IP Performance Measurement
Internet-Draft
Intended status: Informational
Expires: 21 October 2026

N. Ye
W. Sun
Shanghai Jiao Tong University
D. Wang
J. Sun
China Mobile Research Institute

19 April 2026

Switching Efficiency: A Metric Framework for AI Data Center Networks
draft-ye-ippm-switching-efficiency-02

Abstract

This document specifies the Switching Efficiency Framework, a measurement methodology designed to evaluate network efficiency in AI Data Centers (AIDCs). Conventional network metrics, such as bandwidth utilization or network throughput, fail to directly link network activity to computational progress, as they cannot distinguish computationally effective data that directly advances neural network computing from the redundant traffic induced by both multi-hop forwarding and the algorithmic overhead of collective operations.

To address this, this document defines the Switching Efficiency Framework, a measurement methodology for evaluating AIDC network efficiency. The core metric, Switching Efficiency, quantifies the computationally effective data throughput delivered per unit of provisioned switching capacity. To facilitate precise diagnostic analysis, the framework further decomposes this core metric into three fine-grained factors: Data Efficiency, Routing Efficiency, and Port Utilization.

This framework provides metrics that can help operators identify communication bottlenecks and evaluate topology-traffic alignment.

About This Document

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

Status information for this document may be found at https://datatracker.ietf.org/doc/draft-ye-ippm-switching-efficiency/.

Discussion of this document takes place on the ippm Working Group mailing list (mailto:ippm@ietf.org), which is archived at https://mailarchive.ietf.org/arch/browse/ippm/. Subscribe at https://www.ietf.org/mailman/listinfo/ippm/.

Status of This Memo

Copyright Notice

Table of Contents

1. Introduction

In hyperscale AI Data Centers (AIDCs), network communication is often a performance bottleneck for training Large Language Models (LLMs). While diverse network topologies and communication algorithms (e.g., In-Network Computing) are being deployed, operators lack a common quantitative methodology to evaluate how effectively raw physical switching resources are converted into actual training progress.

Conventional performance metrics, such as bandwidth utilization or network throughput, are inadequate for this environment because they measure overall network activity rather than useful work. Specifically, they treat all transferred bytes equally, failing to isolate "computationally effective data" - the net data that directly advances neural network computing. For example, during an All-Reduce operation, large volumes of data are transferred across the fabric only to be discarded after mathematical reduction (algorithmic overhead). Similarly, when the physical topology fails to match the spatial distribution of the workload - such as forcing logically localized, high-volume traffic to cross the broader scale-out fabric - data must traverse an excessive number of forwarding hops (multi-hop overhead). Because traditional metrics conflate these redundancies with effective data delivery, operators cannot accurately quantify how well a specific network architecture aligns with its intended AI traffic patterns.

To bridge this gap, this document defines the Switching Efficiency Framework [SwitchingEfficiencyPaper], which relates the throughput of effective data to the aggregate switching capacity of the network through its core metric, Switching Efficiency (eta). This top-level metric is further decomposed into three diagnostic factors: Data Efficiency (gamma) evaluates the communication algorithm by indicating whether it delivers computationally effective data or generates redundant bytes; Routing Efficiency (delta) evaluates topology-traffic alignment by indicating whether the physical network provides direct paths or forces traffic into excessive multi-hop detours; and Port Utilization (theta) evaluates hardware resource allocation by indicating whether the provisioned switching capacity is actively utilized.

By defining these metrics, this document provides operators and telemetry systems with a common basis for evaluating AIDC network performance and diagnosing communication bottlenecks.

2. Conventions and Definitions

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.

3. Terminology

* *Computationally Effective Data (CED):* The aggregate application-payload volume yielded by a communication primitive and retained by one or more endpoints for subsequent neural network computation. CED excludes transport, network, and link-layer headers; padding; control traffic; unreduced intermediate data; and any bytes that are delivered but not retained as semantic input to the next computation phase.

  - For non-reduction operations (e.g., All-Gather or All-to-All dispatch), CED equals the aggregate newly received application-payload volume retained at the endpoints.

  - For reduction operations (e.g., All-Reduce, Reduce-Scatter, or All-to-All combine), CED equals only the final reduced output volume retained at the endpoints.

* *Switching Capacity:* The aggregate theoretical egress data forwarding rate of all packet-switching ports within the evaluated measurement domain. To reflect the heterogeneous hardware of modern AI Data Centers, this capacity includes all functional transit components within that domain, specifically:

  1. Standalone network switches (e.g., standard Ethernet or InfiniBand switches acting as Top-of-Rack, Leaf, or Spine).

  2. Embedded switching elements within a single compute chassis (e.g., NVSwitch interconnecting GPUs within a server).

  3. Forwarding ports residing natively on the compute accelerators (e.g., Google TPUs).

* *In-Network Computing (INC):* A network architecture paradigm where mathematical or logical operations (such as data reduction in collective communications) are executed within the network data

plane (e.g., by programmable switches) while data is in transit. In the context of AI Data Centers, INC is commonly deployed to offload collective communication reductions (e.g., performing arithmetic operations for All-Reduce directly on the switch), thereby eliminating the transmission of unreduced data and delivering only the reduced results to the endpoints.

* *Observation Window (T):* The common half-open time interval [t0, t1) over which all variables in this document are accumulated. The duration T equals t1 minus t0. All variables used to compute a reported metric instance MUST use the same observation window.

* *Measurement Domain:* The explicitly identified set of compute endpoints, forwarding elements, and forwarding ports included in a reported metric instance.

* *Measured Traffic Set:* The subset of packets, messages, or communication primitives attributed to the workload, job, tenant, or collective class under evaluation. The same traffic-selection rule MUST be applied consistently to V_CED, V_RECV, and V_FWD.

* *Byte Counting Rule:* The declared rule that specifies which bytes are counted and which are not counted when computing V_RECV and V_FWD. A report MUST state this rule explicitly and MUST apply the same rule consistently to V_RECV and V_FWD. V_CED is always counted using only the retained application data because it represents computation input that remains semantically useful to the application.

4. The Switching Efficiency Framework

This section defines the Switching Efficiency Framework. The detailed mathematical derivations supporting this framework are provided in [SwitchingEfficiencyPaper]. For operational measurement, all variables and derived metrics are defined relative to a single measurement domain, a single measured traffic set, a single byte counting rule, and a single observation window T. Two reported results are comparable only if these contextual parameters are the same, or if any differences are explicitly disclosed and normalized.

4.1. Core Variables

The framework relies on four primary operational metrics collected over the measurement window T:

* *V_CED (Total CED Volume):* The aggregate CED yielded by all measured communication primitives whose retained outputs are attributable to the observation window T according to the declared boundary-handling rule.

* *V_RECV (Total Received Volume):* The aggregate byte volume of the measured traffic set successfully accepted at the ingress of all measured compute endpoints during T. Each successful receipt counts once per endpoint receipt. If a payload is received multiple times because of retransmission or duplication, each actual receipt is included in V_RECV.

* *V_FWD (Total Forwarded Volume):* The aggregate byte volume of the measured traffic set emitted on the egress side of all measured forwarding ports during T. Each forwarding event counts once per egress transmission. Therefore, replicated copies, multicast fan-out, load-balancing replicas, retransmissions, and forwarding loops each increase V_FWD according to the number of observed egress transmissions.

* *C_TOTAL (Aggregate Switching Capacity):* The aggregate theoretical egress data forwarding rate of all packet-switching ports within the measurement domain. C_TOTAL equals the sum of the theoretical maximum unidirectional egress data rates of those ports.

4.2. Scope and Accounting Rules

To promote comparable results across implementations and experiments, the following accounting rules apply:

* All volumes defined in this document MUST be reported in bytes. All rates MUST be reported in bytes per second.

* A reported metric instance MUST identify its measurement domain, measured traffic set, byte counting rule, observation window, and boundary-handling rule for communication primitives that overlap the edges of T.

* Only traffic attributable to the measured traffic set MUST be included in V_CED, V_RECV, and V_FWD. Management traffic, storage traffic, unrelated tenant traffic, and background traffic outside the measured traffic set MUST be excluded unless the report explicitly declares that a mixed-traffic domain is being measured.

* The same traffic-selection rule MUST be used for V_CED, V_RECV, and V_FWD.

* The same byte counting rule MUST be used for V_RECV and V_FWD. A report MUST state whether additional non-payload bytes, such as encapsulation or framing overhead, are included. For maximum comparability, experiments that are compared against each other SHOULD use the same byte counting rule across all runs.

* For direct comparison, an implementation SHOULD use operation-aligned observation windows so that measured communication primitives are wholly contained within T. If communication primitives overlap T, the report MUST state whether overlapping primitives are excluded or attributed by a declared attribution point. The attribution point is the completion time of the communication primitive for V_CED, the endpoint receipt time for V_RECV, and the egress transmission time for V_FWD.

4.3. Core Metric: Switching Efficiency (eta)

Switching Efficiency (eta) is the top-level metric quantifying how effectively a network translates its raw physical capacity into computational progress. It is defined as the ratio of the CED throughput over observation window T to the aggregate switching capacity of the network.

$$\eta = \frac{V_{CED} / T}{C_{TOTAL}}$$

A high eta indicates that a large proportion of the network's provisioned hardware capacity is successfully contributing to the delivery of computationally effective data.

4.4. Fine-Grained Efficiency Factors

To enable diagnostic analysis and isolate specific performance bottlenecks, eta is mathematically decomposed into three diagnostic efficiency factors (eta = gamma * delta * theta):

4.4.1. Data Efficiency (gamma)

Data Efficiency evaluates the effectiveness of implementing the communication primitives. It specifies the ratio of Computationally Effective Data (V_CED) to the total received volume (V_RECV).

$$\gamma = \frac{V_{CED}}{V_{RECV}}$$

* *Diagnostic Focus:* Identifies redundant data delivered to endpoints. A value of gamma less than 1 indicates that endpoint ingress traffic contains bytes that do not survive as retained computation input, such as unreduced data, duplicated deliveries, or additional overhead included by the declared byte counting rule. Executing mathematical reductions within the network data plane via INC can improve gamma by reducing non-retained traffic delivered to the endpoints.

4.4.2. Routing Efficiency (delta)

Routing Efficiency quantifies the topological alignment between the physical network architecture and the AI workload traffic patterns.

$$delta = \frac{V\_RECV}{V\_FWD}$$

* *Diagnostic Focus:* Identifies forwarding overhead. In a lossless network with no duplicated in-network copies, delta equals the inverse of the volume-weighted average number of forwarding events incurred per received byte. A value of delta less than 1 indicates that traffic either traverses multiple forwarding stages or experiences extra forwarding caused by retransmission, replication, or looping behavior.

4.4.3. Port Utilization (theta)

Port Utilization measures the spatial and temporal engagement of the provisioned switching capacity.

$$theta = \frac{V\_FWD}{C\_TOTAL * T}$$

* *Diagnostic Focus:* Identifies underutilized switching capacity. A low theta indicates that the provisioned hardware (C_TOTAL) operates below its theoretical maximum data rate over the observation window T, due to either spatial traffic imbalance or temporal idleness.

5. Measurement Methodology

This section specifies the operational procedures for collecting the variables required to compute the efficiency metrics. Accurate measurement requires tight time synchronization (e.g., via the Precision Time Protocol (PTP) [IEEE1588]) across all network and compute endpoints, as well as an observation window T sufficiently large to dilute telemetry polling variance. A report claiming compliance with this specification MUST record the measurement domain, the measured traffic set, the byte counting rule, the observation window, the boundary-handling rule, and the estimated synchronization accuracy of the participating measurement points.

The four core variables span the network, endpoint, and application planes, and are collected as follows:

* *C_TOTAL (Aggregate Switching Capacity):* Derived from the topology inventory. It is computed by summing the theoretical maximum unidirectional egress data rates of all packet-switching ports within the declared measurement domain.

* *V_FWD (Total Forwarded Volume):* Collected from the network plane. Operators MUST extract aggregate egress byte counters from the measured forwarding ports, typically from switch ASIC counters or equivalent forwarding-plane telemetry. Only traffic matching the declared measured traffic set is included. If the same payload is transmitted multiple times on different egress ports, or retransmitted on the same port, each egress transmission counts separately in V_FWD. If communication primitives overlap T, the implementation MUST apply the declared attribution point for V_FWD consistently. Counter wrap, reset, discontinuity, or sampling loss MUST be corrected if possible; otherwise, the affected observation window MUST be reported as invalid or qualified accordingly.

* *V_RECV (Total Received Volume):* Collected from the endpoint plane. Operators MUST extract aggregate ingress byte counters from the host interfaces or accelerators attached to the measured endpoints. Only traffic matching the declared measured traffic set is included. V_RECV counts successful endpoint receipts; bytes dropped before endpoint ingress are excluded. Duplicate deliveries and retransmissions that are actually received at the endpoint each contribute to V_RECV. If communication primitives overlap T, the implementation MUST apply the declared attribution point for V_RECV consistently.

* *V_CED (Total CED Volume):* Collected from the application plane. The implementation MUST count only the retained semantic outputs of measured communication primitives. To avoid the high overhead of parsing verbose logs, operators SHOULD utilize lightweight collection mechanisms such as host-side telemetry agents, eBPF hooks dynamically attached to collective communication APIs, or native metrics endpoints exposed by standard communication libraries (e.g., MPI or vendor-specific equivalents such as NCCL/RCCL). If communication primitives overlap T, the implementation MUST apply the declared attribution point for V_CED consistently, especially when a primitive starts before T or completes after T.

5.1. Reporting Requirements

A comparable measurement report produced using this framework MUST include at least the following items:

* The start time t0, end time t1, and duration T of the observation window.

* A description of the measurement domain, including the set of measured endpoints and forwarding elements.

* A description of the measured traffic set, including any job identifiers, tenant filters, flow selectors, or collective-operation selectors used to isolate the traffic.

* The byte counting rule used for V_RECV and V_FWD, including whether additional non-payload bytes are included.

* The boundary-handling rule used when communication primitives overlap the boundaries of T, including whether overlapping primitives are excluded or attributed by completion time, receipt time, and egress transmission time, respectively.

* The time-synchronization method and the maximum estimated clock error across measurement points.

* The polling or export interval for counters, together with the treatment of counter reset, wrap, or missing samples.

* The set of ports included in C_TOTAL and the theoretical maximum unidirectional egress data rate used for each port or port class.

* The final reported values of V_CED, V_RECV, V_FWD, C_TOTAL, eta, gamma, delta, and theta.

5.2. Uncertainty and Bias

The following effects can materially change the measured values and therefore MUST be disclosed whenever they are present:

* Imperfect isolation of the measured traffic set from unrelated background traffic.

* Incomplete visibility into replicated traffic, dropped packets, or endpoint duplicates.

* Clock error that is large relative to the duration of the communication primitives being measured.

* Counter sampling intervals that are too coarse relative to burst duration, or counter discontinuities caused by reset, wrap, or telemetry loss.

* Application-level instrumentation that cannot unambiguously determine whether partially completed primitives contribute to V_CED.

Two reported results MUST NOT be treated as directly comparable unless the reporting items above are either the same or are normalized to an equivalent basis by the experimenter.

6. Security Considerations

The operational deployment of this measurement framework raises the following security and privacy considerations:

* *Data Confidentiality:* Collecting V_CED and V_RECV can inadvertently expose proprietary AI workload characteristics (e.g., model architecture or training strategies). Telemetry data MUST be transported over encrypted channels, such as Transport Layer Security (TLS) [RFC8446] or Internet Protocol Security (IPsec) [RFC4301], and securely stored.

* *Measurement Integrity:* Falsifying the underlying counters (V_FWD, V_RECV, V_CED) will manipulate the calculated efficiency metrics. Authentication and authorization MUST be enforced for all telemetry endpoints to prevent data poisoning.

7. IANA Considerations

This document has no IANA actions.

8. References

8.1. 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>.

[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>.

8.2. Informative References

[IEEE1588] "IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems", IEEE Std 1588-2019, November 2019.

[RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, DOI 10.17487/RFC4301, December 2005, <https://www.rfc-editor.org/rfc/rfc4301>.

[RFC8446] Rescorla, E., "The Transport Layer Security (TLS) Protocol Version 1.3", RFC 8446, DOI 10.17487/RFC8446, August 2018, <https://www.rfc-editor.org/rfc/rfc8446>.

[SwitchingEfficiencyPaper] Ye, N., Zhu, J., Chen, B., Wang, D., Sun, J., Sun, W., and W. Hu, "Switching Efficiency: A Novel Framework for Dissecting AI Data Center Network Efficiency", arXiv 2604.14690, DOI 10.48550/arXiv.2604.14690, April 2026, <https://doi.org/10.48550/arXiv.2604.14690>.

Acknowledgments

We are grateful for the valuable discussions and input from the community. We also acknowledge support from NSFC.

Authors' Addresses

Niangen Ye
Shanghai Jiao Tong University
China
Email: yng2020@sjtu.edu.cn

Weiqiang Sun
Shanghai Jiao Tong University
China

Email: sunwq@sjtu.edu.cn

Dong Wang
China Mobile Research Institute
Department of Fundamental Network Technology
Beijing
China
Email: wangdongyjy@chinamobile.com

Jiang Sun
China Mobile Research Institute
Department of Fundamental Network Technology
Beijing
China
Email: sunjiang@chinamobile.com