Cryogenic Computer Architecture Modeling with Memory-Side Case Studies

Gyu-hyeon Lee[*]
Department of Electrical and Computer Engineering
Seoul National University
Seoul, Republic of Korea
guhylee@snu.ac.kr

Dongmoon Min[*]
Department of Electrical and Computer Engineering
Seoul National University
Seoul, Republic of Korea
dongmoon.min@snu.ac.kr

Ilkwon Byun[*]
Department of Electrical and Computer Engineering
Seoul National University
Seoul, Republic of Korea
ik.byun@snu.ac.kr

Jangwoo Kim[†]
Department of Electrical and Computer Engineering
Seoul National University
Seoul, Republic of Korea
jangwoo@snu.ac.kr

ABSTRACT

Modern computer architectures suffer from lack of architectural innovations, mainly due to the power wall and the memory wall. That is, architectural innovations become infeasible because they can prohibitively increase power consumption and their performance impacts are eventually bounded by slow memory accesses. 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 ultra-low temperatures. However, cryogenic computers have not been yet realized as computer architects do not fully understand the behaviors of existing computer systems and their cost effectiveness at such ultra-low temperatures.

In this paper, we first develop *CryoRAM*, a validated computer architecture simulation tool to incorporate cryogenic memory devices. For this work, we focus on 77K temperature (easily achieved by applying low-cost liquid nitrogen), at which modern CMOS devices still reliably operate. We also focus on reducing the temperature of memory devices only as a pilot study prior to building a full cryogenic computer. Next, driven by the modeling tool, we propose our temperature-aware memory device and architecture designs to improve the DRAM access speed by 3.8 times or reduce the power consumption to 9.2%. Finally, we provide three promising case studies using cryogenic memories to significantly improve (1) server performance (up to 2.5 times), (2) server power (down to 6% on average), and (3) datacenter's power cost (by 8.4%).

We will release our modeling and simulation tools deliberately implemented on top of only open-source simulators combined, even though some experiments were conducted under industry-confidential environments.

CCS CONCEPTS

• **Hardware** → **Dynamic memory**; **Chip-level power issues**; **Enterprise level and data centers power issues**; **Modeling and parameter extraction**; **Analysis and design of emerging devices and systems**; **Emerging architectures**; *Transistors*; *Temperature simulation and estimation*; *Memory and dense storage*.

KEYWORDS

Cryogenic computing, DRAM, Memory, Simulation, Modeling

ACM Reference Format:
Gyu-hyeon Lee, Dongmoon Min, Ilkwon Byun, and Jangwoo Kim. 2019. Cryogenic Computer Architecture Modeling with Memory-Side Case Studies. In *The 46th Annual International Symposium on Computer Architecture (ISCA '19), June 22–26, 2019, Phoenix, AZ, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3307650.3322219

1 INTRODUCTION

To build a faster computer under the same power budget, both Moore's Law [25] and Dennard Scaling [11] 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* [34]).

To get around the power and memory wall problems, various approaches have been proposed such as the deployment of slow

[*] Three authors contributed equally to the paper.
[†] Corresponding author.

ISCA '19, June 22–26, 2019, Phoenix, AZ, USA

ACM ISBN 978-1-4503-6669-4/19/06…$15.00
https://doi.org/10.1145/3307650.3322219

multi-core designs [20, 22] 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.

In this paper, we resolve the challenges in realizing a cryogenic computer as follows. First, we develop *CryoRAM*, a validated cryogenic memory simulation module, which can be plugged into existing architecture simulators. This work focuses on cryogenic DRAM devices because our current infrastructure can validate only memory modules and it is easy to analyze the cryogenic impact of memory devices isolated from other components. In addition, we set 77K (or -196°C) as our target temperature which can be easily achieved with low-cost liquid nitrogen applied and at which temperature modern CMOS technologies still reliably operate as confirmed by DRAM vendors.

CryoRAM combines three modules (*cryo-pgen*, *cryo-mem*, *cryo-temp*) which model MOSFET, DRAM, and thermal behaviors, respectively. The MOSFET model, when given with the target fabrication technology and operating voltage information, generates the parameters of the result cryogenic CMOS device (e.g., on-channel current and leakage current). The DRAM model takes the cryogenic MOSFET parameters and produces a cryogenic DRAM device architecture optimized for performance and power efficiency with its area, latency, and energy information. The thermal model measures the temperature of the cryogenic DRAM device to see how it maintains the target temperature during run time. We implemented CryoRAM on top of existing open-source memory (CACTI [33]), temperature (HotSpot [36]), transistor (BSIM [35]) models.

To validate CryoRAM, we validated all three sub-models by comparing them with the corresponding transistor and device samples. As we are not allowed to disclose the results obtained at the industry facility, this paper presents results obtained from non-commercial transistor samples fabricated with a large technology and commodity DRAM modules, but we confirmed that what the model presents and projects is reasonably accurate for the latest CMOS technologies. Note that CryoRAM's model projection cannot be validated with real samples because prototyping a cryogenic memory module requires to change the current fabrication process (i.e., doping level, V_{dd}, V_{th}, etc.).

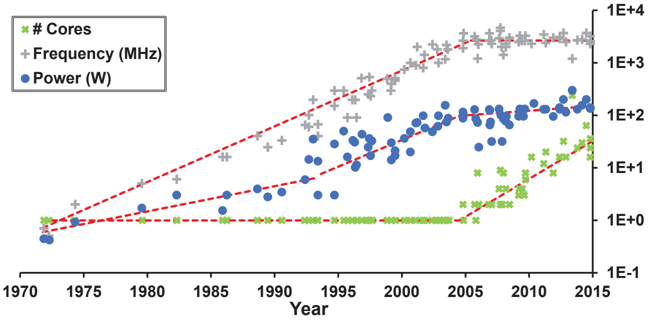


Figure 1: End of single-core performance improvement due to the power wall problem

As a result, our CryoRAM analysis for the 28nm technology produced two cryogenic DRAM memory devices which improve the DRAM access speed by 3.8 times or reduce the power consumption to 9.2%. We call these DRAM devices as *cryogenic low-latency DRAM (CLL-DRAM)* and *cryogenic low-power DRAM (CLP-DRAM)*, respectively. Also, our thermal model confirmed the target cryogenic temperature (i.e., 77K) is well maintained when the DRAM devices operate.

Next, using the cryogenic DRAM devices, we construct three promising case studies. We first construct a server with its L3 cache disabled, but equipped with CLL-DRAM devices. The decreased DRAM access latency comparable to the L3 hit latency can improve the server performance by up to 2.5 times even with the area and power-critical L3 cache disabled. Next, another server structure equipped with CLP-DRAM devices can reduce the memory-side power consumption down to 12.7% when running memory-intensive applications. Our last case study assumes a modern datacenter which replaces 7% of its conventional DRAM devices with CLP-DRAM devices, and dynamically schedules hot memory requests to CLP-DRAM devices to minimize the common-case DRAM access energy. Our cost-effectiveness analysis (e.g., cryogenic cooling cost, memory peak power) shows that it can reduce the modern datacenter's total power cost by 8.4%.

In summary, our work makes the following contributions:

- **Novel Cryogenic Computer Architecture Modeling:** To the best of our knowledge, this is the first work in our community to model, validate, and project the potential of cryogenic computer architectures from the perspective of computer architects.
- **High Impact of Cryogenic Computer Architectures:** We show the high performance and energy-saving impact of three cryogenic computer designs employing cryogenic-optimized DRAM devices.
- **CryoRAM Simulator Release:** We release our CryoRAM tool to the community. The tool can be easily plugged into the existing architecture simulators.

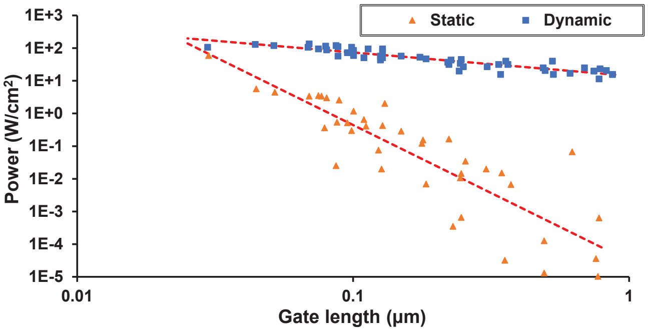
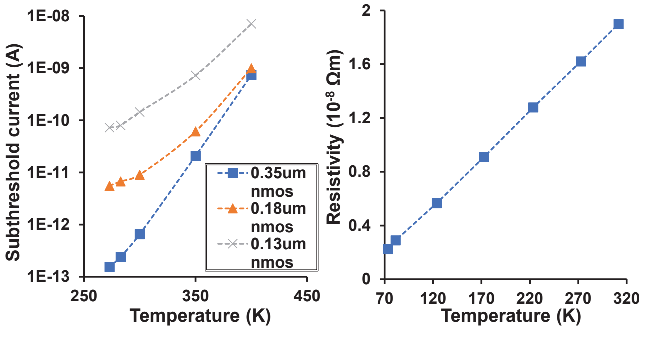


Figure 2: Steep increase of static power with shrinking device size



(a) (b)

Figure 3: Benefits of cryogenic computing (a) Exponentially decreasing subthreshold leakage current; (b) Linearly decreasing wire resistivity

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 [7], as shown in Fig. 1.

$$P_{\mathrm{static}} = V_{\mathrm{dd}} I_{\mathrm{leak}}, \quad I_{\mathrm{leak}} \propto e^{-\frac{qV_{th}}{kT}} \tag{1a}$$

$$P_{\mathrm{dynamic}} \propto V_{\mathrm{dd}}^2 f \tag{1b}$$

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. (1a) shows that static power (P_{static}) is exponentially inversely proportional to V_{th} due to the increasing leakage current (I_{leak}) [19]. In this situation, to increase the chip frequency (f) while maintaining its dynamic power (P_{dynamic} in Eq. (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 [23] (as in Fig. 2). Therefore, computer architects have not been able to improve the single-core 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 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 [20, 22] (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. (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 [26] (as in Fig. 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. 3b shows the wire metal's resistivity (e.g., copper) reduces to 15% of the room temperature [8]. As the circuit delay is mainly determined by the RC delay (=resistance×capacitance), cryogenic computing can significantly improve the speed of circuit computations and wire transfers. 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*).

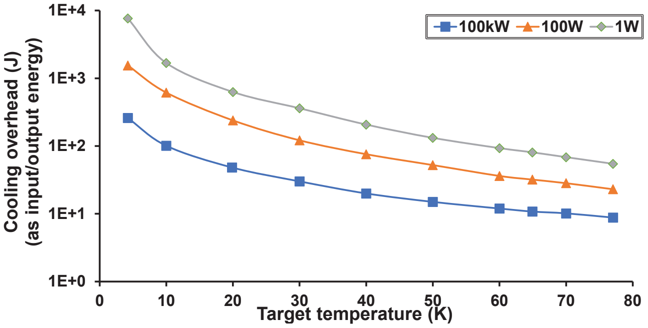


Figure 4: Cooling overhead as the relative amount of the input energy to reach the target temperatures. The legend indicates the efficiency of coolers as their cooling speed.

2.3 Challenges of cryogenic computing

There exist several challenges to realize cryogenic computer systems. The absence of an architectural modeling tool is one of the major challenges. Even though 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. 4 shows the overhead to achieve various target temperatures [17]. 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.

2.4 Research target: 77K-optimized DRAM

In this work, we first develop *CryoRAM*, an accurate computer architecture modeling tool which targets 77K-optimized DRAM modules. After validating our CryoRAM model, we use it to design various 77K-optimized memory modules which can effectively resolve the power and memory walls.

Among two representative cryogenic temperatures, this work focuses on 77K because modern CMOS devices reliably operate at the temperature. 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 [2].

Among the computer architecture modules, this work focuses on memory devices due to the following reasons. First, we assume that it is feasible to cool memory devices isolated from the rest of systems. Second, we narrow down the validation scope to DRAM modules with our current validation infrastructure considered. Third, we believe that cryogenic DRAM modules can significantly improve the performance of memory latency-critical workloads [12], while significantly reducing the energy cost of operating modern datacenters equipped with an increasing number of memory modules [10].

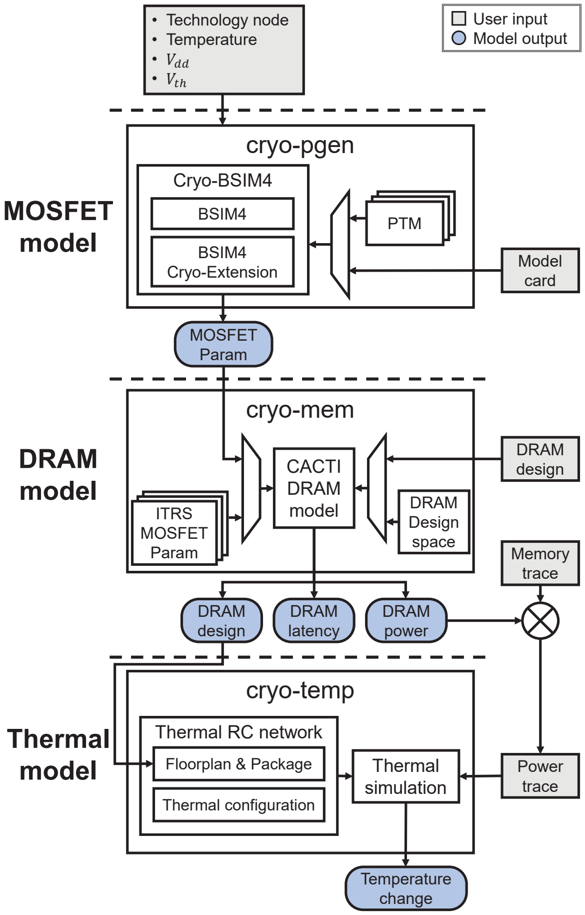
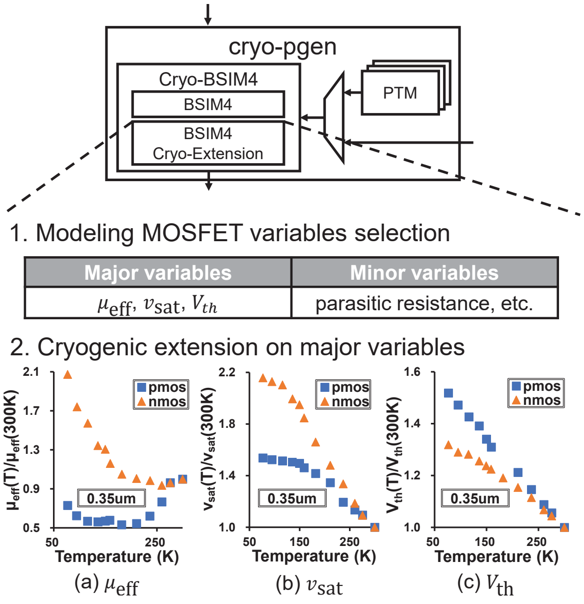


Figure 5: CryoRAM overview

3 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. 5. 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.



(a) μ_{eff} (b) v_{sat} (c) V_{th}

Figure 6: Cryogenic extension to the baseline MOSFET model (a) Carrier mobility model; (b) Saturation velocity model; (c) Threshold voltage model

3.1 MOSFET model

3.1.1 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 [35], 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.

3.1.2 Cryogenic extension. We apply a cryogenic extension to BSIM4 as shown in Fig. 6. 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}).

$$\mu_{eff} = \frac{U_0(T)}{\text{Surface Scattering}(T, E_{eff})} \tag{2}$$

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. (2), 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.

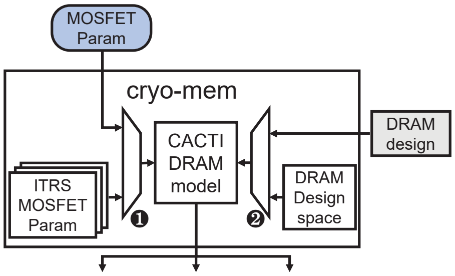


Figure 7: Cryogenic extension to DRAM model to add two input interfaces ❶ MOSFET parameters; ❷ DRAM design

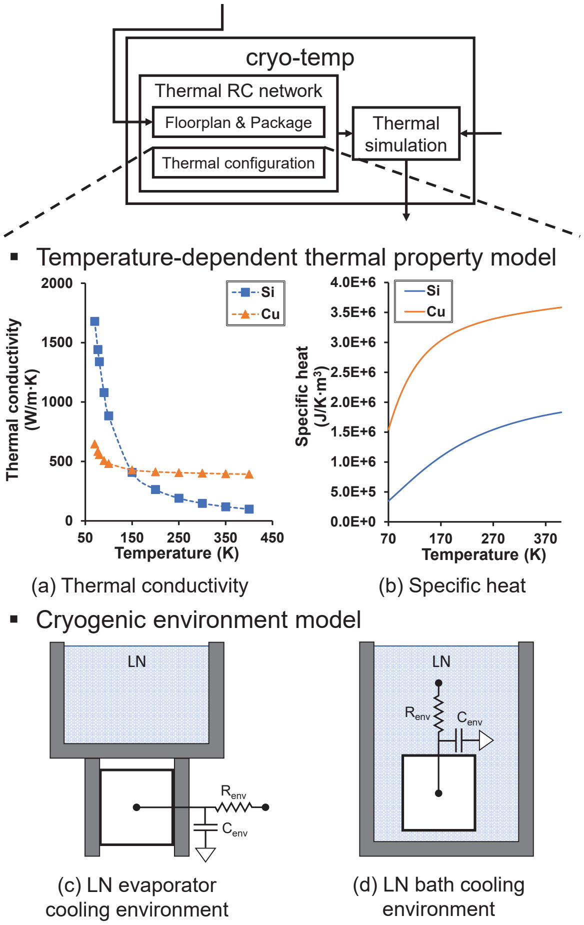
3.1.3 Implementing cryo-pgen. Our cryogenic MOSFET modeling tool, *cryo-pgen*, can generate the output MOSFET parameters at 77K as follows. First, it takes the target fabrication process information from a current room-temperature input model card available as two options: a vendor-driven model card and an open-source PTM model which supports from 180nm to 16nm at 300K [38]. 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 [27, 37] are provided to cryo-pgen as shown in Fig. 6. By assuming that the ratios of three terms at 300K and a low temperature T (i.e., $\mu_{eff}(T)/\mu_{eff}(300K)$, $v_{sat}(T)/v_{sat}(300K)$, $V_{th}(T)/V_{th}(300K)$) 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.

3.2 DRAM model

3.2.1 Baseline memory model. As a baseline memory modeling tool, CryoRAM uses CACTI [33], 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 consumption.

However, CACTI cannot be directly applied to cryogenic memories due to two reasons. First, it uses ITRS [32] MOSFET parameters valid only at 300K-400K. Second, CACTI cannot apply different temperatures to a fixed target memory design.



(a) Thermal conductivity

(b) Specific heat

(c) LN evaporator cooling environment

(d) LN bath cooling environment

Figure 8: Cryogenic extension to the thermal model (a)(b) Temperature-dependent thermal property models; (c)(d) Cryogenic cooling environment models

3.2.2 Implementing cryo-mem. We implement our DRAM modeling tool, *cryo-mem*, by adding cryogenic extensions to CACTI-3dd [9]. Fig. 7 shows the overview of cryo-mem. First, we add an interface to CACTI to accept MOSFET parameters produced by cryo-pgen (Fig. 7❶). Second, we add an interface to CACTI to accept and fix a specific DRAM design while applying different temperatures (Fig. 7❷). 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.

3.3 Thermal model

3.3.1 Baseline thermal model. As a baseline architecture-level temperature modeling tool, CryoRAM uses HotSpot [36] 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. 8. Second, HotSpot does not model the cryogenic-cooling method.

3.3.2 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 [1, 13, 16] as shown in Fig. 8 (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. 8 (c) and (d). The LN evaporator model indirectly cools a target device with temperature conduction via metal plates, which is assumed in Section 4.4. The LN bath cooling directly cools a target device by fully immersing it in LN (Fig. 8) which is assumed in Section 5.1.

4 MODEL VALIDATION

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

4.1 Experimental setup

Fig. 9a 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. 9b shows our experimental setup to validate cryo-mem and cryo-temp. We 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.

4.2 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. 10 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: 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. 10). 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 [19]. 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.

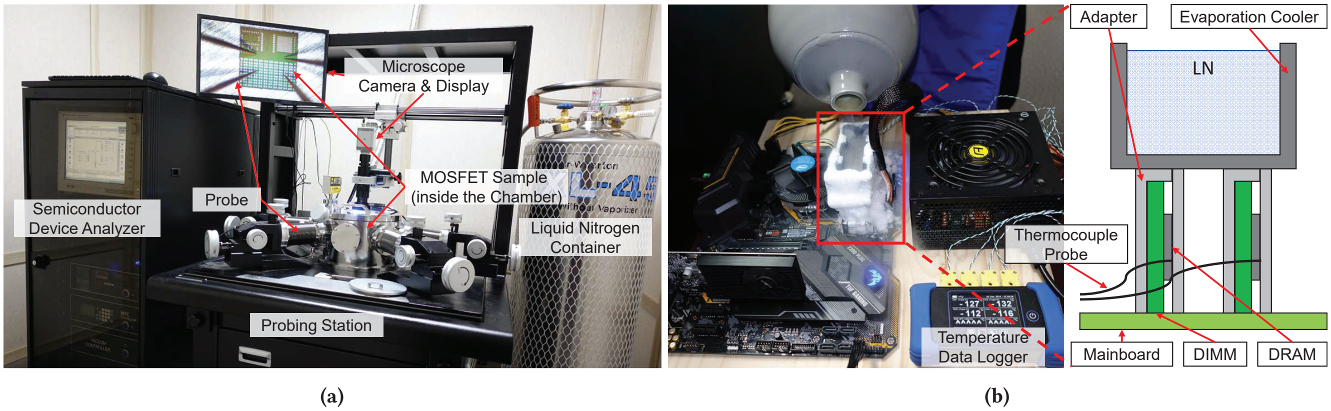
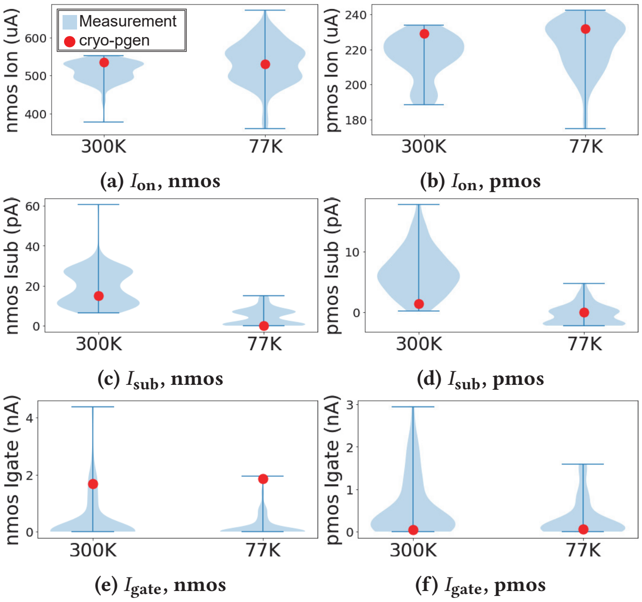


Figure 9: Experimental validation setup for (a) MOSFET modeling (cryo-pgen) and (b) DRAM modeling (cryo-mem, cryo-temp)



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

4.3 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 4.4).

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+ [6] is running (Fig. 9b). 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.

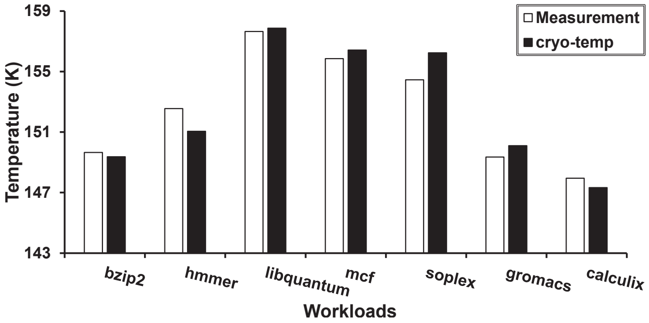


Figure 11: cryo-temp validation results running SPEC2006 workloads: real measurements vs. cryo-temp outcomes

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.

4.4 Thermal model validation

We validate cryo-temp by comparing the DRAM temperature in the real system (Fig. 9b) with the cryo-temp's prediction with the LN evaporator model. We run several SPEC CPU2006 workloads (bzip2, hmmer, libquantum, mcf, soplex, gromacs and calculix) [15] 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 [5] simulation.

Fig. 11 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.

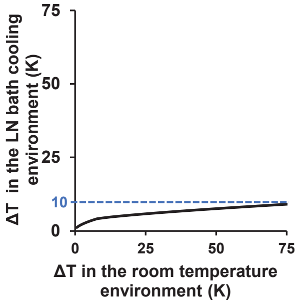


Figure 12: Temperature variations: at room temperature environment vs. LN bath cooling environment

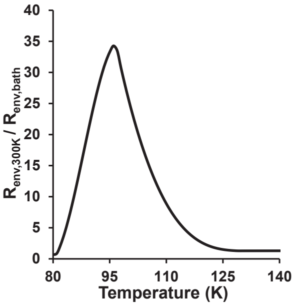


Figure 13: Thermal resistance (R_{env}) variations as the resistance ratio ($R_{env,300K}/R_{env,bath}$)

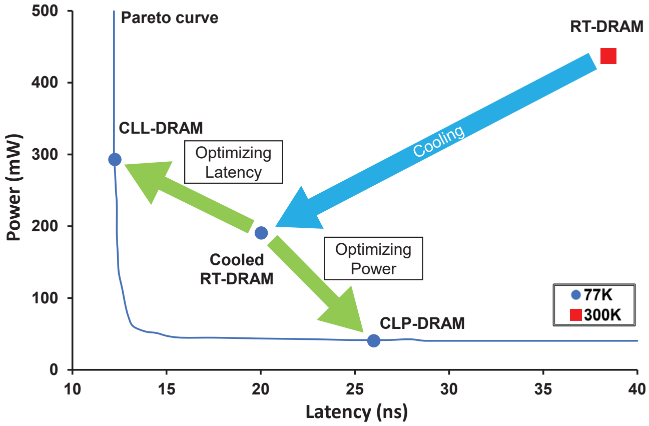


Figure 14: Finding cryogenic-optimal DRAM designs by the design-space explorations with V_{dd} and V_{th} sweeping

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.

5 MODELING RESULTS

In this section, we perform a series of CryoRAM-driven experiments to justify our target 77K environment temperature and to propose cryogenic DRAM device models. First, with cryo-temp, we show the target environment temperature is well preserved. Next, with cryo-pgen and cryo-mem, we show the potential of cryogenic-optimized DRAM devices for higher energy efficiency and/or performance.

5.1 Maintaining the cryogenic temperature

As the main benefits of cryogenic computing come at the target low temperature, it is important to ensure that the cryogenic memory remains in the low-temperature ranges. Therefore, using cryo-temp, we simulate and measure the temperature variations of DRAM in the LN bath cooling environment as well as in the room temperature environment, and compare them.

In Fig. 12, DRAM with the LN bath cooling shows negligible temperature variations (<10K) whereas the counterpart's temperature rises over 75K. The decreasing temperature variations come mainly from the low R_{env} of the LN bath cooling [18]. R_{env} is a thermal resistance contributing to the heat transfer between the device and the surrounding environment. Thus, the low R_{env} increases the heat transfer speed due to the reduced thermal RC delay.

Fig. 13 shows the ratio of R_{env} between the room temperature and the LN bath cooling environment ($R_{env,300K}/R_{env,bath}$) at different temperatures. This graph indicates the heat dissipation speed becomes significantly high near 96K (about 35 in maximum).[1] In this situation, once the temperature reaches 77K, the steep rise of the heat dissipatoin speed prevents the temperature from rising over 96K. Therfore, our memory system will be well maintained at the target low temperature.

Based on this results, 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 as the run-time temperature variations would be minimal.

5.2 Deriving two cryogenic DRAM devices

In this section, we show the 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 the sum of the static power and the dynamic power. We conservatively model the DRAM's refresh using the room-temperature retention time of commercial DRAM (64ms).

Fig. 14 indicates that cooling a commercial DRAM (room temperature DRAM; RT-DRAM) to 77K reduces its latency and power by 48.9% and 43.5%, respectively (Cooled RT-DRAM) thanks to the decreased wire resistivity and static power.

Interestingly, we can further improve the power or the performance by scaling the cooled RT-DRAM's threshold voltage (V_{th}) and operating voltage (V_{dd}). Note that the near-zero leakage current at the low temperature allows aggressive V_{th} and V_{dd} reduction.

To find optimal cryogenic DRAM designs, we explore 150,000+ DRAM designs with different V_{dd} and V_{th}, and obtain a latency-power Pareto optimal curve as the result (Fig. 14). Among various points in the curve, we choose two representative DRAM designs: the power-optimal design (CLP-DRAM) and the latency-optimal design (CLL-DRAM).

Cryogenic Low-Power DRAM (CLP-DRAM): Reducing V_{dd} and V_{th} by half can significantly decrease the dynamic power. We can thus construct an ultra-low power DRAM design (CLP-DRAM) for the cryogenic environment. Fig. 14 shows that CLP-DRAM consumes only 9.2% of power compared to that of RT-DRAM. This huge power reduction comes from the further reduced dynamic power and almost eliminated static power. But, the CLP-DRAM's latency is still only 65.3% of RT-DRAM (or 1.53 times faster).

Cryogenic Low-Latency DRAM (CLL-DRAM): Using high V_{dd} and low V_{th} can significantly improve I_{on} of MOSFET and thus the DRAM speed. For the purpose, we maintain V_{dd} as that of RT-DRAM and scale down only V_{th} by half. As a result, we

[1] The existence of peak follows the physics of the boiling liquid near the hot surface [4].

Table 1: Parameter setup for single-node level case studies

CPU specification		
Cores	Based on the Intel i7-6700 (3.5GHz)	
LLC	12MB, 16way set-assoc, shared, 42cyc (=12ns)	
DRAM access latency[2]		
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

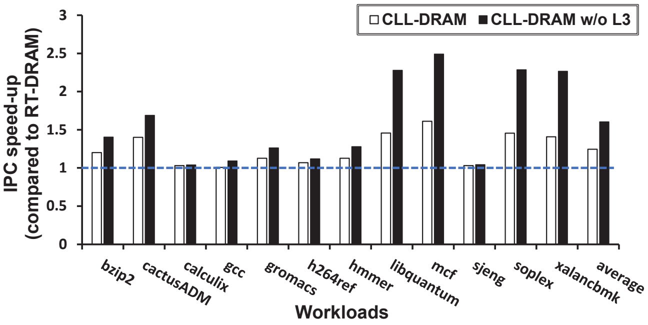


Figure 15: Performance improvement of a single node equipped with CLL-DRAM for SPEC CPU2006 workloads (with L3 cache or without L3 cache)

can make an ultra-low latency DRAM design (CLL-DRAM) for the cryogenic environment. It is 3.80 times faster than RT-DRAM. The huge latency reduction comes from the high I_{on} and low wire resistivity. But, the CLL-DRAM's power consumption remains still lower than that of RT-DRAM (Fig. 14).

6 SINGLE-NODE LEVEL CASE STUDIES

In this section, by using the cryogenic DRAM devices (CLL-DRAM and CLP-DRAM) obtained in Section 5.2, we show two case studies to improve either the performance or the power efficiency of a single server node equipped with the cryogenic memories.

6.1 Evaluation setup

For these case studies, we use gem5 timing simulator [5] with the configuration specified in Table 1. We set CPU-related parameters based on the Intel i7-6700 processor. To set the cryogenic DRAM access latency and power, we use the values derived from Section 5.2 as summarized in Table 1. Our room-temperature baseline node uses RT-DRAM. For workloads, we use 12 workloads chosen from the SPEC CPU2006 benchmark set[15].

6.2 IPC speed-up with CLL-DRAM

Fig. 15 shows the IPC improvement when a server uses 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. But, the performance of less memory-intensive workloads (e.g., calculix, gcc) are insensitive to the reduced memory access latency. Even with the conservative mix of workload choice, CLL-DRAM improves the single node performance by 24% on average.

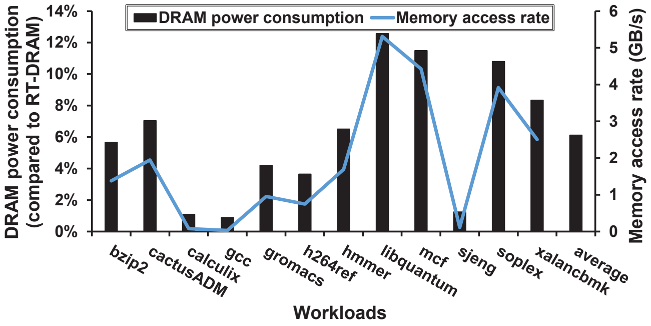


Figure 16: Power consumption of CLP-DRAM (normalized to RT-DRAM) with the memory access rates of SPEC CPU2006 workloads

Next, we measure the server performance with the L3 cache disabled. Note that CLL-DRAM's latency (15.84ns) becomes now comparable to the L3 cache latency (12ns). Then, it can be more beneficial to avoid L3 cache miss penalties by bypassing the L3 cache and directly accessing the CLL-DRAM. Fig. 15 shows that the average performance increases by 60% when L3 cache is disabled (CLL-DRAM w/o L3). For memory-intensive workloads (i.e., libquantum, mcf, soplex, and xalancbmk), the average performance improvement rises up to 2.3 times on average and 2.5 times in maximum, respectively. In addition, with the area- and static-power critical L3 caches removed, architects can invest other logics to the reclaimed die area (e.g., more cores) or save the on-chip power.

6.3 DRAM power reduction with CLP-DRAM

Fig. 16 shows the DRAM power consumption of a node using CLP-DRAM. The power consumption of Fig. 16 is normalized to that of a node using RT-DRAM. To calculate the DRAM power, we add the dynamic power and the static power based on the memory access rate obtained from each workload. As a result, the DRAM power consumption is reduced to 6% on average. For less memory-intensive workloads (e.g., calculix, gcc, sjeng), the power reduction increases up to >100 times because the cryogenic-friendly static power dominates the dynamic power for such workloads.

7 DATACENTER LEVEL CASE STUDY

The cryogenic computing is often considered highly promising for datacenters, in which the reduction of its total power cost can outweigh the cryogenic cooling cost. However, replacing all DRAMs in a datacenter with CLP-DRAMs could be too expensive. Therefore, we propose a Cryogenic Low-Power Architecture for datacenters (CLP-A) which achieves high power reduction by replacing a minimal number of RT-DRAMs with CLP-DRAMs.

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

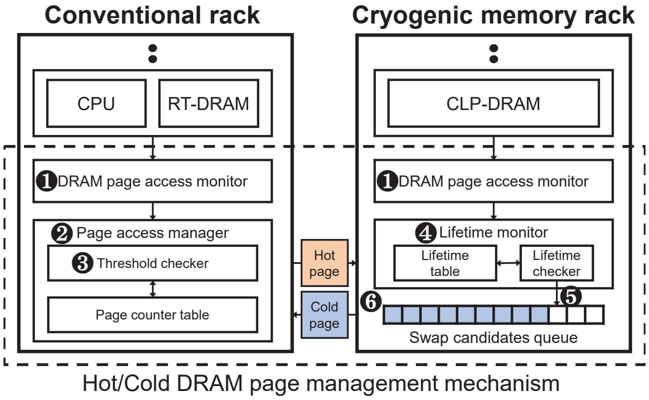


Figure 17: Cryogenic Low-Power Architecture: overview

7.1 Cryogenic Low-Power Architecture (CLP-A)

7.1.1 CLP-A overview. The basic idea of CLP-A is to migrate hot pages to a small set of CLP-DRAMs in order to minimize the overall DRAM power consumption. Fig. 17 shows the CLP-A's overview. CLP-A consists of two types of racks: a large set of conventional racks and a small set of cryogenic memory racks. The datacenter cools down only CLP-DRAMs.

Using on a hot-page management scheme (Section 7.1.2), CLP-A aims to maintain a relatively smaller set of hot pages (i.e., high-locality pages) in the disaggregated CLP-DRAMs. If hot pages are frequently accessed, CLP-A can successfully reduce the total DRAM power consumption with a small number of CLP-DRAMs.

7.1.2 CLP-A's page management mechanism. Our hot page management mechanism is based on [24], and we first summarize key terms as follows.

Hot page: The hot page indicates a frequently-accessed DRAM page. CLP-A maintains an access counter for each page, which is increased at every memory access. A target page becomes hot when its counter value exceeds a "threshold". The counters are reset after their "counter lifetime" from the last access.

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

Fig. 17 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. If the counter exceeds a certain threshold (❸), the threshold checker categorizes the page as a hot page and migrates it to the remote CLP-DRAM.

In cryogenic memory racks, the lifetime monitor manages the lifetime of hot pages. For every page access, it resets the lifetime of the hot page (❹). The lifetime checker registers the lifetime-expired hot pages in the swap candidates queue (❺). When a new hot page arrives, it is swapped with one of swap candidates (❻). If the CLP-DRAMs are full and there are no swap candidates, CLP-A waits until the queue gets a new candidate.

Table 2: Parameter setup for CLP-A

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%	

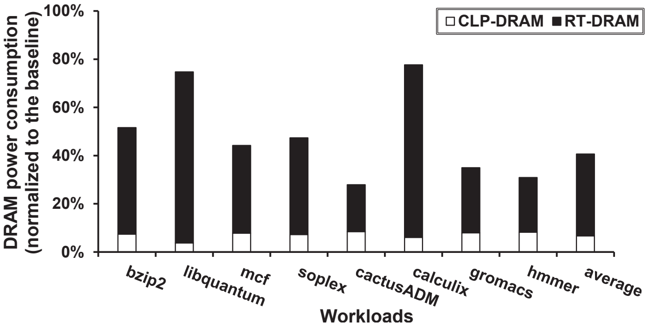


Figure 18: DRAM power consumption of CLP-A for SPEC CPU2006 workloads normalized to that of conventional datacenter

7.2 CLP-A: DRAM power evaluation

To evaluate CLP-A, we compare the CLP-A's DRAM power reduction with a conventional datacenter using only RT-DRAMs. 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 [15]. Table 2 shows the key parameters used for the experiments.

Our detailed parameter setup is explained as follows. First, we set the CLP-DRAM access latency to be the same as the RT-DRAM access latency to conservatively model the inter-rack interconnect latency. Second, we model the page swap overhead by setting the swap latency as 1.2μs [24] and conservatively assume that the RT-DRAM serves memory accesses during the page swap. Third, 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, the hot-page lifetime, and the amount of CLP-DRAMs obtained based on the design-space explorations to find the optimal values. As a result, we take 7% as the ratio of CLP-DRAMs to the total DRAMs, and 200μs as the counter lifetime and the hot page lifetime.

Fig. 18 shows the DRAM power consumption in CLP-A. The results highly vary among the workloads. For example, CLP-A reduces 72% of DRAM power consumption for cactusADM, but only 23% of the power for calculix. Such a large difference results from the memory access patterns of each workload. Moving a page to the CLP-DRAM takes 1.2μs since the page was 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

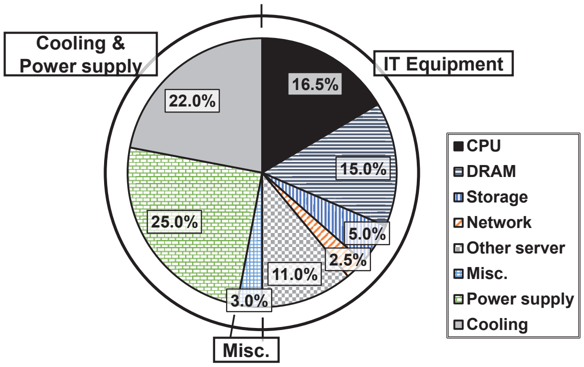


Figure 19: Power breakdown of conventional datacenter [10]

due to the swap overhead. However, even with such workloads, CLP-A exhibits large power reduction (59% on average). With only the small set of CLP-DRAMs used (7% of total DRAMs), CLP-A significantly reduces the DRAM power consumption of the target datacenter.

7.3 Cryogenic datacenter power modeling

In this section, we conduct the datacenter-level power/cost analysis as the cooling overhead would affect the efficiency of power reduction.

7.3.1 Conventional datacenter power modeling. Fig. 19 shows the power breakdown of a typical datacenter [10]. The datacenter power can be divided into three major categories: IT Equipment (50%), Cooling/Power Overhead (47%), and Misc. (3%).

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

$$Cooling = C.O. \times \text{IT Equipment} \tag{3a}$$

$$Power\ Supply = P.O. \times \text{IT Equipment} \tag{3b}$$

Cooling & Power Supply: Cooling is the power consumed to cool the IT components, and Power Supply is the power consumed by power supply. We categorize the two items as a same group because they have a similar relationship with IT Equipment.

To model Cooling & Power supply, we use the linear model shown as Eq. (3). In Eq. (3a), $C.O.$ indicates the cooling overhead (Section 2.3), and $P.O.$ indicates the power overhead as the amount of wasted energy while supplying unit energy (1 J) to the IT components. We use the linear model to make our cost analysis conservative because both of them decrease faster than linearly with the decreasing IT Equipment power [10]. As our evaluations only consider the case of the decreased IT Equipment power, the choice of linear model always exaggerates the cost of Cooling & Power Supply which makes our cost analysis more reliable.

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.

$$\begin{aligned}
&\text{Conventional datacenter power (at room temperature)}\\
&= \text{IT Equipment} + \text{Cooling \& Power Supply} + \text{Misc.}\\
&= \text{IT Equipment} + (C.O._{300\text{K}} + P.O._{300\text{K}}) \cdot \text{IT Equipment} + \text{Misc.}\\
&= \text{IT Equipment} + (\frac{22}{50} + \frac{25}{50}) \cdot \text{IT Equipment} + \text{Misc.}\\
&= 1.94 \cdot \text{IT Equipment} + \text{Misc.} \qquad (4)
\end{aligned}$$

With Eq. (3) applied, the total power consumption of a conventional datacenter can be summarized as Eq. (4). Note that $C.O._{300\text{K}}$ is the ratio of Cooling (22%) and IT Equipment (50%) in Fig. 19. $P.O._{300\text{K}}$ is the ratio of Power Supply (25%) and IT Equipment.

7.3.2 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: The one-time incurred initial cost consists of LN cost and facility cost. We assume LN recycling "stinger system [3]", 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, the one-time cost is paid once.

Recurring cost: The recursively incurred cooling-power consumption (i.e., electricity) accounts for the majority of the cooling cost. The cooling power for a cryogenic datacenter can be modeled also as Eq. (3a). But, $C.O._{77\text{K}}$ is much higher than the conventional datacenter counterpart. Note that the cryogenic cooler's cooling overhead increases if the cooler's efficiency (e.g., the cooling speed) decreases as shown in Fig. 4. Therefore, to conservatively estimate the cooling cost of a modern 10MW system, we use the value of a less-efficient 100kW cryo-cooler ($C.O._{77\text{K}}$ = 9.65 from Fig. 4) [17].

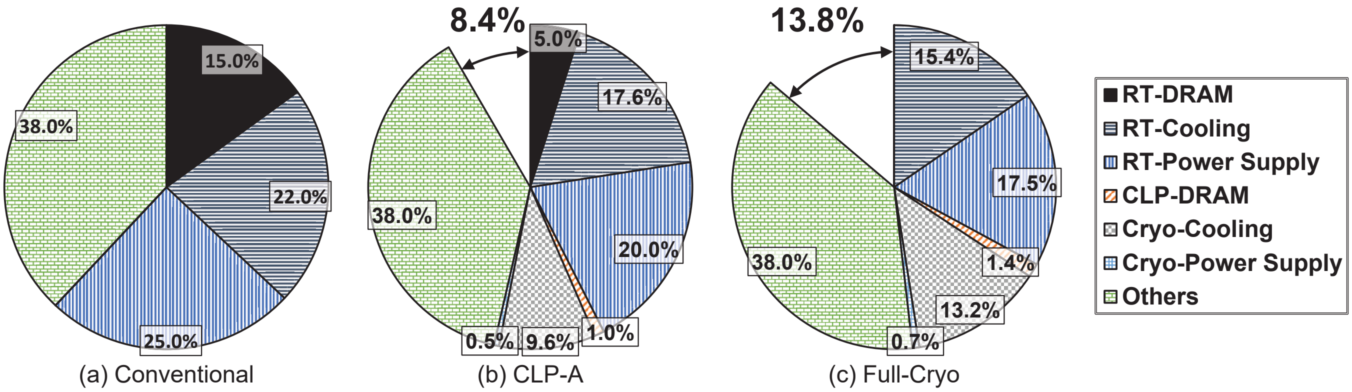
7.3.3 Cryogenic datacenter power modeling. With all costs applied together, we construct our datacenter power model as Eq. (5)

$$\begin{aligned}
&\text{Cryogenic datacenter power}\\
&= (\text{RT-IT} + \text{Cryo-IT}) + (\text{RT-C/P} + \text{Cryo-C/P}) + \text{Misc.}\\
&= (\text{RT-IT} + \text{RT-C/P} + \text{Misc.}) + (\text{Cryo-IT} + \text{Cryo-C/P}) \qquad (5a)\\
&= (1.94 \cdot \text{RT-IT} + \text{Misc.}) + (1 + C.O._{77\text{K}} + P.O._{77\text{K}}) \cdot \text{Cryo-IT}\\
&= 1.94 \cdot \text{RT-IT} + (1 + 9.65 + \frac{22}{50}) \cdot \text{Cryo-IT} + \text{Misc.} \qquad (5b)\\
&= 1.94 \cdot \text{RT-IT} + 11.09 \cdot \text{Cryo-IT} + \text{Misc.} \qquad (5c)
\end{aligned}$$

In the cryogenic datacenter power model, both IT Equipment and Cooling & Power supply are divided into the room temperature parts (RT-IT, RT-C/P) and the cryogenic parts (Cryo-IT, Cryo-C/P). We first replace the "RT-IT + RT-C/P + Misc." with "1.94 · RT-IT + *Misc.*" using Eq. (4), and apply our Cooling & Power Supply model (Eq. (3)) to Eq. (5a). As mentioned in Section 7.3.2, we set $C.O._{77\text{K}}$ to 9.65. The power overhead at 77K ($P.O._{77\text{K}}$) is the same as $P.O._{300\text{K}}$ (= $\frac{22}{50}$) because cryogenic IT components also utilize the existing power supply path (Eq. (5b)). As a result, the total power consumption of the cryogenic datacenter is modeled as Eq. (5c).

7.4 CLP-A: Total power cost evaluation

Based on the power model in Section 7.3, we evaluate the power consumption of our CLP-A's datacenter. Fig. 20 compares the power



(a) Conventional (b) CLP-A (c) Full-Cryo

Figure 20: Total power consumption of a datacenter based on its choice of memory devices: (a) 100% RT-DRAMs (Conventional), (b) 93% RT-DRAMs+7% CLP-DRAMs (CLP-A), and (c) 100% CLP-DRAMs (Full-Cryo)

consumption of a datacenter based on its choice of memory devices: using only RT-DRAMs (Conventional), using 93% RT-DRAMs + 7% CLP-DRAMs (CLP-A), and using only CLP-DRAMs (Full-Cryo). All values are normalized to the power of conventional datacenters. Note that "Others" indicates the power consumed by sources other than DRAM power and Cooling & Power Supply.

For CLP-A, the total power cost is reduced by 8.4% (Fig. 20b) as it significantly reduces the RT-DRAM power (from 15.00% to 5.0%) by migrating hot pages to low-power CLP-DRAMs. The reduced RT-DRAM power also incurs the reduced cost for RT-Cooling and RT-Power Supply (from 47.0% to 37.6%). In addition, the lower power consumption of CLP-DRAM offsets the high cooling overhead as Cryo-Cooling is proportional to CLP-DRAM's power consumption (Eq. (3a)). In this case, the amount of cryogenic cooling cost (Cryo-Cooling; 9.6%) does not exceed the amount of the power reduction.

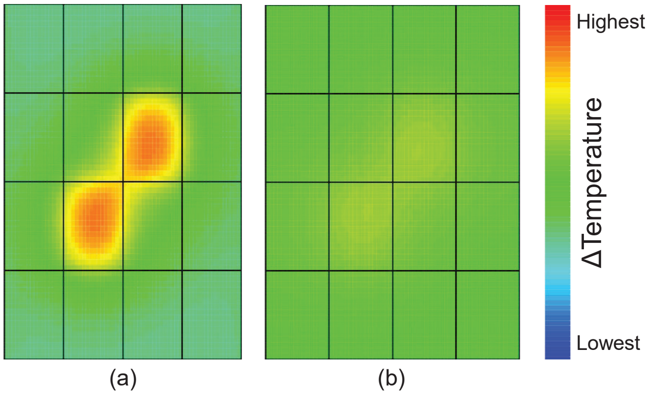
Full-Cryo can achieve the ideal power reduction by replacing all DRAMs in the datacenter with CLP-DRAMs, but it can increase the cooling cost and the memory replacement cost. Thus, we present Full-Cryo results to show the cost-effectiveness of CLP-A using only a small amount of CLP-DRAMs whose power reduction (8.4%) is comparable to Full-Cryo (13.82%).

8 DISCUSSION

8.1 Thermal diffusion

The cryogenic computing is also effective in addressing a heat dissipation problem. In a cryogenic environment, the heat transfer speed of the materials such as silicon and copper increases significantly. For example, the 77K silicon has 39.35 times higher heat transfer speed than the 300K counterpart due to the 9.74 times higher thermal conductivity (shown in Fig. 8a) and the 4.04 times lower specific heat (shown in Fig. 8b). This faster heat transfer can lead to the lower and more flat temperature distribution on a die.

To visualize the effects, we perform thermal simulations for both 300K and 77K environments as shown in Fig. 21. In the 300K environment, two grids show significantly hotter temperature than the nearby cells, as shown in Fig. 21a. On the other hand, Fig. 21b shows that those local hotspots disappear in the 77K environment. This result indicates a great potential to resolve the thermal problems in many aspects (e.g., faster heat dissipations for heat-critical 3D memory designs).



(a) (b)

Figure 21: Simulated temperature distribution (a) 300K environment; (b) 77K environment

8.2 Future directions

To validate our cryogenic-optimized DRAM designs proposed in this work, we plan to create sample prototypes of CLP-DRAM and CLL-DRAM devices.

We expect that cryogenic computing can also improve the performance and power consumption of computation units (e.g., CPU, GPU), interface units (e.g., PCI Express), and memory units other than DRAMs (e.g, SRAM, PRAM, Flash). Therefore, we are currently extending the scope of our cryogenic architecture modeling tool to cover a wider spectrum of cryogenic devices.

We are also investigating the potential of computing at 4K, which is another promising cryogenic temperature domain by enabling superconductor-based logic devices (e.g., RSFQ [21], AQFP [28]).

9 RELATED WORK

In this section, we discuss prior works that focused on the DRAM operating at 77K. In 1991, Henkels et al. [14] fabricated a cryogenic-optimized DRAM and showed its promising characteristics. However, as recent fabrication technologies significantly differ from the technology two decades ago, the findings of the work should be re-evaluated for modern technologies, in particular with the cost effectiveness including the cooling costs.

Recent efforts from both industry and academy target to show the potentials of cryogenic memories. Tannu et al. [29] confirmed that the commodity DRAM chips can work reliably at 80K. Rambus showed that the 77K environment is promising to reduce the DRAM refresh overhead [30], and also suggested that 77K-DRAM is one of the most feasible memory technologies to support superconducting digital processors [31].

Unfortunately, all these studies only focus on commercial DRAMs and specific optimization aspects, and thus they do not fully explore to find the potentials of cryogenic memories and propose new cryogenic-optimized designs. But, they often overlook the high overhead of cooling the devices and how to apply the cooled devices to the existing systems. In addition, they do not cover the system-level performance and power-efficiency improvements in real-world scenarios.

To the best of our knowledge, our work is the first study to develop and validate a modeling tool for cryogenic memory devices, to carefully consider the impact of cooling costs, to design and show the potentials of cryogenic-optimal DRAM devices, and apply them to various system platforms (e.g., server, datacenter).

10 CONCLUSION

The cryogenic computing is a highly promising solution to overcome both power and memory wall challenges. However, cryogenic computers have not been yet realized in the market because computer architects do not fully understand the behaviors of computers running at ultra low temperatures and they do not have proper modeling tools to conduct cryogenic computing researches. This work resolves the challenges as follows. First, we developed and validated CryoRAM, a cryogenic computer architecture modeling tool focusing on memory. Next, we used the tool to derive cryogenic memory designs optimized for either performance or power efficiency. Our case studies with the cooling-cost analysis clearly show that cryogenic computers can significantly improve the performance and power efficiency of a modern datacenter.

ACKNOWLEDGMENTS

This work was partly supported by Samsung Electronics Co., Ltd., Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (MSIT) (NRF-2015M3C4A7065647, NRF-2017R1A2B3011038), Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. R0190-15-2012, No. 1711073912), and Creative Pioneering Researchers Program through Seoul National University. We also appreciate the support from Automation and Systems Research Institute (ASRI) and Inter-university Semiconductor Research Center (ISRC) at Seoul National University.

REFERENCES

[1] J. W. Arblaster. 2015. Thermodynamic Properties of Copper. *Journal of Phase Equilibria and Diffusion* 36, 5 (01 Oct 2015), 422–444. https://doi.org/10.1007/s11669-015-0399-x

[2] F Balestra, L Audaire, and C Lucas. 1987. Influence of substrate freeze-out on the characteristics of MOS transistors at very low temperatures. *Solid-state electronics* 30, 3 (1987), 321–327.

[3] N Balshaw. 1996. Practical cryogenics. And introduction to laboratory cryogenics. (1996).

[4] Randall F Barron. 1999. *Cryogenic heat transfer.* CRC press.

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

[6] Chris Brady. [n.d.]. Memtest86.

[7] Kirk M Bresniker, Sharad Singhal, and R Stanley Williams. 2015. Adapting to thrive in a new economy of memory abundance. *Computer* 48, 12 (2015), 44–53.

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

[9] K. Chen, S. Li, N. Muralimanohar, J. H. Ahn, J. B. Brockman, and N. P. Jouppi. 2012. CACTI-3DD: Architecture-level modeling for 3D die-stacked DRAM main memory. In *2012 Design, Automation Test in Europe Conference Exhibition (DATE).* 33–38. https://doi.org/10.1109/DATE.2012.6176428

[10] Miyuru Dayarathna, Yonggang Wen, and Rui Fan. 2015. Data center energy consumption modeling: A survey. *IEEE Communications Surveys & Tutorials* 18, 1 (2015), 732–794.

[11] Robert H Dennard, Fritz H Gaensslen, V Leo Rideout, Ernest Bassous, and Andre R LeBlanc. 1974. Design of ion-implanted MOSFET's with very small physical dimensions. *IEEE Journal of Solid-State Circuits* 9, 5 (1974), 256–268.

[12] Michael Ferdman, Almutaz Adileh, Onur Kocberber, Stavros Volos, Mohammad Alisafaee, Djordje Jevdjic, Cansu Kaynak, Adrian Daniel Popescu, Anastasia Ailamaki, and Babak Falsafi. 2012. Clearing the clouds: a study of emerging scale-out workloads on modern hardware. In *ACM SIGPLAN Notices*, Vol. 47. ACM, 37–48.

[13] P. Flubacher, A. J. Leadbetter, and J. A. Morrison. 1959. 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* 4, 39 (1959), 273–294. https://doi.org/10.1080/14786435908233340 arXiv:https://doi.org/10.1080/14786435908233340

[14] WH Henkels, D-S Wen, RL Mohler, RL Franch, TJ Bucelot, CW Long, JA Bracchitta, WJ Cote, GB Bronner, Y Taur, et al. 1991. A 4-Mb low-temperature DRAM. *IEEE journal of solid-state circuits* 26, 11 (1991), 1519–1529.

[15] John L Henning. 2006. SPEC CPU2006 benchmark descriptions. *ACM SIGARCH Computer Architecture News* 34, 4 (2006), 1–17.

[16] C. Y. Ho, R. W. Powell, and P. E. Liley. 1972. Thermal Conductivity of the Elements. *Journal of Physical and Chemical Reference Data* 1, 2 (1972), 279–421. https://doi.org/10.1063/1.3253100 arXiv:https://doi.org/10.1063/1.3253100

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

[18] Tao Jin, Jian-ping Hong, Hao Zheng, Ke Tang, and Zhi-hua Gan. 2009. Measurement of boiling heat transfer coefficient in liquid nitrogen bath by inverse heat conduction method. *Journal of Zhejiang University-SCIENCE A* 10, 5 (2009), 691–696.

[19] Nam Sung Kim, Todd Austin, David Baauw, Trevor Mudge, Krisztián Flautner, Jie S Hu, Mary Jane Irwin, Mahmut Kandemir, and Vijaykrishnan Narayanan. 2003. Leakage current: Moore's law meets static power. *computer* 36, 12 (2003), 68–75.

[20] Rakesh Kumar, Dean M Tullsen, Norman P Jouppi, and Parthasarathy Ranganathan. 2005. Heterogeneous chip multiprocessors. *Computer* 38, 11 (2005), 32–38.

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

[22] BA Nayfeh and K Olukotun. 1997. A single-chip multiprocessor. *Computer* 30, 9 (1997), 79–85.

[23] Edward J Nowak. 2002. Maintaining the benefits of CMOS scaling when scaling bogs down. *IBM Journal of Research and Development* 46, 2.3 (2002), 169–180.

[24] Luiz E. Ramos, Eugene Gorbatov, and Ricardo Bianchini. 2011. Page Placement in Hybrid Memory Systems. In *Proceedings of the International Conference on Supercomputing (ICS '11).* ACM, New York, NY, USA, 85–95. https://doi.org/10.1145/1995896.1995911

[25] Robert R Schaller. 1997. Moore's law: past, present and future. *IEEE spectrum* 34, 6 (1997), 52–59.

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

[27] M. Shin, M. Shi, M. Mouis, A. Cros, E. Josse, G. T. Kim, and G. Ghibaudo. 2014. Low temperature characterization of 14nm FDSOI CMOS devices. In *2014 11th International Workshop on Low Temperature Electronics (WOLTE)*. 29–32. https://doi.org/10.1109/WOLTE.2014.6881018

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

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

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

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

[32] Linda Wilson. 2013. International technology roadmap for semiconductors (ITRS). *Semiconductor Industry Association* (2013).

[33] Steven JE Wilton and Norman P Jouppi. 1996. CACTI: An enhanced cache access and cycle time model. *IEEE Journal of Solid-State Circuits* 31, 5 (1996), 677–688.

[34] Wm A Wulf and Sally A McKee. 1995. Hitting the memory wall: implications of the obvious. *ACM SIGARCH computer architecture news* 23, 1 (1995), 20–24.

[35] X Xi, Mohan Dunga, Jin He, Weidong Liu, Kanyu M Cao, Xiaodong Jin, Jeff J Ou, Mansun Chan, Ali M Niknejad, Chenming Hu, et al. 2003. Bsim4. 3.0 mosfet model. *Dept. Elect. Eng. Comput. Sci., Univ. California, Berkeley, CA, Tech. Rep* 94720 (2003), 30.

[36] Runjie Zhang, Mircea R Stan, and Kevin Skadron. 2015. HotSpot 6.0: Validation, acceleration and extension. *University of Virginia, Tech. Rep* (2015).

[37] Hongliang Zhao and Xinghui Liu. 2014. Modeling of a standard 0.35um CMOS technology operating from 77K to 300K. *Cryogenics* 59 (2014), 49 – 59. https://doi.org/10.1016/j.cryogenics.2013.10.003

[38] Wei Zhao and Yu Cao. 2006. New generation of predictive technology model for sub-45nm design exploration. In *7th International Symposium on Quality Electronic Design (ISQED'06)*. 6 pp.–590. https://doi.org/10.1109/ISQED.2006.91