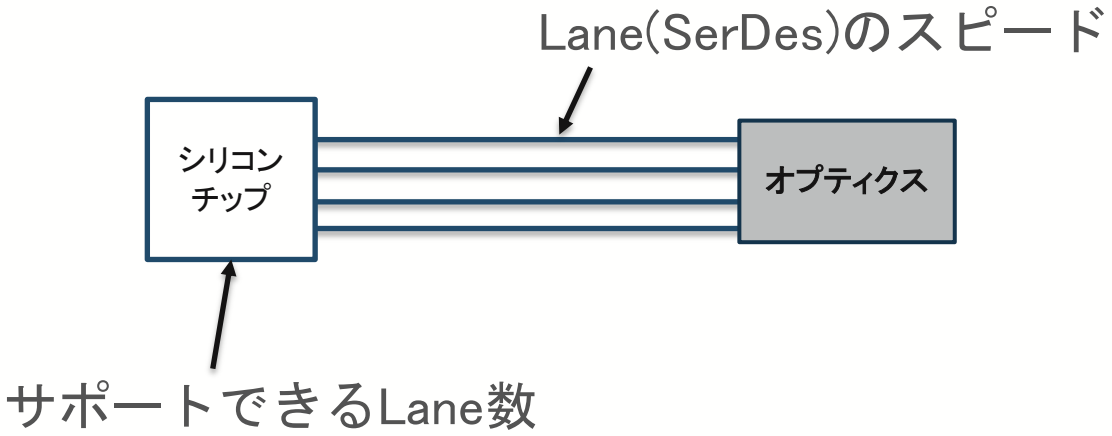
PAM4がサーバーサイドにやってくる

-pam4 changes the server world drastically-

Shishio Tsuchiya
shtsuchi@arista.com

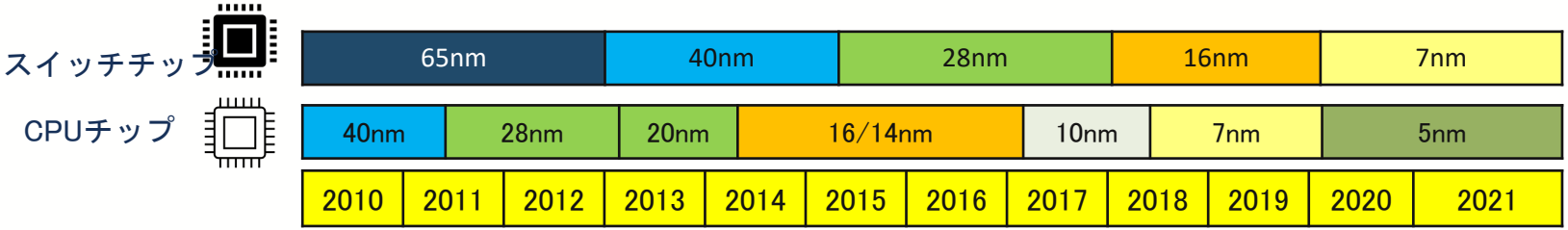
ARISTA

SoC(System On Chip)の容量変化



- SerDesスピードおよび1チップでサポートできるLane数によりSoC容量が決まる
- ムーアの法則(2年間に2倍)をネットワークに

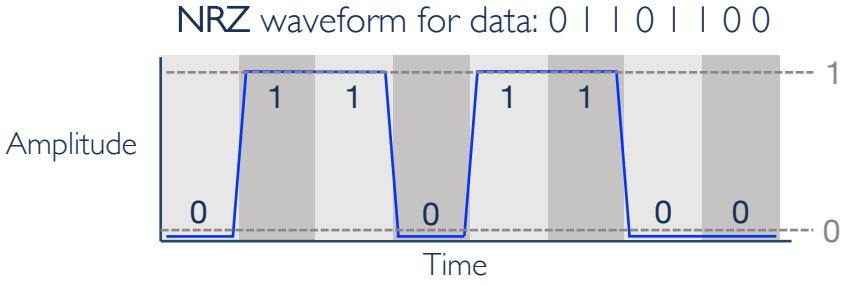
半導体技術の進化とロードマップ

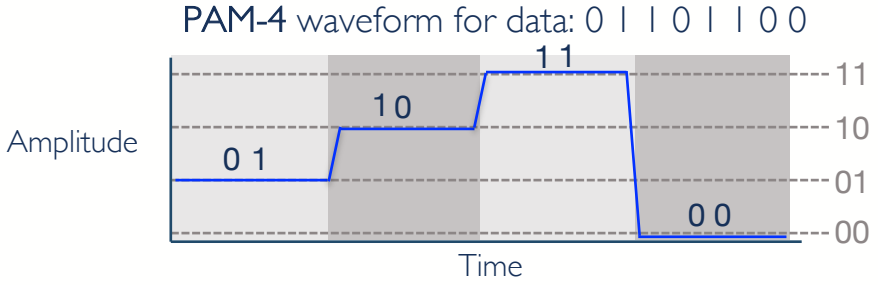


- 半導体技術は進化しており、CPUチップの進化に追従し、スイッチチップも進化していく
- 10年前はかなり遅れていたが、現在は最新のテクノロジーを用いた開発が進んでいる

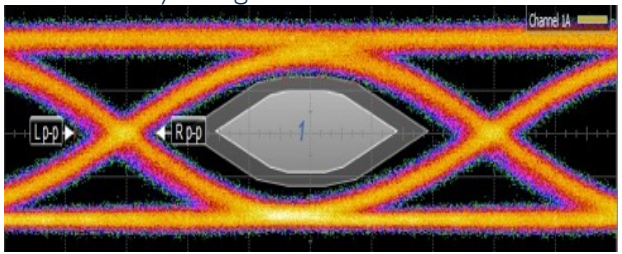
PAM4(Pulse Amplitude Modulation:4パルス振幅変調)

- 400G転送では従来のNRZ(Non Return to Zero)からPAM4に移行
- 01,10,11,00の4つの電圧レベルのパルス信号として伝送

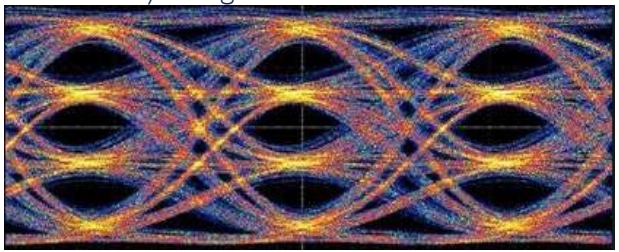




Eye diagram for NRZ data

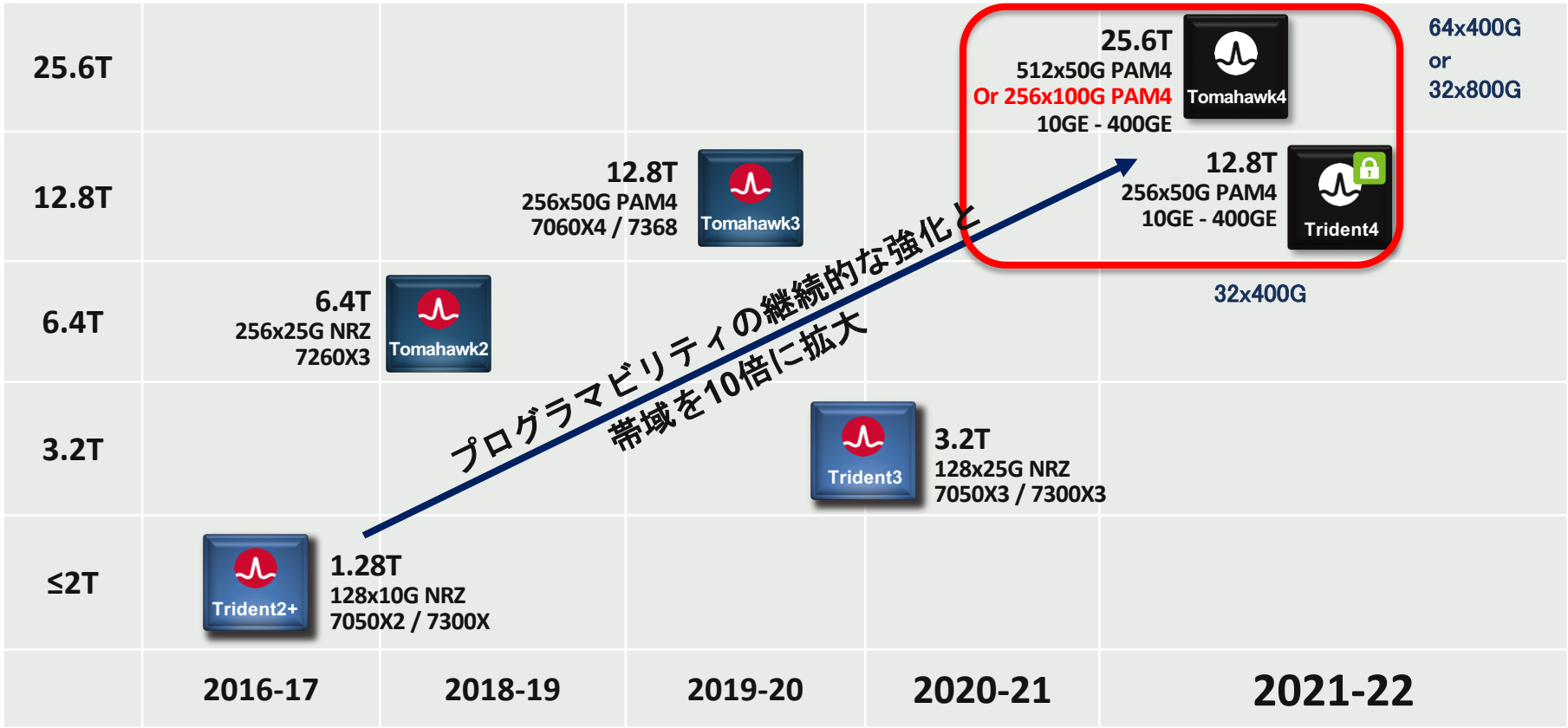


Eye diagram for PAM-4 data



ARISTA

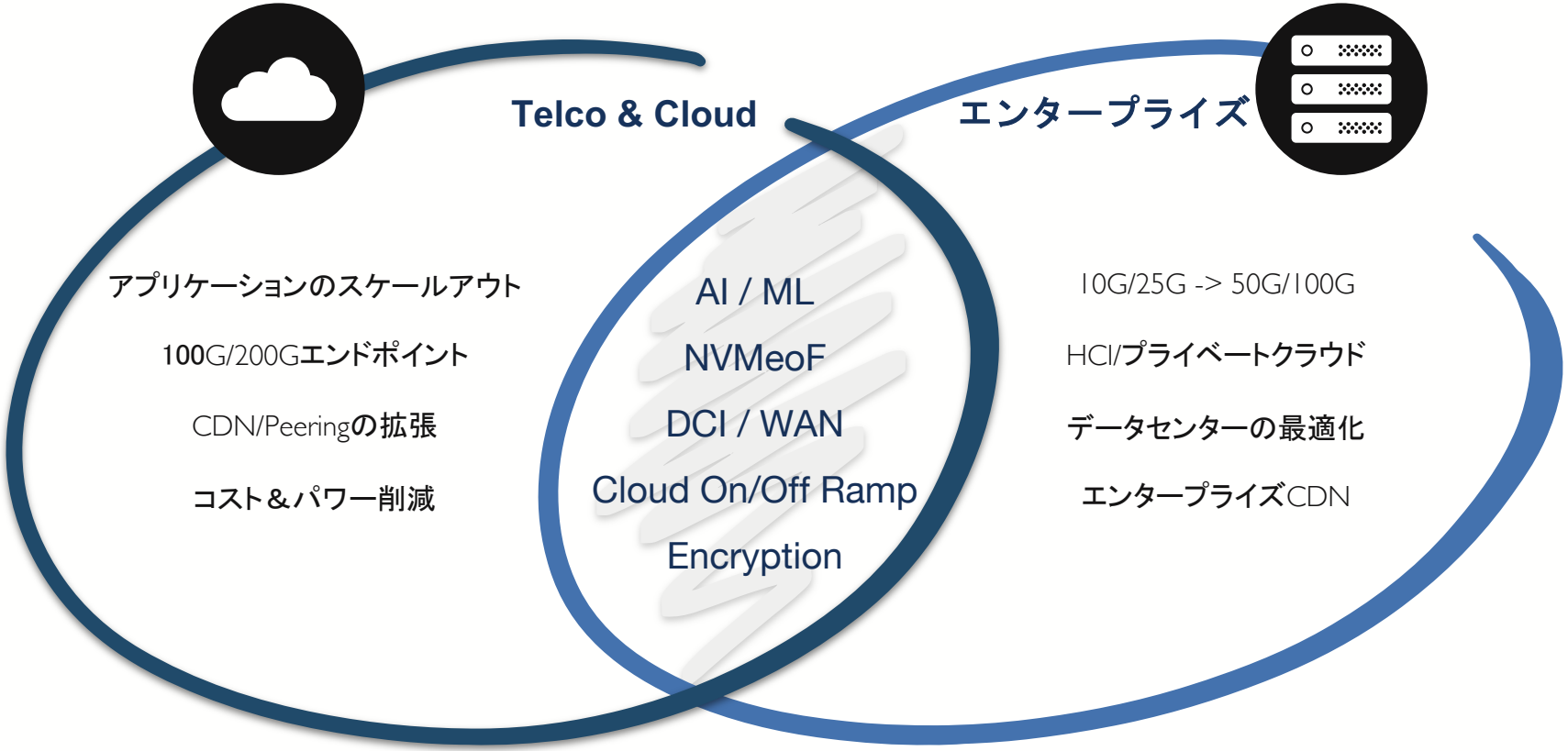
XGSデータセンターポートフォリオ



ARISTA

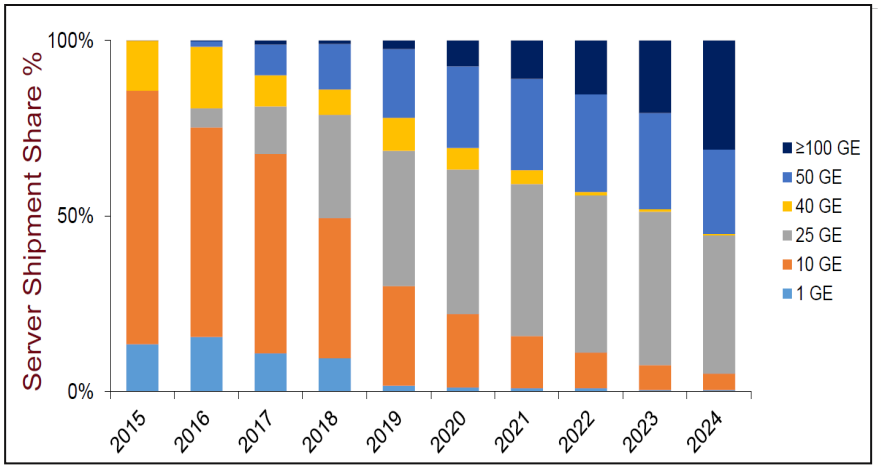
MH

データセンターで400Gを推進するトレンド

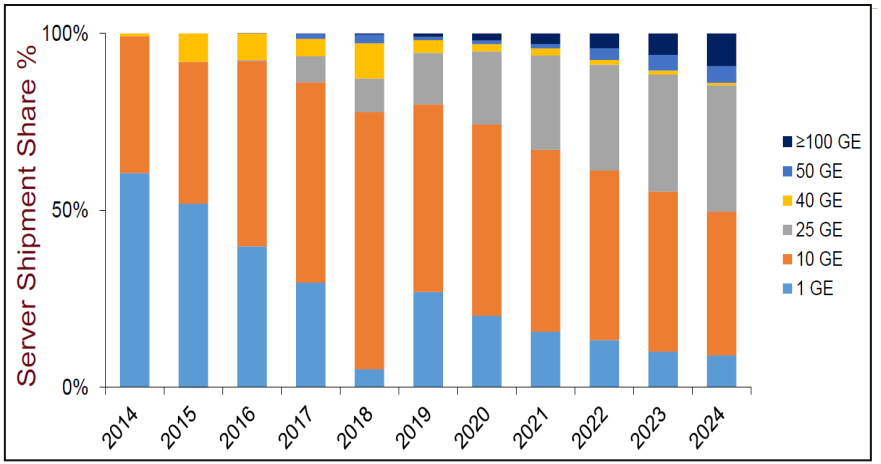


ARISTA

サーバーとNICのスピード移行



クラウドサーバのリンクスピード予測

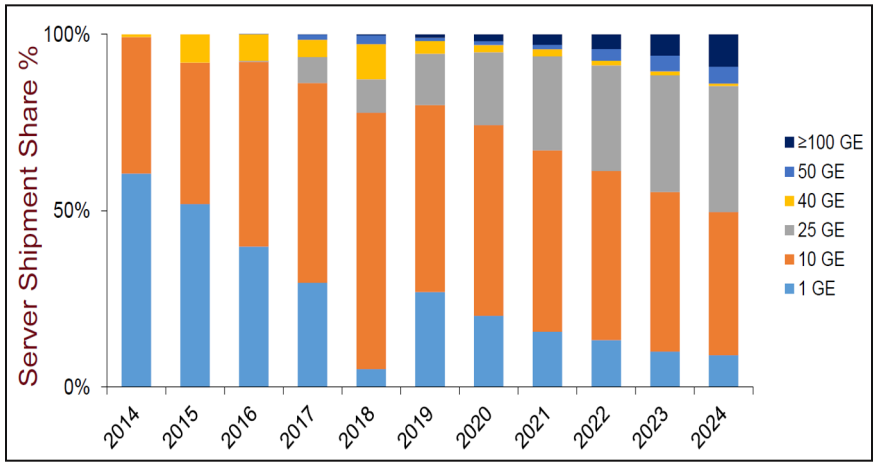


エンタープライズサーバのリンクスピード予測

クラウドとエンタープライズの両方のスピード移行が進行中
コンピュートとストレージの更新と成長に基づくネットワークスイッチの需要
新世代の400GシステムがOpexとCapexを低減

Source: Dell'Oro Group Inc. – Cloud Data Center Capex Servers and Storage Systems Controllers and Adapters 4Q19

エンタープライズ・サーバーとNICスピードの移行



エンタープライズサーバのリンクスピード予測

- PCIe4.0がホスト側の50Gbps化をドライブ 50/100/200GPAM4技術に統合
- 50G SerDesは次世代PAM4ベースのTORで変換
- TOR/Spineスイッチの通信はコンピュートノードにより促進される
- ストレージはNVMeoFへ移行
- 1/10Gから25/50/100Gへの移行が進む
- TORによる低消費電力、収容能力の向上、少ない階層数の実現
- DCIはZRオプティクスで400Gに移行

Source: Dell'Oro Group Inc. – Cloud Data Center Capex Servers and Storage Systems Controllers and Adapters 4Q19

NICの変化

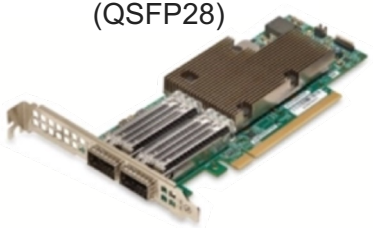
現在、幅広く展開されている
10G / 1G NRZ

10G SFP+ & 1G SFP

広く普及し、成長している
25G NRZ

25G SFP

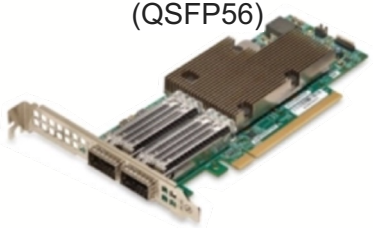
100G & 50G-2 QSFP
(QSFP28)



最近リリースされたもの、またはロードマップ
50G PAM-4

50G SFP

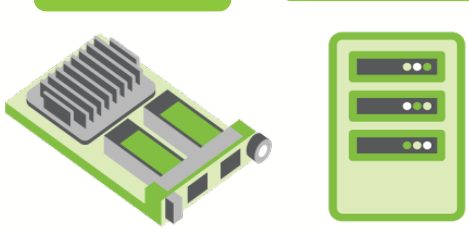
200G & 100G-2 QSFP
(QSFP56)



100G-2 SFP-DD or DSFP

OCP3.0

https://www.opencompute.org/wiki/Server/Mezz



- より高い熱設計電力(TDP)のNIC
- シングルコネクタ(SFF、TSFF)カードは最大80W
- デュアルコネクタ(LFF)カードは最大150W
- 最大 PCIe Gen 4 (16 GT/s:GigaTransfer per second)をサポート。
  - コネクタはPCIe Gen 5 (32 GT/s)と電気的互換性あり
- OCP NIC 3.0カード1枚あたり最大32レーンのPCIeをサポート
- シングルホスト、マルチルートコンプレックス、マルチホスト環境に対応
- より複雑なOCP NIC 3.0カードの設計に対応するため、より大きな基板面積をサポート
- DRAMとアクセラレータを搭載したスマートNICの実装をサポート
- FRUの取り付け、取り外しを簡素化し、全体のダウンタイムを短縮

NetXtreme® E-Series

OCP NIC 3.0 Ethernet Adapters

https://docs.broadcom.com/doc/netxtreme-e-series-pcie-nic-ethernet-adapters-specification-sheet

N2100G

N1100GD

N2200GD

Portfolio and Ordering Information

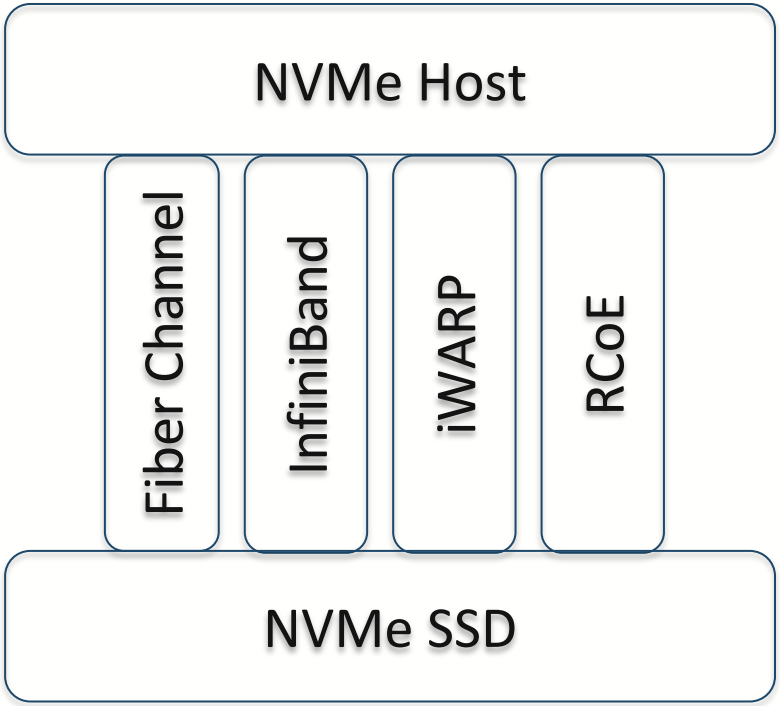
Part Number	Name	Port Speed	I/O	Host I/F	Multihost
BCM95719-N1905C	N41T	4x 1G	RJ-45	PCIe 2.0 x4	No
BCM957412-N4120C	N210P	2x 10G	SFP+	PCIe 3.0 x8	No
BCM957416-N4160C	N210TP	2x 10GBASE-T	RJ-45	PCIe 3.0 x8	No
BCM957414-N4140C	N225P	2x 25G	SFP28	PCIe 3.0 x8	No
BCM957504-N425G	N425G	4x 25G	SFP28	PCIe 3.0/4.0 x16	Yes
BCM957504-N1100G	N1100G	1x 100G	QSFP56	PCIe 3.0/4.0 x16	Yes
BCM957504-N1100GD	N1100GD	1x 100G	DSFP	PCIe 3.0/4.0 x16	No, Multiroot only
BCM957508-N2100G	N2100G	2x 100G	QSFP56	PCIe 3.0/4.0 x16	Yes
BCM957508-N2200G	N2200G	2x 200G	QSFP56	PCIe 3.0/4.0 x16	Yes

NVIDIA MCX623432AC-GDAB ConnectX-6 Dx EN Adapter Card OCP 3.0 50GbE Crypto Enabled

https://store.nvidia.com/en-us/networking/store/product/MCX623432AC-GDAB/NVIDIAMCX623432ACGDABConnectX6DxENAdapterCardOCP3050GbECryptoEnabled/

- SFP56 50Gbps x2

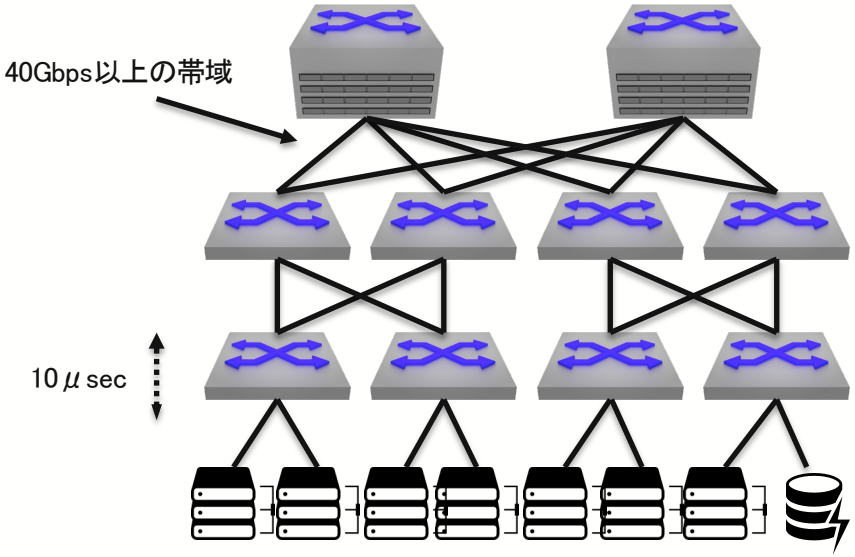
NVMe Over Fabric



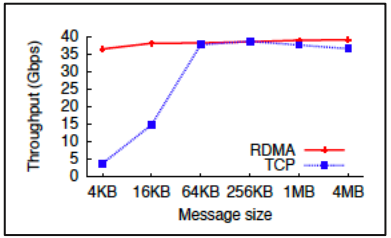
- NVMe Over Fabricとは異なるネットワーキングファブリックを介してNVMeストレージにアクセスするためのアーキテクチャ
- Remote Direct Memory Access(RDMA)

Congestion Control for Large-Scale RDMA Deployments

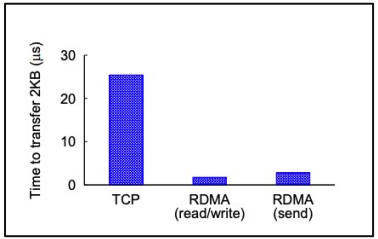
https://conferences.sigcomm.org/sigcomm/2015/pdf/papers/p523.pdf



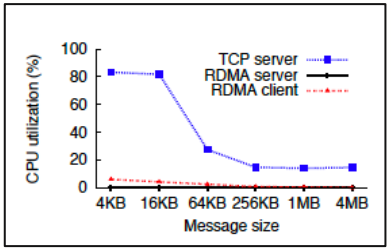
TCP vs RDMA



スループット



転送遅延時間



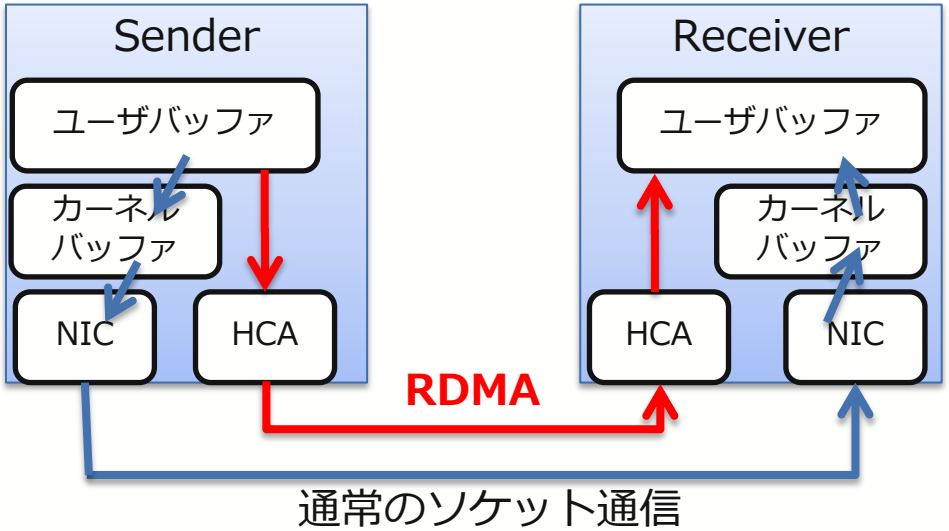
CPU利用率

- 2015年のSIGCOMMの論文
- 昨今のデータセンターにおけるアプリケーション要求は広帯域(40Gbps以上)/超低遅延(10μ秒/hop)/少ないCPU負荷で有ることが求められる

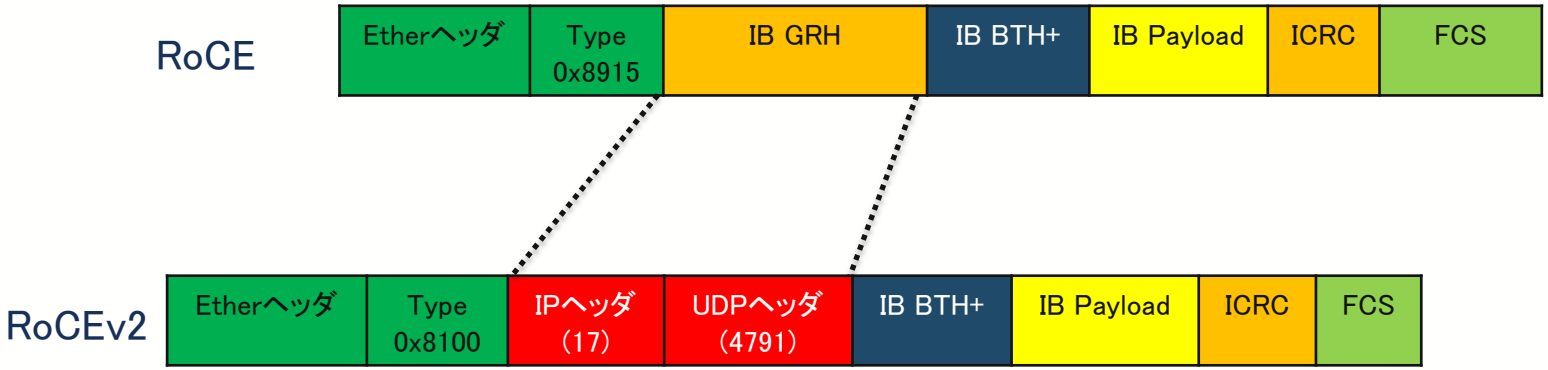
ARISTA

RDMAとは？

- Remote Direct Memory Access(RDMA)
  - アプリケーション間で直接通信
  - システムバス、CPUの負荷を低減
- ロスレスが前提
  - 従来InfiniBandで活用
  - TCPを利用しない=ロスレスが要件
  - RDMA over Ethernet

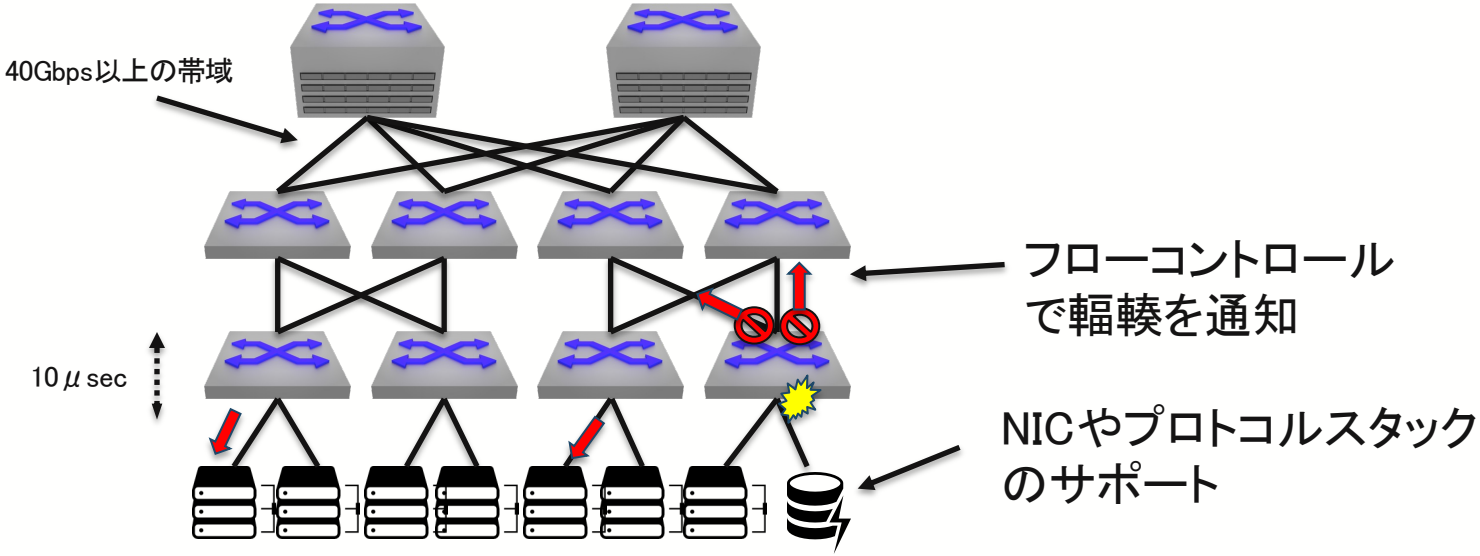


RoCEv2



- RoCEはレイヤー2のみの実装であったのに比較して、RoCEv2はIPもしくはIPv6を使用し、ルーティング可能
  UDPポート4791を使用

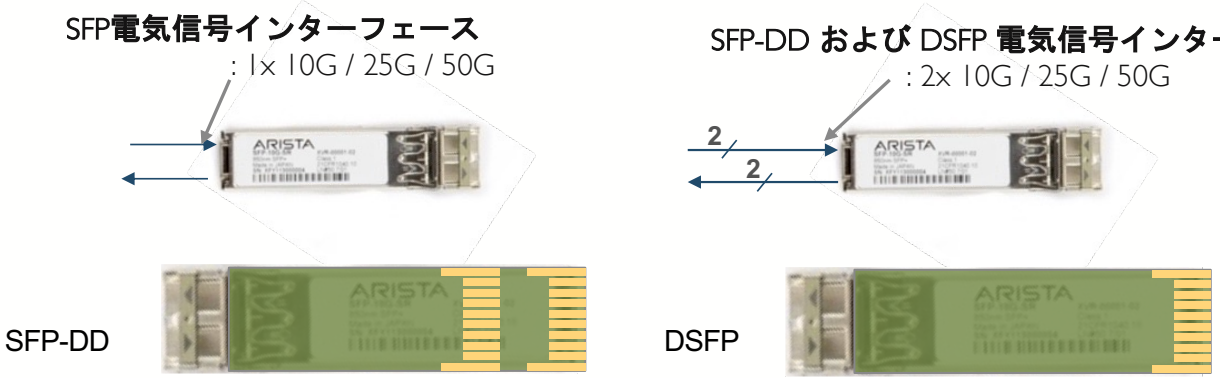
ロスレスイーサネットネットワークを構築するには



- DCBX(Data Center Bridging Capability Exchange)や PFC(Priority-Based Flow Control)が必要になる

SFP-DDとDSFP

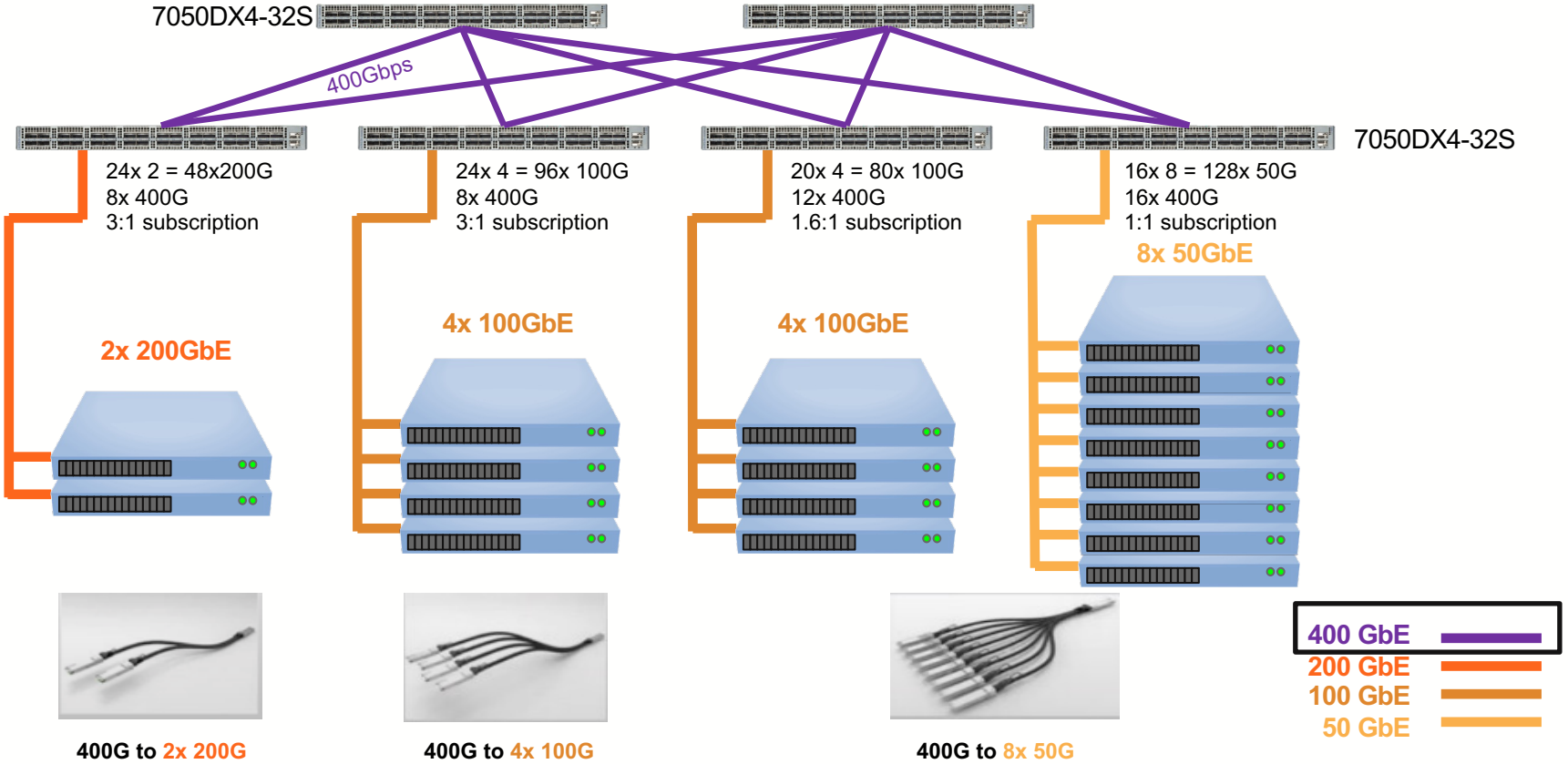
- SFP-DDとDSFPは違うアプローチで同じ結果を
- 2つのMSAでそれぞれマルチベンダーサポート
  - SFP-DD MSA http://sfp-dd.com/
  - DSFP MSA
- 共にSFPと類似のフォームファクターで2x50G 100Gを実現
- 中国のいくつかのクラウドタイタンでDSFPを使用/他はSFP-DDと考えられている



50/100/200G ハイパフォーマンス・コンピューティングの NIC: TORオプション

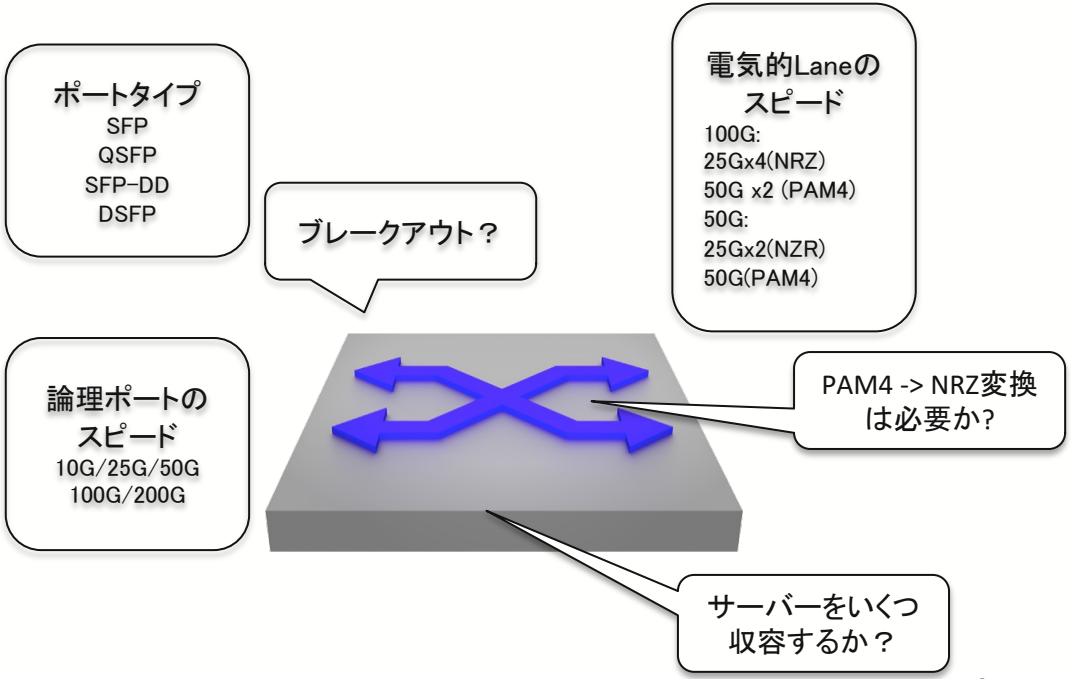
NICの帯域	コネクタ	TORスイッチの構成	スイッチのブレークアウト構成	DACケーブル	Total NIC ports from TOR with oversubscription
50GbE (2x25)	QSFP28 (NRZ)	32x100G 64x100G	2x50G	QSFP28(100G) -> 2xQSFP28(50G)	24x2= 48 2.4Tbps 8x100Gアップリンク (3:1) 48x2= 96 4.8Tbps 16x100Gアップリンク (3:1)
50GbE (1x50)	SFP56 (PAM4) / QSFP56 (PAM4)	32x400G 48x100G + 8x400G	8x50G 4x50G 2x50G	QSFP-DD(400G) -> 8xSFP56(50G) QSFP-DD(400G) -> 4xQSFP28(50G) SFP-DD(100G) -> 1xSFP56(50G)	24x8= 192 9.6Tbps 8x400Gアップリンク (3:1) 24x4= 96 4.8Tbps 8x400Gアップリンク (1.5:1) 48x1= 48 2.4Tbps 8x400G uplinks (1:1)
100GbE (2x50)	QSFP56 (PAM4)	32x400G 40x200G 48x100G + 8x400G	4x100G 2x100G n/a	QSFP-DD(400G) -> 4xQSFP56(100G) QSFP56(200G) -> 1xQSFP56(100G) SFP-DD(100G) -> QSFP56(100G)	24x4= 96 9.6Tbps 8x400Gアップリンク (3:1) 36x1= 36 3.6Tbps 4x200Gアップリンク(4.5:1) 48x1= 48 4.8Tbps 8x400Gアップリンク (1.5:1)
200GbE (4x50)	QSFP56 (PAM4)	32x400G 40x200G	2x200G n/a	QSFP-DD(400G) -> 2xQSFP56(200G) QSFP56(200G) -> QSFP56(200G)	24x2 = 48 9.6Tbps 8x400Gアップリンク(3:1) 32 6.4Tbps 8x200Gアップリンク(4:1)

50G / 100G / 200G / 400Gへの迅速な展開?



ARISTA

これからのTORの選択



- NICが多様化する事でそれをTORも様々スピードおよび電気信号をサポートする必要がある

まとめ

- PAM4の出現によりSerDesのスピードが高速化
  50Gbps/100Gbps時代に突入
- これに対応したPCIe4.0のNICがリリースされる
- ML/AIによる高速も多様化
- サーバースピードのChangeについていけるか?

議論したいこと

- サーバーインターフェース状況とか今どんな感じ？
- インターフェースの選択
  - 50Gbps:SFP-DD or DSFP
  - 200Gbps: QSFP56 or QSFP-DD