AI/ML基盤におけるGPU間ネットワークの負荷と性能影響を探る

トヨタ自動車株式会社　加納浩輝　奥澤智子

目次

- **1部：AI/ML基盤におけるGPU間通信の特性理解**
  - NW性能目標値ってどれくらい？
  - どれくらいトラヒックが流れるの？
  - アプリケーションを意識した性能測定方法は？

- **2部：分散型計算基盤に向けた課題と展望**
  - 分散型計算基盤の構想
  - 長距離RDMAの課題と検討

TOYOTA

第一部

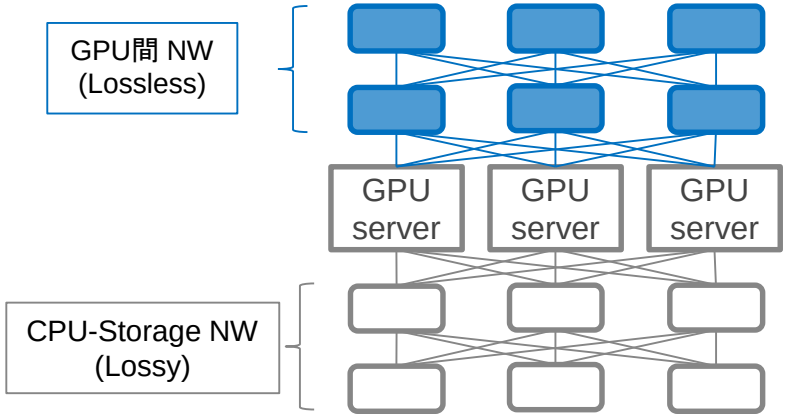
AI/ML基盤における GPU間通信の特性理解

- NW 性能目標値ってどれくらい？
- どれくらいトラヒックが流れるの？
- アプリケーションを意識した性能測定方法は？

GPU間通信 NWの現状 と 悩み

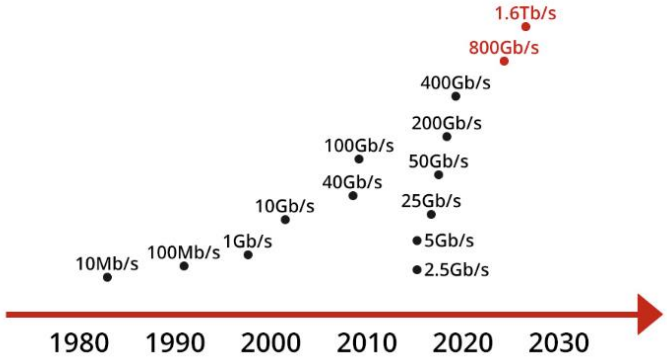
TOYOTA

GPU間通信のために 専用NWの構築



専用NWなら最適化したい
AI/ML Jobにとって最適なNWとは？

物理帯域の増加



出展：400Gから1.6Tへ：光モジュールの進化と革新, Fs.com

AI/ML Jobのためには
どこまで高性能なNWが必要となるのか？

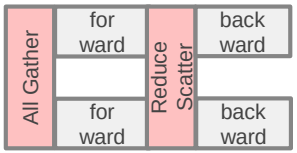
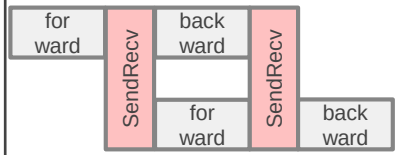
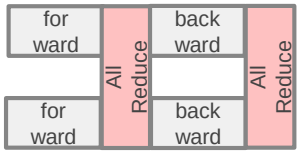
第一部 テーマ

- GPU間通信 NWとAI/ML Jobの関係を検証してみた話
  - NWがAI/ML Jobのボトルネックにならないための性能目標値は？
  - 具体的にどれくらいトラヒックが流れるの？
  - アプリケーションを意識したNW性能試験をするには？

- （話さないこと）
  - Lossless NWの構築方法やDCQCNやDynamic LB等の検証結果

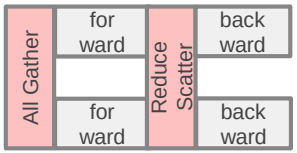
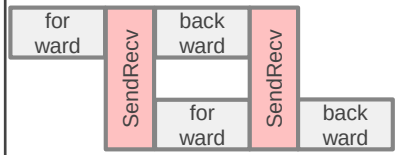
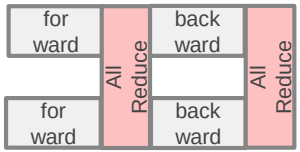
AI/MLで利用される主な分散方式

- 学習高速化やメモリ消費削減などの目的に応じて分散方式を使い分ける
- 分散学習時のGPU間のデータ通信形式・頻度は分散方式によって異なる

	データ並列	FSDP	パイプライン並列	テンソル並列
				
データ	分散	分散	冗長	冗長
モデル	冗長	一時的に分散	分散	分散
通信形式	AllReduce	RdeseScatter + AllGather	SendRecv	AllReduce など
通信頻度	Step毎	レイヤ毎	レイヤ毎	レイヤ毎

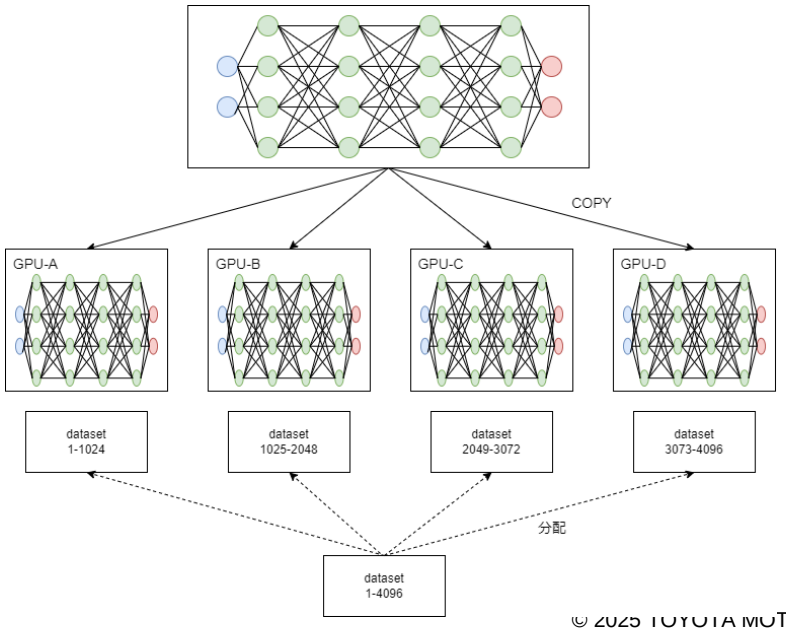
AI/MLで利用される主な分散方式

- 学習高速化やメモリ消費削減などの目的に応じて分散方式を使い分ける
- 分散学習時のGPU間のデータ通信形式・頻度は分散方式によって異なる
- 今回はデータ並列に関して、ネットワーク性能との関係を分析

	データ並列	FSDP	パイプライン並列	テンソル並列
				
データ	分散	分散	冗長	冗長
モデル	冗長	一時的に分散	分散	分散
通信形式	AllReduce	RdeseScatter + AllGather	SendRecv	AllReduce など
通信頻度	Step毎	レイヤ毎	レイヤ毎	レイヤ毎

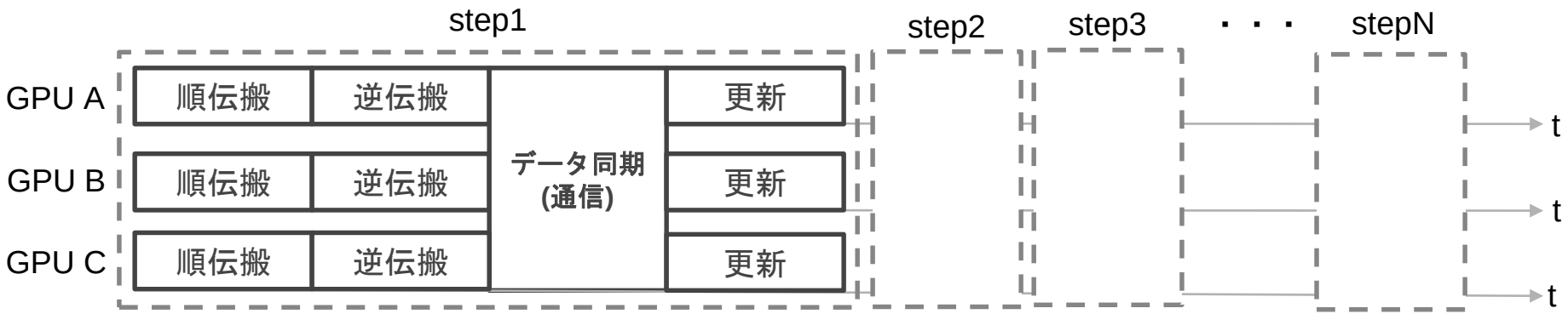
データ並列

- 学習の高速化を目的とした分散方式
- 各GPUへ同一のモデルをコピー ＆ 学習データは分配し分散学習を行う
  - **各GPUのモデルの状態が常に同一となるようにGPU通信によるデータ同期が必要**



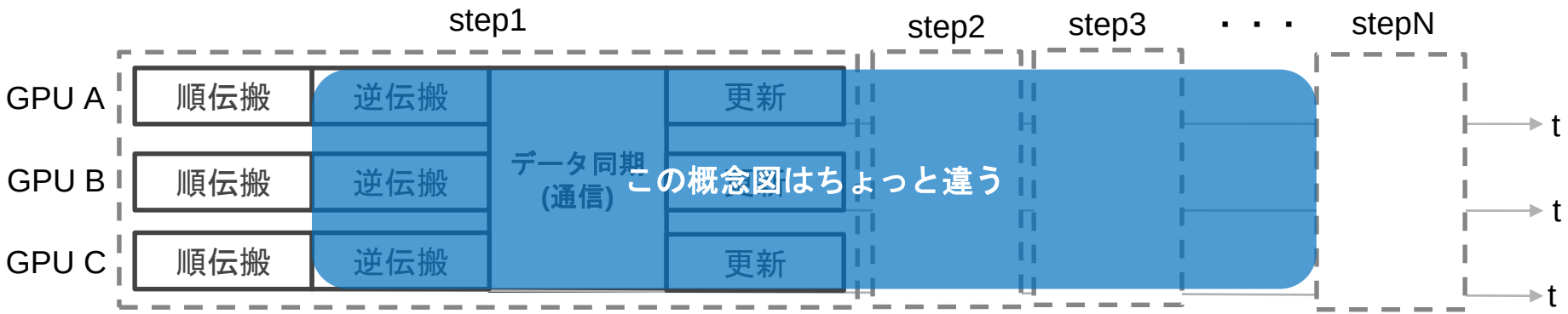
データ並列における計算と通信の関係

- 各stepの 逆伝搬 と パラメータ更新の間でGPU間通信が発生
- データ同期が完了するまで次の処理が開始しない
  → ネットワーク性能が学習全体のボトルネックになる可能性
- **通信時間 = 計算が止まる時間 (GPU idle 時間) なのか？**



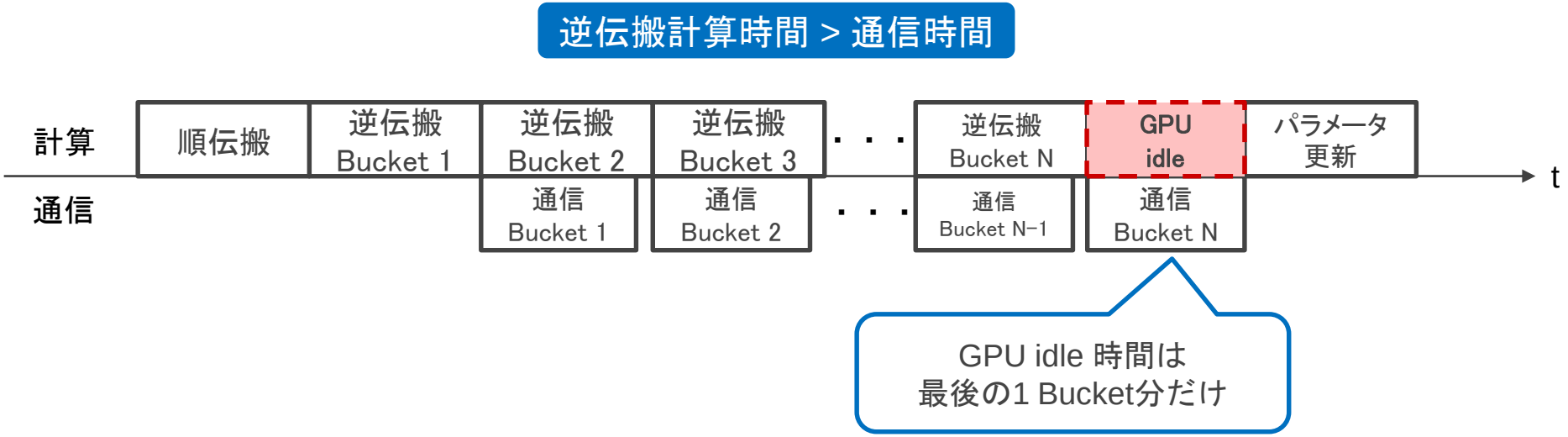
データ並列における計算と通信の関係

- 各stepの 逆伝搬 と パラメータ更新の間でGPU間通信が発生
- データ同期が完了するまで次の処理が開始しない
  → ネットワーク性能が学習全体のボトルネックになる可能性
- **通信時間 = 計算が止まる時間 (GPU idle 時間) なのか？** ➡ **No！**



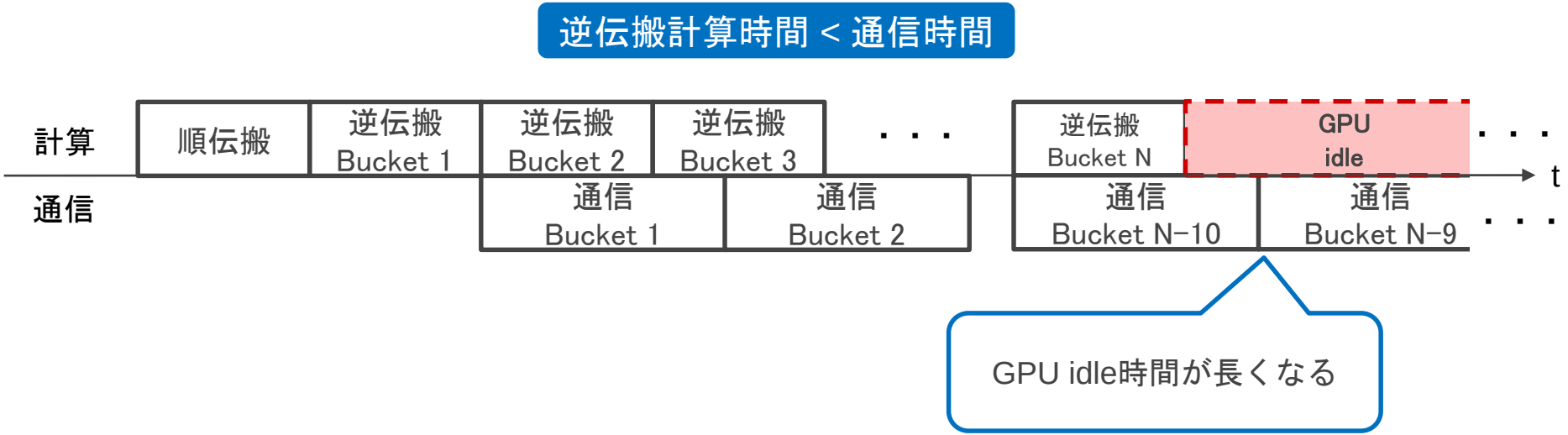
通信-計算 並列実行の仕組み と その効果

- 逆伝搬計算と通信が並列実行されている
  - 更新を行いたい全パラメータのBucketに分割
  - Bucket単位で、計算/通信を並列実行



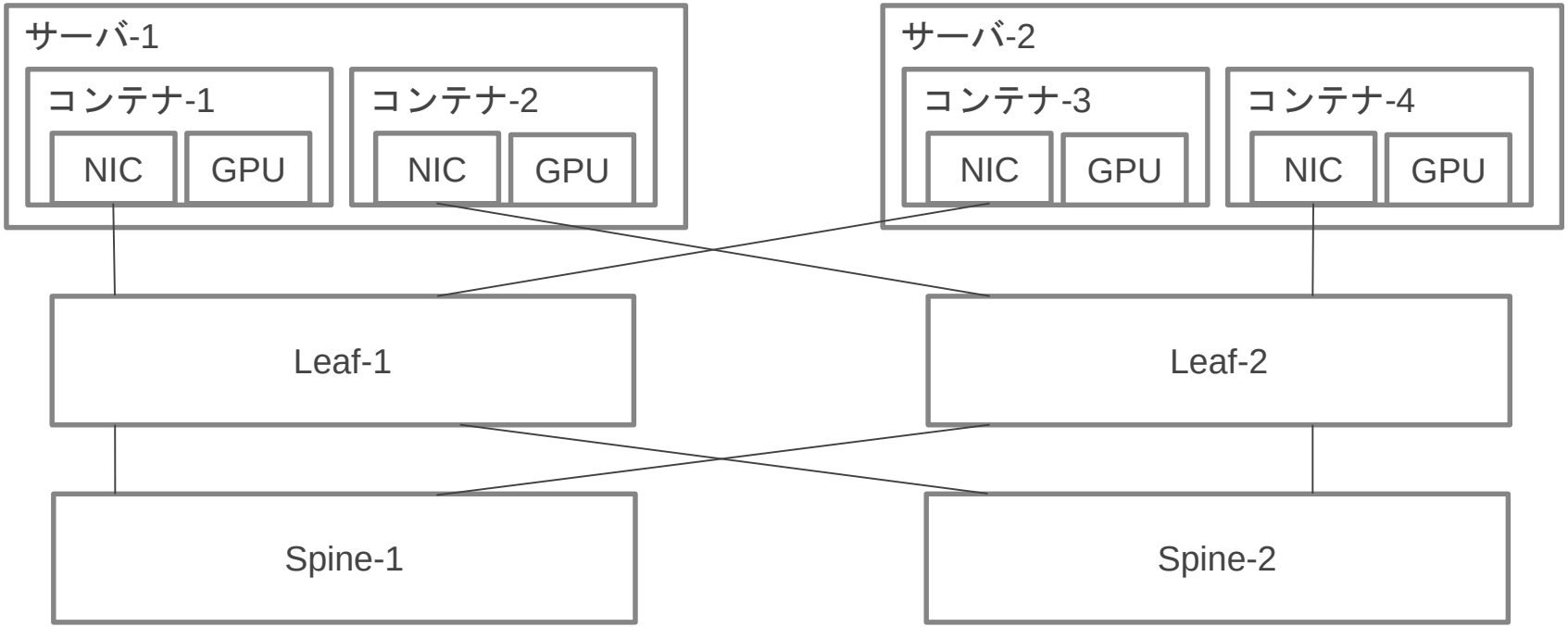
通信-計算 並列実行の仕組み と その効果

- 逆伝搬計算と通信が並列実行されている
  - 更新を行いたい全パラメータのBucketに分割
  - Bucket単位で、計算/通信を並列実行



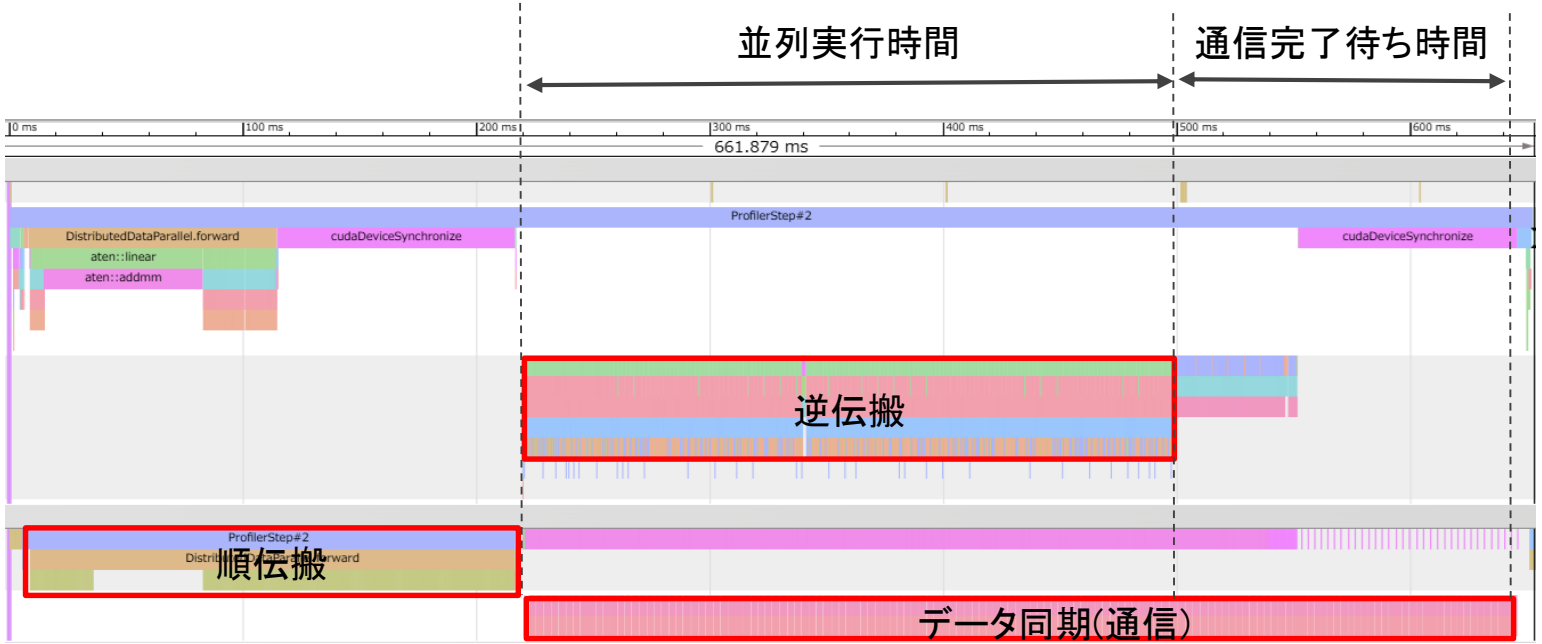
実験環境

- ネットワーク
  - GPU間通信： RoCEv2（ノード内の通信は未利用）
  - サーバNIC帯域：100Gbps/コンテナ
- 学習Job
  - k8s上でPytorch Jobを用いて実行
  - 分散方式：Pytorch DDP



実験結果：通信−計算 並列実行

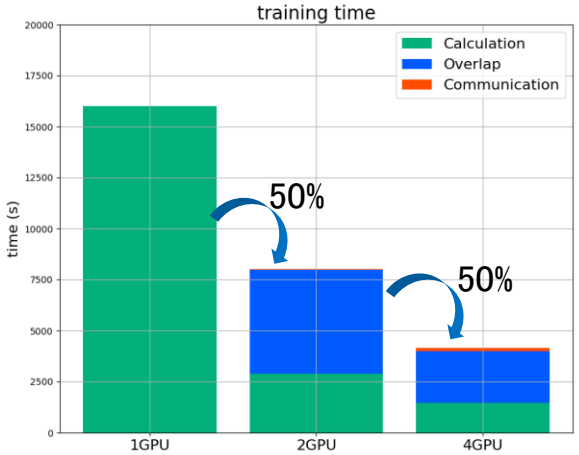
- PyTorch Profileを用いて、逆伝搬と通信の並列実行を確認



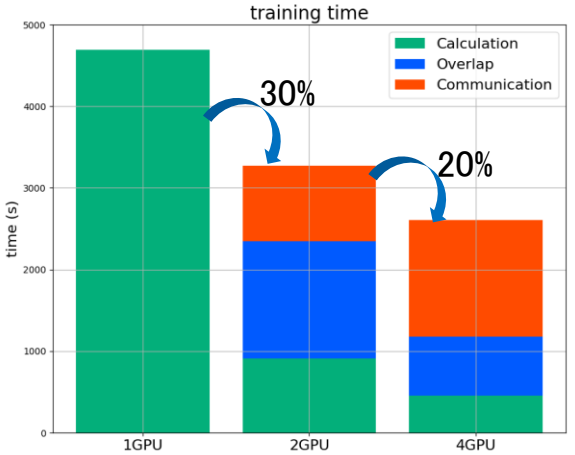
実験結果：通信−計算 並列実行

- 2種類の学習モデルの「計算のみ」、「計算+通信同時」、「通信のみ」の時間を分析
- 逆伝搬時間内に通信が完了するモデルBでは、GPU分散数に対して線形に学習時間が減少

モデルA：
逆伝搬時間 > 通信時間



モデルB：
逆伝搬時間 < 通信時間



※両モデルでパラメータサイズ（同期するデータ量）は1Billion固定
※計算量を変更することで、大小関係を調節

（まとめ）データ転送完了時間の目標値は？

- 計算と通信は並列実行されるため、通信時間 ≠ GPU idle時間
- GPU idle時間は、逆伝搬時間と通信時間の大小関係に依存

（まとめ）データ転送完了時間の目標値は？

- 計算と通信は並列実行されるため、通信時間 ≠ GPU idle時間
- GPU idle時間は、逆伝搬時間と通信時間の大小関係に依存

逆伝搬時間 > 通信時間を達成するには、

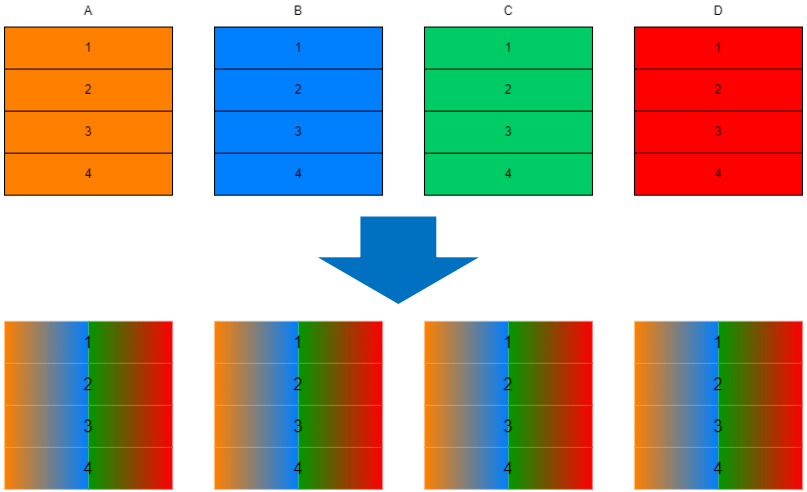
「**逆伝搬時間 > (送信したいデータ量 / 帯域)**」が必要条件

アプリケーション依存

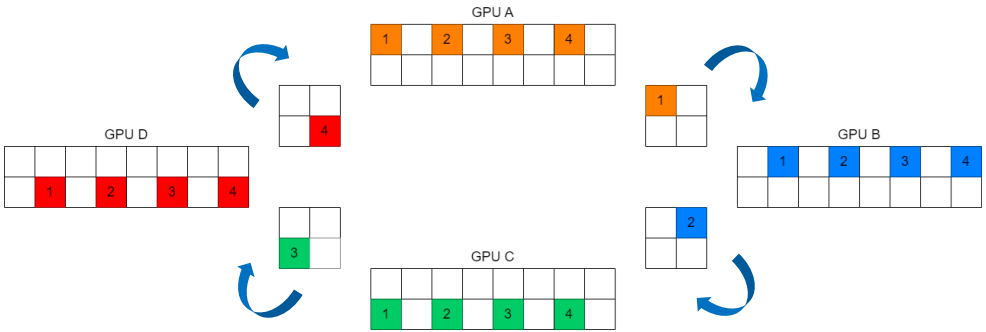
NWエンジニアが構築＆パフォーマンスチューニング

データ並列における通信形式

- AllReduce：各GPUが持つデータを集めて、足し合わせる
  - 各GPUがリング状のトポロジーを組み、データをやり取りすることが一般的（Ring-AllReduce）



All Reduce概念図



RingAll Reduce概念図

具体的にどれくらいトラヒックが流れるの？

- 1 Stepで1GPUが送信するデータ量は以下の式であらわされる

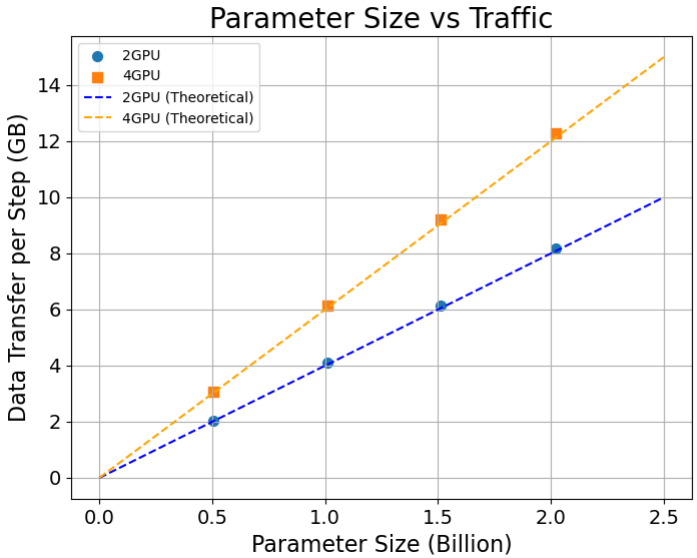
$$\frac{2(N-1)}{N} \times 4P$$

N: 分散GPU 数
P: パラメータ数

- 分散GPU数とパラメータ数がわかれば、ネットワーク負荷は概算できるはず

実験結果：具体的にどれくらいトラヒックが流れるの？

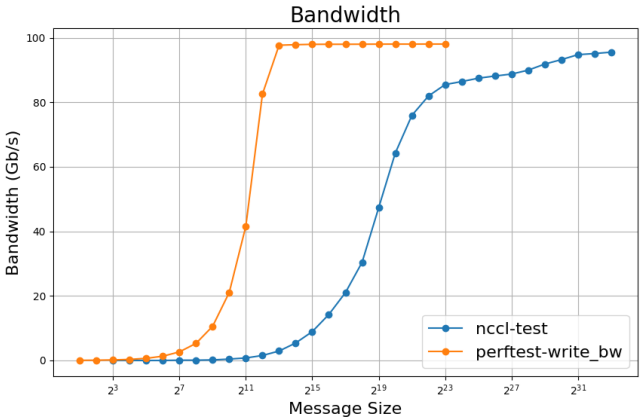
- パラメータ数を変えながら、GPUが送信するトラヒック総量を計測
- 理論値と概ね一致する結果が得られた
  - → **分散GPU数とパラメータ数がわかれば、ネットワーク負荷は概算できる!**



アプリケーションを意識したNW性能試験をするには？

- AIML基盤におけるRDMA性能試験では、主にperf-tests/nccl-testsが利用される
  - Message Sizeに対する、スループットや送信時間を評価する
- perf-tests と nccl-testsで**Message Sizeの定義が異なる**ので注意が必要

	perf-tests	nccl-tests
目的	RMDA通信性能試験	Collective通信性能試験
Message Size	単一のRDMA通信	単一のCollective通信



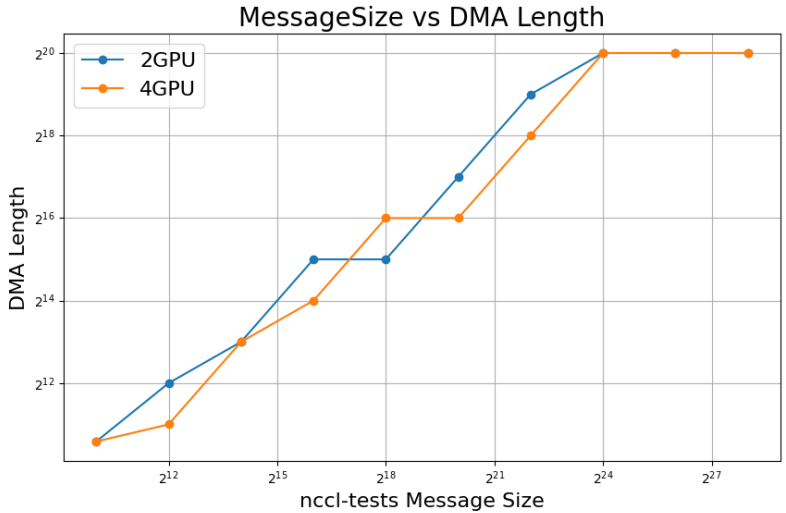
perf-tests におけるMessage Size

- 定義
  - 単一RDMAメッセージ(RDMA Write First ~ RDMA Write Last) の間に転送されるデータ量
  - RDMA Write FirstのRDMA Extended Transport Header内にDMA Length記載

```
> Frame 99: 4170 bytes on wire (33360 bits), 4170 bytes captured (33360 bits)
> Ethernet II, Src: 4a:c7:2c:89:fb:5c (4a:c7:2c:89:fb:5c), Dst: 00:00:00_00:00:10 (00:00:00:00:00:10)
> Internet Protocol Version 4, Src: 192.168.30.179, Dst: 192.168.31.179
> User Datagram Protocol, Src Port: 62337, Dst Port: 4791
v InfiniBand
  v Base Transport Header
      Opcode: Reliable Connection (RC) - RDMA WRITE First (6)
      0... .... = Solicited Event: False
      .1.. .... = MigReq: True
      ..00 .... = Pad Count: 0
      .... 0000 = Header Version: 0
      Partition Key: 65535
      Reserved: 00
      Destination Queue Pair: 0x00036b
      1... .... = Acknowledge Request: True
      .000 0000 = Reserved (7 bits): 0
      Packet Sequence Number: 0
  v RETH - RDMA Extended Transport Header
      Virtual Address: 0x0000000302530000
      Remote Key: 0x0003d2b7
      DMA Length: 1048576 (0x00100000)
    Invariant CRC: 0x5ebf7d56
> Data (4096 bytes)
```

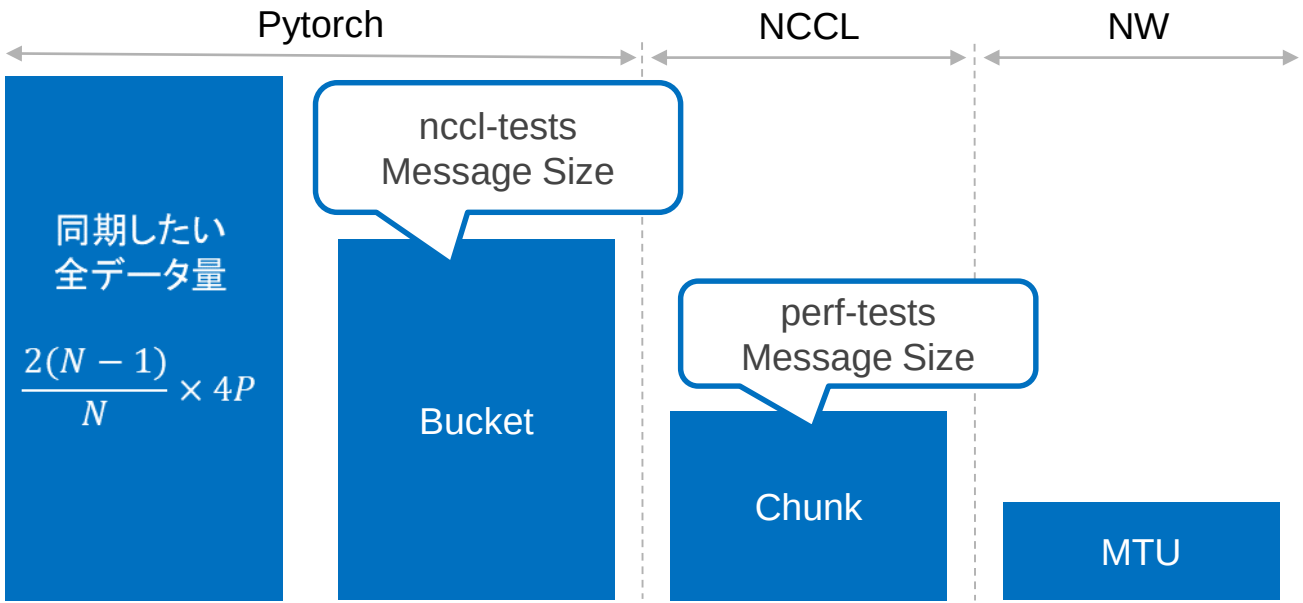
nccl-tests におけるMessage Size

- 定義:
  - 単一のCollective通信で送信されるデータ量
  - RDMAヘッダに記載される値 (DMA Length)とは異なる
- Message Size(perf-tests) = Chunk Size (nccl-tests)
  - ChunkSizeは分散GPU数やオプションパラメータに応じてNCCL側で自動決定される



AI/ML workloadにMessage Size

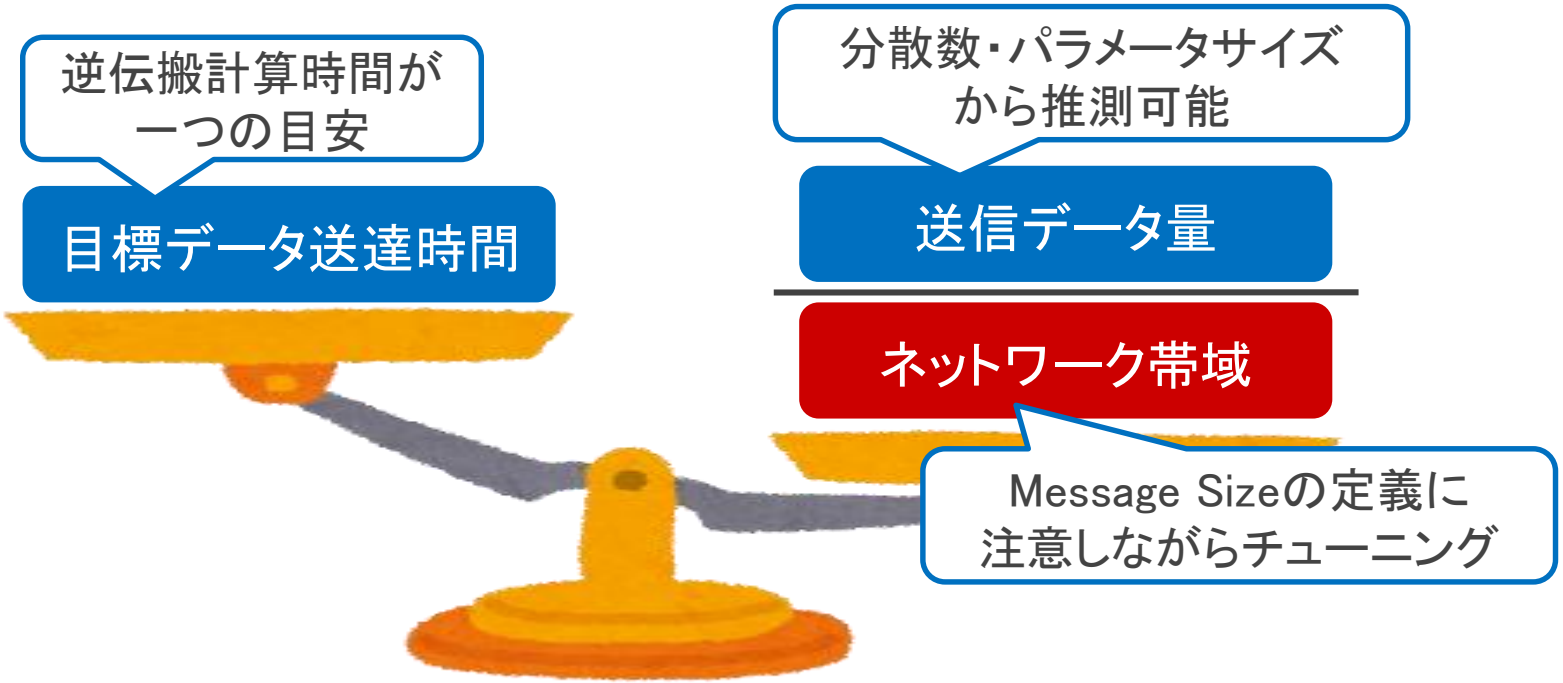
- 全データ→Bucket→Chunk→MTU の順でデータが分割され送信される
  - ベンチマークツール毎にどの段階のメッセージサイズを測っているかが違う
- pytorch/ncclのデフォルト値では Chunkは1M程度



※pytorchではデフォルト値25MB

第1部 まとめ

- AI/ML Jobの特性から決まるパラメータが、ネットワーク性能をどこまで向上していくかを議論する重要な指標になる



TOYOTA

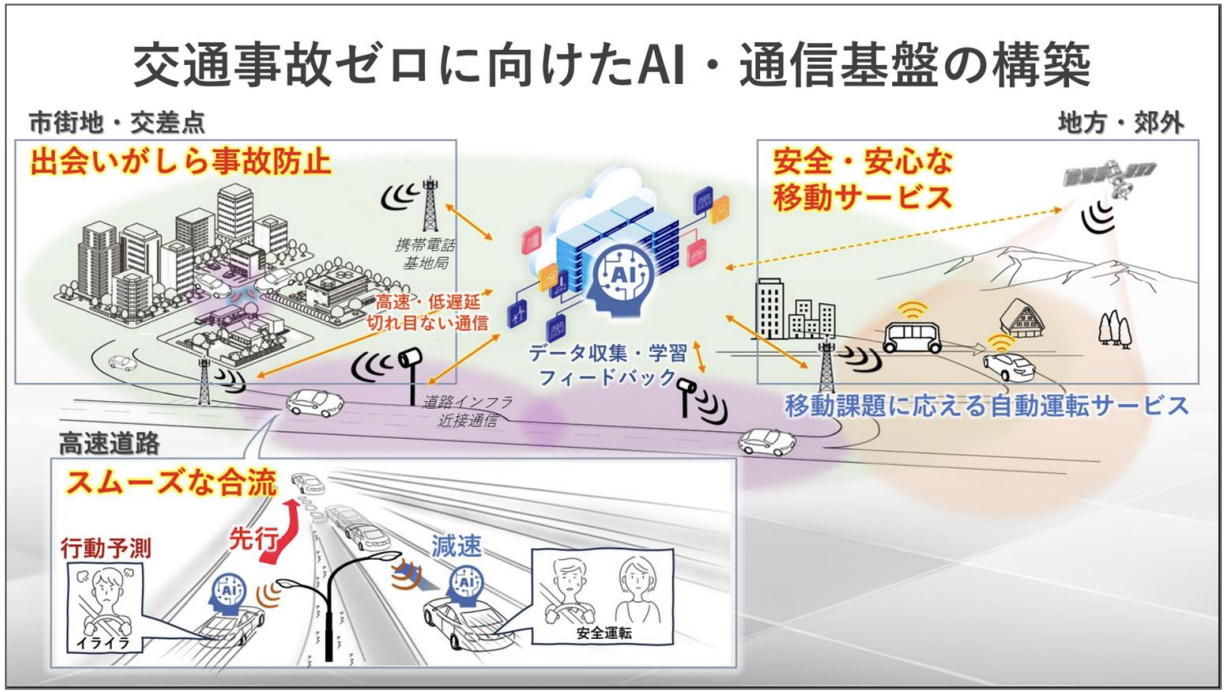
第2部

分散型計算基盤に向けた課題と展望

- 分散型計算基盤の構想
- 長距離RDMAの課題と検討

そもそも、なぜトヨタが計算機基盤を作るのか

- モビリティ x AI・通信



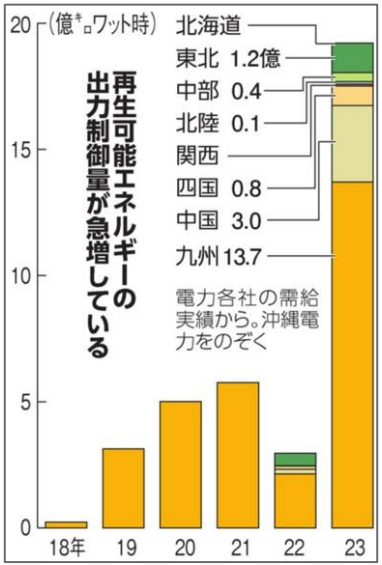
出典：https://global.toyota/jp/newsroom/corporate/41746612.html

余剰グリーン電力を使った分散計算機基盤への取組

- データセンター需要、半導体、AIの利用などで電力消費が増加

→再エネの豊富な地域で計算資源を確保したい

九州：豊富な再エネへの期待



出典：朝日新聞

地域の再エネを活用した地産地消の分散コンピューティング基盤

北海道：風力＋冷却への期待

ユーラス、北海道で31年にも最大級風力　AI電力需要照準

資源エネルギー　＋フォローする

2024年5月21日 10:31 [会員限定記事]

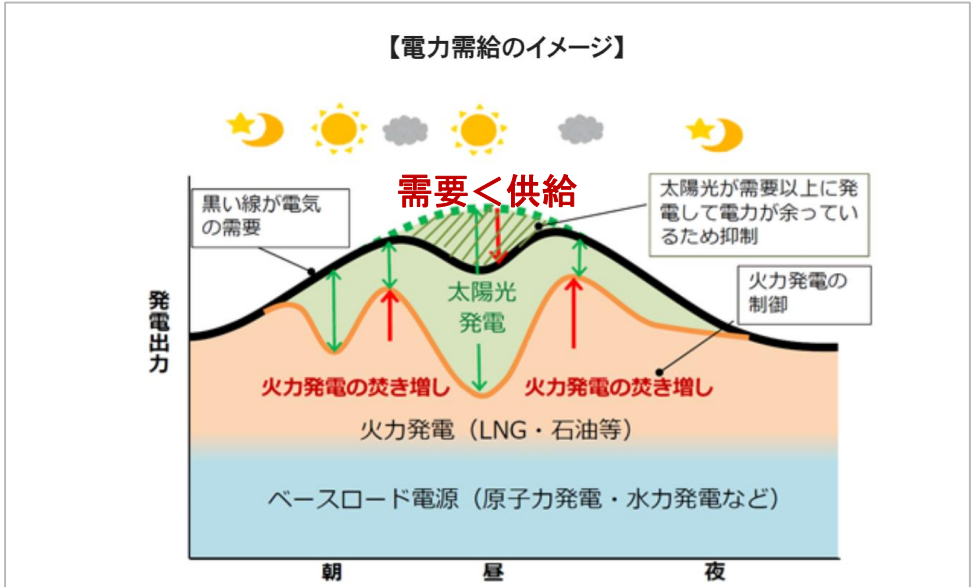
保存

豊田通商子会社で風力発電国内最大手のユーラスエナジーホールディングス（東京・港）は、北海道北部で国内最大級の陸上風力発電事業に乗り出す。総出力は165万キロワットで、早ければ2031年ごろから稼働させる。生成AI（人工知能）で需要が高まるデータセンター（DC）を誘致し、再生可能エネルギー電力を地産地消する体制を築く。

出典：日経新聞

再生可能エネルギーの出力抑制(1)

- 電気の需給バランスを取るため、供給過多時は出力抑制される
  - 太陽光や風力は発電量の制御が難しい



出典：https://www.enecho.meti.go.jp/category/saving_and_new/saiene/grid/08_syuturyokuseigyo.html

再生可能エネルギーの出力抑制(2)

- 抑制されている電力量

2024年度の各エリアの再エネ出力制御見通し等

	北海道	東北	中部	北陸	関西	中国	四国	九州	沖縄
出力制御率見通し（2024年度） 出力制御率(%) ※2 [制御電力量(kWh)]	**0.2%** [0.1億kWh]	**2.5%** [4.0億kWh]	**0.6%** [1.0億kWh]	**1.1%** [0.2億kWh]	**0.7%** [0.8億kWh]	**5.8%** [5.7億kWh]	**4.5%** [2.4億kWh]	**6.1%** [10億kWh]	**0.2%** [87万kWh]
仮に、エリア全体がオンライン化 した場合 出力制御率(%) [制御電力量(kWh)]	0.1% [0.05億kWh]	1.5% [2.4億kWh]	0.5% [0.8億kWh]	1.1% [0.2億kWh]	0.5% [0.5億kWh]	5.2% [5.0億kWh]	3.9% [2.1億kWh]	6.1% [10億kWh]	0.07% [37万kWh]
連系線利用率 ※3	50%	北本-50% /東北東京 80%	-20%	10%	-20%	0%	30%	95%	–
最低需要 ※4 **（2022年度）[万kW]**	280	719	1,056	222	1,190	475	226	718	70.5
変動再エネ導入量（2022年度）[万kW]	300	1,030	1,156	139	716	699	361	1,216	40.2
変動再エネ導入量/最低需要（2022年度）[%]	107%	143%	109%	63%	60%	147%	160%	169%	57%
（参考）出力制御率見通し（2023年度想定更新後） ※5 出力制御率(%)	0.01%	0.93%	0.26%	0.55%	0.20%	3.8%	3.1%	6.7%	0.14%

出典：https://www.meti.go.jp/shingikai/enecho/shoene_shinene/shin_energy/keito_wg/pdf/050_01_00.pdf

トヨタグループでの営み

豊田通商株式会社

Eurus Energy

- トヨタ x 豊田通商 x ユーラスエナジー
- カーボンニュートラルの実現へ
- ユーラスエナジー
  - 北海道だけで21の発電所を操業中
- 再エネの有効利用に関するPoCを実施中
  - 北海道稚内市中心に計画

長距離RDMAの実現可能性は？

- UC例
  - 遠隔地からの学習データ転送
  - 拠点間GPUクラスタの構築、など
- 札幌～稚内間で実施する場合
  - 距離：約300km
  - 光の遅延：1.5ms(片道) ※1km = 5μsで試算

先行研究・事例(1)

- ”Long-distance RDMA-acceleration Frameworks” (NTT)
  - 長距離間でレスポンダからのACKが遅延することによって、性能低下
  - RDMA WAN アクセラレータにて擬似ACKを生成・送信し、距離による遅延の性能影響を軽減させる

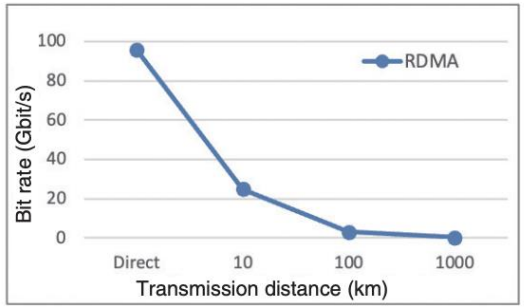


Fig. 3. Impact of RDMA throughput based on communication distance.

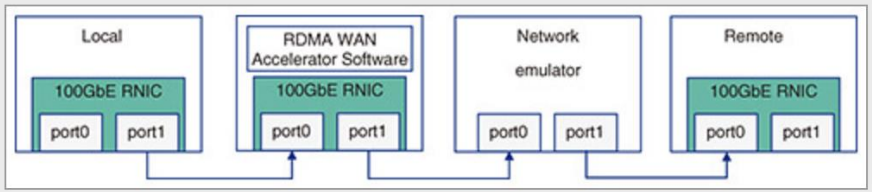
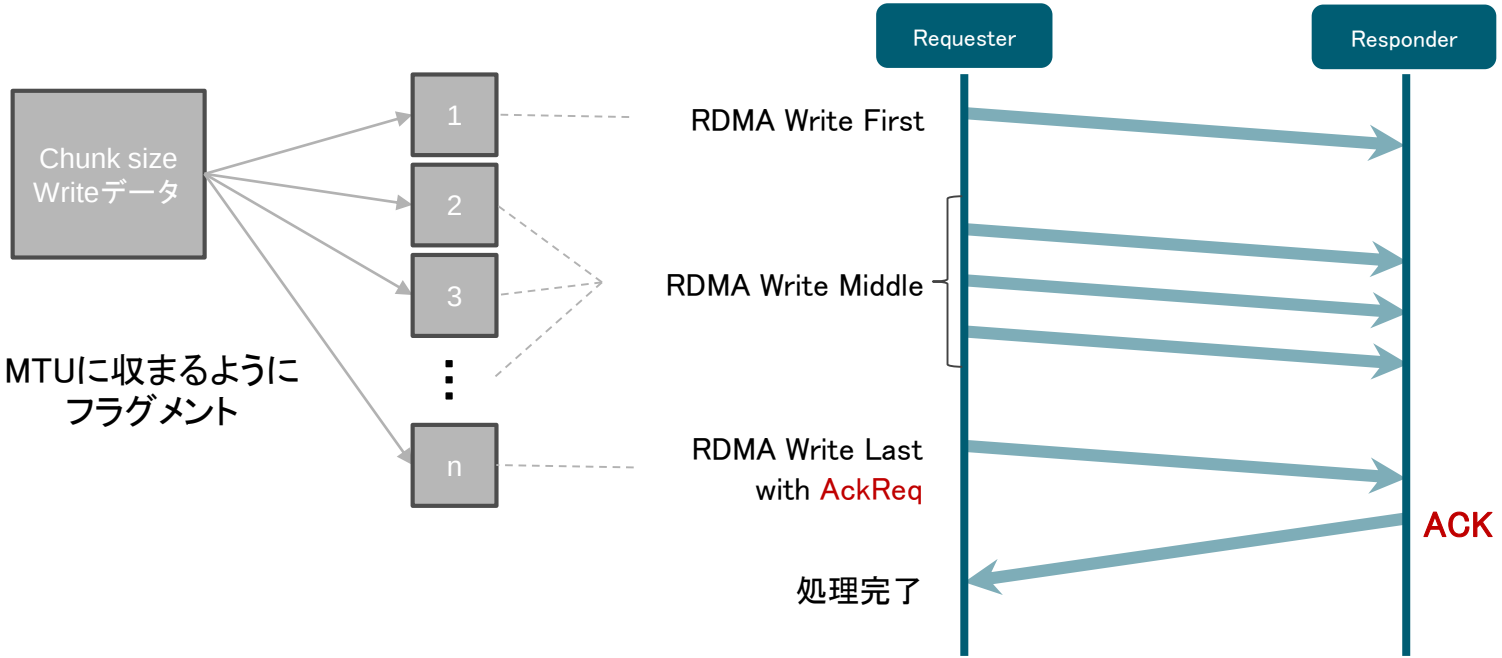


Fig. 5. The experimental setup for evaluating throughput of RDMA WAN accelerator.

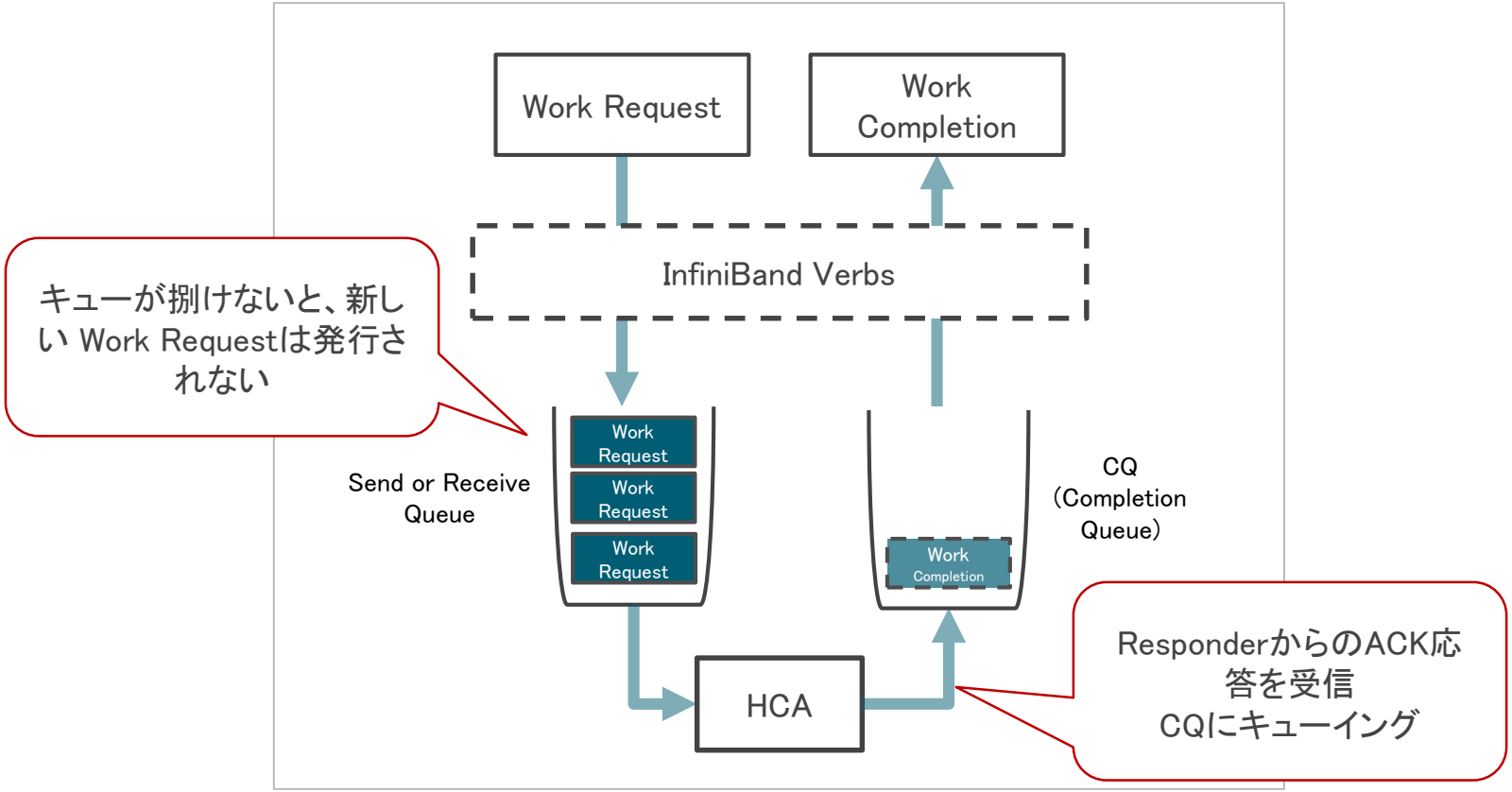
出典 : NTT Technical Review Vol.22, No.3, Mar.2024
https://ntt-review.jp/archive/ntttechnical.php?contents=ntr202403ra1.pdf&mode=show_pdf

RDMAトランスポート / ACK要求応答

- RDMA Write の場合
  - ACK要求(AckReq)のタイミングは実装によるが、Lastは必ずACK要求のbitが立つ
  - Write Last に対するACK応答が帰ってくるまで、処理は完了しない

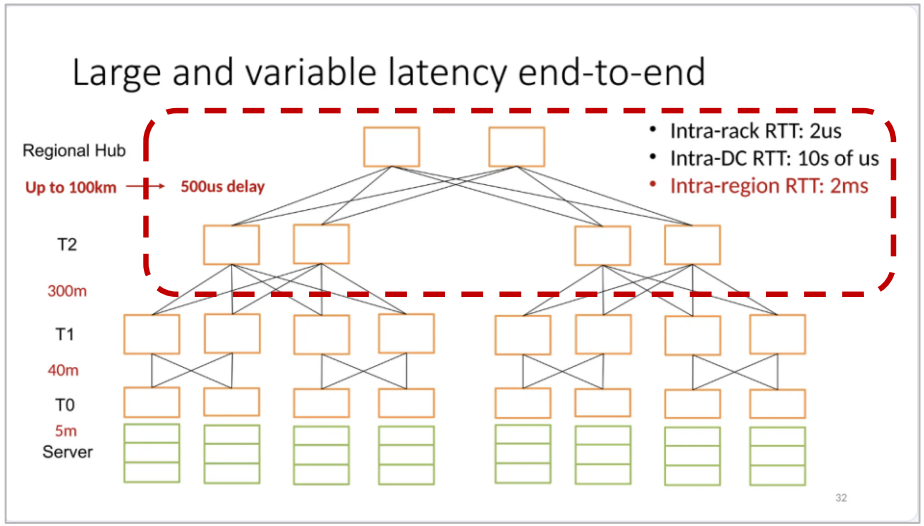


RDMA Work request / Work completion



先行研究・事例(2)

- “Empowering Azure Storage with RDMA” (Microsoft)
  - リージョン間(Up to 100km)のVM～ストレージクラスタ間でRDMAを実施



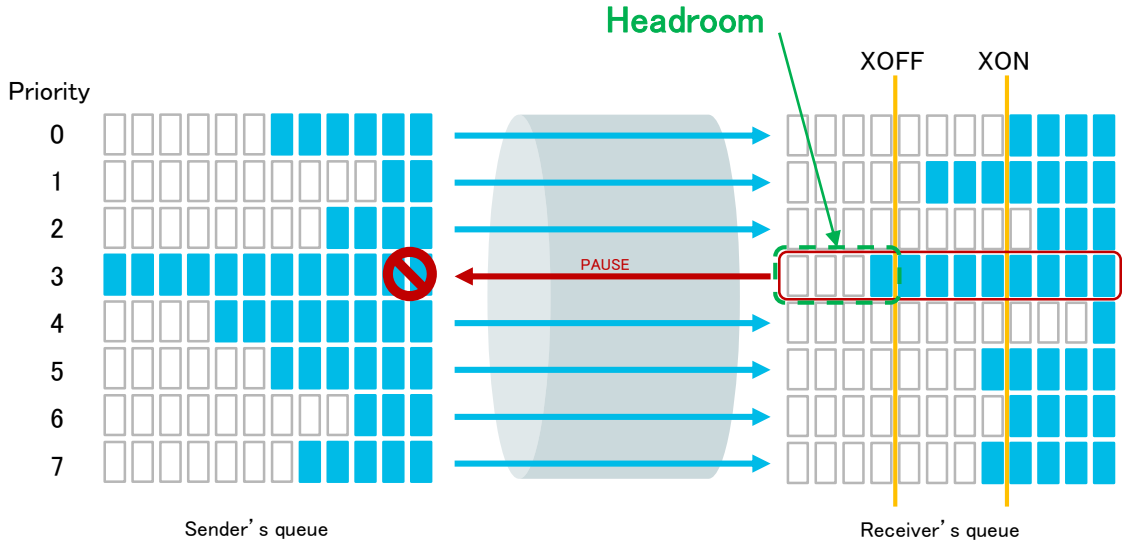
高集積かつ長距離リンクを収容するRHやT2には、数GBのPFC headroom bufferが必要

1. オフチップメモリに、
   RDMA用deep bufferを用意
2. Queue単位ではなく、
   全てのingress lossless queue で、
   PFC headroom pool を共有

出典：
https://www.microsoft.com/en-us/research/uploads/prod/2023/03/RDMA_Experience_Paper_TR-1.pdf
https://conferences.sigcomm.org/events/apnet2023/slides/apnet23/Industry/bai-1500-0630.pdf

PFC headroom

- PAUSEフレーム送信後、送信が停止されるまでタイムラグがある
- その間、in-frightパケット(ACK待ち)を吸収するbufferが必要
- そのbufferが、”Headroom”



Automatic Headroom calculation

- Headroom buffer size の最適化は複雑
  - 距離や帯域・流量によっても変化
- IEEE 802.1 Datacenter
  - PFCの拡張として、Automatic calculation の標準化が進む

AECCについて

- **Automotive Edge Computing Consortium**
  - https://aecc.org/

- トヨタが参加する目的
  - コネクティッド領域の研究開発を加速するには、**オープンイノベーションによる社外知見の取り込み**が必要
  - AECCで**企業間連携をリード**し、コネクティッド基盤のベストプラクティスを迅速に策定

AECC デモイベント in 東京（2024年4月）

デモの様子（**満員御礼!**）

トヨタからデモ**12件**を出展
（来場者の投票で2件が受賞）

授賞式の様子

デモ一覧

D1	Towards Safe Mobility	KDDI, Toyota Motor Corporation
D2	Cooperation for Automation	Toyota Motor North America
D3	Hierarchical Edge AI	Toyota Motor Corporation, Techno-Accel Networks Corporation
D4	Traffic Steering to Optimal Edge Servers	Toyota Motor Corporation, KDDI, Oracle
D5	Energy-Efficient Multi-Tier Edge Simulation	KDDI, Toyota Motor Corporation
D6	Multi-LLM Voice-Interactive AI System	Toyota Motor Corporation
D7	Green Connected Platform Field Trial	Toyota Motor Corporation, KDDI, Cisco
D8	Service-Oriented Vehicle Diagnostics (SOVD)	Toyota Motor Corporation, Denso Corporation, Vector Japan Co., LTD.
D9	Packet Counter in Network Access Device	KDDI
D10	Inter-Vehicle Edge Cloud over Wi-Fi Aware	Toyota Motor Corporation, Denso Corporation
D11	Vehicle Teleoperation in Immersive Digital Twin	Toyota Motor Corporation, Toyota Central R&D Labs
D12	Next Generation of 5G & API Enabled Cars	Ericsson, Toyota Motor Corporation
D13	Dynamic Slice Switching via Telco API	KDDI, Toyota Motor Corporation
D14	Robust Vehicle-to-Cloud Communication	Toyota Motor Corporation

安全支援
エッジ
クルマ通信
マルチパス通信

新メンバが続々と加入中！！

FURUKAWA ELECTRIC　SORACOM　HONDA The Power of Dreams　f5

まとめ(議論したいこと)

1. 専用ネットワークの最適化において、アプリケーションの通信特性分析としてどのような検証を行っているか。
2. ネットワークの設計にどの程度アプリケーション特性を考慮しているか。
   - インフラとしてとにかく高性能化をめざしている？
   - アプリ要件とコスト観点から、インフラ視点では妥協する場所もある？
3. AIワークロードのNW影響を理解するには可視化が大事。実行中のネットワーク負荷をどのように可視化しているか。(特に、マイクロバースト)
4. DC間(100km以上)で、RDMAを考えてますか？
5. 現在提案されているAI/ML基盤ネットワークアーキテクチャの中で、まだ最適化できるポイントはあるのか。