CryoCore: A Fast and Dense Processor Architecture for Cryogenic Computing

Ilkwon Byun*, Dongmoon Min*, Gyu-hyeon Lee, Seongmin Na, and Jangwoo Kim†
Department of Electrical and Computer Engineering
Seoul National University
{ik.byun, dongmoon.min, guhylee, seongmin.na, jangwoo}@snu.ac.kr

***Abstract*—Cryogenic computing can achieve high performance and power efficiency by dramatically reducing the device's leakage power and wire resistance at low temperatures. Recent advances towards cryogenic computing focus on developing cryogenic-optimal cache and memory devices to overcome memory capacity, latency, and power walls. However, little research has been conducted to develop a cryogenic-optimal core architecture despite its high potentials in performance, power, and area efficiency. Once a cryogenic-optimal core becomes available, it will also take full advantage of the cryogenic-optimal cache and memory devices, which leads to a cryogenic-optimal computer.**

In this paper, we first develop *CryoCore-Model* (*CC-Model*), a cryogenic processor modeling framework which can accurately estimate the maximum clock frequency of processor models running at 77K. Next, driven by the modeling tool, we design *CryoCore*, a 77K-optimal core microarchitecture to maximize the core's performance and area efficiency while minimizing the cooling cost. The key idea of CryoCore is to architect a core in a way to reduce the size and number of cooling-unfriendly microarchitecture units and maximize the potential of a voltage and frequency scaling at 77K. Finally, we propose two half-sized, but differently voltage-scaled CryoCore designs aiming for either the maximum performance or power efficiency. With both conventional and our design integrated with cryogenic memories, our high-performance CryoCore design achieves 41% higher single-thread performance for the same power budget and 2x higher multi-thread performance for the same die area. Our low-power CryoCore design reduces the power cost by 38% without sacrificing the single-thread performance.

***Index Terms*—Cryogenic computing, Cryogenic processor, Modeling, Simulation**

I. Introduction

High-performance computing and datacenter industries are requiring the fastest and most power-efficient processors more than ever. However, facing the end of both single-thread and multi-thread performance scalings, server architects are now facing critical challenges to further improve both performance and power efficiency of the current high-end server processors.

As the technology scaling continues, it is getting more difficult to build a faster computer 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 clock frequency. If architects force to increase the clock frequency, it becomes another critical issue to compensate for the correspondingly increasing dynamic power consumption. 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 (i.e., CMP [1], [2], SMT [3]). 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.

Cryogenic computing, which aims to run a computer device at extremely low temperatures, has emerged as a highly promising solution to improve its performance and power efficiency thanks to the significantly reduced leakage current and wire resistance. With the voltage scaling enabled by the almost eliminated leakage current, architects can achieve both higher clock frequency and much lower power consumption. However, at very low temperatures, the cost to cool the device becomes a major challenge because it can easily dominate the obtained performance and power advantages.

Recent proposals have thus focused on developing cryogenic-optimal cache and memory devices [4], [5] whose cost effectiveness at low temperatures can be easily reasoned thanks to their cell-driven array-like structures. With this cryogenic-optimal memory hierarchy available, architects are now in a dire need of a cryogenic-optimal processor core design to achieve the full potential of cryogenic computing, while tolerating the increased cooling cost. However, little research has been conducted to develop a cryogenic-optimal core architecture mainly due to the absence of a performance and cost analysis methodology for processor models running at low temperatures.

In this paper, we propose *CryoCore*, a fast, dense, and cooling-cost efficient cryogenic-optimal processor architecture running at 77K as follows. First, to enable processor design space explorations at 77K, we develop *CryoCore-Model* (*CC-Model*), a validated cryogenic processor modeling framework which can accurately estimate the maximum clock frequency of processor models running at 77K. To obtain the maximum clock frequency, *CC-Model* estimates the critical-path delays for the processor's each pipeline stage and combines them.

CC-Model consists of three submodules which model MOSFET, wire, and processor's pipeline at 77K, respectively. The MOSFET model takes the fabrication-process information and operating voltage as inputs and generates the major MOSFET characteristics at 77K. The wire model takes the metal layer information as inputs and generates the on-chip wire resistivity at 77K. The processor model applies the MOSFET and wire

[*] Both authors contributed equally to this research.
[†] Corresponding author.

978-1-7281-4661-4/20/$31.00 ©2020 IEEE
DOI 10.1109/ISCA45697.2020.00037

models' outputs to the target pipeline design and reports its critical-path delay information. We fully validate CC-Model by validating each submodule using the industry-provided information, previous literatures, and our own experiments.

Second, by applying our modeling framework to two reference core models (i.e., high-performance core vs. low-power core), we identify a critical need of architecting a cryogenic-optimal processor design and its key design principles. We first observe that it is important to minimize the core's dynamic power which significantly increases its cooling cost at 77K. We also observe that it is important to maintain the core's high voltage and frequency design to enable a further voltage and frequency scaling at 77K. In summary, a cryogenic-optimal core should reduce the size and number of dynamic power-consuming units, while targeting a high frequency.

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, compared to high-performance core, CryoCore reduces its dynamic power by 77% and its core area by 48%, while maintaining its high clock frequency. We also assume that the number of cores per chip can be doubled thanks to the half-sized core choice.

Finally, by applying different voltage scalings, we propose two CryoCore designs to further increase either its clock frequency or power efficiency. The high-performance cryogenic core design (*CHP-core*) increases the clock frequency by 51%, whereas the low-power cryogenic core design (*CLP-core*) reduces the power consumption to 2.93%.

For performance evaluation, we compare CHP-core with the high-performance reference core. With both cores integrated with a conventional cache and DRAM, CHP-core improves the single and multi-thread performance of PARSEC workloads by 21.8% (for the same power budget) and 83.2% (for the same die area) on average, respectively. With both cores integrated with a cryogenic-optimal cache [4] and memory [5], CHP-core improves the single and multi-thread performance by 41% and 100%, respectively. For power evaluation, CLP-core reduces the overall power consumption by 38% even with the cooling cost considered, while maintaining the single-thread performance.

In summary, our work makes the following contributions:

- **Cryogenic core performance modeling:** To the best of our knowledge, this is the first work to model, validate, and optimize the core architecture for 77K.
- **Principles for cryogenic-optimal core design:** We identify key design principles to architect a cryogenic-optimal core: (1) minimize the dynamic power and (2) target the maximum frequency.
- **Cryogenic-optimal core designs:** We architect a fast and dense cryogenic-optimal core by making the core half the size, while aiming for the highest frequency. We also present CHP-core and CLP-core to further improve its performance and power efficiency with different voltage scalings.
- **Significant performance improvements:** CHP-core increases the clock frequency by 51%, while roughly doubling the number of cores on the same die. When used with cryogenic memories, CHP-core achieves 41% higher single-thread performance and 100% higher multi-thread performance for 12 PARSEC workloads.
- **Significant cost and power reduction:** CLP-core's dynamic power reduction leads to the 38% of overall power reduction even including the cooling cost, without sacrificing the performance.

II. Background

A. Limitations of CPU performance scaling

Computer architects are now facing critical challenges in improving the performance of CPUs in terms of both single-thread and multi-thread performance.

1) Single-thread performance: Architects have improved a processor's single-thread performance by increasing its clock frequency and improving its microarchitecture. In particular, a clock frequency scaling becomes promising when architectural improvements become difficult. However, as the technology scaling continues, the clock frequency scaling has become extremely difficult due to the significantly increasing wire latency and the power wall problem.

Wire latency problem: Unlike the transistor speed, the latency of wires cannot be easily scaled with the shrinking technology node due to the steeply increasing wire resistivity [6]. As a result, the clock frequency scaling becomes more difficult even when faster and smaller transistors can be deployed.

Power wall problem: Aside from the wire latency problem, the processor's dynamic power is another critical challenge against the clock frequency scaling because the voltage-driven dynamic power quickly increases with the increased clock frequency. In the past, architects could compensate for the increasing dynamic power by satisfying Dennard scaling [7] (i.e., reducing both V_{dd} and V_{th}). However, as the static power exponentially increases with the reduced V_{th}, Dennard scaling (and thus power scaling) stopped.

2) Multi-thread performance: To circumvent the challenges, architects have adopted a multi-threading strategy to improve a processor's throughput instead of its single-thread performance since the early 2000s. There have been two major directions to improve the multi-thread performance: (1) increasing the number of cores on chip or Chip Multiprocessing (CMP) and (2) increasing the number of threads per core or Simultaneous Multithreading (SMT). By applying these two schemes, architects have successfully improved system-level performance until recently. However, even these schemes now suffer from the end of Moore's law [8], [9] together with the wire-latency problem and the power wall problem due to the increasing number of transistors on chip and the increasing critical-path delay in a core, respectively.

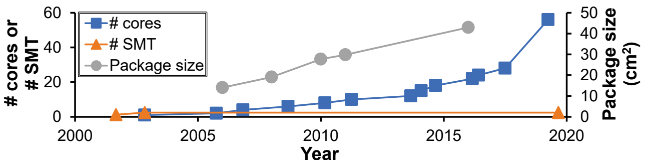


Fig. 1: CMP level, package size, and SMT level of Intel Xeon

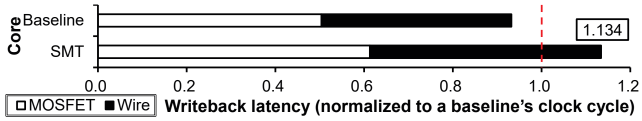


Fig. 2: Increased critical-path delay in SMT processor (derived from our modeling methodology in Section III)

End of CMP scaling: To achieve the CMP scaling, Moore's law should be satisfied because adding more cores on chip requires more transistors. However, integrating more transistors into the same die area is now extremely challenging due to the end of Moore's Law and the increasing on-chip power consumption. Therefore, the number of on-chip cores (i.e., CMP level) cannot be easily increased. For example, Fig. 1 shows the CMP level and the package size of Intel Xeon processors over generations [10]. The figure clearly indicates that architects cannot put more cores on chip without prohibitively increasing the package size or reducing the size of each core.

End of SMT scaling: Increasing the number of threads running on a single core (i.e., SMT level) requires much larger intra-core, memory-like architectural units (e.g., register files, load and store queues, reorder buffer) to keep many architectural states while mitigating intra-unit contentions. However, increasing the size of memory-like modules can incur huge performance degradation due to the increased critical-path delay. Fig. 2 shows our model-driven latency breakdown of a writeback operation in a baseline core and its SMT version. We observe that the SMT core's double-sized register file increases the writeback latency by 13%, which degrades its single-thread performance. We believe that this is one of major reasons together with the intra-unit contention which limits the degree of SMT (Fig. 1).

Therefore, architects are now more than ever in dire need of a novel solution to effectively improve both the single-thread and the multi-thread performance.

B. Potentials of cryogenic computing

To resolve the performance challenges, cryogenic computing has emerged as a highly promising solution as it can dramatically reduce a computer device's leakage current and wire latency. First, as the leakage current shrinks exponentially with the temperature [11], [12], we can significantly reduce both V_{dd} and V_{th}. Second, as the wire resistivity linearly decreases with the temperature, the wire latency is reduced correspondingly. For example, the copper's resistivity decreases by six times when reducing the temperature from 300K to 77K [13]. As a result, at low temperatures, we can safely increase the clock frequency thanks to the faster transistor switching and data transfer with much lower voltages applied.

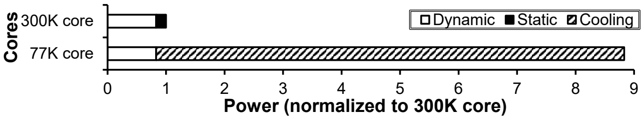


Fig. 3: Conventional core's power consumption with cooling cost included, derived by McPAT [28]

Modern cryogenic computing often aims for two target low temperatures, 77K and 4K, because the two temperatures can be easily achieved by applying Liquid Nitrogen (LN) and Liquid Helium (LHe), respectively. For conventional computing, 77K has been more actively considered than 4K due to the prohibitively higher cooling cost for 4K (e.g., 300-1000 times of device power consumption [15]). Therefore, the 4K computing has been considered mainly for unconventional superconducting and quantum computing (e.g., RSFQ [16]–[18], AQFP [19]–[21], quantum controlling unit [14]).

On the other hand, as the conventional CMOS technology reliably operates at 77K incurring much lower cooling cost [15], [22], computer architects have focused more on CMOS-feasible 77K-based cryogenic computing. Previous works explored the potentials of cryogenic computing, especially for the memory devices operating at 77K [4], [5], [23]–[26]. Among them, recent studies [4], [5] developed a modeling tool to estimate the performance and power consumption of 77K memory devices (i.e., DRAMs and on-chip caches) and used the tool to propose a 77K-optimal memory architecture.

Therefore, it is a straightforward next step to develop a cryogenic-optimal processor core running at 77K which benefits from the reduced power consumption and wire latency. In addition, the cryogenic-optimal core will also get synergistic benefits by using the previously proposed cryogenic caches and memories.

C. Challenges for designing a cryogenic 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) [27], [28]. 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 VI-A2). Fig. 3 shows that lowering a processor's

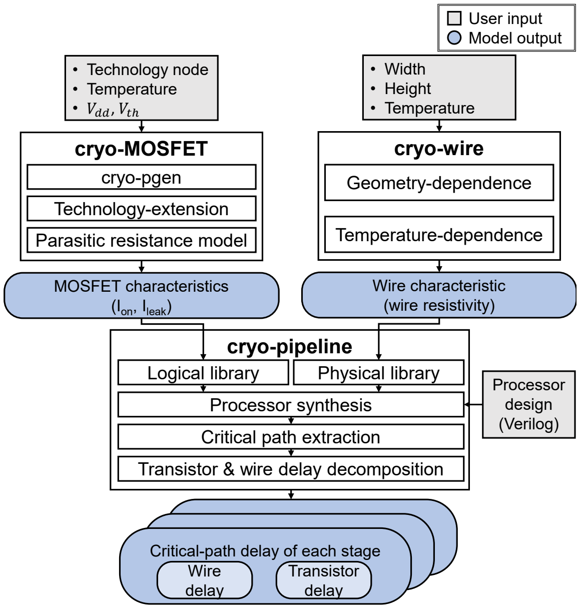


Fig. 4: Cryogenic processor model (CC-Model) overview

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.

In this paper, we resolve the three challenges as follows. We first develop a novel cryogenic processor's performance modeling framework. 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 to provide the highest single-thread and multi-thread performance while satisfying the target die area and cooling cost budget.

III. 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. 4. First, ***MOSFET model (cryo-MOSFET)*** takes fabrication-process information (i.e., model card) as inputs, and then derives the 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.

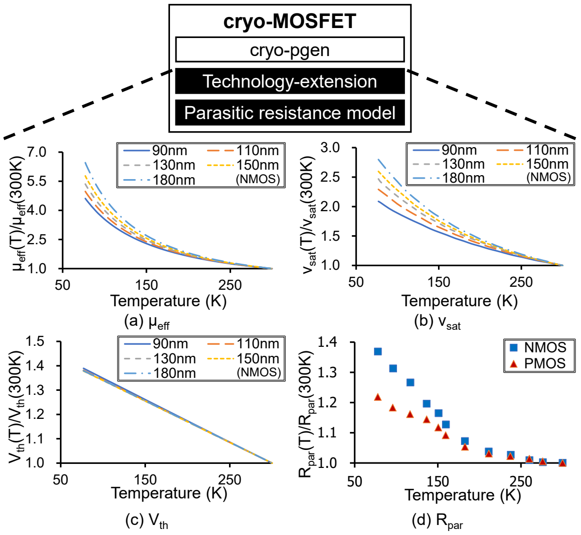


Fig. 5: Extension to the baseline MOSFET model: (a) Carrier mobility; (b) Saturation velocity; (c) Threshold voltage; (d) Parasitic resistance model

A. MOSFET model

To model the major MOSFET characteristics at low temperatures, we utilize cryo-pgen [5] 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 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

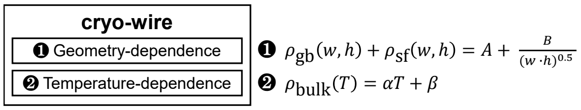

cryo-wire
❶ Geometry-dependence
❷ Temperature-dependence
❶ $\rho_{gb}(w,h) + \rho_{sf}(w,h) = A + \frac{B}{(w \cdot h)^{0.5}}$
❷ $\rho_{bulk}(T) = \alpha T + \beta$

Fig. 6: Wire model

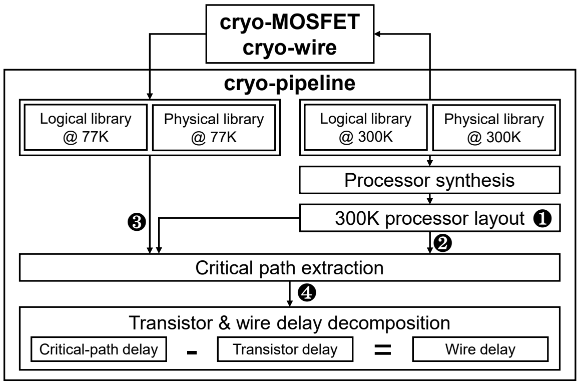


Fig. 7: Processor model

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 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 [29]. 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. 5a-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 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. 5d (i.e., *parasitic resistance model*). We utilize the temperature dependency data of R_{par} from the previous work [29]. With these additional models, cryo-MOSFET can now accurately predict the low-temperature MOSFET characteristics of modern technology nodes.

B. Wire model

$$\rho_{\text{wire}}(T, w, h) = \rho_{\text{bulk}}(T) + \rho_{\text{gb}}(w, h) + \rho_{\text{sf}}(w, h) \qquad (1)$$

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. 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}) [30]–[32]. Eq. (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) [31], [33]–[35]. 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 [31], [33], [36] (Fig. 6❶). We set the purity-related hyperparameters (i.e., A and B) based on the previous studies [33], [37]. Next, we implement temperature-dependent mechanisms (ρ_{bulk}) as the linear model in Fig. 6❷ with the coefficients of coppers [13].

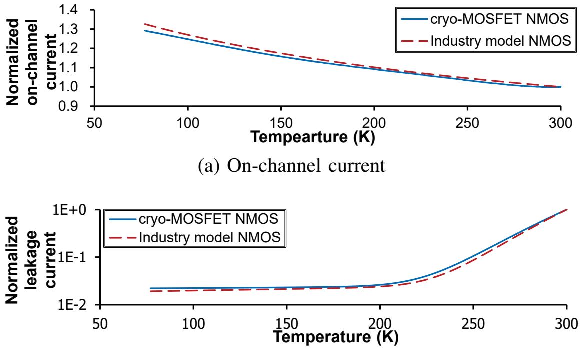
C. Processor model

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 speed-up 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 [38]. 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. 7 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 speed-up at 77K.

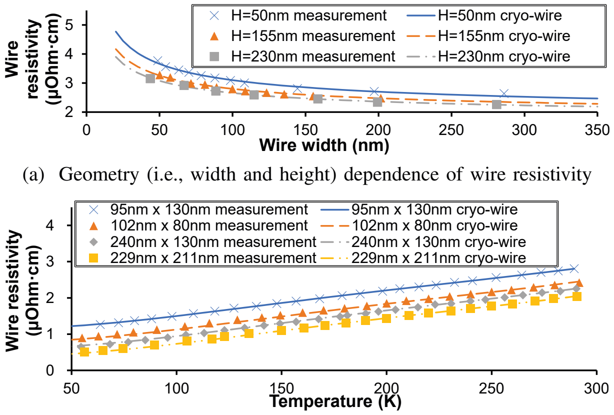
MOSFET/Wire delay decomposition: Cryo-pipeline can fully decompose the critical-path delay into its transistor and wire delay portions by subtracting the impact of the transistor portions from the overall critical-path delay result (❹).



(a) On-channel current

(b) Leakage current

Fig. 8: cryo-MOSFET validation results: industry-validated model vs cryo-MOSFET outcomes



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

(b) Temperature dependence of wire resistivity

Fig. 9: cryo-wire validation results: measurement data from the previous literature vs. cryo-wire outcomes

IV. Model validation

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

A. 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 [39] as its input model card.

Fig. 8 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. 8a). Our MOSFET model not only accurately predicts the trend of increasing I_{on} but also 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. 5a, 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. 8b. 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.

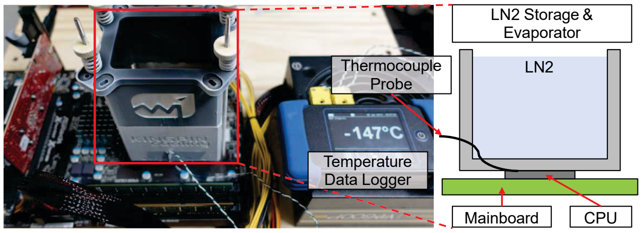


Fig. 10: Experimental setup for the processor model validation

B. Wire model validation

We validate our wire model by comparing the wire resistivity reported by cryo-wire with the measured data from the literature [37], [40], [41]. Fig. 9 shows the validation results for cryo-wire. First, cryo-wire well matches the published resistivity data for various sets of width and height (Fig. 9a) [37]. Second, Fig. 9b shows that our wire model well predicts the linearly decreasing wire resistivity compared to the data from previous literature [40], [41]. 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.

C. 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. 10 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

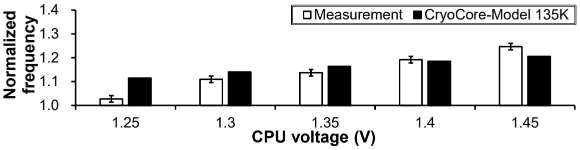


Fig. 11: cryo-pipeline validation results: real measurements vs. cryo-pipeline outcomes

TABLE I: 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

intentionally construct the computer board with the processor fabricated 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. 11 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 [42] with BOOM processor design [43] as the model inputs. Fig. 11 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).

V. Optimizing cryogenic processors

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 V-A). Next, following the directions, we design our 77K-optimal microarchitecture, called *CryoCore* (Section V-B). 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 V-C).

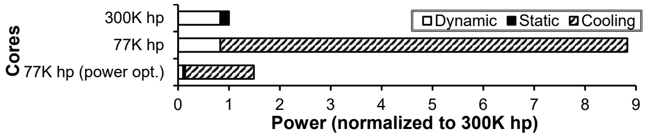


Fig. 12: Power consumption of hp-cores at 300K and 77K

In the following subsections, we conduct performance, power, and area analyses for the processors listed in Table I. We implement the target processors by customizing RISC-V BOOM [43], one of the most representative out-of-order core designs. For performance analysis, we utilize CC-Model with FreePDK 45nm library [42] which can be scaled to 77K by our MOSFET/wire models. For power and die-area analysis, we use McPAT [28] 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 VI-A2 for more details of our power calculation methodology including the cooling cost model.

A. 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 [44] (hp-core in Table I). We set hp-core's frequency at 300K based on the literature [44], and its power and area are calculated from McPAT.

Fig. 12 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. 12 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

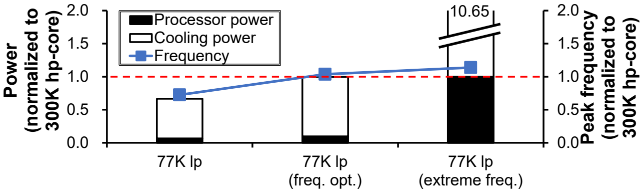


Fig. 13: Maximum frequency and total power consumption of lp-cores operating at 77K

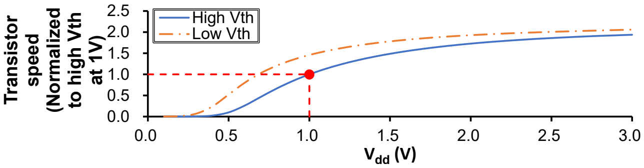


Fig. 14: Saturated transistor speed with the increasing V_{dd}

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 [45], 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 I). The lp-core's frequency is based on the literature [45] and its power consumption and area are derived from McPAT.

Fig. 13 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-core's V_{dd} (and thus its frequency) up to two specific points which form 77K lp (freq. opt) and 77K lp (extreme freq.) as shown in Fig. 13. 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. 14 shows the speed of MOSFET when varying its V_{dd} and V_{th}. We approximate the speed of MOSFET as its transconductance (i.e., $I_{\text{on}}/V_{\text{dd}}$), 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 [46]. 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.

B. CryoCore: Cryogenic-optimal 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 I 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. 14). 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 [27]. 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 I. 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 VI).

C. *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. 15 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. 15 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. 15. 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 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. 15).

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. 15). 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 II). That is, architects can utilize both of their benefits just by applying the dynamic voltage frequency scaling (DVFS) [54] to the single-core design.

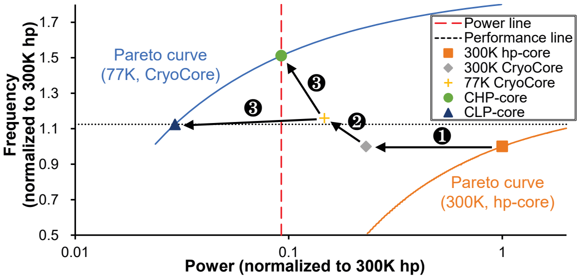


Fig. 15: Deriving cryogenic-optimal processor designs by applying voltage scaling to CryoCore

TABLE II: 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 I</td></tr>
<tr><td>CHP-core</td><td>6.1GHz</td><td colspan="2">0.75V / 0.25V</td><td>CryoCore in Table I</td></tr>
<tr><td>CLP-core</td><td>4.5GHz</td><td colspan="2">0.43V / 0.25V</td><td>CryoCore in Table I</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>

VI. 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 VI-A). Next, we evaluate the single-thread and multi-thread performance of CHP-core (Section VI-B) and power consumption of CLP-core (Section VI-C).

A. *Evaluation methodology*

1) 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 II.

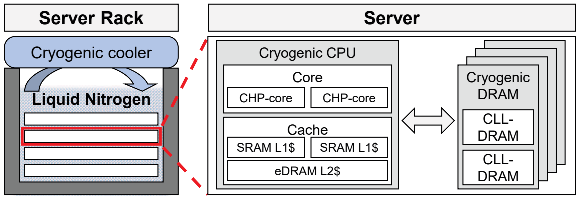


Fig. 16: A full cryogenic computer system which the entire node is cooled down to 77K.

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 [44]. On the other hand, for CHP-core, we set the number of cores to eight based on the area analysis in Table I.

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 VII-A). 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.

For the 77K memory setup, we use CryoCache [4] and CLL-DRAM [5] 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. 16 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.

2) 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 [28] 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 [47]. We obtain the input access trace for McPAT from the gem5 simulations [48] with PARSEC 2.1 workloads [49].

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. 16 shows the overview of our cooling system (i.e., Stinger system [50]), 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) [15], [51]. Therefore, we focus only on the cooling power consumption as the cooling cost.

$$\mathrm{P_{cooling}} = \mathrm{P_{device}} \cdot \mathrm{CO} \tag{2}$$

$$\begin{aligned}\mathrm{P_{77K\text{-}total}} &= \mathrm{P_{77K\text{-}device}} + \mathrm{P_{77K\text{-}cooling}} \\ &= (1 + \mathrm{CO_{77K}})\ \mathrm{P_{77K\text{-}device}} \\ &= 10.65\ \mathrm{P_{77K\text{-}device}} \end{aligned} \tag{3}$$

The cooling power consumption ($\mathrm{P_{cooling}}$) is the electrical power to remove the heat dissipated from the device (Eq. (2)). $\mathrm{P_{device}}$ is the power consumption of the electronic devices and CO is the cooling overhead [15]. 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 [15]. We use 9.65 value for our 77K cooling overhead ($\mathrm{CO_{77K}}$).

Based on Eq. (2), we calculate the total required power for our 77K system ($\mathrm{P_{77K\text{-}total}}$) as Eq. (3). Eq. (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 [15], [52]. The cooling cost model is also conservative considering the continuously increasing power efficiency of cryo-coolers.

B. *Performance evaluation*

1) Single-thread performance: Fig. 17 shows the single-thread performance of the various systems shown in Table II. 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

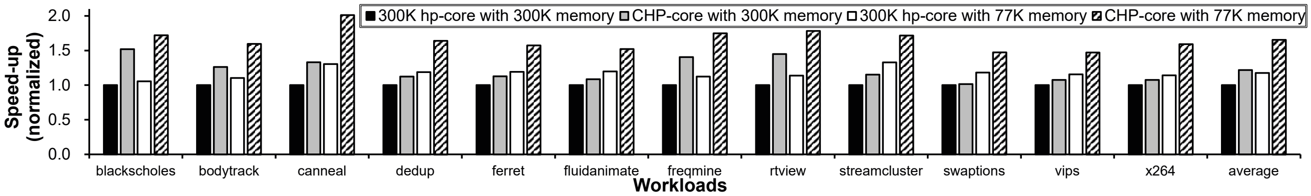


Fig. 17: Single-thread performance of the 300K baseline (300K hp-core with 300K memory), CHP-core with 300K memory, 300K hp-core with 77K memory, and CHP-core with 77K memory.

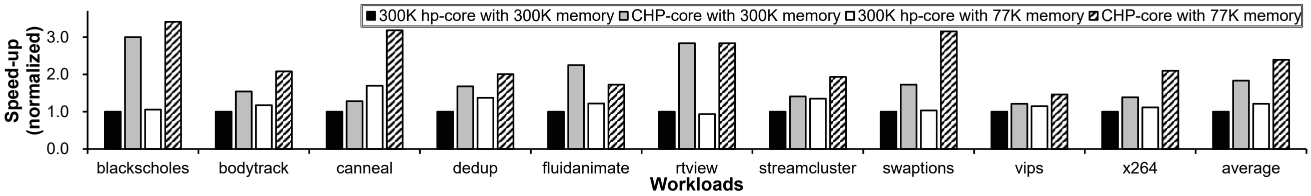


Fig. 18: Multi-thread performance of the 300K baseline (300K hp-core with 300K memory), CHP-core with 300K memory, 300K hp-core with 77K memory, and CHP-core with 77K memory.

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 [49].

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 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.

2) Multi-thread performance: Fig. 18 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.

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.

C. Power evaluation

Fig. 19 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. 3). 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. 12).

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

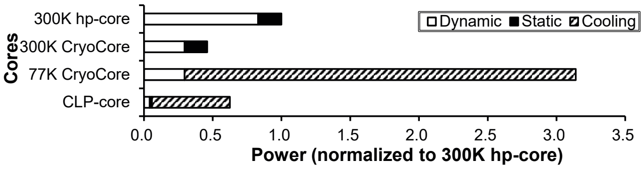


Fig. 19: Total power consumption of the 300K hp-core (baseline), 300K CryoCore, 77K CryoCore, and CLP-core.

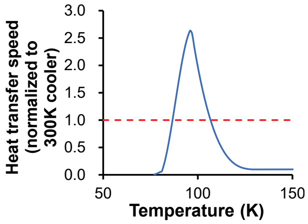


Fig. 20: Heat dissipation speed of LN-bath cooling with temperatures

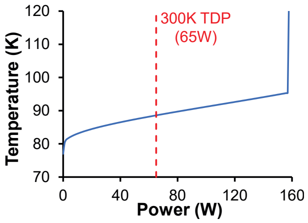


Fig. 21: Temperature variation of the cryogenic processor with power consumption

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. 19 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.

VII. Discussion

A. A thermal budget of the cryogenic processors

The thermal budget analysis is 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. 20 shows the normalized heat-dissipation speed in a low-temperature range. 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 [53]. 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 budget of cryogenic processors. Fig. 21 shows the operating temperature of cryogenic processors with various power consumption (0W-160W). We utilize cryo-temp [5] with HotSpot [53] 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. 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.

VIII. Related work

We discuss the prior works which focused on the processor's critical-path delay modeling and 77K cryogenic computing.

Critical-path delay modeling: Palacharla et al. [27] built the delay model for major pipeline stages to study the impact of the increasing issue width and window on the clock frequency. Li et al. [28] built the power, area, and timing model for given processor configurations. However, there is no previous work that models the critical-path delay at low temperatures.

77K cryogenic computing: Most previous works for 77K computing focused on memory modules. Tannu et al. [26] and Rambus [23] showed the commodity DRAM can reliably work at 77K. Lee et al. [5] built a cryogenic DRAM model and showed the potential of 77K-optimized DRAM. Min et al. [4] analyzed cache technologies at 77K and proposed the cryogenic-optimal cache architecture in terms of performance and energy efficiency. However, the previous works only focused on the on-chip and off-chip memory modules.

To the best of our knowledge, our work is the first study to develop a modeling tool for cryogenic processors, and show the potentials of the full cryogenic computer system where the cryogenic cache [4] and DRAM [5] are also integrated.

IX. Conclusion

Cryogenic computing can significantly improve a computer's performance and power efficiency thanks to the reduced leakage current and wire resistivity at low temperatures. Recent research proposed to build cryogenic-optimal caches and memories. However, little research has been conducted to develop a cryogenic-optimal core due to the lack of performance and cost modeling tool and core design guidelines for low temperatures. To resolve the challenges, 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.

ACKNOWLEDGMENTS

This work was partly supported by Samsung Electronics, SK Hynix, National Research Foundation of Korea (NRF) grant funded by the Korean Government (NRF-2015M3C4A7065647, NRF-2017R1A2B3011038, NRF-2019R1A5A1027055, NRF-2019-Global Ph.D. Fellowship Program), Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (No. 1711080972), 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] B. Nayfeh and K. Olukotun, "A single-chip multiprocessor," *Computer*, vol. 30, no. 9, pp. 79–85, 1997.

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

[3] 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.

[4] 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.

[5] 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

[6] R. Ho, K. W. Mai, and M. A. Horowitz, "The future of wires," *Proceedings of the IEEE*, vol. 89, no. 4, pp. 490–504, 2001.

[7] 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.

[8] R. R. Schaller, "Moore's law: past, present and future," *IEEE spectrum*, vol. 34, no. 6, pp. 52–59, 1997.

[9] L. B. Kish, "End of moore's law: thermal (noise) death of integration in micro and nano electronics," *Physics Letters A*, vol. 305, no. 3-4, pp. 144–149, 2002.

[10] Intel, "Intel xeon processors," 2019. [Online]. Available: https://www.intel.com/content/www/us/en/products/processors/xeon.html

[11] 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.

[12] 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.

[13] 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.

[14] S. R. Ekanayake, T. Lehmann, A. S. Dzurak, R. G. Clark, and A. Brawley, "Characterization of sos-cmos fets at low temperatures for the design of integrated circuits for quantum bit control and readout," *IEEE Transactions on Electron Devices*, vol. 57, no. 2, pp. 539–547, Feb 2010.

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

[16] K. K. Likharev and V. K. Semenov, "Rsfq logic/memory family: A new josephson-junction technology for sub-terahertz-clock-frequency digital systems," *IEEE Transactions on Applied Superconductivity*, vol. 1, no. 1, pp. 3–28, 1991.

[17] D. K. Brock, "Rsfq technology: Circuits and systems," *International journal of high speed electronics and systems*, vol. 11, no. 01, pp. 307–362, 2001.

[18] I. Nagaoka, M. Tanaka, K. Inoue, and A. Fujimaki, "29.3 a 48ghz 5.6 mw gate-level-pipelined multiplier using single-flux quantum logic," in *2019 IEEE International Solid-State Circuits Conference-(ISSCC)*. IEEE, 2019, pp. 460–462.

[19] N. Takeuchi, K. Ehara, K. Inoue, Y. Yamanashi, and N. Yoshikawa, "Margin and energy dissipation of adiabatic quantum-flux-parametron logic at finite temperature," *IEEE Transactions on Applied Superconductivity*, vol. 23, no. 3, pp. 1 700 304–1 700 304, 2013.

[20] N. Takeuchi, D. Ozawa, Y. Yamanashi, and N. Yoshikawa, "An adiabatic quantum flux parametron as an ultra-low-power logic device," *Superconductor Science and Technology*, vol. 26, no. 3, p. 035010, 2013.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

[29] H. Zhao and X. Liu, "Modeling of a standard 0.35um cmos technology operating from 77k to 300k," *Cryogenics*, vol. 59, pp. 49 – 59, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0011227513000969

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

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

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

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

[47] L. Wilson, "International technology roadmap for semiconductors (itrs)," *Semiconductor Industry Association*, vol. 1, 2013.

[48] 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.

[49] 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.

[50] N. Balshaw, "Practical cryogenics. and introduction to laboratory cryogenics," 1996.

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

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

[53] 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.

[54] 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.