DAM: Differentiated Access Memory Systems and Applications

Mark Horowitz, Philip Levis, Subhasish Mitra, Thierry Tambe, Caroline Trippel,
Keith Winstein, H.-S. Philip Wong, and Mary Wootters
https://dam.stanford.edu

April 8, 2024

20 years ago, the practical limits of clock speeds forced processors to move to multicore. The bandwidth and capacity limits of low-latency, random-access memory will force a similarly large change, but this time in how computers organize, provision, and access memory.

Differentiated Access Memories (DAM) is a five-year project at Stanford University to explore this future and new ideas on how to embrace it. Our basic hypothesis is that memory will evolve from a uniform address space of random access memory to a heterogeneous collection of different memories, optimized for different uses. These *Differentiated Access Memory* (DAM) systems will require us to revisit and re-examine computing at all levels, from the design of new memory technologies to high-level algorithms, including

- *Hardware and device architectures*: what new memory technologies will we need, how do we compose them in systems, and what caching and consistency mechanisms do they need?
- *Systems software*: how will low-level software present heterogeneous memories to applications and manage them as a resource?
- *Applications and algorithms*: how can applications and algorithms guide the composition of heterogeneous memories, and how will they evolve to best use them?

Technical limitations necessitate DAMs, and it introduces new research challenges all the way from device physics to algorithms. At the device level, we will have to re-examine how we design, manufacture, and integrate *diverse* memories with optimal connectivity to the compute units. This integration will include on-chip, on-package, off-chip, and far memories. At the architectural level, we will have to explore new layouts, access, and caching structures. We will also have to explore dedicated compute units that bind to diverse memories for application execution and system management. Operating systems software will have to manage differentiated memories and expose them to programs with useful abstractions. Applications will have to adapt to allocate and use differentiated memories for their data structures. Finally, we will see algorithmic space complexity (in terms of read, write, and read-write memory) become just as important as time complexity.

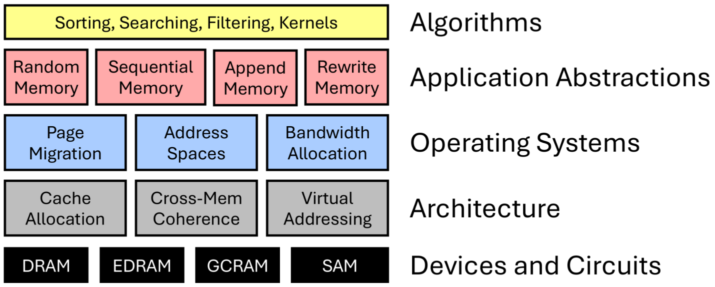


Figure 1: DAM will force changes to multiple components across the technology stack, from devices to applications.

All of these problems are interrelated: caching policies will be informed both by access patterns from applications, access performance from devices, and connectivity with compute units. Similarly, how an operating system exposes memories to applications will be influenced both by algorithmic access patterns as well as how the memories interact architecturally.

Project Goals

Our goal is to, over the five years of the project:

1. Research, define, build, and evaluate new memory architectures that improve energy efficiency, system cost, and application performance.
2. Re-examine how a processor exposes memory to software, tackling the challenges imposed by memory heterogeneity, coherence, consistency, and the variety of application needs.
3. Research, develop, and implement new system architectures and operating systems that expose heterogeneous memories to applications.
4. Design and implement applications and algorithms that leverage the varied capabilities of differentiated access memories to scale better, run faster, and be more efficient.

We will pursue four application domains that emphasize different uses of DAMs: machine learning accelerators, data analytics, high-speed networking, and append-mostly databases. Each application emphasizes a different memory access pattern and set of tradeoffs. Machine learning accelerators, especially those for transformer models, require high bandwidth to move models in and out of SRAM caches. Data analytic applications combine linear scans of large blocks of memory with random-access joins. High-speed networking involves moving page-sized (kilobytes) units of memory to applications based on header fields. Append-mostly databases are large, read-dominated and follow a sequential writing pattern. In each case, we will research, build, and evaluate prototypes to shed light on end-to-end tradeoffs in DAM systems.

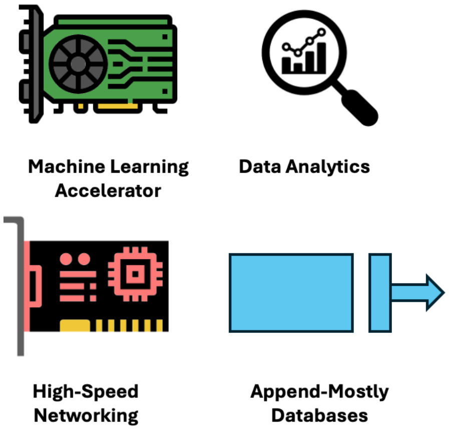


Figure 2: Four example use cases of DAM

Our first year of research focuses on three major topics: embedded DRAMs and long-retention gain cell memory (PIs Wong and Tambe), near memory and heterogeneous memory consistency (PIs Mitra, Trippel, and Horowitz), and OS management of heterogeneous memory (PIs Levis and Winstein).

Affiliates Program

Corporate affiliates are a vital and integral part of the DAM project. Not only do membership fees provide essential funding for the center's activities, but members also engage with project members to explain real world and practical challenges, collaborating with the center on solutions.

Membership fees in DAM's affiliate program are $200,000/year. Joining the affiliate program provides resources for the project to actively engage with affiliates. Affiliates receive invitations to project events.

Faculty

Philip Levis (Faculty Director) is a Professor of Computer Science and Electrical Engineering at Stanford University, where he heads the Stanford Information Networks Group (SING). His research centers on low-level computing systems that interact with the physical world, including low-power computing, operating systems wireless networks, sensor networks, embedded systems, and graphics systems. He has been awarded the Okawa Fellowship, an NSF CAREER award, and a Microsoft New Faculty Fellowship. He's authored over 60 peer-reviewed publications, including three test of time awards and one most influential paper award. His research is the basis for Internet standards on how embedded devices connect to the Internet (RFC6550 and RFC6206).

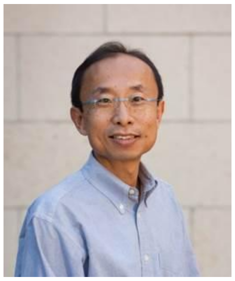
Mark Horowitz is the Fortinet Founders Chair of Electrical Engineering and the Yahoo! Founders Professor in the School of Engineering. He co-founded Rambus, Inc. in 1990 and is a fellow of the IEEE and the ACM and a member of the National Academy of Engineering and the American Academy of Arts and Science. Dr. Horowitz's research interests are quite broad and span using EE and CS analysis methods to problems in molecular biology to creating new design methodologies for analog and digital VLSI circuits.

Subhasish Mitra is William E. Ayer Professor in the Departments of Electrical Engineering and Computer Science at Stanford University. His research ranges across Robust Computing, NanoSystems, Electronic Design Automation (EDA), and Neurosciences. Results from his research group have influenced almost every contemporary electronic system, and have inspired significant government and research initiatives in multiple countries. Prof. Mitra's honors include the Harry H. Goode Memorial Award (by the IEEE Computer Society for outstanding contributions in the information processing field), Newton Technical Impact Award in EDA (test-of-time honor by ACM SIGDA and IEEE CEDA), the University Researcher Award (by the Semiconductor Industry Association and Semiconductor Research Corporation to recognize lifetime research contributions), the Intel Achievement Award (Intel's highest honor), and the US Presidential Early Career Award. He and his students have published over 10 award-winning papers across 5 topic areas (technology, circuits, EDA, test, verification) at major venues. He is an ACM Fellow, an IEEE Fellow, and a Distinguished Alumnus of the Indian Institute of Technology, Kharagpur.

Thierry Tambe is an Assistant Professor of Electrical Engineering at Stanford University. His research interests include hardware and software co-design techniques for domain-specific silicon systems for emerging AI and compute/memory-intensive applications. Prior to debuting his doctoral studies, Thierry was an engineer at Intel where he worked on mixed-signal architectures for high-bandwidth memory and peripheral interfaces on Xeon HPC SoCs. He received a B.S. (2010), M.Eng. (2012) from Texas A&M University, and a PhD (2023) from Harvard University, all in Electrical Engineering. Thierry Tambe is a recipient of the Best Paper Award at the 2020 ACM/IEEE Design Automation Conference, a 2021 NVIDIA Graduate PhD Fellowship, and a 2022 IEEE SSCS Predoctoral Achievement Award.

Caroline Trippel is an Assistant Professor of Computer Science and Electrical Engineering at Stanford University. Her research interests are in the area of computer architecture, with a focus on promoting correctness and security as first-order computer systems design metrics. A central theme of her work is leveraging formal methods techniques to design and verify hardware systems. Caroline's research influenced the design of the RISC-V ISA memory consistency model both via her formal analysis of its draft specification and her subsequent participation in the RISC-V Memory Model Task Group. Additionally, her work produced a novel methodology and tool that synthesized two new variants of the now-famous Meltdown and Spectre attacks. Caroline's research has been recognized with IEEE Top Picks distinctions, an NSF CAREER Award, the 2020 ACM SIGARCH/IEEE CS TCCA Outstanding Dissertation Award, and the 2020 CGS/ProQuest® Distinguished Dissertation Award in Mathematics, Physical Sciences, & Engineering.

Keith Winstein is an Associate Professor of Computer Science and, by courtesy, of Electrical Engineering at Stanford University. His research group creates new kinds of networked systems by rethinking abstractions around communication, compression, and computing. Some of his group's research has found broader use, including the Mosh (mobile shell) tool, the Puffer video-streaming system, the Lepton compression tool, the Mahimahi network emulators, and the gg lambda-computing framework. He has received the SIGCOMM Rising Star Award, the Sloan Research Fellowship, the NSF CAREER Award, the Usenix NSDI Community Award (2020, 2017), the Usenix ATC Best Paper Award, the Applied Networking Research Prize, the SIGCOMM Doctoral Dissertation Award, and a Sprowls award for best doctoral thesis in computer science at MIT. Winstein previously served as a staff reporter at The Wall Street Journal and was the vice president of product management and business development at Ksplice, a startup company now part of Oracle. He did his undergraduate and graduate work at MIT.



H.-S. Philip Wong is the Willard R. and Inez Kerr Bell Professor in the School of Engineering at Stanford University. He joined Stanford University as Professor of Electrical Engineering in 2004. From 1988 to 2004, he was with the IBM T.J. Watson Research Center. From 2018 to 2020, he was on leave from Stanford and was the Vice President of Corporate Research at TSMC, the largest semiconductor foundry in the world, and since 2020 remains the Chief Scientist of TSMC in a consulting, advisory role. He is a Fellow of the IEEE and received the IEEE Andrew S. Grove Award, the IEEE Technical Field Award to honor individuals for outstanding contributions to solid-state devices and technology, as well as the IEEE Electron Devices Society J.J. Ebers Award, the society's highest honor to recognize outstanding technical contributions to the field of electron devices that have made a lasting impact. He is the founding Faculty Co-Director of the Stanford SystemX Alliance – an industrial affiliate program focused on building systems and served as the faculty director of the Stanford Nanofabrication Facility – a shared facility for device fabrication on the Stanford campus that serves academic, industrial, and governmental researchers across the U.S. and around the globe, sponsored in part by the National Science Foundation. He is the Principal Investigator of the Microelectronics Commons California-Pacific-Northwest AI Hardware Hub, a consortium of over 40 companies and academic institutions funded by the CHIPS Act. He is a member of the US Department of Commerce Industrial Advisory Committee on microelectronics.

Mary Wootters is an Associate Professor of Computer Science and Electrical Engineering at Stanford University. She received a PhD in mathematics from the University of Michigan in 2014, and a BA in math and computer science from Swarthmore College in 2008; she was an NSF postdoctoral fellow at Carnegie Mellon University from 2014 to 2016. She works in theoretical computer science, applied math, and information theory; her research interests include error correcting codes and randomized algorithms for dealing with high dimensional data. Her Ph.D. thesis received the Sumner B. Myers Memorial Prize from the UMich Math Department and and the EATCS Distinguished Dissertation award. She is the recipient of an NSF CAREER award, was named a Sloan Research Fellow in 2019 and a Google Research Scholar in 2021; she was awarded the IEEE Information Theory Society James L. Massey award in 2022, and named the IEEE Information Theory Society Goldsmith Lecturer for 2024.