MPLS Japan 2023

HPCC と SRv6 Path Tracing適用性について

26 October 2023

Miya Kohno, Cisco Systems (mkohno@cisco.com)

Abstract

ML/DL workloadを支えるインフラは、ロスレスであることが求められます。しかし、高速・低遅延・高効率を実現しながらロスレスを実現するのは簡単ではなく、高度な輻輳制御が必要になります。（”Fully Scheduled Fabric”によりほぼ完全なロスレスを実現可能ですが、一方、この方式は収容効率を抑えるため、全ての箇所に適用させるのは合理的とは言えません。）

現時点で一般的なのはPFC + ECNによる輻輳制御 (DCQCN: Data Center Quantized Congestion Notification) ですが、この方式では、必要以上にレートを落としてしまったり、逆にレートを落とすタイミングが遅くてバッファオーヴァフローを起こしてしまう、という問題が報告されています。

この問題を解決するために提案されているのがHPCC（High Precision Congestion Control）です。HPCCは、精度の高いリンク負荷情報を得るためにテレメトリ情報を活用し、トラフィックを正確に制御します。HPCCにより、輻輳通知の遅延や輻輳通知への過剰反応といった課題に対処し、高速・低遅延・高効率を実現しながらロスレスを実現することができます。

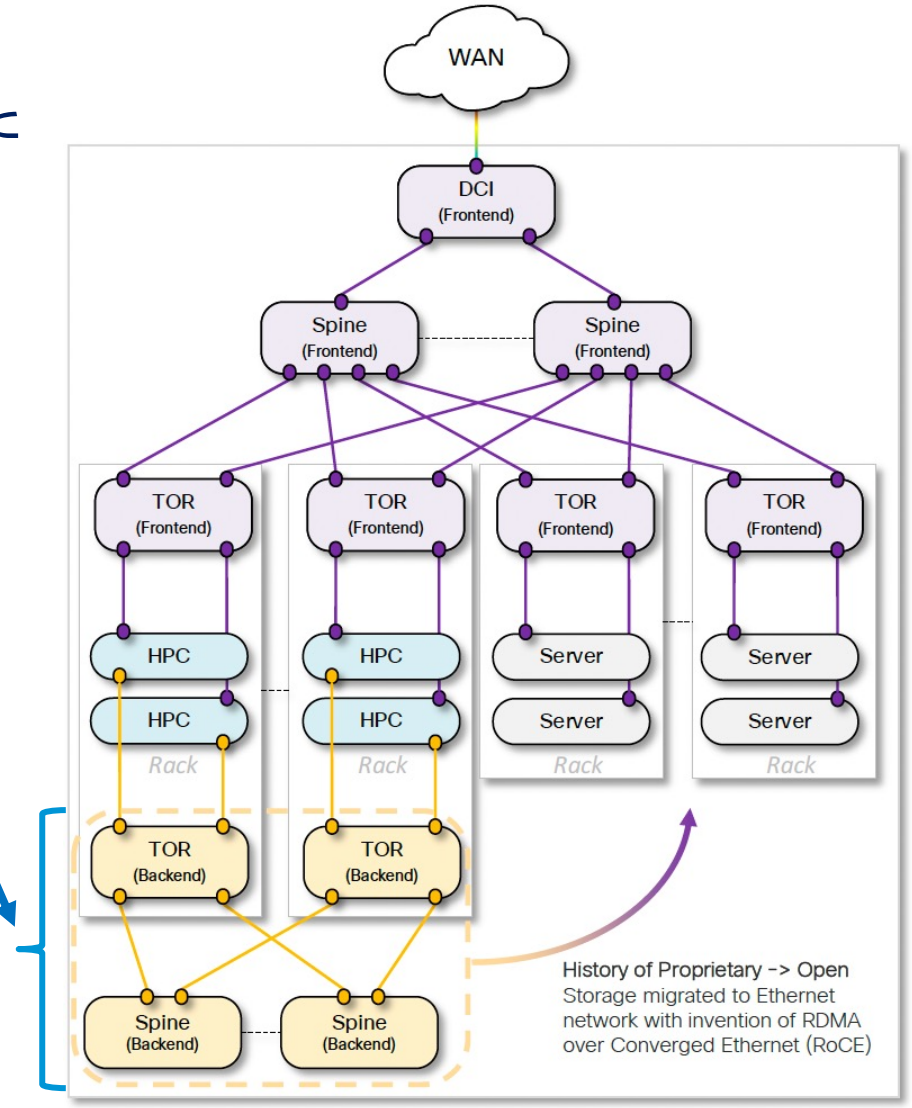
現在は、HPCCのためのテレメトリ情報としてINT（In Network Telemetry）、IOAM（In-situ OAM）などの使用が提唱されていますが、本セッションでは、SRv6 Path TracingによるHPCC適用性とそのメリット、そして、より効率性の高いロスレスファブリックの実現について考察します。

Converged Interconnect へ

- ストレージやGPUを相互接続するためのネットワークは”Backend Interconnect”と呼ばれ、求められるトラフィック要件も異なるため、特殊技術が使われることが殆どだった

- ML/DL 発展による変化
  - 大規模化
    -- テラビット帯域
    -- 数億のトランザクションレート
  - 多様化
    -- AIトレーニング：
    長時間の「エレファント」フロー、バッチモードが特徴
    -- AI推論：
    LLMでは、複数のGPUで推論を実行する
    遅延に敏感

大規模化・多様化に、スケールメリットを活かして対応するために、オープン技術の採用が求められる



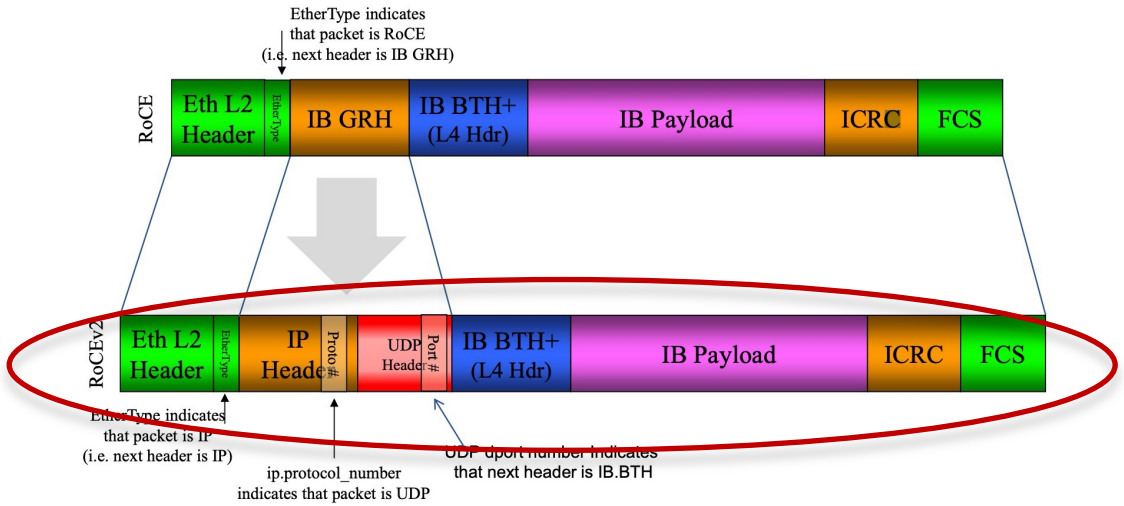
“Evolved Networking: the AI/ML Challege”, Rakesh Chopra, OCP Global Summit 2022

Ethernet で Loss-less, Non-blocking を実現するために

ROCEv2 + DCQCN (DC Quantified Congestion Notification)

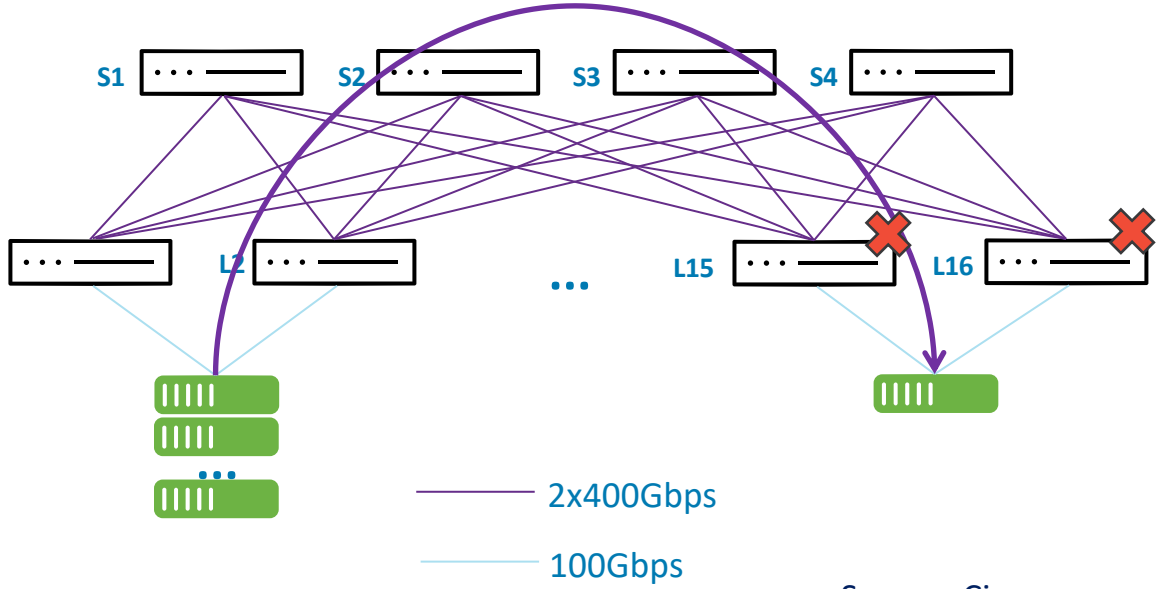
- IP Explicit Congestion Notification (RFC 3168)
- Priority-based Flow Control (802.1Qbb)

- ROCEv2 [UDP/IP + IB BTH + IB Payload]



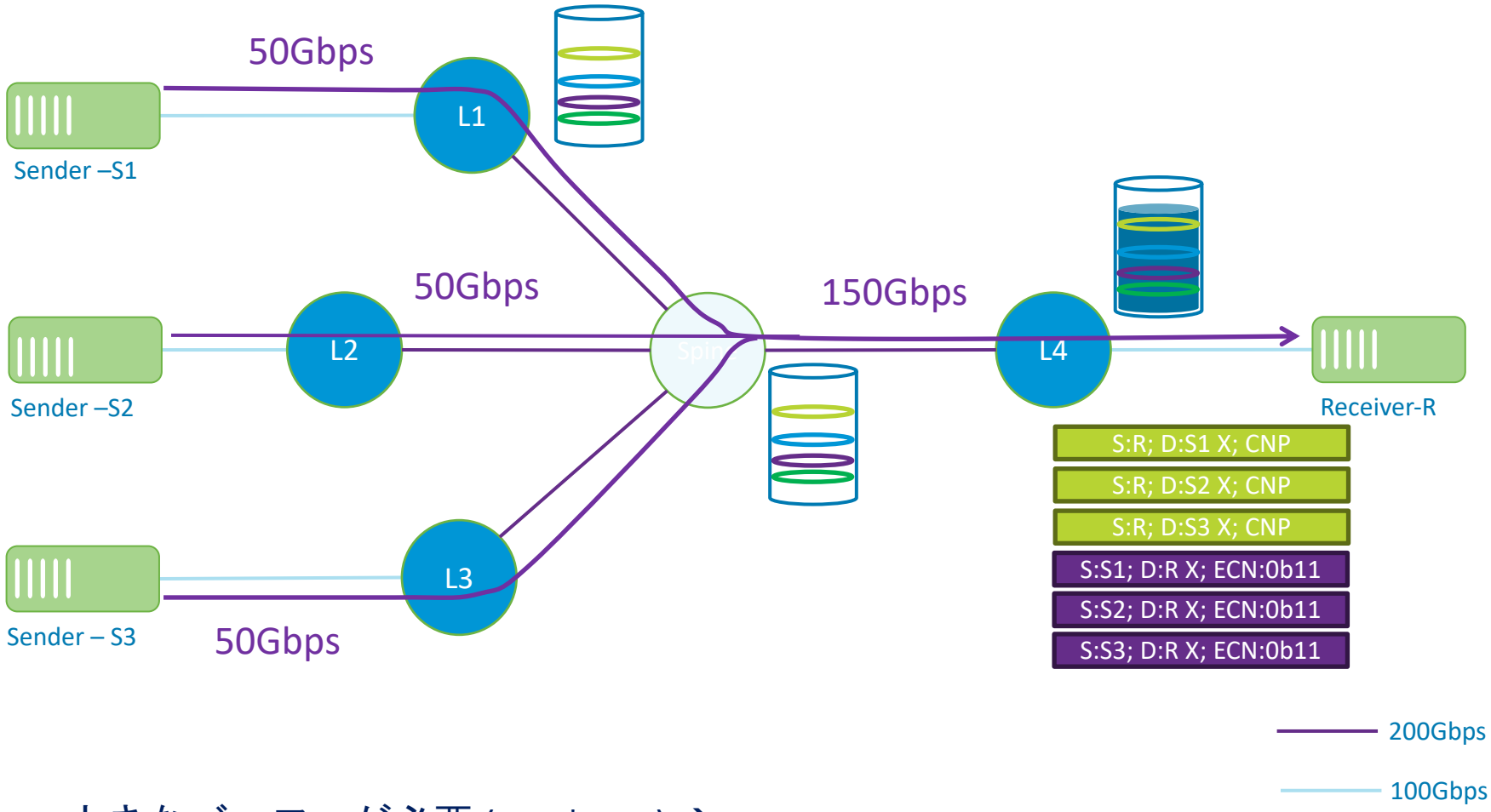
IB: Infiniband, BTH: Base Transport Header
https://bigstep.com/blog/modern-storage-technologies-in-2020

- 輻輳は起こり得る (例) フローの合計 = 300Gbps



Source: Cisco

DCQCN - Priority Flow Control In Action With RoCEv2



- 大きなバッファが必要 (BW * RTT) → Latency
- 過剰にレートを抑えてしまうリスク

Diagram : Cisco

DCQCN の問題点

(1) PFC requires excessive buffering for lossless transport
PFCはロスレスのために過剰なバッファを必要とする

(2) Victim flows, congestion trees, PFC storms, and deadlocks
犠牲者フロー、輻輳ツリー、PFCストーム、デッドロック

(3) Go-back-N retransmission
Go-back-N再送

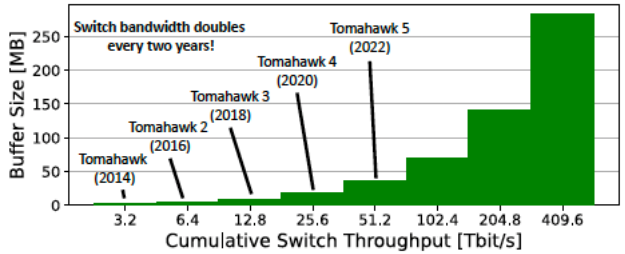
(4) Congestion control and colocation with other traffic
輻輳制御方式と他のトラフィックとの共存

(5) Header sizes, packet rates, scalability
ヘッダーサイズ、パケットレート、スケーラビリティ

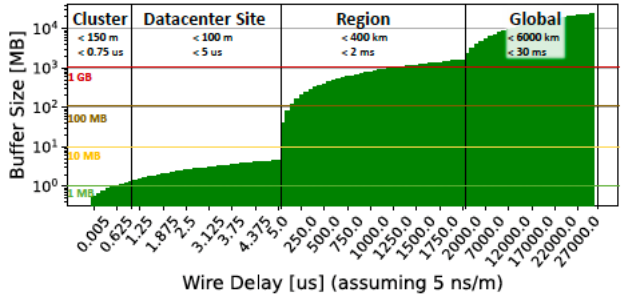
(6) No support for smart stacks
スマートスタック非サポート

(7) Security
セキュリティ

(8) Link-level Reliability
リンクレベルの信頼性



(a) Intra-datacenter per-switch headroom buffer.

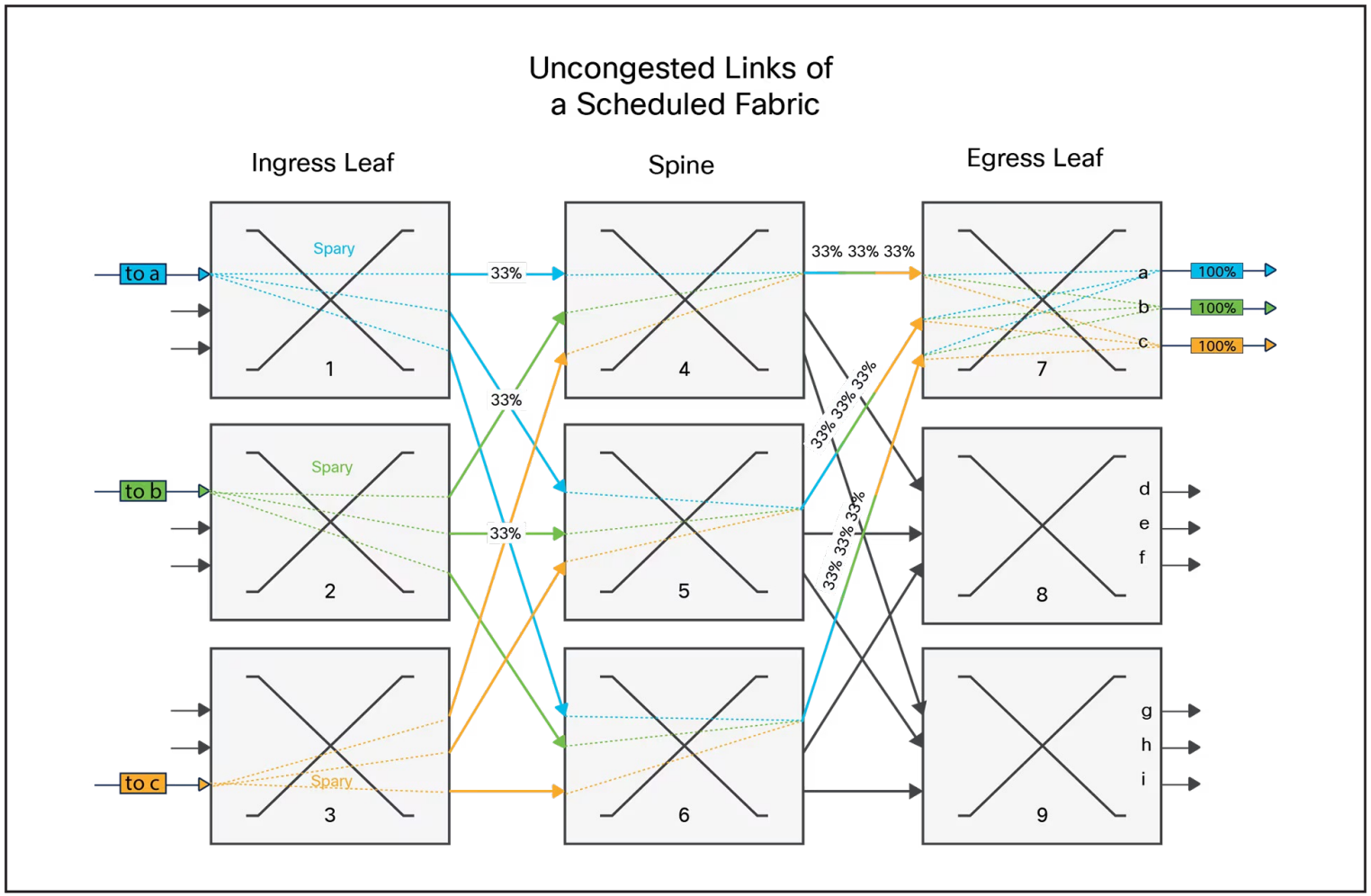


(b) Varying distance per-port headroom buffer.

Figure 1: Headroom Buffer Requirements.

"Datacenter Ethernet and RDMA: Issues at Hyperscale", Torsten Foesler et al, https://arxiv.org/abs/2302.03337

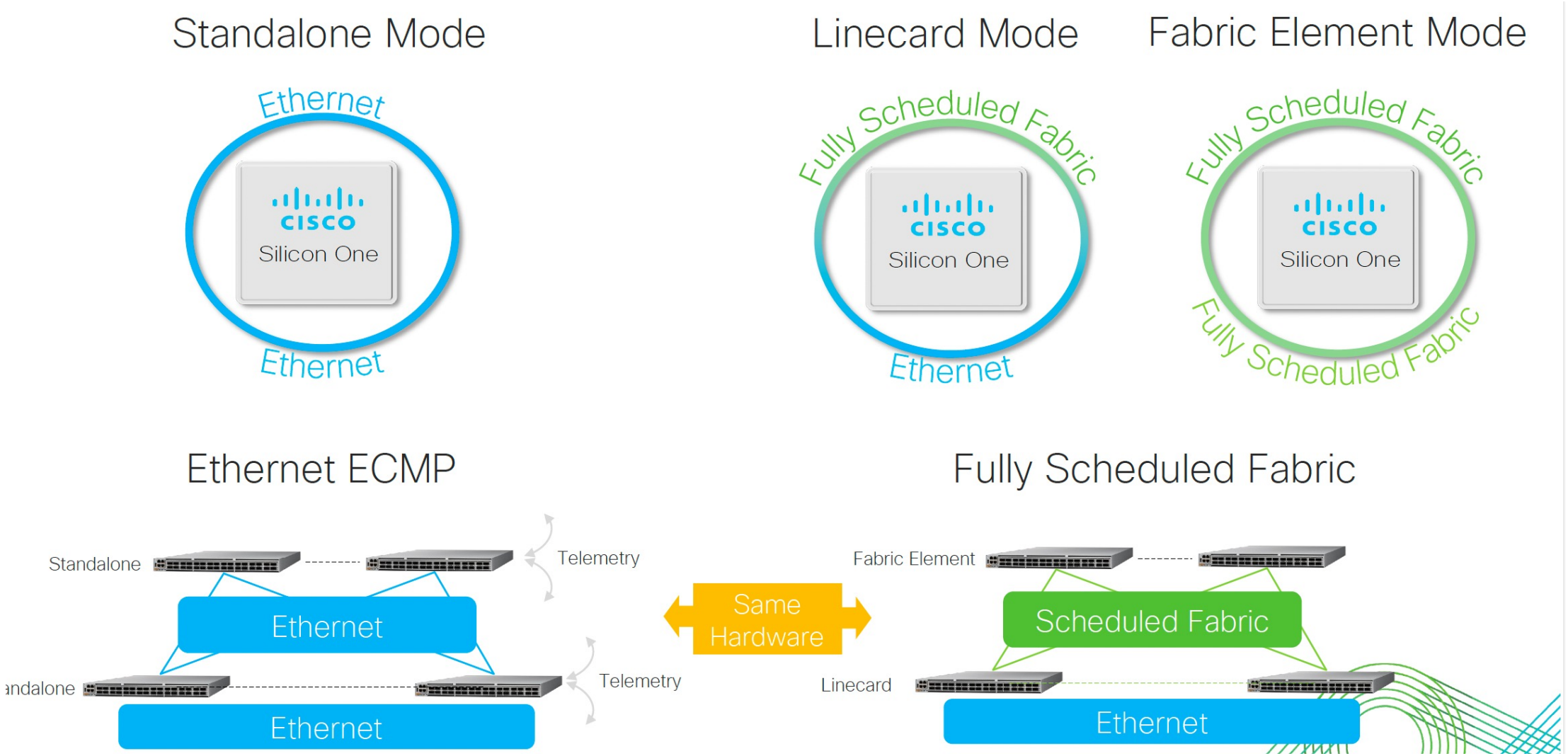
[参考1] Fully Scheduled Fabric による輻輳回避



- Ingress Virtual Output Queue (VoQ) にて、出力ポートとトラフィック・クラス(TC)宛てのパケットを格納する
- ある宛先に対してキューイングされたパケットがあると、Ingress VOQはスケジューラにリクエストを送る
- スケジューラによってGrantされたパケットだけが送信されるので、輻輳は起こらない
- リンクの選択にハッシュは使われない。パケットは、そのパケットがどのフローに関連付けられているかに関係なく、利用可能なすべてのリンクに散布(Spray)される
- パケットがEgressで受信されると、遅延を補正するために再順序付(Re-order)される
  （All Reduce, All-to-Allの場合、最後のパケットの到着だけが問題になる）

https://www.cisco.com/c/en/us/solutions/collateral/silicon-one/evolve-ai-ml-network-silicon-one.html

[参考2] Silicon One による複数モードの実装

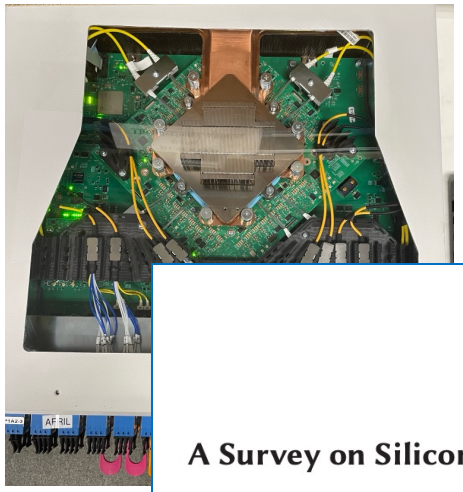


https://www.cisco.com/c/en/us/solutions/collateral/silicon-one/evolve-ai-ml-network-silicon-one.html

[参考3] Silicon One による Silicon Photonics サポート

Co-Packaged Optics Demo @ OFC2023

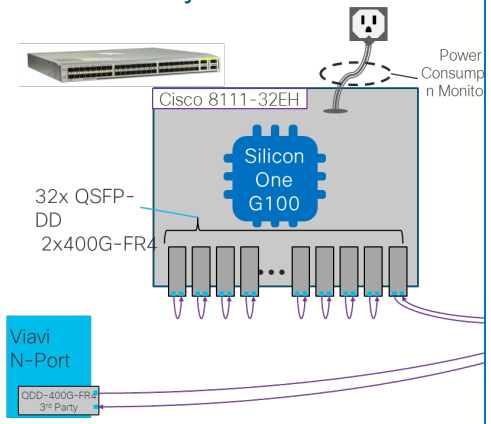
- Cisco CPO proof-of-concept demo:
  - Cisco Silicon One™ 25.6T G100 スイッチ ASIC
  - 8 x 3.2Tb ソケット型 optics tile をスイッチ ASIC と一体化
  - 64 x 400G-FR4 フル機能オプティクス ポートをフロント パネルに搭載
  - マルチベンダーOIF仕様準拠ELSFPモジュール
  - PICにモノリシック集積 optics Mux/Demuxを搭載
  - 消費電力 ~5.5W/800G 光インターコネクト



100x volume reduction

4x OSFP800 => 1x 3.2T CPO

Zoom in: System Demo



https://blogs.cisco.com/sp/cisco-demonstrates-co-packaged-optics-cpo-system-at-ofc-2023

61

A Survey on Silicon Photonics for Deep Learning

FEBIN P. SUNNY, EBADOLLAH TAHERI, MAHDI NIKDAST, and SUDEEP PASRICHA,
Department of Electrical and Computer Engineering, Colorado State University

Deep learning has led to unprecedented successes in solving some very difficult problems in domains such as computer vision, natural language processing, and general pattern recognition. These achievements are the culmination of decades-long research into better training techniques and deeper neural network models, as well as improvements in hardware platforms that are used to train and execute the deep neural network models. Many application-specific integrated circuit (ASIC) hardware accelerators for deep learning have garnered interest in recent years due to their improved performance and energy-efficiency over conventional CPU and GPU architectures. However, these accelerators are constrained by fundamental bottlenecks due to (1) the slowdown in CMOS scaling, which has limited computational and performance-per-watt capabilities of emerging electronic processors; and (2) the use of metallic interconnects for data movement, which do not scale well and are a major cause of bandwidth, latency, and energy inefficiencies in almost every contemporary processor. Silicon photonics has emerged as a promising CMOS-compatible alternative to realize a new generation of deep learning accelerators that can use light for both communication and computation. This article surveys the landscape of silicon photonics to accelerate deep learning, with a coverage of developments across design abstractions in a bottom-up manner, to convey both the capabilities and limitations of the silicon photonics paradigm in the context of deep learning acceleration.

CCS Concepts: • **Hardware → Emerging optical and photonic technologies**; • **Computing methodologies → Machine learning**; • **Hardware → Application-specific VLSI designs**;

"A Survey on Silicon Photonics for Deep Learning"
https://dl.acm.org/doi/pdf/10.1145/3459009

[参考4] Ultra Ethernet Consortium

JOINT DEVELOPMENT FOUNDATION PROJECT

Ultra Ethernet Consortium

WORKING GROUPS NEWS MEMBERSHIP CONTACT US BECOME A MEMBER

The New Era Needs a New Network

As performant as a supercomputing interconnect
Supercomputing 接続のように高性能

As ubiquitous and cost-effective as Ethernet
Ethernet のように高コスト効率

As scalable as a cloud data center
Cloud Data Center のようにスケーラブル

- マルチパスとパケットスプレー
- 柔軟な配信順序
- 最新の輻輳制御メカニズム
- エンドツーエンドのテレメトリー
- 大規模、安定性、信頼性

https://ultraethernet.org

HPCC による改善

HPCC : High Precision Congestion Control

- ネットワークテレメトリ情報を活用して正確かつタイムリーにレート制御を行うことにより、過剰反応しない！高速に反応する！
  →超低遅延、広帯域、ネットワークの安定性を同時に実現

HPCC: High Precision Congestion Control

Yuliang Li[♠♡], Rui Miao[♠], Hongqiang Harry Liu[♠], Yan Zhuang[♠], Fei Feng[♠], Lingbo Tang[♠], Zheng Cao[♠], Ming Zhang[♠],
Frank Kelly[♢], Mohammad Alizadeh[♣], Minlan Yu[♡]
Alibaba Group[♠], Harvard University[♡], University of Cambridge[♢], Massachusetts Institute of Technology[♣]

ABSTRACT

Congestion control (CC) is the key to achieving ultra-low latency, high bandwidth and network stability in high-speed networks. From years of experience operating large-scale and high-speed RDMA networks, we find the existing high-speed CC schemes have inherent limitations for reaching these goals. In this paper, we present HPCC (High Precision Congestion Control), a new high-speed CC mechanism which achieves the three goals simultaneously. HPCC leverages in-network telemetry (INT) to obtain precise link load information and controls traffic precisely. By addressing challenges such as delayed INT information during congestion and overreaction to INT information, HPCC can quickly converge to utilize free bandwidth while avoiding congestion, and can maintain near-zero in-network queues for ultra-low latency. HPCC is also fair and easy to deploy in hardware. We implement HPCC with commodity programmable NICs and switches. In our evaluation, compared to DCQCN and TIMELY, HPCC shortens flow completion times by up to 95%, causing little congestion even under large-scale incasts.

CCS CONCEPTS

• **Networks → Transport protocols; Data center networks;**

demand on high-speed networks. The first trend is new data center architectures like resource disaggregation and heterogeneous computing. In resource disaggregation, CPUs need high-speed networking with remote resources like GPU, memory and disk. According to a recent study [17], resource disaggregation requires 3-5μs network latency and 40-100Gbps network bandwidth to maintain good application-level performance. In heterogeneous computing environments, different computing chips, *e.g.* CPU, FPGA, and GPU, also need high-speed interconnections, and the lower the latency, the better. The second trend is new applications like storage on high I/O speed media, *e.g.* NVMe (non-volatile memory express) and large-scale machine learning training on high computation speed devices, *e.g.* GPU and ASIC. These applications periodically transfer large volume data, and their performance bottleneck is usually in the network since their storage and computation speeds are very fast.

Given that traditional software-based network stacks in hosts can no longer sustain the critical latency and bandwidth requirements [43], offloading network stacks into hardware is an inevitable direction in high-speed networks. In recent years, we deployed large-scale networks with RDMA (remote direct memory access) over Converged Ethernet Version 2 (RoCEv2) in our data centers

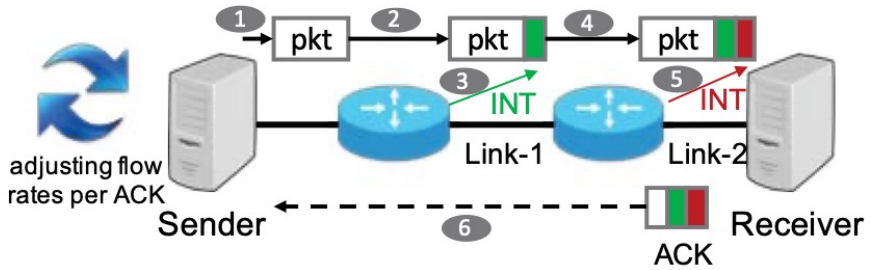
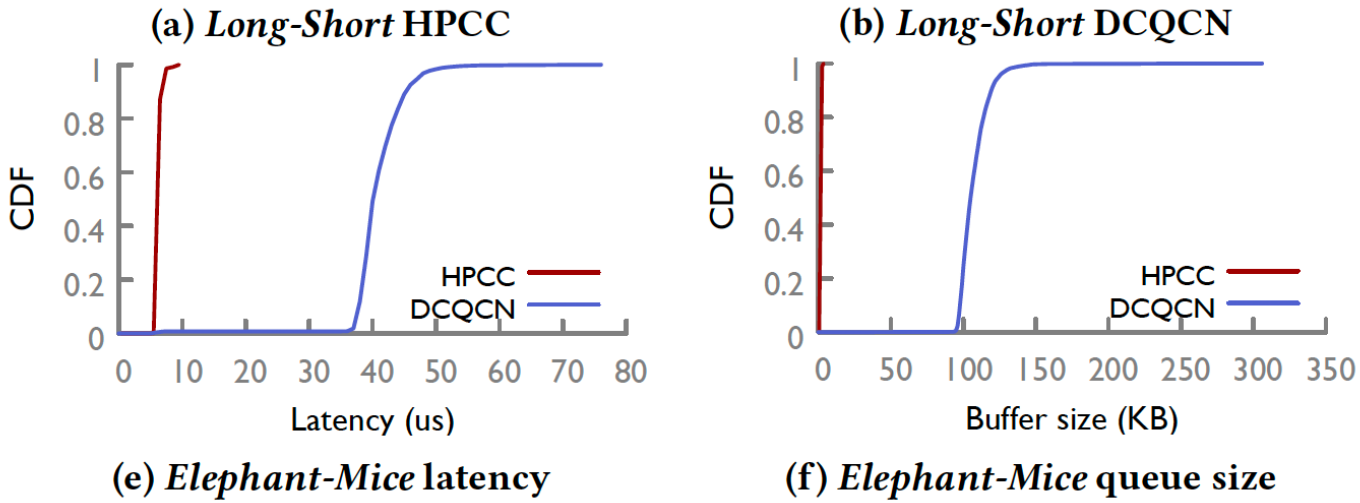


Figure 4: The overview of HPCC framework.

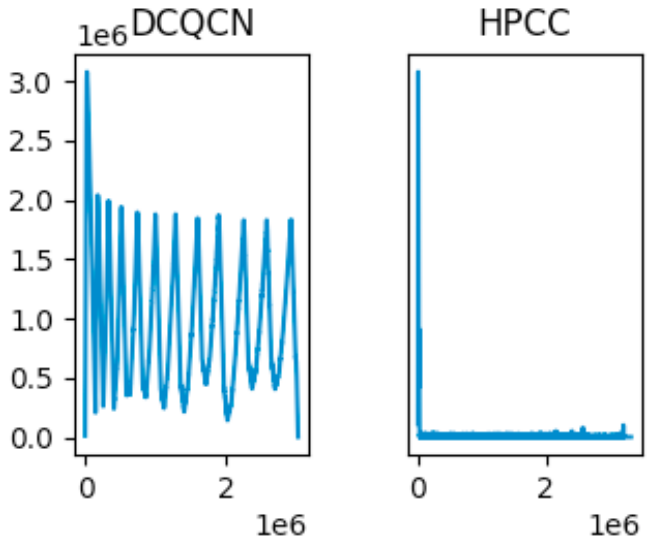
https://dl.acm.org/doi/pdf/10.1145/3341302.3342085

HPCC による改善

- Mice flowの遅延とキューサイズ (Elephant flowを流しながら)



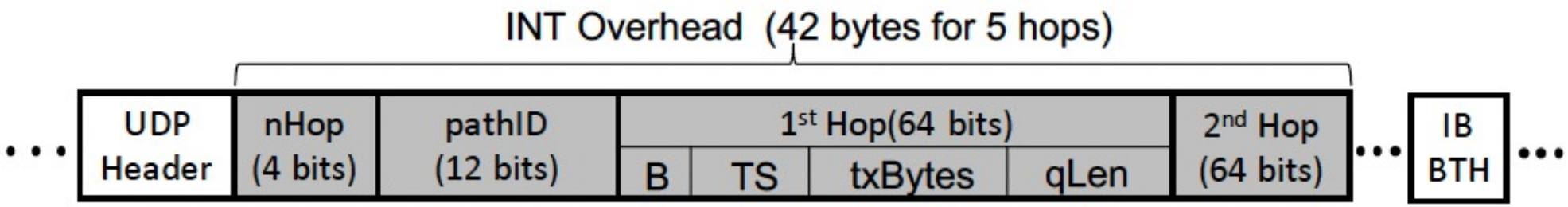
- Incast Traffic におけるキュー長



SIGCOMM '19: Proceedings of the ACM Special Interest Group on Data Communication
HPCC: high precision congestion control
Pages 44–58

T. Khan, S. Rashidi, S. Sridharan, P. Shurpali, A. Akella and T. Krishna, "Impact of RoCE Congestion Control Policies on Distributed Training of DNNs," *2022 IEEE Symposium on High-Performance Interconnects (HOTI)*, CA, USA, 2022, pp. 39-48, doi: 10.1109/HOTI55740.2022.00021.

HPCC packet format

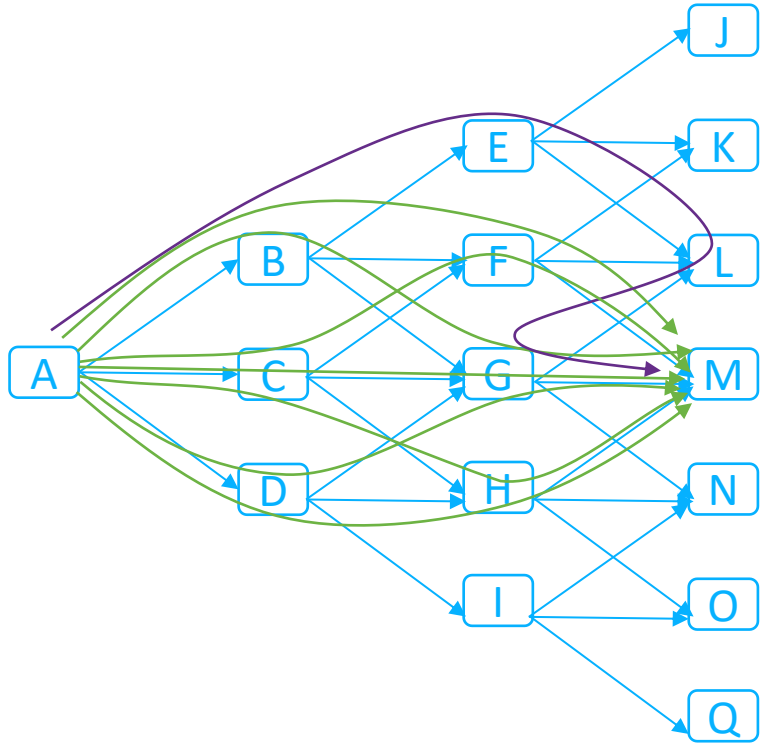


- nHop（4ビット）：　送信側ホストによって0として初期化される
- pathID（12ビット）：　パスの変更を検出するため、パスに沿ったすべてのスイッチIDのXOR
- B（4ビット）：　egressポートの速度タイプ（40Gbps、100Gbpsなど）
- TS（24ビット）：　パケットがegressポートから放出された時のタイムスタンプ
- txBytes（20ビット）：　egressポートから送信された累積合計バイト数
- qLen（16ビット）：　egressポートの現在のキュー長

SIGCOMM '19: Proceedings of the ACM Special Interest Group on Data Communication
HPCC: high precision congestion control
Pages 44–58

SRv6 Path Tracing ?!

- パケット単位のdeterministic（決定論的）なトレース
- HWパイプラインにLinerateで実装
  CPUへのpunt、Co-processor へのオフロードなし
- MTU効率の良さ：3 bytes / hop
  Interface (12 bits), Timestamp (8 bits), Load (4 bits)
- SRv6/IPv6をサポート
- シームレスなデプロイメント
  レガシーノードとのインターワーク



Path Tracing - Packet Format

- IPv6 Header

  SA, DA, Traffic Class, Flow Label…

- IPv6 Hop by Hop Path Tracing Option

  MCD Stack

  - MCD: Midpoint Compressed Data
  - MCD.OIF: 8 or 12-bit Outgoing Interface ID
  - MCD.OIL: 4-bit Outgoing Interface Load
  - MCD.TTS: 8-bit Truncated PTP TimeStamp

- IPv6 Destination Option for Path Tracing

  - T64: 64-bit Timestamp
  - Session ID: プローブを生成する SRC ノードが設定するセッション識別子。同じセッションのプローブを関連付けるために使用される。
  - IF_ID: 12-bit Interface ID
  - IF_LD: 4-bit Interface Load

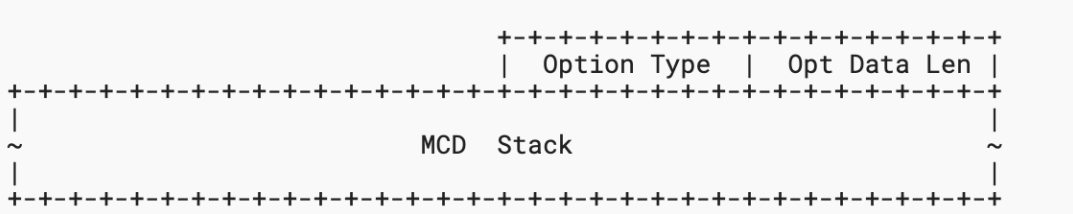


Figure 1: IPv6 Hop-by-Hop Option for Path Tracing (HbH-PT)

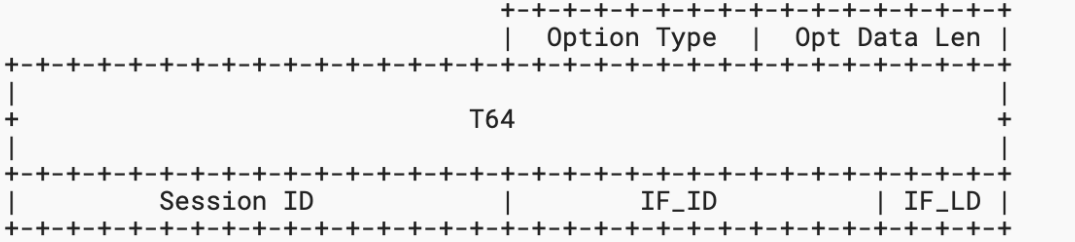


Figure 2: IPv6 Destination Option for Path Tracing (DOH-PT)

https://www.ietf.org/archive/id/draft-filsfils-spring-path-tracing-05.html

Path Tracing の 特徴

- 他の Inband Telemetry 方式 (INT, IFA, iOAM)に比較して、低オーバヘッド
- HW親和性とLine Rate処理
  - 現存のハードウェアでラインレートの処理が可能
  - コプロセッサやスローパスへのオフロード不要
- スケーラブルできめ細かいタイムスタンプ
  - 64 bit (Source Node, Sink Node)
  - 8 bit (Mid Node)
- スケーラブルな負荷測定

- INT (In-Network Telemetry) : https://github.com/p4lang/p4-applications/blob/master/docs/INT_v2_0.pdf
- IFA (Inband Flow Analyzer) : https://datatracker.ietf.org/doc/html/draft-kumar-ippm-ifa
- iOAM (In-Situ OAM) : https://datatracker.ietf.org/doc/html/rfc9197

SRv6 for Data Center

- 共通のデータプレーン (Access, Backbone, DC Fabric, SmartNIC/DPU, Computing…)
- オーバヘッド排除
  - Load Balancing のための UDP/IP 不要 （IPv6 Flow Labelを活用）
  - Multi tenancyのための VXLAN 不要（SRv6 uSID)
  - HPCCのためのTimestamp バイト数縮減 （Path Tracing）
- トラフィックエンジニアリングサポート
  - uSIDにより、6 hopまではSRHも不要
- 豊富なテレメトリーサポート
  - Path Tracing, 3L (Latency, Loss, Liveness) Proving, Demand Matrix
- 柔軟性・拡張性
  - SRv6 Network Programmingにより、SHArP [*] のような In-Network Computingにも対応可能性
- Cisco Silicon OneにおけるSRv6 uSID Native Support

[*] Scalable Hierarchical Aggregation Protocol

Conclusion

- AI/DL トレーニング・推論は、数千のGPUクラスタにより処理され、それを効率的に行うために、ネットワークとの連携はますます重要になっている
- AI/DL Workloadの大規模化・多様化に、スケールメリットを以て対応するために、Ethernet技術を活用したオープンな技術が求められいる
- ロスレスファブリックを、高速・低遅延・高効率に実現するためには、現在のDCQCNでは問題が多く、HPCCなどの高度な輻輳制御が必要になる
- HPCCは、Hop by hopのテレメトリ情報を利用して精緻な輻輳制御を行い、パフォーマンスを大きく改善できる（一方、それだけでは解決できない問題も残る）
- SRv6 Path TracingのHPPCへの適用、およびSilicon One + SRv6 DCにより、より効率的で将来性の高いDC Fabricを構築できる可能性がある

The bridge to possible