



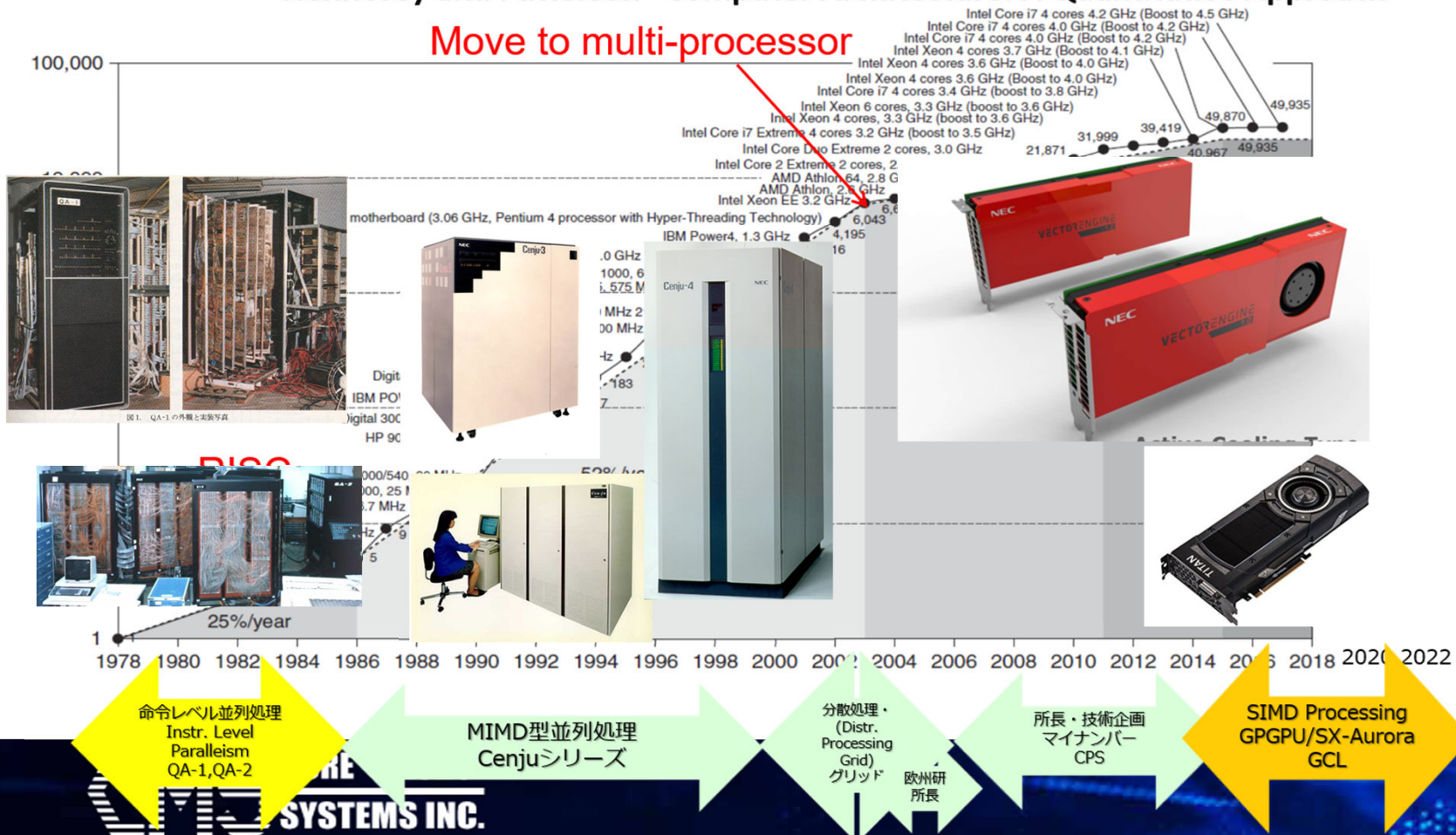
ガード付AIエージェントを用いたHPCベンチマーク最適化フレームワークの検討

July 2, 2026
コアマイクロシステムズ株式会社
中田 登志之

自己紹介

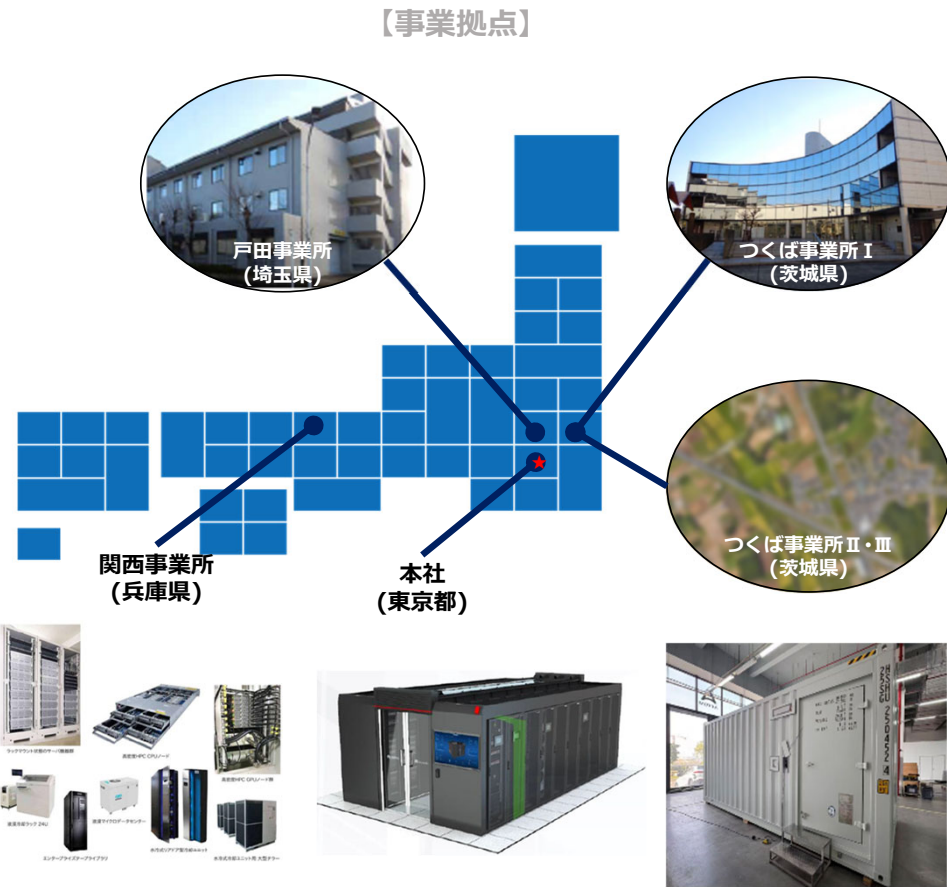
Hennesey and Patterson "Computer Architecture: A Quantitative Approach"

Move to multi-processor



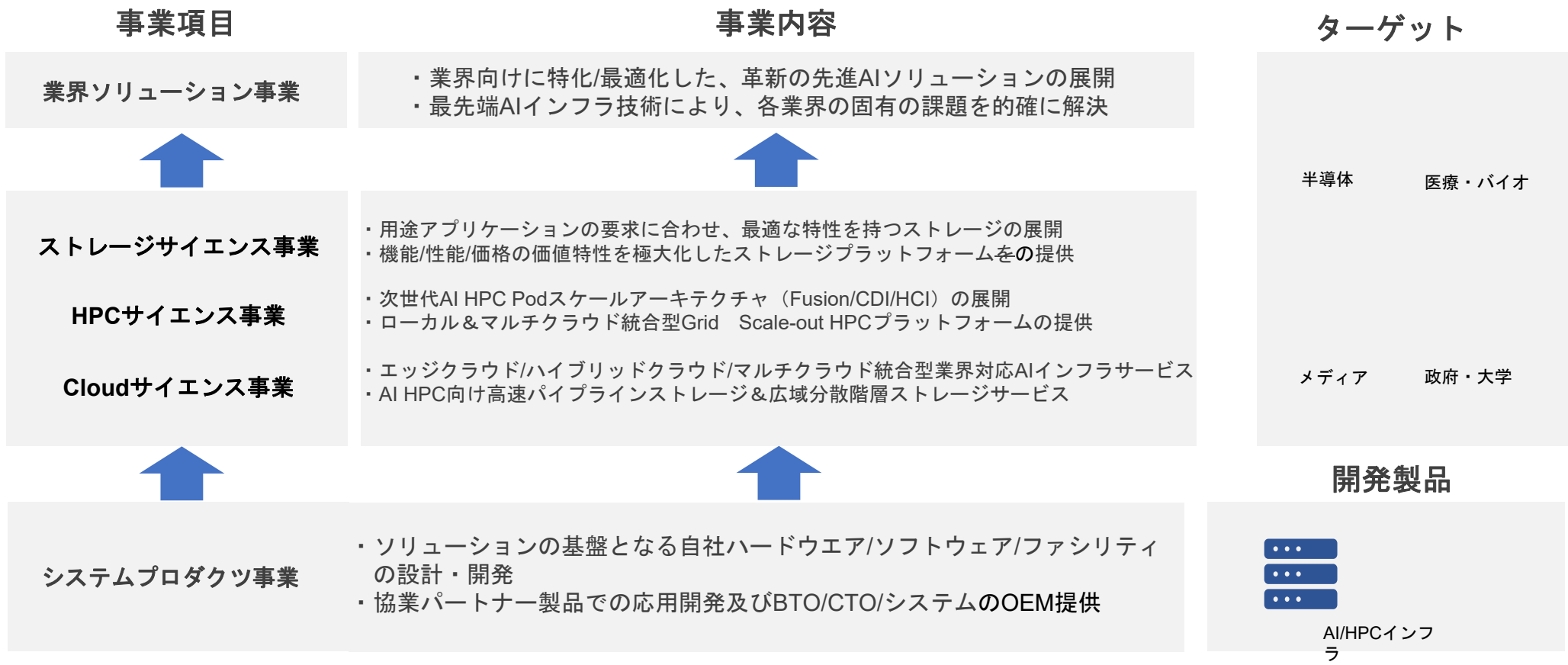
会社概要

社名	コアマイクロシステムズ株式会社 (CORE MICRO SYSTEMS INC.)
所在地	東京都板橋区舟渡1-12-11 ヘリオスIIビル 3F
設立	1992年（平成4年）10月16日
代表者	代表取締役 高橋 晶三
事業内容	- 業界ソリューション事業 - ストレージサイエンス事業 - HPCサイエンス事業 - Cloudサイエンス事業 - システムプロダクツ事業



事業内容

ミッション：最先端AIテクノロジーによる革新的AIソリューションの提供



HPCチューニングの課題

- HPCベンチマークの性能は多くのパラメータに依存する
- 例
 - MPIプロセス数
 - OpenMPスレッド数
 - ランタイム設定
 - 通信設定
- 問題
 - 探索空間が大きく
手動チューニングが困難
 - システム毎に最適値が異なる
- 目的
 - AI Agent による性能最適化支援

フレームワーク全体構造

ベンチマーク計測

2段階の最適化

1) パラメータの広域探索

(Optuna * 利用)



最有力候補選択



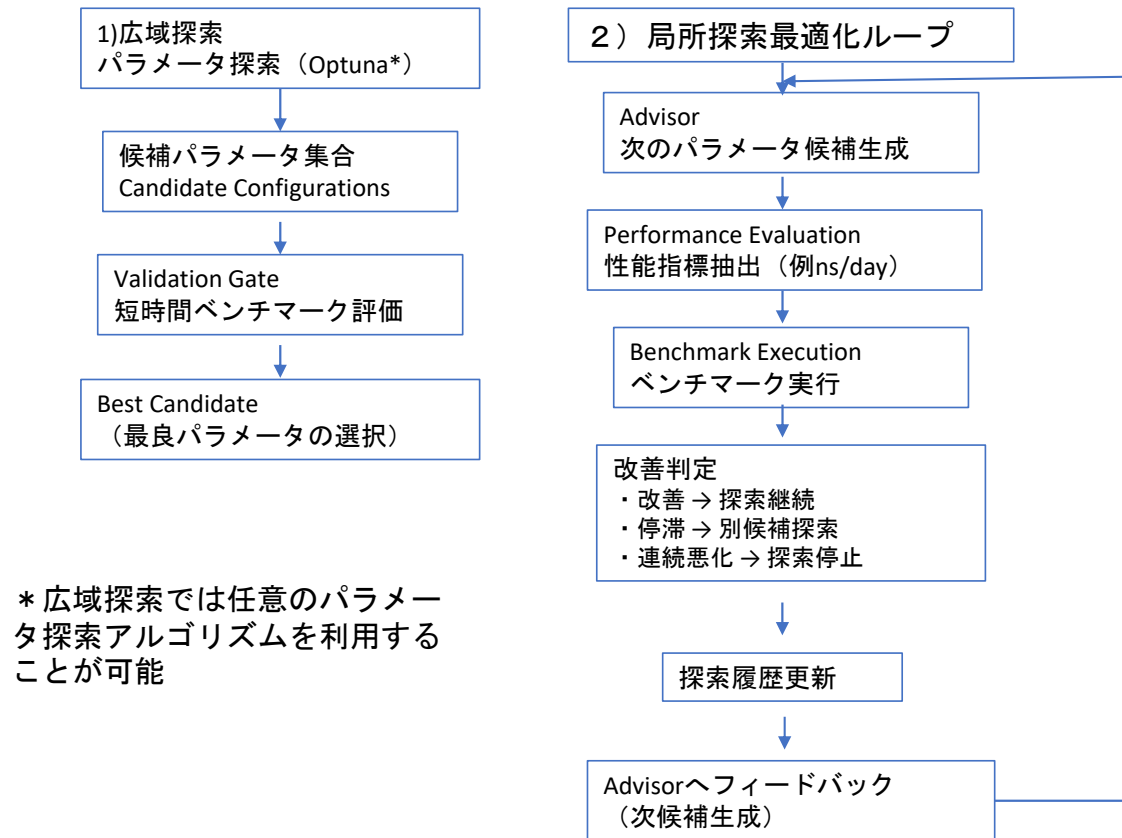
2) 局所探索最適化ループ

* 広域探索では任意のパラメータ探索アルゴリズムを利用することが可能

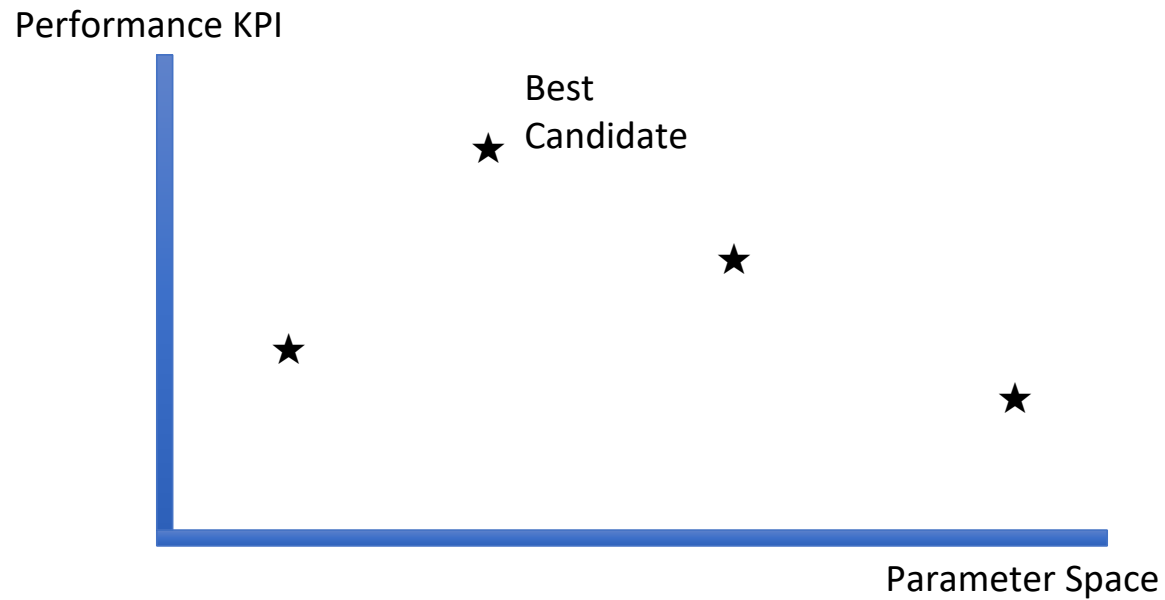
探索空間例

- GROMACS
 - Ntmpi(mpiスレッド数)
 - Ntomp(OpenMPスレッド数)
 - Pin(メモリのピンダウン)
- ONNX
 - Provider (cpu/cuda)
 - Batch_size
- BabelStream
 - Array-size
 - numtimes
- LAMMPS
 - mpi_ranks
 - omp_threads
- OpenFOAM
 - mpi_ranks
 - omp_threads
 - mesh_size

AIによるHPCベンチマーク最適化フレームワーク

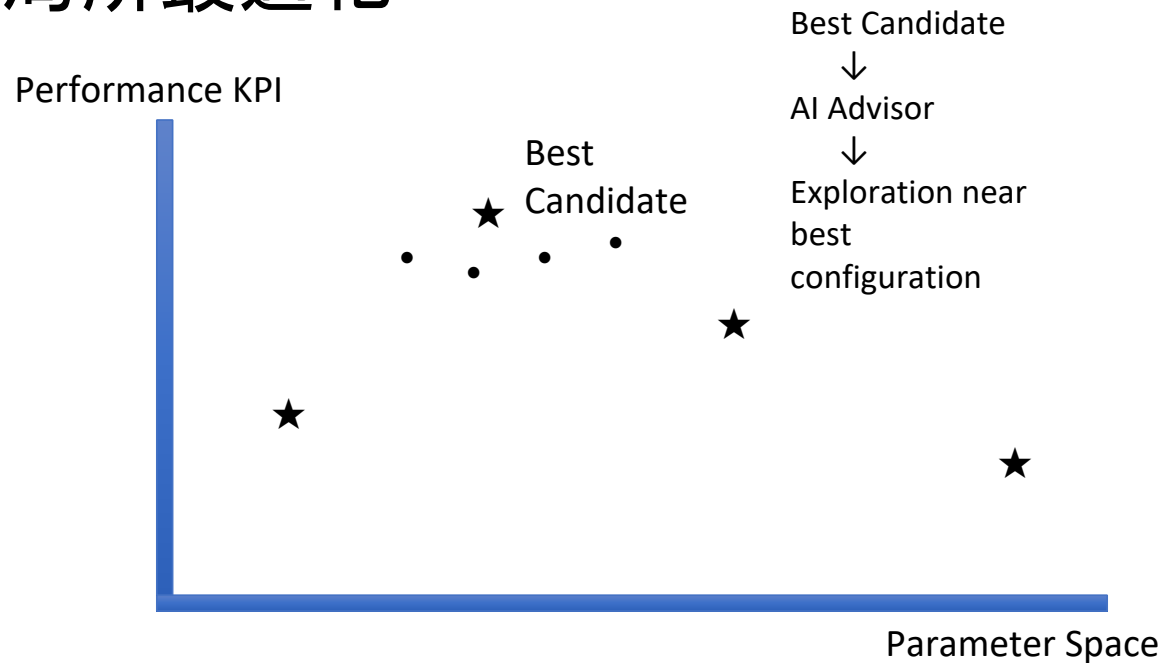


広域探索



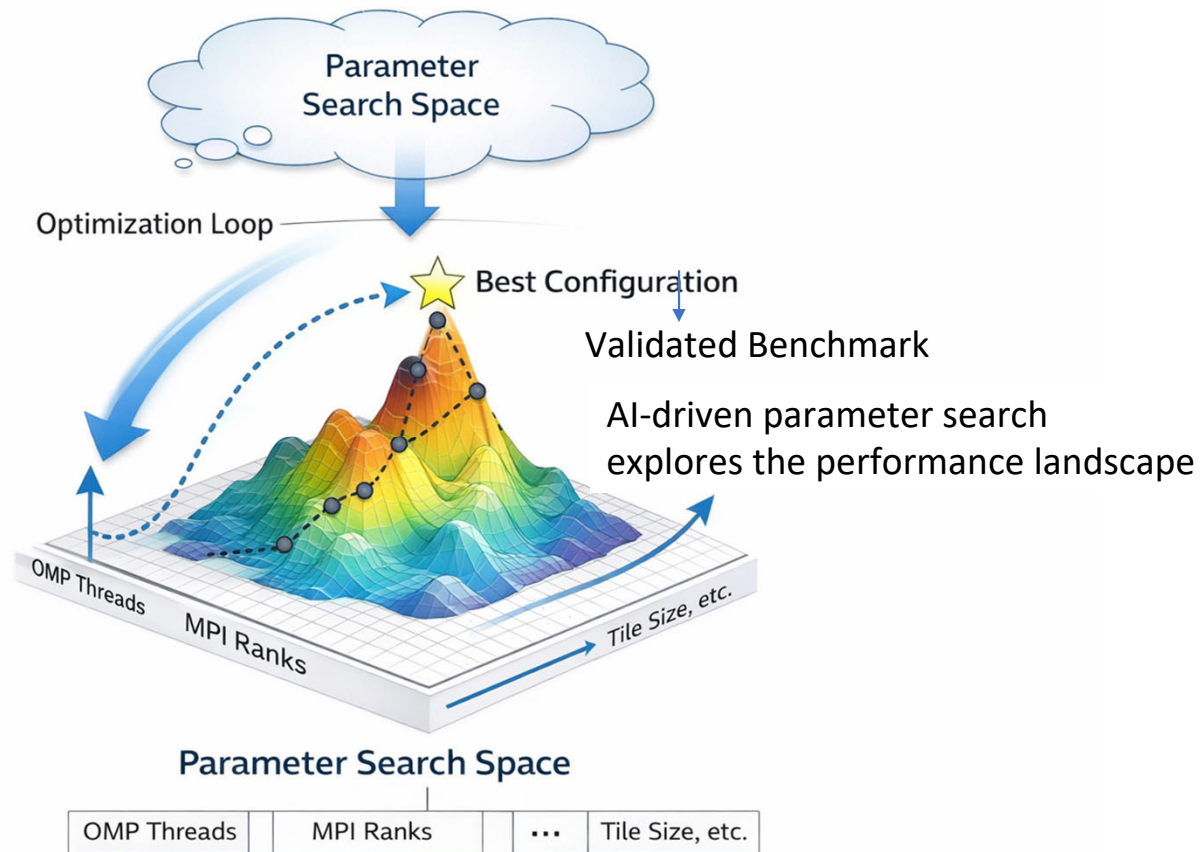
広域探索ではOptuna等を用いて
パラメータ空間全体を広く探索する。
その中から有望候補Top-Kを抽出する。

2)局所最適化



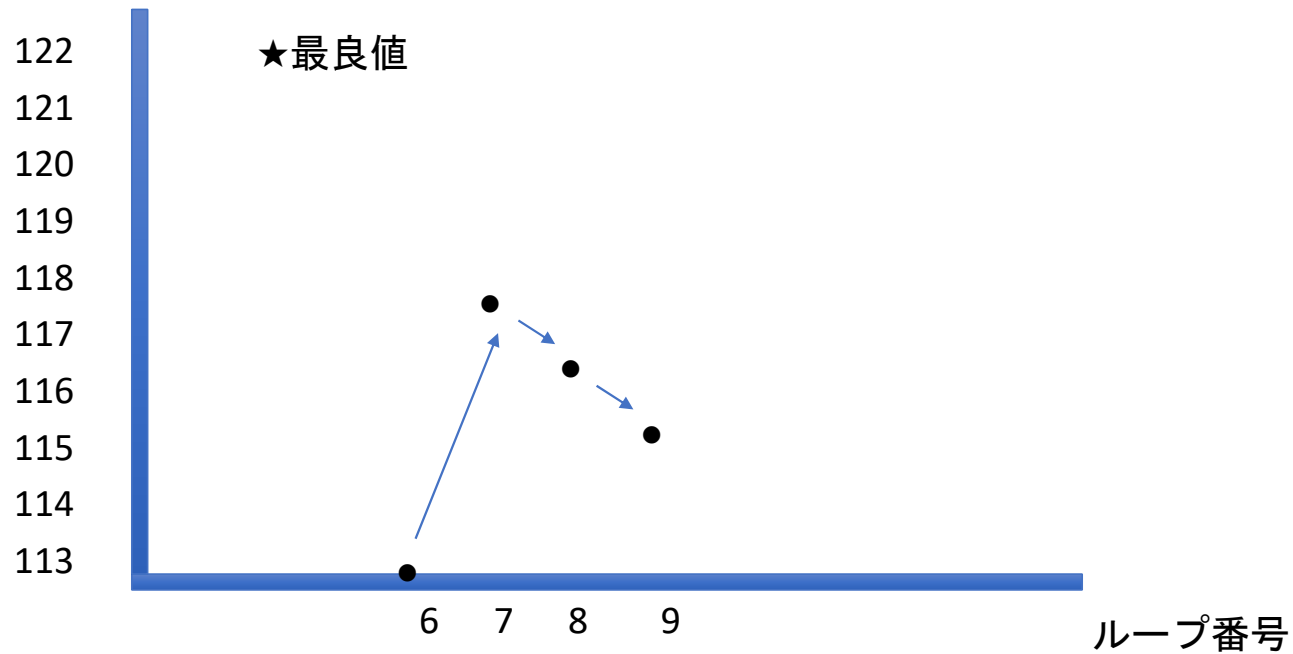
局所最適化では広域最適の最有力候補の近傍パラメータを生成し探索して性能改善を試みる。
性能改善が停止した場合探索を自動終了する。
AI Advisor が探索履歴を解析し最良候補の近傍に新しい構成を提案する。

AI-Driven Benchmark -> Scientific Optimization Framework

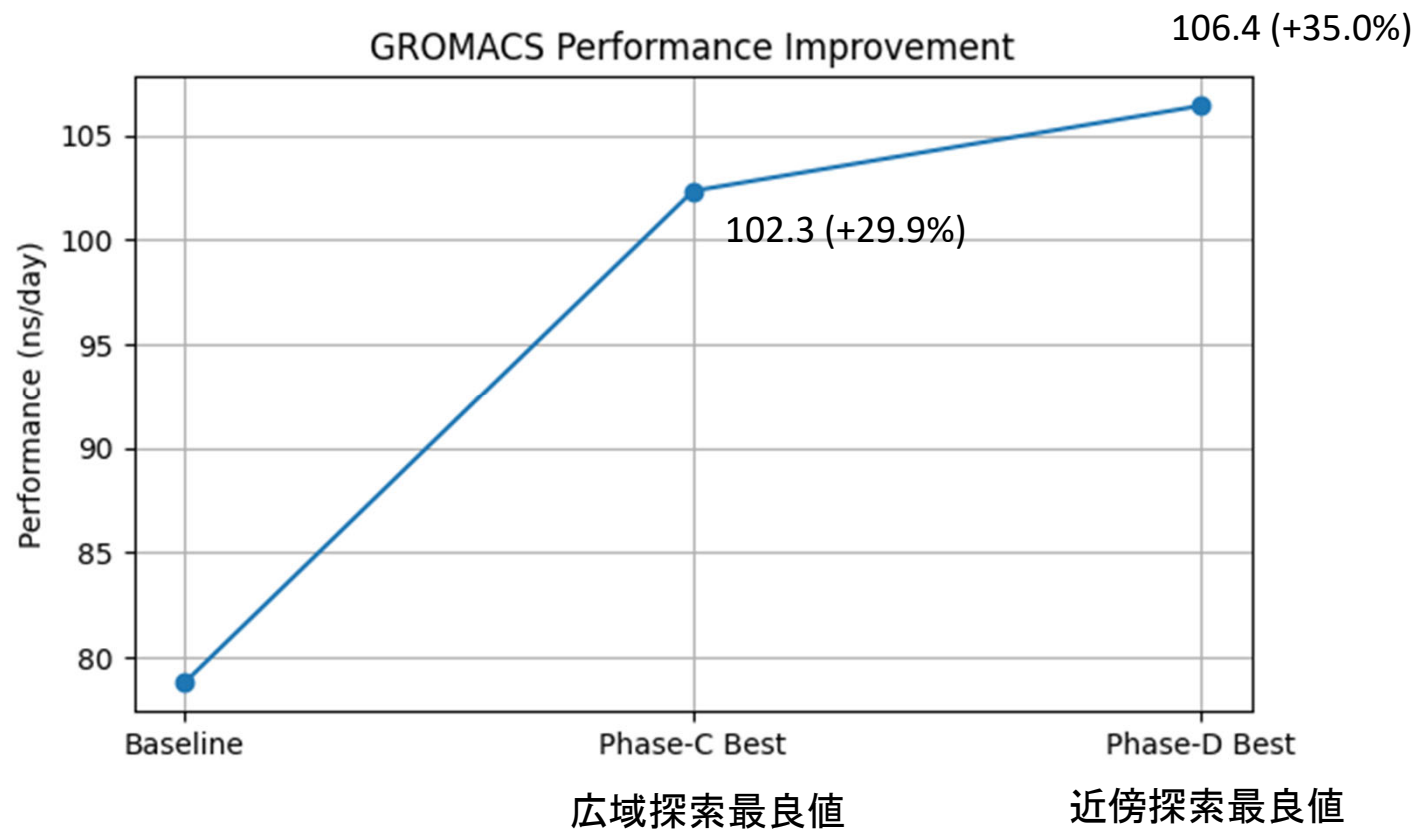


局所最適化の動き（GROMACSの例）

性能目標値KPI ns/day

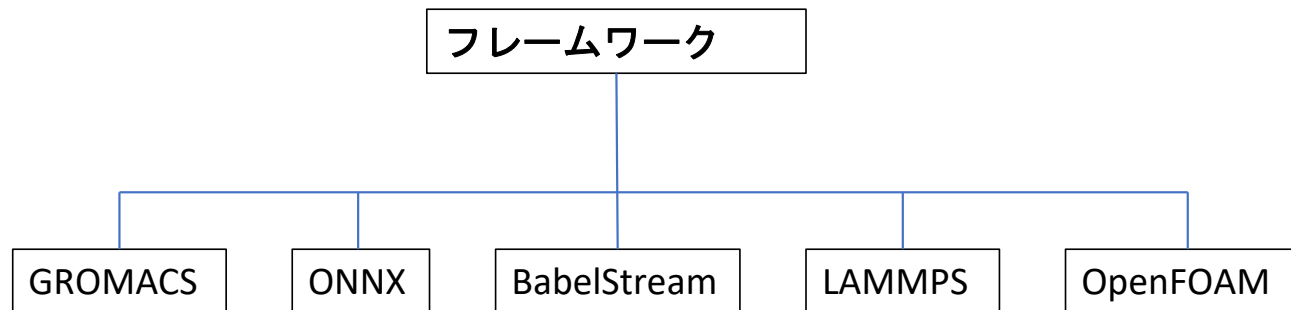


Phase-Dでは履歴に基づき新しいパラメータ候補を生成する。
性能改善が停止した場合探索を自動終了する。（今回は試行9で停止）



Cross Benchmark Result

- 同一フレームワークで複数ベンチマーク実施



マシン特性を考慮した探索例

- GPUメモリ容量を用いた ONNX batch size 探索を実施

- GPU3070搭載マシン：

バッチサイズを1から**32**に増加させることで
スループットを 367.6fps から755.2fpsに

約 2 倍 改善

- GPU5070搭載マシン：

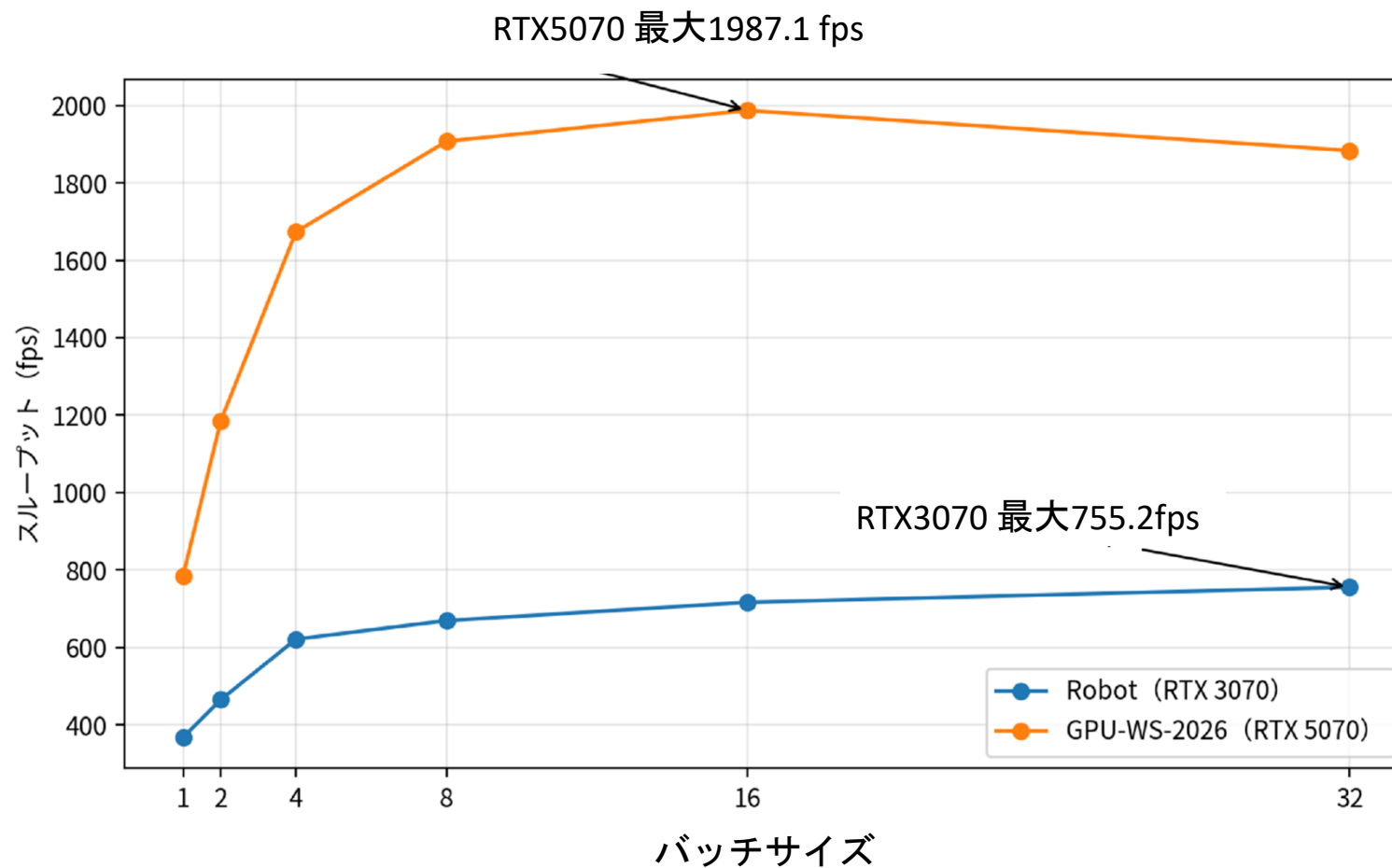
バッチサイズを1から**16**へ増加させることで
スループットを784.9 fps から 1987.1 fps に

約 2.53 倍 改善



GPUメモリ容量だけでなく、GPUアーキテクチャ・実行環境・
kernel効率等により**最適バッチサイズ**が異なることを確認。

ONNX推論スループット性能におけるバッチサイズの効果



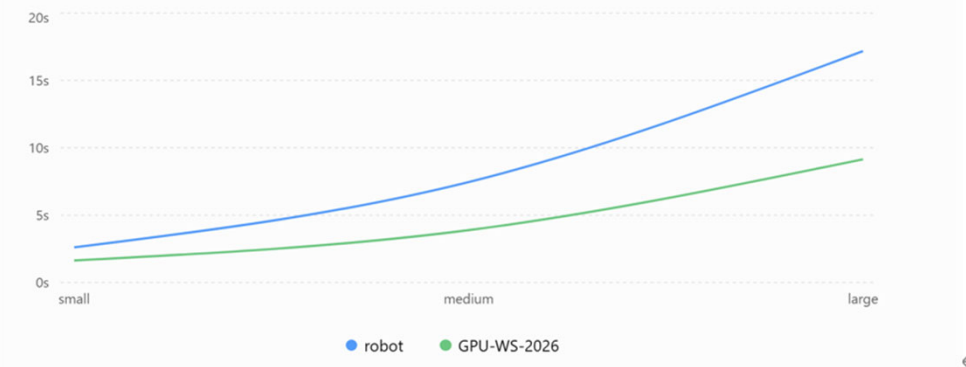
OpenFOAMを2種類のマシンで評価

図1 (メイン図)

OpenFOAM Phase-D 最良値比較

OpenFOAM Phase-D best wall time

Best Phase-D wall time for each mesh size on robot and GPU-WS-2026.



OpenFOAM詳細評価

図2 (robot探索結果)

robot OpenFOAM parameter exploration

Representative wall time results grouped by parallelism configuration.

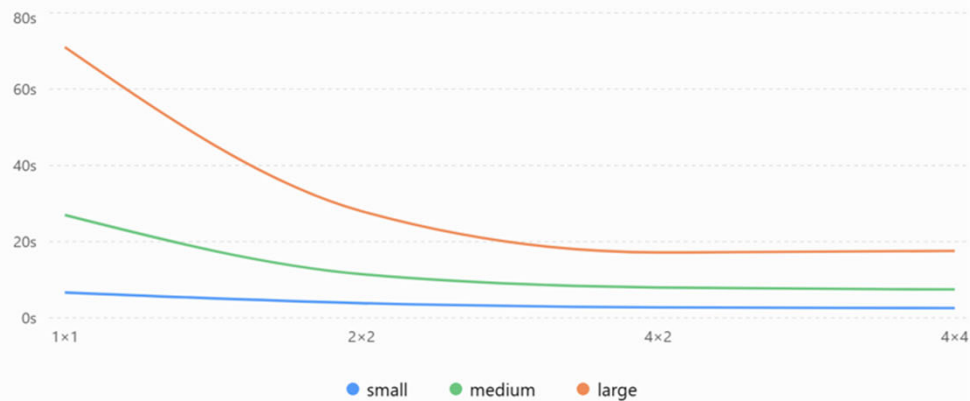
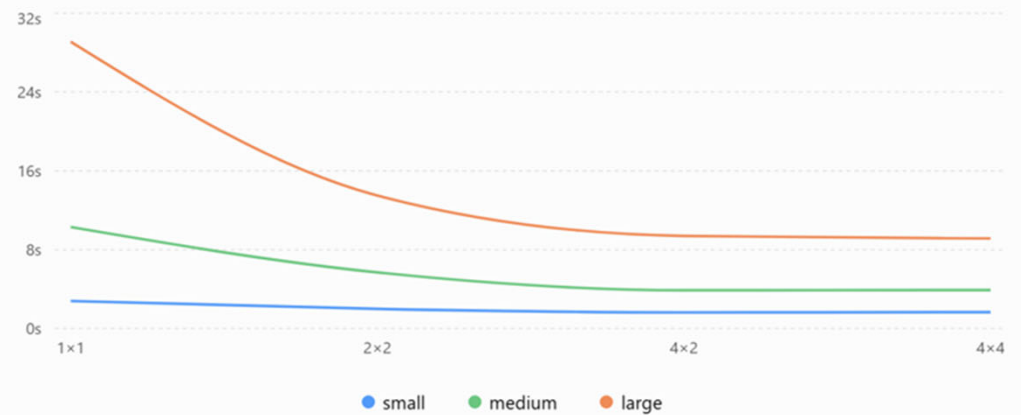


図3 (GPU-WS-2026探索結果)

GPU-WS-2026 OpenFOAM parameter exploration

Representative wall time results grouped by parallelism configuration.



まとめ

- 広域探索と近傍探索を組み合わせたAIベンチマーク最適化手法を確立
- ⇒ 今後より多くのベンチマークで評価を実施予定



HPC4U!

コアマイクロシステムズ株式会社

URL : <https://www.cmsinc.co.jp>