Data Center Resource Disaggregation Drives Need for Cost-Effective 400/800-GbE Interconnects

New challenges arise for cabling complexity, serviceability, and rack power

As new compute-intensive machine learning (ML) and artificial intelligence (AI) workloads drive servers to adopt faster PCI-Express® 5.0 Links, lower-latency cache-coherent protocols like Compute Express Link™, and a dizzying array of memory, storage, AI processor (AIP), smart NIC, FPGA, and GPU elements, so too is heterogenous computing pushing the need for **blazing-fast networks to interconnect the resources**. Distributed compute/memory/storage nodes have spawned requirements for a high-bandwidth, low-latency, and—perhaps most importantly—**scalable and serviceable network topology**, one which can support the explosion of "east-west" traffic brought about by resource disaggregation.

100 GbE, based on 25 Gbps/lane technology, is the workhorse of today's hyperscale datacenter Ethernet networks: connecting servers and storage to leaf switches, and leaf switches to spine switches. In a Clos network topology—common in hyperscale datacenters—scale is achieved by adding more servers under a new leaf switch and connecting this leaf switch to all next-level spine switches (Figure 1). Likewise, bandwidth may be scaled by adding additional spine switches (Figure 2).

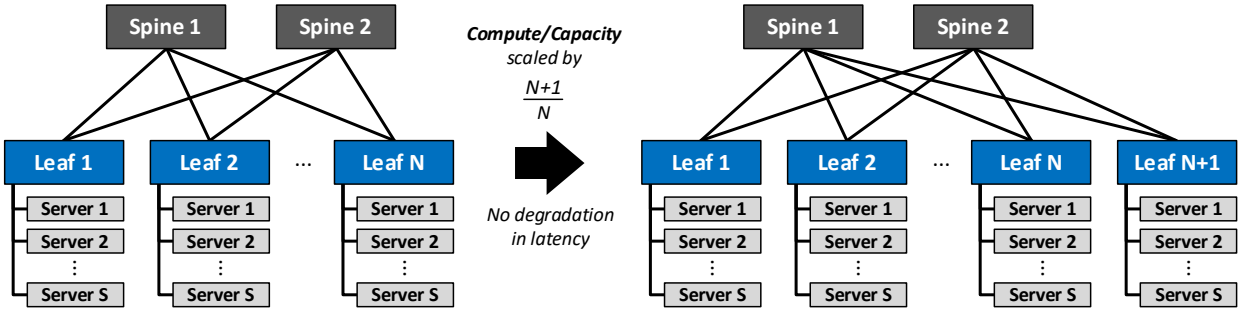

Figure 1: Scaling Compute in a Clos Network

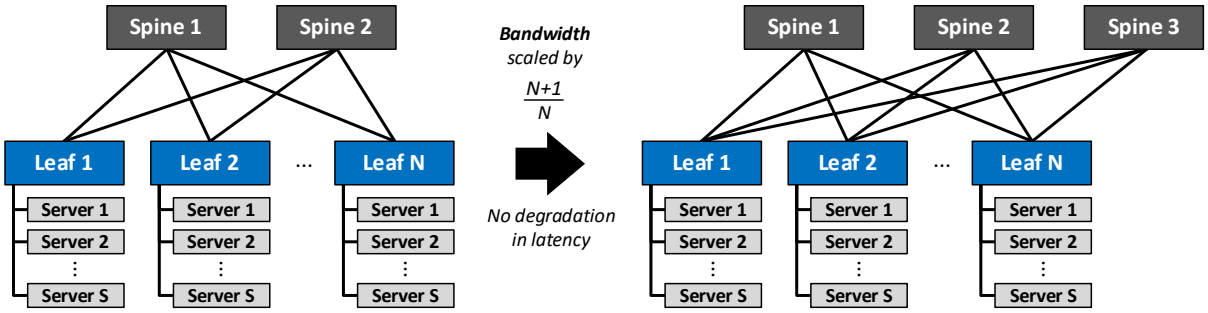

Figure 2: Scaling Bandwidth in a Clos Network

An obvious bottleneck for both the latency of the network (how many "hops" from server to server) as well as the bandwidth of the network segments (port speed) is the leaf and spine switch—often the same physical piece of equipment. Switch silicon providers have been working hard on both fronts: increasing I/O speeds to 50 Gbps/lane today and moving to 100 Gbps/lane in 2021; and increasing the number of ports (the "radix") to 64 eight-lane ports (e.g., QSFP-DD or OSFP form factors). All told, this amounts to 25.6 Tbps of switching bandwidth

in today's most advanced switch chips. These advances in switch chip capabilities are helping to enable *faster* and *flatter* network topologies based on 400-GbE and 800-GbE ports.

Practical Solutions for Switch-to-Switch Interconnects

With an increase in port speed and port count comes new challenges for these lowest layers in the data center network. The first key challenge is *how do you connect 400/800-GbE switch ports to one another in a practical manner?* A simple Clos network topology connects every leaf switch to every spine switch; and in turn, every spine switch is connected to some number "exit leaf" switches (at least two), as shown in Figure 3.

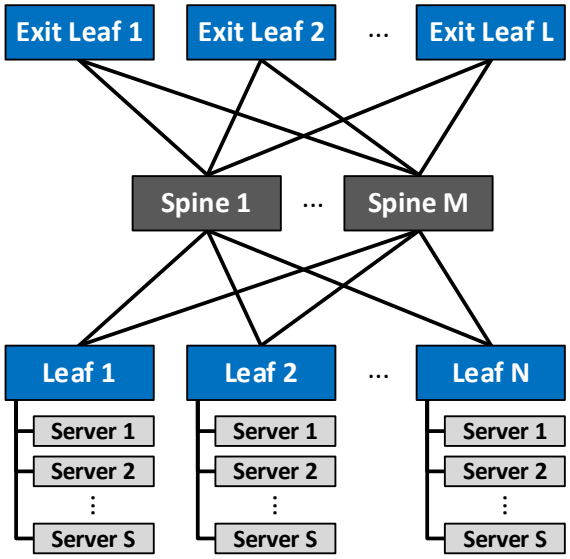

Figure 3: Example Clos Network Topology with N Leaf Switches, M Spine Switches, and L "Exit Leaves"

The leaf and spine switches are generally not co-located (i.e., >10 meters apart). For example, the leaf switches may be at the top (or middle) or each server rack, whereas the spine switches may be at the end of a row or a cluster of rows. The spine switches, on the other hand, are often co-located (i.e., <3 meters apart), as shown in Figure 4.

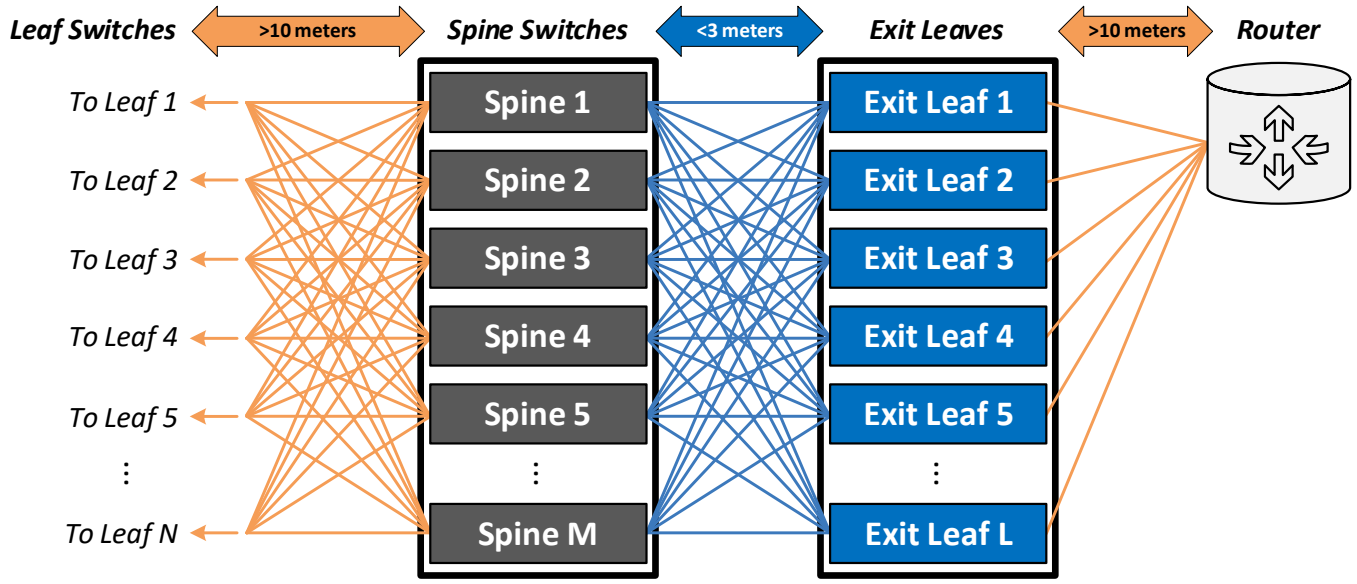

Figure 4: Example Showing Long-Reach Leaf/Spine Interconnects and Short-Reach Spine/Exit-Leaf Interconnects

Spine-to-exit-leaf interconnects (3 meters or less) can be serviced by copper cables, but they are not without their challenges.

Challenges for 400/800-GbE Copper Interconnects

- At 50 Gbps/lane, passive direct-attach copper (DAC) cables barely meet the 3-meter reach requirement.
- At 100 Gbps/lane, experts predict 2 meters may be the practical limit for DACs.
- The switch PCB consumes too much of the overall 35-dB end-to-end channel budget, greatly limiting cable reach and increasing diameter.
- DACs are too rigid, heavy, and bulky. They restrict airflow and make servicing the rack practically impossible.
- At 26 AWG and 16 twinax pairs per cable, there simply is not enough room to achieve a reliable cable attach in the narrower QSFP-DD form factor. OSFP may be the only option.

Various solutions have been proposed to these challenges—some in use already for 50 Gbps/lane, and many expected to be adopted for 100 Gbps/lane.

Potential Solutions to Copper Interconnect Challenges

- Retimers may be used behind certain switch ports to effectively eliminate the switch PCB from the end-to-end channel budget, thereby extending DAC reach and/or reducing DAC thickness. Exotic PCB materials and/or mid-board "fly-over" cables are being considered as well, though their efficacy in extending cable reach would be less.
- Active copper cables may be used to achieve longer, thinner cables—with a side benefit of relaxing the design requirements of the switch PCB (i.e., the switch may be designed for "short reach" instead of "long reach" channels).

Leaf-to-spine interconnects (>10 meters), on the other hand, are resigned to using optical fiber. Although advances in silicon photonics and optical component technology are constantly being made, optical interconnects still present numerous challenges for 400 GbE and 800 GbE.

Challenges for 400/800-GbE Optical Interconnects

- Power consumption of 400-GbE modules is ~12 W, and this is expected to balloon to ~20 W for 800-GbE modules.
- Optical modules require a "short reach" interface to the host switch ASIC and designing such an interface to support 100 Gbps/lane speeds is costly, again requiring exotic PCB materials or mid-board cables, both of which can add significant cost and reliability concerns to a system design.
- The lifespan and reliability of optical modules is historically unimpressive, which means data center operators are constantly having to track down, diagnose, and replace failed modules.
- Industry experts question whether low-cost "VCSEL-based" optics will be possible for 800 GbE.

Potential Solutions to 400/800-GbE Optical Interconnect Challenges

- Again, Retimers may be used on the switch PCB to reduce the equalization burden placed on the optical module and the switch SerDes, thus remove the need for non-mainstream PCB materials or mid-board cabling.
- On-board optical modules ("optical engines") can be used to eliminate some of the PCB and connector losses, thereby modestly decreasing the power consumption of the switch and optical Retimers; though this approach introduces new concerns centered around servicing the optics and internal fibers.

Cost-Effective Switch-to-Server Interconnects

The second key challenge is *how do you effectively distribute switch port bandwidth to servers?* Server network interface cards (NICs) lag switch ports in terms of bandwidth. While 400-GbE switch ports came to market in 2020, server NIC ports are generally at 200 GbE or less. More fundamentally, there is a disparity between per-lane speed of server NIC ports and switch ports. While switch ports now run at 50 Gbps/lane, NIC ports are (for the most part) still at 25 Gbps/lane. When 100 Gbps/lane switch ports arrive in 2021, NIC ports will just be making the transition to 50 Gbps/lane. There is a price to pay for connecting a switch directly to a NIC: switch port bandwidth is underutilized, by a factor of two! Resolving this disparity requires rate conversion on the NIC side up to the higher 100 Gbps/lane speed to fully-utilize the switch port bandwidth.

The most abundant (and cost-sensitive) interconnect in the data center is the top-of-rack (ToR) switch to server interconnect. Painstaking efforts are made to keep this interface low-cost and easy-to-maintain. As such, DAC cables are traditionally used for these 1-3 meter links. As ToR switch ports move to 100 Gbps/lane, such reach becomes questionable and cable thickness, bend radius, and weight become a concern.

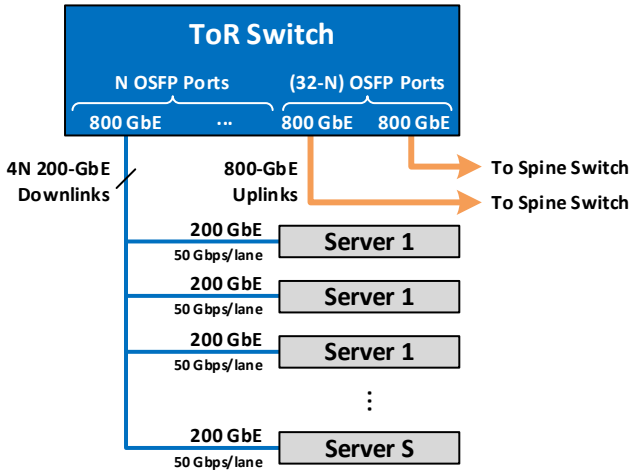

Figure 5: Typical Switch-to-Server Interconnect Using 800-GbE Switch Ports and 200-GbE NICs

Challenges for 400/800-GbE Switch-to-Server Interconnects

- Rate mismatch between NIC (25 or 50 Gbps/lane) and switch (50 or 100 Gbps/lane) leads to wasted switch bandwidth.
- Traditional DAC interconnect becomes too short, thick, and bulky, especially when using one ToR for a grouping of 2-3 racks.

Potential Solutions to 400/800-GbE Switch-to-Server Interconnect Challenges

- Active optical cables (AOCs) can be used to do rate conversion while keeping a slim profile. The major downside is the added cost (~10x compared to DACs) and reliability concerns.
- Active copper cables can likewise be used to do rate conversion and achieve a much thinner diameter cable compared to passive DACs.
- A Retimer with gearbox capabilities can be utilized on the NIC to resolve the per-lane rate disparity and reduce the end-to-end channel loss, thereby increasing cable reach and/or reducing cable diameter.

Conclusions

While heterogenous computing trends have led to the cramming of more processing power and bandwidth into servers and other end-node boxes, these same trends have resulted in an explosion in “east-west” traffic: data moving between servers, GPU trays, storage, and other end-node systems in the network. Luckily, 100 Gbps/lane technology will help carry this increased bandwidth; but with it comes new interconnect challenges and bottlenecks in switch-to-switch and switch-to-server connectivity. Removing these bottlenecks will be key to ushering in a new age for intelligent networks.