資料4-3-1

UTokyo

東京大学情報基盤センター
INFORMATION TECHNOLOGY CENTER, THE UNIVERSITY OF TOKYO

AI for Scienceに不可欠な計算資源の戦略的増強

（ｉ）共用計算資源の大規模増強を図る取組

東京大学　情報基盤センター
事業代表者：　千葉　滋（情報基盤センター長）

事業概要

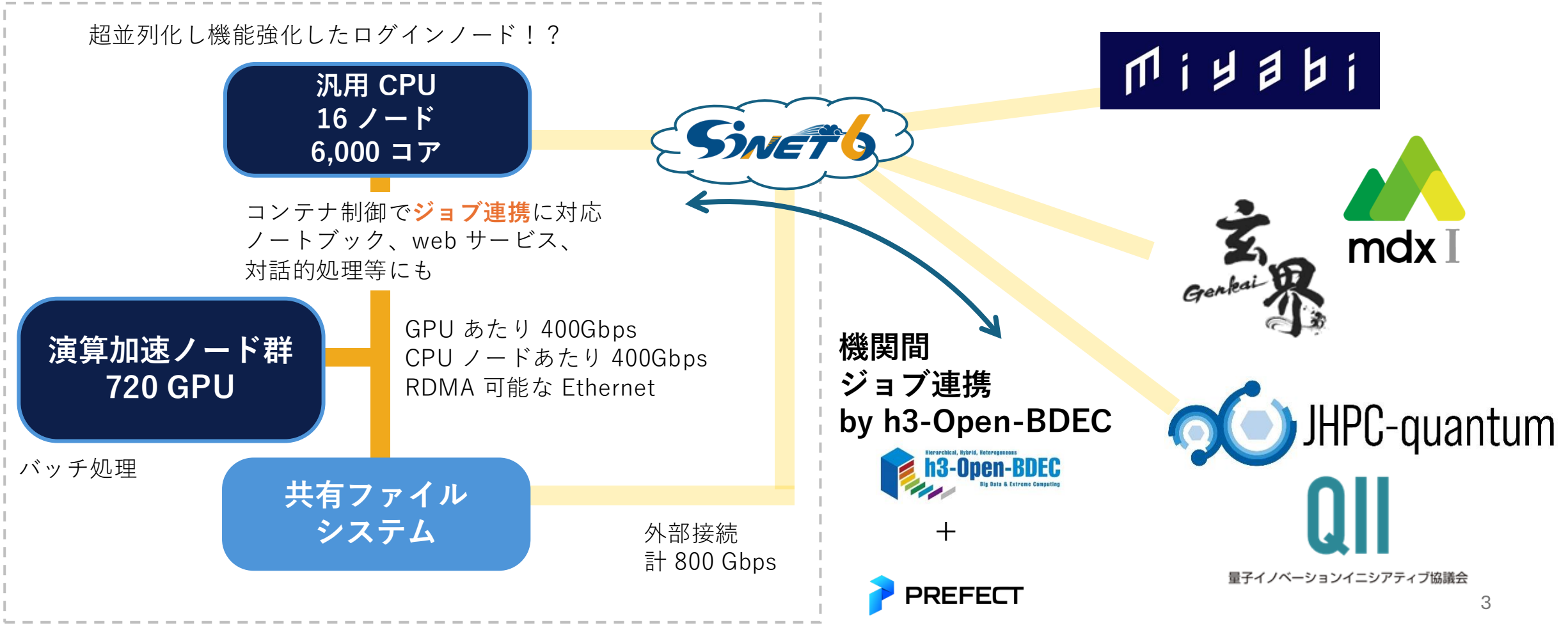
- 「計算・データ・学習・推論」融合基盤システム
  - 補助金額　　36.3 億円
  - 資源提供期間　　**2027年6月**～2033年3月（5年10か月）
  - 提供資源量　　4,976,640 GPU 時間以上（**80% 以上**）
    NVIDIA **Vera-Rubin NVL4** 180 node/720 GPU を予定

  - **HPCI ハブ機能**　　HPCI に資する新たな機能
    - 資源提供機関の間の**ジョブ連携**を実現
      - 本システムに加え、既存 HPCI 資源（Miyabi や他機関の資源）や理研 JHPC-quantum や東大 QII の量子計算機等を連携、ひとつに組織化した計算基盤に
    - Kubernetes + ジョブスケジューラの組合せ
      - バッチ処理に加え、ローカル LLM や AI エージント、対話的処理も
    - AI 法が目指す研究開発の促進、施設の整備・共用、人材育成への貢献

整備する計算基盤

HPC・AI・量子の三位一体で
我が国の GENESIS Mission の
キャッチアップに貢献

- HPCI ハブ機能をそなえた GPU システム



システム構成・必要経費

- GH200 システムを世界的に先行導入した実績をいかし、**Vera-Rubin NVL4 の導入**を目指す
  - GB200 と同等の倍精度性能と圧倒的な AI 性能、confidential computing で AI for Science に貢献
  - すでにベンダーと検討中

- OpenShift コンテナ・仮想化基盤
- RoCE or Ultra Ethernet
- NVMe SSD 2 PB（自己資金）
- 既存の大規模共通ストレージを利用
  - ・Ipomoea-01 (26 PB) をアーカイブとして併用
  - ・HPCI 共用ストレージ東拠点 (100 PB) が近接
  - ・2028年頃を自己資金で SSD 増強予定（SSD価格正常化後）

経費

費目		千円
「計算・データ・学習・推論」融合基盤システム	計算ノード等	3,300,000
	ストレージ	自己資金
管理費(10%)		330,000
合計		3,630,000

汎用 CPU ノード群

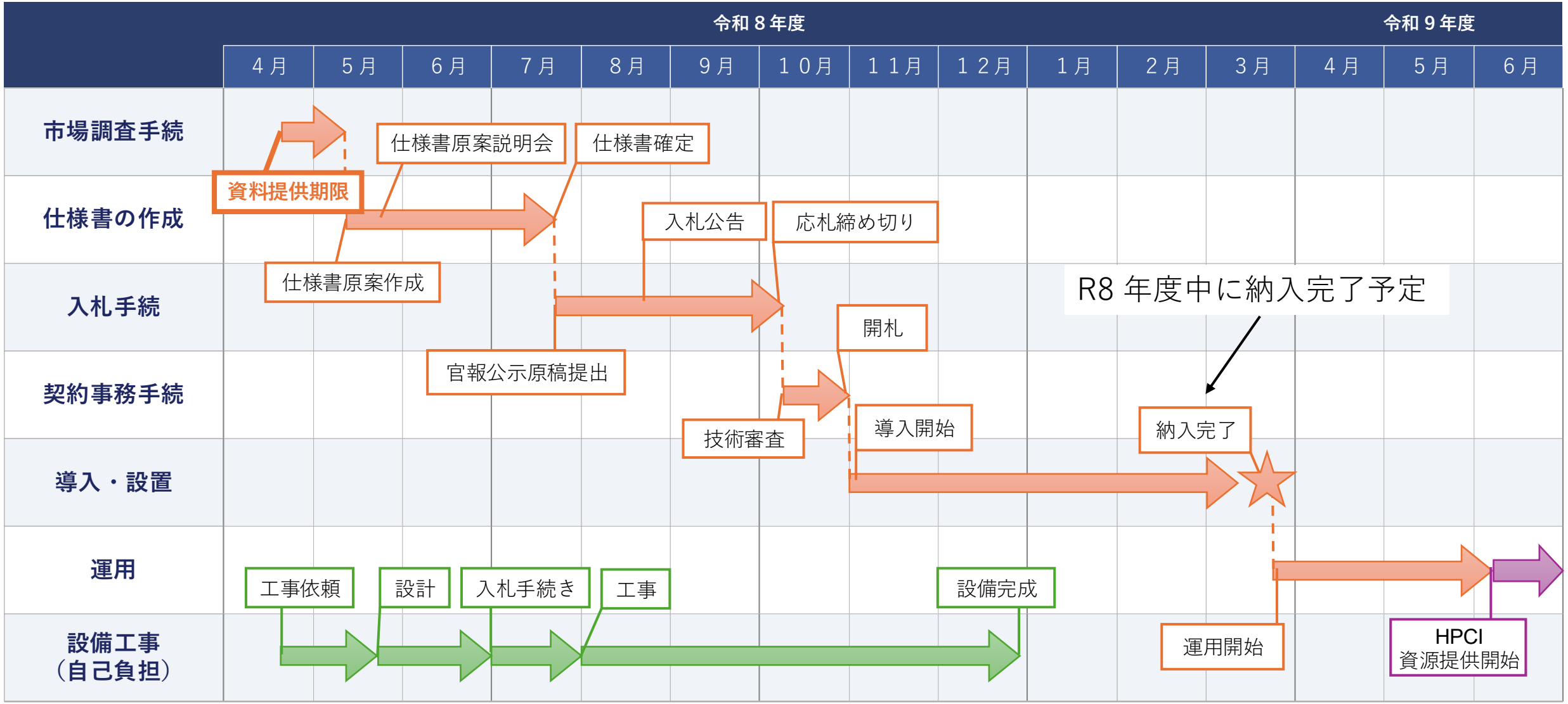
ノード数	16ノード
コア数	約6,000コア
コアあたり倍精度性能	65 GFLOPS以上
コアあたりメモリ	2.0 GB以上、バンド幅 5.0 GB/sec以上
ノード数	16ノード

演算加速 GPU ノード群

項目	Vera-Rubin （主案）	GB200 （代替案）
ノード数 / GPU数	**180ノード / 720 GPU**	240ノード / 960 GPU
倍精度演算性能	140 PFLOPS（エミュレーション）	140 PFLOPS（エミュレーション）
AI演算性能（FP4）	**25 EFLOPS**	9.6 EFLOPS（主案の38%）
メモリ容量	207.3 TByte	165.8 TByte（主案の80%）
メモリバンド幅	15.8 PByte/sec	6.9 PByte/sec（主案の44%）
Confidential Computing	Vera, Rubinともに対応	Graceが非対応なので不可

整備スケジュール

R9 年度内に HPCI に確実に資源提供予定



自己負担の設備工事はすでに設計開始

運用体制・整備環境

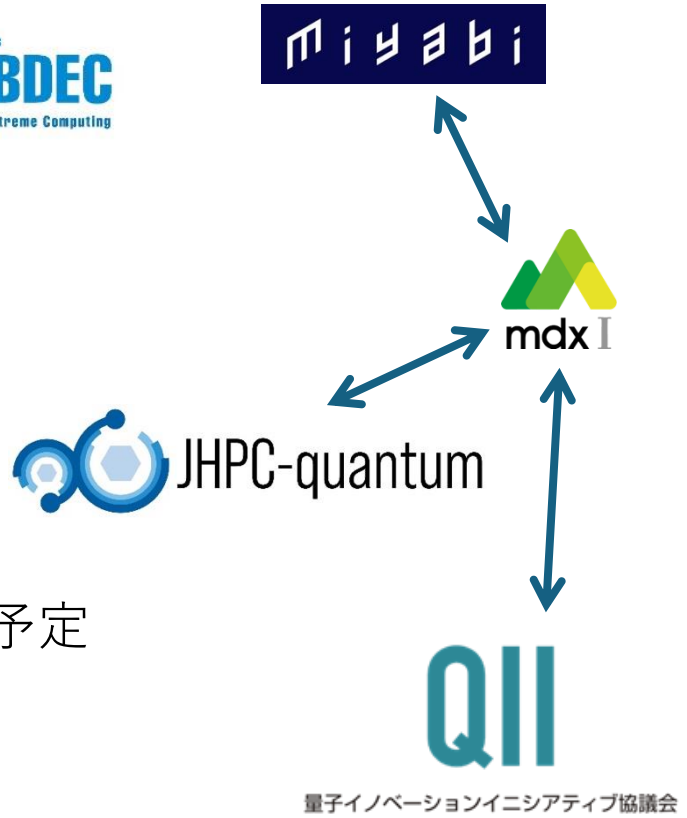
- 利用制度・利用環境
  - 他の採択機関と４大学間の MoU**「AI for Science 推進に関する覚書」**を締結
  - 文部科学省「HPCI 整備計画調査研究」事業の成果を率先して実現予定
    - 当センターの千葉と塙が運用体制と運用技術・セキュリティの研究代表
  - 利用ポイント共通化を他機関と議論中
  - Open OnDemandをユーザポータルとして整備
    - Jupyter Lab 連携機能を強化
- HPCI 利用者拡大
  - 量子・HPC ハイブリッド連携により、最先端の計算基盤を提供
  - Confidential Computing＋テナント分離によりセキュアな環境を提供、HPCI**産業利用**を推進
  - Kubernetes コンテナ制御＋バッチ処理の融合
    - Jupyter Lab 連携等によりバッチ処理に加え**ローカル LLM と AI エージェント**の活用が容易に
    - これまで計算基盤を活用していなかった**人文社会科学等**の研究者を HPCI に誘導

量子・HPC ハイブリッド連携

- H3-Open-BDEC を機能拡張
  - Miyabi、mdx I と量子計算機のジョブ連携
- 経産省・NEDO　JHPC-quantum (2023-2028)
  - 理研、ソフトバンク、東大、阪大
  - IBM（超伝導型）　ibm_kobe　（156+Qubit・神戸）
  - Quantinum （Ion-Trap 型）　Reimei （20+Qubit・和光）

  - **産業応用重視**　2026年7月頃までに連携サービスを本格開始予定
  - ジョブ実行は確認済み、20 社以上の試行課題採択済み
  - 2027 年度 JHPCN 課題で一般に提供を準備中

- 東大・量子 inovation initiative 協議会 (QII)
  - Ibm_kawasaki　（156+Qubit・川崎）



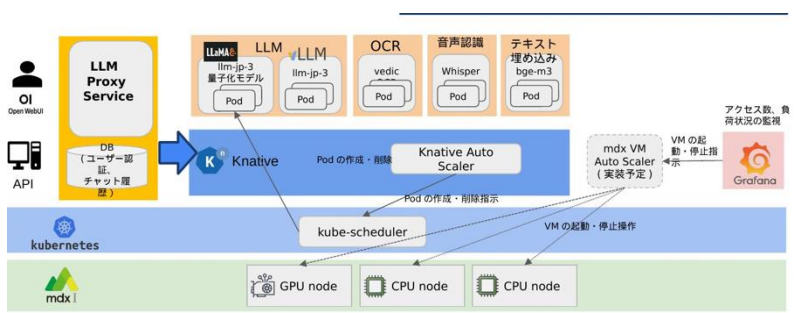
Miyabi は３つの量子計算機に接続する国内唯一のシステム
（QII は 2025年12月より試験サービス開始済み）

mdx-MaaS：オープンモデルの AI 推論サービス

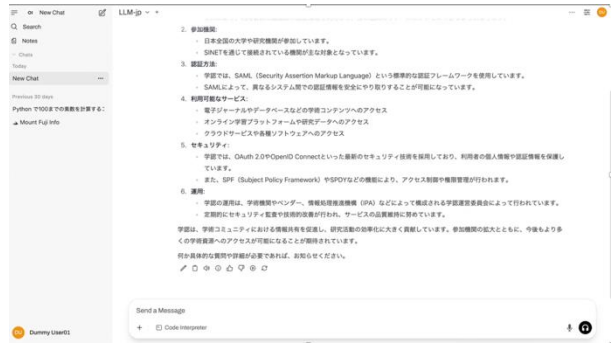
http://sites.google.com/view/mdx-maas/

オープンウェイト の LLM／VLM／OCR／埋込モデル等を、**mdx l 上での推論サービス**として提供。**チャットUI ＋ OpenAI 互換 API**で、環境構築不要・学認認証・履歴保存。

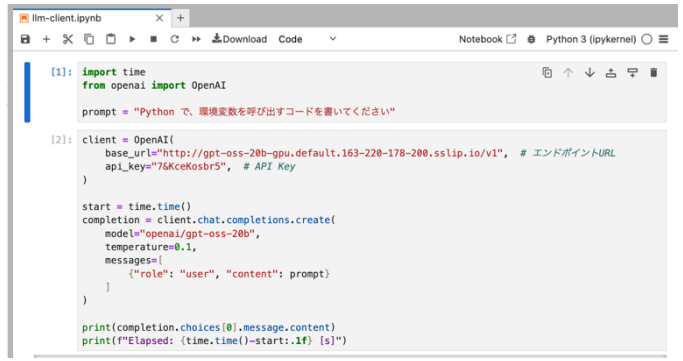
システム構成 / K8s+Knative で自動スケール



チャットUI / Open WebUI＋学認認証



OpenAI 互換 API / Pythonから数行で呼出



提供モデル

SOTAのオープンウェイトモデル・音声認識等も順次追加予定。

LLM

- llm-jp/llm-jp-4-8b-thinking
- llm-jp/llm-jp-3.1-13b-instruct4
- openai/gpt-oss-20b
- google/gemma-4-E4B-it

VLM

- Qwen/Qwen3-VL-8B-Instruct

OCR

- deepseek-ai/DeepSeek-OCR-2

利用ユースケース（実証）

文献スクリーニング
論文の大規模検索・抽出

科学知識の構造化
物性値・条件を表形式抽出

古文書／多言語 OCR
外部送信不可データを処理

★ コミュニティ参加募集
試したいモデル・ユースケースのご要望・改善提案を歓迎。
フィードバックを反映し随時拡張。

申　込　（4ステップ）

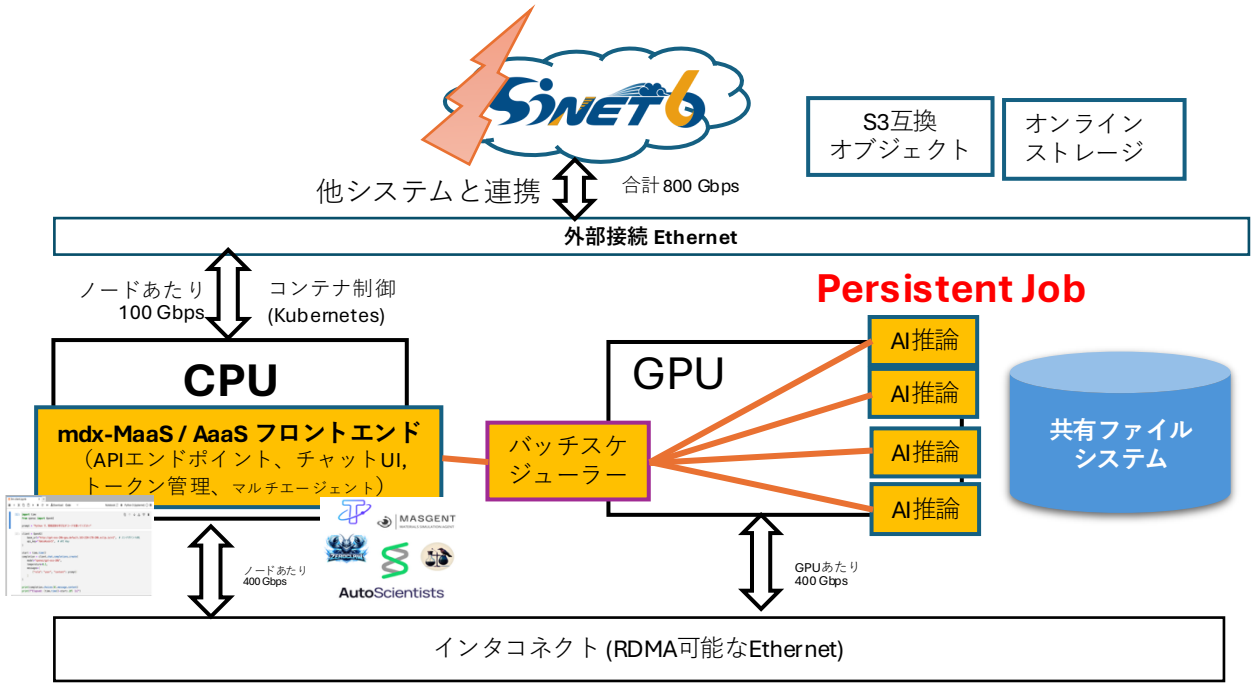
① ポータル **sites.google.com/view/mdx-maas/** で利用申請 → ② 承認後、**学認 ／ mdx ローカルアカウント**でログイン → ③ APIキー発行 → ④ チャットUI/APIですぐ利用開始

基本 無償 提供予定

ただし GPU/CPU 資源が限定的のため、同時実行数・負荷に応じて利用量に制限あり。

AI推論サービス及びエージェント基盤の S/D/L/I基盤での実現構想

- **mdx-MaaSで実現したAI推論サービスを S/D/L/I上に実装予定**
  - AIモデルの推論処理に必要なGPUの動的な確保に関して mdx-MaaSでは Kubernetesで行っていたが、SDLIではジョブスケジューリングシステムのPersistent Jobを使用
  - チャットUI・API・認証/認可・トークン管理など上位機能は 既存実装をそのまま再利用
- **エージェント基盤 (Agent-as-a-Service)もAI推論サービス上に実装する**



既存 mdx-MaaS 実装を再利用　SDLI 向けに新規開発

AaaS 上位層

AaaS（Agent-as-a-Service） — 自律型マルチエージェントによる AI for Science（自律的な研究実行）

- 共通エージェント基盤 ClaudeCode/OpenClaw/OpenCode
- 研究用エージェント AutoResearchClaw / AutoScientists
- ドメイン特化型エージェント Llamp, ChemClaw

mdx-MaaS 共通基盤

チャットUI・API・認証/認可・課金管理（既存実装を再利用）

- チャットUI Open WebUI＋学認認証
- OpenAI 互換 API
- トークン使用量管理
- ユーザー認証・認可

動的スケール

動的リソース制御（オートスケール） リクエスト量に応じて **Persistent Job 数を増減**（従来の K8s+Knative オートスケールに相当する仕組みを Persistent Job で新規実装）

モデルランタイム

AI モデルランタイム 各 Persistent Job 上で **vLLM / llama.cpp** 等を起動し、オープンウェイト LLM／VLM／OCR モデルを推論サービスとして提供

SDLI 基盤

SDLI キューイング基盤（Slurm 等） Kubernetes（従来） → **Persistent Job（本構想）** として推論サービスを常駐

HPCI むけ提供計算資源量・提供期間の見込み

- 4,976,640 GPU 時間（全体の 80% 相当 576 GPU）
  - 残り 20% は独自制度で共用、うち 10% は企業向けを予定
  - HPCI 制度改善でより柔軟な利用が可能になれば**全体の 90%** を提供することも

	CPUノード時間	CPU性能	GPU時間	GPU性能(倍精度)	GPU性能(AI)
2025年度	20,399,040	7.88 PFLOPS	3,179,520	24.3 PFLOPS	675 PFLOPS
2028年度	約 600,000	0.5 PFLOPS程度	**7,879,680**	**135.64 PFLOPS**	**20.6 EFLOPS**
向上比			**約2.5倍**	**約5.6倍**	**約30.6倍**

- 2027年 **6月**から2033年 3月まで（5 年 10ヶ月）
  - 最新・最高性能の **Rubin GPU** を導入し、**いち早く** HPCI に提供
    - 2 ヶ月の準備期間で早期に提供開始
  - 経験上、コンポーネントの一部が保守打ち切りとなるため、6 年以上の運用は好ましくない