Is the Future Cold or Tall? Design Space Exploration of Cryogenic and 3D Embedded Cache Memory

Alexander Hankin*§#, Lillian Pentecost†#, Dongmoon Min‡, David Brooks*, Gu-Yeon Wei*

* Harvard University, Cambridge, MA, USA
† Amherst College, Amherst, MA, USA
‡ Seoul National University, Seoul, Republic of Korea
§ Intel Labs, Hillsboro, OR, USA

Correspondence: ahankin@seas.harvard.edu, lpentecost@amherst.edu

***Abstract*—Memory latency, density, and power efficiency are key bottlenecks in a variety of computing systems, and the need for efficient and dense memory solutions is exacerbated by the continued importance of data-intensive applications such as machine learning, graph processing, and scientific computing. A myriad of emerging technologies and approaches aim to address the limitations of current systems. For example, 3D integration can enable highly dense memory structures, and multiple alternative device technologies such as STT and PCM have emerged as compelling solutions to improve memory system density and efficiency. Additionally, cryogenic operation of computing systems (i.e., ultra-low temperature cooling) is becoming a compelling solution as thermal hotspots have become a primary roadblock to conventional transistor scaling.**

This work probes, evaluates, and compares the potential capabilities of 3D integration, embedded non-volatile memories (eNVMs), and cryogenic operation towards improving future memory systems by presenting the first design space exploration of cryogenic operation and 3D integration applied towards the largest on-chip memory structure, the last level cache, as well as presenting and providing open-source tools for future, related design studies. This work specifically evaluates the application-level benefits or limitations of such proposals by leveraging a cross-computing-stack simulation approach. Our studies reveal that the most compelling solution varies depending on the expected memory traffic patterns and workloads of interest, which in turn exposes several opportunities for future optimization and customization. For example, due to potentially high costs of cooling to cryogenic operation, we find that SRAM or 3T-eDRAM operating at 77K is sub-optimal compared to room-temperature SRAM and eNVM solutions, but exhibits advantages for relatively low-traffic workloads.

I. Introduction

The efficiency of modern-day computing systems continues to be limited by on-chip data movement and energy efficiency. Larger working set sizes imposed by the advent of applications like machine learning trigger on-chip memories to fetch data off-chip, resulting in marked performance and energy penalties. Furthermore, applications which require maximum throughput of a CPU for many cycles force the processor to throttle due to the emergence of advanced thermal hotspots [20], compounding negative performance impacts.

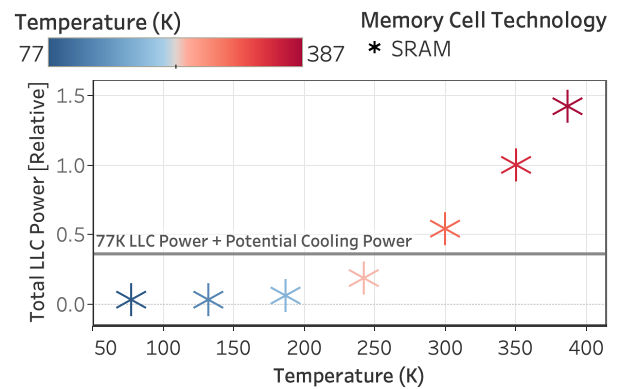


Fig. 1: Total LLC power of a simulated client CPU running SPEC2017.namd (a molecular dynamics benchmark) at temperatures between 77K and 387K, relative to SRAM at typical operating temperature (350K). Reducing operating temperature drastically reduces power, but net benefits are dependent on workload properties and cooling costs.

One among many compelling, emerging approaches to improve memory system efficiency is to leverage cryogenic operation, which refers to operating the computer system in an ultra-low temperature environment. Researchers have been focusing again into temperature as a potential knob to drive performance improvements and keep up with Moore's Law as transistor scaling has become more challenging and thermal hotspots have become a primary bottleneck. In fact, cryogenic computing is an active area that industry focuses on for next-generation computer systems. For example, major companies have already proposed cryogenic-optimized computer devices (e.g., ARM [40], [41], Kioxia [42], IBM [21]) and others are

#Authors contributed equally to this work.

actively working in this area (e.g., Microsoft [46]).[1]

The main advantages of cryogenic operation for traditional silicon systems come from the linearly reduced wire resistivity and virtually nonexistent circuit leakage current [43], [57] which allows for increased clock frequency and reduces overall memory operating power. Typically, for conventional computing with CMOS technology, 77K operation has been explored as an effective temperature to avoid freeze out [38] and high cooling cost [56], while 4K or lower is preferred for superconducting architectures [5], [30], [35], [44], [45], [55]. While cryogenic operation is alluring, the cooling cost associated with it must be carefully considered. For example, studies show that the energy required to maintain a device at cryogenic temperature can be $9.65\times$ higher than the consumed energy of the device being cooled [28], [56], an estimate we will utilize and discuss further in this work.

As an illustration of the promise and limitations of cryogenic operation, Fig. 1 shows total LLC power of a simulated 22nm client processor similar to an Intel Skylake running SPEC2017.namd benchmark at temperatures between 77K and 387K, normalized to 350K operation. The result is derived using the NVMExplorer [37] framework, which we modified to incorporate system configurations from CryoMEM [32], a previously-validated cryogenic memory modeling tool, using memory access traces from Sniper [10]. By reducing operating temperature from 350K to 77K, total operating LLC power can be reduced by more than $50\times$. Even including a conservative estimate of cooling power overhead, there is more than a 50% reduction in total LLC power.

While cryogenic operation shows great promise, the literature is sparse and individual studies are limited in scope, including limited comparison and effective context across disparate technology solutions to improve system efficiency. Most of the cryogenic memory work focuses on dynamic random access memory (DRAM) [21], [28], [46], [48], [51] and very few focus on embedded use cases like the last-level cache (LLC) [32]. Therefore, it is unclear how cryogenic operation fairs when compared with other technology innovations like 3D integration and embedded non-volatile memories (eNVMs); the published benefits of cryogenic memory may be matched or exceeded in general purpose contexts when compared to other technology innovations. If cryogenic operation has the potential to outpace the benefits of orthogonal innovations, it is important to establish the context and extent of that potential. Furthermore, prior art has demonstrated that LLC access behavior affects the optimal memory technology for the LLC [19]. So, any comparison of embedded memory solutions must take into account application behavior using a varied set of benchmarks.

In this work, we present the first design space exploration of cryogenic operation and 3D integration for CPU memory. We consider a broad range of technology options for embedded LLCs, ranging from cryogenically-operated cache memory to 2D and 3D SRAM and eNVMs. We leverage state-of-the-art cryogenic computing and previously-validated 3D eNVM modeling and simulation tools [7], [13], [28], [32], [37], [39] to evaluate each technology. We compare the resulting LLC solutions to determine promising design choices under different optimization goals and workloads of interest. Our studies reveal that while cryogenic operation emerges as a compelling solution under a limited range of application access characteristics, 3D-stacked eNVM technologies offer potential to be more versatile, low-power, and high-performance embedded LLC solutions. These results motivate further studies, among others, in (1) applying cryogenic computing in more specialized computing settings, like classical accelerators, (2) careful consideration and potential optimization of cooling power overheads, (3) tuning 3D integration design and fabrication choices for on-chip CPU memory, and (4) system-level optimizations to support workload characteristics.

[1]In addition, a superconducting-based quantum computer may require a cryogenic-operating processor to aide control of quantum elements, which is another direct application of this work.

II. Background

A. Cryogenic computing for memory devices

1) 77K cryogenic computing overview: Cryogenic computing, which is the concept of running the computers at ultra-low temperatures has emerged as a highly promising idea to maximize the system performance and power efficiency. There exist two representative temperatures and underlying devices for cryogenic computing. First, 77K computing, which is achieved by Liquid Nitrogen (LN), mainly uses CMOS technology because CMOS can operate at a higher speed with relatively low cryogenic-cooling overhead [8], [21], [28], [29], [32], [46], [48]. On the other hand, 4K computing cooled by Liquid Helium (LHe), exploits another device technology called superconducting logic [5], [24], [30], [35], [44], [45], [55]. In this work, we mainly focus on 77K computing using CMOS-device technology.

77K-targeted cryogenic computing can achieve both high performance and power efficiency thanks to the low wire resistivity and leakage current. First, as the wire resistivity is linearly reduced with the temperature [31], we can make much faster computer devices. For example, Copper bulk resistivity is reduced by six times compared to the 300K counterpart [9]. Second, as the transistor leakage current is almost eliminated at low temperatures [43], we can achieve much lower power consumption with aggressive voltage scaling (i.e., reducing supply and threshold voltages at the same time).

2) Previous works in 77K cryogenic memories: For DRAM devices, Tannu et al. showed DRAM reliably operates at cryogenic temperatures [46] and Rambus et al. showed DRAM's cell retention time was significantly prolonged thanks to eliminated leakage current [25], [48]. Based on the observations, Lee et al. developed the performance modeling framework for 77K DRAM (i.e., CryoRAM), and proposed the high-performance and power-efficient DRAM design using the framework [28]. Lee et al. also characterized the row-hammer error behaviors at cryogenic temperatures and proposed near-refresh-free DRAMs with the cryogenic optimal row-hammer

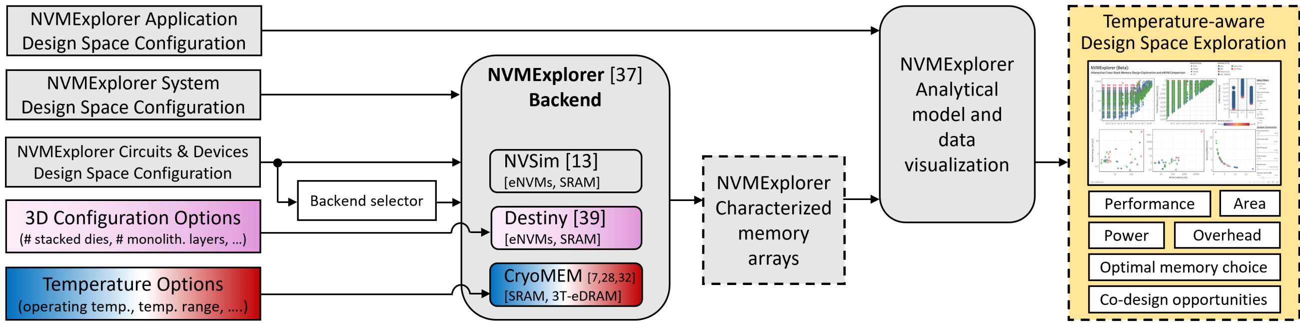


Fig. 2: Block diagram of simulation methodology showing inputs (sharp corners, solid lines), outputs (sharp corners, dashed lines), and functional blocks (rounded corners). This work augments NVMExplorer [37] to interface with alternative memory array characterization tools, Destiny [39] (purple) and CryoMEM [7], [28], [32] (blue/red).

defense [29]. Bae et al. proposed the capacitorless DRAM cell for leakage-free cryogenic memory applications [3].

For cache devices, Min et al. showed SRAM and 3T-eDRAM become highly promising at 77K thanks to the eliminated leakage current, and build the cryogenic SRAM / 3T-eDRAM modeling tools (i.e., CryoMEM) [32]. To build their cache model, they modify CryoRAM by adding SRAM and 3T-eDRAM cache models on its memory-type support. They also validate their modeling with 77K experiment and Hspice simulation using the industry-provided 77K model card. Garzon et al. analyze the operation of 3T-eDRAM and STT-MRAM caches for cryogenic operation [16], [17].

As mentioned above, the most cryogenic memory works focus on DRAM running at 77K. Only a few works focus on embedded use cases such as the last-level cache (LLC) even with its significant impacts on the system performance and power efficiency. Therefore, we focus on LLC in this work, and evaluate the cryogenic-based LLC's feasibility by comparing it with recent technology innovations like 3D integration and embedded non-volatile memories.

3) Promising cell technologies to design 77K LLC: Prior work indicates SRAM and 3T-eDRAM are the most promising cell technologies for 77K cache [32], so we utilize both of these configurations in our evaluations.

First, SRAM is the already widely-used cell technology for cache designs thanks to its relatively faster access speed and reliable, retention-free bit storage, but also suffers from huge static power consumption. Cryogenic temperature drastically reduces static SRAM power thanks to the eliminated leakage current, and thus achieves significant power reduction (as shown in Fig. 1). SRAM also benefits from the low wire resistivity at 77K, which makes the 77K SRAM at least twice faster than the 300K counterpart [32].

3T-eDRAM is also a promising cell technology for 77K LLC, with the cell mechanics described in [32]. The major benefits of 300K 3T-eDRAM are its twice-higher cell density, logic compatibility, and smaller leakage current than SRAM thanks to the three PMOS-only structure, but it also suffers from the huge refresh overhead due to its small storage-node capacitor. In fact, 3T-eDRAM-based LLC cannot run the ordinary workloads at 300K due to its huge IPC reduction (94% for PARSEC workloads [32]) from refreshing. As the eliminated leakage current prolongs the retention time more than 10,000 times and completely resolves refresh overhead, 3T-eDRAM becomes compelling at lower-temperature operation.

B. Embedded Non-Volatile Memory

Emerging, embeddable non-volatile memories (eNVMs) have been demonstrated as beneficial alternatives to traditional embedded memories like SRAM and eDRAM. In addition to their persistent nature—i.e. their ability to maintain their data after power is turned off—eNVMs also provide significant density improvements over SRAM. This is particularly attractive given the high volume of data that is ubiquitous in modern day computing systems. Furthermore, eNVMs use significantly less power than SRAM and eDRAM because of the lack of standby power. This lack of standby power comes from the fact that non-volatile memories do not store charge. Rather, they store information through the process of modulating physical properties of a material.

The most prominent eNVM technologies to emerge include Phase Change Random Access Memory (PCRAM), Resistive RAM (RRAM), Magnetic RAM (MRAM) including Spin-Torque Transfer RAM (STT-RAM) and Spin-Orbit Torque RAM (SOT-RAM), Charge Trap Transistor RAM (CTT-RAM), and ferroelectric-based RAMs (FeFET, FeRAM). Each memory has unique physical properties which manifest themselves at the architecture level which makes certain memories more attractive for different use cases. In particular, STT-RAM has very high endurance (similar to SRAM) which makes it a feasible alternative for write-intensive memory structures like the last-level cache of a CPU; however, this comes at the expense of potentially degraded write performance [19], [37]. A more emerging flavor of magnetic RAM, SOT-RAM improves significantly on the write performance of STT-RAM at the expense of increased read latency [23].

All of the aforementioned eNVM technologies are CMOS compatible and can be relatively easily integrated into modern fabrication processes. For the purposes of this work, since we are focusing on a general purpose CPU last-level cache, we focus on the most promising technologies for this use case:

PCRAM, STT-RAM, and RRAM, with details and specific device properties derived from [37].

1) Previous works in embedded non-volatile memory: Dong et al. [13] developed NVSim, a circuit-level performance model for various eNVM devices including PCM, STT-RAM, and RRAM. By utilizing NVSim, Pentecost et al. [37] developed NVMExplorer, a design space exploration framework to compare and evaluate various eNVM technologies for various application and system configuration. By using NVMExplorer, they evaluated various eNVM technologies for different applications and found the most appropriate eNVM for each case. Korgaonkar et al. [27] proposed the novel STT-RAM LLC design to mitigates the long write latency. Wang et al. [50] propose SRAM and STT-RAM hybrid cache design and develop adaptive placement policy for it. Wang et al. [49], Agarwal et al. [1], and Duan et al. [14] proposed the techniques to improving non-volatile cache lifetime. Wu et al. [53] and Guo et al. [18] proposed PCM-SRAM hybrid cache architecture to achieve high capacity of PCM and reliability of SRAM. By extending NVSim's capability to 3D domain, Poremba et al. [39] present Destiny, the circuit level modeling tool for 3D cache design using SRAM, eDRAM, and eNVM devices. Mittal et al. [34] explore the design space of 3D eNVM and eDRAM caches by using the Destiny tool. However, no previous work investigated the area, total power, and latency of 3D eNVM-based LLC for wide range of the workload traffic, die count, and cell technologies. Also, there is no previous work which compare 3D eNVM-based LLC with emerging cryogenic memories.

C. 3D Integration

3D integration refers to the process of using multiple dies in a vertical fashion to increase the compute resources of a microprocessor. There are multiple methods for 3D integration including face-to-face, face-to-back, and monolithic stacking [4]. Each integration method has different trade-offs associated with it. For example, face-to-face stacking has the potential for higher via density; however, there is a limit of only two layers which can be stacked in this way [39]. Monolithic stacking also allows for higher via density; however, its drawback is that transistors cannot be formed on higher layers as it can destroy transistors formed on lower layers [39].

Recently, modeling and simulation tools have begun to emerge for studying 3D integration. For example, CACTI-3DD [11] is an architecture-level power, area, and timing model for 3D integration of off-chip DRAM. It is based on the previous versions of the CACTI tool which performs architecture-level power, area, and timing modeling for conventional memory systems [52]. In addition to CACTI-3DD, more niche modeling and simulation have emerged as well including Destiny [39]. Destiny is a power, area, and timing model for 3D integration of both on-chip and off-chip SRAM, eDRAM, and eNVMs (including PCM, STT-RAM, and RRAM). Destiny, in turn, is built on top of NVSim [13], which is a similar modeling framework for 2D eNVMs.

III. Methodology

An overview of our simulation methodology is shown in Fig. 2. This work extends NVMExplorer [37], a cross-stack design space exploration framework to evaluate and compare embedded non-volatile memory (eNVM) arrays and identify optimal configurations and characteristics for specified applications. The inputs include application characteristics (e.g., traffic patterns, fault tolerance), system design space and constraints (e.g., capacity, bank organization), and circuits and devices choices (e.g., technology-level details, memory cell properties). However, the original NVMExplorer can only support conventional SRAM and 2D eNVM because it relies specifically on NVSim to simulate and compute memory array characterization of the given technology configurations [13]. To support 3D-stacked dies and various temperatures down to the cryogenic region, this work then modifies NVMExplorer to include a choice of backend simulator for memory array characterization (highlighted in color in Fig. 2). These changes empower additional configuration options and research opportunities, including building context, comparison, and optimization of 3D-stacked dies, monolithic layers, and a range of operating temperatures.

More precisely, this work contributes to the open-source NVMExplorer framework by integrating interfaces for Destiny [39] and CryoMEM [7], [28], [32]. Destiny is a memory array characterization tool which expands the capabilities of NVSim [13] to model fabrication strategies for 3D integration and 1T1C-eDRAM (in addition to the range of eNVMs modeled in NVSim (PCM, STT-RAM, ReRAM)). CryoMEM is an analytical model built around CACTI-3DD [52] for SRAM, 3T-eDRAM and main memory operating at room to cryogenic temperatures (400K down to 77K). Memory array characteristics (e.g., access energy, access latency, leakage power) are generated via one of several simulation backends according to user configuration, and NVMExplorer's analytical model leverages array-level results to compute application-level metrics, which includes a comparison of performance, power, and area to determine the optimal memory choice for a particular design target and to unlock co-design opportunities. This work integrates effective tools for multiple compelling memory solutions using Destiny, CryoMEM and NVMExplorer, to go beyond prior efforts by analyzing the application-level impact of novel technologies for various workload characteristics (e.g., read access/s, write access/s). This work has been integrated into the open-source NVMExplorer framework and is available at http://nvmexplorer.seas.harvard.edu/.

A. System and Technology Configurations

NVMExplorer supplies a database of eNVM cell technology characteristics published in the recent VLSI symposiums (ISSCC, IEDM, and VLSI from 2016—2020), and provides a 'tentpole' approach which selects the extrema of the cell-level characteristics to represent the range of potential behavior of each technology among a large volume of eNVM technologies. For PCM, STT, and RRAM configurations, we utilize NVM-Explorer's provided database [37] to obtain memory cell def-

initions per-technology to identify optimistic and pessimistic design points, with array architectures optimized for energy-delay-product. For the input device parameters of cryogenic SRAM and 3T-eDRAM in CryoMEM-derived results, we utilize 22nm HP devices (i.e., Vdd = 0.8V / Vth = 0.5V) following PTM and ITRS loadmap [58]. We configure all LLC designs as 16-way set-associative, dual-port, and ECC-supported 16 MB caches fabricated with 22nm technology, with key parameters summarized in Table I.

TABLE I: Key CPU model parameters

Class	Desktop (based on Intel Skylake)
Num. cores	8
Process node	22nm
Frequency	5 GHz
L1I$	32 KiB
L1D$	32 KiB
L2$	512 KiB
L3$	shared 16 MiB, 16 ways

B. Benchmarks

We use the full SPECrate CPU2017 suite of workloads [6] in order to represent the modern range of memory traffic that a CPU cache will experience in a general purpose processor. In our performance model, we extrapolate and check based on access latency and SRAM-based traffic statistics per benchmark whether an NVM-based solution will meet the total bandwidth and expected access latencies without incurring slowdown. In this way, you can read any result in Section IV above a relative value of '1' in total LLC latency as a solution that will negatively impact performance, while those solutions falling well below 1 consistently match or outperform the expected latency-per-access and BW potential of the baseline SRAM for that workload. The precise read/write access counts per benchmark are simulated using Sniper with a 1-die-SRAM-based LLC at 350K, and the computed bandwidths use the per-execution access counts and execution time to extrapolate to read accesses/s and write accesses/s under continuous operation of that benchmark.

We normalize array-level and application-level results to those of 350K SRAM. In the case of the SPEC2017 analyses, we normalize all results to those of 350K SRAM results for one reference benchmark (namd shown in Fig. 1). CryoMEM [7], [28], [32] results for power and performance under cryogenic operation have been validated *relatively*, so our comparisons leveraging CryoMEM simulation results and those derived from Destiny are each relative to 350K SRAM per simulation framework. For our analysis, we first compare the performance, power, and area (PPA) of embedded cryogenic SRAM and 3T-eDRAM to their non-cryogenic counterparts to evaluate when cryogenic operation is worthwhile. Then, we compare cryogenic SRAM and 3T-eDRAM with other promising embedded memory proposals like STT-RAM, RRAM, and PCM in both 2D and 3D-stacked configurations. While Destiny also models 1T1C-eDRAM, we exclude it from this analysis as prior work has shown that it is generally slower and exhibits higher dynamic energy than SRAM and 3T-eDRAM [32], [53], [54].

C. Cooling Power Overheads

Previous 77K works use the cooling overhead of 100kW CryoCooler (9.65x) because they target a large-scale server system [7], [28], [32]. Cooling overhead is amortized with increased cooling capacity of the cooling systems, leading to lower relative overhead per system, so we also compare to more conservative scenarios by using the cooling overhead of the lower-capacity systems, for example that of a single desktop. Following the Fig 4.5 of "Case studies in superconducting magnets", we vary the relative operating power overhead from 9.65 (100kW total capacity), 14.3 (1kW total capacity), 21.8 (100W capacity), and 39.6 (10W capacity) to demonstrate the relative potential costs of cryogenic operation at varying computing scales. [56].

IV. Results and Analysis

Beginning with well-established on-chip memory technologies (SRAM and eDRAM), we first examine how memory array characteristics, like latency-per-access and energy-per-access, vary with operating temperature. After evaluating the power and performance of cryogenic computing under realistic application traffic for an LLC, we contextualize the results by comparing to another promising proposal: embedded non-volatile memories (eNVMs). The potentially high density and low leakage of eNVMs have proven beneficial in various use cases including CPU LLC and accelerator scratchpad memories [19], [36], [37].

A. When is cryogenic computing viable?

Fig. 3 shows memory characteristics at varying operating temperature for SRAM and 3T-eDRAM energy-delay-optimized 16MB arrays. The temperature range we select is between 77K (CMOS-compatible cryogenic) to an approximate CPU thermal design point (387K) at intervals of approximately 50 degrees. All results are normalized to 350K SRAM. As the dynamic power mainly depends on the memory-circuit configurations and the operating voltage, the dynamic power has little sensitivity to temperatures, as confirmed by the nominal variation (approximately 10%) in read/write energy-per-bit from 77K to 387K SRAM in Figure 3.

Conversely, static energy consumption at 77K cryogenic operation, as represented by leakage power in Figure 3, right, is approximately 1,000,000$\times$ less than 350K SRAM leakage power thanks to the eliminated subthreshold leakage current [26], [43]. We also observe that, as the temperature increases, the relative leakage power of 3T-eDRAM shifts from approximately 10$\times$ to 100$\times$ less than the leakage power of SRAM. As the PMOS-only 3T-eDRAM cell significantly reduces the subthreshold leakage current [12], its benefit becomes higher at the higher temperature (with the higher leakage current).

Fig. 3 also demonstrates that read/write latency is temperature-dependent, with cryogenic-operation latency about 70% lower than 350K SRAM latency thanks to the reduced wire resistivity at 77K. The lower latency allows for higher bandwidth and faster access to the cryogenic LLCs.

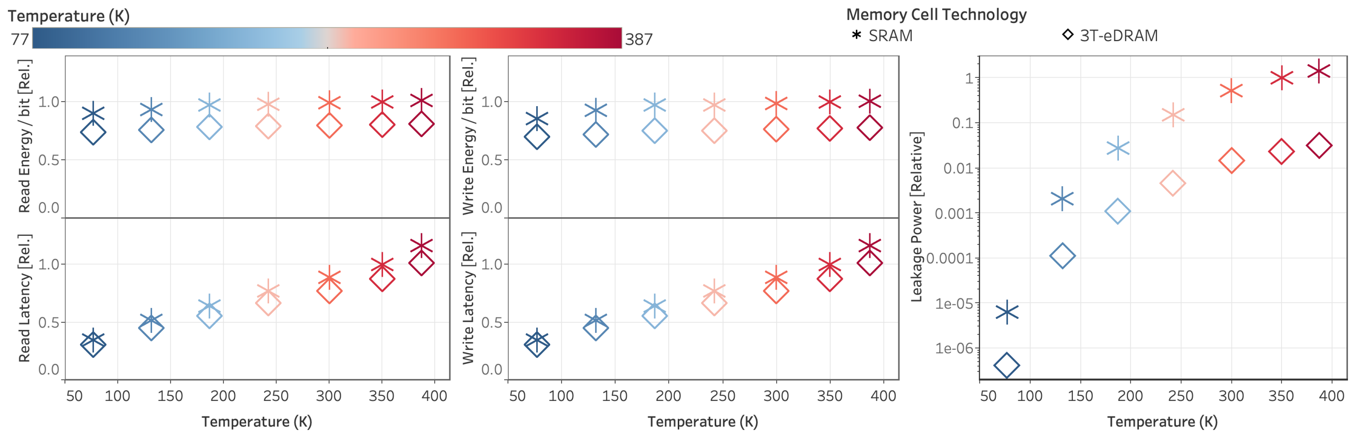


Fig. 3: Array-level characterization for 16MB iso-capacity SRAM and 3T-eDRAM under varying operating temperature (77K–387K), relative to characteristics of SRAM at 350K.

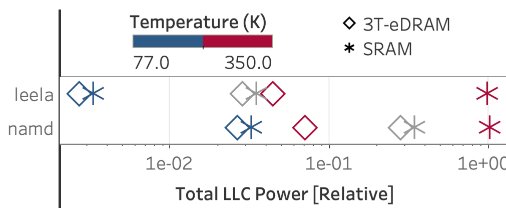
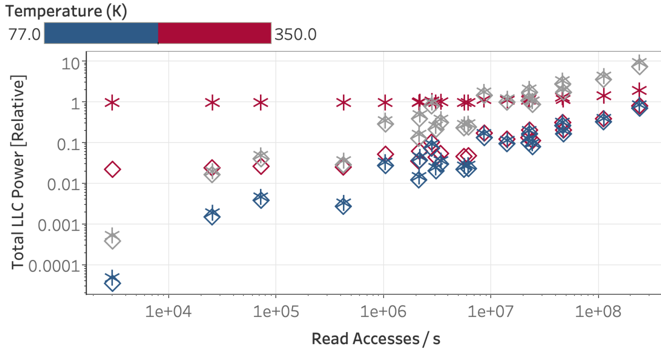


Fig. 4: For two example SPEC 2017 benchmarks, the total LLC operating power under constant operation varies by orders of magnitude at room temperature (red) as compared to cryogenic operation (blue) and cryogenic operation including an estimate of cooling power (gray), as discussed in Section III (shown here relative to room temperature SRAM for the namd benchmark). The full benchmark suite results for operating memory power and access latency are in Figure 5.

To establish the viability of cryogenic LLC solutions, we must consider the realistic operating power of the memory resource as well as any potential performance degradation introduced by changing the LLC technology or operating temperature. Fig. 4 highlights the total LLC power (including leakage + dynamic energy due to memory access pattern) of two example SPEC2017 benchmarks at 350K and 77K, while including an estimated cooling overhead to maintain cryogenic operation at 77K (shown in grey, as discussed in Section III). We note that the traffic pattern and memory load of the workload dictates whether the relative benefits of cryogenic operation are worthwhile for each memory technology. For example, in considering operating a 16MB SRAM LLC for the namd workload (including cooling power overhead estimates), we observe that total relative power reduces by nearly 3× by operating at 77K rather than 350K. Alternatively, the potential benefits of cryogenic operation of an eDRAM cache for the same benchmark are thwarted by the cooling power overhead compared to 350K eDRAM operation due to the huge LLC accesses of the workload. For distinct benchmark memory access patterns, like leela, which executes Monte Carlo simulations for the game Go, cryogenic total operating power is advantageous for both LLC technologies, which prompts further investigation as to under what conditions, constraints, and cooling overheads cryogenic operation is worthwhile.

Fig. 5 shows total LLC power and total LLC latency for the SPEC2017 benchmarks under 77K vs. 350K operating temperature, normalized to those of 350K SRAM. The gray points indicate 77K operation including projected cooling power overheads. Each SPEC2017 benchmark is defined by a specific number of reads/writes-per-second assuming continuous operation of that benchmark and represented as each column of points. Write bandwidth strongly determines the potential slowdown, while the total operating power is shaped by aggregate traffic, so we aim to capture these dependencies in each subfigure while including both the read and write traffic values per benchmark. Fig.5 (left) displays total LLC power vs. read-accesses-per-second for the range of LLC traffic across SPEC2017 workloads, identifying 77K 3T-eDRAM as the lowest power option for all benchmarks. For workloads with read accesses less than 1e4 (povray), cryogenic operation reduces total power significantly, even if cooling power overheads were significantly higher. Between 1e4–1e6 read accesses-per-second, cryogenic operation retains some advantage, and the potential of reduced cooling overheads could enable cryogenic LLC as a power-efficient solution across a range of CPU workloads. Even in this range, 77K 3T-eDRAM exhibits consistently lower power than 350K SRAM. However, for high-bandwidth benchmarks, for example at read access rates about $10^8/s$, the relative power of cryogenic operation and cooling well exceeds the 350K operating baseline, perhaps precluding viability in the absence of significant cooling innovations.

Fig. 5 (right) shows total LLC latency vs. write accesses-per-second for all benchmarks. Across all traffic patterns, 77K 3T-eDRAM is the preferred technology with a slight advantage over 77K SRAM. Furthermore, 77K 3T-eDRAM and 77K SRAM exhibit 2-4× lower aggregate LLC latency than at 350K, consistent with the difference observed in array access latency (Fig. 3). This points to the performance capabilities



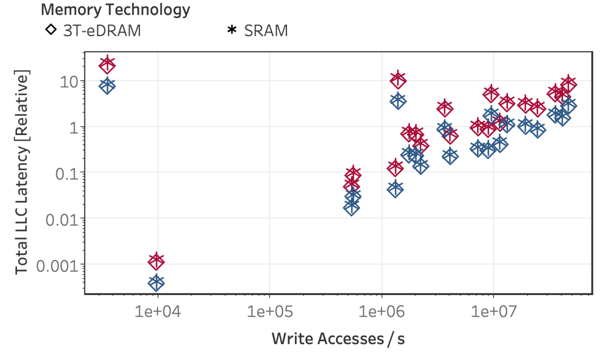


Fig. 5: Total LLC power and total LLC latency for SPEC2017 benchmarks at cryogenic operating temperature (77K, blue), conventional operating temperature (350K, red) and cryogenic operating temperature including cooling overhead (gray), relative to 350K SRAM executing the reference benchmark.

and consistency of cryogenic solutions, a boon when combined with the potential to reduces total memory power.

In summary, we emphasize our key unique observations for the cryogenic LLCs over the previous work. First, we observe that the cryogenic LLCs are not always better than conventional SRAM LLC. Cryogenic LLCs are highly effective for lower-memory-traffic LLC scenarios. However, for the high LLC-access rate of some SPEC2017 workloads, its power efficiency can be even lower than than that of 300K SRAM due to the high dynamic power and cooling cost. Prior work did not uncover this due to their selected benchmarks (PARSEC 2.0) for their evaluation. Second, for the LLC application, we observe that 77K 3T-eDRAM always outperform 77K SRAM for static power, dynamic power, and access latency across workload traffic patterns.

B. How does cryogenic LLC compare to 3D eNVMs?

To compare the disparate benefits of 3D stacking compared to cryogenic operation, we first look at array characteristics as a function of 3D scaling at a conventional operating temperature. We utilize the "tentpole" methodology of NVMExplorer [37] described in Section III to obtain optimistic and pessimistic design points per technology, with array architectures optimized for energy-delay-product.

First, we consider the differences in 2D footprint among iso-capacity LLCs. We vary the number of stacked dies up to 8 and observe that the relative benefit of die stacking changes for different technologies and for different number of dies, as shown in Figure 6. As number of dies increases, the relative benefit of stacking, in terms of area, decreases. Additionally, some technologies benefit more from die stacking than others. 8-die SRAM, for example, achieves more than an 80% reduction in 2D area compared to 1-die SRAM whereas PCM achieves only about a 30% reduction in 2D area from 1-die to 8-dies. This is a function of both the memory cell area of the 2D configuration (2D SRAM starts from a bigger cell area) and the potential array-level area efficiency per technology. Fig. 6 identifies 8-die PCM as the most area efficient memory option, with 8-die STT-RAM and 8-die RRAM as the next best options. 8-die PCM (under optimistic underlying cell assumptions) exhibits over 10× reduction in 2D area footprint compared to 1-die SRAM. It is also interesting to note that compared with 8-die SRAM, all the 8-die eNVMs are at least 2× as dense, and 3D-stacking appears to compound the inherent potential density advantage of various eNVM technologies, at least at a fixed comparison in a 22nm technology node.

Fig. 6 (left) shows that SRAM and PCM exhibit the lowest read/write energy-per-bit. The best read energy-per-bit is achieved by 8-die SRAM and 8-die PCM, while SRAM offers lower write energy-per-bit, regardless of stacking, as expected due to the asymmetry of reads and write energy for these eNVMs. 8-die SRAM and 8-die PCM read energy-per-bit are approximately 75% lower and 55% lower, respectively, than the baseline (1-die SRAM). In terms of read latency, 8-die PCM is the best option, followed by 4-die PCM, 2-die PCM, 8-die STT-RAM and 8-die RRAM, all offering over 80% lower latency than the SRAM baseline. 8-die STT-RAM exhibits lowest write latency. In fact, both 3D and 2D STT-RAM solutions exhibit lower write latency and competitive read latency compared with all other memory technologies.

Fig. 7 shows total LLC power and total LLC latency across SPEC2017 benchmarks, as in Section IV-A, but for the 2D and 3D SRAM, PCM, RRAM, and STT-RAM arrays characterized in Fig. 6. Here, we see that several eNVMs achieve lower total LLC operating power than SRAM across a range of benchmark traffic because even solutions with higher energy access-per-bit exhibit drastically lower leakage power than SRAM. In fact, the eNVM technologies exhibit 2-10× lower power than the SRAM baseline for read accesses-per-second less than 1e7, even considering eNVMs with pessimistic underlying cell properties. The benefits of STT-RAM solutions decrease in the range of 1e4 to 1e8 accesses-per-second as dynamic write power begins to dominate total LLC power. For read accesses greater than 1e7, 8-die PCM emerges as the lowest power technology. Lower power consumption of 3D solutions like 8-die PCM in these regimes of higher access rates is primarily attributed to reduced energy-per-bit of access, made more impactful at highest bandwidth expectations, which is in turn at least partially attributed to the physical area achieved from more stacking, resulting in shorter wire lengths and changes in the internal array architecture and organization.

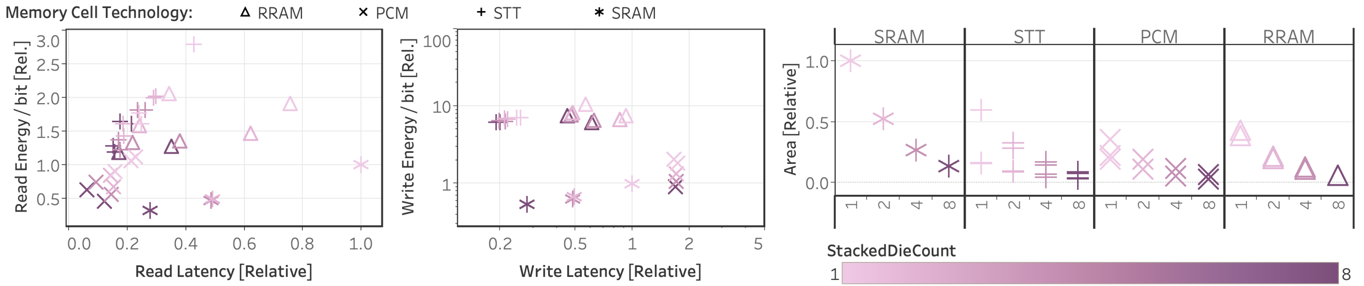


Fig. 6: Array-level characterization for 2D and 3D eNVMs at 350K, relative to characteristics of 16MB iso-capacity 2D SRAM.

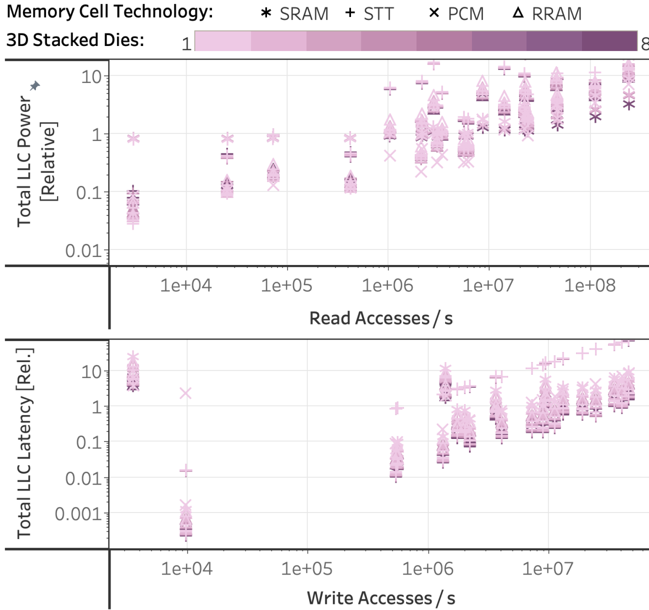


Fig. 7: Total LLC power and total LLC latency vs. workload traffic properties for 2D and 3D eNVMs for the SPEC2017 benchmarks at 350K, relative to SRAM executing the reference benchmark.

Fig. 7 (bottom) shows total LLC latency as a function of writes accesses. As in the comparison of cryogenic vs. room temperature operation in Fig. 7, one LLC solution consistently exhibits the lowest aggregate latency. In this case, it is 8-die STT-RAM (under optimistic underlying cell properties) for all benchmarks except mcf (the lowest write traffic). 8-die STT-RAM exhibits approximately 10× lower latency than the SRAM baseline, followed narrowly by 4-die STT-RAM and 2-die STT-RAM. Other eNVM technologies exhibit total LLC latency within 10× of SRAM. Interestingly, at higher rates of write traffic, PCM and STT-RAM with pessimistic underlying cell properties are consistently higher latency than SRAM, and could thus introduce a negative performance impact as an LLC solution. Due to the asymmetry of read and write costs for eNVM-based solutions, we observe that high-write-traffic and write-dominated benchmarks exhibit diminished relative benefits, though total LLC power for most eNVM solutions remains closer to the SRAM baseline than when we consider cryogenic solutions in the high-traffic regime.

In summary, we emphasize our key unique observations for the 3D stacking over previous work. First, in high-traffic scenarios, higher stacking is better for power efficiency, and in lower-traffic scenarios, lower stacking is better for power efficiency. Second, PCM and STT-RAM are always optimal for the power efficiency and higher performance respectively for every scenario.

C. Summary

We observe that the lowest-power LLC solution, across innovative system solutions like cryogenic operation, eNVMs, and 3D-integration strategies, is dependent on the quantity and read/write balance of workload traffic. We summarize our observations with respect to the volume of read traffic across benchmarks vs. preferred LLC solution under varying optimization goals in Table II. For read traffic less than 10^4 read accesses-per-second, 77K 3T-eDRAM is preferred with more than a 2,500× reduction in power compared to the baseline (350K SRAM) even taking into account a conservative estimate for cooling overhead (shown in gray, Fig. 5). As the volume of read accesses-per-second increases to between 5×10^4 to 8×10^6, 77K 3T-eDRAM and 77K SRAM continue to achieve 20-30× power reduction (including cooling power overhead) compared to 350K SRAM running the same benchmark, but 4-die 3D PCM emerges to achieve slightly more than 20-30× power reduction. Then, for the highest-traffic benchmarks (greater than 8×10^6 read accesses-per-second), 8-die 3D PCM is the lowest power solution. While a significant reduction in cooling overhead compared to what we have modeled could improve the viability of cryogenic operation in this traffic range, we observe a range of 2D and 3D eNVM solutions out-perform 350K SRAM for higher-traffic benchmarks in terms of both power efficiency and potential density. Finally, we note that prior work highlights that eNVMs exhibit varying endurance characteristics, which may be a limitation particularly for PCM and RRAM solutions [37]. In light of this observation, in the cases where PCM is identified as optimal, Table II also lists the second-most-preferred LLC (labeled in the table as “alt”).

These results have several implications for future research directions. While cryogenic operation remains a compelling proposal for ultra-low-leakage memory solutions, more careful study of cooling costs would be warranted in order to make

TABLE II: Summary of optimal LLC solution for different design targets (e.g., total LLC power, performance, or area).

Workload read accesses-per-second	Optimal LLC		
	power (100kW cooling)	performance	area
$<5\times10^4$	77K 3T-eDRAM	8-die 3D STT-RAM / 8-die 3D PCM	3D PCM alt: 3D STT
5×10^4 to 8×10^6	4-die 3D PCM alt: 77K 3T-eDRAM		
$>8\times10^6$	8-die 3D PCM alt: 8-die 3D SRAM		

this a feasible option for a wider range of realistic cache traffic. Alternatively, cryogenic operation might be better-suited to more specialized computing systems and settings where memory traffic is well-understood, relatively lower overall traffic, and perhaps when ambient operating temperatures are advantageously cool (e.g., embedded operation in outer space). On the other hand, 3D-stacked eNVM solutions, particularly STT-RAM- and PCM-based solutions, emerge as compelling, energy-efficient, and performant solutions from these studies across a wider range of workload traffic. The scalability, cost, and effectiveness of 3D integration strategies, in addition to the reliability and lifetime of eNVMs, should be central to further evaluations of these potential LLC solutions.

V. Discussion

A. *Cooling and Thermal Overheads of Cryogenic Computing*

One of the major challenges of cryogenic computing is non-trivial cooling overhead to maintain the temperature of 77K. The cooling overhead indicates the required input energy to remove the unit heat (1J) from the cooling systems. Prior works [28], [29] have modeled the 77K cooling overhead as 9.65 times of device power consumption based on the real data from 235 cryocoolers surveyed by previous works [47], [56]. That is, to achieve power efficiency over 300K systems, 77K systems should consumes 10.65 times less power than 300K systems. In addition, the thermal budget analysis is also crucial because the benefits of cryogenic computing mainly originate from the 77K environments.

As the LLC is located inside the processors, utilizing the cryogenic LLC requires the entire processors to cool down to 77K. In this scenario, other CPU parts (e.g., CPU core, NoC, L1/L2 cache) may incur the significant (1) cooling-power cost or (2) thermal-related problems. However, we note that they do not incur the serious cooling and thermal problems as the previous works already resolve the problems as follows.

Cooling power from other CPU parts. Previous works proposed the cryogenic-optimal microarchitectures for the major computer devices (i.e., CPU core [7], [33], NoC [33], L1/L2 cache [32], DRAM [28], [29]) and already achieved their power efficiency (i.e., lower power compared to 300K device) even including the cooling cost. Therefore, the cryogenic cooling for LLC does not incur the significant cooling power cost of other CPU parts.

Thermal problem from other CPU parts. Thanks to the higher heat transfer speed of materials (e.g., Silicon, Cu, LN) [2], [15], [22], thermal-related problems is negligible in cryogenic computers. For example, the previous works showed that the conventional LN bath-cooling method for cryogenic computing has 2.41 times higher cooling capacity (157W) than the 300K air-cooling method (65W) with 20K of little temperature variation [7]. That is, the heat dissipation from other CPU parts does not affect the reliability of the 77K LLC.

VI. Future Work

During the course of this work, additional research directions have been exposed which we believe the community should pursue. In evaluating cryogenic computing, we observed some opportunities for optimization. For example, the ideal temperature to run the processor at may not be exactly room temperature or cryogenic temperature. Instead, sometimes the optimal temperature is in-between these two operating points. Additionally, the relative benefits of reduced temperature varies according to workload traffic patterns. Therefore, a processor which has the capability to dynamically adjust the operating temperature of the processor may be the optimal method. To achieve this, we believe that temperature should be exposed as a design knob for computer systems.

Furthermore, while 3D stacking proved to be highly beneficial in terms of performance, the resulting power efficiency was very poor in some cases. This may not be a cost that can be incurred (even small power cost) anymore due to the temperature/hotspot bottleneck that is emerging as a result of advanced hotspots [20]. We believe a future interesting work would be to combine both 3D stacking with cryogenic computing to achieve both highly performant and low power/temperature chips for the broadest range of workload traffic patterns.

VII. Conclusion

To overcome the “memory wall” problem, cryogenic operation as well as 3D integration have been proposed as promising options for efficient future memory systems. In this work, we present a design space exploration of embedded cache memory which considers cryogenic operation and 3D integration. We evaluate a range of embedded memory technologies, including SRAM, 3T-eDRAM, PCM, STT-RAM, and RRAM. We take a cross-computing-stack approach to this design space exploration in order to evaluate the impact of application behavior on the optimal memory choice, and find that for different guiding prioritie—i.e., performance, power, and area— the optimal memory choice depends on application traffic. For workloads with low read traffic ($<5\times10^4$ read accesses-per-second) like SPEC2017.povray, 77K 3T-eDRAM is the optimal memory choice for power efficiency, while 3D STT-RAM and 3D PCM are the most performant choices. For read-dominated workloads with high read traffic ($>8\times10^6$ read accesses-per-second) like SPEC2017.mcf, 3D PCM is the optimal memory choice for power efficiency. Our analysis suggests that cryogenic operation is more compelling for more specialized computing systems whereas 3D eNVMs are more versatile, and thus more promising for the LLC.

REFERENCES

[1] S. Agarwal and H. K. Kapoor, "Improving the lifetime of non-volatile cache by write restriction," *IEEE Transactions on Computers*, vol. 68, no. 9, pp. 1297–1312, 2019.

[2] J. Arblaster, "Thermodynamic properties of copper," *Journal of Phase Equilibria and Diffusion*, vol. 36, no. 5, pp. 422–444, 2015.

[3] J.-H. Bae, J.-W. Back, M.-W. Kwon, J. H. Seo, K. Yoo, S. Y. Woo, K. Park, B.-G. Park, and J.-H. Lee, "Characterization of a capacitorless dram cell for cryogenic memory applications," *IEEE Electron Device Letters*, vol. 40, no. 10, pp. 1614–1617, 2019.

[4] B. Black, D. Nelson, C. Webb, and N. Samra, "3d processing technology and its impact on ia32 microprocessors," in *IEEE International Conference on Computer Design: VLSI in Computers and Processors, 2004. ICCD 2004. Proceedings.*, 2004, pp. 316–318.

[5] D. Brock, "Rsfq technology: Circuits and systems," *International Journal of High Speed Electronics and Systems*, 03 2001.

[6] J. Bucek, K.-D. Lange, and J. v. Kistowski, "Spec cpu2017: Next-generation compute benchmark," in *ACM/SPEC International Conference on Performance Engineering*, 2018.

[7] I. Byun, D. Min, G.-h. Lee, S. Na, and J. Kim, "Cryocore: A fast and dense processor architecture for cryogenic computing," in *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)*, 2020, pp. 335–348.

[8] I. Byun, D. Min, G. Lee, S. Na, and J. Kim, "A next-generation cryogenic processor architecture," *IEEE Micro*, vol. 41, no. 3, pp. 80–86, 2021.

[9] W. D. Callister Jr and D. G. Rethwisch, *Fundamentals of materials science and engineering: an integrated approach*. John Wiley & Sons, 2020.

[10] T. Carlson, W. Heirman, S. Eyerman, I. Hur, and L. Eeckhout, "An evaluation of high-level mechanistic core models," *ACM Transactions on Architecture and Code Optimization*, 2014.

[11] K. Chen, S. Li, N. Muralimanohar, J. H. Ahn, J. B. Brockman, and N. P. Jouppi, "Cacti-3dd: Architecture-level modeling for 3d die-stacked dram main memory," in *2012 Design, Automation Test in Europe Conference Exhibition (DATE)*, 2012, pp. 33–38.

[12] K. C. Chun, P. Jain, J. H. Lee, and C. H. Kim, "A 3t gain cell embedded dram utilizing preferential boosting for high density and low power on-die caches," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 6, pp. 1495–1505, 2011.

[13] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, "Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 31, no. 7, pp. 994–1007, 2012.

[14] G. Duan and S. Wang, "Exploiting narrow-width values for improving non-volatile cache lifetime," in *2014 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2014, pp. 1–4.

[15] P. Flubacher, A. Leadbetter, and J. Morrison, "The heat capacity of pure silicon and germanium and properties of their vibrational frequency spectra," *Philosophical Magazine*, vol. 4, no. 39, pp. 273–294, 1959.

[16] E. Garzón, R. De Rose, F. Crupi, A. Teman, and M. Lanuzza, "Exploiting stt-mrams for cryogenic non-volatile cache applications," *IEEE Transactions on Nanotechnology*, vol. 20, pp. 123–128, 2021.

[17] E. Garzón, Y. Greenblatt, O. Harel, M. Lanuzza, and A. Teman, "Gain-cell embedded dram under cryogenic operation—a first study," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 29, no. 7, pp. 1319–1324, 2021.

[18] S. Guo, Z. Liu, D. Wang, H. Wang, and G. Li, "Wear-resistant hybrid cache architecture with phase change memory," in *2012 IEEE Seventh International Conference on Networking, Architecture, and Storage*. IEEE, 2012, pp. 268–272.

[19] A. Hankin, T. Shapira, K. Sangaiah, M. Lui, and M. Hempstead, "Evaluation of non-volatile memory based last level cache given modern use case behavior," in *2019 IEEE International Symposium on Workload Characterization (IISWC)*, 2019.

[20] A. Hankin, D. Werner, M. Amiraski, J. Sebot, K. Vaidyanathan, and M. Hempstead, "Hotgauge: A methodology for characterizing advanced hotspots in modern and next generation processors," in *2021 IEEE International Symposium on Workload Characterization (IISWC)*, 2021, pp. 163–175.

[21] W. Henkels, N. Lu, W. Hwang, T. Rajeevakumar, R. Franch, K. Jenkins, T. Bucelot, D. Heidel, and M. Immediato, "A 12-ns low-temperature dram," *IEEE Transactions on Electron Devices*, vol. 36, no. 8, pp. 1414–1422, 1989.

[22] C. Y. Ho, R. W. Powell, and P. E. Liley, "Thermal conductivity of the elements," *Journal of Physical and Chemical Reference Data*, vol. 1, no. 2, pp. 279–421, 1972.

[23] A. F. Inci, M. M. Isgenc, and D. Marculescu, "Deepnvm: A framework for modeling and analysis of non-volatile memory technologies for deep learning applications," in *Proceedings of the 23rd Conference on Design, Automation and Test in Europe*, ser. DATE '20. San Jose, CA, USA: EDA Consortium, 2020, p. 1295–1298.

[24] K. Ishida, I. Byun, I. Nagaoka, K. Fukumitsu, M. Tanaka, S. Kawakami, T. Tanimoto, T. Ono, J. Kim, and K. Inoue, "Supernpu: An extremely fast neural processing unit using superconducting logic devices," in *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2020, pp. 58–72.

[25] T. Kelly, P. Fernandez, T. Vogelsang, S. McKee, L. Gopalakrishnan, S. Magee, K. Padgett, D. Barrow, J. Rizza, D. Doidge, K. Wright, C. Hampel, and G. Bronner, "Some like it cold: Initial testing results for cryogenic computing components," *Journal of Physics: Conference Series*, vol. 1182, p. 012004, 02 2019.

[26] N. S. Kim, T. Austin, D. Baauw, T. Mudge, K. Flautner, J. S. Hu, M. J. Irwin, M. Kandemir, and V. Narayanan, "Leakage current: Moore's law meets static power," *computer*, vol. 36, no. 12, pp. 68–75, 2003.

[27] K. Korgaonkar, I. Bhati, H. Liu, J. Gaur, S. Manipatruni, S. Subramoney, T. Karnik, S. Swanson, I. Young, and H. Wang, "Density tradeoffs of non-volatile memory as a replacement for sram based last level cache," in *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2018, pp. 315–327.

[28] G.-h. Lee, D. Min, I. Byun, and J. Kim, "Cryogenic computer architecture modeling with memory-side case studies," in *Proceedings of the 46th International Symposium on Computer Architecture*, ser. ISCA '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 774–787. [Online]. Available: https://doi.org/10.1145/3307650.3322219

[29] G.-H. Lee, S. Na, I. Byun, D. Min, and J. Kim, "Cryoguard: A near refresh-free robust dram design for cryogenic computing," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2021, pp. 637–650.

[30] K. Likharev and V. Semenov, "Rsfq logic/memory family: a new josephson-junction technology for sub-terahertz-clock-frequency digital systems," *IEEE Transactions on Applied Superconductivity*, 1991.

[31] R. A. Matula, "Electrical resistivity of copper, gold, palladium, and silver," *Journal of Physical and Chemical Reference Data*, vol. 8, no. 4, pp. 1147–1298, 1979.

[32] D. Min, I. Byun, G.-H. Lee, S. Na, and J. Kim, *CryoCache: A Fast, Large, and Cost-Effective Cache Architecture for Cryogenic Computing*, 2020.

[33] D. Min, Y. Chung, I. Byun, J. Kim, and J. Kim, "Cryowire: wire-driven microarchitecture designs for cryogenic computing," in *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, 2022, pp. 903–917.

[34] S. Mittal, M. Poremba, J. Vetter, and Y. Xie, "Exploring design space of 3d nvm and edram caches using destiny tool," *Oak Ridge National Laboratory, USA, Tech. Rep. ORNL/TM-2014/636*, 2014.

[35] I. Nagaoka, M. Tanaka, K. Inoue, and A. Fujimaki, "29.3 a 48ghz 5.6mw gate-level-pipelined multiplier using single-flux quantum logic," in *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2019.

[36] L. Pentecost, M. Donato, B. Reagen, U. Gupta, S. Ma, G.-Y. Wei, and D. Brooks, "Maxnvm: Maximizing dnn storage density and inference efficiency with sparse encoding and error mitigation," in *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture*, ser. MICRO '52, 2019.

[37] L. Pentecost, A. Hankin, M. Donato, M. Hempstead, G.-Y. Wei, and D. Brooks, "Nvmexplorer: A framework for cross-stack comparisons of embedded non-volatile memories," in *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, 2022, pp. 938–956.

[38] R. Pires, R. Dickstein, S. Titcomb, and R. Anderson, "Carrier freezeout in silicon," *Cryogenics*, 1990.

[39] M. Poremba, S. Mittal, D. Li, J. S. Vetter, and Y. Xie, "Destiny: A tool for modeling emerging 3d nvm and edram caches," in *2015 Design, Automation Test in Europe Conference Exhibition (DATE)*, 2015, pp. 1543–1546.

[40] D. Prasad, M. Vangala, M. Bhargava, A. Beckers, A. Grill, D. Tierno, K. Nathella, T. Achuthan, D. Pietromonaco, J. Myers, M. Walker, B. Parvais, and B. Cline, "Cryo-computing for infrastructure applications: A technology-to-microarchitecture co-optimization study," 12 2022, pp. 23.5.1–23.5.4.

[41] R. Saligram, D. Prasad, D. Pietromonaco, A. Raychowdhury, and B. Cline, "A 64-bit arm cpu at cryogenic temperatures: Design technology co-optimization for power and performance," in *2021 IEEE Custom Integrated Circuits Conference (CICC)*, 2021, pp. 1–2.

[42] T. Sanuki, Y. Aiba, H. Tanaka, T. Maeda, K. Sawa, F. Kikushima, and M. Miura, "Cryogenic operation of 3-d flash memory for storage performance improvement and bit cost scaling," *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits*, vol. 7, no. 2, pp. 159–167, 2021.

[43] O. Semenov, A. Vassighi, and M. Sachdev, "Impact of technology scaling on thermal behavior of leakage current in sub-quarter micron mosfets: perspective of low temperature current testing," *Microelectronics Journal*, 2002.

[44] N. Takeuchi, K. Ehara, K. Inoue, Y. Yamanashi, and N. Yoshikawa, "Margin and energy dissipation of adiabatic quantum-flux-parametron logic at finite temperature," *IEEE Transactions on Applied Superconductivity*, 2013.

[45] N. Takeuchi, D. Ozawa, Y. Yamanashi, and N. Yoshikawa, "An adiabatic quantum flux parametron as an ultra-low-power logic device," *Superconductor Science and Technology*, 2013.

[46] S. S. Tannu, D. M. Carmean, and M. K. Qureshi, "Cryogenic-dram based memory system for scalable quantum computers: A feasibility study," in *Proceedings of the International Symposium on Memory Systems*, ser. MEMSYS '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 189–195. [Online]. Available: https://doi.org/10.1145/3132402.3132436

[47] H. J. ter Brake and G. Wiegerinck, "Low-power cryocooler survey," *Cryogenics*, vol. 42, no. 11, pp. 705–718, 2002.

[48] F. Wang, T. Vogelsang, B. Haukness, and S. C. Magee, "Dram retention at cryogenic temperatures," in *2018 IEEE International Memory Workshop (IMW)*, 2018, pp. 1–4.

[49] J. Wang, X. Dong, Y. Xie, and N. P. Jouppi, "i 2 wap: Improving non-volatile cache lifetime by reducing inter-and intra-set write variations," in *2013 IEEE 19th International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2013, pp. 234–245.

[50] Z. Wang, D. A. Jiménez, C. Xu, G. Sun, and Y. Xie, "Adaptive placement and migration policy for an stt-ram-based hybrid cache," in *2014 IEEE 20th International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 2014, pp. 13–24.

[51] F. Ware, L. Gopalakrishnan, E. Linstadt, S. A. McKee, T. Vogelsang, K. L. Wright, C. Hampel, and G. Bronner, "Do superconducting processors really need cryogenic memories? the case for cold dram," in *Proceedings of the International Symposium on Memory Systems*, ser. MEMSYS '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 183–188. [Online]. Available: https://doi.org/10.1145/3132402.3132424

[52] S. Wilton and N. Jouppi, "Cacti: an enhanced cache access and cycle time model," *IEEE Journal of Solid-State Circuits*, 1996.

[53] X. Wu, J. Li, L. Zhang, E. Speight, R. Rajamony, and Y. Xie, "Hybrid cache architecture with disparate memory technologies," *ACM SIGARCH computer architecture news*, vol. 37, no. 3, pp. 34–45, 2009.

[54] Y. Xie, "Modeling, architecture, and applications for emerging memory technologies," *IEEE Design Test of Computers*, 2011.

[55] N. Yoshikawa, D. Ozawa, and Y. Yamanashi, "Ultra-low-power superconducting logic devices using adiabatic quantum flux parametron," *The Japan Society of Applied Physics*, 2011.

[56] I. Yukikazu, *Case studies in superconducting magnets design and operational issues*. Springer Verlag, 2010.

[57] Y. Zhang, D. Parikh, K. Sankaranarayanan, K. Skadron, and M. Stan, "Hotleakage: A temperature-aware model of subthreshold and gate leakage for architects," 07 2003.

[58] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45 nm early design exploration," *IEEE Transactions on electron Devices*, vol. 53, no. 11, pp. 2816–2823, 2006.