

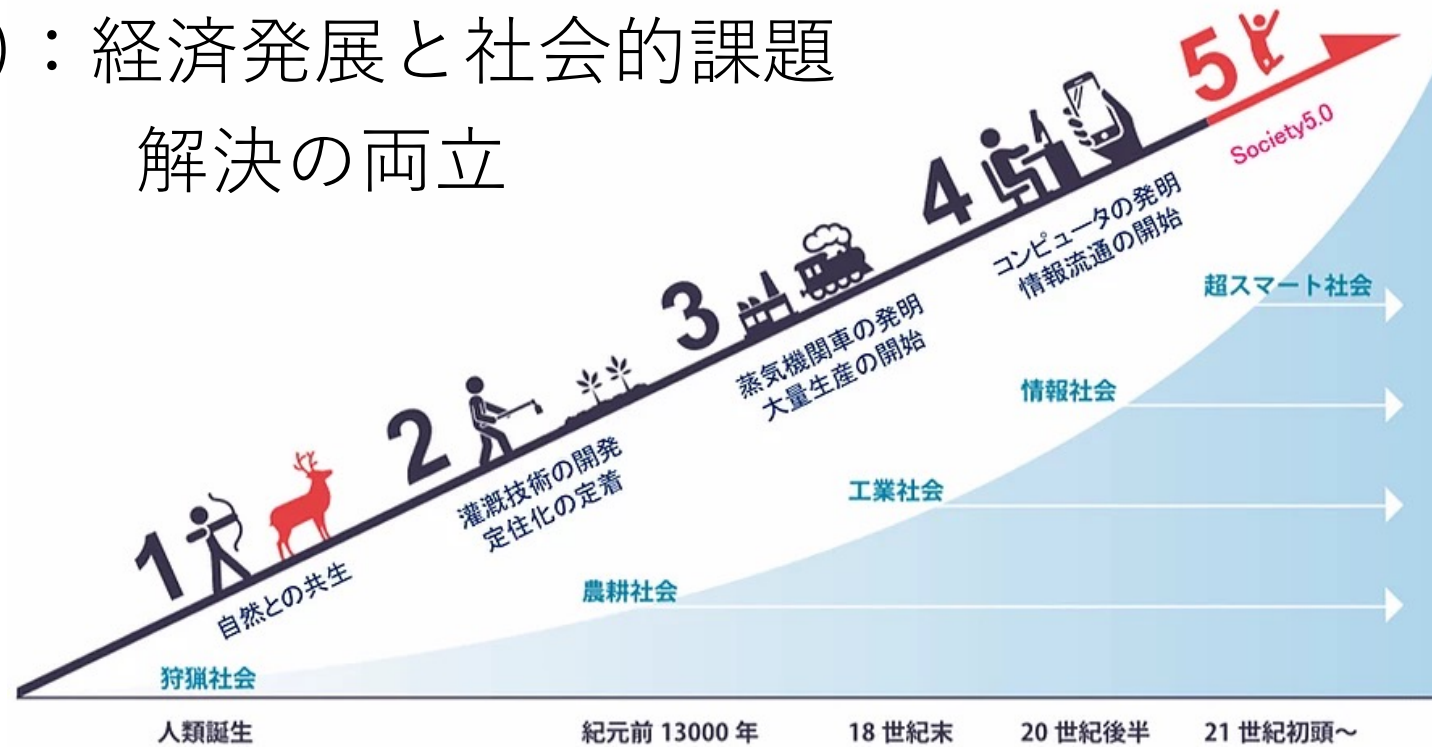
データ利活用に向けた 仮想化基盤mdx

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



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

データ利活用の推進を支える環境

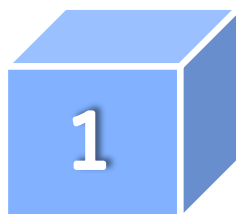
用途に応じてオンデマンドで**短時間に構築・拡張・融合**できる
データ収集・集積・解析機能を提供するプラットフォーム



データ活用社会創成
プラットフォーム計画

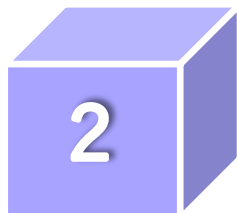


3 本柱



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。

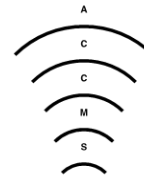
- データ利活用・セキュリティを重視した高性能仮想化環境
- 9大学2研究所が共同運営し、全国共同利用
- 東京大学 柏2キャンパスに設置

研究所 (2)

- 国立情報学研究所、産業技術総合研究所

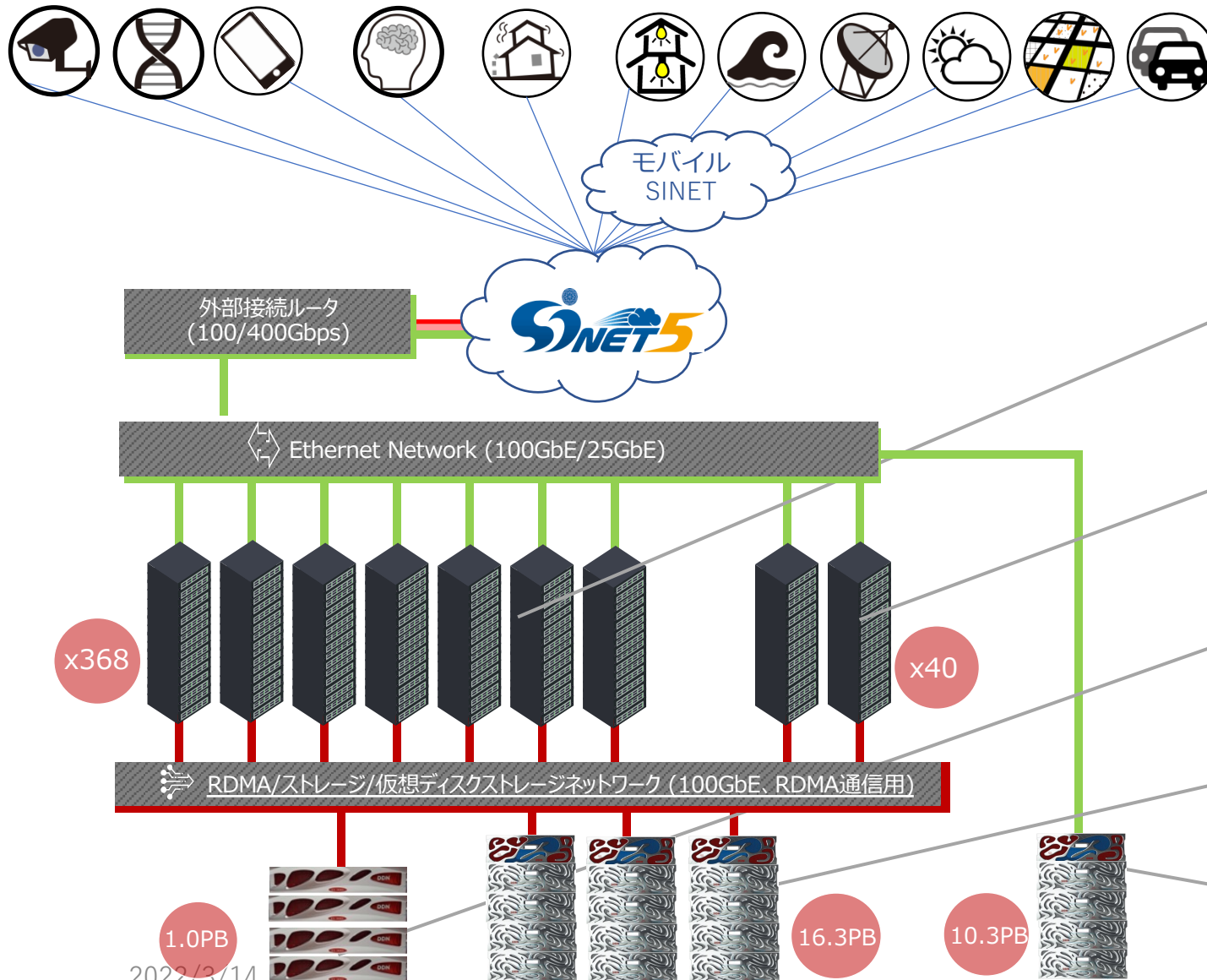
大学 (9)

- 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





mdxの構成



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
 - ➡ 2022年4月 **400G x2** 増速
- 内部高速ネットワーク
 - ✓ RDMA、ストレージ

汎用CPUノード

Intel IceLake x2ソケット **x368ノード**
 理論ピーク性能(FP64): **2.1PFLOPS**
 総メモリバンド幅: 150.7 TB/s

GPU 演算加速ノード

Intel IceLake x2ソケット+NVIDIA A100 x8 **x40ノード**
 理論ピーク性能(FP64): **6.4PFLOPS**
 理論ピーク性能(FP16): **100.7PFLOPS**
 総メモリバンド幅: 496.3 TB/s

高速 NVMe ストレージ

Lustre Filesystem
 1.0 PB (NVMe SSD)
 252 GByte/sec

大容量HDDストレージ

Lustre Filesystem
 16.3 PB (HDD)
 157.5 GByte/sec

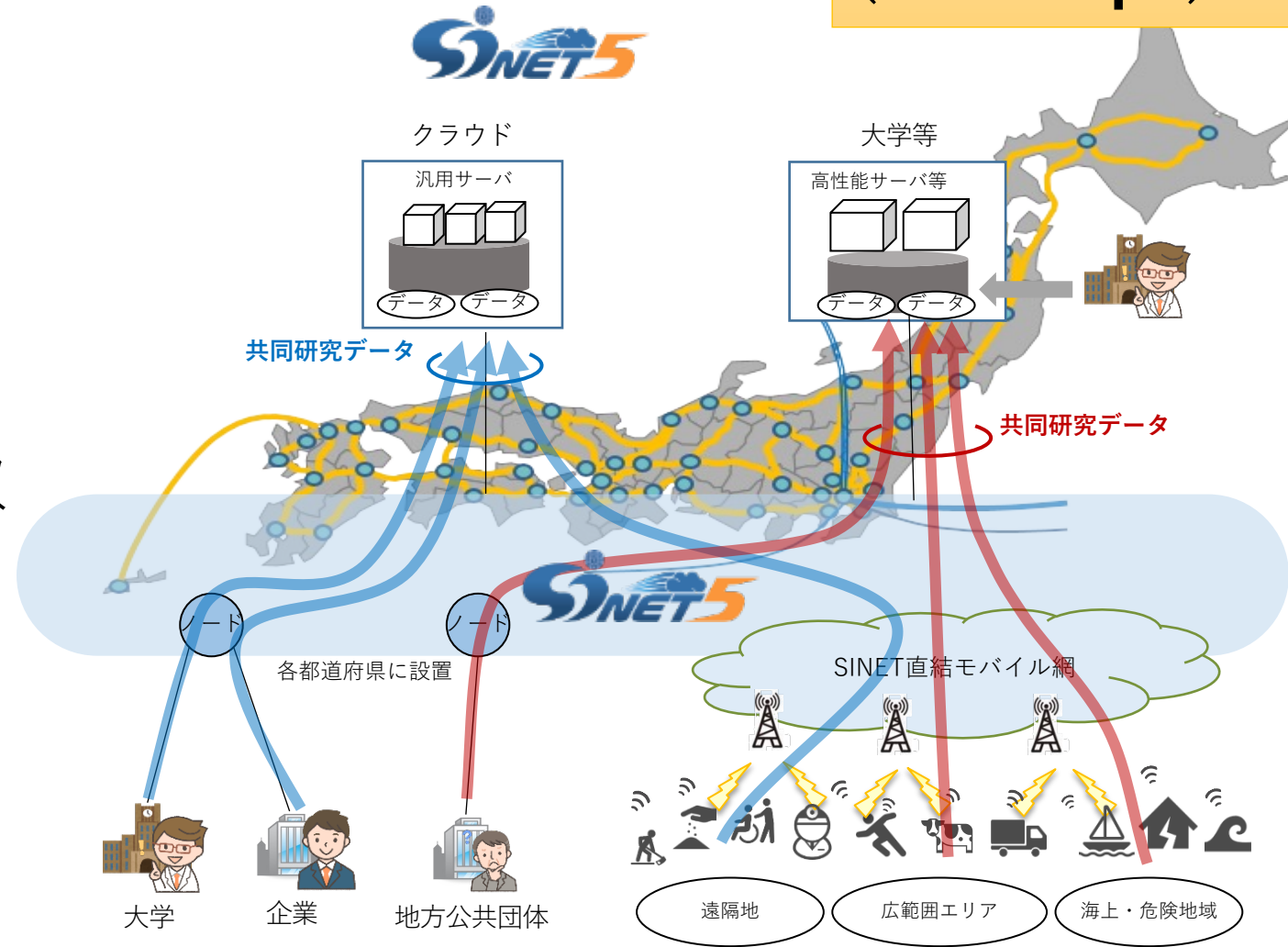
外部共有オブジェクトストレージ

S3 Data Service
 10.3 PB (HDD)
 63.0 GByte/sec

SINET 5・モバイルSINET

SINET 6
(400G bps)へ

- SINET 5 (有線)
 - 全都道府県にノード
 - ノード間は $\geq 100\text{Gbps}$
 - 冗長経路
 - 広域VPN (L3/L2)
- モバイルSINET (広域データ収集基盤)
 - SINET VPNに直結できるモバイル環境
- mdxの個別のプロジェクトからVPNをSINETへ延伸可
 - IoTデバイスのセキュリティ確保に有用



mdx全景



- 免震台(ストレージのみ)
- 入室管理
 - スパコン室 + ケージ



計算ノード

- 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



汎用CPUノード
富士通PRIMERGY CX2550 M6

2022/3/14

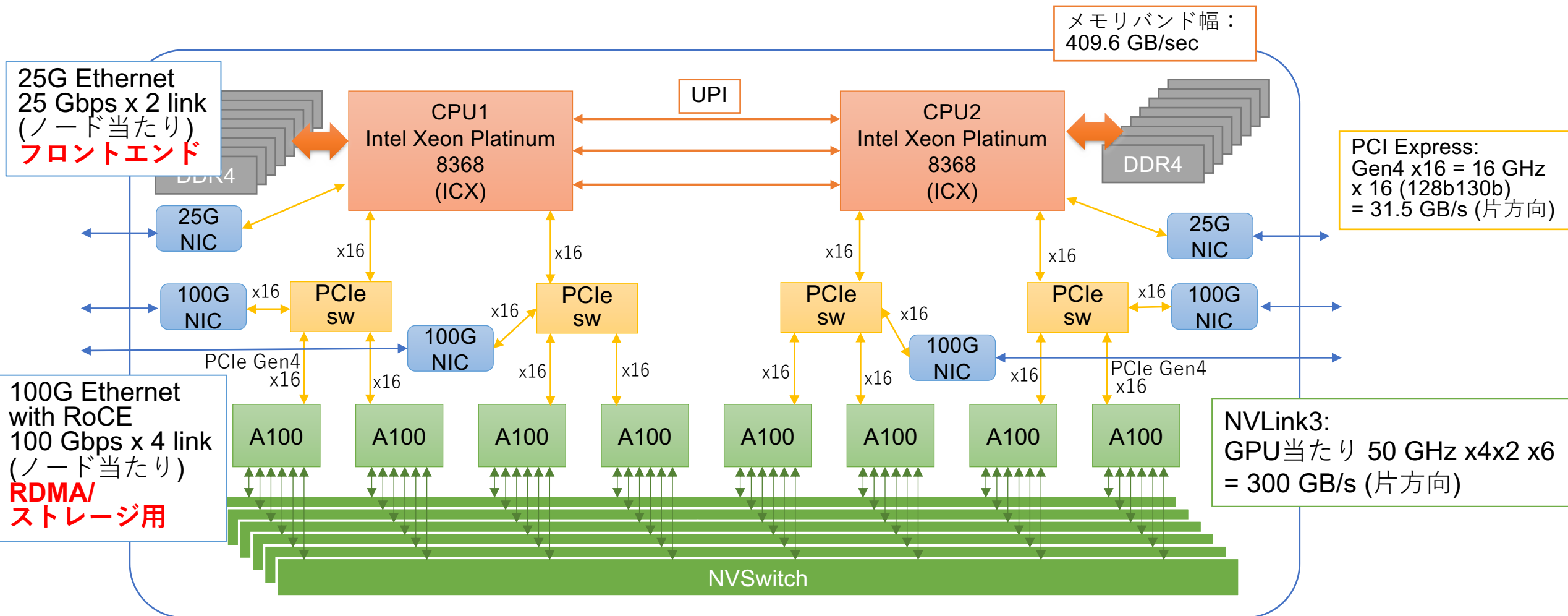


GPUノード
富士通 PRIMERGY GX2570 M6

- 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

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



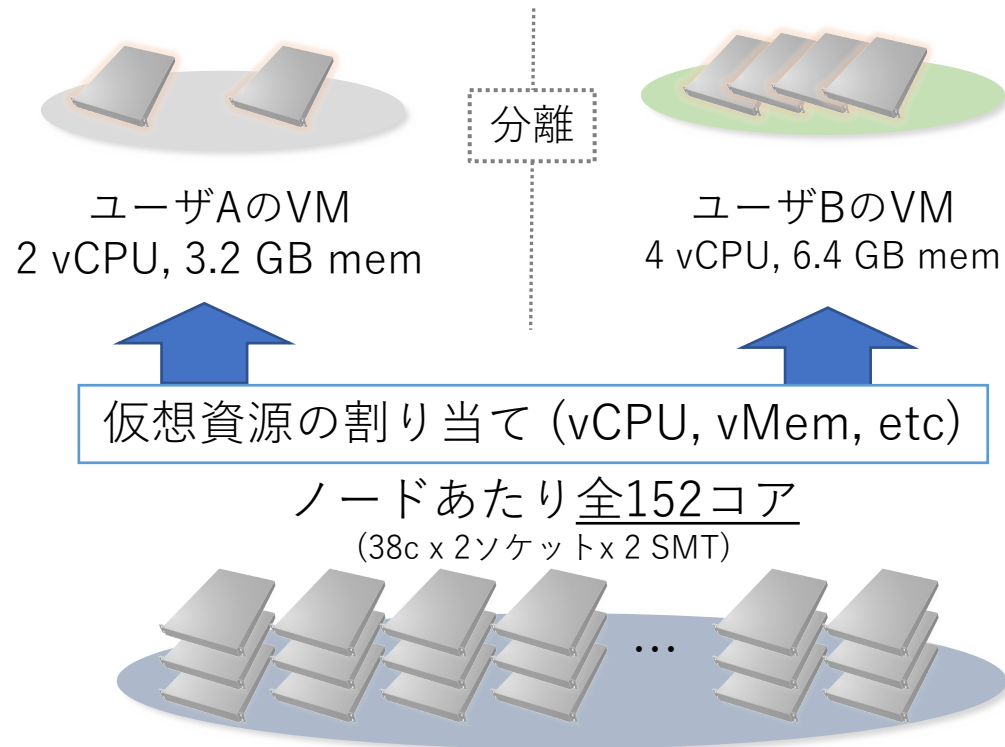
計算ノード

ノード占有も可能

仮想化ソフトウェア：VMware vSphere (vCenter, ESXi)

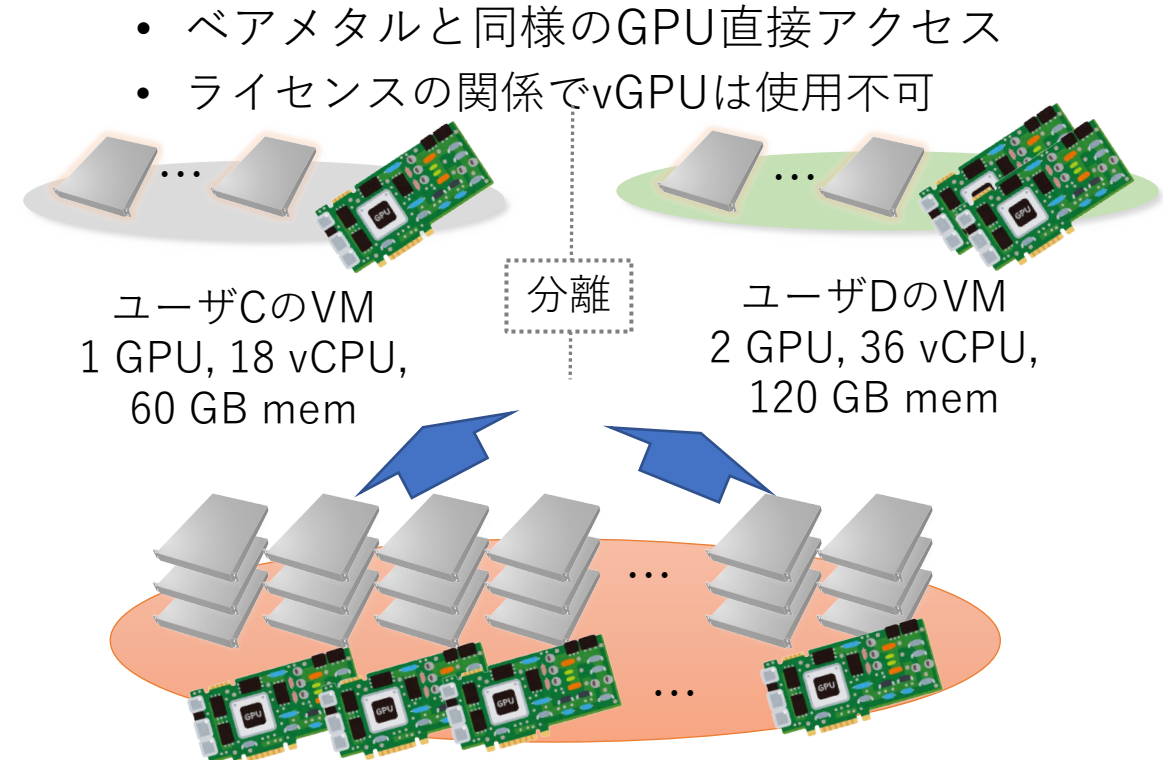
汎用CPUノード

- 1 vCPU+メモリ1.6 GB = 1CPUパック



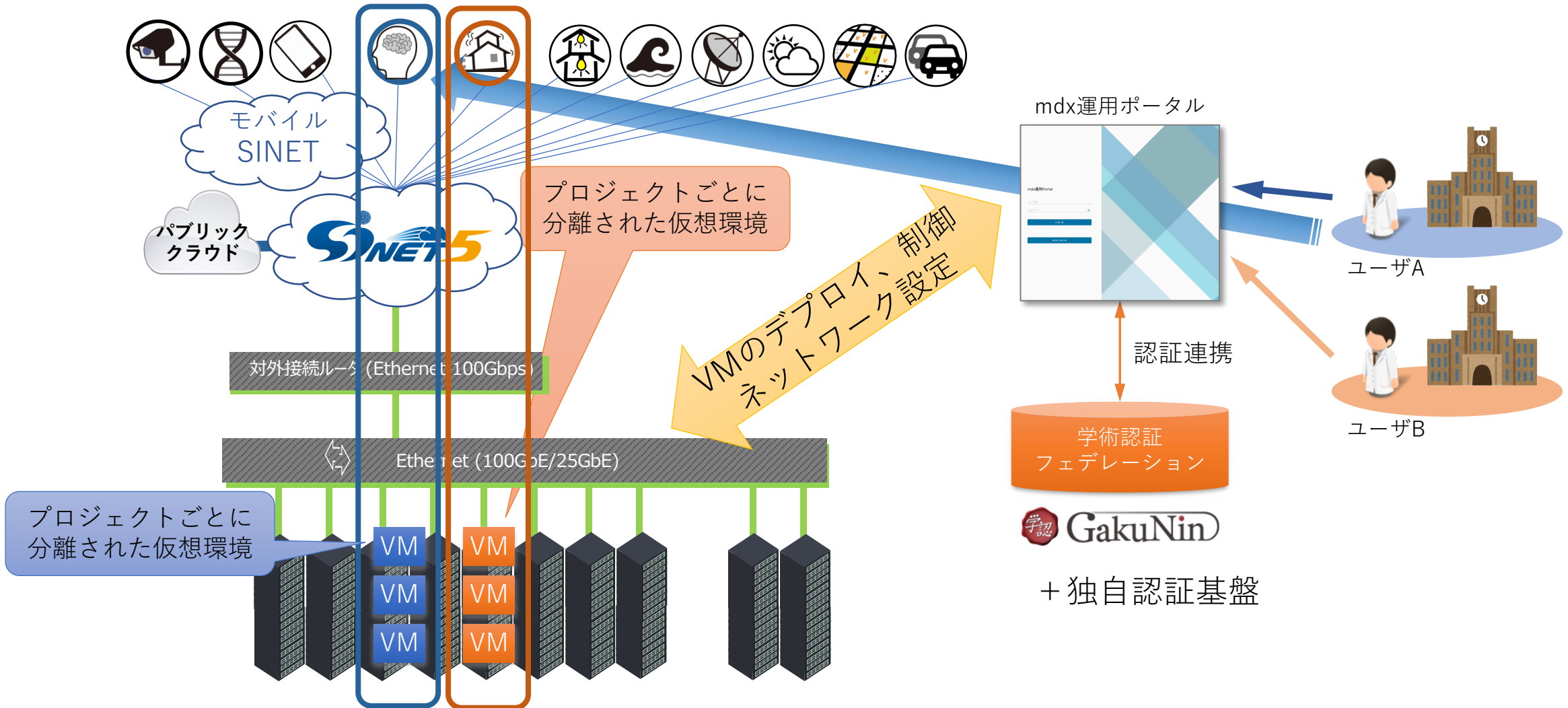
演算加速GPUノード

- 1GPU+18 vCPU+メモリ60GB = 1GPUパック
- GPU接続: PCIe パススルー



「学認」を使える方は <https://mdx.jp> からプロジェクト申請が可能です。
まずはご相談ください！

利用イメージ



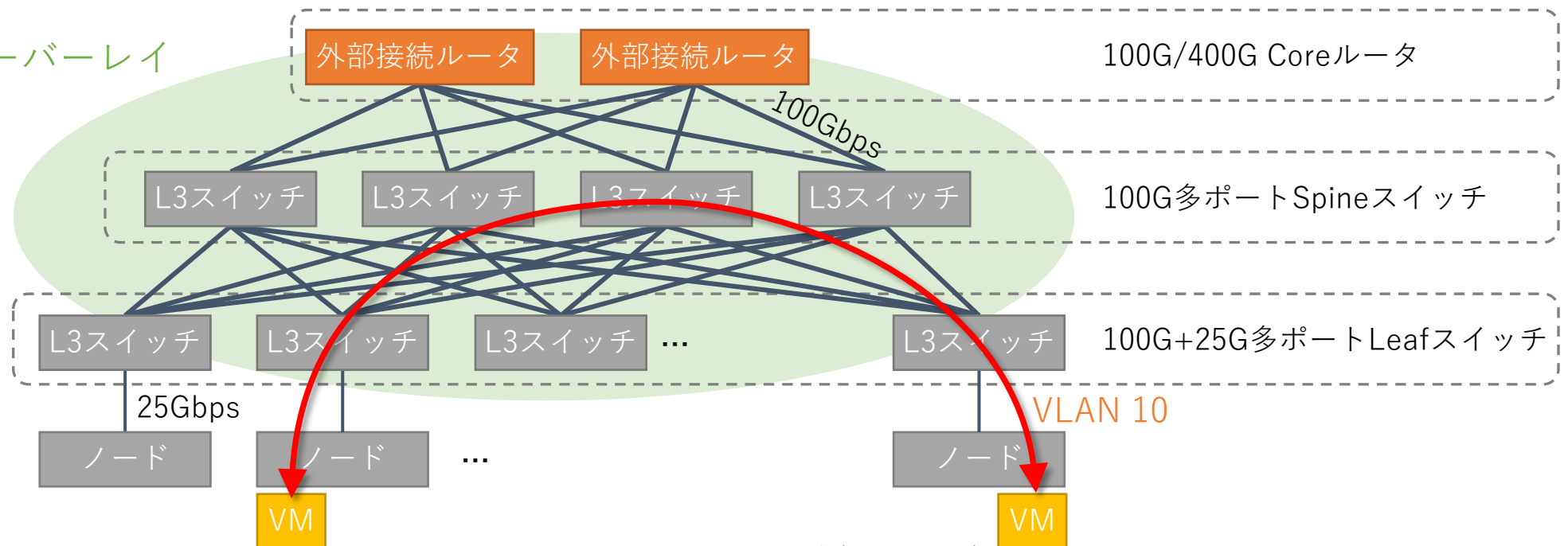
ネットワークインタフェース

- 外部接続ネットワーク：VMXNET3（準仮想化）
 - CPUノード：ノードあたり 25 Gbps Ethernet
 - GPUノード：ノードあたり 25 Gbps Ethernet x2
- 内部高速ネットワーク（ストレージ，ノード間通信用）への接続
 - 各ノードのNICは100Gbps: x1/CPUノード、x4/GPUノード
 - デフォルト (VMXNET3, RDMA無し)： VM間， Lustre共にTCP接続
 - PVRDMA(準仮想化RDMA)： VM間のRDMA通信， Lustreは未対応
 - SR-IOV: VM間， Lustreも含めRDMAによる通信
 - 動作には mlx5ドライバが必要

外部接続ネットワーク

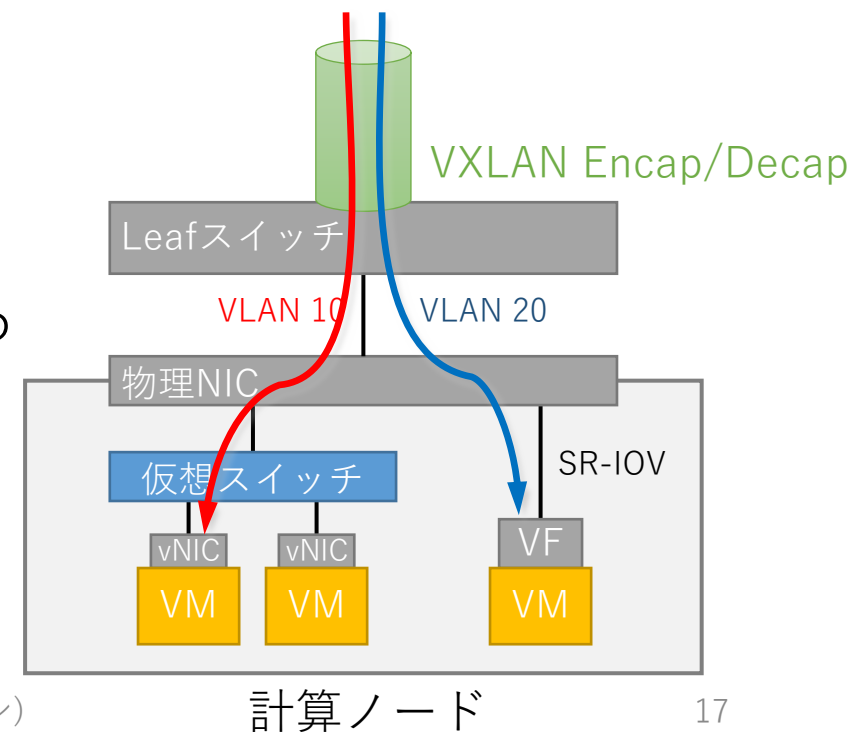
- Virtual eXtensible LAN (VXLAN)によるオーバーレイネットワーク
 - データセンター向けの仮想ネットワーク技術
 - 物理はSpine/LeafのIP Clos Fabric ネットワーク
 - IPネットワーク上にVXLANでVLANを通す(Ethernet over IP)

VXLANオーバーレイ



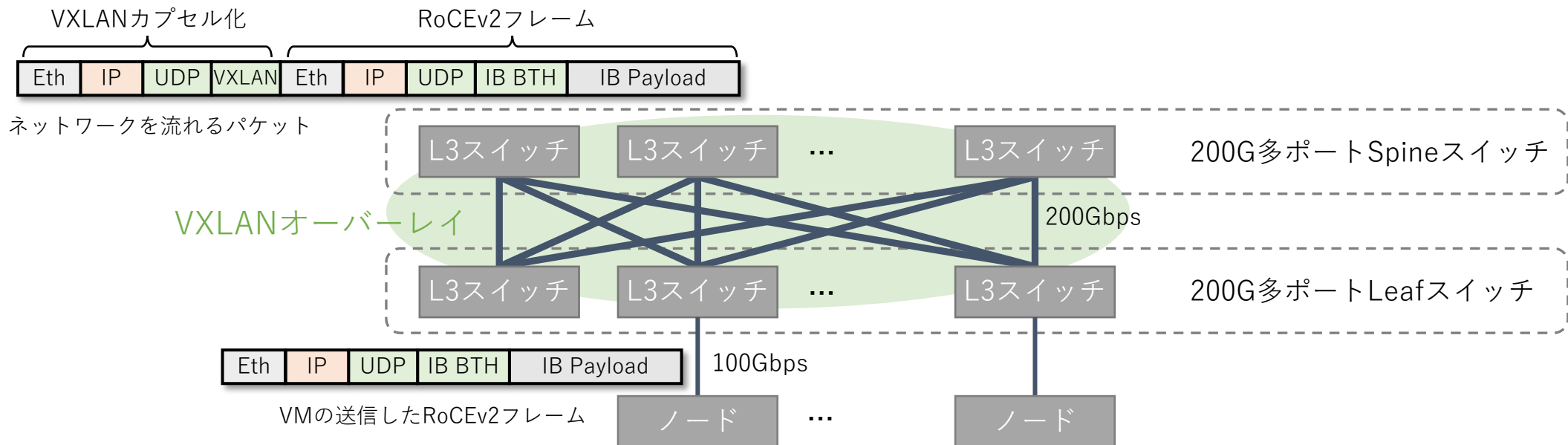
オーバーレイを使う利点

- 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がどのノードに存在しても同じ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
 - Infinibandでは十分なテナント間の分離を実現できないため
 - 外部接続ネットワークと同様、IP Clos topologyにVXLANを利用
 - 各ノードのNICは100Gbps, ネットワークはフルバイセクション
 - Mellanox ConnectX-6 Dx x1/汎用CPUノード、x4/演算加速GPUノード



外部とのデータ連携

- データ利活用・解析のためには…

- データ共有、検索
- 研究データ管理、公開

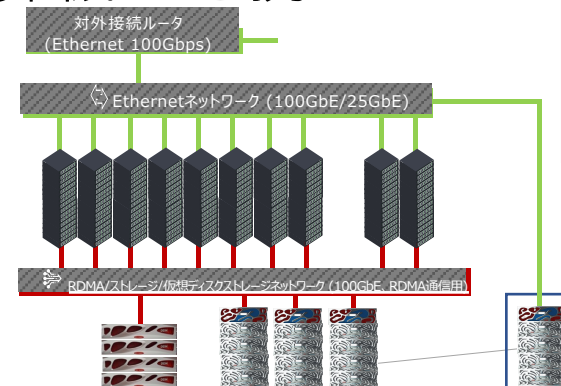


NII Research Data Cloud

- Gakunin RDM : データ管理
- JAIRO Cloud: データ公開
- CiNii: データ検索

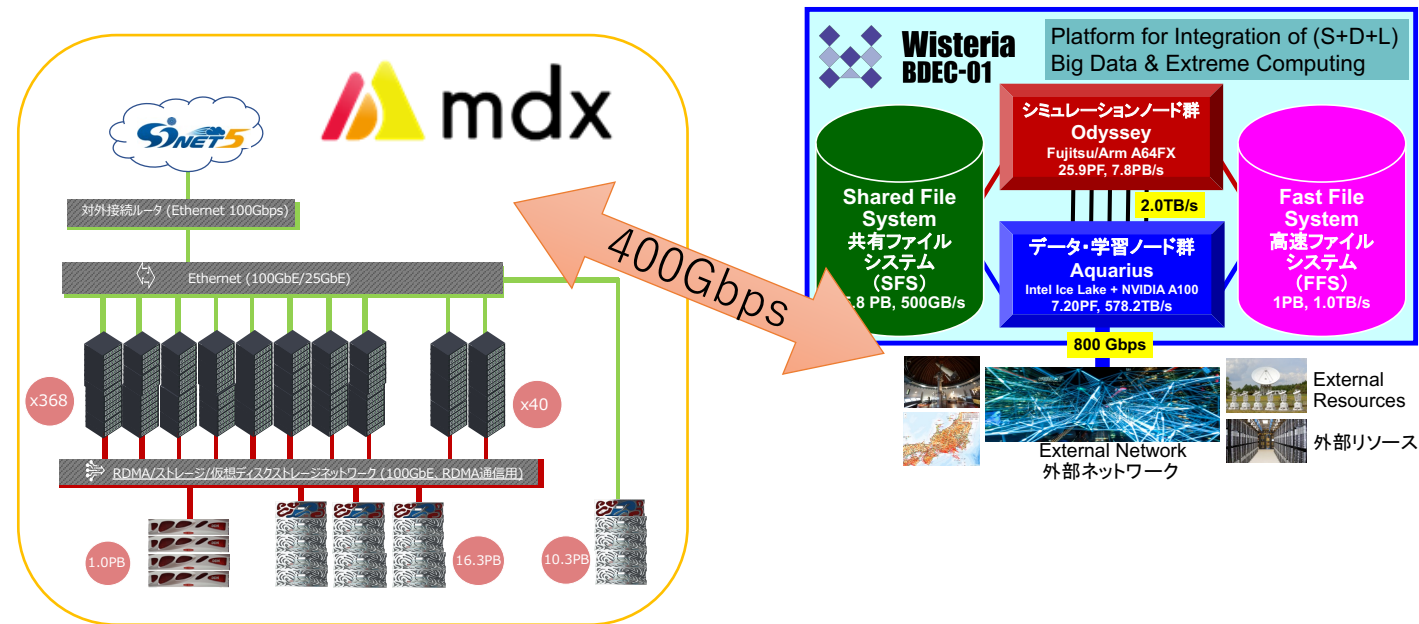
- Gakunin RDMとの連携

- mdxのオブジェクトストレージと連携
- BinderHubを用いたストレージ-計算機の連携
- 内部大容量ストレージ
(Lustreファイルシステム)
との連携



各スパコンとの密な連携

- 目的：高負荷・大容量処理をスパコンにオフロード
 - スパコンはバッチ処理主体：時間的な制約が少ない場合が前提
- SINET6により各スパコンセンターとmdxは高速に接続
 - mdxとWisteria/BDEC-01は 400 Gbps x2 で直結
- 課題
 - ストレージの連携
 - scp, sftpまたはsshfs
 - S3オブジェクトストレージ
 - LustreのNFS export
 - ネットワーク構成
 - スパコン側での実行
 - ジョブのsubmission
 - Python, JupyterLab
 - Container



高付加価値 気象データプラットフォーム

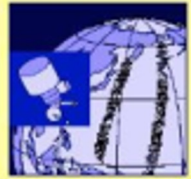
JST共創の場 地域気象データと先端科学による戦略的社会共創拠点 ClimCORE

日本域気象再解析: 観測データと数値モデルの融合による大気場再現

過去の断片的
各種観測データ



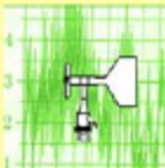
最新の数値予報モデルに取込む「領域再解析」により、①大気状態の全体像を長期にわたり均質に矛盾なく4次元的に再現することで、②時空間的に断片的な観測データに加えて観測不能量も含めた全ての気象要素に関する高時空間分解能データとして社会で広く活用できるよう付加価値を高める。



衛星観測データ



高層観測データ



地上観測データ



海洋観測データ



全球大気再解析

最新の全球数値予報モデルと
データ同化システムによる、
過去観測データの全球再解析

JRA-55 (2013公開*)

* 2013以降も継続中

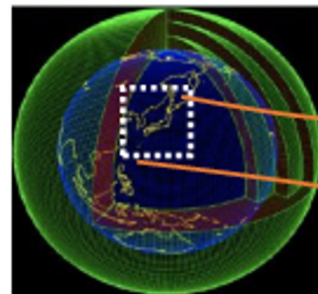
(1958-2012)

60kmメッシュ・6時間毎

JRA-3Q (2021公開予定)

(1947-2021)

40kmメッシュ・6時間毎



世界トップクラス

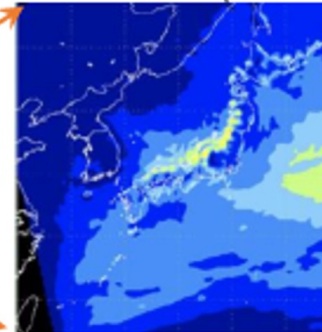
日本域気象再解析

最新の領域数値予報モデルと
データ同化システムによる、
過去観測データの領域再解析

RRJRA (仮称)

従来にはない過去～現在の
詳細な4次元地域気象データ

～5kmメッシュ (1時間毎?)



格段に多くの
潜在的利用者
だが、日本では
未整備

※地表付近は1kmメッシュに
ダウンスケール (15分毎?)

PL、スライド提供:
東大先端研 中村尚先生

mdx: データ活用社会創成 プラットフォーム

- 高性能計算機(*)
スパコン性能：25 PFLOPS
- 大容量データストレージ(*)
26 PB
- 高速ネットワーク
SINET 5

(*スーパーコンピューター：
Oakforest-PACS)

情報基盤センター

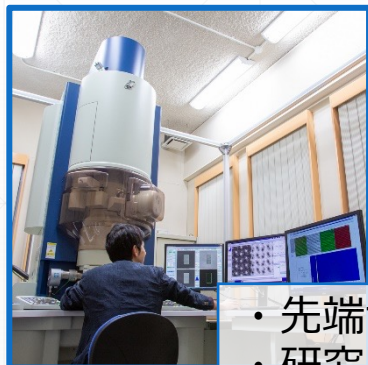


- データ管理・構造化
- データ利活用
- データ人材育成

情報基盤センター mdx

ナノ工学研究セン
ター

微細構造解析PF



- 先端ナノ計測
- 研究開発・支援
- データ創出

武田先端知
スーパークリーンルーム

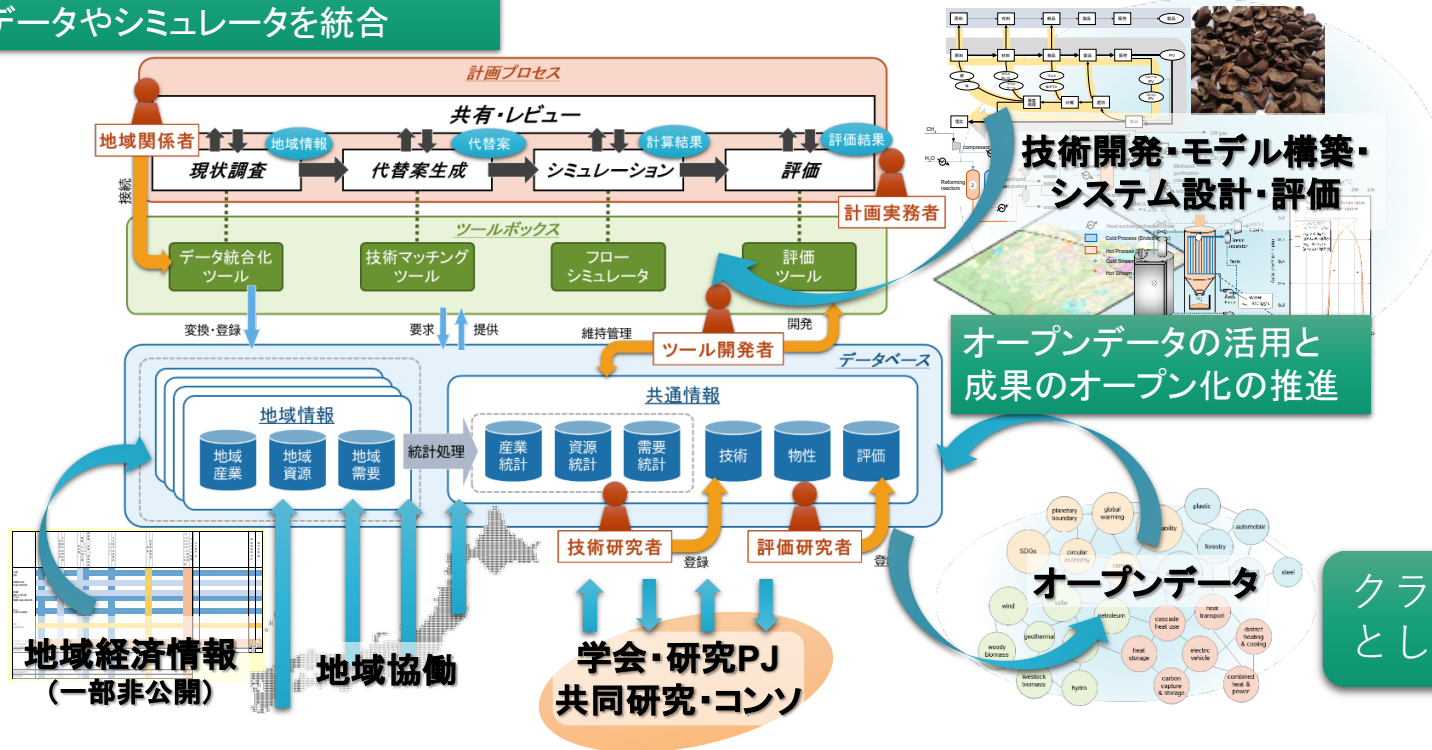
微細加工PF



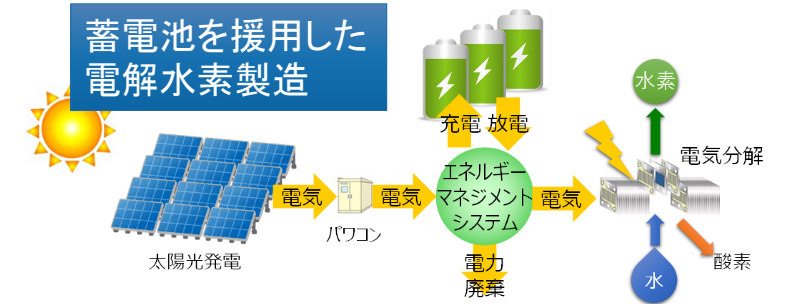
- 先端デバイス加工
- 研究開発・支援
- データ創出

mdx 地域資源を活用した物質・エネルギー生産システム設計のためのシミュレーション基盤

個別事例で蓄積した地域・技術のデータやシミュレータを統合



PL: 東大未来ビジョン研究センター 菊池康紀先生
スライド提供: 東大「プラチナ社会」総括寄付講座 兼松 祐一郎 先生

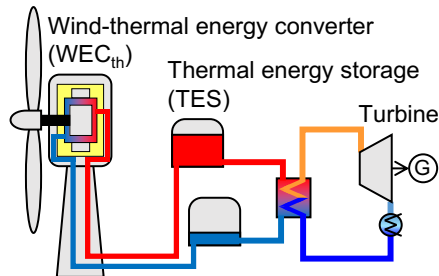


http://www.k.u-tokyo.ac.jp/info/entry/22_entry706/
<https://www.nims.go.jp/news/press/2018/12/201812130.html>

Kikuchi et al., *Int J Hydrogen Energy*, **44**, 1451–1465 (2018)
Sako, Okubo, Kikuchi et al., *J. Clean. Prod.*, **298**, 126809 (2021)

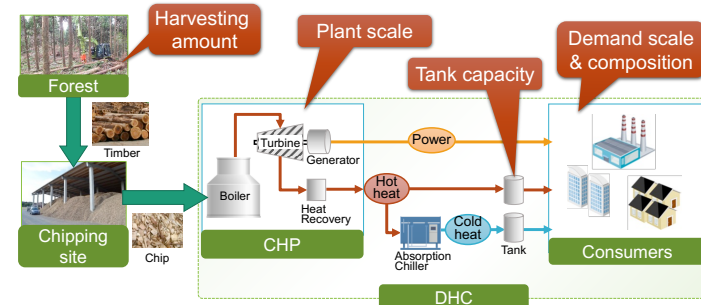
クラウド高速計算環境
としてmdx活用を検討中

風力熱発電+蓄熱



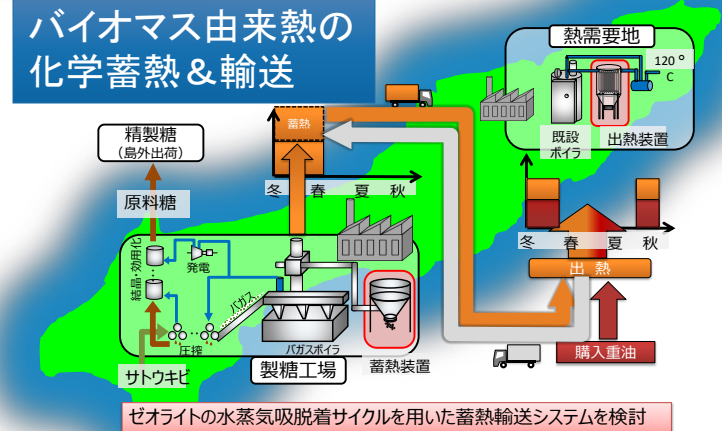
Yamaki, Kanematsu, Kikuchi, *Energy*, **205**, 118056 (2020)

木質バイオマス地域熱供給



Kanematsu, Oosawa, Okubo, Kikuchi, *Applied Energy*, **198**, 160–172 (2017)

バイオマス由来熱の化学蓄熱&輸送



ゼオライトの水蒸気吸脱着サイクルを用いた蓄熱輸送システムを検討

Fujii, Kanematsu, Kikuchi et al., *Appl. Energy*, **238**, 561-571. (2019)

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）、2研究所（NII, AIST）
- mdxの特徴
 - 幅広く利用可能なマルチテナントの仮想化計算基盤
 - 「学認」による利用申請
 - Gakunin RDMとの連携
- VPNによる隔離
- SINET6との連携によるセキュアIoT（モバイルSINET）
- スパコンとmdxとの密な連携

- スパコン見学会録画 (mdx & Wisteria/BDEC-01)
<https://www.youtube.com/watch?v=utCmjkkZPOM>
- mdx紹介と研究事例イメージ
<https://www.youtube.com/watch?v=--7h6tWECX8>