Architecting AI at Scale: From Training Clusters to Inference-Driven Infrastructure

White Paper

Summary

AI infrastructure is evolving beyond a single architectural model. What began as uniform GPU training clusters has expanded into a set of specialized systems aligned to different workload requirements. Training remains important, but inference now accounts for 50–70%[1] of total AI compute demand, introducing new considerations for how data centers are designed and built.

Rather than representing a single workload category, inference spans high-throughput serving, multi-step reasoning, disaggregated pipelines, and emerging agentic systems that maintain persistent context. Each category places different demands on latency, bandwidth, east west traffic, and the underlying optical and Ethernet layers. As a result, modern AI facilities increasingly mix shared infrastructure with workload specific design, and planners can no longer rely on GPU count as the primary design axis.

This paper introduces a six category planning matrix that links workload behavior to system architecture, network characteristics, and fiber requirements. The goal is practical clarity: to help architects, operators, and data center planners align physical infrastructure with the real demands of AI workloads.

The focus is on RoCE-based Ethernet fabrics, now the dominant scale-out model outside vertically integrated GPU systems, and on the optical connectivity supporting these architectures. This paper is the first in a multi-part series moving from high-level architectural analysis to detailed engineering guidance across fiber, network, and physical plant design. Details on the complete series can be found at the end of this white paper.

1. Deloitte
   Mckinsey

Copyright © 2026 AFL. All rights reserved.

Introduction: Why the old model breaks down

Early AI infrastructure followed a simple pattern: build dense GPU training clusters connected across the fastest available fabric and scale according to topology limits. Traffic was predictable, architectures were uniform, and fiber planning focused on reach and port count. That model aligned with an AI landscape dominated by training workloads. However, current AI infrastructure requirements extend beyond those assumptions.

Inference now drives the majority of AI demand, bringing a wider range of behaviors than training. High-throughput serving, multi-step reasoning, disaggregated prefill/decode pipelines, and emerging agentic systems each introduce different demands on latency sensitivity, east-west traffic intensity, and data movement across compute tiers.

Modern AI facilities increasingly combine shared infrastructure in some layers with differentiated design in others. GPU count is no longer the primary planning axis. Instead, architects must understand how each workload category translates into network behavior and fiber requirements.

This paper presents a framework for aligning AI workload categories with system architecture, RoCE-based Ethernet fabrics, and the underlying optical plant. The objective is to provide planners and operators with a clearer basis for matching physical infrastructure to the demands of AI at scale.

Contributors

Dr Alan Keizer
Senior Technology Advisor, AFL

Keith Sullivan
Director of Strategic Innovation, AFL

Ben Atherton
Technical Author, AFL

Paige James
Design Manager, AFL

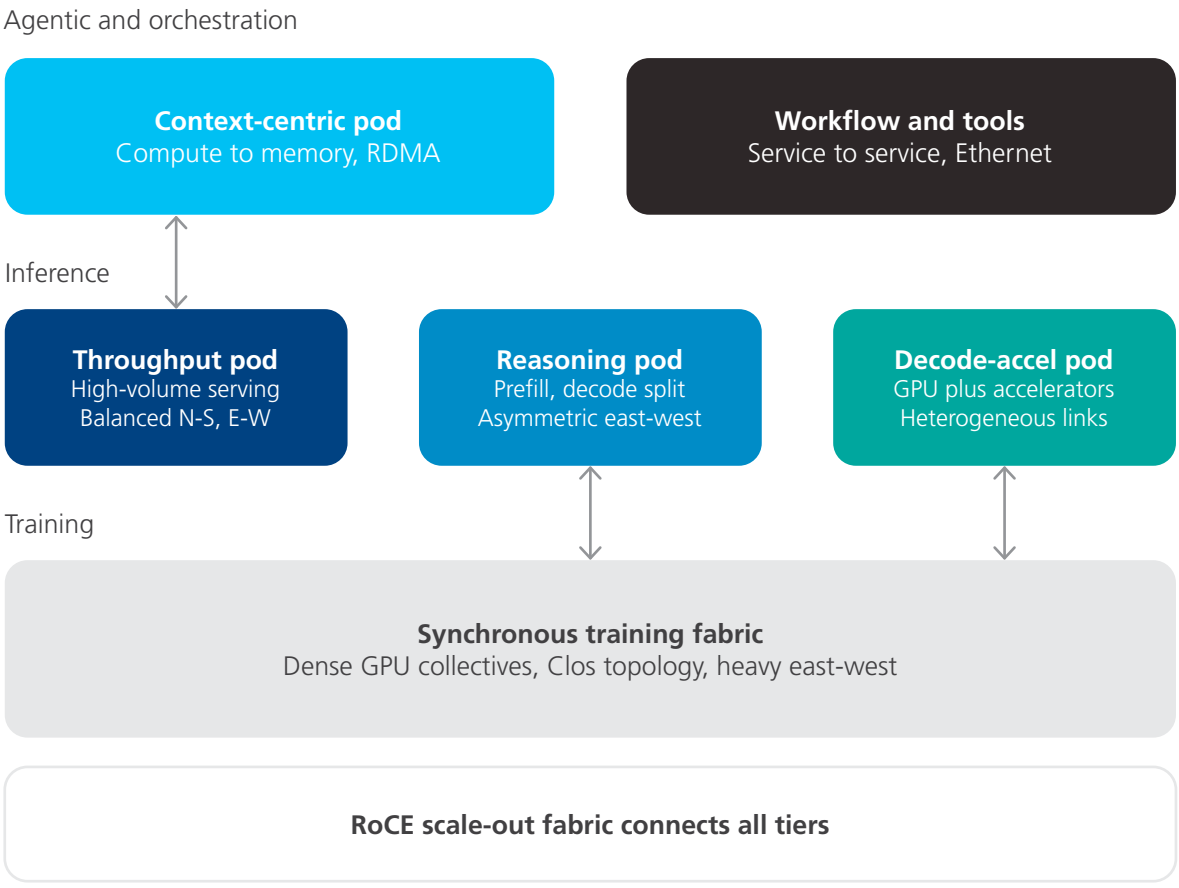
Copyright © 2026 AFL. All rights reserved.

The AI Infrastructure Planning Matrix

AI infrastructure in 2026 can be understood through six architectural categories. These are not theoretical abstractions, but a practical reflection of cluster deployment with direct implications for fiber network planning.

Across the following sections, each category is examined at a high level through four consistent lenses: workload behavior, system design, network characteristics, and fiber infrastructure requirements. This structure provides a comparative view across architectures, enabling clearer alignment between workload function and physical system design, while establishing a foundation for the more detailed engineering analysis developed in subsequent papers in this series.

The architectural stack shown here presents a bottom-up progression from foundational training and scale-out infrastructure to orchestration and context-centric services. This structure reflects the dependency flow between physical network infrastructure and higher-level AI workload functions discussed throughout the following sections.



Graphic 1 *maps the six categories into three tiers with RoCE running vertically as the common scale-out fabric:*

Key: Training · Inference variants · Context-aware · Workflow

Copyright © 2026 AFL. All rights reserved.

01

Synchronous Training Fabric

Workload Behavior

Training defines the phase in which models learn. Large datasets processed through repeated forward and backward passes, adjusting billions or trillions of parameters. GPU-to-GPU synchronizations of gradient data occurs at every step, creating a dependency in which any slowdown propagates across the entire computational phase.

How the System is Built

Training clusters are dense and highly regular, with hundreds or thousands of GPUs arranged across multi-tier Clos networks. Within each node, GPUs communicate through high-bandwidth scale-up links, typically NVLink at 900 GB/s per GPU in current-generation systems. Between nodes, a scale-out fabric carries collective operations such as allreduce, allgather, and alltoall over RoCE.

The Network

The scale-out network operates at near line-rate throughput with tightly controlled tail latency. RoCE fabrics for training commonly operate at 400GbE per port, with 800GbE deployments now emerging. Lossless behavior is required, with Priority Flow Control (PCF) and Explicit Congestion Notification (ECN) used to prevent packet loss during collective operations. Network behavior is deterministic, with engineered paths and defined hop structures.

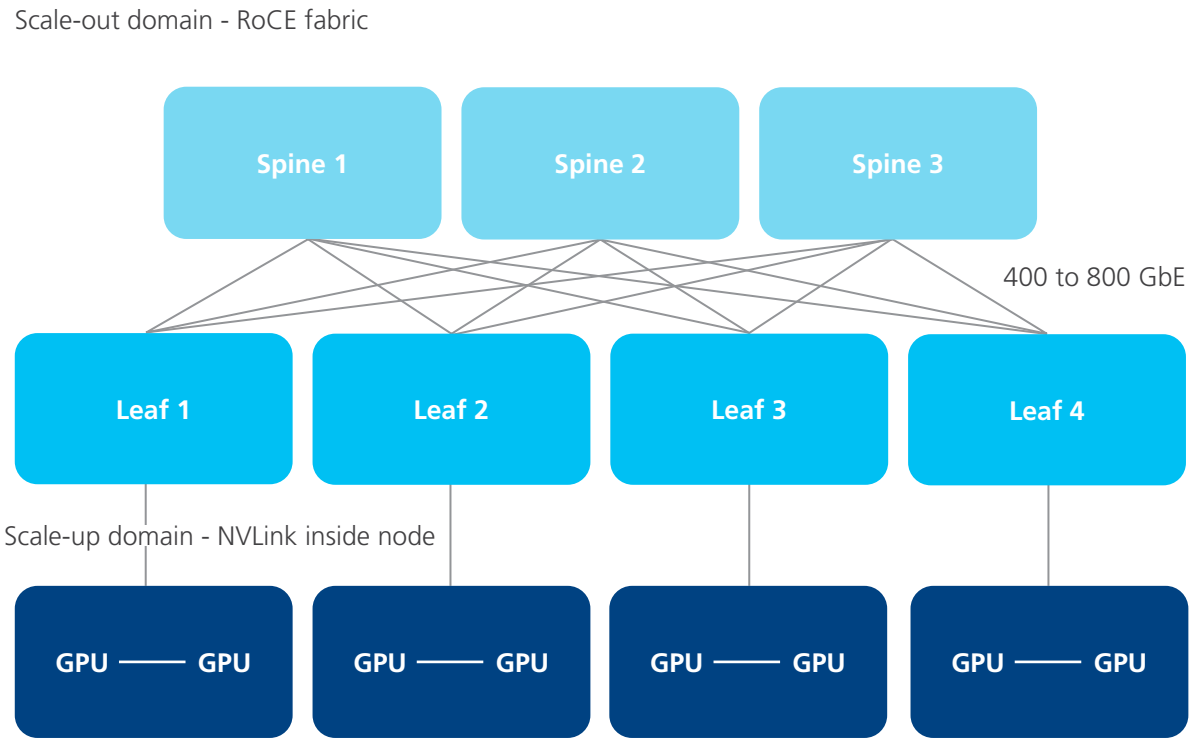
Traffic Pattern

Traffic is predominantly east-west. GPU-to-GPU communication dominates, while storage traffic contributes secondary flows through dataset reads and checkpoint writes. Traffic behavior is synchronized and burst-driven, with simultaneous transmission across nodes during collective operations creating sharp load spikes across the fabric.

Fiber Network Characteristics

Training fabrics represent the highest fiber density environments in the data center. Clos topologies generate large volumes of switch-to-switch connectivity, each requiring discrete optical paths. UHFC structured cabling is essential for managing aggregate fiber counts that scale into tens of thousands of connections within a single cluster. Reach requirements are typically short (under 100 meters within a row and up to 500 meters across a hall), while the density of parallel links makes physical routing and cable management a critical planning constraint.

Copyright © 2026 AFL. All rights reserved.



Traffic Pattern
Dense east-west: synchronised collectives across all GPUs
Fiber: high-density UHFC trunks, short reach under 500m
Thousands of parallel connections per fabric

Graphic 2 *Synchronous Training Fabric. The image renders the Clos structure tangible, showing spine and leaf tiers, the scale-up and scale-out boundary, and the point at which NVLink ends and RoCE begins. The fiber domain spans the leaf switches and all layers above.*

Copyright © 2026 AFL. All rights reserved.

02

Throughput Inference Pod

Workload Function

Throughput inference is the primary production workload in modern AI systems. High volumes of requests are processed for tasks such as classification, embedding generation, short-form text completion, and image recognition. Optimization focuses on maximizing requests per second per unit of compute rather than minimizing per-request latency, with cost efficiency as the governing constraint.

System Design

Throughput inference pods typically deploy lighter compute profiles than training environments. Mid-range or previous-generation GPUs are common, alongside PCIe-attached accelerators in place of tightly coupled NVLink systems, trading intra-node bandwidth for improved cost efficiency. Nodes are grouped into serving pools behind load balancing layers, with requests distributed dynamically across available capacity.

Network Behaviour

RoCE remains the primary transport for inter-node communication, although requirements are less extreme than in synchronous training environments. Many models fit within a single node, removing the need for cross-node tensor parallelism. Where distribution is required, pipeline parallelism introduces sequential stage-to-stage communication, producing more predictable traffic patterns than alltoall-driven training workloads. Standard 100GbE and 200GbE RoCE links are widely deployed. Lossless configuration remains relevant, although tolerance for transient congestion is higher.

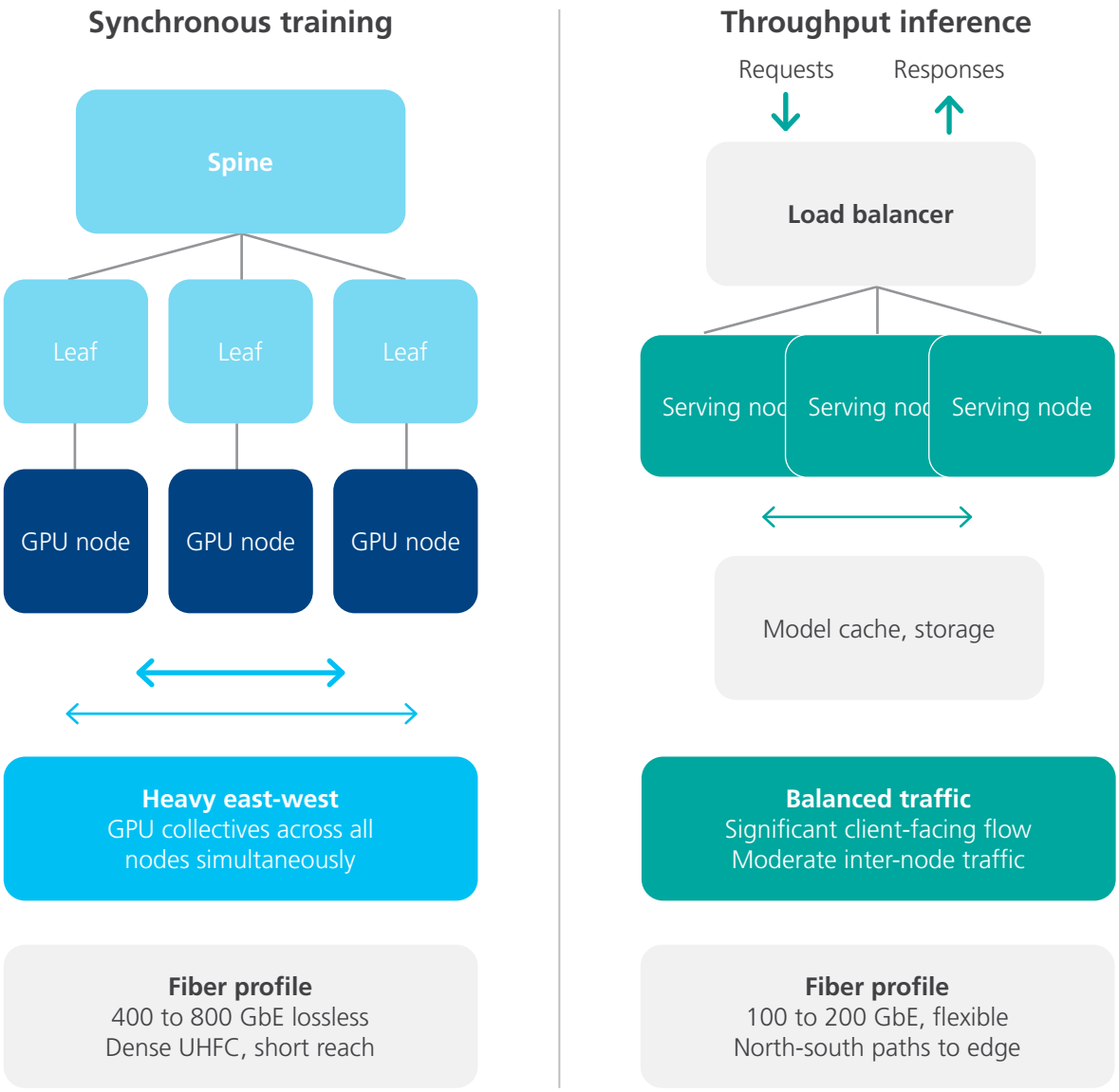
Traffic Pattern

Traffic is more balanced between east-west and north-south flows than any other AI workload category. Ingress traffic carries client requests into the cluster, while egress traffic returns generated responses. East-west traffic supports model distribution and KV-cache movement, although does not dominate overall flow characteristics.

Fiber Network Characteristics

Fiber density is lower than training environments. North-south connectivity introduces longer reach requirements from pods to aggregation layers and onward to data center edge or WAN interconnects. Link speeds are moderate relative to training fabrics. Planning emphasis shifts toward modular expansion and flexible connectivity, with horizontal scaling requiring fiber infrastructure capable of absorbing incremental node growth without extensive rework.

Copyright © 2026 AFL. All rights reserved.



Graphic 3 *translates the contrast into concrete visual terms, with arrow thickness representing relative traffic intensity. Training is dominated by heavy bilateral GPU-to-GPU traffic, while throughput inference exhibits a more balanced, client-facing flow. The result is distinct fiber profiles aligned to distinct workload requirements.*

Copyright © 2026 AFL. All rights reserved.

03

Disaggregated Reasoning Pod

Workload Function

Reasoning inference supports complex queries including multi-step chains of thought, long-context analysis, coding tasks, and document synthesis. These workloads are characterized by extended generation sequences in which models produce large token outputs, often requiring intermediate computation before response completion. Latency per token becomes the critical constraint, with direct impact on user experience.

System Design

The defining architectural pattern is separation of prefill and decode into distinct compute pools. Prefill nodes concentrate GPU resources to process input prompts in parallel under high-throughput conditions. Decode nodes prioritize sequential token generation, optimizing memory bandwidth and KV-cache access rather than peak compute throughput. This separation reflects fundamentally different hardware utilization profiles, where shared execution leads to inefficiency either in compute utilization during decode or memory bandwidth during prefill.

Network Behavior

RoCE assumes a critical architectural role due to the introduction of a workload-specific transfer pattern between compute tiers. The prefill-decode separation generates KV-cache movement, transferring intermediate attention state from prefill nodes to decode nodes. KV-cache payloads can reach gigabyte scale for long-context workloads, creating a transfer that is both latency-sensitive and unidirectional. Typical deployments utilize 400GbE RoCE links between prefill and decode tiers, with strict requirements on congestion control and tail latency.

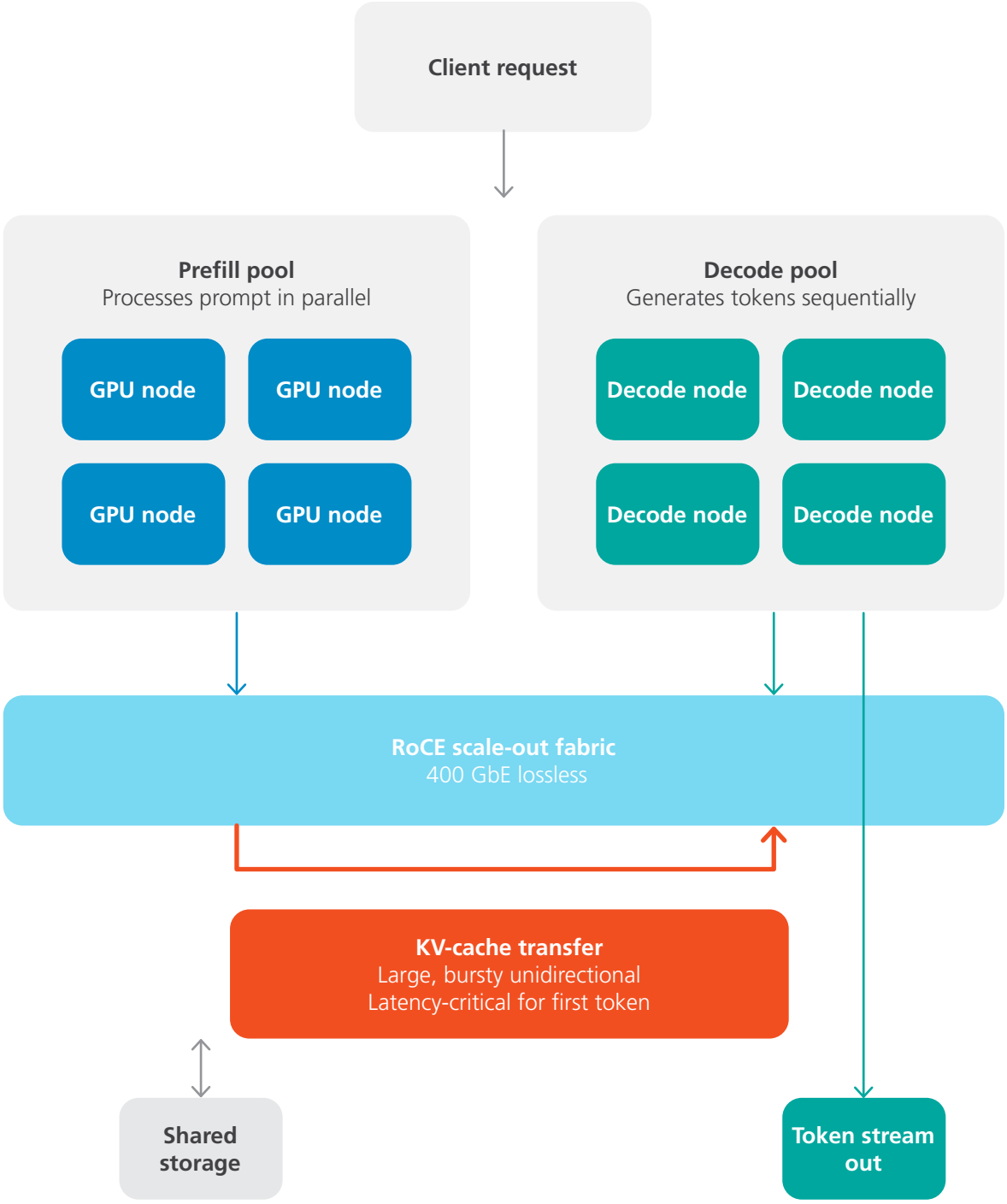
Traffic Pattern

Traffic is strongly asymmetric across the east-west plane. KV-cache transfer from prefill to decode forms a large burst in a single direction. Decode-to-client traffic remains a steady, low-volume stream of token output. Unlike training, no symmetric collective communication pattern exists. The fabric must therefore accommodate large unidirectional bursts without inducing head-of-line blocking or congestion collapse.

Fiber Network Characteristics

The prefill-to-decode interconnect introduces a distinct fiber domain spanning compute pools that may reside across racks, rows, or halls. Latency and jitter on this path directly influence user-perceived responsiveness, creating strict constraints on physical routing. Fiber design prioritizes shortest-path routing, minimized contention, and controlled pathway allocation. This interconnect functions as a discrete performance layer rather than a general extension of existing scale-out connectivity.

Copyright © 2026 AFL. All rights reserved.



Graphic 4 *is arguably the most important in the paper, showing the asymmetric traffic pattern absent in training workloads. The thick orange arrow representing KV-cache transfer contrasts with the thin green token stream, conveying the operational imbalance in a single visual frame.*

Copyright © 2026 AFL. All rights reserved.

04

Heterogeneous Decode-Accelerated Pod

Workload Function

This architecture represents an evolution of disaggregated inference. Prefill remains GPU-based, while decode execution shifts to purpose-built accelerators designed specifically for sequential, memory-bound token generation. The objective is higher decode throughput and lower cost per token relative to GPU-only execution across both phases.

System Design

The prefill tier retains high-end GPU systems comparable to those used in disaggregated reasoning pods. The decode tier introduces specialized accelerators, including custom ASICs, FPGA-based systems, and architectures optimized for high memory bandwidth and energy-efficient token generation. RoCE provides the interconnect between tiers, although hardware heterogeneity introduces differences in NIC capability, MTU configuration, and potentially RoCE implementation variants across endpoints.

Network Behavior

The RoCE fabric operates across heterogeneous compute domains, requiring support for non-uniform endpoint characteristics. Adaptive routing and congestion management increase in importance due to mixed traffic behavior across tiers. Prefill nodes generate large burst transfers, while decode accelerators produce smaller, steadier flows. Queue configuration and traffic shaping require explicit tuning to accommodate these differences, and operational standards continue to evolve as deployments mature.

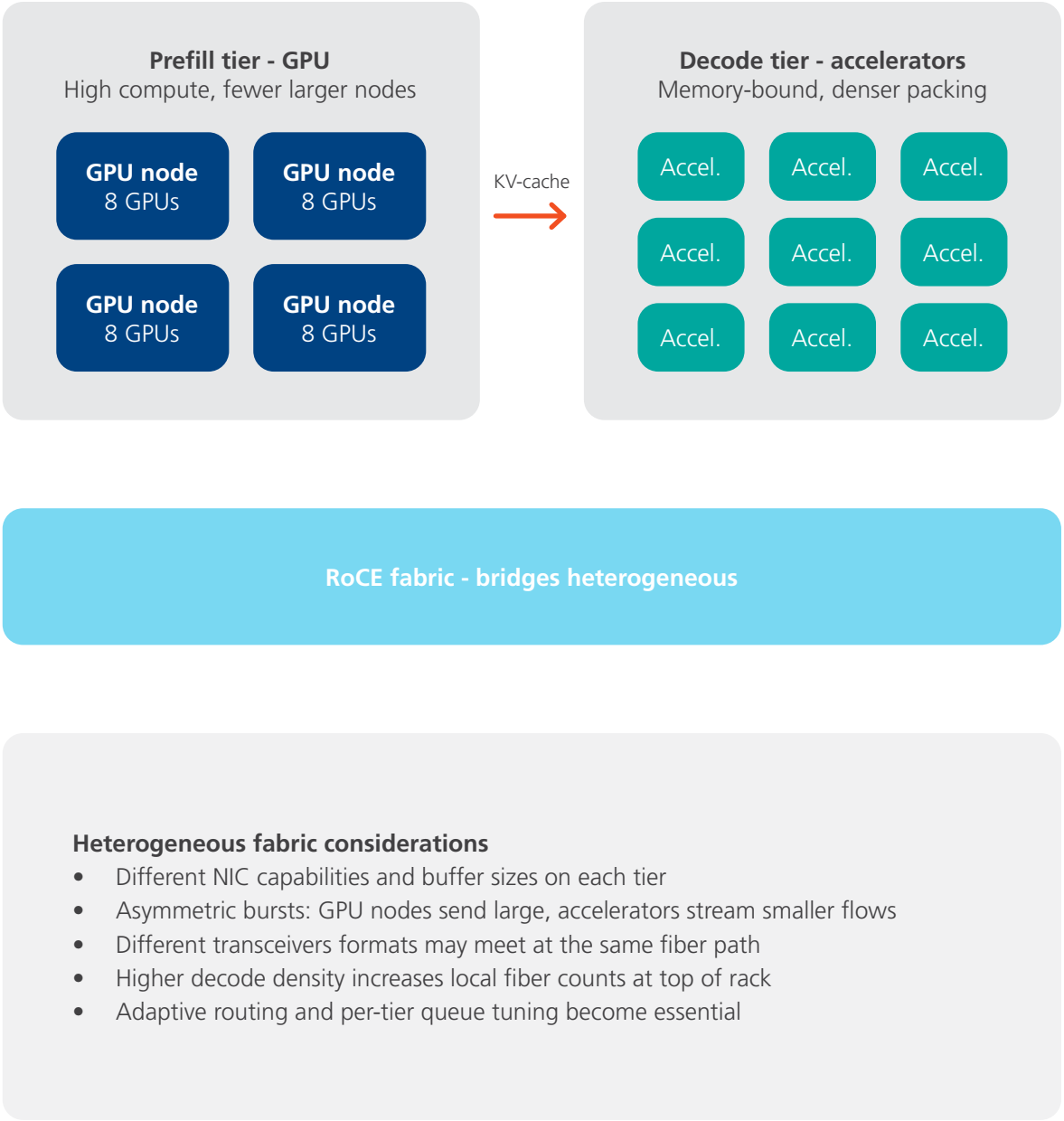
Traffic Pattern

Traffic patterns mirror disaggregated reasoning architectures, with additional variability introduced by heterogeneous execution rates. KV-cache transfer from prefill GPUs to decode accelerators remains the dominant flow. Additional decode-to-decode communication may occur in distributed accelerator configurations. Asymmetry increases due to differences in throughput, buffering, and execution cadence between GPU and decode accelerator tiers.

Fiber Network Characteristics

Fiber requirements align with disaggregated inference architectures, with increased emphasis on physical density within decode layers. Decode accelerators are often deployed at higher per-rack density than GPU systems, increasing local fiber connection density at the rack level. Endpoint heterogeneity introduces additional complexity at the physical layer, including mixed transceiver types across a single logical path and increased variation in optical characteristics across the fabric.

Copyright © 2026 AFL. All rights reserved.



Graphic 5 *shows density asymmetry, with fewer large GPU nodes on the left and compact decode accelerators densely packed on the right. The callout box highlights the practical constraints most relevant to deployment, including heterogeneous NICs, mixed transceivers, and differing buffer behaviour across a shared fabric.*

Copyright © 2026 AFL. All rights reserved.

05

Context-Centric Agentic Pod

Workload Function

Agentic AI systems operate across extended interactions with persistent state. Workloads incorporate prior exchanges, external knowledge sources, and accumulated context to inform ongoing decisions. This model introduces a requirement absent from traditional inference architectures, namely a shared context-memory layer accessible throughout execution.

System Design

The defining architectural element is a dedicated context-memory infrastructure. This infrastructure consists of a networked pool of high-bandwidth memory, commonly implemented through CXL-attached resources or RDMA-accessible DRAM and NVMe. Prefill and decode tiers interact with this memory pool to retrieve contextual data, access cached KV states, and store intermediate results. The context layer operates with memory-like latency in the microsecond range and supports concurrent access across multiple compute nodes.

Network Behavior

The architecture introduces compute-to-context, a third primary traffic class alongside existing flows. Compute-to-context communication generates sustained RDMA traffic between compute tiers and the context-memory pool. Latency sensitivity is critical, as delays in context retrieval directly translate into degraded response times. The RoCE fabric must sustain low and predictable latency across the context path while supporting concurrent prefill-to-decode and north-south traffic. Effective congestion isolation between traffic classes becomes a core design requirement.

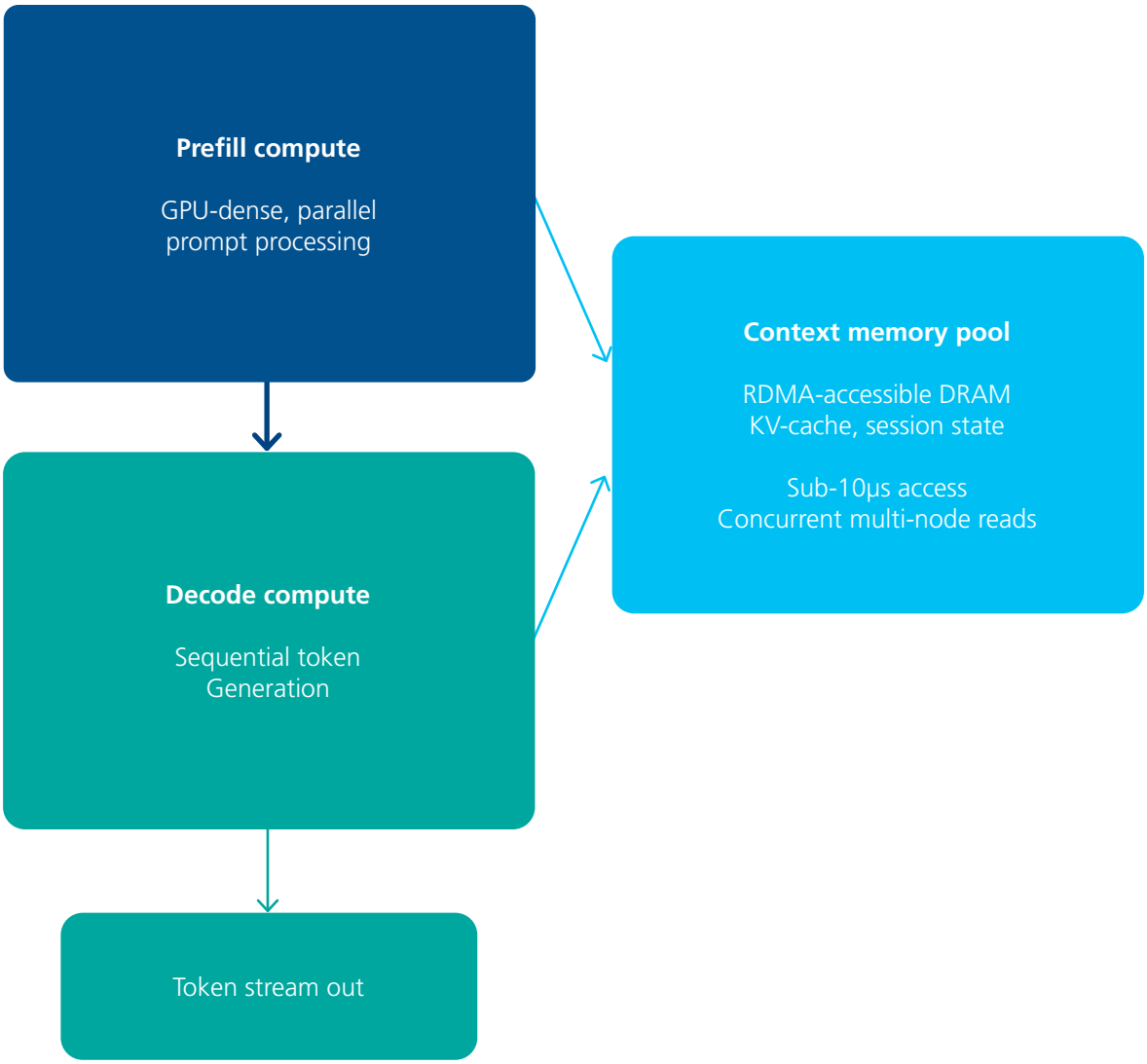
Traffic Pattern

Traffic consists of three intersecting east-west flows, including prefill-to-decode, compute-to-context, and context-to-decode. North-south traffic remains present as a secondary component. The primary constraint is latency predictability rather than aggregate bandwidth, with consistent sub-15-microsecond access required under load conditions.

Fiber Network Characteristics

The context-memory layer introduces a distinct optical domain within the data center. Dedicated fiber paths are required to maintain latency control, avoiding contention with bulk traffic associated with training or general inference. This architecture aligns with emerging references to a context fabric. Implementation may involve separate physical fiber infrastructure or controlled wavelength allocation within shared plant systems. Fiber planning expands to include three discrete connectivity domains, spanning compute-to-compute, compute-to-context, and north-south traffic flows.

Copyright © 2026 AFL. All rights reserved.



Graphic 6 *anchors the agentic architecture, making the third fiber domain visible through the compute-to-context connectivity absent from earlier designs. The graphic provides clear colour-coded reference, reinforcing separation between traffic classes and associated fiber paths.*

Copyright © 2026 AFL. All rights reserved.

06

Agent Workflow and Tool-Execution Estate

Workload Function

Agentic systems extend beyond text generation into orchestration of external tools, including databases, APIs, code interpreters, web services, and retrieval systems. Multi-step workflows execute in sequence, with outputs from one stage determining subsequent actions. This orchestration layer operates alongside inference infrastructure while remaining architecturally distinct.

System Design

The workflow estate is CPU-centric. Orchestration engines, retrieval-augmented generation pipelines, vector databases, tool servers, and API gateways run on conventional server platforms aligned with cloud-native deployment models. Select components may incorporate lightweight GPU resources for embedding generation, although overall architecture remains non-GPU-dominated.

Network Behavior

Standard Ethernet (rather than RoCE) underpins communication across the workflow estate. Traffic consists of service-to-service exchanges, including HTTP and gRPC calls, database queries, and API interactions. Latency requirements operate within low millisecond ranges rather than microsecond constraints. Network architecture aligns with cloud-native service mesh patterns rather than AI-specific fabric design.

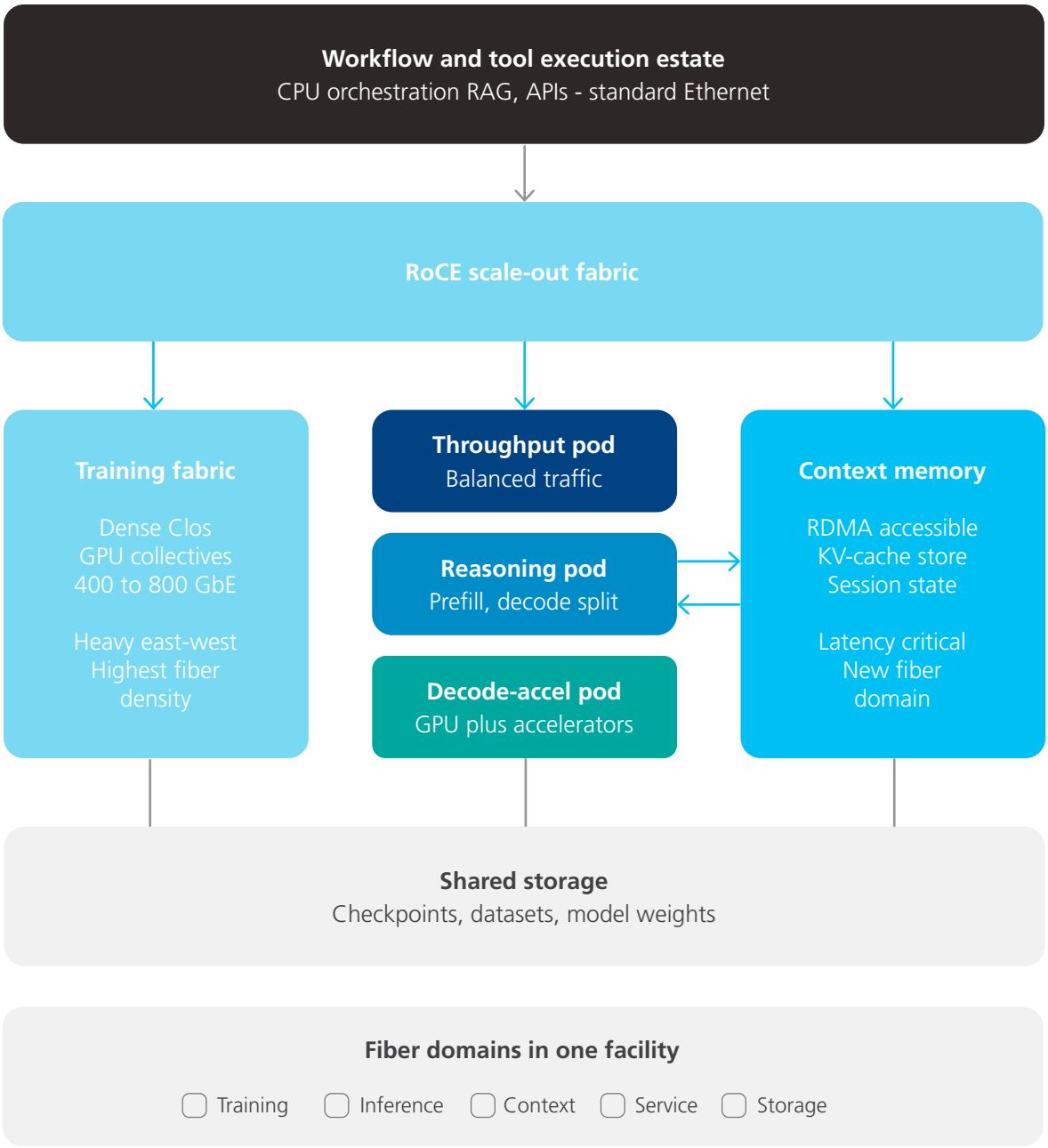
Traffic Pattern

East-west traffic volume is high, although flow characteristics differ from AI compute fabrics. Packet sizes are smaller, flows are shorter-lived, and communication patterns resemble distributed web-scale service environments rather than collective GPU operations.

Fiber Network Characteristics

Fiber infrastructure aligns with conventional data center design, including standard density, reach, and structured cabling practices. Planning considerations arise from colocation with AI compute fabrics within the same facility. Shared trunk routes require coordinated capacity planning across both workflow and inference traffic. Deployment models vary between physical separation and rack-level integration, with the latter introducing additional complexity at top-of-rack connectivity layers.

Copyright © 2026 AFL. All rights reserved.



Graphic 7 *presents all six domains within a unified architectural view, with the RoCE fabric forming the horizontal spine, standard Ethernet positioned above for workflow systems, and storage layers below.*

Core Architectural Insight: From One Fabric to Many

Three structural shifts are driving the direction of AI infrastructure:

1. **Training and inference are separating.** Early systems treated both as a shared, time-sliced cluster. Current deployments define each as a distinct system with different hardware profiles, different network requirements, and different fiber topologies. Infrastructure designed around a single assumption introduces immediate rework risk, particularly at scale.

Copyright © 2026 AFL. All rights reserved.

2. **Inference is fragmenting into multiple architectures.** No single inference model exists. Throughput serving, disaggregated reasoning, heterogeneous decode, and context-centric agentic inference each operate as separate systems with distinct network behaviors. Colocation within a single facility remains common, while shared fabric design does not.

3. **New infrastructure tiers are emerging.** Context-memory systems and workflow orchestration layers function as load-bearing components with independent connectivity requirements. Fiber planning limited to compute-to-compute connectivity omits critical system domains, resulting in incomplete architectural design.

RoCE as the Unifying Scale-Out Protocol

Across all six categories, RoCE-based Ethernet emerges as the primary scale-out interconnect, spanning communication across racks, rows, and halls.

Within tightly coupled environments, scale-up domains inside a node rely on proprietary interconnects such as NVLink. Within traditional service layers, standard Ethernet with TCP/IP is sufficient. At the scale-out boundary, where most fiber infrastructure resides, RoCE provides the transport layer that carries distributed AI workloads.

This position has shifted over time. InfiniBand previously occupied this role across HPC and early AI deployments and continues to feature in vertically integrated NVIDIA DGX and SuperPOD architectures. Broader deployment patterns now favor RoCE. Commodity Ethernet switching, standard optical components, and integration with existing data center tooling enable operational alignment across multi-vendor environments. RoCE combines RDMA performance characteristics with deployment flexibility that InfiniBand-based stacks do not consistently provide at scale.

The trade-off lies in operational complexity. RoCE depends on explicit configuration of lossless Ethernet behavior, since lossless operation is not inherent to standard Ethernet designs. PFC, ECN, and Data Center Quantized Congestion Notification (DCQCN) require careful tuning. Misconfiguration introduces PFC storms, head-of-line blocking, and performance instability that can be difficult to isolate without deep network visibility. These constraints are manageable, although they require specialized operational maturity.

Subsequent papers in this series will examine RoCE network design in detail, including topology selection, switch architecture, lossless configuration, and performance validation under AI workload conditions.

Fiber Infrastructure: From Backbone to Multi-Fabric System

Fiber infrastructure is no longer a uniform backbone interconnecting identical systems. Fiber now operates as a set of specialized fabrics, each aligned to a distinct workload domain.

Training fabrics are dense, regular, and predictable. High fiber counts, short reach requirements, and structured cabling define the environment. The primary design challenge is physical routing, with thousands of fibers requiring management within constrained pathway capacity.

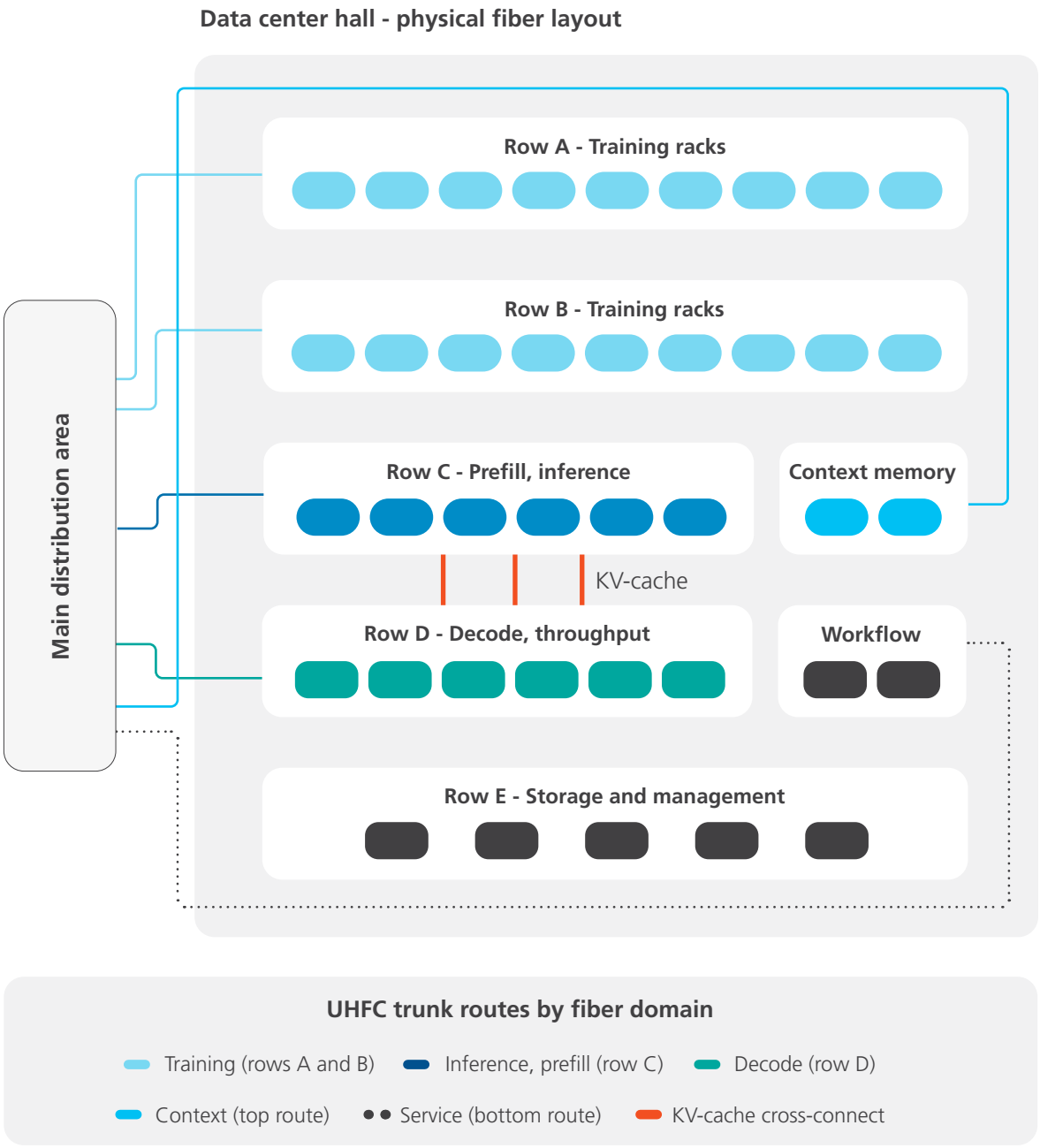
Inference fabrics introduce asymmetry. Prefill-to-decode links carry distinct performance requirements relative to decode-to-client traffic. North-south capacity planning must operate alongside east-west design considerations, with both influencing aggregate fiber architecture.

Context fabrics represent a distinct category. Low-latency, low-jitter connectivity to shared memory systems defines the requirement set. These paths may require physical separation from bulk traffic routes to maintain predictable performance under load.

Copyright © 2026 AFL. All rights reserved.

Service fabrics align more closely with conventional data center cabling models, while operating within the same physical environments as high-density AI infrastructure. Coexistence within shared facilities introduces additional planning complexity across all layers of the fiber plant.

The practical implications are material. Fiber planning now spans multiple link classes with distinct performance envelopes. UHFC structured cabling systems become critical not only for density management but for coordinating overlapping fiber domains. Latency sensitivity in specific paths, particularly context-memory connectivity, introduces physical routing constraints that extend beyond traditional backbone design assumptions.



Graphic 8 *brings the architecture into the physical domain, showing rack rows, fiber trunk routes from the MDA, and color-coded domains illustrating how four fiber types coexist within a single hall. Line weights represent relative density, with training trunks shown as the most substantial paths and context paths as the lightest. The prefill-to-decode cross-connect between rows C and D highlights the KV-cache path as a physical routing constraint within the fiber plant.*

Copyright © 2026 AFL. All rights reserved.

Implications for Optical Architecture

The traditional network model, defined by frontend, backend, storage, and management layers, no longer maps cleanly onto AI infrastructure. Modern AI facilities require explicit architectural definition across multiple connectivity domains:

- **Prefill compute:** GPU-dense environments with high east-west bandwidth carried over RoCE
- **Decode compute:** GPU or accelerator-based systems with asymmetric traffic patterns relative to prefill tiers
- **Context-memory:** RDMA-accessible memory pools with strict latency requirements, emerging as a distinct fiber domain
- **Workflow and services:** CPU-based orchestration layers using standard Ethernet and service-mesh traffic patterns
- **Storage systems:** high-throughput connectivity supporting checkpointing and dataset access, commonly via NVMe-oF or parallel file systems
- **Management networks:** out-of-band control, monitoring, and telemetry infrastructure

Each domain introduces distinct requirements across bandwidth, latency, and reliability. A single-network abstraction leads to uneven resource allocation, with over-provisioning in some areas and insufficient capacity in others. An architectural approach based on discrete domains, with clearly defined interface boundaries, enables more efficient, resilient, and adaptable infrastructure design.

Copyright © 2026 AFL. All rights reserved.

What Comes Next in This Series

This paper presents the evolving AI architecture landscape: six workload categories, multiple coexisting fabrics, and a scale-out layer increasingly built on RoCE. Logically, follow-up questions must focus on physical implementation, specifically the implications for fiber infrastructure. The next two papers in this series move from architecture to engineering, translating workload requirements into physical layer design.

White Paper 2: Building a Large-Scale AI Training Cluster

Training clusters are the most demanding optical environments within a data center. Large-scale deployments require tens of thousands of fiber connections, routed through pathway systems not originally designed for this density and terminated on interfaces operating at 400G and higher. This second White Paper in the series focuses specifically on large-scale AI training clusters (often referred to as "AI factories"), while acknowledging that similar architectural principles apply across both hyperscale environments and emerging neocloud deployments. The paper develops a reference-scale training cluster and derives fiber architecture from first principles. A three-layer cabling model is introduced, consisting of high-count backbone trunks for permanent infrastructure, zone cabling for topology adaptation and meshing, and equipment cords for final device connectivity. This structure separates long-term infrastructure investment from short-term adaptation. Within this framework, analysis covers the relationship between switch topology and fiber count, the role of UHFC structured cabling in managing density within constrained pathways, and the operational disciplines of multi-fiber connectivity, polarity, inspection, and testing required for efficient deployment.

While the underlying hardware and cabling considerations are consistent across deployment models, implementation may vary depending on the operator. Hyperscalers and large AI developers may design and operate these environments directly, whereas neocloud providers (acting as colocation specialists, infrastructure partners, or service providers) may build and manage similar clusters privately or on behalf of customers, often leveraging vendor ecosystems to supplement in-house engineering capabilities.

Copyright © 2026 AFL. All rights reserved.

Paper 3: Heterogeneous Inference and Multi-Fabric Fiber Design

Inference introduces new fiber design challenges. Disaggregated prefill-decode architectures, mixed GPU and accelerator pods, and emerging context-memory tiers create connectivity requirements not present in training environments. These include asymmetric traffic flows, latency-sensitive inter-tier paths, and multiple fabric domains operating within a shared physical plant. Paper 3 addresses co-location across these domains, examining how the three-layer cabling model adapts when backbone infrastructure supports multiple fabrics, when zone cabling must accommodate domain-specific adaptation requirements, and how design choices between direct connection, modular flexibility, and cross-connect architectures align with domain stability and change frequency.

Paper 4: Conclusion

AI infrastructure has progressed beyond a single-architecture model. Training systems, inference platforms, and agentic environments each impose distinct requirements on network design and underlying fiber infrastructure. The scale-out layer is increasingly standardized on RoCE over Ethernet. The optical layer functions as a structured system composed of backbone, zone, and equipment cord layers, each defined by purpose, lifespan, and rate of change. Correct design across these layers determines system capability, deployment speed, and adaptability under evolving workload conditions.

Every fiber route, connector interface, splice, and patch point represents a design decision with direct performance implications. Precision at scale, across density and speed requirements defined by AI workloads, defines the core engineering challenge addressed throughout this series.

Copyright © 2026 AFL. All rights reserved.

Key Terms

Several foundational terms are used throughout this paper and across the series.

Accelerators

Accelerators are any processor optimized for AI workload execution. The category includes GPUs, Google TPUs, custom ASICs, and emerging architectures. Throughout this series, "accelerator" is the general term unless a specific device class is intended.

Accelerator Nodes

Accelerator Nodes are discrete compute units containing multiple accelerators, the associated memory, host CPUs, and local interconnects. The node is the fundamental building block of an AI cluster. Scale-up connectivity operates within the node; scale-out connectivity operates between nodes. Node configurations vary from eight GPUs in a standard server to hundreds of accelerators in tray-based or rack-scale designs.

Clos topology

Clos topology is a multi-tier network architecture widely used in modern data centers. Clos networks are structured as leaf-spine fabrics in which each leaf switch connects to multiple spine switches, creating multiple equal-cost paths between endpoints. In AI clusters, Clos topologies provide the scale-out structure connecting racks of accelerators. Cluster scale is determined by switch radix (port count) and the number of network tiers.

East-west traffic

East-west traffic refers to data moving laterally between servers, accelerators, or compute nodes within the data center, including GPU-to-GPU communication and compute-to-memory tier interactions. This is the dominant traffic pattern inside AI clusters.

GPUs (Graphics Processing Unit)

GPUs (Graphics Processing Unit) were originally designed for rendering graphics, and are now the dominant processor for AI training and inference. GPUs execute thousands of parallel arithmetic operations simultaneously, a structure well matched to the matrix mathematics underlying neural network computation. NVIDIA's data center GPUs (H100, H200, B200/GB200) define the current accelerator landscape. The term is often used interchangeably with "accelerator," though it is technically a subset.

IB (InfiniBand)

IB (InfiniBand) is a high-bandwidth, low-latency interconnect protocol historically dominant in HPC and early AI training clusters. InfiniBand provides native RDMA and has been the preferred fabric in NVIDIA's vertically integrated DGX and SuperPOD architectures. Ethernet with RoCE has emerged as the primary alternative for open, multi-vendor scale-out deployments. This series focuses on Ethernet/RoCE fabrics.

Latency

Latency is the total elapsed time between initiating a data transfer and completion. In AI cluster networking, latency has three principal components: serialization delay (time to place bits on the link), propagation delay (speed-of-light transit through fiber or copper), and switch latency (processing time within each network element). Switch latency, typically 300–500 nanoseconds per hop in modern cut-through switches, accumulates across each tier of a Clos fabric. A three-tier network adds six switch hops per round trip. Total fabric latency directly affects collective communication performance in distributed training and tail latency in inference serving.

Copyright © 2026 AFL. All rights reserved.

North-south traffic

North-south traffic refers to data entering or exiting the cluster, including client requests, returned outputs, and data transfers involving external storage systems. Traditional enterprise data centers were primarily north-south oriented, whereas AI clusters are primarily east-west.

Prefill and decode

Prefill and decode describe the two primary phases of large language model inference. Prefill processes the full input sequence in parallel and is compute-intensive. Decode generates output tokens sequentially and is both latency-sensitive and memory-bandwidth constrained. Increasingly, prefill and decode phases are executed on separate hardware tiers, introducing additional east-west traffic between inference components.

RDMA (Remote Direct Memory Access)

RDMA (Remote Direct Memory Access) is the underlying mechanism implemented by RoCE. RDMA enables one machine to read from or write to another machine's memory without CPU intervention on either side of the transfer. This reduces latency and system overhead, supporting performance requirements in distributed AI workloads.

RoCE (RDMA over Converged Ethernet)

RoCE (RDMA over Converged Ethernet) enables direct memory access between servers over Ethernet networks without involving the operating system or CPU in the data transfer path. This reduces software overhead and supports the low-latency, high-throughput communication required by AI workloads. RoCE has become the dominant protocol for AI scale-out networking outside environments built around proprietary interconnect fabrics such as NVLink-based systems. This paper (and series) examines RoCE in detail.

Scale across

Scale across describes connectivity between separate clusters, sites, or administrative domains. Where scale-up operates within a node and scale-out operates within a cluster, scale-across links distinct clusters or facilities. Use cases include multi-cluster training runs, shared storage tiers, and federated inference serving across geographically distributed sites. Scale-across traffic is primarily north-south in character but may carry east-west workloads when training jobs span cluster boundaries.

Scale-out

Scale-out refers to the network fabric connecting nodes across racks and rows, representing the primary domain for inter-rack connectivity and most of the deployed fiber infrastructure. Ethernet and RoCE operate at this layer.

Scale-up

Scale-up refers to high-bandwidth connectivity within a single node or tightly coupled accelerator domain. This includes interconnects between accelerators within a server or across a single chassis. These links are short-reach, high-speed, and often proprietary or semi-proprietary.

UHFC (Ultra-High Fiber Count) cabling

UHFC (Ultra-High Fiber Count) cabling refers to structured cabling systems with extremely high fiber densities per cable assembly, typically 864 fibers and above (reaching 6,912 or more), with modern deployments extending into the thousands of fibers per trunk. These systems support the aggregation demands of large-scale AI clusters and reduce pathway congestion within high-density infrastructure environments.

Copyright © 2026 AFL. All rights reserved.

ASIC

ASIC (Application Specific Integrated Circuit) refers to a purpose-built semiconductor device designed for a specific function such as packet forwarding in switches or compute acceleration in AI systems.

CXL

CXL (Compute Express Link) is a high-speed, low-latency interconnect standard that enables coherent memory sharing between CPUs and GPUs. CXL extends and operates over PCIe as the physical layer.

DRAM

DRAM (Dynamic Random Access Memory) refers to a form of volatile semiconductor memory used as primary working memory in compute systems, supporting high-bandwidth access to active data and model states. DRAM memory cells are very small but must be continuously refreshed.

FPGA

FPGA (Field Programmable Gate Array) refers to a reconfigurable semiconductor device used for programmable hardware of all types including acceleration, protocol processing, and adaptable compute workloads in networking and AI environments.

gRPC

gRPC (Google Remote Procedure Call) refers to a high-performance communication framework enabling structured service-to-service function calls across distributed microservices and AI orchestration environments. Developed by Google, gRPC is a widely used open source protocol.

MTU

MTU (Maximum Transmission Unit) is an Ethernet control parameter and refers to the largest packet size transmitted over a network interface without fragmentation, influencing throughput efficiency and congestion behavior in data center networks. Longer MTU improves large block transfer while shorter MTU reduces latency.

NIC

NIC (Network Interface Card) refers to a hardware component that provides network connectivity for servers and accelerators, often incorporating RDMA-capable Ethernet or InfiniBand interfaces.

NVMe

NVMe (Non-Volatile Memory Express) refers to a high-performance storage protocol over PCIe designed for low-latency, parallel access to solid-state storage in data center and AI workloads. NVMe supports RDMA allowing remote access to local data.

PCIe

PCIe (Peripheral Component Interconnect Express) refers to a high-speed serial interconnect standard used for communication between CPUs, accelerators, memory subsystems, and peripheral devices within compute systems. Multiple serial lanes can be bonded together to create higher bandwidth channels.

Copyright © 2026 AFL. All rights reserved.

Founded in 1984, AFL is an international manufacturer providing end-to-end network solutions to the energy, service provider, enterprise, hyperscale and industrial markets. The company's products are in use in over 130 countries and include fiber optic cable, assemblies, and hardware, transmission and substation accessories, outside plant equipment, connectivity, test and inspection equipment, fusion splicers, and training. AFL also offers a wide variety of services supporting data center, enterprise, wireless and outside plant applications.

Headquartered in Spartanburg, SC, AFL has operations in the U.S., Mexico, Canada, Europe, Asia and Australia, and is a wholly owned subsidiary of Fujikura Ltd. of Japan.

The information contained within this white paper is accurate and up-to-date to the best of our knowledge at the time of production. All graphs and visual representations are proprietary assets of AFL. These materials are intended for informational purposes only, and may not be used for commercial purposes without express permission from AFL.

Copyright © 2026 AFL. All Rights Reserved E&OE WP-05016