In-Memory Computing: Global Energy Consumption, Carbon Footprint, Technology, and Products Status Quo

Nima TaheriNejad
EclectX Team
TU Wien, Vienna, Austria
Heidelberg University, Heidelberg, Germany
E-mail: nima.taherinejad@ziti.uni-heidelberg.de

***Abstract*—In this paper, we highlight and quantify the importance and potential role of In-Memory Computation (IMC) and memory technologies in the future of humans' global footprint. To this end, we calculate the estimated energy consumption and carbon emission associated with the data movement inside computing systems and put them in perspective using tangible examples. Next, we review various memory technologies as well as their advantages and disadvantages (especially regarding their energy consumption), for usage in computing systems as memory and computing elements. We calculate what their impact is and what would be the potential savings of migrating towards emerging memory technologies. We discuss some of the challenges these emerging memory technologies face, before presenting the highlights of the IMC products on or near the market. This paper aims at providing an insight on the impact of IMC and memory technology on the society at large and clarify the importance of working on IMC and emerging memory technologies to lower the power consumption and overall footprint of computing systems. The status of IMC products show that while moving in the right direction, there is a substantial body of work to be done. We hope this will help engineers to better grasp the extent of the impact they can produce and motivate them further in the pursuit of better computing systems.**

***Index Terms*—In-Memory Computing, Memory Technology, Global Impact, Energy Consumption, Carbon Emission, Industrial Products**

I. Introduction

Reducing energy consumption is a crucial goal in the current circumstances of rapid growth in the computational load and in the number of computing devices. In 2021, two third of the 1.7×10^{15}Wh of energy consumed by the information and communications technology world-wide was due to computations [1]. With the exponential growth in the number of devices, such as Internet of Things (IoT) devices, this number is expected to grow even further. Mobile systems such as smartphones, embedded systems, wearable electronics, and IoT devices, which are often powered by batteries or rely on energy harvesting, require optimal utilization of the available energy [2], [3]. Although energy constraints may be more sever for IoT and embedded devices and present a challenge in providing the necessary energy, the energy consumption of plugged in computers should not be overlooked. Estimated annual energy consumption of data centers [1] has surpassed 200 TWh already in the last decade. With this amount of energy, you can provide the electricity for half of Iranians every year. From the carbon footprint point of view, this puts ICT on par with aviation industry [1]. More importantly, with the current trajectory of the growth in the energy consumption of computing systems, by 2040, their energy consumption will surpass the world's current energy production capacity [4].

However, what do these numbers mean for the key players involved in the design and development of computing systems? In this paper, we will look inside the computers and break these numbers into more computer architecture related details and thus provide helpful insight, direction, and motivation for those who are involved in shaping the future of computing systems, especially computer and electronics engineers.

The rest of this paper is organized as follows. In Section II, we first calculate the energy consumption associated with computing in general and data movement in particular. To better understand these large numbers, in Section III, we put them in perspective by calculating their equivalents in the energy production sector and carbon absorption of trees. Section IV is dedicated to a brief overview of memory technologies and sets the stage for our calculations in Section V, which quantifies the impacts of memory technology on the global energy consumption and carbon dioxide production using a conceptual example. We discuss the limitations of our analyses and the extended footprint of the computing systems and technology in Section VI. Section VII presents some of the challenges that current memory technologies face, in particular regarding their usage for In-Memory Computation (IMC). It also gives an overview of key related products on or near the market, to put in perspective the open space for research and development in the field of IMC and memory technology. Lastly, we conclude the paper in Section VIII.

TABLE I: The energy and carbon footprint of Information & Communications Technology (ICT), computing systems, and data movement inside computing systems.

Year	Item	Energy Consum- -ption [PWh]	Solar Panels [trillion]	# of Nuclear Power Plants	CO_2 [Mt]	Conifer Trees [trillion]
2021	ICT	3	10.4	622	1798	1050
	Computing	1.7	5.90	352	1019	595
	Data Movement	1.07	3.72	222	642	375
2026	ICT	4.9	17.0	1016	2937	1715
	Computing	2.4	8.33	497	1438	840
	Data Movement	1.51	5.25	313	906	529
2030	ICT	8	27.7	1658	4795	2800
	Computing	3.7	12.8	767	2217	1295
	Data Movement	2.33	8.09	483	1397	815

II. ENERGY CONSUMPTION

A key factor in the energy budget of modern computing devices is the data movement between memory and processor, which constitutes a significant portion of the energy consumption of computing systems, e.g., in Google that is about 63% [5]. Taking that as an indicative ratio, we calculate the energy that IMC can save;

$$E_{IMC} = 0.63 \times E_{COM} \tag{1}$$

where E_{COM} is the global energy consumption due to computing and E_{ICT} is the overall energy consumption of ICT. Since in 2021 $E_{COM} = 1.7PWh$ [1], we calculate the global energy consumed in 2021 due to data movement to be 1.26PWh ($E_{IMC} = 2PWh \times 0.63 = 1.07PWh$).

Based on [1], in 2026, we can expect ICT to consume 4.9PWh of energy, 2.4PWh of which is due to computing. If we plug that number into Equation (1), we can estimate that using the traditional computer architectures, 1.51PWh of that energy is spent on data movement inside computers. The projections for 2030 predict that ICT will be responsible for 8PWh of energy consumption, 3.7PWh of which would be due to computing [1]. Using Equation (1), we can estimate the global energy consumption due to data movement, which is 2.33PWh in 2030. It may be hard to grasp these numbers off the bat. Therefore, we put them into context using more tangible examples in the next section.

III. WHAT DOES IT MEAN?

To better understand the meaning of above numbers, we divide the total sum by the produced energy of photovoltaic panels [6] or nuclear power plants [7], to calculate that in 2021 an equivalent of what 3.7×10^9 photovoltaic panels or 222 nuclear power plants were necessary to produce the energy consumed due to data movement. Using the carbon footprint ratio reported in [8], this data movement produces 642Mt of CO_2. Based on the absorption ratios proposed in [9], to absorb the 642Mt CO_2 produced by the data movement, we would need 375 trillion conifer trees.

Using similar calculations and the energy forecast reported in [1], the potential energy consumption of computing in 2026 will be equal to what 8.3 trillion solar panels, or 497 nuclear power plants can produce. The carbon footprint of the 2.4PWh energy consumed due to computing systems will be around 1438Mt, which would require 840 trillion conifer trees to compensate it. From these, 1.5PWh is attributed to the data movement, which is equal to the energy produced by 5.2 trillion solar panels or 313 nuclear power plants. This data movement produces 906Mt of carbon dioxide, whichwould need 529 trillion conifer trees to absorb.

By aggressively integrating IMC into the ICT, by 2030 will be able to save a major portion of the 2.33PWh, which will be otherwise used on data movement. This energy consumption is equal to the energy produced by 483 nuclear power plants. If we do not meaningfully migrate towards IMC, to absorb the 1397Mt CO_2 that will be produced in 2030 by the data movement inside computers, 815 trillion conifer trees will be necessary.

We have inserted the rest of calculated equivalencies and an overall summary of our calculations in Table I. These numbers are strong motivators for computer architects to consider a migration towards IMC.

IV. MEMORY TECHNOLOGY

Memories and memory technology are key contributors to the performance figure of computing systems too [14]–[16]. Admittedly, their role is more pronounced for IMC solutions [17]. Volatile Random Access Memorys (RAMs) such as Dynamic Random Access Memory (DRAM) and Static Random Access Memory (SRAM) are established and mature technologies but provide a less advantageous solution regarding the energy consumption, since they consume power for data retention. In terms of the energy required to access each bit, DRAMs provide a very competitive solution (2 pJ/bit [11]), however, they are difficult and expensive to integrate on chip and when implemented off-chip, they incur data movement costs that takes the edge of this advantage off.

Among non-volatile memories, only Flash memories are equally mature. However, they are three and four orders of magnitude slower than DRAM and SRAM, respectively [12], [13]. Whereas they do not require energy for data retention, they consume a few orders of magnitude more energy than other memory technologies (see Table II). This motivates IMC using emerging memory technologies, such as Phase Change Memory (PCM) [18], Spin Transfer Torque (STT) and

TABLE II: An overview of the characteristics of some of the key memory technologies. Numbers are estimations extracted from [10]–[13].

Characteristic	SRAM	DRAM	NAND Flash	PCM	ReRAM	STT MRAM
Cell size (f^2)	50-150	6-12	4-6	4-12	4-10	6-50
Non-Volatile	No	No	Yes	Yes	Yes	Yes
Read Time	1-8ns	30-50ns	25000-50000ns	30-50ns	1-20ns	1-20ns
Write Time	1-8ns	30-50ns	200000-50000ns	500ns	0.3-30ns	10-20ns
Access Energy / bit	460pJ	2pJ	10000pJ	20-100pJ	1pJ	0.02-10pJ
Endurance	10^{15}-10^{16}	10^{15}-10^{16}	10^5	10^6-10^8	10^6-10^{12}	10^{15}
Byte Operation	Yes	Yes	No	Yes	Yes	Yes
Analog / Multilevel	No	No	Yes	Yes	Yes	No

Magnetoresistive Random Access Memory (MRAM) [19], and Resistive Random Access Memory (ReRAM) [20], which are referred to under the umbrella term of memristors.

In contrast to DRAM, SRAM, and Flash memories, which are charge-based, memristors are resistive-based memories. That is, their state is represented by their resistance. They are non-volatile, fast, Complementary Metal-Oxide Semiconductor (CMOS)-compatible, and have a small form factor [12]. Table II provides an overview of a selection of memory technologies, established and emerging, and their key characteristics. Among memristors, MRAMs seem to be able to provide the most energy efficient option, although not every implementation of them would be more efficient that ReRAMs. ReRAMs are the fastest, the most compact, and the most used memristive technology in practice. Compared to MRAMs, ReRAMs have one key advantage that may play a significant role in their wider adaption, i.e., their ability to store and represent analog values and consequently storing more than one bit of data per cell. The key disadvantage of current ReRAMs technology is their endurance than be orders of magnitude lower than MRAM. It is hard to predict which memristive technology will be the dominant one or whether there will be a winner-takes-it-all scenario or they will all co-exist, which might be a likelier scenario. Nevertheless, it is clear that they are coming to take over, at least a portion of memory and IMC, and their energy efficiency will be a key factor. We note that there are other promising memory technologies such as ferroelectric based memories.

V. Memory Technology Impact

Calculating the effect of memory technology on the overall energy consumption of computing systems is very complex. This complexity stems from the large variety of memory characteristics and technologies used at different hierarchical levels and the high dependency of their energy consumption on the type of application and user behaviors. However, in this paper, we scratch the surface by estimating the effects of replacing SRAM memories used for cache with ReRAMs or MRAMs, consuming approximately a pJ for the access energy. We note that ReRAM technologies with an endurance of 10^{12} could be considered for such a role but those with 10^6 cycles are not likely to take that role. Although they may be used in other locations of the memory hierarchy, or emerging applications such as the weights of Neural Networks (NNs), where the memory is mostly read out and rarely re-written.

For this exercise, we assume an embedded processor as an example. In an embedded processor, we can calculate the role of cache array energy consumption to constitute approximately 35% of the overall processor (computation) power [21]. Taking this ratio as an indication, we can calculate that cache arrays alone were responsible for 602TWh of energy consumption in 2021. In 2026, they are expected to consume approximately 850TWh and in 2030 this number is to rise to 1.31PWh. We will need 176 and 271 nuclear power plants to produce the energy consumed by the caches in 2026 and 2030, respectively. To absorb the 509 and 785 Mt CO_2 produced by this energy consumption in 2026 and 2030, respectively, we will need 297 and 458 trillion conifer trees.

We can calculate that using ReRAM or MRAM caches could have saved 601TWh of energy in 2021 or 360MT of CO_2. That is practically the entire energy consumed by caches. This trend is the same in the foreseeable future, indicating that we can practically eliminate the entire impact of caches that we had calculated above by replacing them with memristors. In this exercise, we assumed MRAM and ReRAM devices with similar energy consumption profiles. Using more efficient memristive technology such as more efficient MRAMs will have a starker effect in further reducing the energy consumption associated with the memory and caches.

VI. Discussions

A. Limitations

Like any study, ours has its limitations too. For instance, we picked 63% reported in [5] as an indicative ratio for the energy consumption ratio due to data movement. In practice, this ratio depends on various aspects of a system, such as type of the system, its computer architecture details, configurations, and the application(s) running on it. In [21], the reported ratio is 70%, whereas in [22] even a 90% ratio is reported. It is impossible to pick a single number that would apply to all computing systems and devices. Hence, we picked [5], which is one of the most comprehensive, well-known, and well-cited studies in this area. We emphasize that our calculations provide a reasonable and quantitative estimation of the **scale** of the energy and carbon footprint of computing system but do not pinpoint the actual numbers. To that end, it is more important to know and understand the order of the numbers rather than their actual absolute values. The conclusions then remain the same.

A similar argument can be built around other number used in this study. Hence, it should not be forgotten that this study is in nature indicative and a first effort to provide an **estimate** of the **scale** of the footprint of computer architecture and memory technology. For more accurate numbers large studies are required and given the large number of variables they may still not manage to pinpoint the footprints. Nevertheless, we contend that estimating the scale of these footprints is sufficient to provide a better perspective and further motivate engineers in moving towards reducing them. More importantly, even such rough estimates can help prioritizing our efforts. For instance, comparing the estimates provided here with those calculated for wearable healthcare devices [23], we can go beyond the intuition that IMC has a larger footprint and consequently is more important. For instance, we can learn that the energy consumption of wearable devices has a completely negligible footprint in comparison. Conducting more of such studies can help both engineers and policy makers to more appropriately prioritize research and development in various fields.

B. Extended footprint

We need to bear in mind that there are other aspects that we did not discuss. For example, the production of computing devices and -when applicable- their recycling has a footprint on the global energy consumption and carbon emission. Whereas the computer architecture (in-memory or out-of-memory computing) may make no difference in the production footprint, the material used in emerging memory technologies or the changes in the fabrication technology to allow novel solution could potentially affect their footprint. However, quantifying this effect is very complex and outside the scope of this work. Moreover, it requires intimate knowledge of manufacturing process (and its associated footprint) in large companies who are not very keen in revealing those details.

VII. Status Quo

A. Challenges of emerging memory technologies

It is important to bear in mind, replacing the established memory technologies with the emerging ones is not without challenges. For instance, the endurance or wear out effect could limit the application of technologies such as PCM or ReRAM. Read and write latencies in the majority of these emerging devices are significantly higher, which mean that one cannot easily replace SRAMs with them, especially for on-chip fast caches of high-performance systems. Stepping beyond the memory applications, other factors such as R_{off} to R_{on} ratio and resistance variation can present serious challenges in designing IMC circuits and systems that use these emerging memory technologies [24]. From a technological point of view, the type and combination of materials used for implementing them and the fabrication techniques used significantly affect their properties and their potential for being adopted at a larger scale. Hence, they are active areas of research. From a maturity point of view, one could arguable compare the state of these emerging technologies with the CMOS in the 1960s and 1970s. Although the speed of scientific advancement is significantly higher, the path to maturity for these technologies is long.

B. IMC, market, and industry

Despite their significant role in computers' performance and environment, IMC products do not have a strong presence in the market. UPmem is the only company with an IMC product commercially available on the open market. Their product, DRAM Processing Unit (DPU), consist of DRAM chips that embed one or two processor per bank in each DRAM chip. Their product can replace DDR4 Dual In-Line Memory Module (DIMM) modules and be used as both memory and computation unit. The DRAM portion of the chip is produced using standard 2X-nm DRAM process and the processor is added to the chip and connect to the memory bank via a 64 bit bus[1].

Next in the line is Samsung, which announced two IMC products in 2021 and 2022; Function-In-Memory DRAM (FIMDRAM) [25] and Acceleration DIMM (AXDIMM) [26]. The former, embeds a number of functions per bank that is 3D integrated with high-bandwidth memory dies. Compared to UPmem DPUs, the functions (processing units) are significantly simpler and more limited and are not as tightly integrated as UPmem (in UPmem the processor is on the same die as the memory bank). AXDIMM [26], however, has a quite different architecture, integrating Field Programmable Gate Arrays (FPGAs) next to the memory ranks and on the same DIMM board. Alibaba seems the be another major company to enter the market with their Hybrid Bonding Process-Near-Memory engine [27], which is another 3D integration of memory and compute dies, similar to FIMDRAM.

The common denominator of these products is using DRAM as their memory technology. An exception to this trend would be Mythic, which uses Flash as their base technology [28]. Mythic has a strong focus on Artificial Intelligence (AI) and their key technology is analog matrix multiplication in the memory array. Another shared property for the list above is that at the moment these products do not seem to be commercially available on the open market, except for UPmem and Mythic. However, it is within reason to think that they soon will be.

When it comes to emerging memory technologies, no major IMC product seem to be on or near the market. Although Hewlett Packard (HP) [29] pioneered research on IMC using memristors (in particular, ReRAM), during the last years it has handed over this pioneering role to International Business Machines corporation (IBM) [30], in particular, using PCM. Other major companies are not necessarily deep into the IMC using memristive technology, but have recently put a foot in the door by researching the base memory technology,

[1]https://www.upmem.com/video-upmem-presenting-its-true-processing-in-memory-solution-hot-chips-2019/ , Last visited: March 2024

e.g., Intel, or Taiwan Semiconductor Manufacturing Company (TSMC) by embedding ReRAMs into their 22nm technology and MRAM to their 16nm technology [31]. Panasonic [32] and Infineon [33] are also interested in the trend of emerging memory technologies and have used ReRAM in some of their microcontrollers [31], [33]. Among the myriad of start-up companies of different size active in both memory and IMC usage of these emerging devices, Crossbar Inc.[2] and Weebit Nano[3] stand out. They offer, among others, high-performance ReRAM memories for integration on chip.

Although products based on emerging memory technologies might have a longer path to the market, they certainly are on their way and it would not be surprising if they cross the line by the end of this decade.

VIII. Conclusions

As the seriousness of climate changes and its catastrophic impacts become more of a common knowledge, engineers and computer architects have become more aware of the potential role of computing systems on our planet and consequently our society. In this work, we quantified this effect by looking into energy consumption of computing systems and what they translate to at the global level. Given that we are in the "big data" era, we focused on the impacts of two important aspects of data, namely data movement and data storage technology.

With our calculations, we showed what is at stake when we choose a computing architecture (in- or out-of-memory computing), given that data movement plays a major role in the overall power consumption of current (out-of-memory) computing architectures. To be specific, every year in this decade 1-2.5PWh of energy will be spent on data movement inside computers. Consequently, 0.6-1.4 giga tons of CO_2 will be produced annually. Although various in-memory computing architectures will have different impacts, it is clear that we need to aim at using IMC much more aggressively, if we want to save on the significant footprint that data movement has on our planet.

In a theoretical scenario analysis, we estimated the potential impacts of memory technologies on the overall energy and carbon footprint of computing systems world-wide. Our calculations showed that even using simplified assumptions that rather underestimate their impact, the numbers are too high to be overlooked. More specifically, the cache arrays will be responsible for an annual energy consumption of 0.6-1.3PWh per year and for 360-784 mega tons of CO_2 production. Using emerging memristive technologies, virtually all that footprint can be eliminated since they can bring more than two to five orders of magnitude savings in energy consumption of the memories.

We note that to pin-point these numbers, if possible at all, significantly larger studies are required. However, the current analysis provides a good base to understand the scale of the problem and the promise of IMC, especially using emerging memory technologies, as its solution. Production footprint is another aspect that was outside the scope of this work and is worth a study. To address the concerns raised above, we need to overcome both technological and architectural challenges. While we have a long way to go, the approach and interests of the industry to emerging memory technologies and IMC show that we are on the right path.

We hope that our quantified evaluations, and putting those numbers in tangible examples, help stakeholder and key players to better understand their impact and motivate engineers and computer architect at their work, given the impact it can have on the society at large and on the future of our planet.

[2]https://www.crossbar-inc.com/products/high-performance-memory/ , Last visited: March 2024

[3]https://www.weebit-nano.com/technology/embedded-reram-ip/ , Last visited: March 2024

References

[1] N. Jones, "How to stop data centres from gobbling up the world's electricity," *Nature*, vol. 561, no. 7722, pp. 163–167, 2018.

[2] R. Iyer and E. Ozer, "Visual IoT: Architectural challenges and opportunities; toward a self-learning and energy-neutral IoT," *IEEE Micro*, vol. 36, no. 6, pp. 45–49, Nov 2016.

[3] N. TaheriNejad, "Wearable medical devices: Challenges and self-aware solutions," in *IEEE Life Sciences*, vol. 2, April 2019, pp. 5–6. [Online]. Available: https://lifesciences.ieee.org/lifesciences-newsletter/2019/april-2019/wearable-medical-devices-challenges-and-self-aware-solutions/

[4] A. Ometov and J. Nurmi, "Towards approximate computing for achieving energy vs. accuracy trade-offs," in *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2022, pp. 632–635.

[5] A. Boroumand, S. Ghose, Y. Kim, R. Ausavarungnirun, E. Shiu, R. Thakur, D. Kim, A. Kuusela, A. Knies, P. Ranganathan, and O. Mutlu, *Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks*. New York, NY, USA: Association for Computing Machinery, 2018, p. 316–331.

[6] P. I. Muoka, M. E. Haque, A. Gargoom, and M. Negnevitsky, "Modeling, simulation and hardware implementation of a pv power plant in a distributed energy generation system," in *2013 IEEE PES Innovative Smart Grid Technologies Conference (ISGT)*. IEEE, 2013, pp. 1–6.

[7] D. Hinds and C. Maslak, "Next-generation nuclear energy: The esbwr," *Nuclear News*, vol. 49, no. 1, pp. 35–40, 2006.

[8] J. Malmodin and D. Lundén, "The energy and carbon footprint of the global ict and e&m sectors 2010–2015," *Sustainability*, vol. 10, no. 9, 2018.

[9] C.-Y. Chung, P.-L. Chung *et al.*, "Assessment of carbon dioxide reduction efficiency using the regional carbon neutral model—a case study in university campus, taiwan," *Low Carbon Econ*, vol. 2, pp. 159–164, 2011.

[10] P. Girard, Y. Cheng, A. Virazel, W. Zhao, R. Bishnoi, and M. B. Tahoori, "A survey of test and reliability solutions for magnetic random access memories," *Proceedings of the IEEE*, vol. 109, no. 2, pp. 149–169, 2021.

[11] H. Bahn and K. Cho, "Implications of nvm based storage on memory subsystem management," *Applied Sciences*, vol. 10, no. 3, 2020.

[12] J. S. Vetter and S. Mittal, "Opportunities for nonvolatile memory systems in extreme-scale high-performance computing," *Computing in Science Engineering*, vol. 17, no. 2, pp. 73–82, 2015.

[13] R. C. Sousa and I. L. Prejbeanu, "Non-volatile magnetic random access memories (mram)," *Comptes Rendus Physique*, vol. 6, no. 9, pp. 1013–1021, 2005, spintronics.

[14] A. Carroll *et al.*, "An analysis of power consumption in a smartphone." in *USENIX annual technical conference*, vol. 14. Boston, MA, 2010, pp. 21–21.

[15] D. Economou, S. Rivoire, C. Kozyrakis, and P. Ranganathan, "Full-system power analysis and modeling for server environments," in *In Proceedings of the 2nd Workshop on Modeling, Benchmarking, and Simulation (MoBS)*, IEEE. Boston, MA: International Symposium on Computer Architecture (IEEE), June 2006, pp. 70–77.

[16] A. Mahesri and V. Vardhan, "Power consumption breakdown on a modern laptop," in *International Workshop on Power-Aware Computer Systems*. Springer, 2004, pp. 165–180.

[17] A. Sebastian, M. Gallo, R. Khaddam-Aljameh, and E. Eleftheriou, "Memory devices and applications for in-memory computing," *Nature Nanotechnology*, vol. 15, 03 2020.

[18] H.-S. P. Wong, S. Raoux, S. Kim, J. Liang, J. P. Reifenberg, B. Rajendran, M. Asheghi, and K. E. Goodson, "Phase change memory," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2201–2227, 2010.

[19] X. Fong, Y. Kim, K. Yogendra, D. Fan, A. Sengupta, A. Raghunathan, and K. Roy, "Spin-transfer torque devices for logic and memory: Prospects and perspectives," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 35, no. 1, pp. 1–22, 2016.

[20] H.-S. P. Wong, H.-Y. Lee, S. Yu, Y.-S. Chen, Y. Wu, P.-S. Chen, B. Lee, F. T. Chen, and M.-J. Tsai, "Metal–oxide rram," *Proceedings of the IEEE*, vol. 100, no. 6, pp. 1951–1970, 2012.

[21] W. J. Dally, J. Balfour, D. Black-Shaffer, J. Chen, R. C. Harting, V. Parikh, J. Park, and D. Sheffield, "Efficient embedded computing," *Computer*, vol. 41, no. 7, pp. 27–32, 2008.

[22] B. Dally, "Challenges for future computing systems," in *Computer Science Lecture*, Colorado State University, 2020.

[23] N. TaheriNejad, P. Perego, and A. M. Rahmani, "Mobile health technology: From daily care and pandemics to their energy consumption and environmental impact," *ACM/Springer Mobile Networks and Applications (Editorial)*, pp. 1–5, 2022.

[24] N. TaheriNejad and D. Radakovits, "From behavioral design of memristive circuits and systems to physical implementations," *IEEE Circuit and Systems (CAS) Magazine*, vol. 19, no. 4, pp. 6–18, Fourthquarter 2019.

[25] Y.-C. Kwon, S. H. Lee, J. Lee, S.-H. Kwon, J. M. Ryu, J.-P. Son, O. Seongil, H.-S. Yu, H. Lee, S. Y. Kim, Y. Cho, J. G. Kim, J. Choi, H.-S. Shin, J. Kim, B. Phuah, H. Kim, M. J. Song, A. Choi, D. Kim, S. Kim, E.-B. Kim, D. Wang, S. Kang, Y. Ro, S. Seo, J. Song, J. Youn, K. Sohn, and N. S. Kim, "25.4 a 20nm 6gb function-in-memory dram, based on hbm2 with a 1.2tflops programmable computing unit using bank-level parallelism, for machine learning applications," in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 64, 2021, pp. 350–352.

[26] L. Ke, X. Zhang, J. So, J. Lee, S. Kang, S. Lee, S. Han, Y. Cho, J. Kim, Y. Kwon, K. Kim, J. Jung, I. Yun, S. Park, H. Park, J. Song, J. Cho, K. Sohn, N. Kim, and H. S. Lee, "Near-memory processing in action: Accelerating personalized recommendation with axdimm," *IEEE Micro*, vol. 42, no. 01, pp. 116–127, jan 2022.

[27] D. Niu, S. Li, Y. Wang, W. Han, Z. Zhang, Y. Guan, T. Guan, F. Sun, F. Xue, L. Duan, Y. Fang, H. Zheng, X. Jiang, S. Wang, F. Zuo, Y. Wang, B. Yu, Q. Ren, and Y. Xie, "184qps/w 64mb/mm23d logic-to-dram hybrid bonding with process-near-memory engine for recommendation system," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 65, 2022, pp. 1–3.

[28] M. Maxfield. (2022, April) Meet mythic ai's soon-to-be-legendary analog ai. [Online]. Available: https://www.eejournal.com/article/meet-mythic-ais-soon-to-be-legendary-analog-ai/

[29] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. Stanley Williams, "The missing memristor found," *Nature*, vol. 453, pp. 80–83, May 2008.

[30] M. Le Gallo, C. Lammie, J. Büchel, F. Carta, O. Fagbohungbe, C. Mackin, H. Tsai, V. Narayanan, A. Sebastian, K. El Maghraoui, and M. J. Rasch, "Using the IBM analog in-memory hardware acceleration kit for neural network training and inference," *APL Machine Learning*, vol. 1, no. 4, p. 041102, 11 2023. [Online]. Available: https://doi.org/10.1063/5.0168089

[31] P. Clarke. (2022, Aug.) Tsmc offers 22nm rram, taking mram on to 16nm. [Online]. Available: https://www.eenewseurope.com/en/tsmc-offers-22nm-rram-taking-mram-on-to-16nm/

[32] P. Release. (2017, Feb.) Panasonic and united microelectronics corporation agreed to develop mass production process for next generation reram. [Online]. Available: https://news.panasonic.com/global/press/en170201-3

[33] R. von Faltin. (2022, Nov.) Infineon and tsmc to introduce rram technology for automotive aurix™ tc4x product family. [Online]. Available: https://www.infineon.com/cms/en/about-infineon/press/market-news/2022/INFATV202211-031.html