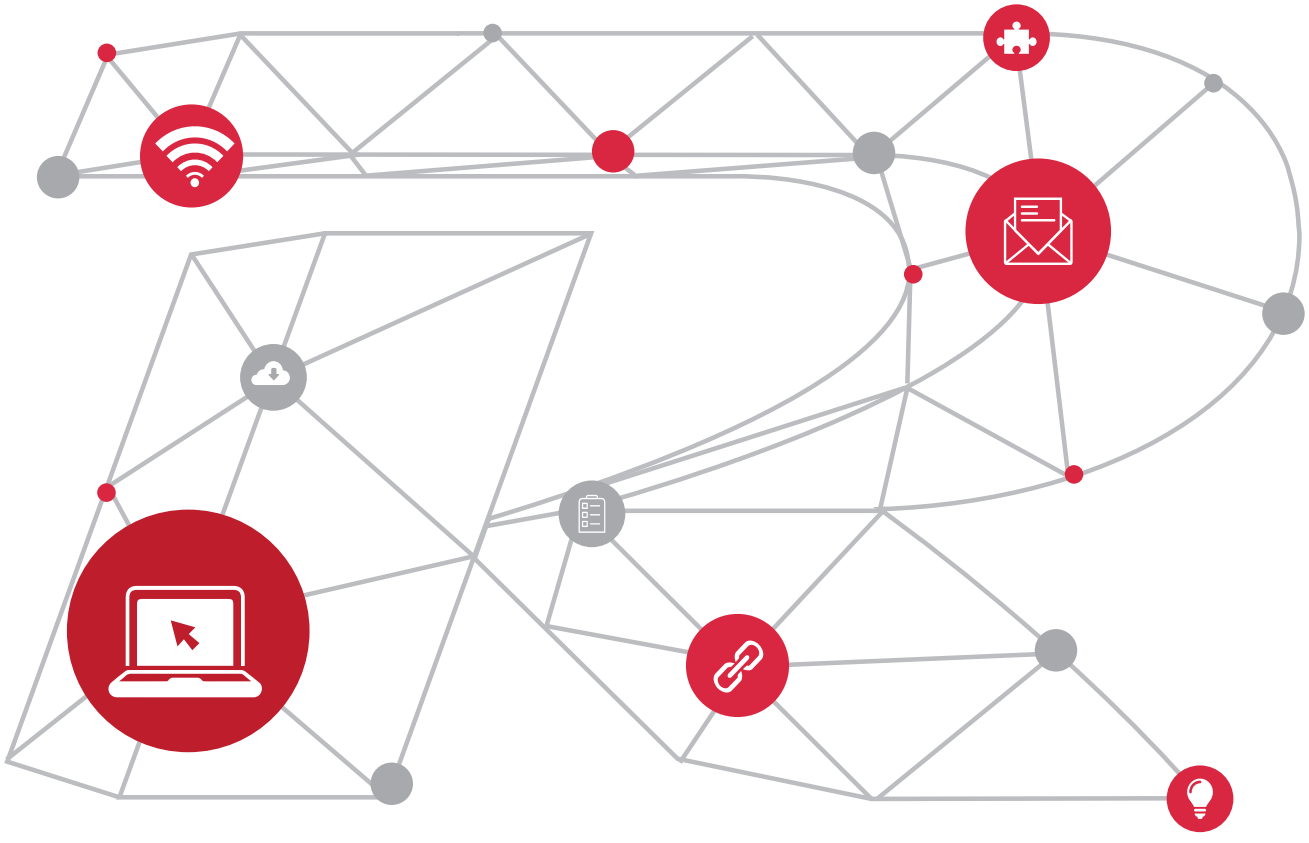
White Paper

A Feast of Technologies
Network Architecture Design for 25G IDCs



Contents

Background 3
Why Upgrade to 25G Ethernets? 3
What Makes Us Choose the 25G Network? 3
Architecture Design Scheme of the 25G Network 4
2-Tier Network Architecture 4
3-Tier Network Architecture 5
Prospects for the Next-Generation IDC Architecture 7

Background

• Why Upgrade to 25G Ethernets?

In the past year, the networks accessed by servers of many Internet data centers (IDCs) had been upgraded from 10G Ethernets to 25G Ethernets. Why did the customers upgrade their networks to 25G?

* Support for high-performance services

The upgrade is needed to support rapid service expansion and performance improvement of the application system. For example, Internet applications based on artificial intelligence (AI) and big data boost exponential growth of service traffic.

* Support for traffic burst

Traffic burst exists in some popular applications. Therefore, the infrastructure on the service side must fully support the traffic burst.

* Matching with server performance upgrade

Performance upgrade of the CPU and storage I/O requires a higher network throughput for each server. Therefore, the 10G network no longer meets the bandwidth requirements.

* Reduced single-bit cost

For public cloud services, 25G networks can reduce the single-bit cost and operation cost.

* Technology bonus

The new generation of 25G switch chips provide rich technical features, such as Telemetry and Remote Direct Memory Access (RDMA). These features greatly improve maintenance efficiency of infrastructure networks and reduce the cost.

What is new in the 25G network in IDCs when it is compared with the 10G network with respect to network architecture? Let's learn more about the architecture of the 25G network.

• What Makes Us Choose the 25G Network?

When a 25G IDC network is deployed, the following two major factors affect the selection of product models and architecture schemes:

* Server scale: the expected number of servers in a single cluster.

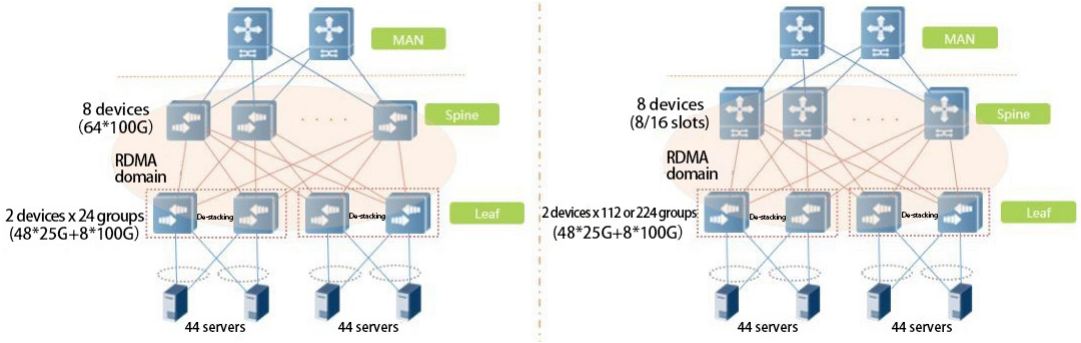
* Service application requirements: the requirements of different types of services on the network convergence ratio, the single uplink of servers, and the dual uplinks of servers.

The most common network architectures are the 2-tier network architecture and 3-tier network architecture. The following analyzes the correspondence between these two architectures and the server scale and requirements of service applications.

Architecture Design Scheme of the 25G Network

• 2-Tier Network Architecture

Figure 1: Topologies of the 2-Tier Network Architectures



Based on the topologies of the two types of 2-tier network architectures shown in Figure 1, the following analyzes their respective single-uplink and dual-uplink modes of servers, the server scales, device models, and the convergence ratios. Table 1 describes the comparison results.

Table 1: Comparison Between the Two Types of 2-Tier Network Architectures

Item	1,000 to 2,000 Servers	5,000 to 20,000 Servers	5,000 to 20,000 Servers
Architecture Type	2-tier box-based multi-core architecture	2-tier chassis-based multi-core architecture	2-tier chassis-based multi-core architecture
Number of Servers — Single Uplink	2,000	10,000 to 20,000	10,000 to 20,000
Number of Servers — Dual Uplinks	1,000	5,000 to 10,000	5,000 to 10,000
Device Model	Leaf device: RG-S6510-48VS8CQ (48 x 25 Gbit/s + 8 x 100 Gbit/s) Spine device: RG-S6520-64CQ (64 x 100 Gbit/s)	Leaf device: RG-S6510-48VS8CQ (48 x 25 Gbit/s + 8 x 100 Gbit/s) Spine device: RG-N18000-X series (CB line cards + 8 or 16 service slots)	Leaf device: RG-S6510-48VS8CQ (48 x 25 Gbit/s + 8 x 100 Gbit/s) Spine device: RG-N18000-X series (DB line cards + 8 or 16 service slots)
Convergence Ratio	Spine device: 3:1 Leaf device: 1.5:1	Spine device: 3:1 Leaf device: 1.5:1	Spine device: 3:1 Leaf device: 1.5:1

When a single cluster includes 1,000 to 2,000 servers, you can use the 2-tier box-based multi-core architecture. This architecture uses the same series of single-chip switches. Therefore, during priority-based flow control (PFC), explicit congestion notification (ECN), and memory management unit (MMU) management, threshold settings are highly consistent and can be conveniently coordinated. In addition, the forwarding delay is low, the throughput rate is high, and RDMA services and network visualization solutions can be deployed across the entire network.

When a single cluster server includes 5,000 to 20,000 servers, the 2-tier chassis-based multi-core architecture is available. At the spine layer, two types of core line cards are available for core devices:

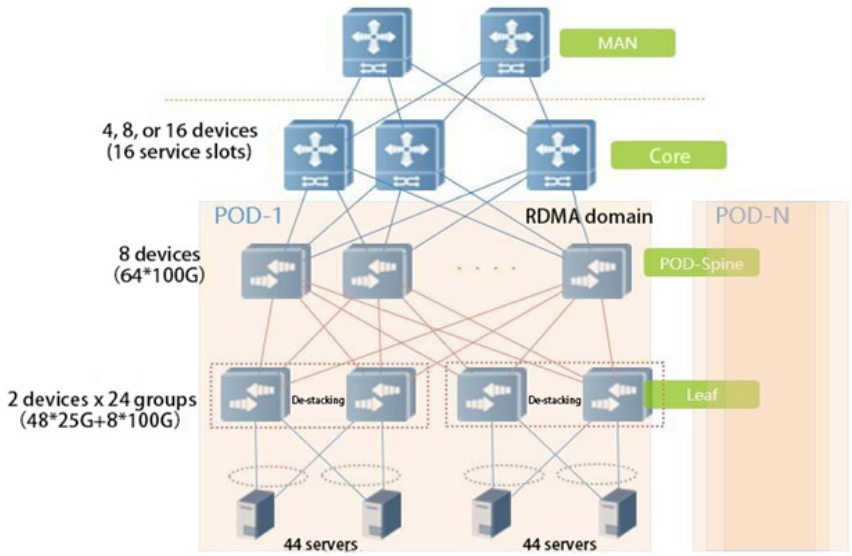
* CB line cards are applicable to service scenarios with multiple ports connected to one port. You can use a large cache to reduce packet loss in these scenarios.

* DB line cards are applicable to scenarios with high requirements for RDMA networking and network visualization. In addition, this architecture inherits the advantages of the 2-tier box-based multi-core architecture.

You can choose one of the 2-tier network architectures based on the server scale of a single cluster and service requirements. You can use External Border Gateway Protocol (eBGP) between spine devices and leaf devices and deploy the same autonomous system (AS) number for all leaf devices. The spine layer replaces the AS number after receiving routes from the leaf layer. This resolves the horizontal split problem of eBGP. When services require dual uplinks of servers, the destacking solution is recommended to deploy the leaf layer. For more information, refer to **How to "Destack" the IDC Network Architecture.**

• 3-Tier Network Architecture

Figure 2: Topology of the 3-Tier Network Architecture



For an ultra-large IDC where a single cluster includes more than 20,000 servers, the 2-tier spine-leaf network architecture no longer meets requirements and the network scalability deteriorates. In this case, a 3-tier architecture that scales out based on the point of delivery (POD, the minimum unit of IDCs) is recommended.

As shown in Figure 2, each POD is a 2-tier spine-leaf network, where the number of servers and the number of network devices are standardized. Then, multiple PODs are interconnected by using core devices. In this way, a larger-scale network is built and scalability is ensured. Table 2 compares two 3-tier architectures with respect to the number of PODs, server scale, device models, and convergence ratio.

Table 2: Comparison Between the Two Types of 3-Tier Network Architectures

Item	Over 20,000 Servers	Over 20,000 Servers
Architecture Type	3-tier architecture that scales out based on PODs	3-tier architecture that scales out based on PODs
Number of PODs	14 to 56	14 to 56
Number of Servers in PODs — Single Uplink	2,000	2,000
Number of Servers in PODs — Dual Uplinks	1,000	1,000
Number of Network-Wide Servers — Single Uplink	30,000 to 120,000	30,000 to 120,000
Number of Network-Wide Servers — Dual Uplinks	15,000 to 60,000	15,000 to 60,000
Device Model	Leaf device: RG-S6510-48VS8CQ (48 x 25 Gbit/s + 8 x 100 Gbit/s) POD-spine device: RG-S6520-64CQ (64 x 100 Gbit/s) Core device: RG-N18000-X series (CB line cards + 16 service slots)	Leaf device: RG-S6510-48VS8CQ (48 x 25 Gbit/s + 8 x 100 Gbit/s) POD-spine device: RG-S6520-64CQ (64 x 100 Gbit/s) Core device: RG-N18000-X series (DB line cards + 16 service slots)
Convergence Ratio	Core device: 3:1 POD-spine device: 3:1 Leaf device: 1.5:1	Core device: 3:1 POD-spine device: 3:1 Leaf device: 1.5:1

Similarly, two types of device models are available for one 3-tier network architecture. A POD includes a standard 2-tier spine-leaf architecture with devices of the same model. At the core layer, devices with different line cards can be selected, that is, the 2-tier chassis-based architecture with CB line cards or with DB line cards can be selected. When RDMA services need to be deployed, it is recommended to deploy the RDMA domain within a POD. Otherwise, if RDMA services are deployed at a larger scale, the control over PFC and ECN messages is much more difficult, and the impact of congestion and backpressure becomes more severe. To plan a larger-scale IDC and deploy more than 100,000 servers in a single cluster, you must upgrade spine switches. The box-type devices with 128 x 100 Gbit/s ports are recommended, so that the server scale can be doubled in the POD.

Prospects for the Next-Generation IDC Architecture

International Data Corporation forecasts that the amount of data to be processed by IDCs will reach 175 ZB in 2025, which is five times that in 2018. In particular, the data amount grows fastest in China, from 7.6 ZB in 2018 to an estimate of 48.6 ZB in 2025. In the face of fast-growing data, the infrastructure network needs to be improved in many aspects, for example, the IP Clos architecture with iterative network bandwidth upgrade and 1:1 convergence ratio. Will the IP Clos network architecture in the next-generation network still use chassis-type devices? How will server access evolve and be upgraded in the future to meet service requirements? The answers will be revealed in the next article.