Data-Center-Oriented Switch Fabric of New Generation

The switch fabric is the core of network equipment and is as vital as the human heart. As a comprehensive application-oriented business platform and the core basic architecture for future cloud computing, data centers have stricter and more extensive requirements on the switch fabric of network equipment. This raises the Ethernet from a traditional "Best effort" network to a more mature "Lossless" network. How will the switch fabric evolve?

I. Introduction

The high-speed development of networks -- especially new generation data centers and cloud computing -- imposes higher requirements on switches’ switch fabric and the routers that form the basic architecture of the Internet. A unified switch fabric with the following features is required: business integration, larger capacity and bandwidth, higher performance and scalability, greater QoS assurance, higher reliability and error-tolerance performance, intelligent and easy management, energy saving, able to realize increasing requirements for new services and applications, improves user experience, and reduces the cost of bandwidth per unit. In turn, progress with network equipment will further promote the popularization and solid development of the Internet.

Switch fabric determines the key attributes of a piece of equipment such as its capacity, performance, scalability and QoS. During the short history of its 20 years or so development there have been different forms of switch fabrics. These include shared bus, shared memory, crossbar matrix and dynamic-route-based CLOS. For large or ultra large capacity rack equipment, depending on the development level of the industry, crossbar matrix or CLOS switch fabrics are usually used, while shared buffering switch fabric is used in line cards as a fully-distributed component for business processing and forwarding.

II. Data Center Requirements for Switch Fabric of New Generation

As an application-oriented comprehensive business platform and the core basic architecture for future cloud computing, the data center’s stricter and more extensive requirements on the switch fabric of network equipment are: Supporting Unified Switch Fabric, Large Capacity and High Scalability, Forwarding Performance, Business Scheduling and Intensified QoS, and Resiliency.

1. Supporting Unified Switch Fabric

Currently there are three relatively independent networks in data centers: Data Network, Storage Area Network (SAN) and High Performance Calculation (HPC). To achieve easier business integration and service availability, simpler management, lower construction costs, and reduced operation and maintenance costs in the future, the three networks will gradually merge. The switch fabric of network equipment must be readily scalable and support FcoE, FC interfaces and their forwarding mechanisms to allow seamless integration with a SAN. It must support new interfaces of the Ethernet, such as CEE (Convergence Enhanced Ethernet), to allow the Ethernet to evolve into a more mature...
“Lossless” network from the traditional “Best-effort” network.

2. Large Capacity and High Scalability

A super bandwidth epoch is coming. Video streams, audio streams, social networks, P2P, and multimedia through applications such as Youtube, iTunes, Facebook, GoogleEarth, Telepresence and mobile video are developing at an annual growth rate of 70%, thus imposing nearly unlimited requirements for bandwidth. A switch must have a high capacity and excellent scalability. As the above applications develop, port numbers, port rates, and port capacity should increase accordingly. New port types are also required to support features such as the virtualization of network resources and cluster systems.

As a key index for the switching capability and future scalability of a system, the switching capacity of a switch is similar to the emission index of an automobile. The switching capacity of a rack switch in a new generation data center is between 1~10Tbps, while that of a cluster system can be as large as tens of Tbps. Port capacity is derived from the port rate multiplied by the relevant linear speed port number and reveals the linear-speed forwarding capability that can be currently supported in a given product. Different versions of products with the same switching capacity may have different port capacities at different stages. Products with the same switching capacity may have different port capacities to support due to different total overheads of the switch fabric.

Port rate: The fabric of a new generation is required to support not only 1G and 10G Ethernet ports, but also to smoothly support one or more 40Gbps and 100Gbps ports in each slot. This is a qualitative leap in the process of bandwidth development.

3. Forwarding Performance

Linear-speed forwarding performance: Usually this refers to 64-byte small packet linear-speed forwarding capability, revealing the system’s capability to process packet heads. given a certain port flow, 64-byte small packets require the system to process more packets in a given unit of time. In terms of forwarding performance, the consistency of linear speed must be ensured. Linear speed must be maintained for both small and large packets to prevent packet loss. Both the Pair and Full Mesh modes are capable of linear-speed forwarding.

Forwarding delay and delay jitter: Currently, port to port delay in store forwarding technology ranges from a few ms to tens of ms, which is satisfactory in most application scenarios. Cut-through forwarding delay can be less than 1 ms, which is suitable for highly delay-sensitive tight coupling and high performance calculations. Delay jitter refers to delay consistency and delay predictability. Real-time services such as VoIP and video require low delays and delay consistency.

4. Business Scheduling and Intensified QoS

Over the last few years, requirements for bandwidth have increased by 50~70% annually, whereas the increase of bandwidth availability has been just 30%. Limited resources cannot give all users and services with sufficient bandwidth, causing network congestion. Network equipment should provide more rigorous QoS. Depending on the different SLA requirements of different users for different services, indices should include guaranteed or predictable bandwidth, packet loss rate, burst buffering capability, delay, and delay jitter.

Business Scheduling & Queuing: A switch fabric without scheduling & queuing is like a junction without traffic lights: collisions and other accidents will result. Like a junction with a lane running in each direction and bi-directional traffic lights, extensive scheduling is better than a junction without traffic lights, but may still be subject to congestion. Intensive scheduling, on the other hand, is obviously more efficient and orderly, as it resembles a junction with three lanes (left turn, straight, and right turn) and tri-directional traffic lights.

In a switch, the lanes become queues and the scheduler serves as the traffic lights. More queues mean that more intensive scheduling and management can be performed over traffic so that different priority services from different egresses can be forwarded with no mutual influence or head-on congestion. More queues result in a more complex scheduler and design. Some equipment also provides hierarchical scheduling (H-QoS). The number of queues that can be supported is one of the key indices for network equipment. Common equipment supports about a dozen to hundreds of queues, while a few high-end products support as many as 1K to tens of Ks of queues.

Stream Classification & Buffering: Closely associated with business scheduling is stream classification & buffering. Stream classification allows the identification of different users and businesses, and then maps them to different priorities and queues. Without buffering or with very little buffering, scheduling is of no use or it use is greatly reduced. Since application is becoming more complex and traffic bursts are both bigger and more frequent (e.g. live search), it is vital for new generation data centers to have a large enough buffering mechanism.

5. Resiliency of Switch Fabric

Resiliency automatically detects and separates a fault so that the
system’s functional performance suffers little or no loss (Graceful Degradation). Resiliency may include Redundancy and Fault Tolerance. Resiliency can be further improved by using N+1 switching network boards, which are physically independent of the master control board, namely, the physical separation of the forwarding plane and the control plane.

III. Traditional CIOQ-based Crossbar Switch Fabric

CIOQ-based Crossbar switch fabric came into being in the 1990s. As shown in Fig. 2, the fabric contains one or more Crossbar chips that work in parallel without buffering. Through switching the network port FP (Fabric Port), each Crossbar chip is connected to the FA port corresponding to all input ports, and is connected to the cross FA port that all output ports correspond to. For scheduling, a centralized Arbitrator connected to all input/output FA chips and Crossbar chips is usually used. The output FA port reports egress congestion to the Arbitrator periodically. A typical switching process will involve three steps: (1) Before the input port transmits packets, the ingress FA port will request that the Arbitrator transmits packets (Request to transmit); (2) Depending on queue congestion at the output port, the Arbitrator will send Request granted to the FA port; (3) Through the switching network, packets are transferred to the output port.

In the ingress direction, in VoQ (Virtual output Queuing) mode, the buffer allocates relevant queues to output ports with different destinations and business streams of different priorities so as to buffer ingress flow. In the egress direction, there is also a buffer, which is used to absorb any burst flow from the switching network. Thus, it is called CIOQ (Combined Input Output Queuing).

Since scheduling is centralized, the Arbitrator scheduling algorithm is highly complex and performance scalability is poor. When system capacity is large, the scheduler is likely to become a bottleneck, thus making accurate scheduling difficult. Extensive scheduling requires a relatively large buffer in the egress direction to perform further scheduling to support system-level QoS at small granularity. To make full use of the egress buffer, the system acceleration ratio should be increased as high as 1.6~2. Increasing the acceleration ratio decreases the effective port capacity that the system can support (acceleration ratio refers to a ratio between the switching network port rate and the actual network port rate).

In the geometric topology of the switch fabric of some products, multiple Crossbars are connected in a form similar to the CLOS fabric, and static routing is used. Before a business stream enters the switching network, a path is designated according to the source port or selected on the basis of the Hash algorithm. So, all packets belonging to the same stream choose the same path to enter the switching network. Evidently, when there is a single business stream in the system, the path selected by the Hash algorithm is likely to cause congestion, while other paths are relatively idle. Similarly, when a business stream switches from level 2 to level 3, congestion occurs easily. This fabric is not a congestion-free CLOS switch fabric and its switching performance is equivalent to a CIOQ-based Crossbar.

To satisfy both large capacity switching and good quality business scheduling, the CIOQ-based Crossbar is the optimum switch fabric. Switching capacity varies from hundreds of Gs to a few Ts and usually supports 10G interfaces, but not 40G and 100G interfaces. As switching capacity is not very large, the switching network is usually integrated with the master control board in 1+1 active/standby mode or load sharing mode. Currently, 10G platform-based high-end rack switching equipment usually uses this fabric. Typical examples are H3C S9500 and Cisco C6500.

IV. Dynamic Routing-based CLOS Switch Fabric of a New Generation

CLOS switch fabric was proposed for the first time by Dr. Charles Clos of the Bell Laboratory in his 1953 paper, Study on Switching Networks. It has since been widely applied in TDM networks. Over the last nearly 20 years as packet switching networks have rapidly developed, a super capacity switch fabric with excellent scalability has been in urgent demand. As a result, CLOS -- an
Data-Center-Oriented Switch Fabric of New Generation

old and new technology -- has gained in importance.

CLOS switch fabric is a multi-level fabric where each switching unit at every level is connected with all the switching units on the next level. A typical 3-level CLOS switch fabric is defined by two parameters: k and n, as shown in Fig. 3. Parameter k is the number of switching units at the middle level, while n is the number of switching units at the 1st level (3rd level). The 1st and 3rd levels are composed of $n \times k$ switching units, and the middle level is composed of $k \times n$ switching units. All these form a $k \times n$ switching network; that is, the network provides $k \times n$ input and output ports.

For a switching network requiring more capacity, the middle level can be a 3-level CLOS network (i.e. a CLOS network can be recursive). For example, 4 $n \times n$ chips of 1st (3rd) level plus 2 $n \times n$ chips on the 2nd level can form a $2n \times 2n$ switching network. Its recursive property theoretically imbues the CLOS network with incomparable scalability, thus supporting smooth expansion (scaling) of the number of switch ports, port rate, and system capacity.

CLOS switch fabric can become literally Non-blocking, Re-arrangeable, and Scalable.

CLOS fabric defines a geometric topology. In the early TDM and voice application, its re-configurability is usually accomplished by software calculation and configuration. For a high-speed packet switching system in which high quantities of business stream destination ports vary frequently and quickly (such as ns level), it is not practical to select and reconfigure forwarding paths with software. Given this, dynamic routing has been specially designed over the last few years for the CLOS fabric used in packet switching systems.

The key point of dynamic routing is to use all accessible paths with balanced load sharing. For the 1st level, each business stream can be evenly transmitted to k paths on the 2nd level by Round-robin or randomly (usually cell-based transmission).

Business streams reaching the 2nd level are switched to 3rd level destination ports through related paths that are selected according to switching network routes based on Cell-based Self-routing. Upon receipt of all 2nd level cells, the 3rd level reorganizes them into packets and ensures correct packet sequencing. Dynamic routing has achieves a non-blocking switch, which helps to reduce the acceleration ratio and give a larger, more effective port capacity.

Dynamic routing features a prominent advantage - it smoothly supports higher-rate network ports, such as 40GE/100GE. This is because it can make full use of all available paths to form a large data stream path, e.g. 24 paths of 3.125Gbps support 100GE data streams. Conversely, static routing is limited by single path bandwidth, e.g. the x AUI interface-based Crossbar switch allows a maximum network port rate of 10Gbps, which fails to support 40GE and 100GE.

The CLOS fabric is based on dynamic routing and, in collaboration with a suitable business scheduling mechanism, supports perfect QoS. Typical equipment using CLOS switch fabric includes: H3C S12500 unified switch fabric core switch and Juniper T1600 core router. At the beginning of February 2009, Juniper advertized its Tx -Matrix Plus and, with the help of the multi-frame interconnection technique, converted 16 T1600 processors into a 25Tbps non-blocking switching system, thus demonstrating the excellent scalability of CLOS fabric. In 2004, Cisco launched its flagship router, CRS-1, which used 3-level dynamic self-routing Benes switch fabric to interconnect 72 racks and realize a system capacity of 46T/92T. The Benes switch is essentially a special type of CLOS switch fabric.

Since the capacity of the CLOS switching system is large, N+1 independent switching network slots are usually used to thoroughly and physically separate the control plane of the master control board. This increases system capacity and scalability and greatly improves the reliability of the forwarding plane, thus preventing possible faults or handover on the control plane from affecting the forwarding plane.

V. H3C S12500 Switch Fabric

H3C S12500 uses dynamic routing-based CLOS switch fabric (which is currently leading the industry, as shown in Fig. 4), a distributed Credit-based business scheduling mechanism, and a large-capacity buffer. These provide flawless QoS and hierarchical business scheduling (H-QoS).
Data-Center-Oriented Switch Fabric of New Generation

In the ingress direction, S12500 supports as many as 48K VoQ queues and large-capacity buffering (average 256MB/10GE). These are mapped onto different VoQ queues according to the business stream destination port, business attributes, and priority to realize business intensive identification and buffering. The ingress FA periodically transmits attributes and the occupying state of the queues sent by each VoQ to the scheduler in the egress direction. Each queue, according to the accumulated number of Credits received from the scheduler, transmits packets of related flows to the switching network and decreases the quantity of remaining Credits depending on the quantity actually transmitted.

In the egress direction, the scheduler schedules all business streams that flow to the egress port according to scheduling policy, available egress port bandwidth, congestion, attributes, priorities, and idle/occupied queues of the streams (VoQ). It allocates Credit for different bandwidths, so that bandwidth can be accurately distributed according to user SLA and QoS guaranteed. Each scheduler schedules business streams that flow to the egress only to form a fully-distributed business scheduling mechanism. Each scheduler corresponds to a Shaper, with the supporting algorithm of Dual Leaky Bucket, and provides a flow Shaping function for shaping each business stream and stream aggregation.

It can be seen that Single-Hop scheduling is more concise, intensive and effective than the CIOQ-based Two-Hop scheduling adopted by Crossbar. Moreover, the required accelerator ratio is much less, which facilitates subsequent port capacity increases.

Scheduling and flow control are end-to-end and not confined to partial congestion control by the switching network. S12500 supports the multi-business scheduling policies of SP/WFQ/FQ. Its combination, SP+WFQ+FQ, supports SP+DWRR and bandwidth reservation for key traffic. It flexibly combines and arranges K scheduler modules to support complex and customized scheduling policies and implements management policies to control congestion on WRED/Tail Drop.

A typical S12500 scheduler module is shown in Fig. 5.

In terms of resiliency design, the S12500 switch fabric uses the switching network N+1 redundancy technique to support a maximum of 9 network boards (each board contains 1 or 2 Fabric switching units). Each business stream is evenly borne by all switching network units. It is also capable of automatically detecting and separating faulty devices and faulty paths (links), thus giving high reliability, resiliency and fault tolerance performance.

The advanced dynamic routing-based CLOS switch fabric, Cell-based self-routing, and distributed scheduling provides the S12500 with ultra-large capacity, full non-blocking forwarding performance, excellent scalability, high-QoS and resiliency. Currently, the S12500 can provide a system switching capacity of up to 6.66Tbps and smoothly support 40GE and 100GE ports. The switching capacity of a single piece of equipment can be
scaled up to 13.32Tbps. Multi-frame (e.g. a centralized switching frame) interconnection/scaling allows the system capacity to be scaled up to 26T/52T/104T or higher, which far exceeds any other switch fabric.

**VI. Summary**

For high-end rack switches and routers, the CIOQ-based Crossbar switch fabric and CLOS switch fabric are considered the optimum fabrics. Of the two, the dynamic routing-based CLOS switch fabric combines Cell-based self-routing and distributed scheduling. It is currently the most advanced switch fabric, and thus ideal for complex applications such as new generation data centers, cloud computing, and large-capacity core switches and core routers.