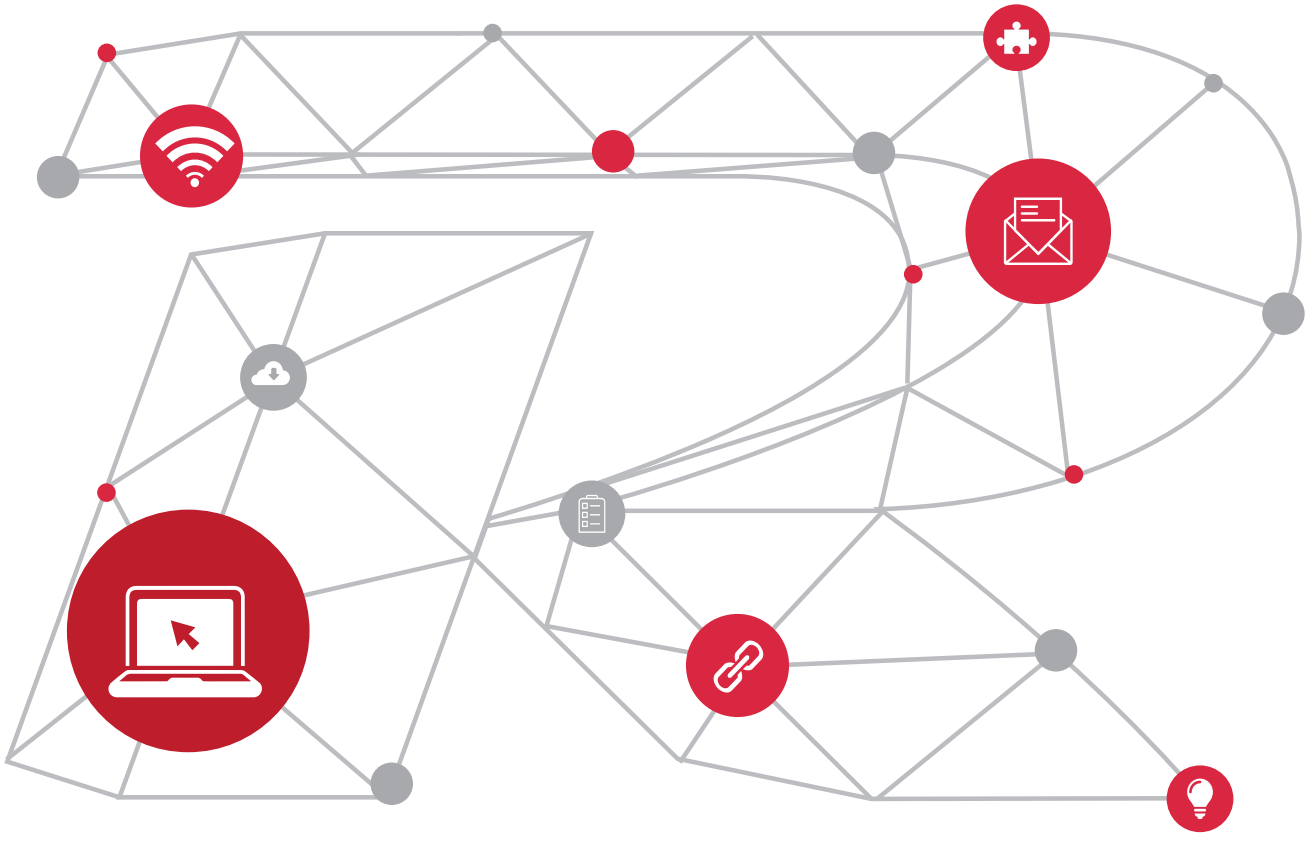
White Paper

A Feast of Technologies
Selecting a Routing Protocol for Networks of Large IDCs



Contents

IDC Network Interconnection Technologies

To meet requirements for layer-2 communication between virtual machines (VMs) and Docker containers in Internet data centers (IDCs), various Internet networking technologies emerged during development of IDC networks. These technologies are implemented based on network device hardware, including routing protocol-based layer-2 networking technologies such as Transparent Interconnection of Lots of Links (TRILL) and Shortest Path Bridging (SPB), and overlay technologies such as Virtual Extensible LAN (VXLAN) and Network Virtualization using Generic Routing Encapsulation (NVGRE). However, complexity and imbalanced device capabilities of these technologies limit their application on network devices.

As we can see, IDC networks are returning to their nature to meet core demands: decoupling from services, simplicity, and reliability. IDCs need to provide only a simple and reliable layer-3 underlay network. Layer-2 overlay networks depend more on host software or intelligent network interface cards (iNICs).

Here comes the question, how do we select an appropriate routing protocol for the layer-3 network of IDCs? This article tries to give you a definite answer by focusing on large IDC scenarios.

Evolution of IDC Network Architecture

It is well-known that economic foundation determines superstructure. Similarly, the physical network architecture of IDCs largely determines routing protocol planning. For more information about architecture design, see **A Feast of Technologies | Network Architecture Design for 25G IDCs.** This article briefly describes the network architecture of IDCs to clarify the relationship between basic architecture and routing protocol selection. of IDCs to clarify the relationship between basic architecture and routing protocol selection.

• Conventional IDC Network Architecture

Figure 1: Conventional IDC Network Architecture (Internal Network Only, Excluding the Gateway Area)

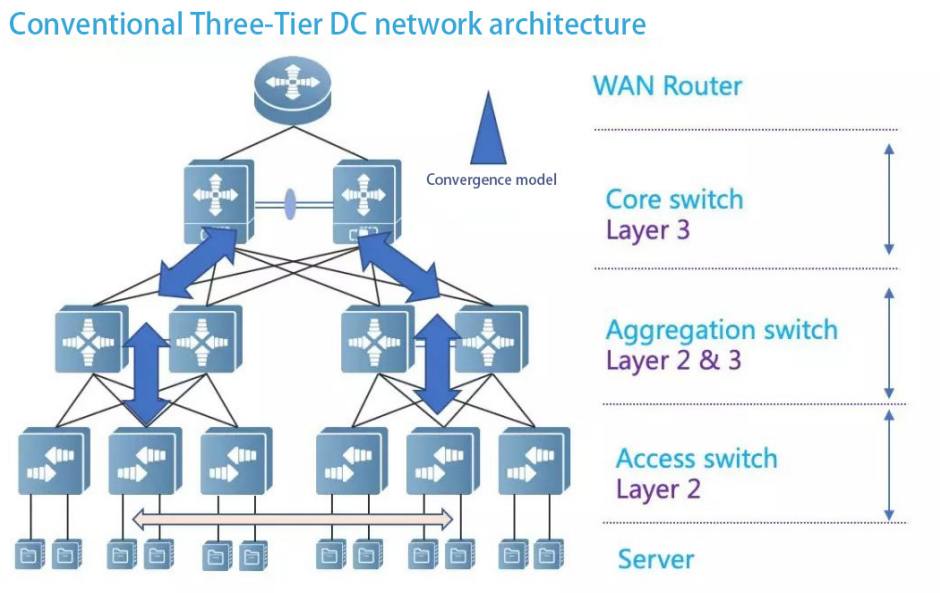


Figure 1 shows the conventional IDC network architecture. The architecture has the following features:

* Conventional IDCs mainly carry services for accessing external networks.

* Traffic distribution conforms to the 80/20 model. Conventional IDCs process more south-north traffic than east-west traffic.

* The network architecture includes the core layer, the convergence layer, and the access layer. The convergence layer and its lower part adopt layer-2 networking. Manufacturers' proprietary virtualization technologies are horizontally deployed at the convergence layer and the core layer to ensure reliability.

* Traffic bottleneck exists at egresses. A convergence ratio of 10:1 or even higher can be maintained within IDCs.

In recent years, as cloud computing, big data, and other services are developing, technologies such as distributed computing and distributed storage are deployed within IDCs in a large scale. From the perspective of networks, east-west traffic within IDCs surge abruptly, and the 80/20 model becomes centered on east-west traffic.

Such situation is beyond the reach of the conventional network architecture, which is exposed to the following shortcomings:

* **Low scalability**: The network scale is limited by the number of core switch ports, and therefore smooth scale-out is unavailable.

* **High convergence ratio**: The traffic model designed for south-north traffic uses a triangular convergence model. In this case, a higher network layer indicates poorer performance, and east-west bandwidth is severely insufficient.

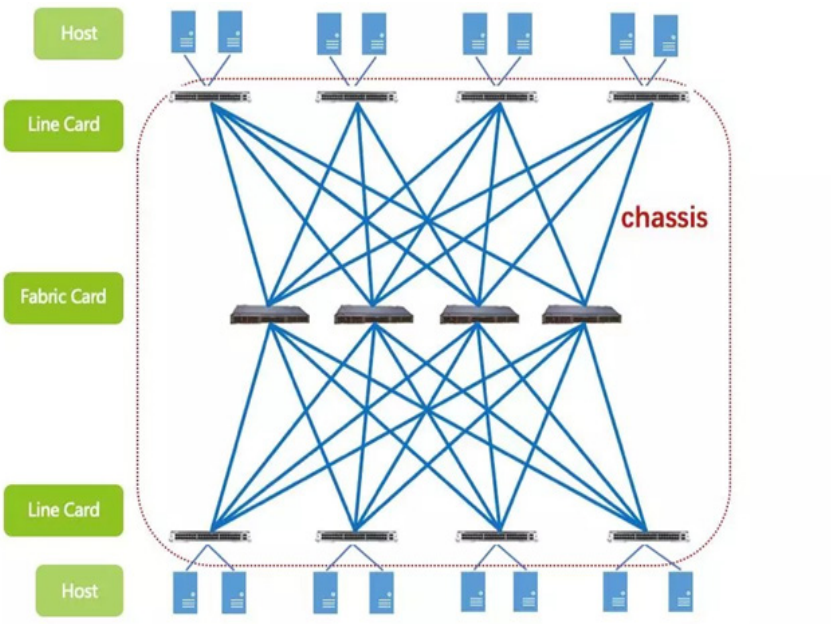
* **Highly complex maintenance of the single control plane**: Reliability of the convergence layer and the core layer depends on the horizontal virtualization technology used by the manufacturer. However, the single control plane applying this technology has apparent weaknesses in ensuring In-Service Software Upgrade (ISSU).

• Fabric Network Architecture

To resolve the problems confronted by conventional IDC networks, the fabric network architecture gradually emerges.

For frame switches in Clos architecture, a fabric module serves as the forwarding bridge between line cards, as shown in Figure 2.

Figure 2: IDC Network Architecture Design — Network as a Fabric

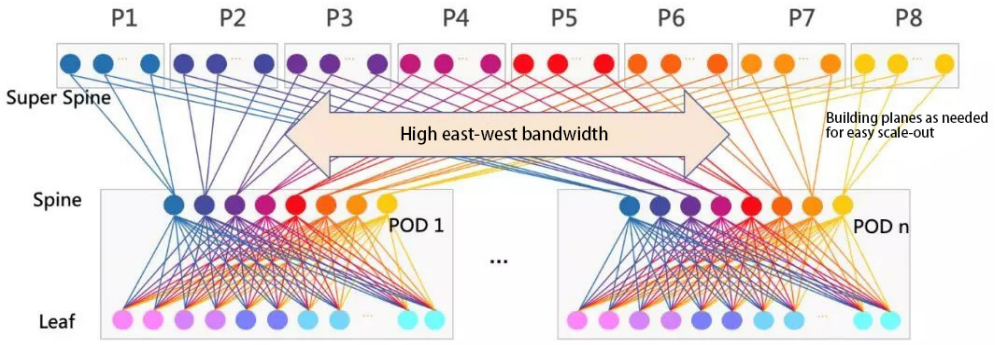


The nowadays popular fabric network architecture for IDCs is similar to Clos switches in many aspects.

* **Line card:** serves as an input/output source to aggregate traffic from all servers. It is equivalent to a top-of-rack (ToR) switch in an IDC.

* **Fabric card**: serves as a high-speed forwarding channel built at the intermediate layer. Cross-ToR traffic is forwarded by the fabric card at a high speed. When you fold Figure 2, you can find that the architecture becomes the most popular leaf-spine network architecture in IDCs.

Figure 3: Leaf-Spine Network Architecture

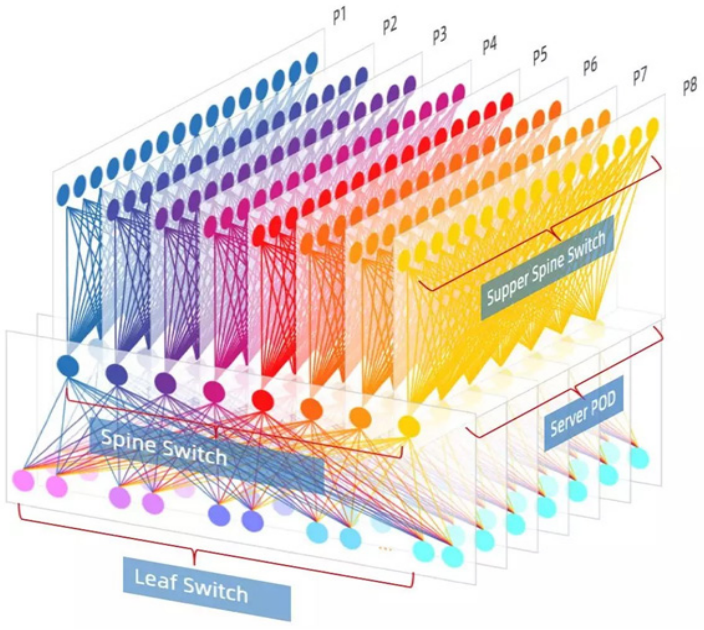


Two layers of the leaf-spine architecture can form a simple leaf-spine network. During IDC deployment, we build the network based on the point of delivery (POD). Certainly, to improve the scale-out capability of this network architecture, we usually add one layer above PODs. This layer is used to horizontally connect different PODs of the IDC, expanding the cluster of the entire IDC.

In addition, the leaf-spine architecture is popular for its powerful scale-out capability, high reliability, and excellent maintainability. Most well-known global Internet giants adopt this network architecture.

Routing Protocols for Fabric Network Architecture

Figure 4: Large Fabric-based IDC Network



Facebook disclosed its IDC network design in 2014. Its network has evolved from F4 to F16, but has a basic architecture similar to that in Figure 4, which shows a classic fabric network. Then, which routing protocol is more appropriate for the fabric network architecture?

In RFC 7938, Use of BGP for Routing in Large-Scale Data Centers, the author proposes that the Border Gateway Protocol (BGP) be used as the only routing protocol within IDCs. The author also provides detailed analysis. Anyone interested in it may read RFC 7938.

Let's analyze why BGP is preferred based on RFC 7938 and practice of BGP networks in Internet companies inside and outside China.

• Routing Design Principle for Large IDC Networks

As an important step in IDC network design, routing design must comply with the general rule of IDCs. The following describes key points of routing design.

Scalability

IDC design: Many large campus networks have 20,000 to 100,000 servers, and a single campus network of large Internet companies can even have more than 300,000 servers. When you design the IDC network, you must make sure that smooth scale-out is supported, the IDC network can be delivered by POD (which reduces initial investment), and the IDC network can be extended to carry large and ultra-large clusters.

Routing protocol design: When switches and servers are deployed at a 1:20 ratio (a typical ratio of 48-port switches to servers in dual-homed networks), an ultra-large IDC can run thousands of network devices. You must ensure that routing protocols used by the network devices are consistent and easy to use. This can ensure quick route transmission and convergence for both a small-scale network in the early stage and a routing domain built with thousands of network elements.

Bandwidth and Traffic Model

IDC design: East-west traffic explodes in IDCs. The conventional high convergence ratio model can no longer meet requirements of east-west traffic. The new network architecture must minimize convergence to zero as far as possible. In Microsoft's network, uplink bandwidth is even higher than downlink bandwidth. To ensure high cost performance, we recommend that you adopt the convergence ratio ranging from 1:1 to 3:1 for each tier.

Routing protocol design: In fabric networks, low convergence is achieved based on uplink load of multiple uplinks. For the typical 25G ToR switch RG-S6510-48VS8CQ, its downlink bandwidth is 48 x 25 Gbit/s = 1,200 Gbit/s, and its uplink bandwidth is 8 x 100 Gbit/s = 800 Gbit/s. When all its ports are used, the convergence ratio is 1.5:1. When you design routing protocols, it is important to implement equal-cost multi-path routing (ECMP) between multiple links in the IDC. In normal cases, ECMP paths can balance traffic. When a path is added or removed, fast convergence can be achieved without affecting live services.

CapEx Minimization

IDC design: Minimize capital expenditure (CapEx) by:

* Standardizing software and hardware requirements of network devices and use the uniform architecture to reduce device types.

* Simplifying network features to reduce development and time costs.

Routing protocol design: Select a mature and general-purpose routing protocol. Ensure that the routing protocol is supported on mainstream models and used on access, core, and backbone devices.

OPEX Minimization

IDC design: Minimize operating expenditure (OPEX). In large IDC networks, OPEX is usually higher than infrastructure construction costs. Reducing OPEX must be considered at the beginning of architecture design as well.

Routing protocol design: Reduce the size of the failure domain in the network by:

* Ensuring that the impact of routing convergence is minimized and convergence is fast when the network is faulty.

* Using only one routing protocol in the entire IDC for more simplified maintenance and reduced learning costs. In addition, the operating knowledge base can be accumulated to help quickly locate problems and recover from failure.

• Routing Protocol Selection for Networks of Large IDCs

Required Capabilities of a Routing Protocol

Based on the preceding key points of routing protocol design, we have reached a conclusion that a routing protocol for large IDCs must be characterized by:

* **Ultra-large scale**: To ensure high scalability, use the same networking protocol from initial building to full configuration of the cluster. The protocol must support scale-out to ultra-large IDCs.

* **Simplicity**: Select a simple, mature, and general-purpose routing protocol and minimize the number of software features to introduce a wider range of device manufacturers.

* **Only one routing protocol**: Try to use only one routing protocol in an IDC to reduce complexity and learning costs, and facilitate accumulation of operating experience.

* **Minimized failure domain**: Minimize the impact of network failures to improve network robustness.

* **Load balancing**: Get rid of dedicated load balancing devices to achieve ECMP within the IDC.

* **Flexible policy control**: Use rich means to control routing policies based on requirements of specified service streams.

* **Fast convergence**: Minimize the impact of network failures to achieve fast convergence.

Existing Routing Protocols

Let's check the application scope of existing routing protocols based on the network requirements.

* **Routing Information Protocol (RIP)**: It is not applicable to large IDCs.

* **Enhanced Interior Gateway Routing Protocol (EIGRP)**: It is proprietary and does not meet requirements 2 and 3.

* **Internal Border Gateway Protocol (IBGP)**: It is generally used with the Interior Gateway Protocol (IGP), and does not meet requirements 2 and 3.

* **Open Shortest Path First (OSPF), Intermediate System to Intermediate System (IS-IS), and BGP**: They can meet all the requirements. IS-IS and OSPF belong to link-state IGP and are similar to each other. OSPF is more widely used. The following provides a comparison between OSPF and BGP.

OSPF vs. BGP

The following definitions are provided based on related information on Wikipedia:

* **OSPF**: It uses a link state routing (LSR) algorithm and falls into the group of IGPs, operating within a single autonomous system (AS). It implements Dijkstra's algorithm to calculate the shortest path tree. It uses "cost" as its routing metric, and further uses a link state database (LSDB) to store the current network topology. The LSDBs for one area are the same on all routers.

* **BGP**: It is a core decentralized protocol used for routing within an autonomous system (AS). It maintains an IP routing table or prefix table to implement network reachability among ASs. It is a path vector protocol. BGP does not use conventional IGP metrics, but makes routing decisions based on paths, network policies, and/or rule sets. For this reason, it is more a vector protocol than a routing protocol.

OSPF and BGP are both widely used and are technologically equal. The following analyzes the application scope of the two routing protocols in large and ultra-large IDCs.

Table 1: Comparison Between Routing Protocols of Large IDCs

Protocol Type / Configuration Item	OSPF	BGP
Routing algorithm	Dijkstra's algorithm	Best path algorithm
Algorithm type	Link-state	Distance-vector
Bearer protocol	IP	TCP, with the retransmission mechanism for ensuring data reliability
Requirement 1: Ultra-large scale	Applicability: ★★★ Theoretically, OSPF does not restrict the number of hops, and therefore can support large-scale routing networks. However, OSPF periodically synchronizes link state information of the whole network. As ultra-large IDCs have a huge link state database, the network devices have high performance costs and large impact of network fluctuation during computing.	Applicability: ★★★★★ BGP delivers only the information about the optimal route. It is applicable to large and ultra-large IDCs and has been widely applied in ultra-large campuses.
Requirement 2: Simplicity	Applicability: ★★★ OSPF is easy to deploy and moderately difficult to maintain.	Applicability: ★★★★ BGP is easy to deploy and easy to maintain.
Requirement 3: Only one routing protocol	Applicability: ★★★★ Meet the requirement You can deploy OSPF alone within an IDC. OSPF is supported by a wide range of software on the server.	Applicability: ★★★★ Meet the requirement You can deploy BGP alone within an IDC. BGP is supported by some software on the server. External ASs are interconnected by using BGP.
Requirement 4: Minimized failure domain	Applicability: ★★ Within the domain, link state information must be synchronized, and all failure information must be synchronously updated.	Applicability: ★★★★ BGP locally delivers only the optimal paths. When the network changes, BGP delivers only incremental information.
Requirement 5: Load balancing	Applicability: ★★★★ When the cost value is planned and multiple paths are available, OSPF can achieve ECMP. When a path is faulty, OSPF must synchronize computing results of devices within the domain.	Applicability: ★★★★★ When the numbers of hops and ASs are planned and multiple paths are available, BGP can achieve ECMP. When a path is faulty, BGP removes the next hop corresponding to the faulty link from the ECMP group.
Requirement 6: Flexible policy control	Applicability: ★★★ OSPF controls route transfer based on the area and link-state advertisement (LSA) type. This is relatively complex.	Applicability: ★★★★ BGP provides rich route selection rules to filter routes and control route receiving and sending.

Protocol Type / Item	OSPF	BGP
Requirement 7: Fast convergence	Applicability: ★★★ When few routes are available, OSPF can work with BFD to reduce convergence to milliseconds. OSPF advertises link state information. When the routing domain is large, computing consumption is high, causing convergence to slow down.	Applicability: ★★★★ When few routes are available, BGP can work with BFD to reduce convergence to milliseconds. BGP advertises routes that are locally calculated. When the routing domain is large, computing performance is not apparently affected. In addition, BGP provides the AS-based quick switching technology.

According to the analysis in Table 1 and some practice in the industry, OSPF is recommended in small and medium IDCs when a small number of network devices exist in the routing domain. BGP is recommended in large and ultra-large IDCs.

Afterword

For brevity, this article describes only the reasons why BGP is preferred in large and ultra-large IDCs without BGP planning. **Ruijie Networks has deployed large and ultra-large IDCs based on BGP in all of the top 3 Internet companies in China.** For BGP planning, here are some questions that I look forward to discussing with you in subsequent articles:

* BGP has a limited quantity of private AS numbers. How do we plan ASs in large IDCs?

* What interface is used in BGP to establish neighbors? How do we plan the interface in ECMP and Link Aggregation Control Protocol (LACP) scenarios?

* BGP provides many route selection rules. How do we make good use of them?

* How do we optimize performance, reliability, and convergence of BGP?