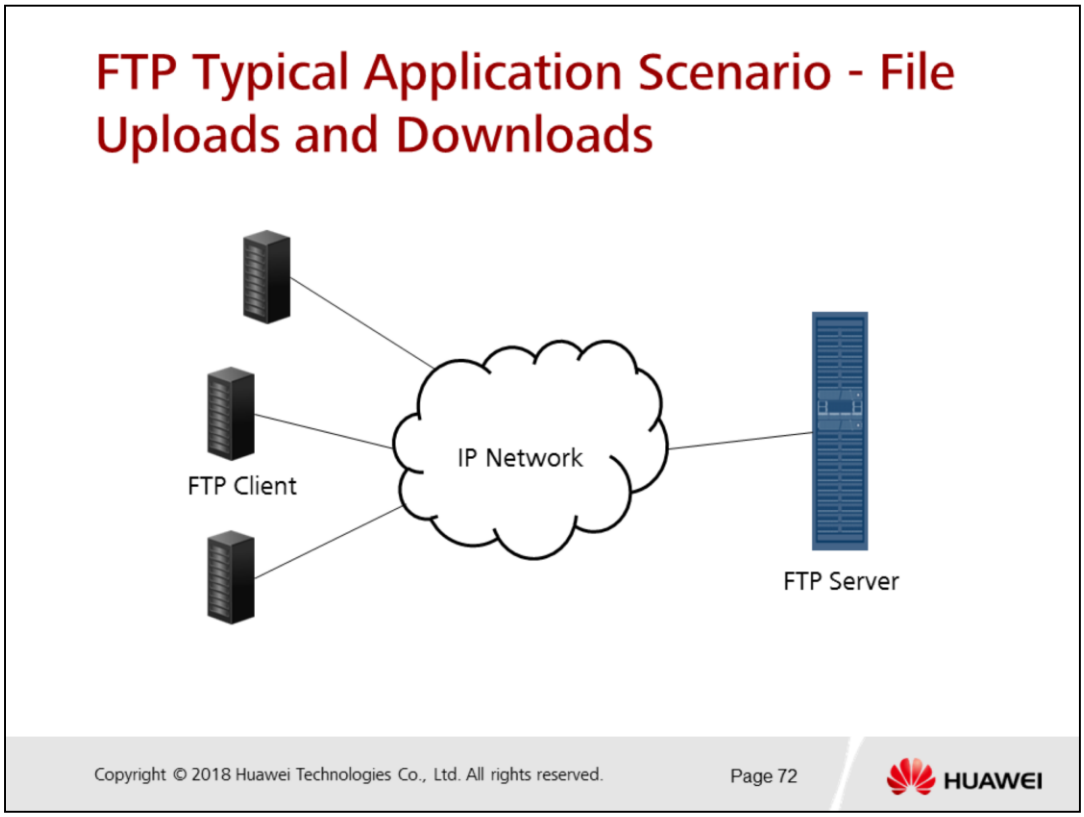
Copyright © 2018 Huawei Technologies Co., Ltd. All rights reserved.

HUAWEI

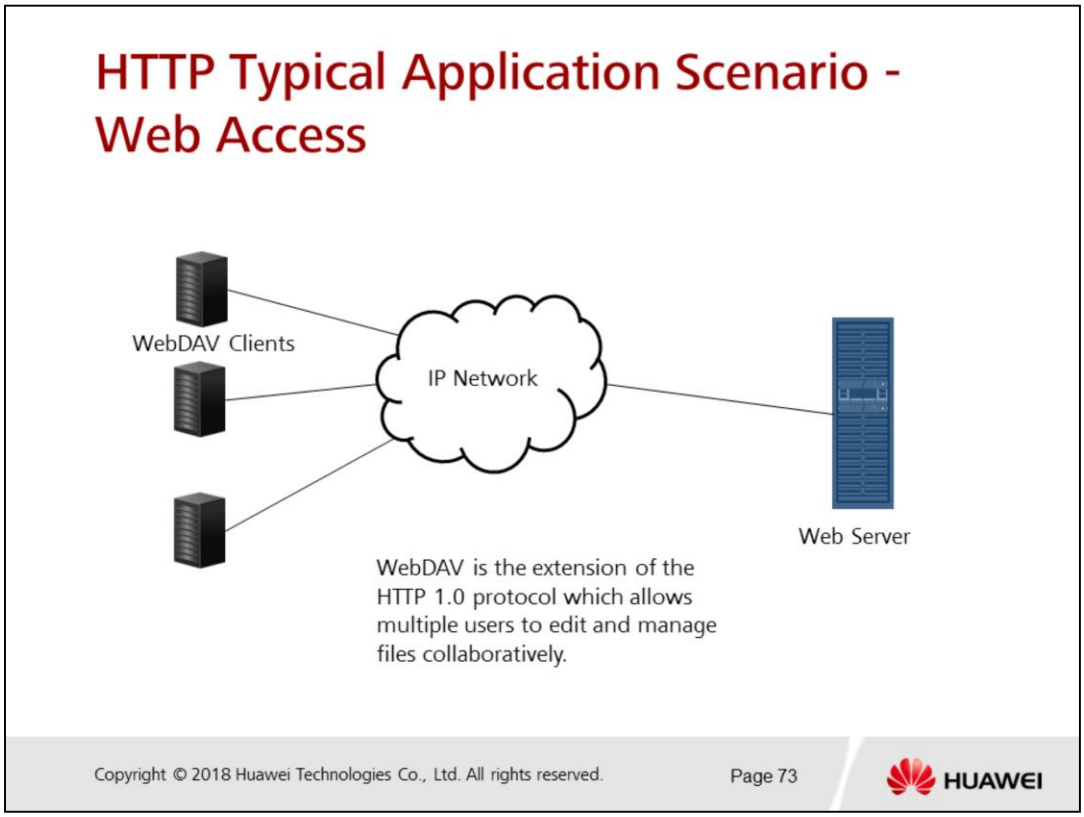
- FTP is the abbreviation of File Transfer Protocol. It is used for controlling the two-way transmission of files on the Internet. At the same time, it is also an application. There are different FTP applications based on different operating systems, and all these applications follow the same protocol to transfer files.
- In the use of FTP, users often encounter two concepts which are: "Download" and "Upload". "Downloading" a file means copying a file from a remote host to your computer and “Uploading" a file means copying the file from your computer to a remote host. In popular Internet terms, users can upload (download) files to (from) remote hosts through client programs.
- FTP has two file transfer modes:
  - Binary mode for transferring program files (e.g. files with the suffixes .app, .bin, and .btm).
  - ASCII mode for transferring text files (such as files with the suffixes .txt, .bat, and .cfg).



- The HTTP protocol defines how a Web client requests a Web page from a Web server and how the server delivers the Web page to the client.
- HTTP uses a short connection for packet transmission and the connection is interrupted/disconnected after each transmission.
- Different from the CIFS and NFS protocols, through the use of WebDav extension protocol and the mounted file system, it enables HTTP share creation, and the upload, download, modification, and locking of resource files.
- HTTPS is a secure version of the HTTP protocol and is based on SSL/TLS security encryption. SSL (Secure Sockets Layer) and its successor Transport Layer Security (TLS) are security protocols that provide security and data integrity for network communications. TLS and SSL encrypt the network connection at the transport layer.



- FTP is used to transfer files between the remote server and the local host.
- The File Transfer Protocol (FTP) is a standard network protocol used for the transfer of computer files between a client and server on a computer network.
- FTP is built on a client-server model architecture and uses separate control and data connections between the client and the server. FTP users may authenticate themselves with a clear-text sign-in protocol, normally in the form of a username and password, but can connect anonymously if the server is configured to allow it. For secure transmission that protects the username and password, and encrypts the content, FTP is often secured with SSL/TLS (FTPS). SSH File Transfer Protocol (SFTP) is sometimes also used instead but it is technologically different.



- The WEB server is also called a WWW (World WIDE WEB) server and provides users with online information browsing services through the HTTP protocol. The user sends a request to the server through the browser to browse the resource information on the server.
- Since the Web has become the basis of the Internet, HTTP 1.1 (Hypertext Transfer Protocol) has proven to be a very flexible and versatile protocol for transmitting data.
- However, there are some obvious disadvantages of HTTP, which restricts it from being adopted as a comprehensive Internet communication protocol: it is well-suited for viewing static documents, but it cannot provide enough complexity (to provide rich authoring capabilities to clients) The way to handle documents. For example, when two authors make changes to a document at the same time without communicating, there will be a "lost update" problem. Only changes made by the last author and re-uploaded to the server will remain, and changes made by another author will be lost.
- WebDAV (Web Distributed Authoring and Versioning) extends the HTTP/1.1 protocol and allows clients to publish, lock, and manage resources on the Web collaboratively.

Summary

- This module mainly introduced:
  - The brief definitions of common storage protocols.
  - In depth introduction of the technical principles of common storage technologies.
  - The brief introduction of application scenarios of common storage protocols.

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

Quiz

1. Which of the following are File Sharing Protocols?
   A. HTTP
   B. iSCSI
   C. NFS
   D. CIFS
2. Which of the following are FC Topologies?
   A. Arbitration Loop Network.
   B. Point-to-Point Network
   C. Switched Network.
   D. Dual-Switched Network.

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

- Answers:
  - ACD.
  - ABC.

Thank You

www.huawei.com