SYSTEM SOFTWARE IN EDGE GPU PLATFORMS FOR URBAN SUSTAINABILITY APPLICATIONS IN DEVELOPING COUNTRIES

OMAIS SHAFI

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
March 2025

SYSTEM SOFTWARE IN EDGE GPU PLATFORMS FOR URBAN SUSTAINABILITY APPLICATIONS IN DEVELOPING COUNTRIES

by

OMAIS SHAFI

Department of Computer Science and Engineering

Submitted

in fulfillment of the requirements of the degree of **Doctor of Philosophy**

to the

Indian Institute of Technology Delhi

March 2025

To Abu, Mummy, Ubaid baya
Thanks to their unwavering support, patience, and belief in me.

Certificate

This is to certify that the thesis titled **System Software in Edge GPU Platforms for Urban Sustainability Applications in Developing Countries**, submitted by **Mr. Omais Shafi**, to the Indian Institute of Technology, Delhi, for the award of the degree of **Doctor of Philosophy** in Computer Science & Engineering, is a record of bonafide work carried out by him under my guidance and supervision at the Department of Computer Science & Engineering, Indian Institute of Technology Delhi. The contents of this thesis, in full or in parts, have not been submitted elsewhere for the award of any degree or diploma.

Dr Rijurekha Sen
Assistant Professor
Dept. of Computer Science and Engineering
Indian Institute of Technology Delhi
New Delhi-110016

Acknowledgements

For indeed, with hardship [will be] ease.

– Quran 94:5/6

All praise is due to Allah, the Almighty, upon whom we depend for sustenance and guidance. I am deeply grateful to Him for granting me the resilience and fortitude needed to persevere until the very end.

I am profoundly thankful to my advisor, Dr.Rijurekha Sen, for her timely and unwavering support, both personally and professionally. Throughout the most challenging periods of my Ph.D. journey, particularly when I faced significant obstacles and uncertainty, her steadfast assistance and encouragement were invaluable. Dr. Sen provided not only insightful guidance in my research but also the motivation and moral support needed to persevere. During times of family commitments or work obligations, her patience and constant push for my Ph.D. work were invaluable. Thank you Riju for being a *great mentor*.

I would also like to express my gratitude to my collaborators, Dr.Gayathri Ananthanarayanan, Dr.Aastha Mehta, Dr.Arpan Gujarati, Dr.Khalid Pandit, and Sachin for their valuable discussions and feedback. Regular meetings with them at various stages of my Ph.D. were crucial. I extend my thanks to my SRC members— Prof.Anshul Kumar, Prof.Kolin Paul, and Prof.Sumantra Dutta Roy—for serving on my committee and providing valuable feedback. Additionally, I am grateful to Prof.Preeti Ranjan Panda for his support during tough times and to Prof.Smruti Ranjan Sarangi for his assistance in the early years of my Ph.D. Thanks also to Chinmai and Amarjeet for their help with many experimental results.

The journey of a Ph.D. cannot be completed without a good circle of friends, and I was fortunate to have many. Mehreen deserves a special mention for her constant motivation, companionship, and all the tea and delicious food we shared. She witnessed my ups and downs and supported me like a rock. She has been my go to person and I owe her a big thanks. I want to thank Janibul Bashir for being a wonderful mentor and brother throughout the earlier years of my

Ph.D. journey. His valuable advice and support were unwavering. I also want to acknowledge Nadeem, Danish, and Zarkab. The times spent at Nilgiri hostel discussing Ph.D. life and everyday matters were some of the best moments of my journey. Nadeem, in particular, was a constant source of support when we were roommates, always taking care of my belongings and being there for me. I would also like to thank Zamir Wani for his valuable advice throughout my Ph.D. and for pushing me through the tough phases when I needed it the most. Thanks to Asifa for all her advices and for listening to my long conversations, and for the unforgettable dinners at Nilgiri hostel and the food she prepared for us.

I have no words to thank Sandeep for his support during my Ph.D. He is one of the best human beings I have come across and always pushed me forward. I will never forget our meetings at CCD, the Lebanese chicken lunches at the SIT faculty lounge, and all the outings we had together. Thanks also to Vinayak and Dishant for the wonderful outings and times shared with Sandeep. I am thankful to Ismi for being there for all the good discussions and motivation she used to give me during my Phd. Additionally, I am grateful to Shubankar, Hameedah, Diksha, and Himanshu for the enriching discussions and memorable times. I also extend my gratitude to my Kashmiri friends— Omar, Adnan, Mujeeb, Saleem, Malik, Shahid, Kaiser,Amir, Faizan, Faiz, Saqib and others. The tea and dinner parties we shared will always be cherished. I may have missed mentioning few names, but I am grateful to everyone for the cherished memories.

Imagining a journey like a Ph.D. without family support is difficult. I thank my entire family for their support and patience over these years. Thank you for your prayers and blessings. I am grateful to all my cousins for always being there when I needed them. I thank my grandparents for their blessings and support. I wish all my grandparents were here to see my growth, but I am sure those who are no longer with us are still sending their blessings.

Finally, I dedicate this thesis to three people without whom this work would never have materialized: my father, my mother, and my brother. Thank you for your unconditional support, encouragement, and the freedom you gave me at every phase of my life. I cannot possibly list everything I want to thank you for. Everything I know or have achieved is because of your efforts. My parents have gone through the toughest phases for providing me with the good education and I have no words to thank them for everything they have done for me. I owe them everything. Special mention to my brother, Ubaid Baya, who sacrificed his career by letting me study outside our hometown while he continued working there despite having job opportunities elsewhere. I owe him a lot and probably can never repay it back. Thank you, Abu, Mummy, and Ubaid Baya, for everything.

A special note of appreciation to my wife, Dr.Hadiya. Meeting you during the final stages of

my PhD has been a blessing. Your love, encouragement, and understanding have given me strength and joy.

The greatest glory in living lies not in never falling, but in rising every time we fall.

–Nelson Mandela

Omais Shafi

Abstract

Road traffic congestion increases vehicular emissions and air pollution, significantly impacting urban sustainability. The excessive pollution contributes to poor air quality and adverse health effects, while the economic costs associated with congestion and accidents strain city resources. In developing countries, where the vehicle-to-road infrastructure ratio is imbalanced, these issues are further magnified. Effective management of traffic and reduction of emissions are crucial for promoting sustainable urban development and improving the quality of life for city dwellers. Additionally, addressing these challenges is essential for reducing the carbon footprint of urban areas, mitigating climate change, and fostering resilient and livable cities. Implementing innovative traffic management solutions and investing in green infrastructure can enhance urban sustainability and create healthier, more efficient environments.

To address these challenges, advanced computational tools are being leveraged to improve road traffic measurement and management. Cameras capture video frames, which are then processed using state-of-the-art machine learning (ML) and deep neural network (DNN) models to analyze traffic patterns and enforce regulations. However, running these heavy computational models on the cloud is challenging due to the lack of reliable broadband connectivity in developing countries. An alternative is to run these models on edge devices equipped with advanced GPU platforms. This thesis explores the system software stack for edge GPU platforms, focusing on both compile-time and runtime aspects to optimize performance and manage computational demands effectively.

This thesis examines two types of system software essential for optimizing neural network inference on Nvidia edge GPUs. The first type is compile-time system software, which processes pre-trained ML models by applying various model compression techniques to generate efficient engines for inferences. The second type is runtime system software, which manages resource utilization, thermal safety, and other critical factors as the ML models perform continuous inferences. For both compile-time and runtime software, the thesis carefully audits existing software stacks from industry and academia, identifying their limitations. Based on this anal-

ysis, new software solutions are designed and implemented to address these shortcomings and enhance overall system performance. Additionally, the thesis explores various aspects of system software for Nvidia edge GPUs, including techniques for managing application dynamism and ambient temperature, and addresses the challenges of deploying applications from multiple vendors.

We first investigate key software frameworks for compressing pre-trained neural networks for edge inference, focusing on TensorRT, AutoTVM, and AutoScheduler. We develop a comprehensive set of metrics to compare these frameworks and design careful experiments to empirically measure those metrics. Our analysis of TensorRT showed notable performance and accuracy improvements but also revealed unexpected non-deterministic behaviors such as varying outputs and increased latency on more powerful hardware. We also evaluated vendor-agnostic frameworks AutoTVM and AutoScheduler, comparing them on performance predictability, inference latency, output reproducibility, and hardware utilization. This comparison led to several interesting observations which have important implications for urban sustainability applications like traffic monitoring. These insights provide recommendations for selecting the most effective DNN compiler for real-time, safety-critical applications, ensuring precise and predictable inferences.

In addition, to address the challenges of managing high throughput and ensuring system longevity in the extreme conditions of developing regions, we developed DynCNN, a controller designed for application dynamism and ambient temperature management. DynCNN leverages processor heterogeneity to regulate thread counts and accelerator frequencies, optimizing application performance while adhering to strict thermal and power constraints. We evaluated DynCNN on three commercially available embedded GPUs using a 40-day dataset from a real traffic intersection, demonstrating its effectiveness in enhancing system efficiency under these demanding conditions.

Furthermore, edge devices face challenges in deploying applications from multiple vendors, each with its own software dependencies. Using Docker containers provides isolation and manages these dependencies, enhancing security and portability. However, applications running together can affect each other's performance and increase device temperature due to shared resource contention. This necessitates advanced scheduling to ensure optimal performance and thermal management. To address these issues, we propose IceEdge, an inference serving framework for CPU-GPU based edge computing servers. IceEdge uses containers for isolation and includes a deadline- and thermal-aware scheduler for ML workloads, ensuring good GPU utilization and maintaining quality-of-service. IceEdge also incorporates a controller that precisely manages CPU-GPU operations, avoiding reliance on OS thermal control and thus

improving efficiency and stability.

In conclusion, this thesis investigates the deployment of deep neural networks (DNNs) on Nvidia edge GPU platforms for sustainability applications in developing countries. Our objective is to provide a comprehensive solution that enhances the longevity, performance, and reliability of edge computing systems by addressing critical challenges such as compiler selection, thermal management, and multi-vendor application deployment. The insights and recommendations from our research will aid in the creation of robust and efficient edge computing solutions, fostering sustainable development and improving the quality of life in underdeveloped regions.

अमूर्त

सड़क यातायात की भीड़भाड़ से वाहनों से होने वाले उत्सर्जन और वायु प्रदूषण में वृद्धि होती है, जिससे शहरी स्थिरता पर महत्वपूर्ण प्रभाव पड़ता है। अत्यधिक प्रदूषण खराब वायु गुणवत्ता और प्रतिकूल स्वास्थ्य प्रभावों में योगदान देता है, जबकि भीड़भाड़ और दुर्घटनाओं से जुड़ी आर्थिक लागत शहर के संसाधनों पर दबाव डालती है। विकासशील देशों में, जहाँ वाहन-से-सड़क अवसंरचना अनुपात असंतुलित है, ये मुद्दे और भी बढ़ जाते हैं। यातायात का प्रभावी प्रबंधन और उत्सर्जन में कमी सतत शहरी विकास को बढ़ावा देने और शहरवासियों के जीवन की गुणवत्ता में सुधार करने के लिए महत्वपूर्ण है। इसके अतिरिक्त, शहरी क्षेत्रों के कार्बन पदचिह्न को कम करने, जलवायु परिवर्तन को कम करने और लचीले और रहने योग्य शहरों को बढ़ावा देने के लिए इन चुनौतियों का समाधान करना आवश्यक है। अभिनव यातायात प्रबंधन समाधानों को लागू करना और हरित अवसंरचना में निवेश करना शहरी स्थिरता को बढ़ा सकता है और स्वस्थ, अधिक कुशल वातावरण बना सकता है।

इन चुनौतियों का समाधान करने के लिए, सड़क यातायात माप और प्रबंधन में सुधार के लिए उन्नत कम्प्यूटेशनल उपकरणों का लाभ उठाया जा रहा है। कैमरे वीडियो फ़्रेम कैप्चर करते हैं, जिन्हें ट्रैफ़िक पैटर्न का विश्लेषण करने और नियमों को लागू करने के लिए अत्याधुनिक मशीन लर्निंग (ML) और डीप न्यूरल नेटवर्क (DNN) मॉडल का उपयोग करके संसाधित किया जाता है। हालांकि, विकासशील देशों में विश्वसनीय ब्रॉडबैंड कनेक्टिविटी की कमी के कारण क्लाउड पर इन भारी कम्प्यूटेशनल मॉडल को चलाना चुनौतीपूर्ण है। एक विकल्प इन मॉडलों को उन्नत GPU प्लेटफ़ॉर्म से लैस एज डिवाइस पर चलाना है। यह थीसिस प्रदर्शन को अनुकूलित करने और कम्प्यूटेशनल मांगों को प्रभावी ढंग से प्रबंधित करने के लिए संकलन-समय और रनटाइम दोनों पहलुओं पर ध्यान केंद्रित करते हुए एज GPU प्लेटफ़ॉर्म के लिए सिस्टम सॉफ़्टवेयर स्टैक की खोज करती है।

यह थीसिस Nvidia एज GPU पर न्यूरल नेटवर्क अनुमान को अनुकूलित करने के लिए आवश्यक दो प्रकार के सिस्टम सॉफ़्टवेयर की जाँच करती है। पहला प्रकार संकलन-समय सिस्टम सॉफ़्टवेयर है, जो अनुमानों के लिए कुशल इंजन बनाने के लिए विभिन्न मॉडल संपीड़न तकनीकों को लागू करके पूर्व-प्रशिक्षित ML मॉडल को संसाधित करता है। दूसरा प्रकार रनटाइम सिस्टम सॉफ़्टवेयर है, जो संसाधन उपयोग, थर्मल सुरक्षा और अन्य महत्वपूर्ण कारकों का प्रबंधन करता है क्योंकि ML मॉडल निरंतर अनुमान लगाते हैं। संकलन-समय और रनटाइम सॉफ़्टवेयर दोनों के लिए, थीसिस उद्योग और शिक्षा जगत से मौजूदा सॉफ़्टवेयर स्टैक का सावधानीपूर्वक ऑडिट करती है, उनकी सीमाओं की पहचान करती है। इस विश्लेषण के आधार पर, इन कमियों को दूर करने और समग्र सिस्टम प्रदर्शन को बढ़ाने के लिए नए सॉफ़्टवेयर समाधान डिज़ाइन और कार्यान्वित किए जाते हैं। इसके अतिरिक्त, थीसिस Nvidia एज GPU के लिए सिस्टम सॉफ़्टवेयर के विभिन्न पहलुओं की खोज करती है, जिसमें एप्लिकेशन की गतिशीलता और परिवेश के तापमान को प्रबंधित करने की तकनीकें शामिल हैं, और कई विक्रेताओं से एप्लिकेशन तैनात करने की चुनौतियों का समाधान करती है।

हम सबसे पहले एज इंफ़रेंस के लिए पूर्व-प्रशिक्षित तंत्रिका नेटवर्क को संपीड़ित करने के लिए प्रमुख सॉफ़्टवेयर फ़्रेमवर्क की जाँच करते हैं, जिसमें TensorRT, AutoTVM और AutoScheduler पर ध्यान केंद्रित किया जाता है। हम इन फ़्रेमवर्क की तुलना करने के लिए मेट्रिक्स का एक व्यापक सेट विकसित करते हैं और उन मेट्रिक्स को अनुभवजन्य रूप से मापने के लिए सावधानीपूर्वक प्रयोग डिज़ाइन करते हैं। TensorRT के हमारे विश्लेषण ने उल्लेखनीय प्रदर्शन और सटीकता में सुधार दिखाया, लेकिन अधिक शक्तिशाली हार्डवेयर पर अलग-अलग आउटपुट और बढ़ी हुई विलंबता जैसे अप्रत्याशित गैर-निर्धारक व्यवहार भी प्रकट किए। हमने विक्रेता-अज्ञेय फ़्रेमवर्क AutoTVM और AutoScheduler का भी मूल्यांकन किया, उनकी तुलना प्रदर्शन की भविष्यवाणी, अनुमान विलंबता, आउटपुट पुनरुत्पादन और हार्डवेयर उपयोग पर की। इस तुलना से कई दिलचस्प अवलोकन हुए, जिनका ट्रैफ़िक मॉनिटरिंग जैसे शहरी स्थिरता अनुप्रयोगों के लिए महत्वपूर्ण निहितार्थ हैं। ये जानकारियां वास्तविक समय,

सुरक्षा-महत्वपूर्ण अनुप्रयोगों के लिए सबसे प्रभावी DNN कंपाइलर का चयन करने के लिए सिफारिशें प्रदान करती हैं, जिससे सटीक और पूर्वानुमानित निष्कर्ष सुनिश्चित होते हैं।

इसके अलावा, विकासशील क्षेत्रों की चरम स्थितियों में उच्च थ्रूपुट को प्रबंधित करने और सिस्टम की दीर्घायु सुनिश्चित करने की चुनौतियों का समाधान करने के लिए, हमने DynCNN विकसित किया, जो अनुप्रयोग गतिशीलता और परिवेश तापमान प्रबंधन के लिए डिज़ाइन किया गया एक नियंत्रक है। DynCNN थ्रेड काउंट और एक्सेलेरेटर आवृत्तियों को विनियमित करने के लिए प्रोसेसर विषमता का लाभ उठाता है, सख्त थर्मल और पावर बाधाओं का पालन करते हुए अनुप्रयोग प्रदर्शन को अनुकूलित करता है। हमने वास्तविक ट्रैफ़िक इंटरसेक्शन से 40-दिवसीय डेटासेट का उपयोग करके तीन व्यावसायिक रूप से उपलब्ध एम्बेडेड GPU पर DynCNN का मूल्यांकन किया, जो इन मांग वाली स्थितियों के तहत सिस्टम दक्षता बढ़ाने में इसकी प्रभावशीलता को प्रदर्शित करता है।

इसके अलावा, एज डिवाइस को कई विक्रेताओं से एप्लिकेशन तैनात करने में चुनौतियों का सामना करना पड़ता है, जिनमें से प्रत्येक की अपनी सॉफ़्टवेयर निर्भरताएँ होती हैं। Docker कंटेनर का उपयोग अलगाव प्रदान करता है और इन निर्भरताओं को प्रबंधित करता है, जिससे सुरक्षा और पोर्टेबिलिटी बढ़ती है। हालाँकि, एक साथ चलने वाले एप्लिकेशन एक-दूसरे के प्रदर्शन को प्रभावित कर सकते हैं और साझा संसाधन विवाद के कारण डिवाइस का तापमान बढ़ा सकते हैं। इसके लिए इष्टतम प्रदर्शन और थर्मल प्रबंधन सुनिश्चित करने के लिए उन्नत शेड्यूलिंग की आवश्यकता होती है। इन मुद्दों को संबोधित करने के लिए, हम CPU-GPU आधारित एज कंप्यूटिंग सर्वर के लिए एक अनुमान सेवारत फ्रेमवर्क, IceEdge का प्रस्ताव करते हैं। आइसएज आइसोलेशन के लिए कंटेनर का उपयोग करता है और इसमें एमएल वर्कलोड के लिए डेडलाइन- और थर्मल-अवेयर शेड्यूलर शामिल है, जो अच्छे GPU उपयोग को सुनिश्चित करता है और सेवा की गुणवत्ता को बनाए रखता है। आइसएज में एक नियंत्रक भी शामिल है जो CPU-GPU संचालन को सटीक रूप से प्रबंधित करता है, OS थर्मल नियंत्रण पर निर्भरता से बचता है और इस प्रकार दक्षता और स्थिरता में सुधार करता है। गतिशील कार्यभार और एज प्लेटफ़ॉर्म की थर्मल सुरक्षा बाधाओं दोनों पर विचार करते हुए क्लॉक फ़्रीक्वेंसी।

निष्कर्ष में, यह थीसिस विकासशील देशों में स्थिरता अनुप्रयोगों के लिए Nvidia एज GPU प्लेटफ़ॉर्म पर डीप न्यूरल नेटवर्क (DNN) की तैनाती की जाँच करती है। हमारा उद्देश्य एक व्यापक समाधान प्रदान करना है जो संकलक चयन, थर्मल प्रबंधन और बहु-विक्रेता अनुप्रयोग परिनियोजन जैसी महत्वपूर्ण चुनौतियों का समाधान करके एज कंप्यूटिंग सिस्टम की दीर्घायु, प्रदर्शन और विश्वसनीयता को बढ़ाता है। हमारे शोध से प्राप्त अंतर्दृष्टि और सिफारिशें मजबूत और कुशल एज कंप्यूटिंग समाधानों के निर्माण, सतत विकास को बढ़ावा देने और अविकसित क्षेत्रों में जीवन की गुणवत्ता में सुधार करने में सहायता करेंगी।

Contents

List of Figures

List of Tables