AddOn Networks
Paths to 100G in the Data Center

addon

Paths to 100G in the Data Center

In our data-driven economy, the data center has become a competitive advantage. Whether the data center supports the business model or whether it is the business model, the data center plays an essential role in business today. Just having data isn't enough. To fully support the business and its customers, that data needs to be accessed, stored, and moved at extremely high speeds. For years, 10 Gbps was the de facto rate in the data center, followed by 25 Gbps and 40 Gbps. Today, 100 Gbps operation is becoming mainstream, even as the technology for 200 Gbps and 400 Gbps transport is the focus of widespread development.

A number of techniques can be used to achieve 100G transmission in the data center. Several types are commercially available for deployment today. Others, like single-lambda 100G, are still in the final phases of development. Long-term, the latter option is of key importance, not just for simplifying 100G deployment but also for easy scalability to 200G and 400G.

DATA CENTER CHALLENGES

A number of factors govern the technology chosen for optical transport in the data center. Given the sheer scale of even a small facility, cost and complexity are of primary concern. Faceplate density and energy consumption likewise influence the decision. The architecture needs to be scalable to address changing needs. Finally, the link as a whole needs to be able to provide sufficient sign integrity all across the required distances.

The traditional model for intra-data center data-center interconnects (DCIs) is the three-level scale-up model (see figure 1). From the top of rack, data passes to a layer of access switches. These switches transfer the data to a layer of aggregation switches. The aggregation layer route stated to the core switches that, in turn, send it to an off-site data center or into the network-area network.

(fig. 1)

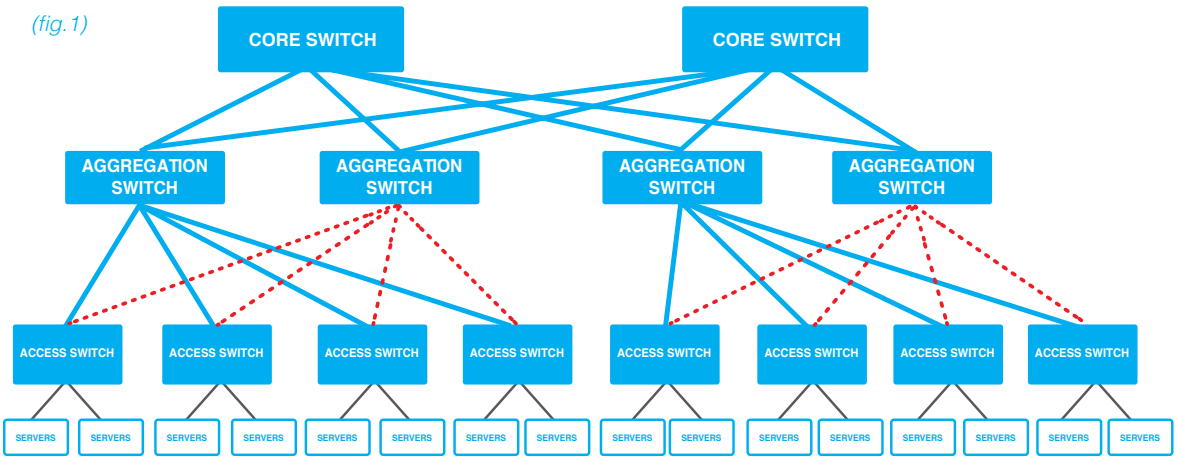

***Figure 1:** The traditional data-center switching topology consists of a layer of access switches that send data from the servers through the next switching layer, known as the aggregation switches. These aggregation switches, in turn, send the data to the core switches for routing off-site.*

The problem with scale-up, three-level switching topology is that it is not particularly scalable. Adding servers means expanding the number of switches at the access and aggregation layers. In the case of hyper-scale data centers, which can have tens of thousands of servers, the architecture increases costs significantly. Worse, network operation depends on the health of the core switches. When one of them goes down, it takes a significant portion of the data center with it. Scale-out architectures provide a more resilient alternative.

The scale-out, or leaf-spine architecture consists of just two switching layers (see figure 2). Data flows from top of rack to a series of access switches known as leaf switches. The leaf switches in turn feed interconnection switches, also known as spine switches. The result is a flatter, more resilient network that is also highly scalable. Adding more servers just requires an additional leaf switch, which is more compact and lower cost than the scale-up architecture.

(fig.2)

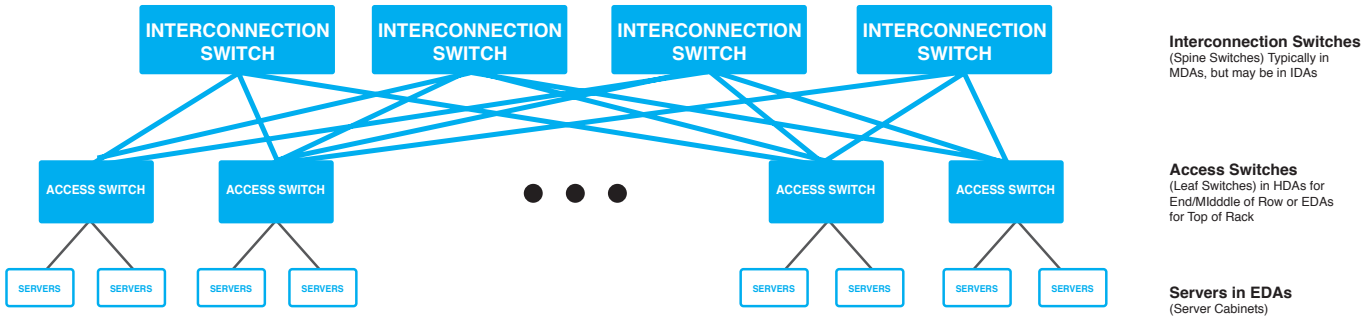

***Figure 2:** In scale-out data-center topology, data moves from the servers to the access switches (leaf switches). From there, it connects to the interconnect switches (spine switches), and then to the Metropolitan-area loop.*

For the past several years, the most common speed for transmitting data from the server stack to the access switch has been 10G. Data-center architects have used 40G links to send data from the access layer to the aggregation layer and out to the core switches. In the tree model, the network again uses 10G data rates for moving data from the server stack to the access switches, and 40G to transport data to the spine switches and out to the Internet.

To satisfy ever-growing bandwidth demand, the speeds are changing from 10G/40G to 25G/100G. Now that we know where 100G technology is being applied in the data center, let's look at how the systems work.

TRANSCEIVER SPEEDS

The most basic modulation scheme for optical communications is on-off keying, commonly implemented as non-return-to-zero (NRZ) modulation. In this scheme, the laser is either fully on (logical 1) or fully off (logical 0). The laser can be switched on and off directly by modulating the drive current; these devices are known as direct-modulated lasers (DMLs). Alternatively, the laser can be driven at a constant current and modulated by an external device known as an electro-absorption modulator; these devices are known as externally modulated lasers (EMLs).

In both cases, the switching speed is controlled by the drive electronics. Until recently, the peak switching speed was 10 GHz, which supported 10G devices. More recently, products with 25-GHz clock speeds have become available and 50-GHz chips are beginning to reach the market. With those drive electronics, NRZ transceivers can achieve base rates of 25G and 50G, respectively. These technologies provide the foundation for 100G data rates.

NRZ PATHS TO 100G

The first generation of 100G transceivers is based on four lanes of 25G traffic. With this basic approach, a transceiver can achieve 100G transmission using various architectures. In theory, it is possible to reach 100G data rates using two lanes of 50G NRZ traffic, but that approach is not being pursued commercially at this time.

The most basic division for 100G NRZ implementations is whether to use single-mode optical fiber or multi-mode fiber. The choice of fiber as implications for the rest of the network, including component types. As a result, the choice of fiber involves a number of trade-offs, including cost, performance, and distance, as well as additional factors like energy consumption and footprint.

SINGLE-MODE FIBER

Single-mode fiber is fabricated with an extremely small core diameter (8 µm to 10 µm). As a result, it supports only one propagation mode; all others will dissipate rapidly. It is typically used with laser-based transceivers operating around 1.3 µm and 1.5 µm. Single-mode fiber is designed to be extremely low loss and provides good signal integrity over long distances.

The next division is whether the architecture is based on transmitting a single spectral channel of 25G data over multiple fibers or multiple 25G wavelengths over a single fiber.

The multi-fiber approach, known as parallel-single-mode (PSM) transmission, involves launching a single wavelength over each of four fiber lanes. The transceivers use NRZ modulation, minimizing the cost of the optoelectronics and electronics. The drawback is fiber cost. Bidirectional communications requires the installation of two fibers for each lane. Past a certain distance, the cost of fiber can become greater than that of the optical components.

An alternative to PSM is the multi-wavelength approach, known as wavelength-division multiplexing (WDM). WDM for 100G involves launching signals into the fiber at four different wavelengths produced by four separate25G lasers inside the transceiver. A multiplexer integrated into the transceiver module combines the wavelengths into a single optical signal that is launched in the fiber. At the far end of the link, the signal is demultiplexed to recover the four initial data streams.

WDM is a mature technology widely used in both metropolitan-area and long-haul networks. It reduces amount of fiber installed by 75% compared to PSM implementations. On the downside, the electronics are both more complex and more expensive than for PSM networks. Below a certain distance, the cost savings are exceeded by the investment required for the electronics.

MULTIMODE FIBER

Multimode fiber provides an alternative to single-mode cabling. Multimode fiber is fabricated with a much larger diameter core (50 µm to 62.5 µm), enabling it to support multiple propagation modes simultaneously. The wider core makes the fiber easier to work with in terms of splicing and alignment.

Multimode fiber is frequently used with shorter wavelength optical equipment such as vertical-cavity surface-emitting lasers operating at 850 nm, which tend to be less expensive than their single-mode fiber counterparts. That cost savings is counterbalanced by a higher cost for the fiber compared to single-mode fiber. Past a certain distance, the cost advantage of using lower-priced optics disappears.

Multimode fiber has significantly higher modal dispersion than single-mode fiber, which shortens the reach. Historically, that was not a problem in the data center. The distances involved fell well within the range of multimode fiber and were sufficiently short that it remained cost-effective. A number of trends are changing those arguments and swinging data-center installations toward single-mode fiber.

. Data centers are getting larger: Hyper-scale data centers are built around tens of thousands of servers with footprints of up to 10,000 ft^2. Intra-data-center interconnects can be as long as 2 km.

. Data rates are getting faster: Early classes of multimode fiber were only rated to support 10G transmission. More recent versions are designed for laser-based data rates of 40G and 100G but the cost increases.

As with single-mode fiber, multimode fiber can be used in parallel-fiber implementations or in spectrally multiplexed implementations. Because of the cost of the fiber, the expense increases rapidly for parallel-fiber implementations. As a result, multimode fiber tends to be reserved for shorter reaches. Another alternative is to use WDM. This can be challenging with multimode fiber, however, because the dispersion characteristics lead to much wider wavelength spacings than for single-mode fiber versions.

PAM-4-based approaches to 100G

Until now, we've been discussing NRZ architecture-based approaches for direct-detection 100G. For second-generation 100G networks, an effort is underway to develop a modulation-based approach that would enable a network to transmit 100G over a single wavelength and a single fiber. This single-lambda 100G approach leverages a technique known as four-level pulse-amplitude modulation (PAM-4).

Recall from our earlier discussion of NRZ, the optical signal varies in amplitude between fully on and fully off. We can consider this a two-level PAM scheme (PAM-2), although in practice that term is not used. In PAM-4, the electronics driving the laser enable it to encode data on the optical signal using four different amplitude levels (see figure 3). As a result, the amount of data a PAM-4 transceiver can send is double that of an NRZ transceiver. This enables a 50G laser to send data at 100G at a single wavelength and over a single duplex fiber lane.

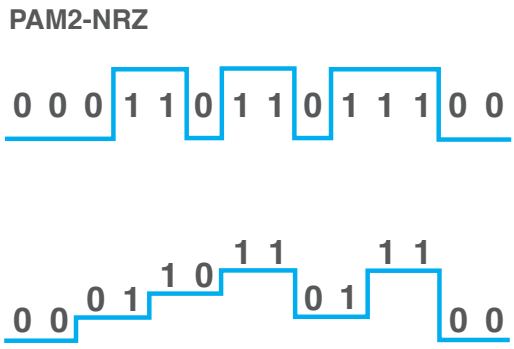

Figure 3: Four-level pulse-amplitude modulation (PAM-4) can send four bits per cycle, doubling the data rate compared to NRZ on-off keying (PAM-2).

The approach offers the best of both worlds. It minimizes the fiber investment and the complexity of the optical components. There is no need for a multiplexer/demultiplexer, for example. Spectral dispersion becomes much less of a concern.

On the downside, PAM-4 is a more complex modulation scheme than NRZ. Transceivers now require digital signal processors, which increases the part count, size, and cost. The addition of two more amplitude levels decreases signal-to-noise ratio. This would reduce range, but for many data center applications, that may not be an issue. The technology is in wide development, with recent announcements around key electronic components such as DSPs, trans-impedance amplifiers, and more.

Development of single-lambda 100G technology is important for another reason. The IEEE has already identified single-lambda 100G as a core enabling technology for the 200G and 400G generations in the near term. Increasing speeds to 200G and 400G would simply require building a multi-lane 100G network using the same PSM or WDM approaches described above.

Although coherent transmission schemes provide alternative paths to 400G, they are far more complex than expensive. Perhaps just important, the electronics and optical components required for coherent detection schemes result in a form factor that is significantly larger than for the direct detection transceivers. Faceplate density and data center applications is enormously important. As a result, data center architects will be pushing to stay with the smaller form factors for as long as possible.

Amid skyrocketing bandwidth demand, data-center architects and operators seek reliable, compact, cost-effective solutions for achieving 100G. In the short term, that is taking the form of 25Gx4 architectures. Whether to choose parallel fiber or WDM implementations depends on the specifics of the actual installation. Long term, single-lambda 100G will fill the need for not just 100G but 200G and 400G operation. Although the eventual roadmap calls for coherent transmission schemes, direct detection will be in place for the foreseeable future.

Contact AddOn Networks to learn more about 100G upgrades and implementation.

AddOn is North America's largest provider of compatible network upgrades and connectivity products, offering compelling value to partners throughout the channel since 1999. AddOn is well known throughout the industry for setting standards of quality and reliability. Providing trusted and tested solutions is really at the heart of what we do. We have invested over a decade of research and resources to creating an internal model to maintain full compliancy for all of our products. AddOn strives to make sure that our partners and their customers have the same confidence in our products as we do.

addonnetworks.com