Mplify White Paper

Carrier Ethernet in Distributed Enterprise AI

The role of predictable, measurable connectivity in distributed AI infrastructure economics

September 2026

Disclaimer

The information in this publication is freely available for reproduction and use by any recipient and is believed to be accurate as of its publication date. Such information is subject to change without notice and Mplify Alliance (Mplify) is not responsible for any errors. Mplify does not assume responsibility to update or correct any information in this publication. No representation or warranty, expressed or implied, is made by Mplify concerning the completeness, accuracy, or applicability of any information contained herein and no liability of any kind shall be assumed by Mplify as a result of reliance upon such information.

The information contained herein is intended to be used without modification by the recipient or user of this document. Mplify is not responsible or liable for any modifications to this document made by any other party.

The receipt or any use of this document or its contents does not in any way create, by implication or otherwise:

a) any express or implied license or right to or under any patent, copyright, trademark or trade secret rights held or claimed by any Mplify member which are or may be associated with the ideas, techniques, concepts or expressions contained herein; nor

b) any warranty or representation that any Mplify members will announce any product(s) and/or service(s) related thereto, or if such announcements are made, that such announced product(s) and/or service(s) embody any or all of the ideas, technologies, or concepts contained herein; nor

c) any form of relationship between any Mplify member and the recipient or user of this document.

Implementation or use of specific Mplify standards, specifications, or recommendations will be voluntary, and no Member shall be obliged to implement them by virtue of participation in Mplify Alliance. Mplify is a non-profit international organization to enable the development and worldwide adoption of agile, assured and orchestrated network services. Mplify does not, expressly or otherwise, endorse or promote any specific products or services.

Contents

1. Executive Summary

Enterprise AI is becoming a distributed infrastructure problem

Enterprise AI infrastructure is becoming distributed. Power and GPU capacity may be available across different colocation facilities, while proprietary or regulated data may need to remain within a specific jurisdiction or controlled environment. As compute and data spread across locations, the network connecting them becomes part of the AI production system.

This is especially relevant for enterprises fine-tuning, adapting and operating domain models using proprietary data. These workloads create an inter-site connectivity requirement distinct from the tightly coupled communication inside a GPU cluster. They need a WAN service that can connect distributed resources with predictable, measurable performance.

For suitable distributed AI workloads, Carrier Ethernet can provide a private inter-site service envelope with committed capacity, measurable delay and delay variation, a defined frame-loss objective, engineered paths and operational accountability.

Three reasons the WAN now matters

- Sovereignty: regulated or sensitive data may need to remain within a country, region or controlled facility.
- Data gravity: moving large volumes of operational data to a distant central location can be slower and more expensive than bringing compute closer to the data.
- Power and GPU availability: using capacity across multiple facilities can allow enterprises to deploy AI resources without waiting for one location to provide all the required power and accelerators.

The implication is strategic: the WAN becomes a managed part of the AI production system, with performance tied directly to workload requirements.

2. Workload Fit

Different AI workloads create different network requirements

NVLink, InfiniBand, and RoCEv2 support tightly coupled, microsecond-scale communication required inside GPU environments. Carrier Ethernet has a different role: connecting enterprise AI sites when the workload can accommodate physical distance and when predictable, measurable transport supports operational, economic or governance requirements.

Why parameter-efficient fine-tuning changes the equation

Full-model training can require participants to exchange very large volumes of model state. Parameter-efficient methods such as LoRA freeze the base model and update a much smaller set of parameters. As an illustrative

example, the 16-bit weights of a 70-billion-parameter model represent approximately 140 GB, while the adapter or gradient state exchanged during parameter-efficient-fine-tuning may be substantially smaller. The actual volume depends on the model, tuning method, configuration and synchronization strategy.

ARCHITECTURAL BOUNDARY

Carrier Ethernet can improve transport predictability, while propagation delay remains determined by physical distance. Model architecture, parallelism strategy, update size, synchronization frequency and fiber distance determine whether a workload is suitable for inter-site training.

Where Carrier Ethernet Fits Best

WORKLOAD	POTENTIAL APPLICABILITY
SOVEREIGN REGIONAL FINE-TUNING	Potentially well suited
METRO MULTI-COLO TRAINING	Potentially well suited
CHECKPOINT / MODEL REPLICATION	Well suited
DISTRIBUTED INFERENCE STAGES	Workload dependent
CONTINENTAL SYNCHRONOUS TRAINING	Highly workload dependent
FRONTIER TENSOR PARALLELISM ACROSS WAN	Generally unsuitable

3. Sovereignty

Keep the data local. Share only approved model updates.

In regulated industries, AI architecture is shaped by where data may be stored, processed, and transferred. A distributed fine-tuning pattern keeps raw records within an approved region or controlled environment while limiting cross-site exchange to specifically reviewed and authorized learning artifacts.

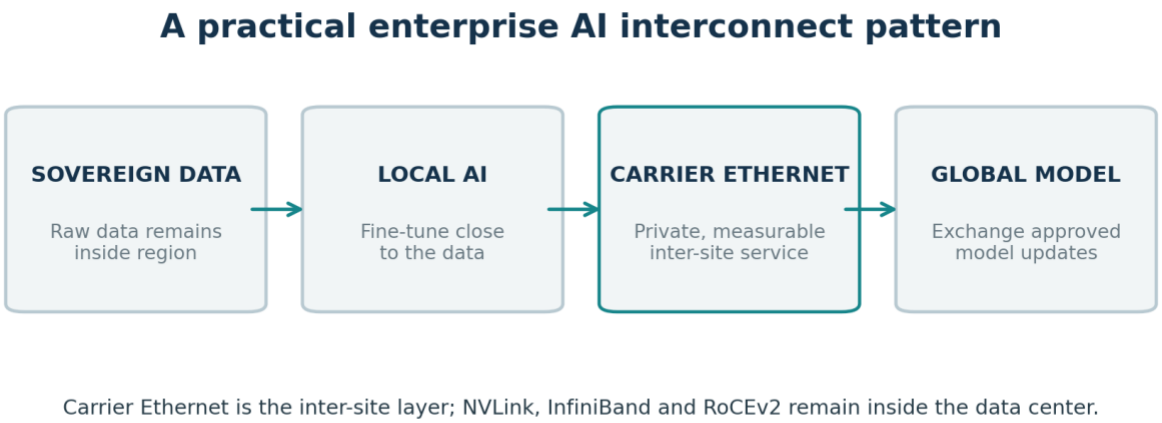

Figure 1. Carrier Ethernet connects sovereign AI sites at the inter-data-center layer.

A practical operating pattern

- Raw customer, patient, financial or proprietary data is processed inside the approved regional AI pod.
- Local GPUs calculate gradients or parameter-efficient adapter updates.
- Reviewed and authorized model updates, gradients, adapters or other learning artifacts may be exchanged with an aggregation or alignment service.
- The network path, security controls and performance are documented and monitored.

COMPLIANCE CAUTION

A private network path alone does not establish regulatory compliance, and model updates are not automatically non-sensitive. Privacy leakage, encryption, retention, route constraints and legal requirements still require workload-specific validation.

4. GPU Economics

The slowest synchronization path can determine the pace of the step

In synchronous distributed training, compute and communication are coupled. A delayed worker or region can extend the synchronization barrier and leave the rest of the participating accelerator pool waiting. The financial impact depends on the workload, the number and cost of the accelerators, the duration of the delay, and how frequently synchronization occurs.

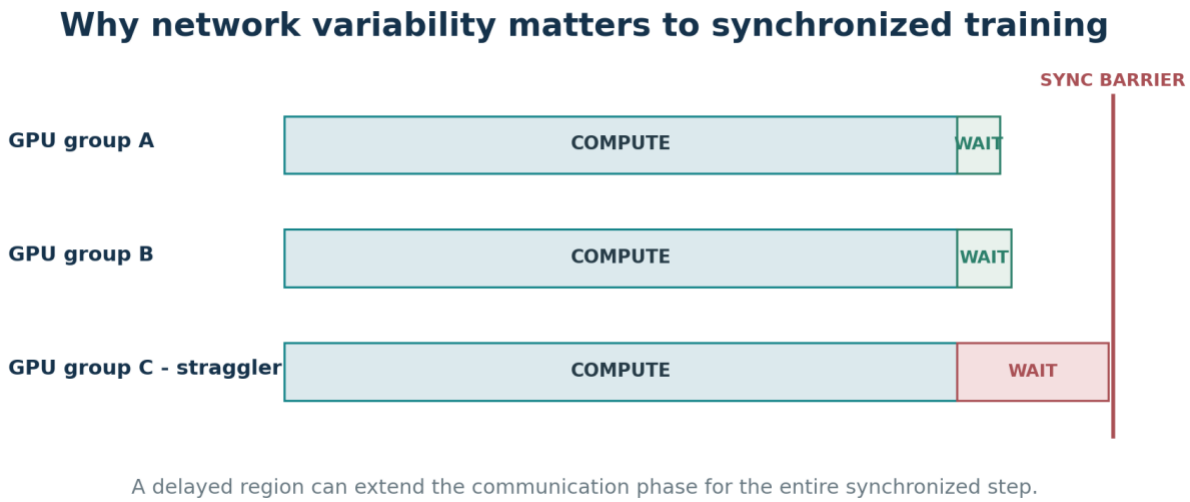

Figure 2. A WAN-induced straggler can extend the communication phase for the entire synchronized step.

Why best-effort transport can become expensive

Congestion, queueing, and competing high-volume flows can introduce loss and variable latency. In synchronous workloads, that variability can extend communication phases and create avoidable GPU idle time. Asynchronous approaches may tolerate more delay; however stale updates can affect convergence and must be evaluated by the AI team.

CFO LENS

Evaluate network costs in the context of the AI infrastructure it supports. A lower-cost connectivity option may increase the effective cost of operating high-value GPU resources if repeated network delays extend training windows or leave accelerators waiting.

5. Carrier Ethernet

Buy a measurable service envelope - not bandwidth in isolation

Carrier Ethernet provides a standardized framework for defining and measuring inter-site service performance. Its value comes from combining private connectivity, committed capacity, agreed performance objectives, engineered paths, service assurance and clear operational accountability.

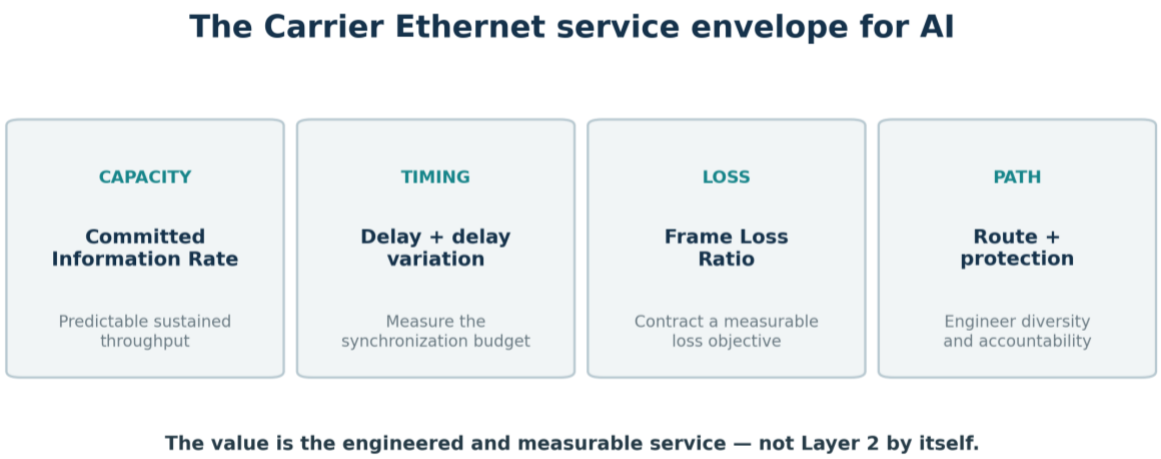

Figure 3. Four Carrier Ethernet attributes that matter directly to distributed AI.

What to engineer

Committed capacity. Use CIR to reserve sustained throughput for model updates, checkpoints and synchronized transfers.

Timing. Measure delay and delay variation against the workload's synchronization budget.

Frame loss. Define an FLR objective and measurement method rather than relying on vague 'lossless WAN' language.

Route and protection. Define primary and diverse paths, jurisdictional constraints, protection mechanisms, and responsibility for restoration.

Edge behavior. Validate MTU, priority mapping, encryption requirements, and the transition from the local RoCEv2, InfiniBand, or Ethernet environment to the WAN service.

6. Reference Architecture

Carrier Ethernet connects the inter-data-center layer

NVLink and other accelerator interconnects support communication within GPU systems, while InfiniBand, RoCEv2, and Ethernet support cluster and storage communication within the data center. In this reference architecture, Carrier Ethernet begins at the data-center edge and connects these local fabrics across metro, regional or selected long-distance routes.

The Clean Boundary

LAYER	TECHNOLOGY	ROLE
GPU / INTRA-NODE	NVLink / accelerator fabric	Local accelerator communication
INTRA-DATA-CENTER	RoCEv2 / InfiniBand / Ethernet	Cluster collectives and storage
DATA-CENTER EDGE	Border gateway / demarc	Burst absorption, policy and mapping
INTER-DATA-CENTER	Carrier Ethernet E-Line / EVPL	Predictable site-to-site transport
CONTROL	NaaS APIs / orchestration	Provision, scale and assure connectivity

DESIGN PRINCIPLE

Design the data-center edge as the transition between the local cluster environment and the WAN service. Depending on the architecture, this boundary may provide buffering, traffic shaping, priority mapping, MTU validation, encryption, and service demarcation.

This separation is important for public positioning: Carrier Ethernet provides the inter-data-center layer for suitable distributed enterprise AI workloads, while local GPU fabrics continue to support tightly coupled communication within each site.

7. Automation

NaaS can make Carrier Ethernet follow the AI workload

Enterprise fine-tuning is often episodic. A nightly training run may need very high bandwidth for several hours, while the same sites need much less capacity during normal operations.

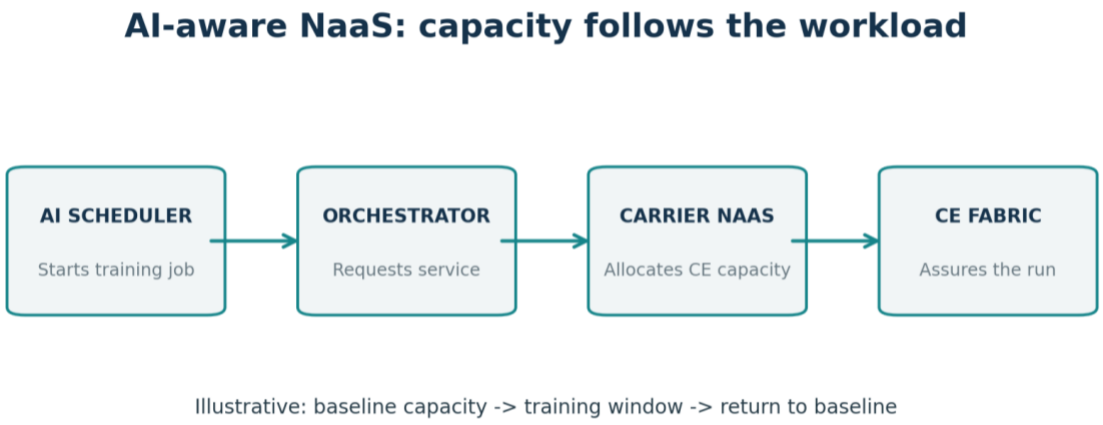

Figure 4. AI scheduler -> orchestration -> carrier NaaS -> assured Carrier Ethernet service.

The degree of automation available today varies by provider, service, and enterprise environment. The following sequence represents a target operating model for workload-driven connectivity.

Target Operating Model

- The AI scheduler launches a training, fine-tuning, replication, or model-alignment job.
- Enterprise orchestration translates the workload's requirements into connectivity intent.
- The required connectivity and service attributes are requested through NaaS APIs.
- The carrier NaaS platform provisions or scales the Carrier Ethernet service.
- Telemetry verifies performance while the job runs.
- Capacity returns to baseline after completion.

STRATEGIC VALUE
Coordinating compute placement, workload schedules, and network capacity can make connectivity a programmable component of the enterprise AI operating environment.

8. Procurement

What an AI interconnect RFP should require

Enterprises should connect network requirements directly to the workload rather than relying on universal "AI network" specifications. The following areas provide a starting point for evaluating an inter-site service.

RFP DIMENSION	WHAT TO EVALUATE
CAPACITY	CIR, oversubscription policy, burst behavior and scaling increments.
LATENCY	One-way and round-trip delay on primary and protection paths.
VARIATION	Frame Delay Variation / jitter objective and measurement method.
LOSS	Frame Loss Ratio objective, measurement window and remediation.
PATH	Physical route, diversity, jurisdictions and failure path.
EDGE	MTU, priority mapping, encryption and demarcation behavior.
AUTOMATION	Ordering, bandwidth changes, telemetry and assurance APIs.

PROCUREMENT PRINCIPLE

Define the workload, validate its communication and synchronization requirements, measure the path, and establish service objectives that support those requirements.

9. Strategic Recommendation

Position Carrier Ethernet as an inter-data-center layer for distributed enterprise AI

Carrier Ethernet is most relevant when enterprises need to connect sovereign AI environments, distributed GPU capacity, and regional data centers with predictable, measured performance and private connectivity.

Its value comes from packaging committed throughput, defined timing and loss objectives, engineered routes, security, protection, assurance, and accountability into a standardized inter-data-center service that enterprises can procure and increasingly automate.

BOARD-LEVEL TAKEAWAY

Enterprise AI will increasingly be constrained by where power, GPUs and data are available. Carrier Ethernet gives the enterprise a practical way to connect those distributed assets with greater predictability and accountability - turning the WAN into a managed part of the AI production system.

Recommended Industry Actions

These requirements point to several opportunities for Mplify and its members to advance a more interoperable, automated approach to distributed AI connectivity:

- Define an AI Interconnect profile with workload-oriented Carrier Ethernet performance attributes.
- Standardize colo-to-colo on-ramps for private GPU cages across operators and regions.
- Advance API-enabled ordering, bandwidth changes, telemetry, and assurance across NaaS environments.
- Publish reference architectures that clearly separate intra-DC GPU fabrics from inter-DC Carrier Ethernet transport.

Source Basis

This public executive edition is based on the supplied architectural content covering enterprise fine-tuning, data sovereignty, GPU synchronization, Carrier Ethernet, intra-/inter-data-center fabric boundaries and NaaS automation. Production thresholds should be validated against the specific workload, carrier design and applicable standards.

Mplify 2026