APPLICATION-AWARE DYNAMIC THERMAL MANAGEMENT IN 3D DRAM

SHAILJA PANDEY

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
JULY 2024

©Indian Institute of Technology Delhi (IITD), New Delhi, 2024

APPLICATION-AWARE DYNAMIC THERMAL MANAGEMENT IN 3D DRAM

by

SHAILJA PANDEY

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

JULY 2024

To

Amma, Kakka, Mummy, and Papa,

who blessed me with Education.

Certificate

This is to certify that the thesis titled "**Application-aware Dynamic Thermal Management in 3D DRAM**" being submitted by **Shailja Pandey** for the award of **Doctor of Philosophy** in Computer Science and Engineering is a record of bonafide work carried out by her under my guidance and supervision in the Department of Computer Science and 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 unless otherwise stated explicitly.

Certain works included in this thesis involved collaboration with other students, which have been explicitly specified/acknowledged in the corresponding chapters and the part done by those collaborators appeared in their respective reports/theses.

Preeti Ranjan Panda
Professor
Department of Computer Science and Engg.
Indian Institute of Technology Delhi
New Delhi- 110016

Acknowledgments

"You are what you believe in. You become that which you believe you can become."

—Bhagavad Gita

I would begin by expressing my deepest gratitude to my Ph.D. supervisor, Prof. Preeti Ranjan Panda, for his patient guidance, enthusiastic encouragement, and invaluable pieces of advice throughout my Ph.D. journey. I joined Prof. Panda's group in February 2020 after three and a half years into the Ph.D. program. Shortly after that, there was a nationwide COVID lockdown while I was restarting my research in a different area with no prior research experience in the area. That was the toughest time in this whole journey as I was navigating through three major tasks: (1) motivate myself to continue my Ph.D., (2) learn a new research area from scratch and define an interesting problem statement, and (3) deal with the anxiety and uncertainty that COVID brought into the world. Amid all the chaos, Prof. Panda showed me a path that I eventually started believing would converge in the future. His expertise and insights were instrumental in shaping my research ideas and methodologies. He is the best mentor anyone can ask for. During my interactions with him, I have learned so many subtle ways of mentoring a student that I will try to replicate when the opportunity comes. As we say among ourselves in the lab – we go into his office with so much stress and come out with so much motivation. Most importantly, I will remain indebted to him for teaching me so much beyond technical lessons and research, indirectly influencing my personal growth as well.

My heartfelt thanks to the members of my Ph.D. committee, Prof. Kolin Paul, Prof. Anuj Grover (IIIT Delhi), and Prof. Abhilash Jindal, for their constructive feedback and suggestions in the semester's progress presentations. Their diverse perspectives improved my research proposals and contributed significantly to their quality.

When Prof. Panda and I defined our problem statement around deep neural networks (DNNs), my friend, Vishal Sharma, helped me understand the fundamentals of DNNs and pointed me to key resources for making faster progress. I am extremely thankful and indebted to him for those countless discussions on how to design efficient Systems for ML workloads. Several insights used in this thesis came from those discussions and helped me build the foundations of my studies.

I would like to express my sincere thanks to everyone I have had the chance to collaborate with in different ways and learn so much along the process. I worked with Lokesh Siddhu, Prof. Rajesh Kedia (IIT Hyderabad), Prof. Anuj Pathania (University of Amsterdam), Martin Rapp, and Prof. Jörg Henkel (Karlsruhe Institute of Technology) for the development of the

simulation infrastructure used in this research. Along with Lokesh, I briefly collaborated with Aritra Bagchi, Issar Ahmad, and Rajesh in one of the research studies. Working with BTech students – Mritunjay Kumar Gupta (helped in modeling HBM architecture), Sayam Sethi (worked on 3D DRAM power budgeting problem), and Aaveg Jain (helped with 12-layer 3D DRAM study) was very interesting as it helped me develop mentoring skills at some level. Overall, these collaborations and interactions contributed greatly to my overall research experience, making it more enriching and rewarding.

My heartfelt thanks to all my friends in our Memory and Embedded Architecture (MARG) Group – Rajesh Kedia, Lokesh Siddhu, Aritra Bagchi, Dharamjeet Choudhary, Sakshi Tiwari, Garima Modi, Divya Praneetha, Ayushi Agarwal, Naman Jain, Dinesh Joshi, Radhika Dharwadkar, and Ravikant Bhardwaj. Lokesh helped and guided me with the initial infrastructure setup required for the research and contributed to the brainstorming sessions to define the problem statement. While Vishal immensely supported me throughout the process when I was looking to change my Ph.D. supervisor, Rajesh and Lokesh gave key suggestions that helped me ground my thoughts for approaching Prof. Panda as a potential advisor. Thank you to all of you for the lovely memories in the lab, *gupshup* over tea and snacks, long chit-chats in the corridors, and fun-filled group lunches. The friendly and positive lab environment greatly helped me settle down quickly and accelerate my research work. Thanks to Sanjana for the joyful time in the initial years, and to Nikki for giving a nice company in the lab. I will deeply cherish all the memories with each one of you for the rest of my life.

I am sincerely thankful to the office staff of the CSE department, particularly Vandana ma'am, Rekha ma'am, Hemant sir, Rajesh sir, and Suresh sir, for their timely efforts toward several administrative tasks involved in my Ph.D. journey. I remember the joyful conversations with Vandana Ma'am that provided me with lots of encouragement when I shifted to a new lab during COVID and was alone in the lab for several months. Thank you to the cleaning staff and guards in the SIT building who spread positivity in their own way by smiling on seeing the students and participating in casual conversations.

I am grateful to have spent several years on the IIT campus and made so many loving memories that will forever remain with me and make me smile. Some of my fondest memories at IIT Delhi during the seven years of stay on campus are in the Kailash Hostel. Thanks to my roommate Khushboo for being there during good and bad times. I would also like to thank two of my old friends, Neetika Singh and Arpita Shah, who supported me during various high and low points in my life.

I would like to express my gratitude to my institute, IIT Delhi, for giving me the opportunity to pursue a Ph.D. and experience the whole process of discovering myself in a way that would

otherwise not have been possible. It gives me immense pride to be a part of an institute that has such deep roots. I would also like to thank the Ministry of Education of India for providing me with a scholarship during my Ph.D. and the Semiconductor Research Corporation for providing me with funding in the later years.

While one can never thank one's parents enough for nurturing them and shaping their lives in the best possible way, I am eternally grateful to my parents, Maya and Banspati Pandey, for providing me with their unwavering support during this journey and breaking several societal norms for me that they grew up with and gotten old with. My brothers, Ankur and Anupam, have been my biggest cheerleaders throughout my education journey, at times prioritizing my seamless education over theirs and trying to eliminate all the hurdles. During my Ph.D., my brothers got married to my lovely sisters-in-law – Kamana Bhabhi and Deepti Bhabhi, who brought so much happiness to the family. Further, both my brothers were blessed with their daughters – Aarya and Aadvika, making me a proud *Bua*. My heart is filled with so much love for my nieces. I would like to thank my *naani* for being a great example of a courageous woman for me. I am also grateful to my *foofaji*, Late Krishna Kishore Shukla, for accompanying me during interviews at various IITs and otherwise while I had still not learned to travel alone.

Finally, I would like to acknowledge that it is a huge privilege to be born in India and, more so, in a village of India that provides an unparalleled life experience. I am thankful to our goddess *Maa Sharda* for blessing me to be born at the epicenter of freedom and spirituality. Maa Sharda is my primary source of strength and inspiration in easy and tough times, and I owe everything to her.

Shailja Pandey

Abstract

3D DRAM is a promising technology to meet the bandwidth demands of data-hungry ML/AI applications and is being actively used in a wide range of hardware platforms such as GPUs, FPGAs, DNN accelerators, and general-purpose cores. Despite the potential of 3D DRAM to bridge the memory bandwidth gaps, the vertically stacked integration of DRAM dies brings a new set of system design challenges. One such challenge is the poor heat dissipation capability, causing excessive memory heating and eventually throttling the memory, slowing down the applications.

Heating in 3D DRAM, due to high power density and close proximity of memory cells in stacked architectures, impacts the data retention capabilities, increasing the refresh rates significantly. 3D stacking enables hundreds of banks/ranks and several independent channels per memory device such that each channel is connected to a separate memory controller; however, the off-chip memory bandwidth is often limited due to limited power budgets that might arise in practical scenarios. Due to this, it is often not possible to activate and run all memory ranks simultaneously, similar to *dark silicon* in multi-core systems. Systems running under a fixed power budget and a thermal constraint employ dynamic power budgeting (*DPB*) and dynamic thermal management (*DTM*) policies, resulting in performance and energy overheads due to thermal-induced CPU and memory throttling.

Existing solutions towards dynamic thermal management are application-agnostic and often lead to sub-optimal performances. Modern applications execute in distinct *phases*, exhibiting varying memory access behavior and showing different sensitivities to architectural parameters such as last level cache (LLC) size, prefetch settings, memory-level parallelism, and memory bandwidth. The indiscriminate approach of prior works for DTM in 3D DRAM may lead to avoidable performance degradation, as different application phases have varying contributions to memory temperature. To this end, this dissertation proposes application-aware DTM techniques targeted towards minimizing the associated performance penalties. We present novel micro-architecture-based approaches and system-level policies for managing heat in 3D DRAM and facilitate faster cooling, leveraging knowledge of the workload. We

focus on deep neural networks (DNNs) in this thesis.

Deep neural network (DNN) implementations are typically characterized by huge data sets and concurrent computation, resulting in a demand for high memory bandwidth due to intensive data movement between compute units and off-chip memory. Performing DNN inference on a general-purpose processor is common, and multi-core CPUs are being used for DNN implementation in servers and commercial SoCs. 3D DRAM is becoming a key enabler for high memory-intensive applications such as DNNs running on general-purpose cores. This dissertation focuses on efficient thermal management on a system consisting of general-purpose CPU cores and a thermally-constrained 3D DRAM.

Firstly, we present a memory temperature-aware prefetcher that jointly utilizes DNN application's memory behavior and 3D DRAM temperature to minimize DTM overheads. The customized prefetching mechanism dynamically computes the best *prefetch settings* for different DNN workload phases, utilizing the unused 3D DRAM bandwidth without adversely affecting the workload's thermal footprint.

Secondly, we present an efficient task mapping based DTM policy for multi-core processors to minimize memory throttling through proactive management of the potential thermal hotspots in 3D DRAM using application-aware optimizations in processing cores. An application-aware DVFS is used to reduce the thermal impact of the application's *phase* when task mapping to an ideal core is not feasible.

Thirdly, we present a thermal and memory access pattern-aware data mapping based DTM policy to map a DNN layer's data to 3D DRAM pseudo-channels. An intelligent mapping of the DNN application's data across memory channels, pseudo-channels, and banks, leveraging the knowledge about each layer's memory access pattern and date reuse, significantly reduces the effective memory latency and the thermal-induced application slowdown.

Finally, we present a reward-based and adjacency-aware dynamic power budgeting which periodically suggests the ideal set of channels that should be enabled under a given power budget and thermal constraint, performing a coordinated thermal and power management for 3D DRAM. The thermal characteristics of 3D DRAM make power budgeting a non-trivial task that requires run-time intervention to minimize the associated performance penalties.

To evaluate the proposed policies, an integrated performance-thermal simulation framework is used. Our results demonstrate significant DNN inference time and memory energy reductions over the state-of-the-art, with minimal run-time cost. This dissertation establishes the significance of leveraging application knowledge towards efficient thermal management in 3D DRAM for a class of well-structured workloads such as deep neural networks.

, ö - |
ô ý -
| | | |
, ö |
ö -

, - |
â ł | ,
- ö , ñ
â | ½ , | -
ö

, ł õ -
- | |
¼ | - ç |
| - ¼

ý , - | |
õ - ñ ñ |
| | õ | ç
ö | | ö ½
ñ

Contents

List of Figures

List of Tables