Non-blocking Switching in the Cloud Computing Era

By Hu Zhongfeng

Foreword:

Data centers are at the core of the cloud computing era that is now beginning. How data centers can better support the fast growing demand for cloud computing services is of great concern to data center owners. Going forth, customers will construct larger data centers, purchase more servers with higher performance, and develop more applications. If data center networks cannot adapt to these changes, they will become bottlenecks to data center growth.

1 Networks Must Go With the Flow in the Cloud Computing Era

The data center is the core of a cloud platform. Growing numbers of services and applications are being deployed, and demand for additional related services and applications is driving data center construction.

The service and resource planning needs of cloud computing data centers differ significantly from those of traditional data centers. Data center networks must change to meet the needs of cloud computing data centers. Changes to the traffic model used by cloud computing data centers have been especially profound and this has created new demands for data center networks.

It is estimated that about 70% of traffic in a cloud computing data center is east-west traffic, while about 80% of traffic in a traditional data center is north-south traffic.

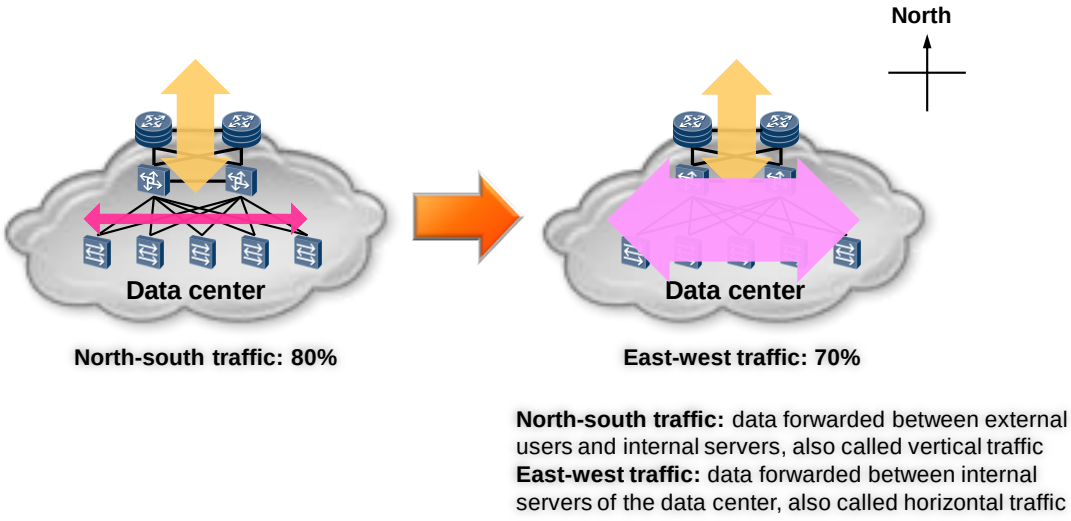

Figure 1 Traffic model evolution in data centers

What is causing these changes to the traffic model?

Traditional data centers mainly provide access for external users, so most traffic moves in a north-south direction. Based on how services are delivered and overall limits on egress bandwidth, bandwidth is allocated to each layer with a convergence ratio between layers. Bandwidth on the access layer is usually several times that of the aggregation layer or the core layer. The common bandwidth convergence ratio ranges from 1:3 to 1:20.

Cloud computing increases demand for both the variety of services available and the volume of service traffic. This increased demand has a great impact on the data center traffic model. Applications such as searching and parallel computing require multiple servers to collaborate, and this increases traffic between servers.

Complex and ever-changing service demands make it difficult to predict traffic on servers or plan network bandwidth. Dynamic virtual machine migration creates complexity and higher east-west traffic volume.

Traditional data center networks are unable to meet the demands of this new traffic model. A cloud computing data center needs a non-blocking network for line-speed traffic transmission between servers within the data center.

2 Fat-tree Architecture Achieves a Non-blocking Data Center Network

The fat-tree architecture developed by Charles E. Leiserson in the 1980s is a widely accepted kind of non-blocking network architecture. This architecture uses a large number of low-performance switches to construct a large-scale non-blocking network.

2. 1 No Bandwidth Convergence in the Fat-tree Architecture

In a traditional tree topology, bandwidth is allocated on each layer and there is a certain convergence rate between layers. The amount of bandwidth at the root is much less than the total sum of bandwidth at all the leaves.

The fat-tree architecture looks more like a real tree. The root has more bandwidth than the leaves. This non-convergence bandwidth planning is the basis for a non-blocking network.

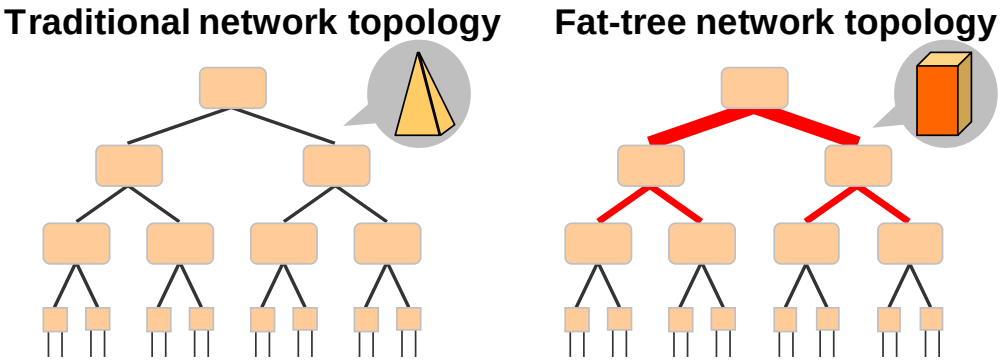

Figure 2 Comparison between a traditional network and a fat-tree network

In the fat-tree network, the uplink bandwidth and the downlink bandwidth on all nodes except the root are the same. Each node on the tree has a line-speed forwarding capability.

Figure 3 shows a 2-element 4-layer fat-tree architecture, in which each leaf switch connects to two terminals and the switches are deployed in four layers. Switches in this networking arrangement all have the same hardware performance specifications.

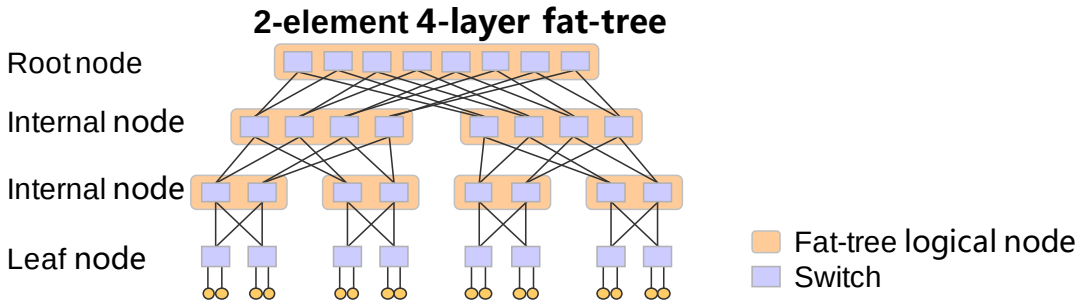

Figure 3 A four-layer fat-tree networking

In Figure 3, each leaf node represents a switch connecting to two terminals. Two switches constitute a logical node in Layer 2 and four switches constitute a logical node on Layer 3. The logical nodes on Layer 2 and Layer 3 are also called internal nodes. There are eight switches on the root node.

The uplink bandwidth and the downlink bandwidth on each logical node are the same. Network bandwidth does not converge.

Only half of the bandwidth on the root node is used for downlink access. Uplink bandwidth is reserved for network scaling in the future, so the network can be easily expanded.

2.2 Fat-tree Needs to be Tailored to Meet Data Center Service Requirements

The root node in fat-tree architecture allows for network scalability by reserving uplink bandwidth equal to the amount of downlink bandwidth in use. The scale of the whole network is usually planned in advance. There are, however, limits (for example, the size of the equipment room) to how much a network can be expanded. The root node does not need to reserve so much uplink bandwidth.

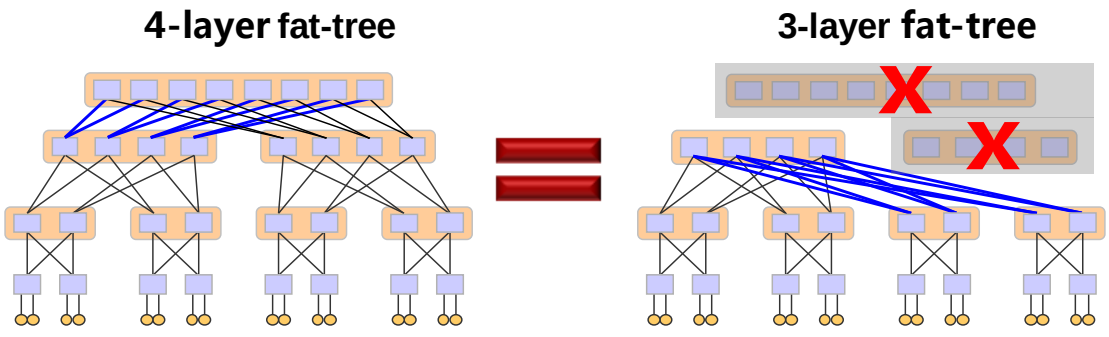

Figure 4 Simplified fat-tree architecture

Figure 4 illustrates a four layer fat-tree architecture that has been simplified to a three layer fat-tree. In the simplified tree, the root is only responsible for non-blocking switching in the network. Bandwidth that was reserved on the four layer fat-tree can also be used for downlink access. The three layer fat-tree requires fewer switches to achieve the same non-blocking switching capability.

In theory, all switches deployed in a fat-tree network have the same hardware performance specifications. In real data center networking, however, access switches only provide services to a few servers and require less forwarding capability than aggregation or core switches. Box-shaped switches are usually used on the access layer, and chassis switches are used on the aggregation layer and core layer.

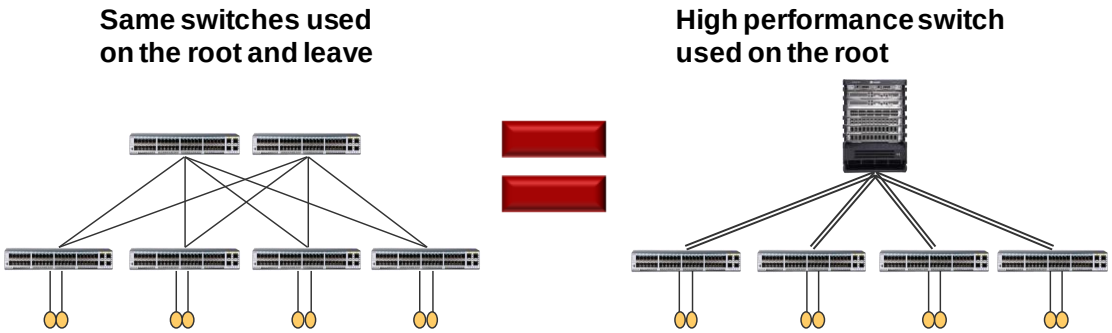

Figure 5 High performance switch used on the root

Figure 5 shows a network that has a high performance chassis switch deployed at the root node. This high performance switch provides larger bandwidth for the servers at the internal nodes, and this networking requires fewer switches. It simplifies network deployment and maintenance, and at the same time improves network performance.

Loops exist in a fat-tree network. Fat-tree networks cannot use Spanning Tree Protocol (STP) technology to implement non-blocking switching because link utilization efficiency is low when STP is used to block redundant links. A technology that makes full use of link resources is required. Routing protocols that run on Layer 3 IP networks and the Transparent Interconnection of Lots of Links (TRILL) protocol that runs on Layer 2 Ethernet networks are commonly used technologies.

3 Huawei Switches Help Build a Non-blocking Network

Huawei's CloudEngine series are next-generation data center switches, including chassis switch CE12800, and box-shaped switches CE6800 and CE5800. These switches support standard TRILL protocol and other widely used routing protocols. They provide GE, 10GE, 40GE, and 100GE access capabilities to support non-blocking switching in data centers.

Figure 6 Huawei next-generation data center switches

1. Each slot of the CE12800 switch supports 24 x 40GE or 96 x 10GE line-speed cards to provide large capacity forwarding capability needed by fat-tree architecture.
2. CE6800 and CE5800 support 40GE uplink interfaces to simplify cable layout and GE or 10GE downlink interfaces to meet a variety of access needs.
3. All CE series switches support standard TRILL protocol. The CE12800 provides Equal-Cost Multiple Path (ECMP) routing and supports a maximum of 32 equal-cost routes, enabling a Layer 2 Ethernet fat-tree network to provide 720 Tbit/s of bidirectional forwarding performance. Networks using CE series switches are highly scalable.

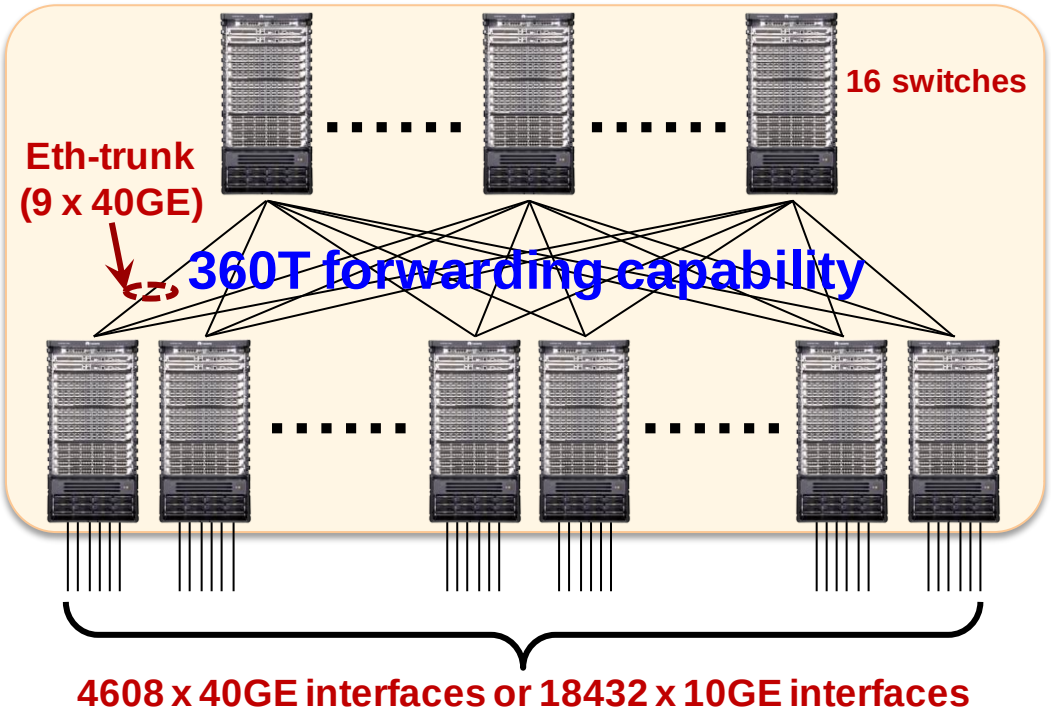

Figure 7 High performance non-blocking fat-tree network constructed with CE12800 switches

Figure 7 shows a high performance non-blocking fat-tree network constructed with CE12800 switches. With the TRILL protocol, a large-scale Layer 2 Ethernet network can be constructed to:

1. Provide 4608 x 40GE line-speed interfaces or 18432 x 10GE line-speed interfaces for external users.
2. Provide 40GE interfaces for internal communication.

Flattened network deployments that use only core switches and TOR switches are the preferred way to reduce network delay in data centers. A fat-tree network can be deployed between the cores switches and TOR switches to form a non-blocking Layer 2 Ethernet network.

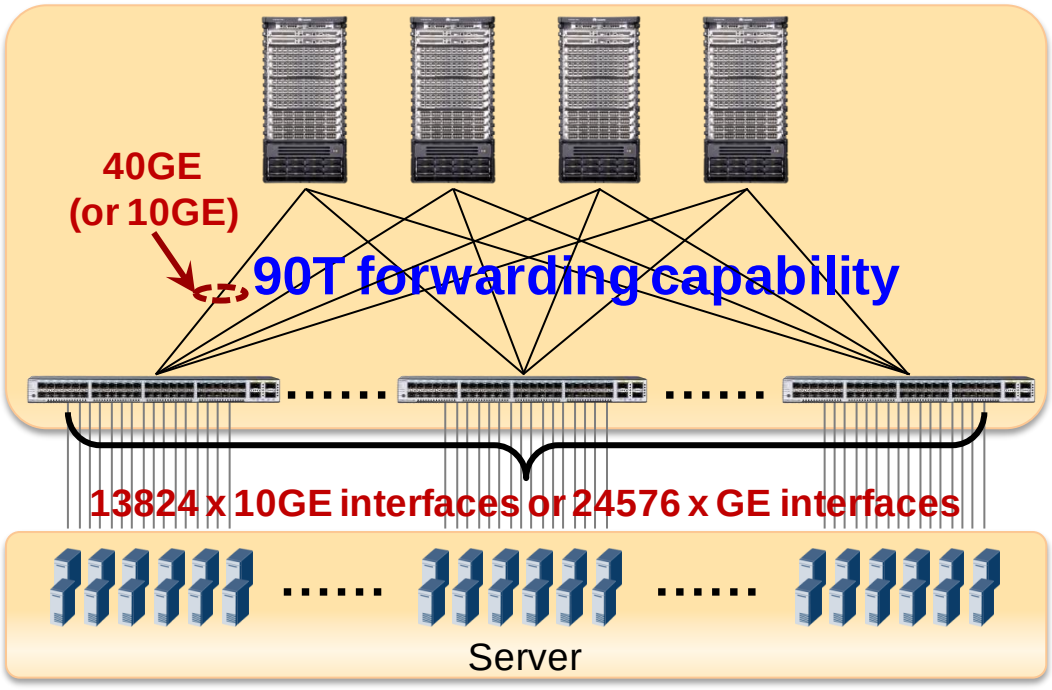

Figure 8 Flattened fat-tree network constructed with CE series switches

Figure 8 shows a flattened non-blocking fat-tree network that is constructed with CE12800 and CE6800 (or CE12800 and CE5800) switches. The flattened network uses TRILL protocol to provide high-density access for 10GE/GE servers.

Because CE12800 supports 32 equal-cost routes and TOR switches support stacking, it is also possible to build a larger flattened fat-tree network that can meet the requirements of large or super large data center networks.

Summary

Cloud computing presents new challenges to data center networks. Non-blocking technology has become the solution of choice to address these challenges. Huawei's CloudEngine series of next-generation data center switches offer a rich mix of service functions and the superior forwarding performance needed to build a non-blocking network for cloud computing data centers.