Quantifying the Benefits of Monolithic 3D Computing Systems Enabled by TFT and RRAM

Abdallah M. Felfel*†, Kamalika Datta*, Arko Dutt*, Hasita Veluri‡, Ahmed Zaky*,
Aaron Voon-Yew Thean‡, Mohamed M. Sabry Aly*
* Nanyang Technological University, Singapore, † Zewail City of Science and Technology, Giza, Egypt,
‡ National University of Singapore. E-mail: s-abdallah.felfel@zewailcity.edu.eg, msabry@ntu.edu.sg

***Abstract*—Current data-centric workloads, such as deep learning, expose the memory-access inefficiencies of current computing systems. Monolithic 3D integration can overcome this limitation by leveraging fine-grained and dense vertical connectivity to enable massively-concurrent accesses between compute and memory units. Thin-Film Transistors (TFTs) and Resistive RAM (RRAM) naturally enable monolithic 3D integration as they are fabricated in low temperature (a crucial requirement). In this paper, we explore ZnO-based TFTs and HfO_2-based RRAM to build a 1TFT-1R memory subsystem in the upper tiers. The TFT-based memory subsystem is stacked on top of a Si-FET bottom tier that can include compute units and SRAM. System-level simulations for various deep learning workloads show that our TFT-based monolithic 3D system achieves up to $11.4\times$ system-level energy-delay product benefits compared to 2D baseline with off-chip DRAM—$5.8\times$ benefits over interposer-based 2.5D integration and $1.25\times$ over 3D stacking of RRAM on silicon using through-silicon vias. These gains are achieved despite the low density of TFT-based RRAM and the higher energy consumption versus 3D stacking with RRAM, due to inherent TFT limitations.**

Keywords: Resistive RAM, Thin-Film Transistor, Monolithic 3D Integration.

I. Introduction

Recent memory-centric workloads, such as deep learning and graph analytics, require both high compute throughput and memory-access bandwidth. Such workloads expose the inefficiencies of limited access to DRAM in current computing systems. 3D integration of compute and memory, where compute and memory tiers are interconnected with vertical links, can potentially overcome this limitation and improve energy efficiency [1]. *Monolithic 3D integration* provides the highest density of such vertical links, as subsequent tiers are connected with interlayer vias (ILVs) with fine pitch—e.g., $100nm$ for $28nm$ technology node [2]. In contrast, 3D stacking connects abut tiers using through-silicon vias (TSVs) of a 5-10μm pitch [3]. While monolithic 3D integration can provide significant energy-efficiency benefits over current baseline systems with off-chip DRAM (and even TSV-based 3D stacking) [1], a key requirement is that all logic and memory devices at upper tiers must be fabricated at low temperature ($< 400^oC$), to preserve the already-fabricated lower tiers and prevent interconnects diffusion [4]. Silicon field-effect transistors (Si-FETs) require very high temperatures in fabrication (up to $1,000^oC$), which cannot be used in upper tiers [4].

New non-volatile memory technologies adopted by industry, such as Resistive and Magnetoresistive RAM (*RRAM* and *MRAM*), naturally enable monolithic 3D integration, thanks to their low-temperature fabrication [5]. While various emerging logic devices also enable monolithic 3D integration, such as Carbon Nanotube FETs (CNFETs) [4], 2D-material based FETs [6], and Si-based CoolCube [7], further efforts are required for each of them to enable very-large scale integration.

This paper aims to quantify the benefits of monolithic 3D integration enabled by *thin-film transistors (TFTs)* [8] and RRAM. To the best of our knowledge, this is the first work that provides a device-to-system quantification of the benefits of a monolithic 3D computing architecture with TFTs and RRAM. Our simulations indicate that the 3D monolithic-TFT-RRAM system achieves up to $11.4\times$ system-level energy-delay-product benefits (EDP: product of execution time and energy consumption) over baseline systems with off-chip DRAM—simultaneous $4.8\times$ speedup and $2.38\times$ energy reduction. Compared to a TSV-based 3D stack with RRAM and Si-FETs, our introduced monolithic 3D configuration achieves up to $1.25\times$ EDP benefits despite having $6.7\times$ lower RRAM density. Contributions of this paper are summarized as follows:

- Modeling RRAM cell with TFT (two fabricated and one projected models) as an access device.
- Designing complete memory-access circuitry with TFTs.
- Evaluating the EDP benefits of the introduced monolithic 3D system against various system configurations.

The paper starts by discussing related work in Section II, then the target architecture in Section III. Section IV elaborates the technology-enablers of this work. Section V discusses the modeling of TFT-based RRAM. Section VI presents the system-level analysis, followed by conclusion in section VII.

II. Related Work

A. Demonstrations of Monolithic 3D Systems

A 2-tier monolithic 3D system has been demonstrated in [9], which comprises a top tier with 26 Mbit TFT-based SRAM, and a bottom tier with Si-CMOS based FPGA fabric. The TFTs are fabricated at low temperature and scaled to $180nm$ gate length. More recent demonstrations include a 2-tier monolithic 3D system built entirely with TFTs and integrating TFT-based SRAM on both tiers [10]. Other monolithic 3D demonstrations have integrated RRAM in upper tiers alongside CNFETs [4] or 2D-FETs [6].

B. Analyzed 3D Computing Systems

Hardware accelerators for deep learning inference, integrated with on-chip 3D-stacked DRAM, leverage high-

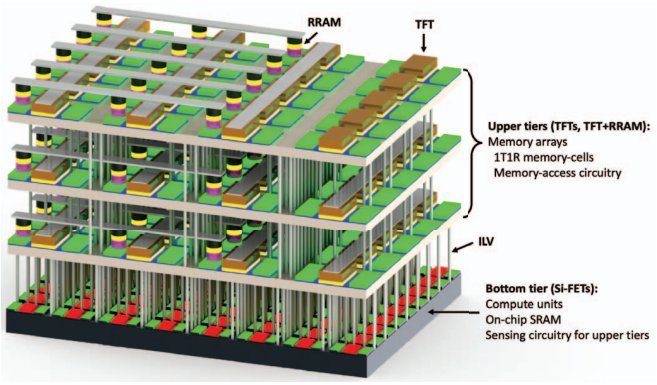

Figure 1: The introduced monolithic 3D architecture.

bandwidth access to DRAM using TSVs to improve execution time and energy consumption [11]. However, the few vertical TSV connections to DRAM, due to large TSV pitch, limits the potential benefits 3D integration can provide. Monolithic 3D computing architectures, enabled by CNFETs and RRAMs, can achieve up to $1000\times$ system-level energy efficiency benefits over current baseline architectures [1], verified by extensive simulations conducted on a wide range of deep learning workloads and architecture templates.

Design-automation frameworks that perform the physical design of monolithic 3D ICs are required to fabricate these analyzed systems. Fortunately, these frameworks can leverage commercial 2D IC design tools to generate the final GDSII output [12]—this is complementary to the work in this paper.

III. Target Architecture

Figure 1 illustrates the introduced monolithic 3D architecture, which comprises the following:

1) A *bottom tier*, built with high-performance Si-CMOS, which includes compute units, SRAM, compute-to-memory interconnects and a subset of memory access circuitry (details in Section V).
2) Multiple upper *memory tiers* built with TFTs and RRAM—both devices can be fabricated at low temperature. These tiers also include portions of the memory access circuitry entirely designed using TFTs.

ILVs vertically connect subsequent tiers. We assume that the TFTs that provide enough current to program RRAM (details later in Section V), can be completely fabricated either in n-type or p-type FETs. Consequently, certain modules—namely, sense amplifiers and latches—are implemented with Si-FETs (on the bottom tier) as such modules require both n-type and p-type FETs. Specific modules designed only with TFTs include: address decoder, voltage selectors, and all multiplexers (MUXes) (details in Section V). In this paper, we assume that we can vertically integrate multiple TFT-based memory tiers.

IV. Technology Enablers

A. TFT

TFT devices widely vary in their material, type, mobility, size, voltage requirement, and driving current. In particular, oxide-based TFTs have good electrical performance, large area uniformity, and fabricated at low temperature—hence they naturally enable monolithic 3D integration. We explore ZnO-based n-type TFTs [13], CuO-based and SnO_2-based p-type TFTs [14], and polysilicon-based n-type or p-type TFTs [10]. We find that n-type TFTs perform better than p-type TFTs because they inherently possess higher mobility [15]. In this paper, we have used ZnO-based TFTs instead of poly-silicon based TFTs because of their lower thermal budget and higher mobility. We deploy ZnO-based TFTs in the memory array for both the access transistors of RRAM cells in a 1-transistor 1-resistor (1T1R)[1] configuration, and the peripheral circuitry (details in Section V). These TFTs would then need to supply sufficient drive currents to RRAM cells.

Table I: Characteristics of the considered TFT models

Model	Mobility	$\mathbf{T}_{ox}$	SS	$\mathbf{V}_T$
v1-F	8.5 $cm^2/(V\text{-}s)$	35 nm	592.00 mV/dec	4.039 V
v2-F	51.2 $cm^2/(V\text{-}s)$	10 nm	160.96 mV/dec	1.000 V
v3-P	100.0 $cm^2/(V\text{-}s)$	10 nm	99.60 mV/dec	1.000 V

Tox: Oxide Thickness, **SS**: Sub-threshold Swing, $\mathbf{V}_T$: Threshold Voltage.

B. ReRAM/RRAM

A typical RRAM cell consists of a metal-insulator-metal stack in which data are stored and retrieved via voltage-controlled resistance switching [5]. By applying a suitable voltage bias, a conductive filament is reformed or broken via RESET (logic 0 at a high-resistance state, HRS) or SET (logic 1 at a low-resistance state, LRS) operations, respectively.

RRAM-based memory systems potentially offer high-density, energy-efficient, low latency, and non-volatile on-chip storage with single- or multi-bit per cell [1], [5]. In this paper, we use a 1T1R, single bit per cell, structure (Figure 2a), which reduces leakage current and solves the sneak path problem. Another challenge of RRAM is the limited endurance at the array level [16]. Therefore, we use RRAM as a read-only memory for the target workloads discussed in Section VI.

V. Memory Cell and Array Modeling

In this section, we discuss 1TFT-1R memory cell modeling, the memory array design with all peripheral circuitry, area, latency, and energy estimation of the memory array, and memory segment modeling. We use calibrated TFT and RRAM Verilog-A compact models along with $28nm$ Si-CMOS PDK for the design, and Cadence Virtuoso to simulate the array.

A. 1TFT-1R Cell Modeling

We use a ZnO-based TFT model [17], [18], where we tune the parameters based on the median of experimental data of 80 fabricated devices (Figure 2c). We use three different TFT model versions, namely, v1-F (fabricated), v2-F (fabricated), v3-P (projected). The RRAM device model is based on a compact physics-based dynamic model [19]. Figure 2a illustrates the 1TFT-1R structure, while Figures 2b-d illustrate I-V characteristics of used RRAM and TFT devices. Table I shows the device characteristics of the three TFT models.

We consider a cylindrical oxide-based RRAM with TiN electrodes, $6nm$ HfO_2 dielectric thickness and $400nm$ diameter. The RRAM model [19] is calibrated based on the median

[1]In this study, 1TFT-1R is identical to 1T1R.

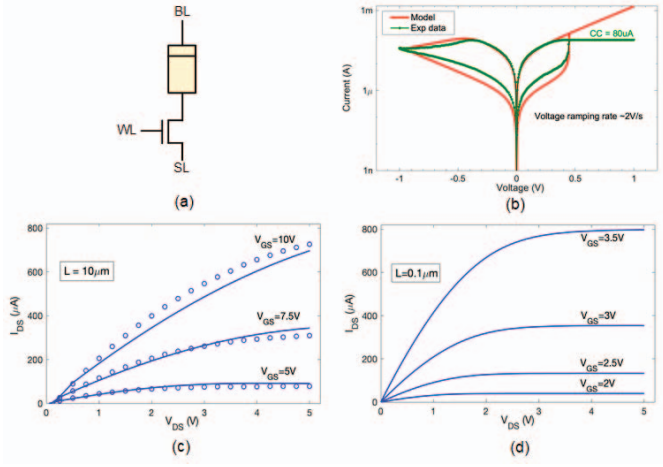

Figure 2: (a) 1TFT-1R memory cell, (b) RRAM experimental I-V DC characteristics (median) used in model validation, (c) TFT transfer characteristics of model v1-F—experimental (markers) and simulated (line), (d) TFT transfer characteristics of model v3-P.

of the experimental data of four similar RRAM cells. For each RRAM cell, 100 DC switching cycles with voltage ramping rate of $\sim 2V/s$ are conducted at room temperature with $80\mu A$ compliance current. The median SET (RESET) voltage is $0.445V$ ($-0.365V$) with a current of $80\mu A$ (87μA), as shown in Figure 2b. Cycle-to-cycle and cell-to-cell variations in SET (RESET) voltage is around $0.4V$ ($0.17V$). With $0.1V$ read voltage, the average measured R_{ON}/R_{OFF} programming window ranges from $4k\Omega/50k\Omega$ to $5.79k\Omega/500k\Omega$. As can be noted, largest variations are in the current levels of HRS. To obtain the experimental butterfly curve, we first simulate the RRAM cell model in series with a $0.2Hz$ sinusoidal voltage source of $1V$ amplitude. With proper RRAM parameters calibration, we then perform a DC hysteresis sweep for $5s$, to replicate the experimental lab setup. With $0.1V$ read voltage, simulation yields a R_{ON}/R_{OFF} of $5.46K\Omega/185K\Omega$, which is within the measured range.

Table II summarizes the required voltages for 1TFT-1R cell terminals, for the three TFT models. These voltages meet the required driving currents for SET/RESET/READ operations and maintain a consistent read ON/OFF ratio of $\sim$7–10 for the three models (akin to the measured ratios), with smallest device dimensions for TFTs (W=$1\mu m$, L=$0.1\mu m$).

B. Array Modeling

In this subsection, we present the design of a 1024×32 subarray, illustrated in Figure 3, that includes a 2D array of 1TFT-1R memory cells, address decoder, voltage MUX for WL/SL, voltage selector for BL, 2-to-1 MUX, and sense amplifiers. The address decoder, voltage MUX, BL voltage selector and 2-to-1 MUX are designed with ZnO-based n-type TFTs. We use static logic design style [20] for these circuits, which allows seamless integration with the 1TFT-1R array. However, this results in additional challenges—higher voltage requirements and interfacing issues. The sense amplifier, designed using $28nm$ Si-CMOS, includes a dynamic comparator, inverter, and a latch.

Table II: Voltage requirements for the 1TFT-1R cell

Model	Operation	BL (V)	SL (V)	WL (V)
v1-F	SET	2.12	0.00	4.23
	RESET	0.00	1.65	6.00
	READ	0.50	0.00	3.15
v2-F	SET	2.00	0.00	3.20
	RESET	0.00	1.13	4.82
	READ	0.50	0.00	2.50
v3-P	SET	1.85	0.00	2.25
	RESET	0.00	1.10	3.60
	READ	0.47	0.00	1.82

We design the 10×1024 address decoder using two 5×32 decoders with active-low outputs, followed by 1024 NOR2 gates (see Figure 3b). Figures 3g and 3h illustrate the schematics of inverter and NOR2 gates, with a VDD of $5.2V$ for v3-P (VDD=$8V$ for v1-F and $6.5V$ for v2-F). We use proper TFT sizing, highlighted in Figures 3c-h, for all components to ensure correct voltage delivery to 1TFT-1R cell terminals and achieve better output swing.

The address decoder selects one of the rows of the memory array—each row activates 32 cells. Voltage (analog) MUXes and selector circuits provide the required voltages to WL, SL, and BL, for SET/RESET/READ memory operations (Table II). We perform transient simulation and analysis of WRITE and READ operations as follows:

WRITE: 1TFT-1R cell is simulated using the RRAM device model. Since SET and RESET operations require different voltages on WL—to write an arbitrary sequence of ones and zeros in a memory word—we perform the write operation as follows. First, we RESET all bits of the word that needs to store '0', then SET the remaining bits within that word.

READ: RRAM cell is replaced by an equivalent resistance R_{on} (logic '1') or R_{off} (logic '0'). We initially pre-charge the bitlines, while simultaneously activate the required row. A sense amplifier for each bitline then compares the bitline voltage with a reference value (median between that of R_{on} and R_{off}). The sense amplifier outputs '1' if the read voltage is higher than the reference (i.e., detected R_{off}) and '0' otherwise. An inverter flips this output and stores it in a latch.

C. Area, Latency and Energy Estimation

We design and simulate the memory macro[2] using three different TFT devices, and estimate area, latency and energy. It may be noted that a TFT is much larger than a RRAM cell, and in an actual layout the access transistors and RRAM cells are vertically stacked. We estimate the 1TFT-1R array area, and the TFT memory-access circuits area, by summing the areas of all TFTs only, with $100nm$ pitch overhead for interconnects (the minimum gate length of used TFT—W=$1\mu m$ and L=$0.1\mu m$). Using v3-P model, the READ latency is $3ns$ with $1.3ns$ precharge time (in which all input signals raise the BL to a stable voltage), $1.1ns$ sensing time, and $0.6ns$ store-to-latch time. The row decoder is activated for $1ns$ during precharge. We estimate the energy consumption of each sub-circuit by multiplying the voltage drop by current

[2]Peripheral circuitry for v1-F and v2-F integrates a dedicated MUX for each WL, contrary to the design in Figure 3 for v3-P.

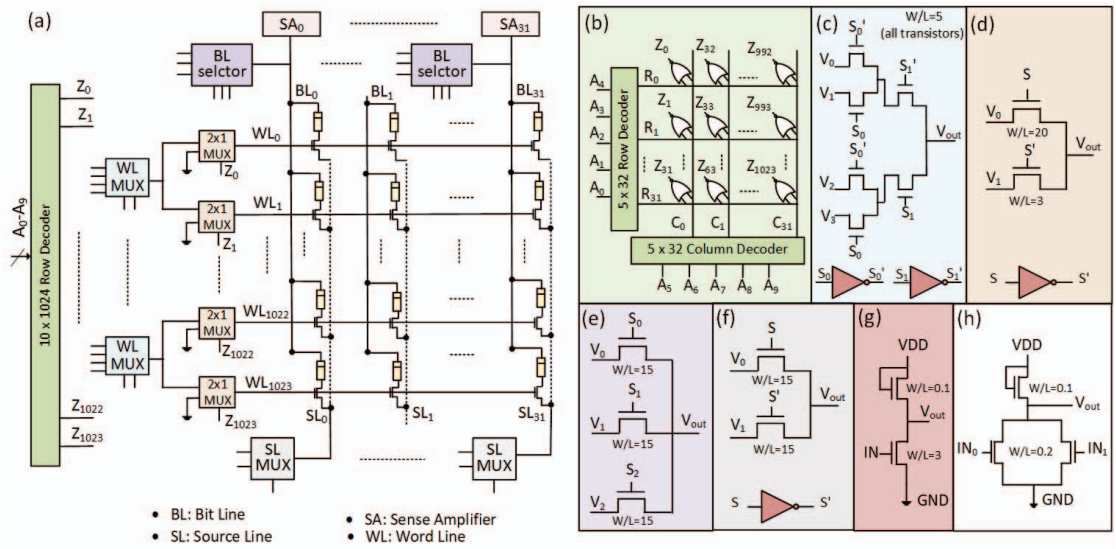

Figure 3: (a) Overall schematic, (b) 10×1024 decoder, (c) Word line multiplexer, (d) 2-to-1 MUX, (e) Bit line voltage selector, (f) Source line multiplexer, (g) NOT gate, (h) NOR2 gate. All S_X are select lines for the shown gates.

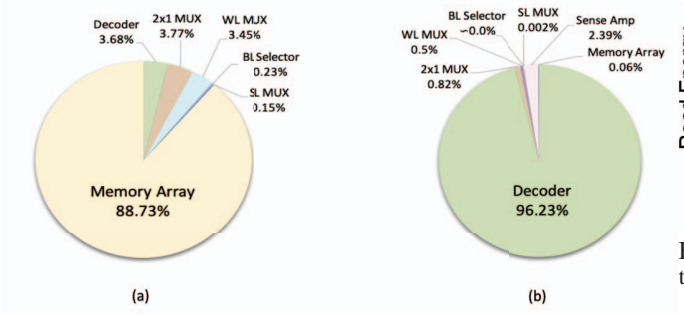

Figure 4: Breakdown of (a) area and (b) read energy for the 1024×32 memory array.

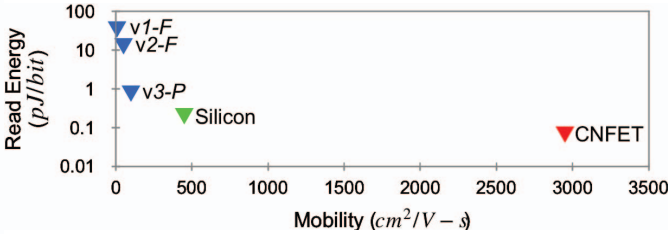

Figure 5: Read energy (log scale) versus mobility for various devices technologies (for a 1Mbit array) [22], [23].

flowing through it, and the operation latency. Figure 4 shows the breakdown of area and energy for a 1024×32 memory array when designed with v3-P. Table III summarizes the area and access energy of different memory sizes for the three models. We find larger arrays improve area efficiency and access energy per bit simultaneously. Access energy per bit is improved primarily due to the low gate capacitance of TFTs (order of 10^{-18}F [21]). So, bigger wordlines do not significantly increase the energy of the peripheral circuitry and their energy is shared across more bits.

As the TFT's carrier mobility improves, we expect lower access energy per bit due to the higher drive current. This is evident from the results illustrated in Figure 5, where we also simulate the memory array using $28nm$ Si-CMOS and CNFET designed at $28nm$ node [1]. Co-optimization of device properties and circuit components leads to lower energy. Moreover, a device with higher mobility enables further shrinking of transistor dimensions and usage of fewer components, e.g., we can use fewer WL MUXes as the transistors can drive multiple rows simultaneously, as shown in Figure 3a.

Table III: Area and energy estimates of different array sizes

TFT Version	Array Size	Area	AE	RE	WE	
					SET	RESET
v1-F	1024×32	0.025	28.60	1088	12276	13143
	1024×256	0.077	74.53	136.30	1553	2420
	1024×1024	0.256	90.01	34.30	404	1271
v2-F	1024×32	0.009	80.10	203.57	478	539
	1024×256	0.059	99.66	31.10	66	127
	1024×1024	0.230	98.85	12.62	22	83
v3-P	1024×32	0.008	88.29	13.94	317	455
	1024×256	0.058	98.00	2.04	46	69
	1024×1024	0.230	99.17	0.76	17	28

Area: in mm^2, **AE**: area efficiency (in %), **RE**: read energy (in pJ/bit), **WE**: write energy (in pJ/bit).

D. Memory Segment Modeling

We define the memory segment as a set of memory arrays that are connected to a single memory-access port. We utilize NVSIM [24] to estimate inter-array interconnect overheads, where H-tree interconnect is used to transfer data and addresses between arrays and the corresponding memory-access port. We represent each memory array with a black box of the same area. Then, for a given memory-segment capacity, we calculate the number of memory arrays within this segment. Finally, interconnect latency and energy values are estimated.

The memory controller, implemented in Si-CMOS, generates the select lines for the voltage MUXes connected to the memory array. To interface the TFT-based memory arrays with the memory controllers, we assume a 7-stage charge pump (e.g., [25]) to translate the voltage levels of logic '1' and '0' from Si-CMOS ($1V$ and $0V$) to the corresponding values of TFTs (e.g., $5.2V$ and $0V$ for v3-P). It consumes an energy

of $408fJ/bit$, and has significantly lower area compared to TFTs on upper tiers.

VI. System-Level Analysis

A. Simulation Methodology

We consider a hardware accelerator for deep learning [26] to quantify the system-level benefits of the introduced TFT-based monolithic 3D system. We summarize the architecture details of this accelerator in Table IV. We use TETRIS [11] to simulate the execution time and energy consumption of different system configurations when executing the inference of various deep learning workloads. This simulation framework takes the deep learning network structure, as well as timing and energy characteristics of each hardware component, to estimate the full execution time and energy consumption. We simulate the inference phase of different deep learning workloads, owing to their proliferation in numerous deep learning platforms [26]—the parameters of the analyzed deep learning networks are summarized in Table V. For timing and energy estimation, we synthesize the PE arrays using a foundry $28nm$ with $1V$ VDD; we assume the access energy and latency of SRAMs as shown in previous works for the same technology node [26]. We extract the area of the PE array from synthesis results (and include a 30% overhead for wiring), and estimate the SRAM area using NVSIM—the accelerator occupies $20mm^2$.

We analyze the following configurations, for which the parameters are summarized in Table IV:

- **2D Baseline**. The accelerator is connected to 1-GByte off-chip DRAM using LPDDR3 interface.
- **2.5D Interposer**. The accelerator is connected to 1-GByte DRAM via an interposer with a $40\mu m$ microbump pitch [1].
- **3D TSV-RRAM**. The accelerator is integrated with a Si-CMOS 1T1R RRAM in a 3D stack with TSVs ($5\mu m$ pitch). PEs and SRAM occupy the bottom tier, whereas RRAM-based memory with all memory-access circuitry occupy upper tiers. Using NVSim [24], we assume a $2.7MBytes/mm^2$ RRAM density.
- **3D Monolithic-TFT** (our introduced configuration). The accelerator is integrated with TFT-based RRAM memory (v3-P, Section V) in a monolithic 3D stack. The vertical ILV pitch is $100nm$ for the considered $28nm$ technology node. The RRAM density is $0.4MBytes/mm^2$.

We map data such that DRAM or RRAM are used to only store the weights, hence they experience only read requests. Nonetheless, weights can be updated into RRAM with a relatively high rate (e.g., once per hour), which extends the operating lifetime to more than 10 years assuming 10^5 RRAM endurance cycles. To accommodate all weights on chip, 3D TSV-RRAM requires 2 RRAM tiers, while 3D monolithic-TFT requires 10 RRAM tiers. Workloads that require significantly higher memory capacity may not be easily mapped to 3D monolithic-TFT due to the low density of TFT-based RRAM. This low-density increases the macro area and routing latency; however, our system-level analysis shows that system benefits are more sensitive to the bandwidth (due to our high density of ILVs connections) than latency or propagation delay.

B. Simulation Results

Table VI summarizes the system-level energy-delay product (EDP: product of application-level energy consumption and execution time) benefits of all configurations over 2D baseline. Our introduced monolithic 3D system achieves superior gains across all considered workload/configuration pairs, with up to $11.4\times$, $5.8\times$, and $1.25\times$ EDP benefits over 2D Baseline, 2.5 interposer, and 3D TSV-RRAM, respectively. These gains are mainly attributed to (a) integration of non-volatile RRAM on-chip, and (b) massive compute-to-memory access bandwidth enabled by monolithic 3D integration.

Figure 6 illustrates the breakdown of execution time and energy consumption for all workload/configuration pairs. 3D monolithic-TFT achieves the highest speedup thanks to the massive connectivity to memory offered by monolithic 3D integration—up to $4.8\times$ speedup versus 2D baseline, which is also $1.26\times$ that of 3D TSV-RRAM. Higher speedups can be achieved with more memory-intensive workloads, assuming that all weights can be stored on-chip in TFT-based RRAM.

3D monolithic-TFT improves energy consumption by up to $2.38\times$ compared to 2D baseline (Figure 6b) by reducing the idle energy consumption of compute and memory units, due to faster execution time enabled by higher connectivity to memory. Additionally, on-chip RRAM consumes less energy versus off-chip DRAM as RRAM arrays are smaller than that of DRAM and RRAM does not require refresh. 3D monolithic-TFT consumes only $\leq 5\%$ higher energy than 3D TSV-RRAM despite the higher RRAM access energy with TFT versus silicon (Figure 5). This is possible since memory-access energy is dominated by the array-to-compute interconnect for both 3D configurations, in addition to the reduced idle energy in 3D monolithic-TFT.

C. Perspective on higher EDP gains

For any workload, maximum EDP benefits—assuming Si-CMOS compute units—would be achieved when the entire

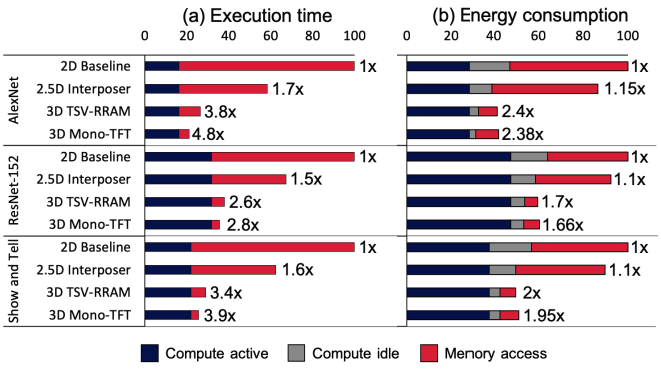

Figure 6: (a) Execution time and (b) energy consumption breakdown of each workload/configuration pair. Results are normalized to that of 2D baseline. We highlight speedup and energy reduction of each configuration versus 2D baseline.

Table IV: Parameters of the examined configurations

		2D Baseline	2.5D Interposer	3D TSV-RRAM	3D Monolithic-TFT
Compute units		4,096 8-bit multiply-and-accumulate processing units (PE), 500 MHz frequency, 0.48 pJ/operation			
Local SRAM		256 Bytes per PE, 2 ns latency, 0.23 pJ/bit energy			
Shared SRAM		2 MBytes, 4 ns latency, 0.32 pJ/bit energy			
Main Memory	Capacity and type	1-GB DRAM (LPDDR3)	1-GB DRAM (HBM)	54-MB-per-tier RRAM	8-MB-per-tier RRAM
	I/O count	32 pins	256 microbumps	1024 TSVs	131,072 ILVs
	Bandwidth (GB/s)	12	32	60	410
	Latency (ns)	70	50	Read: 3.5, Set/Reset: 51	Read: 10, Set/Reset: 54
	Energy (pJ/bit)	15	12	Read: 4.3, Set/Reset: 8.7	Read: 5.2, Set/Reset: 110/117

Table V: Examined deep-learning workloads

Network	Application domain	Number of weights
AlexNet [27]	Image classification	60 Million
ResNet-152 [28]	Image classification	60 Million
Show and tell [29]	Image captioning	75 Million

Table VI: System-level EDP benefits of targeted configurations

Benchmark	2D Baseline	2.5D Interposer	3D TSV-RRAM	3D Mono-TFT
AlexNet	1×	1.96×	9.1×	11.4×
ResNet-152	1×	1.65×	4.4×	4.65×
Show-and-Tell	1×	1.76×	6.8×	7.6×

execution time and energy consumption are consumed only in compute active (Figure 6). For the considered workloads (Table V), maximum EDP benefits are $22\times$ for AlexNet, $6.6\times$ for ResNet-152, and $12\times$ for Show and Tell. By comparing these benefits to those shown in Table VI, 3D monolithic-TFT is $1.4\times - 1.9\times$ lower than the maximum EDP benefits. While further improvements in device properties may improve EDP, it can yield diminishing returns for the considered workloads. However, significant improvements can be obtained with more memory-bound workloads, e.g., benefits can reach $200\times$ for a monolithic 3D system with CNFETs on top of a Si-FET bottom tier [1]—such memory-intensive workloads may require large amounts of on-chip memory that can be difficult to realize due to the low density of TFT-based RRAM. Thus, improving TFT density is key towards greater gains, and places TFT as an alternative enabler for monolithic 3D integration, in addition to emerging logic devices (e.g., CNFETs).

VII. Conclusion

Monolithic 3D computing systems promise significant system-level benefits over current baseline architectures. We have explored the potential of using TFTs in the upper tiers of a monolithic 3D system as memory-access circuitry and selector devices for RRAM. Our analysis shows that we can get significant energy-efficiency benefits and speedups over baseline and TSV-based 3D computing systems, despite the low TFT density. With further developments in TFT device characteristics, we can have higher density TFT-based RRAM. Thus, TFTs enable an alternative promising path towards realizing monolithic 3D systems that will require further investigations at the device, circuit and architecture levels.

VIII. Acknowledgement

This research is supported in part by the NTU Startup Grant (M4082035) and the NRF AME programmatic fund titles Hardware-Software Co-optimisation for Deep Learning (Project No.A1892b0026). We would like to thank E6NanoFab, NUS, and Chen Zhixian from Institute of Microelectronics (IME), A*STAR for providing the experimental measurements of TFT and RRAM, respectively.

References

[1] M. M. S. Aly et al. The N3XT approach to energy-efficient abundant-data computing. *Proc. IEEE*, 107(1), 2019.
[2] T. Tomimatsu et al. Cost-effective 28-nm LSTP CMOS using gate-first metal gate/high-k technology. In *Symposium on VLSI Technology*, 2006.
[3] S. Van Huylenbroeck et al. Small pitch, high aspect ratio via-last TSV module. In *ECTC*, 2016.
[4] M. M. Shulaker et al. Three-dimensional integration of nanotechnologies for computing and data storage on a single chip. *Nature*, 547, 2017.
[5] H. S. P. Wong et al. Metal-oxide RRAM. *Proc. IEEE*, 100(6), 2012.
[6] C. Wang et al. 3D monolithic stacked 1T1R cells using monolayer MoS_2 FET and hBN RRAM fabricated at low (150^oC) temperature. In *IEDM*, 2018.
[7] P. Batude et al. 3DVLSI with CoolCube process: An alternative path to scaling. In *Symposium on VLSI Technology*, 2015.
[8] K. Nomura et al. Thin-film transistor fabricated in single-crystalline transparent oxide semiconductor. *Science*, 300(5623), 2003.
[9] T. Naito et al. World's first monolithic 3D-FPGA with TFT SRAM over 90nm 9 layer Cu CMOS. In *Symposium on VLSI Technology*, 2010.
[10] T. T. Wu et al. High performance and low power monolithic three-dimensional sub-50nm poly Si thin film transistor (TFTs) circuits. *Nature Scientific Reports*, 2017.
[11] M. Gao et al. Tetris: scalable and efficient neural network acceleration with 3D memory. In *ASPLOS*, 2017.
[12] H. Park et al. RTL-to-GDS tool flow and design-for-test solutions for monolithic 3D ICs. In *DAC*, 2019.
[13] R. L. Hoffman. ZnO channel thin film transistor: channel mobility. *J. App. Phys.*, 95(10), 2017.
[14] B. Meyer et al. Binary copper oxide semiconductors: from materials towards devices. *Status Solidi (b)*, 249(8), 2012.
[15] K. Myny et al. *Organic and metal-oxide thin-film transistors*. Cambridge University Press, 2016.
[16] A. Grossi et al. Resistive RAM endurance: array-level characterization and correction techniques targeting deep learning applications. *TED*, 2019.
[17] W. Deng et al. A core compact model for IGZO TFTs considering degeneration mechanism. *IEEE TED*, 65(4), 2018.
[18] J. Fang et al. A surface-potential based DC model of amorphous oxide semiconductor TFTs including degeneration. *Electron Device Letters*, 2017.
[19] P. Chen and S. Yu. Compact modeling of RRAM devices and its applications in 1T1R and 1S1R array design. *IEEE TED*, 62(12), 2015.
[20] J. S. Kim et al. Dynamic logic circuits using a-IGZO TFTs. *TED*, 2017.
[21] Y-C. Yeo et al. MOSFET gate leakage modeling and selection guide for alternative gate dielectrics based on leakage considerations. *TED*, 2003.
[22] C-M. Zhang et al. Mobility degradation of 28-nm bulk MOSFETs irradiated to ultrahigh total ionizing doses. In *IEEE ICTA*, 2018.
[23] G. Hills et al. Understanding energy efficiency benefits of carbon nanotube field-effect transistors for digital VLSI. *IEEE TNANO*, 2018.
[24] X. Dong et al. NVSim: a circuit-level performance, energy, and area model for emerging nonvolatile memory. *IEEE TCAD*, 31(7), 2012.
[25] B. Mohammadi et al. A 128Kb 7T SRAM using a single-cycle boosting mechanism in 28nm FD–SOI. *TCAS-I: Regular Papers*, 2018.
[26] W. Hwang et al. 3D nanosystems enable embedded abundant-data computing: special session paper. In *CODES+ISSS*. ACM, 2017.
[27] A. Krizhevsky et al. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.
[28] K. He et al. Deep residual learning for image recognition. *CVPR*, 2016.
[29] O. Vinyals et al. Show and tell: a neural image caption generator. In *CVPR*, 2015.