資料4-3-2

次世代HPCIに向けた取組について
AI4S計算資源 玄界-D (仮称) 整備計画

HPCI計画推進委員会（第69回）

日時：2026年6月30日(火) 13:00 - 15:00

場所：ハイブリッド開催（文科省東館16F2会議室/ZOOM）※傍聴あり

九州大学情報基盤研究開発センター 美添 一樹

九州大学情報基盤研究開発センター システム概要

Science計算向け玄界システムに
AI計算向け “玄界-D” を接続した
AI for Science 用システム

スーパーコンピュータ玄界（既存資源）

玄界計算ノード

玄界ストレージ

高速ネットワーク

(仮称) 玄界-D 増強GPUノード

ローカルモデル推論

モデル事後学習

推論性能を重視した
GPUを144基以上搭載

**多彩な研究分野
多彩な利用方法に
対応するシステム**

他システム
との連携

常駐型AI活用研究イメージ

増強部分 玄界-Dを主に活用

APIゲートウェイ

OpenAI
互換API

ブラウザ

HTTPS

ユーザ

外部サービス

商用クラウドモデル
(OpenAI, Claude, Gemini...)

API

Pythonコード

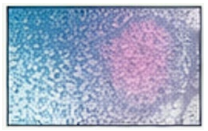
想定する用途

対話型利用、APIで利用
研究データの継続的な解析
Agentic AI による研究遂行

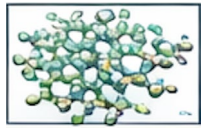
再現性があり**機密データ**も使えるローカルAIモデルと
最高性能を求める商用モデルを容易に併用可能とする

バッチ型AI駆動研究イメージ

既存資源、玄界-D 共に活用



実験データ



シミュレーション

実験ノート

仮説生成

実験データ、シミュレーション、実験ノートの解析
さらにAIモデルの構築や事後学習など

システム概要: (仮称) 玄界 – D

増設AI (推論 + 学習) 環境

既設環境：スーパーコンピュータ玄界

HPC + AI (学習)計算ノードグループ

CPUノードグループ
(FP64 7.4PFLOPS)
- 1024ノード
- 120コア/ノード
- 512GB/ノード

GPUノードグループ
(FP64 5.3PFLOPS)
- 38ノード
- 120コア/ノード
- 512GB /ノード
- 4GPU / ノード

玄界 – D

GPUノード群 (FP4 2880 PFLOPS)
- 144 GPU
  (NVIDIA GB300 NVL72 x 2)
- 72 CPU
- 40 TB GPUメモリ
- 34 TB CPUメモリ

高性能インターコネクト (NVIDIA InfiniBand NDR 200/400 Gbps)

Open OnDemand

Next Cloud

ログイン

大容量HDDストレージ
(55.2 PB)

高速SSDストレージ
(0.7 PB)

AI for Science に対する期待と需要

九大内でアンケートを実施→オールジャパンのニーズを把握

方針 九大内の多様な分野の研究者にリーチ、要望に応えるシステムを構築。
同時にノウハウを普及。フィードバックを全国、さらにグローバルに展開

回答者ほぼ全員が
常駐型AIの需要あり

⇨

既にGPU資源は逼迫中。増強は必要だが、単純な増強ではなく
AI for Science の新たな需要に適したシステムの設計・増強が必要

既に玄界等を活用中の研究テーマ

- 有機材料・デバイス
- 経済 (カーボンニュートラル)
- 経済
- 材料候補検証シミュレーション
- 新奇化合物探索
- 水素・燃料電池
- 農学
- 宇宙・量子 観測解析
- 音響
- 造船
- 医学・生命科学
- 経済

AI4S 利用の需要

- 経済 (カーボンニュートラル) 文書解析・意思決定支援
- デジタルラボ 実験ノート解析
- 新奇化合物探索
- ロボット農場
- 宇宙・量子
- 水素・燃料電池
- 材料: 失敗データ蓄積・再利用
- 海洋・地形
- 有機材料・デバイス
- 音響
- 造船
- 法学
- 医学・生命科学
- 古文書
- 文学・言語学

既に玄界を利用中　非スパコン資源利用中　潜在需要あり

想定する利用形態

バッチ型AI駆動探索ループ

データ準備
- 公開データベースの利用
- 実験データ等を含む非公開のデータベース

モデル学習
- (optional) データベースでAIモデルを訓練

候補探索
- 訓練したモデルを用いて最適化・探索
- 手法は様々（強化学習・ベイズ最適化・MCTS等）

候補評価
- AIモデル、シミュレータなどで候補を評価

データ更新
- 場合によっては評価結果などをデータベースに追加
- 全体のステップを繰り返し

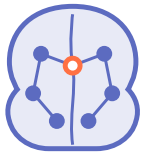
各ステップで、
以下の資源を使い分ける


玄界ストレージ

玄界計算ノード

玄界AIノード

ローカルAIモデル




他システム

クラウドAIモデル

常駐型AI研究支援ループ

研究監視
- 実験・研究ログ監視

継続解析
- 電子実験ノート・データ継続解析
- 24時間または定期的に解析を実行

要約提案
- 履歴要約・報告
- 次の実験を提案

実行連携
- 実験ジョブ実行
- あるいはロボットで自動実験

知識更新
- 実験知識・失敗知識更新
- 研究者へ仮説の提案

横断支援

GPUノード・高速ストレージ・電子実験ノート・既存計算ノード
継続サーベイ・研究者と壁打ち・論文執筆支援

クローズドモデル vs オープンウェイトモデル 最近の状況について

研究データ

- 秘匿性
- 大容量

最高性能では closed model が依然優位だが差は縮小中
reasoning・数学・コーディングでは数か月差以下に肉薄

2024年には“約1年遅れ”という分析があったが、
2025年後半には“数か月差”という分析もある

事後学習等で性能差の補完が可能
例, コーディング、数学、日本語強化モデル

大企業では、Llama/Mistral などのローカル配備可能モデルの採用が相対的に高い
理由は性能そのものより、オンプレ運用・機密保持・コンプライアンス対応だと推測される
用途によって両者を併用している事例が多い

クローズド、オープンウェイト問わず複数モデルを容易に使い分けられるシステムが必要

推論に適した計算機システムの調達・運用が必要
- AI応用と計算機システムの双方に精通した人材
- 継続性のある人員・予算の確保

	クローズドモデル	オープンウェイトモデル
	 ChatGPT Gemini Claude	 Llama DeepSeek Qwen Mistral Swallow LLM-jp
利点	最高性能、導入容易、保守不要	機密データ利用可、継続稼働が容易
	機密データ制約、通信依存	価格予見可能、提供リスク低
	独占・寡占、提供制限・価格リスク	過去モデル保存可能、研究の再現が容易
欠点	モデル更新・挙動変化	性能は数ヶ月遅れ、運用保守必要

AI4S用 LLM常駐サービス: ソフトウェア構成検討

現時点のソフトウェア（調達時には更新の見込み）

レイヤ	ソフトウェア	役割
フロントエンド	Open WebUI	チャットUI, RAG, ファイルアップロード等対応。ユーザ認証も
APIゲートウェイ	LiteLLM Proxy	トークン単位利用量追跡. ガードレール. ローカルとリモートのルーティング
推論エンジン	vLLM (Apptainer)	OpenAI互換API, 高効率KV cache 管理. マルチノードなら Ray を利用
スケジューラ	Slurm	vLLMを通常のバッチジョブとして起動. 専用GPUキューを前提

需要の多い open weight model を選定して常時稼働サービスとして提供
（今後のモデル供給は懸念点）

・1ユーザ占有型よりハードの有効活用
・独自モデルはバッチジョブとして提供

セキュリティ面は重要課題
参考: LiteLLM Proxy への攻撃

他システムとの連携（ローカルサイト、クロスサイト）

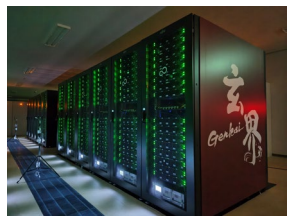
←ローカルサイト
クロスサイト→

Python等

```
from openai impo
client= OpenAI()
resp= client.
chat.completions
.create(..)
```

バッチジョブ

#SBATCH



玄界計算ノード

既に実行中
実現は容易

AI Agent

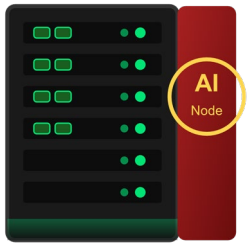
バッチジョブ

#SBATCH

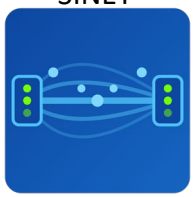
玄界計算ノード

可能だが要検討
特に認証・
セキュリティ

玄界AIノード

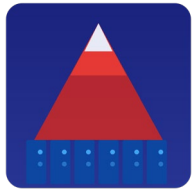


SINET



ホワイトリスト

HPCI



システムA

システムB

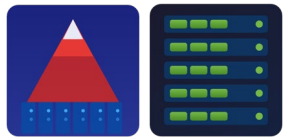
Python等

```
from openai impo
client= OpenAI()
resp= client.
chat.completions
.create(..)
```

バッチジョブ

#SBATCH

ユーザ認証



他システム

他組織と連携して
実験が必要
技術的には可能

AI Agent

バッチジョブ

#SBATCH

ユーザ認証

他システム

責任帰属など
制度面から要検討
限定システム間の
実験から実施

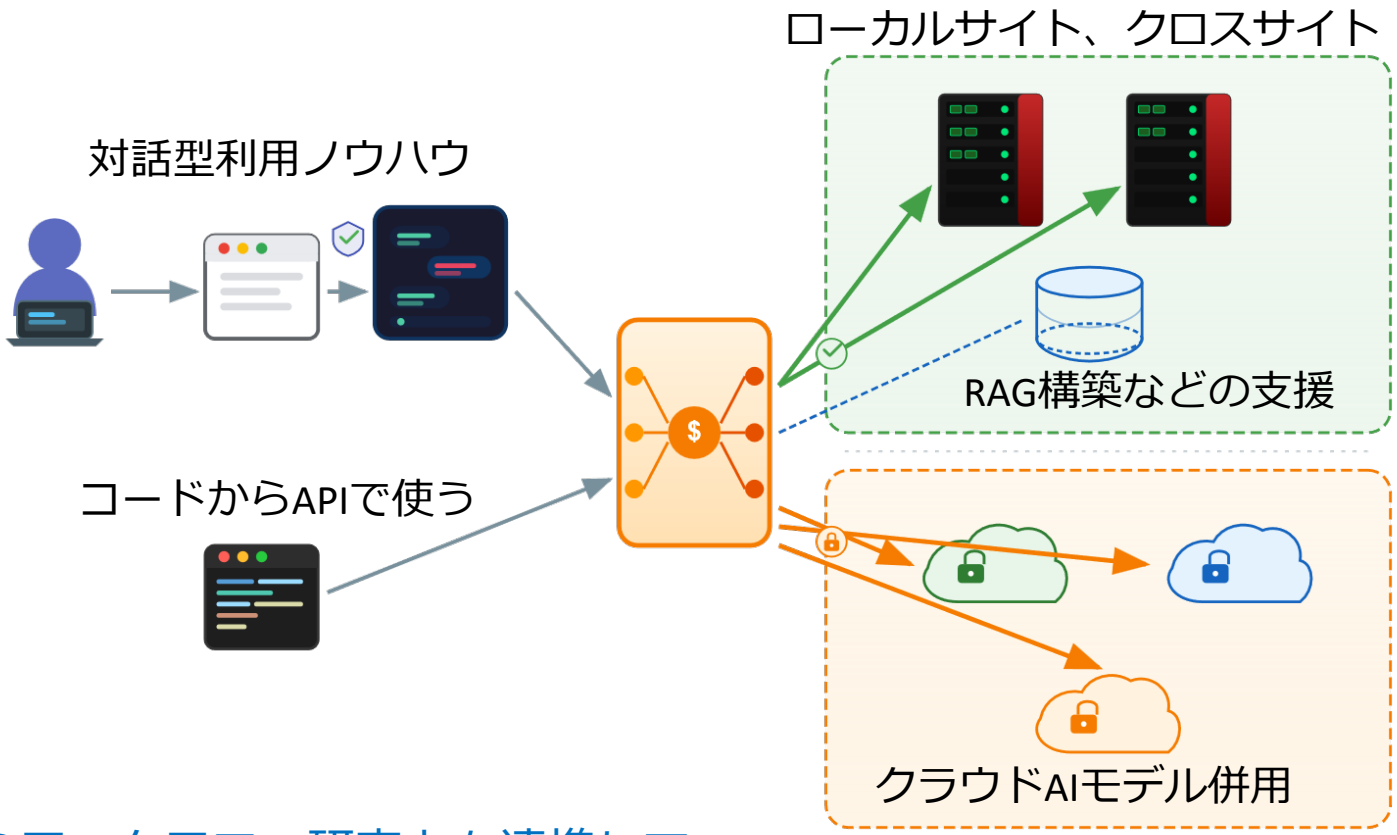
HPCI制度設計について

- LLM常駐サービスに対応したシステム・運用ルールが必要
  - 既存のHPCIの運用は常駐AIサービスと親和性が低い
  - 対応するソフトウェアシステムも存在しない（検討は先ほどのスライド）
- AI分野の事情に合わせた迅速・柔軟な利用申請が必須
  - （審査期間が1週間に短縮されるそうで、非常に良い）
  - 加えてトライアルユースも必要（事前に性能測定などを要求しない）
- 柔軟な運用ルールが必要
  - ローカル open weight model と商用クラウドモデルの併用も必要
- 九州大学では先導して新たな利用スタイルへ対応
  - LLM関連ソフトウェアの整備
  - 講習会などを通じて利用者の獲得
    - そのために資源の30%の利用を希望
  - ノウハウと獲得ユーザをHPCIにフィードバック
    - ノウハウ確立後はさらに多くの割合をHPCIに提供する用意あり

新しい使い方の普及について

ユーザ向け

- AI Jam Q
  - 本家AI JamやJAPAN SCIENTIST AI JAMを倣い（主に）九大におけるAI活用を後押し
- GPUミニキャンプ、講習会など
  - 既存の講習会イベントを強化



システム側を含めて検討

- HPCI整備計画調査研究との連携
  - 運用技術・セキュリティ実証研究のワークフロー研究とも連携して玄界AIノード活用技術を研究、当研究を通じて他機関とも連携
- HAIRDESC（次世代HPC・AI研究開発支援センター）との連携
  - 複数のシステム間を統一的なインターフェースで利用可能とする

注, 高セキュリティデータについて

遺伝データなど、高いセキュリティに対応したストレージは計画に含まれていない。
今後、医学系ユーザとの検討が必要

「玄界-D」の制約（現有体制・納入時期）を一旦外し、特定の調達を離れて近未来に求められるシステムについて検討

AI for Science のための近未来システム像（将来構想）

なぜ AI for Science にローカルモデルが必要か

機密性
実験・医療・ゲノム等
外部に出せないデータを扱える

再現性
商用モデルは更新で挙動が変化
過去モデルを保存し研究を再現

継続性・予見性
価格・提供リスクを回避
24時間の常駐運用が可能

ワークロードの性質が変化 ⇒ 計算資源を管理する基盤ソフトウェアも変わるべき

HPC的バッチ型　資源を Slurm が配分

- run-to-completion（実行完結型）
- 密結合 MPI・FP64／全系を使う大規模ジョブ
- fair-share・占有・アカウンティング
- **計算資源を無駄にしないならこちら**

エージェントAI常駐型　資源を Kubernetes が管理

- 常駐・対話・長寿命のサービス
- 低レイテンシ／バースト・マルチテナント・冗長
- 常駐 エージェントAIを運用しやすい
- **リクエストに応えるサービス提供ならこちら**

どちらも有望 → 複数路線の検討が妥当（多様性は重要） Slurm + K8s の統合方式は複数モデル（Under / Over / Adjacent / Distant＋静的分割）があり、本家 SchedMDも「決定版なし」“no silver bullet” と認めるところ。

- ✓ コンテナ基盤は K8s が事実上の標準（他の選択肢がマイナーに）
- ✓ Slurm 本家は NVIDIA 傘下に / Slinky（K8s基盤の上にSlurm）
- ✓ 密結合MPI、大規模計算は AI・HPC共に Slurm を維持すべき
- ✓ Slurm on K8s もRDMA構成ならほぼベアメタル同等性能（小規模）

エージェントAI基盤の検討・今後の課題

異種ハード混合型 推論基盤（GPU＋推論加速機）

プレフィル
計算律速 → GPU

KVキャッシュ転送

デコード
帯域律速 → 推論加速機 (LPU等)

フェーズごとに最適なハードウェアを割り当てる
特定ベンダー限定のアイデアではなくおそらく今後の標準

セキュリティ／プライバシ（設計段階から専門家の参画が必要）

従来の対策は前提として、その上に *LLM*/エージェント特有の新しい攻撃への対策：

- 間接 prompt injection（RAG・ツール出力・ファイル経由の攻撃）
- エージェントの権限・ツール実行（自律的に行動する主体）
- 利用者間の情報漏洩：KVキャッシュ・GPUメモリ・Proxy の集約ログ

prompt injection 等に精通した人材が必要 → 組織間連携でノウハウを共有

さらなる電力密度への対応

≈600 kW/ラック

Rubin Ultra「Kyber」NVL576・2027年
(GTC2025 NVIDIA 公表値)

40kW(Hopper) → **130kW**(Blackwell)
→ **600kW**（2016年ごろは20kWでも高密度）

800V DC に対応した設備が必要。当然、全液冷
立地・電力供給を考慮して分散立地が望ましい

（私見では）未解決の課題

- 高性能な open weight model の継続的な供給
  （モデル主権・国産モデルの維持に直結）
- ストレージ／データ基盤の設計
  （医学研究などに求められる要件の対応）

組織間連携への期待 それぞれの路線を担う組織が、運用・セキュリティのノウハウを HPCI へ還流。先行している HAIRDESC の枠組みの活用があり得る（国内の複数センターでK8s運用が開始されており、ノウハウの共有に期待）