



分散計算基盤実現に向けた長距離RDMAの高速輻輳制御

NTT株式会社　市川 潤紀
トヨタ自動車株式会社　加納 浩輝

2026年 10月 26日 MPLS Japan 2026

© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

Speaker



市川 潤紀

NTT株式会社
未来ねっと研究所



加納 浩輝

トヨタ自動車株式会社
InfoTech

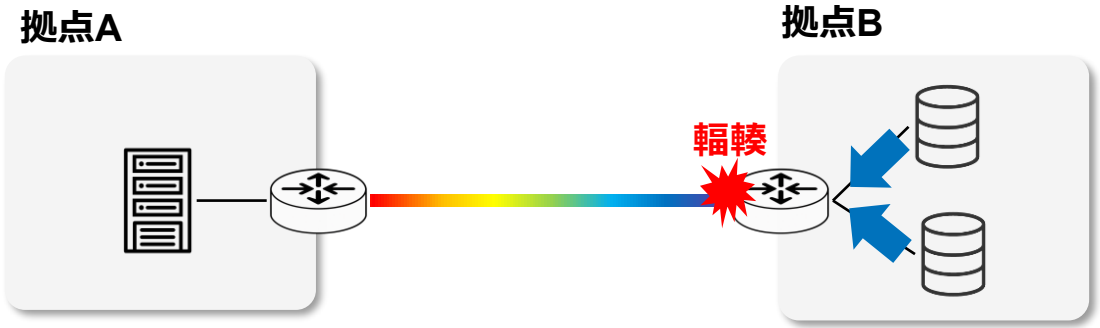
© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

本日のテーマ

NTT TOYOTA

分散計算基盤における遠隔データに高速アクセスしたい

➡ **RDMAを活用した長距離高帯域データアクセスの実現可能性の検証**



© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

パート１：背景

© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

NTT × 分散計算基盤

コア業務へのAI適用、フィジカルAIなど、AIの普及に合わせ求められるニーズが変化

AIネイティブな次世代インフラ「AIOWN」への転換を進めている



- GPU・ネットワーク・電力を用途に応じて最適化する
- DCと主要都市の間はIOWN APNで接続し、低遅延・低消費電力を実現
- **複数拠点のGPU資源やネットワークを柔軟に活用する機能**を順次拡張

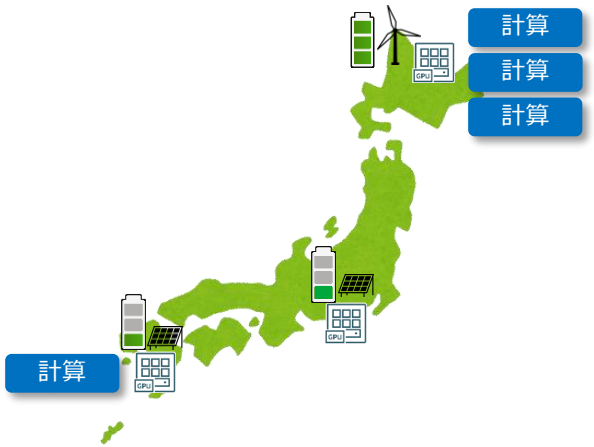
© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

トヨタ自動車 × 分散計算基盤

NTT TOYOTA

AD/ADASシステムの開発を支える分散計算基盤を構築中

再生エネルギー有効活用するために分散化



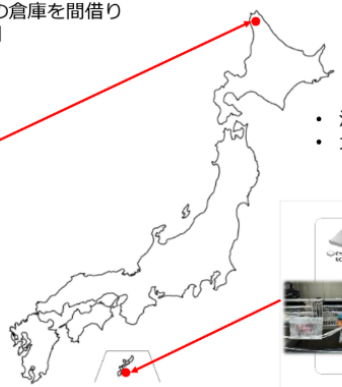
国内で長距離分散実証

TOYOTA INFOTECH

- おそらく**日本最北端**のGPU環境と**日本最南端**のGPU環境

- ユーラスエナジー稚内支社の倉庫を間借り
- IIJさんのマイクロDCを展開





- 沖縄オープンラボのスペースを間借り
- 大手町とIOWNでつなぎ実証



ONIC2025,分散したコンピューティングリソースの活用について

© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

分散計算基盤の課題

NTT TOYOTA

課題：遠隔大規模データセットへの効率的アクセス方法



計算ジョブとデータセットでは移動コストが 数桁異なる

= 計算ジョブは移動できても **データが取り残される**

© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

遠隔データアクセスのボトルネック

NTT TOYOTA

遠隔データアクセス性能 ボトルネックの主要因は？

どちらが主要因になるかはワークロード依存

遅延

性能上限 = **光伝播遅延**

光伝播遅延以下の遅延
が求められるユースケースでは改善不可

帯域

性能上限 = 敷設済 **物理帯域**

コストをかければ
DC内と同等の物理帯域も準備可能

© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

AI学習におけるデータ取得と「遅延」

AI学習実行時，計算の裏側で次のbatchで使うデータを事前取得 (Prefetch)

➡ **「各batch計算時間 > データ取得時間」**を満たしていれば
取得遅延が多少増加しても影響無し



一般に **「光伝播遅延 << 各batch計算時間」** のため

AI学習ではPrefetchにより光伝搬「遅延」の影響を隠蔽しやすい

© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

AI学習におけるデータ取得と「帯域」

NTT TOYOTA

帯域に反比例してデータ取得時間が長くなる
Prefetchによる隠蔽可能時間を超えると、GPUのデータ待ちが発生



分散計算基盤におけるAI学習実行時のデータ取得において
実行「帯域」の確保が必須要件

本発表のテーマ：

RDMAを活用した長距離高帯域データアクセスの実現可能性に関して

© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

パート１：まとめ

- 分散計算基盤の課題の一つは，AI学習に活用するデータセットへの遠隔アクセス

- AI学習ワークロードの特性 (データprefetch)により，
  - 光伝播遅延は隠蔽しやすい
  - 一方，帯域の劣化は学習速度への影響しやすい

- RDMAを活用して，長距離通信の高帯域化の実現可能性を検証

© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

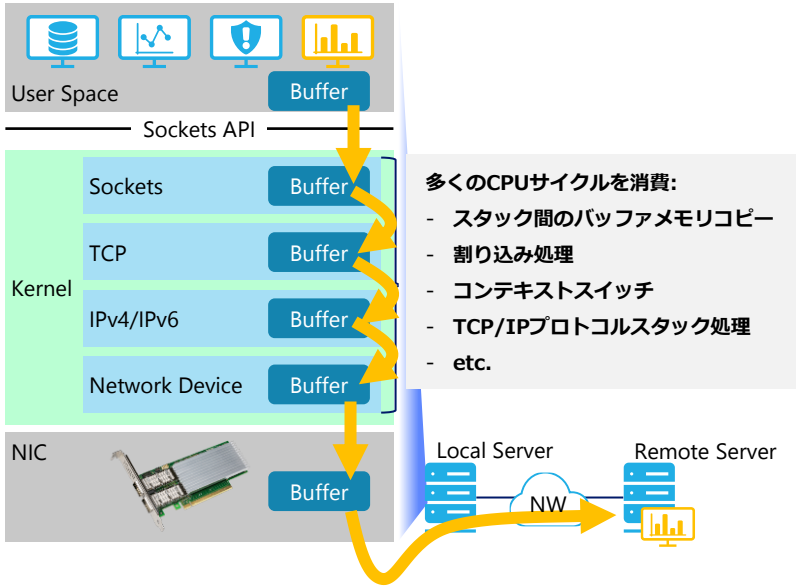
パート２：技術編

© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

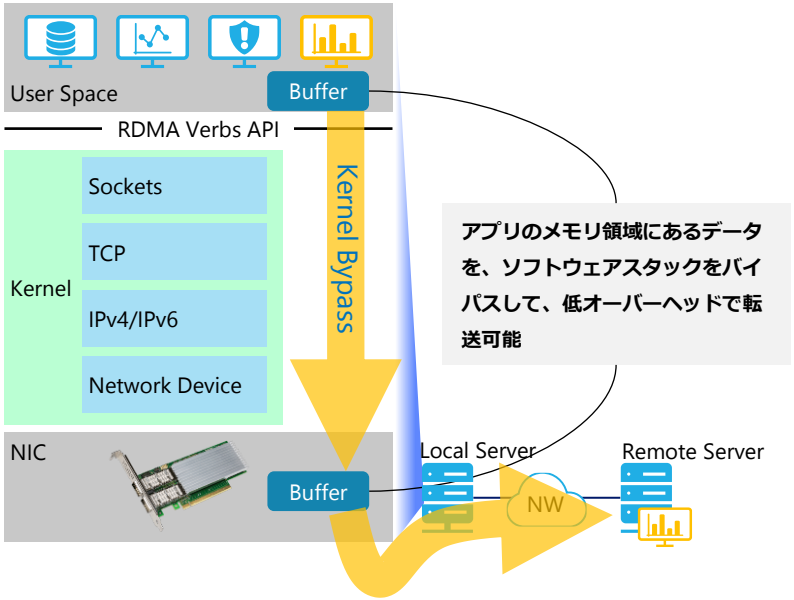
RDMAとは　高帯域低遅延な通信プロトコル

NTT　TOYOTA

従来のソケットベースのデータ転送



RDMAによるデータ転送

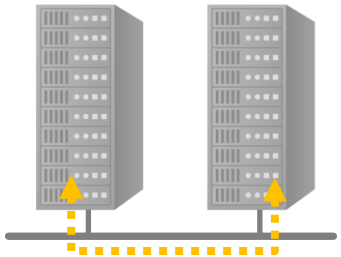


- CPUを介さずメモリを直接転送
- 高スループット・低遅延を実現
- GPU・高速ストレージはRDMA対応が主流

© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

長距離RDMAと分散計算基盤

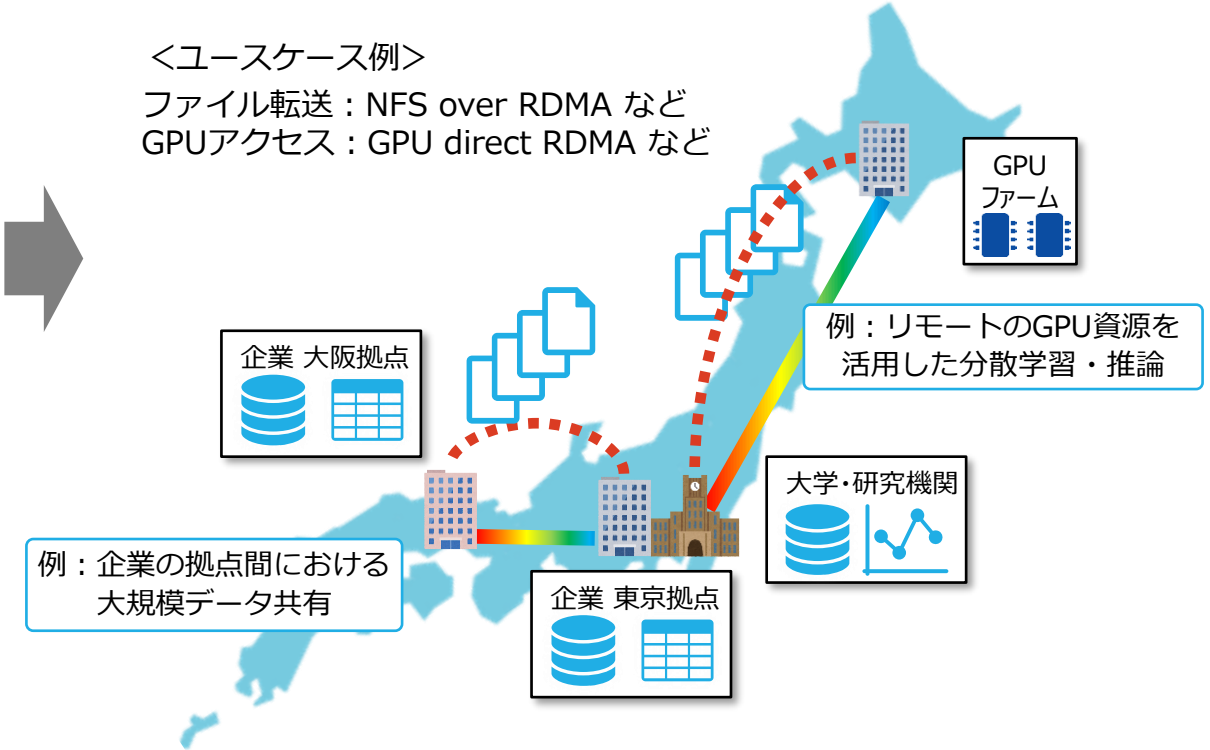
現在のRDMA



HPC向けの高速通信技術として発展
現在ではクラウドやAI基盤を支える
データセンタ内通信へと普及

※HPC: High Performance Computing

<ユースケース例>
ファイル転送：NFS over RDMA など
GPUアクセス：GPU direct RDMA など



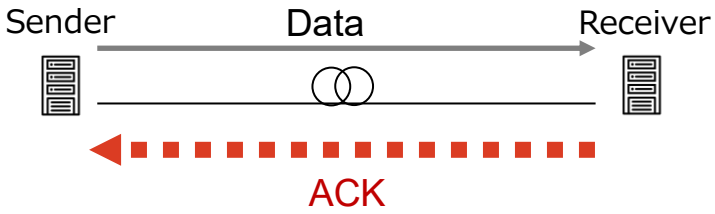
RDMAを活用した遠隔・大規模データへの高速アクセス

© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

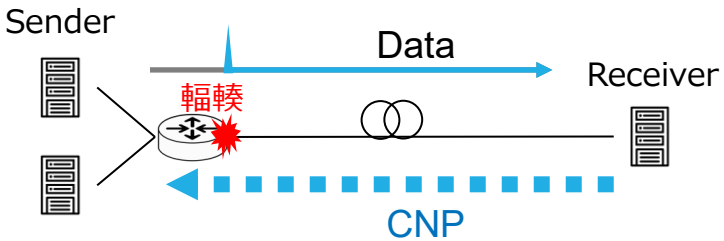
長距離RDMAの課題：フィードバック遅延

長距離RDMAでは，光伝播遅延が原因で帯域が低下する

RDMAでは、メッセージごとに通信を行い、宛先からのACK（確認応答）の受信をもって転送完了とみなし、次の通信に進む



経路上で輻輳が発生すると、SWがECN bitを立てて転送し、宛先がCNP（輻輳通知）を送信して送信元の送信量を下げる



	課題①	課題②
フィードバック	ACK	CNP
遅れると起きること	ACK を待つ間に送信が止まり、スループットが落ちる	輻輳への反応が遅れ、制御が大味になる（反応が足りない場合と効きすぎる場合の両方がある）
メッセージサイズを大きくすると	緩和できる	緩和できない

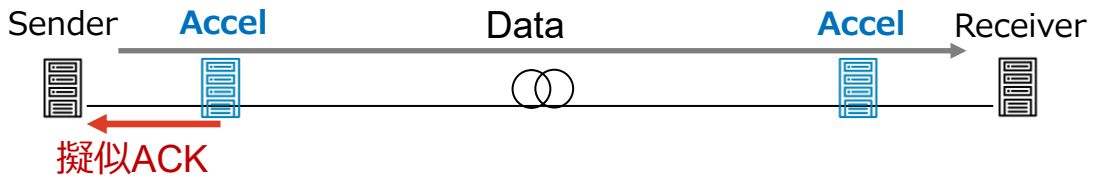
© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

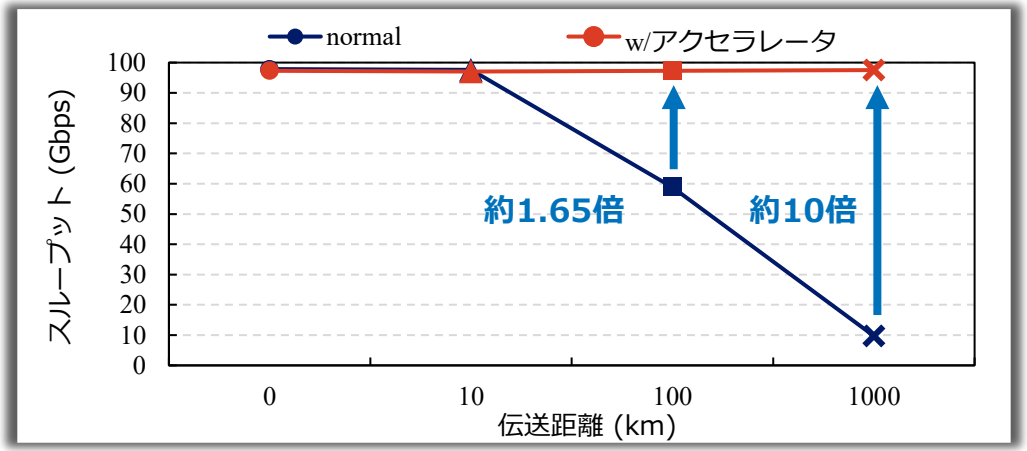
課題①の対策：擬似ACK方式

NTT TOYOTA

NTTのRDMAアクセラレータは、擬似ACK によってスループット低下を解決

送信サーバはパケットを
途切れず連続送信
→ **スループット改善**





メッセージサイズ128KiB, ロス率0%

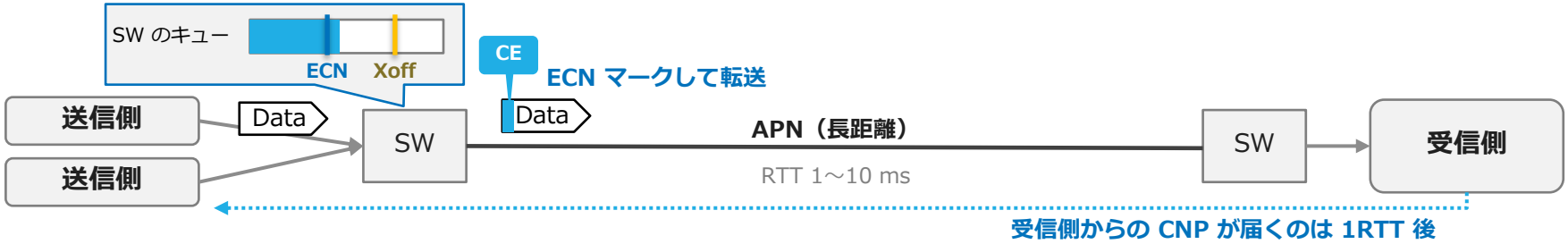
© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

残る課題：長距離 × 輻輳（課題②）

NTT TOYOTA

スループットの問題を解決しても、DC間で輻輳が起きると性能が劣化する

- ECN（Explicit Congestion Notification）：閾値を超えたことをパケットヘッダで通知する仕組み
  - CNPが送信元に届いて減速が始まるまでに1RTTかかる



- 長距離NWでは、ECNが効く前にPFC（Priority Flow Control）が発動してしまう
  - 一定時間の送信停止によりバッファ溢れを防ぐ代わりに、スループットを大きく下げる要因に



© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

RDMAアクセラレータの仕組み

NTT TOYOTA

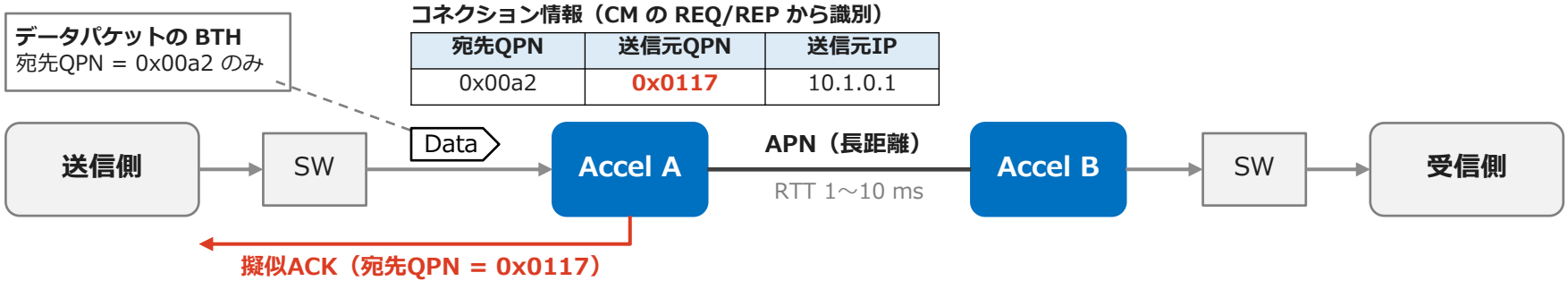
中継パケットを見て、双方向のコネクション情報（QPN）を識別し制御に活用

特徴

- 送受信サーバの近くで、長距離NWを挟み込むように配置
- 送受信サーバに特別な変更は不要で、既存のRDMAアプリをそのまま長距離化できる

仕組み

- 中継するRDMA CMのREQ/REPを見て、双方向のコネクション情報（QPN）を識別する
  - データパケットには宛先のQPNしか載っていない。送信元へ擬似ACKを返すには、この情報が欠かせない
  - **この情報を輻輳制御にも活用する**のが今回の提案



© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

提案：アクセラレータによるFast CNP

NTT TOYOTA

受信側からのCNPを待たずに、通信路の途中から送信元にブレーキをかける

- 従来のECNでは、受信側しかCNPを返せない。長距離では、制御が効くまでに少なくとも1RTTかかる



- アクセラレータは、ECN(CE)マークの付いたパケットを受信すると、すぐに送信元へCNPを返す
  - コネクション情報から送信元のQPを特定し、標準のCNPで返す。NICの改変は不要
  - μsオーダーで送信元の減速を開始できるため、PFCによる一時停止を抑制でき、速度を下げすぎない



© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

擬似ACKとFast CNPの相補作用

アクセル（擬似ACK）とブレーキ（Fast CNP）を、異なるレイヤーで組み合わせる

	擬似ACK	Fast CNP
操作する対象	パケットの送信量：上げる	NICの送信レート：下げる
役割	長距離でのスループット維持	緊急時の制動
応答速度	メッセージ単位	パケット単位

- 2つを組み合わせることで、**スループットを維持しながら、DC間での緊急制動を高速に実現する**

© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

パート２：まとめ

- 光伝播遅延が原因でRDMAの転送帯域が劣化する
  - ACK遅延：非輻輳化でのボトルネック
  - 輻輳通知遅延：輻輳化でのボトルネック

- 「擬似ACK」と「fast CNP」というアイデアで，光伝播遅延の影響を隠蔽可能

© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

パート３：検証結果

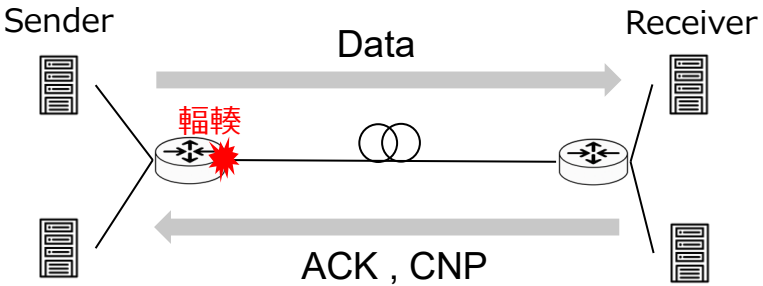
© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

実験構成

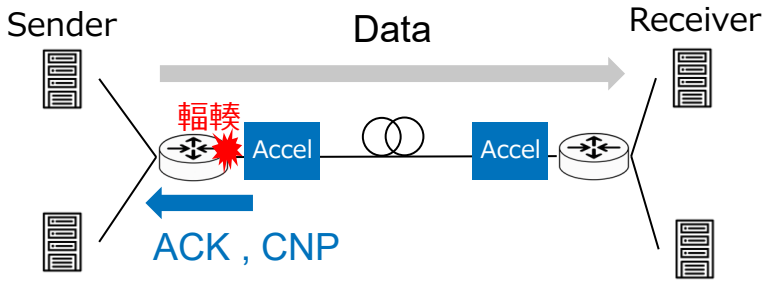
NTT TOYOTA

共通条件： 全区間 100Gbps Link
擬似遅延 20km ~ 1000km相当
測定ツール：perftest (QP数=1, txdepth=128, DMA Length = 1MiB)

基本構成 (default)



提案方式 (proposed)

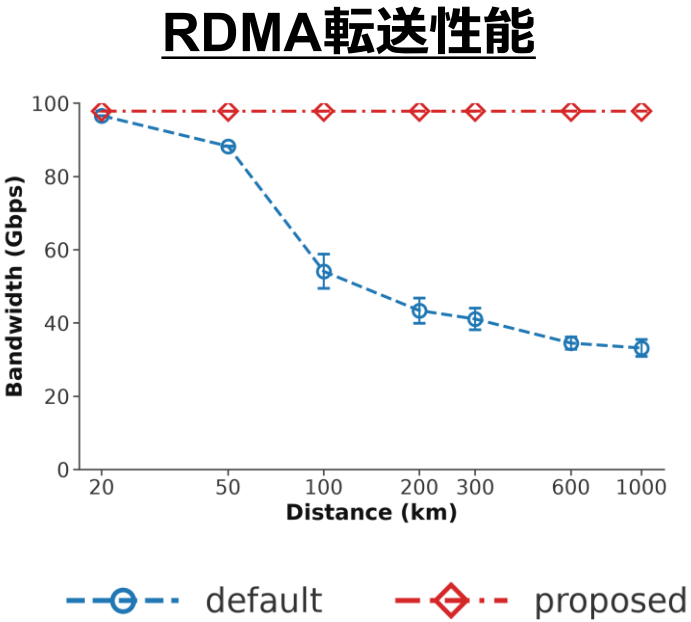


© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

結果

提案方式により，

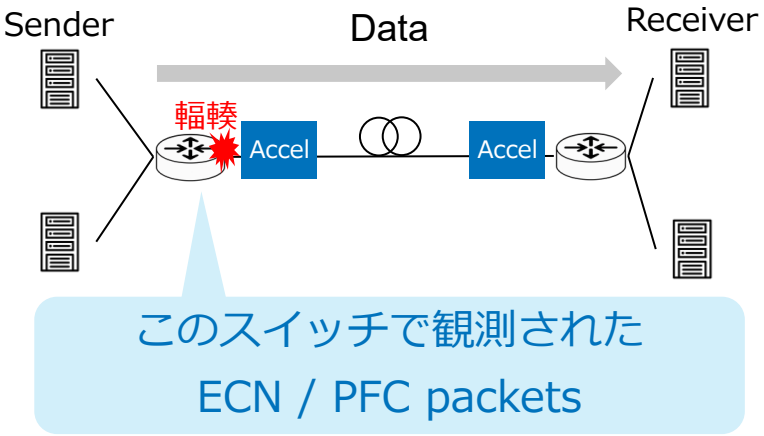
輻輳条件下でも 1000km – 100GbpsのRDMA通信を実現

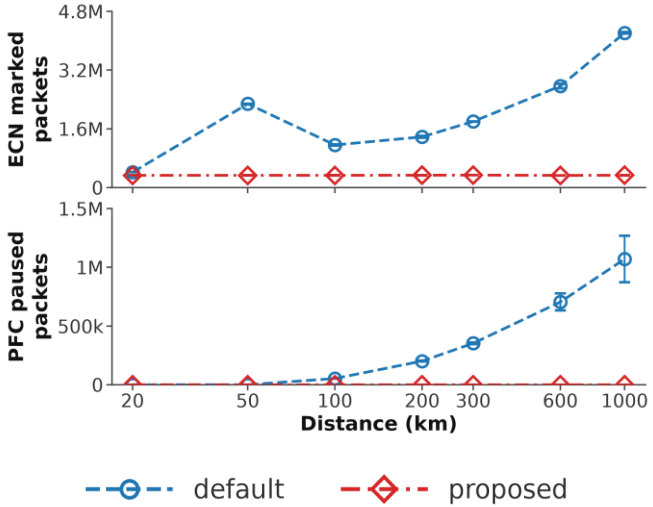


© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

考察：輻輳制御の効果

fastCNPにより，ECN / PFC packetの減少を確認
PFCは全距離で0





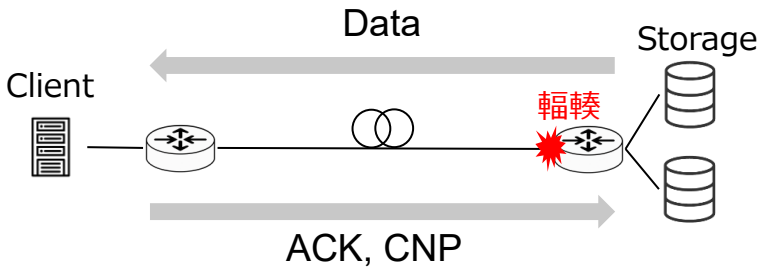
© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

実験構成：NFS over RDMA

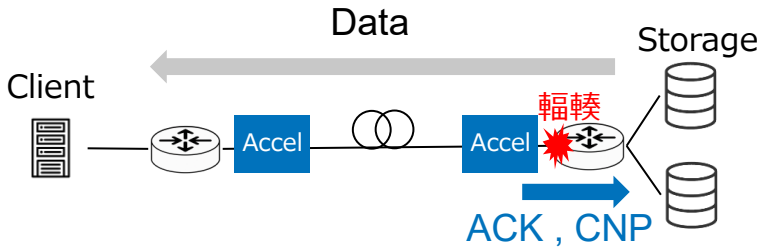
NTT TOYOTA

共通条件： 全区間 100Gbps Link
擬似遅延 20km ~ 1000km相当
測定ツール：fio (block size=256KB, num jobs=32, iodepth=16)

基本構成 (default)



提案方式 (proposed)



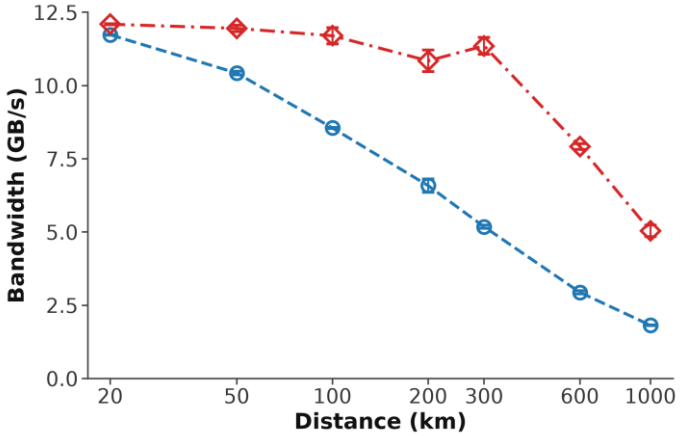
© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

結果

NTT TOYOTA

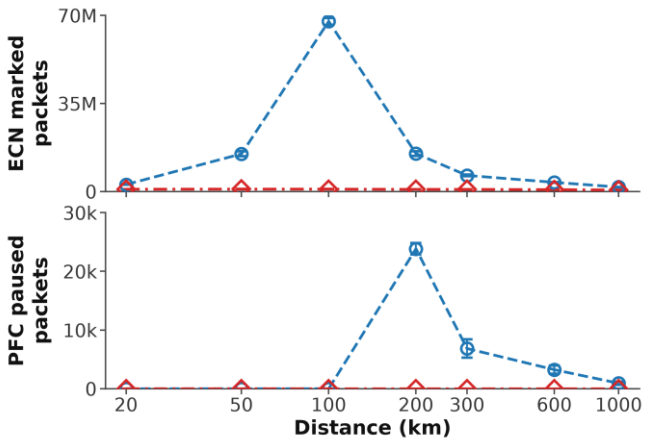
データ読出し性能

300km まで 性能を維持



輻輳制御

ECN / PFC packetsの減少



default proposed

© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

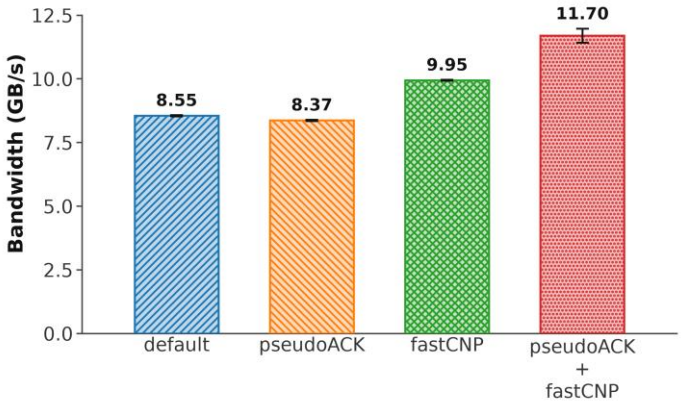
考察：距離別 有効機能

NTT TOYOTA

NFS over RDMAでは距離に応じて有効な機能が異なる

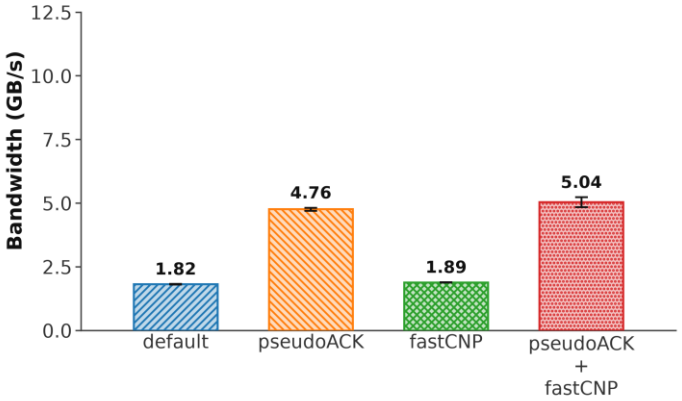
100 km

輻輳ボトルネック
→ fastCNPにより性能向上



1000 km

ACK遅延ボトルネック
→ 擬似ACKにより性能向上



© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

考察：300km以降の性能劣化要因

NTT TOYOTA

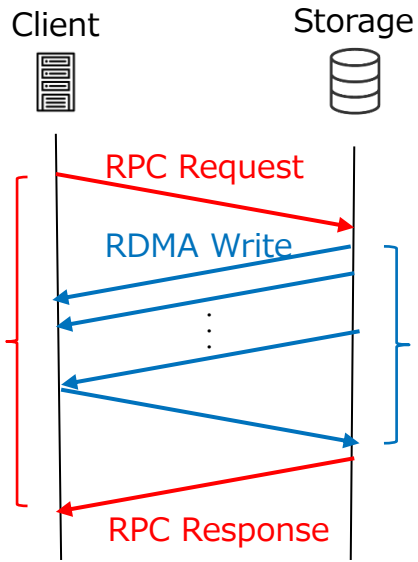
同時発行 RPCリクエスト数がボトルネックとなり，
RDMAトラヒックでネットワーク帯域を埋められないこと発覚

→ 改善方式：今後の課題

性能ボトルネック

RPC over RDMAの仕様により
同時発行 Request数に制限有

= RPC Response待ち状態



提案方式で性能向上

© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.

まとめ

NTT TOYOTA

【モチベーション】分散計算基盤におけるデータアクセス高帯域化

【提案】光伝播遅延がRDMAに与える悪影響を隠蔽し長距離RDMAの高帯域化



fastCNP & 擬似ACKの組合せで長距離高性能通信実現の可能性

© 2026 NTT, Inc. and TOYOTA MOTOR CORPORATION. All Rights Reserved.


NTT


TOYOTA