多様なAI半導体の活用と計算資源の高効率化に関する研究開発

実施者	国立研究開発法人産業技術総合研究所、1FINITY株式会社、株式会社AI福島、株式会社ELEMENTS、富士通株式会社、株式会社テプコシステムズ、株式会社RUTILEA、株式会社ゼウレカ
概要	AIの開発力・供給力を支える国内計算基盤の高度化のため、多様なAI半導体の評価と計算資源の高効率化を実現する技術を研究開発する。高性能化や低消費電力が期待されている複数の新興AI半導体からなるテストベッドを整備し、その性能や省電力性、AI開発の利便性、運用性等の多面的な評価を行い、用途に応じた利活用指針を明らかにする。また、それらAI半導体を含む計算資源の高効率化として、AIワークロードの実行性能予測や高効率・高性能推論システムを開発する。

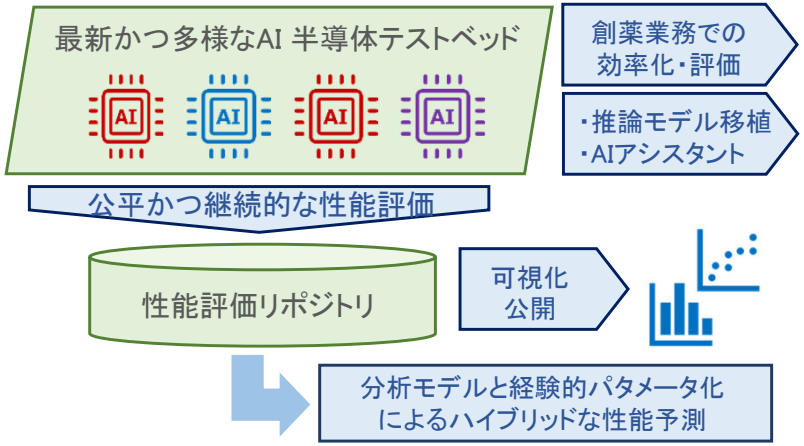
【①多様なAI半導体の評価と効率的な利用技術の研究開発】

- 多様かつ最新のAI半導体の特性・優位性を明らかにするテストベッド構築と公平なAIベンチマークセットの開発
- 実業務（創薬ワークロード）での効率的な利用技術の開発
- 多様なAI半導体での推論モデル移植・実行技術及び生成AIによるアシスタント機能の高度化

【②大規模計算資源利用の高効率化技術の研究開発】

②-1）実行性能予測に基づく資源利用の効率化

- 学習／推論の高精度な実行性能予測技術の開発
- 学習／推論処理の継続的な性能評価と性能予測用データセットの高度化



②-2）高効率・高性能 推論システムの研究開発

a. 入力データ選択制御 ： 多種多様なデータソースからの入力データ量の削減とクエリ処理の効率化・高速化技術の開発
b. LLM推論API ： 日本語とデータ操作言語を組合せた入力に対し投機的デコーディングによる適切な推論エンジン選択技術の開発
c. 構成管理ソフト ： 需要変動に基づく、推論基盤モデル向け計算資源構成の動的適正化技術の開発
d. スケジューラ ： AI半導体の利用効率を上げる高度なスケジューリング技術の開発
e. KV Cacheオフロード ： LLM推論のスケーラビリティ向上と効率化に向けた設計指針の確立