DATA CENTER NETWORK DESIGN AND TECHNOLOGIES

MAHESH SUBRAMANIAM
MICHAL STYSZYNSKI
HIMANSHU TAMBAKUWALA

Foreword by JEFF DOYLE

AI DATA CENTER NETWORK DESIGN AND TECHNOLOGIES

This page intentionally left blank

AI DATA CENTER NETWORK DESIGN AND TECHNOLOGIES

Mahesh Subramaniam, Michal Styszynski,
Himanshu Tambakuwala

Addison-Wesley

Hoboken, New Jersey

Visit us on the Web: informit.com/aw

Library of Congress Control Number: 2025949447

ISBN-13: 978-0-13-543628-8
ISBN-10: 0-13-543628-1

$PrintCode

Head of Enterprise Content and Training, Enterprise Learning and Skills
Julie Phifer

Executive Editor
James Manly

Development Editor
Ellie C. Bru

Managing Editor
Sandra Schroeder

Senior Project Editor
Mandie Frank

Copy Editor
Kitty Wilson

Indexer
Timothy Wright

Proofreader
Donna E. Mulder

Technical Reviewer
Tanveer Dandiana

Designer
Chuti Prasertsith

Compositor
codeMantra

To my brother, Rajesh, who taught me the meaning of connection—long before I ever studied networks. "Your absence echoes in every page, a quiet hum between the lines.… A bright star in my sky, the legacy I live and build for.…You are the architecture of my soul—my reason to dream, my need to create, my home in every connection."

I also want to dedicate this book to my parents, Sarasu and Subramaniam, my wife, Ramya, and my sons, Rithvik and Rithish. They are my steady roots beneath every life storm, the strength that holds the ground I walk upon. Their love is my unwavering light; that light is the one and only thing which makes me stand.

—Mahesh

I want to dedicate this book to my wife, Kasia, and my sons, Ernest and Marcel, for their love, patience, and understanding. Their encouragement, and joy for life inspire me every day and have helped me complete this book.

—Michal

To the ones I owe it all: Neha, Rachit, Aanvi, and my parents, Prakash and Indu. Thank you for every late night and every quiet sacrifice you made.

—Himanshu

This page intentionally left blank

Contents

Foreword

If you've picked up this book, you're likely not someone who has bought into the media hysterics around AI or who thinks that it's a danger to humanity. Those folks have watched too many *Matrix* or *Terminator* movies. Don't get me wrong, as there are reasons to be concerned about AI and reasons there should be intelligent regulations around some of it. But AI is a tool, nothing more. Like any tool it can be used or misused.

Like it or not, AI is integrated into our everyday lives, often in ways that we don't even notice. Arguably, the first widespread use of AI was in spam filters. When you buy spam detection software, you buy a *database* of patterns common to most spam email and an algorithm (*model*) that watches for characteristics in your email that match those patterns. When it thinks it recognizes (*infers*) spam, it moves the email to an isolated mailbox. You actively *train* your spam filter by periodically reviewing your spam mailbox, identifying email that is not spam, so that the app can better infer spam from new data.

There are some fundamental AI/ML concepts in what I just described: a curated reference database, a pattern-matching model, training, and inference based on new data.

Want another example? The first "smart" doorbell I purchased was little more than a camera, a button, and some motion detection capabilities. The one I have now can differentiate a person from random motion. It can tell me if that person dropped off or picked up a package. With a little training (there's that word again), it can identify friends and family members who come to the door. Through my Wi-Fi network, it connects to a database and a pattern recognition model.

And, of course, there's your smartphone with AI apps from face ID and voice commands, to music and entertainment apps that learn your preferences and make recommendations, to targeted advertisements in your social media, to the ability to "Google" (the company name has become a verb) an answer to almost any question you might have. Your new television likely has AI and maybe your new refrigerator as well. Your new car is swimming with it, giving you directions and keeping you in your lane and telling you if you're following too closely. And as I type this foreword, Microsoft Word follows what I'm typing, suggests sentence completion, and catches spelling and grammatical errors.

Then there are the AI apps that are not quietly hiding in your refrigerator or word processor: generative AI that can create images and music, write academic papers or code, or generate designs; predictive AI that can make forecasts based on analysis of historical data (note that predictive AI can make some embarrassingly bad predictions when analyzing particularly chaotic systems, such as human behavior); and natural language models that allow you to communicate naturally with AI systems and receive human-sounding answers.

And, of course, self-driving automobiles utilize a variety of AI models to operate safely on the road. I personally believe self-driving automobiles are not quite ready for prime time, but I also believe they are our future. There could well be a not-so-distant future in which manually driving a car is considered illegally dangerous.

Finally, as network nerds we've all heard the term "autonomous networks." Networks tend to be highly predictable systems, and with machine learning and analytics we are moving closer to network autonomy and further away from the one factor that accounts for most network outages: humans touching the network.

AI, in its many subgroups, can be simple enough for you to create and run programs right on your laptop. But processing massive databases can be hugely expensive in infrastructure, power, and cooling costs. Massive databases require massive, quickly accessible storage infrastructure. Processing that data within reasonable job completion times requires powerful servers running GPU/TPU/DPU clusters in parallel, sometimes requiring tens of thousands of processors running the same model. In the largest-scale cases the storage, processing backends, and inference frontends can involve not just separate types of data centers but also physically separate data centers. Reliably distributing and collecting data among these processors calls for new kinds of data center networking because storage, processing (training), and inference all have different requirements.

That's where this first-of-its-kind book comes in. My friends Mahesh, Himanshu, and Michal have among them a vast level of experience designing AI data centers. They explain in great detail the unique requirements of AI data centers and the technologies created to address these problems. Many of these technologies are already finding their way into high-performance compute (HPC) data centers and will eventually find their way into normal cloud data centers. So even if you are not an AIDC specialist, it behooves you to become familiar with the technologies and design principles covered here.

–Jeff Doyle

Preface

AI Data Center Network Design and Technologies is a comprehensive, practical guide for network engineers, architects, and technology leaders who are building, scaling, or optimizing the infrastructure that supports artificial intelligence. This book bridges the gap between theory and practice, providing an in-depth look at the design, deployment, and management of high-performance AI data centers—where scale-out and scale-up networking, xPU-based computing, storage, and operations work together.

Register your copy of *AI Data Center Network Design and Technologies* on the InformIT site for convenient access to updates and/or corrections as they become available. To start the registration process, go to informit.com/register and log in or create an account. Enter the product ISBN (9780135436288) and click Submit. Look on the Registered Products tab for an Access Bonus Content link next to this product, and follow that link to access any available bonus materials. If you would like to be notified of exclusive offers on new editions and updates, please check the box to receive email from us.

This page intentionally left blank

Acknowledgments

We take this opportunity to sincerely acknowledge those individuals who have provided insightful meets and encouragement from the very start of our careers, helping us lay a strong foundation for the advancements in networking technologies we see today:

Manish Prakash, Shivam Srivastava, Raghu Subramanian, Gouthaman Guna, Tim McCarthy, Rajesh Sunkara, Ansuha P, John Steele, Jim Lowe, Ken D, Varun Arya, Rob Nath, Jim Hurd, Mathias Kokot, Sanjeev Ugalkar, Kaushal Agrawal, Karthik Vugane, Vinod Subramaniam (Vinod), Praful Lalchandani, Mike Bushong, Praveen Jain, A.E. Natarajan, Mansour Karam, Murali Vemula, Kiran K.L., Raj Yavatkar, T. Sridhar, Jay Ramalingam, Arun Viswanathan, Jasmeet Sawhney, Yogesh Kumar, Chirag Kachalia, Rajesh Dhople, Dharshan Ravichandran, Ashwin Balaji Sundaresan, Kumuthini Ratnasingham, Francis Guiller, Jeff Doyle, Amit Sanyal, Arun Gandhi, Ben Baker, Himansu Sahu, Mallikarjun Tallapragada, Suryanarayana MNV, Suraj Kumar, Harisankar Ramalingam, Rajashekhar Reddy, Parthipan T.S., Balaji Palanisami, Rajesh Venmanad, Manoj Kulandaivel, Sureshkumar Chinnappan, Pavan Kurapati, Dilip Sundarraj, Murugan Kanniappan, Russ White, Amogh Vijayakumar, Vince Loschiavo, Gilbert Montanez, Gustavo Castellanos, Nishal Singh, Kevin Oriet, Vinay Kakkar, Praveen Shetty, Sivakumar Gnanasundram, Shankar Nagaraj, Murali Chandran, Santha Kumar, Shahid Ali Khan, Abey Rajan, and Senthilnathan Ramasubbu.

We also extend our sincere appreciation to:

Nipun Chawla, Rahul Mehta, Aroshi Handa, Rajat Setia, Shaleen Vashisht, Mike Albano, Ramesh Kandula, Manish Gupta, Tamas Mondal, Kevin Wang, Antoni Przygienda, Wen Lin, Ajay Gaonkar, Selvakumar Sivaraj, Aquin Mathai, Vrishab Sikand, Kishore Tiruveedhula, Thorbjoern Zieger, Zacharias El Banna, Lars Axeland, Elisabeth Rodrigues, Washid Lootfun, Adrien Desportes, Krzysztof Szarkowicz, Riccardo Belli, Zuhair Makawa, Krishnamachari KR, Ridha Hamidi, Grzegorz Koscian, Piotr Czechowicz, Fathi Ben Zaied, Vikram Nagarajan, Jonathan Scherman, Rene Triana, Danny He, Chris Hackett, Muzamil Khan, Jeff Haas, Andy McCarthy, Pawel Rabiej, Pawel Kocimowski, Ahmed Bilal Khan, Slawomir Karas, Soumyodeep Joarder, and Jeffrey Zhang.

This page intentionally left blank

About the Authors

Mahesh Subramaniam is a proven leader in AI data centers and next-generation networking technologies. He played a key role in defining the advanced software roadmap for AI fabrics, which are now deployed in production networks across various AI data centers worldwide. As the Senior Director of Product Management for AI Data Centers at HPE Juniper Networks, he leads cutting-edge innovations in AI infrastructure and cloud-scale solutions, optimized for both scale-up and scale-out architectures. Mahesh is also an inventor with several technology patents and a recognized speaker at global forums, including the UEC Summit, OCP, and Tokyo MPLS forum. His work has earned him accolades, including the CEO Excellence Award, the Record High Business Award, and the Star Award for the Cloud DC Reference Architecture. With a remarkable history in the networking industry, Mahesh has a strong track record of leading products and managing technical and business strategies across cross-functional teams.

Michal Styszynski is a Product Management Director in the Data Center Networks Business Unit (DC BU) at HPE Juniper Networking. Michal has been with Juniper Networks for more than 13 years. Before his current role, he was a Technical Marketing Engineer (TME) in the DC BU and a Technical Solution Consultant at Juniper. In these roles, he handled data center projects for large-scale enterprises and federal networks and worked closely with Tier 2 cloud and telco-cloud service providers. Before joining Juniper, he spent around 10 years working at Orange, FT R&D, and TPSA Polpak engineering. Michal graduated from the Electronics & Telecommunications department at Wroclaw University of Science & Technology with a master's degree in engineering. He also holds an MBA from Paris Sorbonne Business School and is a JNCIE-DC#523, as well as PEC, PLC, and PMC certified from the Product School in San Francisco.

Himanshu Tambakuwala is a highly accomplished networking expert and certified technical architect whose experience spans the entire product lifecycle—from hands-on engineering to product strategy. He is a JNCIE holder in Data Center and Service Provider technologies and an inventor with four granted technology patents and two additional patents currently filed. As a Product Manager at Juniper Networks, Himanshu was instrumental in defining the feature roadmap for network fabrics that power cutting-edge AI/ML data centers.

What You'll Find Inside

Chapter 1: Wonders in the Workload

Start your journey by examining the complete lifecycle of AI/ML workloads. This chapter explains how raw data is collected, labeled, and preprocessed before being routed into distributed training pipelines. You will understand the fundamentals of forward and backward propagation, gradient descent, and iterative optimization—and see how these processes scale across thousands of GPUs through data, pipeline, and tensor parallelism. The chapter introduces job completion time (JCT) and tail latency as key metrics and discusses how RDMA (in RoCEv2) facilitates the low-latency, high-throughput transfers needed for modern AI. The technical groundwork clarifies why lossless, high-radix, dynamically balanced fabrics are crucial.

Chapter 2: "The Common-Man View" of AI Data Center Fabrics

This chapter examines the different requirements for AI training and inference data centers. It provides a technical comparison of InfiniBand and Ethernet fabrics, emphasizing their trade-offs in latency, scalability, cost, and ecosystem support. The section discusses AI-specific traffic patterns, such as low-entropy flows, elephant bursts, and synchronization-induced congestion, and explains how advanced load-balancing techniques (static, dynamic, and global) and congestion management methods (ECN, PFC, and DCQCN) are used to maintain fabric efficiency and resilience.

Chapter 3: Network Design Considerations

In this chapter, the book becomes a practical blueprint for building scalable AI clusters. It introduces rail-optimized design and rail-unified design (aka non-rail-optimized design), explaining how each approach affects latency, scalability, and fault tolerance. You'll get hands-on guidance for rack layouts (ToR, MoR, and EoR), cable management, and power/cooling planning. The text then expands to advanced topologies—Clos, Dragonfly, and Torus—showing how to scale from a few racks to tens of thousands of GPUs and how to adapt these principles for inference-centric data centers.

Chapter 4: Optics and Cable Management

Dive into the physical layer with a technical tour of optics and cabling. This chapter covers the evolution from 10 Gbps to 400 Gbps/800 Gbps/1.6 Tbps, the role of DSPs, advanced modulation (PAM4 and QAM), and FEC. It compares transceiver types (QSFP, OSFP, and CFP), connectors (MTP/MPO and LC), and cable options (DAC, AEC, AOC, MMF, SMF, and DWDM). The chapter also explores the latest in pluggable, co-packaged, and linear-drive optics, emphasizing the importance of power and thermal management in high-density environments. We have not addressed the upcoming CPO and LRO/LPO optics technologies in this book as they are subject to ongoing architectural development and discussions with multiple vendors regarding power-efficient rack solutions.

Chapter 5: Thermal and Power Efficiency Considerations

AI clusters push power and cooling to their limits. This chapter addresses the engineering required to keep them running. You'll learn about airflow management (front-to-back, back-to-front, and bidirectional), advanced liquid cooling (immersion, cold plate, rear-door heat exchangers, and spray cooling), and the impact of high-density servers and optics on rack design. Real-world examples from hyperscalers illustrate how next-generation clusters are redefining what's possible in data center thermal management.

Chapter 6: Efficient Load Balancing

This chapter is the technical heart of the networking discussion. It starts with the limitations of traditional ECMP and hash-based load balancing and then introduces a suite of modern technologies: static load balancing, dynamic load balancing (DLB), global load balancing (GLB), flowlet-based rebalancing, per-packet spraying, and selective spraying for RDMA. You'll see how each method works, where it excels, and how to tune it for AI/ML traffic patterns, ensuring optimal utilization and minimal congestion.

Chapter 7: RoCEv2 Transport and Congestion Management

This chapter provides an in-depth technical overview of RoCEv2, the primary transport protocol for AI clusters. The chapter details every potential congestion point and explains the mechanisms—including ECN, PFC, DCQCN, SFC, and CSIG—that can be used to manage them. You'll learn how to tune these controls for different environments, how to prevent PFC storms, and how to use emerging techniques for even more precise congestion management.

Chapter 8: IP Routing for AI/ML Fabrics

Routing is the silent backbone of AI performance. This chapter explores BGP, OSPF, IS-IS, and RIFT in the context of AI data centers, detailing advanced features such as BGP unnumbered, add-path, bandwidth communities, and deterministic path forwarding. You'll learn how to design for multi-tenancy, overlay networks, and server-level routing, as well as how to integrate telemetry and controllers for adaptive, performance-aware routing.

Chapter 9: Storage Network Design and Technologies

Training is fundamentally an I/O challenge. This chapter reviews storage technologies—block, file, and object storage; NVMe-oF; parallel file systems; and GPUDirect Storage—and explains the protocols, state machines, and design patterns that enable high-performance, scalable, and resilient storage networks. The chapter provides practical advice on integrating on-premises and cloud storage, as well as on designing for both hot and cold data paths.

Chapter 10: AI Network Performance KPIs

This chapter outlines the key KPIs—throughput, latency, accuracy, power, efficiency, and scalability—and presents the MLCommons/MLPerf benchmarking suite. You'll learn how to create and analyze benchmarks for both training and inference, ensuring that your performance claims are meaningful, reproducible, and actionable.

Chapter 11: Monitoring and Telemetry

Operations is where design meets reality. This chapter covers the tools and techniques for monitoring and telemetry, from SNMP and syslog to streaming telemetry and in-band flow analyzers. You'll see how to combine real-time and historical data, leverage AI-driven analytics, and automate corrective actions to keep your fabric healthy and performant.

Chapter 12: Ultra Ethernet Consortium (UEC)

This chapter introduces the Ultra Ethernet Consortium and its new protocol stack, which is designed for million-node AI clusters. You'll learn about technical innovations in transport (UET), congestion management (NSCC, RCCC, and CBFC), and packet delivery (ROD, RUD, RUDI, and UUD) and see how UEC compares to InfiniBand and RoCEv2. The chapter offers practical guidance for transitioning to UEC-ready fabrics and highlights the challenges and opportunities ahead.

Chapter 13: Scale-Up Systems

The emergence of scale-up systems represents a significant development in AI data center infrastructure. This chapter details advanced integrated high-performance computing platforms for AI data centers, referred to as super-accelerators. These platforms employ more than eight XPUs (accelerators) within a single system rack, interconnected via next-generation technologies such as UALink, which is a viable alternative to Nvidia NVLink. The chapter presents an overview of encapsulation formats and their key characteristics. It also discusses the placement of these new scale-up systems within the data center network and their integration with scale-out backend solutions, including ESUN (Ethernet for Scale-Up Networking), as proposed at a high level by the Open Compute Project Foundation (OCP).

Credits

Cover: BAIVECTOR/Shutterstock

Figures 4.2.1 Dell'Oro Group

This page intentionally left blank

4

Optics and Cable Management

Scaling Optics for AI Clusters

In Chapter 3, "Network Design Considerations," we discussed design concepts involved in building an AI/ML data center. Optics and cables are important components of any data center, so requirements of an AI/ML data center.

With AI/ML clusters, the ports on a server currently support 200 Gbps and 400 Gbps throughput and are moving toward 800 Gbps and 1.6 Tbps throughput. Nvidia is the top vendor for the GPU server supplies for AI/ML clusters. It has multiple generations of GPUs—Volta, Ampere, Hopper, and Blackwell. Nvidia currently has A100 (Ampere) and H100 (Hopper) on the market and is moving toward H200 and further versions of GPU based on Blackwell.

From the GPU, Nvidia uses NVLink to connect to the NVSwitch, which is used within the server to communicate between the GPUs that are internal to the server. Each NVLink is 300 Gbps for the Volta generation, 600 Gbps for the Ampere generation, and 900 Gbps for the Hopper generation, moving toward 1800 Gbps for the Blackwell generation. Other vendors have similar solutions; for example, AMD has Infinity Fabric, and Intel has CXL (Compute Express Link), UCIe (Universal Chiplet Interconnect Express), and PCIe (Peripheral Component Interconnect Express) switches.

Figure 4-1 illustrates the system topology of a server with 8 GPUs, where an internal switch helps with communication across the GPUs in the server.

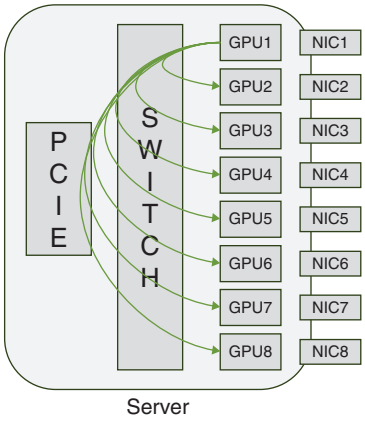

Figure 4-1
System topology

In addition to the NVSwitch for internal communication, there are different NICs that are used for external connectivity between the GPUs across multiple servers. In either case of NVIDIA or AMD-based GPU servers for the scale-out DC use-case, each GPU is connected to a dedicated NIC card, and then each NIC card from the same server connects to different top-of-rack Ethernet or InfiniBand switches. It means for local intra-server communication, the NV switch will be used, and for any server-to-server communication, the external switch is used. GPU-to-NIC card connections are 400 Gbps or 800 Gbps Ethernet or IB. This means a top-of-rack switch is also typically a high-port-density 400Gbps/800Gbps and 1.6Tbps switch. From the AI DC server perspective, besides the GPU-connected NIC cards, there are also storage NIC cards (NVMe-o-F, for example) and out-of-band connections. They are all interconnected via PCIe generation 5 and newer to offer even higher local server interconnects to memory blocks or the CPU.

Figure 4-2 shows a chart from a Dell'Oro report on market adoption of optics for AI clusters from 2020 to 2027. This chart indicates that optics adoption is going to move toward 1.6 Tbps. With the requirement of high-bandwidth optics, there is also a need for high-radix switches that can support a large number of ports per rack unit (RU). In addition, optics need to be power-efficient to reduce the power and thermal budget of a rack. These needs are driving the enhancement of small-form-factor optics, modulation, connectors, and cables.

Figure 4-2.1 shows a 2025 study from Dell'Oro showcasing the adoption of 800 Gbps in 2025 and growing demand for 3.2 Tbps optics by 2029.

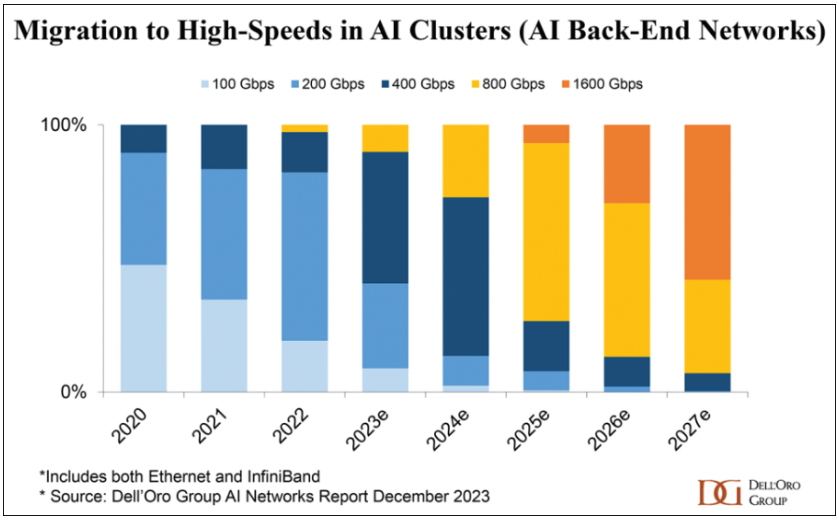

Figure 4-2
Dell'Oro chart on optics evolution

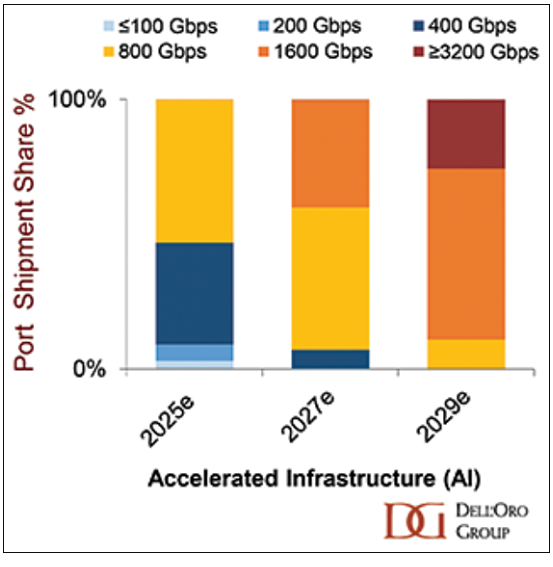

Figure 4-2.1
Dell'Oro chart report on optics evolution 2025

The throughput requirement from server to leaf is expanding to 200 Gbps/400 Gbps/800 Gbps/1.6 Tbps, which is leading vendors to focus on optics to support high-bandwidth connections.

Challenges in Optical Innovation

The optics industry has rapidly advanced from 10 Gbps to 100 Gbps and now to 800 Gbps and beyond. This growth is outpacing traditional models, driven by the need for more bandwidth and faster innovation. Although optical technology doesn't directly follow Moore's Law, it is propelled by data center demands, powerful processors, and expanding data volumes.

These are some of the challenges related to optics technology:

- **Signal quality:** Maintaining signal integrity gets harder as data rates rise. Higher speeds mean more attenuation, dispersion, and noise from crosstalk, all of which can harm signal quality. To tackle these challenges, advanced modulation schemes and error correction are required.
- **Signal conditioning:** At elevated speeds, a signal may experience more severe impairments. To preserve signal integrity, sophisticated methods like digital signal processing and equalization are required.
- **Power consumption and cooling requirements:** Higher-speed optics usually require more power, which can create significant challenges in data centers, where energy efficiency is vital. In addition, increased power usage results in high heat dissipation and requires cooling.
- **Availability and cost:** As AI/ML data centers are demanding higher-speed optics, different organizations are working toward coming up with standards for high-speed optics. Optics technology is lagging, and challenges include availability and cost of high-speed optics units.

Packet Flow

Whenever a network device receives data, it is in the form of either electrical signals or optical signals. A signal goes through multiple stages before it reaches the packet-forwarding engine. Similarly, a packet goes through multiple stages before it is transmitted out. Figure 4-3 illustrates the major components related to the optics in a switch.

In the following sections, we'll discuss packet flow for the newer 400 Gbps and onward optics.

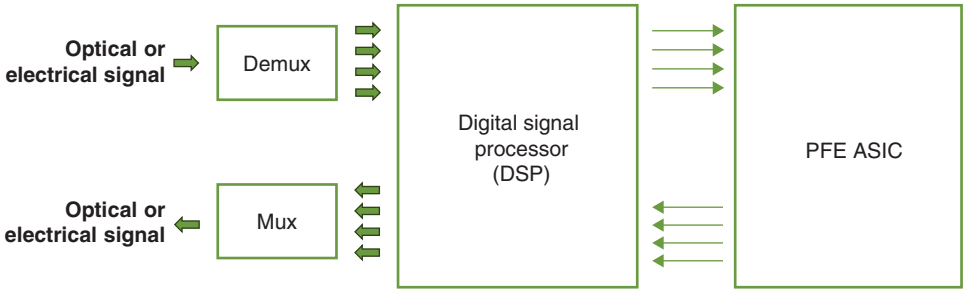

Figure 4-3
Packet flow from optics toward the PFE ASIC

Demultiplexers and Multiplexers

As shown in Figure 4-4, when a signal—either in electrical form when using copper cables or optical form when using optical fiber—reaches the pluggable optics, it goes through a demultiplexer (demux) function that splits the signals.

Figure 4-4
Demultiplexer

The PFE ASIC has multiple SerDes (serializer/deserializers), which form a high-speed interface used to convert data streams received between serial and parallel forms. As the PFE ASIC processes packets serially, the SerDes performs the job of serializing the data received from multiple SerDes links in parallel. If the PFE ASIC supports ~50 Gbps SerDes, a demux could perform the demultiplexing to 8 × 50 Gbps from 400 Gbps optics. If the PFE ASIC supports ~100 Gbps SerDes, a demux could perform the demultiplexing to 4 × 100 Gbps from 400 Gbps optics or 8 × 100 Gbps from 800 Gbps optics.

After the PFE ASIC processes the data, the SerDes again parallelizes the data into separate streams. Either single or multiple SerDes can be mapped to a single optics unit to achieve the required rate. This process is used for creating N:1 conversion of signals, and it is achieved with the help of multiplexing, as illustrated in Figure 4-5.

Figure 4-5
Multiplexer

With AI/ML requirements of 400 Gbps, 800 Gbps, and beyond, the SerDes links are now at 200 Gbps and moving toward higher speeds. Further improvements in the mux and demux are required to be able to achieve higher speeds.

Digital Signal Processors (DSPs)

The signals from the demultiplexer are passed to the digital signal processor (DSP), which carries out several functions. Let's investigate each of them in detail.

Modulation and Demodulation

DSPs are responsible for dealing with sophisticated modulation methods in high-speed optical communications, as illustrated in Figure 4-6. They encode digital data onto the optical carrier wave by converting it into diverse amplitude and phase states. On the receiving side, DSPs decode the optical signal by interpreting the detected changes in amplitude and phase to retrieve the original data.

Modulation is the process of converting data or information to electrical or optical signals. Modulation is required from the ASIC to the electrical signals and then later from electrical signals to optical signals. The modulation at each layer may be different. NRZ, also known as PAM-2, is a traditional modulation technique that does not support higher-bandwidth requirements. PAM-4 and above are being used for newer optics to support the higher-bandwidth requirements of AI/ML data centers.

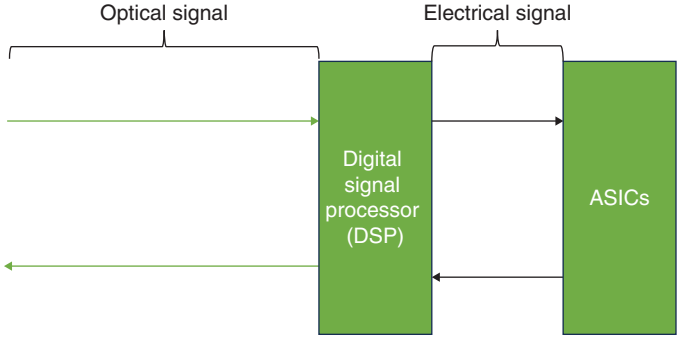

Figure 4-6
Conversion of optical to electrical signals and vice versa

The following modulations are used in different types of network connectivity:

- **NRZ:** Non-Return-to-Zero (NRZ), which used to be a widely used scheme, has two voltage levels to represent 0 and 1. It is commonly used in the 28 GHz range and for some 56 GHz channels.
- **PAM-4:** Pulse Amplitude Modulation with four levels (PAM-4) is a signal-encoding technique that uses four distinct signal levels (that is, voltage levels) to represent four combinations of two-bit logic (00, 01, 10, and 11). PAM-4 modules double the bandwidth of a connection, with each signal level representing 2 bits of logic information. PAM-4 is used for some 56 GHz channels and all 112 GHz channels.
- **Higher-order modulation:** For greater speeds, modulation schemes such as PAM-8 or QAM (quadrature amplitude modulation) are considered. These methods boost data rates by using additional signal levels or combining various amplitude and phase states.

- **DWDM (dense wavelength division multiplexing):** Coherent DWDM technology uses amplitude, phase, and polarization of light. It supports much higher bit rates on a single wavelength when DSPs are used. DWDM is used for transmitting multiple data channels over a single fiber, expanding data capacity in high-speed connections between data centers. 400G ZR modules have ushered in a new era of DWDM technology marked by open, standards-based, and pluggable DWDM optics, enabling true IP-over-DWDM. 400G ZR modules are used for connectivity between data centers (up to 80 km).

To achieve higher data transmission rates, the industry is trying out different modulation methods that make it possible to transmit more data at a time. We can compare it to adding more seats to carry a larger number of passengers on a flight, where the downside is that privacy is reduced. Similarly, transmitting more data at a time results in more noise to the signals and creates a need for more sophisticated mechanisms to deal with it.

Error Detection and Correction

DSPs are responsible for error detection and handling retransmission or correction of received packets to maintain data accuracy. They use forward error correction (FEC) algorithms to rectify errors that occur during transmission. Methods like low-density parity-check (LDPC) codes and Bose-Chaudhuri-Hocquenghem (BCH) codes are often used to enhance data integrity and reliability. At 400 Gbps and above, FEC is needed for reliability, although it introduces latency.

Clock Data Recovery

DSPs play a critical role in synchronizing the transmitter and receiver to maintain data integrity and minimize errors. A DSP extracts clock signals from the incoming data, which is vital for precise data sampling and decoding. This function becomes especially important at high data rates, where accurate timing is essential.

Equalization

DSPs use equalization to improve the signal-to-noise ratio (SNR). Equalization is a signal processing technique that restores the shape of a signal waveform in optics. Equalization algorithms include feed-forward equalization (FFE) and decision-feedback equalization (DFE).

Transmission Modes

Beyond the DSP and mux/demux, we move more toward optics. Two decades ago, most network gear, including switches and routers, relied on copper cables for data transmission. But these cables are limited in terms of how fast data can move. As data speeds picked up, these cables couldn't keep up, as they only reached about 3 meters for 100 Gbps Ethernet links. They were good for connections in the same rack, though, such as connecting servers to the top of rack (TOR) switch.

Today, we have high-performance switches, routers, and SmartNICs that can handle speeds of 200 Gbps, 400 Gbps, and even 800 Gbps. To handle these speeds, fiber-optic cables have taken over from copper ones. Instead of using electrical signals, optical fiber uses light to send data.

The advantage in fiber-optic cables is that they can go much farther—up to 80 km or even 120 km—without losing signal strength. They're also much more reliable than copper because they use total internal reflection to carry light and therefore aren't affected by electromagnetic interference. In addition, fiber-optic cables handle changes in temperature and pressure well.

The following sections detail the different cable options available for connectivity in data center fabrics.

Multi-Mode Fiber (MMF)

MMF is optical fiber that is designed for the transmission of multiple rays of light at a time (see Figure 4-7). Normally, the core diameter of MMF is 50 µm (micrometers) and 62.5 µm. It propagates 850 nm (nanometers) and 1300 nm wavelengths from low-cost light sources like LEDs or VCSELs and has more attenuation. MMF is used for short-to medium-range connections, as well as for connecting devices within the same rack or in nearby racks.

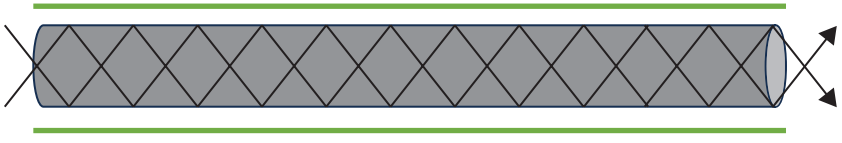

Figure 4-7
Multi-mode fiber

There are five different grades of MMF:

- **OM1:** OM1 supports 1 Gbps speeds up to a distance of 300 m and 10 Gbps speeds up to 33 m.
- **OM2:** OM2 supports 1 Gbps speeds up to a distance of 550 m and 10 Gbps speeds up to 82 m.
- **OM3:** OM3 supports 40 Gbps speeds up to a distance of 240 m and 100 Gbps to 400 Gbps speeds up to a distance of 100 m.
- **OM4:** OM4 supports speeds of 100 Gbps to 400 Gbps up to a distance of 150 m.
- **OM5:** OM5 supports speeds of 100 Gbps to 400 Gbps up to a distance of 150 m. It can accomplish 400 Gbps with just four transmit and receive cables and wavelength division multiplexing (WDM) technology. OM5 uses the wavelength 953 nm along with the wavelength of 850 (OM4).

OM1 has a glass core diameter of 62.5 µm. The other grades of MMF have a glass core diameter of 50 µm.

Single-Mode Fiber (SMF)

SMF is an optical fiber that is designed for the transmission of a single ray of light at a time, eliminating distortion from overlapping light pulses (see Figure 4-8). Normally, SMF has a core diameter of 8 µm to 10 µm, which can propagate higher wavelengths of 1310 nm and 1550 nm from the laser. It is an expensive type of cable that is used for long-distance data transmission between different buildings or data centers.

Figure 4-8
Single-Mode Fiber

Dense Wavelength Division Multiplexing (DWDM)

DWDM enables dense wavelength multiplexing by multiplexing multiple electrical signals to a single optical lane, as illustrated in Figure 4-9. DWDM technology helps in transmitting data over long distances, as described earlier in this chapter. This is specially required in case of data center interconnects.

Figure 4-9
DWDM

AI Server Connectivity Options

On the switch side, the connection could be either single-mode or multi-mode fiber and optics, depending on the distance. Since the server ports an OSFP transceiver with 800 Gbps, the switch side must be a 800 Gbps port or two 400 Gbps ports with the use of breakout cables. Figure 4-10 illustrates both of these connectivity options. For rail-optimized design (ROD), another option is able to connect 8 GPUs to 8 different switches.

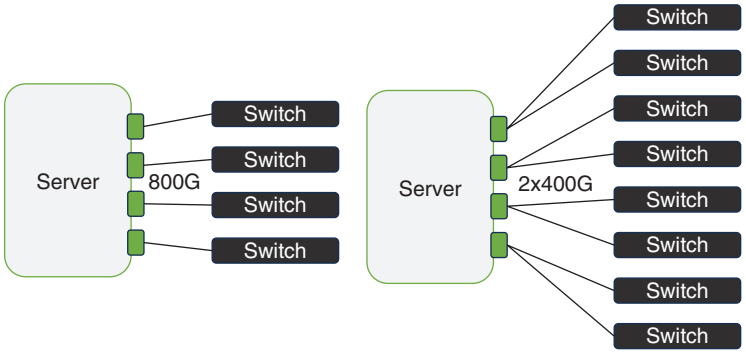

Figure 4-10
Connection options with the Nvidia H100

Figure 4-11 illustrates the connectivity option with the A100 server. There are 8 ports on the server that can connect to 8 different switches in ROD.

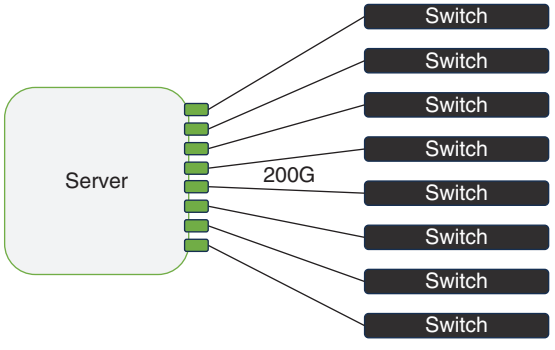

Figure 4-11
Connection option with the Nvidia A100

Transceiver Types

The choice of optics is based on the connectivity distance between two nodes. In general, optics meant for smaller distances are cheaper than those meant for longer distances. Each type of optical fiber has a suffix to denote the reach and optical lanes. For example, in 400G-SR8, SR stands for *short reach*, which is up to 100 m, and 8 denotes the number of optical lanes; 400G-SR8 can have 8 optical lanes of 53 Gbps each, which are multiplexed to support 400 Gbps bandwidth.

Table 4-1 lists the transceiver types and the reach and mode of optical fiber with which it can be used.

Index

Numerics

A

B

C

D

E

F

G

H

I

J-K

L

M

N

O

P

Q

R

S

T

U

V

W

X-Y-Z