

LA&S Senior Seminar
Literature Review
9 September, 2025

Literature Review #1: A review of magnetic random-access memory technologies and their feasibility for energy-efficient, datacenter-focused applications.

As of 2025, datacenters consume 1.5% of global electricity, a figure projected to double or triple by 2030, driven by interest in generative AI. This exponential growth stems from the computational demands of large language models and generative AI systems, intensified by competitive pressures among technology companies to achieve AI supremacy. And, while datacenters are built all over the world as long as there is cheap land and good connectivity, this is an especially pressing issue in the United States, where these trends, federal policies prioritizing domestic semiconductor manufacturing and AI competitiveness have incentivized rapid datacenter expansion, often prioritizing speed-to-market over green solutions, placing strain on aging utilities infrastructure and entrenching the cheapest available energy production sources. Given these entrenched economic and policy drivers, addressing datacenter energy consumption requires fundamental advances in hardware efficiency rather than relying solely on grid modernization or building optimization.

Heat dissipation represents the primary constraint on datacenter efficiency and performance scaling. While thermal management poses minor inconveniences for consumer devices, it becomes a critical bottleneck at datacenter scale, where processors operate at maximum load continuously. To put the magnitude of the problem into perspective, datacenters of even a modest size can be 90,000 times computationally powerful than the latest smartphone, so every ounce of lost efficiency matters. Given the ecological concerns shared by many scientists, if more datacenters are to be built, then they must be made cleaner. This thermal wall

coincides with the slowing of Moore's Law scaling, creating what researchers term the 'Beyond-CMOS imperative'—the urgent need for alternative device technologies that can deliver performance improvements without proportional increases in power consumption (Beyond CMOS).

A historically viable alternative to charge-based devices is held by spintronic devices. Spintronic devices, which leverage electron spin as an additional degree of freedom in electronic circuits, have long been a focus of innovation, particularly in data storage technologies such as hard drives. The discovery of giant magnetoresistance in 1988, for which the Nobel Prize would be awarded, would be hugely impactful for bringing large storage sizes to commercial market, kick starting the field of spintronics and bringing gigabyte, and later terabyte sized hard drives to the common person. Furthermore, the discovery of spin-transfer torque, spin-orbit torque, and giant tunneling magnetoresistance would form the basis of magnetic random-access memory (MRAM) technologies, which have been integrated in cache memories for over a decade (Dieny et al.). In the pressing demands for high-speed, high-density, and high-efficiency set forth by datacenters, physicists look towards the integration of MRAM as a solution.

There are multiple application spaces which MRAM has the potential to shine in. The first is on-chip applications, where small, extremely fast MRAM devices would be implemented. This is an extremely important application, as improvements in embedded memory directly improve computing efficiency in two ways. The first is as a replacement for flash memory, responsible primarily for reading and writing operations related to program execution. Second, if one looks into the CPU of their computer, they would also find a few kinds of on-chip caches, designed to be the very fastest memory the computer uses. These are designated L1 to L3, with L1 being the fastest and L3 being the slowest. Unlike other embedded or primary memories, the

CPU cache must remain in use at all times, and pulling data from these can be expensive from a power usage standpoint. L1, at the smallest size, is also the most power-efficient, as L2 and L3 at their larger sizes consume anywhere from 2-5x more power, respectively (Lam).

Indeed, MRAM is already showing viability as a replacement for flash, in the form of spin-transfer torque (STT)-MRAM (Nyugen et al.). STT-MRAM devices are built on magnetic tunnel junctions (MTJs), composed of two magnetic layers separated by a thin gap. One layer works to store the information as a one or a zero, while the other acts as a reference for which the state can be read. In terms of performance, STT-MRAM is three times smaller than conventional flash memory, and is at best able to switch in the few-nanosecond regime (Yang et al.) (Hellenbrand et al.). Recent improvements to switching efficiency and signal-to-noise-ratio have enabled fabs to replace conventional embedded flash memory and static random-access memory with STT-MRAM in high-end electronics. Recent reviews suggest viability in cache applications as well, as new generations of the technology push read-write speeds even lower. Currently, only slow L3 cache is viable at the commercial level due in part to speed demands, but largely due to material endurance.

At the cache level, researchers instead expect spin-orbit torque (SOT)-MRAM to dominate due to the potential to circumvent the aforementioned speed and endurance issues with STT-MRAM (Nyugen et al.). Instead of a MTJ, SOT-MRAM utilizes an in-plane charge current, which injects a spin-polarized current into the ferromagnetic layer via the spin-Hall and/or Rashba-Edelstein effect (Manchon et al.). The construction of SOT-MRAM devices demands the separation of a read and write lane, which effectively reduces device wear by halving the amount of total operations done. Currently, the barrier to commercial implementation lies in material optimization. (something about SHE materials, topological materials)

While cache and flash memory devices pose an important area for improvement, it is actually slower, large scale memory which is responsible for the largest power consumption in terms of read/write operations, albeit they are less frequent. For this purpose, voltage controlled magnetic anisotropy (VCMA)-MRAM is an extremely promising emerging technology. This technique is fundamentally differentiated from the current controlled SOT-MRAM and STT-MRAM. By eliminating this factor, Ohmic energy dissipation, often the predominant inefficiency, may be significantly reduced. VCMA-MRAM devices are therefore highly energy efficient, and have been demonstrated to operate with switching energy of 40 fJ at sub nanosecond speeds (Khalili Amiri et al.).

While memory serves as a critical component to modern day computing, there is an inherent disadvantage in specialized use-cases, such as the training of AI models, in that the physical separation of memory and compute units reduces performance and power-efficiency. Moreover, features such as “non-volatility, stochasticity, and oscillations in MRAM, provide new feasibility for novel computing to solve combinatorial optimization problems, which are notoriously difficult for conventional computers” (Nguyen et al.). Given the emphasis on AI-focused datacenters in the last decade, it is highly desireable, then, to approach the issue of power usage and cooling using these technologies (Shehabi et al.). This unique architectural suitability towards AI workloads positions MRAM technology as a cornerstone to reducing power inefficiency and furthering technological progress.

The commercial success of STT-MRAM in flash memory and solid-state drive applications represents only the beginning of spintronics' transformative potential for datacenter efficiency. As fabrication techniques mature and materials science advances accelerate, next-generation MRAM technologies are poised to revolutionize computing architectures from cache

memory to novel processing paradigms. Recent breakthroughs in achieving room-temperature perpendicular magnetic anisotropy, critical to dense memory architectures, demonstrates that the fundamental materials challenges are yielding to sustained research efforts.

More importantly, the convergence of spintronic devices with AI workloads creates unprecedented opportunities for hardware-software co-optimization. Unlike conventional memory technologies that merely store and retrieve data, spintronic systems offer inherent support for probabilistic computing and combinatorial optimization, precisely the computational primitives driving datacenter growth. As the datacenter industry faces unsustainable energy consumption as a barrier to continued AI advancement, spintronic technologies provide a pathway to transcend traditional performance-power tradeoffs. The question is certainly not whether MRAM technology will transform computing, but how rapidly they can be deployed to alleviate the energy demands of modern computing architecture.

Works cited

- Beyond CMOS: The Future of Semiconductors - IEEE IRDSTM.
<https://irds.ieee.org/home/what-is-beyond-cmos>. Accessed 7 Sep. 2025.
- Bi, Xiuyuan, et al. "STT-RAM Cell Design Considering CMOS and MTJ Temperature Dependence." *IEEE Transactions on Magnetics*, vol. 48, no. 11, Nov. 2012, pp. 3821–24. IEEE Xplore, <https://doi.org/10.1109/TMAG.2012.2200469>.
- Cai, Kaiming, et al. "Spin-Based Magnetic Random-Access Memory for High-Performance Computing." *National Science Review*, vol. 11, no. 3, Oct. 2023, p. nwad272. PubMed Central, <https://doi.org/10.1093/nsr/nwad272>.
- Diény, B., et al. "Opportunities and Challenges for Spintronics in the Microelectronics Industry." *Nature Electronics*, vol. 3, no. 8, Aug. 2020, pp. 446–59. www.nature.com, <https://doi.org/10.1038/s41928-020-0461-5>.
- Hellenbrand, Markus, et al. "Progress of Emerging Non-Volatile Memory Technologies in Industry." *MRS Communications*, vol. 14, no. 6, Dec. 2024, pp. 1099–112. Springer Link, <https://doi.org/10.1557/s43579-024-00660-2>.
- Hu, G., et al. "Key Parameters Affecting STT-MRAM Switching Efficiency and Improved Device Performance of 400°C-Compatible p-MTJs." 2017 IEEE International Electron Devices Meeting (IEDM), 2017, p. 38.3.1-38.3.4. IEEE Xplore, <https://doi.org/10.1109/IEDM.2017.8268515>.
- Joshi, Vinod Kumar, et al. "From MTJ Device to Hybrid CMOS/MTJ Circuits: A Review." *IEEE Access*, vol. 8, 2020, pp. 194105–46. IEEE Xplore, <https://doi.org/10.1109/ACCESS.2020.3033023>.
- Khalili Amiri, Pedram, et al. "Electric-Field-Controlled Magnetoelectric RAM: Progress, Challenges, and Scaling." *IEEE Transactions on Magnetics*, vol. 51, no. 11, Nov. 2015, pp. 1–7. IEEE Xplore, <https://doi.org/10.1109/TMAG.2015.2443124>.
- Kim, Woojin, et al. "VCMA-MTJ: Towards Ghz Operation Low Power MRAM." 2023 IEEE International Magnetic Conference - Short Papers (INTERMAG Short Papers), 2023, pp. 1–2. IEEE Xplore, <https://doi.org/10.1109/INTERMAGShortPapers58606.2023.10228494>.
- Lam, Chester. Alder Lake's Caching and Power Efficiency. 24 Aug. 2025, <https://chipsandcheese.com/p/alder-lakes-caching-and-power-efficiency>.
- Li, Yueling, et al. "Work-in-Progress: Toward Energy-Efficient Near STT-MRAM Processing Architecture for Neural Networks." 2022 International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), 2022, pp. 13–14. IEEE Xplore, <https://doi.org/10.1109/CODES-ISSS55005.2022.00013>.
- Manchon, A., et al. "Current-Induced Spin-Orbit Torques in Ferromagnetic and Antiferromagnetic Systems." *Reviews of Modern Physics*, vol. 91, no. 3, Sep. 2019, p. 035004. DOI.org (Crossref), <https://doi.org/10.1103/RevModPhys.91.035004>.
- Nguyen, V. D., et al. "Recent Progress in Spin-Orbit Torque Magnetic Random-Access Memory." *Npj Spintronics*, vol. 2, no. 1, Oct. 2024, p. 48. www.nature.com, <https://doi.org/10.1038/s44306-024-00044-1>.
- Shehabi, Arman, et al. United States Data Center Energy Usage Report. LBNL--1005775, 1372902, 1 Jun. 2016, p. LBNL--1005775, 1372902. DOI.org (Crossref), <https://doi.org/10.2172/1372902>.

- Song, Cheng, et al. "Recent Progress in Voltage Control of Magnetism: Materials, Mechanisms, and Performance." *Progress in Materials Science*, vol. 87, Jun. 2017, pp. 33–82. ScienceDirect, <https://doi.org/10.1016/j.pmatsci.2017.02.002>.
- Yang, Hyunsoo, et al. "Two-Dimensional Materials Prospects for Non-Volatile Spintronic Memories." *Nature*, vol. 606, no. 7915, Jun. 2022, pp. 663–73. www.nature.com, <https://doi.org/10.1038/s41586-022-04768-0>.
- Yusuf, Alaba, et al. "Domain-Specific STT-MRAM-Based In-Memory Computing: A Survey." *IEEE Access*, vol. 12, 2024, pp. 28036–56. IEEE Xplore, <https://doi.org/10.1109/ACCESS.2024.3365632>.