論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	System Software Studies for Seamlessly Overcoming GPU Resource Limitations
著者(和文)	MARKTHUBPAK
Author(English)	Pak Markthub
出典(和文)	学位:博士(理学), 学位授与機関:東京工業大学, 報告番号:甲第10943号, 授与年月日:2018年9月20日, 学位の種別:課程博士, 審査員:松岡 聡,遠藤 敏夫,脇田 建,額田 彰,横田 理央
Citation(English)	Degree:Doctor (Science), Conferring organization: Tokyo Institute of Technology, Report number:甲第10943号, Conferred date:2018/9/20, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

専攻： Department of	数理・計算科学　専攻	申請学位（専攻分野）： Academic Degree Requested	博士　（　理学　） Doctor of
学生氏名： Student's Name	MARKTHUB Pak	指導教員（主）： Academic Supervisor(main)	特任教授　松岡　聡
		指導教員（副）： Academic Supervisor(sub)	

要旨（英文 800 語程度）
Thesis Summary（approx.800 English Words ）

Heterogeneous computing with GPU (Graphic Processing Unit) accelerators is growing in importance in high performance computing (HPC), Machine Learning (ML), and other areas. With the convergence of AI, Big Data, and HPC, GPU has become an indispensable component for thousands of applications. Large number of supercomputers, clusters, and Cloud-service providers are equipped with GPUs to serve the rapidly-growing need of their customers. Various studies and solutions have emerged to address problems regarding using GPU for computation. However, the common treatment that view a GPU as a separate entity from the host machine still leaves many problems unsolvable.

Our study focuses on two important problems that hinder the development of GPU applications and lower the utilization on heterogeneous resource-sharing system. The first problem originates from small GPU memory capacity. Recently, application datasets have expanded beyond the memory capacity of GPUs, and often beyond the capacity of their hosts. In order to process those large datasets, applications have to employ domain decomposition and orchestrate data movement between storage, host memory, and GPU memory. The development becomes more demanding, and in many cases further studies are needed to discover how to partition the datasets and manage the data movement in the memory hierarchy efficiently.

To address the memory capacity problem, we propose DRAGON, a solution that enables all classes of GPU applications to transparently compute on terabytes of datasets residing on Non-Volatile Memory (NVM) storage, while ensuring the integrity of data buffers as necessary. DRAGON leverages the page-faulting mechanism on the recent NVIDIA GPUs by extending the capabilities of CUDA Unified Memory (UM). Further, DRAGON improves overall end-to-end application performance by dynamically optimizing accesses to NVM with direct page-cache usage and two-level prefetching. Our empirical evaluation on an NVIDIA P100 GPU and a Micron 9100 NVMe card using traditional HPC kernels and popular deep-learning workloads shows that DRAGON improves application execution times up to 2.3x compared to standard UM-based execution with conventional file operations and manual data transfers.

The second problem commonly takes place in multi-GPU batch-queue node-sharing systems. The problem stems from the fact that job schedulers treats GPUs as rigid resources. When a batch job requests for a service, it specifies how many GPUs per node it needs. The job schedulers cannot assign any nodes that have fewer number of unoccupied GPUs to the job. This situation occasionally leads to high number of idle GPUs on those systems. We call this problem the scattered idle-GPU problem.

A solution for solving the scattered idle-GPU problem is to virtually increase the number of unoccupied GPUs on a node with GPU remoting. We study the state-of-the-art GPU remoting technology called rCUDA. We mathematically and empirically evaluate rCUDA's overhead in various scenarios and demonstrate that GPU remoting is not enough to efficiently solve the scattered idle-GPU problem. We propose mrCUDA, a GPU middleware for transparently migrating execution from remote GPUs to local GPUs. We mathematically study the overhead of mrCUDA and empirically show that the overhead is negligible on real applications (less than 1%). We then take the first-come-first-serve scheduling algorithm as a case study and demonstrate a simple way to integrate mrCUDA without changing the scheduling policy. We call the mrCUDA-integrated algorithm MRQ. By ways of simulation using both synthetic and recorded job sets in various situations, we show that MRQ can reduce the makespan of a job by up to 30% with insignificant increasing in the execution time (less than 1%) on average.

This thesis provides several contributions to communities that use GPU for computation. The study highlights a few approaches from the system level that can transparently, universally, and efficiently solve the limited GPU memory capacity and the scattered idle-GPU problems.

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800