BMWG Working Group
Internet-Draft
Intended status: Informational
Expires: 23 October 2026

F. Calabria
Cisco
C. Pignataro
Blue Fern Consulting
Q. Wu
G. Fioccola
Huawei

21 April 2026

Benchmarking Terminology for AI Network Fabrics
draft-calabria-bmwg-ai-fabric-terminology-01

Abstract

This document defines benchmarking terminology for evaluating Ethernet-based network fabrics used in distributed Artificial Intelligence (AI) training and inference workloads. It provides a unified vocabulary consolidating and extending terms from RFC 1242, RFC 8238, and the companion AI fabric methodology documents, establishing precise, vendor-neutral definitions for collective communication primitives, RDMA transport mechanisms (RoCEv2 and Ultra Ethernet Transport), congestion control behaviors, AI-specific Key Performance Indicators (KPIs), and fabric topology concepts.

This document is a companion to [I-D.calabria-bmwg-ai-fabric-training-bench] and [I-D.calabria-bmwg-ai-fabric-inference-bench]. Those documents SHOULD NOT be applied without first consulting the terminology defined herein. Where definitions herein overlap with RFC 1242 or RFC 8238, the AI fabric context definition in this document takes precedence.

About This Document

This note is to be removed before publishing as an RFC.

The latest revision of this draft can be found at https://fcalabri.github.io/bmwg-ai-fabric-terminology/draft-calabria-bmwg-ai-fabric-terminology.html. Status information for this document may be found at https://datatracker.ietf.org/doc/draft-calabria-bmwg-ai-fabric-terminology/.

Source for this draft and an issue tracker can be found at https://github.com/fcalabri/bmwg-ai-fabric-terminology.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction 3
1.1. Requirements Language 3
1.2. Scope and Purpose 3
1.3. Relationship to Existing BMWG Work 3
1.4. Relationship to Companion Documents 4
2. General Benchmarking Terms 4
3. Collective Communication Terms 6
4. Distributed Parallelism Strategy Terms 8
5. Network Transport Terms 10
5.1. RoCEv2 and RDMA Terms 10
5.2. Ultra Ethernet Transport (UET) Terms 11
5.2.1. UET Transport Services Comparison 13
6. Congestion Control and Fabric Behavior Terms 14
6.1. Load Balancing Strategy Comparison 16
7. Fabric Topology and Infrastructure Terms 16

8. Training-Specific Terms 19
9. Inference-Specific Terms 20
9.1. Inference Phase Characteristics 24
10. KPI Classification Terms 25
10.1. KPI Tier Summary 26
11. Referenced Standards Abbreviations 26
Acknowledgments 28
References 28
Normative References 28
Informative References 28
Appendix A: Term Cross-Reference to Companion Documents 29
Appendix B: Term Taxonomy Summary 31
Authors' Addresses 32

1. Introduction

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

1.2. Scope and Purpose

This document defines terminology specifically for benchmarking Ethernet-based AI network fabrics in controlled laboratory environments. The defined terms cover: distributed AI training collective communication patterns, LLM inference serving architectures, RDMA transport semantics (RoCEv2 and UET), congestion control mechanisms, fabric topology characteristics, and performance metric definitions.

This document does not define acceptance criteria, performance requirements, or configuration recommendations. It does not address benchmarking of live operational networks, intra-node (NVLink/PCIe) interconnects, or storage networking.

1.3. Relationship to Existing BMWG Work

This document extends the foundational BMWG terminology established in [RFC1242] (network interconnect benchmarking terminology) and [RFC8238] (data center benchmarking terminology). Where terms are defined in those RFCs, this document provides AI fabric context extensions; the core definitions remain as established. This document also extends the test methodology framework of [RFC2544] and [RFC8239] as applied in the companion AI fabric methodology

documents.

1.4. Relationship to Companion Documents

This document is one of three companion Internet-Drafts addressing AI fabric benchmarking:

* [I-D.calabria-bmwg-ai-fabric-terminology] (this document): Terminology definitions.

* [I-D.calabria-bmwg-ai-fabric-training-bench]: Benchmarking methodology for AI training workloads.

* [I-D.calabria-bmwg-ai-fabric-inference-bench]: Benchmarking methodology for AI inference serving workloads.

Implementers and evaluators SHOULD read this terminology document before applying the companion methodology documents. Terms defined here are used normatively in those documents and are not redefined there unless the specific workload context introduces a substantive difference, which is noted explicitly.

2. General Benchmarking Terms

The following terms establish the general measurement framework applicable to all AI fabric benchmarking activities.

Term	Definition
AI Fabric	The dedicated Ethernet backend network interconnecting accelerators (GPUs/XPUs) for distributed AI training and inference workloads. Typically implemented as a non-blocking Clos (fat-tree) topology running RoCEv2 or UET transport. Distinct from the front-end (management/storage) network.
DUT	Device Under Test. The network element(s) whose performance characteristics are being measured. In AI fabric benchmarking the DUT is one or more fabric elements: leaf switches, spine switches, NICs, or the complete fabric assembly.
SUT	System Under Test. The complete AI compute system including accelerators, NICs, the fabric DUT, and serving/training software, when end-to-end metrics are the measurement objective.

RT	Router Tester / Traffic Generator. Test equipment capable of generating and receiving network traffic at specified rates with nanosecond-resolution timestamping sufficient for the measurements defined in the companion methodology documents.
JFI	Jain's Fairness Index. A scalar measure of flow-level throughput fairness across n flows: JFI = (Σ x) / (n Σ x) where x is the throughput of flow i. A value of 1.0 indicates perfect fairness; lower values indicate disparity. *SHOULD* be computed per [RFC1242] reporting conventions.
Offered Load	The total traffic rate presented to the DUT from test equipment, expressed as a fraction of line rate (0100%) or as absolute bit/s. Offered load is controlled independently of DUT absorption, enabling characterization of saturation behavior.
Trial Duration	The time interval over which a single measurement is conducted. For AI fabric tests, the *RECOMMENDED* minimum is 60 seconds for throughput tests and 300 seconds for soak/stability tests, per the methodology in [RFC2544] as extended herein.
Warmup Period	A mandatory pre-measurement interval during which traffic is sent but results are not recorded. Ensures adaptive routing tables, PFC watermarks, and DCQCN/UET congestion controllers reach steady state before measurement begins. *RECOMMENDED* minimum: 10 seconds.
Binary Search	An iterative test procedure for determining the maximum offered load at which a DUT meets a specified acceptance criterion (e.g., zero packet loss). The search halves the candidate load range at each iteration, converging to a resolution of 0.1% offered load within 10 iterations.
Percentile Latency	A latency statistic expressing that the specified fraction of all measured latency samples fall at or below the reported value. Denoted Pxx (e.g., P50, P95, P99, P99.9). Tail latency (P99 and above) is especially relevant for AI fabric

	benchmarking because SLO violations are determined by worst-case, not median, performance.

Table 1: General Benchmarking Terms

3. Collective Communication Terms

The following terms define the collective communication operations that are the primary traffic sources in distributed AI workloads.

Term	Definition
Collective Operation	A coordinated communication pattern executed simultaneously across all accelerators in a training or inference group. Core collectives: AllReduce (gradient aggregation), AllGather (parameter distribution), ReduceScatter (partial reduction + scatter), and AllToAll (expert dispatch in MoE models).
AllReduce	A collective in which each participant contributes a tensor and all participants receive the element-wise sum (or other reduction) of all contributions. The dominant communication primitive in data-parallel and tensor-parallel training. BusBW is the primary KPI.
AllGather	A collective in which each participant contributes a shard of a tensor and all participants receive the concatenation of all shards. Used in tensor-parallel (Megatron-style) layers to reconstruct distributed activations or parameters.
ReduceScatter	A collective combining an element-wise reduction with a scatter, so each participant receives a distinct slice of the reduced result. Used in ZeRO-stage optimizer strategies and as the first half of a ring-AllReduce.
AllToAll	A collective in which each participant sends a distinct payload to every other participant and receives a distinct

	payload from every other participant. The critical collective for Mixture-of-Experts token dispatch. Generates N−1 independent point-to-point flows for N participants.
Ring Algorithm	An AllReduce (or AllGather/ReduceScatter) algorithm structured as a logical ring of participants. Each participant sends to its right neighbor and receives from its left neighbor in 2(N−1) steps. Bus bandwidth efficiency = 2(N−1)/N, approaching 100% for large N. Standard baseline for BusBW calculation.
BusBW	The effective data throughput per accelerator during a collective operation, computed as: BusBW = (data_size × algo_factor) / time where algo_factor normalizes for the collective type and algorithm: Collective Algorithm algo_factor AllReduce Ring / recursive doubling 2 × (n−1) / n AllReduce Binary / double-binary tree 2 × log(n) / n AllGather Ring (n−1) / n ReduceScatter Ring (n−1) / n AllToAll Direct (n−1) / n n = number of participating accelerators. Ring AllReduce is the conventional comparison baseline. Note: collective libraries commonly select the algorithm dynamically based on message size (e.g., tree-based for small messages, ring for large messages); algo_factor therefore varies with message size and MUST be reported per message-size bucket when dynamic selection is active. Reports MUST state: collective type, algorithm, algo_factor value, collective

	library name and version, and n. Units: Gbps per accelerator.
CCL	Collective Communication Library. A software library providing optimized implementations of collective operations (AllReduce, AllGather, etc.) over a specific transport. The CCL implementation *MUST* be documented in the test report.
SPMD	Single Program Multiple Data. The execution model underlying bulk-synchronous distributed training, in which all accelerators execute identical computation on distinct data partitions, synchronizing at collective barriers between steps.
Bulk Synchronous Parallel (BSP)	A distributed computation model structured as alternating compute and communicate phases with a global synchronization barrier between phases. Standard training workloads follow BSP: forward pass → backward pass → AllReduce gradient sync → optimizer step.

Table 2: Collective Communication Terms

4. Distributed Parallelism Strategy Terms

The following terms define the parallelism strategies used in distributed AI model training and inference, which determine traffic patterns and fabric requirements.

Term	Definition
Data Parallelism (DP)	A distributed training strategy replicating the full model on each accelerator, partitioning the training dataset across replicas. Gradient synchronization after each backward pass requires an AllReduce across all DP ranks. Memory-efficient for small models; communication overhead scales with parameter count.
*Tensor	A distributed training and inference strategy

Parallelism (TP)*	partitioning individual weight matrices across multiple accelerators. Each rank computes a partial result; AllGather or ReduceScatter collectives are required within each layer to aggregate results. Dominant parallelism within a node (intra-node).
Pipeline Parallelism (PP)	A distributed strategy assigning contiguous groups of transformer layers to distinct stages (accelerators or nodes). Each stage processes one microbatch and forwards activations to the next stage. Generates point-to-point inter-stage traffic across the fabric (activations and gradients).
Expert Parallelism (EP)	A parallelism strategy for Mixture-of-Experts models distributing expert sub-networks across accelerators. Each token is routed to its designated experts (typically top-K of E total experts), requiring AllToAll communication for dispatch. Wide EP (e.g., 96-way) generates dense inter-node AllToAll at every MoE layer.
MoE	Mixture of Experts. A transformer architecture replacing dense feed-forward layers with a set of E expert sub-networks, of which only top-K experts (typically K=2 or K=4) are activated per token via a learned router. MoE enables large model capacity with sub-linear compute, but introduces AllToAll communication requirements proportional to E and sequence length.
DP Attention	Data Parallelism applied to the attention computation, where the KV cache is partitioned across data-parallel ranks. Each rank holds 1/ DP_SIZE of the KV cache; AllToAll communication exchanges attention outputs. Used in inference to reduce per-accelerator memory footprint for long contexts.
ZeRO	Zero Redundancy Optimizer. A memory optimization strategy for data-parallel training that shards model states (parameters, gradients, optimizer states) across DP ranks instead of replicating them. Stage 1 shards optimizer states; Stage 2 adds gradient sharding; Stage 3 adds parameter sharding. Each stage increases AllGather/ReduceScatter communication.

Table 3: Distributed Parallelism Strategy Terms

5. Network Transport Terms

5.1. RoCEv2 and RDMA Terms

The following terms define RDMA and RoCEv2 transport semantics as used in AI fabric benchmarking. UET, PDC, and ROD are included here for direct comparison with their RoCEv2 counterparts; full UET-specific terms are defined in Section 5.2.

Term	Definition
RDMA	Remote Direct Memory Access. A transport mechanism enabling direct memory-to-memory data transfer between hosts without involving the destination CPU, providing zero-copy semantics and kernel bypass. Implementations include InfiniBand Verbs (native IB), iWARP (RDMA over TCP), and RoCEv2 (RDMA over Converged Ethernet v2).
RoCEv2	RDMA over Converged Ethernet version 2. An RDMA transport encapsulating InfiniBand transport layer (BTH) over UDP/IP, enabling RDMA semantics on standard Ethernet infrastructure. Requires lossless fabric operation (PFC or equivalent) for correctness. Standardized in IBTA Annex 16; transported over UDP destination port 4791.
QP	Queue Pair. The fundamental RDMA communication endpoint comprising a Send Queue (SQ) and Receive Queue (RQ). QPs are connection-oriented in Reliable Connected (RC) mode. Multiple QPs per source-destination pair are used to increase ECMP entropy in fabric load balancing.
Reliable Connected (RC)	An RDMA QP transport service type providing reliable, in-order delivery between exactly two endpoints. The primary QP type for AI collective operations via RoCEv2. Requires connection setup before data transfer and maintains per-QP state for retransmission.
RDMA Verb	An operation primitive of the RDMA programming model. Key verbs: SEND/RECV (two-sided, receiver

	must post a buffer), WRITE (one-sided, target memory written directly), READ (one-sided, remote memory read), and Atomic (compare-and-swap, fetch-and-add). AI collectives predominantly use WRITE and SEND.
UET	Ultra Ethernet Transport. A transport protocol defined by the Ultra Ethernet Consortium (UEC) Specification 1.0 as a next-generation AI/HPC fabric transport. UET is connectionless, supports native packet spraying (RUD), and integrates multipath load balancing and congestion control. Transported over UDP destination port 4793 (pending IANA verification).
PDC	Packet Delivery Context. The ephemeral, lightweight transport endpoint in UET, analogous to but distinct from an RDMA Queue Pair. PDCs are connectionless (no setup handshake), enabling low-latency initiation and reduced per-flow state in the NIC and switch.
ROD	Reliable Ordered Delivery. A UET transport service providing reliable, in-order packet delivery, semantically equivalent to RoCEv2 RC mode. Suitable for legacy RDMA applications requiring strict ordering guarantees.

Table 4: RoCEv2 and RDMA Terms

5.2. Ultra Ethernet Transport (UET) Terms

The following terms define UET-specific concepts introduced by the Ultra Ethernet Consortium (UEC) Specification 1.0 [UEC-SPEC-1.0].

Term	Definition
RUD	Reliable Unordered Delivery. A UET transport service providing reliable delivery without maintaining packet order across paths. Enables native packet spraying across ECMP paths without reorder-buffer overhead at the receiver NIC. The preferred UET service class for AI training collectives.
RUDI	Reliable Unordered Delivery for Idempotent

	operations. A UET transport service optimized for operations safe to execute more than once (e.g., RDMA Writes to non-accumulating targets), allowing simplified retransmission logic with reduced state overhead.
UUD	Unreliable Unordered Delivery. A UET transport service providing best-effort, unordered packet delivery with minimal overhead. Suitable for telemetry, speculative operations, or workloads with application-layer loss tolerance.
UEC Profile	A defined subset of UET features targeting a specific use case: AI Base (core AI training/ inference, mandatory feature set), AI Full (AI Base plus deferred send, exact-match tagging, extended atomics), or HPC (latency-optimized for traditional HPC workloads with fine-grained synchronization).
LLR	Link Layer Retry. An optional UEC link-layer enhancement providing fast per-hop error recovery at the Ethernet link layer. LLR detects symbol errors at the FEC level and retransmits the affected frame before it is dropped, reducing the frequency of transport-layer retransmission and improving tail latency.
Packet Trimming	An optional UEC link-layer behavior in which a congested switch, rather than dropping the full packet, transmits only the packet header (trimmed packet) to the receiver. Trimming enables the receiver to detect loss and initiate selective retransmission more rapidly, reducing bandwidth waste versus silent drop.
CBFC	Credit-Based Flow Control. An optional UEC link-layer buffer management mechanism using explicit credit grants from downstream to upstream devices. CBFC provides backpressure without transmitting PFC PAUSE frames, eliminating the head-of-line blocking and storm propagation risks associated with PFC.
Entropy Value	A per-packet field in the UET header used to distribute packets of a single message across available ECMP paths, providing explicit spray entropy independent of the IP 5-tuple. Enables hardware-assisted packet spraying without requiring transport-layer state in the switch.

GIN	GPU-Initiated Networking. A communication paradigm in which GPU threads directly initiate network RDMA operations (sends, one-sided writes/reads) to the NIC hardware without CPU involvement, eliminating the CPU-GPU synchronization round-trip. Reduces effective latency by several microseconds for fine-grained operations.
KVCXL	KV Cache Transfer Library. A software library providing standardized point-to-point data transfer primitives (register, transfer, notify) for inference engines, abstracting underlying transport mechanisms (intra-node interconnect, RDMA, PCIe, storage interfaces). Enables transport-agnostic KV cache migration in disaggregated serving architectures.

Table 5: Ultra Ethernet Transport (UET) Terms

5.2.1. UET Transport Services Comparison

Service	Ordered	Reliable	Retransmission Complexity	Primary Use Case
ROD	Yes	Yes	Full per-QP state	Legacy RDMA / ordered AI ops
RUD	No	Yes	Reduced (unordered)	AI training collectives with spray
RUDI	No	Yes	Minimal (idempotent)	RDMA Writes; simple retransmit
UUD	No	No	None	Telemetry, speculative ops

Table 6: UET Transport Services Comparison

6. Congestion Control and Fabric Behavior Terms

The following terms define congestion management mechanisms and associated fabric behaviors critical to AI workload performance.

Term	Definition
PFC	Priority Flow Control (IEEE 802.1Qbb). A lossless Ethernet mechanism in which a receiver transmits a PAUSE frame to its upstream neighbor on a specific priority class when its ingress buffer approaches a configured threshold, temporarily halting transmission of that priority. Required for lossless RoCEv2 operation. PFC operates hop-by-hop and can propagate congestion upstream (PFC storm risk).
PFC Storm	A pathological condition in which PFC PAUSE frames propagate across multiple hops, causing widespread throughput degradation or deadlock unrelated to the original congestion source. Detection and mitigation *SHOULD* be part of soak test evaluation per the companion methodology documents.
PFC Deadlock	A circular PFC dependency in which sets of flows mutually pause each other indefinitely, resulting in zero progress for affected traffic classes. Deadlock risk is elevated in non-tree topologies and *MUST* be evaluated in fabric-level soak tests.
ECN	Explicit Congestion Notification ([RFC3168]). An IP-layer mechanism in which a congested router marks packets with the Congestion Experienced (CE) codepoint in the IP ECN field instead of dropping them. The receiver echoes congestion feedback to the sender via the transport protocol, triggering rate reduction. Used with RoCEv2 as part of DCQCN.
DCQCN	Data Center Quantized Congestion Notification. An end-to-end congestion control algorithm for RoCEv2 flows, combining ECN marking at congested switches with rate-based sender reduction using an AIMD scheme. Note: PFC serves as a separate, orthogonal backstop to

	prevent packet loss during DCQCN convergence; PFC is *not* a component of the DCQCN algorithm itself.
ECN Marking Ratio	The fraction of packets (expressed as a percentage) that are marked with the CE codepoint in the IP ECN field over a measurement interval. A high ECN Marking Ratio indicates persistent congestion and is a primary Fabric Health Indicator.
Incast	A traffic pattern in which multiple sources simultaneously send to a single destination, potentially overwhelming the destination's NIC receive buffer and the switch's egress port buffer. Incast is a dominant congestion mechanism in AllReduce and collective operations.
Incast Ratio	The ratio of concurrent senders to receivers in an incast communication pattern (N:1). The incast ratio determines the oversubscription factor at the destination port and is a primary test parameter for congestion characterization.
Packet Spray	A load balancing strategy distributing individual packets of a single RDMA message across all available ECMP paths, maximizing link utilization at the cost of potential out-of-order delivery at the receiver. Native in UET (RUD mode); requires NIC reorder buffering for RoCEv2 RC mode.
DLB / Flowlet	Dynamic Load Balancing using flowlet detection. A per-flow rerouting mechanism that reassigns a flow to a new ECMP path when the flow has been idle longer than the flowlet gap threshold (typically 500 ns2 s), reducing out-of-order packet risk compared to packet spray while improving utilization over static per-flow ECMP.
ECMP	Equal-Cost Multi-Path routing. A forwarding mechanism distributing traffic across multiple equal-cost paths, typically via hash of the IP 5-tuple (or entropy field in UET). ECMP imbalance (MMR > 1.0) is a primary fabric efficiency metric for AI traffic.

MMR	Max-Mean Ratio. The ratio of the flow count (or traffic load) on the most heavily utilized link to the average flow count per link across all fabric links. MMR = 1.0 indicates perfect ECMP balance; MMR > 1.0 quantifies imbalance that degrades effective fabric bandwidth.

Table 7: Congestion Control and Fabric Behavior Terms

6.1. Load Balancing Strategy Comparison

Strategy	Granularity	Reorder Risk	Utilization	Complexity
ECMP (5-tuple hash)	Per-flow	None	Low (elephant flow bias)	Low
DLB / Flowlet	Per-flowlet	Low	Medium	Medium
Packet Spray (RoCEv2)	Per-packet	High	High	High (NIC reorder buffer)
Packet Spray (UET RUD)	Per-packet	None (transport tolerates OOO)	High	Low

Table 8: Load Balancing Strategy Comparison

7. Fabric Topology and Infrastructure Terms

The following terms define fabric topology architectures and infrastructure components referenced in the companion methodology documents.

Term	Definition
Fabric DUT Boundary	The precise measurement boundary for BMWG AI fabric benchmarks. Defined as the NIC Ethernet port (transmit side at source,

		receive side at destination). All benchmarked metrics (throughput, latency, loss, congestion) are measured at or between NIC Ethernet ports. Intra-node segments (NVLink, PCIe Gen4/5, CXL) are outside the DUT boundary and MUST NOT be included in fabric benchmark results without explicit labelling as a separate measurement component.
Intra-Node Transfer Overhead		The latency and bandwidth consumed by data movement within a single server node: specifically, the GPU-to-NIC path via PCIe or CXL, and GPU-to-GPU communication via NVLink. Intra-node transfer overhead is a contextual measurement reported alongside fabric benchmarks in end-to-end decomposition tests but is not itself the benchmarked entity in any test in this document
Clos / Fat-Tree Topology		A multi-stage switch topology providing non-blocking or oversubscribed connectivity between all leaf-to-leaf pairs. In AI fabric deployments, a two-tier (leaf-spine) or three-tier (leaf-spine-superspine) Clos is standard. Full bisection bandwidth (1:1) is the target for training fabrics; 2:1 or 4:1 oversubscription may be acceptable for inference fabrics.
Rail-Optimized Topology		A topology in which the NIC ports of each server are distributed across multiple ToR switches (one NIC port per switch), such that collective traffic between adjacent servers traverses different physical paths. Minimizes switch-to-switch traffic during ring AllReduce, maximizing effective BusBW. Requires ECMP-aware collective placement.
Bisection Bandwidth		The aggregate bandwidth across the minimum cut that divides the fabric into two equal halves. Non-blocking fabrics provide bisection bandwidth equal to half the total edge (server-facing) bandwidth. Limits worst-case all-to-all

	communication throughput.
Oversubscription Ratio	The ratio of total edge (server-facing) bandwidth to total bisection bandwidth in a Clos fabric. A 1:1 ratio is non-blocking; higher ratios (e.g., 2:1, 4:1) reduce fabric cost but may bottleneck all-to-all and AllReduce patterns when all server ports are active simultaneously.
ToR Switch	Top-of-Rack switch. The first-hop aggregation switch connecting accelerator servers in a rack to the spine layer of the fabric. In rail-optimized topologies, multiple ToR switches serve a single rack, with each server's NICs distributed across ToRs.
Spine / Superspine	Intermediate and top-layer switches in a multi-tier Clos fabric, providing inter-rack and inter-pod connectivity respectively. Spine switches aggregate multiple ToR switches; superspine switches aggregate multiple spine pods.
NIC	Network Interface Controller. The hardware device providing network connectivity for an accelerator host. AI fabric NICs support RDMA (RoCEv2 or UET), hardware offload for collective operations, and, optionally, GPU-Initiated Networking (GIN). NIC model and firmware version *MUST* be documented in all benchmark reports.
Buffer Occupancy	The instantaneous or time-averaged fill level of a switch port's packet buffer, expressed in bytes or as a fraction of total buffer capacity. Elevated sustained buffer occupancy indicates congestion. P99 buffer occupancy is a Fabric Health Indicator in the companion methodology documents.
Zero-Impact Failover	Sub-microsecond automatic path convergence upon a link or switch failure resulting in no measurable increase to

		JCT or TTFT. Requires pre-programmed alternate paths and hardware-level fast reroute (FRR) with sub-microsecond detection, not relying on routing protocol convergence.
Link Utilization		The fraction of the nominal link capacity actually used for data transmission over a measurement interval, expressed as a percentage. Reported as mean, P95, and P99 per link. High asymmetric link utilization (low average but high peak) is characteristic of bursty AI inference traffic.

Table 9: Fabric Topology and Infrastructure Terms

8. Training-Specific Terms

The following terms are specific to AI training workload benchmarking and are used normatively in [I-D.calabria-bmwg-ai-fabric-training-bench].

Term	Definition
JCT	Job Completion Time. The wall-clock elapsed time from the start of a training job (or benchmark iteration) until all participating accelerators complete their work, inclusive of all forward pass, backward pass, and collective communication phases. JCT is the primary end-to-end training efficiency KPI.
Roofline JCT	The theoretical minimum JCT assuming perfect (zero-contention, zero-queuing) network behavior: Roofline JCT = computation_time + serialization_delay, where serialization_delay = message_size / link_rate. Provides a reference for evaluating fabric overhead.
JCT Ratio	The ratio of measured JCT to Roofline JCT. A value of 1.0 indicates no network-induced overhead. Values > 1.0 quantify fabric inefficiency: JCT Ratio = JCT_measured / JCT_roofline. The JCT Ratio is the primary

	comparative metric for AI training fabric benchmarking.
Gradient Synchronization	The AllReduce collective operation performed after the backward pass of each training step to sum the locally computed gradients across all data-parallel replicas. The dominant communication event in data-parallel training, occurring once per training step per layer.
Step Time	The wall-clock duration of a single training iteration (forward pass + backward pass + gradient synchronization + optimizer step). Step time = computation time + communication time, where the communication time is dominated by the AllReduce collective.
Soak Test	A sustained-load test run for an extended period (minimum 24 hours for stability evaluation) at a defined offered load fraction (e.g., 70% or 90% of maximum throughput). Soak tests detect buffer leaks, ECMP imbalance drift, PFC storm initiation, and long-tail error accumulation not visible in short-duration tests.

Table 10: Training-Specific Terms

9. Inference-Specific Terms

The following terms are specific to AI inference serving workload benchmarking and are used normatively in [I-D.calabria-bmwg-ai-fabric-inference-bench].

Term	Definition
TTFT	Time to First Token. The elapsed time from receipt of an inference request by the serving system to emission of the first output token. Encompasses prompt processing (prefill), KV cache generation, optional KV cache transfer (in disaggregated architectures), and the initial decode step. Interactive serving target: TTFT < 500 ms at P99.

ITL	Inter-Token Latency. The elapsed time between successive output tokens during the autoregressive decode phase. Measured at P50, P95, P99, and P99.9 to characterize tail latency behavior. Interactive serving target: ITL < 50 ms at P99.
TPS	Tokens Per Second. Aggregate throughput of the inference serving system, measured as the total number of output tokens generated per second across all concurrent requests. Reported separately for input-side (prefill) TPS and output-side (decode) TPS.
KV Cache	Key-Value Cache. The intermediate attention state (key and value projection matrices from multi-head attention layers) computed during the prefill phase and reused during each decode step to avoid redundant recomputation. KV cache size scales with: layers × attention_heads × head_dim × sequence_length × precision. The attention head configuration *MUST* be reported in all benchmark results.
Prefill Phase	The compute-bound phase of LLM inference in which the entire input prompt is processed in parallel to generate the KV cache and the first output token. Characterized by high arithmetic intensity (200400 ops/byte), high accelerator utilization (9095%), and large activation tensors. Prefill latency dominates TTFT for long prompts.
Decode Phase	The memory-bandwidth-bound phase of LLM inference in which output tokens are generated autoregressively, one token per forward pass, by reading the KV cache. Characterized by low arithmetic intensity (6080 ops/byte), lower accelerator utilization (2040%), and memory-bandwidth-limited KV cache reads. Decode throughput limits TPS.
Disaggregated Serving	An inference serving architecture in which the prefill phase and decode phase are

	executed on physically separate groups of accelerators (workers), connected by a network fabric. Allows independent scaling of prefill and decode resources (xPyD) but introduces KV cache transfer as a fabric-critical data movement.
xPyD Ratio	The allocation ratio of x prefill workers to y decode workers in a disaggregated serving cluster. Example: 3P9D denotes 3 prefill nodes and 9 decode nodes. The optimal xPyD ratio depends on model size, prompt/output length distributions, and TTFT/ITL SLO targets.
Continuous Batching	A dynamic inference scheduling technique that inserts new requests into an active decode batch as slots become available (without waiting for the current batch to complete), improving accelerator utilization compared to static batching. Generates variable batch sizes that affect fabric traffic burstiness.
PagedAttention	A KV cache memory management technique storing attention keys and values in fixed-size, non-contiguous virtual pages (typically 1664 KB), inspired by OS virtual memory management. Reduces memory fragmentation and enables efficient KV cache sharing across requests with common prefixes.
Prefix Caching	Reuse of previously computed KV cache segments for inference requests sharing a common prompt prefix (e.g., a fixed system prompt), eliminating redundant prefill computation. Prefix cache hit rate is a secondary KPI for inference serving efficiency.
Normal Dispatch	An AllToAll MoE dispatch communication mode optimized for the prefill phase. Payload sizes are variable (depending on token-to-expert routing), generating dynamic tensor shapes incompatible with static graph capture. Maximizes throughput for large batches at the cost of higher per-dispatch

	latency.
Low-Latency Dispatch	An AllToAll MoE dispatch communication mode optimized for the decode phase. Payload sizes are padded to fixed maximum dimensions (compatible with static graph capture), enabling lower kernel-launch overhead at the cost of slight bandwidth inefficiency. Target: < 200 s per dispatch round trip.
Expert Choice Routing	A token routing strategy in which experts select which tokens to process, rather than tokens selecting experts. Each expert accepts its top-C tokens by affinity score, producing perfect load balance but non-uniform AllToAll message sizes across EP ranks.
Auxiliary Loss Top-k	A top-k routing variant that adds a load-balancing auxiliary loss during training to encourage uniform token distribution across experts. Produces near-uniform AllToAll traffic in inference and reduces hot-spot risk on the fabric.
Top-k with Token Drop	A top-k routing variant in which tokens destined for overloaded experts are dropped or redirected to a fallback. Reduces worst-case dispatch traffic volume at the cost of model output quality under load.
T_dispatch	The dispatch payload per accelerator per MoE layer, computed as: T_dispatch = (B * k * H_model * P_bytes) / N where B = batch size (tokens), k = top-k routing count, H_model = hidden dimension, P_bytes = bytes per element (BF16=2, FP8=1), N = EP group size. Used as the canonical traffic volume parameter in the MoE test matrix (see Section 7.1 of the companion inference benchmarking draft).
SLO	Service Level Objective. A quantitative target for an inference serving KPI. AI inference SLOs typically specify maximum TTFT (e.g., < 500 ms P99) and maximum ITL (e.g., < 50 ms P99) under a specified

	request arrival rate.
Speculative Decoding	An inference acceleration technique using a small draft model to generate candidate token sequences verified in parallel by the target model. Reduces effective ITL but generates bursty, variable-length KV cache traffic; noted as a future benchmarking area not fully specified in the current companion documents.
S_KV	The total size in bytes of the KV cache state generated by a single inference request across all transformer layers and all context tokens, computed as: S_KV = 2 x L x H_kv x D x C x P_bytes. Where: L = number of transformer layers; H_kv = number of KV attention heads per layer (H_kv <= H_total for GQA/MQA); D = per-head key/value dimension (head_dim), typically model_dim / H_total; C = context length in tokens (prompt + generated tokens); P_bytes = precision in bytes per element (FP16/BF16 = 2, FP8/INT8 = 1); Factor 2 accounts for both K and V tensors, each of shape [H_kv, D] per layer per token.

Table 11: Inference-Specific Terms

See Section 7.1 of [I-D.calabria-bmwg-ai-fabric-inference-bench] for the MoE test matrix referenced by T_dispatch above.

9.1. Inference Phase Characteristics

Phase	Compute Bound	Arithmetic Intensity	Accelerator Util.	Primary KPI
Prefill	Yes	200400 ops/byte	9095%	TTFT
Decode	No (memory BW bound)	6080 ops/ byte	2040%	ITL, TPS

Table 12: Inference Phase Characteristics

10. KPI Classification Terms

The following terms define the three-tier KPI taxonomy used across both companion methodology documents.

Term	Definition
Primary KPI	A top-level performance indicator directly representing end-user experience or training efficiency. In training: JCT Ratio and BusBW. In inference: TTFT and ITL. Primary KPIs are the principal reporting metric and the basis for comparative benchmarking across DUT implementations.
Secondary KPI	A fabric-level performance indicator providing mechanistic explanation for primary KPI values. Examples: collective operation throughput (BusBW), KV cache transfer goodput, AllToAll dispatch latency, ECMP imbalance (MMR), and link utilization. Secondary KPIs enable root-cause analysis of Primary KPI deviations.
Fabric Health Indicator (FHI)	An operational metric characterizing fabric stability and anomaly conditions rather than peak performance. FHIs include: PFC event rate, PFC storm occurrence, ECN marking ratio, packet loss rate, buffer occupancy (P99), and retransmission rate. FHIs *SHOULD* be continuously monitored and reported throughout all test categories.
Goodput	The application-useful data delivered per unit Benchmark reports MUST use the qualified term to avoid ambiguity. *Fabric_Goodput:* RDMA message payload bytes successfully delivered per unit time at the DUT boundary, excluding transport headers, framing overhead, padding, and retransmitted bytes. This is the numerator quantity in KV_xfer_bandwidth and EP_alltoall_bandwidth. Units: GB/s or Gbps; reports MUST state which. *Inference_Goodput:* Output tokens successfully delivered per unit time, counting only requests that complete without preemption, eviction, or error. Corresponds to TPS_output over successfully completed requests only. Units: tokens/second. The two planes MUST NOT be conflated. KV_BW

	measures Fabric_Goodput; it does not measure Inference_Goodput.	
Zero Packet Loss	time, excluding retransmitted packets, protocol. A test acceptance criterion requiring that no packets are dropped by the DUT during the measurement interval. For RoCEv2 and UET transports, zero packet loss is the target operating condition. The binary search procedure in the companion methodology documents determines the maximum offered load satisfying this criterion.	

Table 13: KPI Classification Terms

10.1. KPI Tier Summary

Tier	Training Examples	Inference Examples	Purpose
Primary KPI	JCT Ratio, BusBW	TTFT, ITL, TPS	Direct end-user experience / business impact
Secondary KPI	AllReduce BusBW, MMR, Link Utilization	AllToAll dispatch latency, KV transfer goodput	Root cause analysis of Primary KPI deviations
Fabric Health Indicator (FHI)	PFC events, ECN ratio, packet loss, buffer P99, retx rate	PFC events, ECN ratio, packet loss, buffer P99	Ongoing fabric stability and anomaly detection

Table 14: KPI Tier Summary

11. Referenced Standards Abbreviations

The following abbreviations refer to normative and informative IETF documents referenced throughout this document and the companion methodology documents.

Reference	Definition
RFC 1242	"Benchmarking Terminology for Network Interconnect Devices" (Bradner, 1991). Defines foundational benchmarking terms (throughput, latency, frame loss rate, back-to-back frames). The baseline terminology reference for BMWG work. Where terms in this document overlap with RFC 1242 definitions, the AI fabric context definitions herein take precedence.
RFC 2544	"Benchmarking Methodology for Network Interconnect Devices" (Bradner & McQuaid, 1999). Defines test methodologies for throughput, latency, frame loss rate, and back-to-back measurements. The AI fabric methodology documents extend RFC 2544 procedures for AI-specific traffic patterns and test durations.
RFC 8238	"Data Center Benchmarking Terminology" (Bitar et al., 2017). Extends RFC 1242 with data center-relevant terms including forwarding table scaling, congestion, and VM/SDN. Incast, ECN, and buffer occupancy concepts in this document align with RFC 8238 definitions.
RFC 8239	"Data Center Benchmarking Methodology" (Bitar et al., 2017). Defines test methodologies for data center network functions including incast, ECN marking, and lossless behavior. The AI fabric companion methodology documents extend RFC 8239 for distributed AI collective traffic patterns.
RFC 2119 / RFC 8174	"Key words for use in RFCs to Indicate Requirement Levels" (Bradner, 1997; Leiba, 2017). Define the normative requirement language: MUST, MUST NOT, REQUIRED, SHALL, SHALL NOT, SHOULD, SHOULD NOT, RECOMMENDED, MAY, and OPTIONAL. RFC 8174 clarifies that these terms are normative only when in uppercase; lowercase uses are not normative.

Table 15: Referenced Standards Abbreviations

Acknowledgments

This work has benefited from the discussions that occurred during IPPM&BMWG joint meeting and on BMWG mailing list. Thanks Carsten Rossenhoevel, Mohamed Boucadair , Sowjanya Reddy for valuable review and comments.

References

Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/rfc/rfc2119>.

[RFC2544] Bradner, S. and J. McQuaid, "Benchmarking Methodology for Network Interconnect Devices", RFC 2544, DOI 10.17487/RFC2544, March 1999, <https://www.rfc-editor.org/rfc/rfc2544>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/rfc/rfc8174>.

Informative References

[I-D.calabria-bmwg-ai-fabric-inference-bench] Calabria, F., Pignataro, C., Wu, Q., and G. Fioccola, "Benchmarking Methodology for AI Inference Serving Network Fabrics", Work in Progress, Internet-Draft, draft-calabria-bmwg-ai-fabric-inference-bench-00, 26 February 2026, <https://datatracker.ietf.org/doc/html/draft-calabria-bmwg-ai-fabric-inference-bench-00>.

[I-D.calabria-bmwg-ai-fabric-terminology] Calabria, F., Pignataro, C., Wu, Q., and G. Fioccola, "Benchmarking Terminology for AI Network Fabrics", Work in Progress, Internet-Draft, draft-calabria-bmwg-ai-fabric-terminology-00, 26 February 2026, <https://datatracker.ietf.org/doc/html/draft-calabria-bmwg-ai-fabric-terminology-00>.

[I-D.calabria-bmwg-ai-fabric-training-bench]
Calabria, F., Pignataro, C., Wu, Q., and G. Fioccola, "Benchmarking Methodology for AI Training Network Fabrics", Work in Progress, Internet-Draft, draft-calabria-bmwg-ai-fabric-training-bench-00, 26 February 2026, <https://datatracker.ietf.org/doc/html/draft-calabria-bmwg-ai-fabric-training-bench-00>.

[IBTA-ROCE]
InfiniBand Trade Association, "InfiniBand Architecture Specification, Annex 16: RoCE", 2010, <https://www.infinibandta.org>.

[RFC1242] Bradner, S., "Benchmarking Terminology for Network Interconnection Devices", RFC 1242, DOI 10.17487/RFC1242, July 1991, <https://www.rfc-editor.org/rfc/rfc1242>.

[RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, DOI 10.17487/RFC3168, September 2001, <https://www.rfc-editor.org/rfc/rfc3168>.

[RFC8238] Avramov, L. and J. Rapp, "Data Center Benchmarking Terminology", RFC 8238, DOI 10.17487/RFC8238, August 2017, <https://www.rfc-editor.org/rfc/rfc8238>.

[RFC8239] Avramov, L. and J. Rapp, "Data Center Benchmarking Methodology", RFC 8239, DOI 10.17487/RFC8239, August 2017, <https://www.rfc-editor.org/rfc/rfc8239>.

[UEC-SPEC-1.0]
Ultra Ethernet Consortium, "Ultra Ethernet Specification 1.0", 2024, <https://ultraethernet.org>.

Appendix A: Term Cross-Reference to Companion Documents

The following table identifies which terms from this document are used in each companion methodology document.

Term Category	Used in Training Bench	Used in Inference Bench
General Benchmarking Terms (§2)	All terms	All terms
Collective	AllReduce,	AllToAll, BusBW

Communication (§3)	AllGather, ReduceScatter, AllToAll, BusBW, CCL, Ring Algorithm, BSP, SPMD		
Parallelism Strategies (§4)	DP, TP, PP, EP, MoE, ZeRO	EP, MoE, DP Attention	
RDMA / RoCEv2 (§5.1)	RDMA, RoCEv2, QP, RC mode, RDMA Verb	RDMA, RoCEv2, QP, RC mode, GIN, KVCXL	
UET Terms (§5.2)	UET, PDC, ROD, RUD, RUDI, UUD, LLR, Packet Trimming, CBFC, UEC Profile, Entropy Value	UET, RUD, GIN	
Congestion Control (§6)	PFC, PFC Storm, PFC Deadlock, ECN, DCQCN, ECN Marking Ratio, Incast, Incast Ratio, Packet Spray, DLB/ Flowlet, ECMP, MMR	PFC, ECN, DCQCN, Incast, Packet Spray, ECMP	
Fabric Topology (§7)	Clos, Rail-Optimized, Bisection BW, Oversubscription, ToR, Spine, NIC, Buffer Occupancy, Zero-Impact Failover, Link Utilization	Clos, Bisection BW, ToR, NIC, Buffer Occupancy, Link Utilization	
Training-Specific (§8)	JCT, Roofline JCT, JCT Ratio, Gradient Sync, Step Time, Soak Test	Soak Test	
Inference-Specific (§9)	—	TTFT, ITL, TPS, KV Cache, Prefill, Decode, Disaggregated Serving, xPyD,	

			Continuous Batching, PagedAttention, Prefix Caching, Normal/Low-Latency Dispatch, SLO
KPI Classification (§10)	Primary KPI (JCT Ratio, BusBW), Secondary KPI, FHI, Goodput, Zero Packet Loss	Primary KPI (TTFT, ITL), Secondary KPI, FHI, Goodput, Zero Packet Loss	

Table 16: Term Cross-Reference to Companion Documents

Appendix B: Term Taxonomy Summary

The following table provides a concise summary of all defined terms organized by category, with the section reference for the full definition.

Section	Term(s)	Category
2	DUT, SUT, RT, JFI, Offered Load, Trial Duration, Warmup Period, Binary Search, Percentile Latency, AI Fabric	General Benchmarking
3	Collective Operation, AllReduce, AllGather, ReduceScatter, AllToAll, Ring Algorithm, BusBW, CCL, SPMD, BSP	Collective Communication
4	Data Parallelism, Tensor Parallelism, Pipeline Parallelism, Expert Parallelism, MoE, DP Attention, ZeRO	Parallelism Strategies
5.1	RDMA, RoCEv2, QP, Reliable Connected (RC), RDMA Verb, UET, PDC, ROD	Transport — RDMA / RoCEv2
5.2	RUD, RUDI, UUD, UEC Profile, LLR, Packet Trimming, CBFC,	Transport — UET

	Entropy Value, GIN, KVCXL	
6	PFC, PFC Storm, PFC Deadlock, ECN, DCQCN, ECN Marking Ratio, Incast, Incast Ratio, Packet Spray, DLB/Flowlet, ECMP, MMR	Congestion Control
7	Clos/Fat-Tree, Rail-Optimized, Bisection Bandwidth, Oversubscription Ratio, ToR Switch, Spine/Superspine, NIC, Buffer Occupancy, Zero-Impact Failover, Link Utilization	Fabric Topology
8	JCT, Roofline JCT, JCT Ratio, Gradient Synchronization, Step Time, Soak Test	Training-Specific
9	TTFT, ITL, TPS, KV Cache, Prefill Phase, Decode Phase, Disaggregated Serving, xPyD Ratio, Continuous Batching, PagedAttention, Prefix Caching, Normal Dispatch, Low-Latency Dispatch, SLO, Speculative Decoding	Inference-Specific
10	Primary KPI, Secondary KPI, Fabric Health Indicator, Goodput, Zero Packet Loss	KPI Classification
11	RFC 1242, RFC 2544, RFC 8238, RFC 8239, RFC 2119/8174	Referenced Standards

Table 17: Complete Term Taxonomy

Authors' Addresses

Fernando Calabria
Cisco
Email: fcalabri@cisco.com

Carlos Pignataro
Blue Fern Consulting
Email: carlos@bluefern.consulting

Qin Wu
Huawei
Email: bill.wu@huawei.com

Giuseppe Fioccola
Huawei
Email: giuseppe.fioccola@huawei.com