Building a Next-Generation Data Center

Data center networking has changed significantly during the last few years with the introduction of 10 Gigabit Ethernet (10GE), unified fabrics, high-speed non-blocking core switches, and novel data-center architectures (example: Clos fabrics).

If you're planning a new data center rollout, or an upgrade or expansion of your existing data center, you should consider using these new technologies and reevaluating your design and deployment options in light of the benefits they offer. The cost savings you could realize in capital and operating expenditures are usually significantly larger than the increased investment necessary to introduce such technologies into the data center design process.

Each data center has a unique set of requirements that influence its design options. This whitepaper describes the typical requirements you should collect during the preparation phase, the criteria you should consider during the design process, and the key decision points you'll have to make.

Collecting the Requirements

During the requirements-collection phase, you should consider numerous aspects, including the number of virtual versus physical (bare-metal) servers, the typical number of virtual machines running on each physical server, the traffic distribution in your data center (the ratio between server-to-user and server-to-server traffic), the amount of network and storage traffic per physical host (bare-metal or hypervisor), and the private or public cloud deployment plans.

Server virtualization has become a fact of life in modern data centers. Its hardware-consolidation capabilities and associated cost reductions are simply impossible to ignore. Most data centers are already heavy users of server virtualization, many of them virtualizing almost all workload. If you aren't using server virtualization yet, you should strongly consider it during the data center design phase.

Virtualization ratio (the number of virtual machines running on a single physical server) has increased dramatically with the next-generation servers, which hold up to 40 processor cores and up to 1TB of RAM. These servers can easily host between 50 and 100 virtual machines, significantly changing the traffic patterns in your data center.

Inter-server versus server-to-user traffic. In the past, traditional single-server (scale-up) applications generated little inter-server traffic, as most processing occurred within a single physical or virtual server. The amount of inter-server traffic generated by modern distributed (scale-out) applications far exceeds the server-to-user traffic; in some cases, more than 80% of the traffic stays within the data center.

Amount of network and storage traffic per server. High virtualization ratios require adequately sized server uplinks – especially important if you plan to consolidate storage and user traffic over Ethernet uplinks using technologies like iSCSI, NFS or Fibre Channel over Ethernet (FCoE).

Private cloud deployment. From the technology standpoint, private clouds are no different from server virtualization. However, these approaches differ significantly in the application deployment model: application teams should be able to deploy their own applications, creating their own virtual resource pools and networks on demand. The number of application-specific virtual networks in a medium-sized private cloud environment could easily exceed the limitations of commonly used virtual network technologies such as VLANs.

Designing for the Future

In combination with the specific needs of your application environment or business, consider the following general design criteria:

Workload mobility. One benefit of server virtualization is the ability to move running virtual machines between physical servers. The IP address assigned to a virtual machine must not change during the move (or the virtual machine would lose all TCP and UDP sessions), usually requiring layer 2 (VLAN) connectivity between origin and target servers.

Most hypervisor vendors offer Ethernet over IP solutions (VXLAN, NVGRE or STT) that you can use to implement workload mobility across IP subnets.

The Ethernet over IP technologies are still in a nascent phase. For example, none offers direct connectivity to physical networks or appliances.

Equidistant endpoints. The performance of the scale-out applications must not depend on the physical proximity of the virtual machines that are part of the application stack. The potential bandwidth between any two virtual machines in the data center should be the same regardless of their physical location.

Virtual machines running in the same physical server or connected to the same top-of-rack (ToR) switch are an obvious exception.

You can build a data center with equidistant endpoints if you use Clos fabric architecture (where every ToR switch connects to every core switch) with a non-blocking core.

Non-blocking network core. Traditional data centers of the past had numerous chokepoints, including oversubscribed links between top-of-rack and aggregation switches, and between aggregation and core switches, as well as non-wire-speed aggregation/core switch performance, making it impossible to implement the *equidistant endpoints* requirement.

A modern data center design should include a reasonable amount of oversubscription on the links between ToR and aggregation/core switches (a 3:1 oversubscription ratio is commonly used), but no oversubscription in the network core. Aggregation/core switches should be non-blocking and capable of forwarding traffic at wire speed.

Lossless transport is a prerequisite for unified fabric (integration of network and storage traffic on the same physical infrastructure). Other high-volume applications, such as backup, also can benefit.

Because lossless transport significantly improves the performance of high-volume (elephant) TCP flows that are mixed with regular user traffic, it should be a mandatory requirement in any modern data center design.

Lossless transport is implemented with Data Center Bridging (DCB) standards. The data center switches should support at least Priority Flow Control (PFC, 802.1Qbb) and Data Center Bridging Exchange (DCBX, part of 802.1Qaz), with Enhanced Transmission Selection (ETS, 802.1Qaz) being highly desirable.

Simplified provisioning and management. Traditional data center switches were managed as isolated individual devices; each had its own configuration and its own set of network management counters.

Some modern data center solutions implement a shared *management plane* for a cluster of switches, making them appear as a single switch to the network operator and network management system. Other solutions use a

shared *control plane* with a single controller (sometimes known as a *supervisor* or *routing engine*) that controls multiple physical devices.

Shared control plane solutions are inherently more brittle than shared management plane versions, as the physical devices don't have autonomy and thus cannot operate independently under transient failure conditions (for example, link or node failures).

KEY DECISION POINTS

During the network design process, you will have to consider several significant issues. For example:

- Keep the storage and network traffic separate, or build a unified fabric?
- Deploy servers with 1GE or 10GE uplinks?
- What's a realistic size for the VM mobility domain?
- How easily can we make changes and additions to the new data center network?

The following paragraphs explain the decision points in more detail.

Unified fabric (storage and network traffic integration) should be implemented at least between the servers and top-of-rack switches to minimize the cabling and network interface card (NIC) requirements.

Unified fabric on the network edge doesn't force to you to choose a particular SAN protocol or change your existing storage solution. You can use FCoE between servers and ToR switches, or use iSCSI or NFS for an Ethernet-only solution.

You can split storage and network traffic at the ToR switch (for example, switching from FCoE into Fiber Channel SAN), or keep them integrated throughout the network (for example, with multihop FCoE), but it's important to use unified fabric where it matters most – on server uplinks.

Gigabit or 10 Gigabit Ethernet. You already made this decision if you chose to use unified fabric – it requires 10 Gigabit Ethernet (10GE). You should also consider 10GE technology when deploying large modern servers. A VM running on a single Xeon core can generate up to 4 Gbps of traffic; multiple 10GE uplinks will barely suffice for servers with dozens of cores.

Finally, VMware makes good use of 10GE uplinks – vMotion can consume up to 8 Gbps of bandwidth on a single 10GE uplink (vSphere 4.1 and later) and use multiple parallel 10GE links (starting with vSphere 5.0). You can migrate your virtual machines up to 10 times faster by migrating from Gigabit to 10 Gigabit Ethernet.

Size of mobility domain. In an ideal world, you could deploy your virtual machines on any physical server, moving them across the whole data center or even multiple data centers at will. In today's data center, however, you're limited by a number of technology and implementation considerations:

- The vSphere high-availability (HA) cluster has at most 32 hosts.
- The vSphere virtual distributed switch (vDS) that you should use in large-scale deployments can connect at most 350 hosts, and you cannot move a running virtual machine between two distributed switches.
- VLANs are commonly used to implement VM mobility, and a typical VLAN shouldn't have more than a few hundred hosts.
- A bridged (Layer 2) domain (a collection of VLANs) shouldn't span more than approximately 1,000 hosts.

Furthermore, every bridged domain is a single failure domain. If you implement your data center as a single bridged network, then a single software bug in one of the switches, or a misconfigured virtual machine, could bring down the whole data center.

You should therefore split your data center into multiple failure domains (you can also call them *availability zones* or *swim lanes*) connected with Layer 3 or Layer 4 to Layer 7 devices (routers, firewalls and/or load balancers). VM mobility will be limited to a single availability zone unless you use MAC over IP technologies (VXLAN, NVGRE or STT), but you'll have a more robust and reliable data center.

Ease of moves, additions and changes. Most data centers are continuously evolving (growing, changing or upgrading), and you should consider the impact of these changes on the data center network. Ideally, you should be able to add new top-of-rack switches with no reconfiguration of core or aggregation switches, increasing the capacity of the network core without forklift upgrades or extensive rewiring.

CONCLUSIONS AND RECOMMENDATIONS

You should strongly consider the following recommendations during the network design process:

- Unified fabrics significantly reduce the requirements for cabling and device management. At minimum, use unified fabrics between the servers and top-of-rack switches.
- 10 Gigabit Ethernet reduces the number of NICs in the servers – and thus the complexity of wiring and management. Choose 10GE over 1GE whenever possible.
- Limitations imposed by the hypervisor vendors make it impossible to implement unlimited virtual machine mobility in the data center. Keep your requirements realistic.
- A transparently bridged Layer 2 network that is usually required to implement virtual machine mobility is a single failure domain. Split your data center into multiple VM mobility domains, isolated with routers, load balancers and/or firewalls.

Ivan Pepelnjak (CCIE #1354 Emeritus) is Chief Technology Advisor at NIL Data Communications. He is the author of numerous webinars and advanced networking books and a prolific blogger, focusing on data center and cloud networking, network virtualization and scalable application design.