Monolithic 3D Compute-in-Memory Accelerator with BEOL Transistor based Reconfigurable Interconnect

Yandong Luo[1], Sourav Dutta[2], Ankit Kaul[1], Sung-kyu Lim[1], Muhannad Bakir[1], Suman Datta[2], Shimeng Yu[1]
[1]Georgia Institute of Technology, Atlanta, GA, USA [2]University of Notre Dame, Notre Dame, IN, USA
Email: shimeng.yu@ece.gatech.edu

2021 IEEE International Electron Devices Meeting (IEDM) | 978-1-6654-2572-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IEDM19574.2021.9720690

***Abstract*—**CIM-based inference engine with FeFET is projected to exhibit excellent energy efficiency, but its area scaling is still limited by availability of logic voltage compatible FeFET at leading-edge node. This work performs system-technology co-design (STCO) of a monolithic 3D (M3D) CIM accelerator using back-end-of-line (BEOL) compatible oxide channel MOSFET and FeFET, where W-doped In_2O_3 (IWO) NMOS is utilized to design area-efficient M3D write circuit and the IWO-based FeFET is adopted as the routing switch for reconfigurable interconnect. From a system-level evaluation, our M3D IWO FeFET design (utilizing a hybrid 22nm/7nm M3D partition) shows 2.9× times higher energy efficiency than a 7nm SRAM design with comparable chip area. The chip area can be further reduced by improving the electron mobility of the oxide channel.

I. Introduction

CIM is one of the promising paradigms for deep neural network (DNN) acceleration. A CIM array (Fig. 1(a)) consists of three parts: memory array with either 8T-SRAM or emerging non-volatile memories (eNVMs); write circuit for weight programming, which requires I/O transistor to provide high write voltage; and mixed-signal peripheral circuit for partial sum processing including analog-to-digital converter (ADC) and shift-add. Among eNVMs, ferroelectric field effect transistor (FeFET) stands out due to its high on-state resistance (R_{ON}>100kΩ) and low write energy (<10fJ). According to DNN+NeuroSim [1], the energy per op for 22nm FeFET CIM array with 2-bit per cell outperforms that of 8T-SRAM at 7nm node (Fig. 1(b)). However, the area cost for FeFET CIM array is 9× larger than 7nm SRAM, which is limited by the availability of FeFET today only at legacy node (e.g. 22nm). One solution is to harness M3D integration [2] where the memory cell and write circuit are fabricated on the top tier utilizing design rules for a legacy node, while the mixed signal and logic peripheral circuits are fabricated on the bottom tier with a leading-edge node. In such M3D scheme, the chip area is still limited by the FeFET write circuit as it is difficult to reduce the FeFET write voltage below 1V.

Besides the technological challenges mentioned above, another limitation of the architectural design arises from the rigidity of today's CIM design, which is usually customized for one specific DNN model. The hardware overhead to achieve reconfigurability between different DNN models is significant. For example, in FPGA, the connection box and switching box consume up to 70%-80% of the total chip area [3], because the routing element based on NMOS pass transistor + SRAM configuration bit cell is not area efficient.

We address both technological and architectural challenges by using IWO NMOS [4] and IWO FeFET [5] to design area-efficient M3D write circuit for CIM array and routing switch at BEOL for reconfigurable CIM architecture, respectively.

II. IWO Transistor Technology

In the sequential M3D scheme, the processing temperature of top tier transistors should be constrained below 400℃ in order not to degrade the bottom silicon CMOS transistors. Technology options include silicon recrystallization by laser annealing [6] and amorphous oxide semiconductor, where IWO is one of the promising channel materials [7]. In the IWO material, the tungsten acts as both electron donor and stabilizer by absorbing the oxygen vacancies. Fig. 2(a) shows the device structure of a double-gated IWO transistor on the top tier. The IWO channel is sandwiched between the top and back gate electrodes. Thanks to the large bandgap of the IWO, IWO transistor can endure high voltage (2.5V) even with a short gate length of 30nm~50nm. By integrating hafnium-zirconium oxide (HZO) into the gate stack, IWO FeFET is obtained, which can be utilized as a memory device for storing configuration bit and as a pass transistor concurrently. It should be noted that only NMOS is available for IWO. Hence, to design M3D circuit, transistor-level partitioning is needed by using IWO transistor on the top tier and PMOS on the Si, which is interconnected by the inter-tier vias (MIV).

Fig. 2(b) shows the I-V curve of the IWO transistor, CMOS logic and 2.5V I/O transistor. The performance of IWO transistor is noticeably inferior to CMOS logic due to its relative low electron mobility (~20cm^2/V·s). Nevertheless, under high V_{gs} (2.5V), IWO transistor shows competitive performance over I/O transistor. Therefore, it can be regarded as a good replacement for NMOS I/O transistor on the top tier. A comparison between IWO transistor and laser recrystallized Si transistor is made in Table I. IWO transistor shows lower leakage and the capability to be processed at even lower temperature. More importantly, it offers higher endurance to high voltage operation thus benefits the write circuit design.

III. Peripheral Circuit Design with IWO Device

A. Level Shifter and Output Driver Design

The high programming voltage (2~4V) of FeFET is delivered by the write circuit using level shifter and the output driver. M3D write circuit is designed with IWO NMOS and Si PMOS, as shown in Fig. 3(a). The parasitic capacitance and resistance of the 3D interconnect are estimated to be 0.18fF and 91Ω, respectively, assuming 6 metal layers at the bottom tier with M1~M4 pitch=40nm, M5~M6 pitch=64nm and MIV diameter of 30nm [8].

The timing diagram of the write circuit is shown in Fig. 3(b). The M3D design with the same transistor size as CMOS (IWO-same) and a scaled-up design (IWO-sized_up) are considered. 1V (I/O voltage at 7nm node) and 2.5V is assumed as the VDDL and VDDH, respectively. For IWO-same, the delay of the input stage of the driver is increased by about 0.52ns compared to the CMOS baseline, which is attributed to the degraded I_{ON} of IWO transistor when operating at VDDL. Therefore, when the IWO transistor width of the input stage is sized up to 720nm (in IWO-sized_up), almost the same performance as CMOS design is achieved.

From the layout in Fig. 3(c), it is observed that the width of the IWO-based level shifter is reduced by 41% compared with the CMOS design due to the reduced gate length (50nm vs. 270nm). The height of the IWO-based level shifter is slightly increased by 0.27μm due to the increased transistor width and the additional gate contact for double-gate device structure. In the M3D write circuit, the layout is partitioned at half-height and then folded. MIVs are inserted at the cross-section, which induces height overhead of half metal pitch and therefore the area of the M3D design is about 51.2% of the 2D unfolded layout.

B. *Routing Switch Design with BEOL FeFET*

The schematic of a 6T-routing switch is shown in Fig. 4(a). The typical NMOS+SRAM routing element is replaced by an IWO FeFET on the top tier and the peripheral circuits are fabricated at the bottom tier. The circuit schematic of an input-output pair is shown at the bottom of Fig. 4(a). A minimum-sized PMOS level restorer is utilized to restore the signal and then the output drivers deliver the signal across the wires to the next routing switch. The width of FeFET and PMOS level restorer should be carefully chosen so that the PMOS is not too strong to pull down the output of FeFET. Fig. 4(b) shows the timing diagram of the IWO FeFET (22nm) and CMOS routing switch (7nm). The periphery of IWO FeFET switch is assumed to be 7nm at the bottom tier. The width of the IWO FeFET is 400nm so that its output can be pulled down. The output delay of the IWO FeFET is only increased by 0.1ns compared to CMOS. To further scale down FeFET width, the electron mobility in the oxide channel should be increased (Fig. 4(c)). It is noted that the routing switch requires low R_{ON} (<10kΩ) as opposed to that of weight memory cell. To use W=100nm BEOL FeFET, a mobility increase of 4~5× is needed. The layout of 7nm CMOS 6T-routing switch and 22nm BEOL FeFET routing switch with W=100nm are shown in the right of Fig. 5. Placing routing switch at BEOL saves the bottom tier silicon area to fabricate other circuits. The unit area for each circuit module is summarized in Table II.

IV. RECONGIRUABLE CIM ARCHITECTURE DESIGN

A similar design principle to FPGA is adopted for the reconfigurable CIM architecture (Fig. 5). It consists of processing element (PE), switch box (S) and connection box (C). An 8×8 PE array with BEOL FeFET weight memory is assumed in this work. In each PE, there are 4×2 CIM array with each array size being 144×128. The input is delivered to the read word line (RWL) by the input driver. The partial sum current is sensed by the peripheral ADC. The write operation is conducted through the write word line (WWL) and write bit line (WBL), which are driven by the BL and WL level shifters with M3D design, respectively. Each PE also consists of input buffer, accumulation units, special function units for batch norm (BN), ReLU and pooling. All those modules are placed at the bottom tier with 7nm process. The 256-bit bandwidth routing switch is placed at the top tier using IWO FeFET. Its level restorers, output drivers and programming circuit are placed at the bottom tier with 7nm process. It is assumed that CMOS circuit can be fabricated underneath the routing switch.

Fig. 6 show the design automation flow to achieve the compiling-time reconfigurability for DNN mapping. The input is a dataflow graph (DFG) of the DNN model, which consists of the operations including vector-matrix multiplication, BN, ReLU and pooling. Next, the DFG is converted to a hardware DFG with the architecture-specific information such as the PE size. In the hardware DFG, the large weight blocks are partitioned according to the PE size and split into multiple PEs. Additional edges are induced for the partial sum reduction from different PEs. For each layer, the BN and ReLU are mapped into the PE that produces the final output. In this design automation flow, simulated annealing is utilized to minimize the total communication distance between PEs. Finally, the routes are set up using the Pathfinder algorithm. If a link is shared by multiple PEs, they are properly scheduled to avoid the conflicts.

V. SYSTEM LEVEL PERFORMANCE EVALUATION

To evaluate the system-level performance, we first update the technology file of NeuroSim with the device parameters of the IWO devices. New circuit modules such as the M3D level shifter and IWO FeFET-based routing switch are also included. The M3D temperature profile is estimated using the integrated compact thermal model [2] assuming air-cooling. Next, a customized simulator is built using the reconfiguration information from the compiler and the hardware performance from the NeuroSim. In this work, 7nm 2D SRAM, 22nm 2D FeFET and hybrid M3D design (22nm top/7nm bottom) are included. The device parameters and architecture configurations are included in Table III. The 7nm 2D SRAM design uses SRAM-CIM array with CMOS routing switch. The cell-bit precision of the FeFET and SRAM are assumed to be 2-bit and 1-bit, respectively. To demonstrate the reconfigurability, the performance of ResNet-20 (w/ 0.27M weights) and ResNet-32 (w/ 0.46M weights) using the same hardware are evaluated for CIFAR-10 dataset.

Fig. 7 shows that compared to 7nm SRAM design, hybrid M3D design with BEOL FeFET achieves 2.9× higher energy efficiency due to a lower ADC energy consumption. With the reported IWO FeFET [9], the area is 11% higher than 7nm SRAM due to the sized-up routing switch at the top tier. With projected 5× improved mobility, the area becomes 18% smaller than 7nm SRAM. Further device engineering is needed to enhance the transport in the oxide channel as well as to reduce the interface contact resistance for BEOL transistor.

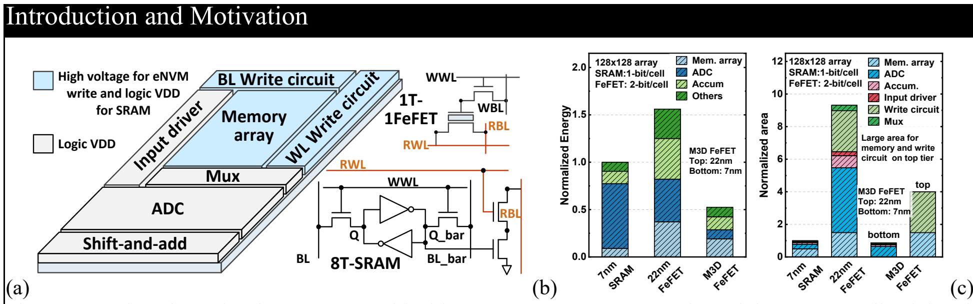
ACKNOWLEDGMENT

This work is supported by ASCENT, one of the SRC/DARPA JUMP centers, and by IMPACT, one of the SRC nCORE centers.

REFERENCES

[1] X. Peng et al., IEDM 2019. [2] X. Peng et al., IEDM 2020. [3] X. Chen et al., TCAS-I, 2019. [4] H. Ye et al., IEDM 2020. [5] S. Dutta et al., IEDM 2020. [6] F.K. Hsueh, et al., IEDM 2017. [7] W. Chakraborty et al., VLSI 2020. [8] Y.J. Lee et.al., ICCAD 2012. [9] S. Dutta et al., arXiv:2105.11078, 2021.

Introduction and Motivation



(c)

Summary of the issues

- The area for write circuit and memory cell does not scale down well as the other periphery operating at logic voltage
- Reconfigurability vs. circuit overhead of the reconfigurable circuit (e.g. routing switch)

Solutions proposed

- IWO transistor for M3D write circuit (level shifter and the output driver) design to reduce the area overhead
- BEOL IWO based FeFET as M3D routing switch
- Reconfigurable M3D CIM architecture and its compile flow

Figure 1 (a) The schematic of a CIM array with either 8T-SRAM or FeFET as the weight memory cell. High voltage (2~4V) is provided by the write circuit to program FeFET. The other peripheral circuits operate at the logic voltage. (b) The subarray level energy and area breakdown. For FeFET M3D design, the FeFET memory array and write circuit are fabricated on the top tier with 22nm node while the bottom tier circuits are using 7nm node. Here we assume regular Si CMOS and its I/O process are available on the top tier, which may not be feasible in practice, thus we just use this option for illustration purpose. The ADC area are reduced with 7nm process. But the write circuit area does not scale down if the programming voltage remains the same. (c) A summary of the issues and the solutions proposed by this paper.

IWO Transistor and IWO FeFET Technology at BEOL

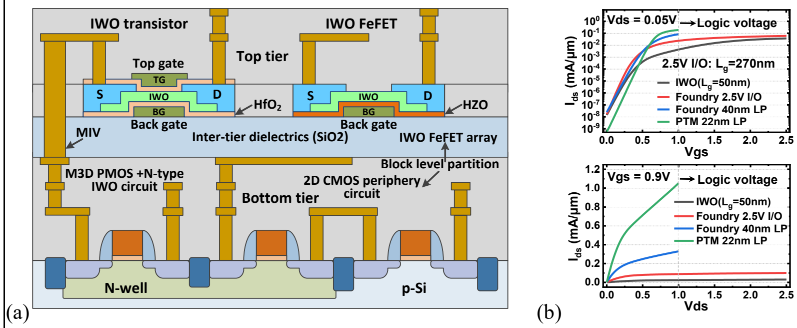


Table I Comparison of IWO transistor and laser recrystallized Si

Tech.	IWO [4]	Recry-Si [6]
μ(cm^2/V·s)	20	N/A
$T_{process}$ (°C)	<250	~400
I_{ON} (logic V_{DD})	<50 μA/μm	~400 μA/μm
I_{ON} (V_{gs}=2.5V)	~550 μA/μm	N/A
Leakage	<1fA/μm	<10pA/μm

Figure 2 (a) An illustration of the device structure of a double-gated IWO transistor and a single-gated IWO FeFET. The M3D circuit with transistor level partition uses IWO NMOS on the top tier and a PMOS on the Si. M3D circuit with block level partition can be achieved with IWO FeFET memory array on the top tier and 2D CMOS circuit on the Si. (b) The I-V curve of the IWO, CMOS 2.5V I/O and CMOS logic transistor. IWO transistor shows comparable performance to CMOS 2.5V I/O transistor when driving high voltage (e.g. V_{gs}=2.5V). However, its performance is not satisfactory under logic V_{DD} (e.g. 0.9V).

M3D Level Shifter Design with IWO Transistor

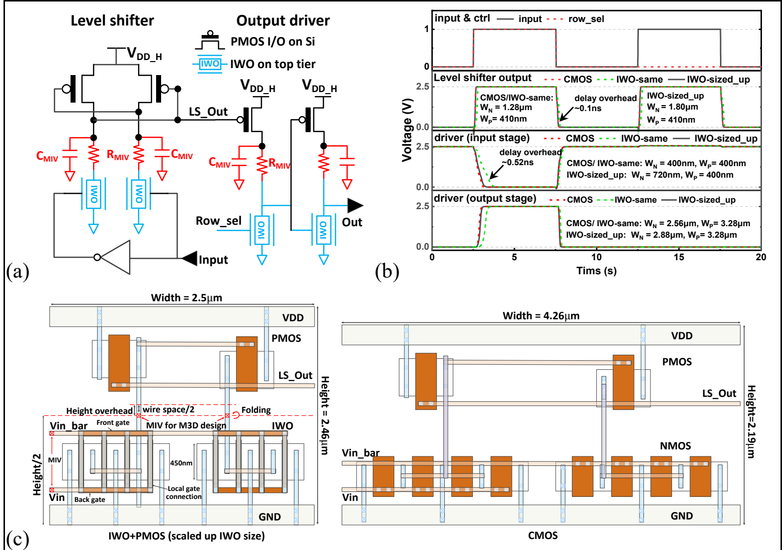


Figure 3 (a) The schematic of the M3D level shifter. It uses transistor level partition with IWO NMOS on the top tier and PMOS on Si. Parasitic capacitance and resistance are introduced by the metal stacks and MIV. (b) The timing diagram during the level shifter operation. For IWO transistor, two schemes are considered: one with the same size as Si NMOS transistor (IWO-same) and the other with increased width to match the delay of CMOS design (IWO-sized_up). (c) The layout of the unfolded M3D and 2D CMOS level shifter design with 2.5V I/O transistor design rule. The M3D design will be folded at the half-height of 2D design by inserting MIVs, which introduces height overhead. The area of the M3D design is 34% of 2D CMOS design, as listed in Table II.

M3D Routing Switch Design with IWO FeFET

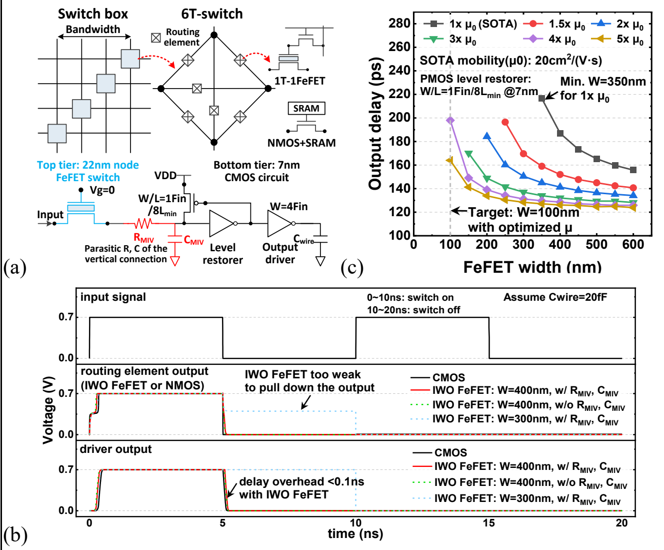


Figure 4 (a) The schematic of a switch box, 6T routing switch (top) and the circuit design for an input-output pair. (b) The timing diagram of the routing switch. The impact of R, C from M3D interconnect is negligible. The width ratio of the PMOS and pass gate should be properly chosen for correct circuit operation. Therefore, 400nm width FeFET is chosen with reported IWO FeFET device parameters [9]. (c) The output delay vs. electron mobility improvement. To use W=100nm BEOL FeFET, the mobility should be improved by 4~5 times. For each mobility value, the BEOL FeFET width starts from the min. value that gives the correct operation.

Architecture of M3D Reconfigurable CIM Accelerator Design with IWO Transistor and IWO FeFET Technology

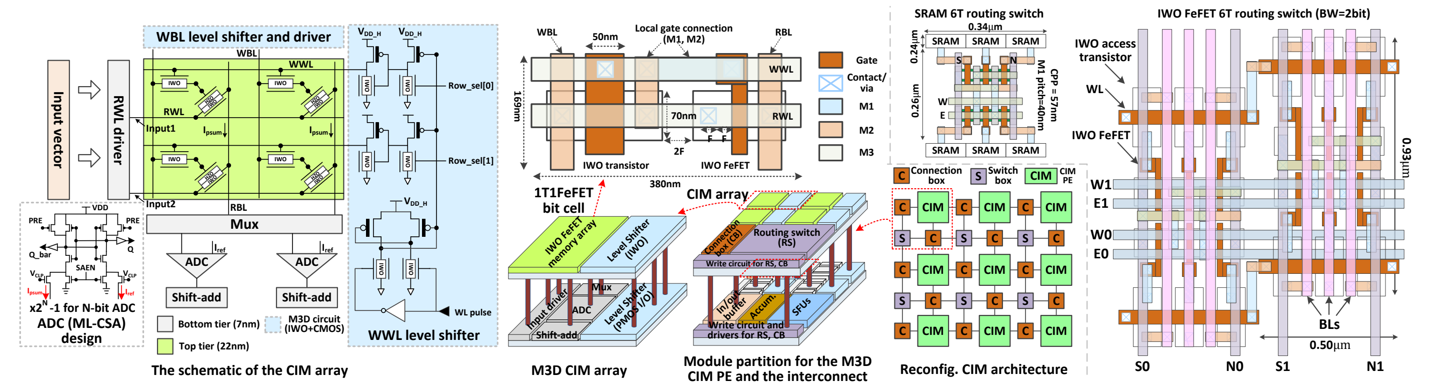


Figure 5 Left: The array level and architecture level design for reconfigurable CIM accelerator. The architecture follows a similar design principle as FPGA with connection box (mux) and switch box as the reconfigurable interconnect. In each CIM PE, there are 4×2 arrays with 144×128 array size. The IWO FeFET as weight memory has bit-cell size ~132F^2 (considering the IWO access transistor for 1T-1FeFET bit cell at 22nm process). The IWO access transistor is single-gate for a compact bit-cell design. Right: the layout of FeFET and NMOS+6T-SRAM routing element. It should be noted that the height of a routing box is smaller than the summation of height of each routing element due to the overlapping.

Flow for Compiling Time Reconfigurability and Hardware Performance Evaluation Methodology

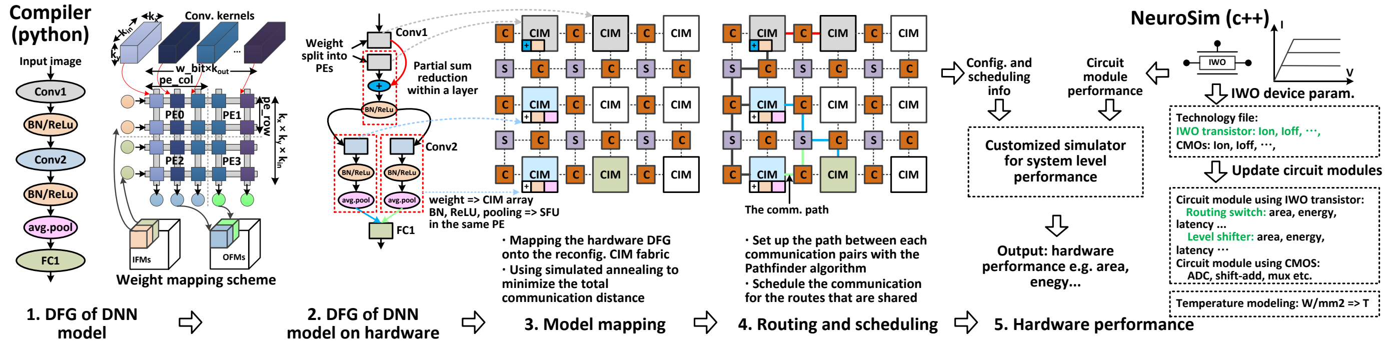

Compiler (python)
Input image
Conv1
BN/ReLu
Conv2
BN/ReLu
avg.pool
FC1
Conv. kernels
w_bit×k_out
pe_col
pe_row
PE0
PE1
PE2
PE3
IFMs
OFMs
Weight mapping scheme
Weight split into PEs
Conv1
Partial sum reduction within a layer
Conv2
FC1
weight => CIM array
BN, ReLU, pooling => SFU in the same PE
CIM
The comm. path
NeuroSim (c++)
IWO
IWO device param.
Config. and scheduling info
Circuit module performance
Customized simulator for system level performance
Output: hardware performance e.g. area, enegy...
Technology file:
IWO transistor: Ion, Ioff, ···,
CMOs: Ion, Ioff, ···,
Update circuit modules
Circuit module using IWO transistor:
Routing switch: area, energy, latency ...
Level shifter: area, energy, latency ···
Circuit module using CMOS:
ADC, shift-add, mux etc.
Temperature modeling: W/mm2 => T
1. DFG of DNN model
2. DFG of DNN model on hardware
3. Model mapping
4. Routing and scheduling
5. Hardware performance
· Mapping the hardware DFG onto the reconfig. CIM fabric
· Using simulated annealing to minimize the total communication distance
· Set up the path between each communication pairs with the Pathfinder algorithm
· Schedule the communication for the routes that are shared

Figure 6. The flow to compile the reconfigurable CIM accelerator. The input is the data flow graph (DFG) of the DNN model. Then the weight block of each layer is partitioned according to the weight mapping scheme and the PE size. The BN, ReLU and pooling operations are mapped onto the special function unit of the PE that produce the final OFM of a layer. The hardware performance is evaluated by integrating with M3D NeuroSim with the integrated thermal model [2].

Area of Circuit Modules

Table II Area of Circuit Modules (1 Unit)

Unit: μm^2	IWO+CMOS (2D unfolded)	IWO+CMOS (M3D)	2D CMOS
LS[1]	6.15	3.15	9.33
LS driver	9.65	4.96	9.74
RS[1,2]	0.46[3]	0.46[3]	0.25

1. LS=level shifter RS=routing switch
2. The area of RS is for one routing element. For M3D IWO FeFET, the area is the maximum of the top and bottom tier.
3. Assume W=100nm for BEOL FeFET with improved mobility

Hardware Performance Evaluation Set-up

Table III Device Parameters and Hardware Configurations

	2D 8T-SRAM	2D FeFET	M3D FeFET	M3D FeFET (Opt.)
Tech node.	7nm	22nm	7nm(btm)/22nm(top)	7nm(btm)/22nm(top)
R_{ON}	--	110kΩ (Si channel)	180kΩ (IWO channel)	180kΩ (IWO channel)
Bit cell size	0.0337μm^2(688F^2)	0.101μm^2(210F^2)	0.064μm^2(132F^2)	0.064μm^2(132F^2)
Cell precision	1-bit	2-bit	2-bit	2-bit
ADC precision	4-bit	5-bit	5-bit	5-bit
Routing switch	NMOS+SRAM	2D FeFET (W=400n)	IWO FeFET (W=400n)	BEOL FeFET (W=100n with 5×μ_0)

Energy and Chip Area Breakdown

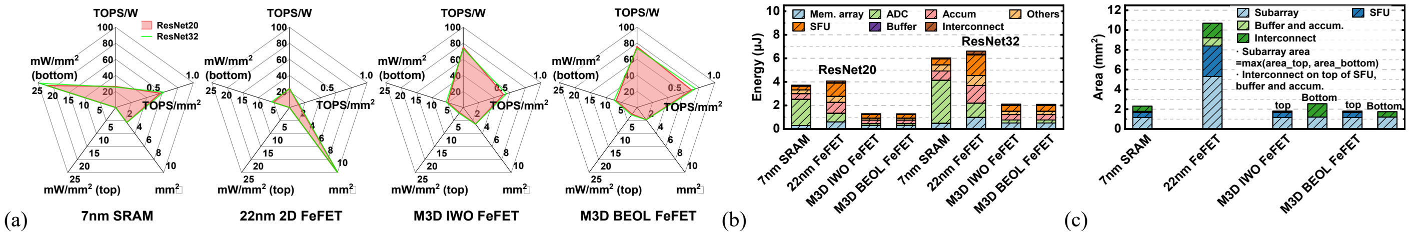


Figure 7 (a) The performance metrics of the reconfigurable CIM accelerator with different device options. The M3D BEOL FeFET refers to the design with optimized BEOL FeFET for routing switch. The number of CIM arrays is 4×4 in 7nm SRAM PE to obtain the same PE size as FeFET-based designs with 2-bit per cell precision. The operating frequency is assumed to be 200MHz for all designs. The PE utilization is about 34% for ResNet-20 and 53% for ResNet-32. ResNet-32 uses more resources and achieves 1% higher accuracy for CIFAR-10 image inference. The weight and activations are both 8-bit. The difference of the performance metrics is small when running these two DNN models. (b) The system level energy breakdown for one image inference. (c) The chip area breakdown. The area of the subarray is max(area_top, area_bottom). For the M3D designs, the interconnect (connection box, switch box) is on top of the SFUs (BN, ReLU and pooling units), buffer and accumulation units. The operating temperature for M3D FeFET based design is about 40℃ due to the low power density.