

mdx: データ活用社会に向けた 産学官連携のための 共創プラットフォーム

東京大学 情報基盤センター
埴 敏博

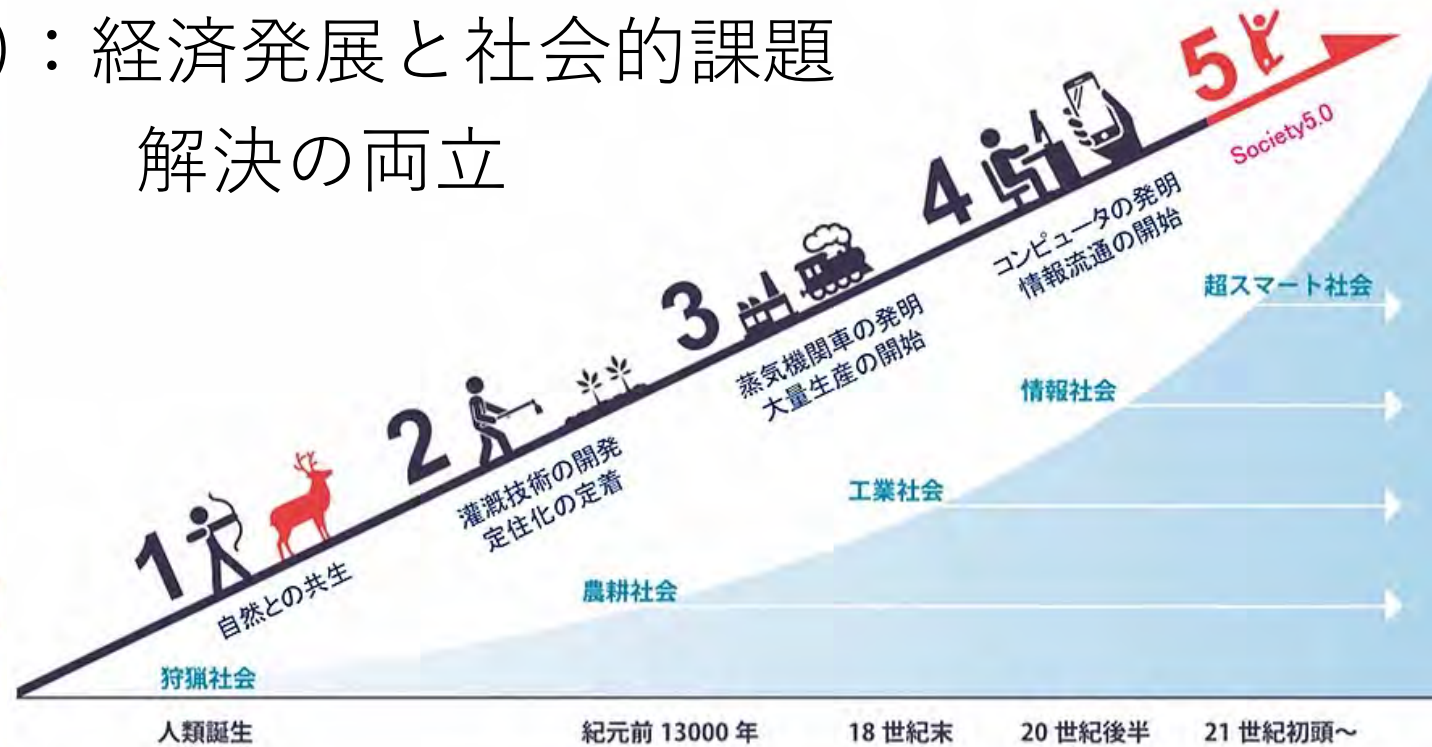
Society 5.0：未来社会のコンセプト

- 超スマート社会：AI, IoT, ビッグデータなどの革新技术を社会全体に活用し、サイバー空間（仮想）とフィジカル空間（現実）を高度に融合させた社会
- 持続可能な開発目標(SDGs)：経済発展と社会的課題

解決の両立



2021/6/18



PCCC WS in 柏2021

経団連HP <http://www.keidanrensds.com>

Society 5.0が実現する知識集約社会



データの収集・集約

高度な処理
高速な計算基盤

知の抽出・
データの活用

データ活用社会創成プラットフォーム

「データ活用社会創成プラットフォーム」は
用途に応じてオンデマンドで**短時間に構築・拡張・融合**できる
データ収集・集積・解析機能を提供するプラットホーム。

研究所 (2)

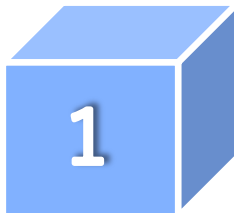
・ 国立情報学研究所、産業技術総合研究所

大学 (9)

・ 北海道大、東北大、筑波大、東京大、
東京工業大、名古屋大、京都大、大阪大、
九州大

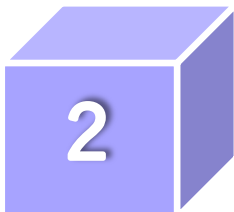


データ活用社会創成プラットフォーム 3本柱



SINETを活かしたリアルタイム収集・集積・解析環境の動的な構築

遠隔地のセンサーやストレージ、データプラットフォームの計算資源、ストレージをつないで、リアルタイムに
入力から出力を得られるアプリケーションごとの収集・集積・解析環境（仮想データプラットフォーム：仮想DP）を、使いたいときに即時に構築する
SINETモバイル基盤によりセンサー等のデータを安定してセキュアにつなぐ



高性能計算環境によるデータ科学と計算科学の融合

データ科学、計算科学の手法を融合し、さらに国内最高の計算環境を用いて他に無い
高精度の予測を行えるようにする

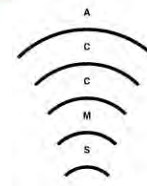


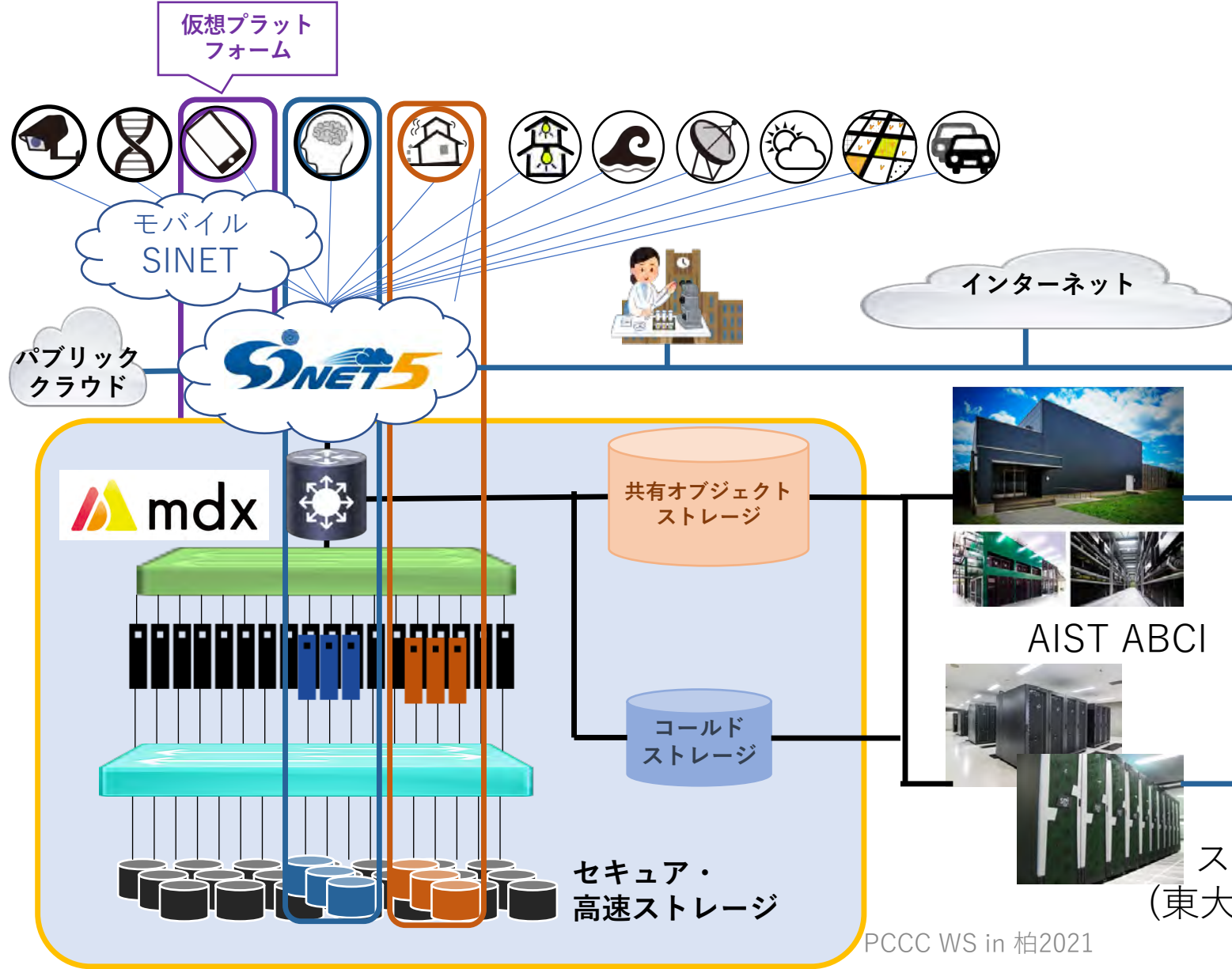
異種データ・異種知識の融合活用の推進と利用者支援

様々な分野のデータ保持者、解析者、利用者が産学にまたがって連携するコミュニティを形成し、新たな価値創造につなげる。
データ活用を目指す利用者へのコンサルティングや開発支援を実施する。

mdxとは

- データ利活用・セキュリティを重視した高性能仮想化環境
- 9大学2研究所が共同運営し、全国共同利用
- 東京大学 柏2キャンパスに設置

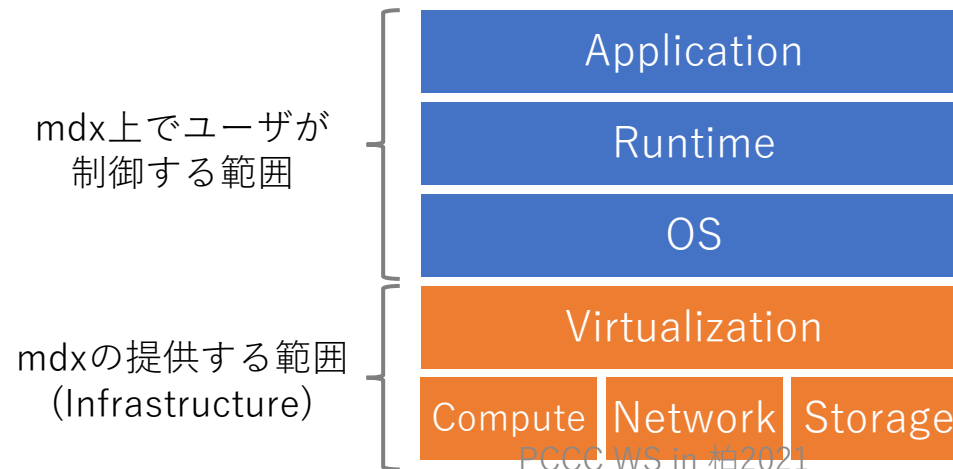




- 仮想プラットフォーム
 - 柔軟・セキュアな環境の構築が可能
 - 高性能な「マルチテナント」環境
- 大容量ストレージ
 - セキュア
 - データ共有
- SINET・モバイルSINETと接続
 - セキュアなIoT環境の構築が可能
 - ABCI, BDECなど既存・将来設置のスパコンと連携

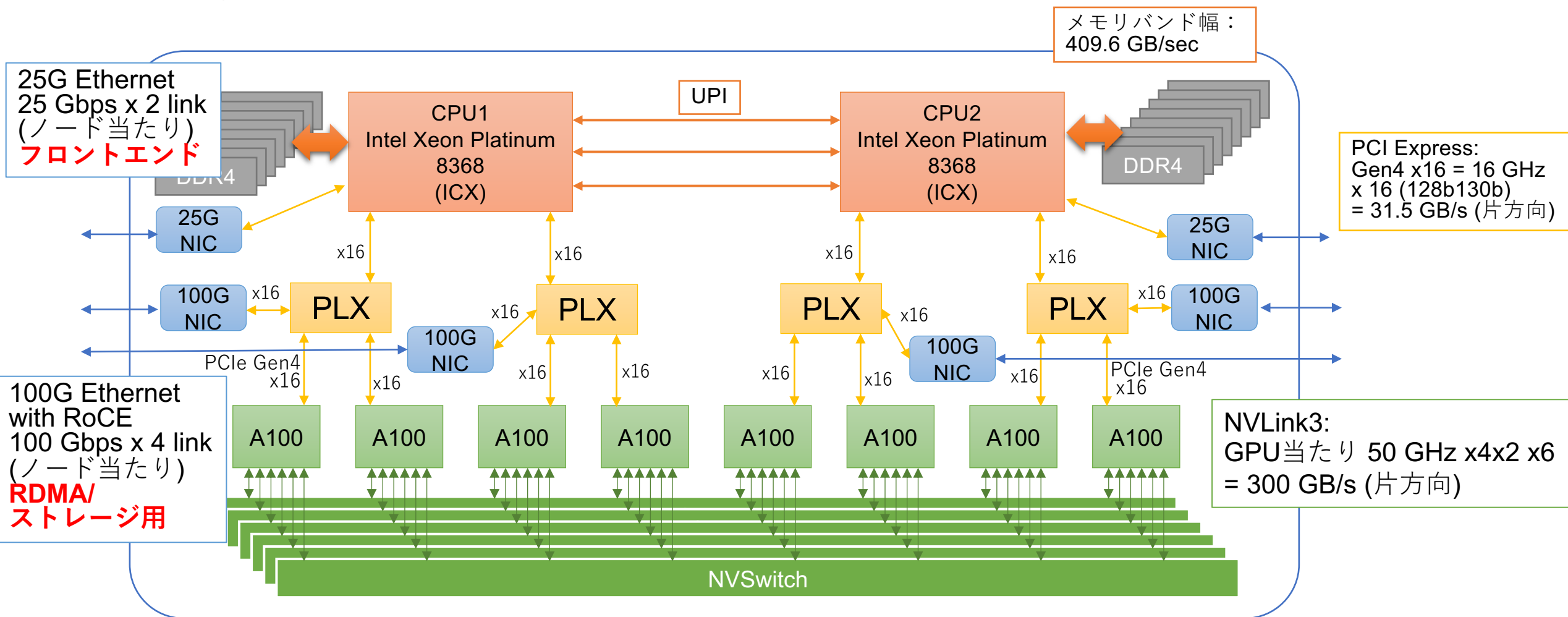
mdxが提供するもの

- 高性能と分離を両立した **Infrastructure as a Service (IaaS)** 基盤
 - Platform as a Service (PaaS) も提供予定
- 計算環境, データ集積のインフラをユーザに提供する
 - 高速なネットワーク、大容量のストレージで支える
- ユーザはインフラ上で仮想マシン(VM)を作成してアプリを作ったりそれを利用する
 - ユーザは仮想マシンに好きなOSを使い、root権限を持って好きにできる
 - ユースケースに則したテンプレートを提供



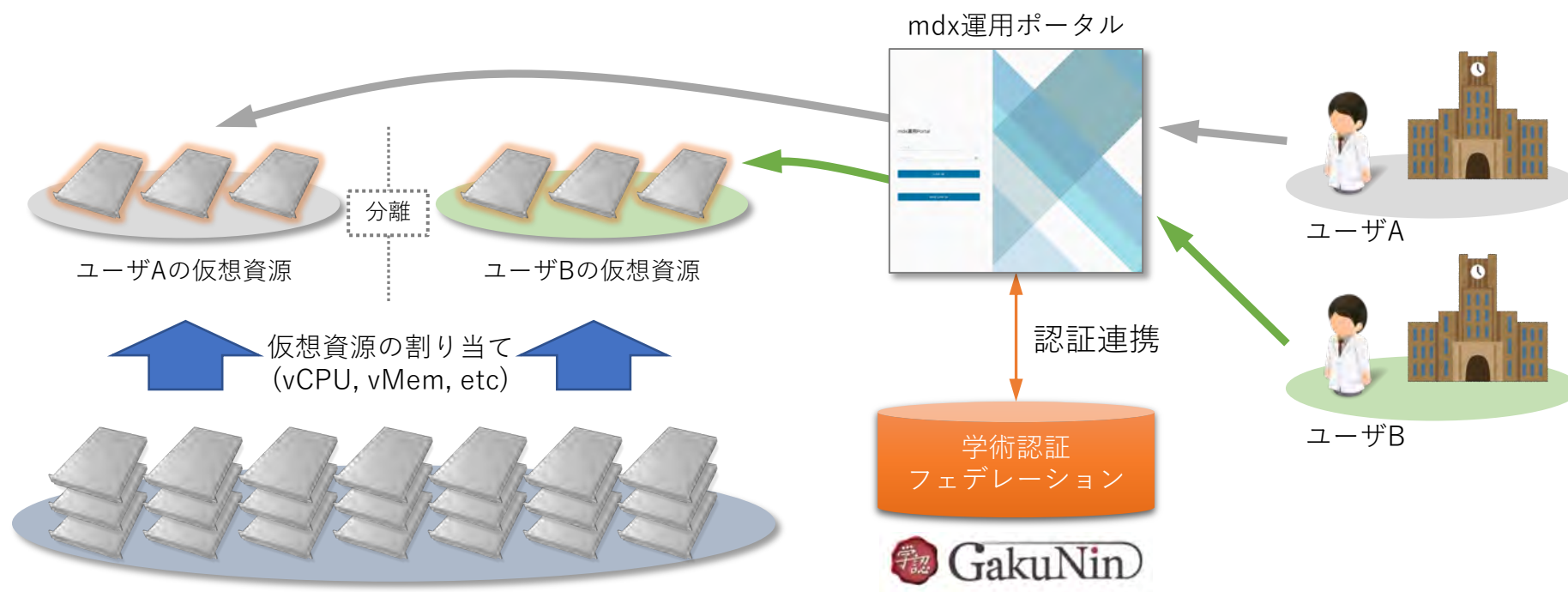
mdx GPU演算加速ノード

- Intel Xeon Platinum 8368 (38c 2.4GHz) x 2ソケット, 512GBメモリ
- ノード当たり8基のNVIDIA A100 GPU



利用イメージ

- ユーザごとに分離された仮想環境を構成
- mdx運用ポータルを通じて仮想環境を操作(VMのデプロイ、設定、etc)
 - 運用ポータルへのログインはNIIが運用する学認と連携(予定) + 独自の認証基盤



- 2021年3月 稼働開始

- 設置場所： 柏IIキャンパス

- <2.0 MVA(冷却込み)、<170 m²
- 東大BDECシステムと同部屋 (2021年5月稼働予定)
- 産総研ABCIとも連携
- 富士通が設置・運用

- 計算ノード

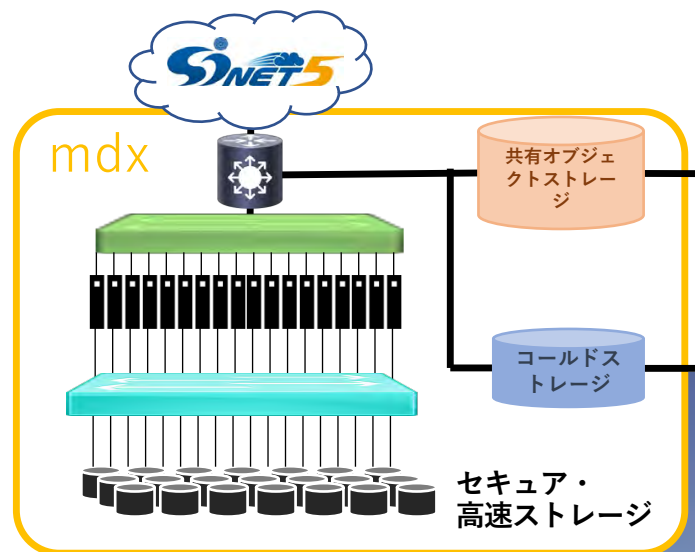
- 汎用CPUノード：Intel Xeon (IceLake-SP) x2ソケット 368ノード, 2.1 PFLOPS (FP64)
- GPUノード：Intel Xeon (IceLake-SP) x2ソケット +NVIDIA A100 x8 GPU, 40ノード
6.4 PF (FP64), 6.7 PF (FP32), 100 PF (FP16)

- ストレージ

- 高速ストレージ：1 PB, 250 GB/s, NVMe SSD, **Lustre** (DDN)
- 内部ストレージ：16 PB, 150+ GB/s, **Lustre** (DDN)
- 共有オブジェクトストレージ: 10 PB, 60+ GB/s, S3 互換 (DDN **Lustre** + S3 Data Service)
- 光ディスクドライブ (Sony ODS)

- ネットワーク

- フロントエンド(Juniper)：25G Ethernet ベース
 - 100G to SINET (400G アップデート予定)
 - 400G to BDEC
- ストレージ, RDMA (Mellanox/NVIDIA)：100G Ethernet with RoCEv2
- EVPN-VXLANによるオーバーレイ



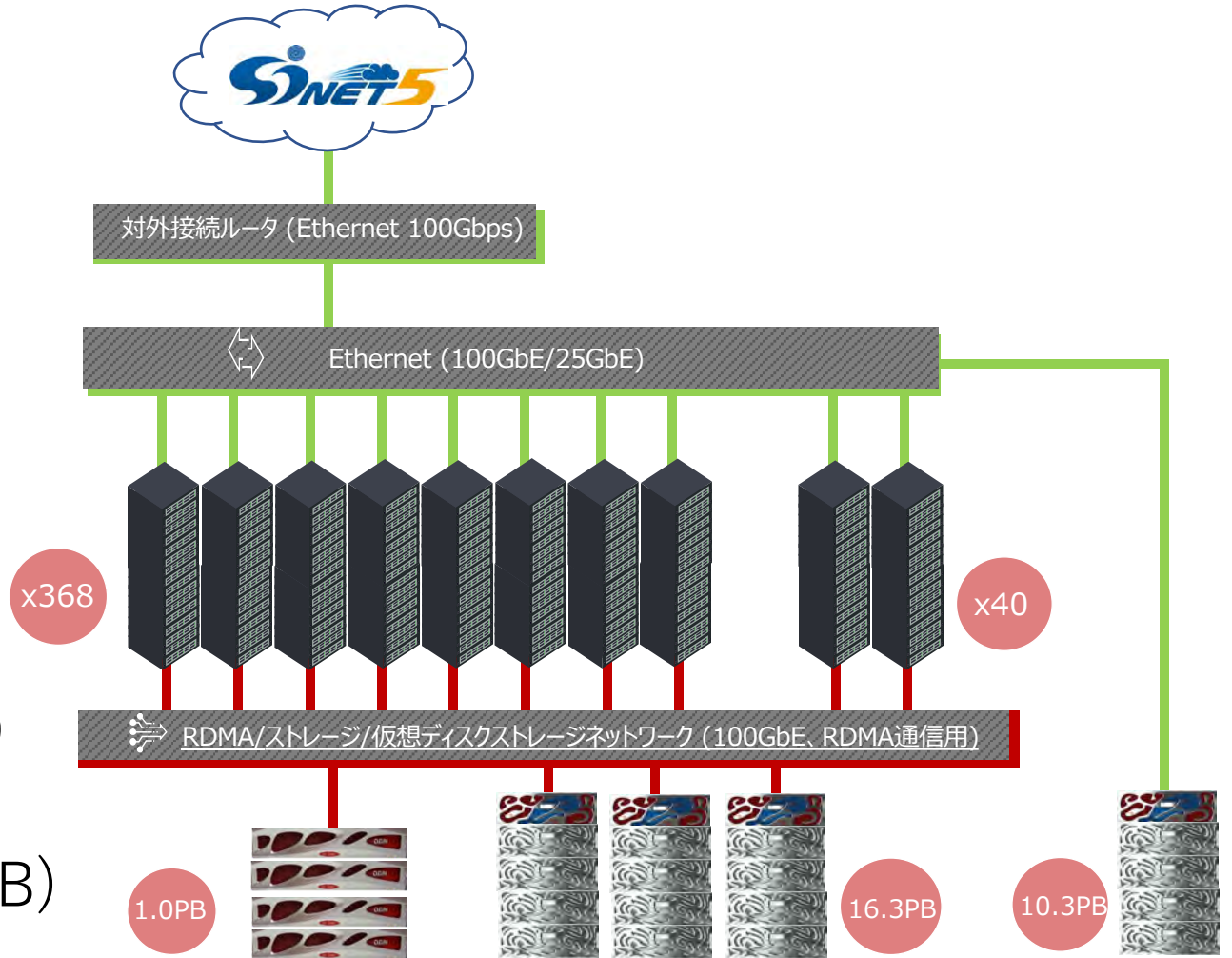
- ソフトウェア

- 仮想化：VMware vSphere
- オーケストレーション



物理構成

- 2つのネットワーク
 - 外部接続ネットワーク
 - SINETと100G x2で接続
 - SINET6で増速予定
 - 内部高速ネットワーク
 - RDMA、ストレージ
- 複数のストレージ
 - 高速ストレージ (1PB, NVMe)
 - 内部ストレージ (16PB)
 - オブジェクトストレージ(10PB)



計算ノード構成

- 汎用CPUノード, 368台
 - Intel Xeon Platinum 8368 (38c, 2.4GHz, IceLake-SP) x2ソケット
- GPUノード, 40台
 - Intel Xeon Platinum 8368 (38c, 2.4GHz, IceLake-SP) x2ソケット
+ NVIDIA A100 x8 GPU
- Hyper Visor
 - VMware vSphere



汎用CPUノード
富士通PRIMERGY CX2550 M6

2021/6/18



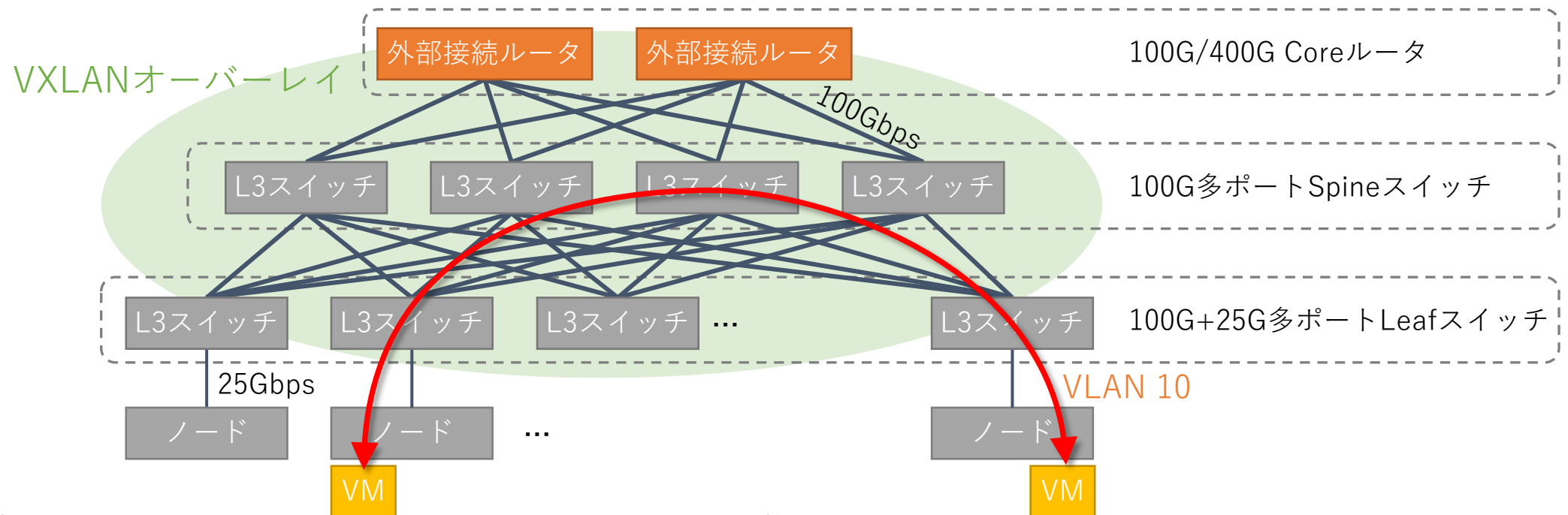
GPUノード
富士通 PRIMERGY GX2570 M7

PCCC WS in 柏2021

- 外部接続ネットワーク向けNICは25Gbps
- 内部高速ネットワーク向けNICは100Gbps

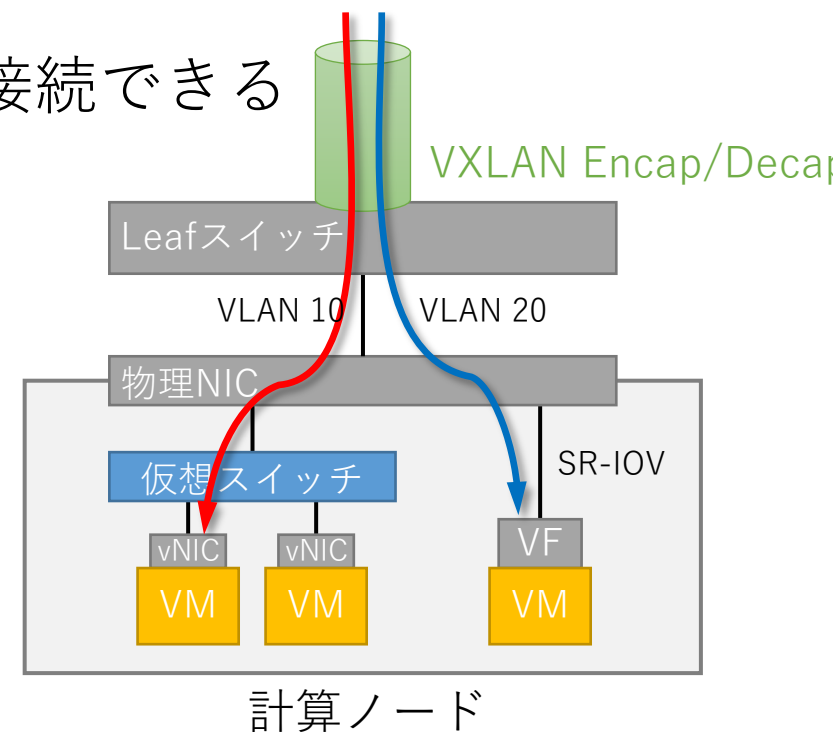
外部接続ネットワーク

- 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)



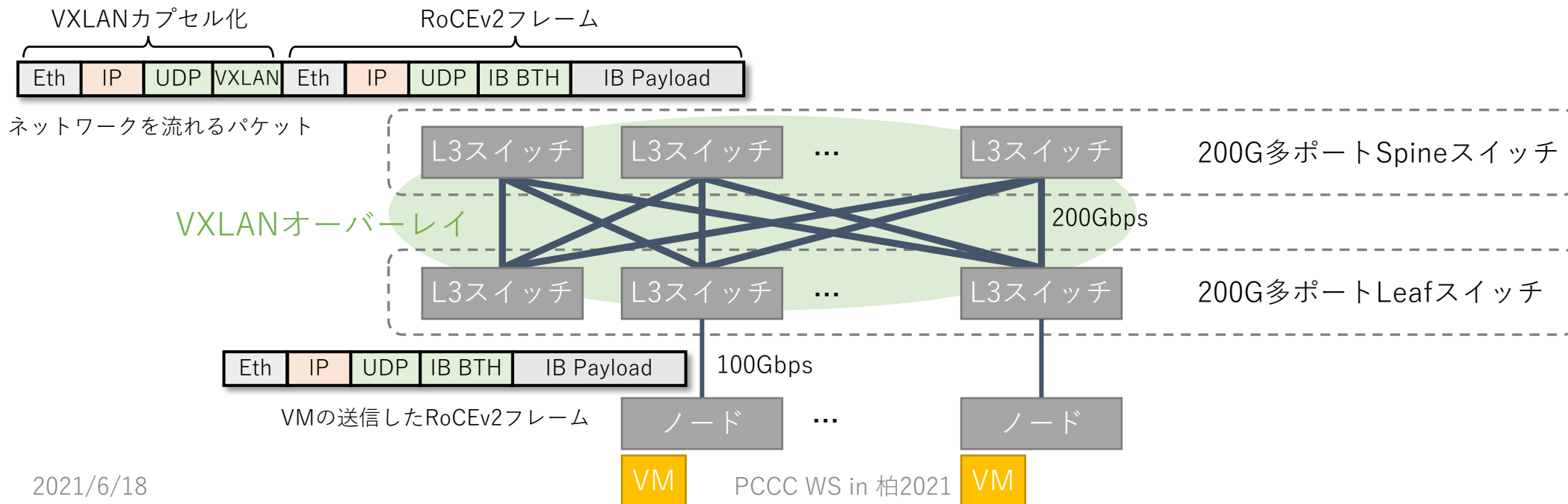
オーバーレイを使う利点

- IPネットワークの利点を維持したままVLANを延伸
 - 複数パスを使ったロードバランスと冗長
 - ユーザのVMがどのノードにいても同じVLANに接続できる
- 仮想マシンのネットワーク性能と分離
 - トンネルやルーティングの処理をHVのS/Wではなく外部のH/Wスイッチで実施
 - 仮想マシンのNICが仮想NICでもSR-IOVでもネットワーク側は1つのロジックで制御できる
 - 万が一HVが乗っ取られても、他のVLANにはアクセスできない



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



VMの割り当てポリシー（予定）

資源割り当て

- 汎用CPUノード：コア単位で指定、
コア当たりメモリ量を比例配分
- 演算加速GPUノード：GPU単位で指定、
コア数、メモリ量はノード全体の 1/8ずつ
- ノード占有：ノード単位での割り当て

内部高速ネットワーク、GPUの接続

内部高速ネットワークへの接続

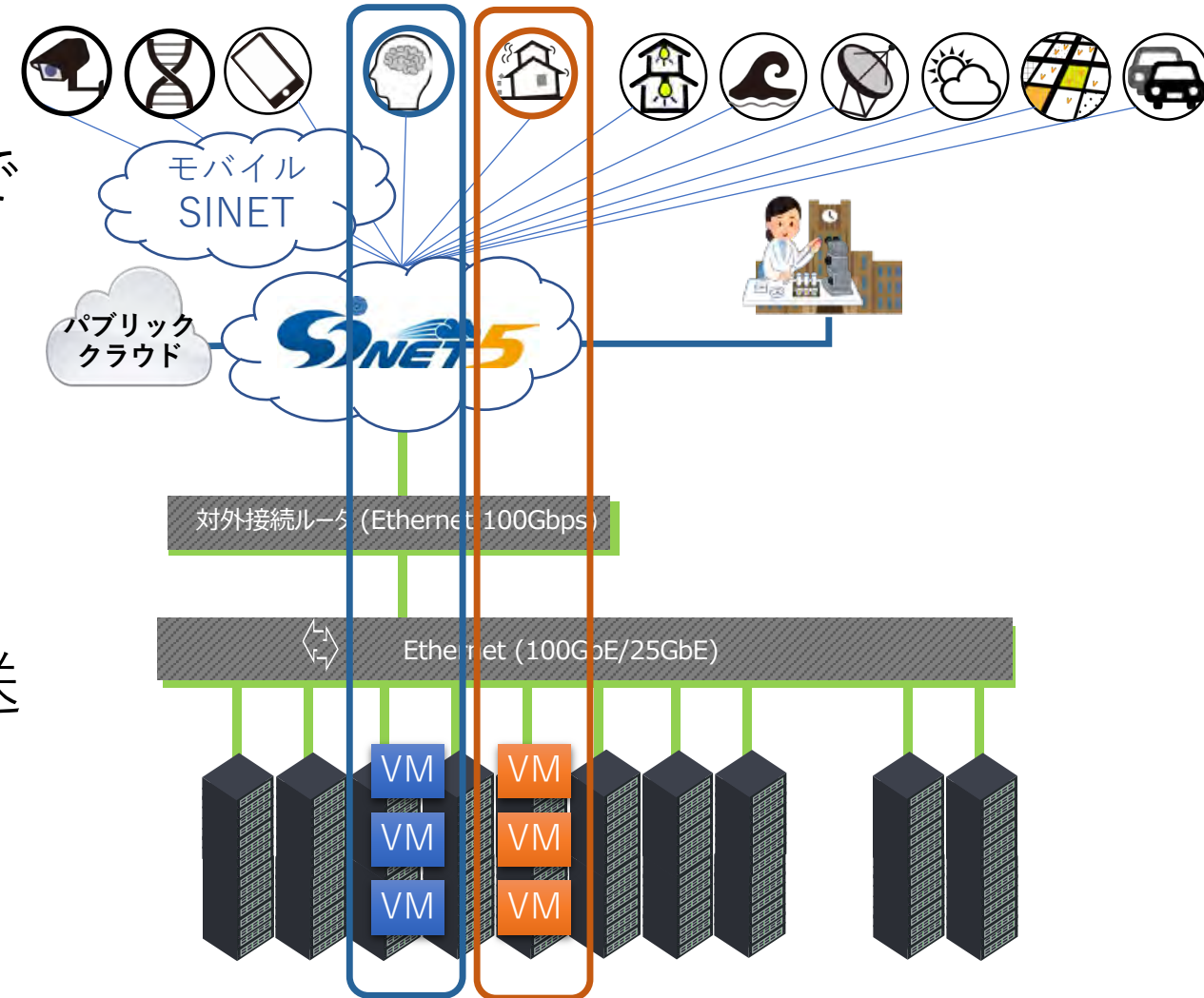
- デフォルト: VM間のグループ、LustreはTCP接続
- PVRDMA (準仮想化RDMA): ノード間のRDMA通信は可能
- SR-IOV: ノード間、Lustreも含めRDMAによる通信

GPU接続

- PCIe パススルー: ベアメタルと同様のGPU直接アクセス

SINETとの連携

- SINETのL2VPNサービスを通じて、セキュアな閉域網経由でVLANをmdxに接続可能
- 他のSINETサイト(大学・研究機関)ともセキュアに接続
- モバイルSINET経由で、IoTデバイスからmdx上の計算環境へセキュアにデータを転送



まとめ

- データ科学・活用のための基盤 **mdx:データ活用社会創成プラットフォーム**を導入
 - 9大学（北大、東北大、筑波大、東大、東工大、名大、京大、阪大、九大）
 - NII, AIST
- mdxの特徴
 - 幅広く利用可能なIaaSベースの仮想化計算基盤
 - 高性能計算機・ストレージ
 - VPNによる隔離
 - SINETとの連携による、セキュアIoT（モバイルSINET）、他のスパコンとの密な連携
- mdxの利用について
 - <https://mdx.jp/>
 - 利用開始に向けて鋭意準備中
 - mdxを利用したデータ集積およびデータ提供、データ解析、高性能処理などのプラットフォーム構築を、幅広い分野の方々が容易に実現できるように