DESIGN OF SECURE HARDWARE ARCHITECTURES FOR NEURAL NETWORKS

NIVEDITA SHRIVASTAVA

INDIAN INSTITUTE OF TECHNOLOGY DELHI
JULY 2025

DESIGN OF SECURE HARDWARE ARCHITECTURES FOR NEURAL NETWORKS

by

NIVEDITA SHRIVASTAVA

Department of Electrical Engineering

Submitted

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

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

JULY 2025

Certificate

This is to certify that the thesis titled **DESIGN OF SECURE HARDWARE ARCHITECTURES FOR NEURAL NETWORKS** being submitted by **Ms. NIVEDITA SHRIVASTAVA** for the award of **Doctor of Philosophy** in Electrical Engineering is a record of bona fide work carried out by her under my guidance and supervision at the Department of Electrical Engineering, Indian Institute of Technology Delhi. The work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma.

Smruti Ranjan Sarangi
Professor
Department of Electrical Engineering
Indian Institute of Technology Delhi
New Delhi- 110016

For my mother,
Late Mrs. Babita Shrivastava

Acknowledgements

I wish to begin by expressing my deepest gratitude to **Prof. Smruti R. Sarangi** for his guidance, encouragement, and unwavering support throughout my PhD journey. I am beyond words to express my gratitude to him for granting me the invaluable opportunity to pursue a PhD and for guiding me through this transformative journey. I am immensely grateful for his patience and wisdom during moments when I felt utterly lost, and for his ability to steer me back on track with clarity and purpose. His relentless pursuit of excellence and perfection, has been a constant source of motivation for me to push my boundaries, strive for improvement, and excel in both academic and personal endeavors. This thesis would have been inconceivable without his valuable guidance and support.

I would like to express my heartfelt gratitude to my Student Research Committee members: **Prof. Harshan Jagdeesh**, **Prof. Mukul Sarkar**, and **Prof. Huzur Saran**, for their invaluable insights, constructive feedback, and thoughtful guidance throughout my research journey. I am also deeply thankful to **Prof. Sparsh Mittal** for his collaboration on my first paper. His guidance, suggestions, and encouragement provided crucial direction and helped me improve the quality of my work significantly.

I would like to express my sincere gratitude to the administrative staff in the department for their continuous support and assistance. I am especially thankful to Ms. Aditi, who went above and beyond her duties to help me navigate and complete the formalities for various processes.

I would like to extend my thanks to all of my friends at IIT Delhi for the many wonderful discussions we shared, both technical and non-technical. I began my research alongside a senior student Dr. Diksha Moolcandani, and I am deeply thankful to her for being a mentor in my early days, helping me grasp the entire research flow—how to approach problems, conduct in-depth analysis, and structure my work for a paper submission. My gratitude extends to the senior students—Dr. Priyanka Singla, Dr. Shubhankar Singh and Dr. Hameedah Sultan—for their invaluable support. Their guidance and

insights made my initial days at the institute far more manageable. I am also thankful to Akanksha, Rohith, and Shruti for always being available for technical discussions and for fostering a friendly and stress-free environment in the lab. The sense of community made the journey not only productive but also enjoyable. I am also grateful to Ani Sunny for being an incredible teammate. Our long discussions and late-night brainstorming sessions during our recent collaboration were truly memorable and impactful.

I am deeply grateful to my father for his unwavering support, motivation, and unconditional love. His presence has been a constant source of strength throughout this journey. I also extend my heartfelt thanks to my husband and mother-in-law for their incredible support and for being my pillars of strength during this endeavor. Their understanding, patience, and encouragement helped me a lot. I am especially thankful to my mother for her unwavering belief in me and her constant motivation to complete my PhD. This accomplishment would not have been possible without her. I deeply wish she was here to witness this milestone.

I would also like to acknowledge the use of OpenAI's ChatGPT for assisting with language refinement in this thesis.

Nivedita Shrivastava

Abstract

Securing neural processing units (NPUs) is critical, as neural network (NN) models embody valuable intellectual property (IP), requiring substantial investments of time, effort, and resources. For example, data collection for a mid-sized AI project via platforms like Amazon Mechanical Turk can exceed $100K, while training methodologies often involve multi-million dollar investments. Large-scale training on computational clusters can take weeks, further emphasizing their value. This makes ML models prime targets for attacks, where theft can compromise IP, privacy, and competitive advantage.

Securing NPUs necessitates enforcing key security properties: confidentiality (preventing unauthorized access), integrity (detecting and preventing tampering), authentication (ensuring access for genuine users), freshness (preventing replay attacks), and resilience to side-channel attacks. These properties defend against threats such as eavesdropping, tampering, cold boot, replay attacks, and memory and timing-based side-channel leaks.

To ensure the confidentiality of a model, it is essential to encrypt the model weights and biases. Given the large size of modern NN models and increasing demand for fast computation, achieving high-throughput and low-latency encryption becomes critical. AES is among the most widely used and highly secure encryption ciphers. However, its software implementation based on T-tables is susceptible to cache side-channel attacks (CSCA). In our first work, we systematically analyzed more than 100 research articles but we only consider a representative subset for detailed discussion. This selection focuses on techniques with significant impact or novel design, while omitting redundant or obsolete approaches. We show that while address-based countermeasures (CMs) introduce obfuscation, they can still be broken in $n^{O(\log(\log(n)))}$ time for popular ciphers like AES.

However, this time bound translates to billions of encryptions, posing a substantial challenge for practical feasibility. To reduce the required number of attack iterations, first we propose a theoretical framework to systematically classify the state-of-the-art countermeasures. Then, we propose a two-phase attack framework to efficiently retrieve the

secret key even in the presence of these advanced countermeasures in just 51,840 tries. This highlights the need for a secure and fast encryption engine that can withstand current and futuristic attacks.

Sadly, the optimization efforts for designing fast encryption engines have focused solely on architecture, as security remained unquantified. This prevented modifications to the cipher's core functionality, such as reducing rounds — a tempting shortcut that risks compromising security guarantees. However, recent works have demonstrated that security can now be quantified. Leveraging their insights, we exploit the fact that if the plaintext distribution has sufficient randomness, fewer encryption rounds are required to generate ciphertexts that maintain the same level of security as AES. Building on this foundation, we introduce a novel encryption engine that dynamically generates ephemeral keys for each data block and provides the same security as AES with nearly 5× faster speed.

Next, we observe that large model sizes hinder the fast implementation of basic security measures. Trusted execution environments (TEEs) like Intel SGX and TrustZone offer a protected on-chip execution environment. However, their high overheads make them inefficient for memory-intensive NPUs. To address the broader security challenges, we first conducted an extensive survey of neural processing unit (NPU) architectures to understand their design. The main issue lies in securing the off-chip data. To solve this, we extensively analyzed and characterized the off-chip memory traffic pattern of different NPUs and designed a fast hardware circuit to predict these patterns. Using these predictions, we implemented mechanisms to ensure the integrity and freshness of the off-chip data at the layer level. This helped us to design an ultra-fast and secure NPU that is 20% faster.

In addition to securing model weights, we identified that the model's parameters itself are attractive targets for adversaries as also they pave the path for mounting other attacks. These parameters are susceptible to leaks through memory- and timing-based side channels. To address this, we classified existing countermeasure into four categories and observed that these countermeasures primarily introduce additive noise to obfuscate the leaked patterns. We highlight the shortcomings of these countermeasures by demonstrating three attacks on them. So, to further enhance security, we introduced an additional layer of defense by incorporating multiplicative noise in the form of compression. Leveraging space-filling curves, our approach securely tiles, compresses, and bins data, effectively mitigating side-channel attacks and expanding the search space to 10^{70}, ensuring robust protection.

In this thesis, we propose a comprehensive and structured approach to tackle the diverse security threats faced by NPUs. By integrating a combination of novel solutions, we aim to develop NPUs that are not only secure but also fast, ensuring robust protection against modern attack vectors.

संक्षेप

तंत्रिका प्रसंस्करण इकाइयों (या एनपीयू) को सुरक्षित करना महत्वपूर्ण है, क्योंकि तंत्रिका नेटवर्क (या एनएन) मॉडल में मूल्यवान बौद्धिक संपदा (या आईपी) शामिल है, जिसमें समय, प्रयास और संसाधनों के पर्याप्त निवेश की आवश्यकता होती है। उदाहरण के लिए, अमेज़न मैकेनिकल टर्क जैसे प्लेटफार्मों के माध्यम से एक मध्यम आकार की कृत्रिम बुद्धिमत्ता परियोजना के लिए डेटा संग्रह १ लाख डॉलर से अधिक हो सकता है, जबकि प्रशिक्षण पद्धतियों में अक्सर बहु-मिलियन डॉलर का निवेश शामिल होता है। कम्प्यूटेशनल समूहों पर बड़े पैमाने पर प्रशिक्षण में हफ्तों लग सकते हैं, जो उनके मूल्य पर और जोर देता है। यह एमएल मॉडल को हमलों के लिए प्रमुख लक्ष्य बनाता है, जहां चोरी आईपी, गोपनीयता और प्रतिस्पर्धी लाभ से समझौता कर सकती है।

एनपीयू को सुरक्षित करने के लिए प्रमुख सुरक्षा गुणों को लागू करने की आवश्यकता होती हैः गोपनीयता (अनधिकृत पहुंच को रोकना), अखंडता (छेड़छाड़ का पता लगाना और रोकना), प्रमाणीकरण (वास्तविक उपयोगकर्ताओं के लिए पहुंच सुनिश्चित करना), ताजगी (रिप्ले हमलों को रोकना) और साइड-चैनल हमलों के लिए संघर्ष क्षमता। ये गुण जासूसी, छेड़छाड़, कोल्ड बूट, रिप्ले हमलों, और स्मृति और समय-आधारित साइड-चैनल लीक जैसे खतरों से बचाव करते हैं।

मॉडल की गोपनीयता सुनिश्चित करने के लिए, मॉडल के वेट्स और पूर्वाग्रहों को कूटबद्ध करना आवश्यक है। आधुनिक एनएन मॉडल के बड़े आकार और तेजी से गणना की बढ़ती मांग को देखते हुए, उच्च-थ्रूपुट और कम-विलंबता एन्क्रिप्शन प्राप्त करना महत्वपूर्ण हो जाता है। जबकि एईएस, सबसे लोकप्रिय और सुरक्षित साइफर में से एक, व्यापक रूप से अपनाया जाता है, इसका टी-टेबल आधारित सॉफ्टवेयर कार्यान्वयन कैश साइड-चैनल हमलों (या सीएससीए) के लिए असुरक्षित है। हमारे पहले काम में, हमने १०० से अधिक शोध लेखों का व्यवस्थित रूप से विश्लेषण किया और दिखाया कि जबकि पता-आधारित प्रतिरोधात्मक उपाय (या सीएम) अस्पष्टता का परिचय देते हैं, लेकिन उन्हें अभी भी एईएस जैसे लोकप्रिय साइफर के लिए ($n^{O(\log(\log(n)))}$) समय में तोड़ा जा सकता है।

हालांकि, यह समय सीमा अरबों एन्क्रिप्शन में बदल जाती है, जो व्यावहारिक व्यवहार्यता के लिए एक महत्वपूर्ण चुनौती पेश करती है। हमले के पुनरावृत्तियों की आवश्यक संख्या को कम करने के लिए, पहले हम अत्याधुनिक सीएम को व्यवस्थित रूप से वर्गीकृत करने के लिए एक सैद्धांतिक ढांचे का प्रस्ताव करते हैं। फिर, हम केवल ५१,८४० प्रयासों में इन उन्नत सीएम्स की उपस्थिति में भी गुप्त कुंजी को कुशलता से पुनः प्राप्त करने के लिए दो-चरणीय हमले के ढांचे का प्रस्ताव करते हैं। यह एक सुरक्षित और तेज कूटलेखन इंजन की आवश्यकता पर प्रकाश डालता है जो वर्तमान और भविष्य के हमलों का सामना कर सकता है।

अफसोस की बात है कि तेज एन्क्रिप्शन इंजनों को डिजाइन करने के लिए अनुकूलन प्रयासों ने पूरी तरह से वास्तुकला पर ध्यान केंद्रित किया है, क्योंकि सुरक्षा अगणनीय रही है। इसने साइफर की मुख्य कार्यक्षमता में संशोधनों को रोक दिया, जैसे कि राउंड को कम करना-एक आकर्षक शॉर्टकट जो सुरक्षा गारंटी से समझौता करने का जोखिम उठाता है। हालाँकि, हाल के कार्यों ने प्रदर्शित किया है कि सुरक्षा की मात्रा अब निर्धारित की जा सकती है। उनकी अंतर्दृष्टि का लाभ उठाते हुए, हम इस तथ्य का फायदा उठाते हैं कि यदि प्लेनटेक्स्ट वितरण में पर्याप्त यादृच्छिकता है, तो सिफरटेक्स्ट उत्पन्न करने के लिए कम एन्क्रिप्शन राउंड की आवश्यकता होती है जो एईएस के समान स्तर की सुरक्षा बनाए रखते हैं। इसके आधार पर, हम एक नए एन्क्रिप्शन इंजन का प्रस्ताव करते हैं जो गतिशील रूप से प्रत्येक डेटा ब्लॉक के लिए अल्पकालिक कुंजी उत्पन्न करता है और ५ गुना तेज गति के साथ एईएस के समान सुरक्षा प्रदान करता है।

इसके बाद, हम देखते हैं कि बड़े मॉडल आकार बुनियादी सुरक्षा उपायों के तेजी से कार्यान्वयन में बाधा डालते हैं। जबकि इंटेल एसजीएक्स और ट्रस्टजोन जैसे विश्वसनीय निष्पादन वातावरण (या टीईई) एक सुरक्षित ऑन-चिप निष्पादन वातावरण प्रदान कर सकते हैं, उनके पास बहुत अधिक ओवरहेड्स हैं और स्मृति-गहन एनपीयू के लिए उपयुक्त नहीं हैं। व्यापक सुरक्षा चुनौतियों का समाधान करने के लिए, हमने पहले एनपीयू आर्किटेक्चर का एक व्यापक सर्वेक्षण किया ताकि उनके डिजाइन को समझा जा सके। मुख्य मुद्दा ऑफ-चिप डेटा को सुरक्षित करने में निहित है। इसे हल करने के लिए, हमने विभिन्न एनपीयू के ऑफ-चिप मेमोरी ट्रैफिक पैटर्न का व्यापक रूप से विश्लेषण और विशेषता की और इन पैटर्न का पूर्वानुमान करने के लिए एक तेज हार्डवेयर सर्किट तैयार किया। इन पूर्वानुमानों का उपयोग करते हुए, हमने लेयर स्तर पर ऑफ-चिप डेटा की अखंडता और ताजगी सुनिश्चित करने

के लिए तंत्र को लागू किया। इसने हमें एक अल्ट्रा-फास्ट और सुरक्षित एनपीयू डिजाइन करने में मदद की जो २०% तेज है।

मॉडल भार को सुरक्षित करने के अलावा, हमने पहचान की कि मॉडल के पैरामीटर अपने आप में विरोधियों के लिए आकर्षक लक्ष्य हैं और साथ ही वे अन्य हमलों को बढ़ाने का मार्ग भी प्रशस्त करते हैं। ये पैरामीटर स्मृति-और समय-आधारित साइड चैनलों के माध्यम से रिसाव के लिए अतिसंवेदनशील हैं। इसे संबोधित करने के लिए, हमने मौजूदा सीएम्स को चार श्रेणियों में वर्गीकृत किया और देखा कि ये सीएम मुख्य रूप से लीक हुए पैटर्न को अस्पष्ट करने के लिए एडिटिव नॉइज़ पेश करते हैं। हम इन सीएम्स पर तीन हमले करके उनकी कमियों को उजागर करते हैं। इसलिए, सुरक्षा को और बढ़ाने के लिए, हमने संपीड़न के रूप में गुणक नॉइज़ को शामिल करके रक्षा की एक अतिरिक्त परत पेश की। स्पेस-फिलिंग कर्व्स का उपयोग करके, हमारा दृष्टिकोण डेटा को सुरक्षित रूप से टाइल करता है, संपीड़ित करता है और बिन करता है, जिससे साइड-चैनल हमलों को प्रभावी रूप से कम किया जाता है और खोज स्थान को $१०^{७०}$ तक विस्तारित किया जाता है, जिससे मजबूत सुरक्षा सुनिश्चित होती है।

इस थीसिस में, हम एनपीयू द्वारा सामना किए जाने वाले विविध सुरक्षा खतरों से निपटने के लिए एक व्यवस्थित और समग्र दृष्टिकोण का प्रस्ताव करते हैं। नए समाधानों के संयोजन को एकीकृत करके, हमारा लक्ष्य एनपीयू विकसित करना है जो न केवल सुरक्षित हैं बल्कि तेज भी हैं, जो नवीनतम आक्रमण विधियों के खिलाफ ठोस सुरक्षा सुनिश्चित करना।

Contents

List of Figures

List of Tables