論文 / 著書情報
Article / Book Information

題目(和文)	FPGAとOpenCLによる高性能計算
Title(English)	High Performance Computing with FPGAs and OpenCL
著者(和文)	ハミド レザ ゾフーリ
Author(English)	Hamid Reza ZOHOURI
出典(和文)	学位:博士(学術), 学位授与機関:東京工業大学, 報告番号:甲第11010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,額田 彰,横田 理央,中原 啓貴
Citation(English)	Degree:Doctor (Academic), Conferring organization: Tokyo Institute of Technology, Report number:甲第11010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	Mathematical and Computing Sciences	専攻	申請学位（専攻分野）： Academic Degree Requested	博士 Doctor of	（Philosophy）
学生氏名： Student's Name	Hamid Reza ZOHOURI		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 ）

With the impending death of Moore's law, the High Performance Computing (HPC) community is actively exploring new options to satisfy the never-ending need for faster and more power efficient means of computing. Even though GPUs have been widely employed in world-class supercomputers in the past few years to accelerate different types of computation, the high power usage and slow power efficiency improvements of these devices remains a limiting factor in deploying larger supercomputers, especially on the path to Exascale computing. Field-Programmable Gate Arrays (FPGAs) are an emerging alternative to GPUs for this purpose. These devices, despite being older than GPUs, have rarely been used in the HPC industry and have mainly been employed in embedded and low-power markets. Since the traditional applications of FPGAs have vastly different requirements compared to typical HPC application, the usability, productivity and performance of FPGAs for HPC applications is unknown.

In this study, our aim is to determine whether FPGAs can be considered as a viable solution for accelerating HPC applications, and if so, how they fare against existing processors in terms of performance and power efficiency in different HPC workloads. We take advantage of the recent improvements in High Level Synthesis (HLS) that, unlike traditional Hardware Description Languages (HDL) that are known to be notoriously hard to use and debug, allow FPGAs to be more easily programmed by software programmers using familiar software programming languages. Specifically, we use Intel FPGA SDK for OpenCL that allows modern Intel FPGAs to be programmed as an accelerator, similar to GPUs.

In the first step, we evaluate the performance and power efficiency of FPGAs in different benchmarks, each being a representative of a specific HPC workload. For this purpose, we port a subset of the Rodinia benchmark suite for two generations of Intel FPGAs, and then optimize each benchmark based on the specific architectural characteristics of these FPGAs. Then, we compare the performance and power efficiency of these devices against same-generation CPUs and GPUs. We show that even though a direct port of CPU and GPU kernels for FPGAs usually performs poorly on these devices, with FPGA-specific optimizations, up to two orders of magnitude performance improvement can be achieved, resulting in better performance to that of CPUs in all cases, and competitive performance to that of GPUs in most. Furthermore, we show that FPGAs have a clear power efficiency edge

in every case, reaching up to 16.7 and 5.6 times higher power efficiency compared to their same-generation CPUs and GPUs, respectively.

Based on our experience from the initial evaluation, we determine that for stencil computation, which is one of the most important computation patterns in HPC, FPGAs can not only compete with GPUs in terms of power efficiency, but also in terms of pure performance. Taking advantage of the unique architectural advantages of FPGAs for stencil computation, we design and implement a parameterized OpenCL-based template kernel that can be used to accelerate 2D and 3D star-shaped stencils on FPGAs regardless of stencil order. Our design, apart from using both spatial and temporal blocking, also employs multiple HLS-specific optimizations for FPGAs to maximize performance. Moreover, we devise a performance model that allows us to quickly tune the performance parameters of our design, significantly reducing the time and resources necessary for placement and routing. We show that our design allows FPGAs to achieve superior performance to that of CPUs, GPUs and Xeon Phi devices in 2D stencil computation, and competitive performance for 3D. Specifically, we show that using an Intel Arria 10 GX 1150 device, for 2D and 3D star-shaped stencils of first to fourth-order, we can achieve over 700 and 270 GFLOP/s of compute performance, respectively. Furthermore, we show that our implementation outperforms all existing implementations of stencil computation on FPGAs.

This thesis makes multiple contributions to the emerging field of using FPGAs in HPC, and the optimization techniques discussed in this work can be used as guidelines for optimizing most types of applications on FPGAs using HLS, even for non-HPC applications.

（博士課程）
Doctoral Program

東京工業大学
Tokyo Institute of Technology