LEAKAGE AWARE DYNAMIC THERMAL MANAGEMENT FOR 3D MEMORY ARCHITECTURES

LOKESH SIDDHU

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
December 2022

LEAKAGE AWARE DYNAMIC THERMAL MANAGEMENT FOR 3D MEMORY ARCHITECTURES

by

Lokesh Siddhu

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

December 2022

Certificate

This is to certify that the thesis titled "**Leakage Aware Dynamic Thermal Management for 3D Memory Architectures**" being submitted by **Lokesh Siddhu** for the award of **Doctor of Philosophy** in Computer Science and Engineering is a record of bonafide work carried out by him 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 or use of measurement data obtained by other students, which have been explicitly specified/acknowledged in the corresponding chapters and the part done by those collaborators appeared in their respective thesis.

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

Acknowledgments

"Gratitude is not an attitude. Gratitude is something that flows out of you when you are overwhelmed by the recognition of what you have received."

—Sadhguru

I would like to express my gratitude to Prof. Preeti Ranjan Panda for his support, guidance, and insight. His unwavering commitment to quality has deeply touched me. Also, his keen attention, careful listening, and understanding are qualities worth emulating in every aspect of life. His constructive feedback on various presentation aspects has improved my skills by leaps and bounds. Prof. Panda has been a joy to share time and work with.

I have had the privilege of getting insightful and valuable feedback from Prof. Anshul Kumar, Prof. Kolin Paul, and Prof. Mukul Sarkar throughout my research. Their suggestions, understanding, and critique have been very helpful. Prof. M. Balakrishnan's vision has been a constant source of inspiration. His ability to present complex problems clearly and simply provided ease in understanding and imbibing. Also, I had a wonderful time TA-ing. It offered me the experience of conducting and managing labs and marking examinations and assignments. I would like to thank all the related professors.

Rajesh Kedia has been a wonderful friend, collaborator, and colleague at IIT Delhi. Almost daily, we discussed research problems and solutions, resulting in a regular knowledge update. His support, both in personal and professional capacities, has been of immense value to me. Over the years, I have had the good fortune of collaborating with many professors and students - Anuj Pathania, Rajesh Kedia, Shailja Pandey, Aritra Bagchi, Martin Rapp, Isaar Ahmad, Hameedah Sultan. Their ideas and suggestions have helped me thoroughly. Discussions with them have resulted in improved research quality and presentation. I would also like to thank the members of the Memory and Architecture Research Group (MARG) - Rajesh Kedia, Neetu Jindal, Shailja Pandey, Sakshi Tiwari, Hadi Brais, Divya Praneetha, Isaar Ahmad, Aritra Bagchi, Ayushi Agarwal, Garima Modi, Zilmarij Iqbal, for their everlasting support and company. The technical feedback in our weekly meeting ensured steady progress and learning. In addition, our informal conversations over lunch/tea outings provided a daily dose of humor, joy, and research ideas.

In the Ph.D. journey, my labmates have played a crucial role in creating a conducive atmosphere for research. I would like to thank Kunal Dahiya, Ankit Anand, Dinesh Khandelwal, Vishal Sharma, Dilpreet Kaur, Harsh Das, Samuel Wedaj, Solomon Abera, Richa Gupta, Himanshu Gandhi and Iqra Altaf. Our discussion ranged from technical topics to

global leadership and polity, always resulting in a more thorough understanding of research or any other situation. Many parts of my research's foundation were established with the help of Sandeep Chandran, Namita Sharma, Rajshekar Kalayappan, and Prathmesh Kallurkar. Time spent with them was invaluable.

During my Ph.D., I had the opportunity to organize volunteer activities for VLSID 2019. It was an immensely enriching experience. In many ways, it depicts how the crux of the academic culture lies in several volunteers. Various selfless volunteers have built the community, which allows us to function. I wish to thank them from the bottom of my heart.

The office staff at Amar Nath and Shashi Khosla School of Information Technology and the Department of Computer Science have been most helpful. I would like to thank Ms. Vandana, Ms. Rekha, Mr. Hemant, Mr. Suresh, Mr. Rajesh, Mr. Som Dutt, and many others - they have been most helpful and streamlined many administrative activities. My stay at the hostel placed me in a youthful and exuberant environment. Providing a stay and food on the campus is a luxury. I thank the hostel for their diligent effort. Much of my research assistance was funded by the Ministry of Electronics and Information Technology (MeitY) with the Visvesvaraya Ph.D. Scheme. Their support and efficiency are commendable.

I am fortunate to have friends connected dearly. Every moment spent with them was a joy. Immense gratitude to Ankit Gupta, Swati Gupta, Gaurav Singh, Disha Singh, Ruchik Shastri, Manvi Sharma, Amit Mishra, Namrata Mishra, Srijith Haridas, Nagraj Swami, Karthika Ravindran, Sumeet Mali, Nimash Choudhary, Chinmay Jain, Mehul Agrawal. Their feedback and understanding have positively shaped various aspects of my thought and emotion.

My family has been a constant source of support. Their commitment to education, learning, and knowledge has been exemplary. My parents have especially done their best to take care of me and provide a comfortable environment for my pursuit of knowledge. As I reflect, life is connected in several myriad ways, and countless people have contributed to enabling me. All the masters over generations have passed their knowledge and wisdom. My sincere gratitude and love to all life. It has been an immensely fulfilling and joyful journey.

Lokesh Siddhu

Abstract

Memory architects have relied on process scaling (making transistors, wires smaller) to gain transistor density and performance in the last few decades. However, the growing fabrication/production costs and variability in the manufacturing process have motivated them to explore 3D architectures, which allow the stacking of 2D memories to gain an order of magnitude of bandwidth. However, the increased stacking and power density in 3D memories make heat dissipation a challenge. At elevated temperatures, the transistors (and chips) may operate unreliably. Higher temperatures than the safe operating temperature might also cause permanent damage to the chip. Cooling costs increase significantly when on-chip temperatures are higher. Leakage current grows exponentially with temperature causing a further rise in temperature and may lead to thermal runaway. Hence, there is a need to study low overhead thermal-aware system-level solutions for 3D memory architectures.

While prior works have looked at reducing dynamic/switching power through reduced memory accesses, in these memories, both leakage and dynamic power consumption are comparable. Furthermore, as the temperature rises, the leakage power increases, creating a thermal-leakage loop. We present the first study of the thermal-leakage loop resulting in a system-level DTM policy, *FastCool*, based on data migration to the 2D memory. We propose a detailed analytical model to quantify the memory delay, including queuing delays for a hybrid (3D + 2D) memory system. We suggest an improved DTM strategy, *Energy-Efficient FastCool (EEFC)*, which incorporates additional designer insights to dynamically select the channels to be closed, improving the performance and energy further. We also identify that EEFC is resilient to process variability.

To lower DTM overheads, we present a proactive DTM policy, *PredictNcool*, which uses a lightweight and accurate steady-state temperature predictor for 3D memories. *PredictNcool* is the first work exploiting various designer insights about the layout/floorplan (including different symmetries) to reduce model parameters of a machine learning-based predictor, significantly reducing the required training data and prediction time. Further, we discuss low power state based DTM approaches, decreasing energy overheads, and energy-delay-product by avoiding

data migration and 2D memory usage. We suggest a DTM policy to reduce the performance overheads by integrating low power state and partial channel turn-off based approaches.

As cores and 3D memory get heated in modern computing systems, we suggest a joint thermal management approach, *CoreMemDTM*, avoiding inefficiencies caused by independently running DTMs and minimizing performance overheads. We suggest employing DTM depending on the thermal margin since safe temperature thresholds might differ for the two subsystems. We present a stall-balanced core DVFS policy for core thermal management that enables distributed cooling, decreasing overheads. Further, we introduce *CoMeT*, the integrated Core and Memory Thermal simulation toolchain, which supports various core and 3D memory configurations. The source code of *CoMeT* has been made open for public use under the *MIT* license.

तीन-डी मेमोरी आर्किटेक्चर के लिए लीकेज अवेयर डायनामिक थर्मल मैनेजमेंट

सार:

मेमोरी आर्किटेक्ट्स ने पिछले कुछ दशकों में ट्रांजिस्टर घनत्व और प्रदर्शन हासिल करने के लिए प्रोसेस स्केलिंग (ट्रांजिस्टर, तारों को छोटा बनाना) पर भरोसा किया है। हालांकि, निर्माण प्रक्रिया में बढ़ती निर्माण/उत्पादन लागत और परिवर्तनशीलता ने उन्हें तीन-डी आर्किटेक्चर का पता लगाने के लिए प्रेरित किया है, जो बैंडविड्थ की परिमाण के क्रम को प्राप्त करने के लिए दो-डी मेमोरी के स्टैकिंग की अनुमति देता है। हालांकि, तीन-डी मेमोरी में बढ़ी हुई स्टैकिंग और पावर डेंसिटी गर्मी अपव्यय को एक चुनौती बनाती है। ऊंचे तापमान पर, ट्रांजिस्टर (और चिप्स) अविश्वसनीय रूप से काम कर सकते हैं। सुरक्षित ऑपरेटिंग तापमान से अधिक तापमान भी चिप को स्थायी नुकसान पहुंचा सकता है। जब ऑन-चिप तापमान अधिक होता है तो कूलिंग की लागत काफी बढ़ जाती है। लीकेज करंट तापमान के साथ तेजी से बढ़ता है जिससे तापमान में और वृद्धि होती है और इससे थर्मल पलायन हो सकता है। इसलिए, तीन-डी मेमोरी आर्किटेक्चर के लिए कम ओवरहेड थर्मल-जागरूक सिस्टम-स्तरीय समाधानों का अध्ययन करने की आवश्यकता है।

पिछले कार्यों ने कम मेमोरी एक्सेस के माध्यम से गतिशील/स्विचिंग पावर को कम करने पर ध्यान दिया है, इन मेमोरी में, रिसाव और गतिशील बिजली की खपत दोनों तुलनीय हैं। इसके अलावा, जैसे-जैसे तापमान बढ़ता है, रिसाव की शक्ति बढ़ती है, जिससे थर्मल-रिसाव लूप बनता है। हम थर्मल-लीकेज लूप का पहला अध्ययन प्रस्तुत करते हैं जिसके परिणामस्वरूप सिस्टम-स्तरीय डीटीएम नीति, फास्टकूल, दो-डी मेमोरी में डेटा माइग्रेशन पर आधारित है। हम एक हाइब्रिड (तीन-डी + दो-डी) मेमोरी सिस्टम के लिए क्यूइंग देरी सहित मेमोरी देरी को निर्धारित करने के लिए एक विस्तृत विश्लेषणात्मक मॉडल का प्रस्ताव करते हैं। हम एक बेहतर डीटीएम रणनीति, ऊर्जा-कुशल फास्टकूल का सुझाव देते हैं, जो प्रदर्शन और ऊर्जा को और बेहतर बनाने के लिए चैनलों को गतिशील रूप से चुनने के लिए अतिरिक्त डिज़ाइनर अंतर्दृष्टि को शामिल करती है। हम यह भी पहचानते हैं कि ऊर्जा-कुशल फास्टकूल प्रक्रिया परिवर्तनशीलता की उपस्थिति में भी अच्छा प्रदर्शन करती है।

डीटीएम ओवरहेड्स को कम करने के लिए, हम एक सक्रिय डीटीएम नीति प्रस्तुत करते हैं, प्रिडिक्टएनकूल, जो तीन-डी मेमोरी के लिए हल्के और सटीक स्थिर-अवस्था तापमान पूर्वसूचक का उपयोग करती है। प्रेडिक्टएनकूल मशीन लर्निंग-आधारित प्रेडिक्टर के मॉडल मापदंडों को कम करने के लिए लेआउट/फ्लोरप्लान (विभिन्न समरूपता सहित) के बारे में विभिन्न डिज़ाइनर अंतर्दृष्टि का उपयोग करने वाला पहला काम है, आवश्यक प्रशिक्षण डेटा और भविष्यवाणी समय को महत्वपूर्ण रूप से कम करता है। इसके अलावा, हम डेटा माइग्रेशन और दो-डी मेमोरी उपयोग से बचने के द्वारा कम बिजली की स्थिति आधारित डीटीएम दृष्टिकोण, कम ऊर्जा खर्च और ऊर्जा-देरी-उत्पाद पर चर्चा करते हैं। हम कम बिजली की स्थिति और आंशिक चैनल टर्न-ऑफ आधारित दृष्टिकोणों को एकीकृत करके प्रदर्शन ओवरहेड्स को कम करने के लिए एक डीटीएम नीति का सुझाव देते हैं।

चूंकि आधुनिक कंप्यूटिंग सिस्टम में कोर और तीन-डी मेमोरी गर्म हो जाती है, हम एक संयुक्त थर्मल प्रबंधन दृष्टिकोण, कोर मेम डीटीएम का सुझाव देते हैं, ताकि स्वतंत्र रूप से डीटीएम चलाने और प्रदर्शन ओवरहेड्स को कम करने के कारण होने वाली अक्षमताओं से बचा जा सके। हम थर्मल मार्जिन के आधार पर डीटीएम को नियोजित करने का सुझाव देते हैं क्योंकि सुरक्षित तापमान थ्रेसहोल्ड दो सबसिस्टम के लिए भिन्न हो सकते हैं। हम कोर थर्मल प्रबंधन के लिए एक स्टाल-संतुलित कोर डीवीएफएस नीति प्रस्तुत करते हैं जो वितरित कूलिंग, घटते ओवरहेड्स को सक्षम बनाता है। इसके अलावा, हम कॉमेट, एकीकृत कोर और मेमोरी थर्मल सिमुलेशन टूलचेन पेश करते हैं, जो विभिन्न कोर और तीन-डी मेमोरी कॉन्फ़िगरेशन का समर्थन करता है। कॉमेट के स्रोत कोड को एम.आई.टी लाइसेंस के तहत सार्वजनिक उपयोग के लिए खुला कर दिया गया है।

Contents

List of Figures

List of Tables

List of Acronyms

API	Application programming interface
CMPs	Chip multi-processors
DRAM	Dynamic random access memory
DTM	Dynamic thermal management
DVFS	Dynamic voltage and frequency scaling
EDA	Electronic design automation
EDP	Energy-delay product
EEFC	Energy-Efficient FastCool
FC	FastCool
HBM	High Bandwidth Memory
HMC	Hybrid Memory Cube
JEDEC	Joint Electron Device Engineering Council
LR	Linear regression
MPSoC	Multi-Processor System on Chip
NN	Neural network
NVM	Non-Volatile memory
PCB	Printed circuit board
PCM	Phase change memory
PID	Proportional-Integral-Derivative
PNC	PredictNcool
TAM	Thermal-aware migration
WID	Within-die
WIO	Wide I/O