Architecture Evolution and Development of Core Switches in Data Centers

Written By Qu Zhijun

As the Internet continues to develop, the quantity of data has grown explosively, and data management and transmission have become increasingly important. With such vast quantities of data, high data security, centralized data management, reliable data transmission, and fast data processing are required. To meet these requirements, data centers have come into being. During data center construction, core switches play an important role in meeting construction requirements.

Ethernet has been in existence for over 30 years, and over that time, bandwidth has increased from 10 Mbit/s to 100 Mbit/s to 1000 Mbit/s to 10 Gbit/s to 40 Gbit/s, and now to 100 Gbit/s. 10GE Ethernet is currently in widespread use, and 40GE Ethernet and 100GE Ethernet are gradually being adopted.

As data center networks and Ethernet standards rapidly develop, next-generation core switches need to meet data center network construction requirements and adapt to Ethernet standards evolution. To address these challenges, next-generation core switches require a well-designed system architecture that can meet requirements on hardware architecture, high-speed links, heat dissipation, interface density, and interface rate.

Hardware Architecture Development of Core Switches

Core switches undergo three phases of development: providing FE aggregation and GE uplink interfaces in phase 1, providing GE aggregation and 10GE uplink interfaces in phase 2, and providing 10GE/40GE aggregation and 100GE uplink interfaces in phase 3. The three phases require different hardware architectures of core switches.

**Phase 1 (from 2000 to 2006): Core switches provide high-density FE/GE access and aggregation, and a few 10GE uplink interfaces.** In phase 1, core switches use a hardware
architecture consisting of central switching fabrics and LSW chips. Central switching fabrics use Ethernet switching chips, and LSW chips that provide FE/GE and a few 10GE interfaces are used as interface chips on line cards. Backplane links provide bandwidth of 1.25 Gbit/s to 6.25 Gbit/s. Line cards provide 48GE line-rate forwarding and process Layer 2 and Layer 3 services. Simple priority-based queue scheduling is used for QoS. Typical core switches in phase 1 include Cisco 4500/6500 and H3C 6500/7500.

Figure 1 Core switch hardware architecture in phase 1

Central

1.25Gbit/s to 6.25Gbit/s

FE LSW

GE LSW

10GE

Core switches in phase 1 apply to the following scenarios:
- FE/GE access and a few 10GE uplink interfaces are required.
- A few GE servers comprise a small data center network.

Phase 2 (from 2006 to 2012): Core switches provide high-density GE/10GE access and aggregation, and a few 10GE/40GE uplink interfaces. In phase 2, central switching fabrics have many models, including simple switching fabrics sharing the buffer, switching fabrics with central arbitration, and CLOS switching fabric at a rate of 6.25 Gbit/s. LSW chips that provide GE/10GE interfaces and a few 40GE uplink interfaces are used as interface chips on line cards. Backplane links provide bandwidth of 5 Gbit/s to 10 Gbit/s. The bandwidth of a slot on each line card is less than 480 Gbit/s. Line cards provide a maximum of 16–48*10GE line-rate forwarding and process Layer 2 and Layer 3 services. Core switches use simple HQoS scheduling, and have variable caching capacities depending on LSW chips, for example, 2 ms per interface and 10 ms per interface.

Figure 2 Core switch hardware architecture in phase 2
Core switches in phase 2 apply to the following scenarios:

- GE access and 10GE uplink interfaces are required.
- GE servers comprise small- and medium-scale data center networks.

**Phase 3 (from 2012 to 2020):** Core switches provide high-density 10GE/40GE access and aggregation, and a few 10GE uplink interfaces. In phase 3, core switches use CLOS dynamic routing and switching fabrics. Line cards use PP chips that can process complex services, support high-density 10GE/40GE line-rate forwarding and service processing, and provide a few 100GE uplink interfaces. Backplane links provide bandwidth of 10 Gbit/s, which can be increased to 25 Gbit/s in the future. Line cards provide 1 Tbit/s to 4 Tbit/s of bandwidth per slot, support a maximum of 48–96*10GE or 24*40GE line-rate forwarding, have comprehensive QoS capabilities, and provide a large caching capacity of 100 ms per interface. A typical core switch in phase 3 is Huawei CE12800.

Figure 3 Core switch hardware architecture in phase 3

Core switches in phase 3 apply to scenarios where high-density 10GE/40GE access is required. 10GE servers are currently in widespread use, but will gradually be replaced by 40GE
servers by 2015. Therefore, data center core switches in phase 3 must be capable of providing high-density 10GE/40GE interfaces to meet requirements of large- and medium-scale data center networks comprising 10GE/40GE servers.

Huawei CE12800 is the flagship next-generation core switch that provides high-density 10GE/40GE interfaces, 1 Tbit/s to 4 Tbit/s of bandwidth, comprehensive QoS capability, and a large caching capacity of at least 100 ms per interface.

The following table lists the characteristics of core switches in all three phases.

<table>
<thead>
<tr>
<th></th>
<th>Phase 1</th>
<th>Phase 2</th>
<th>Phase 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interface density</td>
<td>48FE to 48GE</td>
<td>48GE to 48*10GE</td>
<td>48<em>10GE to 48</em>100GE</td>
</tr>
<tr>
<td>Slot bandwidth</td>
<td>&lt; 128 Gbit/s</td>
<td>&lt; 480 Gbit/s</td>
<td>1 Tbit/s to 4 Tbit/s</td>
</tr>
<tr>
<td>QoS capability</td>
<td>Low</td>
<td>Medium</td>
<td>High</td>
</tr>
<tr>
<td>Typical product</td>
<td>H3C 7500/9500</td>
<td>Cisco N7000</td>
<td>Huawei CE12800</td>
</tr>
<tr>
<td></td>
<td>Cisco 6500</td>
<td>H3C 12500/10500</td>
<td></td>
</tr>
</tbody>
</table>

End-to-End High-speed Link Development of Core Switches

Key performance indicators for core switches are high interface density and high bandwidth, which are implemented by high-speed links. If the backplane of a core switch uses 10G links, a board can provide 48*10GE line-rate forwarding. If the backplane of a core switch uses 25G links, a board can provide 48*100GE line-rate forwarding. Figure 4 shows which components comprise an end-to-end high-speed link on a core switch.

Figure 4 Components comprising an end-to-end high-speed link on a core switch
As shown in Figure 4, the high-speed link signal is outputted from the chip, travels through the line card PCB, backplane connector, backplane PCB, backplane connector, and line card PCB, and is then received at the input end of the chip. Reducing end-to-end cabling length and minimizing interference from backplane connectors can help improve high-speed link performance during signal transmission.

**Reducing end-to-end cabling length**

Figure 5 illustrates the impact of channel attenuation on the bit error rate, and the need to reduce end-to-end cabling length.

![Figure 5 Impact of channel attenuation on the bit error rate](image)

In Figure 5, IL indicates the channel attenuation (dB) of signals, and BER indicates the bit error rate (BER). Cabling length determines the attenuation of a high-speed channel. When the
BER is fixed, a shorter channel can provide a higher signal rate. When the signal rate is fixed, a shorter channel ensures a lower BER.

On a core switch, the cabling length includes the line card cabling length and backplane cabling length. If the backplane cabling length is reduced to 0, the end-to-end cabling length is minimized. Figure 6 shows the orthogonal architecture of a next-generation core switch, in which the backplane cabling length is 0.

Figure 6 Orthogonal architecture of a next-generation core switch

![Orthogonal Architecture of a Next-Generation Core Switch](image)

Line cards are installed only on switching fabrics of a core switch using orthogonal connectors. End-to-end cabling length is minimized because no connection is needed to the backplane. This ensures channel availability in the case of higher bandwidth and allowing for a higher signal rate.

**Minimizing interference from backplane connectors**

Interference from backplane connectors can severely degrade the performance of a high-speed link. To ensure performance of a high-speed link, interference from backplane connectors should be less than 35 dB when the bandwidth is between 10 Gbit/s and 25 Gbit/s.

Interference from backplane connectors is determined by the distance between connectors' wafers. Generally, the distance between wafers on a mainstream backplane is approximately 2 mm, which cannot meet interference requirements of high-speed backplane connectors. Additionally, next-generation core switches need to support a higher signal rate. To meet these requirements,
next-generation backplane connectors are required since existing backplane connectors cannot support the ongoing development and evolution of core switches. Currently, mainstream core switches are designed based on existing backplane connectors. A high-speed link provides a maximum rate of 10 Gbit/s, which cannot be increased in the future and cannot reach 48*40GE/100GE.

Huawei CE12800 uses a high-performance orthogonal architecture and next-generation orthogonal high-speed backplane connectors. Huawei and industry-leading connector vendors have developed these next-generation backplane connectors to increase the distance between wafers to 2.7 mm. Huawei CE12800 supports the evolution of backplane links from 10G links to 25G links and can provide up to 4 Tbit/s slot bandwidth.

**Heat Dissipation System Evolution of Core Switches**

The heat dissipation system of core switches is developing from traditional left-to-right and optimized front/rear-to-rear ventilation channels to a strict front-to-rear ventilation channel. For a core switch providing FE/GE interfaces and non-line-rate 10GE forwarding, the board power is low, so the left-to-right or front/rear-to-rear ventilation channel can meet heat dissipation requirements. For a core switch providing high-density 10GE/40GE line-rate forwarding, the board power is high, so a strict front-to-rear ventilation channel is required.

Figure 7 illustrates the heat dissipation design and requirements in data center equipment rooms. To meet these heat dissipation requirements, core switches must use the front-to-rear ventilation channel.

Figure 7 Front-to-rear ventilation channel design, isolating cold air from hot air channels
During data center construction, energy consumption and emissions need to be reduced. To improve heat dissipation efficiency and reduce power consumption in equipment rooms, ensure that cold air channels are isolated from hot air channels. Heat exchangers blow cold air into the device from the front side and blow hot air out from the rear side. Therefore, core switches in data center equipment rooms must use the front-to-rear ventilation channel to meet heat dissipation requirements.

**Traditional left-to-right ventilation channel**

Traditional core switches often use the left-to-right ventilation channel for heat dissipation. When these core switches are mounted in cabinets, hot air recirculates into the cabinets because the left-to-right ventilation channel is used, resulting in low heat dissipation efficiency. Heat dissipation requirements of high-power boards cannot be met; only traditional core switches with power lower than 300 W can meet heat dissipation requirements. Figure 8 shows the traditional left-to-right ventilation channel on a core switch.

*Figure 8 Traditional left-to-right ventilation channel on a core switch*
As shown in Figure 8, when a core switch using the traditional left-to-right ventilation channel is mounted in a cabinet, hot air recirculates into the air intake vent along the side walls of the cabinet, increasing the temperature of the air intake vent by at least 15°C. If the core switch runs for a long period, system reliability and board heat dissipation capability are affected, and heat dissipation requirements of high-power, high-density 10GE/40GE boards cannot be met.

Therefore, cold air channels must be isolated from hot air channels, and air intake vents must be isolated from air exhaust vents to prevent cold air from mixing with hot air. This ensures a lower air intake vent temperature, improved heat dissipation efficiency, and lower power consumption. Core switches using the traditional left-to-right ventilation channel cannot meet data center construction requirements. In addition, on campus networks, these core switches cannot meet heat dissipation requirements for high-power boards.

Core switches using the traditional left-to-right ventilation channel include H3C S10500/9500E/7500E, and Cisco N7018/7009.

**Optimized front/rear-to-rear ventilation channel**

To isolate cold air channels from hot air channels, some vendors change the ventilation channel of the core switch system to front/rear-to-rear, as shown in Figure 9.

Figure 9 Optimized front/rear-to-rear ventilation channel on a core switch
As shown in Figure 9, the optimized ventilation channel design blows cold air into the device from the front and rear of the device, and blows hot air out from the rear. However, this design still cannot meet heat dissipation requirements in data center equipment rooms. This is because some boards on the device still require cold air to be blown into the device from the rear for heat dissipation. Consequently, cold air mixes with hot air, and the device is therefore unable to meet heat dissipation requirements of high-density boards.

**Strict front-to-rear ventilation channel**

Left-to-right and front/rear-to-rear ventilation channel designs cannot meet heat dissipation requirements in data center equipment rooms. Therefore, next-generation core switches use the strict front-to-rear ventilation channel design to isolate cold air from hot air channels. This design improves the heat dissipation efficiency of core switches, provides a heat dissipation capability of 1000 W per slot, and supports high-density, high-bandwidth 48*40GE/100GE boards. Figure 10 shows the strict front-to-rear ventilation channel on a next-generation core switch.
The strict front-to-rear ventilation channel blows cold air into the device from the front, blows hot air out from the rear, and uses non-blocking ventilation. This design improves heat dissipation efficiency, with measurement data showing that the non-blocking ventilation channel can improve heat dissipation efficiency by 30%.

Huawei CE12800 uses the strict front-to-rear non-blocking ventilation channel to meet heat dissipation requirements of high-density 40GE/100GE boards. This high heat dissipation efficiency allows Huawei CE12800 to operate at high temperatures (up to 45°C) for long periods. If the ambient temperature of data center equipment rooms is increased to 45°C, power consumption in equipment rooms can be reduced by two-thirds.

**Interface Rate and Density Evolution of Core Switches**

Core switches are developed based on Ethernet standards. As the Ethernet interface rate has increased from 10M to 100GE, next-generation core switches must be capable of supporting high-density 40GE and 100GE interfaces.

**Figure 11 Evolution of Ethernet standards**
Next-generation core switches support 40GE and 100GE interfaces in addition to FE, GE, and 10GE interfaces, and support high interface density (24 ports and 48 ports). Line-rate forwarding on 24*40GE, 48*40GE, and 24*100GE interfaces is also supported.

Huawei CE12800 is a core switch designed to provide high-density 100GE interfaces. The first CE12800 version (which has an interface density six times that of the industry average) supports 24*40GE line cards. In addition, CE12800 uses advanced architecture, supports 25 Gbit/s bandwidth, and will provide 48*40GE, 24*100GE, and 48*100GE interfaces in the future.

**Huawei CE12800, a Model Next-generation Core Switch**

Huawei CE12800 is designed for data center networks comprising 10GE/40GE servers, and is the core device used to construct large or super large data center networks. CE12800 provides bandwidth of up to 4 Tbit/s, uses a strict front-to-rear ventilation channel, has a heat dissipation capability of 1000 W per slot, and provides a large buffering capability of 100 ms and comprehensive QoS capability. CE12800 is a model next-generation core switch which meets network construction requirements in the cloud computing era, and it will become an industry-leading core switch within the next ten years.