CryoCache: A Fast, Large, and Cost-Effective Cache Architecture for Cryogenic Computing

Dongmoon Min
Department of Electrical and Computer Engineering
Seoul National University
dongmoon.min@snu.ac.kr

Ilkwon Byun
Department of Electrical and Computer Engineering
Seoul National University
ik.byun@snu.ac.kr

Gyu-Hyeon Lee
Department of Electrical and Computer Engineering
Seoul National University
guhylee@snu.ac.kr

Seongmin Na
Department of Electrical and Computer Engineering
Seoul National University
seongmin.na@snu.ac.kr

Jangwoo Kim*
Department of Electrical and Computer Engineering
Seoul National University
jangwoo@snu.ac.kr

Abstract

Cryogenic computing, which is to run a computer at extremely low temperatures (e.g., 77K), is a highly promising solution to dramatically improve the computer's performance and power efficiency thanks to the significantly reduced leakage power and wire resistance. However, computer architects are facing fundamental challenges in developing and deploying cryogenic-optimal architectural units due to the lack of understanding about its cost-effectiveness and feasibility (e.g., device and cooling costs vs. speedup, energy and area saving) and thus how to architect such cryogenic-optimal units.

In this paper, we propose *CryoCache*, a cost-effective, technology-feasible cryogenic-optimal cache architecture running at 77K. For this goal, we first thoroughly analyze the cost-effectiveness and feasibility of various on-chip memory cell technologies running at 77K. Based on the analysis, we architect cryogenic-optimal caches with conventional technology-feasible 6T-SRAM and 3T-eDRAM cells whose performance, area, and power benefits at 77K clearly outweigh their cooling costs. Our evaluations show that our example CryoCache architecture achieves 2× faster cache access and 2× larger capacity compared to conventional caches running at the room temperature. To the best of our knowledge, this is the first work to propose a fast, large, and cost-effective cache architecture which can be applied to cryogenic computing.

CCS Concepts. • **Computer systems organization** → **Architectures**; • **Hardware** → **Modeling and parameter extraction**; **Semiconductor memory**; **Power estimation and optimization**; **Analysis and design of emerging devices and systems**; **Memory and dense storage**.

Keywords. Cryogenic computing; Cryogenic cache; Technology comparison and analysis; Modeling; Simulation

ACM Reference Format:
Dongmoon Min, Ilkwon Byun, Gyu-Hyeon Lee, Seongmin Na, and Jangwoo Kim. 2020. CryoCache: A Fast, Large, and Cost-Effective Cache Architecture for Cryogenic Computing. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '20), March 16–20, 2020, Lausanne, Switzerland.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3373376.3378513

1 Introduction

Cryogenic computing, which is to run a computer device at extremely low temperatures, has been considered as a highly promising solution to improve the computer device's performance and power efficiency thanks to the devices' dramatic reduction of leakage power and resistance. In particular, architects have been making efforts to take the best advantage of the two key benefits of cryogenic computing to resolve the memory walls from the perspective of the memory access latency and energy efficiency.

To apply the cryogenic computing to resolve memory walls, a recent study [29] developed a validated architecture modeling tool to accurately estimate the performance and power efficiency of DRAM devices at 77K (-196 °C), and proposed cryogenic-optimal DRAM devices whose performance and power efficiency are significantly higher than those of conventional DRAM devices at the room temperature, 300K (27 °C).

*Corresponding author.

ASPLOS '20, March 16–20, 2020, Lausanne, Switzerland
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7102-5/20/03...$15.00
https://doi.org/10.1145/3373376.3378513

Therefore, it is a straightforward next step to apply the cryogenic computing to on-chip caches to further contribute to resolving on-chip memory walls. In fact, as on-chip caches are more frequently accessed by computing cores than the off-chip DRAM, improving the capacity and latency of caches is more important than those of DRAMs. However, it raises fundamental challenges to construct cryogenic-optimal caches because their advantages and disadvantages, feasibility, and applicability have not been actively explored.

To resolve the challenges, architects should first determine the most appropriate cache-cell technology for the target temperature by analyzing the candidate cell's size and access latency, dynamic and static energy efficiency, and cooling cost. In particular, the cost to maintain the target low temperature should be carefully analyzed to avoid misleading messages regarding cryogenic caches. Next, with the cryogenic-optimal cell technology determined, architects should construct a cryogenic-optimal cache architecture whose overall advantages clearly outperform the device and cooling costs.

In this paper, we propose *CryoCache*, a cost-effective, technology-feasible cryogenic-optimal cache architecture running at 77K. Our example cryogenic cache roughly doubles the cache capacity and performance of conventional SRAM-based caches, while its overall power consumption including the cooling cost decreases significantly.

To achieve this goal, we first perform a thorough analysis to estimate the performance, energy consumption and per-bit area of major cache cell technologies (i.e., 6T-SRAM, 3T-eDRAM, 1T1C-eDRAM, and STT-RAM cells), and then choose technology-feasible 6T-SRAM and 3T-eDRAM cells as highly promising candidates to build cryogenic caches.

Second, we show that the latency and power consumption of SRAM caches can be significantly reduced, and the degree of reduction increases with the cache's physical size. For example, the access latency of an 8MB L3 cache reduces from 42 cycles (the room temperature) to 18 cycles (77K) with its leakage power nearly eliminated.

Third, we show that we can double the capacity of a conventional SRAM cache by replacing its 6T-SRAM cells with technology-feasible, roughly half-sized 3T-eDRAM cells. At the room temperature, we cannot build a cache with 3T-eDRAM cells due to their extremely short retention time (i.e., 2.5μs at 300K) and thus frequent cell refreshes. However, at 77K, as the retention period dramatically increases by 10,000 times (i.e., >30ms at 77K) thanks to the leakage current reduction, we can build nearly refresh-free 3T-eDRAM caches.

Fourth, we show that the cooling cost quickly increases with the cache's power consumption at 77K. The power consumption at 77K comes mainly from the dynamic power consumption. Therefore, we minimize the cache's dynamic power by reducing their supply and threshold voltages. Even though the voltages cannot be scaled down at the room temperature due to the severely increased leakage power consumption, we can safely reduce the voltages at 77K thanks to the nearly-eliminated leakage power. We further reduce the cooling cost by constructing the L2 and L3 caches with 3T-eDRAM cells made of only static-power negligible PMOS transistors.

Finally, we propose and evaluate our example CryoCache architecture for 77K. The example cryogenic cache consists of 6T-SRAM cell-based L1 caches and 3T-eDRAM cell-based L2 and L3 caches. The proposed cache architecture effectively resolves critical on-chip memory walls as follows. By replacing 6T-SRAM cells in L2 and L3 caches with 3T-eDRAM cells, CryoCache increases the L2 cache capacity from 256KB to 512KB and the L3 cache capacity from 8MB to 16MB. At the same time, the reduced temperature decreases the L1, L2, and L3 cache access latencies down to 50%, 67%, and 50%, respectively. To compensate for the cache-cooling cost, we decrease the cache's supply and threshold voltages which can be safely done at the low temperature.

For evaluation, we measure the performance of a modern high-performance processor (i.e., Intel i7 6700) equipped with our example CryoCache architecture. The results show that CryoCache improves the performance of 11 PARSEC workloads by 80% on average (and up to 4.14 times), thanks to the increased cache-access performance and capacity. In addition, CryoCache reduces its overall power consumption by 34.1% even with the cooling cost considered, thanks to the cryogenic-optimal voltage scaling.

In summary, this paper makes the following contributions:

- **Analysis and findings**. We carefully analyze the size, performance, power consumption, and feasibility of various on-chip memory cells at 77K. Based on the analysis, we find 6T-SRAM and 3T-eDRAM cells as the promising candidates to build cryogenic-optimal caches.
- **Cryogenic cache design**. To realize the findings, we develop a novel cryogenic cache architecture consisting of 6T-SRAM cell-based L1 caches, and 3T-eDRAM cell-based L2 and L3 caches.
- **Capacity and performance improvements**. CryoCache roughly doubles the performance of L1, L2, and L3 caches at 77K, while doubling the capacity of L2 and L3 caches. It leads to the 80% average performance improvement for 11 PARSEC workloads.
- **Cost and power reduction**. CryoCache's dynamic power reduction with voltage scaling achieves the 34.1% overall power reduction even including the cooling cost.

To the best of our knowledge, this is the first work to propose a fast, large, technology-feasible, and cost-effective cache architecture which can be applied to cryogenic computing.

2 Background & Motivation

2.1 Need for fast and large caches

Computer architects are facing critical challenges in developing next-generation caches: latency, capacity, and power consumption. In the past, with both Moore's Law and Dennard scaling satisfied, architects could increase both the capacity and performance of caches while maintaining the power budget by making SRAM cells with smaller transistors and reducing their supply and threshold voltages (V_{dd} and V_{th}).

However, it is becoming extremely difficult to increase the cache capacity or performance as neither Moore's Law nor Dennard scaling is valid anymore. As a result, we cannot make SRAM cells smaller nor faster.

Cache performance challenge. The cache performance problem is getting worse as a processor's performance gets quickly bounded by its cache performance due to the relatively slower instruction and data reads than the computing speed. The performance gap between computing and caching will keep increasing because the cache access latency is dominated by its wire latency, which is difficult to be reduced for smaller technology nodes [39].

Cache capacity and power challenges. The cache capacity problem is also getting worse as modern processors deploy an increasing number of cores which require an increasing capacity of caches on a chip (e.g., many L2 caches, larger L3 cache). In addition, it is not effective either to increase the die size for increasing the cache capacity due to the inhibitive increase of chip cost and power requirement.

Therefore, architects are in dire need of faster and larger caches, while satisfying the die area and power consumption budget.

2.2 Potentials of cryogenic computing

Cryogenic computing, which means computing at extremely low temperatures, has emerged as a breakthrough for resolving the performance and power challenges. The major benefits of cryogenic computing are the dramatically reduced leakage current and wire resistivity. First, thanks to the almost eliminated leakage current at low temperatures [48, 64], we can increase the clock frequency while achieving much lower power consumption (by reducing both V_{dd} and V_{th}). Second, as the wire resistivity linearly decreases with the temperature, we can correspondingly reduce the wire latency and thus build much faster memory devices. For example, the copper's resistivity at 77K is six times lower than the resistivity at 300K [37].

Regarding the target temperature, cryogenic computing focuses on two representative ultra-low temperatures, 77K and 4K. 77K computing, which can be easily achieved by applying Liquid Nitrogen (LN), can utilize CMOS technology reliably operating with higher performance at 77K. On the other hand, CMOS technology is unsuitable for 4K computing due to the freeze-out effect of CMOS [45] and much higher cooling cost [24]. For that reason, the 4K cryogenic computing is only feasible with unconventional superconducting-based devices (e.g., RSFQ [7, 32, 41], AQFP [52, 53, 63]). Therefore, computer architects have focused more on 77K-based cryogenic computing than 4K-based computing.

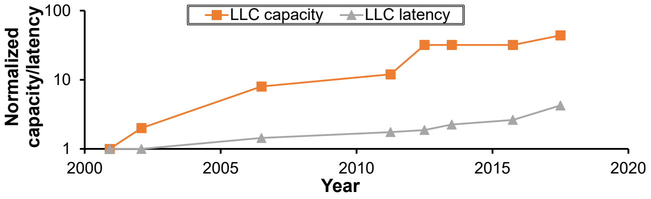


Figure 1. LLC latency and capacity of CPUs over generations

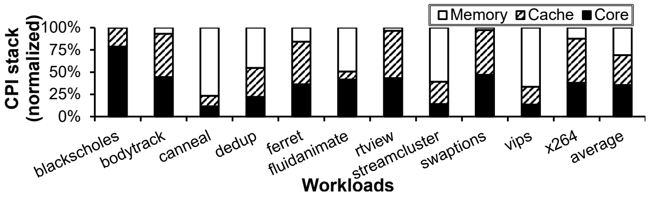


Figure 2. Normalized CPI stacks of PARSEC 2.1 workloads

Previous works explored the potentials of 77K computing, especially focusing on DRAM devices [19, 29, 54, 56, 59]. Among them, a recent study [29] developed a modeling tool to estimate the performance and power consumption of 77K-based DRAM devices and used the tool to propose 77K-optimal DRAM devices.

2.3 Focusing on cryogenic caches at 77K

In this work, we aim to improve the latency and capacity of on-chip caches under the same power and die budget by applying the 77K-based cryogenic computing.

To computer architects, it is more challenging, but also more important to improve the performance and capacity of on-chip caches than those of DRAMs. First, increasing the cache capacity under the same die and power budget significantly improves both the single-thread performance and multi-thread throughput. For example, Fig. 1 shows the access latency and capacity of Last-Level Caches (LLC) over generations, normalized to those of Pentium 4 (180 nm) in early 2000 [1]. This figure clearly indicates that both the latency and capacity of LLC have been significantly increased over generations even though they might not still meet the desires.

Second, as caches are frequently accessed by computing cores, reducing their access latency will also significantly improve the processor's overall performance. Fig. 2 shows the normalized CPI stacks of PARSEC 2.1 workloads [4] obtained by our gem5 simulations [5]. The figure clearly indicates that the cache performance significantly contributes to the modern application performance.

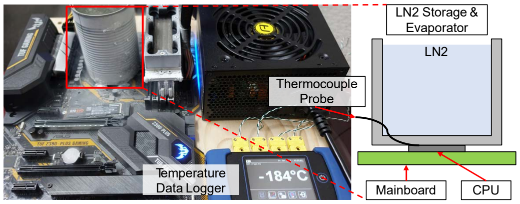


Figure 3. Our setup to run the whole processor at 77K

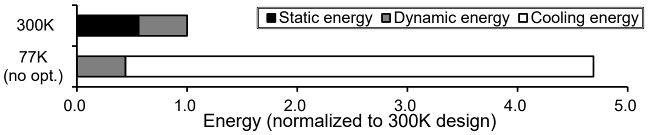


Figure 4. Total required energy of caches with 77K cooling

Cryogenic computing can be highly promising to resolve these performance, capacity, and power issues. To quickly estimate its performance benefit, we reduced the temperature of an Intel i7 8700K processor to 77K by applying Liquid Nitrogen to our test board which allows the frequency adjustment and cache-latency measurement, as shown in Fig. 3. From this experiment, we observe that the on-chip caches can run faster by 20% at 77K. This result matches our modeling result shown in later sections (32KB L1 speed-up in Fig. 13b).

In fact, we can further improve the cryogenic-cache's performance by applying cryogenic-friendly cache cell technologies and temperature-optimal cache architecture designs, which are described in the following sections.

2.4 Challenges of cryogenic caches

Even with the cryogenic cache's promising aspects, there exist several critical challenges to be resolved as follows.

Cryogenic-optimal cell technology. Architects should determine the most appropriate cache-cell technology for the target temperature. Researchers have explored various memory cell technologies and proposed caches with different tradeoffs (e.g., 6T-SRAM, 3T-eDRAM, 1T1C-eDRAM, STT-RAM cells [18, 30, 35, 57]). However, all these works assumed operations at the room temperature (300K), which make their trade-off analyses invalid for 77K operations. Therefore, architects are in dire need of analyzing the candidate cell's size, access latency, and dynamic and static energy efficiency to build cryogenic-optimal caches with the cells and run for 77K.

Cooling cost analysis and compensation. Architects must carefully analyze the cooling cost and propose a way to compensate for the cost. In fact, many works to propose cryogenic computing have overlooked the cooling cost, which leads incorrect cost-effectiveness. However, the cooling cost can be severe enough to make the advantages of cryogenic computing ineffective. For example, to maintain the device temperature at 77K, we should apply 9.65 times higher energy than the energy consumed by the cooled device [24, 29]. Fig. 4 shows the severely increased cooling power consumption (driven by the dynamic energy at 77K) for running *swaptions* from PARSEC workloads. To compensate for this cost, cryogenic caches should consume only 10% of the energy consumed by caches running at 300K.

Cryogenic-optimal cache architecture. Once the accurate tradeoffs of candidate cache cell technologies are available for 77K, architects should find the best cache architecture. The cryogenic-optimal cache architecture should provide the highest speed to the latency-critical workloads and the largest capacity to the capacity-critical workloads, while keeping their overall die area and power consumption under the budget.

To resolve these challenges, we carefully analyze and select the most appropriate cache-cell technologies for the target temperature in terms of performance, power, cost, and feasibility. Next, by exploiting the analyses, we architect and propose our cryogenic-optimal, technology-feasible cache design which achieves both the high performance and the energy efficiency, while satisfying the die-area and cooling cost budget.

3 Cell technologies for cryogenic caches

To determine the 77K-optimal memory technology, we analyze major cache-cell technologies (i.e., 6T-SRAM, 3T-eDRAM, 1T1C-eDRAM, STT-RAM) as summarized in Table 1. Our analysis focuses on two points: (1) each technology's tradeoffs and (2) how they are affected by the temperature reduction. We analyze the cell-level characteristics (e.g., cell density, retention time) in this section, and the cache-level characteristics (e.g., dynamic power, performance) in Section 5.

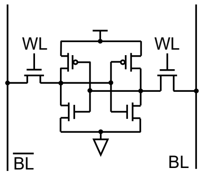
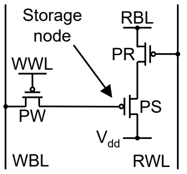
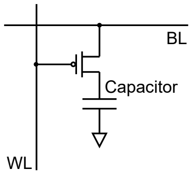
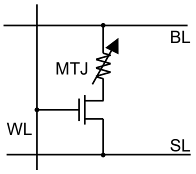
3.1 6T-SRAM

Behaviors at 300K. The 6T-SRAM cell is the conventional technology for cache designs at the room temperature. The main advantage of 6T-SRAM is its relatively faster access speed and more reliable, retention-free bit storage than other candidates. However, SRAM has several shortcomings in terms of the cell size and static power [11, 15, 46]. As each 6T-SRAM cell uses six transistors per bit, its cell size is larger than other candidates consisting of a smaller number of transistors. In addition, as each 6T-SRAM cell contains multiple leakage paths, it consumes a huge amount of static power.

Behaviors at 77K. The SRAM remains as a promising design choice for 77K caches. First, the SRAM's access latency decreases with the temperature reduction thanks to the reduced wire latency and the mobility improvement [49]. We confirm the latency reduction with our modeling results in Section 5.

Second, the SRAM's static power nearly disappears at 77K thanks to the greatly reduced subthreshold leakage current, which is the dominant source of leakage power consumption

Table 1. Comparison of memory technologies for on-chip caches

	(a) 6T-SRAM	(b) 3T-eDRAM	(c) 1T1C-eDRAM	(d) STT-RAM
Cell schematic		 Storage node RBL PR WWL PW PS Vdd WBL RWL		
Major advantage	Fast read/write	High density Logic compatible Small leakage Fast read/write	High density	High density Non-volatility Near-zero leakage
Critical drawback	High leakage power Large cell area	Short retention time	Extra process (Cap) Slow read/write High access energy	Extra process (MTJ) Write overhead
Cryogenic effect	(+) Faster speed (+) Near-zero leakage	(+) Faster speed (+) Improved retention time	(-) Cannot resolve the process problems	(-) Higher write overhead

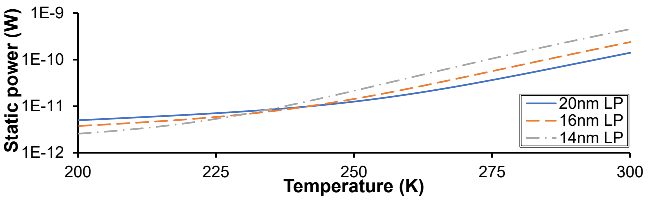


Figure 5. Static power of differently scaled SRAM cells

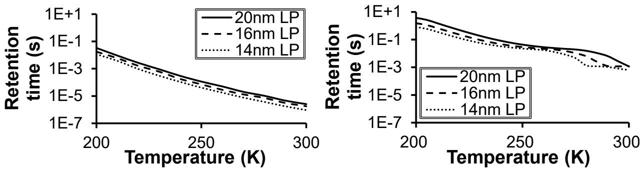
at 300K. Our simulations using Hspice and PTM models [66] (Fig. 5) show the static power of differently-scaled SRAM cells operating at different temperatures. The simulation limits the minimum temperature to 200K, the minimum temperature validated by PTM [65]. With the temperature reduction, the static power quickly disappears (e.g., 89.4 times reduction for 14nm at 200K) and its reduction degree is higher for the leakage-subject smaller technologies. At 200K, the static power of the 20nm node is higher than the smaller nodes by applying higher V_{dd} to the larger nodes and thus incurring relatively higher gate tunneling current [27].

Therefore, we consider SRAM to remain as a promising candidate to build a cryogenic cache.

3.2 3T-eDRAM

Behaviors at 300K. The 3T-eDRAM cell consists of three PMOS transistors: a write access transistor (PW), a storage transistor (PS), and a read access transistor (PR). Table 1b shows the 3T-eDRAM cell's key characteristics.

A 3T-eDRAM cell stores a bit value on PS's gate capacitance (or *storage node*). For a write, the write bitline (WBL) is pulled up to the desired voltage level, while the write wordline (WWL) drives PW to store the value to the storage node. For a read, the read bitline (RBL) is pulled down to the zero voltage, and then the read wordline (RWL) is switched from V_{dd} to 0V to activate PR. If '0' is stored in the storage node, the pre-discharged RBL is pulled up to V_{dd}. If '1' is stored, the RBL remains discharged. The sense amplifier quickly translates the stored value based on the RBL's voltage level.



(a) 3T-eDRAM **(b)** 1T1C-eDRAM

Figure 6. Retention time of (a) 3T-eDRAM and (b) 1T1C-eDRAM cells

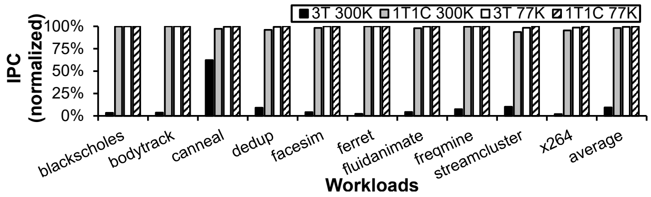


Figure 7. Performance impact of different eDRAM cells (3T, 1T1C) at different temperatures (300K, 77K). IPC values are normalized to IPC without refreshing

The main advantages of 3T-eDRAM are its seamless implementation on a logic die (i.e., logic compatibility), 2× higher cell density over the 6T-SRAM cell by using only three transistors per bit, fast access speed (even comparable to SRAM), and smaller static power consumption by using only low-leakage PMOS transistors [11, 15].

Cell refresh overhead. However, at 300K, the 3T-eDRAM cell is not feasible for a cache design due to its prohibitive refresh overhead. As the 3T-eDRAM cell's value is gradually leaked away, the cell should be refreshed. Fig. 6a shows the 3T-eDRAM's retention time with the technology and temperature variations. We obtain the results with Hspice Monte Carlo simulations as done by [14], and the overall trend of the prohibitive refresh overhead matches the results of [27]. For example, the 3T-eDRAM's retention time for the

14nm node is 927ns, which is almost 70,000 times shorter than that of DRAM (64ms).

Making a cache with 3T-eDRAM cells leads to severe performance degradation as shown in Fig. 7. We set the retention time of 3T-eDRAM to 2.5μs (20nm LP), the longest value at 300K. The graph compares the performance of a processor having 3T-eDRAM caches to a baseline having conventional 6T-SRAM caches (described in Table 2). The refresh operation of 3T-eDRAM cells at 300K unacceptably degrades the performance down to 6% on average. Such huge refresh overheads prevent modern processors from implementing 3T-eDRAM caches at 300K.

Behaviors at 77K. Interestingly, we observe that the cryogenic environment effectively eliminates the refresh overhead by dramatically extending the retention time. Even at 200K, the retention time is extended by more than 10,000 times thanks to the reduced leakage current (Fig. 6a). Note that the retention time will be further reduced for 77K due to the more reduction of the leakage currents (e.g., subthreshold current, GIDL, gate-tunneling current) below 200K [3, 48, 50, 64].

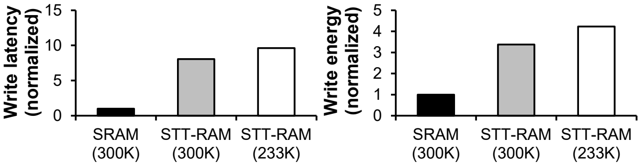
Therefore, making a 3T-eDRAM cache at 77K becomes highly promising thanks to the nearly eliminated refreshing overhead. To measure the application performance, we use the shortest retention time (11.5ms in 14nm LP) at 200K for conservatively applying the reduced refresh overhead. Fig. 7 shows that the 3T-eDRAM cache's performance becomes similar to that of SRAM cache under the cryogenic temperature. Based on its promising behaviors at 77K (e.g., doubled density, faster access speed, lower power, longer retention), we choose the 3T-eDRAM as another promising candidate to build a cryogenic cache.

3.3 1T1C-eDRAM

Behaviors at 300K. Each 1T1C-eDRAM cell consists of an access transistor and a capacitor (Table 1c), which makes its cell density roughly three times higher than the 6T-SRAM cell (i.e., 2.85 times [12]). Another advantage is its reasonable refresh overhead even at 300K. As the capacitor of 1T1C-eDRAM is much larger than the storage node of 3T-eDRAM, its retention time at 300K is 100 times longer than 3T-eDRAM (Fig. 6). The performance degradation due to the refresh overhead is acceptable (i.e., 2.2%) at 300K as shown in Fig. 7.

However, the 1T1C-eDRAM cell suffers from fundamental limitations as follows. First of all, its fabrication is incompatible with the conventional transistor-only logic process due to its per-cell capacitor. In addition, the 1T1C-eDRAM cell is slower and consumes more access energy than SRAM and 3T-eDRAM [61, 62]. Therefore, 1T1C-eDRAM has been used to build extremely large, but slow off-chip caches (e.g., 128MB off-chip cache of IBM Power 8).

Behaviors at 77K. Unfortunately, the temperature reduction does not resolve the 1T1C-eDRAM's key disadvantages, as the eDRAM's main advantage at a low temperature is the reduced refreshing overhead. Fig. 6b shows that the 1T1C-eDRAM's retention time at 300K is already as long as the refresh-tolerable 77K 3T-eDRAM's retention time. Therefore, the application performance of 77K 1T1C-eDRAM caches is the same as those of 77K 3T-eDRAM and 300K 6T-SRAM caches (Fig. 7).



(a) Write latency **(b)** Write energy

Figure 8. Write overhead of STT-RAM at 300K and 233K.

Due to its characteristics inferior to the 77K 3T-eDRAM cells (i.e., logic incompatibility, slower access, higher energy), we exclude the 1T1C-eDRAM cell as a candidate to build a cryogenic cache.

3.4 STT-RAM

Behaviors at 300K. The STT-RAM cell is an emerging memory cell technology thanks to its high density (i.e., 2.94 times higher than SRAM), near-zero leakage, and non-volatility [16]. Table 1d shows the STT-RAM's structure. Each STT-RAM cell consists of one transistor and one magnetic tunneling junction (MTJ). The MTJ consists of two magnetic layers, whose polarity determines its resistance. Applying a high voltage to MTJ changes the polarity and thus changes the stored data.

However, the STT-RAM cell comes with two critical limitations. First, to build the MTJ, the cell implementation requires additional fabrication process. Second, it suffers from severe write overhead. To write a value, it should apply a high voltage to MTJ for a long time enough to change the layer's magnetic polarity. Fig. 8 shows the write latency and energy of 22nm 128KB STT-RAM at 300K and 233K. We used NVSim [17] to obtain the 300K STT-RAM values, and scaled the values for 233K according to [10]. The write overhead is normalized to that of 22nm 128KB SRAM values obtained with CACTI [40]. The results indicate that the STT-RAM's write latency is 8.1 times longer and its energy is 3.4 times higher than those of the SRAM baseline.

Behaviors at 77K. Unfortunately, the temperature reduction increases the STT-RAM's write overhead. Fig. 8 shows that the write latency and energy overheads increase with the temperature reduction. The reason is the MTJ's increased thermal stability which makes the polarity change more difficult at the low temperature [60]. This write overhead will further increase at lower temperatures as the thermal stability is inversely proportional to the temperature [25].

Therefore, due to its increasing write overhead at low temperatures, we exclude the STT-RAM cell as a candidate to build a cryogenic cache.

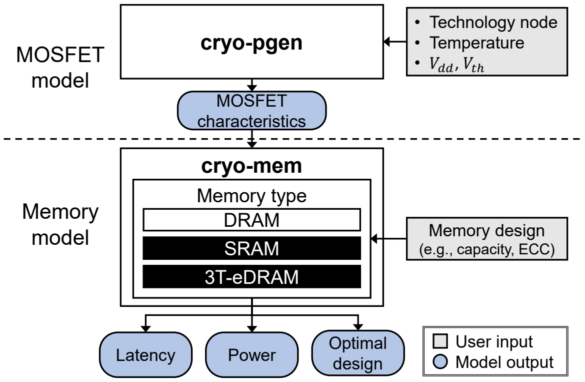


Figure 9. Our cryogenic cache modeling tool based on [29]

4 Cryogenic cache modeling and analysis

In the previous section, we chose 6T-SRAM and 3T-eDRAM cells as the promising candidates to be used for cryogenic caches. Therefore, we develop a cryogenic cache model in this section, in order to accurately estimate the latency and energy consumption of the two candidate cells at 77K.

To measure the access latency and power consumption, we modify CryoRAM [29], a state-of-the-art cryogenic memory modeling tool, to implement the representative 6T-SRAM and 3T-eDRAM caches. CryoRAM consists of the cryogenic MOSFET model (cryo-pgen) and DRAM-device memory model (cryo-mem). We add 6T-SRAM and 3T-eDRAM cache models to the tool's cryo-mem component. We also modify CryoRAM to estimate the latency and power consumption of the two new memory cells with the cryogenic MOSFET properties obtained by cryo-pgen. Fig. 9 shows our modified modeling methodology with the newly added memory models marked as the black-colored components.

4.1 300K cache modeling

We first develop our 6T-SRAM cache model for 300K by applying the CACTI's SRAM model to our cryogenic modeling tool. As we do not find any 3T-eDRAM cache models available in public [11, 26], we develop our own 3T-eDRAM cache model by modifying the SRAM cache model as follows.

(1) Decoder. The 3T-eDRAM cache's decoder can be modeled from the SRAM cache model by considering their differences in the cell structure. For example, the SRAM cache uses one output port per cell because read and write operations share the same wordlines. On the other hand, the 3T-eDRAM cache uses two output ports per cell because read and write operations use different wordlines (Table 1b). The higher number of output ports increases the number of transistors in the decoder and thus makes the decoder slower. We take the differences into account to model our 3T-eDRAM cache model. Fig. 10a compares the decoder structures of SRAM and 3T-eDRAM.

(2) Cell size. As the cell size directly affects various key physical structures (e.g., Htree, decoder, bitline, wordline) and thus the cache's performance and energy consumption, we carefully estimate the size of 3T-eDRAM cell. For the purpose, we derive its relative size to 6T-SRAM by drawing and comparing both cell layouts with Magic [42] (Fig. 10b). W and H in Fig. 10b indicate the width and height of an SRAM cell, respectively. Our result shows that the 3T-eDRAM cell is 2.13 times smaller than the 6T-SRAM cell. The smaller cell size reduces the size of decoder and the length of wordlines.

(3) Bitline RC model. As the bitline RC model determines the bitline latency, we carefully extract the 3T-eDRAM's bitline RC model from the SRAM model by changing NMOS resistance (R_{nmos}) to PMOS resistance (R_{pmos}) (Fig. 10c). The bitline RC model of SRAM consists of two R_{nmos} because two serialized NMOS transistors drive the bitline. On the other hand, 3T-eDRAM charges the bitline with two serialized PMOS transistors. Note that R_{pmos} is higher than R_{nmos} due to the lower mobility of PMOS [23]. We apply the differences to our model.

(4) Sense amplifier. The 3T-eDRAM's sense amplifier differs from that of SRAM. However, the latency and energy consumption of the sense amplifier are negligible compared with those of the decoder, bitlines, and other peripheral circuits [26]. Therefore, we apply the SRAM cache's sense amplifier model to our 3T-eDRAM cache model.

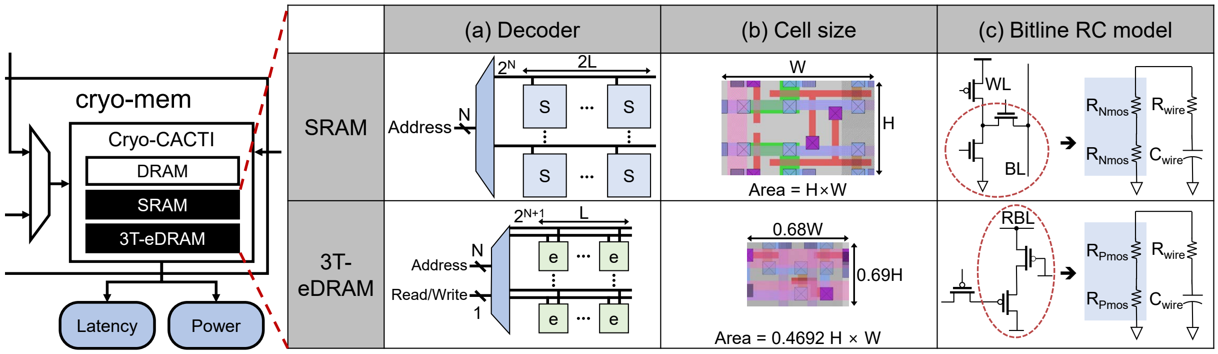


Figure 10. SRAM and 3T-eDRAM cache modeling overview

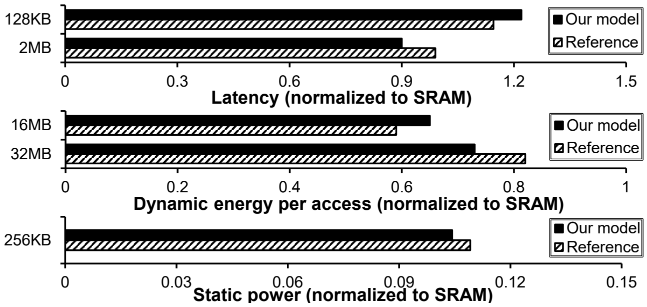


Figure 11. 300K 3T-eDRAM model validation results

4.2 300K cache model validation

To validate the 3T-eDRAM model's latency and static power, we compare our model results against the publicly available reference values obtained from 65nm fabricated chips [14]. Fig. 11 shows our validation results, in which all results are normalized to those of the same capacity SRAM. To validate the 3T-eDRAM's dynamic energy per access values, we compare our model results against the publicly available reference values obtained from the 32nm process modeling [11].

The validation results show that the latency, static power, and the dynamic energy of our model closely match the reference results with 8.4% difference on average. Therefore, we conclude that our 3T-eDRAM cache modeling is reasonably validated for the room-temperature operations. Note that we verify only the relative ratios between 3T-eDRAM and SRAM rather than the absolute values in terms of latency and energy consumption because we only utilize the relative values in the following sections.

4.3 Cryogenic environment modeling

We expect that the latency and power consumption of SRAM and 3T-eDRAM caches will be significantly lower at 77K due to the reduced wire resistivity and subthreshold current. For instance, the wire resistivity is reduced to 17.5% with the temperature reduction from 300K to 77K [37], which will improve the performance of the caches. At the same time, the nearly eliminated subthreshold current allows aggressive V_{dd} and V_{th} scaling (or *V_{dd}/V_{th} scaling*), which will greatly improve the energy efficiency of the caches as well. Therefore, to develop our cryogenic cache model, we apply the CryoRAM's low-temperature MOSFET model (cryo-pgen) [29] which can accurately estimate the wire resistivity, the leakage current, and the impact of $V_{\mathrm{dd}}/V_{\mathrm{th}}$ scaling.

4.4 77K cache model validation

To validate our cryogenic cache model, we compare the model's prediction with the results of Hspice simulations. For the Hspice simulations, we utilize an industry-provided MOSFET model card designed for the 65nm technology at 77K. Note that we evaluate the speed-up of 77K caches which have the same circuit design as 300K-optimized caches.

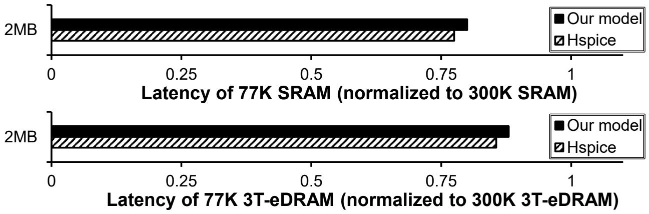


Figure 12. 77K cache model validation results

We do not additionally validate the static and dynamic energy model for the cryogenic caches due to the following reasons. First, the temperature model for the cache's static energy is the same as the already validated DRAM's model because the temperature dependence of the subthreshold leakage current does not depend on the memory cell type. In addition, regardless of the target temperature, the dynamic energy per access remains the same because the dynamic energy only depends on the supply voltage and capacitance of the circuit.

Fig. 12 shows the validation results of 2MB 77K caches, in which all the 77K latency values are normalized to the latency of the same caches operating at 300K. Based on our modeling, the SRAM and 3T-eDRAM caches become 20% and 12% faster at 77K, respectively. The speed-up values closely match the Hspice simulation results with 2.4% of the maximum error rate. Therefore, we conclude that the accuracy of our cryogenic model is well validated.

With our validated cache model, we perform aggressive design-space explorations to find our optimal cache architecture at 77K. The following section provides more details to find the optimal cryogenic cache architecture.

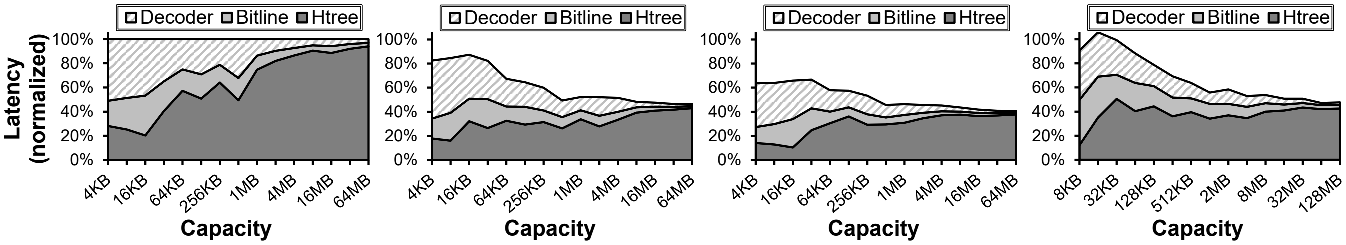
5 CryoCache: 77K-optimal cache design

With the cryogenic cache model described in the previous section, we propose an optimal cache architecture for the cryogenic environment. To achieve the goal, we first perform exhaustive experiments to figure out the optimal V_{dd} and V_{th} values to compensate for the cooling cost (Section 5.1). Next, we analyze the SRAM and 3T-eDRAM-based cryogenic caches in terms of the latency (Section 5.2) and the energy consumption (Section 5.3). Based on the analysis, we propose an optimal cache hierarchy by selectively using different cache configurations for different levels (Section 5.4).

For a fair comparison, we compare 3T-eDRAM and SRAM caches which occupy the same die area. For example, as 3T-eDRAM is twice denser than SRAM, we compare 16MB 3T-eDRAM and 8MB SRAM caches.

5.1 V_{dd} and V_{th} scaling

Our baseline cache design is an 8-way set-associative, dual-port, and ECC-supported SRAM cache fabricated with 22nm



(a) 300K SRAM caches **(b)** 77K SRAM (no opt.) caches **(c)** 77K SRAM (opt.) caches **(d)** 77K 3T-eDRAM (opt.) caches

Figure 13. Latency breakdown of (a) 300K SRAM, (b) 77K SRAM (no opt.), (c) 77K SRAM (opt.), (d) 77K 3T-eDRAM (opt.) caches in various capacity. All of the latency values are normalized to the latency of the 300K SRAM caches with same area.

technology. V_{dd} and V_{th} of the baseline are 0.8V and 0.5V, respectively, which are the 22nm PTM default values [66]. We use the same design for our cryogenic caches, except the detailed circuit design (e.g., placement of repeaters, number of subarrays) and V_{dd} and V_{th} values.

As shown in Fig. 4, the cryogenic cache cannot achieve the target cost-effectiveness without reducing its dynamic energy consumption due to the cooling cost. Therefore, V_{dd} and V_{th} should be reduced to minimize the dynamic energy without losing its performance. However, scaling down the V_{dd} and V_{th} level increases the static energy consumption [27]. Therefore, we should find the optimal voltages to build cost-effective cryogenic caches.

We scale down V_{dd} and V_{th} under the following constraints. First, the access latency of the voltage-scaled 77K caches should be shorter than that of the baseline cache at 77K. Second, among the satisfied V_{dd} and V_{th} sets, we select a set which minimizes the total cache energy consumption. As a result, we set V_{dd} and V_{th} of cryogenic caches to 0.44V and 0.24V, respectively.

In the following sections, we compare the voltage-optimized 77K caches with 77K SRAM caches without voltage scaling. "Opt." means the voltage-optimized cache design, while "No opt." indicates the 77K cache without voltage scaling.

5.2 Latency analysis

Fig. 13 shows the latency breakdown of 300K SRAM, 77K SRAM (no opt.), 77K SRAM (opt.), and 77K 3T-eDRAM (opt.) caches for various capacities. The access latency consists of the decoder, bitline, and Htree latencies. The decoder latency includes the wordline latency. The Htree latency means the global interconnect latency. The irregular points (e.g., 512KB in Fig. 13a) exist because the model proposes differently optimized circuit designs for each capacity.

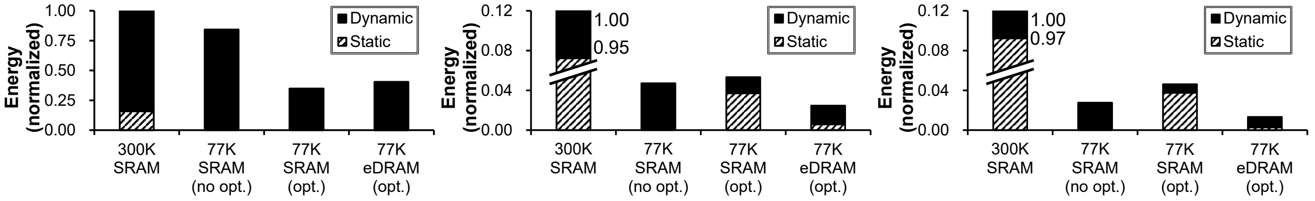
In summary, cryogenic caches are faster than the 300K baseline caches. 77K SRAM (opt.) caches serve the fastest access speed among the cryogenic caches. On the other hand, 77K 3T-eDRAM (opt.) caches can provide twice a larger capacity with the comparable access speed than 77K SRAM (opt.) caches.

First, Fig. 13a shows the latency breakdown of 300K SRAM caches. For the 4KB capacity, the decoder latency dominates the access latency. However, the ratio of decoder latency decreases for larger capacity caches because the decoder latency is proportional to the log of the memory capacity [47]. The ratio of bitline latency also decreases for larger capacity caches because our cache model regulates the bitline latency by splitting one bank to many subarrays. However, the Htree latency portion continually increases and becomes dominant for larger capacity caches. As the Htree latency is proportional to the area, the model cannot regulate the latency by circuit-level optimizations (e.g., number of subarrays). Htree latency occupies 93% of the access latency in the 64MB 300K SRAM cache.

Next, Fig. 13b shows the latency breakdown of 77K SRAM (no opt.) caches. All of the latency components are significantly reduced thanks to the wire resistivity reduction at 77K. Among them, Htree latency greatly decreases because Htree is mostly composed of wires. Therefore, the latency reduction becomes more effective for larger capacity caches where the Htree latency is dominant. The latency of the 64MB 77K SRAM (no opt.) cache is 45.6% of the 64MB 300K SRAM latency.

Fig. 13c shows the latency breakdown of 77K SRAM (opt.) caches. 77K SRAM (opt.) caches are always faster than 77K SRAM (no opt.) caches by scaling down V_{th} (2.1 times) more than V_{dd} (1.8 times) which makes the transistors run faster [23]. The latency of the 64MB 77K SRAM (opt.) cache is 40.6% of the 64MB 300K SRAM latency.

Finally, Fig. 13d shows the latency breakdown of 77K 3T-eDRAM (opt.) caches. Due to the high bitline latency, 77K 3T-eDRAM caches are much slower than the same-area 77K SRAM caches for small capacities. However, the latency of 77K 3T-eDRAM (opt.) caches becomes comparable to the same-area 77K SRAM cache latency for the large capacity range. As the Htree latency is proportional to the area, the 77K 3T-eDRAM's latency becomes comparable to the same-area 77K SRAM cache latency. The access latency of the 128MB 77K 3T-eDRAM (opt.) cache is 47.7% of the 64MB 300K SRAM latency.



(a) L1 (32KB SRAM & 64KB 3T-eDRAM) **(b)** L2 (256KB SRAM & 512KB 3T-eDRAM) **(c)** L3 (8MB SRAM & 16MB 3T-eDRAM)

Figure 14. Energy breakdown of four caches (300K SRAM, 77K SRAM (no opt.), 77K SRAM (opt.), 77K 3T-eDRAM (opt.)) for (a) L1, (b) L2, and (c) L3 design. 3T-eDRAM caches have twice a larger capacity than SRAM caches.

5.3 Energy consumption analysis

Fig. 14 shows the energy breakdown of caches for L1, L2, and L3 design when executing 11 PARSEC 2.1 workloads (i.e., *blackscholes, bodytrack, canneal, dedup, ferret, fluidanimate, rtview, streamcluster, swaptions, vips, x264*) [4] with the baseline setting in Table 2. We use the cache access rate of the baseline for calculating the dynamic energy of each cache.

First, Fig. 14a shows the energy breakdown of L1 caches. The dynamic energy dominates the L1 energy consumption due to its high access rate. The dynamic energy of 77K SRAM (no opt.) cache is the same as that of 300K SRAM cache (84.3%) because the cryogenic cache has the same V_{dd} as the 300K cache. On the other hand, dynamic energies of other 77K caches (33.6% in 77K SRAM (opt.), 40.3% in 77K 3T-eDRAM (opt.)) are lower than that of 300K SRAM cache due to their reduced V_{dd}.

Among the two voltage-optimized cryogenic caches, 77K SRAM (opt.) cache has lower dynamic energy consumption than 77K 3T-eDRAM (opt.) cache. As the 3T-eDRAM cache is twice denser than the SRAM cache, more transistors are connected with the 3T-eDRAM's wordline and bitline. For this reason, 3T-eDRAM caches should drive larger capacitance for switching and consume more dynamic energy than SRAM caches. Therefore, 77K SRAM (opt.) cache has the lowest energy consumption (34.9%) for the L1 design.

Figs. 14b and 14c show the energy consumption of L2 and L3 caches, respectively. The static energy dominates the energy consumption in 300K SRAM caches. The huge area occupancy induces the significant static energy consumption. The L2 and L3 caches occupy 8 and 256 times larger area than L1 caches, respectively. Therefore, their static energies are 8 and 256 times higher and dominate the overall energy consumption.

The static energy of cryogenic caches is lower than that of 300K SRAM caches because the static energy is exponentially proportional to the temperature. Among the cryogenic caches, 77K SRAM (opt.) caches have the highest static energy consumption. Due to the reduced V_{th}, 77K SRAM (opt.) consumes higher static energy than 77K SRAM (no opt.). On the other hand, 77K 3T-eDRAM (opt.) cache has negligible static energy thanks to the low leakage current of PMOS. As the leakage current of PMOS is about ten times lower than that of NMOS, the PMOS-based 3T-eDRAM cache consumes much lower static energy than SRAM caches consisting of NMOS [15]. At the same time, 77K eDRAM (opt.) caches have lower dynamic energy than 77K SRAM (no opt.) caches thanks to its lower V_{dd}. Therefore, 77K 3T-eDRAM (opt.) caches have the lowest energy consumption for L2 and L3 designs. For the L2 design, energy consumption of 77K 3T-eDRAM (opt.) cache (2.5%) is 1.9 and 2.2 times lower than 77K SRAM (no opt.) (4.7%) and 77K SRAM (opt.) (5.3%), respectively. For the L3 design, the energy consumption of 77K 3T-eDRAM (opt.) (1.3%) is lower than that of 77K SRAM (no opt.) (2.8%) by 2.1 times, and that of 77K SRAM (opt.) (4.6%) by 3.5 times, respectively.

Table 2. Evaluation setup

Common specification				
CPU		Based on the Intel i7-6700		
Memory		DDR4 2400		
Cache specification				
Design	Level	Type	Capacity	Latency
Baseline (300K)	L1	SRAM	32KB	4cyc
	L2	SRAM	256KB	12cyc
	L3	SRAM	8MB	42cyc
All SRAM (77K, no opt.)	L1	SRAM	32KB	3cyc
	L2	SRAM	256KB	8cyc
	L3	SRAM	8MB	21cyc
All SRAM (77K, opt.)	L1	SRAM	32KB	2cyc
	L2	SRAM	256KB	6cyc
	L3	SRAM	8MB	18cyc
All eDRAM (77K, opt.)	L1	3T-eDRAM	64KB	4cyc
	L2	3T-eDRAM	512KB	8cyc
	L3	3T-eDRAM	16MB	21cyc
CryoCache	L1	SRAM	32KB	2cyc
	L2	3T-eDRAM	512KB	8cyc
	L3	3T-eDRAM	16MB	21cyc

5.4 Selecting the 77K-optimal cache architecture

Based on the latency and the energy analyses, we propose *CryoCache*, the cryogenic-optimal cache design for high performance and energy efficiency. First, we select 77K SRAM (opt.) for the L1 cache design. The short access latency is the most important factor for L1 design because the system performance is more sensitive to the L1 access latency than the L1 capacity [20, 44]. Reducing the L1 dynamic energy is also important because the dynamic energy dominates the L1 energy consumption. For these reasons, 77K SRAM (opt.) is the best choice for the L1 cache design because it provides the fastest access speed with the minimum dynamic energy consumption.

Second, we select 77K 3T-eDRAM (opt.) for the L2 and L3 cache designs. The system performance is more sensitive to the L3 capacity than the L3 latency due to the huge L3 miss penalty. Reducing the static energy is also important because the static energy dominates both L2 and L3 energy consumption. Therefore, the low-leakage and high-density 77K 3T-eDRAM (opt.) is the best choice for L2 and L3 design because it provides the highest cache capacity with minimum static power consumption.

6 Evaluation using 77K caches

In this section, we show the system-level performance and energy-efficiency of the proposed cache design. We first introduce our evaluation methodology (Section 6.1). Next, we evaluate our cache design in terms of the performance (Section 6.2) and energy consumption (Section 6.3).

6.1 Evaluation methodology

6.1.1 Evaluation setup. For the evaluation, we use Gem5 timing simulator [5]. Our simulation setup is based on Intel i7 6700 processor's specification which has four cores, private L1 and L2 caches, and a shared L3 cache [2]. We utilize 11 PARSEC 2.1 workloads [4].

We evaluate CryoCache by comparing it with the baseline (Baseline (300K)). We also evaluate three other 77K cache-based system designs: systems with 77K SRAM (no opt.) caches (All SRAM (77K, no opt.)), with 77K SRAM (opt.) caches (All SRAM (77K, opt.)), and with 77K 3T-eDRAM (opt.) caches (All eDRAM (77K, opt.)).

We set the latency of 77K caches based on the relative speed-up obtained in Section 5.2. For example, our model predicts that the 8MB 77K SRAM (opt.) cache is 2.3 times faster than the 8MB 300K SRAM cache. Therefore, we set the latency of 77K-optimized SRAM to 18 cycles, which is 2.3 times shorter than the baseline latency. We summarize the setup in Table 2.

6.1.2 Energy evaluation methodology. We include the energy consumption for the cryogenic cooling because the cooling energy dominates the overall energy consumption at 77K. The cooling energy consumption ($E_{cooling}$) can be represented as the electrical energy to remove the heat dissipated from the device (Eq. (1)).

$$E_{cooling} = E_{device} \cdot CO \tag{1}$$

E_{device} is the energy consumption of the electronic devices and CO is the cooling overhead [24]. The cooling overhead indicates the required energy to remove unit heat (1J) from the cooling system. The cooling overhead significantly increases as the target temperature decreases and it reaches 9.65 in the 77K cooling system [24]. Therefore, we use 9.65 value for our 77K cooling overhead (CO_{77K}).

$$\begin{aligned} E_{77K\text{-}total} &= E_{77K\text{-}device} + E_{77K\text{-}cooling} \\ &= (1 + CO_{77K})\, E_{77K\text{-}device} \\ &= 10.65\, E_{77K\text{-}device} \end{aligned} \tag{2}$$

Based on the cooling energy model, we calculate the total required energy for our 77K system ($E_{77K\text{-}total}$) as Eq. (2). Eq. (2) indicates that the 77K cache should consume at most 10.65 times less energy than the 300K cache to achieve the energy efficiency. Note that we exclude the cooling cost for the 300K baseline system to conservatively show the cryogenic cache's energy efficiency.

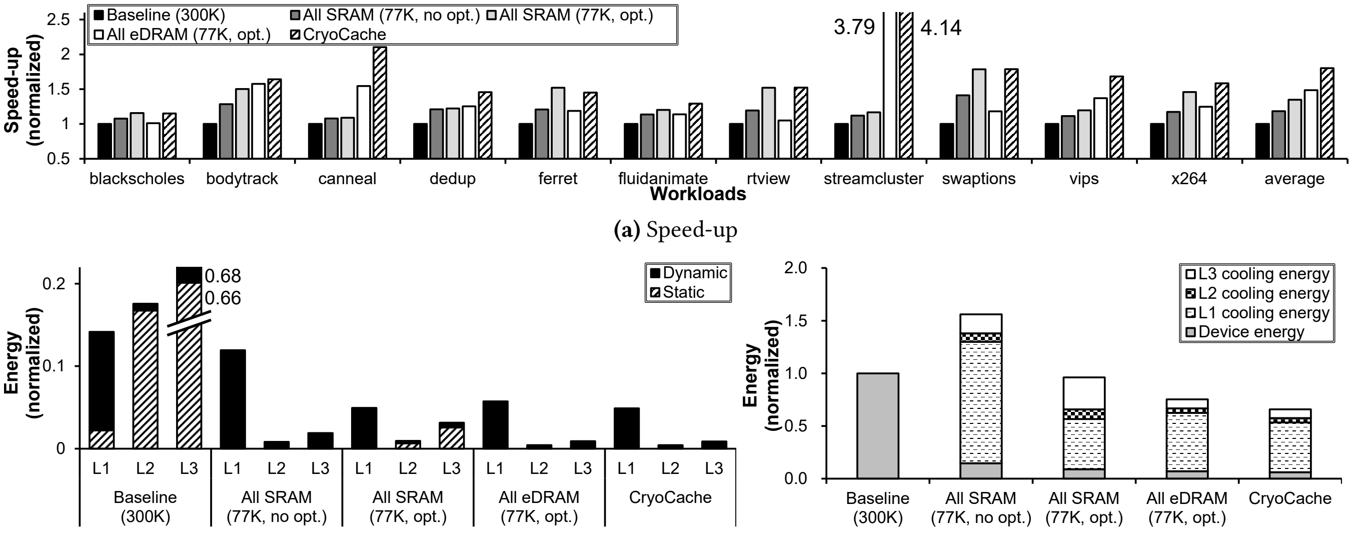
The 77K cooling system also needs the LN cost and the cooling facility cost. However, we focus only on the cooling energy consumption because the LN cost and the cooling facility cost are the one-time cost to build LN recycling systems. The recurring cooling energy cost is much higher than the one-time cost and dominates the cryogenic cooling cost [36]. For this reason, our energy evaluation reflects the realistic cooling cost for 77K.

6.2 Performance evaluation

Fig. 15a shows the speed-up of cryogenic caches. The speed-up is inversely proportional to the execution time normalized to that of the baseline. In our performance evaluation, CryoCache achieves the highest speed-up (80%) compared to others.

First, All SRAM (77K, no opt.) achieves the speed-up of 18.3% on average, up to 41.0% for *swaptions*. The speed-up purely results from the reduced access latency. Each workload has a different speed-up, due to the differences in the performance bottleneck. For example, *swaptions* shows the highest speed-up (41.0%) because *swaptions* has the largest cache portion in the CPI stack (Fig. 2). On the other hand, *canneal* shows a marginal speed-up (7.9%) because its performance is not much affected by the cache latency.

Next, All SRAM (77K, opt.) achieves the speed-up of 34.7% on average, up to 78.5% in *swaptions*. All SRAM (77K, opt.) achieves the higher speed-up than All SRAM (77K, no opt.) because the voltage-optimized caches are faster than the unoptimized caches. *Swaptions* shows the highest speed-up (78.5%) for the same reason as the All SRAM (77K, no opt.) case.



(a) Speed-up

(b) Cache energy breakdown

(c) Total energy consumption

Figure 15. (a) Speed-up, (b) cache energy breakdown, and (c) total energy consumption (including the cooling cost) of the five cache designs. The values are normalized to those of Baseline (300K).

All eDRAM (77K, opt.) shows the speed-up of 48.6% on average, up to 3.79 times for *streamcluster*. All eDRAM (77K, opt.) achieves 13.9% higher speed-up compared to All SRAM (77K, opt.) and it comes mainly from the doubled capacity. Among workloads, *streamcluster* achieves the highest speed-up (3.79 times) because its working set (16MB) fits for the new LLC capacity [4]. The doubled capacity also significantly improves other capacity-sensitive workloads such as *canneal*.

Unfortunately, All SRAM (77K, opt.) and All eDRAM (77K, opt.) cannot improve the performance of capacity-critical and latency-critical workloads, respectively. In All SRAM (77K, opt.), the performance of *streamcluster* and *canneal* remains nearly the same because the reduced access latency cannot benefit these workloads, as shown in the CPI stack (Fig. 2). On the other hand, All eDRAM (77K, opt.) greatly improves the performance of capacity-critical workloads (i.e., *streamcluster, canneal*), but cannot benefit the latency-critical workloads (i.e., *blackscholes, ferret, rtview, swaptions, x264*).

Different from two cases, CryoCache can boost both the latency-critical workloads and the capacity-critical workloads. CryoCache provides both the low access latency and the large capacity by utilizing faster SRAM in L1 design and denser 3T-eDRAM in L2 and L3 designs. Therefore, our cache architecture outperforms other designs for most of the workloads. CryoCache achieves the speed-up of 80% on average, up to 4.14 times for *streamcluster*. For some workloads (i.e., *blackscholes, ferret*), the speed-up of CryoCache is slightly smaller than All SRAM (77K, opt.) because the relatively long access latency of L2 and L3 3T-eDRAM more strongly affects the performance than the doubled capacity. Except for these workloads, CryoCache outperforms other designs thanks to its carefully designed cache architecture.

6.3 Energy evaluation

Figs. 15b and 15c show the cache energy breakdown and the total energy consumption including the cooling cost, respectively. Energy values are normalized to the total energy consumption of Baseline (300K). In our energy evaluation, CryoCache has the lowest cache energy consumption (6.2%) and total energy consumption (65.9%).

In Baseline (300K), the L1 dynamic energy occupies 11.9% of the cache energy consumption. The L2 and L3 static energy consumptions are 16.8% and 66.4% of the total cache energy consumption.

In All SRAM (77K, no opt.), the static energy consumption is almost eliminated thanks to the low temperature. However, the L1 dynamic energy consumption (11.9%) dominates the cache energy (Fig. 15b) and induces the huge cooling energy consumption (Fig. 15c). Therefore, the total energy consumption of All SRAM (77K, no opt.) is 56% higher than that of the baseline.

In All SRAM (77K, opt.), the overall dynamic energy is significantly reduced thanks to the V_{dd} and V_{th} scaling. However, the L2 and L3 static energy consumptions increase due to the reduced V_{th}. The L2 and L3 static energy consumptions occupy 35.6% of total cache energy and incur the huge cooling energy consumption (31.0%).

On the other hand, All eDRAM (77K, opt.) has the significantly reduced energy consumption thanks to the low static power of 3T-eDRAM. The cache energy consumption of All eDRAM (77K, opt.) is 7.1% of the baseline energy. As a result, All eDRAM (77K, opt.) consumes 24.6% less total energy than the baseline.

However, CryoCache consumes much less energy than others. Unlike All SRAM (77K, opt.) case, CryoCache uses

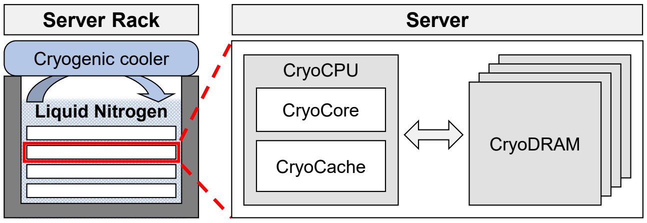


Figure 16. Overview of the full cryogenic computer system

3T-eDRAM for L2 and L3 design which greatly reduces their static energy. For the dynamic energy-critical L1 design, we select the 32KB SRAM cache which consumes much less dynamic energy than the 64KB 3T-eDRAM. Therefore, the cache energy consumption is reduced to 6.19% of the baseline cache energy. The total energy consumption is also 34.1% lower than that of the baseline. That is, by utilizing our proposed cache design, architects can increase the system's performance up to 4.14 times, with 34.1% lower cost.

7 Discussion

7.1 Full cryogenic computer system

Our cryogenic cache study is an intermediate step prior to building the full cryogenic computer systems. Fig. 16 shows the overview of a full cryogenic computer system where the entire computing nodes are cooled down to 77K. For a higher cooling efficiency, the system recycles LN by using a cryogenic cooler. Both CPUs and DRAMs are fabricated with their V_{dd} and V_{th} scaled down for a higher energy efficiency. As a result, the entire system is expected to achieve a huge performance gain by utilizing the reduced wire resistivity and increased carrier mobility. Therefore, the 77K cryogenic computer system will greatly improve both the system's performance and energy efficiency.

7.2 Cache-to-pipeline interface

This work focuses on cooling the cache hierarchy only and analyzing its impact isolated from the rest of the processor, while the entire processor is being cooled (Fig. 16). Note that the entire processor reliably works at 77K (Fig. 3). The result indicates that the cache-to-pipeline interface can reliably operate at the low temperature and is likely to run even faster. Therefore, we do not address the reliability of cache-to-pipeline interface in this work.

For the fair and conservative performance analysis, we used the un-cache part's performance at 300K for the evaluation (Section 6). However, the low temperature can greatly improve the pipeline's performance and energy efficiency, the same as the cache part. As our next work, we are currently architecting cryogenic-optimal pipelines.

7.3 V_{th} scaling of FinFET devices

We scale down V_{dd} and V_{th} to achieve the power efficiency of cryogenic caches. However, previous works show that V_{th} of FinFET is difficult to control, compared to that of the planar devices [21]. Despite such difficulties, previous works suggest solutions to control the V_{th} of FinFET, such as carefully choosing the gate metal [55], optimizing dopant profile in the channel [33], or changing the pocket (halo) doping level of the devices [13]. Therefore, we believe that our power reduction scheme for cryogenic computing (i.e., V_{dd}/V_{th} scaling) is feasible not only for the planar MOSFET-based chips, but also for modern FinFET-based processors.

7.4 Relevant to quantum computing

CryoCache is also highly promising for the low-temperature quantum computing as its controller design. There exist two representative designs as quantum computer's controllers: CMOS-based quantum controller [22, 43], and superconducting-logic-based quantum controller [31, 38]. Regardless of the controller's type, CryoCache contributes to both types of quantum controllers. First, the CMOS-based controller needs CMOS-based caches to store the requests from users and the results of quantum computers. Second, the superconducting-logic-based quantum controllers also might build their memory parts with CMOS because architects cannot build large memory modules with the superconducting circuits yet [8, 9]. For these reasons, CryoCache can also contribute to quantum computing as a memory module for quantum controllers.

8 Related work

In this section, we discuss prior works which focused on heterogeneous cache design and 77K cryogenic computing.

Heterogeneous cache design. Chang et al. [11] and Xie [62] analyzed several types of memory cells and compared them in terms of performance and energy efficiency. Also, various hybrid cache designs [6, 34, 51, 61] utilized several memory technologies for the 3D stacked architecture. Wang et al. [58] and Kotra et al. [28] proposed the hybrid cache design for LLC. However, there is no previous work which compares and utilizes the various memory technologies at 77K.

77K cryogenic computing. Tannu et al. [54] confirmed that the commodity DRAM chips can work reliably at 80K. Rambus [56] showed that the 77K environment can greatly reduce the DRAM refresh overhead and also suggested that DRAM is the most appropriate memory technology to support the 4K superconducting processors at 77K domain [59]. Recently, Lee et al. [29] built a cryogenic DRAM model (i.e., CryoRAM) which calculates access latency and power consumption of most optimized DRAM design for various target temperatures. They also showed the potential of 77K-optimized DRAM in the datacenter level. However, the previous works only focus on the cryogenic DRAM.

To the best of our knowledge, our work is the first study to compare and utilize various cache technologies at 77K. Our work is also the first study to show the cryogenic cache's potential in terms of performance and power efficiency.

9 Conclusion

Cryogenic computing is a highly promising solution to dramatically improve both performance and power efficiency. However, computer architects are facing challenges in developing and deploying cryogenic on-chip caches due to the lack of understanding about its cost-effectiveness and feasibility. To resolve the problem, we thoroughly analyzed the cost-effectiveness and feasibility of various on-chip memory technologies running at 77K. Based on the analysis, we architected CryoCache, a cost-effective, technology-feasible cryogenic-optimal cache architecture with 6T-SRAM and 3T-eDRAM technologies. Our evaluation clearly indicates that cryogenic computing can effectively double the cache access speed and capacity with greatly reduced total energy cost.

Acknowledgments

This work was partly supported by Samsung Electronics Co., Ltd., National Research Foundation of Korea (NRF) grant funded by the Korean Government (NRF-2015M3C4A7065647, NRF-2017R1A2B3011038, NRF-2019R1A5A1027055, NRF-2019-Global Ph.D. Fellowship Program), Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (No.2018-0-00769), and Creative Pioneering Researchers Program through Seoul National University. We also appreciate the support from Automation and Systems Research Institute (ASRI) and Inter-university Semiconductor Research Center (ISRC) at Seoul National University.

References

[1] [n.d.]. *Cache specifications from 7-cpu.* https://www.7-cpu.com/

[2] [n.d.]. *Intel i7 6700 processor.* https://ark.intel.com/content/www/us/en/ark/products/88196/intel-core-i7-6700-processor-8m-cache-up-to-4-00-ghz.html

[3] Francis Balestra and Gérard Ghibaudo. 2001. *Device and circuit cryogenic operation for low temperature electronics.* Springer.

[4] Christian Bienia, Sanjeev Kumar, Jaswinder Pal Singh, and Kai Li. 2008. The PARSEC benchmark suite: Characterization and architectural implications. In *Proceedings of the 17th international conference on Parallel architectures and compilation techniques.* ACM, 72–81.

[5] Nathan Binkert, Bradford Beckmann, Gabriel Black, Steven K Reinhardt, Ali Saidi, Arkaprava Basu, Joel Hestness, Derek R Hower, Tushar Krishna, Somayeh Sardashti, et al. 2011. The gem5 simulator. *ACM SIGARCH Computer Architecture News* 39, 2 (2011), 1–7.

[6] Bryan Black, Murali Annavaram, Ned Brekelbaum, John DeVale, Lei Jiang, Gabriel H Loh, Don McCaule, Pat Morrow, Donald W Nelson, Daniel Pantuso, et al. 2006. Die stacking (3D) microarchitecture. In *Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture.* IEEE Computer Society, 469–479.

[7] Darren K Brock. 2001. RSFQ technology: Circuits and systems. *International journal of high speed electronics and systems* 11, 01 (2001), 307–362.

[8] Darren K Brock. 2001. RSFQ technology: Circuits and systems. *International journal of high speed electronics and systems* 11, 01 (2001), 307–362.

[9] Paul Bunyk, Konstantin Likharev, and Dmitry Zinoviev. 2001. RSFQ technology: Physics and devices. *International journal of high speed electronics and systems* 11, 01 (2001), 257–305.

[10] Hao Cai, Wang Kang, You Wang, Lirida Naviner, Jun Yang, and Weisheng Zhao. 2017. High performance MRAM with spin-transfer-torque and voltage-controlled magnetic anisotropy effects. *Applied Sciences* 7, 9 (2017), 929.

[11] Mu-Tien Chang, Paul Rosenfeld, Shih-Lien Lu, and Bruce Jacob. 2013. Technology comparison for large last-level caches (L 3 Cs): Low-leakage SRAM, low write-energy STT-RAM, and refresh-optimized eDRAM. In *2013 IEEE 19th International Symposium on High Performance Computer Architecture (HPCA).* IEEE, 143–154.

[12] Yunji Chen, Tao Luo, Shaoli Liu, Shijin Zhang, Liqiang He, Jia Wang, Ling Li, Tianshi Chen, Zhiwei Xu, Ninghui Sun, et al. 2014. Dadiannao: A machine-learning supercomputer. In *Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture.* IEEE Computer Society, 609–622.

[13] Hye Jin Cho, Jeong Dong Choe, Jeongnam Han, Dongchan Kim, Heungsik Park, Doohoon Goo, Ming Li, Chang Woo Oh, Dong-Won Kim, Tae Yong Kim, et al. 2005. The Vth controllability of 5nm body-tied CMOS FinFET. In *IEEE VLSI-TSA International Symposium on VLSI Technology, 2005.(VLSI-TSA-Tech).* IEEE, 116–117.

[14] Ki Chul Chun, Pulkit Jain, Jung Hwa Lee, and Chris H Kim. 2009. A sub-0.9 V logic-compatible embedded DRAM with boosted 3T gain cell, regulated bit-line write scheme and PVT-tracking read reference bias. In *2009 Symposium on VLSI Circuits.* IEEE, 134–135.

[15] Ki Chul Chun, Pulkit Jain, Jung Hwa Lee, and Chris H Kim. 2011. A 3T gain cell embedded DRAM utilizing preferential boosting for high density and low power on-die caches. *IEEE Journal of Solid-State Circuits* 46, 6 (2011), 1495–1505.

[16] Ki Chul Chun, Hui Zhao, Jonathan D Harms, Tae-Hyoung Kim, Jian-Ping Wang, and Chris H Kim. 2012. A scaling roadmap and performance evaluation of in-plane and perpendicular MTJ based STT-MRAMs for high-density cache memory. *IEEE Journal of Solid-State Circuits* 48, 2 (2012), 598–610.

[17] Xiangyu Dong, Cong Xu, Yuan Xie, and Norman P Jouppi. 2012. Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 31, 7 (2012), 994–1007.

[18] Xuanyao Fong, Yusung Kim, Karthik Yogendra, Deliang Fan, Abhronil Sengupta, Anand Raghunathan, and Kaushik Roy. 2015. Spin-transfer torque devices for logic and memory: Prospects and perspectives. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 35, 1 (2015), 1–22.

[19] WH Henkels, NCC Lu, W Hwang, TV Rajeevakumar, RL Franch, KA Jenkins, TJ Bucelot, DF Heidel, and MJ Immediato. 1989. A 12-ns low-temperature DRAM. *IEEE Transactions on Electron Devices* 36, 8 (1989), 1414–1422.

[20] Farrukh Hijaz, Qingchuan Shi, and Omer Khan. 2013. A private level-1 cache architecture to exploit the latency and capacity tradeoffs in multicores operating at near-threshold voltages. In *2013 IEEE 31st International Conference on Computer Design (ICCD).* IEEE, 85–92.

[21] Digh Hisamoto, Wen-Chin Lee, Jakub Kedzierski, Hideki Takeuchi, Kazuya Asano, Charles Kuo, Erik Anderson, Tsu-Jae King, Jeffrey Bokor, and Chenming Hu. 2000. FinFET-a self-aligned double-gate MOSFET scalable to 20 nm. *IEEE transactions on electron devices* 47, 12 (2000), 2320–2325.

[22] JM Hornibrook, JI Colless, ID Conway Lamb, SJ Pauka, H Lu, AC Gossard, JD Watson, GC Gardner, S Fallahi, MJ Manfra, et al. 2015. Cryogenic control architecture for large-scale quantum computing. *Physical Review Applied* 3, 2 (2015), 024010.

[23] Chenming Hu. 2010. *Modern semiconductor devices for integrated circuits.* Vol. 2. Prentice Hall Upper Saddle River, NJ.

[24] Yukikazu Iwasa. 2009. *Case studies in superconducting magnets: design and operational issues.* Springer Science & Business Media.

[25] Jodi M Iwata-Harms, Guenole Jan, Huanlong Liu, Santiago Serrano-Guisan, Jian Zhu, Luc Thomas, Ru-Ying Tong, Vignesh Sundar, and Po-Kang Wang. 2018. High-temperature thermal stability driven by magnetization dilution in CoFeB free layers for spin-transfer-torque magnetic random access memory. *Scientific reports* 8, 1 (2018), 14409.

[26] Naifeng Jing, Yao Shen, Yao Lu, Shrikanth Ganapathy, Zhigang Mao, Minyi Guo, Ramon Canal, and Xiaoyao Liang. 2013. An energy-efficient and scalable eDRAM-based register file architecture for GPGPU. In *ACM SIGARCH Computer Architecture News*, Vol. 41. ACM, 344–355.

[27] Nam Sung Kim, Todd Austin, David Blaauw, Trevor Mudge, Jie S Hu, Mary Jane Irwin, Mahmut Kandemir, Vijaykrishnan Narayanan, et al. 2003. Leakage Current: Moore. *computer* 12 (2003), 68–75.

[28] Jagadish B Kotra, Mohammad Arjomand, Diana Guttman, Mahmut T Kandemir, and Chita R Das. 2016. Re-NUCA: A practical nuca architecture for reram based last-level caches. In *2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 576–585.

[29] Gyu-hyeon Lee, Dongmoon Min, Ilkwon Byun, and Jangwoo Kim. 2019. Cryogenic Computer Architecture Modeling with Memory-side Case Studies. In *Proceedings of the 46th International Symposium on Computer Architecture* (Phoenix, Arizona) *(ISCA '19)*. ACM, New York, NY, USA, 774–787. https://doi.org/10.1145/3307650.3322219

[30] Xiaoyao Liang, Ramon Canal, Gu-Yeon Wei, and David Brooks. 2007. Process variation tolerant 3T1D-based cache architectures. In *Proceedings of the 40th Annual IEEE/ACM International Symposium on Microarchitecture*. IEEE Computer Society, 15–26.

[31] Per J Liebermann and Frank K Wilhelm. 2016. Optimal Qubit Control Using Single-Flux Quantum Pulses. *Physical Review Applied* 6, 2 (2016), 024022.

[32] Konstantin K Likharev and Vasilii K Semenov. 1991. RSFQ logic/memory family: A new Josephson-junction technology for sub-terahertz-clock-frequency digital systems. *IEEE Transactions on Applied Superconductivity* 1, 1 (1991), 3–28.

[33] C-H Lin, R Kambhampati, RJ Miller, TB Hook, A Bryant, W Haensch, P Oldiges, I Lauer, T Yamashita, V Basker, et al. 2012. Channel doping impact on FinFETs for 22nm and beyond. In *2012 Symposium on VLSI Technology (VLSIT)*. IEEE, 15–16.

[34] Gabriel H Loh, Yuan Xie, and Bryan Black. 2007. Processor design in 3D die-stacking technologies. *Ieee Micro* 27, 3 (2007), 31–48.

[35] Kristen Lovin, Benjamin C Lee, Xiaoyao Liang, David Brooks, and Gu-Yeon Wei. 2009. Empirical performance models for 3T1D memories. In *2009 IEEE International Conference on Computer Design*. IEEE, 398–403.

[36] William L Luyben. 2017. Estimating refrigeration costs at cryogenic temperatures. *Computers & Chemical Engineering* 103 (2017), 144–150.

[37] Richard Allen Matula. 1979. Electrical resistivity of copper, gold, palladium, and silver. *Journal of Physical and Chemical Reference Data* 8, 4 (1979), 1147–1298.

[38] R McDermott and MG Vavilov. 2014. Accurate qubit control with single flux quantum pulses. *Physical Review Applied* 2, 1 (2014), 014007.

[39] Naveen Muralimanohar and Rajeev Balasubramonian. 2007. Interconnect design considerations for large NUCA caches. In *ACM SIGARCH Computer Architecture News*, Vol. 35. ACM, 369–380.

[40] Naveen Muralimanohar, Rajeev Balasubramonian, and Norman P Jouppi. [n.d.]. CACTI 6.0: A tool to model large caches. ([n. d.]).

[41] Ikki Nagaoka, Masamitsu Tanaka, Koji Inoue, and Akira Fujimaki. 2019. 29.3 A 48GHz 5.6 mW Gate-Level-Pipelined Multiplier Using Single-Flux Quantum Logic. In *2019 IEEE International Solid-State Circuits Conference-(ISSCC)*. IEEE, 460–462.

[42] John K Ousterhout, Gordon T Hamachi, Robert N Mayo, Walter S Scott, and George S Taylor. 1985. The magic VLSI layout system. *IEEE Design & Test of Computers* 2, 1 (1985), 19–30.

[43] Bishnu Patra, Rosario M Incandela, Jeroen PG Van Dijk, Harald AR Homulle, Lin Song, Mina Shahmohammadi, Robert Bogdan Staszewski, Andrei Vladimirescu, Masoud Babaie, Fabio Sebastiano, et al. 2017. Cryo-CMOS circuits and systems for quantum computing applications. *IEEE Journal of Solid-State Circuits* 53, 1 (2017), 309–321.

[44] David A Patterson and John L Hennessy. 2013. *Computer organization and design MIPS edition: the hardware/software interface.* Newnes.

[45] RG Pires, RM Dickstein, SL Titcomb, and RL Anderson. 1990. Carrier freezeout in silicon. *Cryogenics* 30, 12 (1990), 1064–1068.

[46] Masood Qazi, Mahmut Sinangil, and Anantha Chandrakasan. 2011. Challenges and directions for low-voltage SRAM. *IEEE design & test of computers* 28, 1 (2011), 32–43.

[47] Glen Reinman and Norman P Jouppi. 2000. CACTI 2.0: An integrated cache timing and power model. *Western Research Lab Research Report* 7 (2000).

[48] Oleg Semenov, Arman Vassighi, and Manoj Sachdev. 2002. Impact of technology scaling on thermal behavior of leakage current in sub-quarter micron MOSFETs: perspective of low temperature current testing. *Microelectronics Journal* 33, 11 (2002), 985–994.

[49] M Shin, M Shi, M Mouis, A Cros, E Josse, Gyu-Tae Kim, and G Ghibaudo. 2014. Low temperature characterization of 14nm FDSOI CMOS devices. In *2014 11th International Workshop on Low Temperature Electronics (WOLTE)*. IEEE, 29–32.

[50] Richard G Southwick, Justin Reed, Christopher Buu, Hieu Bui, Ross Butler, G Bersuker, and William B Knowlton. 2008. Temperature (5.6-300K) Dependence Comparison of Carrier Transport Mechanisms in HfO 2/SiO 2 and SiO 2 MOS Gate Stacks. In *2008 IEEE International Integrated Reliability Workshop Final Report*. IEEE, 48–54.

[51] Guangyu Sun, Xiangyu Dong, Yuan Xie, Jian Li, and Yiran Chen. 2009. A novel architecture of the 3D stacked MRAM L2 cache for CMPs. In *2009 IEEE 15th International Symposium on High Performance Computer Architecture*. IEEE, 239–249.

[52] N Takeuchi, K Ehara, K Inoue, Y Yamanashi, and N Yoshikawa. 2013. Margin and energy dissipation of adiabatic quantum-flux-parametron logic at finite temperature. *IEEE Transactions on Applied Superconductivity* 23, 3 (2013), 1700304–1700304.

[53] Naoki Takeuchi, Dan Ozawa, Yuki Yamanashi, and Nobuyuki Yoshikawa. 2013. An adiabatic quantum flux parametron as an ultra-low-power logic device. *Superconductor Science and Technology* 26, 3 (2013), 035010.

[54] Swamit S Tannu, Douglas M Carmean, and Moinuddin K Qureshi. 2017. Cryogenic-DRAM based memory system for scalable quantum computers: a feasibility study. In *Proceedings of the International Symposium on Memory Systems*. ACM, 189–195.

[55] Narendar Vadthiya, Ramanuj Mishra, Sanjeev Rai, and R Mishra. 2012. Threshold Voltage Control Schemes in Finfets. *International Journal of VLSI Design and Communication Systems* 3 (2012).

[56] Fiona Wang, Thomas Vogelsang, Brent Haukness, and Stephen C Magee. 2018. DRAM Retention at Cryogenic Temperatures. In *2018 IEEE International Memory Workshop (IMW)*. IEEE, 1–4.

[57] G Wang, D Anand, N Butt, A Cestero, M Chudzik, J Ervin, S Fang, G Freeman, H Ho, B Khan, et al. 2009. Scaling deep trench based eDRAM on SOI to 32nm and Beyond. In *2009 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 1–4.

[58] Zhe Wang, Daniel A Jiménez, Cong Xu, Guangyu Sun, and Yuan Xie. 2014. 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, 13–24.

[59] Fred Ware, Liji Gopalakrishnan, Eric Linstadt, Sally A McKee, Thomas Vogelsang, Kenneth L Wright, Craig Hampel, and Gary Bronner. 2017. Do superconducting processors really need cryogenic memories?: the case for cold DRAM. In *Proceedings of the International Symposium on Memory Systems*. ACM, 183–188.

[60] Bi Wu, Pengcheng Dai, Yuanqing Cheng, Ying Wang, Jianlei Yang, Zhaohao Wang, Dijun Liu, and Weisheng Zhao. 2019. A Novel High Performance and Energy Efficient NUCA Architecture for STT-MRAM LLCs with Thermal Consideration. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (2019).

[61] Xiaoxia Wu, Jian Li, Lixin Zhang, Evan Speight, Ram Rajamony, and Yuan Xie. 2009. Hybrid cache architecture with disparate memory technologies. In *ACM SIGARCH computer architecture news*, Vol. 37. ACM, 34–45.

[62] Yuan Xie. 2011. Modeling, architecture, and applications for emerging memory technologies. *IEEE design & test of computers* 28, 1 (2011), 44–51.

[63] N Yoshikawa, D Ozawa, and Y Yamanashi. 2011. Ultra-low-power superconducting logic devices using adiabatic quantum flux parametron. In *Extended Abstracts of the 2011 International Conference on Solid State Devices and Materials (SSDM 2011), Nagoya*.

[64] Yan Zhang, Dharmesh Parikh, Karthik Sankaranarayanan, Kevin Skadron, and Mircea Stan. 2003. Hotleakage: A temperature-aware model of subthreshold and gate leakage for architects. *University of Virginia Dept of Computer Science Tech Report CS-2003* 5 (2003).

[65] Hongliang Zhao and Xinghui Liu. 2014. Modeling of a standard 0.35 μm CMOS technology operating from 77 K to 300 K. *Cryogenics* 59 (2014), 49–59.

[66] Wei Zhao and Yu Cao. 2006. New generation of predictive technology model for sub-45 nm early design exploration. *IEEE Transactions on Electron Devices* 53, 11 (2006), 2816–2823.