addon

Data Center Interconnects

Tackling the Special Challenges of DCIs

White Paper

Tackling the Special Challenges of DCIs

Last year, the United States alone created 2.66 GB of data per minute. By the year 2025, analysts estimate that the world will generate 163 ZB of data annually. It may be in the form of videos, photographs, and social-media communications or business records such as online transactions, or communications between the many devices linked in the Internet of things. For businesses around the globe, data is their most valuable and, frequently, their biggest challenge. It is not enough to simply capture the data. It needs to be transported over data-center interconnects (DCIs) to the point of use, whether across the data center for business analytics, across town for back-up, or served up to end-users as streaming video and social media feeds.

In terms of hardware, data centers operate at a completely different scale from long-haul networks. A central office in a long-haul network might require dozens of transceivers; a hyperscale data center can contain more than 1000. DCIs can be classed as intra-data center¾between racks in the same building or facilities on the same campus or inter-data center, linking facilities that are geographically separated by as much as 80 km.

In long-haul networks, performance trumps all. Inter-data center interconnects operate under the same constraints as metropolitan area networks, putting the focus on speed, distance (2 km to 80 km), and spectral multiplexing. Intra-data center DCIs (up to 2 km) must operate within a different set of constraints, including reliability, form factor, and energy efficiency, as well as practical issues like availability, ease-of-use, and above all, cost effectiveness. Here, we discuss each of these factors in turn, with a focus on current solutions.

DCI performance requirements

HIGH DATA RATE

The data rates vary depending on the application. Intra-data center DCIs for a traditional data center would probably use 10 Gb Ethernet while a hyperscale data center might need 40 Gbps or even 100 Gbps. The type of transceivers used depend on both speed and application.

Understanding data rate starts with the concepts of baud and bit rate. The baud is the number of state changes (symbols) in a single cycle. The bit rate is the number of bits transmitted per second. One of the simplest modulation approaches is on-off keying, most commonly implemented in non-return-to zero (NRZ) format. In this binary amplitude-modulation scheme, the laser generates a data stream by operating either at either maximum intensity (logical 1) or zero intensity (logical 0). NRZ transmits one bit per baud and two baud (or two bits) per cycle. In other words, baud and bit rate are identical.

The problem with NRZ is that the bit rate is limited by the switching speed of the lasers. Semiconductor lasers can be modulated by switching the current (direct-modulated lasers) or by use of external electro-absorption modulators (externally modulated lasers). In both cases, the modulation rate is limited by the clock speed of the drive electronics. For years, the rate was stalled out at 10 GHz. 25 GHz devices are now available and some 50 GHz products are beginning to reach certain markets.

Because of the speed increases in electronics, DCIs can operate at 10 Gbps and 25 Gbps using NRZ modulation. The jump to higher speeds requires either more advanced architectures or more sophisticated modulation schemes. Architecture-based approaches provide an easier route, achieving 100 Gbps data rates by running four lanes of 25 Gbps simultaneously.

The two most common architectures are wavelength-division multiplexing (WDM) and parallel single-mode (PSM) architectures. In WDM, multiple tightly spaced (5 to 50 nm) spectral channels run down the same fiber, which is typically single mode. PSM delivers the same capacity by running one channel down each of several (typically four) individual fibers. In both cases, bidirectional communica-tion requires duplex fibers.

DCIs generally use coarse WDM (CWDM) formats defined for four channels, and with wider channel spacings than seen in the dense WDM grids used in the metro and long-haul space. An alternative is short-wave WDM (SWDM) which is defined for devices operating around 850 nm (see figure 1). These systems typically use multimode fiber, which is more costly than single-mode fiber, but the optoelectronic components are less expensive. The combination makes SWDM a viable alternative for short-reach networks.

(fig.1)

Figure 1: Short-wave WDM takes advantage of transceivers operating at 850 nm, like the 0061004008-AO, to support economical, short reach interconnects over multi-mode fiber.

CWDM is an effective approach to achieve a 100-Gbps bit rate but it requires four times the amount of optoelectronic components, plus multiplexer/multiplexer (MUX/DEMUX) stages for each link. This adds cost, complexity, and points of failure.

PSM provides a good alternative to CWDM for the right application. PSM can be implemented using either a single laser source for the link with a separate modulator for each fiber, or a dedicated DML for each fiber. The approach uses fewer and less sophisticated components compared to CWDM. Because each of the four lanes must be duplex, however, the infrastructure cost is considerably higher, particularly as distance increases. PSM is best used below 500 m.

Modulation-based approaches

We can also boost bit rate by applying more sophisticated modulation techniques to transmit multiple bits per symbol. One of those techniques is based on pulse-amplitude modulation (PAM).

NRZ on-off keying is a binary, or two-level PAM format. Strictly speaking, we could refer to it as PAM-2, although that is rarely done. If we modulate the laser to a total of four different amplitudes, instead of just fully on and fully off, the system can now code for 00, 01, 10, and 11. This enables it to send four bits per cycle or two bits per symbol (see figure 2). This modulation format is referred to as PAM-4. Compared to NRZ modulation, PAM-4 delivers double the bit rate for the same baud.

(fig.2)

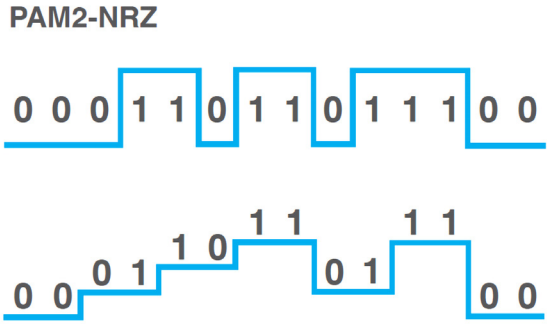

Figure 2: Four-level pulse-amplitude modulation (PAM-4) can send four bits per cycle, doubling the data rate compared to NRZ on-off keying (PAM-2).

With PAM-4 modulation, a transceiver with a single laser driven by 25 GHz electronics can operate at 50 Gbps. A two-laser version implemented with a two-channel CWDM system or a two-fiber PSM link can reach 100 Gbps (two lanes at 50 Gbps each). With 50 GHz electronics, a PAM-4 transceiver can reach 100 Gbps in a single lane. A 4 x 100 Gbps architecture provides a straightforward avenue to 400 Gbps operation; indeed, PAM4 is one of the technologies identified by the IEEE P802.3bs Task Force to achieve 400 Gbps in the DCI space.

PAM-4 is a relatively simple modulation scheme. It provides a path to higher data rates while minimizing hardware. On the downside, it requires four different optical amplitudes, which reduces signal-to-noise ratio (SNR). As a result, the approach is best applied to shorter distances.

Until now, we have focused on incoherent modulation schemes - those that do not involve phase. Schemes have been developed that are coherent, meaning that they modulate both the phase and amplitude of the signal, as well as in some cases the polarization. In quadrature phase shift keying (QPSK), the system sends two separate amplitude-modulated data streams with phases in quadrature relative to one another (offset by 90°). The quadrature approach doubles the number of bits sent per symbol, sending two bits per symbol instead of one. If the data streams are sent over two different polarizations, the scheme adds two more bits per symbol for a total of four.

We can combine phase and amplitude modulation to achieve even greater data rates (see table). 16-level quadrature amplitude modulation (16-QAM), for example, adds two bits per symbol via QPSK and another two bits via amplitude modulation for a total of four bits per symbol. Sent over two distinct polarizations, 16-QAM results in eight bits per symbol. The highest capacity currently available is with 64-QAM, which sends six bits per symbol over two distinct polarizations, for a 12-fold increase in data rate compared to NRZ on-off keying. The trade-off for this increase in bit rate is higher complexity, higher cost, and significantly larger module sizes, all of which can be problematic in the DCI space.

Table 1: Additional bits per symbol provided by higher-order modulation schemes

GBaud Rate	Polarization States	Coding	Bits/Symbol	Bits/Second (Gbps)	Transmission Capacity (Gbps)
32	2	BPSK (OOK)	16	45	0
		QPSK	2	128	100
		16-OAM	4	256	200
		64-QAM	6	384	300
64	2	PPS K1		128	100
		QPSK	2	256	200
		16-QAM	4	512	400

Reliability

Even the fastest modulation scheme is useless if the network is down. Reliability is essential to network performance. Choosing and end-to-end proprietary solution may seem like the obvious choice, but that may involve some components with substandard performance. Specifying individually selected best-in-class hardware addresses performance concerns but introduces challenges of its own.

One of the most common misconceptions in designing a network is that all systems are generic, that those best-in-class components can be put together easily. The reality is that each network is unique and purpose-built for the application. The process of qualifying parts needs to take that into account.

Build the actual system and test it end-to-end under actual operating conditions. Temperature, for example, can be a problem. Semiconductor lasers exhibit wavelength drift as a function of operating temperature. Parts expand and contract, so ensure that the system as designed will both perform to specifications and deliver on lifetime requirements.

Firmware can be another problem. Even within the same vendor, different product lines may have different firmware, or the firmware may need to be handled differently depending upon the application. It may need to call different lines of code or examine different bytes. If fixes are necessary, it is better to determine that in the laboratory than in the data center or central office.

Practical DCI requirements

Many of the requirements for intra-data center DCI hardware are driven less by performance demands than by practical considerations. The equipment needs to fit in the allotted space, be easy enough to use that it doesn't delay installation or repairs. Above all, it needs to be cost-effective.

HIGH DENSITY

The high volumes of transceivers and switches in the data center create a constant pressure for both small size and interoperability. Interfaces need to support multiple lanes of traffic. At the same time, the form factors need to be designed to maximize the faceplate density the number of devices that can be installed across the width of a standard rack. These motivations have led to the development of several multi-source agreements specifying packaging, which encompasses the housing dimensions, electrical interfaces, power budgets, optical interfaces, etc.

The most common form factor in the intra-data center DCI space is the quad small-form factor pluggable (QSFP) family of packaging specifications (see figure 3). QSFP interfaces support four channels operating at various speeds for aggregate data rates of 4 Gbps (QSFP, 4 x 1 Gbps), 40 Gbps (QSFP+, 4 x 10 Gbps), 200 Gbps (QSFP14, 4 x 14 Gbps), and 400 Gbps (QSFP28, 4 x 28 Gbps). In addition, work is underway on QSFP-DD, which stacks another layer of contacts on top for a total of eight lanes. This will support 200 Gbps NRZ implementations (8 x 25 Gbps) and 400 Gbps PAM-4 implementations (8 x 50 Gbps) while maintaining high faceplate densities.

(fig.3)

Figure 3: The QSFP28-100GB-LR4-AO is a 100 Gbps QSFP transceiver suitable for distances of up to 10 km.

Coherent transceivers require more components, and more space, than amplitude-modulated devices. To accommodate these needs, the Compact Form Factor Pluggable multi-source agreement (MSA) the created the compact form factor pluggable (CFP) package, an 82-mm-wide package that could support 100 Gbps coherent transmission (10 x 10 Gbps or 4 x 25 Gbps). Although the CFP worked well for coherent devices, it was far too big for use in DCIs. In response to popular demand in the metropolitan-area networking space, the group released the CFP2 specification, which called out a package roughly half the size of its predecessor. It could deliver 100 Gbps (10 x 10 Gbps or 4 x 25 Gbps), 200 (8 x 25 Gbps), or 400 Gbps (8 x 50 Gbps) networks. The form factor is more appropriate for inter-data center DCIs (see figure 4).

More recently, the MSA has delivered two additional options: the CFP4 and the CFP8. With a width of 21.5 mm, the CFP4 offers very nearly the same faceplate density as the QSFP package. It can support 4 x 10 Gbps or 4 x 25 Gbps coherent transmission. The CFP8 is roughly the same width as the CFP2 and supports 400 Gbps operation (16 x 25 Gbps or 8 x 50 Gbps). Among the key differences between the CFP2 and the CFP8 are that the CFP8 supports a higher pinout and double the power usage.

(fig.4)

Figure 4: Network designers have a choice of transceiver form factors, including (from left) the CFP, the CFP2, the CFP4, and the QSFP28.

In addition to the packages above, work is underway on another specification, the Octal Small Form Factor Pluggable (OSFP). It is designed to accommodate eight lanes of high-speed signaling. It can be operated in one of two modes. When used with 25 GHz PAM4, it can house a 400 Gbps transceiver (8 x 50 Gbps). When used with 25 GHz NRZ signaling, the OSFP can deliver 200 Gbps (8 x 25 Gbps).

It is important to note once again that coherent transition transmission remains a rarity in the intra-data center DCI space. As time goes on and bandwidth demand rises, however, data center owners will need to adapt. These form factor packaging options will help.

EASE-OF-USE

Another key requirement for the DCI space is ease of use. The form factors described above are all pluggable and hot-swappable. For DCIs incorporating CWDM architectures, tunable lasers are essential to minimize inventory for sparing, etc.. This is a case in which being able to choose best in breed components rather than an end-to-end proprietary system can provide significant benefits. Look to suppliers who are able to customize components for specific system requirements.

ENERGY EFFICIENCY

The components used in DCIs need to be as efficient as possible. The reasons are twofold. As mentioned previously, heat degrades laser performance, impacting performance. It also reduces lifetime of the electronics so that the net result is both reduced signal quality and increased failures. The high packing density complicates heat dissipation even with the use of active technologies such as fans. Minimizing loss reduces the scope of the problem. Energy efficiency is also a concern as far as cost of operation. Data centers consume a tremendous amount of energy, both to run the equipment and to power the environmental controls. Look for equipment with good efficiency levels for optimal results.

INTEROPERABILITY AND OPENNESS

With the sheer volumes involved in DCIs, as well as the great variety in use cases, network designers need the freedom to choose the best component for their application, whether in terms of performance, availability, or cost. Multi-source agreements and standards have been developed to support that level of interoperability. This enables vendors to provide cost-effective alternatives to OEM hardware, giving asset owners the opportunity to choose more economical solutions (see figure 5).

(fig.5)

Figure 5: The CFP4-100G-LR4-AO, a CFP4 transceiver, provides 100GBase-LR4 throughput up to 10km over sin-gle-mode fiber (SMF) at a wavelength of 1310nm using an LC connector. It is guaranteed to be 100% compatible with the equivalent Cisco® transceiver. It has been programmed, uniquely serialized and data-traffic and applica-tion tested to ensure that it will initialize and perform identically.

One common misconception with OEM systems holds that swapping in components from a third-party supplier will void the warranty. This is not true. The Sherman Antitrust Act makes it illegal for OEMs to invalidate a warranty or other support if an asset owner buys an upgrade from a different company. This law was put into place to protect consumers and end-users. As mentioned above, however, it is essential to perform systems-level testing with the candidate components to ensure that the network will operate as intended with the new hardware.

The data center environment imposes a number of very specific requirements on hardware, software, and implementation. Speed is important but bit rate alone is not sufficient. The industry has developed a number of options for solving the challenges of DCIs. By choosing the component that will best serve the application and testing to ensure performance, end-users and service organizations can expect reliable performance for a reasonable price.

Contact AddOn Networks to learn more about economical, high-performance components to support your DCIs.

About AddOn Networks

AddOn Networks is the global leader in optical connectivity solutions serving data center, enterprise, government, education, and healthcare provider networks.

We operate in over 25 countries through our long-standing commercial channels to provide continuity of supply and world-class service.

Find an expansive network catalog from legacy GBICs to cutting-edge 100G, 400G active and passive solutions. Every optic we ship is first programmed and tested to a 99.98% reliability rating in house. We carry solutions that are compliant with the most stringent global compliance standards including ISO 9001:2015, TAA, RoHS, NEBS Level 3, and more. Get 100% compatible optics in form and functionality across 100 OEM manufacturers, covering more than 20,000 systems and platforms.

At AddOn, we're invested in your complete success. All optical solutions are backed by lifetime warranty and 24/7/365 global field engineering support.

North America
sales@addonnetworks.com

15775 Gateway Circle
Tustin, CA 92780
+1 877 292 1701

EMEA
salessupportemea@addonnetworks.com

Eagle House, Lakeside Business Park,
South Cerney, Gloucestershire, GL7 5XL
+44 1285 842070

For more information, please visit **www.addonnetworks.com**