Fledge: Flexible Edge Platforms Enabled by In-memory Computing

Kamalika Datta*, Arko Dutt*, Ahmed Zaky*, Umesh Chand†, Devendra Singh†, Yida Li†,
Jackson Chun-Yang Huang†, Aaron Thean†, Mohamed M Sabry Aly*
*Nanyang Technological University Singapore, †National University of Singapore
Email: kamalika.datta@ntu.edu.sg, arko001@e.ntu.edu.sg, ahmed.zaky@ntu.edu.sg, umesh.chand@nus.edu.sg,
devendra.singh@nus.edu.sg, li.yida@nus.edu.sg, jackson.huang@nus.edu.sg, aaron.thean@nus.edu.sg, msabry@ntu.edu.sg

***Abstract*—The proliferation of advanced analytics and artificial intelligence has been driven by huge volumes of data that are mostly generated at the edge. Simultaneously, there is a rising demand to perform analytics on edge platforms (i.e., *near-sensor data analytics*). However, conventional architectures of such platforms may not execute the targeted applications in an energy-efficient manner. Emerging near and in-memory computing paradigms can increase the energy efficiency of edge platforms by relying on emerging logic and memory devices. More importantly, these paradigms enable the possibility of performing computations on unconventional platforms, namely *flexible computing systems*. In this paper, we explore the benefits of in-memory computing at the edge on a flexible substrate enabled by thin-film transistors (TFTs) and resistive RAM (RRAM). As a case study, we consider bio-signal processing application workloads, i.e., compressive sensing and anomaly detection. We model the device, circuit, and architecture of our targeted platform and evaluate the corresponding system-level performance. Preliminary results indicate that in-memory computing enabled by flexible electronic devices enables a new class of edge platforms with lower power consumption, compared to that of rigid TFT devices.**

Keywords: RRAM, Thin-Film Transistor, Flexible Electronics, Edge Computing.

I. Introduction

Electronic and computing systems are ubiquitous nowadays and are deeply integrated in our daily activities. The abundance of these devices has led to the generation of huge volumes of data that have been the fuel of a new class of abundant-data applications. These applications apply complex analytics on such large amounts of data, e.g., artificial intelligence (AI), to provide new classes of services [1]. There is a rising demand to push these services on deeply embedded systems for more personalized AI applications, and perform *near-sensor data analytics*. A key feature for such systems is to execute corresponding workloads very close to the sensors. In this regard, *flexible electronics* can be a promising path.

Recent years have witnessed massive improvements in fabricating various electronic components using devices on flexible substrate [2], such as sensors [3], energy storage, and even new application frontiers (e.g., textile [4]). One of the most promising devices for flexible electronics, is thin-film transistor (TFT) [5], [6]. Various TFT technologies exist, like polycrystalline silicon based TFTs [6], oxide-based TFTs [7],carbon nanotube based TFTs [8], and organic TFTs [9]. These devices have weaker characteristics than conventional Si-CMOS—i.e., higher driving voltage, lower current, higher delays, and larger dimensions—which renders building a computing system with logic and memory devices a major challenge. Conventional computing systems separate both computing and memory and sequentially fetch instructions and data from memory to perform any operation. With the higher delays of flexible electronics, conventional computing may not be a feasible approach. Indeed, one has to leverage the benefits of new computing paradigms, that embraces parallelism and can merge computing and memory in very close proximity, to enable computing on flexible electronics.

Neuromorphic [10] and *in-memory* [11] computing significantly improve energy consumption by reducing (or totally eliminating) the data transfer from one memory array to compute units. These computing paradigms are proliferating nowadays as they are naturally enabled by new non-volatile memory technologies such as resistive RAM (RRAM), magneto-resistive RAM (MR-RAM), phase-change RAM (PCRAM), etc. [12]. These memory devices, i.e., RRAM, have been recently demonstrated on a flexible substrate [13], [14], which paves the way to integrate them with flexible TFTs and build a full neuromorphic computing unit.

In this paper, we explore the prospects of flexible neuromorphic computing systems and introduce a modeling methodology to analyze the system-level functionality and potential benefits. Figure 1 illustrates the introduced flexible computing system. Using device models calibrated by experimental measurements, we design all corresponding circuits where their characteristics (latency, power, area) are fed into a system-level simulation infrastructure to deduce the total execution time and energy consumption when running application workloads. We have explored two workloads for personalized healthcare to assess the feasibility of the proposed design, viz. *compressive sensing* and *anomaly detection* of an electrocardiograph (ECG) signal. Vector-matrix multiplication is a major operation for both workloads, which can naturally be adopted in neuromorphic computing system.

This paper starts in Section II by presenting a brief overview on the state-of-the-art in flexible device technologies, with specific emphasis on flexible TFT (FTFT) and flexible RRAM

978-3-9819263-4-7/DATE20/©2020 EDAA

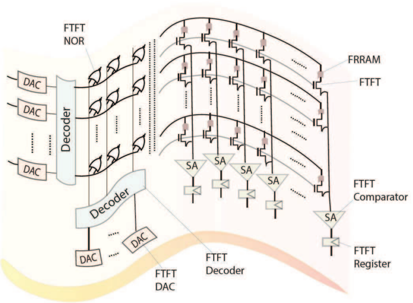

Figure 1: Flexible Computing System

(FRRAM) and device characteristics used in the design of the flexible system. In Section III, we discuss the architectural modeling of the TFT and RRAM array designed using flexible device models, and the corresponding customization carried out on the NVSIM tool [15] and MNSIM tools [16]. In Section IV we analyze area, latency and power of two applications. Finally, we conclude the paper in Section V.

II. Technology Enablers

In this section, we discuss the two considered devices, namely flexible thin film transistors (FTFT) and flexible RRAM (FRRAM), which are required for the targeted flexible computing system. Additionally, we briefly discuss recent demonstrations of computing circuits on a flexible substrate.

A. Flexible Thin Film Transistor (FTFT)

TFTs are fabricated by forming thin layers of materials on various substrates. For realizing bendable or flexible FTFT devices, we require flexible dielectric and semiconducting materials, as well as a flexible substrate on which the layers are formed. Various FTFT implementations exist, which are based on amorphous-Si [17], organic materials [18], oxides [19], and carbon nanotube [20].

Amporhous-Si based FTFTs can operate with a supply voltage of $1-10$V, a threshold voltage of 3.24V, field-effect mobility of 0.46cm^2/V.s, sub-threshold swing of 1.31V/dec, and on-off ratio of over 10^8. [17]. Organic FTFTs can have a higher mobility (1.0cm^2/V.s) and lower threshold voltage (0.53V) [18]. In this paper we use oxide-based FTFTs as they offer high carrier mobility, low process temperature, good transparency and good stability, compared to amorphous-Si and organic FTFTs. In particular, we model a ZnO-based FTFT (Figure 2a) based on measured experimental data. This FTFT uses a bottom-gate structure with layers of chromium (10nm) and gold (40nm) on a flexible polyimide substrate. On top of this, an Al_2O_3 layer is deposited as the gate dielectric. A 70nm-thick ZnO layer is created next, that forms the active channel. For the source and drain contacts, 10nm-thick titanium layer is used, followed by 30nm-thick platinum capping. The fabricated device has channel dimensions of $W = 25\mu$m, and $L = 20\mu$m. Figures 2(b) and 2(c) show the transfer and output characteristics of the device at drain voltage of $V_D = 1$V—tThe device exhibits a very high on/off current ratio of $> 10^8$.

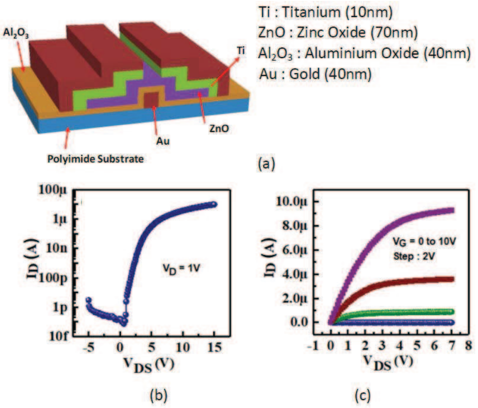

Figure 2: (a) Schematic of Flexible TFT, (b) Transfer characteristics, and (c) Output characteristics of the ZnO TFT device

In addition to oxide FTFTs, carbon nanotube based TFTs (CNT-TFT) provide superior carrier mobility and very good mechanical flexibility, and usually exhibit p-type characteristics [20]—we will consider CNT-TFT in our future work.

B. Flexible RRAM (FRRAM)

Similar to conventional RRAM, FRRAM is a metal-insulator-metal stack. FRRAM, however, is fabricated on a flexible substrate, but still retains similar properties of RRAM. A ITO/HfO$_x$/ITO based RRAM on a flexible polyethylene terephthalate (PET) substrate has an on/off resistance ratio of 40, set and reset voltages of 0.4V and 0.2V respectively, and good mechanical stability [21]. A transparent FRRAM with a TiO_2 dielectric, fabricated on flexible ITO/PET substrate, demonstrated a stable switching behavior (on/off ratio of 10, set/reset voltage of 2V) and good bending endurance [22]. In [13], authors use aerosol-jet-printed technology to create Ag/MoS_2/Ag RRAM cells in a crossbar structure, which exhibits low switching voltage (< 0.2V), high on-off resistance ratio (10^7), switching energy of 4.5fJ/bit. A multiple of such FRRAM cells were integrated to demonstrate a 4×4 crossbar, which can successfully withstand 1000 bending cycles under various bending radii.

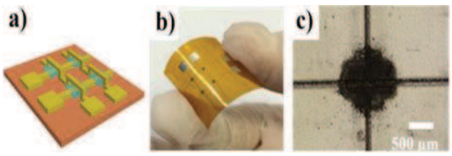

Figure 3: (a) Schematic diagram, (b) photo image, and (c) microscopic image of a 2×2 WSe_2 RRAM with Ag electrodes ($70\mu m$ line width) [14]

In this paper we use a FRRAM model using parameters from a WSe_2-based device fabricated using aerosol jet printing [14]. The device exhibits forming-free, unipolar behavior, < 1V switching voltage, on/off ratio of > 100, and set and reset voltages of 0.7V and 0.3V, respectively. This FRRAM uses Ag (silver) as the top/bottom contacts. An FRRAM array is also demonstrated. Figures 3(a)-(c) show the schematic diagram, a photo image, and a zoomed-in microscopic image of a fabricated 2×2 WSe_2 RRAM array. The size of a cell is basically limited by the width of the printed electrodes (70×70 μm for Ag) [14]. Figure 4a shows the DC-sweep characteristics for the modeled FRRAM with a set current of 2μA, where this FRRAM exhibits non-volatile behavior. Under this scenario, a unipolar behaviour is observed with a reset voltage (current) of 0.3V (80μA), that is, an operating power of 24μW (Figure 4(b)). A retention time in excess of 2.5 hours at room temperature is observed, when the device is in the stable LRS state, with > 100 on/off resistance ratio (Figure 4(c)).

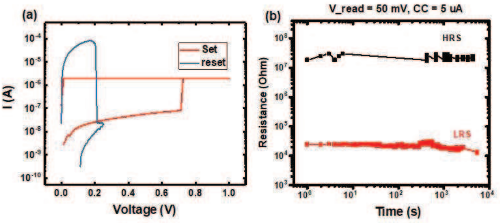

Figure 4: (a)DC sweep of WSe_2 RRAM with Ag electrodes at set current of $2\mu A$, (b) retention curve of the WSe_2 RRAM [14]

C. Flexible Electronics Circuits and Systems

In recent years, there has been an increase in the number of circuit demonstrations of flexible electronics.

Basic circuit constructs, such as ring oscillators, adders, and even data-storing elements such as SRAM have been demonstrated [23], [24], [25], [26]. The ring oscillator in [24] achieved 16ps delay with the accompanying SRAM operating at 0.6V, where both n-type and p-type FTFTs were fabricated using extremely thin silicon-on-insulator technology. Full systems in FTFTs may not require both types. For instance, the single-bit adder used single-wall p-type CNT-FTFT [25], whereas the 128-bit SRAM macro used n-type a-IGZO FTFTs [26] for the cell and all peripheral circuitry (both digital and analog). This SRAM consumes $100\mu W$ with read (write) latency of $280\mu s$ ($110\mu s$).

Additionally, various demonstrations showcased full small-scale systems built entirely on a flexible substrate. A signal conditioning and data-storing circuit was fabricated and integrated with a temperature sensor [27]. While this circuit was a simplistic one—a J-K flip-flop and an analog differential amplifier—it showcased the possibility of near-sensor processing. This can be attached to other bendable sensors that have been demonstrated in several applications, such as pH sensing [3], X-ray detector [28], and general biosensors [29]. Moy et al. [30] demonstrated a flexible electroencephalogram sensing and signal-processing systems on a flexible substrate. They have used an amplifier, followed by compressive sensing to scale down the acquired signal. Compressive sensing here is achieved via sampling and integration of the input signal. An in-memory computing circuit on a flexible substrate has been recently demonstrated [31] that implements an 8×8 crossbar array fabricated on a PES substrate, where the FRRAM had 10^7 on/off ratio, $10^5 s$ retention, and $3\text{V}/0.5\text{V}$ set/reset voltages.

While further improvements in flexible technology will enable scalable systems, it is important to evaluate—at design time—the operating conditions of such systems via proper modeling and simulation infrastructure that can leverage current device modeling frameworks (e.g., [32]).

III. Device and Architecture Modeling

To analyze a complete system using flexible electronics, we need to model both logic operations and memory functionality using flexible devices. We introduce an exploration framework, illustrated in Figure 5, that is tailored towards analyzing neuromorphic or in-memory flexible computing systems (Figure 6). The framework starts by modeling the considered devices, i.e., FTFT and FRRAM. In particular, we tune parameters of the IGZO-based flexible TFT model [33], based on experimental data, and extract the FRRAM parameters from the measured data of the fabricated device reported in [14]. We have also used device data of ZnO-based TFT device on a rigid substrate [34] and a RRAM device used in TFT-based RRAM memory macro design [35]. These device models are then used to design basic constructs, i.e., memory arrays, as well as digital and analog circuitry, that constitute the required modules in a full system. We then feed the characteristics of these modules (power, latency and area) to a system simulator that deduces the total energy and time of the examined computing system. We also leverage a memory simulator to rapidly estimate the access time and energy of various memory-block sizes.

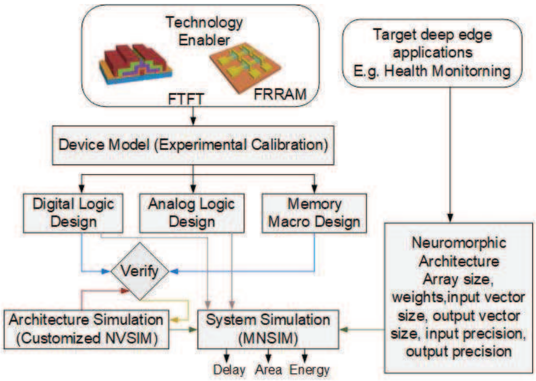
Technology Enabler
FTFT
FRRAM
Target deep edge applications E.g. Health Monitorning
Device Model (Experimental Calibration)
Digital Logic Design
Analog Logic Design
Memory Macro Design
Verify
Neuromorphic Architecture Array size, weights,input vector size, output vector size, input precision, output precision
Architecture Simulation (Customized NVSIM)
System Simulation (MNSIM)
Delay
Area
Energy

Figure 5: Our evaluation framework

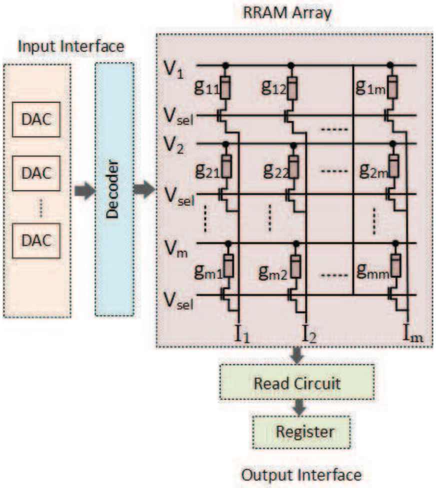

Figure 6: Memory macro for neuromorphic computing

A. Circuit Design

As shown in Figure 6, a neuromorphic circuit comprises a number of modules, which are as follows:

- memory array with 1T-1R cell structure
- access decoder and multiplexers
- a number of digital-to-analog converters (DAC)
- read circuit with a data-storing register

All circuits are designed using Verilog-a model of a rigid ZnO-based TFT [35], which only supports n-type devices. To account for FTFTs, we tune the results of this model using the measured data (Figure 2).

1) Memory array design: The basic memory array is designed using RRAM cells with access transistors in one-transistor-one-rram (1T1R) configuration. In this paper, we use two different cell structures, i.e., a rigid ZnO-based TFT stacked with HfO_2 RRAM [35], and FTFT adjacent to FRRAM (Section II). The peripheral circuits constitute access decoder, multiplexers, DACs, comparators and registers.

2) Access decoder and multiplexers: These components use a set of NOR gates, with a NOT gate at each NOR output [35]. Figure 7 shows the schematics of NOT and NOR (or NOR2 using two logic inputs) gates, using n-type TFTs only, where we have used the same transistor sizing as in [35].

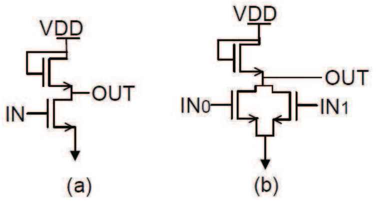

Figure 7: N-type TFT-based schematics of— (a) NOT, (b) NOR2

3) DAC and read circuitry: We build a comparator using n-type TFTs only, akin to the design introduced in [36], to be used in both DAC and to transfer the analog aggregated current for each bitline to a digital value. We use a fixed device length L=$1\mu m$, keeping W/L (device width-to-length) ratios of all TFTs same as the comparator design in [36]. As illustrated in Figure 8, the comparator cascades three amplification stages (each stage uses a differential amplifier circuit) to ensure sufficient DC gain in order to overcome offset of the next PFL stage (having positive feedback analog latch). The PFL stage, only used in the read circuit, enables speedup in comparison of inputs by using two n-type analog inverters adjacent to two analog latches, respectively. This stage also improves output regeneration by cross-coupling devices with inputs of analog inverters (for more details on circuit operation, please refer to [36]). The final stage that constitutes the four logic inverters implements a fully dynamic digital latch. The capacitors provide dynamic memory to the last set of inverters. The comparator is activated when enable, EN in Figure 8, goes high to low. It consumes a power of 1.84mW at $30\mu s$ latency, and has an area overhead of $1100\mu m^2$—results obtained from simulation on Cadence Virtuoso.

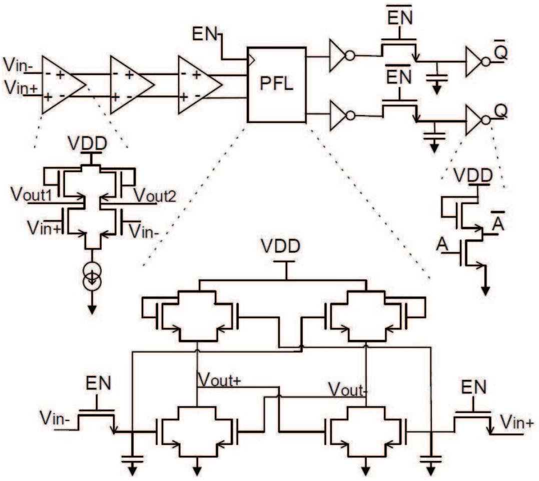

Figure 8: Comparator design constituting 3 amplifiers, a positive feedback analog latch (PFL), 4 inverters with a TFT-capacitor pair

B. Architectural Modeling

To enable rapid system exploration, it is important to quickly deduce the energy and latency of the module that can significantly vary among architectures, i.e., the memory array. We have modified NVSIM [15] and integrated it to our framework, and model various TFT (or FTFT) based RRAM macros. NVSIM [15] can estimate the access latency, energy and even area of non-volatile memories, assuming Si-CMOS devices for memory peripherals. However, it does not support TFTs, which have much larger technology feature sizes ($\geq 1\mu$m) or has only n-type or p-type device realizations. We summarize the alterations made to NVSIM below.

- We have added a new technology node needed to match the feature size of the ZnO-based TFT device (1μm) and ZnO-based FTFT device (25μm).
- We have modified the decoder design of TFTs to account that only n-type transistors are used. We altered the design to be based on NOR gates, instead of NAND gates. We have also tuned the gate sizing and delay.
- We have changed all the p-type pullup networks to n-type pullup networks as required in TFT based design.
- We have updated the formula to calculate various device capacitances consistent with TFT device geometries.
- We have included the calculations for static power that is predominant in TFT based static logic design.

We modelled various TFT- and FTFT-based macros. The ZnO-based TFT has carrier mobility of 8.5cm^2/Vs, dielectric oxide thickness of 35nm, sub-threshold swing of 592mV/dec, and threshold voltage of 4.039V. The corresponding values for the ZnO-based flexible TFT are 5.3cm^2/Vs, 40nm, 485mV/dec, and 7V respectively—the electrical parameters for the rigid TFTs and FTFTs do not vary significantly.

Results for various memory array sizes are summarized in Table I for both rigid and flexible devices. We have validated the results for the rigid devices against the corresponding results obtained through circuit simulation using Cadence Virtuoso with the Verilog-a device models.

C. System Simulation

For the system-level simulation of neuromorphic or in-memory flexible computing systems, we need to map the target applications to a computing fabric defined by the application needs. We have used MNSIM [16] to model and simulate the system, with proper customization as required for the devices used in our designs.

Figure 6 illustrates the full system simulated in MNSIM. DACs convert digital input signals to analog voltages that are fed into the RRAM array. The *read circuit* converts the column currents to voltages, digitizes them using threshold detectors, and stores the output vector in a register.

MNSIM models a multi-layer neural network by cascading a number of neuromorphic macros (Figure 6), each realized using one or more RRAM arrays, with necessary interfaces. A basic $m \times m$ 1T1R array (Figure 6) can be used to multiply a vector $V = \{v_1, v_2, \ldots, v_m\}$ with an $m \times m$ co-efficient matrix $G = \{g_{ij}\}$ to generate the result vector $I = \{I_1, I_2, \ldots, I_m\}$:

$$I_{1\times m} = V_{1\times m} \times G_{m\times m}$$

where v_i is an input analog voltage, g_{ij} is the conductance value of the RRAM cell in row i and column j, and I_i is the current flowing out of column i. The conductance values are often discrete, and may be initialized by applying *set* and *reset* pulses across the cell. Such in-memory vector-matrix multiplication is essential to realize neuromorphic operations. As shown in the figure, the voltages v_i are directly applied to the bit lines, and all the source lines of the cells in a column are wired together to aggregate the currents $I_k = \Sigma_{1\leq i\leq m}(v_i * g_{ik})$. Also, all the V_{sel} lines driving the access transistors are simultaneously enabled. Comparator circuits convert the aggregated current value to a digital signals as explained earlier.

Table I: Simulation results for various subarray sizes

Array Size	TFT [34] & RRAM [37]			FTFT [33] & FRRAM [14]		
	Latency (ns)	Power (mW)	Energy (pJ)	Latency (us)	Power (mW)	Energy (nJ)
8 × 8	1.48	26.8	39.6	11.30	11.0	124.0
16 × 16	2.75	29.9	81.8	29.70	8.3	247.0
32 × 32	2.91	60.8	177.0	42.60	13.3	569.0
64 × 64	3.11	108.6	338.0	90.00	11.12	1001.0

We have modified MNSIM to account for that targeted devices as follows.

- We have incorporated the array latency, power, and area from NVSIM.
- We have fed the area, latency and power consumption of the decoder and comparator circuits using the values obtained from circuit simulation in Cadence Virtuoso.

IV. Case Study:Health monitoring at the edge

We analyze two workloads related to health monitoring, viz. compressive sensing [38] and anomaly detection [39], which require vector-matrix multiplication during computation. In [38], a *sparse binary sensing* approach is reported, which is based on multiplying a vector with a sparse binary matrix containing at most d non-zero entries in each column. Experiments with MIT-NIH arrhythmia ECG database show an optimum value of $d = 12$ for matrices of size 512×64 (88% compression). We have carried out system-level simulation of a neuromorphic array of size 512×64, for both rigid and flexible TFTs. The results of simulation are summarized in Table II, which shows the area, latency and power consumption for the various constituent subsystems. FTFT-based system occupies reasonable footprint and execution time (despite being $> 100\times$ larger than rigid TFTs). Furthermore, FTFT consumes less power than TFTs.

We also consider another application in the area of anomaly detection, where a neuromorphic computing circuit performs cardiac arrhythmia analysis and classifies five different beat types, one normal and four anomalous, in real-time [39]. They use a 3-layer feed-forward neural network, which can be realized using two RRAM arrays of sizes 300×210 and 210×5 to achieve 91% classification accuracy. For simulating this system, we use four 512×64 arrays in the first layer, connected in cascade with another array of size 256×8. We simulate the two layers comprising of individual arrays, and consider the interfacing circuitry as well in the calculation. Table II shows the results for both rigid TFT and FTFT.

V. Conclusion

Flexible circuit technologies can support the ubiquity of computing systems. Despite the relatively weak performance of these devices versus commensurate devices on a rigid

Table II: Area, latency and power analyses

TFT	Metric	Array	Decoder	DAC	Read	Total
Workload 1: Compressive Sensing [38] (512×64)						
Rigid	Area (mm^2)	0.002	0.0056	0.563	0.070	0.64
	Time (μs)	0.003	0.0016	30.0	30.0	60.0
	Power (mW)	786	21	942	118	1870
Flexible	Area (mm^2)	16.4	0.557	140.8	17.6	175.29
	Time (μs)	105.0	85.2	106500	106500	213190
	Power (mW)	0.0033	1.54	68.6	8.58	78.7
Workload 2: Anomaly Detection [39] ($300 \times 210 \times 8$)						
Rigid	Area (mm^2)	0.008	0.0251	2.53	0.29	2.85
	Time (μs)	0.005	0.0026	60.0	60.0	120.0
	Power (mW)	3190	87.9	4240	531.0	8049
Flexible	Area (mm^2)	66.62	6.275	632.5	72.5	777.9
	Time (μs)	129.0	102.0	213000	213000	426231
	Power (mW)	0.0152	7.05	309.0	38.6	354.67

substrate, emerging in-memory computing paradigms and flexible logic and memory devices can both enable full flexible computing systems. We have explored the potential of neuromorphic computing using FTFTs and FRRAMs via a framework capable of simulating application workloads on systems that leverage experimentally-calibrated flexible device models. System analyses for two healthcare-targeted workloads—compressive sensing and anomaly detection in ECG signals—show that systems with flexible devices consume lower power than a rigid counterpart. Advancement in technologies for flexible applications will ensure improvement in area and performance, thereby enabling new frontiers in edge computing.

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, for providing the experimental measurements of FTFT and FRRAM, respectively.

REFERENCES

[1] Yann LeCun et al. Deep learning. *nature*, 521(7553):436–444, 2015.

[2] A. Nathan et al. Flexible electronics: the next ubiquitous platform. *Proc. IEEE*, 100, 2012.

[3] N. Liu et al. Flexible sensory platform based on oxide-based neuromorphic transistors. *Science Report*, 5, 2015.

[4] T. Carey et al. Fully inkjet-printed two-dimensional material field-effect heterojunctions for wearable and textile electronics. *Nature Communications*, 8(1202), 2017.

[5] P. K. Weimer. The TFT – a new thin film transistor. *Proc. IRE*, 50, 1962.

[6] Y. Kuo. Thin film transistor technology – past, present and future. *The Electrochem. Soc. Interface*, 2013.

[7] R. L. Hoffman. ZnO channel thin film transistor: Channel mobility. *J. App. Phys.*, 95(10), 2004.

[8] L. Shao et al. Compact modeling of carbon nanotube thin film transistors for flexible circuit design. In *DATE*, 2018.

[9] K. Myny et al. *Organic and metal-oxide thin-film transistors*. Cambridge University Press, 2016.

[10] H. Wu et al. Device and circuit optimization of rram for neuromorphic computing. In *International Electron Devices Meeting (IEDM)*, 2017.

[11] B. Chen et al. Efficient in-memory computing architecture based on crossbar arrays. In *IEDM*, 2015.

[12] P. L. Thangkhiew et al. Efficient mapping of boolean functions to memristor crossbar using MAGIC NOR gates. *IEEE TCAS-I*, 2018.

[13] X. Feng et al. A fully printed flexible MoS_2 memristive artificial synapse with femtojoule switching energy. *Advanced Electronic Materials*, 1900740, 2019.

[14] Y. Li et al. Aerosol jet printed WSe_2 based RRAM on Kapton suitable for flexible monolithic memory integration. In *FLEPS*. IEEE, 2019.

[15] X. Dong et al. NVSim: A circuit-level performance, energy, and area model for emerging nonvolatile memory. *IEEE TCAD*, 31(7), 2012.

[16] L. Xia et al. MNSIM: Simulation platform for memristor-based neuromorphic computing system. *IEEE TCAD*, 37(5), 2018.

[17] J. K. Lee et al. a-Si: H thin-film transistor-driven flexible color E-paper display on flexible substrates. *IEEE Electron Device Lett.*, 31(8), 2010.

[18] K. Fukuda et al. Fully-printed high-performance organic thin-film transistors and circuitry on one-micron-thick polymer films. *Nature Communications*, 5, 2014.

[19] S. Jeong et al. Bendable thin-film transistors based on sol-gel derived amorphous Ga-doped In_2O_3 semiconductors. *Superlattice Microstructures*, 59, 2013.

[20] H. Wang et al. Tuning the threshold voltage of carbon nanotube transistors by n-type molecular doping for robust and flexible complementary circuits. In *Proc. National Academy of Sciences*, 2014.

[21] J. Shang et al. Highly flexible resistive switching memory based on amorphous-nanocrystalline hafnium oxide films. *Nanoscale*, 2017.

[22] K. N Pham et al. TiO_2 thin film based transparent flexible resistive switching random access memory. *Nanoscience and Nanotech.*, 7, 2016.

[23] J. Tang et al. Flexible cmos integrated circuits based on carbon nanotubes with sub-10 ns stage delays. *Nature Electronics*, 2018.

[24] D. Shahrjerdi et al. Advanced flexible cmos integrated circuits on plastic enabled by controlled spalling technology. In *IEDM*, 2012.

[25] J. Noh et al. Fully gravure-printed flexible full adder using swnt-based tfts. *IEEE Electron Device Letters*, 33(11):1574–1576, 2012.

[26] F. De Roose et al. A thin-film, a-igzo, 128b sram and lprom matrix with integrated periphery on flexible foil. *IEEE JSSC*, 52(11), 2017.

[27] W. Honda et al. Bendable cmos digital and analog circuits monolithically integrated with a temperature sensor. *Advanced Materials Technologies*, 1(5):1600058, 2016.

[28] J. T. Smith et al. Optically seamless flexible electronic tiles for ultra large-area digital X-ray imaging. *IEEE Tran. Compon. Packag. Manuf. Tech.*, 4(6), 2014.

[29] S. Shah et al. Biosensing platform on a flexible substrate. *Sensors and Actiators, B: Chem.*, 210, 2015.

[30] T. Moy et al. 16.4 a flexible eeg acquisition and biomarker extraction system based on thin-film electronics. In *ISSCC*, pages 294–295, 2016.

[31] B. Jang et al. Zero-static-power nonvolatile logic-in-memory circuits for flexible electronics. *Nano Research*, 10(7):2459–2470, Jul 2017.

[32] A. Vilouras et al. Modeling of cmos devices and circuits on flexible ultrathin chips. *IEEE Transactions on Electron Devices*, 64(5), 2017.

[33] H-H. Tsu et al. A flexible IGZO thin-film transistor with stacked TiO_2-based dielectricc fabricated at room temperature. *IEEE Electron Device Lett.*, 34(6), 2013.

[34] W. Deng et al. A core compact model for IGZO TFTs considering degeneration mechanism. *IEEE TED*, 65(4):1370–1376, April 2018.

[35] A. Felfel et al. Quantifying the benefits of monolithic 3D computing systems enabled by TFT and RRAM. In *DATE*, 2020.

[36] A. Correia et al. Design of a robust general-purpose low-offset comparator based on igzo thin-film transistors. In *(ISCAS)*, 2015.

[37] P. Chen and S. Yu. Compact modeling of RRAM devices and its applications in 1T1R and 1S1R array design. *IEEE TED*, 62(12), 2015.

[38] H. Mamaghanian et al. Compressed sensing for real-time energy-efficient ECG compression on wireless body sensor nodes. *IEEE Trans. Biomedical Engg.*, 58(9), 2011.

[39] A. M. Hassan et al. Real-time cardiac arrhythmia classification using memristor neuromorphic computing system. In *EMBC*, 2018.