# A study on accelerating LLM inference using KV cache sharing with IOWN APN

NTT株式会社、デバイスイノベーションセンタ、田仲顕至

2025年 10月 30日

# 自己紹介：田仲 顕至（Kenji Tanaka）

- Communication-efficient distributed deep learning with GPU-FPGA heterogeneous computingCommunication-efficient distributed deep learning with GPU-FPGA heterogeneous computing（IEEE Hot Interconnects 2020）
- CiraaS: cloud computing with programmable logicCiraaS: cloud computing with programmable logic（ACM SIGCOMM’22 Poster）
- Transparent Broadband VPN Gateway: Achieving 0.39 Tbps per Tunnel with Bump-in-the-WireTransparent Broadband VPN Gateway: Achieving 0.39 Tbps per Tunnel with Bump-in-the-Wire（IEEE INFOCOMM’24 Oral）
- IOWN Global Forum 検討中の GPUDirect Async Kernel Initiated Network を用いた、低遅延・高効率ビデオ処理AI（GTC’24 → GTC’25 → OCP Global Summit’25 Poster）

- **KV Cache Sharing over IOWN All-Photonics Network: Building a Sustainable and High-Performance Nation-Wide Distributed AI for LLM Inference (OPC APAC Summit’25 Oral)**
- **システム全般、Smart NIC、NIC-GPU連携 に関する技術領域を中心に研究活動中**

# 背景：推論需要と電力網

- LLM 推論需要は年々拡大し、電力効率の改善が追いつかない。
- 増設のボトルネックは発電に加え、送電・変電にもある。
- 集中型DCのみでは需要の速度に追いつけない。

AMD, OCP Global Summit'25 Keynote より

（参考）千葉県印西市のデータセンター

印西市は東京都心から約 40km の場所に位置し、従来から金融機関のコンピュータセンターが立地してきた。近年、国内外の IT 企業などによるデータセンター建設が相次ぎ、電力契約の申し込みも大幅に増加した。

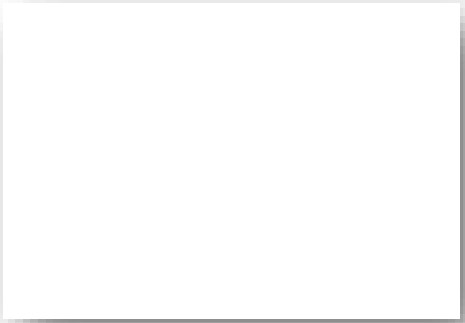
東京電力パワーグリッドは、印西地域への電力供給増強を計画し、50 万 V 送電線に接続する新京葉変電所（船橋市）から千葉ニュータウン変電所（印西市）を経て千葉印西変電所（印西市・新設）に至る 27.5 万 V の地下ケーブル線を敷設した。

地下ケーブル線を敷設するための約 10km の洞道は、シールド工法により 2 年半で建設するなど、早期の建設に努めた結果、2019 年 4 月から約 5 年間で工事を完工させることができた。千葉印西変電所の容量は 2024 年 6 月の運用開始時には 60 万 kW であり、将来的には更に増加させることを計画している。

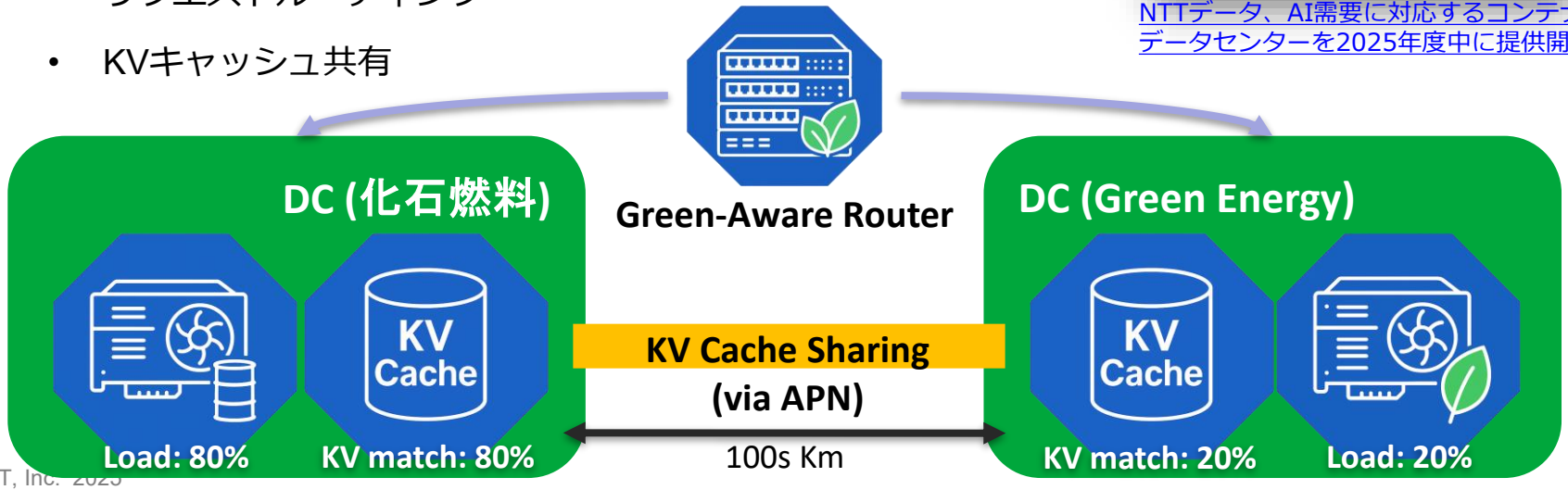
電力・ガス取引監視等委員会、局地的電力需要増加と送配電ネットワークに関する研究会報告書より

# ゴールイメージ

- 小規模DCを分散展開し電力網への負担を軽減
  - 再生可能電源近傍に展開することで CO2 排出量削減にも寄与
- IOWN All-Photonics Network で DC を束ね有効活用
  - リクエストルーティング
  - KVキャッシュ共有



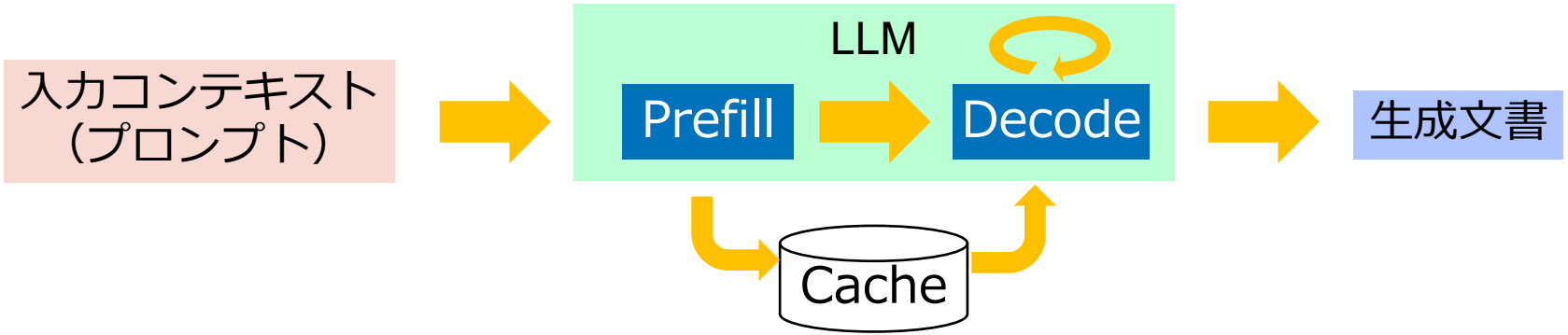
NTTデータ、AI需要に対応するコンテナ型データセンターを2025年度中に提供開始



# KVキャッシュの再利用・共有

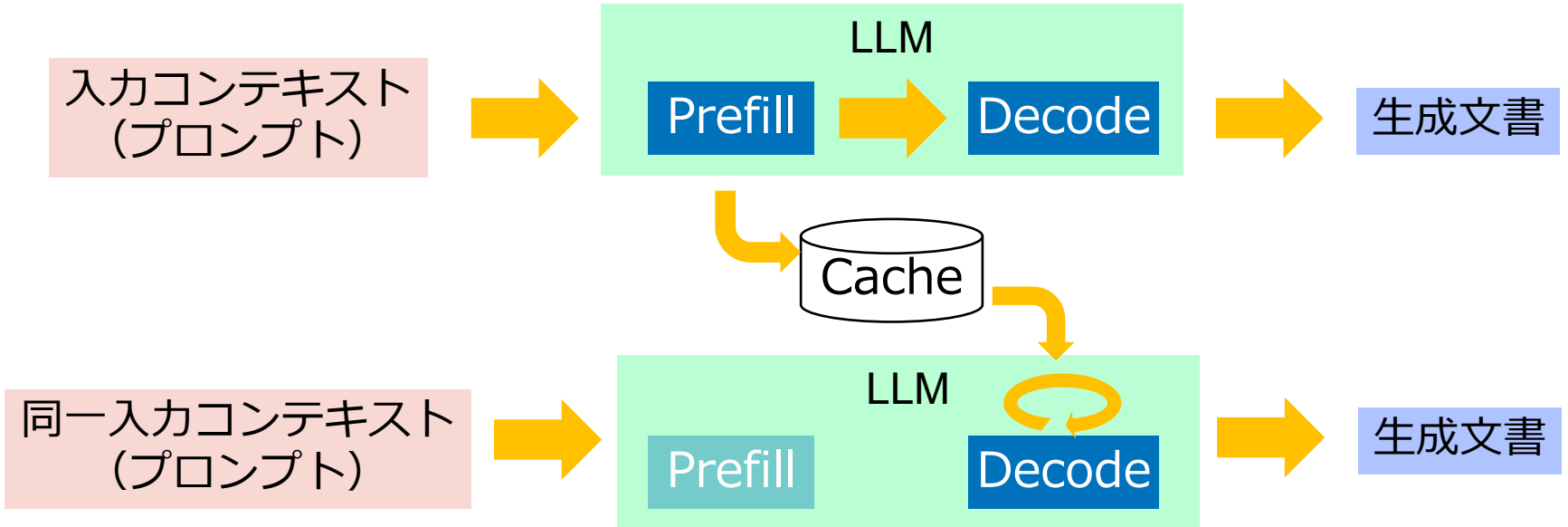
# KVキャッシュとは

- KV キャッシュとは、Prefill 段階で各層が生成した Key/Value の保存である。



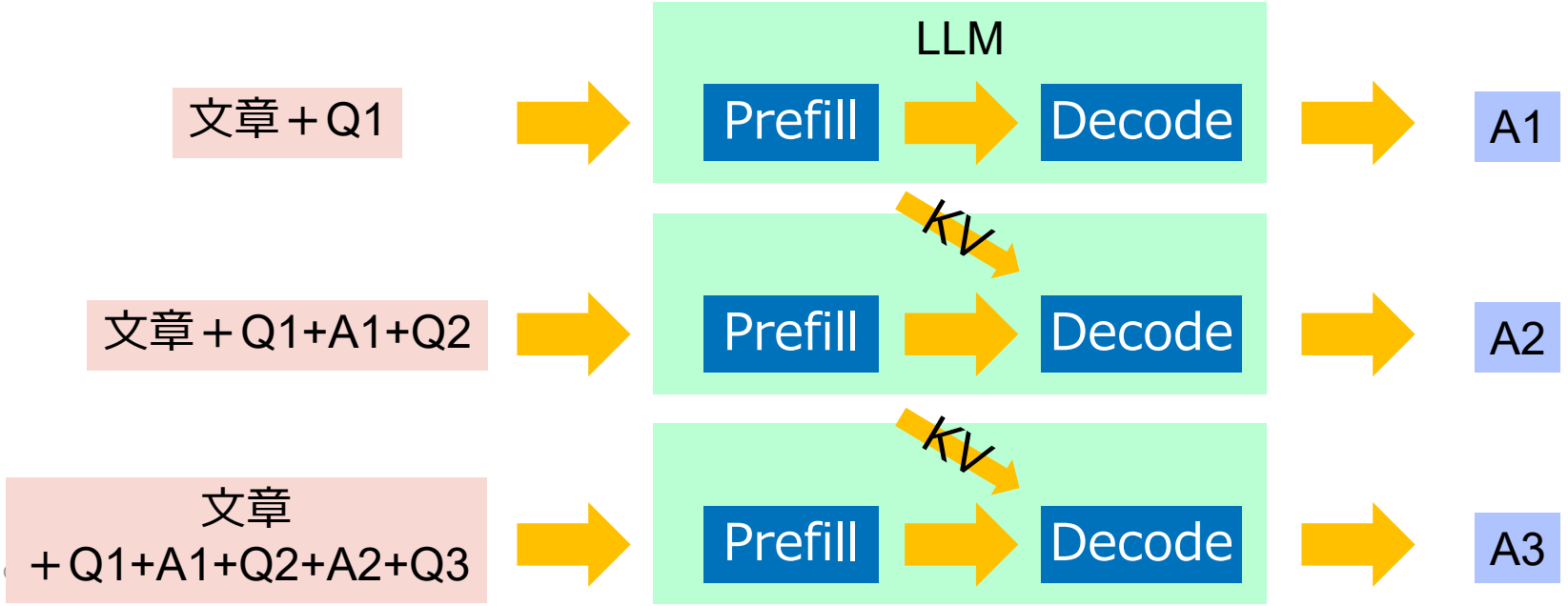
# KVキャッシュとは

- KV キャッシュとは、Prefill 段階で各層が生成した Key/Value の保存である。
- KV を再利用すると、Prefill の計算を大幅に削減できる。



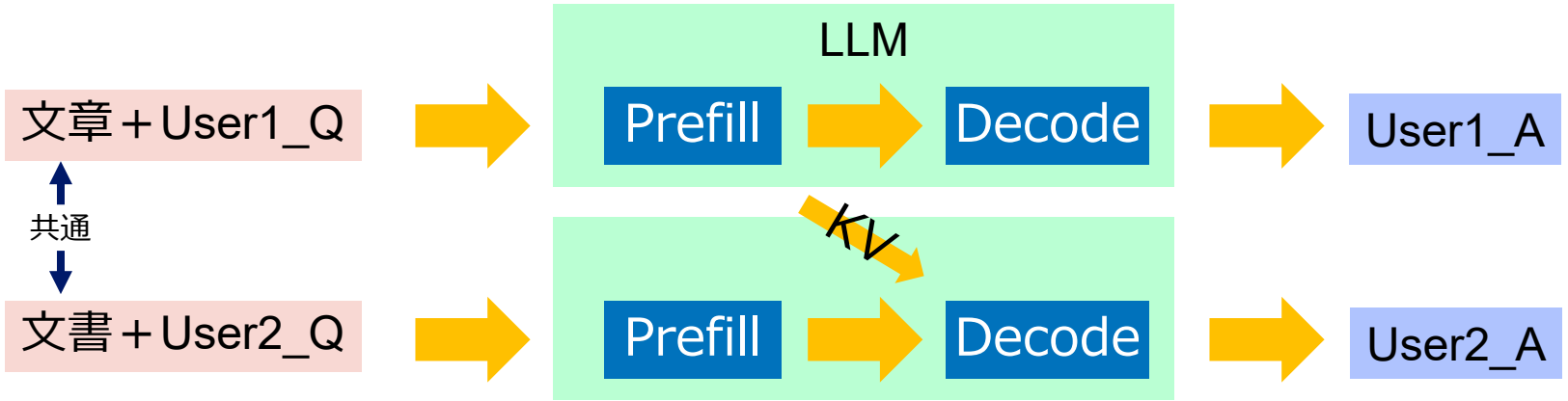
# KVキャッシュの再利用とは

- KV キャッシュとは、Prefill 段階で各層が生成した Key/Value の保存である。
- 「KV再利用＝同一ユーザ内」



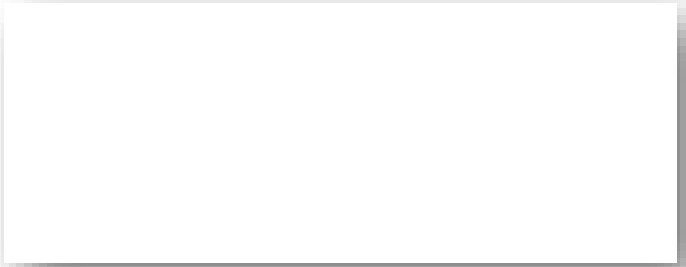
# KVキャッシュの共有とは

- KV キャッシュとは、Prefill 段階で各層が生成した Key/Value の保存である。
- 「KV共有＝ユーザ間」

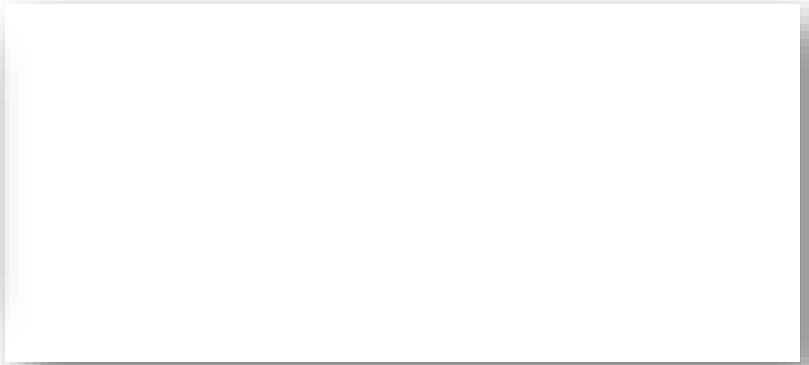


# KVキャッシュの再利用・共有の実体

- Deepseek、OpenAI API、Azure などでは KVキャッシュヒット割引を提供
- Alibaba Cloud の報告[1]によると
  - KVキャッシュ再利用率は、 良いケースで 6 割
  - **KVキャッシュ共有は、ほぼ実施されていない**



Deepseek料金表より



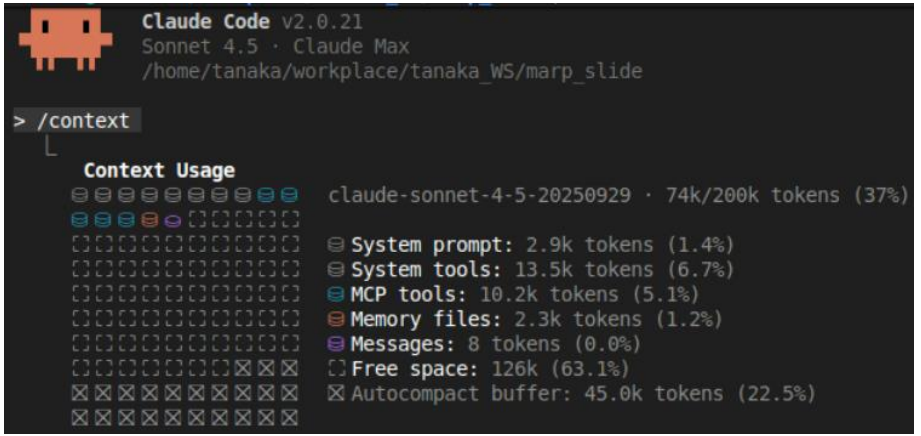
- なぜ KVキャッシュ共有は難しいのか？



[1] J. Wang+, “KVCache Cache in the Wild; Characterizing and Optimizing KVCache Cache at a Large Cloud Provider,” USENIX ATC’25.

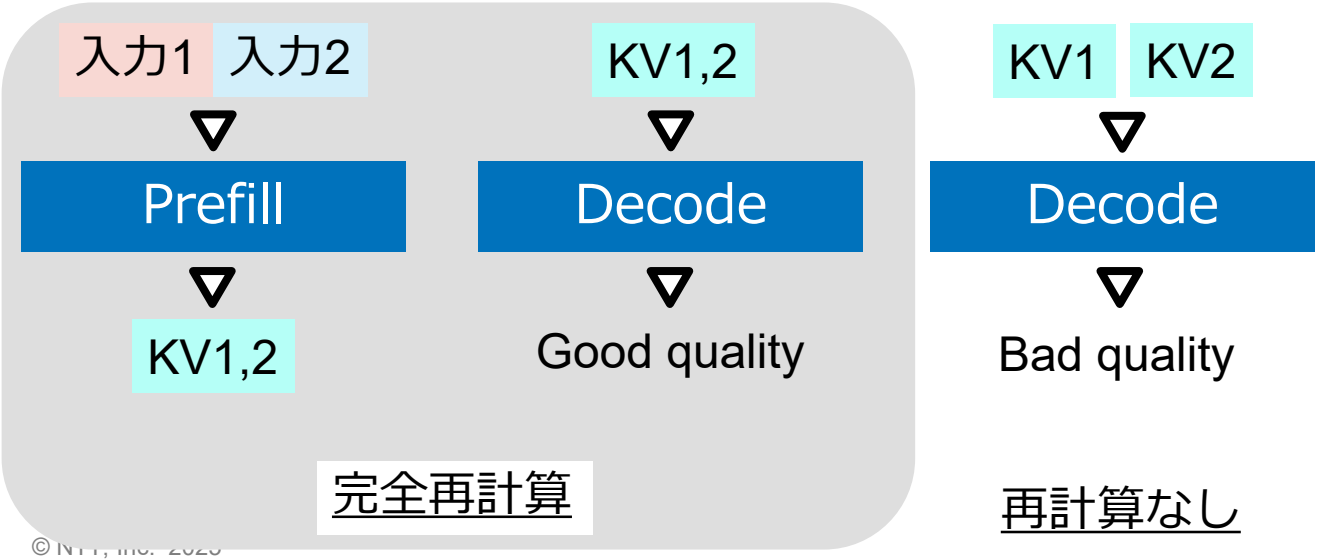
# KVキャッシュの前方一致

- 入力トークン列が ”前方一致” しなければKVキャッシュは再利用できない
  - Transformer を含む decoder タイプのLLMでは、ある位置のトークンのKVキャッシュは、その位置より前のトークンの情報を持つため
- 通常、 “ユーザ情報” や “メモリ” を含むシステムプロンプト（ガードレール、ツール仕様書）が先頭に入るため、ユーザ間の共有は難しくなる
- コンテキストを個別にキャッシュして機械的に結合すると、コンテキスト間の関係性が破壊され精度が落ちる。



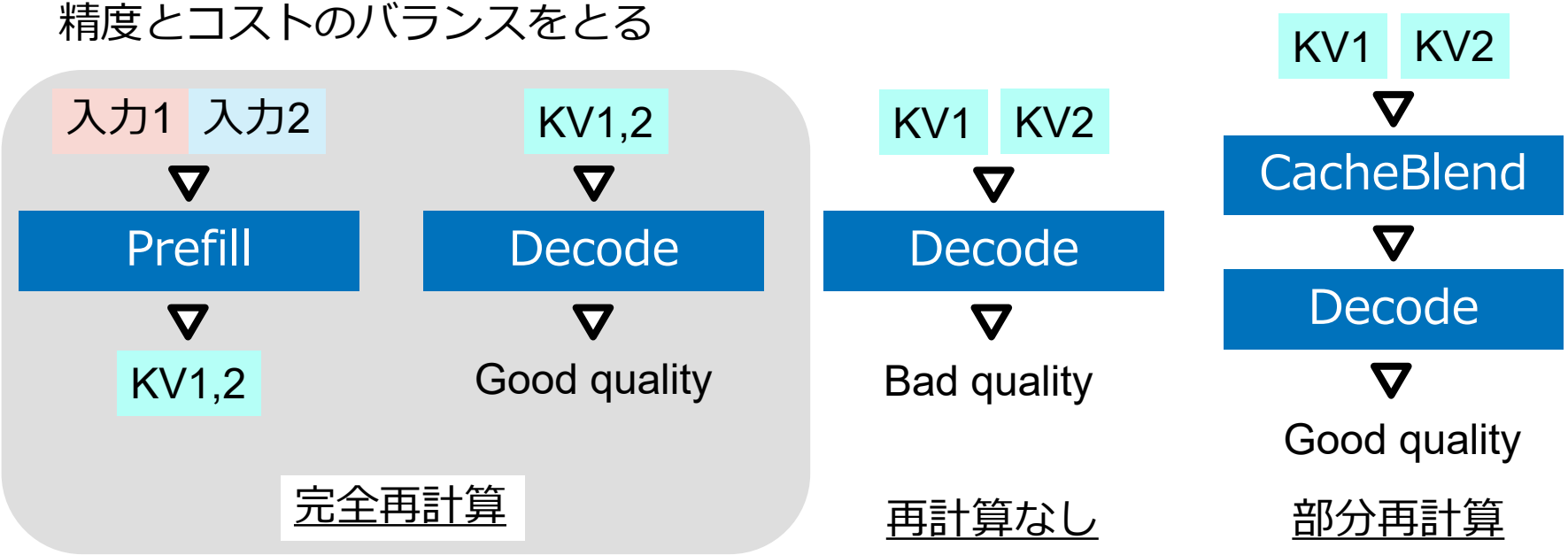
# KVキャッシュ間の依存の再計算

- KVキャッシュ間の関係性には再計算が必要。再計算にはコストがかかる。
- 個別のKVキャッシュを再計算なしに再利用した場合は精度劣化



# KVキャッシュ間の依存の再計算

CacheBlend では、KVキャッシュ間の関係性の強い部分のみを再計算することで、精度とコストのバランスをとる






J. Yao+, “CacheBlend: Fast Large Language Model Serving for RAG with Cached Knowledge Fusion,” EuroSys’25.

# KVキャッシュ間の依存の再計算

- CacheBlend [1] や、改良版の KVShare [2] を用いた場合、
  - 定形の入力（例：system＋文書＋質問）では文書の100%共有が可能 [1]
  - 会話ログのような不定形入力でも共有率 >50% が可能 [2]

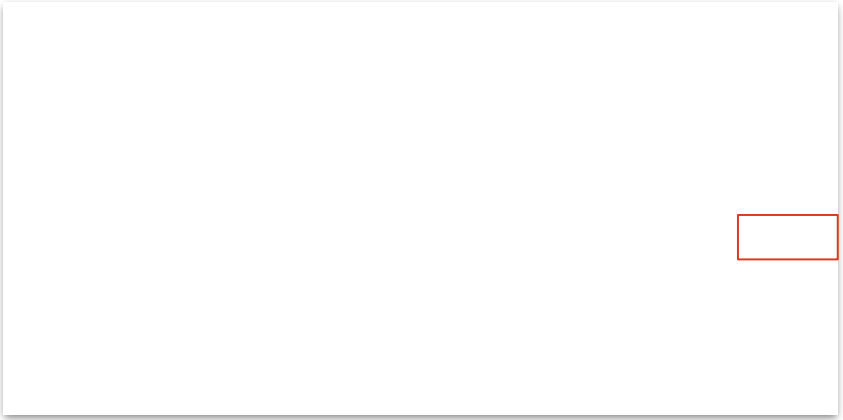
  } Prefill の大幅削減

- 再利用では精度劣化なし、共有でも劣化は小さい [3] } 精度も維持


[1] J. Yao+, “CacheBlend: Fast Large Language Model Serving for RAG with Cached Knowledge Fusion,” EuroSys’25.
[2] H. Yang+, “KVShare: An LLM Service System with Efficient and Effective Multi-Tenant KV Cache Reuse,” arXiv:2503.16525.
[3] Y. Li+, “SCBench: A KV Cache-Centric Analysis of Long-Context Methods,” arXiv:2412.10319.

# CacheBlend を使う

- CacheBlend 著者らは、CacheBlend を LMCache という名のOSSで公開中
- LMCache は
  - vLLM の KVキャッシュマネージャモジュールという位置づけ
  - Nvidia Dynamo においても KVキャッシュマネージャとしてサポート
  - llm-d にとってのコアの KVキャッシュマネージャ として貢献
  - 本日は解説しないが、以下の強力な機能も持つ
    - KVキャッシュの圧縮転送
    - 計算と通信のオーバラップ
    - NIXL との連携



LMCache & Nvidia, Scaling KV Caches for LLMs: How LMCache + NIXL Handle Network and Storage Heterogeneity,”PyTorch Conference’25.

# 現在の取り組み

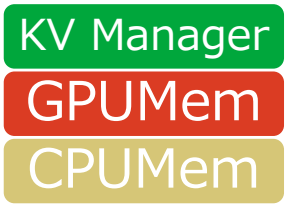
# CacheBlend 登場による要求されるシステムの変化

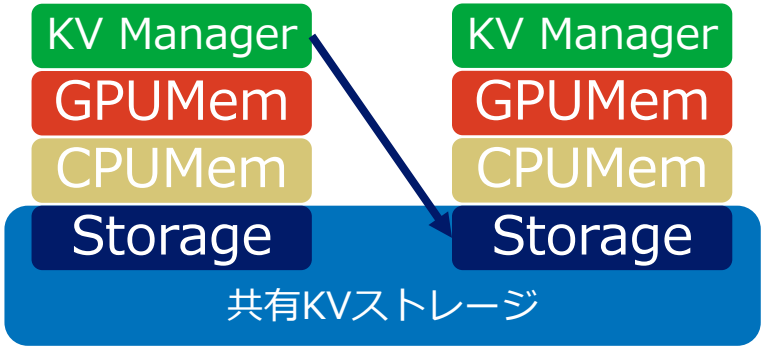
これまで：KVキャッシュは前方一致で、読むKVキャッシュは１つ

- メモリプーリングのような中規模の記憶空間で十分
- Hot（GPU Memory)、Warm（Host Memory）の記憶階層

CacheBlend 後：複数断片の再利用が可能、記憶容量の限界効用が高い

- KVキャッシュ母集団が広いほどヒット率が上がり、Prefill削減効果が上昇
- Storageを共有し、大容量な Cold （Remote Storage）なKVキャッシュストレージプールが有効？



# CacheBlend 登場による要求されるシステムの変化

これまで：KVキャッシュは前方一致で、読むKVキャッシュは１つ

- メモリプーリングのような中規模の記憶空間で十分
- Hot（GPU Memory)、Warm（Host Memory）の記憶階層

CacheBlend 後：複数断片の再利用が可能、記憶容量の限界効用が高い

- KVキャッシュ母集団が広いほどヒット率が上がり、Prefill削減効果が上昇
- Storageを共有し、大容量な Cold （Remote Storage）なKVキャッシュストレージプールが有効？

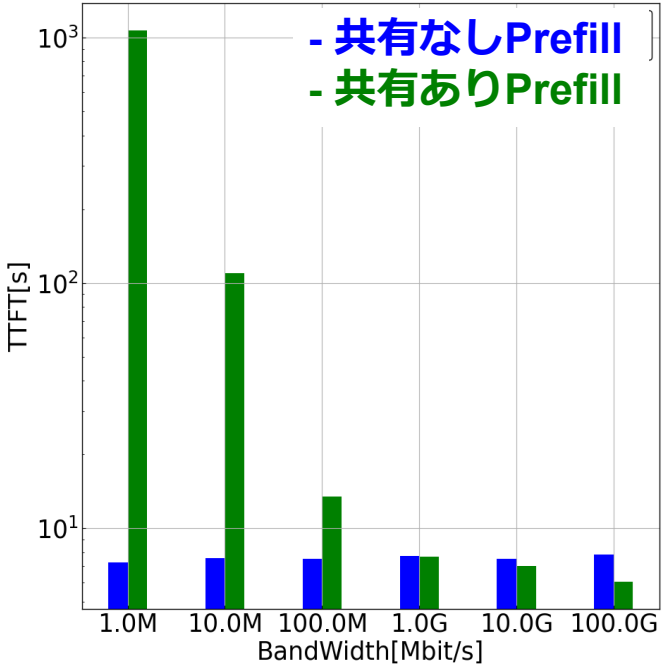
# 検証1：KVキャッシュリモートアクセス帯域の影響

- KVキャッシュは “入力トークン長” × “モデルサイズ” のため大容量
- アクセス帯域がパフォーマンスに与える影響を評価

モデル：Mistral-7B-Instruct-v0.2
入力トークンサイズ：1K
KVキャッシュサイズ：120 MB
GPU：A100

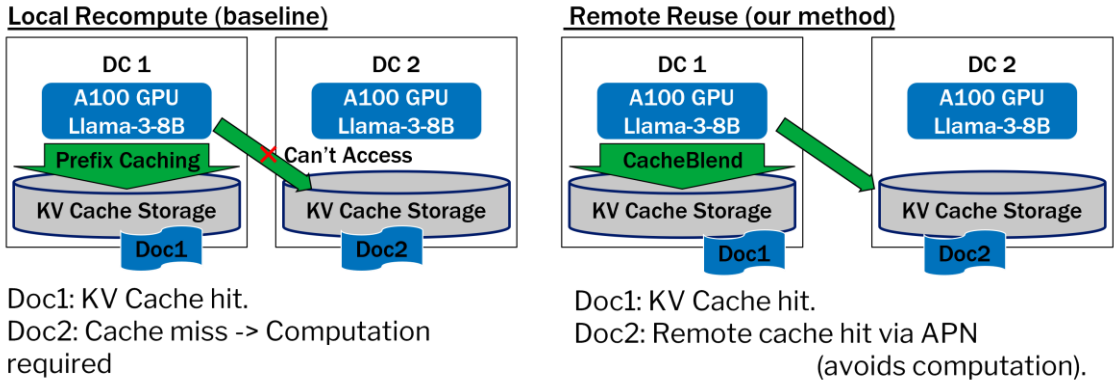
結果：

- 10G で約9割のTTFT 削減
- 100G で約8割の TTFT 削減



# 検証2：KVキャッシュリモートアクセス遅延の影響

- 前実験を踏まえ、長距離通信として APN をエミュレート
- レイテンシ（距離）がTTFT・電力効率に与える影響を評価
- 文書のKVキャッシュをユーザ間共有するシナリオ
  - モデル：Llama-3.1-8B
  - 入力トークンサイズ：平均約65K [1]
  - KVキャッシュサイズ：約8 GB
  - A100 x1 / node
  - 遅延挿入（tc コマンド）
  - レイテンシ – 距離特性：
    Open APN Report 参照 [2]

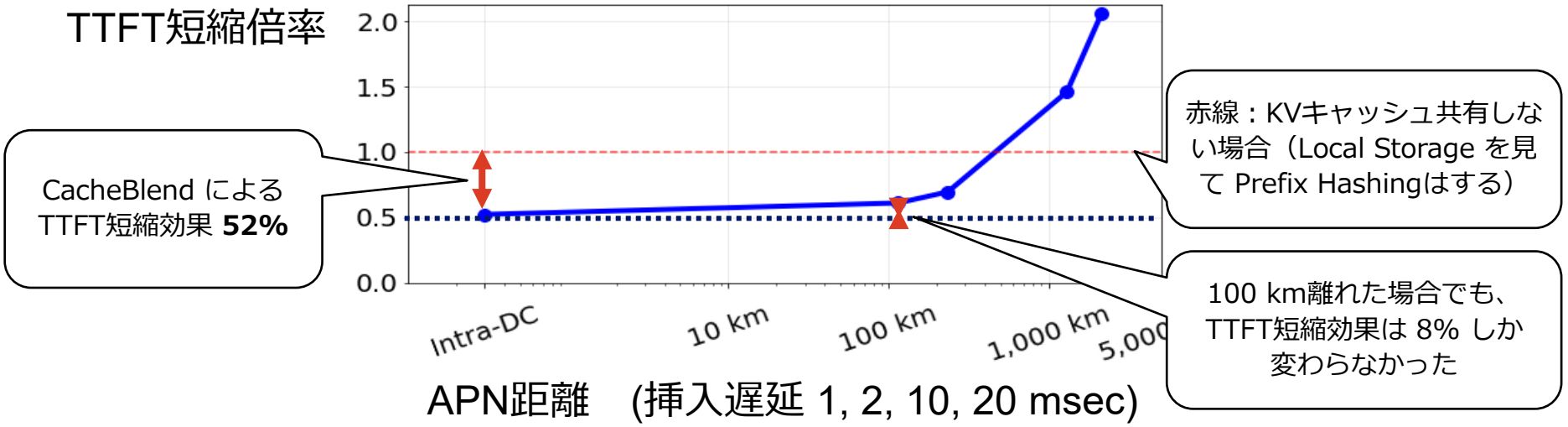





[1] M. Wang+, "Leave No Document Bdhind: Benchmarking Long-Context LLMs with Extended Multi-Doc QA," EMNLP'24.
[2] IOWN GF PoC Report Phase 1 -- Sensor Data Aggregation and Ingestion ver.1.1

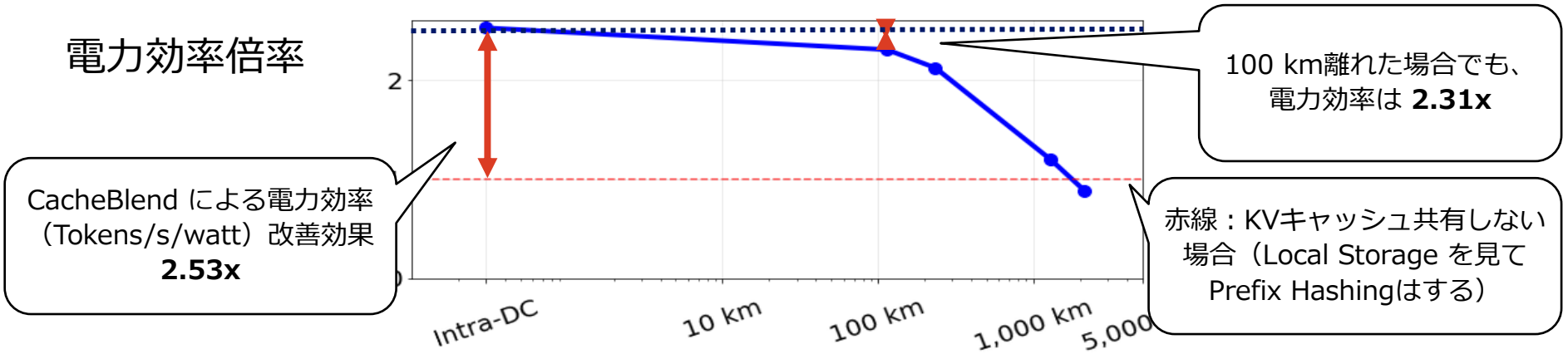
# 検証2：KVキャッシュリモートアクセス遅延の影響

- KVキャッシュを共有することで TTFT（Time-to-First-Token）は削減される
- ある程度（100km）距離を離しても、TTFT短縮効果は維持される



# 検証2：KVキャッシュリモートアクセス遅延の影響

- 電力効率（Tokens/s/watt）は、KVキャッシュ共有によるPrefill計算量削減で向上
- ある程度（100km）距離を離しても、電力効率メリットは維持される




APN距離　(挿入遅延 1, 2, 10, 20 msec)

# まとめ

- LLM推論の需要増、電力インフラへの負担増
- LLM推論の計算量を削減するKVキャッシュの再利用は大きく注目されている
- KVキャッシュの共有を可能にする CacheBlend に本発表では注目
- KVキャッシュ共有時代には、大きなリモートストレージでの
  KVキャッシュプールが必要と考え、リモートアクセス要件をエミュレート
- 低遅延・広帯域ネットワークにて、小型DCを繋ぎKVキャッシュ共有することで、
  100 km 圏内で DC 内比 8%オーバヘッドの推論性能を確認
- 電力インフラ負荷の低い小型分散DCを束ねた、LLM推論基盤の可能性を示した

# Innovating a Sustainable Future for People and Planet