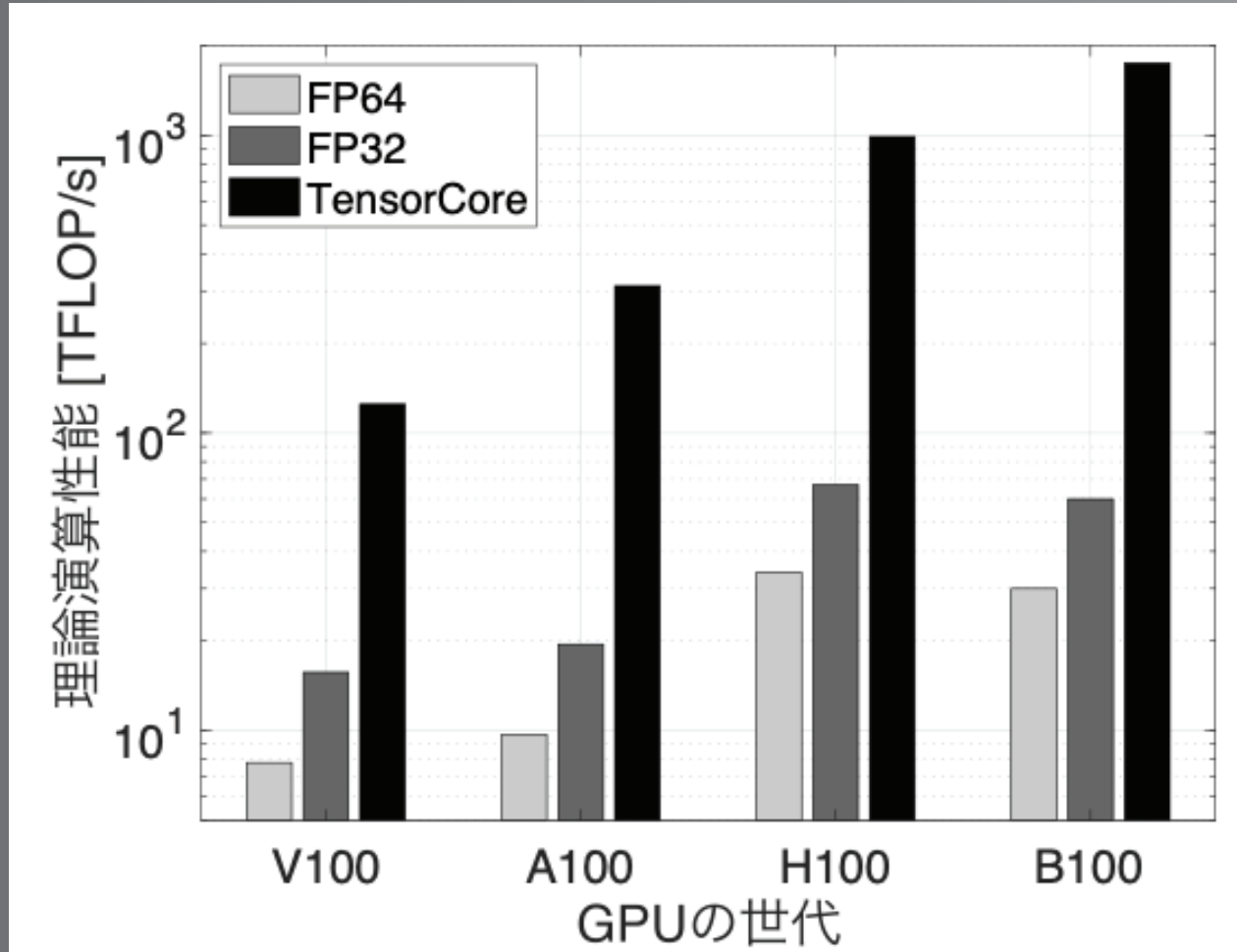


次世代計算機の潜在能力を引き出すための 科学技術アプリケーションの刷新

背景

これからの半導体は深層学習分野の巨大な需要に応えるべくTensor Coreのような低精度行列演算器にますます特化することが予想される

右図にあるように、H100からB100にかけてTensor Coreのみの性能が向上しておりFP64やFP32の性能はむしろ低下している



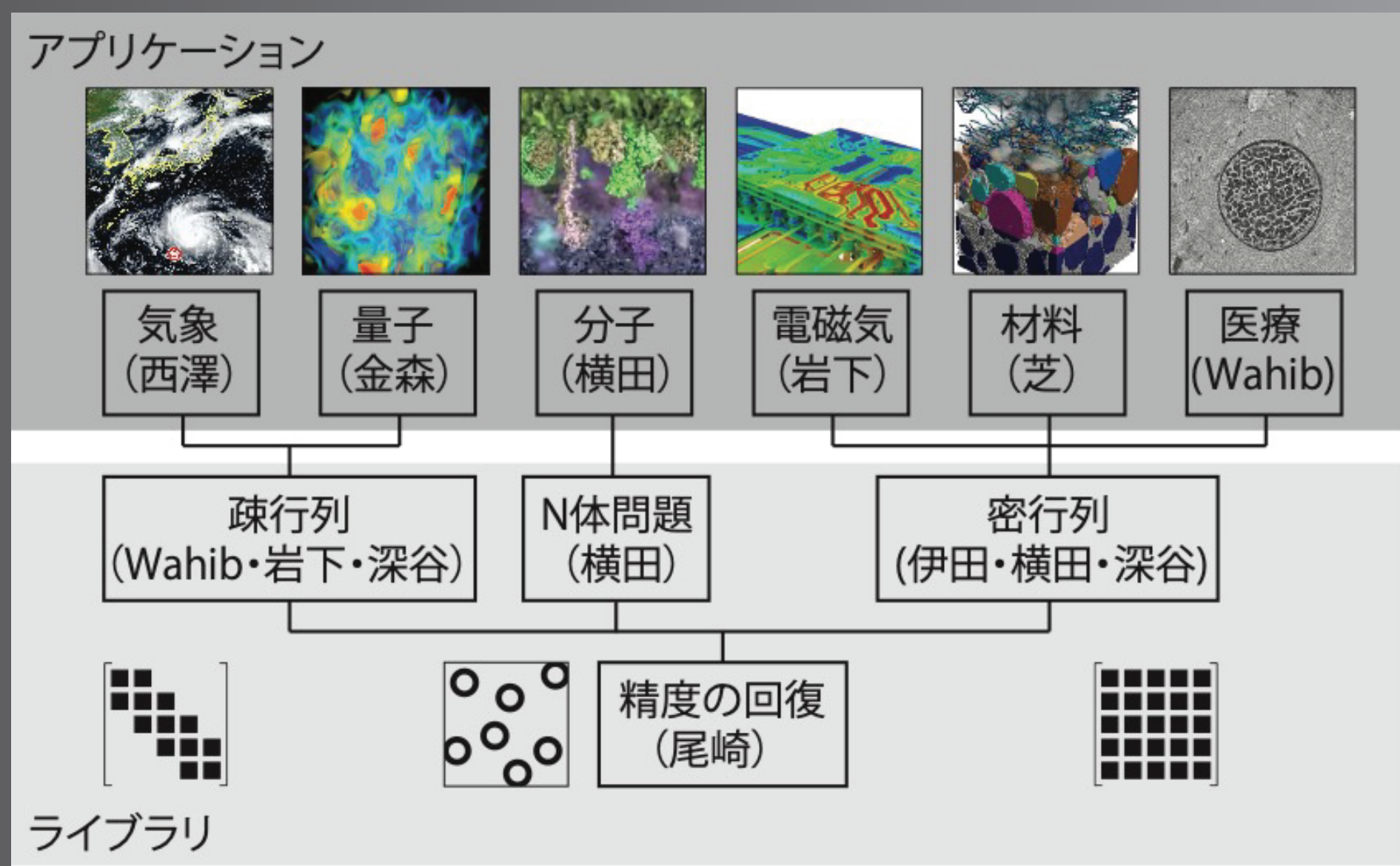
Tensor Coreが利用できないアプリケーションコードは今後半導体の進歩による性能向上が期待できない

欧米では以前からスパコンにGPUが導入されているため主要アプリケーションのGPU化は既に行われているが、Tensor Core化まで行われているものはまだ少ない

国産の計算科学アプリケーションのTensor Core化に今から着手すればそこから世界を牽引する研究が多く派生することが期待される

研究目的

本研究の目的は、これから主流になる深層学習向けのプロセッサに対し、代表的な6つの計算科学アプリケーションの高速化を世界に先駆けて行い、その技術を国内に広く波及させることで、次世代計算基盤における科学技術成果創出の最大化を図ることである



選定したアプリケーションは上図に示す気象・電磁気・分子・量子・材料・医療の6分野に関するものであり、計算科学アプリケーションで扱う代表的な分野からなる。また、それぞれが用いる手法も有限体積法・境界要素法・粒子法・LQCD・深層学習・トモグラフィーと全く性質の違うアルゴリズムがベースになっている

準備状況

6つのアプリケーションのTensor Core(TC)化に向けた準備状況を右表に示す

分野	手法	担当者	GPU 実装	GPU 最適化	密行列積化	TC 化
気象	有限体積法	西澤	○	×	×	×
電磁場	境界要素法	岩下・伊田	○	×	×	×
分子	N 体問題	横田	○	△	○	×
量子	LQCD	金森	○	×	×	×
材料	深層学習	芝	○	○	○	△
医療	トモグラフィー	Wahib	○	△	○	△

いずれのアプリケーションもGPU実装自体はできており、2025年度にはその性能最適化に専念できる状況にある

材料と医療のアプリケーションはTensor Core実装自体もできているが、低精度化をどこまで行えるかが未検証であり、性能に関しても最適化の余地がある

一方、気象・電磁場・量子のアプリケーションは密行列積の形にアルゴリズムを変形する必要があり、材料・医療・分子のアプリケーションに比べて難易度は高いと予想される

研究計画

Tensor Coreを用いないGPU化

気象・電磁気・分子・量子・材料・医療のそれぞれのアプリケーションに対して、まずはTensor Coreを用いないGPU化を行う

これが後にTensor Core化を行った際の効果を定量的に測るベースラインとなる

密行列積の形にアルゴリズムを変形

Tensor Coreを活用するうえで密行列積の形になるようにアルゴリズムや離散化手法を改良しCPU実装を行う

気象アプリ：高次要素や不連続基底を導入することで疎行列を密ブロック化

電磁気アプリ：複数の右辺を同時に処理することで行列・ベクトル積が行列・行列積になる

分子アプリ：FMMを用いることで主要な計算の部分を密行列積の形に変換

医療アプリ：投影行列を再利用することで密行列積化

Tensor Coreを用いた実装の開発

気象・電磁気・分子・量子・材料・医療のそれぞれのアプリケーションに対して密行列積の形に変形したアルゴリズムを元にTensor Core実装を行う

低精度演算の影響の評価

Tensor Coreで扱える精度はTF32, FP16, BF16, FP8, INT8, FP4, INT4など様々なものがある

Ozaki schemeを用いることでアプリに必要な精度を最小限のオーバーヘッドで再現

密行列・疎行列・N体問題のライブラリ化

密行列・疎行列・N体問題のそれぞれのTensor Core実装をオープンソースライブラリとして提供

関連分野の研究動向

疎行列の計算にTensor Coreを用いるもの[1, 2]

ステンシル計算にTensor Coreを用いるもの[3, 4, 5]

マルチグリッド法にTensor Coreを用いるもの[6]

[1] Y. Lu, W. Liu, “DASP: Specific Dense Matrix Multiply-Accumulate Units Accelerated General Sparse Matrix-Vector Multiplication”, Proc. of SC23, Nov. 2023.
[2] P. Okanovic, G. Kwasniewski, P. S. Labini, M. Besta, F. Vella, T. Hoefler, “High Performance nstructured SpMM Computation Using Tensor Cores”, Proc. of SC24, Nov. 2024.
[3] X. Liu, Y. Liu, H. Yang, J. Liao, M. Li, Z. Luan, D. Qian, “Toward Accelerated Stencil Computation by Adapting Tensor Core Unit on GPU”, Proc. of ISC22, Jun. 2022.
[4] Y. Chen, K. Li, Y. Wang, D. Bai, L. Wang, L. Ma, L. Yuan, Y. Zhang, T. Cao, M. Yang, “ConvStencil: Transform Stencil Computation to Matrix Multiplication on Tensor Cores”, Proc. of PPOPP24, Feb. 2024.
[5] Y. Zhang, K. Li, L. Yuan, J. Cheng, Y. Zhang, T. Cao, M. Yang, “LoRASStencil: Low-Rank Adaptation of Stencil Computation on Tensor Cores”, Proc. of SC24, Nov. 2024.
[6] Y. Lu, L. Zeng, T. Wang, X. Fu, W. Li, H. Cheng, D. Yang, Z. Jin, M. Casas, W. Liu, “AmgT: Algebraic Multigrid Solver on Tensor Cores”, Proc. of SC24, Nov. 2024.