生成AIを用いたHPCコードのGPU化を支援の中核に

資料3-2-2
HPCI計画推進委員会
（第70回）令和8年9月29日

2025年度HPCI調査検討WG提言 | HPCI計画推進委員会　2026年9月29日　横田 理央

背景：科学コードの継続性

・長年蓄積したFortran/Cコードは各研究分野の財産であり、GPU移行でも科学的な意味を維持する必要がある

・少人数の開発では、コード理解・移植・性能調整・引継ぎを同じ人が担い、支援を受ける準備自体も負担となる

・テスト不足は大きな変更の正当性確認を難しくし、AI生成コードでも誤りを見逃す原因になる

・GPU化は手段であり、診断結果に応じて全面GPU化、部分GPU化、CPUのままの高度化を選ぶ

重点：保守用コードと生成物

・生成AIを用いて複雑なコードをリファクタリングし、開発者が理解・保守できる簡潔なコードへ整理する

・簡潔なコードからGPUコードを自動生成する「コンパイラのような」使い方を検討する

・可読性と継続開発を保ちながら、対象機種や最適化方針の変更に対応することを狙う

・自動変換は提言上の可能性であり、任意のHPCコードで完全自動化を保証するものではない

支援：HAIRDESCとの連携

・診断・相談：計算時間、依存関係、データ転送、移植工数を整理し、対象と進め方を選ぶ

・人材育成：段階別研修、ハンズオン、AIによる学習支援と専門家の助言を組み合わせる

・検証環境：共用GPUテストベッドで性能・スケーラビリティを確認する

・知識循環：成功例だけでなく期待性能が得られなかった事例とサンプルも共有し、教材・相談へ戻す

提言の柱	数値・期限の例（提言／KPI上の目標）	生成AIによるGPU化との接点
人材・アプリ支援	**診断2027年度開始／GPU対応2028年50%**	移植だけでなく、診断・検証・保守を継続支援
システム・運用	**チェックリスト2027年度／共通基盤2030年度**	再現可能な検証環境と移行期の利用継続
分野拡大と評価	**共同利用枠2027年度／相談100件以上/年**	AIとHPCの知見を共有し、成果をKPIで追跡

生成AIによるGPU化：簡素化 → 自動生成 → 検証 → 保守

§4.3.1の提案を主図に整理｜単発のコード生成ではなく、科学コードの継続開発を支える

① 既存の科学コード

Fortran/C等の研究資産
仕様・入力・参照結果
依存関係と性能の把握

② 簡潔な保守用コード

生成AIでリファクタリング
可読性・継続開発を重視
科学的意味を開発者が確認

③ GPUコードを生成

簡潔なコードを起点に変換
対象機種に合わせて最適化
必要に応じて再生成

④ 実行・検証・改善

正当性と性能を確認
人が変更を採否判断
知見をコード・事例へ還元

保守の中心＝人が理解できるコード　／　GPU生成物＝検証して利用　／　失敗・性能不足は診断へ戻す

GPU化の対象と方針を診断

原文の提案

・重いカーネル、ループ依存、転送やMPI通信の負担を調べ、移植方針と工数を助言する

・全面GPU化／部分アクセラレーション／CPU高度化を比較し、無理なGPU化を避ける

・疎行列、FFT、N体、格子法等のパターン別ガイドとサンプルを整備する

・OpenACC、OpenMP offload、CUDA、HIPや既存GPUライブラリの活用を支援する

採用条件を明示して検証する

発表用の具体化案（数値基準は未設定）

・正しさ：参照CPU結果、許容誤差、保存量、境界条件、代表入力での回帰を確認する

・性能：カーネルだけでなく通信・転送を含むアプリ全体を同条件で比較する

・再現性：入力、変更差分、生成条件、コンパイラ、GPUと計測条件を記録する

・保守性：開発者が変更の意図を説明でき、再生成時にもテストできる形で残す

「生成件数」から「利用成果」へ

既存KPIとの接続と補助指標案

・原文KPI：GPU対応率、相談解決率、公開成功事例、研修後のスキル活用率

・補助指標案：検証合格率、移植に要した人の工数、再生成後の回帰成功率

・補助指標案：性能向上と継続保守に要した負担を合わせて確認する

・追加指標には新たな数値目標を置かず、最初の支援事例で測定方法を固める

第4章の提言枠：GPU化を支えるサービスと到達目標

枠内の全5項目を抽出｜人材・診断・実例をセットで整備し、生成AI活用を現場に定着させる

枠内の提言項目	期限・目標値（原文）	実行内容・生成AI活用との接続
オンライン教材ポータル・相談窓口	**2026年度に開設**	入門・FAQ・質問対応を用意し、コードの理解やGPU化の入口を支える
コードGPU化診断サービス	**2027年度に運用開始**	AI変換に着手する前に、GPU化の対象、難所、効果と工数を診断する
GPU化成功・失敗事例集・サンプル集	**2026年度 5件** **2027年度 20件**	変換例と性能不足の原因を蓄積し、研修・生成・検証で再利用する
GPU化可能な主要アプリのGPU対応	**2028年度 50%** **2030年度 100%**	母集団は原文で「GPU化可能な主要アプリ」と記載されている
GPUに精通した人材の育成	**2030年までに100名**	自分の科学コードを理解し、生成物の妥当性と性能を判断できる人材へ

数値の扱い：対象の違いを踏まえて読む

§7.3は「HPCIで主要と位置づけるアプリ」の全面・部分GPU対応を2030年80%以上とする

GPU化可能性の判定、対象リスト、部分対応の基準が統一済みとは確認できないため、100%と80%以上の両方を残す

支援の運営：HAIRDESCと連携

段階別研修・ハンズオン・教材、診断・専門家相談、GPUテストベッド、成功・失敗事例の共有を接続する

コードの正当性と保守性を支援に含め、少人数開発の負担を軽減するという原文の方向を具体化する

第5・6章の提言枠：GPU化を支える基盤とコミュニティ

両章の枠内9項目を全て抽出｜移行環境と知見の共有を整え、生成AI活用を支援サービスへつなぐ

第5章：システム・運用	期限・頻度
富岳NEXT技術ワークショップ	**2026年度から年1回開催**
富岳NEXT移行チェックリスト	**2027年度までに提供**
富岳NEXT試作機	**2029年度に提供開始**
統一的UI/UX・統合認証基盤・データ共有基盤	**2030年度までに実現**

第6章：コミュニティ・分野拡大	期限・頻度
HPC+AI分野横断ワークショップ	**2026年度から年1回**
AIモデル・データの共有基盤	**2026年度から提供開始**
産業利用ワンストップ窓口	**2026年度に設置**
HPC+AI共同利用プロジェクト枠	**2027年度に新設**
量子ハイブリッド計算パイロット	**2028年度に開始**

移行を止めない運用へ（第5章本文）

・技術情報の早期提供とチェックリストで、現行コードの対応状況・移植方針を見通せるようにする

・端境期は代替の計算資源と利用支援を含め、研究の継続性に配慮する

・システム間のUI/UX、認証、データ共有を連携し、異なる基盤での利用負担を下げる

・セキュリティとデータ管理を両立する環境で、診断・生成コードの試行・結果確認を支える

知識を循環させ、利用の入口を広げる（第6章）

・分野横断ワークショップと共同利用で、科学コードの課題とAI活用の知見を交換する

・モデル・データの共有を、AIを駆使できるHPC分野の人材育成と実践につなぐ

・産業利用のワンストップ相談、実証、共用プラットフォームにより導入の障壁を下げる

・量子との連携は発展的施策として位置づけ、既存分野の継続性と新規分野の拡大を両立する

補正予算が大学等に求めた要件：共用・連携・運転・制度・支援

AI for Scienceに不可欠な計算資源の戦略的増強 76億円｜公募要領・選定結果・基本的な戦略方針の公開一次資料から6項目を抽出

求められた要件	公開資料の根拠	公開資料が示す実例
①共用前提の規模と提供量	公募要領(i)①③④：数100GPU規模を整備し令和9年度内にHPCI共用開始、整備後5年以上の提供と資源量の計画	東大：GPU時間の80%以上をHPCIへ、2027年6月～2033年3月提供／北大：28GPU年を2027年度から
②HPCI連携と相互運用性	公募要領(i)⑤・(ii)⑤：採択機関間の相互運用性確保を含め、今後のHPCIの在り方の検討に参画し協力	北大：東大・筑波大・九大と運用面を含め連携／東大：h3-Open-BDECで機関間ジョブ連携
③運転環境と5年の継続運営	公募要領(i)②：光熱水費・高圧電源・設置場所・人的体制と整備後5年以上の継続運営。補助は単年度	東大：高温水冷却を自己資金で整備、運転費は利用負担金・自己資金で確保。柏IIに電源2MW
④外部GPU・機微データ対応	戦略方針⓬：産学の計算資源(GPU)も含めた機動的な利用・調達を支援。申請では機関外制度と内規を説明	九大：商用クラウドモデルをAPI経由で併用／東大：Confidential Computingで産業利用に対応
⑤新しいHPCI利用制度へ追随	公募要領：整備資源は今後設置される新たなHPCI利用制度で共用。戦略方針⓭：2026年度中に有償利用	補正予算資料：現行の利用申請システムの抜本的改修をRISTへ委託。UI改善・スマホ対応・管理の一元化
⑥AI活用支援体制と人材	申請内容②④：提供資源量・利用者数・論文数の目標値と過去3年の実績。人材は2030年度までに200人以上	東大：HAIRDESCのAI for Science担当中核機関／北大：知識生成基盤研究部門がAI技術支援

数値の位置づけ：2030年度までが集中改革期間

共用計算資源を2030年度までに10倍以上（戦略方針⓭）
次世代HPCIの新配分システムを2030年度までに構築
(i)は上限40億円・2件程度に申請7件→東大・九大
(ii)は上限5億円・3件程度に申請7件→北大・筑波大

本提言との接続

共用GPUテストベッドと診断は(ii)共用の効率化と同方向
統一的UI/UX・統合認証は利用申請システム改修と同じ課題
産業利用ワンストップ窓口は内規説明の要求と接続できる

共用計算資源増強への提案：Agentic AI推論を支える計算基盤の要件

米中フロンティア企業の公開資料と実測トレースから6項目｜ピーク性能ではなく「メモリ×キャッシュ×隔離実行」が律速

要件	根拠となる公開実測値	批判的注記：何が未検証か
①【HW】HBM容量×帯域で調達	エージェント処理の91.0〜98.6%がデコードで メモリ帯域律速。GB300 NVL72=20TB/576TB/s	91〜98%はGemma4-31B等の公開2モデルでの実測。 ラック電力は非公開、独自ASICはMLPerf不参加
②【SW】KVキャッシュを一級資源に	Claude Code/Codex実トレース4,300セッション： 前置文脈中央値12.6万、キャッシュヒット95.7%	別実験の3.8倍はヒット1.7%からの比較で過大。 文脈圧縮を入れるとヒット率は半減する
③【HW/SW】KVの階層記憶と広帯域	8万トークン文脈では3.5GB/sで再計算に勝つ (NVMeは13.5GB/s)。DeepSeekはKVヒット56.3%	Dynamoは4階層を文書化するが実測値を非公開。 CXLは+153〜211nsの遅延増を伴う
④【SW】プログラム単位の管理	負荷0.9で待ち時間がReAct 87%・MCTS 91% (Chatbot 80%)。ツール実行が実時間の59.8%	4〜15倍の基準はvLLM 0.6.1・A100・公開モデル。 チャンクドプレフィル等と競合し加算不可
⑤【HW】隔離実行と高CPU:GPU比	rStar2-Agentは1ノード1,024実行ワーカーで 1RLステップ4.5万並行呼出・平均0.3秒	GPU:CPU比と外向き帯域の設計値を公開した ベンダーはゼロ。論文もノード数を記載せず
⑥【SW】疎注意・MLAでKV削減	MLAのKVは70.3KB/token（GQA比4.66倍小）。 DeepSeekはDSAでAPI価格50%超減と公表	70.3KBはDeepSeek自身のISCA'25論文の値。 長文精度の独立再現とCANN比較は未公開

批判的な前提：公開資料の限界

本番の利用率・バッチ・トークン単価を公開する米企業はない
完全な本番開示はDeepSeekの1件のみ（ディスクKVヒット56.3%）
「10GW」等は契約・計画値であり通電量ではない
CM384の電力性能2.5倍劣は第三者試算（論文にkW記載なし）

日本の共用基盤への含意

追うべきはGW規模でなくソフト側。KV階層・キャッシュ認識スケジューラ・疎注意は公開実装で第三者再現済み
MLPerf Agenticのデータセンタ版は未リリース（v6.1はEdge版のみ）
評価はトークン単価でなく完了タスク当たり費用へ