Ph.D. DISSERTATION

Fast and Power-Efficient Cryogenic CMOS Computer Architecture

빠르고 저전력인
극저온 CMOS 기반 컴퓨터 설계

BY

Dongmoon Min

FEBRUARY 2024

DEPARTMENT OF ELECTRICAL AND
COMPUTER ENGINEERING
SEOUL NATIONAL UNIVERSITY

Ph.D. DISSERTATION

Fast and Power-Efficient Cryogenic CMOS Computer Architecture

빠르고 저전력인
극저온 CMOS 기반 컴퓨터 설계

BY

Dongmoon Min

FEBRUARY 2024

DEPARTMENT OF ELECTRICAL AND
COMPUTER ENGINEERING
SEOUL NATIONAL UNIVERSITY

Fast and Power-Efficient Cryogenic CMOS Computer Architecture

빠르고 저전력인
극저온 CMOS 기반 컴퓨터 설계

지도교수 김 장 우

이 논문을 공학박사 학위논문으로 제출함

2024년 1월

서울대학교 대학원

전기 · 정보 공학부

민 동 문

민동문의 공학박사 학위 논문을 인준함

2024년 1월

위 원 장: 김 재 준 (인)

부위원장: 김 장 우 (인)

위 원: 심 재 웅 (인)

위 원: 이 진 호 (인)

위 원: 김 한 준 (인)

Abstract

Modern computer architectures suffer from a lack of architectural innovations, mainly due to the power wall and memory wall problems. That is, architectural innovations become infeasible because they can prohibitively increase the power consumption (i.e., power wall) and their performance impacts are eventually bounded by slow memories (i.e., memory wall). To address the challenges, making computer systems run at ultra-low temperatures (or cryogenic computer systems) has emerged as a highly promising solution as both power consumption and wire resistivity are expected to significantly reduce at low temperatures. However, cryogenic computers have not been yet realized due to the lack of understanding about their cost-effectiveness and feasibility (e.g., device and cooling costs vs. speedup, energy and area saving) and thus how to architect such cryogenic-optimal computer units.

In this dissertation, we introduce the CryoServer project to design a fast and power-efficient cryogenic CMOS-based computer system. In this project, we focus on 77K temperature (-196°C; easily achieved by applying low-cost liquid nitrogen), at which modern CMOS devices reliably operate with moderate cooling cost. To realize full cryogenic computer systems, we develop 77K-optimized computer units for three major computer devices (i.e., DRAM, cache, and core). First, we develop *CryoRAM*, a validated cryogenic DRAM performance modeling tool, and propose two cryogenic-optimal DRAM designs (i.e., CLL-DRAM, CLP-DRAM) targeting for high performance and low power consumption, respectively. Second, we propose *CryoCache*, a fast, large, and power-efficient 77K-optimized cache architecture. Finally, we develop *CryoCore*, a 77K-optimized core architecture, which maximizes core's performance and area efficiency while minimizing the cooling cost. The full cryogenic computer systems equipped with our 77K-optimized DRAM, cache, and core designs achieves significant performance gain and power efficiency even including the cooling cost.

keywords: Cryogenic Computing, DRAM, Cache, Processor, Performance Modeling, Computer Architecture

student number: 2018-29413

Contents

List of Tables

List of Figures

Chapter 1

Introduction

High-performance computing and datacenter industries always require the fastest and most power-efficient computer systems. However, facing the end of both Moore's Law and Dennard Scalings, server architects are now facing critical challenges to further improve performance and power efficiency of the current high-end server systems.

To build a faster computer under the same power budget, both Moore's Law [88] and Dennard Scaling [28] must be satisfied so that both the size and operating voltage of a transistor can be reduced simultaneously. Only then, architects can place more logics and memories on the same sized chip, and increase the chip's frequency without increasing its power consumption. However, we are now experiencing the end of the both trends, mainly due to the difficulty in reducing the transistor's supply and threshold voltage without prohibitively increasing its leakage power (*power wall problem*).

On the other hand, even if the power wall problem were magically resolved, the unimproved memory performance would remain as another critical problem because the memory access latency is bounded by the wire latency rather than the transistor speed. Then, any architectural innovations do not contribute to the system's overall performance improvement (*memory wall problem* [112]).

To get around the power wall and memory wall problems, various approaches have been proposed such as the deployment of slow multi-core designs [58, 74] and the information processing close to or within the memory. But, these circumventions can suffer from the parallelization overhead, the increasing on-chip power consumption and the requirement of radical architectural innovation.

Therefore, computer architects are now more than ever in dire need of effectively resolving the power and memory walls. To achieve the goal, the concept of running a computer at ultra-low temperatures (e.g., -200°C) (or *cryogenic computer*) has emerged as a highly promising idea because reducing the temperature leads to the exponential decrease of leakage power and the linear decrease of wire resistivity at the same time. The reduced leakage power (allowing the reduced operating voltages as well) and wire resistivity can realize extremely low-power computer modules and low-latency memory accesses.

However, cryogenic computers have not been yet realized as realistic solutions in the field due to the following reasons. First, computer architects do not fully understand how computer systems behave at such ultra-low temperatures and how the cost effectiveness of the systems would be affected with the cooling cost considered. Second, there is no modeling tool available to the architects which can be used to evaluate the performance, power and cost of the cryogenic architecture designs. Finally, there is no cryogenic-optimal architecture for major computer devices (e.g., DRAM, cache, CPU core) which achieves the highest performance and power efficiency at 77K.

To this end, we initiate the CryoServer project, in which we model and design a fast and power-efficient cryogenic computer system running at 77K. In this project, we model and develop 77K-optimized DRAM, cache, and CPU core devices as follows.

- **CryoRAM:** First, to model and design the 77K-optimized DRAM architecture, we develop *CryoRAM*, a validated cryogenic memory simulation tool, which accurately predicts the access latency and power consumption of DRAM devices running at 77K. CryoRAM consists of MOSFET, DRAM, and thermal models, and we thoroughly validate each of them with real cryogenic experiments. Next, driven by the modeling tool, we propose two cryogenic-optimized memory devices (i.e., CLL-

DRAM and CLP-DRAM) which improve the DRAM access speed and reduce the DRAM power consumption, respectively. Finally, we provide three promising case studies using the cryogenic memories. In single server-level case studies, we show that the cryogenic DRAM can significantly improves server performance and power efficiency. In datacenter-level case study, we show that our power-efficient DRAM can greatly reduce the datacenter's total power even including the cooling cost.

- **CryoCache:** CryoCache is a cost-effective, technology-feasible cryogenic-optimal cache architecture running at 77K. To develop CryoCache, 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 build a modeling framework to estimate the latency and power of SRAM and 3T-eDRAM cell-based caches running at 77K. We validate our cache model by comparing it with Hspice simulation using industry-validated cryogenic MOSFET model. Third, driven by the modeling tool, we show that the latency and power consumption of SRAM caches can be significantly reduced at 77K, and the degree of reduction increases with the cache's physical size. Fourth, 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. Finally, based on the analysis, we develop CryoCache, which consists of 6T-SRAM cell-based L1 caches and 3T-eDRAM cell-based L2 and L3 caches. CryoCache significantly improves performance thanks to the twice improved cache-access speed and capacity, while reducing total required power including cooling cost.
- **CryoCore:** CryoCore is a fast, dense, and cooling-cost efficient cryogenic-optimal processor architecture running at 77K. To build CryoCore, we first develop CryoCore-Model (CC-Model), a validated cryogenic processor modeling framework which can accurately estimate the maximum clock frequency of processors running at 77K.

Second, by applying our modeling framework to two reference core models (i.e., high-performance Intel core vs. low-power ARM core), we identify two design principles in designing processors running at 77K. We first observe that it is important to minimize the core's dynamic power at the microarchitectural level, because the dynamic power significantly increases its cooling cost at 77K. We also observe that it is important to maintain the core's high clock frequency at the microarchitectural level to enable a further voltage and frequency scaling at 77K. Third, by following the principles, we design CryoCore, our cryogenic-optimal core architecture design. CryoCore first takes the high-performance reference core's pipeline depth and operating voltage to maintain its peak frequency. CryoCore then takes the low-power reference core's narrower pipeline width, and smaller and fewer microarchitecture units. As a result, CryoCore achieve high frequency while minimizing its dynamic power consumption at the microarchitecure level. Finally, by applying different voltage scalings, we propose two CryoCore designs which maximize the clock frequency or power efficiency, respectively. CryoCore significantly improves both single-thread and multi-thread performance with little total power consumption.

To evaluate the full cryogenic computer system, we incorporate our 77K-optimized DRAM, cache, and core design into a single cryogenic computer, and evaluate its performance against that of conventional room-temperature systems by using Gem5 simulator [12]. The simulation result shows that our cryogenic computer system equipped with our DRAM, cache, core architectures achieves significant performance gain over conventional room-temperature computers with lower power consumption. We also find that the speed-up of 77K memory with CryoCore is much higher than that of 77K memory or CryoCore only, which indicates their synergistic impact.

The rest of the dissertation is organized as follows. Chapter 2 illustrates a limitation of current computers and key benefits of cryogenic computing. Chapter 3 describes how we model and optimize DRAM devices running at 77K by using the key benefits. Chapter 4 describes how we model and develop the optimal cache design (CryoCache) running at 77K. Chapter 5 describes how we model and develop the 77K-optimized processor architecture (CryoCore), and shows performance evaluation of the full cryogenic computer system. Chapter 6 introduces the future research directions to realize the cryogenic computers. Chapter 7 discusses the related publications, and Chapter 8 concludes the dissertation.

Chapter 2

Background

2.1 Limitations of current computer scaling

To improve the computer performance with the same power budget, both Moore's Law and Dennard Scaling should be satisfied. With Moore's Law, architects can place more transistors on the same sized chip, which provides an opportunity to increase the chip performance. Dennard scaling aims to enable such transistor scaling while maintaining its power density, by reducing V_{dd} and V_{th} at the same ratio. However, as Dennard Scaling stopped in the early 2000s, the device scaling has been ineffective in increasing the chip frequency (and thus the chip performance) since then [14], as shown in Fig. 2.1.

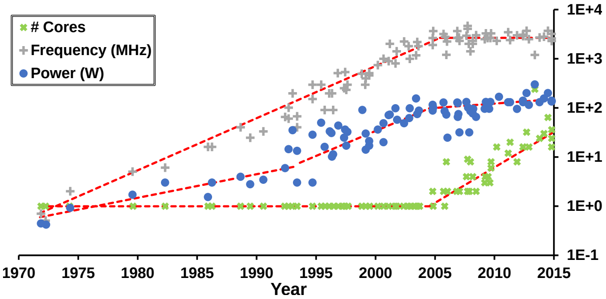


Figure 2.1: End of single-thread performance scaling due to the power wall problem

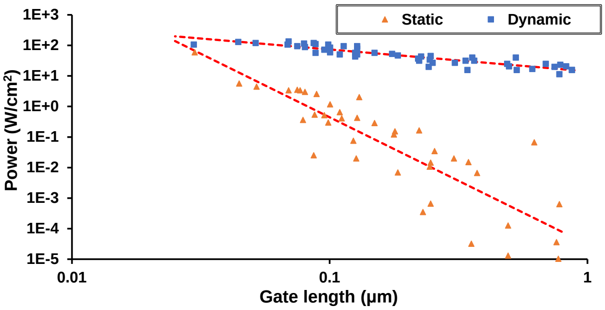


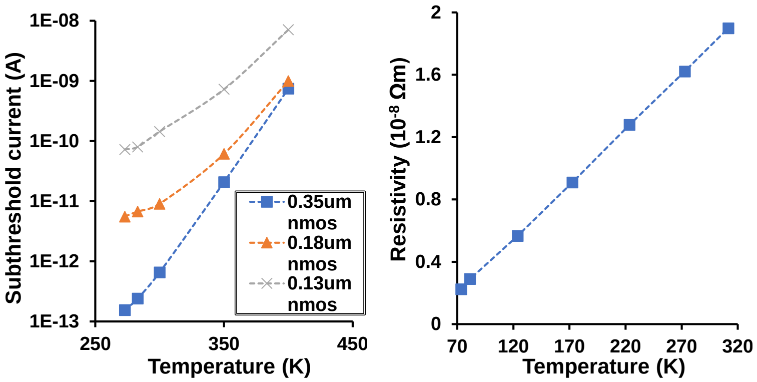
Figure 2.2: Steep increase of static power with shrinking device size

$$P_{\text{static}} = V_{\text{dd}} I_{\text{leak}}, \quad I_{\text{leak}} \propto e^{-\frac{qV_{th}}{kT}} \tag{2.1a}$$

$$P_{\text{dynamic}} \propto V_{\text{dd}}^2 f \tag{2.1b}$$

This problem originates from the prohibitively increasing static power when the transistor's V_{dd} and V_{th} are proportionally reduced to maintain the power density. Eq. (2.1a) shows that static power (P_{static}) is exponentially inversely proportional to V_{th} due to the increasing leakage current (I_{leak}) [57]. In this situation, to increase the chip frequency (f) while maintaining its dynamic power (P_{dynamic} in Eq. (2.1b)), the architects must reduce V_{dd} and V_{th}. But, the voltage reduction leads to the unacceptable increase of static power. As a result, the current generation of transistors suffer from the increased static power as well as the dynamic power [75] (as in Fig. 2.2). Therefore, computer architects have not been able to meaningfully improve the single-thread performance since the early 2000s, which indicates a critical performance challenge, the 'power wall' problem.

On the other hand, even if the power wall problem were magically resolved, the computer performance is bounded by the memory performance. As the memory access latency depends more on the wire latency than the transistor performance, even the successful device scaling cannot improve the memory performance. Therefore, regardless of the transistor speed, the computer's overall performance is eventually limited by



(a) Subthreshold current

(b) Wire resistivity

Figure 2.3: Benefits of cryogenic computing: (a) Exponentially decreasing subthreshold leakage current; (b) Linearly decreasing wire resistivity

the memory performance, which indicates another critical performance challenge, the ‘memory wall’ problem.

Therefore, in order to mitigate the power and memory walls, architects have proposed various approaches such as the deployment of slow multi-core designs [58, 74] (e.g., chip multiprocessor) and the information processing close to or within the memory (e.g., processor-in-memory). But, these circumventions can suffer from other challenges such as the parallelization overhead, the increasing on-chip power consumption and the requirement of radical architectural innovation.

2.2 Benefits of cryogenic computing

The concept of cryogenic computer systems has emerged as a highly promising idea to solve the power and memory wall challenges. The cryogenic computing literally means operating computers at extremely low temperatures such as 77K and 4K. These two representative temperatures, which categorize the domains of the cryogenic computing, can be achieved by applying liquid nitrogen (LN) and liquid helium (LH) respectively. Cryogenic computing is highly promising because it can resolve the fundamental challenges of the conventional computing.

One of the main advantages of cryogenic computing is the ability to eliminate static power. Eq. (2.1a) shows I_{leak} significantly decreases at cryogenic temperatures because reducing the temperature leads to the exponential decrease of the largest leakage component, subthreshold leakage [89] (as in Fig. 2.3a). As previously mentioned, static power is the main reason why V_{dd} and V_{th} cannot be reduced and the frequency cannot be increased. However, using cryogenic computing, architects can increase the chip frequency without increasing the dynamic power (=*solve the power wall problem*).

The other major advantage of cryogenic computing is the linearly decreasing wire resistivity. Fig. 2.3b shows the wire metal's resistivity (e.g., copper) reduces to 15% of the room temperature [16]. As the circuit delay is mainly determined by the RC delay (=resistance×capacitance), cryogenic computing can significantly improve the speed of circuit computations and data transfers of wires. In particular, because memory latency is dominated by the wire latency, architects can greatly improve the memory performance with the cryogenic computing (=*solve the memory wall problem*).

Therefore, cryogenic computing is a highly promising solution to effectively resolve both the power wall and memory wall problems. By using the cryogenic environment, architects can make an extremely low power system with higher performance, which was difficult to achieve in conventional computing.

Among two representative cryogenic temperatures, this work focuses on 77K because modern CMOS devices reliably operate at the temperature with relatively low cooling power cost. On the other hand, CMOS technology is considered rather inappropriate for 4K computing due to the higher cooling cost and the freeze-out effect of 4K environment [8].

Chapter 3

CryoRAM: Modeling and Optimizing DRAM at 77K

3.1 Motivation, Challenge, and Goal

3.1.1 Targeting a 77K-optimized DRAM

As the first step to realize cryogenic computers, we focus on memory devices (DRAM) due to the following reasons. First, the memory takes the highest benefits of cryogenic environment due to its huge static power and wire-delay portion in total latency. Second, even without other 77K-optimized devices (e.g., cache, core), it is feasible to cool memory devices isolated from the rest of systems, similar to the disaggregated memory architectures [64]. Finally, we believe that cryogenic DRAM modules can significantly improve the performance of memory latency-critical workloads [31], while significantly reducing the energy cost of operating modern datacenters equipped with an increasing number of memory modules [27].

3.1.2 Challenges in designing 77K-optimized DRAM

However, there exist several challenges in designing 77K-optimized DRAM. First, the absence of an architectural modeling tool is one of the major challenges. Even though

[1] CryoRAM was published on the 46th ACM/IEEE International Symposium on Computer Architecture (ISCA).

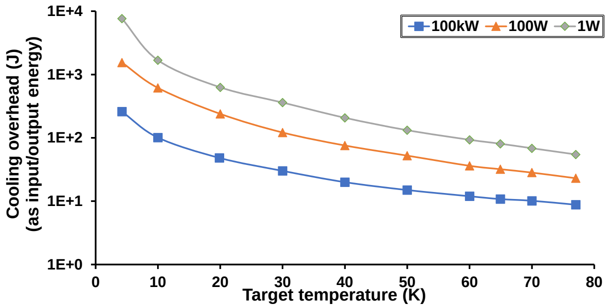


Figure 3.1: Cooling overhead as the relative amount of the input energy to reach the target temperatures. The legend indicates the efficiency of coolers as their cooling speed.

architects depend on high-level architecture modeling tools to design and evaluate their architectural innovations, to the best of our knowledge, a reliable cryogenic computer architecture modeling tool is currently unavailable.

Another major challenge is the non-trivial cost of cooling computer systems. Fig. 3.1 shows the overhead to achieve various target temperatures [50]. The cooling overhead indicates the relative amount of input energy required to remove unit heat (1J) from the cooling systems with three different coolers having different cooling efficiencies. The figure shows that the cooling overhead rapidly increases as the target temperature decreases, which makes the cryogenic computing more expensive. Therefore, cryogenic computer systems must be designed with comprehensive consideration of the cooling cost, and their performance and power advantages should outweigh the non-trivial cooling cost. Note that cooling overhead of 100kW-scale coolers is 9.65 at 77K. It indicates we should achieve at least 10.65 times lower power to achieve power advantage at the cryogenic temperature.

3.1.3 Research goal

We resolve the challenges in designing a 77K-optimized DRAM as follows. First, we develop CryoRAM, a validated cryogenic memory simulation framework, which can

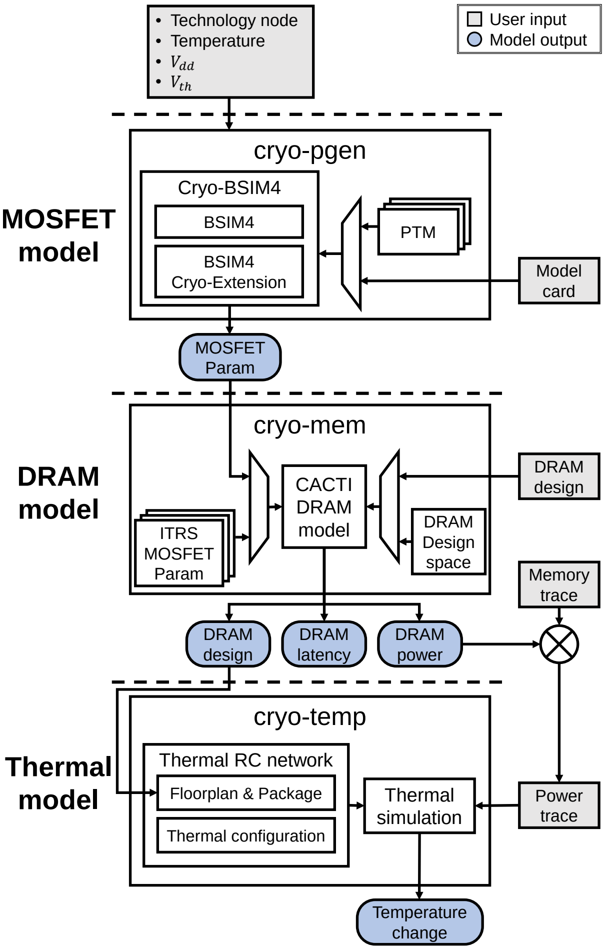


Figure 3.2: CryoRAM overview

derive a 77K-optimized DRAM design and evaluate its latency and power consumption. Second, we propose two 77K-optimized DRAM devices targeting for high performance and power efficiency. Finally, by using the DRAM devices, we show three promising case studies and show the potential of 77K-optimized DRAMs.

3.2 CryoRAM: Cryogenic DRAM Modeling Framework

In this section, we describe *CryoRAM*, our cryogenic memory modeling framework to explore the potential of the 77K-optimized memory. CryoRAM consists of three sub-models as shown in Fig. 3.2. First, the ***MOSFET model*** takes fabrication process information (e.g., model card) as inputs, and then derives the major electrical properties (e.g, MOSFET parameters) for a wide range of temperatures including 77K. Next, with the MOSFET parameters obtained from the MOSFET model, the ***DRAM model*** generates a target temperature-optimal DRAM design and reports its latency and power consumption. Finally, the ***thermal model*** reports dynamic temperature changes of the output DRAM design, while running target applications (e.g., by injecting memory/power traces). In this figure, the shaded square boxes and the rounded-square boxes indicate the inputs and outputs of each sub-model, respectively.

We implemented CryoRAM on top of existing models supporting only conventional temperatures by modifying them to accurately work for low temperatures. In this way, CryoRAM can be easily applied to existing computer architecture modeling tools. In the following sections, we explain the limitations of conventional models and how we added cryogenic supports to the models.

3.2.1 MOSFET model

Baseline MOSFET model

The MOSFET model is the most important component in CryoRAM because the main benefits of cryogenic computing come from the low-level MOSFET properties. As a baseline model, we use BSIM4 [113], a widely-used MOSFET model which takes a model card as an input, solves a set of equations, and derives the MOSFET parameters. The input model card is a set of parameters related to the MOSFET fabrication process (e.g., doping concentration, gate dielectric thickness). The output MOSFET parameters are high-level MOSFET electrical properties which affect the transistor performance significantly (e.g., on-channel current (I_{on}), subthreshold leakage current (I_{sub}), gate tunneling current (I_{gate})). However, BSIM4 does not provide accurate MOSFET parameters below 200K due to its simple temperature model.

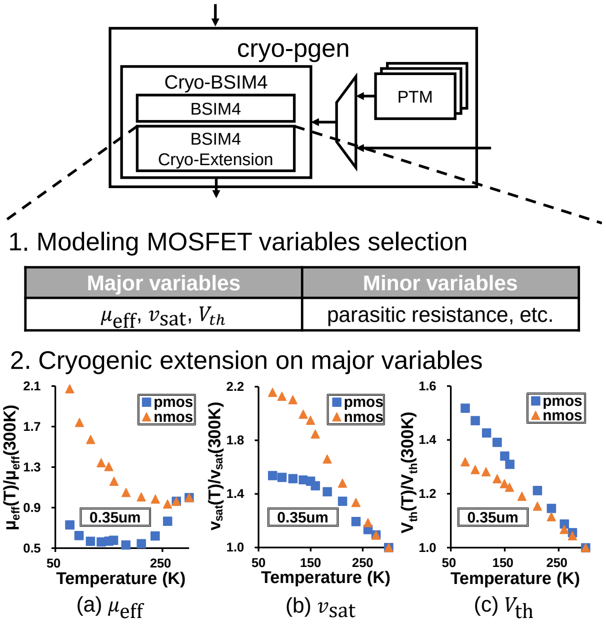


Figure 3.3: Cryogenic extension to the baseline MOSFET model: (a) Carrier mobility model; (b) Saturation velocity model; (c) Threshold voltage model

Cryogenic extension

We apply a cryogenic extension to BSIM4 as shown in Fig. 3.3. First, we select major fabrication-related, temperature-dependent MOSFET variables which would significantly affect the output parameters at low temperatures. Based on our analysis, we choose three major MOSFET variables as carrier mobility (μ_{eff}), carrier's saturation velocity (v_{sat}), and threshold voltage (V_{th}).

$$\mu_{\text{eff}} = \frac{U_0(T)}{\text{Surface Scattering}(T, E_{\text{eff}})} \tag{3.1}$$

Carrier mobility: Carrier mobility (μ_{eff}) is the ratio of the carrier velocity to the electric field strength in the channel, and a higher mobility increases the critical MOSFET properties such as I_{on} and I_{sub}. BSIM4 models the carrier mobility as Eq. (3.1), where U_0 indicates the carrier mobility with zero gate voltage. For non-zero gate voltages, carrier mobility becomes lower than U_0 due to the increased carrier collision at the interface (i.e., surface scattering). A lower temperature leads to a higher mobility thanks to the increased U_0, while decreasing the surface scattering.

Saturation velocity: The saturation velocity (v_{sat}) is the maximum velocity of the carriers inside the channel, and I_{on} increases with the velocity. The saturation velocity can be decreased when carriers and atoms collide. A lower temperature leads to a higher velocity by reducing the carrier-atom collisions.

Threshold voltage: The threshold voltage (V_{th}) is a minimum difference between the gate and source voltages to form a channel. A higher threshold voltage reduces I_{on} and I_{sub}. A lower temperature leads to a higher threshold voltage.

Implementing cryo-pgen

Our cryogenic MOSFET modeling tool, *cryo-pgen*, can generate the output MOSFET parameters at 77K as follows. First, it takes the target fabrication process information from a current room-temperature input model card available as two options: a vendor-driven model card and an open-source PTM model which supports from 180nm to 16nm at 300K [119]. Cryo-pgen can also adjust the process parameters automatically according to the given V_{dd}, V_{th} and target temperature.

Next, to estimate μ_0, v_{sat}, and V_{th} at low temperatures for the target technology, the baseline sensitivity data constructed from various literatures [90, 118] are provided to cryo-pgen as shown in Fig. 3.3. By assuming that the ratios of three terms at 300K and a low temperature T (i.e., $\mu_{\text{eff}}(T)/\mu_{\text{eff}}(300\text{K})$, $v_{\text{sat}}(T)/v_{\text{sat}}(300\text{K})$, $V_{\text{th}}(T)/V_{\text{th}}(300\text{K})$) are preserved across different technologies, cryo-pgen can estimate the value of each term for a target cryogenic temperature by referring to the model card's target process information at 300K and adjusting the value to 77K following the baseline data ratio.

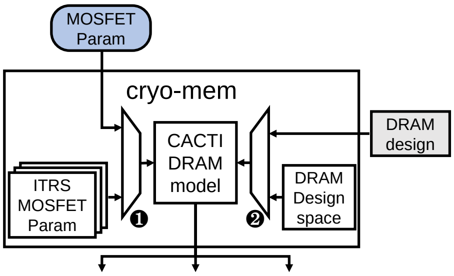


Figure 3.4: Cryogenic extension to DRAM model to add two input interfaces: ❶ MOSFET parameters; ❷ DRAM design

3.2.2 DRAM model

Baseline memory model

As a baseline memory modeling tool, CryoRAM uses CACTI [108], which takes memory specifications from users (e.g., memory capacity, the number of input/output ports), explores a large space of circuit-level designs, finds an optimal memory design for the underlying MOSFET parameters, and reports its latency and power.

However, CACTI cannot be directly applied to cryogenic memories due to two reasons. First, it uses ITRS [107] MOSFET parameters valid only at 300K-400K. Second, CACTI cannot apply different temperatures to a fixed target memory design.

Implementing cryo-mem

We implement our DRAM modeling tool, *cryo-mem*, by adding cryogenic extensions to CACTI-3dd [21]. Fig. 3.4 shows the overview of cryo-mem. First, we add an interface to CACTI to accept MOSFET parameters produced by cryo-pgen (Fig. 3.4❶). Second, we add an interface to CACTI to accept and fix a specific DRAM design while applying different temperatures (Fig. 3.4❷). In addition, we separately model peripheral circuit transistors and DRAM cell access transistors in our MOSFET model because DRAM access transistors use thicker gate dielectric than peripheral transistors to increase the data retention time.

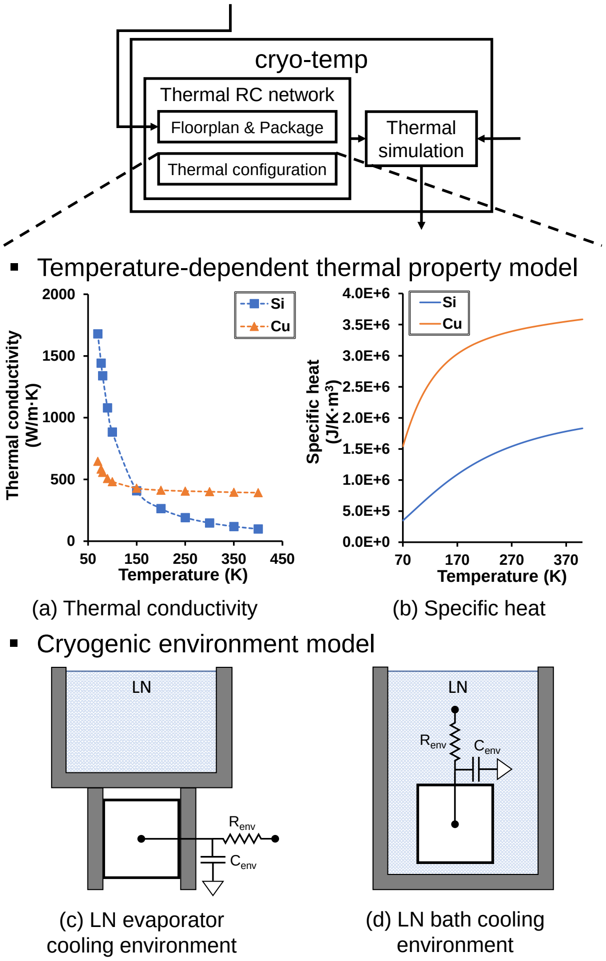


Figure 3.5: Cryogenic extension to the thermal model: (a),(b) Temperature-dependent thermal property models; (c),(d) Cryogenic cooling environment models

3.2.3 Thermal model

Baseline thermal model

As a baseline architecture-level temperature modeling tool, CryoRAM uses HotSpot [115] to take input power traces, construct a thermal resistor-to-capacitor (RC) circuit network, simulate the heat flow based on the RC delay model, and report the target device's dynamic temperatures.

However, the baseline HotSpot does not support cryogenic temperatures due to two reasons. First, the R and C values change significantly at low temperatures because the R-C critical thermal properties (e.g., thermal conductivity and specific heat) are highly sensitive to temperature changes as shown in Fig. 3.5. Second, HotSpot does not model the cryogenic-cooling method.

Implementing cryo-temp

We implement our temperature modeling tool, *cryo-temp*, by adding two cryogenic extensions to HotSpot. First, it collects the R-C critical thermal properties for primary materials (e.g., silicon (Si), copper (Cu)) from previous literatures [6, 32, 43] as shown in Fig. 3.5 (a) and (b). Cryo-temp then refers to the information at every temperature simulation step. Second, cryo-temp supports two cooling models, LN evaporator model and LN bath cooling model, as shown in Fig. 3.5 (c) and (d). The LN evaporator model indirectly cools a target device with temperature conduction via metal plates, which is assumed in Section 3.3.4. The LN bath cooling directly cools a target device by fully immersing it in LN (Fig. 3.5) which is assumed in Section 3.4.1.

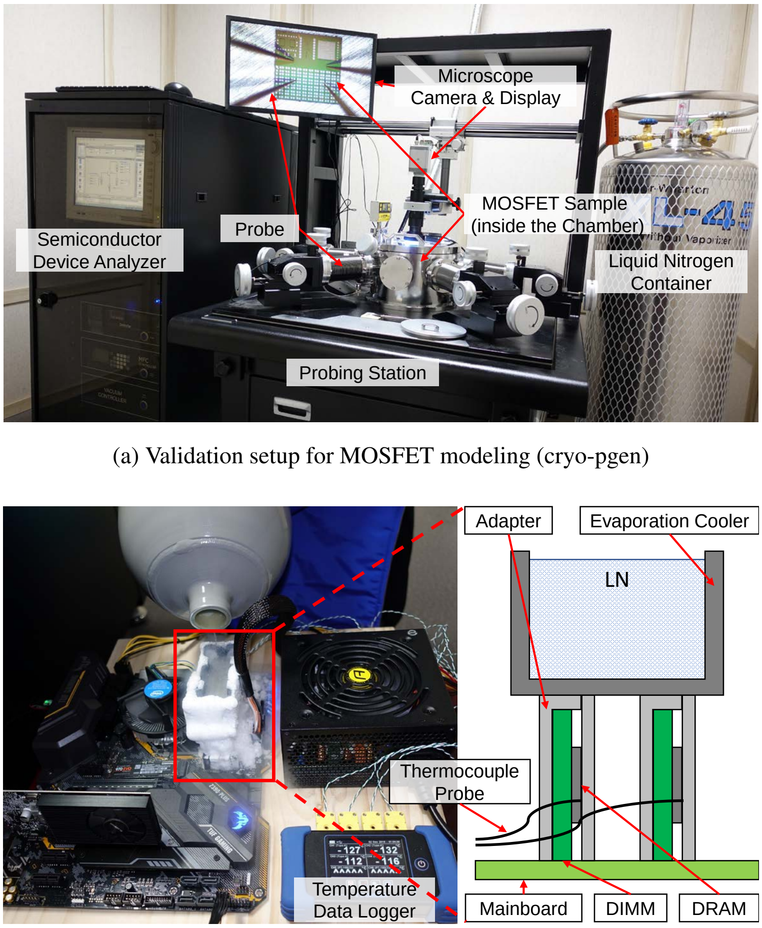
3.3 Model Validation

In this section, we validate our models by comparing their outputs with measurements. We first describe our experimental setup and then show the validation results.

3.3.1 Experimental setup

Fig. 3.6a shows our experimental setup to validate cryo-pgen. We use a custom-built MOSFET probing station consisting of a Keysight B1500A semiconductor device analyzer and an LN-based cooling unit. By placing our MOSFET sample inside the station's chamber, we can measure its gate, source, and drain currents via the probes.

Fig. 3.6b shows our experimental setup to validate cryo-mem and cryo-temp. We

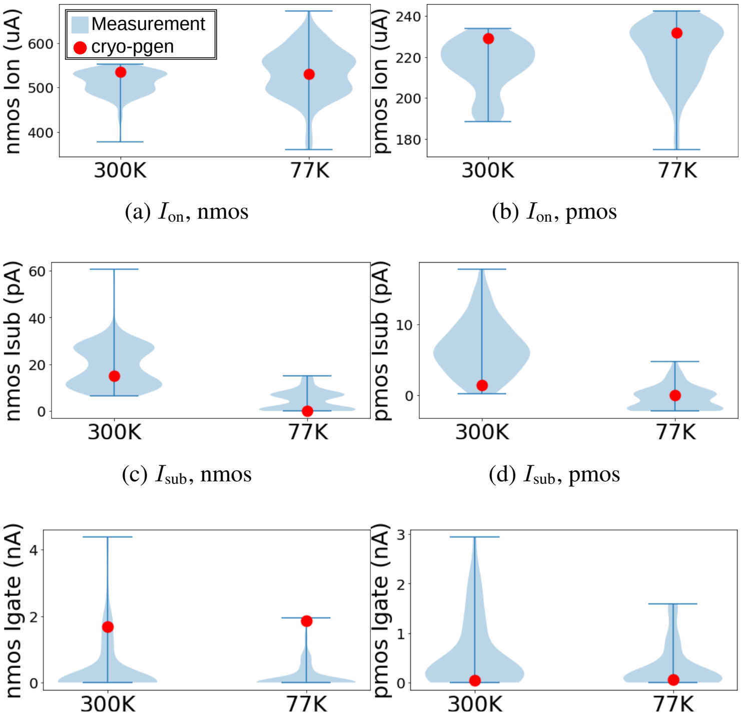


(a) Validation setup for MOSFET modeling (cryo-pgen)

(b) Validation setup for DRAM modeling (cryo-mem, cryo-temp)

Figure 3.6: Experimental validation setup for (a) MOSFET modeling (cryo-pgen) and (b) DRAM modeling (cryo-mem, cryo-temp)

construct a sample computer board using various commodity parts (i.e., Intel Z390 mainboard, Intel i7-8700 CPU, and two Micron DDR4 8G PC4-21300 DIMMs). With this setup, we can reduce the DIMM's temperature by applying LN to the container placed on top of them. It also allows us to control the board's memory clock frequency with Intel XMP.



(a) I_{on}, nmos

(b) I_{on}, pmos

(c) I_{sub}, nmos

(d) I_{sub}, pmos

(e) I_{gate}, nmos

(f) I_{gate}, pmos

Figure 3.7: cryo-pgen validation results: real measurements (violin distribution) vs. cryo-pgen outcomes (dot)

3.3.2 MOSFET model validation

We validate our MOSFET model by comparing the three MOSFET parameters reported by cryo-pgen (i.e., I_{on}, I_{sub}, I_{gate}) with the real measurements obtained from 220 180nm MOSFET samples. Fig. 3.7 shows the validation results for cryo-pgen. The violin-like distributions indicate the measurements from 200 MOSFET samples with their variance, whereas the dots indicate the results of cryo-pgen. The graphs show that cryo-pgen accurately models the target MOSFET parameters by placing the dots inside the distributions. These validations also provide the projections of the MOSFET parameters when the temperature decreases: such as slightly increased I_{on}, significantly reduced I_{sub}, and constant I_{gate}.

I_{gate} is at least 10 times higher than I_{sub} and dominates overall leakage current in 180nm technology (Fig. 3.7). However, I_{gate} has become 100 times lower than I_{sub} since high-K materials were adopted as a gate dielectric in MOSFETs below 45nm [57]. For example, with 22nm PTM, I_{sub} and I_{gate} per unit gate length (1um) are 85nA/um and 0.5nA/um, respectively. In modern technology, I_{sub} is dominant in overall leakage current. As I_{sub} is practically eliminated at 77K, the overall leakage current greatly reduces.

3.3.3 DRAM model validation

In this section, we validate cryo-mem for its DRAM performance prediction. We measure the speed-up of the memory in the cryogenic temperature. We perform the validation for the power in the following section (Section 3.3.4).

We measure the DRAM performance as the maximum DRAM frequency at 160K and 300K. Note that 160K is the minimum temperature achievable with the LN evaporation cooler while Memtest86+ [13] is running (Fig. 3.6b). We sweep the DRAM clock frequency to find the maximum frequency at which the system still reliably operates. At 300K, Our DRAM reliably operates at up to 2666MHz frequency. At 160K, the maximum frequency is safely increased to 3333Hz. These results imply that the speed-up lies in the range of 1.25 to 1.30.

Next, we take the results to validate the memory performance prediction of cryo-mem. Using cryo-mem, we derive the 300K-optimized DRAM circuit design and estimate its latency at 160K. Cryo-mem predicts that 300K-optimized DRAM becomes 1.29 times faster at 160K. Since the prediction is within the range of measurements, we conclude that the experiment shows the accuracy of cryo-mem.

3.3.4 Thermal model validation

We validate cryo-temp by comparing the DRAM temperature in the real system (Fig. 3.6b) with the cryo-temp's prediction with the LN evaporator model. We run several SPEC

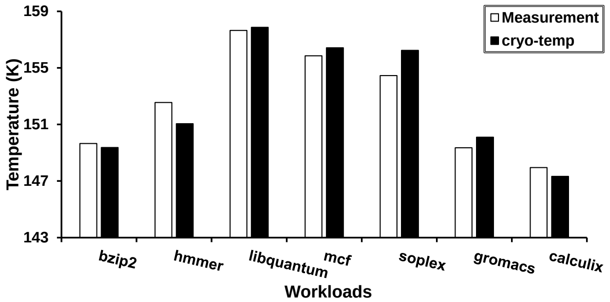


Figure 3.8: Cryo-temp validation result comparing temperature change measurements with cryo-temp's predictions for SPEC2006 workloads

CPU2006 workloads (bzip2, hmmer, libquantum, mcf, soplex, gromacs and calculix) [41] and measure the DRAM temperature using the temperature data logger. For the cryo-temp, we generate the power trace for each workload by combining cryo-mem's power output with the memory traces extracted from gem5 [12] simulation.

Fig. 3.8 shows the accuracy of cryo-temp by comparing the measured DRAM temperature and the cryo-temp's prediction for each workload. First, the graphs show that cryo-temp well match the measurements. Second, the small errors observed (i.e., 0.82K on average and 1.79K in maximum) are tolerable because few-Kelvin of errors can be easily introduced during the measurement process and they do not affect the overall prediction accuracy.

It should be also noted that the accuracy of cryo-temp indirectly validates the cryo-mem's power prediction. In the validation process of cryo-temp, we use the cryo-mem's output power prediction to generate the input power traces and feed them to cryo-temp. Therefore, the cryo-temp validation implicitly validates the accuracy of cryo-mem as well.

3.4 77K-Optimized DRAM Design

Using validated CryoRAM, we show the result of two experiments in this section. First, by using cryo-temp, we check the temperature variation when DRAM operates

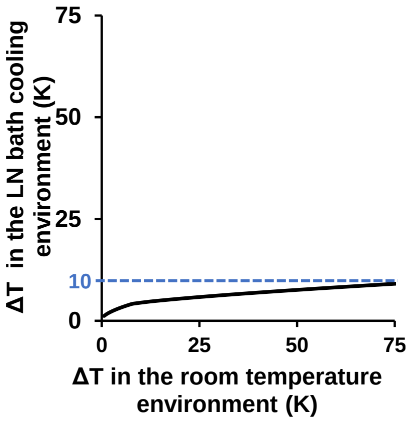


Figure 3.9: Temperature change comparison between the room temperature environment and LN bath cooling environment

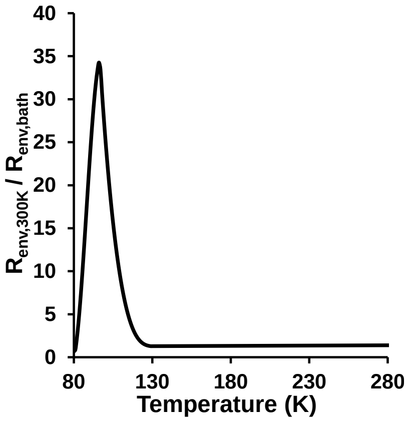


Figure 3.10: High ratio of R_{env} between the room temperature and the LN bath cooling environment ($R_{\text{env,300K}}/R_{\text{env,bath}}$) near 77K

in the LN bath cooling environment. Next, we use cryo-pgen and cryo-mem to show the potential of cryogenic DRAM in terms of power and latency.

3.4.1 Maintaining the cryogenic temperature

As every benefit of cryogenic computing comes from low temperature, it is crucial to ensure that the cryogenic memory remains in low-temperature ranges. Therefore, we simulate the temperature change of DRAM in the LN bath cooling environment using cryo-temp. We also provide the simulation in the room temperature environment using the same power trace for the comparison.

In Fig. 3.9, DRAM with the LN bath cooling shows negligible temperature variation (below 10K) whereas the counterpart's temperature rises over 75K. Such a huge difference results from the low R_{env} of the LN bath cooling [52]. Note that R_{env} is a thermal resistance indicating the heat transfer between the device and the surrounding environment. Therefore, low R_{env} indicates the high heat transfer speed due to the reduced thermal RC delay.

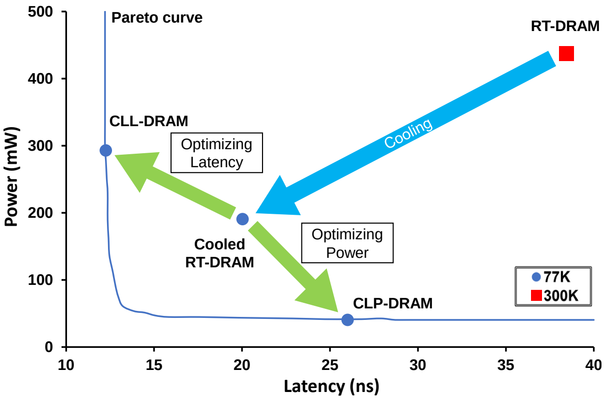


Figure 3.11: Optimal DRAM design exploration with V_{dd} and V_{th} sweeping

Fig. 3.10 shows that the ratio of R_{env} between the room temperature and the LN bath cooling environment (=$R_{\text{env,300K}}/R_{\text{env,bath}}$) is significantly high near 77K (about 35 in maximum). It means that heat transfer is up to 35 times faster in the LN bath cooling compared to the room temperature. Such high heat transfer speed prevents the temperature from increasing in the LN bath cooling environment.

Based on the result, we conclude that the low temperature (77K) would be maintained well in the cryogenic environment. We thus focus only on 77K in the following sections without consideration about run-time temperature changes.

3.4.2 Deriving two cryogenic DRAM devices

In this section, we show potentials of the 77K-optimized DRAM in terms of latency and power consumption. The latency in this section means the random access latency. The power consumption is sum of the static power and the dynamic power, weighted by the memory access rate from Micron DDR4 power calculator [70]. We also conservatively model DRAM refresh power, with the same retention time of commercial DRAM (64ms).

As Fig. 3.11 shows, when the commercial DRAM (room temperature DRAM; RT-DRAM) is cooled down to 77K (cooled RT-DRAM), the latency is reduced to 48.9%

and power is reduced to 43.5%. Such great improvements result from the reduction of wire resistivity and static power, respectively. However, we can achieve more significant improvement by scaling threshold voltage (V_{th}) and operating voltage (V_{dd}) at 77K. The near-zero leakage current enables the V_{th} reduction, which is not achievable due to exploding static power at the room temperature. We explore 150,000+ DRAM designs with different V_{dd} and V_{th} to find optimal cryogenic DRAM design, and obtain a latency-power Pareto optimal curve as the result (Fig. 3.11). Among various points in the curve, we discuss two representative DRAM designs, the power-optimal design, and the latency-optimal design.

Cryogenic Low-Power DRAM (CLP-DRAM): We scale down V_{dd} and V_{th} by half to reduce dynamic power at 77K. As a result, we can make ultra-low power DRAM (CLP-DRAM) in a cryogenic environment. Fig. 3.11 shows that CLP-DRAM consumes only 9.2% of power compared to that of RT-DRAM. This huge reduction comes from reduced dynamic power and eliminated static power. At the same time, its latency is only 65.3% (1.53 times faster) compared to that of RT-DRAM.

Cryogenic Low-Latency DRAM (CLL-DRAM): Using high V_{dd} with low V_{th} can greatly improve I_{on} of MOSFET, and it reduces DRAM latency significantly. We set V_{dd} the same as RT-DRAM's and scale down V_{th} by half. As a result, we can make ultra-low latency DRAM (CLL-DRAM) for cryogenic environments. The latency of CLL-DRAM is significantly reduced while keeping the power consumption less than RT-DRAM (Fig. 3.11). CLL-DRAM is 3.80 times faster than RT-DRAM, thanks to high I_{on} and reduced wire resistivity.

3.5 Single-Node Level Case Studies

In Section 3.4.2, we proposed a CLL-DRAM specialized for latency and a CLP-DRAM specialized for power. In this section, by using these cryogenic DRAMs, we show two case studies on a single-node level. First, we cover the IPC speed-up through CLL-DRAM and then show the power reduction through CLP-DRAM.

CPU specification		
Cores	Based on the Intel i7-6700 (3.5GHz)	
LLC	12MB, 16way set-assoc, shared, 42cyc (=12ns)	
DRAM access latency[1]		
RT-DRAM	60.32ns (with tRAS=32ns, tCAS=tRP=14.16ns)	
CLL-DRAM	15.84ns (with tRAS=8.4ns, tCAS=tRS=3.72ns)	
DRAM power (per chip)		
	Static power	Dynamic energy
RT-DRAM	171 mW	2 nJ/access
CLP-DRAM	1.29 mW	0.51 nJ/access

Table 3.1: Parameter setup for single-node level case studies

3.5.1 Evaluation setup

In the case studies, we use gem5 timing simulator [12] with the configuration specified in Table 3.1. We set CPU-related parameters based on the Intel i7-6700 processor. For DRAM access latency and power, we use values derived from Section 3.4.2 as summarized in Table 3.1. Baseline of the two case studies is a node with the RT-DRAM. Also, we use 12 workloads chosen from the SPEC CPU2006 benchmark [41]. To focus on DRAM devices only, we assume that maximum DRAM channel bandwidth is the same as the maximum DRAM bandwidth, the inverse of DRAM device latency.

3.5.2 IPC speed-up with CLL-DRAM

Fig. 3.12 shows the IPC speed-up in a node with the CLL-DRAM. With 3.8 times faster DRAM access speed over RT-DRAM, memory-intensive workloads (e.g., mcf, libquantum) benefit from the low L3 cache miss penalty. On the other hand, the IPC of workloads such as calculix and gcc remains nearly constant due to their low memory access rate. Despite including such workloads, CLL-DRAM improves the IPC by 24% on average (CLL-DRAM in Fig. 3.12).

Additionally, we achieve a more significant speed-up by eliminating the L3 cache. Note that L3 cache latency is 12ns and CLL-DRAM's latency is 15.84ns in Table 3.1. Since CLL-DRAM's latency becomes comparable to the L3 cache latency, it is better

[1] DRAM access latency is calculated by the sum of tRAS, tCAS, and tRP

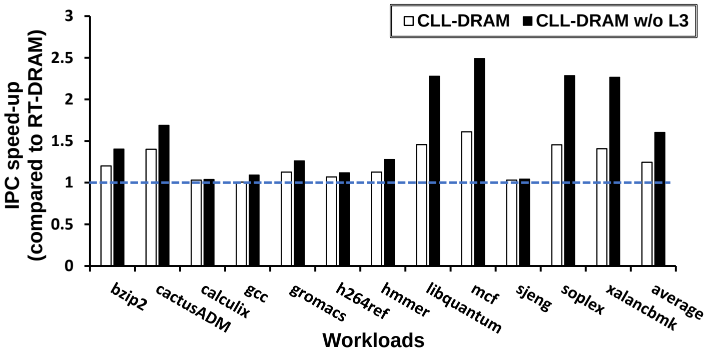


Figure 3.12: IPC speed-up in CLL-DRAM nodes (with L3 cache or without L3 cache) for SPEC CPU2006 workloads

to bypass the L3 cache and directly access CLL-DRAM. Fig. 3.12 shows the IPC speed-up of the CLL-DRAM without L3 cache (CLL-DRAM w/o L3). The IPC of the node without L3 cache is increased by 60% on average by eliminating L3 cache miss penalties. Especially, for memory-intensive workloads (i.e., libquantum, mcf, soplex, and xalancbmk), the figure shows the speed-up of 2.3 on average and 2.5 in maximum, respectively. In addition, because the L3 cache occupies significant area of a chip, removing the L3 cache also allows architects to search for new designs, such as adding more cores on the chip.

3.5.3 DRAM power reduction with CLP-DRAM

Fig. 3.13 shows the DRAM power consumption of a node with the CLP-DRAM. Note that the power consumption of Fig. 3.13 is normalized to that of a node with the RT-DRAM. To calculate the DRAM power, we add the dynamic power and the static power using the memory access rate obtained from each workload. DRAM power consumption is reduced to 6% on average and some workloads (e.g., calculix, gcc, sjeng) show significant power reduction (more than 100 times). These workloads have a small memory access rate as the line graph in Fig. 3.13 shows. Therefore, the static power dominates their DRAM power consumption. For such cases, the CLP-DRAM's

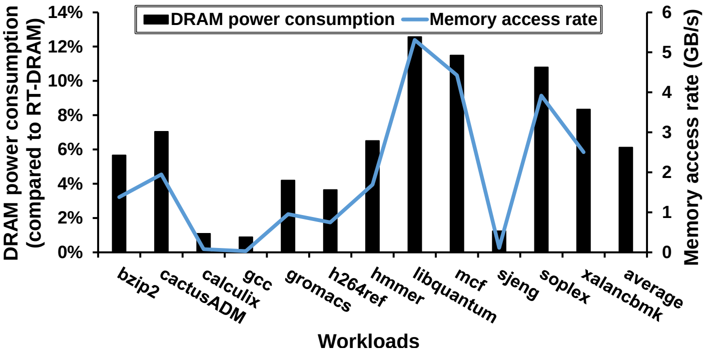


Figure 3.13: Normalized DRAM power consumption of a CLP-DRAM node and the memory access rate for SPEC CPU2006 workloads

negligible static power significantly reduces the DRAM power consumption. Without these workloads, the DRAM power consumption is reduced to 12.7% on average.

3.6 Datacenter Level Case Study

The cryogenic computing is expected to be used primarily in datacenters because it requires huge cooling facilities to maintain a cryogenic environment. In Section 3.5.3, we show CLP-DRAM's great power reduction which can significantly decrease datacenters' total power cost. However, replacing all DRAMs in a datacenter with CLP-DRAMs incurs huge replacement overhead. Therefore, we need a solution to achieve high power reduction while using a minimal number of cryogenic memories.

In this section, we propose a Cryogenic Low-Power Architecture for datacenters (CLP-A), as a use case of CLP-DRAM. CLP-A achieves huge power reduction by adding only the small number of CLP-DRAM to conventional datacenters.

To show its potential, we first introduce CLP-A by explaining its key idea and how CLP-A operates (Section 3.6.1). Next, we show CLP-A's DRAM power reduction in Section 3.6.2. Finally, we evaluate CLP-A's datacenter-level total power reduction using our cryogenic power model (Sections 3.6.3 and 3.6.4).

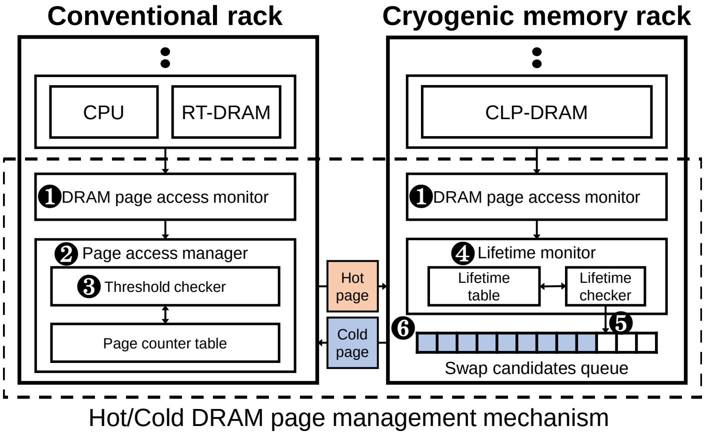


Figure 3.14: Cryogenic low-power architecture (CLP-A) overview

3.6.1 Cryogenic Low-Power Architecture (CLP-A)

CLP-A overview

The basic idea of CLP-A is to migrate hot pages to low-power CLP-DRAM in order to minimize DRAM power consumption. Fig. 3.14 shows the CLP-A's overview. CLP-A consists of two types of racks: conventional racks (at 300K) and cryogenic memory racks (at 77K) where CLP-DRAM resides in. Using its hot page management mechanism (Section 3.6.1), CLP-A moves hot pages to the disaggregated CLP-DRAM. Thanks to the localities, there are only a few hot pages at a certain point. As a result, CLP-A needs only a few CLP-DRAMs to store the hot pages. Since most of the memory accesses target hot pages, CLP-A can successfully reduce the total DRAM power consumption, using only a small amount of low-power CLP-DRAMs.

CLP-A's page management mechanism

We make the hot page management mechanism inspired by [82]. Before introducing its detailed mechanism, we summarize important terms as follows.

Page (or DRAM page): As CLP-A applies the page management on the DRAM page level, the term, ”page” in this section, means a DRAM page (i.e., DRAM row).

Hot page: The hot page means a frequently accessed page, and CLP-A categorizes a page as a hot page when the page's counter value exceeds a ”threshold”. CLP-A stores a counter for each page and increases the target page's counter in every memory access. The counters are reset after ”counter lifetime” from the last access.

Cold page: A hot page becomes a cold page if the page is not accessed during ”hot page lifetime.” Every page starts as a cold page.

Fig. 3.14 shows the CLP-A's detailed page management mechanism. Every rack has a DRAM page access monitor which monitors every memory access and notifies the accesses to page access manager or lifetime monitor ❶. In conventional racks, for every memory access, a page access manager increases the corresponding counter in a page counter table ❷. Counters are reset when counter lifetime elapses from the last access. Consequently, the counter in the table infers the number of accesses to the corresponding page since the last counter reset. If the counter of a page exceeds a threshold ❸, the threshold checker categorizes the page as a hot page and migrates it to the CLP-DRAM.

In cryogenic memory racks, the lifetime monitor manages a lifetime of the hot page. For every page access, the monitor resets the lifetime of the hot page ❹. A lifetime checker finds hot pages whose lifetime is expired and registers it in a swap candidates queue ❺. When a new hot page comes from a conventional rack, the cold page in the queue is swapped out ❻. If there is no page in the swap candidate queue while all CLP-DRAMs are full, CLP-A does not swap pages until a new page enters the queue.

As we mentioned, the above mechanism needs only the small number of CLP-DRAM to hold enough hot pages. Consequently, CLP-A significantly reduces power with minimal cooling and additional equipment costs.

Mechanism specification		
DRAM latency	Based on Micron MT40A2G4	
	Latency	Energy
Swap overhead	1.2μs	8 × (RT-DRAM access energy + CLP-DRAM access energy)
Counter lifetime	200μs	
Hot page lifetime	200μs	
Hot page ratio	7%	

Table 3.2: Parameter setup for CLP-A

3.6.2 CLP-A: DRAM power evaluation

To evaluate CLP-A, we show the CLP-A's DRAM power reduction compared to a conventional datacenter which consists of RT-DRAMs only. For the experiment, we first implement an architectural memory trace-based simulator and then simulate CLP-A's hot/cold DRAM page management mechanism for eight SPEC CPU2006 workloads [41].

In order to define CLP-A's mechanism in detail, we configure parameters as shown in Table 3.2. Detailed parameter setup procedures are provided as follows. First, we set the access latency of disaggregated CLP-DRAM to the same as the latency of RT-DRAM by assuming that CLP-DRAM's speed-up offsets the inter-rack interconnect overhead. Next, we consider the page swap overhead. We set the swap latency to 1.2μs with reference to [82] and conservatively assume that the RT-DRAM serves memory accesses during the page swap. The swap energy overhead is 8×(RT-DRAM access energy + CLP-DRAM access energy) because moving a 512B DRAM page requires eight 64B-CAS operations. Lastly, we set the counter lifetime, hot-page lifetime, and the amount of CLP-DRAMs by sweeping these parameters and choosing the optimal values. As a result, we take 7% as the ratio of CLP-DRAMs to the total DRAMs in CLP-A and 200μs as the counter lifetime and the hot page lifetime.

Fig. 3.15 shows the DRAM power consumption in CLP-A. There is a large difference in the DRAM power among the workloads. For example, CLP-A reduces 72% of

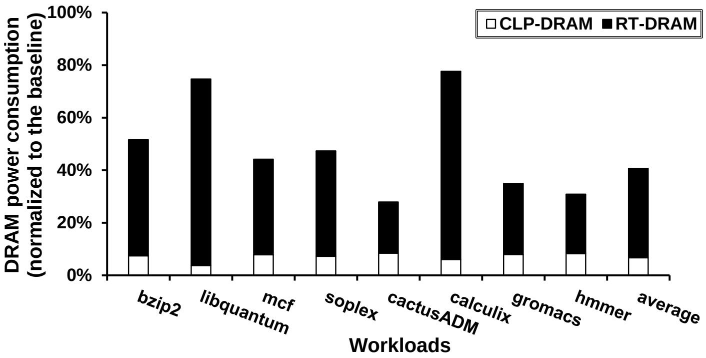


Figure 3.15: DRAM power consumption of CLP-A for SPEC CPU2006 workloads normalized to that of conventional datacenter

DRAM power consumption for cactusADM, but only 23% of the power for calculix. Such a large difference results from the memory access pattern of each workload. In the CLP-A's page management mechanism, moving a page to the CLP-DRAM takes 1.2μs after the page is categorized as a hot page. If a workload does not access the page after the migration, it cannot benefit from the CLP-DRAM but consumes more power due to the swap overhead. However, even with such workloads, CLP-A exhibits large power reduction (59% on average). With only the small number of CLP-DRAM (7% of total DRAMs), CLP-A significantly reduces the DRAM power consumption compared to the conventional datacenter.

3.6.3 Cryogenic datacenter power modeling

The energy efficiency of CLP-DRAM does not guarantee the efficiency in the overall system. The high cooling overhead of the cryogenic system may incur higher cooling cost even with the reduced power consumption. Therefore, we need to conduct datacenter-level power evaluation considering the high cooling overhead. However, to the best of our knowledge, there is no power model for cryogenic datacenters.

To make the model, we first model the power of conventional datacenters which covers the room-temperature parts of CLP-A (conventional rack in Fig. 3.14). Next,

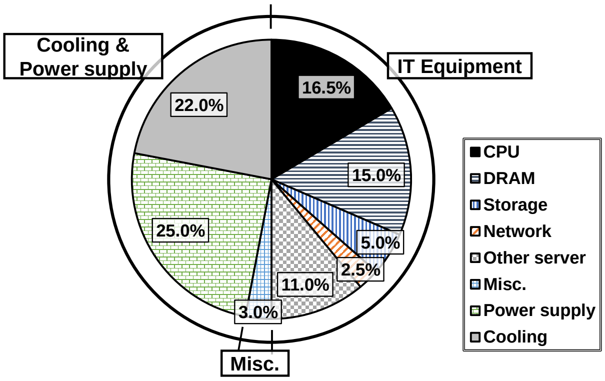


Figure 3.16: Power breakdown of conventional datacenter

we extend it to CLP-A by adding the cryogenic-cooling cost model. We use the derived power model in Section 3.6.4 to evaluate CLP-A's datacenter-level power reduction.

Conventional datacenter power modeling

Fig. 3.16 shows a typical datacenter's power breakdown [27]. The datacenter power can be divided into three categories: IT Equipment (50%), Cooling/Power Overhead (47%), and Misc. (3%).

IT Equipment: IT Equipment means the power consumption of IT components (e.g., CPU, DRAM, storage, network) which accounts for the largest portion in the conventional datacenter's power.

$$Cooling = C.O. \times \text{IT Equipment} \tag{3.2a}$$

$$Power\ Supply = P.O. \times \text{IT Equipment} \tag{3.2b}$$

Cooling & Power Supply: Cooling is the power for cooling IT components and Power Supply is the power loss occurred during power supply. We categorize them as the same group (Cooling & Power Supply) because they have a similar relationship with IT Equipment. To model Cooling & Power supply, we use the linear model shown as Eq. (3.2). $C.O.$ in Eq. (3.2a) means the cooling overhead discussed in Section 3.1.2.

$P.O.$ in Eq. (3.2b) is power overhead, the amount of wasted energy while supplying unit energy (1 J) to the IT components. Such a linear model is conservative because, in the real world, both of them decrease faster than linear as the IT Equipment decreases at 300K [27]. As our evaluations only consider the IT power decreasing cases, this linear model always exaggerates the model's Cooling & Power Supply at 300K.

Misc.: Misc. is the power consumed for miscellaneous reasons (e.g., lighting). Therefore, Misc. is not related to the power consumption of the other two categories.

$$\begin{aligned}
&\text{Conventional datacenter power (at room temperature)}\\
&= \text{IT Equipment} + \text{Cooling \& Power Supply} + \text{Misc.}\\
&= \text{IT Equipment} + (C.O._{\text{300K}} + P.O._{\text{300K}}) \cdot \text{IT Equipment} + \text{Misc.}\\
&= \text{IT Equipment} + (\frac{22}{50} + \frac{25}{50}) \cdot \text{IT Equipment} + \text{Misc.}\\
&= 1.94 \cdot \text{IT Equipment} + \text{Misc.}
\end{aligned} \tag{3.3}$$

By applying Eq. (3.2), the total power consumption in a conventional datacenter is summarized as Eq. (3.3). Note that $C.O._{\text{300K}}$ is the ratio of Cooling (22%) and IT Equipment (50%) in Fig. 3.16. $P.O._{\text{300K}}$ is the ratio of Power Supply (25%) and IT Equipment.

Cryogenic-cooling cost analysis

In this subsection, we analyze the cryogenic-cooling cost for CLP-A. The cryogenic-cooling cost consists of two parts: one-time cost and recurring cost.

One-time cost: One-time cost consists of LN cost and facility cost. We assume LN recycling "stinger system [10]", which requires only a small LN cost for the initial setup (0.5 $/L). The facility cost is proportional to the size of the computing environment. However, they are only one-time costs, therefore the cooling cost is dominated by the recurring cost.

Recurring cost: Recursively incurred cooling-power consumption (i.e., electricity) accounts for the majority of the cooling cost. Cooling power for cryogenic datacenter also can be modeled as Eq. (3.2a), however, $C.O._{77\mathrm{K}}$ is much higher than the conventional datacenter counterpart. For the cooling overhead targeting 77K, we conservatively use 100kW cryo-cooler's value ($C.O._{77\mathrm{K}}$ = 9.65 from Fig. 3.1) to estimate the cost of modern 10MW system [50]. Note that the cryogenic cooler's cooling overhead decreases as the cooling speed of the cooler increases (as shown in Fig. 3.1).

Cryogenic datacenter power modeling

Finally, by combining all together, we make the cryogenic data center power model.

$$\begin{aligned}
&\text{Cryogenic datacenter power}\\
&= (\text{RT-IT} + \text{Cryo-IT}) + (\text{RT-C/P} + \text{Cryo-C/P}) + \text{Misc.}\\
&= (\text{RT-IT} + \text{RT-C/P} + \text{Misc.}) + (\text{Cryo-IT} + \text{Cryo-C/P}) && (3.4a)\\
&= (1.94 \cdot \text{RT-IT} + \text{Misc.}) + (1 + C.O._{77\mathrm{K}} + P.O._{77\mathrm{K}}) \cdot \text{Cryo-IT}\\
&= 1.94 \cdot \text{RT-IT} + (1 + 9.65 + \frac{22}{50}) \cdot \text{Cryo-IT} + \text{Misc.} && (3.4b)\\
&= 1.94 \cdot \text{RT-IT} + 11.09 \cdot \text{Cryo-IT} + \text{Misc.} && (3.4c)
\end{aligned}$$

In the cryogenic datacenter power model, both IT Equipment and Cooling & Power supply are divided into room temperature parts (RT-IT, RT-C/P) and cryogenic parts (Cryo-IT, Cryo-C/P). We first replace the "RT-IT + RT-C/P + Misc." to "$1.94 \cdot$ RT-IT + $Misc.$" using Eq. (3.3) and apply our Cooling & Power Supply model (Eq. (3.2)) to Eq. (3.4a). As mentioned in Section 3.6.3, we set $C.O._{77\mathrm{K}}$ to 9.65. The power overhead at 77K ($P.O._{77\mathrm{K}}$) is the same as $P.O._{300\mathrm{K}}$ ($= \frac{22}{50}$) because cryogenic IT components also utilize the existing power supply path (Eq. (3.4b)). As a result, the total power consumption of the cryogenic datacenter is modeled as Eq. (3.4c).

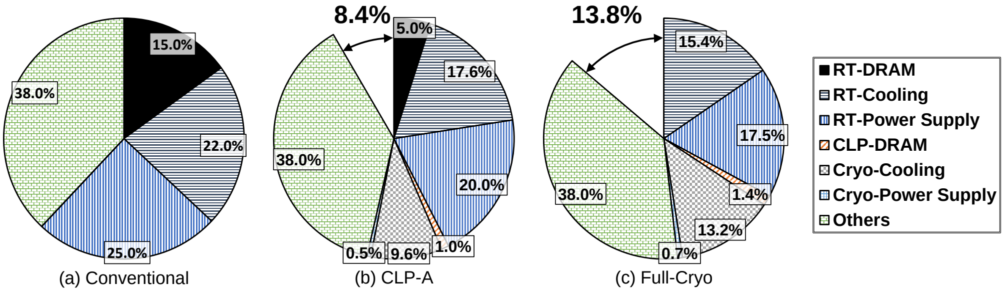


Figure 3.17: Total power consumption of three datacenters normalized to the conventional datacenter. ”Others” means the sum of power except DRAM, and Cooling & Power Supply. (a) Conventional datacenter with only RT-DRAM; (b) CLP-A with 7% CLP-DRAM; (c) Cryogenic datacenter with only CLP-DRAM.

3.6.4 CLP-A: Total power cost evaluation

Based on the power model in Section 3.6.3, we evaluate CLP-A’s datacenter-level power consumption. Fig. 3.17 compares three types of datacenter: conventional datacenter with only RT-DRAM (Conventional), CLP-A with 7% CLP-DRAM (CLP-A), and cryogenic datacenter with only CLP-DRAM (Full-Cryo). All values are normalized to the power of conventional datacenters. Note that ”Others” means the sum of power except DRAM power and Cooling & Power Supply.

In CLP-A, the total power cost is reduced by 8.4% (Fig. 3.17(b)). The result can be explained by reduced RT-DRAM power and CLP-DRAM’s low power consumption. First, CLP-A significantly reduces RT-DRAM power (from 15.00% to 5.0%) by migrating hot pages to low-power CLP-DRAMs (as shown in Fig. 3.15). The power reduction in RT-DRAM also incurs the sum of RT-Cooling and RT-Power Supply to decrease from 47.0% to 37.6%. In summary, RT-DRAM’s power reduction greatly reduces CLP-A’s total power cost by 19.4%. Second, CLP-DRAM’s low power consumption offsets the high cooling overhead. Note that Cryo-Cooling is proportional to CLP-DRAM’s power consumption (Eq. (3.2a)). Even with the high cooling overhead, cryogenic cooling cost (Cryo-Cooling; 9.6%) does not exceed the RT-DRAM’s power reduction thanks to the extremely low-power CLP-DRAM.

Compared with Full-Cryo, CLP-A shows a huge benefit in terms of the power reduction per replacement overhead. Even though CLP-A uses only the small number of CLP-DRAMs (7%), its total power reduction (8.4%) does not significantly differ from the Full-Cryo's case (13.82%). Note that Full-Cryo provides ideal power reduction of a cryogenic-memory datacenter by replacing all DRAMs with the CLP-DRAMs.

The power consumption of DRAM will continue to increase due to the increasing number of memory-intensive workloads in datacenters [55]. Therefore, the motivation to use cryogenic DRAMs will also continue to increase correspondingly.

3.7 CryoRAM: Conclusion

In this chapter, we first modeled and validated CryoRAM, a cryogenic memory modeling tool. Next, driven by CryoRAM, we derived two cryogenic-optimized memories, Cryogenic Low-Latency DRAM (CLL-DRAM) and Cryogenic Low-Power DRAM (CLP-DRAM). Lastly, by using the cryogenic memories, we provided three case studies in which the cryogenic memories can significantly improve both performance and power efficiency, or reduce the cost of running a modern datacenter.

Chapter 4

CryoCache: Modeling and Optimizing Caches at 77K

4.1 Motivation, Challenge, and Goal

4.1.1 Targeting a 77K-optimized cache

As a next step to realize full cryogenic computer systems, we target to build a 77K-optimized cache architecture. Specifically, 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. 4.1 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 architects do their best effort to improve both latency and capacity of LLC even though they might not still meet the desires.

Second, as caches are frequently accessed by computing cores, reducing their

[1] CryoCache was published on the 25th ACM Architectural Support for Programming Languages and Operating Systems (ASPLOS).

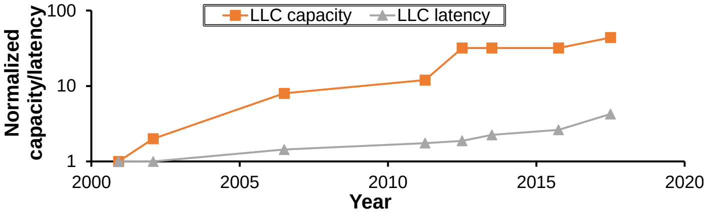


Figure 4.1: LLC latency and capacity of CPUs over generations

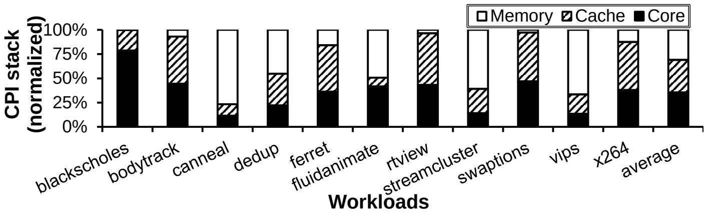


Figure 4.2: Normalized CPI stacks of PARSEC 2.1 workloads

access latency will also significantly improve the processor's overall performance. Fig. 4.2 shows the normalized CPI stacks of PARSEC 2.1 workloads [11] obtained by our gem5 simulations [12]. The figure clearly indicates that the cache performance significantly contributes to the modern application performance.

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. 4.3. From this experiment, we observe that the on-chip caches can run

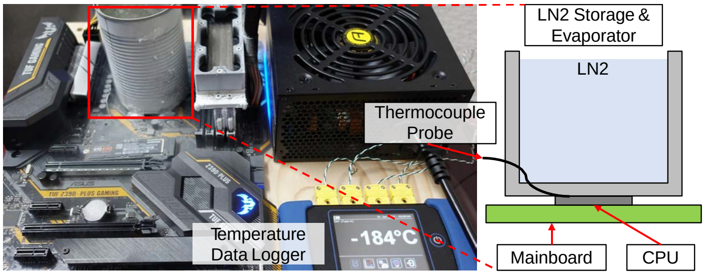


Figure 4.3: Our setup to run the whole processor at ~77K

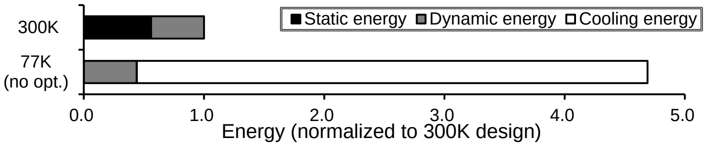


Figure 4.4: Total required energy of caches with 77K cooling

faster by 20% at 77K. This result matches our modeling result shown in later sections (32KB L1 speed-up in Fig. 4.13b).

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.

4.1.2 Challenges in designing a 77K-optimized cache

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 [33, 63, 65, 105]). 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 [50, 60]. Fig. 4.4 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.

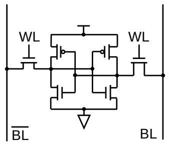
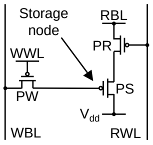
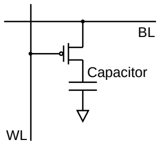
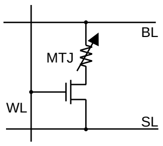
4.1.3 Research goal

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.

4.2 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 4.1. 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 4.4.

Table 4.1: 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 V$_{dd}$ 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

4.2.1 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 [20, 24, 81]. 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 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 [91]. We confirm the latency reduction with our modeling results in Section 4.4.

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 at 300K. Our simulations using Hspice and PTM models [119] (Fig. 4.5) show the static power of differently-scaled SRAM cells operating at different temperatures. The simulation limits the minimum temperature to 200K, the lowest temperature validated by PTM [118]. 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

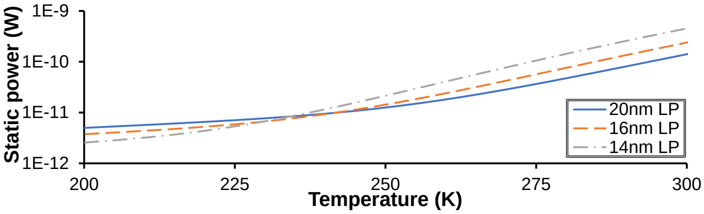


Figure 4.5: Static power of differently scaled SRAM cells

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 [57].

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

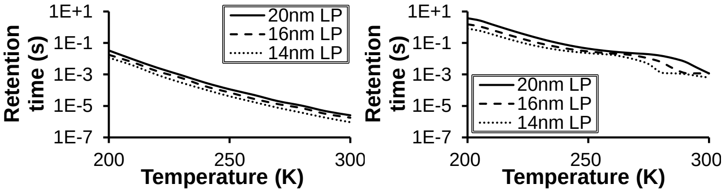
4.2.2 3T-eDRAM

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 4.1b 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.

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 [20, 24].

Cell refresh overhead. However, at 300K, the 3T-eDRAM cell is not feasible for a



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

Figure 4.6: Retention time of (a) 3T-eDRAM and (b) 1T1C-eDRAM cells

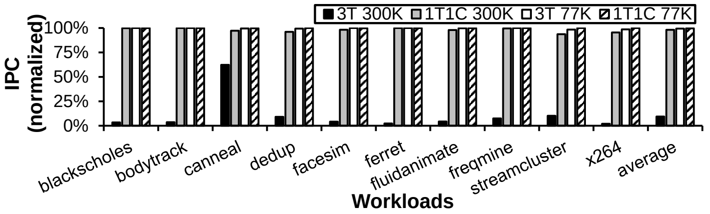


Figure 4.7: Performance impact of different eDRAM cells (3T, 1T1C) at different temperatures (300K, 77K). IPC values are normalized to IPC without refreshing

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. 4.6a 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 [25], and the overall trend of the prohibitive refresh overhead matches the results of [57]. 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. 4.7. 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 4.2). 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.

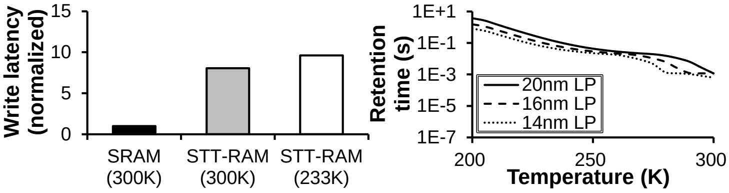
Behaviors at 77K. Interestingly, we observe that the cryogenic environment effectively eliminates the refresh overhead by dramatically extending the retention time. Even at 200K, the retention time is extended by more than 10,000 times thanks to the reduced leakage current (Fig. 4.6a). Note that the 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 [9, 89, 96, 117].

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. 4.7 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.

4.2.3 1T1C-eDRAM

Behaviors at 300K. Each 1T1C-eDRAM cell consists of an access transistor and a capacitor (Table 4.1c), which makes its cell density roughly three times higher than the 6T-SRAM cell (i.e., 2.85 times [22]). 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. 4.6). The performance degradation due to the refresh overhead is acceptable (i.e., 2.2%) at 300K as shown in Fig. 4.7.

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. Even though per-cell capacitor can be easily constructed in the fabrication process for DRAMs, it is difficult and expensive to build the capacitors inside processors using the transistor-only logic process. In addition,



(a) Write latency

(b) Write energy

Figure 4.8: Write overhead of STT-RAM at 300K and 233K.

the 1T1C-eDRAM cell is slower and consumes more access energy than SRAM and 3T-eDRAM [111, 114]. 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. 4.6b 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. 4.7).

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. Note that 1T1C-eDRAM also becomes faster at cryogenic temperatures, similar to the case of cryogenic DRAMs.

4.2.4 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 [26]. Table 4.1d 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. 4.8 shows the write latency and energy of 22nm 128KB STT-RAM at 300K and 233K. We used NVSim [29] to obtain the 300K STT-RAM values, and scaled the values for 233K according to [15]. The write overhead is normalized to that of 22nm 128KB SRAM values obtained with CACTI [73]. 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. 4.8 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 [109]. This write overhead will further increase at lower temperatures as the thermal stability is inversely proportional to the temperature [51].

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

4.3 Cryogenic Cache Modeling Framework

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 [60], our 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

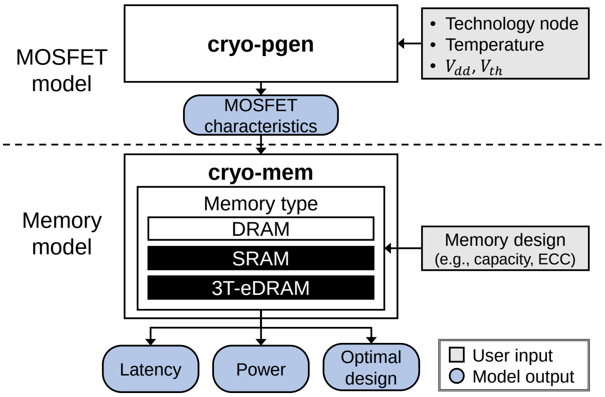


Figure 4.9: Our cryogenic cache modeling tool based on CryoRAM [60]

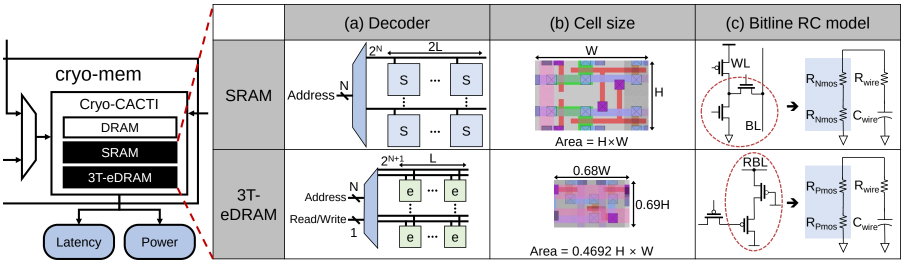


Figure 4.10: SRAM and 3T-eDRAM cache modeling overview

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. 4.9 shows our modified modeling methodology with the newly added memory models marked as the black-colored components.

4.3.1 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 [20, 53], 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 4.1b). 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. 4.10a 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 [76] (Fig. 4.10b). W and H in Fig. 4.10b 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. 4.10c). 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 [47]. 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 [53]. Therefore, we apply the SRAM cache's sense amplifier model to our 3T-eDRAM cache model.

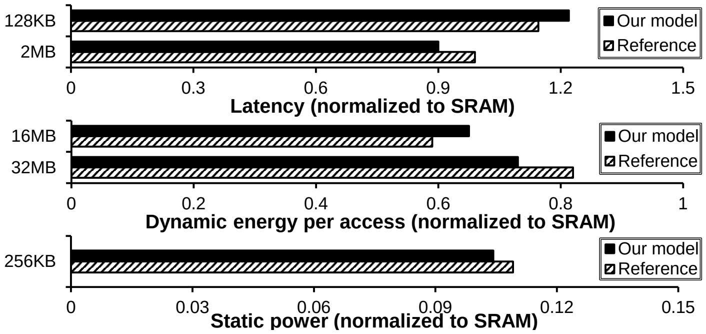


Figure 4.11: 300K 3T-eDRAM model validation results

4.3.2 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 [25]. Fig. 4.11 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 [20].

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.

4.3.3 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

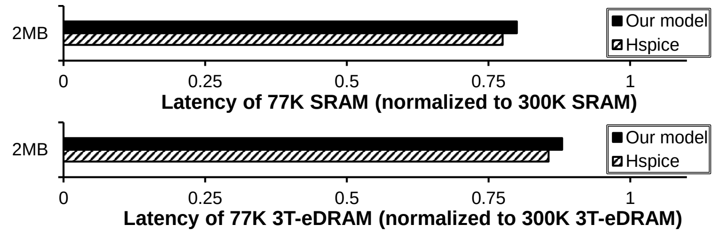


Figure 4.12: 77K cache model validation results

reduction from 300K to 77K [67], 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) [60] which can accurately estimate the wire resistivity, the leakage current, and the impact of $V_{\mathrm{dd}}/V_{\mathrm{th}}$ scaling.

4.3.4 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.

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. 4.12 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.

4.4 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 4.4.1). Next, we analyze the SRAM and 3T-eDRAM-based cryogenic caches in terms of the latency (Section 4.4.2) and the energy consumption (Section 4.4.3). Based on the analysis, we propose an optimal cache hierarchy by selectively using different cache configurations for different levels (Section 4.4.4).

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.

4.4.1 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 technology. V_{dd} and V_{th} of the baseline are 0.8V and 0.5V, respectively, which are the 22nm PTM default values [119]. 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. 4.4, 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 [57]. 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.

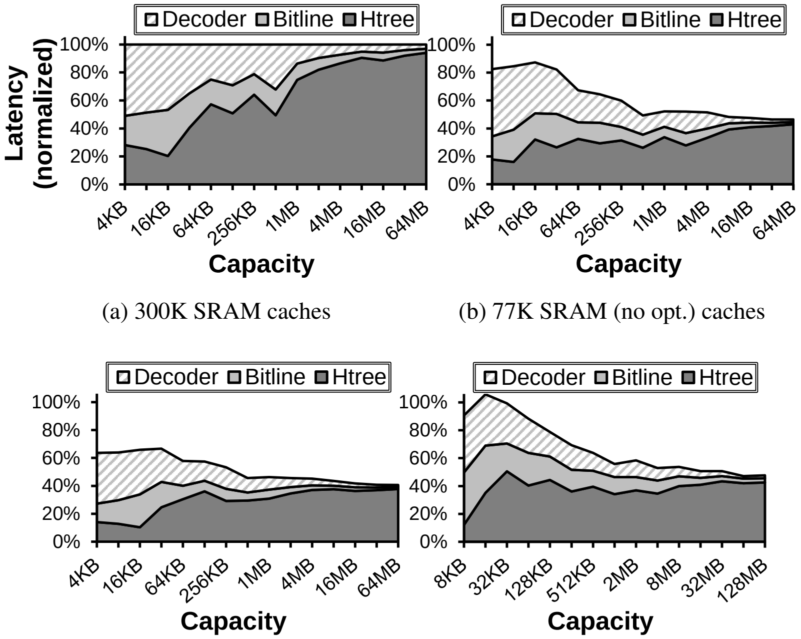
In the following sections, we compare the voltage-optimized 77K caches with 77K SRAM caches without voltage scaling. ”Opt.” means the voltage-optimized cache design, while ”No opt.” indicates the 77K cache without voltage scaling.

4.4.2 Latency analysis

Fig. 4.13 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. 4.13a) 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. 4.13a shows the latency breakdown of 300K SRAM caches. For the 4KB capacity, the decoder latency dominates the access latency. However, the ratio



(a) 300K SRAM caches

(b) 77K SRAM (no opt.) caches

(c) 77K SRAM (opt.) caches

(d) 77K 3T-eDRAM (opt.) caches

Figure 4.13: Latency breakdown of (a) 300K SRAM, (b) 77K SRAM (no opt.), (c) 77K SRAM (opt.), (d) 77K 3T-eDRAM (opt.) caches in various capacity. The latency values are normalized to the latency of the 300K SRAM caches with same area.

of decoder latency decreases for larger capacity caches because the decoder latency is proportional to the log of the memory capacity [83]. 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. 4.13b 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.

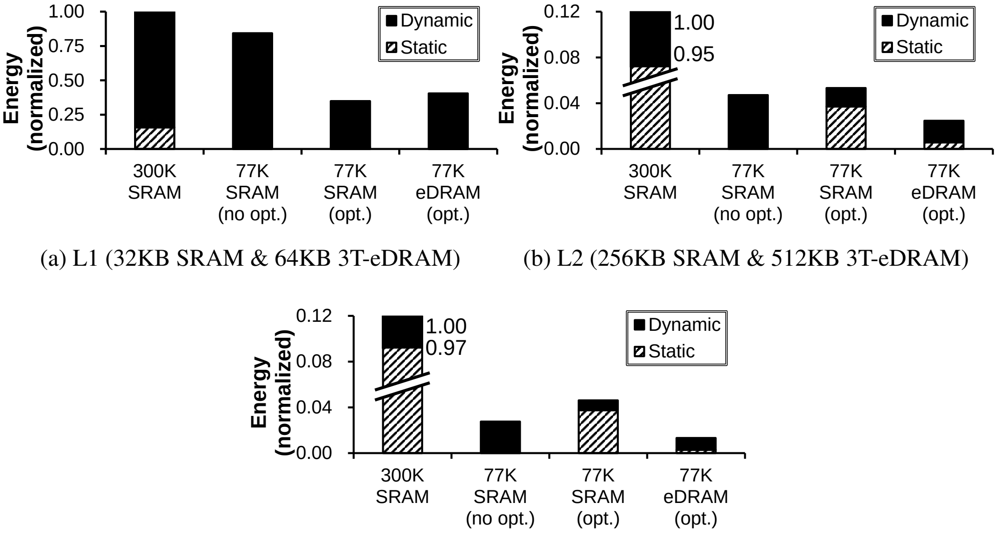
Fig. 4.13c shows the latency breakdown of 77K SRAM (opt.) caches. 77K SRAM (opt.) caches are always faster than 77K SRAM (no opt.) caches by scaling down V_{th} (2.1 times) more than V_{dd} (1.8 times) which makes the transistors run faster [47]. The latency of the 64MB 77K SRAM (opt.) cache is 40.6% of the 64MB 300K SRAM latency.

Finally, Fig. 4.13d 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.

4.4.3 Energy consumption analysis

Fig. 4.14 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*) [11] with the baseline setting in Table 4.2. We use the cache access rate of the baseline for calculating the dynamic energy of each cache.

First, Fig. 4.14a 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



(a) L1 (32KB SRAM & 64KB 3T-eDRAM)

(b) L2 (256KB SRAM & 512KB 3T-eDRAM)

(c) L3 (8MB SRAM & 16MB 3T-eDRAM)

Figure 4.14: 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.

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. 4.14b and 4.14c 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 [24]. 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.

4.4.4 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, CryoCache selects 77K SRAM (opt.) for its 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 [42, 79]. 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, CryoCache selects 77K 3T-eDRAM (opt.) for its 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.

4.5 Evaluation

In this section, we show the system-level performance and energy-efficiency of the proposed cache design. We first introduce our evaluation methodology (Section 4.5.1). Next, we evaluate our cache design in terms of the performance (Section 5.5.2) and energy consumption (Section 4.5.3).

4.5.1 Evaluation methodology

Evaluation setup

For the evaluation, we use Gem5 timing simulator [12]. 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 [49]. We utilize 11 PARSEC 2.1 workloads [11].

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 4.4.2. 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 4.2.

Table 4.2: Evaluation setup

<table>
<tr><th colspan="5">Common specification</th></tr>
<tr><td colspan="2">CPU</td><td colspan="3">Based on the Intel i7-6700</td></tr>
<tr><td colspan="2">Memory</td><td colspan="3">DDR4 2400</td></tr>
<tr><th colspan="5">Cache specification</th></tr>
<tr><th>Design</th><th>Level</th><th>Type</th><th>Capacity</th><th>Latency</th></tr>
<tr><td rowspan="3">Baseline
(300K)</td><td>L1</td><td>SRAM</td><td>32KB</td><td>4cyc</td></tr>
<tr><td>L2</td><td>SRAM</td><td>256KB</td><td>12cyc</td></tr>
<tr><td>L3</td><td>SRAM</td><td>8MB</td><td>42cyc</td></tr>
<tr><td rowspan="3">All SRAM
(77K, no opt.)</td><td>L1</td><td>SRAM</td><td>32KB</td><td>3cyc</td></tr>
<tr><td>L2</td><td>SRAM</td><td>256KB</td><td>8cyc</td></tr>
<tr><td>L3</td><td>SRAM</td><td>8MB</td><td>21cyc</td></tr>
<tr><td rowspan="3">All SRAM
(77K, opt.)</td><td>L1</td><td>SRAM</td><td>32KB</td><td>2cyc</td></tr>
<tr><td>L2</td><td>SRAM</td><td>256KB</td><td>6cyc</td></tr>
<tr><td>L3</td><td>SRAM</td><td>8MB</td><td>18cyc</td></tr>
<tr><td rowspan="3">All eDRAM
(77K, opt.)</td><td>L1</td><td>3T-eDRAM</td><td>64KB</td><td>4cyc</td></tr>
<tr><td>L2</td><td>3T-eDRAM</td><td>512KB</td><td>8cyc</td></tr>
<tr><td>L3</td><td>3T-eDRAM</td><td>16MB</td><td>21cyc</td></tr>
<tr><td rowspan="3">CryoCache</td><td>L1</td><td>SRAM</td><td>32KB</td><td>2cyc</td></tr>
<tr><td>L2</td><td>3T-eDRAM</td><td>512KB</td><td>8cyc</td></tr>
<tr><td>L3</td><td>3T-eDRAM</td><td>16MB</td><td>21cyc</td></tr>
</table>

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. (5.2)).

$$E_{cooling} = E_{device} \cdot CO \tag{4.1}$$

E_{device} is the energy consumption of the electronic devices and CO is the cooling overhead [50]. 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 [50]. Therefore,

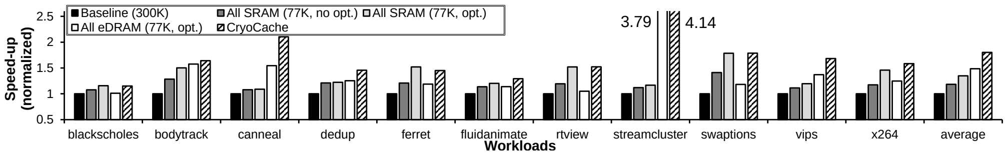


Figure 4.15: Speed-up of CryoCache

we use 9.65 value for our 77K cooling overhead ($\mathrm{CO}_{77\mathrm{K}}$).

$$\begin{aligned}\mathrm{E}_{77\mathrm{K\text{-}total}} &= \mathrm{E}_{77\mathrm{K\text{-}device}} + \mathrm{E}_{77\mathrm{K\text{-}cooling}} \\ &= (1 + \mathrm{CO}_{77\mathrm{K}})\, \mathrm{E}_{77\mathrm{K\text{-}device}} \\ &= 10.65\, \mathrm{E}_{77\mathrm{K\text{-}device}} \end{aligned} \tag{4.2}$$

Based on the cooling energy model, we calculate the total required energy for our 77K system ($\mathrm{E}_{77\mathrm{K\text{-}total}}$) as Eq. (5.3). Eq. (5.3) 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 [66]. For this reason, our energy evaluation reflects the realistic cooling cost for 77K.

4.5.2 Performance evaluation

Fig. 4.15 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.

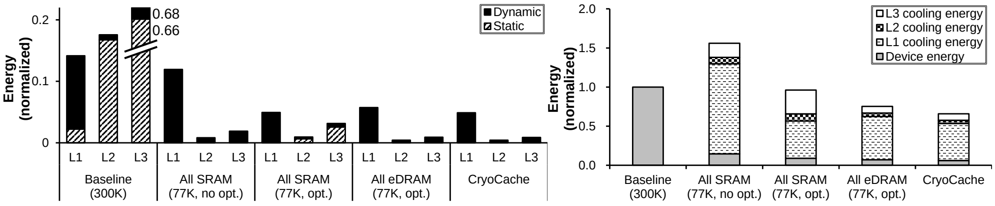
First, All SRAM (77K, no opt.) achieves the speed-up of 18.3% on average, up to 41.0% for *swaptions*. The speed-up purely results from the reduced access latency. Each workload has a different speed-up, due to the differences in the performance bottleneck. For example, *swaptions* shows the highest speed-up (41.0%) because *swaptions* has the largest cache portion in the CPI stack (Fig. 4.2). 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.

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 [11]. 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. 4.2). 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*).

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



(a) Cache energy breakdown

(b) Total energy consumption

Figure 4.16: (a) cache energy breakdown and (b) total energy consumption (including the cooling cost) of the five cache designs, normalized to those of Baseline (300K).

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.

4.5.3 Energy evaluation

Figs. 4.16a and 4.16b 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. 4.16a) and induces the huge cooling energy consumption (Fig. 4.16b). 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 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.

4.6 CryoCache: Conclusion

In this chapter, we first analyzed the cost-effectiveness and feasibility of various on-chip memory technologies running at 77K. Next, we developed our cryogenic cache modeling framework to estimate latency and power of various cache designs. Finally, based on the analysis and our framework, we architected CryoCache, a fast, large, power-efficient, and technology-feasible cache architecture running at 77K.

Chapter 5

CryoCore: Modeling and Optimizing Cores at 77K

5.1 Motivation, Challenge, and Goal

5.1.1 Targeting a 77K-opimized core

After resolving on-chip and off-chip memory walls in the previous chapters, we target to build a 77K-optimized core architecture as a next step. Specifically, we aim to improve single-thread and multi-thread performance of CPU pipelines under the same power budget by applying the 77K-based cryogenic computing.

As the technology scaling continues, it is getting more difficult to build a faster processor mainly due to the significantly increasing wire resistance and leakage current. As the wire delay cannot be scaled with the shrinking device size, it is now extremely challenging to increase the core clock frequency. Also, if architects force to increase the clock frequency, the processor cannot compensate for the correspondingly increasing dynamic power consumption due to the prohibitively increasing leakage current with voltage scaling (i.e., *end of single-thread performance scaling*).

To get around this single-thread performance challenge, architects have instead improved a chip's multi-thread performance by using more cores and hardware threads

[1] CryoCore was published on the 47th ACM/IEEE International Symposium on Computer Architecture (ISCA) and IEEE Micro Top Picks (TopPicks).

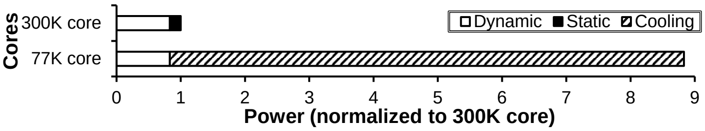


Figure 5.1: Conventional core's power consumption with cooling cost included, derived by McPAT [62]

(i.e., CMP [58, 74], SMT [103]). However, these circumventions are hitting the physical and economic limits (e.g., the increasing chip power consumption or dark silicon) as well as the programming burden (i.e., *end of multi-thread performance scaling*).

As cryogenic computing resolves the fundamental problems (i.e., increasing wire resistance and leakage current) of processors' single-thread and multi-thread performance scaling, we target to build a 77K-optimized core architecture as our next step. In addition, the cryogenic-optimal core will also get synergistic benefits by using the previously proposed cryogenic caches and memories.

5.1.2 Challenges in designing a 77K-optimized core

To develop a 77K-optimal cryogenic core, architects should resolve the three following challenges.

Absence of a core performance model: To design a performance-optimal core, architects need a performance model to accurately estimate a target core's per-pipeline critical-path delays and its maximum core frequency. Researchers have proposed various critical-path delay models for major pipeline stages (e.g., renaming, issue selection, bypass logic) [62, 77]. However, as all these models assume the room temperature (i.e., 300K), so they cannot be used for core designs running at 77K.

Cooling cost analysis and compensation: To estimate the cost effectiveness of a cryogenic core, architects should carefully analyze and reduce its cooling cost. To maintain a device's temperature at 77K, a conventional cryogenic cooler consumes 9.65 times higher energy than the cooled device (see Section 5.5.1). Fig. 5.1 shows that lowering a processor's temperature from 300K to 77K can significantly increase its overall power consumption due to the cooler's increased power consumption which is approximately 10 times of the processor's dynamic power at 77K. Therefore, such cooling costs can make ineffective the most of the advantages obtained by cryogenic computing. Therefore, to compensate for the cooling cost, a 77K-optimal cryogenic core should reduce its dynamic power by 10 times compared to the a core running at 300K.

Cryogenic-optimal core architecture: With a cryogenic core performance, power, and cooling-cost modeling tool available, architects should design a 77K-optimal core architecture. The optimal cryogenic core architecture should provide the highest single-thread and multi-thread performance while keeping its overall area and power overhead under the budget. However, to the best of our knowledge, neither such analysis nor the proposed core architecture exists.

5.1.3 Research goal

In this chapter, we resolve the three challenges as follows. We first develop a novel cryogenic processor's performance modeling framework (*CC-Model*). Next, we analyze a core's maximum frequency, power consumption, and cooling costs for the target cryogenic temperature. Finally, we architect and propose our cryogenic-optimal processor design (*CryoCore*) to provide the highest single-thread and multi-thread performance while satisfying the target die area and cooling cost budget.

5.2 CC-Model: Cryogenic Core Modeling Framework

In this section, we describe our cryogenic processor modeling framework, CryoCore-Model (CC-Model), to explore and design our 77K-optimized processors. CC-Model consists of three sub-models as shown in Fig. 5.2. First, *MOSFET model (cryo-MOSFET)* takes fabrication-process information (i.e., model card) as inputs, and then derives the

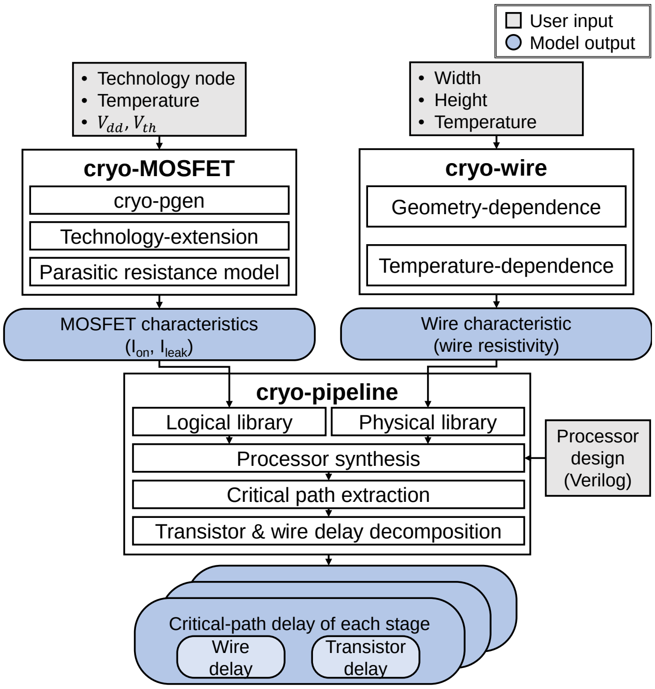
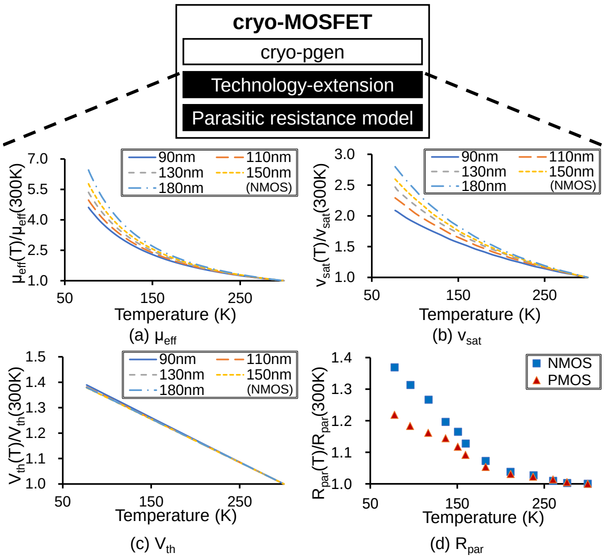


Figure 5.2: Cryogenic processor model (CC-Model) overview

major MOSFET characteristics (i.e., on-channel current (I_{on}), leakage current (I_{leak})) for a wide range of temperatures including 77K. Second, based on the given metal layer's information, *wire model (cryo-wire)* generates the on-chip wire characteristic (i.e., wire resistivity) at cryogenic temperatures. Finally, *processor model (cryo-pipeline)* reports the critical-path delay of each pipeline stage by utilizing the output low-temperature MOSFET/wire properties from MOSFET/wire models. In the following sections, we explain each model's role and implementation details.

5.2.1 MOSFET model

To model the major MOSFET characteristics at low temperatures, we utilize cryo-pgen [60] as a baseline model. Cryo-pgen is a validated cryogenic MOSFET model which takes a model card as an input, automatically adjusts the model card for given V_{dd} and V_{th}, and derives the MOSFET characteristics at the target temperature. The



(a) μ_{eff}

(b) v_{sat}

(c) V_{th}

(d) R_{par}

Figure 5.3: Extension to the baseline MOSFET model: (a) Carrier mobility; (b) Saturation velocity; (c) Threshold voltage; (d) Parasitic resistance model

input model card is a set of low-level MOSFET variables related to the MOSFET fabrication process (e.g., gate-oxide thickness, doping concentration). The output MOSFET characteristics include the on-channel current (I_{on}) and the leakage current (I_{leak}). Cryo-pgen predicts the MOSFET characteristics at 77K by adjusting the three highly temperature-dependent MOSFET variables (i.e., effective carrier mobility (μ_{eff}), saturation velocity (v_{sat}), threshold voltage (V_{th})) to 77K values.

However, cryo-pgen has two challenges to predict the low-temperature MOSFET characteristics of modern technology nodes. First, cryo-pgen cannot accurately predict the values of temperature-dependent variables for small technology nodes. Cryo-pgen estimates μ_{eff}, v_{sat}, and V_{th} at low temperatures by assuming that the ratios of three variables between 300K and the target temperature (T) (i.e., μ_{eff}(T)/μ_{eff}(300K), v_{sat}(T)/v_{sat}(300K), V_{th}(T)/V_{th}(300K)) are preserved in every technology node. However, the simple assumption is insufficient to predict the complex impact of technology scaling on the temperature model. Second, cryo-pgen does not model the temperature

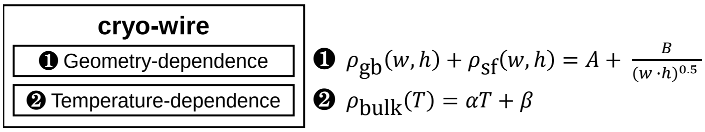

cryo-wire
❶ Geometry-dependence
❷ Temperature-dependence
❶ $\rho_{gb}(w,h) + \rho_{sf}(w,h) = A + \frac{B}{(w \cdot h)^{0.5}}$
❷ $\rho_{bulk}(T) = \alpha T + \beta$

Figure 5.4: Wire model

dependency of the parasitic resistance (R_{par}). The absence of R_{par} model makes it difficult for cryo-pgen to accurately predict the MOSFET characteristics in small technology nodes because the impact of R_{par} grows with technology scaling [118]. The problems become more critical for processors because CPU's transistors are much smaller than other memory devices.

To resolve the challenges, we build cryo-MOSFET by implementing two additional models on the top of cryo-pgen. First, we separately model the temperature dependency in each gate length, based on the industry-provided MOSFET model (i.e., *technology-extension model*). Fig. 5.3a-c shows the temperature dependency of μ_{eff}, v_{sat}, and V_{th} for various gate lengths ranging from 180nm to 90nm. Each graph in Fig. 5.3 is extracted from the industry-validated device model. Cryo-MOSFET can also predict the MOSFET characteristics of smaller nodes because it extrapolates the variables for smaller technologies.

Second, we add the temperature dependence model for R_{par} as shown in Fig. 5.3d (i.e., *parasitic resistance model*). We utilize the temperature dependency data of R_{par} from the previous work [118]. With these additional models, cryo-MOSFET can now accurately predict the low-temperature MOSFET characteristics of modern technology nodes.

5.2.2 Wire model

$$\rho_{\text{wire}}(T, w, h) = \rho_{\text{bulk}}(T) + \rho_{\text{gb}}(w, h) + \rho_{\text{sf}}(w, h) \tag{5.1}$$

The goal of the wire model is to accurately predict the wire resistivity at low temperatures for each on-chip metal layer, which has a different wire width and height.

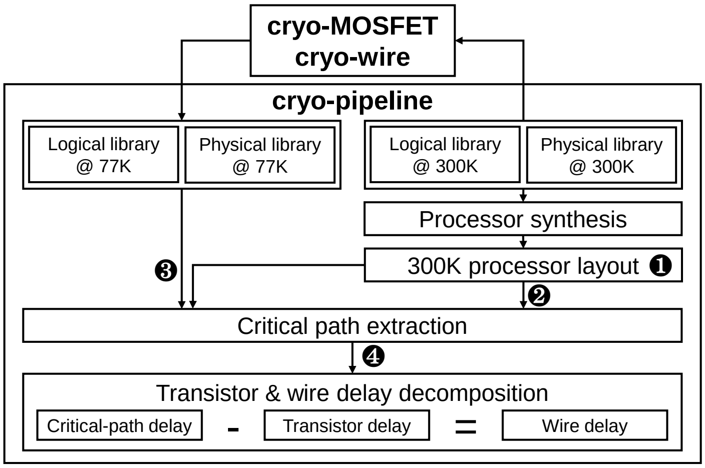


Figure 5.5: Processor model

The wire resistivity (ρ_{wire}) is mainly determined by the three physical mechanisms: geometry-independent scattering (ρ_{bulk}), grain boundary scattering (ρ_{gb}), and surface scattering (ρ_{sf}) [54,68,95]. Eq. (5.1) shows the relationship where T, w, and h indicate the wire's temperature, width, and height, respectively. Among the three mechanisms, ρ_{bulk} depends only on the temperature. On the other hand, ρ_{gb} and ρ_{sf} are mainly determined by the width, height, and purity of wires (i.e., wire geometry) [46,69,78,95]. Therefore, we should consider both the geometry and the temperature dependency in the wire model.

We implement these two dependencies on cryo-wire as follows. First, we build geometry-dependent mechanisms (i.e., ρ_{gb} and ρ_{sf}) by utilizing simple physics-based models [45,46,95] (Fig. 5.4❶). We set the purity-related hyper-parameters (i.e., A and B) based on the previous studies [46,97]. Next, we implement temperature-dependent mechanisms (ρ_{bulk}) as the linear model in Fig. 5.4❷ with the coefficients of coppers [67].

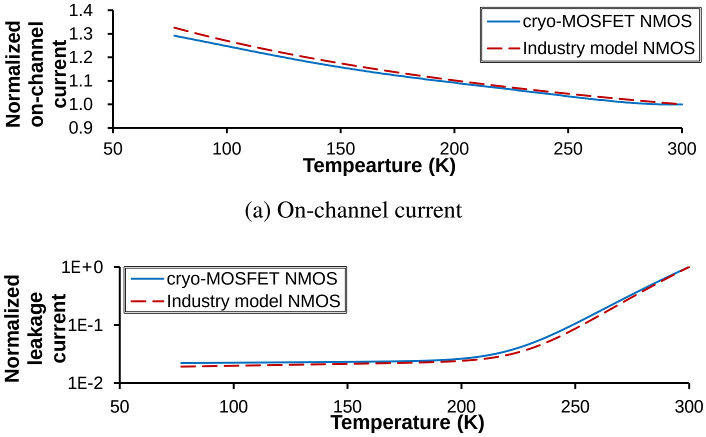
5.2.3 Processor model

Based on the given processor design, our processor model (cryo-pipeline) predicts the critical-path delay of each pipeline stage at low temperatures, by taking the MOSFET and wire characteristics from cryo-MOSFET and cryo-wire, respectively. In addition, cryo-pipeline can decompose each critical-path delay to the transistor and the wire delay portion. Therefore, with cryo-pipeline, architects can predict the frequency speedup at cryogenic temperatures and analyze how the low temperatures affect the delay of each pipeline stage.

For cryo-pipeline implementation, we utilize Synopsys Design Compiler Topographical Mode [99]. Design Compiler Topographical Mode can synthesize a Verilog design based on the logical library (i.e., transistor/gate information) and the physical library (i.e., metal-layer information). In addition, Design Compiler provides an interface to fix a specific layout design while applying different libraries. Finally, Design Compiler Topographical Mode can report critical-path delay of each stage and extract the transistor-only delay of target paths (with no-wire option). By using Design Compiler, we implement cryo-pipeline as follows.

Critical-path delay of each pipeline stage: Fig. 5.5 shows the detailed overview of our processor model. First, cryo-pipeline synthesizes a processor layout by utilizing an input processor design (Verilog) and 300K logical/physical libraries (❶). Next, with the processor layout, cryo-pipeline extracts the critical-path delay of each pipeline stage at 300K (❷). Finally, cryo-pipeline derives the delays at 77K with the same layout by using the 77K libraries generated by MOSFET/wire models (❸). By doing so, cryo-pipeline accurately predicts the absolute delay and relative frequency speedup at 77K.

MOSFET/Wire delay decomposition: Cryo-pipeline can fully decompose the critical-path delay into its transistor and wire delay portions by subtracting the impact of the transistor portions from the overall critical-path delay result considering all modeling aspects (❹).



(a) On-channel current

(b) Leakage current

Figure 5.6: cryo-MOSFET validation results: industry-validated model vs cryo-MOSFET outcomes

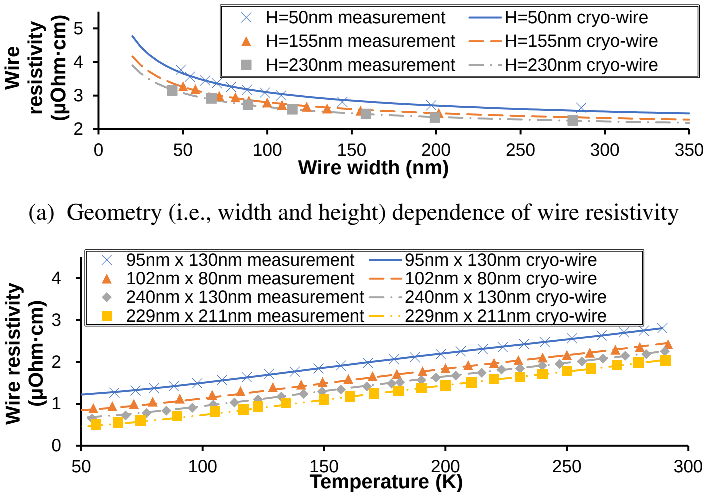
5.3 Model Validation

In this section, we validate our models by comparing their outputs with industry-provided information, previous literature, and our own experiments.

5.3.1 MOSFET model validation

We validate our MOSFET model by comparing major MOSFET characteristics (i.e., I_{on}, I_{leak}) predicted by cryo-MOSFET with those obtained from our industry-provided MOSFET model card. The industry model card for Hspice simulation is based on MOSFET samples fabricated with 2z nm technology, and the data was pre-validated by actual measurements for the 77K-to-300K temperature range. To match the technology, cryo-MOSFET uses 22nm PTM [119] as its input model card.

Fig. 5.6 shows the cryo-MOSFET's accuracy in terms of I_{on} and I_{leak}. I_{on} and I_{leak} values are normalized to the 300K value of each model. First, cryo-MOSFET well matches the industry model's I_{on} improvement at low temperatures (Fig. 5.6a). Our MOSFET model not only accurately predicts the trend of increasing I_{on} but also



(a) Geometry (i.e., width and height) dependence of wire resistivity

(b) Temperature dependence of wire resistivity

Figure 5.7: cryo-wire validation results: measurement data from the previous literature vs. cryo-wire outcomes

shows the small errors for every temperature, 3.3% in maximum. The I_{on} improvement stems from the increase in μ_{eff} and v_{sat} (as shown in Fig. 5.3a, b). Our MOSFET model never overestimates the increase in I_{on}.

Second, cryo-MOSFET's prediction for I_{leak} is also accurate as shown in Fig. 5.6b. Cryo-MOSFET accurately models the exponentially decreasing leakage current from 300K to 200K, and the nearly constant leakage current below 200K. The exponentially decreasing and nearly constant trends originate from the temperature dependence of subthreshold current and gate leakage current, respectively. In addition, our MOSFET model's predictions are slightly higher than the industry model's results. Therefore, we conclude that our MOSFET model accurately and conservatively predicts the target MOSFET characteristics at the low temperatures.

5.3.2 Wire model validation

We validate our wire model by comparing the wire resistivity reported by cryo-wire with the measured data from the literature [97, 110, 116]. Fig. 5.7 shows the validation

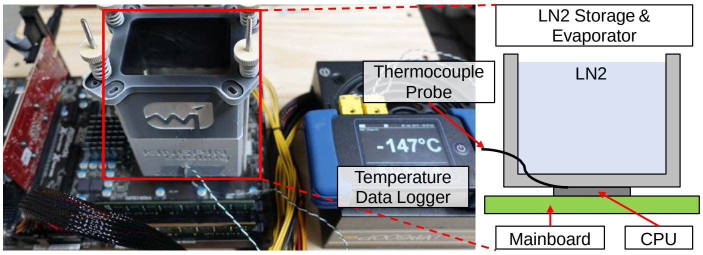


Figure 5.8: Experimental setup for the processor model validation

results for cryo-wire. First, cryo-wire well matches the published resistivity data for various sets of width and height (Fig. 5.7a) [97]. Second, Fig. 5.7b shows that our wire model well predicts the linearly decreasing wire resistivity compared to the data from previous literature [110, 116]. In addition, cryo-wire always reports slightly higher resistivity values for the given temperatures. Therefore, the results indicate that cryo-wire accurately and conservatively predicts the resistivity.

5.3.3 Processor model validation

In this section, we validate cryo-pipeline for its frequency speed-up prediction with various voltage setup. For the ideal validation, we should compare the model's prediction and the measurement data for the exactly same processor design. However, the ideal experiment is almost impossible because the Verilog source file of a commercial processor is usually unavailable. As an alternative approach, we use a representative processor design for the model's input and show the frequency speed-up prediction reasonably matches with the measured value for a commercial processor.

Fig. 5.8 shows our experimental setup for validating cryo-pipeline. We construct a sample computer board using various commodity parts (i.e., AMD 970 mainboard, AMD Phenom2 X4 960T CPU, and two Samsung DDR3 2G DIMMs) and the evaporator for LN cooling. With the setup, we can separately cool-down the CPU socket. This setup also allows us to adjust the CPU's voltage and frequency independently. Note that we intentionally construct the computer board with the processor fabricated

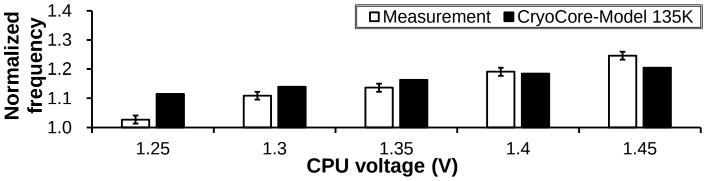


Figure 5.9: cryo-pipeline validation results: real measurements vs. cryo-pipeline outcomes

with 45nm technology to validate cryo-pipeline targeting the 45nm technology.

With the experimental setup, we measure the frequency speed-up at 135K compared to the maximum frequency at 300K. Note that 135K is the average temperature achieved with our indirect cooling system during the experiment. We find the maximum frequency of each temperature by increasing CPU frequency until the booting process fails or the CPU does not reliably operate.

Fig. 5.9 shows the validation results of cryo-pipeline. The error bars indicate the last succeeded frequency and the first failed frequency from the experiments. To derive cryo-pipeline's speed-up results, we use FreePDK 45nm library [98] with BOOM processor design [17] as the model inputs. Fig. 5.9 shows that cryo-pipeline reports a reasonably accurate frequency speed-up at 135K, with 4.5% of the maximum error at 1.45V, even with two processor designs use different microarchitectures (i.e., AMD and Boom processors).

5.4 CryoCore: Cryogenic-Optimal Core Design

In this section, with our validated modeling framework, we architect a 77K-optimized core design in terms of performance and power efficiency. In general, a complex design such as a microprocessor core has an extremely wide design space which cannot be fully explored by a modeling tool. Therefore, we first draw key design directions to architect a core microarchitecture running at 77K (Section 5.4.1). Next, following the directions, we design our 77K-optimal microarchitecture, called *CryoCore* (Sec-

Table 5.1: Hardware specifications of hp, lp, and CryoCore

	Hp-core (i7-6700)	Lp-core (Cortex-A15)	**CryoCore**
# cache load/store ports	4	1	1
Pipeline width	8	4	4
Load queue size	72	24	24
Store queue size	56	24	24
Issue queue size	97	72	72
Reorder buffer size	224	96	96
# physical integer registers	180	100	100
# physical float registers	168	96	96
Max frequency	4.0GHz	2.5GHz	4.0GHz
Power per core (45nm)	24W	1.5W	5.5W
Core area (45nm)	44.3mm^2	11.54mm^2	22.89mm^2
Core & L1/L2 area (45nm)	97.51mm^2	17.51mm^2	38.89mm^2
Supply voltage (V_{dd})	1.25V	1.0V	1.25V

tion 5.4.2). Finally, by applying different voltage scalings, we propose two CryoCore designs which are optimized for either higher performance (CHP-core) or power efficiency (CLP-core), respectively (Section 5.4.3).

In the following subsections, we conduct performance, power, and area analyses for the processors listed in Table 5.1. The pipeline width in Table 5.1 indicates the fetch or issue width of processor pipelines. We implement the target processors by customizing RISC-V BOOM [17], one of the most representative out-of-order core designs. For performance analysis, we utilize CC-Model with FreePDK 45nm library [98] which can be scaled to 77K by our MOSFET/wire models. For power and die-area analysis, we use McPAT [62] based on the 45nm technology node. Note that we use 45nm technology because FreePDK 45nm is the smallest technology library which we find among various open-source physical/logical libraries. See Section 5.5.1 for more details of our power calculation methodology including the cooling cost model.

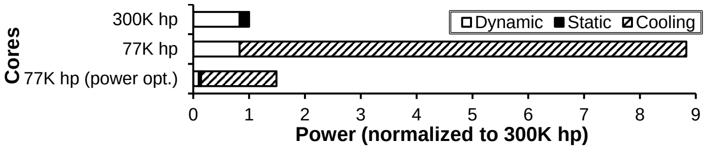


Figure 5.10: Power consumption of hp-cores at 300K and 77K

5.4.1 Design principles for 77K-optimal core microarchitecture

In this section, we introduce our power and performance-side design principles by performing case studies with two reference core models: high-performance core (hp-core) and low-power core (lp-core).

Principle 1. Minimize dynamic power consumption at the microarchitectural level.

We first emphasize the importance of reducing the dynamic power at the microarchitectural level. To draw the principle, we first start from our high-performance core model running at 77K to target the high-performance datacenter market. We set the hardware specification of hp-core based on Intel i7-6700 Skylake processor [30] (hp-core in Table 5.1). We set hp-core's frequency at 300K based on the literature [30], and its power and area are calculated from McPAT.

Fig. 5.10 shows the power consumption of hp-cores operating at various temperatures and voltages. 300K hp and 77K hp in the figure indicate two hp-core designs running at 300K and 77K without any voltage optimization, respectively. First, we observe that dynamic power (83%) dominates the power consumption of hp-core running at 300K (300K hp). Unfortunately, as the cryogenic temperature does not affect the dynamic power, the dynamic power remains and incurs huge cooling power consumption (800%) at 77K (77K hp).

To reduce the dynamic power, we can decrease the V_{dd} and V_{th} level simultaneously at 77K. However, even though the aggressive voltage scaling is applied, hp-core cannot achieve the power efficiency at 77K. 77K hp (power opt.) in Fig. 5.10 indicates the lowest power design obtained by voltage scaling while maintaining the clock frequency at 300K. Even with the aggressive voltage scaling, the huge dynamic power cannot be removed and incurs the significant cooling cost at 77K. As the graph shows, the power consumption of 77K hp (power opt.) is still higher than the total power of 300K hp. That is, there exists a limit in dynamic power reduction with voltage scaling, and thus naively adopting hp-core's microarchitecture cannot achieve the power efficiency at 77K. Therefore, we should minimize the dynamic power at the microarchitectural level for power efficiency.

Principle 2. Maximize the clock frequency at the microarchitectural level.

We now emphasize the importance of achieving the high frequency at the microarchitectural level. To draw the principle, we perform an analysis with a low-power reference core (i.e., lp-core) because we highlighted the importance of lower power consumption in the previous section. We set the hardware specification of lp-core based on ARM Cortex-A15 processor [59], whose power consumption (1.5W) and maximum clock frequency (2.5GHz) are lower than hp-core by 93.7% and 37.5%, respectively (lp-core in Table 5.1). The lp-core's frequency is based on the literature [59] and its power consumption and area are derived from McPAT.

Fig. 5.11 shows the result of frequency and power analysis for three lp-core designs running at 77K. The three designs (77K lp, 77K lp (freq. opt), and 77K lp (extreme freq.)) share the same core design, but apply different voltage scalings to adjust their frequencies. For this analysis, we include the cooling power overhead to maintain the low temperature. To directly compare the lp-cores with high-performance server processors, we normalize the values to those of hp-core operating at 300K (i.e., 300K hp-core).

First, lp-core with the nominal voltage (77K lp) consumes 33.5% less power compared to 300K hp-core, even with the cooling cost included. The improved power efficiency results from lp-core's dynamic power-optimized microarchitecture. However, 77K lp-core's baseline clock frequency (2.9GHz) is 27.5% lower than the 300K hp-core's frequency.

To achieve a higher frequency enabled by the reduced temperature, we increase lp-

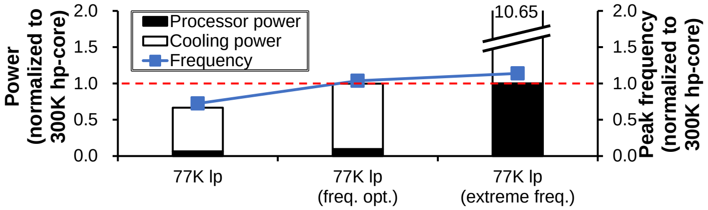


Figure 5.11: Maximum frequency and total power consumption of lp-cores operating at 77K

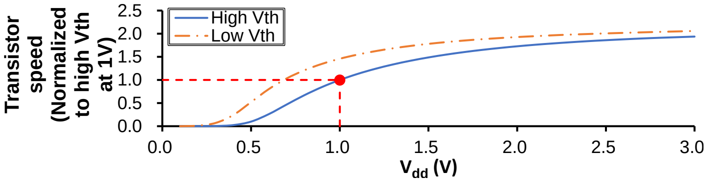


Figure 5.12: Saturated transistor speed with the increasing V_{dd}

core's V_{dd} (and thus its frequency) up to two specific points which forms 77K lp (freq. opt) and 77K lp (extreme freq.) as shown in Fig. 5.11. 77K lp (freq. opt.) is the design point to keep its total power consumption (including its cooling cost) the same as the hp-core's power at 300K. 77K lp (extreme freq.) is the design point to keep the core's device power (ignoring its cooling cost) the same as the hp-core's power at 300K. Even with the same power consumed, the frequency of 77K lp (freq. opt.) is only 3.75% higher than 300K hp-core's frequency. Furthermore, the frequency improvement of 77K lp (extreme freq.) is only 13.75% even with the aggressively increased V_{dd} and the severely increased power cost due to the cooling (1065%).

The limited frequency improvement with the voltage scaling originates from the saturated MOSFET speed at high V_{dd}. Fig. 5.12 shows the speed of MOSFET when varying its V_{dd} and V_{th}. We approximate the speed of MOSFET as its transconductance (i.e., $I_{\mathrm{on}}/V_{\mathrm{dd}}$), and derive it from Hspice simulations with industry-validated MOSFET model cards. High V_{th} means the MOSFET model with high V_{th} for 300K operation, and Low V_{th} means V_{th}-reduced MOSFET targeting for 77K operations. First, the MOSFET speed of High V_{th} is saturated at high V_{dd} domain because I_{on} is linearly proportional to V_{dd} in the high-voltage region [47]. Even though we reduce the V_{th} level (Low V_{th}), the maximum MOSFET speed at high-voltage region does not change significantly. That is, the peak frequency at 77K is mainly determined by the frequency at the nominal voltage. Therefore, we should maximize the clock frequency at the microarchitectural level for higher performance.

5.4.2 CryoCore: Cryogenic-optimal core microarchitecture design

The design principles to architect a cryogenic-optimal core are summarized as follows. First, the cryogenic-optimal core should consume much lower dynamic power than conventional high-performance cores. Next, the cryogenic-optimal core should apply much higher frequency than conventional low-power cores.

Following the principle, we design *CryoCore*, our cryogenic-optimal core design. CryoCore has the same pipeline structure (e.g., the number of pipeline stages), operating voltage, and clock frequency with the high-performance core (hp-core), but its overall sizes of microarchitectural units are the same as those of the low-power core (lp-core). By doing so, CryoCore reduces its power consumption significantly, while maintaining its maximum frequency high. Table 5.1 summarizes the frequency, power, area, and microarchitectural specifications of CryoCore at 300K.

First, CryoCore's power consumption (5.5W) is much lower than hp-core's power consumption (24W). The smaller pipeline width and size of microarchitectural units greatly reduce CryoCore's dynamic power. They also reduce the static power consumption because the static power is proportional to the chip area.

Next, CryoCore's voltage level and maximum frequency (4.0GHz) are the same as those of hp-core. We set CryoCore's V_{dd} to the same with the hp-core's voltage because higher V_{dd} cannot effectively improve the peak frequency (as shown in Fig. 5.12). Also, CryoCore adopts the pipeline structure of hp-core, which makes CryoCore have the high frequency. In fact, CryoCore's frequency can be much higher than the hp-core's frequency because CryoCore's smaller size of microarchitectural units can reduce its critical-path delay significantly [77]. However, we set CryoCore's frequency to the same as hp-core's frequency to conservatively show CryoCore's performance improvement.

Finally, CryoCore's area (22.89mm^2) is only 50% of that of hp-core (44.3mm^2), thanks to its narrow pipeline, a fewer number of units, and the reduced sizes of units. When L1 and L2 caches are added, CryoCore's area is only 40% of that of hp-core as shown in Table 5.1. This area advantage indicates that we can integrate twice more cores under the same area budget, and we evaluate the increased core density in our evaluation (Section 5.5).

5.4.3 Deriving two cryogenic-optimal processors

In this section, we derive two 77K-optimal processors by applying V_{dd} and V_{th} scaling to CryoCore. Fig. 5.13 summarizes the whole optimization process including the voltage scaling. The frequency and power values are normalized to those of 300K hp-core. Note that the power values in Fig. 5.13 do not include the cooling power consumption.

We start from 300K hp-core, which is on its power-frequency Pareto curve. First, we adopt CryoCore's microarchitecture and reduce the power consumption to 23% (❶). Next, we cool down CryoCore to 77K and increase its clock frequency by 16%. At the same time, we reduce CryoCore's power consumption by 14.7%, by taking an advantage of the eliminated static power (❷). Finally, we explore 25,000+ design points of different V_{dd} and V_{th}, and obtain the power-frequency Pareto-optimal curve as shown in Fig. 5.13. Among the optimal design points, we choose the two representative 77K processor designs: the power-optimal design (Cryogenic Low-Power core; *CLP-core*) and the frequency-optimal design (Cryogenic High-Performance core; *CHP-core*) (❸).

Cryogenic Low-Power core (CLP-core): Reducing both V_{dd} and V_{th} decreases the dynamic power while maintaining the same maximum frequency. By doing so, we obtain the ultra low-power processor design (CLP-core) without any performance degradation. CLP-core consumes only 2.93% of power compared to hp-core operating

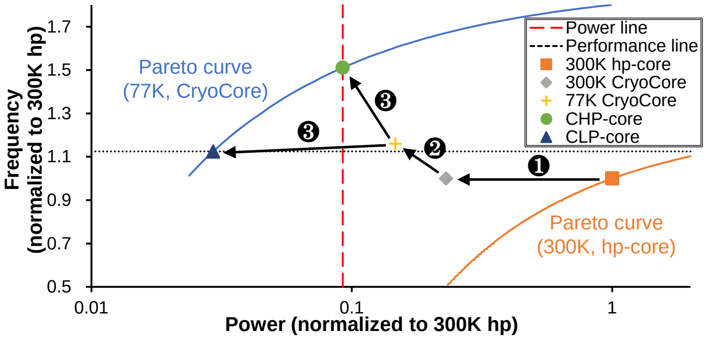


Figure 5.13: Deriving cryogenic-optimal processor designs by applying voltage scaling to CryoCore

at 300K. Note that CLP-core's clock frequency is 13% higher than hp-core's frequency which keeps the processor's performance similar to that of hp-core (Performance line in Fig. 5.13).

Cryogenic High-Performance core (CHP-core): We can improve the processor's clock frequency by applying higher V_{dd}. In this manner, we obtain the high-performance core design (CHP-core) by increasing V_{dd} within the cooling power budget (Power line in Fig. 5.13). As a result, CHP-core has 1.5 times higher peak frequency with 9.2% of device power consumption. CHP-core's total power consumption including cooling cost is the same as that of hp-core at 300K.

Note that architects can build the two proposed processors (i.e., CLP-core, CHP-core) with single hardware design because their microarchitecture (CryoCore) and V_{th} values are exactly the same with each other (as shown in Table 5.2). That is, architects can utilize both of their benefits just by applying the dynamic voltage frequency scaling (DVFS) [93] to the single-core design.

5.5 Evaluation

In this section, we show the system-level performance gain and power efficiency of our proposed core design. We first introduce our evaluation methodology (Section 5.5.1).

Table 5.2: Evaluation setup

<table>
<tr><th colspan="5">Evaluation setup</th></tr>
<tr><th>Design</th><th>Core type</th><th colspan="2"># cores</th><th>Memory type</th></tr>
<tr><td>300K hp-core with 300K memory</td><td>300K hp-core</td><td colspan="2">4</td><td>300K memory</td></tr>
<tr><td>CHP-core with 300K memory</td><td>CHP-core</td><td colspan="2">8</td><td>300K memory</td></tr>
<tr><td>300K hp-core with 77K memory</td><td>300K hp-core</td><td colspan="2">4</td><td>77K memory</td></tr>
<tr><td>CHP-core with 77K memory</td><td>CHP-core</td><td colspan="2">8</td><td>77K memory</td></tr>
<tr><th colspan="5">Core specification</th></tr>
<tr><th>Design</th><th>Frequency</th><th colspan="2">V_{dd} / V_{th0}</th><th>μ-arch specification</th></tr>
<tr><td>300K hp-core</td><td>3.4GHz</td><td colspan="2">1.25V / 0.47V</td><td>Hp-core in Table 5.1</td></tr>
<tr><td>CHP-core</td><td>6.1GHz</td><td colspan="2">0.75V / 0.25V</td><td>CryoCore in Table 5.1</td></tr>
<tr><td>CLP-core</td><td>4.5GHz</td><td colspan="2">0.43V / 0.25V</td><td>CryoCore in Table 5.1</td></tr>
<tr><th colspan="5">Memory specification</th></tr>
<tr><th rowspan="2">Design</th><th colspan="3">Cache specification</th><th rowspan="2">DRAM random access latency</th></tr>
<tr><th>L1</th><th>L2</th><th>L3</th></tr>
<tr><td>300K memory</td><td>32KB
4cyc</td><td>256KB
12cyc</td><td>8MB
42cyc</td><td>60.32ns</td></tr>
<tr><td>77K memory</td><td>32KB
2cyc</td><td>512KB
8cyc</td><td>16MB
21cyc</td><td>15.84ns</td></tr>
</table>

Next, we evaluate the single-thread and multi-thread performance of CHP-core (Section 5.5.2) and power consumption of CLP-core (Section 5.5.3).

5.5.1 Evaluation methodology

Performance evaluation methodology

We evaluate CHP-core's single-thread and multi-thread performance by considering four combinations of core and memory designs: (1) 300K hp-core with 300K memory, (2) CHP-core with 300K memory, (3) 300K hp-core with 77K memory, (4) CHP-core with 77K memory. We summarize the setup in Table 5.2.

Core design. We compare (1) 300K hp-core and (2) 77K CHP-core for evaluation. We set 300K hp-core's number of cores to four following the Intel i7-6700 specification [30]. On the other hand, for CHP-core, we set the number of cores to eight based on the area analysis in Table 5.1.

We set 77K CHP-core's clock frequency to its maximum frequency (6.1GHz), and 300K hp-core's frequency to its nominal clock frequency (3.4GHz) following Intel i7-6700 specification. In our performance evaluation, we fully utilize all on-chip cores. In that case, the 300K baseline cores should operate at the nominal frequency (3.4GHz) instead of the maximum frequency (4.0GHz), due to the thermal budget constraint. On the other hand, 77K CHP-core can reliably operate with the maximum frequency (6.1GHz) because they consume much less power (8.92W) with the much higher thermal budget (according to Section 5.6). Therefore, we set CHP-core to operate at 6.1GHz, which is 1.5 times higher than the 300K maximum frequency, or 1.8 times higher than the 300K nominal frequency.

Memory and cache hierarchy. We evaluate CHP-core's performance by adding two different memory hierarchy designs to the core: (1) a conventional memory hierarchy operating at room temperature (300K memory) and (2) a cryogenic-optimal memory hierarchy designed and optimized for 77K (77K memory). For the 300K memory setup, we use Intel i7-6700 processor's cache specifications and DDR4-2400's DRAM access latency. For this setup, we assume that only CHP-core's pipeline structure benefits from the low temperature, making the core evaluation conservative.

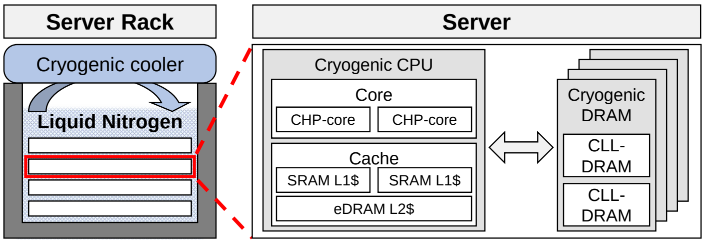


Figure 5.14: A full cryogenic computer system which the entire node is cooled down to 77K.

For the 77K memory setup, we use CryoCache [71] and CLL-DRAM [60] for its cache and DRAM designs, respectively. At 77K, CryoCache provides twice higher density and performance than conventional room-temperature caches, whereas CLL-DRAM provides 3.8 times higher speed than conventional room-temperature DRAMs. Fig. 5.14 shows the overview of our full cryogenic computer system in which the entire node is fully immersed in Liquid Nitrogen. Using this setup, we assume that CHP-core can take full advantages of cryogenic-optimal core, cache, and DRAM designs.

Power evaluation methodology

We evaluate the power consumption of CLP-core by comparing the power consumption of the four processor designs: (1) 300K hp-core, (2) 300K CryoCore, (3) 77K CryoCore, and (4) 77K CLP-core. To calculate the power consumption of each processor running at 300K and 77K, we utilize McPAT [62] integrated with cryo-MOSFET. For example, to calculate 77K CLP-core's power consumption, we first get the voltage level and leakage current at 77K from cryo-MOSFET, and then utilize them as inputs for McPAT to calculate the corresponding power. Note that 300K processors' power values derived from our methodology are similar to the McPAT's default values because the 300K transistor model of cryo-MOSFET and McPAT are both based on the ITRS loadmap [107]. We obtain the input access trace for McPAT from the gem5 simulations [12] with PARSEC 2.1 workloads [11].

Cooling cost model. In our evaluation, we include the power consumption for the cryogenic cooling because the cooling power dominates the overall power consumption at 77K. Fig. 5.14 shows the overview of our cooling system (i.e., Stinger system [10]), which recycles Liquid Nitrogen (LN) by using the cryogenic cooler. In the cooling system, the recurring electricity cost for cooling is much higher than other one-time cooling costs (e.g., cooling-facility cost, LN cost) [50, 66]. Therefore, we focus only on the cooling power consumption as the cooling cost.

$$\mathrm{P_{cooling}} = \mathrm{P_{device}} \cdot \mathrm{CO} \tag{5.2}$$

$$\begin{aligned}\mathrm{P_{77K\text{-}total}} &= \mathrm{P_{77K\text{-}device}} + \mathrm{P_{77K\text{-}cooling}} \\ &= (1 + \mathrm{CO_{77K}})\ \mathrm{P_{77K\text{-}device}} \\ &= 10.65\ \mathrm{P_{77K\text{-}device}}\end{aligned} \tag{5.3}$$

The cooling power consumption ($\mathrm{P_{cooling}}$) is the electrical power to remove the heat dissipated from the device (Eq. (5.2)). $\mathrm{P_{device}}$ is the power consumption of the electronic devices and CO is the cooling overhead [50]. The cooling overhead indicates the required power to remove unit heat (1W) from the cooling system. The cooling overhead significantly increases with the target temperature reduction, and it reaches 9.65 in 100KW-scale 77K cooling systems [50]. We use 9.65 value for our 77K cooling overhead ($\mathrm{CO_{77K}}$).

Based on Eq. (5.2), we calculate the total required power for our 77K system ($\mathrm{P_{77K\text{-}total}}$) as Eq. (5.3). Eq. (5.3) indicates that the cryogenic core should consume at least 10.65 times less power than the 300K processor to achieve the power efficiency. We exclude the cooling cost for the 300K system to conservatively show the cryogenic core's power efficiency.

Note that our cooling cost model is accurate and realistic because the cost model and modeling parameters are derived from the real data of 235 cryocoolers in 2002 [50, 102]. The cooling cost model is also conservative considering the continuously increasing power efficiency of cryo-coolers.

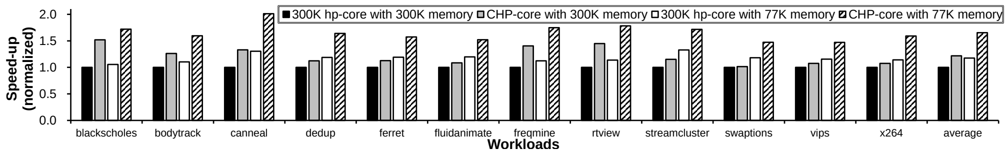


Figure 5.15: Single-thread performance of the 300K baseline (300K hp-core with 300K memory), CHP-core with 300K memory, 300K hp-core with 77K memory, and CHP-core with 77K memory.

5.5.2 Performance evaluation

Single-thread performance

Fig. 5.15 shows the single-thread performance of the various systems shown in Table 5.2. The performance is calculated by the inverse of the execution time and is normalized to that of the 300K hp-core with 300K memory system.

First, CHP-core with 300K memory achieves 21.9% of speed-up on average, up to 51.9% in *blackscholes*. Even though CHP-core's IPC is reduced due to the smaller microarchitectural units, all the workloads become faster thanks to the significantly increased clock frequency. Among the workloads, *blackscholes* achieves the highest speed-up (51.9%). On the other hand, several workloads (e.g., *fluidanimate*, *swaptions*, *vips*, *x264*) show a marginal speed-up (less than 8%) because their performance is highly bounded on memory performance [11].

Next, 300K hp-core with 77K memory achieves 17.6% of speed-up on average, up to 32.9% in *streamcluster*. The 77K memory system boosts all memory-bounded workloads as the 77K memory provides a faster access with a larger cache. However, even with the promising aspects, the cryogenic memory cannot boost the computing-bounded workloads. For example, the speed-ups of *blackscholes*, *bodytrack* and *rtview* are negligible because they cannot take the benefits of the larger and faster memory. That is, we cannot achieve the highest performance only with the 77K memory.

Different from two cases, CHP-core with 77K memory can achieve the highest

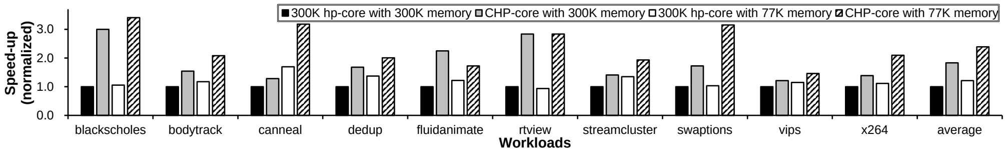


Figure 5.16: Multi-thread performance of the 300K baseline (300K hp-core with 300K memory), CHP-core with 300K memory, 300K hp-core with 77K memory, and CHP-core with 77K memory.

performance of all workloads with 65.4% speed-up on average, up to 2.01 times in *canneal*. Also, the system is 41% faster than the 300K hp-core with 77K memory. Such a significant speed-up comes from the synergetic effect of the cryogenic processor and memory. As the 77K memory resolves the memory-side bottleneck, the slow on-chip core becomes the major performance bottleneck in the system with the 77K memory. In that case, the high-performance CHP-core can fully exploit its potential. *Canneal* clearly shows the synergetic effect of the cryogenic memory and processor with 2.01 times of speed-up. The results of other workloads also support the synergetic effect by achieving their highest speed-up.

Multi-thread performance

Fig. 5.16 shows the multi-thread performance of the target systems. The multi-thread performance improvement of CHP-core is much higher than single-thread speed-up because CryoCore can fully utilize twice many cores for multi-thread execution.

First, with the 300K memory system, CHP-core achieves the speed-up of 83.2% on average, up to three times in *blackscholes*. For the computing-bounded workloads (e.g., *blackscholes*, *rtview*), CHP-core effectively doubles the multi-thread speed-up, compared to their single-thread performance gain. In addition, CHP-core also boosts the memory-bounded workloads (e.g., *dedup*, *vips*, *x264*). However, their performance improvement is much less than double because the increasing number of cores incurs higher cache contention which degrades the performance.

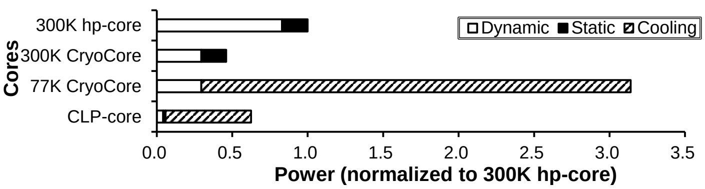


Figure 5.17: Total power consumption of the 300K hp-core (baseline), 300K CryoCore, 77K CryoCore, and CLP-core.

Next, with the 77K memories, CHP-core improves the performance by 2.39 times on average, up to 3.41 times in *blackscholes*. CHP-core with 77K memory is 100% faster than 300K hp-core with 77K memory (21.0%), which indicates the synergetic effect of using cryogenic core and memory system together. Note that the multi-thread speed-up of 300K hp-core with 77K memory (21.0%) is similar to its single-thread speed-up (17.6%). It indicates the 77K memory system cannot meaningfully improve the multi-thread performance compared to the single-thread performance. That is, CHP-core is necessary to effectively improve the system's throughput.

In summary, by utilizing CHP-core, architects can improve both the single-thread and multi-thread performance up to 2.01 times, and 3.41 times, respectively, with the same power budget even including the huge cooling cost.

5.5.3 Power evaluation

Fig. 5.17 shows the total required power consumption (including the cooling cost) of various cores. The values are normalized to the 300K hp-core's power.

In the 300K hp-core, the dynamic power occupies 83% of the total power and incurs huge cooling power consumption at 77K (as shown in Fig. 5.1). Due to its huge initial dynamic power consumption, the hp-core design cannot achieve the power efficiency at 77K, even applying the aggressive voltage scaling (as shown in Fig. 5.10).

Next, 300K CryoCore has significantly reduced power consumption thanks to the reduced pipeline width and microarchitectural units' size. The reduced size of units greatly decreases dynamic power because it decreases the power consumption per access and the CAM search overhead. Therefore, 300K CryoCore consumes 53% less dynamic power and 54% less total power consumption than the 300K hp-core.

However, reducing the number of memory entries is insufficient to achieve power efficiency at 77K. 77K CryoCore in Fig. 5.17 indicates the CryoCore without voltage scaling. Even though CryoCore design consumes 71% reduced device power than the baseline, remaining dynamic power (29.5%) incurs significant cooling power consumption (284.5%). Therefore, the total power consumption of the 77K CryoCore design is 3.1 times higher than the 300K hp-core power.

On the other hand, CLP-core consumes much less total power than the 300K hp-core. CLP-core has small initial dynamic power thanks to the smaller microarchitectural units. In addition, CLP-core further reduces the remaining dynamic power with the voltage scaling. Therefore, CLP-core consumes 37.5% less total power than the 300K hp-core. That is, architects can achieve the same single-thread performance and doubled throughput (i.e., twice more cores) with 37.5% lower power consumption by utilizing the proposed core design, CLP-core.

5.6 Thermal budget of the cryogenic processors

The thermal budget and runtime temperature analyses are crucial because the benefits of cryogenic computing come from the low-temperature environment. Thanks to the high heat-dissipation speed of LN-based cooling, the thermal budget of the cryogenic processor greatly increases at 77K. Fig. 5.18 shows the normalized heat-dissipation speed of LN-bath cooling in a low-temperature range [52]. The heat dissipation speed is defined as the heat transfer coefficient, and its value is normalized to the value of the IBM Power7 in HotSpot [94]. The dissipation speed significantly increases and becomes 2.64 times higher at 100K compared to the 300K baseline speed.

The steeply increasing heat dissipation speed can greatly increase the thermal

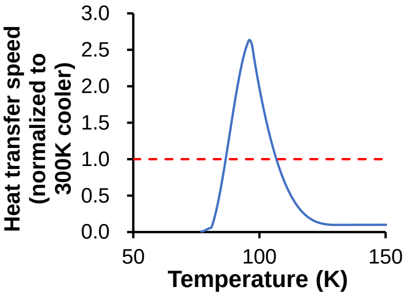


Figure 5.18: Heat dissipation speed of LN-bath cooling with temperatures

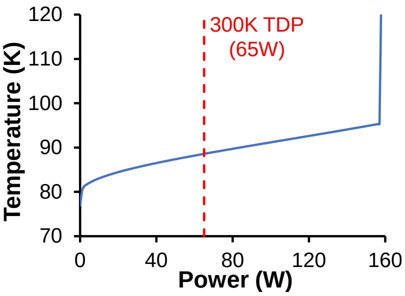


Figure 5.19: Temperature variation of the cryogenic processor with power consumption

budget of cryogenic processors. Fig. 5.19 shows the operating temperature of cryogenic processors with various power consumption (0W-160W). We utilize the validated cryo-temp (in Section 3.2.3) with HotSpot [94] and set the initial temperature to 77K. The cryogenic processor can reliably operate with 157W of power consumption, which is 2.41 times higher than the TDP of i7-6700 processors (65W). Note that the power consumption of 77K-optimal processors operating at 100K does not change significantly because the dynamic power is not affected by the temperature and the static power is still near-zero level at 100K. In addition, the power consumption of 77K-optimized processors is much lower than room-temperature processors thanks to 77K-enabled voltage scaling. That is, thermal-related problems (e.g., power wall, dark-silicon), which have been the biggest challenges for modern architects, are negligible in cryogenic processors.

5.7 CryoCore: Conclusion

In this chapter, we developed and validated CryoCore-Model (CC-Model), a cryogenic processor's performance modeling and cost analysis framework. Next, we used the tool to design CryoCore, our novel 77K optimal core microarchitecture which minimizes the core's dynamic power and area, while achieving a high clock frequency. Finally, we proposed two half-sized, differently voltage-scaled CryoCore designs aiming for either high performance or power efficiency. Our evaluation clearly indicates that cryogenic computing can significantly improve a core's single-thread and multi-thread performance or reduce its total power cost for the same die area. We also confirm the synergetic effect of the cryogenic processor and memory.

Chapter 6

Towards a Full Cryogenic Computer System

The main purpose of our CryoServer project was to design a fast and power-efficient full cryogenic computer architecture. As the first potential study of cryogenic computing, we propose the 77K-optimized DRAM (i.e., CryoRAM), cache (i.e., CryoCache) and core architectures (i.e., CryoCore), and showed the significant performance gain and power reduction when integrating them together.

Beyond the potential and architectural studies conducted in our CryoServer project, we introduce the remaining challenges in realizing cryogenic computer systems.

Device fabrication: To overcome the huge cooling power cost at 77K, we assume aggressive V_{dd} and V_{th} scaling enabled by cryogenic environment. Even though our model already considers the performance impact of the voltage scaling, it is crucial to develop the fabrication process to enable the voltage-scaled cryogenic devices. As the CMOS devices optimized for 77K operation requires completely different fabrication setup (e.g., different doping concentration), it requires non-trivial efforts to develop the 77K-optimized CMOS fabrication process. However, we believe it eventually becomes feasible in the future considering the recent interests and demonstrations of the voltage-scaled CMOS chips at cryogenic temperatures (e.g., TSMC [23], ARM [80]).

Reliability issue: As shown in Fig. 5.14, the currently most effective way to maintain 77K is immersing entire servers into Liquid Nitrogen (i.e., LN bath cooling). In the cooling system, not only the CMOS components but also other on-board units (e.g., PCB board, capacitor) are cooled down to 77K. Therefore, it is crucial to investigate their long-term reliability to realiably operate the servers at cryogenic temperatures. Due to the lack of such a reliability study, researchers should more focus on this topic to realize the full cryogenic computer system.

Cryogenic cooling system: Third, it is essential to develop the cooling systems for cryogenic computers. Many of the current cryogenic coolers have low cooling speed and cooling capacity, because most of their previous applications (e.g., superconductor magnet, quantum computer) dissipate little heat inside the cooling system. Therefore, it is important to build the large-scale cryogenic cooling systems that reliably maintain the low temperature under the huge power dissipation of servers.

Need of more circuit-level and architectural studies: Finally, even though we provide valuable insights in designing cryogenic computers, it is essential to conduct further circuit and architecture-level studies to realize cryogenic computers. For example, we should analyze the maximum DRAM channel bandwidth at cryogenic temperatures to best utilize 77K-optimized DRAM. To focus on DRAM devices only, we naively set the maximum DRAM channel bandwidth to the same as DRAM device bandwidth in Chapter 3. In addition, we should further optimize the memory hierarchy of cryogenic computer systems holistically, because we optimized each computing and memory device individually (i.e., DRAM, cache, core) and just merged them in Section 5.5. Meanwhile, we should also build 77K-optimized 3D NAND flash storages, network cards, accelerators, and graphic processing units (GPU) to realize the true sense of full cryogenic computer systems.

We believe that architects can realize the fast and power-efficient cryogenic computer systems if they resolve the aforementioned crucial challenges.

Chapter 7

Related Work

In this section, we discuss prior works related to the CMOS-based cryogenic computer architectures targeting 77K.

7.1 77K-targeted cryogenic DRAM

In 1991, Henkels et al. [39,40] from IBM fabricated 77K-optimized DRAM which was much faster than the conventional DRAM at that period. Even though their fabrication technology significantly differs from that of the latest fabrication technology, the paper showed promising characteristics of 77K-optimized DRAM at 77K (e.g., three [40] to six [39] times faster DRAM access speed, eight hours of long retention time). Tannu et al. [101] showed a critical need of cryogenic memory for quantum computers and confirmed that the commodity DRAM chips can work reliably at 80K. Wang et al. from Rambus [56, 104] studied the DRAM retention time at 77K and showed that the cryogenic environment is highly promising to reduce the DRAM refresh overhead. Ware et al. in Rambus [106] suggested that 77K DRAM is one of the most feasible memory technology to support superconducting digital processors . Lee et al. [61] found out the critical reliability issue of cryogenic DRAM due to row-hammer failure, and proposed a cryogenic-friendly row-hammer mitigation technique to resolve the problem. By using reduced leakage current and longer retention time at 77K, Bae et al. [7] Chakraborty et al. [18, 19] developed a 77K-optimized capacitor-less DRAM whose cell density and access speed are much higher than those of conventional DRAMs using huge capacitors.

7.2 77K-targeted cryogenic on-chip memory

Several previous works investigated the characteristics of various cell technologies at cryogenic temperatures for high-performance server and quantum computing applications. Garzón et al. [34–37] investigated the characteristics of various cell technologies (e.g., 3T-eDRAM, STT-MRAM) at cryogenic temperature and confirmed their superiority as on-chip cache memories at 77K. Yuhao et al. [92] and Saligram et al. [85] fabricated a 77K-optimized 3T-eDRAM memory and confirmed its potential of longer retention time and lower power consumption. Hankin et al. [38] investigated the latency and runtime power of 77K-optimized SRAM and 3T-eDRAM under various memory-access patterns, and derived key implications in designing cryogenic caches. Hu et al. [48] proposed a high-density 4T SRAM as a promising cryogenic on-chip memory technology and showed its potential of low read & write access latency, low power consumption, and cell area reduction.

7.3 77K-targeted cryogenic processor

Min et al. [72] proposed high-performance cryogenic processor and on-chip network designs by exploiting fast cryogenic wires. They observed that the faster cryogenic wires make the CPU superpipelining and shared bus feasible at 77K, and proposed the superpipelined core and shared-bus-based on-chip network architectures to improve the performance of cryogenic computers. Major industries have also been interested in developing cryogenic processors. Saligram et al. [86] from ARM developed the processor performance model using the FinFET model card and ARM Cortex-A53

RTL design, and evaluated the frequency, performance, and energy efficiency of their cryogenic processors. Prasad et al. [80] from ARM and IMEC developed a more sophisticated cryogenic processor model based on the device measurement of fabricated 14nm/16nm FinFET technologies, and showed significant performance gain and power reduction of cryogenic processors.

7.4 Other 77K-targeted cryogenic computer devices

Researchers also have been interested in processing-in-memory (PIM) and 3D NAND flash-memory architectures running at 77K. Resch et al. [84] suggested that cryogenic computing highly benefits processing-in-memory, and evaluated the performance of the PIM architectures using various cell technologies (i.e., SRAM, DRAM, STT-MRAM). Alam et al. [5] proposed the PIM architecture using a novel twisted bilayer graphene (tBLG) cell technology, and Hou et al. [44] proposed the cryogenic PIM architecture design with STT-MRAM technology. For 3D NAND flash devices, Aiba et al. [3] from Kioxia investigated cell characteristics of 3D NAND flash memories at 77K, and observed their significantly improved cell saturation current, retention time, endurance, and runtime temperature variation at 77K. By utilizing the improved cell characteristics, Aiba et al. [4] and Sanuki et al. [87] from Kioxia fabricated a 6-bit-per-cell (HLC) NAND flash memory and demonstrated their high reliability at 77K. Tanaka et al. [100] and Aiba et al. [2] further improved their cell reliability by adopting single-crystal channel and recovery annealing technologies, respectively, and successfully demonstrated the 7-bit-per-cell 3D NAND flash memories.

Chapter 8

Conclusion

High-performance computing and datacenter industries always require the fastest and most power efficient computer system. However, computer architects are now facing critical challenges to further improve performance and power efficiency of the current high-end server systems. Specifically, modern computer architectures suffer from lack of architectural innovations, mainly due to the power wall and memory wall problems. That is, architectural innovations become infeasible because they can prohibitively increase the power consumption (i.e., power wall) and their performance impacts are eventually bounded by slow memories (i.e., memory wall).

To address the challenges, making computer systems run at ultra-low temperatures (or cryogenic computer systems) has emerged as a highly promising solution as both power consumption and wire resistivity are expected to significantly reduce at the low temperatures. Thanks to the reduced leakage current of cryogenic computing, architects can increase the chip frequency without increasing the dynamic power (i.e., overcome the power wall problem). In addition, because memory latency is dominated by the wire latency, architects can greatly improve the memory performance with the cryogenic computing (i.e., solve the memory wall problem). That is, cryogenic computing have huge potential to improve performance and power efficiency of current computer systems.

However, cryogenic computers have not been yet realized mainly due to the following reason. First, there is no modeling tool available to the architects which can be used to evaluate the performance, power, and cost of the cryogenic architecture design. In addition, due to the lack of understanding about its cost-effectiveness and feasibility (e.g., device and cooling costs vs. speedup, energy and area saving), architects do not know how to build a cryogenic-optimal computer architecture.

In this dissertation, we introduced our CryoServer project, which resolves the fundamental challenges of designing a fast and power-efficient cryogenic computer system. Specifically, to realize full cryogenic computer systems, we built the performance modeling tool and developed 77K-optimized computer units for three major computer devices (i.e., DRAM, cache, and core).

First, we developed *CryoRAM*, a validated cryogenic DRAM performance modeling tool, and proposed two cryogenic-optimal DRAM designs (i.e., CLL-DRAM, CLP-DRAM) targeting for high performance and low power consumption, respectively. We also presented three promising case studies using cryogenic memories, in which we improve server performance, server power efficiency, and datacenter power efficiency, respectively.

Second, we proposed *CryoCache*, a fast, large, and power-efficient 77K-optimized cache architecture. To build CryoCache, we selected the promising cell technology candidates (i.e., SRAM, 3T-eDRAM) for cryogenic caches, developed the performance modeling framework for the selected cache technologies, and designed our 77K-optimized cache architecture which consists of 6T-SRAM cell-based L1 caches and 3T-eDRAM cell-based L2 and L3 caches.

Finally, we developed *CryoCore*, a 77K-optimized core architecture, which maximizes core's performance and area efficiency while minimizing the cooling cost. To achieve the goal, we first developed and validated cryogenic processor modeling framework, CC-Model. Then, using the framework, we identified two design principles in designing cryogenic processors. Finally, we architected our 77K-optimized core microarchitecture, which takes the narrower pipeline width and deeper pipeline depth to maximize the clock frequency and power efficiency, respectively.

We also showed the potential of full cryogenic computer systems. The full cryogenic computer systems equipped with our 77K-optimized DRAM, cache, and core designs achieves significant performance gain and power efficiency even including the cooling cost.

Bibliography

[1] Cache specifications from 7-cpu. [Online]. Available: https://www.7-cpu.com/

[2] Y. Aiba, Y. Higashi, H. Tanaka, H. Tanaka, F. Kikushima, T. Fujisawa, H. Mukaida, M. Miura, and T. Sanuki, "Demonstration of recovery annealing on 7-bits per cell 3d flash memory at cryogenic operation for bit cost scalability and sustainability," in *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*. IEEE, 2023, pp. 1–2.

[3] Y. Aiba, H. Tanaka, T. Maeda, K. Sawa, F. Kikushima, M. Miura, T. Fujisawa, M. Matsuo, H. Horii, H. Mukaida *et al.*, "Bringing in cryogenics to storage: Characteristics and performance improvement of 3d flash memory," in *2021 IEEE International Memory Workshop (IMW)*. IEEE, 2021, pp. 1–4.

[4] Y. Aiba, H. Tanaka, T. Maeda, K. Sawa, F. Kikushima, M. Miura, T. Fujisawa, M. Matsuo, and T. Sanuki, "Cryogenic operation of 3d flash memory for new applications and bit cost scaling with 6-bit per cell (hlc) and beyond," in *2021 5th IEEE Electron Devices Technology & Manufacturing Conference (EDTM)*. IEEE, 2021, pp. 1–3.

[5] S. Alam, M. M. Islam, M. S. Hossain, A. Jaiswal, and A. Aziz, "Cryocim: Cryogenic compute-in-memory based on the quantum anomalous hall effect," *Applied Physics Letters*, vol. 120, no. 14, 2022.

[6] J. W. Arblaster, “Thermodynamic properties of copper,” *Journal of Phase Equilibria and Diffusion*, vol. 36, no. 5, pp. 422–444, Oct 2015. [Online]. Available: https://doi.org/10.1007/s11669-015-0399-x

[7] 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.

[8] F. Balestra, L. Audaire, and C. Lucas, “Influence of substrate freeze-out on the characteristics of mos transistors at very low temperatures,” *Solid-state electronics*, vol. 30, no. 3, pp. 321–327, 1987.

[9] F. Balestra and G. Ghibaudo, *Device and circuit cryogenic operation for low temperature electronics*. Springer, 2001.

[10] N. Balshaw, “Practical cryogenics. and introduction to laboratory cryogenics,” 1996.

[11] C. Bienia, S. Kumar, J. P. Singh, and K. Li, “The parsec benchmark suite: Characterization and architectural implications,” in *Proceedings of the 17th international conference on Parallel architectures and compilation techniques*. ACM, 2008, pp. 72–81.

[12] N. Binkert, B. Beckmann, G. Black, S. K. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. R. Hower, T. Krishna, S. Sardashti *et al.*, “The gem5 simulator,” *ACM SIGARCH Computer Architecture News*, vol. 39, no. 2, pp. 1–7, 2011.

[13] C. Brady, “Memtest86.”

[14] K. M. Bresniker, S. Singhal, and R. S. Williams, “Adapting to thrive in a new economy of memory abundance,” *Computer*, vol. 48, no. 12, pp. 44–53, 2015.

[15] H. Cai, W. Kang, Y. Wang, L. Naviner, J. Yang, and W. Zhao, "High performance mram with spin-transfer-torque and voltage-controlled magnetic anisotropy effects," *Applied Sciences*, vol. 7, no. 9, p. 929, 2017.

[16] W. D. Callister Jr and D. G. Rethwisch, *Fundamentals of materials science and engineering: an integrated approach.* John Wiley & Sons, 2012.

[17] C. Celio, D. A. Patterson, and K. Asanovic, "The berkeley out-of-order machine (boom): An industry-competitive, synthesizable, parameterized risc-v processor," *EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2015-167*, 2015.

[18] S. Chakraborty and J. P. Kulkarni, "Cryo-tram: Gated thyristor based capacitor-less dram for cryogenic computing," in *2022 Device Research Conference (DRC).* IEEE, 2022, pp. 1–2.

[19] W. Chakraborty, R. Saligram, A. Gupta, M. San Jose, K. A. Aabrar, S. Dutta, A. Khanna, A. Raychowdhury, and S. Datta, "Pseudo-static 1t capacitorless dram using 22nm fdsoi for cryogenic cache memory," in *2021 IEEE International Electron Devices Meeting (IEDM).* IEEE, 2021, pp. 40–1.

[20] M.-T. Chang, P. Rosenfeld, S.-L. Lu, and B. Jacob, "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, 2013, pp. 143–154.

[21] 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)*, March 2012, pp. 33–38.

[22] Y. Chen, T. Luo, S. Liu, S. Zhang, L. He, J. Wang, L. Li, T. Chen, Z. Xu, N. Sun *et al.*, "Dadiannao: A machine-learning supercomputer," in *Proceedings of the*

47th Annual IEEE/ACM International Symposium on Microarchitecture. IEEE Computer Society, 2014, pp. 609–622.

[23] H.-L. Chiang, R. Hadi, J.-F. Wang, H.-C. Han, J.-J. Wu, H.-H. Hsieh, J.-J. Horng, W.-S. Chou, B.-S. Lien, C.-H. Chang *et al.*, "How fault-tolerant quantum computing benefits from cryo-cmos technology," in *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*. IEEE, 2023, pp. 1–2.

[24] K. Chun, P. Jain, J. Lee, and C. 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.

[25] K. C. Chun, P. Jain, J. H. Lee, and C. H. Kim, "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, 2009, pp. 134–135.

[26] K. C. Chun, H. Zhao, J. D. Harms, T.-H. Kim, J.-P. Wang, and C. H. Kim, "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*, vol. 48, no. 2, pp. 598–610, 2012.

[27] M. Dayarathna, Y. Wen, and R. Fan, "Data center energy consumption modeling: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 732–794, 2015.

[28] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted mosfet's with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, 1974.

[29] 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.

[30] J. Doweck, W.-F. Kao, A. K.-y. Lu, J. Mandelblat, A. Rahatekar, L. Rappoport, E. Rotem, A. Yasin, and A. Yoaz, "Inside 6th-generation intel core: New microarchitecture code-named skylake," *IEEE Micro*, vol. 37, no. 2, pp. 52–62, 2017.

[31] M. Ferdman, A. Adileh, O. Kocberber, S. Volos, M. Alisafaee, D. Jevdjic, C. Kaynak, A. D. Popescu, A. Ailamaki, and B. Falsafi, "Clearing the clouds: a study of emerging scale-out workloads on modern hardware," in *ACM SIGPLAN Notices*, vol. 47, no. 4. ACM, 2012, pp. 37–48.

[32] P. Flubacher, A. J. Leadbetter, and J. A. Morrison, "The heat capacity of pure silicon and germanium and properties of their vibrational frequency spectra," *The Philosophical Magazine: A Journal of Theoretical Experimental and Applied Physics*, vol. 4, no. 39, pp. 273–294, 1959. [Online]. Available: https://doi.org/10.1080/14786435908233340

[33] X. Fong, Y. Kim, K. Yogendra, D. Fan, A. Sengupta, A. Raghunathan, and K. Roy, "Spin-transfer torque devices for logic and memory: Prospects and perspectives," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 35, no. 1, pp. 1–22, 2015.

[34] E. Garzón, R. De Rose, F. Crupi, M. Carpentieri, A. Teman, and M. Lanuzza, "Simulation analysis of dmtj-based stt-mram operating at cryogenic temperatures," *IEEE Transactions on Magnetics*, vol. 57, no. 7, pp. 1–6, 2021.

[35] 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.

[36] E. Garzón, R. De Rose, F. Crupi, L. Trojman, A. Teman, and M. Lanuzza, "Relaxing non-volatility for energy-efficient dmtj based cryogenic stt-mram," *Solid-State Electronics*, vol. 184, p. 108090, 2021.

[37] 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.

[38] A. Hankin, L. Pentecost, D. Min, D. Brooks, and G.-Y. Wei, "Is the future cold or tall? design space exploration of cryogenic and 3d embedded cache memory," in *2023 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. IEEE, 2023, pp. 134–144.

[39] 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.

[40] W. Henkels, D.-S. Wen, R. Mohler, R. Franch, T. Bucelot, C. Long, J. Bracchitta, W. Cote, G. Bronner, Y. Taur *et al.*, "A 4-mb low-temperature dram," *IEEE journal of solid-state circuits*, vol. 26, no. 11, pp. 1519–1529, 1991.

[41] J. L. Henning, "Spec cpu2006 benchmark descriptions," *ACM SIGARCH Computer Architecture News*, vol. 34, no. 4, pp. 1–17, 2006.

[42] F. Hijaz, Q. Shi, and O. Khan, "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, 2013, pp. 85–92.

[43] 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. [Online]. Available: https://doi.org/10.1063/1.3253100

[44] Y. Hou, W. Ge, Y. Guo, L. Naviner, Y. Wang, B. Liu, J. Yang, and H. Cai, "Cryogenic in-mram computing," in *2021 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*. IEEE, 2021, pp. 1–6.

[45] C.-K. Hu, J. Kelly, J. H. Chen, H. Huang, Y. Ostrovski, R. Patlolla, B. Peethala, P. Adusumilli, T. Spooner, L. Gignac *et al.*, "Electromigration and resistivity in on-chip cu, co and ru damascene nanowires," in *2017 IEEE International Interconnect Technology Conference (IITC)*. IEEE, 2017, pp. 1–3.

[46] C.-K. Hu, J. Kelly, H. Huang, K. Motoyama, H. Shobha, Y. Ostrovski, J. H. Chen, R. Patlolla, B. Peethala, P. Adusumilli *et al.*, "Future on-chip interconnect metallization and electromigration," in *2018 IEEE International Reliability Physics Symposium (IRPS)*. IEEE, 2018, pp. 4F–1.

[47] C. Hu, *Modern semiconductor devices for integrated circuits*. Prentice Hall Upper Saddle River, NJ, 2010, vol. 2.

[48] V. P.-H. Hu, C.-J. Liu, H.-L. Chiang, J.-F. Wang, C.-C. Cheng, T.-C. Chen, and M.-F. Chang, "High-density and high-speed 4t finfet sram for cryogenic computing," in *2021 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2021, pp. 8–6.

[49] Intel, "i7-6700 skylake," 2015. [Online]. Available: https://www.7-cpu.com/cpu/Skylake.html

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

[51] J. M. Iwata-Harms, G. Jan, H. Liu, S. Serrano-Guisan, J. Zhu, L. Thomas, R.-Y. Tong, V. Sundar, and P.-K. Wang, "High-temperature thermal stability driven by magnetization dilution in cofeb free layers for spin-transfer-torque magnetic random access memory," *Scientific reports*, vol. 8, no. 1, p. 14409, 2018.

[52] T. Jin, J.-p. Hong, H. Zheng, K. Tang, and Z.-h. Gan, "Measurement of boiling heat transfer coefficient in liquid nitrogen bath by inverse heat conduction method," *Journal of Zhejiang University-SCIENCE A*, vol. 10, no. 5, pp. 691–696, 2009.

[53] N. Jing, Y. Shen, Y. Lu, S. Ganapathy, Z. Mao, M. Guo, R. Canal, and X. Liang, "An energy-efficient and scalable edram-based register file architecture for gpgpu," in *ACM SIGARCH Computer Architecture News*, vol. 41, no. 3. ACM, 2013, pp. 344–355.

[54] D. Josell, S. H. Brongersma, and Z. Tőkei, "Size-dependent resistivity in nanoscale interconnects," *Annual Review of Materials Research*, vol. 39, pp. 231–254, 2009.

[55] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)*. IEEE, 2017, pp. 1–12.

[56] T. Kelly, P. Fernandez, T. Vogelsang, S. A. McKee, L. Gopalakrishnan, S. Magee, K. Padgett, D. Barrow, J. Rizza, D. Doidge *et al.*, "Some like it cold: Initial testing results for cryogenic computing components," in *Journal of Physics: Conference Series*, vol. 1182, no. 1. IOP Publishing, 2019, p. 012004.

[57] 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.

[58] R. Kumar, D. M. Tullsen, N. P. Jouppi, and P. Ranganathan, "Heterogeneous chip multiprocessors," *Computer*, vol. 38, no. 11, pp. 32–38, 2005.

[59] T. Lanier, “Exploring the design of the cortex-a15 processor,” *URL: http://www. arm. com/files/pdf/atexploring the design of the cortex-a15. pdf (visited on 12/11/2013)*, 2011.

[60] 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: ACM, 2019, pp. 774–787. [Online]. Available: http://doi.acm.org/10.1145/3307650.3322219

[61] 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.

[62] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, “Mcpat: an integrated power, area, and timing modeling framework for multicore and manycore architectures,” in *Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture*. ACM, 2009, pp. 469–480.

[63] X. Liang, R. Canal, G.-Y. Wei, and D. Brooks, “Process variation tolerant 3t1d-based cache architectures,” in *Proceedings of the 40th Annual IEEE/ACM International Symposium on Microarchitecture*. IEEE Computer Society, 2007, pp. 15–26.

[64] K. Lim, J. Chang, T. Mudge, P. Ranganathan, S. K. Reinhardt, and T. F. Wenisch, “Disaggregated memory for expansion and sharing in blade servers,” *ACM SIGARCH computer architecture news*, vol. 37, no. 3, pp. 267–278, 2009.

[65] K. Lovin, B. C. Lee, X. Liang, D. Brooks, and G.-Y. Wei, "Empirical performance models for 3t1d memories," in *2009 IEEE International Conference on Computer Design*. IEEE, 2009, pp. 398–403.

[66] W. L. Luyben, "Estimating refrigeration costs at cryogenic temperatures," *Computers & Chemical Engineering*, vol. 103, pp. 144–150, 2017.

[67] 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.

[68] A. Mayadas and M. Shatzkes, "Electrical-resistivity model for polycrystalline films: the case of arbitrary reflection at external surfaces," *Physical review B*, vol. 1, no. 4, p. 1382, 1970.

[69] M. Mehrpoo, B. Patra, J. Gong, J. van Dijk, H. Homulle, G. Kiene, A. Vladimirescu, F. Sebastiano, E. Charbon, M. Babaie *et al.*, "Benefits and challenges of designing cryogenic cmos rf circuits for quantum computers," in *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2019, pp. 1–5.

[70] Micron, "Technical note. calculating memory power for ddr4 sdram," 2017. [Online]. Available: https://www.micron.com/-/media/documents/products/technical-note/dram/tn4007_ddr4_power_calculation.pdf

[71] D. Min, I. Byun, G.-H. Lee, S. Na, and J. Kim, "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*, 2020, pp. 449–464.

[72] 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.

[73] N. Muralimanohar, R. Balasubramonian, and N. P. Jouppi, “Cacti 6.0: A tool to model large caches.”

[74] B. Nayfeh and K. Olukotun, “A single-chip multiprocessor,” *Computer*, vol. 30, no. 9, pp. 79–85, 1997.

[75] E. J. Nowak, “Maintaining the benefits of cmos scaling when scaling bogs down,” *IBM Journal of Research and Development*, vol. 46, no. 2.3, pp. 169–180, 2002.

[76] J. K. Ousterhout, G. T. Hamachi, R. N. Mayo, W. S. Scott, and G. S. Taylor, “The magic vlsi layout system,” *IEEE Design & Test of Computers*, vol. 2, no. 1, pp. 19–30, 1985.

[77] S. Palacharla, N. P. Jouppi, and J. E. Smith, *Complexity-effective superscalar processors*. ACM, 1997, vol. 25, no. 2.

[78] B. Patra, M. Mehrpoo, A. Ruffino, F. Sebastiano, E. Charbon, and M. Babaie, “Characterization and analysis of on-chip microwave passive components at cryogenic temperatures,” *arXiv preprint arXiv:1911.13084*, 2019.

[79] D. A. Patterson and J. L. Hennessy, *Computer organization and design MIPS edition: the hardware/software interface*. Newnes, 2013.

[80] D. Prasad, M. Vangala, M. Bhargava, A. Beckers, A. Grill, D. Tierno, K. Nathella, T. Achuthan, D. Pietromonaco, J. Myers *et al.*, “Cryo-computing for infrastructure applications: A technology-to-microarchitecture co-optimization study,” in *2022 International Electron Devices Meeting (IEDM)*. IEEE, 2022, pp. 23–5.

[81] M. Qazi, M. Sinangil, and A. Chandrakasan, “Challenges and directions for low-voltage sram,” *IEEE design & test of computers*, vol. 28, no. 1, pp. 32–43, 2011.

[82] L. E. Ramos, E. Gorbatov, and R. Bianchini, “Page placement in hybrid memory systems,” in *Proceedings of the International Conference on Supercomputing*, ser. ICS ’11. New York, NY, USA: ACM, 2011, pp. 85–95. [Online]. Available: http://doi.acm.org/10.1145/1995896.1995911

[83] G. Reinman and N. P. Jouppi, “Cacti 2.0: An integrated cache timing and power model,” *Western Research Lab Research Report*, vol. 7, 2000.

[84] S. Resch, H. Cilasun, and U. R. Karpuzcu, “Cryogenic pim: Challenges & opportunities,” *IEEE Computer Architecture Letters*, vol. 20, no. 1, pp. 74–77, 2021.

[85] R. Saligram, S. Datta, and A. Raychowdhury, “Cryomem: A 4k-300k 1.3 ghz edram macro with hybrid 2t-gain-cell in a 28nm logic process for cryogenic applications,” in *2021 IEEE Custom Integrated Circuits Conference (CICC)*. IEEE, 2021, pp. 1–2.

[86] 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)*. IEEE, 2021, pp. 1–2.

[87] 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.

[88] R. R. Schaller, “Moore’s law: past, present and future,” *IEEE spectrum*, vol. 34, no. 6, pp. 52–59, 1997.

[89] 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*, vol. 33, no. 11, pp. 985–994, 2002.

[90] M. Shin, M. Shi, M. Mouis, A. Cros, E. Josse, G. T. Kim, and G. Ghibaudo, "Low temperature characterization of 14nm fdsoi cmos devices," in *2014 11th International Workshop on Low Temperature Electronics (WOLTE)*, July 2014, pp. 29–32.

[91] M. Shin, M. Shi, M. Mouis, A. Cros, E. Josse, G.-T. Kim, and G. Ghibaudo, "Low temperature characterization of 14nm fdsoi cmos devices," in *2014 11th International Workshop on Low Temperature Electronics (WOLTE).* IEEE, 2014, pp. 29–32.

[92] Y. Shu, H. Zhang, H. Sun, Q. Deng, and Y. Ha, "Csdb-edram: A 16kb energy-efficient 4t csdb gain cell edram with over 16.6 s retention time and 49.23 uw/kb at 4.2 k for cryogenic computing," in *2023 IEEE International Symposium on Circuits and Systems (ISCAS).* IEEE, 2023, pp. 1–5.

[93] M. Själander, M. Martonosi, and S. Kaxiras, "Power-efficient computer architectures: Recent advances," *Synthesis Lectures on Computer Architecture*, vol. 9, no. 3, pp. 1–96, 2014.

[94] K. Skadron, M. R. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan, "Temperature-aware microarchitecture," in *30th Annual International Symposium on Computer Architecture, 2003. Proceedings.* IEEE, 2003, pp. 2–13.

[95] R. Smith, E. Ryan, C.-K. Hu, K. Motoyama, N. Lanzillo, D. Metzler, L. Jiang, J. Demarest, R. Quon, L. Gignac *et al.*, "An evaluation of fuchs-sondheimer and

mayadas-shatzkes models below 14nm node wide lines," *AIP Advances*, vol. 9, no. 2, p. 025015, 2019.

[96] R. G. Southwick, J. Reed, C. Buu, H. Bui, R. Butler, G. Bersuker, and W. B. Knowlton, "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, 2008, pp. 48–54.

[97] W. Steinhögl, G. Schindler, G. Steinlesberger, M. Traving, and M. Engelhardt, "Comprehensive study of the resistivity of copper wires with lateral dimensions of 100 nm and smaller," *Journal of Applied Physics*, vol. 97, no. 2, p. 023706, 2005.

[98] J. E. Stine, I. Castellanos, M. Wood, J. Henson, F. Love, W. R. Davis, P. D. Franzon, M. Bucher, S. Basavarajaiah, J. Oh *et al.*, "Freepdk: An open-source variation-aware design kit," in *2007 IEEE international conference on Microelectronic Systems Education (MSE'07)*. IEEE, 2007, pp. 173–174.

[99] Synopsys, "Synopsys dc ultra," 2019. [Online]. Available: https://www.synopsys.com/implementation-and-signoff/rtl-synthesis-test/dc-ultra.html

[100] H. Tanaka, Y. Aiba, T. Maeda, K. Ota, Y. Higashi, K. Sawa, F. Kikushima, M. Miura, and T. Sanuki, "Toward 7 bits per cell: Synergistic improvement of 3d flash memory by combination of single-crystal channel and cryogenic operation," in *2022 IEEE International Memory Workshop (IMW)*. IEEE, 2022, pp. 1–4.

[101] 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*. ACM, 2017, pp. 189–195.

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

[103] D. M. Tullsen, S. J. Eggers, and H. M. Levy, “Simultaneous multithreading: Maximizing on-chip parallelism,” in *ACM SIGARCH computer architecture news*, vol. 23, no. 2. ACM, 1995, pp. 392–403.

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

[105] G. Wang, D. Anand, N. Butt, A. Cestero, M. Chudzik, J. Ervin, S. Fang, G. Freeman, H. Ho, B. Khan *et al.*, “Scaling deep trench based edram on soi to 32nm and beyond,” in *2009 IEEE International Electron Devices Meeting (IEDM)*. IEEE, 2009, pp. 1–4.

[106] 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*. ACM, 2017, pp. 183–188.

[107] L. Wilson, “International technology roadmap for semiconductors (itrs),” *Semiconductor Industry Association*, 2013.

[108] S. J. Wilton and N. P. Jouppi, “Cacti: An enhanced cache access and cycle time model,” *IEEE Journal of Solid-State Circuits*, vol. 31, no. 5, pp. 677–688, 1996.

[109] B. Wu, P. Dai, Y. Cheng, Y. Wang, J. Yang, Z. Wang, D. Liu, and W. Zhao, “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.

[110] W. Wu, S. Brongersma, M. Van Hove, and K. Maex, “Influence of surface and grain-boundary scattering on the resistivity of copper in reduced dimensions,” *Applied physics letters*, vol. 84, no. 15, pp. 2838–2840, 2004.

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

[112] W. A. Wulf and S. A. McKee, “Hitting the memory wall: implications of the obvious,” *ACM SIGARCH computer architecture news*, vol. 23, no. 1, pp. 20–24, 1995.

[113] X. Xi, M. Dunga, J. He, W. Liu, K. M. Cao, X. Jin, J. J. Ou, M. Chan, A. M. Niknejad, C. Hu *et al.*, “Bsim4. 3.0 mosfet model,” *Dept. Elect. Eng. Comput. Sci., Univ. California, Berkeley, CA, Tech. Rep*, vol. 94720, p. 30, 2003.

[114] Y. Xie, “Modeling, architecture, and applications for emerging memory technologies,” *IEEE design & test of computers*, vol. 28, no. 1, pp. 44–51, 2011.

[115] R. Zhang, M. R. Stan, and K. Skadron, “Hotspot 6.0: Validation, acceleration and extension,” *University of Virginia, Tech. Rep*, 2015.

[116] W. Zhang, S. Brongersma, Z. Li, D. Li, O. Richard, and K. Maex, “Analysis of the size effect in electroplated fine copper wires and a realistic assessment to model copper resistivity,” *Journal of applied physics*, vol. 101, no. 6, p. 063703, 2007.

[117] Y. Zhang, D. Parikh, K. Sankaranarayanan, K. Skadron, and M. Stan, “Hotleakage: A temperature-aware model of subthreshold and gate leakage for architects,” *University of Virginia Dept of Computer Science Tech Report CS-2003*, vol. 5, 2003.

[118] H. Zhao and X. Liu, “Modeling of a standard 0.35 μm cmos technology operating from 77 k to 300 k,” *Cryogenics*, vol. 59, pp. 49–59, 2014.

[119] 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.

초 록

처리해야 하는 데이터의 규모와 프로그램이 요구하는 계산량이 계속해서 많아짐에 따라, 데이터센터와 슈퍼컴퓨터 업계에서는 더욱 고성능이면서 저전력인 서버를 계속해서 필요로 하고 있다. 하지만 컴퓨터를 설계하는 아키텍트들은 2000년대 중반부터 누설전류로 인한 회로의 발열 문제 (i.e., power wall)와 와이어 저항 증가로 인한 메모리의 성능 문제 (i.e., memory wall)로 인해, 더이상 컴퓨터의 성능과 전력을 유의미하게 개선하지 못하고 있다. 저온에서 누설전류와 와이어 저항이 크게 줄어들기 때문에, 컴퓨터를 극저온 환경에서 동작시키는 극저온 컴퓨팅은 컴퓨터 성능 개선의 근본적인 문제를 극복할 수 있는 유용한 해결책으로서 각광을 맞고 있다. 하지만 극저온 환경에서 컴퓨터 부품들의 성능과 전력을 평가할 수 있는 모델링 프레임워크의 부재와, 극저온을 유지하는 냉각 비용을 극복할 수 있는 적절한 설계의 부재로, 아키텍터들은 아직 극저온 컴퓨팅을 실현하지 못하고 있다.

본 학위논문에서는 컴퓨터의 주요 부품인 DRAM, 캐쉬, 코어에 대해서 극저온에서의 성능을 평가하는 모델링 프레임워크를 개발하고, 극저온에서 전력을 최소로 하면서 고성능을 달성하는 회로 설계를 소개한다. 첫째로, 우리는 DRAM 성능 모델링 툴인 CryoRAM을 만들고 검증했고, 이를 이용하여 고성능과 저전력을 목표로 한 2가지 DRAM 디자인을 제안했다. 둘째로, 우리는 극저온 최적인 캐쉬 구조인 CryoCache를 설계했다. CryoCache는 상온 SRAM 기반 캐쉬보다 빠른 속도, 큰 capacity, 더 낮은 전력을 동시에 제공한다. 마지막으로, 우리는 극저온 최적인 코어 구조인 CryoCore를 설계했다. CryoCore는 고성능과 저전력을 달성하면서도 더 작은 면적만을 차지한다. 우리의 극저온 최적인 DRAM, 캐쉬, 코어 설계를 적용했을 때, 극저온 컴퓨터로 높은 성능과 저전력을 동시에 달성할 수 있음을 보였다.

주요어: 극저온 컴퓨팅, 차세대 컴퓨팅, DRAM, 캐쉬, 프로세서, 성능 모델링, 컴퓨터 구조, 시뮬레이션

학번: 2018-29413

ACKNOWLEGEMENT

제가 박사과정을 무사히 마칠 수 있었던 과정에는 정말로 많은 분들의 도움이 있었습니다. 이 글을 통해 감사의 인사를 드리고자 합니다.

가장 먼저, 평범한 학부생이었던 저의 가능성을 발굴해주시고 지도해주신 김장우 교수님께 진심으로 감사드립니다. 교수님께서 물심양면으로 지원해주신 덕분에 재미있고 멋진 연구를 원없이 할 수 있었고 박사 졸업까지 무사히 잘 도달할 수 있었습니다. 교수님의 제자로서 지도를 받을 수 있었던 것은 제게 정말 큰 행운이었습니다. 교수님 덕분에 연구자로서의 마음가짐과 삶을 대하는 긍정적인 자세를 배울 수 있었고, 성장으로 충만한 대학원 생활을 보낼 수 있었습니다. 박사 졸업이 끝이 아니라 시작이라는 생각으로 꾸준히 저의 부족함을 채워나가면서, 저만이 할 수 있는 멋진 연구를 해나갈 수 있도록 노력하겠습니다.

또한, 이 긴 여정을 함께 해준 SNU HPCS 연구실 동료들에게도 깊은 감사의 인사 전하고 싶습니다. 특히 학부생 인턴부터 박사 졸업까지 6년 가까운 시간 동안 동고동락하며, 친구이자 동료로서 12편의 논문을 함께 써온 변일권에게 정말 고맙습니다. 또한, 극저온팀의 믿음직한 후배이자 동료 연구자로서 저를 믿고 함께해준 김준표, 최준혁, 조정민, 정현성, 홍주원, 그리고 함께 연구했던 이규현, 정유진에게도 감사의 마음 전합니다. 그리고 저의 첫 사수로서 제게 정말 많은 것을 가르쳐주신 이재원 선배, 성공한 연구자로서 본보기가 되어주신 김영석, 장한휘, 김준성, 권동업 선배, 저보다 동생이지만 더 어른스러운 모습으로 항상 의지가 되었던 이헌준 선배에게도 감사합니다. 또한 ISCA'19 제출 10시간 전, 엉망이었던 제 첫 논문의 글을 밤을 새워가며 함께 고쳐주시며, 진정한 동료애가 무엇인지 느끼게 해주신 안재형, Amin, 배수헌 선배에게도 정말 감사합니다.

아울러, 학위논문을 심사해주신 김재준, 심재웅, 이진호, 김한준 교수님께 깊은 감사의 인사를 드리고 싶습니다. 교수님들 덕분에 저의 부족한 부분들을 깨닫고 채워나갈 수 있었고, 앞으로도 훌륭한 연구를 해 나갈 용기를 얻을 수 있었습니다.

학위 기간동안 항상 큰 힘이 되어주었던 가족들에게도 사랑한다고 전하고 싶습니다. 무엇보다도, 전화로 힘들다는 투정 뿐이었던 저를 응원해주시고, 저의 성공을 기도해주신 아버지와 어머니께 진심으로 감사드립니다. 그리고 같은 대학원생이자 학자로서 같은 길을 걷고 있는 동생에게도 고맙고 힘내라고 전하고 싶습니다. 또한 박사 기간이 생각보다 짧으니 좀더 열심히 하라고 전하고 싶습니다.

마지막으로, 비록 전공은 다르지만 밤마다 산책을 하며 연구 얘기를 나누던 동료 연구자이자, 가장 힘든 시기에 가장 큰 힘이 되었던 김주연 양에게도 진심으로 감사의 인사 전합니다.