

# Hi-Speed Wide-band Storage Networking for IP networkers

**Tajima Hirotaka**

JANOG45

Jan.24,2020

~~(好きなことが言えるようにする)~~

## お約束のスライド

- tajiの所属組織の見解ではありません
- tajiの妄想を含みます

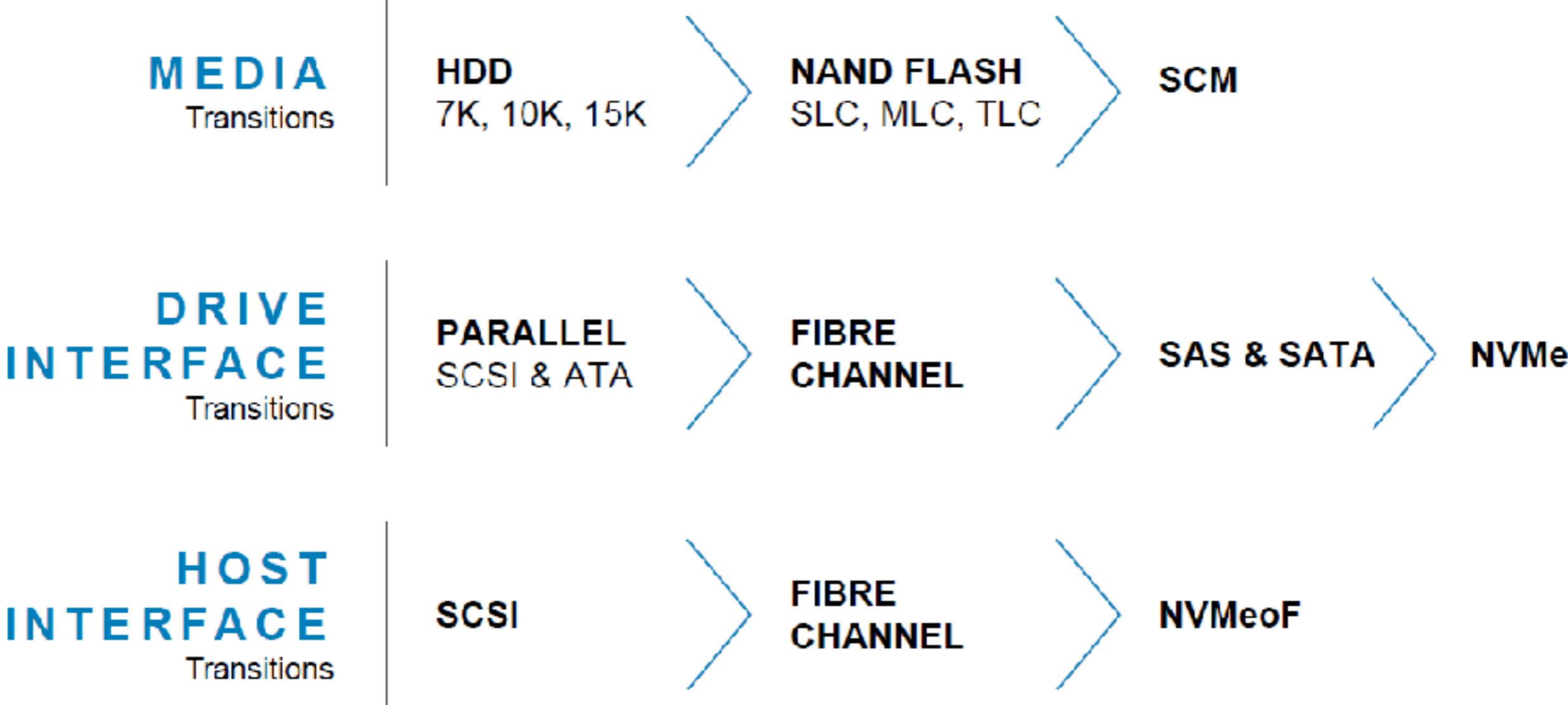
# Agenda

1. Storage Networking 基礎知識
2. 知っておきたい○○ over IP/Ethernet

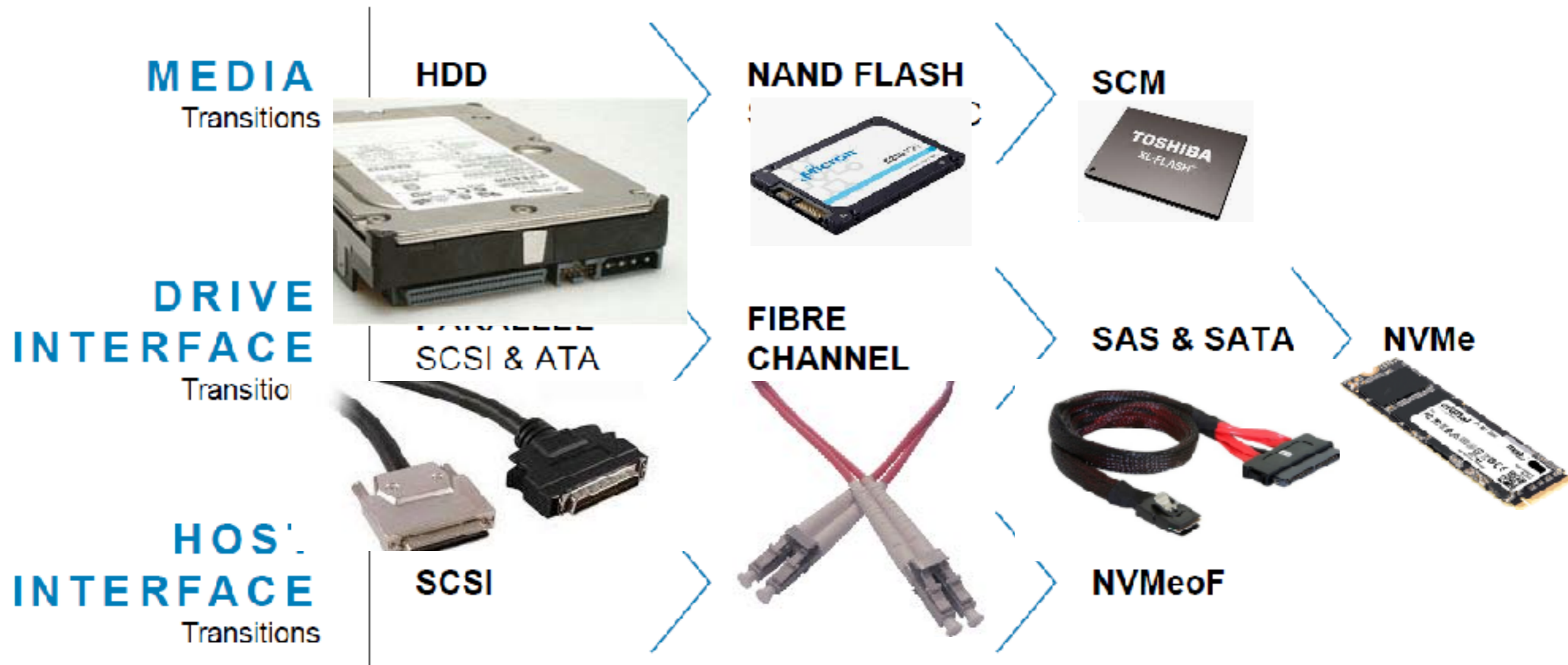
# Agenda

1. Storage Networking 基礎知識
2. 知っておきたい○○ over IP/Ethernet

# Storage technology transitions



# Storage technology transitions

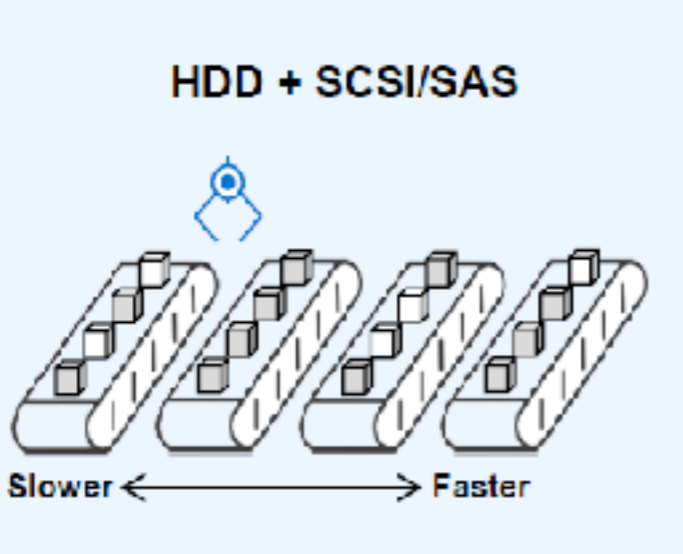
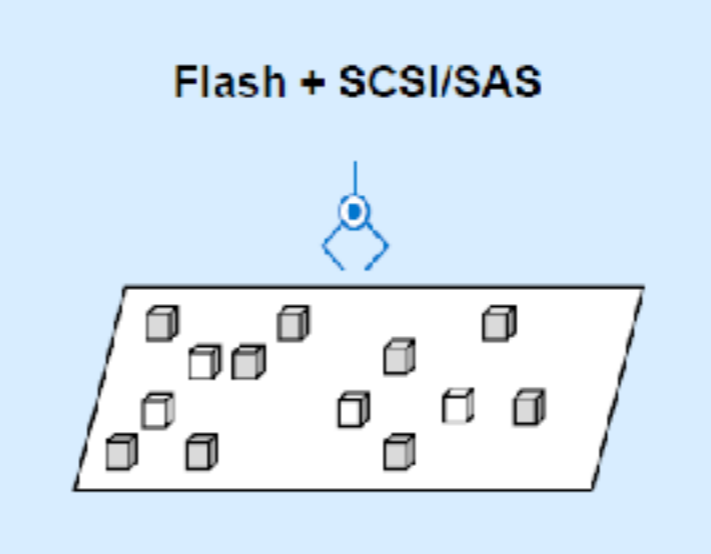
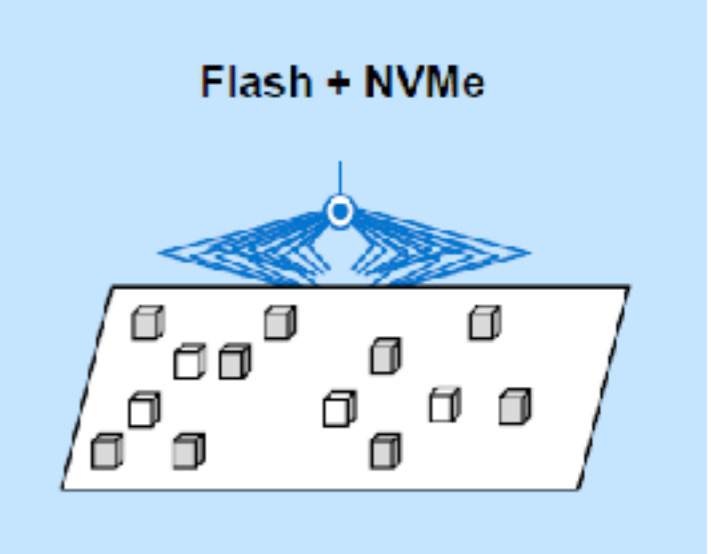



# **NVMe of (over Fabrics)**

# **NVMe of** **(over Fabrics)**



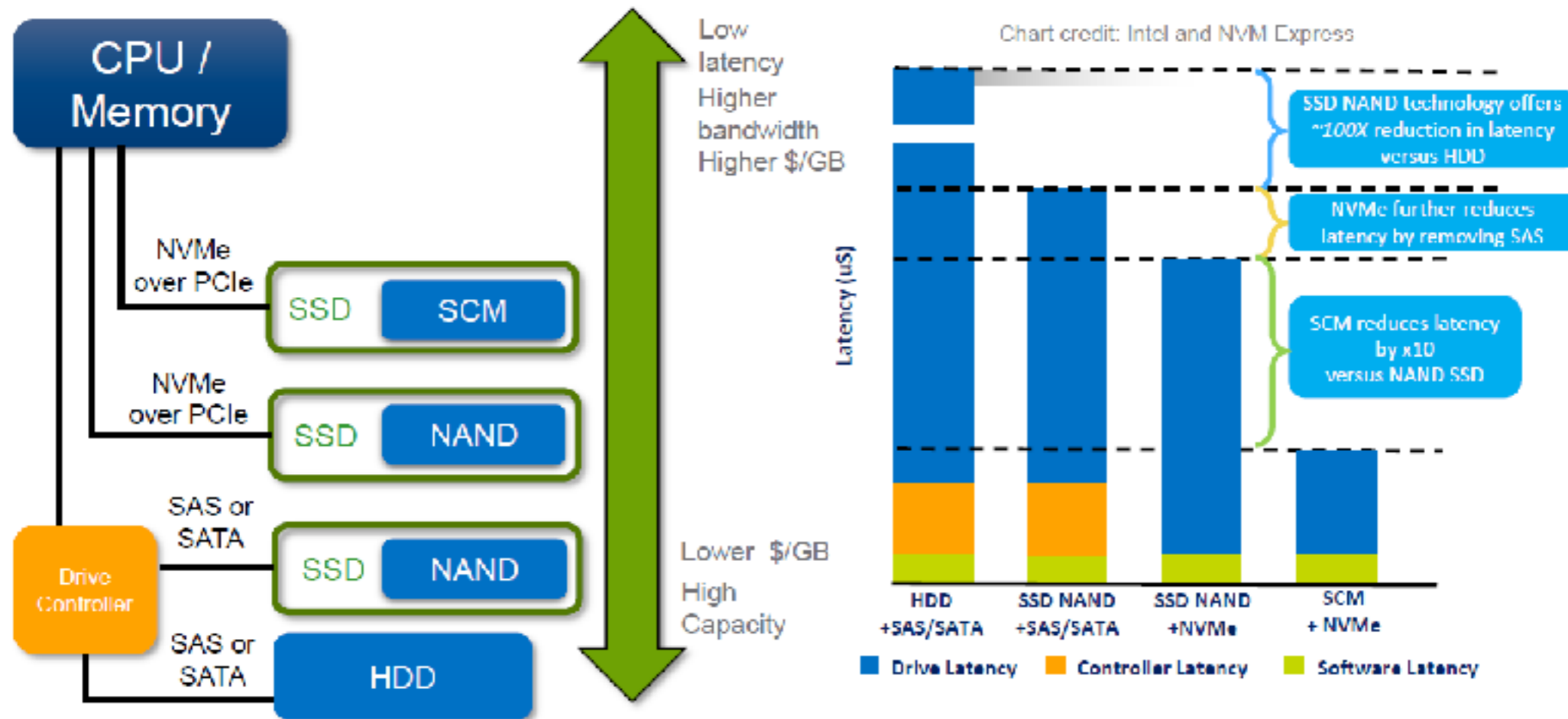
# Conceptualizing NVMe

	HDD + SCSI/SAS	Flash + SCSI/SAS	Flash + NVMe
			
Media	HDD: varying speed conveyer belts carrying data blocks (faster belts = lower seek time & latency)	Flash: all data blocks available at the same seek time & latency	Flash: all data blocks available at the same seek time & latency
Protocol	SCSI / SAS: pick & place robot w/tracked, single arm executing 1 command at a time, 1 queue	SCSI / SAS: pick & place robot w/single arm executing 1 command at a time, 1 queue	NVMe / PCIe: pick & place robot with 1000s of arms, all processing & executing commands simultaneously, over 64K queues

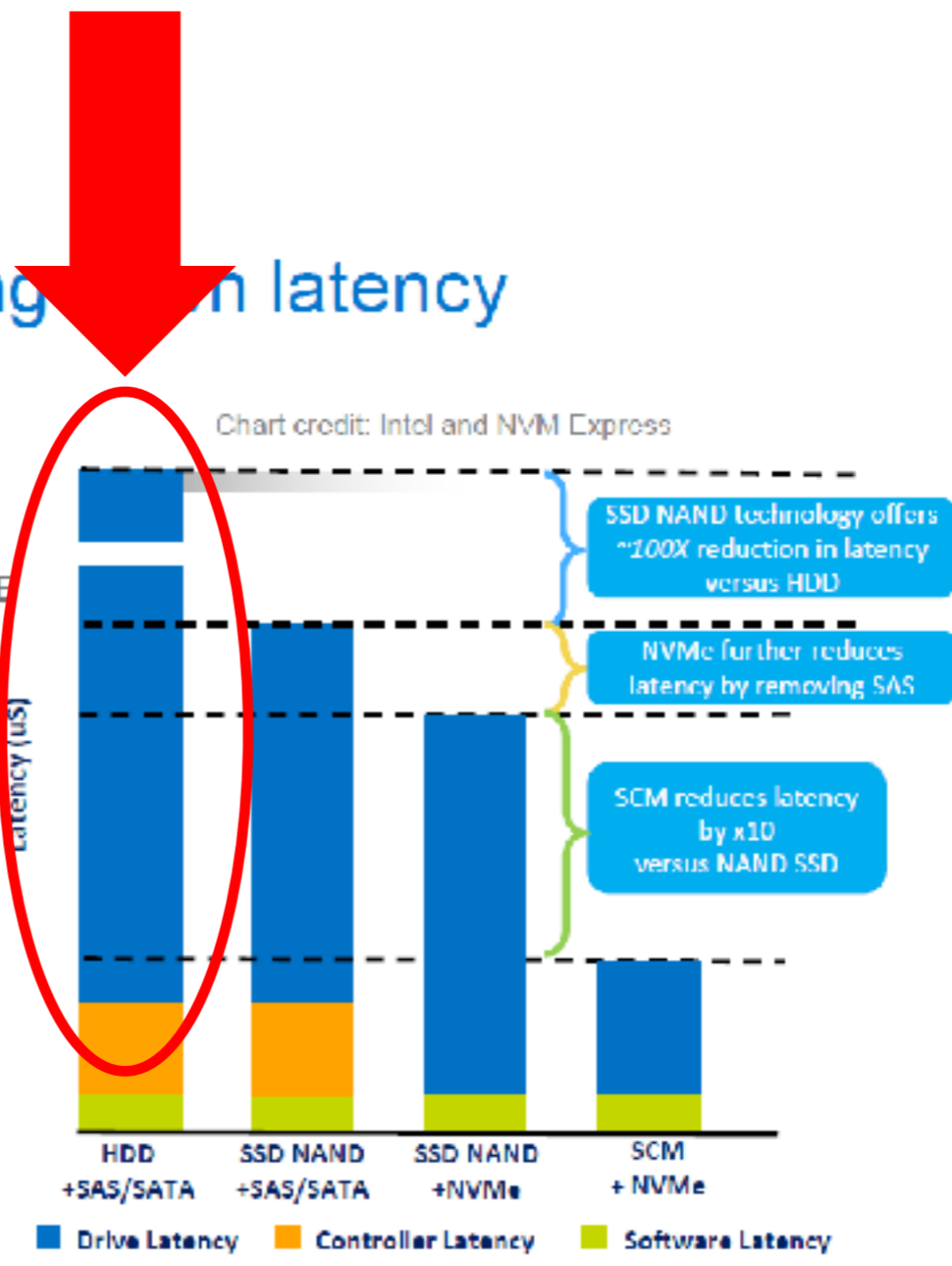
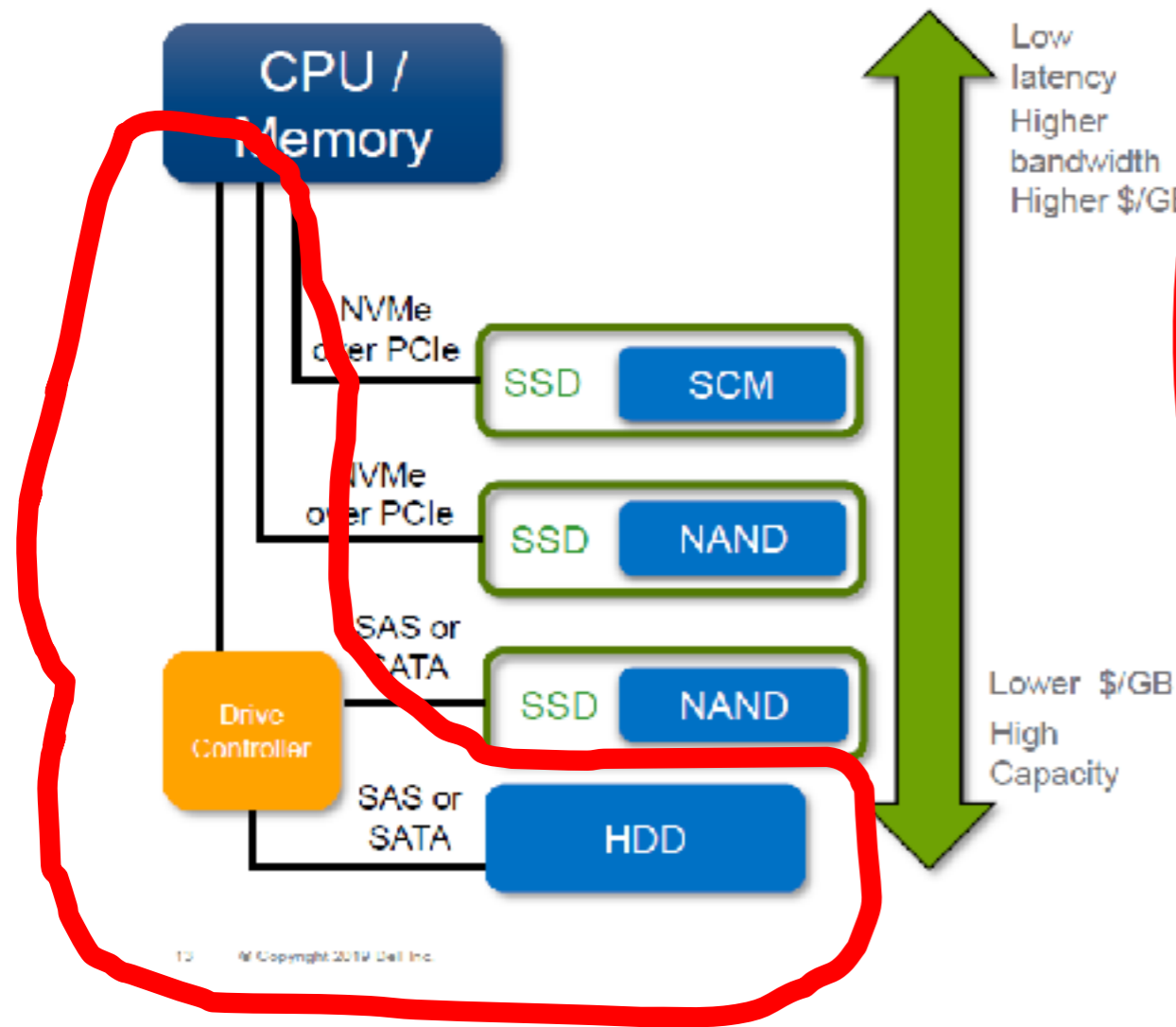

 Concept attribution:  
 J Meiz, "The Data Robot Parable"

# どこがチンタラ遅いんだ？

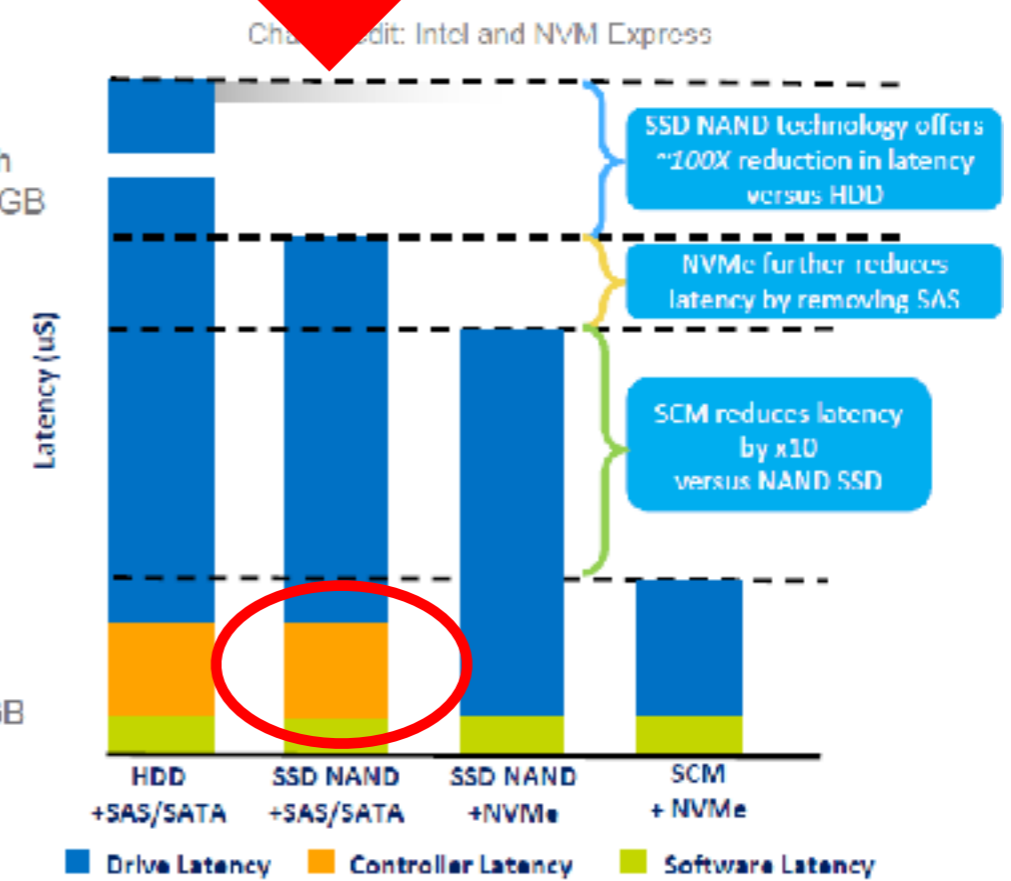
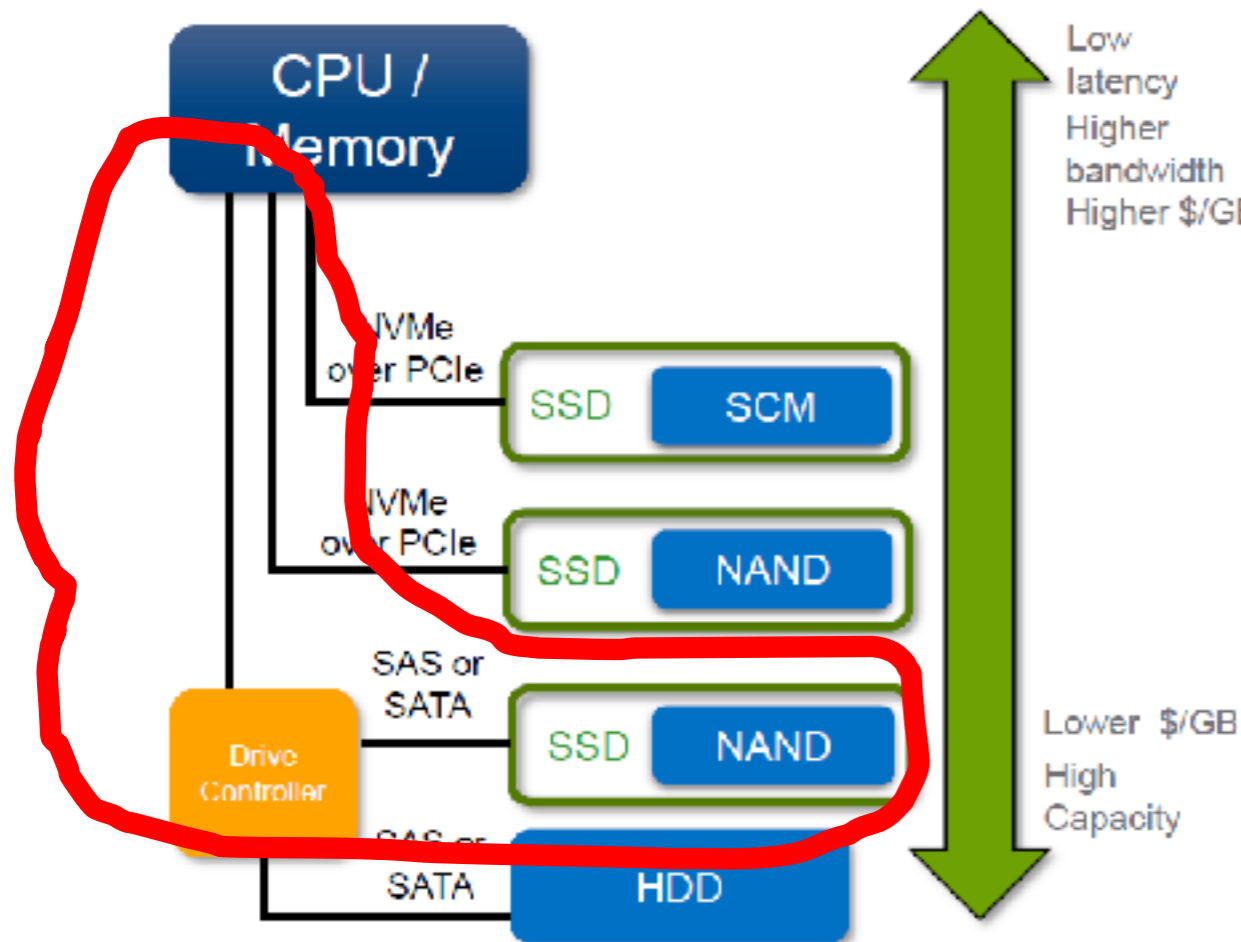
## NVMe Technical Basics: Driving down latency



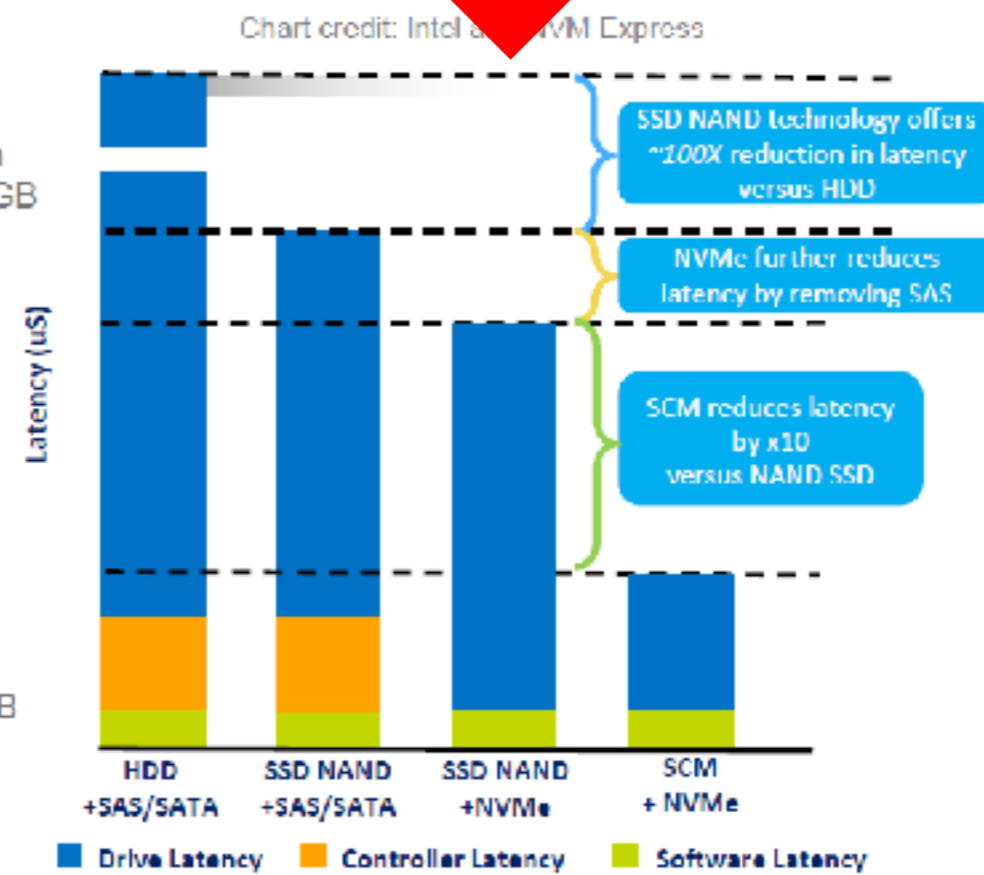
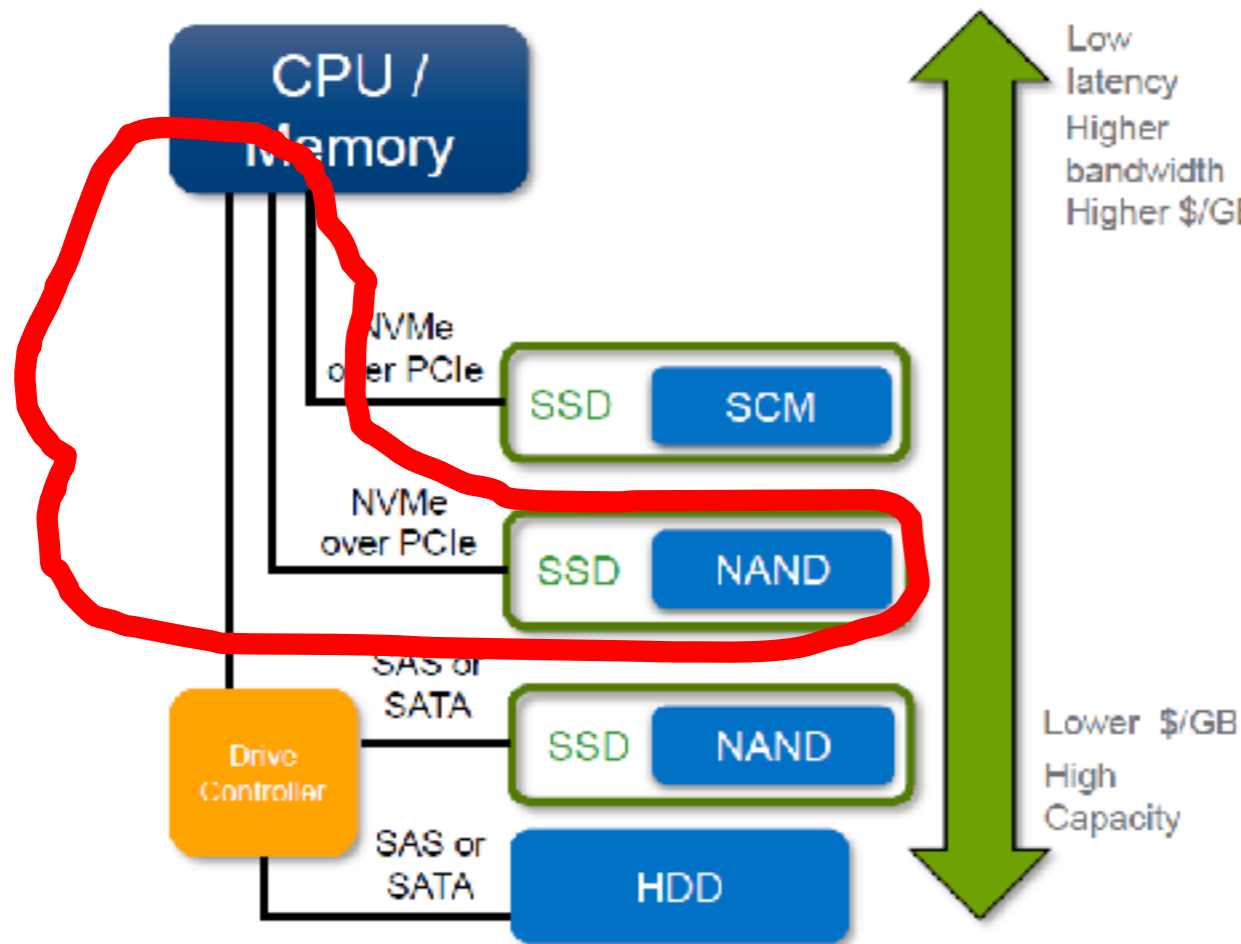
# NVMe Technical Basics: Driving down latency



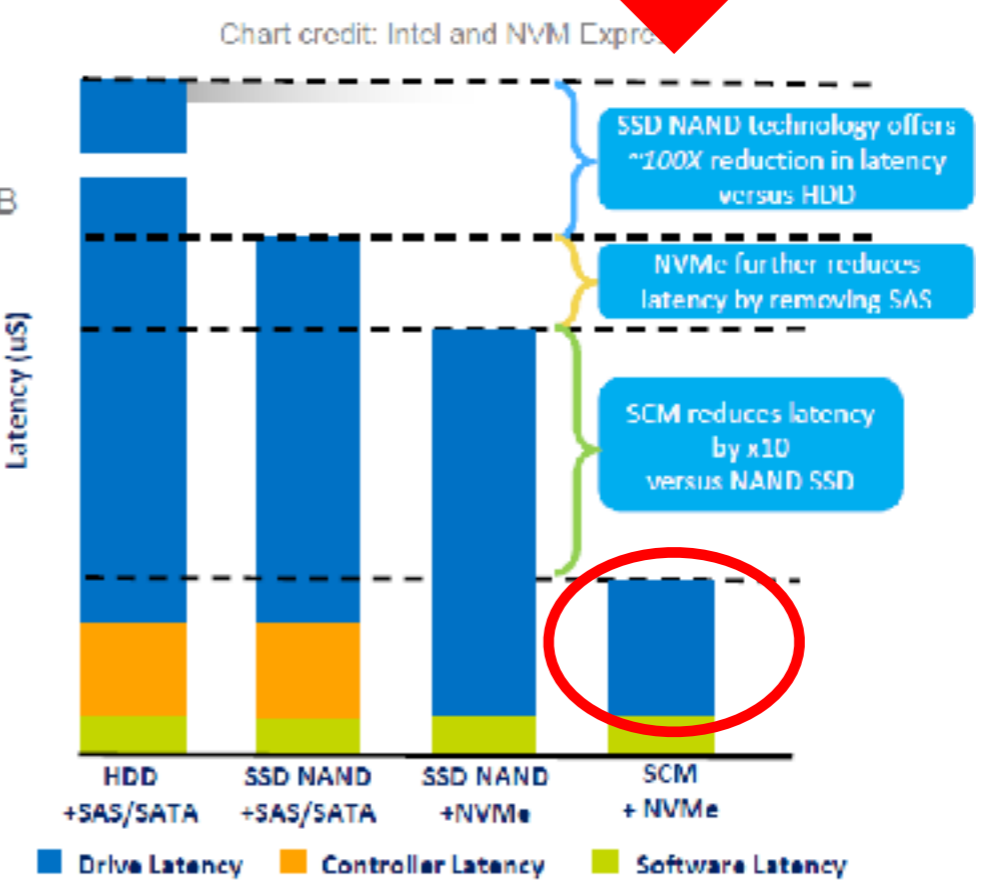
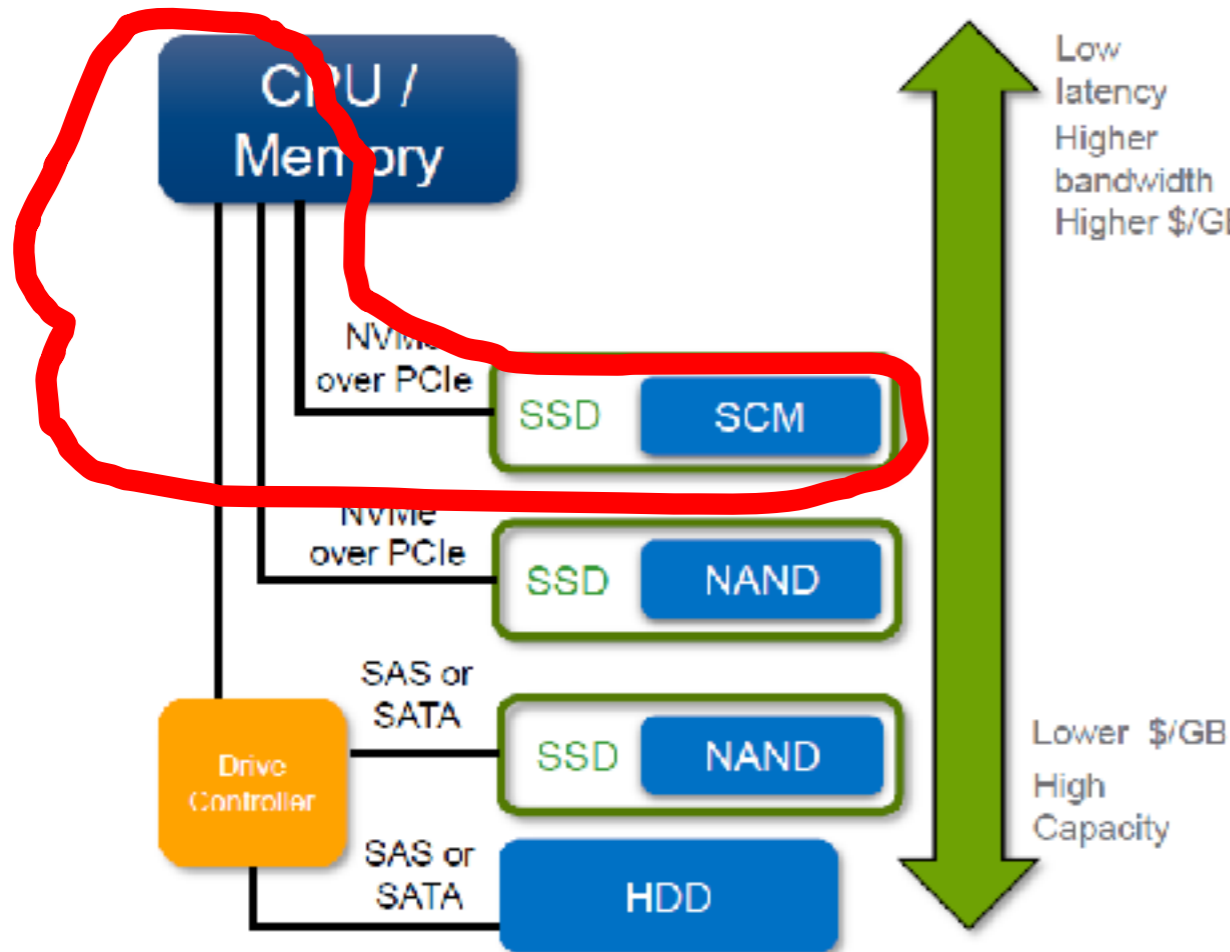
# NVMe Technical Basics: Driving down latency



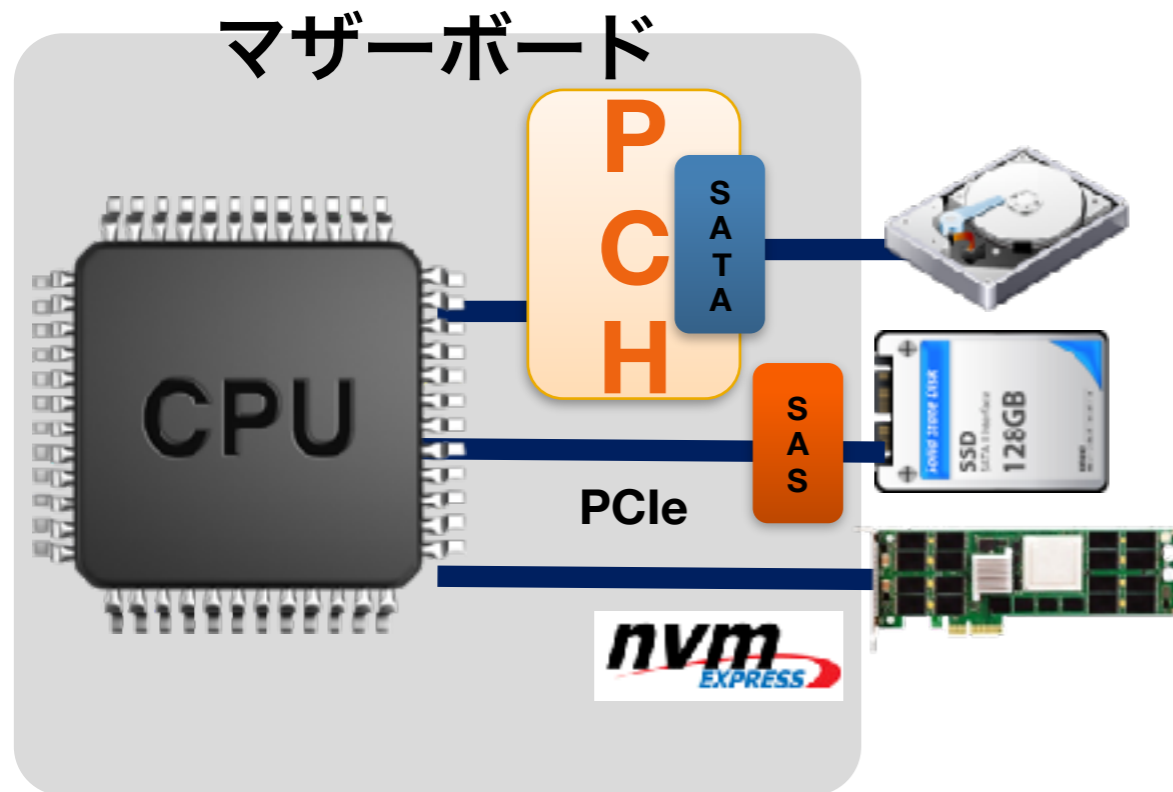
# NVMe Technical Basics: Driving down latency



# NVMe Technical Basics: Driving down latency



# NVMeってなんだ？



- NVMe は、レジスターインターフェイス～HW、OS、ソフトウェア間の通信の仕組みを標準化したもの。プロトコルと呼んでも良い
  - 従来のレジスターインターフェイスで対応する用語は、AHCI (Advanced Host Controller Interface)
  - NVMe Expressも当初は NVMeHCI という名前で検討が始まった
- SATA 3
  - 6Gbps (実効 600MB/s)
- SAS 3.0
  - 12Gb/s (実効1200MB/s)
  - 最近のUnityやXtremIO X2などは、SAS 3.0を利用
- PCIe Gen3 で 1レーン 1GB/s
  - PCIe Gen3 x4 だと、4GB/s
  - PCIe Gen3 x8 だと、8GB/s
- SATA はもともとHDD用に作られた。速度的にも進化してきたSSDには力不足

# 用語的には。 。 。

- **Non-Volatile Memory (NVM):**

不揮発メモリ

- **NVMe = NVM Express:**

不揮発メモリの速達(超訳)

- **NVMe-oF = NVMe over Fabrics:**

NVMeをネットワークで運ぶ



# NVMe を理解する

## PCI Express SSDの業界標準インターフェイス

不揮発メモリ用に設計・構築された

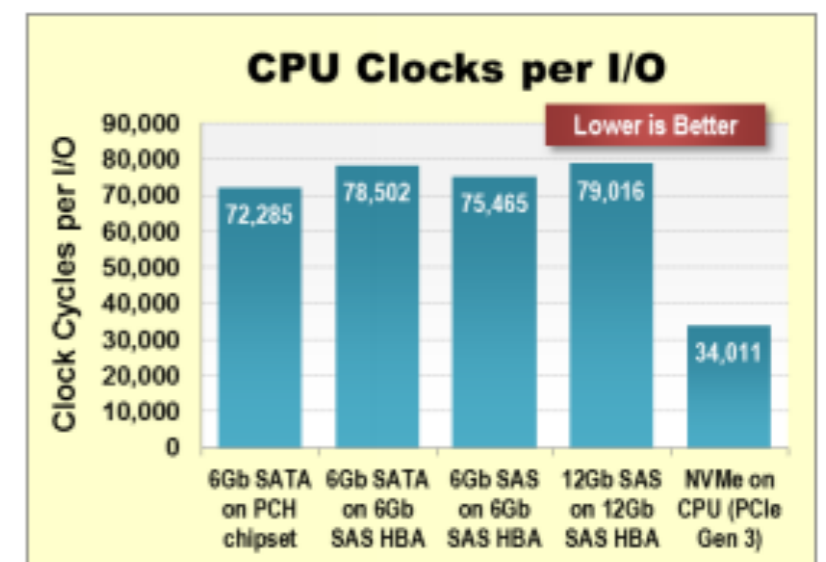
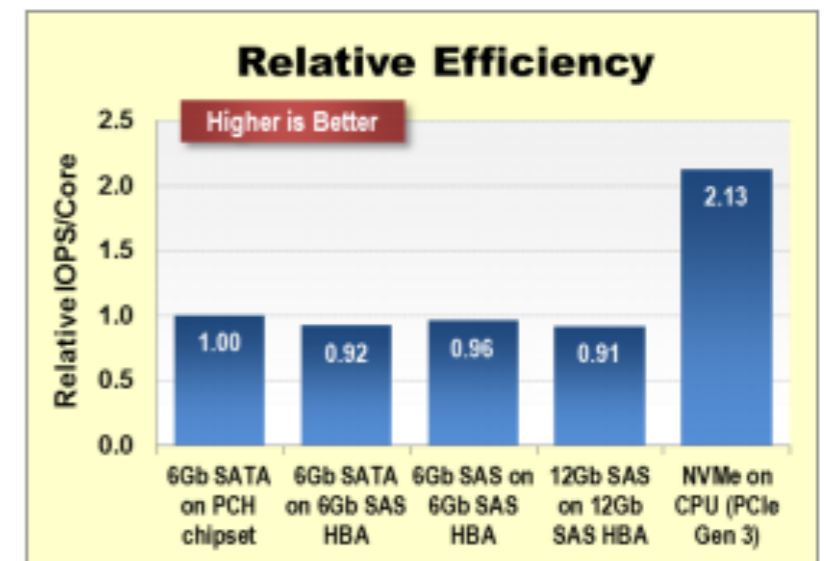
(SASやSATAがその生い立ちとして保持している) HDDに関する「遺産」を排除

不揮発メモリに最適化

低レイテンシで、オーバーヘッドの小さい、新たなストレージスタック

効率的でスケラブル

合理化されたプロトコルとコマンドセット、I/Oごとに必要となるクロックサイクルを削減



# NVMe を理解する

## PCI Express SSDの業界標準インターフェイス

不揮発メモ

(SASやSATAが  
する「遺産」を扱

不揮発メモ

低レイテンシで、  
ジスタック

効率的です

合理化されたプロ

なるクロックサイクルを削減

Flashを使い倒すには  
NVMeが必要



# NVMeをワンスライドでいうと

- PCIe経由のSSDアクセスに特化した
- ナウイコマンドセットの
- すごい並列化と深いキューで処理するやつ

**NVMe is for local PCIe storage access!**

ここまではローカル接続のストレージのはなし。

(なんでNVMeが必要なのかの説明)

一方、リモートストレージ接続の観点でみると、、、

# Agenda

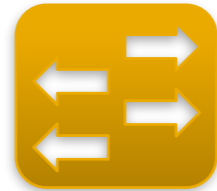
1. Storage Networking 基礎知識
2. 知っておきたい○○ over IP/Ethernet

# ボトルネックはどこだ？

CLIENTS /  
HOSTS



FRONT-END  
CONNECT



STORAGE  
CONTROLLER

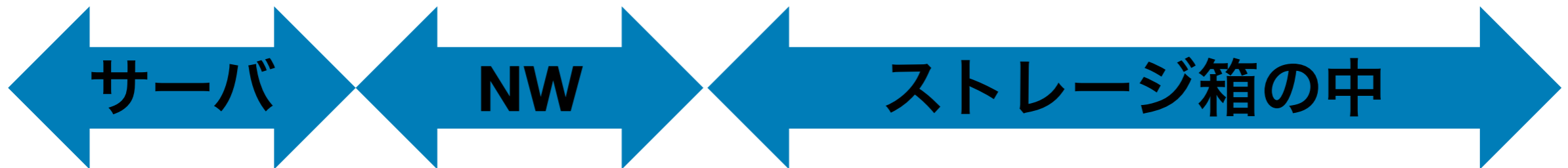


BACK-END  
CONNECT



PHYSICAL  
STORAGE

7200 RPM



ボトルネックはどこだ? → ちょっと前までディスクだった

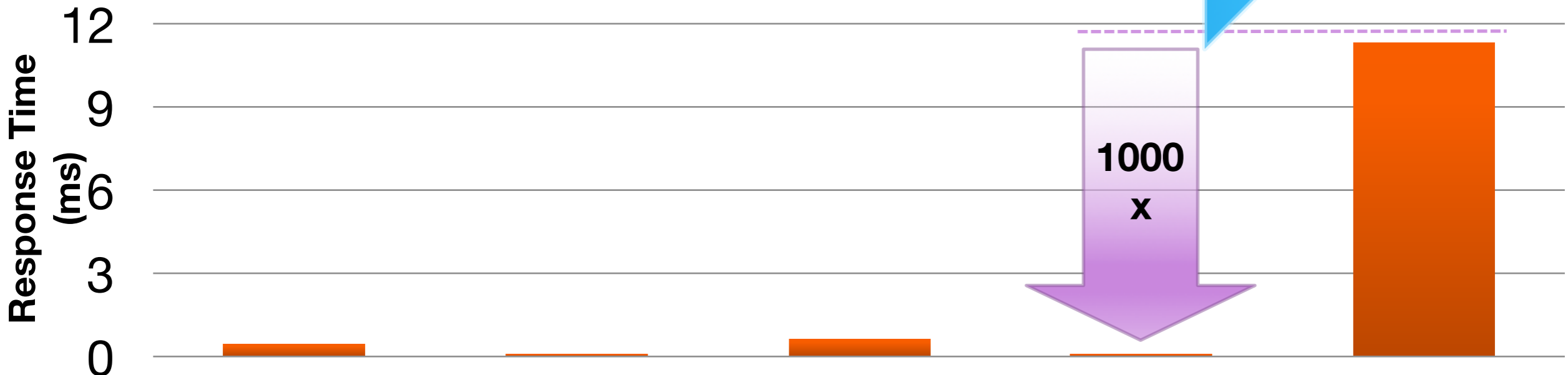
CLIENTS /  
HOSTS

FRONT-END  
CONNECT

STORAGE  
CONTROLLER

BACK-END  
CONNECT

回転ディ  
スク  
STORAGE



# ボトルネックはどこだ? →クライアント側になった

CLIENTS /  
HOSTS

FRONT-END  
CONNECT

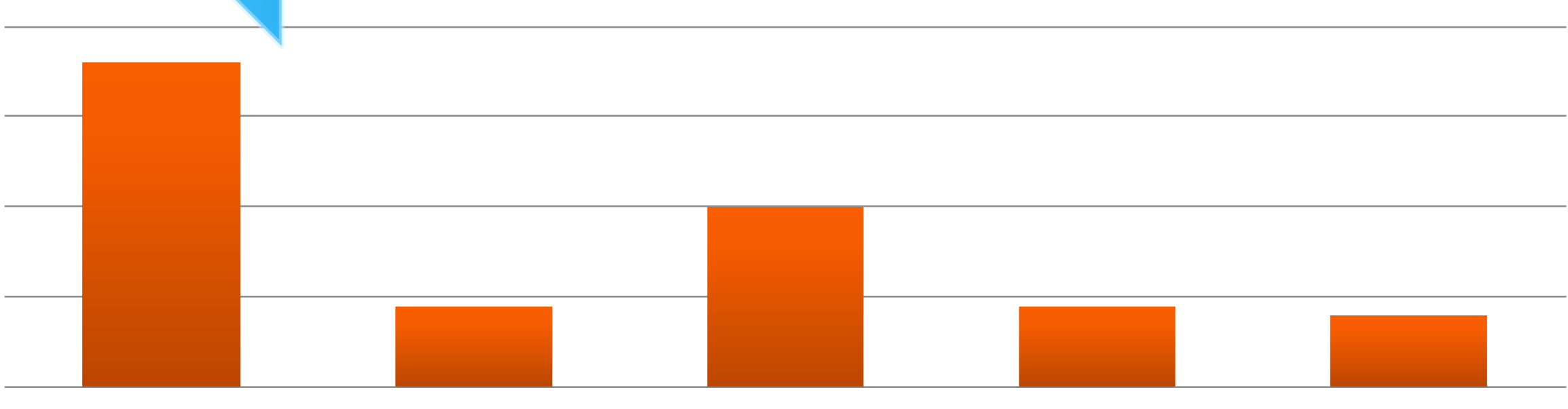
STORAGE  
CONTROLLER

BACK-END  
CONNECT

**Flash**  
STORAGE



Response Time  
(ms)





# より広帯域が求められる時代に

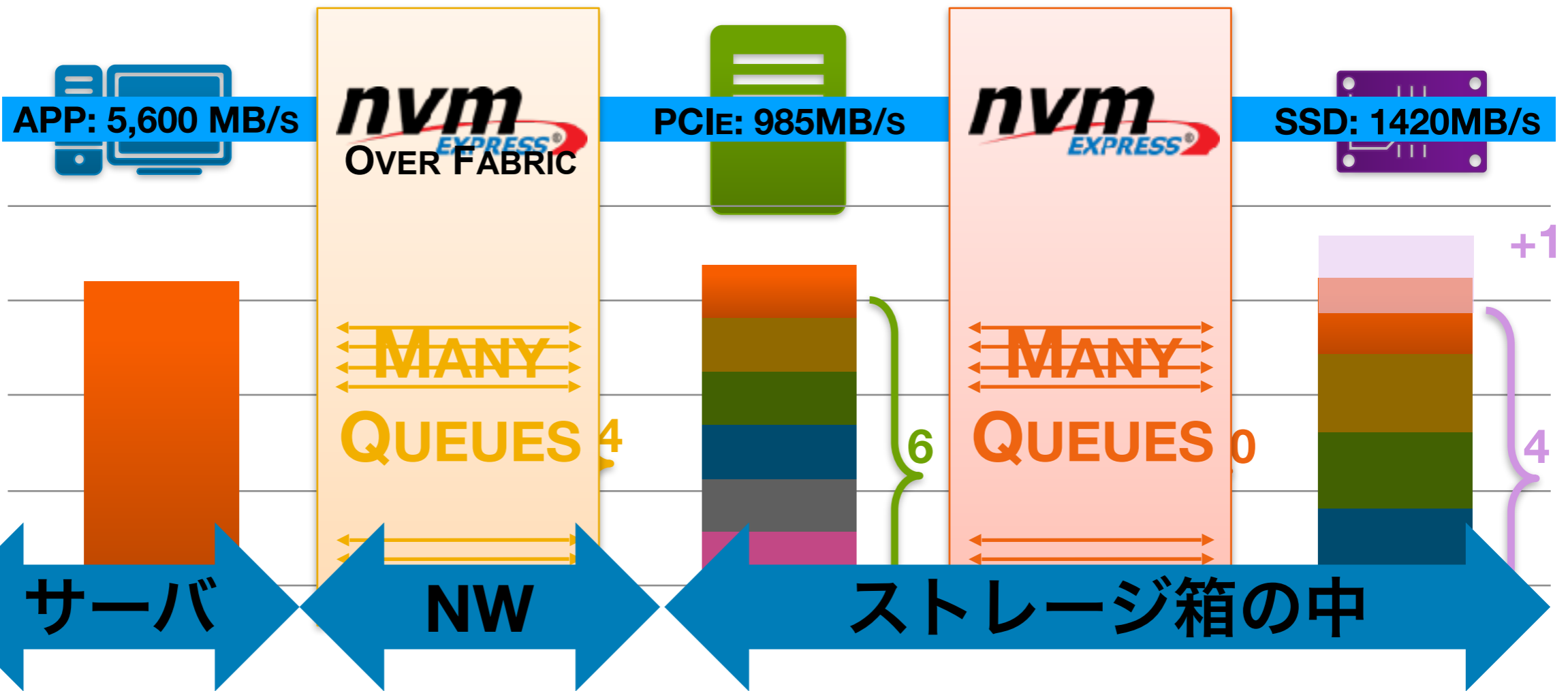
CLIENTS /  
HOSTS

FRONT-END  
CONNECT

STORAGE  
CONTROLLER

BACK-END  
CONNECT

PHYSICAL  
STORAGE



# より広帯域が求められる時代に

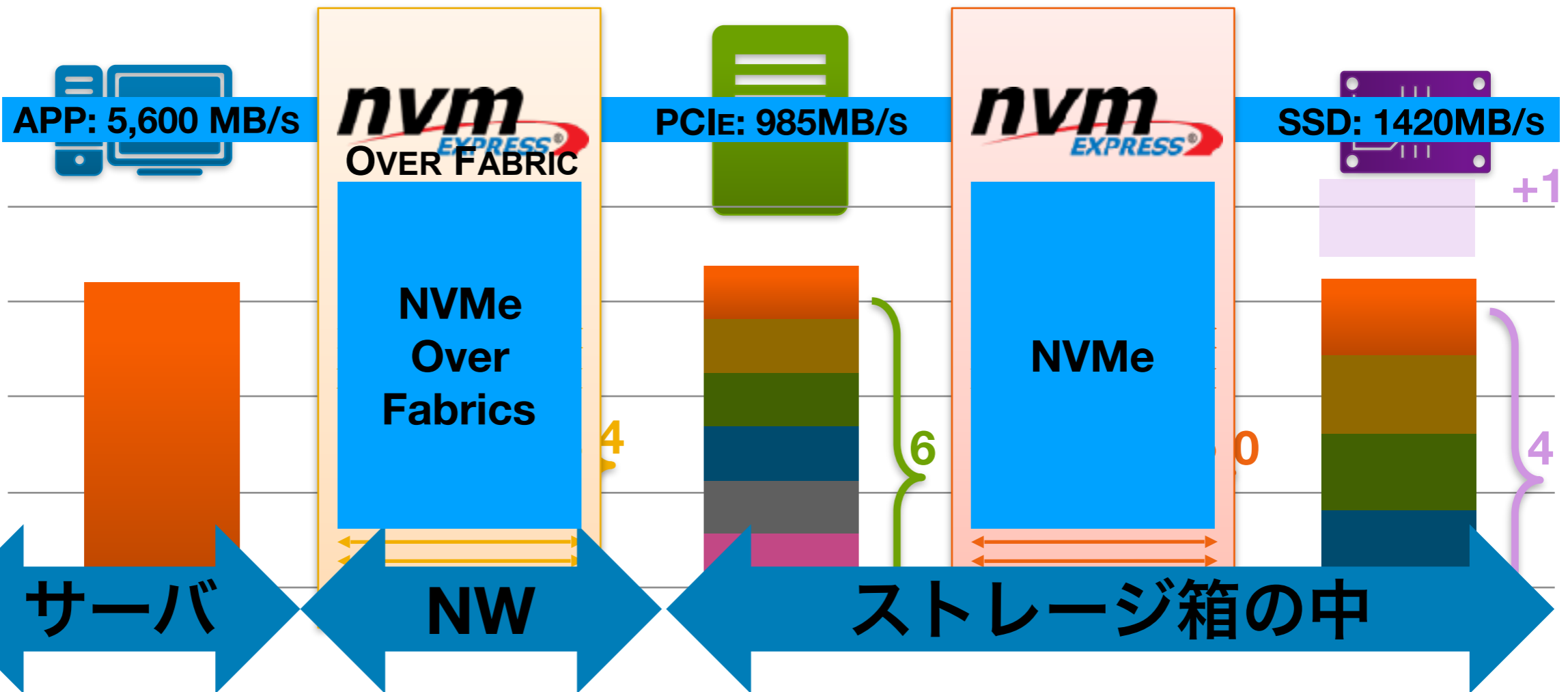
CLIENTS /  
HOSTS

FRONT-END  
CONNECT

STORAGE  
CONTROLLER

BACK-END  
CONNECT

PHYSICAL  
STORAGE



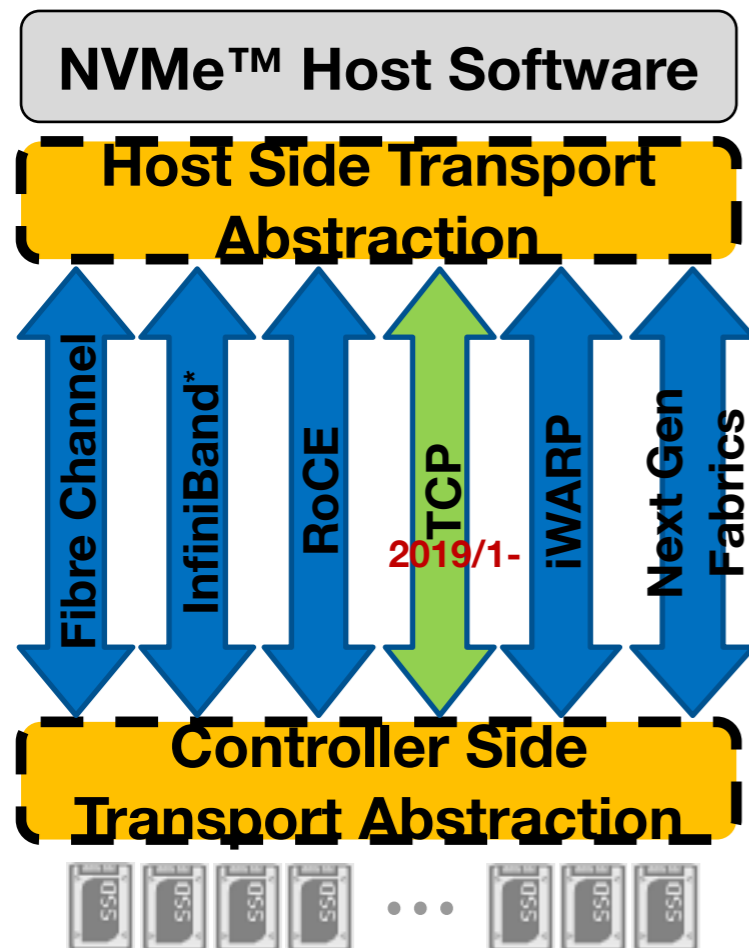
**Spindle時代: Media+Drive I/F速度 >> NW(~1G)**

**Flash時代 : Media+SATA/SAS ≈ NW(10G)**

**SCM+NVM時代: Media+NVM << NW?  
(100G)**

**Networkの選択肢がより多彩な時代に**

# NVMe over Fabrics



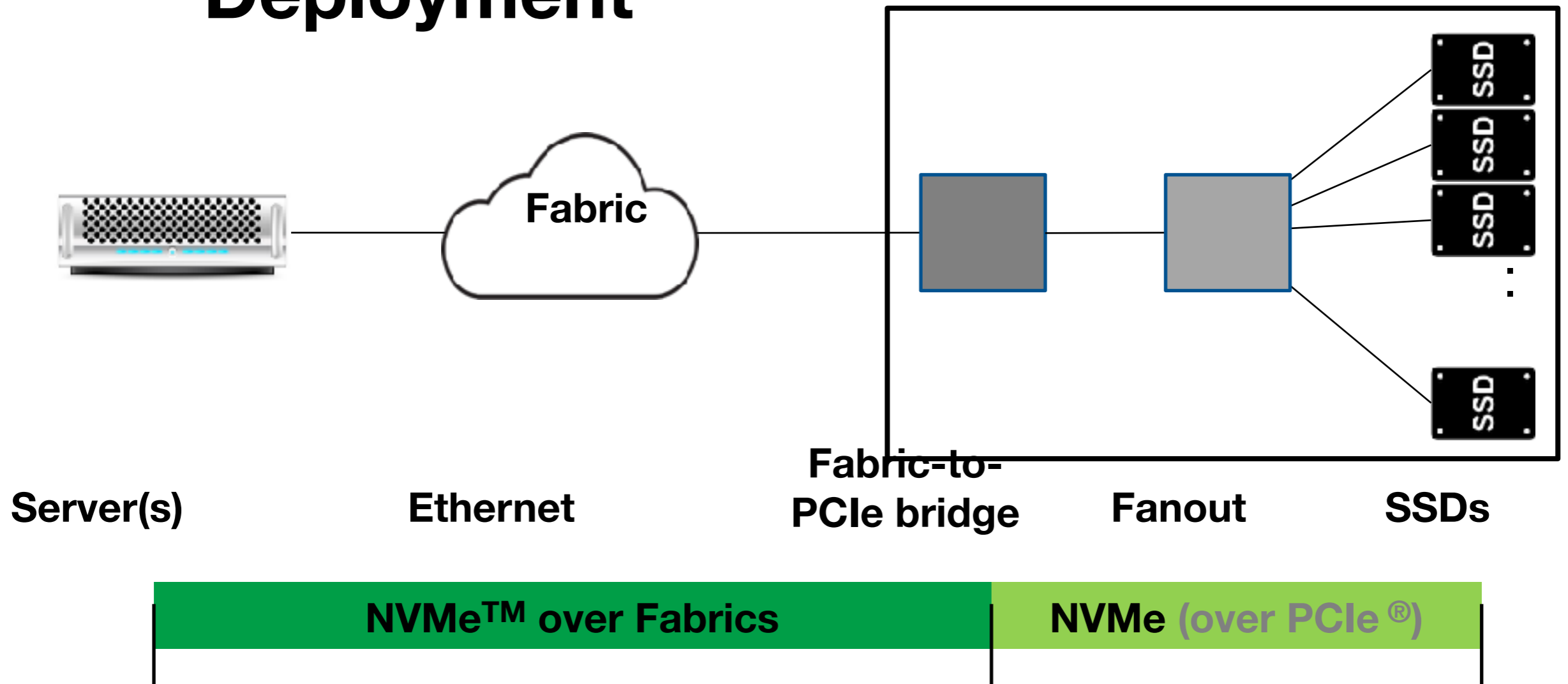
- NVMe over Fabrics
  - NVMeブロックストレージ・プロトコルを、ストレージネットワーク・ファブリックに拡張する
  - 2016年6月に仕様 1.0 が公開
  - NVMeデバイスを大量に扱う、（データセンター内で）離れた場所のNVMeデバイスにアクセスする、、、等
- 2019年1月に、あらたにNVMe over TCP(NVMe/TCP) が批准された
  - TP 8000 として
  - NVMe-oF 1.1 の仕様ドキュメントに統合予定
    - Later 2019 ?

[http://www.nvmexpress.org/wp-content/uploads/NVMe\\_Over\\_Fabrics.pdf](http://www.nvmexpress.org/wp-content/uploads/NVMe_Over_Fabrics.pdf)

<https://nvmexpress.org/answering-your-questions-nvme-tcp-what-you-need-to-know-about-the-specification-webcast-qa/>

# Typical NVMe-oF™ Deployment

JBOF / EBOF / FBOF



# 「ファブリック (Fabrics)」ってなに？

- **Fibre Channel**
  - 昔も今も広く使われてる。
- **InfiniBand (RDMA)**
  - 広帯域、低遅延をいかして、組み込み用途、特にHPC
  - ストレージ内部だけでなくサーバとの接続にも
- **IP/Ethernet RDMA: RoCEv2 , iWARP**
  - RoCEv2: 発音ロッキー、UDP/IPベース、**ロスレス** Converged Ethernet推奨
    - V1はL2、v2と互換性なし
  - iWARP: TCP/IPベース、実質的にHW実装が必要、普及していない
  - NICインプリが推奨
- **IP/Ethernet non-RDMA: TCP/IP**
  - ソフトウェアベース、NICのTCPオフロードで高速化



**New!**

# 「ファブリック (Fabrics)」 ってなに？

- **Fibre Channel**

- 昔も今も広く使われてる。

- **InfiniBand (RDMA)**

- 広帯域、低遅延をいかして、組み込み用途、特にHPC
- ストレージ内部だけでなくサーバとの接続にも

- **IP/Ethernet RDMA: RoCEv2 , iWARP**

- RoCEv2: 発音ロッキー、UDP/IPベース、**ロスレス** Converged Ethernet推奨
  - V1はL2、v2と互換性なし
- iWARP: TCP/IPベース、実質的にHW実装が必要、普及してない
- NICインプリが推奨

- **IP/Ethernet non-RDMA: TCP/IP**

- ソフトウェアベース、NICのTCPオフロードで高速化



**New!**

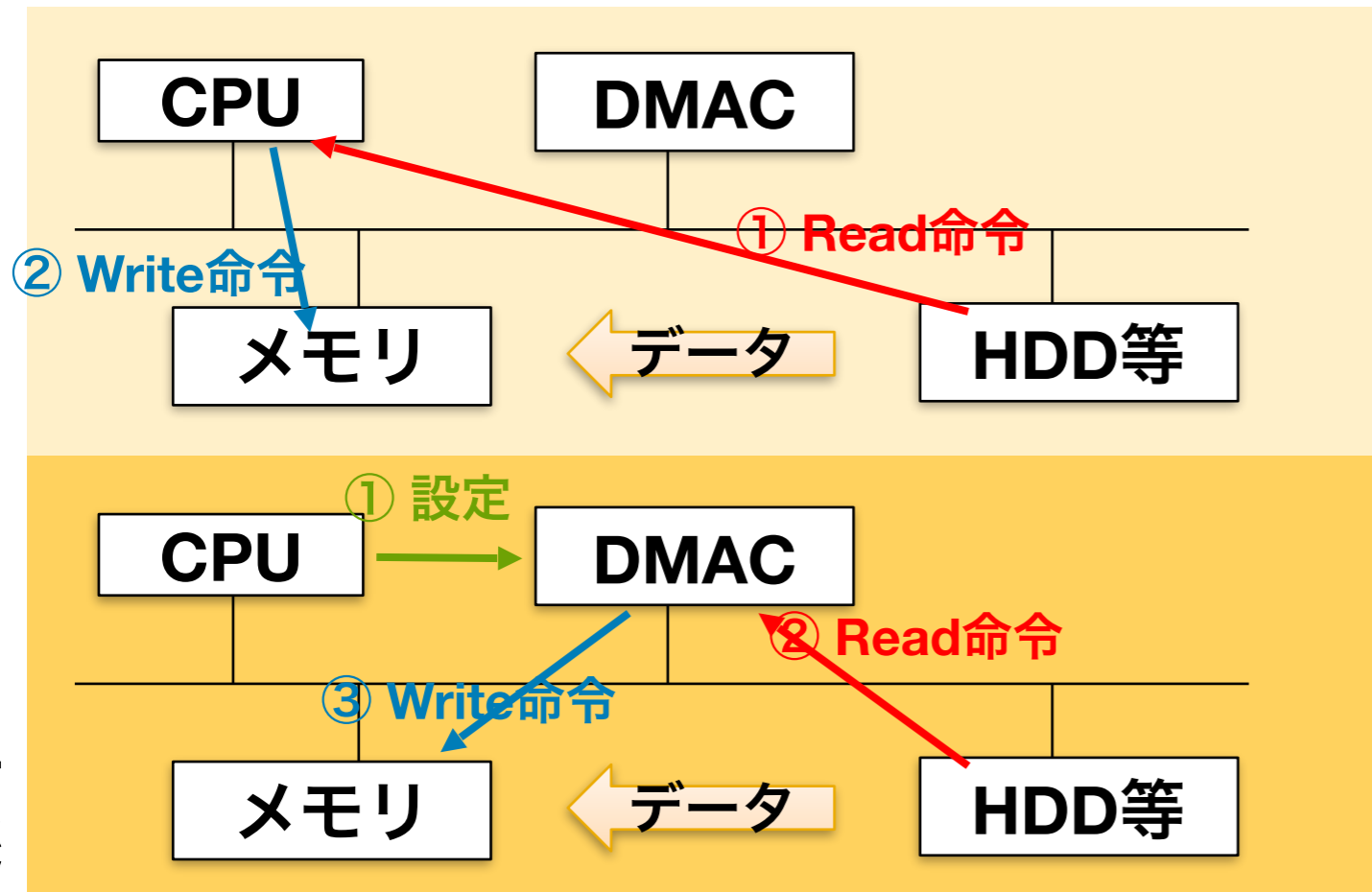


**RDMAについて補足**

# RDMAのまえにDMA

- DMA(Direct Memory Access)方式とは、CPUを介さずに「直接」メモリにアクセスすることをいう
- CPUの代わりにしてくれるのが「DMAコントローラ」というデバイスで、略してDMAC（でいーまっく）とも呼ばれたりする
- DMACは、チップセット（サウスブリッジに含まれています）

PIO転送



DMA転送

Programmed I/O ↔ DMA

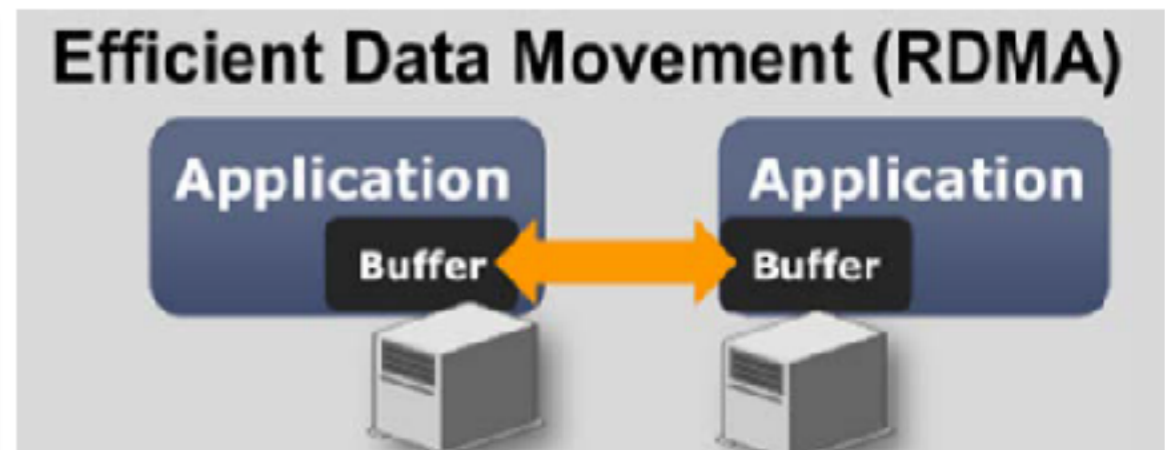
## 利用例

- HDD, USBメモリ、グラフィックス他多数

# RDMA

- コンピューティングにおいて、**Remote Direct Memory Access(RDMA, リモートDMA)**とは、あるコンピュータの主記憶から、異なる（リモートの）コンピュータの主記憶への DMA 転送を行うことである。RDMAでは両コンピュータのオペレーティングシステムに入ることなく転送が行われる。これにより高スループット、低レイテンシの通信を行うことができるため、特に大規模並列のコンピュータ・クラスターにおいて有用である。
- 対象のアプリケーションメモリとオペレーティングシステム中のデータバッファの間でのデータの転送をなくし、CPUやキャッシュを用いることなく、あるいはコンテキストスイッチを行うこともなく、他の処理と並列にデータを転送できる。RDMAリードまたはライト要求を行うアプリケーションでは、アプリケーションメモリのデータは直接ネットワークに配送されるため、レイテンシが削減され高速なデータ転送が可能となる。

異なるコンピュータのメモリ空間同士で（リモートで）、DMA転送を行うということ



# 「ファブリック(Fabrics)」ってなに？

- Fibre Channel
- InfiniBand
- RoCE
  - RDMA over Converged Ethernet
  - RoCEv1 Link Layer Protocol
  - RoCEv2 internet Layer Protocol
  - v1とv2では、プロトコルヘッダも異なる
  - 従来のEthernetも利用可能ではあるが、標準として採用されていない
- iWARP
  - Internet Wide Area RDMA Protocol
  - TOEのようなHWでの実装でないといけない
  - RDMAコンソーシアムによって策定されている
- TCP

**8G/16Gが主流**

**32Gも徐々に**

**DC内**

**実はメタル線も可**



rd)  
奨

# 「ファブリック (Fabrics)」ってなに？

- Fibre Channel
- InfiniBand
- RoCE
  - RDMA over Converged Ethernet (RoCE)
  - RoCEv1 Link Layer Protocol
  - RoCEv2 internet Layer Protocol
  - v1とv2では、プロトコルヘッダも異なる
  - 従来のEthernetも利用可能ではあるが、遅延を低減させる必要があるとする
- iWARP
  - Internet Wide Area RDMA Protocol
  - TOEのようなHWでの実装でない
  - RDMAコンソーシアムによって策定
- TCP

**HPC以外にストレージ接続も**

**10~40Gbpsが主流**

**低遅延(10GbEの1/10)**

**DC内**



# 「ファブリック (Fabrics)」ってなに？

- Fibre Channel
- InfiniBand
- **RoCE**
  - RDMA over Converged Ethernet, or RoCE (発音 : "rocky")
  - RoCEv1 Link Layer Protocol
  - RoCEv2 internet Layer Protocols(ルーティングできる)
  - v1とv2では、プロトコルヘッダも異なり互換性がない(RNICは下位互換あり?) (RNIC ::= RDMA network interface card)
  - 従来のEthernetも利用可能ではあるが、ロスレスの Converged Ethernet を必須(**shall**)とする
- **iWARP**
  - Internet Wide Area RDMA Protocol (iWARP) とは、RDMA over TCPを実現する通信プロトコル群の総称である
  - TOEのようなHWでの実装でないと、速度的に問題あり
  - RDMAコンソーシアムによって策定された標準が Internet Engineering Task Force (IETF) によって改版された
- TCP

# 「ファブリック (Fabrics)」ってなに？

- Fibre Channel
- InfiniBand
- RoCE
  - RDMA over Converged Ethernet, or RoCE (発音 : "rocky")
  - RoCEv1 Link Layer Protocol
  - RoCEv2 internet Layer Protocols(ルーティングできる)
  - v1とv2では、プロトコルヘッダも異なり互換性がない(RNICは下位互換あり?) (RNIC ::= RDMA network interface card)
  - 従来のEthernetも利用可能ではあるが、ロスレスの Converged Ethernet が必須(Shall)とする
- **iWARP**
  - Internet Wide Area RDMA Protocol (iWARP) とは、RDMA over TCPを実現する通信プロトコル群の総称である
  - TOEのようなHWでの実装でないと、速度的に問題あり
  - RDMAコンソーシアムによって策定された標準が Internet Engineering Task Force (IETF) によって改版された
- TCP

# 余談:shall

## 7.1 Transport Requirements

- The NVMe Transport **shall provide reliable delivery of capsules between a host and NVM subsystem (and allocated controller) over each connection**. The NVMe Transport may deliver command capsules in any order on each queue except for I/O commands that are part of fused operations (refer to section 4.10 of the NVMe Base specification).

信頼性のあるトランスポートが必須

RFC2111

1. 「しなければならない (MUST)」

この語句、もしくは「要求されている (REQUIRED)」および「**することになる (SHALL)**」は、その規定が当該仕様の**絶対的な 要請事項**であることを意味します。

もっと余談: 「Shall we dance?」 「I shall return」



# NVMe over Ethernet – longer term options

RoCE (RDMA over Converged Ethernet) – ベストなオペレーションには、ロスレスネットワーク（特別なハードウェア）の用意が必要。（Mellanox and Emulexがサポート）

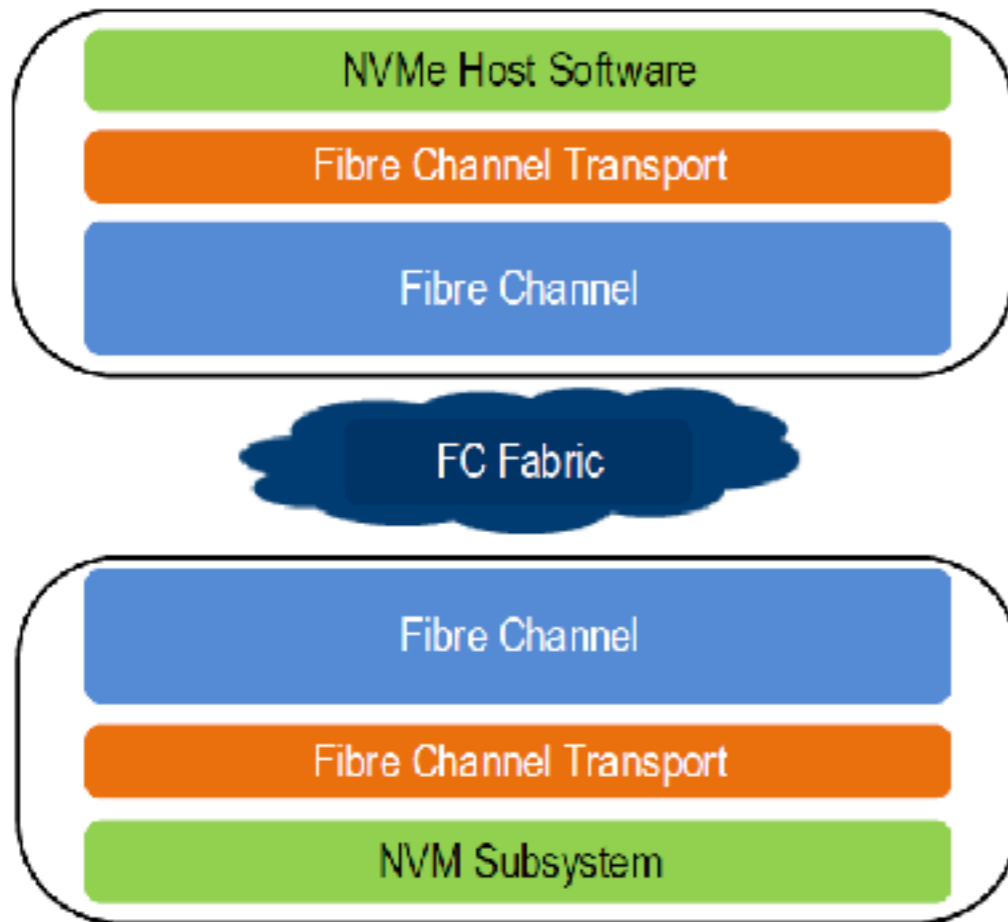
iWARP (Internet Wide-Area RDMA Protocol) – 複雑なハード/ソフトのスタックでよりCPUリソースの消費と、現時点では10Gbのみ。（Chelsio and Intelがサポート）

NVMe over TCP – 標準のスイッチとシンプルなTCPスタックを利用。プロポーザルが内部で批准されたばかり。NVMe oFの V1.1 に盛り込まれる予定

# 2種類の NVMe over Fabrics

## Over Fibre Channel

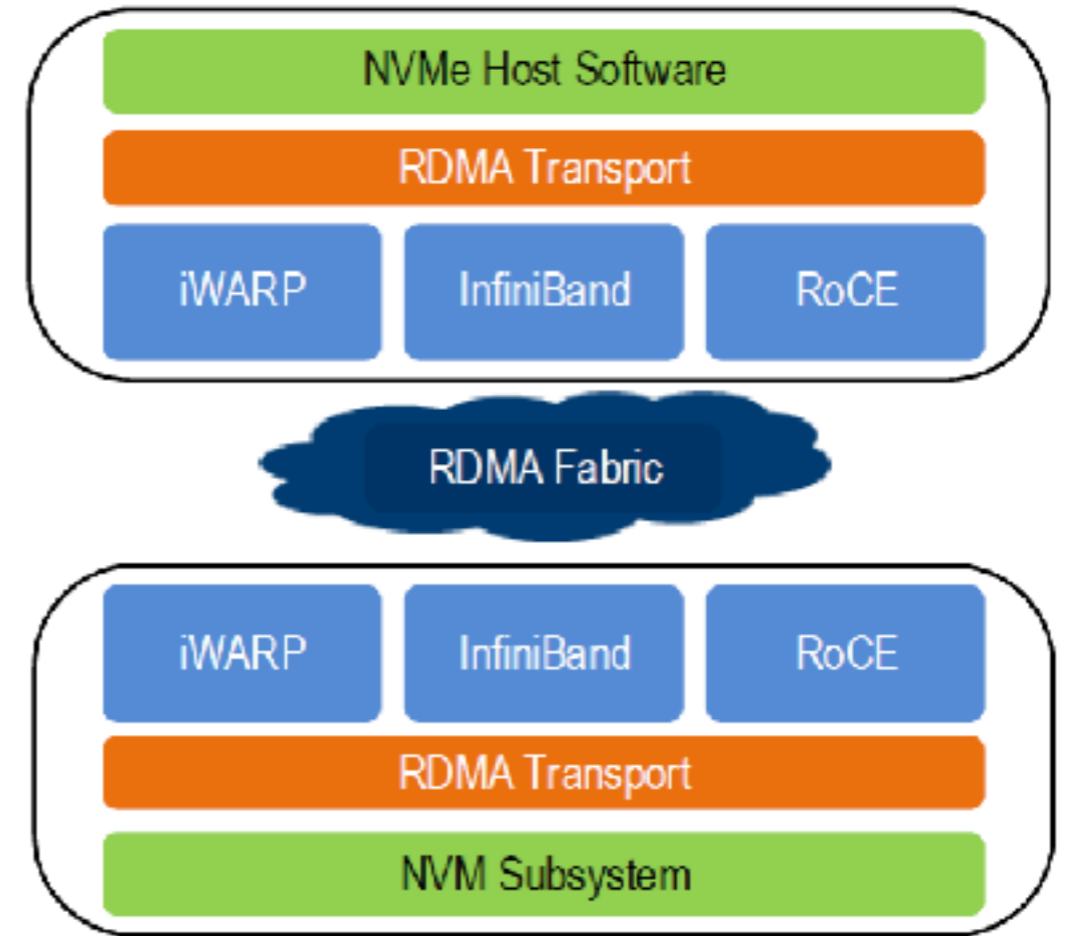
Figure 36: Fibre Channel Transport Protocol Layers



NVMe  
over  
TCPも  
こちら  
と同様

## Over RDMA

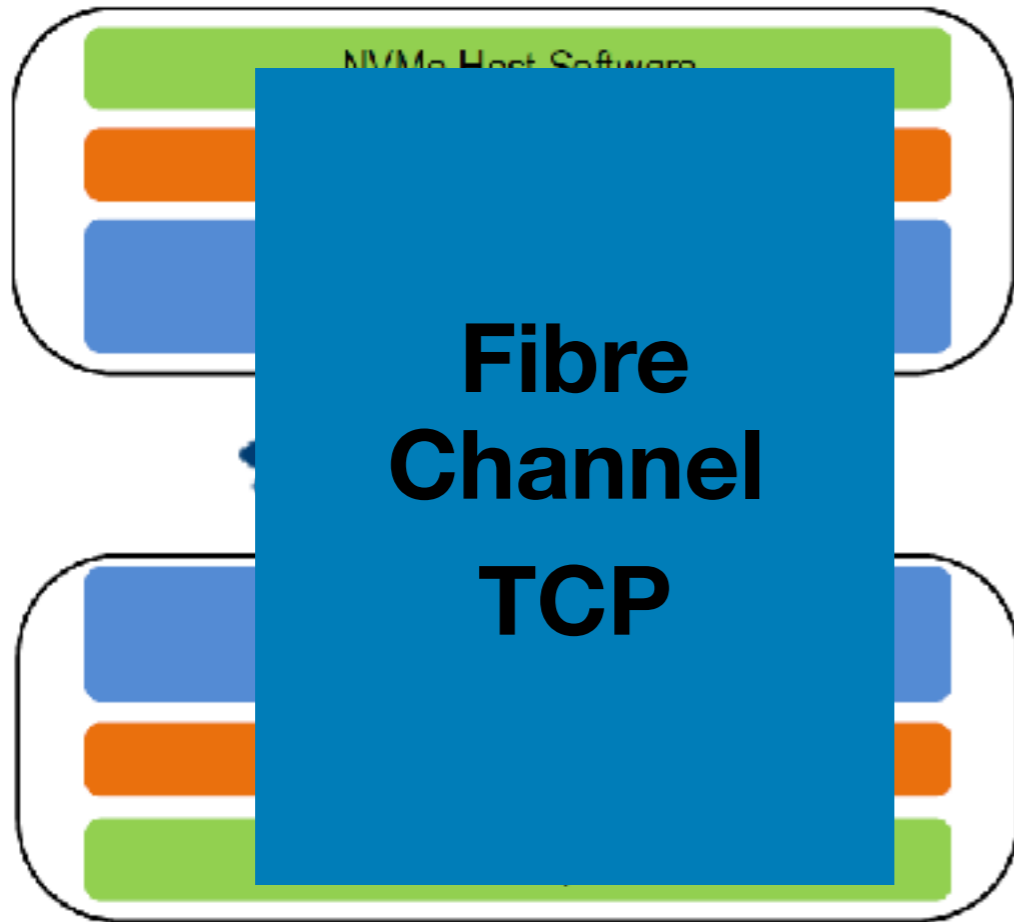
Figure 38: RDMA Transport Protocol Layers



# 2種類の NVMe over Fabrics

## Over Fibre Channel

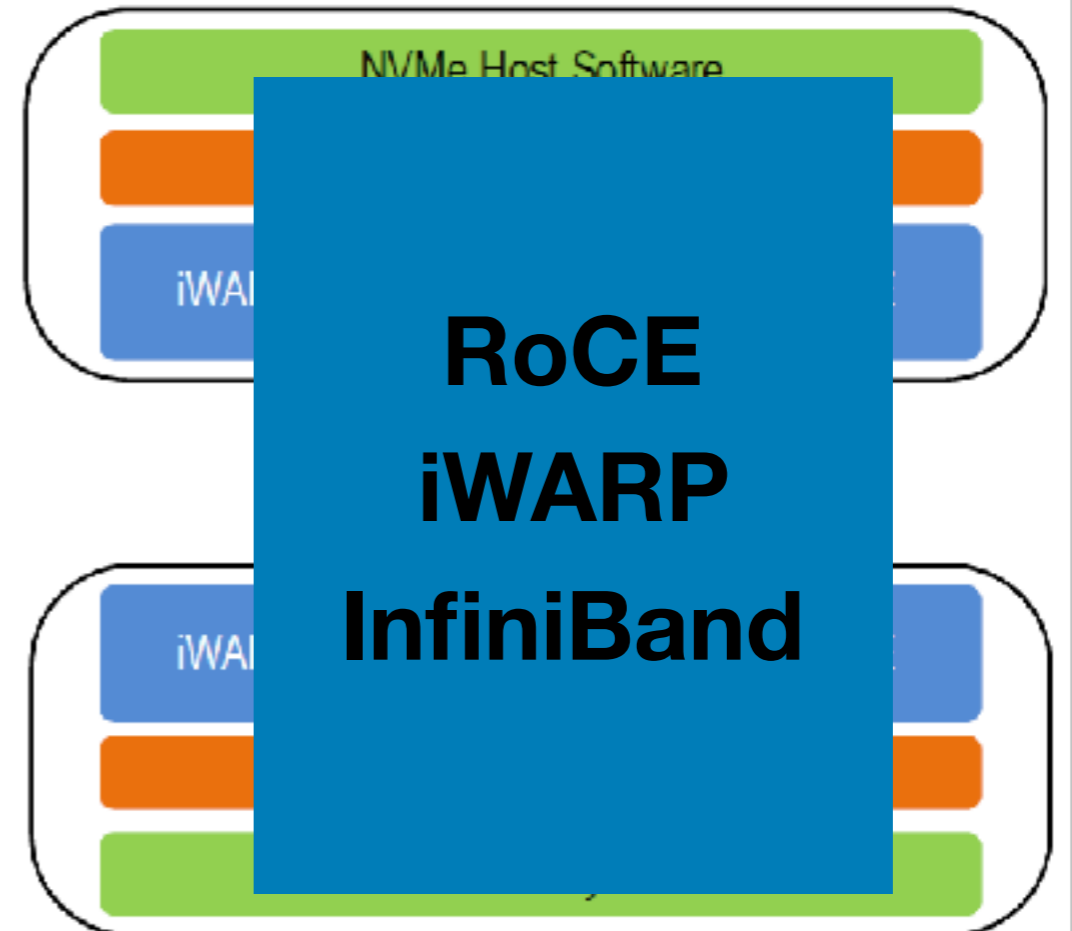
Figure 36: Fibre Channel Transport Protocol Layers



NVMe  
over  
TCPも  
こちら  
と同様

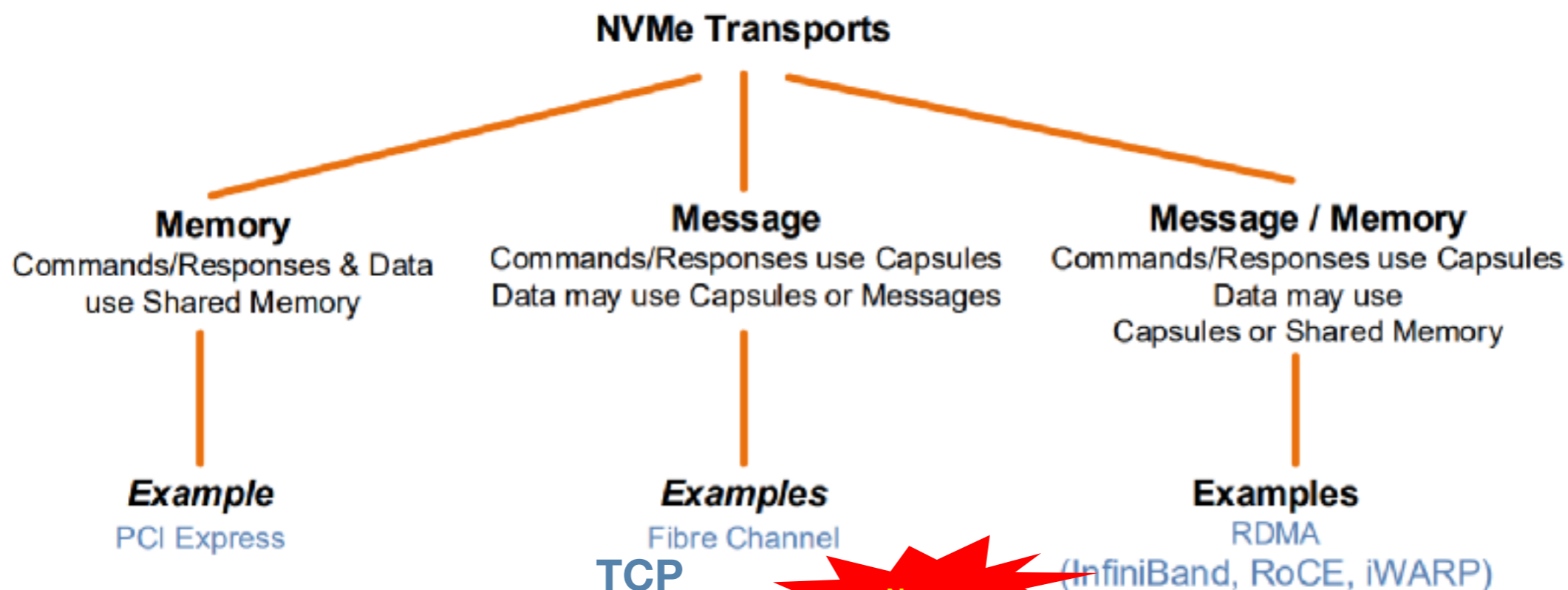
## Over RDMA

Figure 38: RDMA Transport Protocol Layers



# 2種類のNVMe over Fabrics + (ローカルの) NVMe

Figure 1: Taxonomy of Transports



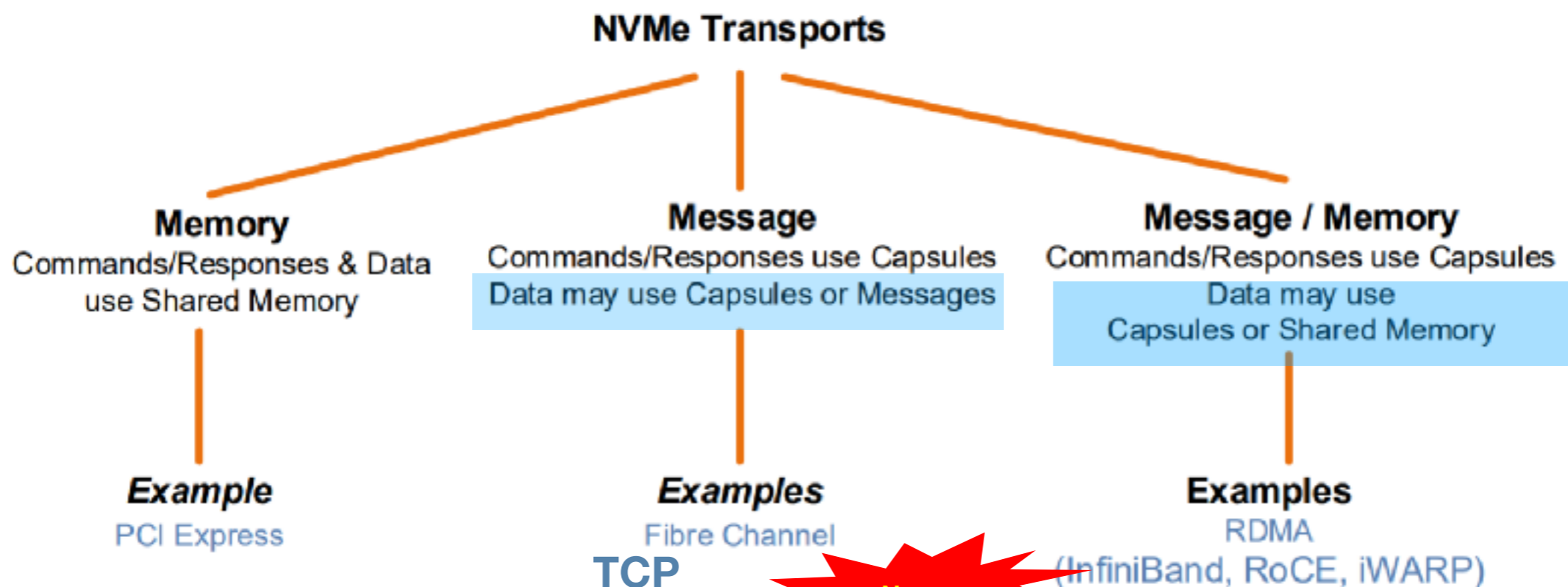
ローカル

Fibre Channel  
TCP

InfiniBand  
RoCE  
iWARP

# 2種類のNVMe over Fabrics + (ローカルの) NVMe

Figure 1: Taxonomy of Transports



ローカル

Fibre Channel  
TCP

InfiniBand  
RoCE  
iWARP

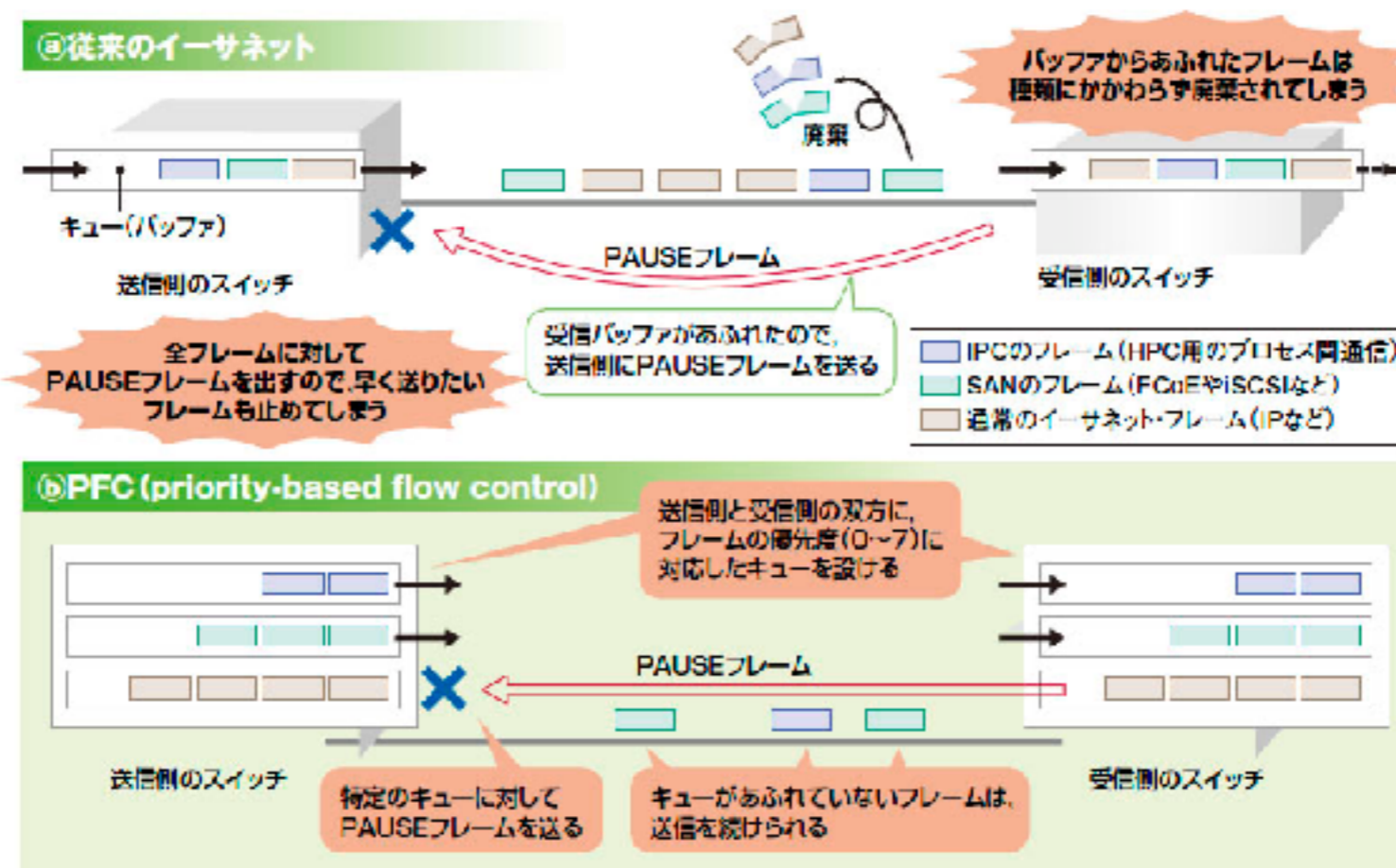
**Fabricsに求められるのは「ロスレス」  
ネットワーク**

# Fabricsに求められる「ロスレス」ネットワーク

- NVMe over Fabricsには「ロスレス」なネットワークが必要
  - データの順序の乱れやデータの欠損は許されない
- ロスレスを実現するネットワーク技術
  - FC / Infiniband = PCI Expressと同等
  - CEE: Converged Enhanced Ethernet (Ethernetの拡張)
  - TCP/IP (再送ベース)

# RoCEでロスレスを実現するPFC

- CEEのPFC:Priority-based Flow Control

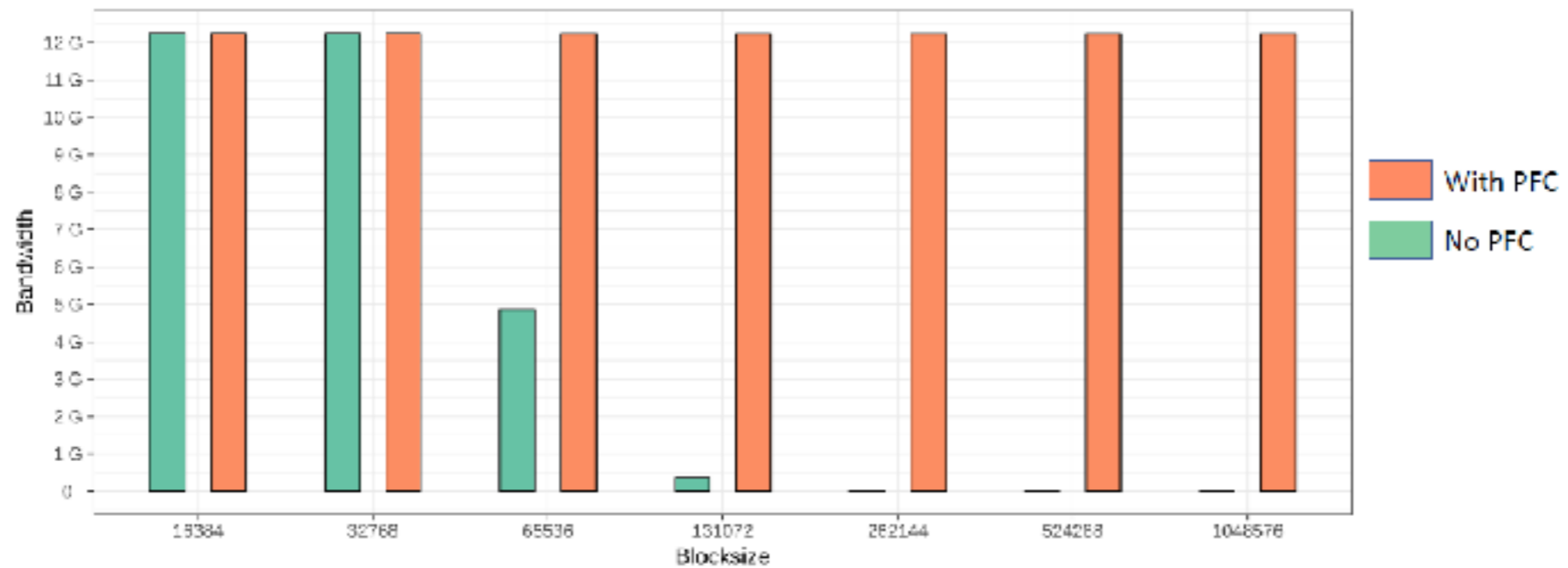




# PFC無しだと再送多発

## Do I Really Need DCB (Lossless Net) with RoCE?

BW vs. IO Size



Source: Western Digital Performance Tests

Western Digital

Flash Memory Summit 2019, Santa Clara, CA  
© 2019 Western Digital Corporation or its affiliates. All rights reserved.

11/26/2019

9

<https://www.flashmemorysummit.com/Proceedings2019/08-08-Thursday/>

それでも **Ethernet** はむずい。。。。

- Parking lot problem
- Victim Flow problem

# Parking lot problem

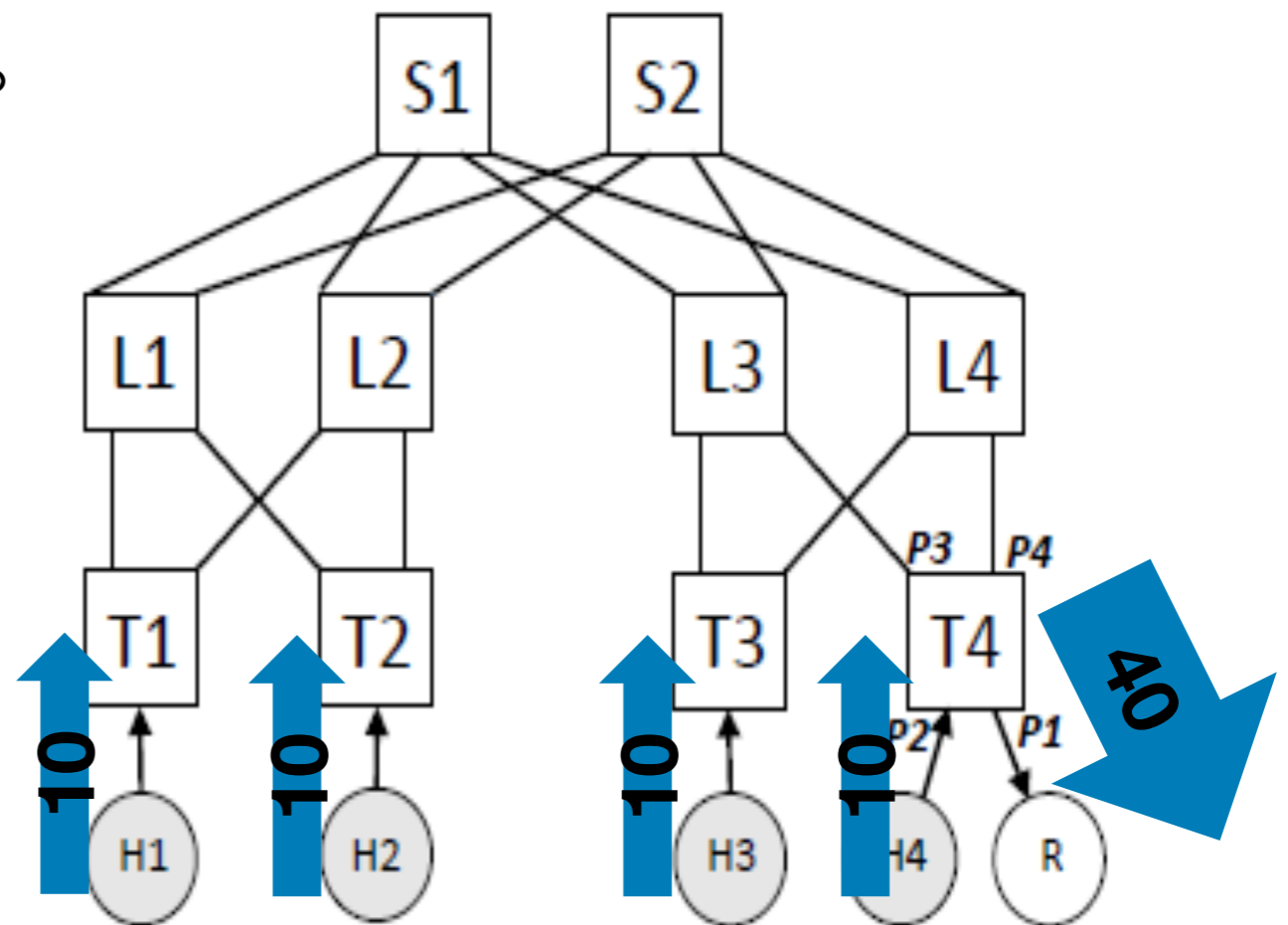
\*すべて40Gbpsリンク

\*BGP ECMP

{H1, H2, H3, H4} → (各10Gbps) → R

↓↓↓

T4→R は40Gが出る、はずだが。。。。



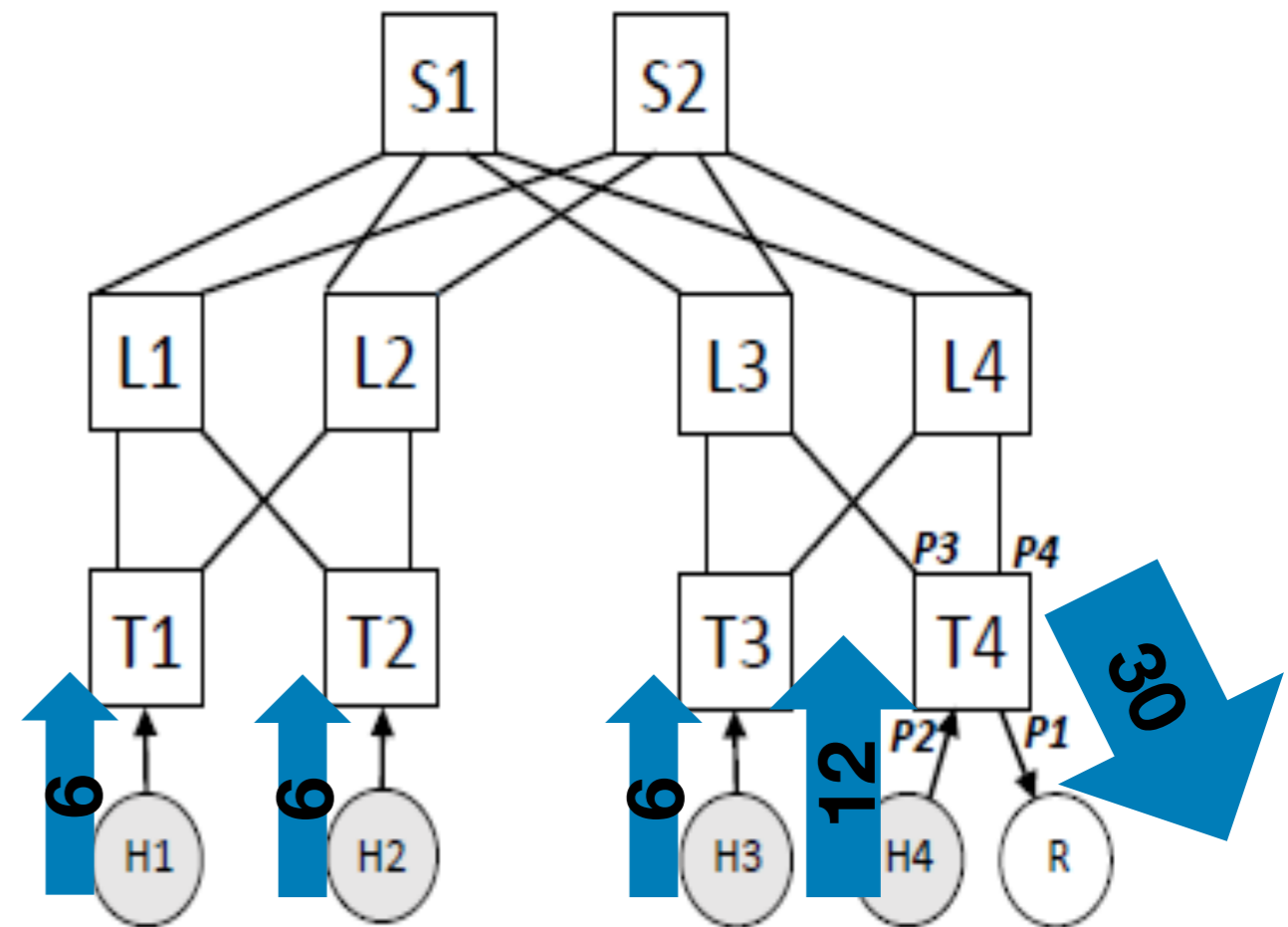
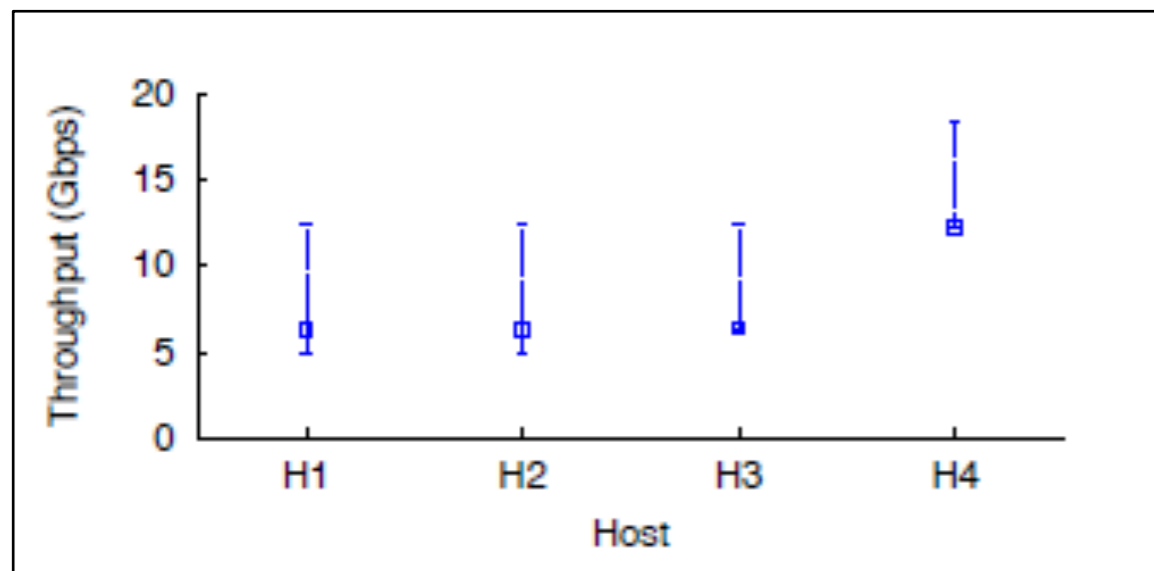
# Parking lot problem(cont.)

\*すべて40Gbpsリンク

\*BGP ECMP

そうならない。

Rに近い H4 → T4 が優先される。



# Parking lot problem(cont.)

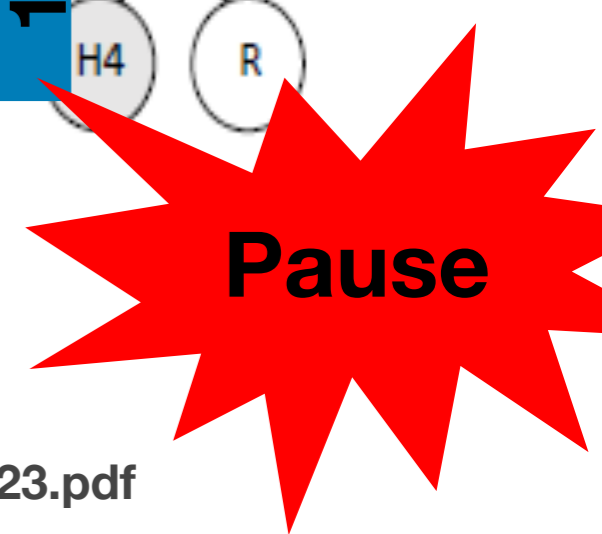
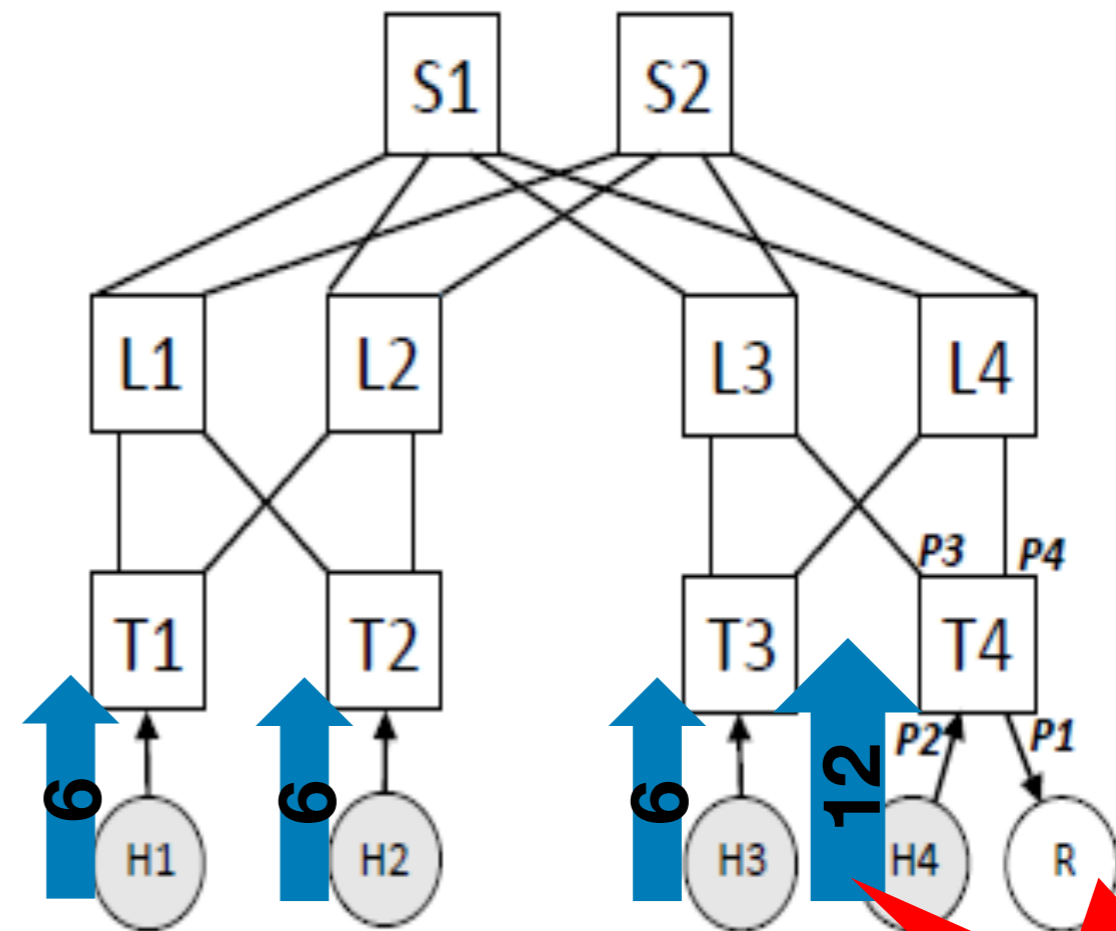
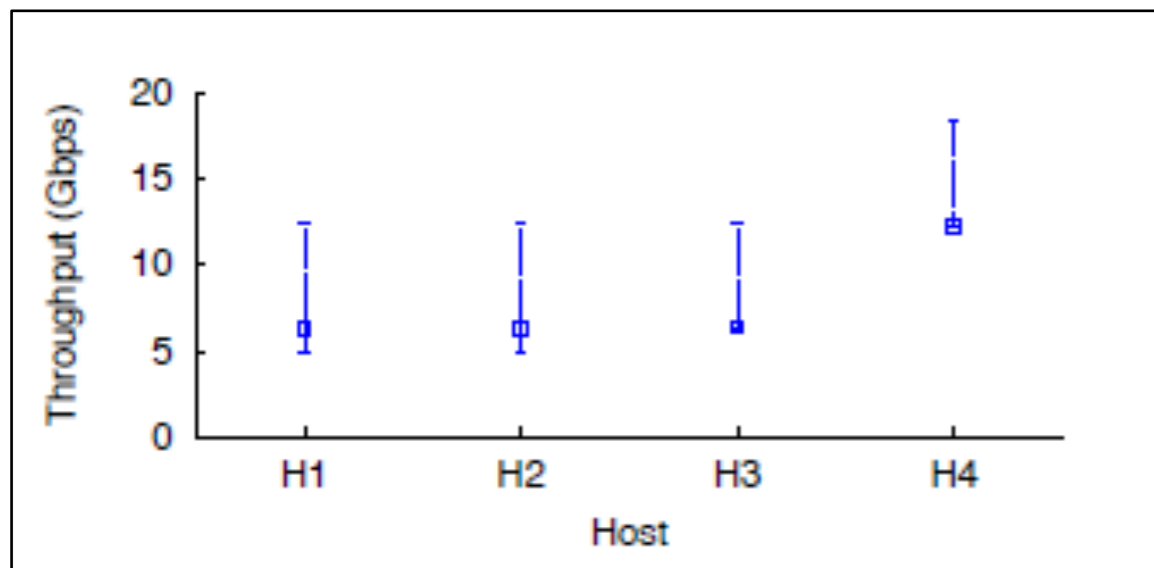
\*すべて40Gbpsリンク

\*BGP ECMP

P2リンクは1フロー

{P3,P4}リンクは複数フロー(ECMP)

→P2リンクはPauseの影響が少ない



# Victim Flow problem

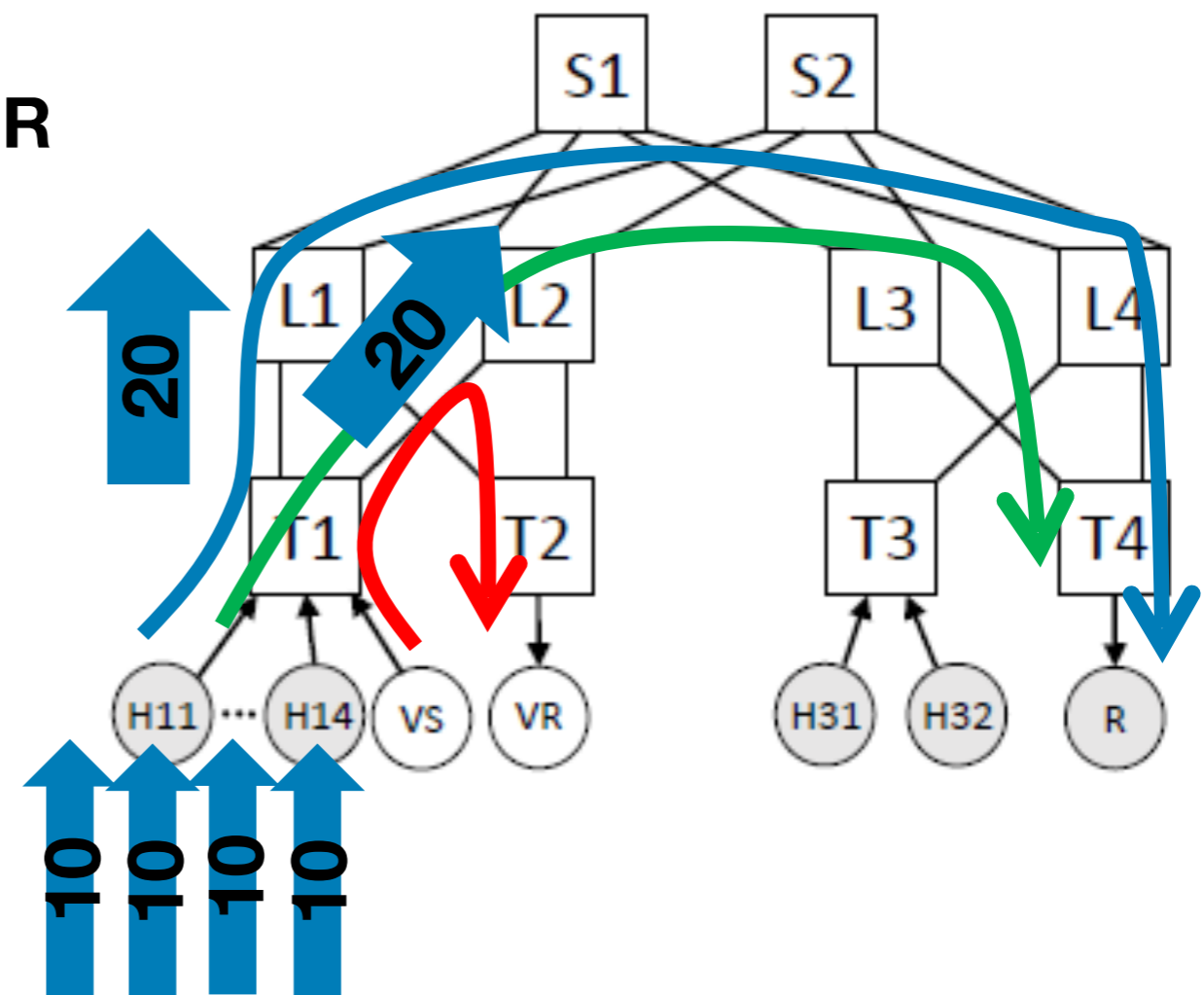
\*すべて40Gbpsリンク

\*BGP ECMP

{H11, H12, H13, H14} → (各10Gbps) → R

↓↓↓

VS → VRは20Gが出る、はずだが。。。。

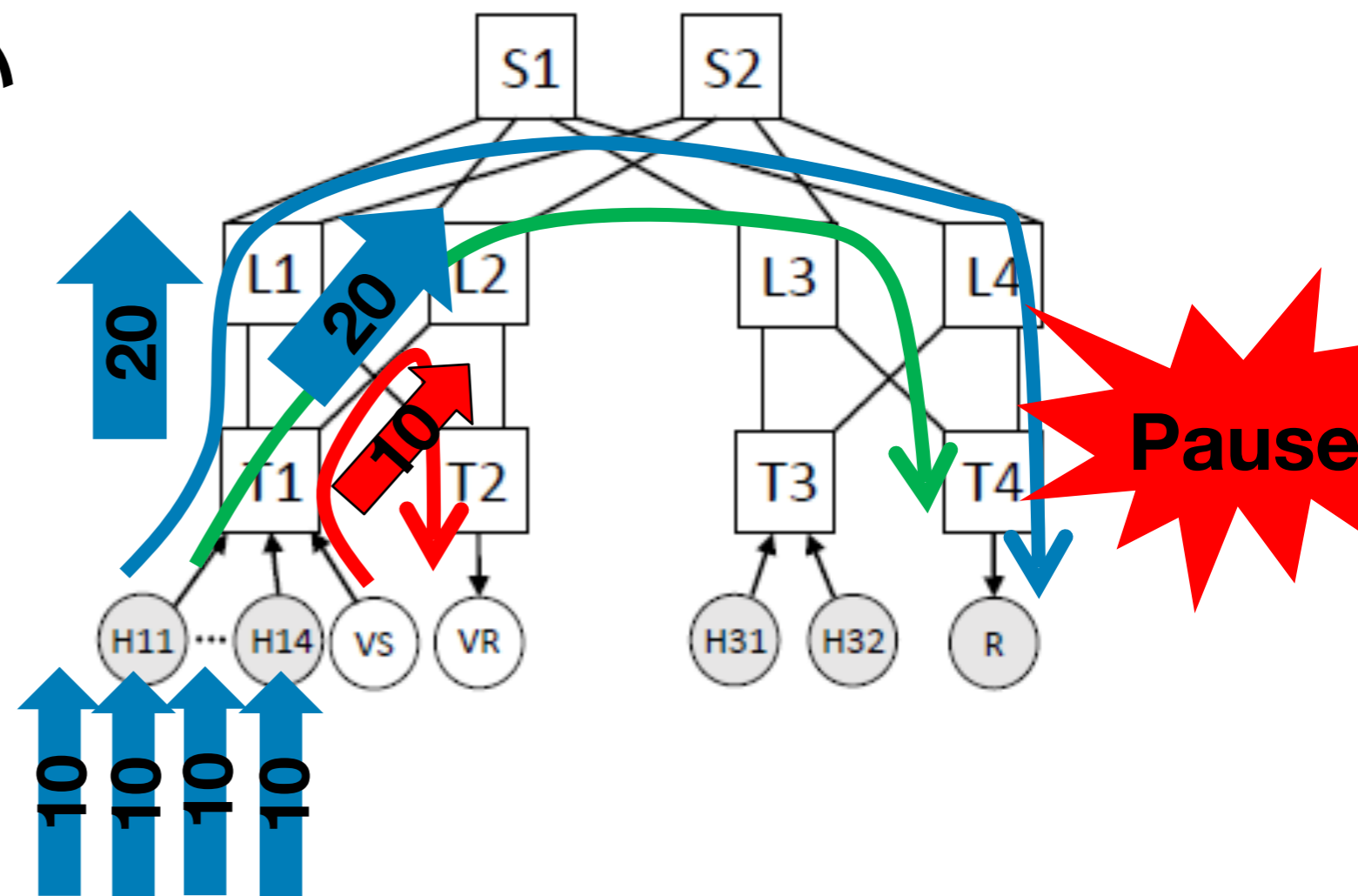


# Victim Flow problem

\*すべて40Gbpsリンク

\*BGP ECMP

VS→VR は10Gしかでない



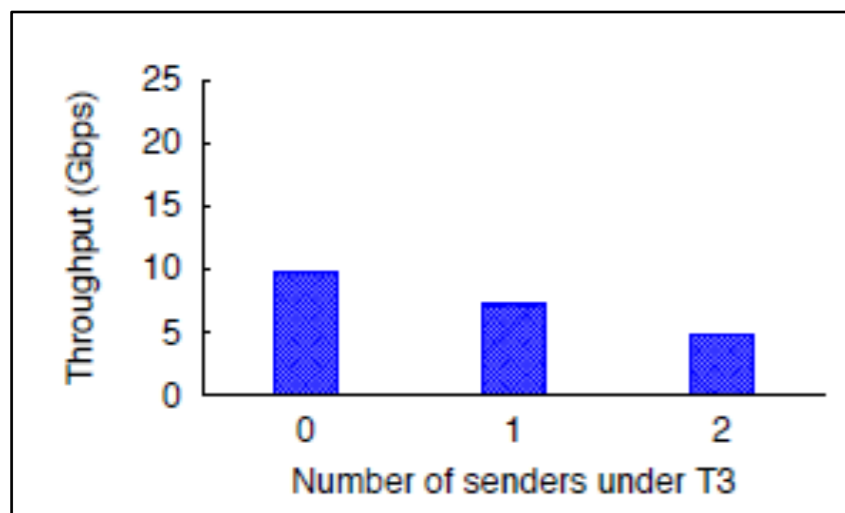
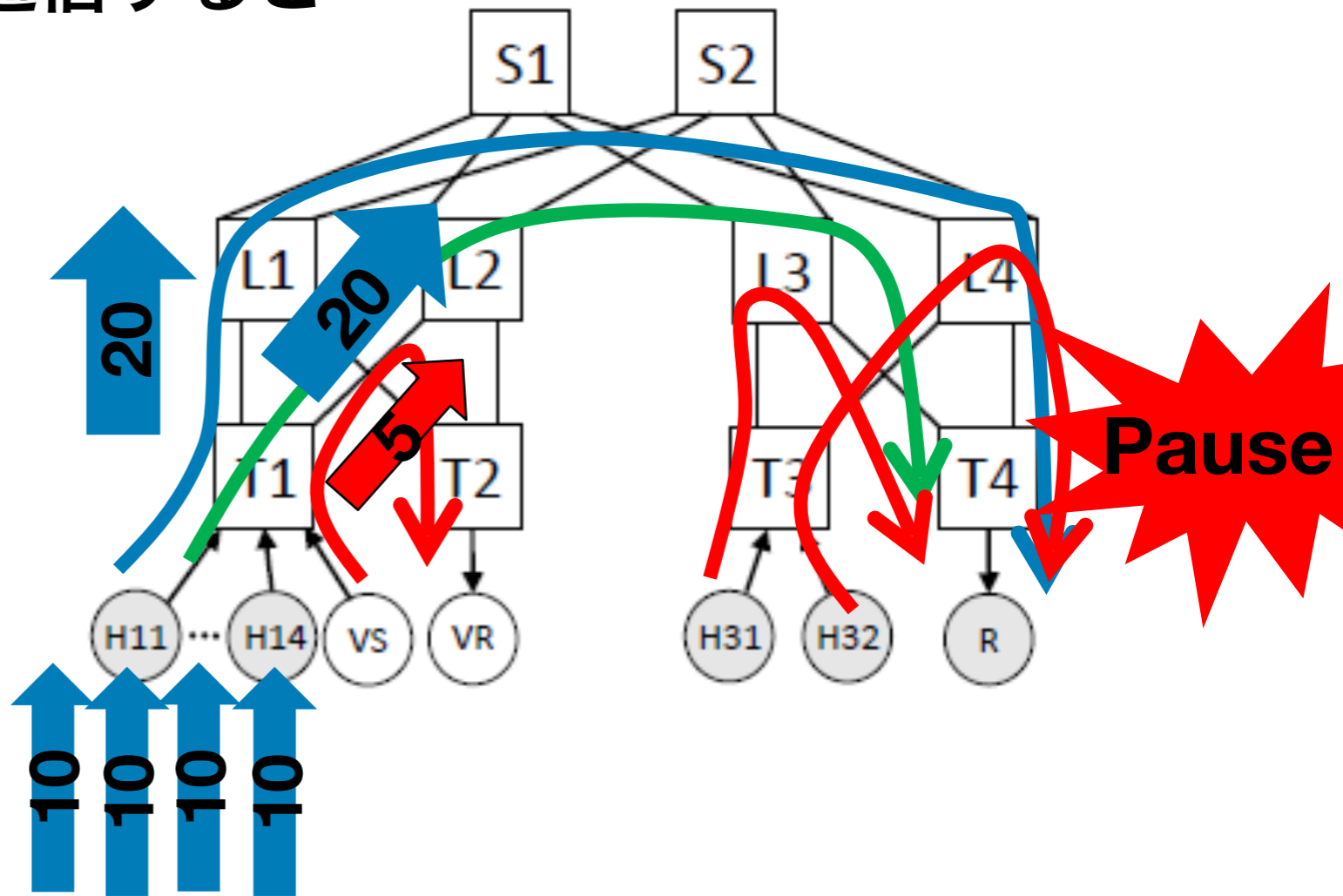
# Victim Flow problem

\*すべて40Gbpsリンク

\*BGP ECMP

さらに、H31→R, H32→R が通信すると

VS→VR はより減少する



T3配下の送信ノード数



# まとめ

- 今どきのストレージメディア(Flash, SCM)の性能を引き出すにはNVMeが必須
- 高速ストレージメディアを高速ネットワークで運ぶ複数の選択肢あり
- ローカルNVMeをネットワークで運ぶのがNVMe over Fabrics (NVMeoF)
- FabricsにはFC以外にInfiniband, Ethernet など選択肢がある
- 超高速Ethernetはとても魅力的だが、用途を踏まえた十分な検討とPoCを
- (ぼそっ)FCは安パイ、RoCEはハードのケアが、TCPは今後注目