Form 3

Abstract of Thesis

Name (Yuechao Lu)	
Title	A Study on Reducing the Amount of Out-of-Core Data Access for GPU-Accelerated Applications (GPUにより加速された応用のためのアウトオブコアデータ参照量の削減に関する研究)

Abstract of Thesis

The computing architectures are shifting quickly as we move forward to exascale computing. Heterogeneous computing architectures employing accelerators like graphics processing unit (GPU) have become the mainstream approach for high-performance computing (HPC) systems from supercomputers to mobile devices. However, the increased layers of the memory hierarchy in heterogeneous architectures have hindered the users to extract the full potential and maximum scalability of the hardware. Communication cost for transferring data between memory hierarchy dominates the overall processing time for most applications and the floating-point operations per second (flop/s) has become comparatively irrelevant.

In this work, we look into two applications to address the problems in accelerating applications with GPUs. The first application is the cone beam computed tomography (CT) reconstruction, which is an widely used by medical imaging devices. The second application is a matrix decomposition algorithm called randomized singular value decomposition (RSVD). We show that the CPU-GPU data transfer is the main bottleneck for processing large scale data on GPU-enabled systems. We first propose methods to accelerate RSVD by reducing the data transfer between the CPU and GPU. We then propose algorithms which modify the original RSVD to fit into the heterogeneous computing architecture. The pro- posed methods successfully move the performance bottleneck from CPU-GPU bandwidth bound to compute bound, so that the computation ability of the GPUs can be fully utilized for acceleration.

This thesis is divided into three parts. In the first part of this work, we propose a cache- aware optimization method to accelerate the out-of-core cone beam CT reconstruction on a GPU. Out-of-core data here are data that are too large to fit into the GPU memory at once. Utilizing the GPU in reconstructing CT images has gained its popularity for its high performance and low cost implementation compared to other methods. The proposed method extends a GPU-based previous method by increasing the cache hit rate to speed up the reconstruction of high-resolution volumes that exceed the capacity of GPU memory. More specifically, our approach accelerates the well-known Feldkamp, Davis, and Kress (FDK) algorithm by utilizing the following three strategies: (1) a loop organization strategy that identifies the best trade-o↵ point between the cache hit rate and the number of o↵-chip memory accesses; (2) a data structure that exploits high locality within a layered texture; and (3) a fully pipelined strategy for hiding file input/output (I/O) time of GPU execution and data transfer time. We implement our proposed method on NVIDIA' s Maxwell architecture and provide a tuning guideline for adjusting the execution parameters, which include the granularity and shape of thread blocks as well as the granularity of I/O data to be streamed through the pipeline, which maximizes reconstruction performance. Our experimental results show that it took less than three minutes to reconstruct a 2048^3-voxel volume from 1200 2048^2-pixel projection images on a single GPU; this translates to a speedup of approximately 1.47× as compared to a previous method. We also make clear a trade-off between the texture cache hit rate and the number of memory accesses. Concerning GPU optimization, we found that it is not necessarily efficient to compact as many tasks as possible into kernel execution to decrease total execution time. Instead, proper tuning is required to identify the optimum number of tasks that will minimize the overall time. With the aid of

texture interpolation and cache-aware strategies, our presented GPU implementation achieves performance advantages over other computing platforms.

In the second part, we propose two acceleration methods, namely Fused and Gram, for reducing the out-of-core data access when performing RSVD on GPUs. Both methods accelerate GPU-enabled RSVD using the following three schemes: (1) a highly tuned general matrix-matrix multiplication (GEMM) scheme for processing out-of-core data on GPUs; (2) a data-access reduction scheme based on one-dimensional (1D) data partition; and (3) a first-in, first-out (FIFO) scheme that reduces CPU-GPU data transfer using a reverse iteration. The Fused method further reduces the amount of out-of-core data access by merging two GEMM operations into a single operation. In contrast, the Gram method reduces both in- core (i.e., all the working data can be held on the GPU memory) and out-of-core data access by explicitly forming the Gram matrix. According to our experimental results, the Fused and Gram methods improved the RSVD performance by up to 1.9× and 5.2×, respectively, compared with a straightforward method that deploys schemes (1) and (2) on the GPU. In addition, we present a case study of deploying the Gram method for accelerating robust principal component analysis (RPCA), a convex optimization problem in machine learning.

In the third part, we propose a two-pass RSVD, named block randomized SVD (BRSVD), designed for matrices with a slow-decay singular spectrum that is often observed in image data. BRSVD fully utilizes the power of modern computing system architectures and efficiently processes large-scale data in a parallel and batched fashion. Our experiments show that BRSVD effectively moves the performance bottleneck from data transfer to computation, so that outperforms existing RSVD methods in terms of speed with retaining similar accuracy. We also show an application of randomized SVD to convex RPCA on a GPU, which shows significant speedup in computer vision applications.

Our work demonstrates that communication cost is an important factor to influence the overall performance. The first application demonstrates that cache-aware optimization improves overall performance effectively. The second application shows that reducing the communication cost at the expense of increased computational cost is a viable approach in a computing environment where communication cost exceeds the computational cost. Furthermore, our work shows that redesigning algorithms to fit for the heterogeneous computing architecture is a feasible approach in dealing with bandwidth bound problems.

様式７

論文審査の結果の要旨及び担当者

氏　名　（　Yuechao Lu　）		
	（職）	氏　名
論文審査担当者	主 査　教 授	伊野　文彦
	副 査　教 授	増澤　利光
	副 査　教 授	松下　康之

論文審査の結果の要旨

グラフィクス処理を加速するためのハードウェアGPU（Graphics Processing Unit）は，汎用計算の加速器としても注目されている．GPUは，多くの応用においてCPUに対する高速化を果たしているが，その小さなメモリ容量が大規模問題を加速するときの障壁となっている．したがって，メモリ容量の制約を超えて大規模問題を高速化できれば，加速器としての有用性を高めることができる．本研究では，GPUメモリ容量を超える大規模データ（アウトオブコアデータ）に対する高速処理を実現するために，CPU・GPU間のデータ転送量を削減する手法を提案し，いくつかの実用的な応用を用いて評価している．本学位論文の主要な成果は以下の3点である．

(1) Cone Beam Computed Tomography（CBCT）撮像装置のための再構成問題に対し，GPUのためのキャッシュ最適化技術を提案している．CPU・GPU間のデータ転送量を削減するために，GPUメモリ上のデータに対する参照局所性を高めるだけでなく，データ構造ならびにループ構成を改善することにより，再構成処理の性能ボトルネックをGPUメモリの帯域幅からGPU演算器に移すことを実現している．結果として，提案手法は既存手法を上回る再構成性能を達成できることを示した．

(2) 大規模行列に対する特異値分解問題に対し，CPU・GPU間のデータ転送量を削減することによりGPU上で乱択アルゴリズムを高速化するための手法を提案している．提案手法には2つの方式があり，それらはGPUメモリに転送したデータを再利用することでデータ転送量を削減する方式，ならびにGPUにおける計算量の増加を許容する代わりにデータ転送量を削減する方式である．これらの方式を主成分分析問題に適用し，既存手法に対する計算性能の優位性を示した．

(3) 大規模行列に対する特異値分解問題に対し，近似解を許容するGPU向け乱択アルゴリズムの高速化手法を提案している．提案手法は，対象とする大規模行列が同様の特異値を持つ部分行列で構成されているものと仮定し，大規模行列を列方向に分割し，それらを独立に処理する．これによりCPU・GPU間におけるデータ転送量の削減を実現している．この仮定がいくつかの応用において成り立つことを，厳密解との計算誤差を実験的に検証することで確認している．さらに，顔認証で必要とされる固有顔問題に提案手法を適用し，既存手法に対する計算性能の優位性を示した．

以上のように，本論文はメモリ容量に関して制約の強いGPU上で大規模計算を加速することに関して重要な研究成果をあげており，情報科学の発展に寄与するところが大きい．よって，本論文は博士（情報科学）の学位論文として価値のあるものと認める．