Runtime System for GPU-based Hierarchical LU Factorization

Qianxiang Ma, Tokyo Institute of Technology. ma@rio.gsic.titech.ac.jp
Rio Yokota, Global Scientific Information and Computing Center. rioyokota@gsic.titech.ac.jp

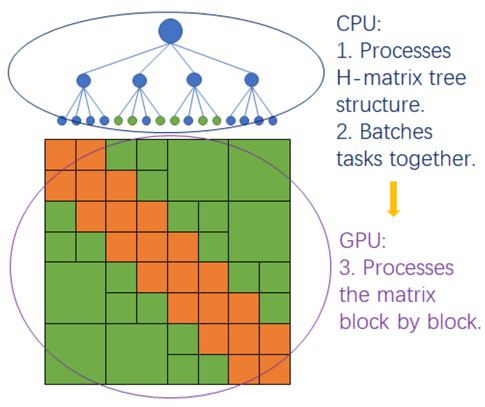
INTRODUCTION

Hierarchical Low-Rank Approximation of Matrices reduces not only the storage requirement from O (n^2) to O (n log n), but also the number of floating point operations (FLOPS) of matrix calculations.

GPUs have considerably more cores inside and greater calculation potentials than CPUs, but fully utilizing such potential is a also challenging task.

DESIGN

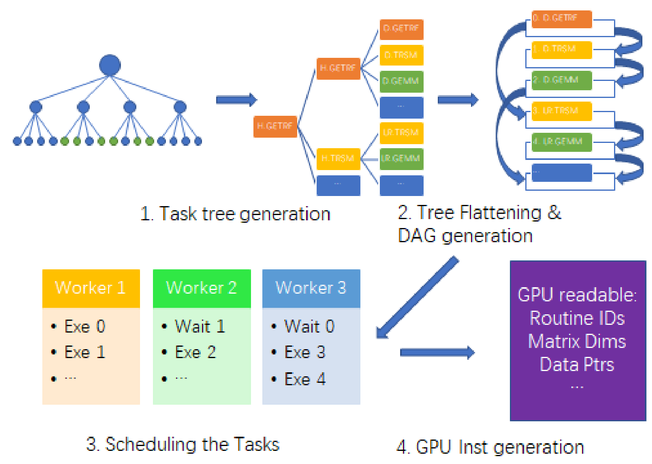
When CPU and GPU are collaborating, they have different strong points: CPU can handle branches, tree structures and recursions effectively, but GPU is faster doing repetitive tasks. To best utilize their strong points, we have a highly modularized and pipelined design.



HOST (CPU)

Different from Matrix multiplications, In-place Hierarchical LU factorization has dependencies amongst the tasks, that fetching premature data leads to incorrect factorizations results and race conditions.

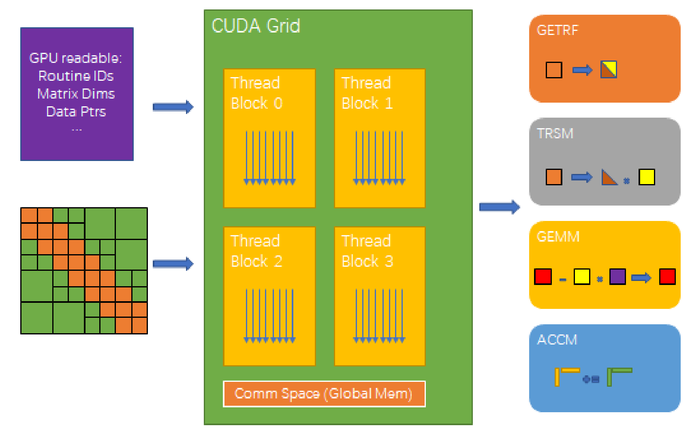
An Illustration of the Host:



DEVICE (GPU)

GPU uses a unified kernel with multiple device functions that implements the BLAS routines. All threads in one thread block execute a device function together.

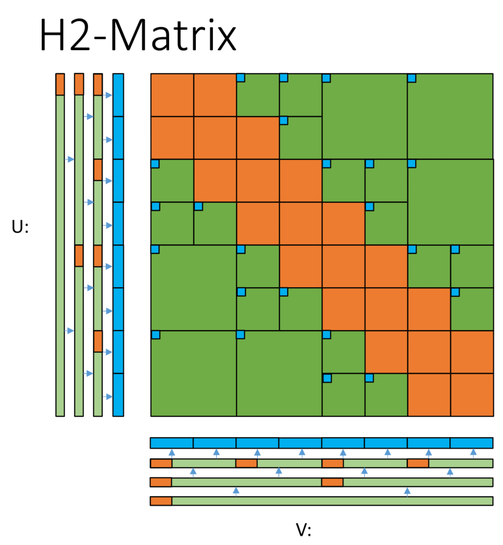
Kernel Level & Thread Block Level:



Warp level: Vectorized mem I/O & Warp shuffling.

H2-EXTENDIBILITY

- Factorizing an H2-matrix uses even less floating point operations when comparing with factorizing an H-matrix.
- Fewer total elements stored.
- Strong connection between blocks located on the same rows / columns, as well as among different layers in the hierarchy.
- Very careful updating to prevent unwanted side effects due to the shared (nested) bases among blocks and layers.



RESULTS

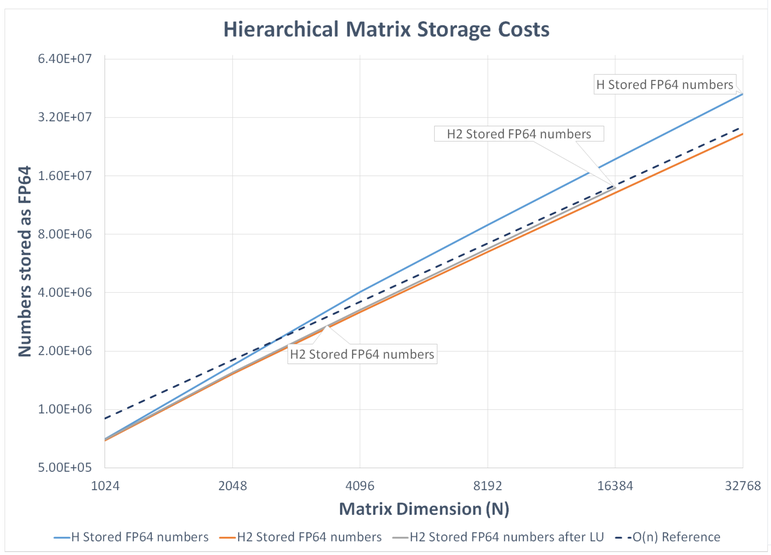


Figure 1: Linear Storage Cost

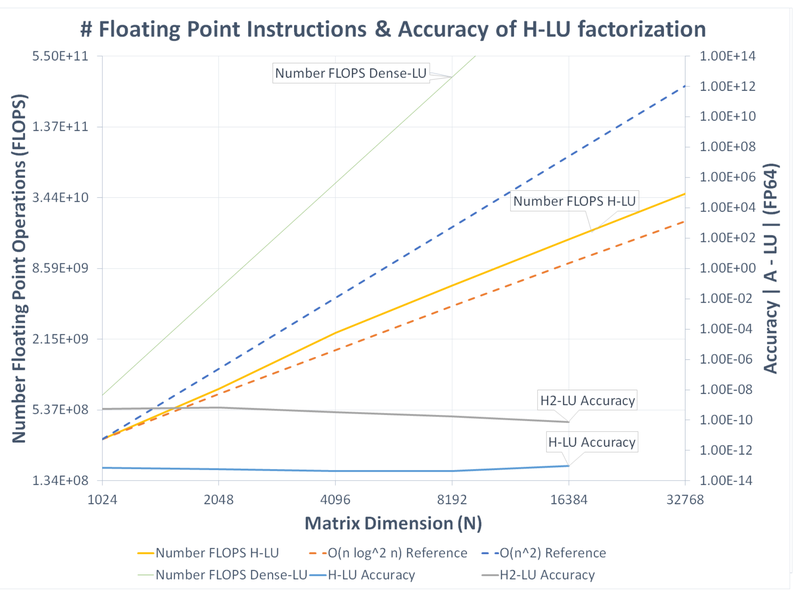


Figure 2: Log Linear Factorization + High Accuracy

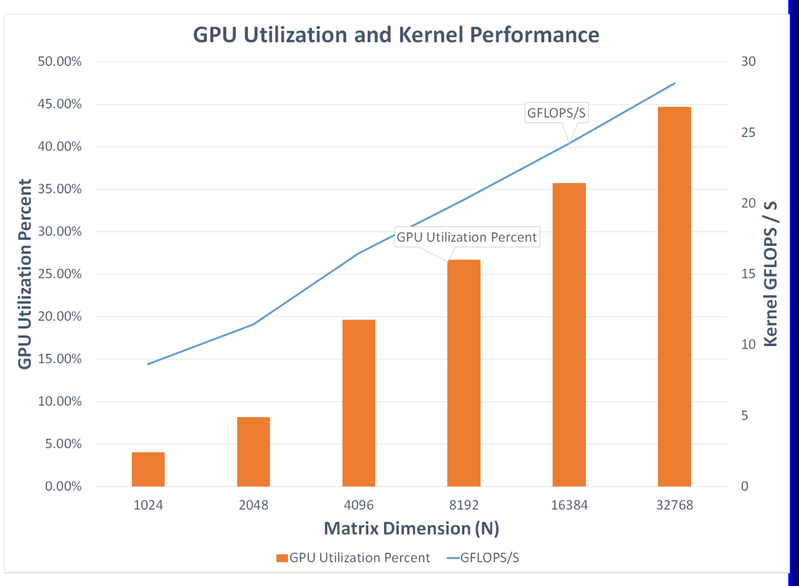


Figure 3: Utilization goes up as problem sizes grow

CONCLUSION & ACKOWLEDGEMENT

- GPUs are typically considered ineffective handling tree structures and recursions, but with enough preprocessing from the CPU, trees can be transformed into batched tasks.
- Hierarchical Low-rank Approximations of matrices compresses the FLOPS required for matrix calculations very significantly, which is not only LU factorization. We believe that our approach could be developed further to accommodate even more kinds of H-matrix calculations.
- The factorization of H2 formats is still under improvement, so only preliminary results are presented.

- This work was supported by JST CREST Grant Number JPMJCR19F5.