JANOG57 Day3 2026年2月13日(金)

広域分散計算基盤を拓く長距離RDMAの課題と可能性

トヨタ自動車株式会社　加納浩輝

自己紹介

氏名： 加納 浩輝

所属： トヨタ自動車株式会社 InfoTech

- トヨタでの業務 (2023~)
  - GPU基盤に関する研究開発 (NW・kubernetes)
  - パブリッククラウド運用管理 (AWS・Azure・GCP・OCI)

- JANOG歴
  - JANOG54: 初参加
  - JANOG55: 初登壇

発表概要

広域分散GPU基盤間でのデータアクセス効率化を目的に
長距離RDMAへの挑戦

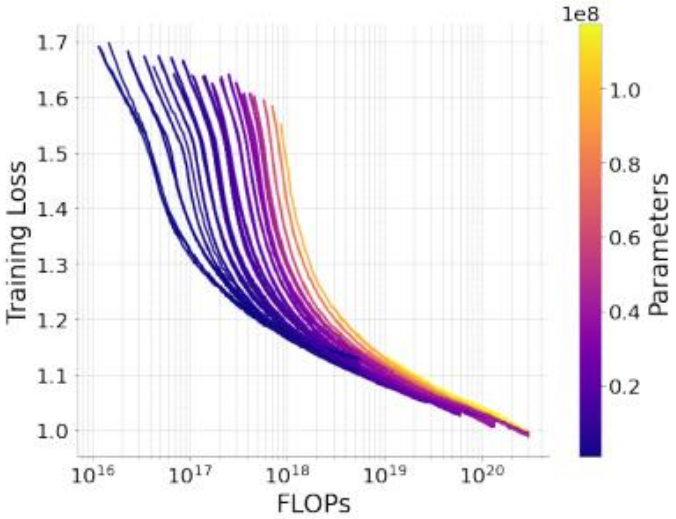
長距離RDMAの課題は？
既存技術でどこまで性能改善できる？

なぜトヨタがGPU基盤？ - AD/ADAS開発 -

TOYOTA

AI技術を活用したAD/ADASシステムの開発へのGPU利用

- AD: Autonomous Driving (自動運転)
- ADAS: Advanced Driver-Assistance Systems (先進運転支援システム)



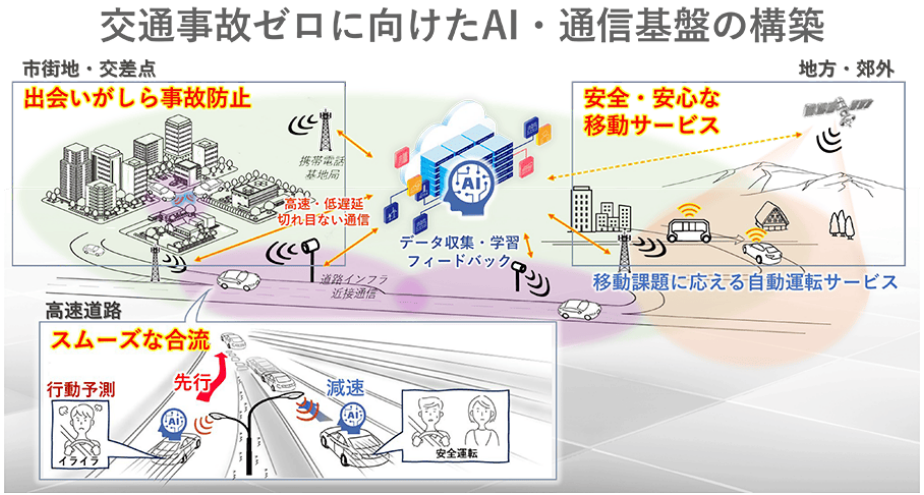
自動運転モデル精度に対する計算量スケール則の報告あり
→ 計画的な計算資源の確保が必要

引用：Scaling Laws of Motion Forecasting and Planning Technical Report

なぜトヨタがGPU基盤？ - NTT様との協業 -

分散計算基盤：交通事故ゼロ実現に向けたモビリティAI基盤の構成要素

IOWN APNで接続された分散計算基盤構築プロジェクト進行中

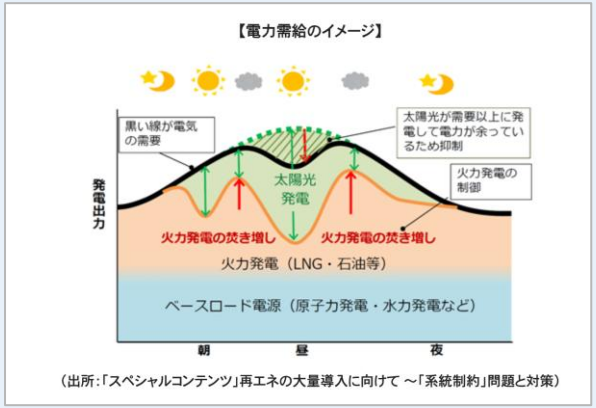


引用：NTTとトヨタ自動車、交通事故ゼロ社会の実現に向けた「モビリティ×AI・通信」の共同取り組みに合意
https://global.toyota/jp/newsroom/corporate/41746612.html

なぜ分散？

長期視点

再生エネルギー有効活用



Keyword: ワットビット連携

短期視点（現実）

設備拡張時

① 電力・スペース売切

② 社外サービスの活用

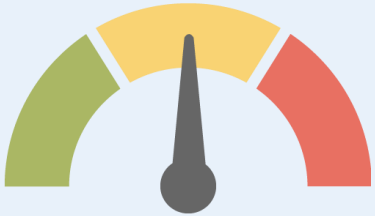
もっと詳しく
知りたい方

Open Networking Conference Japan 2025
分散したコンピューティングリソースの活用について

分散計算基盤の課題

TOYOTA

性能



遠隔データへのアクセス性能

分散学習性能

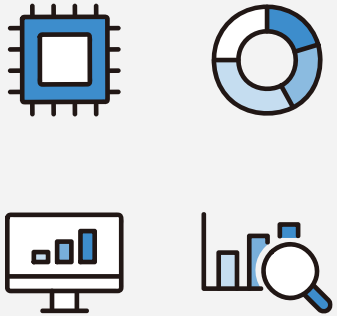
今日はこのお話

UX

ジョブの実行先決定

拠点毎の差異隠蔽

運用



可視化

リソース配分公平性

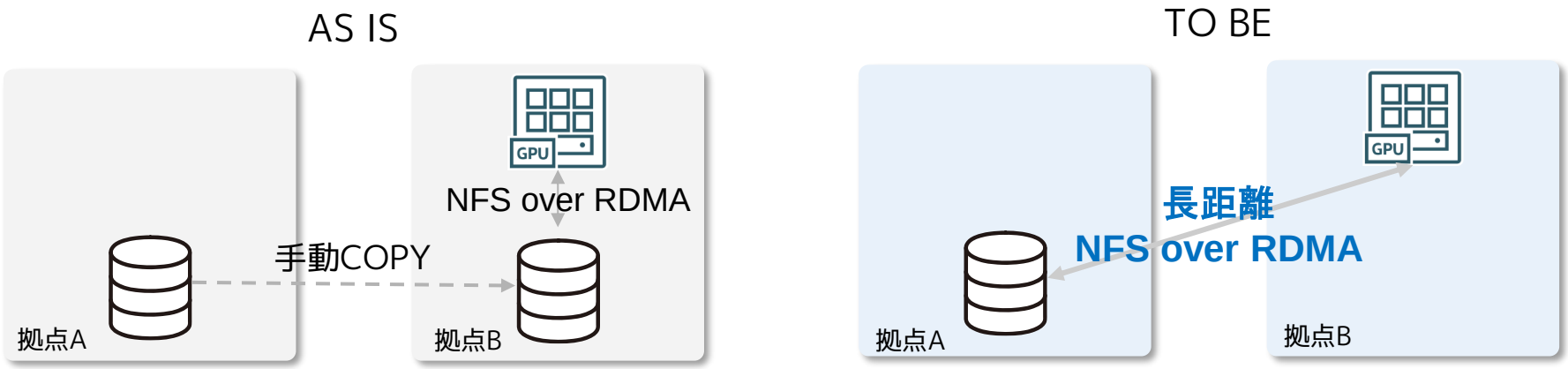
沖縄クラウドネイティブ勉強会2025
GPUを効率的に使うための分散型計算基盤

UX・運用課題が気になる方

遠隔データへのアクセス

DC内/外で**同じよう**に高性能なデータアクセスは可能か？

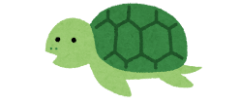
NFS over RDMA

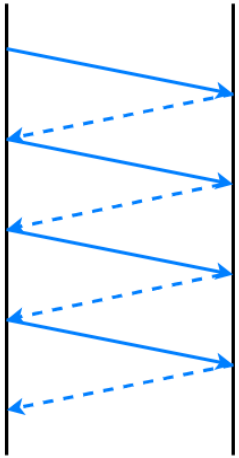


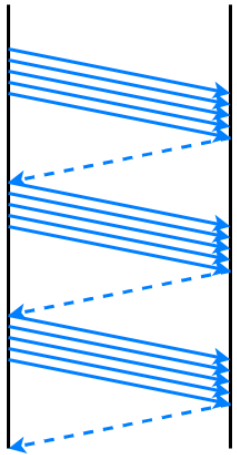
→ RDMAの長距離化への挑戦

遅延 vs 帯域

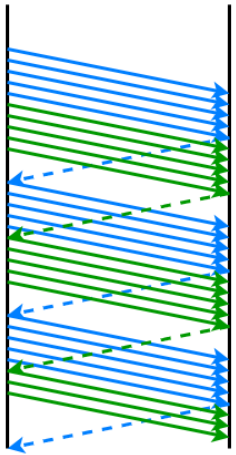
データ転送速度　∝　一括送信量 x 多重度







一括送信



多重化

高遅延(長距離)下では、一括送信量・多重度を上げないと隙間が埋まらず性能劣化

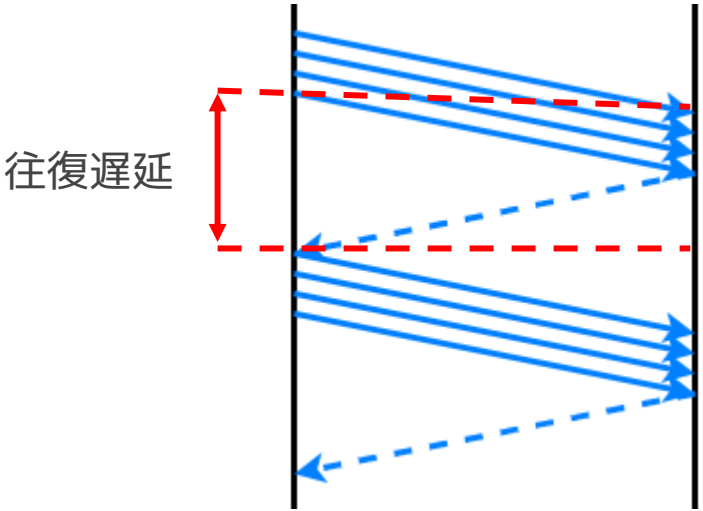
どこまで一括送信量・多重度を上げればいいのか？

TOYOTA

一括送信量・多重度をどこまで上げればいい？

→　帯域遅延積 (BDP)を埋められるまで

帯域遅延積 (**B**andwidth **D**elay **P**roduct) = 目標帯域 (物理帯域) x **往復遅延**

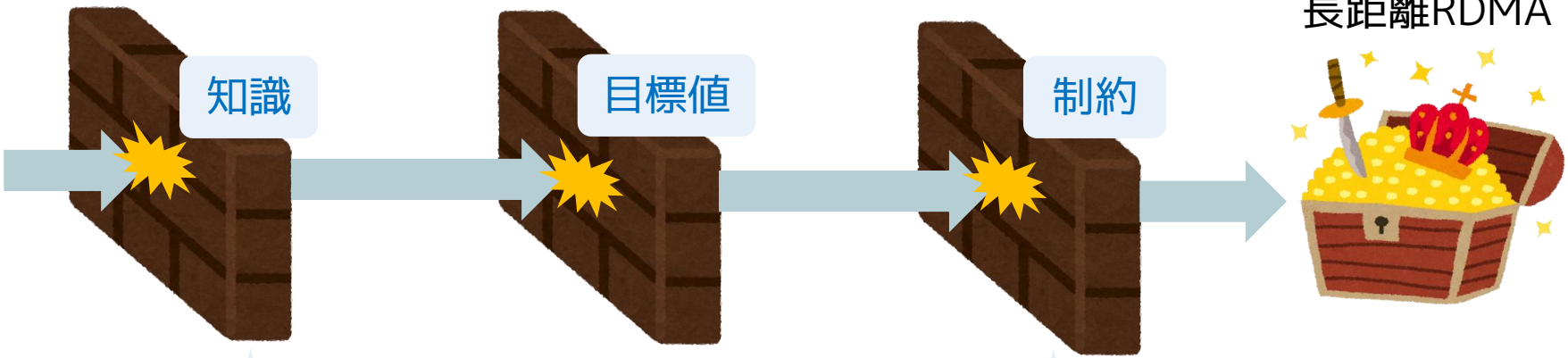


本来この時間で
BDP相当量のデータを送信可能

【例】
400Gbps, RTT 1msの環境
BDP = 400Gbps x 1ms = 50MB
BDPを埋めるには，　一括送信量 x 多重度 > 50MB

RDMAでBDPを埋める壁

BDP = 目標帯域 x 往復遅延
大 (~数百Gbps)　大



高性能
長距離RDMA

RDMAで一括送信量・多重度に
該当するパラメータは？

NICのハードウェア
リソース制約

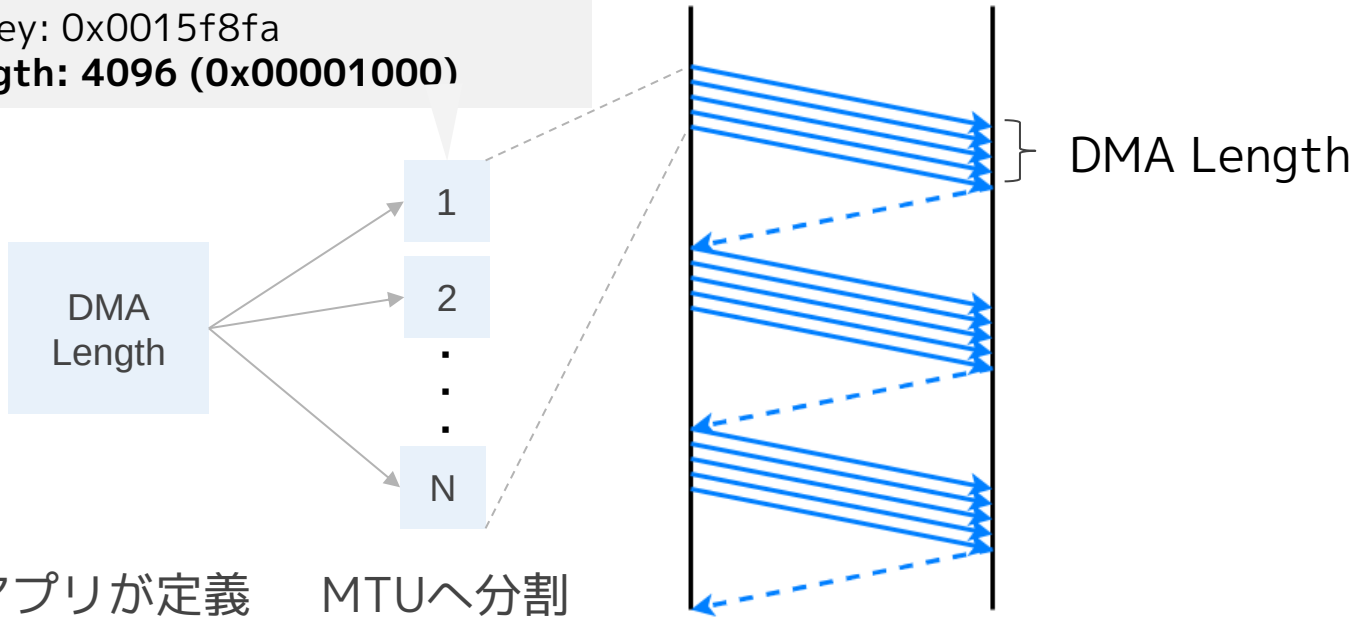
RDMAにおける一括送信量

一括送信量 = DMA Length

→　単一DMA Operationで送信されるデータ長

パケットキャプチャ抜粋

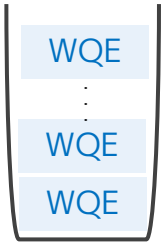
```
RETH - RDMA Extended Transport Header
      Virtual Address: 0xda17631a7648d000
      Remote Key: 0x0015f8fa
      DMA Length: 4096 (0x00001000)
```



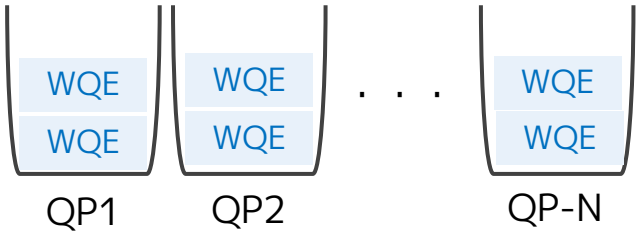
RDMAにおける多重度

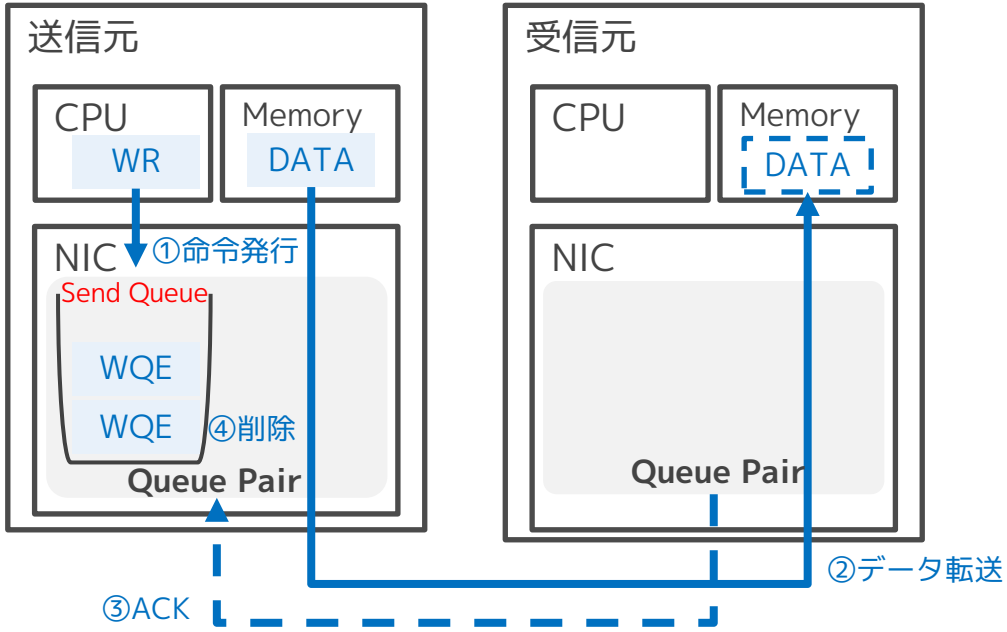
多重度上限 ＝ **Send Queue サイズの総量**　※ RDMA Writeの場合
→ 実行中命令(WQE)の保存Queue

対策１
Queue長を伸ばす



対策2
Queue数を増やす(= QP数を増やす)



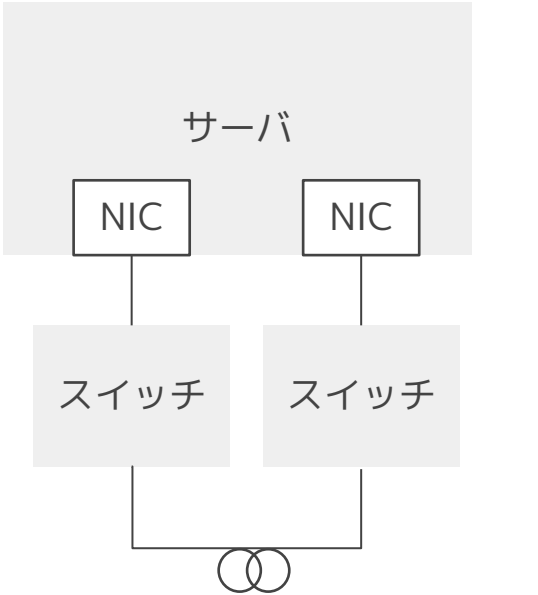


実験環境

TOYOTA

ボビンファイバを用いてラックサイズでデータセンタ間距離を摸擬

機器	製品	備考
NIC (サーバ)	ConnectX-7	Firmware: 28.43.3608
スイッチ	AIS800-64D	NOS: Edgecore SONiC 202311.3

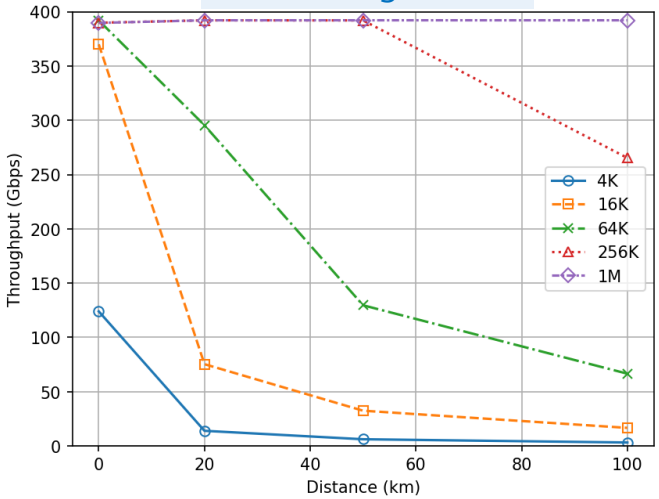


1m, 20km, 50km, 100km

結果 (一括送信量)

DMA Lengthの調整で 100km-400GbpsのRDMA通信達成

DMA Length変更



Send Queue: 128
QP数: 1

400Gbps – 100km (RTT 1ms)では,

BDP = 400Gbps x 1ms = 50MB

BDPを埋めるには,

一括送信量 x 多重度 > 50MB

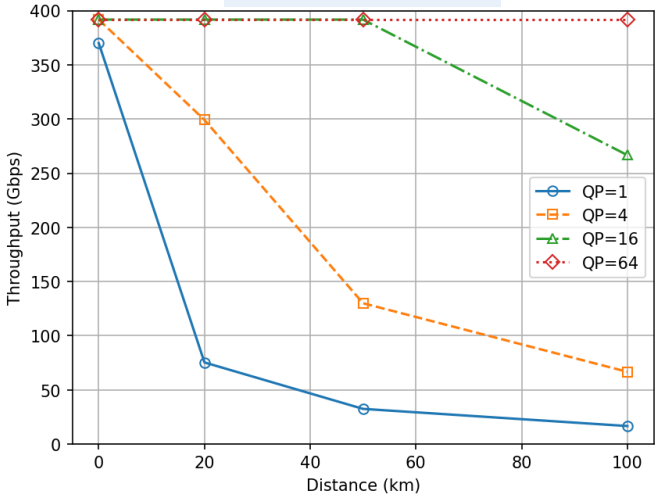
DMA Length　SQ長 x QP数

QP数：1, SQ長：128 の条件下では

DMA Length > 400KB

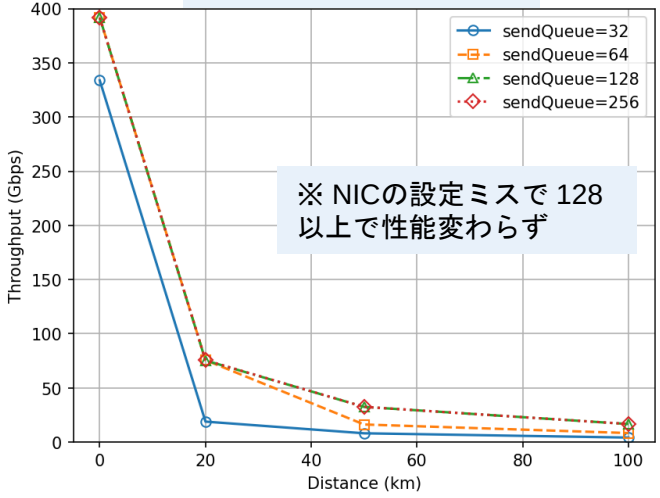
結果(多重度)

QP数変更



DMA Length: 16KB
Send Queue: 128
↓
400Gbps-100km
必要条件 QP数 > 25

SQ長 変更



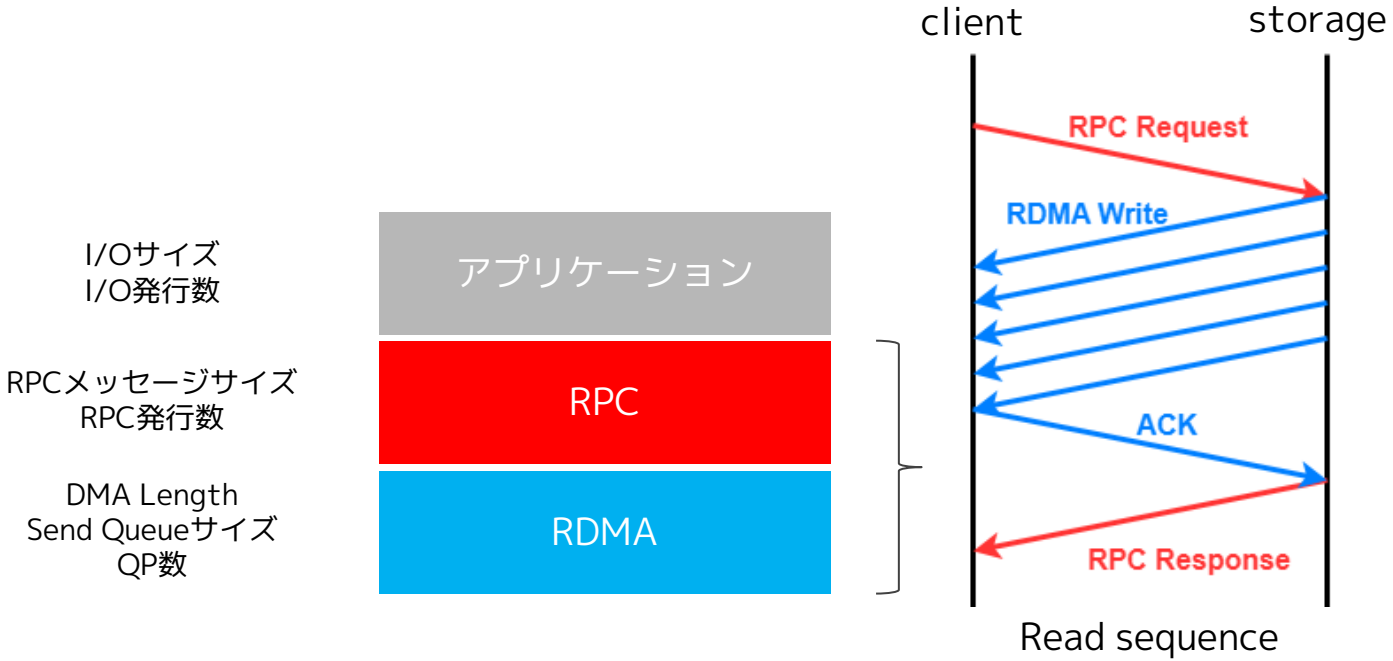
DMA Length: 16KB
QP数: 1
↓
400Gbps-100km
必要条件 SQ長 > 3,200

NFS over RDMAのチューニング

NFS over RDMA実行時において，

① RMDA関連パラメータはどうなってるの？

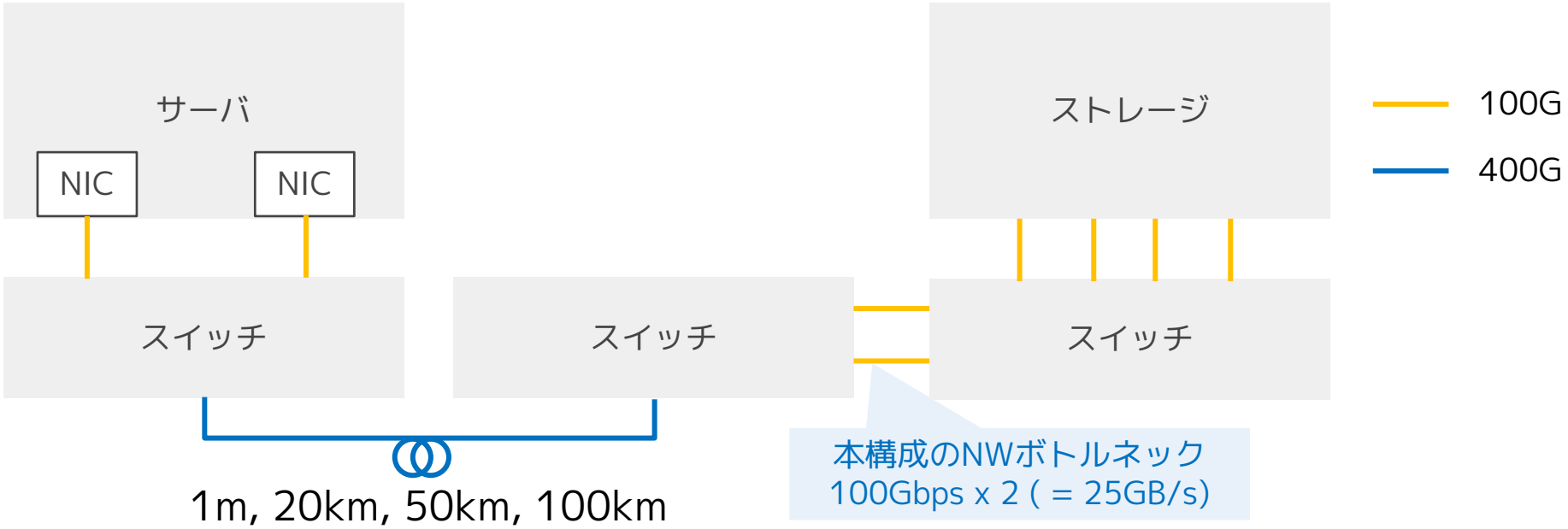
② 他のレイヤにボトルネックは無いのか？



Read sequence

実験環境 (NFS over RDMA)

クライアント /　ストレージともに全NICに負荷を分散可能な構成 (= pNFSプロトコル)



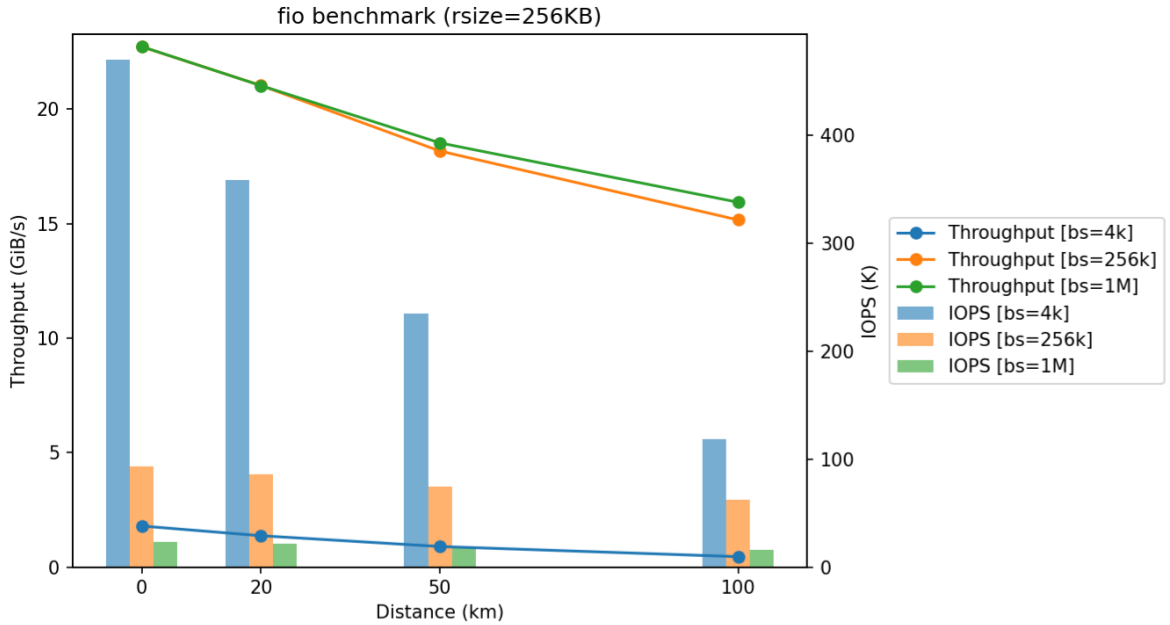
結果 (fio)

TOYOTA

距離に反比例した性能劣化を確認

fioパラメータ

- mode: read
- numjobs = 32
- iodepth = 16
- file size = 1GB



ボトルネック分析

RDMAレイヤは十分な多重度・一括送信量がありそう
アプリ・RPCレイヤもチューニング済み

RDMA関連パラメータ

		値	備考
一括送信量	DMA Length	120KB	NIC仕様で上限が決定
多重度	QP数	4 (/storage NIC) → 合計 16	構成次第
	Send Queue	?	ベンダー非公開

※ 各値の詳細説明は補足資料(2)に記載

ボトルネックはクライアント・ストレージ以外？

PFCが大量発生

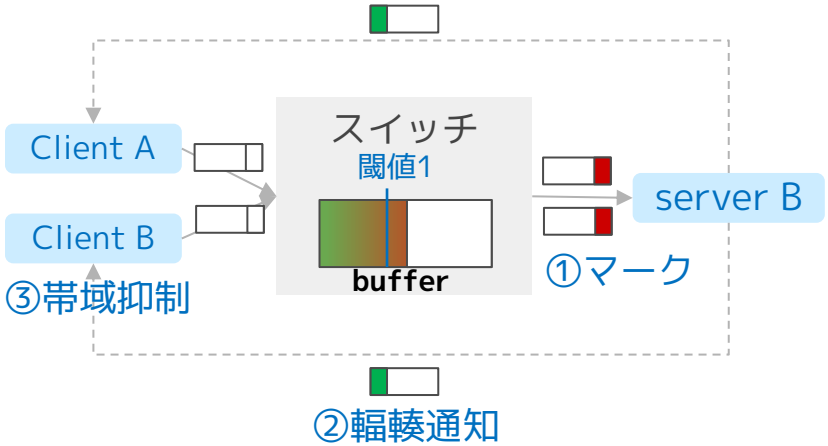
長距離構成 (> 20km)の時のみPFCが多量発生

なんで距離が長いときだけ？？

ECN・PFC おさらい

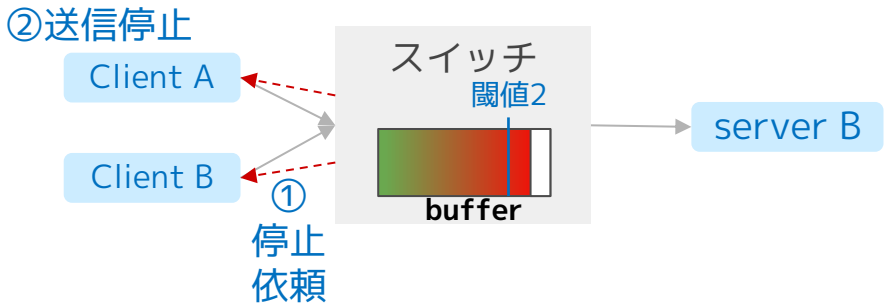
ECN

帯域抑制し輻輳を緩和



PFC

パケット送信を止めパケロス回避
性能影響大

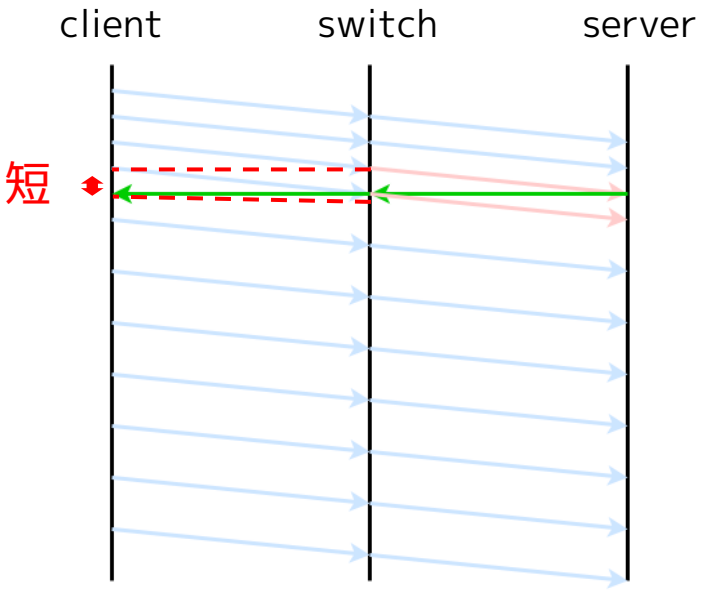


各閾値を調整しPFCを発動させないのが理想

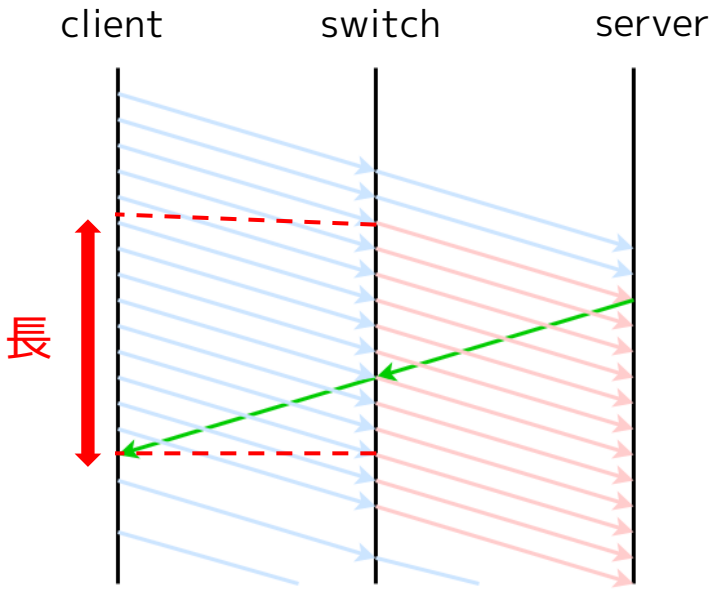
ECNと遅延の関係

高遅延環境では，輻輳発生 → 帯域抑制 までに時間がかかる

低遅延環境



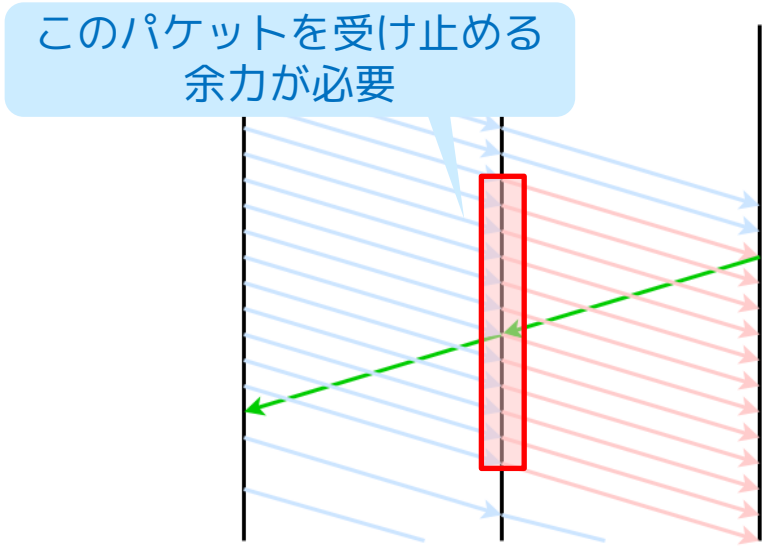
高遅延環境

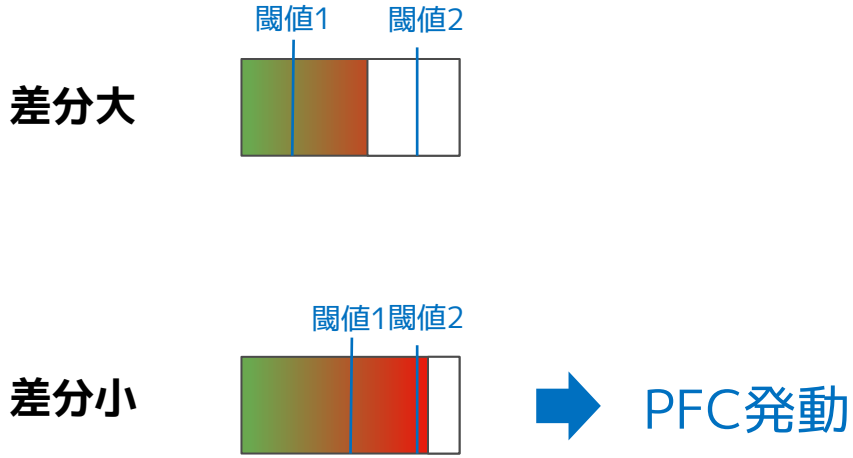


高遅延下での ECN-PFCチューニング

ECN発動閾値 ⟷ PFC発動閾値

BDPに比例した大きな差分が必要





ECNパケットを受け取る頻度でレート制御するので，実際はもっと複雑

検証環境の振り返り

今回は短距離経路でチューニングしたパラメータで実験

→　長距離ECNの輻輳制御が間に合わずPFCが多発

ECN (gmin, gmax)	150KB, 1.5MB
PFC (xoff)	35MB

BDPよりは大きな値に設定
もっと広げる必要あった？

長距離向けのチューニングはできるのか？

既存DCQCNですべてを満たすチューニングは可能なのか

非輻輳時の性能

ECN閾値を下げる
→ 平時の性能劣化

ハードウェア制約

PFC閾値を上る
→ パケットロスのリスク

短距離-長距離通信の性能両立

UEC・Fast CNPなどの新規格で解決可能か？

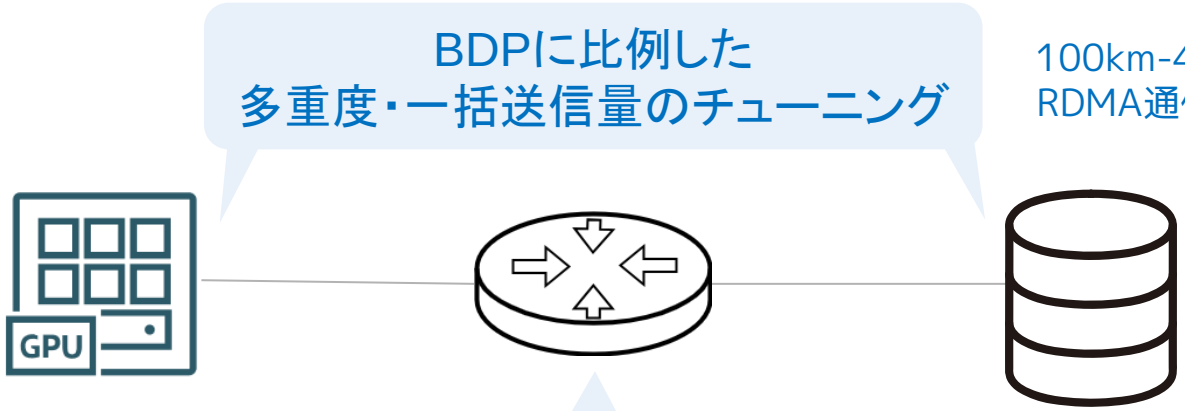
みなさんどう思います？

まとめ

TOYOTA

長距離RDMAはBDPとの闘い

帯域遅延積 (**B**andwidth **D**elay **P**roduct) = 目標帯域 (物理帯域) x 往復遅延
大 大



議論

- **既存技術で，長距離RDMAの性能を担保するQoS設定は可能か？**
  - 次世代技術に期待しているが，機器総入れ替えは大変
- **PoCで終わらせないために，超えなくてはいけない壁は他にあるか？**
  - 拠点間ネットワークのTCP-RDMA通信の共存
- **そもそも，拠点間RDMAは本当に必要か？ TCPで十分？**
  - データアクセス以外にユースケースはあるか？
  - 拠点間で分散学習(GPU間通信)したいのってどんな時だろう？
- **ワットビット連携とネットワークエンジニアの役割は何か？**
  - 長距離RDMAに限らず，再生可能エネルギーの効率的利用に向けてネットワークエンジニアができることは.

補足

(補足1) Perftestにおけるパラメータ調整

RDMA基礎検証に用いたベンチマークソフトウェア

Perftest: RDMAパフォーマン測定スツール(https://github.com/linux-rdma/perftest)

遅延対策の為に変更したいパラメータオプションは下記

`-t, --tx-depth=<dep>` SendQueue長さの変更

`-q, --qp=<num of qp's>` Qpair数の変更

`-s, --size=<size>` RDMAメッセージサイズの変更

(補足2-1) NFS over RDMAにおけるDMA Length

TOYOTA

DMA Length = min(block Size, 120KB)

最大DMA Length = ページサイズ(4KB) x max_SGE(30) = 120KB

※ max_SEG:
NICでサポートしている最大SGE (Scatter Gather Element). NVIDIA NICでは30となっている

※ SGE (Scatter Gather Element):
非連続なアドレスのデータを、まとめて受取る(Scatter), まとめて送る(Gather)ためにメモリのアドレス・サイズを格納する構造体.

Block size = 4KB

RETH - RDMA Extended Transport Header
Virtual Address: 0xda17631a7648d000
Remote Key: 0x0015f8fa
DMA Length: 4096 (0x00001000)

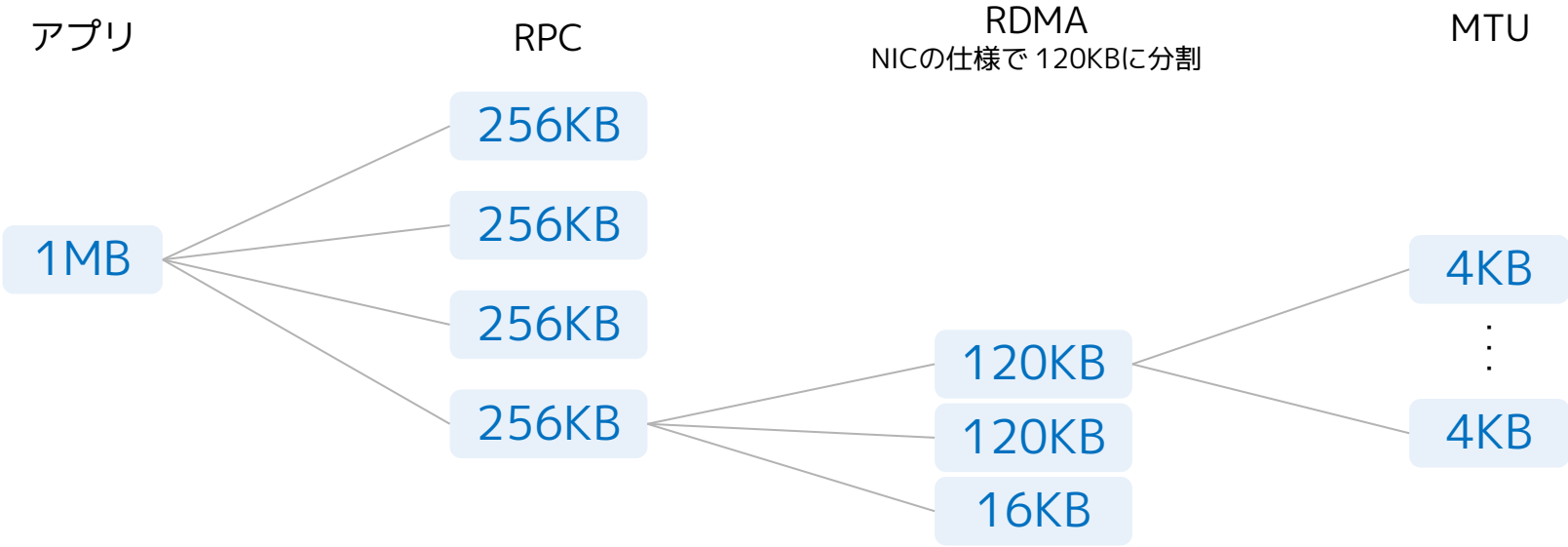
Block size = 256KB

RETH - RDMA Extended Transport Header
Virtual Address: 0xd5a3ed907a41d000
Remote Key: 0x002083b6
DMA Length: 122880 (0x0001e000)

(補足2-2) NFS over RDMAのチューニング

メッセージサイズは下のレイヤほど分割されていく

上位レイヤの多重度 > 下レイヤの多重度 になっているか？
端数が出て効率が悪くなってないか？

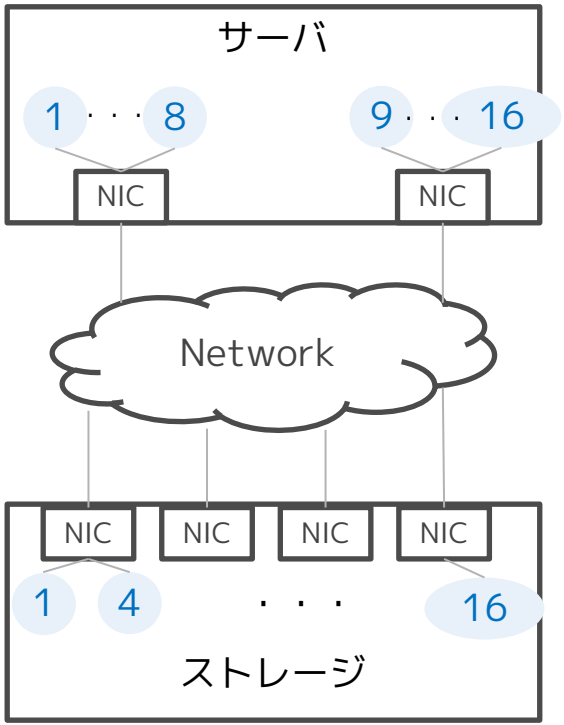


アプリが１MBのread発行 , nfs mount rsize = 256KBのケース 例

（補足2-3）NFS over RDMAにおけるQP数

- ストレージに 4個 / NIC →　合計 16個のIPアドレスを設定

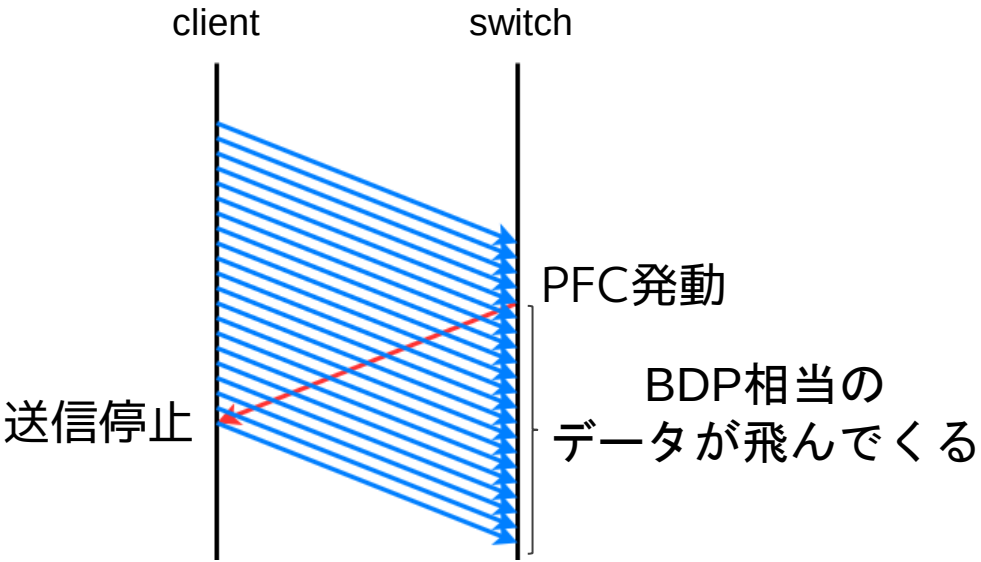
→ IPアドレス毎に異なるQPを生成

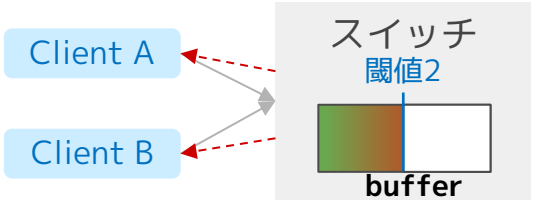


システム構成を変えれば
QP数も変わる

(補足3) 長距離リンクにおけるPFC

PFC発動～送信停止までに飛んでくるデータを受け止めるバッファが必要





PFC発動閾値を下げる
↓
PFC発動確率高
バッファサイズ足りるの？