

「理究」のアーリーアクセスにおける活用事例

第70回HPCI計画推進委員会

理化学研究所 計算科学研究センター (R-CCS)
運用技術部門 先端運用技術ユニット 山本啓二

2026年9月29日

「理研(RIKYU)」 概要

2026 年 3 月に理化学研究所 計算科学研究センターに導入された、AI for Science のための新しいスーパーコンピュータ

- 世界最高水準の AI 性能と「富岳」等との協調で科学研究を加速
- NVIDIA Grace Blackwell を搭載した計算ノード **400 台**(B200、**1,600 GPU**)
- FP64 演算性能: **64.160 PFLOPS**
- FP8 演算性能: **15.539 EFLOPS** ※富岳の約 **7.23 倍** ⇒ AI 学習・推論に強み

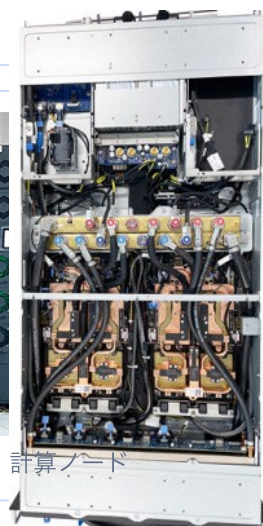
ストレージ: NVMe の Lustre 1.9 PB(ホーム・共有)+ HDD の Lustre 10 PB(課題データ) / ネットワーク: InfiniBand XDR 800 Gbps × 4 レール

冷却: 温水冷却(供給 32 °C / 還り 42 °C)、pPUE 1.03(2026 年 8 月)

2026 年 3 月納入 → 5 月アーリーアクセス開始 → 10 月本格運用開始予定



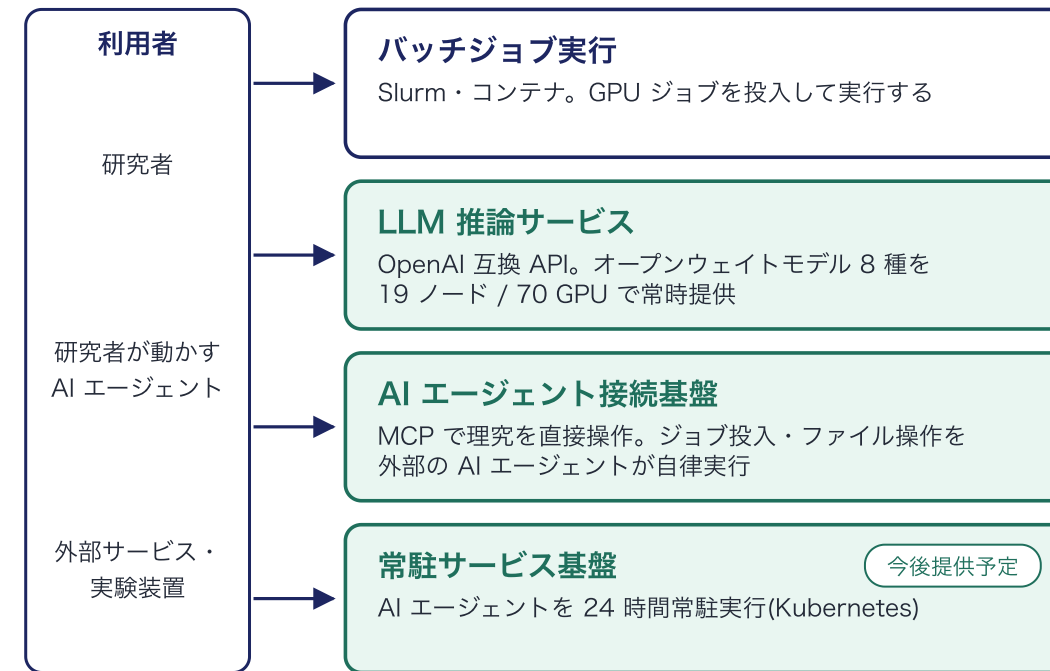
計算ノードラック 20 本 + ネットワークラック 2 本。1 ラック 20 ノード



計算ノード

HPC と AI を 1 つの基盤で

シミュレーションと AI の学習・推論を、**同じマシン・同じアカウント・同じデータ**の上で行き来する



従来の HPC センターとの違い — バッチジョブを動かすだけの計算機ではない

バッチジョブ実行に加えて、OpenAI 互換の **LLM 推論サービス**、外部の AI エージェントがつながる **AI エージェント接続基盤(MCP)**、AI エージェントを 24 時間動かす **常駐サービス基盤**(今後提供予定)を提供する。4 つは並列の関係で、同じ計算機・同じアカウント・同じデータの上に載る。

バッチジョブの利用状況

第 1 期(2026/5/12~7/6)と、継続中の第 2 期(7/7~)。同一のハードウェア・同一の定義で集計。第 2 期の数値は 2026 年 9 月 19 日までの集計

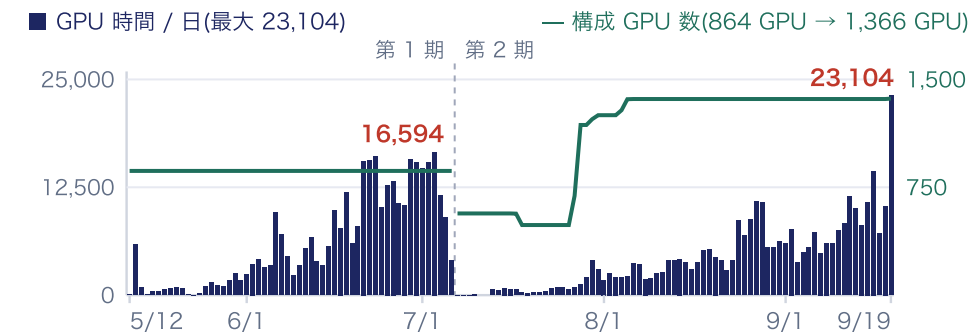
	第 1 期(5/12~7/6)	第 2 期(7/7~、継続中)
提供規模	216 ノード / 864 GPU	342 ノード / 1,366 GPU
GPU 利用時間	340,397 時間	321,280 時間
ジョブ実行者(累計)	101 人	157 人
登録ユーザ(累計)	—(課題制の前)	221 人(課題 76 件)
日別アクティブ(最大)	45 人	63 人(8/27)

10 月以降の需要見通し — 日あたりの GPU 時間



- ・理研 AGIS 課題: 28 課題で 10 月~翌 3 月に 約 230 万 GPU 時間の利用計画
- ・ARISE 事業: 10 月以降の利用を想定し、ある程度の利用枠を確保済み
- ・SPReAD1000: 第 1 回公募の 5 課題が利用中。第 2 回公募でも同程度の課題数を見込む(資源量としては上記 2 つに比べ小規模)

日別の GPU 利用時間



日別アクティブユーザと、ジョブ実行者の累計



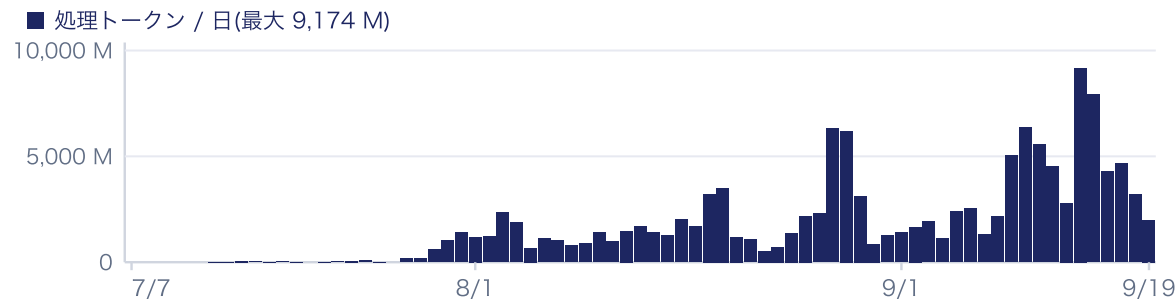
理研 アーリーアクセス 第 1 期(EA1)

- ・本格運用開始に先立つ**試験利用**。限られた利用者にさまざまなユースケースで試験的に利用してもらい、性能面・運用面の問題を洗い出して**本運用サービスの高度化**を図った
- ・2026 年 5~7 月の 56 日間にわたり、**最大 216 ノード(864 GPU)**をアーリーアクセス第 1 期として供用
- ・**101 人の利用者だけで 864 GPU が飽和し、30 時間を超える待ち行列が発生**(実際には同時アクティブユーザ 40 人で既に飽和)
- ・第 2 期は 1,366 GPU に増強して待ち行列を解消。一方で **9/19 には 23,104 GPU 時間**に達し、第 1 期の最大日 16,594 GPU 時間(占有率 80 %)を上回った

RIKEN R-CCS LLM 推論サービスの利用状況

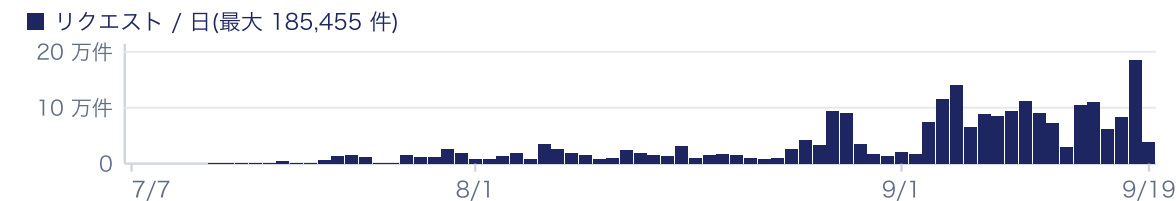
2026 年 7 月 7 日にサービス開始。第 1 期では提供していない。数値は 2026 年 9 月 19 日までの集計

日別の処理トークン(入力) — 第 2 期(7/7～)



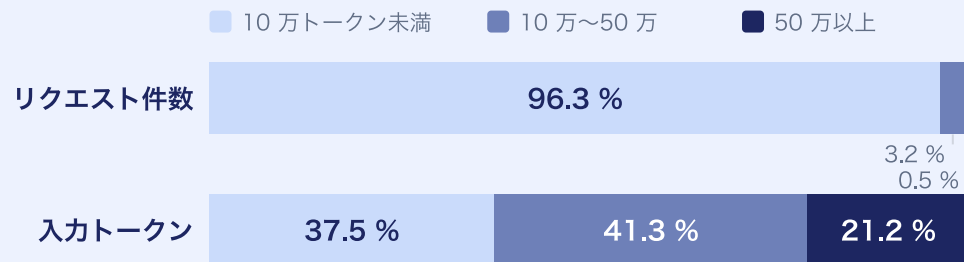
月別: 7 月(7/7～31)3,796 M → 8 月 56,941 M → 9 月(1～19 日)70,055 M。第 2 期の累計 130,792 M トークン

日別のリクエスト数 — 第 2 期(7/7～)



9/13～14 は特定の 1 利用者が 1 モデルへ短い入力の呼び出しを大量に行っており(2 日で 305 万件、期間全体の 56 %)、グラフから除いている。入力トークンでは期間の 2.4 %

少数の長いリクエストが、大半のトークンを占める



2026 年 9 月 1～19 日。長い前置きを保ったまま繰り返し呼ぶ、エージェント型の利用を示す

使われ方は「チャット」ではなく「エージェント」

- ・平均文脈長 **80 K トークン**(2026 年 8 月、中央値 37 K)
- ・同じ長い前置きを繰り返す利用が支配的(キャッシュ供給率 93 %)
- ・人手のチャットではなく、**AI エージェントが自動で回している形**

バッチジョブサービスを利用せず、LLM 推論サービスだけを利用するユーザがいる

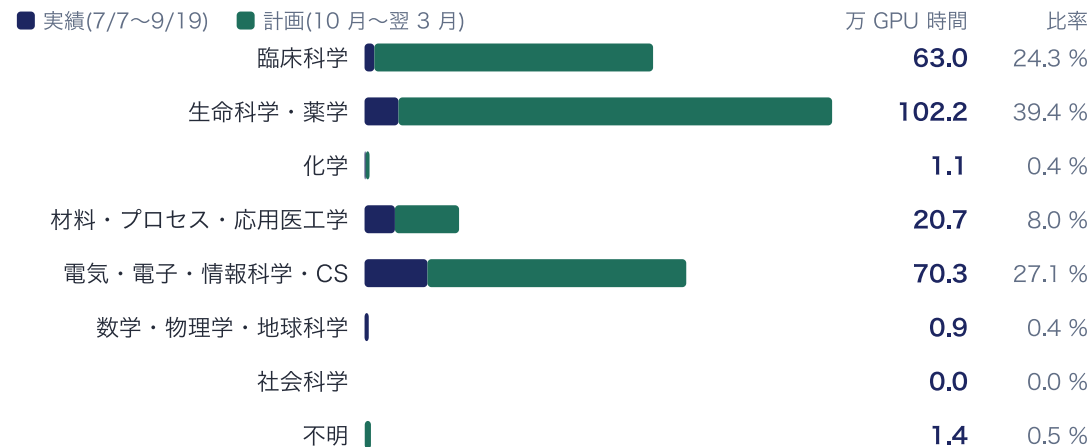
第 2 期に LLM 推論サービスを使った利用者は **68 人**。そのうち **11 人** は一度もジョブを投入していない(バッチジョブサービスの利用者は 157 人、両方使うのは 57 人)。課題単位でも **4 課題**が計算ジョブを 1 件も投入せず LLM 推論サービスだけを使っている。

GPU 時間 5 時間未満で **1 万 M トークン・67 万リクエスト**を使う課題もあり、計算量では見えない利用がある。

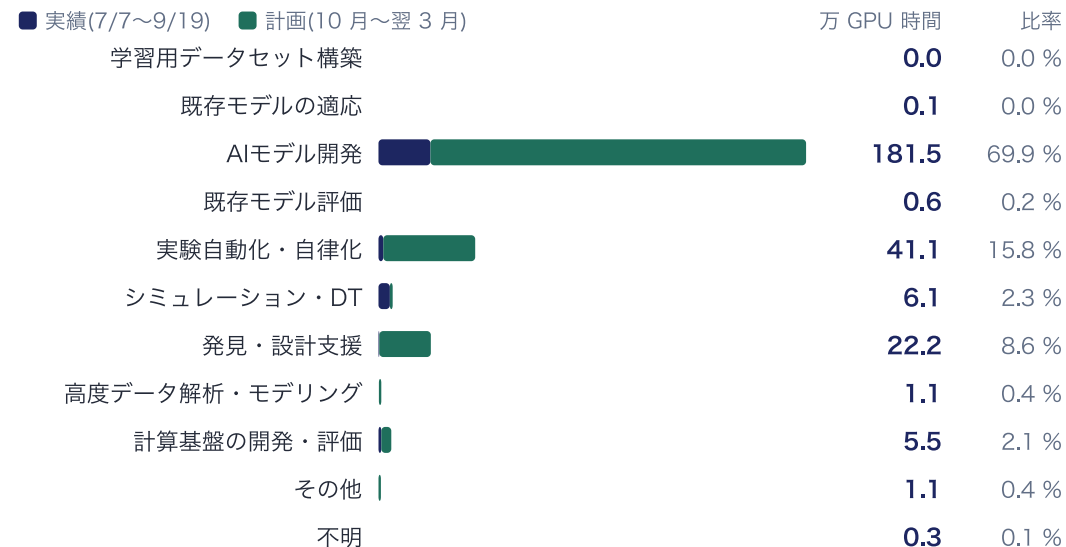
分野別・ユースケース別の課題の内訳

実績(2026/7/7～9/19、課題 75 件)と、理研 AGIS 課題 28 件の利用計画(10 月～翌 3 月、課題申請に記載された約 230 万 GPU 時間)を重ねた。区分は SPReAD1000 に準拠

分野別(研究領域)の GPU 時間



ユースケース別の GPU 時間



主な傾向

- ・実績では **情報科学が GPU 時間の 44 %** を占めるが、今後の計画では **医学・生命系(生命科学・薬学 + 臨床科学)が 68 %** に上がる。基盤モデルの本格的な学習がこれから始まる
- ・ユースケースでは **AI モデル開発が実績・計画とも 7 割**。一方で **実験自動化・自律化は 6.6 % → 17.1 %** と大きく伸び、AI が研究そのものを実行する使い方が計画段階から織り込まれている
- ・計画を立てた課題側の申告でも、**推論・エージェントが 21.9 % (約 50 万 GPU 時間)**、学習が 73.2 %。LLM 推論サービスへの需要は一過性ではない
- ・計画の値は**課題申請に記載されたもの**のみで、申請外の利用は含まない
- ・実績では **16 課題(21 %)**がバッチジョブも LLM 推論サービスも未使用で、需要はまだ顕在化しきっていない

次世代半導体パッケージング用加工技術の確立 — ガラス基板への高アスペクト比穴加工をシミュレーションで最適化する

1. 課題

- 次世代半導体パッケージング材料としてガラス基板が有望視されているが、基板の両面を接続する配線用の高アスペクト比穴加工が課題
- 理化学研究所が開発したフェムト秒レーザー加工技術により超高速で高品質・高アスペクト比なガラス穴あけが可能に
プレスリリース「世界最高速のガラス微細貫通穴加工技術を開発—次世代半導体デバイス製造のキーテクノロジー—」
(理化学研究所光量子工学研究センターと株式会社エンプラスとの共同研究)
https://www.riken.jp/press/2026/20260610_1/
- レーザー技術の進展によりレーザー加工の制御パラメーター数が大きく増加し、パラメーター空間が指数関数的に発散
- シミュレーションによる最適化が望まれていたが、マルチスケールにまたがる加工現象の定量化が困難であった

2. 研究内容

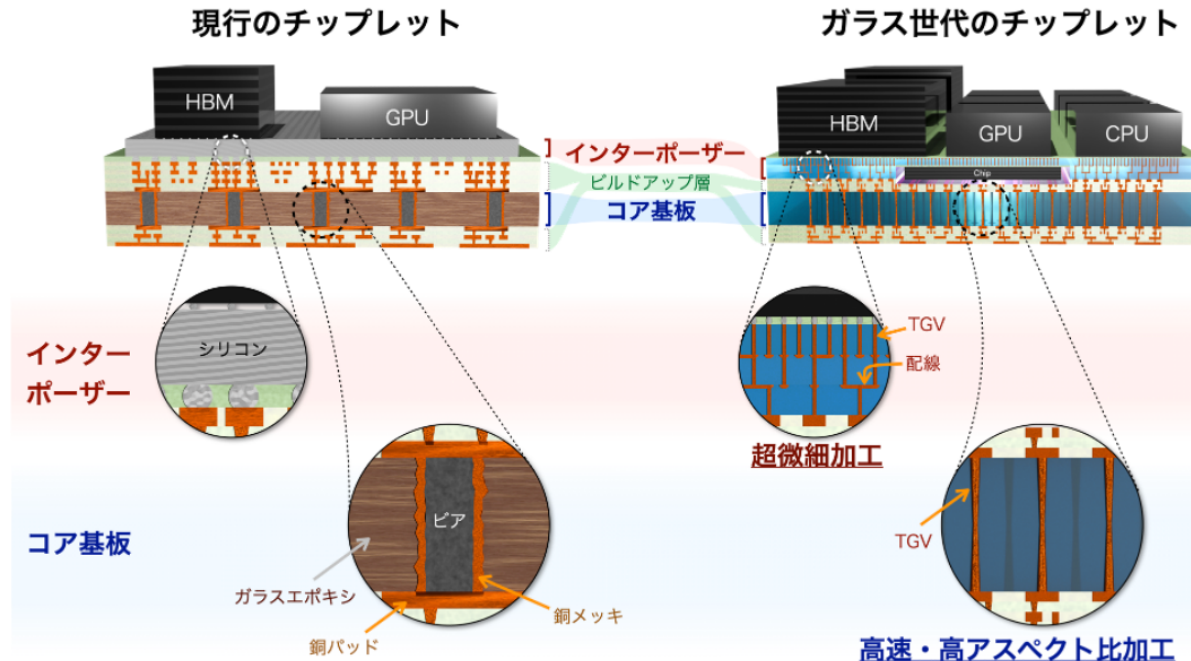
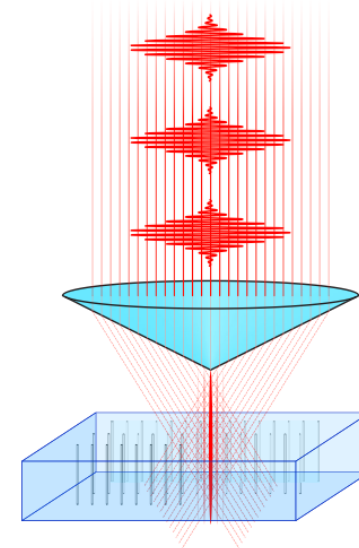
- 実験とシミュレーションを連携させるシミュレーターの構築
- プロセス基盤モデル開発のための学習データ生成

3. 計算資源量

- 250台のGPUを24時間 × 3回占有して大規模シミュレーションを実施 → 18,000 GPU時間

4. 将来展望

- 実験に裏付けされたシミュレーターに基づき、軽量・高速にパラメーター最適化が可能な材料プロセス基盤モデルの構築を行う
- 広範囲に適用可能なシミュレーターのパラメーター調整に100,000 GPU時間、深層学習モデル学習に30,000 GPU時間を想定



その他の利用事例の概要

第 2 期(2026/7/7～9/19)の消費が大きい課題から、分野の広がりが見えるように 8 件を選んだ

課題	分野	内容と、投入されたジョブの形	ジョブ数	GPU 時間
材料科学の基盤モデル開発	材料・プロセス・ 応用医工学	第一原理計算・分子動力学・自動実験・学術文献を連携させ、材料科学の基盤モデルと自律的なデータ生成基盤を開発。1～32 GPU の短めのジョブを多数投入する形	17,861	45,741
シミュレーションを加速する サロゲートモデルの開発	電気・電子・情報 科学・CS	シミュレーション出力・実験計測・観測データから学習するサロゲートモデルを開発。最大 16 GPU、1 ジョブ平均 21 GPU 時間の学習ジョブを継続的に実行	1,851	38,563
創薬・生命科学の マルチモーダル基盤モデル開発	生命科学・薬学	化合物・タンパク質・RNA・ゲノム・医療関連知識を統合した分子基盤モデルを開発。4 GPU の学習ジョブを 1 ジョブ平均 29 GPU 時間と長時間まわす	1,123	32,229
疾患ゲノムの解析に向けた 言語モデル開発	臨床科学	多数の生物種とヒト集団のゲノムを学習した言語モデルを構築し、既存手法では捉えにくい病因候補の変異を絞り込む。4 GPU のジョブを 9 月に集中投入	1,580	20,426
大規模言語モデルの 新しいアーキテクチャの研究	電気・電子・情報 科学・CS	独自アーキテクチャの事前学習を行い、訓練データ規模の拡大に伴う能力の発現を検証。最大 16 GPU・1 ジョブ平均 38 GPU 時間の長時間学習を 5 日間に集中	485	18,642
ロボット基盤モデルの学習	電気・電子・情報 科学・CS	実世界のロボットを自然言語や画像で制御するため、オフライン強化学習で制御モデルを学習。4 GPU の学習ジョブを 1 ジョブ平均 24 GPU 時間で継続	612	14,890
研究を自動化する AI エージェント基盤の開発	生命科学・薬学	複数のエージェントと MCP サーバ群を統合し、仮説生成から実験・解析・知識更新までを支援する基盤を構築。短時間のジョブを多数投入	3,993	13,048
次世代計算基盤に向けた ソフトウェアの GPU 移植・最適化	電気・電子・情報 科学・CS	主要なシミュレーションソフトウェアを GPU 環境へ移植・最適化。短時間の性能評価ジョブを多数投入し、単一ジョブでは 最大 768 GPU のスケーリング試験も実施	6,418	7,383

大規模な基盤モデルの事前学習が資源消費の中心だが、**研究そのものを AI エージェントに実行させる課題**や、**次世代計算基盤のソフトウェアを準備する課題**も立ち上がっている。ジョブの形は課題によって大きく異なり、4 GPU を長時間まわすものから、短時間のジョブを数千件投入するものまである。