MSCA Individual Fellowships
Final Report

Project Number: 898296
Project Acronym: TECTONIC
Project Title: Towards Employing Compilers for Thermal Management and Optimal Data Placement in Hybrid Cache

Periodic Technical Report

PART B

Period covered by the report: from 01/01/2021 to 31/12/2022
Periodic report – FINAL

1 Explanation of the work carried out by the beneficiaries and Overview of the progress

With the onset of the TECTONIC project, we started working on thermal properties of the registers built in state-of-the-art FinFET technology nodes. We investigated the thermal properties of FinFET based cores, caches, and registers and established relationships between data accesses and hotspots, by executing a set of benchmark applications in our simulation framework. Towards conducting this research, we collaborated with Cyprus University of Technology. The detailed outcomes of this investigation are published in *Computing Frontiers 2022* conference[1].

We also started investigating non-volatile memory (NVM) based cache with our collaborators at IIT-BHU (India), Newcastle University (UK) and at The University of Edinburgh (UK). Among all types of NVMs, we focused on the STT-RAM (Spin Transfer Torque RAM) and ReRAM (Resistive RAM) for our investigation, as these two NVMs offer more promising features than the other NVMs.

We started exploring the write related issues of the STT-RAM and realized the necessity of dynamic swapping of the cache blocks between the cache sets. Moreover, the relation between temperature, write endurance, retention time, device lifetime have also been investigated. The preliminary results are published in ASAP 2021[2] conference, whereas the detailed version of this work is still under review and yet to be published. Our comprehensive analysis includes a detailed design space exploration based on thermal analysis of the STT-RAM's fundamental properties, and their mitigation techniques are also proposed. In addition, we have also investigated the write issues in STT-RAM based multi-retention cache in case of heterogeneous systems, equipped with CPU and GPU cores. Our investigation reveals that, managing write issues are more complex in heterogeneous systems, due to the variation in lifetimes of the CPU and GPU applications. We proposed a solution to overcome this issue and preliminary results have been accepted for a publication in a premium venue, *60-th Design Automation Conference (DAC 2023)*[3]. A followup publication that will include a comprehensive thermal analysis of STT-RAM cache is in the final preparations.

In our exploration of ReRAM based caches, we are currently working with Newcastle University (UK) to develop a power-performance-thermal simulation framework. The architectural analytical framework is developed at NTNU,

[1] Shounak Chakraborty, Vassos Soteriou, Magnus Själander, "STIFF: Thermally Safe Temperature Effect Inversion aware FinFET based Multi-core", CF 2022.

[2] Sukarn Agarwal, Shounak Chakraborty, "ABACa: Access Based Allocation on Set Wise Multi-Retention in STT-RAM Last Level Cache", ASAP 2021.

[3] Sukarn Agarwal, Shounak Chakraborty, Magnus Själander, "Architecting Selective Refresh based Multi-Retention Cache for Heterogeneous System (ARMOUR)", DAC 2023 (accepted)

whereas circuit level verification and modeling of the ReRAM is being done at the Newcastle University. The simulation framework is already developed and is undergoing final verification for correctness. On completion of the verification process, the simulation framework will be released as an open-source tool (named as HotReRAM) for public use. The documentation for the framework is under preparation, and will be submitted to one of the premium journals, e.g., IEEE TVLSI, IEEE TCAD, ACM TACO, ACM TECS, etc.

We also proposed to develop a loop-splitting based technique to overcome write limitations of NVM based caches. During the last six months of the project, we started investigating this topic with our collaborator at the Florida State University (USA). This work considers an STT-RAM based multi-retention cache, where write intensive data will be segregated at the compilation level, and will be placed in the appropriate retention zone by employing linker and cache controller together. The exploration is still under progress and will be finished by end of April 2023, and subsequently the results will be communicated at a premium conference (e.g., CASES, PACT, HPCA, ASPLOS) and/or journal (e.g., ACM TACO, ACM TOCS, IEEE TC, IEEE TPDS, etc.) for publication.

Objectives: The objectives of the TECTONIC project are as follows:

1. To establish a relationship between data accesses and hot-spots at the CPU registers by investigating compute-intensive loops of contemporary large-scale applications (e.g., multimedia based applications, heavy algebraic computation, etc.).
2. To employ loop splitting towards ameliorating the write endurance problem in non-volatile memories (NVMs).
3. To investigate the benefits of dynamic swapping of the contents of cache ways based upon write intensity.
4. The knowledge deduced by the above investigations will enable us to analyze power-performance-thermal trade-offs and develop an optimal thermal-aware data placement strategy for NVM-based hybrid cache.
5. The implementation of TECTONIC also includes the creation of a performance-power-thermal simulation framework for the NVM.

Overall assessment: *The project has achieved most of its objectives and milestones for the period, with relatively minor deviations.*

The project has successfully achieved its first objective. The third objective has also been completed, and the corresponding publication is still under review. We have also completed the work related to fourth and fifth objectives,

whereas the respective papers are still under preparation, and will be finished soon. The second objective of the project is still in progress which will also be finished soon.

The primary goal of the TECTONIC was to investigate thermal properties of the contemporary multi-cores equipped with NVM based last-level caches. TECTONIC also aimed to investigate the thermal properties and solutions of the write limitations of the NVMs. By considering the project objectives, we have successfully completed first and third objectives whereas rest two are still under progress, as we have detailed in the beginning of the section. The simulation framework is yet to be released, and the respective article is also under preparation and will be communicated soon for publication.

Explanation of the work carried out per WP (Work Package)

WP1: Design and develop the thermal simulator

We started with investigation of several fundamental properties of ReRAM and its dependency on temperature. We enhanced an existing thermal simulator, HotSpot, with ReRAM's thermal characteristics to support the thermal properties of ReRAM. This extended version of HotSpot that supports ReRAM's characteristics is named as HotReRAM. To enable detailed verification of the thermal model of ReRAM at the circuit level, we collaborated with Newcastle University (UK). Finally, to develop the complete simulation framework, that can simulate performance-power-temperature of a ReRAM-cache based system, we integrated the gem5, McPAT, NVSim, and HotReRAM simulators. We integrated McPAT to simulate power of the core area, whereas NVSim is employed to derive the power consumption of the ReRAM based caches. However, this work is now at its final stage and yet to be completed. We are currently verifying the correctness of the thermal model of the HotReRAM we developed. This simulation framework will be released as an open-source tool for public use upon its completion. The respective documentation is still under process and will shortly be communicated for the publication.

WP2: Loop Investigation and Wear Leveling

We started with investigating performance monitoring counters (PMCs) in determining on-chip thermal status for contemporary FinFET based multi-cores. The outcomes have been published in the *CF'22* conference, and our investigation was performed collaboratively with Cyprus University of Technology.

We also explored the write limitations of STT-RAM based caches, and a basic write-aware block migration technique was developed and published in the *ASAP'21* conference. The extended version of this work is still under review, in which we investigated the thermal-aware block management for STT-RAM to

improve the lifetime of the device, along with a novel solution to combat the write endurance. In fact, our solution prudentially manages the writes counts across the cache while maintaining thermal safety. The simulation shows how this novel approach surpasses the state-of-the-art techniques. However, to the best of our knowledge, this is the first write balancing mechanism that also considers thermal properties of the STT-RAM based caches. We also investigated the impact on fundamental properties (write endurance, retention times, power/thermal efficiency, etc.) of NVM based shared caches in state-of-the-art heterogeneous systems (equipped with CPUs and GPUs). The preliminary results have been documented in a paper, which has been accepted for publication in premium conference, *DAC 2023*. The detailed investigation that includes comprehensive analysis of the thermal properties of STT-RAM is in its final stage and yet to be communicated on completion. We collaborated with IIT-BHU (India) and the University of Edinburgh (UK) to investigate the STT-RAM's write limitations and its thermal properties with development of novel mitigation techniques.

WP3: Optimal data placement strategy
We are currently developing a compiler based write intensive data segregation technique jointly with the Florida State University (USA). On completion of this work, the write limitations of the multi-retention STT-RAM cache will be ameliorated by placing the blocks at the appropriate retention zones by developing novel architectural as well as system programming (that includes compilers and linkers) based mechanisms. This work is still under progress and will be finished tentatively by the end of April 2023. On completion, we will communicate the outcomes at the premium conference and/or journal.

WP4: Dissemination & Exploitation
During the entire project duration, we have communicated our outcomes for publications at premium conferences and journals, out of which we have published 4 conferences and 6 journals, and some of the outcomes are still under review. Note that, conferences are the premier venue for the fields of Computer Architecture, Optimizing Compilers, and Digital Design with the top-tier conferences being much more prestigious than journals in the same field. We have presented our work at the conferences like *ASAP'21*, *CF'22*, *CODES+ISSS'21* and HiPEAC 2022, and will present our recently accepted paper in *DAC 2023* in July 2023. The work has also been presented in several weekly seminars of the *CAL* and *EECS* research groups at the *Department of Computer Science, NTNU*, and at our collaborators' research groups in Cyprus, UK, USA, and India.

Details on the (non-scientific) management activities of the project

1. Was the researcher involved in all management aspects of the fellowship?
   » *Yes, the researcher was involved in all management aspects of the fellowship.*

2. Did the researcher manage the financial part of the project?
   » *The researcher was fully aware regarding management of the financial part of the project. NTNU (the host institute) has efficient teams who take care about this both at the department level and at the faculty level.*

3. Did the researcher receive support from the administrative staff at the host institution?
   » *Yes, in all aspects he received support from the administrative staffs of NTNU, the host institute.*

4. How was the integration of the researcher within the host/department? Did the researcher supervise Master/PhD students?
   » *The researcher was very much integrated within the host, especially with the Computing research group of the Department of Computer Science, NTNU. In fact, before his MSCA-IF fellowship period, he worked within the same group for 2 years as a post-doctoral scholar with ERCIM fellowship. He also presented a lot of his research outcomes at several weekly seminars of the group. He is currently working with one PhD scholar in the department at the host institute, and is remotely supervising one PhD scholar in India.*

5. Were there weekly meetings with the supervisor?
   » *Yes, there were weekly meetings with the supervisor.*

6. Was the researcher involved in setting up external collaboration (if any), and in the publication of the results?
   » *The researcher has established many international (academic and industrial) collaborations. He has successfully published many journal articles and conference papers with his collaborators.*

Impact

MSCA-IF certainly boosted the researcher's career. At present Shounak is working as a *Researcher* at the same research group at the **Department of Computer Science, NTNU, Norway,** while focusing in the same research domain. The research experience gained during the fellowship has also helped him to get the *Computer System Architect*[4] position at **ZeroPoint Technologies AB,**

[4]received offer in *June 2022*, and joined in *January 2023*

Gothenburg, Sweden, a well known organization working towards improving power efficiency of the memory systems by incorporating novel memory compression techniques. He has also established many research collaborations with several universities in UK, USA, Cyprus, India, Sweden, Finland, and with industries. Currently, the fellow is also looking forward to apply for the *ERC Starting Grant*, *FRIPRO grant of The Research Council of Norway*, *UKRI-RCN Collaborative Research Grant*, and also looking for some higher academic positions.

The outcomes of the TECTONIC project includes a simulation framework for ReRAM based systems, which can simulate performance-power-thermal status of the underlying architectural components with cycle accuracy. Unavailability of such a simulation infrastructure was a prime bottleneck for advancement of the NVM research from a thermal perspective. This simulator will enable researchers in both academia and industry to carry on several design space exploration before deploying their final product/IP. Moreover, our other explorations in the field of NVMs and thermal management has the potential to initiate future research avenues. Additionally, during the fellowship, Shounak explored on-chip heat transfer for the modern chip multi-processors and proposed techniques to combat the thermal breakdown[5]. These explorations will enable designers and architects to strengthen their next generation chips to mitigate the thermal issues, and offer a set of future research directions for on-chip thermal management.

During the fellowship period, Shounak collaboratively (with University of Essex, UK, and IIIT Guwahati, India) developed several thermal management policies[6] for contemporary multi-cores, targeted for real-time system paradigms. As technology scaling has made the power and thermal aspects the prime design concern for chip designers and architects, managing temperature while processing real-time tasks will be a fundamental requirement for the automa-

[5]Published articles/conference papers on on-chip thermal management:

- Shounak Chakraborty, Vassos Soteriou, Magnus Själander, "STIFF: thermally safe temperature effect inversion aware FinFET based multi-core." CF, 2022.
- Shounak Chakraborty, Magnus Själander, "WaFFLe: Gated Cache-Ways with Per-Core Fine-Grained DVFS for Reduced On-Chip Temperature and Leakage Consumption." ACM TACO 2021.

[6]Published articles/conference papers on thermal management for the real-time systems:

- Sangeet Saha, Shounak Chakraborty, Sukarn Agarwal, Rahul Gangopadhyay, Magnus Själander, Klaus D. McDonald-Maier, "DELICIOUS: Deadline-Aware Approximate Computing in Cache-Conscious Multicore." IEEE TPDS, 2023.
- Yanshul Sharma, Sanjay Moulik, Shounak Chakraborty, "RESTORE: Real-Time Task Scheduling on a Temperature Aware FinFET based Multicore" DATE, 2022.
- Yanshul Sharma, Shounak Chakraborty, Sanjay Moulik, "ETA-HP: an energy and temperature-aware real-time scheduler for heterogeneous platforms" Journal of Supercomputing, 2022.
- Sangeet Saha, Shounak Chakraborty, Xiaojun Zhai, Shoaib Ehsan, Klaus D. McDonald-Maier, "ACCURATE: Accuracy Maximization for Real-Time Multicore Systems With Energy-Efficient Way-Sharing Caches" IEEE TCAD, 2022.
- Shounak Chakraborty, Sangeet Saha, Magnus Själander, Klaus D. McDonald-Maier, "Prepare: Power-Aware Approximate Real-time Task Scheduling for Energy-Adaptive QoS Maximization" ACM TECS, 2021 (accepted in CODES+ISSS 2021).

tion industry, so for the society.

Finally, it can be stated that, completion of TECTONIC will enhance the research scope in the domain of NVMs. Concurrently, the thermal management of the contemporary real-time systems will also open up numerous opportunities for designers and architects to build energy-efficient, high-performance, next-generation computing systems.

Note that, the publications related to TECTONIC followed the *EU's open access policy for research*[7]. All of these open-access documents, publications (IEEE/ACM), and presentations acknowledged the MSCA-IF grant in accordance with the H2020 mandate (EU) and the NTNU's IPR policy. The future publications, related to TECTONIC, will also follow the same.

Maintaining thermal safety while executing heavy computational workloads is one of the biggest challenge for the software and hardware industries. Implementation of TECTONIC can also contribute towards **European Policy Objectives**, directly or indirectly, as TECTONIC improves the performance of modern computing systems while maintaining thermal safety. Hence, implementing TECTONIC can be useful for efficient execution of certain applications having heavy computational workloads, like, Weather Prediction or Climate Action, Traffic Management Software for Smart Cities and Railway Communication, State-of-the-art Communication Devices, Healthcare, etc., which are basically the significant parts of the **European Policy Objectives**.

2 Update of the plan for exploitation and dissemination of results (if applicable)

- **List the conferences attended:**
  - 32nd IEEE International Conference on Application-specific Systems, Architectures and Processors (ASAP 2021)
  - International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS'21)
  - 19th ACM International Conference on Computing Frontiers (CF '22)
  - Design, Automation & Test in Europe Conference & Exhibition (DATE 2022)
  - HiPEAC conference 2022
- **List of conferences to be attended:** Design Automation Conference 2023 (DAC 2023) in July 2023.

[7] for example, this is an open access paper of the researcher from the TECTONIC project:
Shounak Chakraborty, Magnus Själander, "WaFFLe: Gated Cache-Ways with Per-Core Fine-Grained DVFS for Reduced On-Chip Temperature and Leakage Consumption." ACM TACO 2021.

- *Did you disseminate project results in **scientific publications** as planned in – or in addition to – the DoA (including the deposition of publications in open access repositories)? Do they include a reference to EU funding?:* Yes, the project results are disseminated in the form of publications which include a reference to EU funding, and they automatically follow the NTNU's open access policy for the published research outcomes.
- List all the **outreach activities** undertaken (visit to schools, Researchers' Night, etc.): presented several research outcomes at (i) the weekly seminars of the CAL research group at NTNU, (ii) research talks at IIIT Guwahati (India), (iii) several research meetings with Newcastle University (UK), University of Essex (UK), The University of Edinburgh (UK), Cyprus University of Technology (Cyprus), and Florida State University (USA).
- *Did you disseminate and communicate project activities and results by **other means than scientific publications** (social media, press-release, the project web site, video/film, etc.) as planned in – or in addition to – the DoA? Do they include a reference to EU funding?* Project results are published for a broader audience at the known professional social network, like *LinkedIn* by mentioning about the associated EU funding.

Note that, all future publications related to TECTONIC will also acknowledge the associated EU funding.

3 Update of the Data Management Plan (if applicable)

In case of the TECTONIC project, a set of outcomes are already available in the form of publications, which also follow the NTNU's open access policy. Additionally, the research results can be reproduced by following the experimental setup elaborated in each of the published items. All of the research publications related to TECTONIC, those are currently either under review or under preparation will also contain the research outcomes in detail (which can also be reproduced) and will also follow the NTNU's open access policy. TECTONIC's outcomes further include a simulator, which will be released soon as open source, that will not only help generate several of the research outcomes of TECTONIC, but will also enable the architects and researchers to verify and deploy emerging NVM based memories.