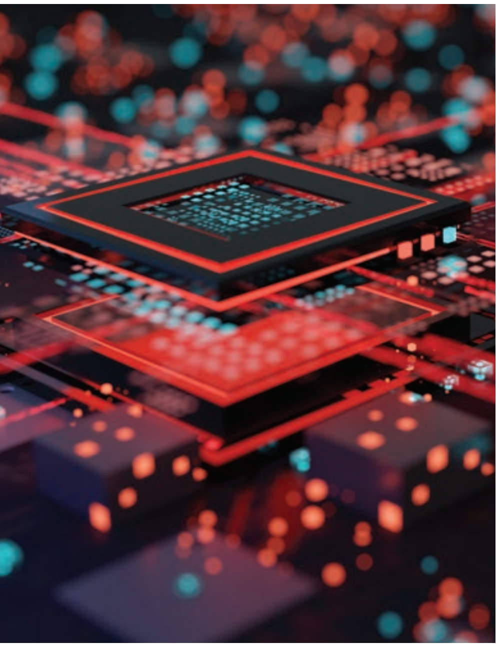
Generative AI Solutions Guide

EXCEEDING THE REQUIREMENTS OF EMERGING AI NETWORKS

www.siemon.com/ai

Connecting the World to a Higher Standard

The Generative AI Revolution is Here. Are you Ready? We Are.



For years, Artificial Intelligence (AI) and Machine Learning (ML) have reshaped industries, empowered lives, and tackled complex global issues. These transformative forces, formerly identified as HPC (high-performance computing), have fueled digital transformations across organizations of all sizes, boosting productivity, efficiency, and problem-solving prowess. Now, the emergence of highly innovative machine generative AI (GenAI) models, powered by deep learning and neural networks, is further disrupting the game. By generating original content and tackling intricate challenges, GenAI is poised to revolutionize not just how organizations operate, but the very fabric of innovation itself.

Increased use of these data- and compute-intensive ML and GenAI applications is placing unprecedented demands on data center infrastructure, requiring reliable high-bandwidth, low-latency data transmission, significantly higher cabling and rack power densities, and advanced cooling methods. As data centers gear up for AI, users need innovative, robust network infrastructure solutions that will help them to easily design, deploy, and scale back-end, front-end, and storage network fabrics for complex high-performance computing (HPC) AI environments. Thankfully, Siemon has everything you need to embark on the Generative AI revolution – and it's all backed by experience and expert Data Center Services.

> **The AI boom across various industries is fueling 35%+ market growth, with GenAI alone expected to hit $1.3 trillion by 2032.**
>
> (Precedent Research, 2023)

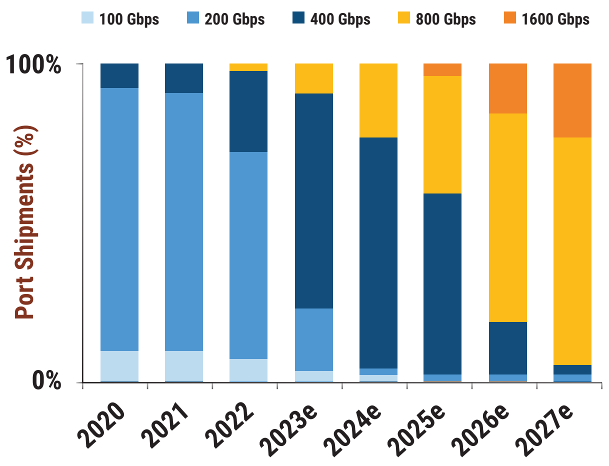
Advanced AI Calls for a Data Center Design Rethink

Accelerated GenAI and ML models consist of training (learning new capabilities) and inference (applying the capabilities to new data). These deep-learning and neural networks mimic the human brain's architecture and function to learn and generate new, original content based on analyzing patterns, nuances, and characteristics across massive, complex data sets. Large language models (LLM), such as ChatGPT and Google Bard, are examples of these GenAI models trained on vast amounts of data to understand and generate plausible language responses.

General-purpose CPUs that perform control and input/output operations in sequence cannot effectively pull vast amounts of data in parallel from various sources and process it quickly enough. Therefore, accelerated ML and GenAI models rely on graphical processing units (GPUs) that use accelerated parallel processing to execute thousands of high-throughput computations simultaneously. The compute capability of a single GPU-based server can match the performance of dozens of traditional CPU-based servers!

Unique ML and GenAI Characteristics

Migration to High-Speed in AI Networks (Back-End)

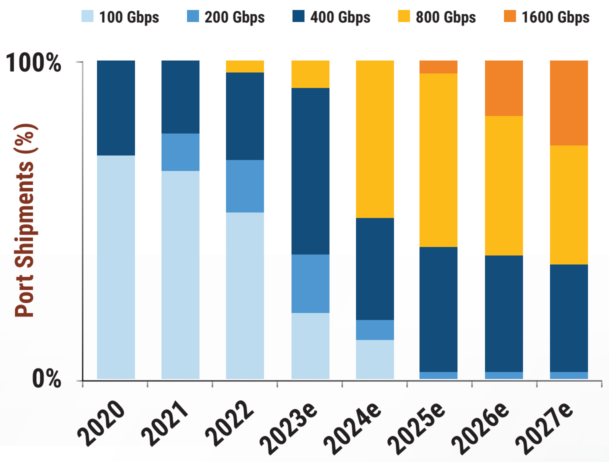


- Includes both Ethernet and InfiniBand
- Source: Dell'Oro Group AI Networks Report

Preliminary Forecast (2023-2027)

- InfiniBand and Ethernet will coexist.
- Nearly all ports will be at 800G speeds and above by 2027.
- Triple-Digit CAGR for network bandwidth.

Migration to High-Speed in AI Networks (Front-End)



- Source: Dell'Oro Group AI Networks Report

Preliminary Forecast (2023-2027)

- All ports are Ethernet.
- Nearly 2/3 of all ports will be at 800G speeds and above by 2027.

To pull information from massive data sets, ML and GenAI models run within back-end HPC multi-node clusters with hundreds or thousands of interconnected GPUs that come with unique requirements:

- **Very High Bandwidth –** 100G, 200G, 400G, and even 800G speeds at the server level, with switch-to-switch links rapidly migrating to 800G and 1.6T speeds.
- **Extremely Low Latency –** Real-time (< 20 milliseconds) east-west data transmission between nodes.
- **Dramatically Increased Power Consumption** – GPU-based servers require up to 10x more power, resulting in rack power densities of 30-100kW or more.
- **Advanced Cooling –** Data centers are evaluating more efficient cooling methods, such as direct-to-chip liquid cooling and liquid immersion cooling to handle the high heat generation.
- **InfiniBand and Ethernet Protocols –** InfiniBand's high-throughput and lower latency performance dominate back-end inter-GPU connectivity, while Ethernet's compatibility, security, and management capabilities are ideal for front-end interfaces. Ongoing Ethernet advancements will allow both to exist in the back-end.

  Front-end refers to the user-facing elements of a website or app, like design and interactivity, while back-end encompasses the server-side logic and data architecture that powers it.

- **High-Density, High-Performance Cabling –** More high bandwidth connectivity is required for high-speed connections between nodes and for storage, management, and switching.

Due to the high power & cooling requirements, supporting AI often requires distributing GPUs across cabinets and shifting from Top of Rack (ToR) configurations to End of Row or Middle of Row (EoR/MoR). The higher power requirements may also necessitate advanced, efficient liquid cooling methods.

We Are AI-Ready

Whether for an AI cloud service provider with the power and cooling to support higher rack power densities and ToR configurations within large HPC clusters consisting of thousands of interconnected GPUs, or a large enterprise looking to build their own business-specific AI infrastructure within an existing on-site or colocation data center. Siemon has everything users need to support the design, delivery, and day-to-day operation of AI networks including:

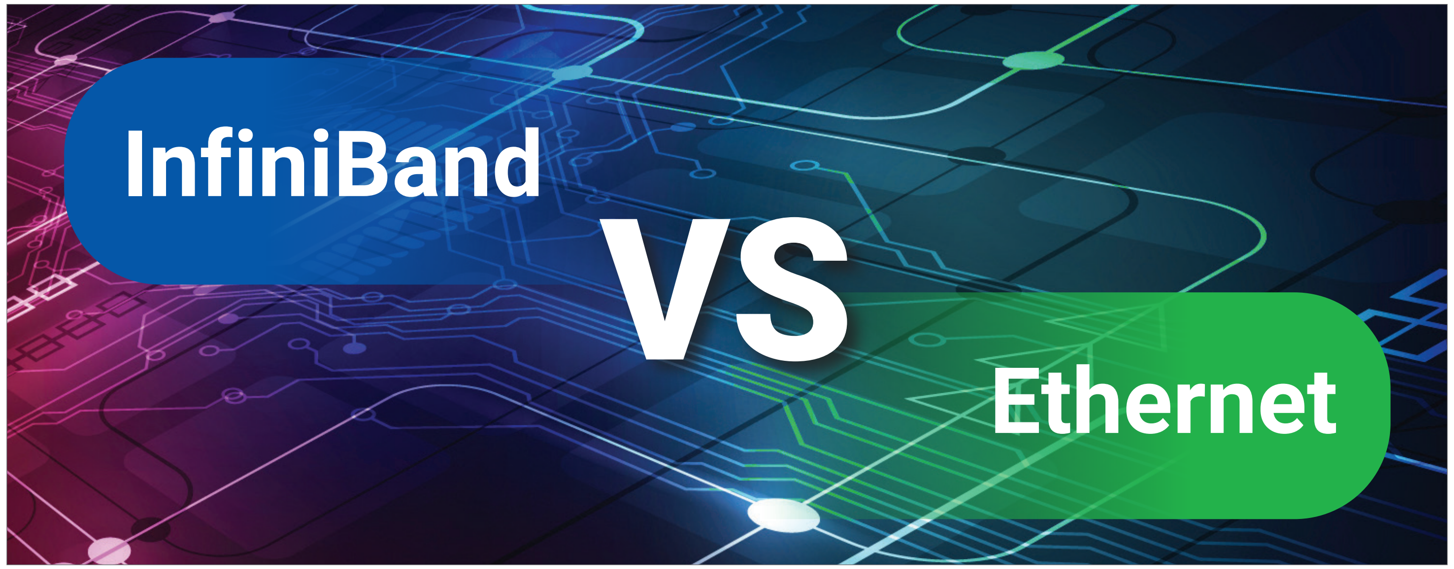
High-density, end-to-end LightVerse singlemode and multimode MTP, MTP Pro® or MPO fiber systems that deliver high-performance Ultra-Low-Loss (ULL) transmission to 800G and beyond for back-end, front-end, and storage fabrics.

A comprehensive line of direct attach cables (DACs) and active optical cables (AOCs) for point-to-point high-speed, low-latency connections within back-end AI clusters for Ethernet, RoCE and InfiniBand networks.

The ability to support both InfiniBand and Ethernet protocols in ToR and EoR/MoR configurations.

Expertise and experience helping companies design and deploy optimized AI-focused InfiniBand and Ethernet copper and fiber infrastructure for all required networks – compute, storage, and out-of-band and in-band management.

Data center services to provide infrastructure design recommendations based on specific use cases, budget, existing infrastructure, and future needs.



Bringing Together Two Data Center Giants

While Ethernet has long been the de-facto protocol for most networking needs, InfiniBand has been common in HPC networks and is already established as a powerful alternative for back-end GPU interconnects within AI clusters due to its Remote Direct Memory Access (RDMA) technology that can handle multiple high-bandwidth, low-latency parallel connections for massive data transfer. The latency for an InfiniBand switch is about 100 nanoseconds versus 230 for an Ethernet switch. With its improved support for AI, the global InfiniBand market is expected to grow by more than 40% between 2023 and 2028.

In contrast, Ethernet's TCP/IP broad ecosystem support, scalability, ease of use, and security and management features make it ideal for front-end AI networks, including switch-to-switch connections, storage fabrics, and out-of-band and in-band management networks. As switch technology and protocols such as RDMA over converged Ethernet (RoCE) and Ultra Ethernet Transport (UET) advance to meet the requirements of AI workloads, both InfiniBand and Ethernet will compete and reside within back-end AI clusters while frontend AI networks will remain Ethernet. Ultimately, both technologies have their strengths within AI architecture and can support up to 800G speeds with roadmaps to 1.6 Terabit.

	InfiniBand			Ethernet			
Lanes	Form Factor	Lane Speed	Link Speed	Form Factor	Lane Speed	Link Speed	Connectivity Options
2	QSFP	EDR HDR NDR	50G 100G 200G	QSFP	25G 50G 100G	50G 100G 200G	1. Optical Transceivers + Fiber Cabling 2. Direct Attach Cable (DAC) 3. Active Optical Cable (AOC)
4	QSFP OSFP	EDR HDR NDR	100G 200G 400G	QSFP OSFP	25G 50G 100G	100G 200G 400G	
8	OSFP	HDR NDR XDR	400G 800G 1.6T	QSFPDD OSFP	50G 100G 200G	400G 800G 1.6T	

Siemon Data Center Services

Supporting Organizations to Harness the True Potential of their Data Center AI Environments.

Whether you're a service provider entrusted with multiple clients' IT and delivering on SLAs, or an organization investing in AI and High-Performance Computing (HPC) networks to accelerate your business, your ability to thrive and grow is only as good as your data center's underlying network infrastructure.

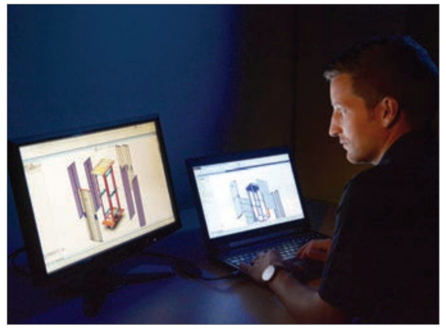
Your data center's cabling infrastructure is at the very core of meeting internal and external customer expectations for top-level AI networks and HPC availability and performance. Your teams also need worry-free uptime, reliability, and scalability assurance so they can focus on what it takes to be successful.

How Can We Support You?

We've focused our data center expertise into a global service network, designed to guide you through the process of selecting and designing the underlying physical infrastructure you need to ensure your data center is AI-ready while offering you the ongoing support you need to respond quickly to changing needs, prevent downtime and maintain peak performance.

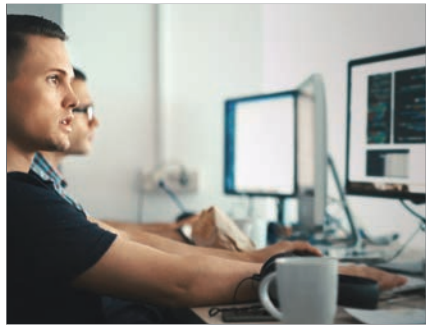
Data Center Cabling Audits

Get a comprehensive snapshot of your current cabling infrastructure strategies, as well as detailed analysis and guidance on improvements that can be made and potential savings that can be achieved.



Data Center Design Services

Are you looking to design, deploy or upgrade a data center space and need expert advice and assistance to help you through the process? Our team of DC Design experts are ready to support you.



Technical Services Group

Our expert technical services teams are available around the world to support our clients throughout their Siemon experience. Our teams are with you every step of the way.

Supported by Industry Leading Partnerships

Over the years we've developed an ecosystem of data center partners who are all specialists in what they do. Siemon prides itself on partnering with global AI and HPC leaders who provide complementary products and services that combine with our best-in-class IT infrastructure solutions to deliver additional value and support to our customers.

ARISTA

CISCO™

The Need for Ultra-Low Loss Connectivity

InfiniBand and Ethernet links in back-end and front-end AI networks leverage the bandwidth capabilities of cost-effective multimode and short-reach singlemode optical transceiver technology. For high-speed structured cabling links, multimode fiber supports up to 50 meter (m) for 200G and 400G, while short-reach singlemode supports much longer distances of 500 and 2000m. For Ethernet deployments, multimode (SR and VR) and short-reach singlemode applications (DR and FR) have stringent insertion loss requirements, with a maximum channel loss of 1.9 dB for multimode, 3 dB for DR singlemode, and 4 dB for FR singlemode.

For structured cabling within AI networks, Ultra-Low Loss (ULL) MPO/MTP connectivity ensures maximum channel distances while ensuring margin to accommodate installation variables and deliver the flexibility to support convenient cross-connects that help improve manageability, scalability, and speed of deployment. When selecting connectivity, loss values can vary from vendor to vendor, with many offering standard loss, low loss, and ULL connectivity. Due to the variability, it's essential to ensure third-party verified maximum insertion loss values that provide a true indication of performance.

Siemon's ULL LightVerse® MTP/MPO multimode and singlemode MTP/MPO cabling systems have been third-party verified to provide considerable insertion loss margin for enhanced performance of AI networks delivering speeds of 100, 200, 400 and 800G.

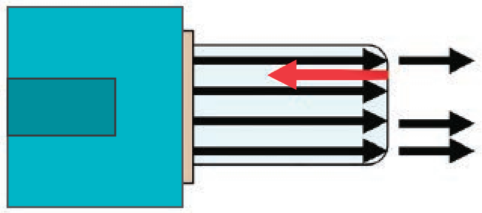
intertek

Total Quality. Assured.

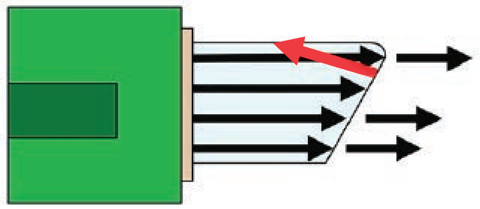
The Need for Multimode and Singlemode APC Connectivity

In addition to stringent insertion loss requirements, high-speed 400 and 800G applications and future 1.6 Terabit applications are more susceptible to reflectance due to a higher signal-to-noise (SNR) ratio. Short-reach DR and FR singlemode applications are especially susceptible, so industry standards have specified reflectance (return loss) values based on the number of mated pairs in the channel.

Poor reflectance performance can adversely impact channel insertion loss and transmission performance. As a result, many cloud data centers are specifying the use of angled physical contact (APC) for multimode, in addition to the traditional singlemode MTP/MPO connectivity for high-speed AI links. Unlike ultra-physical contact (UPC) connectors that feature a rounded fiber end face, APC connectors are polished at an 8-degree angle to reduce the amount of reflected signal. Multimode MTP/MPO/UPC connectors typically have a reflectance value of -20 dB, while multimode MTP/MPO/APC connectors have an improved reflectance value of -35 dB.

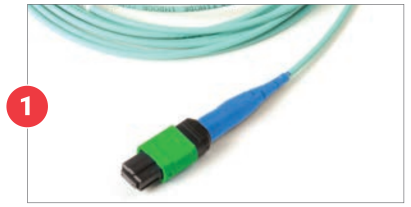


MTP/MPO MM/UPC Connectors
(-20 dB)



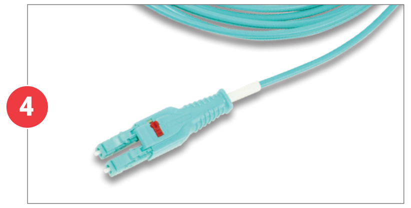
MTP/MPO MM/APC Connectors
(-35 dB)

Connectivity Solutions for AI Applications

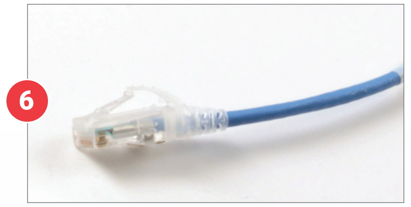

1 MTP/MPO Fiber Equipment Cords

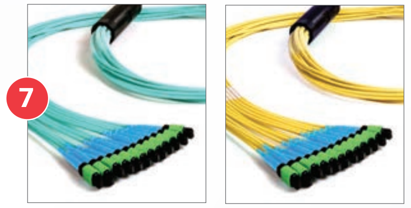
2 MTP/MPO Fiber Jumpers

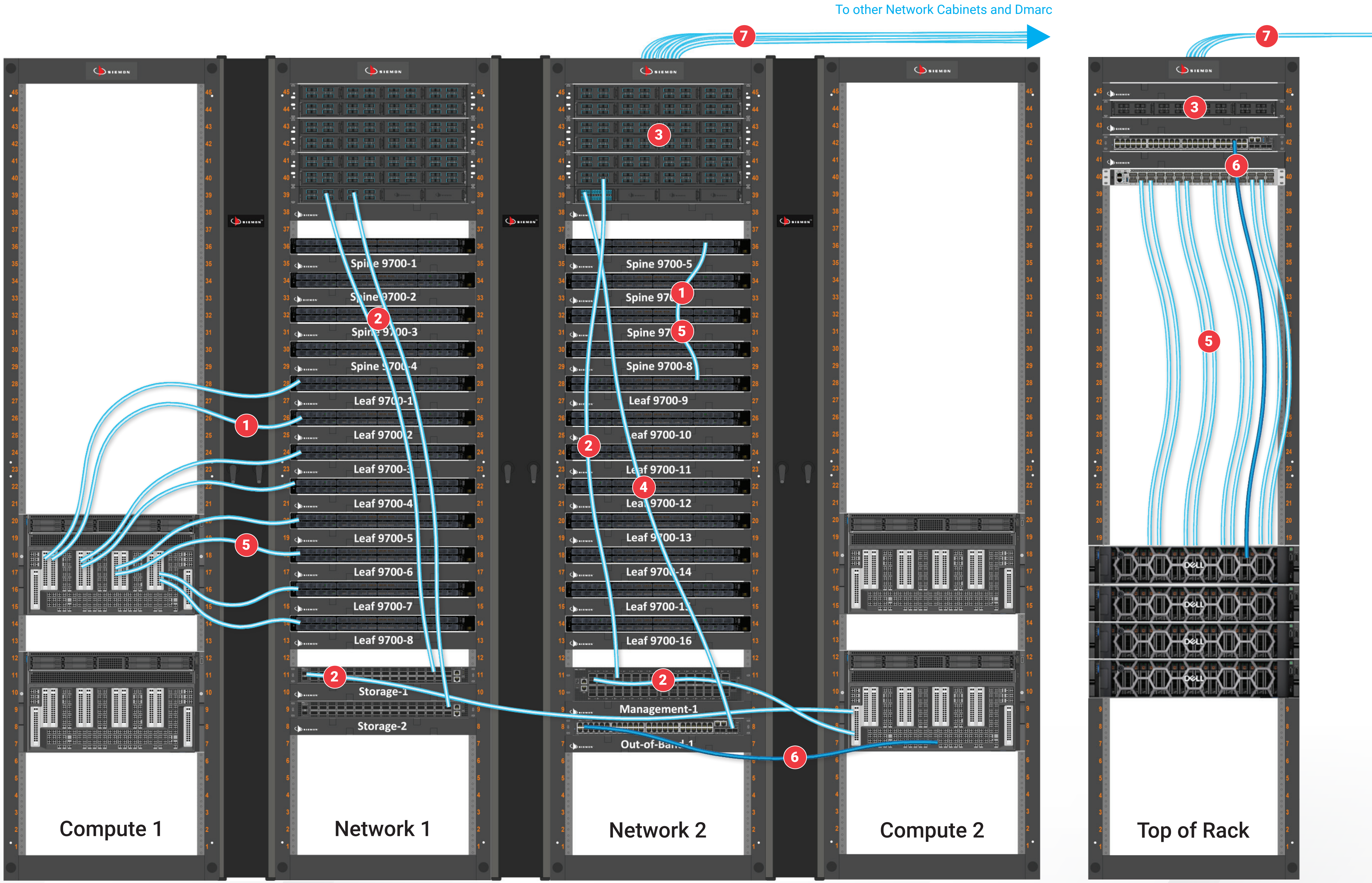
3 LightVerse® Fiber Optic Cabling System


4 LC BladePatch® Fiber Jumpers

5 Direct Attach Cables and Active Optical Cables


6 Category 6A and 6 Modular Cords


7 Plug and Play Fiber Trunks and Assemblies



Example for Training Model

Example for Inference Model

The Siemon Advantage

Established in 1903, Siemon's is the trusted go-to industry leader in the manufacturing and innovation of high-quality, high-performance data center solutions for customers around the world. Siemon Advanced Data Center Solutions are engineered from the ground up to meet the demanding needs of AI and HPC environments, both today and in the future. Our deep understanding of data center requirements has fueled our strategic pivot to address the latest AI revolution. All backed by Siemon's industry-leading quality, performance, and reliability they combine to help you reduce risk, maximize uptime, and successfully deliver new AI applications and services.

- **Siemon has successfully worked with several other companies** to design and deploy HDR 200G and NDR 400G InfiniBand compute systems.
- **We have extensive knowledge of leading hardware vendor's** reference architecture designs and can provide cabling design recommendations for particular use cases.
- **Siemon has unique solutions that meet AI's low latency requirements**, simplify deployment, and provide flexibility after initial installation.
- **Siemon can support all common networks:** Compute, Storage, Out-of-band, and In-band management.
- **Continually advancing product line** to meet evolving data center needs, driven by a culture of continuous improvement, significant investment in R&D, and leading participation in industry standards.
- **Backed by global sales coverage and a comprehensive data center partner ecosystem** with leading providers of complimentary products and services for a total value-added approach.
- **Our broad solutions span InfiniBand and Ethernet** fiber, copper, DACs and AOCs to support an end-to-end implementation.

Hybrid Cabling for GenAI: Beyond Traditional Point-to-Point

While long-distance links in HPC clusters benefit from structured cabling, short, low-latency connections for GPUs often rely on point-to-point solutions like DACs. But AI clusters with spread-out GPUs and racks push lengths beyond DAC limits. AOCs and individual fiber cables offer up to 100m for cabinet-to-cabinet connections, but managing hundreds or thousands in large clusters becomes a hassle. Some of the benefits of using structured cabling are as follows:

Max Flexibility: Connect any GPU to any other with patch panel fiber jumpers, adapting easily to your evolving AI needs.

Dense Switch Connections: Manage high-density leaf-to-spine switch connections seamlessly for rail-optimized designs.

Protect Key Equipment: Avoid touching critical, expensive equipment ports during moves, adds, and changes.

Cost-Effective & Scalable: Simplify Day 2 MAC work and scale to higher speeds without re-cabling, saving time and money.

Easy Troubleshooting: Standardized labeling and documentation at patch panels simplifies troubleshooting and reduces downtime.

Improved Airflow & Space: Reduce cable congestion for better airflow and easier equipment access.

Certified Performance: Industry-compliant system guarantees cabling performance.

Pioneering AI with Leading Associations

Siemon is an active member with AI's Leading Associations to Shape a Responsible Future.

INFINIBAND
TRADE ASSOCIATION

"Siemon's partnership with InfiniBand reinforces our commitment to advancing network infrastructure solutions globally. We recognize the pivotal role InfiniBand plays in meeting the escalating demands of Artificial Intelligence and accelerator cards. Siemon is committed to providing innovative cabling and connectivity solutions that enable the advancement and adoption of this technology."

– Gary Bernstein | Sr. Director of Global Data Center Sales, Siemon

ethernet alliance

"The Ethernet Alliance has long been committed to supporting Ethernet development through industry standards and multivendor interoperability. This falls perfectly in line with Siemon's longstanding participation in industry standards and commitment to delivering standards-based, quality solutions."

– John Siemon | Chief Technology Officer, Siemon

IEEE

"IEEE 802.3 Working Group develops standards for Ethernet networks. Siemon is active in all the key task forces developing standards for 200, 400, 800G & 1.6T that are being used by many AI networks."

– Dave Valentukonis | North America Technical Services Manager, Siemon

Siemon Advanced Data Center Solutions

High-Density Fiber Enclosures

Siemon's high-density LightVerse® Fiber Optic Cabling System includes high-density enclosures including Core, Plus and Pro options as well as MTP/MPO-to-LC modules, MTP/MPO and LC adapter plates to support singlemode and multimode fiber patching, and two different fiber splice options – all designed to support network deployments of 400G and beyond.

go.siemon.com/LightVerseEnclosures

Ultra-High-Density Enclosures

Designed to easily integrate into any standard 19" rack or cabinet, Siemon's LightStack™ Fiber Enclosures are constructed of high-quality steel and are designed for easy installation and quick deployment of plug and play fiber solutions for today's advanced data center and storage area network environments.

LightStack will be available in North America in Spring 2024.

go.siemon.com/LightStackEnclosures

Direct Attach Cables

Ideal for short reach switch-to-server connections Siemon (DAC) support a variety of QSFP28, SFP28, QSFP+, SFP+ form factors, and come in half meter increments from 0.5m to 5m, with breakout options and multiple colors available. Ask about higher speeds coming later this year.

go.siemon.com/DACs

Active Optical Cables

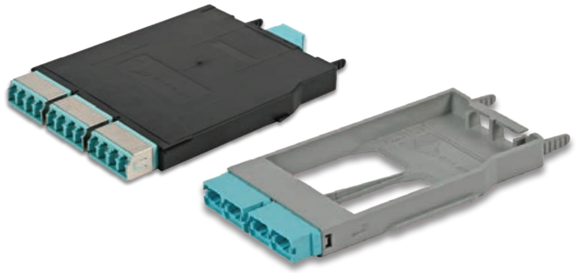
With lengths up to 100 meters, Siemon's multimode fiber AOC cable assemblies are ideal for longer reach, point to point connections in data centers, cabinets and equipment outside of the rack. The energy efficient design requires less power than transceiver assemblies with smaller bundles promote better airflow and lower cooling costs.

go.siemon.com/AOCs

Modules and Adapters

The Siemon family of Plug and Play modules support MTP/MPO to LC/SC fiber transition, allowing for quicker and easier deployment of up to 24 fibers in a single module. The LightVerse adapter plates are designed for simple one-handed installation from the front of the enclosure and can be passed from the rear to the front and back again without leaving the enclosure.

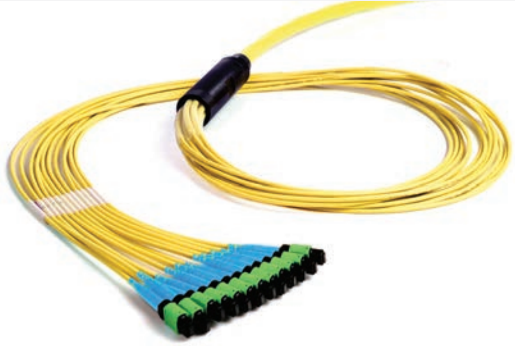
go.siemon.com/LightVerseModAdapt



Ultra-High-Density Modules and Adapters

LightStack MTP/MPO-to-LC Ultra-Low-Loss (ULL) Plug and Play modules deliver a quick and efficient way to deploy high-performance fiber cabling in a low-profile, high-density package. LightStack ULL MTP/MPO pass-through adapters are available in 6-port designs supporting up to 72 fibers per adapter plate and are offered in both aligned and opposed key orientations to accommodate all polarity methods.

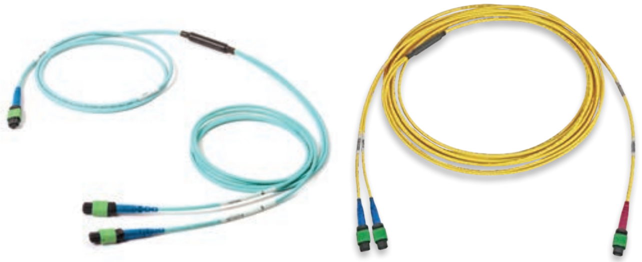
go.siemon.com/LightStackModAdapt



MTP, MTP Pro® or MPO Fiber Trunks

Siemon's wide range of singlemode and multimode plug and play trunks and assemblies include Base-8 or Base-12 MTP to MTP, MTP Pro to MTP Pro or MPO to MPO trunks in a variety of fiber counts, hybrid MTP/MPO-to-LC trunks, MTP/MPO conversion cords, MTP/MPO jumpers and LC or SC duplex and simplex breakout assemblies. Newly released 16-fiber MTP/MPO jumpers and (1) 16-fiber MTP/MPO to (2) 8-fiber MTP/MPO Y conversion cords are also available.

go.siemon.com/DCMTPTrunks



MTP, MTP Pro or MPO Fiber Conversion Cords

Siemon's Conversion cord ensures 100 percent fiber utilization in 40 and 100G applications. Multiple versions of conversion cords are available, including transitioning (2)12-fiber MTP/MPO connectivity from the backbone trunk to (3) 8-fiber MTP/MPO connectors and transitioning (2) 12-fiber MTP/MPO to MTP/MPO trunks from the backbone to (1) 24-fiber MTP/MPO connector to connect to the active equipment.

go.siemon.com/DCConversionCords

Siemon Advanced Data Center Solutions

Copper/Fiber Combo Patch Panels

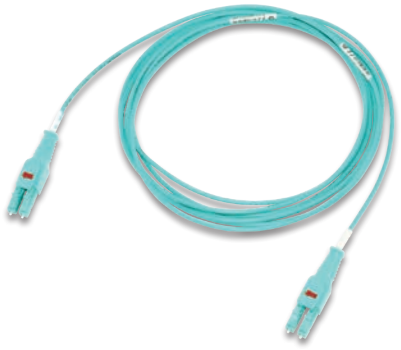
LightVerse® Copper/Fiber Combo patch panels are designed to allow users to easily mix their high-performance fiber optic and copper connectivity within the same 1U of rack space. Available in flat or angled configurations, the LightVerse Copper/Fiber Combo patch panel supports both singlemode and multimode fiber, or shielded category 6A and unshielded category 6A/6 copper connectivity.

go.siemon.com/CopperFiberPanels

Modular High-Density Fiber Patch Panels

Streamline fiber deployment in data centers and intelligent buildings with LightVerse Modular Patch Panels. These 1U, 2U, or 4U panels offer flexible, high-density solutions for up to 96 fibers, supporting seamless patch, transition, & splice terminations.

go.siemon.com/LightVerseHDPanels



High-Density Fiber Jumpers

Siemon's LC BladePatch® singlemode and multimode duplex jumpers offer a unique solution for high-density fiber optic patching environments with a revolutionary push-pull UniClick™ boot design to control the latch, enabling easy access and removal in tight-fitting areas.

go.siemon.com/DCFiberJumpers

MTP/MPO Fiber Jumpers

Siemon's MTP, MTP Pro® or MPO jumpers are used to connect the MTP/MPO trunk backbone to the active equipment. The compact design of the MTP/MPO footprint and Siemon's 2mm diameter RazorCore cable achieves greater connectivity access, reduction in cable pathway congestion and improved airflow around the active equipment. MTP, MTP Pro or MPO jumpers are available in Base-8 and Base-12 versions.

go.siemon.com/DCJumpers

Advanced Copper Cabling Solutions

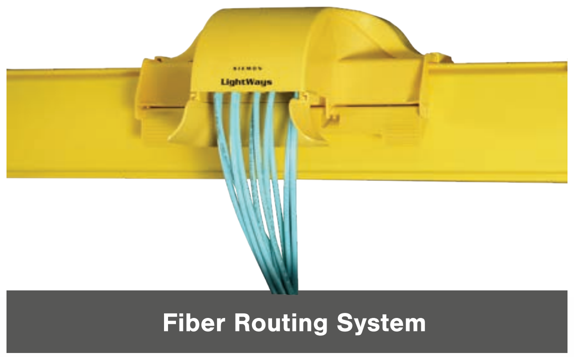
Unleash network excellence with Siemon's Advanced Copper Cabling Systems. Our UltraMAX™ and Z-MAX™ systems deliver end-to-end next-gen performance for Cat 5e, 6, 6A, 7A, and 8.2 options. Explore twisted pair cables, RJ45/TERA® outlets, field-installable plugs, pre-terminated trunks, patch panels, and diverse patch cords. Design, deploy, and connect with confidence.

go.siemon.com/Copper

Copper Patch Cords

SkinnyPatch® Modular Cords deliver superior performance with a reduced cable diameter for improved airflow and increased flexibility in high-density patching areas. The cord's smaller 28 AWG stranded copper construction offers a significantly tighter bend radius for easier cable routing and enhanced cable management. SkinnyPatch is available in category 6A shielded, 6A UTP and category 6 UTP versions.

go.siemon.com/DCSkinnyPatch



Fiber Routing System

Manufactured from halogen free, flame-retardant UL94/V0 plastic and available in four different sizes. LightWays™ is easy to assemble and includes a wide variety of straight duct, elbows, tees, crosses, reducers and innovative outlets ideal for designing a system that meets the precise needs of your data center space.

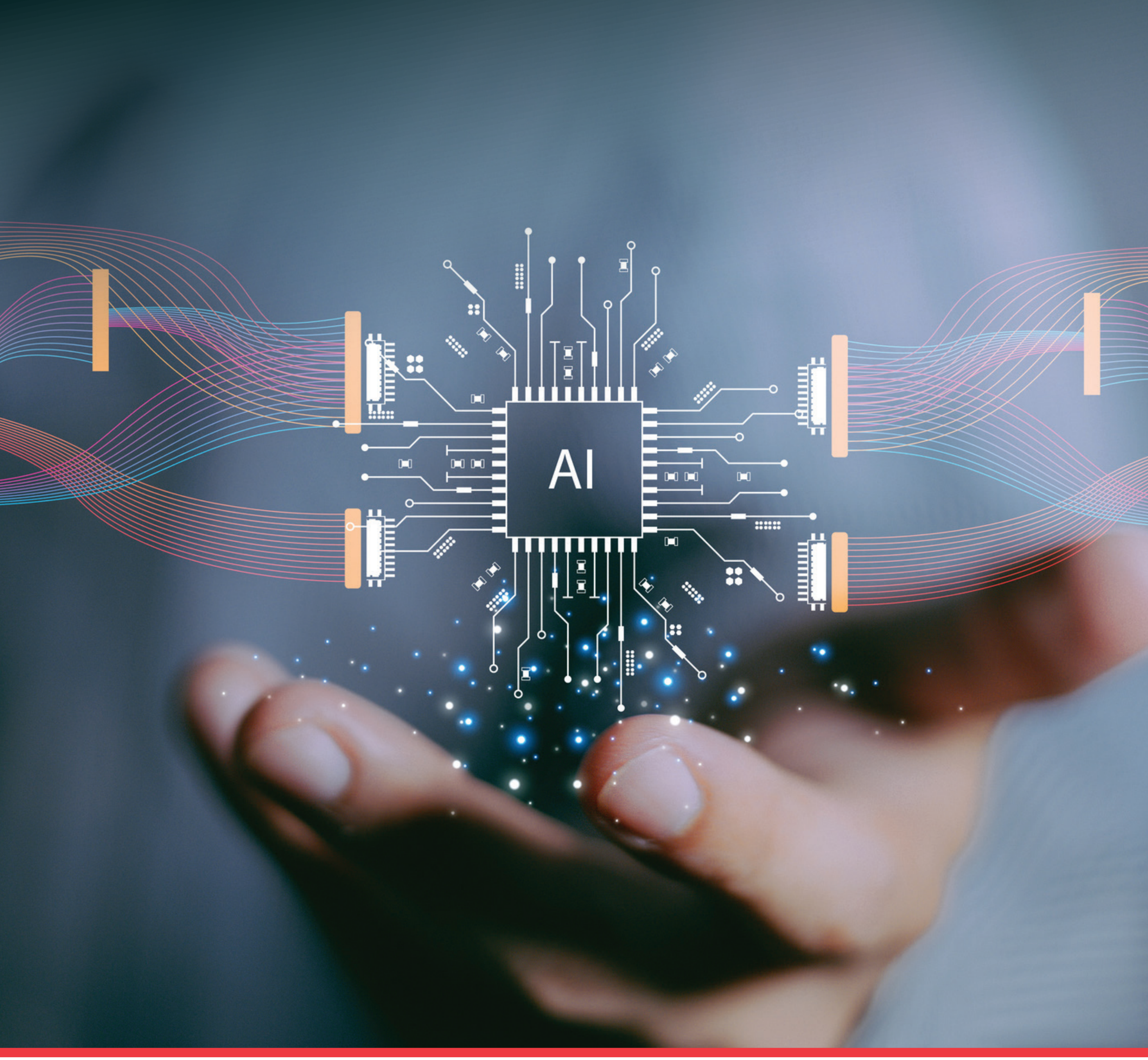
LightWays is available in limited geographies.
Visit www.siemon.com for availability.

go.siemon.com/DCFiberRouting

Racks and Cable Management

Siemon offers a wide range of racks, vertical and horizontal cable management and accessories to meet a variety of data center needs.

go.siemon.com/DCRacks



START YOUR SIEMON AI JOURNEY TODAY!

For more information visit:
www.siemon.com

Find your local Siemon distributor:
go.siemon.com/distributor

24/7 Customer Support:
customer_service@siemon.com

For inquiries specific to wireless/wireline:
go.siemon.com/cellular

SIEMON™