mdx: アカデミックHPCクラウドmdxの紹介と今後の技術課題

空閑洋平 / Yohei Kuga
東京大学情報基盤センター
NII LLM研究開発センター(客員)
sora@ds.itc.u-tokyo.ac.jp

mdx

本日の内容

★HPCクラウド?

☆**クラウド基盤**(仮想マシン,仮想ネットワーク) + **RDMA**という意味なら、mdxはおそらくHPCクラウド

★mdxのインフラ紹介

★事例紹介: LLM-jpによるLLM事前学習クラスタ構築

☆仮想システム+GPUクラスタの話が多いので、ニッチな話題かもしれません

★HPCクラウド視点での課題と理想

☆mdxとあまり関係ないかもですが…

mdxとは

★9大学2研究所が共同運営し,

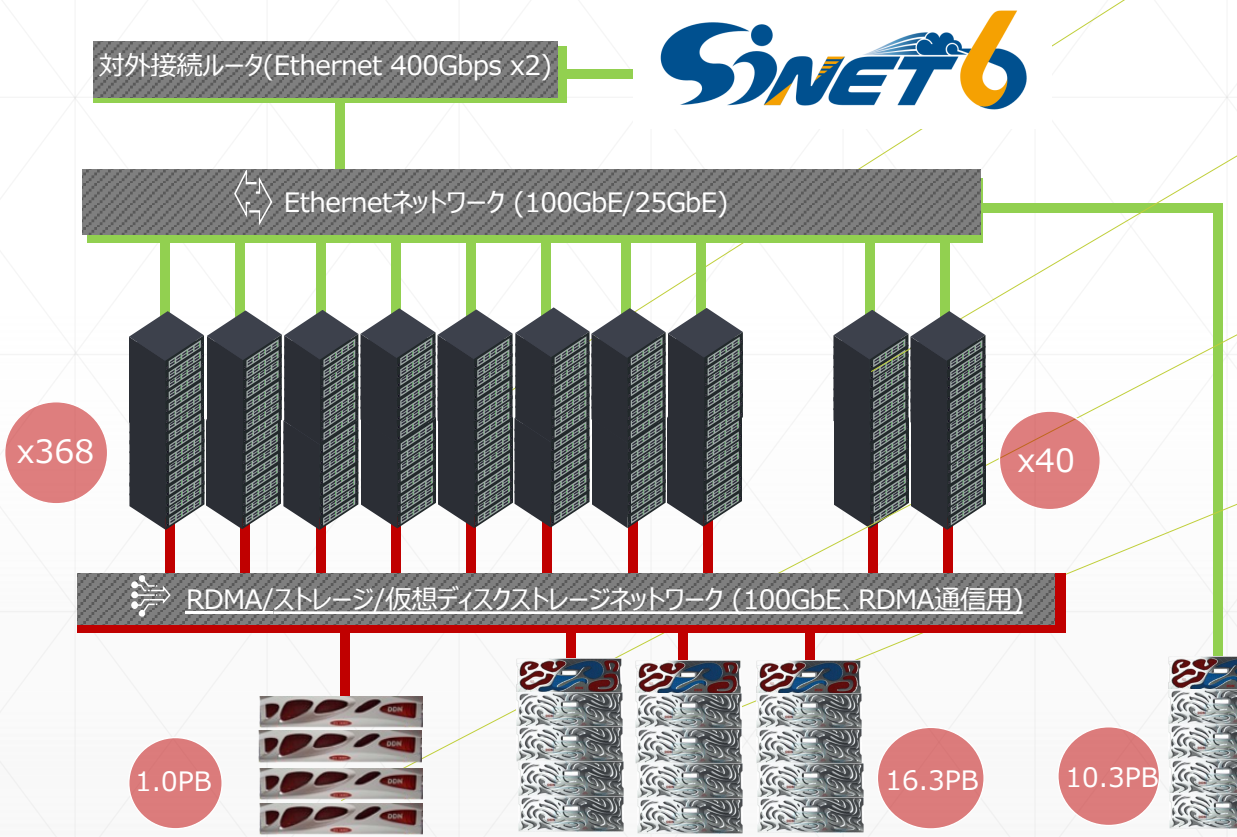
★全国共同利用に供する,

★データ科学・データ駆動科学・データ活用応用にフォーカスした

★高性能仮想化環境 https://mdx.jp/

★@ 東京大学柏IIキャンパス

mdxハードウェアスペック



汎用CPUノード
Ice Lake （76仮想コア）x 2 x 368ノード
総理論演算性能(倍精度)：2.1PFLOPS
総メモリバンド幅：150.7TByte/秒

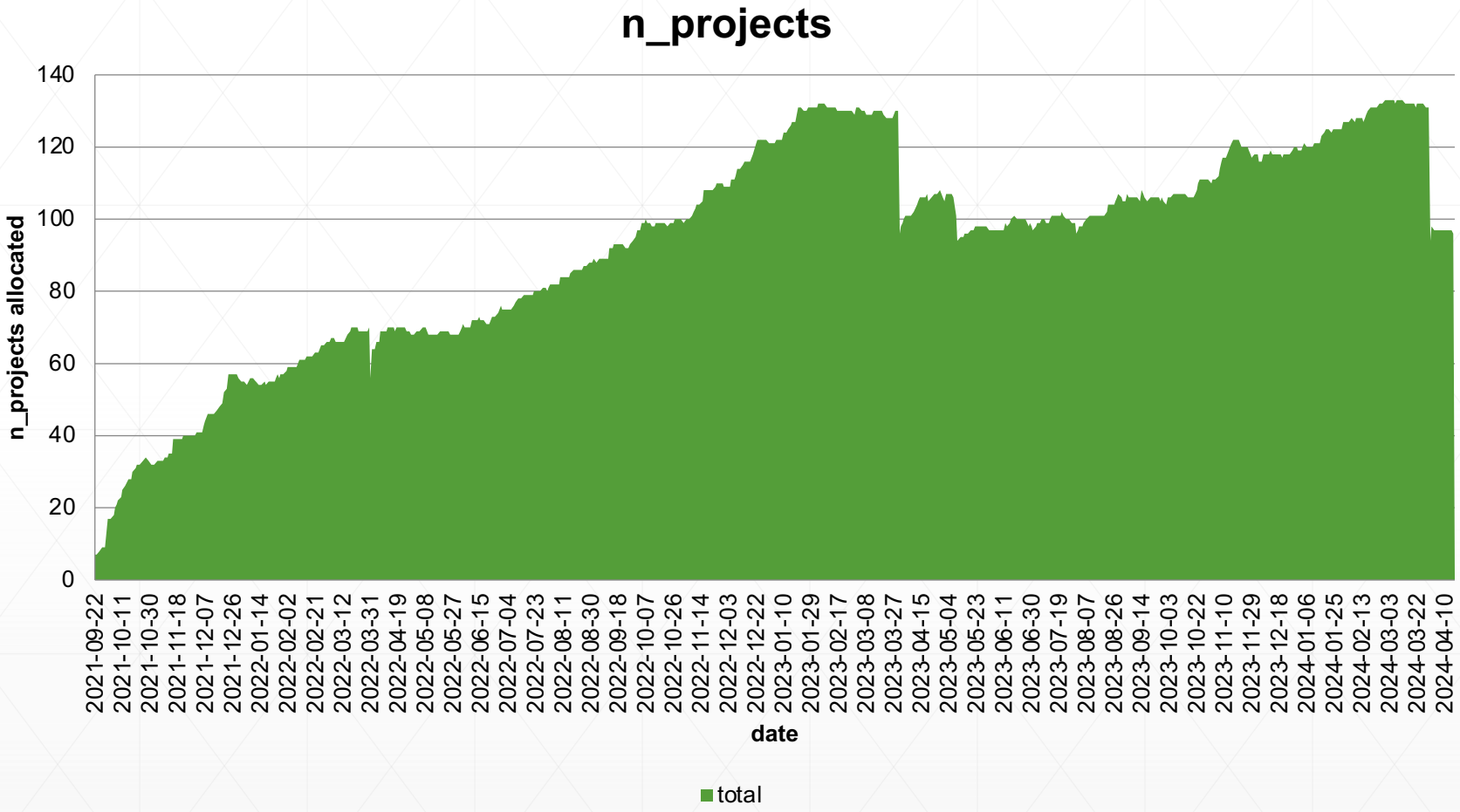
演算加速GPUノード
NVIDIA **A100** x **8 x 40**ノード
総理論演算性能(倍精度)：6.4PFLOPS
総理論演算性能(半精度)：100.7PFLOPS
総メモリバンド幅：496.3TByte/秒

高速NVMeストレージ
Lustreファイルシステム（マルチテナント）
1.0PB (NVMe SSD)
252GByte/秒

大容量HDDストレージ
Lustre並列ファイルシステム（マルチテナント）
16.3PB
157.5GByte/秒

外部共有オブジェクトストレージ
S3 Data Service
10.3PB
63.0 GByte/秒

mdxプロジェクト(テナント)数

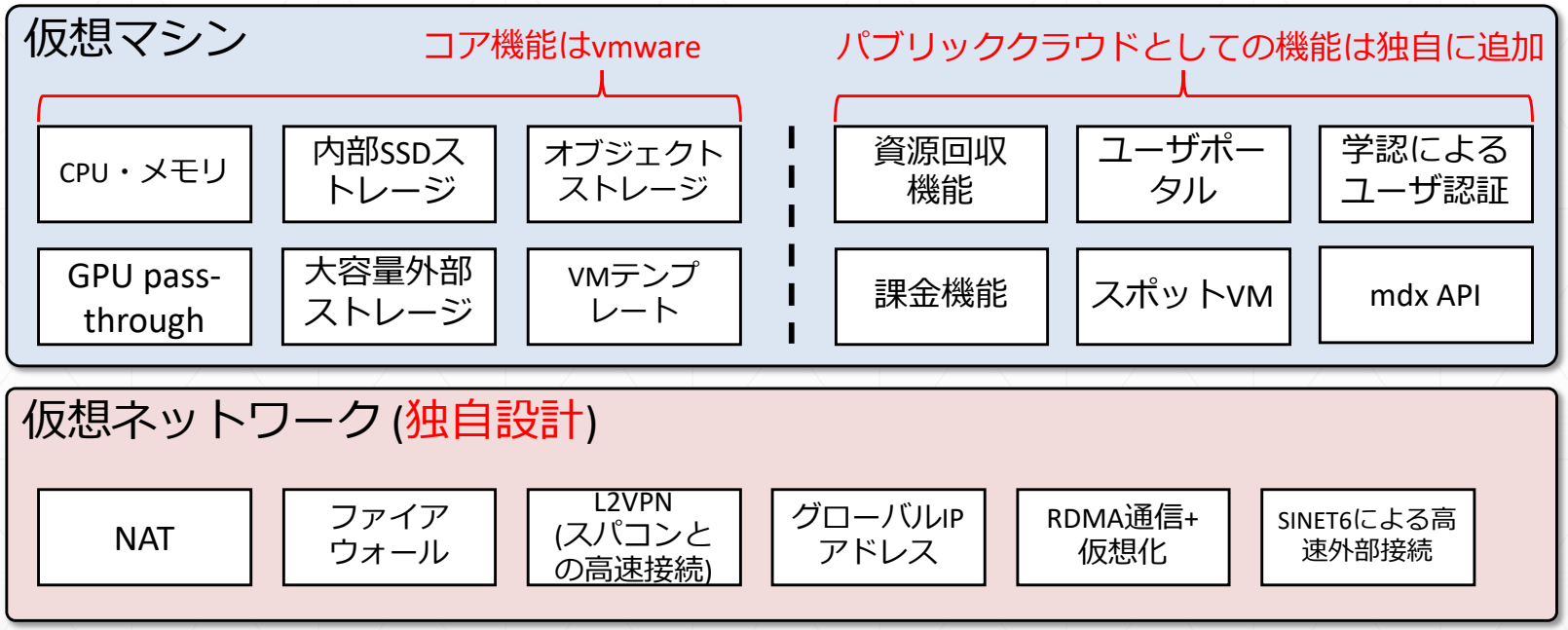


mdxソフトウェアスペック

★仮想マシンと仮想ネットワークを導入

☆インフラをプログラマブルに操作・管理

☆1テナント = 1研究プロジェクトで管理し、テナント間で環境を分離



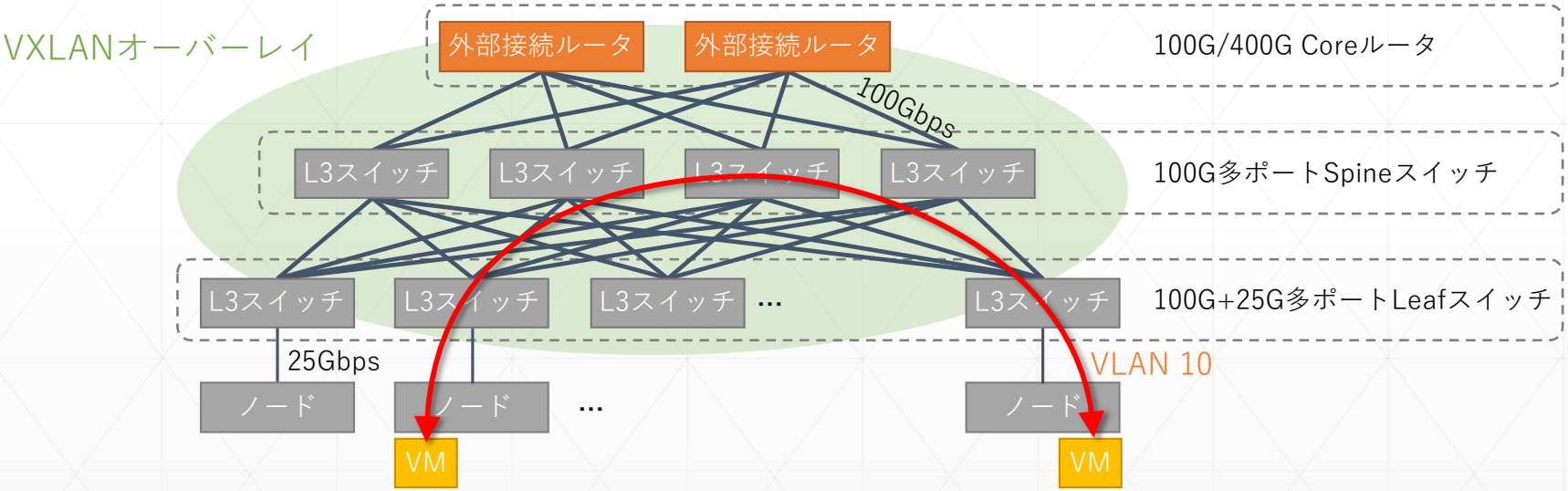
外部接続ネットワーク

★Virtual eXtensible LAN (VXLAN)によるオーバーレイネットワーク

☆データセンター向けの仮想ネットワーク技術

☆物理はSpine/LeafのIP Clos Fabric ネットワーク

☆IPネットワーク上にVXLANでVLANを通す(Ethernet over IP)



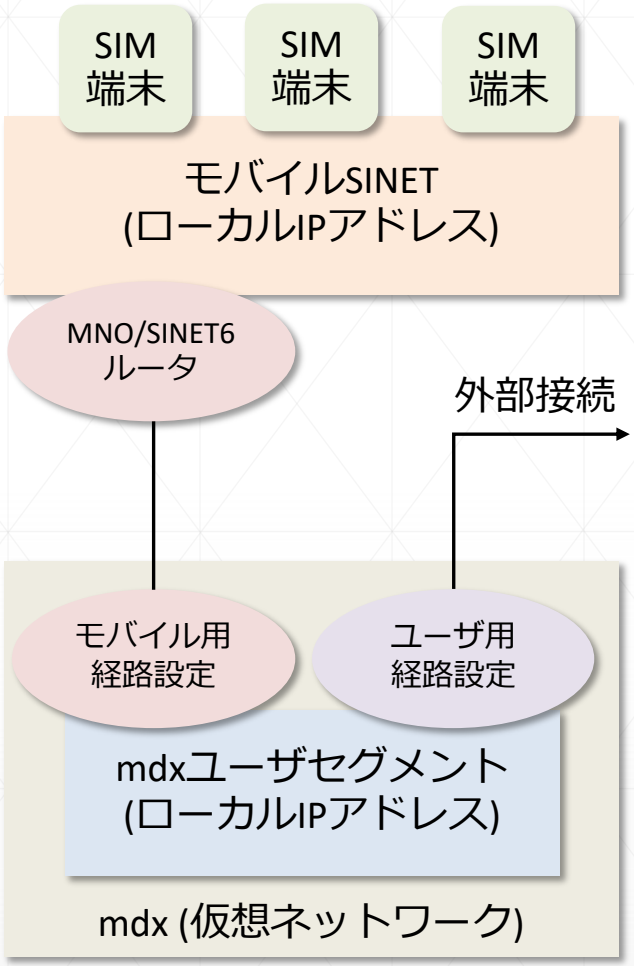
(外部接続ネットワーク・ユースケース)
IoT・センサーデータの蓄積・解析基盤の構築

★NII SINETモバイルとの連携

★mdxテナント単位でL3経路の設定が可能

☆mdxコアルータでユーザセグメントごとのL3VRFを追加作成

★その他、テナント単位でNII L2VPNやパブリッククラウドとのVPN接続にも対応



内部高速ネットワーク

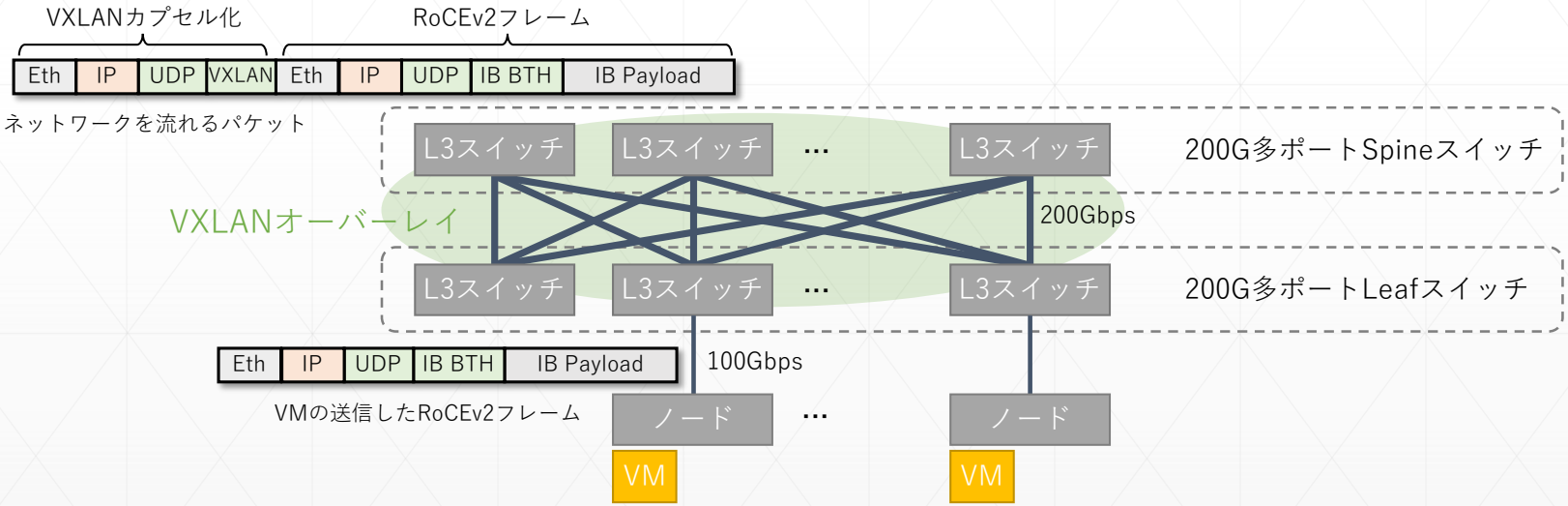
★RDMA over Converged Ethernet (RoCE) over VXLAN

☆RoCE: RDMAをIPネットワーク越しに行う技術

☆外部接続ネットワークと同様、IP ClosトポロジにVXLANを利用

Infinibandでは十分なテナント間の分離を実現できなかった

☆各ノードのNICは100Gbpsで、ネットワークはフルバイセクション



(内部高速ネットワーク・ユースケース)
LLM-jpによるLLM事前学習クラスタの構築

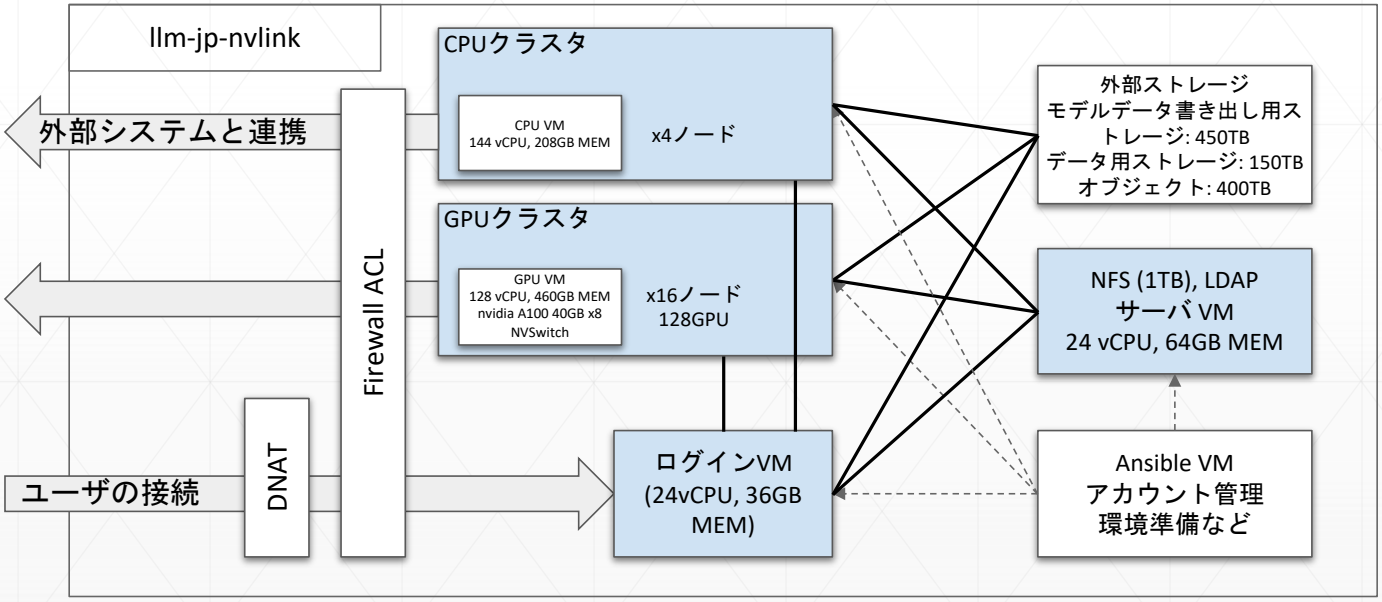
★2023/6) LLM-jpがmdx上で事前学習クラスタを構築開始

☆Nvidia A100 GPU (40GB) 128台を利用、GPU間はRDMAで通信

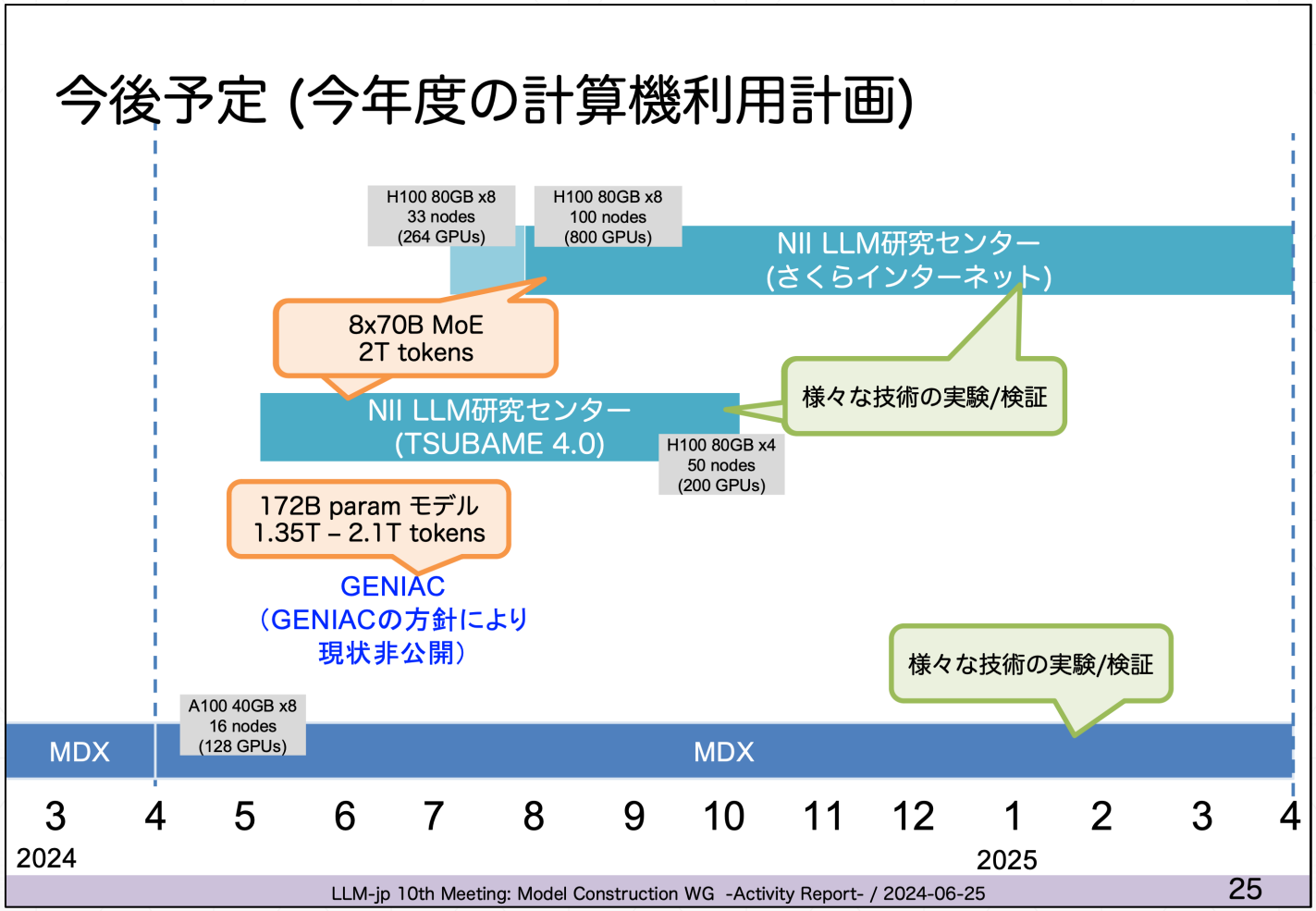
★2023/10) 事前学習、チューニングしたLLM-jp-13Bモデルを公開

☆https://www.nii.ac.jp/news/release/2023/1020.html

★以降は、小規模実験(スポットVM併用)、コーパスデータの収集などに利用



LLM-jpでの今後の計算機利用計画



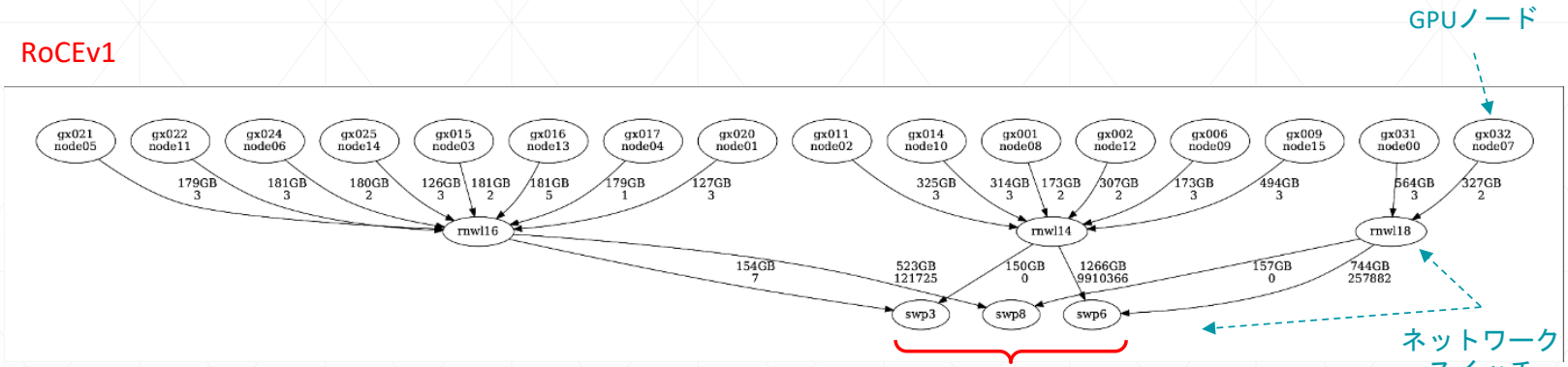
https://drive.google.com/file/d/1iL_cHpyIfqx4aKbs_xCNO8eSUnZ51Uq-/view?usp=drive_link

mdxで発生した LLMクラスタの構築時の様々なトラブル

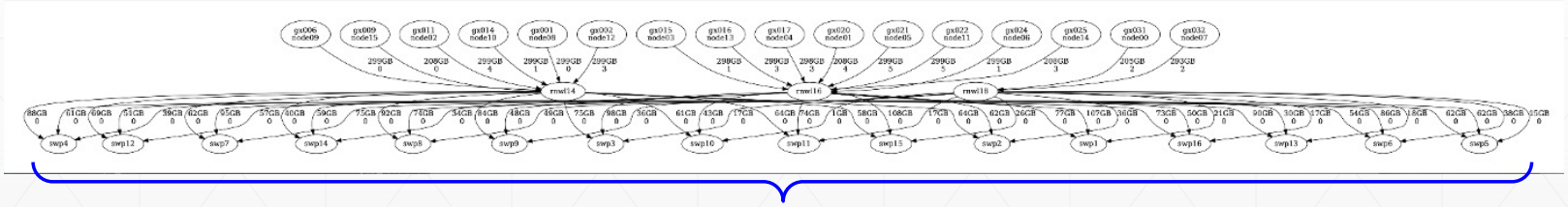
★RoCEの運用経験不足でクラスタ性能がボトルネックに

★仮想マシンと今どきのGPU機能の相性が大変悪い

RoCEv1



RoCEv2



仮想マシンとGPUクラスタの相性が悪い

★GPUクラスタのようなヘテロ構成が得意ではない

- ☆IOがPassthrough接続だらけでVM運用の良さが失われる
- ☆VMの仮想PCIeトポロジと物理トポロジの不一致
- ☆仮想マシンでの扱いが難しい技術: GPUDirect, Intel PCM (PCIeリンク等の監視), VMへのNVSwitchの割り当て,など
- ☆最近のGPUマシンはBIOSで仮想化支援機能が有効化できないもある

★ワーカラウンドな対応

- ☆物理マシン＝1VMで運用、NCCLに物理PCIeトポロジを設定, など

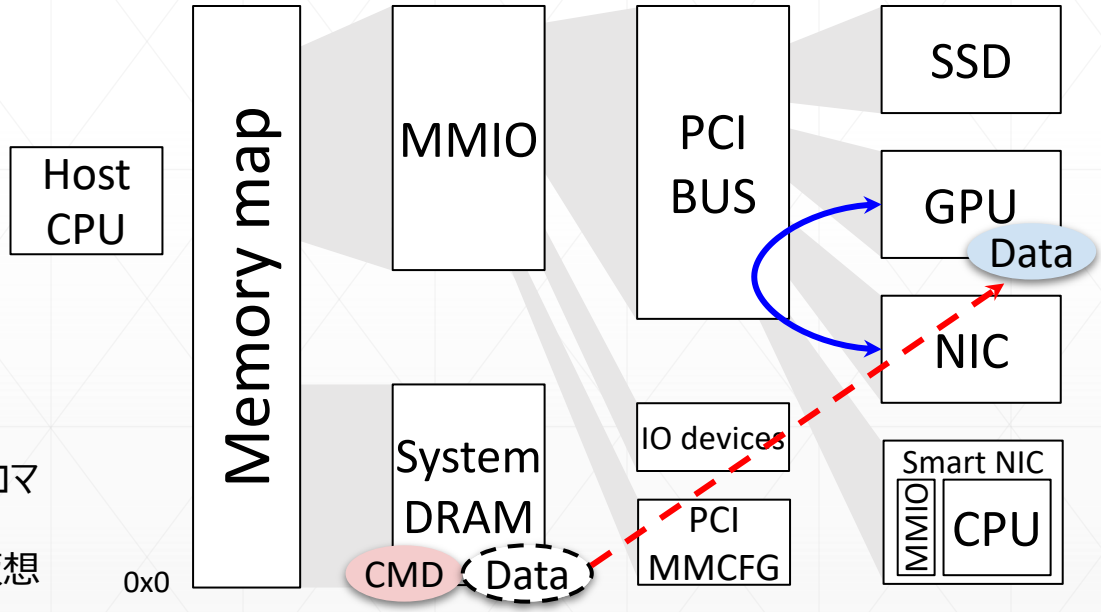
仮想化システムでのRDMA対応

★Linux Device Memory TCP (Google GPUDirect-TCPX)

☆GPU-RDMA NIC連携事例。Google Cloud A3インスタンスで利用

☆NICのHeader Split機能を使用して、NICメタデータ(Descriptor ring)はホストメモリのまま、受信パケットペイロードをGPUメモリに直接DMA

```
enp0s12       tx 52.56 Mbps 78.43 Kpps
enp0s12       rx 107.10 Mbps 82.03 Kpps
enp12s0       tx 80.38 Gbps 312.76 Kpps
enp12s0       rx 78.51 Gbps 1.32 Mpps
enp134s0      tx 79.89 Gbps 339.59 Kpps
enp134s0      rx 77.96 Gbps 1.31 Mpps
enp140s0      tx 80.81 Gbps 346.28 Kpps
enp140s0      rx 78.91 Gbps 1.35 Mpps
enp6s0        tx 78.58 Gbps 324.67 Kpps
enp6s0        rx 76.66 Gbps 1.28 Mpp
```



- メタデータはホストメモリにあるため、既存のLinuxコマンドでGPU間トラフィックが観測可能
- Lossy EthernetでRDMAできるため、既存の仮想システムの資源が利用可能

CMD: パケットヘッダ+メタデータ
Data: パケットペイロード

そのほかのLLM構築基盤でよくあるインフラ側の課題をざっくばらんに…

★具体的な数値は今後LLM-jpから報告

★GPUメモリサイズやモデルのパラメータ数が増え、データ、ファイルの取り扱いが隅々まで大変になることを実感

☆例) LLM事前学習中のチェックポイントファイルのサイズが大きくなり、ストレージやネットワーク性能の要求、費用が跳ね上がる

☆PCIe, RDMA通信のクレジットベースの輻輳制御に起因して、ノード内・ホスト間データリンク経路でバーストなトラフィックが発生

★GPUDirect RDMAのようなCPUをバイパスするデータ転送が多用され、CPU側からの監視やパフォーマンス解析が困難

☆構成にもよるが、メモリ帯域やPCIeリンク帯域に十分余裕がある想定が難しくなってきているため、様々なデータ転送経路の監視が必要

(個人的に) どのようなHPCクラウドが理想か ＝＞ ヘテロジニアス環境への対応

★データセンタでは今後も様々な新デバイスが登場予定

☆CXLメモリ, SmartNIC, Accerator等

★クラウド技術がヘテロ構成に対応するためには、CPU・GPU・NIC・ストレージ等さまざまなデバイス連携を自由にオーケストレーションできることが望ましい

☆現状は、GPUDirect RDMA, GPUDirect Storageのように個別ごとに機能が実装

☆NICなどのデバイスが、データごとにDMA先を制御できるようになるとデバイス連携の幅が広がる。組み合わせ方は無数にある

☆理想は、デバイスのメモリ領域についてもOSのメモリマネージャで管理し、ホストメモリと同等のレベルでmallocできるようになると、利用ユースケースが広がる?

(個人的に) どのようなHPCクラウドが理想か ＝＞ ヘテロジニアス環境への対応

★具体例: [ASPLOS23 BaM]

★GPU-SSD連携事例

★NVMeコマンドのSQ/CQをホストメモリからGPUメモリに移動し、NVMeコマンド発行をGPUで実施

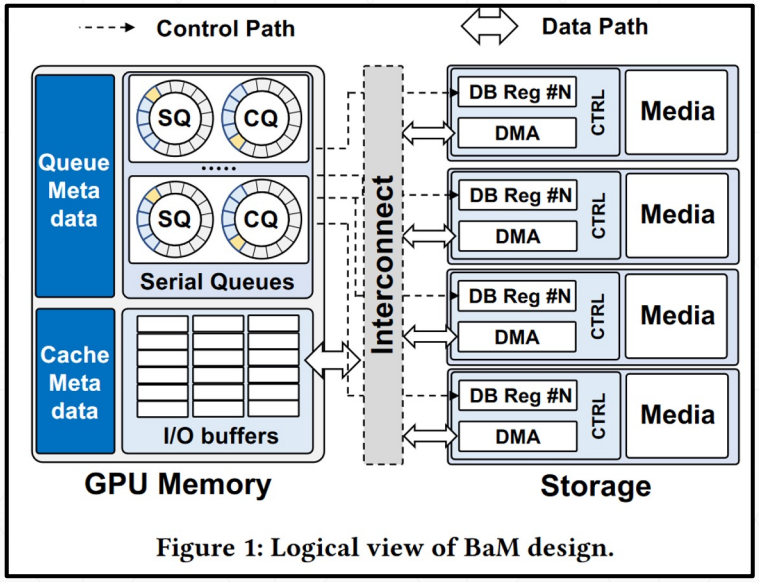
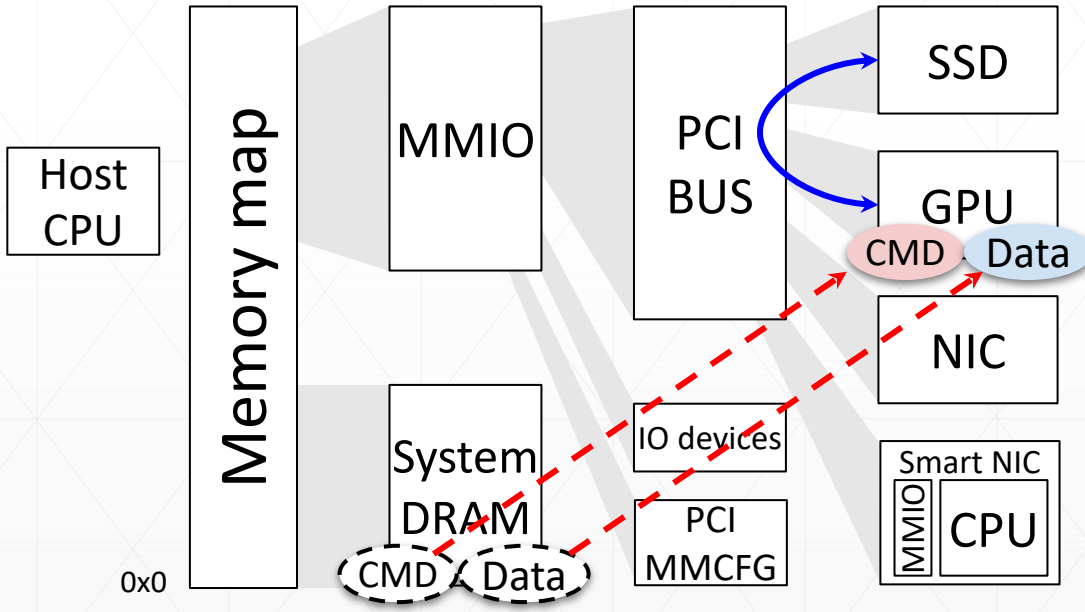


Figure 1: Logical view of BaM design.



まとめ

★mdxは、大学、研究機関、企業の方による｛学術研究、教育、技術開発、社会実装｝などに利用できます！

☆https://mdx.jp

★LLM構築基盤では、GPUメモリサイズやモデルのパラメータ数の増加に伴って、ノード内・ノード間様々なデータ転送経路の性能確保、監視が重要

★今後登場するであろう様々なデータセンタデバイスの使用したヘテロ構成への対応はまだまだ課題が多そう

☆デバイス間連携のデザインパターンが見えてきている段階

☆(1) ハードウェアのDMA先の制御機能 (2)OSメモリマネージャによるハードウェアメモリの払い出し (3)アプリケーションレベルのデバイス連携のオーケストレーション対応、などが理想?