論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Cross-Architecture Performance Prediction Using Collaborative Filtering
著者(和文)	Salaria Shweta
Author(English)	Shweta Salaria
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11318号, 授与年月日:2019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 (Academic), Conferring organization: Tokyo Institute of Technology, Report number:甲第11318号, Conferred date:2019/9/20, 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	Shweta SALARIA	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 ）

Over the last decade, in response to the slowing or end of Moore's law, the High Performance Computing (HPC) community has turned towards heterogeneous systems. The trend in the recent decade has been to extend the application of HPC systems beyond the computationally intensive scientific domain to data-intensive domains, or the domain of Big Data. While HPC has its roots in solving compute-intensive scientific and large-scale distributed problems, Big Data problems have been proven to benefit from typical HPC environments: high processing power, low-latency networks, and non-blocking communications. However, the use of HPC systems is limited to scientists and researchers who have access to supercomputing labs and centers. With HPC extending its application space from scientific applications to big data analytics, the increasing demand for resources in HPC centers may not be met immediately. Hence, it is important to study alternative HPC solutions and the extent to which these solutions can be employed for different classes of computing and their applications.

In the first step, we performed the evaluation of an HPC mini-app, NICAM-DC-MINI and data-intensive benchmark, Graph500 on three systems: 1) our in-house production supercomputer TSUBAME2.5, 2) the energy-efficient TSUBAME-KFC supercomputer and 3) AWS EC2's latest generation of compute-optimized instances, C4. We conducted this study in an attempt to measure the capabilities of C4 instances for representative HPC and Big Data applications. Our finding demonstrates that Amazon EC2 C4 instances exhibit good compute performance with low performance variability. The 10 Gbps Ethernet network in these instances is the chief bottleneck for scaling the workloads. So, C4 instances can be a good fit for compute-intensive and memory-intensive application with little communication. We optimized the Graph500 benchmark for C4 instances by finding the optimal number of threads and MPI ranks and achieved performance comparable to our in-house supercomputers. Similarly, we achieved good performance using both 10 and 20 MPI ranks per C4 instance for NICAM-DC-MINI. Thus, C4 instances can be adopted as a replacement to on-premises hardware deployments for such workloads.

Performance prediction of scientific applications across systems becomes increasingly important in today's diverse computing environments. A wide range of choices in execution platforms pose new challenges to researchers in choosing a system which best fits their workloads and

administrators in scheduling applications to the best performing systems. While previous studies have employed simulation- or profile-based prediction approaches, such solutions are time-consuming to be deployed on multiple platforms. To address this problem, we use two collaborative filtering techniques to build analytical models, which can quickly and accurately predict the performance of workloads across different multicore systems. The first technique leverages information gained from performance observed for certain applications on a subset of systems and use it to discover similarities among applications as well as systems. The second collaborative filtering based model learns latent features of systems and workloads automatically and uses these features to characterize the performance of applications on different platforms. We evaluated both the methods using 30 workloads chosen from NAS Parallel Benchmarks, BOTS and Rodinia benchmark suites on ten different systems. Our results show that such collaborative filtering methods can make predictions with RMSE as low as 0.6 and with an average RMSE of 1.6.

The graphic processing units (GPUs) have become a primary source of heterogeneity in today's computing systems. With the rapid increase in number and types of GPUs available, finding the best hardware accelerator for each application is a challenge. For that matter, it is time consuming and tedious to execute every application on every GPU system to learn the correlation between application properties and hardware characteristics. To address this problem, we extend our previously proposed collaborating filtering based modeling technique, to build an analytical model which can predict performance of applications across different GPU systems. Our model learns representations, or embeddings (dense vectors of latent features) for applications and systems and uses them to characterize the performance of various GPU-accelerated applications. We improve state-of-the-art collaborative filtering approach based on matrix factorization by building a multi-layer perceptron. In addition to increased accuracy in predicting application performance, we can use this model to simultaneously predict multiple metrics such as rates of memory access operations. We evaluate our approach on a set of 30 well-known micro-applications and seven NVIDIA GPUs. As a result, we can predict expected instructions per second value with 90.6% accuracy in average.