Ph.D. in Information Technology
Thesis Defense

January 19th, 2024
at 9:00 am
Sala Seminari Nicola Schiavoni

Guido Walter DI DONATO – XXXVI Cycle

LEVERAGING HETEROGENEOUS HARDWARE ACCELERATION FROM HIGH-LEVEL PROGRAMMING LANGUAGES: THE CASE FOR BIOMEDICAL INFORMATICS

Supervisor: Prof. **Marco Domenico Santambrogio**

Abstract:

In recent years, the field of biomedical informatics has witnessed an explosion in data volume and complexity, demanding innovative approaches to accelerate computational tasks. Graphics Processing Units (GPUs) have emerged as potent hardware accelerators, driving substantial performance gains in various domains. However, their adoption within biomedical informatics has been mainly limited to the deep learning domain, due to the availability of libraries that mask the GPU computation. This is mainly due to the scarce integration of GPUs with high-level programming languages, such as R or Python, commonly used in the biomedical context. This dissertation addresses the critical need to extend GPU utilization to such high-level languages in the context of biomedical informatics.

The first part of this thesis presents a series of applications that leverage GPUs to expedite the training of deep learning models for different tasks, namely drug repurposing, dementia detection, and lung cancer identification. By harnessing the parallel processing capabilities of GPUs, significant reductions in training times are achieved, thereby facilitating quicker iterations, and enhancing the pace of research and clinical decision-making. Nonetheless, the aforementioned applications involve other compute-intensive processing steps that could benefit from heterogeneous hardware acceleration. A bridge between high-level languages and GPU programming is imperative to enable researchers and practitioners to exploit the full potential of GPU hardware in biomedical informatics applications.

To address such a need, the centerpiece of this dissertation presents an innovative framework to simplify GPU programming, named GrCUDA. Initially developed by NVIDIA and Oracle, GrCUDA serves as an intermediary layer that enables seamless access to GPU resources from the high-level programming languages supported by the polyglot GraalVM ecosystem. This thesis introduces a novel multi-GPU asynchronous scheduler for GrCUDA, which facilitates the development of multi-GPU applications by abstracting low-level GPU details and providing a high-level interface that is both efficient and user-friendly. Experimental results prove that the proposed scheduler transparently provides speedups comparable to what an expert programmer can achieve by hand, making multi-GPU computations easier to approach while minimizing performance compromises. Thus, the GrCUDA framework emerges as a crucial bridge, enabling biomedical informatics researchers to harness GPUs' substantial computational capabilities without compromising programming ease and language flexibility.

Building upon the GrCUDA framework, the final part of the thesis showcases two applications that underscore its utility. The first application introduces GPJSON, a GPU-accelerated JavaScript Object Notation (JSON) processing engine, exemplifying how GrCUDA can enhance data parsing and manipulation tasks prevalent in biomedical informatics. The second application presents a sequence-to-graph aligner tailored for genomics applications, demonstrating how GrCUDA can handle complex systems inherent to the domain.

In conclusion, this dissertation advocates for the integration of GPUs into high-level programming languages to address the computational challenges in biomedical informatics. By enhancing the GrCUDA framework, this work provides a novel solution that empowers researchers and practitioners to harness the full potential of GPUs while developing applications for critical tasks within the biomedical field. Through concrete applications and empirical results, this thesis underscores the transformative impact of GPU acceleration on the landscape of biomedical informatics and, subsequently, of precision medicine.

Alberto ZENI – XXXVI Cycle

A FRAMEWORK FOR THE AIDED DESIGN OF HIGH-PERFORMANCE GENOME ANALYSIS APPLICATIONS ON HETEROGENEOUS ARCHITECTURES

Supervisor: Prof. **Marco Domenico Santambrogio**

Abstract:

Technological innovation and the declining cost of Next Generation Sequencing (NGS) have driven an explosion in the quantity of genomic research.
However, analyzing the massive amount of sequencing data generated from this research has revealed a great computational challenge.
The rate of data generation in genomics is outpacing the rate at which it can be processed,
and demand more computational power than what current processors can deliver.
Indeed, Central Processing Unit (CPU) performance is not following the same trend as it is becoming extremely difficult to integrate more transistors on a chip.
Consequently, as we reach the end of Moore's Law predictions, we need new architectural solutions to satisfy the continuously growing performance demand of High-Performance Computing (HPC) applications like genomic algorithms.
In this context, hardware, namely Graphics Processing Units (GPUs) and Field Programmable Gate Arrays (FPGAs), incarnate
an effective solution to offload compute-intensive tasks from the CPU.
However, in order to fully exploit the computing power of GPUs and FPGAs, applications must be developed from the ground up, specifically targeting these architectures.
Moreover, the process of developing highly performing heterogeneous applications still requires both domain-specific knowledge and expertise to leverage the architecture effectively.
For this reason, the focus of this dissertation is a framework for the development of genomic applications exploiting high-performance computing with heterogeneous hardware architectures.
Indeed, the objective of this work is to both provide a customizable, fully hardware-accelerated genomic pipeline and help the end user when developing genomic applications.
The key aspect of this work is to focus on flexibility; therefore, the integrated pipeline provides multiple implementations for each pipeline stage that are compatible with a wide range of GPUs and FPGAs.
Furthermore, knowing the rapid change in the scope of algorithms, genomic pipelines need to be frequently updated.
Because of this, the framework enables and provides guidance for users to implement their own applications on GPU.
For these reasons, this dissertation explores the crucial issues and peculiar characteristics of the different architectures employed in the framework.
It begins with a Chapter providing an overview of genome assembly and heterogeneous architectures.

Then, the dissertation dives into the framework description, providing an in-depth analysis of every accelerated step of the genome assembly pipeline.
For each step, the dissertation includes design methodologies, design innovations, literature overview, and performance analysis.
Finally, this dissertation depicts how the framework can be exploited for developing highly performance GPU algorithms that can be exploited in the genome-assembly pipeline, providing various performance analyses and examples to prove its effectiveness.

PhD Committee

Prof. **Michele Carminati,** Politecnico di Milano

Prof. **Dionysios Pnevmatikatos,** National Technical University of Athens

Prof. **Juergen Becker**, Karlsruhe Institute of Technology