

AI Fabric Benchmarking Methodology

Three Companion Internet-Drafts: Terminology · Training · Inference

Fernando Calabria (Cisco Systems)

Carlos Pignataro (Blue Fern Consulting)

Giuseppe Fioccola (Huawei)

Qin Wu

`draft-calabria-bmwg-ai-fabric-terminology-00`

`draft-calabria-bmwg-ai-fabric-training-bench-00`

`draft-Calabria-bmwg-ai-fabric-inference-bench-00`

Motivation, Document Suite & Key Metrics

The Gap in Existing BMWG Coverage

- AI/ML traffic patterns absent from RFC 2544 / RFC 8239 scope
- Training (AllReduce/AllGather) & Inference (disaggregated prefill/decode, MoE KV cache) add novel traffic patterns
- No vendor-neutral lab methodology covers RoCEv2/RDMA PFC, ECN, DCQCN, KPI

AI WORKLOAD CLASS EXAMPLES

Class	Topology	Primary Traffic
AI Training	8-8192 GPU, rail-optimised	AllReduce / AllGather
AI Inference	Disagg. Prefill + Decode	KV cache RDMA, MoE routing
Fine-Tuning (LoRA)	Smaller GPU cluster	Gradient all-reduce
Embedding Infer.	Single-node/small cluster	Low-latency P2P messaging

Three Companion Documents

draft-calabria-pignataro-bmwg-ai-fabric-terms-00

Terminology

Shared vocabulary: accelerator topologies (spine-leaf, rail-optimised), collective primitives, RDMA transport semantics, congestion mechanisms (PFC, ECN, DCQCN), inference constructs (KV cache, MoE routing).

draft-calabria-pignataro-bmwg-ai-fabric-training-bench-00

Training Benchmarks

Test procedures for bulk-synchronous collective ops over RoCEv2/UET fabrics. Covers LLM pre-training, fine-tuning, gradient sync. Scenarios: AllReduce ring/tree, AllGather, ReduceScatter, multi-job interference.

KPIs: CCT · JCT · PFC rate · ECN marks

draft-calabria-pignataro-bmwg-ai-fabric-inference-bench-00

Inference Benchmarks

Test procedures for disaggregated prefill/decode and MoE expert routing fabrics. Covers LLM serving, embedding inference, multi-tenant. Scenarios: KV cache RDMA transfer, MoE all-to-all, congestion under burst.

KPIs: TTFT · ITL · KV BW · ECN marks

Motivation, Document Suite & Key Metrics

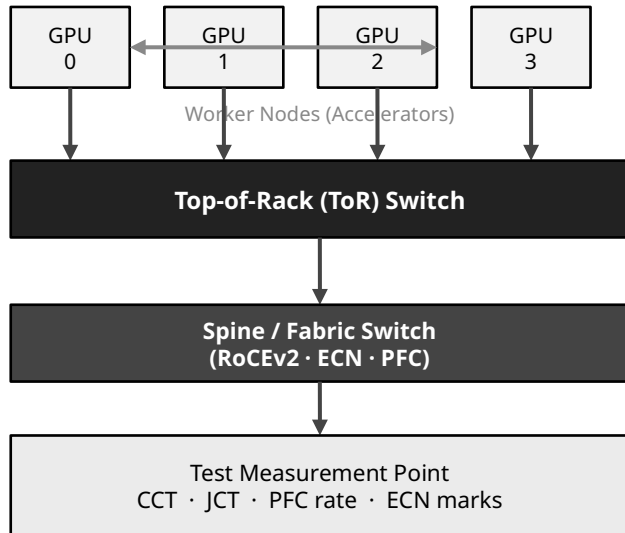
Key Performance Indicators (KPIs)

KPI	Workload	Definition
CCT	Training	Collective Completion Time — AllReduce/AllGather op duration across all participating ranks
JCT	Training	Job Completion Time — full training step including all collectives and compute phases
TTFT	Inference	Time to First Token — from request receipt to first output token (network latency contribution)
ITL	Inference	Inter-Token Latency — time between consecutive decode-phase token emissions
KV BW	Inference	KV-cache transfer bandwidth — sustained RDMA throughput during disaggregated prefill/decode
PFC / ECN	Both	Pause frame incidence (PFC) and congestion notification rate (ECN) — fabric health indicators

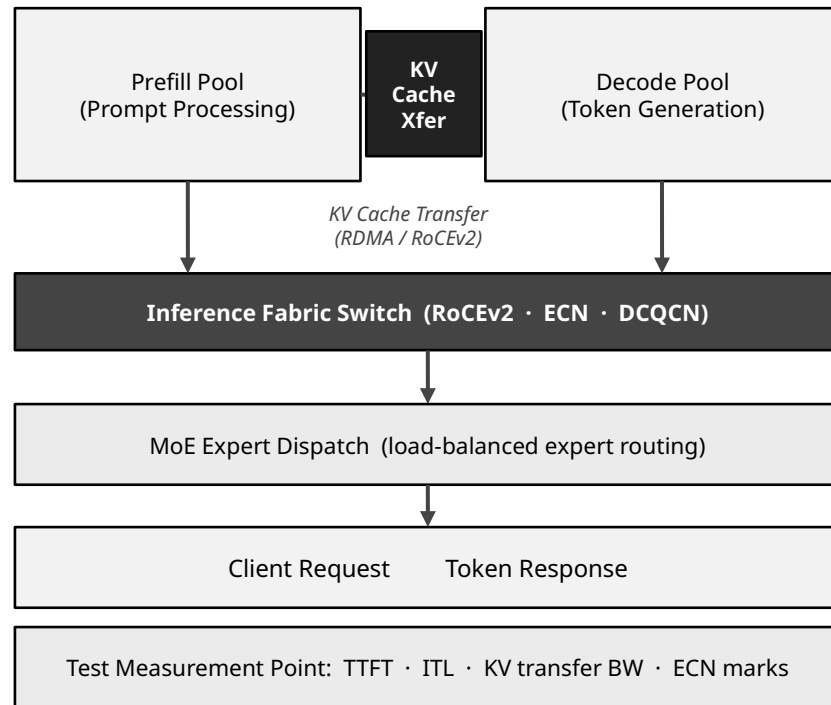
High-Level Architecture: Training vs. Inference Fabrics

TRAINING FABRIC (Bulk-Synchronous Collective)

AllReduce / AllGather (ring & tree topologies)

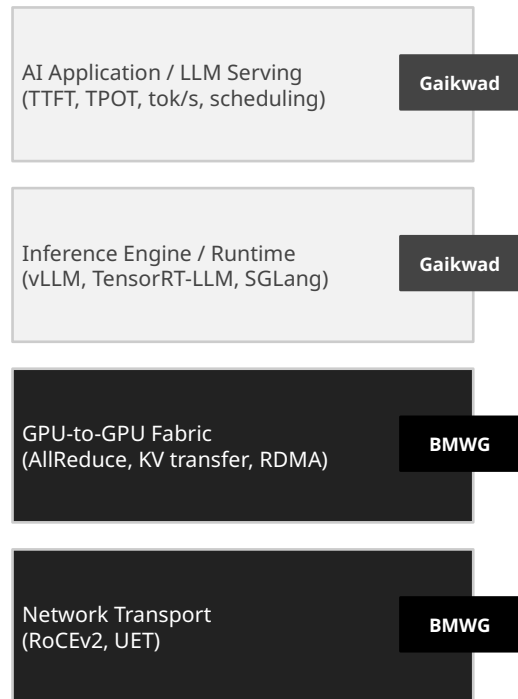


INFERENCE FABRIC (Disaggregated Prefill / Decode)



Scope Comparison vs. draft-gaikwad-llm-benchmarking-methodology-00

AI Stack — Where Each Draft Operates



↑ Complementary, not competitive
Different layers of the same AI system

Detailed Scope Comparison

Dimension	BMWG AI Fabric (This Work)	draft-gaikwad-llm-benchmarking-methodology
Primary question	How does the network fabric perform under AI workload traffic?	How does an LLM serving system perform under inference load?
System Under Test	Switches, links, RDMA adapters, congestion control engines	Model engine, application gateway, compound (RAG / agentic) system
OSI Layer focus	L2-L4 (data link, transport, congestion)	L7 (HTTP/gRPC token streaming, application-level latency)
Workload origin	Collective ops (AllReduce, AllGather), KV cache RDMA transfers	Token distributions: ShareGPT, HumanEval, synthetic patterns
KPI examples	CCT, JCT, PFC rate, ECN marks, KV transfer BW	TTFT (full e2e), ITL, tok/s, scheduling fairness, prefix cache hit rate
TTFT scope	Network fabric contribution to TTFT only	Full end-to-end TTFT (compute + memory + network)
AI Training workloads	Core — AllReduce, AllGather, ReduceScatter, multi-job interference	✗ Not addressed
MoE expert routing	Core — EP all-to-all latency & bandwidth (inference draft)	✗ Not addressed
IETF WG status	Individual submission targeting BMWG chartered WG adoption	Individual I-D — no WG affiliation, not endorsed by IETF

The two drafts are complementary — fabric benchmarks establish network substrate capability; LLM benchmarks establish application-layer efficiency given that sub

Scope Comparison vs. draft-gaikwad-llm-benchmarking-profiles-00

draft-gaikwad-llm-benchmarking-profiles-00 defines standard reference workload profiles (token length distributions, concurrency levels, request patterns) for reproducible LLM serving benchmarks. It provides the workload parameterization layer consumed by the Gaikwad methodology draft. It does NOT define any network test procedures.

Scope Comparison

Dimension	BMWG AI Fabric (Training + Inference)	draft-gaikwad-llm-benchmarking-profiles
Purpose	Define test procedures and KPIs for network fabric evaluation under AI workloads	Define standard reference workload profiles for reproducible LLM serving benchmarks
Workload specification	Collective op sizes (MB), RDMA message rates, congestion injection levels per test case	Token length distributions (input/output), request concurrency, arrival rate patterns (Poisson, closed-loop)
Network test procedures	Full test suite — CCT, JCT, TTFT network contribution, KV BW, MoE all-to-all	✗ No network test procedures — profiles specify app-layer load only
Traffic model	Bulk-sync collective ops (training); bursty RDMA point-to-point (inference)	HTTP/gRPC request arrival distributions driving LLM serving engines
Tokenization / LLM config	✗ LLM model parameters are DUT context (documented, not varied)	Core — tokenizer normalization, reproducible workload construction (§4.3)
Disaggregated prefill/decode	Network path between pools: KV cache RDMA latency and bandwidth	✗ Architecture context mentioned; not benchmarked at application layer
MoE expert parallelism	Core — EP all-to-all latency, bandwidth, fabric congestion behavior	✗ Not addressed in profiles or companion methodology draft
Coordination opportunity	Workload parameterization could bridge fabric load and application serving load	Profile concurrency levels could directly inform fabric congestion test matrix

Opportunity: coordinate with Gaikwad profile authors on workload parameterization that bridges network fabric load levels and application-layer serving load

BMWG Charter Alignment & Compliance

Charter Compliance Summary

BMWG Charter Principle	Requirement	This Work	Reference
Lab environment only	Measurements MUST be from controlled testbeds; no live network	All test procedures explicitly scoped to controlled laboratory environments	BMWG Charter §1; RFC 2544 §6.1
Vendor neutral	Benchmarks SHALL have universal applicability; no product names	No vendor/product names anywhere in three docs; technology-class general throughout	BMWG Charter §2; RFC 1242 §3
No acceptance thresholds	WG MUST NOT define pass/fail or minimum performance requirements	Explicitly stated in §1.4 (Scope) of both methodology documents	BMWG Charter §3
Terminology companion doc	Each methodology SHOULD be paired with a companion terminology document	Follows RFC 1242 + RFC 8239 precedent; companion terminology draft submitted jointly	RFC 1242; RFC 8239; RFC 8238
Forwarding plane scope	WG covers management, control, and forwarding plane benchmarks	Covers forwarding plane (fabric switch) and data-plane transport (RoCEv2, UET)	BMWG Charter §1
VNF / virtual infra scope	Scope extended to virtual network functions and supporting infrastructure	AI accelerator fabrics are the natural extension of VNF/cloud infrastructure benchmarking scope	BMWG Charter §2 (VNF)

Next Steps & Requested Actions

Requested from BMWG participants

1

Working Group Technical Review

Feedback requested on KPI definitions, test procedure design, traffic model parameterization, and DUT/SUT configuration requirements across all three documents.

Complementary scope — no conflict

2

Coordination with Gaikwad Authors

Authors plan to engage Madhava Gaikwad to align scope boundaries, especially around the TTFT network contribution measurement interface and workload parameterization bridge.

Individual submission → BMWG WG Item

3

Path to Working Group Adoption

Subject to WG interest, authors seek adoption under the BMWG charter. Three-document structure follows established BMWG practice (cf. RFC 8238/8239 precedent).