jh190043-NAHI

横田　理央 (東京工業大学)

Hierarchical low-rank approximation methods on distributed memory and GPUs

背景

H行列、H^2行列、HSS行列などの階層的低ランク近似法はO(N^2)の要素を持つ密行列をO(N)の要素を持つ行列に圧縮することができる。圧縮された行列を用いることで、行列積、LU分解、固有値計算をO($N\log^2 N$)で行うことができるため、従来密行列の解法が用いられてきた分野では階層的低ランク近似法の導入が近年盛んに行なわれている。また、密行列のみならず、疎行列の直接解法におけるSchur補元の圧縮に用いることもできるため、流体、構造、電磁界解析において前処理法として用いる研究も盛んに行なわれている。しかし、これらの階層的低ランク近似法は比較的新しい手法であるため、高性能な並列実装は少なく、GPUなどへの実装も未成熟である。これらの階層的低ランク近似法に内在する並列度は高く、高性能な分散メモリ・GPU実装に大きな期待が寄せられている。

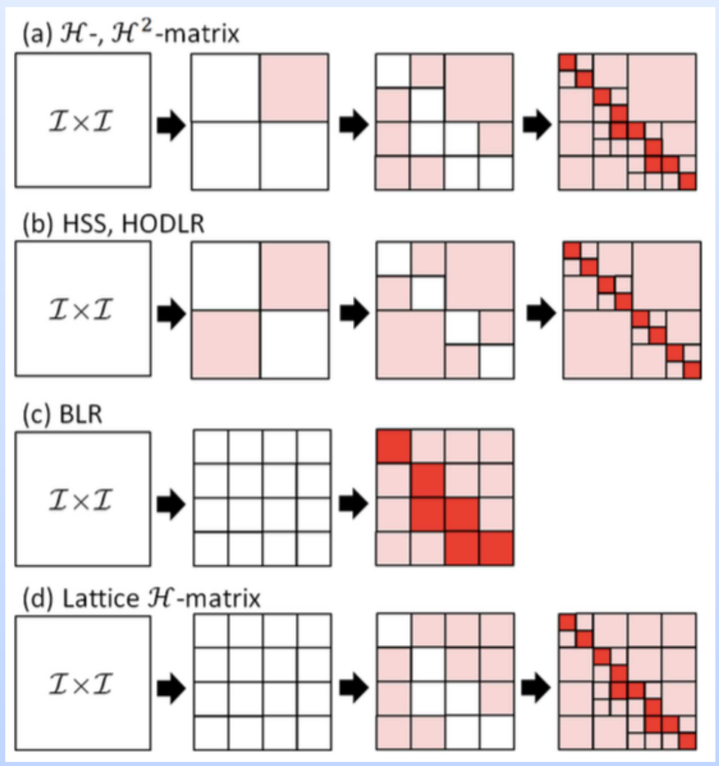
目的

本研究では,エクサスケールを視野に入れた階層的低ランク近似法の分散メモリ・GPU上での高性能な実装を行うことを目的とする。このとき重要になるのが比較的小さな密行列の高速な処理である。Tennessee大学のDongarraグループではまさにこのような小さな密行列のバッチ処理をGPU上で高速に行うライブラリを開発しており、JHPCNの国際共同研究として行うことでこの技術をいち早く導入できる。

昨年度は分散並列でH行列のLU分解を行う際の負荷分散アルゴリズムやバッチ処理を用いる際の階層的データのストリーム化について開発を行ったが、今年度はStarPUなどのランタイムを用いたH行列のLU分解やマルチGPU上でのH行列のLU分解、さらにH行列のQR分解も行う。

Lattice H行列によるLU分解

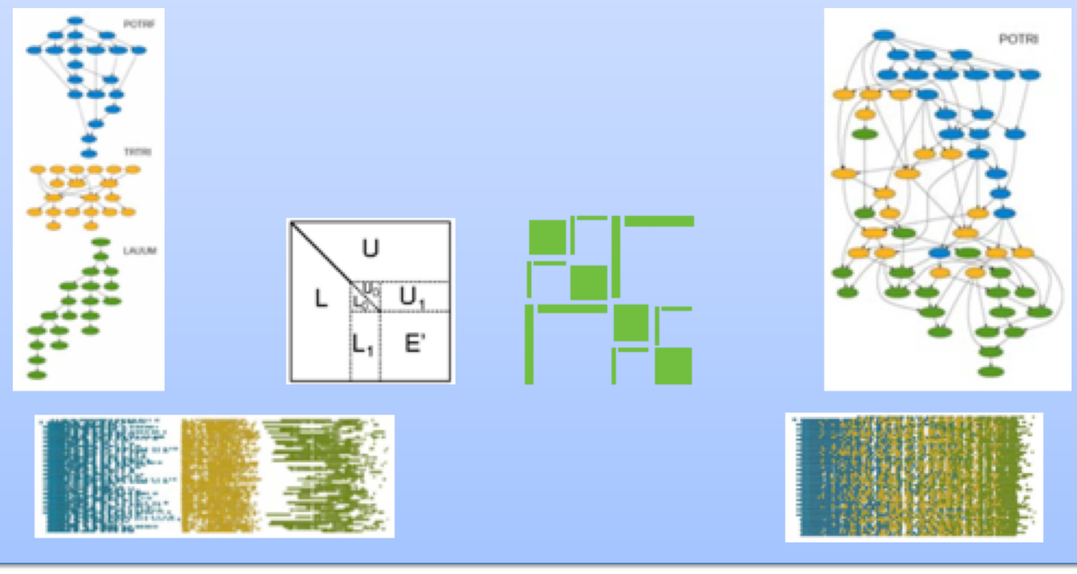
右図に様々な低ランク近似法の違いを図示する。(a)にあるH-matrixやH^2-matrixなどの手法は(b)のHSSやHODLRとは異なり、非対角ブロックをより細かく分割するのが特徴である。(c)のBLRは階層的でない低ランク近似法であり、(d)はBLRとH-matrixのハイブリッドである。(d)のlattice H-matrixはハイブリッド化により、BLRのもつ並列度とH-matrixのもつO($N\log^2 N$)の計算コストの両方を有する。下の表にLattice H行列の主要な関数の呼ばれる回数と計算コスト、その積である全体の計算コストを示す。



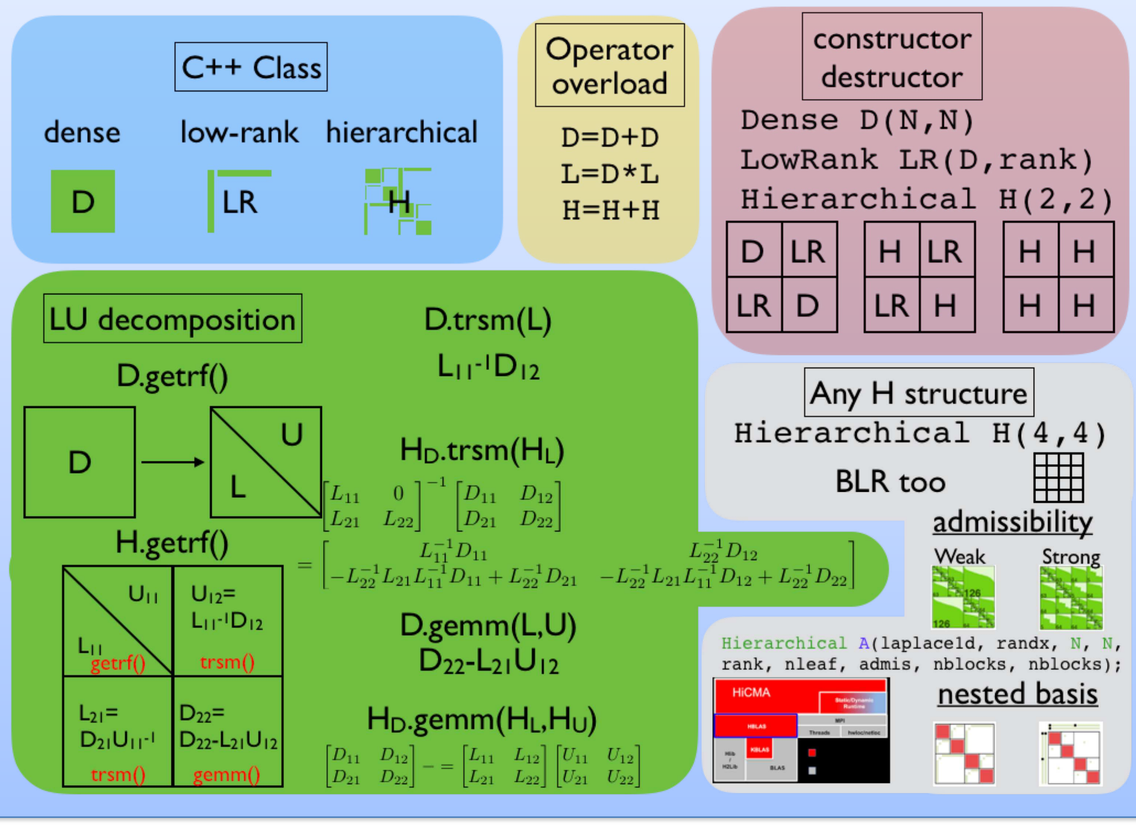
Function	# of calls	Complexity	Total complexity
getrf($\mathcal{H}$)	$\mathcal{O}(n/l)$	$\mathcal{O}(l\log^2 l)$	$\mathcal{O}((n/l) * (l\log^2 l))$
trsm(low-rank)	$\mathcal{O}((n/l)^2)$	$\mathcal{O}(l)$	$\mathcal{O}((n/l)^2 * l)$
trsm($\mathcal{H}$)	$\mathcal{O}(n/l)$	$\mathcal{O}(l\log l)$	$\mathcal{O}((n/l) * (l\log l))$
gemm(low-rank)	$\mathcal{O}((n/l)^3)$	$\mathcal{O}(l)$	$\mathcal{O}((n/l)^3 * l)$
gemm($\mathcal{H}$)	$\mathcal{O}((n/l)^2)$	$\mathcal{O}(l\log^2 l)$	$\mathcal{O}((n/l)^2 * (l\log^2 l))$

H行列のLU分解におけるランタイムの活用

H行列を圧縮する計算やH行列・ベクトル積は個々ののブロックを独立に計算できるため、依存関係のないバッチ処理になる。一方、LU分解では個々のブロックの計算が依存関係にあるため、単純なバッチ処理を行うことができない。密行列のブロック並列化では既に広く行われているランタイムによる依存関係の処理を利用した動的負荷分散は下図に示すように、ブロックLU分解などの処理において大幅な計算時間の短縮につながる。ただし、H行列の場合は、これに階層構造と、低ランクブロックという要素が追加されるため、動的負荷分散は容易ではない。



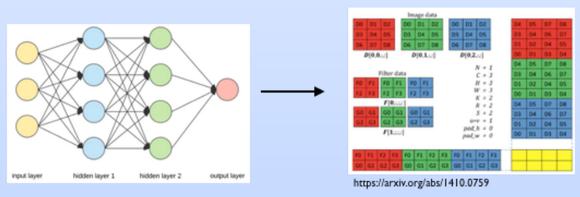
新たなC++コードの設計



深層学習におけるH行列の利用

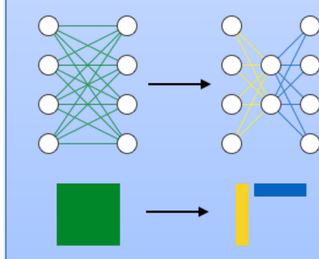
深層ニューラルネットは密行列演算になる → H行列は密行列の高速近似解法

畳み込みNNは密行列積になる



- 行列が非常に長細い
- ブロックのランクが高い
- 圧縮した行列が１回しか使われない

低ランク近似を用いてNN自体を圧縮



- 汎化性能と敵対的データに対する堅牢性が維持できるのか？
- 全結合層にしか有効でない
- AutoMLの出現