ABSTRAK

Penelitian ini menganalisis permasalahan pemborosan sumber daya dalam komputasi heterogen yang menggabungkan CPU dan GPU. Dengan melambatnya Hukum Moore, industri komputasi modern telah beralih ke arsitektur heterogen untuk meningkatkan kinerja. Namun, integrasi CPU dan GPU menciptakan dua masalah signifikan: (1) CPU Idle Time, di mana CPU menganggur saat GPU menjalankan komputasi, dan (2) Underutilization GPU yang disebabkan oleh bottleneck bandwidth PCI-Express yang membatasi transfer data antara memori Host dan Device.

Penelitian ini membandingkan tiga skema eksekusi pada beban kerja General Matrix Multiplication (GEMM): CPU-only, CPU-GPU tanpa stream, dan CPU-GPU dengan CUDA Streams. Pengujian dilakukan pada berbagai dimensi data (matriks 2D, 3D dengan variabel kedalaman, dan 4D dengan variabel batch size) menggunakan hardware NVIDIA RTX 3050 Mobile yang terhubung via PCIe 4.0 x8 pada sistem Intel Core i5-11400H.

Hasil penelitian menunjukkan bahwa skema CPU-GPU dengan CUDA Streams secara konsisten mengungguli kedua skema lainnya. Teknik overlapping data transfer dan N-way concurrency berhasil menyembunyikan latensi transfer PCIe di balik eksekusi kernel, mencapai throughput puncak pada matriks 3D dengan besaran matriks 512 dan pada matriks 4D di besar matriks 512. Penelitian ini memvalidasi bahwa penggunaan CUDA Streams efektif memaksimalkan utilisasi Execution Engine GPU dan core CPU secara bersamaan, mengurangi idle time, dan meningkatkan performa keseluruhan sistem heterogen walaupun efektifitasnya kurang jika digunakan pada matriks skala besar.

Kata Kunci: Komputasi Heterogen, CUDA Streams, Overlapping Transfer Data, Bottleneck PCI-Express, Perkalian Matriks (GEMM), Ko-pemrosesan CPU-GPU, N-way Concurrency, Utilisasi Sumber Daya, RTX 3050 Mobile.

ABSTRACT

Resource underutilization and performance bottlenecks in heterogeneous CPU-GPU systems present critical challenges as traditional scaling approaches reach physical limitations. This study investigates the effectiveness of CUDA Streams in mitigating PCI-Express bandwidth bottlenecks and CPU idle time during heterogeneous computing operations. Three execution paradigms—pure CPU, CPU-GPU without stream, and CPU-GPU with CUDA Streams—are evaluated using General Matrix Multiplication (GEMM) workloads across multiple dimensional configurations: 2D matrices, 3D matrices with variable depth dimensions, and 4D tensors with variable batch sizes. Experiments were conducted on an NVIDIA RTX 3050 Mobile GPU (4GB GDDR6, 2048 CUDA cores) connected via PCIe 4.0 x8 to an Intel Core i5-11400H processor (6 cores/12 threads, 16GB DDR4).

Results demonstrate that CPU-GPU heterogeneous execution with CUDA Streams achieves consistent performance improvements over homogeneous CPU or GPU alternatives through effective overlap of data transfer and kernel execution. Peak throughput values of 696.06 GFLOPs and 636.76 GFLOPs were achieved on 3D and 4D matrices respectively at intermediate matrix sizes (N=512). However, effectiveness diminishes on large-scale matrices where computation time dominates total execution time, resulting in performance convergence between CPU-GPU CUDA Streams and pure GPU implementations. The findings validate that CUDA Streams optimization is context-dependent and most beneficial for memory-bound workloads with moderate data dimensions.

Keywords: Heterogeneous Computing, CUDA Streams, Data Transfer Overlapping, PCI-Express Bottleneck, General Matrix Multiplication, CPU-GPU Co-processing, N-way Concurrency, Resource Utilization, Mobile GPU Architecture.