SAFARI
SAFARI Research Group

Newsletter
January 2021

Think Big, Aim High, and Have a Wonderful 2021!

Dear SAFARI friends,

Happy New Year! We are excited to share our group highlights with you in this second edition of the SAFARI newsletter (You can find the first edition from April 2020 **here**). 2020 has brought many challenges and changes to all of our lives. We have adapted our scientific research and teaching to an online global presence. We continue to test and find creative ways to continue collaborations in our research and improve our online education. Thanks to the efforts of many people who have organized various events, we have had the pleasure of contributing to many virtual global conferences and meetings, spanning topics in computer architecture, computing systems, circuits and devices, bioinformatics, security, reliability, machine learning, reconfigurable computing, and more. We have continued to enjoy having fruitful interactions with our collaborators around the globe. We have also been honored and humbled to receive multiple prestigious awards, like the Best Paper Award at **MICRO 2020** and **IEEE Security and Privacy 2020** conferences, as well as the **Pwnie Award** for Most

Innovative Research.

Our new building, ETZ, Gloriastrasse 35

Our group moved to the **Department of Information Technology and Electrical Engineering** (aka Electrical and Computer Engineering, D-ITET) in early October, where we have more space and a brand-new hardware research lab. We are very positive about this change and look forward to developing new and diverse collaborations across the computing stack, all the way from algorithms to devices, with our colleagues in our new department.

We have also expanded our educational activities at the bachelor and master's levels, adding new **hands-on project courses** in memory systems, processing in memory and genome analysis. We will continue to be affiliated with the Department of Computer Science, extending our collaborations and continuing to teach our freshman-level **Digital Design and Computer Architecture** class.

In this second edition of the SAFARI newsletter, we share our research, teaching and outreach highlights from 2020, and look ahead to a new and inspiring future in 2021.

We wish you a wonderful 2021, in all aspects of your lives!

Onur Mutlu

Fun *SAFARI* fact: Have you ever wondered what SAFARI actually means? You are not alone. Many people have asked us, and we would like to share the meaning of SAFARI with you. SAFARI is the name first given to the research group Onur Mutlu started at Carnegie Mellon University in 2009. It originally stood for the research vision of the group at the time: SAfe, FAir, Robust and Intelligent computer architectures! The vision still forms a part of the research we do, but the group's focus has expanded over the years. Onur likes to think about the group's research as a SAFARI for new ideas and breakthroughs in computer architecture and bioinformatics.

News & Awards

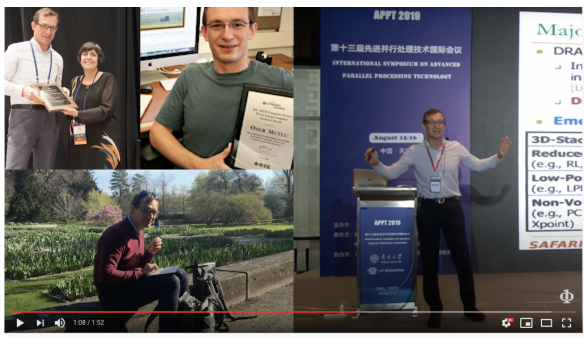
BEER wins the Best Paper Award at MICRO 2020

Congratulations to **Minesh Patel, Jeremie Kim, Taha Shahroodi, Hasan Hassan, and Onur Mutlu** for the Best Paper Award at this year's IEEE/ACM International Symposium on Microarchitecture for their paper:

Bit-Exact ECC Recovery (BEER): Determining DRAM On-Die ECC Functions by Exploiting DRAM Data Retention Characteristics

MICRO Oct 2020

Slides (pptx) (pdf)
Lightning Talk Slides (pptx) (pdf)
Talk Video (15 minutes)
Short Talk Video (5.5 minutes)
Lightning Talk Video (1.5 minutes)
Lecture Video (52.5 minutes)
BEER Source Code
Paper



Onur Mutlu Awarded 2020 Edward J. McCluskey Technical Achievement

TRRespass wins the Best Paper Award at IEEE S&P

TRRespass wins the Pwnie Award for "Most Innovative Research"

Congratulations to our co-authors and collaborators at the Systems and Network Security Group at VU Amsterdam: **Pietro Frigo, Emanuele Vannacci, Hasan Hassan, Victor van der Veen, Onur Mutlu, Cristiano Giuffrida, Herbert Bos, and Kaveh Razavi**

TRRespass: Exploiting the Many Sides of Target Row Refresh

IEEE S&P May 2020

Lecture Video (59 minutes), Onur Mutlu
Lecture Slides (pptx) (pdf), Onur Mutlu
TRRespass Source Code
Project Overview
Paper
SAFARI News Best Paper Award S&P
SAFARI News Pwnie Award

Nomination for the Stamatis Vassiliadis Memorial Best Paper Award at FPL 2020!

Congratulations to **Gagandeep Singh** and co-authors on their nomination for the

IEEE Computer Society Edward J. McCluskey Technical Achievement Award 2020 video

Onur received the IEEE Computer Society prestigious 2020 Edward J. McCluskey Technical Achievement Award for "innovative and impactful contributions to computer memory systems." In honour of this achievement, the IEEE Computer Society have produced a commemorative video.

Watch the video

IEEE Computer Society News

Jisung Park was awarded a Postdoctoral Research Fellowship from the National Research Foundation of Korea for his project on "Storage System Design for Machine Learning Applications". The NRF award "Fostering the Next Generation of Researchers" supports postdocs to maintain research continuity and foster their independent research capacity.

co-authors on their nomination for the Stamatis Vassiliadis Memorial Award at FPL2020!

Gagandeep Singh, Dionysios Diamantopoulos, Christoph Hagleitner, Juan Gómez-Luna, Sander Stuijk, Onur Mutlu, and Henk Corporaal

NERO: A Near High-Bandwidth Memory Stencil Accelerator for Weather Prediction Modeling

FPL September 2020

Talk Video

Paper

Mohammad Sadrosadati, an affiliated postdoc with SAFARI, was awarded the Khwarizmi Youth Award. The award is a national version of the Khwarizmi International Award, open to Iranians under 35 years old "to honor young scientists and embolden them to keep taking even bigger steps in their research career". Congratulations Mohammad!

Read more

Onur was elected ACM SIGMICRO Chair!

ACM SIGMICRO Officers

ACM SIG Elections

Congratulations Jisung!

Read more

Congratulations to all our SAFARI group members on their HiPEAC Paper Awards!

ASPLOS 2020

Evanesco: Architectural Support for Efficient Data Sanitization in Modern Flash-Based Storage Systems

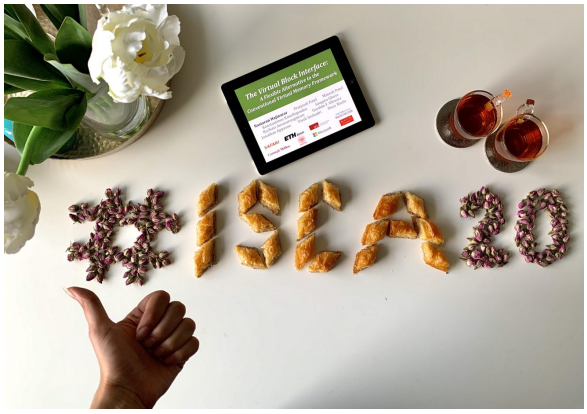
Myungsuk Kim, Jisung Park, Genhee Cho, Yoona Kim, Lois Orosa, Onur Mutlu, Jihong Kim

HPCA 2020

Techniques for Reducing the Connected-Standby Energy Consumption of Mobile Devices

Jawad Haj-Yahya, Yanos Sazeides, Mohammed Alser, Efraim Rotem, Onur Mutlu

---



Congratulations to **Nastaran Hajinazar** for the Best Picture Award at ISCA 2020!

Read more SAFARI news

ISCA 2020

CLR-DRAM: A Low-Cost DRAM Architecture Enabling Dynamic Capacity-Latency Trade-Off

Haocong Luo, Taha Shahroodi, Hasan Hassan, Minesh Patel, Abdullah Giray Yaglikçi, Lois Orosa, Jisung Park, Onur Mutlu

Revisiting RowHammer: An Experimental Analysis of Modern DRAM Devices and Mitigation Techniques

Jeremie S. Kim, Minesh Patel, Abdullah Giray Yaglikçi, Hasan Hassan, Roknoddin Azizi, Lois Orosa, Onur Mutlu

SysScale: Exploiting Multi-domain Dynamic Voltage and Frequency Scaling for Energy Efficient Mobile Processors

Jawad Haj-Yahya, Mohammed Alser, Jeremie S. Kim, Abdullah Giray Yaglikçi, Nandita Vijaykumar, Efraim Rotem, Onur Mutlu

The Virtual Block Interface: A Flexible Alternative to the Conventional Virtual Memory Framework

Nastaran Hajinazar, Pratyush Patel, Minesh Patel, Konstantinos Kanellopoulos, Saugata Ghose, Rachata Ausavarungnirun, Geraldo F. Oliveira, Jonathan Appavoo, Vivek Seshadri, Onur Mutlu

MICRO 2020

GenASM: A High-Performance, Low-Power Approximate String Matching Acceleration Framework for Genome Sequence Analysis

Damla Senol Cali, Gurpreet S. Kalsi, Zulal Bingol, Can Firtina, Lavanya Subramanian, Jeremie S. Kim, Rachata Ausavarungnirun, Mohammed Alser, Juan Gomez-Luna, Amirali Boroumand, Anant Nori, Allison Scibisz, Sreenivas Subramoney, Can Alkan, Saugata Ghose, and Onur Mutlu

MICRO 2020

Bit-Exact ECC Recovery (BEER): Determining DRAM On-Die ECC Functions by Exploiting DRAM Data Retention Characteristics

Minesh Patel, Jeremie S. Kim, Taha Shahroodi, Hasan Hassan, and Onur Mutlu

FIGARO: Improving System Performance via Fine-Grained In-DRAM Data Relocation and Caching

Yaohua Wang, Lois Orosa, Xiangjun Peng, Yang Guo, Saugata Ghose, Minesh Patel, Jeremie S. Kim, Juan Gómez Luna, Mohammad Sadrosadati, Nika Mansouri Ghiasi, and Onur Mutlu

FlexWatts: A Power- and Workload-Aware Hybrid Power Delivery Network for Energy-Efficient Microprocessors

Jawad Haj-Yahya, Mohammed Alser, Jeremie S. Kim, Lois Orosa, Efraim Rotem, Avi Mendelson, Anupam Chattopadhyay, and Onur Mutlu

---

We thank our industry partners for their continued and new support in 2020!

Our group was generously supported with new donations this year from Google, Huawei, Intel, SRC and VMware. We thank all of our industry partners for their continued support of our research, and our partners who have generously supported us with new donations this year. We also wish to thank the **ETH Foundation** for facilitating and supporting our industry relations and gift funding.

ASML
Facebook
Google
HiSilicon
Huawei
IMEC
Intel
Microsoft
SRC
VMware

We always welcome generous donations and gifts to support our research. Please contact **Onur** for more information on donating and being a part of SAFARI supporters.

Selected Recent Publications

A Modern Primer on Processing in Memory

Onur Mutlu, Saugata Ghose, Juan Gomez-Luna, and Rachata Ausavarungnirun

Invited Book Chapter in Emerging Computing: From Devices to Systems - Looking Beyond Moore and Von Neumann, Springer, to be published in 2021.

Tutorial Video on "Memory-Centric Computing Systems" (1 hour 51 minutes)

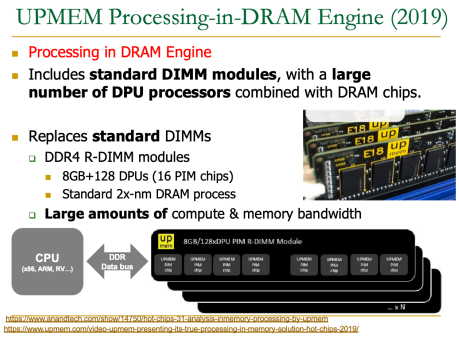
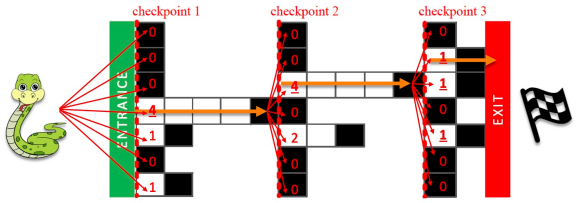


Figure 26: UPMEM PIM architecture and hardware. Reproduced from [214].

SneakySnake: A Fast and Accurate Universal Genome Pre-Alignment Filter for CPUs, GPUs, and FPGAs

Mohammed Alser, Taha Shahroodi, Juan-Gomez Luna, Can Alkan, and Onur Mutlu

Bioinformatics, December 2020.
Paper PDF
Paper in Bioinformatics
Source Code



An example of the chip maze for a reference sequence *R* = 'GGTGCAGAGCTC', a query sequence *Q*= 'GGTGAGAGTTGT', and an edit distance threshold (*E*) of 3. Our SneakySnake algorithm *quickly* finds an optimal signal net (highlighted in orange) with 3 obstacles, each of which is located at the end of each escape segment, and hence SneakySnake decides that sequence alignment for *R* & *Q* is needed, as the number of obstacles ≤ *E*.

Accelerating Genome Analysis: A Primer on an Ongoing Journey

Mohammed Alser, Zulal Bingol, Damla Senol Cali, Jeremie Kim, Saugata Ghose, Can Alkan, and Onur Mutlu

IEEE MICRO, Volume: 40, Issue: 05, Pages: 65-75, September/October 2020.
Paper PDF
Paper in IEEE Micro
Slides (pptx) (pdf)

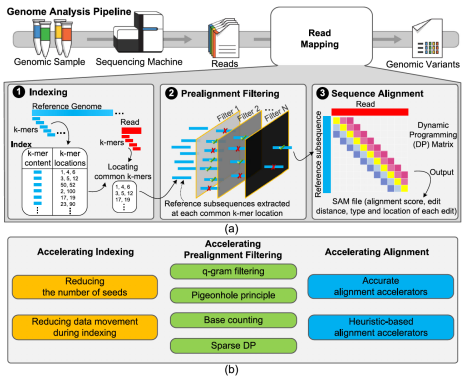
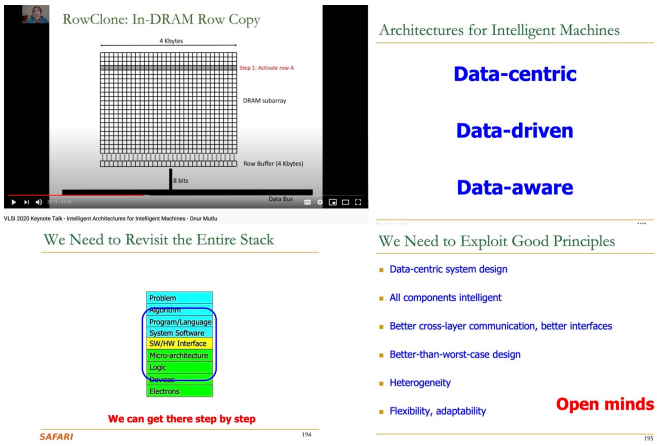


Figure 1. (a) Three steps of read mapping in genome analysis: 1) indexing, 2) pre-alignment filtering, and 3) sequence alignment. (b) Overview of the existing approaches to accelerating each step of read mapping.

Intelligent Architectures for Intelligent Machines

Onur Mutlu

Keynote Paper in Proceedings of the **2020 International Symposia on VLSI** (VLSI), Hsinchu City, Taiwan, August 2020.
Paper
Slides (pptx) (pdf)
Keynote Talk Video



Bit-Exact ECC Recovery (BEER): Determining DRAM On-Die ECC Functions by Exploiting DRAM Data Retention Characteristics

Minesh Patel, Jeremie S. Kim, Taha Shahroodi, Hasan Hassan, and Onur Mutlu

Proceedings of the 53rd International Symposium on Microarchitecture (MICRO).
Virtual, October 2020.

Best Paper Award!

Paper
Slides (pptx) (pdf)
Short Talk Slides (pptx) (pdf)
Talk Video (15 minutes)
Short Talk Video (5.5 minutes)
Lightning Talk Video (1.5 minutes)
Lecture Video (52.5 minutes)
BEER Source Code

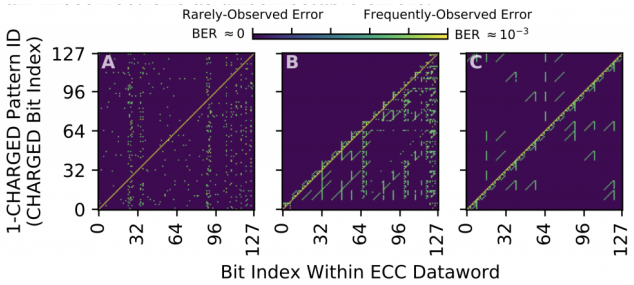


Figure 3: Errors observed in a single representative chip from each manufacturer using the 1-CHARGED test patterns, showing that manufacturers appear to use different ECC functions.

GenASM: A High-Performance, Low-Power Approximate String Matching Acceleration Framework for Genome Sequence Analysis

Damla Senol Cali, Gurpreet S. Kalsi, Zulal Bingol, Can Firtina, Lavanya Subramanian, Jeremie S. Kim, Rachata Ausavarungnirun, Mohammed Alser, Juan Gomez-Luna, Amirali Boroumand, Anant Nori, Allison Scibisz, Sreenivas Subramoney, Can Alkan, Saugata Ghose, and Onur Mutlu

Proceedings of the 53rd International Symposium on Microarchitecture (MICRO).
Virtual, October 2020.

Paper
Slides (pptx) (pdf)
Short Talk Slides (pptx) (pdf)
MICRO 2020 Talk Video
MICRO 2020 Lighting Talk Video
ARM Research Summit Talk Video
ARM Research Summit Short Talk Video and Q&A
Lecture Video
GenASM Source Code

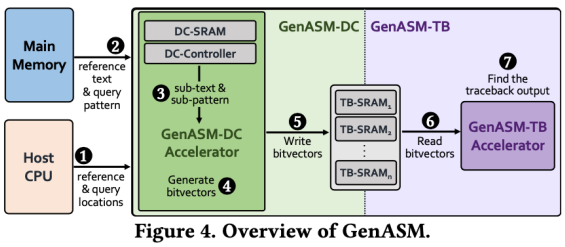


Figure 4. Overview of GenASM.

FIGARO: Improving System Performance via Fine-Grained In-DRAM Data Relocation and Caching

Yaohua Wang, Lois Orosa, Xiangjun Peng, Yang Guo, Saugata Ghose, Minesh Patel, Jeremie S. Kim, Juan Gómez Luna, Mohammad Sadrosadati, Nika Mansouri Ghiasi, and Onur Mutlu

Proceedings of the 53rd International Symposium on Microarchitecture (MICRO).
Virtual, October 2020.

Read more
Slides (pptx) (pdf)
Short Talk Slides (pptx) (pdf)
Talk Video
Lightning Talk Video
Short Talk Video
Lecture Video

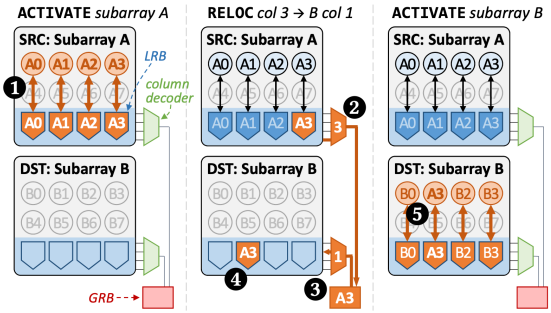


Figure 4. An example of data relocation using FIGARO.

FlexWatts: A Power- and Workload-Aware Hybrid Power Delivery Network for Energy-Efficient Microprocessors

Jawad Haj-Yahya, Mohammed Alser, Jeremie S. Kim, Lois Orosa, Efraim Rotem, Avi Mendelson, Anupam Chattopadhyay, and Onur Mutlu

Proceedings of the 53rd International Symposium on Microarchitecture (MICRO).
Virtual, October 2020.

Read more
Slides (pptx) (pdf)
Short Talk Slides (pptx) (pdf)
Talk Video
Lightning Talk Video
Short Talk Video and Q&A

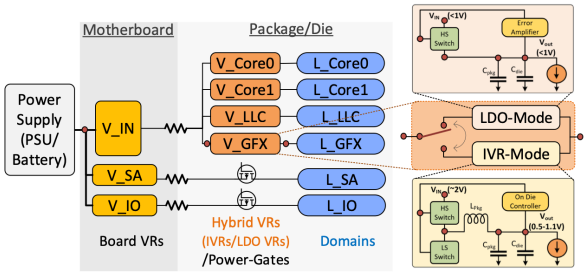


Figure 6: Our hybrid adaptive PDN (FlexWatts). FlexWatts uses an off-chip VR to each system domain with a low and narrow power consumption range (i.e., SA and IO domains). For

NATSA: A Near-Data Processing Accelerator for Time Series Analysis

Ivan Fernandez, Ricardo Quislant, Christina Giannoula, Mohammed Alser, Juan Gómez-Luna, Eladio Gutiérrez, Oscar Plata, and Onur Mutlu

Proceedings of the 38th IEEE International Conference on Computer Design (ICCD), Virtual, October 2020.

Paper

Slides (pptx) (pdf)

Talk Video

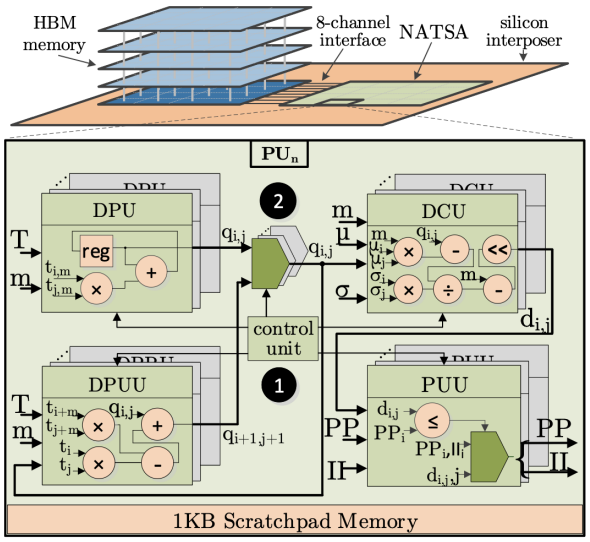


Figure 5: NATSA design and integration next to HBM memory. NATSA is connected directly to the HBM interface.

WoLFRaM: Enhancing Wear-Leveling and Fault Tolerance in Resistive Memories using Programmable Address Decoders

Leonid Yavits, Lois Orosa, João Dinis Ferreira, Mattan Erez, Ran Ginosar, and Onur Mutlu

Proceedings of the 38th IEEE International Conference on Computer Design (ICCD), Virtual, October 2020.

Paper

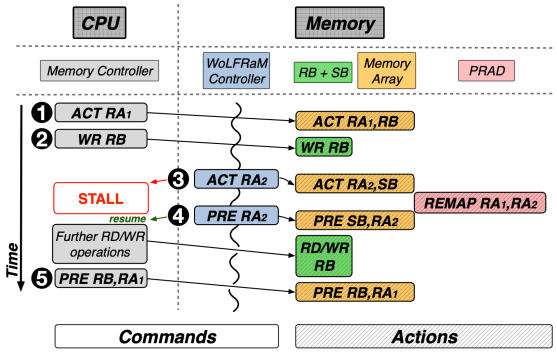
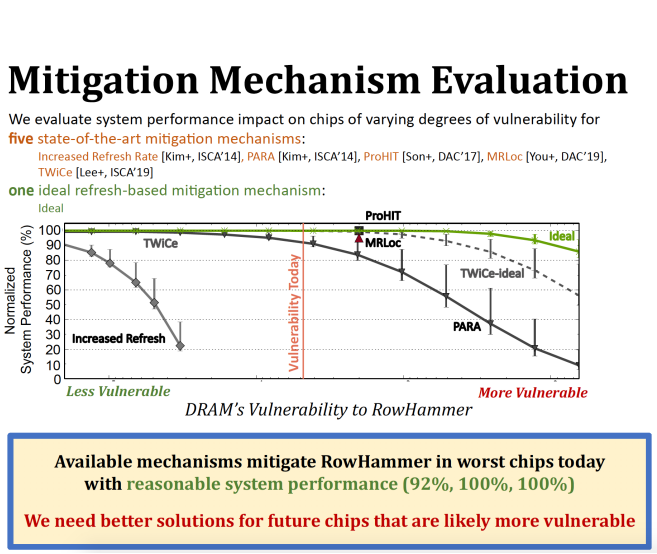


Figure 5: Sequence of commands issued by the memory controller and the WoLFRaM controller to remap and swap two memory blocks.

Revisiting RowHammer: An Experimental Analysis of Modern Devices and Mitigation Techniques

Jeremie S. Kim, Minesh Patel, A. Giray Yaglikci, Hasan Hassan, Roknoddin Azizi, Lois Orosa, and Onur Mutlu

Proceedings of the 47th International Symposium on Computer Architecture (ISCA), Valencia, Spain, June 2020.
Paper
ARM Research Summit Poster
Slides (pptx) (pdf)
Lightning Talk Slides (pptx) (pdf)
Talk Video
Lightning Talk Video
Lecture Video



CLR-DRAM: A Low-Cost DRAM Architecture Enabling Dynamic Capacity-Latency Trade-Off

Haocong Luo, Taha Shahroodi, Hasan Hassan, Minesh Patel, A. Giray Yaglikci, Lois Orosa, Jisung Park, and Onur Mutlu,

Proceedings of the 47th International Symposium on Computer Architecture (ISCA), Valencia, Spain, June 2020.
Paper
Slides (pptx) (pdf)
Lightning Talk Slides (pptx) (pdf)
Talk Video
Lightning Talk Video

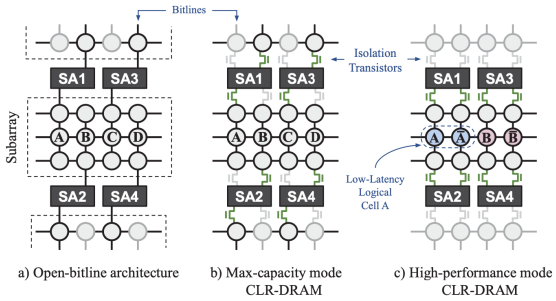


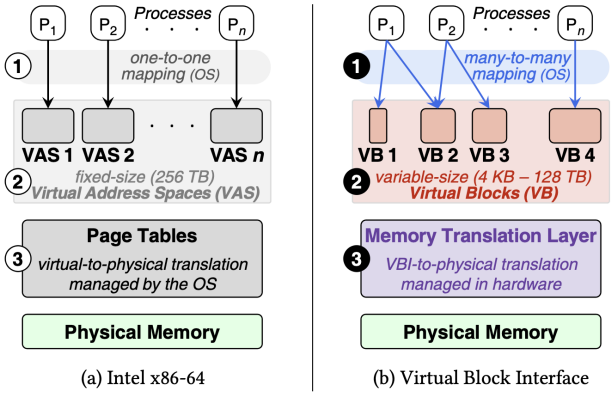
Figure 1: Comparison of a) the conventional open-bitline architecture and our CLR-DRAM architecture operating in b) max-capacity mode and c) high-performance mode. Isolation transistors highlighted in green are enabled, faded grey disabled.

The Virtual Block Interface: A Flexible Alternative to the Conventional Virtual Memory Framework

Nastaran Hajinazar, Pratyush Patel, Minesh Patel, Konstantinos Kanellopoulos, Saugata Ghose, Rachata Ausavarungnirun, Geraldo Francisco de Oliveira Jr., Jonathan Appavoo, Vivek Seshadri, and Onur Mutlu

Proceedings of the 47th International Symposium on Computer Architecture (ISCA), Valencia, Spain, June 2020.

Paper
ARM Research Summit Poster
Slides (pptx) (pdf)
Lightning Talk Slides (pptx) (pdf)
Talk Video
Lightning Talk Video
Lecture Video



(a) Intel x86-64 (b) Virtual Block Interface

Figure 1: Virtual memory management in x86-64 and in VBI.

SysScale: Exploiting Multi-domain Dynamic Voltage and Frequency Scaling for Energy Efficient Mobile Processors

Jawad Haj-Yahya, Mohammed Alser, Jeremie Kim, A. Giray Yaglikci, Nandita Vijaykumar, Efraim Rotem, and Onur Mutlu,

Proceedings of the 47th International Symposium on Computer Architecture (ISCA), Valencia, Spain, June 2020.

Paper
Slides (pptx) (pdf)
Lightning Talk Slides (pptx) (pdf)
Talk Video
Lightning Talk Video

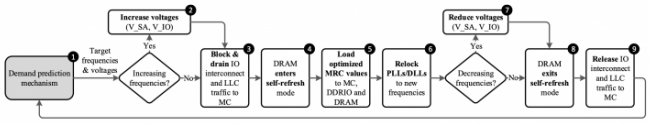


Figure 5: SysScale power management flow which carry out the DVFS of SoC domains.

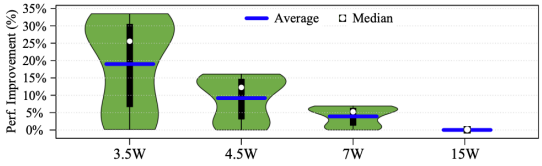


Figure 10: SysScale performance benefit vs. TDP on SPEC CPU2006 workloads.

Refresh Triggered Computation: Improving the Energy Efficiency of Convolutional Neural Network Accelerators

Syed M. A. H. Jafri, Hasan Hassan, Ahmed Hemani, and Onur Mutlu

ACM Transactions on Architecture and Code Optimization (TACO), December 2020.
Paper PDF
Paper in TACO

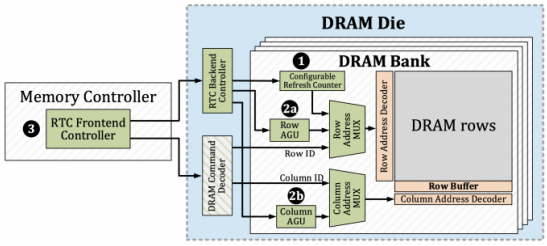


Fig. 6. Modifications in the memory controller and DRAM to support Full-RTC

Highly Concurrent Latency-Tolerant Register Files for GPUs

Mohammad Sadrosadati, Amirhossein Mirhosseini, Seyed Borna Ehsani, Hamid Sarbazi-Azad, Mario Drumond, Babak Falsafi, Rachata Ausavarungnirun, and Onur Mutlu

ACM Transactions on Computer Systems (TOCS), to appear in 2021.
Paper PDF

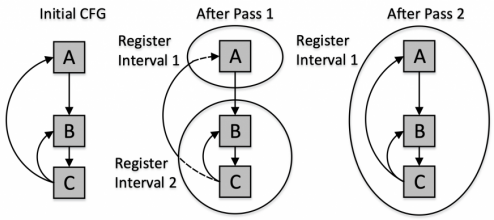


Fig. 5. Register-interval formation for a simple nested loop example. A,B,C represent basic blocks.

TRRespass: Exploiting the Many Sides of Target Row Refresh

Pietro Frigo, Emanuele Vannacci, Hasan Hassan, Victor van der Veen, Onur Mutlu, Cristiano Giuffrida, Herbert Bos, and Kaveh Razavi

Proceedings of the 41st IEEE Symposium on Security and Privacy (S&P), San Francisco, CA, USA, May 2020.

Best paper award

Pwnie Award 2020 for Most Innovative Research

Paper

Lecture Video (59 minutes), Onur Mutlu

Lecture Slides (pptx) (pdf), Onur Mutlu

TRRespass Source Code

Web Article

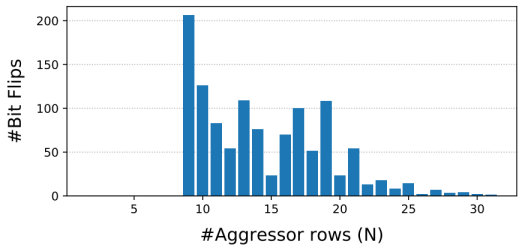


Fig. 13: Bit flips vs. number of aggressor rows. Module $\mathcal{A}_{10}$: Number of bit flips triggered with *N-sided* RowHammer for varying number of *N* on Intel Core i7-7700K. Each aggressor row is one row away from the closest aggressor row (i.e., `VAVAVA`... configuration) and aggressor rows are hammered in a round-robin fashion.

An Experimental Study of Reduced-Voltage Operation in Modern FPGAs for Neural Network Acceleration

Behzad Salami, Erhan Baturay Onural, Ismail Emir Yuksel, Fahrettin Koc, Oguz Ergin, Adrian Cristal Kestelman, Osman S. Unsal, Hamid Sarbazi-Azad, and Onur Mutlu

Proceedings of the 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), Valencia, Spain, June 2020.

Paper

Slides (pptx) (pdf)

Lightning Talk Slides (pptx) (pdf)

Talk Video

Lightning Talk Video

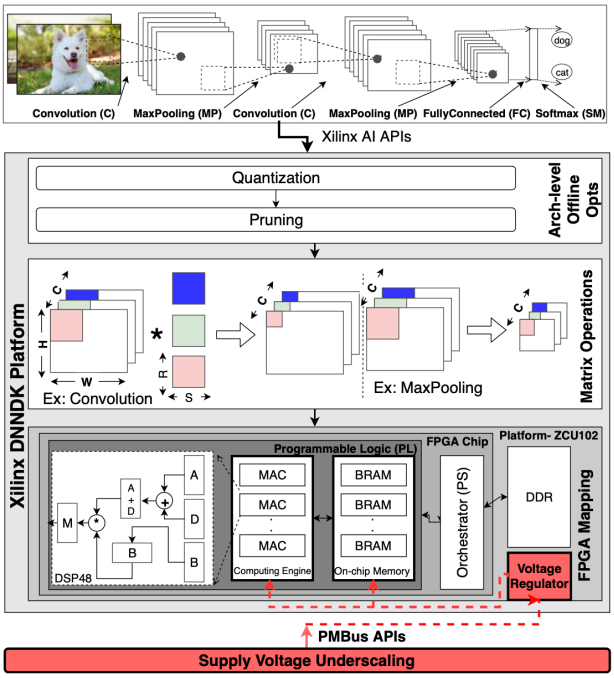


Figure 1: Our overall methodology (for simplicity, we show a simplified block diagram of Xilinx DNNDK [122].).

The Non-IID Data Quagmire of Decentralized Machine Learning

Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip B. Gibbons

Proceedings of the 37th International Conference on Machine Learning (ICML), Virtual, July 2020.

Paper

Slides (pptx) (pdf)

Talk Video

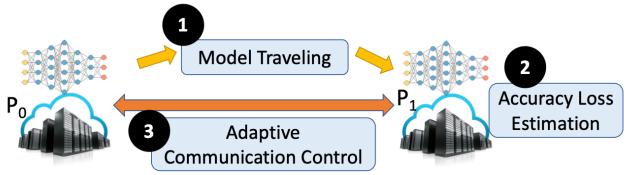


Figure 7. Overview of SkewScout

Are We Susceptible to Rowhammer? An End-to-End Methodology for Cloud Providers

Lucian Cojocar, Jeremie Kim, Minesh Patel, Lillian Tsai, Stefan Saroiu, Alec Wolman, Onur Mutlu

Proceedings of the 41st IEEE Symposium on Security and Privacy (S&P), San Francisco, CA, USA, May 2020.

Paper

Slides (pptx) (pdf)

Talk Video

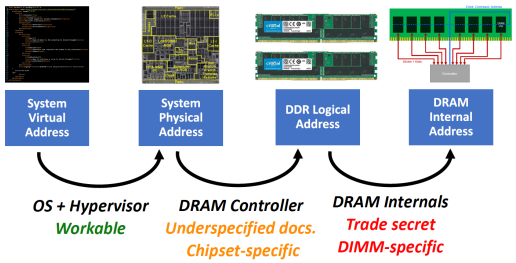


Figure 2: Three remapping layers from virtual address to DRAM internal address.

Optically Connected Memory for Disaggregated Data Centers

Jorge Gonzalez, Alexander Gazman, Maarten Hattink, Mauricio G. Palma, Meisam Bahadori, Ruth Rubio-Noriega, Lois Orosa, Madeleine Glick, Onur Mutlu, Keren Bergman, and Rodolfo Azevedo

Proceedings of the 32nd International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD), Porto, Portugal, September 2020.
Paper
Slides (pptx) (pdf)
Talk Video

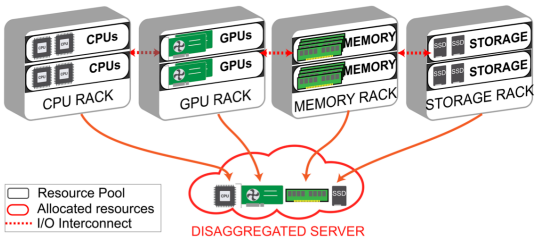


Fig. 1. Disaggregation concept for data centers.

NERO: A Near High-Bandwidth Memory Stencil Accelerator for Weather Prediction Modeling

Gagandeep Singh, Dionysios Diamantopoulos, Christoph Hagleitner, Juan Gómez-Luna, Sander Stuijk, Onur Mutlu, and Henk Corporaal

Proceedings of the 30th International Conference on Field-Programmable Logic and Applications (FPL), Gothenburg, Sweden, September 2020

Nominated for the Stamatis Vassiliadis Memorial Best Paper Award
Paper
Slides (pptx) (pdf)
Talk Video

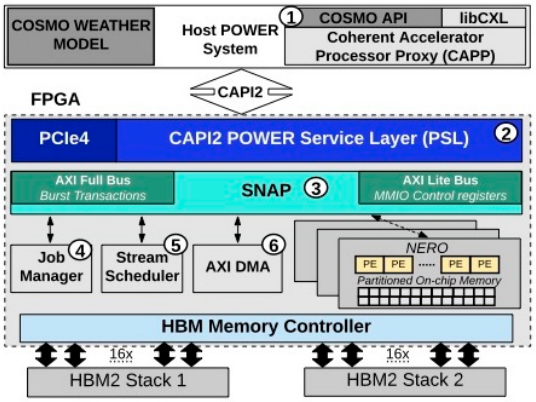


Figure 4: NERO application framework. We co-design our software and hardware using the SNAP framework. COSMO API allows the host to offload kernels to our FPGA platform.

Aging-Aware Request Scheduling for Non-Volatile Main Memory

Shihao Song, Anup Das, Onur Mutlu, and Nagarajan Kandasamy

Proceedings of the 26th Asia and South Pacific Design Automation Conference (ASP-DAC), Virtual, January 2021.
Paper

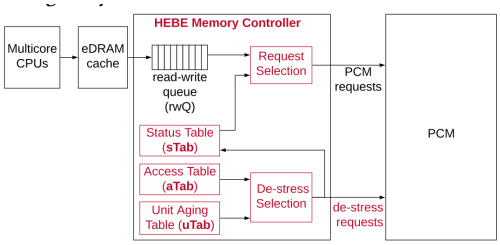


Figure 6: Request and de-stress scheduling in HEBE.

Improving Phase Change Memory Performance with Data Content Aware Access

Shihao Song, Anup Das, Onur Mutlu, and Nagarajan Kandasamy,

Proceedings of theACM SIGPLAN International Symposium on Memory Management (ISMM), London, UK, June 2020.
Paper
Slides (pptx) (pdf)
Talk Video

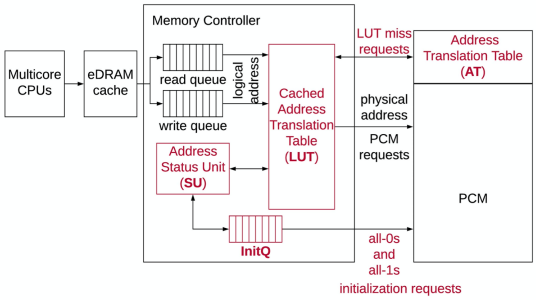


Figure 8. DATACON's address translation mechanism.

Apollo: A Sequencing-Technology-Independent, Scalable, and Accurate Assembly Polishing Algorithm

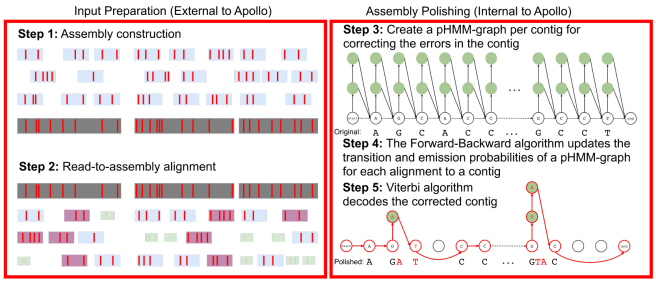
Can Firtina, Jeremie S. Kim, Mohammed Alser, Damla Senol Cali, A. Ercument Cicek, Can Alkan, Onur Mutlu

Bioinformatics, Volume 36, Issue 12, 15 June 2020, Pages 3669–3679

Paper PDF

Paper in Bioinformatics

Source Code



Accelerating B-spline Interpolation on GPUs: Application to Medical Image Registration

Orestis Zachariadis, Andrea Teatini, Nitin Satpute, Juan Gomez-Luna, Onur Mutlu, Ole Jakob Elle, and Joaquin Olivares

Computer Methods and Programs in Biomedicine, 193, September 2020.

Paper PDF

Paper in Computer Methods and Programs in Biomedicine

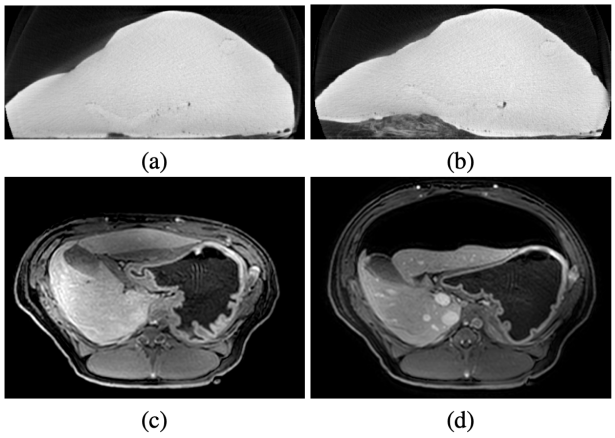


Figure 4: Medical images used for pre-clinical evaluation of our optimized image registration through FFD. (a) and (b) show two DynaCT scans of the liver phantom, and (c) and (d) are MRI scans of the porcine model, respectively without (c) and with pneumoperitoneum applied (d).

NoM: Network-on-Memory for Inter-Bank Data Transfer in Highly-Banked Memories

Seyyed Hossein SeyyedAghaei Rezaei, Mehdi Modarressi, Rachata Ausavarungnirun, Mohammad Sadrosadati, Onur Mutlu, and Masoud Daneshtalab

IEEE Computer Architecture Letters (CAL): 2020, Volume: 19, Issue: 01, Pages: 80-83

Paper

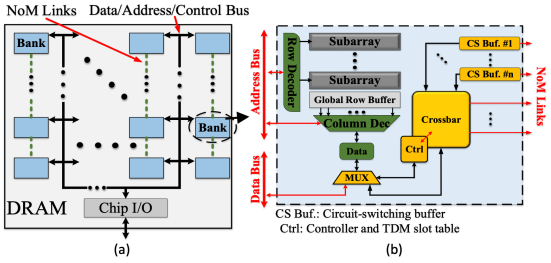


Fig. 1. (a) DRAM structure with NoM links, and (b) The modified architecture of a memory bank. The components that are unique to NoM are colored in yellow

Robust Machine Learning Systems: Challenges, Current Trends, Perspectives, and the Road Ahead

Muhammad Shafique, Mahum Naseer, Theocharis Theocharides, Christos Kyrkou, Onur Mutlu, Lois Orosa, and Jungwook Choi

IEEE Design & Test, Volume: 37, Issue: 2, April 2020

Paper

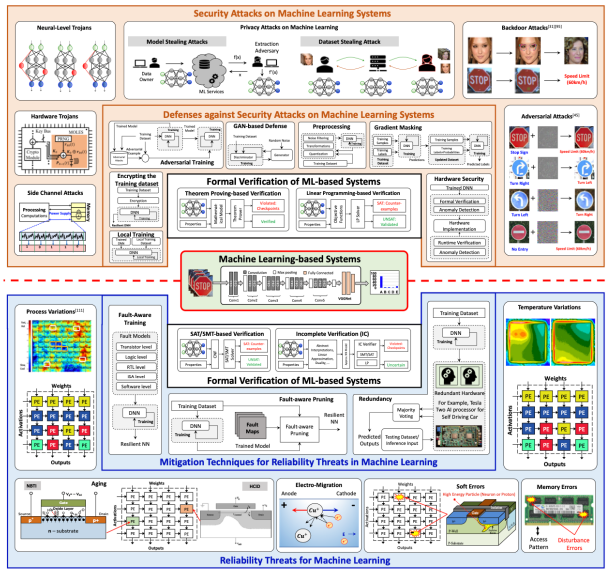
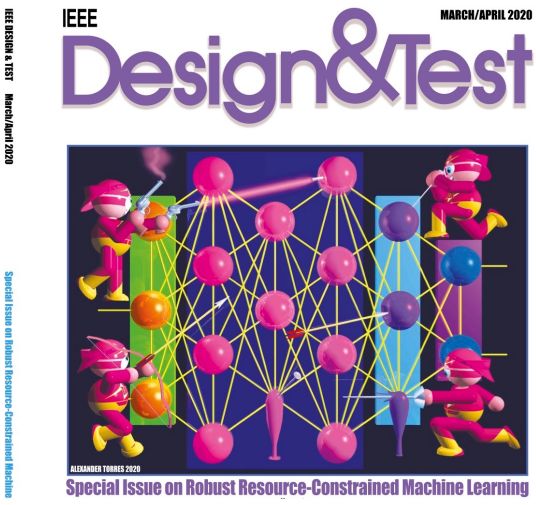


Fig. 1: Overview of threats and challenges associated with ML-based systems: reliability threats and corresponding mitigation techniques (bottom), and security attacks and corresponding defenses (top).

Guest Editorial: Robust Resource-Constrained Systems for Machine Learning

Theocharis Theocharides, Muhammad Shafique, Jungwook Choi, Onur Mutlu

IEEE Design & Test, March/April 2020
Editorial



TADOC: Text Analytics Directly on Compression

Feng Zhang, Jidong Zhai, Xipeng Shen, Dalin Wang, Zheng Chen, Onur Mutlu, Wenguang Chen, and Xiaoyong Du

VLDB Journal, September 2020.
Paper

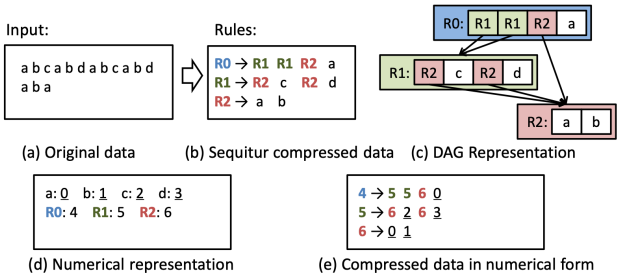


Fig. 1: A compression example with Sequitur.

Enabling Efficient Random Access to Hierarchically-Compressed Data

Feng Zhang, Jidong Zhai, Xipeng Shen, Onur Mutlu, and Xiaoyong Du

Proceedings of the 36th International Conference on Data Engineering (ICDE), Dallas, TX, USA, April 2020.

Paper

Slides (pptx) (pdf)

Talk Video

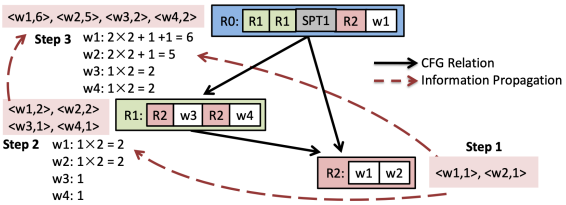


Fig. 2. An example of counting word frequencies with TADOC.

Evanesco: Architectural Support for Efficient Data Sanitization in Modern Flash-Based Storage Systems

Myungsuk Kim, Jisung Park, Geonhee Cho, Yoona Kim, Lois Orosa, Onur Mutlu, Jihong Kim

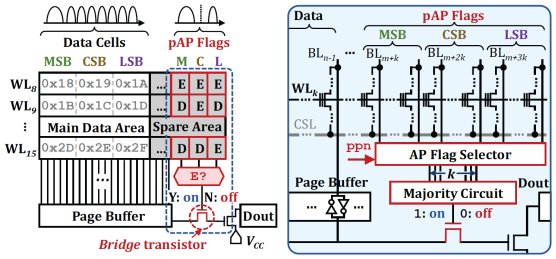
ASPLOS, March 16 – 20 2020, Lausanne, Switzerland

Paper

Slides (pptx) (pdf)

Talk Video

Source Code



(a) Organizational overview. **(b)** A pAP implementation.

Figure 8. Our pLock implementation.

Boyi: A Systematic Framework for Automatically Deciding the Right Execution Model of OpenCL Applications on FPGAs

Jiantong Jiang, Zeke Wang, Xue Liu, Juan Gómez-Luna, Nan Guan, Qingxu Deng, Wei Zhang, and Onur Mutlu

Proceedings of the 28th International Symposium on Field-Programmable Gate Arrays (FPGA), Seaside, CA, USA, February 2020.

Paper

Slides (pptx) (pdf)

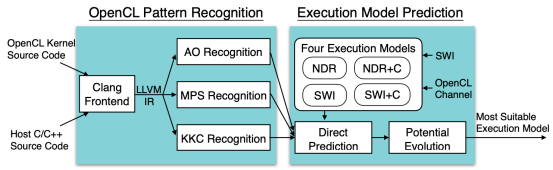


Figure 2: Overview of Boyi.

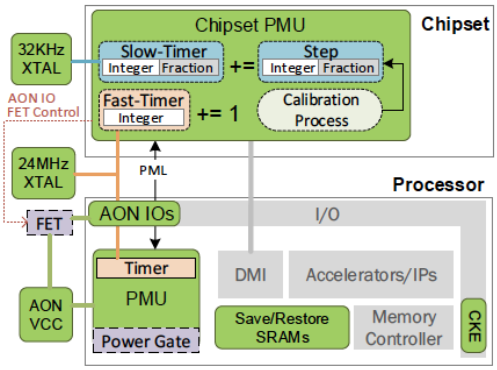
Techniques for Reducing the Connected-Standby Energy Consumption of Mobile Devices

Jawad Haj-Yahya, Yanos Sazeides, Mohammed Alser, Efraim Rotem, Onur Mutlu

Proceedings of the 26th International Symposium on High-Performance Computer Architecture (HPCA), San Diego, CA, USA, February 2020.

Paper

Slides (pptx) (pdf)



All SAFARI Publications

PhD Defense

Congratulations to **Amirali Boroumand** on his successful PhD defense in November!

Practical Mechanisms for Reducing Processor-Memory Data Movement in Modern Workloads

Advisors: Onur Mutlu, Saugata Ghose

Abstract Data movement between the memory system and computation units is one of the most critical challenges in designing high performance and energy-efficient computing system. The high cost of data movement is forcing architects to rethink the fundamental design of computer systems. Recent advances in memory design enable the opportunity for architects to avoid unnecessary data movement by performing Processing-In-Memory (PIM), also known as Near-Data Processing (NDP). While PIM can allow many data-intensive applications to avoid moving data from memory to the CPU, it introduces new challenges for system architects and programmers. Our goal in this thesis is to make PIM effective and practical in conventional computing systems. Toward this end, this thesis presents three major directions: (1) examining the suitability of PIM across key workloads, (2) addressing major system challenges for adopting PIM in computing systems, and (3) re-designing applications aware of PIM capability. In line with these three major directions, we propose a series of practical mechanisms to reduce processor-memory data movement in modern workloads.

Overview of Our Approach

1. Identifying key primitives by analyzing the benefit of PIM for key workloads (*work 1*)
2. Mitigating key system challenges for communication with PIM logic (*work 2*)
3. Redesigning applications aware of PIM using software-hardware co-design (*works 3 & 4*)

SAFARI 6

Thesis Contributions

1. Mitigating data movement bottlenecks in Google Consumer workloads
2. Efficient Cache Coherence Support for Near-Data Accelerators (CoNDA)
3. Efficiently Accelerating Edge ML Inference by Exploiting Layer Heterogeneity (Mensa)
4. Enabling Effective HTAP Databases with Specialized HW/SW Co-Design (Polynesia)

SAFARI 8

PhD thesis talk slides (pdf) (pptx)

Thesis Papers:

CoNDA: Efficient Cache Coherence Support for Near-Data Accelerators, Amirali Boroumand, Saugata Ghose, Minesh Patel, Hasan Hassan, Brandon Lucia, Rachata Ausavarungnirun, Kevin Hsieh, Nastaran Hajinazar, Krishna T. Malladi, Hongzhong Zheng, and Onur Mutlu, ISCA, June 2019. Paper | Lightning Talk Video | Slides (pptx) (pdf)

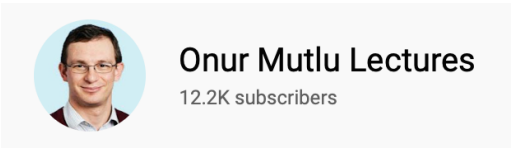
Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks, Amirali Boroumand, Saugata Ghose, Youngsok Kim, Rachata Ausavarungnirun, Eric Shiu, Rahul Thakur, Daehyun Kim, Aki Kuusela, Allan Knies, Parthasarathy Ranganathan, and Onur Mutlu, ASPLOS, March 2018.
Paper | Lightning Talk Video | Full Talk Video (21 minutes) | Slides (pptx) (pdf)

LazyPIM: An Efficient Cache Coherence Mechanism for Processing-in-Memory, Amirali Boroumand, Saugata Ghose, Minesh Patel, Hasan Hassan, Brandon Lucia, Kevin Hsieh, Krishna T. Malladi, Hongzhong Zheng, and Onur Mutlu, IEEE Computer Architecture Letters (CAL), June 2016. Paper

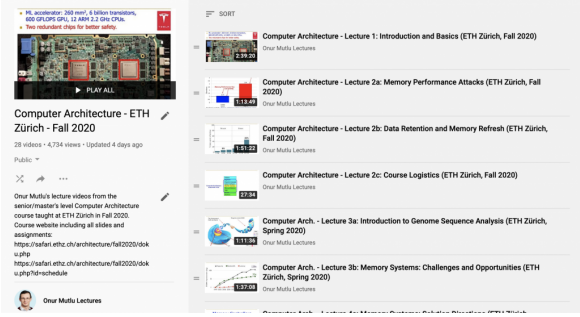
---

Virtual Talks & Courses

We regularly post new videos of talks and lectures on our YouTube Channels. We want everyone to be able to access and benefit from our educational materials. We encourage you to subscribe, comment and give us feedback on our courses and talks! Please also help spread the word so that more people in the world can benefit from the free materials.

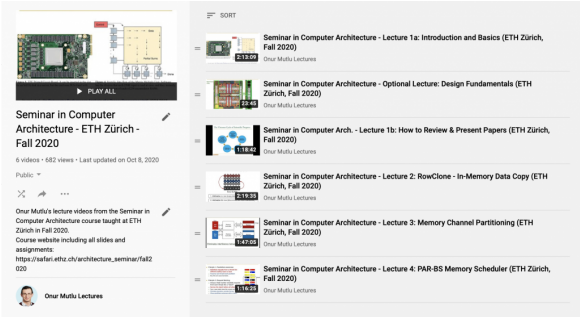


Recent Courses



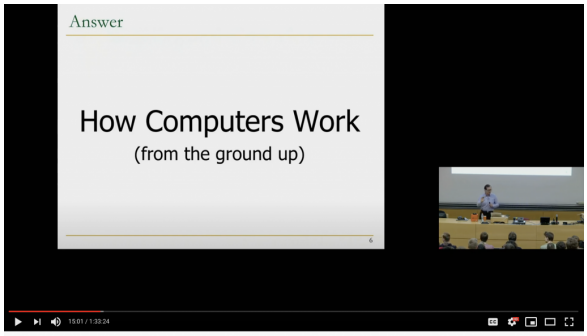
Lecture Videos: Computer Architecture, Fall 2020, ETH

Course Website



Lecture Videos: Seminar in Computer Architecture Fall 2020, ETH

Course Website



Digital Design & Computer Architecture: Lecture 1: Introduction and Basics (ETH Zürich, Spring 2020)

Lecture Videos: Digital Design & Computer Architecture: Spring 2020, ETH

Course Website

Recent Talks - Onur Mutlu

Onur Mutlu - IEDM Tutorial Executive Summary: Memory-Centric Computing Systems, 12 December 2020

Onur Mutlu - Memory-Centric Computing Systems Invited Tutorial at 66th International Electron Devices Meeting (IEDM), Virtual, 12 December 2020.

Slides (pptx) (pdf)

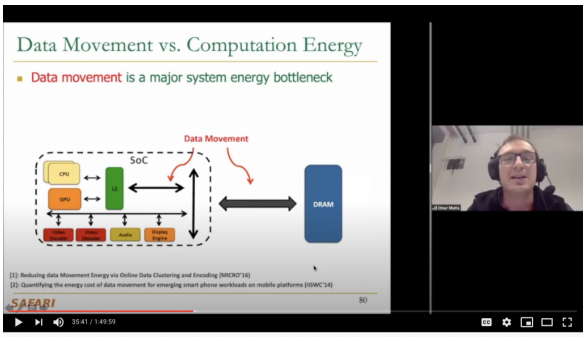
Tutorial Video (1 hour 51 minutes)

Executive Summary Slides (pptx) (pdf)

Executive Summary Video (2 minutes)

Related Keynote Paper from VLSI-DAT 2020

Related Review Paper on Processing in Memory



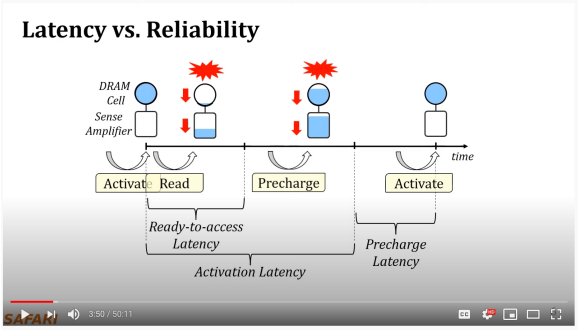
Texas State Univ. CS Seminar - Intelligent Architectures for Intelligent Machines - Onur Mutlu

Onur Mutlu - Intelligent Architectures for Intelligent Machines

Related Keynote Paper from VLSI-DAT 2020

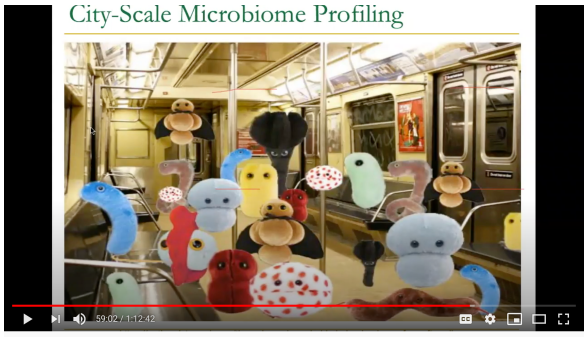
1 - Invited Talk at Texas State University Computer Science Seminar, 13 November

Projects & Seminars Courses Fall 2020:
We have introduced several hands-on project-based courses this semester. They aim to encourage independent experimentation, creative thinking and design, and explorative learning:
Course Website



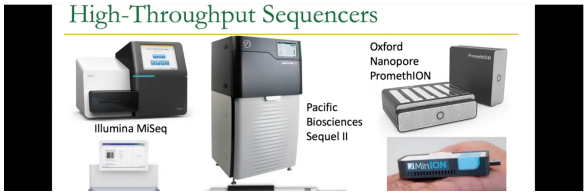
SoftMC Project Course - Introduction to SoftMC (ETH Zürich, Fall 2020)

Hasan Hassan - Understanding and Improving Modern DRAM Performance, Reliability, and Security using SoftMC



Accelerating Genomics Project Course - Lecture 1: Accelerating Genome Analysis (ETH Zürich, Fall'20)

Mohammed Alser - Accelerating Genome Analysis with FPGAs, GPUs, and New Execution Paradigms



2020.
Slides (pptx) (pdf)
Talk Video (1 hour 50 minutes, including Q&A)

2 - Keynote Talk at 13th ACM International Systems and Storage Conference (SYSTOR), 13 October 2020.
Slides (pptx) (pdf)
Talk Video (60 minutes)

3 - Distinguished Lecture at HKUST Engineering and HKSTP Distinguished Speaker Series, 7 October 2020.
Slides (pptx) (pdf)
Talk Video (1 hour 24 minutes)

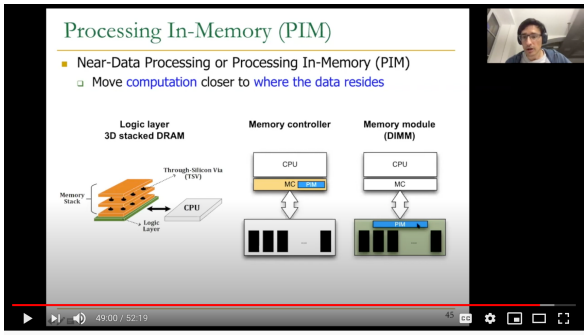
4 - Keynote Talk at National Science Foundation Workshop on Processing-In-Memory Technology (NSF-PIM), 26 October 2020.
Slides (pptx) (pdf)
Talk Video (63 minutes + Q&A)

5 - Plenary Keynote Talk at the 2020 International Symposia on VLSI (VLSI), 11 August 2020.
Slides (pptx) (pdf)
Talk Video (55 minutes)
Keynote Paper

6 - Keynote Talk at Huawei Compute and Storage Technology Workshop, 2 December 2020.
Slides (pptx) (pdf)
Talk Video (56 minutes)
Related Keynote Paper from VLSI-DAT 2020
Related Review Paper on Processing in Memory

Mobile Genomics Project Course - Lecture 1: Genome Sequencing on Mobile Devices (ETH Zürich, Fall20)

Mohammed Alser - Genome Sequencing on Mobile Devices



PiM Project Course - Lecture 1: Exploring the Processing-in-Memory Paradigm (ETH Zürich, Fall 2020)

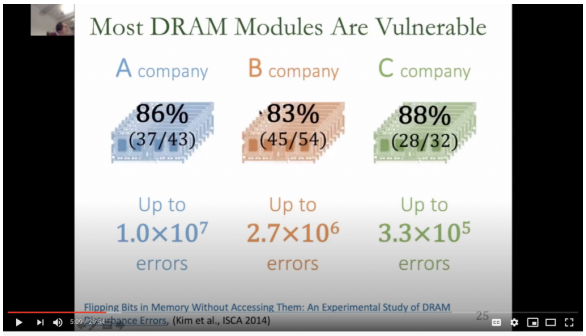
Juan Gomez Luna - Exploring the Processing-in-Memory Paradigm for Future Computing Systems

---

More of Onur Mutlu's Keynote Talks

Onur Mutlu, **Computation in Memory: An Architectural Perspective,** 57th Design Automation Conference Tutorial on New Era of Compute-In-Memory (DAC), Virtual, 20 July 2020. Slides (pptx) (pdf)

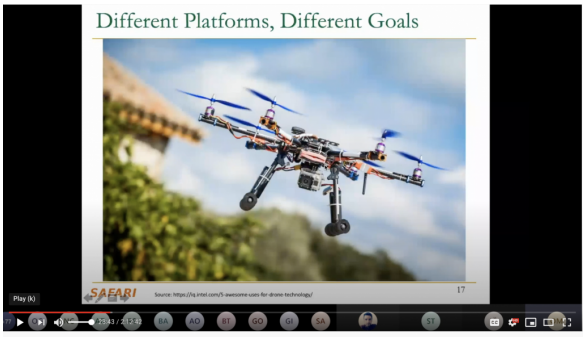
Onur Mutlu, **How to Build an Impactful Research Group,** 57th Design Automation Conference Early Career Workshop (DAC), Virtual, 19 July 2020. Slides (pptx) (pdf)



Onur Mutlu - Revisiting RowHammer (Invited Talk at VLSI-SOC 2020)

Onur Mutlu - Revisiting RowHammer

Invited Talk at 28th IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC), Special Session on AI Frameworks, 9 October 2020.
Slides (pptx) (pdf)
Talk Video (30 minutes)
Related Survey Paper from IEEE TCAD 2019



Onur Mutlu - Invited Lecture at Izmir Institute of Technology - Computer Architecture: Why Today?

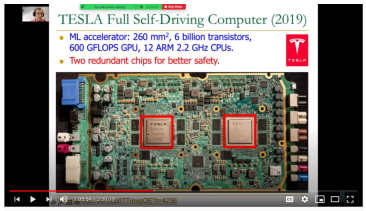
Onur Mutlu - Computer Architecture: Why Is It So Important and Exciting Today?

Invited Lecture at Izmir Institute of Technology (IYTE), 16 October 2020.
Slides (pptx) (pdf)
Talk Video (2 hours 12 minutes)

For a full list of lectures and keynote talks, visit Onur's website

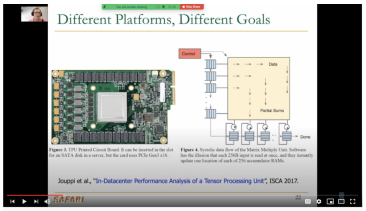
---

Selected Introductory Lectures



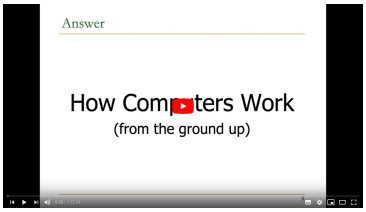
Computer Architecture - Lecture 1: Introduction and Basics (ETH Zürich, Fall 2020)

Onur Mutlu - Computer Architecture - Introduction and Basics, Fall 2020



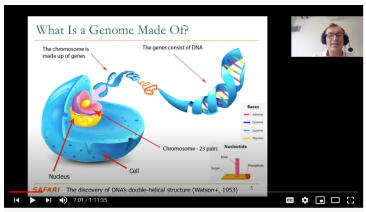
Seminar in Computer Architecture - Lecture 1a: Introduction and Basics (ETH Zürich, Fall 2020)

Onur Mutlu - Seminar in Computer Architecture Lecture 1a: Introduction and Basics, Fall 2020



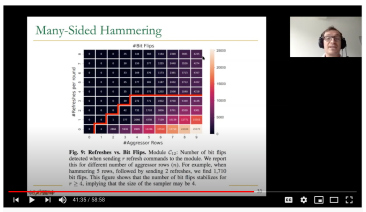
Onur Mutlu - Digital Design & Computer Architecture: Lecture 1: Introduction and Basics, Spring 2020

Selected Lectures: Computer Architecture ETH Fall 2020



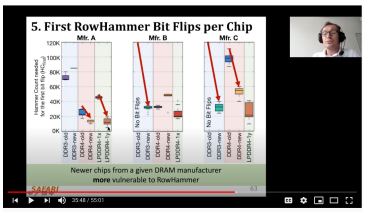
Computer Arch. - Lecture 3a: Introduction to Genome Sequence Analysis (ETH Zürich, Spring 2020)

Onur Mutlu - Lecture 3a: Introduction to Genome Sequence Analysis



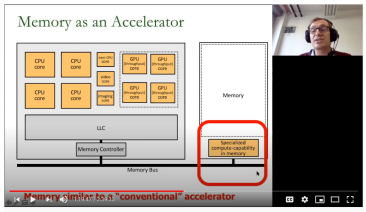
Computer Architecture - Lecture 5a: RowHammer in 2020: TRRespass (ETH Zürich, Fall 2020)

Onur Mutlu - Lecture 5a: RowHammer in 2020: TRRespass



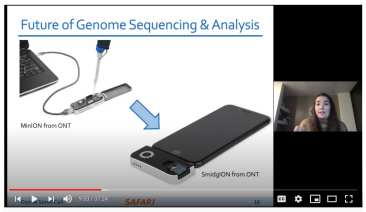
Computer Arch. - Lecture 5b: RowHammer in 2020: Revisiting RowHammer (ETH Zürich, Fall 2020)

Onur Mutlu - Lecture 5b: RowHammer in 2020: Revisiting RowHammer



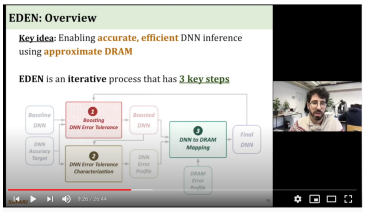
ETH ZÜRICH HAUPTGEBÄUDE
Computer Architecture - Lecture 6: Computation in Memory (ETH Zürich, Fall 2020)

Onur Mutlu - Lecture 6: Computation in Memory



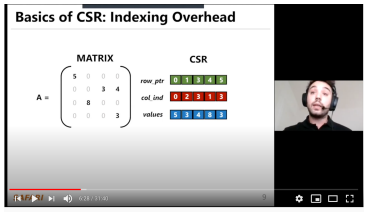
Comp. Architecture - Lecture 9a: GenASM: Approx. String Matching Accelerator (ETH Zürich, Fall 2020)

Damla Senol Cali - Lecture 9a: GenASM: Approx. String Matching Accelerator



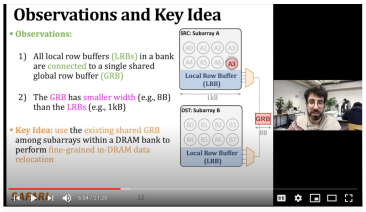
Comp Arch - Lecture 9b: EDEN: Efficient DNN Inference w/ Approximate DRAM (ETH Zürich, Spring 2020)

Lois Orosa - Lecture 9b: EDEN: Efficient DNN Inference w/ Approximate DRAM



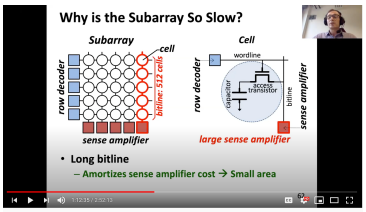
Comp Architecture - Lecture 9c: SMASH: Accelerating Sparse Matrix Operations (ETH Zürich, Fall 2020)

Konstantinos Kanellopoulos - Lecture 9c: SMASH: Accelerating Sparse Matrix Operations



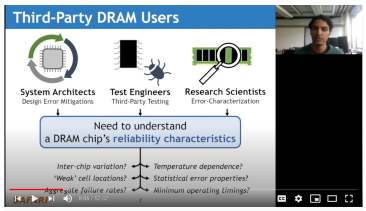
Comp Architecture - Lecture 9d: FIGARO: Fine-Grained In-DRAM Data Relocation (ETH Zürich, Fall 2020)

Lois Orosa - Lecture 9d: FIGARO: Fine-Grained In-DRAM Data Relocation



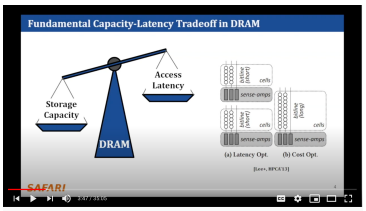
Computer Architecture - Lecture 10: Low-Latency Memory (ETH Zürich, Fall 2020)

Onur Mutlu - Lecture 10: Low-Latency Memory



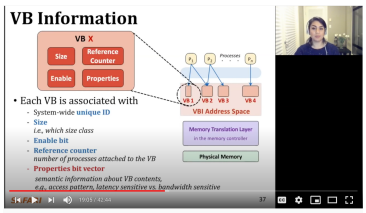
ETH ZÜRICH HAUPTGEBÄUDE
Computer Architecture - Lecture 12a: Bit-Exact ECC Recovery (BEER) (ETH Zürich, Fall 2020)

Minesh Patel - Lecture 12a: Bit-Exact ECC Recovery (BEER)



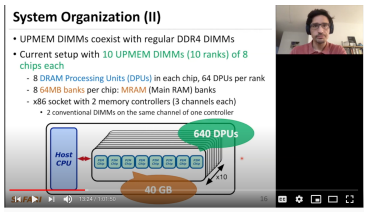
ETH ZÜRICH HAUPTGEBÄUDE
Computer Arch. - Lecture 12b: CLR-DRAM: Capacity-Latency Reconfigurable DRAM (ETH Zürich, Fall 2020)

Haocong Luo - Lecture 12b: CLR-DRAM: Capacity-Latency Reconfigurable DRAM



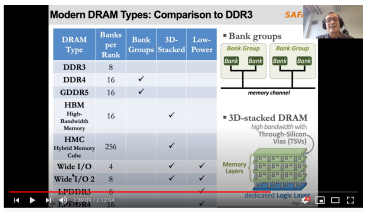
ETH ZÜRICH HAUPTGEBÄUDE
Computer Architecture - Lecture 12c: The Virtual Block Interface (ETH Zürich, Fall 2020)

Nastaran Hajinazar - Lecture 12c: The Virtual Block Interface



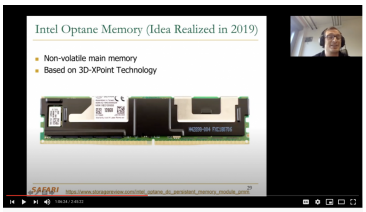
ETH ZÜRICH HAUPTGEBÄUDE
Computer Architecture - Lecture 12d: Real Processing-in-DRAM with UPMEM (ETH Zürich, Fall 2020)

Juan Gomez Luna - Lecture 12d: Real Processing-in-DRAM with UPMEM



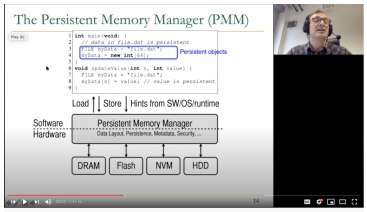
Computer Architecture - Lecture 14: Simulation (with a Focus on Memory) (ETH Zürich, Fall 2020)

Onur Mutlu - Lecture 14: Simulation (with a Focus on Memory)

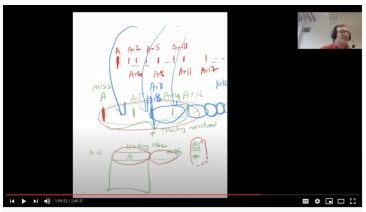


Computer Architecture - Lecture 15: Emerging Memory Technologies (ETH Zürich, Fall 2020)

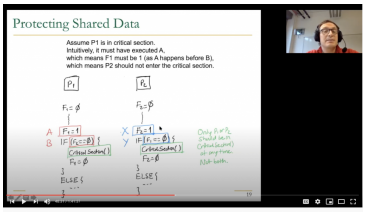
Onur Mutlu - Lecture 15: Emerging Memory Technologies



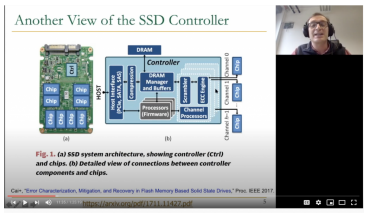
Onur Mutlu - Lecture 16a: Opportunities & Challenges of Emerging Memory Technologies



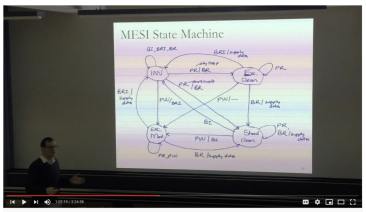
Onur Mutlu - Lecture 18: Prefetching



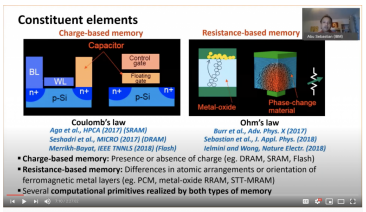
Onur Mutlu - Lecture 20: Memory Ordering (Memory Consistency)



Onur Mutlu - Lecture 11a: Memory Controllers



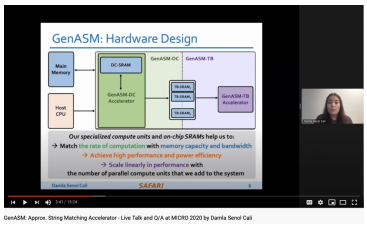
Onur Mutlu - Lecture 21: Cache Coherence



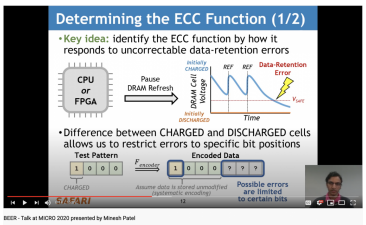
Guest Lecture by Abu Sebastian, IBM Research: In-Memory Computing: Memory Devices & Applications

---

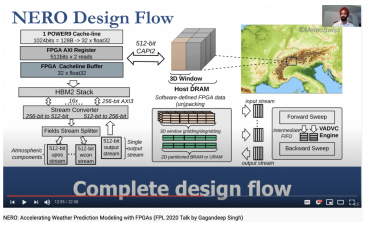
Selected Conference Talks - SAFARI Group Members



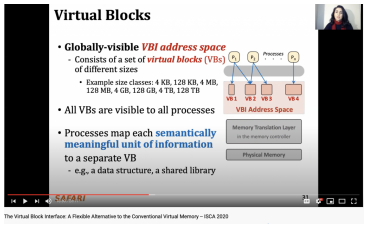
Damla Senol Cali - GenASM: Approx. String Matching Accelerator, Live Talk and Q/A, MICRO 2020 (15.25)
ARM Research Summit Short Talk Video and Q&A (31 minutes)



Minesh Patel - Bit-Exact ECC Recovery (BEER), Live Talk and Q/A, MICRO 2020

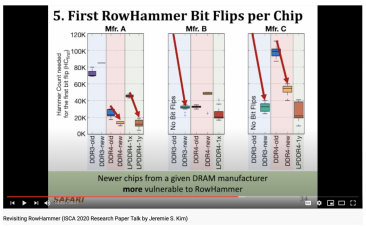


Gagandeep Singh - NERO: Accelerating Weather Prediction Modeling with FPGAs, FPL 2020



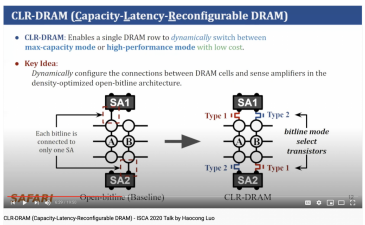
The Virtual Block Interface: A Flexible Alternative to the Conventional Virtual Memory – ISCA 2020

Nastaran Hajinazar - The Virtual Block Interface: A Flexible Alternative to the Conventional Virtual Memory, ISCA 2020



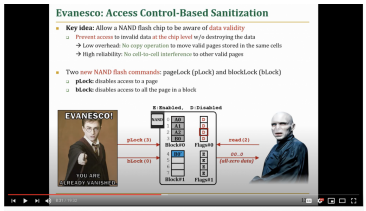
Revisiting RowHammer (ISCA 2020 Research Paper Talk by Jeremie S. Kim)

Jeremie Kim - Revisiting RowHammer, ISCA 2020



CLR-DRAM (Capacity-Latency-Reconfigurable DRAM) - ISCA 2020 Talk by Haocong Luo

Haocong Luo - CLR-DRAM (Capacity-Latency-Reconfigurable DRAM), ISCA 2020

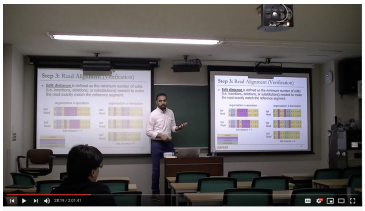


Evanesco: Efficient Data Sanitization in Modern Flash-Based Storage Systems - ASPLOS 2020 Talk

Jisung Park - Evanesco: Efficient Data Sanitization in Modern Flash-Based Storage Systems, ASPLOS 2020

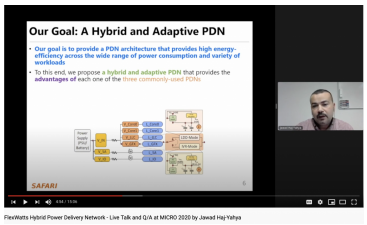
Mohammed Alser - Accelerating Genome Analysis Using New Algorithms and Hardware Designs (Tokyo)

Mohammed Alser - Accelerating Genome Analysis Using New Algorithms and Hardware Designs, Invited Talk, University of Tokyo, December 2019.



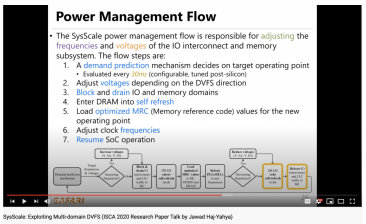
Mohammed Alser - A Roadmap for Fast and Efficient Genome Analysis (The University of Tokyo)

Mohammed Alser - A Roadmap for Fast and Efficient Genome Analysis Invited Talk, University of Tokyo, December 2019.



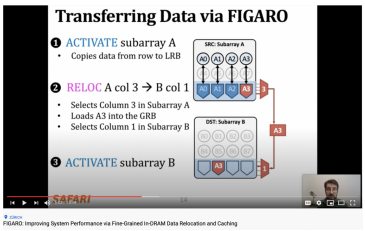
FlexWatts Hybrid Power Delivery Network - Live Talk and Q/A at MICRO 2020 by Jawad Haj-Yahya

Jawad Haj-Yahya - FlexWatts Hybrid Power Delivery Network - Live Talk and Q/A, MICRO 2020



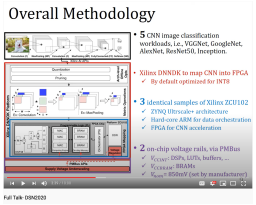
SysScale: Exploiting Multi-domain DVFS (ISCA 2020 Research Paper Talk by Jawad Haj-Yahya)

Jawad Haj-Yahya - SysScale: Exploiting Multi-domain DVFS, ISCA 2020

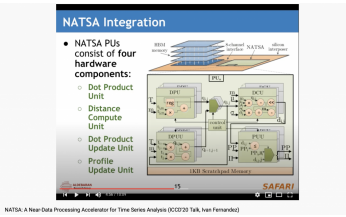


FIGARO: Improving System Performance via Fine-Grained In-DRAM Data Relocation and Caching

Lois Orosa - FIGARO: Improving System Performance via Fine-Grained In-DRAM Data Relocation and Caching, MICRO 2020



Behzad Salami - An Experimental Study of Reduced-Voltage Operation in Modern FPGAs for Neural Network, DSN 2020



Ivan Fernandez - NATSA: A Near-Data Processing Accelerator for Time Series Analysis, ICCD'20

Kevin Hsieh - The Non-IID Data Quagmire of Decentralized Machine Learning

---

Selected Open Source Releases

We have released several research artifacts and tools in 2020 -- all are available on GitHub: **CMU-SAFARI**

BEER

BEER determines an ECC code's parity-check matrix based on the uncorrectable errors it can cause. BEER targets Hamming codes that are used for DRAM on-die ECC but can be extended to apply to other linear block codes (e.g., BCH, Reed-Solomon).

Described in our MICRO 2020 paper:

Patel et al., Bit-Exact ECC Recovery (BEER): Determining DRAM On-Die ECC Functions by Exploiting DRAM Data Retention Characteristics

CLRDRAM

Circuit-level model for the Capacity-Latency Reconfigurable DRAM (CLR-DRAM) architecture. This repository contains the SPICE models of the CLR-DRAM architecture and the baseline architecture described in our ISCA 2020 paper:

Luo et al., CLR-DRAM: A Low-Cost DRAM Architecture Enabling Dynamic Capacity-Latency Trade-Off.

SneakySnake?

The first and the only pre-alignment filtering algorithm that works efficiently and fast on modern CPU, FPGA, and GPU architectures. SneakySnake greatly (by more than two orders of magnitude) expedites sequence alignment calculation for both short (Illumina) and long (ONT and PacBio) reads.

Described in our Bioinformatics paper:

Alser et al., SneakySnake: A Fast and Accurate Universal Genome Pre-Alignment Filter for CPUs, GPUs, and FPGAs.

PDNspot

PDNspot is a framework that models the three commonly-used power delivery network (PDN) architectures in modern client processors in multiple metrics of interest. PDNspot provides a versatile framework that enables multi-dimensional architecture-space exploration of modern processor PDNs.

Described in our MICRO 2020 paper:

Haj-Yahya et al., FlexWatts: A Power- and Workload-Aware Hybrid Power Delivery Network for Energy-Efficient Microprocessors.

AirLift

AirLift is a tool that updates mapped reads from one reference genome to another. Unlike existing tools, It accounts for regions not shared between the two reference genomes and enables remapping across all parts of the references.

Described in our preprint:

Kim et al., AirLift: A Fast and Comprehensive Technique for Translating Alignments between Reference Genomes.

GenASM

GenASM is an approximate string matching (ASM) acceleration framework for genome sequence analysis. GenASM is a fast, efficient, and flexible framework for both short and long reads, which can be used to accelerate multiple steps of the genome sequence analysis pipeline.

Described in our MICRO 2020 paper:

Senol Cali et. al., GenASM: A High-Performance, Low-Power Approximate String Matching Acceleration Framework for Genome Sequence Analysis.

SMASH

Source code of the sparse matrix kernels and utilities used to evaluate the schemes presented in our MICRO 2019 paper: **Kanellopoulos et al., SMASH: Co-designing Software Compression and Hardware-Accelerated Indexing for Efficient Sparse Matrix Operations**

Apollo

Apollo is an assembly polishing algorithm that attempts to correct the errors in an assembly. It can take multiple set of reads in a single run and polish the assemblies of genomes of any size.
Described in our Bioinformatics paper: **Firtina et al., Apollo: A Sequencing-Technology-Independent, Scalable, and Accurate Assembly Polishing Algorithm**

Meet our members:
on research, education and career

Minesh Patel is a PhD student and the lead author of **Bit-Exact ECC Recovery (BEER): Determining DRAM On-Die ECC Functions by Exploiting DRAM Data Retention Characteristics** that won the Best Paper Award at MICRO 2020.

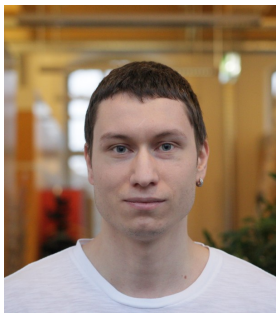
You recently won the Best Paper Award at MICRO, congratulations! Can you tell us about the significance of this paper?

This paper addresses the larger problem that hidden proprietary features implemented by DRAM manufacturers impede end-users from bringing out the best of DRAM technology. We believe BEER takes an important step towards bridging the gap between industry and end-users, starting by focusing on a key example of such features: on-die ECC. Our work discusses how and why on-die ECC limits third-party DRAM consumers and then introduces techniques that the consumers can use to overcome these limitations. We have released our tools in an **open-source project** and look forward to having the community use and extend them.

What were the biggest challenges for you during the writing and review process?

I would say that the biggest challenge we faced when writing this paper was to clearly articulate the problem of on-die ECC limiting third-party users. This includes both (i) describing how and why this limitation arises and (ii) providing concrete examples that the reader can relate to. We spent considerable effort in crafting these arguments such that both we and the reader have a clear understanding of the problem we tackle, our goal in this work, and the final value of our contributions.

Minesh Patel, Jeremie S. Kim, Taha Shahroodi, Hasan Hassan, and Onur Mutlu,
Bit-Exact ECC Recovery (BEER): Determining DRAM On-Die ECC Functions by Exploiting DRAM Data Retention Characteristics, Proceedings of the 53rd International Symposium on Microarchitecture (MICRO), Virtual, October 2020.
Short Talk Slides (pptx) (pdf) | Lecture Slides (pptx) (pdf)
Talk Video (15 minutes) | Short Talk Video (5.5 minutes) | Lightning Talk Video (1.5 minutes)
Lecture Video (52.5 minutes) | BEER Source Code
Best Paper Award MICRO 2020



Hasan Hassan is a PhD student and a co-author of **TRRespass: Exploiting the Many Sides of Target Row Refresh** that won the Best Paper Award at S&P 2020.

You were a co-author on TRRespass, which recently won a Best Paper Award at IEEE S&P as well as a prestigious Pwnie Award for Most Innovative Research. What is the significance of this paper?

Shortly after the discovery of the RowHammer vulnerability of DRAM, DRAM vendors announced RowHammer-free DRAM devices that implement in-DRAM solutions to protect against RowHammer. However, in TRRespass, we find that such solutions, commonly referred to as Target Row Refresh (TRR), do not effectively protect against RowHammer attacks when many rows are hammered at the same time. We show that the RowHammer vulnerability is not only still intact on the current DDR4 devices, but it has also become worse due to technology node scaling.

How was your experience in collaborating with the Systems and Network Security Group at VU Amsterdam on this work?

I am glad that our combined effort with the Systems and Network Security Group at VU Amsterdam won us the Best Paper Award at IEEE S&P. It has been a great experience for me to collaborate with experts in hardware security. I hope there will be more such collaborations that result in impactful research.

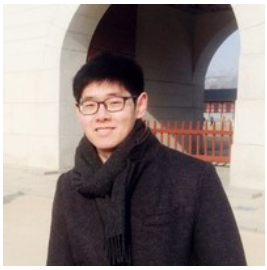
Which tools did you use in this work?

I think SoftMC, our FPGA-based DRAM testing infrastructure, was one of the key enablers of this research. We used SoftMC to interface with DDR4 DRAM chips in a much more flexible way than anyone can do using commodity desktop and mobile systems. Specifically, we used SoftMC to communicate with DRAM chips using low-level DDR4 commands as opposed to using load/store instructions provided by typical instruction set architectures. In a way, SoftMC lets us be the memory controller and provides the flexibility of issuing any DDR4 command at any time, which is not possible with commodity systems.

An earlier version of SoftMC that supports DDR3 devices is open-source and can be accessed **here**. In 2017, we published a **paper** that describes the design of SoftMC in detail.

I am also involved in maintaining **Ramulator**, a cycle-accurate DRAM simulator that we describe in **this paper**, and **Scarab**, which is a cycle-accurate simulator for state-of-the-art multicore CPUs.

Pietro Frigo, Emanuele Vannacci, Hasan Hassan, Victor van der Veen, Onur Mutlu, Cristiano Giuffrida, Herbert Bos, and Kaveh Razavi, TRRespass: Exploiting the Many Sides of Target Row Refresh, Proceedings of the 41st IEEE Symposium on Security and Privacy (S&P), San Francisco, CA, USA, May 2020.
Talk Video (17 minutes) | Lecture Video (59 minutes) | Lecture Slides (pptx) (pdf)
Source Code
Best paper award IEEE S&P
Pwnie Award 2020 for Most Innovative Research



Jeremie Kim is a PhD student and lead author of **Revisiting RowHammer: An Experimental Analysis of Modern Devices and Mitigation Techniques** presented at ISCA 2020.

Can you tell us why it was important to revisit RowHammer now and how you extended the original study?

Since the first rigorous analysis of RowHammer on real DRAM chips in 2014, there have been many advances in DRAM technology including newer DRAM generations and protocols and reduced DRAM technology node sizes. While the original paper showed some correlation of more recently manufactured DRAM chips having a higher vulnerability to RowHammer bit flips, it was important to revisit RowHammer on more modern devices to see how the RowHammer vulnerability has worsened since then.

So are there good mitigation techniques for RowHammer, and have manufacturers been able to really solve the problem?

While our goal was not to determine whether mitigation mechanisms in modern devices prevented RowHammer bit flips, there have been many papers (e.g., TRRespass) that demonstrate weaknesses in existing mechanisms across many generations of DRAM. Until a patent or white paper is published by a DRAM manufacturer with proofs of RowHammer-free operation is released, I will continue to be skeptical, as DRAM manufacturer claims of RowHammer-free chips have been repeatedly refuted in the literature.

What is the overall significance of the paper?

We make two significant observations in this paper. First, the RowHammer vulnerability has significantly worsened in recent years, with the weakest observed chip exhibiting RowHammer bit flips after only 9600 accesses. Second, state-of-the-art academic proposals for RowHammer mitigation either are not scalable or have prohibitively large performance overheads in projected future devices given our observed trends of RowHammer vulnerability. Therefore, it is critical to continue research in more effective solutions to RowHammer.

Jeremie S. Kim, Minesh Patel, A. Giray Yaglikci, Hasan Hassan, Roknoddin Azizi, Lois Orosa, and Onur Mutlu, Revisiting RowHammer: An Experimental Analysis of Modern Devices and Mitigation Techniques, Proceedings of the 47th International Symposium on Computer Architecture (ISCA), Valencia, Spain, June 2020.
Slides (pptx) (pdf) | Lecture Slides (pptx) (pdf) |
Talk Video (20 minutes) | Lightning Talk Video (3 minutes) | Lecture Video (55 minutes)

SAFARI January 2021 Newsletter - Interview with Damla Senol Cali

Damla Senol Cali is a PhD student with SAFARI at **CMU** co-advised by **Onur Mutlu** and **Saugata Ghose**. She presented her interview as the first video contribution to the SAFARI Meet our Members section!

Watch Damla's video interview **here** | Read the transcript **here**

Damla Senol Cali, Gurpreet S. Kalsi, Zulal Bingol, Can Firtina, Lavanya Subramanian, Jeremie S. Kim, Rachata Ausavarungnirun, Mohammed Alser, Juan Gomez-Luna, Amirali Boroumand, Anant Nori, Allison Scibisz, Sreenivas Subramoney, Can Alkan, Saugata Ghose, and Onur Mutlu, GenASM: A High-Performance, Low-Power Approximate String Matching Acceleration Framework for Genome Sequence Analysis, Proceedings of the 53rd International Symposium on Microarchitecture (MICRO), Virtual, October 2020.
Full paper link | Talk Video (18 mins) | Talk Slides (pptx) (pdf) |
Lecture Video (37 mins) | Lecture Slides (pptx) (pdf) | GenASM Source Code |
More information here

Damla Senol Cali, Jeremie Kim, Saugata Ghose, Can Alkan, and Onur Mutlu, Nanopore Sequencing Technology and Tools for Genome Assembly: Computational Analysis of the Current State, Bottlenecks and Future Directions, Briefings in Bioinformatics (BIB), 2018.
Paper link | Paper PDF | AACBB'19 Talk Video | Slides (ppt) (pdf) |
More information here

Christina Giannoula is a PhD student at the **National Technical University of Athens (NTUA)**, and was a visiting researcher with SAFARI in 2019.

How was your experience with SAFARI, and how do you think your PhD has benefitted from your stay in the group?

Regarding my experience with the SAFARI research group, I identify three important characteristics in the SAFARI group culture:

1. Collaboration : I realized that research can vastly benefit from close collaboration. Impactful and innovative research requires a lot of experience, a robust understanding of theoretical principles and deep knowledge surrounding the addressed problem. Considering the above, without the significant help and support from the other group members, who willingly and generously shared their expertise with me, providing me with rigorous feedback and offering their constructive criticism, my efforts would be much less impactful. Their valuable advice, consisting of concrete guidelines and useful methodologies helped me to work efficiently and stay focused in tackling my research obstacles.

2. Learning from group members : The best way to grow at a personal and professional level (both being key components of the PhD learning process) is to share knowledge and experience with others. In the context of academic research, this is particularly important for young students who have little prior research experience. Prof. Mutlu dedicates a lot of time providing feedback to his students, offering significant help by being engaged in frequent technical discussions with them, while at the same time encouraging them to develop their critical thinking skills and training them to evaluate, write and communicate their work efficiently and effectively. This attitude of mentoring and solidarity is commonplace in all SAFARI members (both senior and junior), who spend substantial effort on sharing feedback through discussing and interacting with each other within the group.

3. Motivation : Publishing a great paper is hard and time-consuming. Sometimes, the efforts invested in research may take long to come to fruition. For this reason, it is crucial for students to stay motivated and above all enjoy what they do. In SAFARI, motivation is achieved by: (i) carefully selecting the project students are working on to fit their interests and skills, (ii) interacting with highly-motivated group members (an inspiring work environment significantly elevates the student's productivity), and (iii) being encouraged from your mentors and collaborators to always set high goals and persevere against all obstacles to attain them.

The aforementioned characteristics of the SAFARI group drove my mindset to be highly collaborative, made me be more open to receiving feedback and accepting criticism to improve my work, and be motivated to explore new fields/topics, without exclusively limiting myself to familiar ideas and concepts. As a result, my visit at SAFARI was an invaluable experience in my PhD, as it greatly helped me to acquire a proper research mindset by collaborating with a team of excellence to carry out high-quality research.

Would you recommend such a visit to other PhD students?
When I left Zurich, I had a new research mindset and a lot of motivation to pursue new research directions. During my stay, I greatly expanded my knowledge in the field of computer architecture (specifically in the topics of software/hardware cooperation, architectural support for synchronization and near-data-processing) and acquired new perspectives in approaching research problems. Last but not least, my luggage was full of new experiences, collaborators and friends. Furthermore, I will be grateful to Prof. Mutlu for treating me as if I were his own PhD student and the people at SAFARI group for welcoming me and making me a part of their superb team. Considering the above, I would definitely and highly recommend other PhD students to visit the SAFARI research group.

Mohammed Alser is a Senior Researcher and Lecturer with SAFARI. He was previously a PhD student in SAFARI, co-advised with **Can Alkan**. Mohammed co-teaches two **Projects and Seminars courses** on Genome Sequencing Analysis and Mobile Genomics along with the **Seminar on Computer Architecture**.

You have been busy this past year, and have published quite a few papers. Your recent work, SneakySnake, was recently published in Bioinformatics. This is an important work in improving computations for genome analysis. Can you tell us more about the significance of this work, and what broader impacts you hope for it?

SneakySnake is one of the projects that I enjoyed the most working on. We try in this work to significantly reduce the time spent on finding the similarities and differences between two genomic sequences without sacrificing solution optimality. Finding the similarities and differences between two sequences is a well-known computer science problem, called approximate string matching (ASM), which is solved using computationally expensive algorithms.

SneakySnake quickly finds the sequence pairs that have a large (greater than a user-defined threshold) number of differences and prevents applying computationally expensive algorithms for these sequence pairs, as such sequence pairs are usually not useful for genomic studies. SneakySnake is inspired by the single net routing (SNR) problem in VLSI design that was introduced in 1976. SneakySnake is the first work that proposes to convert the ASM problem into an instance of the SNR problem, which provides several key benefits as we discussed in the paper, and proposes a new efficient algorithm for comparing genomic sequences at scale.

SneakySnake is very beneficial for analyzing both short (e.g., Illumina) and long (e.g., nanopore) sequences as it accelerates the analysis of genomic sequences by up to two orders of magnitude compared to the state-of-the-art algorithms. SneakySnake works efficiently and fast on modern CPU, FPGA, and GPU architectures, which can potentially enable new applications of genome sequencing such as rapid surveillance of disease outbreaks including Ebola and COVID-19, near-patient testing, and bringing precision medicine to remote locations, without the need for large infrastructure.

One of the Bioinformatics journal's reviewers states that: "SneakySnake is a valuable contribution to bioinformatics and it was innovative to reduce the ASM problem to the SNR problem in VLSI CAD".

You also recently published Accelerating Genome Analysis, which reviews the improvements made in hardware accelerators for genome analysis. What are your take away messages from this paper, and what do you see as future priorities in hardware improvements for genome analysis?

Most speedup comes from parallelism enabled by novel architectures and algorithms. We need to develop acceleration solutions that exploit new efficient hardware-aware algorithms, hardware/software co-design, and hardware accelerators to achieve a high degree of parallelism.

Accelerating the entire genome analysis pipeline is important. Accelerating only a single step of genome analysis is not an effective acceleration approach as it limits the overall achieved speedup according to Amdahl's Law.

Genome analysis is currently heavily bottlenecked by data movement. We need to reduce the high amount of data movement that takes place during genome analysis. Moving data (1) between compute units and main memory, (2) between multiple hardware accelerators, and (3) between the sequencing machine and the computer performing the analysis incurs high costs in terms of execution time and energy. These costs are a

significant barrier to enabling efficient analysis that can keep up with sequencing technologies.

The need for flexible hardware architectures. We need to develop flexible hardware architectures that do not conservatively limit the range of supported parameter values at design time. Rapid changes in sequencing technologies (e.g., those that result in high sequencing error rates and longer read lengths) can quickly make specialized hardware with restricted parameter values obsolete.

The need for new genomic data formats. We need to adapt existing genomic data formats for hardware accelerators or develop more efficient file formats to maximize the benefits of hardware accelerators and reduce resource utilization.

Looking into the future, building a genome sequencing machine that provides the entire genome as a single string, rather than its short subsequences, might be possible. However, we believe that the need for hardware acceleration of whole-genome analysis will continue to remain necessary. We also believe performing genome analysis inside the sequencing machine itself can significantly improve efficiency by eliminating sequencer-to-computer data movement.

Your work has many topical applications that are highly relevant to society, including COVID modeling. Can you talk a bit about this, and your future research directions?

As the entire world is largely negatively impacted by the recent COVID-19 outbreak, we believe that everyone can help to end this pandemic based on their skills, expertise, and available resources. At SAFARI research group, we are helping with two main directions.

We are working on developing an accurate and configurable prediction model that evaluates the existing mitigation measures that the government applies in a region and provides suggestions on what strength the future mitigation measures should be. We are quantifying the spread of COVID-19 in Switzerland (as a use-case) by calculating the daily reproduction number of COVID-19, which quantifies how many people are infected on average by an infected person. The reproduction number is directly affected by the mitigation measures that the government applies to a region. We are also considering other important factors such as daylight temperature that significantly affect the spread of COVID-19 as we observed during the year 2020.

We are also working on developing new algorithms and hardware accelerators that perform fast and accurate metagenomic profiling for assessing microbial diversity, identifying potential new species, and investigating microbiomes associated with COVID-19 and other diseases.

Performing genomic tests at scale during a pandemic highlights the dire need for building efficient specialized hardware that is both scalable and portable to enable genome analysis anywhere and anytime. We hope that the progress we make in this direction will also enable new applications that benefit human life and society.

Mohammed Alser, Taha Shahroodi, Juan-Gomez Luna, Can Alkan, and Onur Mutlu, SneakySnake: A Fast and Accurate Universal Genome Pre-Alignment Filter for CPUs, GPUs, and FPGAs, Bioinformatics, December 2020.
Paper PDF | Paper link Bioinformatics | Source Code

Mohammed Alser, Zulal Bingol, Damla Senol Cali, Jeremie Kim, Saugata Ghose, Can Alkan, and Onur Mutlu, Accelerating Genome Analysis: A Primer on an Ongoing Journey, IEEE MICRO, September/October 2020.
Paper | Slides (pptx) (pdf)

Behzad Salami is an affiliated researcher with SAFARI, currently at **Barcelona Supercomputing Center.**

You recently joined the group remotely. Can you tell us what you are working on, and how your experience with collaborating with SAFARI has been?

I joined SAFARI in September 2020 as an Affiliated Researcher. In my former research effort and in my Ph.D. thesis, I improved hardware accelerators' energy-efficiency. More recently, in collaboration with Prof. Onur Mutlu, I proposed an effective Undervolting technique for FPGA-based Convolutional Neural Network (CNN) accelerators, which delivered more than 3X energy-efficiency gain. In my opinion, this work, published in **DSN'2020**, opened a new door for using such undervolting techniques in accelerators especially in power-constrained domains like mobile environments. With this experience as a starting point and jointly with other SAFARI Groups' members, I have started exploring the energy-efficiency, resiliency, and security aspects of novel accelerator architectures like Processing-near & in-Memory systems. I believe that the cutting-edge research in the group, which I am thrilled to be part of, plays an important role in enabling such architectures in future embedded and HPC systems.

And how has it been working remotely with the group?

I started working with the group remotely, right in the middle of the Coronavirus pandemic,

which created a Volatile, Uncertain, Complex, and Ambiguous (VUCA) condition. Under this condition, starting to work with new people, in a new atmosphere, and on new projects could be challenging. However, this phase was quite smooth for me, thanks to the group's friendly atmosphere, the warm welcome of Prof. Onur Mutlu and other group members, and the projects' collaborative nature. In this way, I see online meetings as a key way to stay motivated to explore and build new ideas in collaboration with other members of the group.

Behzad Salami, Erhan Baturay Onural, Ismail Emir Yuksel, Fahrettin Koc, Oguz Ergin, Adrian Cristal Kestelman, Osman S. Unsal, Hamid Sarbazi-Azad, and Onur Mutlu, An Experimental Study of Reduced-Voltage Operation in Modern FPGAs for Neural Network Acceleration, Proceedings of the 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), Valencia, Spain, June 2020.
Paper | Slides (pptx) (pdf) | Lightning Talk Slides (pptx) (pdf) |
Talk Video | Lightning Talk Video

Haiyu Mao joined SAFARI in October as a postdoc.

When you arrived, most of us were working remotely due to Corona, which must have been quite challenging for you, having just arrived from China. Can you tell us about your first impressions in SAFARI, and your impressions now, after a few months with SAFARI?

When I arrived in September, SAFARI happened to move to the new office. It was a little challenging for me to start during the COVID-19 time. However, I felt very happy to receive the help from Prof. Onur Mutlu and SAFARI group members. Onur met me in person. We discussed the projects that I was interested in. He gave me a lot of suggestions on how to start the research and how to do high-impact research. I was very impressed with Onur's serious attitude towards research and teaching, which drove me to work hard, think big, and aim high. Tracy helped me a lot with my setup at ETH and the new office, which enabled me to start the research in a short time. Other group members, Nika, Jisung, Lois, Hasan, Minesh, Geraldo, and so on, helped me a lot with the setup. I am very happy to have these friendly and helpful group members.

Research: Since we have lots of group meetings and brainstorming, which has continued over Zoom, I didn't feel much inconvenience with regard to collaborating during COVID-19. I am getting to know the group members and their research through these meetings. I am glad to work with our brilliant and self-motivated group members. I hope I can have several high-impact research outcomes in such an amazing and collaborative group.

Teaching: I was very nervous about the teaching at the beginning because of my lack of teaching experience. After three months, I already feel much better with the help of other TAs.

Can you tell us what you are working on, and what your future research plans/ directions are?

My research interests mainly include non-volatile memories, processing-in-memory, memory security, and machine learning. I am now working on the security of processing-in-memory and heterogeneous processing-in-memory architectures. Besides this, I am also interested in neural network training in processing-in-memory architectures. I hope I can contribute a lot to the processing-in-memory research and bring these ideas into industry to have wonderful commercial chips.

Jawad Haj-Yahya joined SAFARI in 2019 as a Senior Researcher.

You had several papers published this year, can you tell us more about the significance of each of these papers?

I have published multiple papers on energy efficiency and power management this year.

In our HPCA 2020 paper that introduces **ODRIPS, i.e., optimized-deepest-runtime-idle-power state**, we developed techniques for reducing the connected-standby energy consumption of mobile devices. ODRIPS dynamically: 1) offloads the monitoring of wake-up events to low-power off-chip circuitry, which enables turning off all of the processor's clock sources, 2) offloads all of the processor's input/output functionality off-chip and power gates the corresponding on-chip input/output functions, and 3) transfers the processor's context to a secure memory region inside DRAM, which eliminates the need to store the context using high-leakage on-chip SRAMs, thereby reducing leakage power. Overall, our proposed three-pronged strategy reduces platform average power in connected-standby mode by 22%.

In our ISCA 2020 paper that introduces **SysScale**, we proposed the first work to enable coordinated and highly-efficient dynamic voltage and frequency scaling (DVFS) across all system-on-chip (SoC) domains to increase the energy efficiency of mobile SoCs. SysScale

introduces the ability to redistribute the total power budget across all SoC domains according to the performance demands of each domain. We implement SysScale on the Intel Skylake SoC for mobile devices. On a 2-core Skylake with a 4.5W TDP, SysScale improves the performance of SPEC CPU2006 and 3DMark workloads by up to 16% and 8.9% (9.2% and 7.9% on average), respectively.

In our MICRO 2020 paper that introduces **FlexWatts**, we design a novel adaptive hybrid Power Delivery Network (PDN) that maintains high efficiency and high performance in metrics of interest in client processors across a wide spectrum of power consumption and workloads. To our knowledge, FlexWatts is the first hybrid PDN to use two types of on-chip voltage regulators (IVR and LDO) to simultaneously leverage the advantages of both. During this project, we also develop a versatile framework, PDNspot, that enables multi-dimensional architecture-level exploration of modern processor PDNs. To our knowledge, PDNspot is the first tool that can evaluate the effects of multiple PDN parameters, TDP, and workload characteristics on prominent system metrics such as energy consumption, performance, board area, and bill of materials (BOM). **We have open-sourced PDNspot**. For a 4W thermal design power (TDP) processor, FlexWatts improves the average performance of the SPEC CPU2006 and 3DMark06 workloads by 22% and 25%, respectively. For battery life workloads, FlexWatts reduces the average power consumption of video playback by 11% across all tested TDPs (4W–50W).

What were the biggest challenges in publishing these papers (if any)?

All our three recent works focus on optimizing the energy-efficiency of modern processors from three different angles by optimizing 1) idle energy consumption, 2) dynamic energy consumption, and 3) power delivery networks. Proposing improvements to these highly-optimized commodity processor components is very challenging. It requires a deep understanding of modern processor power management architecture and bottlenecks. For example, in ODRIPS the baseline processors already consume a few milliwatts in the deepest idle state. Reducing this already-low power consumption requires further out-of-the-box thinking. We introduce the idea of dynamically offloading the processor context into DRAM, which saves the entire processor context once the processor is idle.

In SysScale, we observed that the power budget of the processor is inefficiently managed between various SoC domains. The most challenging part was to figure out how to dynamically redistribute the power-budget between domains with minimal overhead. Therefore, building the transition flow and the performance prediction algorithm were the main challenging parts.

In FlexWatts, we observed that the commonly-used power delivery networks in client

processors have multiple pros and cons with respect to multiple metrics. The challenging part is to develop a model that accurately and efficiently quantifies these metrics. To this end, we developed PDNspot. Next, we observed the need for a hybrid PDN that provides the benefits of the commonly-used PDNs. The challenging part of this PDN is to make it low overhead in terms of area and cost while maintaining low transition time between the different modes of operation of the hybrid PDN.

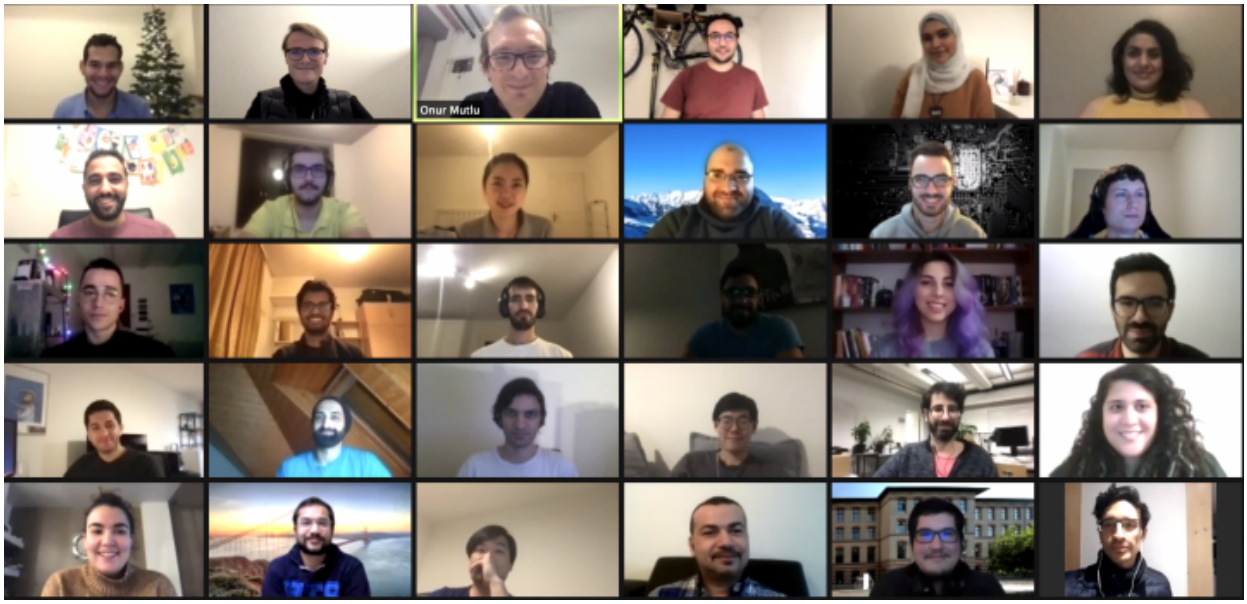
What are you working on next?

I'm working on multiple projects related to energy-efficiency and hardware security. In particular, we observed a fundamental bottleneck in the modern display subsystem of modern processors and we are working to reduce this bottleneck which will save significant energy for essential workloads such as video streaming (which probably almost all of us rely on today). Moreover, we found that there are many security vulnerabilities in the energy management mechanisms of modern processors. We are working on demonstrating security attacks on these mechanisms and propose new architectures to make these processors more secure.

Jawad Haj-Yahya, Yanos Sazeides, Mohammed Alser, Efraim Rotem, Onur Mutlu, Techniques for Reducing the Connected-Standby Energy Consumption of Mobile Devices, Proceedings of the 26th International Symposium on High-Performance Computer Architecture (HPCA), San Diego, CA, USA, February 2020.
Paper | Slides (pptx) (pdf)

Jawad Haj-Yahya, Mohammed Alser, Jeremie Kim, A. Giray Yaglikci, Nandita Vijaykumar, Efraim Rotem, and Onur Mutlu, SysScale: Exploiting Multi-domain Dynamic Voltage and Frequency Scaling for Energy Efficient Mobile Processors, Proceedings of the 47th International Symposium on Computer Architecture (ISCA), Valencia, Spain, June 2020.
Paper | Slides (pptx) (pdf) | Lightning Talk Slides (pptx) (pdf) |
Talk Video | Lightning Talk Video

Jawad Haj-Yahya, Mohammed Alser, Jeremie S. Kim, Lois Orosa, Efraim Rotem, Avi Mendelson, Anupam Chattopadhyay, and Onur Mutlu, FlexWatts: A Power- and Workload-Aware Hybrid Power Delivery Network for Energy-Efficient Microprocessors, Proceedings of the 53rd International Symposium on Microarchitecture (MICRO). Virtual, October 2020.
Paper | Slides (pptx) (pdf) | Short Talk Slides (pptx) (pdf) | Talk Video | Lightning Talk Video
Short Talk Video and Q&A



SAFARI
SAFARI Research Group

Subscribe to Our Newsletter
safari.ethz.ch

Prepared by Tracy Ewen and Onur Mutlu,
with content from many SAFARI members.

Past SAFARI Newsletters: **April 2020**