BROADCOM®

White Paper

CPO BiDi

An Efficient Solution for Scale-up of AI/ML Clusters Using Optics

Executive Summary

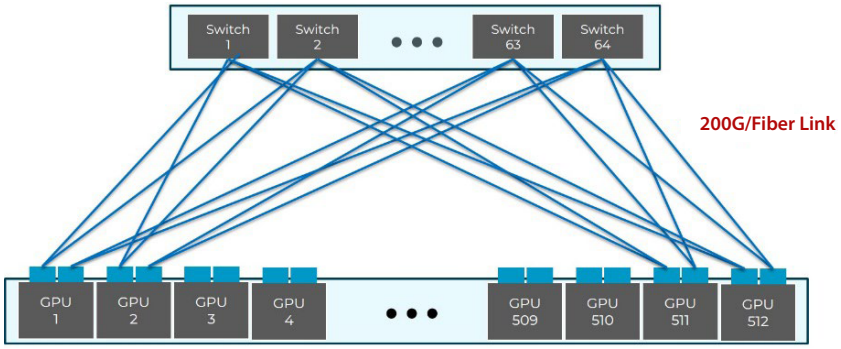
As the need for AI/ML clusters with higher GPU counts grows, there will be a meaningful shift to use optics for building out these clusters. For a 512 to 1024 scale-up GPU cluster, compared to a traditional DR-based co-packaged optics (CPO) breakout solution, CPO using a bi-directional (Bi-Di) technology can lower the cost of optics by as much as *15%* and enable a more efficient way to construct scale-up networks.

Large Cluster Size Scale-up with Optics

Scale-up network architectures for latency-sensitive applications, such as training large language models (LLMs), require a single switching stage fabric for all-to-all connectivity between the compute nodes that form a cluster. Although optics are used extensively in the back-end network fabric of AI/ML clusters for scaling out the network across multiple nodes, their use has been limited from a scale-up perspective. This is because up until recently, compared to optics, copper interconnects have been able to support the reach and bandwidth density requirements for scale-up networks at a lower cost point. However, with serialized data rates pushing past 200G/lane, passive copper interconnects are unable to extend beyond the physical distances within a single rack. Efforts to increase the reach of copper interconnects using electrical retimers come with significant power and latency penalties. Additionally, the power dissipation of individual GPUs limits the quantity of GPUs that can be placed in a single server rack without requiring elaborate cooling methods that add to the overall cost of deployment and ownership. Recent examples of 72 GPUs interconnected in a rack have shown to require not only liquid cooling but also extensive engineering of the copper links that interconnect them (*NVIDIA Blackwell Architecture Technical Brief*).

Given these challenges, it is inevitable that fiber optics, with a negligible loss budget penalty for traversing longer reaches, will replace copper interconnects for scaling up GPU clusters beyond the constraints of a rack. As illustrated in Figure 1, a 102.4T switch with a radix of 512 could support 12.8 Tb/s of all-to-all connectivity (over fiber) between 512 GPUs with a single layer of switching. In this case, each GPU would have 12.8 Tb/s of interconnect bandwidth (in each direction) connecting it to each of the 64 switches via a 200-Gb/s fiber link. By not being limited to the physical distances within a single rack, optics can also ease constraints on the number of GPUs that need to be placed in a rack and reduce the cost of cooling while simultaneously supporting even larger scale-up clusters with high-radix switches.

Figure 1: 512 All-to-All GPU Cluster with Breakout of 200G/Lane Enabled by DR Optics and a 512-Radix 102.4T Switch (Note: The bandwidth of each GPU is 12.8T per direction.)

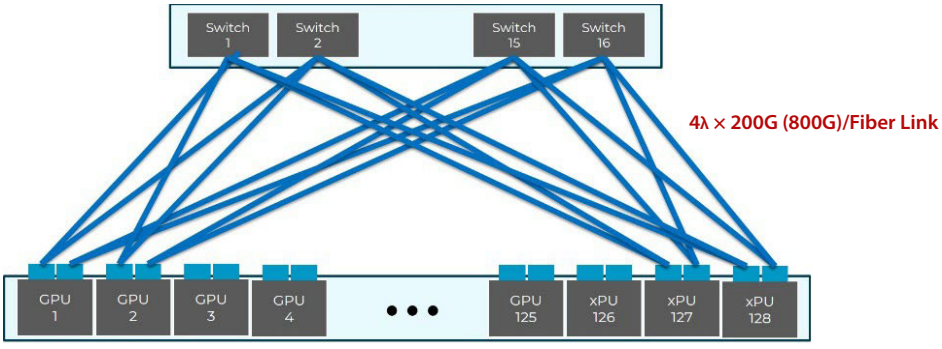


Having outlined the advantages of optics and their need to scale up large GPUs clusters, the rest of this white paper compares the different ways to build these scale-up clusters where the bandwidth density per fiber of an optical link connecting a compute node to the switching layer must be properly matched with switch capabilities from a perspective of the radix, bandwidth density, fiber plant management, and overall cost of ownership. We show that for optimal efficiency and total cost of ownership (TCO), a bi-directional approach is best suited to support scale-up clusters with optics.

DR vs FR4 for Scale-up

To scale bandwidth per fiber in a cost-effective manner, the FR4 standard was adopted. The FR4 standard consists of four wavelengths on a fiber with each wavelength carrying 100 or 200 Gb/s depending on the line rate. Although FR4 is efficient from a bandwidth and fiber-plant-management perspective, it limits the size of the scale-up cluster. Contrasting FR4 with the example shown in Figure 1, which is based on the DR optical standard, Figure 2 shows that with FR4 optics and the same switch port density, the scale-up cluster size is limited to 128 GPUs each with 800G (4×200G) of connectivity to 16 102.4T switches. Clearly, FR4, which is optimal for scale-out architectures from a perspective of fiber bandwidth efficiency, limits the size of scale-up architectures. However, the increased granularity afforded by an optical breakout solution, such as DR optics, comes at the cost of increased fiber plant complexity (4×). In the next section, we look at the different options for optical breakout while preserving bandwidth granularity per fiber.

Figure 2: 128 All-to-All GPU Cluster with Breakout of 800G/Lane Enabled by FR Optics and a 512-Radix 102.4T Switch



PMD Options for High-Radix Scale-up: DR vs Bi-Di

There are two ways in which the physical media dependent (PMD) layer of the network can be constructed via optical breakout to support the bandwidth granularity needed for high-radix scale-up clusters with optics: 1) Using DR optics on parallel single mode (PSM) fiber or 2) Using bi-directional optics on PSM fiber. The following is a brief description of each technology.

1. **DR:** DR optics technology is widely deployed in data centers and leverages PSM fiber to enable high-speed data transmission for reaches up to 500m. Each DR-optics-based link consists of a pair of fibers, where one fiber is used exclusively for transmit and the other fiber is used exclusively for receive. Given the separation of transmit and receive transmission onto two fibers, DR optics use a single (common) wavelength for communication.
2. **Bi-Di:** Bi-Di technology refers to the capability of transmitting and receiving data simultaneously over the same optical fiber. This technology enhances the efficiency and capacity of optical communication systems by enabling bi-directional communication along a single strand of fiber. Bi-Di systems typically use wavelength division multiplexing (WDM) techniques, allowing different wavelengths of light to carry data in both directions simultaneously. With this approach, Bi-Di technology optimizes the utilization of available optical infrastructure and lowers overall costs. It is ubiquitous in fiber-to-the-home (FTTH) installations, reducing the need for additional fiber installations and delivering reliable and high-performance connectivity. We believe that Bi-Di technology can revolutionize, in a similar way, how optics can be used to scale up GPU clusters.

DR vs Bi-Di: Network-Level and Device-Level Architecture Differences

The block diagrams in Figure 3 and Figure 4 illustrate the differences in fiber I/O between a DR and Bi-Di implementation of a 12.8T optical engine (OE). In the case of DR, because the transmit and receive signals are on separate fibers, a total of 128 fibers (64 Tx + 64 Rx) are required to connect the 12.8T engine to the network. The same connectivity can be achieved by 64 fibers in the equivalent Bi-Di implementation because each fiber carries two wavelengths: λ_1 = 1270 nm; λ_2 = 1310 nm. Also, as shown in Figure 4, there are two equal sets of fibers escaping the engine for Bi-Di Group A with transmit on λ_1 and receive on λ_2 and Group B with transmit on λ_2 and receive on λ_1.

Figure 3: 12.8T Optical Engine Fiber I/O with DR Optics

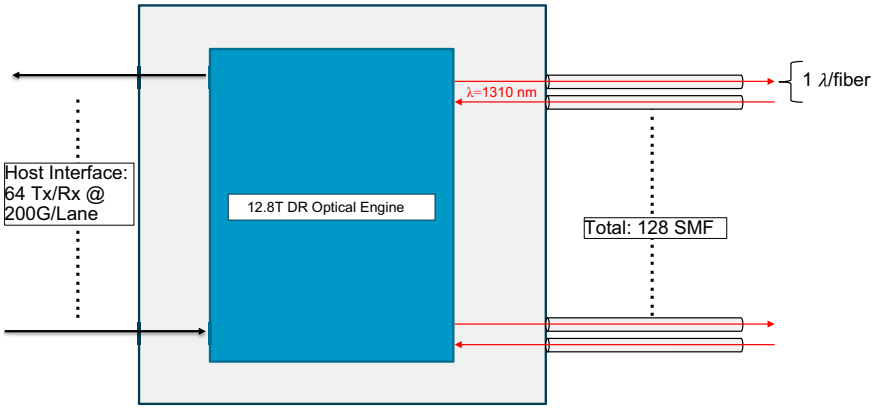
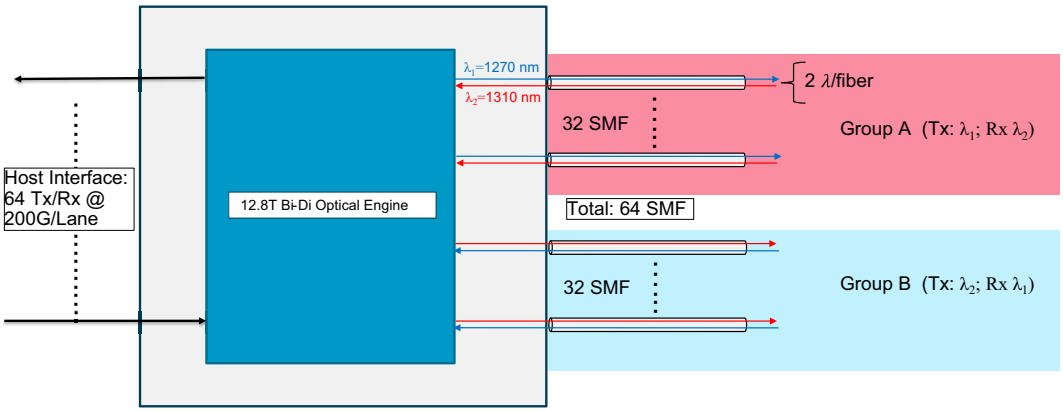


Figure 4: 12.8T Optical Engine Fiber I/O with Bi-Di Optics



It is important to note that in a network deployment of a Bi-Di solution using CPO, the fact that there are two groups of fiber I/O based on the Tx and Rx wavelength does not reduce the radix connectivity compared to DR. Nor does it impose different hardware requirements on the two sides of the link; that is, both sides of the link will have the same set of hardware (optics, electronics, and so on) on the OE. The only constraint imposed is that the each fiber port from Group A must be connected to a Group B fiber port on the other side of the link. Figure 5 shows the interconnection between the Group A and Group B fiber ports for two OE pairs using Bi-Di.

Figure 5: Point-to-Multipoint Connectivity Using Bi-Di

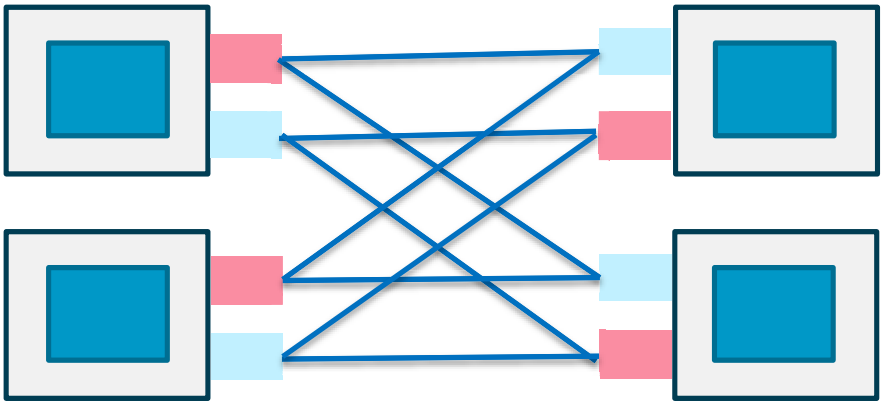
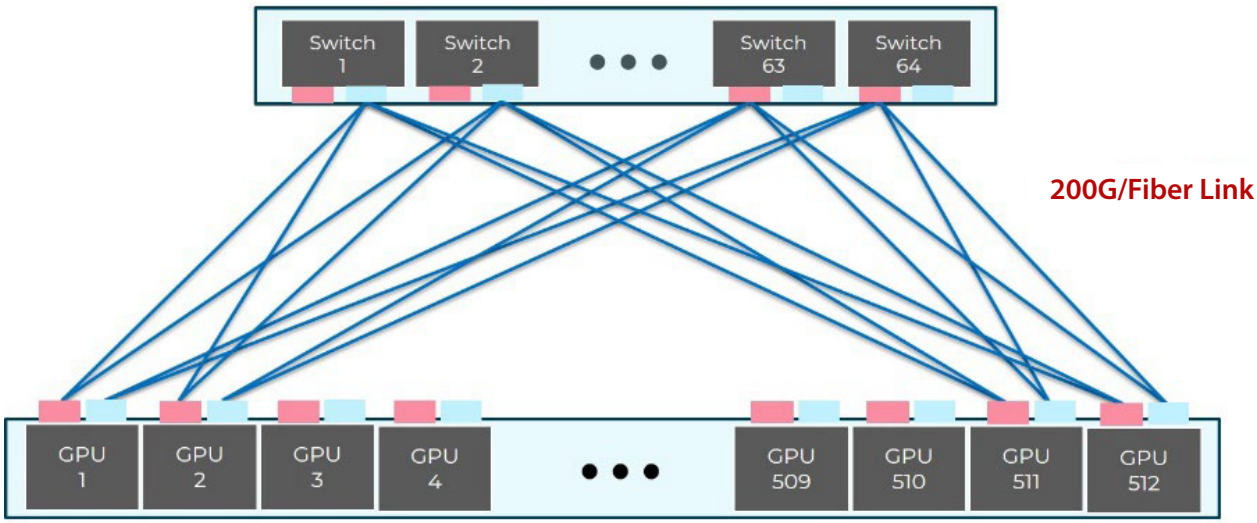


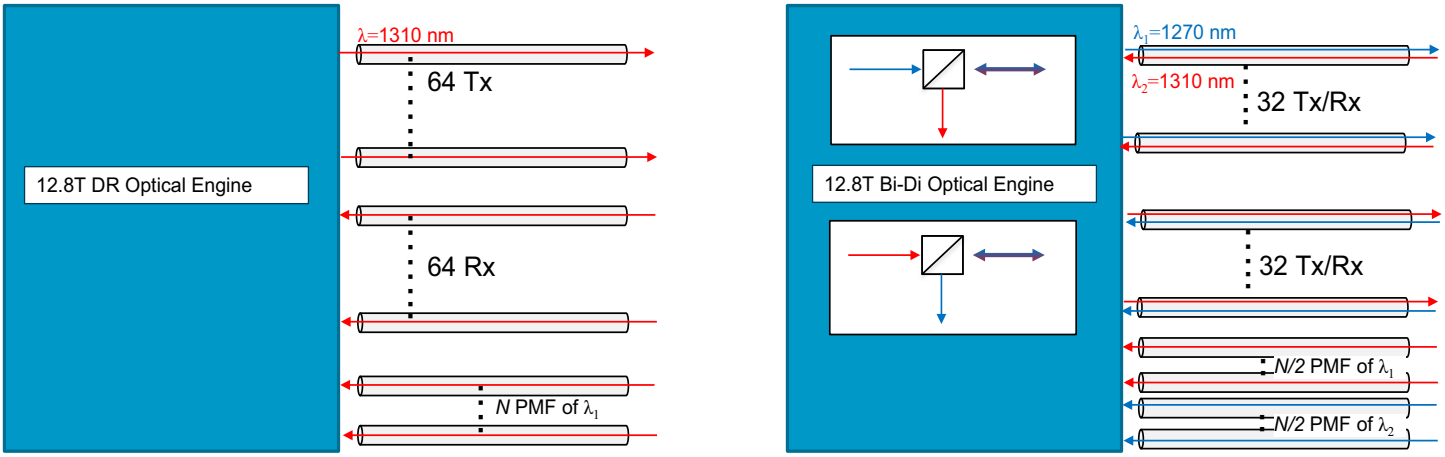
Figure 6 shows a scale-up cluster with Bi-Di optics that is identical in size to the one shown in Figure 1. Each switch and GPU has half its ports allocated for Group A and the other half allocated for Group B. Even in the case of a non-CPO implementation (that is, pluggable modules) on the switch side, as long as the Bi-Di pluggable transceiver module has an even number of lanes, no separate SKU is required. Half of the lanes will be allocated to Group A and the other half to the Group B type fiber port.

Figure 6: 512 All-to-All GPU Cluster with Breakout of 200G/Lane Enabled by Bi-Di Optics and a 512-Radix 102.4T Switch



At the engine level, the optics differ between the two implementations, as shown in Figure 7. In the case of the DR implementation, a MUX/DMUX is not required because the transmit and receive paths are not shared. In the case of the Bi-Di implementation, a diplexer is required in the photonic integrated circuit (to separate the wavelengths of the transmit and receive signal). This difference does add a slight power penalty (< 1 dB) due to the added loss of the diplexer, which shows up as an ~10% increase in the laser optical power into the engine. Additionally, assuming that the optical engine is serviced by an external light source, both implementations will require polarization maintaining fiber (PMF) connections to an external light source. In the case of Bi-Di, half of the lasers will be of wavelength λ_1 and the remaining half will be of wavelength λ_2. The 40-nm spacing between the two wavelengths ensures that there is sufficient wavelength margin to account for any temperature differences in the optics between the two endpoints of a Bi-Di link.

Figure 7: Device-Level Differences between DR and Bi-Di Optics



DR vs Bi-Di: TCO Comparison

To realize the full benefit of Bi-Di for building scale-up clusters, CPO should be implemented on both ends of the link, that is at both the GPU side and the switch side. Table 1 captures the complexity and cost-factor differences between DR and Bi-Di for a 512-GPU cluster where the average link length is assumed to be 30m. Even considering a 10% higher cost for the laser modules in the Bi-Di implementation, the overall systems cost savings on the optics is ~15% and scales with the average link length from the GPUs to the switches in the cluster, as shown in Figure 8. The primary cost savings from a Bi-Di implementation comes from the reduction in fiber infrastructure and a cost-effective fiber-attach solution at the engine level. The cost-and-complexity comparison does not take into consideration the operational challenges of routing such a large number of fibers in a physically constrained space where DR optics are even more complex than Bi-Di with 2× the fiber quantity required. Also, as noted earlier, the number of external lasers servicing an engine does not change between a Bi-Di and DR implementation. Furthermore, the wavelengths for a Bi-Di implementation are chosen such that they conform to CH0 (1264.5 nm to 1277.5 nm) and CH2 (1304.5 nm to 1317.5 nm) of the CWDM wavelength grid. This allows for Bi-Di technology to leverage a mature market for laser sources whose cost can be driven down in time.

Table 1: Comparison of the Cost and Fiber Complexity between a DR and Bi-Di Implementation of a 512-GPU Cluster

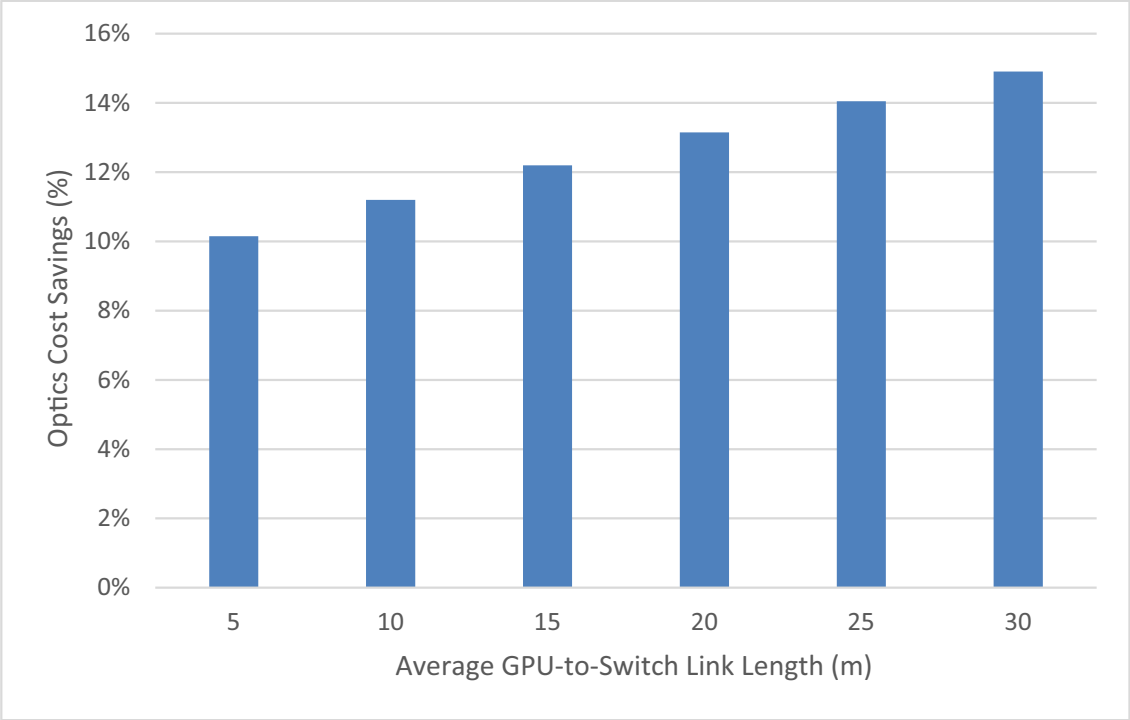
512-GPU Scale-up Cluster with 64 CPO Switches and 512 12.8T CPO Engines		
	DR	**Bi-Di**
Number of SMF fiber cable bundles inside CPO switches and 12.8T engines[a]	2048	2048
Number of fibers per cable bundle	64	32
Cost factor of fiber cable bundles	2x	x
Number of 30m fiber links in cluster	65,536	32,768
Cost factor of SMF links in cluster	2y	y
Laser modules count[b]	2048	2048
Cost factor of laser modules[c]	z	1.1z

a. Assumes that each 12.8T engine requires two cable bundles.

b. Assumes that each 12.8T engine requires two laser modules.

c. Assumes that an additional 10% laser output power for diplexer loss translates to a 10% cost overhead on the laser modules for Bi-Di.

Figure 8: Optics Cost Savings (%) as a Function of GPU-to-Switch Fiber Link Length Using Bi-Di Optics for a 512-GPU Scale-up Cluster



Conclusion

As the need for AI/ML clusters with higher-count GPUs grows, there will be a meaningful shift to optics for scaling up these clusters. In this paper, we have shown that CPO using Bi-Di technology can provide up to a 15% cost savings on the optics compared to conventional DR optics. Additionally, as GPU clusters for scale-up networks grow in size with longer link lengths and higher-radix switches, the cost savings will only increase further, making Bi-Di even more attractive.