DAM: Differentiated Access Memory Systems and Applications

Chris Gregg, Mark Horowitz, Philip Levis, Subhasish Mitra, Thierry Tambe, Caroline Trippel,
Keith Winstein, H.-S. Philip Wong, and Mary Wootters
https://dam.stanford.edu
Year 2: October 4, 2025

25 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. Memory will no longer be a strict hierarchy of random-access technologies: there will be memories optimized for reads, memories optimized for short-term scratch pads, and memories optimized for large streams of parallel requests. Growing and emerging technologies, such as High Bandwidth Memory (HBM), High Bandwidth Flash (HBF), gain-cell RAM, and CXL memory expanders are some of the first steps towards this computing future.

Differentiated Access Memories (DAM) is a seven-year project at Stanford University, started in 2024, to research the next generation of computer systems that will include and support heterogeneous memories. The project focuses on the fact that memory is evolving 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 computations need, how do we compose them in systems, how will hardware handle coherence and consistency across them, and what new architectural support does software need?
- *Systems software*: how will operating systems and runtime systems 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?

Two such memories we are focusing on are Long-Term RAM (LtRAM) and Short-Term RAM (StRAM). LtRAM is designed for read-mostly data that is written rarely (e.g., minutes or longer). Examples of application data that fits well with LtRAM include executable code, append-mostly database tables, cold memory, large key-value stores, ML inference weights, and transformer key-value caches. LtRAM can be implemented with a variety of device technologies, including RRAM and FeRAM. StRAM is designed for transient data that is accessed that is written then read, or accessed frequently. Examples of data that fits well with StRAM includes caches, packet buffers, and ML activations. We are primarily exploring gain cells as a device technology for StRAM.

Our primary research goal for 2025-2026 is to research, design, and evaluate StRAM and LtRAM devices, their integration into full computing systems, and how software can leverage their benefits.

Projects in Year 2: 2025-2026

This is a brief summary of some of the leading projects DAM students and faculty are working on this year. If you would like further information on any of them, or to discuss them with the PIs and students, please contact Philip Levis <pal@cs.stanford.edu> and Caroline Trippel <trippel@cs.stanford.edu>, the DAM co-directors.

- A methodology for deriving provably optimal multi-chip(let) architectures via joint optimization of inter-chiplet and chip traffic, per-chiplet memory and compute, heterogeneous inter- and intra-chiplet interconnects, and fine-grained spatiotemporal power management.

- BRIDGE (Blended Retention-Indexed Diverse Gain cEll), an innovative on-chip memory solution that offers high density, low energy consumption, and high speed, with tunable retention.

- A complete operational model of the C11 memory model and proving its equivalence with the axiomatic model. Such an operational model can improve the performance of model checkers for parallel C programs, in addition to benefitting other program verification techniques.

- Flow addressing, a new memory addressing architecture for network packet processing. Using flow addressing, software can access many disjoint payloads (e.g., application data spread across packets) as a single contiguous region of memory.

- GainSight, an evaluation framework that primarily examines how well existing workloads map to accelerators with gain cell memories given the cells' limited data retention time.

- GEMMA, an accelerator prototype that targets state-space ML models where model states, weights and activations are stored on StRAM macros on chip.

- SimSpect is a framework for validating hardware Spectre defense prototypes to their security guarantees. It provides a formal specification language for defenses as well as a pipeline that uses the specification to generate and run minimal litmus tests.

- Memory fit, a hardware mechanism for deciding which memory is best suited to an access pattern. Using the information theory of compression, memory fit can accurately place data by embedding state into extended page table entries while processing accesses at memory speeds.

- Building an expansion card that connects LtRAM to a commodity processor, to explore its tradeoffs and operating system management.

MemoryDAX Affiliates Program

Corporate affiliates are a vital and integral part of the DAM project. Membership fees provide essential funding for the center's activities. Members engage with project members to collaborate, explain real world and practical challenges, and to develop new technologies that meet real needs. There are two levels of membership in the MemoryDAX affiliate program. Founding Members support the program at $200,000/year, while Contributing Members support the program at $100,000/year.

Faculty

Philip Levis (Faculty Co-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).

Caroline Trippel (Faculty Co-Director) 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 participation in the RISC-V Memory Model Task Group. 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.

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.

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.