# Design-Technology Co-optimization for Cryogenic Tensor Processing Unit

Dong Suk Kang, Shimeng Yu
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
Email: shimeng.yu@ece.gatech.edu

***Abstract*— The cryogenic silicon complementary-metal-oxide-semiconductor (CMOS) technology and its application in tensor processing unit (TPU) design are explored. Using the 22 nm fully-depleted-silicon-on-insulator (FDSOI) transistor model that was calibrated at 70 K, this study provides insights into the design/technological knobs to achieve a superior performance at cryogenic temperature (cryo-TPU) by exploiting threshold voltage ($V_{th}$) engineering, gain-cell embedded DRAM (GC-eDRAM) and true-single phase clock (TSPC) D flip-flop. Benchmark shows that cryo-TPU using GC-eDRAM based global buffer and TSPC D flip-flop based register surpasses conventional TPU architecture operating at the room temperature: over 33% chip area reduction in iso-power condition, over 94% power reduction in iso-area condition and over 40% power reduction even when the refrigerator cooling power is included.**

***Keywords—cryogenic computing, TPU, gain-cell, embedded DRAM, DTCO, hardware accelerator***

## I. INTRODUCTION

The development of TPU, a domain-specific hardware accelerator for the efficient processing of deep neural networks, made it possible to accelerate inference operations tens of times faster than CPU and GPU [1]. Nevertheless, the soaring demand for the use of neural networks and the increase of the depth and size of the neural network further requires better TPU performance in terms of power and latency. As the conventional CMOS technology is approaching and end of the 2D scaling, technological breakthrough is required (e.g., 3D integration).

Cryogenic CMOS is gaining recent attention as a strong candidate for the solution to high performance computing. The liquid nitrogen temperature (77 K) is considered a suitable environment for cryogenic computing because of a tradeoff between the energy saving at lower temperature and a rising cooling power. The cooling power from refrigerator is roughly 10× of the chip power at 77 K [2], which indicates that 10× chip power reduction is required to show the system-level benefits in the energy carbon footprint. This paper proposes cryogenic-optimized TPU, targeting a faster, denser, and more energy-efficient system than room-temperature TPU. Innovations at transistor characteristics, memory cells and circuit topologies were proposed for the cryogenic system to achieve the targets listed above. Section II is about the system and simulation setup. Firstly, digital systolic TPU-like architecture is described. Then, the 70 K properties of the 22nm MOSFET [3] are described. Note that the available experimental data was measured at 70 K in the proximity of 77 K. Lastly, the simulation setup and flow is described. Section III presents design-technology co-optimization (DTCO) methods to improve the performance of cryo-TPU. To be detailed, 1) GC-eDRAM replacing SRAM in

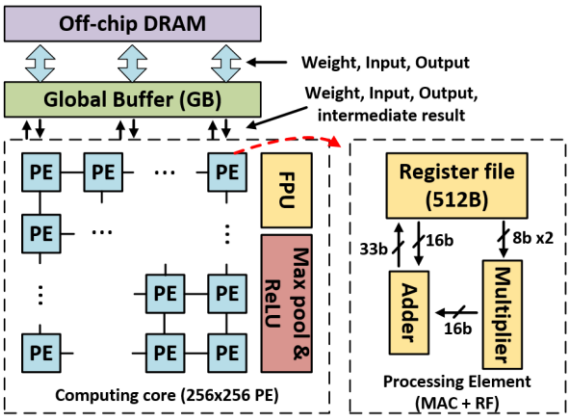



Figure 1. TPU-like architecture for the benchmark

the global buffer and its dedicated $V_{th}$ engineering, 2) another $V_{th}$ engineering for logic transistors, 3) TSPC D flip-flop replacing conventional master-slave D flip-flop used as register, and 4) wire width optimization strategy for cryogenic temperature, are suggested. Section IV shows a benchmark simulation on the performance improvement of the global buffer and finally the entire TPU system, in which the techniques of Section III are applied. Finally, Section V draws the conclusion.

## II. Background and Methods

### A. TPU-like architecture

Fig.1 shows the systolic architecture of the digital multiply-and-accumulate (MAC) array. Processing elements (PEs) receive weights and inputs from the DRAM and stores them in the global buffer (GB). When the inference starts, weights, inputs are fetched from a GB and deliver to the register files so they can be used by the MAC units and perform the systolic operation. Functional modules such as pooling, ReLU, and floating-point unit (FPU) are integrated for the activation and normalization. The LPDDR4 interface (~1 pJ/bit) is assumed.

### B. Cryogenic MOSFET

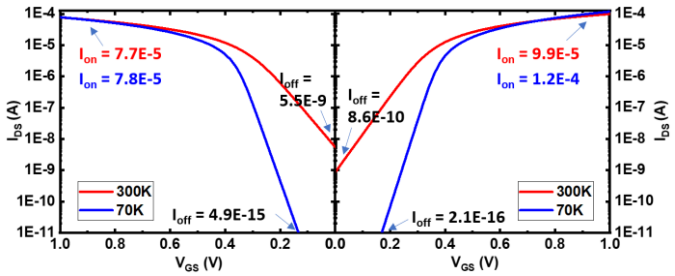



Figure 2. $I_{DS}$-$V_{GS}$ curve of PMOS (left) and NMOS (right) at 70 K and 300 K (W/L = 90 nm/22 nm). Calibrated with experimental data of 22nm FDSOI [3].

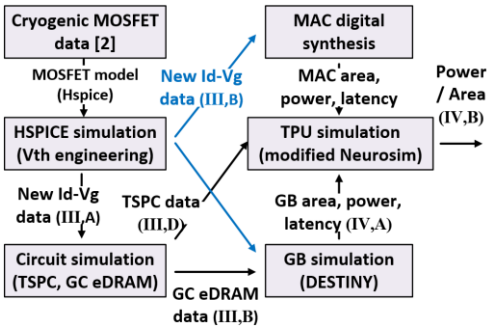



Figure 3. Flowchart of the simulation for TPU-like architecture

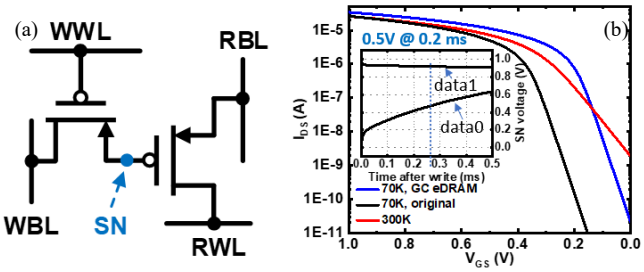



Figure 4. (a) GC-eDRAM schematic (b) $I_{DS}$-$V_{GS}$ curve of the original and $V_{th}$-engineered NMOS for GC-eDRAM (W/L = 30 nm/22 nm, projected); (inset) storage node voltage during hold, showing retention time 0.2 ms at 70 K.

Fig. 2 plots the drain current versus the gate voltage of the MOSFET at 300 K and 70 K. The virtual source based compact model based on the experimental measurement data of 22 nm FDSOI [3] was used. Fig. 2 shows the MOSFET off-current ($I_{off}$) at the two temperatures. The subthreshold slope at 70 K is 2.3 times steeper than that of 300 K, and the $V_{th}$ is increased by 60 mV, resulting in several orders smaller $I_{off}$ compared to the 300 K. Regarding the on-current ($I_{on}$), the increased carrier mobility conducts a larger $I_{on}$ than the 300 K device even the 70 K device enter the inversion region later. The two features enable device engineering to further boost its cryogenic performance. In this paper, two $V_{th}$ engineering methods were adapted toward high-speed for GC-eDRAM and toward low power for the logic transistors in the MAC units.

### C. Simulation setup

Fig. 3 is the simulation flow. Based on the calibrated transistor model from measurement data [2], HSPICE simulation is performed to obtain the expected new $I_{DS}$-$V_{GS}$ curve after $V_{th}$ parameter is modified for $V_{th}$ engineering (in practice by metal work function engineering). Transistor-level simulation at cryogenic temperature is done to obtain the performance of TSPC and GC eDRAM. Digital standard cell library is modified to support synthesizing the MAC units. Based on the GC eDRAM cell-level result and $V_{th}$ engineered periphery, the performance of global buffer was estimated with DESTINY, a memory simulator [4]. Finally, the

area/ power performance of the TPU is estimated by a modified NeuroSim simulator [5] with the GB, MAC and TSPC simulation data.

## III. DTCO AT CRYOGENIC TEMPERATURE

### A. *Gain-Cell embedded DRAM with $V_{th}$ engineering*

So far, the global buffer in TPU has been comprised with 6T-SRAM. However, apart from the advantage of fast read/write

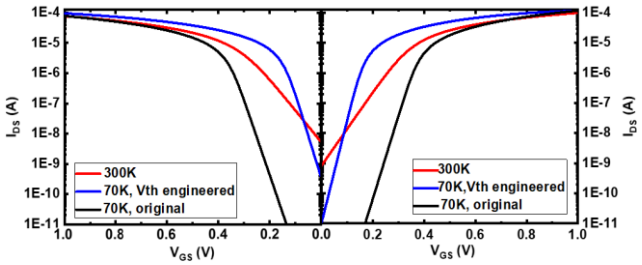



Figure 5. $I_{DS}$-$V_{GS}$ curve of the $V_{th}$-engineered PMOS (left) and NMOS (right) for logic transistor (W/L = 90 nm/22 nm) ($V_{th}$ engineered MOSFET $V_{DS}$ = 0.3V)

speed, the large cell size of SRAM (~160 $F^2$) limits the memory array area and power performance scaling. The GC-eDRAM, which comprises only two transistors while maintaining a similar read scheme as SRAM, has an advantage on those aspects. The GC-eDRAM is made of write transistor M0, read transistor M1, and the storage node SN, which is usually the gate capacitance of M1. The write operation of the GC-eDRAM is held by charging/discharging the SN by turning on the M0. The voltage stored in the SN turns on/off the M1, in read operation. The characteristic of cryogenic MOSFET with low $I_{off}$ enables GC-eDRAM to be logic compatible, allowing faster operation speed under small cell size and power consumption. Also, recent papers including [6] suggested Vth engineering towards a negative direction for a low-temperature process, employing the steeper SS slope in the lower temperature. Such Vth engineering can be also adjusted in GC-eDRAM for a further improvement. Fig. 4(b) shows the $I_{DS}$-$V_{GS}$ curve of the $V_{th}$-engineered NMOS that is optimized for GC-eDRAM cell. $V_{th}$ is reduced to 350 mV when the retention time of GC-eDRAM (0.5 V voltage difference at 0.2 msec after write) is compared to that of the room-temperature eDRAM design [7]. The $I_{on}$ of the engineered NMOS is 1.13 mA/μm, which is 1.5 times larger than that at 300 K, allowing even faster read-out.

### B. *Logic transistor $V_{th}$ engineering*

As aforementioned, 10 times chip power reduction is required to offset the cooling cost to operate the chip at 77K. To meet this requirement, $V_{th}$-engineering is also applied to the logic transistors as shown in Fig. 4. $V_{th}$ is reduced to 150mV so that the power supply $V_{DD}$ can be reduced from 1.0 V to 0.3 V. Still, the $I_{off}$ (N/P :11 pA/390 pA) is only about 1.1% and 7.1% of the 300 K N/PMOS (850 pA/5.4 nA), respectively. The new logic transistor model is applied to the registers, peripheral circuits of the GB, and PEs. Although the $I_{on}$ was also reduced, the reduced $V_{DD}$ still guarantees that the digital circuits can operate sufficiently at the frequency of the cryo-TPU.

### C. *Determining wire width of cryogenic system*

When a load is connected to the end of the wire, the RC delay and energy consumption in transmitting a signal from the beginning of the wire to the load are approximated as follows.

$$R_{wire} = \rho\frac{L}{W},\ C_{wire} = \varepsilon * \frac{W*L}{d} \quad (1)$$

$$\tau = R_{wire} * (C_{wire} + C_{load}) = \rho\varepsilon\frac{L^2}{d} + \rho\frac{L*C_{load}}{W} \quad (2)$$

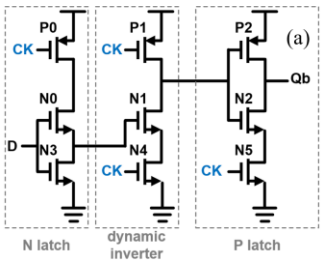
$$E_{wire} = \frac{1}{2}C_{wire}{V_{dd}}^2 = \frac{\varepsilon WL}{2d}{V_{dd}}^2 \quad (3)$$

The W and L is width and length of the wire, $\rho$ and $\varepsilon$ are wire resistivity and permittivity, respectively. $d$ is the distance from wire to neighboring metal or substrate, which contributes to wire capacitance. Reducing the wire width will increase RC delay in the local routing because $C_{load}$ is relatively larger than $C_{wire}$. On the other hand, it is no longer the case in global routing where $C_{wire}$ becomes relatively larger. Based on this, the optimal wire width is set to trade-off between latency and power consumption. At 77 K, the reduction of wire resistivity (×0.6) [8] can enable new design knobs. If the wire width is reduced following the trend of the resistivity, the delay of the local wire will become equal to the 300 K (assuming $C_{load}$ is much larger than $C_{wire}$), but the global wire will maintain the reduced RC delay (with the same factor of the reduced resistivity). The important thing is that the reduced wire width can reduce the dynamic energy consumed in both the local and global wire. As discussed earlier, the cryogenic GC-eDRAM based GB already guarantees sufficient latency due to increased transistor $I_{on}$ and cell size reduction by using GC-eDRAM, reducing the width can help save entire system's power, which is strictly limited by the cooling power. To prove the viability, latency, and energy consumption of the same 70 K GC-eDRAM based memory array were estimated by reducing wire width (×0.6). The The

bank-level simulation was conducted with DESTINY. While the latency slightly increased from 901 ps to 915 ps, the energy consumption was significantly reduced by 25 %, from 25 pJ/op to 20 pJ/op. 1 op is defined as read/write 128 bits of a word.

### *D. TSPC D flip-flop*

The TSPC D flip-flop replaces the conventional master-slave D flip-flop to reduce the area and the power consumption caused by the registers in PE. Fig. 6(a) is the circuit schematic of the TSPC D flip flop. When the clock is low, the input data is stored at the dynamic inverter output. When the clock rises, P-latch becomes transparent and output is changed following the dynamic inverter output, while N-latch is closed to block the input data. At the 300 K, floated dynamic inverter output node (D = 1, CK = 0) has a risk of data loss due to high leakage current. Since leakage current is significantly reduced at 70 K, the risk of data loss is now reduced and only advantages remain. TSPC D flip-flop has a shorter propagation delay as it has a lower logic depth than the master-slave D flip-flop, and it is advantageous in terms of area and power consumption because it uses less transistors. The comparison result between the two flip flops is shown in Fig. 6 (b). The TSPC D flip-flop consumes less than half of the dynamic and static power compared to the conventional D flip flop, while the latency is relatively similar (23 ps for conventional model, 27 ps for TSPC model).




| (b) | 300K, master-slave | 70K, TSPC |
|---|---|---|
| Dynamic energy (fJ/op) | 1.5 | 1.0 |
| Standby power (nW) | 4.3 | 0.01 |
| Area (μm²) | 1.23 | 0.58 |

Figure 6. (a) Circuit schematic of the TSPC D flip-flop (b) Comparison table of master-slave D flip-flop (300 K, 1 V) and TSPC D flip-flop (70 K, 0.3 V)

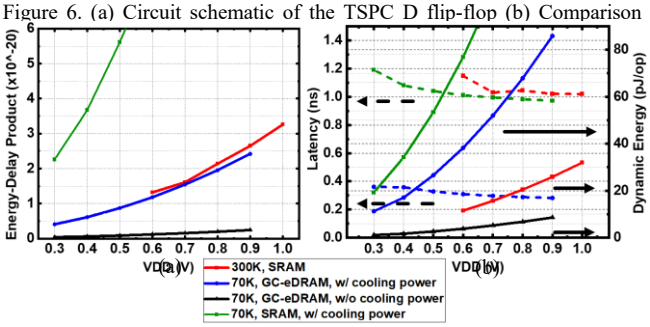



Figure 7. (a) Energy-delay product (EDP) (b) latency, dynamic energy consumption of the 2 MB global buffer, as a function of the supply voltage.

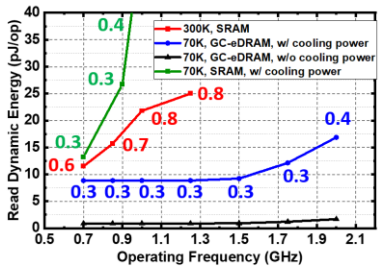



Figure 8. Dynamic energy consumption of the global buffer according to operating frequency at the respective peripheral supply voltage ($V_{DD}$ labeled).

## IV. BENCHMARK SIMULATION RESULTS

### *A. Global buffer benchmark*

Global buffer performances using SRAM and GC-eDRAM at 300 K and 70 K were estimated using the DESTINY. For global buffers at 70 K, aforementioned $V_{th}$-engineered logic transistor, and narrower wire width were adapted. The energy-delay product (EDP), energy consumption per operation, and latency according to the peripheral $V_{DD}$ are plotted in Fig. 7. Throughout the simulation, the cooling power was estimated by using the following equation:

$$P_{cool} = P_{chip}\left(\frac{300-T}{T}\right) * \eta \qquad (4)$$

$P_{chip}$ is the power consumed by chip at temperature T and $P_{cool}$ is the refrigerator power to tolerate $P_{chip}$. Finally, $\eta$ is the cooling efficiency. Collecting the data on commercial refrigerators [2], $\eta$ was set to 3. Even including the refrigerator power, the GC-eDRAM based GB at 70 K (cryo-GB) shows better EDP until 0.6V. Furthermore, $V_{DD}$ can be reduced to 0.3 V at 70 K, thanks to the $V_{th}$

engineering. It makes the buffer achieve a 31% lower EDP than 300 K (Fig. 7(a)). Regardless of the supply voltage, the latency is always better in cryo-GB due to the less cell size and wire resistivity (Fig. 7(b)). The energy consumption per operation according to operating frequency at both temperatures are plotted in Fig. 8. The energy consumption is relatively similar at 700 MHz, the clock frequency of the room temperature TPU [1]. As frequency rises to 1.25 GHz, the 300 K SRAM based GB's energy consumption increases drastically and its frequency could not reach further increase in given bank size limit. On the other hand, the energy consumption of the cryo-GB is lower in the overall frequency region, and even could reach the operation frequency up to 2 GHz. This simulation result suggests that the cryo-GB is a more suitable candidate when TPU pursues higher performance

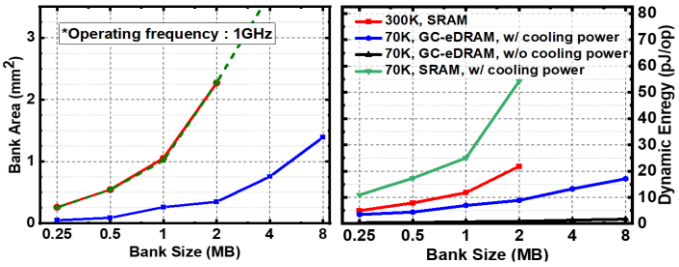



Figure 9. (a) Area and (b) dynamic energy consumption of the global buffer, according to the bank size.

TABLE I. AREA AND POWER BREAKDOWN OF 300 K TPU WITH 12 MB/4 MB SRAM BUFFER AND CRYO-TPU WITH 12 MB EDRAM BUFFER

| | | 300 K, SRAM, 12 MB | 300 K, SRAM, 4 MB | 70 K, SRAM, 12 MB | 70 K, GC-eDRAM, 12 MB |
|---|---|---|---|---|---|
| Area ($mm^2$) | MACs | 44.7 | 44.7 | 44.7 | 44.7 |
| | Register | 41.5 | 41.5 | 19.6 | 19.6 |
| | GB | 13.6 | 4.5 | 13.7 | 2.1 |
| | Total | **99.8** | 91.0 | 77.9 | **66.4** |
| Energy (μJ) | MACs | 67.8 | 67.8 | 7.1 (72.0) | 7.1 (72.0) |
| | Register | 33.1 | 33.1 | 2.4 (24.1) | 2.4 (24.1) |
| | GB | 23.0 | 38.2 | 5.6 (56.7) | 0.9 (9.3) |
| | DRAM | 3.7 | 74.0 | 1.8 (18.4) | 1.8 (18.4) |
| | Total | 127.6 | **213.1** | 16.9 (171.2) | **12.3 (123.7)** |

* Values in parentheses are energy consumption w/ cooling power.

in the future. To make it more consistent with room temperature TPU, 1 GHz operating frequency was used for the following simulations.

To show the area efficiency of the cryo-GB, the area and energy consumption at the two temperatures were compared according to the memory bank size and plotted in Fig. 9. The reduced cell size is translated to the bank-level area, showing that even 4 MB bank area of cryo-GB is almost a third of the 2 MB 300 K SRAM based GB. Due to the reduced wire length and wire capacitance, the number of the repeaters decreases correspondingly. It is confirmed that the energy consumption of the 2 MB cryo-GB is less than half the 2 MB 300 K SRAM based GB. In the following TPU simulation, 2 MB bank size was used for both 70 K and 300 K global buffer designs.

### *B. TPU benchmark*

With the $V_{th}$ engineered logic transistors and proposed global buffer, cryo-TPU area and energy consumption were estimated using the modified NeuroSim simulator and listed in Table I. The VGG-8 network was used for the simulation. Throughout the simulations, MAC units and registers were kept as configured similarly as in the reported TPU [1], but cryo-registers used TSPC D flip-flops. Global buffer settings were varied with three options: 12 MB SRAM, 4 MB SRAM and 12 MB GC-eDRAM.

Firstly, comparing SRAM and GC-eDRAM at 70 K shows that the GC-eDRAM can successfully replace SRAM in the cryo-TPU, both in terms of energy and area. Next, comparing the 300 K SRAM based GB and 70 K GC-eDRAM based GB (cryo-GB) with the same capacity (12 MB) shows how much area can be saved with the same configuration and total power budget. The results show that a cryo-TPU can save close to 33% of the chip area while consuming comparable power. Lastly, the advantage will be more evident as the buffer array capacity increases and can store more data for targeting a larger neural network. Comparing a small SRAM based GB (4 MB) and larger cryo-GB (12 MB) can show such benefits, because off-chip DRAM access to reload the weight is reduced. The cryo-GB could store all the weights in the simulation case, while the SRAM based GB could store only 1/3 of them. While the LPDDR interface is only needed for the output & input transmission at 70 K operation, the 300 K operation needs frequent DRAM access to update the rest 8 MB of the weight. The simulation result shows that less DRAM access reduced 40% of the total power (even after considering the cooling cost). If the area of cryo-GB is further increased until it occupies the same TPU area, more energy saving can be expected to support larger neural network models.

## V. CONCLUSION

The benefits of operating a TPU in cryogenic temperature are showcased in this paper. The DTCO techniques including $V_{th}$ engineering, TSPC D flip-flop, and GC-eDRAM based global buffer were suggested to maximize the performance improvement. With the $V_{th}$ engineered NMOS, GC-eDRAM replacing SRAM of the global buffer showed improvement over 60 % in latency, over 59 % in power consumption and over 80 % in memory density in a 2 MB bank size. Finally, the TPU benchmark simulation shows that the proposed TPU designed for 70 K operation has better energy/area performance. Even when the refrigerator's cooling cost is considered, the area can be saved in iso-energy conditions when targeting a small neural network, or energy can be saved by increasing global buffer size when TPU targets a large neural network.


## ACKNOWLEDGMENT

This work is supported by NSF-CCF-2218604. The authors thank Prof. Suman Datta for providing the cryogenic model.


## REFERENCES


[1] N. P. Jouppi, et al, "In-datacenter performance analysis of a tensor processing unit," *ACM/IEEE International Symposium on Computer Architecture (ISCA)*, 2017.

[2] Report on Cryogenic Electronics and Quantum Information Processing, *International Roadmap for Devices and Systems (IRDS)*, 2020.

[3] W. Chakraborty, et al., "Cryogenic RF CMOS on 22nm FDSOI platform with record $f_T$=495GHz and $f_{MAX}$=497GHz," *IEEE Symposium on VLSI Technology,* 2021.

[4] M. Poremba, S. Mittal, D. Li, J. S. Vetter, Y. Xie, "DESTINY: A tool for modeling emerging 3D NVM and eDRAM caches," *ACM/IEEE Design, Automation & Test in Europe Conference (DATE)*, 2015.

[5] A. Lu, X. Peng, W. Li, H. Jiang, S. Yu, "NeuroSim simulator for compute-in-memory hardware accelerator: validation and benchmark," *Frontiers in Artificial Intelligence*, vol. 4, 659060, 2021.

[6] P. Wang, X. Peng, W. Chakraborty, A. Khan, S. Datta and S. Yu, "Cryogenic Performance for Compute-in-Memory Based Deep Neural Network Accelerator," *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021.

[7] K. C. Chun, P. Jain, T. Kim, C. H. Kim, "A 1.1V, 667MHz random cycle, asymmetric 2T gain cell embedded DRAM with a 99.9 percentile retention time of 110µsec," *IEEE Symposium on VLSI Circuits*, 2010.

[8] H. L. Chiang et al., "Cold CMOS as a power-performance-reliability booster for advanced FinFETs," *IEEE Symposium on VLSI Technology*, 2020.